**Project Description:**

With the dataset provided for IMDB ratings, my job is to analyze it and provide insights that can predict the success of the upcoming movies. In the later pages of this report, I will apply statistical measures on the data which will help movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects. The approach and the corresponding insights and results will be listed down in detail in the coming pages of this report.

**Tech-Stack Used:**

To implement this project, I have used Microsoft Excel 365 Version 2302. There are two reasons for this:

* MS Excel 365 allows a lot of features that are not available in the previous version such as there are a lot of automated charts which help me to quickly use them on my data without having to perform the calculations manually.
* This subscription is offered to me by the company I currently work in, so I am quite familiar with its ins and outs.

Kindly download the dataset from the link below and view in Excel to see all the charts and tables for the problems (Google sheets doesn’t show all the analytics done).

[IMDB\_movies\_analyzed](https://docs.google.com/spreadsheets/d/1ECbe2GjnIpMw1q_bkP5aXRyvl9pTqZj-/edit?usp=sharing&ouid=104611970421205783778&rtpof=true&sd=true)

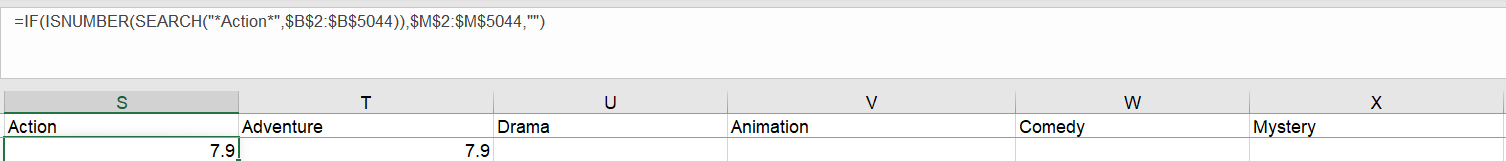
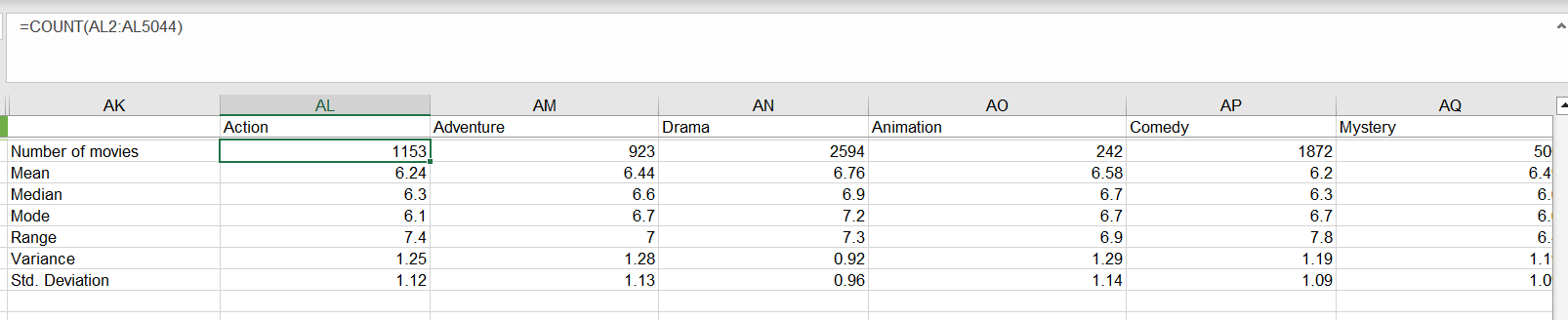
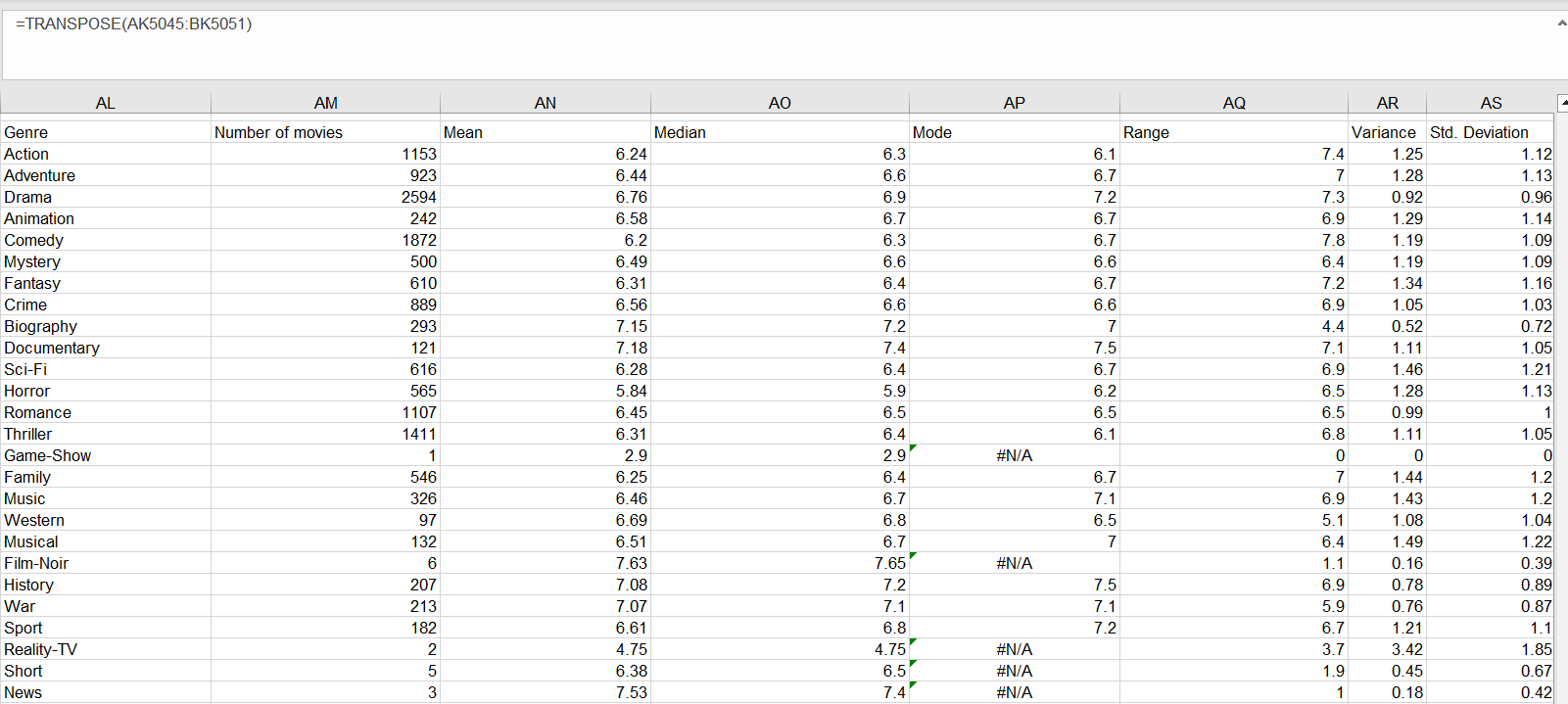
**Data Cleaning:**

1. First, we will do some visual formatting to the given dataset.
2. Look at the columns and remove those columns that are not needed for our analysis. I have removed columns such as actor\_likes, movie link, etc.
3. After the above step I get a total of 9 useful columns.

Problem Description No. 1:

Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.  
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

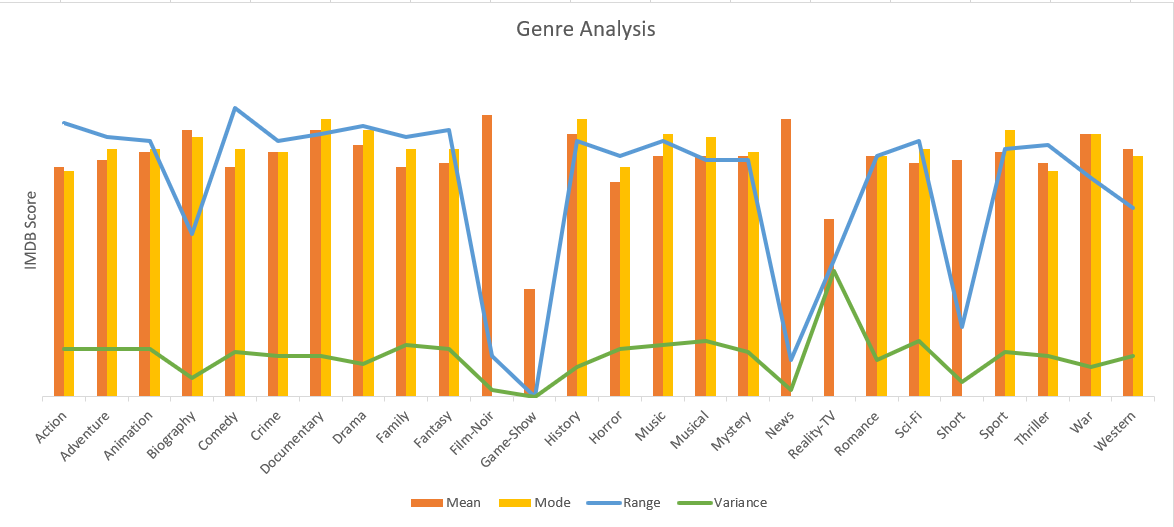
**Approach:**

1. Since we need to calculate the aggregations for each genre separately, we will separate the genres for each movie given in ‘genres’ column into genre1, genre2 and so on.
2. To do this, we will use the TEXTSPLIT() function. Since this function will require new columns, we will keep inserting new columns in the dataset provided until we see Blanks as the only option when we filter the new column created each time.
3. After the above step, we will get 8 columns for genres in total, but the total number of genres are more than 8 as we will see in the upcoming steps.
4. To create a list of unique genres, we will use 3 functions in combinations:
   1. VSTACK() – to combine the 8 columns of genres into a single column.
   2. FILTER() – to filter out the cells with “0” from the list obtained above.
   3. UNIQUE() – to select only the unique genres.
   4. TRANSPOSE() – to create a horizontal list from the result obtained above.
5. After this, we will get a total of 26 genres.
6. Next, we store the IMDB scores for each genre separately since we will need these IMDB scores for calculating more than 2 statistical measures.
7. The formula below shows how this is calculated. Since we cannot use wildcards with IF() function, I have used two more functions – SEARCH and ISNUMBER to achieve my goal.  
   
8. The SEARCH function searches the genre mentioned within double quotes in the **original** genre column and returns 1 or #VALUE! error. I have searched in the original genre column because the same genre can be present in 8 different columns if we search in columns C to J.
9. The ISNUMBER function then checks whether the result returned by SEARCH is a number or not, in our case 1 or #VALUE! And returns TRUE or FALSE.
10. Finally, the IF function takes TRUE or FALSE as its argument and returns the IMDB score from column M if TRUE or nothing if FALSE.
11. Note that we had split the genres column into separate columns so we could get the name of each genre and not miss any single one because just applying the filter in this column won’t do the job.
12. This function is copied to the other columns with respect to the different genres.
13. Once this is done, we can simply use the COUNT(), AVERAGE(), MEDIAN(), MODE(), MAX(), MIN(), VAR.P() and STDEV.P() over the respective ranges and calculate our statistical measures. I have also used ROUND to correct the result to 1 decimal place.  
    
14. Note that, in the “Mode” row we have some genres for which the MODE() function returns #NA error. This is because for these genres there are no duplicate values. Hence, the function cannot find which value is duplicated the maximum number of times. We will keep these #NA as it is since it signifies that this particular genre has not maintained a consistent imdb score.
15. Because we want to visualize only these statistical results and not the IMBD scores, we will create another range from the above result using TRANSPOSE() function:  
    
16. These values, I have copied in another sheet to visualize the results after creating a table.

**Insights:**

1. The most popular genre can be found by looking at the Number of movies column for the maximum value. Here, it is Drama followed by Comedy, Thriller, Action, and Romance.
2. But when we look at the Mean column, we find that the genre Film-Noir has the highest average imdb score, followed by the other unpopular genres such as News, Documentary, Biography.
3. Looking at the mean and median columns we see that the values are almost similar for one genre. This means that we don’t have any outliers in our dataset and that the data-spread is symmetrical.
4. Looking at the mode column, we see that the genre Documentary has the highest imdb score repeated more than once. This is different from the finding in insight 2 because even though Film-Noir has the highest imdb score overall, it is not consistent in maintaining that score.
5. Looking at the range column, we see that the genre Comedy has the largest difference in imdb scores – 7.8, followed closely by the genres Action, Drama and Fantasy. This means that the movie producers cannot rely on only the genres when it comes to analyzing the imdb scores. **Two movies with the same genre can have very different IMDB score.**
6. Finally, to look for a reliable IMDB score in terms of genre, we will look at the variance and std. deviation columns. We see that it is the **unpopular** genres such as Film-Noir, News, and Biography for which the datapoints are not far apart from the average imdb rating. This means that the movie producers can rely on the imdb scores of these genres to predict the success of movies made on these genres.

To summarize all the findings, the **most popular movie genres are Drama**, Comedy, Thriller, Action, and Romance. The **most successful genre based on imdb ratings is Biography** because the average imdb score is 7.2, it has maintained a consistent score of 7, the ratings are not too far apart from the average rating and finally the range of the lowest and highest rating is also a nominal value of 4.4.



From the visualized chart we will first traverse through the "Mean" bars. Film-Noir is the tallest, but "Mode" bar is absent which means we don't have consistent data for this genre.

Next, we can see there is News genre, but the same issue exists there - Mode bar is absent.

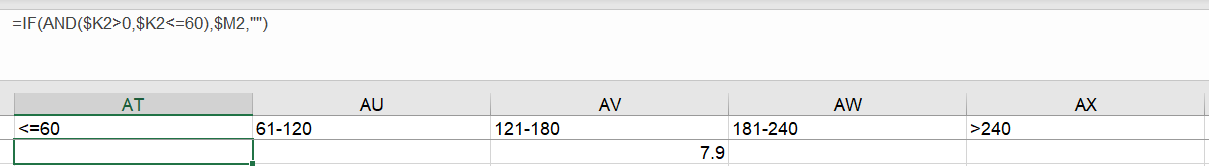
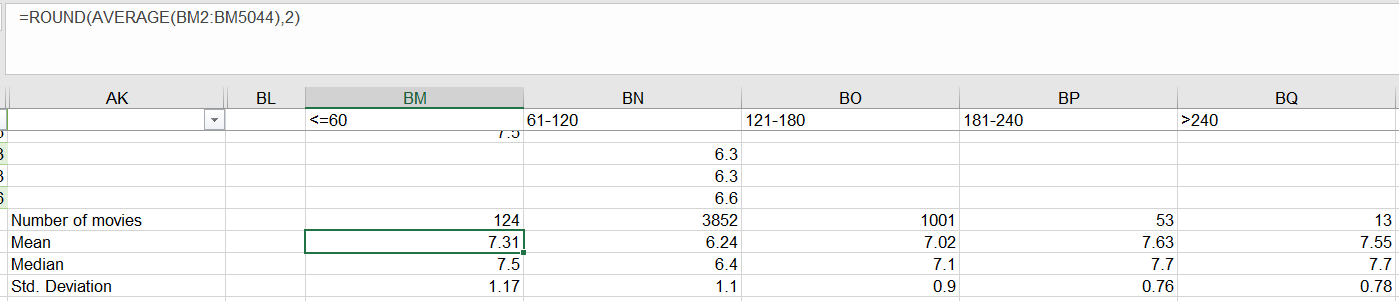
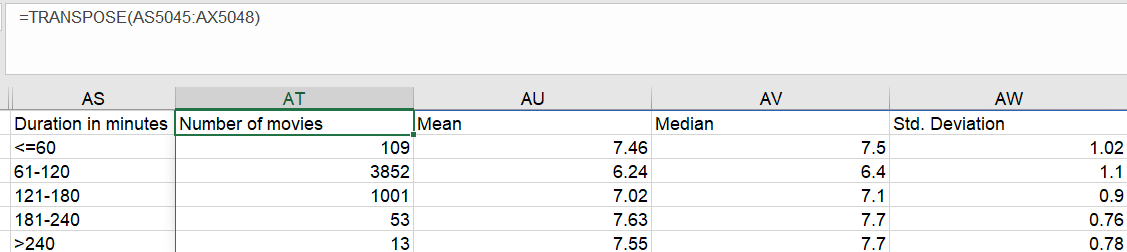
So, now we see the Biography and Documentary genres and see that even though the Mode bar is higher for Documentary, the range is also comparatively higher than Biography.

**This concludes that based on genre, Biography is the most successful genre in terms of IMDB rating.**

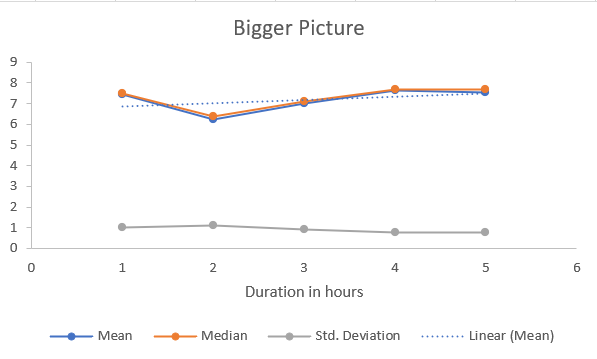
Problem Description No. 2:

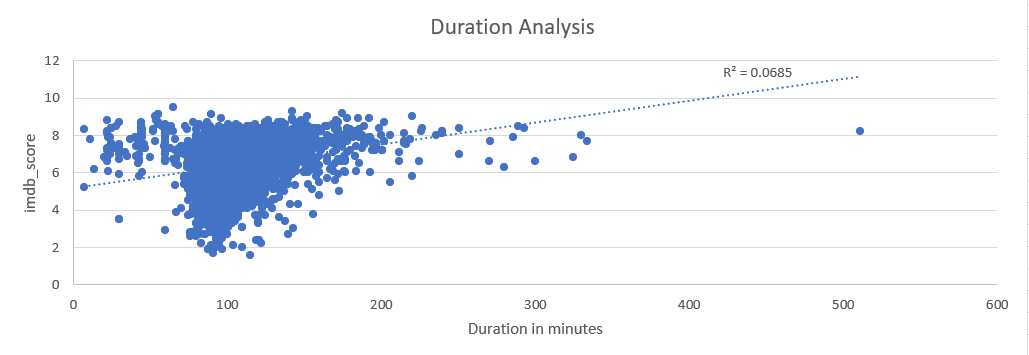
Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

**Approach:**

1. In the duration column, we see that the duration for some of the movies is missing. We will remove these movies from our analysis when we select the interval.
2. To analyze the duration of the movies, we will create class intervals to see the count of the movies according to the duration along with other statistical measures.
3. This interval will be of 60 minutes.
4. As in problem 1, we will extract the imdb scores for these intervals as shown below:
5. For the intervals <=60, 61-120, 121-180 and 181-240, I will use the IF() function together with the AND() because there are two criteria. To make copying the formulas easier, I had hidden the rows 3-5043 and then dragged down from row 2 to row 5044. After that I unhid the rows again.  
   
6. In the formula above, the column is the same, but the rows will keep changing so I have made only the column K absolute with the help of $ symbol. When both conditions are satisfied, I am taking the IMDB score from column M, if not then I am keeping the cell empty.
7. Once we have the imdb scores, we can calculate the mean, meadian and std. deviation using the excel functions as in problem 1:  
   
8. Again, since we want to visualize the statistical measures and not the IMDB scores separately, I have created a new range from these measures using the TRANSPOSE function and taken this range in a new sheet.  
   

**Insights:**

From the bigger picture chart, we can see that movies that are **between 2 to 4 hours long have a considerably higher IMDB score** as there is an increase in the average IMDB score. The std. deviation plot tells us that the duration available for the movies is not widely spread or in other words we can rely on the duration of the movies to predict the future IMDB score.  


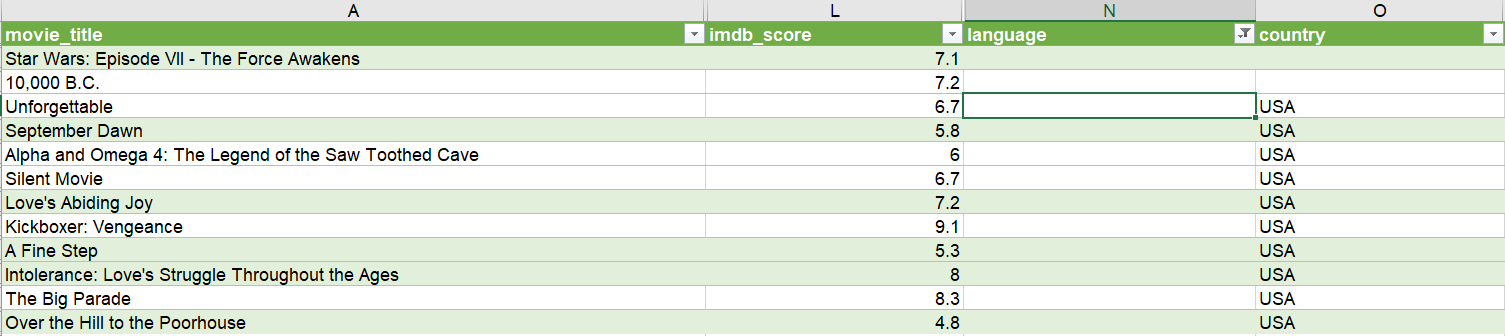
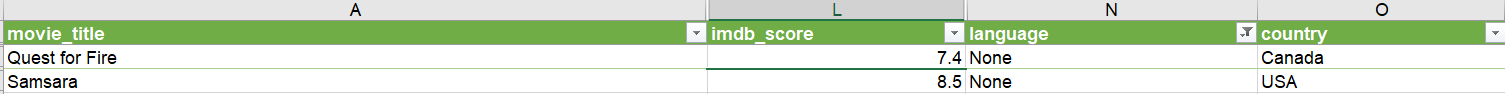
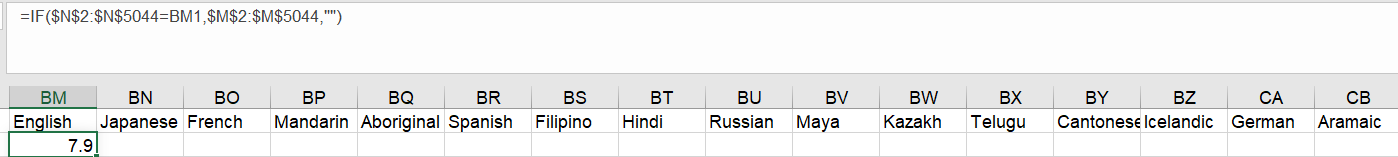
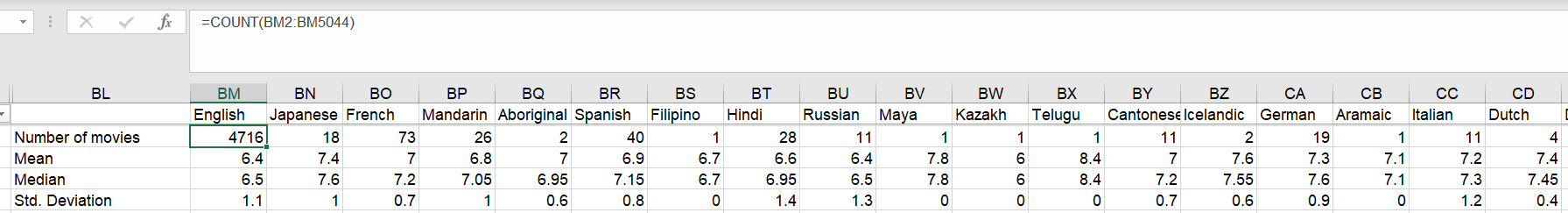
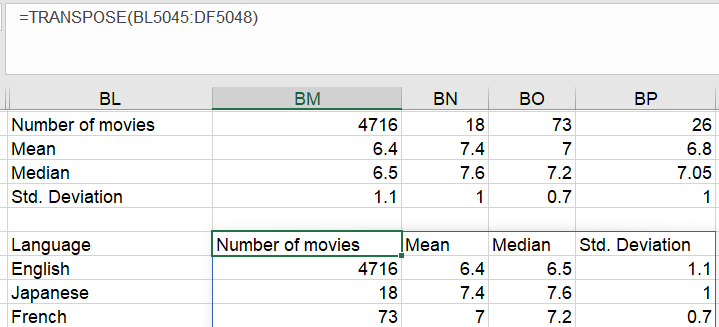
When we look at the in-depth scatter plot for **all durations** available in the dataset, we can see that there is greater density between 90-180 minutes or 1.5-3 hours long which means that most movies are between 1.5-3 hours long. The trendline shows that as the duration of a movie increases, so does its IMDB score. But the catch here is this trendline is not reliable as is evident from the R-squared value which is not close to 1.  


Therefore, to conclude, in terms of duration, movies lengthened between 2 to 4 hours will have a higher IMDB score. Specifically, those between 3-4 hours long as they have the highest IMDB average score while also the lowest std. deviation or in other words, most of the movies between 3-4 hours long have the rating close to 7.63 – the average IMDB score.  

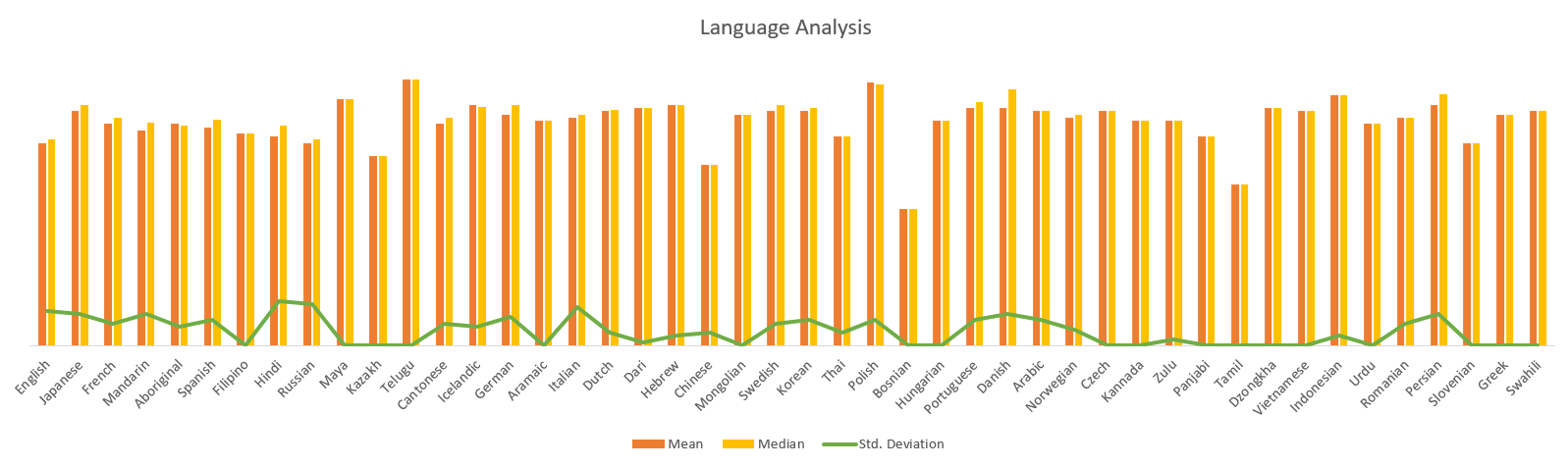

Problem Description No. 3:

Language Analysis: Situation: Examine the distribution of movies based on their language.  
Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

**Approach:**

1. In language column we can see that there are some empty cells. When we filter these out, we see that except for 2 movies, the others are released in USA. Since in the US English is the native, we will fill these empty cells with English.  
   
2. The first 2 empty cells we will ignore from our analysis.
3. We also have two movies, where language specified is “None”. We will also replace these cells with English because in both Canada and USA, the native is English.   
   
4. As in problem no. 1 we will use the UNIQUE, FILTER and TRANSPOSE functions to create a horizontal list of languages.
5. After this, we get 46 language categories and then we find out the IMDB scores as in above problems as shown below:  
     
   In the formula above, I am looking down the “language” (N) column for the categories defined in row 1 (BM1) and if it matches, I have selected the IMDB score from column M else I have kept the cell as empty. Since my looking range and the value selection range is the same, I have made it absolute using $ symbol. The category changes according to the looking value so I have kept it as it is and copied the formula towards the right.
6. Once we have the IMDB scores, we can simply use the COUNT(), MEAN(), MEDIAN() and STDEV.P() over the respective ranges and calculate our statistical measures. I have also used ROUND to correct the result to 1 decimal place.  
   
7. And finally, since for visualizing the results we need only these statistical findings and not the separate IMDB scores, I will copy this range into a new sheet by using TRANSPOSE() function.  
   

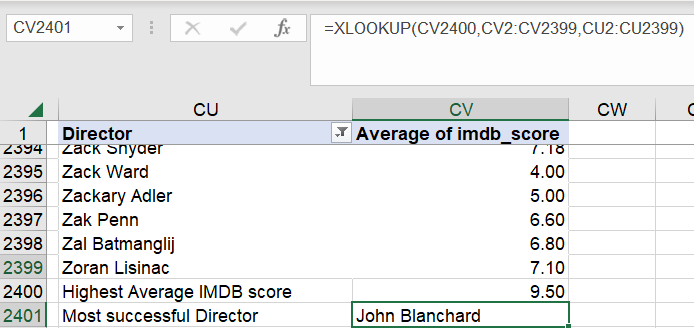
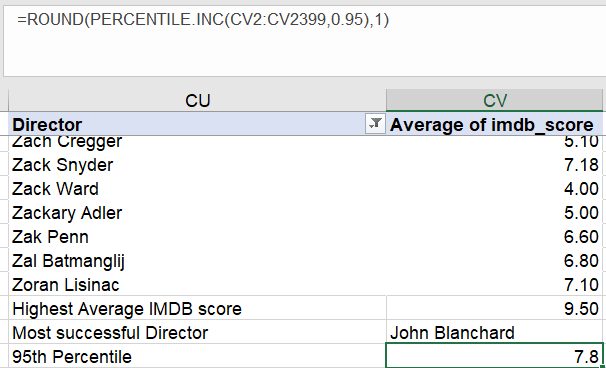
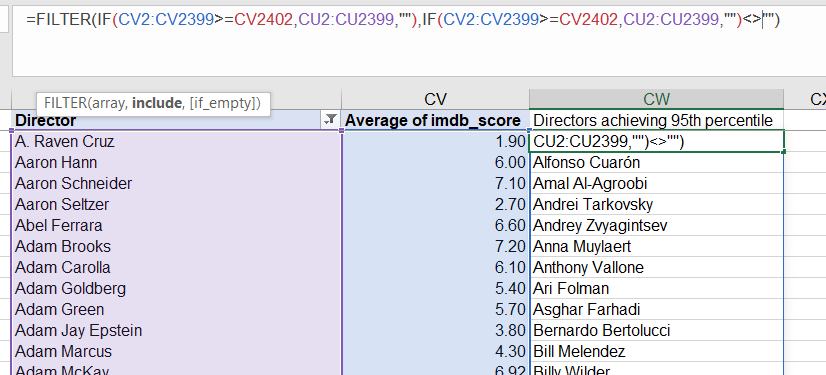
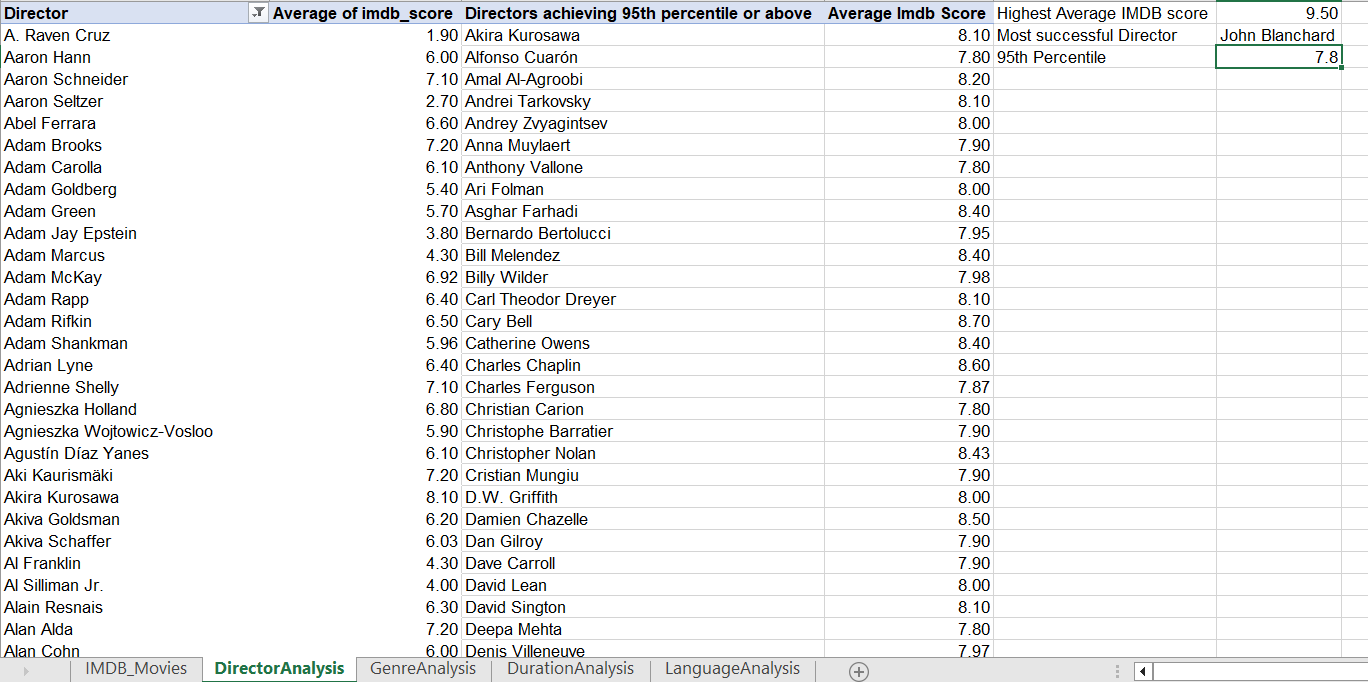
**Insights:**

1. From the first look at the result set, we can see that English is most common language across movies. No other language is even close to English.
2. Next, we will plot the mean, median and std. deviation to see if English is also the language which is the most successful in terms of IMDB score.  
   
3. From the chart above we have the following findings:
   1. When we look at the “Mean”/”Median” bars, we see that though English is the most common language, it is in fact not the most successful in terms of IMDB score.
   2. We see a close competition between Polish and Telugu for the most successful language.
   3. To choose the winner, we will look at the std. deviation green line and see that while both these languages have a close average IMDB score - evident from the mean, the std. deviation value or the reliability value is better for Telugu (0) as compared to Polish (0.8). Hence, **movies made in Telugu language have shown the most successful IMDB score**.

Problem Description No. 4:

Director Analysis: Influence of directors on movie ratings.  
Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

**Approach:**

1. In the director\_name column also, we have a lot of missing names. We will ignore these from our analysis.
2. We will create a pivot table this time with director\_name and imdb\_score since we don’t need the imdb scores to perform other statistical calculations.
3. Next, we will find the highest average imdb score with the help of MAX() function and with this score, the most successful director with the help of XLOOKUP() function:  
   
4. Since, the highest average imdb score is 9.5, we will find the 95th percentile with the help of PERCENTILE.INC() function over the average imdb score range. I have corrected the result to 1 decimal place:  
   
5. Next, we will see which directors have achieved this 95th percentile mark with the help of IF() function:  
     
   In the above formula, I have selected the range of directors if they have average imdb score of >= 95th percentile, i.e., 7.8 and later using FILTER() function, filtered out the blank cells.
6. In total out of 2398 directors, we get only 119 who have achieved an imdb score of 95th or above percentile.
7. I have moved this pivot table to a new sheet for deriving insights:  
   

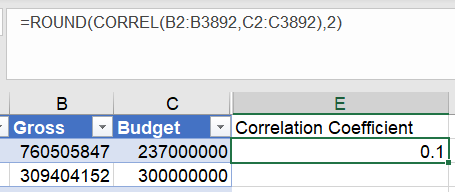
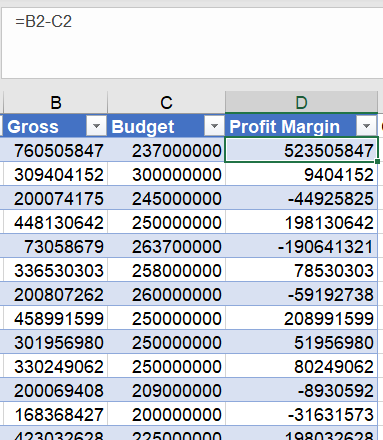
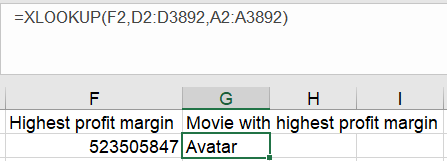
**Insights:**

The most successful director in the dataset given is John Blanchard with an average IMDB score of 9.5. There are only 119 directors out of 2398 which have an average IMDB rating of 95th percentile or above.

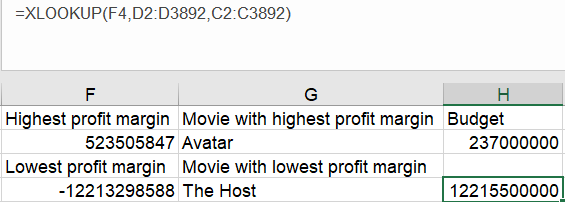
Problem Description No. 5:

Budget Analysis: Explore the relationship between movie budgets and their financial success.  
Task: Analyze the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

**Approach:**

1. There are some movies for which the budget and/or gross values are missing, we will skip these movies from our analysis with the help of applying filters first on the “gross” column and next on the “budget” column.
2. Then we will copy the resultant records in a new sheet along with the movie names. We will be left with 3891 records in total
3. To analyze the success of a movie based on its budget we will find the correlation between the budget incurred and the gross accumulated with the help of CORREL() function.  
     
   I have corrected the result to two decimal places.
4. The correlation coefficient 0.1 tells us that there is a weak relation between the budget of a movie and its success.
5. Profit margin can be simply calculated by subtracting the budget from the gross earnings [columnB- columnC]  
   
6. To find the movie with the highest profit margin, we will apply MAX() function over this resultant column D and lookup this value to return the movie name with the help of XLOOKUP() function:  
   

**Insights:**

When we look at the correlation coefficient, we see that the value is 0.1 which implies that there isn’t any relation between what budget has been spent on the making of a movie and its success in terms of gross earnings. In other words, **a movie can have the maximum budget but still it’s not necessary that it will secure the maximum profits too**. This we can see from the comparison of the budgets and margins of the most successful and the least successful movies:  
  
Even though the budget of “The Host” was much more than that of “Avatar”, the former has incurred huge losses.

**Data Story:**

If we summarize all 5 findings, we can create a short story here for the movie producers, directors, and investors to predict the success of their future projects.

**Why does Biography genre have the most reliable IMDB rating?** – because of its authenticity which pleases the viewers.

Why does authenticity please the viewers? – Because they connect with the characters more when they are more human with flaws rather than idolized.

Why does viewer opinion matter? – Because they rate the movies leading to greater influx of other viewers if these ratings are positive leading to a better gross earning.

**Why do movies within 2-4 hours have a better IMDB rating?** – because they can better involve the viewers and allow them time to become a part of the story.

Why do viewers want to become a part of the story? – so they can feel that through the movie they are living their own life. It gives them a sense of satisfaction when they feel included rather than just watching the movie from a fence which generally happens in case of movies with lesser duration.

Why do movies more than 4 hours have lesser IMDB rating? – because too lengthy a movie can fail to grasp viewers’ attention span for too long, they might feel distracted or will generally not have that much time to invest in a saga.

**Why does Telugu** **language have better IMDB rating?** – because they cater to the native viewers’ experience instead of focusing on a larger population. This incorporates authentic experiences mixed with emotional connect which is not only appreciated by the native speakers but also creates a better impression on the other viewers which leads to better ratings overall.

**Why do directors influence movie ratings?** – When a director makes a movie which is successful, the viewers like their favourite actors also start following their favourite directors. They start watching more movies by that director which increases their popularity. These directors also get better investments from more producers and because of better budget they can create better movies. This becomes a cycle, which leads to better IMDB ratings ultimately.

**Why does budget not always decide the success of a movie?** – because as we have seen there a lot of factors which affect a movie’s success such as genre, language, duration, and director. Only after the producers have these checkboxes filled with the right variables can the budget have an impact on the success of the movie. In other words, the budget acts as an X-factor which may or may not be responsible for the overall success of a movie.

**Result:**

This project by far was the most complicated one. On the first look, it seems easy but as stated – a Data Analyst doesn’t just answer questions but predicts the future success of a project or in this case a movie. Taking the data, cleaning, and sampling it had allowed me to apply my critical thinking skills to derive better insights that fit the whole data. Visualizing the data was another challenge which I found has had a significant impact on analytical thinking. Overall, this project had covered all the aspects of using Microsoft Excel – allowing me work around the various functions and shortcuts, along with the numerous different charts and selecting the best fit that explains my story.