Data Understanding

Describe the data being used for this project.

Questions to consider:

- Where did the data come from, and how do they relate to the data analysis questions?
- What do the data represent? Who is in the sample and what variables are included?
- What is the target variable?
- What are the properties of the variables you intend to use?

In [95]:

```
# Import standard packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from pprint import pprint
import operator
from itertools import groupby
%matplotlib inline
```

In [104]:

```
# FILE_PATH = (Path(__file__).with_name("title.basics.csv")).absolute()

def load_datasets():
    """
    Load various data sets

    args: None
    return (dataframe): title_basics, title_ratings
    """

    title_basics = pd.read_csv("title.basics.csv")
    title_ratings = pd.read_csv("title.ratings.csv")

    return title_basics, title_ratings

def merge_datasets(dataframe_a, dataframe_b, common_column):
    """
    Merges provided dataframes and returns a common dataframe

    args (dataframes, string): dataframe_a, dataframe_b, common_column

    returns (dataframe): merged_dataframe
    """
    return pd.merge(dataframe_a, dataframe_b, on=common_column)


def remove_duplicates(dataframe, column):
    """
    Remove duplicates based on passed column

    args: dataframe, column
    """
    return dataframe.drop_duplicates(subset=column, keep="last")

def drop_empty_titles(dataframe):
    """
    Remove empty genres from the dataframe

    args (dataframe): dataframe

    returns (dataframe): dataframe
    """

    return dataframe.dropna(subset=["genres"]) # inplace=True

def group_data_by_genre_and_start_year(data_list):
    """
    Group data based on genre

    args: (list): data_list

    return: (list of lists): grouped_data_list
    """
    print(data_list[:5])
    data_list = sorted(data_list, key=operator.itemgetter("start_year"), revers
    return [list(group_item[1]) for group_item in groupby(data_list, key=operat
def group_data_by_year(data_list):
    """
    """

def process_data(grouped_data_list):
    """
    Process data

    return (list of dictionaries): processed_data_list

    Example Response:
    processed_data = [
        {"Year": "2012", "Genre":"War", "Votes": 200000, "AverageRating": 7},
        {"Year": "2013", "Genre":"War", "Votes": 200000, "AverageRating": 5},
        {"Year": "2014", "Genre":"Action", "Votes": 200100, "AverageRating": 7}
        ]
    """
    grouped_list = []

    for data_list in grouped_data_list:
        processed_data = []

        for key, data in groupby(data_list, key=operator.itemgetter("genres"))
            grouped_data = list(data)
            year = grouped_data[0].get("start_year")
            total_votes = sum(data.get("numvotes") for data in grouped_data)
            average_votes = total_votes / len(grouped_data)
            average_rating = sum(data.get("averagerating") for data in grouped_
            dic = {"Year": year, "Genre": key, "Votes": total_votes,"AverageVo
            processed_data.append(dic)
        lst = sorted(processed_data, key=operator.itemgetter("AverageVotes", "
        grouped_list.append(lst[0])
    return grouped_list


title_basics, title_ratings = load_datasets()
df = merge_datasets(title_basics, title_ratings, "tconst")
df = remove_duplicates(df, "original_title")
df = drop_empty_titles(df)
data_list = df.to_dict("records")
grouped_data = group_data_by_genre_and_start_year(data_list)
grouped_data_list = process_data(grouped_data)
column_names = ["Year", "Genre", "Votes", "AverageVotes", "AverageRating"]
df = pd.DataFrame(grouped_data_list, columns=column_names)

df.to_excel("file.xlsx", index=False)
df.head(15)
enres': 'Drama', 'averagerating': 6.9, 'numvotes': 4517}, {'tconst': 'tt006
9204', 'primary_title': 'Sabse Bada Sukh', 'original_title': 'Sabse Bada Su
kh', 'start_year': 2018, 'runtime_minutes': nan, 'genres': 'Comedy,Drama',
'averagerating': 6.1, 'numvotes': 13}, {'tconst': 'tt0100275', 'primary_tit
le': 'The Wandering Soap Opera', 'original_title': 'La Telenovela Errante',
'start_year': 2017, 'runtime_minutes': 80.0, 'genres': 'Comedy,Drama,Fantas
y', 'averagerating': 6.5, 'numvotes': 119}]
```

Out[104]:

	Year	Genre	Votes	AverageVotes	AverageRating
0	2019	Action,Adventure,Sci-Fi	737360	368680.0	7.9
1	2018	Action,Adventure,Sci-Fi	670926	670926.0	8.5
2	2017	Action,Drama,Sci-Fi	560270	560270.0	8.1
3	2016	Action,Adventure,Comedy	820847	820847.0	8.0
4	2015	Action,Adventure,Fantasy	784780	784780.0	8.0
5	2014	Adventure,Drama,Sci-Fi	1299334	1299334.0	8.6

Data Preparation

Describe and justify the process for preparing the data for analysis.

Questions to consider:

- Were there variables you dropped or created?
- How did you address missing values or outliers?
- Why are these choices appropriate given the data and the business problem?

In [6]:

```
# Here you run your code to clean the data
```

Data Modeling

Describe and justify the process for analyzing or modeling the data.

Questions to consider:

- How did you analyze or model the data?
- How did you iterate on your initial approach to make it better?
- Why are these choices appropriate given the data and the business problem?

In [ ]:

```
# Here you run your code to model the data
```

Evaluation

Evaluate how well your work solves the stated business problem.

Questions to consider:

- How do you interpret the results?
- How well does your model fit your data? How much better is this than your baseline model?
- How confident are you that your results would generalize beyond the data you have?
- How confident are you that this model would benefit the business if put into use?

Conclusions

Provide your conclusions about the work you've done, including any limitations or next steps.

Questions to consider:

- What would you recommend the business do as a result of this work?
- What are some reasons why your analysis might not fully solve the business problem?
- What else could you do in the future to improve this project?