Introduction-

Business Problem : Mubi.com is an online movie platform which tries to show movies most preferred by movie audience.It earns profits through number of people using the platform and therefore wants to include such movies on its platform so that maximum people are attracted to use it.For this purpose it wants to analyse audience movie preference and the earning potential of movies so as to maximise its users.We as data analyst,will use process of analytics life cycle to solve this problem.

STEP 1: Problem Identification

We are given this problem statement from the business side-

Analysing audience movie preference and earning potential of movies'

They need to have a solution to this project as it will help in understanding what movies should be marketed to their audience for increasing their earnings.

The analytics team in the company has come up with the following initial problem statement based on above requirements -

- Analyse audience movie preference across different dimensions like rating system, genre, language etc.
- Analyse movie earning based on audience movie preference and other dimensions like genre, year released etc.
- Build an interactive dashboard based on above analysis so that business stakeholder can interact with data on their own and take quick decisions or do further research using this preliminary analysis.

STEP 2: Data Identification

We will choose the imdb and rotten dataset, below we are stating the reason why we are going with this data.

Ratings Data

The purpose of choosing imdb and rotten tomatoes rating was that these are the two most popular rating systems that people use to watch movies. So they can be taken as indicators for audience movie preference.

In particular

- imdb score shows what a general moviegoer audience prefer
- rotten tomatoes score shows how much critics prefer a movie

Earnings Data

Movie earnings data is available in the worldwide_gross_income column of IMDB dataset. It basically tells us the box office earnings of movies in millions.

STEP 3: Data Collection

We have collected our data from Kaggle for this problem because this is the easiest to get good quality data for our initial problem statement. By good quality we mean, a decent sized data representing the whole problem area.

For imdb dataset we had 4 files in it namely IMDB names.csv, IMDB ratings.csv,IMDB movies.csv and IMDB title_principles.csv but for our analysis we used only IMDB movies.csv

For the rotten tomatoes dataset also we had 2 sub-files namely rotten_tomaotoes_movies.csv and rotten_tomatoes_critic_review.csv . Here also we used only the rotten_tomatoes_movies.csv file for the analysis.

There are 2 more option also available for data collection

- Using an API (application programming interface)
- Building a dataset yourself using market study

We did not utilise these because this was a preliminary study and we already have a good dataset with easy access from Kaggle.

STEP 4: Data Cleaning and Manipulation

Final Dataset for the project

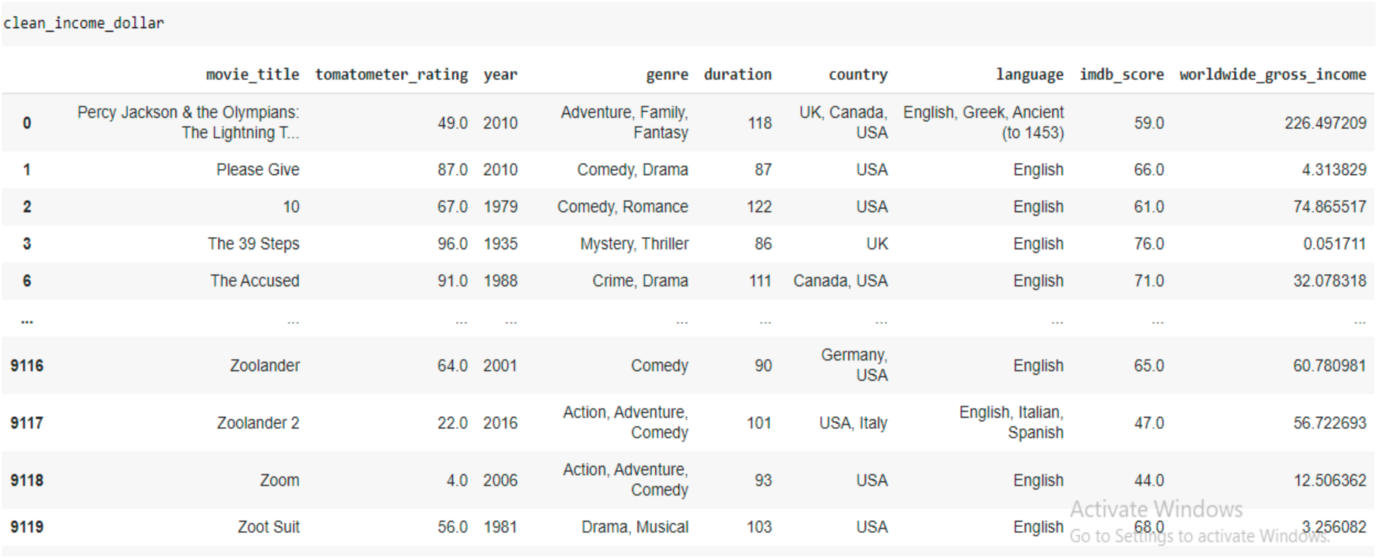
Initially Imdb movies dataframe had 85855 rows and 22 columns like year,title,review from users,critic reviews,imdb score ,genre etc but we selected only few of these columns for analysis

For the rotten tomatoes movies dataframe we had 17711 rows and 22 columns like actor ,director ,audience status ,top critics count etc out of which we had to drop many columns that were not important for analysis.

We combined the IMDB and rotten tomato movie dataset in one data file.This was required as for the analysis purpose some important column information was in imdb and some was in rotten tomatoes.

- We cleaned the worldwide income column by removing null values and have taken earnings only in $.
- We scaled the imdb ratings by 10 times so as to make it comparable to rotten tomatoes ratings.
- For countries and languages columns we have taken values which have more than 100 movies at least.

The final dataset after merging the imdb and rotten tomatoes dataframes included only 7148 rows and 9 columns,a picture of this dataframe is shown below:



```
clean_income_dollar
```

	movie_title	tomatometer_rating	year	genre	duration	country	language	imdb_score	worldwide_gross_income
0	Percy Jackson & the Olympians: The Lightning T...	49.0	2010	Adventure, Family, Fantasy	118	UK, Canada, USA	English, Greek, Ancient (to 1453)	59.0	226.497209
1	Please Give	87.0	2010	Comedy, Drama	87	USA	English	66.0	4.313829
2	10	67.0	1979	Comedy, Romance	122	USA	English	61.0	74.865517
3	The 39 Steps	96.0	1935	Mystery, Thriller	86	UK	English	76.0	0.051711
6	The Accused	91.0	1988	Crime, Drama	111	Canada, USA	English	71.0	32.078318
...	...	...	...	...	...	...	...	...	...
9116	Zoolander	64.0	2001	Comedy	90	Germany, USA	English	65.0	60.780981
9117	Zoolander 2	22.0	2016	Action, Adventure, Comedy	101	USA, Italy	English, Italian, Spanish	47.0	56.722693
9118	Zoom	4.0	2006	Action, Adventure, Comedy	93	USA	English	44.0	12.506362
9119	Zoot Suit	56.0	1981	Drama, Musical	103	USA	English	68.0	3.256082

STEP 5: Data Analysis and Visualisation

This step is all about solving for the initial problem statement with the Final dataset that we have created.

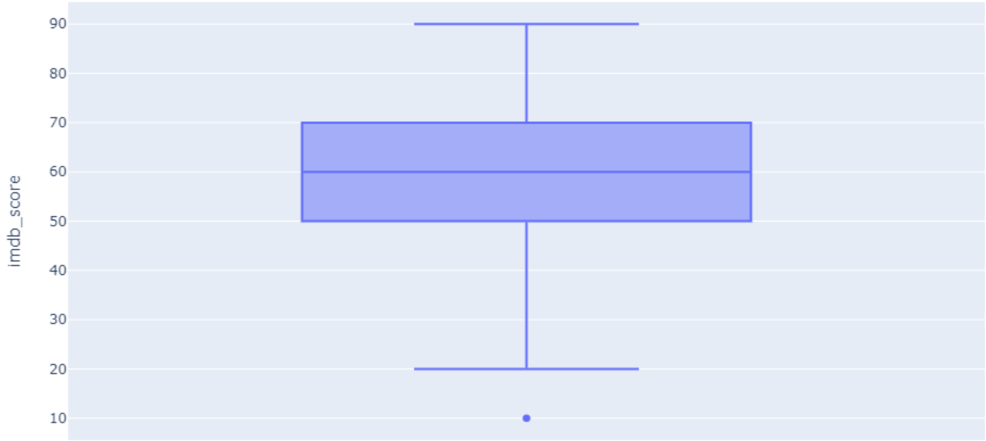
- Analyse audience movie preference across different dimensions like rating system, genre, language etc.
- Analyse movie earnings based on audience movie preference and other dimensions like genre, year released etc.
- Build an interactive dashboard based on the above analysis so that business stakeholders can interact with data on their own and make quick decisions or do further research using this preliminary analysis.

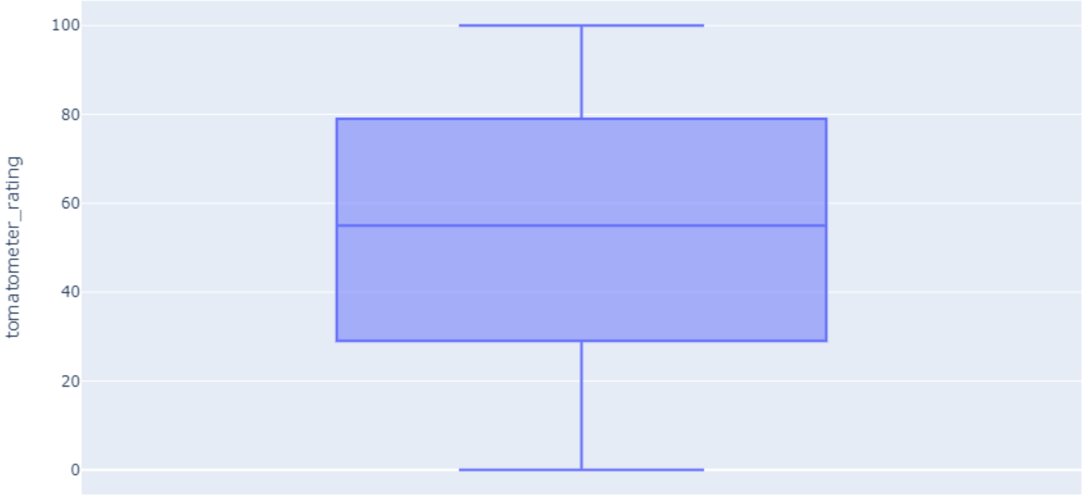
Assessment of Audience movie preference

We can analyse the two ratings of IMDB and rotten tomatoes across multiple dimensions to assess audience movie preference.

We will implement following analysis points

Comparison of rotten tomatoes and IMDB scores - this helps in understanding how movies are preferred across different rating systems.





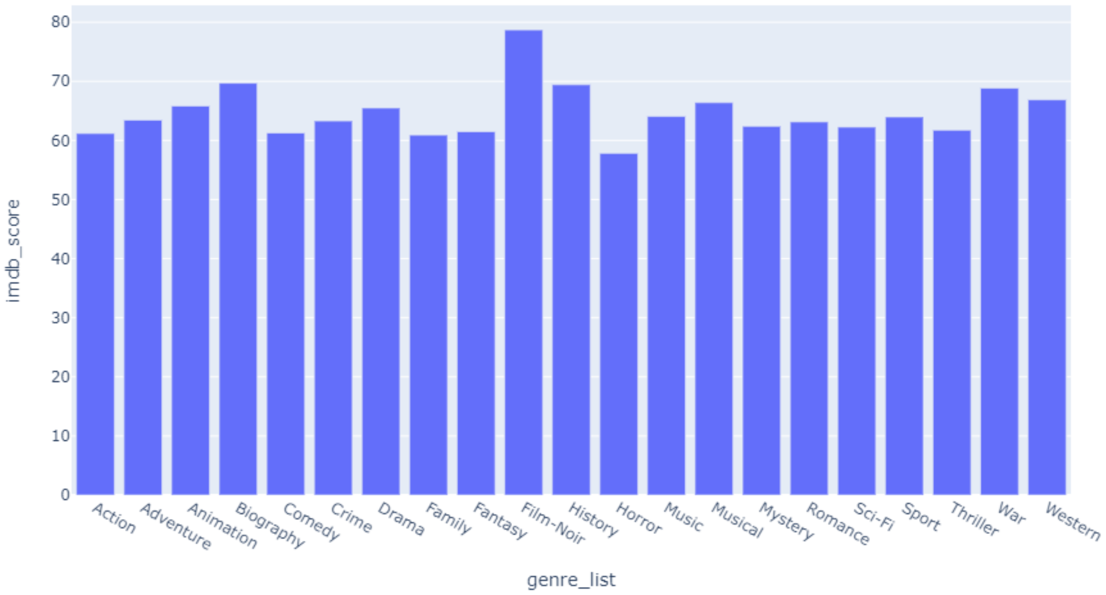
Analysis Point

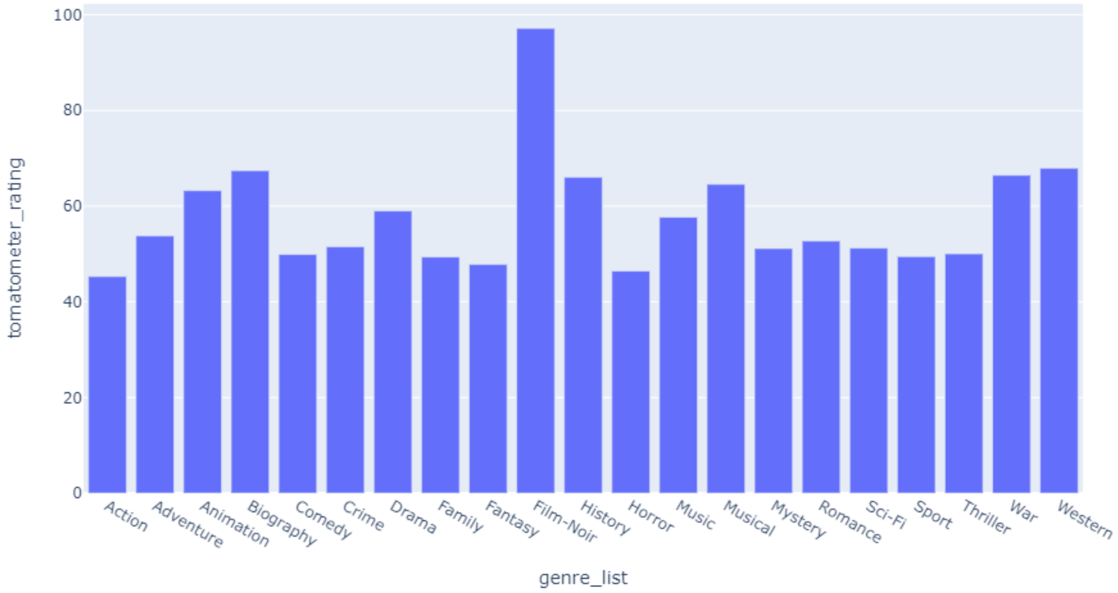
From the above two plots we can observe that most of the IMDB scores are in a narrow band of 60 - 70 while most of the rotten tomatoes scores lie in a bigger band of 30 - 80.

This might be due to the fact that audiences usually rate most of the movies averagely with less strictness i.e. movies are not usually rated at extremes

but critics have a more strict rating criteria so despite many movies being rated average, many of the other movies are either rated very high or very low.

Rating preference across genre - which genres are more preferable



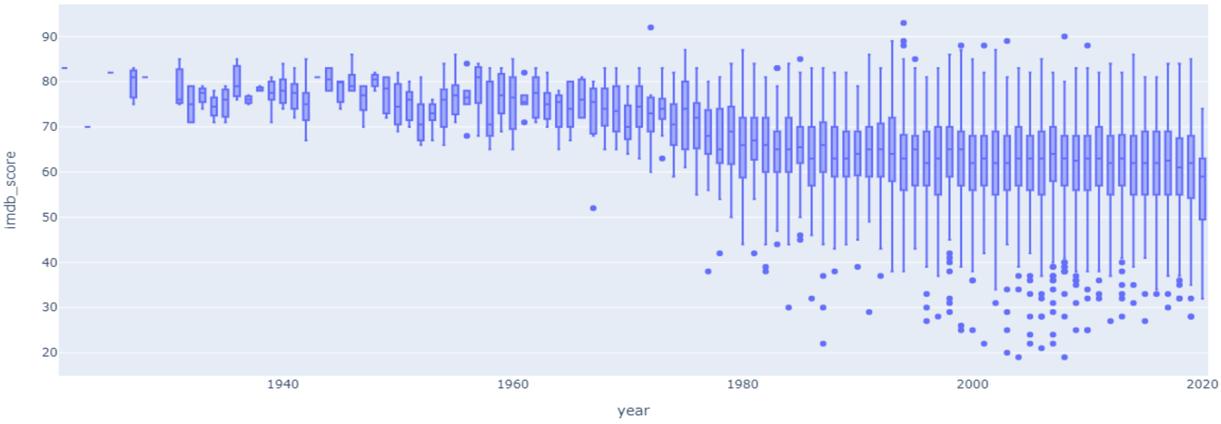


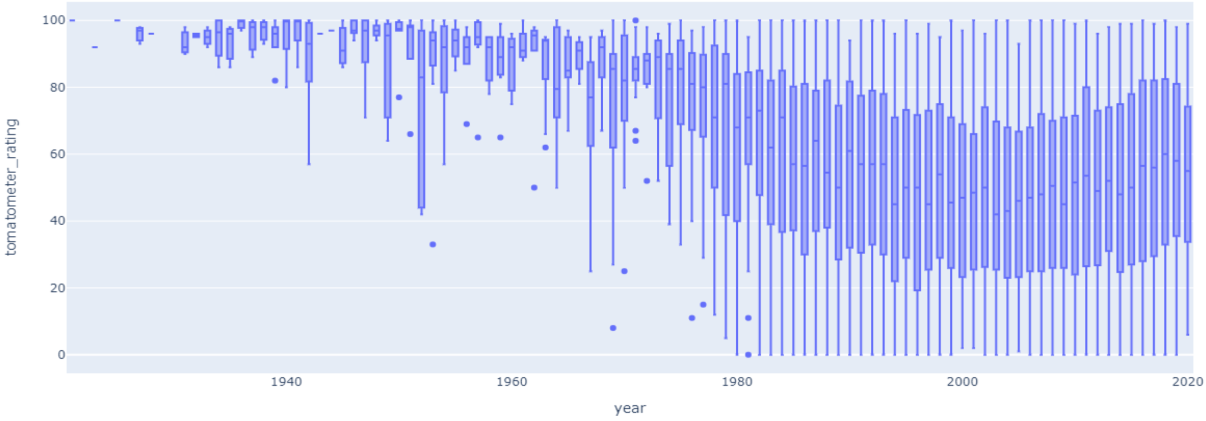
Analysis Points

Based on the above 2 plots, we can clearly following points about genre preference according to two rating systems:

1. film-noir genre is more preferred in both of the rating system
2. Some genres like Animation, War, Western are equally preferred in both rating systems.
3. Overall it seems that in tomatometer rating(usually most of the genres are rated around 50) the genre is rated less than Imdb ratings(usually most of the genres are rated around 60).
4. A good difference in rating can be observed for following genre (usually for these genres, imdb mean rating is higher than rotten tomato mean rating)
   - Action
   - Comedy
   - Crime
   - Family
   - Thriller

Rating preference across year released - are old movies liked more or newer movies are liked more

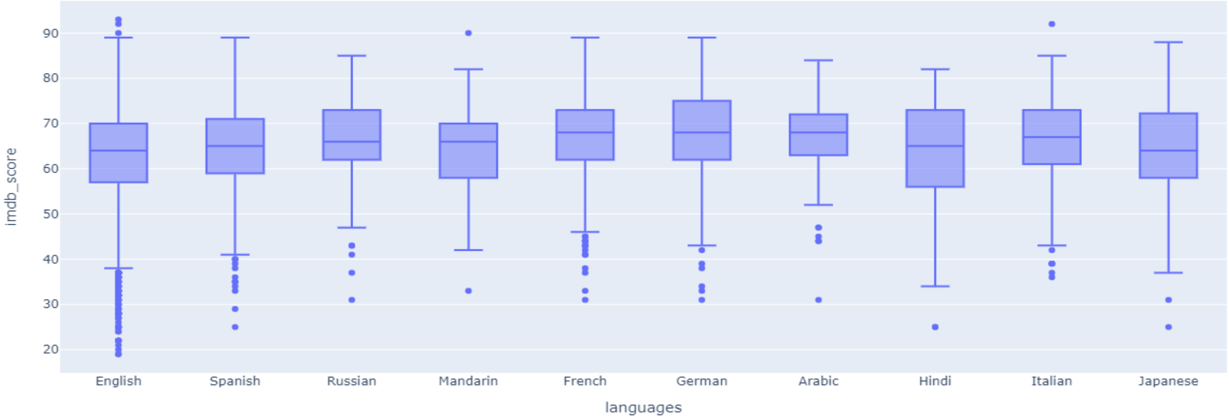


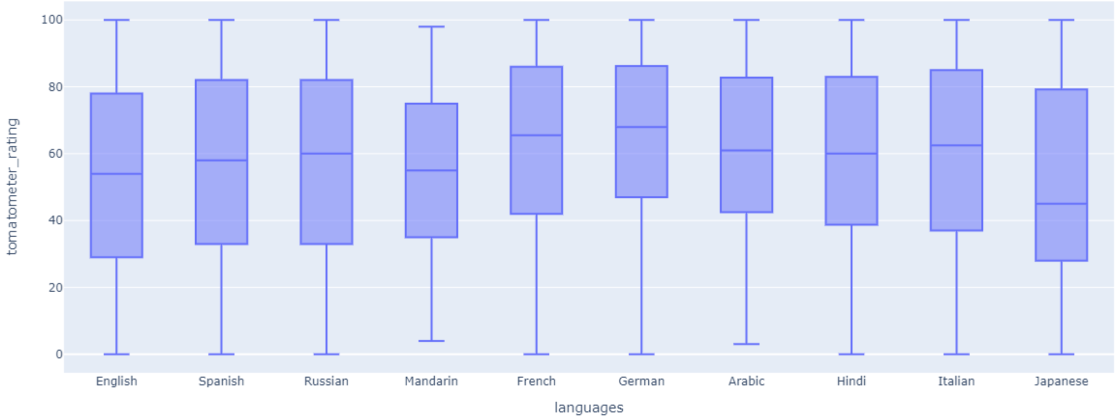


Analysis

- For imdb score , no film released before 1967 was given a score less than 50. for films released after 1979 audience has given scores as low as 0, therefore films released before 1967 can be shown again to audience and they will like them,as for films released later than 1975,there are almost equal number of movies in each band of 10 from 30-80 i.e. audience liked many movies and disliked many movies. Audience likes all the old movies in the dataset.
- For rotten tomatoes critics have judged older movies(released before 1960) linently or they have given good ratings to them , but for films released after 1960 has extreme ratings ranging all the way from 0-100. critics have rated extremely for films released later than 1960s,so they prefer older movies more.

Rating preference across language - Comparison of audience preference across languages

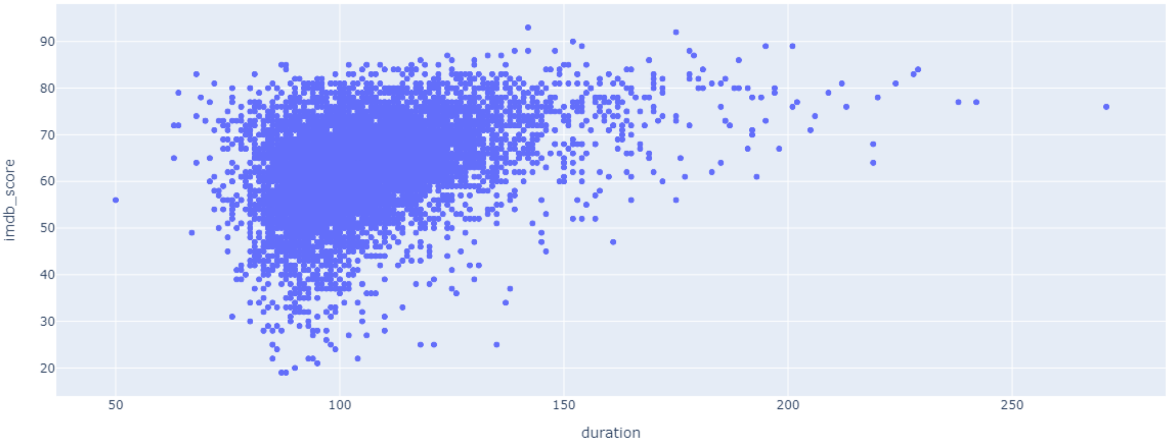


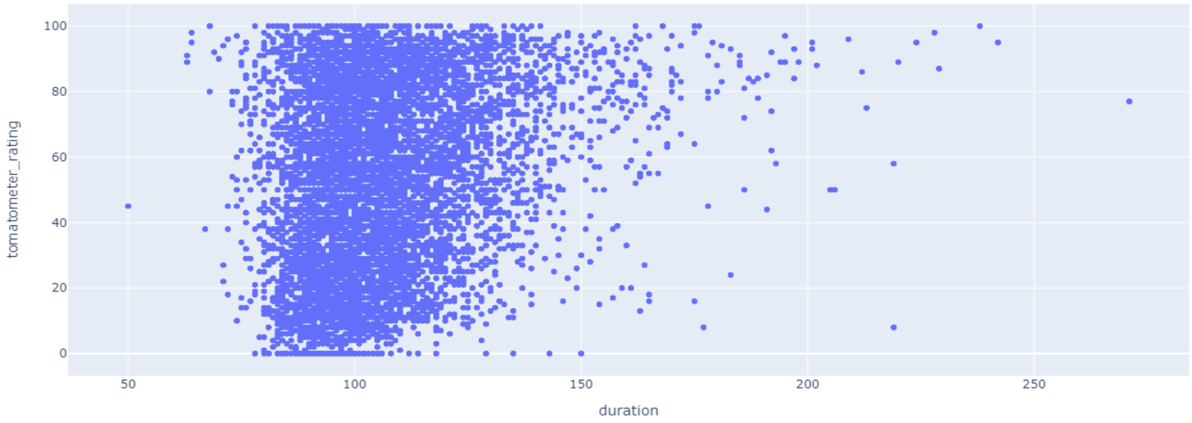


Analysis:

1. In imdb box plots,most of the movies in all languages are scored between 55-75 which is a good score .all languages have outlier movies below the minimum but only english ,mandarin and italian have outliers above maximum and that is 90+ here ,whereas english is the only language getting score less than 20
2. For tomatometer ratings most of the movies in all languages are scored between 25-85 which shows a wide spread,there are no outliers,overall, scoring is extreme ranging from 0-100.

<u>Rating preference across duration</u> - long duration movies are liked or disliked



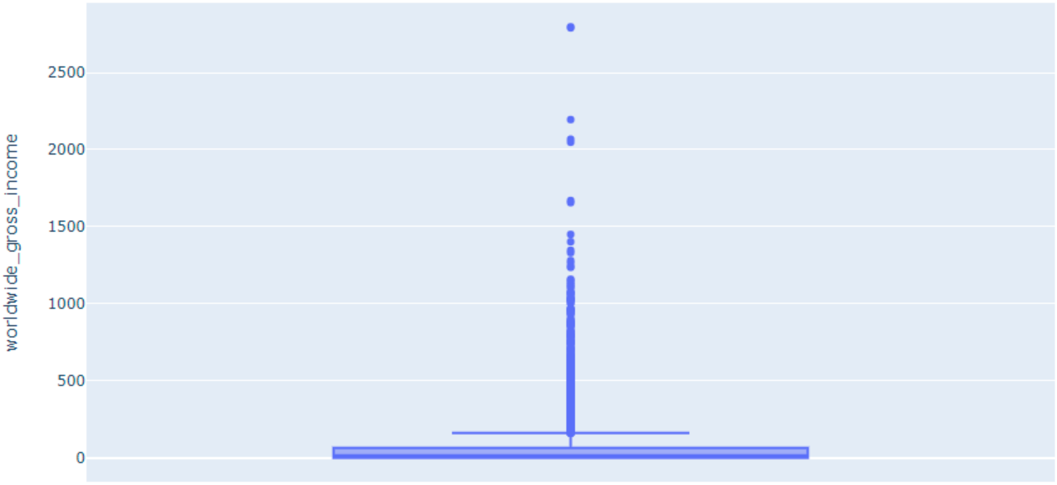


Analysis:

- Most of the movies have a duration between 60 - 160 minutes.
- For imdb scores ,the audience likes all movies of all duration type.there is some positive relation between longer duration movies and the high preference for it by audience.
- For critics ratings, they have rated most of the movies between 60 -160 minutes extremely giving 0 to 100 rating to movies, we also see critics have given high ratings for movies longer than 180 minutes.

Assessment of the movie earning potential

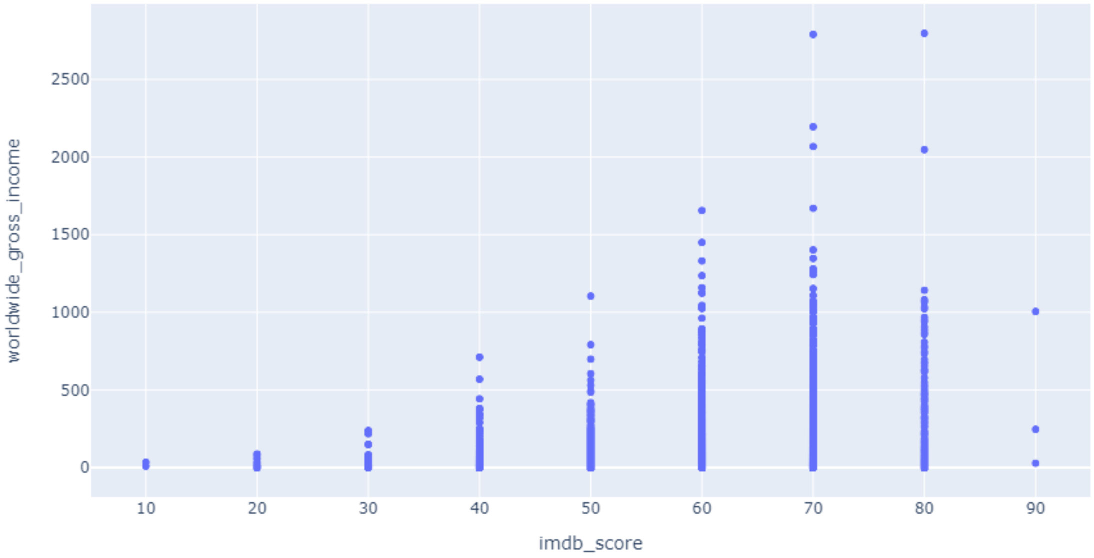
Using worldwide_gross_income column data we will analyse following points-

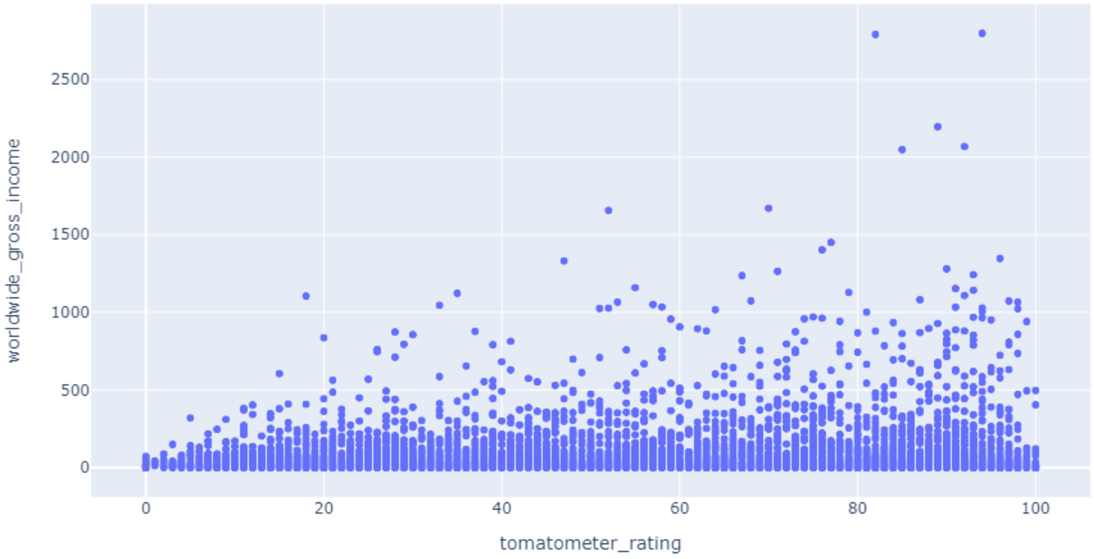


Analysis:

From the above box plot we see that half the movies earned less than 16 million dollars as 16 is the median but there are a large number of outliers in the data which have earned much more than others,but very few movies have earned more than 1500 million.

Earning potential Vs Rating Comparison - does movies with higher rating have higher earning potential





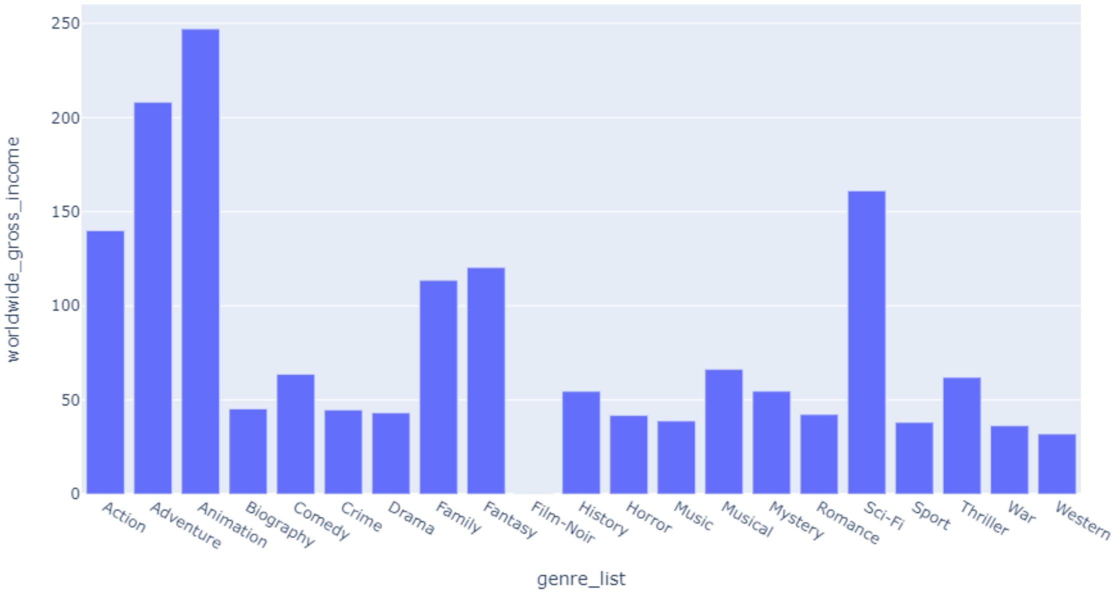
Analysis Points

- if we see the first plot of gross income and imdb score, it shows that there is a bit of a relation between income and score values. As the score increases, so does the earnings of movies.
- The second plot shows a much weaker relation between tomatometer_rating and worldwide_gross_income.
- We may have higher relation in case of imdb and not in case of tomatometer because imdb rating is influenced by layman audience while tomatometer is

more influenced by critics. And it is the layman audience who pays for movies, so that is why we see that higher earning movies are usually rated higher.

- Movies with very high earnings are usually rated better in both rating systems for eg. movies like 'Avengers','Avatar','Titanic' have very high earnings and they are rated good in both the rating system.

Earning potential Vs Genre Comparison - does genre affect the earning of a movie



Analysis Points:

Here we can clearly see that some genres have very earning potential. Most of them have mediocre earning potential.

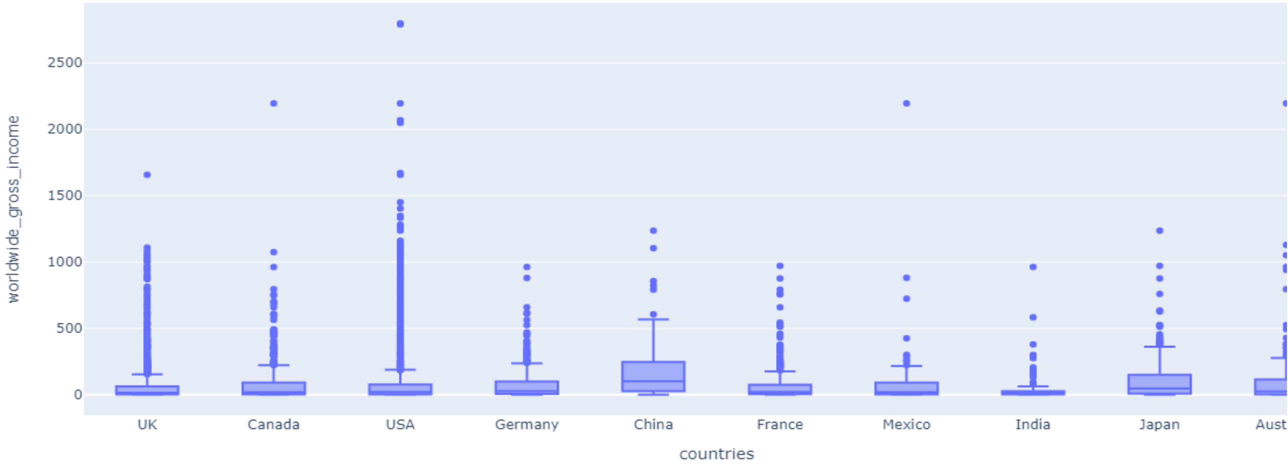
Genres with earning potential more than 100 million.

- Action
- Adventure
- Animation
- Family
- Fantasy
- Sci-Fi

Most of the other genres have mediocre earning potential of around 50 million.

Film-noir is the genre with very little earning potential.

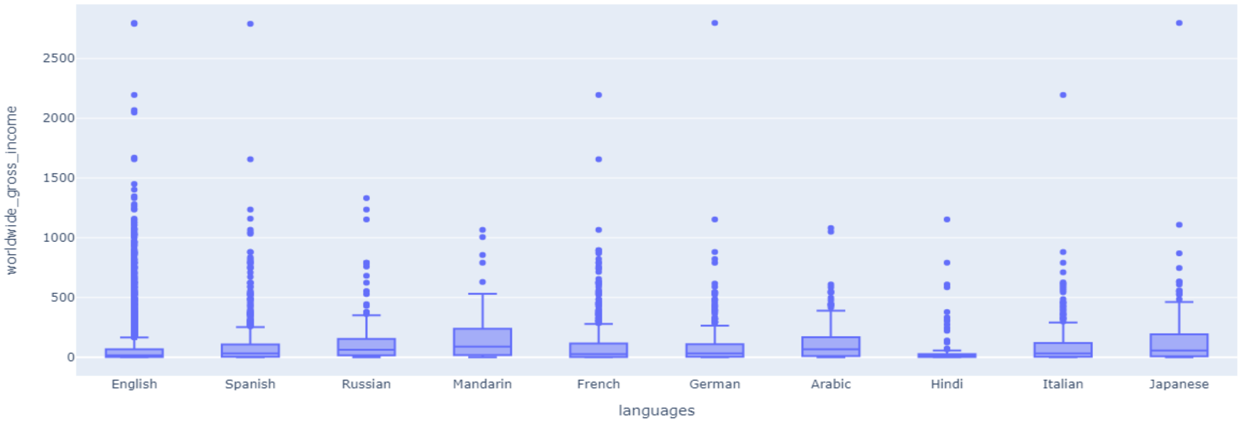
Earning potential Vs Country Comparison - does the earning of a movie differ based on the country it is getting released into.



Analysis:

- China and Japan has big boxes relative to others which means most of the movies made in China and Japan earns in this box range,which means average movies will earn more in these 2 countries as compared to others.
- Only USA,UK,Canada,Mexico,Australia have movies earning more than 1500 million dollars worldwide.
- Indian movies have many outliers but could not cross $1000 million in earnings.this maybe due to the fact that most of the viewers are Indians and pay in rupees,also Indian movies are not produced worldwide.
- USA has highest number of outlier movies and only country where movies could earn more than $2000 million

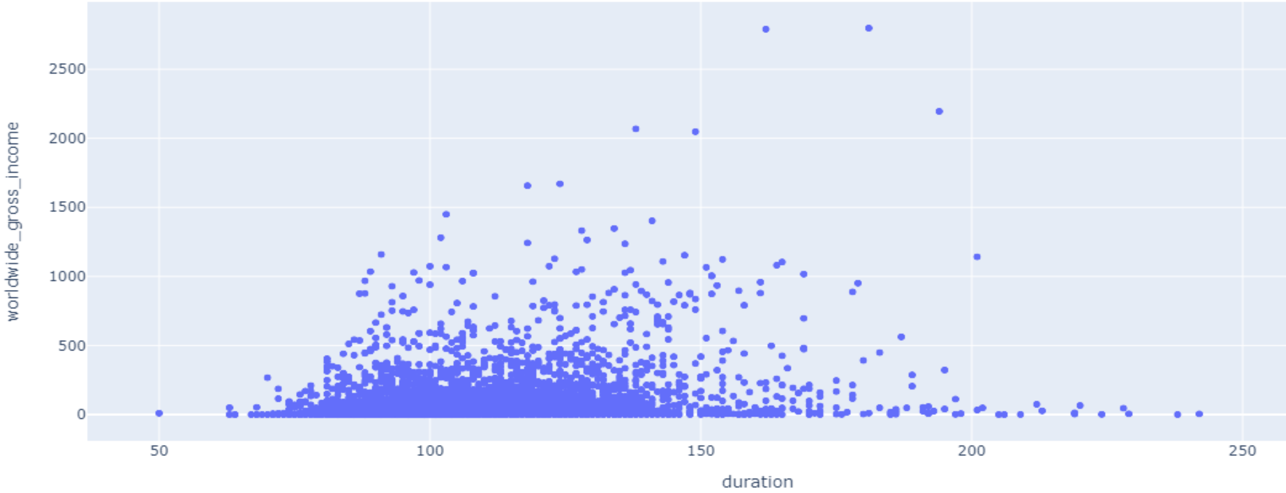
Earning potential Vs language Comparison - how does language affect earning of a movie



Analysis

1. All these languages have outlier earning movies. Most movies in these languages earn less than $500 million but outliers are earning much more than other languages and there are many outliers in these languages.
2. English and hindi languages have very short boxes implying there are very few movies in the box whereas there large number of outliers
3. English ,Spanish,German, and Japanese are only languages in which movies earned more than $2000 million.

Earning potential Vs duration Comparison - do the earning of high and low duration movie differ



Analysis:

1. Movies earning more than $1000 million are at least longer than 88 minutes .
2. Most of the movies are between 60-150 minutes and earned less than $500 million.
3. Movies earning more than $1500 million are at least 118 minutes in duration with Avengers-endgame being the highest grosser with 181 minutes .

STEP 6: Solution and Presentation

For this step we have made a dashboard showing imdb and tomatometer ratings based on income,year and genres, link for the dashboard is shared below:

https://ali-dashboard-movies.herokuapp.com/

- We have also prepared a powerpoint presentation for the business stakeholders so that they are able to understand the insights from the analysis in an easy language.

Future Work : We are creating a new dashboard which shows movies by top 10 directors in which we can filter out movies based on year and worldwide gross income. We will deploy this dashboard on a public server using Heroku app.