Full Stack Development with AI
Lab 7.3 – Data Visualisation

Lab Overview

In the previous Lab 7.2, you have performed some data preparation, preprocessing and analysis tasks using the Pandas and NumPy libraries. In this lab, you will continue to use these two libraries in conjunction with the Matplotlib and Seaborn libraries to perform data visualisation.

Data visualisation is one of two techniques of exploratory data analysis (EDA), the other being statistical measures. It is very useful to help develop a better understanding of your data being further analysis is performed. When performed appropriately, EDA can help to improve the efficiency and effectiveness of downstream data analysis tasks.

Similar to Lab 7.1 and 7.2, you are strongly encouraged to use the Jupyter Notebook format for this lab.

Exercise 1 – Data Visualisation with Seaborn

This exercise is based on the combined panel dataset of the Forbes 2000 list of companies from year 2008 to 2022 that you have worked with in Lab 7.2. In particular, you will focus on the latest year 2022 Forbes data.

1. Import the combined panel dataset that you have generated in Exercise 2 into a Pandas DataFrame.

   Filter the financial data for Forbes 2000 companies in year 2022.

2. Seaborn (https://seaborn.pydata.org/) is a Python data visualisation library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. One useful visualisation supported by Seaborn is heat map.

   Seaborn can be installed with the following command using `pip`:

   ```
   python -m pip install seaborn
   ```

   Generate a heat map for Market_Value using Country as the row header and Industry as the column header.

   ***Hint**: You need to use Pandas to create a pivot table that aggregates the summation of Market_Value along the two required dimensions before generating the heatmap.*

   What information or insights can you obtain from this heat map?

Exercise 2 – Data Visualisation with Matplotlib

Who have survived the Titanic shipwreck? The Titanic dataset hosted on Kaggle (https://www.kaggle.com/competitions/titanic) is a legendary dataset used for predicting which passengers survived the Titanic shipwreck with machine learning models. You are given titanic.csv in data.zip, which is the training dataset of the Kaggle competition.

In this exercise, instead of performing prediction, you will be using Pandas and Matplotlib to perform exploratory data analysis with data visualisation to get a sensing which passengers had a higher likelihood of surviving.

titanic.csv consists of 12 variables and 891 records:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Some additional notes about the following variables in titanic.csv:

- pclass – A proxy for socio-economic status (SES)
  - 1st = Upper
  - 2nd = Middle
  - 3rd = Lower
- age:
  - Age is fractional if less than 1.
  - If the age is estimated, it is in the form of xx.5

- sibsp:
    - The dataset defines family relations as:
        - Sibling = brother, sister, stepbrother, stepsister
        - Spouse = husband, wife (mistresses and fiancés were ignored)
- parch:
    - The dataset defines family relations as:
        - Parent = mother, father
        - Child = daughter, son, stepdaughter, stepson
        - Some children travelled only with a nanny, therefore parch=0 for them

Study the Titanic shipwreck dataset carefully and generate some useful data visualisations to help obtain insights on which passengers had a higher likelihood of surviving the shipwreck.

Explain the intuition behind each insight, e.g., why a particular sub-group of passengers had a higher observed likelihood of survivability compared to other sub-groups.

***Hint**: You need to generate pivot tables to aggregate the data before creating charts.*

-- End of Lab --