1. Table

Description automatically generated

Ans: 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:

First step entails data extraction and data gathering, both of which are ongoing activities.

• The first step is to convert the excel file into a CSV file.

• In the second step, a project structure is developed to organize the data collection and project script.

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

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

STAGE 2: DATA PROCESSING:

* Create 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

Description automatically generated

# Cleaning the rows to remove all Not Available cases where weight.pounds<150

cleaned\_data<-na.omit(data[data$Weight.Pounds.>=150,])

#Writing clean\_data into the new CSV file

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

Graphical user interface, table

Description automatically generated

A screenshot of a computer

Description automatically generated

STAGE 3: DATA ANALYSIS:

• Save the clean data in a variable.

• Do a field-by-field testing analysis on clean data to evaluate if there is a significant difference in Age and Grip.

# Read the clean\_data csv into clean\_data

cleaned\_data<-na.omit(data[data$Weight.Pounds.>=150,])

Graphical user interface, table

Description automatically generated

# testing of fraility by fields

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

Graphical user interface, text, application, email

Description automatically generated

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

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

Graphical user interface, text, application, email

Description automatically generated

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")

Chart, scatter chart

Description automatically generated

CODE:

#data Acquisition

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

#data Processing

cleaned\_data<-na.omit(data[data$Weight.Pounds.>=150,])

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

#data Analysis

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

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

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 given in the link below (create 5 different visualizations). Explain what kind analysis has become easier with each of the visualizations. Create the folder structure for this question similar to question 1.

Ans:

1. **Histogram:**

• Histogram is a visual representation of the frequency distribution of categorized continuous classes and in other ways histogram graphically displays the frequency distribution of a few data points from a single variable. In histograms, data is frequently segmented into "bins" or "range groups," which count the number of data points in each bin.

When to use:

• The histogram is an excellent tool for visualizing distributions of numerical or quantitative data.

• Converting interval data into continuous or discrete data.

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

Chart, histogram

Description automatically generated

1. **Plot:**

A plot is a graphic representation of a data collection that typically takes the form of a graph and demonstrates the relationship between two or more variables. Either manually or digitally creating the plot is possible. Plotter types in the past have included mechanical and electronic.

When to use:

* plot can be used to display the relationships between the values in a dataset.
* To illustrate correlation and clustering, use sizable datasets.
* If each point in your dataset has two values.
* When the sequence of the points in the dataset is unimportant.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

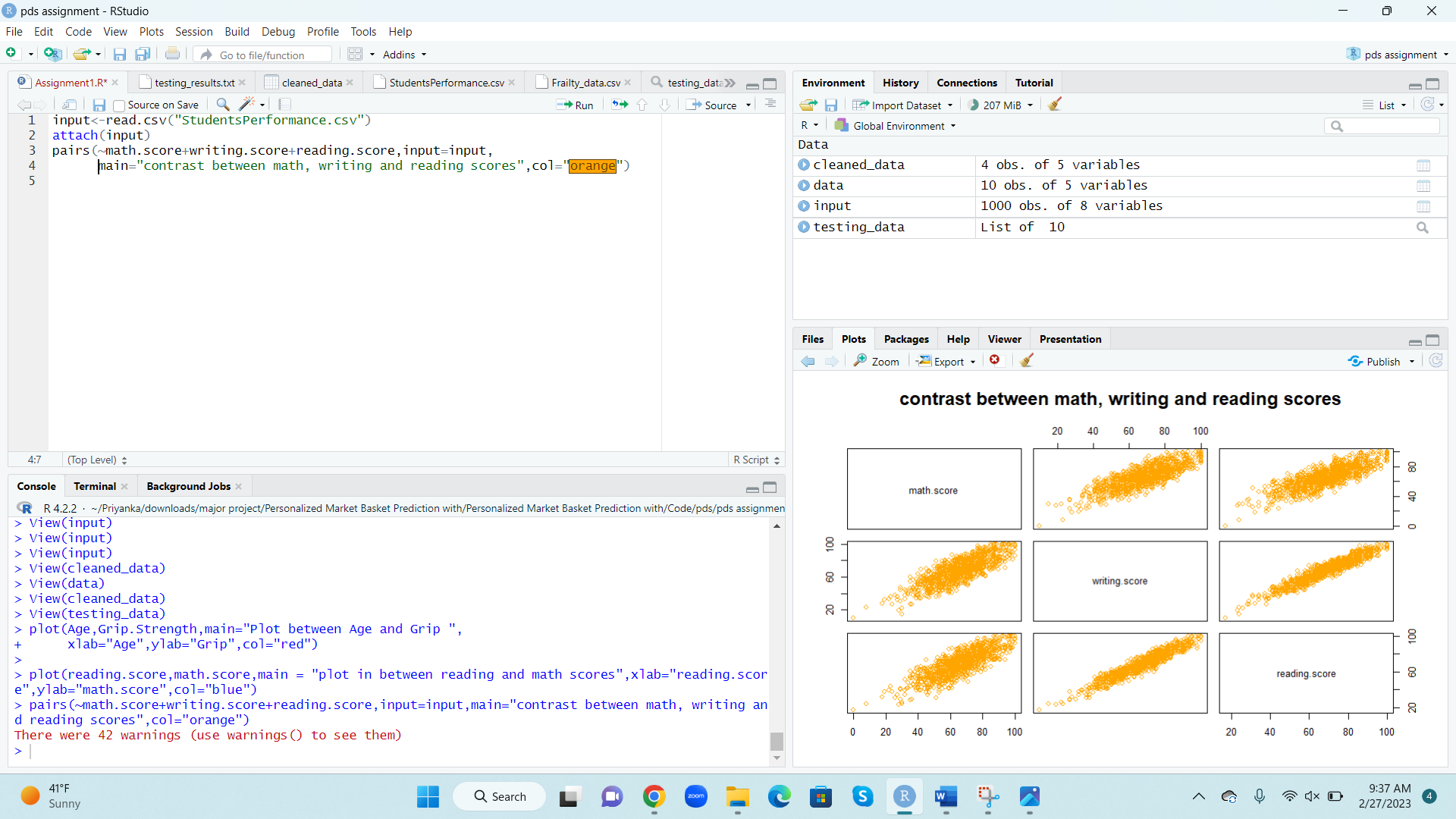
Description automatically generated

1. **Pairs:**

An ordered pair contains one point's coordinates in the coordinate system. An ordered pair of coordinates can be used to pinpoint a location (x, y). The first value is the x-coordinate, and the second is the y-coordinate. At the locations of the ordered pair, place a dot on the graph.

When to use:

* The pairs() function in R can be used to make a pair plot. The pairs function, which comes pre-installed with R Language, produces a scatterplot matrix. The pairs() function creates a scatter plot matrix for each pair of variables in the input data frame.
* The pair plot displays both the distribution of single variables as well as the relationships between two variables. They provide a wonderful method for identifying interrelationships between variables that can be further investigated. Each panel of a multipanel scatter plot known as a pair plot contains a scatter plot between a pair of variables.



Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated

1. **Sunflowerplot:**

Use a scatterplot with a sunflower plot to reduce overplotting. Sunflower plots only show one point when multiple points with the same x, y values exist; however, the number of points that are actually present is indicated by the small edges or petals radiating from the point.

When to use:

* Sunflower plots are scatter plots with a bivariate distribution presentation. When the density of data in a certain location of a plot rises, reading gets harder.
* Sunflower plots can handle high-density data since each petal represents an observation.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

1. **Boxplot:**

A box plot is a graphical representation of statistical data that uses the minimum, first quartile, median, third quartile, and maximum values. A "box plot" is the name given to the graph's design, which resembles a rectangle with lines extending from the top and bottom. It is a common type of descriptive statistics graphic used in explanatory data analysis is the box plot. A box and whisker plot is another name for it.

When to use:

* Box plots show the distribution and skewness of numerical data by displaying averages and data quartiles or percentiles.
* When should you use a box plot? Box plots are used to display numerical data distributions, particularly when comparing data from various groups. They are made to quickly display key details like symmetry, skew, variance, and outliers of a group of data.

A screenshot of a computer

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Histogram:**

In my opinion, the histograms are adaptable and simple. It can be used in a variety of circumstances to provide an educational glimpse into a frequency distribution. It can be used in sales and marketing to create the best price plans and marketing campaigns, for example.

A frequency histogram's advantage is that it enables visual comparison of data to a frequency distribution.

Histograms can display the typical distribution of a well-run process. But because histograms are produced frequently, any deviation can be quickly identified. Organizations gain a lot from this because it enables them to quickly identify and manage process variance.

But there is a disadvantage of the frequency histogram is that it requires more time and effort to create than the frequency distribution and is more difficult to do so.

**Histogram vs plot:**

A histogram can be used to display a large amount of data along with its frequency. The density plot is a continuous and smoothed representation of the data-estimated histogram. It is calculated using an estimation of kernel density.

Plots are used to display the values of the set. Contrarily, the median is less visible than it is in the box plot.

**Histogram vs pairs:**

The two primary graphs on which the pairs plot is based are the scatter plot and the histogram.

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

**Histogram vs sunflowerplot:**

It can help with the identification of unusual results outliers or data omissions. A histogram classifies or groups the likely values of a data set.

Unpleasant, and difficult to spot important trends in huge data sets, and possibly confusing for viewers unfamiliar with the concept.

**Histogram vs boxplot:**

Histograms accurately depict the distribution of a variable. Box plots make an effort to achieve the same objective, but the distribution of the variable is less clearly shown.

The disadvantage of using a boxplot instead of a histogram is that the mean is displayed as the measure of center. A boxplot does not depict variation. A boxplot presents less data.