Programming for Data Analytic
SOFT8032
Second Examination

November 2022

1 Second Assessment. First Project

This project contributes 30% in your final mark. This is an individual project and has to be all done by yourself. You are not allowed to disclose your code to anyone else. You may be called to explain different parts of your submission, if needed.

Any question regarding the project should be communicated with farshad.toosi@mtu.ie or Canvas message.

1.1 Dataset Overview

For this project we are going to perform a number analytical tasks on the **movie_metadata.csv** file.

1.2 Project Specification

The objective of this project is to provide an insight into the underlying pattern of the dataset such as statistical details of different features and etc. Please perform the following tasks:

1. Each movie has a feature that shows the imdb score of the movie. For this task you are required to find the average of imdb score for the following sub-groups:

   - A group of movies where the **director** of the movie is also the first **actor** of the movie.
   - A group of movies where the **director** of the movie is also the second **actor** of the movie.
   - A group of movies where the **director** of the movie is also the third **actor** of the movie.
   - A group of movies where the **director** of the movie is not acting as the first, second or third actor.

   Use an appropriate visualization technique and visually compare the average of imdb score of the above groups. Interpret your findings.

2. Each movie is either *color* or *black and white*. Use appropriate visualization technique and display the number of **color** movies and the number of **black and white** movies. Perform appropriate data cleaning before visualization so at the end you will have only two unique values in the dataset. If a movie type is Unknown, they need to be excluded from this analysis. Explain your decisions for data cleaning and finally interpret the results.

3. Each movie may contain one or more than one genre(s). If a movie has more than one genre, the genres are separated by pipe symbol (i.e., |). You are required to first extract all the unique genres and then find the top 5 popular genres within the dataset. And finally apply an appropriate visualization to visually depict the population of each genre.

4. Each movie has a length (duration). For this task, you are required to analyse this feature of the movies and visually depict the collection of movies' duration as follows:

   (a) Remove all cells that are either empty of have non numerical values.

   (b) Apply boxplot and visually depict the distribution of the movies' duration, e.g., what is the minimum, maximum, median, first quartile and third quartile.

   (c) Apply another visualization (simple diagram) to display all the movies' duration in a sorted ordering. The low-outliers and high-outliers should be visualized with different colors. Blue indicates the low-outliers (duration that are smaller than the minimum) and red indicates high-outliers (duration that are larger than the maximum) See Figure 1 as a reference for this task.

5. Each movie has a budget in the input file. In this task, you are required to extract a subset of all movies (**targetSet**) where their budget falls in the range of A and B. A and B are identified as follows:

   (a) The median of the entire collection of movie budgets is identified and named C.

   (b) Two sub-collections are created: **Sub1:** movies with budget less than C and **Sub2:** movies with budget greater than C.

   (c) The median of **Sub1** is called A and the median of **Sub2** is called B.

   Use an appropriate visualization technique and visually depict the common movie-budgets in **targetSet**. Explain your finding.

Note that all visualization plots need to have proper labels and annotations if needed. Lack of visualization features attracts penalty.

1.3 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function with interpretation as a comment below the function. You may define extra functions if needed.

Please write your name and student ID as a comment in the designated area in the provided template python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the python file should be named: s1234567.py

The deadline for this project is 20th Nov 2022. One-week late submission with 10 marks penalty would be accepted and the deadline would be 27th Nov 2022.

Any question about this project should be communicated with Farshad Ghassemi Toosi farshad.toosi@mtu.ie or via Canvas.

Please submit your project via Canvas and submit only and only **ONE PYTHON FILE**.

1.4 Rubric

This rubric is subject to change.

1. Correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (100%)
2. Relatively correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (70%)
3. Partly correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (40%)
4. Wrong task implementation. (0%)

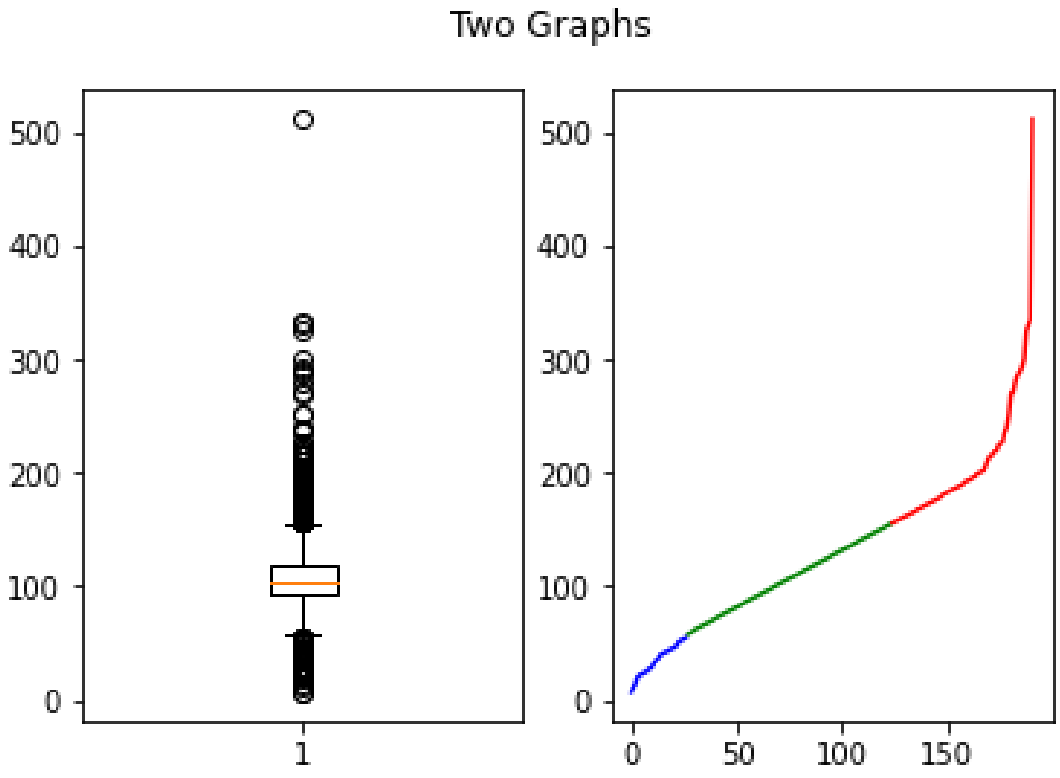

Figure 1: Movie Lengths