Data Wrangling

Capstone Project 1: Wine Quality Data

Audrey Rager

For my Capstone Project 1, I will be analyzing a Wine Quality data set from UC Irvine Machine Learning Repository.

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.

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.

Methods

Data Wrangling

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?

   Outliers in many variables. I eliminated values greater than or equal to 1.5 times the Interquartile Range (IQR) above the third quartile (Q3) and less than 1.5 times the IQR below the first quartile (Q1).

Exploratory Data Analysis

- Summary of findings
    a. Before outliers removed
        - 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.

- Visuals and Statistics to Support Findings

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

Descriptive statistics for Wine Data Before Eliminating Outliers

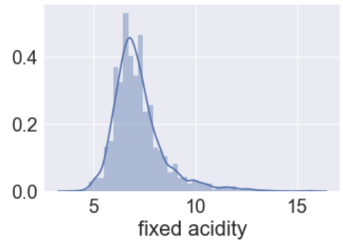
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497	6497
mean	7.2153	0.3397	0.3186	5.4432	0.0560	30.525	115.74	0.9947	3.2185	0.5313	10.492	5.8184
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

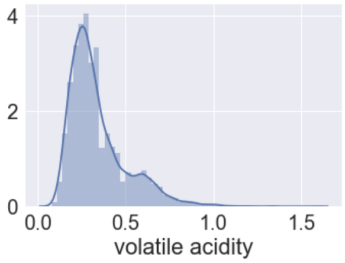
Descriptive statistics for Wine Data After Eliminating Outliers

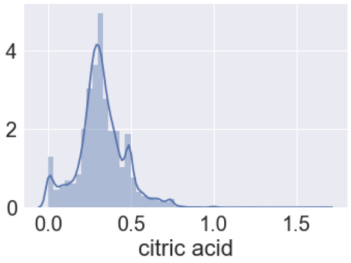
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4514	4514	4514	4514	4514	4514	4514	4514	4514	4514	4514	4514
mean	6.880472	0.283068	0.318347	5.979098	0.044407	33.145215	130.33828	0.993979	3.201232	0.493051	10.590274	5.922907
std	0.790118	0.100359	0.088194	4.764865	0.012567	15.265536	46.222992	0.002822	0.143035	0.107871	1.215803	0.864799
min	4.800	0.080	0.090	0.600	0.012000	2.00	9.00	0.987110	2.8200	0.2200	8.40	3.00
25%	6.300	0.210	0.260000	1.800	0.03600	22.00	101.0000	0.991700	3.10	0.4100	9.500	5.00
50%	6.80	0.2700	0.3100	4.700	0.043000	32.00	128.00000	0.993805	3.190	0.480	10.500	6.00
75%	7.400	0.3300	0.3700	9.00	0.051000	44.00	162.00	0.996058	3.300	0.560	11.400	6.00
max	9.100000	0.640000	0.550000	19.300000	0.081000	76.000000	252.000000	1.001700	3.590000	0.790000	14.200000	9.000000

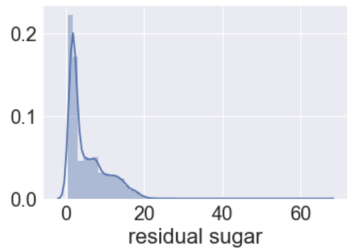
Wine Quality Histograms

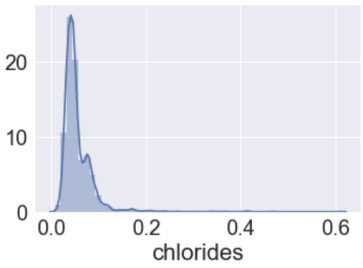
Wine Data Histograms Before Eliminating Outliers

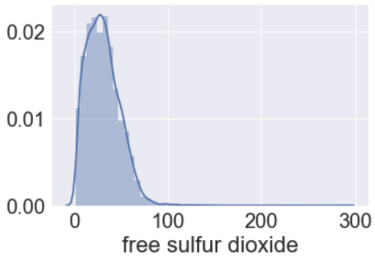


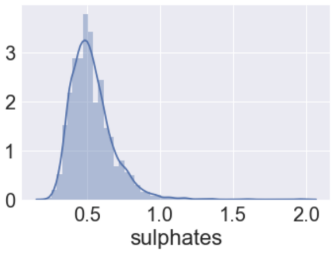


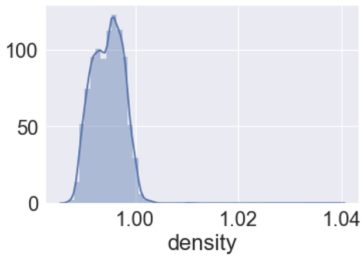


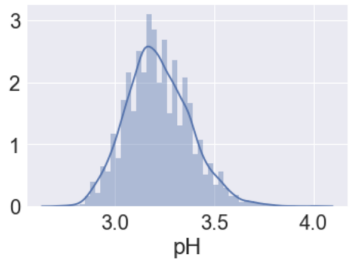


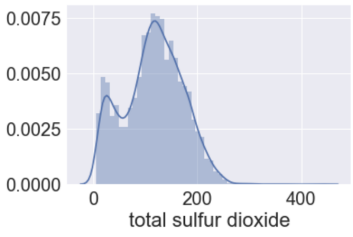


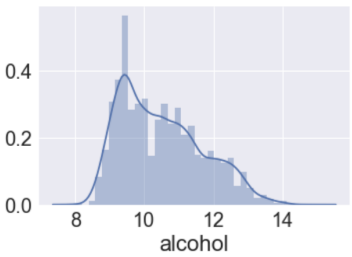


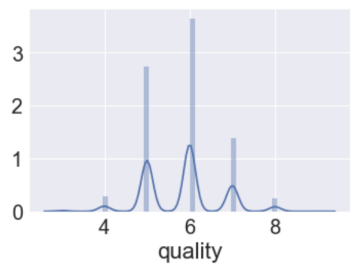




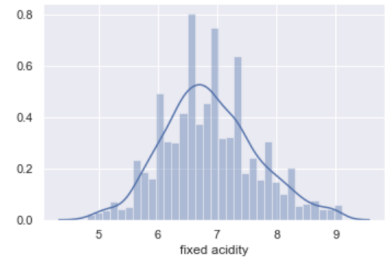


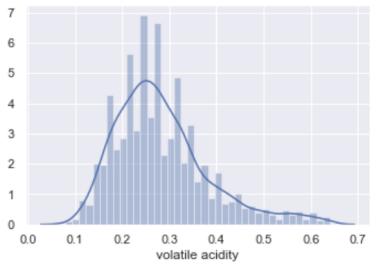


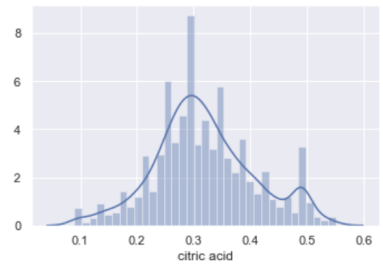


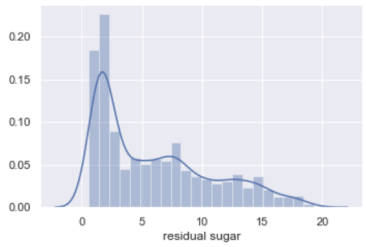


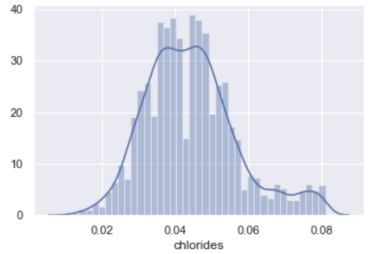
Wine Data Histograms After Eliminating Outliers

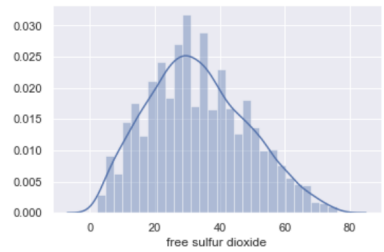


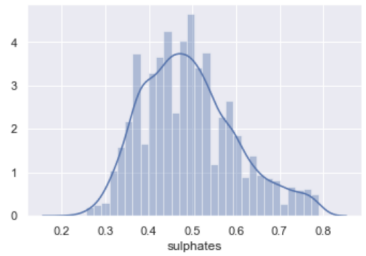


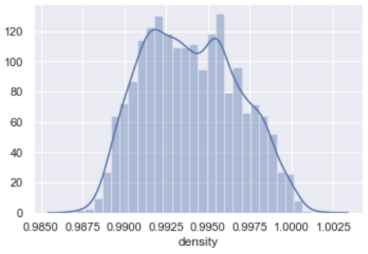


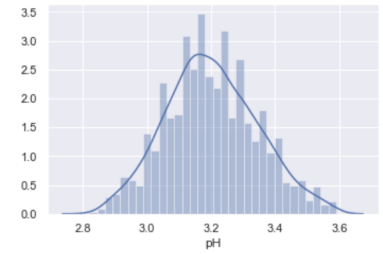


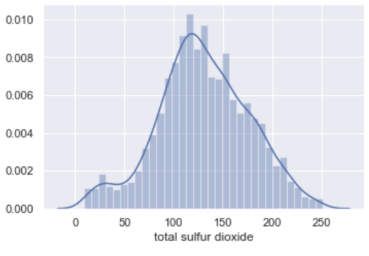


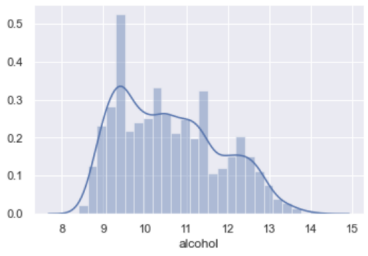


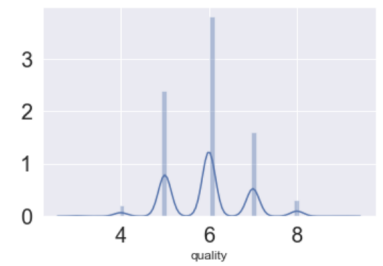








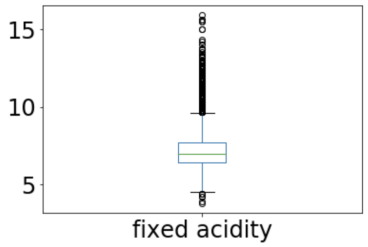


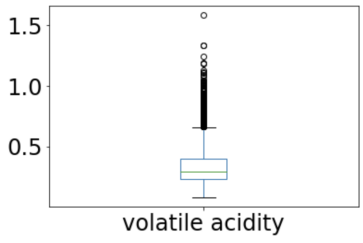


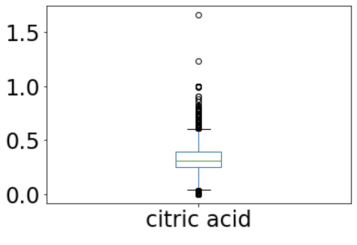
Skewness of Wine Data Attributes Before Eliminating Outliers		
Attribute	**Skewness**	**Skew Category**
Fixed Acidity	1.722892	Heavily Skewed
Volatile Acidity	1.494751	Heavily Skewed
Citric Acid	0.471622	Light Skewed
Residual Sugar	1.435073	Heavily Skewed
Chlorides	5.398581	Heavily Skewed
Free Sulfur 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

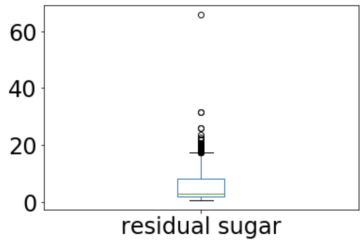
Skewness of Wine Data Attributes After Eliminating Outliers		
Attribute	**Skewness**	**Skew Category**
Fixed Acidity	0.287594	Light Skewed
Volatile Acidity	1.032212	Heavily Skewed
Citric Acid	0.218450	Light Skewed
Residual Sugar	0.819261	Heavily Skewed
Chlorides	0.643093	Heavily Skewed
Free Sulfur Dioxide	0.294768	Light Skewed
Total Sulfur Dioxide	-0.071066	Light Skewed
Density	0.167702	Light Skewed
pH	0.168798	Light Skewed
Sulphates	0.523771	Heavily Skewed
Alcohol	0.420021	Light Skewed
Quality	0.202390	Light Skewed

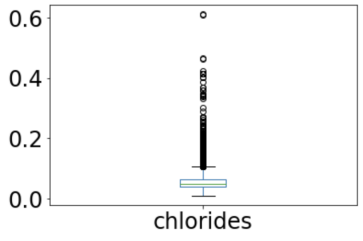
Wine Data Box Plots Before Eliminating Outliers

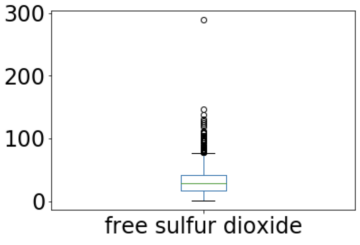


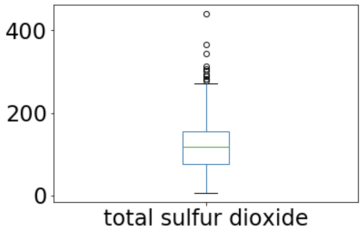


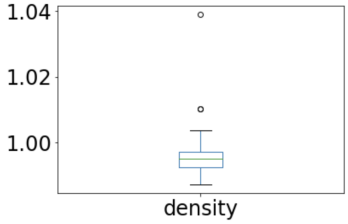


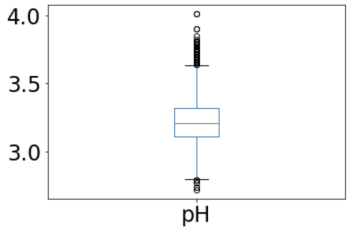


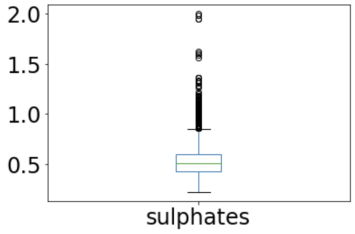


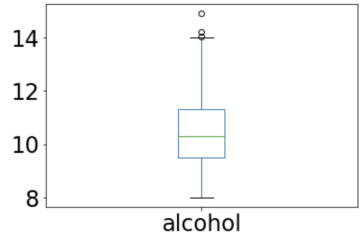


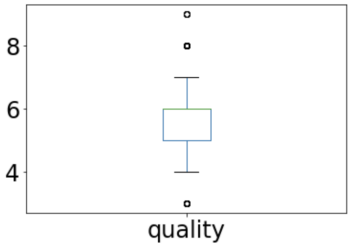


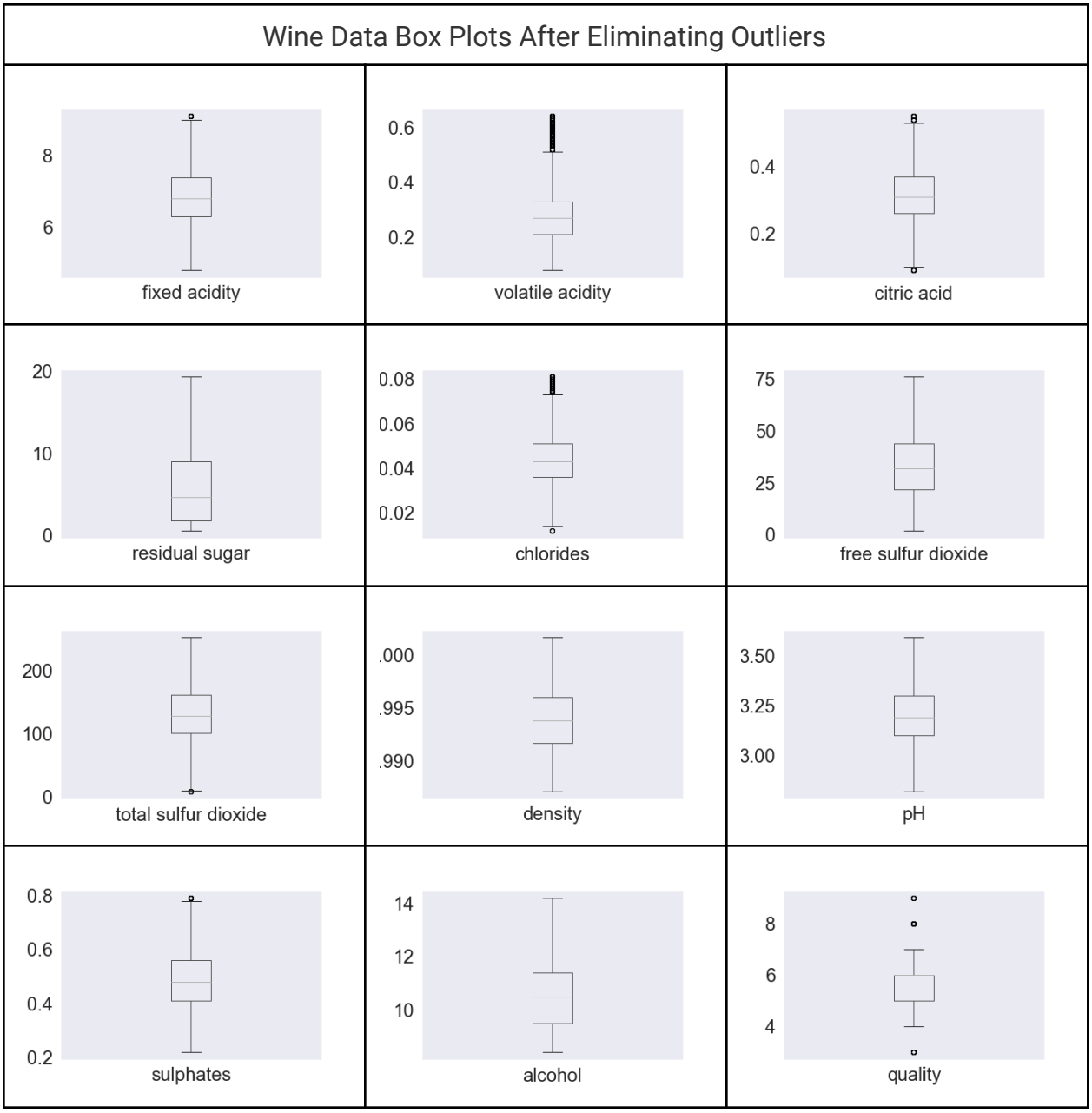




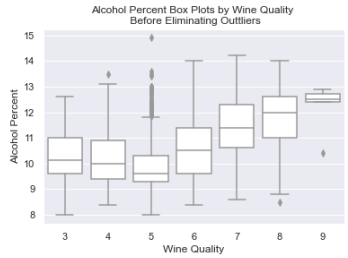


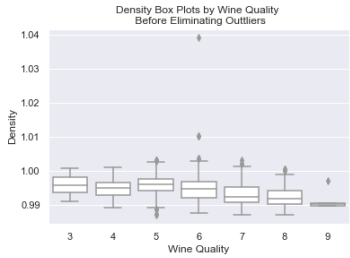


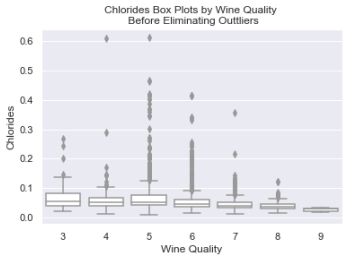


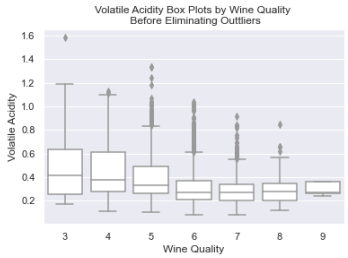

Wine Data Box Plots After Eliminating Outliers
8
6
fixed acidity
0.6
0.4
0.2
volatile acidity
0.4
0.2
citric acid
20
10
0
residual sugar
0.08
0.06
0.04
0.02
chlorides
75
50
25
0
free sulfur dioxide
200
100
0
total sulfur dioxide
.000
.995
.990
density
3.50
3.25
3.00
pH
0.8
0.6
0.4
0.2
sulphates
14
12
10
alcohol
8
6
4
quality

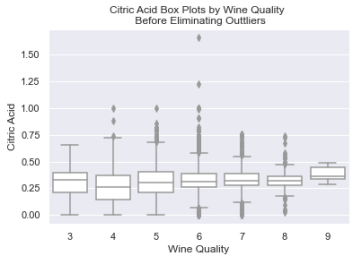
Wine Data Attributes by Quality Box Plots Before Eliminating Outliers

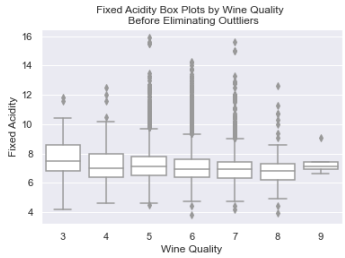


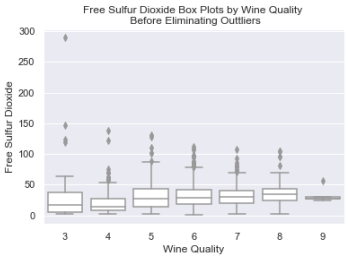


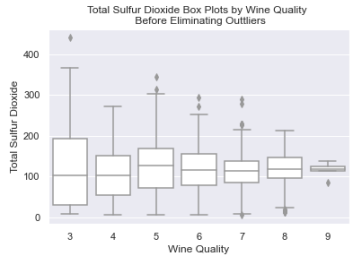


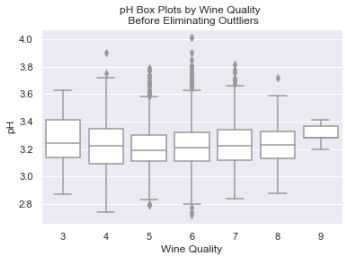


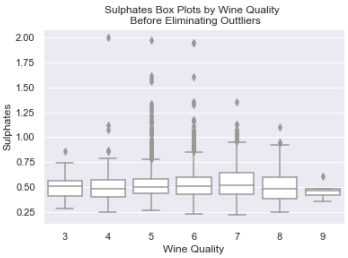


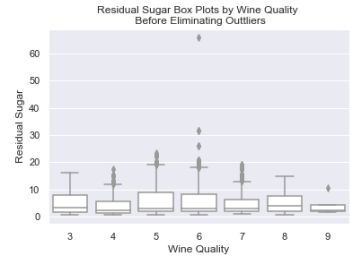




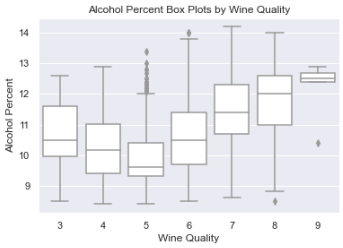


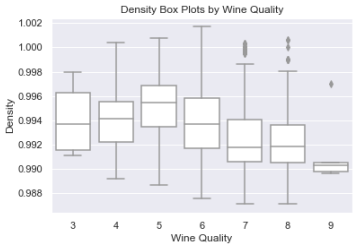


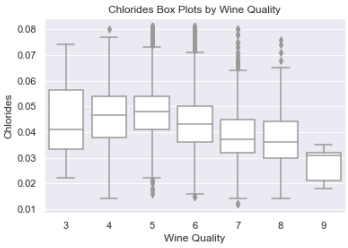


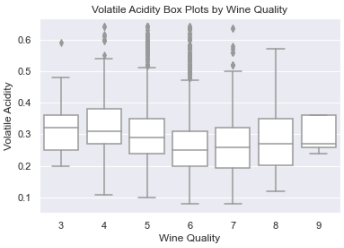


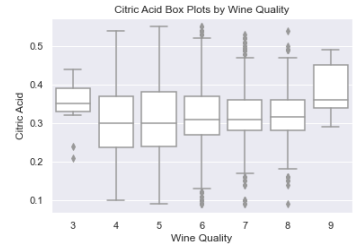
Wine Data Attributes by Quality Box Plots After Eliminating Outliers

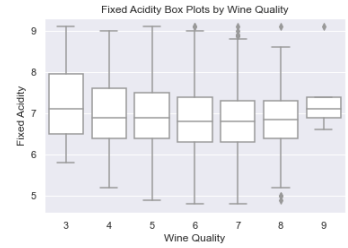


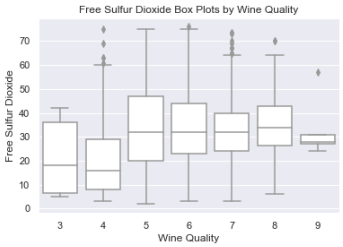


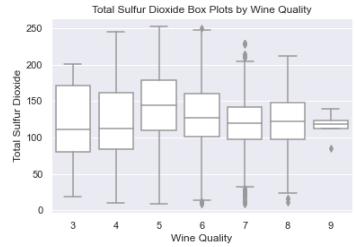


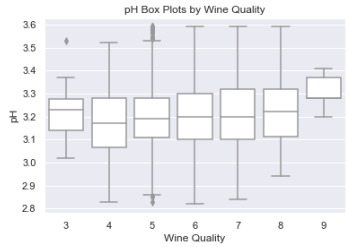


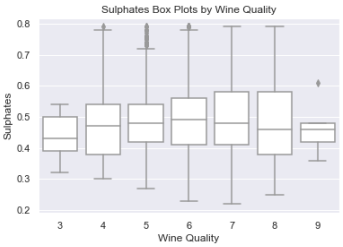


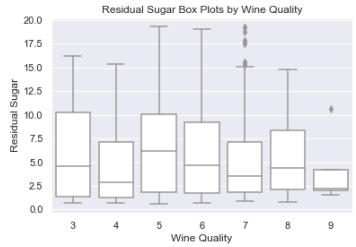


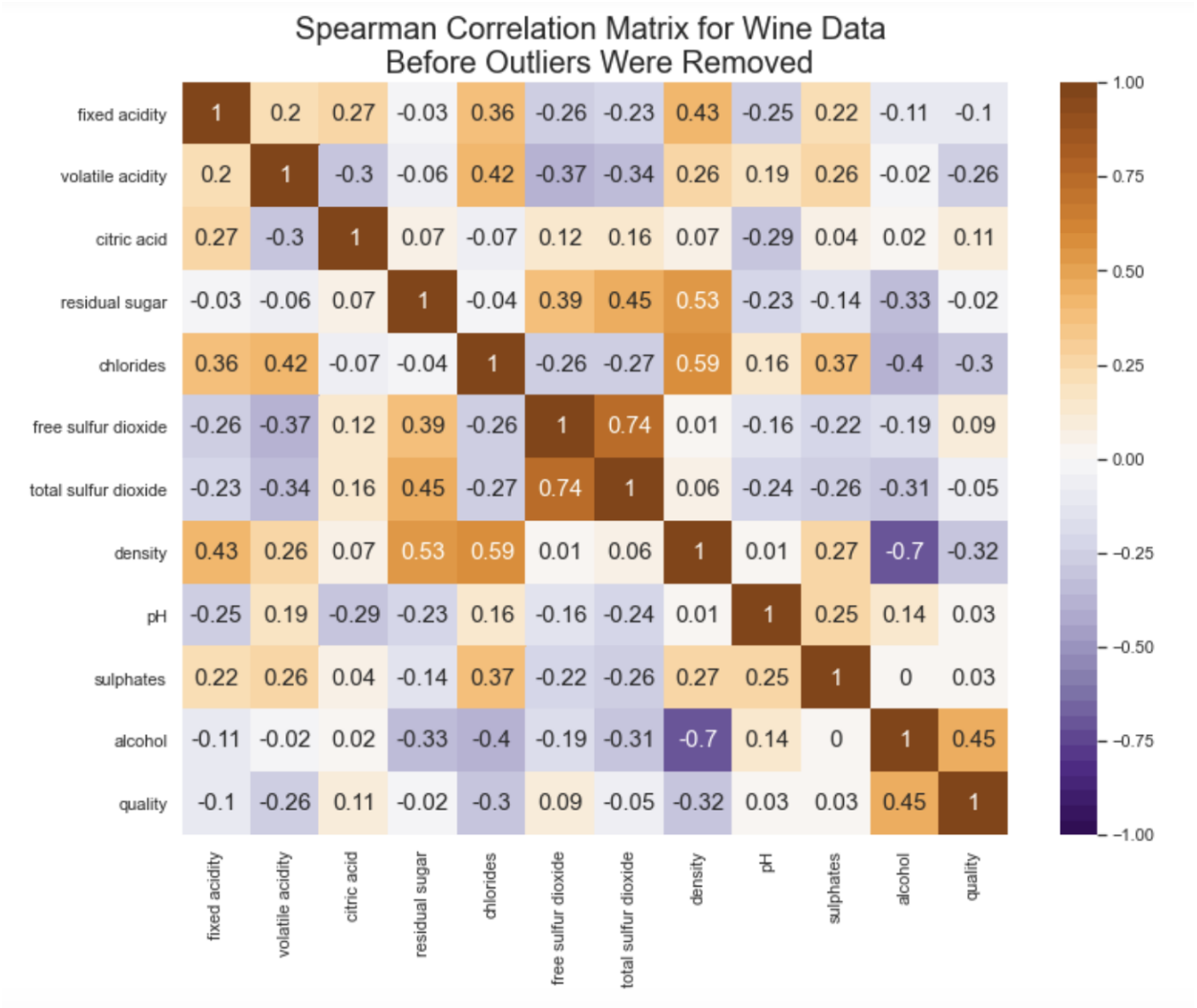


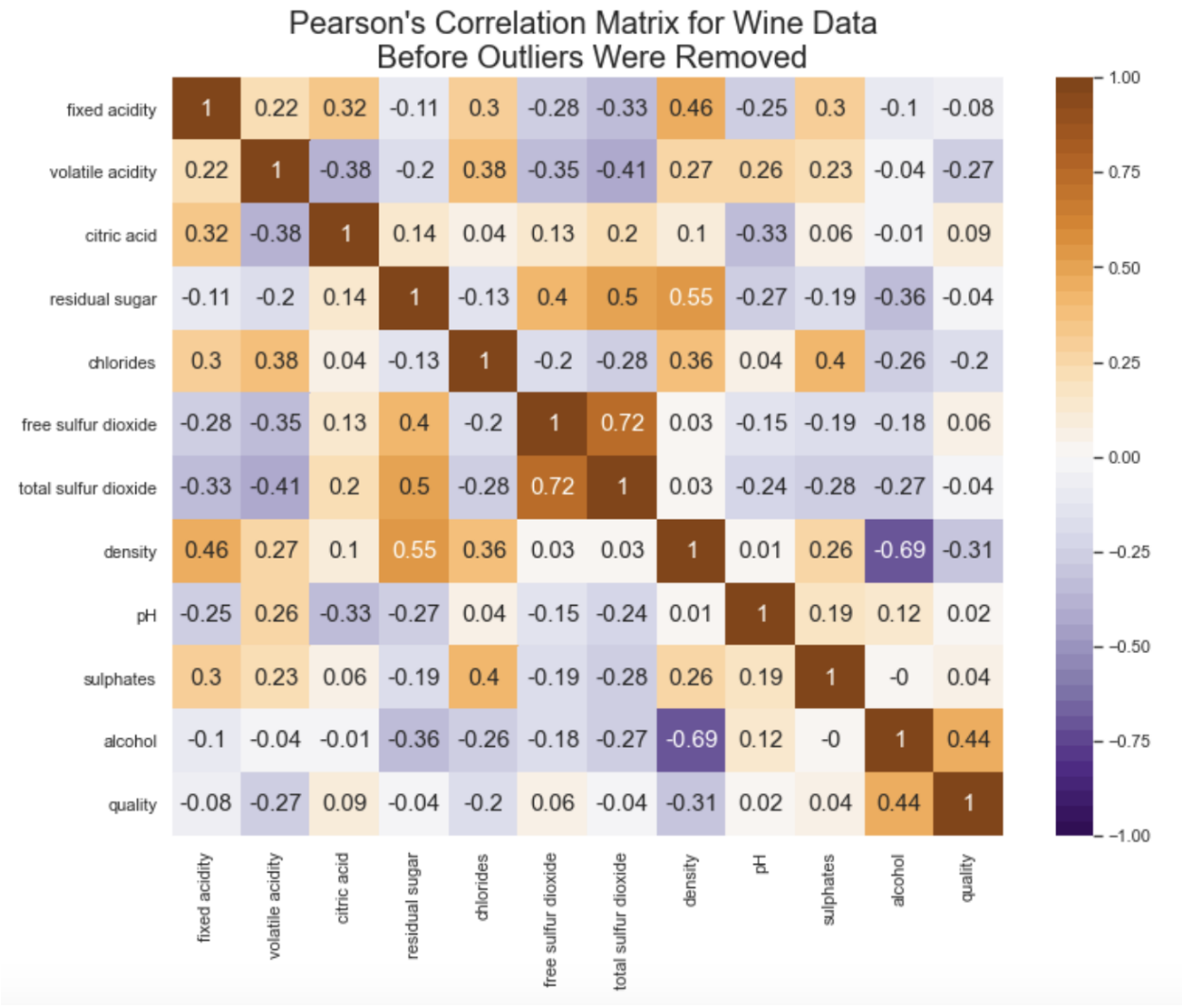


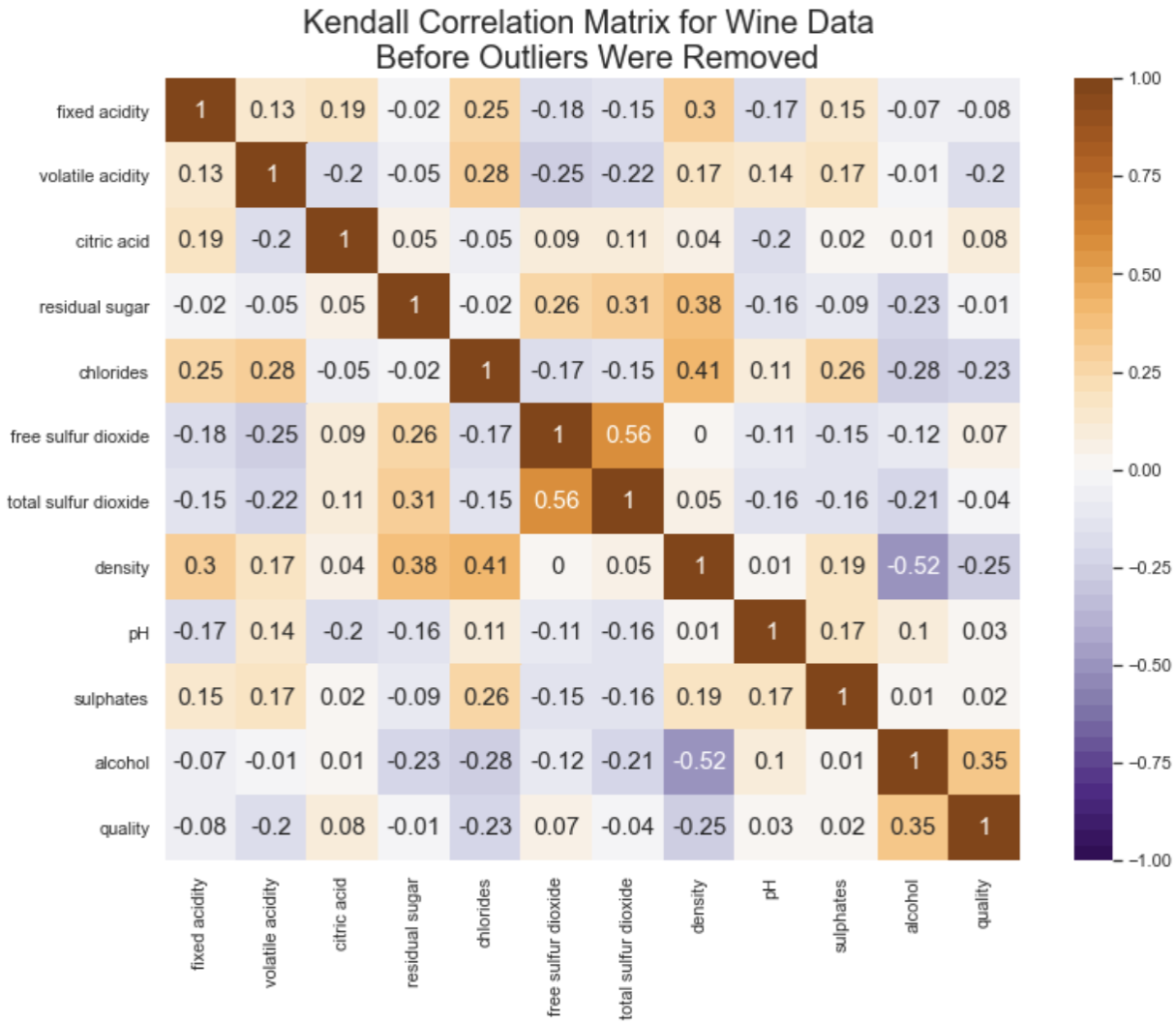






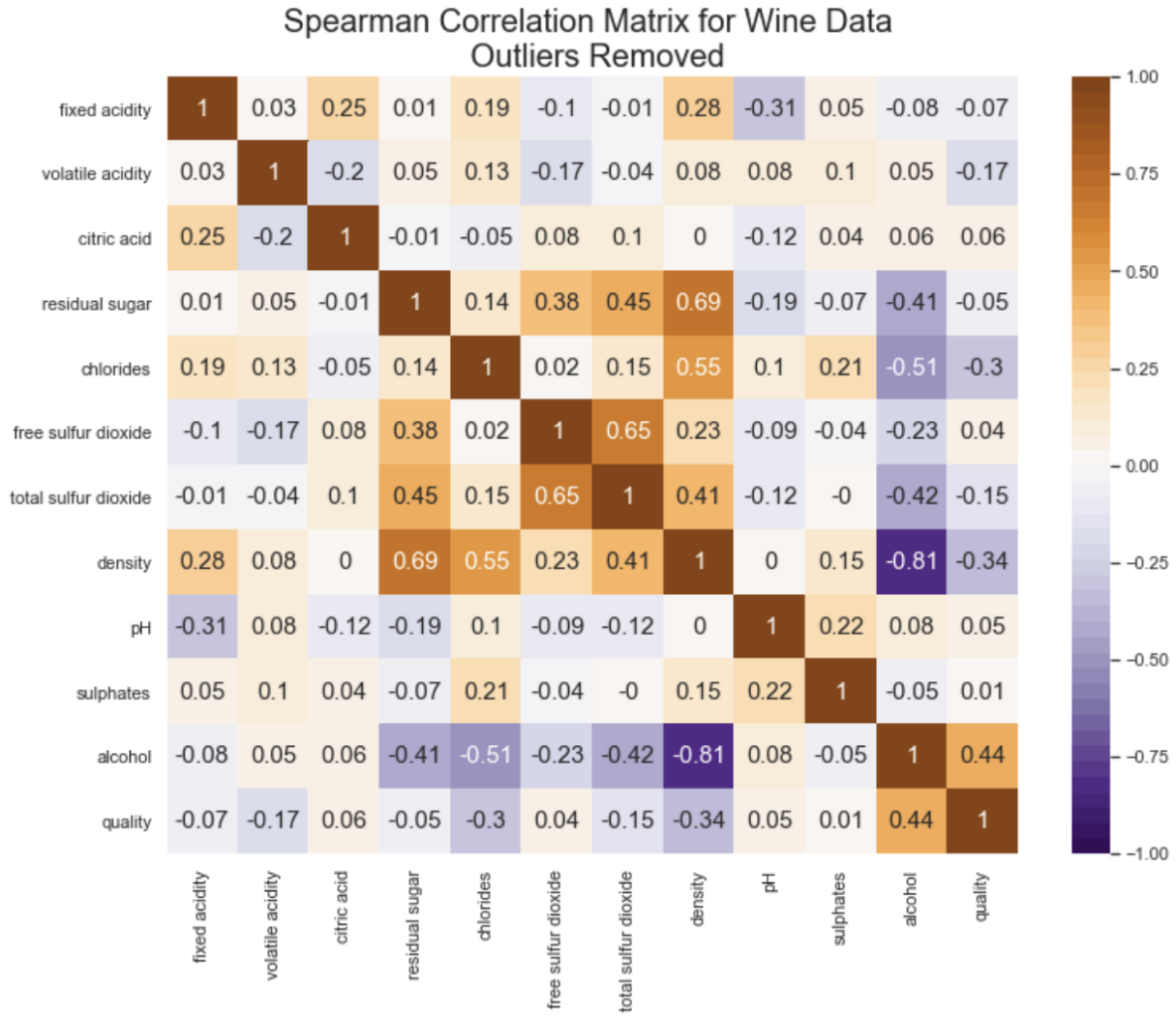

Spearman Correlation Matrix for Wine Data
Before Outliers Were Removed
fixed acidity 1 0.2 0.27 -0.03 0.36 -0.26 -0.23 0.43 -0.25 0.22 -0.11 -0.1
volatile acidity 0.2 1 -0.3 -0.06 0.42 -0.37 -0.34 0.26 0.19 0.26 -0.02 -0.26
citric acid 0.27 -0.3 1 0.07 -0.07 0.12 0.16 0.07 -0.29 0.04 0.02 0.11
residual 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
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

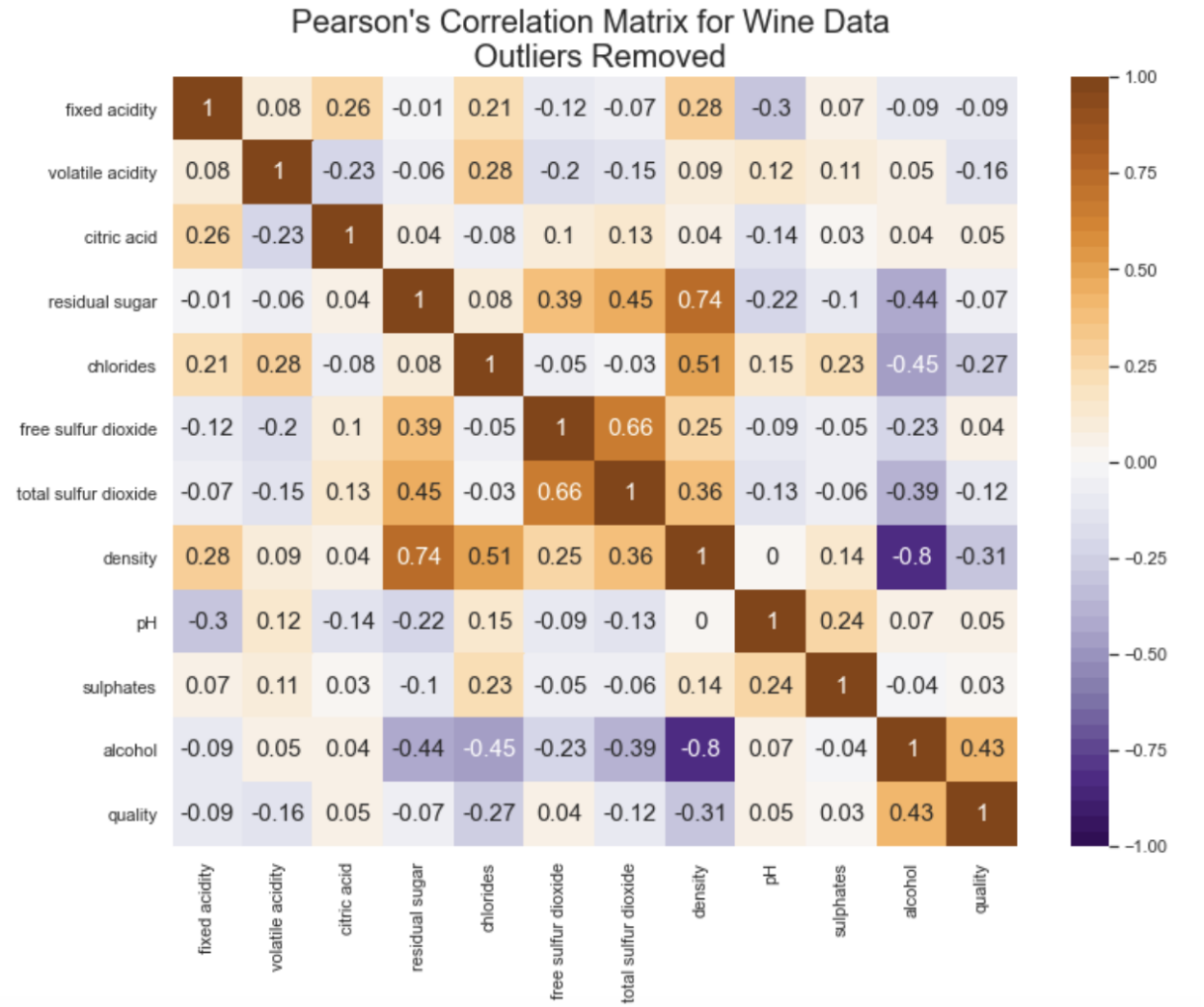

Pearson's Correlation Matrix for Wine Data
Before Outliers Were Removed
fixed acidity 1 0.22 0.32 -0.11 0.3 -0.28 -0.33 0.46 -0.25 0.3 -0.1 -0.08
volatile acidity 0.22 1 -0.38 -0.2 0.38 -0.35 -0.41 0.27 0.26 0.23 -0.04 -0.27
citric acid 0.32 -0.38 1 0.14 0.04 0.13 0.2 0.1 -0.33 0.06 -0.01 0.09
residual 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
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

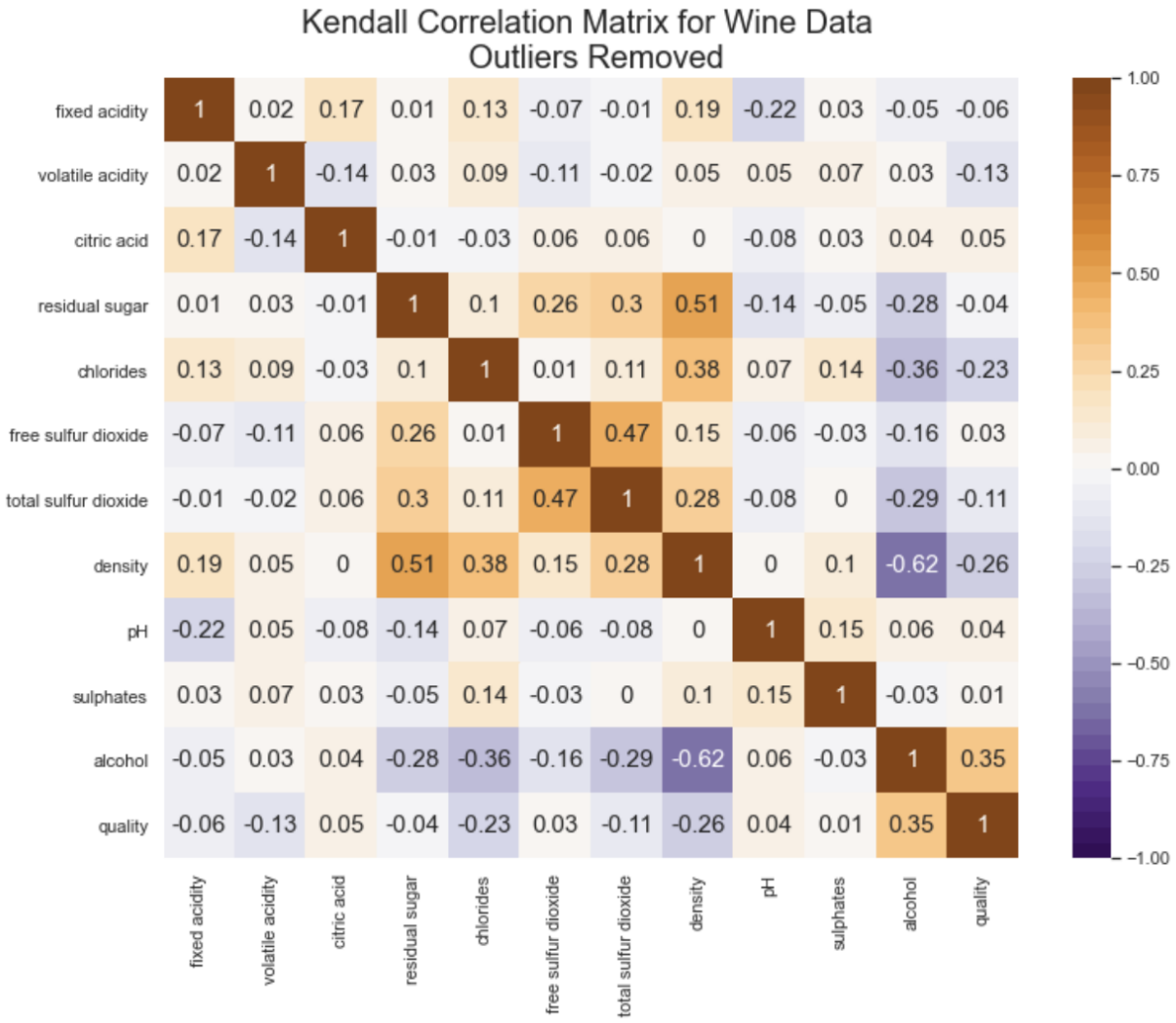

Kendall Correlation Matrix for Wine Data
Before Outliers Were Removed
fixed acidity 1 0.13 0.19 -0.02 0.25 -0.18 -0.15 0.3 -0.17 0.15 -0.07 -0.08
volatile acidity 0.13 1 -0.2 -0.05 0.28 -0.25 -0.22 0.17 0.14 0.17 -0.01 -0.2
citric acid 0.19 -0.2 1 0.05 -0.05 0.09 0.11 0.04 -0.2 0.02 0.01 0.08
residual 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
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Summary of Correlation Results for Wine Data (Before Outliers Removed)			
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 (Before Outliers Removed)			
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)


Spearman Correlation Matrix for Wine Data
Outliers Removed
fixed acidity 1 0.03 0.25 0.01 0.19 -0.1 -0.01 0.28 -0.31 0.05 -0.08 -0.07
volatile acidity 0.03 1 -0.2 0.05 0.13 -0.17 -0.04 0.08 0.08 0.1 0.05 -0.17
citric acid 0.25 -0.2 1 -0.01 -0.05 0.08 0.1 0 -0.12 0.04 0.06 0.06
residual sugar 0.01 0.05 -0.01 1 0.14 0.38 0.45 0.69 -0.19 -0.07 -0.41 -0.05
chlorides 0.19 0.13 -0.05 0.14 1 0.02 0.15 0.55 0.1 0.21 -0.51 -0.3
free sulfur dioxide -0.1 -0.17 0.08 0.38 0.02 1 0.65 0.23 -0.09 -0.04 -0.23 0.04
total sulfur dioxide -0.01 -0.04 0.1 0.45 0.15 0.65 1 0.41 -0.12 -0 -0.42 -0.15
density 0.28 0.08 0 0.69 0.55 0.23 0.41 1 0 0.15 -0.81 -0.34
pH -0.31 0.08 -0.12 -0.19 0.1 -0.09 -0.12 0 1 0.22 0.08 0.05
sulphates 0.05 0.1 0.04 -0.07 0.21 -0.04 -0 0.15 0.22 1 -0.05 0.01
alcohol -0.08 0.05 0.06 -0.41 -0.51 -0.23 -0.42 -0.81 0.08 -0.05 1 0.44
quality -0.07 -0.17 0.06 -0.05 -0.3 0.04 -0.15 -0.34 0.05 0.01 0.44 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

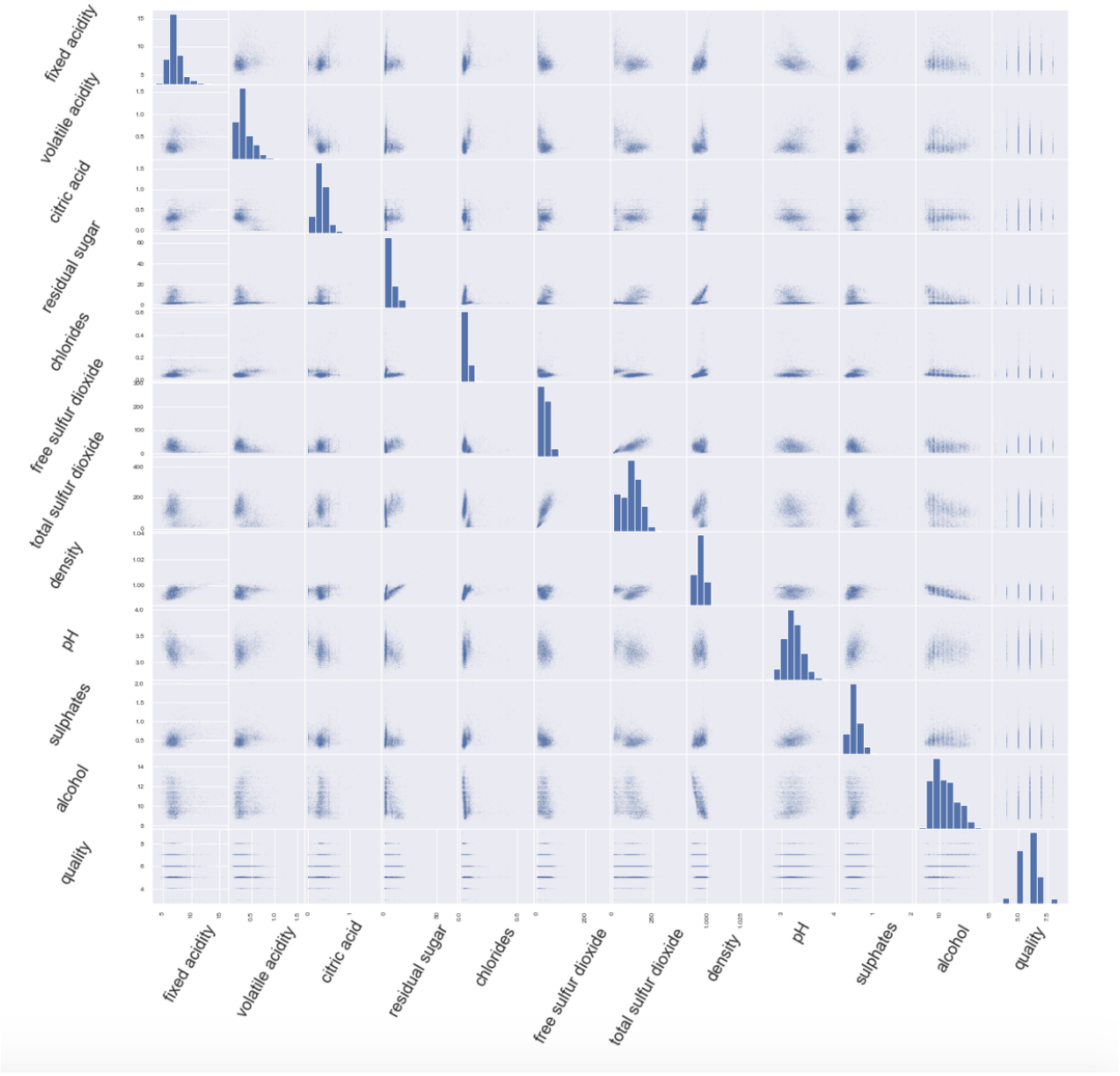

Pearson's Correlation Matrix for Wine Data
Outliers Removed
fixed acidity 1 0.08 0.26 -0.01 0.21 -0.12 -0.07 0.28 -0.3 0.07 -0.09 -0.09
volatile acidity 0.08 1 -0.23 -0.06 0.28 -0.2 -0.15 0.09 0.12 0.11 0.05 -0.16
citric acid 0.26 -0.23 1 0.04 -0.08 0.1 0.13 0.04 -0.14 0.03 0.04 0.05
residual sugar -0.01 -0.06 0.04 1 0.08 0.39 0.45 0.74 -0.22 -0.1 -0.44 -0.07
chlorides 0.21 0.28 -0.08 0.08 1 -0.05 -0.03 0.51 0.15 0.23 -0.45 -0.27
free sulfur dioxide -0.12 -0.2 0.1 0.39 -0.05 1 0.66 0.25 -0.09 -0.05 -0.23 0.04
total sulfur dioxide -0.07 -0.15 0.13 0.45 -0.03 0.66 1 0.36 -0.13 -0.06 -0.39 -0.12
density 0.28 0.09 0.04 0.74 0.51 0.25 0.36 1 0 0.14 -0.8 -0.31
pH -0.3 0.12 -0.14 -0.22 0.15 -0.09 -0.13 0 1 0.24 0.07 0.05
sulphates 0.07 0.11 0.03 -0.1 0.23 -0.05 -0.06 0.14 0.24 1 -0.04 0.03
alcohol -0.09 0.05 0.04 -0.44 -0.45 -0.23 -0.39 -0.8 0.07 -0.04 1 0.43
quality -0.09 -0.16 0.05 -0.07 -0.27 0.04 -0.12 -0.31 0.05 0.03 0.43 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Kendall Correlation Matrix for Wine Data
Outliers Removed
fixed acidity 1 0.02 0.17 0.01 0.13 -0.07 -0.01 0.19 -0.22 0.03 -0.05 -0.06
volatile acidity 0.02 1 -0.14 0.03 0.09 -0.11 -0.02 0.05 0.05 0.07 0.03 -0.13
citric acid 0.17 -0.14 1 -0.01 -0.03 0.06 0.06 0 -0.08 0.03 0.04 0.05
residual sugar 0.01 0.03 -0.01 1 0.1 0.26 0.3 0.51 -0.14 -0.05 -0.28 -0.04
chlorides 0.13 0.09 -0.03 0.1 1 0.01 0.11 0.38 0.07 0.14 -0.36 -0.23
free sulfur dioxide -0.07 -0.11 0.06 0.26 0.01 1 0.47 0.15 -0.06 -0.03 -0.16 0.03
total sulfur dioxide -0.01 -0.02 0.06 0.3 0.11 0.47 1 0.28 -0.08 0 -0.29 -0.11
density 0.19 0.05 0 0.51 0.38 0.15 0.28 1 0 0.1 -0.62 -0.26
pH -0.22 0.05 -0.08 -0.14 0.07 -0.06 -0.08 0 1 0.15 0.06 0.04
sulphates 0.03 0.07 0.03 -0.05 0.14 -0.03 0 0.1 0.15 1 -0.03 0.01
alcohol -0.05 0.03 0.04 -0.28 -0.36 -0.16 -0.29 -0.62 0.06 -0.03 1 0.35
quality -0.06 -0.13 0.05 -0.04 -0.23 0.03 -0.11 -0.26 0.04 0.01 0.35 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

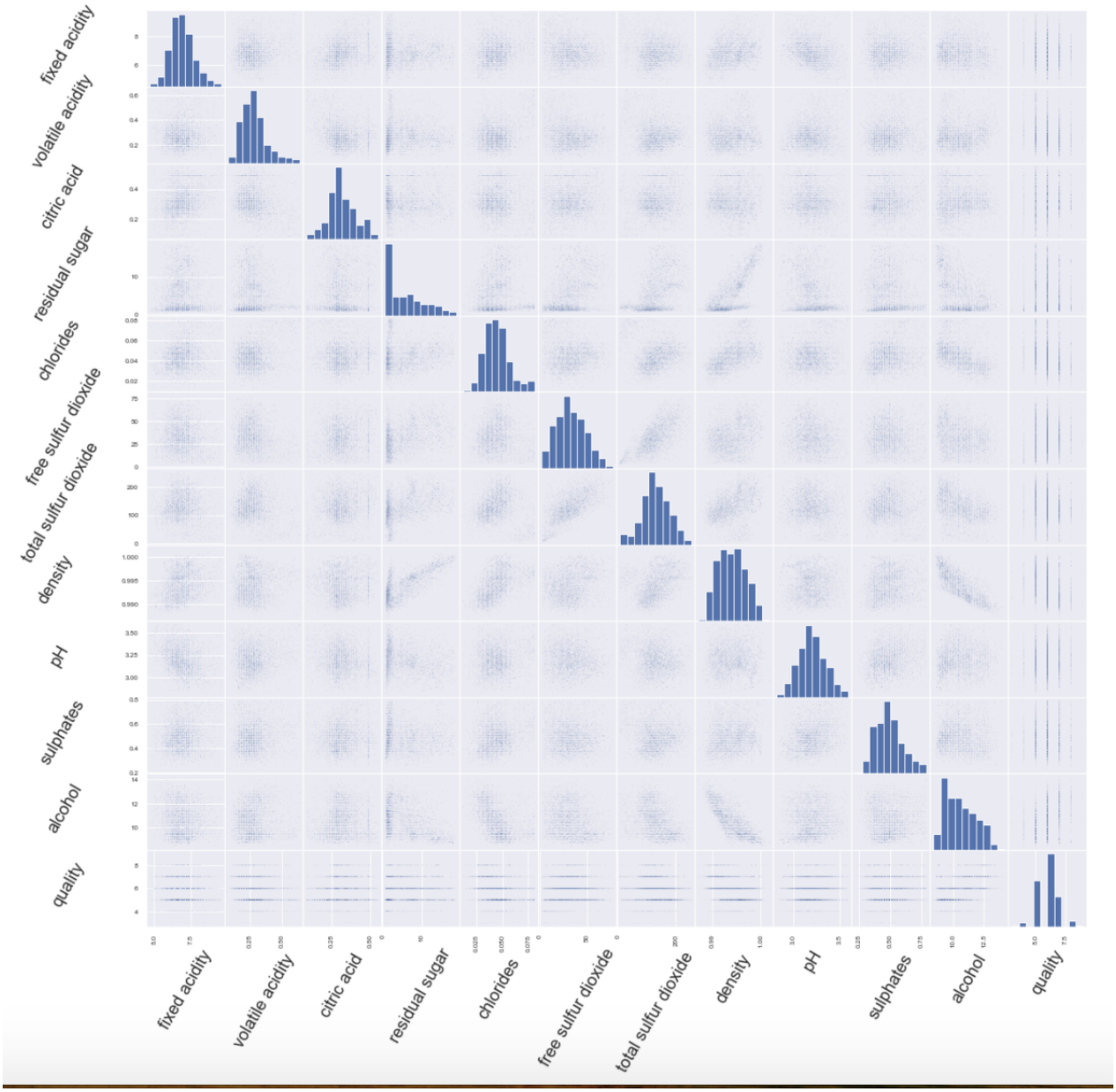
Summary of Correlation Results for Wine Data (Outliers Removed)			
Attribute	Spearman	Pearson’s	Kendall
Alcohol	0.438278	0.432710	0.345778
Density	-0.339716	-0.310473	-0.260015
Chlorides	-0.300207	-0.267661	-0.234052
Volatile Acidity	-0.170554	-0.163855	-0.133324
Citric Acid	0.063361	0.048409	0.049765
Fixed Acidity	-0.073996	-0.087174	-0.057232
Free Sulfur Dioxide	0.042364	0.041818	0.032023
Total Sulfur Dioxide	-0.147229	-0.119894	-0.112757
pH	0.049999	0.049083	0.038388
Sulphates	0.014279	0.030125	0.011683
Residual Sugar	-0.051808	-0.073186	-0.039669

Correlation Ranking for Wine Data (Outliers Removed)			
Correlation Ranking	Spearman	Pearson's	Kendall
1	Alcohol (0.438278)	Alcohol (0.432710)	Alcohol (0.345778)
2	Density (-0.339716)	Density (-0.310473)	Density (-0.260015)
3	Chlorides (-0.300207)	Chlorides (-0.267661)	Chlorides (-0.234052)
4	Volatile Acidity (-0.170554)	Volatile Acidity (-0.163855)	Volatile Acidity (-0.133324)
5	Total Sulfur Dioxide (-0.147229)	Total Sulfur Dioxide (-0.119894)	Total Sulfur Dioxide (-0.112757)
6	Fixed Acidity (-0.073996)	Fixed Acidity (-0.087174)	Fixed Acidity (-0.057232)
7	Citric Acid (0.063361)	Residual Sugar (-0.073186)	Citric Acid (0.049765)
8	Residual Sugar (-0.051808)	pH (0.049083)	Residual Sugar (-0.039669)
9	pH (0.049999)	Citric Acid (0.048409)	pH (0.038388)
10	Free Sulfur Dioxide (0.042364)	Free Sulfur Dioxide (0.041818)	Free Sulfur Dioxide (0.032023)
11	Sulphates (0.014279)	Sulphates (0.030125)	Sulphates (0.011683)

Wine Scatter Matrix (Before Outliers Removed)



Wine Scatter Matrix (Outliers Removed)



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