Batch:H2-4	Roll No.:16010122257
Experiment No. 4	

Title: Exploratory Data Analysis

Aim: Use R libraries to implement exploratory data analysis on chosen datasets.

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:

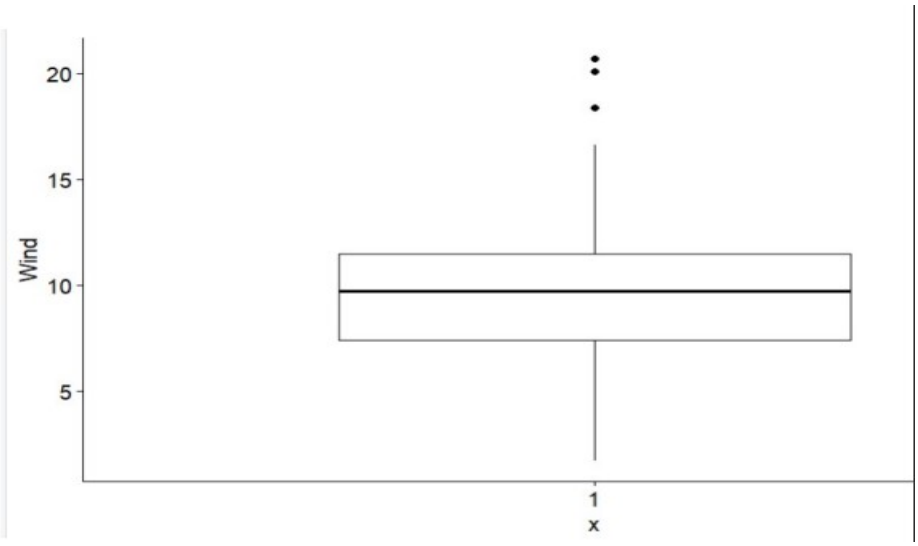
1. Data Mining Concepts and Techniques Jiawei Han, Michelin Kamber, Jian Pie, 3rd edition

---

How to Remove Outliers from Data Including Multi-Variables in R

In our case, we select the quantitative variables from iris data. We have four variables – sepal length, sepal width, petal length and petal width and 150 observations.
The outliers can be observed using the boxplot() function.

INITIAL BOXPLOT WITH OUTLIERS:



NEXT:
Outliers are either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile.
We find first (Q1) and third (Q3) quartiles by using the quantile() function.
Then, the interquartile range (IQR) is found by the IQR() function.
Then, we calculate Q1 – 1.5*IQR to find the lower limit for outliers.

After that, we calculate Q3 + 1.5*IQR to find the upper limit for outliers.
Then, we use the subset() function to eliminate outliers.

CODE:

```
ggboxplot(my_data,y="Wind",width=0.6)
data<-airquality[,1:4]
dim(data)
```

OUTPUT:

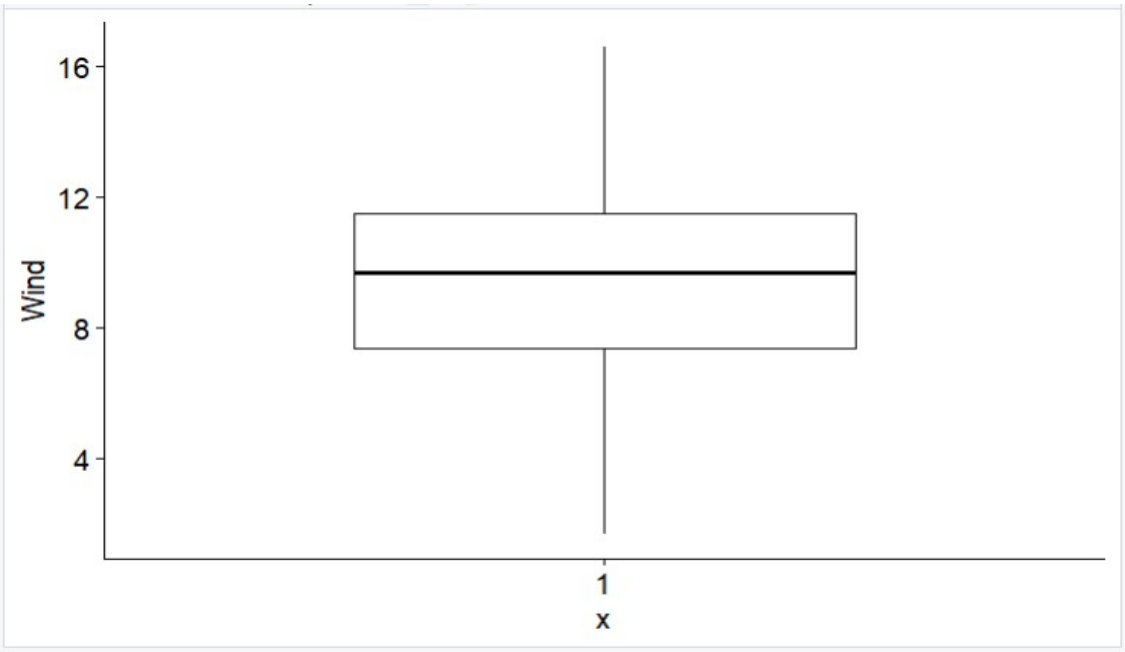
```
[1] 153   4
```

CODE:

```
quartiles<-quantile(data$Wind,probs=c(.25,.75),na.rm=FALSE)
quartiles
IQR<-IQR(data$Wind)
IQR
Lower<-quartiles[1]-1.5*IQR
Upper<-quartiles[2]+1.5*IQR
Lower
Upper
data_no_outlier<-subset(data,data$Wind>Lower & data$Wind<Upper)
dim(data_no_outlier)
ggboxplot(data_no_outlier,y="Wind",width=0.6)
```

OUTPUT:

```
> quartiles<-quantile(data$Wind,probs=c(.25,.75),na.rm=FALSE)
> quartiles
 25%  75%
 7.4 11.5
> IQR<-IQR(data$Wind)
> IQR
[1] 4.1
> Lower<-quartiles[1]-1.5*IQR
> Upper<-quartiles[2]+1.5*IQR
> Lower
 25%
1.25
> Upper
  75%
17.65
> data_no_outlier<-subset(data,data$Wind>Lower & data$Wind<Upper)
> dim(data_no_outlier)
[1] 150   4
> ggboxplot(data_no_outlier,y="Wind",width=0.6)
>
```



Handling missing values

1. Multiple NA or NAN values can exist in a vector.
2. To deal with NA type of missing values in a vector we can use is.na() function by passing the vector as an argument.
3. To deal with the NAN type of missing values in a vector we can use is.nan() function by passing the vector as an argument.
4. Generally, NAN values can be included in the NA type but the vice-versa is not true.

Removing Missing Data/ Values

1. na.omit – It simply rules out any rows that contain any missing value and forgets those rows forever.
2. na.exclude – This ignores rows having at least one missing value.
3. na.pass – Take no action.
4. na.fail – It terminates the execution if any of the missing values are found.

CODE:

```
na.omit
```

OUTPUT:

```
> na.omit
function (object, ...)
UseMethod("na.omit")
<bytecode: 0x00000224eabbf820>
<environment: namespace:stats>
```

CODE + OUTPUT:

```
> na.exclude(5)
[1] 5
```

Filling Missing Values with Mean or Median

Consider a dataframe that contains NA values for example:

Create a list of columns having at least one NA value.

CODE:

```
dataframe<-data.frame(Name=c("Tiger","Lion","Leopard","Cheetah"),
                      India=c(3000,700,10000,30),
                      SouthAfrica=c(NA,2500,3000,2600),
                      Nepal=c(900,NA,1250,NA))
print(dataframe)
listMissingColumns<-colnames(dataframe)[apply(dataframe,2,anyNA)]
print(listMissingColumns)
```

OUTPUT:

```
> dataframe<-data.frame(Name=c("Tiger","Lion","Leopard","Cheetah"),
+                       India=c(3000,700,10000,30),
+                       SouthAfrica=c(NA,2500,3000,2600),
+                       Nepal=c(900,NA,1250,NA))
> print(dataframe)
     Name India SouthAfrica Nepal
1   Tiger  3000          NA   900
2    Lion   700        2500    NA
3 Leopard 10000        3000  1250
4 Cheetah    30        2600    NA
> listMissingColumns<-colnames(dataframe)[apply(dataframe,2,anyNA)]
> print(listMissingColumns)
[1] "SouthAfrica" "Nepal"
>
```

Compute the mean and median of the corresponding columns. Since we need to omit NA values in the missing columns, therefore, we can pass the "na.rm = True" argument to the apply() function.

```
> meanMissing <- apply(dataframe[,colnames(dataframe)%in%listMissingColumns],
+                      2,mean,na.rm=TRUE)
> print(meanMissing)
SouthAfrica       Nepal
       2700        1075
>
```

```
> medianMissing<-apply(dataframe[,colnames(dataframe)%in% listMissingColumns],
+                      2,median,na.rm=TRUE)
> print(medianMissing)
SouthAfrica       Nepal
       2600        1075
>
```

Now our mean and median values of corresponding columns are ready. In this step, we will replace NA values with mean and median using the mutate() function which is defined under the “dplyr” package.

```
> newDataFrameMean<-dataframe%>%mutate(
+   SouthAfrica=ifelse(is.na(SouthAfrica),meanMissing[1],SouthAfrica),
+   Nepal=ifelse(is.na(Nepal),meanMissing[2],Nepal)
+ )
> print(newDataFrameMeanMean)
Error: object 'newDataFrameMeanMean' not found
> print(newDataFrameMean)
     Name India SouthAfrica Nepal
1   Tiger  3000        2700   900
2    Lion   700        2500  1075
3 Leopard 10000        3000  1250
4 Cheetah    30        2600  1075
>
```

```
> newDataFrameMedian<-dataframe%>%mutate(
+   SouthAfrica=ifelse(is.na(SouthAfrica),medianMissing[1],SouthAfrica),
+   Nepal=ifelse(is.na(Nepal),medianMissing[2],Nepal)
+ )
> print(newDataFrameMedian)
     Name India SouthAfrica Nepal
1   Tiger  3000        2600   900
2    Lion   700        2500  1075
3 Leopard 10000        3000  1250
4 Cheetah    30        2600  1075
>
```

What is Exploratory Data Analysis (EDA) ?

Exploratory data analysis (EDA) is the process of analysing data to uncover their key features. 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

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 and spread including skewness and kurtosis

Estimate Skewness and Kurtosis

Load the moments library

```
> install.packages("moments")
```

```
> library(moments)
```

Calculate skewness. Skewness is a measure of symmetry.

Negative skewness: mean of the data < median and the data distribution is left-skewed.

Positive skewness: mean of the data > median and the data distribution is right-skewed.

```
> skewness(airquality$Wind)
[1] 0.3443985
>
```

Distribution is skewed towards the right.

The normal distribution has zero kurtosis and thus the standard tail shape. It is said to be mesokurtic.
Negative kurtosis would indicate a thin-tailed data distribution, and is said to be platykurtic.
Positive kurtosis would indicate a fat-tailed distribution, and is said to be leptokurtic.

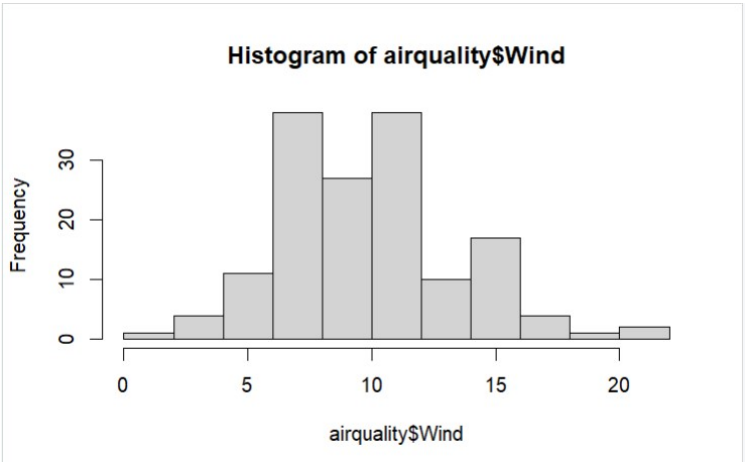
```
> kurtosis(airquality$Wind)
[1] 3.068849
>
```

Uni-variate graphical EDA

Common sorts of univariate graphics are:

1. ***Histogram***: The foremost basic graph is a histogram, which may be a barplot during which each bar represents the frequency (count) or proportion (count/total count) of cases for a variety of values. Histograms are one of the simplest ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.
2. ***Stem-and-leaf plots***: An easy substitute for a histogram may be stem-and-leaf plots. It shows all data values and therefore the shape of the distribution.
3. ***Box Plots***: Another very useful univariate graphical technique is the boxplot. Boxplots are excellent at presenting information about central tendency and show robust measures of location and spread also as providing information about symmetry and outliers, although they will be misleading about aspects like multimodality. One among the simplest uses of boxplots is within the sort of side-by-side boxplots.
4. ***Quantile-normal plots***: The ultimate univariate graphical EDA technique is the most intricate. It's called the quantile-normal or QN plot or more generally the quantile-quantile or QQ plot. it's wont to see how well a specific sample follows a specific theoretical distribution. It allows detection of non-normality and diagnosis of skewness and kurtosis

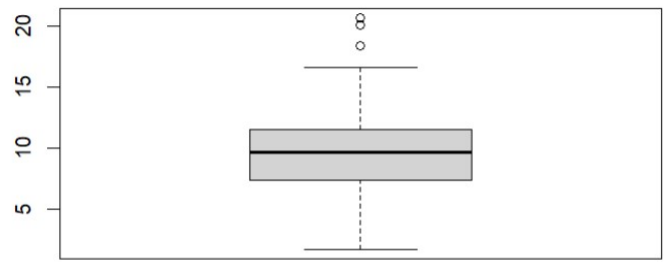
```
> hist(airquality$Wind)
>
```



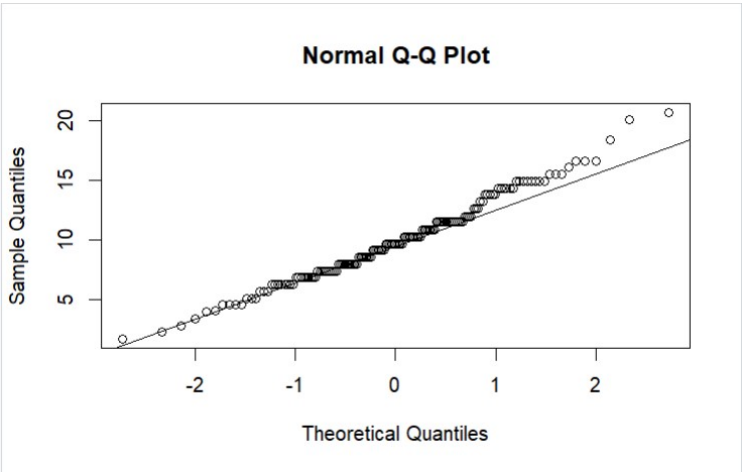

```
> stem(airquality$Wind)
```

```
  The decimal point is at the |

   1 | 7
   2 | 38
   3 | 4
   4 | 016666
   5 | 111777
   6 | 33333333999999999
   7 | 4444444444
   8 | 00000000000666666666
   9 | 22222222777777777777
  10 | 3333333333399999999
  11 | 555555555555555
  12 | 0000666
  13 | 2288888
  14 | 33333399999999
  15 | 555
  16 | 1666
  17 |
  18 | 4
  19 |
  20 | 17

>
```

```
> boxplot(airquality$Wind)
>
```




```
> qqnorm(airquality$Wind)
> qqline(airquality$Wind)
>
```



Multi-variate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation (please refer to experiment 2) or by calculating the correlation coefficient.

```
> a <- c(7,6,9,12,14)
> b <- c(1,4,19,23,24)
> print(cor(a,b))
[1] 0.9037814
> print(cor(a,b,method="spearman"))
[1] 0.9
>
```

SAMPLE FOR REFERENCE:

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
-4	-9.6	16	16	38.4
-2	0.4	4	4	-0.8
0	-7.6	0	0	0
2	22.4	4	4	44.8
4	-5.6	16	16	-22.4
0	0	**40** (SS_x)	**683.2** (SS_y)	**60** (SP_{xy})

Pearson's correlation coefficient

$$S_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$S_{XY} = \frac{60}{5 - 1} = 15$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}}$$

$$r = \frac{60}{\sqrt{(40*683.2)}} = \mathbf{0.363}$$

Ranks

X	Y
1	1
2	4
3	2
4	5
5	3

X does not contain ties.
Y does not contain ties.

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
-2	-2	4	4	4
-1	1	1	1	-1
0	-1	0	0	0
1	2	1	1	2
2	0	4	4	0
0	0	**10** (SS_x)	**10** (SS_y)	**5** (SP_{xy})

Spearman's rank correlation coefficient

$$S_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$S_{XY} = \frac{5}{5 - 1} = 1.25$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}}$$

$$r = \frac{5}{\sqrt{(10*10)}} = \mathbf{0.5}$$

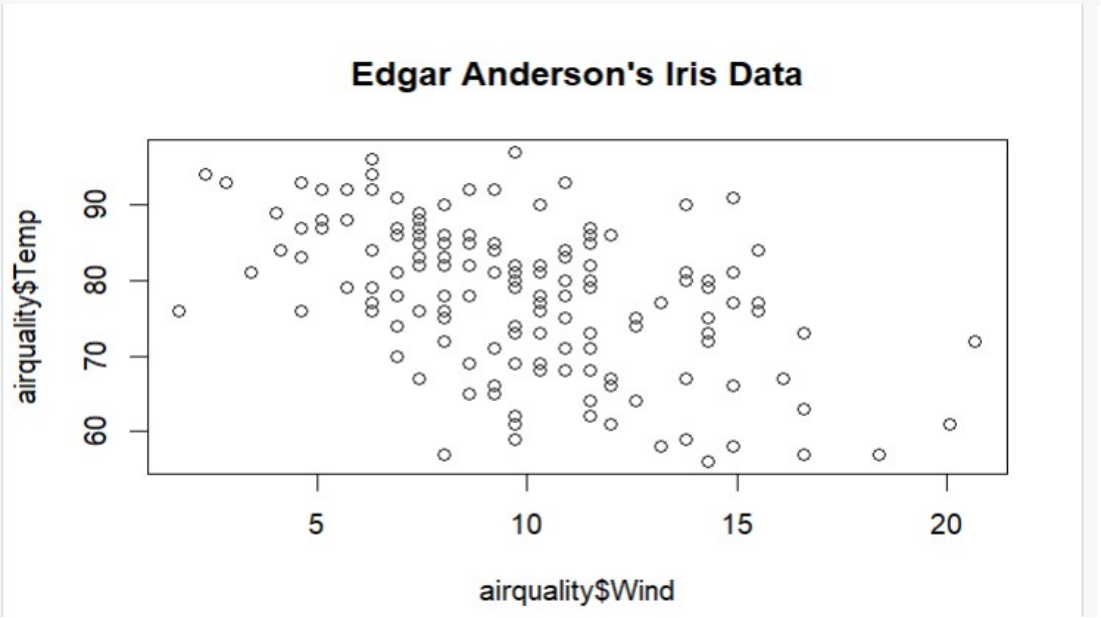
You can also perform **chi-squared tests** for multivariate graphical exploratory data analysis (EDA). For further reading on chi-squared tests in R, you may refer to the article Chi-Squared Test | R-bloggers.

Multi-variate graphical EDA

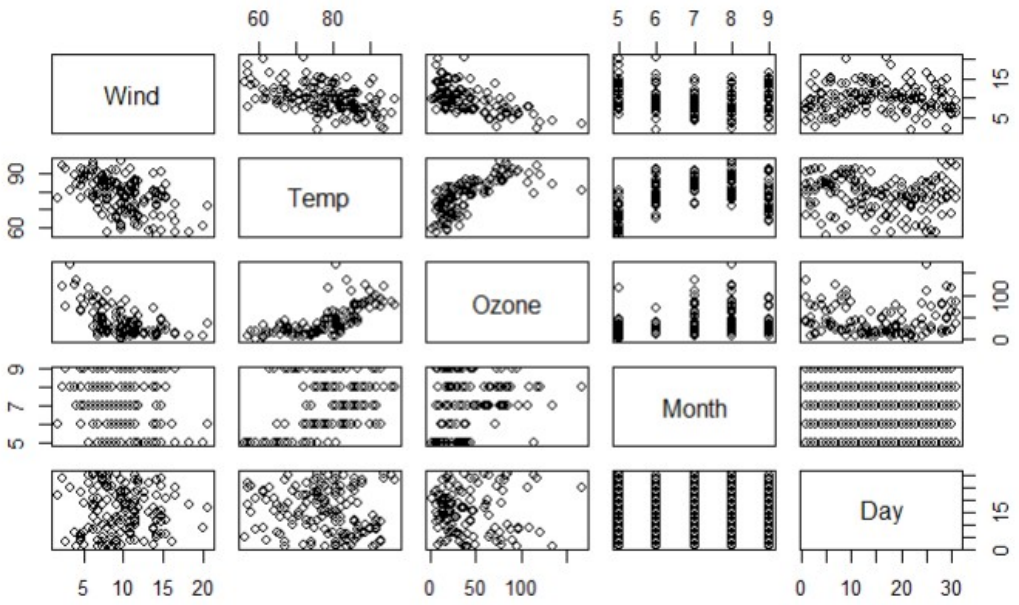
Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable. **Stripcharts** can also be used. Please refer to experiment 2 for the code and examples.

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 convention to put the outcome on the y (vertical) axis.
Scatter plots can be extended to *n* attributes, resulting in a ***scatter-plot matrix***.

```
> plot(airquality$Wind,airquality$Temp,main="Edgar Anderson's Iris Data")
>
```



```
> pairs(~Wind+Temp+Ozone+Month+Day,data=airquality)
>
```



Procedure for Implementation in lab :

1. Identify a Dataset for Exploratory Data Analysis
2. Handle the missing values appropriately
3. Detect and remove outliers
4. Perform the following:
    a. univariate non-graphical EDA
    b. univariate graphical EDA
    c. multivariate non-graphical EDA
    d. multivariate graphical EDA
5. What is your understanding of the data after implementing steps of EDA identified above?

From the implemented steps of Exploratory Data Analysis (EDA) outlined above, it is apparent that the dataset has undergone comprehensive scrutiny to understand its characteristics, structure, and underlying patterns. Through various statistical and visual techniques, key insights have been derived, such as:

1. Data Distribution: The distribution of variables has been explored using histograms, box plots, or density plots, revealing the central tendency, spread, and skewness of the data.

2. Data Relationships: Relationships between variables have been examined through correlation matrices, scatter plots, or pair plots. This has provided insights into potential dependencies or associations between different features.

3. Data Quality: Checks for data quality issues such as missing values, outliers, or inconsistencies have been conducted. Strategies for handling these issues have been identified, which may include imputation, transformation, or removal of problematic data points.

4. Feature Importance: Techniques like feature importance ranking or correlation analysis have been employed to identify the most influential variables for modeling purposes.

5. Data Preprocessing Needs: Based on the observed patterns and anomalies, requirements for data preprocessing have been identified. This may involve data cleaning, feature engineering, or scaling to prepare the data for further analysis or modeling.

6. Initial Hypotheses: Initial hypotheses about the relationships within the data or potential drivers of certain outcomes have been formulated. These hypotheses can guide further analysis or experimentation.

7. Scope for Further Analysis: Areas where further analysis or investigation is warranted have been identified. This could include deeper dives into specific subsets of data, exploration of additional variables, or conducting advanced statistical modeling.

Overall, the EDA process has equipped us with a solid understanding of the dataset's characteristics and provided valuable insights that can inform subsequent steps in the data analysis pipeline, such as feature selection, model building, and decision-making.
Students should add their R code and screenshots of output. Also students should provide the following details of the dataset:

Data set used:**Air Quality Dataset**
Source:**R data sets**

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.:III. Histogram - This is an appropriate way to visualize a list of eye colors as it allows you to see the frequency or count of each eye color category within the dataset.

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.:I. Median household income - Median is a better summary measure than mean when comparing incomes because it is less affected by extreme values.

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.:II. The median household income doesn't change - Increasing all incomes by the same percentage does not affect the relative ordering of the incomes, so the median remains the same.

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.:II. The interquartile range of the household incomes doesn't change - Adding the same amount to all incomes does not affect the spread of the data, so the interquartile range remains the same.

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.:III. The houses that sold for more than $342,000 represent less than 50% of all sales - Since the median is less than $342,000, more than half of the houses sold for less than $342,000, implying that those sold for more represent less than 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?
   I. the standard deviation of the household incomes doesn't change
   II. cannot be determined from the information given
   III. the standard deviation of the household incomes goes up by $5,000

ANS.:I. The standard deviation of the household incomes doesn't change - Adding the same amount to all incomes does not change the variability of the data, so the standard deviation remains the same.

7. Which of the following graphical displays can be used to understand the distribution of data?
   I. Box Plot
   II. Quantile-Normal plot
   III. Histogram
   IV. Scatter Plot

ANS.:I. Box Plot, III. Histogram - Both box plots and histograms are graphical displays commonly used to understand the distribution of data.

8. Correlation between two variables X&Y is 0.85. Now, after adding the value 2 to all the values of X, the correlation coefficient will be
   a. 0.85
   b. 0.87
   c. 0.65
   d. 0.82

ANS.:a. 0.85 - Adding a constant to all values of one variable does not change the correlation coefficient between that variable and another. So, the correlation coefficient remains the same, which is 0.85.