**PROJECT REPORT**

**ETL PROJECT TO DETERMINE IF THERE IS A CORRELATION BETWEEN THE BAD WORDS USED IN MOVIE SCRIPTS AND THE RATING OF THE MOVIE**

**OBJECTIVE**:

* The main objective of the project is used to determine if there is a correlation between the frequency of bad words used in a movie script and the rating of the movie.
* Secondary objectives include determining the trend of usage of bad words over the years, most commonly used bad words in movies, determining which countries contribute to movies with more bad words and the likely influence of IMDB ratings

**METHODS USED**:

The ETL part of the project is implemented as follows:

1. **Extraction** –

Data Sources:

* IMSDB (scripts)
* Fandango/ Rottentomatoes
* Metacritic
* OMDB

IMSDB – The movie script was extracted from the IMSDB website using Beautiful Soap.

OMDB – The IMDB rating, metascore, movie running time, release date etc was extracted from OMDB api

1. **Transformation** -

* Convert DataFrame to .csv file
* Cleaning up non-movie links from movie URLs
* Aggregating DataFrames to pull all values in one .csv file
* Filtering out Null values from data
* Importing data to MySQL Database (relational)

1. **Load** –

The table from the MySql Database was loaded into Tableau Desktop and insights were drawn from the following

**FINDINGS & CONCLUSION**:

* The movies with the highest bad word count had a decent IMDB rating ranging between 6-8
* The usage of bad words is normally distributed with the peak at 1990-2000s, the decade also accounts for the lowest IMDB rating average when compared to other decades
* Geographical distribution of movies country wise showed that India produced movies with lowest bad word count and highest IMDB rating
* Films with higher IMDB ratings had bad word count ranging between an avg of 50-100 implying that there’s a threshold for the amount of bad words that audiences will tolerate