|  |  |  |  |
| --- | --- | --- | --- |
| **Controle de Versões** | | | |
| **Versão** | **Data** | **Autor** | **Notas da Revisão** |
| 1.0 | 28/01 | Felipe Marques | Criação do documento. |
| 1.1 | 28/01 | Felipe Maques | Atualização da introdução e estatística. |

# **Introdução**

A base de dados utilizada neste estudo foi obtida através de uma pesquisa no site de [site de pesquisa de dados do google](https://datasetsearch.research.google.com/).

A base de dados selecionada lista características de vinhos tintos e brancos separadamente. Estas características são listadas na tabela 1.1.

|  |  |
| --- | --- |
| Características | Descrição |
| Acidez fixa | Os ácidos orgânicos da uva (tartárico, málico e cítrico) e os originados da fermentação alcoólica (succínico, lático, etc.), que formam a acidez fixa do vinho. |
| Acidez volátil | O ácido acético é elaborado na fermentação alcoólica, formando a chamada acidez volátil. |
| Ácido Cítrico | Ácido cítrico está presente nas uvas em menor quantidade, sendo que nos vinhos apresenta pouca ou nenhuma presença. Por exemplo, nos tintos ele desaparece devido à ação de bactérias láticas (fermentação malolática). Também é possível dizer que é fresco, porém, em alguns casos, pode conferir um paladar amargo no final. |
| Açúcar residual | A concentração residual de açúcar refere-se à quantidade de sólidos de açúcar em um determinado volume de vinho após o final da fermentação, além de qualquer adição de açúcar. A concentração residual de açúcar é expressa em gramas por litro ou como uma porcentagem do peso em volume. Por exemplo, um vinho com açúcar residual de 0,2% contém dois gramas de açúcar em um litro de vinho.  Os vinhos secos costumam estar próximos dessa faixa de 0,2% até 0,5%. No entanto, cada país tem a sua classificação. No Brasil, a lei diz que para vinhos tranquilos e frisantes serem considerados secos, o máximo seria 4 gramas por litro de glicose, e, para espumantes, ditos nature (ou seja, sem adição de açúcar no chamado licor de expedição), até 3 gramas (confira a tabela da legislação brasileira no quadro). No entanto, nos Estados Unidos, por exemplo, não há qualquer indicação sobre esse tema. Já nas regulamentações da União Europeia, os vinhos secos com acidez moderada podem conter não mais que 9 gramas por litro de açúcar residual, exceto quando o ácido estiver acima de 7 g/l. s ácidos orgânicos da uva (tartárico, málico e cítrico) e os originados da fermentação alcoólica (succínico, lático, etc.), que formam a acidez fixa do vinho. |
| Cloretos | O teor de cloretos nos vinhos é muito variável, normalmente inferior a 50 mg L-1. Os vinhos obtidos de vinhedos situados mais próximos  do mar apresentam teores mais elevados. O teor de cloretos nos vinhos pode aumentar em função de colagens realizadas, ou também em virtude da adição de ácido clorídrico – que não é permitida. |
| Dióxido de enxofre livre | O dióxido de enxofre livre corresponde àquele encontrado na forma de SO2 e de combinações minerais do tipo H2 SO3 , HSO3 - e SO2- 3.  O “dióxido de enxofre livre” encontra-se presente nos vinhos e mostos sob diversas formas quimicamente definidas e diferenciadas com teores dependentes de fatores de ordem física e físico-química – o pH, a temperatura e cinética dos diversos equilíbrios estabelecidos (Curvelo-Garcia, A.S.,1988). |
| Dióxido de enxofre total | O dióxido de enxofre total corresponde à soma do dióxido de enxofre livre mais o combinado existente no vinho. |
| Densidade | A densidade do vinho está relacionada principalmente ao seu teor alcoólico e de açúcares residuais. Assim, os valores médios de densidade das amostras de vinhos tintos se mantiveram entre 0,9961 (amostra Ft) e 0,9981 (amostra Dt). |
| PH | O pH representa a concentração de íons de hidrogênio livres dissolvidos no vinho. O valor é expresso pelo logaritmo da concentração de íons hidrogênio, que, no caso dos vinhos brasileiros, é variável de 3,0 até 3,8, dependendo do tipo de vinho (branco ou tinto), da cultivar e da safra. |
| Sulfato | Os sulfatos são ânions minerais, sempre presentes nos vinhos e provenientes da própria uva como constituinte normal e da oxidação do ácido sulfuroso, assumindo maior importância nos vinhos fortemente sulfitados e submetidos depois a arejamentos. Nesse sentido, o teor de sulfatos aumenta progressivamente durante a conservação do vinho.  Outra eventual causa de incorporação de sulfato no vinho consiste na aplicação do gesso (CaSO 4 ) para a correção da acidez. A adição de ácido sulfúrico é rigorosamente proibida. |
| Álcool | O álcool é um resultado natural da produção do vinho.  Ou seja, ele não é acrescentado durante o processo, mas sim passa a existir graças a determinadas reações químicas.  A glicose e a frutose são dois tipos de açúcar presentes nas uvas, que podem ter um sabor mais adocicado ou mais ácido na hora de degustar a fruta de acordo com a sua concentração.  Quando uvas mais doces são utilizadas para a produção de um vinho, isso quer dizer que a bebida tende a ser mais alcoólica, já que os açúcares servirão de alimento para microorganismos chamados de leveduras.  Essas leveduras existem tanto nas próprias uvas quanto nas vinhas e nos barris de carvalho, utilizados para o amadurecimento da bebida.  Depois de se alimentarem de açúcar, elas sintetizam dióxido de carbono e etanol, que é o álcool dos vinhos.  Esse é o tão famoso processo de fermentação: a sintetização das leveduras, que transformam a glicose e a frutose em álcool.  Logo, o teor alcoólico do vinho existe naturalmente, graças a esse processo.  Existe uma exceção bastante conhecida, que é o caso dos vinhos do Porto, que recebem aguardente vínica para que o processo de fermentação seja interrompido.  Nesse caso, o álcool é adicionado “artificialmente”, embora ainda pudesse surgir naturalmente caso a fermentação prosseguisse da maneira que ocorre com os outros tipos de vinho. |
| Qualidade | Avaliação da qualidade do vinho. |

**Tabela 1.1:** Características analisadas do vinho.

Em relação a análise dos dados, a mesma foi dividida em duas partes. A primeira é feita uma analise estatística dos dados dos vinhos tintos e banco. A segunda parte é realizado uma analise aplicando técnicas de data mining.

# Análise d**os Dados**

## ***Análise Estatística com R***

A análise estatística foi realizada utilizando a linguagem r, sendo esta uma é uma linguagem de programação multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados. foi criado originalmente por ross ihaka e por robert gentleman no departamento de estatística da universidade de auckland, nova zelândia.

### ***Análise de Dados de Vinho Tinto***

Seção responsável por apresentar os dados da análise estatística dos dados do vinho tinto.

#### ***Estatísticas***

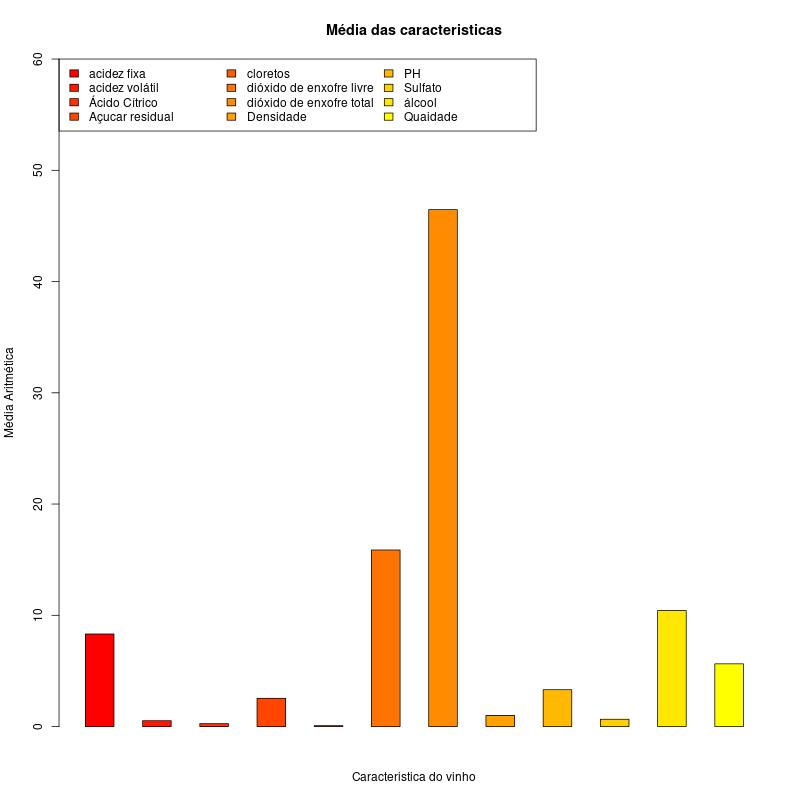
Nesta seção são apresentados resultados de medidas de posição e de dispersão dos dados. Estas estatísticas são listadas na tabela 2.1.

|  |  |
| --- | --- |
| Medidas | Descrição |
| Média | Esse tipo de média funciona de forma mais adequada quando os valores são relativamente uniformes. Por ser sensível aos dados, nem sempre fornece os resultados mais adequados. Isso porque todos os dados possuem a mesma importância (peso). |
| Mediana | A mediana é uma medida de tendência central da Estatística que corresponde ao valor central de um conjunto de valores ordenados.  O termo “mediana” refere-se a “meio”. Dado um conjunto de informações numéricas, o valor central corresponde à mediana desse conjunto. Dessa forma, é importante que esses valores sejam colocados em ordem, seja crescente ou decrescente. Se houver uma quantidade ímpar de valores numéricos, a mediana será o valor central do conjunto numérico. Se a quantidade de valores for um número par, devemos fazer uma média aritmética dos dois números centrais, e esse resultado será o valor da mediana. |
| Desvio Padrão | O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Ou seja, o desvio padrão indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados. |
| Variância | Dado um conjunto de dados, a variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central (médio).  Quanto menor é a variância, mais próximos os valores estão da média; mas quanto maior ela é, mais os valores estão distantes da média. |
| Maior Valor | Apresenta o maior valor de cada característica. |
| Menor Valor | Apresenta o menor valor de cada característica. |

**Tabela 2.1:** Medidas de posição e de dispersão dos dados.

O arquivo “DCWine/DadosGerados/EstatisticaRedWine.csv” contém os resultados das medidas listadas na tabela 2.1.

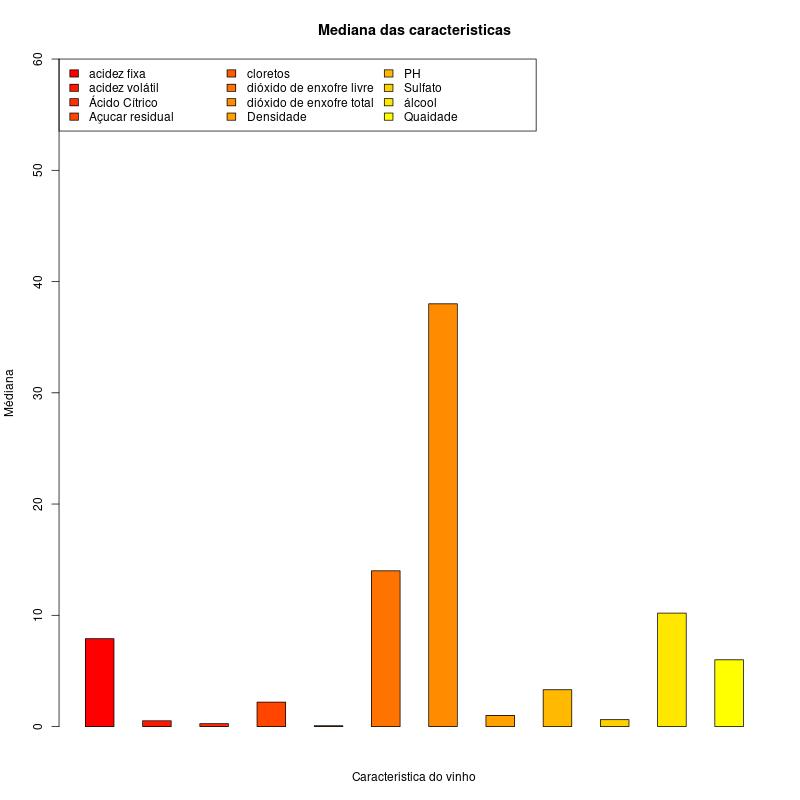
Os gráficos gerados a partir destes resultados são apresentados a seguir. A figura 2.1 apresenta a média das características.



**Figura 2.1:** Media das características do vinho tinto

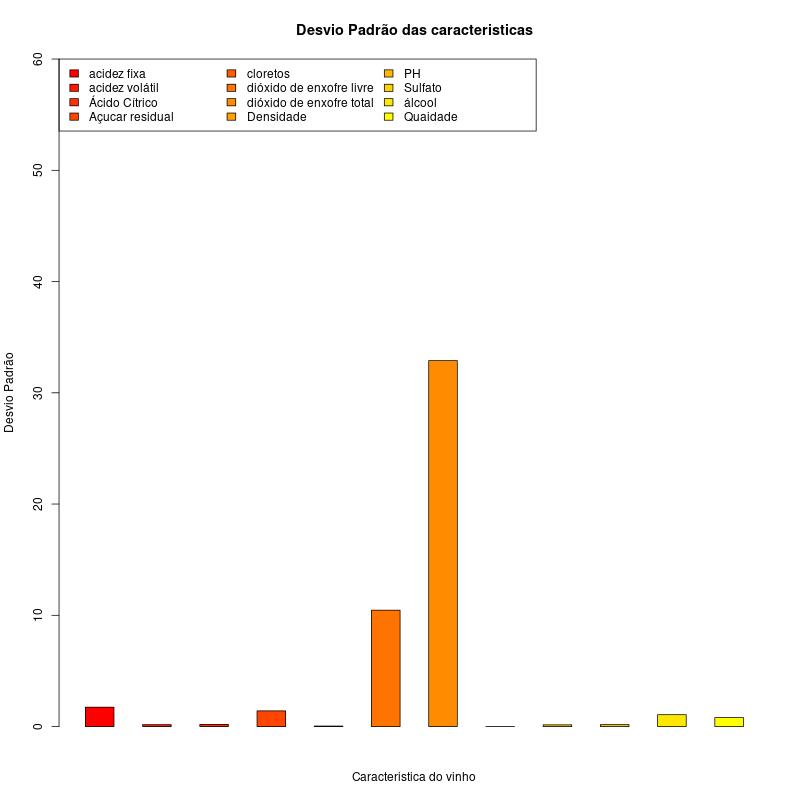
#### [Realizar analise dos dados apresentados na figura 2.1]

Na figura 2.2 é apresentado a mediana das características.

**Figura 2.2:** Mediana das características do vinho tinto

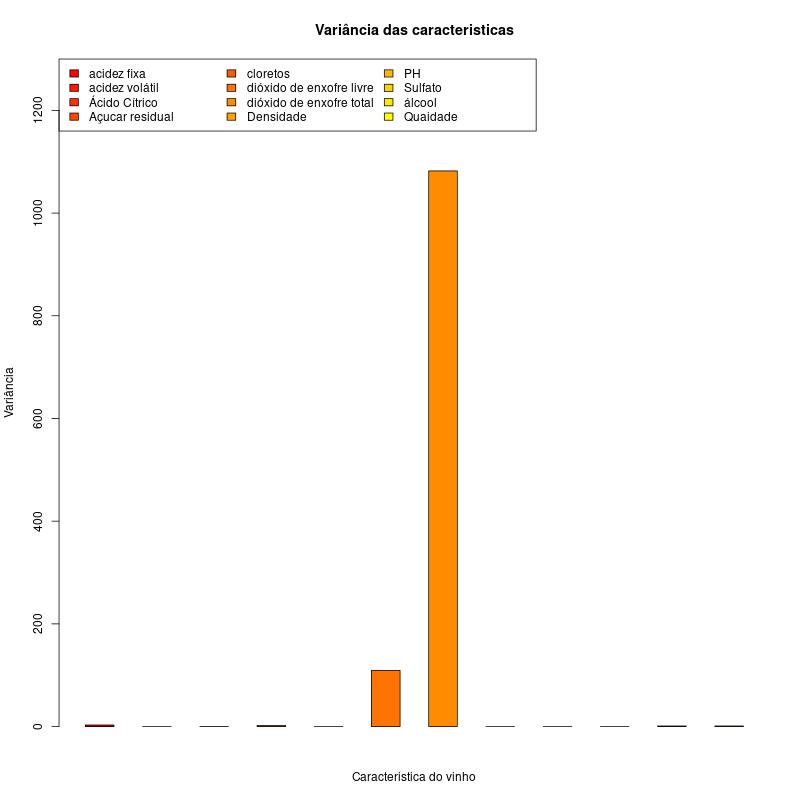
#### [Realizar analise dos dados apresentados na figura 2.2]

O próximo gráfico, apresentado na figura 2.3, é responsável por apresentar os dados do desvio padrão, onde estes dados apresentam o quão homogêneo são as característica dos vinhos tintos. Indica quanta observação de um conjunto de dados difere de sua média.

**Figura 2.3:** Desvio padrão das características do vinho tinto

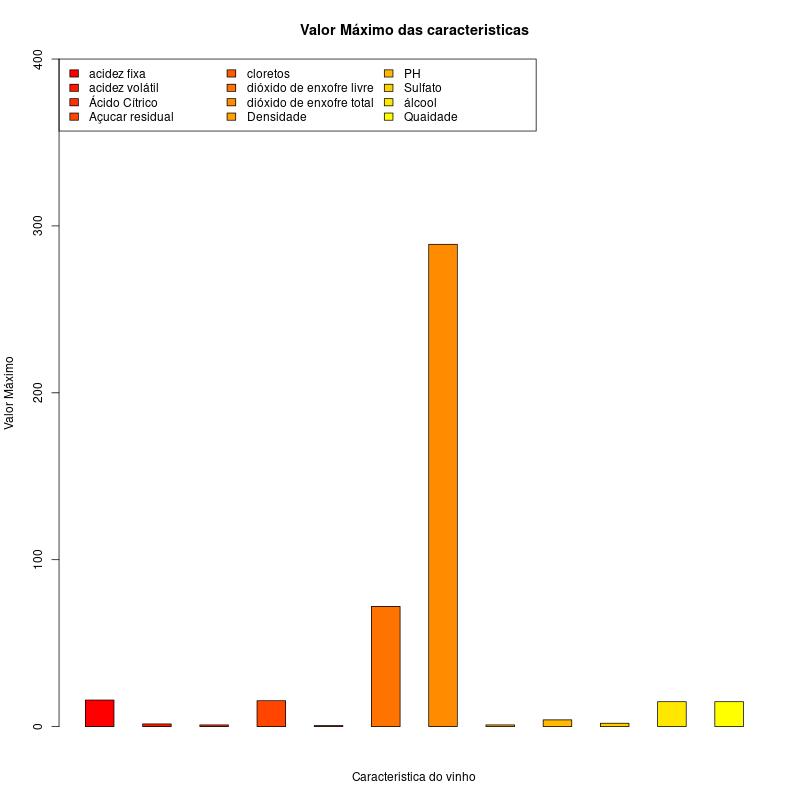
#### [Realizar analise dos dados apresentados na figura 2.3]

Agora são apresentado, no gráfico 2.4, os dados da variância dos dados das características dos vinhos tintos, onde a variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central. Indicando até que ponto os indivíduos de um grupo estão espalhados.

**Figura 2.4:** Variância das características do vinho tinto

#### [Realizar analise dos dados apresentados na figura 2.4]

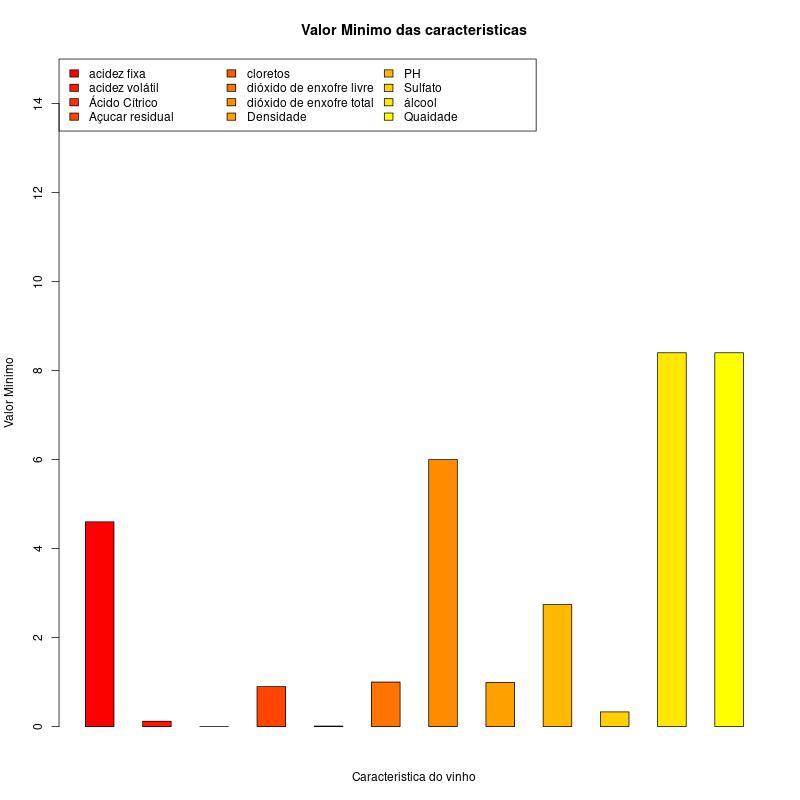
No próximo gráfico, da figura 2.5, é apresentado os resultados do maior valor de cada característica, com o objetivo de poder comparar com o menor valor de cada característica, sendo este valor apresentado na figura 2.6.



**Figura 2.5:** Maior das características do vinho tinto

#### [Realizar analise dos dados apresentados na figura 2.5]

O gráfico a seguir apresenta os menores valores das características, visando realizar a comparação com os maior valor.

**Figura 2.6:** Maior das características do vinho tinto

#### [Realizar analise dos dados apresentados na figura 2.6]

#### ***Dispersão e Correlação dos Dados***

Nesta seção são apresentados os resultados da dispersão de cada característica em relação à qualidade do vinho, com o objetivo de verificar quais características influenciam mais na qualidade dos dados.

A correlação pode ser visualizada no arquivo “DCWine/DadosGerados/CorrelacaoRedWine.csv”. A corrlação busca verificar a correlação entre todas as caracteristicas.

##### ***Correlação dos Dados***

A tabela 2.2 apresenta as correlações entre todas as características, sendo estes dados os mesmos do arquivo citado na seção 2.1.1.2.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Acidez fixa** | **Acidez volátil** | **Ácido Cítrico** | **Açúcar residual** | **Cloretos** | **D.E.L** | **D.E.T** | **Densidade** | **PH** | **Sulfato** | **Álcool** | **Qualidade** |
| Acidez fixa | 1,000 | -0,256 | 0,672 | 0,115 | 0,094 | -0,154 | -0,113 | 0,668 | -0,683 | 0,183 | -0,062 | 0,124 |
| Acidez volátil | -0,256 | 1,000 | -0,552 | 0,002 | 0,061 | -0,011 | 0,076 | 0,022 | 0,235 | -0,261 | -0,202 | -0,391 |
| Ácido Cítrico | 0,672 | -0,552 | 1,000 | 0,144 | 0,204 | -0,061 | 0,036 | 0,365 | -0,542 | 0,313 | 0,110 | 0,226 |
| Açúcar residual | 0,115 | 0,002 | 0,144 | 1,000 | 0,056 | 0,187 | 0,203 | 0,355 | -0,086 | 0,006 | 0,042 | 0,014 |
| Cloretos | 0,094 | 0,061 | 0,204 | 0,056 | 1,000 | 0,006 | 0,047 | 0,201 | -0,265 | 0,371 | -0,221 | -0,129 |
| D.E.L | -0,154 | -0,011 | -0,061 | 0,187 | 0,006 | 1,000 | 0,668 | -0,022 | 0,070 | 0,052 | -0,069 | -0,051 |
| D.E.T | -0,113 | 0,076 | 0,036 | 0,203 | 0,047 | 0,668 | 1,000 | 0,071 | -0,066 | 0,043 | -0,206 | -0,185 |
| Densidade | 0,668 | 0,022 | 0,365 | 0,355 | 0,201 | -0,022 | 0,071 | 1,000 | -0,342 | 0,149 | -0,496 | -0,175 |
| PH | -0,683 | 0,235 | -0,542 | -0,086 | -0,265 | 0,070 | -0,066 | -0,342 | 1,000 | -0,197 | 0,206 | -0,058 |
| Sulfato | 0,183 | -0,261 | 0,313 | 0,006 | 0,371 | 0,052 | 0,043 | 0,149 | -0,197 | 1,000 | 0,094 | 0,251 |
| Álcool | -0,062 | -0,202 | 0,110 | 0,042 | -0,221 | -0,069 | -0,206 | -0,496 | 0,206 | 0,094 | 1,000 | 0,476 |
| Qualidade | 0,124 | -0,391 | 0,226 | 0,014 | -0,129 | -0,051 | -0,185 | -0,175 | -0,058 | 0,251 | 0,476 | 1,000 |

**Tabela 2.2.**: Correlação entre todas as características

Para auxiliar a analise será definido um valor de corte para determinar as características que possuem correlação. Para considerar que existe uma correlação entre as características os valores devem ser maior que 0,45, ou maior que -0,45. Sendo que estes valores estão marcados de amarelo.

#### ***BloxPlot***

[Apresentar gráficos Bloxplot]

### ***Análise de Dados de Vinho Branco***

[Apresenta os resultados de vinho branco.]

#### ***Estatísticas***

[Apresentar dados estatísticos, como média, mediana, desvio padrão, etc.]

#### ***Dispersão e Correlação dos Dados***

[Apresentar os gráficos de dispersão e e a correlação dos dados]

#### ***BloxPlot***

[Apresentar gráficos Bloxplot]

## ***Data Mining***

[Apresentar resultados coletados através da aplicação de técnicas de data mining]

### ***Data Mining dos Dados de Vinho Tinto***

### ***Data Mining dos Dados de Vinho Branco***

|  |  |  |
| --- | --- | --- |
| **Aprovações** | | |
| **Participante** | **Assinatura** | **Data** |
| Patrocinador do Projeto |  |  |
| Gerente do Projeto |  |  |