K. J. Somaiya College of Engineering,
Mumbai-77
(A Constituent College of Somaiya Vidyavihar University)
Department of Computer Engineering)

Batch:D2 **Roll No.:16010122323**

Experiment No. 4

Title : Exploratory data analysis

Aim: Use R tool to implement Exploratory data analysis for a given case study.

Expected Outcome of Experiment:
CO3: Explain the significance of exploratory data analysis (EDA) in data science
CO5: Apply basic tools to carry out EDA for the Data Science process.

Books/ Journals/ Websites referred:
1. Data Mining Concepts and Techniques Jiawei Han, Michelin Kamber, Jian Pie, 3rd edition

---

What is Exploratory Data Analysis (EDA)?
Exploratory data analysis (EDA) is the process of analysing data to uncover their key features. EDA refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis is used to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed

Main reasons for using exploratory data analysis

1. Detection of mistakes
2. Checking for assumptions
3. Preliminary selection of appropriate models
4. Determining relationships among the explanatory variables, and
5. Assessing the direction and rough size of relationships between explanatory and outcome variables

Approach to detecting outlier in the attribute using boxplot approach
Outliers are either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile.

4 types of EDA

- Univariate non-graphical
- Multivariate non-graphical
- Univariate graphical
- Multivariate graphical

Univariate non-graphical EDA

A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

For numerical data find the measure of central tendency spread, skewness, kurtosis

Univariate graphical EDA

With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers. Quantile Normal plots allow detection of non-normality and diagnosis of skewness and kurtosis.

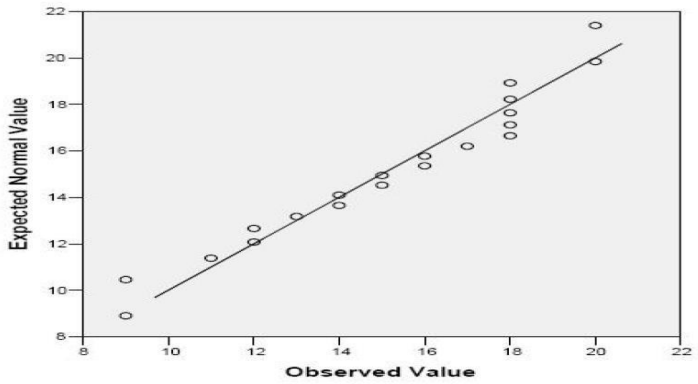


Figure 4.10: Quantile-normal plot with ties.

Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

Chi-square test, Pearson product moment coefficient, covariance, Spearman rank correlation coefficient can be used to measure the correlation between the attributes

Correlation between the nominal attributes-

X^2 (chi-square) test

Test to check whether the nominal attributes are independent or correlated

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality

Eg:

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
- It shows that like_science_fiction and play_chess are correlated in the group

Pearson's product moment coefficient

It is used to find the correlation between numerical attributes.

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A \sigma_B}$$

- Correlation coefficient (also called Pearson's product moment coefficient) where n is the number of tuples, and are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.
  - If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
  - $r_{A,B} = 0$: independent;
  - $r_{AB} < 0$: negatively correlated

Covariance

Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

Where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:

Spearman's Rank Correlation Co-efficient

To check the correlation between ordinal attributes.
The formula for Spearman's Rank Correlation Co-efficient is:

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where: d = difference between the ranking of each item
n = the number of paired observations

Spearman's Rank Correlation Co-efficient can take any value between +1 and -1.
+1 = Perfect positive correlation
0 = No correlation
-1 = Perfect negative correlation

Example- There are 2 judges, ranking the participants in dance competition. We used Spearman's rank correlation co-efficient to check whether ranks given by 2 judges are positively or negatively correlated**.**

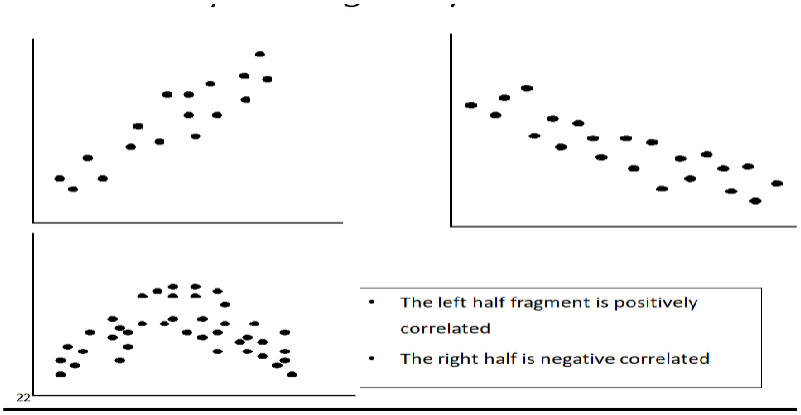
Multi-variate non-graphical EDA

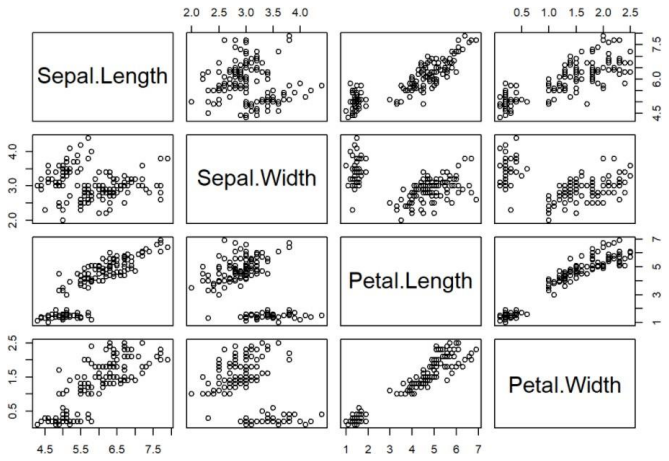
Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.

For two quantitative variables, the basic graphical EDA technique is the scatterplot which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. If one variable is explanatory and the other is outcome, it is a very, very strong convention to put the outcome on the y (vertical) axis.
Scatter plots can be extended to *n* attributes, resulting in a *scatter-plot matrix.*

Scatter Plot and scatter plot matrix

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. It is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships. Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure below shows examples of positive and negative correlations between two attributes.





Procedure for Implementation in lab:

Identify a Dataset for Exploratory data analysis

1. Get the insight into the dataset. Remove the missing values.
2. Detect outliers and remove outliers
3. Perform Univariate analysis using statistical and graphical tools.
4. Find the correlation between the attributes- Nominal, Numerical ordinal using Pearson correlation coefficient, Rank correlation coefficient and Chi-square test and also graphically using scatter plot and pairwise plot

Implementation (with R code and snap shot of output)

Data set used: Wine quality dataset
Title: Red wine quality
Source: Kaggle
Number of instances: 1599

Code and output screenshots:-

install.packages("tidyverse")
library(tidyverse)
wine <- read.csv('winequality.csv')
str(wine)

```
> wine <- read.csv('winequality.csv')
> str(wine)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

names(wine)

```
> names(wine)
 [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
 [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
 [7] "total.sulfur.dioxide" "density"              "pH"
[10] "sulphates"            "alcohol"              "quality"
```

dim(wine)

```
dim(wine)
] 1599   12
```

head(wine)

```
> head(wine)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1           7.4             0.70        0.00            1.9     0.076
2           7.8             0.88        0.00            2.6     0.098
3           7.8             0.76        0.04            2.3     0.092
4          11.2             0.28        0.56            1.9     0.075
5           7.4             0.70        0.00            1.9     0.076
6           7.4             0.66        0.00            1.8     0.075
  free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
1                  11                   34  0.9978 3.51      0.56     9.4       5
2                  25                   67  0.9968 3.20      0.68     9.8       5
3                  15                   54  0.9970 3.26      0.65     9.8       5
4                  17                   60  0.9980 3.16      0.58     9.8       6
5                  11                   34  0.9978 3.51      0.56     9.4       5
6                  13                   40  0.9978 3.51      0.56     9.4       5
```

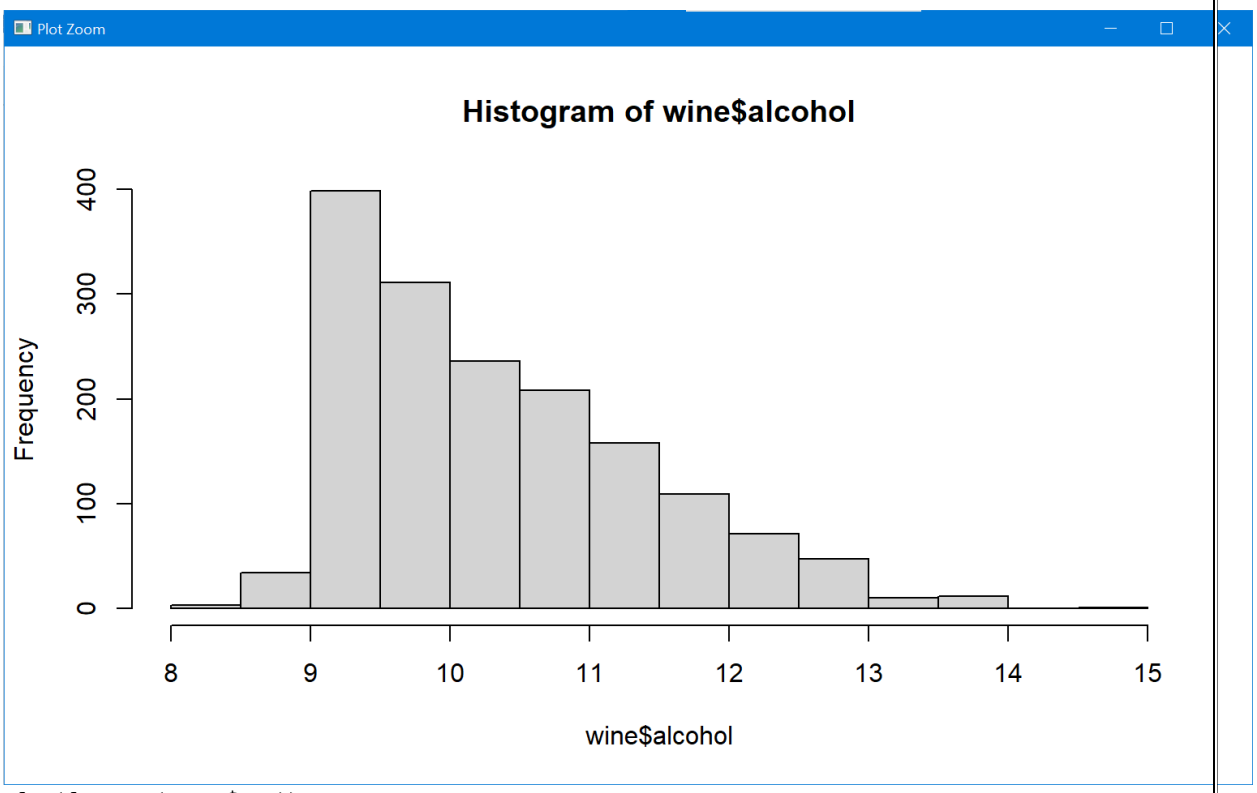
summary(wine)

```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100
 free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740
 1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210
 Median :14.00       Median : 38.00       Median :0.9968   Median :3.310
 Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311
 3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400
 Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010
   sulphates         alcohol         quality
 Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :0.6200   Median :10.20   Median :6.000
 Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

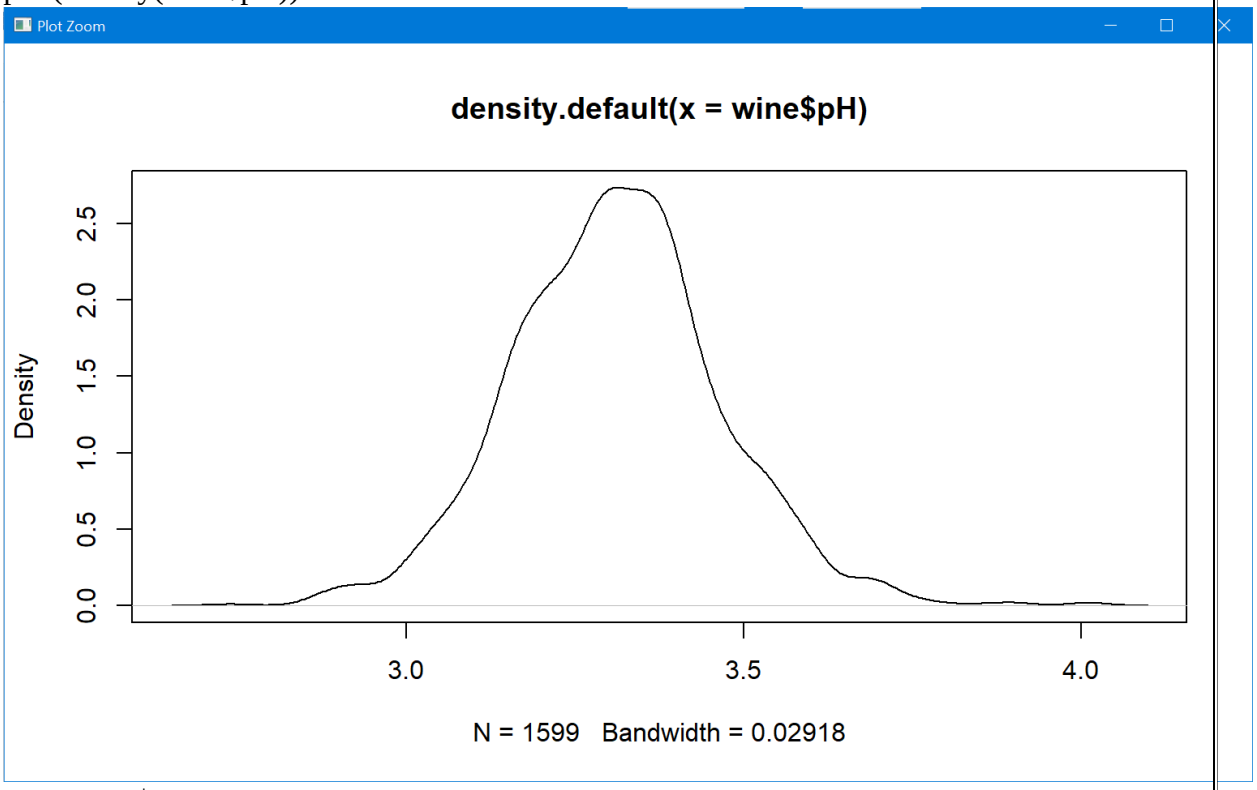
quantile(wine$pH)

```
> quantile(wine$pH)
   0%   25%   50%   75%  100%
2.74  3.21  3.31  3.40  4.01
```

hist(wine$alcohol)



plot(density(wine$pH))

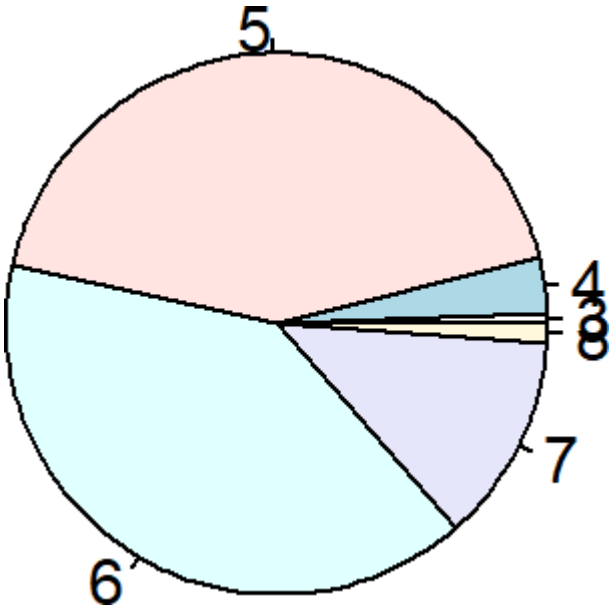


table(wine$quality)

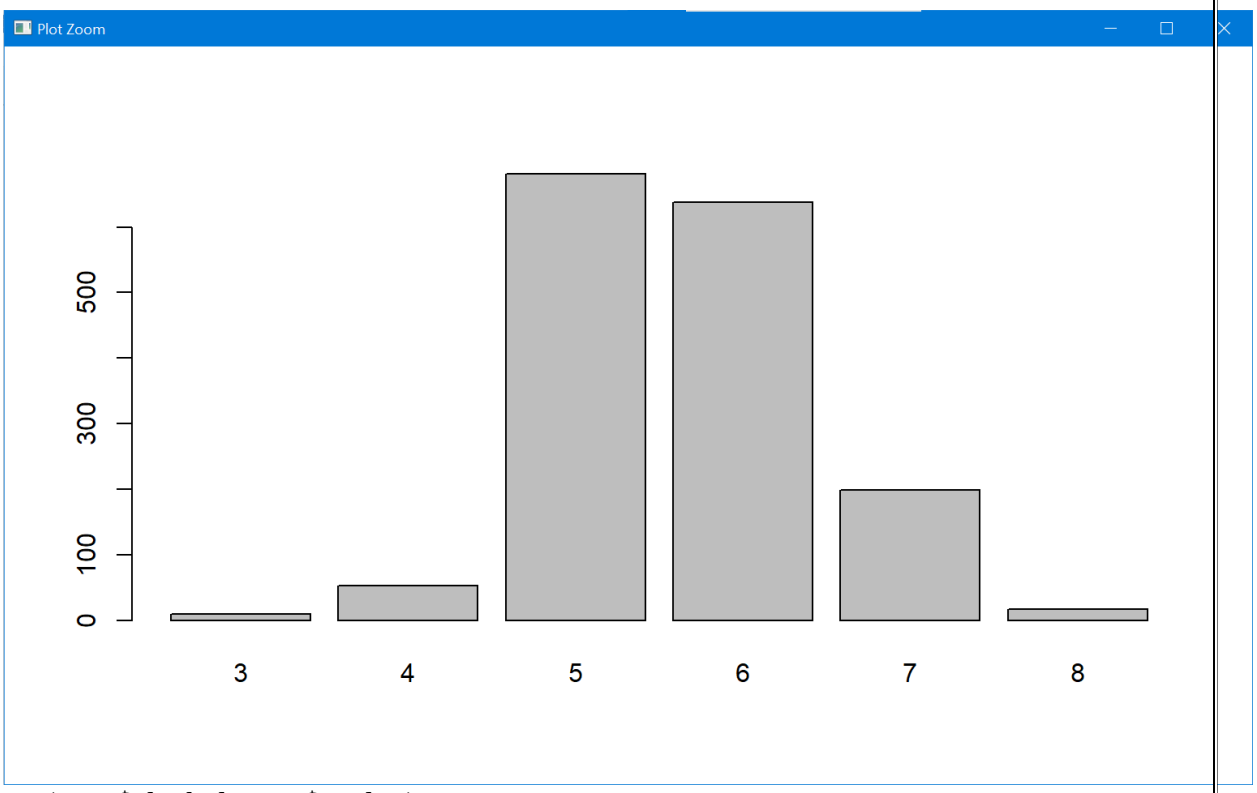
```
> table(wine$quality)

   3    4    5    6    7    8
  10   53  681  638  199   18
```

pie(table(wine$quality))



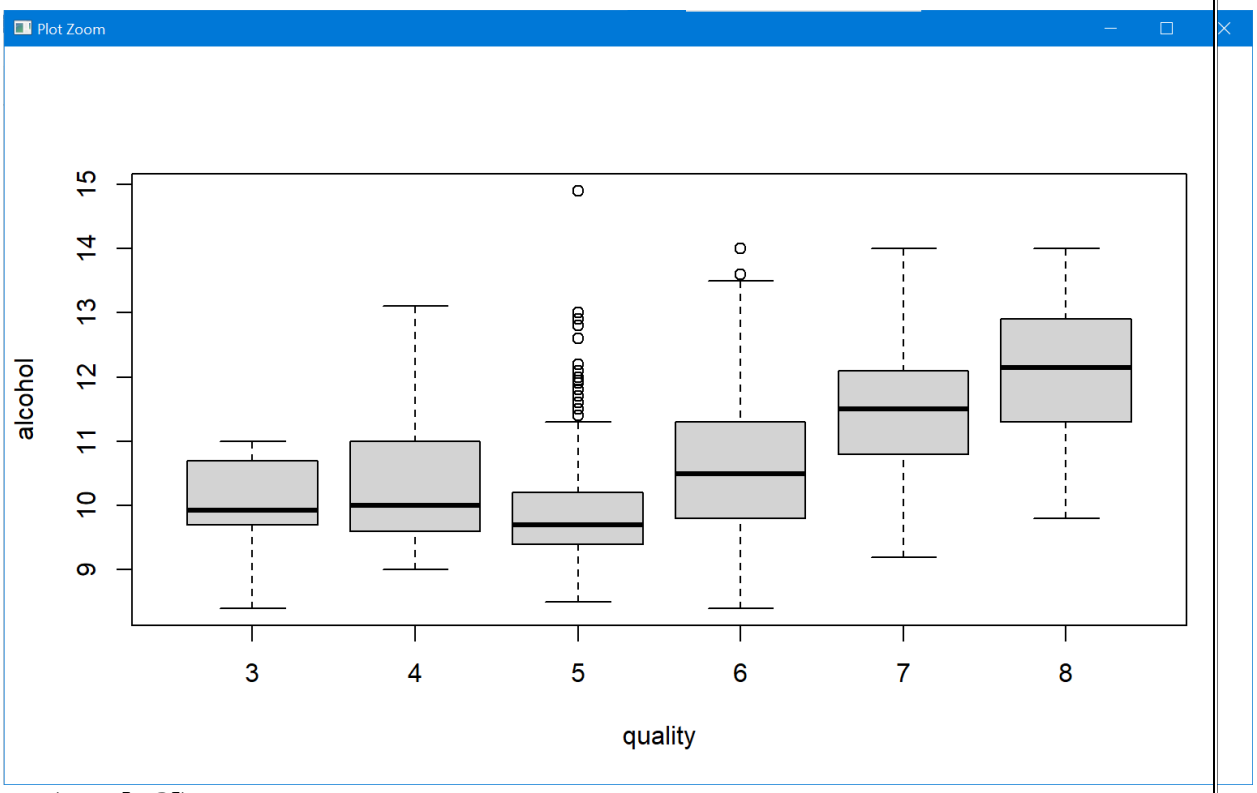
barplot(table(wine$quality))



cor(wine$alcohol, wine$quality)

```
> cor(wine$alcohol, wine$quality)
[1] 0.4761663
```

boxplot(alcohol~quality,data = wine)



cov(wine[1:3])

```
> cov(wine[1:3])
                  fixed.acidity volatile.acidity citric.acid
fixed.acidity        3.03141639      -0.07985142  0.22782000
volatile.acidity    -0.07985142       0.03206238 -0.01927162
citric.acid          0.22782000      -0.01927162  0.03794748
```

table(wine$fixed.acidity,wine$volatile.acidity,wine$citric.acid)

	0.12	0.16	0.18	0.19	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27
4.6	0	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0	0

	0.28	0.29	0.295	0.3	0.305	0.31	0.315	0.32	0.33	0.34	0.35	0.36
4.6	0	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	2	0	0

	0.365	0.37	0.38	0.39	0.395	0.4	0.41	0.415	0.42	0.43	0.44	0.45
4.6	0	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	1	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0	0

	0.46	0.47	0.475	0.48	0.49	0.5	0.51	0.52	0.53	0.54	0.545	0.55
4.6	0	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0	0

	0.56	0.565	0.57	0.575	0.58	0.585	0.59	0.595	0.6	0.605	0.61
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	1	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.615	0.62	0.625	0.63	0.635	0.64	0.645	0.65	0.655	0.66	0.665
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	1	0	0	0	0

	0.67	0.675	0.68	0.685	0.69	0.695	0.7	0.705	0.71	0.715	0.72
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.725	0.73	0.735	0.74	0.745	0.75	0.755	0.76	0.765	0.77	0.775
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.78	0.785	0.79	0.795	0.8	0.805	0.81	0.815	0.82	0.825	0.83
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.835	0.84	0.845	0.85	0.855	0.86	0.865	0.87	0.875	0.88	0.885
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.89	0.895	0.9	0.91	0.915	0.92	0.935	0.95	0.955	0.96	0.965
4.6	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0

	0.975	0.98	1	1.005	1.01	1.02	1.025	1.035	1.04	1.07	1.09	1.115
4.6	0	0	0	0	0	0	0	0	0	0	0	0
4.7	0	0	0	0	0	0	0	0	0	0	0	0
4.9	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
5.1	0	0	0	0	0	0	0	0	0	0	0	0
5.2	0	0	0	0	0	0	0	0	0	0	0	0

```
     1.13 1.18 1.185 1.24 1.33 1.58
4.6     0    0     0    0    0    0
4.7     0    0     0    0    0    0
4.9     0    0     0    0    0    0
5       0    0     0    0    0    0
5.1     0    0     0    0    0    0
5.2     0    0     0    0    0    0

[ reached getOption("max.print") -- omitted 90 row(s) and 79 matrix sl
ce(s) ]
```

chisq.test(wine$fixed.acidity,wine$volatile.acidity,wine$citric.acid)

```
> chisq.test(wine$fixed.acidity,wine$volatile.acidity,wine$citric.acid)

        Pearson's Chi-squared test

data:  wine$fixed.acidity and wine$volatile.acidity
X-squared = 16081, df = 13490, p-value < 2.2e-16

Warning message:
In chisq.test(wine$fixed.acidity, wine$volatile.acidity, wine$citric.ac
id) :
  Chi-squared approximation may be incorrect
```

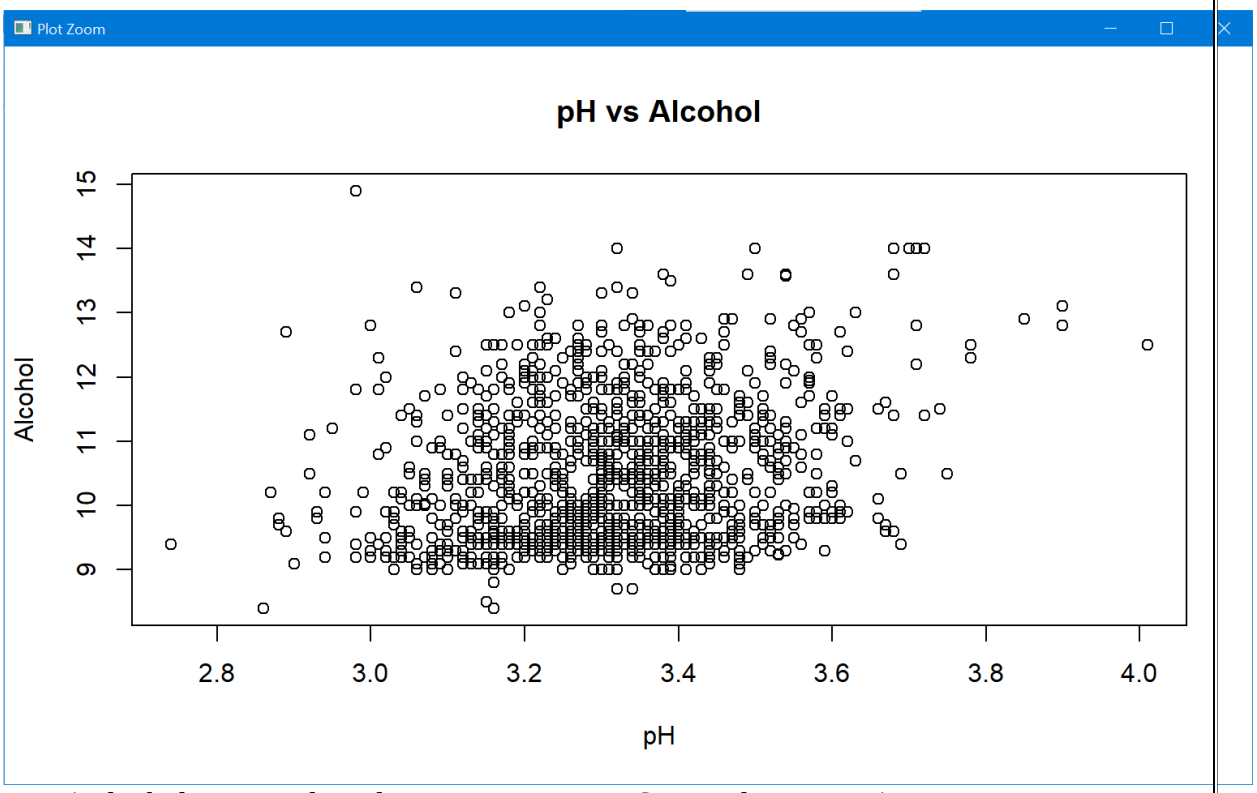
rankcorr = cor.test(x=wine$chlorides,y=wine$total.sulfur.dioxide,method = "spearman")

```
> rankcorr = cor.test(x=wine$chlorides,y=wine$total.sulfur.dioxide,meth
od = "spearman")
Warning message:
In cor.test.default(x = wine$chlorides, y = wine$total.sulfur.dioxide
 :
  Cannot compute exact p-value with ties
```

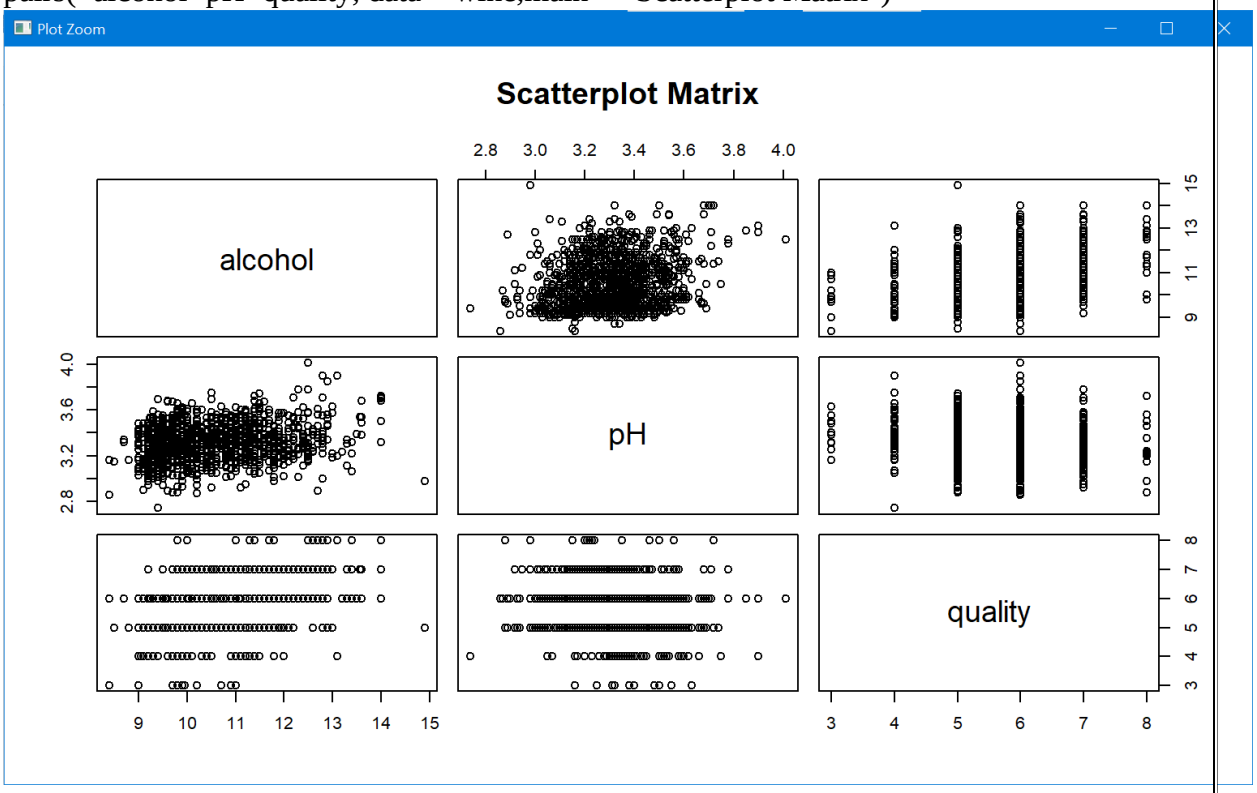
rankcorr$estimate

```
> rankcorr$estimate
      rho
0.1300333
```

plot(x = wine$pH, wine$alcohol,xlab = "pH",ylab = "Alcohol", main = "pH vs Alcohol")



pairs(~alcohol+pH+quality, data = wine,main = "Scatterplot Matrix")



Comments on these for your dataset:

What is your understanding of the data after implementing steps of EDA identified above?

- All the given steps were followed. We successfully loaded the dataset. We found pearson product moment coefficient, found covariance, did a chi-square test for independence, spearman rank correlation and plotted a scatter plot matrix.

Post lab Questions:

1. What is an appropriate way to visualize a list of the eye colors of 120 people?
   I. Boxplot
   II. Pie-chart
   III. Histogram
   IV. Scatterplot

Ans. Pie-chart

2. You want to investigate whether households in California tend to have a higher income than households in Massachusetts. Which summary measure would you use to compare the two states?

   I. median household income
   II. mean household income
   III. 3rd quartile of household income
   IV. IQR

Ans. median household income

3. Suppose all household incomes in California increase by 5%. How does that change the median household income?

   I. median household income goes up by 5%
   II. the median household income doesn't change
   III. cannot be determined from the information given

Ans. median household income goes up by 5%

4. Suppose all household incomes in California increase by $5,000. How does that change the interquartile range of the household incomes?
   I. cannot be determined from the information given
   II. the interquartile range of the household incomes doesn't change
   III. the interquartile range of the household incomes goes up by $5,000

Ans. the interquartile range of the household incomes goes up by $5,000

5. The median sales price for houses in a certain county during the last year was $342,000. What can we say about the percentage of sales represented by the houses that sold for more than $342,000?

I. the houses that sold for more than $342,000 represent more than 50% of all sales
II. the houses that sold for more than $342,000 represent exactly 50% of all sales
III. the houses that sold for more than $342,000 represent less than 50% of all sales

Ans. the houses that sold for more than $342,000 represent exactly 50% of all sales

6. Suppose all household incomes in California increase by $5,000. How does that change the standard deviation of the household incomes? the standard deviation of the household incomes doesn't change

I. cannot be determined from the information given
II. the standard deviation of the household incomes goes up by $5,000

Ans. cannot be determined from the information given

7. Which of the following graphical displays be used understand the distribution of data?

I. Box Plot
II. Quantile-Normal plot
III. Histogram
IV. Scatter Plot

Ans. Histogram

8. Correlation between two variables X&Y is 0.85. Now, after adding the value 2 to all the values of X, the correlation co-efficient will be

a. 0.85
b. 0.87
c. 0.65
d. 0.82

Ans. 0.85