Steve Abonyo Data Science Part-Time Class-Phase 1 Project

A.) *Business Understanding*

This business-project aims at helping **Microsfot** make an **informed decision while entering into the movie industry**. They've noticed a number big companies joining the industry and they also want to try their luck on this new fun. We have three sets of data from which we want to derive our conclusions from. At they end of the study, they may know which **genre is the best, what kind of income to they expect, what other competitors are reaping from the industry** etc.,

B.) *Data Understanding*

We have sets of data from which we are supposed to derive our insights. The three sets of data are the following `imdb.title.basics` , `imdb.title.ratings` , `bom.movie_gross`

- From the title.basics data, we have the movie title, start year , run time in minutes and the genre of the movies
- From the title.ratings data, can get the rating of the movies and the number of people who voted for the movies
- From the bom.movies data, we can get both the domestic and the gross income for the movies per year

C.) *Data Preparation*

NB: We are required to merge the three datasets so because they aree linked. The link is using the primary keys.

Title.basics dataset and ***Title.ratings dataset*** both have a common variable called `tconst` , From which will use the `primary_title` column to join with the ***bom.movie dataset*** `title` column. NB: because the title and primary key are different labels, we will rename so as to stndardize them.

So the primary keys we have are the following;

- *tconst*
- *primary_title*

Nevertheles, we are going to use the `merge` function from pandas to join the 3 dataframes together

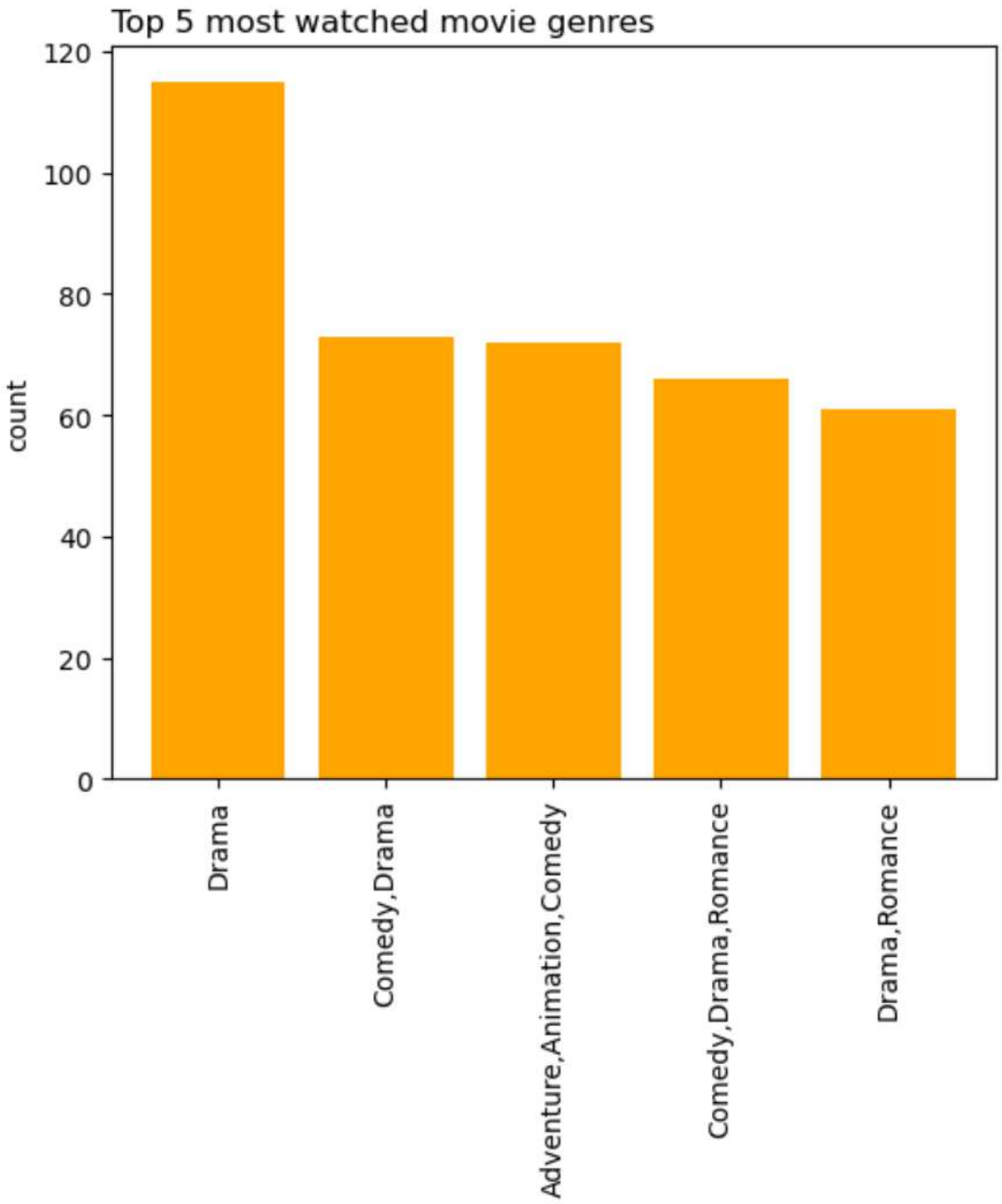
Steps that were taken so as to ensure we have clean data are the following;

- Dealing with the missing values
- Ensuring the variables are of the correct type
- Checking for duplicates
- Stripping white spaces
- Dropping irrelevant columns
- Dealing with the outliers

D.) *Results*

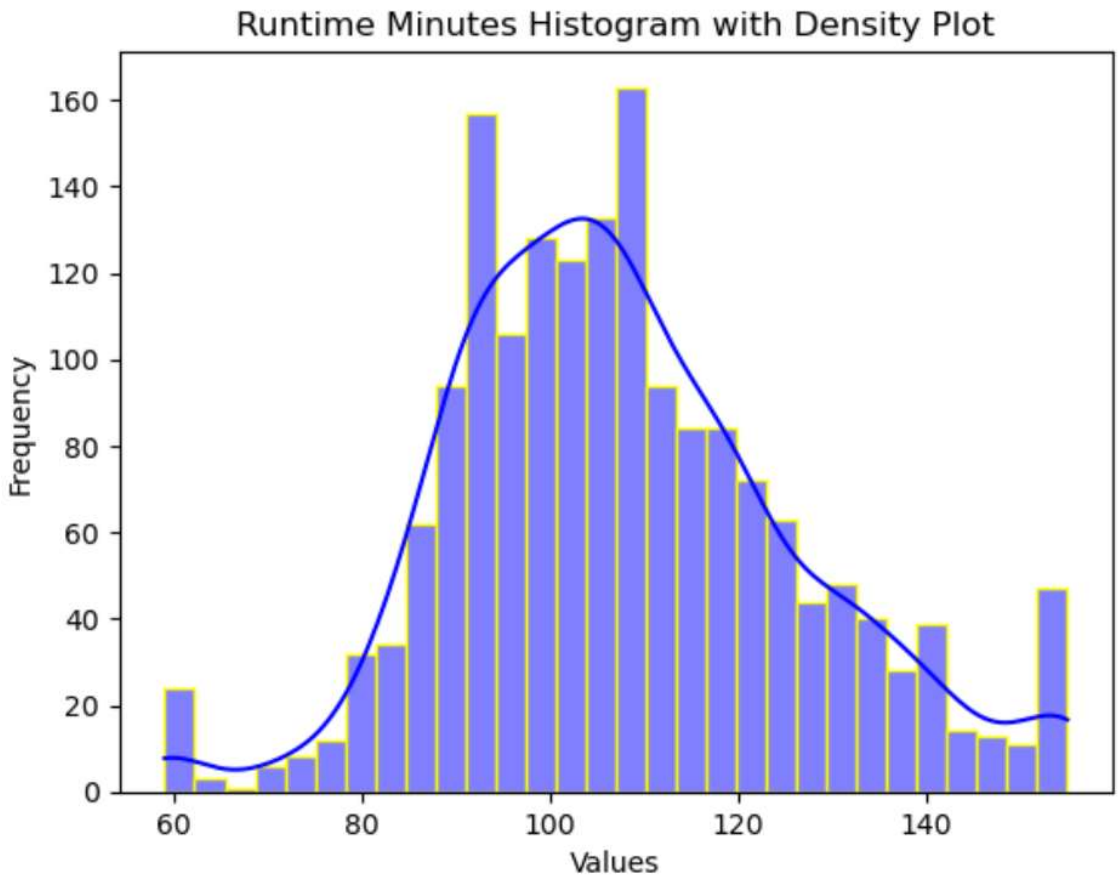
Below are the some of the major findings we saw from the data

Genre



In the top 5 movies, Drama genre appears the most. So drama kind of movies is the best field to venture into

Movie runtime in minutes



Most of the movies run in the range of time between 90 mins and 120 mins

Correlation

	averagerating	numvotes	runtime_minutes	domestic_gross	foreign_gross
averagerating	1.000000	0.376249	0.267372	0.148042	0.134273
numvotes	0.376249	1.000000	0.340455	0.663636	0.582207
runtime_minutes	0.267372	0.340455	1.000000	0.141650	0.186008
domestic_gross	0.148042	0.663636	0.141650	1.000000	0.777419
foreign_gross	0.134273	0.582207	0.186008	0.777419	1.000000

There is a close strong positive correlation between number of votes, foreign gross and domestoc gross income

E.) *Conclusions*

The three major reccomendations are the following;

- **Microsoft should produce movies that are less than 100 mins**:, from the analysis it was discovered that movies less than 100 mins even had a higher movie rating than there other counterparts of time above 100mins
- **Microsft should target international market in their movies; not only local market**: from the analysis between 2010 and 2019, we've seen that in the latter years, the foreign gross income is increasing where are the domestic gross income is decreasing
- **The best studios to invest in are the following;** `Fox` , `Uni` and `BV` : From the analysis it was discovered that movies produced in the mentioned studios produced the best income both domestic and foreign gross

In [ ]: