**STAT 46700/ CS 59000 Topics in Data Science Spring 2025**

**Homework 2**

**[Vaishak Balachandra]**

*Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations. For example, suppose I ask for how many numbers are greater than 5 in the vector, x=c(1,9,2,8,10,12,15). Do not simply count the number of by hand, instead let the R count the numbers by using appropriate code.*

**Q.N. 1)** Consider points **A** (1,1,7), **B** (2,9,5), **C** (9,6,3), **D**(3,5,7) in a three-dimensional space. Using R

a) Calculate the Euclidean distance matrix

b) Calculate the Manhattan distance matrix.

> #### Q1

>

> # a -- Eucledian

> A = c(1,1,7)

> B = c(2,9,5)

> C = c(9,6,3)

> D = c(3,5,7)

> Q1 <- rbind(A,B,C,D)

> Q1

[,1] [,2] [,3]

A 1 1 7

B 2 9 5

C 9 6 3

D 3 5 7

>

>

> cat("Euclidean Distances between the points are given by:")

Euclidean Distances between the points are given by:

> eucl\_distance = dist(Q1, method = "euclidean")

> # or

> # install.packages("philentropy")

> # library(philentropy)

> # eucl\_distance = distance(Q1, method = 'euclidean')

> eucl\_distance

A B C

B 8.306624

C 10.246951 7.874008

D 4.472136 4.582576 7.280110

> #

> # Euclidean Distance between A and B is 8.306624 units

> # Euclidean Distance between A and C is 10.246951 units

> # Euclidean Distance between A and D is 4.472136 units

> # Euclidean Distance between B and C is 7.874008 units

> # Euclidean Distance between B and D is 4.582576 units

> # Euclidean Distance between C and D is 7.280110 units

>

>

> # b -- Manhattan

> cat("Manhattan Distances between the points are given by:")

Manhattan Distances between the points are given by:

> manhattan\_distance = dist(Q1, method = "manhattan")

> # or

> # manhattan\_distance = distance(Q1, method = "manhattan")

> manhattan\_distance

A B C

B 11

C 17 12

D 6 7 11

> #

> # Manhattan Distance between A and B is 11 units

> # Manhattan Distance between A and C is 17 units

> # Manhattan Distance between A and D is 6 units

> # Manhattan Distance between B and C is 12 units

> # Manhattan Distance between B and D is 7 units

> # Manhattan Distance between C and D is 11 units

**Q. N. 2)** The "mtcars" data set, readily available in R, provides information about 32 different car models, including details like miles per gallon (mpg), number of cylinders (cyl), displacement (disp), horsepower (hp), weight (wt), and other performance metrics. Select the variable *mpg* from the dataset and rescale the values using z-score and minmax (0,1) method.

> #### Q2

>

> # a

> data(mtcars)

> head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

> names(mtcars)

[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear" "carb"

> dim(mtcars)

[1] 32 11

> Q2 <- mtcars[,1]

> Q2

[1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7

[18] 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4

> length(Q2)

[1] 32

>

> cat("Zscore representation of the mpg's are given by:")

Zscore representation of the mpg's are given by:

> zscore <- scale(Q2)

> zscore

[,1]

[1,] 0.15088482

[2,] 0.15088482

[3,] 0.44954345

[4,] 0.21725341

[5,] -0.23073453

[6,] -0.33028740

[7,] -0.96078893

[8,] 0.71501778

[9,] 0.44954345

[10,] -0.14777380

[11,] -0.38006384

[12,] -0.61235388

[13,] -0.46302456

[14,] -0.81145962

[15,] -1.60788262

[16,] -1.60788262

[17,] -0.89442035

[18,] 2.04238943

[19,] 1.71054652

[20,] 2.29127162

[21,] 0.23384555

[22,] -0.76168319

[23,] -0.81145962

[24,] -1.12671039

[25,] -0.14777380

[26,] 1.19619000

[27,] 0.98049211

[28,] 1.71054652

[29,] -0.71190675

[30,] -0.06481307

[31,] -0.84464392

[32,] 0.21725341

attr(,"scaled:center")

[1] 20.09062

attr(,"scaled:scale")

[1] 6.02694

>

>

> # b

> cat("Min-Max(0,1) representation of the mpg's are given by:")

Min-Max(0,1) representation of the mpg's are given by:

> install.packages("scales")

> library(scales)

> min\_max\_score <- rescale(Q2)

> # or

> # min\_max\_score <- scale(Q2, center = min(Q2), scale = max(Q2) - min(Q2))

> min\_max\_score

[1] 0.4510638 0.4510638 0.5276596 0.4680851 0.3531915 0.3276596 0.1659574 0.5957447

[9] 0.5276596 0.3744681 0.3148936 0.2553191 0.2936170 0.2042553 0.0000000 0.0000000

[17] 0.1829787 0.9361702 0.8510638 1.0000000 0.4723404 0.2170213 0.2042553 0.1234043

[25] 0.3744681 0.7191489 0.6638298 0.8510638 0.2297872 0.3957447 0.1957447 0.4680851

**Q. N. 3)** An article entitled“Investigating the Effect of Task Complexities on the Response Time of Human Operators to Perform Emergency Tasks of Nuclear Power Plants” by J. Park and S. Cho was published in *Annals of Nuclear Energy in 2010*. Times for NPP workers to complete 35 emergency tacks of varying complexity (measured by TACOM score) are recorded. Tasks were completed by US and non-US workers. The data are provided in the link below:

<http://users.stat.ufl.edu/~winner/data/nuclear_time.dat>

The variables included in the data are: Task, Nationality (1= US, 0= non-US), Time to complete the task and TACOM complexity score.

1. Import the data in R and print first 5 observations.
2. Display the completion time of the task based on nationality.
3. Is there a significance difference in the completion time between US and non-US workers?

> #### Q3

>

> # a

> Q3 <- read.table("https://users.stat.ufl.edu/~winner/data/nuclear\_time.dat")

> head(Q3,5)

V1 V2 V3 V4

1 1 1 60 3.788

2 2 1 336 6.108

3 3 1 432 6.283

4 4 1 516 6.373

5 5 1 624 6.459

> names(Q3) = c("Task","Nationality","Time\_to\_Complete","TACOM\_Complexity\_Score")

> head(Q3,5)

Task Nationality Time\_to\_Complete TACOM\_Complexity\_Score

1 1 1 60 3.788

2 2 1 336 6.108

3 3 1 432 6.283

4 4 1 516 6.373

5 5 1 624 6.459

> attach(Q3)

>

>

> # b

> boxplot(Time\_to\_Complete~Nationality, main = "Boxplot of Time to Complete based on Nationality", xlab = "Nationality (1= US, 0= non-US)", ylab = "Time to Complete", col = c(5,12), col.main = "orange", col.lab = "purple")

>

>

> # c

> t.test(Time\_to\_Complete~Nationality)

Welch Two Sample t-test

data: Time\_to\_Complete by Nationality

t = -3.8266, df = 52.494, p-value = 0.0003475

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-311.99847 -97.37296

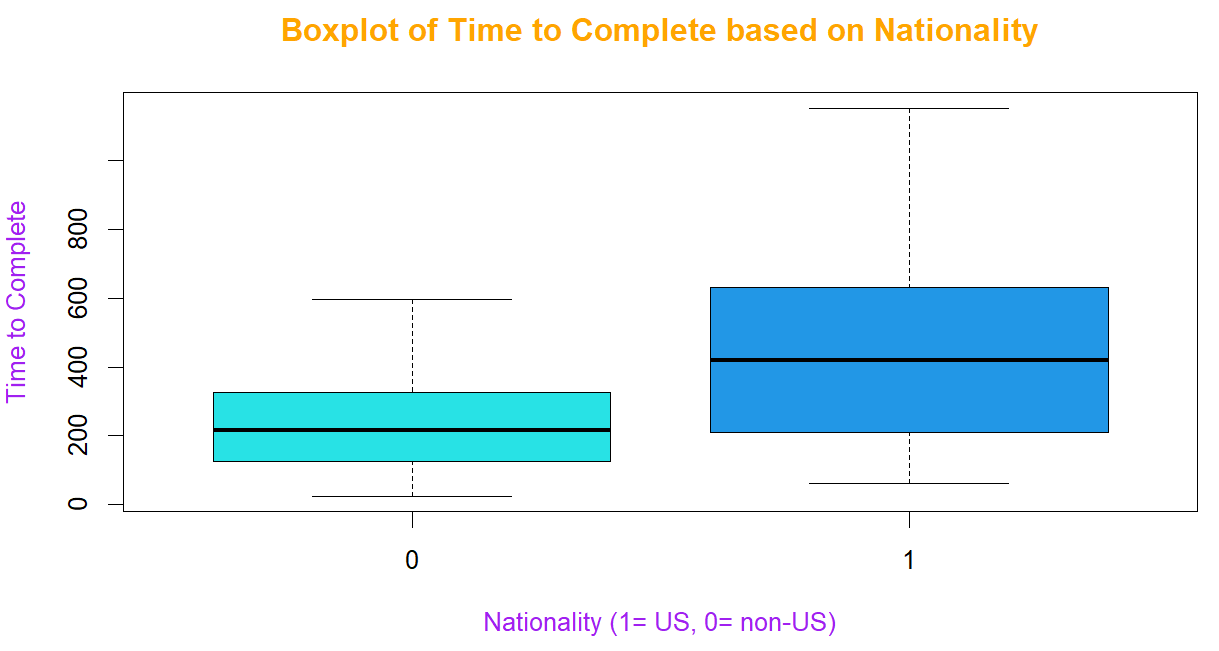
sample estimates:

mean in group 0 mean in group 1

235.2000 439.8857

> cat("Here, the pvalue = 0.0003475, which is less than 0.05. Thus, we reject the null hypothesis and conclude that there is a significant difference in the time to complete the task between the US and the non-US individuals.")

Here, the pvalue = 0.0003475, which is less than 0.05. Thus, we reject the null hypothesis and conclude that there is a significant difference in the time to complete the task between the US and the non-US individuals.



**Q. N. 4)** The dataset provided in the link below contains the prices of ladies' diamond rings and the carat size of their diamond stones. The rings are made with gold of 20 carats purity and are each mounted with a single diamond stone. The article associated with this dataset appears in the Journal of Statistics Education, Volume 4, Number 3 (1996).

<http://jse.amstat.org/datasets/diamond.dat.txt>

1. Import the data in R
2. The first column represent the size of the diamond in carats and second column is the Price of the ring (in Singapore dollars). Insert the variable names: Size and Price for column 1 and column 2 respectively.
3. Display the Size and Price of the rings using scatterplot.
4. Fit a simple linear regression model using Price as a response variable. State the equation of the linear regression model.
5. Test for significance of the size to determine the price of the ring.
6. Predict the price of a ring of size 0.24 carat. Construct a 95% confidence interval and prediction interval for the predicted price.

> #### Q4

>

> # a

> Q4 <- read.table("https://jse.amstat.org/datasets/diamond.dat.txt")

> head(Q4)

V1 V2

1 0.17 355

2 0.16 328

3 0.17 350

4 0.18 325

5 0.25 642

6 0.16 342

>

> # b

> colnames(Q4) = c("Size","Price")

> head(Q4)

Size Price

1 0.17 355

2 0.16 328

3 0.17 350

4 0.18 325

5 0.25 642

6 0.16 342

> attach(Q4)

>

> # c

> plot(Size, Price, main = "Prize against Size", pch = 17, col = 2, cex = 1.5, xlab = "Size(in carats)", ylab = "Prize(in Singapore Dollar)", col.lab = "orange", col.main = "maroon",ylim = c(100, 1250))

>

> # d

> model = lm(Price~Size)

> model

Call:

lm(formula = Price ~ Size)

Coefficients:

(Intercept) Size

-259.6 3721.0

> abline(model, lwd = 2, col = "green")

> cat("Fitted Model:

+ Price = -259.6 + 3721\*Size")

Fitted Model:

Price = -259.6 + 3721\*Size

>

> # e

> summary(model)

Call:

lm(formula = Price ~ Size)

Residuals:

Min 1Q Median 3Q Max

-85.159 -21.448 -0.869 18.972 79.370

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -259.63 17.32 -14.99 <2e-16 \*\*\*

Size 3721.02 81.79 45.50 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778

F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

> cat("Here, the pvalue = 2.2e-16, which is less than 0.05. Thus we reject the null hypothesis, and conclude that the size of the diamond (in carats) has a significant positive relationship with its price.")

Here, the pvalue = 2.2e-16, which is less than 0.05. Thus we reject the null hypothesis, and conclude that the size of the diamond (in carats) has a significant positive relationship with its price.

>

> # f

> predict(model, data.frame(Size = 0.24))

1

633.4201

> cat("For 0.24 carat, the predicted Price is 633.4201 Singaporean Dollar")

For 0.24 carat, the predicted Price is 633.4201 Singaporean Dollar

> predict(model, data.frame(Size = 0.24), interval = "confidence", level = 0.95)

fit lwr upr

1 633.4201 622.4484 644.3917

> cat("95% Confidence Interval: for 0.24 carat, the predicted price lies in [622.4484, 644.3917]")

95% Confidence Interval: for 0.24 carat, the predicted price lies in [622.4484, 644.3917]

> predict(model, data.frame(Size = 0.24), interval = "prediction", level = 0.95)

fit lwr upr

1 633.4201 568.3961 698.444

> cat("95% Prediction Interval: for 0.24 carat, the predicted price lies in [568.3961, 698.444]")

95% Prediction Interval: for 0.24 carat, the predicted price lies in [568.3961, 698.444]

