Title Module 2: Elementary Probability Theory

Requirement: The dataset must be stored on the desktop.

Introduction: We have a dataset with 4 attributes and 50 observations. The dataset is about the projects that a company has undertaken in past and now wants to analyses the impact and value each one of them holds. Each observation lies in a separate category depending on all the attributes and their values. The 4 attributes are as follows: 1- PIP - A unique key to each observation. 2- Quality - That defines how good the quality of each project was. 3- Speed- That defines as to how much time each project took to be built. 4- Cost- The last one defines how much money company has spent in that project.

With analysis of these factors the manager can submit the report to CEO and can give away a proposal to draw some guidelines before executing any project in future.

Here I am going to submit my report and the findings. There will be three separate files, one csv file that I created that holds the data table, second is excel file that defines all the notations, and third one is report file in as a markdown document.

Part1:

We need to find the probability of each of the three categories our project lies into. Our universal set consists of 50 data points and all are into different category as we will see further.

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
q <- length(which(df$QUALITY > 500))
s <- length(which(df$SPEED < 13))
c <- length(which(df$COST < 234000))
count <- sum(q,s,c)
prob_q <- (q/count)* 100
prob_s <- (s/count) * 100
prob_c <- (c/count) * 100

print( paste("Count is:",q,"Probability of QUALITY:",round(prob_q,)))
## [1] "Count is: 29 Probability of QUALITY: 38"

print(paste("Count is:",s,"Probability of SPEED:",round(prob_s)))
## [1] "Count is: 22 Probability of SPEED: 29"

print(paste("Count is:",c,"Probability of COST:",round(prob_c)))
## [1] "Count is: 26 Probability of COST: 34"
```

We see that each probability represents the number of respective data points lies in that category. Note that the probability is rounded off to nearest integer so adding them all might exceed the 100% value by one integer.

Project with desired quality are 38. Projects that are done under desired timeframe are 29. Projects that does not exceed the allotted money expenditure are 34.

Part2:

There are 8 different scores that are assigned to each observation based on certain they meet. They are as follows:

1- None of the three=0
2- Quality only = 1
3- Speed only = 2
4- Cost only = 3
5- Quality and speed but not the Cost = 4
6- Quality and cost but not the speed = 5 7
7- Speed and cost but not the Quality = 6
8- All three are met = 8

Now we are going to look as to how much count each one of them has in order to find their probability.

```
#None of the three

df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] < 500) && (df$SPEED[i] >= 13) && (df$COST[i] >
234000)){score$scores[i] <- 1
    }
            }
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 6 and probability is: 12"



#Quality only
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] > 500) && (df$SPEED[i] >= 13) && (df$COST[i] >
234000)){score$scores[i] <- 1
    }
        }
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 9 and probability is: 18"
```

```r
#Speed only
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
 for(i in 1:nrow(df)) {
   if((df$QUALITY[i] < 500) && (df$SPEED[i] < 13) && (df$COST[i] >
234000)){score$scores[i] <- 1
    }

        }
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 5 and probability is: 10"



##Cost only
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] < 500) && (df$SPEED[i] >= 13) && (df$COST[i] <
234000)){score$scores[i] <- 1
    }
        }
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 4 and probability is: 8"



##Quality and speed but not the Cost

df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] > 500) && (df$SPEED[i] < 13) && (df$COST[i] >
234000)){score$scores[i] <- 1
}

}
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 4 and probability is: 8"



#Quality and cost but not the speed
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
```

```
   if((df$QUALITY[i] > 500) && (df$SPEED[i] >= 13) && (df$COST[i] <
234000)){score$scores[i] <- 1
}

}
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 9 and probability is: 18"



#SPeed and cost but not the Quality

df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] < 500) && (df$SPEED[i] < 13) && (df$COST[i] <
234000)){score$scores[i] <- 1
}

}
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 6 and probability is: 12"



#All three are met

df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
score <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(score) <- c("scores")
for(i in 1:nrow(df)) {
   if((df$QUALITY[i] > 500) && (df$SPEED[i] < 13) && (df$COST[i] <
234000)){score$scores[i] <- 1
}
}
counts<- sum(as.numeric(score$scores), na.rm=TRUE)
probability <- (counts/50)*100
print(paste("count is:",counts,"and","probability is:",probability))

## [1] "count is: 7 and probability is: 14"
```

We can see that the highest probability are of those that are “meeting the Quality criterion only” and “Quality and cost but not the speed”. Least are the probability that meets “cost only” criterion and “Quality and speed but not the Cost” criterion.

Part3:

To clearly define the findings of part 2 analysis, I am creating a Venn diagram with each section clearly defining how many data points lies in each one of them. To see the Venn diagram open the excel file and tally part2 analysis from that Venn diagram.

Part4:

a) Of those who satisfied Cost, what percentage also satisfied Speed?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(c) <- c("scores")
colnames(cs) <- c("scores")
for(i in 1:nrow(df)) {
  if((df$COST[i] < 234000)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i] < 234000) && (df$SPEED[i] < 13)){cs$scores[i] <- 1
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
prob <- (count_cs/count_c)*100
print(paste("percentage is:",prob,"%"))
```

```
## [1] "percentage is: 50 %"
```

We see that 50% of those who satisfies cost also satisfies speed. That means half of those who stratifies cost criteria also satisfies speed criterion.

b) Of those who satisfied Quality, what percentage also satisfied Cost?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(c) <- c("scores")
colnames(cs) <- c("scores")
for(i in 1:nrow(df)) {
  if((df$QUALITY[i] > 500)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$QUALITY[i] > 500) && (df$COST[i] < 234000)){cs$scores[i] <- 1
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
prob <- (count_cs/count_c)*100
print(paste("percentage is:",round(prob,2),"%"))
```

```
## [1] "percentage is: 55.17 %"
```

55% of those who satisfies quality do satisfies cost also.

c) Of those who satisfied Quality, what percentage also satisfied Speed but did not satisfy the Cost?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(c) <- c("scores")
colnames(cs) <- c("scores")
for(i in 1:nrow(df)) {
  if((df$QUALITY[i] > 500)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$QUALITY[i] > 500) && (df$COST[i] > 234000)&& (df$SPEED[i] <
13)){cs$scores[i] <- 1
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
prob <- (count_cs/count_c)*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 13.79 %"
```

d) Of those who satisfied Cost, what percentage also satisfied Speed but did not satisfy the Quality?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))

colnames(c) <- c("scores")
colnames(cs) <- c("scores")

for(i in 1:nrow(df)) {
  if((df$COST[i] < 234000)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i]<234000) &&(df$SPEED[i] < 13) && (df$QUALITY[i] <
500)){cs$scores[i] <- 1
    }
        }

count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
prob <- (count_cs/(count_c))*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 23.08 %"
```

e) Of those who did not satisfy Speed, what percentage satisfied Quality and Cost?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
ex <- data.frame(matrix(0,ncol=1,nrow=50))

colnames(c) <- c("scores")
colnames(cs) <- c("scores")
colnames(ex) <- c("scores")

for(i in 1:nrow(df)) {
  if((df$SPEED[i] >= 13)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i]<234000) &&(df$SPEED[i] >= 13) && (df$QUALITY[i]
>500)){cs$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i]>234000) && (df$SPEED[i] >= 13) && (df$QUALITY[i] <
500)){ex$scores[i] <- 1
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
count_ex <-sum(as.numeric(ex$scores),na.rm=TRUE)

prob <- ((count_cs)/(count_c-count_ex))*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 40.91 %"
```

f) What percentage satisfied exactly two of the three criteria?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
css <- data.frame(matrix(0,ncol=1,nrow=50))
ex <- data.frame(matrix(0,ncol=1,nrow=50))

colnames(c) <- c("scores")
colnames(cs) <- c("scores")
colnames(css) <- c("scores")
colnames(ex) <- c("scores")

for(i in 1:nrow(df)) {
  if((df$SPEED[i] < 13) && (df$QUALITY[i]>500) ){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)) {
    if((df$COST[i]<234000) &&(df$SPEED[i] < 13)){cs$scores[i] <- 1
    }
        }

for(i in 1:nrow(df)) {
    if((df$COST[i]<234000) && (df$QUALITY[i] >500)){css$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)) {
    if((df$COST[i]<234000) && (df$QUALITY[i] >500) && (df$SPEED[i]<13)){
ex$scores[i] <- 1
```

```
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
count_css <- sum(as.numeric(css$scores),na.rm=TRUE)
count_ex <- sum(as.numeric(ex$scores),na.rm=TRUE)

total <- 50
numerator <- (count_c+count_cs+count_css) - 3*count_ex
prob <- (numerator/total)*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 38 %"
```

g) Of those who satisfied at least one of the three criteria, what percentage satisfied exactly one criterion?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
css <- data.frame(matrix(0,ncol=1,nrow=50))
ex <- data.frame(matrix(0,ncol=1,nrow=50))

colnames(c) <- c("scores")
colnames(cs) <- c("scores")
colnames(css) <- c("scores")
colnames(ex) <- c("scores")

for(i in 1:nrow(df)){
  if((df$COST[i] > 234000) &&(df$SPEED[i] < 13) && (df$QUALITY[i] <
500)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i] < 234000) &&(df$SPEED[i] >= 13) && (df$QUALITY[i] <
500)){cs$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i] > 234000) &&(df$SPEED[i] >= 13) && (df$QUALITY[i] >
500)){css$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i] > 234000) &&(df$SPEED[i] >= 13) && (df$QUALITY[i] <
500)){ex$scores[i] <- 1
    }
        }
total<- 50
count_c <-sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
count_css <-sum(as.numeric(css$scores),na.rm=TRUE)
count_ex <-sum(as.numeric(ex$scores),na.rm=TRUE)
single <- sum(count_c,count_cs,count_css)
inverse <- total-count_ex
prob <- (single/inverse)*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 40.91 %"
```

h) Of those who did not satisfy Cost, what percentage satisfied the Speed criterion?

```
df <- read.table("project_data.csv", sep=",", header=TRUE)
colnames(df) <- c("PIP","QUALITY","SPEED","COST")
c <- data.frame(matrix(0,ncol=1,nrow=50))
cs <- data.frame(matrix(0,ncol=1,nrow=50))
colnames(c) <- c("scores")
colnames(cs) <- c("scores")
for(i in 1:nrow(df)) {
  if((df$COST[i] > 234000)){c$scores[i] <- 1
    }
        }
for(i in 1:nrow(df)){
    if((df$COST[i] > 234000) && (df$SPEED[i] < 13)){cs$scores[i] <- 1
    }
        }
count_c <- sum(as.numeric(c$scores),na.rm=TRUE)
count_cs <-sum(as.numeric(cs$scores),na.rm=TRUE)
prob <- (count_cs/count_c)*100
print(paste("percentage is:",round(prob,2),"%"))

## [1] "percentage is: 37.5 %"
```

Takeaways: We see that our company might been spending more money on projects then we should. My claim is based on the fact that there are only four observation that satisfies only cost criteria. On the positive side, we have seen that most of our projects are satisfying the quality criteria so we. We can be rest assured by that and can start focusing on other factors and not the quality part.

Almost equal half of the projects did not get completed on time that implies that we can work with our managing teams to implement some new norms in our policies, for example, if we need to increase the number of employees we have per team.