Movie Modeling

Overview

This data science project intent to analysize movies trends from available dataset available from last 100+ years.

Asking appropriate question or performing EDA (Exploratory Data Analysis) with the available dataset will help answer various trends in the past, present and probable future trends.

We expect that we will able to produce/answer following answer with the available dataset that will be of great help to the target audience.

We are asking the following question with the available dataset

- What's has been the trend of Movie Genre being made in last 100+ yrs ?
- What's has been the production budget trend ?
- What's has been the production budget per Genre ?
- What's has been the domestic box office collection ?
- What's type of genre tend to maximize profit for production houses ?
- What's month of year sees minimum/maximum movie releases ?

Target Audience

This exploratory study is intent to target focus group who are related to entertainment industry.

1. Production Houses
    - Disney
    - Universal Studio
2. Entertainment Media Industry
    - Film Magzines.
    - Film Critics.
3. New entrepreneur to the Movie Industry.
4. Theatre Owner.

Individual client can see ongoing trends and apply predictive analysis with answer available for the above question. With the limited budget available for each production houses, better and wise decision can be made to fund genre to maximize profit.

Theater Owner can better decide on leasing theater space as volume of movie release varies for each month.

Film Critics/Film Magzines can analyzes further with changing times how the taste of audience have occurred and can be a dataset for other datascience studies being done for behaviour changes occuring in the society.

Data Acquisition

We will acquire the data from http://www.the-numbers.com/movies (http://www.the-numbers.com/movies). The webcrawler
will acquire the dataset for each year and perform DataWrangling and present it as a single source dataset.

Release Date	Movie	Genre	Production Budget	Domestic BoxOffice To Date
September 5	Intolerance	Adventure	$200,000	$8,000,000
December 31	The Cabinet of Dr. Caligari	Horror	$150,000	$300,000

There are source for data acquisition and with project webcrawler can be extend for other sources and data wrangling process can mitigate/resolve difference appropriately.

Deliverables

This project will include R code and will be posted here **github.com (https://github.com/mohankri/datascience/tree/master/springboard/project)**

It will also include result html (using R-markdown) have answer to the problem
stated above.

Deliverables will include complete dataset, webCrawler for performing any update of dataset.

Goal

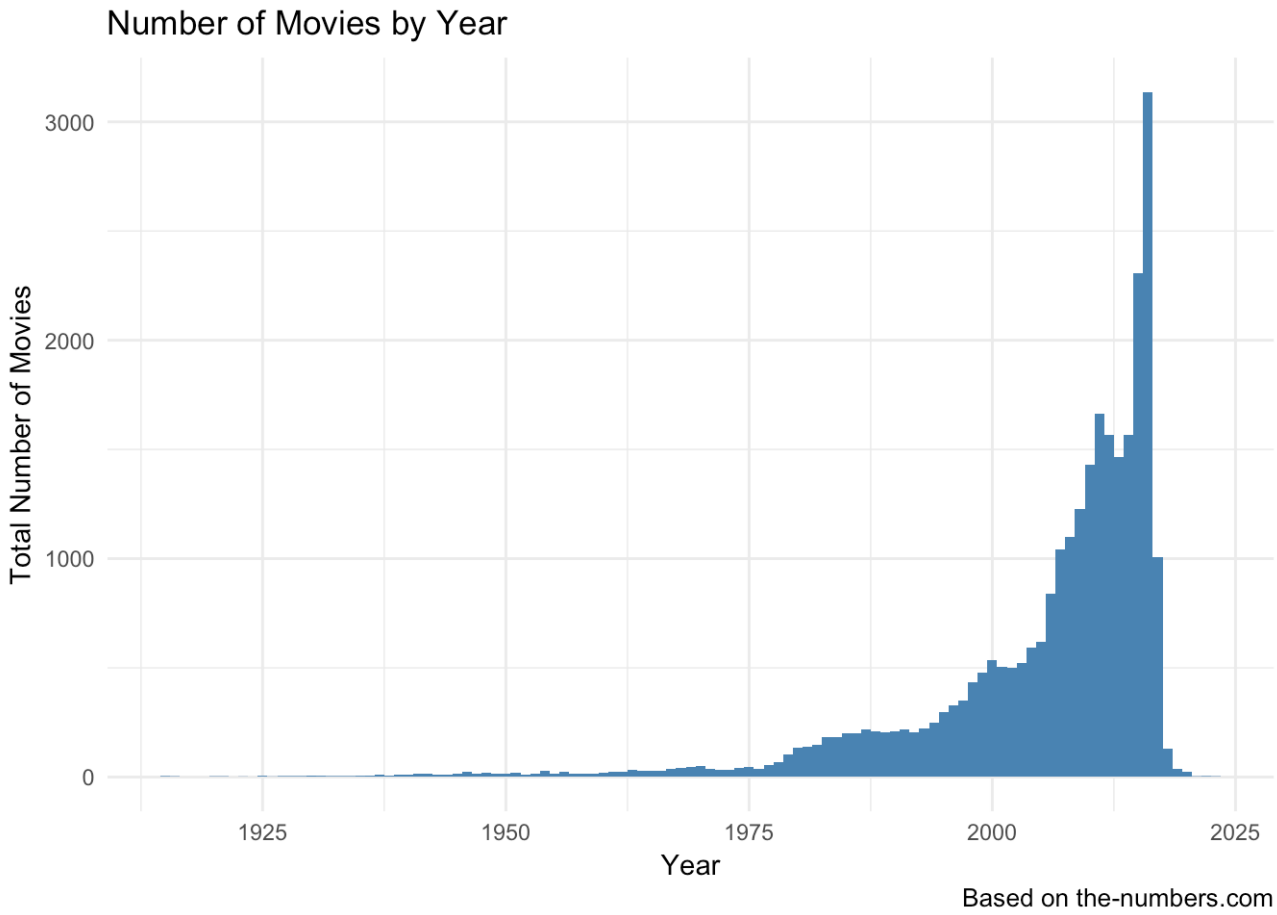
The Goal is to model prediction for the following

1. Number of movies to be made in each individual genre.
2. Production Budget Trend
3. Domestic Box-Office Collection Trend.

Required Libraries

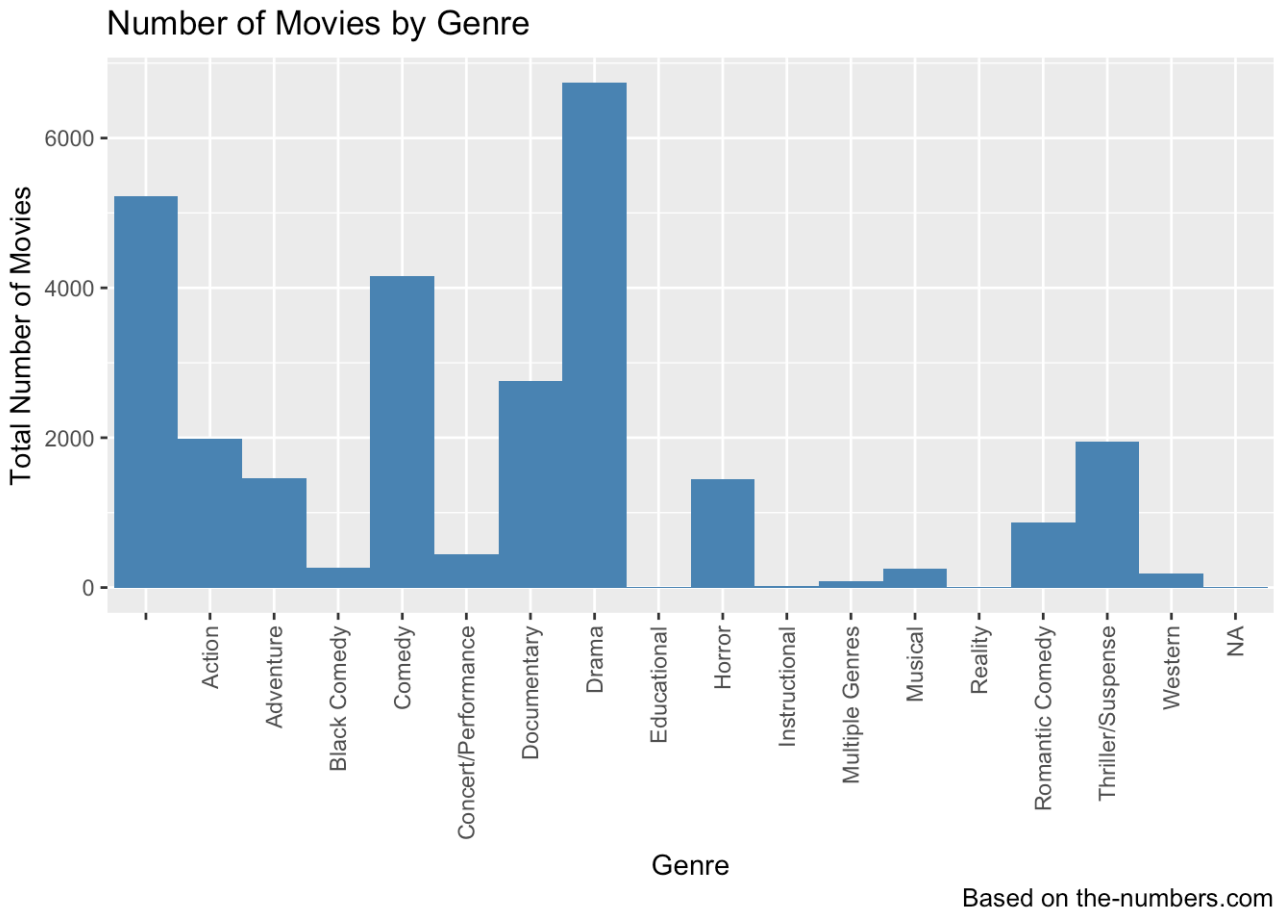
This project is developed using R language. Install dplyr, ggplot2, lattice, devtools, statiscalModeling and rpart using install.packages()

Overall Trend for Number of Movies




```
## # A tibble: 6 × 2
##    Year `n()`
##   <int> <int>
## 1  1915     3
## 2  1916     2
## 3  1920     2
## 4  1921     2
## 5  1923     1
## 6  1925     4
```

Overall Trend of Movies by Genre



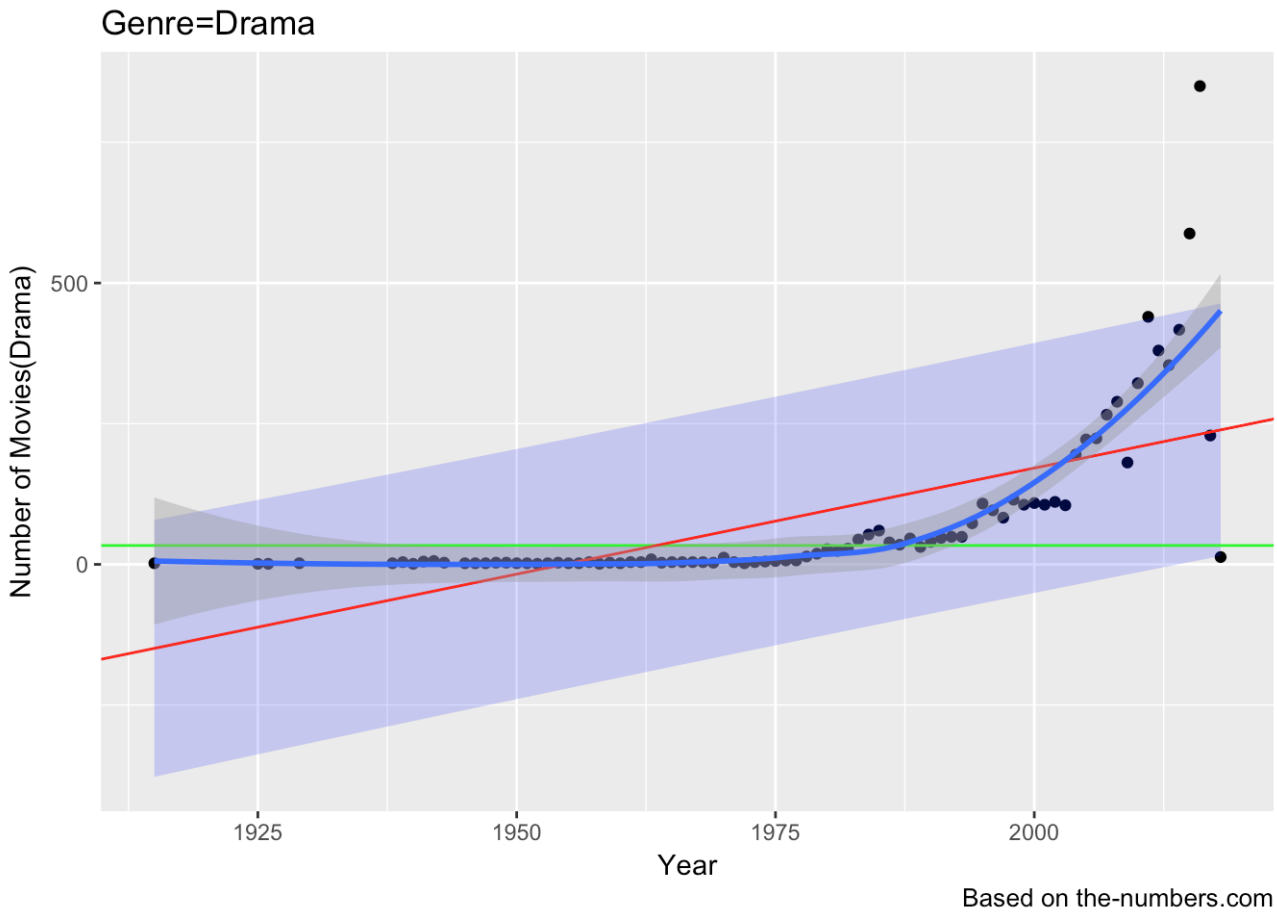

```
## # A tibble: 6 × 2
##                  Genre `n()`
##                 <fctr> <int>
## 1                       5218
## 2               Action  1986
## 3            Adventure  1461
## 4         Black Comedy   268
## 5               Comedy  4160
## 6 Concert/Performance   440
```

Model Prediction Genre="Drama"

```
train_data <- tbl_df(df)
genre<-train_data %>% group_by(Year, Genre) %>% summarise(n())
drama <- subset(genre, Genre=="Drama")

model1 <- lm(drama$`n()`~Year, data=drama)
mean.num_of_movie=mean(genre$`n()`, na.rm=T)

drama.df=data.frame(drama)
mp <- cbind(drama.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=drama$Year, y=drama$`n()`)) +   geom_point(aes(y = drama$`n(
)`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
          slope=model1$coefficients[2], color="red") +
          stat_smooth(method="loess", formula = y ~ x, size=1) +
          labs(title="Genre=Drama") + labs(x="Year") +
          labs(y="Number of Movies(Drama)") + caption

print(p)
```




```
print(summary(drama))
```

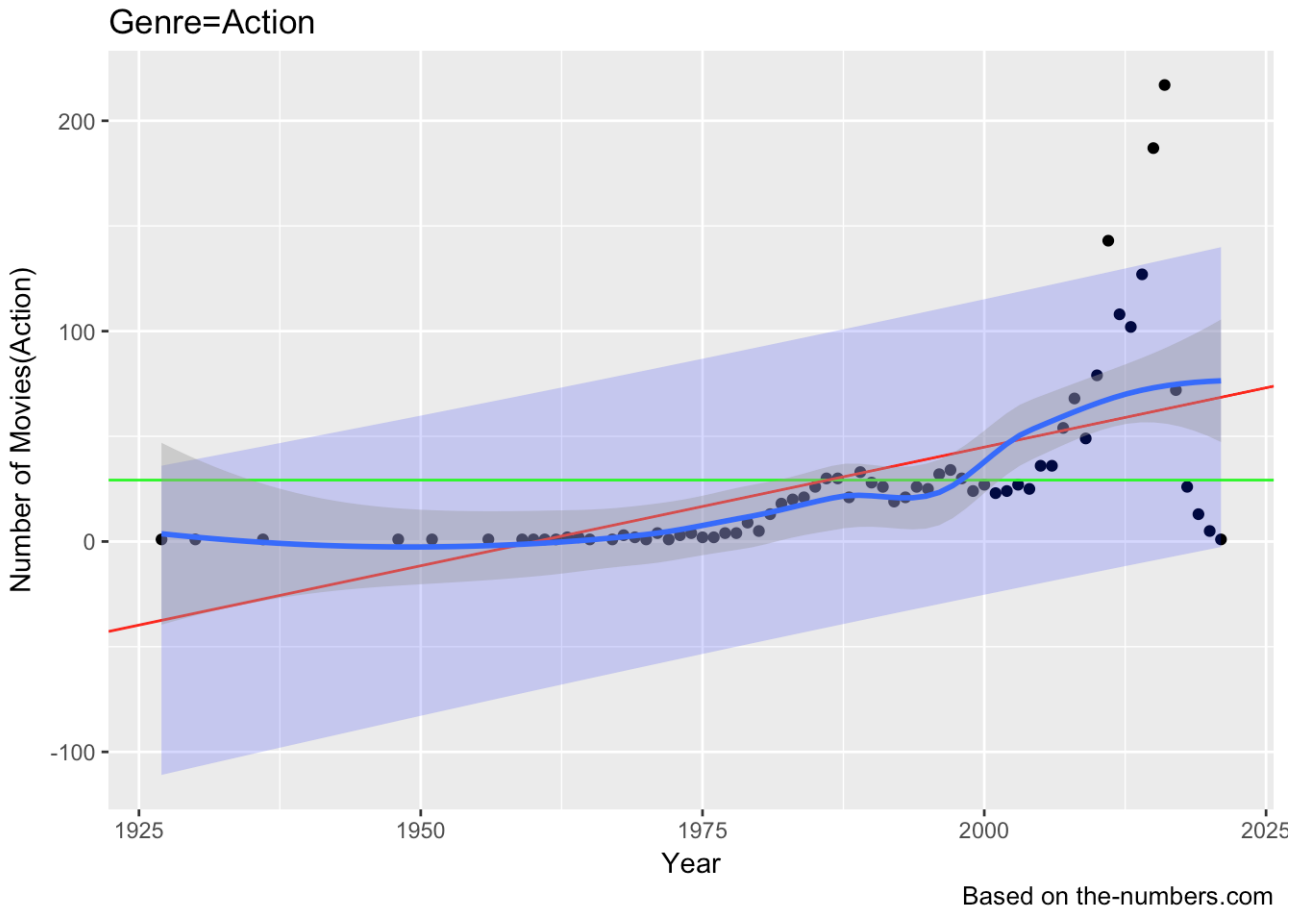
```
##       Year                Genre         n()        
##  Min.   :1915   Drama       :84   Min.   :  1.00  
##  1st Qu.:1956               : 0   1st Qu.:  3.00  
##  Median :1976   Action      : 0   Median : 10.50  
##  Mean   :1976   Adventure   : 0   Mean   : 80.19  
##  3rd Qu.:1997   Black Comedy: 0   3rd Qu.: 98.25  
##  Max.   :2018   Comedy      : 0   Max.   :850.00  
##                 (Other)     : 0
```

Model Prediction Genre="Action"

```
action <- subset(genre, Genre=="Action")
model1 <- lm(action$`n()`~Year, data=action)
mean.num_of_movie=mean(action$`n()`, na.rm=T)

action.df=data.frame(action)
mp <- cbind(action.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=action$Year, y=action$`n()`)) + geom_point(aes(y = action$`n
()`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
          slope=model1$coefficients[2], color="red") +
          stat_smooth(method="loess", formula = y ~ x, size=1) +
          labs(title="Genre=Action") + labs(x="Year") +
          labs(y="Number of Movies(Action)") + caption

print(p)
```




```
print(summary(action))
```

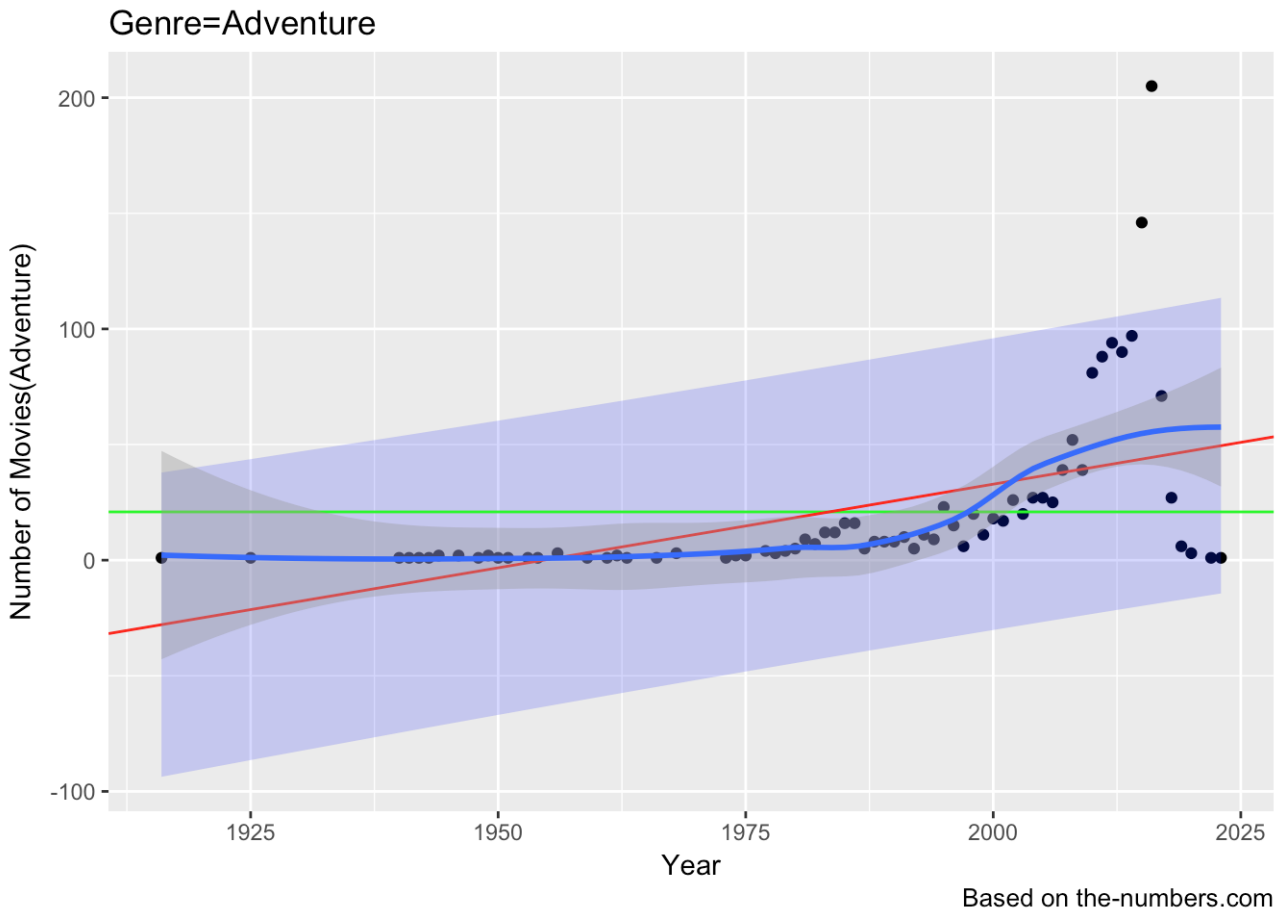
```
##       Year                       Genre         n()        
##  Min.   :1927   Action              :68   Min.   :  1.00  
##  1st Qu.:1971                       : 0   1st Qu.:  2.00  
##  Median :1988   Adventure           : 0   Median : 20.50  
##  Mean   :1986   Black Comedy        : 0   Mean   : 29.21  
##  3rd Qu.:2004   Comedy              : 0   3rd Qu.: 30.00  
##  Max.   :2021   Concert/Performance: 0   Max.   :217.00  
##                 (Other)             : 0
```

Model Prediction Genre="Adventure"

```
adven <- subset(genre, Genre=="Adventure")

model1 <- lm(adven$`n()`~Year, data=adven)
mean.num_of_movie=mean(adven$`n()`, na.rm=T)

adven.df=data.frame(adven)
mp <- cbind(adven.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=adven$Year, y=adven$`n()`)) + geom_point(aes(y = adven$`n()`
)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
           slope=model1$coefficients[2], color="red") +
           stat_smooth(method="loess", formula = y ~ x, size=1) +
           labs(title="Genre=Adventure") + labs(x="Year") +
           labs(y="Number of Movies(Adventure)") + caption

print(p)
```




```
print(summary(adven))
```

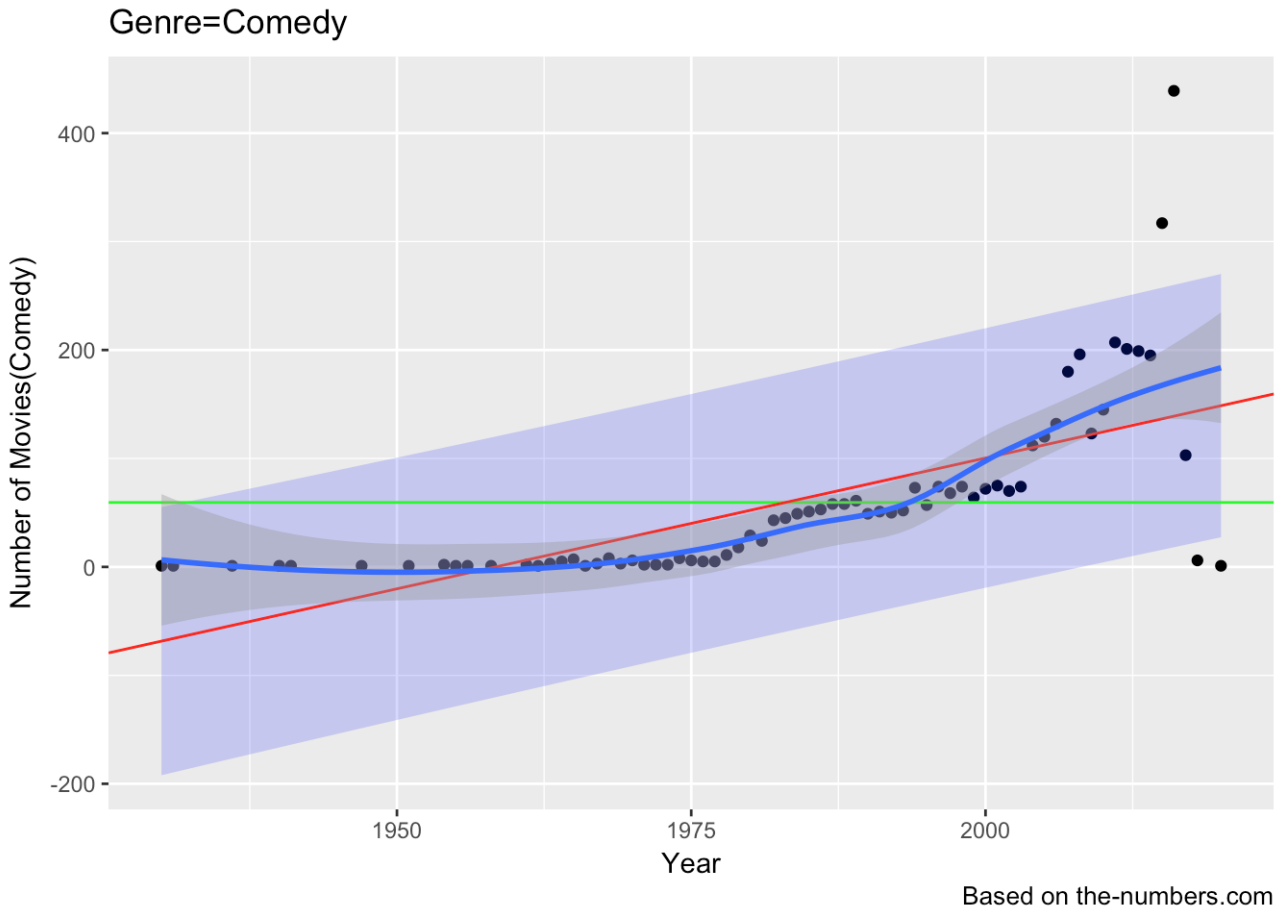
```
##       Year                      Genre         n()        
##  Min.   :1916   Adventure          :70   Min.   :  1.00  
##  1st Qu.:1962                      : 0   1st Qu.:  1.25  
##  Median :1988   Action             : 0   Median :  6.50  
##  Mean   :1983   Black Comedy       : 0   Mean   : 20.87  
##  3rd Qu.:2005   Comedy             : 0   3rd Qu.: 20.00  
##  Max.   :2023   Concert/Performance: 0   Max.   :205.00  
##                 (Other)            : 0
```

Model Prediction Genre="Comedy"

```
comedy <- subset(genre, Genre=="Comedy")

model1 <- lm(comedy$`n()`~Year, data=comedy)
mean.num_of_movie=mean(comedy$`n()`, na.rm=T)

comedy.df=data.frame(comedy)
mp <- cbind(comedy.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=comedy$Year, y=comedy$`n()`)) + geom_point(aes(y = comedy$`n
()`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
          slope=model1$coefficients[2], color="red") +
          stat_smooth(method="loess", formula = y ~ x, size=1) +
          labs(title="Genre=Comedy") + labs(x="Year") +
          labs(y="Number of Movies(Comedy)") + caption

print(p)
```




```
print(summary(comedy))
```

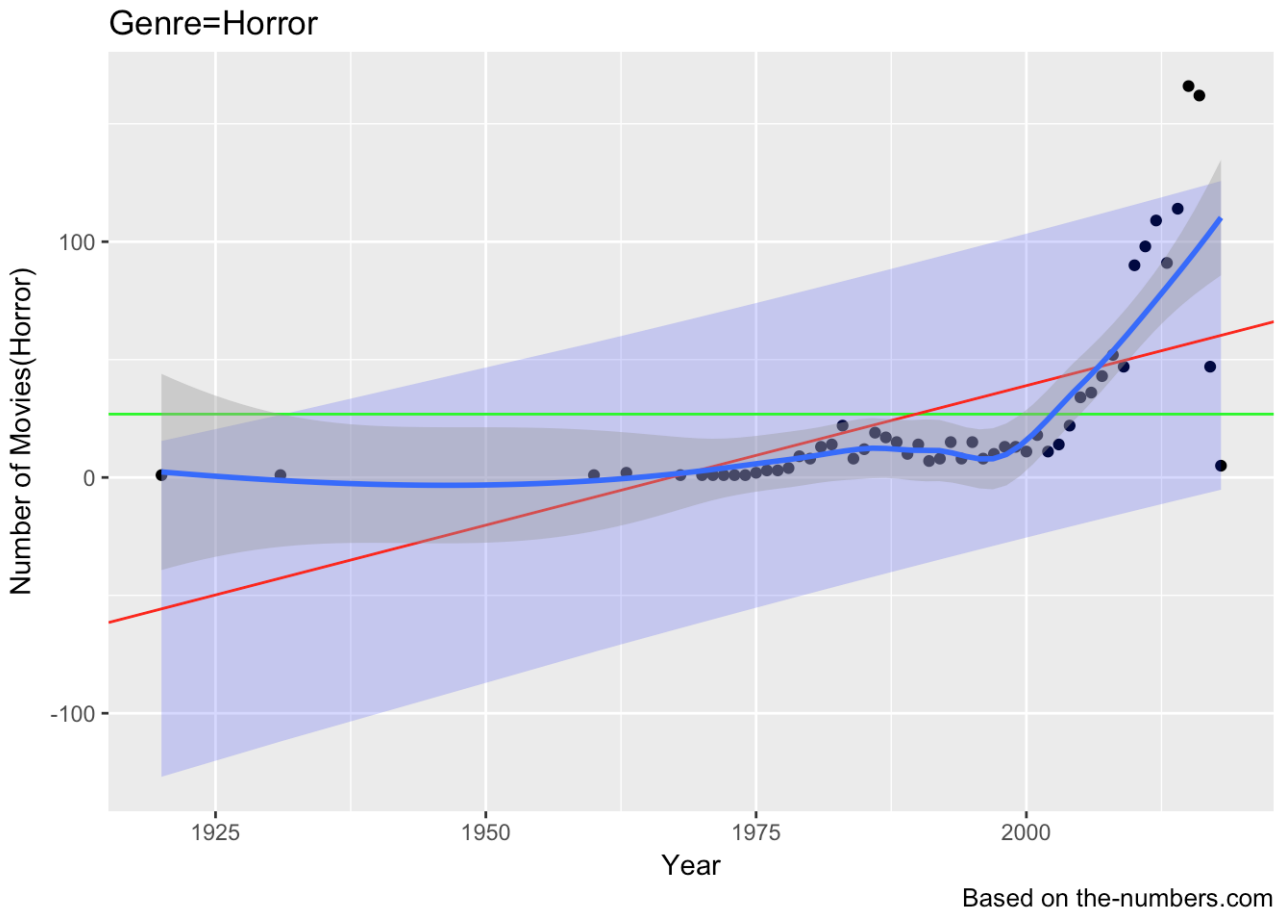
```
##       Year                       Genre         n()        
##  Min.   :1930   Comedy             :70   Min.   :  1.00  
##  1st Qu.:1967                      : 0   1st Qu.:  2.25  
##  Median :1984   Action             : 0   Median : 44.00  
##  Mean   :1983   Adventure          : 0   Mean   : 59.43  
##  3rd Qu.:2002   Black Comedy       : 0   3rd Qu.: 73.75  
##  Max.   :2020   Concert/Performance: 0   Max.   :439.00  
##                 (Other)            : 0
```

Model Prediction Genre="Horror"

```
horror <- subset(genre, Genre=="Horror")

model1 <- lm(horror$`n()`~Year, data=horror)
mean.num_of_movie=mean(horror$`n()`, na.rm=T)

horror.df=data.frame(horror)
mp <- cbind(horror.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=horror$Year, y=horror$`n()`)) + geom_point(aes(y = horror$`n
()`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
           slope=model1$coefficients[2], color="red") +
           stat_smooth(method="loess", formula = y ~ x, size=1) +
           labs(title="Genre=Horror") + labs(x="Year") +
           labs(y="Number of Movies(Horror)") + caption

print(p)
```




```
print(summary(horror))
```

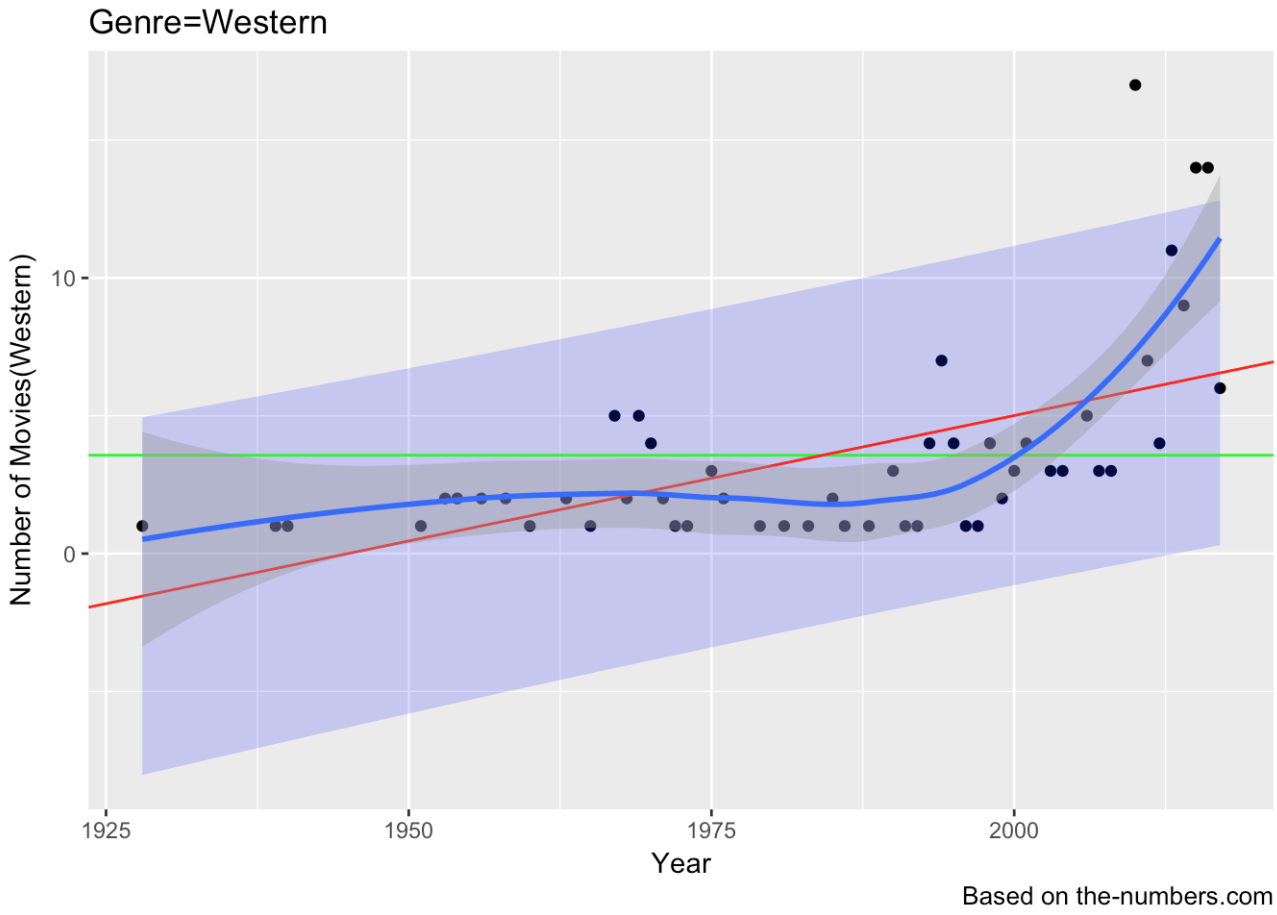
```
##       Year                Genre         n()        
##  Min.   :1920   Horror      :54   Min.   :  1.00  
##  1st Qu.:1978               : 0   1st Qu.:  4.25  
##  Median :1992   Action      : 0   Median : 12.50  
##  Mean   :1990   Adventure   : 0   Mean   : 26.87  
##  3rd Qu.:2005   Black Comedy: 0   3rd Qu.: 22.00  
##  Max.   :2018   Comedy      : 0   Max.   :166.00  
##                 (Other)     : 0
```

Model Prediction Genre="Western"

```
western <- subset(genre, Genre=="Western")

model1 <- lm(western$`n()`~Year, data=western)
mean.num_of_movie=mean(western$`n()`, na.rm=T)

western.df=data.frame(western)
mp <- cbind(western.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=western$Year, y=western$`n()`)) + geom_point(aes(y = western
$`n()`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
          slope=model1$coefficients[2], color="red") +
          stat_smooth(method="loess", formula = y ~ x, size=1) +
          labs(title="Genre=Western") + labs(x="Year") +
          labs(y="Number of Movies(Western)") + caption

print(p)
```




```
print(summary(western))
```

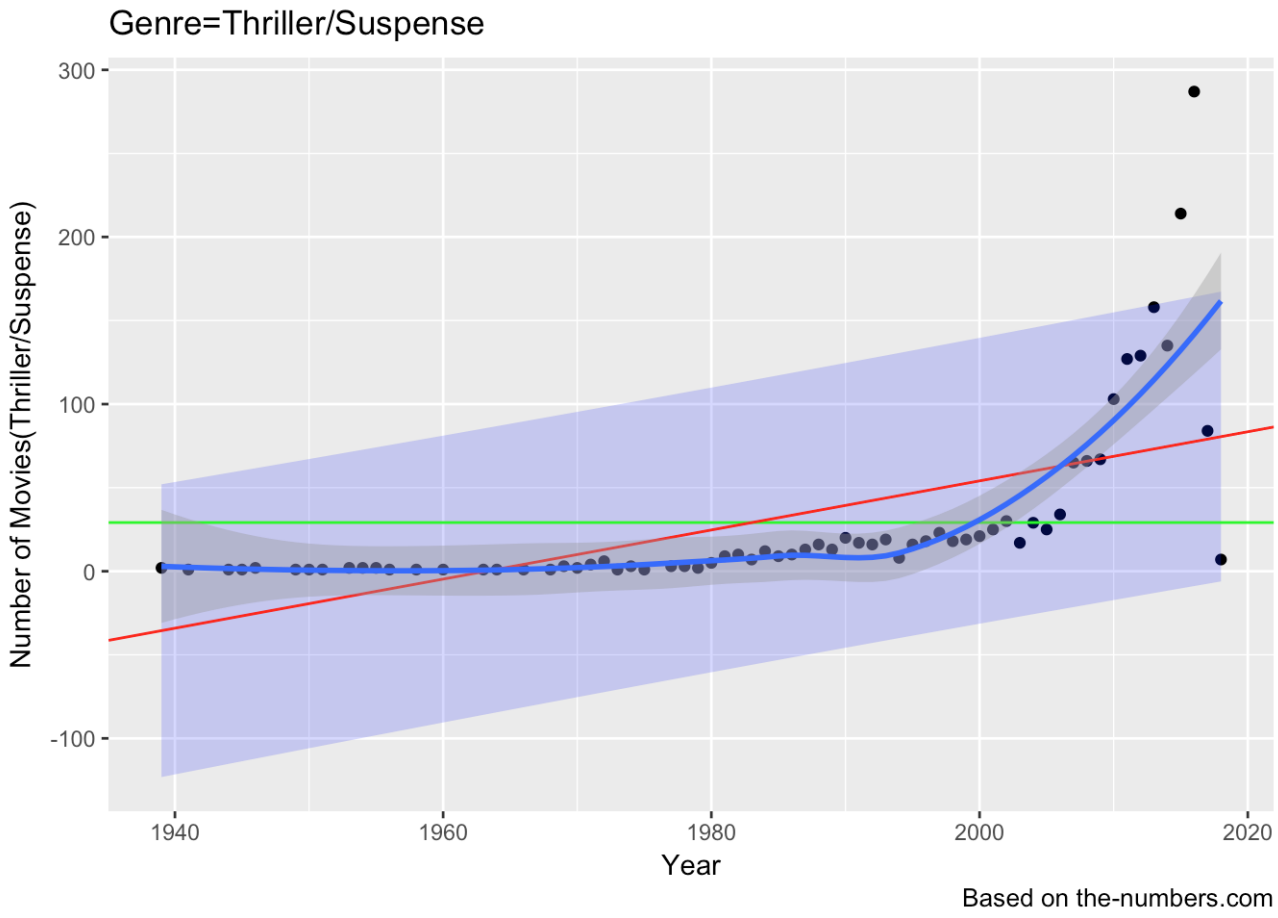
```
##       Year               Genre         n()        
##  Min.   :1928   Western     :51   Min.   : 1.000  
##  1st Qu.:1968               : 0   1st Qu.: 1.000  
##  Median :1988   Action      : 0   Median : 2.000  
##  Mean   :1984   Adventure   : 0   Mean   : 3.569  
##  3rd Qu.:2002   Black Comedy: 0   3rd Qu.: 4.000  
##  Max.   :2017   Comedy      : 0   Max.   :17.000  
##                 (Other)     : 0
```

Model Prediction Genre="Thriller/Suspense"

```
thriller <- subset(genre, Genre=="Thriller/Suspense")

model1 <- lm(thriller$`n()`~Year, data=thriller)
mean.num_of_movie=mean(thriller$`n()`, na.rm=T)

thriller.df=data.frame(thriller)
mp <- cbind(thriller.df, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

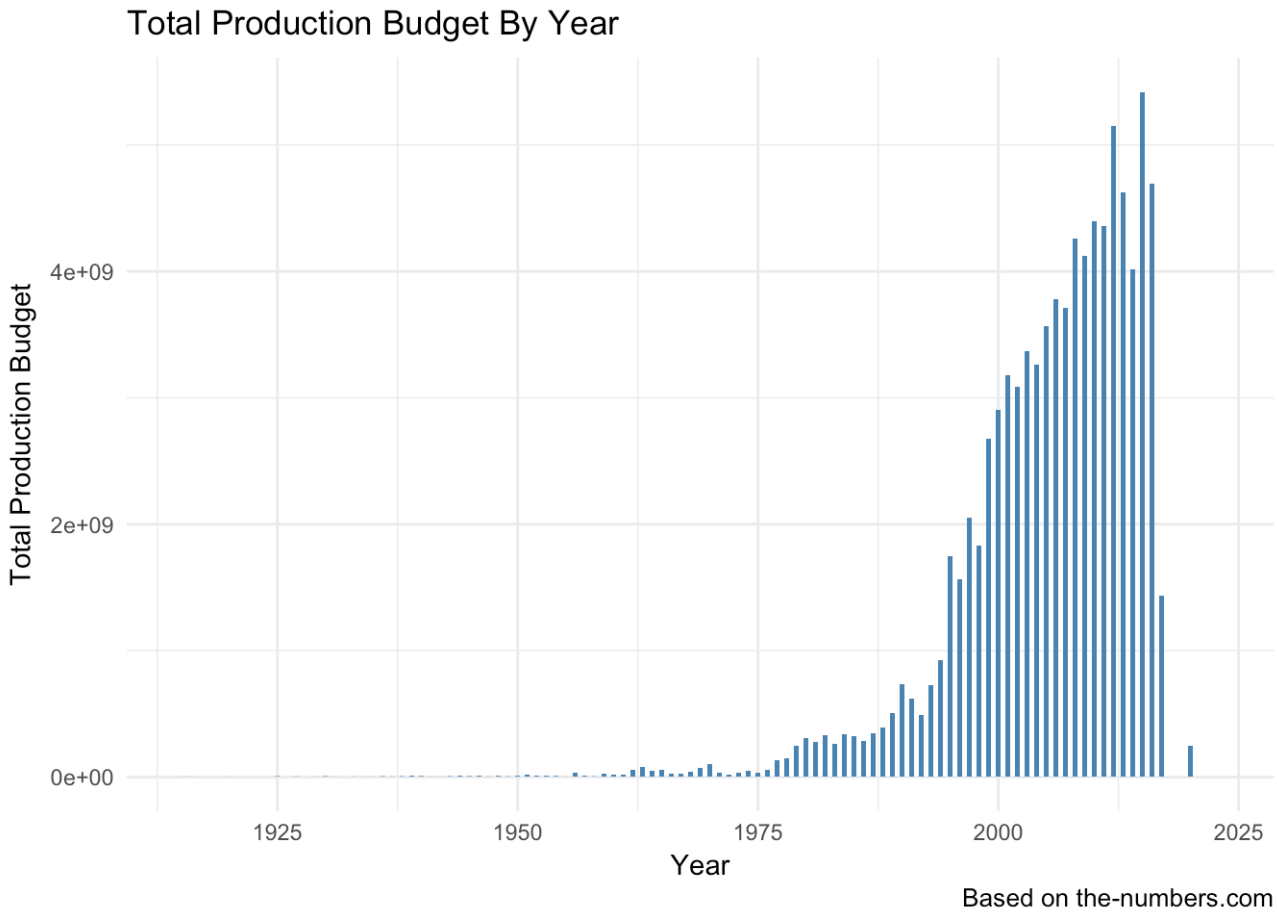
```
p<-ggplot(mp, aes(x=thriller$Year, y=thriller$`n()`)) + geom_point(aes(y = thril
ler$`n()`)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.num_of_movie, color="green") +
  geom_abline(intercept=model1$coefficients[1],
           slope=model1$coefficients[2], color="red") +
           stat_smooth(method="loess", formula = y ~ x, size=1) +
           labs(title="Genre=Thriller/Suspense") + labs(x="Year") +
           labs(y="Number of Movies(Thriller/Suspense)") + caption

print(p)
```




```
print(summary(thriller))
```

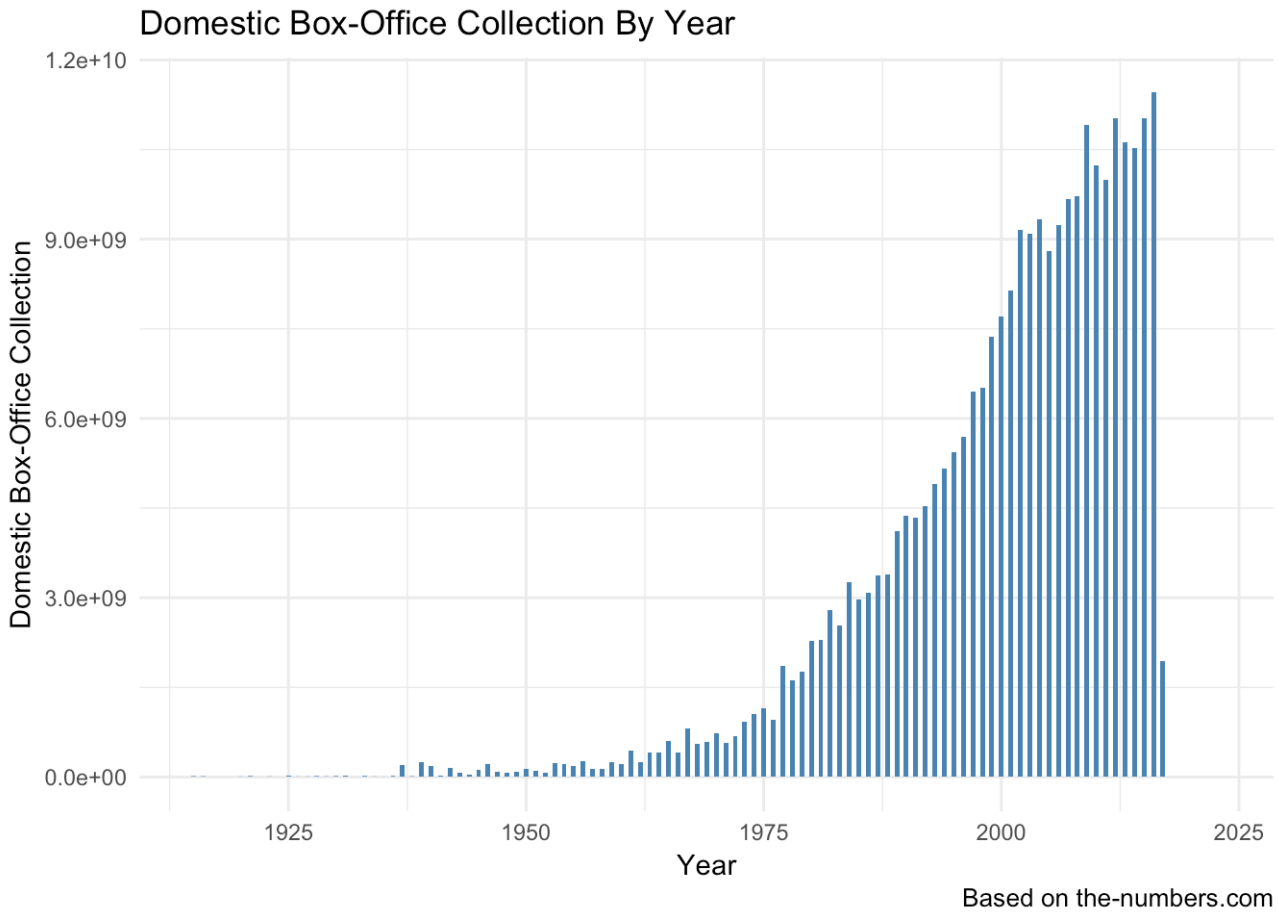
```
##       Year                    Genre         n()        
##  Min.   :1939   Thriller/Suspense:67   Min.   :  1.00  
##  1st Qu.:1967                    : 0   1st Qu.:  2.00  
##  Median :1985   Action           : 0   Median :  9.00  
##  Mean   :1983   Adventure        : 0   Mean   : 29.13  
##  3rd Qu.:2002   Black Comedy     : 0   3rd Qu.: 22.00  
##  Max.   :2018   Comedy           : 0   Max.   :287.00  
##                 (Other)          : 0
```

Production Budget Trend




```
##       Year      ProductionBudget        n()
##  Min.   :1915   Min.   :1.10e+03   Min.   :   1.00
##  1st Qu.:1995   1st Qu.:3.25e+06   1st Qu.:   1.00
##  Median :2004   Median :1.40e+07   Median :   1.00
##  Mean   :1999   Mean   :3.28e+07   Mean   :  10.02
##  3rd Qu.:2010   3rd Qu.:4.20e+07   3rd Qu.:   2.00
##  Max.   :2023   Max.   :4.25e+08   Max.   :2953.00
##                 NA's   :102
```

Domestic Box-Office Collection Trend



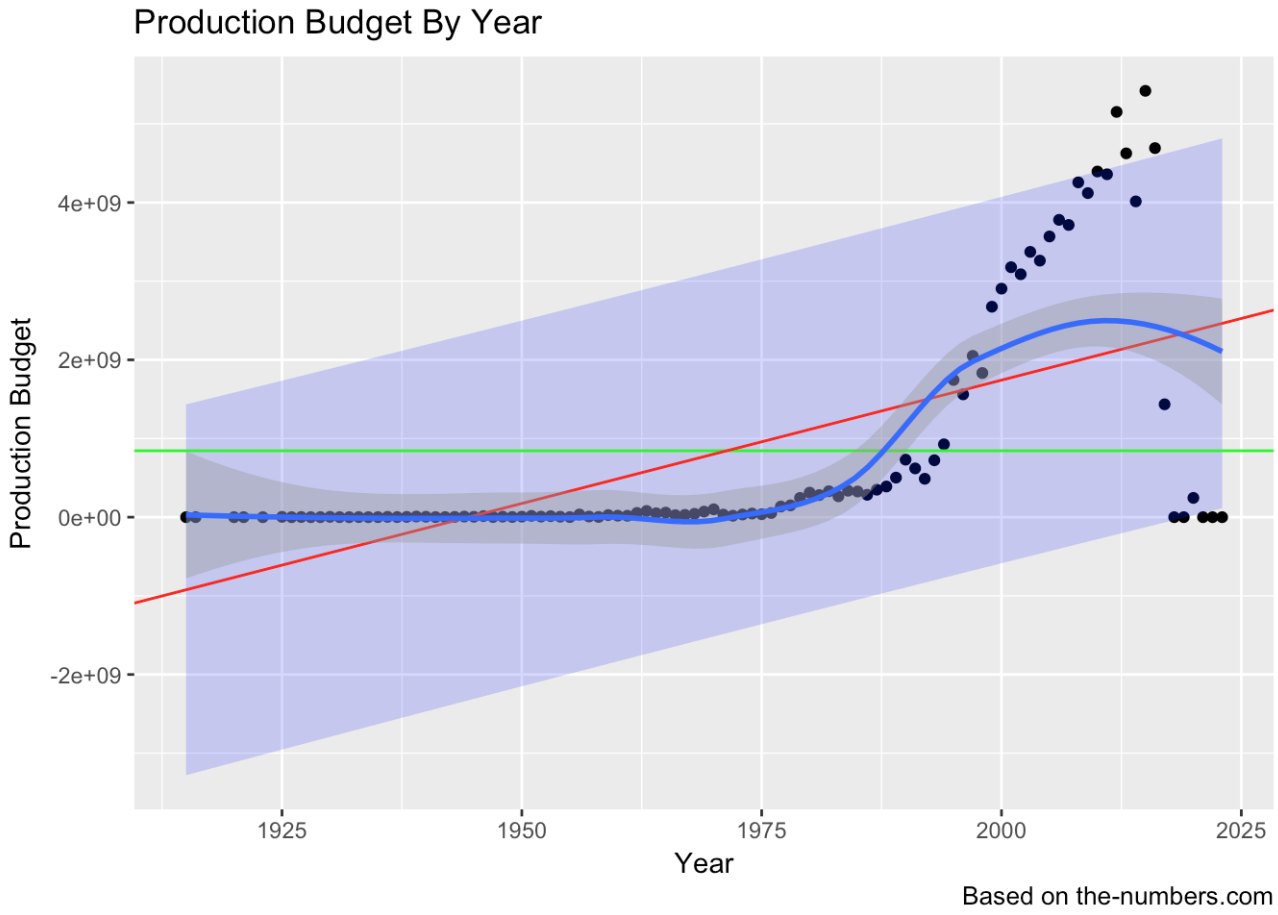

```
##       Year            x
##  Min.   :1915   Min.   :0.000e+00
##  1st Qu.:1946   1st Qu.:5.965e+07
##  Median :1972   Median :5.020e+08
##  Mean   :1971   Mean   :2.623e+09
##  3rd Qu.:1997   3rd Qu.:4.354e+09
##  Max.   :2023   Max.   :1.146e+10
```

Model Prediction for Production Budget

```
mean.pbudget<-mean(aggrP$x, na.rm=T)
model1 <- lm(aggrP$x~Year, data=aggrP)

mp <- cbind(aggrP, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=aggrP$Year, y=aggrP$x)) + geom_point(aes(y = aggrP$x)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.pbudget, color="green") +
  geom_abline(intercept=model1$coefficients[1], slope=model1$coefficients[2], co
lor="red") +
  stat_smooth(method="loess", formula = y ~ x, size=1) +
  labs(title="Production Budget By Year") + labs(x="Year") + labs(y="Production
Budget") + caption

print(p)
```




```
print(summary(aggrP))
```

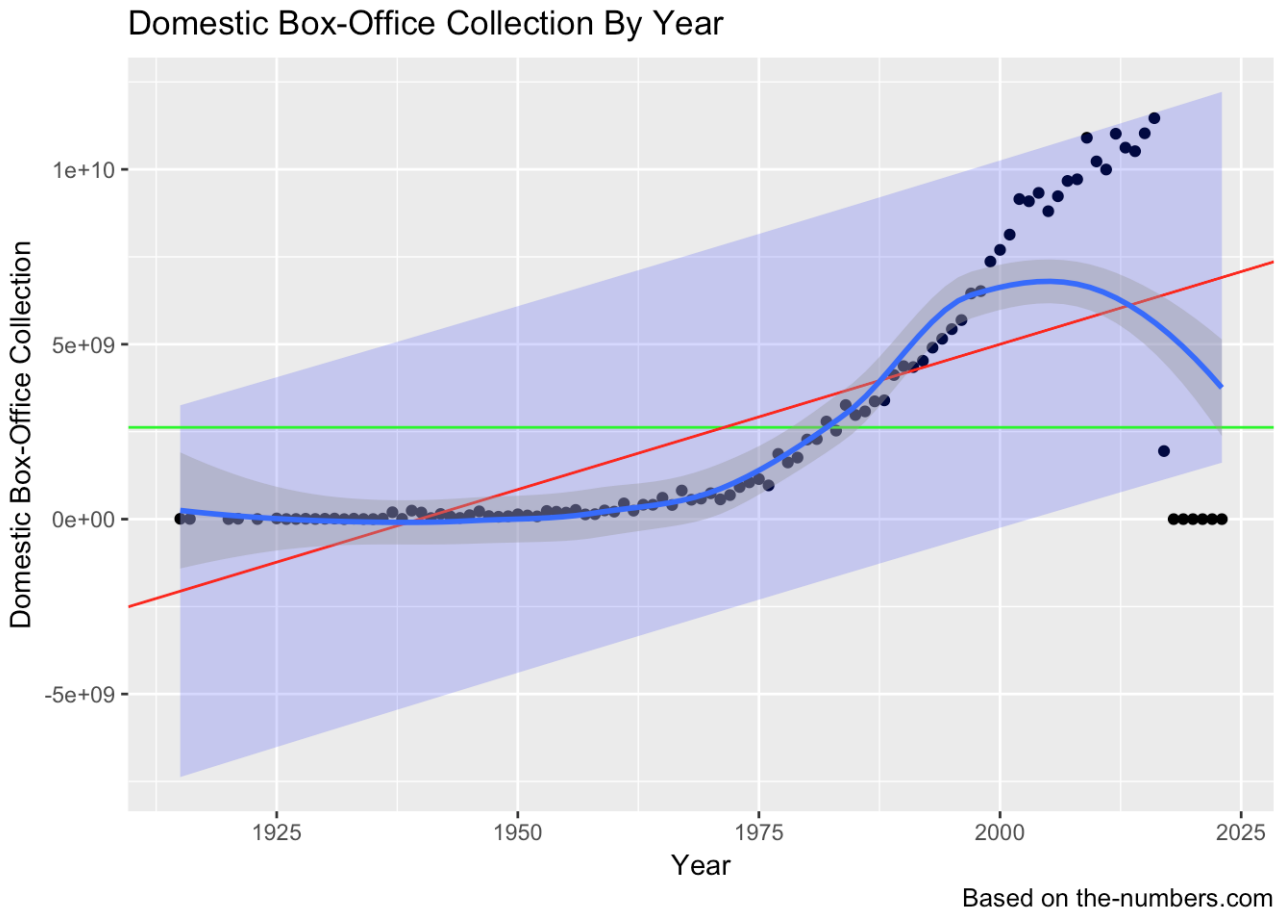
```
##       Year            x
##  Min.   :1915   Min.   :0.000e+00
##  1st Qu.:1946   1st Qu.:3.483e+06
##  Median :1972   Median :3.569e+07
##  Mean   :1971   Mean   :8.443e+08
##  3rd Qu.:1997   3rd Qu.:6.459e+08
##  Max.   :2023   Max.   :5.420e+09
```

Model Prediction for Domestic Box-Office Collection

```
mean.pbudget<-mean(aggrD$x, na.rm=T)
model1 <- lm(aggrD$x~Year, data=aggrD)

#horror.df=data.frame(horror)
mp <- cbind(aggrD, predict(model1, interval = "prediction"))
```

```
## Warning in predict.lm(model1, interval = "prediction"): predictions on curren
t data refer to _future_ responses
```

```
p<-ggplot(mp, aes(x=aggrD$Year, y=aggrD$x)) + geom_point(y=aggrD$x) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +
  geom_hline(yintercept=mean.pbudget, color="green") +
  geom_abline(intercept=model1$coefficients[1], slope=model1$coefficients[2], co
lor="red") +
  stat_smooth(method="loess", formula = y ~ x, size=1) +
    labs(title="Domestic Box-Office Collection By Year") + labs(x="Year") + labs
(y="Domestic Box-Office Collection") + caption

print(p)
```



Acknowledgement

Many thanks to my mentor Sneha Runwal for the guidance and help. I really appreciate her guidance in navigating introductory data science project.