Python Final Project

Karim Ashraf – Ali Ayman – Marwan Saad

Overview

This dataset is for the British Airways and contains data about the customers and customer reviews. Where these customers answered questions about their experience consuming the service provided by the company.

Code Documentation:

* In the beginning the libraries Pandas, NumPy and seaborn were imported and given short names to be used later in calling their functions
* Then the dataset was loaded in order to see the full data before any work was done
* For the original dataset, we found the shape of the data (rows , columns)
* Then we printed the first and last 5 instances of the original dataset to get a clear overview of the data in it
* At this point information about the data is needed to know about the data types of each column
* Then we used the describe function to provide an overview on the statistics of the data and find the count, mean, standard deviation, min value, max value, mode, first quartile and third quartile
* Then we determined the count of null values in each column to prepare for the cleansing process
* Then we printed all the column names to use them later

Our First class (DataProcessor):

* There are 2 methods in this class and a constructor function where we define the attributes which is the file name
* The first method is the function named load\_data which uses pandas to read the csv file and load the data
* The second method named clean\_data removes the columns 'food\_and\_beverages', 'wifi\_and\_connecticity' and 'wifi\_and\_connectivity’ due to the high quantity of null values and the incorrect type of data in the column 'food\_and\_beverages' where it is stored as an object and it’s purpose is to be an int or float. The function also removes the rows that haas null values and resets the index returning a dataframe named data
* An instance is then created to load the clean data ready to be used
* Then we checked for the new data shape and also checked for the sum of null values in each column which should be zero now

Our Second class (DataAnalyzer):

* There are 4 methods in this class, the first one gets the average of a column in relation to another column using the mean function where it is given in its parameters both column names and the data
* The second method find the distribution of data in a certain column using the column data and its value counts where it’s given in it’s parameters the data and the column name
* The third gets the median of a certain column using the median function after getting the value counts where it’s given in it’s parameters the data and the column name
* The last method gets the mode of a certain column by finding the value counts and displaying them where the one with the most frequency is the mode and it’s given in it’s parameters the data and the column name
* Then instances are made to find some data using the methods

Our Third class (DataVisualizer):

* There 7 methods in the class, the first method prints a line chart in order to find the relationship or compare between 2 columns or find the change among a certain timeline. This method takes the number of instances to work with, the x column and the y column as parameters
* The second method plots a histogram to find the distribution of data of a column and summarize the data on an interval scale. This method takes the data, the column name and the number of bins as parameters
* The third method plots a pie chart to show the percentage of a certain column from the whole. This method takes the data and the column name as parameters
* The fourth method plots a scatter plot and a line of best fit in order to show the relationship between 2 numeric variables and uses the pallet tab10 to colour the diagram and the hue to differentiate between 2 columns used. This method takes the data, the number of instances to work with, the x and y columns and the hue column as parameters
* The fifth methods plots a heatmap to show the correlation between all columns and each other using the correlation of each column to the other and classifying them on an interval from 0 to 1.This method takes the data as a parameter
* The sixth method is a box plot to provide a quick visual summary of the data and show the median, upper and lower quartiles, minimum and maximum values and any outliers in the dataset using the box plot function . This method takes the data and the x and y columns as parameters.
* The last method is the bar plot to find the relation between to columns in the dataset visually using the bar plot function where the classification index = none. This method takes the x and y columns, the hue column, the number of instances to work with and the data as parameters
* Then instances are created for some data