COMP 3610- Big Data Analytics

The University of The West Indies, St. Augustine

Assignment 1

Date Distributed: February 4, 2022 **Date Due**: February 18, 2022 (11:55pm)

Problem Description

Your firm has been hired by one of the top major film studios at Hollywood (Universal Pictures) to analyze data on the top highest grossing Hollywood movies. The film studio would like to know some characteristics on the highest grossing films and you are tasked with making recommendations on what type of film they should produce in future.

The dataset originated out of scraping IMBD, Rotten Tomatoes and many other websites. It contains the top 900 highest grossing films in Hollywood. Table 1 provides a description of each field present within the dataset.

Column	Description
Title	Title of the movie
Movie Information	A brief synopsis of the movie
Distributor	Film operation/distribution company
Release Date	The date at which the movie was released
Domestic Sales	Self explanatory.
International Sales	Self explanatory.
World Wide Sales	Self explanatory.
Genre	The style or category of the movie.
Run Time	The length of time(hours and minutes) the movie runs for.
License	The movie's rating.

Table 1: Description of the fields present within the dataset.

You are required to clean and investigate the data and address the concerns of the film studio. You must be mindful of missing data and are also required to perform data imputation where necessary (**run-time** should be in minutes, **genre** should be one hot encoded and **release date** and **license** are up to your discretion). Please give the necessary explanations for all your decisions. The film studio also expects that you provide at least two investigations of your own based on the data provided. Furthermore, you are also required to perform outlier analysis on any feature of your choice. Are there any additional data that you can use to supplement the current data and to further enhance the reliability of your recommendations and analysis?

Tip: Regarding the number of plots, there should be no less than 6 plots and no more than 10. Use appropriate narratives (via text/markdown code) for each plot in order to describe your findings and tell your story based on the data.

The dataset can be downloaded from the following link: *Click here to access file*.

Guidelines

1. You may use any of the python visualization libraries used in tutorials for graphs/plots.
2. You are required to submit a Jupyter notebook file.
3. There would be no need to submit the dataset file.
4. Mark scheme for Assignment 1:

Item	**Marks**
Data Ingestion	1mk
Data Summary & Investigation	5mks
Data Manipulation and Imputation	20mks (**Genre, Release Date, Run Time, License**)*
Reasoning for Data Imputation	9mks
Data Visualisation	18mks
Inferences and Analysis	16 mks
Outlier Analysis	5mks
Inclusion of Additional Data	5mks

Submission Details

- Ensure that your notebook is named according to the following format: **firstname_lastname_idnumber.ipynb**
- Export/download your file from Jupyter notebook.
- Email your file to: **comp3610@gmail.com**
  Put as the title of the email: **COMP 3610-A1**
- There would be a 10% penalty per day for late submissions, up to 5 days.