

UNIVERSIDADE FEDERAL DO PIAUÍ
CIÊNCIA DA COMPUTAÇÃO

DELPHINO LUCIANI DE PAULA ARAUJO FILHO
GUILHERME OLIVEIRA DESIDÉRIO
SAMIA ISABEL DA ROCHA GONCALVES

TRABALHO FINAL

TERESINA
MARÇO DE 2023

Sumário

1.0 Introdução	3
1.1 Preparação do grupo	3
1.2 Material utilizado	3
1.3 Dicionário de dados	4
1.4 Tratamento de dados	6
2.0 Gráficos e análise de resultados	8
2.1 Gráfico de barras de variáveis	8
2.2 Correlação entre qualidade e outras variáveis	9
2.3 Detalhamento Variável PH	11
2.4 Detalhamento Variável Densidade	12
2.5 Mapa de calor	14
2.5.1 Correlação:	15
3.0 Conclusão	16
4.0 Fonte:	16

1.0 Introdução

A análise de gráficos de características de vinhos é uma técnica comum em degustações e avaliações sensoriais de vinhos. Esse relatório apresenta informações visuais sobre os atributos sensoriais de um vinho, como aroma, sabor, acidez e corpo, permitindo a identificação de características específicas e a comparação entre diferentes vinhos. A análise desses gráficos requer conhecimento e experiência em degustação de vinhos, bem como uma compreensão dos atributos sensoriais e suas interações. É uma ferramenta útil para produtores de vinho, sommeliers e entusiastas que buscam aprimorar suas habilidades de degustação e avaliação de vinhos.

O presente relatório visa detalhar o Projeto final de datascience do grupo, que corresponde a 3 nota da disciplina de Laboratório de Programação do Departamento de Computação - CCN da Universidade Federal do Piauí, período 2022.2.

Para a realização da nossa análise foi decidido em consenso com os integrantes do grupo a utilização do dataset *Red Wine Quality* presente no link: (<https://archive.ics.uci.edu/ml/datasets/Wine+Qualit>) que tem a intenção de analisar a preferências de consumidores de vinho e analisar qual o vinho escolhido com base no perfil do consumidor, que dentre os disponibilizados pela professora Cledjan foi o escolhido para ser utilizado no projeto.

O projeto do google colab encontra-se no link:

<https://colab.research.google.com/drive/1l--vc3WcMEz-7tG6Hg563HAW-FuoRV6n?authuser=1>

1.1 Preparação do grupo

Todos os integrantes do grupo se reuniram em uma chamada online onde todos contribuíram para a criação do trabalho, dividindo em parte de código e simulação/elaboração do relatório, criação de gráficos e estudo dos resultados.

1.2 Material utilizado

Para análise do dataset foi utilizado a linguagem python para utilizar de sua extensão matplotlib e seaborn para gerar os gráficos do projeto, Pandas para análise de dados, Numpy para gerar arrays de n dimensões e a plataforma google colab como

compiladora e IDE devido a facilidade e a possibilidade de desenvolver em conjunto e conferir os resultados em tempo real.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1.3 Dicionário de dados

Neste trabalho iremos analisar alguns dados importantes para o estudo das variáveis sendo elas:

Acidez volátil: Acidez volátil são os ácidos gasosos presentes no vinho.

Acidez fixa: Os ácidos fixos primários encontrados no vinho são tartárico, succínico, cítrico e málico.

Açúcar residual: Quantidade de açúcar que resta após a fermentação.

ácido cítrico: É um ácido orgânico fraco, encontrado naturalmente nas frutas cítricas.

Cloretos: Quantidade de sal presente no vinho.

Dióxido de enxofre livre: SO_2 é usado para prevenção de oxidação e deterioração microbiana do vinho.

Dióxido de enxofre total: O dióxido de enxofre (SO_2) é um composto muito importante utilizado na conservação de vinhos, devido às suas propriedades antioxidantes

pH: No vinho, o pH é usado para verificar a acidez.

Densidade: A densidade do vinho está relacionada principalmente ao seu teor alcoólico e de açúcares residuais.

Sulfatos: Os sulfitos adicionados preservam a frescura e protegem o vinho da oxidação e bactérias.

Álcool: Porcentagem de álcool presente no vinho.

Qualidade: A qualidade do vinho deve ser condizente à sensação experimentada pelo degustador ao cheirar a amostra.

Importar Dataset

```
[ ] input_file_wine = 'winequality-red.csv'
    df_wine = pd.read_csv(input_file_wine)
    df_wine
```

	Unnamed: 0	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
0	0	7.4	0.700	0.00	1.9	0.076	11.0
1	1	7.8	0.880	0.00	2.6	0.098	25.0
2	2	7.8	0.760	0.04	2.3	0.092	15.0
3	3	11.2	0.280	0.56	1.9	0.075	17.0
4	4	7.4	0.700	0.00	1.9	0.076	11.0
...
1594	1594	6.2	0.600	0.08	2.0	0.090	32.0
1595	1595	5.9	0.550	0.10	2.2	0.062	39.0
1596	1596	6.3	0.510	0.13	2.3	0.076	29.0
1597	1597	5.9	0.645	0.12	2.0	0.075	32.0
1598	1598	6.0	0.310	0.47	3.6	0.067	18.0

1599 rows x 13 columns

total sulfur dioxide	density	pH	sulphates	alcohol	quality
-0.117850	-0.368372	0.136005	-0.125307	0.245123	0.066453
-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000

```
# Tipos de variáveis
df_wine.dtypes
```

```
Unnamed: 0          int64
fixed acidity      float64
volatile acidity   float64
citric acid        float64
residual sugar     float64
chlorides          float64
free sulfur dioxide float64
total sulfur dioxide float64
density           float64
pH               float64
sulphates         float64
alcohol           float64
quality           int64
dtype: object
```

```
# Tipos de Dados
df_wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0             1599 non-null   int64
1   fixed acidity           1599 non-null   float64
2   volatile acidity        1599 non-null   float64
3   citric acid             1599 non-null   float64
4   residual sugar          1599 non-null   float64
5   chlorides              1599 non-null   float64
6   free sulfur dioxide     1599 non-null   float64
7   total sulfur dioxide    1599 non-null   float64
8   density                1599 non-null   float64
9   pH                    1599 non-null   float64
10  sulphates              1599 non-null   float64
11  alcohol                1599 non-null   float64
12  quality                1599 non-null   int64
dtypes: float64(11), int64(2)
memory usage: 162.5 KB
```

1.4 Tratamento de dados

Para ter a certeza de que não existem dados faltantes, duplicados ou vazios é necessário fazer um tratamento anterior a visualização e análise dos gráficos sendo necessário então executar as seguintes linhas de código e seus resultados:

```
# Registros duplicados
duplicados = df_wine.duplicated()
duplicados
```

```
0      False
1      False
2      False
3      False
4      False
...
1594    False
1595    False
1596    False
1597    False
1598    False
Length: 1599, dtype: bool
```

```
duplicados.sum()
```

```
0
```

```
# Checa valores vazios no dataset
df_wine.isnull().sum()
```

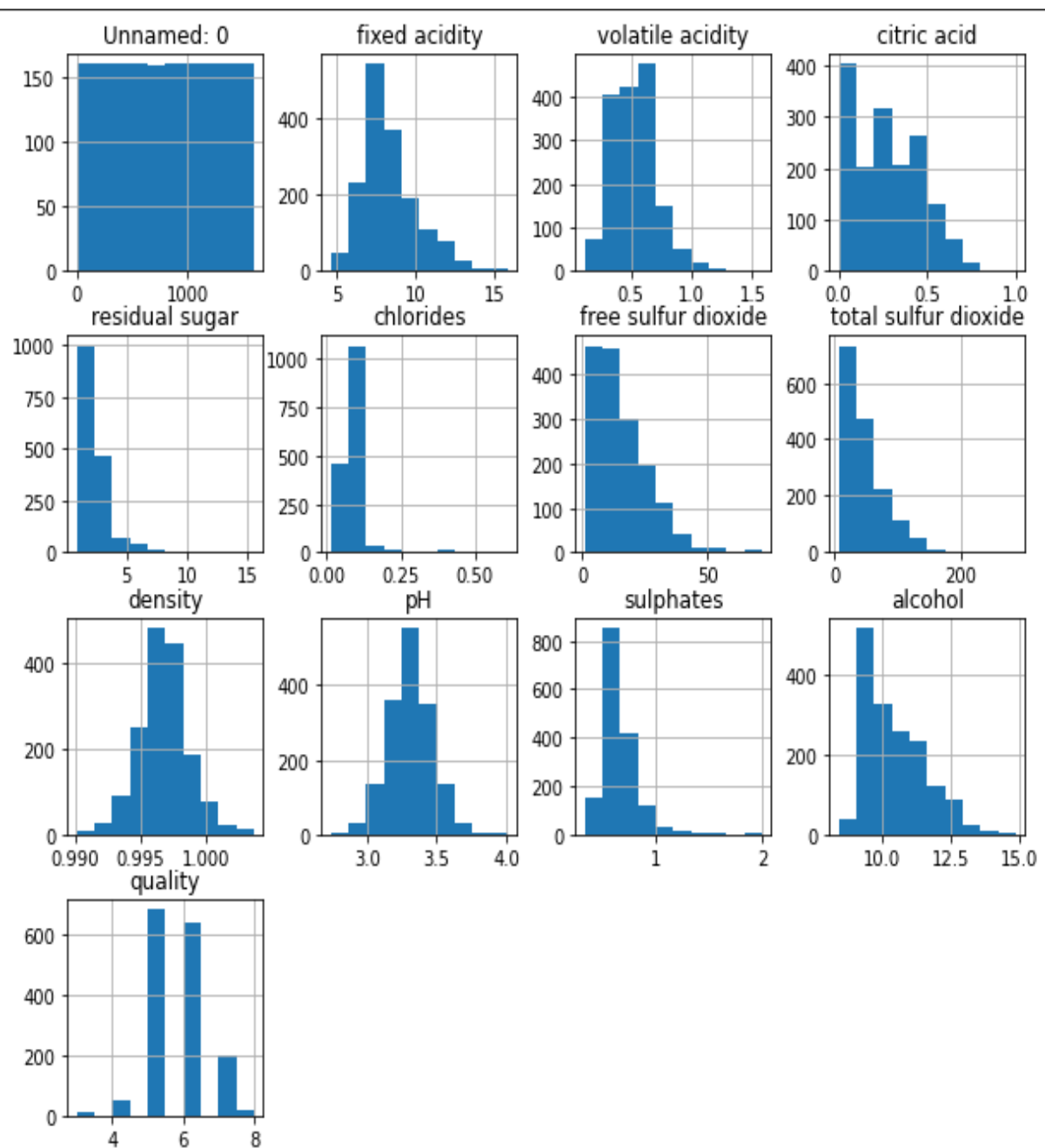
```
Unnamed: 0          0
fixed acidity        0
volatile acidity     0
citric acid          0
residual sugar       0
chlorides            0
free sulfur dioxide  0
total sulfur dioxide 0
density              0
pH                  0
sulphates            0
alcohol              0
quality              0
dtype: int64
```

```
# Percentual de dados faltantes  
round(df_wine.isnull().mean() *100,2)
```

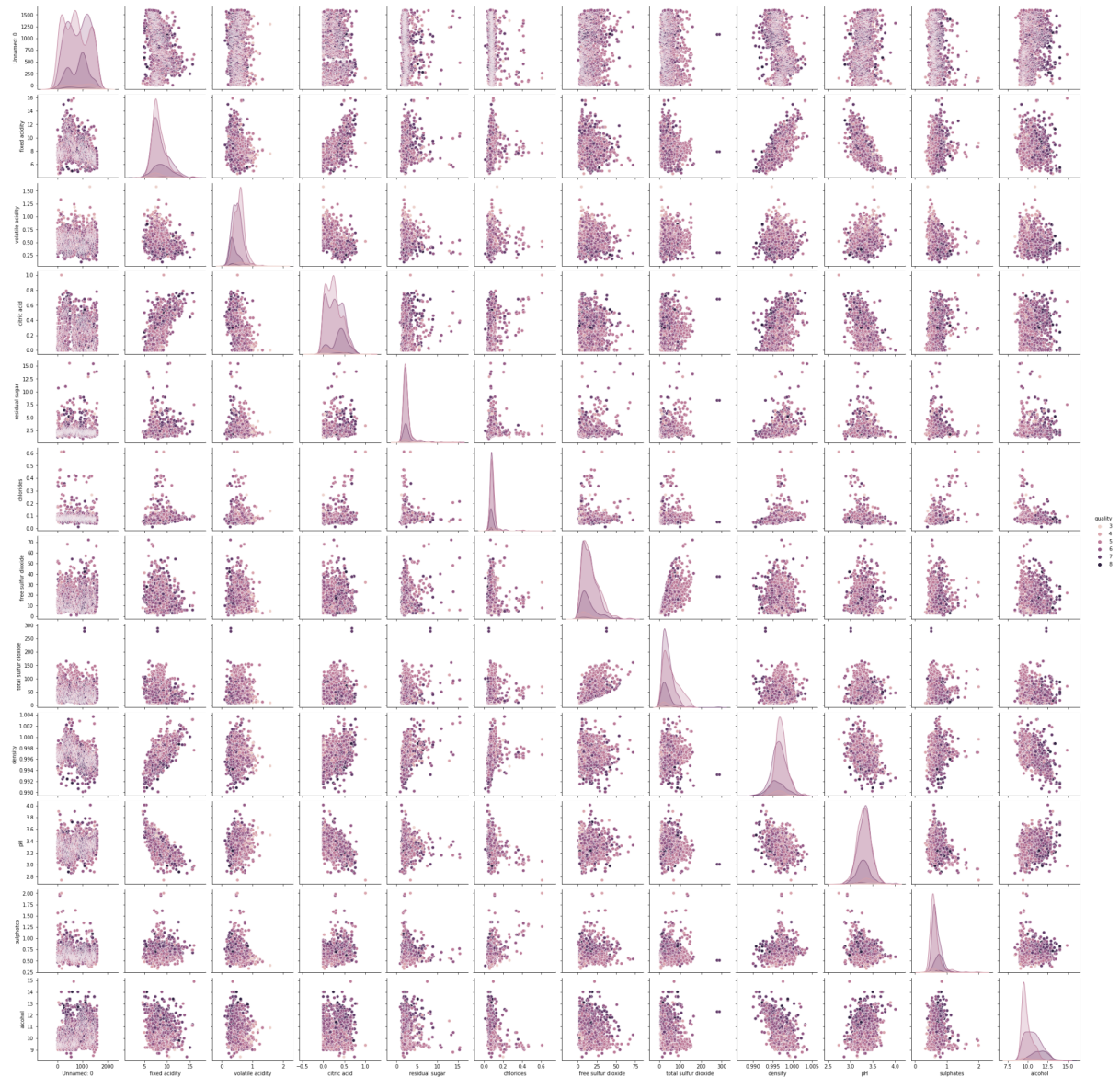
```
Unnamed: 0      0.0  
fixed acidity    0.0  
volatile acidity 0.0  
citric acid      0.0  
residual sugar   0.0  
chlorides        0.0  
free sulfur dioxide 0.0  
total sulfur dioxide 0.0  
density          0.0  
pH              0.0  
sulphates        0.0  
alcohol          0.0  
quality          0.0  
dtype: float64
```

2.0 Gráficos e análise de resultados

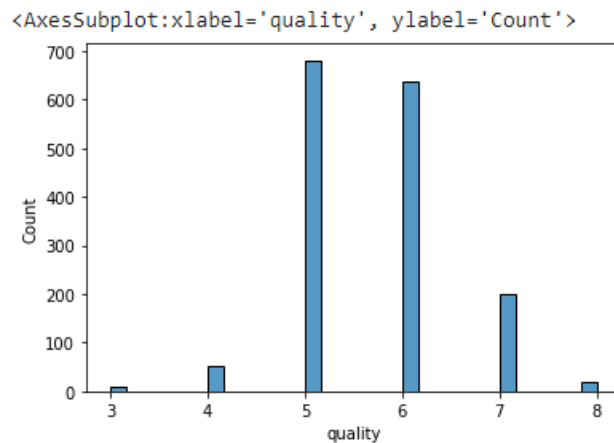
2.1 Gráfico de barras de variáveis



2.2 Correlação entre qualidade e outras variáveis



```
# Distribuição de qualidade
sns.histplot(df_wine, x='quality')
```



```
# Tabela de Distribuição
tabela = df_wine['quality'].value_counts().to_frame().reset_index()
tabela.columns = ['quality', 'count']
tabela.style.format({'count': '{:,}'})
```

	quality	count
0	5	681
1	6	638
2	7	199
3	4	53
4	8	18
5	3	10

```
# Métricas Estatísticas
df_wine['quality'].describe()
```

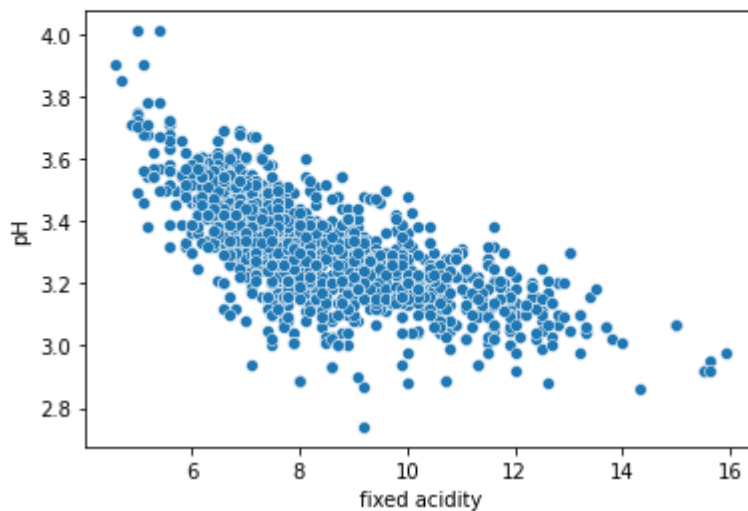
```
count    1599.000000
mean       5.636023
std        0.807569
min         3.000000
25%         5.000000
50%         6.000000
75%         6.000000
max         8.000000
Name: quality, dtype: float64
```

2.3 Detalhamento Variável PH

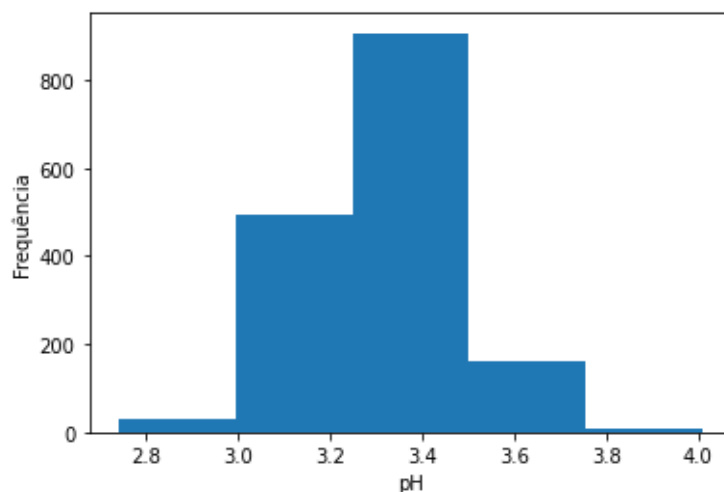
A análise do pH do vinho é importante para garantir que o vinho esteja em condições ideais para a fermentação, estabilidade, sabor e conservação. Isso permite que os produtores ajustem a acidez do vinho, se necessário, para obter um produto final de alta qualidade.

```
# Relação entre Acidez Fixa e pH  
# Gráfico de dispersão  
sns.scatterplot(data=df_wine, x='fixed acidity', y='pH')
```

```
<AxesSubplot:xlabel='fixed acidity', ylabel='pH'>
```

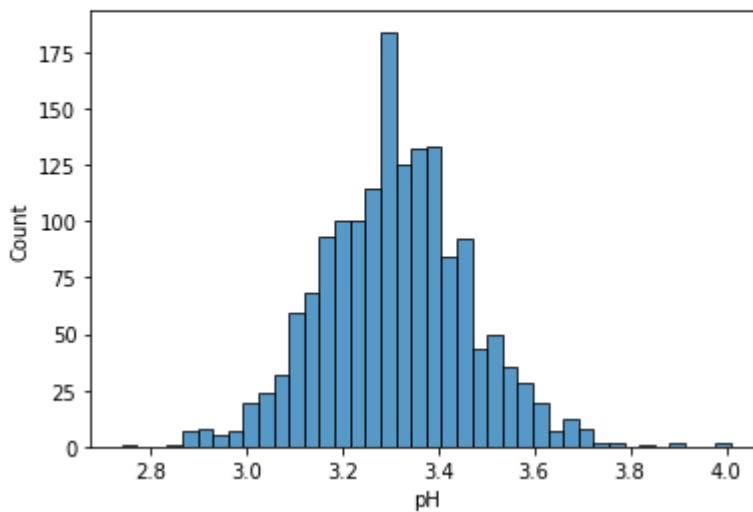


```
# Histograma de frequência do pH  
plt.hist(df_wine['pH'], bins=5)  
plt.xlabel('pH')  
plt.ylabel('Frequência')  
plt.show()
```



```
sns.histplot(df_wine, x='pH')
```

```
<AxesSubplot:xlabel='pH', ylabel='Count'>
```



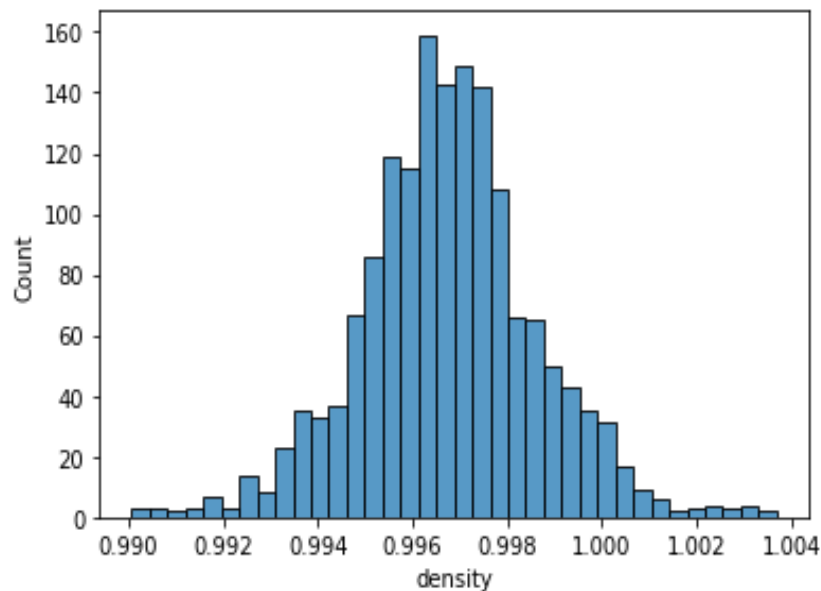
2.4 Detalhamento Variável Densidade

A análise da densidade do vinho é importante para garantir que o vinho esteja fermentando corretamente e para calcular o teor alcoólico.

Além disso, é importante analisar a relação entre acidez volátil e densidade ocorre porque o ácido acético tem uma densidade menor do que a da água. Isso significa que quanto maior a quantidade de ácido acético presente no vinho, menor será a densidade do vinho. Assim, a análise da densidade pode ser usada para avaliar a quantidade de acidez volátil presente no vinho. Se a densidade estiver anormalmente baixa, isso pode indicar a presença de uma quantidade excessiva de acidez volátil. Portanto, a análise da densidade é uma ferramenta importante para avaliar a qualidade do vinho e garantir que ele esteja dentro dos padrões desejados.

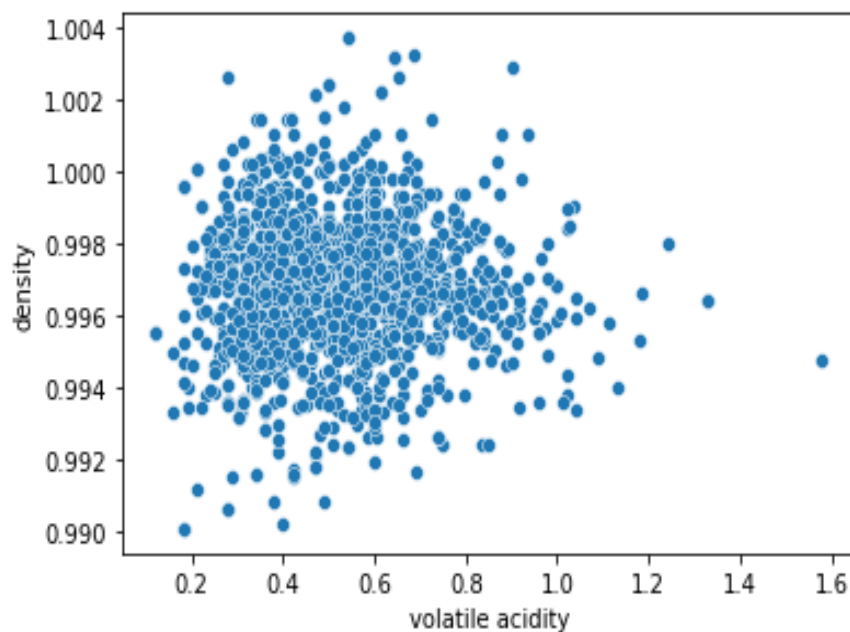
```
# Distribuição de densidade  
sns.histplot(df_wine, x='density')
```

<AxesSubplot:xlabel='density', ylabel='Count'>



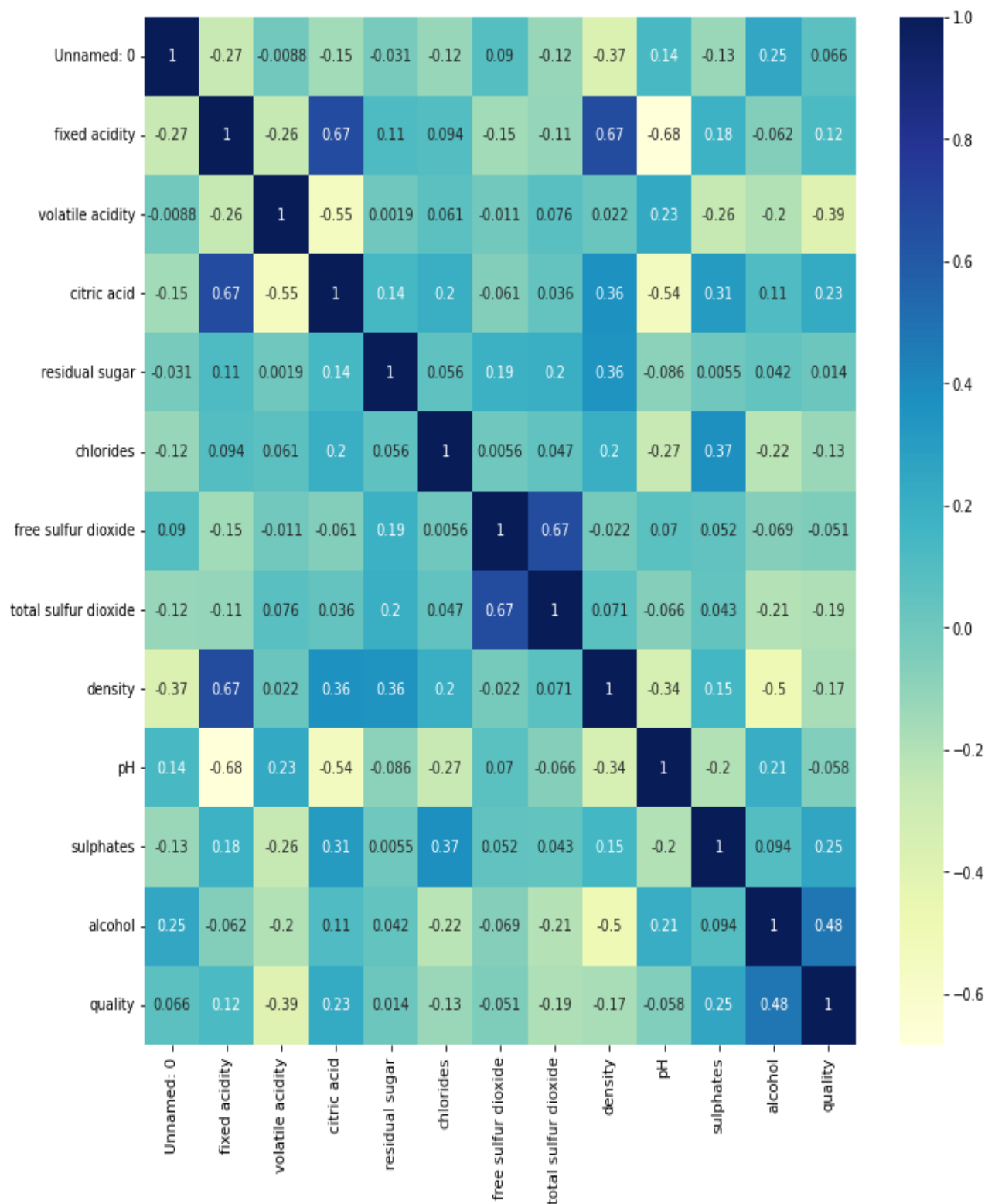
```
# Relação entre Acidz Volátil e Densidade  
sns.scatterplot(data=df_wine, x='volatile acidity', y='density')
```

<AxesSubplot:xlabel='volatile acidity', ylabel='density'>



2.5 Mapa de calor

Um mapa de calor é uma ferramenta visual que ajuda os analistas de dados a identificar rapidamente padrões, tendências e correlações nos dados. Ele é usado para tornar os dados mais fáceis de entender e analisar, e é especialmente útil em datasets grandes e complexas.



2.5.1 Correlação:

	Unnamed: 0	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
Unnamed: 0	1.000000	-0.268484	-0.008815	-0.153551	-0.031261	-0.119869
fixed acidity	-0.268484	1.000000	-0.256131	0.671703	0.114777	0.093705
volatile acidity	-0.008815	-0.256131	1.000000	-0.552496	0.001918	0.061298
citric acid	-0.153551	0.671703	-0.552496	1.000000	0.143577	0.203823
residual sugar	-0.031261	0.114777	0.001918	0.143577	1.000000	0.055610
chlorides	-0.119869	0.093705	0.061298	0.203823	0.055610	1.000000
free sulfur dioxide	0.090480	-0.153794	-0.010504	-0.060978	0.187049	0.005562
total sulfur dioxide	-0.117850	-0.113181	0.076470	0.035533	0.203028	0.047400
density	-0.368372	0.668047	0.022026	0.364947	0.355283	0.200632
pH	0.136005	-0.682978	0.234937	-0.541904	-0.085652	-0.265026
sulphates	-0.125307	0.183006	-0.260987	0.312770	0.005527	0.371260
alcohol	0.245123	-0.061668	-0.202288	0.109903	0.042075	-0.221141
quality	0.066453	0.124052	-0.390558	0.226373	0.013732	-0.128907

free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0.090480	-0.117850	-0.368372	0.136005	-0.125307	0.245123	0.066453
-0.153794	-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
-0.010504	0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
-0.060978	0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
0.187049	0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
0.005562	0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
1.000000	0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
0.667666	1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
-0.021946	0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
0.070377	-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
0.051658	0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
-0.069408	-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
-0.050656	-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000

3.0 Conclusão

A partir da análise do dataset de vinhos, podemos concluir que existem diversas variáveis que afetam a qualidade do vinho, tais como a acidez, o teor alcoólico, o pH e o tipo de uva. Foi possível observar que vinhos de alta qualidade tendem a ter um teor alcoólico maior e uma acidez mais baixa. É importante destacar que a análise dos dados é apenas um passo inicial na compreensão da produção de vinhos. É necessário que sejam realizados testes sensoriais e químicos adicionais para confirmar e aprofundar os resultados obtidos na análise de dados. Os resultados obtidos através da análise de dados podem ser valiosos para a produção de vinhos de alta qualidade, uma vez que os produtores podem utilizar as informações para ajustar os processos de produção e selecionar as melhores uvas e técnicas de fermentação. Por fim, a análise de dados pode ser uma ferramenta poderosa para aprimorar a produção e a qualidade do vinho.

4.0 Fonte:

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>