Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-21М
Ся Бэйбэй

Москва — 2022 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Требуется выполнить следующие действия [1]:

• Выбрать набор данных (датасет).

• Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

• Сформировать отчет и разместить его в своем репозитории на GitHub

3. Ход выполнения работы

3.1. Текстовое описание набора данных

В качестве набора данных использовались метрологические данные физико-химических исследований качества вина. [2]. Данный набор данных доступен по следующему адресу:Wine Quality Dataset | Kaggle. Этот набор данных состоит из одного файла WineQT.csv, содержащего все данные датасета. Данный файл содержит следующие колонки:

1 - fixed acidity（фиксированная кислотность）
2 - volatile acidity （летучая кислотность）
3 - citric acid（лимонная кислота）
4 - residual sugar（остаточный сахар）
5 – chlorides（хлориды）
6 - free sulfur dioxide（свободный диоксид серы）
7 - total sulfur dioxide（общий диоксид серы）
8 – density（плотность）
9 - pH
10 – sulphates（сульфаты）
11 – alcohol（алкоголь）
Output variable (based on sensory data):
12 - quality (score between 0 and 10)（качество）

3.2. Основные характеристики набора данных

Подключим все необходимые библиотеки & Загрузим непосредственно данные:

```
In [4]:
# import the libraries for analysis
import numpy as np
# excel
import pandas as pd
#Drawing
import seaborn as sns
#Drawing 2D
import matplotlib.pyplot as plt
#using style ggplot
plt.style.use("ggplot")
#内嵌画图
%matplotlib inline
#开源的可视化框架  Visualization framework
import plotly.graph_objects as go
import plotly.express   as px

#importing the dataset

df=pd.read_csv("./WineQT.csv")

#Looking the data set
df.head()
```

Показать часть данных о физико - химических параметрах вина:

```
In [7]: GlobalTemp.head(5)
```

Out[7]:

	dt	LandAverageTemperature	LandAverageTemperatureUncertainty	LandMaxTemperature	LandMaxTemperatureUncertainty	LandMinTemperature	LandMinTe
0	1750-01-01	3.034	3.574	NaN	NaN	NaN	
1	1750-02-01	3.083	3.702	NaN	NaN	NaN	
2	1750-03-01	5.626	3.076	NaN	NaN	NaN	
3	1750-04-01	8.490	2.451	NaN	NaN	NaN	
4	1750-05-01	11.573	2.072	NaN	NaN	NaN	

Показывать размер данных:

```
In [5]:
#print the shape dataset
print("Shape the dataset",df.shape)
```

```
Shape the dataset (1143, 13)
```

проверка типов данных и пустоты данных：

```
In [6]: #Checking the dtypes of all the columns
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64  
 12  Id                    1143 non-null   int64  
dtypes: float64(11), int64(2)
memory usage: 116.2 KB
```

```
In [7]: #checking null value
        df.isna().sum()
```

```
Out[7]: 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
        Id                      0
        dtype: int64
```

Показывать многомерные данные：

```
In [8]: # describe value data set  count,averge,standard deviation
        df.describe().round(2)
```

Out[8]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
count	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00
mean	8.31	0.53	0.27	2.53	0.09	15.62	45.91	1.00	3.31	0.66	10.44	5.66	804.97
std	1.75	0.18	0.20	1.36	0.05	10.25	32.78	0.00	0.16	0.17	1.08	0.81	464.00
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00	0.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	21.00	1.00	3.20	0.55	9.50	5.00	411.00
50%	7.90	0.52	0.25	2.20	0.08	13.00	37.00	1.00	3.31	0.62	10.20	6.00	794.00
75%	9.10	0.64	0.42	2.60	0.09	21.00	61.00	1.00	3.40	0.73	11.10	6.00	1209.50
max	15.90	1.58	1.00	15.50	0.61	68.00	289.00	1.00	4.01	2.00	14.90	8.00	1597.00

```
In [9]: #Drop column ID, cause we don't need
        df.drop(columns="Id",inplace=True)
```

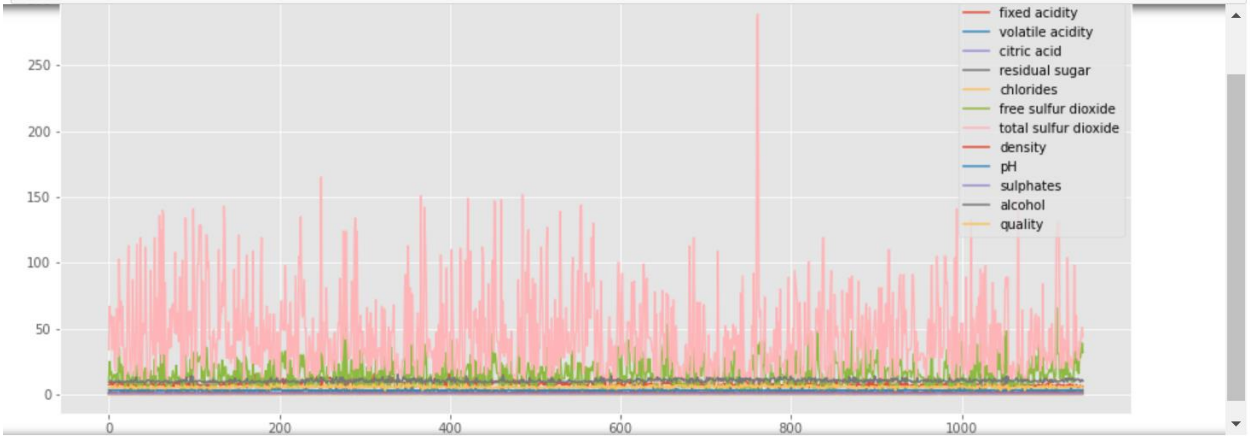
```
In [10]: # the unique quality-distinct
         print("The value Quality",df["quality"].unique())
```

```
The value Quality [5 6 7 4 8 3]
```

3.3. Визуальное исследование датасета

Показать распределение данных по всем параметрам：

```
#graph all the data set-just for looking
#每一列绘制一条曲线  figsize 画布尺寸
df.plot(figsize=(15,6))
```



Очевидно, что, за исключением свободной диоксид серы и общей диоксид серы, другие параметры не претерпели значительных изменений.

Показывать среднее значение параметров с одинаковым качеством красного вина：

```
In [12]:
# making Group by
ave_qu =df.groupby("quality").mean()
ave_qu
```

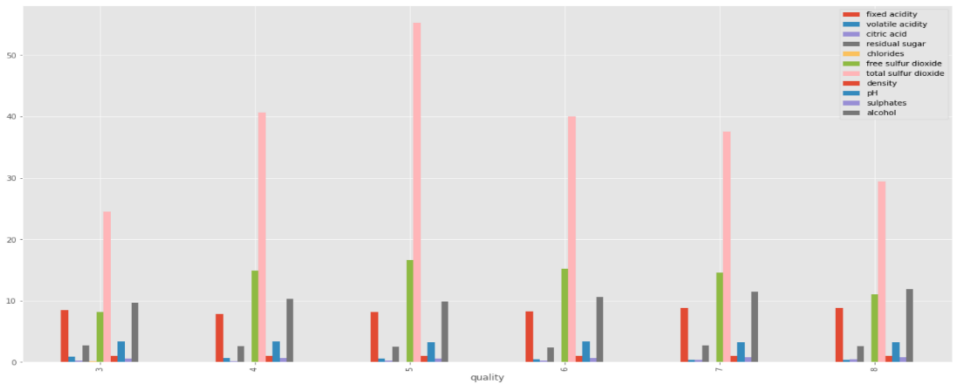
Out[12]:

quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
3	8.450000	0.897500	0.211667	2.666667	0.105333	8.166667	24.500000	0.997682	3.361667	0.550000	9.691667
4	7.809091	0.700000	0.165758	2.566667	0.094788	14.848485	40.606061	0.996669	3.391212	0.637879	10.260606
5	8.161077	0.585280	0.240124	2.540476	0.091770	16.612836	55.299172	0.997073	3.302091	0.613375	9.902277
6	8.317749	0.504957	0.263680	2.444805	0.085281	15.215368	39.941558	0.996610	3.323788	0.676537	10.655339
7	8.851049	0.393671	0.386573	2.760140	0.075217	14.538462	37.489510	0.996071	3.287133	0.743566	11.482634
8	8.806250	0.410000	0.432500	2.643750	0.070187	11.062500	29.375000	0.995553	3.240625	0.766250	11.937500

Очевидно, качество вина и степень летучести алкоголя отрицательны.
Качество спирта и концентрация хлорида в вине были отрицательными.
Разница в плотности между разными видами вина невелика.
Качество вина прямо связано с концентрацией алкоголя.
Среднее значение параметра с одинаковым качеством красного вина, показанное на колонке:

```
In [13]:
# graph the group by
ave_qu.plot(kind="bar",figsize=(20,10))
```

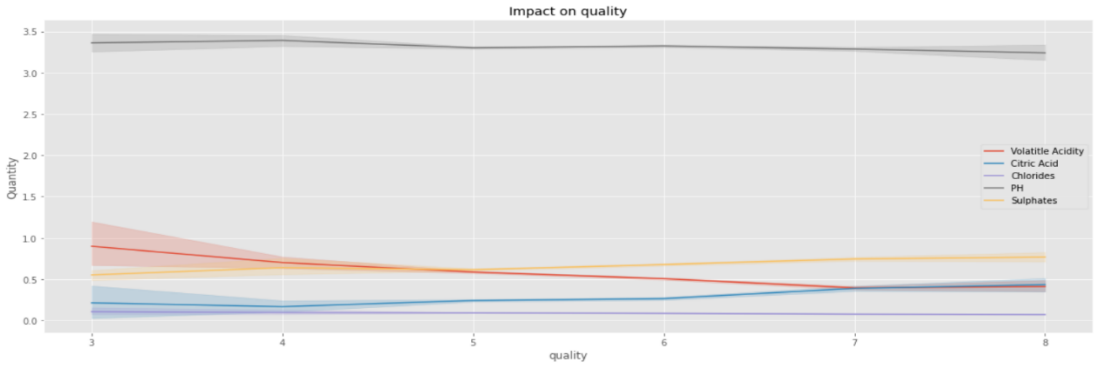
Out[13]: <AxesSubplot:xlabel='quality'>



Очевидно, что когда качество вина составляет 5, наибольшие значения имеют диоксид серы и общий диоксид серы.

Изучение влияния летучей кислотности, лимонной кислоты, хлорида, PH, сульфата на качество вина:

```
#Now we can see the effect of the elements on the quality
#Let see effect some of elements on the quality -detail
#绘图 drawing
plt.figure(figsize=(20,7))
sns.lineplot(data=df,x="quality",y="volatile acidity",label="Volatitle Acidity")
sns.lineplot(data=df,x="quality",y="citric acid",label="Citric Acid")
sns.lineplot(data=df,x="quality",y="chlorides",label="Chlorides")
sns.lineplot(data=df,x="quality",y="pH",label="PH")
sns.lineplot(data=df,x="quality",y="sulphates",label="Sulphates")
#y轴的名字
plt.ylabel("Quantity")
plt.title("Impact on quality")
plt.legend()
plt.show()
```

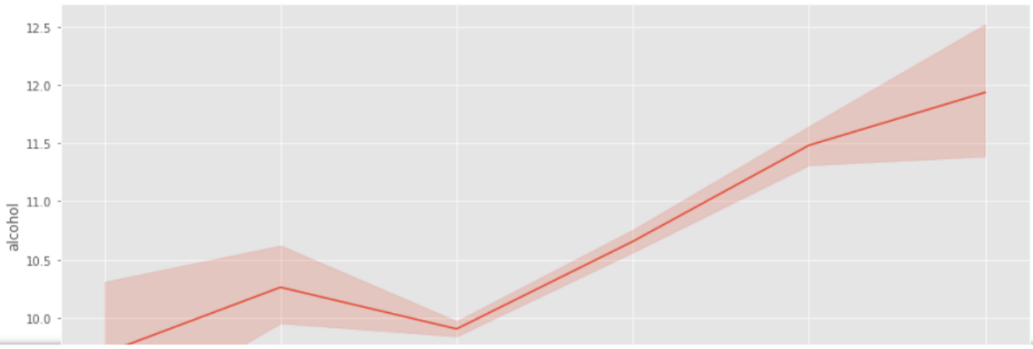


Видно, что значение вышеприведенных параметров не сильно влияет на качество вина.

Показать на ломаной диаграмме соотношение между качеством вина и количеством алкоголя:

```
In [18]: # effect the aicohol on the quality
         plt.figure(figsize=(15,7))
         sns.lineplot(data=df,x="quality",y='alcohol')
```

```
Out[18]: <AxesSubplot:xlabel='quality', ylabel='alcohol'>
```



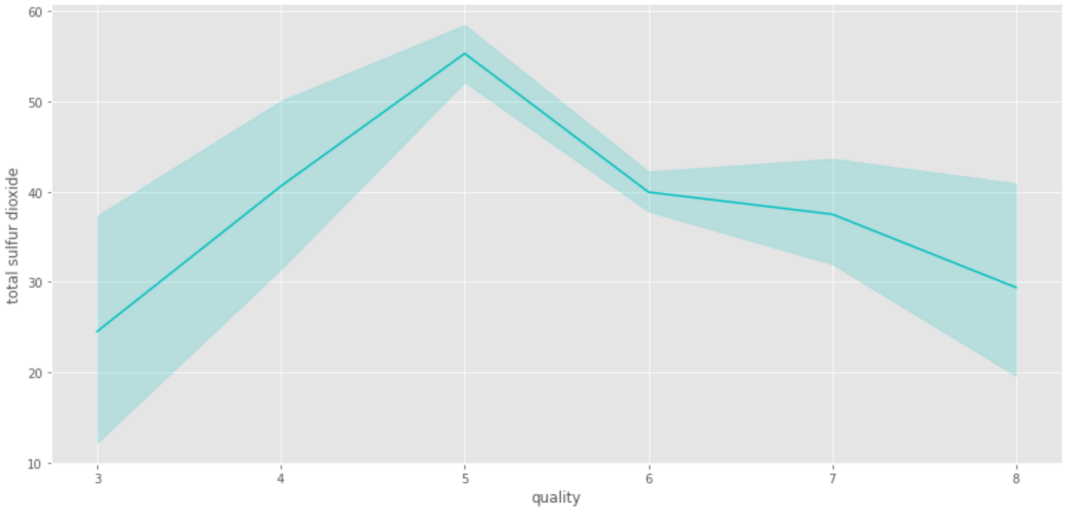
Результаты как раз подтверждают вышеприведенные выводы:

Качество вина прямо связано с концентрацией алкоголя.

Изучение с помощью ломаной диаграммы соотношения между качеством вина и концентрацией свободной диоксида серы и общей концентрацией диоксида серы:

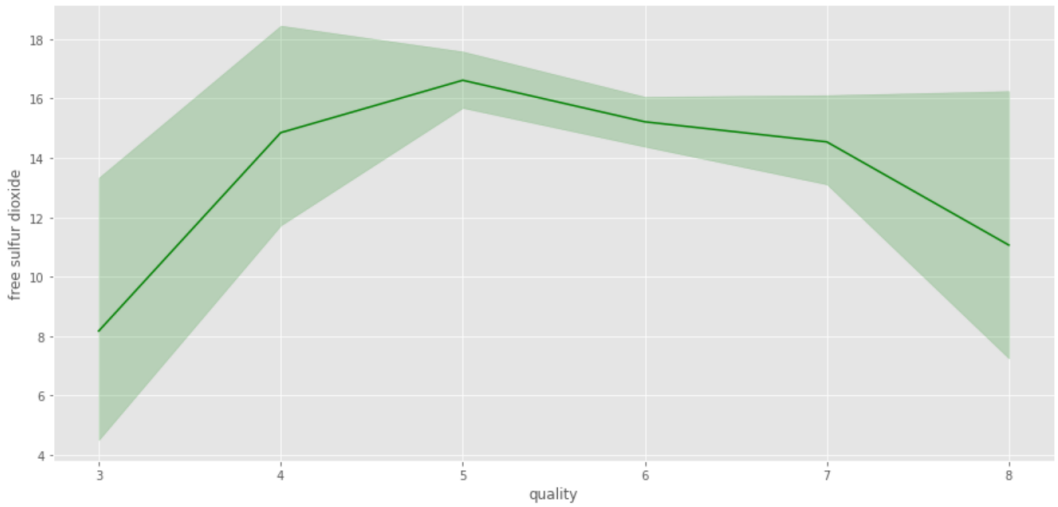
```
In [21]: # effect the total sulfur dioxide on the quality
         plt.figure(figsize=(15,7))
         sns.lineplot(data=df,x="quality",y="total sulfur dioxide",color="c")
```

```
Out[21]: <AxesSubplot:xlabel='quality', ylabel='total sulfur dioxide'>
```




```
In [27]: #effect the free sulfur dioxide on the quality
         plt.figure(figsize=(15,7))
         sns.lineplot(data=df, x="quality",y="free sulfur dioxide",color="g")
```

```
Out[27]: <AxesSubplot:xlabel='quality', ylabel='free sulfur dioxide'>
```

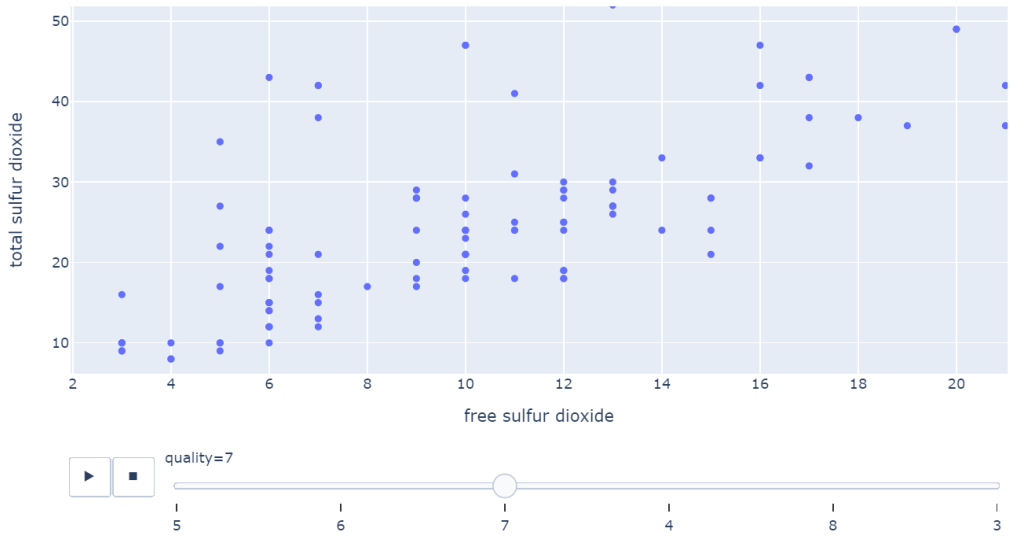


Результаты как раз подтверждают вышеприведенные выводы:

Очевидно, что между качеством вина и приведенными выше параметрами существует нелинейная связь.

Изучение с помощью растрескивания соотношения между качеством вина и концентрацией свободной диоксида серы и общей концентрацией диоксида серы:

```
In [28]:
# using gragh interactive the show the effect free and the total-sulfer dioxide in the quality
px.scatter(df,x="free sulfur dioxide",y="total sulfur dioxide",animation_frame="quality")
```



Очевидно, что существует положительная корреляция между двумя параметрами.

Построим корреляционную матрицу по всему набору данных:

```
In [30]:
df.corr()
```

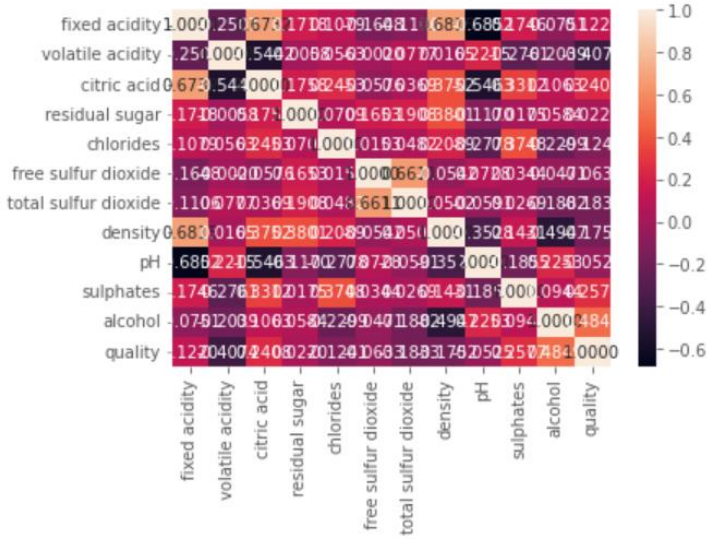
Out[30]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.250728	0.673157	0.171831	0.107889	-0.164831	-0.110628	0.681501	-0.685163	0.174592	-0.075055	0.121970
volatile acidity	-0.250728	1.000000	-0.544187	-0.005751	0.056336	-0.001962	0.077748	0.016512	0.221492	-0.276079	-0.203909	-0.407394
citric acid	0.673157	-0.544187	1.000000	0.175815	0.245312	-0.057589	0.036871	0.375243	-0.546339	0.331232	0.106250	0.240821
residual sugar	0.171831	-0.005751	0.175815	1.000000	0.070863	0.165339	0.190790	0.380147	-0.116959	0.017475	0.058421	0.022002
chlorides	0.107889	0.056336	0.245312	0.070863	1.000000	0.015280	0.048163	0.208901	-0.277759	0.374784	-0.229917	-0.124085
free sulfur dioxide	-0.164831	-0.001962	-0.057589	0.165339	0.015280	1.000000	0.661093	-0.054150	0.072804	0.034445	-0.047095	-0.063260
total sulfur dioxide	-0.110628	0.077748	0.036871	0.190790	0.048163	0.661093	1.000000	0.050175	-0.059126	0.026894	-0.188165	-0.183339
density	0.681501	0.016512	0.375243	0.380147	0.208901	-0.054150	0.050175	1.000000	-0.352775	0.143139	-0.494727	-0.175208
pH	-0.685163	0.221492	-0.546339	-0.116959	-0.277759	0.072804	-0.059126	-0.352775	1.000000	-0.185499	0.225322	-0.052453
sulphates	0.174592	-0.276079	0.331232	0.017475	0.374784	0.034445	0.026894	0.143139	-0.185499	1.000000	0.094421	0.257710
alcohol	-0.075055	-0.203909	0.106250	0.058421	-0.229917	-0.047095	-0.188165	-0.494727	0.225322	0.094421	1.000000	0.484866
quality	0.121970	-0.407394	0.240821	0.022002	-0.124085	-0.063260	-0.183339	-0.175208	-0.052453	0.257710	0.484866	1.000000

Визуализируем корреляционную матрицу с помощью тепловой карты:

```
In [33]: sns.heatmap(df.corr(),annot=True,fmt=".4f")
```

```
Out[33]: <AxesSubplot:>
```



Кроме того, можно видеть, что высокая зависимость между фиксированной кислотой и величиной PH соответствует теоретическим выводам. также существует тесная связь между постоянной кислотностью и плотностью.

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets