University of Stirling MATPMD0

Computing Science & Mathematics 2022

**MATPMD0 INTRODUCTORY STATISTICS FOR DATA SCIENCE**

**PROJECT : AUTUMN SEMESTER 2022**

**Submission due 19th December 17:00**

**Student Number: <Student id>**

**Declaration: In submitting this project I declare that this is all my own work and I did not seek help to complete it.**

**For each project question, insert answers below.**

1. Perform an exploratory data analysis, taking care to describe the type of variables in the data set.

**Ans**: In given data set variables are quantitative, numeric, and continuous.

**Code in R language:**

#loading Libraries

library(stats)

library(dplyr)

#1 Data summarizaiton

a= r$BodyFat

b=r$Abdomen

# Calculation  mean of the given data

mean(a)

mean(b)

# Calculation  median of the given data

median(r$BodyFat)

median(r$Abdomen)

# Calculation  mode of the given data

Mode(a)

Mode(b)

#calculate standard deviation

sd(a)

sd(b)

#calculate Variance

var(a)

var(b)

#calculate IQR

quantile(a)

quantile(b)

#Data visualization

str(a)

str(b)

#Numeric Data -Histogram

hist(a)

hist(b)

#Catagorical Data -Barplot

barplot(a)

barplot(b)

**Mode Function in R language:**# Calculation  mode of the given data

Mode=function(x){

  ta=table(x)

  tam=max(ta)

  if(all(ta==tam))

    mod=NA

  else

    if(is.numeric(x))

      mod=as.numeric(names(ta)[ta==tam])

  else

    mod=names(ta)[ta==tam]

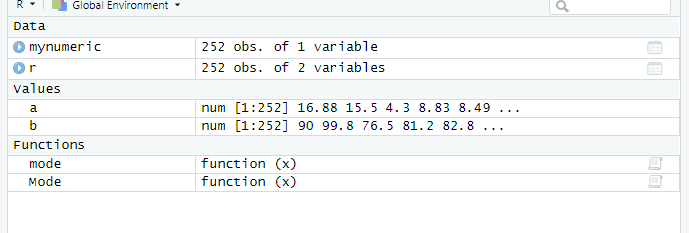
  return(mod)

}

**Outputs:**

**a=Bodyfat**

**b=Abdomen**

**A picture containing text

Description automatically generatedApplication

Description automatically generated with low confidenceA picture containing graphical user interface

Description automatically generatedA picture containing graphical user interface

Description automatically generatedGraphical user interface, application

Description automatically generatedText, letter

Description automatically generatedText

Description automatically generated with medium confidence**

**Histogram of a(BodyFat)& Histogram of b(Abdomen)**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

**Bar plot of a(Bodyfat)& Barplot of b(Abdomen)**

**Chart, bar chart

Description automatically generatedChart, bar chart

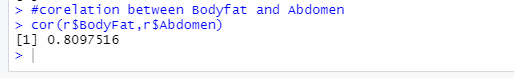
Description automatically generated**

1. Calculate the correlation coefficient for the two variables given and comment on the relationship between body fat and abdomen size.

**Code:**

#corelation between Bodyfat and Abdomen

cor(r$BodyFat,r$Abdomen)



3. Investigate a model to test the relationship between body fat and abdomen size. You must include output from R to support your findings. Details you should include are:

* **a description of the model.**

Alternative hypothesis :True difference in means is equal to 0

* **a summary of the fitted model with interpretation of test statistics and parameter estimates;**

**T-test**

**t.test(r$BodyFat,r$Abdomen)**

**Text

Description automatically generated**

* **evidence as to whether assumptions of the model have been met;**

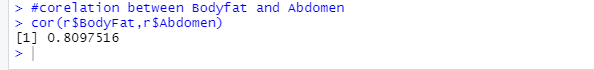
T-Test

* **conduct a formal test to question whether there is a significant linear relationship between body fat and abdomen size.**

**Corelation**

#corelation between Bodyfat and Abdomen

cor(r$BodyFat,r$Abdomen)

****

4. Use the equation to predict the percentage of body fat for a male whose abdomen measures 100cm.

Graphical user interface, text

Description automatically generated

Chart, scatter chart

Description automatically generated

5. Assess the predictive performance of the model.

**Prediction:**

v=lm(r$BodyFat~r$Abdomen,data = r)

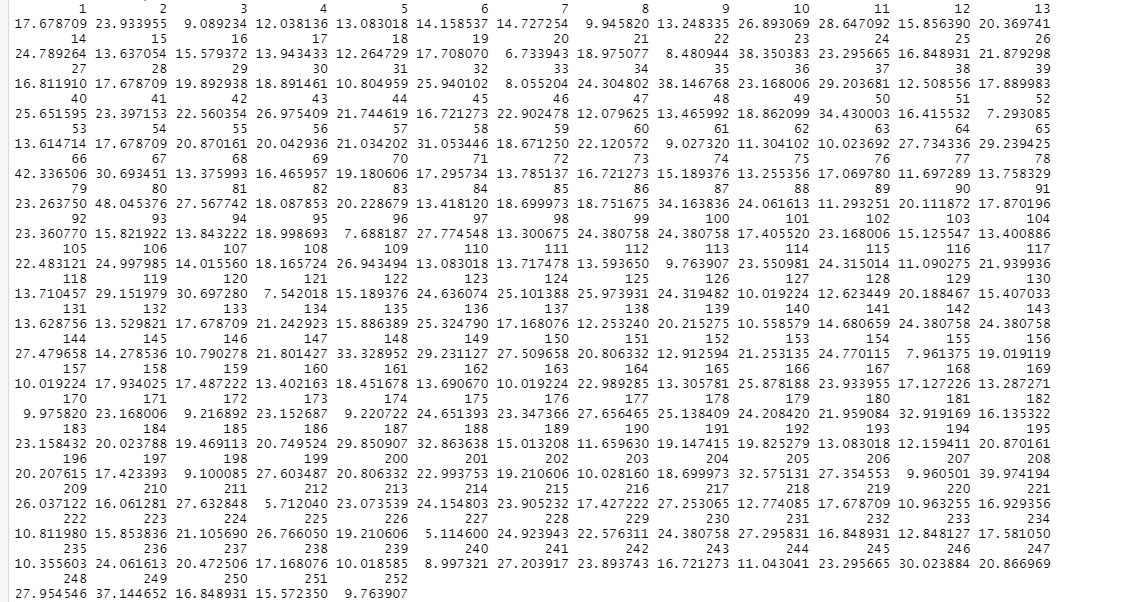
base\_model <- lm(a ~ b + a, data = r)

base\_model\_output <- predict(base\_model, newdata = r)

base\_model\_output

#Prediction

cbind(base\_model\_output)



6. In this final section include all R code that you have used for this project verbatim. Ensure that:

* the code for each question can be easily found;
* all code is adequately commented;
* variable names are sensible.

#loading Libraries

library(stats)

library(dplyr)

#1 Data summarizaiton

a= r$BodyFat

b=r$Abdomen

# Calculation  mean of the given data

mean(a)

mean(b)

# Calculation  median of the given data

median(r$BodyFat)

median(r$Abdomen)

# Calculation  mode of the given data

Mode(a)

Mode(b)

#calculate standard deviation

sd(a)

sd(b)

#calculate Variance

var(a)

var(b)

#calculate IQR

quantile(a)

quantile(b)

#Data visualization

str(a)

str(b)

#Numeric Data -Histogram

hist(a)

hist(b)

#Catagorical Data -Barplot

barplot(a)

barplot(b)

cor(a,b)

#T test

t.test(r$BodyFat,r$Abdomen)

#linear Regression Model

v=lm(r$BodyFat~r$Abdomen,data = r)

base\_model <- lm(a ~ b + a, data = r)

base\_model\_output <- predict(base\_model, newdata = r)

base\_model\_output

#Prediction

cbind(base\_model\_output)

**Code of Mode:**

# Calculation  mode of the given data

Mode=function(x){

  ta=table(x)

  tam=max(ta)

  if(all(ta==tam))

    mod=NA

  else

    if(is.numeric(x))

      mod=as.numeric(names(ta)[ta==tam])

  else

    mod=names(ta)[ta==tam]

  return(mod)

}