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

**Homework 2-Solution Due: February 12, 2025**

*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

*Solution:*

> A=c(1,1,7)

> B=c(2,9,5)

> C=c(9,6,3)

> D=c(3,5,7)

> x=rbind(A,B,C,D)

> M=dist(x,method='euclidean')

> as.matrix(M)

A B C D

A 0.000000 8.306624 10.246951 4.472136

B 8.306624 0.000000 7.874008 4.582576

C 10.246951 7.874008 0.000000 7.280110

D 4.472136 4.582576 7.280110 0.000000

b) Calculate the Manhattan distance matrix.

*Solution:*

> M=dist(x,method='manhattan')

> as.matrix(M)

A B C D

A 0 11 17 6

B 11 0 12 7

C 17 12 0 11

D 6 7 11 0

**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.

Solution:

> data(mtcars)

> attach(mtcars)

> head(mpg)

[1] 21.0 21.0 22.8 21.4 18.7 18.1

> # z-score method

> round(scale(mpg),3)

[,1]

[1,] 0.151

[2,] 0.151

[3,] 0.450

[4,] 0.217

[5,] -0.231

[6,] -0.330

[7,] -0.961

[8,] 0.715

[9,] 0.450

[10,] -0.148

[11,] -0.380

[12,] -0.612

[13,] -0.463

[14,] -0.811

[15,] -1.608

[16,] -1.608

[17,] -0.894

[18,] 2.042

[19,] 1.711

[20,] 2.291

[21,] 0.234

[22,] -0.762

[23,] -0.811

[24,] -1.127

[25,] -0.148

[26,] 1.196

[27,] 0.980

[28,] 1.711

[29,] -0.712

[30,] -0.065

[31,] -0.845

[32,] 0.217

attr(,"scaled:center")

[1] 20.09062

attr(,"scaled:scale")

[1] 6.026948

> round(scale(mpg, center = min(mpg), scale = (max(mpg)-min(mpg))),3)

[,1]

[1,] 0.451

[2,] 0.451

[3,] 0.528

[4,] 0.468

[5,] 0.353

[6,] 0.328

[7,] 0.166

[8,] 0.596

[9,] 0.528

[10,] 0.374

[11,] 0.315

[12,] 0.255

[13,] 0.294

[14,] 0.204

[15,] 0.000

[16,] 0.000

[17,] 0.183

[18,] 0.936

[19,] 0.851

[20,] 1.000

[21,] 0.472

[22,] 0.217

[23,] 0.204

[24,] 0.123

[25,] 0.374

[26,] 0.719

[27,] 0.664

[28,] 0.851

[29,] 0.230

[30,] 0.396

[31,] 0.196

[32,] 0.468

attr(,"scaled:center")

[1] 10.4

attr(,"scaled:scale")

[1] 23.5

**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?

*Solution:*

1. *We use R code below to import and print first 5 observations*

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

> attach(Q3)

> head(Q3,5)

V1 V2 V3 V4

1 1 1 60 3.788

2 2 1 336 6.108

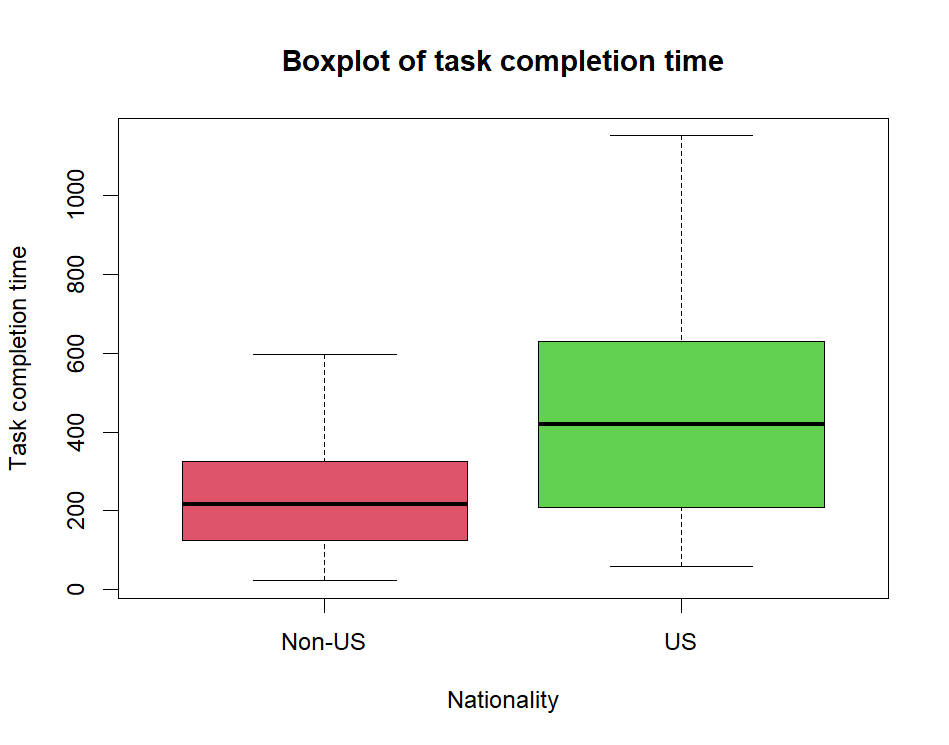
3 3 1 432 6.283

4 4 1 516 6.373

5 5 1 624 6.459

1. We are given that V2 represents the nationality and V3 represent the completion time.

> boxplot(V3~V2, names=c("Non-US", "US"), col=c(2,3), xlab="Nationality",ylab="Task completion time", main="Boxplot of task completion time")



**Using ggplot**

|  |
| --- |
| > Q3=data.frame(V1,V2,V3,V4)  > head(Q3)  V1 V2 V3 V4  1 1 US 60 3.788  2 2 US 336 6.108  3 3 US 432 6.283  4 4 US 516 6.373  5 5 US 624 6.459  6 6 US 708 6.572  > V2=factor(V2,labels=c("Non-US","US"))  > library("ggplot2")  > color <- c("purple", "green")  > line <- "goldenrod2"  > ggplot(Q3,aes(group=V2,x=V2, y=V3))+ geom\_boxplot(fill=color)  + scale\_x\_discrete(name="Nationality") + scale\_y\_continuous("Task complition time")+  + ggtitle("Task completion time by nationality")+theme(plot.title=element\_text(hjust=0.5)) |
|  |
| |  | | --- | |  | |

Chart, box and whisker chart

Description automatically generated

1. *We want to test*

*H0: Average task completion time for non-US worker = Average completion time for US worker*

*Ha: Average task completion time for non-US worker ≠ Average completion time for US worker*

*Based on the computer output below we have p-value=0.0003475 which is less than 0.05 so we reject the null hypothesis and conclude that there is sufficient evidence that the average completion time differ based on the nationality.*

> t.test(V3~V2)

Welch Two Sample t-test

data: V3 by V2

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

alternative hypothesis: true difference in means 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

**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.

*Solution:*

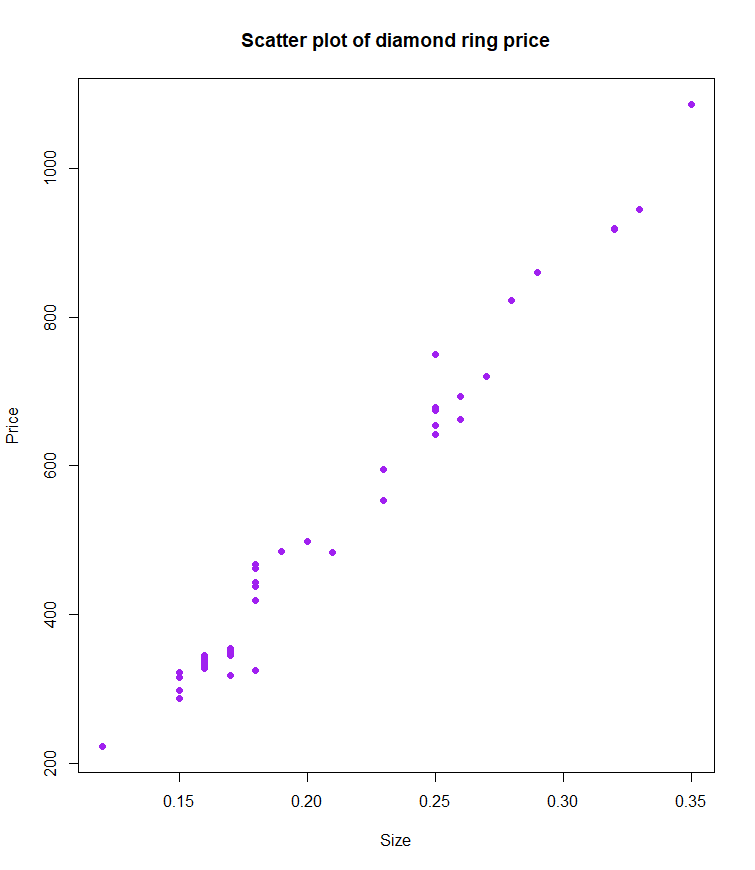
1. *We use R code below to import the data in R*

|  |
| --- |
| > Q4=read.table("http://jse.amstat.org/datasets/diamond.dat.txt")  > attach(Q4)  > 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 |
|  |
| |  | | --- | |  | |

1. *We have updated the variable names as below*

|  |
| --- |
| > names(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 |
|  |
| |  | | --- | |  | |

1. > plot(Size, Price, pch=19, col="purple", main="Scatter plot of diamond ring price")



**Using ggplot**

>p=ggplot(Q4, aes(x=Size, y=Price))

>p+geom\_smooth(method="lm")+geom\_point(size=2)

Chart, scatter chart

Description automatically generated

1. *The fitted model is*

*Price=-259.6+3721\*size*

> model=lm(Price~Size)

> model

Call:

lm(formula = Price ~ Size)

Coefficients:

(Intercept) Size

-259.6 3721.0

1. *We used R code below to print the results from the test.*

*We would like to test whether the size is a significant variable to determine the price. Which is equivalent to test*

*Based on the p-value of the test we can reject the null hypothesis and conclude that Size of the diamond is a significant variable to determine the price of the ring.*

> summary(model)$coefficients

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

(Intercept) -259.6259 17.31886 -14.99094 2.523271e-19

Size 3721.0249 81.78588 45.49715 6.751260e-40

1. *Based on the R output below it is predicted that the Price of 0.24 carat ring is 633.4201 with 95% confidence and prediction interval (622.4484, 644.3917) and (568.3961, 698.444) respectively.*

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

1

633.4201

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

fit lwr upr

1 633.4201 622.4484 644.3917

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

fit lwr upr

1 633.4201 568.3961 698.444