Assignment_1

Ajay Kanubhai Patel

2022-09-30

#Loading the package

```
#Load packages to convert file in PDF.

if(!require(tinytex)){install.packages("tinytex")}
```

```
## Loading required package: tinytex
```

#This sets the working directory

This section is for the basic set up. It will clear all the plots, the console and the workspace. It also sets the overall format for numbers.

```
if(!is.null(dev.list())) dev.off()
```

```
## null device
##           1
```

```
cat("\014")
```

```
rm(list=ls())
options(scipen=9)
```

#To read Excel file in R data frame.

```
if(!require(readxl)){install.packages("readxl")}
```

```
## Loading required package: readxl
```

```
library("readxl")
```

Q1. 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.

People are downloading our app more and more, but our current users are using the app less and less frequently.

Answer:

- Are users increasing who use our application or just number of downloads increasing? Number of downloads is increasing does not mean that number of users are growing.
- How many times our application is downloaded in particular time frame? which gives insight when our app is downloaded the more/less.
- How many hours (or any different time measuring units such as minutes, seconds, etc. . .) are being spent on our application? – It gives idea about customer actively using app or just they have downloaded on their device.
- Are there number of new registrations increasing?

This gives insight whether actual number of users are increasing because it can be possible that someone downloads application on multiple devices but using same registration credentials/log in credentials.

- Which platform has more downloads, and do they use same log in credentials? Users are downloading on Android, IOS, or desktop version. However,they sign- in with same credentials.

Q2. Consider the following three arrays of data. Each array is data for one customer of a streaming service. The numbers in the array represent the number of videos the customer streamed in a day (for example, customer A streamed 21 videos on the first day, 20 on the second and so on).

Customer A: (21 20 19 18 21 20 18 22 20 18) Customer B: (22 19 18 21 27 21 22 19 21 24) Customer C: (9 10 8 11 8 7 10 11 7 10)

Based on the data provided, answer the following questions. Make sure to provide evidence for your answers.

```
a) Which customer streams the least on a typical day?
b) Which customer is the most inconsistent in the usage of the streaming
   service?
```

Answer:

Here, number of videos streamed by three different customers in ten days are given. If Mean(average) is calculated for given customers, then,

1. Average number of videos streamed by customer A = 21 + 20 +19 +18 +21 +20 + 18 +22 +20 +18 / 10 = 19.7
2. Average number of videos streamed by customer B = 22 + 19 + 18 + 21 + 27 + 21 + 22 + 19 + 21 + 24 / 10 = 21.4
3. Average number of videos streamed by customer C = 9 + 10 + 8 + 11 + 8 + 7 + 10 + 11 + 7 + 10 / 10 = 9.1

```
Customer C streams the least on a typical day and it is on an average 9.1
videos per day.
```

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

```
If we calculate standard deviation for each customer,

Standard deviation for customer A = 1.34

Standard deviation for customer B = 2.49

Standard deviation for customer C = 1.44

Max(A) = 22
Min(A) = 18
Range(A) = Max(A)- Min(A) = 4

Max(B) = 27
Min(B) = 18
Range(B) = Max(B)- Min(B) = 9

Max(C) = 11
Min(C) = 7
Range(C) = Max(C)- Min(C) = 4
```

Answer: Customer B is the most inconsistent in the usage of the streaming service.

PART2:

Q1. Basic Manipulation

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

Verify working directory # to read our excel file located at ('D:/Final Assignment/DATA') namely "2014 and 2015 CSM 22F.xlsx". # to convert our file into data structure in which each component form column and content of the component form the row.

```
getwd()
```

```
## [1] "D:/Final Assignment/DATA"
```

```
Assignment01_AP <- read_excel("2014 and 2015 CSM 22F.xlsx")


Assignment01_AP  <- as.data.frame(Assignment01_AP )
```

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

to change all column name by appending my initials (Ajay Patel = AP) and separate it by "_". # head(data,n) displays 1st n rows present in our excel file. head(Assignment01_BDSA,08)= shows first 8 rows. If number is not provided by default is shows 1st 6 rows

```
colnames(Assignment01_AP) <- paste(colnames(Assignment01_AP), "AP", sep = "_")


head(Assignment01_AP)
```

```
##                            Movie_AP Year_AP Ratings_AP Genre_AP  Gross_AP
## 1                           13 Sins    2014        6.3        8      9130
```

```
## 2                   22 Jump Street    2014        7.1        1 192000000
## 3                   3 Days to Kill    2014        6.2        1  30700000
## 4            300: Rise of an Empire    2014        6.3        1 106000000
## 5                 A Haunted House 2    2014        4.7        8  17300000
## 6 A Million Ways to Die in the West    2014        6.1        8  42600000
##   Budget_AP Screens_AP Sequel_AP Sentiment_AP Views_AP Likes_AP Dislikes_AP
## 1   4000000         45         1            0  3280543     4632          425
## 2  50000000       3306         2            2   583289     3465           61
## 3  28000000       2872         1            0   304861      328           34
## 4 110000000       3470         2            0   452917     2429          132
## 5   3500000       2310         2            0  3145573    12163          610
## 6  40000000       3158         1            0  3013011     9595          419
##   Comments_AP Aggregate Followers_AP
## 1         636                1120000
## 2         186               12350000
## 3          47                 483000
## 4         590                 568000
## 5        1082                1923800
## 6        1020                8153000
```

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

to return dimension of data frame as [no, of rows] [no. of column].

```
dim(Assignment01_AP)
```

```
## [1] 187  14
```

```
print("The dimensions of the dataset: 187  14")
```

```
## [1] "The dimensions of the dataset: 187  14"
```

Q2. Summarizing Data 1. Means and Standard Deviations a. Calculate the mean and standard deviation for Gross.

```
# to calculate mean of Gross_AP.
Gross_mean_AP <- mean(Assignment01_AP$Gross_AP)
# to print the mean of Gross_AP as we have assigned it as variable of mean(Assignment01_AP$Gross_AP).
Gross_mean_AP
```

```
## [1] 77646944
```

```
# to calculate standard deviation of Gross_AP.
Gross_sd_AP<-sd(Assignment01_AP$Gross_AP)
# to print value of the standard deviation of Gross_AP as we have assigned it as variable of sd(Assignm
Gross_sd_AP
```

```
## [1] 93899208
```

b. Use the results above to calculate the coefficient of variation (rounded to 2 decimal places).

```
# to calculate coefficient of variance for Gross_AP.
#(dividing standard deviation by mean)
Gross_CV_AP = Gross_sd_AP / Gross_mean_AP
# to print value of coefficient of variance for Gross_AP.
Gross_CV_AP
```

```
## [1] 1.20931
```

```r
# to print value of Gross_CV_AP up to two decimal points.
round(Gross_CV_AP,2)
```

```
## [1] 1.21
```

c. Calculate the mean and standard deviation for Budget. Also calculate the coefficient of variation (rounded to 2 decimal places).

```r
#to calculate mean of Budget_AP.
Budget_mean_AP <- mean(Assignment01_AP$Budget_AP)

#it print mean of Budget_AP.
Budget_mean_AP
```

```
## [1] 53844373
```

```r
# to calculate standard deviation of Budget_AP.
Budget_sd_AP <- sd(Assignment01_AP$Budget_AP)

#to print standard deviation of Budget_AP.
Budget_sd_AP
```

```
## [1] 57100689
```

```r
#to calculate coefficient of variance for Budget_AP.
Budget_CV_AP <- Budget_sd_AP/Budget_mean_AP

#to print value of coefficient of variance for Budget_AP.
Budget_CV_AP
```

```
## [1] 1.060476
```

```r
#to get value of coefficient of variance for Budget_CV_AP up to two
#decimal points.
round(Budget_CV_AP, 2)
```

```
## [1] 1.06
```

d. Does the budget or the gross sales of a movie have more variation?

```r
print("There is more variation in budget and gross sales of the movie.")
```

```
## [1] "There is more variation in budget and gross sales of the movie."
```

2. Calculate the 32nd percentile of the number of Likes given. This calculation should be rounded to the nearest whole number (no decimal places).

```r
# to calculate 32nd percentile of the no. of likes.
Likes_quantile_AP <- quantile(Assignment01_AP$Likes_AP, c(.32))

# to print 32nd percentile of the no. of likes.
Likes_quantile_AP
```

```
##     32%
## 3354.88
```

```r
#to get whole number value of 32nd percentile of the no. of
#likes.quantile(Assignment01_AP$Likes_AP, .32)
round(Likes_quantile_AP,0)
```

```
##  32%
```

```
## 3355
```

Q3.Organizing Data 1. Summary Table a. Create a table showing average rating by year. This should be rounded to two decimal places.

```
# to get average rating of movies by year.
avgRating_AP <- aggregate(Assignment01_AP[,3], by=list(Assignment01_AP$Year_AP), FUN=mean, na.rm=TRUE)

#to print the above result.
avgRating_AP
```

```
##   Group.1        x
## 1    2014 6.435338
## 2    2015 6.403704
```

```
# to round up value of avgRating_AP with two decimals.
round(avgRating_AP,2)
```

```
##   Group.1    x
## 1    2014 6.44
## 2    2015 6.40
```

b. Which year's movies have the highest rating? What is it?

```
print("2014's movies have the highest rating , which is 6.44.")
```

```
## [1] "2014's movies have the highest rating , which is 6.44."
```

Q3 (2)

Cross Tabulation a. Create a table counting all genres of movies and which sequel number it is.

```
#to create table of genres of movies with sequel number.
genseq_AP <- with(Assignment01_AP, table(Genre_AP, Sequel_AP))
```

b. Change the table to show the percentage of each genre that is the 1st, 2nd, etc. movie in the series. These should be rounded to two decimal places.

```
#to get percentage og each genre with two decimal places.
genseq_AP<- round(prop.table(genseq_AP)*100,2)
genseq_AP
```

```
##         Sequel_AP
## Genre_AP     1     2     3     4     5     6     7
##       1  19.25  5.35  1.07  1.60  1.07  0.00  1.07
##       2   2.67  0.53  1.60  0.00  0.53  0.53  0.00
##       3  17.65  1.07  0.53  0.00  0.53  0.00  0.00
##       6   1.07  0.00  0.00  0.00  0.00  0.00  0.00
##       7   0.53  0.00  0.00  0.00  0.00  0.00  0.00
##       8  18.18  4.28  0.53  0.00  0.00  0.00  0.00
##       9   5.35  0.00  0.00  0.00  0.00  0.00  0.00
##       10  4.28  0.53  0.00  0.00  0.00  0.00  0.00
##       12  4.81  1.07  0.53  0.00  0.00  0.00  0.00
##       15  3.74  0.00  0.00  0.00  0.00  0.00  0.00
```

c. What percentage of movies in genre number 8 are not sequel?

```
print("18.18% of movies in genre number 8 are not sequel.")
```

```
## [1] "18.18% of movies in genre number 8 are not sequel."
```

Q3(3).

Bar Plot a. Create a bar plot of genre of movies. b. The plot should be: i. Rank ordered by highest count of genre. ii. Properly labeled (title, x-axis, etc) iii.The bars should have a different colour than the one shown in class.

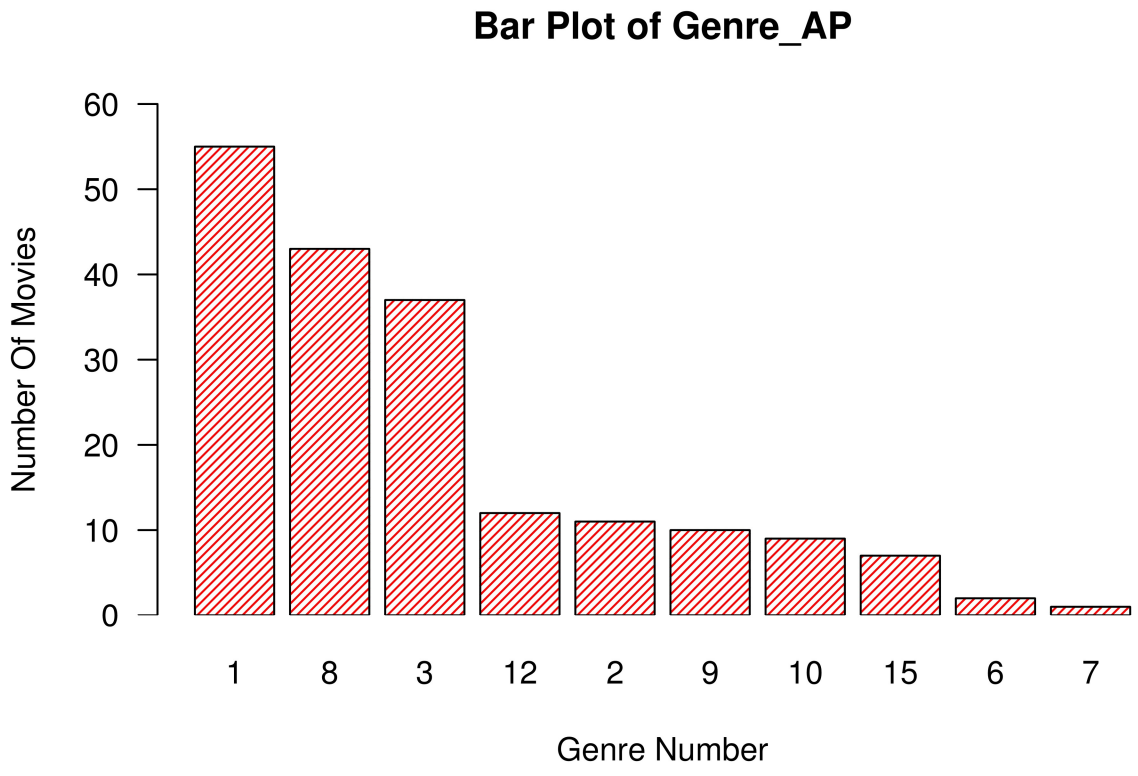
c. Based on the bar plot, (approximately) how many movies are there in genre number 8?

```
#a

# to create table for Genre_AP.
Genres_AP <- table(Assignment01_AP$Genre_AP)
#to set Genres_AP in decreasing order as we need the highest count 1st in the Bar Plot
Genres_AP <- Genres_AP[order(Genres_AP,decreasing=TRUE)]

#to get Bar Plot For Genres_AP, (col has default code.However, I prefer to write
#color name),(main gives title to graph, xlab and ylab give lable to x axis and
#y axis respectively),(ylim is set from 0 to 60 to get all the point clearly),
#(las=1 gives unit of y labled horizontally)

bar_AP <- barplot(Genres_AP,
        col="red",
        density = 30,
        main="Bar Plot of Genre_AP",
        xlab="Genre Number",
        ylab= "Number Of Movies",
        ylim = c(0,60), las=1)
```




```
print("Based on barplot there are approximately 43 movies fall under genre number 8.")
```

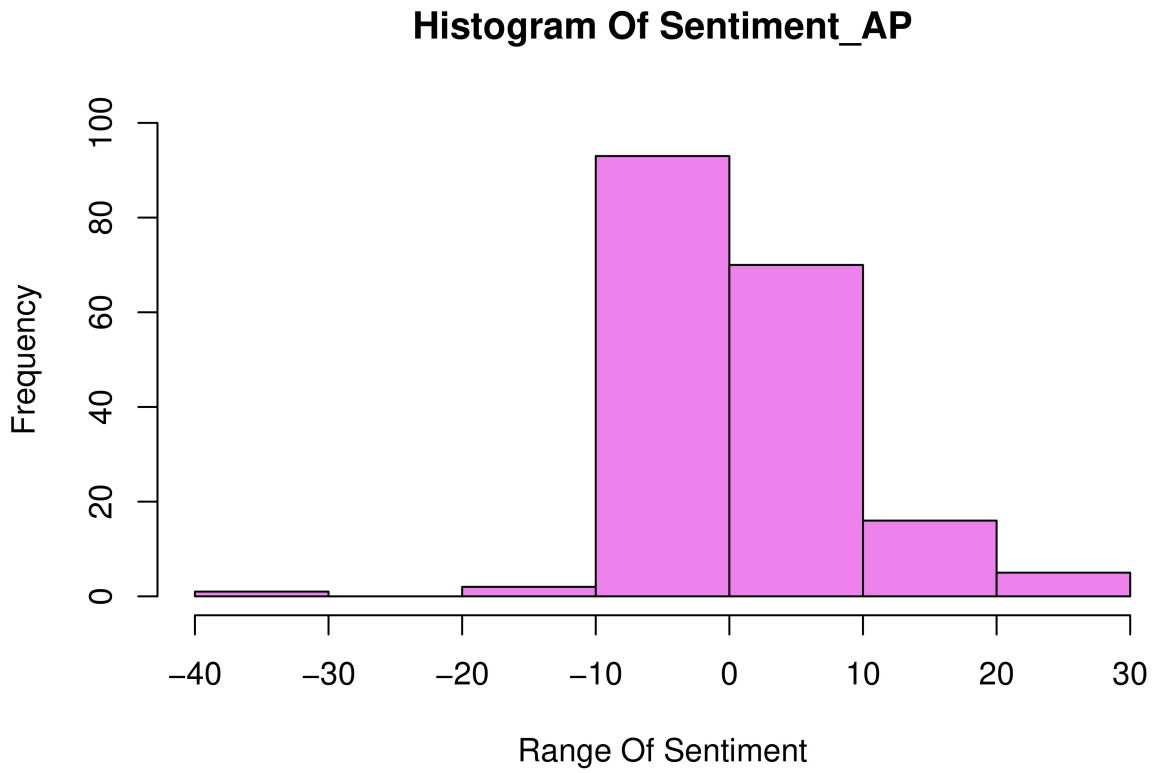
```
## [1] "Based on barplot there are approximately 43 movies fall under genre number 8."
```

Q3(4)

Histogram a. Create a histogram of sentiment. b. The plot should be properly labeled and a unique colour. c. Which range of sentiment is the most common?

```
# To get Histogram of Sentiment_AP with properly label and unique color.

hist_AP<-hist(Assignment01_AP$Sentiment_AP,col = "violet",
     xlab="Range Of Sentiment",
     main = "Histogram Of Sentiment_AP",
     xlim=c(-40,30),
     ylim=c(0,100))
```




```
print("From the derived histogram, from -10 to 0 range is the most common.")
```

```
## [1] "From the derived histogram, from -10 to 0 range is the most common."
```

Q3(5)

Box plot a. Create a horizontal box plot of number of screens the movies were shown on. b. The plot should be properly labeled and a unique colour. c. Based on the box plot, approximately how many movies were on fewer than 775
screens?

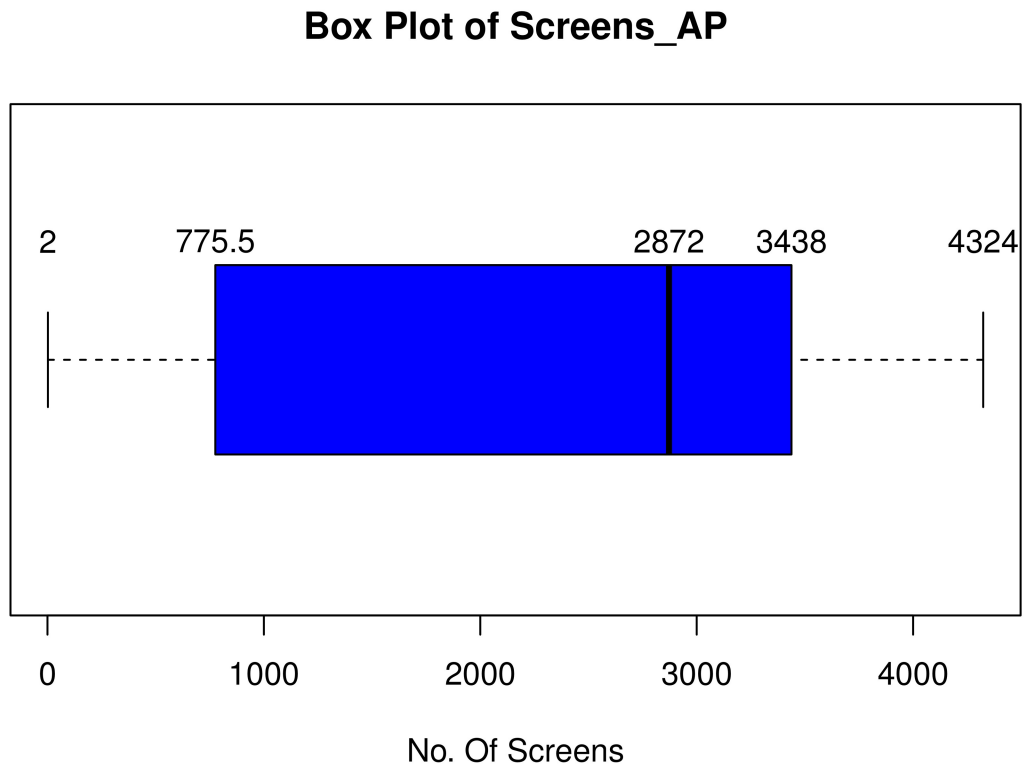
```
# To get Box Plot for Screens_AP

box_AP<-boxplot(Assignment01_AP$Screens_AP,
       main="Box Plot of Screens_AP",
```

```
        xlab="No. Of Screens",
        col="blue",
        horizontal=TRUE,
        pch=19)

# fivenum() function in R Language is used to return Tukey's five-number summary
#of input data i.e., minimum value, lower-hinge value, median value, upper-hinge
#value and maximum value of the input data.

text(x=fivenum(Assignment01_AP$Screens_AP), labels =fivenum(Assignment01_AP$Screens_AP), y=1.25)
```




```
#c

print("There were approximately 25% movies on fewer than 775 screens as it was in 1st quartile.")
```

```
## [1] "There were approximately 25% movies on fewer than 775 screens as it was in 1st quartile."
```

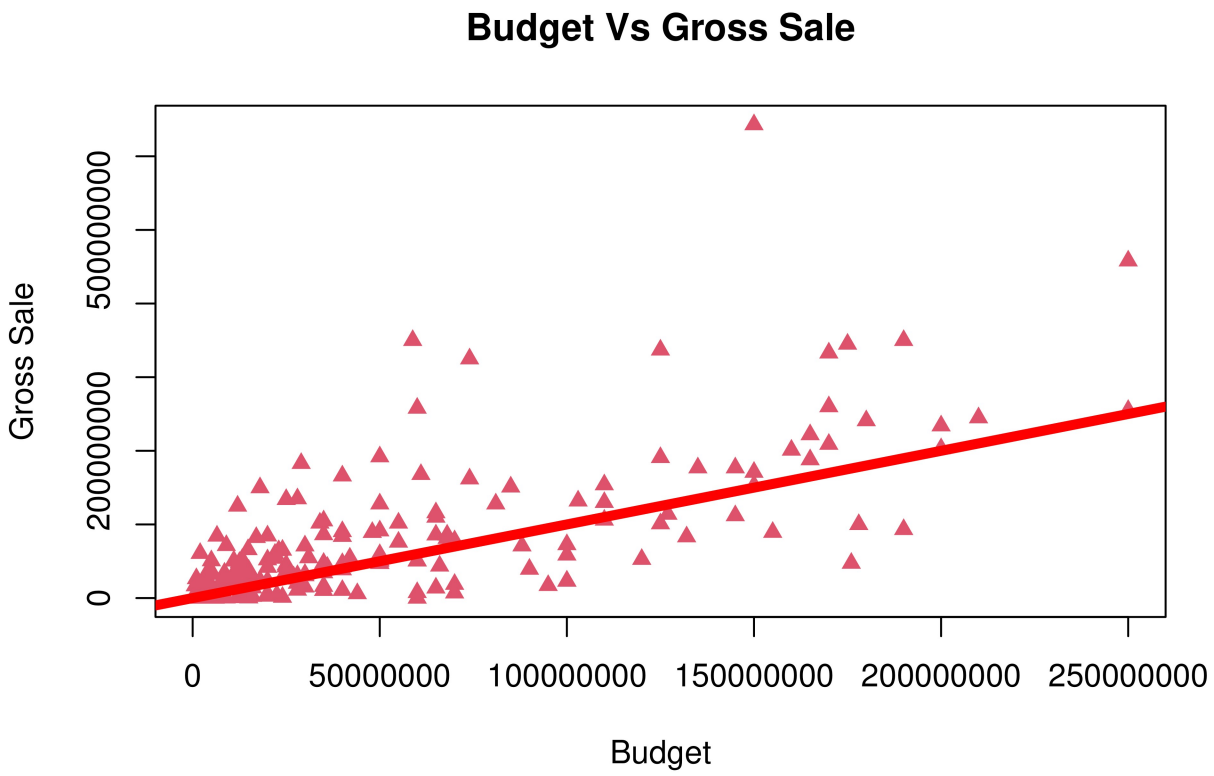
Q3(6)

Scatter Plot a. Create a scatter plot comparing budget and gross sales. b. The plot should be properly labeled with a marker type different than the one demonstrated in class. c. Add a line at 45 degrees to the chart.

```
# To plot Scatter Plot of budget and gross sales.

plot(Gross_AP~Budget_AP,
     data=Assignment01_AP,
     col=10,
```

```
     pch=17,
     main="Budget Vs Gross Sale",
     xlab="Budget",
     ylab="Gross Sale")
abline(coef = c(0, 1),
       col = "red",
       lwd = 5)
```



d. Does there appear to be an association between budget and gross sales for movies? From the Scatter Plot, in most of the cases budget and gross sales are
positively correlated.However, there are a few points that represents weak correlation between budget and gross sales.

e. What does it mean if a movie is plotted below the line? A movie is plotted below the line is considered poor(flop) in terms of gaining profit as we can see budget is high but gross sell is less. In other words, Weak correlation between budget and gross sale."