# EX.NO: 01

**Date:**

**ODD\_or\_EVEN, ADDITION AND SUBTRACTION PROGRAM**

**USING R TOOLS**

**PROBLEM STATEMENT:**

Download the dataset from the UCI repository (or) any other appropriate website and perform (or) implement the central tendency measures. (mean, median, mode and midrange) and Data dispersion technique including summary.

**DESCRIPTION:**

This data comes from the 2010 census profile of general population and housing characteristics. Zip codes and limited to those that fall at least partially within LA city boundaries. The dataset will be updated after the next census in 2020.

**CENTRAL TENDENCY:**

1. **Odd or Even :** Odd numbers are those numbers that cannot be divided into two equal parts, whereas

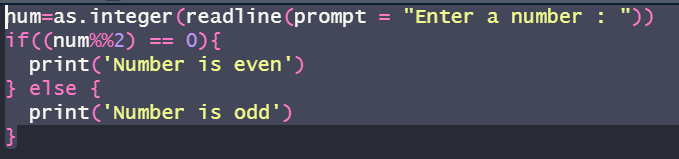
even numbers are those numbers that can be divided into two equal parts.

1. **Addition :** Adding something, especially two or more numbers.
2. **Subtraction** : Subtracting something, especially two or more numbers.

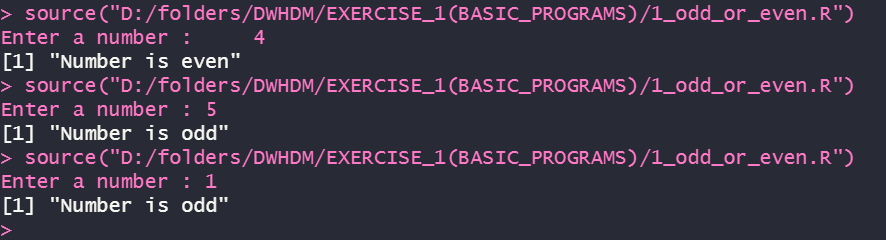
**INPUTS AND OUTPUTS OF BASIC PROGRAM :**

**ODD or EVEN :**

**INPUT :**

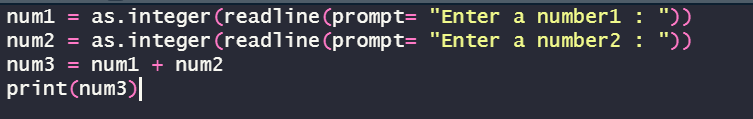


**OUTPUT :**

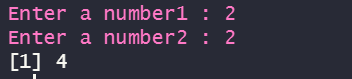


**ADDITION :**

**INPUT :**

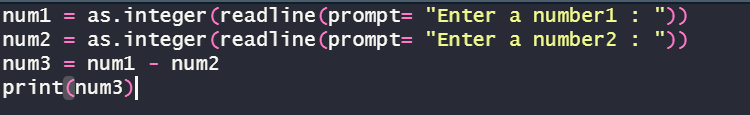


**OUTPUT :**

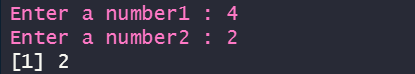


**SUBTRACTION :**

**INPUT :**



**OUTPUT :**



**RESULT:**

Thus the basic programs like odd or even, addition and subtraction are executed successfully.

**EX. NO: 02**

**Date:**

**CENTRAL TENDENCY AND DATA DISPERSION MEASURES**

**USING R-TOOL**

**PROBLEM STATEMENT:**

Download the dataset from the UCI repository (or) any other appropriate website and perform (or) implement the central tendency measures. (mean, median, mode and midrange) and Data dispersion technique including summary.

**DESCRIPTION:**

This data comes from the 2010 census profile of general population and housing characteristics. Zip codes and limited to those that fall at least partially within LA city boundaries. The dataset will be updated after the next census in 2020.

**CENTRAL TENDENCY:**

1. **Mean** : The mean is the average of the numbers: a calculated "central" value of a set of

numbers.

1. **Median :** The median is a statistical term that is one way of finding the 'average' of a set of

data points.

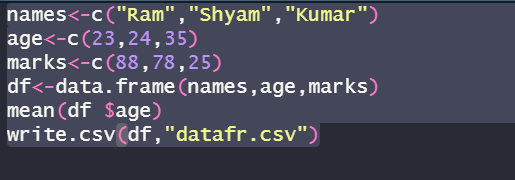
1. **Mode :** The mode of a set of data values is the value that appears most often.
2. **Summary :** A summary table stores data that has been aggregated in a way that answers a

mean common (or resource-intensive) business query.

**INPUTS AND OUTPUTS OF CENTRAL TENDENCY AND DATA DISPERSION:**

**MEAN:**

**INPUT :**

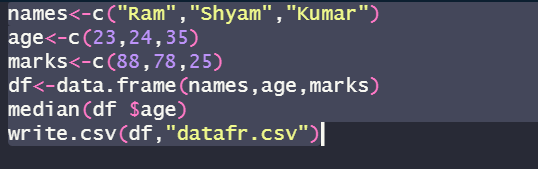
****

**OUTPUT :**

****

**MEDIAN :**

**INPUT :**

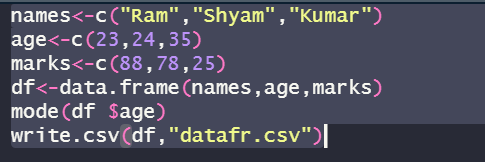
****

**OUTPUT :**

****

**MODE :**

**INPUT :**

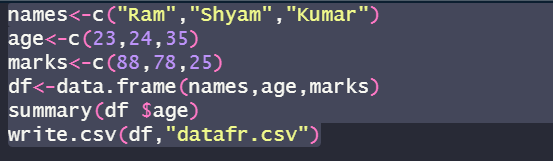
****

**OUTPUT :**

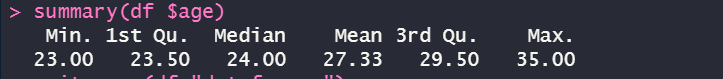
****

**SUMMARY :**

**INPUT :**

****

**OUTPUT :**

****

**RESULT:**

Thus the central tendency and measures of dispersion have been executed successfully. The outlier values are from more than upper fence there are no lower fence values.

**EX.NO: 03**

**Date:**

**CENTRAL TENDENCY AND DATA DISPERSION MEASURES**

**USING R-TOOL**

**PROBLEM STATEMENT:**

Download the dataset from the UCI repository (or) any other appropriate website and perform (or) implement the central tendency measures.(mean, median, mode and midrange) and

Data dispersion technique including summary.

**DESCRIPTION:**

This data comes from the 2010 census profile of general population and housing characteristics. Zip codes and limited to those that fall at least partially within LA city boundaries. The dataset will be updated after the next census in 2020.

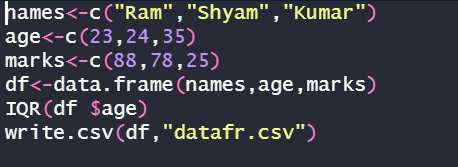
**MEASURES OF DISPERSION:**

1. **Inter Quartile Range :**The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts.
2. **Quartiles :**A quartile is a statistical term describing a division of observations into four defined intervals based upon the values of the data and how they compare to the entire set of observations.
3. **Mid Range :**The arithmetic mean of the largest and the smallest values in a sample or other group.

**INPUTS AND OUTPUTS OF CENTRAL TENDENCY AND DATA DISPERSION:**

**IQR :**

**INPUT :**

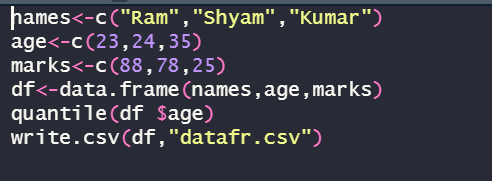
****

**OUTPUT :**

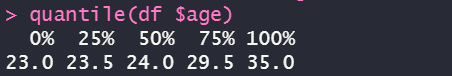
****

**QUANTILE :**

**INPUT :**

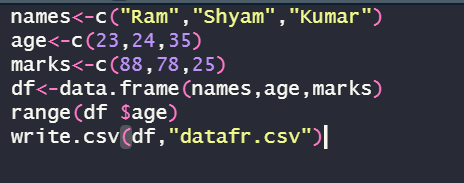
****

**OUTPUT :**

****

**RANGE :**

**INPUT :**

****

**OUTPUT :**



**RESULT:**

Thus the central tendency and measures of dispersion have been executed successfully. The outlier values are from more than upper fence there are no lower fence values.

**EX.NO: 04**

**Date :**

**PLOTTING GRAPHS USING R-TOOL**

**PROBLEM STATEMENT:**

Plot the boxplot, barplot and horizontal barplot for the dataset which was taken in the previous exercise.

**DESCRIPTION:**

Consider a dataset diabetes.csv, where it contains the attributes are Pregnancies,

Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcomes.

**IMPLEMENTATION:**

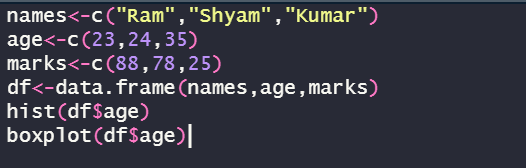
1. BoxPlot
2. BarPlot

**BOXPLOT :**

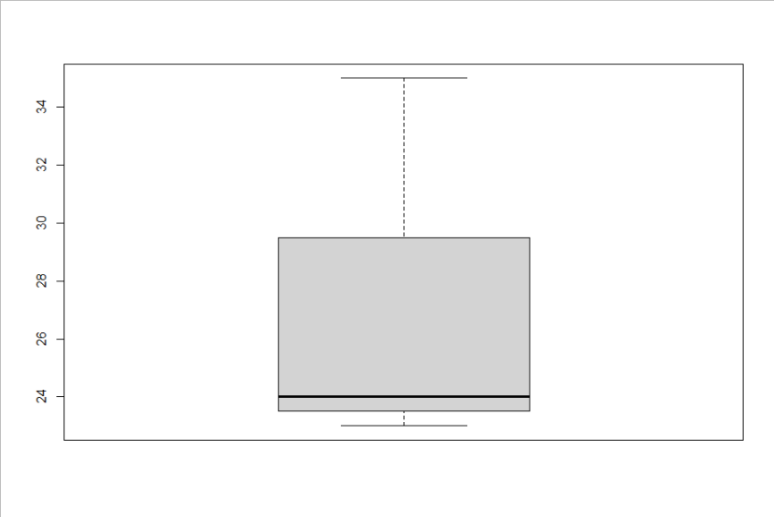
A box plot is a graphical rendition of statistical data based on the minimum,

fir00000000000000000000000000st quartile, median, third quartile, and maximum. The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the s1top and bottom. Because of the extending lines, this type of graph is sometimes called a box-andwhisker plot.

INPUT :



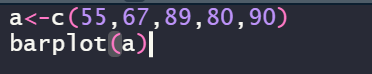
OUTPUT:



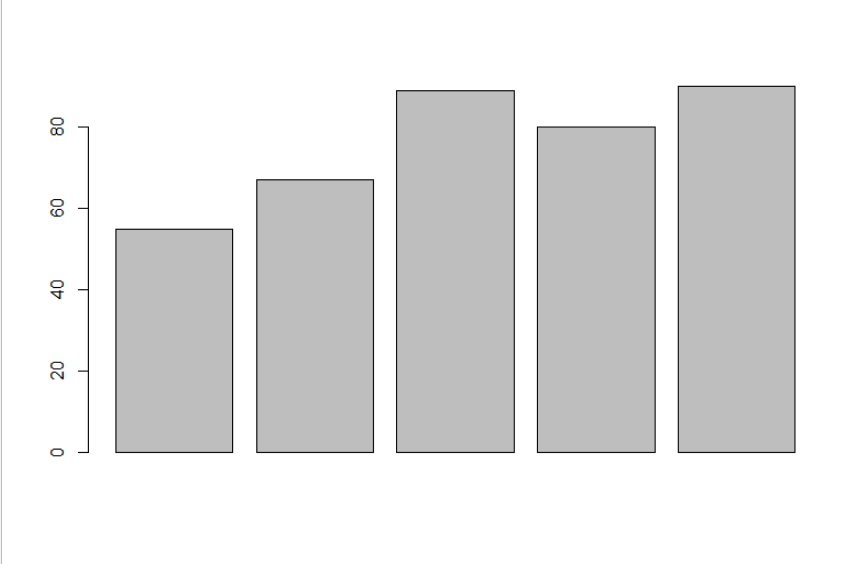
**BARPLOT :**

A barplot (or barchart) is one of the most common types of graphic. It shows the relationship between a numeric and a categoric variable. Each entity of the categoric variable is represented as a bar. The size of the bar represents its numeric value.

**INPUT :**

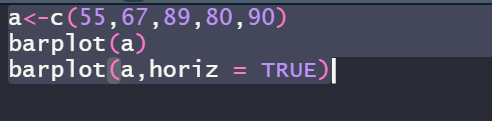


**OUTPUT :**

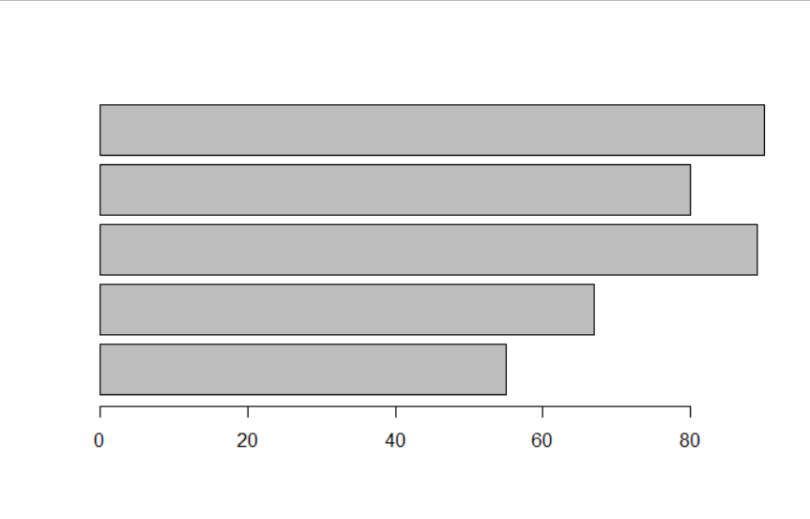


**HORIZONTAL BARPLOT :**

**INPUT :**

****

**OUTPUT :**



**RESULT:**

Thus, the plotting of graphs like boxplot, barplot and horizontal barplot for the given dataset has been successfully completed.

**EX.NO: 05**

**Date :**

**PLOTTING GRAPHS USING R-TOOL**

**PROBLEM STATEMENT:**

Plot the histogram and scatterplot for the dataset which was taken in the previous exercise.

**DESCRIPTION:**

Consider a dataset diabetes.csv, where it contains the attributes are Pregnancies,

Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcomes

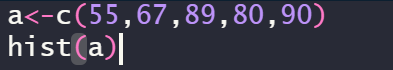
**IMPLEMENTATION:**

1. Histogram
2. Scatterplot (Scatter Smooth)

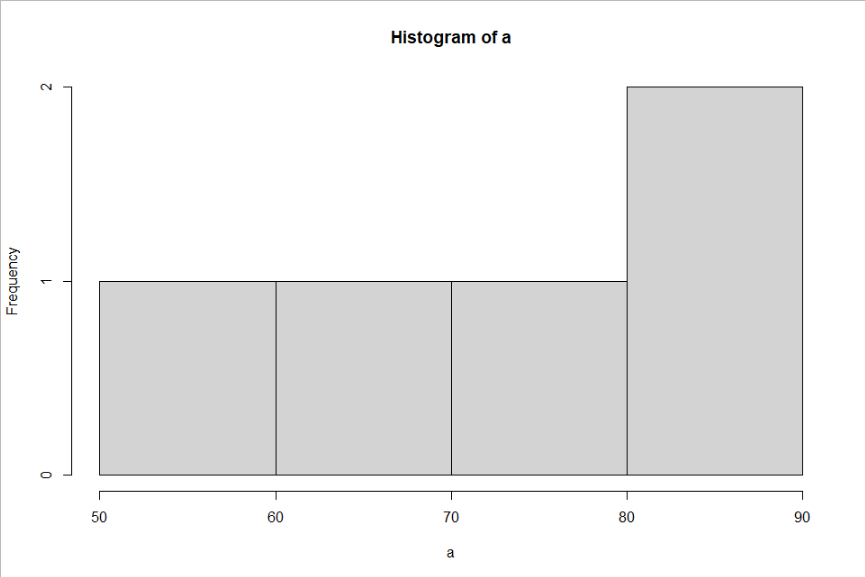
**HISTOGRAM :**

A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

INPUT :

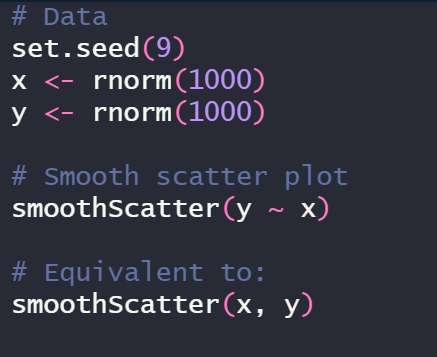


OUTPUT :

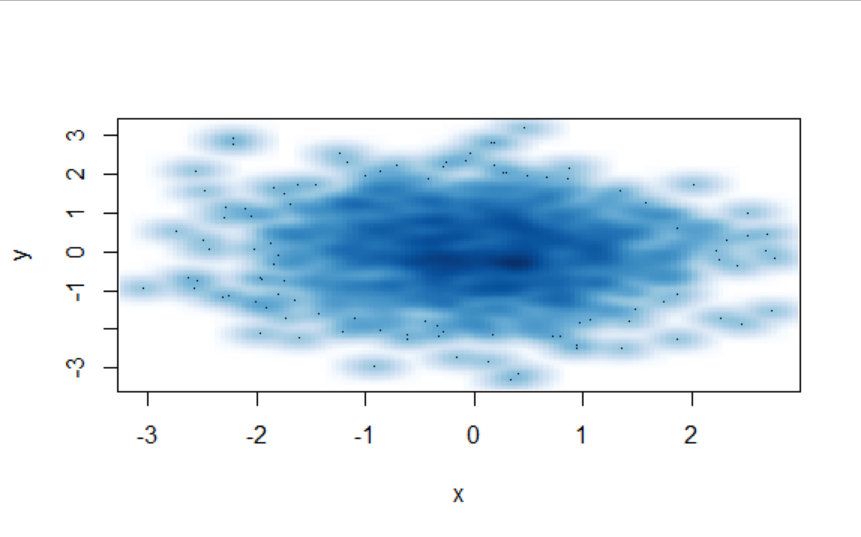


**SCATTERPLOT :**

**INPUT :**

****

**OUTPUT :**



**RESULT:**

Thus, the plotting of graphs like histogram and scatterplot for the given dataset has been successfully completed.

**EX.NO : 06**

**Date :**

**PERFORM CORRELATION ANALYSIS AND NORMALIZATION USING R-TOOL**

**PROBLEM STATEMENT :**

Perform the correlation analysis for the numerical attribute using pearson coefficient and for categorical attribute using chi-square and also, perform the normalization technique using z score for the given data frames of particular dataset.

**DESCRIPTION :**

A dataset of name diabetes.csv is given for the correlation analysis, to calculate or to correlate between Age and Insulin and the same dataset for the performance of normalization technique.

* **CORRELATION ANALYSIS:**

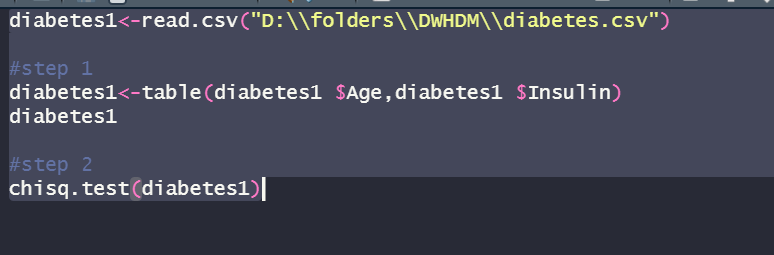
**STEPS INVOLVED:**

1. Create a new table with required dataframes.
2. After that apply the formula or query for the chi-square test.

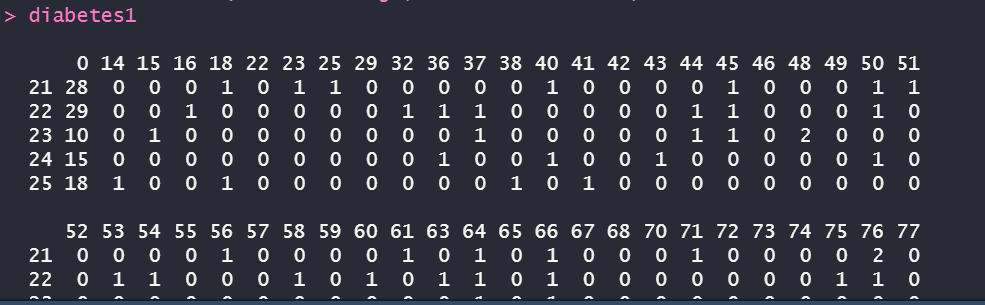
**QUERIES:**

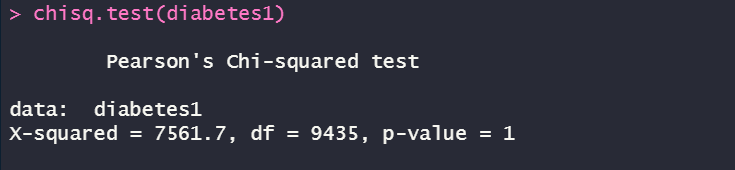
* diabetes1<-table(diabetes$Age,diabetes$Insulin)
* diabetes1
* chi sq.test(diabetes1)

**INPUT :**

****

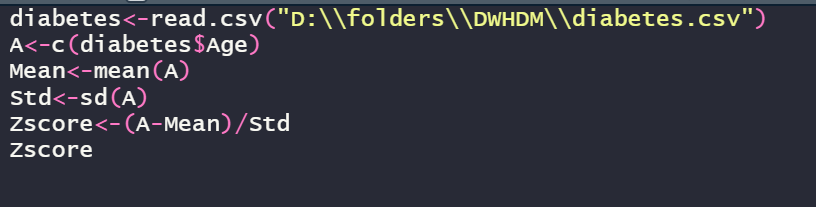
**OUTPUT :**

****

****

* **Z SCORE NORMALIZATION :**
* A<- c(diabetes$Age)
* Mean<- mean(A)
* Std<- sd(A)
* Zscore<- (A-Mean)/Std
* Zscore

**INPUT :**

****

**OUTPUT :**



**RESULT:**

Thus, the correlation analysis and normalization for the given dataset has been successfully executed and observed.

**EX.NO : 07**

**Date :**

**PERFORM CORRELATION ANALYSIS AND NORMALIZATION USING R-TOOL**

**PROBLEM STATEMENT :**

Perform the correlation analysis for perform the normalization technique for the given data frames of particular dataset.

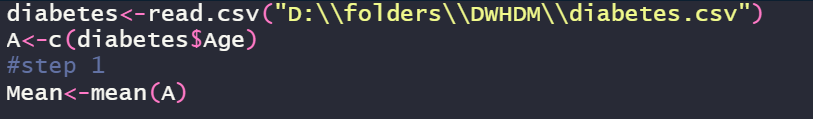
**DESCRIPTION :**

A dataset of name diabetes.csv is given for the correlation analysis, to calculate or to correlate between Age and Insulin and the same dataset for the performance of normalization technique.

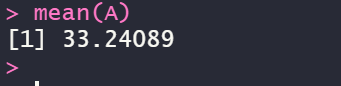
* **NORMALIZATION :**

1. **MEAN** **NORMALIZATION**
   * A<-c(diabetes$Age)
   * Mean<-mean(A)

**INPUT :**

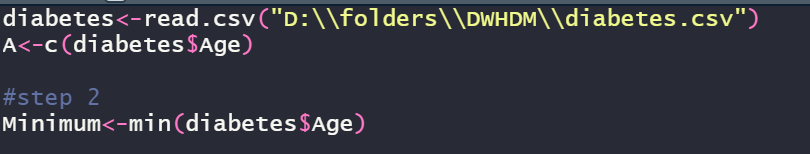


**OUTPUT :**

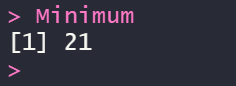


1. **MINIMUM NORMALIZATION**
   * A<-c(diabetes$Age)
   * Minimum<-min(diabetes$Age)

**INPUT :**

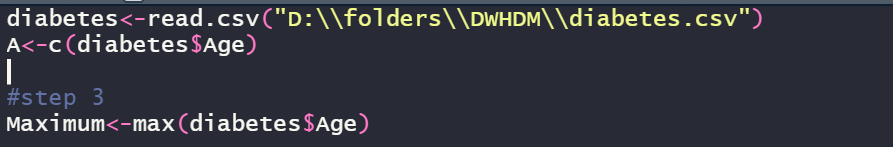
****

**OUTPUT :**



1. **MAXIMUM NORMALIZATION**
   * A<-c(diabetes$Age)
   * Maximum<-max(diabetes$Age)

**INPUT :**

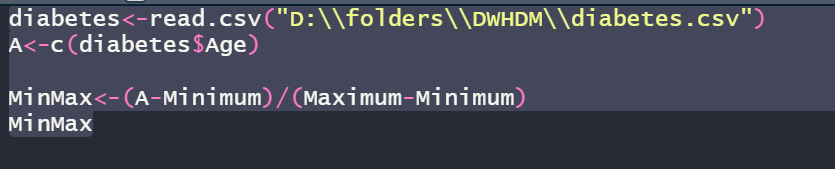
****

**OUTPUT :**

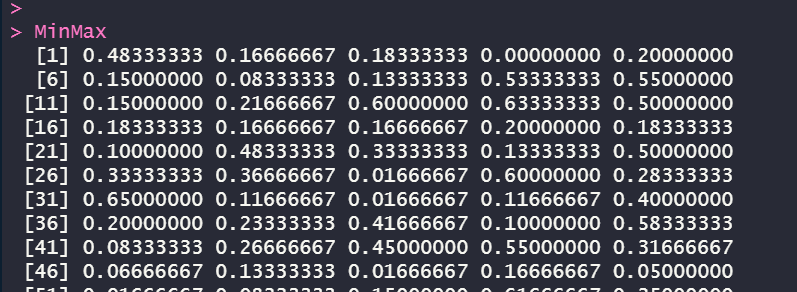
****

1. **MINMAX NORMALIZATION**
   * A<-c(diabetes$Age)
   * MinMax<-(A-Minimum)/(Maximum-Minimum)
   * MinMax

**INPUT :**

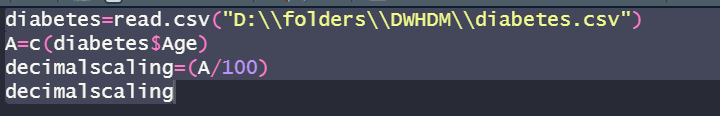


**OUPTUT :**

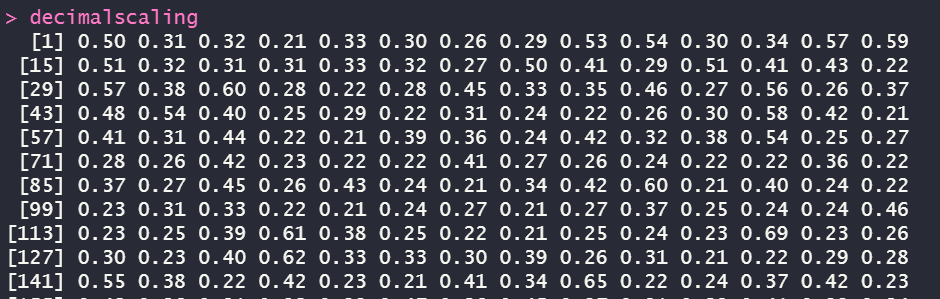
****

1. **DECIMAL** **SCALING** **NORMALIZATION**
   * A=c(diabetes$Age)
   * decimalscaling=(A/100)
   * decimalscaling

**INPUT :**

****

**OUTPUT :**

****

**RESULT:**

Thus, the correlation analysis and normalization for the given dataset has been successfully executed and observed.

**EX.No: 08**

**Date :**

**REGRESSION ANALYSIS USING R TOOL**

**PROBLEM STATEMENT :**

Perform the linear regression and multiple regression for the given dataset.

**DESCRIPTION :**

Consider a dataset of diabetes.csv with the attributes pregnancies, Glucose, BloodPressure, SkinThickness, BMI, Diabetes, Age, Outcome for the analysis. There will be linear regression analysis between Age and BloodPressure. Where, for the multiple regression, the analysis is between Age, BloodPressure, Glucose from the dataset.

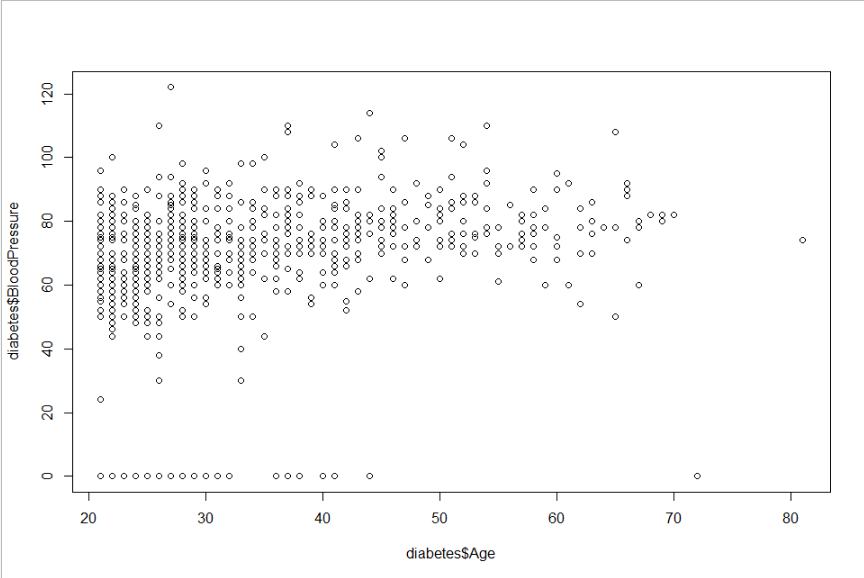
* **LINEAR REGRESSION :**

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.

**INPUT :**

* Relation <- lm(diabetes$BloodPressure~diabetes$Age)
* Png<- (file=”linear regression.png”)
* Plot(diabetes$Age, diabetes$BloodPressure, col=”green”, main= “ Linear Regression Analysis” , abline= (lm(diabetes$BloodPressure~ diabetes$Age)), xlab = “BloodPressure”, ylanb= “Age”)

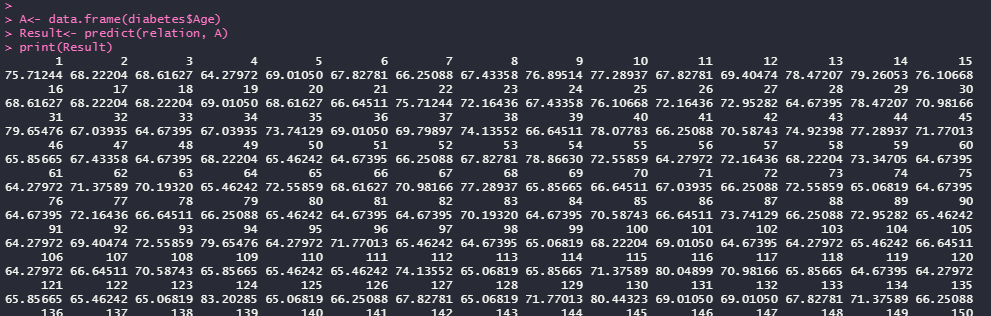
**OUTPUT:**



**INPUT :**

* A<- data.frame(diabetes$Age)
* Result<- predict(relation, A)
* Print(Result)

**OUTPUT :**

****

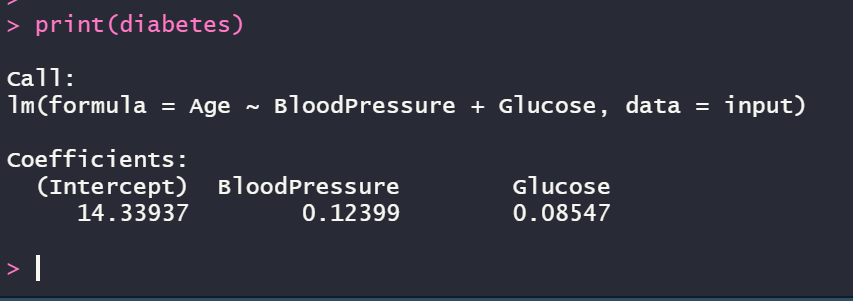
* **MULTIPLE REGRESSION :**

Multiple regression is a statistical tool used to derive the value of a criterion from several other independent, or predictor, variables. It is the simultaneous combination of multiple factors to assess how and to what extent they affect a certain outcome.

**INPUT :**

* Input <- diabetes[,c(“Age”, “BloodPressure”, “Glucose”)]
* Model <- lm(Age~ BloodPressure+Glucose,data=input)
* Print(model)

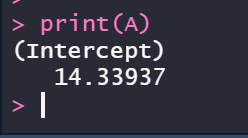
**OUTPUT:**

****

**INPUT :**

* A<- coef(model)[1]
* Print(A)

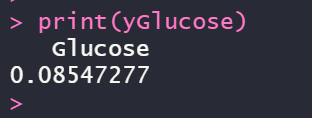
**OUTPUT:**

****

**INPUT :**

* xBloodPressure<- coef(model)[2]
* yGlucose<- coef(model)[3]
* print(xBloodPressure)
* print(yGlucose)

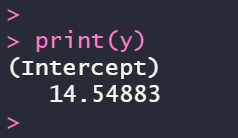
**OUTPUT:**

****

**INPUT :**

* y = A+xBloodPressure + yGlucose
* print(y)

**OUTPUT:**

****

**RESULT :**

Thus, the linear regression and the multiple regression analysis for the given dataset has been successfully completed.