TMDb Data Analytics Project

- Olayinka James

PROJECT OUTLINE

- The Movie Database (TMDb) Dataset Background
- TMDb Dataset Question Statement
- TMDb Dataset Exploration
- TMDb Dataset Statistical Analysis
- TMDb Dataset Cleaning
- TMDb Exploratory Data Analysis
- Visualization and Inferences
- Conclusion and Bottlenecks
- References

UDACITY | alx

TMDb DATASET BACKGROUND

The Movie Database (TMDb) Dataset contains information about more than 10,000 movies. This information includes ratings based on popularity, the number of votes cast, the average vote, casts, directors, and production companies, as well as additional information regarding various movies' budgets and the revenue generated by respective movies.

It is important to note that in 2010, as a direct result of the inflation in the value of the dollar, a new feature called budget inflation and revenue inflation was introduced in an effort to adequately account for this monetary index.

UDACITY | alx

TMDb DATASET QUESTION STATEMENT

Some of that will be explored in the analysis of the TMDb dataset includes:

- Genres of the top 10 movies with the highest inflation revenue
- Original title of the top 10 movies with the highest inflation revenue
- Top cast of the top 10 movies in a movies with highest inflation revenue
- Directors of the top 10 movies of movies with the highest inflation revenue
- Production companies of the top 10 movies with the highest inflation revenue
- Top 10 movies with their popularity rating
- Top 10 movies with their average vote
- Percentage Change in budget and revenue due to 2010 dollar inflation
- Correlation analysis table
- Release year of Top 10 movies with Inflated Revenue

TMDb DATSET EXPLORATION

TMDb dataset contains exactly entries 10866 with 21 features and 19 features except "ID" and "IMDB_ID" missing values and other information to be explored which includes but not limited to:

S/N	Features	Missing Values	Data Types	% of Missing Values	Number of Unique Element
1	Popularity	-	Float	-	10814
2	Budget	-	Integer	-	557
3	Revenue	-	Integer	-	4702
4	Title	-	Object	-	10571
5	Homepage	7390	Object	73%	2896
6	Casts	76	Object	0.7%	10719
7	Director	44	Object	0.4%	5067
8	Tagline	2824	Object	26%	7997
9	Keywords	1493	Object	13.7%	8804
10	Overview	4	Object		10847
11	Runtime	-	Integer	-	247
12	Genres	23	Object	0.2%	2039
13	Production companies	1030	Object	9.5%	7445
14	Release date	-	Object	-	5909
15	Release year	-	Integer	-	56
16	Number of votes	-	Integer	-	1289
17	Average votes	-	Float	-	72
18	Budget inflation	-	Float	-	2614
19	Revenue inflation	-	Float	-	4840

TMDb DATSET STATISTICAL ANALYSIS

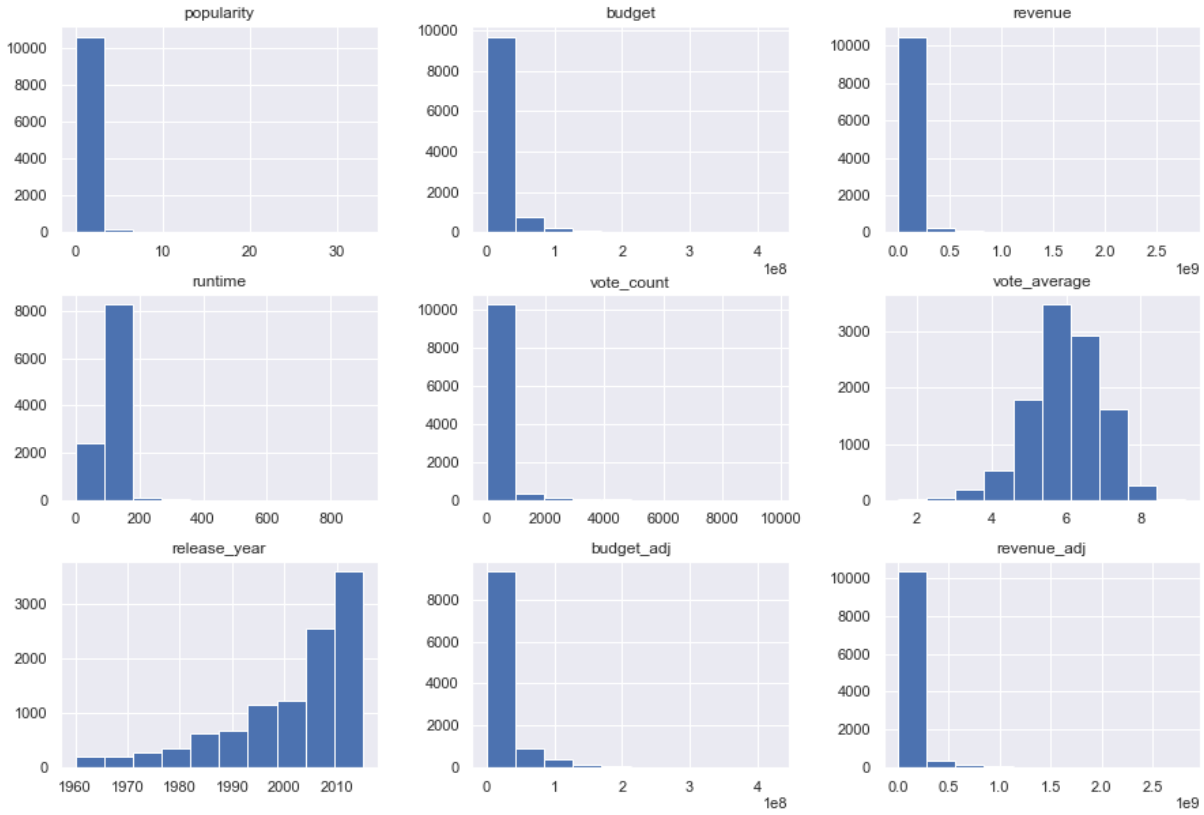


Fig 1 – Histogram Distribution of Numerical Features

Fig 1 shows the distribution of important numerical features in the TMDb dataset.

Fig 2 explains the distribution of the frequency curve with respect to their median, mode, mean and skewness is x-rayed in the chart below.

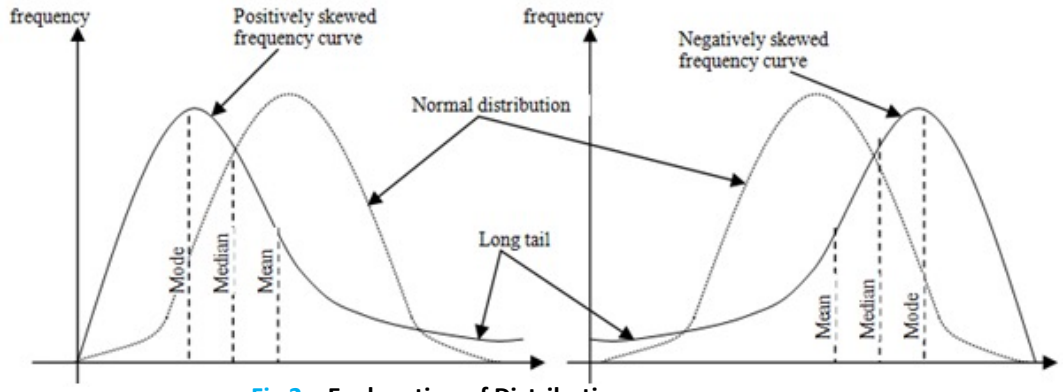
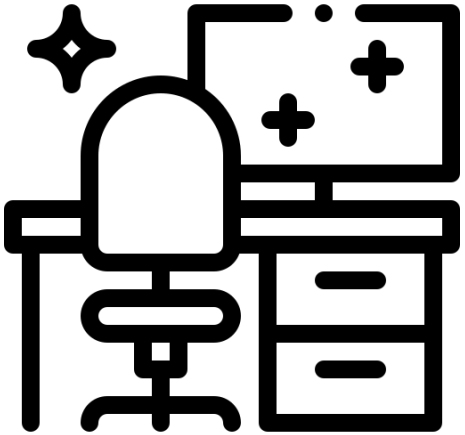


Fig 2 – Explanation of Distribution

TMDb DATSET CLEANING



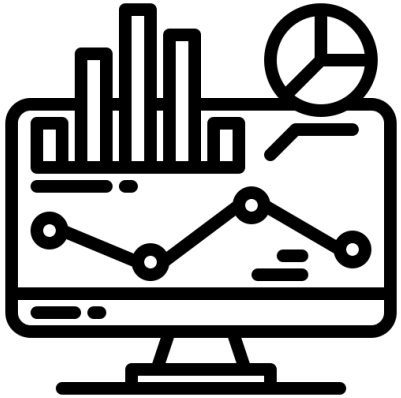
Some of the data cleaning operations includes:

- Removing columns that are not important to the analysis
- Drop duplicate rows in the dataset
- Dropping missing values
- Dropping null rows
- Dropping duplicated columns
- Changing to the required data types in preparation for analysis
- Separation of columns with elements combined by '|'. Features such as Cast, Genres and Production companies
- Renaming “revenue_adj” and “budget_ad” columns to suffix of “_inflation”

TMDb EXPLORATORY DATA ANALYSIS

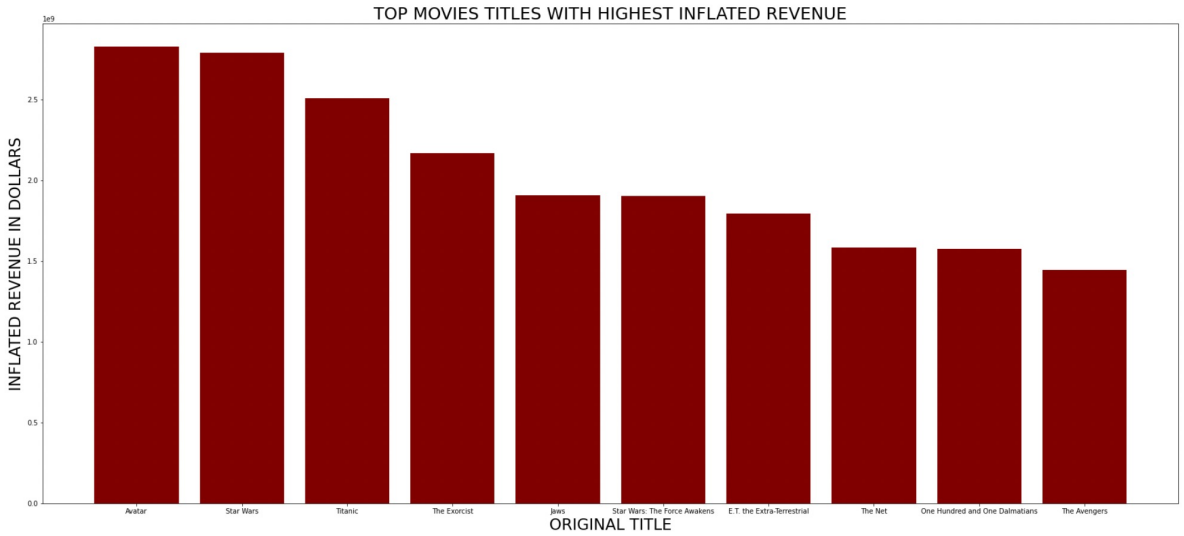
Key Questions to be Explored and Visualized

- Genres of the top 10 movies with the highest inflation revenue
- Movies of the top 10 movies with the highest inflation revenue
- Top cast of the top 10 movies in a movies with highest inflation revenue
- Directors of the top 10 movies of movies with the highest inflation revenue
- Production companies of the top 10 movies with the highest inflation revenue
- Top 10 movies with their popularity rating
- Top 10 movies with their average vote
- Percentage Change in budget and revenue due to 2010 dollar inflation
- Correlation analysis table and heatmap
- Revenue versus Revenue Inflation to complement percentage change Calculation
- Budget versus Budget Inflation to complement percentage change Calculation
- Release year of Top movies with Inflated Revenue



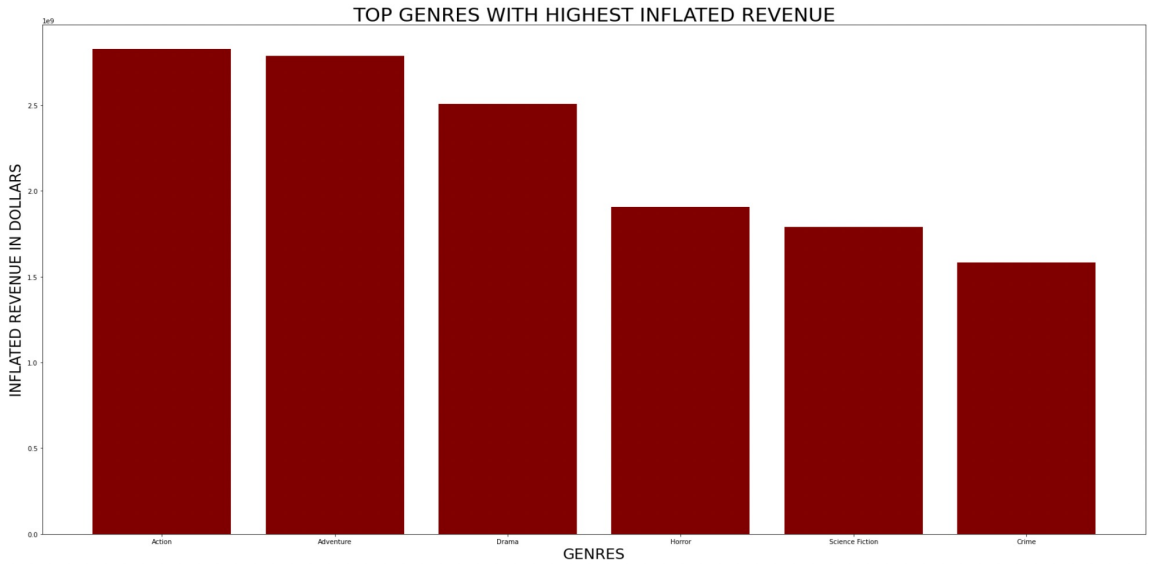
UDACITY | alx

TMDb VISUALIZATION AND INFERENCES



Inference:

Avatar, Star Wars and Titanic are the top 3 movies with the highest inflation revenue



Inference:

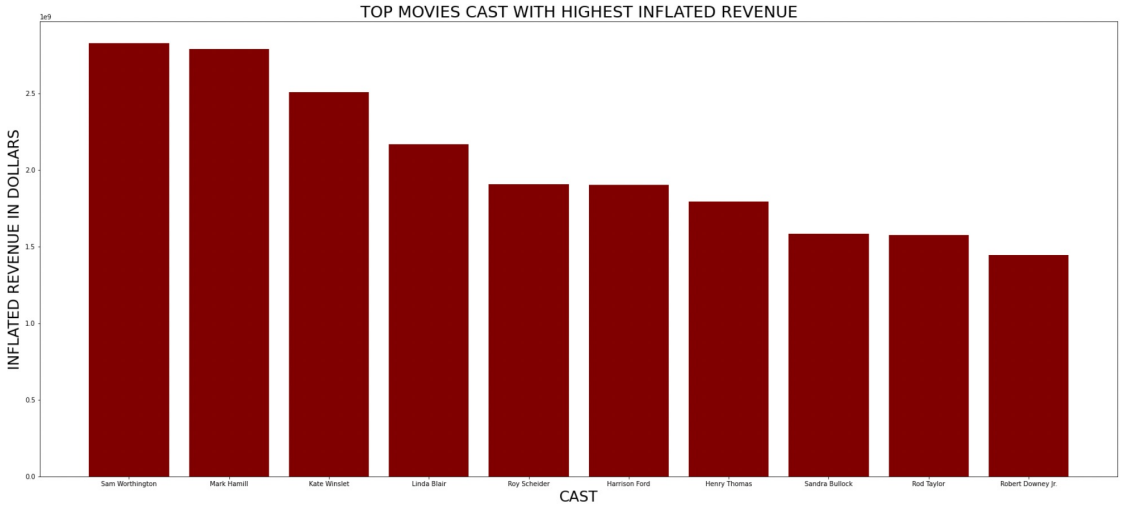
Action, Adventure and Drama are the top 3 genres of movies with the highest inflation revenue

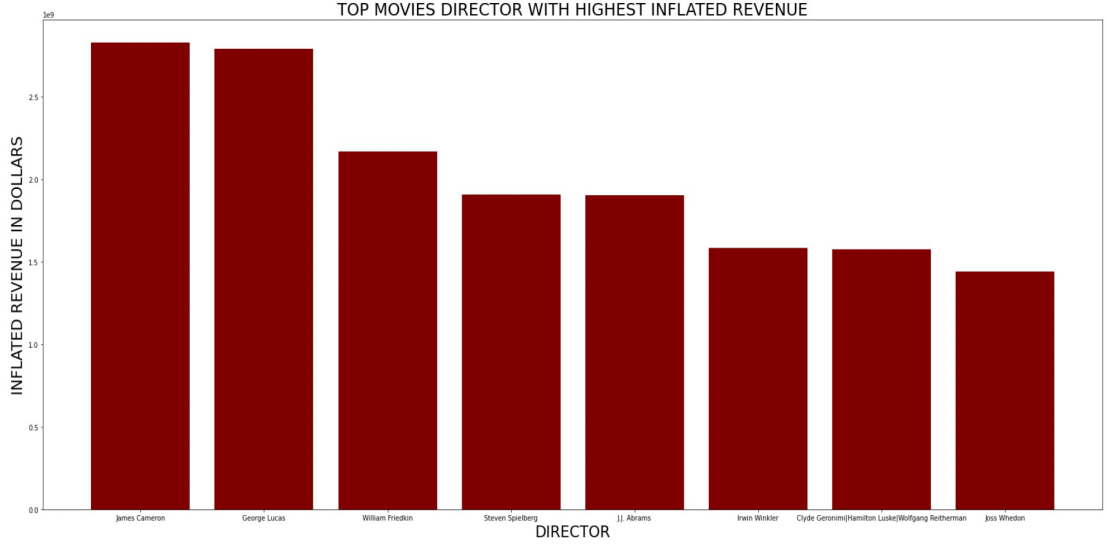
UDACITY | alx

TMDb VISUALIZATION AND INFERENCES

Inference:

Sam, Mark and Kate are the top 3 casts that featured in the movies with the highest inflation revenue

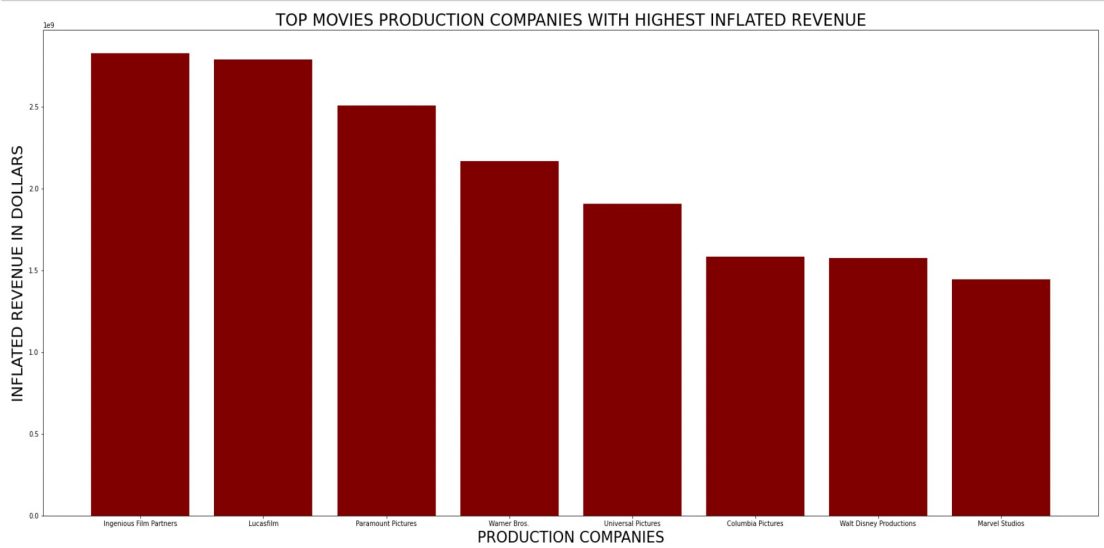




Inference:

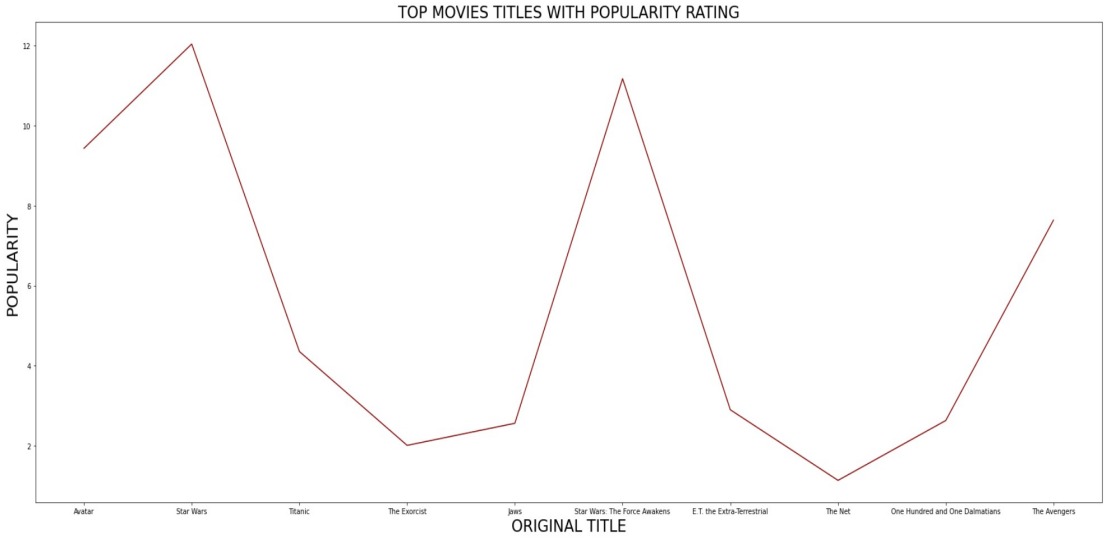
James, George and Williams are the top 3 directors that directed the movies with the highest inflation revenue

TMDb VISUALIZATION AND INFERENCES



Inference:

Indigenous Films partner, Lucasfilm and Paramount Pictures are the top 3 production companies that produced the movies with the highest inflation revenue



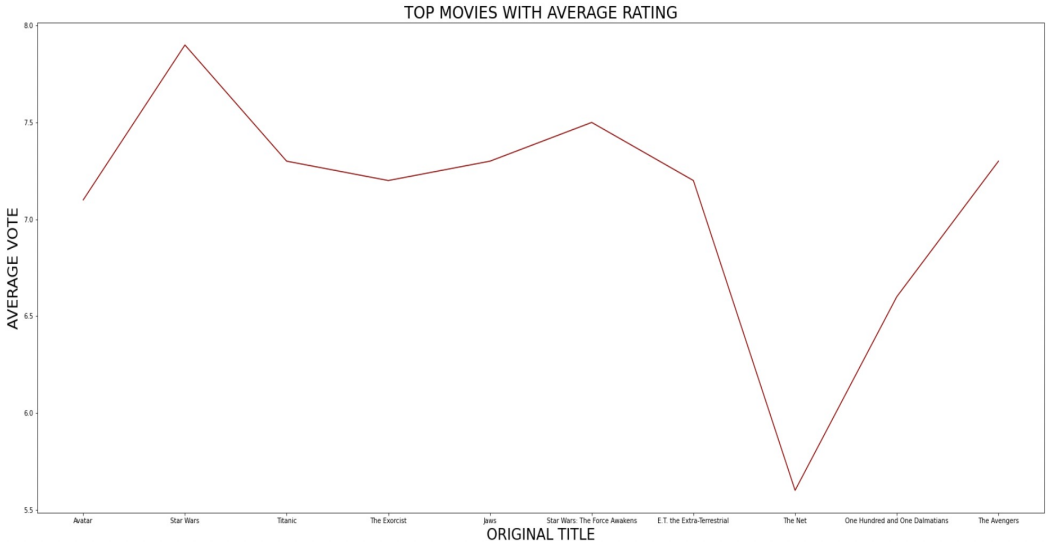
Inference:

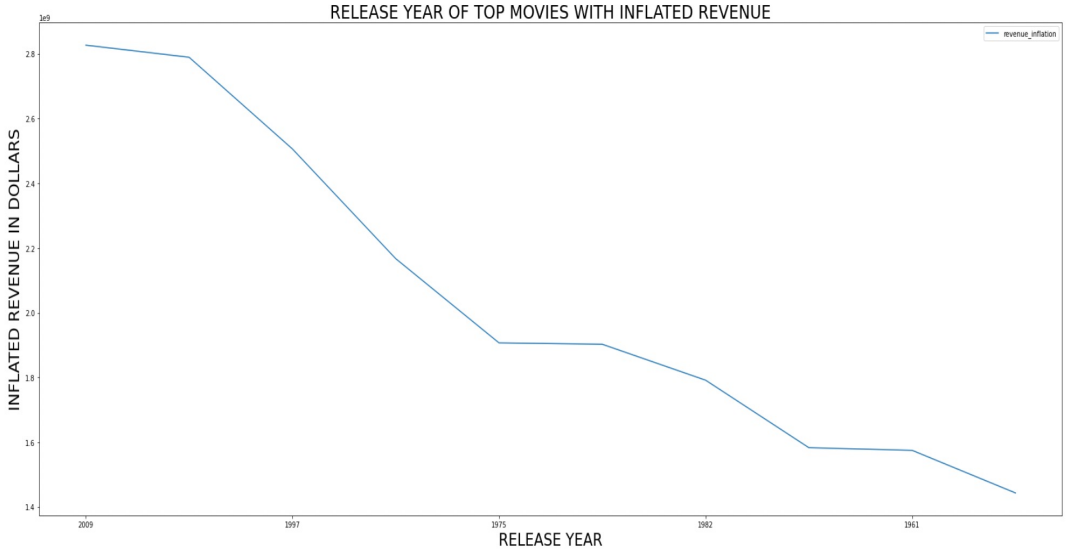
Star wars, Star wars – The Force awaken and Avengers are the top 3 movies with the highest popularity among other movies

TMDb VISUALIZATION AND INFERENCES

Inference:

Star wars, Star wars – The Force awaken and Avengers are the top 3 movies with the highest vote average

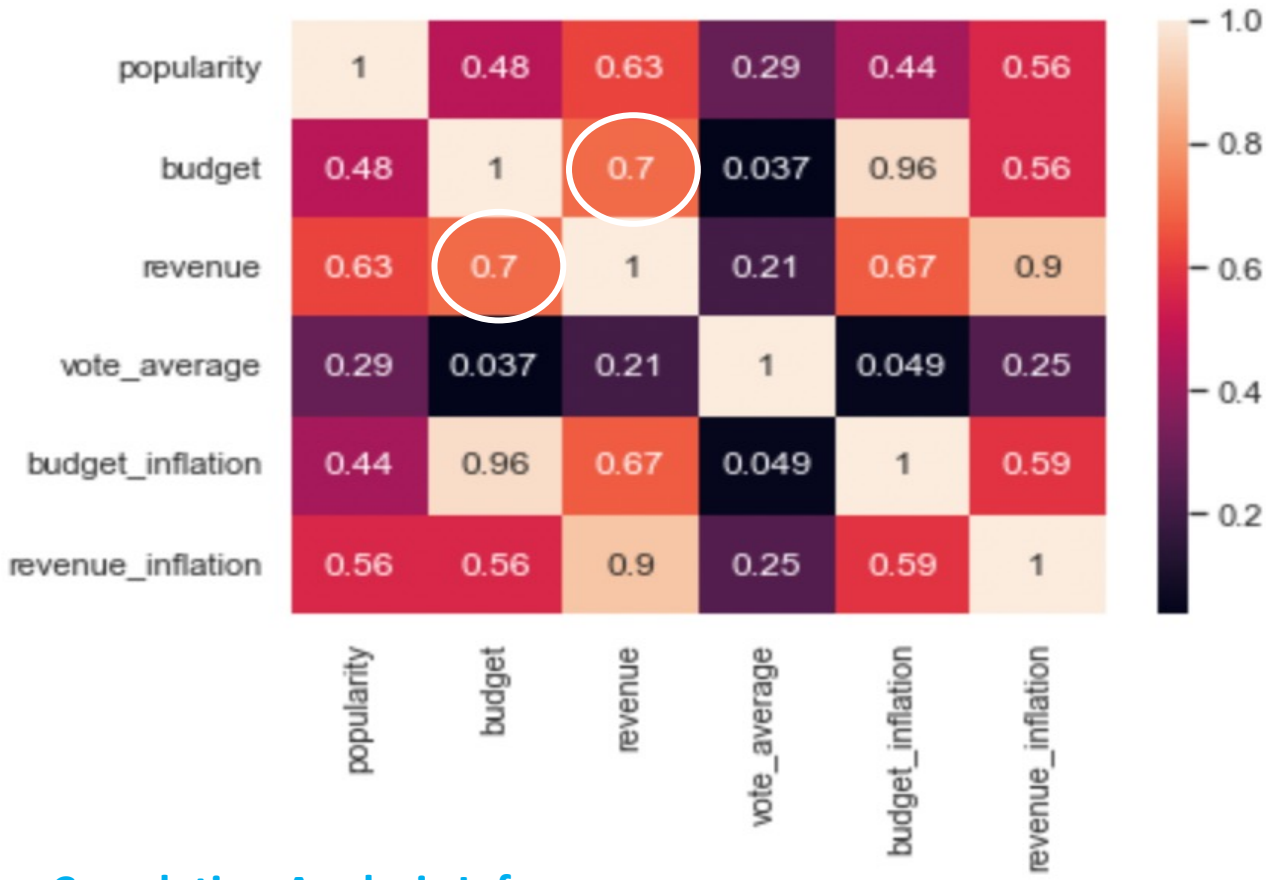




Inference:

The movie with the highest revenue which is Avatar was released in the year 2009. Movies released after 1961 has a progressive increase in revenue.

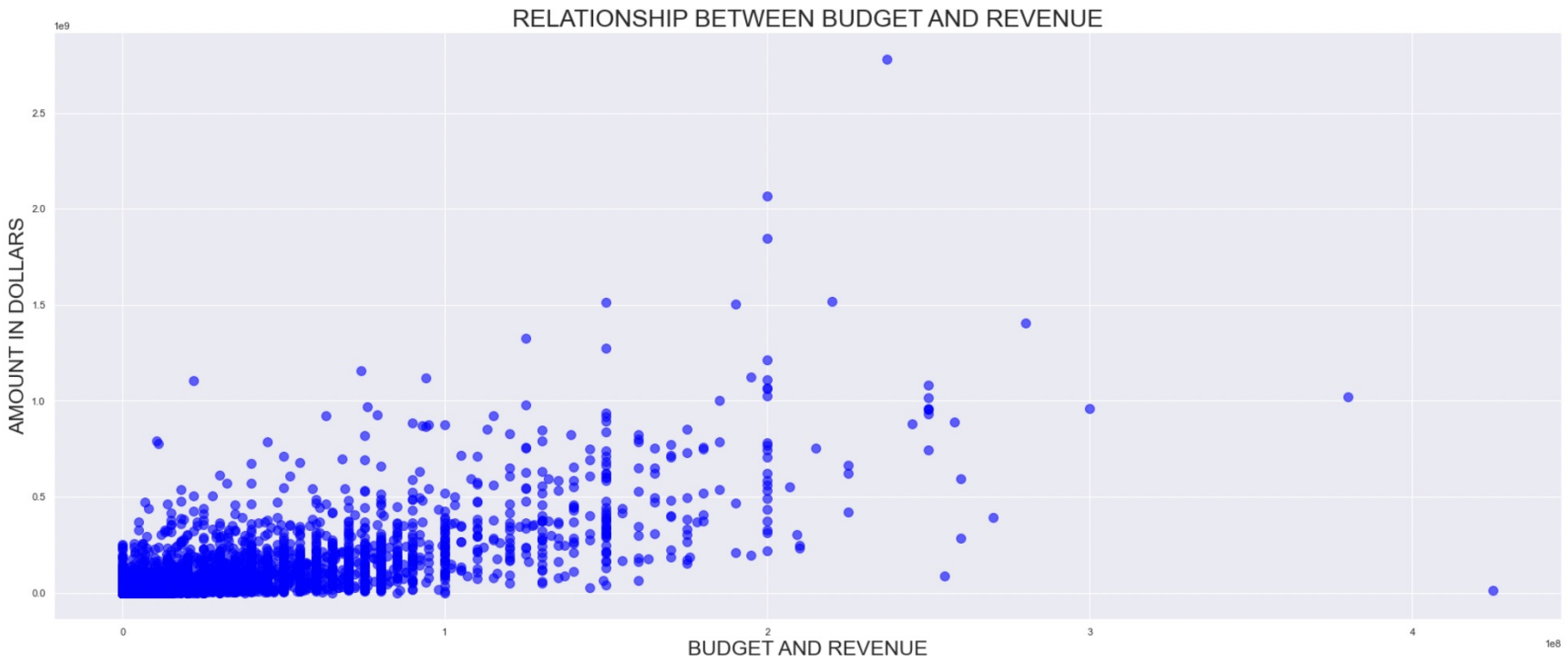
TMDb VISUALIZATION AND INFERENCES



Heatmap Correlation Analysis Inferences:

- *There is a strong correlation between popularity of movie and revenue*
- *There is a strong correlation between **budget** and **revenue***
- *Correlation between budget and budget inflation as well as revenue and revenue inflation was 1, which informs that the percentage change due to 2010 inflation was the same for those features*

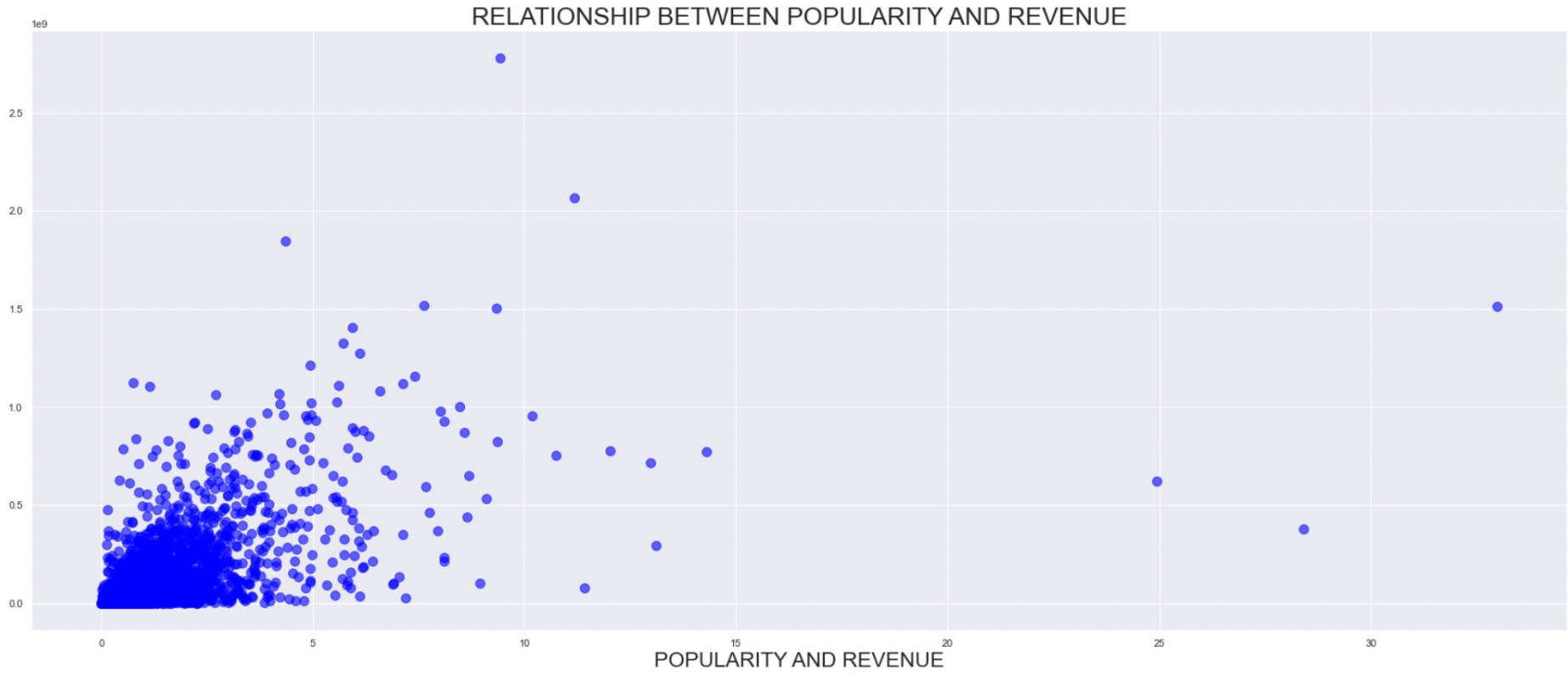
TMDb VISUALIZATION AND INFERENCES



Scatterplot relationship between the budget and Revenue

*This scatterplot shows is a strong correlation between **popularity** and **revenue** which is indicative of the fact that a movie with high popularity rating will most likely generated a high revenue*

TMDb VISUALIZATION AND INFERENCES



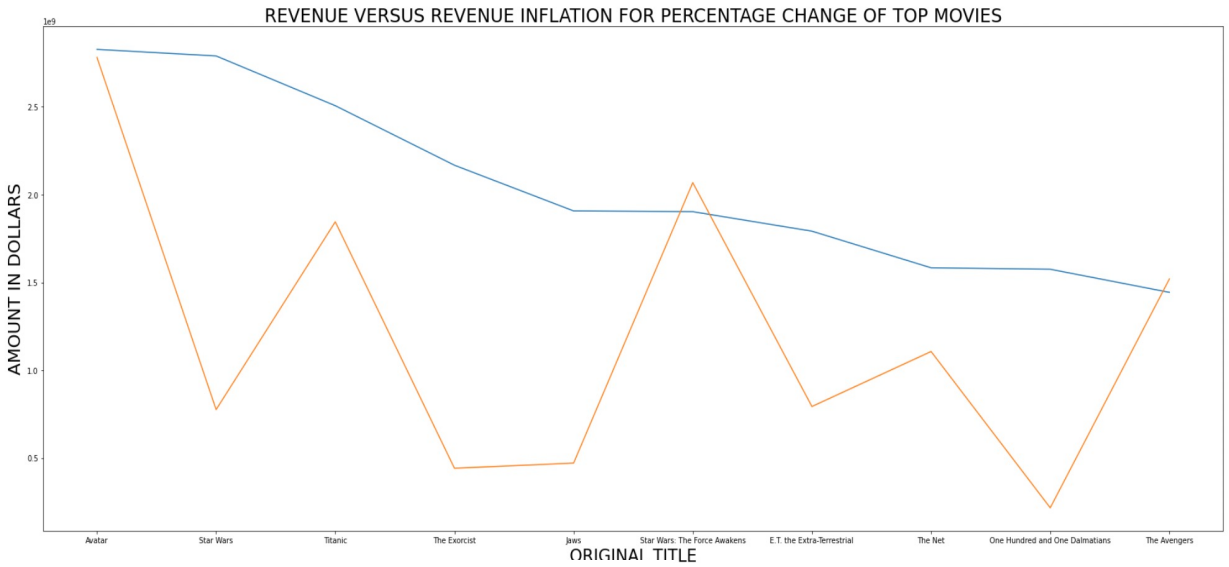
Scatterplot relationship between the popularity and Revenue

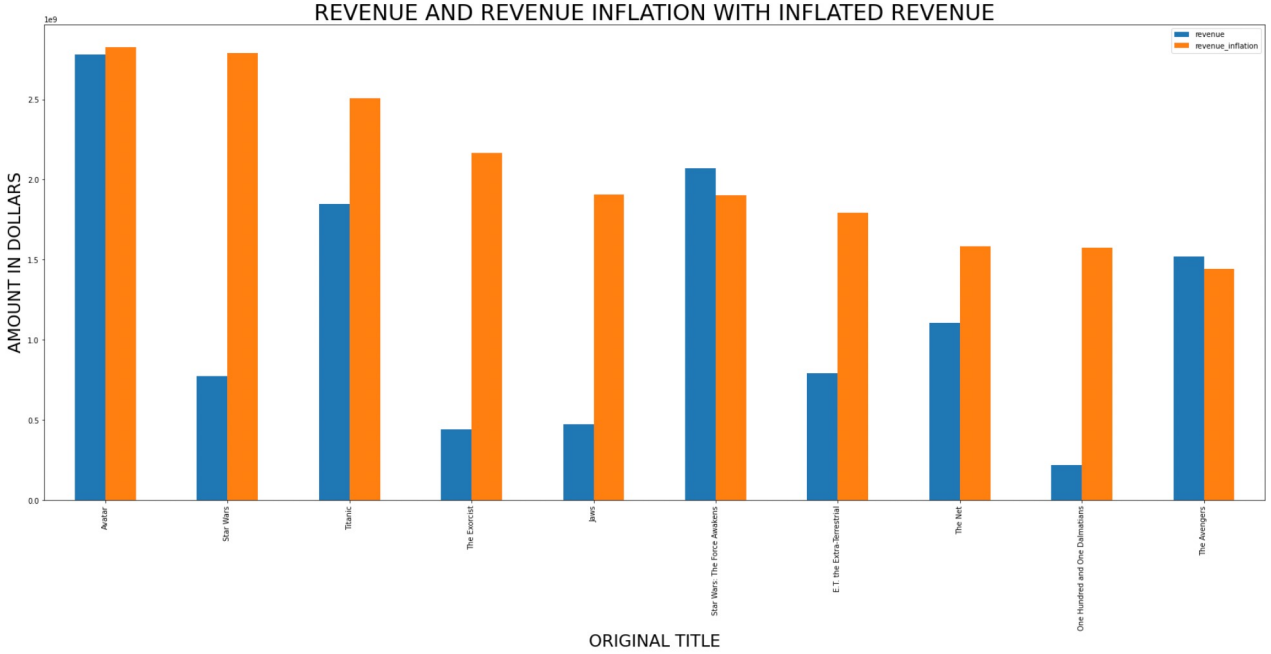
*This scatterplot shows is a strong correlation between **popularity** and **revenue** which is indicative of the fact that a movie with high budget will most likely generated a high revenue*

TMDb VISUALIZATION AND INFERENCES

Inference:

Due to 2010 dollar inflation this chart shows the increment in the revenue to revenue inflation for top 10 movies in a line chart



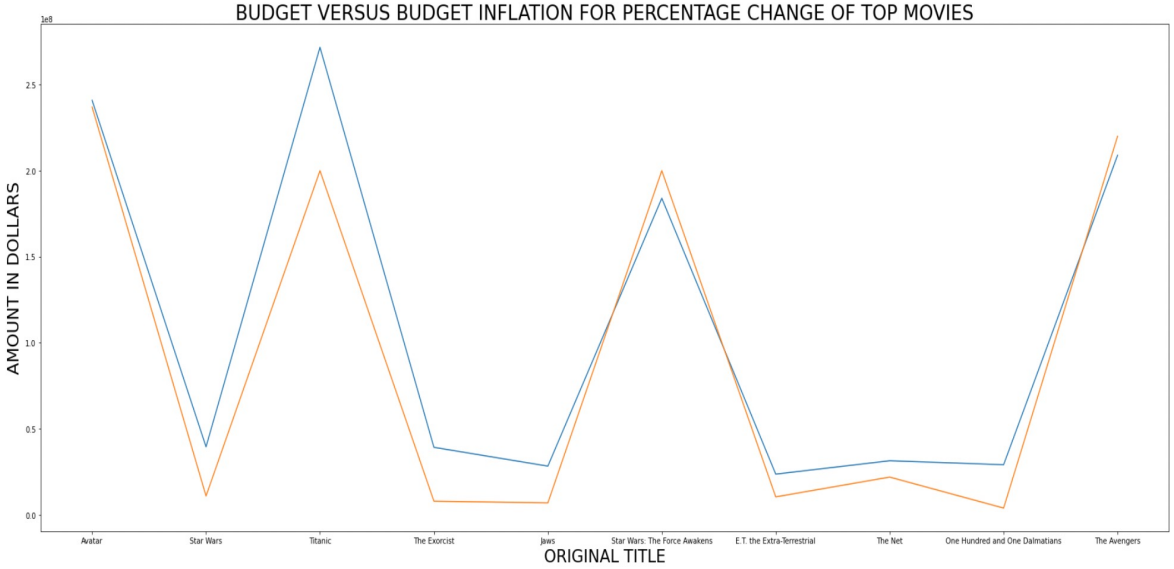


Inference:

Due to 2010 dollar inflation this chart shows the increment in the revenue to revenue inflation for top 10 movies in a bar chart

UDACITY | alx

TMDb VISUALIZATION AND INFERENCES

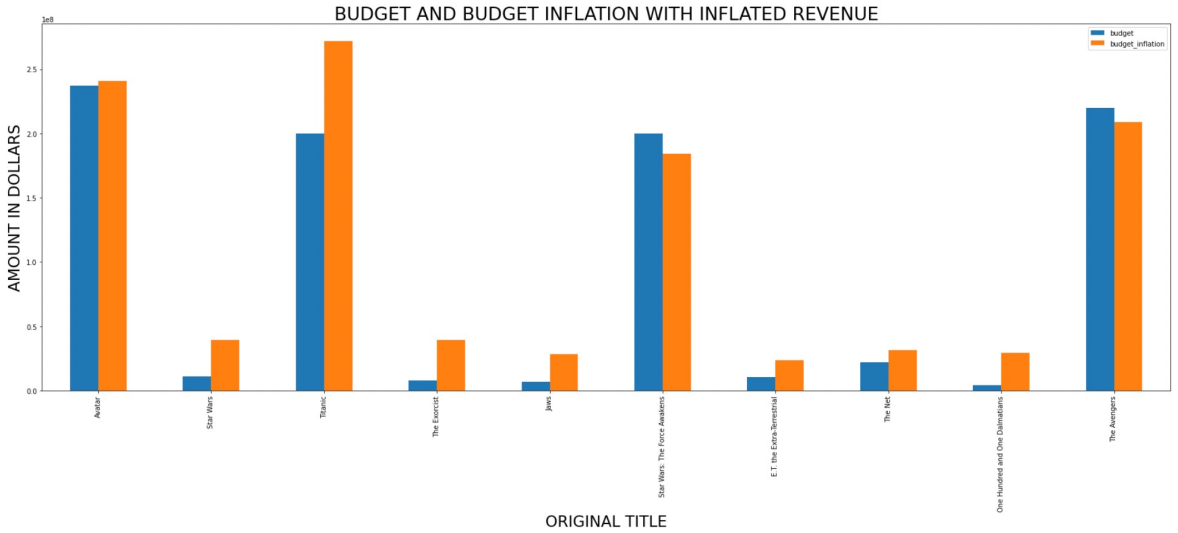


Inference:

Due to 2010 dollar inflation this chart shows the increment in the budget to budget inflation for top 10 movies in a line chart

Inference:

Due to 2010 dollar inflation this chart shows the increment in the budget to budget inflation for top 10 movies in a bar chart



TMDb CONCLUSION AND BOTTLENECKS

Visualization Conclusion

- *Avatar, Star Wars and Titanic are the top 3 movies with the highest inflation revenue. Avatar has the highest inflation revenue which is $2,827,124,000*
- *Action, Adventure and Drama are the top 3 genres of movies with the highest inflation revenue*
- *Sam, Mark and Kate are the top 3 casts that featured in the movies with the highest inflation revenue*
- *James, George and Williams are the top 3 directors that directed the movies with the highest inflation revenue*
- *Indigenous Films partner, Lucasfilm and Paramount Pictures are the top 3 production companies that produced the movies with the highest inflation revenue*
- *Star wars, Star wars – The Force awaken and Avengers are the top 3 movies with the highest popularity among other movies*
- *Star wars, Star wars – The Force awaken and Avengers are the top 3 movies with the highest vote average*
- *The movie with the highest revenue which is Avatar was released in the year 2009. Movies released after 1961 has a progressive increase in revenue*

Heatmap Correlation Analysis Inferences:

- *There is a strong correlation between popularity of movie and revenue*
- *The scatterplot shows is a strong correlation between **popularity** and **revenue** which is indicative of the fact that a movie with high budget will most likely generated a high revenue*
- *There is a strong correlation between budget and revenue*
- *The scatterplot shows is a strong correlation between **budget** and **revenue** which is indicative of the fact that a movie with high budget will most likely generated a high revenue*
- *Correlation between budget and budget inflation as well as revenue and revenue inflation was 1, which informs that the percentage change due to 2010 inflation was the same for those features*

Bottlenecks

- *The separation of some features with **"|"** made it complicated to fully factor in other elements in such feature. This also increased the time taken for data cleaning.*

UDACITY | alx

TMDb REFERENCES

- **Statistical Distribution (Skewness)** - Investopedia Link
- **Pandas Library Documentation** - Pandas Documentation
- **Numpy Library Documentation** - Numpy Documentation
- **Matplotlib Library Documentation** - Matplotlib Documentation
- **Seaborn Library Documentation** - Seaborn Documentation
- **Geek for Geek Website** - Geek for Geek
- **Stack Overflow** - Stack Overflow