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

**Experiment No.: 2**

|  |
| --- |
| **Title:** To apply the descriptive statistics techniques. |

**Aim:**

To apply various descriptive statistics techniques, such as measures of central tendency, variability, and distribution, to analyse and summarize the key features of a dataset.

**Expected Outcome of Experiment:**

**CO1:** Develop an understanding of data science and business analytics.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

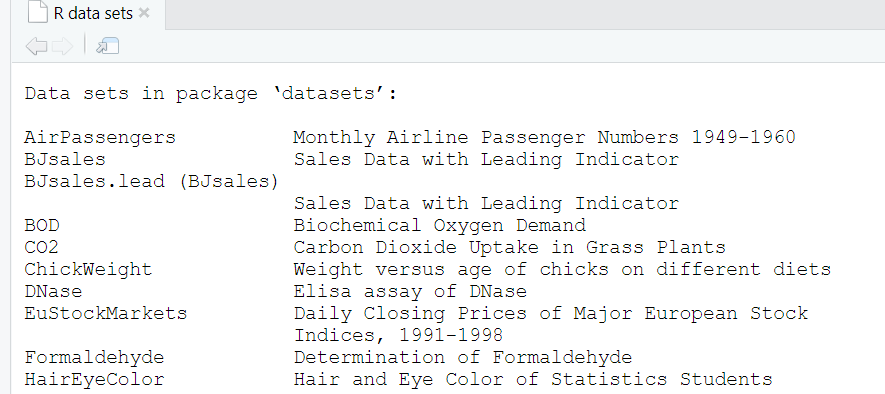
**Books / Journals / Websites Referred:**

1. <https://www.programiz.com/r>

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Select a built-in R dataset:**



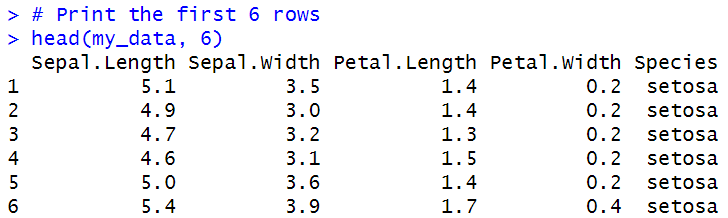


Here, we’ll use the built-in R data set named *iris*. Every student in the batch has to choose a unique dataset.



**Check your data:**

You can inspect your data using the functions **head**() and **tails**(), which will display the first and the last part of the data, respectively.

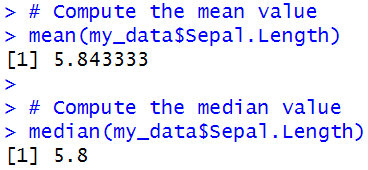


**R functions for computing descriptive statistics:**

|  |  |
| --- | --- |
| **Description** | **R Function** |
| Mean | mean() |
| Standard deviation | sd() |
| Variance | var() |
| Minimum | min() |
| Maximum | maximum() |
| Median | median() |
| Range of values (minimum and maximum) | range() |
| Sample quantiles | quantile() |
| Generic function | summary() |
| Interquartile range | IQR() |

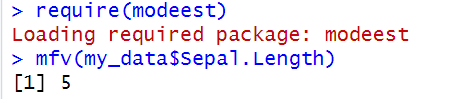
**Descriptive statistics for a single group:**

1. Measure of central tendancy: mean, median, mode –



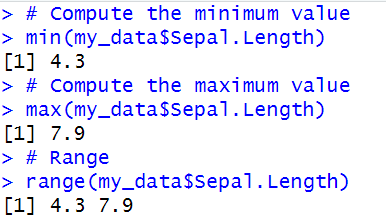
1. The function mfv() [in the modeest R package] can be used to compute the mode of a variable –

****

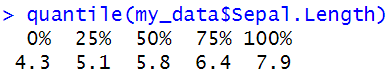
****

**Measure of Variability:**

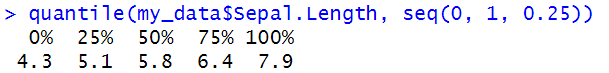
1. Range: minimum & maximum –

****

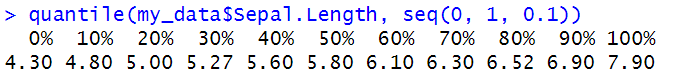
1. Quantiles **–**



By default, the function returns the minimum, the maximum and three **quartiles** (the 0.25, 0.50 and 0.75 quantiles).



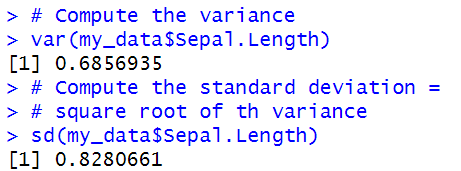
To compute deciles (0.1, 0.2, 0.3, …., 0.9), use this:



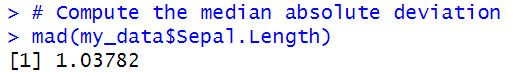
1. Interquartile Range –



1. Variance and Standard Deviation –



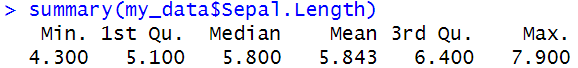
1. Median Absolute Deviation –



**Computing an overall summary of a variable and an entire data frame:**

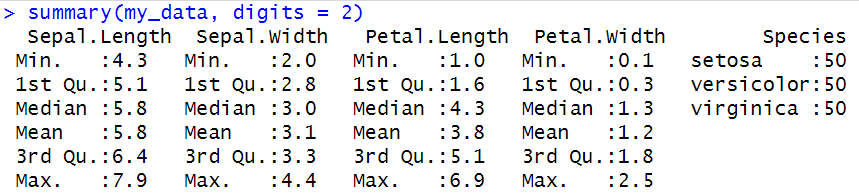
1. summary() function –

**Summary of a single variable.** Five values are returned: the mean, median, 25th and 75th quartiles, min and max in one single line call:



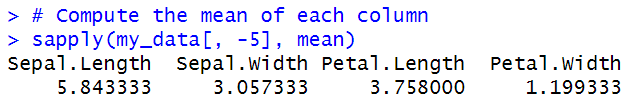
**Summary of a data frame.** In this case, the function summary() is automatically applied to each column. The format of the result depends on the type of the data contained in the column. For example:

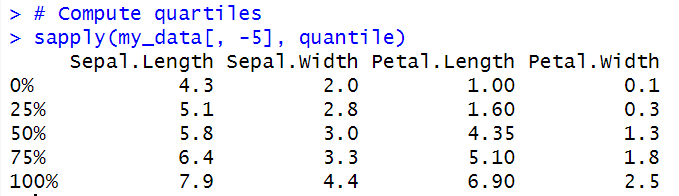
* If the column is a numeric variable, mean, median, min, max and quartiles are returned.
* If the column is a factor variable, the number of observations in each group is returned.



1. saaply() funtction –

It’s also possible to use the function sapply() to apply a particular function over a list or vector. For instance, we can use it to compute for each column in a data frame, the mean, sd, var, min, quantile, …





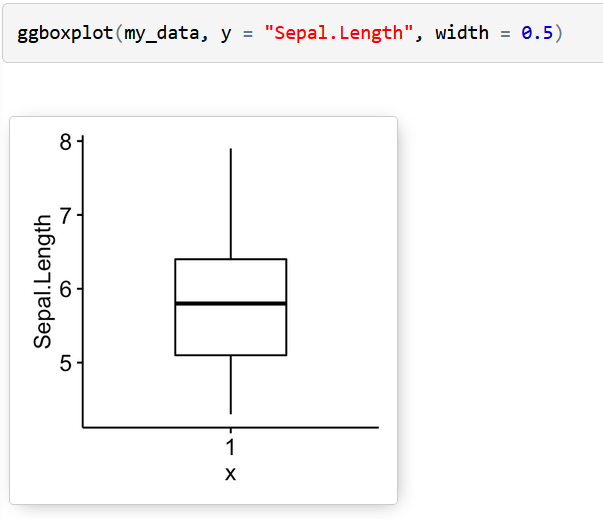
**Graphical Display of Disctibutions:**

The R package **ggpubr** will be used to create graphs.

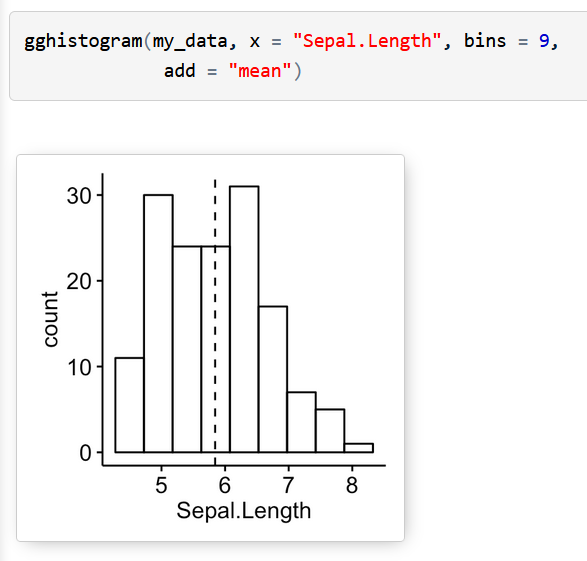




1. Box Plot –

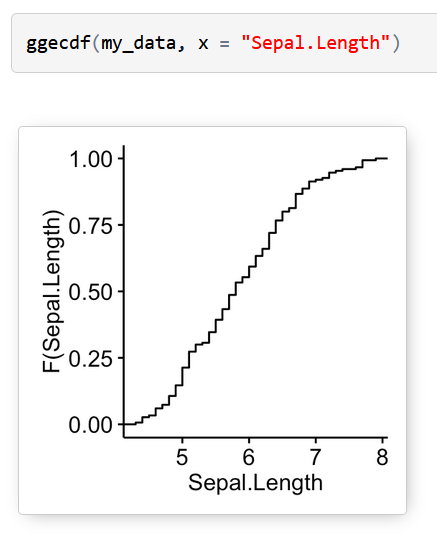


1. Histogram with mean line –

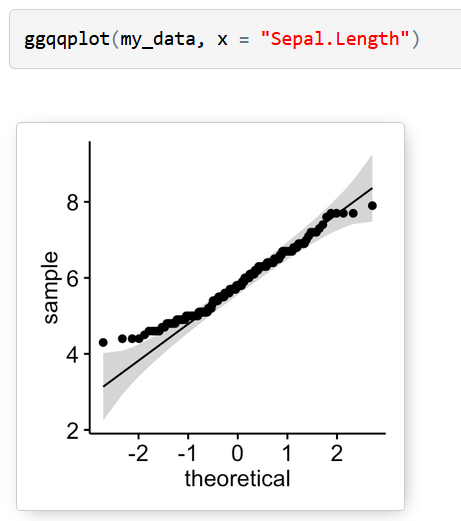


1. Empirical cumulative distribution function (ECDF) –

ECDF is the fraction of data smaller than or equal to x.



1. QQ plots are used to check whether the data is normally distributed –

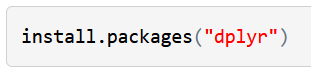


**Descriptive statistics by groups:**

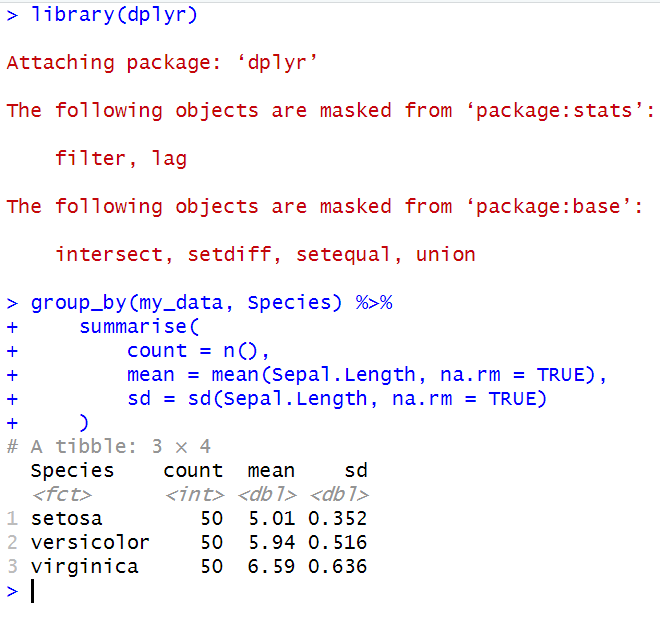
To compute summary statistics by groups, the functions **group\_by**() and **summarise**() [in **dplyr** package] can be used.

* We want to group the data by *Species* and then:
  + compute the number of element in each group. R function: **n**()
  + compute the mean. R function **mean**()
  + and the standard deviation. R function **sd**()

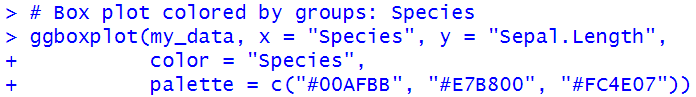
Install **ddplyr** as follow:

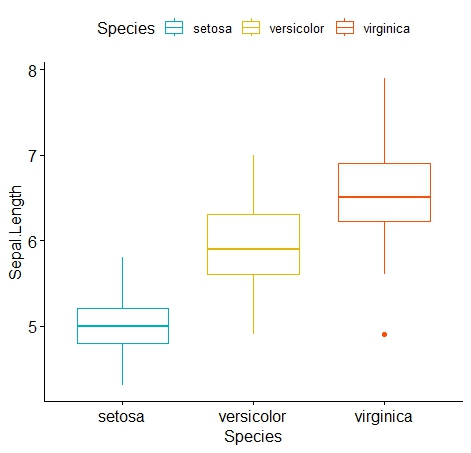


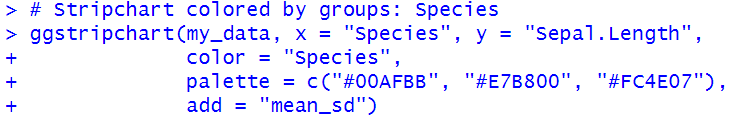
**%>%** is used to chain the operations.

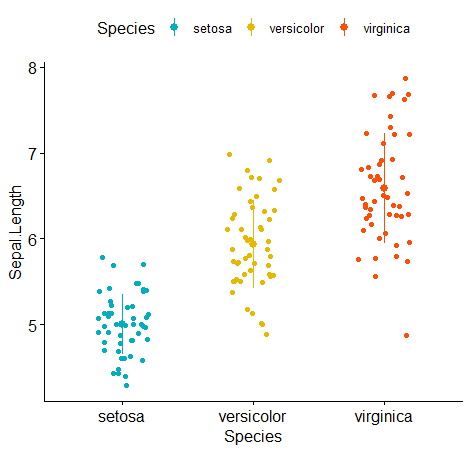


Graphics for grouped data:







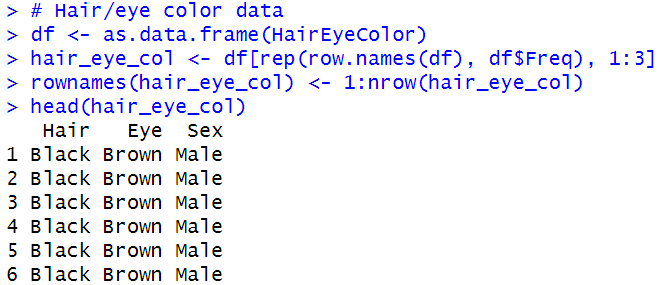


**Frequency Tables:**

A frequency table (or contingency table) is used to describe categorical variables. It contains the counts at each combination of factor levels.

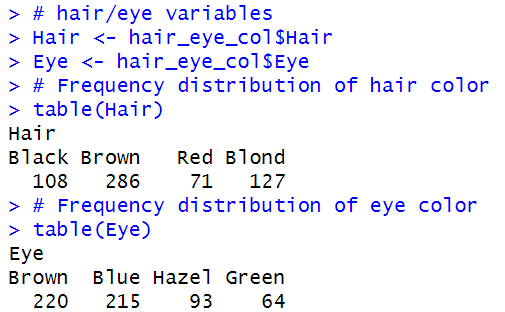
R function to generate tables: **table**()

For this section we will use the built-in R dataset that contains the distribution of hair and eye color by sex of 592 students:

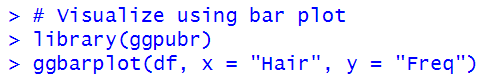


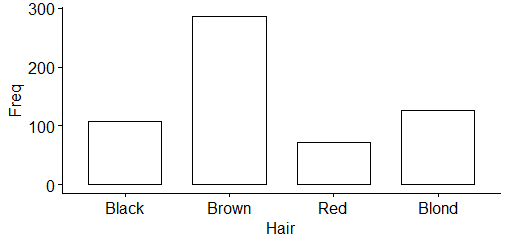
1. Simple frequency distribution: one categorical variable –

Table of counts:

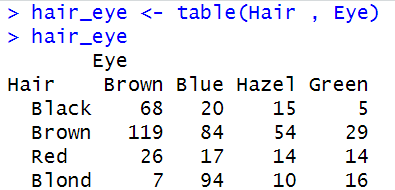


Visualization:



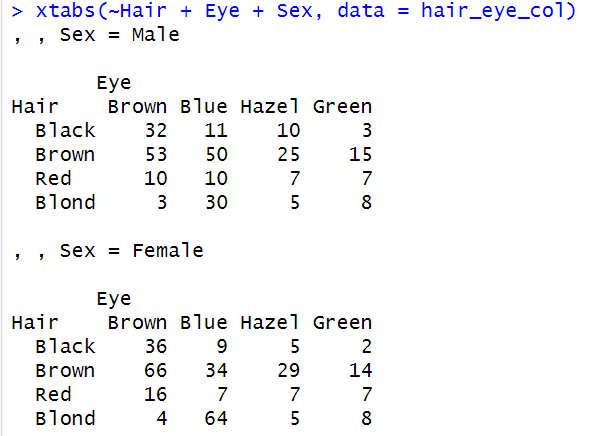


1. Two-way contingency table: Two categorical variables –

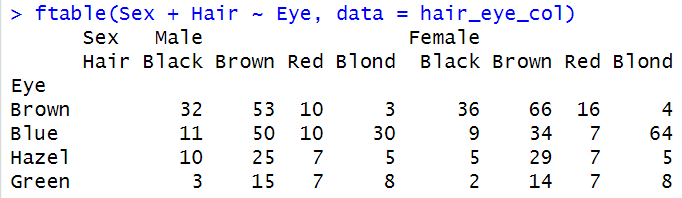


1. Multiway tables: More than two categorical variables –

Hair and Eye color distributions by sex using **xtabs**():

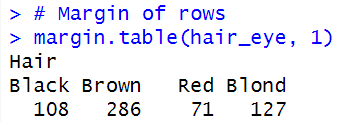


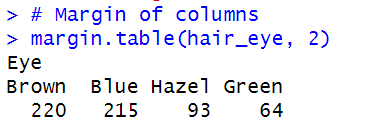
You can also use the function ftable() [for flat contingency tables]. It returns a cleaner looking output compared to xtabs() when you have more than two variables:



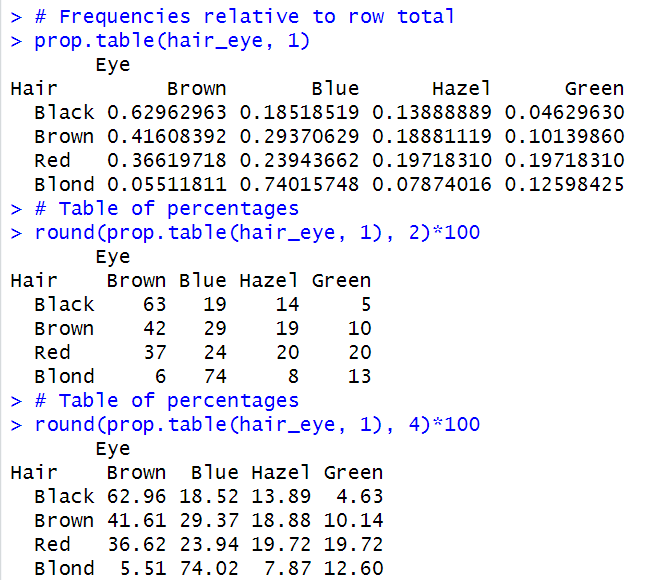
**Compute table margins and relative frequency:**

1. **Table margins** correspond to the sums of counts along rows or columns of the table.





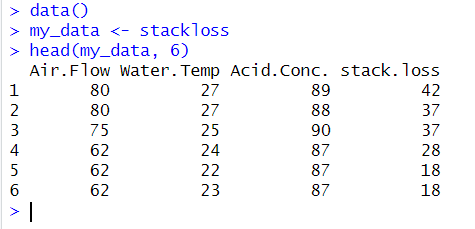
1. **Relative frequencies** express table entries as proportions of table margins (i.e., row or column totals).

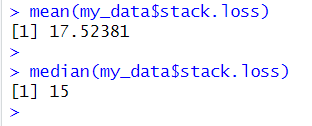


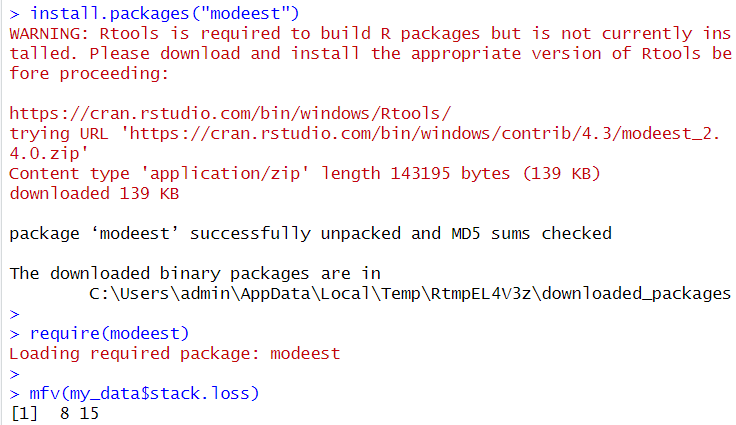
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

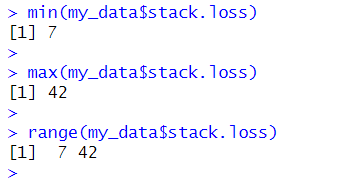
**Implementation:**

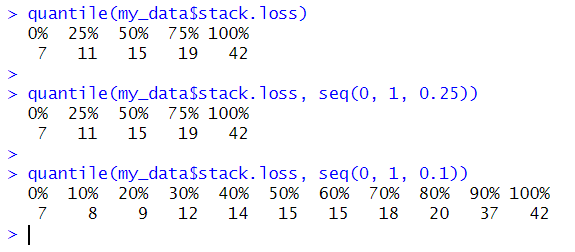
Students have to perform **ALL** the operations shown above on the dataset of their choice and add their screenshots here.



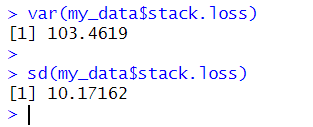


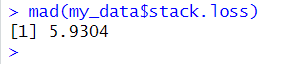


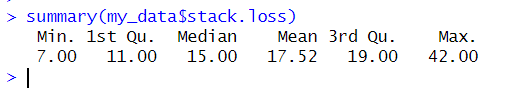


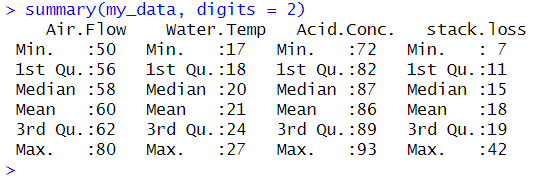


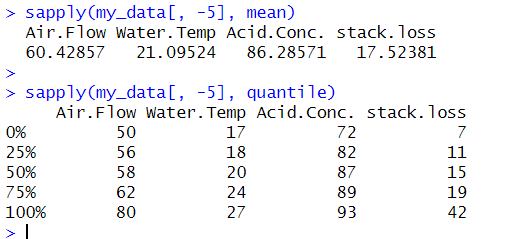


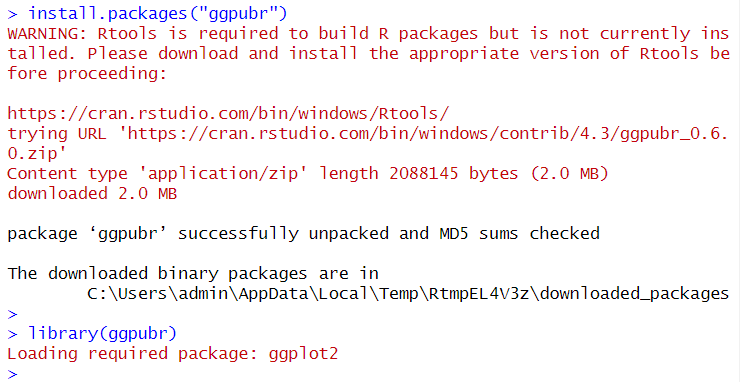




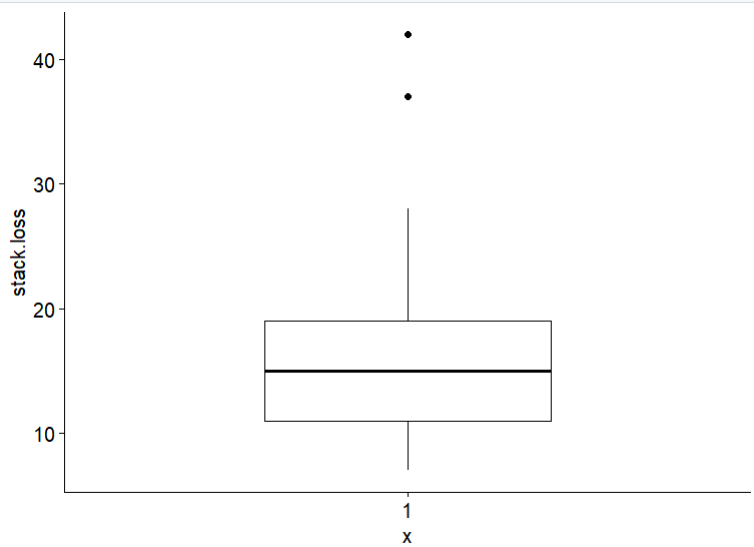




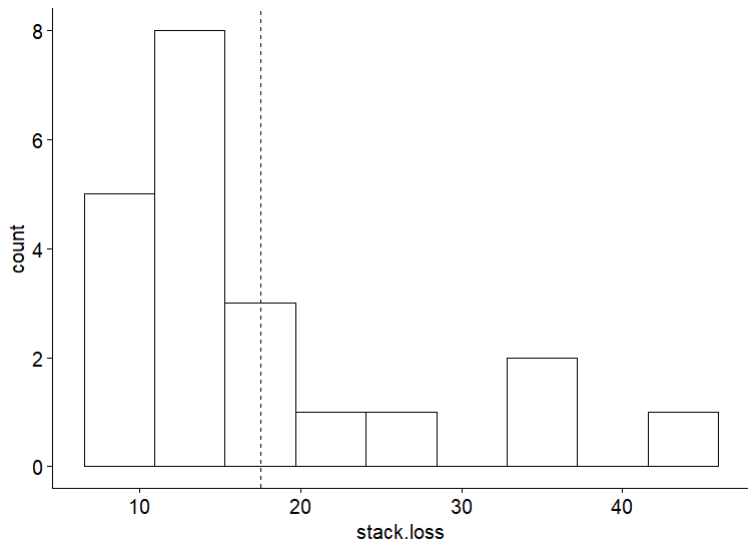




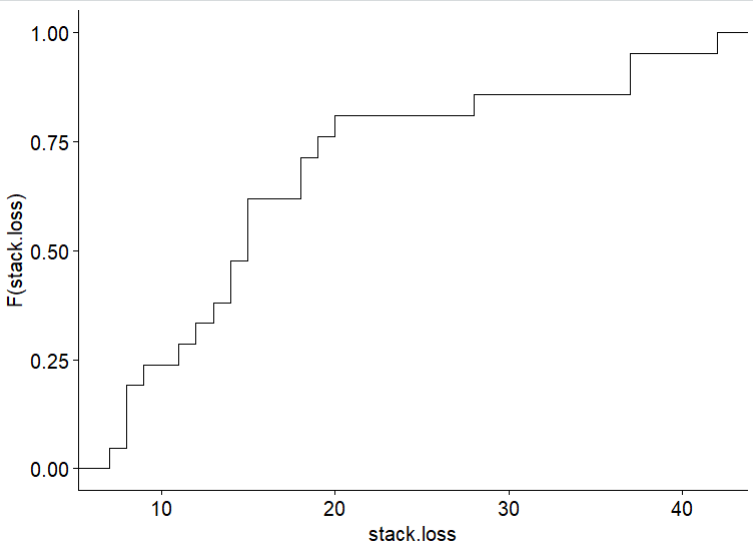




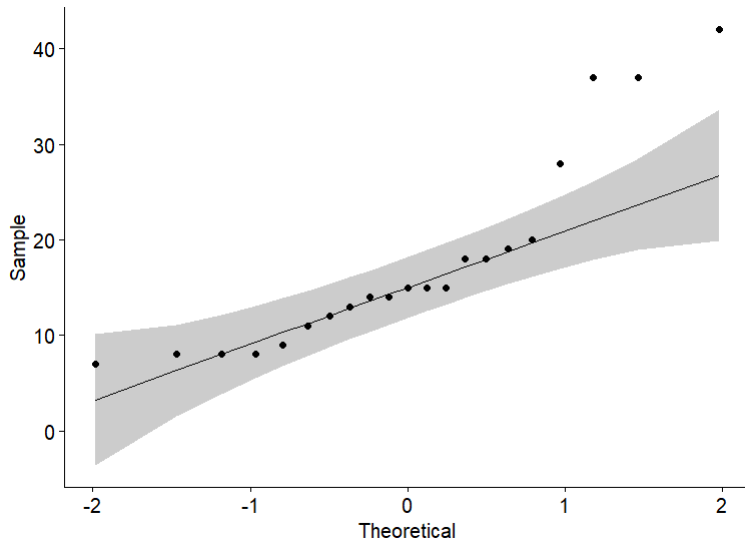


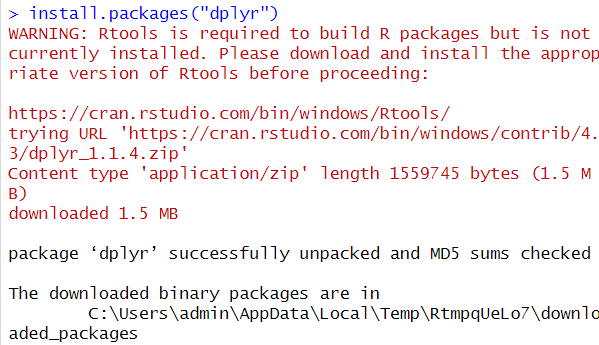


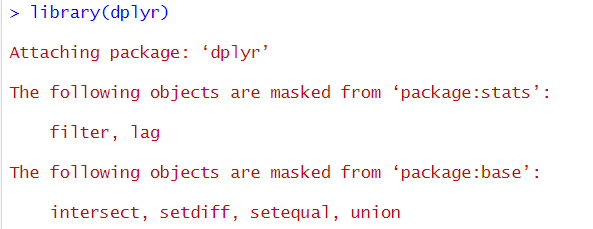




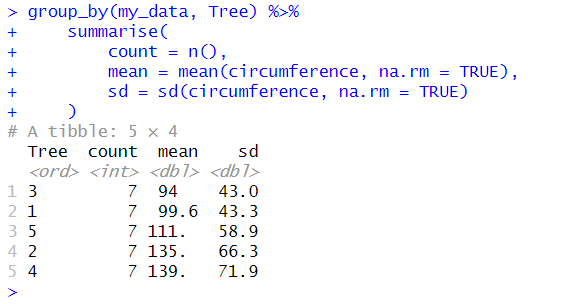


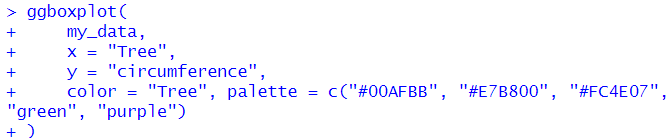


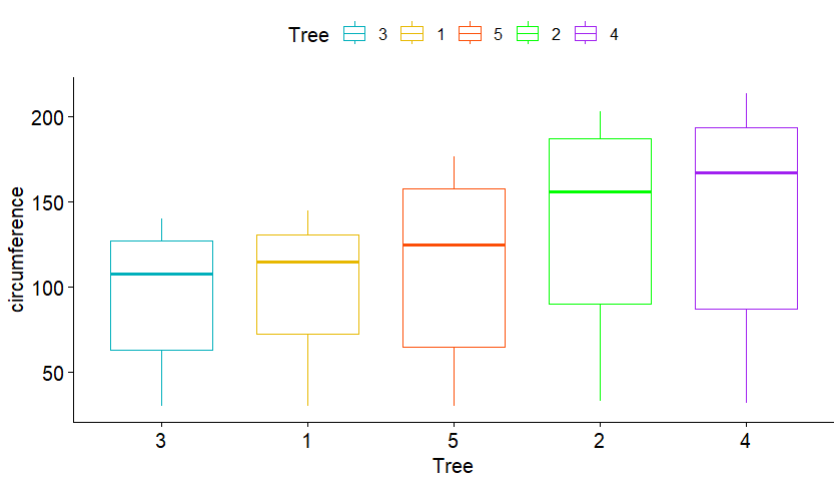


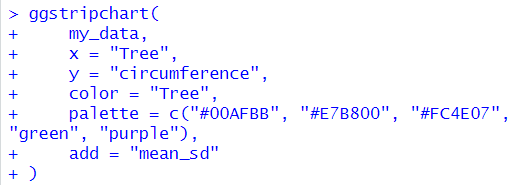


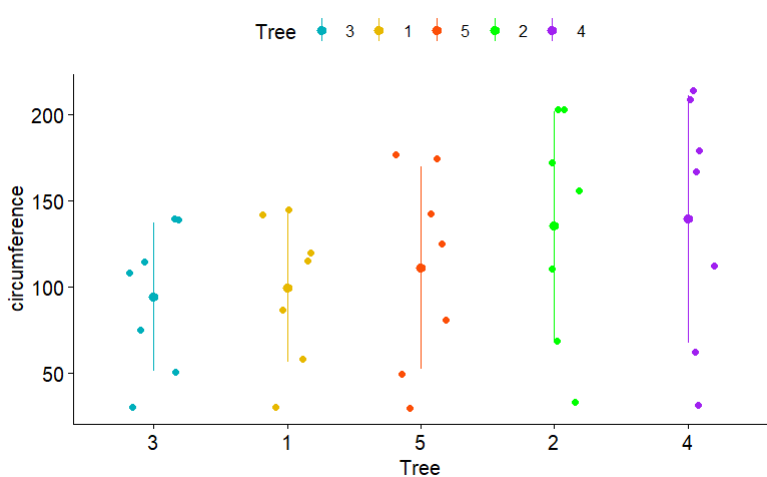
(Changed data to orange dataset to create graphs based on groupings)



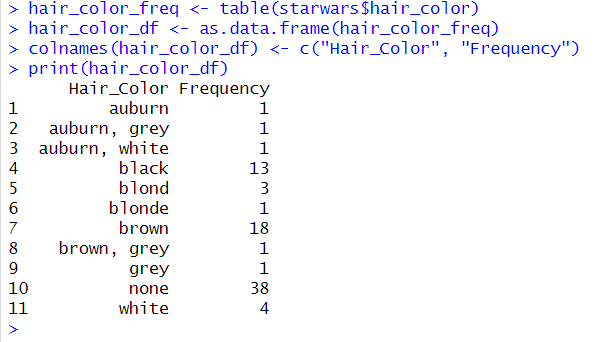


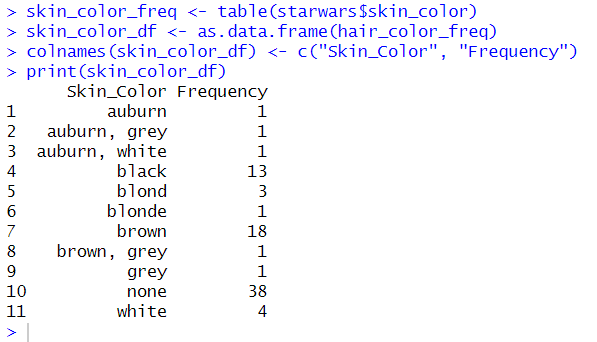


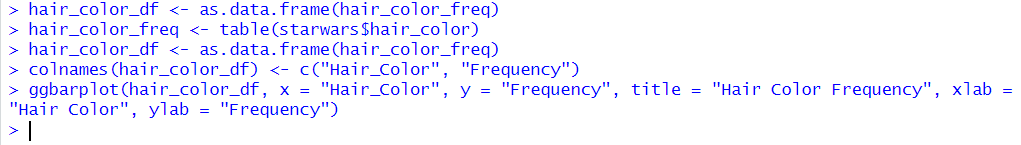


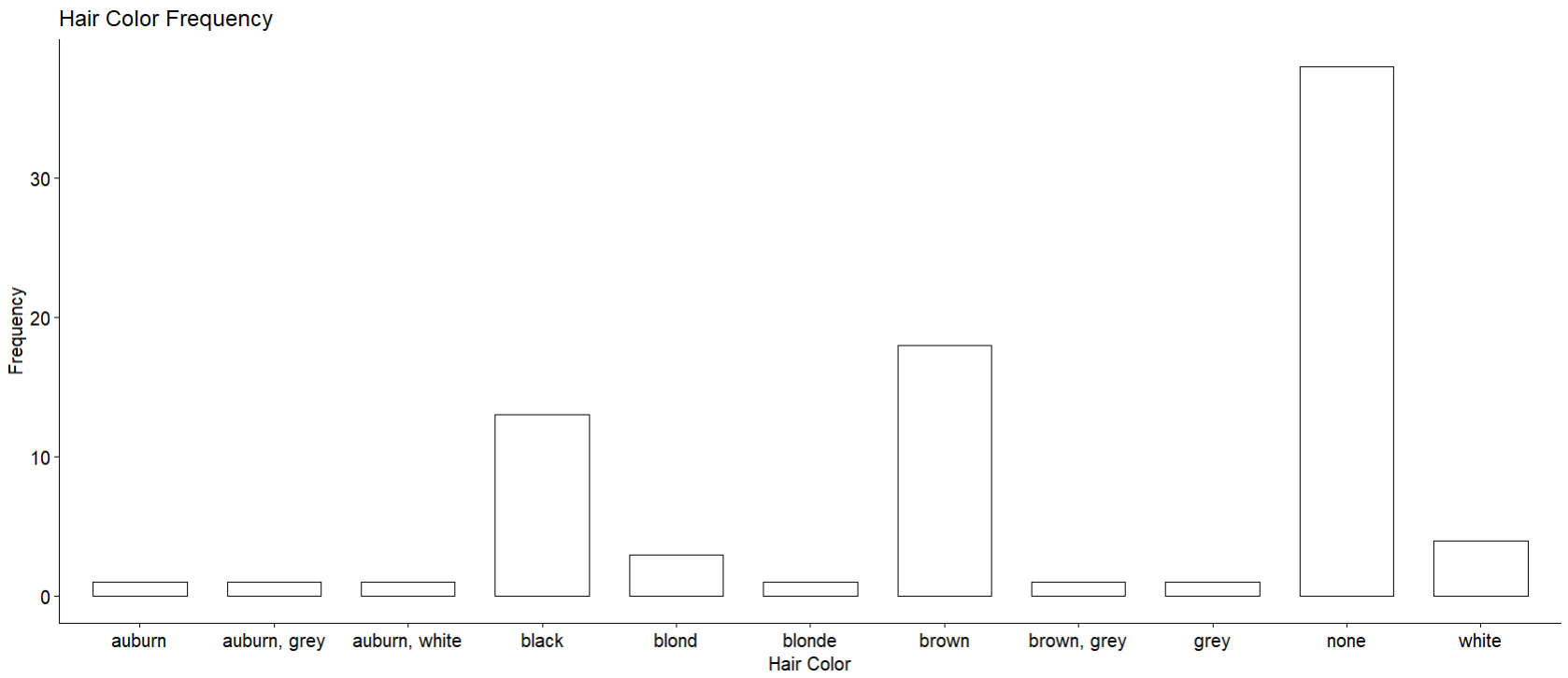


(dataset used is starwars from dplyr library)

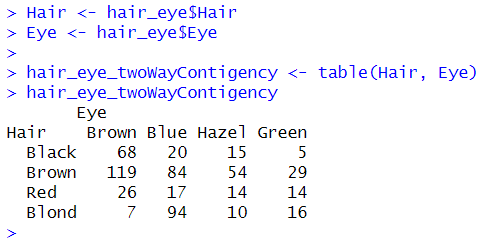


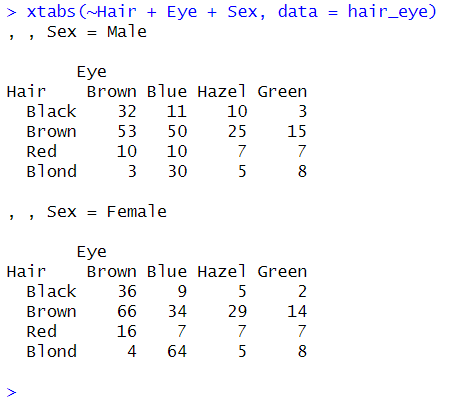


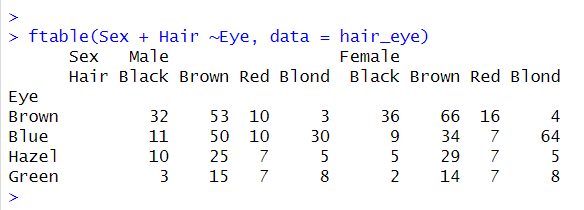


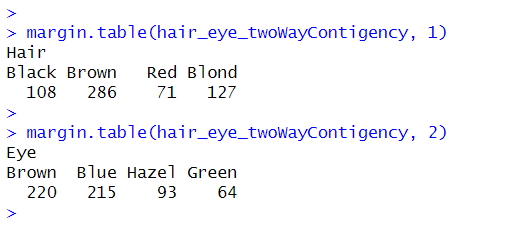


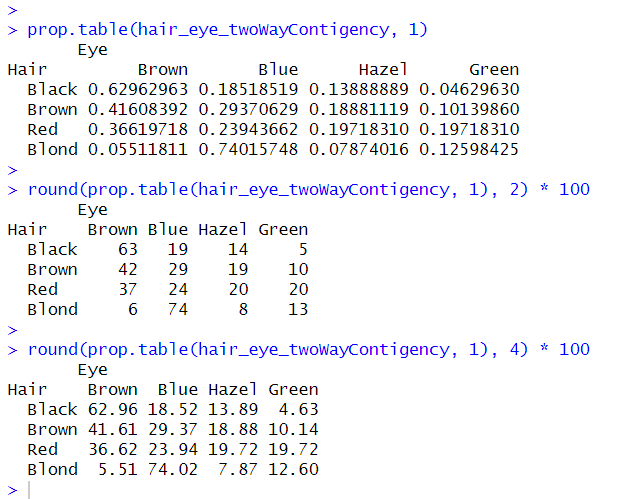
(dataset used is HairEyeColor to create Two-Way Contingency and Multiway tables)











**Post Lab Question:**

1. **Critically assess the limitations of using only measures of central tendency in data analysis.**

* Neglect of Distribution Shape:

Measures of central tendency (mean, median, mode) provide a summary of the center of the data but do not convey information about the distribution's shape. Two datasets with the same mean can have very different distributions.

* Sensitive to Outliers: The mean, in particular, is sensitive to extreme values (outliers). A few outliers can significantly impact the mean, leading to a distorted representation of the typical values in the dataset.
* Loss of Individual Variation: Central tendency measures do not provide insights into the individual variation within the dataset. Data points may vary widely around the central value, and this information is lost when focusing only on measures of central tendency.

1. **Compare and contrast the different measures of variability, with the focus on when one measure might be more informative than the other.**

* Range: The range is the simplest measure of variability, representing the difference between the maximum and minimum values. It is sensitive to outliers but provides a quick overview of the spread.
* Variance and Standard Deviation: Variance and standard deviation quantify the average deviation of each data point from the mean. They are more robust than the range but can be influenced by outliers.
* Interquartile Range (IQR): IQR is less sensitive to outliers than the range and provides a measure of the spread within the central 50% of the data. It is useful when extreme values are present.
* Coefficient of Variation (CV): CV expresses the standard deviation as a percentage of the mean, allowing for comparison of variability between datasets with different scales.

The choice of a measure depends on the characteristics of the dataset and the specific goals of the analysis. For example, if outliers are a concern, IQR or robust measures may be preferred.

1. **Imagine you are presented with a dataset from a research study. Discuss how applying descriptive statistics techniques could aid in understanding the key features and trends in the data. Take any real-life examples to aid your analysis.**

Example: Let's consider a dataset of household incomes in a country. Descriptive statistics can provide insights:

* Measures of Central Tendency: Mean income can give an average income level, and the median can indicate the income level at which 50% of households fall below.
* Measures of Variability: Standard deviation or IQR can reveal how spread out the incomes are, providing information on income inequality.
* Distribution Shape: A histogram or density plot can visually represent the distribution of incomes, identifying if it is skewed, bimodal, or normally distributed.
* Outliers: Identifying outliers can be crucial, as extremely high or low incomes may impact policy decisions or social programs.

Descriptive statistics, in this case, aid in summarizing, visualizing, and interpreting the key features and trends in the income data, contributing to a more comprehensive understanding for decision-making and policy formulation.

**Conclusion:**

In this experiment, we used different ways to study and understand a set of data. We looked at things like average values, how spread out the data is, and the overall shape of the information. By doing this, we gained a detailed picture of the dataset, helping us uncover important patterns and features within the data. This thorough exploration using various statistical techniques allowed us to make better sense of the information at hand.