UNIVERSIDADE ABERTA

ANÁLISE ESTATÍSTICA MULTIVARIADA DE UMA BASE DE VINHOS NO AMBIENTE COMPUTACIONAL RSTUDIO UTILIZANDO ANÁLISE DE COMPONENTES PRINCIPAIS

Geraldo Lima Almendra Filho

Dissertação de Mestrado em Estatística, Matemática e Computação

Orientadora: Prof.ª Doutora Catarina S Nunes

ABRIL - 2021

AGRADECIMENTOS

Aos professores e professoras da Universidade Aberta de Portugal pelo apoio e orientações durante o Curso de Estatística, Matemática e Computação, pelos conhecimentos transmitidos, assim como pelo rigor acadêmico sempre exigido.

À minha orientadora Doutora Catarina S Nunes pela dedicação na orientação do meu trabalho sem poupar esforços em me alertar sobre todos os pontos que deveriam ser reavaliados tanto na essência como na forma de apresentação.

À minha esposa e melhor amiga Ana Maria um agradecimento muito especial por ter me incentivado nos estudos, sempre me cobrando seriedade e disciplina e, sobretudo, pela paciência com minha dedicação ao Mestrado que subtraiu dezenas de horas do nosso convívio familiar.

Aos meus filhos Vinícius e Luciana que sempre foram meus maiores incentivos nas minhas lutas na vida para justificar perante Deus a responsabilidade dada para participar da construção humana de pessoas tão especiais.

Dados Internacionais de Catálogo de Publicação (CIP)

Almendra, Geraldo Lima Filho

Análise estatística multivariada de uma base de vinhos no ambiente computacional RStudio utilizando Análise de Componentes Principais – Petrópolis, 2022

Orientadora: Prof.ª Doutora Catarina S Nunes

Dissertação (Mestrado em Estatística, Matemática e Computação com especialização em Estatística Computacional)

Universidade Aberta de Portugal

Inclui bibliografia

1. Análise Multivariada de uma base de vinhos.

2. Linguagem RStudio

ABSTRACT

This work aims at the study of a database with 1599 red wine records through multivariate statistics using Principal Components Analysis (PCA), and an assessment of relations between wine components and quality measured by a sensory evaluation by tasting specialists.

PCA is a method of multivariate analysis that uses a linear transformation to reduce the dimension of databases, transforming large sets of variables with its associated instances (data collected), in non-correlated subsets. These subsets will form alternative groups of the original variables defining new variables (main components) with the same data input records, to analyze and explain the data total variability through new components or factors, which are defined as linear combinations of the original variables that influence their behavior, this is the main objective of the PCA.

The merit of dimensionality reduction is to change a little precision by mainly simplicity for the ease of explore, visualize, analyze, and explain, when we choose few variables that explain the largest parcels of data variability.

The database does not have the names of the wine brands, these are substituted by numeral labels in an increasing order, which allows us to perform a multivariable analysis using PCA.

The wine database is from the study of P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis [1] (*csv* file), containing de data associated with 12 variables – chemical wine attributes – and a categorical variable specifying the type of wine.

Our study focusses on descriptive statistical evaluations (in a PCA context) and also explores the relation between the wine quality and changes in the metrics of the attributes. This is one of the objectives, since one cannot correlate the wine brands with the grades obtained by the sensorial evaluation.

We shall see that the quality, taste and other wine characteristics from the sensorial evaluation [3] (Parts 8 and 9), are correlated in some degree with changes in the wine attributes metrics – chemical attributes: (i) are highly correlated with the alcohol level and (ii) are in a smaller extent correlated with chlorides, sulfates, density, fixed acidity and citric acid.

Regarding the principal components (the new variables associated with the wine records) required to statistically evaluate the data base (on a multivariate perspective), this work demonstrates that only 4 principal components are responsible for 69% of the explained variability, which validates the use of PCA to analyze the data.

All mathematical and statistics concepts required for the development of this work will be explained throughout the text, as necessary. However, we should emphasize that the algorithms (packages) of the software RStudio have embedded the necessary programs and calculations for the development of PCA.

This dissertation achieves its objectives, since it was possible to develop a PCA with the proposed database and to use in detail the RStudio software (and its packages) in a multivariate analysis.

Keywords: PCA, Multivariate Analysis, Wine Quality, UCI Data Set, RStudio an Linear Algebra.

RESUMO

Análise de Componentes Principais (Principal Component Analysis – PCA), é um método de Análise Multivariada que utiliza uma transformação linear para reduzir a dimensão de bancos de dados, transformando grandes conjuntos de variáveis com suas instâncias associadas (dados coletados), em subconjuntos não correlacionados com agrupamentos alternativos das variáveis originais, formando novas variáveis com os mesmos registros de entrada de dados visando analisar e explicar a variação total dos dados através de componentes ou fatores, que são definidos como combinações lineares das variáveis originais que mais influenciam o seu comportamento.

O mérito da redução da dimensionalidade, é trocar um pouco de precisão por, principalmente, simplicidade, pela facilidade de explorar, visualizar, analisar e explicar, ao escolhermos poucos CPs (Componentes Principais) que explicam as maiores parcelas de variabilidade dos dados.

A partir de uma base de dados que contém 1599 registros de vinhos tintos (Data Folder da fonte [1]), vamos reagrupar suas variáveis, eliminando as redundâncias ou semelhanças por força das correlações identificadas no conjunto de dados.

O objetivo é formar subconjuntos não correlacionados, ou com mínimas correlações e alta dispersão o que, naturalmente, vem às custas relativas da precisão analítica pontual.

Como o banco de dados não informa os nomes de *rótulos (marcas dos vinhos) por decisão do autor da pesquisa,* estes serão substituídos por *"rótulos numéricos"* em ordem crescente para melhor conduzir e auxiliar a Análise Multivariada através da PCA.

A base citada tem como origem o estudo de P. Cortez, A. Cerdeira, F. Almeida, T. Matos e J. Reis [1] e apresenta em arquivo *csv* os registros de dados associados a 12

variáveis - atributos químicos dos vinhos – e a uma variável categórica, no caso o tipo de vinho tinto.

Nosso estudo contém, a par de avaliações estatísticas descritivas no contexto do PCA, uma abordagem das relações da qualidade com mudanças nas métricas dos atributos como um dos objetivos do projeto já que não podemos correlacionar *"marcas formais"* mas sim suas instâncias com as notas obtidas na avaliação sensorial.

Veremos que a qualidade, sabor e outras características, caracterizadas através de avalição sensorial feita [3] por um grupo de provadores, estão correlacionados em maior ou menor grau com alterações métricas nos atributos – componentes químicos: (i) altamente correlacionados com o grau alcoólico e (ii), em menor escala, com os atributos cloretos, sulfatos, densidade, acidez fixa e ácido cítrico.

Em termos dos CPs necessários para avaliar estatisticamente de forma multivariada a base de dados, será demonstrado que apenas 4 CPs resultantes do processo PCA detém cerca de 69 % das variações explicadas validando o uso do modelo PCA para analisar estatisticamente e de forma multivariada a base de dados.

Todos os conceitos matemáticos e estatísticos necessários para o desenvolvimento deste trabalho foram expostos ao longo do texto, devendo-se enfatizar que os algoritmos (*packages*) da linguagem RStudio trazem embutidos as rotinas de cálculos necessárias para o desenvolvimento do processo PCA.

Como será descrito nos Comentários Conclusivos, os objetivos da Dissertação foram atingidos pois foi possível desenvolver o processo PCA sobre a base de dados escolhida, para além da utilização detalhada do RStudio e seus "packages" para esta finalidade de Análise Multivariada, como pode ser visto no script do **Anexo I.**

Termos Chaves: PCA, Análise Multivariada, Wine Quality, UCI Data Set, RStudio e Álgebra Linear.

ÍNDICE

Parte da bibliografia teve utilização direta com replicações e/ou adaptações de texto. Outra parte foi base para estudos conceituais que fundamentaram o desenvolvimento da Dissertação.

ABREVIATURAS

PCA – Principal Componentes Analysis

CP ou PC – Componente Principal

CPs ou PCs – Componentes Principais

LISTA DE FIGURAS

LISTA DE QUADROS COM TABELAS

INTRODUÇÃO – MOTIVAÇÃO E JUSTIFICATIVA DA ESCOLHA DO OBJETO DA DISSERTAÇÃO

A motivação principal de explorar a *Análise de Componentes Principais* é fortalecer um conhecimento estatístico visando um aprofundamento de estudos na *Análise Multivariada* em todo o seu processo, objetivando o exercício de atividades profissionais e acadêmicas utilizando a linguagem R como base inicial de apoio computacional para a análise estatística de banco de dados.

O advento da computação moderna, de grande capacidade de armazenamento e velocidade de processamento de cálculo matemáticos complexos, é que permitiu, a partir da década de 30, o nascimento de linguagens e algoritmos próprios para o desenvolvimento de programas estatísticos para processar, analisar e interpretar grandes massas de dados com incrível velocidade, capacidade analítica e gráfica.

A linguagem R com seu interface (IDE) RStudio vem crescendo exponencialmente sua utilização na área estatística, motivando dezenas de especialistas e programadores a desenvolverem pacotes para a execução de tarefas complexas na área de processamento de dados, que necessitam de cálculos matemáticos e estatísticos de grande complexidade exigindo algoritmos cada vez mais sofisticados.

Para abordagens mais amplas no contexto da Análise Multivariada, outros procedimentos podem ser executados: Análise de Agrupamentos (Cluster), Análise Fatorial e Análise Discriminante conforme o objetivo a ser atingido.

Será necessário abordar dois algoritmos do RStudio no contexto da análise fatorial para validar a utilização de CPs na base utilizada. Algumas passagens pela Estatística Descritiva de qualificação de dados, por motivos pontuais, serão necessárias.

A *Estatística Multivariada* se utiliza de métodos estatísticos que permitem a análise de banco de dados com múltiplas variáveis medidas simultaneamente para cada registro amostral, o que não seria possível, ou muito complexo, fazer por métodos de Estatística Univariada para análise individual ou aos pares de variáveis.

Figura 1 – Principais Métodos de Avaliação de Componentes de uma Base de Dados Multivariados [Tradução e Adaptação de [10]]

A PCA é um método de análise estatística multivariada desenvolvido originalmente por Karl Pearson em 1901 e trata da *modelagem da estrutura de variância* (dispersão estatística), da *covariância* (interdependência ou relação linear entre variáveis) e da *correlação* (força de associação e direção do relacionamento linear entre duas variáveis).

Os *métodos computacionais* avançados e práticos para o cálculo da PCA somente surgiram bem mais tarde com os trabalhos desenvolvidos por Hotelling (1933,1936) que usou essa técnica estatística para estudar as *estruturas de correlação* de uma base de dados [45 - 48]. O RStudio foi desenvolvido como uma IDE (Integrated Development Environment) com código aberto para o desenvolvimento integrado de um ambiente estatístico e computacional na linguagem R que permite a contribuição de uma comunidade de estatísticos e programadores profissionais.

Além da análise multivariada através da PCA vamos relacionar a qualidade - *medida por uma avaliação sensorial* - com as métricas dos componentes químicos dos vinhos.

Como síntese da abordagem da Dissertação temos:

Começamos na **Parte 1** descrevendo aspectos da produção de vinhos com uma descrição da base de dados; na **Parte 2** fazemos uma explanação sobre a metodologia da Dissertação; nas **Partes 3 e 4** fundamentamos conceitualmente a PCA e abordamos seu processo de desenvolvimento; nas **Partes 5 e 6** procedemos à importação da base de dados e fazemos testes de adequação da base ao processo PCA; uma abordagem que procura caracterizar as variáveis recorrendo ao cálculo de medidas de estatística descritiva resumida dos dados é feita na **Parte 7**; os aspectos qualitativos dos vinhos e as relações com seus atributos é feita na **Parte 8**; uma análise gráfica do comportamento da qualidade é feita na **Parte 9**; a aplicação do processo PCA na base de dados, um resumo de suas etapas e suas execuções por pacotes estatísticos são feitos nas **Partes 10, 11 e 12**; fechamos as etapas com a **Parte 13** que apresenta uma análise gráfica das CPs, dos atributos e das correlações. Terminamos a Dissertação com os Comentários Conclusivos.

PARTE 1 – SOBRE O VINHO E ASPECTOS DESCRITIVOS DA BASE DE DADOS

"O **vinho** é produzido através da fermentação das uvas (esmagadas) por ação das leveduras (fungo responsável pela fermentação), que transformam o açúcar da fruta em álcool e dióxido de carbono. Embora a Saccharomyces Cerevisiae não seja a única levedura envolvida na produção de vinho, é a mais importante. [1, 2, 3 e 4]

Cumpre destacar o comentário extraído do trabalho de Paulo Cortez [1] para melhor entendimento do contexto factual da análise estatística multivariada aplicada sobre o seu *"Data Set Wine da UCI - archive.ics.uci.edu"* [2] disponibilizado para pesquisa pública: (Texto traduzido do inglês):

"Antes considerado um bem de luxo, hoje o vinho é cada vez mais apreciado por um leque mais amplo de consumidores. Portugal é um dos dez principais exportadores de vinho com 3,17% da quota de mercado em 2005. As exportações do seu vinho verde (da região noroeste) aumentaram 36% entre 1997 e 2007. Para apoiar o seu crescimento, a indústria do vinho está investindo em novas tecnologias para os processos de vinificação e venda. A certificação de vinhos e a avaliação da qualidade são elementos fundamentais neste contexto. A certificação evita a adulteração ilegal de vinhos (para salvaguardar a saúde humana) e garante qualidade para o mercado vitivinícola.

A certificação de vinhos é geralmente avaliada por testes físico-químicos e sensoriais. Os testes de laboratório físico-químico usados rotineiramente para caracterizar o vinho incluem a determinação dos valores de densidade, álcool ou pH, enquanto os testes sensoriais dependem principalmente de especialistas humanos. Ressalta-se que o paladar é o menos compreendido dos sentidos humanos, portanto a classificação de vinhos é uma tarefa difícil.

Além disso, as relações entre a análise físico-química e sensorial são complexas e ainda não totalmente compreendidas."

Parte 1.1 – Visão geral ilustrativa do formato matricial da base de dados

Na figura 2 vemos uma ilustração de um Data Frame padrão como um exemplo de um *espelho visual* dos dados a serem coletados da base original para importação em arquivo *excel.csv.* para a execução do processo PCA no ambiente RStudio:

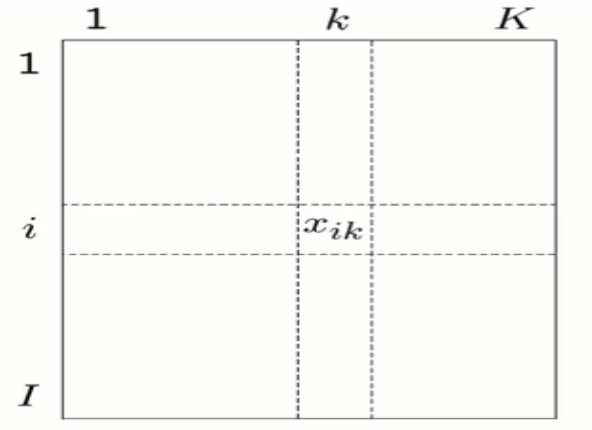

Figura 2 – Modelo de um Data Frame para gerar uma matriz [26].

- Dado um conjunto $\{x_i\}_i^n$

- Sendo ***k*** o nome das variáveis/atributos considerados, e

- $\boldsymbol{x_{ik}}$**,** sendo ***i*** o valor dos dados desses atributos associados a uma amostra de vinho (como por exemplo um *rótulo fictício* também denominado *instância* ou *indivíduo*).

- cada observação x_i tem *k* ou *p* dimensões sendo $X_i = (X_{i1}, X_{i2}, \ldots\ldots, X_{ip})$

Principais características do banco de dados importado da fonte consultada:

- Website - https://archive.ics.uci.edu/ml/datasets/wine+quality http://www3.dsi.uminho.pt/pcortez/wine/ - CSV Files – Wine Red and White Quality Data Sets.

- Banco de dados: Multivariado.

- Atributos: Quantitativo real.

- Tarefas associadas: Análise Multivariada e classificação.

- Número de observações total: 1599.

- Número de observações sem duplicidades: (verificado no estudo).

- Número de atributos (variáveis definidas pelo autor da pesquisa consultada): 13 incluindo o tipo de vinho (variável categórica).

- Data de publicação: 07/10/2009.

Os dados têm como referência as variantes tinto (Vinho Verde**)** português.

Na terminologia estatística/PCA as variáveis/indivíduos (atributos) de 1 a 12 serão consideradas quimicamente *independentes* e *ativas* para as análises.

A numeração das linhas da matriz ou data frame representativo da base de dados são "rótulos/labels fictícios" (marcas) não explicitados conforme critério do levantamento/pesquisa original:

> *"Devido a questões de privacidade e logística, apenas variáveis físico-químicas (entradas) e sensoriais (a saída) estão disponíveis (por exemplo, não há dados sobre os tipos de uva, marca de vinho, preço de venda do vinho, etc.)."* Paulo Cortez.

Parte 1.2 – Denominação técnica dos atributos químicos (variáveis) componentes dos vinhos

Variáveis constantes com seus nomes originais e unidades de medida (métricas) [1]:

1	- fixed acidity (acidez fixa)	(g(tartaric acid)/dm3)
2	- volatile acidity (acidez volátil)	(g(acetic acid)/dm3)
3	- citric acid (ácido cítrico)	(g/dm3)
4	- residual sugar (açúcar residual)	(g/dm3)
5	- chlorides (cloretos)	(g(sodium chloride)/dm3)
6	- free sulfur dioxide (fsd)	(mg/dm3)
7	- total sulfur dioxide (tsd)	(mg/dm3)

8 - density (densidade) (g/cm3)

9 - pH escala de 0 a 14

10 – sulfates (sulfatos) (g(potassium sulphate)/dm3)

11 - alcohol (grau alcoólico) (% vol.)

Variável de Saída (baseada nos dados sensoriais):

12 - quality (qualidade) (classificação entre 0 and 10)

13 – tipo (red wine)

PARTE 2 – **INTRODUÇÃO GERAL SOBRE A METODOLOGIA DA DISSERTAÇÃO**

Como a Dissertação, conforme o ***Plano de Estudos,*** *tem como meta tratar da utilização do* ***RStudio*** *para fazer a Análise dos Componentes Principais, ressaltando-se seus principais fundamentos matemáticos e estatísticos do processo*, alguns dos comandos contidos no *RScript* serão copiados ao longo do texto – à exceção de comandos com muitas linhas que ficarão registrados apenas no *RScript*.

Este caminho foi escolhido para não comprometer a estética do trabalho, mas apresentando de forma seletiva o recurso computacional para que seja atingido plenamente os objetivos gerais da Dissertação: apresentar a metodologia estatística e matemática que fundamentam a análise multivariada através do *software RStudio* e seus recursos em uma aplicação PCA.

Para complementar nossas análises - além do processo PCA - estudaremos as relações entre as notas obtidas na avalição sensorial e a métrica de cada atributo, interpretando tendências de mudança de qualidade conforme variações quantitativas nos atributos individuais.

Para essa finalidade serão analisados os box plot (**Parte 8 e 9**) individuais de cada atributo e sua relação com as notas obtidas na avaliação sensorial de todos os 1599 "*rótulos numéricos*".

Poderemos, também, associar estatisticamente variações de qualidade dos vinhos amostrados conforme mudanças métricas nos seus atributos através de avaliações das correlações identificadas entre os componentes dos vinhos e a qualidade.

Vamos pressupor que para confirmar as tendências de alterações de qualidade em outros tipos de vinhos, resultantes de mudanças planejadas nas métricas dos atributos, seria necessária uma nova avaliação sensorial com a atribuição de novas notas, o que poderia refletir as mudanças na qualidade.

Não consideramos este processo adicional como objetivo da Dissertação já que não temos acesso a condições laboratoriais para fazer esta abordagem.

Como já existem estudos químicos definindo padrões de efeitos de variações métricas dos atributos na composição de novos tipos de vinhos, provocando alterações no sabor resultante, vamos nos ater apenas nas tendências de variação qualificada pela provável nota de uma nova teórica avaliação sensorial já efetuada, mas apenas para efeito estatístico.

Somente a comparação das notas com todas as amostras, analisando cada atributo separadamente, será o indicador escolhido para estudar as relações dos atributos com as notas obtidas.

O ***tipo de vinho*** será uma variável categórica única *red wine* ou *vinho tinto.* As outras variáveis serão consideradas atributos com maior ou menor correlação com a qualidade e com outros atributos.

A *avaliação sensorial* é influenciada pelas características físicas e químicas de 11 variáveis a que denominamos de *atributos componentes dos vinhos***.**

Não estará em julgamento a *qualidade da análise sensorial original* partindo-se do pressuposto que o processo de degustação foi qualificado como profissional e correto.

PARTE 3 – FUNDAMENTAÇÃO CONCEITUAL DA ANÁLISE DOS COMPONENTES PRINCIPAIS

"*O objetivo principal do PCA é o de explicar a estrutura de variância e covariância de um vetor, composto de p-variáveis aleatórias, através da construção de combinações lineares das variáveis originais. Estas combinações lineares são chamadas de componentes principais. A informação contida nas p-variáveis originais é substituída pela informação contida em k (k<p) componentes principais não correlacionadas."* [5]

A PCA se trata de uma metodologia estatística, que se utiliza amplamente da Álgebra Linear associada a processos estatísticos, para que um conjunto de atributos contidos nos registros individuais da base de dados sejam representados por meio de matrizes que serão a base do processo da Análise Multivariada.

"Na Análise de Componentes Principais serão desenvolvidas técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados". [5]

Conforme anteriormente citado, o PCA abrange técnicas de normalização de dados e redução de dimensionalidade, que são rotinas executadas pelos comandos do RStudio.

Premissas básicas necessárias para a execução do processo PCA:

- Os dados originais estão representados por características ou atributos – variáveis correlacionadas.

- O objetivo é transformar essas variáveis em novas variáveis através de uma mudança de base do espaço vetorial, sendo que essas variáveis novas chamadas de componentes principais não sejam correlacionadas e retenham em *ordem decrescente* a maior parte da variação das variáveis originais.

Pontos críticos a serem enfatizados sobre o processo PCA:

1) Com o PCA a interpretabilidade adequada do banco de dados será viabilizada com

a *criação de novas variáveis*, subconjuntos denominados *componentes principais* que representam *combinações lineares de todas as variáveis com pesos diferentes* consideradas para caracterizar estatisticamente as amostras coletadas, mas que contêm todos os rótulos originais apresentados em uma nova forma de dispersão;

2) O PCA evita a multicolinearidade (alta correlação entre variáveis independentes) conduzindo a análise à *compreensão das relações entre as variáveis e à identificação dos padrões ocultos na base de dados*;

3) É necessário um teste confirmatório de esfericidade e adequação da amostra para um processo PCA;

4) Em termos de expectativa de avaliação sensorial**,** a qualidade é uma nota entre 0 (a pior) e 10 (a melhor) para cada registro de vinho.

O atributo qualidade será mais explorado no PCA com a análise descritiva dos efeitos dos *outliers e das correlações da qualidade com os atributos componentes dos vinhos.*

Explorando o contexto do comentário de Paulo Cortez na página **Wine Quality Data Set da UCI**: *"Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods."*

*("Algoritmos de detecção de outlier poderiam ser usados para detectar poucos vinhos excelentes ou ruins. Além disso**,** não temos certeza se todas as variáveis de entrada são relevantes. Portanto, pode ser interessante testar métodos de seleção de recursos"). (não utilizados no nosso estudo).*

A verificação da existência de *outliers* de atributos que não tenham uma boa correlação com o item qualidade poderá ser um bom indicativo a priori de má qualidade do vinho.

Ao longo da Dissertação *não serão desenvolvidas* provas ou demonstrações complexas da base teórica matemática e estatística dos instrumentos computacionais de avaliação estatística implícitos nos comandos e funções do *RStudio*, mas sim demonstrações objetivas de suas aplicações e interpretações no contexto das

utilizações de recursos matemáticos e estatísticos pelos pacotes da linguagem RStudio.

Contudo, havendo necessidade expositiva, algumas fórmulas e conceitos importantes poderão ser mais enfatizadas/explicadas de forma mais específica seguindo as fontes consultadas e citadas.

O Script R – **Anexo I** - é parte fundamental da execução do projeto de Dissertação. Sendo assim, serão feitas ao longo do texto referências às linhas de comando utilizadas para obter os resultados descritos nas análises. *Contudo algumas instruções e comentários não serão transcritos do R para o texto por não terem, aplicabilidade para o entendimento do mesmo mas apenas como recurso técnico de elaboração do modelo/algoritmo de análise PCA.*

As imagens dos *outputs* dos comandos e funções do RStudio mais importantes colocados no texto, foram copiadas diretamente do ambiente RStudio, sendo preservados, preferencialmente, à exceção de legendas alteradas em alguns gráficos, suas simbologias e descrições naturais no processo computacional, sendo desenvolvido na língua inglesa.

PARTE 4 – INTRODUÇÃO AO DESENVOLVIMENTO DO PROCESSO PCA

O processo PCA se inicia com a definição de uma base matricial que considera um vetor coluna aleatório $X\ (x_1, x_2, \ldots, x_n)$ formado por todas as instâncias x_i. No nosso caso representados por um base de vinhos com seus 1599 rótulos, em que cada um dos rótulos está associado a um conjunto de atributos = variáveis.

No PCA as variáveis são geralmente padronizadas como rotina interna do RStudio com o argumento *scale* quando utilizamos o package *FactoMineR*. Isto é particularmente recomendado quando as variáveis são medidas em diferentes escalas. Em alguns casos podemos querer redimensionar os dados quando a média e/ou o desvio padrão das variáveis são em grande parte muito diferentes. O objetivo é fazer escalas comparáveis.

Após os dados normalizados serem associados em um novo sistema de coordenadas seguido de um processo de rotação [19], são definidos *eixos ortogonais* para configurar independência entre os componentes que ficam perpendiculares entre si.

A redução da dimensionalidade nos conduz ao menor número possível de fatores novos cada um representando uma equação (tratada na **Parte 10**) que reflete as combinações lineares formadoras das CPs [7,9].

O PCA gera no seu processo quantas componentes (equações) forem suas variáveis. Esses componentes são subconjuntos das variáveis originais associados às amostras e que irão representar as CPs não correlacionadas (independentes), objetivando explicar uma parcela relevante e suficiente do conteúdo informativo e variacional dos dados originais.

Conforme acima mencionado, por detrás de cada componente principal se encontra uma combinação linear das variáveis e seus pesos *(loadings* da matriz de covariância) associados a todas as variáveis.

Em relação à *perda de precisão*, citada na introdução, espera-se, que seja plenamente compensada pela capacidade estatística explicativa obtida, mesmo com a redução da dimensionalidade dos dados originais**.**

A PCA, como um dos instrumentos de Análise Multivariada nos conduz aos seguintes objetivos:

- geração de subconjuntos a partir de combinações lineares alternativas e interpretáveis em subconjuntos das variáveis originais associadas às amostras dos vinhos pesquisados que possam pontualmente serem avaliados nas suas características em termos de atributos associados a cada CP.
- redução das dimensões do conjunto de dados, mas mantendo a variação máxima dos dados originais no processo de separação das principais componentes**,** variações obtidas por processos computacionais matematicamente padronizados.
- encontrar direções (vetores) mutuamente ortogonais representativos das Componentes Principais**,** nos quais os pontos das amostras selecionadas serão projetados nesses novos eixos/vetores: linhas retas em que os dados deverão estar os mais dispersos possíveis, e que serão rotacionados para coincidirem com os eixos teóricos ***x, y, z****...,* etc sem alterações das distâncias entre os pontos relacionados a uma determinada CP.
- descrição e entendimento da estrutura de correlação/covariância entre as variáveis, permitindo matematicamente a definição de novas componentes principais não correlacionadas.

Como importante referência do processo PCA, a primeira *CP* deve explicar a maior parcela da variabilidade total do conjunto original de dados. A segunda CP explica a segunda maior parcela de variabilidade total dos dados originais, e assim por diante. Esse fato será demonstrado na apresentação do processo PCA (**Parte 10**) como uma premissa do modelo de redimensionamento de uma base de dados [5,10 e 17].

PARTE 5 – IMPORTAÇÃO DA BASE DE DADOS E IDENTIFICAÇÃO DE VARIÁVEIS

A base original (*arquivo UCI*.csv [11]) é composta por 1599 registros que podemos denominar de *instâncias/rótulos numéricos,* cada uma associada a 12 variáveis contínuas descritivas de características físico-químicas, uma variável sensorial de qualidade (nota de 0 a 10), e uma variável categórica não utilizada que representa o tipo de vinho – nesse caso apenas um tipo: vinho tinto (*vinho verde*) associado a diversos rótulos.

O processo PCA investigará todas as dissimilaridades (ausências de correlações) entre as amostras de vinhos com seus atributos para separar os registros conforme a dispersão calculada nos dados, identificando, em ordem decrescente de importância, as novas variáveis denominadas de Componentes Principais – subconjuntos das variáveis originais.

Esses subconjuntos, cada um com sua própria direção de variabilidade a que chamamos, como já assinalado, de *vetores próprio*s, são representados pelos *escores* de cada registro, resultado de uma combinação linear de seus loadings (pesos) ponderados pelos valores das variáveis (atributos).

Os passos seguintes dizem respeito à preparação da base computacional para o processamento PCA com o auxílio do RStudio.

Começamos com a titulação das colunas dos atributos com seus nomes técnicos abreviados seguida da leitura do banco de dados para criação do *arquivo vinhos* acompanhado da definição do vetor de titulação (***atributos_vinhos***) (**R69**)

> *OBS: **R69** é a referência à linha correspondente no Script R no **Anexo I**. Seguindo a limitação do editor de fontes do R foi utilizada a fonte Lucida Console em **negrito** e não a fonte Arial que é utilizada em todo o texto que não se refere a comandos do R. Em alguns casos os quadros outputs do R serão*

ampliados ou reduzidos manualmente para uma melhor apresentação estética de seus conteúdos.

Os comandos R apresentam, muitas vezes argumentos alternativos ou complementares com diversos níveis de sofisticação nos seus outputs, principalmente quando se referem a recursos de apresentação gráfica. Contudo somente argumentos estritamente necessários serão utilizados visando apresentar os outputs desejados [12].

Vamos ressaltar que em toda a Dissertação o nome dos atributos constantes dos inputs e outputs do R serão abreviados, destacando-se seu nome completo, quando assim se fizer necessário para um melhor entendimento do ponto em questão. Segue o comando R utilizado para a criação do vetor de titulação necessário para o processamento PCA: (**R69 a R75**)

```
atributos_vinhos<- c("label","aci_fix","aci_vol","ac_citr",
                     "acu_res","clor", "fsd","tsd","dens","PH",
                     "sulf","gr_alc","qual","tipo")
```

OBS: As colunas label e tipo não serão utilizadas de forma explícita.

Os nomes completos dos atributos estão descritos na **Parte 1.2** que descreve os componentes químicos do vinho. Após a leitura do banco de dados e a criação do vetor titulação é feita a importação do arquivo *csv* com os dados da base baixada da ***UCI Data Set*** para o ambiente R.

A criação do arquivo vinhos é feita com o seguinte comando R:

```
vinhos<-read.csv2("vinhos_tintos_UCI.csv",
                  col.names = atributos_vinhos,header=FALSE,
                  skip=1)
```

Resumo da base de dados importada – espelho parcial com sua métrica original. São apresentadas as seis primeiras linhas de uma tabela de dados. **(R98)**

	aci_fix	aci_vol	ac_citr	acu_res	clor	fsd	tsd	dens	PH	sulf	gr_alc	qual
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5

Quadro 1 – Espelho parcial da base de dados (seis primeiros rótulos) com sua métrica original conforme as unidades documentadas na **Parte 1.2**

As variáveis após a leitura do banco de dados entram como *chr* (caracteres) e são transformadas através de comandos R em variáveis numéricas (*num*) para serem processadas no RStudio. Filtrando variáveis numéricas: (**R91** a **R95**)

```
'data.frame':   1599 obs. of  12 variables:
 $ aci_fix: num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ aci_vol: num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ ac_citr: num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ acu_res: num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ clor   : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ fsd    : num  11 25 15 17 11 13 15 15 9 17 ...
 $ tsd    : num  34 67 54 60 34 40 59 21 18 102 ...
 $ dens   : num  0.998 0.997 0.997 0.998 0.998 ...
 $ PH     : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulf   : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ gr_alc : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ qual   : num  5 5 5 6 5 5 5 7 7 5 ...
```

Quadro 2 – Data frame com as variáveis caracterizadas como numéricas

Temos que complementar a formalização do arquivo vinhos testando a presença de registros totalmente nulos, o que resulta no seguinte output R: (**R111**)

```
aci_fix aci_vol ac_citr acu_res    clor     fsd     tsd    dens      PH    sulf
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
 gr_alc    qual
  FALSE   FALSE
```

Como observado (FALSE) não são identificados *registro de vinhos totalmente nulos*, mas examinando-se o Banco de Dados original observa-se a ausência de registros de alguns atributos para alguns rótulos sem comprometer, como veremos mais tarde, a análise do bancos de dados.

PARTE 6 – TESTES DE ADEQUAÇÃO DA BASE DE DADOS PARA O PROCESSO PCA

Através de recursos da *Análise Fatorial Exploratória* faremos uma verificação da adequação dos dados para o nosso estudo PCA. Usamos esta técnica para *"investigar as relações entre um grande número de variáveis e organizá-las em um conjunto menor de fatore*s" [13,14].

São as seguintes condições de análise para a aplicação dos testes de adequação:

- Tamanho da amostra: pelo menos 5 vezes maior do que o número de variáveis ou pelo menos mais do que 100, requisitos atendidos pela base de dados.(**R120 - R121**) [14]

- Variáveis utilizando escalas numérica cuja transformação foi feita no contexto da **Parte 5.** (**R123**)

- Verificação da matriz de correlação. Menos de 21 % dos coeficientes de correlação apresenta valores maiores do que 0.3. Por esse fato, Isoladamente, o quadro de correlações compromete parcialmente a aplicação do PCA mas sem afetar de forma relevante o processo. (**R128**)

	aci_fix	aci_vol	ac_citr	acu_res	clor	fsd	tsd	dens	PH	sulf	gr_alc	qual
aci_fix	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
aci_vol	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
ac_citr	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
acu_res	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
clor	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
fsd	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
tsd	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
dens	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.18
PH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulf	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
gr_alc	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
qual	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.18	-0.06	0.25	0.48	1.00

Quadro 3 – Matriz de correlações entre os atributos

A **Figura 3** permite uma visão gráfica da Matriz de Correlações de Person sendo que a escala de cores reflete o índice de correlação no intervalo de -1 a +1.
(R130)

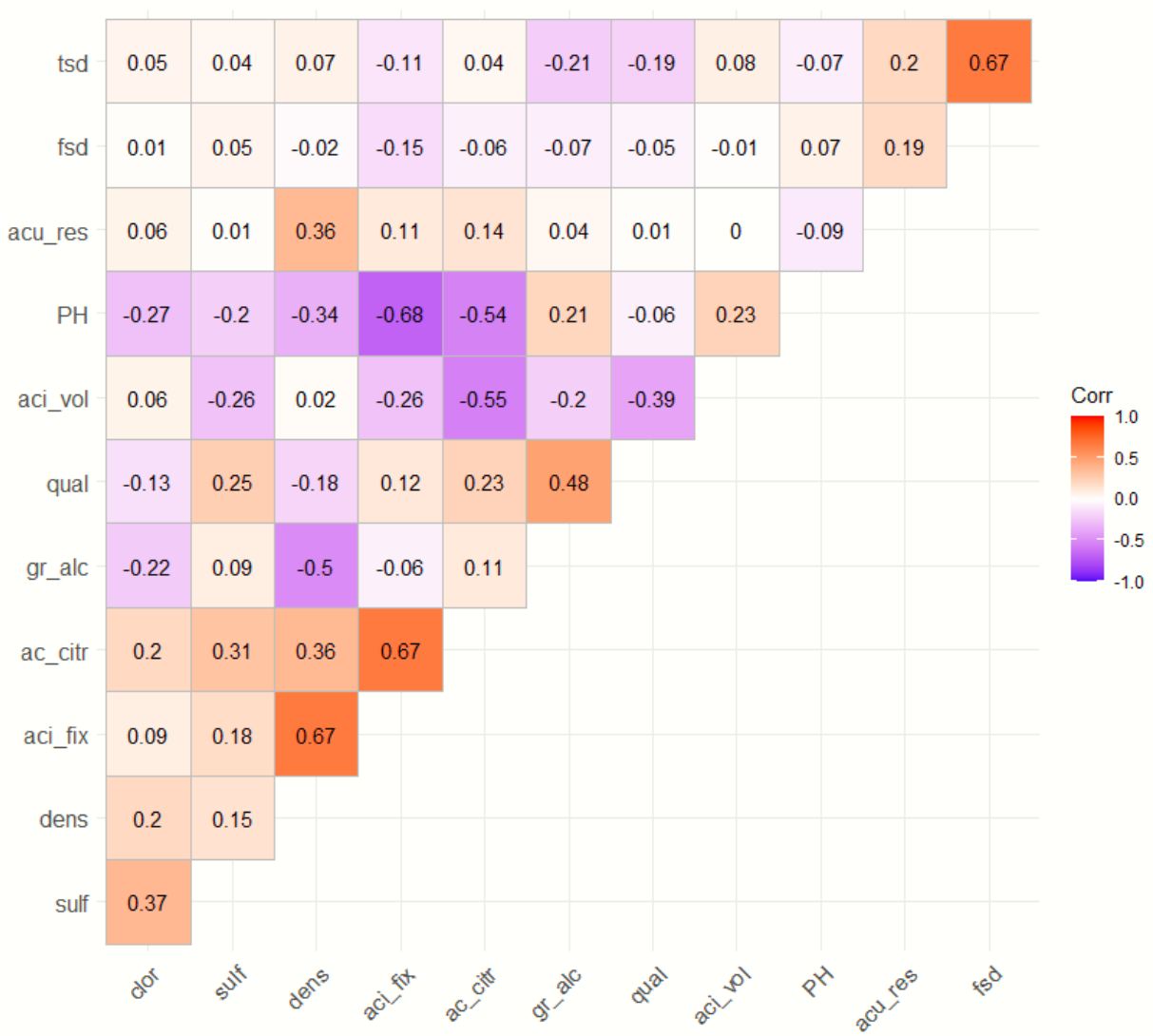

Figura 3 – Gráfico das correlações entre os atributos da Base de Dados

- Teste de esfericidade de Barlett

> “Estatística de teste usada para examinar a hipótese de que as variáveis não sejam correlacionadas na população, ou seja, a matriz de correlação da população é uma matriz identidade onde cada variável se correlaciona

perfeitamente com ela própria (*r=1*), mas não apresenta correlação com as outras variáveis (*r=0*).” [20,14,33]

O teste de esfericidade de Bartlett examina se a matriz de correlação deve ser fatorada, isto é, os dados não são *independentes*.

Este teste compara a matriz de correlação observada à matriz de identidade. Por esse processo podemos verificar se há uma certa redundância entre as variáveis que podemos resumir com alguns fatores (Componentes Principais). Se as variáveis estiverem perfeitamente correlacionadas, apenas um CP é suficiente. Se eles são ortogonais, precisamos de componentes no mesmo número de variáveis. Neste último caso, a matriz de correlação é a mesma que a matriz de identidade. Uma estratégia simples é visualizar a matriz de correlação. Se os valores fora da diagonal principal são frequentemente altos (em valor absoluto), algumas variáveis estão correlacionadas. *Se a maioria desses valores estiverem perto de zero, o PCA não é realmente útil.*

Como na *verificação da matriz de correlações* não temos incidência de muitos valores próximos de zero podemos usar o modelo PCA.

Utilizando o RStudio para fazer o teste de esfericidade: (**R135**)

```
$chisq
[1] 8731.122

$p.value
[1] 0

$df
[1] 66
```

Quadro 4 – Output do R para avaliar o teste de esfericidade da Base de Dados

OBS:

O nível de significância foi pequeno o suficiente para o R definir que é 0. Assumindo que valores menores que 0.05 indicam que uma parcial abordagem

fatorial pode ser útil para validar a aplicação do PCA na base de dados, então se mostraram adequados para esta metodologia de análise multivariada.

A literatura nos informa que o teste do Bartlett tem uma forte desvantagem. Tende a ser sempre estatisticamente significante quando o número de instâncias *n* aumenta.

- KMO Medida de adequação de uma amostra

A medida de amostragem Kaiser-Meyer-Olkin (KMO) tem valores entre 0 e 1, com pequenos valores indicando que, em geral, as variáveis têm pouco em comum para justificar uma análise e valores de componentes principais. Acima de 0,5 são considerados satisfatórios para uma análise de CP [14, 33].

O índice de KMO tem o mesmo objetivo do teste de Barlett. Ele verifica se podemos fatorizar eficientemente as variáveis originais. Contudo é fundamentado em outra abordagem.

A matriz de correlação é sempre o ponto de partida. Sabemos que as variáveis são mais ou menos correlacionadas, mas a correlação entre duas variáveis pode ser influenciada pelos outras. [20 – Cap. 2] [14]. É também um índice usado para avaliar a adequacidade da análise fatorial. Valores altos (entre 0,5 e 1,0) indicam que a análise fatorial é apropriada. Valores abaixo de 0,5 indicam que a análise fatorial pode ser inadequada.

Se o índice de KMO for alto (≈1), a PCA pode agir de forma eficiente. Se o KMO é baixo (≈0), a PCA não é relevante. Algumas referências dão uma tabela para a interpretação do valor do índice de KMO obtido em um conjunto de dados.

Utilizando o RStudio para fazer o KMO (**R139**) obtemos o *output* do R que mostram os resultados obtidos (**Quadro 5**). Podemos verificar que tanto o KMO total (0.43) entre 0 e1 assim como o KMO de cada um dos componentes químicos não foram altos, mas em um patamar com uma média que não compromete (próximo do valor mínimo = 0.5) a validação do método PCA.

```
> KMO(vinhos[2:12])
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = vinhos[2:12])
Overall MSA =  0.43
MSA for each item =
aci_fix aci_vol ac_citr acu_res    clor     fsd     tsd
   0.45    0.52    0.70    0.20    0.46    0.48    0.45
   dens      PH    sulf  gr_alc
   0.37    0.45    0.51    0.23
```

Quadro 5 – Output do comando KMO no R

Parte 6.1 – Conclusão dos testes de adequação da base.

O teste de esfericidade do Bartlett e o índice de KMO permitem detectar se podemos ou não resumir as informações fornecidas pelas variáveis do banco de dados em alguns fatores (Componentes Principais). Mas eles não dão indicação sobre o número apropriado de fatores não fornecendo uma associação direta com o número de PC's do PCA.

O resultado dos dois testes e suas conclusões individuais não compromete o uso do PCA da base na dados utilizada, principalmente pelo tamanho da base escolhida. Esta conclusão será reforçada na aplicação do modelo PCA na **Parte 10** que nos informa que 4 Componentes Principais serão suficientes para fazer uma avaliação da variância total da base de dados.

PARTE 7 – INFORMAÇÕES ESTATÍSTICAS DESCRITIVAS PRINCIPAIS DA BASE DE DADOS.

Começamos com a apresentação de um histograma da base de dados dos vinhos.

Os atributos Ácido Fixo, Ácido Volátil, fsd e tsd (free sulfur dioxides) além de Sulfatos, se apresentam com uma forma de distribuição parecida, assimétrico positivo com uma cauda à direita, apresentando mais *outliers* que todos os atributos. (**R155 a R16**)

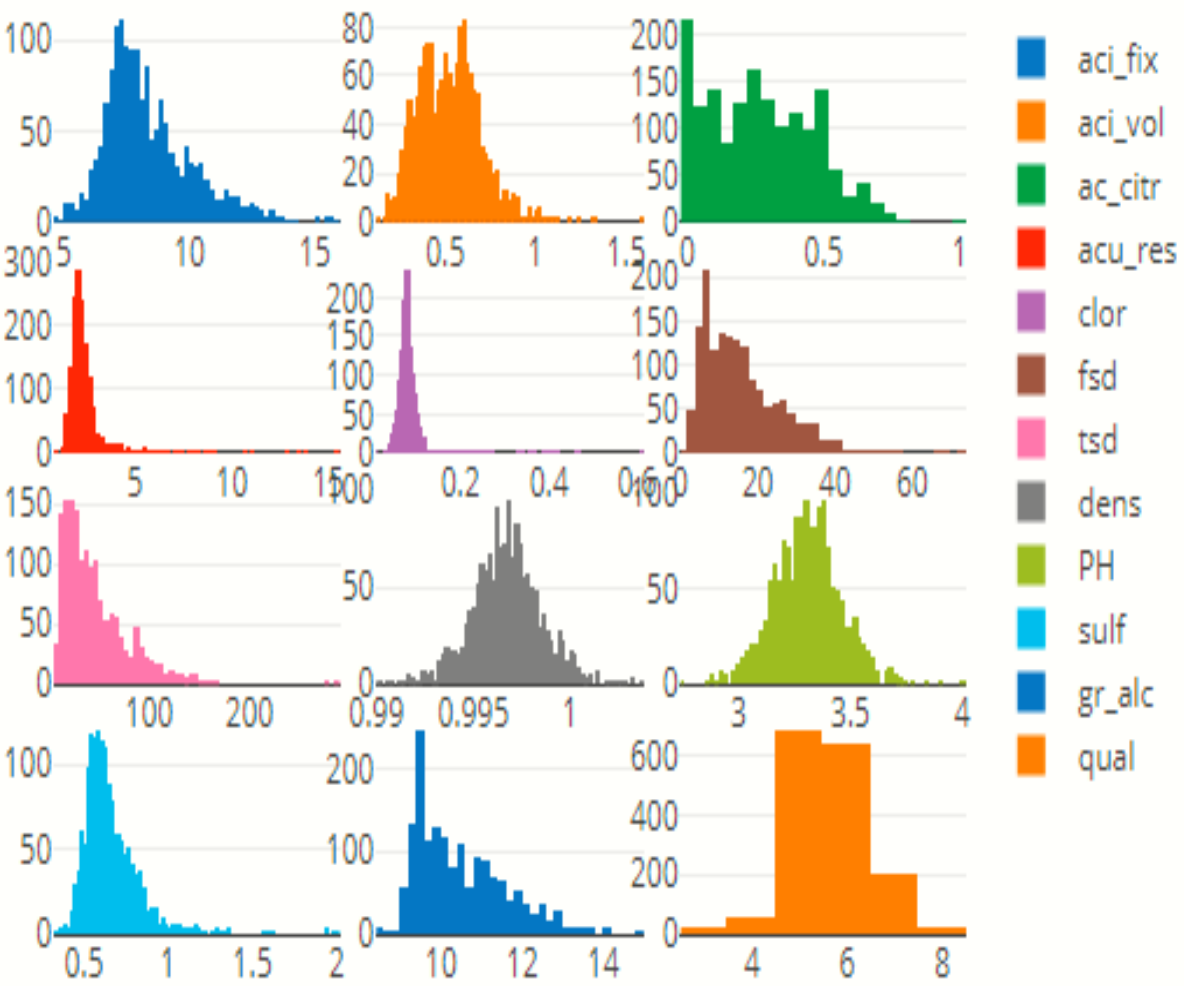

Figura 4 – Histogramas dos atributos e da qualidade.

Já a Densidade e o pH se apresentam com um padrão normal de distribuição com reduzido número de *outliers*. Açúcar Residual e Cloretos parecem apresentar muitos *outliers* positivos.

De acordo com a base de dados o Ácido Cítrico apresenta alguns valores zerados o que compromete uma definição de um padrão de distribuição. Açúcar Residual e Cloretos parecem apresentar muitos outliers positivos.

O sentido estatístico dos *outliers* deve ser avaliado com cuidado tendo em vista que as métricas escolhidas para os atributos de cada rótulo numérico são planejadas e não feitas ao acaso. Isso muda o sentido dos *outiliers* que não representam necessariamente vieses estatísticos quando avaliados sensorialmente, mas sim resultados de escolhas de misturas previamente estabelecidas que não devem ser alteradas para não mudar o sabor planejado para o vinho.

A qualidade dos vinhos não é bem distribuída concentrando-se nas notas médias entre 5 e 6. O Quadro 6 de notas apresenta a quantidade de rótulos identificados com cada nota e a concentração citada (notas de 0 a 10 sendo a menor nota observada 3 e a maior nota observada igual a 8). **(R646)**

3	4	5	6	7	8
10	53	681	638	199	18

Quadro 6 – Tabela de notas resultado da avaliação sensorial

O **Quadro 7** nos permite analisar o sumário estatístico descritivo da base de dados e observar a posição dos indicadores em relação à média dos dados e interpretar as variações IQR.

aci_fix	aci_vol	ac_citr	acu_res
Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900
1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900
Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200
Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539
3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600
Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500

clor	fsd	tsd	dens
Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901
1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956
Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968
Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967
3rd Qu.:0.09000	3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9978
Max. :0.61100	Max. :72.00	Max. :289.00	Max. :1.0037

PH	sulf	gr_alc	qual
Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
Median :3.310	Median :0.6200	Median :10.20	Median :6.000
Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

Quadro 7 – Sumário estatístico para análise IQR (variação interquartil). (**R165**)

A **Figura 5** apresenta um modelo que permite a leitura e interpretação das variações interquartis que influenciam a formação dos box plot da base de dados.

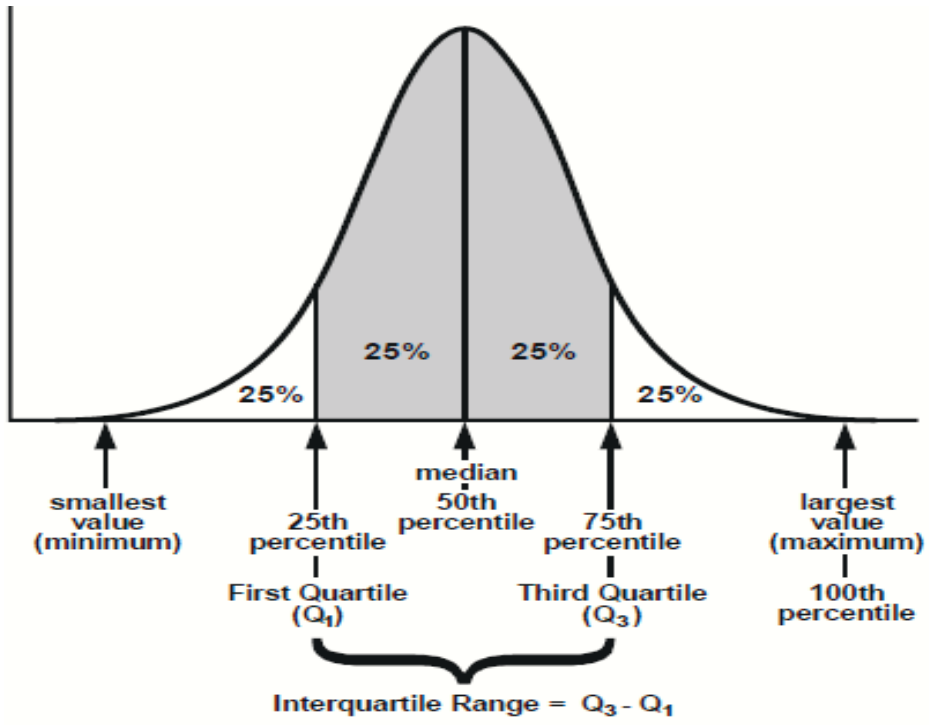

Figura 5 – Gráfico ilustrativo para análise IQR [15].

Parte 7.1 – Visão geral dos boxplots dos atributos para avaliação de outliers

A **Figura 6** apresenta o box plot de todos os atributos da base de vinhos sinalizando as maiores distorções relação à média de cada um (**R176 – R184**).

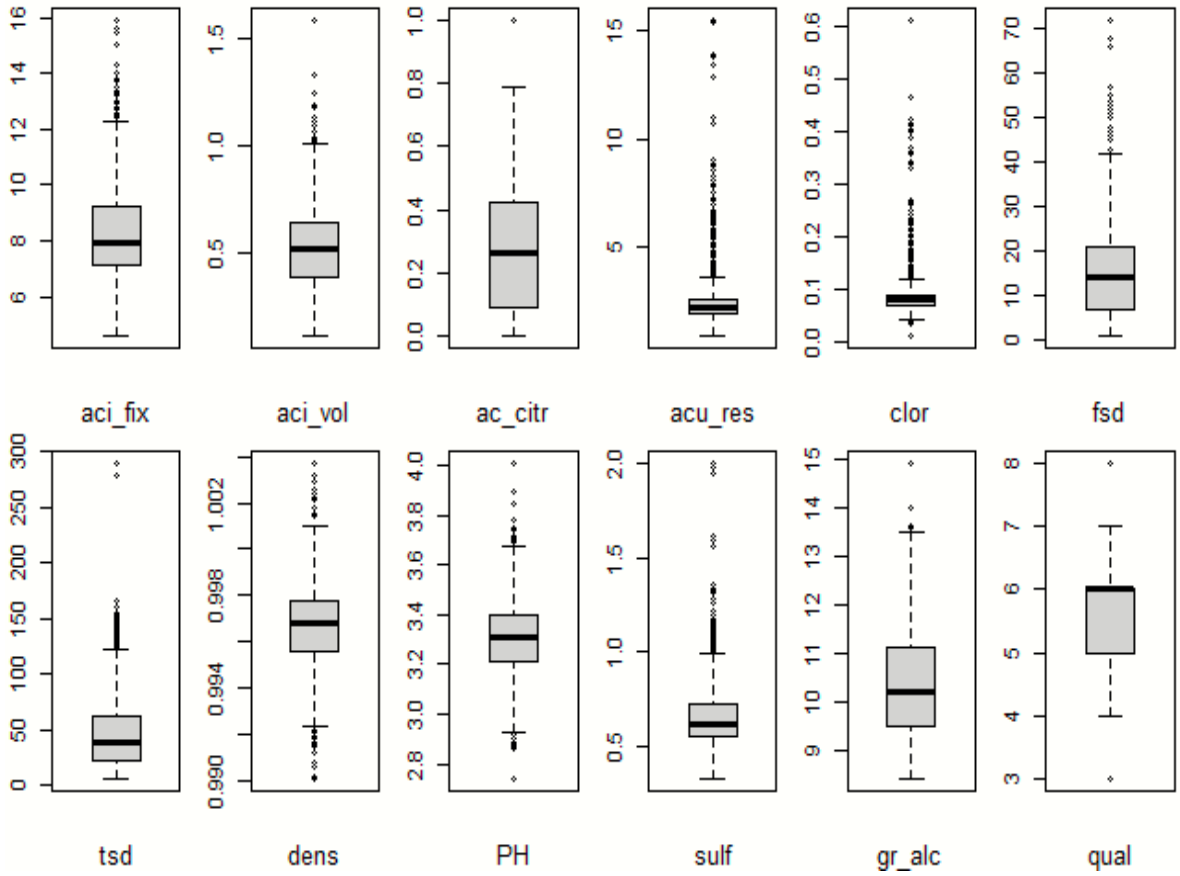

Figura 6 – Quadro geral individualizado por atributo.

Legendas da **Figura 6** *(output* do R):

1 - fixed acidity (acidez fixa) - aci_fix
2 - volatile acidity (acidez volátil) - aci_vol
3 - citric acid (ácido cítrico) - ac_citr
4 - residual sugar (açúcar residual) - acu_res
5 - chlorides (cloretos) - clor

6 - free sulfur dioxide - fsd
7 - total sulfur dioxide - tsd
8 - density (densidade) - dens
9 - pH - pH
10 – sulfates (sulfatos) - sulf
11 - alcohol (grau alcoólico) - gr_alc
12 - quality (qualidade) - qual

Para uma melhor visualização dos outliers as **Figuras 8** apresentam uma separação de componentes para ressaltar os outliers que não são claramente identificados na **Figura 7**. As legendas são as mesmas da **Figura 6**.

> **OBS:** "*Os outiliers recebem diversos outros nomes: dados discrepantes, pontos fora da curva, observações fora do comum, anomalias, valores atípicos e outros."* [48]

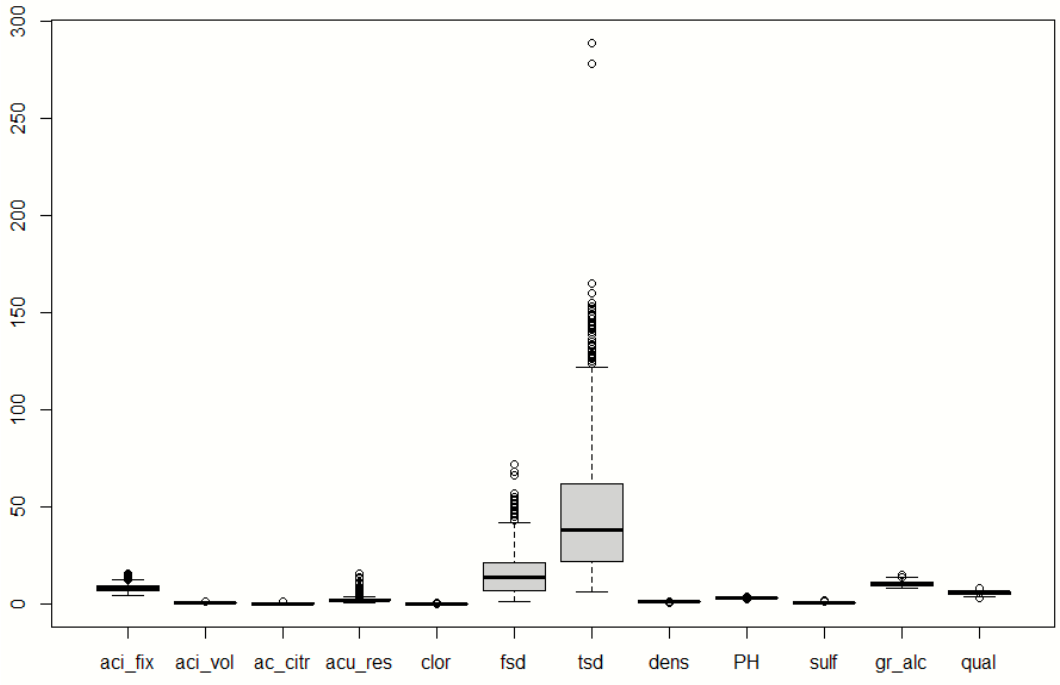

Figura 7 – Quadro geral alternativo de todos os atributos.

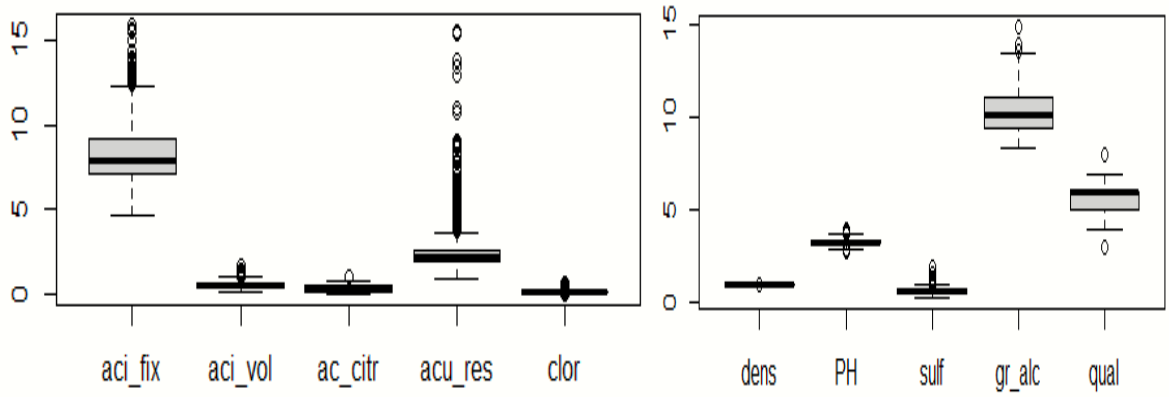

Figura 8: Quadro da Figura 7 particionado

A **Figura 8** espelha o quadro da **Figura 7** particionado para realçar alguns atributos com outiliers mais relevantes.

Nos box plot menores vemos os atributos não ressaltados no boxplot geral devido aos outliers dos atributos fsd e do atributo tsd que mascaram a análise em termos de representação. Uma análise *dos desvios padrões* nos leva às seguintes considerações: existem diferenças relevantes em termos de desvios padrão dos atributos confirmando os *outliers* que foram detectados nos boxplots que fizemos anteriormente. (**R186 – R189 – R190**)

Comparativamente os atributos *acu_res*, *fsd* e *tsd* possuem um desvio padrão acima dos demais. Os dados mostram que o desvio padrão do atributo qualidade pode nos indicar que considerar a influência de *outliers* de outros atributos com impacto sensorial significativo deve ser relevante.

Considerando que a variação métrica dos atributos em cada tipo de vinho é um determinante de seu sabor/qualidade não faz sentido tentarmos reformular a base de dados com a eliminação de *outliers* pois isso exigiria avaliações sensoriais sucessivas para analisar os efeitos desta eliminação alterando toda a caracterização obtida com

a avaliação sensorial original que é parte relevante da avaliação qualitativa dos vinhos da base de dados original. (**R186 – R189 – R190**)

Apenas no âmbito da estatística descritiva os outliers têm importância analítica na avaliação individual de cada atributo, comparativamente às medidas dos outros atributos. Cabe ressaltar que uma análise do *arquivo csv* dos vinhos mostrou que apenas 18 “rótulos” foram classificados com nota 8 e nenhum acima desta nota, fortalecendo a ideia de ausência de qualidade a nível de excelência dos vinhos pesquisados na linha da *análise sensorial* adotada, ou seja, nenhum *outlier* apresentou um desvio suficiente para influenciar de forma relativa a concentração dos vinhos entre as notas 5 e 6.

O sumário estatístico e os box plot mostram que a maioria das variáveis possui ampla variação em relação ao IQR, o que indica dispersão nos dados (como esperado na formação das CPs e presença de vários outliers em todas as variáveis, exceto no álcool e na variável de saída qualidade.

Complementando a avaliação geral dos outliers podemos observar desvios relevantes nas medidas descritivas conforme o **Quadro 8**. (**R196**).

Os boxplots demonstraram que os outiliers existentes em alguns atributos são derivados de diferenças em algumas de suas medidas descritivas em relação a outros devido à diferença das métricas dos componentes químicos utilizados, ressaltando-se fsd, tsd e gra_alc.

	vars	n	mean	sd	median	trimmed	mad	min	max	range
aci_fix	1	1599	8.32	1.74	7.90	8.15	1.48	4.60	15.90	11.30
aci_vol	2	1599	0.53	0.18	0.52	0.52	0.18	0.12	1.58	1.46
ac_citr	3	1599	0.27	0.19	0.26	0.26	0.25	0.00	1.00	1.00
acu_res	4	1599	2.54	1.41	2.20	2.26	0.44	0.90	15.50	14.60
clor	5	1599	0.09	0.05	0.08	0.08	0.01	0.01	0.61	0.60
fsd	6	1599	15.87	10.46	14.00	14.58	10.38	1.00	72.00	71.00
tsd	7	1599	46.47	32.90	38.00	41.84	26.69	6.00	289.00	283.00
dens	8	1599	1.00	0.00	1.00	1.00	0.00	0.99	1.00	0.01
PH	9	1599	3.31	0.15	3.31	3.31	0.15	2.74	4.01	1.27
sulf	10	1599	0.66	0.17	0.62	0.64	0.12	0.33	2.00	1.67
gr_alc	11	1599	10.42	1.07	10.20	10.31	1.04	8.40	14.90	6.50
qual	12	1599	5.64	0.81	6.00	5.59	1.48	3.00	8.00	5.00

Quadro 8 – Medidas estatísticas descritivas da Base de Dados

PARTE 8 – ESTUDO DO COMPORTAMENTO DA VARIÁVEL QUALIDADE COM OS RECURSOS DO RSTUDIO

Um exame preliminar da base de dados no item qualidade nos mostra que poucos vinhos são na realidade ruins, isto é, com uma qualidade inferior a 3. De maneira semelhante também temos uma pequena quantidade (18 rótulos) de vinhos bons (mas não excelentes), ou seja, com resultado de avaliação sensorial com nota máxima 8.

O **Quadro 8** nos permite verificar que, a princípio, os atributos fsd, tsd, acu_res, aci_fix e gr_alc (*com base na média, no desvio padrão e na faixa de variação*) são os que têm mais potencial de influenciar o comportamento estatístico da base, no caso de mudanças no perfil individual de composições químicas dos vinhos.

Revendo a tabela de notas espelhadas no **Quadro 6** como resultado da avaliação sensorial, e fazendo classificação de vinhos com notas abaixo ou igual a 6 como ruins e maior que 6 como bons, encontramos 217 bons e 1382 ruins. A maior parte dos vinhos (82%) tem suas notas concentradas entre 5 e 6, incluindo essas notas. **(R200 e R650)**

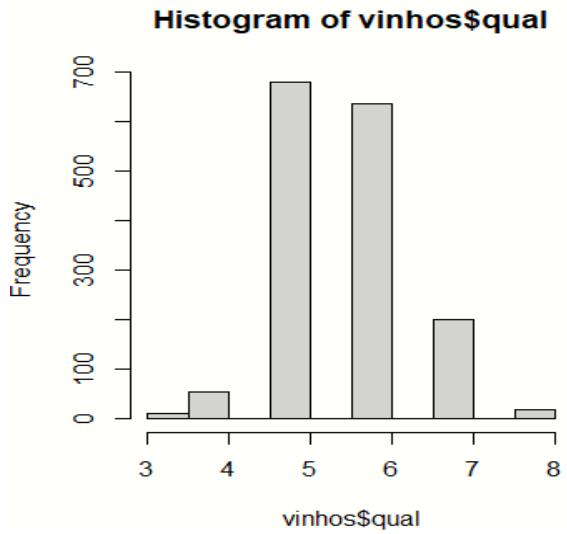

Figura 9 – Histograma da variável categórica qualidade dos vinhos (**R654**).

O atributo qualidade e suas correlações com outros atributos terão especial atenção e relevância. Dessa forma, vamos primeiro entender sua distribuição através de um gráfico de barras (**Figura 9**), já que a qualidade é um dado ordinal. (**R637 a 641**)

Devemos destacar que a qualidade final de um vinho está bastante associada aos seguintes fatores naturais: clima, solo e ao tipo de uva e, também, à sua composição química de atributos intencionais (definidos com suas métricas na fabricação dos vinhos). Contudo neste trabalho vamos investigar estatisticamente através da matriz de correlações como é refletida na qualidade a composição química dos rótulos.

Reforçando o que foi dito na introdução da **Parte 8**, pelos quadros de separação de vinhos bons e ruins vemos que a quantidade de vinhos classificados como ruins devido a notas abaixo de 6 é bem superior aos vinhos classificados como bons o que, certamente, influenciará negativamente o item qualidade na avaliação da base de dados. Para análises posteriores, os 217 vinhos considerados bons terão papel importante na avaliação de qualidade como resultado da análise sensorial com os grupos avaliados isoladamente.

Como a qualidade recebe grande influência das métricas dos componentes químicos que compõem os vinhos, é importante a forma de como os atributos afetam a avaliação sensorial dos vinhos: Fontes de estudo: [1, 2, 12, 22, e 32]

- “Acidez fixa: a maioria dos ácidos envolvidos com vinho ou fixos ou não voláteis (não evaporam prontamente).

- Acidez volátil: quantidade de ácido acético no vinho, que com muito alto dos níveis pode levar a um gosto desagradável e vinagre.

- Ácido cítrico: encontrado em pequenas quantidades, o ácido cítrico pode adicionar 'frescor' e sabor aos vinhos.

- Açúcar residual: a quantidade de açúcar remanescente após paradas de fermentação.

- Cloretos: quantidade de sal no vinho.

- Dióxido de enxofre livre: forma livre de dióxido de enxofre. Existe em equilíbrio entre a molecular dióxido de enxofre (como um gás dissolvido) e íon bissulfito.

- Dióxido total de enxofre: quantidade de formas livres e ligadas de S02.

- Densidade: densidade no vinho está próxima à da água, dependendo do álcool e do teor de açúcar percentual.

- pH: descreve como o vinho ácido ou básico é em uma escala de 0 (muito ácido) a 14 (“muito básico”).

- Sulfatos: um aditivo de vinho que pode contribuir para os níveis de gás de dióxido de enxofre (S02) (“sulfato de potássio”).

- Álcool: o teor percentual de Grau Alcoólico (% em volume).

Parte 8.1 – Gráficos da matriz de correlações entre atributos e a qualidade.

Fonte principal desta parte: [28]

Os box plot da **Parte 9** irão apresentar a evolução horizontal da análise sensorial (notas) relacionada com a métrica individual dos atributos (vertical) permitindo uma visão da correlação existente (ou não).

Pela análise da matriz de correlações espelhada na **Figura 3** podemos comprovar a importância da porcentagem do álcool (cor. = 0.5) e da acidez volátil (cor. = - 0.6) para a qualidade do vinho. Em observação anterior vimos que a acidez volátil é responsável por um gosto desagradável no vinho. Outros aspectos podem ser observados no exame da matriz de correlações**.**

										qual
									gr_alc	0.5
								sulf	0.1	0.3
							PH	-0.2	0.2	-0.1
						dens	-0.3	0.1	-0.5	-0.2
					tsd	0.1	-0.1	0	-0.2	-0.2
				fsd	0.7	0	0.1	0.1	-0.1	-0.1
			clor	0	0	0.2	-0.3	0.4	-0.2	-0.1
		acu_res	0.1	0.2	0.2	0.4	-0.1	0	0	0
	ac_citr	0.1	0.2	-0.1	0	0.4	-0.5	0.3	0.1	0.2
aci_vol	-0.6	0	0.1	0	0.1	0	0.2	-0.3	-0.2	-0.4
aci_fix	-0.3	0.7	0.1	0.1	-0.2	-0.1	0.7	-0.7	0.2	-0.1	0.1

Quadro 9 - Visão alternativa da matriz de correlações sendo as legendas as mesmas da **Figura 6** (**R607 a R616)**

Observações a partir do **Quadro 9**:

- o atributo pH é negativamente correlacionado com a acidez, uma vez que os valores mais baixos na escala de pH significa uma maior acidez.

- nota-se que existe alguma correlação entre sulfatos e cloretos (0.4) sem correlação relevante com a qualidade (0.3 e - 0.1).

- a matriz de correlação demonstra correlações entre densidade e outros atributos, como álcool, açúcar residual, ácido cítrico e acidez fixa.

Podemos ainda criar um gráfico pairplot que procura resumir as informações já enfatizadas anteriormente.Com base nisso, nossa próxima tarefa é criar alguns pares relacionados a todos esses atributos escolhidos. Os pairplots das **Figuras 10** e **11** a seguir resumem a maioria das informações que temos até agora, como por exemplo a distribuição desequilibrada de qualidade - ou seja, há mais vinhos marcados como 5 ou 6 do que os abaixo e acima desses valores.

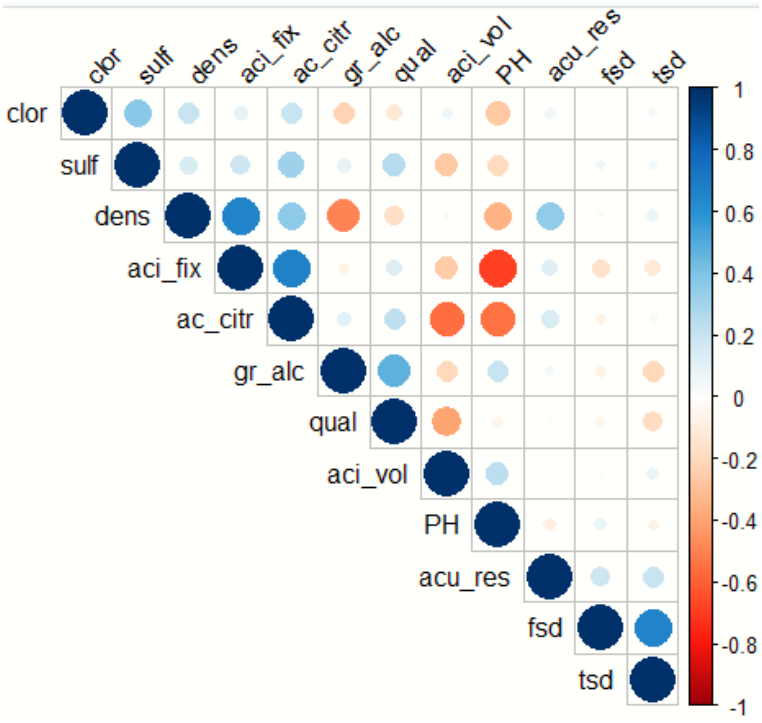

Figura 10 – Resumo gráfico da Matriz de Correlações

De acordo com as **Figuras 10**, álcool e acidez volátil são os atributos com a relação mais relevante com a qualidade. Em seguida vamos explorar os boxplot (**Parte 9**) que

trazem muitas informações incluindo tendências de comportamento da qualidade avaliando as diferentes métricas entre os vinhos pesquisados.

Nas matrizes de correlações alternativas mostradas na Fig. 10 os círculos maiores representam as maiores correlações entre os atributos cabendo-se destacar as correlações entre gr_alc e dens, aci_vol e aci_citr, PH e aci_fix, PH e aci_citr. Notamos que a maior influência isolada sobre a qualidade é exercida pela gr_alc e sulf. As bolas azuis cheias representam correlação =1 pois se referem à diagonal principal da matriz de correlações. (**R210 a R212**).

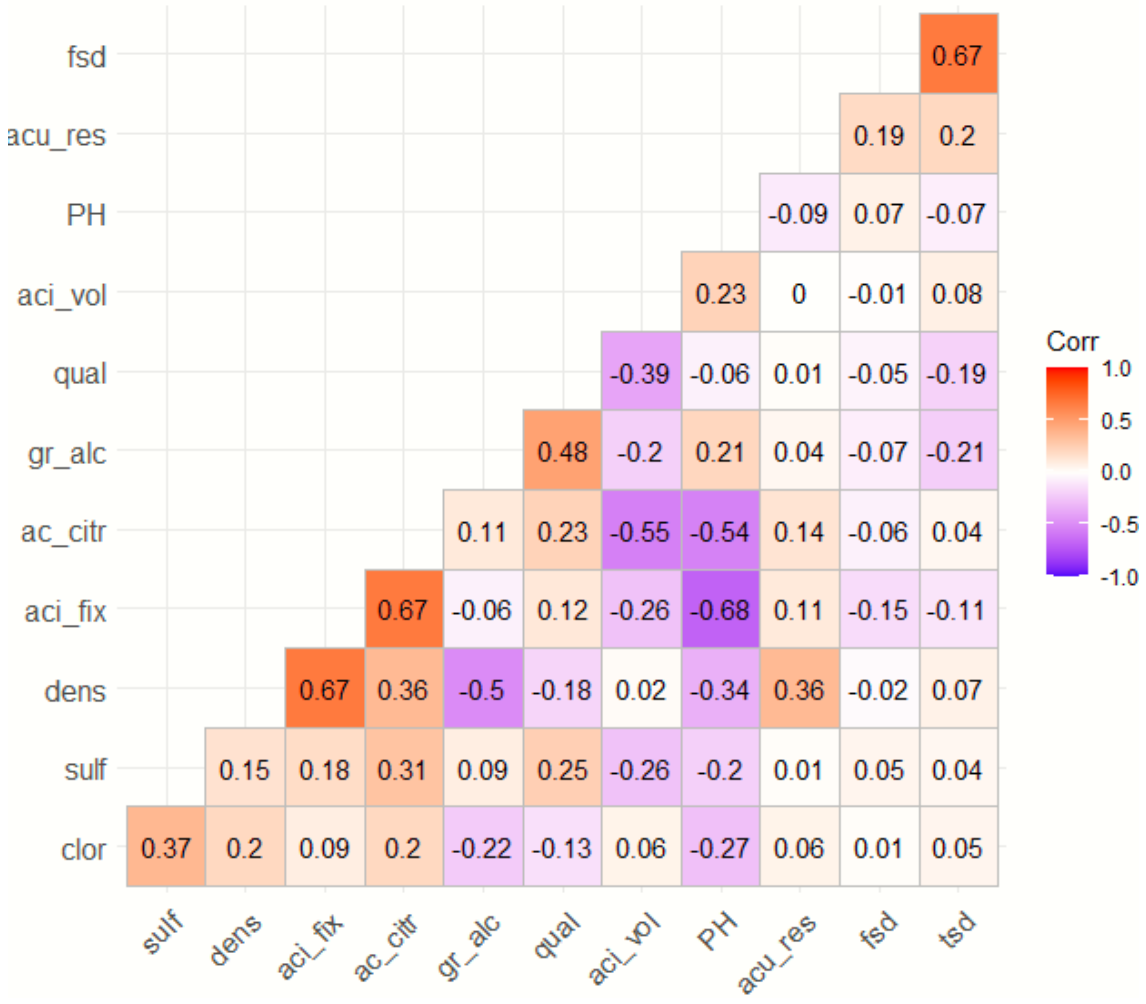

Figura 11 – Resumo gráfico alternativo da matriz de correlações.

PARTE 9 – ANÁLISE GRÁFICA DO COMPORTAMENTO DA QUALIDADE E SUAS CORRELAÇÕES COM OUTROS ATRIBUTOS ATRAVÉS DE BOXPLOTS

Para termos uma visão melhor das relações entre os atributos vamos examinar graficamente as correlações entre os componentes químicos do vinho e a qualidade.

Cabe destacar que, com base nas **Figuras 10** e **11**, assim como nos boxplots a serem analisados, temos outras observações pertinentes a serem enfatizadas pela sua importância no quadro de correlações:

- O alcohol e o volatile acidity apresentam a maior correlação com a qualidade;

- O atributo pH se apresenta com uma forte correlação negativa com os ácidos (quanto menor o pH maior acidez), no entanto com o atributo volatile acidity verificamos uma correlação positiva;

- Os atributos Sulphates e Alcohol tem correlações positivas com a qualidade;

- O atributo Density se apresenta com uma forte correlação com os atributos ácidos, Fixed Acidity e Citric Acid, residual sugar além de uma forte correlação negativa com Alcohol.

As figuras seguintes mostram a evolução horizontal da análise sensorial (notas) relacionada com a métrica individual dos atributos (vertical), permitindo uma visão da relação existente (ou não) do uso “coletivo” de um determinado atributo e as notas recebidas por todos os rótulos pesquisados devido a esse atributo, não cabendo se falar de índice de correlação positivo ou negativo, mas somente de uma tendência de mudança de qualidade devido à métrica diferenciada de um mesmo atributo em todos os rótulos avaliados coletivamente.

Fonte de Estudo: [1, 2, 12]

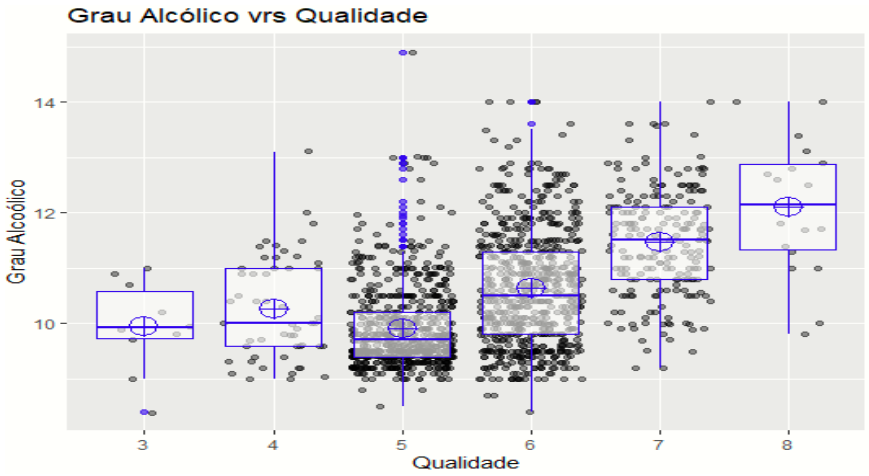

Figura 12 – Comportamento das notas de qualidade de todos os rótulos conforme o Grau Alcóolico de acordo com as métricas deste atributo. (**R226 – R233**)

Pelo gráfico da figura 12 pode-se perceber que vinhos com notas de avaliação sensorial média entre 5 e 6 se apresentam com teor de Grau Alcoólico concentrado aproximado entre 9% e 11% podendo serem classificados por esse motivo como suaves ou doces.

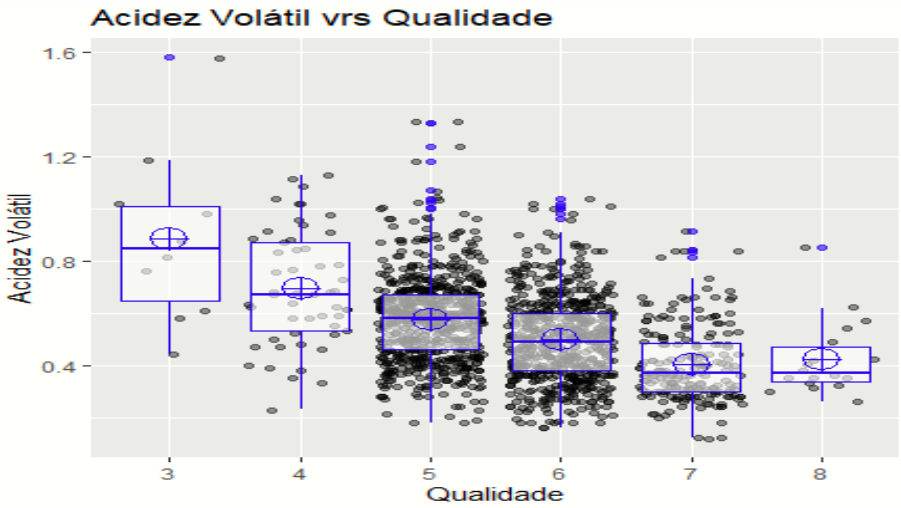

Figura 13 – Comportamento das notas de qualidade de todos os rótulos conforme a Acidez Volátil de acordo com as métricas deste atributo. (**R211 - R218**)

A figura 13 demonstra relação contrária entre o aumento da métrica do atributo Acidez Volátil e o comportamento da avaliação sensorial da qualidade. Qualidade diminui com o aumento da acidez no conjunto dos rótulos pesquisados.

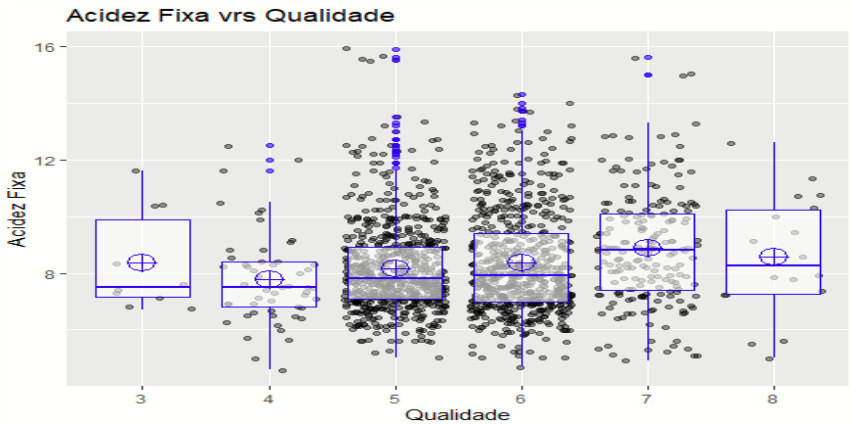

Figura 14 – Comportamento das notas de qualidade de todos os rótulos conforme a Acidez Fixa de acordo com as métricas deste atributo. (**R246 – R253**)

Alterações na métrica do atributo Acidez Fixa mostrado na **Figura 14** não tem efeito relevante sobre mudanças coletivas na avaliação sensorial da qualidade.

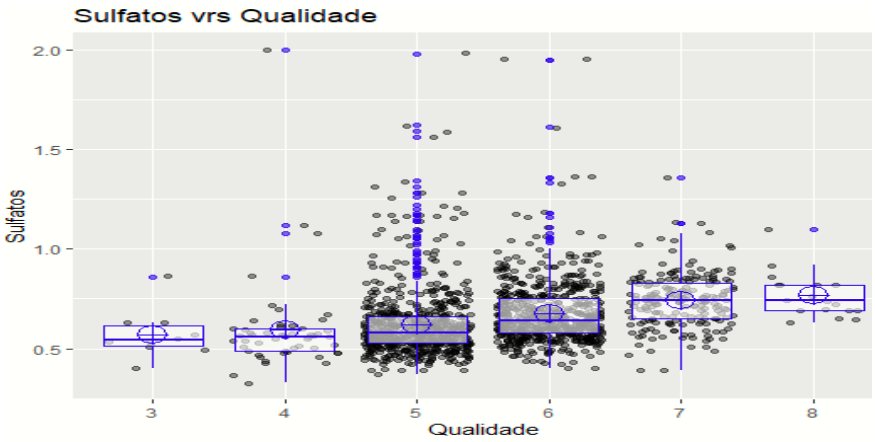

Figura 15 – Comportamento das notas de qualidade de todos os rótulos conforme os Sulfatos de acordo com o comportamento da métrica deste atributo. **(R256 – R263)**

A **Figura 15** mostra a existência de uma leve tendência positiva entre o aumento coletivo do atributo Sulfatos e a Qualidade. Quantidades maiores do atributo Sulfatos provocam aumento de qualidade.

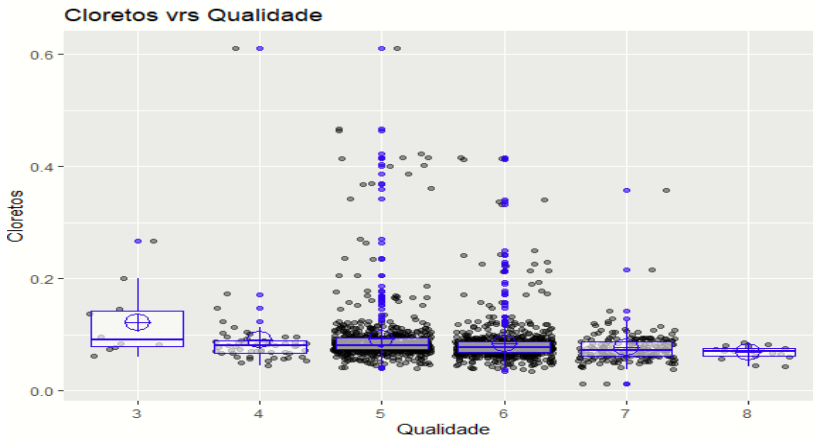

Figura 16 - Comportamento das notas de qualidade de todos os rótulos conforme os Cloretos de acordo com as métricas deste atributo. **(R266 – R273)**

Na **Figura 16** observa-se que a maior parte dos vinhos que recebem notas entre 5 e 7 pela análise sensorial apresenta valores reduzidos de Cloretos.

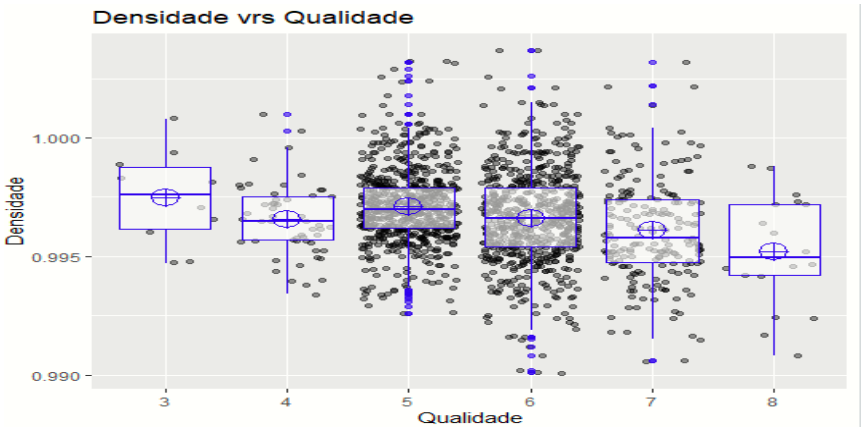

Figura 17 – Comportamento das notas de qualidade de todos os rótulos conforme a Densidade de acordo com o comportamento da métrica deste atributo. (**R276 – R283**)

A maior parte dos vinhos com notas médias entre 5 e 7 apresentam uma tendência média de manutenção da Qualidade com as mudanças de densidade.

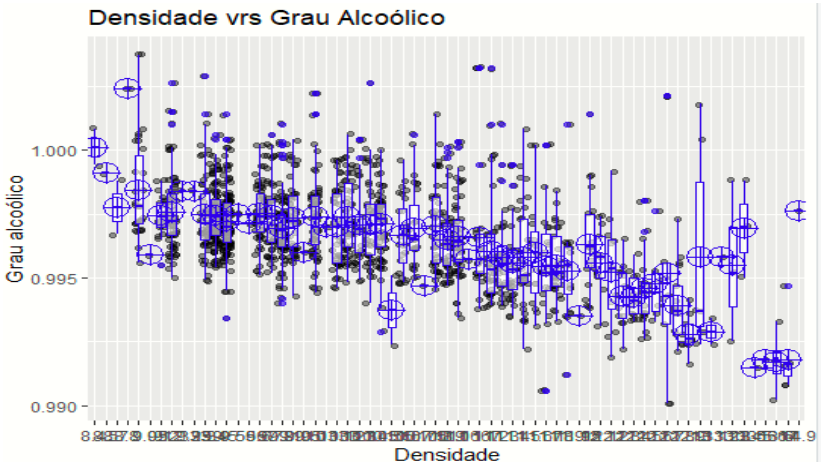

Figura 18 – Comportamento da Densidade de todos os rótulos conforme o Grau Alcóolico de acordo com o comportamento da métrica deste atributo. (**R286 – R283**)

O box plot da **Figura 18** mostra que o valor da densidade decresce na medida que o Grau Alcoólico é reduzido. Conforme visto no box plot anterior podemos concluir que os vinhos com melhor qualidade tem uma menor Densidade, logo um menor grau alcoólico.

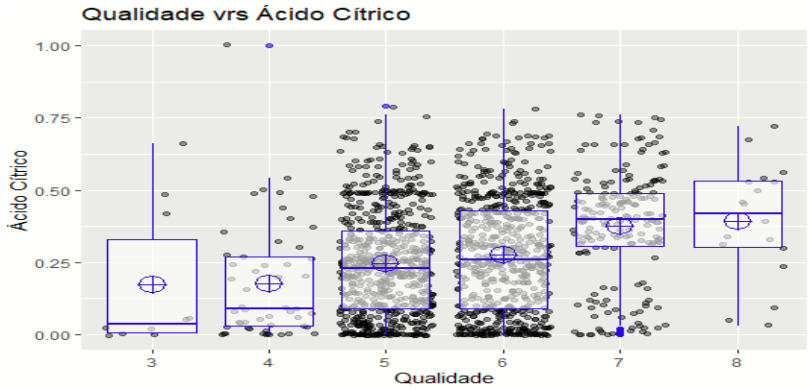

Figura 19 – Comportamento das notas de Qualidade de todos os rótulos conforme o Ácido Cítrico de acordo com o comportamento da métrica deste atributo. **(R296 – R303)**

O box plot na **Figura 19** mostra que um aumento no Ácido Cítrico resulta em um aumento da qualidade.

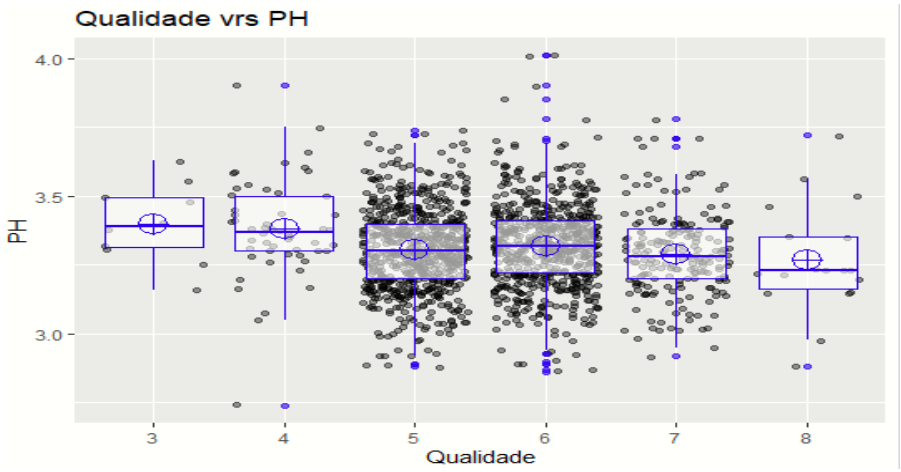

Figura 20 – Comportamento das notas de Qualidade de todos os rótulos conforme o pH de acordo com o comportamento da métrica deste atributo. **(R306 – R313)**

O plot na Figura 20 mostra que o pH tem pouca influência na qualidade de todos os rótulos

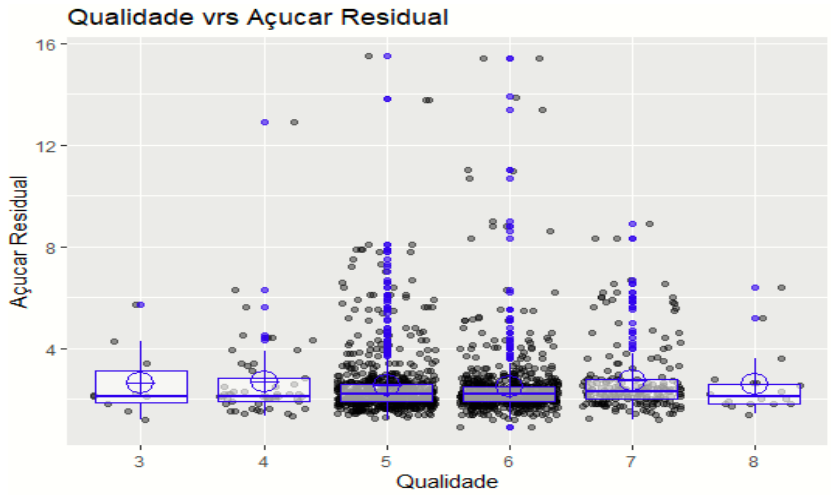

Figura 21 – Comportamento das notas de Qualidade de todos os rótulos conforme o Açúcar Residual de acordo com o comportamento da métrica deste atributo. **(R316 – R323)**

O box plot na **Figura 21** nos mostra que o atributo Açúcar Residual não tem relevância na avaliação sensorial da qualidade que se mantém estável no conjunto de “rótulos”.

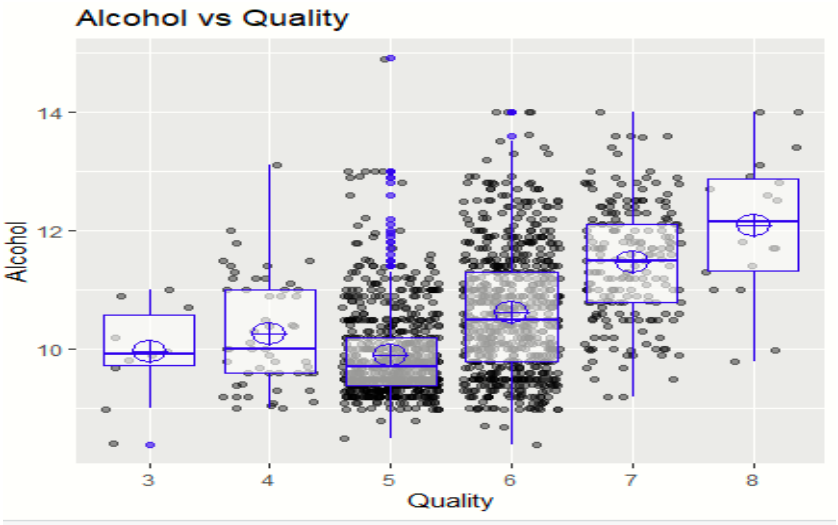

Figura 22 – Comportamento das notas de Qualidade de todos os rótulos conforme o Alcohol de acordo com o comportamento da métrica deste atributo. **(R326 a R333)**

O box plot na **Figura 22** identifica uma forte relação entre o aumento da quantidade de Alcohol com a qualidade observada na avaliação sensorial.

PARTE 10 – UMA APLICAÇÃO DA ANÁLISE DOS COMPONENTES PRINCIPAIS NA BASE DE VINHOS

Na Parte 3 abordamos de forma introdutória os objetivos da Análise Multivariada e um dos seus instrumentos, a PCA, que tem como meta encontrar uma adequada representação para extensas bases de dados viabilizando análises estatísticas com muitas variáveis.

O processo começa com uma redução de dimensão da base de dados em subconjuntos com novas dimensões e diversas variáveis não correlacionadas, resultando em uma significativa vantagem analítica.

Neste ponto, para um melhor entendimento do processo PCA, vamos abordar adicionalmente um pouco do formalismo matemático envolvido destacando alguns conceitos de álgebra linear e matricial necessários para as explanações pertinentes ao tema, sem preocupação com a demonstração de fórmulas complexas mas, essencialmente, com suas interpretações e aplicações. Algumas repetições textuais em relação ao que já foi escrito serão necessárias para evitar vazios de explicação dificultando a lógica da exposição e seu entendimento.

Parte 10.1 – Loadings e Scores

A PCA mais formalmente é baseado em uma decomposição da matriz de dados ***X*** em duas matrizes ***V*** e ***U***. As duas matrizes ***V*** e ***U*** são ortogonais. A matriz ***V*** é geralmente chamada de *matriz de carregamento* e a matriz U é chamada de *matriz de pontuações*.

A Eq. 1 mostra um exemplo de uma componente principal resultante do processo de cálculo das CPs que tem como um dos seus fundamentos realizar uma combinação linear para cada componente principal formada pelos dados originais e pelos *loadings* produzidos durante o processo que define um *modelo para a primeira componente principal;*

$$PCA1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p \qquad \text{Eq.1} \qquad \text{Fonte : [7,9]}$$

Conforme citado anteriormente (Parte 4) as constantes w_{ij} (*loadings*) podem ser entendidas como os pesos para cada variável original ao calcular cada componente principal.

Em termos gráficos a **Figura 23** mostra que a partir de uma matriz ***U*** que contém os dados originais e uma matriz ***V*** que contém as variáveis associadas aos dados, obtemos, através de uma operação matricial (*transposição, rotação e multiplicação*) um conjunto ***A*** de componentes principais espelhados por combinações lineares à semelhança da Eq.1.

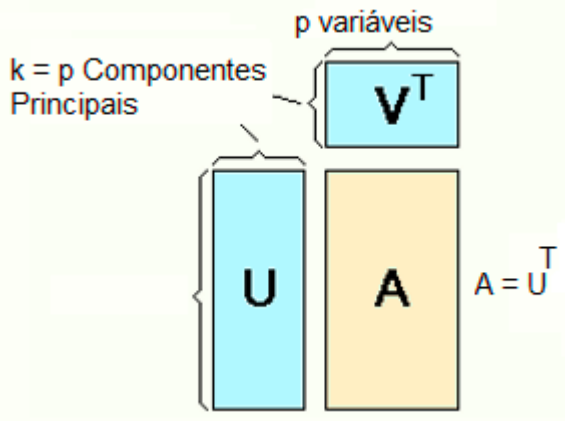

Figura 23 – Representação do processo de cálculo das operações matriciais citadas

Fonte: adaptado da página Fundamentals of Statistics de H. Lohninger.

O resultado é um quadro com um conjunto de novas variáveis chamadas de *scores* associadas a combinações lineares formados por pesos (*loadings*) combinados com os valores das variáveis originais com suas respectivas métricas conforme o modelo mostrado pela Eq. 1.

Em um contexto de uma extensa base de dados, os registros são *matricialmente e linearmente* transformados em subconjuntos de fatores denominados *Componentes Principais.* Estas têm como objetivo explicar a maior parte da variação das variáveis originais de forma mais simples do que avaliar dados em um cenário multidimensional com *p* variáveis, sem um prévio estudo e separação de suas redundâncias estatísticas.

A *quantidade da variação* retida por cada CP – representado por uma linha da matriz ***A*** – é medida por aquilo que se convencionou em termos matemáticos denominar ***valor próprio (eigenvalue)*** *ou escore do CP.*

Este processo de criação de CPs alternativos como subconjuntos dos registros originais recebe a denominação de *Redução de Dimensionalidade* que resulta em uma matriz construída com as transformações lineares conforme equação Eq.1.

Parte 10.2 – Matriz de Covariância

A matriz de covariância representa os valores da covariância de cada par de variáveis dos dados multivariados. Além disso, a covariância entre as mesmas variáveis é igual à variância, portanto, a diagonal principal mostra a variância de cada variável.

A finalidade da covariância é encontrar um valor que indica através do seu sinal como duas variáveis variam juntas sendo calculadas através da Eq. 2, conhecida a média da amostra total (população = 1599 registros numéricos).

$$cov(X,Y) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right] \qquad \text{Eq.2}$$

Fonte: [10]

A covariância sendo positiva significa que a variável ***X*** e a variável ***Y*** variam na mesma direção. Se houver uma covariância negativa, isso é interpretado como o oposto, ou seja, há uma relação negativa entre as duas variáveis. Resumindo: a covariância representa a direção de variação e o grau de associação linear entre duas variáveis.

Covariância zero significa que as variáveis (dimensões) são independentes entre si. É possível que ***X*** e ***Y*** não sejam independentes e tenham covariância zero, sendo as

chamadas variáveis descorrelacionadas. Na matriz de covariância, os valores fora da diagonal principal são diferentes de zero. Isso indica a presença de redundância nos dados indicando que existe uma certa correlação entre as variáveis.

O processo de construção da matriz de covariância ocorre em duas etapas através dos recursos (algoritmos) do RStudio, exemplificando:

-Transformação dos dados originais em uma matriz como no exemplo simplificado a Eq. 3 apresenta um espelho resumido da base de dados com suas variáveis.

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \qquad \text{Eq. 3}$$

- Construção de uma matriz de variância – covariância. A Eq. 4 apresenta um modelo da matriz de variância – covariância.

Matriz Variância – Covariância

$$\sum = \begin{bmatrix} Var_{(x_1)} & Cov_{(x_1x_2)} & Cov_{(x_1x_3)} \\ Cov_{(x_2x_1)} & Var_{(x_2)} & Cov_{(x_2x_3)} \\ Cov_{(x_3x_1)} & Cov_{(x_3x_2)} & Var_{(x_3)} \end{bmatrix} \qquad \text{Eq. 4}$$

Para o cálculo da matriz será necessário redefinir/padronizar nossas variáveis iniciais (X_1, X_2, X_3,….), uma rotina realizada automaticamente pelo algoritmo do R que calcula a covariância. A padronização evita que, por exemplo, que dados com métricas de unidades diferentes, ou por valores altos ou mínimos extremos, possam comprometer o processo de interpretação e de análise estatística

Diagonalizar significa que queremos alterar a matriz de covariância de modo que os elementos fora da diagonal sejam próximos de zero (ou seja, correlação zero entre pares de variáveis distintas). Temos ainda a vantagem que matrizes diagonalizáveis se tornam especialmente fáceis de manusear. O processo de diagonalização permite

o cálculo dos autovetores e autovalores de uma matriz quadrada. Também chamado de “eigendecomposition”.

> OBS: *Na álgebra linear, a “eigendecomposição” é a fatoração de uma matriz em uma forma canônica, em que a matriz é representada em termos de seus autovalores e autovetores. Somente matrizes diagonalizáveis podem ser fatoradas dessa maneira. Quando a matriz é fautorizada se torna uma matriz simétrica normal ou real. A decomposição é chamada de "decomposição espectral", derivada do teorema espectral que é um processo inerente do algoritmo do RStudio usado para calcular os componentes principais.*

Parte 10.3 – Autovalores autovetores da matriz de covariância.

Parte adaptada de [17]

A análise multivariada requer uma matriz de covariância e têm nos conceitos de *autovalores e autovetores* parte essencial do processo PCA, sendo que os autovalores representam a magnitude da propagação da direção dos componentes principais representados pelos seus autovetores. As linhas da matriz representam os autovetores da matriz de covariância estimada dos dados.

Um vetor $\boldsymbol{V}$ é um autovetor de uma matriz quadrada $\boldsymbol{M}$ se $\boldsymbol{Mv}$ resulta num múltiplo de $\boldsymbol{V}$, ou seja, em $\boldsymbol{\lambda V}$ (multiplicação de um escalar pelo vetor). Nesse caso, $\boldsymbol{\lambda}$ é denominado autovalor de $\boldsymbol{M}$ associado ao autovetor $\boldsymbol{V}$.

A validação da aplicação do processo PCA em uma base de dados precisa, fundamentalmente, além de reduzir a dimensão dos dados, identificar um padrão de comportamento das variáveis reagrupadas permitindo que novas variáveis chamadas de componentes principais estabeleçam subconjuntos alternativos de variáveis para um adequado processo de análise.

Geometricamente, as componentes principais podem ser definidas como localizações dos pontos amostrais (rótulos numéricos) em um novo sistema de eixos, resultado da

rotação do sistema original de eixos, na direção da máxima variabilidade dos registros das amostras de vinho.

Parte 10.4 – Demonstração gráfica resumida do processo PCA

[18 – Adaptado]

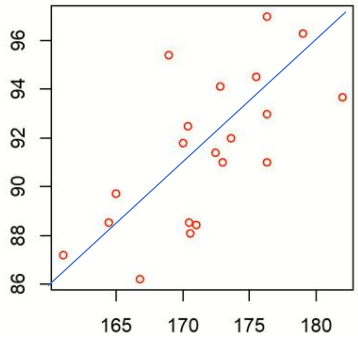

Figura 24 – Dispersão de dados e eixo componente principal

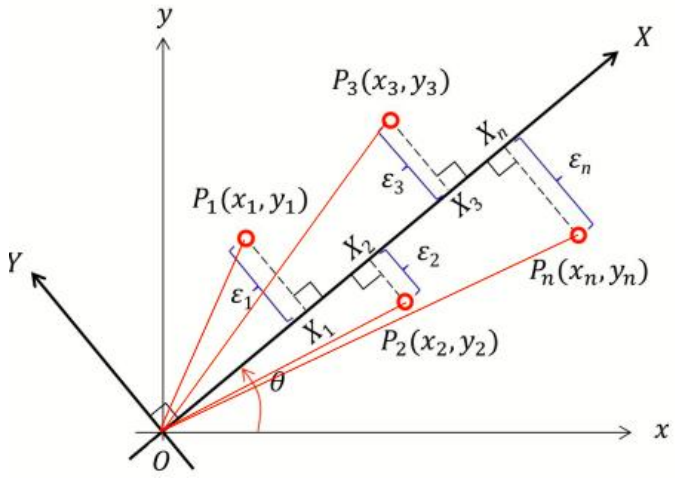

Figura 25 – Maximização das distâncias dos pontos em relação ao eixo *X*.

O eixo da componente principal é aquele que maximiza a equação Eq. 5, resultado que é obtido com a aplicação do Teorema de Pitágoras nos triângulos formados.

$$E = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2. \qquad \text{Eq. 5}$$

O resultado desse processo é a obtenção das máximas variações que irão definir cada componente principal.

A validação da aplicação do processo PCA em uma base de dados precisa, fundamentalmente, além de reduzir a dimensão dos dados, identificar um padrão de comportamento das variáveis reagrupadas permitindo que novas variáveis chamadas de Componentes Principais estabeleçam subconjuntos alternativos de variáveis para um adequado processo de análise.

Parte 10.5 – Processamento da PCA e obtenção das Componentes Principais

A *decomposição espectral* expressa uma matriz simétrica ***A*** explicitamente em termos de seus autovalores e autovetores. Isso nos fornece uma maneira de construir uma matriz com os autovalores e autovetores (ortonormais) dados. Este resultado é obtido ao escrevermos uma matriz ***A*** na forma da Eq. 6***.***

Adaptado de [23]

$$\boldsymbol{A = P\Lambda P^T} \qquad \text{Eq. 6}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \qquad \text{Eq. 7}$$

A Eq. 7 apresenta a matriz diagonal dos autovalores de ***A.***

A decomposição espectral da matriz de covariâncias nos permite calcular:

- autovalores: $\lambda_1, \lambda_2, \ldots, \lambda_p$

- autovetores padronizados: $w_1, w_2, \ldots, w_p$

- matriz formada pelos autovetores ortonormais $\boldsymbol{P} = [w_1, w_2, \ldots, w_p]$

- equações geradoras

$$w_i = (w_{i1}, w_{i2} \ldots, w_{ip})^T \quad \text{Eq. 8}$$

$$X = (X_1, X_2, \ldots, X_p) \quad \text{Eq. 9}$$

$$PC1 = w_i^T X \quad \text{Eq. 10}$$

Com a aplicação do processo PCA as variáveis originais $X_1, X_2, \ldots, X_p$ são transformadas com a decomposição espectral e pelo algoritmo do RStudio nas Componentes Principais:

$PC1, PC2, \ldots, PCp$ gerando a Matriz dos componentes principais

$$PC1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p$$

$$PC2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p \quad \text{Eq. 11}$$

$$PCp = w_{p1}X_1 + w_{22}X_2 + \cdots + w_{pp}X_p$$

Na **Figura 26**, o eixo CP**1** é a primeira direção principal ao longo da qual as amostras mostram a maior variação. O eixo **CP2** é a segunda direção mais importante e é ortogonal ao eixo **CP1,** sendo feita depois uma *adequada rotação dos eixos* para caracterizar sua ortogonalidade com novas coordenadas.

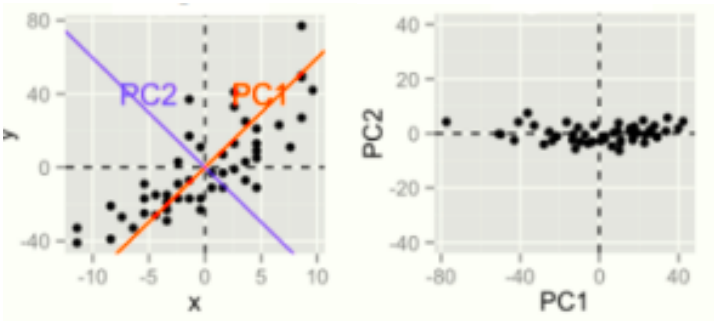

Figura 26: Exemplo gráfico da definição e rotação de CP1 e CP2.
Fonte: [10]

Tecnicamente falando, a quantidade de variância retida por cada componente principal é medida pelo chamado eigenvalue (autovalor ou valor próprio).

O processo de rotação que permite a mudança de eixos da **Figura 26** é o resultado da aplicação da matriz da Transformada de Hotelling que apresenta suas linhas *"formadas a partir dos autovetores de covariância arranjados de modo que a primeira linha, o elemento (0,0), seja o autovetor correspondente ao maior autovalor, e assim sucessivamente até que a última linha corresponda ao menor autovalor"* [10]. O processo de rotação muda as coordenadas para maximizar a soma das variâncias dos quadrados dos loadings resultando no cálculo dos *escores*. Os "inputs" do Varimax são os valores rotacionados constantes da matriz de loadings do PCA.

Como métodos de rotação devemos citar o Varimax, o Quartimax e o Equimax sendo o Varimax o mais utilizado no processo PCA para produzir uma ortogonalidade para cada componente principal conforme a seguinte definição:

> *" é um método de rotação ortogonal e pretende que, para cada componente principal, deva existir apenas alguns pesos significativos e todos os outros sejam próximos de zero, isto é, o objetivo é maximizar a variação entre os pesos de cada componente principal, daí o nome Varimax"* [18, 19]

Como resultado principal deste movimento de rotação são definidas, conforme a Eq. 10 as componentes principais que resultam de combinações lineares das variáveis originais.

> *OBS: um biplot PCA (***Figura 26***) mostra ambos os escores de PC de amostras (pontos) e cargas de variáveis (vetores). Quanto mais longe esses vetores são de uma origem do PC, mais influência que eles têm naquele PC. A Equação 1 representativa de um componente principal exibe quanta variação o componente principal captura dos dados.*

As direções dos novos eixos rotacionados são chamados de autovetores da matriz de covariância. Quando se fala em autovetores, subentende-se "autovetores de comprimento 1", (não nulos) já que a propriedade desejada é apenas a direção do vetor. Os autovetores estão associados aos autovalores, sendo os autovalores classificados em ordem decrescente para formar as *Componentes Principais*. Os autovalores correspondem ao *peso* de cada variável nas "componentes principais".

As primeiras *Componentes Principais* dos dados são as primeiras direções que explicam as variâncias máximas. Isso é equivalente aos primeiros autovetores da matriz de covariância da Base de Dados. A redução das dimensões dos dados é obtida identificando as direções principais, nas quais os dados variam agrupados em novas variáveis não correlacionadas. Os valores das Componentes Principais são o somatório dos *scores* dos rótulos fictícios dos vinhos tintos que formam uma componente.

Os números na diagonal da matriz de covariância diagonalizada são chamados de valores próprios da matriz de covariância. Os maiores valores próprios correspondem a variações significativas. Representam a informação que cada componente principal carrega.

Parte 10.6 - Interpretação geométrica dos autovalores e autovetores

Um exemplo da imagem de autovetores é apresentada na Figura 27.

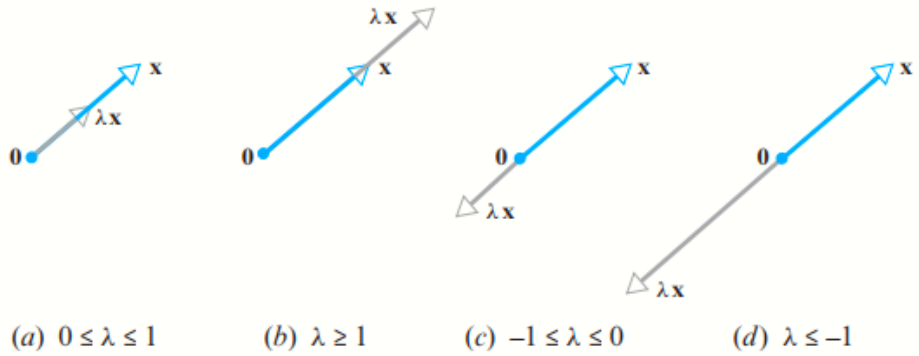

Figura 27 – Exemplo geométrico de vetores [21].

"*Em geral, a imagem de um vetor* ***x*** *pela multiplicação com uma matriz quadrada* ***A*** *difere de* **x** *em magnitude e direção. No entanto, no caso especial em que* ***x*** *é um autovetor de A, a multiplicação por A deixa a direção inalterada. Por exemplo, em* $\boldsymbol{R^2}$ *ou* $\boldsymbol{R^3}$ *, a multiplicação por A aplica cada autovetor* ***x*** *de* ***A*** *(se houver) ao longo da mesma reta pela origem determinada por* ***x****. Dependendo do sinal e da magnitude do autovalor λ associado a* ***x****, a operação* $\boldsymbol{Ax = \lambda x}$ *comprime ou expande* ***x*** *pelo fator* ***λ,*** *com uma inversão de direção no caso em que* ***λ é negativo.***" [62]

Parte 10.7 – Decomposição espectral como método de cálculo dos autovalores e autovetores

Fonte utilizada: [14]

A utilização das funções *"eigen" ou "princomp"* do RStudio executam a aplicação dos recursos do cálculo matricial e da álgebra linear para executarmos a análise de componentes principais usando a *decomposição espectral*.

Como já descrito no processo PCA as novas variáveis (CP1, CP2, ... , CPp) são uma combinação linear das variáveis originais.

O cálculo dos autovalores e dos autovetores é feito através da *Decomposição Espectral* da matriz de covariâncias. Este processo consiste em escrever uma matriz diagonalizável em termos de seus autovetores e autovalores que são calculados através de um algoritmo do RStudio.

Através da decomposição da matriz de covariâncias obtemos os componentes principais. Seus valores numéricos denominados *scores* podem ser calculados para cada elemento amostral e são analisados e interpretados através da análise de variância. Os autovetores representam o peso de cada variável nos componentes principais e os autovalores são a informação que cada componente principal carrega, isto é, seu *score*.

Este processo expressa (decompõe) uma matriz em termos de seus autovalores e autovetores – ortonormais. Pelo Teorema da Decomposição Espectral a matriz de correlação ou matriz de variâncias e covariâncias pode ser decomposta como a soma de ***P*** matrizes, cada uma relacionada com um autovetor da matriz de correlação ou da matriz de variância/covariância.

Os autovalores da matriz (**Equação 11**) formada pelos autovetores ortonormais (linearmente independentes) são representados pela diagonal principal e o determinante dessa matriz é igual ao produto dos autovalores. “Cada autovalor representa a variância de uma componente principal sendo o primeiro o de maior variabilidade” [14].

PARTE 11 – EXECUÇÃO DAS ETAPAS DA ANÁLISE DAS COMPONENTES PRINCIPAIS

As etapas principais do processo PCA podem ser assim enfatizadas:

I) Preparação dos dados: padronização da base de dados.

II) Cálculo da matriz de covariância / correlação.

III) Cálculo dos autovetores e os autovalores através da matriz de covariância pela decomposição espectral que examina as covariâncias / correlações entre as variáveis. Se usarmos a matriz de covariância os dados não precisam ser normalizados (automático pelo R).

Definição dos componentes principais em ordem decrescente: O número de autovetores escolhidos será o número de dimensões do novo conjunto de dados. autovetores = $(w_{i1}, w_{i2} \ldots, w_{ip})$. O autovetor com o maior autovalor associado, corresponde à primeira Componente Principal do conjunto de dados usado.

- Transposição dos autovetores: as linhas são os autovetores.
- Transposição dos dados ajustados: as linhas são as instâncias = rótulos numéricos, e as colunas são as variáveis/ atributos.
- Novos dados = autovetores transpostos multiplicados pelos dados ajustados transpostos.

A PCA é realizada na matriz de covariância ou matriz de correlação com dados padronizados. A covariância entre o mesmo par de dimensões resulta na variância. Uma matriz de covariância contém as covariâncias entre todos os pares possíveis de variáveis no conjunto de dados.

A padronização da base de dados é fundamental ao executar a análise da componente principal. A PCA tenta obter os dados com variância máxima, sabendo-se que a variação é alta para dados de alta magnitude. Isso inclina a PCA para características de alto valor o que justifica plenamente a padronização.

Este processo está implícito quando a base do processo PCA é uma matriz de variância e covariância. A análise de componentes principais depende somente da matriz de covariância (ou de correlação) não importando a forma da distribuição multivariada das variáveis.

Parte 11.1 – Cálculo da matriz de covariância da base de dados

Como estamos comparando atributos químicos não temos dados técnicos que possam explicar suas interdependências estatísticas. A análise do atributo qualidade estabelecido pela análise sensorial terá seu comportamento avaliado pelas influências de alterações dos atributos químicos conforme correlação verificada entre os atributos.

A matriz de covariância (**Quadro 10**) nos mostra as direções que os vetores representativos dos atributos assume quando comparados. A matriz de covariância é uma matriz quadrada que contém as variâncias e covariâncias associadas a diversas variáveis. Os elementos diagonais da matriz contêm os desvios (variâncias) das variáveis, e os elementos fora da diagonal contêm as covariâncias entre todos os possíveis pares de variáveis. A variância mede o quanto os dados estão espalhados em torno da média.

> *"A matriz de covariância exibe os valores de covariância, que medem a relação linear de cada par de itens ou variáveis. Valores de covariância positivos indicam que valores acima da média de uma variável estão associados a valores médios acima da outra variável e que valores abaixo da média de uma variável estão associados com valores abaixo da média de outra variável. Valores de covariância negativos indicam que valores acima da média de uma variável estão associados com valores médios abaixo da outra variável."* [Fonte: tutorial Software Minitab]

	aci_fix	aci_vol	ac_citr	acu_res	clor	fsd	tsd	dens	PH	sulf	gr_alc	qual
aci_fix	3.0314	-0.0799	0.2278	0.2818	0.0077	-2.8009	-6.4823	0.0022	-0.1836	0.0540	-0.1144	0.1744
aci_vol	-0.0799	0.0321	-0.0193	0.0005	0.0005	-0.0197	0.4504	0.0000	0.0065	-0.0079	-0.0386	-0.0565
ac_citr	0.2278	-0.0193	0.0379	0.0394	0.0019	-0.1243	0.2277	0.0001	-0.0163	0.0103	0.0228	0.0356
acu_res	0.2818	0.0005	0.0394	1.9879	0.0037	2.7586	9.4164	0.0009	-0.0186	0.0013	0.0632	0.0156
clor	0.0077	0.0005	0.0019	0.0037	0.0022	0.0027	0.0734	0.0000	-0.0019	0.0030	-0.0111	-0.0049
fsd	-2.8009	-0.0197	-0.1243	2.7586	0.0027	109.4149	229.7375	-0.0004	0.1137	0.0916	-0.7737	-0.4279
tsd	-6.4823	0.4504	0.2277	9.4164	0.0734	229.7375	1082.1024	0.0044	-0.3377	0.2395	-7.2093	-4.9172
dens	0.0022	0.0000	0.0001	0.0009	0.0000	-0.0004	0.0044	0.0000	-0.0001	0.0000	-0.0010	-0.0003
PH	-0.1836	0.0065	-0.0163	-0.0186	-0.0019	0.1137	-0.3377	-0.0001	0.0238	-0.0051	0.0338	-0.0072
sulf	0.0540	-0.0079	0.0103	0.0013	0.0030	0.0916	0.2395	0.0000	-0.0051	0.0287	0.0169	0.0344
gr_alc	-0.1144	-0.0386	0.0228	0.0632	-0.0111	-0.7737	-7.2093	-0.0010	0.0338	0.0169	1.1356	0.4098
qual	0.1744	-0.0565	0.0356	0.0156	-0.0049	-0.4279	-4.9172	-0.0003	-0.0072	0.0344	0.4098	0.6522

Quadro 10 – Matriz de Covariância (**R374 – R375**)

Parte 11.2 - Cálculo da Matriz de Correlações

Correlação, dependência ou associação é qualquer relação estatística entre duas variáveis dentro de uma ampla classe de relações que envolvam dependência entre duas variáveis. Muito utilizada quando avaliarmos os gráficos de círculos de correlação.

A matriz de correlação (**Quadro 11**) pode ser interpretada como a matriz de variância/covariância dos dados normalizados em relação à média e ao desvio padrão.

	aci_fix	aci_vol	ac_citr	acu_res	clor	fsd	tsd	dens	PH	sulf	gr_alc	qual
aci_fix	1.0000	-0.2561	0.6717	0.1148	0.0937	-0.1538	-0.1132	0.6678	-0.6830	0.1830	-0.0617	0.1241
aci_vol	-0.2561	1.0000	-0.5525	0.0019	0.0613	-0.0105	0.0765	0.0229	0.2349	-0.2610	-0.2023	-0.3906
ac_citr	0.6717	-0.5525	1.0000	0.1436	0.2038	-0.0610	0.0355	0.3641	-0.5419	0.3128	0.1099	0.2264
acu_res	0.1148	0.0019	0.1436	1.0000	0.0556	0.1870	0.2030	0.3551	-0.0857	0.0055	0.0421	0.0137
clor	0.0937	0.0613	0.2038	0.0556	1.0000	0.0056	0.0474	0.2004	-0.2650	0.3713	-0.2211	-0.1289
fsd	-0.1538	-0.0105	-0.0610	0.1870	0.0056	1.0000	0.6677	-0.0226	0.0704	0.0517	-0.0694	-0.0507
tsd	-0.1132	0.0765	0.0355	0.2030	0.0474	0.6677	1.0000	0.0712	-0.0665	0.0429	-0.2057	-0.1851
dens	0.6678	0.0229	0.3641	0.3551	0.2004	-0.0226	0.0712	1.0000	-0.3412	0.1477	-0.4966	-0.1752
PH	-0.6830	0.2349	-0.5419	-0.0857	-0.2650	0.0704	-0.0665	-0.3412	1.0000	-0.1966	0.2056	-0.0577
sulf	0.1830	-0.2610	0.3128	0.0055	0.3713	0.0517	0.0429	0.1477	-0.1966	1.0000	0.0936	0.2514
gr_alc	-0.0617	-0.2023	0.1099	0.0421	-0.2211	-0.0694	-0.2057	-0.4966	0.2056	0.0936	1.0000	0.4762
qual	0.1241	-0.3906	0.2264	0.0137	-0.1289	-0.0507	-0.1851	-0.1752	-0.0577	0.2514	0.4762	1.0000

Quadro 11 – Matriz de Correlações (**R374 – R375**)

PARTE 12 – APLICAÇÃO DOS PACOTES RESIDENTES NO RStudio NO PROCESSO PCA

Como já citado anteriormente, por padrão, a função *PCA( )* [**no FactoMineR**] permite, com o parâmetro "scale", a padronização dos dados a *priori*. Vamos aplicar o processo PCA nas variáveis com seus registros de instâncias ("rótulos fictícios") associadas de forma individualizadas, simulando "registros" de dados ativos. As componentes no PCA são obtidas através da diagonalização da matriz de covariância *que extrai os autovetores e autovalores associados às PC*.

Os *valores próprios* medem a quantidade de variação retida por cada componente principal. Os valores próprios mais significativos estarão nas primeiras PC e menores para as subsequentes. Ou seja, as primeiras PC correspondem às direções com a quantidade máxima de variação do conjunto de dados. Através da determinação dos valores próprios podemos identificar o número de PC a serem avaliadas.

Reforçando a abordagem na **Parte 10.1**, chamamos de *"loadings"* os *coeficientes* ($\boldsymbol{w_{1i}}$) das combinações lineares associadas às variáveis originais a partir dos quais os componentes principais são construídos.

O valor da *CP1*, por exemplo, é denominado o escore**,** que é a informação da componente para as marcas "numéricas" no grupo de vinhos associadas à PC.

Os coeficientes são denominados os *autovetores* da matriz de covariância da base de dados. Utilizando o comando ***prcomp*** calcularemos os *coeficientes* de todos os vinhos em relação às CP e seus atributos para localizar o valor da CP1 parcial em relação ao primeiro tipo de vinho e seguir um processo de confirmação de seu valor.

Após a criação do arquivo de dados, a função ***prcomp()*** do R será utilizada para calcular as componentes principais da base de dados. Teremos *n* CP escores individuais para a primeira componente principal pois temos *n* registros associados a esta componente mas que são totalizados no escore total da CP1.

É necessário definir *scale* = *TRUE*, o que resulta no dimensionamento do conjunto de dados para ter uma média 0 e um desvio padrão igual a 1, antes do cálculo das componentes principais (padronização de dados).

Parte 12.1 – PCA gráficos das variáveis e seus vetores para as dimensões 1 e 2 ou componentes principais 1 e 2

O círculo de correlações de raio unitário é apresentado na Figura 28.

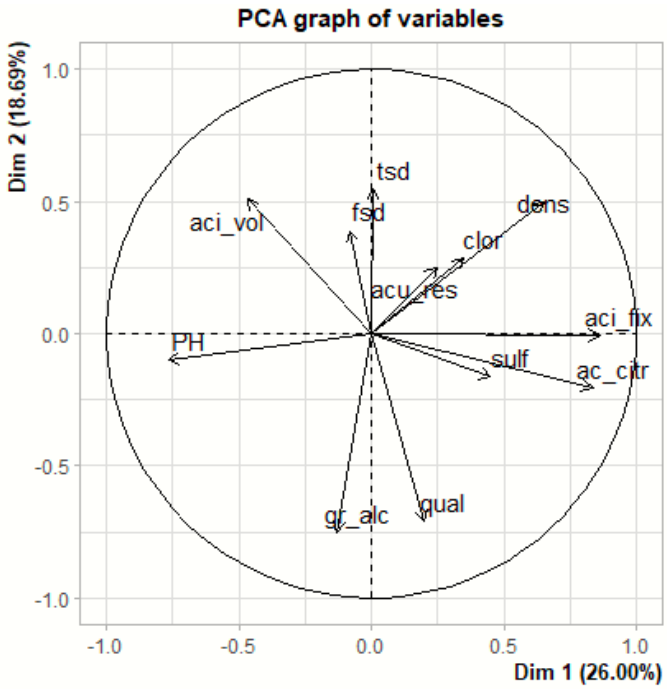

Figura 28 – Círculo de Correlações (**R392**)

Interpretação do Círculo de Correlações da Figura 28:

- A Dim 1 (Primeira Componente Principal) capta 26.00 % da variância total dos dados.

- A Dim 2 (Segunda Componente Principal) capta 18.69 % da variância total dos dados.

- No eixo vertical temos a componente 2 e no horizontal a componente 1.

- O sinal dos quadrantes determinam uma relação entre os valores das variáveis com as componentes.

- As setas captam a relação da variável e a componente.

- Quanto mais comprida a seta, maior é a importância daquela variável para a componente.

- Uma seta mais próxima do eixo representado pelo diâmetro horizontal do círculo e maior é mais importante apenas para o eixo da Dim 1.

- O inverso para uma seta maior mas paralela ao eixo da dimensão 2 é uma variável mais importante apenas para o eixo da Dim 2.

- As setas com aproximadamente 45 graus são importantes para os dois eixos.

- O quadrante 1 é positivo, portanto positivo para as duas variáveis. Maiores valores são importantes para as duas componentes.

- Por outro lado para a variável no segundo quadrante, um maior valor desta variável está associado a menores valores da dimensão 1 e maiores valores para a dimensão 2 pois o quadrante é positivo para a dimensão 2, embora esta variável seja menos importante para a dimensão 2 pois está mais próxima ou mais paralela ao eixo 1.

Esta análise permite avaliar quais são as principais variáveis (mais importantes) para cada Componente Principal. Podemos continuar a fazer comparativos entre as 4 primeiras componentes principais que captam cerca de 69 % da variância total explicada. O círculo de correlações da **Figura 28** mostra ambos os scores das PC1 e PC2 de amostras (pontos) e loadings de variáveis (vetores). Quanto mais distante esses vetores são da origem da CP, mais influência que eles têm naquela PC.

As parcelas de correlação também sugerem como as variáveis se correlacionam umas com as outras: um pequeno ângulo implica correlação positiva, um grande sugere correlação negativa, e um ângulo de 90° indica nenhuma correlação entre dois vetores.

Uma outra forma de avaliar graficamente as componentes principais é feita pelo scree plot da base de dados espelhado na **Figura 29**, que exibe quanta variação cada CP captura dos dados, sendo uma curva íngreme que se dobra rapidamente achatando em seguida.

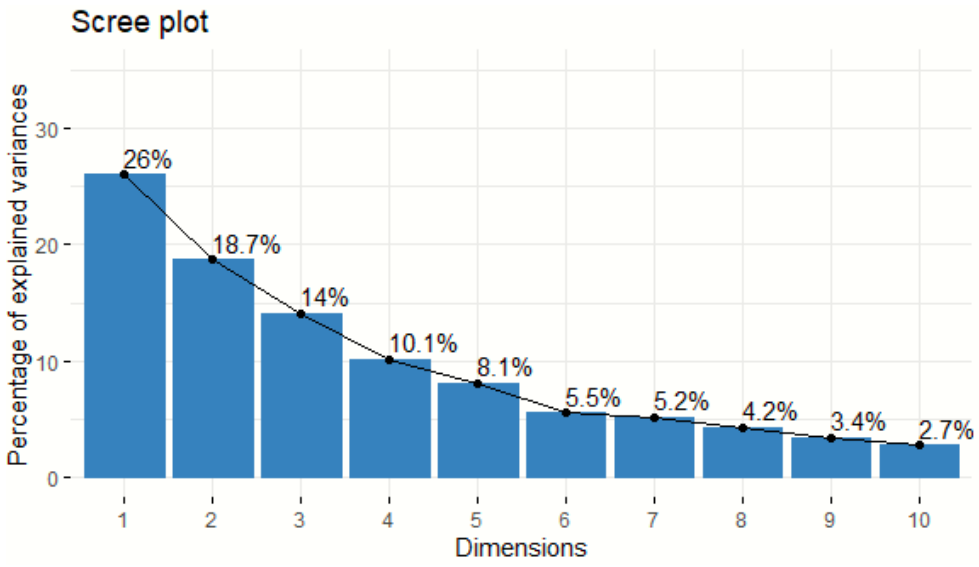

Figura 29 - Percentuais de variação explicada para cada componente principal mostrando o comportamento decrescente da variância **(R388)**.

Parte 12.2 – Desvios padrões e rotação dos componentes/loadings:

Os coeficientes dos atributos (autovetores da matriz de correlação) são descritos na tabela de autovetores do **Quadro 12** para cada componente principal. Cada variável original tem a sua participação (peso) em cada CP que devem ser analisados em módulo, que determina o grau de importância da variável. Os sinais negativos indicam, quando comparamos dois atributos de uma componente, o grau de correlação entre eles.

Através das combinações lineares dentro do contexto da PCA, conseguimos explicar a estrutura de variância e covariância de seus vetores aleatórios representantes das p-variáveis/atributos.

Com o quadro obtido pela rotação ***n x k***, com os dados padronizados alocados nas CPs correspondentes podemos avaliar a importância de cada componente. Esses valores são os coeficientes das combinações lineares que nos dizem quanto das variáveis originais são usadas na criação de cada PC. Quanto maior o coeficiente, mais importante é a variável relacionada.

```
Standard deviations (1, .., p=12):
 [1] 1.7664369 1.4973998 1.2974061 1.1023411 0.9865651 0.8139989 0.7863675 0.7112313
 [9] 0.6413429 0.5726717 0.4246212 0.2436600

Rotation (n x k) = (12 x 12):
               PC1          PC2         PC3          PC4         PC5         PC6
aci_fix  0.48799824 -0.003927653  0.16475581  0.230973208 -0.07885785  0.05492934
aci_vol -0.26514314  0.339096302  0.22687492 -0.041960132  0.29930857  0.29616967
ac_citr  0.47339567 -0.137365698 -0.10028652  0.056730530 -0.12014353  0.13705358
acu_res  0.13910990  0.167654289 -0.24355551  0.383277816  0.70919314  0.11001071
clor     0.19743676  0.189728437  0.02607924 -0.654776763  0.26643284  0.33758479
fsd     -0.04606357  0.258980019 -0.61630175  0.033924827 -0.15930224 -0.04287968
tsd      0.00397275  0.363601882 -0.54102631  0.028665877 -0.21823368  0.11518818
dens     0.36994651  0.331226297  0.16909952  0.200900070  0.20880577 -0.42580168
PH      -0.43281157 -0.065505137 -0.06939592  0.005713592  0.25770401 -0.47988493
sulf     0.25451171 -0.109505687 -0.21302230 -0.560203712  0.21519962 -0.40435462
gr_alc  -0.07317557 -0.502831357 -0.22439932  0.092019265  0.25978171  0.39203021
qual     0.11254217 -0.473196011 -0.22281198  0.037094812  0.13787034 -0.14325786
```

Quadro 12 – Desvios padrões e loadings das PC por atributo, de um total de 12 componentes **(R392)**

Com esses valores podemos simular o cálculo de qualquer score de um registro como explicado na Parte 12.8 .

Parte 12.3 – Resumo dos componentes principais

O resumo mostrado no **Quadro 13** fornece o desvio padrão de cada componente, e a proporção de variância explicada por cada componente.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.766 1.4974 1.2974 1.1023 0.98657 0.81400
Proportion of Variance 0.260 0.1868 0.1403 0.1013 0.08111 0.05522
Cumulative Proportion  0.260 0.4469 0.5871 0.6884 0.76952 0.82474
                          PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     0.78637 0.71123 0.64134 0.57267 0.42462 0.24366
Proportion of Variance 0.05153 0.04215 0.03428 0.02733 0.01503 0.00495
Cumulative Proportion  0.87627 0.91842 0.95270 0.98003 0.99505 1.00000
```

Quadro 13 – Resumo dos Componentes Principais **(R399).**

O **Quadro 14** fornece uma tabela alternativa dos autovalores com variação percentual**.**

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.12029939	26.0024949	26.00249
Dim.2	2.24220628	18.6850524	44.68755
Dim.3	1.68326270	14.0271892	58.71474
Dim.4	1.21515589	10.1262991	68.84104
Dim.5	0.97331071	8.1109226	76.95196

Quadro 14 - Tabela alternativa de autovalores com variação percentual, de um total de 12 componentes **(R402)**

Devemos observar que a soma dos ***n*** autovalores é a mesma que o traço da matriz de variância/covariância dos dados normalizados em relação à média e ao desvio.

A tabela de autovetores mostrada no **Quadro 14** para cada componente principal **(novas variáveis**) representa o banco de dados redimensionado e apresenta cada componente refletindo o comportamento conjunto dos atributos envolvidos conforme seus níveis de dispersão máxima. Essas componentes contém em ordem decrescente a maior parte da variância dos dados originais.

Uma maior correlação entre as variáveis permite que sejam usados um número menor de componentes para explicar a maior parte da variância. Conforme demonstrado, para interpretarmos a variância devemos calcular os autovalores que representarão a variância explicada por cada componente. Podemos identificar que 58.71 % da

variação total é representada pelas 3 primeiras componentes, e 68.74 % pelos quatro primeiras componentes.

Com o processo PCA iremos estudar o comportamento de 2, 3 ou 4 novas variáveis e seus atributos ao invés de analisarmos o comportamento de 12 variáveis – atributos dos vinhos – , e considerando a qualidade como a **variável de saída.**

Para um entendimento mais simplificado do processo PCA faremos algumas análises utilizando somente a CP1 e a CP2 que acumulam 44.69 % da variação total, pois não vamos explorar predições estatísticas mas, preferencialmente, a influência conjugada dos atributos na definição das CPs e suas relações com a qualidade pós avaliação sensorial.

O número de CPs para chegar a 80 %, se esse fosse um objetivo complementar de análise, alternativamente, precisaríamos estudar a **redução das variáveis** menos importantes através de um estudo de suas correlações com a qualidade se essa variável categórica tiver esse tratamento.

Esses dados complementam o biplot da **Figura 30** que considerou nas projeções apenas dois CPs para efeito de análise simplificada.

Pela tabela contida no **Quadro 14** a proporção da variância explicada pela primeira componente principal é de (3.12029939/12)*100 = 26.00 % aproximadamente.

Um autovalor > 1 indica que as PCs são responsáveis por mais variância do que uma das variáveis originais nos dados padronizados. Isso é comumente usado como um ponto de corte para o qual as PC são retidas. Isso é verdadeiro apenas quando os dados são padronizados.

Com o auxílio do RStudio podemos agora examinar os CPs destacando-se o peso de cada atributo na sua formação no Quadro 15.

	aci_fix	aci_vol	ac_citr	acu_res	clor	fsd
PC1	0.487998245	-0.26514314	0.47339567	0.13910990	0.197436761	-0.04606357
PC2	-0.003927653	0.33909630	-0.13736570	0.16765429	0.189728437	0.25898002
PC3	0.164755813	0.22687492	-0.10028652	-0.24355551	0.026079243	-0.61630175
PC4	0.230973208	-0.04196013	0.05673053	0.38327782	-0.654776763	0.03392483
PC5	-0.078857851	0.29930857	-0.12014353	0.70919314	0.266432837	-0.15930224
PC6	0.054929340	0.29616967	0.13705358	0.11001071	0.337584788	-0.04287968
PC7	-0.307234026	-0.62675834	0.24375369	0.28390966	0.229621283	-0.13820718
PC8	0.200550117	0.14601846	0.29629824	-0.17048062	-0.187178345	-0.01950877
PC9	-0.174433970	-0.06009140	-0.22149689	0.27826369	-0.420056716	-0.31732048
PC10	0.183599526	-0.15471487	-0.34556045	0.05211434	0.004011682	0.58587026
PC11	-0.256085792	0.37743761	0.62441566	0.08809686	-0.208620477	0.23752601
PC12	-0.638579962	-0.00485901	0.07040597	-0.18353764	-0.053956055	0.05268969
	tsd	dens	PH	sulf	gr_alc	qual
PC1	0.00397275	0.36994651	-0.432811573	0.25451171	-0.07317557	0.112542167
PC2	0.36360188	0.33122630	-0.065505137	-0.10950569	-0.50283136	-0.473196011
PC3	-0.54102631	0.16909952	-0.069395918	-0.21302230	-0.22439932	-0.222811976
PC4	0.02866588	0.20090007	0.005713592	-0.56020371	0.09201926	0.037094812
PC5	-0.21823368	0.20880577	0.257704007	0.21519962	0.25978171	0.137870339
PC6	0.11518818	-0.42580168	-0.479884927	-0.40435462	0.39203021	-0.143257856
PC7	-0.11067827	-0.12173130	0.186684230	-0.23242579	-0.12273258	-0.412224809
PC8	0.08998758	0.07947679	0.314804322	0.27584523	0.47092164	-0.612218730
PC9	0.12095506	-0.24906695	-0.462274939	0.45265446	-0.09665166	-0.240270954
PC10	-0.58957122	-0.04350525	-0.206553272	0.07174384	0.11013654	-0.259880974
PC11	-0.35472850	-0.23155534	-0.005091120	0.09751063	-0.32030708	0.052752121
PC12	-0.07034440	0.56646828	-0.341310456	-0.06743997	0.31766402	-0.008685953

Quadro 15 - Tabela de escores transposta para associar cada PC aos atributos

(R407 a R409)

Podemos também, através da tabela contida no **Quadro 16**, avaliar a contribuição percentual de cada atributo na formação de cada uma das componentes principais servindo de base para uma análise do efeito da alteração dos atributos, considerando-se a avaliação sensorial como estabelecida. Após a alteração de um atributo uma nova avaliação sensorial precisará ser feita ou, então, a correlação entre o atributo e a qualidade poderá servir como indicativo de impacto da alteração do atributo individualizado.

	Dim.1	Dim.2	Dim.3
aci_fix	23.814228677	0.001542646	2.71444780
aci_vol	7.030088625	11.498630182	5.14722304
ac_citr	22.410345786	1.886933506	1.00573859
acu_res	1.935156332	2.810796046	5.93192875
clor	3.898127474	3.599687980	0.06801269
fsd	0.212185291	6.707065034	37.98278532
tsd	0.001578274	13.220632892	29.27094632
dens	13.686042235	10.971085982	2.85946470
PH	18.732585790	0.429092302	0.48157934
sulf	6.477621191	1.199149557	4.53785016
gr_alc	0.535466381	25.283937361	5.03550560
qual	1.266573944	22.391446512	4.96451768

Quadro 16 - Contribuição percentual das variáveis nos CPs, de um total de 12 variáveis por componente. (**R413)**

Os valores na tabela do **Quadro 16** representam percentuais de importância dos atributos em cada dimensão investigada.

Se avaliarmos a Dim 1 (PC1) fica claro que os atributos aci_fix, aci_citr, dens e PH têm grande influência na formação desta componente que reflete a maior participação na variância total da base de dados.

Podemos também avaliar a contribuição média de cada atributo se quisermos fazer um corte e considerar as quatro primeiras CPs como mais relevantes para explicar a variação total do banco de dados.

A variância das variáveis originais dos dados normalizados/padronizados é menos importante do que a variância explicada por autovalores das novas variáveis maiores que 1, que são utilizados para a separação dos dados originais em novas variáveis representadas pelas CPs (pontos de corte).

Podemos, também, decidir quantos PCs serão consideradas para explicar a variância total através do “Scree Plot” da **Figura 29** da página 74, um gráfico de linha representativo do comportamento das componentes. A quantidade de componentes escolhida no scree plot depende da variação acumulada que se pretende aceitar como suficiente para avaliar a variação total dos dados.

Conforme a tabela dos valores próprios do **Quadro 13** que nos apresenta a importância das componentes, vemos que a variância está superior a 1 para as CPs 1, 2, 3 e 4. Portanto, usando o critério de Kaiser [27], manteríamos as quatro primeiras CP's.

Uma maneira alternativa de decidir quantas CPs vamos reter é manter o número de componentes necessários para explicar pelo menos alguma *quantidade mínima* da variância total.

Por exemplo, se for importante explicar pelo menos 69 % da variância (aprox. 70%), reteríamos as primeiras quatro CPs, como podemos ver na saída de *summary(res.PCA)* conforme o **Quadro 13** onde as primeiras quatro CPs explicam este porcentual da variância (enquanto as primeiras três componentes explicam apenas 58.7 % e podem não ser suficientes). **(R399)**

Parte 12.4 – Pontuações (escores) das CPs para cada tipo de vinho associado às componentes principais

O Quadro 17 apresenta uma tabela contendo os escores individuais de cada *"rótulo fictício"* que influencia cada PC da base de dados. Em seguida mostraremos um exemplo do processo manual de confirmação do cálculo de um *score* específico associado ao primeiro rótulo que influencia o escore da CP1.

	PC1	PC2	PC3	PC4	PC5
[1,]	-1.7798976	1.1571968	1.385831476	0.04388731	0.23289730
[2,]	-1.0048582	2.0702587	-0.009875897	-0.46680257	0.18193456
[3,]	-0.9163141	1.3926813	0.697634995	-0.34621926	0.09454233
[4,]	2.4028192	-0.2137471	-0.066748310	0.88823861	-1.52464895
[5,]	-1.7798976	1.1571968	1.385831476	0.04388731	0.23289730
[6,]	-1.7428117	1.1813614	1.135350897	0.05170367	0.03981016

Quadro 17 - Pontuações (escores) dos componentes principais, de um total de 1599 tipos de vinhos e 12 componentes principais **(R438)**

Observação: Processo de confirmação do cálculo das componentes individuais dos atributos por um algoritmo alternativo: **R440 a 450**

1) Padronização dos dados (score z)

```
   aci_fix    aci_vol    ac_citr    acu_res       clor
-0.5281944  0.9615758 -1.3910371 -0.4530767 -0.2436305
       fsd        tsd       dens         PH       sulf
-0.4660467 -0.3790141  0.5563463  1.2882399 -0.5790254
    gr_alc       qual
-0.9599457 -0.7875763
```

2) Em seguida calculamos os escores da PC1 em relação a todos os atributos do primeiro tipo de vinho (pesos = w_{1i}) – Pela rotação n x k

```
                PC1
aci_fix  0.48799824
aci_vol -0.26514314
ac_citr  0.47339567
acu_res  0.13910990
clor     0.19743676
fsd     -0.04606357
tsd      0.00397275
dens     0.36994651
PH      -0.43281157
sulf     0.25451171
gr_alc  -0.07317557
qual     0.11254217
```

3) Multiplicar o escore de cada componente do vinho1 pelo peso corresponde do PC1 para este componente. O resultado deverá ser aproximado/igual ao escore individual do primeiro "*rótulo numérico*" componente da PC1 do vinho1:

```
               PC1
[1,] -1.7798975758
```

PARTE 13 - ANÁLISE GRÁFICA DAS COMPONENTES, DOS ATRIBUTOS E DE SUAS CORRELAÇÕES

Nesta parte iremos analisar o Biplot - gráfico que projeta cada uma das observações do conjunto de dados em um gráfico de dispersão que usa a primeira e a segunda CPs como os eixos. Pela quantidade e instâncias ou indivíduos (rótulos fictícios) este gráfico não é o mais adequado para análise pois a quantidade/concentração dos registros dificulta a observação permitindo apenas uma visão melhor dos vetores dos atributos.

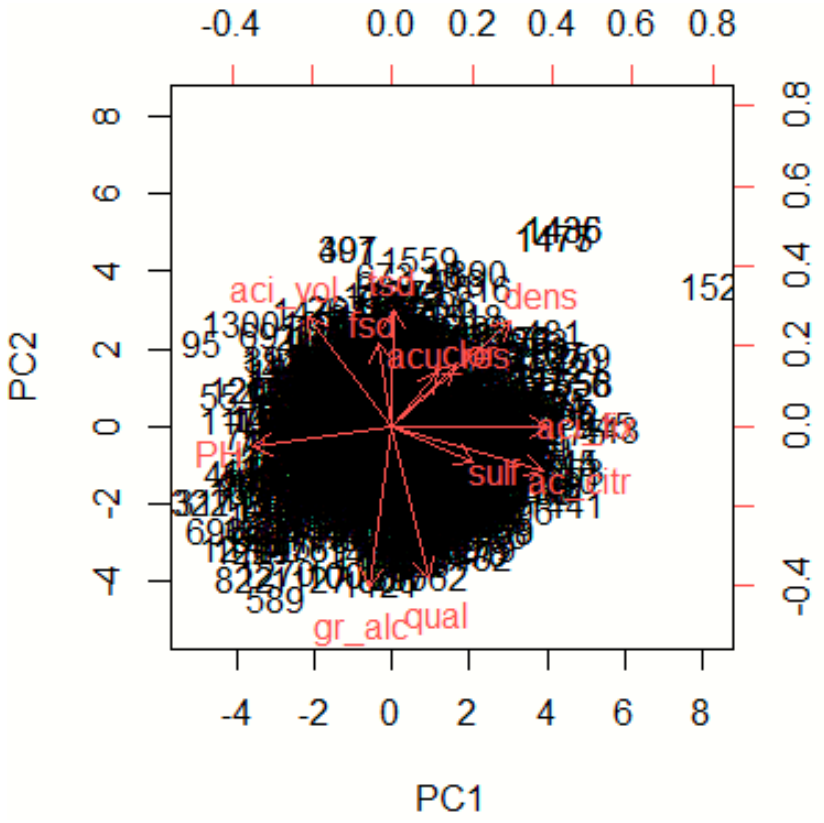

Figura 30 – Biplot das duas primeiras PC **(R458)**

Observe na **Figura 30** que o ponto (0,0) reflete a origem de um círculo de correlações garantindo que as setas no gráfico sejam dimensionadas para representar a importância dos atributos.

Podemos, também, decidir quantas CPs serão consideradas para explicar a variância total através do “Scree Plot” já visto na **Figura 29**, um gráfico de linha representativo do comportamento dos valores próprios das componentes. O gráfico que mostra o total da variância explicada por cada componente, para melhor visualização **(R372).** O mesmo gráfico da pode ser interpretado como demonstrativo da proporção de informação (variância) retida por cada PC.

Para auxiliar a análise do círculo de correlações conforme a **Figura 30** que apresenta um gráfico de dispersão temos como gerar uma Tabela de Coordenadas:

(R468)

	Dim.1	Dim.2	Dim.3	Dim.4
aci_fix	0.862018116	-0.005881268	-0.21375520	-0.254611259
aci_vol	-0.468358637	0.507762748	-0.29434892	0.046254378
ac_citr	0.836223584	-0.205691375	0.13011235	-0.062536395
acu_res	0.245728857	0.251045505	0.31599042	-0.422502888
clor	0.348759585	0.284099331	-0.03383537	0.721787334
fsd	-0.081368399	0.387796639	0.79959368	-0.037396730
tsd	0.007017612	0.544457401	0.70193085	-0.031599574
dens	0.653487179	0.495978204	-0.21939075	-0.221460404
PH	-0.764534342	-0.098087382	0.09003469	-0.006298327
sulf	0.449578886	-0.163973799	0.27637644	0.617535574
gr_alc	-0.129260026	-0.752939594	0.29113706	-0.101436617
qual	0.198798639	-0.708563632	0.28907763	-0.040891135

Quadro 18 – Tabela de coordenadas das variáveis (até dimensão 12) utilizadas na **Figura 30** que contém o gráfico de dispersão

Considerando, por exemplo, duas dimensões (2 PCs) e com as coordenadas do **Quadro 18** podemos criar um gráfico de dispersão dos atributos utilizando um gráfico circular de correlações para visualizar as relações entre as componentes principais e os atributos que influenciaram nas suas escolhas de representatividade de parcelas das variações totais (**Figura 31**).

A tabela constante do **Quadro 18** apresenta como coordenadas de cada variável suas **correlações** com as duas CPs escolhidas permitindo uma análise do tipo de influência

de cada variável na formação dessas componentes. As observações são representadas pelas suas projeções nos eixos (**Figura 31**). Vemos, por exemplo que dentro da PC1 (eixo horizontal) o **pH** está positivamente correlacionado com o **gr_acl** e na PC2 (eixo vertical) a **qualidade** está negativamente correlacionada com a densidade.

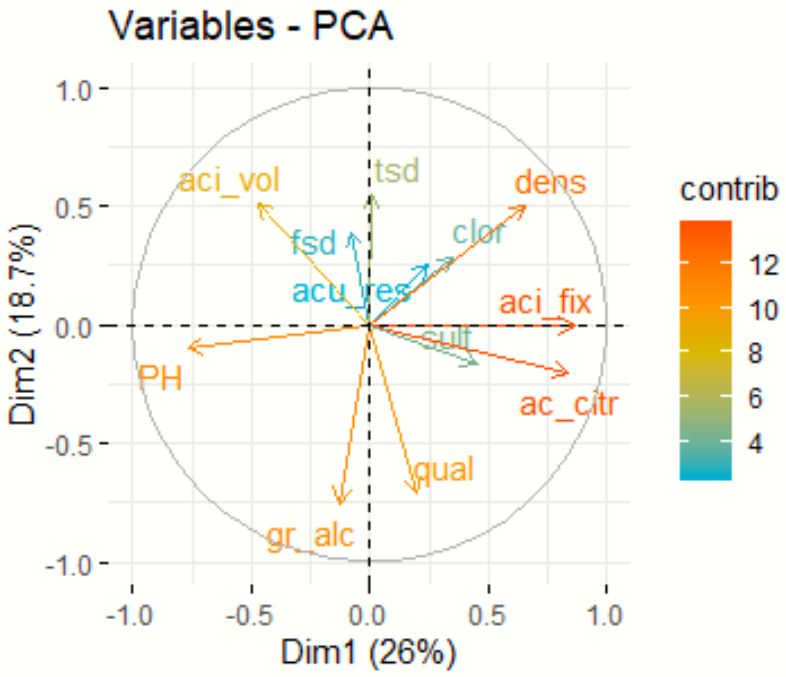

Figura 31 – Gráfico circular das variáveis que mais contribuem para cada dimensão principal em que o escurecimento das cores refletem o peso das contribuições **(R474)**

O gráfico da **Figura 31** também é conhecido como parcelas de correlação variável. Isso mostra as relações entre todas as variáveis. Pode ser interpretado como segue:

- Variáveis positivamente correlacionadas são agrupadas.

- Variáveis correlacionadas negativamente são posicionadas em lados opostos da origem do gráfico (quadrantes opostos).

- A distância vetorial entre os pontos e a origem mede a qualidade das variáveis no mapa das CPs. Pontos que estão longe da origem estão bem representadas no mapa das CPs.

Os vetores dos gráfico na **Figura 31** que estão próximos têm contribuição similar na análise das variações estudadas**.** **(R484 - R486 - R488)**

Podemos também avaliar as contribuições das variáveis através de gráficos de barras conforme a **Figura 32,** com as duas primeiras dimensões. Acima das linhas tracejadas se encontram as maiores contribuições.

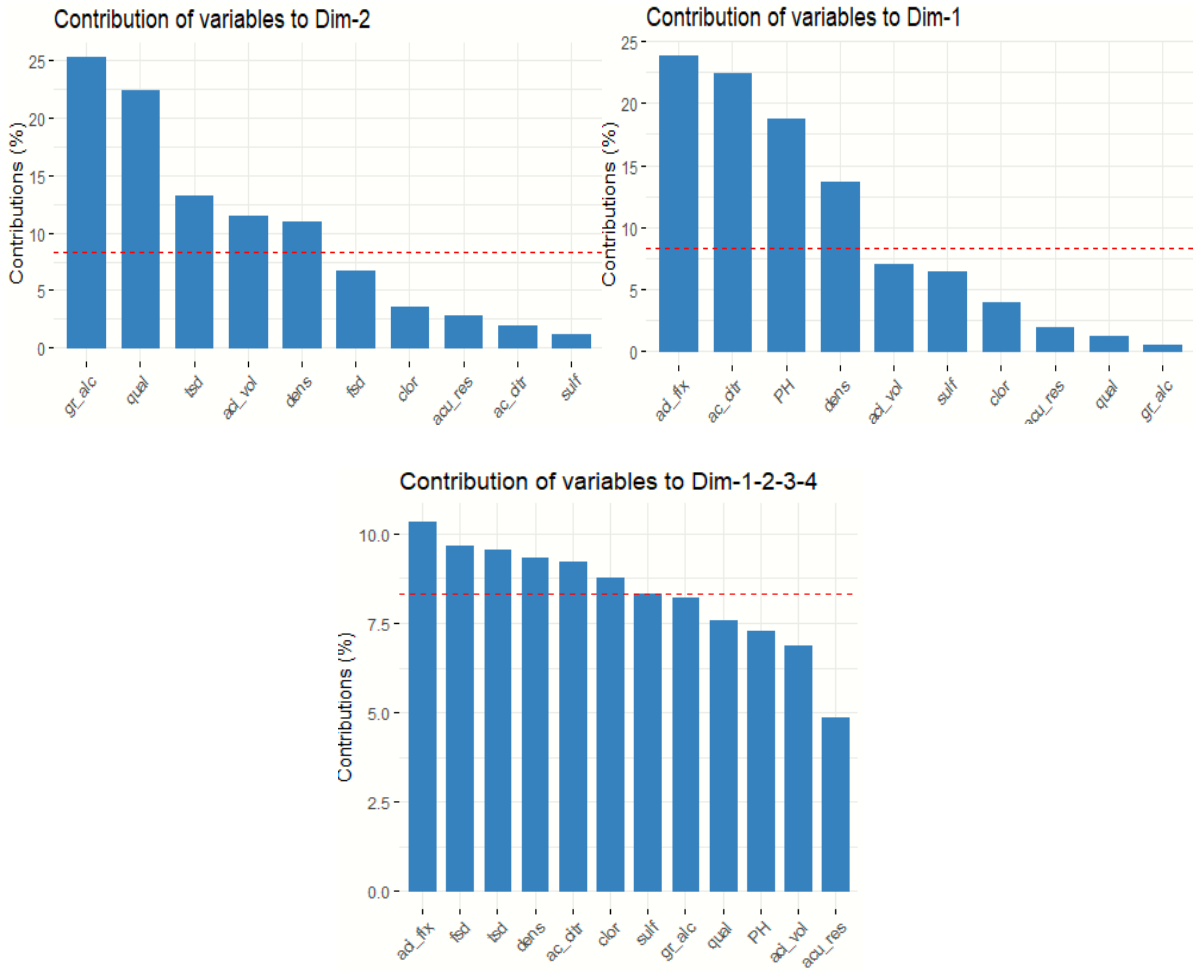

Figura 32 - Gráficos de barras com as duas primeiras e quatro dimensões espelhando de forma alternativa o peso das contribuições dos atributos.

Se quisermos considerar de forma alternativa os aspectos acima avaliados podemos também utilizar gráficos com as instâncias (“rótulos fictícios”).

Pela quantidade de rótulos (1599) e a concentração dos mesmos, teremos dificuldades de avaliação, conforme poderá ser comprovado as **Figuras 33** e **34.** Nos dois gráficos espelhados nas figuras, observa-se limites para a análise à exceção da dispersão ou concentração dos rótulos nas dimensões 1 e 2.

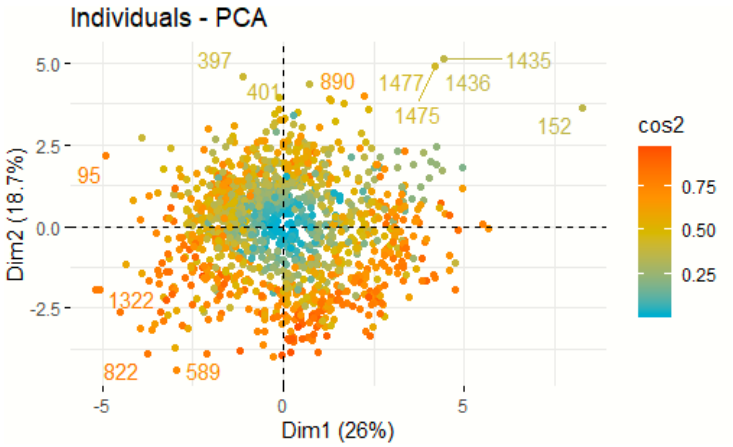

Figura 33 - Biplot de indivíduos (rótulos fictícios) e variáveis nas duas primeiras dimensões **(R491).**

.

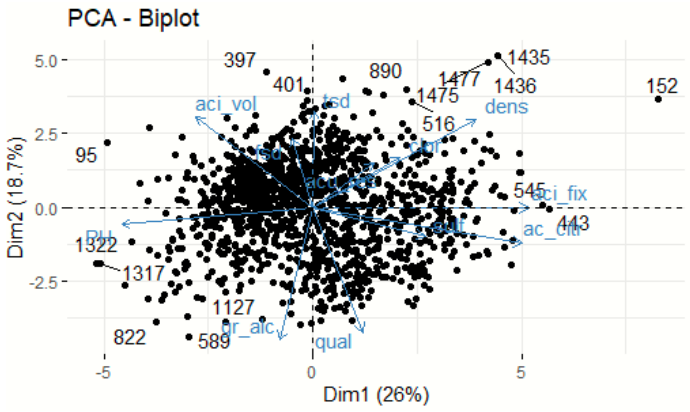

Figura 34 - Biplot de indivíduos (rótulos fictícios) e variáveis nas duas primeiras dimensões **(R496)**

Apesar das limitações pela concentração visual dos registros podemos avaliar o cosseno quadrado (***cos2***) que mostra a importância de uma componente para uma determinada observação (vetor de variáveis originais)

A visão analítica da nomenclatura ***cos2*** utilizada pelo RStudio se encontra abordada no trabalho ***Principal componente analysis*** desenvolvido pelos autores Herve Abdi e Lynne J. [49]:

> **"Squared Cosine of a Component with an Observation**
> *The squared cosine shows the importance of a cosine indicates the contribution of a component to the squared distance of the observation to the origin. It corresponds to the square of the cosine of the angle from the right triangle made with the origin, the observation, and its projection on the component and is computed as: "*

$$\cos^2_{i,\ell} = \frac{f^2_{i,\ell}}{\sum_{\ell} f^2_{i,\ell}} = \frac{f^2_{i,\ell}}{d^2_{i,g}} \qquad \textit{Eq. 11}$$

Observações adicionais sobre a variável computacional ***cos2*** ($cos^2_{i,l}$ na Eq. 11) do package "factoextra" (**Figura 33**):

- Um valor alto de ***cos2*** (cores mais escuras) indica uma boa influência de um atributo na formação das dimensões de interesse. Geometricamente, nesse caso, a variável é posicionada próximo à circunferência do círculo de correlação.
- Um valor baixo de ***cos2*** significa que a variável não é bem representada pelas PC. Nesse caso, a variável está próxima ao centro do círculo de correlações.

- Se considerarmos uma determinada variável, a soma dos valores de seus ***cos2*** em todas as CP's é igual a um.

- Para medirmos a qualidade da representação de uma variável devemos somar o valor de seus ***cos2*** em suas CPs avaliando quantos são necessários para que a soma resulte no valor 1. Quando isso acontece podemos afirmar que as componentes consideradas a representam perfeitamente e as variáveis serão posicionadas no círculo de correlações. Se uma variável se posicionar ***dentro*** do círculo significa que várias componentes são necessárias para sua boa representação.

- Se uma variável é perfeitamente representada por apenas duas CPs principais (Dim.1 e Dim.2), a soma do ***cos2*** nessas duas PC é igual a um. Nesse caso, as variáveis serão posicionadas no círculo de correlações

Utilizando agora um gráfico de barras **(Figura 35)** do cos2 e com as dimensões de interesse de 1 a 3 que espelha a qualidade da representação [10].

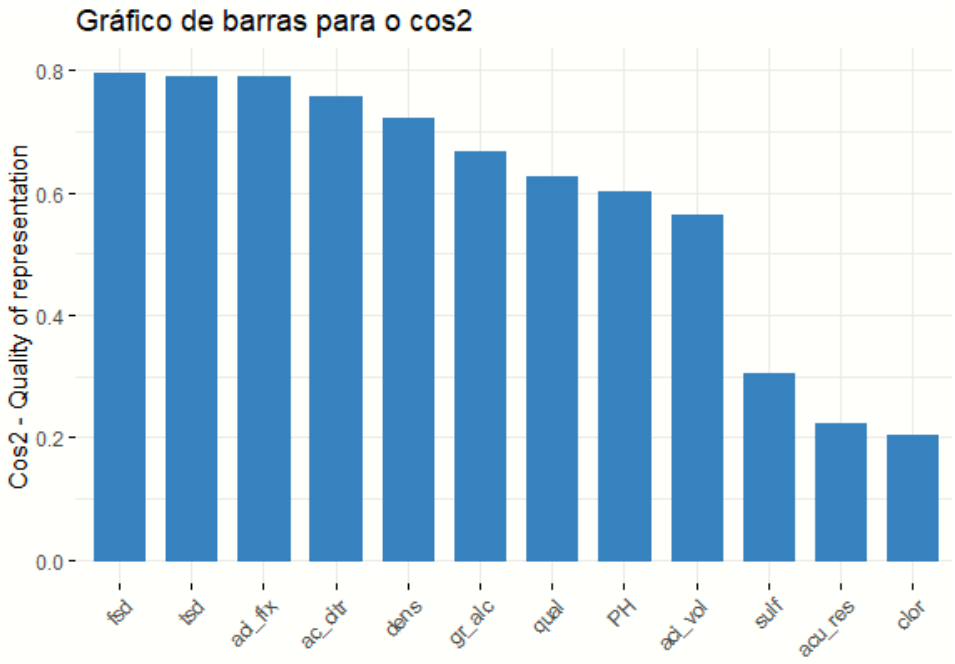

Figura 35 – Gráfico de barras para o ***cos2*** (**R517**).

PARTE 13.1 – Estudo do Círculo de Correlações

Uma alternativa para avaliarmos as contribuições das variáveis para a formação das CPs é estudarmos o Círculo de Correlações

O **Quadro 19** evidencia as coordenadas dos atributos considerando a faixa de -1 a + 1. Nas figuras as cores mais escuras da legenda revelam a proximidade das correlações por grupo de atributos e permite uma análise do tipo de influência de cada variável na formação desses componentes. As observações são representadas pelas suas projeções nos eixos. Vemos, por exemplo que dentro da PC1 (eixo horizontal) o **pH** está negativamente correlacionado com o **gr_acl** e na PC2 (eixo vertical) a **qual** (qualidade) está negativamente correlacionada com a **dens** (densidade). As correlações entre os atributos estão demonstradas no **Quadros 9** e **11** e nas **Figuras 10** e **11**.

	Dim.1	Dim.2
aci_fix	0.862018116	-0.005881268
aci_vol	-0.468358637	0.507762748
ac_citr	0.836223584	-0.205691375
acu_res	0.245728857	0.251045505
clor	0.348759585	0.284099331
fsd	-0.081368399	0.387796639
tsd	0.007017612	0.544457401
dens	0.653487179	0.495978204
PH	-0.764534342	-0.098087382
sulf	0.449578886	-0.163973799
gr_alc	-0.129260026	-0.752939594
qual	0.198798639	-0.708563632

Quadro 19 - Coordenadas das variáveis considerando as duas primeiras componentes principais **(R501)**.

Na **Figura 36** podemos constatar que os atributos positivamente correlacionados aparecem agrupados próximos uns dos outros dentro do círculo e os negativamente correlacionados aparecem em posição oposta à origem. Quanto mais distante da origem, melhor serão representados no mapa. Neste caso devemos destacar os

atributos **aci_fix, ac_citr, gr_alc, dens, PH** e **qual** como influenciadores principais das CPs (CP1 e CP2). **(R506 – R513)**

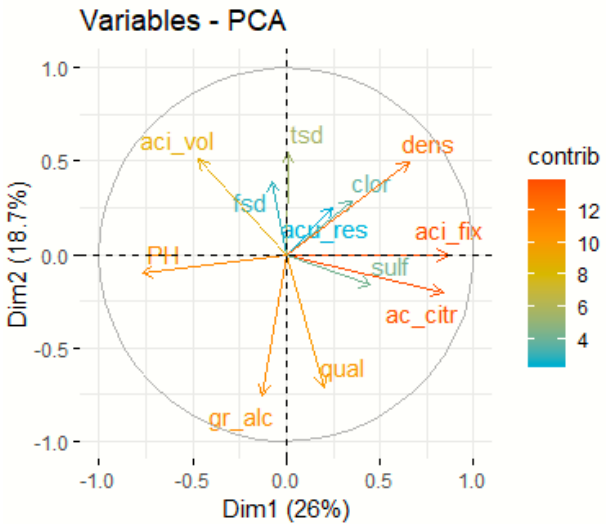

Figura 36 – Qualidade da representação (contribuição).

Podemos complementar a análise gráfica das contribuições dos atributos em termos de importância para a composição das primeiras cinco componentes principais com um mapa geral de fatores espelhado na **Figura 37.**

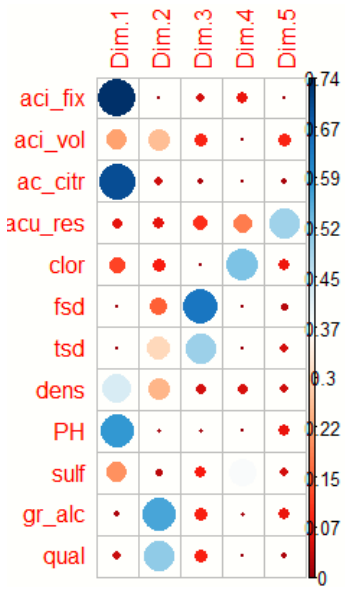

Figura 37 – Mapa Geral de Fatores **(R523)**

O mapa nos indica os atributos mais importantes que influenciam a definição de cada CP da base de dados de vinhos tintos, mostrando que a importância de alguns componentes pode se diluir à medida em que a ordem dos componentes aumenta.

Observe-se que a qualidade tem pouca importância na formação das componentes, sendo muito mais consequência da escolha dos componentes químicos (atributos) de cada “rótulo”. Destaque-se a importância dos atributos **aci_vol, fsd, tsd, dens** e **gr_alc** na definição da CP1 e CP2.

CONCLUSÕES

A Dissertação teve como principais objetivos:

i) aplicar a metodologia PCA em uma extensa base de dados para selecionar Componentes Principais que permitem análises numéricas e gráficas. O que permitiu refletir as características estatísticas da base com apenas quatro Componentes Principais que representaram cerca de 69 % da variação total observada nos dados.

Para avaliar com mais profundidade as quatro componentes principais podemos, de forma complementar, comentar a participação de cada atributo em cada componente de maneira comparativa.

Como cada leitor pode ter um interesse específico em avaliar o papel de cada componente químico na composição do vinho escolhido, deixamos para que cada leitor explore este aspecto conforme seu próprio interesse.

Como não conhecemos as marcas, mas somente suas composições associadas a "rótulos fictícios", estudos estatísticos entre eles dificultará comparativos de valor que estão restritos à avaliação sensorial.

ii) como importante objetivo complementar, aproveitando-se da *variável de saída qualidade*, estudou-se as relações entre todos os vinhos amostrados e seus atributos com as notas obtidas através de avaliação sensorial. Esta abordagem permitiu que fosse caracterizada uma verificação de tendência de mudança de qualidade associada às variações nas métricas dos atributos, caso isso fosse desejado com base nos registros dos vinhos.

iii) o objetivo e motivação em utilizar o RStudio como linguagem computacional utilizada nesta Dissertação, foram fundamentados na constatação da crescente amplitude da utilização deste software na Análise Estatística Multivariada, seja pela simplicidade de sua utilização, pela diversidade da documentação disponível (na ampla bibliografia teórica e aplicada) e pela multiplicidade de pacotes estatísticos (algoritmos) já desenvolvidos, que abrangem todos os instrumentos matemáticos e

estatísticos necessários, atendendo aos diversos níveis de complexidade de estudos e aplicações nessas áreas do conhecimento.

BIBLIOGRAFIA – LIVROS E OUTRAS FONTES DE CONSULTA

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos e J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Página: http://www3.dsi.uminho.pt/pcortez/Home.html

[2] Vinho: wikipedia https://pt.wikipedia.org/wiki/Vinho

[3] https://blog.famigliavalduga.com.br/entenda-o-que-e-a-analise-sensorial-do-vinho-e-como-ela-e-feita/

[4] https://www.clubedosvinhos.com.br/o-esmagamento-da-uva-e-sua-influencia-no-vinho/

[5] Análise de dados através de métodos de estatística multivariada – Uma abordagem aplicada – Sueli Aparecida Mingoti – Editora ufmg

[6] Red and White Wine Quality – https://rpubs.com/Daria/57835

[7] Principal Components Analysis – Steven M. Holland Tutorial with R - Department of Geology, University of Georgia, Athens, GA 30602-2501 – Page 3

[8] Clovis Paiva – Porquê e quando é necessário normatizar os dados:

https://medium.com/tentando-ser-um-unic%C3%B3rnio/porqu%C3%AA-e-quando-%C3%A9-necess%C3%A1rio-normalizar-os-dados-92e5cce445aa

[9] An Introduction to Applied Multivariate Analysis with R (Chapter 3) - Brian Everitt and Torsten Hothorn - Springer

[10] Alboukadel Kassambara, Practical Guide To Principal Component Methods in R. sthda.com. Edition 1

[11] https://archive.ics.uci.edu/ml/datasets/wine+quality
http://www3.dsi.uminho.pt/pcortez/wine/ - CSV Files – Wine Red and White Quality Data Sets.

[12] ggplot2 – Elegant Graphics for Data Analysis – Hadley Wickham – Second Edition - Springer

[13] Methods of Multivariate Analysis – Alvin C. Rencher (pag. 447- 448) – Wiley Interscience

[14] Analise Fatorial Daniel Abud Seabra Matos/Erica Castilho Rodrigues (pag. 39...) - Enap

[15] http://makemeanalyst.com/explore-your-data-range-interquartile-range-and-box-plot/

[16] Álgebra Matricial em R – Adelmo https://rpubs.com/adelmofilho/AlgebraMatricial

[17] Análise de Componentes Principais – Simone Vasconcelos http://www.ic.uff.br/~aconci/PCA-ACP.pdf

[18] Basics of Multivariate Analysis II – Principal Component Analysis -Takeshi Furuhashi]

[19] Rotação e interpretação das componentes principais

https://www.google.com/search?q=rota%C3%A7%C3%A3o+no+pca&oq=&aqs=chrome.0.35i39i362l8...8.53600j0j7&sourceid=chrome&ie=UTF-8

[20] Denize Ivete. Reis https://smolski.github.io/livroavancado/analisf.html

[21] Elementary Linear Algebra 11th edition – Howard Anton | Chris Rorres – Wiley

[22] Red wine quality Data Analysis – Jose A Dianes - https://github.com/jadianes/data-science-your-way/blob/master/apps/wine-quality-data-analysis/README.md

[23] Analise Multivariada- Hedibert Freitas Lopes

http://hedibert.org/wp-content/uploads/2015/02/AnaliseMultivariada-aula3.pdf

[24] http://sillasgonzaga.com/material/cdr/ggplot2.html

[24] Learn ggplot2 Using Shiny App – Keon-Woong Moon – Springer - 2016

[25] Brian Everitt, Torsten Hothorn**.** An introduction to Applied Multivariate Analysis with R. Springer New York Dordrecht Heidelberg London - 2011

[26] **Exploratory Multivariate Analysis by Example Using r** – François Husson, Sébastien Lê e Jérome Pagès – CRC Press

[27**] Gilbert Strang. Linear Algebra and Learning From Data** – MIT

[28] https://www.kaggle.com/tsilveira/wine-r

[29] Thomas D. Wickens, The Geometry of Multivariate Statistics. Lawrence Eribaum Associates, Inc.

[30] https://blog.famigliavalduga.com.br/entenda-o-que-e-a-analise-sensorial-do-vinho-e-como-ela-e-feita/

[31] https://www.clubedosvinhos.com.br/o-esmagamento-da-uva-e-sua-influencia-no-vinho/

[32] https://github.com/jtsou/Red-Wine-Analysis-with-R - Chemical Properties

[33] https://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_KMO_Bartlett.pdf

[34] J. Douglas Carroll, Paul E. Green, Mathematical Tools for Applied Multivariate Analysis. Revised Edition. Academic Press

[35] [11] Package factoextra : Extract and Visualize the Results of Multivariate Data Analyses: http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analyses-and-elegant-visualization

[36] Package FactoMine: *http://factominer.free.fr/*

[37] STHDA – Statistical tools for high-throughput data analysis. http://www.sthda.com/english/

[38] Debbie l. Hans-Vaughn, Applied Multivariate Statistical Concepts. Routledge,Taylor & Francis Group

[39] Joseph F. Hair, William C. Black, Barry J. Babin e Rolph E. Anderson ,Multivariate Data Analysis. Multivariate Data Analysis. Pearson New International Edition.

[40] François Husson, Sébastien Lê, Jérôme Pagès. Exploratory Multivariate Analysis by Example Using R. Second Edition. CRC Press. A Chapman & Hall Book

[41] Zakaria Jaadi. A step by step explanation of Principal Components Analysis**.** https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[42] Jeff Jauaregui, Principal component analysis with linear algebra. Associate professor in the Department of Mathematics at Union College. http://www.math.union.edu/~jaureguj/PCA.pdf

[43] Kaiser, Henry F. 1961. “A Note on Guttman’s Lower Bound for the Number of Common Factors.” *British Journal of Statistical Psychology* 14: 1–2.

[44] Jekaterina Novikova. Predicting Red Wine Quality: Exploratory Data Analysis *https://api.rpubs.com/jeknov/redwine*

[45] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520. Hotelling, H (1936). "Relations between two sets of variates"

[46] Hotelling, Harold (December 1936). "Relation between two sets of variates".

[47] WikpediA - Harold Hotelling

[48] Outliers, o que são e como tratá-los em uma análise de dados. https://www.aquare.la/o-que-sao-outliers-e-como-trata-los-em-uma-analise-de-dados/

[49] Harold Hotelling – Analysis of a complex of statistical variables into Principal Components. https://babel.hathitrust.org/cgi/pt?id=wu.89097139406&view=1up&seq=5

[50] Principal component analysis - Herve Abdi Lynne J. Williams
https://personal.utdallas.edu/~herve/abdi-awPCA2010.pdf

[51] Paulo Cortez: Modern Optimization with R, Use R! series, Springer, 2014

Anexo I – Script RStudio do processo PCA desenvolvido na Dissertação na forma de arquivo R.

```
#=============================================
# SCRIPT R DA DISSERTACAO
#=============================================
# Pacotes carregados e rotinas de inicialização
#=============================================
rm(list=ls(all = TRUE))
graphics.off() # Clear data objects values
shell("cls")   # and windows or plots
#=============================================
library(ggplot2)        # Utilizado
library(gridExtra)
library(readr)
library(dplyr)
library(corrplot)       # Utilizado
library(plotly)         # Utilizado
library(GGally)         # Utilizado
library(caret)
library(rpart)
library(rpart.plot)
library(shiny)
library(plotly)
library(devtools)       # Utilizado
library(readxl)
library(reprex)
library(psych)          # Utilizado
library(rminer)
library(MASS)           # Utilizado
library(ggcorrplot)     # utilizado
library(GGally)
library(FactoMineR)     # utilizado
library(Factoshiny)
library(factoextra)     # utilizado
library(missMDA)
library(paran)
library(ggthemes)
library(ade4)
library(BBmisc)         # Utilizado
library(gghalves)
```

```r
library(base)
library(data.table)
library(xlsx)
library(jmv)
library(magrittr)
library(jmvconnect)
library(pcaMethods)
library(stats)
library(ltm)
library(psych)
library(ggalt)
library(foreign)
library(fBasics)
library(sp)
library(tidyverse)
library(rgl)
library(plot3D)
library(NMF)

#=====================================================
# Alguns comandos de verificação (testes/conferência)
#  serão utilizados mas não farão parte explícita do
#  código da Dissertação.
#=====================================================
# Parte 5- IMPORTAÇÃO DA BASE DE DADOS E IDENTIFICAÇÃO
# VARIÁVEIS
#=====================================================
# Criacao do vetor de titulacao dos atributos para os
#  vinhos tintos.

atributos_vinhos<- c("label","aci_fix","aci_vol","ac_citr",
                     "acu_res","clor", "fsd","tsd","dens","PH",
                     "sulf","gr_alc","qual")

vinhos<-read.csv2("vinhos_tintos_UCI.csv",
                     col.names = atributos_vinhos,header=FALSE,
                     skip=1)

vinhos        # Print base de dados
vinhos[2:13] # A coluna 1 (label) e 14 (variável categórica)
              # não serão utilizadas de forma explícita
#-------------------------------------------------------------
```

```
vinhos2<-vinhos[2:13]  # Cópia 2
vinhos2
vinhos3<-vinhos[2:13]  # Cópia 3
vinhos3
#-----------------------------------------------------------------
vinhos<-data.frame(vinhos[,c(2:13)]) # Arquivo que será utilizado
                                     # sem label e tipo
vinhos
#-----------------------------------------------------------------
# Transformando todas as variáveis em numéricas
index<-c(1:12)
vinhos[index]<-apply(vinhos[,index],2,
      function(x) as.numeric(as.character(x)))
# vinhos$label=as.factor(vinhos$label)
str(vinhos[1:12])
#-----------------------------------------------------------------
# Resumo da base de dados - espelho parcial
head(vinhos[1:12]) # Corresponde às 6 primeiras linhas da tabela
                   # de dados incluindo todas as colunas de variáveis.
str(vinhos) # Data Frame
#=================================================================
```

```
# Normalização da base de dados.Arquivos criados para uso posterior.
score_z = scale(vinhos[1:12],center=T)
score_z
DATA<-vinhos[1:12]
str(DATA)
DATA.scaled<-scale(DATA,center=TRUE,scale=TRUE)
DATA.scaled
#=================================================================
```

```
# Checar presença de registros totalmente nulos
sapply(vinhos[1:12], function(x)all(is.na(x)))
#=================================================================
dados_pca<-vinhos       # arquivo de dados auxiliar para uso
                        #  posterior.
dados_pca
#=================================================================
# Parte 6 - Testes de adequação da base dados para o processo PCA
# Parte 6.1 - Condições de validação dos testes de adequação
# Dimensões dos dados
dim(vinhos)
dim(vinhos)[1]>5*dim(vinhos)[2]
# Confirmação escalas numéricas
str(vinhos)
# Analisando a matriz de correlação
# Cálculo da matriz de correlações dos dados originais
vinhos
cor.c<-round(cor(vinhos),digits=2)
cor.c
# Verificação gráfica da matriz de correlações
ggcorrplot(cor(vinhos[1:12]),hc.order =TRUE,type="upper",
       lab = TRUE, insig ="blank")
# describe(vinhos)

# Teste de Esfecidade de Barlett
cortest.bartlett(cor.c,n=1599)
```

```
136
137 # KMO
138 vinhos
139 KMO(vinhos)
140 # Parte 6.2 - Conclusão dos testes de adequação da base.
141 #============================================================
142 # PARTE 7 - Informações estatísticas descritivas principais
143 #          da base de dados
144 #============================================================
145 # Histograma da Base Geral de vinhos
146 atributos_vinhos<- c("label","aci_fix","aci_vol","ac_citr",
147                      "acu_res","clor", "fsd","tsd","dens","PH",
148                      "sulf","gr_alc","qual")
149
150 vinhos<-read.csv2("vinhos_tintos_UCI.csv",
151                   col.names = atributos_vinhos,header=FALSE,
152                   skip=1)
153 vinhos
154 attach(vinhos)
155 grafico_lista <- vector("list",
156                         length = length(atributos_vinhos)-2)
157 for(i in 2:13){
158   grafico_lista[[i-1]] <- plot_ly(x = as.formula(vinhos[i]),
159       type = 'histogram', name = atributos_vinhos[i])
160 }
```

```
161 subplot(grafico_lista,nrows = 4)
162 detach(vinhos)
163 #============================================================================
164 #
165 summary(vinhos)
166 vinhos
167 vinhos<-data.frame(vinhos[,c(2:13)]) # Recuperando Data Frame
168 vinhos
169 index<-c(1:12)
170 vinhos[index]<-apply(vinhos[,index],2,
171                      function(x) as.numeric(as.character(x)))
172 # vinhos$label=as.factor(vinhos$label)
173 str(vinhos[1:12])
174 #============================================================================
175 # Parte 7.1 - Box Plot da base geral de vinhos
176 graphics.off()
177 par(mar=c(2,2,2,1))
178 oldpar = par(mfrow = c(3,6))
179 for ( i in 1:12 ) {
180   boxplot(vinhos[[i]])
181   mtext(names(vinhos)[i], cex = 0.8, side = 1,
182         line = 2)
183 }
184 par(oldpar)
185
186 # Box plot da base de dados geral
187 boxplot(vinhos,scale.= TRUE)
```

```r
# Box plot vinhos classificados
boxplot(vinhos[1:6],scale.= TRUE)
boxplot(vinhos[9:12],scale.= TRUE)

# Efeitos da variação métrica dos atributos
# nas medidas descritivas da base de dados
# Quadro comparativo da média, desvio padrão, mediana
# e amplitude dos dados
psych::describeBy(vinhos[1:12])

# Parte 8 - Estudo do comportamento da variávelqualidade
# Tabela de notas
table(vinhos$qual)
#
# notas e suas medidas estatísticas descritivas
vinhos$classe_vinho<-ifelse(vinhos$qual<=6,'ruim','bom')
psych::describeBy(vinhos,vinhos$classe_vinho)
print(vinhos$classe_vinho)

# Gráficos da matriz de correlações
#  com a qualidade.
# corrplot(cor.c,type="lower")
corrplot(cor.c, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
ggcorrplot(cor(vinhos[1:12]),hc.order =TRUE,type="lower",
           lab = TRUE, insig ="blank")
# Checar presença de registros duplicados
# vinhos
# vinhos[duplicated(vinhos[2:13],fromLast=TRUE)]
```

```r

# Parte 9 - Análise gráfica do comportamento da
#  qualidade e suas correlações através de boxplots.
# Parte 8.4.1 -  Análise complementar das relações
# individuais da qualidade com alguns atributos
# através de plots bivariados.

str(vinhos)
# Alcohool vrs Quality
ggplot(aes(x= factor(qual), y= gr_alc), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Qualidade',
       y= 'Grau Alcoólico',
       title= 'Grau Alcoólico vrs Qualidade')

# Volatile Acidity vrs Quality
ggplot(aes(x= factor(qual), y= aci_vol), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Qualidade',
       y= 'Acidez Volátil',
       title= 'Acidez Volátil vrs Qualidade')
```

```r

# Fixed Acidity vrs Quality
ggplot(aes(x= factor(qual), y= aci_fix), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Qualidade',
       y= 'Acidez Fixa',
       title= 'Acidez Fixa vrs Qualidade')

# Sulphates vrs Quality
ggplot(aes(x= factor(qual), y= sulf), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Qualidade',
       y= 'Sulfatos',
       title= 'Sulfatos vrs Qualidade')

# Chrolides vrs Quality
ggplot(aes(x= factor(qual), y= clor), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +


  labs(x= 'Qualidade',
       y= 'Cloretos',
       title= 'Cloretos vrs Qualidade')

# Density vrs Quality
ggplot(aes(x= factor(qual), y= dens), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Qualidade',
       y= 'Densidade',
       title= 'Densidade vrs Qualidade')

# Alcohol vrs Density
ggplot(aes(x= factor(gr_alc), y= dens), data = vinhos[1:12]) +
  geom_jitter( alpha = .4) +
  geom_boxplot( alpha = .6,color = 'blue2')+
  stat_summary(fun = "mean", geom = "point", color = "blue2",
               shape = 10, size = 6) +
  labs(x= 'Densidade',
       y= 'Grau Alcoólico',
       title= 'Densidade vrs Grau Alcoólico')
```

```
294
295 # Ácido Cítrico vrs Qualidade
296 ggplot(aes(x= factor(qual), y= ac_citr), data = vinhos[1:12]) +
297   geom_jitter( alpha = .4) +
298   geom_boxplot( alpha = .6,color = 'blue2')+
299   stat_summary(fun = "mean", geom = "point", color = "blue2",
300                shape = 10, size = 6) +
301   labs(x= 'Qualidade',
302        y= 'Ácido Cítrico',
303        title= 'Qualidade vrs Ácido Cítrico')
304
305 # PH vrs Qualidade
306 ggplot(aes(x= factor(qual), y= PH), data = vinhos[1:12]) +
307   geom_jitter( alpha = .4) +
308   geom_boxplot( alpha = .6,color = 'blue2')+
309   stat_summary(fun = "mean", geom = "point", color = "blue2",
310                shape = 10, size = 6) +
311   labs(x= 'Qualidade',
312        y= 'PH',
313        title= 'Qualidade vrs PH')
```

```
314
315 # Acúcar Residual vrs Qualidade
316 ggplot(aes(x= factor(qual), y=acu_res), data = vinhos[1:12]) +
317   geom_jitter( alpha = .4) +
318   geom_boxplot( alpha = .6,color = 'blue2')+
319   stat_summary(fun = "mean", geom = "point", color = "blue2",
320                shape = 10, size = 6) +
321   labs(x= 'Qualidade',
322        y= 'Açucar Residual',
323        title= 'Qualidade vrs Açucar Residual')
324
325 # Quality vrs Alcohol
326 ggplot(aes(x= factor(qual), y= gr_alc), data = vinhos[1:12]) +
327   geom_jitter( alpha = .4) +
328   geom_boxplot( alpha = .6,color = 'blue2')+
329   stat_summary(fun = "mean", geom = "point", color = "blue2",
330                shape = 10, size = 6) +
331   labs(x= 'Quality',
332        y= 'Alcohol',
333        title= 'Alcohol vs Quality')
334
335 # PARTE 10 - Execução das etapas da análise dos componentes principais
336
```

```r
# Padronização da base de dados
score_z = scale(vinhos[1:12],center=T,scale=TRUE)
round(head(score_z),4)
summary(score_z) # Sumário com dados padronizados
# write.xlsx(score_z,file,"padronizado.xlsx")

vinhos.n<-score_z   # Arquivo auxiliar
vinhos.n

# Cálculo da matriz de covariância dos dados originais
#=============================================================
atributos_vinhos<- c("label","aci_fix","aci_vol","ac_citr",
                     "acu_res","clor", "fsd","tsd","dens","PH'
                     "sulf","gr_alc","qual")

vinhos<-read.csv2("vinhos_tintos_UCI.csv",
                  col.names = atributos_vinhos,header=FALSE,
                  skip=1)

vinhos       # Print base de dados
vinhos[2:13] # A coluna 1 (label) e 14 (variável categórica)
             # não serão utilizadas de forma explícita
index<-c(1:13)
vinhos[index]<-apply(vinhos[,index],2,
                     function(x) as.numeric(as.character(x)))
```

```r
# vinhos$label=as.factor(vinhos$label)
str(vinhos[2:13])
vinhos<-vinhos[2:13]
vinhos
str(vinhos)
#=============================================================
# Cálculo da Matriz de Covariância/variância
vinhos
cov.c<-round(cov(vinhos),digits=4)
cov.c
#
# Cálculo da matriz de correlações dos dados originais
cor.c<-round(cor(vinhos),digits=4)
cor.c

# Cálculo da matriz de correlações
cor.cPR<- round(cor(score_z),digits=4)
cor.cPR

# Aplicação dos pacotes residentes no RStudio
#  no processo PCA

# PCA - Gráficos das variáveis e seus vetores
res.pcag<-PCA(vinhos) # PCA graph of variables
res.pcag
summary(res.pcag) # Estudar saida do comando
fviz_eig(res.pca2,addlabels = TRUE, ylim = c(0,35))
```

```r

# Desvios padrões e rotação dos componentes/loadings
res.pca<-prcomp(vinhos,center = TRUE,scale=TRUE)
res.pca
res.pca2 <- PCA(vinhos)
res.pca2 # Resultados disponiveis para geração de tabelas
fviz_eig(res.pca)


# Resumo dos componentes principais
summary(res.pca)
# Componentes Principais com o peso de
# cada atributo na sua formação
eig.val <- get_eigenvalue(res.pca2) # Obtemos tambem as coordenadas
eig.val

# Mesma tabela mas transposta para associar cada PC aos atributos
# Matriz transposta
pca_corr<-prcomp(dados_pca,center=TRUE,scale=TRUE)
M_transp = t(pca_corr$rotation)
print(M_transp)
#============================================================
# Contribuição percentual das
#  variáveis nos Componentes Principais
res.pca2$var$contrib
head(res.pca2$var$contrib, 12)
corrplot(res.pca2$var$contrib, is.corr=FALSE)
```

```r


fviz_pca_ind(res.pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
# Centering and scaling the supplementary individuals
ind.scaled <- scale(ind.sup,
                    center = res.pca$center,
                    scale = res.pca$scale)
# Coordinates of the individividuals
coord_func <- function(individuals, loadings){
  r <- loadings*individuals
  apply(r, 2, sum)
}
pca.loadings <- res.pca$rotation
pca.loadings
# Parte 13.7 - As pontuações (scores) dos componentes
# principais para cada tipo de vinho associado aos
# componentes principais

```

```r
head(res.pca$x)
#=========================================================
# Processo de confirmação do cálculo dos componentes individuais
# dos atributos por um algoritmo alternativo
pca_corr<-prcomp(dados_pca,center=TRUE,scale=TRUE)
score_z_vinho1 <- score_z[1,]
pca_corr # Loadings (pesos)
valores<-score_z_vinho1 # referente ao vinho1
valores
calculos<-summary(pca_corr)$x
calculos
score_z = scale(vinhos,center=T)
pca_corr$x
#=========================================================

# Análise Gráfica dos componentes, dos atributos
#  e de suas correlações
# Analisando um biplot das duas primeiras
#                componentes principais
# Visualizando os resultados com um Biplot
biplot(res.pca, scale = 0)

# Scree Plot
res.pca <- prcomp(vinhos[1:12],center=TRUE,scale=TRUE)
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 35))
# corrplot(res.pca$var$cos2,is.corr = FALSE)
fviz_screeplot(res.pca, addlabels = TRUE, ylim = c(0, 35))
```

```r

# Coordenadas das variáveis
head(res.pca2$var$coord)
res.pca2$var$coord

# Gráfico de dispersão dos atributos
#  dentro dos componentes principais.

# Gráfico das variáveis
fviz_pca_var(res.pca, col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800",
            "#FC4E07"),repel = TRUE
            #  Avoid text overlapping
)

# Gráficos de Barras das Contribuições das
# variáveis para as duas primeiras dimensões

# Contributions of variables to PC1
fviz_contrib(res.pca, choice = "var", axes = 1, top = 10)
# Contributions of variables to PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 10)
# Considerando os componentes de 1 a 4
fviz_contrib(res.pca, choice = "var", axes = 1:4, top = 12)
```

```r
# Gráficos das instâncias (rótulos fictícios)
fviz_pca_ind(res.pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping (slow if many points)
)
# Biplot de indivíduos (rótulos) e variáveis
fviz_pca_biplot(res.pca, repel = TRUE)

# Círculo de correlações.
# Distribuição das variáveis nas
#  duas primeiras comopnentes principais
res.pca2$var$coord

# Parte 1
fviz_pca_var(pca_corr,col.var = "steelblue")
# Parte 2
fviz_pca_var(pca_corr,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)

```

```r
# Quadro de corrrelações entre atributos ???????????
ggcorrplot(cor(vinhos), hc.order = TRUE, type = "lower",
           lab = TRUE, insig = "blank")

# Qualidade da representação
fviz_cos2(res.pca,choice="var",axes=1:3,
          title=("Gráfico de barras para o cos2"))
?fviz_cos2


# Mapa Geral de Fatores
corrplot(res.pca2$var$cos2, is.corr=FALSE)

res.pca2$var$contrib
# corrplot(res.pca2$var$contrib, is.corr=FALSE)




# Resumo dos componentes principais
summary(res.pca2)

eig.val <- get_eigenvalue(res.pca2) # Obtemos tambem as coordenadas
eig.val

```

```r
# Revertendo os sinais
# res.pca$rotation <- 1*res.pca$rotation
# res.pca$rotation
# res.pca$rotation <- -1*res.pca$rotation
# res.pca$rotation

# Desvios padrões e rotação dos componentes
# res.pca2 <- prcomp(vinhos,center=TRUE,scale=TRUE)
# res.pca2
# print(res.pca)

# Obtendo os autovalores e os autovetores
# Mostrando os principais componentes por vinho
res.pca$rotation

# Scores dos componentes principais e revertendo os sinais
res.pca2$x<--1*res.pca$x
res.pca2$x

# Espelho parcial dos primeiros scores por tipo de vinho
head(res.pca2$x)

# Resumo dos Componentes Principais e variância explicada
# por cada componente

# Tabela alternativa de autovalores com percentuais de
#  variação percentual
get_eig(pca_corr)
```

```r

# Mesma tabela mas transposta para associar cada PC aos atributos
# Matriz transposta
M_transp = t(pca_corr$rotation)
print(M_transp)

#
# contribuição percentual das variáveis nos quatro primeiros
#  Componentes Principais

res.pca2$var$contrib
head(res.pca2$var$contrib,12)
corrplot(res.pca2$var$contrib, is.corr=FALSE)

head(res.pca2$var$contrib,12)

res.pca2$var$contrib
head(res.pca2$var$contrib, 4)
corrplot(res.pca2$var$contrib, is.corr=FALSE)

cor(vinhos[2:12])
# Mudanças na qualidade versus matriz de correlações (R406)
cor.c

# pairs(vinhos[2:12], col = vinhos[2:12]$qual,
#       pch = vinhos[2:12]$qual)

# Tabela dos loadings (coeficientes/pesos)
head(res.pca$x)
```

```
#============================================================

vinhos
vinhos<-data.frame(vinhos[,c(1:12)]) # Recuperando Data Frame
vinhos

index<-c(1:12)
vinhos[index]<-apply(vinhos[,index],2,
                     function(x) as.numeric(as.character(x)))
# vinhos$label=as.factor(vinhos$label)
str(vinhos[1:12])

ggplot(aes(x=aci_fix), data = vinhos) +
  geom_bar(fill = 'steelblue', colour='darkgrey', alpha= 0.8)
vinhos$aci_fix

ggcorr(vinhos[1:12], geom = "blank", label = TRUE,
       hjust = 0.9, layout.exp = 2) +
  geom_point(size = 8, aes(color = coefficient > 0,
                           alpha = abs(coefficient) > 0.35)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = FALSE, alpha = FALSE)
```

```
options(repr.plot.width=7.5, repr.plot.height=6)  #Setting the plot size

theme_set(theme_minimal(10))
## Defining the variables we will explore:
variables <- c('qual', 'gr_alc','aci_vol',
               'dens', 'acu_res')

## Plotting ggpairs for the selected attributed:
ggpairs(vinhos[1:12][variables], aes(alpha=0.3))

## I experimented adding jitter, but the result was
# not satisfactory:
# ggpairs(wines[variables], aes(alpha=0.3),
#   lower = list(continuous=wrap("points", position="jitter")))

## Changing plot size to 4x3
#(Ref: http://blog.revolutionanalytics.com/2015/09/
# resizing-plots-in-the-r-kernel-for-jupyter-notebooks.html)
options(repr.plot.width=4, repr.plot.height=3)
## Plotting the histogram
ggplot(aes(x=qual), data = vinhos[1:12]) +
  geom_bar(fill = 'steelblue', colour='darkgrey', alpha= 0.8)
```

```
#================================================
# Estudo do comportamento da variávelqualidade
# Tabela de notas
table(vinhos$qual)
# Classificação dos vinhos por faixa de
# notas e suas medidas estatísticas descritivas
vinhos$classe_vinho<-ifelse(vinhos$qual<=6,'ruim','bom')
psych::describeBy(vinhos,vinhos$classe_vinho)
print(vinhos$classe_vinho)

#================================================
```