

Movies

2016-2019

Eusebia ,Winifred, Hamzata






⇒ Worked with R Studio and downloaded packages to accurately create and display a set of data.

⇒ Used over 2,000 movies released from 2016-2019 to create a dataframe by web scraping.

⇒ Generated 5 hypotheses and tested them using our data frame.

⇒ Specifically looked into release periods, genres, tickets sold, and movie distributors to validate our hypotheses.

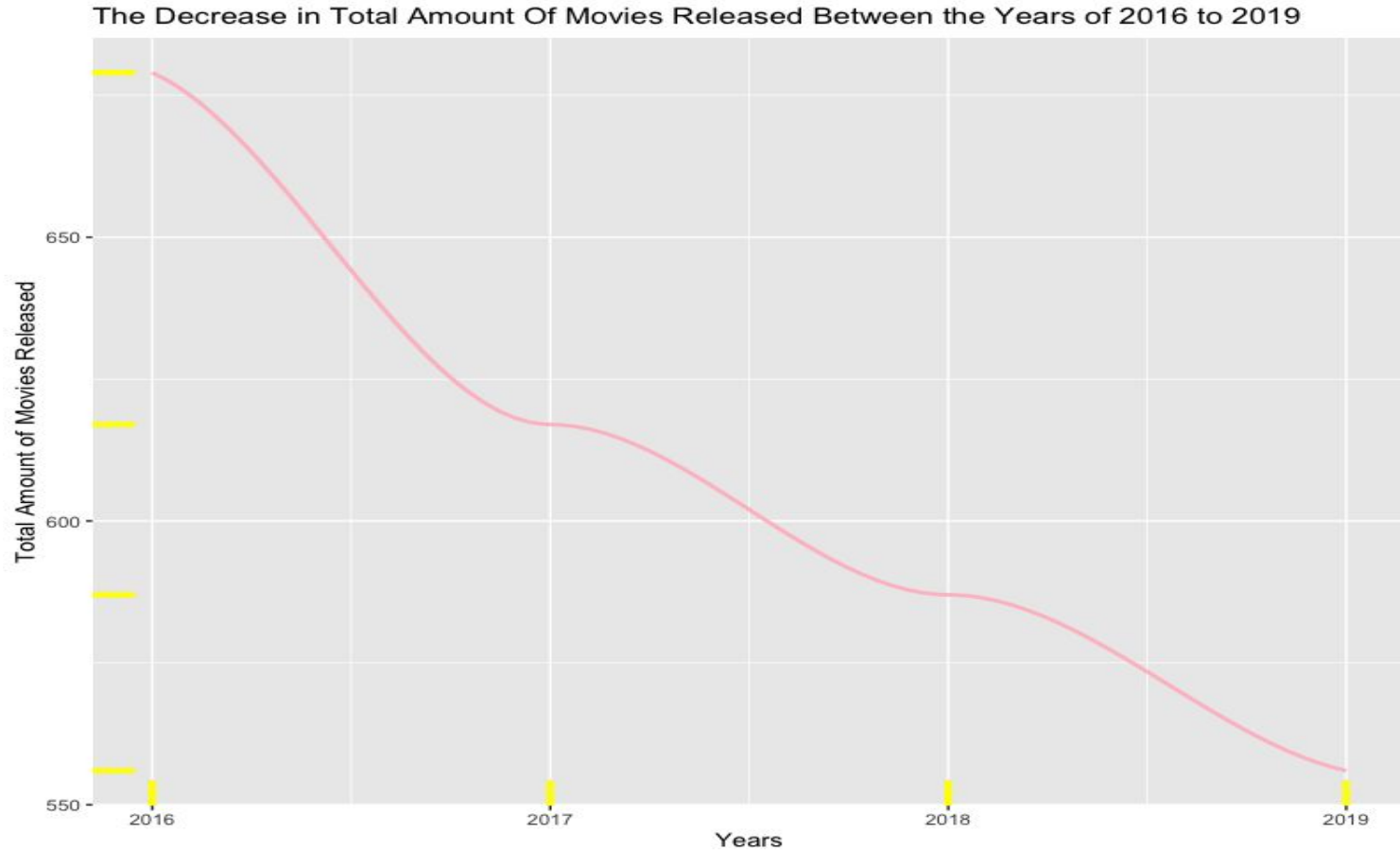
⇒ Used R Studio to create quadratic graphs, line graphs, bar graphs and a table to display information from our sample data.



Disclaimer:

These findings are based on just over 2,000 movies that we have web scraped to form a data frame. [See attached data frame](#). If the data frame was extensive the results would be have been different.

Hypothesis 1: The number of movies being released has increased from 2016 to 2019.



R- Studio Code for Plotting Graph 1.

```
library(readxl)
Movies <- read_excel("~/Desktop/Movies (1).xlsx",
                     col_types = c("numeric", "text", "date",
                                   "text", "text", "numeric", "numeric"))

View(Movies)
Movies <- data.frame(Movies)
View(Movies)

library(ggplot2)
library(dplyr)

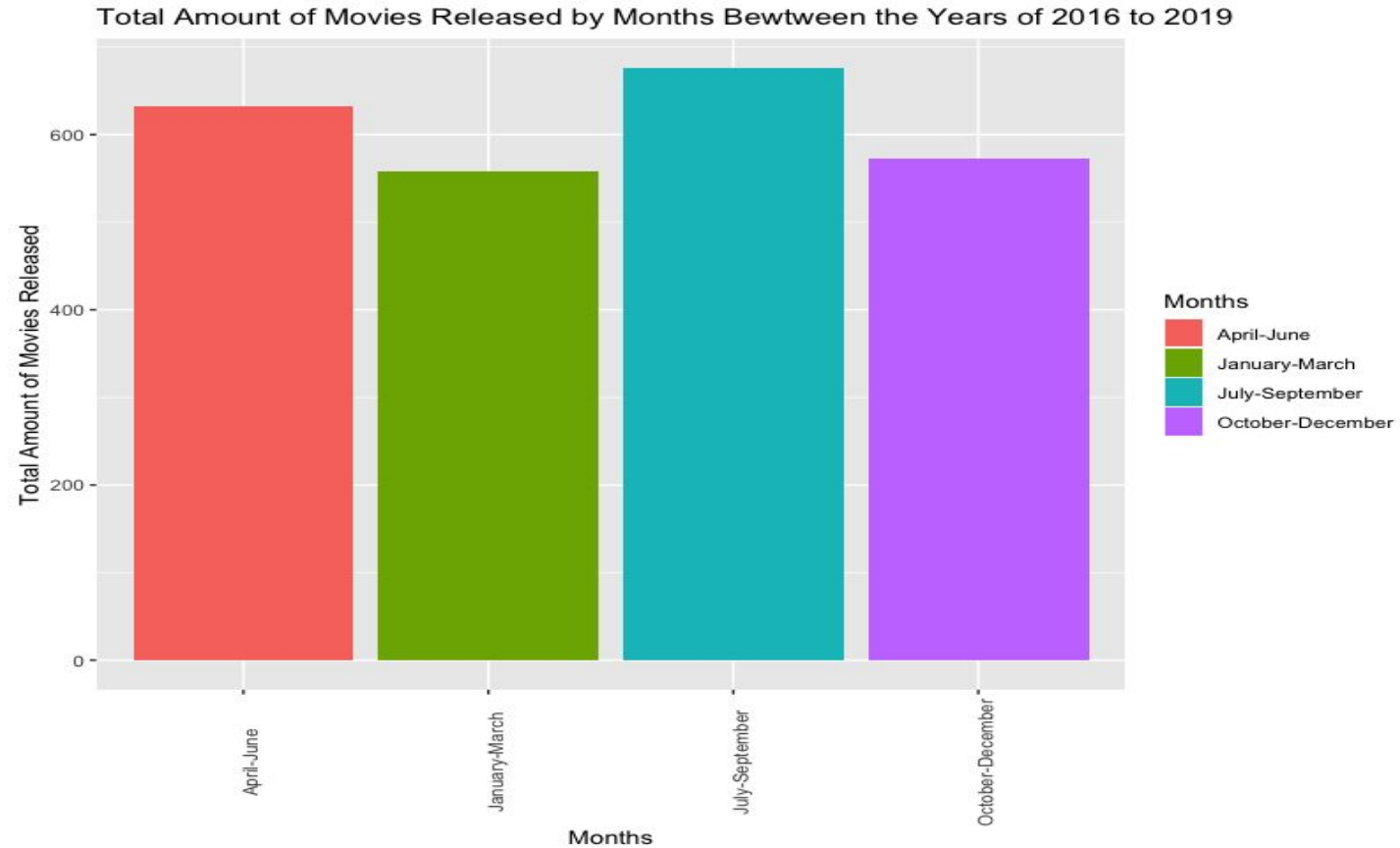
#Graph 1
Years <- c(2016,2017,2018,2019)
Amount_of_Movies_Released <- c(679,617,587,556)
One <- data.frame(Years,Amount_of_Movies_Released)

One.Graph <- ggplot(data=One, aes(x=Years, y=Amount_of_Movies_Released)) +
  geom_smooth(color="pink") +
  ggtitle("The Decrease in Total Amount Of Movies Released Between the Years of 2016 to 2019")
Y <- print(One.Graph+labs (y="Total Amount of Movies Released", x= "Years"))
```

Analysis and Finding for Graph 1

- ❑ Our first hypothesis was proven not valid. In fact, according to graph 1 the numbers of movies that were being released decreased from 2016 to 2019. The most drastic drop happened during 2016 where the slope dropped dramatically compared to the decrease from 2017-2019
- ❑ The amount of movies that were released kept decreasing per year. This resulted in a negative slope, instead of a positive slope.

Hypothesis 2: Most movies released in 2016-2019 were released during the months of July, August, September.



R- Studio Code for Plotting Graph 2

```
#Graph 2
Months<-months.Date(Movies$Release.Date)
Movies <-cbind(Movies,Months)

Months[Months=="January"] <- "January-March"
Months[Months== "February"] <- "January-March"
Months[Months== "March"] <- "January-March"

Months[Months=="April"] <- "April-June"
Months[Months=="May"] <- "April-June"
Months[Months=="June"] <- "April-June"

Months[Months=="July"] <- "July-September"
Months[Months=="August"] <- "July-September"
Months[Months=="September"] <- "July-September"

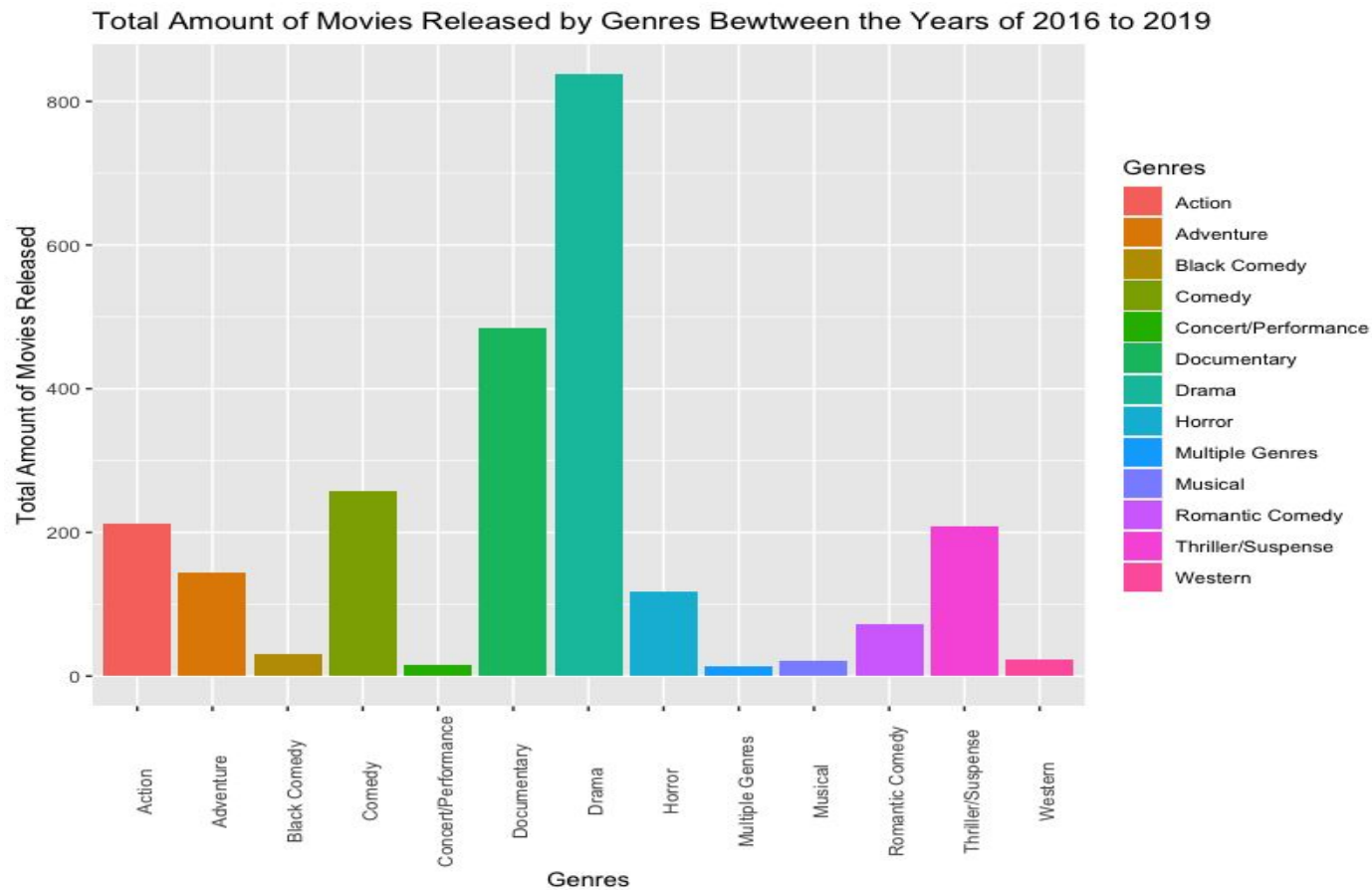
Months[Months=="October"] <- "October-December"
Months[Months=="November"] <- "October-December"
Months[Months=="December"] <- "October-December"
Months.2<-Months
Months.2
Movies <- cbind(Movies,Months.2)
Movies

Two.Graph<- ggplot(Movies, aes(x=Months.2,fill=Months.2)) +
  geom_bar()+theme(axis.text.x = element_text(angle = 90))+
  ggtitle("Total Amount of Movies Released by Months Between the Years of 2016 to 2019"
)
X<- print(Two.Graph+labs (y="Total Amount of Movies Released", x= "Months", fill = "Months"))
```


Analysis and Finding for Graph 2

- ★ Our second hypothesis was proven valid. According to the second graph most movies released between the years of 2016 to 2019 were released during the months of July, August, September.
- ★ Based on our sample data and the data displayed in the second graph we can infer that July, August, and September were popular months when it came to releasing movies. January, February, and March on the other hand were not so popular months.

Hypothesis 3: During the years of 2016-2019, most movie genres fell under the categories of Drama and Action.



R- Studio Code for Plotting Graph 3

```
#Graph 3
Movies$Genre[Movies$Genre=="Concert/Perfor,Ä¶"] <- "Concert/Performance"
Three.Graph <- ggplot(Movies, aes(x=Genre,fill=Genre)) +
  geom_bar()+theme(axis.text.x = element_text(angle = 90))+
  ggtitle("Total Amount of Movies Released by Genres Bewtween the Years of 2016 to 2019"
)
Z<- print(Three.Graph+labs (y="Total Amount of Movies Released", x= "Genres", fill = "Ge
nres"))
Three.Graph
```



Analysis and Finding for Graph 3

- Our third hypothesis was proven not valid. According to graph 3 during the years of 2016 to 2019, most movie genres did fall under the category of Drama, but at the same time most movie genres did not fall under the category of Action.
- Based on our data and the data displayed on graph 3 we can infer that Drama and Documentary were popular movie genres from 2016 to 2019 as most movies released fell under these two genres.

Hypothesis 4: Walt Disney , Warner Brothers, and Universal Studios were Apart of the Top 50 Movie Distributors During the Years of 2016-2019.

Top 50 Movie Distributors 2016 - 2019.

Movie Distributors	Total Amount of Movies Distributed
Magnolia Pictures	87
Kino Lorber	86
Sony Pictures	83
Warner Bros.	81
IFC Films	81
Lionsgate	78
Well Go USA	77
Universal Studios	68
Indican Pictures	64
A24	56
Film Movement	56
Sony Pictures Classics	55
20th Century Fox	52

Abramorama Films	50
Strand	46
First Run Features	46
Paramount Pictures	45
The Orchard	43
Walt Disney	42
Oscilloscope Pictures	40
Roadside Attractions	39
China Lion Film Distribution	38
Focus Features	33
Film-Rise	33
Cohen Media Group	33
Music Box Films	32
STX Entertainment	31
GKIDS	28

Bleecker Street	27
CJ Entertainment	27
Neon	27
Fox Searchlight	26
Janus Films	25
Cinema Guild	22
FIP	21
Eros Entertainment	21
Greenwich	19
Gravitas Ventures	18
Amazon Studios	18
Zeitgeist	17
Distrib Films US	17
Open Road	16
Self-Distributed	16
Trafalgar Releasing	15
Mongrel Media	14
Paladin	13
Weinstein Co.	12

Rialto Pictures	12
Entertainment Studios	12
Pure Flix Entertainment	11

R- Studio Code for Creating Table 1.

```
# Table 1
duplicated(Movies$Distributor)
New <- Movies$Distributor[!duplicated(Movies$Distributor)]
New

for (i in 1:length(New)){
  print(New[i])
  print(count(Movies[which(Movies$Distributor==New[i]),]))
}

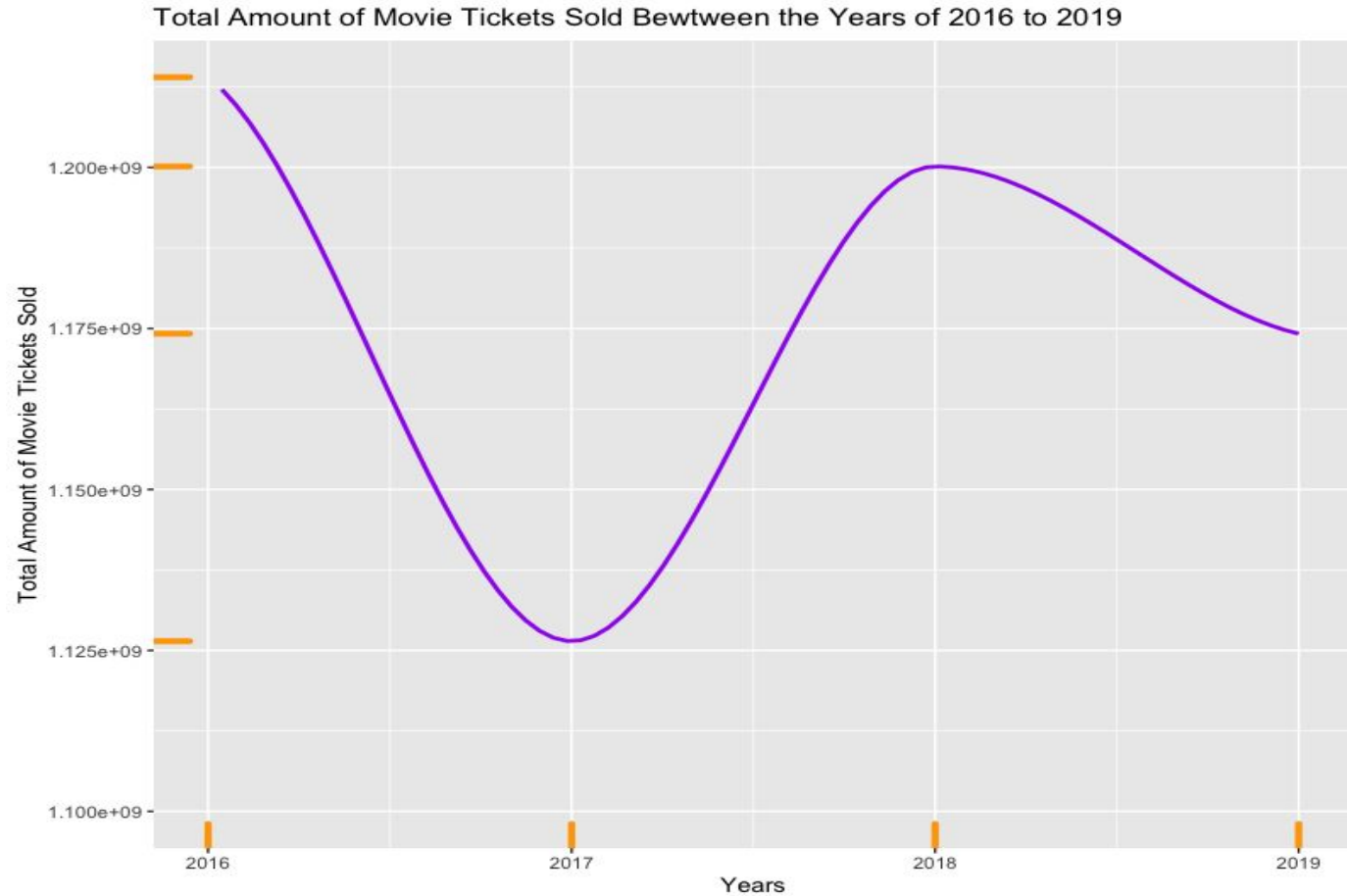
New.2 <- c( 42, 68, 52, 81, 45, 83, 31, 78, 33, 16, 39, 1, 56, 11, 3, 27, 10,
            2, 26, 2, 5, 55, 7, 10, 11, 12, 43, 4, 21, 3, 81, 11, 32, 50, 4, 77, 1, 2, 1, 8, 87,
            8, 1, 40, 21, 4, 1, 9, 6, 1, 1, 27, 1, 3, 1, 2, 38, 7, 1, 7, 1, 8, 33, 64, 33, 1, 28, 6,
            5, 3, 86, 4, 1, 2, 1, 5, 1, 25, 17, 16, 1, 46, 1, 18, 1, 56, 1, 6, 2, 3, 7, 8, 1, 1,
            4, 1, 1, 13, 7, 1, 8, 14, 1, 1, 17, 1, 2, 12, 4, 46, 1, 7, 11, 1, 9,
            3, 1, 6, 1, 9, 1, 1, 1, 1, 7, 1, 1, 1, 5, 3, 1, 1, 1, 1, 1, 5, 1, 1,
            2, 1, 1, 1, 2, 1, 12, 5, 2, 8, 4, 8, 7, 1, 1, 27, 1, 2, 1, 7, 18, 2, 6, 4, 1, 1, 2, 1, 3, 2, 3, 22,
            1, 1, 4, 1, 4, 1, 1, 4, 1, 1, 1, 1, 3, 4, 4, 19, 2, 3, 2, 3, 2, 4, 1, 1, 1, 2, 4, 1, 1, 2, 1,
            1, 1, 1, 1, 8, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 6, 1, 4, 15, 1,
            1, 1, 1, 1, 5, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1)
```

```
Dis.2 <- data.frame(New, New.2)
Dis.2
Dis.2 <- Dis.2[order(-Dis.2$New.2),]
write.csv(Dis.2, "Dis.2.csv")
```

Analysis and Finding for Table 1.

- ❖ Our fourth hypothesis was proven to be valid. According to the table Walt Disney, Warner Brothers, and Universal Studios were apart of the top 50 movie distributors during the years of 2016-2019.
- ❖ Based on our sample data and the data displayed on table 1 we can infer that Warner Brothers, and Universal Studios were very popular distributors during the years of 2016 to 2019. Walt Disney on the other hand was not as popular as Warner Brothers and Universal Studio.

Hypothesis 5: The number of tickets being sold have drastically decreased from 2016 to 2019.



R- Studio Code for Plotting Graph 4

```
#Graph 4
sum(Movies$Tickets.Sold[1:679])
sum(Movies$Tickets.Sold[680:1296])
sum(Movies$Tickets.Sold[1297:1883])
sum(Movies$Tickets.Sold[1884:2439])

Years <- c(2016,2017,2018,2019)
TATS <- c(1214020806,1126430739,1200161994,1174192136)
Graph.4 <- data.frame(Years,TATS)

Four.Graph <- ggplot(Graph.5,aes(x=Years,y=TATS)) +
  geom_smooth(color="Purple")+
  ylim(1.100e+09,1214020806)+
  geom_rug(col="orange",alpha=10, size=1.5)+
  ggtitle("Total Amount of Movie Tickets Sold Between the Years of 2016 to 2019")
A<- print(Four.Graph+labs (y="Total Amount of Movie Tickets Sold", x= "Years", fill="Years"))
Four.Graph
```

Analysis and Finding for Graph 4

- Our fifth hypothesis was proven not valid. According to our fourth graph the number of tickets being sold decreased from 2016 to 2017 and then increased from 2017 to 2018 and then decreased from 2018 to 2019.
- We expected our fourth graph to have been linear with a negative slope but instead our fourth graph was quadratic with a maximum at 2018 and minimum at 2017.

References

1. These are the websites that were used in gathering information for our sample data that contained 2,000 plus movies.
 - <https://www.the-numbers.com/market/2016/top-grossing-movies>
 - <https://www.the-numbers.com/market/2017/top-grossing-movies>
 - <https://www.the-numbers.com/market/2018/top-grossing-movies>
 - <https://www.the-numbers.com/market/2019/top-grossing-movies>