# TRAINITY

## PROJECT 5

**IMDB MOVIE ANALYSIS**

**Project Description:**

* IMDB Movie Analysis Project aim is to explore and analyze the dataset of movies provided by the IMDB platform .As a Data Analyst, the task is to extract meaningful insights from the data and to provide the detailed report of the IMDB platform that can help movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.
* This dataset includes director names, movie titles, movie genre, movie duration, movie language,gross ,budget ,IMDB\_ratings and etc…

**Approach:**

* I went through the dataset and Understood the insights of it. After analyzing the dataset I thought many columns like critic\_num\_views etc… will create meaningful insights, but they are not related to the tasks given in this project. So I removed all the columns except those which are related to this project.
* In my opinion,Removing the blank cell rows may lead to loss of data. So I thought that cleaning the data according to the requirements of tasks is the best way to handle the blank cells.

1. **Movie Genre Analysis:**

**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.

* To do this task, I’ve splitted the genre column using the text\_to\_columns option in google sheets. Then I removed duplicates of Genre and found the number of movies for each genre and average, median, mode, max, min, variance,standard deviation for each genre by using excel functions.

| **GENRE\_TYPE** | **MOVIE\_COUNT** | **AVERAGE** | **MEDIAN** | **MODE** | **MAX** | **MIN** | **VARIANCE** | **Standard Deviation** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Adventure | 923 | 6.441170098 | 6.6 | 6.7 | 8.9 | 1.9 | 1.279604703 | 1.131196138 |
| Thriller | 1411 | 6.314245216 | 6.4 | 6.1 | 9 | 2.2 | 1.111619625 | 1.054333735 |
| Animation | 242 | 6.576033058 | 6.7 | 6.7 | 8.6 | 1.7 | 1.298676314 | 1.139594803 |
| Family | 546 | 6.245054945 | 6.4 | 6.7 | 8.7 | 1.7 | 1.443837887 | 1.201598055 |
| Fantasy | 610 | 6.30704918 | 6.4 | 6.7 | 8.9 | 1.7 | 1.347191607 | 1.160685835 |
| Romance | 1107 | 6.450587173 | 6.5 | 6.5 | 8.6 | 2.1 | 0.9920860021 | 0.996035141 |
| Crime | 889 | 6.564791901 | 6.6 | 6.6 | 9.3 | 2.4 | 1.053612597 | 1.02645633 |
| Comedy | 1872 | 6.195245726 | 6.3 | 6.7 | 9.5 | 1.7 | 1.189656701 | 1.090713849 |
| Drama | 2594 | 6.763762529 | 6.9 | 7.2 | 9.3 | 2 | 0.9165266786 | 0.9573539986 |
| Sci-Fi | 616 | 6.281818182 | 6.4 | 6.7 | 8.8 | 1.9 | 1.466075388 | 1.210816001 |
| Horror | 565 | 5.843539823 | 5.9 | 6.2 | 8.7 | 2.2 | 1.277959079 | 1.130468522 |
| Mystery | 500 | 6.4864 | 6.6 | 6.6 | 8.6 | 2.2 | 1.189754549 | 1.090758703 |
| Western | 97 | 6.689690722 | 6.8 | 6.5 | 8.9 | 3.8 | 1.086767612 | 1.042481468 |
| History | 207 | 7.083574879 | 7.2 | 7.5 | 8.9 | 2 | 0.7883696825 | 0.8879018428 |
| Musical | 132 | 6.507575758 | 6.7 | 7 | 8.5 | 2.1 | 1.502384918 | 1.225718123 |
| Music | 214 | 6.45797546 | 6.6 | 6.5 | 8.5 | 1.6 | 1.389659076 | 1.178838019 |
| War | 213 | 7.070422535 | 7.1 | 7.1 | 8.6 | 2.7 | 0.7651116131 | 0.8747065868 |
| Biography | 293 | 7.150170648 | 7.2 | 7 | 8.9 | 4.5 | 0.5220290804 | 0.7225157994 |
| Sport | 182 | 6.606043956 | 6.8 | 7.2 | 8.7 | 2 | 1.214272661 | 1.101940407 |
| Reality-TV | 2 | 4.75 | 4.75 | 4.8 | 6.6 | 2.9 | 6.845 | 2.61629509 |
| Documentary | 121 | 7.180165289 | 7.4 | 7.8 | 8.7 | 1.6 | 1.116269972 | 1.056536782 |
| Short | 5 | 6.38 | 6.5 | 6.7 | 7.1 | 5.2 | 0.557 | 0.7463243263 |
| Film-Noir | 6 | 7.633333333 | 7.65 | 7.7 | 8.2 | 7.1 | 0.1866666667 | 0.4320493799 |
| News | 3 | 7.533333333 | 7.4 | 7.1 | 8.1 | 7.1 | 0.2633333333 | 0.5131601439 |
| Action | 1153 | 6.239895924 | 6.3 | 6.4 | 9.1 | 1.7 | 1.25179235 | 1.118835265 |

**B. Movie Duration Analysis:**

**Task:** Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

* For this task, I took the required columns only i.e., Duration and IMDM\_score. This makes it easy to complete this task.
* We can tell the relationship between those two columns by mean, median and standard deviation.

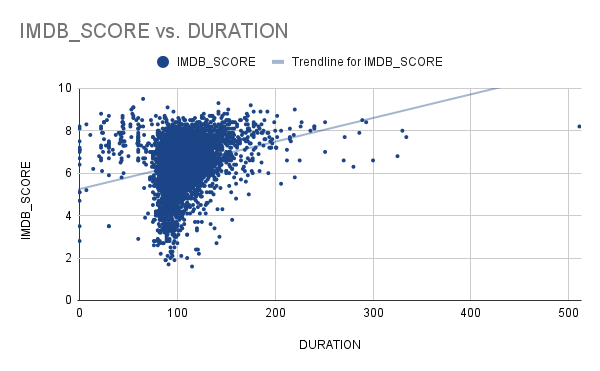
**Statistics of Movie duration:**

| **AVERAGE** | **MEDIAN** | **Standard Deviation** |
| --- | --- | --- |
| 107.201074 | 103 | 25.19744081 |

**Statistics of IMDB\_Score:**

| **AVERAGE** | **MEDIAN** | **Standard Deviation** |
| --- | --- | --- |
| 6.442137616 | 6.6 | 1.125115866 |

**Relation between Duration and IMDB\_Score:**

****

**C. Language Analysis:**

**Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

| **LANGUAGE** | **MOVIE\_COUNT** | **AVERAGE** | **MEDIAN** | **Standard Deviation** |
| --- | --- | --- | --- | --- |
| English | 4704 | 6.398426871 | 6.5 | 1.934412283 |
| Japanese | 18 | 7.394444444 | 7.6 | 0.4447626131 |
| French | 73 | 7.038356164 | 7.2 | 0.8452227213 |
| Mandarin | 26 | 6.788461538 | 7.05 | 0.4917270108 |
| Aboriginal | 2 | 6.95 | 6.95 | 0.138825246 |
| Spanish | 40 | 6.9375 | 7.15 | 0.6200408285 |
| Filipino | 1 | 6.7 | 6.7 | 0.09434748301 |
| Hindi | 28 | 6.632142857 | 6.95 | 0.5033784925 |
| Russian | 11 | 6.363636364 | 6.5 | 0.3032379656 |
| Maya | 1 | 7.8 | 7.8 | 0.1098373683 |
| Kazakh | 1 | 6 | 6 | 0.0844902833 |
| Telugu | 1 | 8.4 | 8.4 | 0.1182863966 |
| Cantonese | 11 | 6.954545455 | 7.2 | 0.3259957224 |
| Icelandic | 2 | 7.55 | 7.55 | 0.1508962954 |
| German | 19 | 7.342105263 | 7.6 | 0.4534567169 |
| Aramaic | 1 | 7.1 | 7.1 | 0.09998016857 |
| Italian | 11 | 7.227272727 | 7.3 | 0.3417285575 |
| Dutch | 4 | 7.425 | 7.45 | 0.2093202606 |
| Dari | 2 | 7.5 | 7.5 | 0.1493575982 |
| Hebrew | 5 | 7.58 | 7.6 | 0.23876798 |
| Chinese | 3 | 5.666666667 | 5.7 | 0.1386186548 |
| Mongolian | 1 | 7.3 | 7.3 | 0.1027965113 |
| Swedish | 5 | 7.44 | 7.6 | 0.2351439135 |
| Korean | 8 | 7.3875 | 7.5 | 0.295637122 |
| Thai | 3 | 6.633333333 | 6.6 | 0.1620056471 |
| Polish | 4 | 8.25 | 8.25 | 0.2335097138 |
| Bosnian | 1 | 4.3 | 4.3 | 0.0605513697 |
| None | 2 | 7.95 | 7.95 | 0.1586835151 |
| Hungarian | 1 | 7.1 | 7.1 | 0.09998016857 |
| Portuguese | 8 | 7.4875 | 7.7 | 0.2998275707 |
| Danish | 5 | 7.5 | 8.1 | 0.2380050417 |
| Arabic | 5 | 7.38 | 7.4 | 0.2336183824 |
| Norwegian | 4 | 7.15 | 7.3 | 0.2017956887 |
| Czech | 1 | 7.4 | 7.4 | 0.1042046827 |
| Kannada | 1 | 7.1 | 7.1 | 0.09998016857 |
| Zulu | 2 | 7.1 | 7.1 | 0.141435391 |
| Panjabi | 1 | 6.6 | 6.6 | 0.09293931163 |
| Tamil | 1 | 5.1 | 5.1 | 0.0718167408 |
| Dzongkha | 1 | 7.5 | 7.5 | 0.1056128541 |
| Vietnamese | 1 | 7.4 | 7.4 | 0.1042046827 |
| Indonesian | 2 | 7.9 | 7.9 | 0.1574227789 |
| Urdu | 1 | 7 | 7 | 0.09857199718 |
| Romanian | 2 | 7.2 | 7.2 | 0.1440468003 |
| Persian | 4 | 7.575 | 7.95 | 0.2152864699 |
| Slovenian | 1 | 6.4 | 6.4 | 0.09012296885 |
| Greek | 1 | 7.3 | 7.3 | 0.1027965113 |
| Swahili | 1 | 7.4 | 7.4 | 0.1042046827 |

* By Statistical Analysis, The most common language used in movies is **English** with 4704 movies.

**D. Director Analysis:**

**Task:** Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

| **DIRECTOR\_NAME** | **AVG\_IMDBSCORE** |
| --- | --- |
| James Cameron | 7.914285714 |
| Gore Verbinski | 6.985714286 |
| Sam Mendes | 7.5 |
| Christopher Nolan | 8.425 |
| Doug Walker | 7.1 |
| Andrew Stanton | 7.733333333 |
| Sam Raimi | 6.907692308 |
| Nathan Greno | 7.8 |
| Joss Whedon | 7.925 |
| David Yates | 7.05 |
| Zack Snyder | 7.175 |
| Bryan Singer | 7.2875 |
| Marc Forster | 7.15 |
| Andrew Adamson | 7.08 |
| Rob Marshall | 6.6 |
| Barry Sonnenfeld | 6.457142857 |
| Peter Jackson | 7.675 |
| Marc Webb | 7.133333333 |
| Ridley Scott | 7.070588235 |

And more rows up to 2400.

* I calculated Average IMDB\_score per each director.I took a specific range(95%=0.95) to identify the top directors with high imdb\_scores.

| **95 percentile** | **Director\_count>95%** |
| --- | --- |
| 7.769166667 | 120 |

* From the above table, 95% of all IMDB\_Scores is 7.769. I considered the directors having the IMDB\_scores of above 95% as top directors and there are 120 top directors having average imdb\_score more than 95%.

**E. Budget Analysis:**

**Task:** Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

* For this task, I used specific columns like movie title, gross, budget.
* To calculate the profit margin,I used the formula **Profit Margin=Gross-Budget.**
* For cleaning the data having blank cells, I removed the rows having both gross and budget as blank cells.
* The result is :

| **MOVIE\_TITLE** | **BUDGET** | **GROSS** | **PROFIT\_MARGIN** |
| --- | --- | --- | --- |
| Avatar | 237000000 | 760505847 | 523505847 |
| Pirates of the Caribbean: At World's End | 300000000 | 309404152 | 9404152 |
| Spectre | 245000000 | 200074175 | -44925825 |
| The Dark Knight Rises | 250000000 | 448130642 | 198130642 |
| John Carter | 263700000 | 73058679 | -190641321 |
| Spider-Man 3 | 258000000 | 336530303 | 78530303 |
| Tangled | 260000000 | 200807262 | -59192738 |
| Avengers: Age of Ultron | 250000000 | 458991599 | 208991599 |
| Harry Potter and the Half-Blood Prince | 250000000 | 301956980 | 51956980 |
| Batman v Superman: Dawn of Justice | 250000000 | 330249062 | 80249062 |
| Superman Returns | 209000000 | 200069408 | -8930592 |
| Quantum of Solace | 200000000 | 168368427 | -31631573 |
| Pirates of the Caribbean: Dead Man's Chest | 225000000 | 423032628 | 198032628 |
| The Lone Ranger | 215000000 | 89289910 | -125710090 |
| Man of Steel | 225000000 | 291021565 | 66021565 |
| The Chronicles of Narnia: Prince Caspian | 225000000 | 141614023 | -83385977 |
| The Avengers | 220000000 | 623279547 | 403279547 |
| Pirates of the Caribbean: On Stranger Tides | 250000000 | 241063875 | -8936125 |

And many rows up to 5044.

| **MOVIE\_TITLE** | **BUDGET** | **GROSS** | **PROFIT\_MARGIN** | **Highest Profit Margin** |
| --- | --- | --- | --- | --- |
| Avatar | 237000000 | 760505847 | 523505847 | 523505847 |

* From the above table, The movie with the highest profit margin is **Avatar.**

| **CORRELATION** |
| --- |
| 0.1021794535 |

* The correlation coefficient between movie budgets and gross is 0.1021794535

**Tech-Stack Used:**

* Google Sheets

**Insights:**

I gained knowledge of using data statistics and data visualization in Excel to extract meaningful conclusions from the dataset. I came to know more about data cleaning strategies depending on the situation of data and charts helps to convey the insights of data more efficiently and quickly.

**Result:**

* Learned more data cleaning strategies practically in this project.

**GoogleSheets link:**

[**https://docs.google.com/spreadsheets/d/10Hifz7fUVr90PmCDszhGp3k5MN\_YJCFSp1kn7hVGJZs/edit#gid=539557811**](https://docs.google.com/spreadsheets/d/10Hifz7fUVr90PmCDszhGp3k5MN_YJCFSp1kn7hVGJZs/edit#gid=539557811)