**Mthetheleli Mbeje  
Project 1(Used Titanic dataset on R)**  
  
1.Opening the Dataset  
install.packages("openxlsx")  
require("openxlsx")

> data1<-read.csv("C:\\Users\\Simphiwe Mbeje\\OneDrive\\Desktop\\tested.csv" , sep=",")

> data1

head(data1)

PassengerId Survived Pclass Name

1 892 0 3 Kelly, Mr. James

2 893 1 3 Wilkes, Mrs. James (Ellen Needs)

3 894 0 2 Myles, Mr. Thomas Francis

4 895 0 3 Wirz, Mr. Albert

5 896 1 3 Hirvonen, Mrs. Alexander (Helga E Lindqvist)

6 897 0 3 Svensson, Mr. Johan Cervin

Sex Age SibSp Parch Ticket Fare Cabin Embarked

1 male 34.5 0 0 330911 7.8292 Q

2 female 47.0 1 0 363272 7.0000 S

3 male 62.0 0 0 240276 9.6875 Q

4 male 27.0 0 0 315154 8.6625 S

5 female 22.0 1 1 3101298 12.2875 S

6 male 14.0 0 0 7538 9.2250 S

2.To Handle Missing Values  
> new<-na.omit(data1)

> new

3.Basic Statistical Analysis  
  
> summary(new)

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare

Min. : 892.0 Min. :0.0000 Min. :1.000 Length:331 Length:331 Min. : 0.17 Min. :0.0000 Min. :0.0000 Length:331 Min. : 0.00

1st Qu.: 992.5 1st Qu.:0.0000 1st Qu.:1.000 Class :character Class :character 1st Qu.:21.00 1st Qu.:0.0000 1st Qu.:0.0000 Class :character 1st Qu.: 8.05

Median :1100.0 Median :0.0000 Median :2.000 Mode :character Mode :character Median :27.00 Median :0.0000 Median :0.0000 Mode :character Median : 16.00

Mean :1100.2 Mean :0.3837 Mean :2.142 Mean :30.18 Mean :0.4834 Mean :0.3988 Mean : 40.98

3rd Qu.:1210.5 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:39.00 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.: 40.63

Max. :1307.0 Max. :1.0000 Max. :3.000 Max. :76.00 Max. :8.0000 Max. :6.0000 Max. :512.33

Cabin Embarked

Length:331 Length:331

Class :character Class :character

Mode :character Mode :character

4.To View Data

> str(new)

'data.frame': 331 obs. of 12 variables:

$ PassengerId: int 892 893 894 895 896 897 898 899 900 901 ...

$ Survived : int 0 1 0 0 1 0 1 0 1 0 ...

$ Pclass : int 3 3 2 3 3 3 3 2 3 3 ...

$ Name : chr "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...

$ Sex : chr "male" "female" "male" "male" ...

$ Age : num 34.5 47 62 27 22 14 30 26 18 21 ...

$ SibSp : int 0 1 0 0 1 0 0 1 0 2 ...

$ Parch : int 0 0 0 0 1 0 0 1 0 0 ...

$ Ticket : chr "330911" "363272" "240276" "315154" ...

$ Fare : num 7.83 7 9.69 8.66 12.29 ...

$ Cabin : chr "" "" "" "" ...

$ Embarked : chr "Q" "S" "Q" "S" ...

- attr(\*, "na.action")= 'omit' Named int [1:87] 11 23 30 34 37 40 42 48 55 59 ...

..- attr(\*, "names")= chr [1:87] "11" "23" "30" "34" ...

5. Correlation(between Age & Passanger ID)  
  
> cor.test(new)

Error in cor.test.default(new) : 'x' must be a numeric vector

> cor.test(new$Age,new$PassengerId)

Pearson's product-moment correlation

data: new$Age and new$PassengerId

t = -0.56944, df = 329, p-value = 0.5694

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.13870998 0.07668098

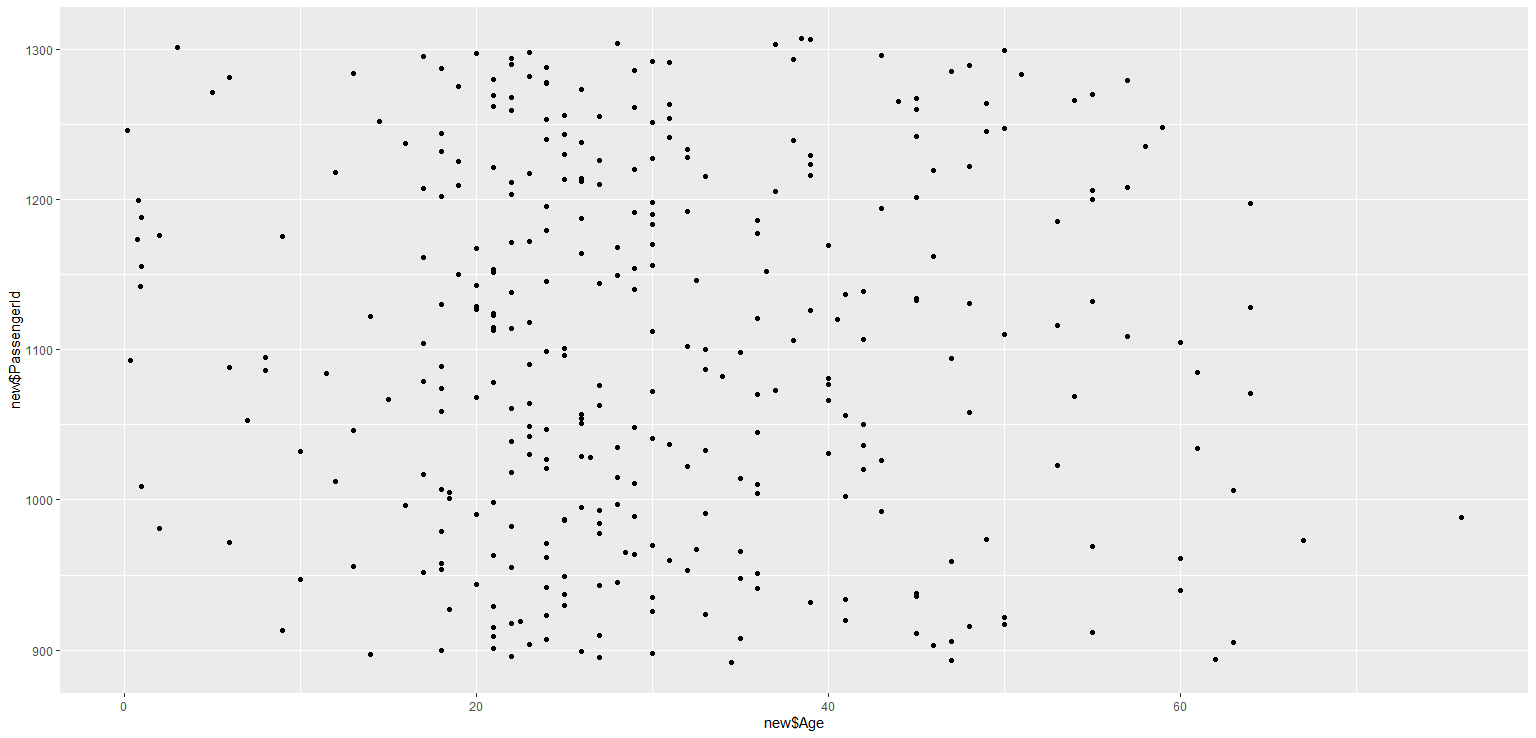
sample estimates:

cor

-0.0313788  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
6. Ploting a Scatter Plot  
  
install.packages("ggplot2")  
require("ggplot2")

> PassengerId<-as.numeric(new$PassengerId)

> new<-data.frame(PassengerId,new)

> plot<-gglot(new,aes.(new$age,new$PassengerId)) + geom\_point()  
  
  


7. Ploting Line plot  
