# Course Project – Assignment

**Mohammed Alnadi (ma1322) - 172007414**
**Umar Khattak (uk50) - 177009705**

1. **Data Collection**

   **Describe the source of your data or provide details of how you collected the data**

   The source of our data was a dataset containing a vast amount of analytical data about over 5,000 movies. The specific data we used from this dataset was the budget, gross income, and total cast facebook likes of each movie. We were able to collect the data from these columns by isolating them and creating a new dataset in which included one or two of the given column's data.

2. **Data Format Description**

   **Describe the format of your dataset in 1-2 paragraphs. E.g.: What files are included? What is the file format / structure of the data? What are the most relevant attributes within each file and what do they mean?**

The dataset we used (movie_metadata.csv) contained a total of 28 columns, where each column provided a different detail about each movie. For example, one column provides the actor 1 name, another provides the time duration of the movie, etc. There are a total of 5,043 movies, which equals out to 141,204 cells of data from our dataset. We chose to use the most relevant columns which we thought would work best with our project and would provide the best insight as to predicting the imdb score of a movie.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5043 entries, 0 to 5042
Data columns (total 28 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   color                      5024 non-null   object
 1   director_name              4939 non-null   object
 2   num_critic_for_reviews     4993 non-null   float64
 3   duration                   5028 non-null   float64
 4   director_facebook_likes    4939 non-null   float64
 5   actor_3_facebook_likes     5020 non-null   float64
 6   actor_2_name               5030 non-null   object
 7   actor_1_facebook_likes     5036 non-null   float64
 8   gross                      4159 non-null   float64
 9   genres                     5043 non-null   object
 10  actor_1_name               5036 non-null   object
 11  movie_title                5043 non-null   object
 12  num_voted_users            5043 non-null   int64
 13  cast_total_facebook_likes  5043 non-null   int64
 14  actor_3_name               5020 non-null   object
 15  facenumber_in_poster       5030 non-null   float64
 16  plot_keywords              4890 non-null   object
 17  movie_imdb_link            5043 non-null   object
 18  num_user_for_reviews       5022 non-null   float64
 19  language                   5031 non-null   object
 20  country                    5038 non-null   object
 21  content_rating             4740 non-null   object
 22  budget                     4551 non-null   float64
 23  title_year                 4935 non-null   float64
 24  actor_2_facebook_likes     5030 non-null   float64
 25  imdb_score                 5043 non-null   float64
 26  aspect_ratio               4714 non-null   float64
 27  movie_facebook_likes       5043 non-null   int64
dtypes: float64(13), int64(3), object(12)
memory usage: 1.1+ MB
```

For our graphs, we cross-referenced the budget with the average imdb score of the movies within a given budget restriction (ex. Budget $1 million to $10 million) and cross-referenced the gross income with the average amount of facebook likes per movie within a given gross income restriction (ex. Gross Income $0 to $1 million). We used these specific attributes because when users vote towards an imdb rating of a movie, factors that are taken into consideration include the total facebook likes, the net profit based on a budget, and the actors total facebook likes which is included in the total amount of facebook likes. Using these attributes will help up to accurately predict an imdb rating of a movie.

3. **<u>Descriptive Statistics</u>**
   **Analyze some of the basic characteristics of the data values. For example: What is the range of values? What is the mean, standard deviation, etc.? Can you plot the distribution of values for numerical variables? What is the distribution of categories for categorical variables? You can optionally visualize some of these.**

   **Note: This is an open-ended task and the kinds of statistics you can compute will depend on a lot on your dataset. For example, for textual data, you could report top keywords, average length, etc.**

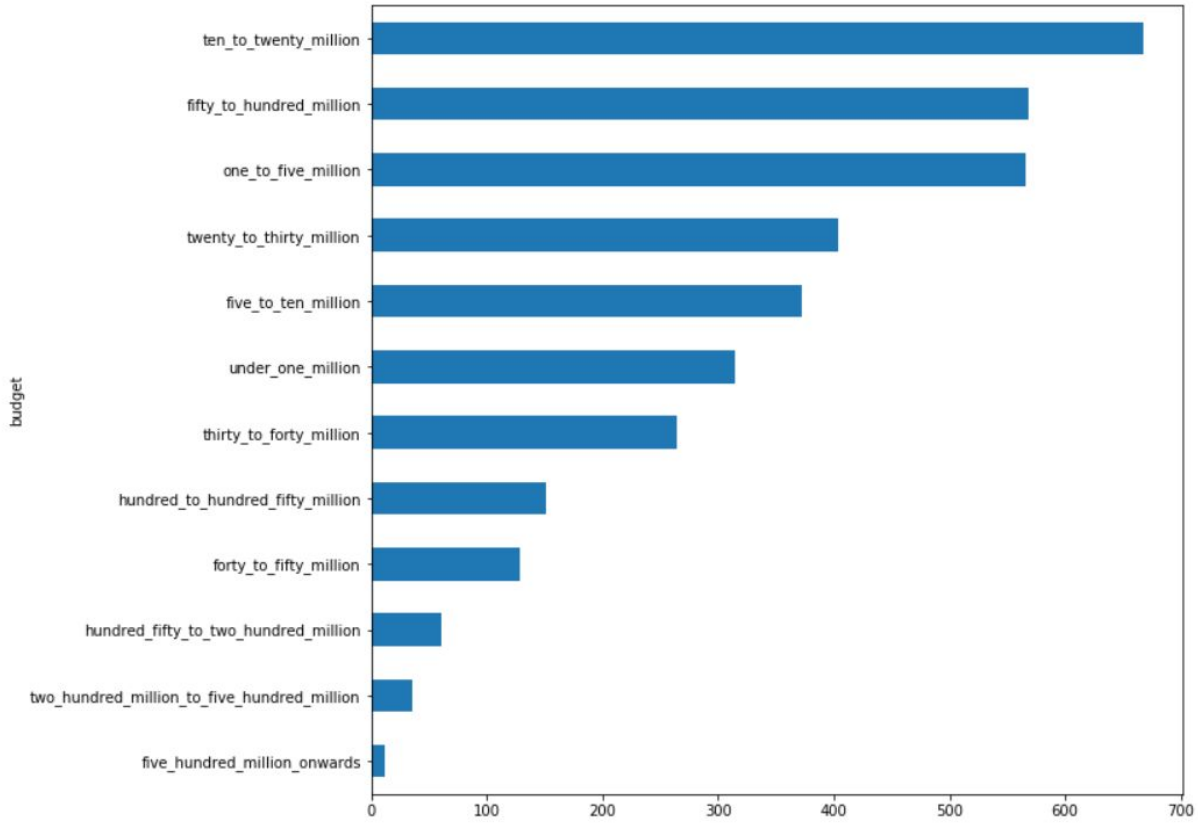
From the Dataset, the following columns are the most important:

- Genre
- gross
- budget
- facebook likes
- IMDB Rating

For this portion of the report, we will show statistics in regards to these columns:

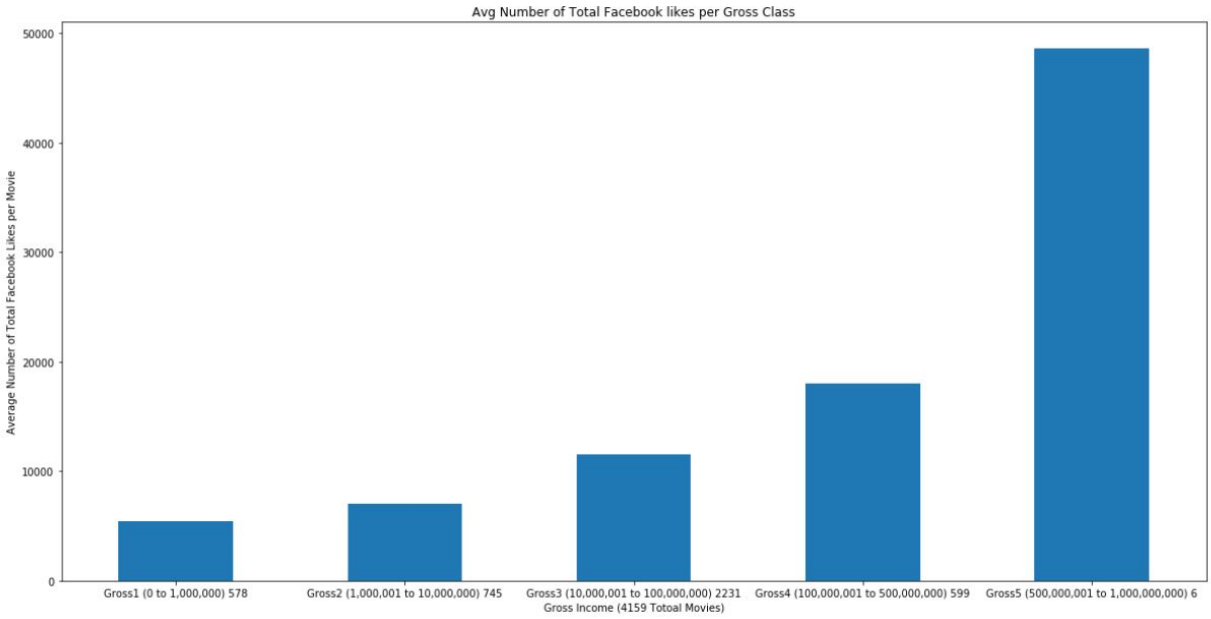
| | gross | budget | movie_facebook_likes | imdb_score |
|---|---|---|---|---|
| **count** | 4.159000e+03 | 4.551000e+03 | 5043.000000 | 5043.000000 |
| **mean** | 4.846841e+07 | 3.975262e+07 | 7525.964505 | 6.442138 |
| **std** | 6.845299e+07 | 2.061149e+08 | 19320.445110 | 1.125116 |
| **min** | 1.620000e+02 | 2.180000e+02 | 0.000000 | 1.600000 |
| **25%** | 5.340988e+06 | 6.000000e+06 | 0.000000 | 5.800000 |
| **50%** | 2.551750e+07 | 2.000000e+07 | 166.000000 | 6.600000 |
| **75%** | 6.230944e+07 | 4.500000e+07 | 3000.000000 | 7.200000 |
| **max** | 7.605058e+08 | 1.221550e+10 | 349000.000000 | 9.500000 |

The number of movies based on budget:




## 4. Data Analysis, Visualization, and Insights

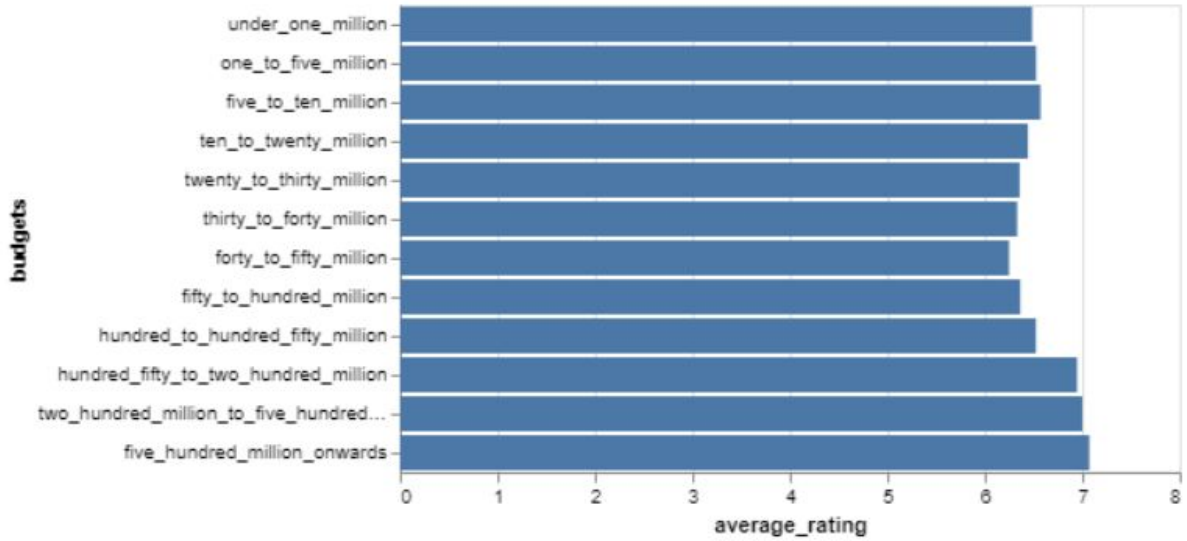
For our graph where we cross-referenced the gross income with the total facebook likes, we separated the movies into 5 different classes based on income. For example, Gross1 is the total amount of movies with an income from $0-$1,000,000. Gross2 is from $1,000,001-$10,000,000. And so on. Because some of the movies in the dataset did not have a budget or gross income, there were only a total of 4,159 movies that were used, which is why we had to separate the movies first, and take the facebook likes from there. If we were to take the facebook likes from the original dataset, it would take the average of the movies that did not have a gross income, and would therefore give us an inaccurate visualization.




Gross1 had a total of 578 movies, with an average of 5401 facebook likes per movie
Gross2 had a total of 745 movies, with an average of 7021 facebook likes per movie
Gross3 had a total of 2231 movies, with an average of 11563 facebook likes per movie
Gross4 had a total of 599movies, with an average of 18056 facebook likes per movie
Gross5 had a total of 6 movies, with an average of 48618 facebook likes per movie

As you can see by the numerical values, the average amount of facebook likes per movie increases as the gross income increases.

For our other graph, we separated the amount of movies per budget class:



315 movies had a budget under $1 million with an average imdb score of 6.48
566 movies had a budget between $1 to $5 million with an avg imdb score of 6.52
372 movies had a budget between $5 to $10 million with an avg imdb score of 6.57
668 movies had a budget between $10 to $20 million with an avg imdb score of 6.44
404 movies had a budget between $20 to $30 million with an avg imdb score of 6.35
264 movies had a budget between $30 to $40 million with an avg imdb score of 6.43
128 movies had a budget between $40 to $50 million with an avg imdb score of 6.24
569 movies had a budget between $50 to $100 million with an avg imdb score of 6.36
151 movies had a budget between $100 to $150 million with an avg imdb score of 6.52
60 movies had a budget between $150 to $200 million with an avg imdb score of 6.94
35 movies had a budget between $200 to $500 million with an avg imdb score of 7.00
11 movies had a budget of $500 million plus with an avg imdb score of 7.07

| budget | average_rating | count | std |
|---|---|---|---|
| under_one_million | 6.486667 | 315 | 1.229991 |
| one_to_five_million | 6.526325 | 566 | 1.249848 |
| five_to_ten_million | 6.571237 | 372 | 1.118266 |
| ten_to_twenty_million | 6.441916 | 668 | 1.158080 |
| twenty_to_thirty_million | 6.357426 | 404 | 1.120736 |
| thirty_to_forty_million | 6.332197 | 264 | 0.948155 |
| forty_to_fifty_million | 6.249219 | 128 | 1.080318 |
| fifty_to_hundred_million | 6.363620 | 569 | 1.006438 |
| hundred_to_hundred_fifty_million | 6.524503 | 151 | 0.911846 |
| hundred_fifty_to_two_hundred_million | 6.946667 | 60 | 1.031449 |
| two_hundred_million_to_five_hundred_million | 7.000000 | 35 | 0.707938 |
| five_hundred_million_onwards | 7.072727 | 11 | 0.825943 |

The graph preceding this chart shows the change in the average movie ratings per the budget of the movie. Movies with relatively low budgets seem to do better than mid-range budget movies, which is somewhat surprising. It is possible that this is the case because people knew they had such low budgets and thus, they took that into consideration when rating the movie.

## 5. Future Plans

The plan for now is as follows: We need to continue to find more correlations in the data. Once we deem that there are enough data correlations, we will need to build a neural network that will be able to predict the IMDB rating of a movie based on: Genre, budget,

and facebook likes. We are in particular lacking when it comes to finding correlations between IMDB ratings and movie genres. This is something that we wish to explore next. Once this is done, should we have sufficient time, we can also use our data to predict the gross of a movie.