1)

Table

Description automatically generated

The three stages of reproducible workflow are:

1. Stage 1: Data Acquisition
2. Stage 2: Data Processing
3. Stage 3: Data Analysis

STAGE 1: DATA ACQUISITION:

The first steps which includes the extraction of the data  and the collecting of data, which are continuing tasks.

• The Excel file that must be the first be converted to a CSV file. In the second stage, a project structure is to be created to organize the data collection and the script of the project.

# data<-read.csv("Frailty\_data.csv") #

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

STAGE 2: DATA PROCESSING:

* Creating a script to process the raw data.

# Reading the raw data into an object called Frality\_data

data<-read.csv("Frailty\_data.csv")

Graphical user interface, text, application

Description automatically generated

# Now by Cleaning the rows to remove all Not Available cases where Weight.Pounds <120

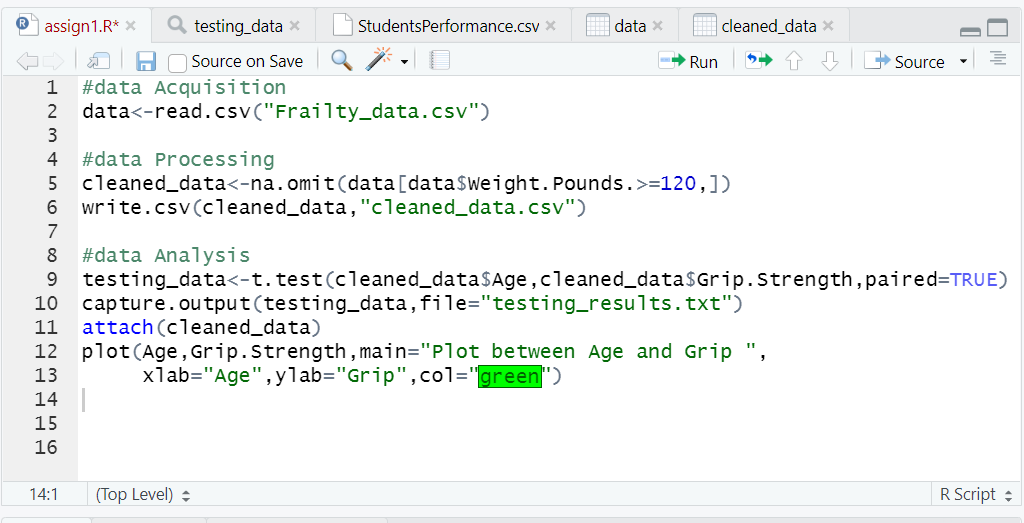
cleaned\_data<-na.omit(data[data$Weight.Pounds.>=120,])

#Writing clean\_data into the new CSV file

write.csv(cleaned\_data,"cleaned\_data.csv")

A screenshot of a computer

Description automatically generated



STAGE 3: DATA ANALYSIS:

* Clean data should be saved in a variable.
* To Analyze the clean data from field-by-field testing to determine whether age and grip significantly differ from one another.

# To Read the clean\_data csv into clean\_data

cleaned\_data<-na.omit(data[data$Weight.Pounds.>=120,])

A screenshot of a computer

Description automatically generated

# The testing of fraility by fields

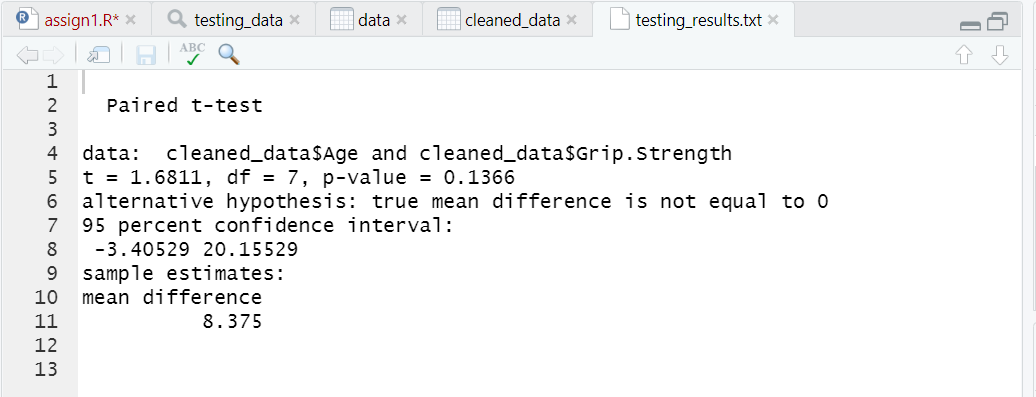
testing\_data<-t.test(cleaned\_data$Age,cleaned\_data$Grip.Strength,paired=TRUE)

Graphical user interface, table

Description automatically generated with medium confidence

#Saved the derived result in to a text file in a result directory.

capture.output(testing\_data,file="testing\_results.txt")



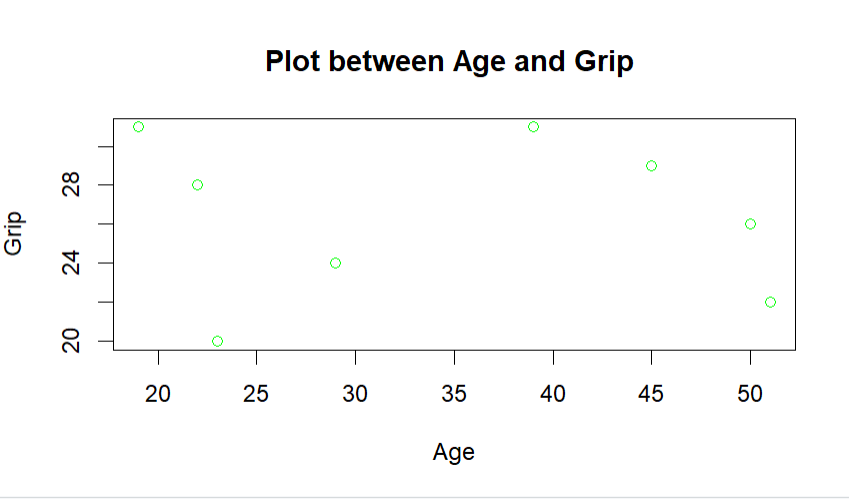
Plotting the Age vs Grip data:

* Plot between the Age and Grip data

attach(cleaned\_data)

plot(Age,Grip.Strength,main="Plot between Age and Grip ",

xlab="Age",ylab="Grip",col="red")

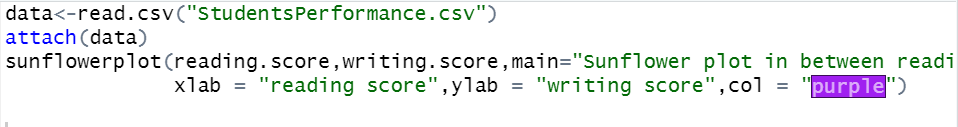


2) Perform 5 data visualization tasks on the student performance dataset linked below (create 5 different visualizations). Explain how each visualization has made analysis easier. Create a folder structure for this question similar to question 1.

**1. Sunflowerplot:**

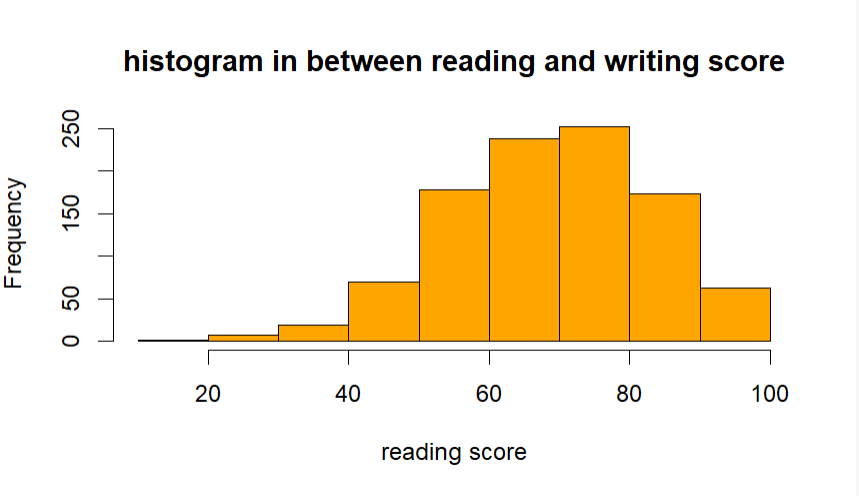
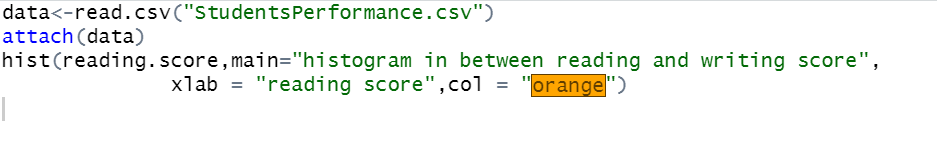
**\*** To avoid overplotting, combine a scatterplot and a sunflower plot. A sunflower plot only shows one point when multiple points with the same x, y values exist, but the tiny borders or petals radiating from the point show how many points are truly present.

\* Sunflower plots are data-display bivariate scatter plots. Reading becomes more difficult as the amount of data in a given plot area grows. Because each petal on a sunflower represents a single observation, sunflower plots can handle massive data densities.





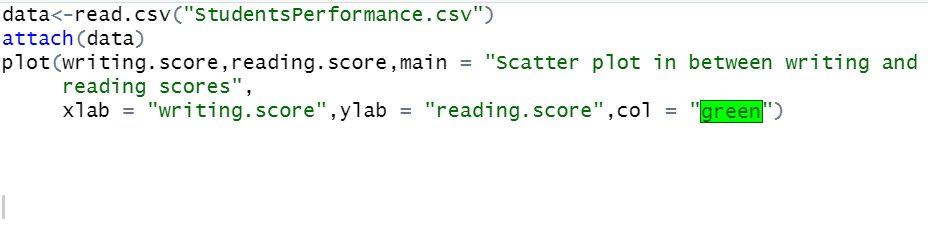
**2. Histogram:**

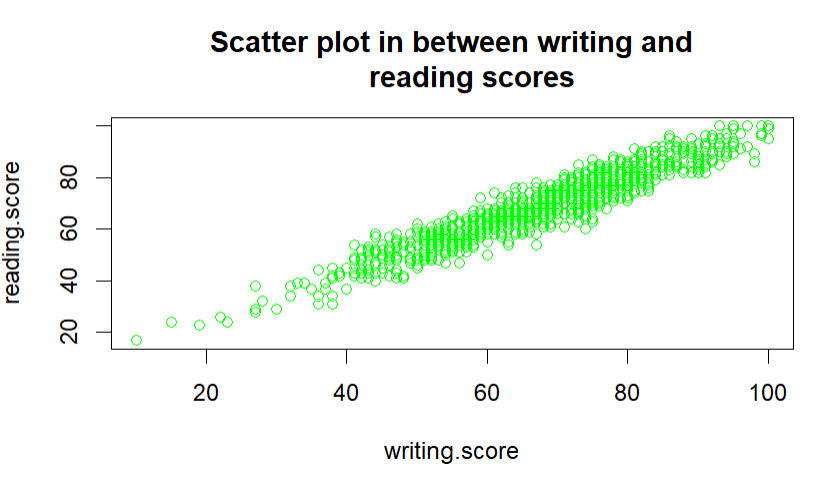
* A histogram is a visual representation of the frequency distribution of categorize continuous classes, as well as a graphic representation of the frequency distribution of a small number of data points from a single variable.
* A wonderful technique to finding out the distributions of the numerical or quantitative data is to use the histogram that to convert interval data into continuous or discrete data.

**3. Plot:**

• A plot, which is often in the form of a graph, depicts the relationship between two or more variables graphically. The storyline could be written by hand or created digitally. Historically, the two main types of plotters were mechanical and electronic.

• A plot that can be used that to demonstrate how the values in a dataset relate to one another. Use large datasets to demonstrate correlation and clustering. If each point in your dataset has two values. When the order of the points in the dataset is unimportant.



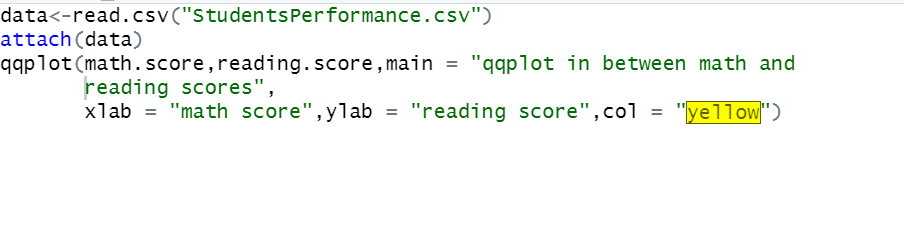


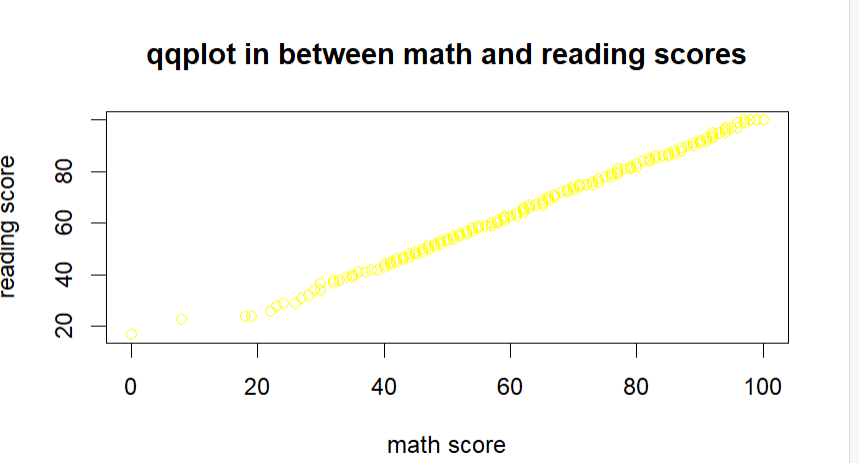
**4. QQplot:**

\* The quantile-quantile (q-q) plot is a graphical method for determining whether two data sets are from populations with a common distribution. In a q-q graphic, the quantiles from the first data set are plotted against the quantiles from the second data set.

\* The quantile-quantile (QQ) plot determines whether two sets of data are drawn from the same distribution.

\* Q-Q plots are used to determine the type of distribution for a random variable, which can be Gaussian, Uniform, Exponential, Pareto, or another. The power of the Q-Q plot can be used to determine the type of distribution just by looking at it.





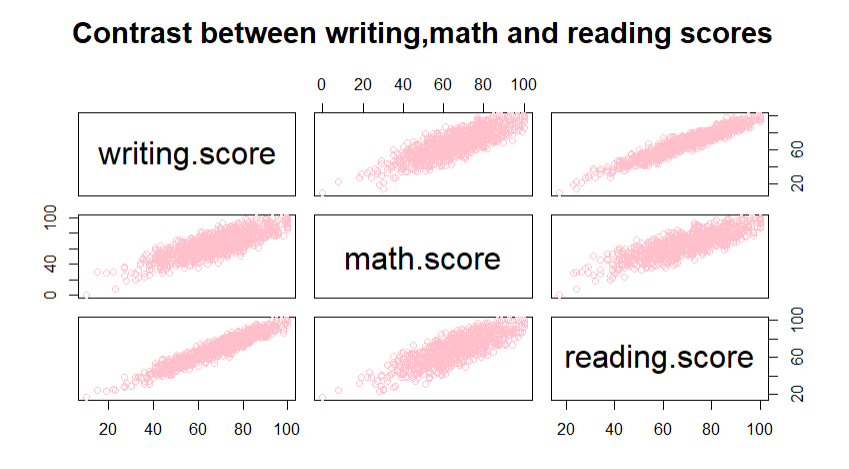
1. **Pairs:**

• An ordered pair contains the coordinates of one point. Use an ordered pair of coordinates to find a location (x, y). The first value is the x-coordinate, and the second is the y-coordinate. Put a dot on the graph to represent the location of the ordered pair.

•The pairs() function in R can be used to generate a pair plot. The pairs function, which comes pre-installed with R Language, generates a scatterplot matrix. The pairs() function generates a scatter plot matrix for each pair of variables in the input data frame

Graphical user interface, text

Description automatically generated



* **Histogram:**

In my opinion, the histograms are adaptable and straightforward. It can be used to provide an informative look at a frequency distribution in a variety of situations. It can be used to develop the best price plans and marketing strategies, for example, in sales and marketing.

A frequency histogram is useful because it allows you to visually compare data to a frequency distribution. Histograms show the typical distribution of a well-run process. However, because histograms are constantly generated, any variation can be easily identified. Organizations benefit greatly from this because it allows them to quickly identify and manage process deviation.

The frequency histogram, on the other hand, has the disadvantage of taking more time and effort to generate than the frequency graph.

* **Histogram vs sunflowerplot:**

It can aid in the detection of anomalous results, data omissions, and outliers. A histogram categorizes or groups the likely values in a data set.

Unfavorable, difficult to identify significant trends in vast data sets, and possibly perplexing for viewers unfamiliar with the concept.

* **Histogram vs plot:**

A histogram can show the frequency of a large amount of data. The density plot represents the data-estimated histogram in a continuous and smooth manner. A kernel density estimate is used to compute it.

Plots are used to show the set's values. In contrast to the box plot, the median is less obvious.

* **Histogram vs qqplot:**

The histogram shows that this distribution is skewed to the left. Given that the data sets are drawn from various distributions, the QQ plot shows that the points do not line up in a straight line.

Histograms, on the other hand, are frequently used to conveniently depict the key characteristics of the data's distribution.

* **Histogram vs pairs:**

The scatter plot and histogram are the two main graphs upon which the pairs plot is built.

The upper and lower triangular scatter plots depict the relationship, or lack thereof, between two variables, whereas the diagonal histogram depicts the distribution of a single variable..