**Batch: A - 2 Roll No.: 16014022050**

**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:**

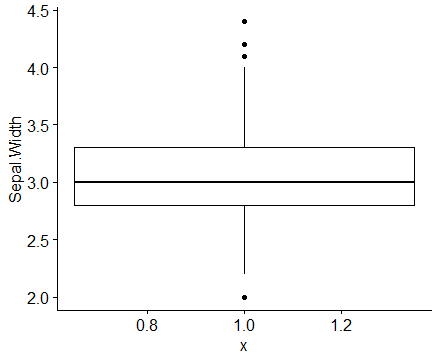
1. Data Mining Concepts and Techniques Jiawei Han, Michelin Kamber, Jian Pie, 3rd edition

**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.





There exist four outliers seen in the sepal width variable.

Outliers are either 1.5 × IQR or more above the third quartile or 1.5×IQR or more below the first quartile.

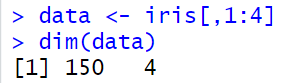
We find first (Q1) and third (Q3) quartiles by using the quantile() function.

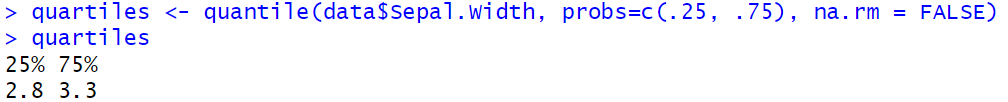
Then, the interquartile range (IQR) is found by the IQR() function.

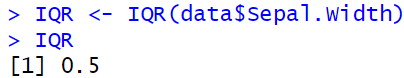
Then, we calculate Q1 – 1.5\*IQR to find the lower limit for outliers.

After that, we calculate Q3 + 1.5\*IQR to find the upper limit for outliers.

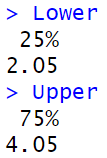
Then, we use the subset() function to eliminate outliers.

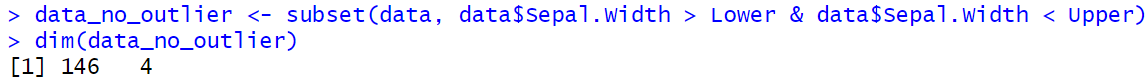
****

****

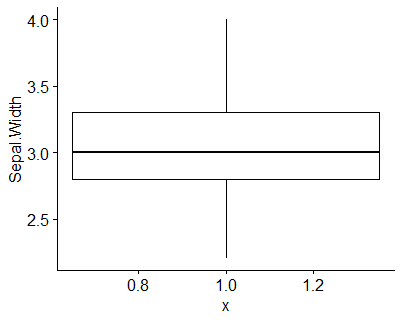
****

****

****

****

****

****

**Handling missing values:**

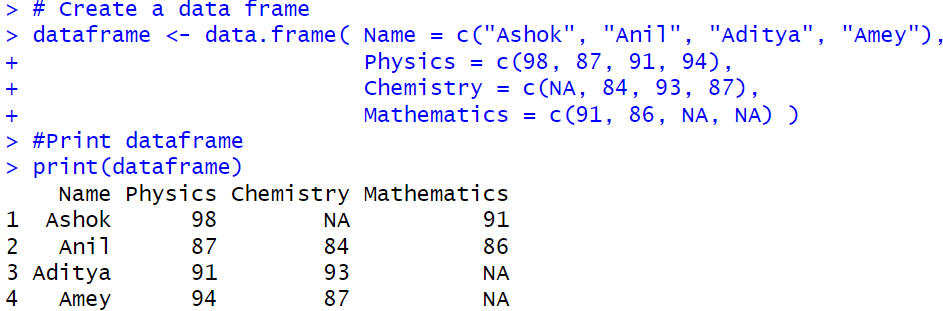
1. Multiple NA or NAN values can exist in a vector.
2. To deal with NA type of missing values in a vector we can use is.na() function by passing the vector as an argument.
3. To deal with the NAN type of missing values in a vector we can use is.nan() function by passing the vector as an argument.
4.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

**Removing Missing Data/ Values:**

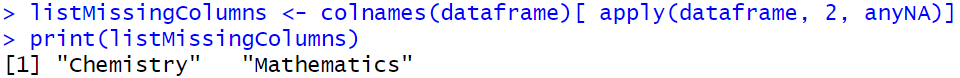
1. na.omit − It simply rules out any rows that contain any missing value and forgets those rows forever.
2. na.exclude − This ignores rows having at least one missing value.
3. na.pass − Take no action.
4. na.fail − It terminates the execution if any of the missing values are found.

**Filling Missing Values with Mean or Median:**

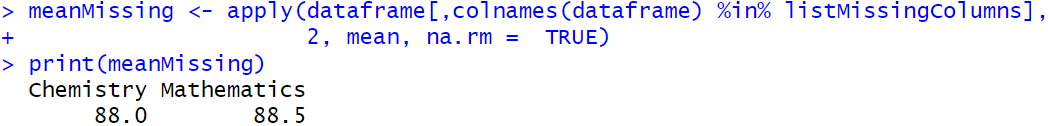
Consider a dataframe that contains NA values for example:

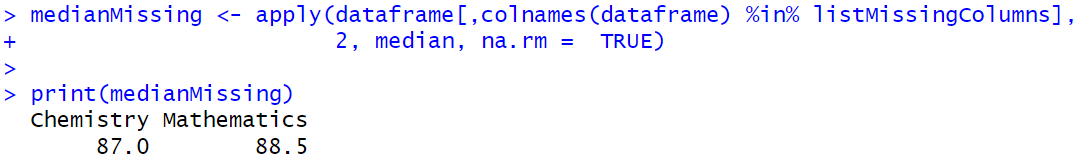


Create a list of columns having at least one NA value.

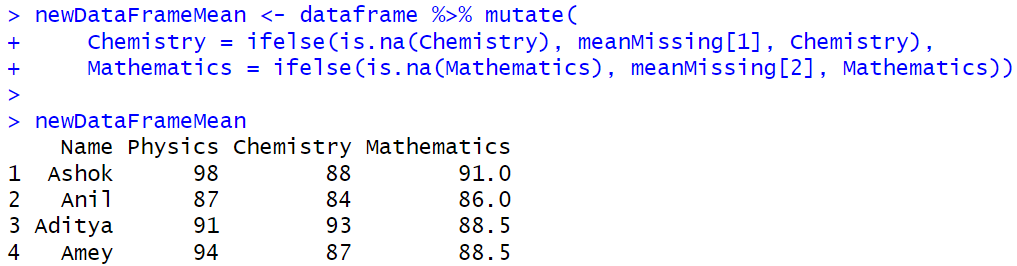


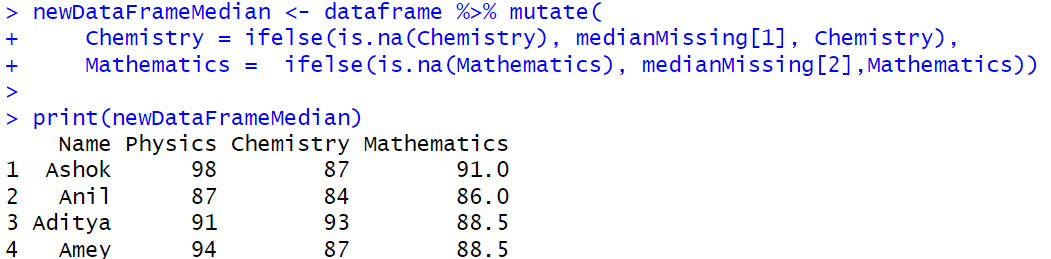
Compute the mean and median of the corresponding columns. Since we need to omit NA values in the missing columns, therefore, we can pass the "na.rm = True" argument to the apply() function.





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.





**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





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.



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.



It is leptokurtic.

**Uni-variate graphical EDA:**

Common sorts of univariate graphics are:

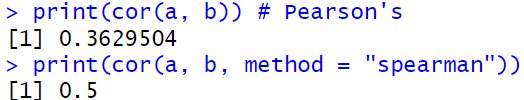
1. ***Histogram***: The foremost basic graph is a histogram, which may be a barplot during which each bar represents the frequency (count) or proportion (count/total count) of cases for a variety of values. Histograms are one of the simplest ways to quickly learn a lot about your data, including central 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

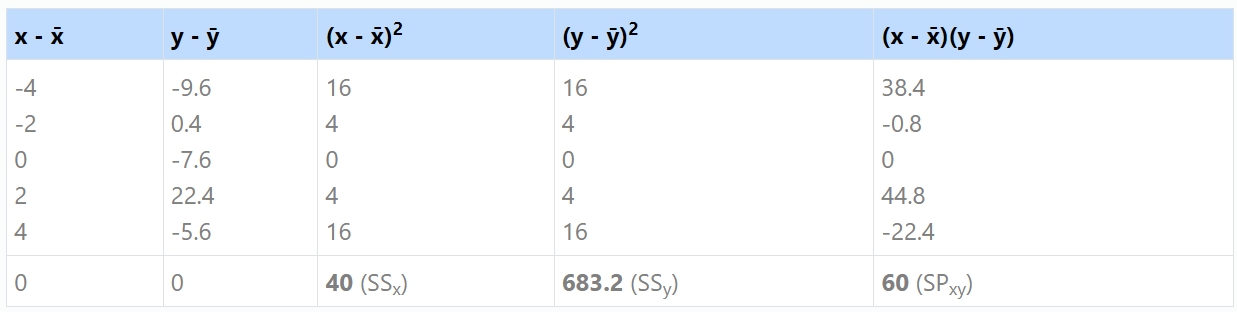
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

**Multi-variate non-graphical EDA:**

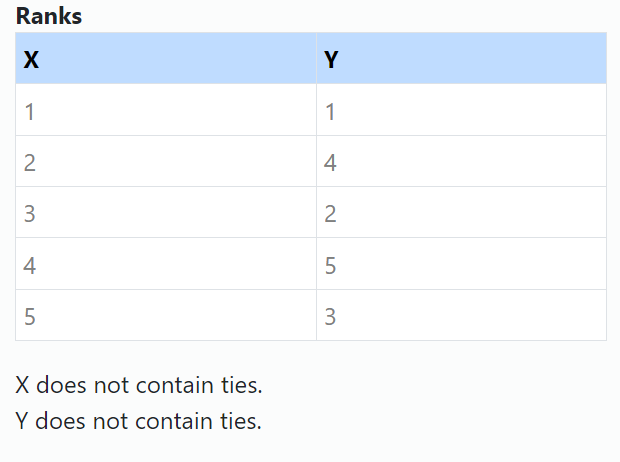
Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation (please refer to experiment 2) or by calculating the correlation coefficient.

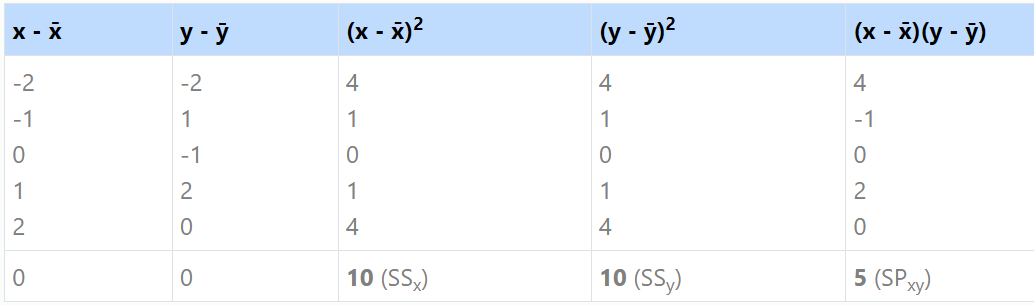






|  |  |
| --- | --- |
| Pearson’s correlation coefficient |  |





|  |  |
| --- | --- |
| Spearman’s rank correlation coefficient |  |

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](https://www.r-bloggers.com/2016/08/chi-squared-test/).

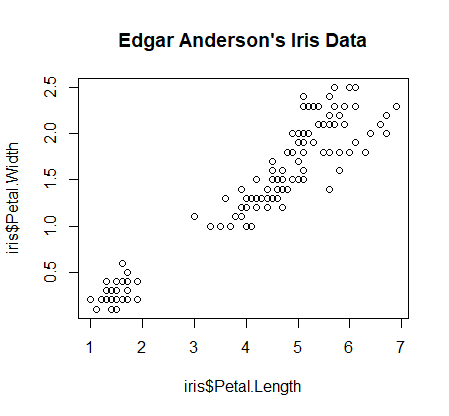
**Multi-variate graphical EDA:**

**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.

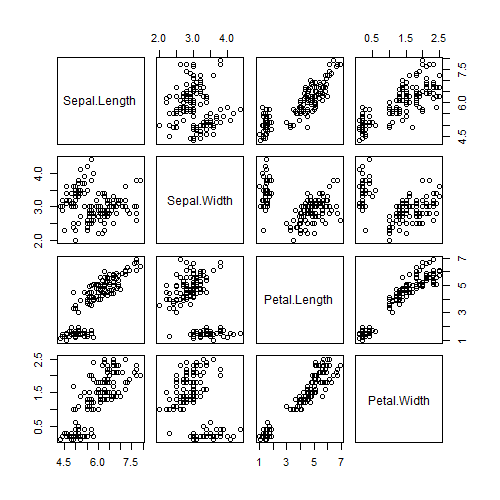
For two quantitative variables, the basic graphical EDA technique is the **scatterplot** which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. If one variable is explanatory and the other is outcome, it is a convention to put the outcome on the y (vertical) axis.

Scatter plots can be extended to *n* attributes, resulting in a ***scatter-plot matrix***.









**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:
   1. univariate non-graphical EDA
   2. univariate graphical EDA
   3. multivariate non-graphical EDA
   4. multivariate graphical EDA
5. What is your understanding of the data after implementing steps of EDA identified above?

**Students should add their R code and screenshots of output. Also, students should provide the following details of the dataset:**

Data set used: stack.loss

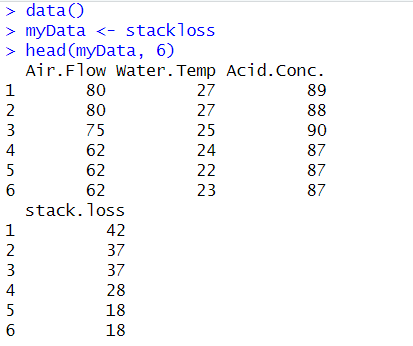
Title: Brownlee's Stack Loss Plant Data

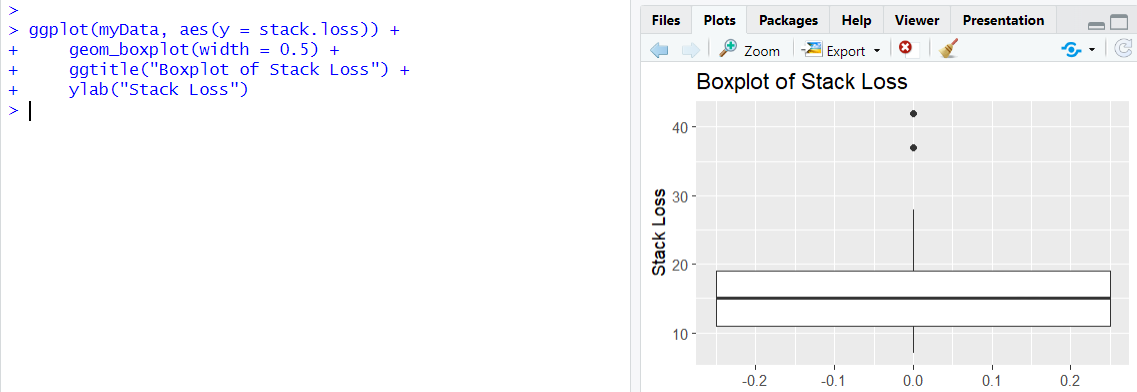
Source: R Datasets

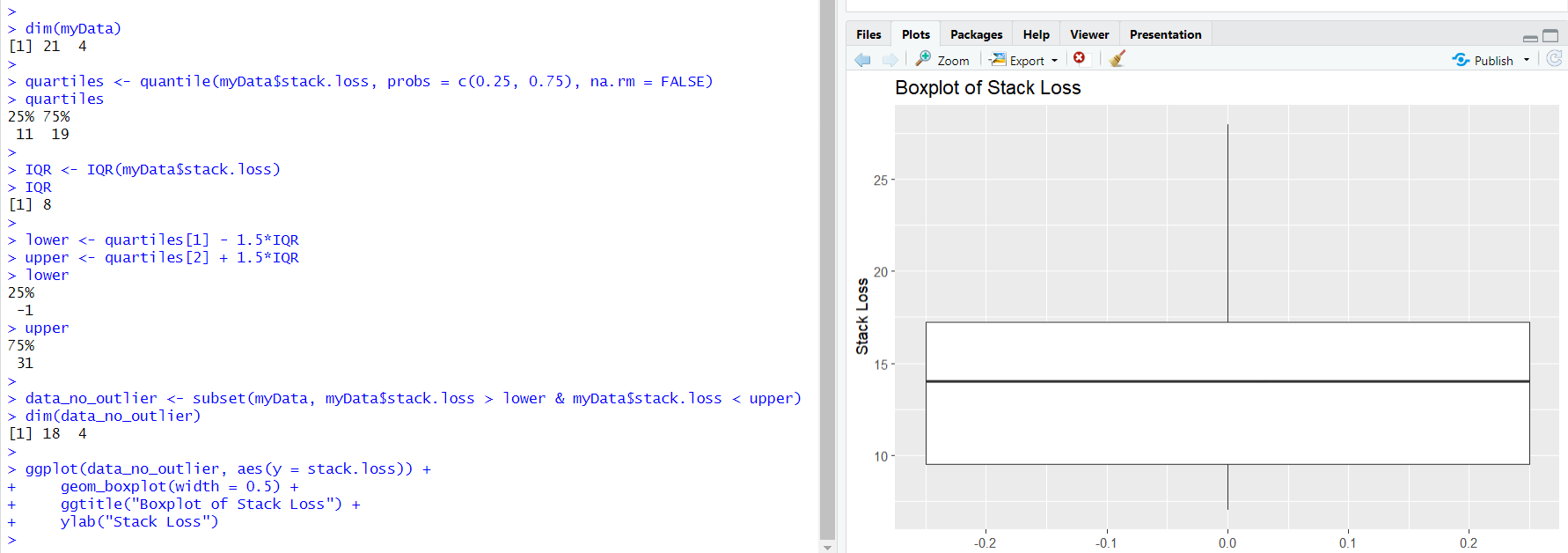
Number of instances:

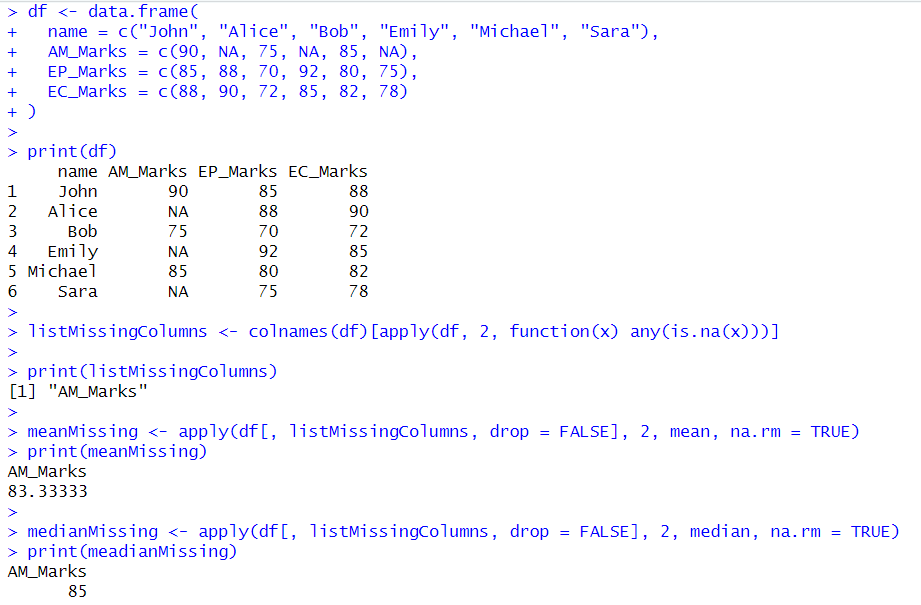
Number of attributes:

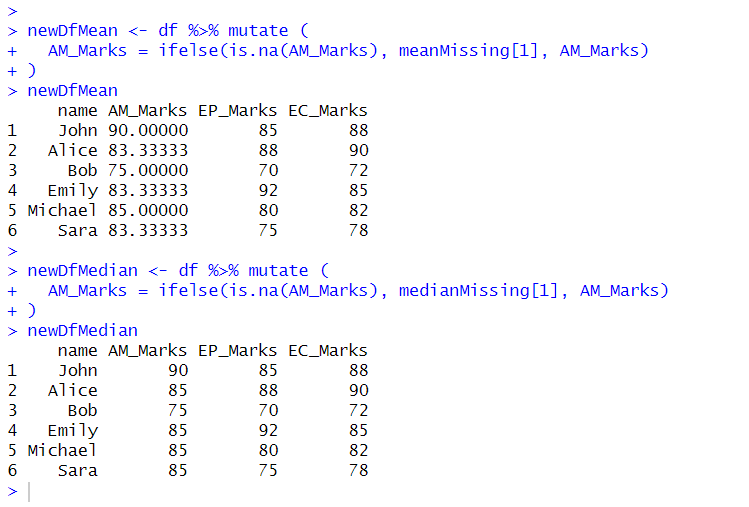
Attribute information**:**

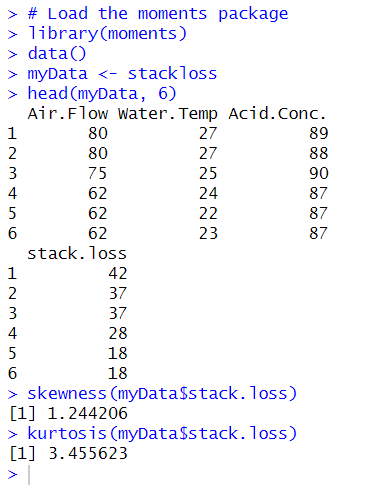
****

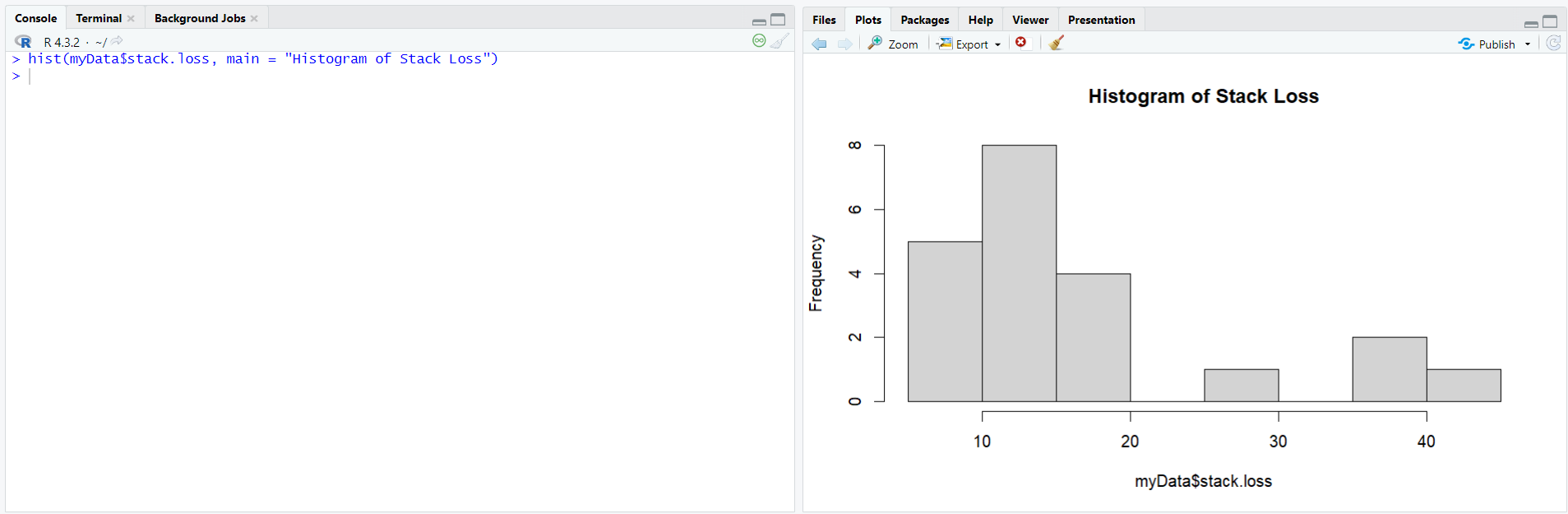
****

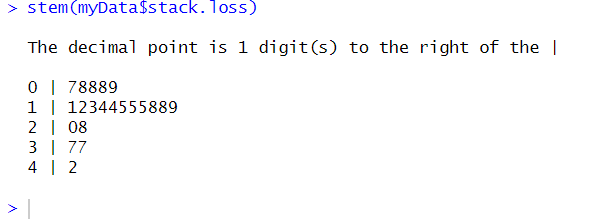
****

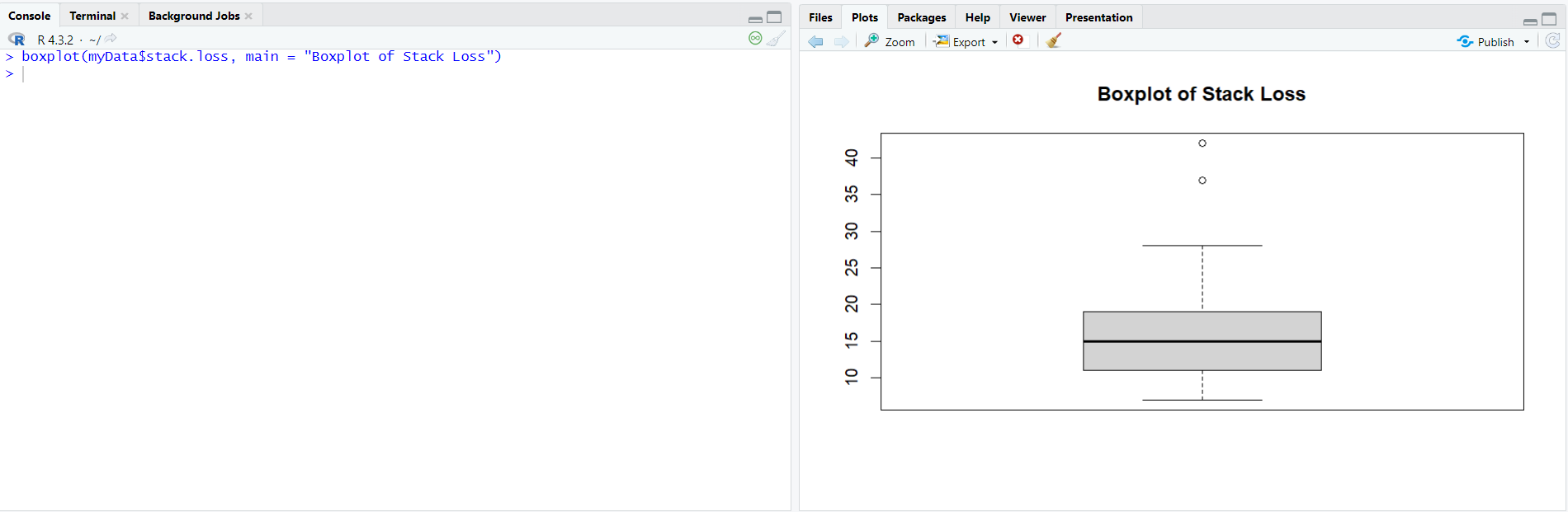
****

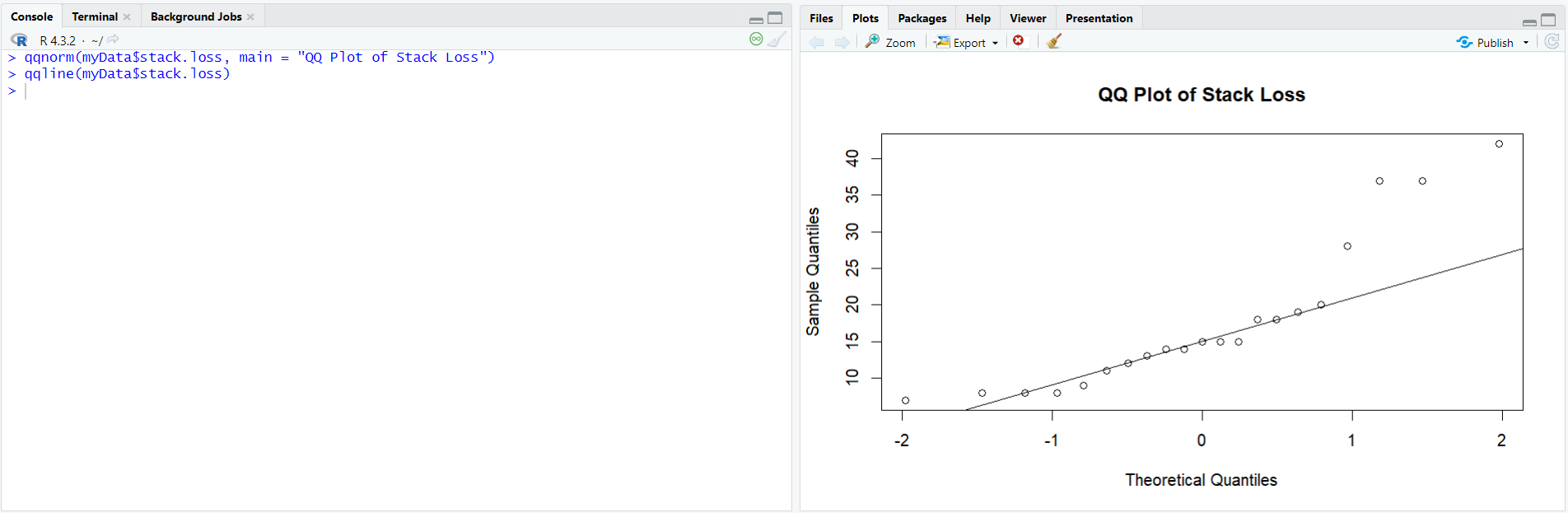
****

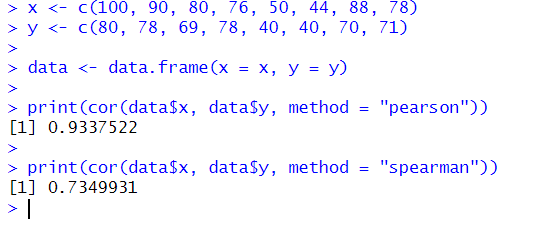
****

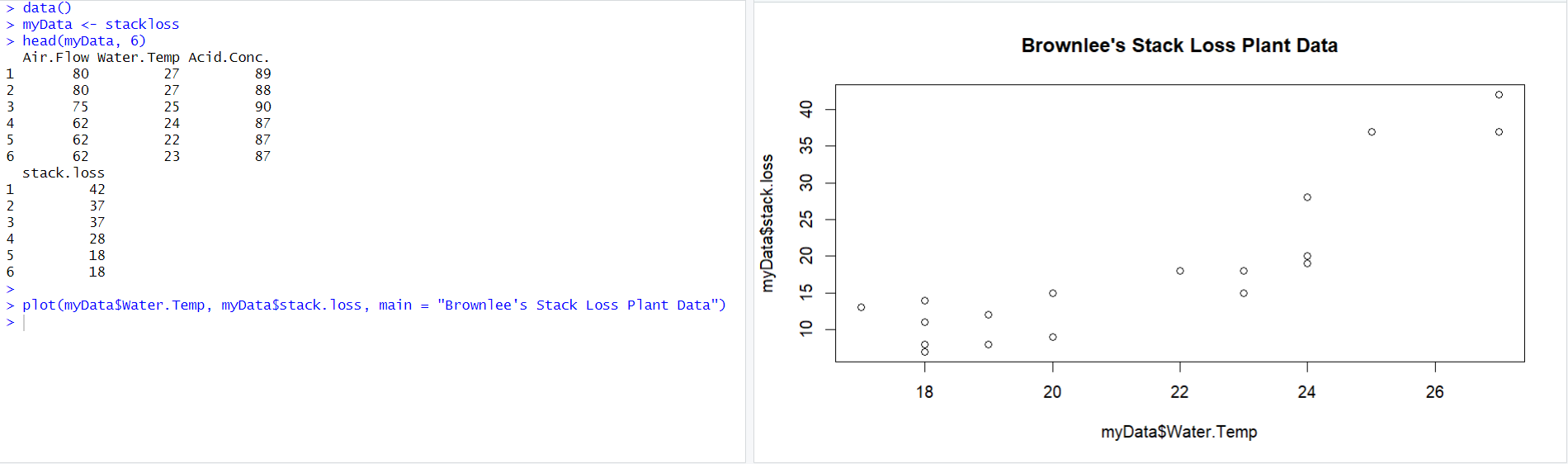
****

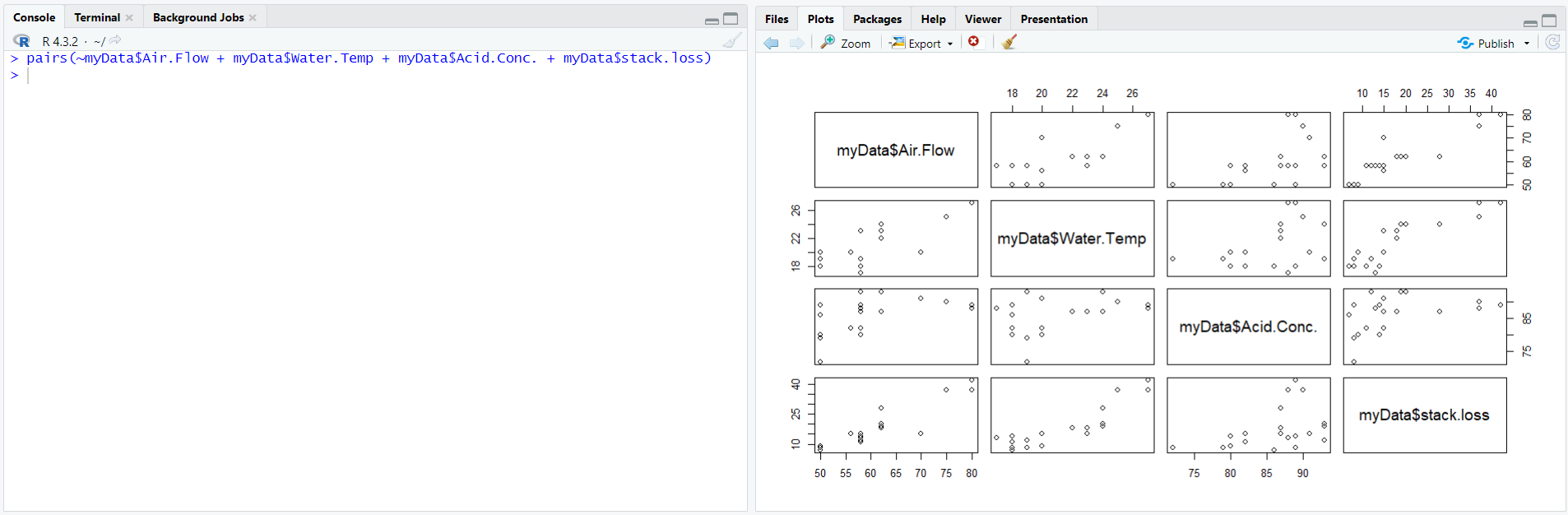
****

****

****

****

****

****

**Post lab Questions:**

1. **What is an appropriate way to visualize a list of the eye colors of 120 people?**
2. **Boxplot**
3. **Pie-chart**
4. **Histogram**
5. **Scatterplot**
6. **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?**
7. **median household income**
8. **mean household income**
9. **3rd quartile of household income**
10. **IQR**
11. **Suppose all household incomes in California increase by 5%. How does that change the median household income?**
12. **median household income goes up by 5%**
13. **the median household income doesn't change**
14. **cannot be determined from the information given**
15. **Suppose all household incomes in California increase by $5,000. How does that change the interquartile range of the household incomes?**
16. **cannot be determined from the information given**
17. **the interquartile range of the household incomes doesn't change**
18. **the interquartile range of the household incomes goes up by $5,000**
19. **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?**
20. **the houses that sold for more than $342,000 represent more than 50% of all sales**
21. **the houses that sold for more than $342,000 represent exactly 50% of all sales**
22. **the houses that sold for more than $342,000 represent less than 50% of all sales**
23. **Suppose all household incomes in California increase by $5,000. How does that change the standard deviation of the household incomes?**
24. **the standard deviation of the household incomes doesn't change**
25. **cannot be determined from the information given**
26. **the standard deviation of the household incomes goes up by $5,000**
27. **Which of the following graphical displays can be used to understand the distribution of data?**
28. **Box Plot**
29. **Quantile-Normal plot**
30. **Histogram**
31. **Scatter Plot**
32. **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**
33. **0.85**
34. **0.87**
35. **0.65**
36. **0.82**