Project 5- IMDB Movie Analysis

By Visharad Singh Chauhan

Project Description: For your Final Project, we are providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.
We can do this by asking 'What?' This is where you frame the problem i.e. what is the problem?
Use these questions to guide your thinking:

- What do you see happening?
- What is your hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Answering these questions will help you define a problem you are trying to solve and will allow you to find the right data to solve it.

Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

Make sure to use 5 Whys Analysis in your analysis and use this to create a report which conveys a data story.

Once you have framed the problem and gathered initial insights from the data, you can ask the following questions as you dig deeper into your analysis.

- What do you see happening?
- What are the specific symptoms of the problem?
- What is your hypothesis for the cause of the problem?

Five 'Whys' approach
Once you have the problem better defined, you can use 5 Whys technique to determine its root cause by repeatedly asking the question “Why”.
It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here’s an example of how this technique could be used to figure out the cause of the following problem: A business went over budget on a recent project.

Q: "Why did we go over budget on our project?"
A: It took much longer than we expected to complete.

Q: "Why did it take longer than expected to complete?
A: We had to redesign several elements of the product.

Q: "Why did we have to redesign elements of the product?"
A: Features of the product were confusing to use.

Q: "Why were the features of the product confusing to use?"
A: We made incorrect assumptions about what users wanted.

Q: "Why did we make incorrect assumptions about what users wanted?"
A: Our user experience research team didn't ask effective questions.

As you see above, what looked like a budgeting problem turned out to be a problem with the user experience team not working effectively.

While asking Why is easy, what we're interested in is the answer. Each time you answer why the next time gets more difficult as you must think deeper behind the reasons for this. As you ask why, you may find that you have multiple answers for the same question.

You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

A. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
   Your task: Clean the data

B. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
   Your task: Find the movies with the highest profit?

C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.
   Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
   Your task: Find IMDB Top 250

D. **Best Directors:** TGroup the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

F. **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio,
and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
Append the rows of all these columns and store them in a new column named Combined
Group the combined column using the actor_1_name column.
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.
Your task: Find the critic-favorite and audience-favorite actors

Cleaning the data: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

This is the most imoprtant step of any Data analysis project. The process included in this step varies from question to question and Dataset to Dataset. To clean the dataset we will

1. First dropping the columns which have no use for the analysis that we will be doing

2. Columns like ‘Color’, ‘director_facebook_likes’, ‘actor_3_facebook_likes, ‘actor_2_name’, ‘actor_1_facebook_likes’, ‘cast_total_facebook_likes, ‘actor_3_name’, ‘facenumber_in_posts’, ‘plot_keywords’, ‘movie_imdb_link’, ‘content_rating’ ,’actor_2_facebook_likes’, ‘aspect_ratio’,

'movie_facebook_likes' are the columns containing lest important data for the analysis tasks provided. So, these columns needs to be deleted.

3. After dropping the these columns now we need to remove the rows from the dataset having anyone of its column value as blank/NULL using Find & Select.

4. Then we need to get rid of the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab.

The dataset is now cleaned.

Link for the dataset

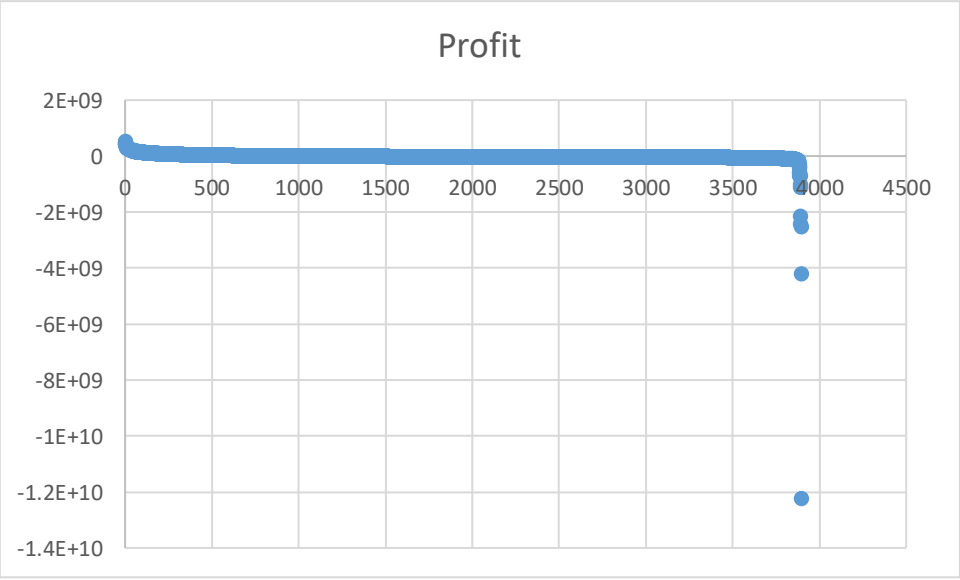
Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
Your task: Find the movies with the highest profit?

To find the movies with the highest profit: -

1. First we need to subtract the budget value from the gross value to get the profit in the profit column.

2. Sort the Profit column values in descending order, then, by using the scatter plot option we will plot values of profit(y_axis) and budget(x_axis)

3. Then with the help of graph we will be finding the outliers



A. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
Your task: Find IMDB Top 250

To find the IMDB Top 250 we will:-

1. First we will filter out those rows whose num_voted_users > 25000 using the sort and filter option

2. Then we will arrange the dataset on the basis of imdb_score in descending order

3. Then we will select only the top 250 rows for the further analysis

4. Then we will create a new column rank using the RANK() function and using the formula =RANK(N2,$N$2:$N$251,0)+COUNTIFS($N$2:N2,N2)-1 ('N' refers to the column that contains the imdb score)

5. Then we will filter out (unselect 'English') from the language column and we will get the desired output.

Visit the repository for the excel files

Best Directors: Group the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

To find the best top 10 directors on the basis of mean of imdb_score we will:-

1. First select the imdb_score column of the cleaned dataset

2. Then we will click and create a pivot table for the whole dataset

3. We will add director_name into the rows section of the pivot table

4. Then we will add average imdb_score into the values section of the pivot table

5. Then we will first sort the data on the basis of average of imdb_score in descending order and then on the basis of director name alphabetically.

It will give us the following resultant table:

Director Name	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4

Popular Genres: Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

To find the Popular Genres we will:-

1. First select the genres column of the cleaned dataset

2. Then we will go for the pivot table option

3. Then we will select the genres name as row labels

4. Then we will the values as the count of the number of genres and then sort it in descending order on the basis of count of the number of genres

Resultant table:

Genres	Count of genres
Drama	158
Comedy|Drama|Romance	152
Comedy|Drama	149
Comedy	148
Comedy|Romance	136
Drama|Romance	120
Crime|Drama|Thriller	83
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46

Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
Append the rows of all these columns and store them in a new column named Combined
Group the combined column using the actor_1_name column.
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.
Your task: Find the critic-favorite and audience-favorite actors

To find the critic-favorite and audience-favorite actors we will create a pivot table and:-

1. First, filter actors namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors from the actor_1_name column

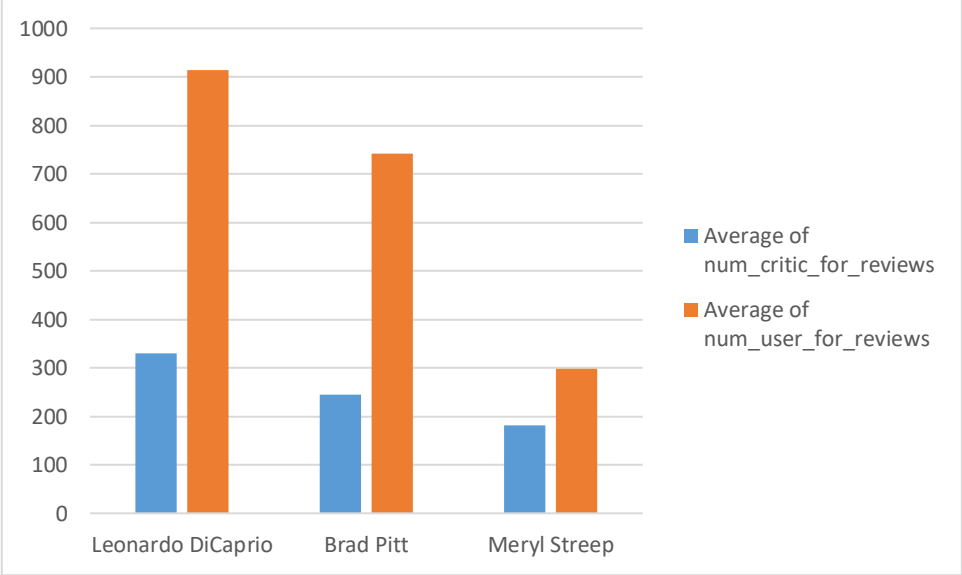
2. Then we will append the above 3 created columns into 1 column named actor_1_name_combine

3. Then we will group the 3 columns of critic-favorite and audience-favorite actors

4. Then using the pivot table we will find the average, sum and count of critic-favorite and audience-favorite actors

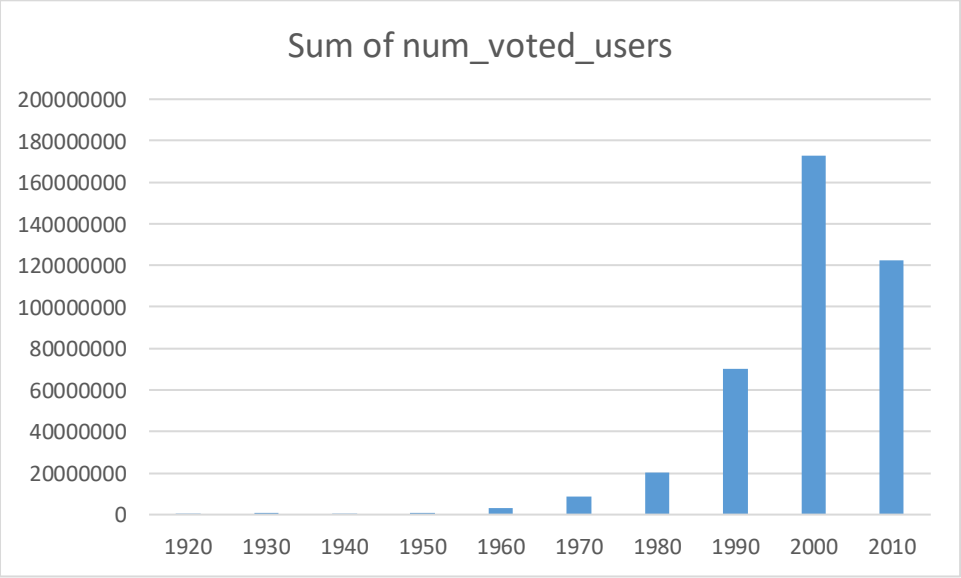
Resultant Table:

Actor_name	Average of num_critic_for_reviews	Average of num_user_for_reviews
Leonardo DiCaprio	330.1904762	914.4761905
Brad Pitt	245	742.3529412
Meryl Streep	181.4545455	297.1818182



For the decades and the data associated to them, the resultant will be:

Decade	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336
1960	2985581
1970	8704723
1980	20101705
1990	70090204
2000	173033966
2010	122492496



Tech Stack Used- MS Excel 2016

Link for the datasheets