Assignment

Exploratory Data Analysis of Wine Quality Dataset

Prepared for:
Dr. Rumana Rois
Course: PMASDS04
Introduction to Data Science with Python

Prepared by:

Name: A. M. Abeid Hassan
ID # 016
Section-B
Batch - 02
PM-ASDS Program

Date of Submission: 13-12-2019

Table of Contents

1.1 Data Description

This dataset contains **white wine samples** of the Portuguese “Vinho Verde” wine. It is part of the two datasets created, using red and white wine samples, in an effort to model human taste preferences based on physicochemical properties of the wine. It contains **4,898 observations** concerning different attributes (citric acid, chlorides, density, quality, etc.) of the “VinhoVerde” white wine. It has been obtained from the UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

The original dataset contains 12 columns. For time constraint, only **6 variables** were selected for analysis including the output variable, quality. The selected variables have been listed below. The level of measurement, appropriate measures, and suitable plots along with the description of the variables are summarized in table-1.

Variable Name	**Variable Description**	**Value level**	**Level of Measurements**	**Appropriate measures (of central tendency)**	**Appropriate measures (of dispersion)**	**Appropriate plots**
citric_acid	Amount of citric acid (g / dm^3)		Ratio	Mean, Median, Mode	Standard Deviation	Boxplot, Histogram
chlorides	Amount of salt, sodium chloride, in wine (g / dm^3)		Ratio	Mean, Median, Mode	Standard Deviation	Boxplot, Histogram
density	Density of wine (g / cm^3)		Ratio	Mean, Median, Mode	Standard Deviation	Boxplot, Histogram
pH	Describes how acidic or basic a wine is on a pH scale from 0 to 14.	0 = very acidic 14 = very basic	Interval	Mean, Median, Mode	Standard Deviation	Boxplot, Histogram
alcohol	It represents the percent alcohol content of the wine		Ratio	Mean, Median, Mode	Standard Deviation	Boxplot, Histogram
quality (output variable)	It represents the quality of wine graded by experts on a scale of 0 to 10	For our analysis we have considered: 9-10 = excellent 8 = very good 6-7 = good 5 = moderate 3-4 = bad 1-2 = very bad	Ordinal	Median, Mode		Pie chart, Bar diagram

Table 1: Variables summary information of White Wine Quality Dataset

1.2 Objectives

The main objective is to perform **Exploratory Data Analysis** for initial investigations on the data. To that extent, we would like to discover patterns and spot anomalies with the help of **summary statistics** and **graphical representations**. Quality is the output variable of our dataset, as it has to be predicted using the remaining (input) variables. Our objectives could be laid down as follows:

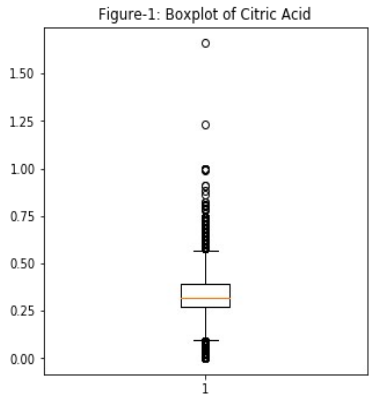
- To perform data cleaning by identifying missing values and outliers.
- To specify the distribution of the input variables.
- To reveal the different proportion of Quality grades (very bad to excellent) of white wine samples
- To calculate different summary measures of the variables.
- To find the association between variables.

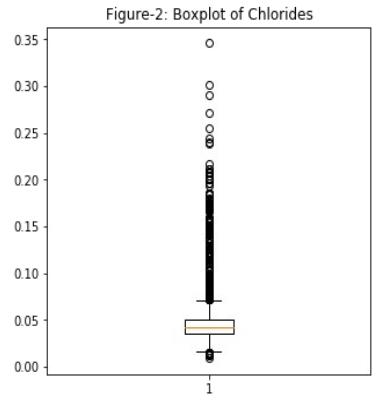
1.3 Data Cleaning

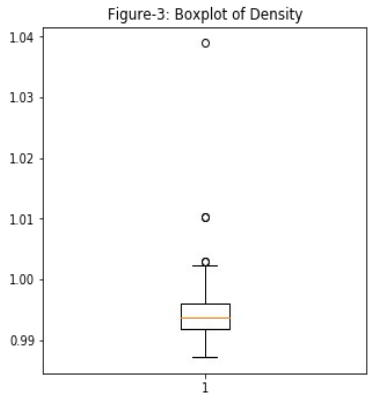
We want to identify any **missing values** and **outliers** in the dataset. Using .info() we found that there were no values missing in the white wine dataset. The output is as follows:

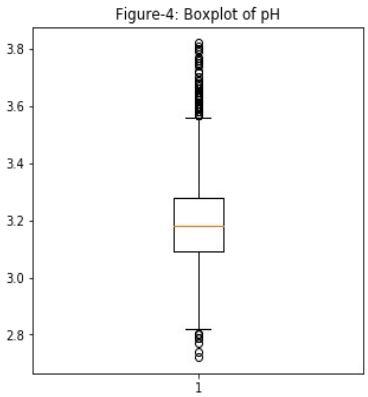
citric_acid 4898 non-null float64
chlorides 4898 non-null float64
density 4898 non-null float64
pH 4898 non-null float64
alcohol 4898 non-null float64
quality 4898 non-null int64

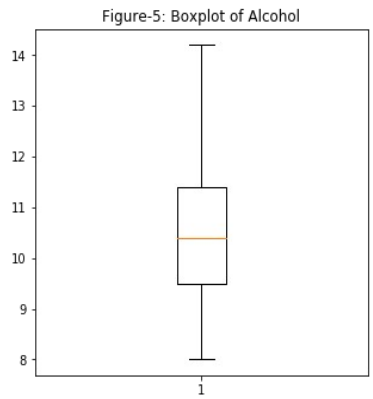
From the boxplots below we can easily identify the outliers. We can conclude that Alcohol has no outliers. Citric Acid, Chlorides, pH, and Density does have some outliers.











1.4 Univariate Analysis:

As the physicochemical (input) variables are quantitative, we could represent our data by plotting histograms and boxplots. The output variable, quality, is Ordinal. Hence, we could make a pie chart or a bar diagram for the ‘quality’ variable.

	citric_acid	chlorides	density	pH	alcohol	quality
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	0.334192	0.045772	0.994027	3.188267	10.514267	5.877909
std	0.121020	0.021848	0.002991	0.151001	1.230621	0.885639
min	0.000000	0.009000	0.987110	2.720000	8.000000	3.000000
25%	0.270000	0.036000	0.991723	3.090000	9.500000	5.000000
50%	0.320000	0.043000	0.993740	3.180000	10.400000	6.000000
75%	0.390000	0.050000	0.996100	3.280000	11.400000	6.000000
max	1.660000	0.346000	1.038980	3.820000	14.200000	9.000000

Table 2: Descriptive Statistics

1.4.1 Graphical Representations:

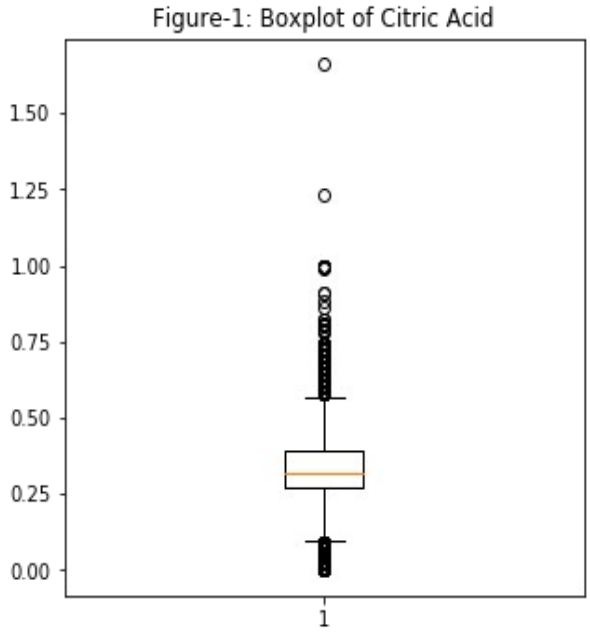
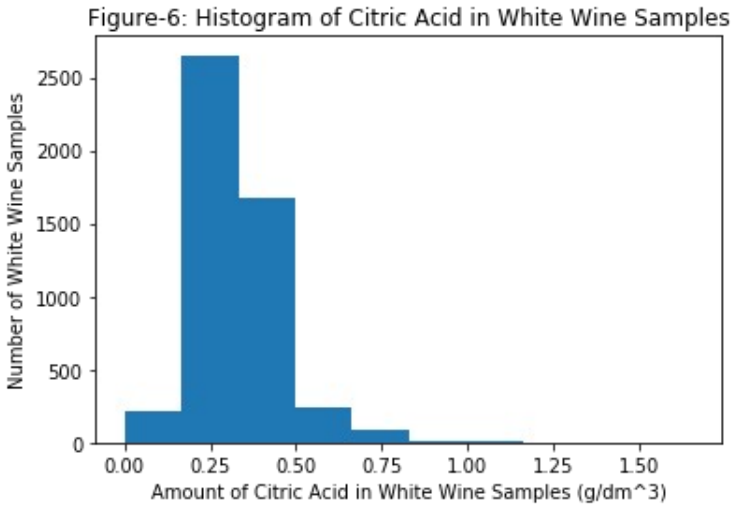


Figure-1 shows that the middle 50% of white wine samples contain Citric Acid between 0.27 g/dm^3 and 0.39 g/dm^3. The distribution of the citric acid is symmetric because the length of the whisker above 0.39 g/dm^3 is about the same length as the whisker below 0.27 g/dm^3. Also the area in the box between 0.27 g/dm^3 and the median of 0.32 is about the same as the area between the median and 0.39 g/dm^3. The histogram in Figure-6 shows a positively skewed distribution due to the presence of the outliers.



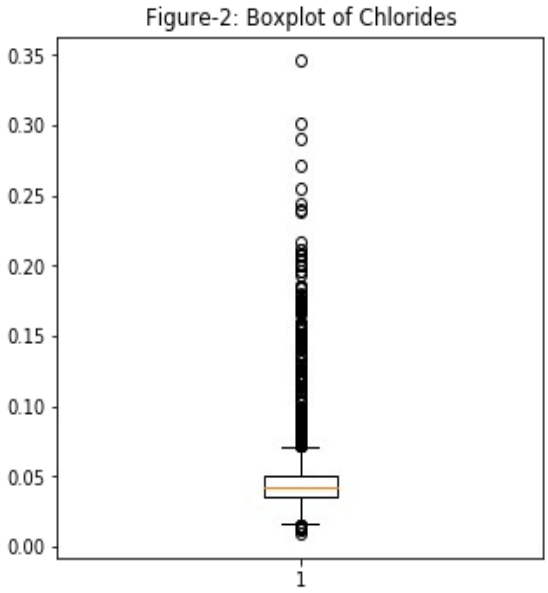
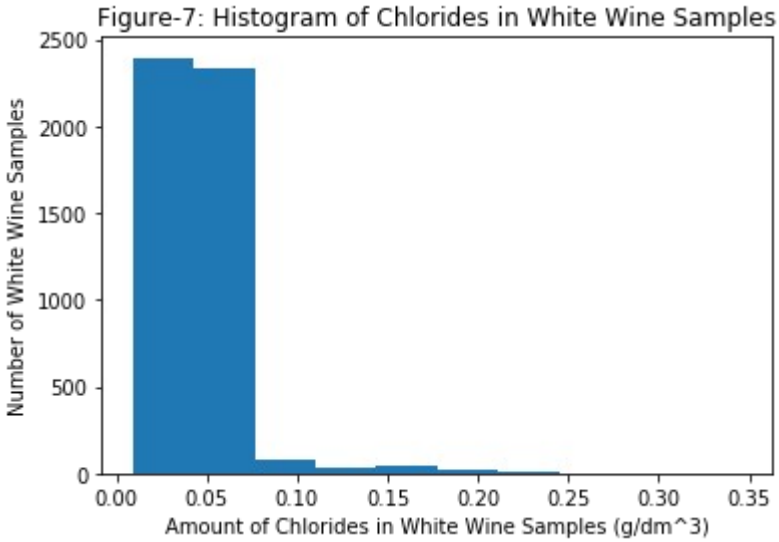
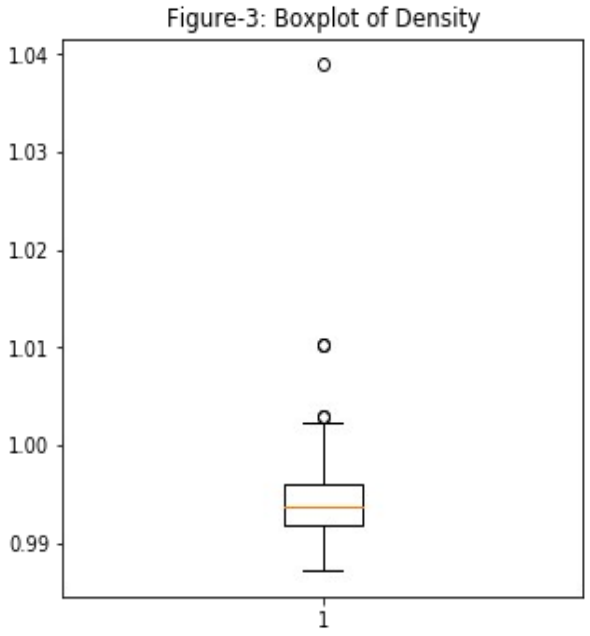
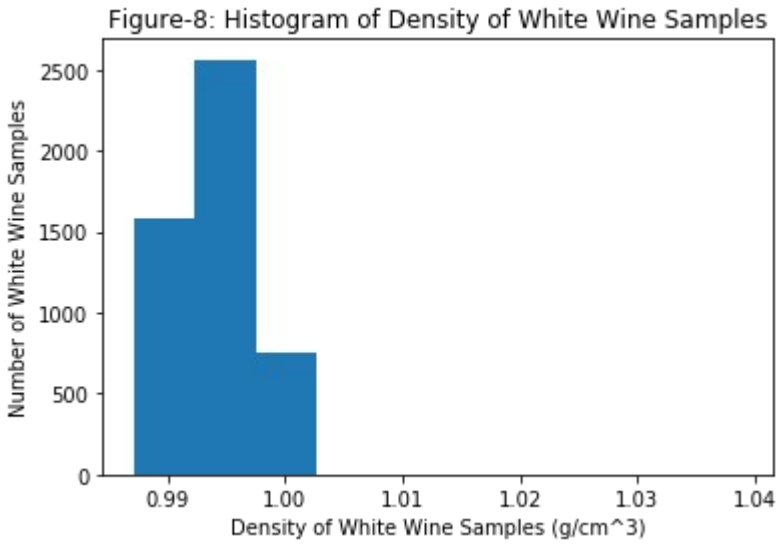


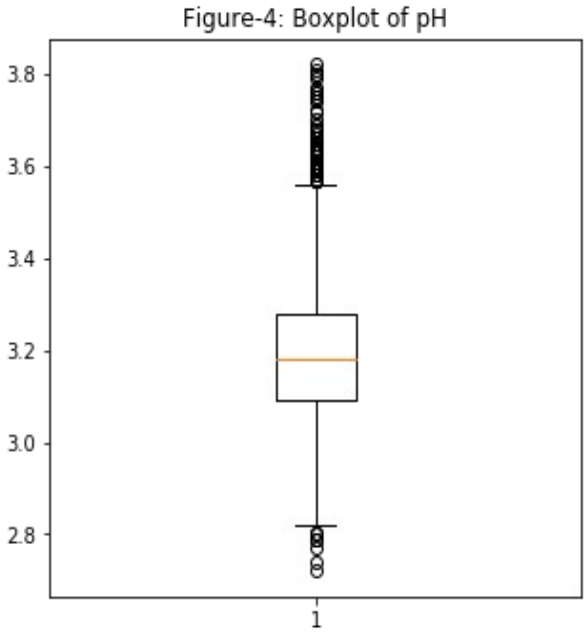
Figure-2 explores that Chlorides has a symmetric distribution with many outliers. It is observed in Figure-7 that majority of white wine samples contain salt between 0.009 g/dm^3 and 0.075 g/dm^3.



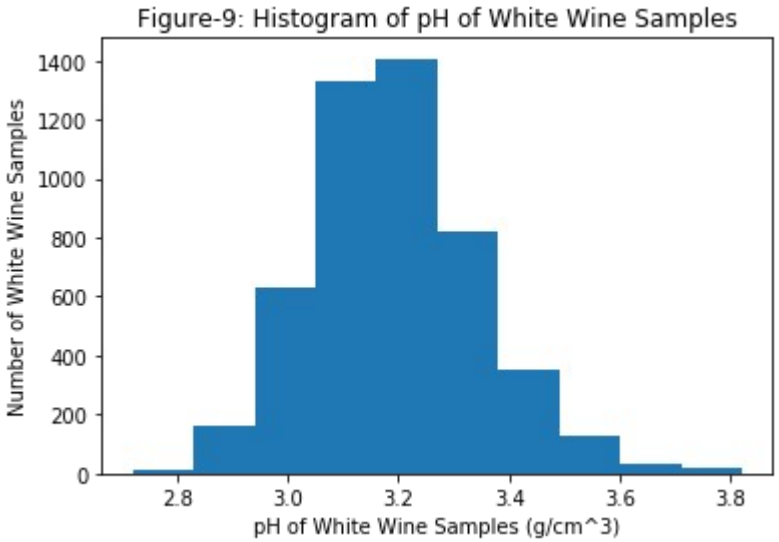


The boxplot in Figure-3 reveals that 50% of the wine samples have a Density between 0.993 g/cm^3 and 0.996 g/cm^3. The median is in the middle of the box, from which we can say that Density follows a symmetric distribution. The shape of the distribution in Figure-8 is positively skewed because of the presence of outliers above the maximum value.





It can be concluded from Figure-4 that 50% of the wine samples has a pH value between 3.09 and 3.28, and the distribution of pH is symmetric. Figure-9 confirms that.



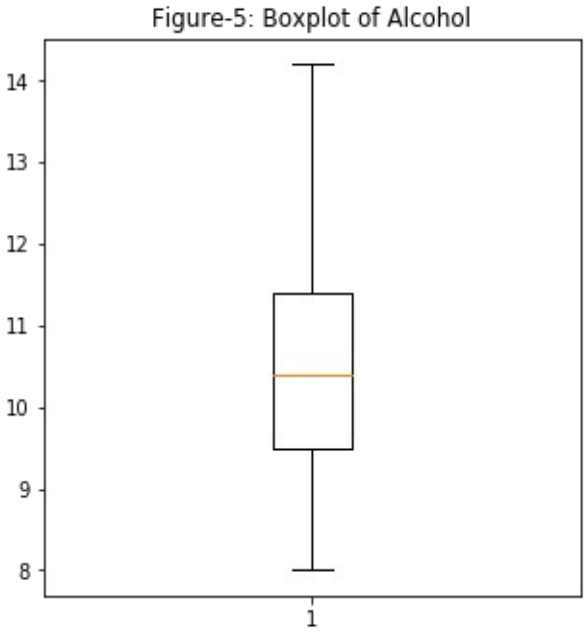


Figure-5 represents the boxplot of Alcohol content of white wine samples, which suggest that the middle 50% of wine samples contain alcohol between 9.5% and 11.4%. The overall distribution of wine samples is positively skewed. There are no outliers. Similar information about the shape of the distribution is found in Figure-10.

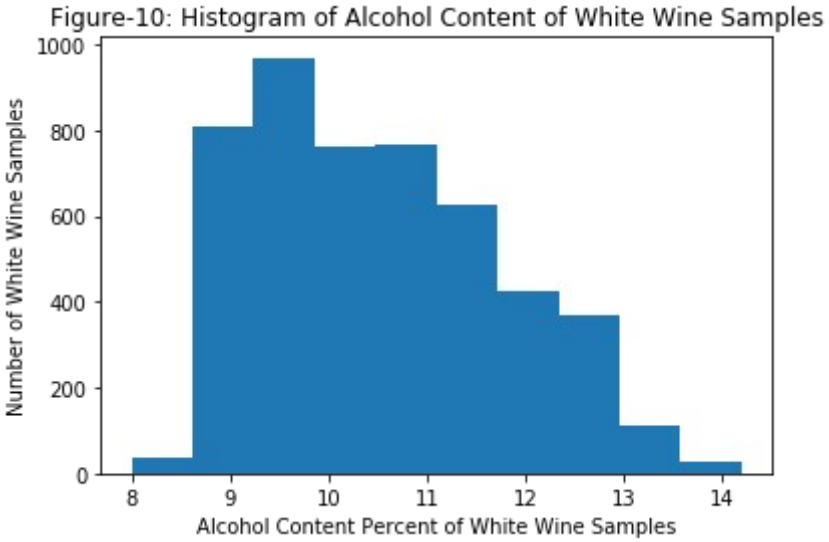


Figure-11: Pie Chart of Quality of White Wine Samples

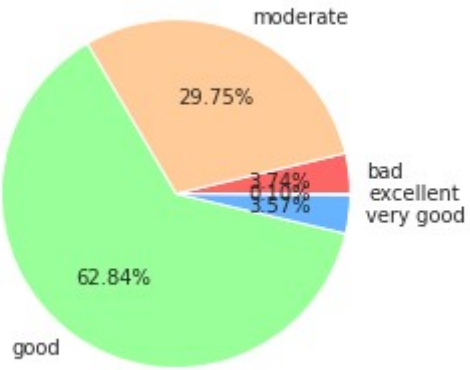
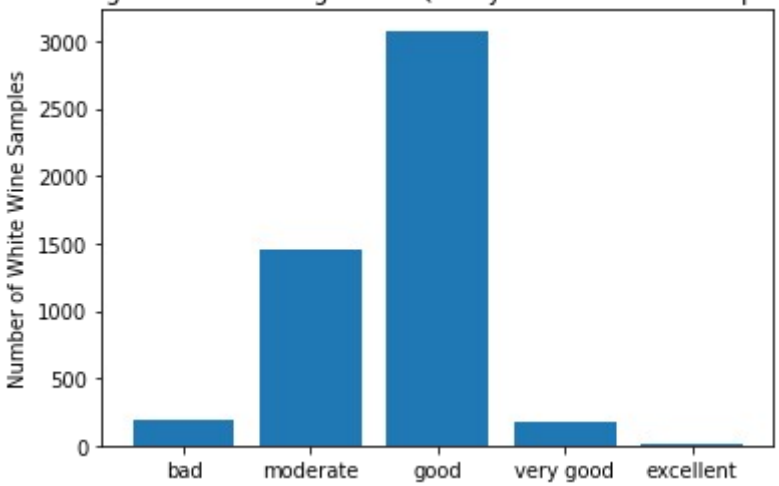


Figure-11 illustrates Pie Chart of the Quality of white wine samples. Majority of the observations, 62.84% are wines with “Good” grade. 29.75% of the samples are of “Moderate” grade, 3.57% are rated as very good and 3.74% are are rated as bad. The dataset consists of only 0.10% “Excellent” white wine samples.

Figure 12 - Bar Diagram of Quality of White Wine Samples



1.4.2 Summary Measures:

The table below shows the the mean, median, mode, and standard deviation of the variables in the dataset.

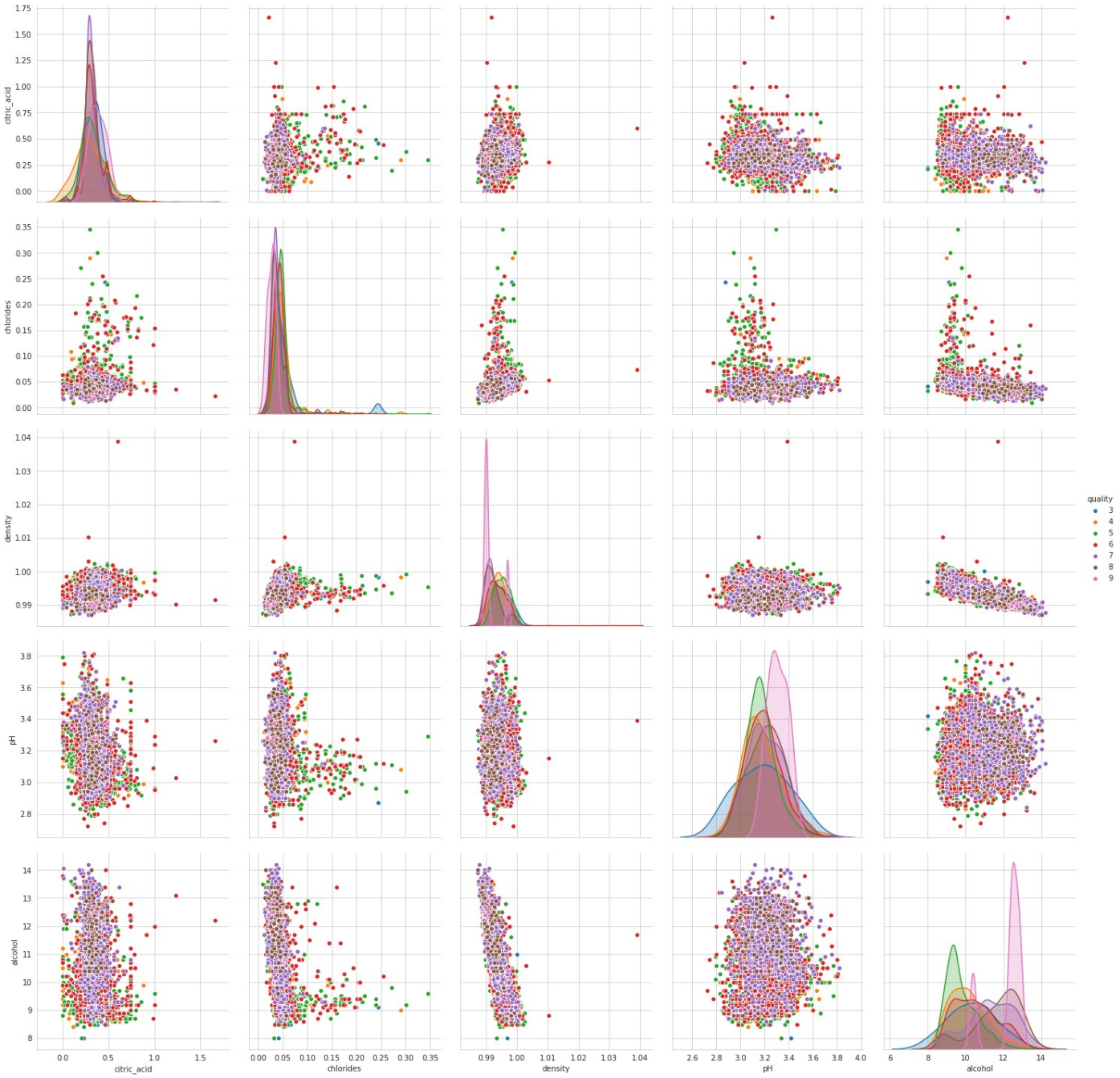
	Mean	**Median**	**Mode**	**Standard Deviation**
alcohol	10.51	10.40	9.40	1.23
chlorides	0.05	0.04	0.04	0.02
citric_acid	0.33	0.32	0.30	0.12
density	0.99	0.99	0.99	0.00
pH	3.19	3.18	3.14	0.15
quality	NaN	6.00	6.00	NaN

Table 3: Summary Measures of the Variables of White Wine Quality Dataset

1.5 Bivariate Analysis:

1.5.1 Graphical Representations:

The Pairplot diagram summarizes scatter diagrams between all the input variables, Citric Acid, Chlorides, Density, pH and Alcohol Content of White Wine Samples at Quality. From Figure-14, it can interpreted that Excellent wines (in light pink) contain the highest amount of alcohol and pH-value as shown by the peak of the histograms. They are also the most dense among all the other grades.



Figure–14: Pairplots of Citric Acid, Chlorides, Density, pH and Alcohol Content of White Wine Samples at Quality

1.5.2 Summary Measures:

Pearson's Correlation coefficients show the magnitude of correlation between the input variables in the correlation matrix below. Chlorides and density have a weak positive correlation with a correlation coefficient of 0.26. Alcohol and density have a strong negative correlation with a correlation coefficient of -0.78.

	citric_acid	**chlorides**	**density**	**pH**	**alcohol**
citric_acid	1.00	0.11	0.15	-0.16	-0.08
chlorides	0.11	1.00	0.26	-0.09	-0.36
density	0.15	0.26	1.00	-0.09	-0.78
pH	-0.16	-0.09	-0.09	1.00	0.12
alcohol	-0.08	-0.36	-0.78	0.12	1.00

Table 4: Correlation Matrix of different variables of White Wine Quality Dataset

1.6 Appendix

1.6.1 Python codes for EDA (written in jupyter notebook):

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

# read csv data file
wine = pd.read_csv("/home/abeid/Documents/pm-asds/04-intro-to-data-science-with-python/assignment-1/
winequality-white.csv")

# drop column ‘quality’
wine_new = wine.drop(['quality'], axis=1)


## Data Overview (without output variable)
wine_new.head()


## Data Overview (with output variable)
wine.head()

# data size
wine.shape

# 4898 rows and 12 columns


## Data Cleaning

# dataframe of selected variables
# wine_new2 = wine.drop(['fixed_acidity','volatile_acidity','residual_sugar','free_sulfur_dioxide',
'total_sulfur_dioxide', 'sulphates'], axis=1)
# wine_new2

#1 missing values
wine_new2.info()

# no missing values


## Descriptive Statistics
wine_new2.describe()


## Univariate Analysis

# boxplots

# figure 1 - boxplot of citric_acid
```

```
plt.figure(figsize=(5,5))
plt.boxplot(wine_new.citric_acid)
plt.title('Figure-1: Boxplot of Citric Acid')
plt.show()


# figure 2 - box plot of chlorides
plt.figure(figsize=(5,5))
plt.boxplot(wine_new.chlorides)
plt.title('Figure-2: Boxplot of Chlorides')
plt.show()

# figure 3 - box plot of density
plt.figure(figsize=(5,5))
plt.boxplot(wine_new.density)
plt.title('Figure-3: Boxplot of Density')
plt.show()

# figure 4 - box plot of pH
plt.figure(figsize=(5,5))
plt.boxplot(wine_new.pH)
plt.title('Figure-4: Boxplot of pH')
plt.show()

# figure 5 - box plot of alcohol
plt.figure(figsize=(5,5))
plt.boxplot(wine_new.alcohol)
plt.title('Figure-2: Boxplot of Alcohol')
plt.show()

# figure 6 - histogram of citric_acid
plt.hist(wine_new.citric_acid, bins=10)
plt.xlabel('Amount of Citric Acid in White Wine Samples (g/dm^3)')
plt.ylabel('Number of White Wine Samples')
plt.title('Figure-6: Histogram of Citric Acid in White Wine Samples')
plt.show()

# figure 7 - histogram of chlorides
plt.hist(wine_new.chlorides, bins=10)
plt.xlabel('Amount of Chlorides in White Wine Samples (g/dm^3)')
plt.ylabel('Number of White Wine Samples')
plt.title('Figure-7: Histogram of Chlorides in White Wine Samples')
plt.show()

# figure 8 - histogram of density
plt.hist(wine_new.density, bins=10)
plt.xlabel('Density of White Wine Samples (g/cm^3)')
plt.ylabel('Number of White Wine Samples')
plt.title('Figure-8: Histogram of Density of White Wine Samples')
plt.show()

# figure 9 - histogram of pH
plt.hist(wine_new.pH, bins=10)
```

```
plt.xlabel('pH of White Wine Samples (g/cm^3)')
plt.ylabel('Number of White Wine Samples')
plt.title('Figure-9: Histogram of pH of White Wine Samples')
plt.show()

# figure 10 - histogram of alcohol
plt.hist(wine_new.alcohol, bins=10)
plt.xlabel('Alcohol Content Percent of White Wine Samples')
plt.ylabel('Number of White Wine Samples')
plt.title('Figure-10: Histogram of Alcohol Content of White Wine Samples')
plt.show()

# figure 11 - pie chart of quality

# frequency of quality
# wine['quality'].value_counts().sort_index()

quality_fre = [183, 1457, 3078, 175, 5]
labels = ['bad', 'moderate', 'good', 'very good', 'excellent']
colors = ['#ff6666', '#ffcc99', '#99ff99', '#66b3ff', '#c2c2f0']

plt.pie(quality_fre,labels=labels, colors=colors, autopct='%.2f%%', radius = 1)
plt.title('Figure-11: Pie Chart of Quality of White Wine Samples')
plt.show()

# figure 12 - bar diagram of quality

objects = ('bad', 'moderate', 'good', 'very good', 'excellent')
x_pos = np.arange(len(objects))
quality_fre = [183, 1457, 3078, 175, 5]

plt.bar(x_pos, quality_fre)
plt.xticks(x_pos, objects)
plt.ylabel('Number of White Wine Samples')
plt.title('Figure 12 - Bar Diagram of Quality of White Wine Samples')
plt.show()


# summary statistics

# mean
mean = wine_new.mean() # mean dataframe
mean_sv = mean.loc[['citric_acid', 'chlorides', 'density', 'pH', 'alcohol']] # mean of selected variables
mean1 = mean_sv.to_frame(name='Mean')
mean1 # mean1 dataframe

# median
median = wine.median() # median dataframe
median_sv = median.loc[['citric_acid', 'chlorides', 'density', 'pH', 'alcohol', 'quality']] # median of selected
variables
median1 = median_sv.to_frame(name='Median')
median1 # median1 dataframe
```

```
# mode
mode = wine.mode() # mode series
mode_t = mode.transpose() # transpose of mode dataframe
mode_sv = mode_t.loc[['citric_acid', 'chlorides', 'density', 'pH', 'alcohol', 'quality']] # mode of selected
variables
mode1 = mode_sv.rename(columns = {0: "Mode"})
mode1 # mode1 dataframe

# standard deviation
std_dev = wine.std() # std_dev dataframe
std_dev_sv = std_dev.loc[['citric_acid', 'chlorides', 'density', 'pH', 'alcohol']] # std_dev of selected variables
std_dev1 = std_dev_sv.to_frame(name='Standard Deviation' )
std_dev1

# data frame of summary statistics

# merging mean1, median1, mode1, and std_dev dataframes
merge1 = mean1.join(median1, how='outer')
merge2 = merge1.join(mode1, how='outer')
summary_stat = merge2.join(std_dev1, how='outer')
summary_stat.round(decimals=2)


## Bi-variate Analysis

# pairplots

# dataframe of selected variables
wine_new2 = wine.drop(['fixed_acidity','volatile_acidity','residual_sugar','free_sulfur_dioxide',
'total_sulfur_dioxide', 'sulphates'], axis=1)
# wine_new2

sns.set_style('whitegrid');
sns.pairplot(wine_new2, vars= ['citric_acid', 'chlorides', 'density', 'pH', 'alcohol'], hue = 'quality', height = 4)
# plt.title('Figure-13: Pairplots of Citric Acid, Chlorides, Density, pH and Alcohol Content of White Wine
Samples')
plt.show()

# how to map the quality scale to grades?

# summary measures

# Correlation Matrix
corr_matrix = wine_new2.corr()
corr_matrix.round(decimals=2)
```