Institute of Systems Science

National University of Singapore

GRADUATE CERTIFICATE

BUSINESS ANALYTICS PRACTICE

Supplementary Workshop Guide

Subject: *NICF- Statistics Bootcamp*

Workshop 1.6

Background

The below table lists out the U.S. top 15 grossing movies in year 2017. The unit for TopGross is USD 1 million.

Table 1 :U.S. Top 15 grossing movies in year 2017

Film	Studio	TopGross	OpenQuarter
The Last Jedi	Disney	620	4
Beauty and the Beast	Disney	504	1
Wonder Woman	Warner	413	2
Jumanji	Sony	405	4
Guardians of the Galaxy	Disney	390	2
Spider-Man	Sony	334	3
It	Warner	328	3
Thor: Ragnarok	Disney	315	4
Despicable Me 3	Universal	265	2
Justice League	Warner	229	4
Logan	Fox	226	1
The Fate of the Furious	Universal	226	2
Coco	Disney	210	4
Dunkirk	Warner	188	3
Get Out	Universal	176	1

Task

Complete the below tasks:

a. Create four vectors: Film, Studio, TopGross, OpenQuarter.
b. Create a dataframe TopMovies; strings should not be encoded as factor.
c. Encode Studio as factor.
d. In R, categorical variables are usually represented by factors. Encode 1,2,3,4 in OpenQuarter as factors '1st', '2nd', '3rd', '4th'.
e. Use R command to find out how many Disney movies were there in the top 15 grossing.
f. Use R command to find out how many Disney movies were there in the top **10** grossing.
g. How much Disney has earned from the movies in the top **15** grossing?
h. How much Disney has earned from the movies in the top **10** grossing?
i. How much Disney has earned from the movies in the top **5** grossing?
j. Use R command to find out how many Warner movies were there in the top **15** grossing.
k. Use R command to find out how many Warner movies were there in the top **10** grossing.
l. How much Warner has earned from the movies in the top **15** grossing?
m. How much Warner has earned from the movies in the top **10** grossing?
n. How much Warner has earned from the movies in the top **5** grossing?
o. Plot a graph that illustrates the number of movies made by each studio in the top **15** grossing list
p. Plot a graph that illustrates the number of movies made by each studio in the top **10** grossing list
q. Plot a graph that illustrates the total revenue each studio has earned from the movies in the top **15** grossing
r. Plot a graph that illustrates the total revenue each studio has earned from the movies the top **10** grossing
s. Plot a graph that illustrates the total revenue each studio has earned from the movies in the top **15** grossing (Using ggplot)
t. Assume you want to know which quarter in a year is the best time to launch a movie. Plot a graph to illustrate the mean revenue received in each quarter (Using ggplot)

Solutions

a. Create four vectors: Film, Studio, TopGross, OpenQuarter.

```
Film = c('The Last Jedi',
          'Beauty and the Beast',
          'Wonder Woman',
          'Jumanji',
          'Guardians of the Galaxy',
          'Spider-Man',
          'It',
          'Thor: Ragnarok',
          'Despicable Me 3',
          'Justice League',
          'Logan',
          'The Fate of the Furious',
          'Coco',
          'Dunkirk',
          'Get Out')

Studio = c('Disney',
            'Disney',
            'Warner',
            'Sony',
            'Disney',
            'Sony',
            'Warner',
            'Disney',
            'Universal',
            'Warner',
            'Fox',
            'Universal',
            'Disney',
            'Warner',
            'Universal')

TopGross = c(620,
             504,
             413,
             405,
             390,
             334,
             328,
             315,
             265,
             229,
             226,
             226,
             210,
             188,
             176)

OpenQuarter = c(4,
                1,
                2,
                4,
                2,
                3,
                3,
                4,
                2,
                4,
                1,
                2,
                4,
                3,
                1)
```

b. Create a dataframe TopMovies; strings should not be encoded as factor.

```
TopMovies = data.frame(Film,Studio,TopGross,OpenQuarter,
                       stringsAsFactors = FALSE)
```

c. Encode `Studio` as factor.

```
TopMovies$Studio = factor(TopMovies$Studio)
```

d. In R, categorical variables are usually represented by factors. Encode 1,2,3,4 in OpenQuarter as factors ‘1st’, ‘2nd', ‘3rd', ‘4th'.

```
TopMovies$OpenQuarter = factor(TopMovies$OpenQuarter,
                               levels = c(1,2,3,4),
                               labels = c('1st','2nd','3rd','4th'))
```

e. Use R command to find out how many Disney movies were there in the top 15 grossing.

```
sum(TopMovies$Studio == 'Disney')
```

f. Use R command to find out how many Disney movies were there in the top 10 grossing.

```
sum(TopMovies$Studio[1:10] == 'Disney')
```

g. How much Disney has earned from the movies in the top 15 grossing?

```
sum(TopMovies$TopGross[TopMovies$Studio == 'Disney'])
```

h. How much Disney has earned from the movies in the top 10 grossing?

```
Top10 = TopMovies[1:10,]
sum(Top10$TopGross[Top10$Studio == 'Disney'])

# Question: Why this command is incorrect?
# sum(TopMovies$TopGross[TopMovies$Studio[1:10] == 'Disney'])
```

i. How much Disney has earned from the movies in the top 5 grossing?

```
Top5 = TopMovies[1:5,]
sum(Top5$TopGross[Top5$Studio == 'Disney'])
```

j. Use R command to find out how many Warner movies were there in the top 15 grossing.

```
sum(TopMovies$Studio == 'Warner')
```

k. Use R command to find out how many Warner movies were there in the top 10 grossing.

```
sum(TopMovies$Studio[1:10] == 'Warner')
```

l. How much Warner has earned from the movies in the top 15 grossing?

```
sum(TopMovies$TopGross[TopMovies$Studio == 'Warner'])
```

m. How much Warner has earned from the movies in the top 10 grossing?

```
sum(Top10$TopGross[Top10$Studio == 'Warner'])
```

n. How much Warner has earned from the movies in the top 5 grossing?

```
sum(Top5$TopGross[Top5$Studio == 'Warner'])
```

o. Plot a graph that illustrates the number of movies made by each studio in the top 15 grossing list

```
barplot(table(TopMovies$Studio))
```

p. Plot a graph that illustrates the number of movies made by each studio in the top 10 grossing list

```
barplot(table(Top10$Studio))
```

q. Plot a graph that illustrates the total revenue each studio has earned from the movies in the top 15 grossing

```
Total15 = aggregate(TopMovies$TopGross, by=list(Studio=TopMovies$Studio),
FUN=sum)
barplot(Total15$x,names.arg = Total15$Studio)
```

r. Plot a graph that illustrates the total revenue each studio has earned from the movies the top 10 grossing

```
Total10 = aggregate(Top10$TopGross, by=list(Studio=Top10$Studio),FUN=sum)
barplot(Total10$x,names.arg = Total10$Studio)
```

s. Plot a graph that illustrates the total revenue each studio has earned from the movies in the top 15 grossing (Using ggplot)

```
library(ggplot2)
ggplot(Total15,aes(x=Studio,y=x))+geom_bar(stat='identity')
```

t. Assume you want to know which quarter in a year is the best time to launch a movie. Plot a graph to illustrate the mean revenue received in each quarter (Using ggplot)

```
QuarterMean = aggregate(TopMovies$TopGross,
by=list(Quarter=TopMovies$OpenQuarter),FUN=mean)

ggplot(QuarterMean,aes(x=Quarter,y=x))+geom_bar(stat='identity')
```