Capstone Project 1: Milestone Report
Wine Quality
Audrey Rager

Problem statement: Why it's a useful question to answer and for whom (get this from your proposal)

Baby Boomers account for the majority of wine consumption in the U.S. (McMilan, R., 2020). However, due to mortality, their numbers are declining. Gen Xers do not make up enough of the population to offset this decline. Millennials offer the best opportunity for growth in the wine industry (McMilan, R., 2020).

Using results of various physiochemical tests, it may be possible to predict wine quality before wine reaches maturity. Vineyards may be able to use the results of data analysis to predict wine quality to price the vintage and market to appropriate distributors (Cortez et al, 2009; McMilan, R., 2020).

In addition to the principle of supply and demand, the price of wine depends opinion of the wine's quality by experts and quality assessments by physicochemical tests. Expert opinions can vary widely.

Correlating expert human quality assessment to the chemical properties of wine could be of benefit to the wine industry. Doing so would make the certification and quality assessment and assurance process more controlled and predictable. This would make the process of pricing and marketing wine more economical (McMilan, R., 2020).

Description of the dataset, how you obtained, cleaned, and wrangled it (get this from your data wrangling report)

For my Capstone Project 1, I will be analyzing a Wine Quality data set from UC Irvine Machine Learning Repository. The data are related to red and white wine variants of Portuguese "Vinho Verde" wine (Cortez et al, 2009). The data set includes 11 physicochemical variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) and one output variable of quality (score between 0 and 1).

1. What kind of cleaning steps did you perform?

   I didn't have to perform any data cleaning for this data set. I simply imported the two csv files, one for red wine and one for white wine, into Pandas dataframes using the pd.csv_read() function with a sep=';' option because the data values were separated with a semicolon instead of a comma. I combined the two data

frames into one using the .append() method. Before combining them, I added a 'wine type' column to each and populated the column with the values 'red' or 'white' using the following code.

reddf['wine type'] = 'red'

whitedf['wine type'] = 'white'

winedf = reddf.append(whitedf, ignore_index=True)

2. How did you deal with missing values, if any?

I searched for missing values with the following code:

```
# Compute the percentage of missing values in the data

def missing_percentage(data1, data2, col_name = "Missing value (%)"):

    # Calculating the missing percentage

    missing_red = pd.DataFrame(data1.isnull().sum() /len(data1)*100, columns =
[col_name])

    missing_white = pd.DataFrame(data2.isnull().sum() /len(data2)*100, columns =
[col_name])

    # Forming the output dataframe

    missing_df = pd.DataFrame({'Red Wine': missing_red.iloc[:, 0], 'White wine':
missing_white.iloc[:, 0]})

    return missing_df

missing_percentage(reddf, whitedf)
```

There were no missing values in either the red wine or white wine data sets.

3. Were there outliers, and how did you handle them?

There were outliers in several of the attributes. I read the article "Modeling wine preferences by data mining from physicochemical properties" by the authors who provided the data (Cortez, et al., 2009). The authors did not indicate that these values were erroneous, therefore, I will leave them in for future data analysis.

Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)

1. Summary of findings
    - Quality is normally distributed with a slight left skew. Most values in Quality are concentrated in 5, 6 and 7. Values range from 3 to 9, with no values in 1, 2, or 10.
    - All variables have outliers, mostly on the large side (left-skewed).
    - The distributions for fixed acidity, volatile acidity and citric acid would be symmetrical if outliers were removed.
    - In contract, removing outliers from residual sugar would have little or no effect on its skewness; it will remain positively skewed.
    - Density and alcohol have only a few outliers, making it different from the other variables
    - Alcohol does not have a normal distribution. It has an irregular, step-down pattern toward higher alcohol levels. It does not have any pronounced outliers.
    - Total Sulfur Dioxide and Density appear to have bimodal distributions.
    - Range is much larger compared to the IQR. Mean is usually greater than the median. These observations indicate that there are outliers in the data set and before any analysis is performed outliers must be taken care of.

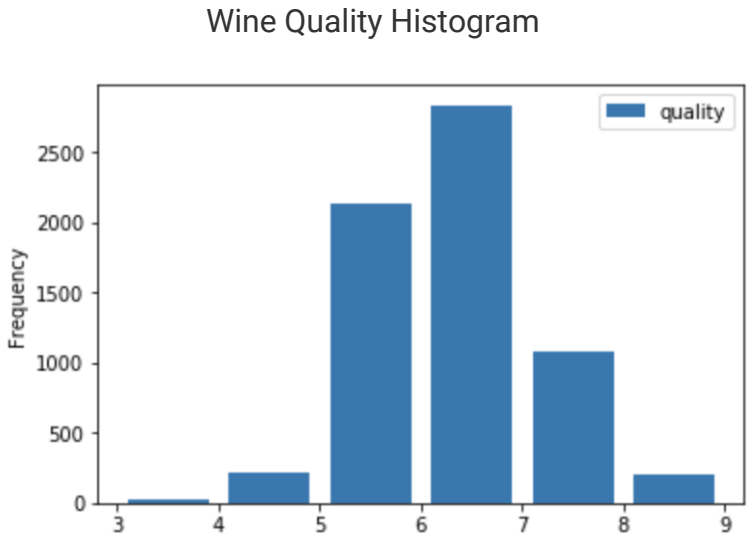
2. Visuals and Statistics to Support Findings

    https://github.com/ahrager/Springboard/blob/master/Capstone1WineQualityAnalysis_Combined_AudreyRager20200826.ipynb

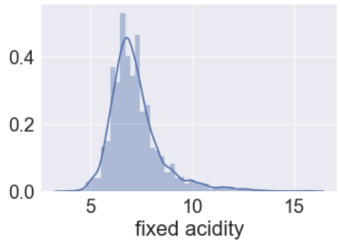
Descriptive statistics for Wine data

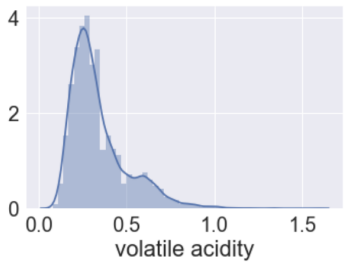
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497
mean	7.2153	0.3397	0.3186	5.4432	0.0560	30.525	115.74	0.9947	3.2185	0.5313	10.492	5.8184
std	1.2964	0.1646	0.1453	4.7578	0.0350	17.749	56.522	0.0030	0.1608	0.1488	1.1927	0.8733
min	3.8	0.08	0	0.60	0.0090	1.0	6.0	0.99	2.7	0.22	8.0	3.0

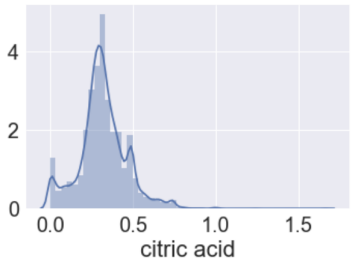
25%	6.4	0.23	0.25	1.8	0.038	17.0	77.0	0.99	3.1	0.43	9.5	5.0
50%	7.0	0.29	0.31	3.0	0.047	29.0	118.0	0.99	3.2	0.51	10	6.0
75%	7.7	0.40	0.39	8.1	0.065	41.0	156.0	1.0	3.3	0.6	11	6.0
max	15.9	1.58	1.66	65.8	0.611	289	440	1.04	4.0	2.0	14.9	9.0

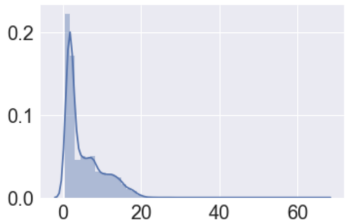


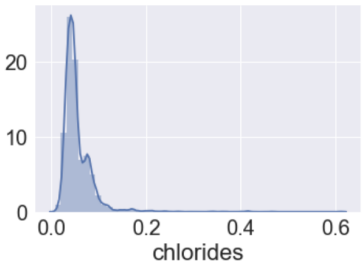
Wine Data Histograms

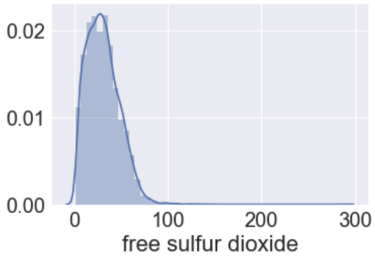


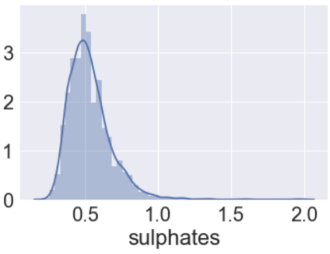


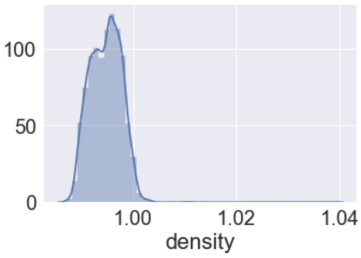


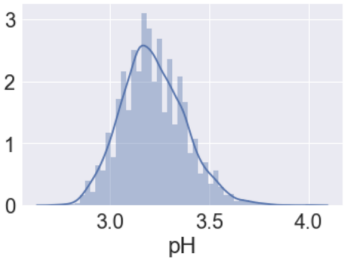


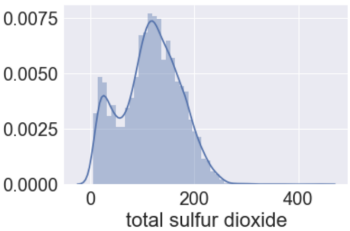


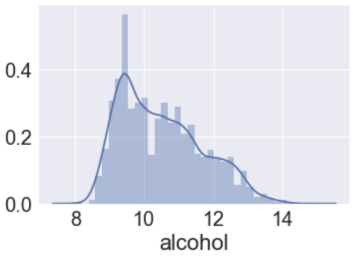


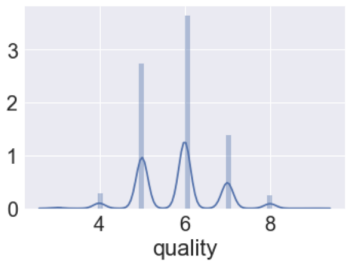




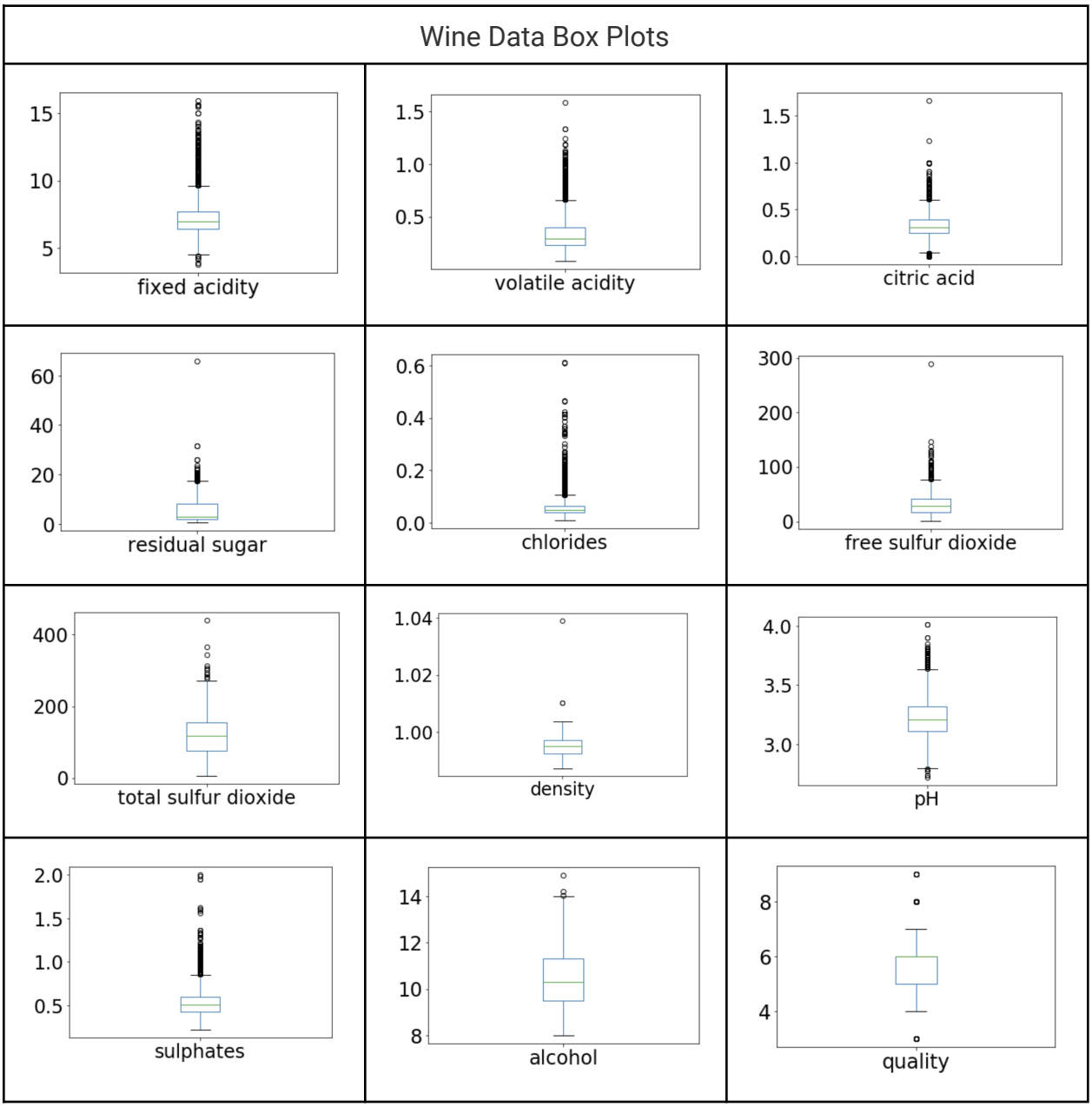


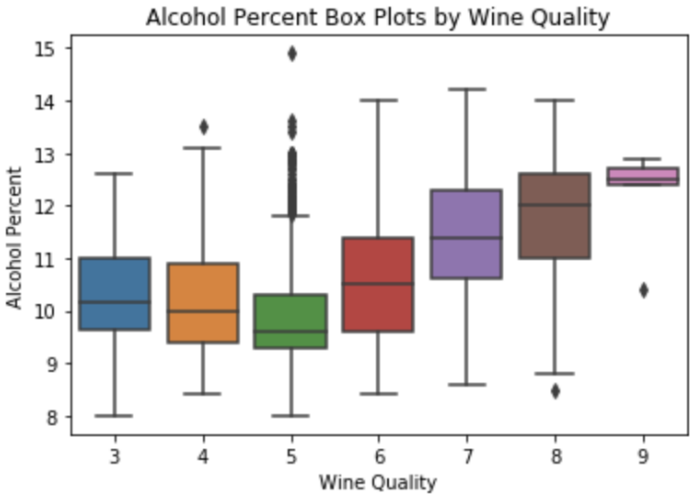


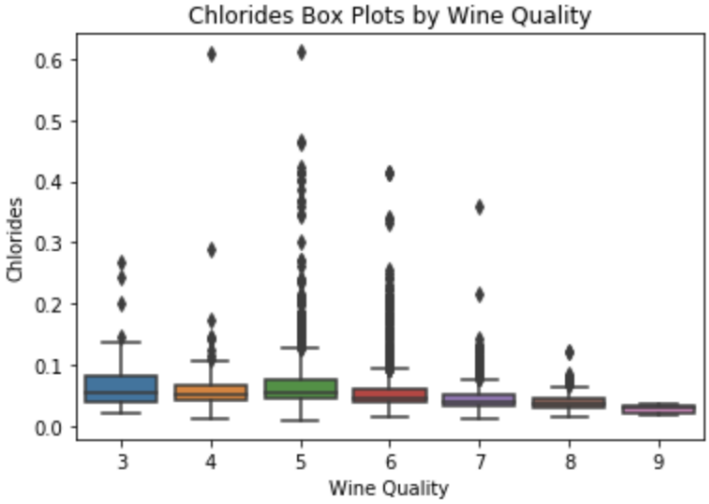


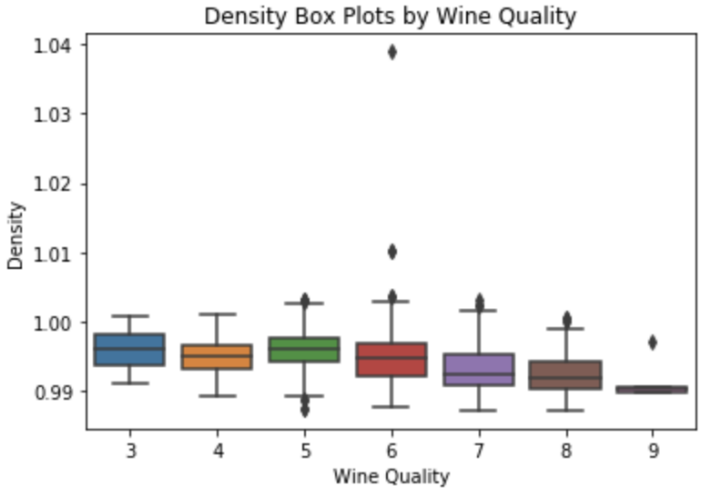


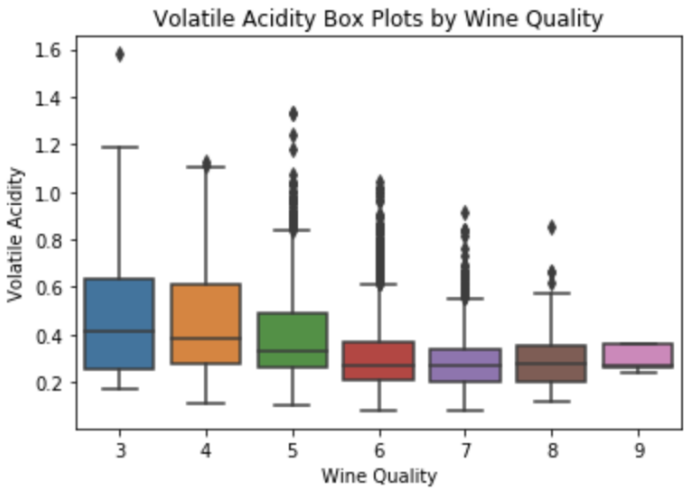
Skewness of Wine Data Attributes		
Attribute	**Skewness**	**Skew Category**
Fixed Acidity	1.722892	Heavily Skewed
Volatile Acidity	1.494751	Heavily Skewed
Citric Acid	0.471622	Light Skewed
Residual Sugar	1.435073	Heavily Skewed
Chlorides	5.398581	Heavily Skewed
Free Sulfur Dioxide	1.219784	Heavily Skewed
Total Sulfur Dioxide	-0.001177	Light Skewed
Density	0.503485	Heavily Skewed
pH	0.386749	Light Skewed
Sulphates	1.796855	Heavily Skewed
Alcohol	0.565587	Heavily Skewed
Quality	0.189579	Light Skewed

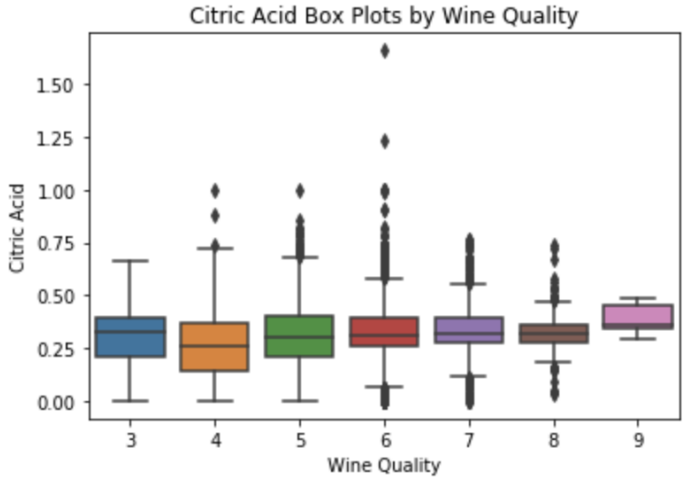

Wine Data Box Plots
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality

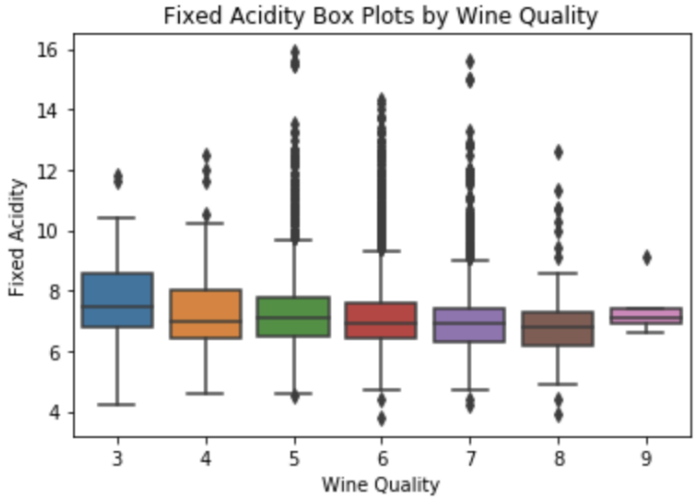

Alcohol Percent Box Plots by Wine Quality
Alcohol Percent
Wine Quality
15
14
13
12
11
10
9
8
3
4
5
6
7
8
9

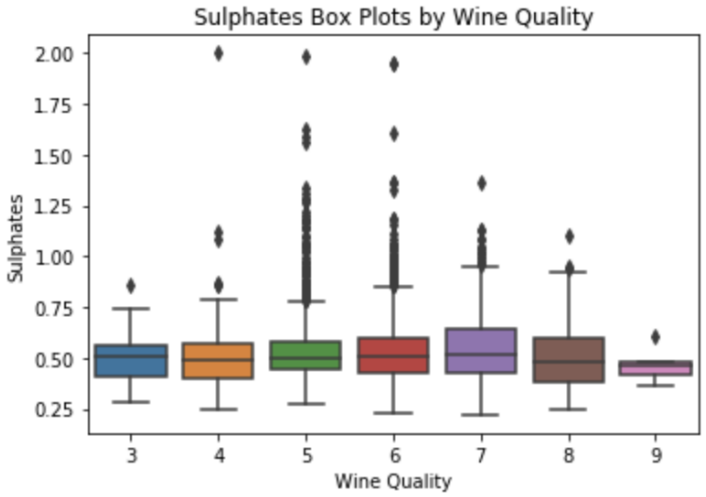

Chlorides Box Plots by Wine Quality
Chlorides
Wine Quality
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9


Density Box Plots by Wine Quality
Density
Wine Quality
1.04
1.03
1.02
1.01
1.00
0.99
3
4
5
6
7
8
9

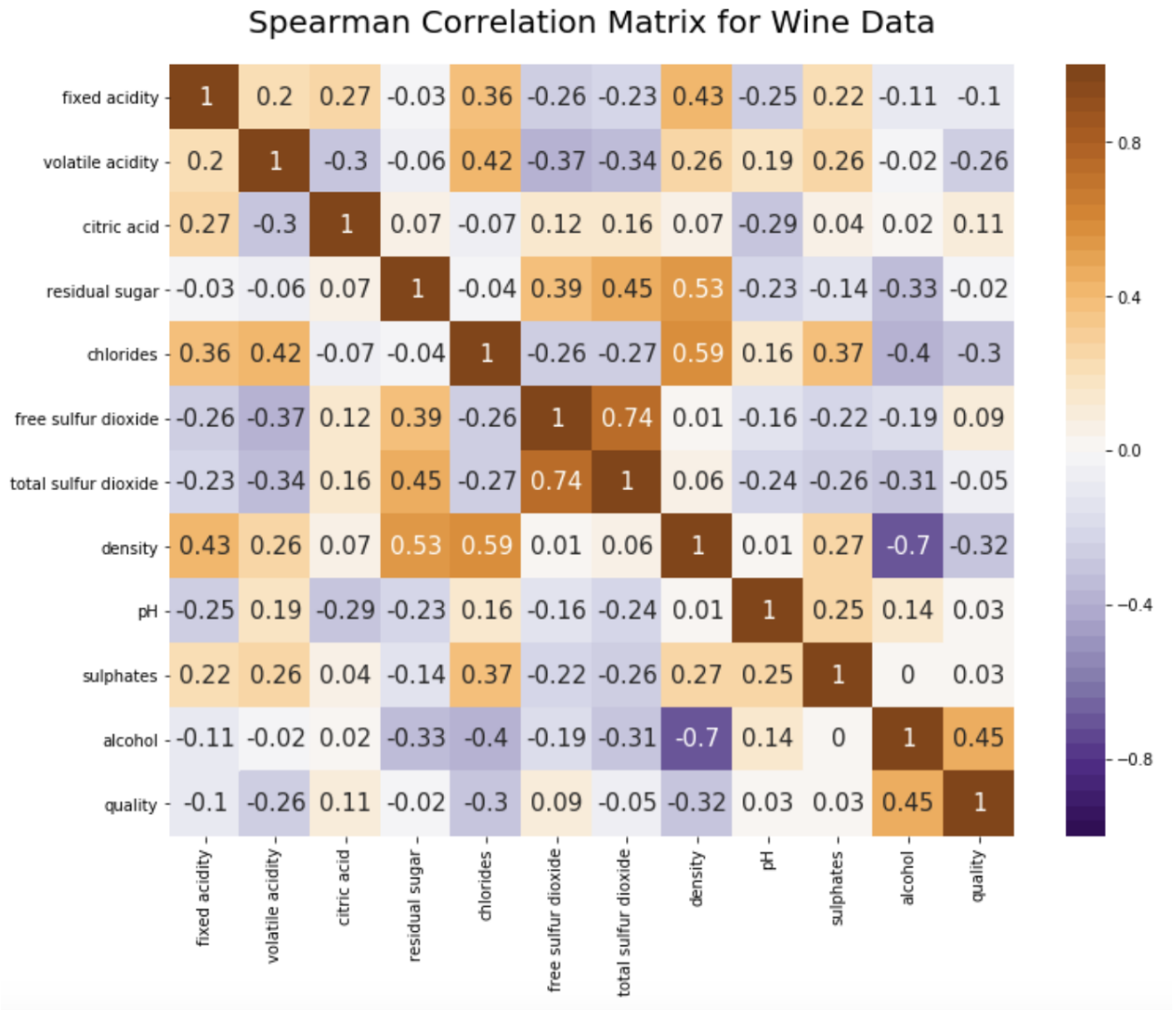

Volatile Acidity Box Plots by Wine Quality
Volatile Acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
8
9
Wine Quality

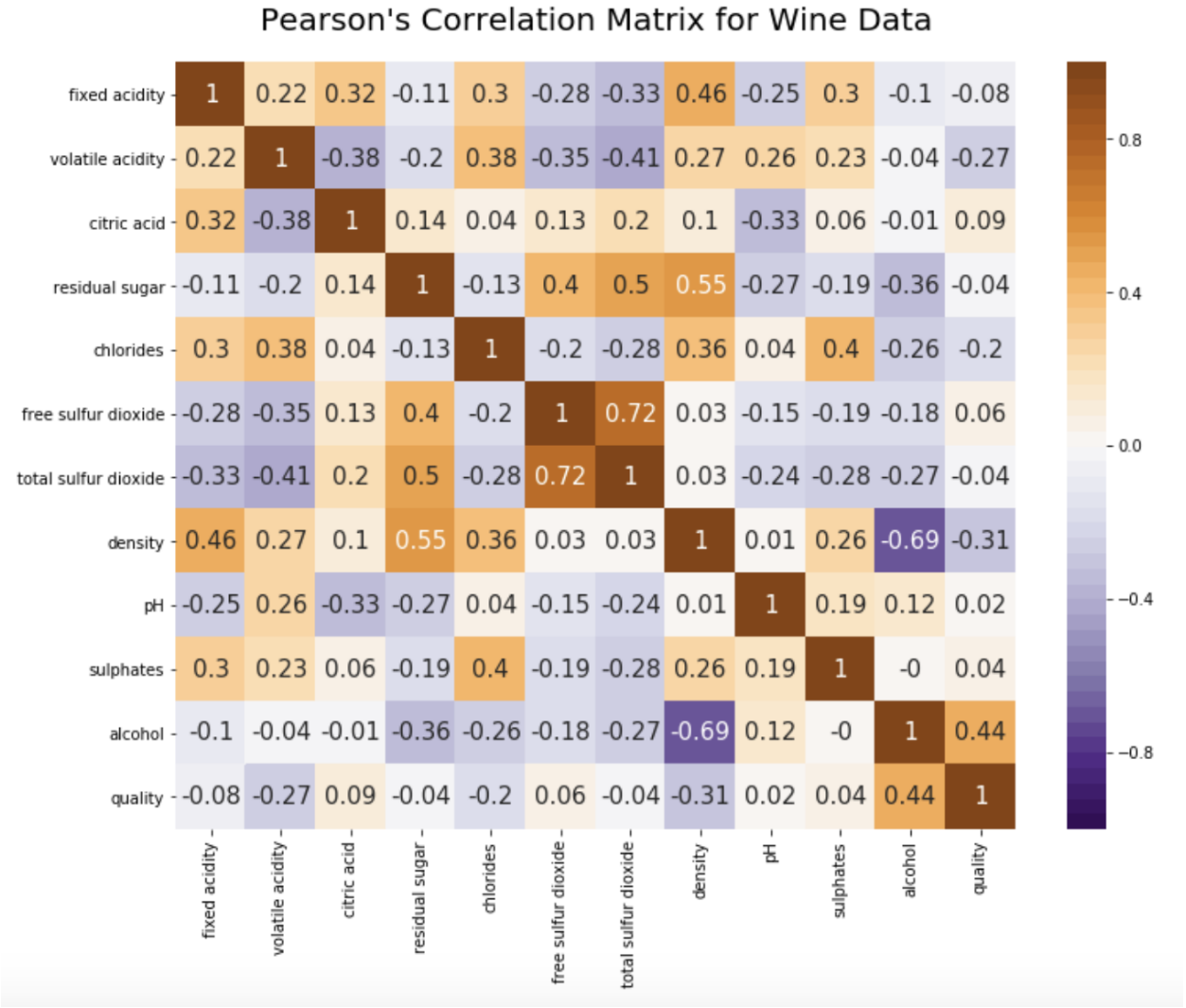

Citric Acid Box Plots by Wine Quality
Citric Acid
1.50
1.25
1.00
0.75
0.50
0.25
0.00
3
4
5
6
7
8
9
Wine Quality

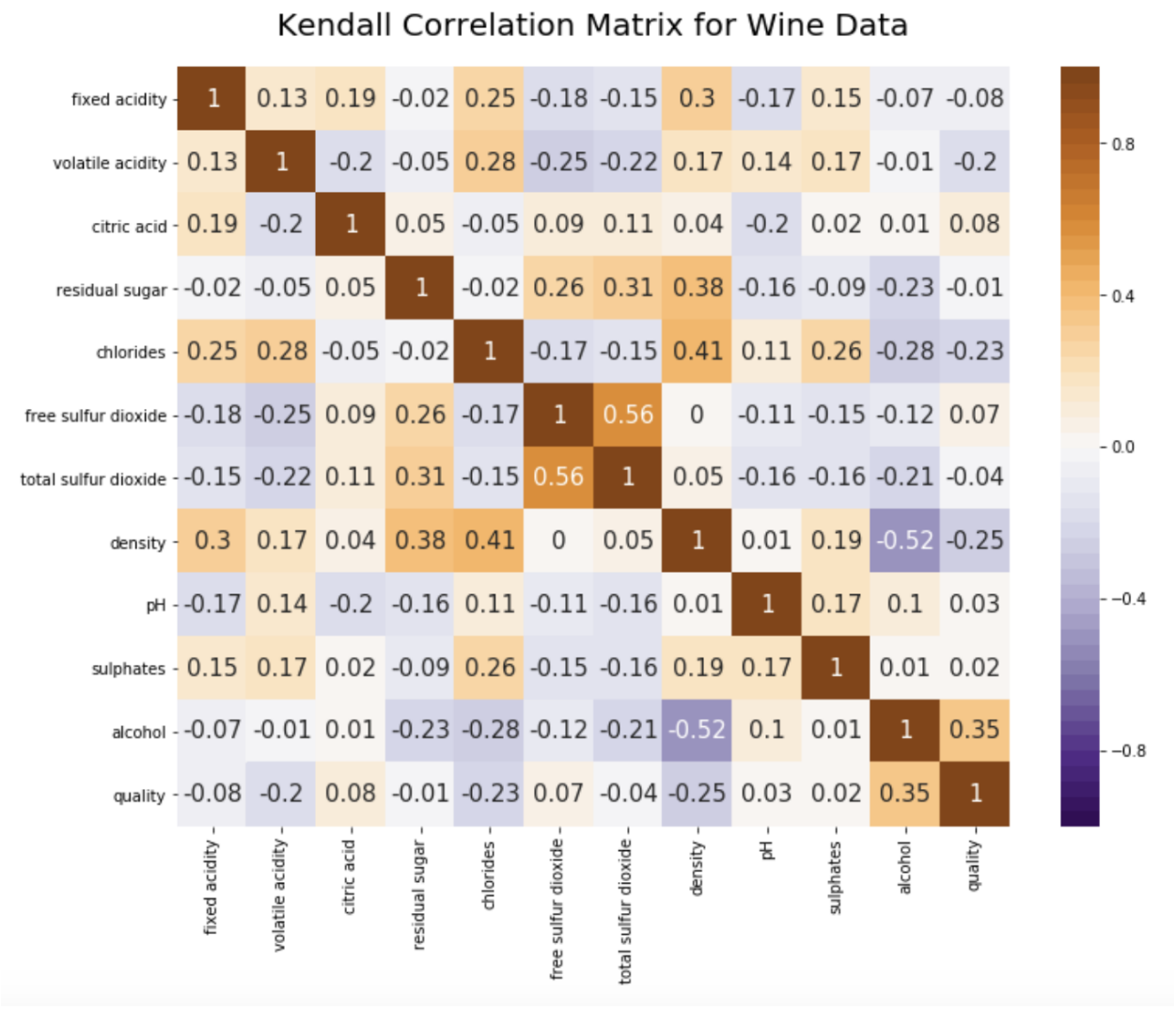

Fixed Acidity Box Plots by Wine Quality
Fixed Acidity
16
14
12
10
8
6
4
3
4
5
6
7
8
9
Wine Quality


Sulphates Box Plots by Wine Quality
Sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
3
4
5
6
7
8
9
Wine Quality


Residual Sugar Box Plots by Wine Quality
Residual Sugar
60
50
40
30
20
10
0
3
4
5
6
7
8
9
Wine Quality


Spearman Correlation Matrix for Wine Data
fixed acidity 1 0.2 0.27 -0.03 0.36 -0.26 -0.23 0.43 -0.25 0.22 -0.11 -0.1
volatile acidity 0.2 1 -0.3 -0.06 0.42 -0.37 -0.34 0.26 0.19 0.26 -0.02 -0.26
citric acid 0.27 -0.3 1 0.07 -0.07 0.12 0.16 0.07 -0.29 0.04 0.02 0.11
residual sugar -0.03 -0.06 0.07 1 -0.04 0.39 0.45 0.53 -0.23 -0.14 -0.33 -0.02
chlorides 0.36 0.42 -0.07 -0.04 1 -0.26 -0.27 0.59 0.16 0.37 -0.4 -0.3
free sulfur dioxide -0.26 -0.37 0.12 0.39 -0.26 1 0.74 0.01 -0.16 -0.22 -0.19 0.09
total sulfur dioxide -0.23 -0.34 0.16 0.45 -0.27 0.74 1 0.06 -0.24 -0.26 -0.31 -0.05
density 0.43 0.26 0.07 0.53 0.59 0.01 0.06 1 0.01 0.27 -0.7 -0.32
pH -0.25 0.19 -0.29 -0.23 0.16 -0.16 -0.24 0.01 1 0.25 0.14 0.03
sulphates 0.22 0.26 0.04 -0.14 0.37 -0.22 -0.26 0.27 0.25 1 0 0.03
alcohol -0.11 -0.02 0.02 -0.33 -0.4 -0.19 -0.31 -0.7 0.14 0 1 0.45
quality -0.1 -0.26 0.11 -0.02 -0.3 0.09 -0.05 -0.32 0.03 0.03 0.45 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
0.8
0.4
0.0
−0.4
−0.8

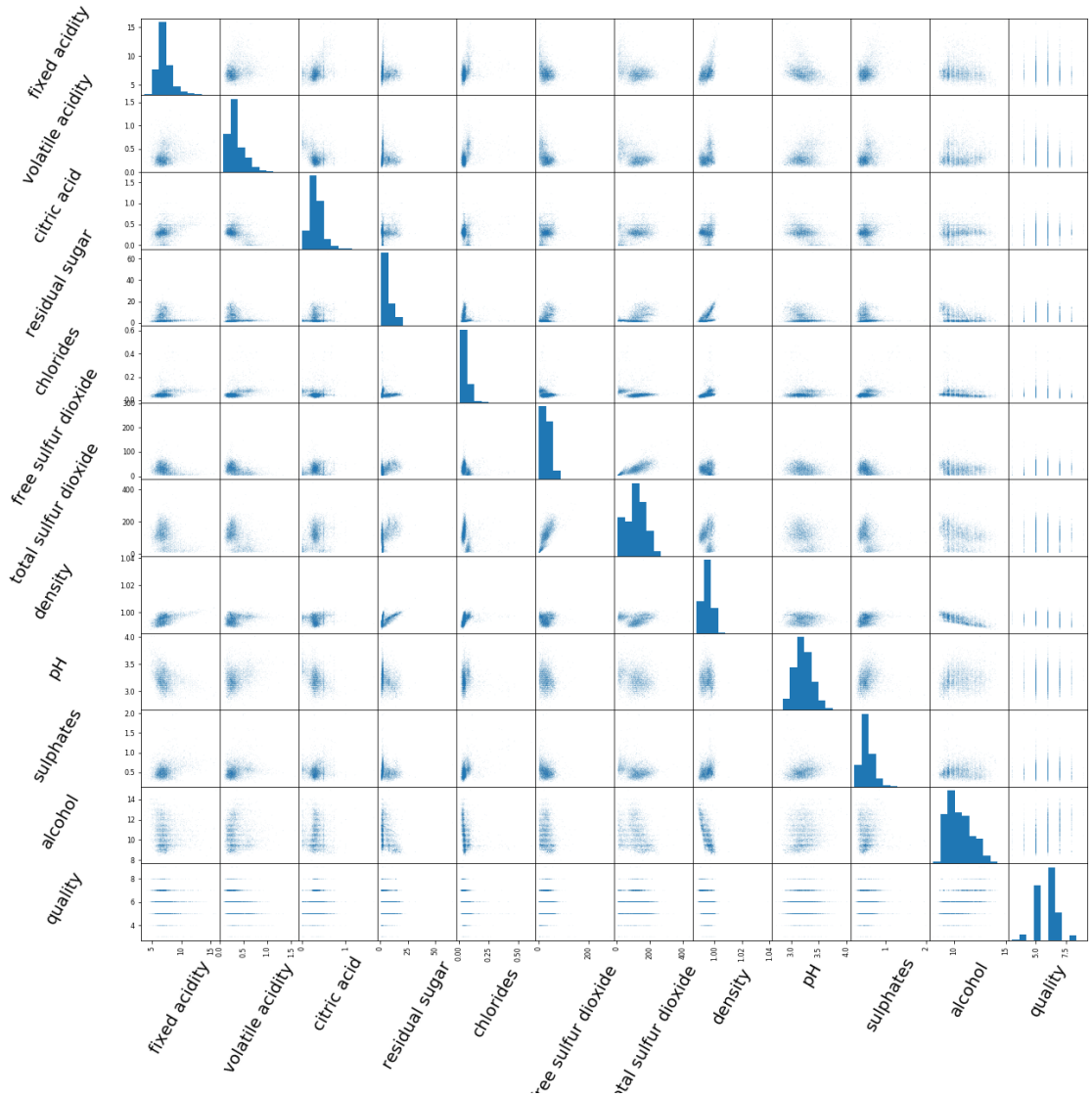

Pearson's Correlation Matrix for Wine Data
fixed acidity 1 0.22 0.32 -0.11 0.3 -0.28 -0.33 0.46 -0.25 0.3 -0.1 -0.08
volatile acidity 0.22 1 -0.38 -0.2 0.38 -0.35 -0.41 0.27 0.26 0.23 -0.04 -0.27
citric acid 0.32 -0.38 1 0.14 0.04 0.13 0.2 0.1 -0.33 0.06 -0.01 0.09
residual sugar -0.11 -0.2 0.14 1 -0.13 0.4 0.5 0.55 -0.27 -0.19 -0.36 -0.04
chlorides 0.3 0.38 0.04 -0.13 1 -0.2 -0.28 0.36 0.04 0.4 -0.26 -0.2
free sulfur dioxide -0.28 -0.35 0.13 0.4 -0.2 1 0.72 0.03 -0.15 -0.19 -0.18 0.06
total sulfur dioxide -0.33 -0.41 0.2 0.5 -0.28 0.72 1 0.03 -0.24 -0.28 -0.27 -0.04
density 0.46 0.27 0.1 0.55 0.36 0.03 0.03 1 0.01 0.26 -0.69 -0.31
pH -0.25 0.26 -0.33 -0.27 0.04 -0.15 -0.24 0.01 1 0.19 0.12 0.02
sulphates 0.3 0.23 0.06 -0.19 0.4 -0.19 -0.28 0.26 0.19 1 -0 0.04
alcohol -0.1 -0.04 -0.01 -0.36 -0.26 -0.18 -0.27 -0.69 0.12 -0 1 0.44
quality -0.08 -0.27 0.09 -0.04 -0.2 0.06 -0.04 -0.31 0.02 0.04 0.44 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
0.8
0.4
0.0
−0.4
−0.8


Kendall Correlation Matrix for Wine Data
fixed acidity 1 0.13 0.19 -0.02 0.25 -0.18 -0.15 0.3 -0.17 0.15 -0.07 -0.08
volatile acidity 0.13 1 -0.2 -0.05 0.28 -0.25 -0.22 0.17 0.14 0.17 -0.01 -0.2
citric acid 0.19 -0.2 1 0.05 -0.05 0.09 0.11 0.04 -0.2 0.02 0.01 0.08
residual sugar -0.02 -0.05 0.05 1 -0.02 0.26 0.31 0.38 -0.16 -0.09 -0.23 -0.01
chlorides 0.25 0.28 -0.05 -0.02 1 -0.17 -0.15 0.41 0.11 0.26 -0.28 -0.23
free sulfur dioxide -0.18 -0.25 0.09 0.26 -0.17 1 0.56 0 -0.11 -0.15 -0.12 0.07
total sulfur dioxide -0.15 -0.22 0.11 0.31 -0.15 0.56 1 0.05 -0.16 -0.16 -0.21 -0.04
density 0.3 0.17 0.04 0.38 0.41 0 0.05 1 0.01 0.19 -0.52 -0.25
pH -0.17 0.14 -0.2 -0.16 0.11 -0.11 -0.16 0.01 1 0.17 0.1 0.03
sulphates 0.15 0.17 0.02 -0.09 0.26 -0.15 -0.16 0.19 0.17 1 0.01 0.02
alcohol -0.07 -0.01 0.01 -0.23 -0.28 -0.12 -0.21 -0.52 0.1 0.01 1 0.35
quality -0.08 -0.2 0.08 -0.01 -0.23 0.07 -0.04 -0.25 0.03 0.02 0.35 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
0.8
0.4
0.0
−0.4
−0.8

Summary of Correlation Results for Wine Data			
Attribute	Spearman	Pearson's	Kendall
Alcohol	0.446925 (1)	0.444319 (1)	0.352430 (1)
Density	-0.322806 (2)	-0.305858 (2)	-0.247978 (2)
Chlorides	-0.295054 (3)	-0.200666 (4)	-0.228872 (3)
Volatile Acidity	-0.257806 (4)	-0.265699 (3)	-0.199101 (4)
Citric Acid	0.105711 (5)	0.085532 (5)	0.082160 (5)
Fixed Acidity	-0.098154 (6)	-0.076743 (6)	-0.075990 (6)
Free Sulfur Dioxide	0.086865 (7)	0.055463 (7)	0.066713 (7)
Total Sulfur Dioxide	-0.054777 (8)	-0.041385 (8)	-0.042283 (8)
pH	0.032538 (9)	0.019506 (11)	0.025223 (9)
Sulphates	0.029831 (10)	0.038485 (9)	0.023679 (10)
Residual Sugar	-0.016891 (11)	-0.036980 (10)	-0.013097 (11)

Correlation Ranking for Wine Data			
Correlation Ranking	Spearman	Pearson's	Kendall
1	Alcohol (0.446925)	Alcohol (0.444319)	Alcohol (0.352430)
2	Density (-0.322806)	Density (-0.305858)	Density (-0.247978)
3	Chlorides (-0.295054)	Volatile Acidity (-0.265699)	Chlorides (-0.228872)
4	Volatile Acidity (-0.257806)	Chlorides (-0.200666)	Volatile Acidity (-0.199101)
5	Citric Acid (0.105711)	Citric Acid (0.085532)	Citric Acid (0.082160)
6	Fixed Acidity (-0.098154)	Fixed Acidity (-0.076743)	Fixed Acidity (-0.075990)
7	Free Sulfur Dioxide (0.086865)	Free Sulfur Dioxide (0.055463)	Free Sulfur Dioxide (0.066713)
8	Total Sulfur Dioxide (-0.054777)	Total Sulfur Dioxide (-0.041385)	Total Sulfur Dioxide (-0.042283)
9	pH (0.032538)	Sulphates (0.038485)	pH (0.025223)
10	Sulphates (0.029831)	Residual Sugar (-0.03698)	Sulphates (0.023679)
11	Residual Sugar (-0.016891)	pH (0.019506)	Residual Sugar (-0.013097)

Wine Scatter Matrix



REFERENCES

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

Wine Quality Dataset from UC Irvine Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/Wine+Quality

McMilan, R., 2020, State of the U.S. Wine Industry, Silicon Valley Bank Wine Division, https://www.svb.com/globalassets/library/uploadedfiles/reports/svb-2020-state-of-the-wine-industry-report-final.pdf, 71 p.

PSU, 2020, Analysis of Wine Quality Data, Applied Data Mining and Statistical Learning (STAT 508), Department of Statistics, Eberly College of Science, Pennsylvania State University, open.ed@psu, https://online.stat.psu.edu/stat508/lesson/analysis-wine-quality-data, accessed February 12, 2020.