Predicting Success of a Movie - Data Visualization

Madhusudan

11/2021

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("ggplot2")
```

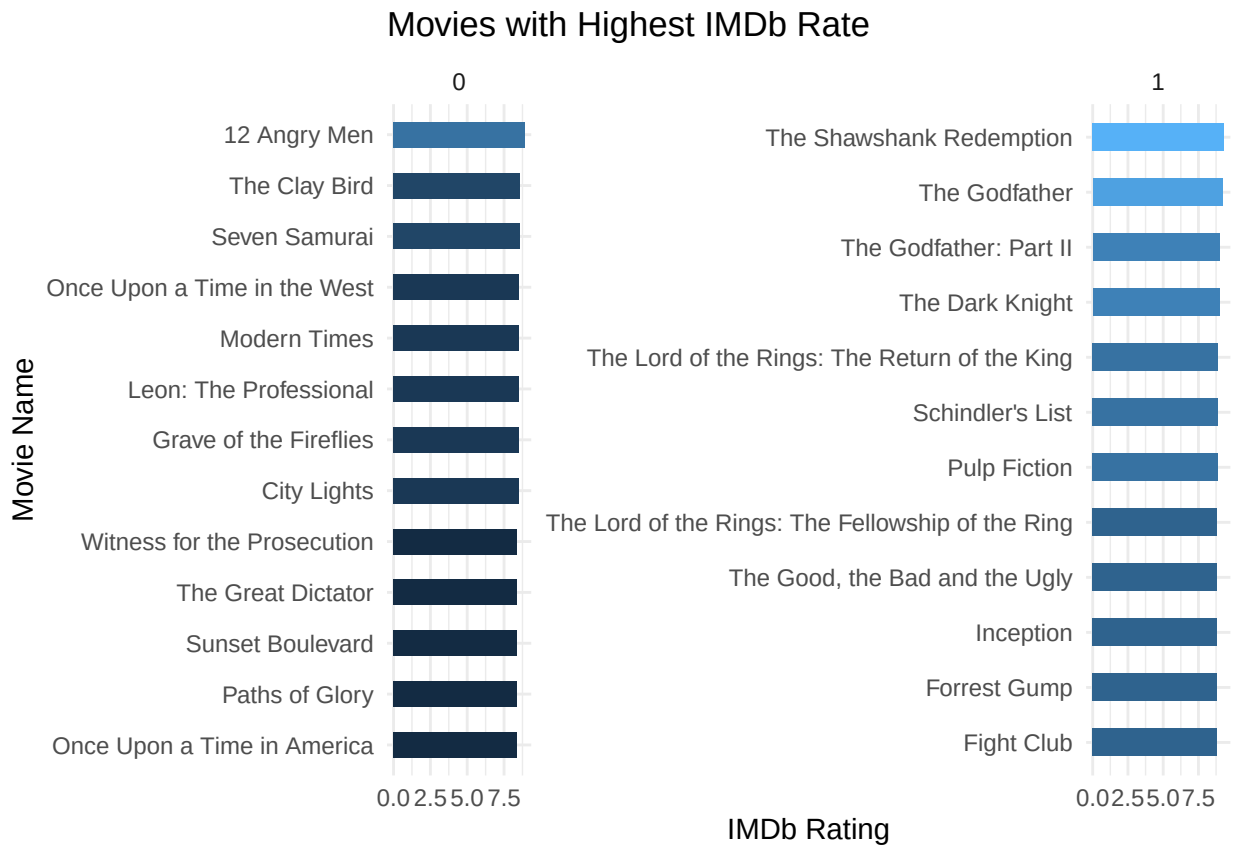
```
#Importing the cleaned data
Cleaned_data <- as_tibble(read_csv("cleaned_movies_database.csv"))
```

```
## Rows: 8992 Columns: 147

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (7): genres, imdb_id, production_countries, original_language, title, ...
## dbl (139): popularity, runtime, vote_count, year, budget, worlwide_gross_inc...
## lgl   (1): adult

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#The top 20 Movies with High IMDb Ratings, For  hit and not hit movies

Cleaned_data %>%
  group_by(hit) %>%
  top_n(10,wt=weighted_average_vote) %>%
  summarise(title, weighted_average_vote,hit) %>%
  arrange(desc(weighted_average_vote)) %>%
  ggplot(aes(x=reorder(title, weighted_average_vote), y=weighted_average_vote, fill=weighted_average_vo
  geom_col(width=0.5 ,show.legend = FALSE)+
```

```
  facet_wrap(~hit,scales = 'free')+
  coord_flip()+
  labs(x="Movie Name", y="IMDb Rating", title = "Movies with Highest IMDb Rate")+
  theme_minimal()
```

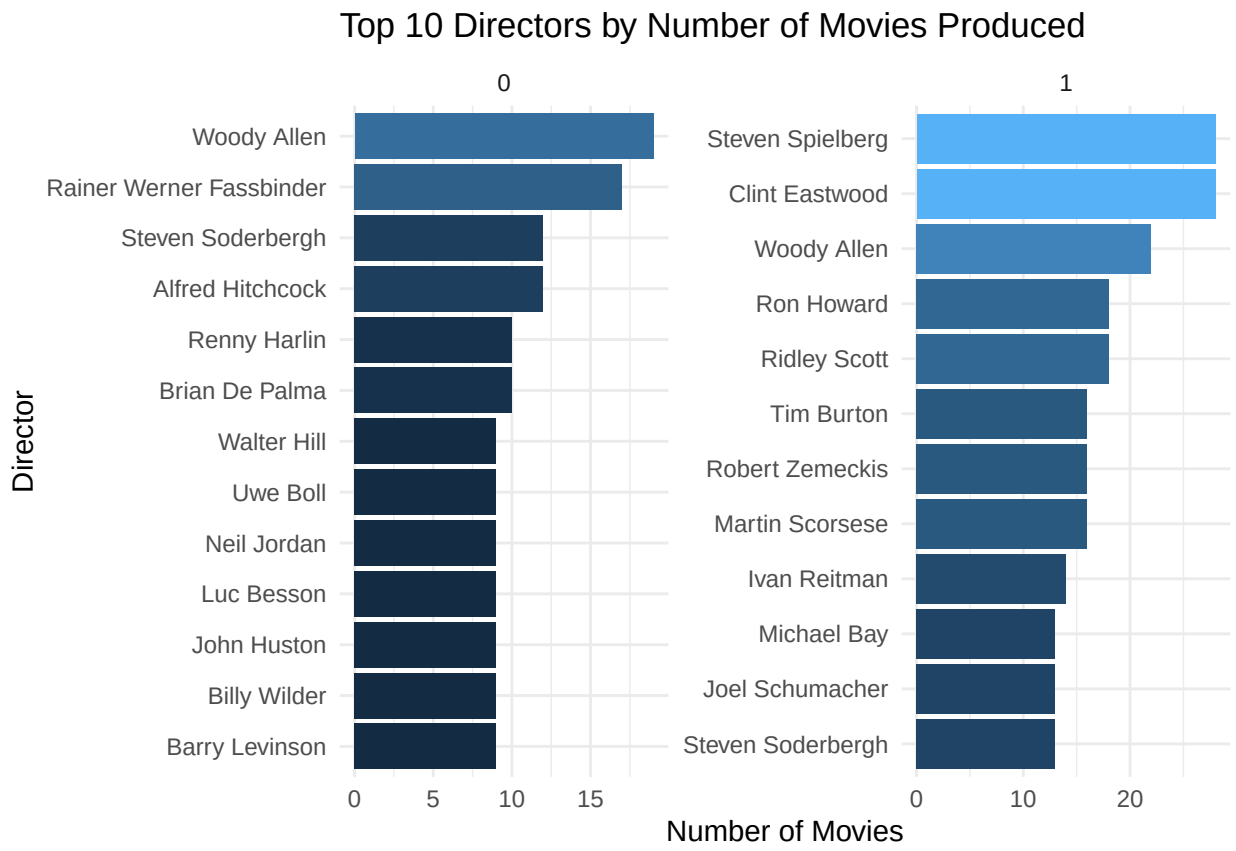
```
## `summarise()` has grouped output by 'hit'. You can override using the `.groups` argument.
```



From the above Figure we can see that the hit movies, have higher IMDb ratings compared to the non-hit movies

```
#Top 10 directors for hit and not hit movies

Cleaned_data %>%
  group_by(hit) %>%
  count(director, sort=TRUE) %>%
  top_n(10)%>%
  ggplot(aes(x=reorder(director, n), y=n, fill=n)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~hit,scales = 'free')+
  labs(x="Director", y="Number of Movies",title="Top 10 Directors by Number of Movies Produced") +
  coord_flip() +
  theme_minimal()
```

```
## Selecting by n
```



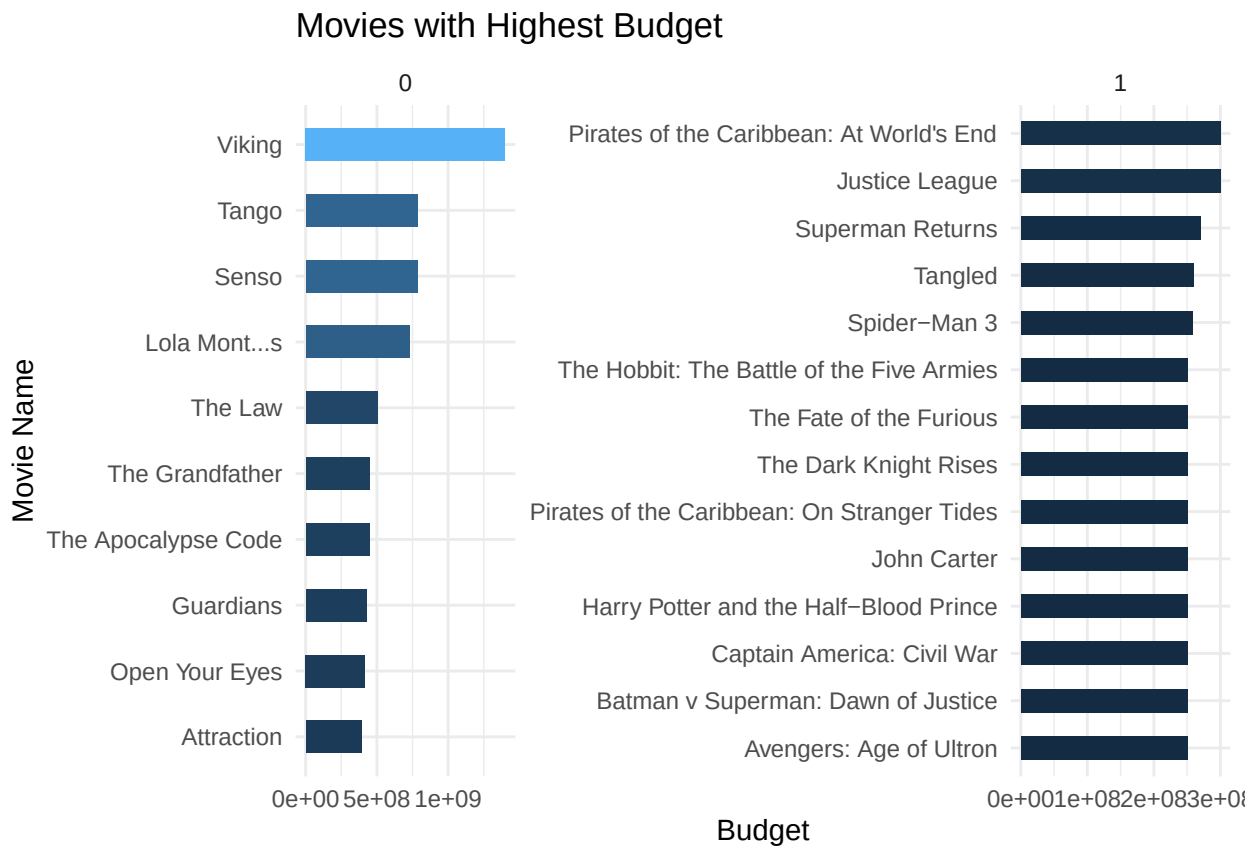

```
#Movies with Highest budget faceted by hit or not hit

Cleaned_data %>%
  mutate(title = str_replace_all(title,"[^[:graph:]]", " ")) %>%
  group_by(hit) %>%
  top_n(10,wt=budget) %>%
  summarise(title, budget,hit) %>%
  arrange(desc(budget)) %>%
  ggplot(aes(x=reorder(title, budget), y=budget, fill=budget))+
  geom_col(width=0.5 ,show.legend = FALSE)+
  facet_wrap(~hit,scales = 'free')+
  coord_flip()+
  labs(x="Movie Name", y="Budget", title = "Movies with Highest Budget") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'hit'. You can override using the `.groups` argument.

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <ef>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bf>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Lola Montï¿½s' in 'mbcsToSbcs': dot substituted for <bd>
```

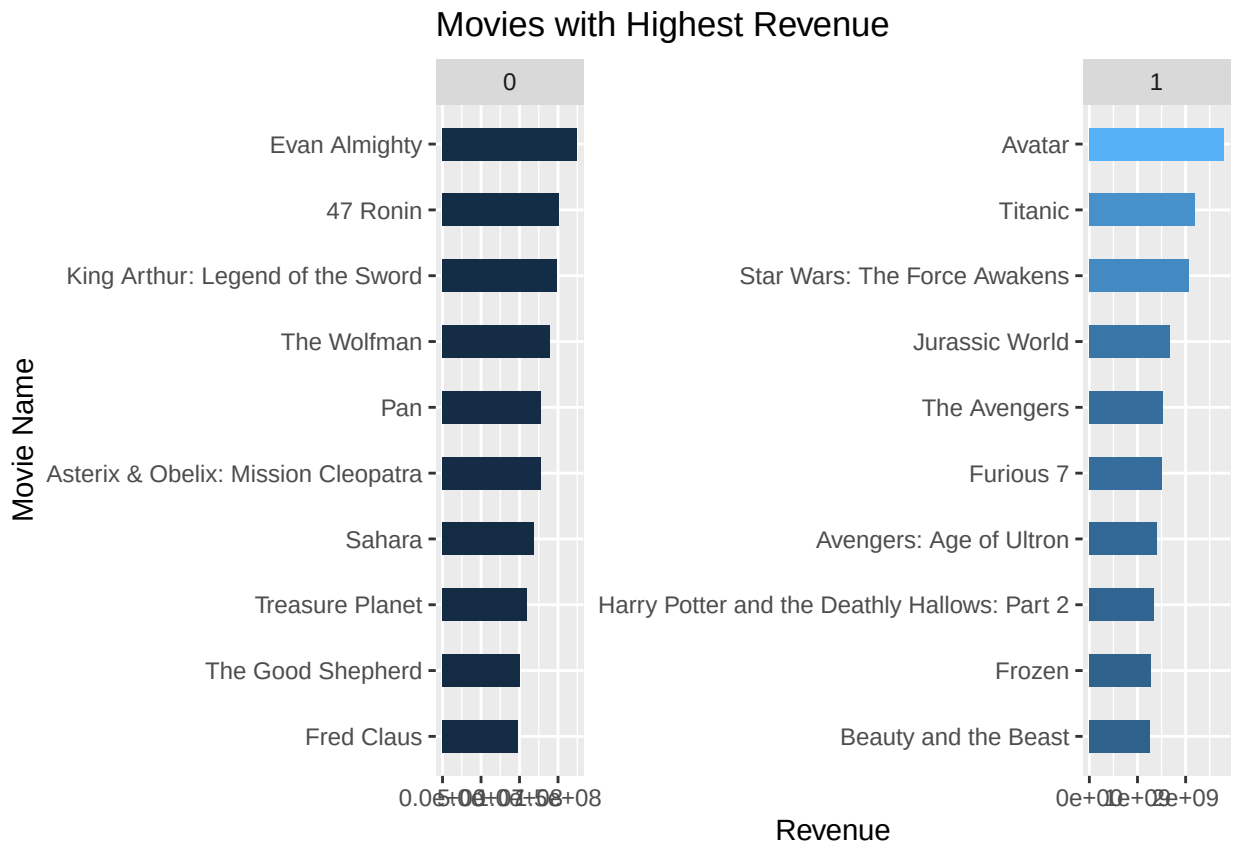


It can be see that the top 10 movies in terms of budget, have higher budget for non-hit movies than hit movies

```
#Movies with Highest Revenues faceted by hit or not hit

Cleaned_data %>%
  mutate(title = str_replace_all(title,"[^[:graph:]]", " ")) %>%
  group_by(hit) %>%
  top_n(10,wt=worlwide_gross_income) %>%
  summarise(title, worlwide_gross_income,hit) %>%
```

```
  arrange(desc(worlwide_gross_income)) %>%
  ggplot(aes(x=reorder(title, worlwide_gross_income), y=worlwide_gross_income, fill=worlwide_gross_incor
  geom_col(width=0.5 ,show.legend = FALSE)+
  facet_wrap(~hit,scales = 'free')+
  coord_flip()+
  labs(x="Movie Name", y="Revenue", title = "Movies with Highest Revenue") +
  theme()
```

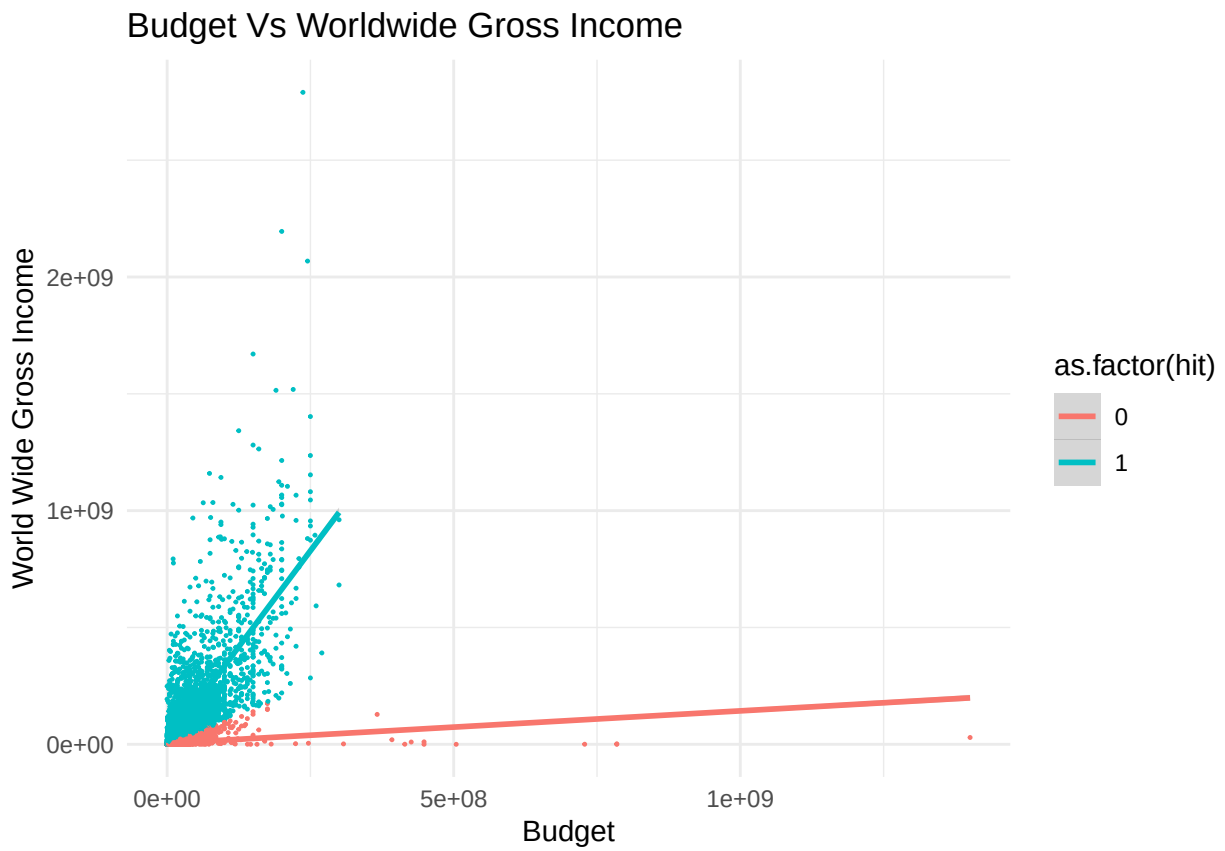
```
## `summarise()` has grouped output by 'hit'. You can override using the `.groups` argument.
```



It can be see that the top 10 movies in terms of revenue, have higher revenue for hit movies than non-hit movies

```
#Scatter plot to see if there is any relationship between budget and worldwide gross income faceted by

Cleaned_data %>%
  ggplot(mapping = aes(x = budget, y = worlwide_gross_income, color = as.factor(hit))) +
  geom_point(size = 0.2) +
  geom_smooth(method = lm) +
  labs(x="Budget", y="World Wide Gross Income", title= "Budget Vs Worldwide Gross Income")+
  scale_fill_discrete(name = "Hit Movie") +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

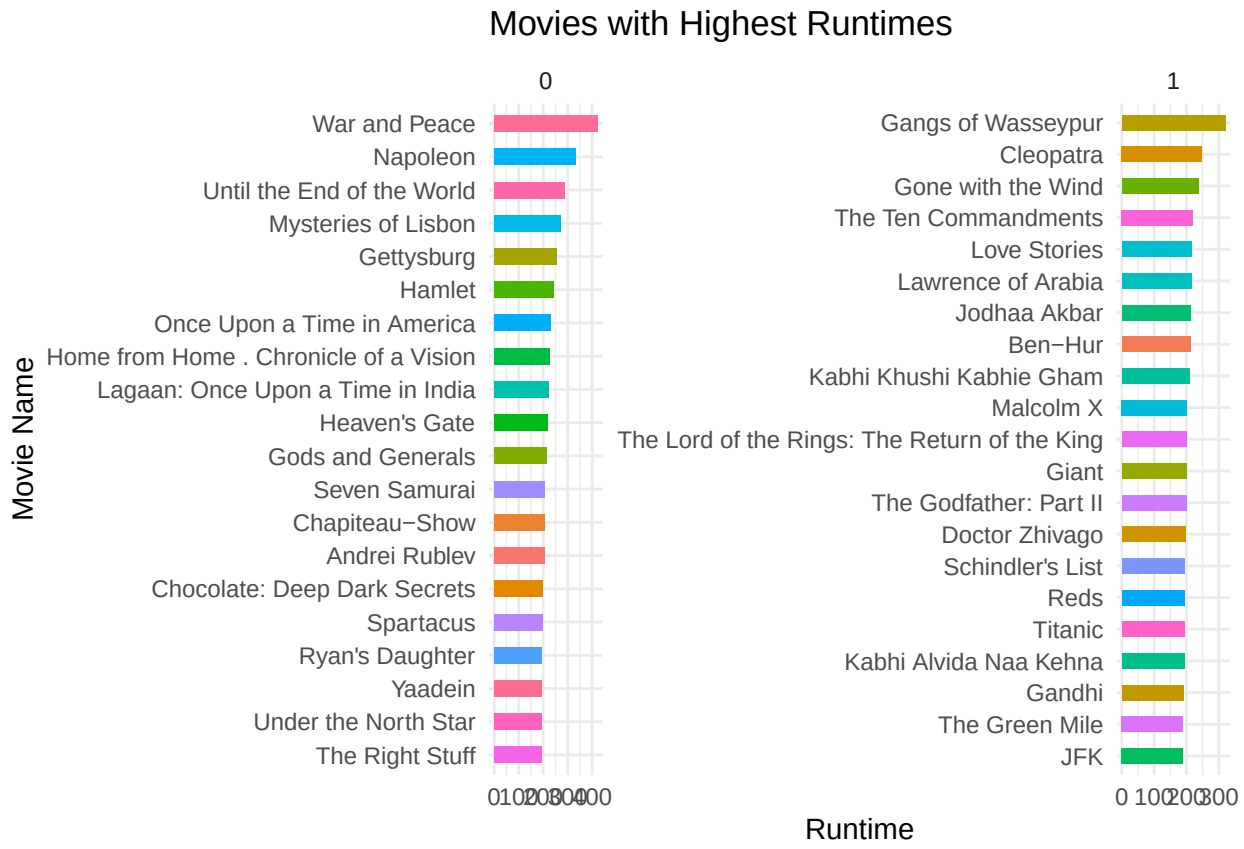


From the above we can see that the revenue and budget both increase for hit movies. However, for non-hit movies, the increasing budget does not increase the revenue after a certain point, which is expected

```
#Movies with highest run times in terms of hit and non hit
Cleaned_data %>%
  group_by(hit)%>%
  top_n(20,wt=runtime) %>%
  summarise(title, runtime, hit) %>%
  arrange(desc(runtime))%>%
  ggplot(aes(x=reorder(title,runtime), y=runtime, fill=title))+
  geom_col(width=0.5 ,show.legend = FALSE)+
  facet_wrap(~hit,scales = 'free')+
  coord_flip()+
  labs(x="Movie Name", y="Runtime", title = "Movies with Highest Runtimes") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'hit'. You can override using the `.groups` argument.

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>
```

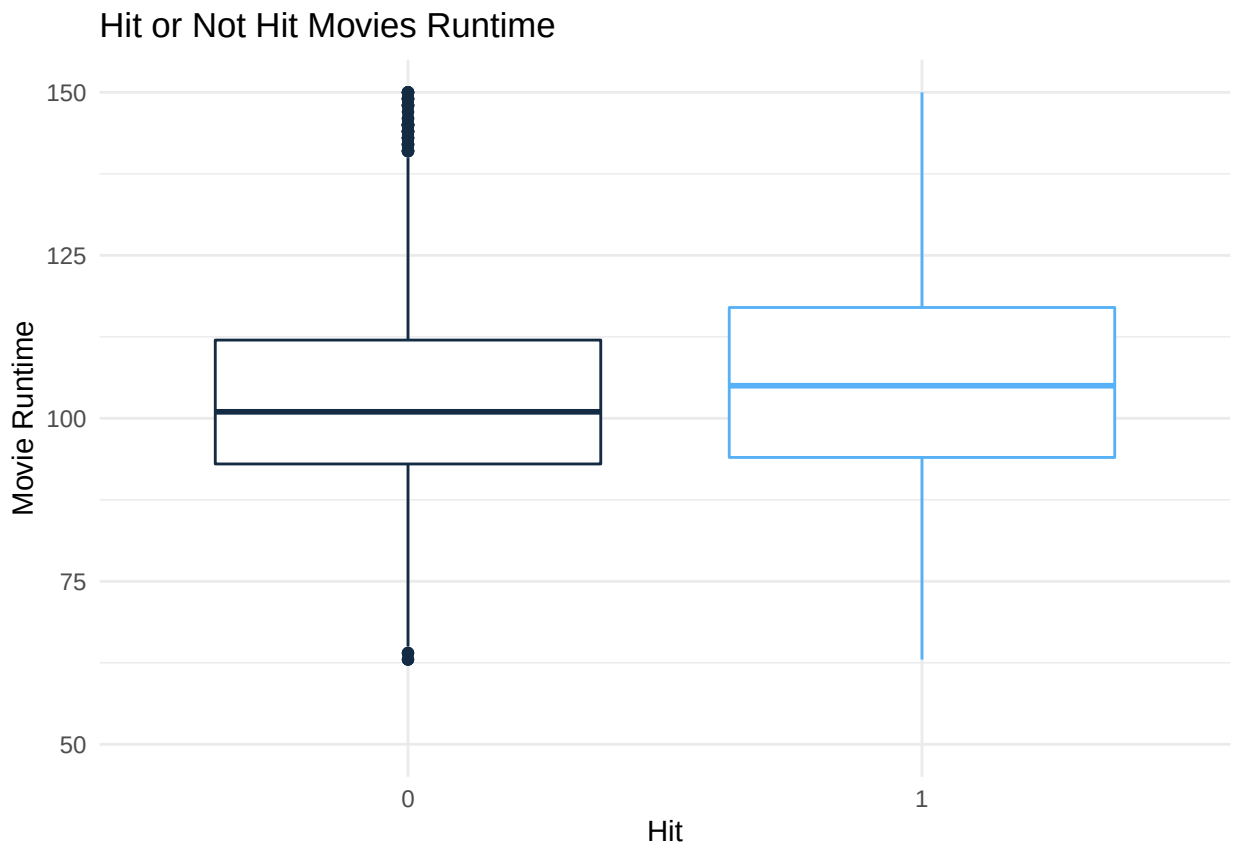
```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Home from Home - Chronicle of a Vision' in 'mbcsToSbcs':
## dot substituted for <96>
```




```
#Run time of hit and not-hit movies

Cleaned_data %>%
  ggplot(mapping = aes(x = as.factor(hit), y=runtime, color= hit)) +
  geom_boxplot(show.legend = "False") +
  labs(y="Movie Runtime",x = "Hit", title= "Hit or Not Hit Movies Runtime")+
  ylim(50,150)+
  theme_minimal()
```

```
## Warning: Removed 366 rows containing non-finite values (stat_boxplot).
```

```
## Warning: `show.legend` must be a logical vector.
```

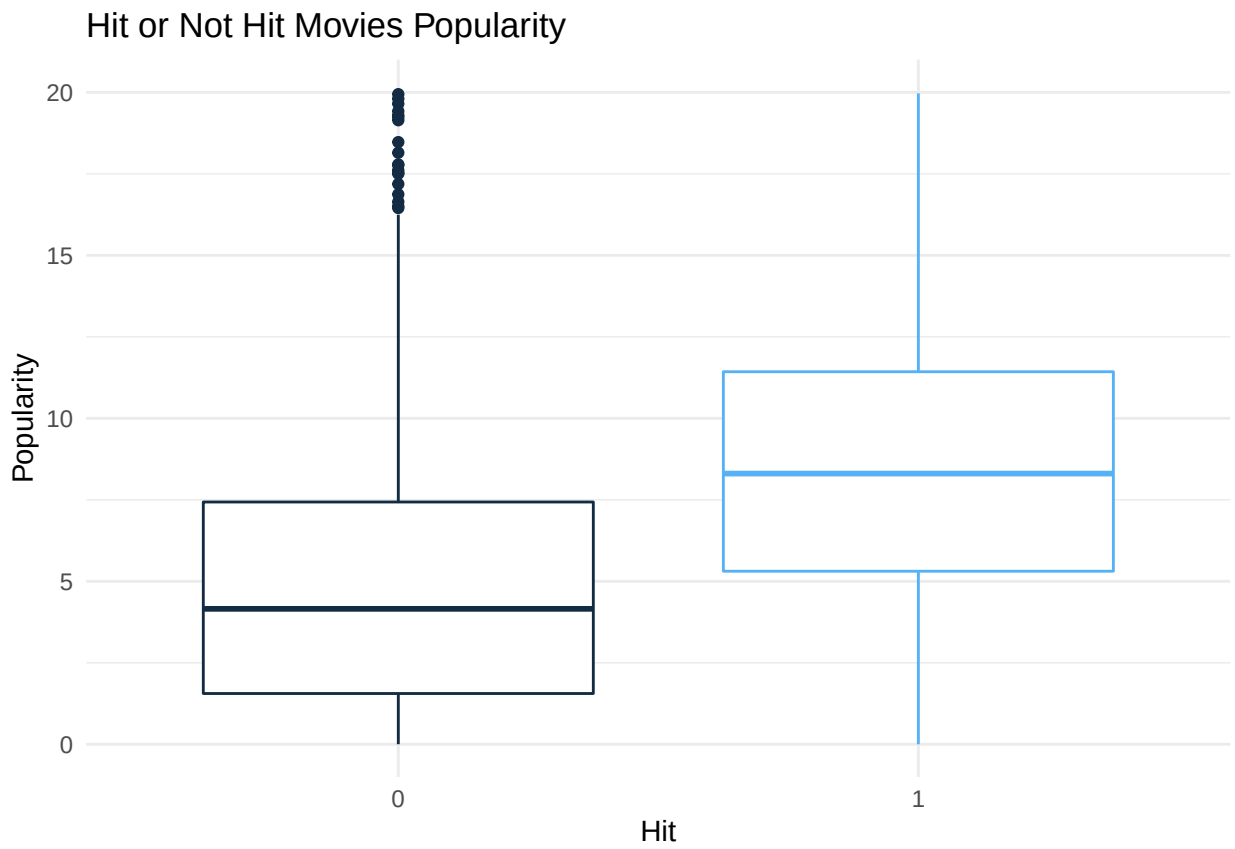


From the above boxplots we can see that the run-times are mostly similar for both hit and non-hit movies

```
#Popularity Score of hit and not-hit movies

Cleaned_data %>%
  ggplot(mapping = aes(x = as.factor(hit), y=popularity, color= hit)) +
  geom_boxplot(show.legend = "False") +
  labs(y="Popularity",x = "Hit", title= "Hit or Not Hit Movies Popularity")+
  ylim(0,20)+
  theme_minimal()
```

```
## Warning: Removed 248 rows containing non-finite values (stat_boxplot).
```

```
## Warning: ‘show.legend‘ must be a logical vector.
```

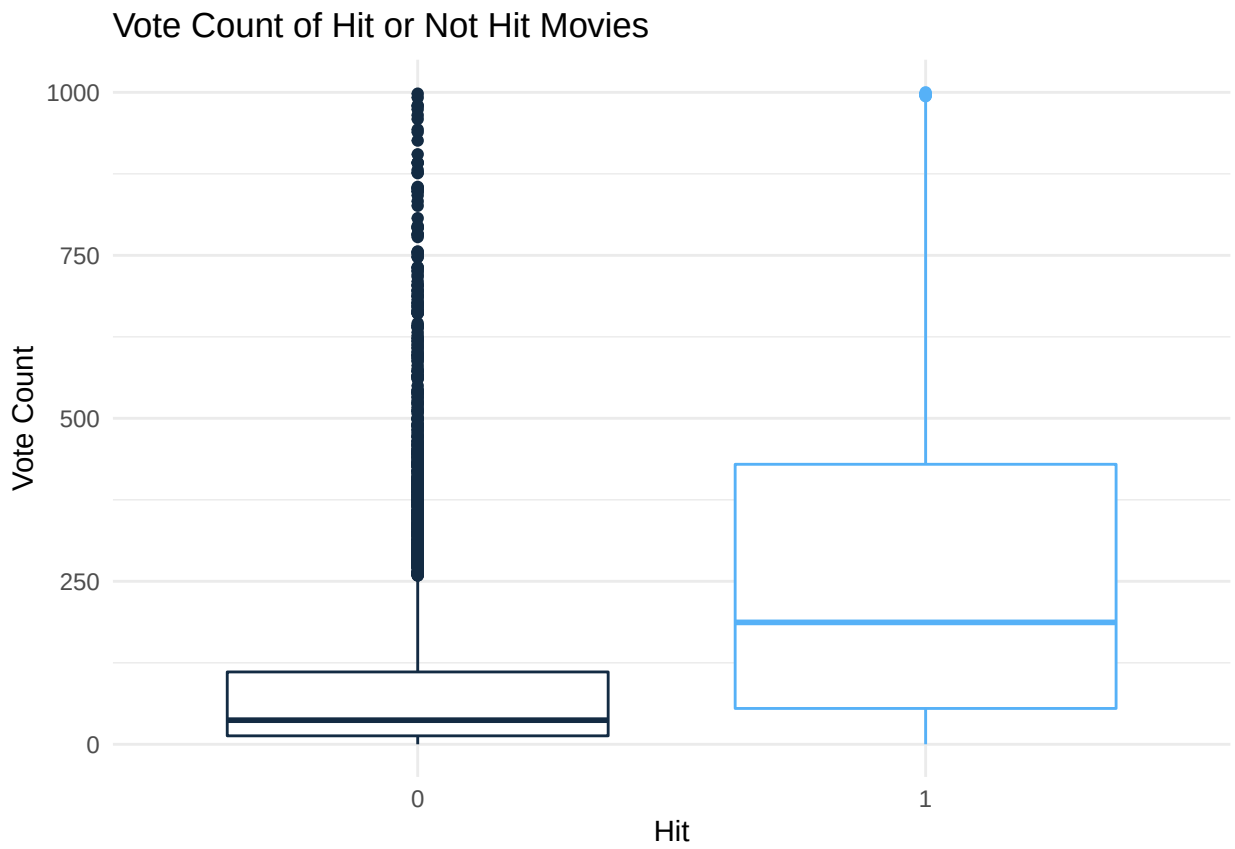


From the above boxplots we can see that that the popularity score for hit movies is significantly higher compared to that of non-hit movies

```
#Vote Count of hit and not-hit movies

Cleaned_data %>%
  ggplot(mapping = aes(x = as.factor(hit), y=vote_count, color= hit)) +
  geom_boxplot(show.legend = "False") +
  labs(y="Vote Count",x = "Hit", title= "Vote Count of Hit or Not Hit Movies")+
  ylim(0,1000)+
  theme_minimal()
```

```
## Warning: Removed 1107 rows containing non-finite values (stat_boxplot).
```

```
## Warning: `show.legend` must be a logical vector.
```

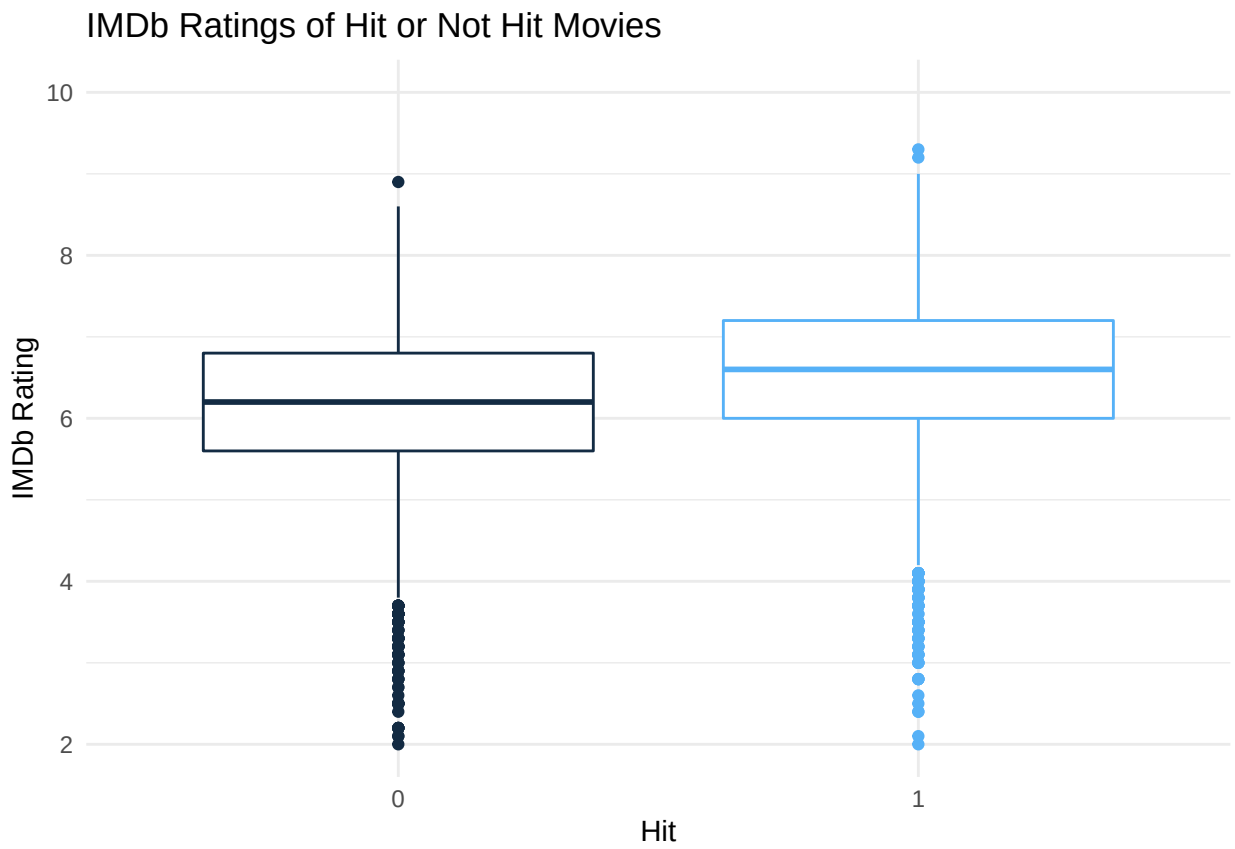


From the above boxplots we can see that that the vote count for hit movies is significantly higher compared to that of non-hit movies

```
#IMDB Ratings of hit and not-hit movies

Cleaned_data %>%
  ggplot(mapping = aes(x = as.factor(hit), y=weighted_average_vote, color= hit)) +
  geom_boxplot(show.legend = "False") +
  labs(y="IMDb Rating",x = "Hit", title= "IMDb Ratings of Hit or Not Hit Movies")+
  ylim(2,10)+
  theme_minimal()
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```

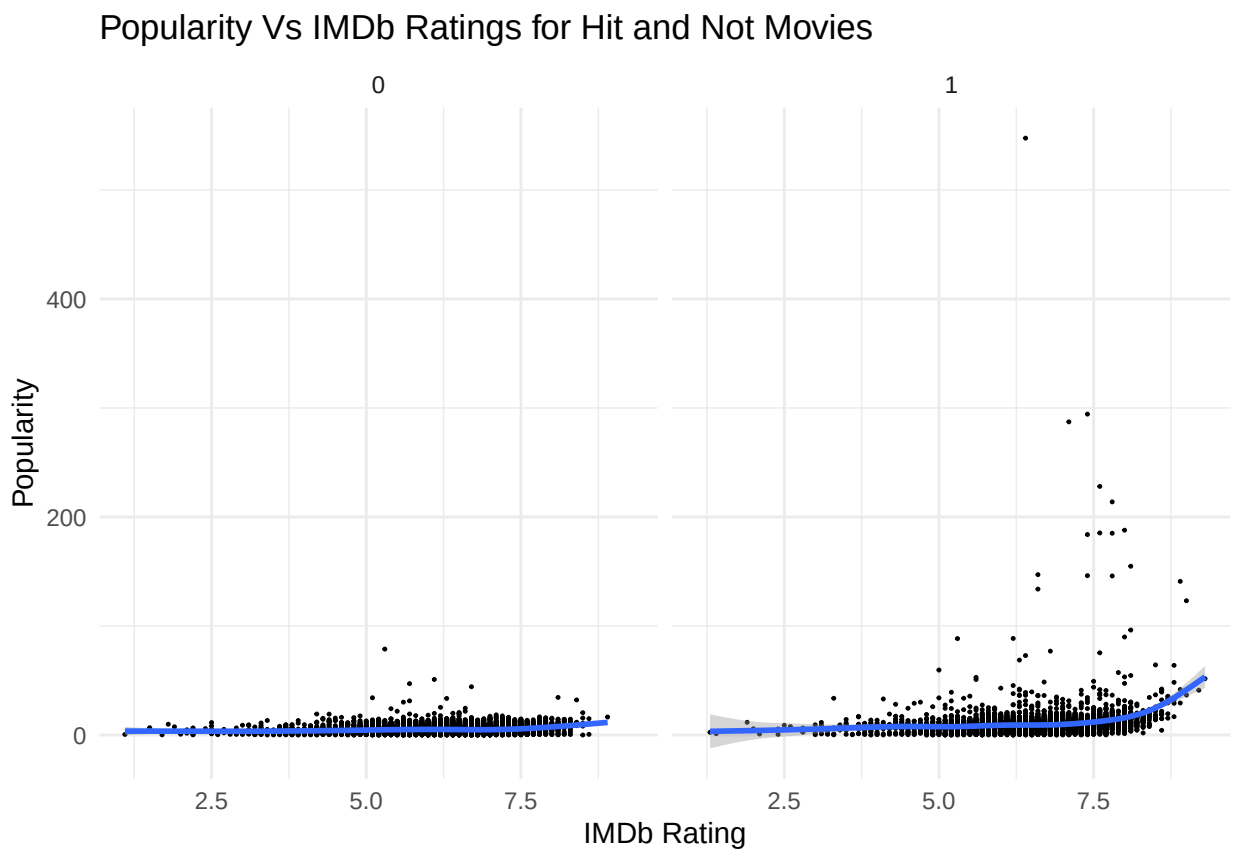
```
## Warning: `show.legend` must be a logical vector.
```



From the above boxplots we can see that that the IMDB Ratings for hit movies is slightly higher compared to that of non-hit movies

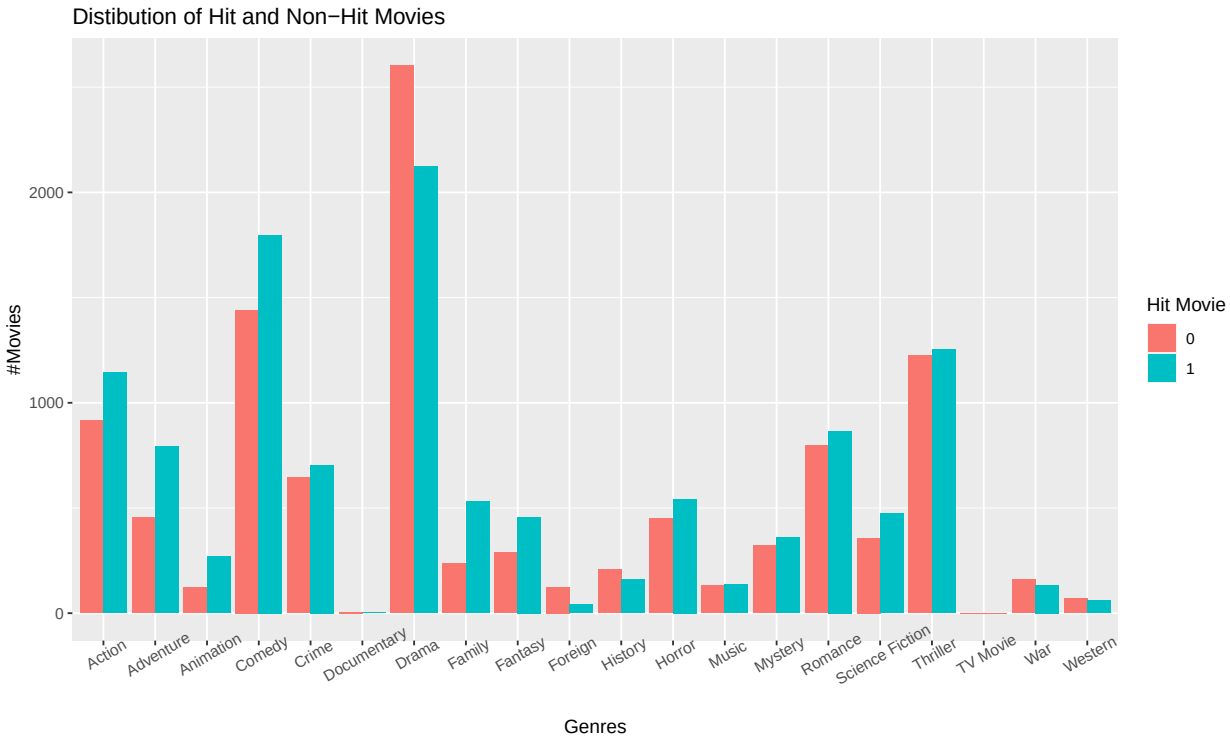
```
#Popularity vs IMDB Ratings

Cleaned_data %>%
  ggplot(mapping = aes(x =weighted_average_vote, y = popularity)) +
  geom_point(size=0.2)+
  facet_wrap(~hit) +
  labs(x="IMDb Rating",y="Popularity", title = "Popularity Vs IMDb Ratings for Hit and Not Movies")+
  geom_smooth()+
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



It can be seen that there is no correlation between popularity score and IMDB rating

```
#Impact of Genre

Cleaned_data %>%
  select(c(imdb_id, 126:145, hit)) %>%
  gather('genre','has_genre',2:21) %>%
  filter(has_genre == 1) %>%
  ggplot(mapping = aes(x = genre)) +
  geom_bar(aes(fill = as.factor(hit)), position = 'dodge') +
  theme(axis.text.x = element_text(angle = 30)) +
  labs(x = 'Genres', y = '#Movies', title = 'Distibution of Hit and Non-Hit Movies') +
  scale_fill_discrete(name = "Hit Movie")
```



As seen from the above graph, Genre `Drama` has significantly higher number of non-hit movies, when compared to other genres. On the other hand, genres `Action`, `Adventure`, `Animation`,`Comedy`, `Family`, `Fantasy` have higher number of hit movies, compared to Non-hit movies