

Analyzing IMDB Movies

Justin Biancamano

2022-10-27

Contents

Summary	1
How has the amount of movies released each year changed over time?	3
Questions about movie run times	4
What is the distribution of movie runtimes?	4
Do animated movies run shorter than non-animated movies?	4
Do the highest grossing films have different run times than the lowest grossing films	5
Questions about top actors	7
Of the Actors with the most lead Roles, who averages the highest IMDB_Ratings?	7
Al Pacino and Robert De Niro are considered to be two of the greatest Actors of all time but whose movies have a higher IMDB Rating	8
Questions about Movie Certificates and Genres	9
How has the amount of movies by certficate changed over time?	9
What are the most common genres and least common genres	10
Tables	11
Conclusions	12

Summary

In this report I will use the IMDB Movies Dataset from Kaggle in order to answer various questions about the top 1000 films by IMDB_Rating. After eliminating all the observations with null values the dataset contains 688 films with 17 variables. The variables are as follows:

- *Poster_Link*: The link to the movie's poster on IMDB
- *Series_Title*: The title of the film
- *Released_Year*: The year the movie was released
- *Certificate*: Audience Rating of G, PG, Pg-13, or R
- *Runtime*: Length of the movie in minutes

- *Genre*: Genre(s) of the movie
- *IMDB_Rating*: IMDB's Rating of the Movie
- *Overview*: Description of the movie
- *Meta_score*: A weighted average of critics reviews
- *Director*: Name of the Director
- *Star1*: Name of Lead Actor -*Star2,Star3,Star4*: Name of Supporting Actors -*No_of_Votes*: Total number of Votes made by users -*Gross*: Total money earned by that movie

I have also created a variable *Decade* which groups together films' release year into different decades. This variable was created to make some of the charts and patterns easier to notice. *Certificate* was originally more than 4 ratings because overtime the rating names and guidelines changed, and it is better for people to understand if the certificates are standardized to the modern ones we know today. Because we do not use *Overview* or *Poster_Link* in any of the analysis it has been removed to look at the data which will actually be used.

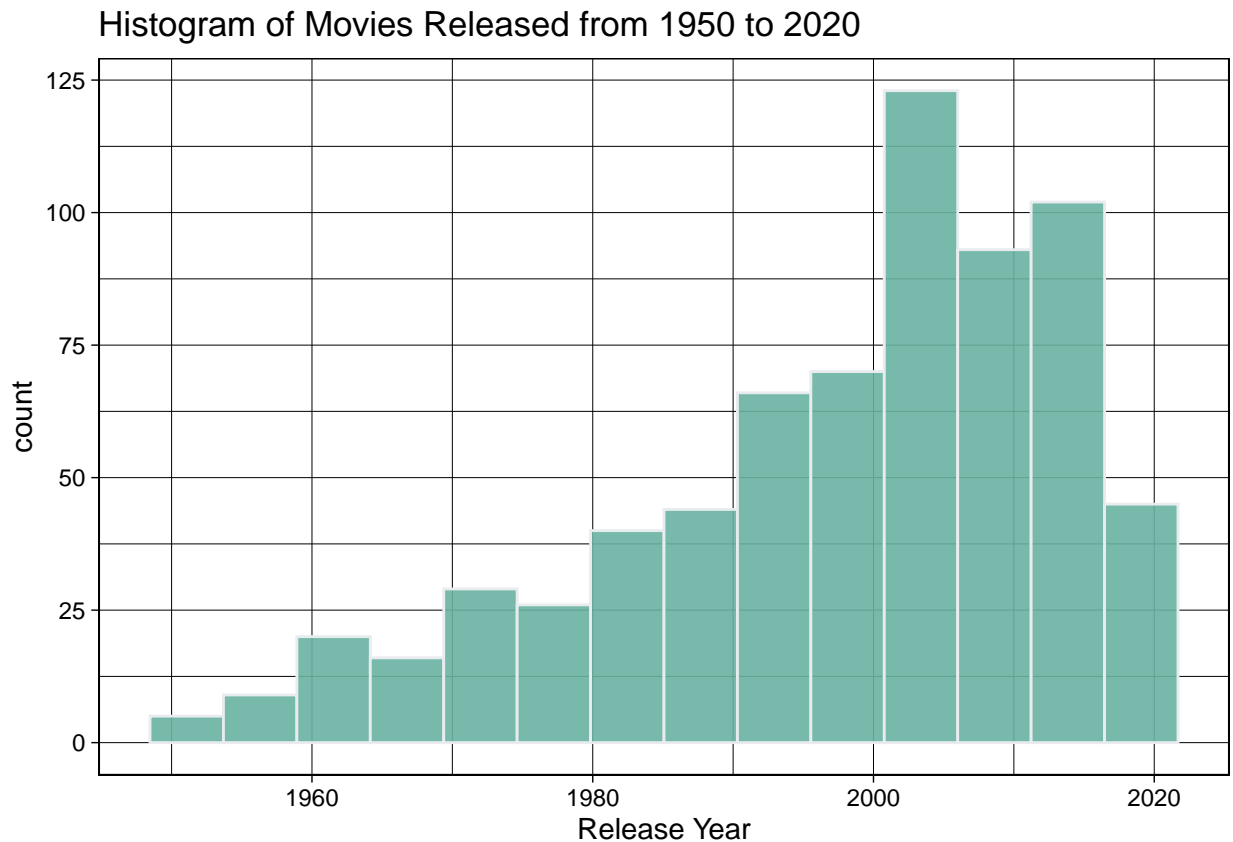
```
# Head the dataset but remove the columns that are not necessary
head(imdb[,-c(1,9)], n=10)
```

```
##              Series_Title Released_Year Decade Certificate
## 297 Star Wars: Episode VII - The Force Awakens      2015    2010         G
## 49              Avengers: Endgame      2019    2010        PG-13
## 395              Avatar      2009    2000        PG-13
## 50      Avengers: Infinity War      2018    2010        PG-13
## 422              Titanic      1997    1990        PG-13
## 216              The Avengers      2012    2010        PG-13
## 595      Incredibles 2      2018    2010        PG-13
## 3      The Dark Knight      2008    2000        PG-13
## 365              Rogue One      2016    2010        PG-13
## 53      The Dark Knight Rises      2012    2010        PG-13
##      Runtime              Genre IMDB_Rating Meta_score
## 297      138      Action, Adventure, Sci-Fi      7.9      80
## 49      181      Action, Adventure, Drama      8.4      78
## 395      162      Action, Adventure, Fantasy      7.8      83
## 50      149      Action, Adventure, Sci-Fi      8.4      68
## 422      194              Drama, Romance      7.8      75
## 216      143      Action, Adventure, Sci-Fi      8.0      69
## 595      118      Animation, Action, Adventure      7.6      80
## 3      152              Action, Crime, Drama      9.0      84
## 365      133      Action, Adventure, Sci-Fi      7.8      65
## 53      164              Action, Adventure      8.4      78
##      Director              Star1              Star2              Star3
## 297      J.J. Abrams      Daisy Ridley      John Boyega      Oscar Isaac
## 49      Anthony Russo      Joe Russo      Robert Downey Jr.      Chris Evans
## 395      James Cameron      Sam Worthington      Zoe Saldana      Sigourney Weaver
## 50      Anthony Russo      Joe Russo      Robert Downey Jr.      Chris Hemsworth
## 422      James Cameron      Leonardo DiCaprio      Kate Winslet      Billy Zane
## 216      Joss Whedon      Robert Downey Jr.      Chris Evans      Scarlett Johansson
## 595      Brad Bird      Craig T. Nelson      Holly Hunter      Sarah Vowell
## 3      Christopher Nolan      Christian Bale      Heath Ledger      Aaron Eckhart
## 365      Gareth Edwards      Felicity Jones      Diego Luna      Alan Tudyk
## 53      Christopher Nolan      Christian Bale      Tom Hardy      Anne Hathaway
##      Star4 No_of_Votes      Gross
## 297      Domhnall Gleeson      860823 936662225
```

## 49	Mark Ruffalo	809955	858373000
## 395	Michelle Rodriguez	1118998	760507625
## 50	Mark Ruffalo	834477	678815482
## 422	Kathy Bates	1046089	659325379
## 216	Jeremy Renner	1260806	623279547
## 595	Huck Milner	250057	608581744
## 3	Michael Caine	2303232	534858444
## 365	Donnie Yen	556608	532177324
## 53	Gary Oldman	1516346	448139099

How has the amount of movies released each year changed over time?

```
ggplot(imdb, aes(x=Released_Year)) + geom_histogram(bins = 14, fill="#69b3a2", color="#e9ecef", alpha=0
```

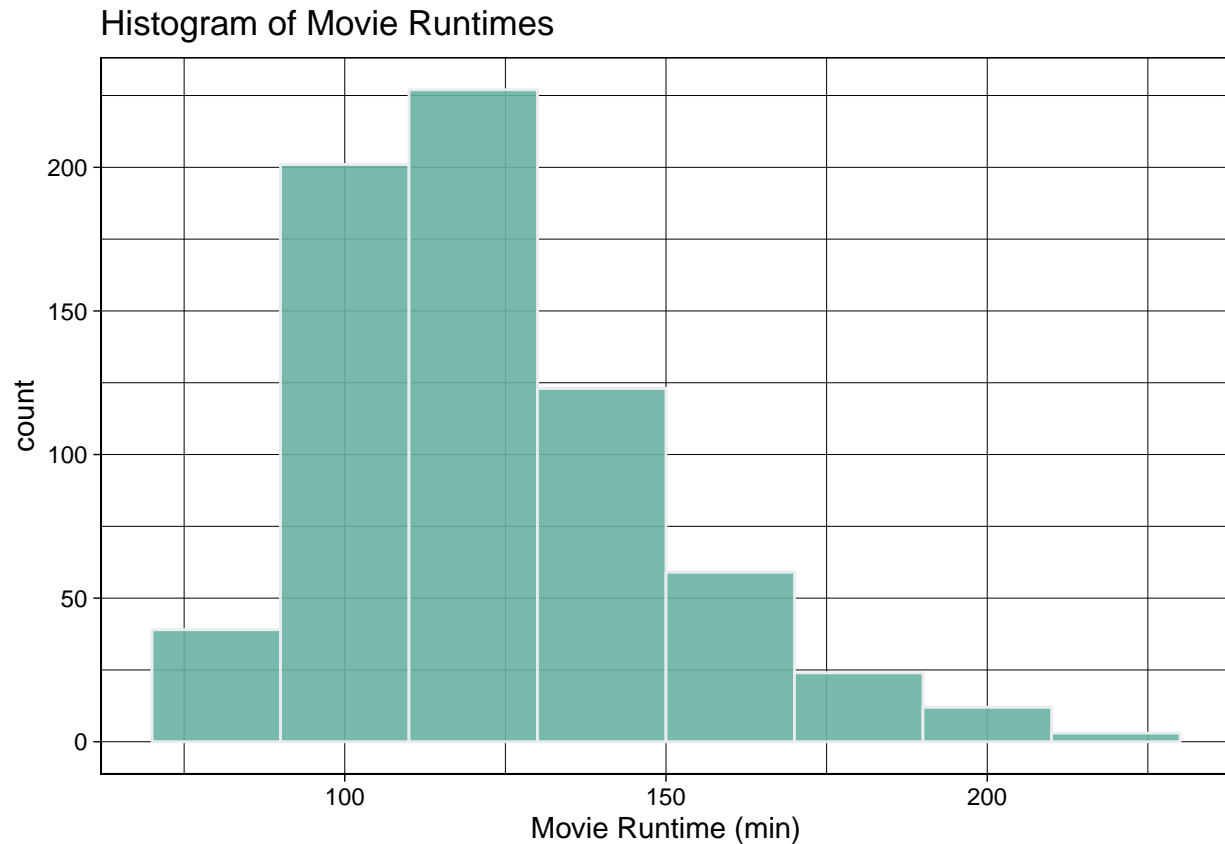


The amount of movies released each year has increased with there being some decent sized drops notably in the early 2010s and the late 2010s. Overall, the number of movies released each year has increased from the start of the period to the end of the period. It appears that with every drop in movie releases from the previous period there has always been an increase in the next 5 year period. That is a good sign for increase in movies post-Covid which negatively impacted the mmovie industry.

Questions about movie run times

What is the distribution of movie runtimes?

```
ggplot(imdb, aes(x=Runtime)) + geom_histogram(binwidth = 20, fill="#69b3a2", color="#e9ecef", alpha=0.9)
```



As we can see movie run time is significantly skewed right which makes sense considering all movies are going to be at least an hour and can end whenever the director deems the story is finished. The mean run time appears to be at around 125 minutes or 2 hours and 5 minutes.

Do animated movies run shorter than non-animated movies?

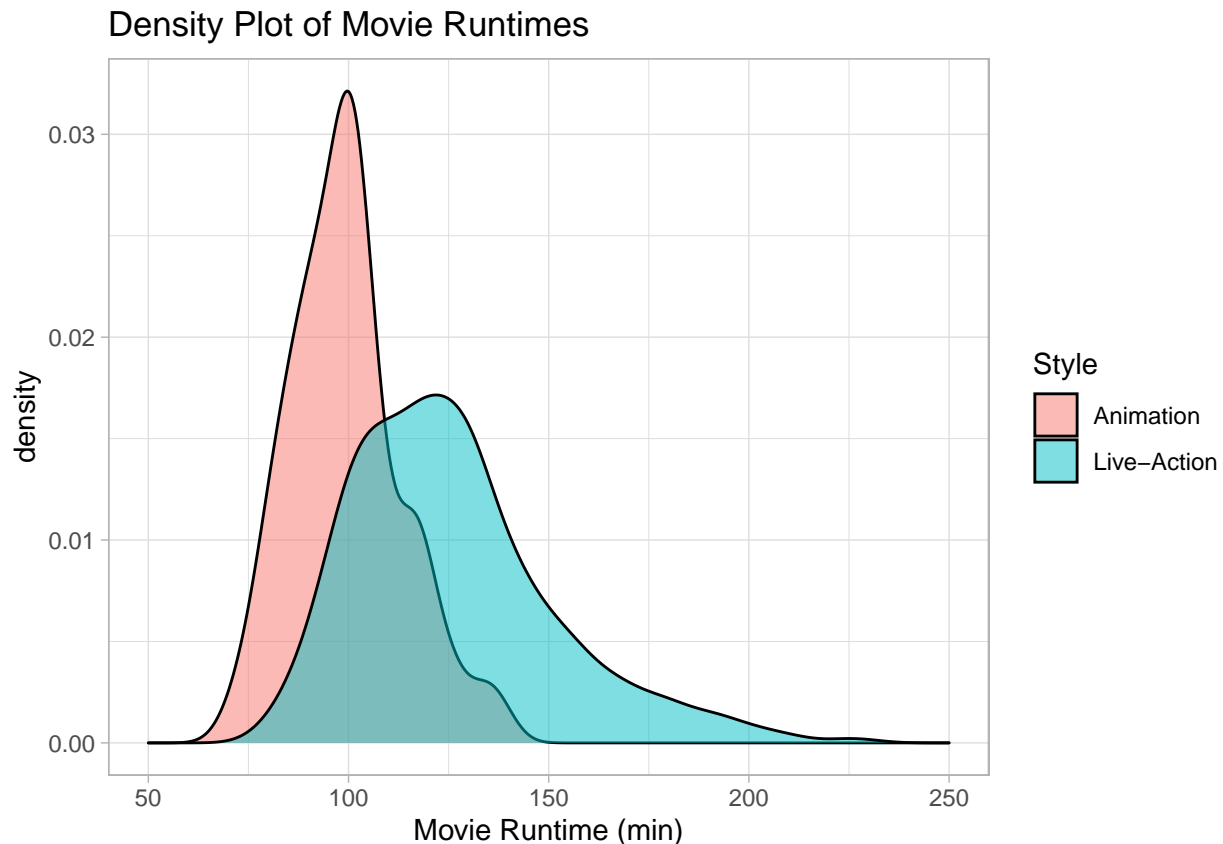
```
# Filter the data into animated and non_animated movies
animation <- filter(imdb, grepl("Animation", Genre, fixed = TRUE))
no_animation <- filter(imdb, !grepl("Animation", Genre, fixed=TRUE))

# Create a new column style to distinguish between the two
animation$Style <- rep("Animation", nrow(animation))
no_animation$Style <- rep("Live-Action", nrow(no_animation))

# Rejoin the separate datasets
imdb_style <- rbind(animation, no_animation)
```

```
# Make the Style variable a two level factor
imdb_style$Style <- as.factor(imdb_style$Style)

# Plot the densities of the two movie styles
ggplot(imdb_style, aes(x=Runtime, group=Style, fill=Style)) +
  geom_density(alpha=0.5) + xlim(50, 250) + labs(title="Density Plot of Movie Runtimes") + xlab("Movie Runtime (min)")
```

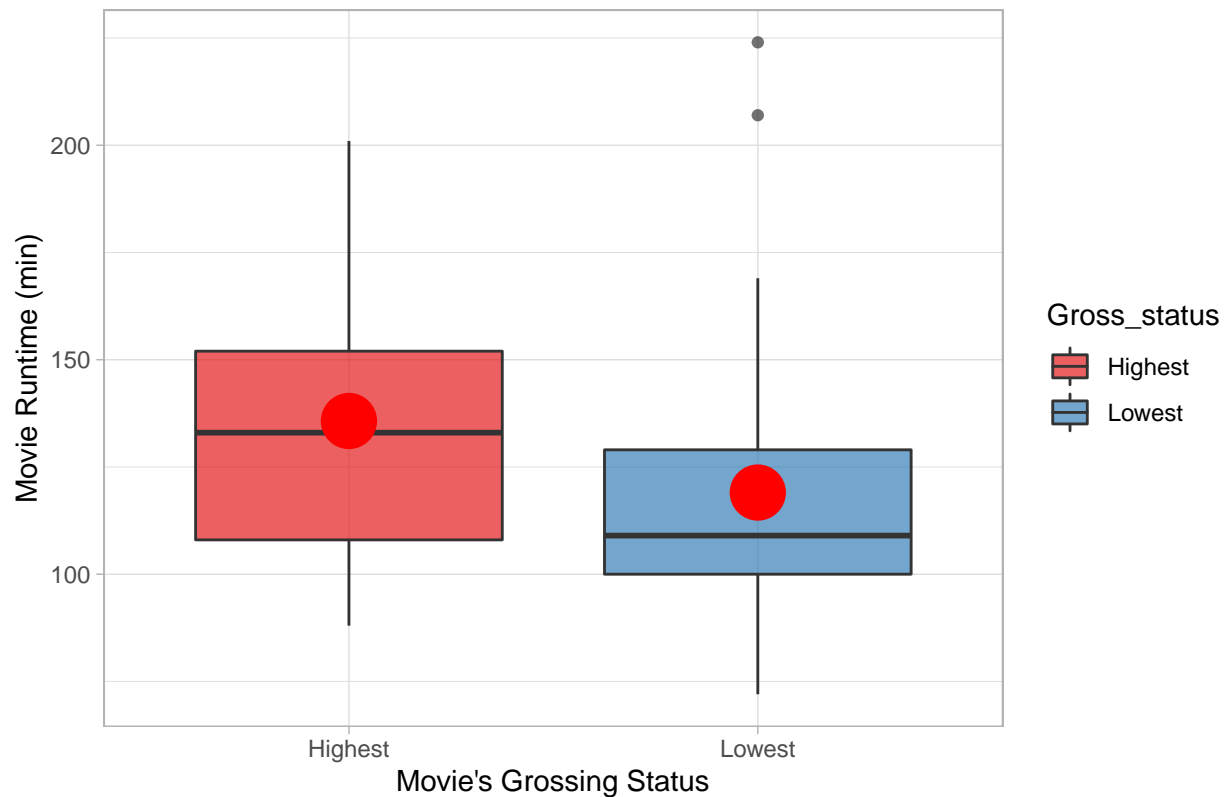


Yes, we can see that animated movies have a mean runtime of about 100 minutes and no runtimes greater than 150, while the non animated movies have a mean runtime of about 125 with the longest movie being almost 240 minutes long.

Do the highest grossing films have different run times than the lowest grossing films

```
# Create side by side boxplots to show the spread of runtimes for the highest and lowest grossing films
ggplot(gross, aes(x=Gross_status, y=Runtime, fill=Gross_status)) +
  geom_boxplot(alpha=0.7) +
  stat_summary(fun=mean, geom="point", shape=20, size=14, color="red", fill="red") +
  theme(legend.position="none") +
  scale_fill_brewer(palette="Set1") + labs(title="Box plots of Runtimes for 25 Highest and Lowest Grossing Films")
```

Box plots of Runtimes for 25 Highest and Lowest Grossing Films of all Time



As we can see the highest grossing films have a longer average runtime than the lowest grossing films and the lowest grossing films have two extreme points with films running longer than 200 minutes. We will now look at the summary to find the exact differences.

```
# Show the summary statistics for the highest grossing films and lowest grossing films
sapply("25 Highest Grossing Films", function(x) summary(top_gross$Runtime))
```

```
##          25 Highest Grossing Films
## Min.          88.00
## 1st Qu.       108.00
## Median       133.00
## Mean        135.72
## 3rd Qu.     152.00
## Max.       201.00
```

```
sapply("25 Lowest Grossing Films", function(x) summary(bottom_gross$Runtime))
```

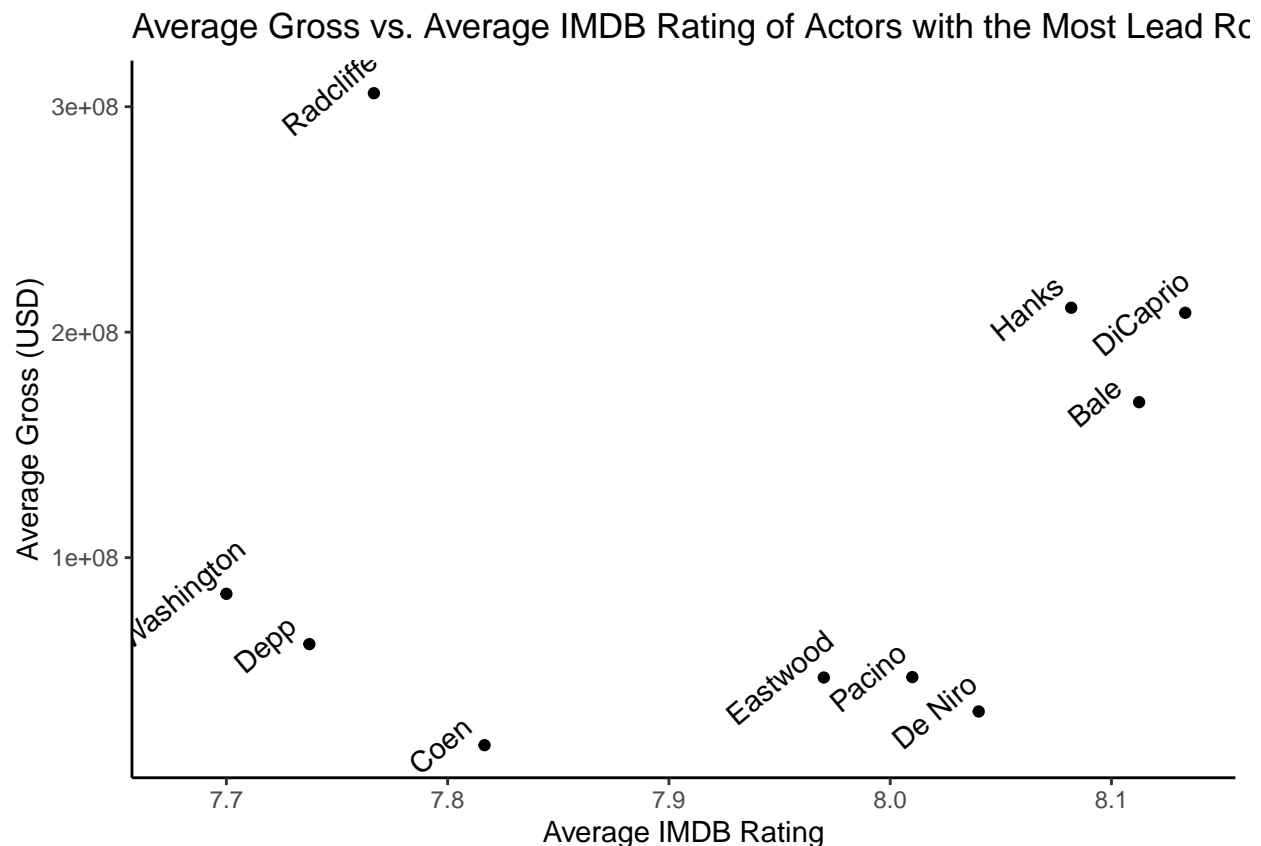
```
##          25 Lowest Grossing Films
## Min.          72.00
## 1st Qu.       100.00
## Median       109.00
## Mean        119.04
## 3rd Qu.     129.00
## Max.       224.00
```

The highest grossing films run an average of 136 minutes while the lowest grossing films run an average of 119 minutes. The highest grossing film's median movie run time is 133 minutes while the median run time for the lowest grossing films is 109. The highest grossing films run over 15 minutes longer than the lowest grossing films and have a smaller range than the lowest grossing films. It is interesting to note that the shortest low grossing film is 16 minutes shorter than the shortest highest grossing film, and the longest low grossing film is 23 minutes longer than the longest high grossing film.

Questions about top actors

Of the Actors with the most lead Roles, who averages the highest IMDB_Ratings?

```
# Plot the ten actors with the most leading roles and show their names next to their point
ggplot(leads, aes(Average_IMDB_Score, Average_Gross)) + geom_point() +
  geom_text(aes(label = Actor), nudge_x = -0.02, angle=40, check_overlap = T) + theme_classic() + labs(t
```

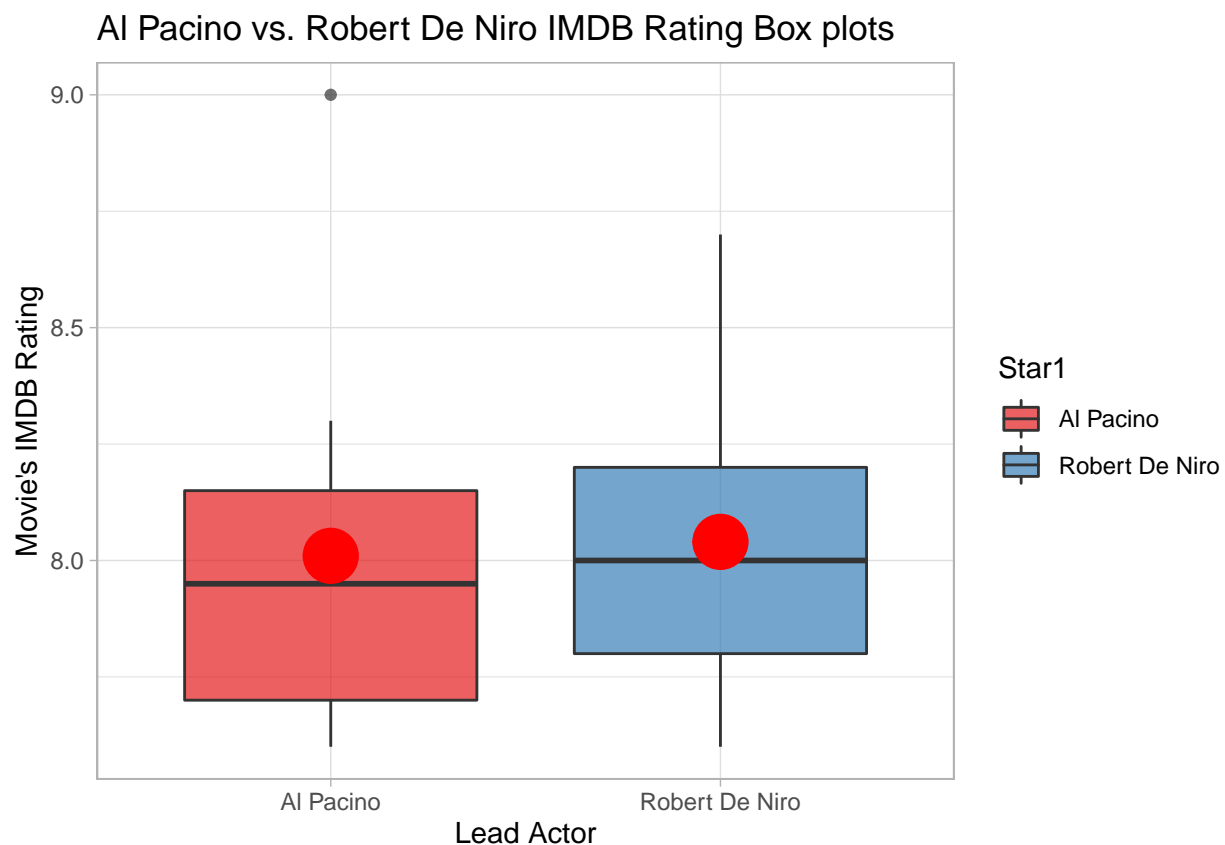


It does not appear that a higher IMDB rating means a higher gross. Daniel Radcliffe has the highest average gross because he plays the main character in the Harry Potter franchise, and that series has a large fan base. Pacino and DeNiro usually do the same type of movies which are crime and mob related so it makes sense they have similar average gross and ratings. It is interesting to note that from the 10 actors with the most lead roles, Denzel Washington, Johnny Depp, Daniel Radcliffe, and Ethan Coen have the lowest average IMDB rating. This could suggest a positive correlation between number of movies as the leading actor and average IMDB rating.

Al Pacino and Robert De Niro are considered to be two of the greatest Actors of all time but whose movies have a higher IMDB Rating

```
# Combine the data where Al Pacino is the lead actor and Robert De Niro is the lead actor
mobsters <- rbind(pacino, niro)
# Make the actor's name a factor to group by for the plot
mobsters$Star1 <- as.factor(mobsters$Star1)

# Create side by side box plots to show the actors spread of ratings
ggplot(mobsters, aes(x=Star1, y=IMDB_Rating, fill=Star1)) +
  geom_boxplot(alpha=0.7) +
  stat_summary(fun=mean, geom="point", shape=20, size=14, color="red", fill="red") +
  theme(legend.position="none") +
  scale_fill_brewer(palette="Set1")+ labs(title="Al Pacino vs. Robert De Niro IMDB Rating Box plots")
```



This box plot only shows the distribution of movies where Al Pacino or Robert De Niro were the lead actors. As a lead actor Robert De Niro has a slightly higher mean IMDB rating, but Pacino's rating distribution is thrown off by the one movie in which he received a rating of 9. We will now look at the summary to examine it more in depth

```
sapply("Al Pacino IMDB Ratings", function(x) summary(pacino$IMDB_Rating))
```

```
##           Al Pacino IMDB Ratings
## Min.                7.60
```



```
## 1st Qu.          7.70
## Median           7.95
## Mean             8.01
## 3rd Qu.          8.15
## Max.             9.00
```

```
sapply("Robert De Niro IMDB Ratings", function(x) summary(niro$IMDB_Rating))
```

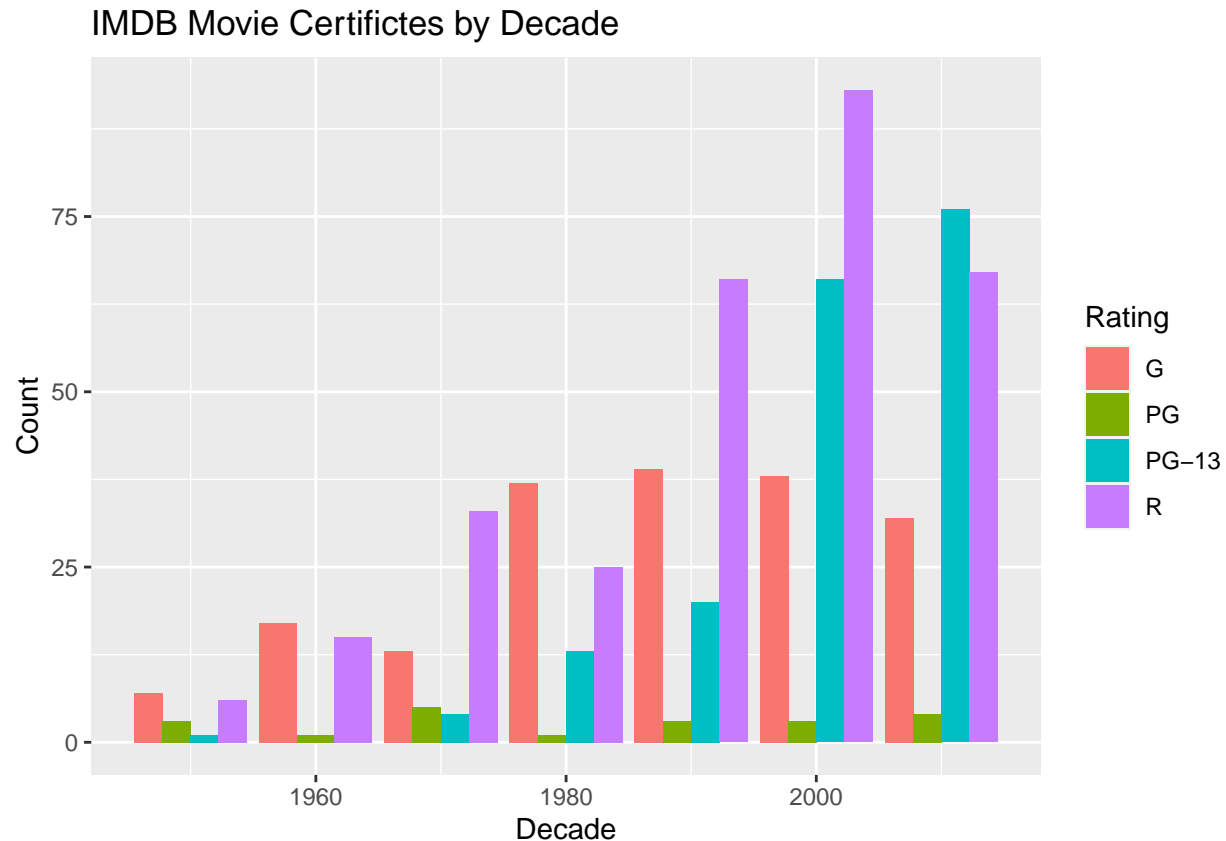
```
##           Robert De Niro IMDB Ratings
## Min.              7.60
## 1st Qu.           7.80
## Median            8.00
## Mean              8.04
## 3rd Qu.           8.20
## Max.              8.70
```

Both actor's lowest rated movie was a 7.6 while Pacino's highest rated movie received a score 0.3 points higher than that of De Niro's highest Scoring films. Pacino does have the highest movie rating by 0.3 points, but the mean and medians for the tow actors are almost identical. Considering that IMDB scores are always rounded to one decimal place we can conclude that both actors average the same rating for movies where they have the lead role. This makes sense because Al Pacino and Robert De Niro typically star in crime/mafia movies.

Questions about Movie Certificates and Genres

How has the amount of movies by certificate changed over time?

```
ggplot(imdb_freq, aes(Decade, Count, fill = Rating)) +
  geom_bar(stat="identity", position = "dodge") +
  labs(title="IMDB Movie Certifictes by Decade")
```

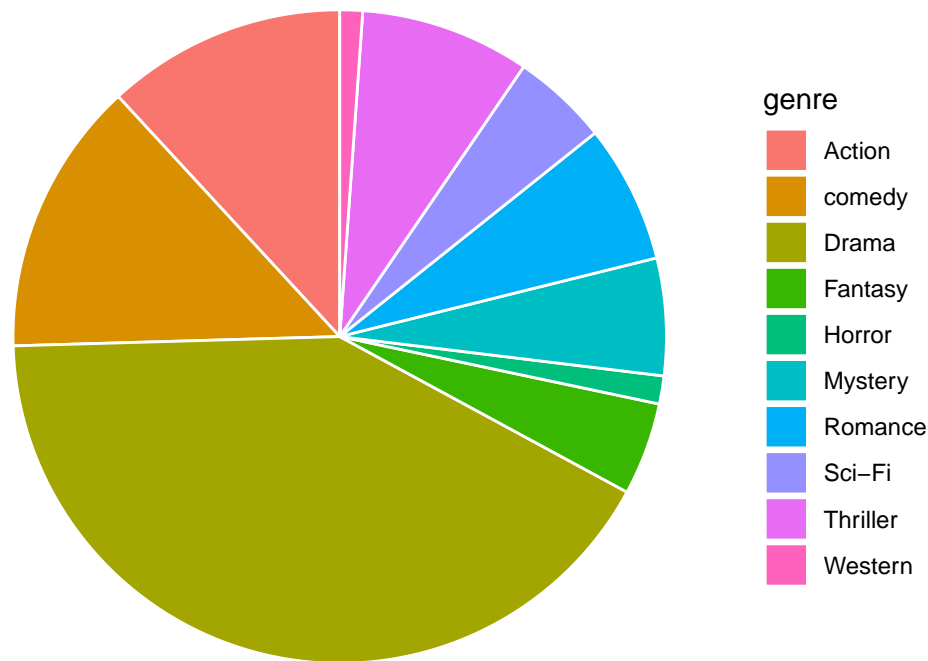


As we have seen from the histogram showing the number of movies released each year we are not surprised to see higher columns in more recent decades. PG-13 and R rated movies have shown a significant increase from the 1980s to the late 90s and early 2000s. In the 50s, 60s, and 80s the predominant movie rating has been G, and in 1970 it being second to R rated movies. After the 1980s, though there have been almost 50 more R and PG-13 rated movies than G and PG. Overall the number of G rated movies has increased until the 90s, and has begun to decrease slightly. The number of PG movies released each decade has remained the same. Finally the number of PG_13 and R rated movies released each decade has increased significantly. This means that over time the highest rated IMDB movies are more adult and teenager audience focused.

What are the most common genres and least common genres

```
ggplot(genre_df, aes(x="", y=count, fill=genre)) + geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) + theme_void() + labs(title="Pie Chart of Movie Genres")
```

Pie Chart of Movie Genres



As we can see from the pie chart Drama movies make up nearly half of all the movies in the data set. The top three genres in movies are Drama, Comedy, and Action, in that order. The genres in the least amount of movies are Western, Horror, and Fantasy, with Sci-Fi having a bit more than Fantasy. It is interesting that there are significantly more Thriller movies than there are Horror when it seems like they usually go hand in hand.

Tables

```
# Look at the number of movies for each rating group
head(scores)
```

##	IMDB_Rating Group	Number of Movies	Percent of Total
## 1	7.0 to 7.9	405	0.59
## 2	8.0 to 8.9	278	0.4
## 3	9.0 to 10.0	5	0.01

It seems very hard to have a movie receive an IMDB rating above 9 with 1% of all movies in the data set having a rating higher than 9. All but 5 movies have an IMDB rating between 7 and 8.9. There is almost 20% more movies with ratings between 7 and 7.9 than there are movies between 8 and 8.9. These are the highest rated movies on IMDB, and yet there are still very few to have a rating over

```
# We did not round before to make the graph easier to interpret but now we need to round everything so
leads[,-c(1:4)] <- round(leads[,-c(1:4)],1)
head(leads)
```

##	Actor	First_Release	Last_Release	Highest_Grossing_Film	Average_Runtime
## 1	Hanks	1993	2019	Toy Story 4	131.4
## 2	Pacino	1973	1997	Heat	148.6
## 3	Eastwood	1964	2008	Gran Torino	126.6
## 4	De Niro	1976	2019	Awakenings	145.7
## 5	DiCaprio	1997	2019	Titanic	156.9
## 6	Bale	1987	2015	The Dark Knight	133.9

##	Average_IMDB_Score	Average_Meta_Score	Average_Number_of_votes	Average_Gross
## 1	8.1	79.9	704054.9	210841775
## 2	8.0	74.6	399344.0	47028078
## 3	8.0	74.0	266495.8	46844780
## 4	8.0	80.2	369827.2	31786403
## 5	8.1	74.4	1023275.4	208591306
## 6	8.1	70.8	955656.5	168948929

Here is the table showing the actors with the most lead roles, their first and most recent release years, name of their highest grossing films, and averages for other important variables.

Conclusions

Although these questions were based on the top IMDB Rated movie, this sample can offer good insights into the population of all movies on IMDB. But here is the summary of all questions asked and their answers:

- **Has the number of movies released each year increased over time?** Yes, there has been some periods in which the number has decreased but there are significantly more movies released in the past 20 years than there have been in the first 20 years of the period. Maybe movies have gotten better or there really are just more movies released each year.
- **What is the mean movie runtime for the top rated movies?** The mean runtime is 2 hours and 5 minutes with some movies running up to almost almost 2 hours longer.
- **Do animated movies run shorter than non animated movies?** Yes, they run on average almost half an hour shorter probably because most animated movies are made for kids.
- **Do the highest grossing films have different runtimes than the lowest grossing films?** Yes, the highest grossing films run about 15 minutes longer than the lowest grossing films.
- **For actors with the most lead roles, who has the highest average gross, and who has the highest average IMDB rating?** Daniel Radcliffe, Tom Hanks, and Leonardo DiCaprio have the highest average gross at over 200 million dollars. Leonardo DiCaprio, Christian Bale, and Tom Hanks have the highest average IMDB rating with all averaging about 8.1
- **Who averages higher IMDB ratings Pacino or DeNiro?** They average the same rating probably because they star in similar movies.
- **How has movie certificates changed over time?** There has been a big increase in the number of PG-13 and R films while PG has remained constant. This means that the highest rated movies on IMDB are predominantly R and PG-13.

- **What are the most and least common movie genres?** Many movies have multiple genres, and the most common are Drama, Comedy and Action. The least common movie genres are Western, Horror, and Fantasy movies suggesting that the most common genres for this sample earn higher average IMDB ratings than the least common genres.

It is interesting to see how many more films for adult and teen audience have been released over the years, and to see the makeup of the movie genres going forward. To perform further research I would like to acquire a bigger dataset that expands the IMDB movie ratings to include lower rated movies. I would like to see if with a wider range of observations, if there is a correlation between IMDB ratings and Gross. I would also like to see if the majority of the lower rated movies are the genres of the least common genres of the highest rated movies, or if there really are just significantly less Western, Horror, and Fantasy films released.