**Batch: H2\_3**

**Roll No.: 16010122221**

**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 analyze and summarize the key features of a dataset.

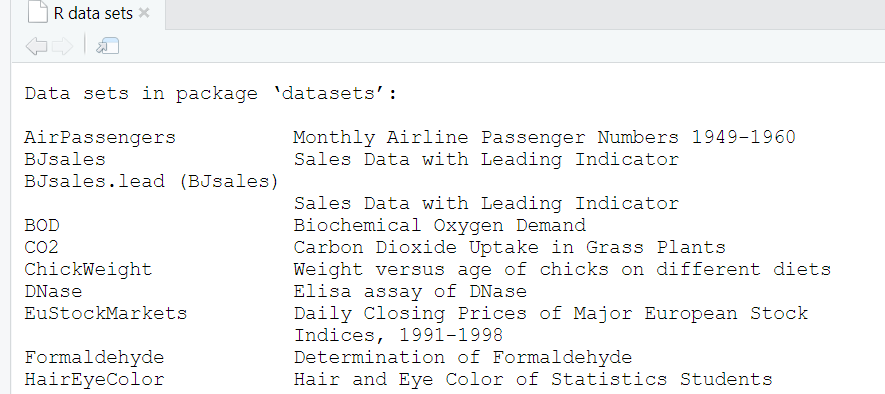
# Expected Outcome of Experiment:

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

# Books/ Journals/ Websites referred:

**Select a built-in R dataset**

You can see a list of all the built-in datasets using the data() function.

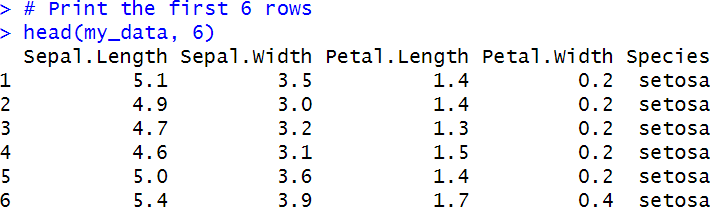


Here, we’ll use the built-in R data set named *iris*. Every student in the batch has to choose a unique dataset.



# 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.



# To find number or row in data

**>nrow(my\_data) What will be output of**

**>ncol(my\_data)**

# R functions for computing descriptive statistics:

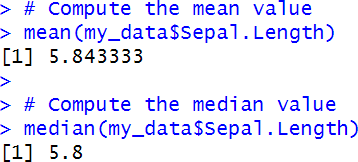
|  |  |
| --- | --- |
| **Description** | **R function** |
| **Mean** | **mean**() |
| **Standard deviation** | **sd**() |
| **Variance** | **var**() |
| **Minimum** | **min**() |
| **Maximum** | **maximum**() |
| **Median** | **median**() |
| **Range of values** (minimum and maximum) | **range**() |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

**Descriptive statistics for a single group Measure of central tendency:** mean, median, mode

**Mean:** Mean is nothing but the average of the given set of values.

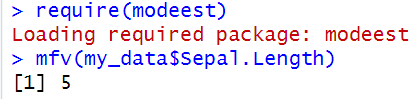
**Median:** The median of a set of data is the middlemost number or centre value in the set. The median is also the number that is halfway into the set. To find the median, the data should be arranged first in order of least to greatest or greatest to the least value.

**Mode:** The mode is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.



# The function mfv() [in the modeest R package] can be used to compute the mo de of a variable.





**Measure of variability**

# Range: minimum & maximum

**Quantiles**

[A quantile is a particular part of a data set that determines how many values in a distribution](https://www.statista.com/statistics-glossary/definition/356/quantile/) [are above or below a certain limit](https://www.statista.com/statistics-glossary/definition/356/quantile/). [Quantiles are cut points that divide the range of a probability](https://en.wikipedia.org/wiki/Quantile) [distribution into continuous intervals with equal probabilities, or divide the observations in a](https://en.wikipedia.org/wiki/Quantile) [sample in the same way.](https://en.wikipedia.org/wiki/Quantile) [The word "quantile" comes from the word quantity, and it refers to](https://www.statisticshowto.com/quantile-definition-find-easy-steps/) [dividing a sample or a probability distribution into equal-sized, adjacent subgroups](https://www.statisticshowto.com/quantile-definition-find-easy-steps/).

So given data set is arranged in the ascending order. Assume there are 100 samples; then, 25% means THE value of 25th sample; in other words the value below which 25% of the samples lie.



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:

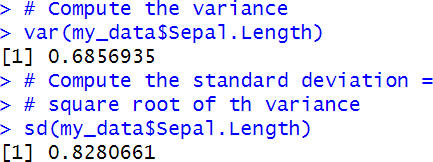


# Interquartile range

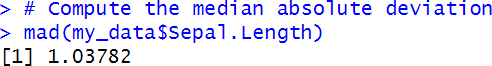
The difference between the **upper and lower quartile** is known as the interquartile range.



# Variance and standard deviation



**Median absolute deviation**



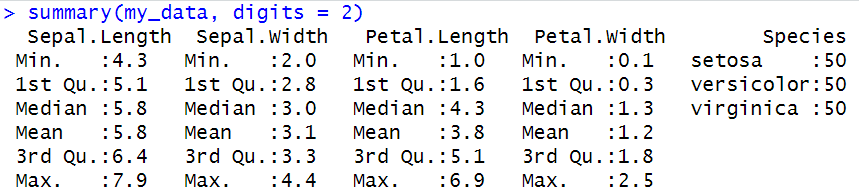
# Computing an overall summary of a variable and an entire data frame summary() function

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



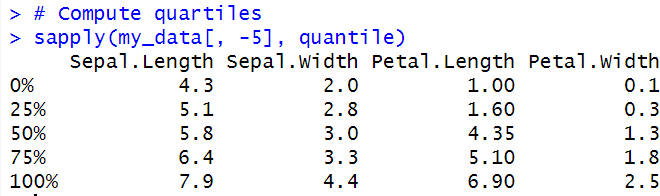
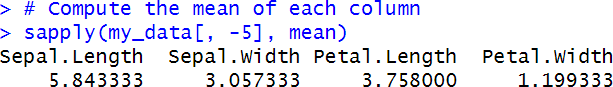
**Summary of a data frame:** In this case, the function **summary**() is automatically applied to each column. The format of the result depends on the type of the data contained in the column. For example:

* 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.



**sapply() function**

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, …

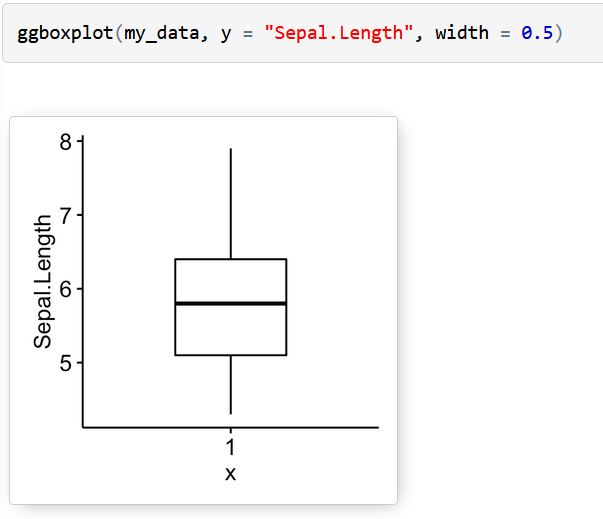


# Graphical display of distributions:

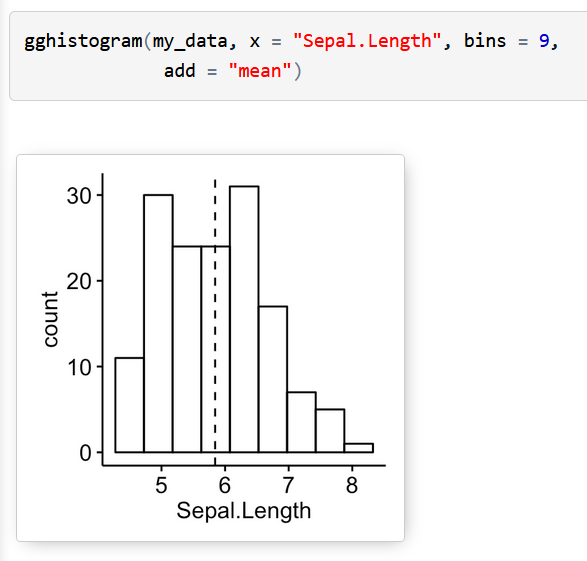
The R package **ggpubr** will be used to create graphs.



## Box Plot

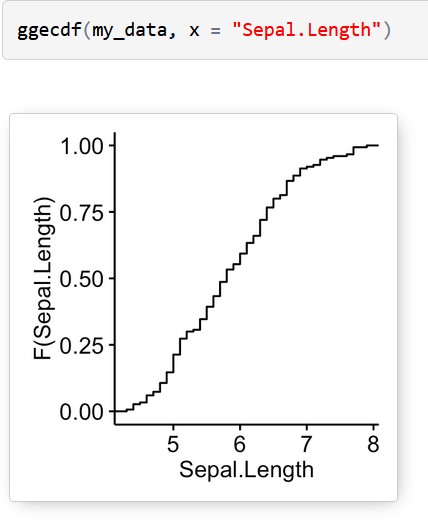


Histogram with mean line



# Empirical cumulative distribution function (ECDF)

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



# 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:



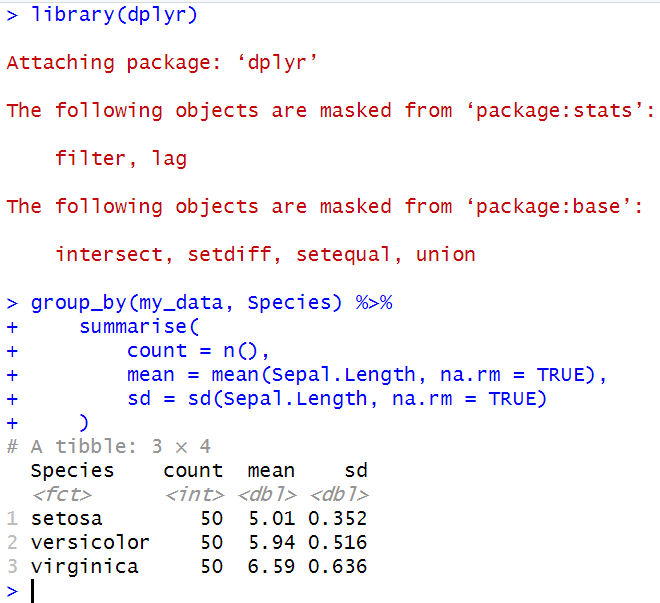
**Descriptive statistics by groups:**

To compute summary statistics by groups, the functions **group\_by**() and **summarise**() [in

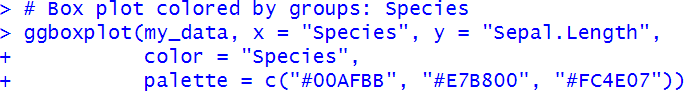
**dplyr** package] can be used.

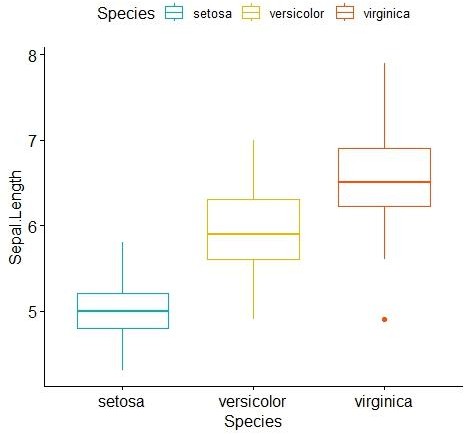
* We want to group the data by Species and then:
  + compute the number of element in each group. R function: **n**()
  + compute the mean. R function **mean**()
  + and the standard deviation. R function **sd**()

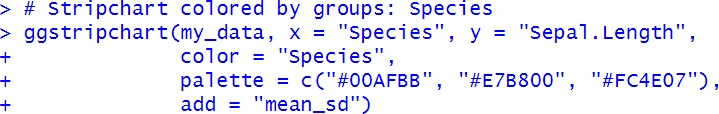
**%>%** is used to chain the operations.

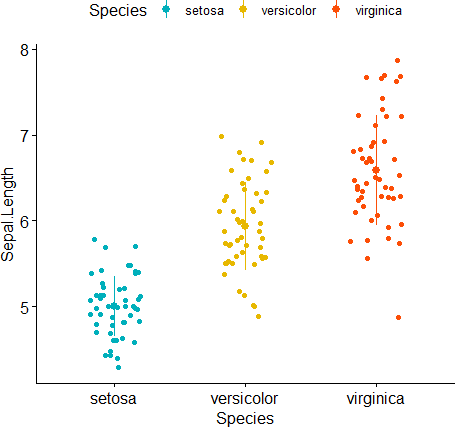


**Graphics for grouped data:**







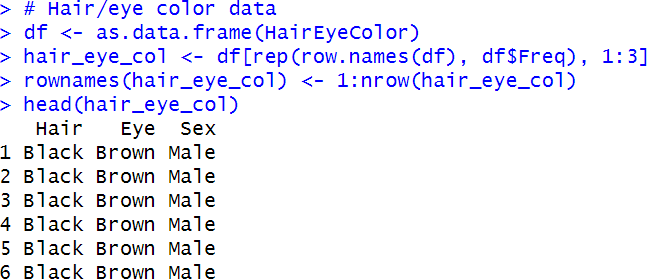


# Frequency tables:

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

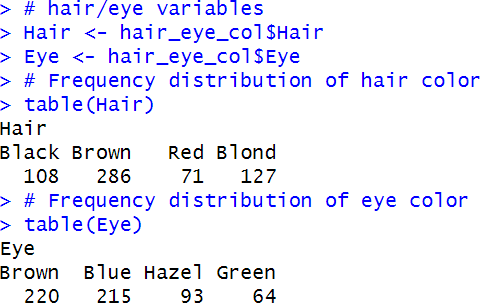
R function to generate tables: **table**()

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:

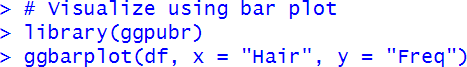


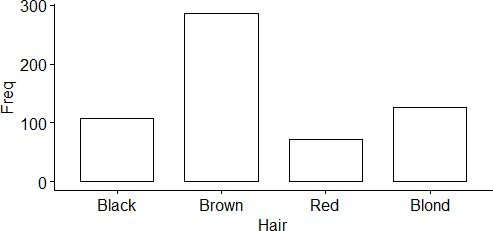
# Simple frequency distribution: one categorical variable

## Table of counts:



**Visualization:**

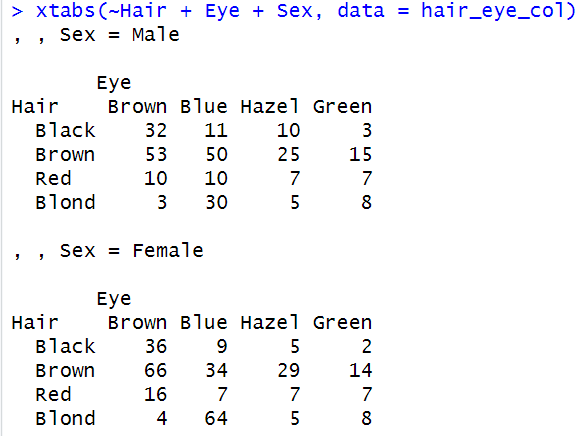




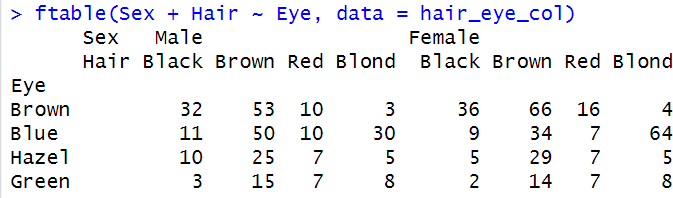
# Two-way contingency table: Two categorical variables

**Multiway tables: More than two categorical variables**

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

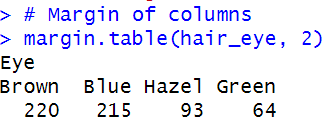
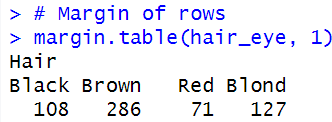


You can also use the function **ftable**() [for flat contingency tables]. It returns a cleaner looking output compared to xtabs() when you have more than two variables:

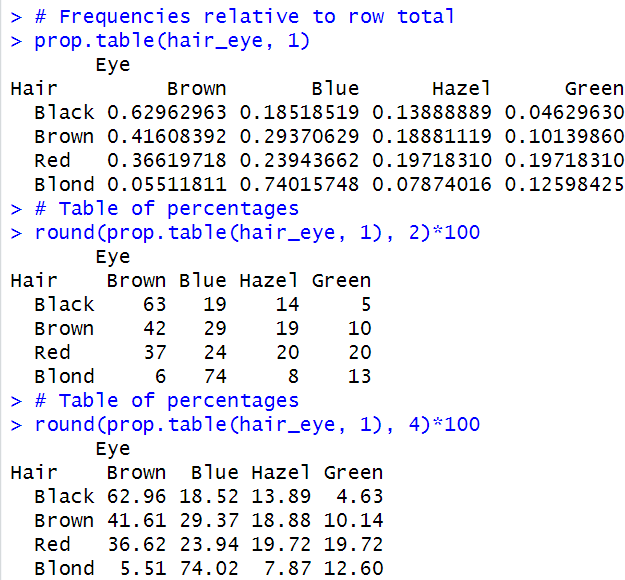


# Compute table margins and relative frequency

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



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



# EXECUTION:

**CODE:**

my\_data <- trees print(my\_data) nrow(my\_data) ncol(my\_data) typeof(my\_data)

#Mean mean(my\_data$Girth)

#Median median(my\_data$Height)

#Mode install.packages("modeest") require(modeest) mfv(my\_data$Volume)

#Min min(my\_data$Height)

#Max max(my\_data$Height)

#Range range(my\_data$Girth)

#Quantiles quantile(my\_data$Girth) quantile(my\_data$Height) quantile(my\_data$Volume)

#Interquartile Range IQR(my\_data$Girth)

#Variance & Standard Deviation var(my\_data$Height) sd(my\_data$Volume)

#Median Absolute Deviation mad(my\_data$Girth)

#Summary summary(my\_data$Volume) summary(my\_data, digits = 2)

#Sapply

sapply(my\_data[, -5], mean)

sapply(my\_data[, -1], quantile)

#Box Plot

ggboxplot(my\_data, y = "Girth", width = 0.5)

#Histogram

gghistogram(my\_data, x = "Height", bins = 12, add = "mean")

#ecdf

ggecdf(my\_data, x = "Volume")

#qqplot

ggqqplot(my\_data, x = "Height")

#Descriptive Statistics dplyr::group\_by(my\_data, Girth) %>% dplyr::summarise(

count = n(),

mean = mean(Girth, na.rm = TRUE), sd = sd(Girth, na.rm = TRUE)

)

#Box Plot colored by groups my\_data1 <- iris print(my\_data1)

ggstripchart(my\_data1, x = "Species", y = "Sepal.Length", color = "Species", palette = c("#00AFBB", "#E7B800", "#FC4E07"))

#Frequency

my\_data2 <- HairEyeColor

df <- as.data.frame(HairEyeColor)

hair\_eye\_col <- df[rep(row.names(df), df$Freq), 1:4] rownames(hair\_eye\_col) <- 1:nrow(hair\_eye\_col) head(hair\_eye\_col)

#Table of counts of categorical variables Hair <- table(hair\_eye\_col$Hair)

Eye <- table(hair\_eye\_col$Eye) print(Hair)

print(Eye)

#Visualization of the the counts ggbarplot(df, x = "Hair", y = "Freq")

#Two-way contingency table

hair\_eye <- table(hair\_eye\_col$Hair, hair\_eye\_col$Eye) hair\_eye

#Multiway Tables

xtabs(~Hair + Eye + Sex, data = hair\_eye\_col)

#Flat Contingency Table

ftable(Sex + Hair ~ Eye, data = hair\_eye\_col)

#Table Margins margin.table(hair\_eye, 1)

margin.table(hair\_eye, 2)

#Relative Frequencies of row total prop.table(hair\_eye, 1)

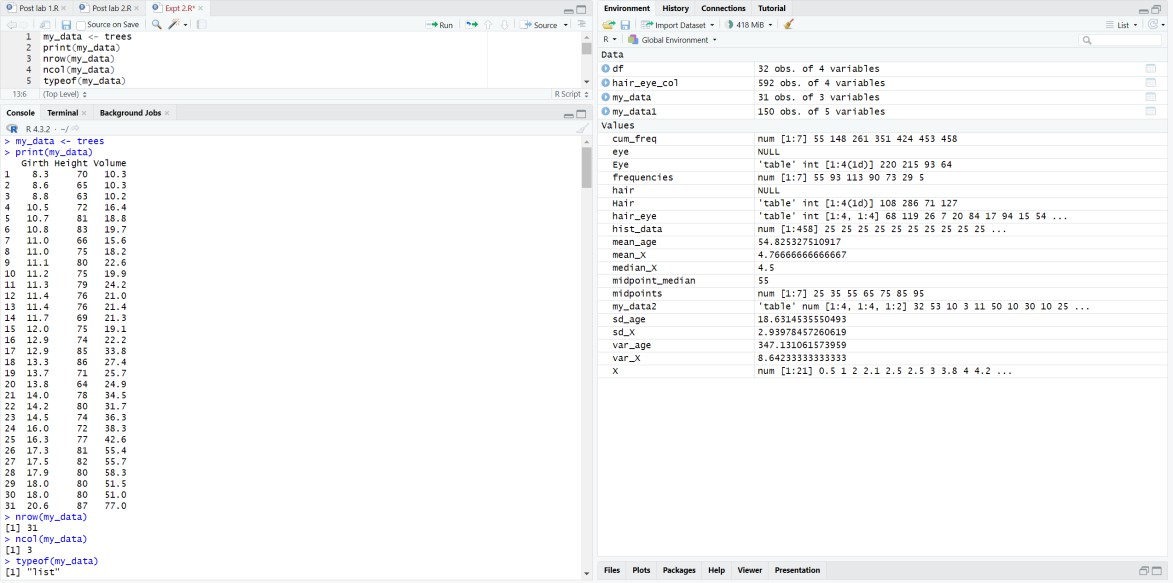
#Table of Percentages

hair\_eye <- table(hair\_eye\_col$Hair, hair\_eye\_col$Eye) round(prop.table(hair\_eye, 1), 2)\*100

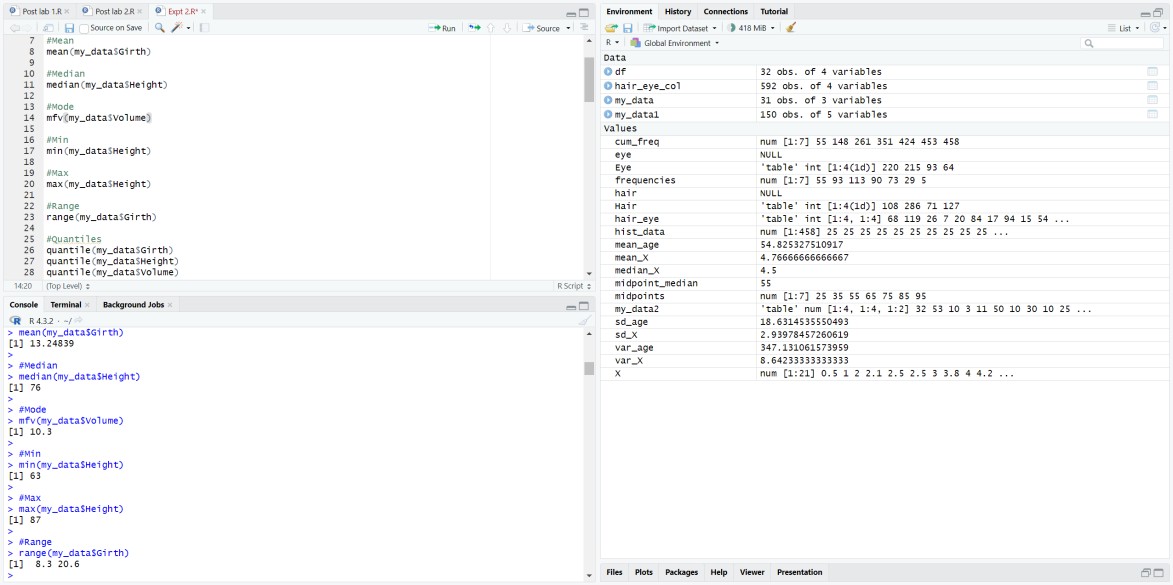
round(prop.table(hair\_eye, 2), 4)\*100

# EXECUTION SCREENSHOTS:

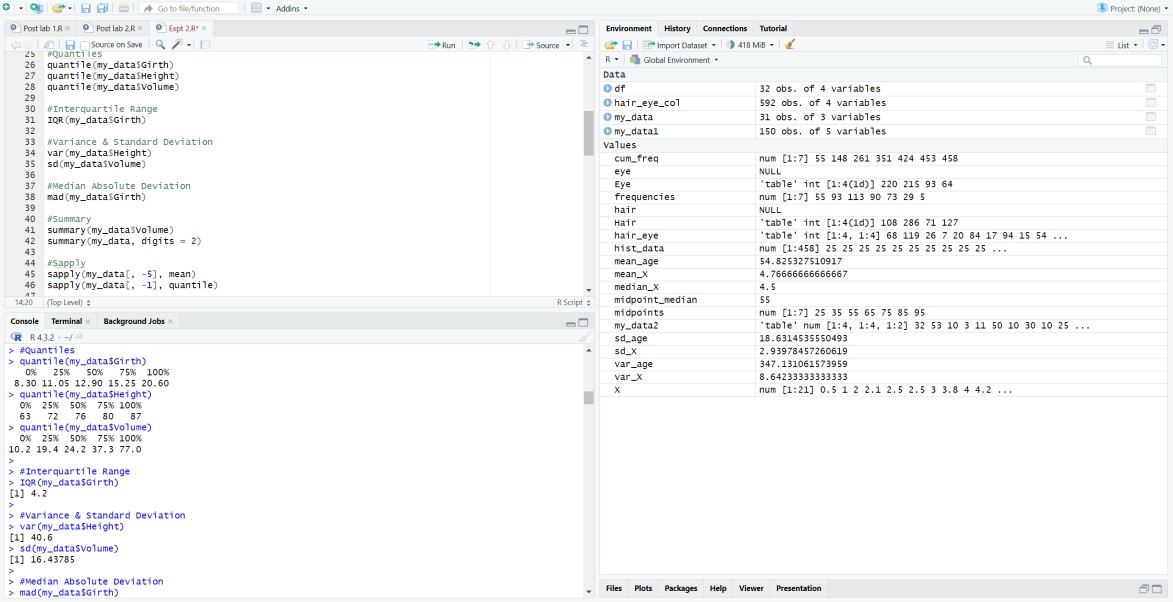
**1)**



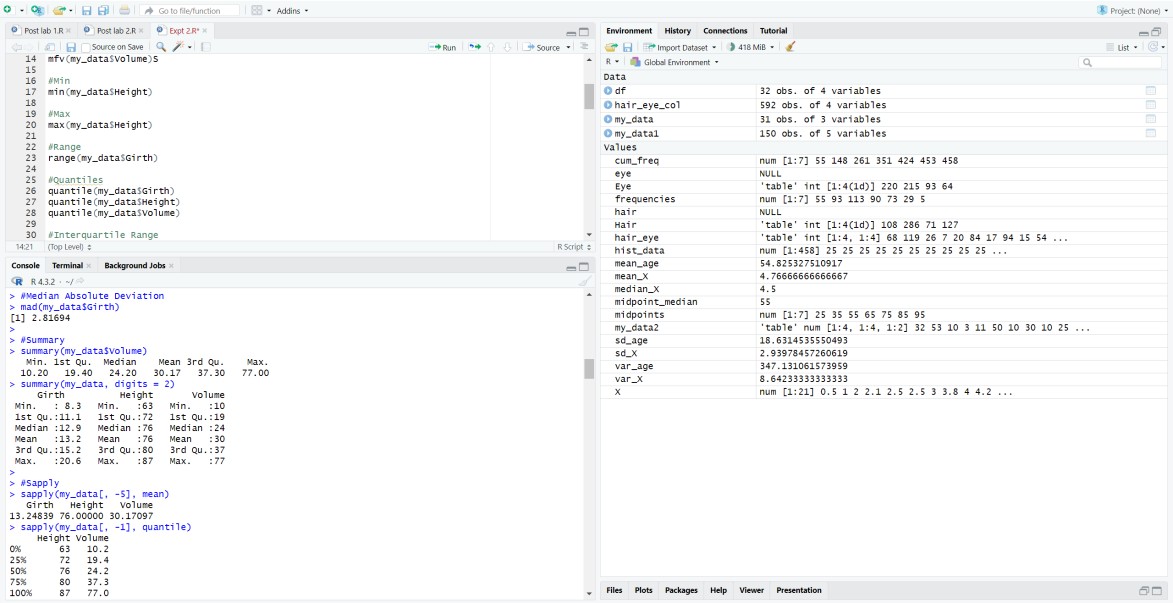
# 2)



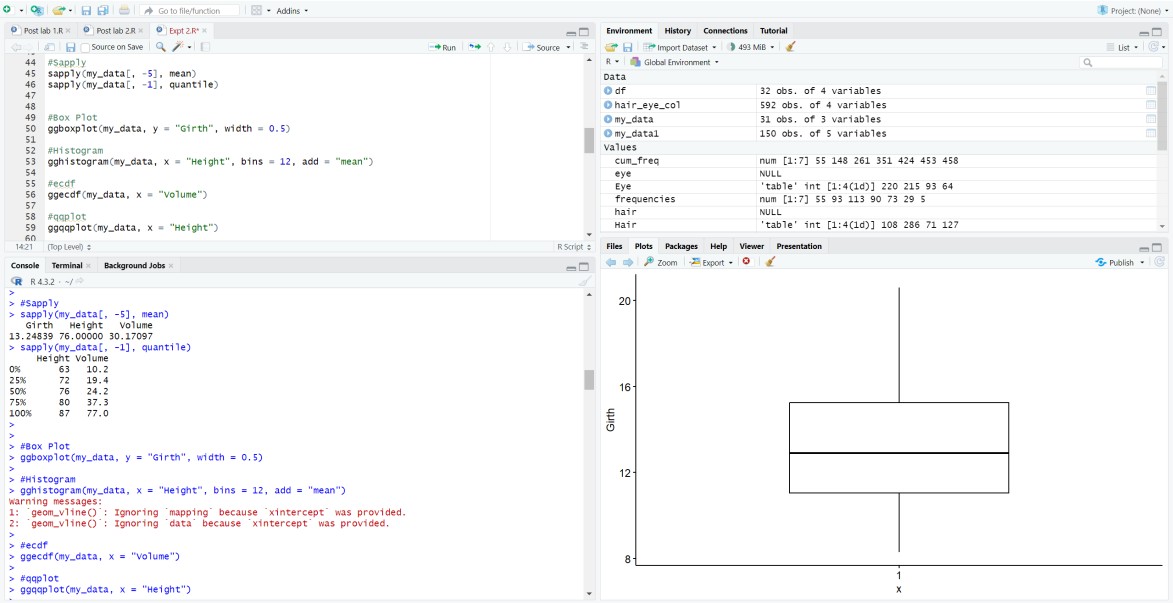
**3)**



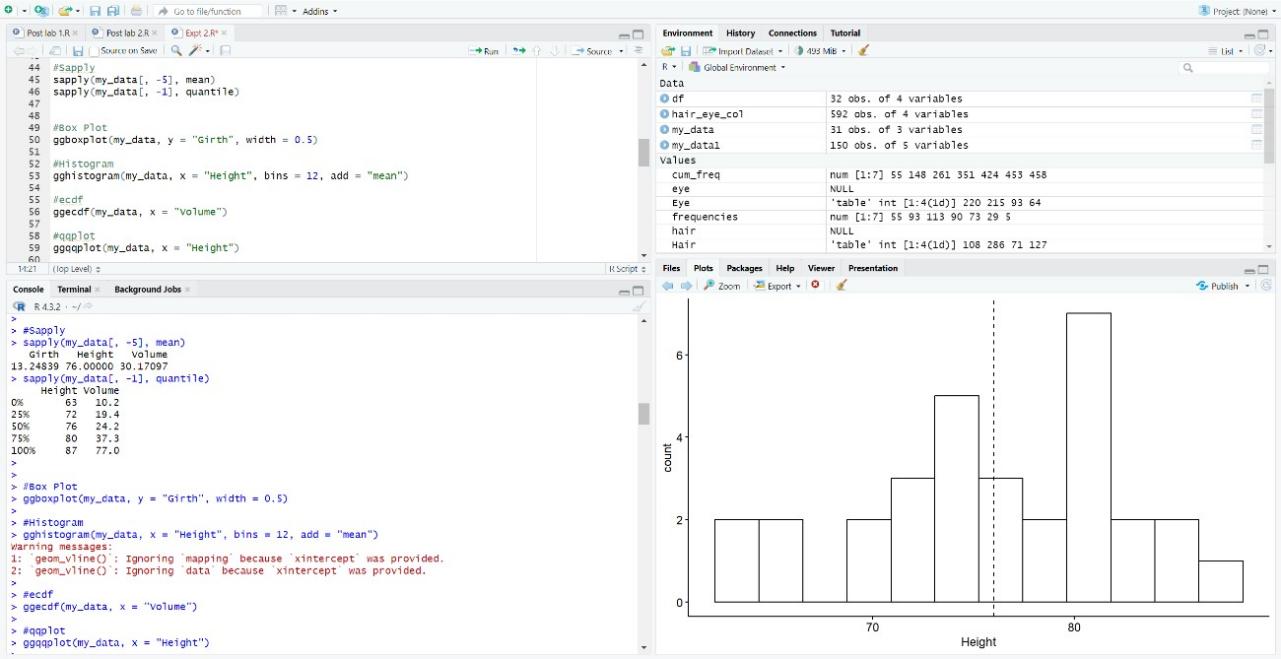
# 4)



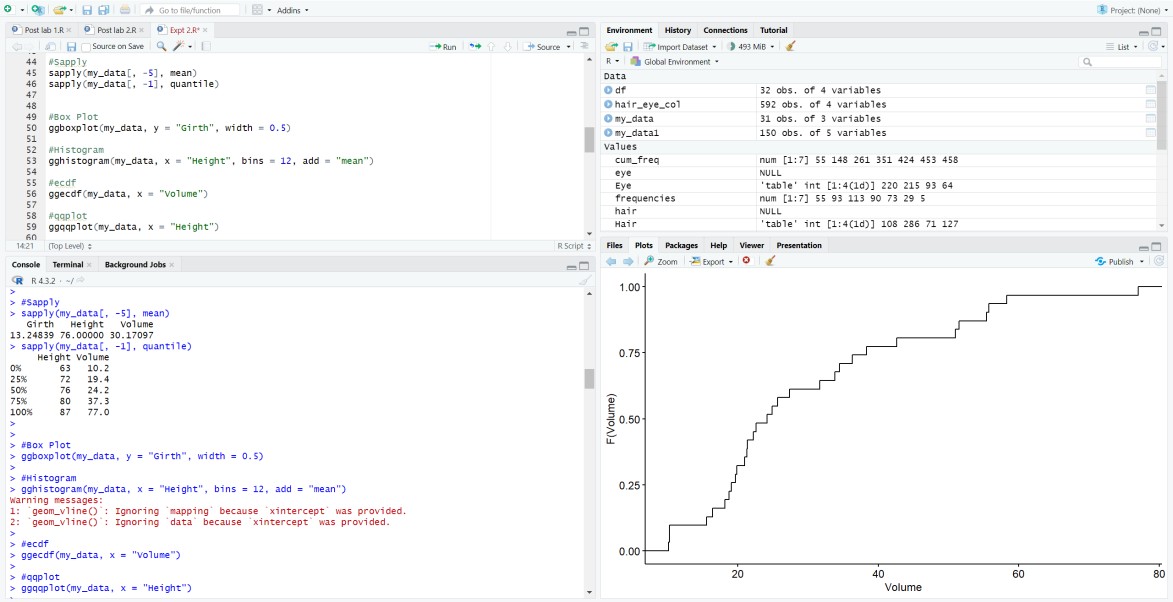
**5)**



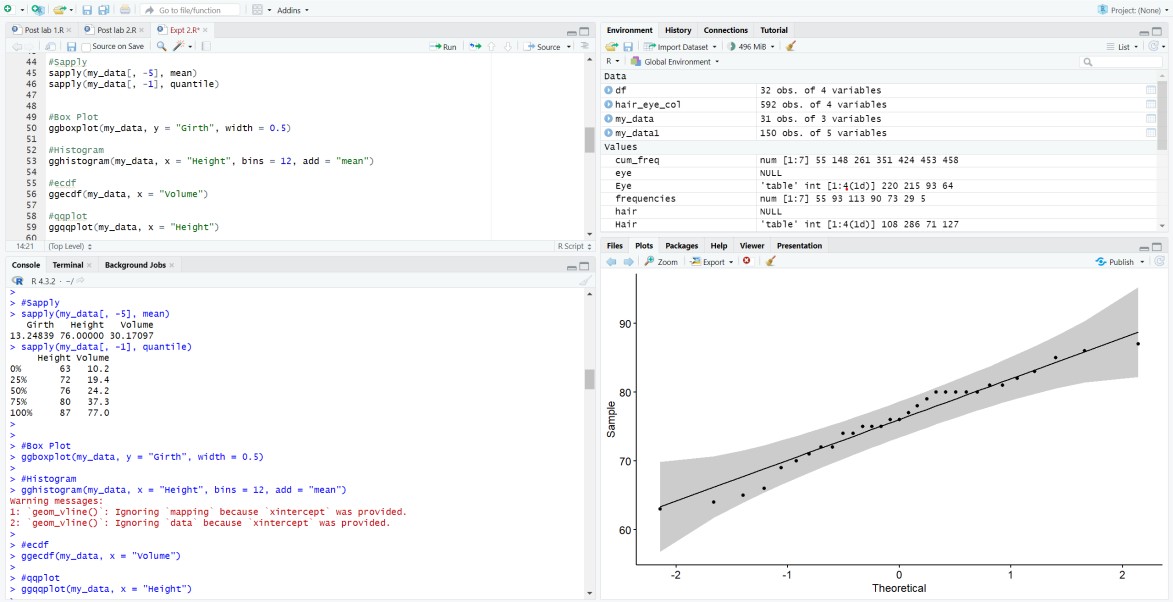
# 6)



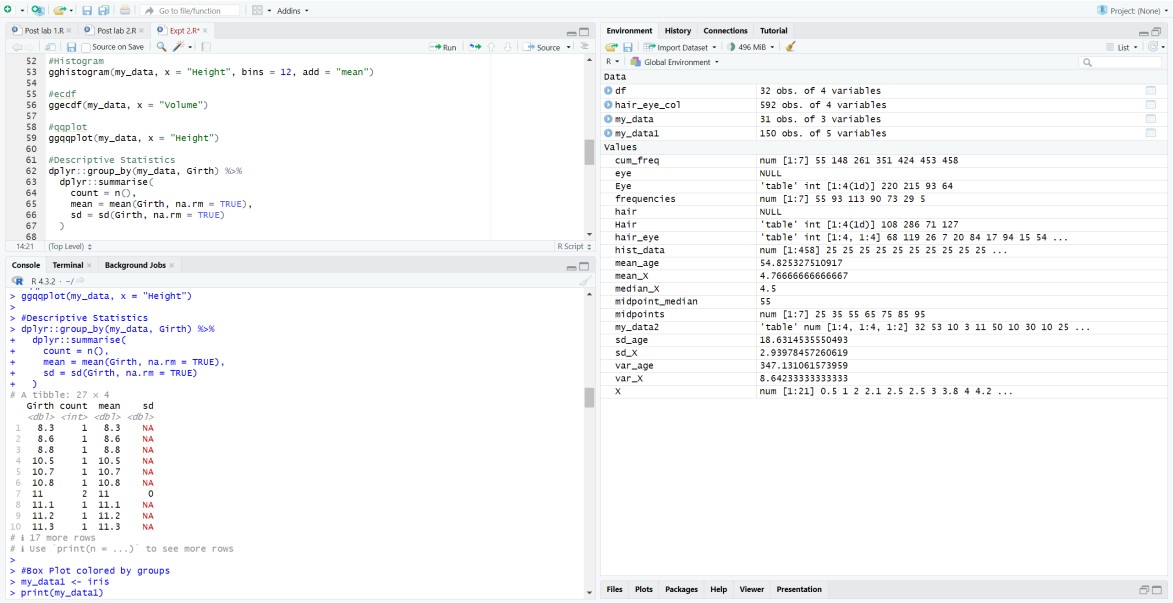
**7)**



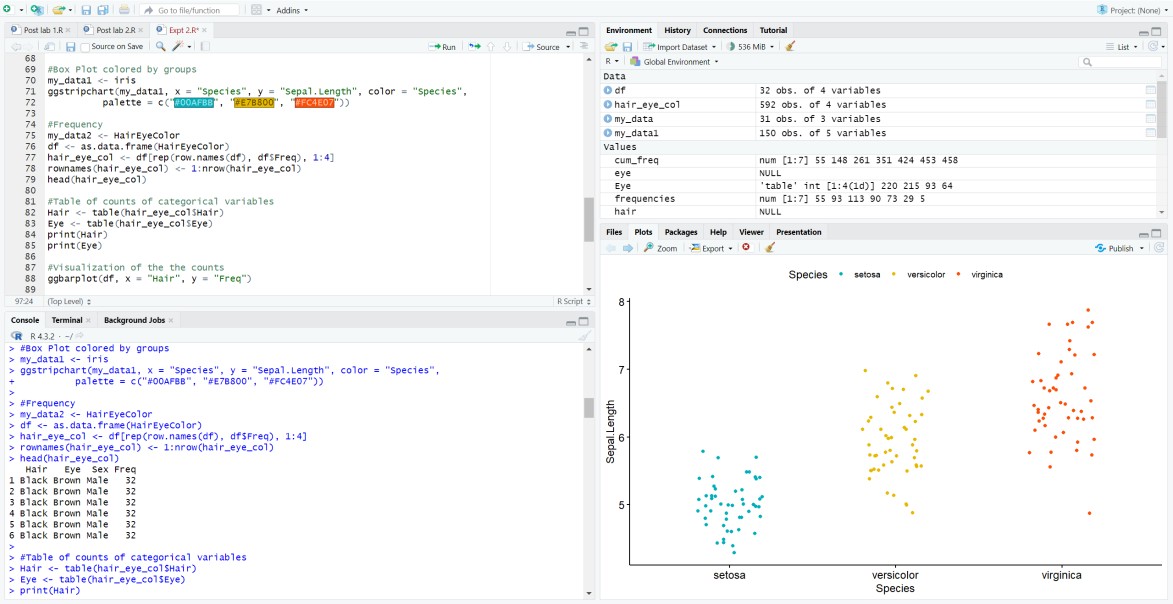
# 8)



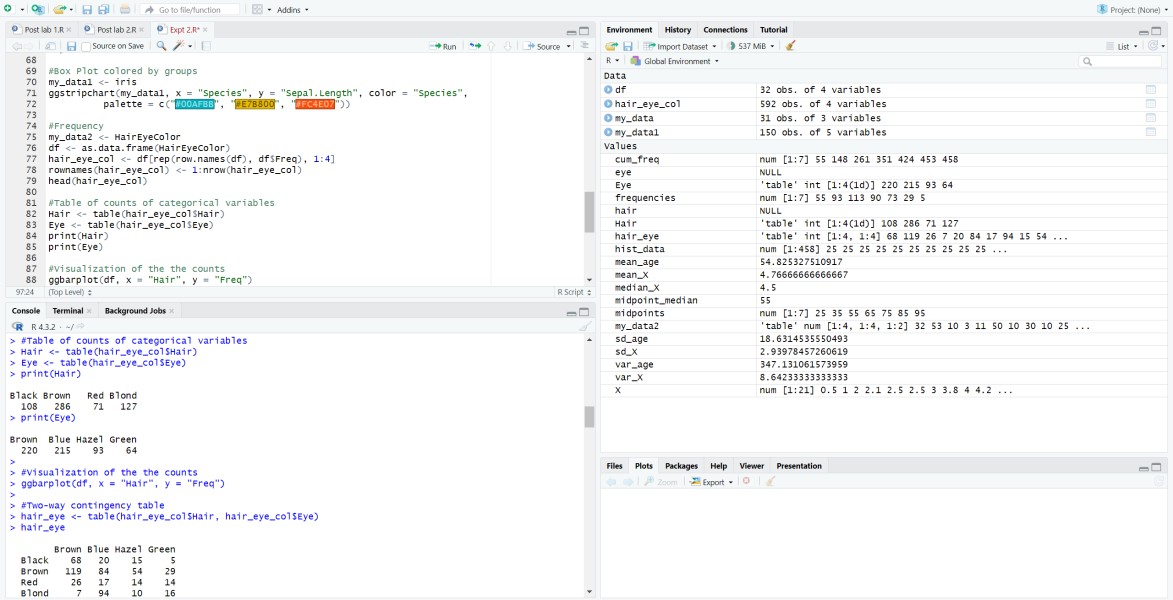
**9)**



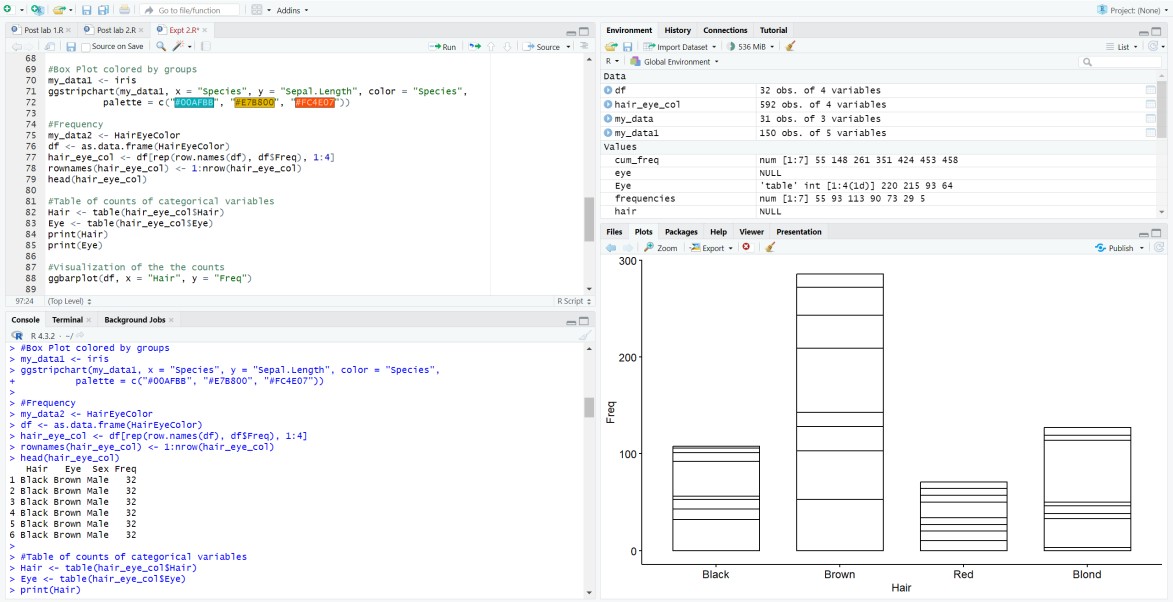
# 10)



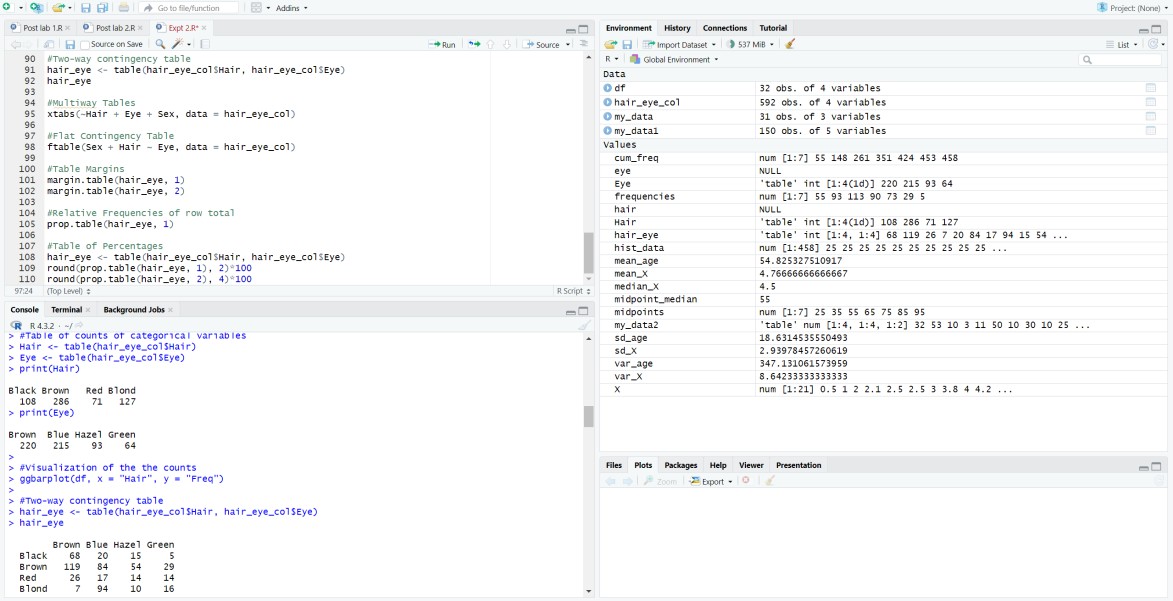
**11)**



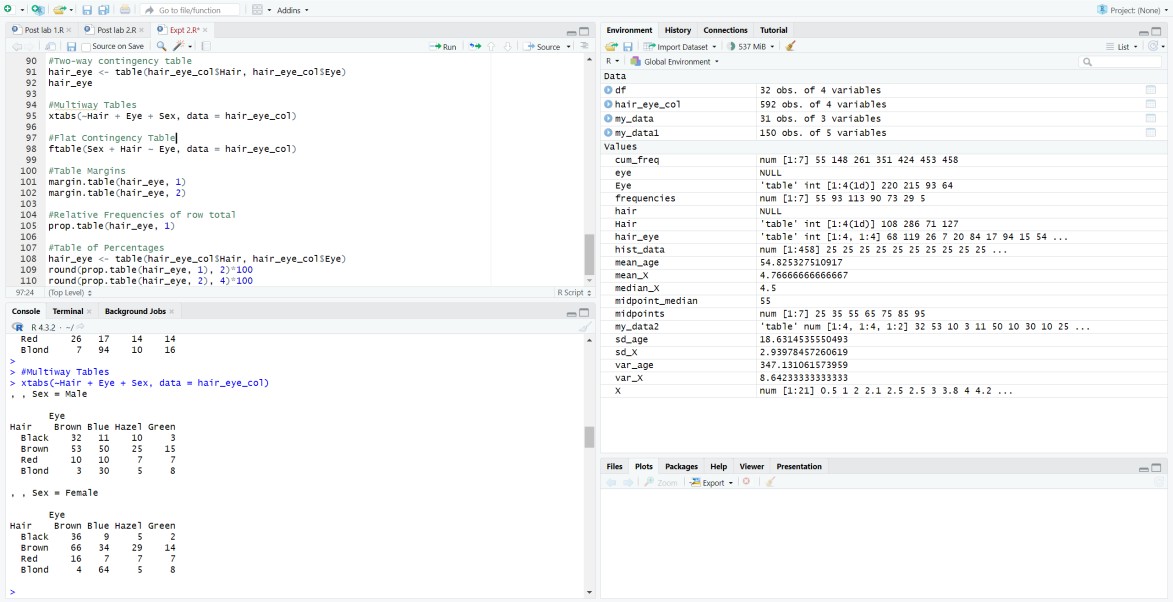
# 12)



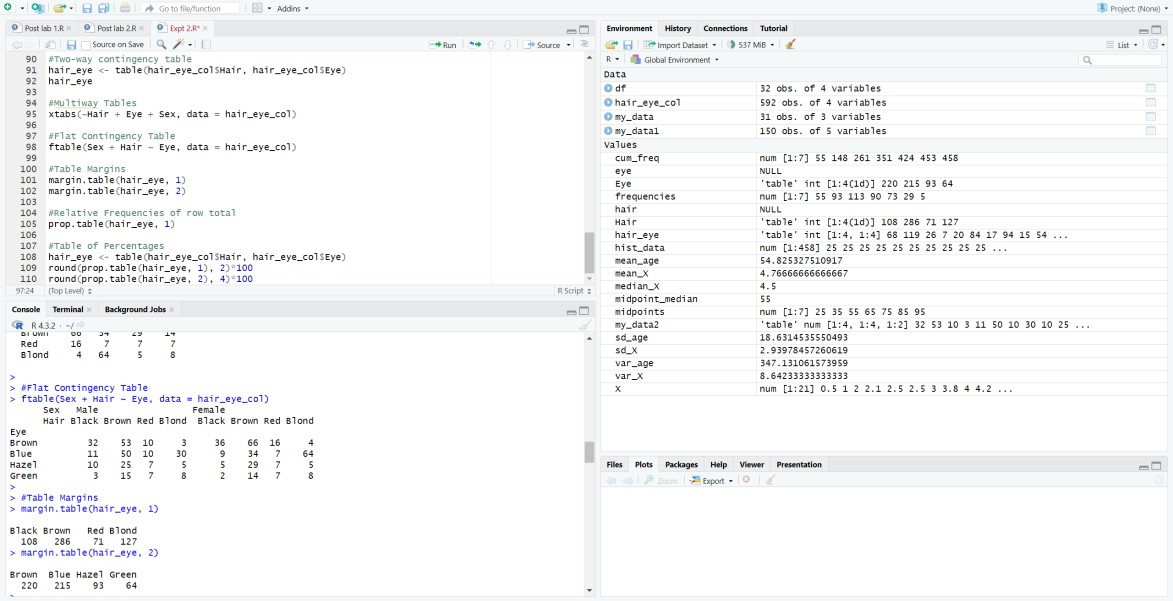
**13)**



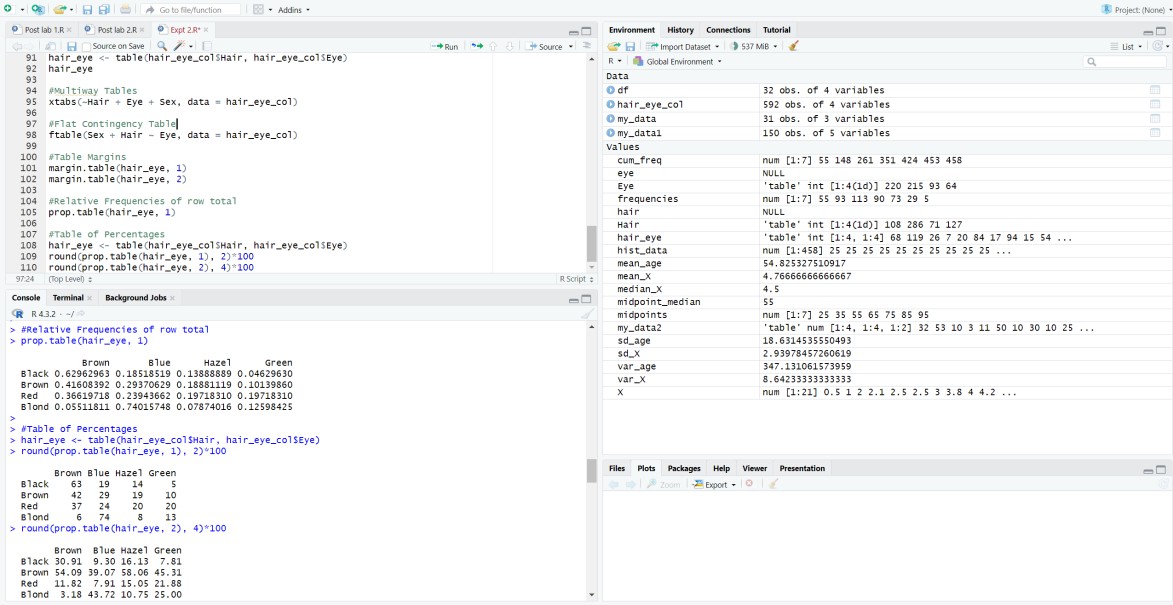
# 14)



**15)**



# 16)



**Conclusion:** Applying descriptive statistics in Applied Data Science reveals insights into central tendencies, variabilities, and distributions. It forms a foundational step for informed decision-making and strategic data-driven approaches.

# Post Lab questions

**Write R commands for the following ?**

1. In an article in American Journal of Pathology, Pitts et al (2001) have taken the measurements on diameters in centimetres of the neoplasm removed from the breasts of 20 subjects with pure sarcoma. Following is the dataset: 0.5, 1, 2, 2.1, 2.5, 2.5, 3.0, 3.8, 4.0, 4.2, 4.5, 5.0, 5.0, 5.0, 5.0, 6.0, 6.5, 7.0, 8.0, 9.5, 13.0
   1. Enter the dataset using scan function and store in the variable X
   2. Find the mean, median, variance and standard deviation of x
   3. Create the boxplot

# SOLUTION:

**CODE:**

# a. Enter the dataset using scan function and store in the variable X

X <- scan(text = "0.5 1 2 2.1 2.5 2.5 3.0 3.8 4.0 4.2 4.5 5.0 5.0 5.0 5.0 6.0 6.5 7.0 8.0

9.5 13.0")

# b. Find the mean, median, variance, and standard deviation of X mean\_X <- mean(X)

median\_X <- median(X) var\_X <- var(X)

sd\_X <- sd(X)

# Print the results cat("Mean:", mean\_X, "\n")

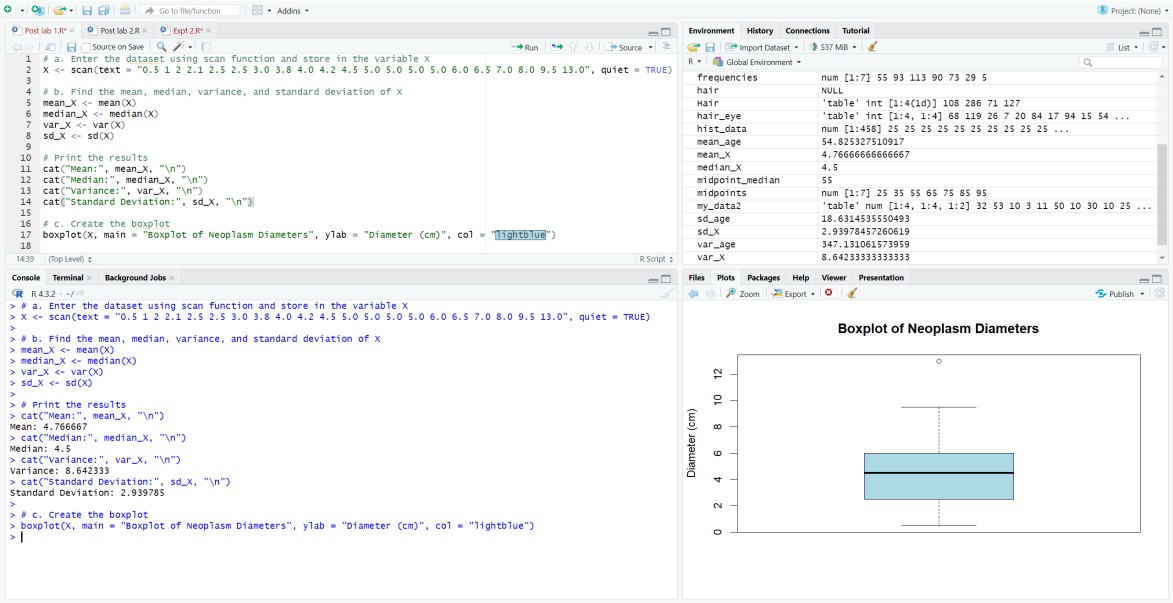
cat("Median:", median\_X, "\n")

cat("Variance:", var\_X, "\n") cat("Standard Deviation:", sd\_X, "\n")

# c. Create the boxplot

boxplot(X, main = "Boxplot of Neoplasm Diameters", ylab = "Diameter (cm)", col = "lightblue")

# EXECUTION SCREENSHOT:



1. American Journal of psychiatry conducted a study of the presence of significant psychiatric illness in heterozygous carriers of the gene for the Wolfram syndrome. Among the subject studied were 543 blood relatives of patients of Wolfram syndrome. Following is the frequency distribution of ages of these blood relatives:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age(Mid-point) | 25 | 35 | 55 | 65 | 75 | 85 | 95 |
| Number(Frequency) | 55 | 93 | 113 | 90 | 73 | 29 | 5 |

* 1. Enter the dataset using data.frame command
  2. Add a column cumulative frequency
  3. Add a column of relative frequency (frequency/total frequency)
  4. Add a column of relative cumulative frequency (cumulative frequency/total frequency)
  5. Plot cumulative frequency vs mid points

# SOLUTION:

**CODE:**

# Given frequency distribution midpoints <- c(25, 35, 55, 65, 75, 85, 95)

frequencies <- c(55, 93, 113, 90, 73, 29, 5)

# a. Enter the dataset using data.frame command

df <- data.frame(AgeMidpoint = midpoints, Frequency = frequencies)

# b. Add a column cumulative frequency df$CumulativeFrequency <- cumsum(df$Frequency)

# c. Add a column of relative frequency (frequency/total frequency) df$RelativeFrequency <- df$Frequency / sum(df$Frequency)

# d. Add a column of relative cumulative frequency (cumulative frequency/total frequency) df$RelativeCumulativeFrequency <- df$CumulativeFrequency / sum(df$Frequency)

# e. Plot cumulative frequency vs midpoints

plot(df$AgeMidpoint, df$CumulativeFrequency, type = "o", main = "Cumulative Frequency vs Midpoints",

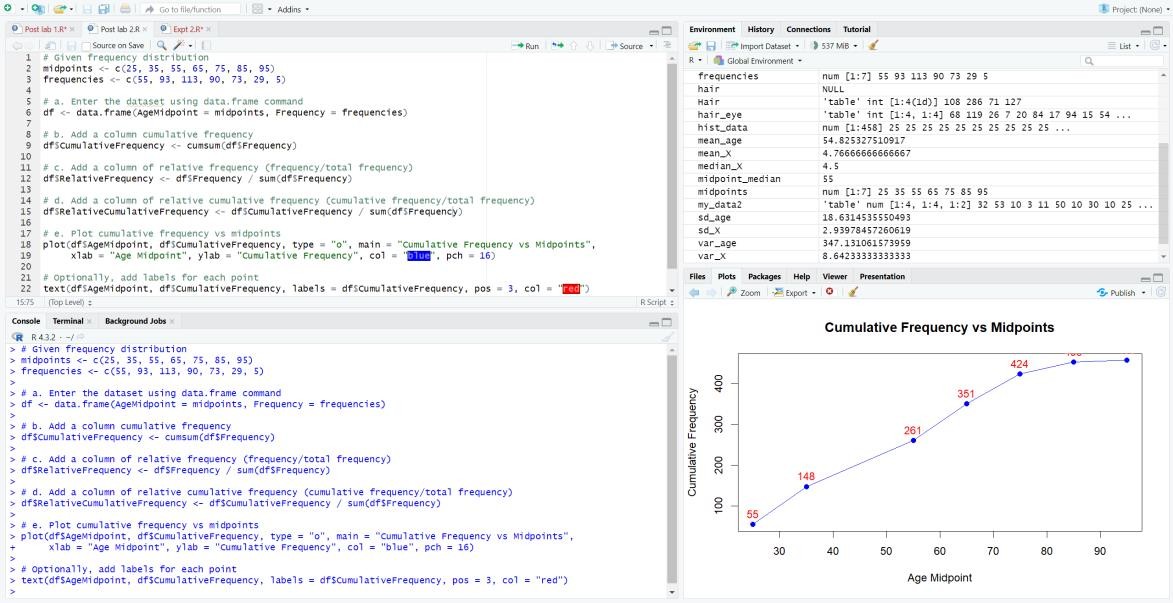
xlab = "Age Midpoint", ylab = "Cumulative Frequency", col = "blue", pch = 16)

# Optionally, add labels for each point

text(df$AgeMidpoint, df$CumulativeFrequency, labels = df$CumulativeFrequency, pos = 3, col

= "red")

# EXECUTION SCREENSHOT:



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

# Limitations of Using Only Measures of Central Tendency:

Central tendency measures (mean, median, mode) have limitations:

**Ignoring Distribution Shape:** They don't provide information about the shape of the distribution. Two datasets with the same mean might have very different distributions.

**Sensitivity to Outliers:** The mean is sensitive to extreme values (outliers), and a few outliers can significantly distort the mean. Median is less affected, but still may not be entirely robust.

**Not Descriptive of Spread:** Central tendency measures don't give insights into the spread or dispersion of data. Two datasets with the same mean can have different levels of variability.

**Applicability to Different Distributions:** Different central tendency measures might be more suitable for different types of data (e.g., median for skewed data, mean for symmetric).

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

**Comparison of Measures of Variability:**

Measures of variability (range, variance, standard deviation, interquartile range) have distinct characteristics:

**Range:** Simple but sensitive to outliers; it doesn't capture the overall spread effectively.

**Variance and Standard Deviation:** Provide a more nuanced understanding of the spread around the mean; sensitive to outliers.

**Interquartile Range (IQR):** Captures the spread of the middle 50% of the data, less sensitive to extreme values.

# When to Use One Measure Over Another:

**Use Range for Simplicity:** When simplicity is crucial and extreme values are not a significant concern.

**Use Variance/SD for Precision:** When a more precise measure of spread is needed and outliers need to be considered.

**Use IQR for Robustness:** When you want a measure that is less sensitive to extreme values.

**3.** 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.

# Descriptive Statistics Techniques in Understanding Data:

**Example Scenario:** Examining Exam Scores

**Mean:** Provides an average score, indicating the overall performance.

**Median:** Shows the middle point of the scores, helpful if there are extreme scores.

**Mode:** Identifies the most common score.

**Variance/SD:** Indicates the spread of scores around the mean.

**Histogram/Boxplot:** Visual representations to grasp the distribution shape and detect outliers.

Understanding these descriptive statistics aids in identifying trends, variations, and potential outliers, helping researchers make informed decisions and interpretations.

**Real-Life Example:** Examining Income Distribution

**Mean Income:** Gives an average income level.

**Median Income:** Provides the income level at the middle point.

**Mode Income:** Shows the most frequently occurring income range.

**Standard Deviation:** Indicates the variation in income levels.

**Boxplot:** Visualizes the income distribution and identifies potential outliers.

Analysing these descriptive statistics helps policymakers understand income disparities, target interventions, and make evidence-based decisions.