

COMP 6940: Big Data and Visual Analytics

Assignment 1

University of the West Indies, St. Augustine

Due: Feb. 23rd 2024 @ 11:59PM

1 Introduction

This assignment is designed to provide hands-on experience with the end-to-end process of working with big data, from scraping web data to preprocessing, analysis, and visualization. You will work with real-world data from two distinct sources: the Trinidad and Tobago Lotto Plus and the highest grossing Hollywood movies.

2 Part 1: Trinidad and Tobago Lotto Plus

You are tasked with creating a small containing the results of the Trinidad and Tobago Lotto Plus spanning the years of 2010 to 2023.

2.1 Tasks

2.1.1 Data Collection

- Visit the NLCB Lotto Plus results page: <https://www.nlcbplaywhelotto.com/nlcb-lotto-plus-results/>
- Scrape all Lotto Plus results from January 2010 to December 2024. Your dataset should contain the following information: Draw#, Numbers, Power Ball, Multiplier, Jackpot, Wins, Draw Date.

2.1.2 Data Preprocessing

- Clean and preprocess the data as necessary. An 'X' in the data signifies that the specified value is not available.
Note: Draw date should be a date in the format dd-mm-yyyy, numbers should be transformed to be either a numerical list or encoded as multiple numerical values, and draw#, powerball, multiplier, jackpot and wins should be numerical values.
- Clearly explain all preprocessing decisions made.

- Save the final dataset to `a1_lotto_plus_idnumber.csv` with headers.
- Provide summary statistics for the dataset.

3 Part 2: Analysis of Highest Grossing Hollywood Movies

You are hired by Universal Pictures to analyze the top highest grossing Hollywood movies, providing insights and recommendations for future film productions. The dataset, ‘top_hollywood_grossing_movies.csv’ can be found alongside this document.

3.1 Dataset Description

| 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 |

Table 1: Description of the fields present within the dataset

3.2 Tasks

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.](#)

4 Guidelines

1. You may use any of the Python visualization and data processing libraries used in tutorials for graphs/plots and data manipulation.

4.1 Mark Scheme

4.1.1 Part 1: Trinidad and Tobago Lotto Plus Data Analysis

| Item | Marks |
|-----------------------------|------------|
| Scraping | 10 marks |
| Preprocessing + Explanation | x 16 marks |
| Summary Statistics | 2 marks |
| Saving Data | 2 marks |

4.1.2 Part 2: Analysis of Highest Grossing Hollywood Movies

| Item | Marks |
|----------------------------------|----------|
| Data Ingestion | 1 mark |
| Data Summary & Investigation | 5 marks |
| Data Manipulation and Imputation | 15 marks |
| Reasoning for Data Imputation | 9 marks |
| Data Visualisation | 16 marks |
| Inferences and Analysis | 18 marks |
| Outlier Analysis | 5 marks |
| Inclusion of Additional Data | 10 marks |

**For Data Manipulation and Imputation, special attention should be given to Genre, Release Date, Run Time, and License.*

5 Deliverables

Students are required to submit the following:

1. A detailed Python notebook containing all code written, documenting the process for part 1 named **a1_lotto_plus_idnumber.ipynb**.
2. A single CSV file for the scraped and cleaned dataset, named **a1_lotto_plus_idnumber.csv**.
3. A detailed Python notebook containing all code written, documenting the process for part 2 named **a1_movies_idnumber.ipynb**.
4. Zip all files into a single file called **a1_idnumber.zip**.
5. Do **not** submit the movies dataset.
6. See the Course Github Page for further submission details.