East West University
Department of Computer Science and Engineering

CSE 303: Statistics for Data Science
Course Instructor: Dr. Mohammad Rezwanul Huq

Exploratory Data Analysis using Pandas

Lab Assignment – 1 (Tasks on Lab 04)

Instructions:

- Download the dataset from the course repository (dataset_lab04.csv)
- Create a new python script (.py) file.
- Start solving the following tasks. You must define a function for solving each task. The name of the function should be like - Lab04_Task1_<your-student-id>. Then you must call each function sequentially at the end of the same script one after another, arranged in a different cell.
- A sample script is given below.

```
#%%
# Student ID: AAAA-B-CC-DDD
# Student Name: XXXXXXXXXXX
# Lab Assignment - 1

import pandas as pd
df = pd.read_csv('dataset.csv')
df.info()

#%%
# Task 1
lab04_task1_DMRH()

#%%
# this cell must be executed before calling the function
def lab04_task1_DMRH():
    print ('Number of rows: ', df.shape[0])
    print ('Number of columns: ', df.shape[1])
```

Lab Tasks:

1. How many rows and columns this dataframe has? Print this information.
2. Describe (numerical summary) the time and amount column. Print this information.
3. There are 31 columns in the dataset. Compute some statistical measures like mean, median, standard deviation, variance using Pandas Function for at least two columns. Print this information.
4. Show the Box Plot of Time and Amount column. Also print the value of Q1, Median, Q3, IQR. Are there any outliers? Explain your answer and print it.
5. Show the Histogram of Time and Amount column. Print the value of the Skewness and Kurtosis using appropriate Pandas functions. Comment on the type of the data distribution and print it.

6. Find the percentage of records with class value = 0 (Non-Fraudulent) and class value = 1 (Fraudulent). Print this information.
7. Show the result you have got in 6 using a Histogram.
8. Show the result you have got in 6 using a Bar chart. Create the bar chart on the percentage value, not on the total number of occurrences.
9. Show the Histrogram (data distribution) of a few other columns (your choice) showing both positive and negative skew and also leptokurtic and platykurtic data distribution. So, you should display at least four Histograms.
10. Find the highest positive correlation among all attributes. While finding the correlation, use appropriate code, not manually. Print this information accordingly.
11. Support your findings in Question 10 using a Scatter Plot.
12. Find the highest negative correlation among all attributes. While finding the correlation, use appropriate code, not manually. Print this information accordingly.
13. Support your findings in Question 12 using a Scatter Plot.
14. Create a Box Plot of the Amount Column.
15. Now create two other box plots side by side. The first one will show the Amount column value for which the class value = 0 (Non-Fraudulent) and the second one will show the Amount column value for which the class value = 1 (Fraudulent). Do you find any particular pattern by just considering Amount column. Explain your answer and print it accordingly.

Submission Link:

https://forms.gle/pd46v4G36vn1XRkS9

Submission Deadline:

Thursday, 01 April 2021, by 11:50 AM (Before Thursday's Class).

Late Submission will be automatically graded as ZERO.