Skills for Hire
Data Analytics
Assignment 2 - Python

Overview

Deadline: October 15, 2023 - 11:59 pm Atlantic Time
Credits: 40% of your assignment grade

Instructions

This assignment comprises 3 questions on the following topics: Data Visualization, Sentiment Analysis and Statistics. You should answer **all 3 questions**. Each question has an associated number of marks included at the end of the question. The dataset required for the question is located below the question topic. Each question begins on a new page below.

You are **required** to write the code in Python. You should write your code in a notebook (Google Colab or Jupyter). You are free to use any library you like. Prior to your submission, you need to run your notebook and ensure that the submitted version has your cells run.

Recommended Libraries: pandas, numpy, matplotlib, seaborn, nltk, scipy, etc.

Submission Requirements

File format: Notebook (.ipynb)
File name: DA_**Group**_Assignment2_**StudentID_FirstName_LastName_EmailAddress**
** Make sure to replace any bolded placeholders with the appropriate data, eg. group name

Submission Process

Submit via Google Form:
https://forms.gle/VhVQ3XXzqGa7Xusc8

Datasets and Sample Code

We have provided your datasets and sample code in Google Drive:
https://drive.google.com/drive/u/0/folders/1F1ya9jq_35Rn4uplGdgLBxtutzmRppt_

Questions begin on the next page.

Question 1: Data Visualization [15 marks]

Dataset: titanic.csv

Part 1: Data Exploration [4 points]

1. Load the Titanic dataset into a pandas DataFrame. [1]
2. Explore the first few rows of the dataset to understand its structure. [1]
3. Calculate and visualise the basic statistics (mean, median, etc) for the numeric columns [1 correct numeric + 1]

Part 2: Visualization [11 points]

1. Create a bar chart to show the distribution of passengers by class (1st, 2nd, 3rd). [1]
2. Create a histogram to visualise the age distribution of passengers. [1]
3. Calculate and visualise the survival rate of passengers. Create a pie chart to represent this information. [1]
4. Create a bar chart to show the count of passengers by gender. [1]
5. Create a stacked bar chart to visualise the survival rate by gender. Explain any gender-based differences in survival. [2]
6. Create a count plot (bar chart) to show the number of passengers who embarked from each location (S, C, Q). What do you notice about the embarkation points? [2]
7. Create a box plot to visualise the distribution of family sizes among passengers. [1]
8. Create a scatter plot to visualise the relationship between fare and age. Do passengers who paid higher fares tend to be older? [2]

Question 2: Sentiment Analysis [25 marks]

Dataset: Corona_NLP.csv

Part 1: Data Loading [2 points]

1. Load the dataset Corona_NLP.csv [1]
2. Show the last 10 rows of the dataframe [1]

Part 2: Manipulation and Visualisations [8 points]

The independent attributes that convey information are:

1. *Location*
2. *Time (TweetedAt)*
3. *Text (OriginalTweet)*

It should be noted that although we will not be using the TweetedAt and Location columns in our analysis, we will still conduct a minor exploratory data analysis (EDA) on these columns.

1. Check if there are any duplicate rows. If there are any, remove the duplicates. [1]
2. (a) In the Location column, find unique values and their corresponding frequencies (number of times they appear) [1]

   (b) Sort the values in descending order. [1]

   (c) Using a bar chart, visualise your results for the top 20 locations. [1]

3. You will see that Sentiment can take 5 values: "Extremely Negative", "Negative", "Neutral", "Positive", "Extremely Positive". Using an appropriate color palette and an appropriate type of graph, visualise how Sentiment is distributed. [2]
4. Change the labels of part of the data, combining "Extremely Negative" and "Negative" into “Negative” and "Positive" and "Extremely Positive" into positive. [2]

Part 3 - Cleaning [8 points]

Perform the following text cleaning steps. I suggest that you do subparts 1, 2 and 3 in the same function:

1. Remove hyperlinks [1]
2. Remove hashtags [1]
3. Replace \n and \r with an empty space [2]
4. Using TweetTokenizer in nltk.tokenize, perform tokenization. Sample code is provided. You can use another function if you wish. [1]
5. In the same function, remove stopwords and punctuation signs. [1]
6. Perform stemming using PorterStemmer in nltk.stem [1]
7. Apply these functions to the dataset and visualise the first 5 rows. [1]

Part 4 - Model development [7]

1. Create a tf-idf vectorizer and fit the function. [1]
2. Keeping only the columns Tweet and Sentiment, split your dataset into training (80%) and test (20%) sets. [1]
3. Build a logistic regression model and fit the data. [3]
4. What is the accuracy of the model? [1]
5. Generate a classification report. [1]
6. BONUS: Comment on the metrics. [0]

Question 3: Statistics [10 points]

Dataset: q3_data.csv

1. What is the mean, mode, median and interquartile range for x1 and x2? [4]
2. What is the correlation coefficient between x1 and x2? [1]
3. Fit a regression line and visualise the data points. [2]
4. Comment on the correlation coefficient, the regression line and your scatter plot. [3]