Resources / Assignments (/COMP9321/20T1/resources/41975) / Week 3 (/COMP9321/20T1/resources/41976) / Assignment-1

Assignment-1

Specification | Make Submission | Check Submission | Collect Submission

The assignment data has been extracted from a Movie dataset on Kaggle (https://www.kaggle.com/rounakbanik/the-movies-dataset) , with some minor modification to make things interesting. The dataset is split into two CSV files **credits (https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/credits.csv)** and **movies (https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/movies.csv)** . Use the datasets to answer the following questions:

- **Question 1: (based on the both datasets) (0.5 Mark)**
  Join the two datasets based on the "id" columns in the datasets, keeping the rows as long as there is a match between the id columns of both dataset (do not concatenate the datasets).

- **Question 2: (based on the dataframe created in Question-1) (0.5 Mark)**
  Keep the following columns in the resultant dataframe (remove the rest of columns from the result dataset):
  ' *'id', title', 'popularity', 'cast', 'crew', 'budget', 'genres', 'original_language', 'production_companies', 'production_countries', 'release_date', 'revenue', 'runtime', 'spoken_languages', 'vote_average', 'vote_count'*'

- **Question 3: (based on the dataframe created in Question-2) (0.5 Mark)**
  Set the index of the resultant dataframe as 'id'.

- **Question 4: (based on the dataframe created in Question-3) (0.5 Mark)**
  Drop all rows where the budget is 0

- **Question 5: (based on the dataframe created in Question-4) (1 Mark)**
  Assume that there is a ranking scheme for movies defined by " *(revenue - budget)/budget* ". Add a new column for the dataframe, and name it "success_impact", and calculate it for each movie based on the given formula.

- **Question 6: (based on the dataframe created in Question-5) (1 Mark)**
  Normalize the " *popularity* " column by scaling between 0 to 100. The least popular movie should be 0 and the most popular one must be 100. It is a float number.

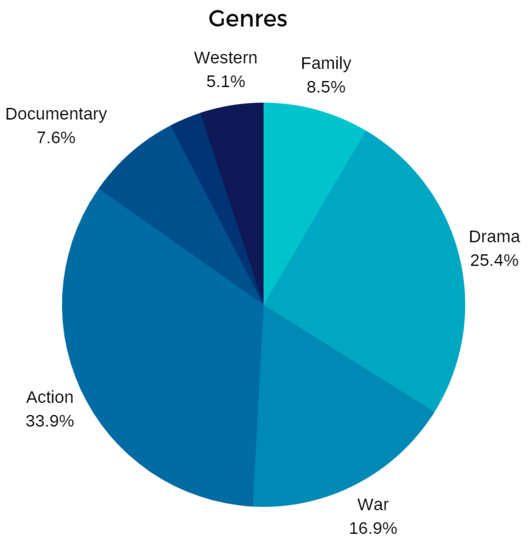
- **Question 7: (based on the dataframe created in Question-6) (0.5 Mark)**
  Change the data type of the "popularity" column to (int16).

- **Question 8: (based on the dataframe created in Question-7) (1.5 Marks)**
  Clean the "cast" column by converting the complex value (JSONs) to a comma separated value. The cleaned "cast" column should be a comma-separated value of alphabetically sorted characters (e.g., Angela, Athena, Betty, Chester Rush) . NOTE: keep unusual characters e.g., '(uncredited)' as they are; no need for further cleansing.

- **Question 9: (based on the dataframe created in Question-8) (1.5 Marks)**
  Return a list, containing the names of the top 10 movies according to the number of movie characters (Harry Potter! is one character! do not count the letters in the title of movies!). The first element in the list should be the movie with the most number of characters.
  UPDATE: You can assume that there is no COMMA in the characters.

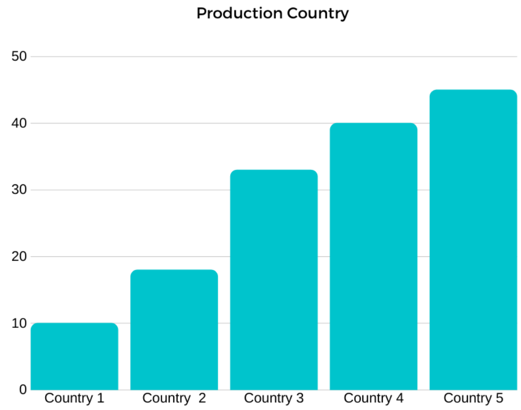
- **Question 10 : (based on the dataframe created in Question-8) (1 Marks)**
  Sort the dataframe by the release date (the most recently released movie should be first row in the dataframe)

- **Question 11: (based on the dataframe created in Question-10) (2 Marks)**
  - (**1 .5 Mark**) Plot a pie chart, showing the distribution of genres in the dataset (e.g., Family, Drama).
  - (**0.5 Mark**) Show the percentage of each genre in the pie chart. Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all genres in the dataset.
  UPDATE: You can add a legend to your chart if labels overlap.You can also merge the some of the infrequent labels (up to 4) and name them "other genres".

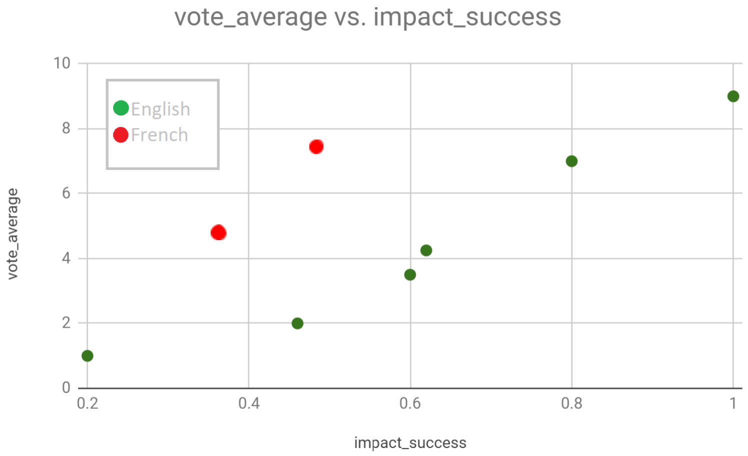


- **Question 12 : (based on the dataframe created in Question-10) (2 Marks)**
  - (1.5 Marks) Plot a bar chart of the countries in which movies have been produced. For each county you need to show the count of movies.

- **(0.5 Mark)** Countries should be alphabetically sorted according to their names.
Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all countries in the dataset.



- **Question 13: (based on the dataframe created in Question-10) (2.5 Marks)**
- (1.5 Marks) Plot a scatter chart with x axis being "vote_average" and y axis being "success_impact".
- (0.5 Marks) Ink bubbles based on the movie language (e.g, English, French); In case of having multiple languages for the same movie, you are free to pick any one as you wish.
- (0.5 Marks) Add a legend showing the name of languages and their associated colors. **UPDATE: You can use both "original_language" (e.g. "en", "fr") or "spoken_languages" .**



Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all countries in the dataset. (also the x and y axis should be swapped in the figure)

What not to forget!

Due Date: Friday the 13th of March 2020 17:59

Submit your script named " YOUR_ZID .py" (z2123232.py) which contains your code.
You are required to use the following code template (**it is not complete; please download the file**) for your submission: