Introduction to Data Science — Assignment 2

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TA(s): Shahrzad Javidi, Melika Sadeghi

Deadline: Friday, Farvardin 17th, 11:59 PM

Introduction

This assignment includes specific tasks to investigate open-ended questions. The open-ended questions ask you to think creatively and critically about how the plots you have created provide insight into the data.

Task 1

TA: Melika Sadeghi

The provided dataset(task1.csv) contains information about the passengers of the sunken ship 'RMS Lusitania'. In this task, you will become familiar with working with numpy, pandas, and matplotlib. This is a brief explanation of the columns in the given dataset:

- survived: Indicates if a passenger survived (1) or not (0).
- pclass: The ticket class (1 = First class, 2 = Second class, 3 = Third class).
- sex: The passenger's sex (male or female).
- age: The passenger's age in years.
- sibsp: The number of siblings or spouses the passenger had aboard the ship.
- parch: The number of parents or children the passenger had aboard the ship.
- fare: The fare the passenger paid for the ticket.
- embarked: The port where the passenger embarked (C = Cherbourg, Q = Queenstown, S = Southampton).
- class: The class of the ticket the passenger had (First, Second, or Third).
- who: Categorizes passengers as 'man', 'woman', or 'child', likely derived from age and sex.
- adult_male: A boolean indicating if the passenger is an adult male or not.
- deck: The deck the passenger's cabin was on, indicated by letters.
- embark_town: The town from which the passenger embarked, corresponding to the 'embarked' codes.
- alive: Indicates if the passenger survived ('yes') or not ('no')

- alone: A boolean indicating if the passenger was traveling alone (no siblings, spouses, parents, or children aboard)

Questions

1. First, read the file using the pandas library and save it in a pandas dataframe. Then, using the methods 'info', 'head', 'tail', 'describe' from the pandas library, examine the general structure of the data, and explain what information each of the outputs shows.

2. Show the type of each data column. Some columns are of type categorical and some are of type numerical from the pandas library info. To process the non-numerical columns, one of the possible methods is labeling; in such a way that each of the categories is replaced by a number. For example, in this dataset, there is a categorical column named sex, which includes values Male and Female. Modify the values of this column, so that each of these models is mapped to one of the numbers in the range [0, 1].
3. Generate a heatmap of the correlation matrix for numerical features in the dataset.
4. Columns that are not on the main diagonal and have a correlation of 1 mean that one column is extra and can be removed. Therefore, based on this, delete the extra columns that meet these conditions.

5. How many passengers survived (survived == 1) the disaster?

6. Find all female passengers aged more than 30 years. How many are there?

7. Identify passengers who embarked from Cherbourg ('C') and paid a fare greater than $100.

8. Identify columns with missing values. Propose and apply a strategy for handling these missing values.

9. What is the average age of passengers on the ship? How does it differ between males and females?

10. Is there a correlation between fare paid and survival rate? Provide a statistical summary.

11. Use Matplotlib to plot the proportion of passengers that survived by class.

12. Plot the age distribution of passengers, distinguishing between those who survived and those who didn't.

13. Create a scatter plot showing the relationship between age and fare paid, color-coded by survival.

14. Create a pivot table to show the average fare and survival rate for each class and sex combination.

15. Plot a grouped bar chart using Matplotlib to show the average fare paid by passengers, grouped by their class and survival status.

Task 2

TA: Shahrzad Javidi

Welcome to this dataset(task2.csv), which focuses on data scientist salaries across different regions from 2020 to 2024. Together, we'll delve into its insights. Let's get started!

Questions

- At first, it's better to do some pre-analysis to delete duplicates and NA (missing) data if they exist. Then, since salaries are inherently tied to each respective country's currency, we need to standardize them to a single currency for meaningful comparisons. Let's begin by identifying the currencies present in the dataset. Given the many currencies, let's examine their frequencies and remove data associated with currencies represented fewer than ten times.
- In this phase, we'll convert these currencies to USD. You can do this in two ways: manually searching for the exchange rates via online resources such as Google or utilizing software packages like Forex-Python for streamlined conversion.
- Now, leveraging the insights gleaned from your dataset, let's employ various Exploratory Data Analysis(EDA) techniques to extract valuable insights. For instance, we could identify the top 10 most popular job titles or the top 10 highest salaries. Wishing you the best of luck!
    - It's important to note that thorough analysis and interpretation of your findings and visualizations are crucial for a comprehensive evaluation in this section.
    - Hint: Based on what you have learnt so far, you need to apply different visualization techniques. Also, you need to check the distribution of variables. It's up to you to decide which plot suits well for which variable(s). Your task will be considered 'successful' if you have gained enough insights from this dataset.

Notes

- Upload your work as a zip file in this format on the website: DS_CA2_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.