National College of Ireland

School of Computing
Module: Data Mining and ML
Week: 1 Lab: 1

The purpose of the lab is to load a sample data and test the python environment install. To complete this lab, students must first install the Anaconda software using the link provided in Moodle.

1. Following the steps below to setup you environment for the lab:
   a) Import the *pandas* library by clicking on Environment in Anaconda.
   b) Then select "All" from the package dropdown.
   c) Then search for pandas in the search area the right top conner.
   d) Select pandas from the populated list and click on Apply to install the it in your environment.
2. Download the Sample.csv file provided in Moodle to a location on your PC
3. Open the Jupyter notebook from the Anaconda.
4. Browse to the location where you downloaded the Sample.csv file and create a new notebook file using python 3.

4. Copy and paste the code below:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

5. Run the following commands in different blocks:

   a) data = pd.read_csv('Sample.csv') -> Load dataset
   b) data.describe() -> Display summary
   c) *data.columns -> Check for column names*
   d) data.drop('product_id', axis=1, inplace=True) -> Remove product Id Column
   e) *data.head()* -> Display the top 5 rows. The default value of 5 can be changed to desired value by passing a different integer as a parameter(eg.: data.head(7))
   f) data.info() -> Display basic information
   g) data['total'] = data.qty * data.unit_price -> Add a new column by multiplying qty column with unit price column.
   h) *data.head()* -> Run the display again to show the new column
   i) *data.isnull().sum() - > Check for missing values*
   j) *Run the code below in one block*

```
plt.figure(figsize=(14, 7))
plt.subplot(2, 2, 1)
sns.histplot(data['qty'], kde=True)
plt.title('Distribution of Quantity')

plt.subplot(2, 2, 2)
sns.histplot(data['unit_price'], kde=True)
plt.title('Distribution of Unit Price')

plt.subplot(2, 2, 3)
sns.histplot(data['total'], kde=True)
plt.title('Distribution of Total Price')
plt.tight_layout()
```

k.

```
plt.figure(figsize=(14, 7))
plt.plot(data['date'], data['qty'],
label='Quantity')
plt.plot(data['date'], data['unit_price'],
label='Unit Price')
plt.plot(data['date'], data['total'],
label='Total Price')
plt.xlabel('Date')
plt.ylabel('Values')
plt.title('Time Series of Coffee Sales')
plt.legend()
plt.show()
```

6. Create a new notebook file and import your data from https://raw.githubusercontent.com/mwaskom/seaborn-data/master/iris.csv

7. Repeat the code in 4 and 5. Save your answer and submit your results for assessment on Moodle.