IMDB MOVIE ANALYSIS

Problem Statement: The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Approach:

1. Approach-1: Data Cleaning and transformation using MS Excel Power Query and data analysis using Excel pivot tables, functions, descriptive statistics. Data Visualization MS Excel/Power Bi.
2. Approach-2: Data cleaning in MS Excel/SQL, and Data analysis using MySQL Workbench. Data Visualization MS Excel/Power Bi.

I choose the Approach-1 as I've completed previous projects in MySQL, I want to get hands-on practice in MS Excel.

Tech-Stack Used: MS Excel 2022.

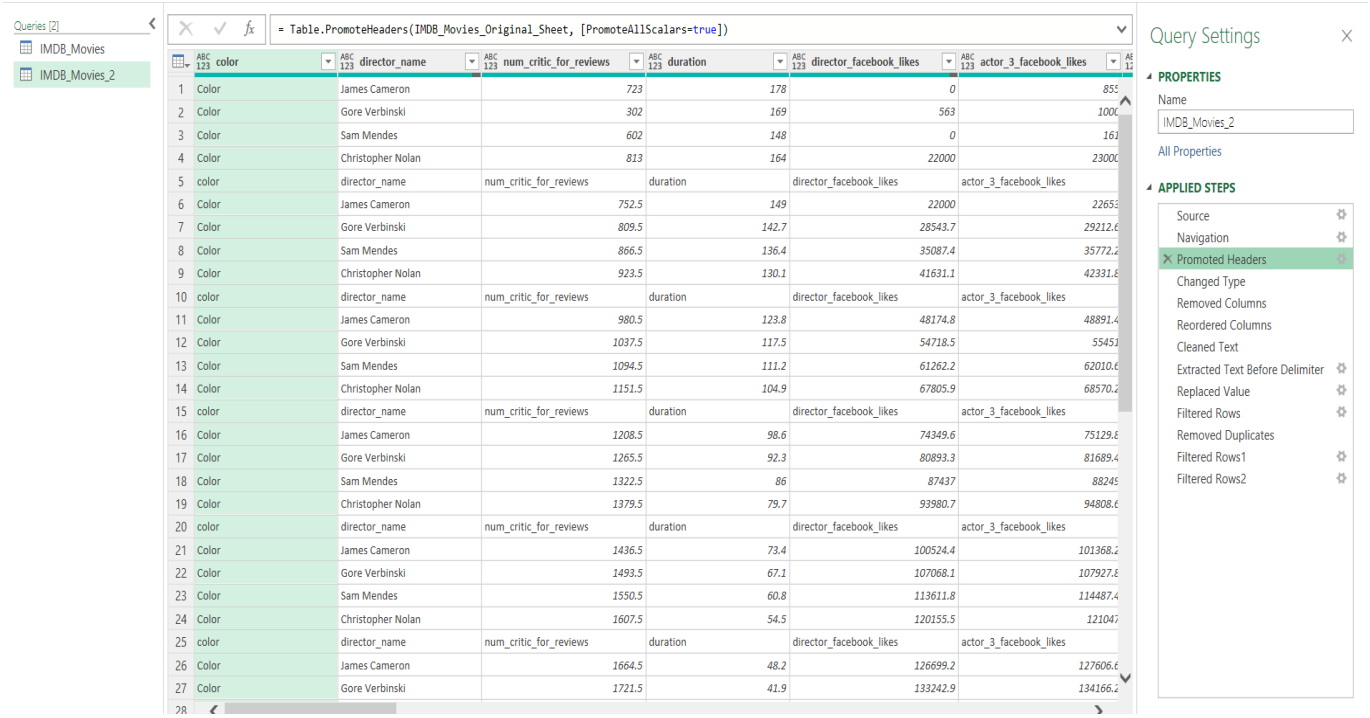
Topics Overview:

- Movie Genre Analysis.
- Movie Duration Analysis.
- Language Analysis.
- Director Analysis.
- Budget Analysis.

DATA CLEANING:

1. Drop columns which are not required for analysis.
2. Load the Data to Power Query in Excel.
3. Remove duplicate rows in the data.
4. Handle Empty Rows:
    - Rows where most of the columns are empty removed.
    - Empty rows in **Language** column are filled with **English** language.
5. Clean the Text Columns with special characters.
6. Change the data type for the columns where needed.
7. 1st normalization: Transform the column **genre** using split the column by rows option with custom delimiter "|".
8. IMDB_Movies sheet: Final rows for analysis:11232 rows after separating multiple genres.
9. IMDB_Movies_2 sheet: Final rows for analysis :3789
10. Final columns: 8 columns
    - Imdb_score
    - movie_title
    - director_name
    - genres
    - language

- duration
- gross
- budget



EXCEL SHEET ATTACHMENT:

- [https://docs.google.com/file/d/11HuxBIbceqMjGVGUq42ae2_XIaQjHzbS/edit?filetype=msexcel]

DATA ANALYSIS:

MOVIE GENRE ANALYSIS:

APPROACH:

- Here I've fetched unique rows from the column **Genre** using unique function**.**

Formula: =UNIQUE(IMDB_Movies[genres])

- Calculated count of genre in the movies using COUNT(IF()) Function, Added IFERROR() to handle the error scenario.

Formula: =IFERROR(COUNT(IF(IMDB_Movies[[#All],[genres]]='Movie Genre Analysis'!E4,IMDB_Movies[[#All],[imdb_score]])),0)

- Calculated Descriptive statistics of **Imdb_score (mean,median,mode,standard deviation,variance,range**)[Refer excel sheet for formulas].
- Highlighted top 5 using conditional formatting.
- Additionally count, mean,stdevp are calculated using pivot table.

1) Descriptive Statistics of IMDB Score for Genre

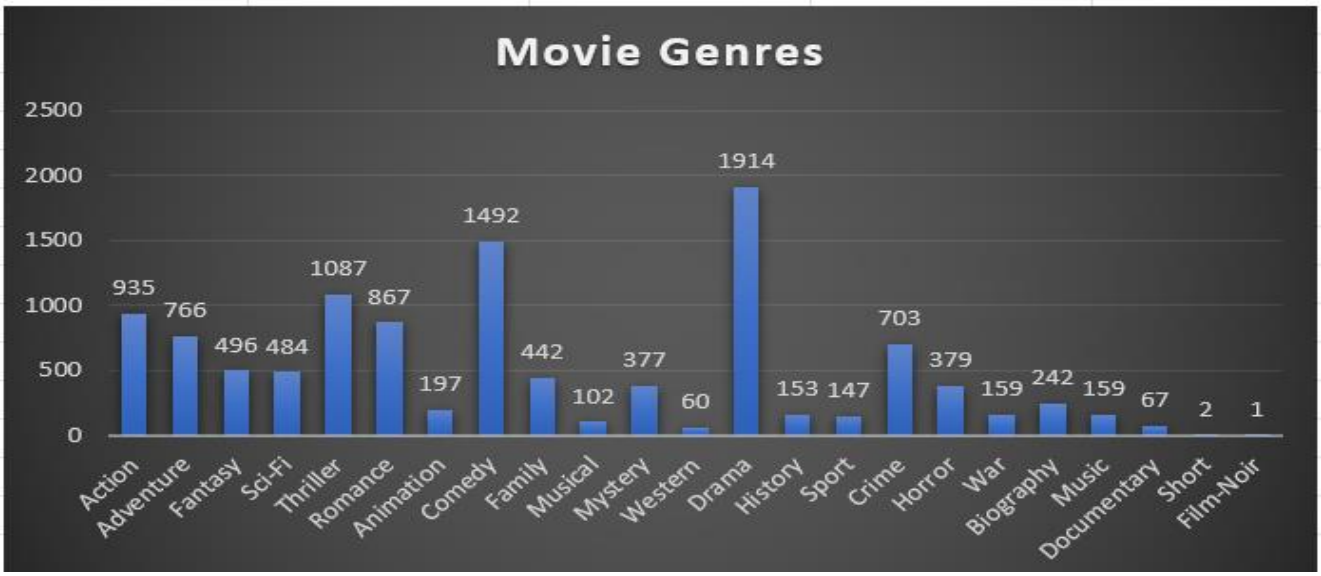
Genre	Count of Genre	Mean	Median	Mode	STDEV	VAR	RANGE
Action	935	6.29	6.3	6.6	1.04	1.25	6.9
Adventure	766	6.45	6.6	6.6	1.12	1.3	6.6
Fantasy	496	6.29	6.4	6.7	1.14	1.36	6.7
Sci-Fi	484	6.33	6.4	7	1.17	0.94	6.9
Thriller	1087	6.37	6.4	6.5	0.97	0.94	6.3
Romance	867	6.43	6.5	6.5	0.97	0.98	6.4
Animation	197	6.7	6.8	7.3	0.99	1.08	5.8
Comedy	1492	6.18	6.3	6.3	1.04	1.36	6.9
Family	442	6.2	6.3	5.4	1.17	1.29	6.7
Musical	102	6.55	6.7	7.1	1.14	1.01	6.4
Mystery	377	6.47	6.5	6.6	1.01	0.97	5.5
Western	60	6.76	6.8	6.8	0.98	0.8	4.8
Drama	1914	6.79	6.9	6.7	0.89	0.46	7.2
History	153	7.12	7.2	7.7	0.68	1.09	3.4
Sport	147	6.6	6.8	7.2	1.04	0.97	6.4
Crime	703	6.55	6.6	6.6	0.98	0.98	6.9
Horror	379	5.9	5.9	6.2	0.99	0.65	6.3
War	159	7.05	7.1	7.1	0.81	0.5	4.3
Biography	242	7.14	7.2	7	0.71	1.46	4.4
Music	159	6.37	6.5	6.5	1.21	1.42	6.9
Documentary	67	7.01	7.2	6.6	1.19	0.09	6.9
Short	2	6.8	6.8	0	0.3	0	0.6
Film-Noir	1	7.7	7.7	0	0	0	0

2) Analysis Using pivot

Row Labels	Count of genres	Mean imdb_score	StdDevp of imdb_score
Action	935	6.29	1.04
Adventure	766	6.45	1.12
Animation	197	6.70	0.99
Biography	242	7.14	0.71
Comedy	1492	6.18	1.04
Crime	703	6.55	0.98
Documentary	67	7.01	1.19
Drama	1914	6.79	0.89
Family	442	6.20	1.17
Fantasy	496	6.29	1.14
Film-Noir	1	7.70	0.00
History	153	7.12	0.68
Horror	379	5.90	0.99
Music	159	6.37	1.21
Musical	102	6.55	1.14
Mystery	377	6.47	1.01
Romance	867	6.43	0.97
Sci-Fi	484	6.33	1.17
Short	2	6.80	0.30
Sport	147	6.60	1.04
Thriller	1087	6.37	0.97
War	159	7.05	0.81
Western	60	6.76	0.98
Grand Total	**11231**	**6.46**	**1.04**

INSIGHTS:

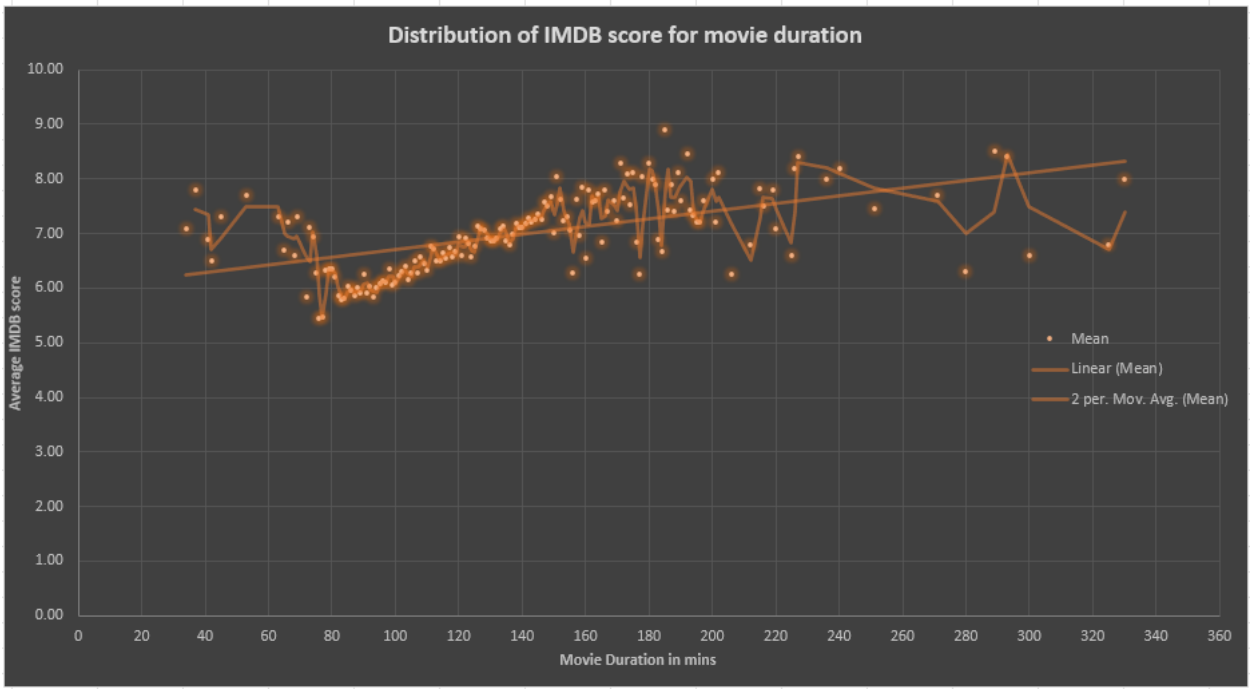
- From the analysis, most of the movies are from this genre **Drama, Comedy.** Genre like **Short, Film-Noir** has very least number of movies.
- By calculating mean, median, we can identify that although **Biography** genre has highest rating it is considered as an outlier and might have impact on further analysis.
- Average movie ratings for the genres are between 5.9-7.14.
- From the mode calculation, the genre that appears most frequently in the movies is **Animation, History, sport** have highest rating. Most of the genres have consistent rating. **Short, Film-Noi**r can be considered as outliers.
- High variance and standard deviation indicate that, Genre has mixed review with both highest and lowest ratings included. From this we can identify genres like **Music, Fantasy, Comedy, Adventure** tends to have mixed reviews.
- There are also many genres with consistent ratings based on range.

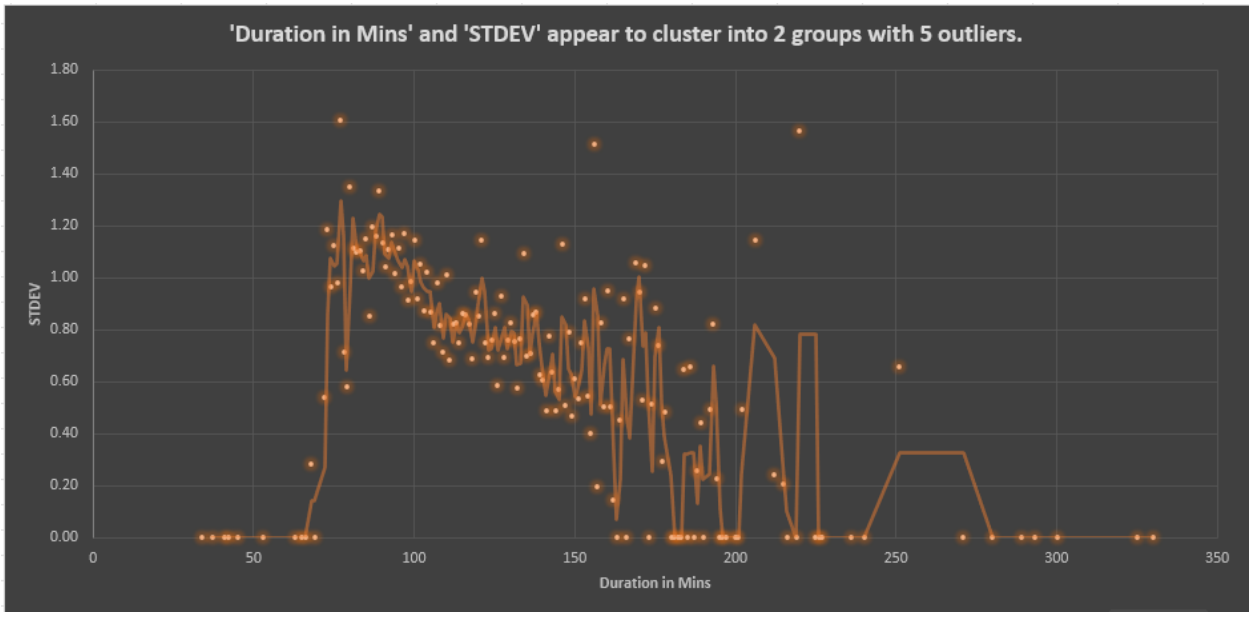


MOVIE DURATION ANALYSIS:

APPROACH:

- Here I've fetched unique rows from the column **duration** using unique function**.** Formula: =UNIQUE(IMDB_Movies_2[duration])
- Calculated Descriptive statistics of **Imdb_score (mean,median,standard deviation**)[Refer excel sheet for formulas].
- Highlighted top 5 using conditional formatting.
- Plotted scatter plot chart for movie duration in mins and average imdb_score.





INSIGHTS:

- From this Scatter plot chart, we can identify the group where more values are distributed, in this case we can say that duration of 80 mins to 150 mins can be considered as a group, this indicates the movies with similar duration and imdb_score.

- Duration of 185 mins, with highest mean, duration between 70-80 mins have least mean can be considered as a potential outlier.
- From the second chart, durations with mins like 68,77,156,220,251 can be potential outliers.

LANGUAGE ANALYSIS:

APPROACH:

- Here I've fetched unique rows from the column **language** using unique function**.** Formula: =UNIQUE(IMDB_Movies_2[language])
- Calculated Descriptive statistics of **Imdb_score (mean,median,standard deviation**)[Refer excel sheet for formulas].
- Highlighted top 3 using conditional formatting.

Language statistics				
Language	Count	Mean	Median	STDEV
English	3609	6.42	6.5	1.04
Mandarin	14	7.02	7.25	0.74
Aboriginal	2	6.95	7.5	0.54
Spanish	26	7.05	7.2	0.82
French	37	7.29	7.2	0.54
Filipino	1	6.7	6.7	0
Maya	1	7.8	7.8	0
Kazakh	1	6	6	0
Telugu	1	8.4	8.4	0
Cantonese	8	7.24	7.3	0.38
Japanese	12	7.63	7.7	0.88
Aramaic	1	7.1	7.1	0
Italian	7	7.19	6.9	1.12
Dutch	3	7.57	7.8	0.33
Dari	2	7.5	7.5	0.1
German	13	7.69	7.7	0.59
Mongolian	1	7.3	7.3	0
Thai	3	6.63	6.6	0.3
Bosnian	1	4.3	4.3	0
Korean	4	7.88	7.5	0.53
Hungarian	1	7.1	7.1	0
Hindi	10	6.76	7.3	1
Icelandic	1	6.9	6.9	0
Danish	3	7.9	7.3	0.43
Portuguese	5	7.76	8.1	0.68
Norwegian	4	7.15	7	0.45
Czech	1	7.4	7.4	0
Russian	1	6.5	6.5	0
None	1	8.5	8.5	0
Zulu	1	7.3	7.3	0
Hebrew	3	7.5	8	0.37
Dzongkha	1	7.5	7.5	0
Arabic	1	7.2	7.2	0
Vietnamese	1	7.4	7.4	0
Indonesian	2	7.9	8.2	0.3
Romanian	1	7.9	7.9	0
Persian	3	8.13	8.4	0.38
Swedish	1	7.6	7.6	0

INSIGHTS:

- From this analysis we can identify the frequency of language, English (3609) movies has High popularity compared to other languages, followed by French (37), Spanish (26) as 'Medium popularity'.
- From the statistical derivatives we can understand that most of the movie ratings are consistent.
- Regional ratings also impact the Language analysis.

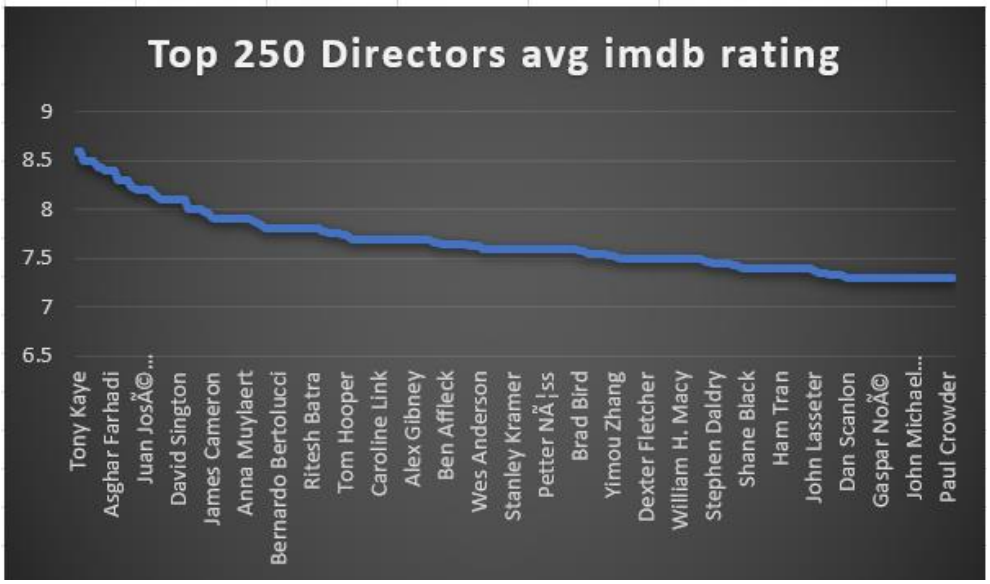
- While English, Hindi and Italian also have high STDEV, indicating wide range of ratings.

DIRECTOR ANALYSIS:

APPROACH:

- Here I've fetched unique rows from the column **Director** using unique function**.** Formula: =UNIQUE(IMDB_Movies_2[directors])
- Calculated Descriptive statistics of **Imdb_score (mean,median,standard deviation**)[Refer excel sheet for formulas].
- Calculated 95th percentile. **=PERCENTILE.INC(IF(IMDB_Movies_2[director_name]='Director Analysis'!B4,IMDB_Movies_2[imdb_score]),0.95)**
- Calculated Top 250 directors using filter-sort functions.

Top 5 Director	95th percentile	avg imdb rating
Frank Darabont	9.18	7.98
Francis Ford Coppola	9.12	7.66
Christopher Nolan	8.93	8.43
Peter Jackson	8.86	7.89
Sergio Leone	8.85	8.43



INSIGHTS:

- From this analysis, we can find the top 5 directors who have highest average imdb rating among the other directors.
- Top 250 Directors average rating lies between 7 to above 8.5.
- Frank Darabont being the highest rated Director.
- Director vande Curtis-hall has a very least rating of 2.10.

BUDGET ANALYSIS:

APPROACH:

- Here I've fetched unique rows from the column **Movie** using unique function**.** Formula: =UNIQUE(IMDB_Movies_2[movie_title]).
- Calculated profit (gross earnings-budget).
- Calculated Correlated coefficient using CORREL Function for gross and budget. **=CORREL(D4:D3791,C4:C3791)**

correlation cofficient	0.223252814

- Calculated Top movie with max profit using max function. **=MAX(E4:E3792)**

Max Profit	
Avatar	52,35,05,847.00

- Calculated Top 5 movies, using sort and filter functions. **=SORT(FILTER(B4:E3791,E4:E3791>=LARGE(E4:E3791,5)),4,-1)**

Top 5 Movies based on profit

Movie	Gross	Budget	Profit
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	1100000000%	449935665
E.T. the Extra-Terrestrial	434949459	10500000	424449459

INSIGHTS:

- Avatar movie has highest profit of 523 million based on our analysis, followed by Jurassic world, Titanic, Star Wars Episode IV-Anew Hope, E.T.the Extra Terrestrial bagging the next top movies based on the profit.
- Lady Vengenance movie with very less profit.

RESULTS:

- **Drama, Comedy are the most popular genres in the movies.**
- **Most of the movies are in English language, non-English language movies also performed well, we can this consider because of the regional audience ratings being higher.**
- **Duration between 80-150 mins movies have consistent ratings.**
- **Avatar movie has highest profit of 523 million, Titanic highest profitable movie in the 90's.**
- **Frank Darabont Is the Director with highest imdb Score.**

ATTACHMENTS:

Excel Sheet:
https://docs.google.com/file/d/11HuxBIbceqMjGVGUq42ae2_XIaQjHzbS/edit?filetype=msexcel

Loom Video:
https://www.loom.com/share/a46933b185b04b31b1aeca539ccf5a37?sid=74406924-19af-4557-b13e-97dc27211124