**#QMB Assignment 1**

**#Name - Manish Sanjay Talekar**

**#U Number – U54500765**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**#Preprocessing:**

**#Question 1.**

> main\_data <- read\_xlsx("6304 Module 1 Assignment Data.xlsx")

> main\_data

# A tibble: 155 x 5

Film `Production Costs` `Distribution Costs` `Domestic Revenue` `Foreign Revenue`

<chr> <dbl> <dbl> <dbl> <dbl>

1 STREET GIRL 211 293 806 198

2 VAGABOND LOVER 204 217 671 85

3 THE SAINT IN NEW YORK 128 137 350 310

4 BACHELOR MOTHER 509 639 1170 805

5 TOP HAT 609 1268 1782 1420

6 LITTLE WOMEN 424 776 1337 663

7 RIO RITA 678 787 1775 625

8 CUCKOOS 407 121 662 201

9 FIVE CAME BACK 225 231 441 280

10 KITTY FOYLE 738 778 1710 675

# ... with 145 more rows

**#Question 2.**

> set.seed(54500765)

> sub\_data\_set <- sample\_n(main\_data, 45)

> sub\_data\_set

# A tibble: 45 x 5

Film `Production Costs` `Distribution Costs` `Domestic Revenue` `Foreign Revenue`

<chr> <dbl> <dbl> <dbl> <dbl>

1 LAST DAYS OF POMPEII 818 399 489 491

2 HAVING A WONDER 966 309 771 237

3 SIN TAKES A HOLIDAY 450 213 463 160

4 LOVE AFFAIR 860 669 975 775

5 ROOKIE COP 77 67 108 54

6 CROSS FIRE 26 42 74 24

7 BRINGING UP BABY 1073 401 715 394

8 LIFE OF VERGIE WINTERS 331 236 506 148

9 MY LIFE WITH CAROLINE 503 359 530 300

10 ANNIE OAKLEY 354 218 435 185

# ... with 35 more rows

**#Question 3.**

> sub\_data\_set$TotalRevenue <- sub\_data\_set$`Domestic Revenue` +

+ sub\_data\_set$`Foreign Revenue`

> sub\_data\_set

# A tibble: 45 x 6

Film `Production Costs` `Distribution Cos~ `Domestic Revenu~ `Foreign Revenu~ TotalRevenue

<chr> <dbl> <dbl> <dbl> <dbl> <dbl>

1 LAST DAYS OF PO~ 818 399 489 491 980

2 HAVING A WONDER 966 309 771 237 1008

3 SIN TAKES A HOL~ 450 213 463 160 623

4 LOVE AFFAIR 860 669 975 775 1750

5 ROOKIE COP 77 67 108 54 162

6 CROSS FIRE 26 42 74 24 98

7 BRINGING UP BABY 1073 401 715 394 1109

8 LIFE OF VERGIE ~ 331 236 506 148 654

9 MY LIFE WITH CA~ 503 359 530 300 830

10 ANNIE OAKLEY 354 218 435 185 620

# ... with 35 more rows

**#Question 4.**

> sub\_data\_set$TotalProfitGenerated <- sub\_data\_set$TotalRevenue -

+ sub\_data\_set$`Production Costs` -

+ sub\_data\_set$`Distribution Costs`

> sub\_data\_set

# A tibble: 45 x 7

Film `Production Costs` `Distribution Cost~ `Domestic Revenu~ `Foreign Revenu~ TotalRevenue TotalProfitGenera~

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 LAST DAYS OF PO~ 818 399 489 491 980 -237

2 HAVING A WONDER 966 309 771 237 1008 -267

3 SIN TAKES A HOL~ 450 213 463 160 623 -40

4 LOVE AFFAIR 860 669 975 775 1750 221

5 ROOKIE COP 77 67 108 54 162 18

6 CROSS FIRE 26 42 74 24 98 30

7 BRINGING UP BABY 1073 401 715 394 1109 -365

8 LIFE OF VERGIE ~ 331 236 506 148 654 87

9 MY LIFE WITH CA~ 503 359 530 300 830 -32

10 ANNIE OAKLEY 354 218 435 185 620 48

# ... with 35 more rows

**#Analysis:**

**#Question 1.**

> str(sub\_data\_set)

tibble [45 x 7] (S3: tbl\_df/tbl/data.frame)

$ Film : chr [1:45] "LAST DAYS OF POMPEII" "HAVING A WONDER" "SIN TAKES A HOLIDAY" "LOVE AFFAIR" ...

$ Production Costs : num [1:45] 818 966 450 860 77 ...

$ Distribution Costs : num [1:45] 399 309 213 669 67 42 401 236 359 218 ...

$ Domestic Revenue : num [1:45] 489 771 463 975 108 74 715 506 530 435 ...

$ Foreign Revenue : num [1:45] 491 237 160 775 54 24 394 148 300 185 ...

$ TotalRevenue : num [1:45] 980 1008 623 1750 162 ...

$ TotalProfitGenerated: num [1:45] -237 -267 -40 221 18 30 -365 87 -32 48 ...

**#Question 2.**

> mean(sub\_data\_set$TotalProfitGenerated)

[1] 14.35556

> median(sub\_data\_set$TotalProfitGenerated)

[1] 30

> sd(sub\_data\_set$TotalProfitGenerated)

[1] 184.2036

> skewness(sub\_data\_set$TotalProfitGenerated)

[1] -0.1487643

> kurtosis(sub\_data\_set$TotalProfitGenerated)

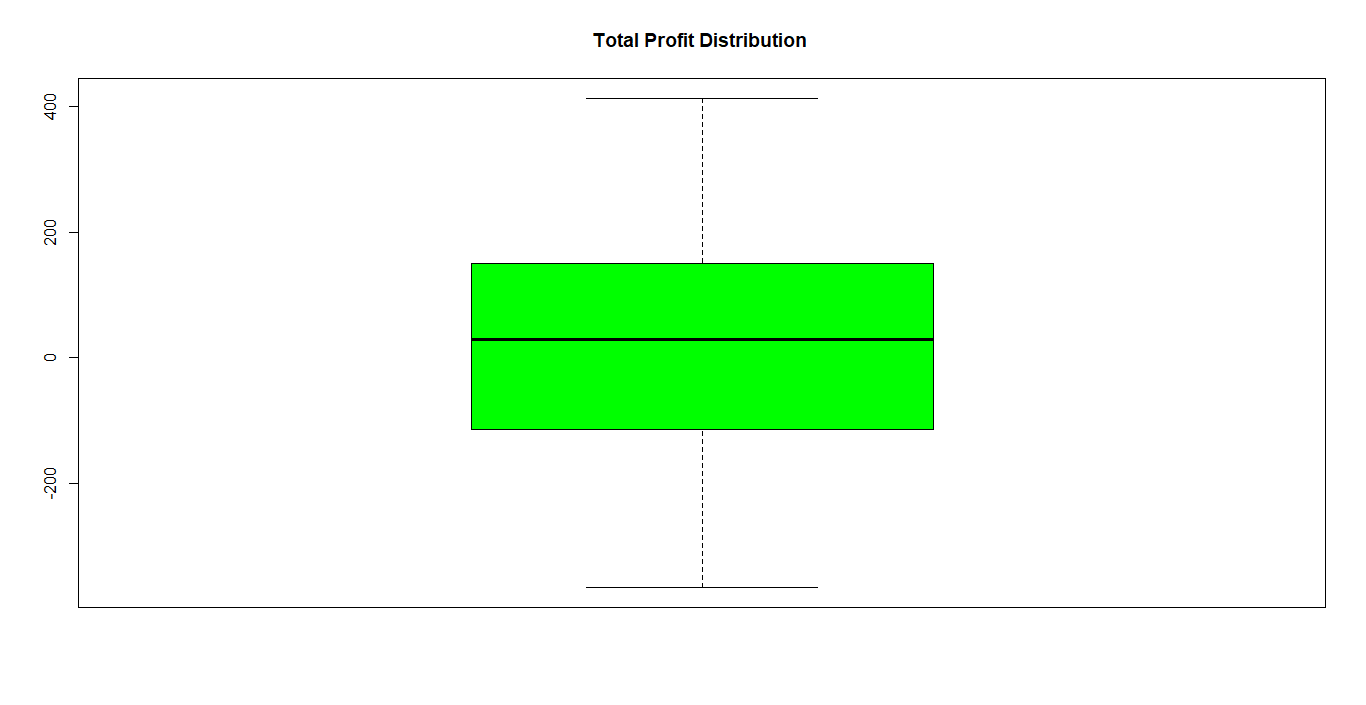
[1] 2.653156

INTERPRETATION – Distribution of data is not perfectly symmetric; the data seems to be slightly skewed to the right.

**#Question 3.**

> boxplot(sub\_data\_set$TotalProfitGenerated, col="green",

+ main="Total Profit Distribution ")



**#Question 4.**

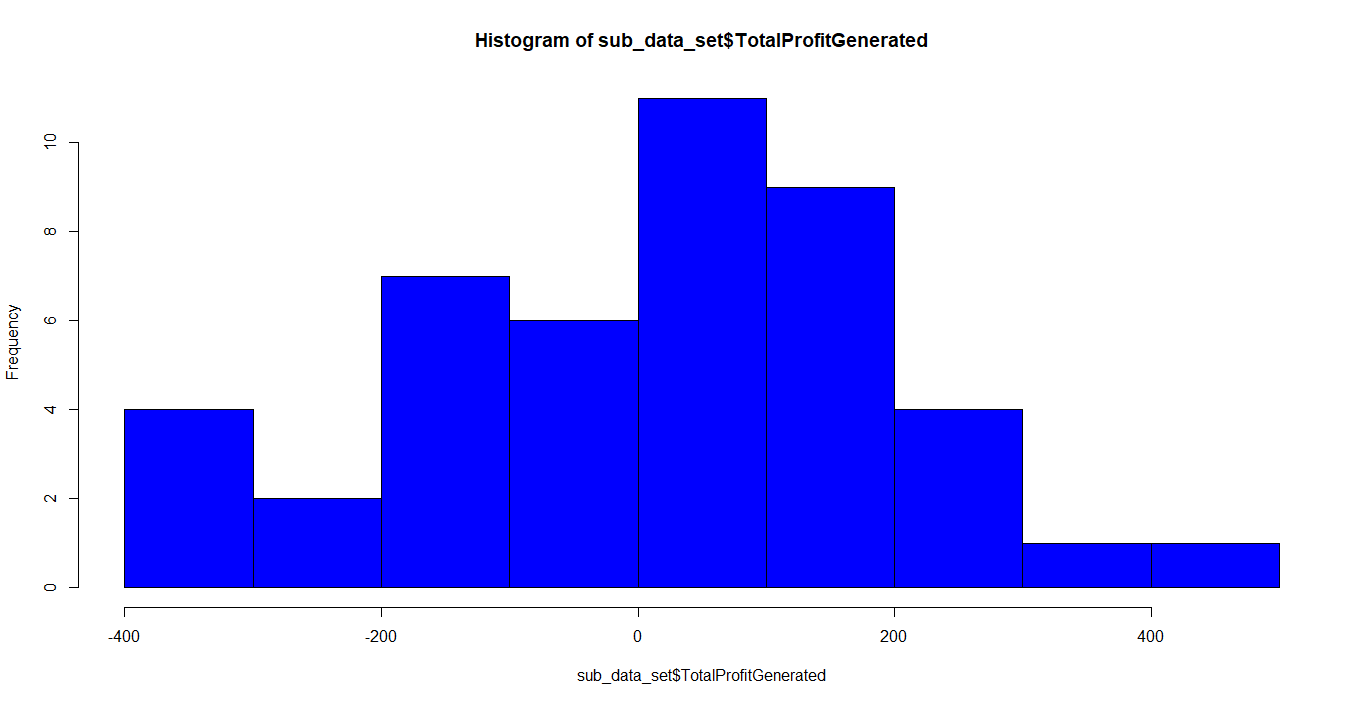
> quantile(sub\_data\_set$`Domestic Revenue`,probs=seq(0,1,.1))

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

29.0 130.4 321.6 379.6 449.8 489.0 568.2 705.0 888.2 1059.8 1751.0

**#Question 5.**

> hist(sub\_data\_set$TotalProfitGenerated, col="blue")



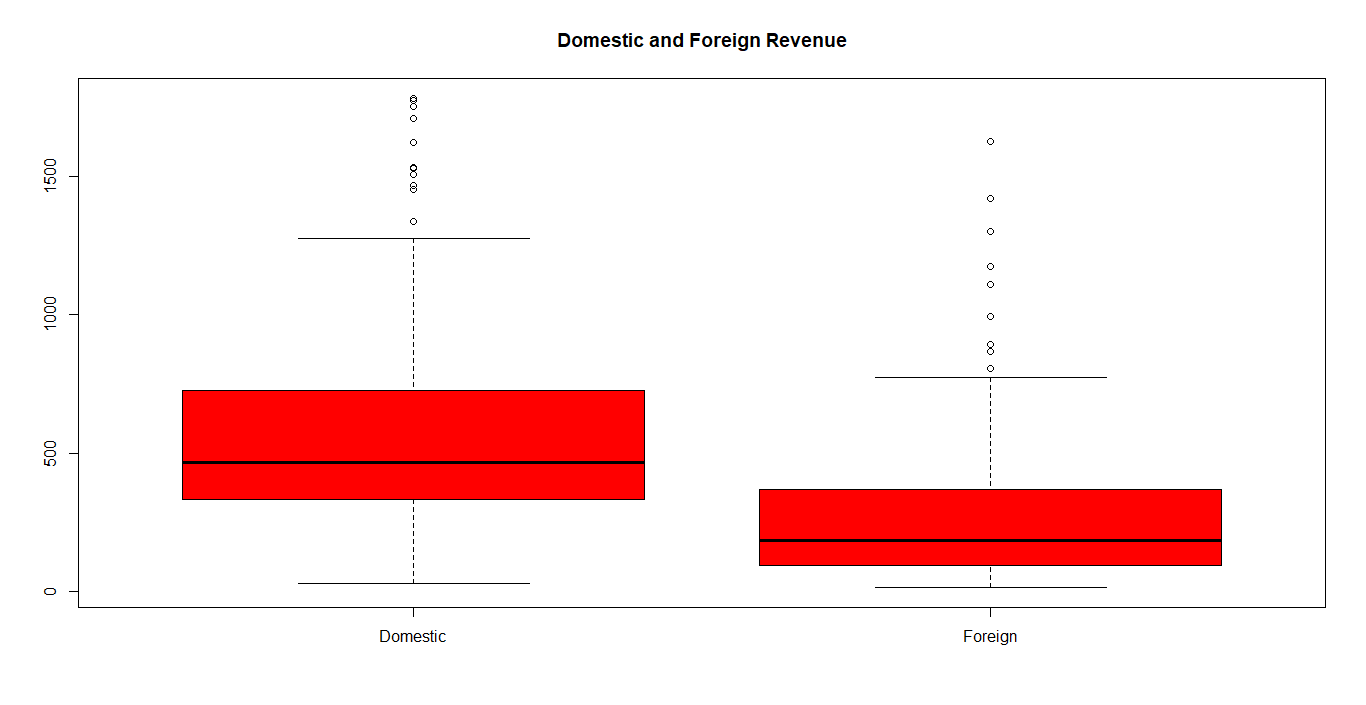
INTERPRETATION – From the histogram I would like to conclude that majority of RKO films were profitable.

**#Question 6.**

> boxplot(main\_data$`Domestic Revenue`, main\_data$`Foreign Revenue`,

+ col="red", main="Domestic and Foreign Revenue",

+ names=c('Domestic', 'Foreign'))



R Script –

