Amal Koodoruth
Harsh Khilnani

Assignment 2 - Python

Skills for Hire Atlantic

Due date: April 16 2023

Overview

This assignment is broken into two parts:

1. **Part 1 [20 marks]**: Analysis on New York City Restaurants
2. **Part 2 [30 marks]**: Sentiment Analysis on Tweets

You will be graded on the accuracy of your results and on the quality of your plots. Please make sure to clearly indicate the question parts before answering them and to comment your code. The marks for each problem is indicated between brackets.
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.
Recommended libraries: pandas, numpy, matplotlib, seaborn, etc.

Note: This assignment will count for 30% of your *Assignments* grade.

Submission Instructions

Submit through Google Form:
`https://forms.gle/wNRhp8ZfxQdQNmL29`

Submission Requirements

File format: Notebook (.ipynb)
File name: DA_**Group**_Assignment2_**FirstName**_**LastName**_**EmailAddress**

** MAKE SURE TO REPLACE **Group** with your Group colour, **FirstName** with your first name, **LastName** with your last name and **EmailAddress** with your email address**

Precaution

In this assignment, you will be asked to draw conclusions from your results. It is very common to misinterpret data. While completing this assignment, one important concept that should considered is the Simpson's Paradox. For more information on this, we invite you to read this short 7-minute-read article.
`https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765`

1 Part 1

In this part you are given a dataset on restaurants in New York City.
The aim of this part is to make some visualizations and draw insights from them.
Dataset for this part: `food_order.csv`

1.1 Data Loading

Task 1: Load the dataset `food_order.csv` into a pandas DataFrame. [1 mark]
Task 2: Display the first 10 rows of your dataset. This will give you a broad idea of how your data looks like. [2 mark]

1.2 Data Description

Task 3: Identify the data types for each feature (please do not eyeball it). You will need to include that in your report. [1 mark]

Before doing further analysis, we need to clean our dataset. More specifically, we want to identify (and adjust for) missing values.

Task 4: Write a function that calculates the ”*null rate*”for each column. The *null rate* is the number of null values as a percentage of the total number of samples in the dataset. [1 mark]

Task 5: Identify the number of unique values for each column. [1 mark]

1.3 Data Manipulation

Task 6: Add one column, `total_time`, that is defined by the sum of `food_preparation_time` and `delivery_time`. Units for `food_preparation_time` and `delivery_time` are in minutes. So should the units for `total_time` be. [2 mark]

1.4 Satistics

Task 7: Identify the different cusine types. [1 mark]

Task 8: Find the number of restaurants per cuisine type. [1 mark]

1.5 Visualization

In this section, you will be required to make plots that can be used to draw conclusions. Unless mentioned otherwise, you can use any plot that you believe is appropriate. Note that you will be graded on whether plot conveys the message that you want to convey. Remember to choose an appropriate (and consistent) colour palette to make your plots visually pleasing. Feel free to add anything that you think will make the graph look better.

Task 9:

(a) Using your results from Task 8, and using a pie chart display the count of restaurants per cuisine type. [1 mark]

(b) Show only the 5 cuisine types with most restaurants. [1 mark]

Task 10:

(a) Plot a graph that shows the 10 most popular `restaurant_name`. [1 mark]

(b) Highlight the top 3. [1 mark]

Task 11: Plot a graph that shows the 10 most popular restaurants for each `cuisine_type` and highlight the top 3. If there if less than 10 restaurants for a particular cuisine type, just show the maximum number of restaurants and highlight the top 3. [2 mark]

Task 12:

(a) What is the proportion of reviews for *Shake Shack*? [1 mark]

(b) Use a pie chart to display your results. [1 mark]

Task 13: Is there a link between `ratings` and `food_preparation_time`? Justify your answer. You may use an appropriate graph. [2 marks]

[Total marks: 20]

2 Part 2

The aim of this part is to carry out sentiment analysis.
Dataset for this part: `tweets.csv`

2.1 Data Loading

Task 1: Load the dataset `movies_shows.csv` into a pandas DataFrame. [1 mark]

Task 2: Display the last 10 rows of your dataset. This will give you a broad idea of how your data looks like. [1 mark]

2.2 Data Cleaning

An important step in Data Science is to clean data. In this sub part, you will be preparing the dataset for future analysis. You might want to use the libraries `re` and `TextBlob` for this part. Complete these tasks and store the cleaned tweet in a column called `cleaned_tweet`.

regex: `https://docs.python.org/3/library/re.html`
TextBlob: `https://textblob.readthedocs.io/en/dev/`

Task 3: Remove hyperlinks. [1 mark]

Task 4: Remove stopwords. Stopwords are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. [2 mark]

Task 5: Remove mentions. Here you want to remove the symbol '@' and the word that comes after it. [2 marks]

Task 6: Remove the hashtag. Here you want to remove the symbol '#' and keep the text that comes after it. [2 marks]

Task 7: Remove punctuation signs. [1 mark]

Task 8: Remove the word "RT". [1 mark]

Task 9: Remove emojis. [2 marks]

Task 10: Remove leading and trailing whitespaces. [2 marks]

Task 11: Apply lemmatization on every word. [2 mark]

Task 12: Store the cleaned tweet in a new column called `tweet_cleaned`. [2 marks]

2.3 Sentiment Analysis

Sentiment Analysis is process of classifying text into categories- positive, negative or neutral. One way to do this is by manual annotation. However, we would not do that in this assignment.

Task 13:

(a) Can you think of reasons why we would or would not manually annotate our whole dataset? [2 marks]

(b) Can you propose alternative ways to do this? [1 mark]

Instead of manual annotation, we would first calculate the *Polarity* of each tweet. In a nutshell, polarity measures how positive and how negative a text is. Please refer to this link for more details:
`https://pythonalgos.com/natural-language-processing-what-is-text-polarity/#:~:text=In%20short%2C%20text%20polarity%20is,negative%20emotions%20in%20a%20sentence.`

```python
from textblob import TextBlob

#Create a function to get the polarity
def getPolarity(twt):
  return TextBlob(twt).sentiment.polarity
```

Listing 1: Code snippet to get polarity

```python
#Create a new column to save the results of the created function
df['Polarity'] = df['cleaned_tweets'].apply(getPolarity)
```

Listing 2: Code snippet to apply the function on the relevant column

The code above will get the polarity for each tweet in the dataset.

Task 14: Add a column `sentiment` to the dataframe. If the polarity is >0, the sentiment is positive. If it is <0, the sentiment is negative. Else, the sentiment is neutral. [2 marks]

Task 15: Use an appropriate plot(s) to show how sentiment is distributed in the dataset? [1 mark]
Question: Comment on this. [1 mark]

Task 16:

(a) What are the 10 most common words in each sentiment? [1 mark]

(b) Use an appropriate plot(s) to visualize this. [1 mark]

Question: Comment on this. [1 mark]

Task 17: Use an appropriate plot to show the number of `Likes` and `Retweets` per sentiment. [1 mark]
Question: Comment on this. [1 mark]

[Total marks: 30]