MGT – 6203: Homework 3

100 points: Part 1 (60 points Quiz) + Part 2 (40 points Peer-Corrected)

Part 1 (all questions 5 points)

Q.1 From 2012 to 2018, which pricing model (Performance, CPM, Hybrid) has brought in the most revenue?

a. Performance pricing model
b. CPM pricing model
c. Hybrid pricing model
d. All models have performed the same

Explanation: Week 7, Video 3, Slide 13

Q.2 What does CPM stand for?

a. Cost Per Million
b. Cost Per Thousand
c. Counts Per Hundred
d. None of the Above

Explanation: CPM = Cost Per Mille = Cost Per Thousand

Q.3 Which of the following statements is correct with respect to Bounce Rate?

A. Bounce Rate gives an indication of the proportion of visitors who did not interact with the website.

B. Bounce Rate tells us how long, on average, visitors are staying on our website.

C. Bounce Rate increases when someone loads a page and decreases after 30 minutes of inactivity.

D. A high Bounce Rate generally indicates that the website entrance pages are very relevant to the website's visitors.

Answer: A
Explanation: Statement A is correct.
B) Average Session Duration defines how long, on average, visitors are staying on the website.
C) A session starts right away when someone loads a page and ends after 30 minutes of inactivity.
D) A high Bounce Rate generally indicates that the website entrance pages are not relevant to the website's visitors.

Q.4 A company is doing an ad campaign where the details of the ad are presented below.

(Assume that a customer would purchase **twice** in his/her life-time upon watching the advertisement once)

Metric	Value
Avg CPC (Cost per click)	$1.05
Conversion Rate	7%
Avg.Sale Value	$80
Profit Margin	20%

What is the break-even price of average CPC per customer over lifetime?

A. $1.12
B. $2.24
C. $1.19
D. $2.48

Solution: 32*0.07 = 2.24

PPC Conversion cost = $1.05/0.07 = 15
Profit Margin per sale = $80*0.2 = 16
Profit Margin per customer over lifetime (excluding advertisement costs) = 32
Total profit per customer = 32 – 15 = 17

For break even the cost = 32*0.07 = 2.24

Q.5 The objective of Conversion Rate Optimization (CRO) is to:

A. Increase number of website visitors

B. Increase website sales

C. Enhance engagement

D. All the above

Answer: D
Explanation: From the slides, we learnt that Conversion rate optimization (CRO) is the systematic process of increasing the percentage of website visitors who take a desired action. Desired actions are different based on website goals. So above choices are general goals of all websites, and they are CRO goals.

Q.6 A website that uses Google Analytics wants to know the percentage of visitors that do not interact with the website. Which metric should be used?

A. Page per sessions

B. Pageviews

C. Users

D. Bounce Rate

Answer: D
Explanation: The following is how every choice may be defined –
A: Page Per Session-dividing the total number of pageviews by the total number of sessions. It is good indicator of overall user engagement.
B: Pageviews -any view of a page that is being tracked by Google Analytics.
C: Users -Total number of unique visitors to the website
D: Bounce Rate -It is the number of single-page sessions (bounces) divided by the total number of sessions. It shows the proportion of visitors who did not interact with the website.

Questions 7-8 can be answered using case study: **Chase**

Q.7 In the Chase case, Chase segmented customers based on the types of rewards they preferred. Which segmentation strategy does Chase use?

A. Behavioural method

B. Demographic method

C. Psychographic method

Answer: A
Explanation: The following line from the case study may be used to answer the above question–
"*Behavioral/attitudinal segmentation provided insight into how consumers used their cards andhow much they valued rewards and/or what types of rewards they preferred (cashback, miles, points) as well as their channel preferences."*

Q.8 A complete economics of credit card transaction includes:

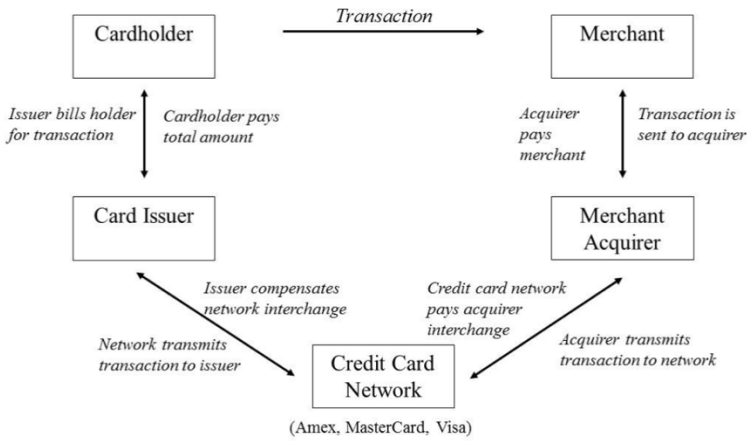
A. Card Issuer; Merchant Acquirer; Merchant

B. Card Issuer; Cardholder; Merchant Issuer; Merchant

C. Card Issuer; Cardholder; Merchant; Merchant Acquirer; Credit Card Network

D. Card Issuer; Cardholder; Merchant; Merchant Issuer; Credit Card Network

Answer: C
Explanation:



Q9

The following questions are based on the **Advertising** dataset (Advertising_Updated.csv). The sales are in thousands of units, while the advertising budgets (TV, Radio, Newspaper) are in thousands of dollars.

Load the data:

```
ad = read.csv('P:\\6203 TA\\Advertising.csv')
```

Run the following linear regression model:

```
lm <- lm(Sales~., data=ad)
```

Now that we have our linear regression model, let's try to make a prediction for the sales given a new set of advertising budgets as follows:

```
new.dat <- data.frame(TV=200, Radio=10, Newspaper=20)
```

You are required to report the predicted sales as well as the lower and upper bound for the 95% prediction interval. What will you report?

A. The predicted sales value is $13,543.06, with a 95% prediction interval of $10,210.25 and $16,875.87.
B. The predicted sales value is $13,956.37, with a 95% prediction interval of $10,613.31 and $17,299.43.
C. The predicted sales value is $15,852.04, with a 95% prediction interval of $12,508.44 and $19,195.64.
D. The predicted sales value is $9,379.90 with a 95% prediction interval of $6,038.61 and $12,721.20.

Answer: B

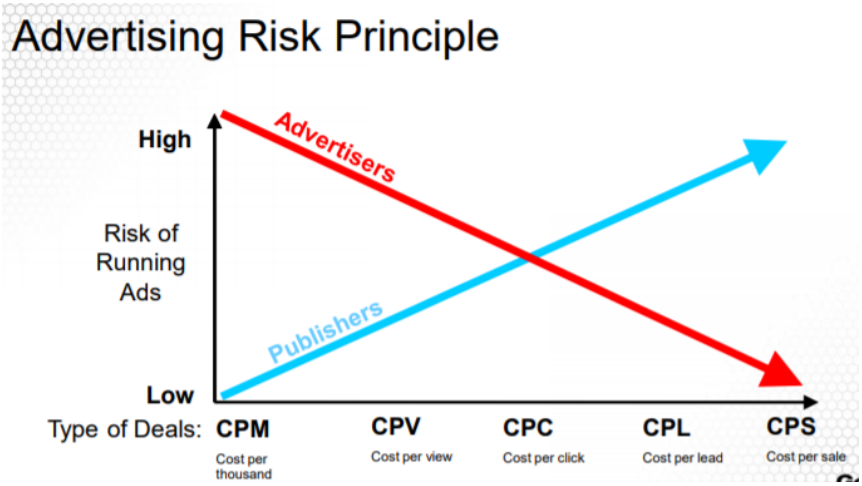
Explanation: Use the predict function in R and change the interval to “prediction” and level to “0.95”.

```
new.dat <- data.frame(TV=200, Radio=10, Newspaper=20)
predict(lm, newdata=new.dat, interval='prediction', level=0.95)
       fit      lwr      upr
1 13.95637 10.61331 17.29943
```

Q.10 If a company is planning on investing an amount X in advertising. Which of the following is the safest investment option? (Assuming they all cost the same per unit)

a. CPM
b. CPV
c. CPC
d. CPS

Answer:



Instructions for Question 11 and 12:
A popular vegan restaurant is known to have long waiting lines from 12-2 pm in the afternoon. Recently, due to an increase in the demand, the amount of time that customers wait in the queue has increased. The manager does not want to lose customers due to this and hence decides to set up another counter to increase the overall service rate. The arrival rate has increased to 58 customers/hour. The current service rate with 4 counters in the restaurant is 60 customers/hour.

Q.11 What is the average amount of time customers will wait in line under the current scenario? (in minutes)

A. 19 minutes
B. 25 minutes
C. 29 minutes
D. 33 minutes

$W_q = \frac{\lambda \times 60}{\mu(\mu-\lambda)} = \frac{58 \times 60}{60(60-58)}$

29 mins

Q.12 On average, how many customers will be waiting in the queue after the manager introduces another counter? Total service rate with 5 counters is 65 customers/hour. (Round to the nearest integer)

A. 5
B. 7
C. 9
D. 11

$$W_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{58^2}{65(65-58)} = 7.39 \approx 7$$

Part 2 (Each question 5 marks)

Instructions for Q.1 to 2

Please use the Facebook Ad dataset KAG_data.csv for the next set of questions. We advise to solve these questions using R (preferably using dplyr library wherever applicable) after reviewing the code provided for Week 11 and other resources provided for learning dplyr in R Learning Guide.

Load the dataset as below:

```
data <- read.csv("KAG_data.csv", stringsAsFactors = FALSE)
```

Q.1
a) (2marks) Which ad (provide ad_id as the answer) among the ads that have the least CPC led to the most impressions?

A. 1121094

Solution:

```{r}
data <- read.csv("downloads/KAG.csv",stringsAsFactors = FALSE)
data_sub <- data %>%
  filter(CPC == min(CPC)) %>%
  filter(Impressions == max(Impressions)) %>%
  select(ad_id)
```

b) (3 marks) What campaign (provide campaign_id as the answer) had spent least efficiently on brand awareness on an average (i.e. most Cost per mille or CPM: use total cost for the campaign / total impressions in thousands)?

A. 936

Solution:

```{r}
data_partb <- data %>%
  group_by(campaign_id) %>%
  summarise(Total_Impressions = sum(Impressions), Total_Spent = sum(Spent))%>%
  mutate(CPM = Total_Spent/Total_Impressions*1000)%>%
  filter(CPM == max(CPM))%>%
  select(campaign_id)

```

Q.2 Assume each conversion ('Total_Conversion') is worth $10, each approved conversion ('Approved_Conversion') is worth $50. ROAS (return on advertising spent) is revenue as a percentage of the advertising spent. Calculate ROAS and round it to two decimals. (Use 'Spent' as the Cost in the given ROAS formula)

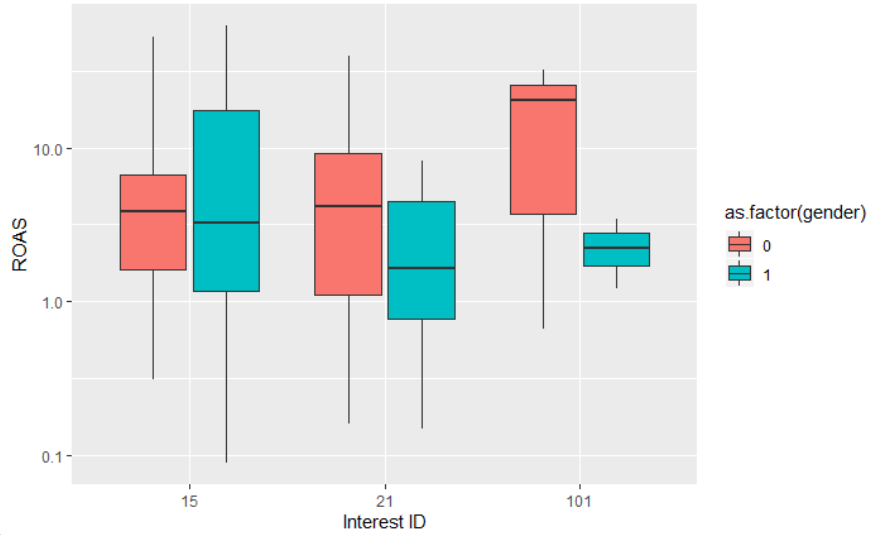
A) 3 marks Make a boxplot of the ROAS grouped by gender for interest_id = 15, 21, 101 in one graph. Try to use the function '+ scale_y_log10()' in ggplot to make the visualization look better. The x-axis label should be 'Interest ID' while the y-axis label should be ROAS.

B) 2 marks Summarize the median and mean of ROAS by genders when campign_id == 1178.

(Note: Make sure to remove the advertisements where there is no advertising spent)

Hint: $ROAS = \frac{Revenue}{Cost} = \frac{10 \times Total\_Conversion + 50 * Approved\_Conversion}{Cost}$

Solution:



S

gender <int>	medianROAS <dbl>	meanROAS <dbl>
0	1.59	3.142053
1	0.92	1.922794

A)

```
```{r}
data_question2 <- data %>%
  filter(Spent != 0)%>%
  mutate(ROAS = round((10*Total_Conversion + 50*Approved_Conversion)/Spent,2))%>%
  filter(interest == 15 |interest == 21 | interest == 101 )
```
```{r}
ggplot(data=data_question2, aes(x=factor(interest), y = ROAS, fill=factor(gender))) +
geom_boxplot() + scale_y_log10() + ggtitle("BoxPlot for the ROAS grouped by gender vs.
Interest Id") + labs(x="Interest Id", y="ROAS")
```

B)

```
data_question2 <- data %>%
  filter(Spent != 0)%>%
  mutate(ROAS = round((10*Total_Conversion + 50*Approved_Conversion)/Spent,2))

data_question2_partb <- data_question2 %>%
  filter(campaign_id == 1178)%>%
  group_by(gender)%>%
  summarise(medianROAS = median(ROAS), meanROAS = mean(ROAS))
```

Instructions for Questions 3-5:

Using the Advertising.csv dataset and the following setup instructions to solve the questions.

Library(pROC)
Library(caret)
Library(dplyr)
Library(ggplot2)
```

```
Data <- read.csv('Advertising.csv', stringsAsFactors=FALSE, headers=TRUE)
Data$Clicked.on.Ad <- as.factor(Data$Clicked.on.Ad)
Head(Data)
```

3) Make a scatter plot for Daily.Internet.Usage against Age. Separate the datapoints by different shapes and/or color based on if the datapoint has clicked on the ad or not.
Based off the general trends in the scatter plot you created, consider a new data point where an individual has a Daily.Internet.Usage less than or equal to 150, and an age of 40. Would this new individual be likely to click the ad or not click the ad?

Solution:

```
area_age_scatter <- ggplot(data=data_2,aes(x=Age,y=Daily.Internet.Usage,shape=Clicked.on.Ad,color=Clicked.on.Ad))
area_age_scatter <- area_age_scatter + geom_point()
area_age_scatter
```

Solution: The individual should classify this new individual as Clicked.on.Ad = 1 or the individual is likely to click the ad

4) Create a logistic regression model using the variables 'Daily.Time.Spent.on.Site', 'Age', and 'Area.Income' to predict the variable 'Clicked.on.Ad'. Display the summary output of this logistic regression model.
Now that we have created our logistic regression model, we must test the model. When testing such models, it is always recommended to split the data into a training (from which we build the model) and test (on which we test the model) set. This is done to avoid bias, as testing the model on the data from which it is originally built from is unrepresentative of how the model will perform on new data.

- That said, for the case of simplicity, test the model on the full original dataset.
    - Use type = ”response” to insure we get the predicted probabilities of clicking the advert
    - Append the predicted probabilities to a new column in the original dataset or simply to a new data frame. The choice is up to you, but ensure you know how to reference this column of probabilities.

Using a threshold of 80% (0.8), create a new column in the original dataset that represents if the model predicts a click or not for that person. Note this means probabilities above 80% should be treated as a click prediction. Now, using the caret package, create a confusion matrix for the model predictions and actual clicks. Print and/or plot this output.

Solution:

```
logistic_reg_model<-
glm(Clicked.on.Ad~Daily.Time.Spent.on.Site+Area.Income+Age,data=data_2,fam
ily=binomial(link                                                    ='logit'))
summary(logistic_reg_model)
##
##                                                                  Call:
## glm(formula = Clicked.on.Ad ~ Daily.Time.Spent.on.Site + Area.Income +
##          Age,  family  =  binomial(link  =  "logit"),  data  =  data_2)
##
##                              Deviance                          Residuals:
##          Min               1Q       Median               3Q            Max
##   -2.59567       -0.30198       -0.06632          0.09470          2.84467
##
##                                                          Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.504e+01   1.443e+00   10.420    <2e-16 ***
## Daily.Time.Spent.on.Site  -2.048e-01   1.565e-02  -13.085    <2e-16 ***
## Area.Income                  -1.173e-04   1.274e-05   -9.206    <2e-16 ***
## Age                             1.630e-01   1.785e-02    9.132    <2e-16 ***
##                                                                  ---
## Signif. codes:   0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
##
##   (Dispersion   parameter   for   binomial   family   taken   to   be   1)
##
##           Null  deviance:  1386.29     on   999     degrees   of   freedom
##   Residual   deviance:      391.98      on   996     degrees   of   freedom
##                                AIC:                                399.98
##
## Number of Fisher Scoring iterations: 7
```

Solution:

```
data_3                    <-                    data_2                    %>%
   select(Daily.Time.Spent.on.Site,Age,Area.Income,Clicked.on.Ad)
pred <- predict(logistic_reg_model,data_3[,1:3], type = 'response')
Data_3                          <-                          cbind(data_3,pred)
data_4                    <-                    data_3                    %>%
   mutate(pred_class                =                as.integer(pred>=.80))
  <-
confusionMatrix(data=as.factor(data_4$pred_class),reference=data_4$Clicked
.on.Ad)
print(confusion)
```

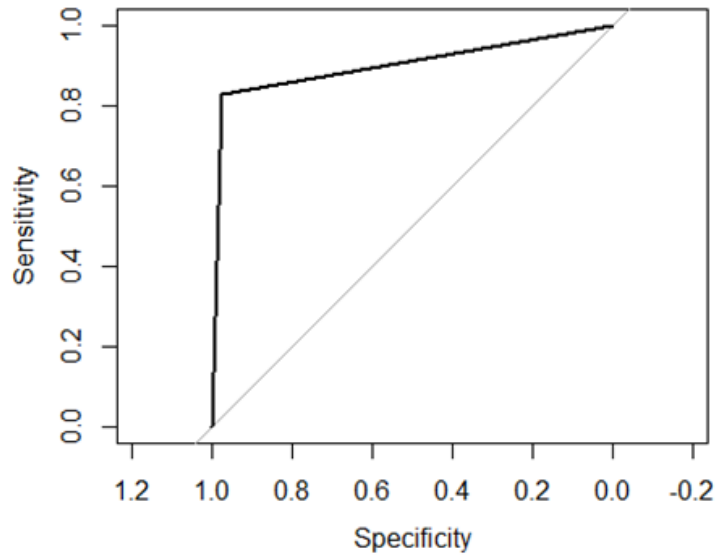
```
##           Confusion           Matrix           and           Statistics
##
##                                                                 Reference
##         Prediction                              0                       1
##                                                 0     488              87
##                                                 1          12          413
##
##                                                      Accuracy    :    0.901
##                                          95%  CI   :  (0.8808,   0.9188)
##                             No      Information      Rate      :      0.5
##                     P-Value     [Acc     >     NIR]     :     <     2.2e-16
##
##                                                        Kappa    :    0.802
##
##               Mcnemar's       Test       P-Value       :       1.028e-13
##
##                                                  Sensitivity    :    0.9760
##                                                  Specificity    :    0.8260
##                                       Pos    Pred    Value    :    0.8487
##                                       Neg    Pred    Value    :    0.9718
##                                                  Prevalence    :    0.5000
##                                     Detection    Rate    :    0.4880
##                          Detection       Prevalence       :       0.5750
##                                   Balanced     Accuracy     :     0.9010
##
##                                       'Positive'     Class     :     0
##
```

5) Given the output above, how many false negative occurrences do you observe? Recall false negative means the instances where the model predicts the case to be false when in reality it is true. For this example, this refers to cases where the ad is clicked but the model predicts that it isn't. Using the Proc() library, use the roc() function to create and plot a ROC curve of our predictions and true labels.

Solution: False negatives would be when Reference = 1 and Prediction = 0 which is 87 false negatives

```
roc<-roc(data_4$Clicked.on.Ad,data_4$pred_class)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot(roc)
```



Instructions for Questions 6-8: In response to the ongoing pandemic, a local restaurant has implemented social distancing measures, which include the closing its in person dining areas. In order to keep the business going, the restaurant will now rely on drive through lines to handle customer ordering and service. After implementing the new ordering system, management observes that customers arrive at the rate of 62 customers per hour. Under the current system the restaurant has only 5 servers with a total service rate of 70 customers/hour.

6)

A)2marks Given the description above, what is the average amount of time customers will wait in line under the current restaurant scenario? (round to closest integer)

Solution: The answer is 7 minutes; This is derived using the formula for average waiting time and is solved below.

Average wait= (lambda * 60 minutes)/(mu(mu-lambda))
= (62*60)/(70(70-62))
=6.64 minutes ~ 7

Alternatively, in R:

```
rm(list=ls())
arrival_rate <- 62
service_rate <- 70
average_wait <- (arrival_rate*60)/(service_rate *(service_rate-
arrival_rate))
round(average_wait,0)
```

```
## [1] 7
```

B)3marks Now consider that the restaurant manager from above is wanting to reduce customer waiting times. In order to do this, the manager decides to add another server bringing the total number of servers to 6 and the total service rate to 84 customer/hour. On average, how many customers will be waiting in the queue after the manager introduces this extra server? (round to closest integer) (average arrival remains 62/hour)

Solution: The answer is 2; This is derived using the formula for average queue length, and solved below:

Average queue length = (arrival rate^2)/ (service rate (service rate – arrival rate)
Average queue length = (62^2) / (84 (84 – 62))
Average queue length = 2.08 ~ 2

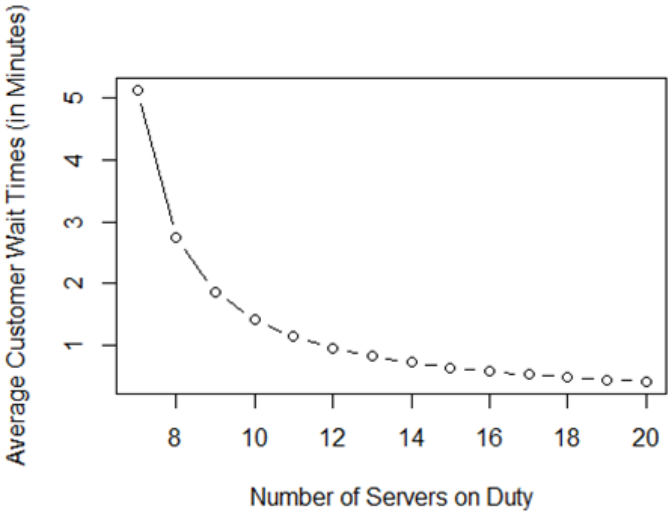
Alternatively, in R:

```
rm(list=ls())
arrival_rate <- 62
service_rate <- 84
average_queue_length <- (arrival_rate^2)/(service_rate *(service_rate-
arrival_rate))
round(average_queue_length,0)
## [1] 2
```

Q7) Now consider that every sever on duty for the restaurant adds a constant 14 customer/hour to the total service rate of the restaurant. Given that the rate of arrivals for the restaurant remains constant at 82 customers/hour, plot (using R) the average amount of time customers will wait in line with various numbers of servers. The minimum servers that can be on duty is 7 and the maximum is 20. Display the plot as output.

Solution:

```
rm(list=ls())
library(dplyr)
arrival_rate <- 82
service_rate_per_server <- 14
number_of_servers <- seq(7,20)
service_rates <- number_of_servers*service_rate_per_server
df <- as.data.frame(number_of_servers)
df <- cbind(df,service_rates)
df<-  df %>%
   mutate(average_waits= (arrival_rate*60)/(service_rates *(service_rates-
arrival_rate)))
wait_plot <- plot(x=df$number_of_servers,y=df$average_waits,type = "b",
xlab = 'Number of Servers on Duty', ylab = 'Average Customer Wait Times
(in Minutes)')
```



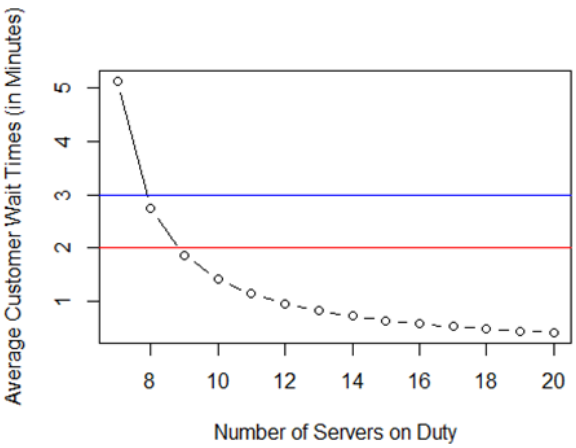
8)

A)2marks Based on your plot above, at what number of servers does the average customer wait time drop below 3 minutes?

B)3 marks Based on your plot above, Describe the behavior of the chart and give some commentary about the relationship between the two variables. For example, is the chart increase or decreasing? What value does average wait time approach if we continue to add more and more servers (to infinity)?

Solution:

```
plot(x=df$number_of_servers,y=df$average_waits,type = "b", xlab = 'Number
of Servers on Duty', ylab = 'Average Customer Wait Times (in Minutes)')
abline(h=3, col='blue')
abline(h=2,col='red')
```



It is apparent from the chart above (or numerically) that average wait times reach below 3 minute levels when 8 severs are on duty)

Solution: Generic commentary on the graph, must talk about the limits of the graph as severs on duty approach infinity, must reference decreasing over the interval