Assn1

Shivam

2023-01-31

1.You are working Streaming Service. The following statement is made by your manager. Based on the examples and discussion in Lecture 1, transform it in to a question that can be answered with data analytics. Make sure you discuss the logic and reasoning you use to transform it and what questions you might ask.

"We have more customers than before, but our new customers are streaming less than before."

-1. The first question that came in my mind after reading the above statement was what percentage of customers increased as compared to previous timeline?

```
By answering this question we can quantify what chunk of customers were affecting the streaming time.
```

-2. How many hours of streaming on average do our new customers do compared to our old customers?

```
Once we get to know the numbers in detail, we can find the reason behind to improve the results and may
```

-3. How do the streaming preferences of new and returning customers differ across various demographic categories (such as age, gender, and location)?

```
If we analyse how the streaming history of this new or returning customers is scattered over the geograp

Similarly if the company can figure out which age group or gender  is helping to increase the average s
```

-4. What factors may be contributing to the decrease in usage among new customers?

This question helps to identify any external or internal factors that may be contributing to the decrease in usage. This may help later to solve the issue and increase the streaming hours.

```
setwd("C:/Users/holys/OneDrive/Desktop/Data Analytics,Mathamatics,Algor/Assignment01Explonatory")
```

2.Consider the following three arrays of data. Each array is data for one file sharing site. The numbers in the array represent the number of downloads for each site in a day (for example, Site A had 28 downloads on the first day, 29 on the second and so on).

```
Site A: (28 29 31 28 30 30 30 32 28 33)
Site B: (23 19 23 33 32 27 20 24 42 32)
```

Based on the data provided, and using the skills learned in this class, answer the following questions. Make sure to provide evidence for your answers.

a) Which customer streams the least on a typical day?

```
A_SJ <- c(28,29,31,28,30,30,30,32,28,33)
B_SJ <- c(23,19,23,33,32,27,20,24,42,32)
C_SJ <- c(27,26,28,25,27,27,30,30, 28,26)

#On a typical day, average number of customers who stream Site A:
  mean(A_SJ)                                                                    # the mean for site A
```

```
## [1] 29.9
```

```
#On a typical day, average number of customers who stream Site B:
  mean(B_SJ)                                                                    # the mean for site B
```

```
## [1] 27.5
```

```
#On a typical day, average number of customers who stream Site C:
  mean(C_SJ)                                                                    # the Mean for site C
```

```
## [1] 27.4
```

```
#Site C has the lowest average of viewers in ten days, hence site c has the least number of customers o
```

b) Which customer is the most inconsistent in the usage of the streaming service?

To measure the most inconsistent Site, we have to measure the standard deviation is the square root of the variance:

```
  sd(A_SJ)
```

```
## [1] 1.72884
```

```
sd(B_SJ)
```

```
## [1] 7.168604
```

```
sd(C_SJ)
```

```
## [1] 1.646545
```

```
#Site B has the highest standard deviation value, which implies that it is the most inconsistent usage
```

PART 2 ##########
Question 1

1. Read in the text file and change to a data frame.

```
setwd("C:/Users/holys/OneDrive/Desktop/Data Analytics,Mathamatics,Algor")

test_SJ <- read.table("PROG8430-23W-Assign01 (1).txt", sep=",", header=TRUE)

    test_SJ <- as.data.frame(test_SJ)
    head(test_SJ)
```

```
##        Manufacturer Server        DC   SMBR  SMBT Conn
## 1              Lled MG9696  Waterloo 102479 43473 6625
## 2            Ovonel RX8838  Waterloo 103678 62534 7580
## 3              Lled MB3406 Cambridge 102003 35916 5957
## 4              Lled MB3406 Kitchener  98889 40245 6120
## 5 Highway-Passenger DF6726 Cambridge 104907 25422 5839
## 6 Highway-Passenger DF6726 Kitchener 102659 53168 7076
```

2. Append your initials to all variables in the data frame (Note – you will need to do this in all your subsequent assignments).

```
 colnames(test_SJ) <- paste(colnames(test_SJ), "SJ", sep = "_")
  head(test_SJ)
```

```
##     Manufacturer_SJ Server_SJ     DC_SJ SMBR_SJ SMBT_SJ Conn_SJ
## 1              Lled    MG9696  Waterloo  102479   43473    6625
## 2            Ovonel    RX8838  Waterloo  103678   62534    7580
## 3              Lled    MB3406 Cambridge  102003   35916    5957
## 4              Lled    MB3406 Kitchener   98889   40245    6120
## 5 Highway-Passenger    DF6726 Cambridge  104907   25422    5839
## 6 Highway-Passenger    DF6726 Kitchener  102659   53168    7076
```

3. Change each character variable to a factor variable.

```
test_SJ$Manufacturer_SJ <- factor(test_SJ$Manufacturer_SJ)
test_SJ$Server_SJ       <- factor(test_SJ$Server_SJ)
test_SJ$DC_SJ           <- factor(test_SJ$DC_SJ)
```

4. What are the dimensions of the dataset (rows and columns)?

```
 dim(test_SJ)
```

```
## [1] 90000     6
```

```
# dim() function is used to return the dimensions of the table/ datasets / number of rows and columns
```

Question 2 Summarizing Data

1. Means and Standard Deviations:-

a. Calculate the mean and standard deviation for Server Message Blocks Received.

```
    mean(test_SJ$SMBR_SJ)
```

```
## [1] 100017.5
```

```
    sd(test_SJ$SMBR_SJ)
```

```
## [1] 10002.46
```

```
# b. Use the results above to calculate the coefficient of variation
  (rounded to 3 decimal places).
```

```
      CV <- (sd(test_SJ$SMBR_SJ) / mean(test_SJ$SMBR_SJ))
      round(CV,3)
```

```
## [1] 0.1
```

```

```

c. Calculate the mean and standard deviation for Server Message Blocks Transmitted.

```
      mean(test_SJ$SMBT_SJ)
```

```
## [1] 49966
```

```
      sd(test_SJ$SMBT_SJ)
```

```
## [1] 10024.44
```

```
# d. Also calculate the coefficient of variation (rounded to 3
#   decimal places).
```

```
    CV_of_SMBT <- (sd(test_SJ$SMBT_SJ) / mean(test_SJ$SMBT_SJ))
    CV_of_SMBT
```

```
## [1] 0.2006251
```

```
# e. Does the SMBT or SMBR have more variation?
```

```
    var_of_SMBR <- var(test_SJ$SMBR_SJ)

    var_of_SMBT <- var(test_SJ$SMBT_SJ)

    var_of_SMBR
```

```
## [1] 100049172
```

```
    var_of_SMBT
```

```
## [1] 100489304
```

```
# sd_of_SMBT > sd_of_SMBR

# SMBT have more variation.
```

2. Calculate the 45th percentile of the number of Server Message Blocks Transmitted. This calculation should be rounded to the nearest whole number (no decimal places).

To fing the 45 th percentile of number of SMBT we will use quantile() function. We will also use round() function for rounding up the value to the nearest value.

```
    quantile(test_SJ$SMBR_SJ,probs = 0.45)
```

```
##   45%
## 98732
```

```
    round(quantile(test_SJ$SMBR_SJ,probs = 0.45))
```

```
##   45%
## 98732
```

Question 3

#Organizing Data

1. Summary Table

   a. Create a table showing the average Server Message Blocks Transmitted by Manufacturer. This should be rounded to two decimal places.

```
    Average_SMBT_SJ <- aggregate(test_SJ$SMBT_SJ,
                                 by=list(test_SJ$Manufacturer_SJ),
                                 FUN=function(m){round(mean(m),digits=2)})

    colnames(Average_SMBT_SJ) <- c("Manufacturer","Average")

    Average_SMBT_SJ
```

```
##        Manufacturer  Average
## 1 Highway-Passenger 49916.14
## 2              Lled 50008.12
## 3            Ovonel 49973.76
```

   b. Which Manufacturer's Servers have, on average, transmitted the most server message blocks? Which manufacturer is it?

    Lled transmitted the the most server message blocks with average of 50008.12.

2. Cross Tabulation

a. Create a table counting all Servers by Data Centre.

```
Servers_And_DCs_SJ <- table(test_SJ$Server_SJ,test_SJ$DC_SJ)
Servers_And_DCs_SJ
```

```
##
##          Bridgeport Cambridge Elmira Kitchener Waterloo
##   DF6726       2971      4385   5869      7363     8827
##   DJ3756         60        87    118       157      163
##   MB3406       2188      3433   4534      5634     6882
##   MG9696        719      1128   1435      1810     2237
##   RQ8547       2082      3184   4161      5248     6421
##   RX8838        925      1365   1734      2191     2689
```

b. Change the table to show the percentage of each Server in each Data Centre . This should be rounded to three decimal places.

```
Servers_in_percentage_SJ <- prop.table(Servers_And_DCs_SJ)
 round(Servers_in_percentage_SJ,3)
```

```
##
##          Bridgeport Cambridge Elmira Kitchener Waterloo
##   DF6726      0.033     0.049  0.065     0.082    0.098
##   DJ3756      0.001     0.001  0.001     0.002    0.002
##   MB3406      0.024     0.038  0.050     0.063    0.076
##   MG9696      0.008     0.013  0.016     0.020    0.025
##   RQ8547      0.023     0.035  0.046     0.058    0.071
##   RX8838      0.010     0.015  0.019     0.024    0.030
```

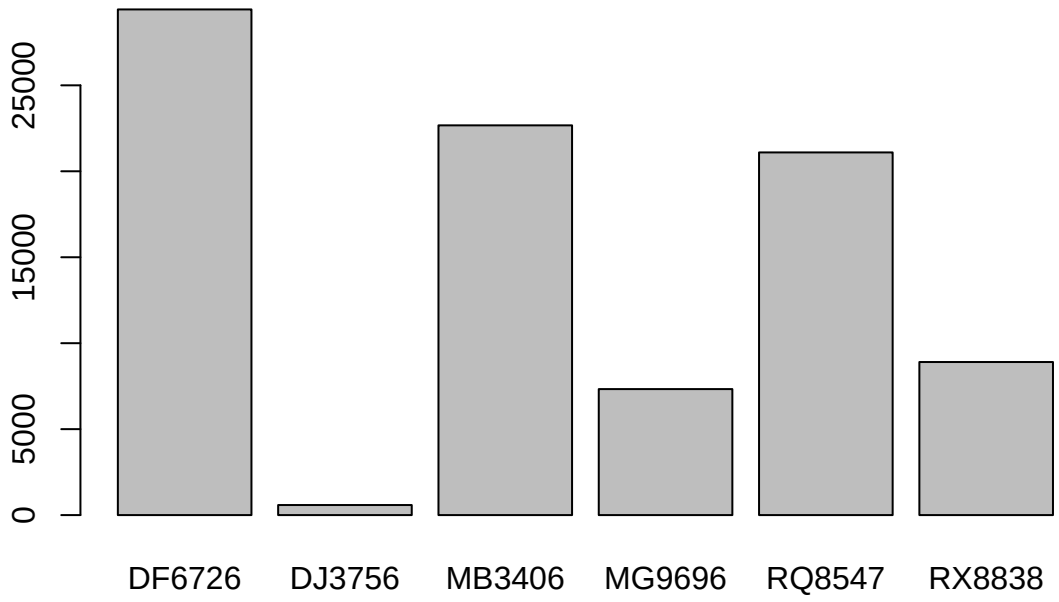
c. What percentage of servers at Elmira are MG9696?

MG9696 was about 1.6% at Elmira.

3. Bar Plot

a. Create a bar plot of count of Servers Models.

```
Count_Servers_SJ <- table(test_SJ$Server_SJ)
barplot(Count_Servers_SJ)
```



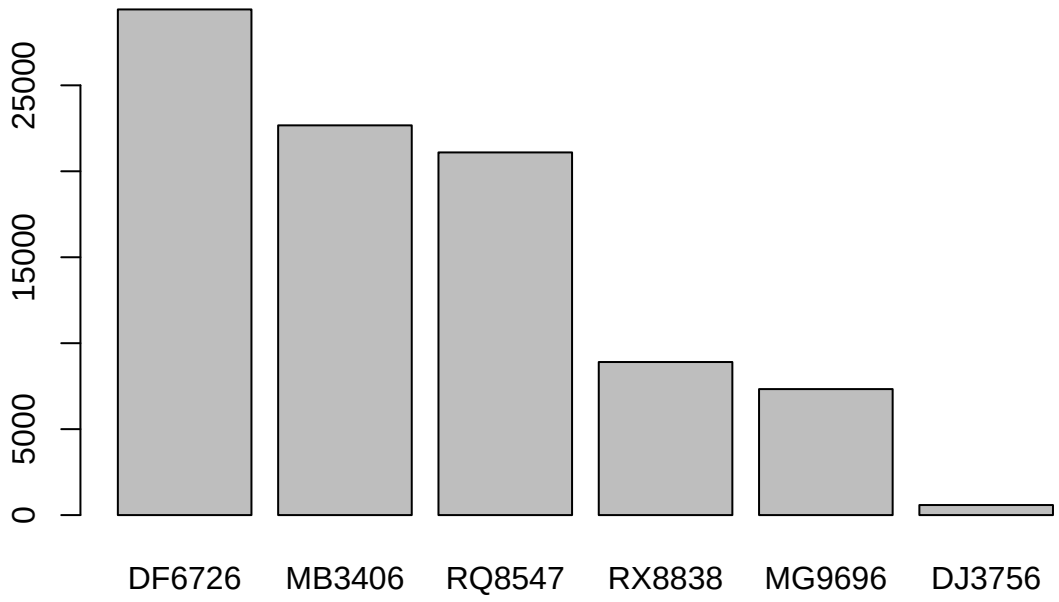

```
test_SJ <- as.data.frame(test_SJ)
    head(test_SJ)
```

```
##     Manufacturer_SJ Server_SJ     DC_SJ SMBR_SJ SMBT_SJ Conn_SJ
## 1              Lled    MG9696  Waterloo  102479   43473    6625
## 2            Ovonel    RX8838  Waterloo  103678   62534    7580
## 3              Lled    MB3406 Cambridge  102003   35916    5957
## 4              Lled    MB3406 Kitchener   98889   40245    6120
## 5 Highway-Passenger    DF6726 Cambridge  104907   25422    5839
## 6 Highway-Passenger    DF6726 Kitchener  102659   53168    7076
```

b. The plot should be:

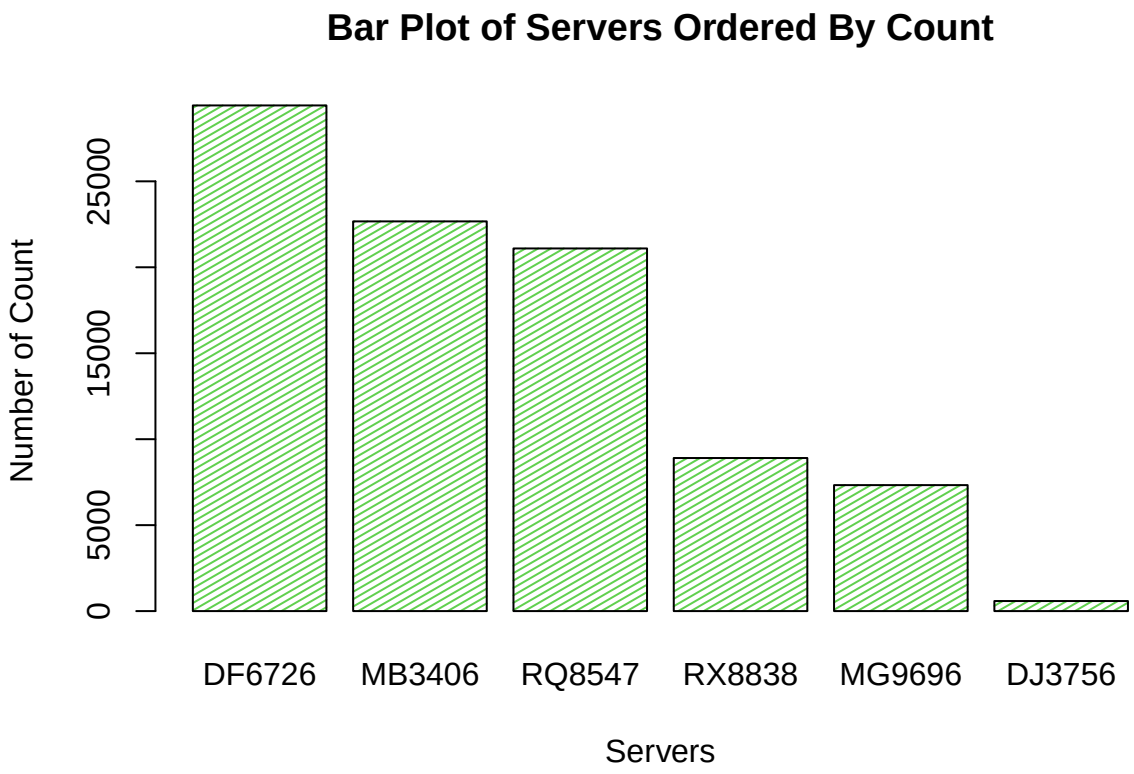
   i. Rank ordered by highest count of Server Model.

```
Servers_Ordered_by_count_SJ <- Count_Servers_SJ[order(Count_Servers_SJ,decreasing=TRUE)]
barplot(Servers_Ordered_by_count_SJ)
```



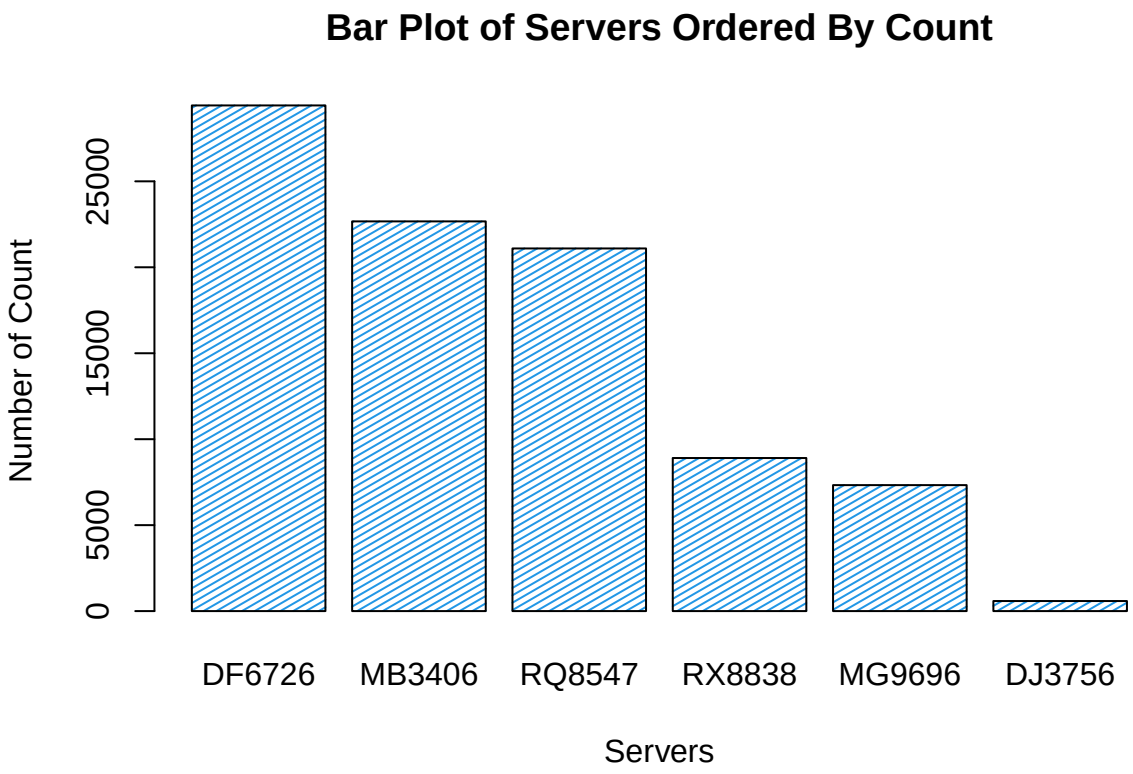
ii. Properly labeled (title, x-axis, etc)

```
barplot(Servers_Ordered_by_count_SJ,
        col=3,
        density = 30, angle = 30,
        main="Bar Plot of Servers Ordered By Count",
        xlab="Servers",
        ylab = "Number of Count")
```



iii. The bars should have a different colour than the one
shown in class.

```
barplot(Servers_Ordered_by_count_SJ,
        col=108,
        density = 30, angle = 30,
        main="Bar Plot of Servers Ordered By Count",
        xlab="Servers",
        ylab = "Number of Count")
```



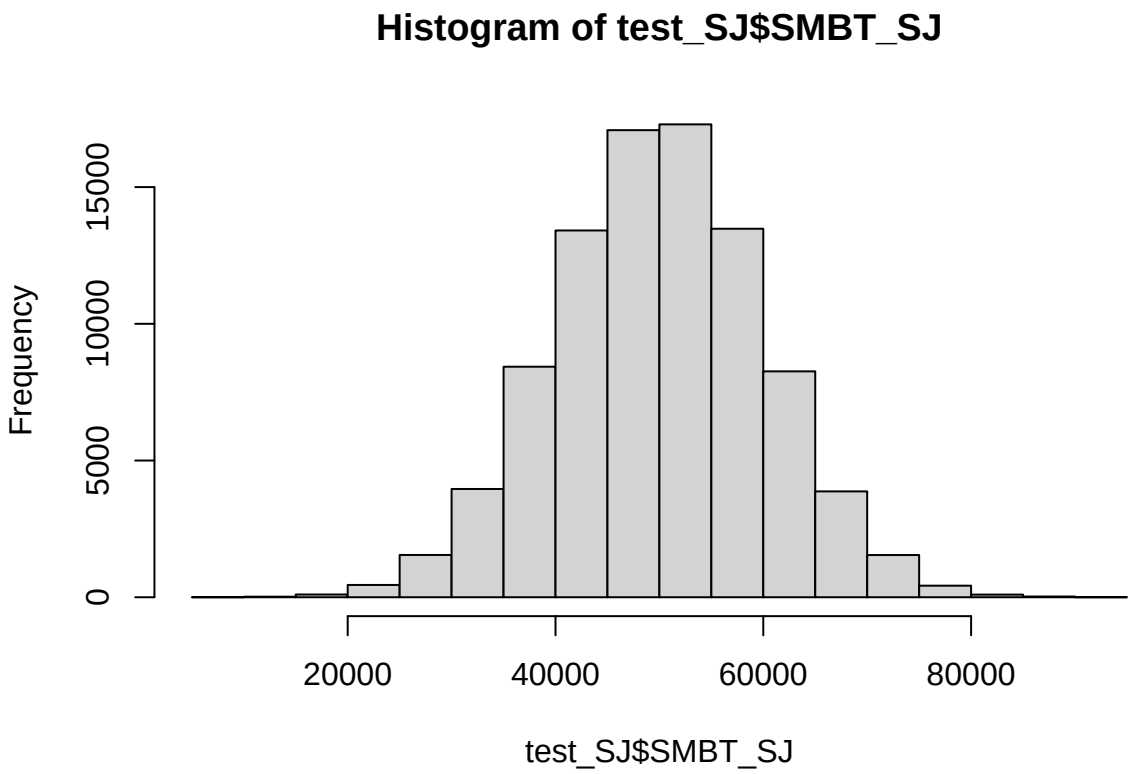
c. Based on the bar plot, (approximately) how many of Server RX8838 are there?

In the above graph plotted we can say that approximately 8,500 Server RX8838 are there.

4. Histogram

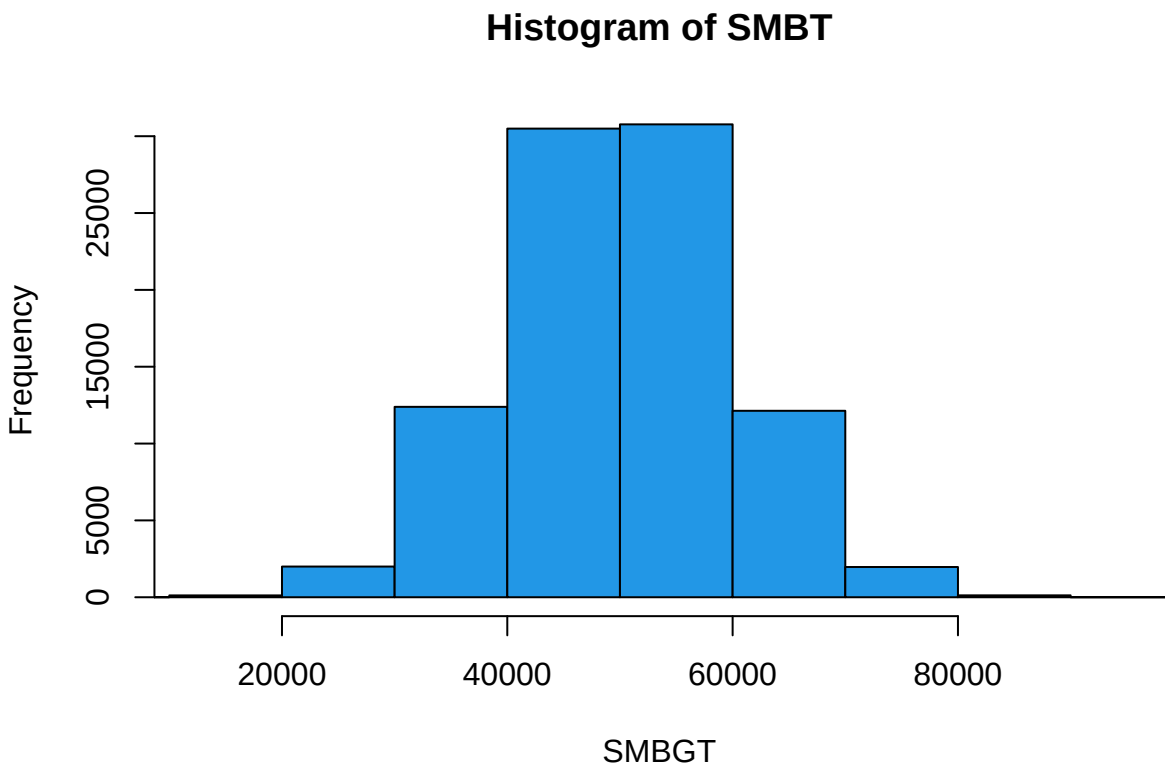
a. Create a histogram of Server Message Blocks Transmitted.

```
hist(test_SJ$SMBT_SJ)
```



b. The plot should be properly labeled and a unique colour and have 10 breaks.

```
hist(test_SJ$SMBT_SJ,
     col = 100,
     breaks = 10,
     xlab="SMBGT",
     xlim = c(12000,95000),
     main="Histogram of SMBT")
```

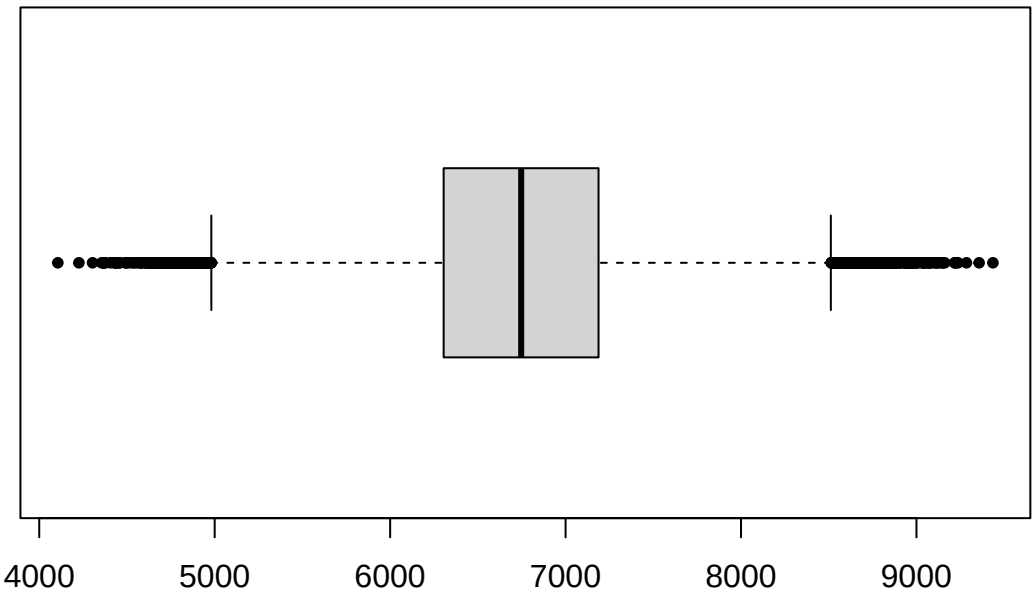


c. Which range of SMBT is the most common? -By the above histogram we can depict the most common range is 40,000-60,000.

5. Box plot

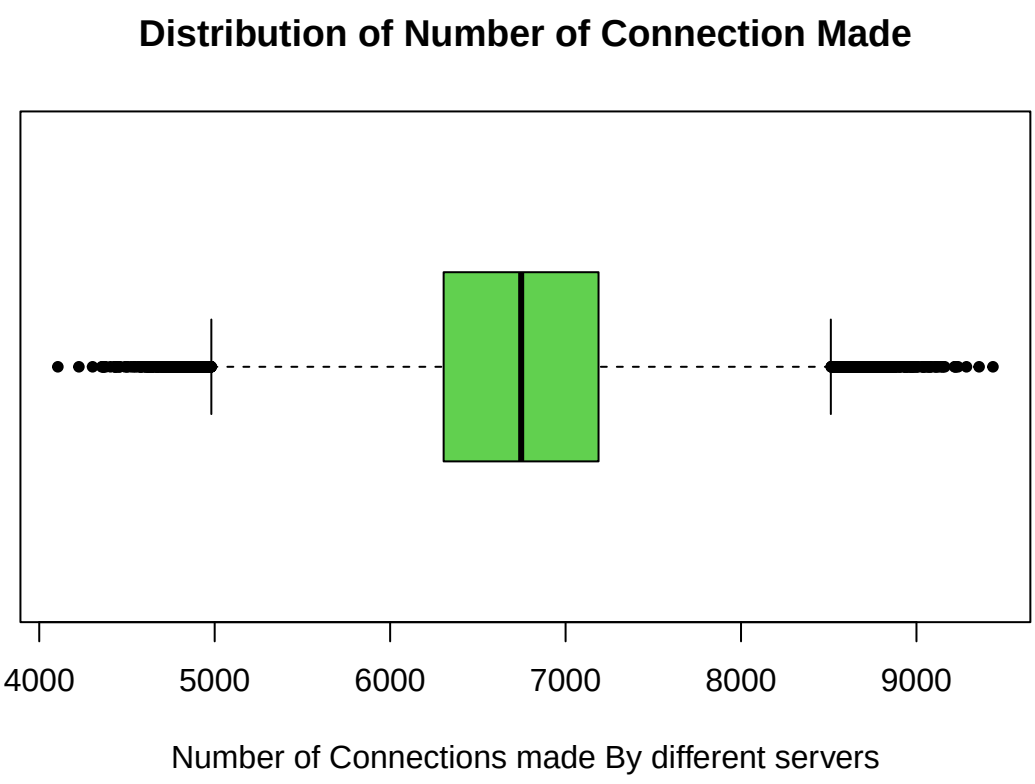
a. Create a horizontal box plot of number of Connections Made.

```
boxplot(test_SJ$Conn_SJ,
        horizontal=TRUE,
        pch=20)
```



b. The plot should be properly labeled and a unique colour.

```
boxplot(test_SJ$Conn_SJ,
        main="Distribution of Number of Connection Made",
        xlab="Number of Connections made By different servers",
        col=67,
        horizontal=TRUE,
        pch=20)
```



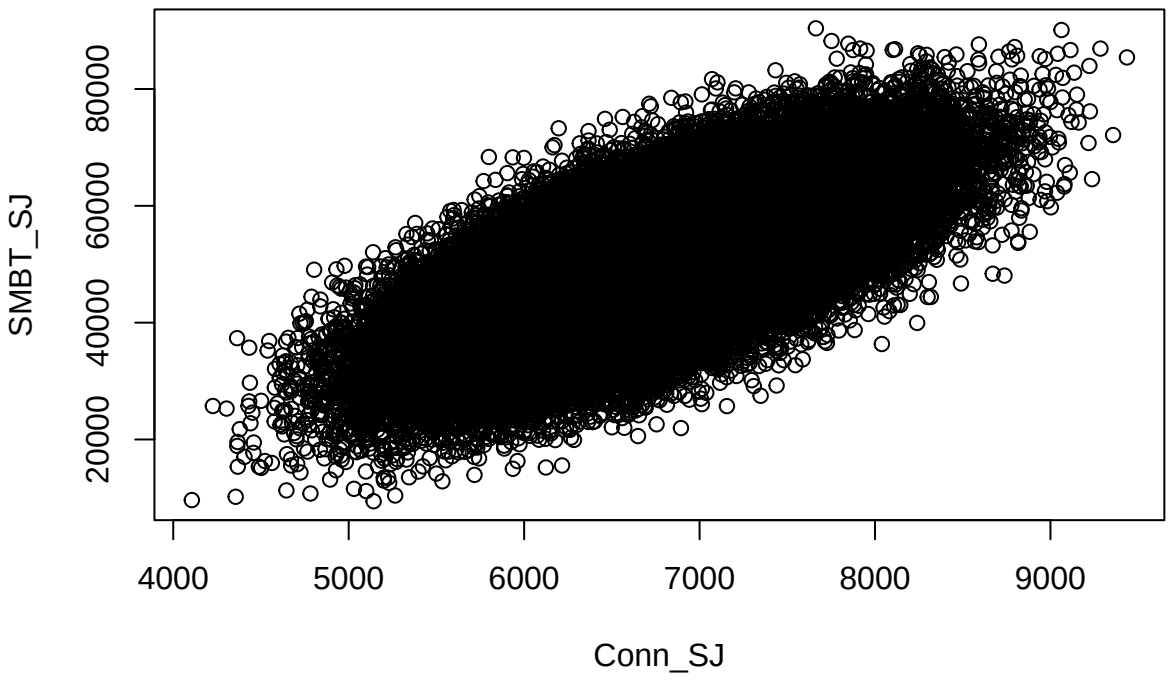
c. Based on the box plot, approximately how many servers made fewer than 6160 connections?

```
According to the box plot nearly  25% of the total servers that is; 22,500 servers made less connection
```

6. Scatter Plot

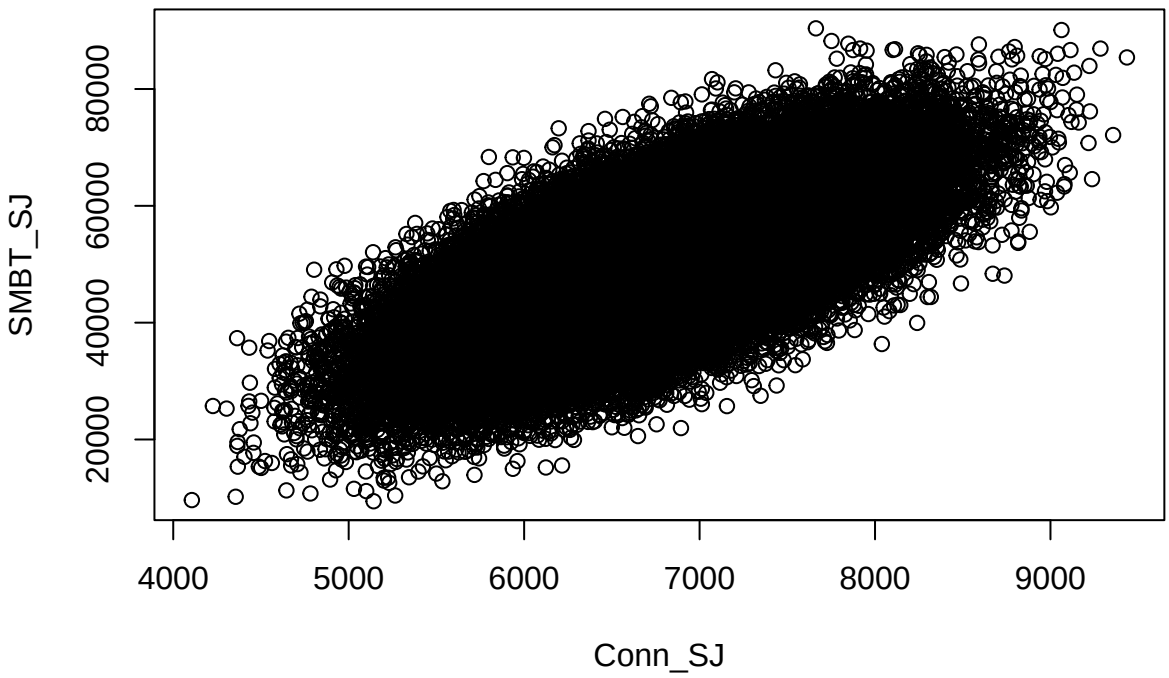
a. Create a scatter plot comparing Server Message Blocks Transmitted and Connections Made.

```
plot(SMBT_SJ ~ Conn_SJ,
     data=test_SJ,
     main="  ")
```



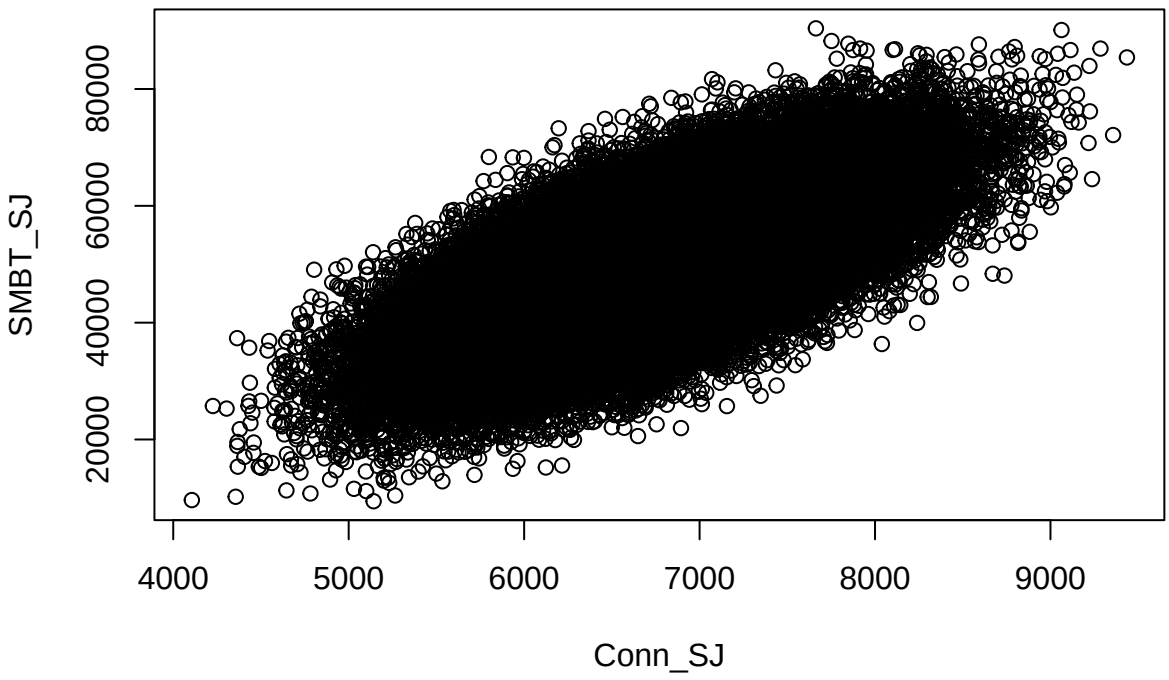

```
#You can change some formating

plot(SMBT_SJ ~ Conn_SJ,
     data=test_SJ)
abline(coef = c(6,0)) #overlays a line, intercept=60, slope=0
```



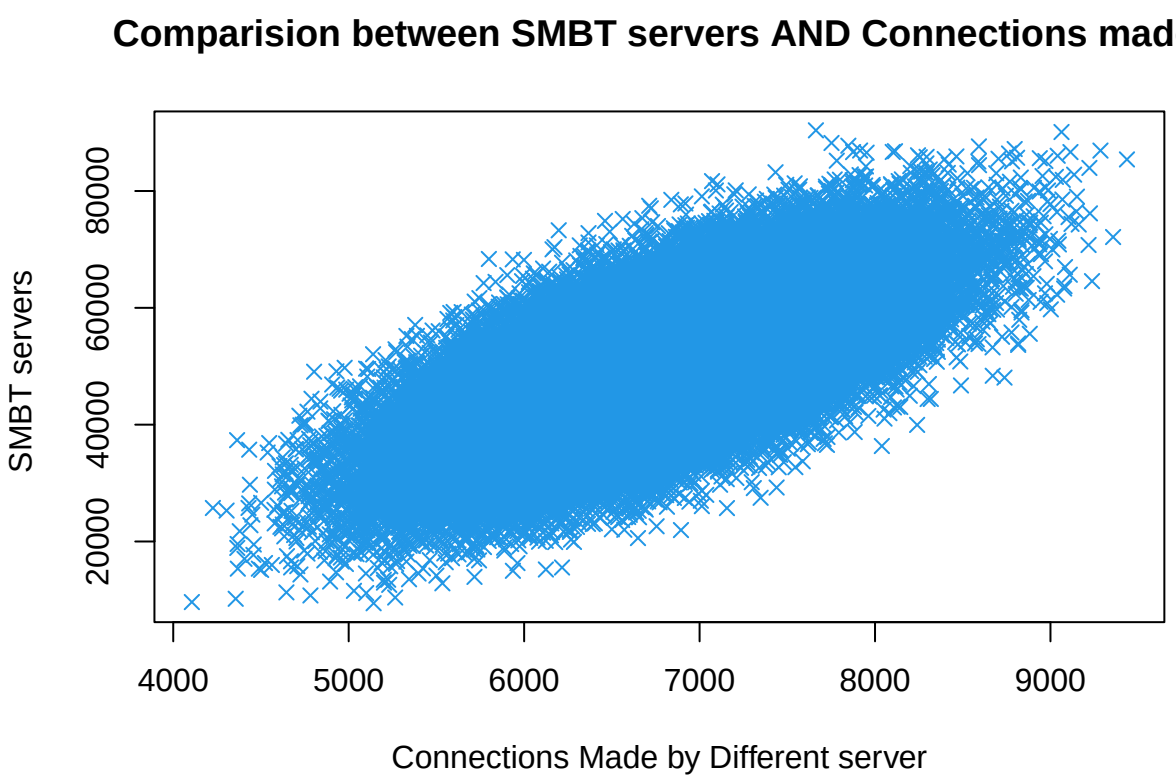
b. The plot should be properly labeled with a marker type different than the one demonstrated in class.

```
plot(SMBT_SJ ~ Conn_SJ,
     data=test_SJ,
     main="  ")
```




```
#You can change some formating

plot(SMBT_SJ ~ Conn_SJ,
     data=test_SJ,
     col=100,
     pch=4,
     main="Comparision between SMBT servers AND Connections made ",
     xlab="Connections Made by Different server",
     ylab="SMBT servers ")
abline(coef = c(6,0)) #overlays a line, intercept=60, slope=0
```



c. Does there appear to be an association between Server Message Blocks Transmitted and Connections Made?

```
By watching the Scatter plot we can say that the data points are dense enough so one
can conclude that their is strong correlation between SMBT servers and Connections made.
```