PHASE ONE PROJECT

PROJECT TITLE: ANALYSING MOVIE DATA USING DIFFERENT DATASETS

PRESENTER: PAMELA JEPKORIR CHEBII

DATE: 03/06/2024

STUDENT PACE: PART- TIME

INTRODUCTION

- DATA UDERSTANDING

- Microsoft sees all the big companies creating original video content and they want to get in on the fun.
- They have decided to create a new movie studio, but they don't know anything about creating movies.
- You are charged with exploring what types of films are currently doing the best at the box office.
- You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

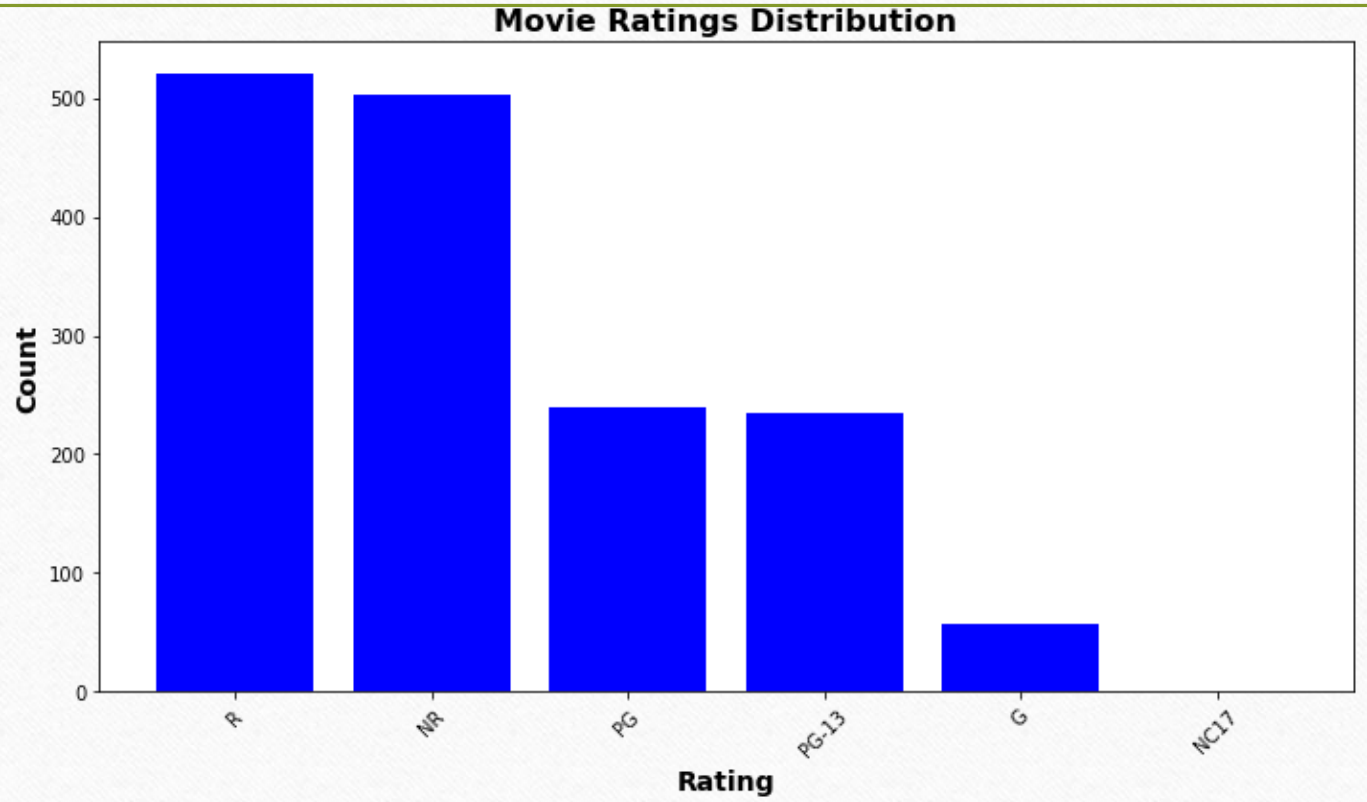
Objectives

- To read different data sets
- To recommend on the best type of movies to produce

Analysing Data

- **Info dataset**
- The movie dataset has a total of 1560 rows with eleven datasets.
- All the columns are of object data type.
- The only column that doesn't have missing values is the runtime column with the rating column having the least missing values at 3
- The R movie type has the highest rating at 1557 out of a total of 1560 values.

Graph for rating against value counts

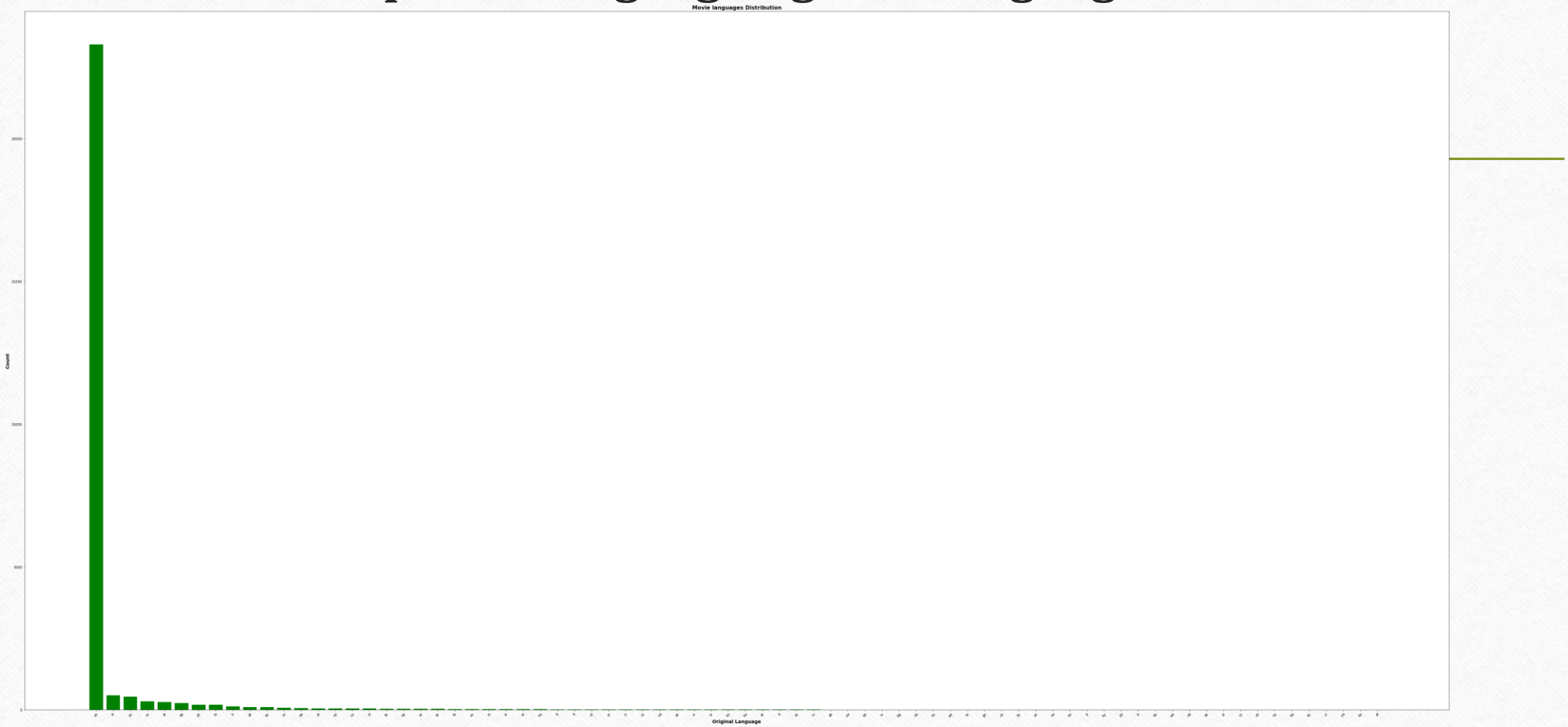


Original Langage

Languages in which the datasets were produced

- The datasets were produced in 76 different languages with English as the highest at 87.8%
- This indicates that English Language is the most preferred language for movie production
- Also it indicates that most people prefer to watch English movies

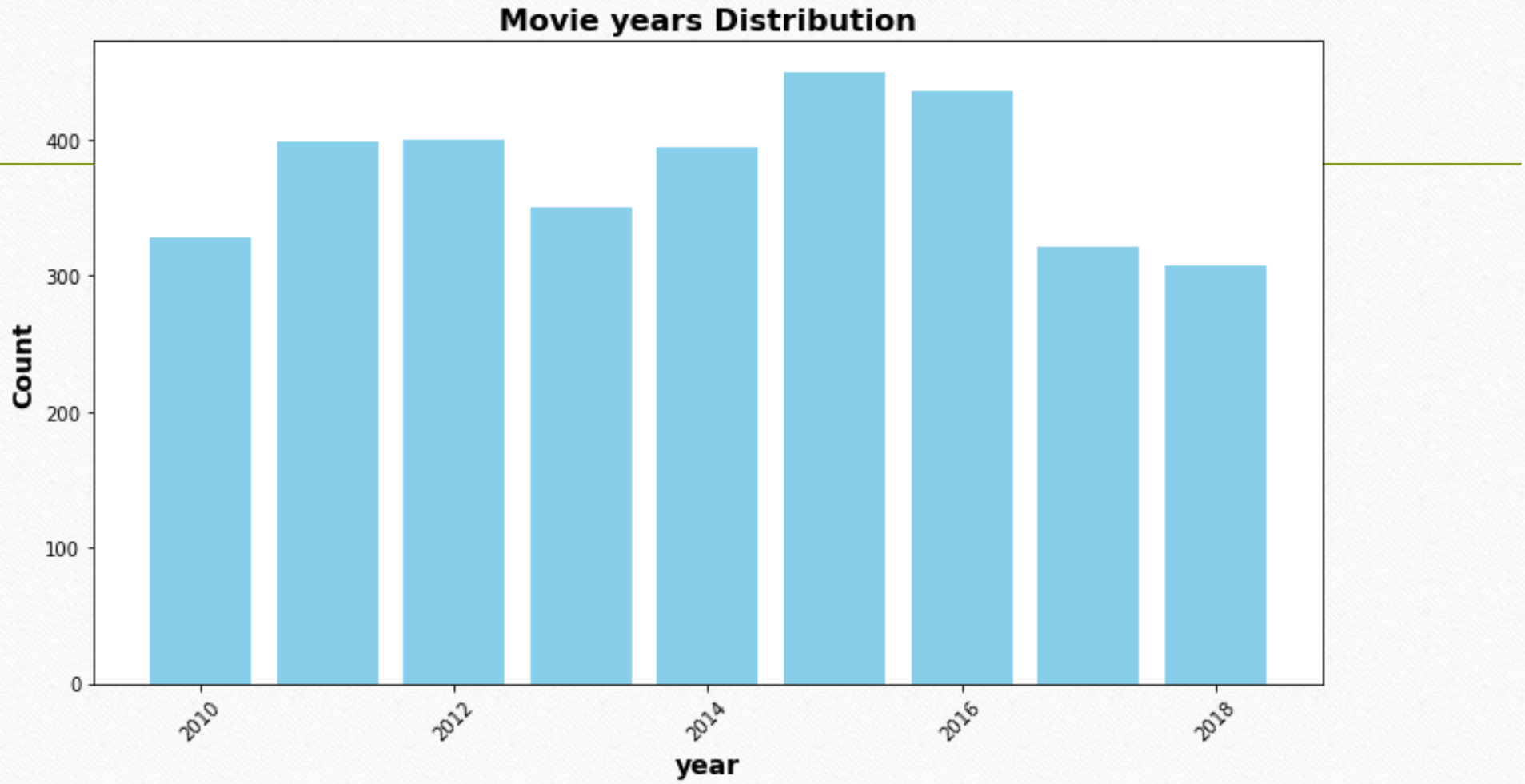
Graph for Language Against language counts



RRODUCTION YEARS

- The movie dataset was produced between the year 2010 and 2018.
- 2015 had the highest number of movies produced at 450 with a percentage of 13.28%
- In 2018 the least numbers of movies with a number of 301 were produced. The percentage was 9.09%

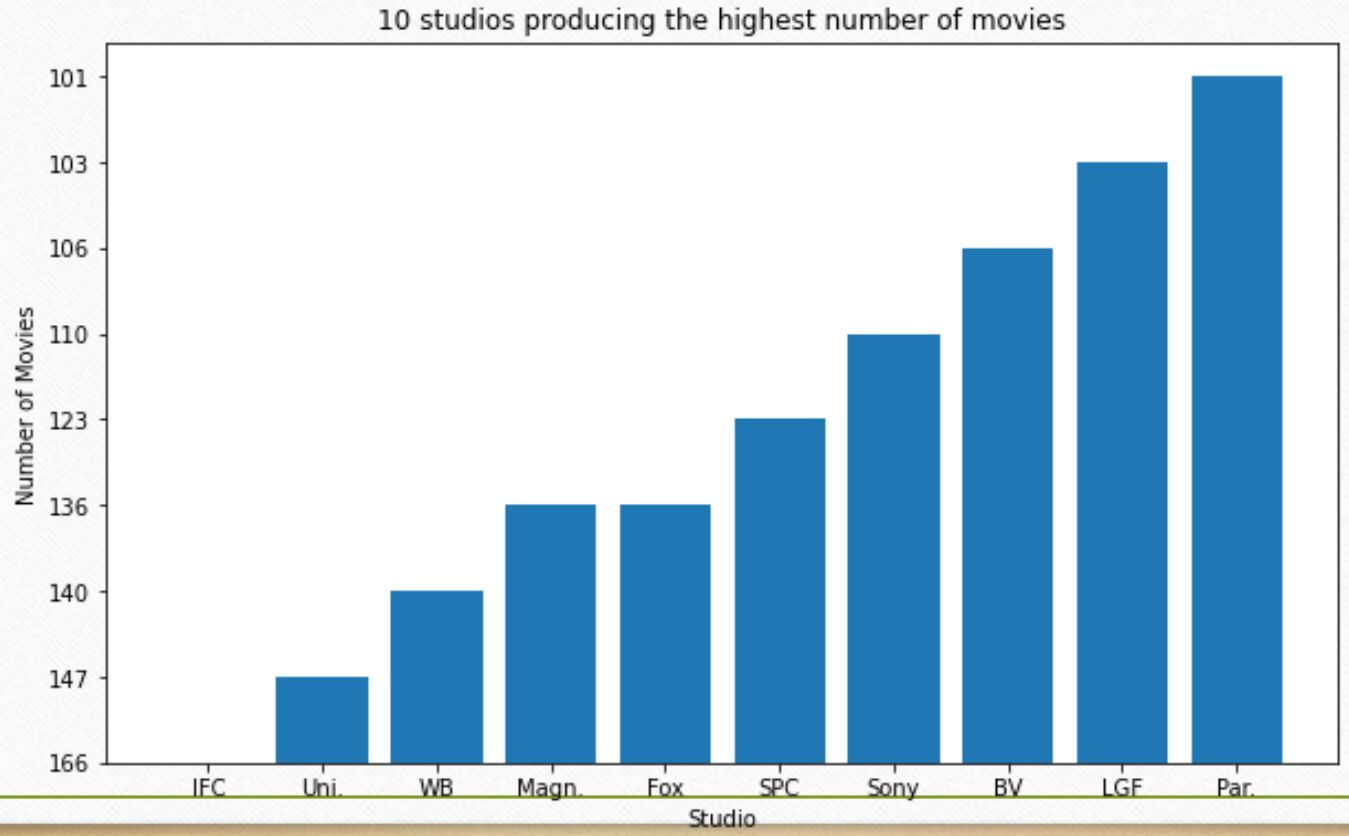
Movie Production Year Counts



Studios Producing movies

- The total number of studios in the dataset is 257 with IFC producing the highest number of movies at 166.
- The total sum of movies produced by all the 257 studios is 3387 movies.

Movie Studios Representation



Dealing with missing data

- 39% of data in the foreign_gross column is missing which means that 39% of the data are sold locally only whereas 61% are sold both locally and internationally

Conclusion

- The best type of movies to produce are R movies
- Nc17 are the least preferred movies