Report IDS

Students ID:

Khaled Saleh : 2232890

Mohanad Assaf : 2233114

Doctor :

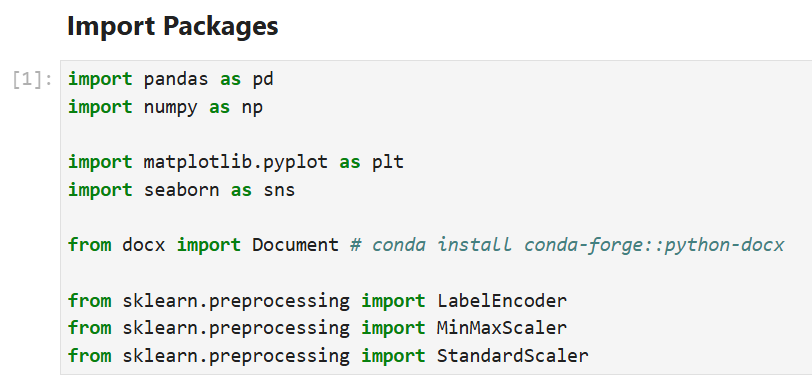
Zaher Ibrahim Saleh Salah.

Data set used :

All.

Date:

5/17/2024

Import several libraries used for data analysis (Pandas and Numpy), data visualization (matplotlib and seaborn), document creation (docx), and data preprocessing (LabelEncoder, MinMaxScalar, and StandardScalar).

We read all the provided CSV files and merge them into a single data frame. This allows for easier data manipulation and function application.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generatedA close up of a computer screen

Description automatically generatedWe took a general overview of the data, examined the datatypes of each column to know what methods should be followed to deal with this column, and then checked for the presence of any null values.

***A1:*** We completed the task as instructed in the assignment, classifying the type of each feature, and then stored the results in a file named A1.docx.

`Website\_Activity` is an ordinal datatype because there is a specific order to the data. 'Seldom' indicates less activity than 'regular', and 'regular' indicates less activity than 'frequent'.

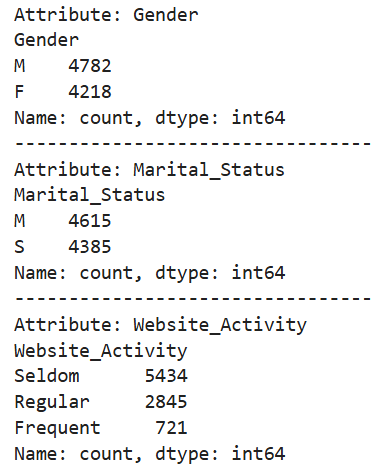
`Age` is a ratio datatype because it has a true zero point. A value of zero represents the absence of age. And there are no negative ages in the world.

The rest of the features are nominal because they have no order and are not numeric, with the exception of `User\_ID`. Although `User\_ID` is numeric, it is also considered nominal because no calculations can be performed on it.

***A2:***

A screen shot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated

We identified the frequencies of values for object datatypes. For the numeric datatype Age (we will not perform any calculations on User\_ID because it is merely an ID and it doesn't make any sense to perform calculations on it),

A screenshot of a computer

Description automatically generated

we calculated location measures (median, mean) and spread measures (variance, standard deviation, range, Q1, Q2, Q3, IQR). After completing these calculations, we stored the results in a docx file named A2.docx.

A graph of a bar plot

Description automatically generatedA graph of bar plot for gender

Description automatically generatedWe created bar plots for the object features to visualize the data distribution.

We plot the gender, as we can see that the males are higher than females.

And in the second plot the married people and single people are very high. So, you should be careful when you do digital marketing for each one of them.

A bar graph with different colored squares

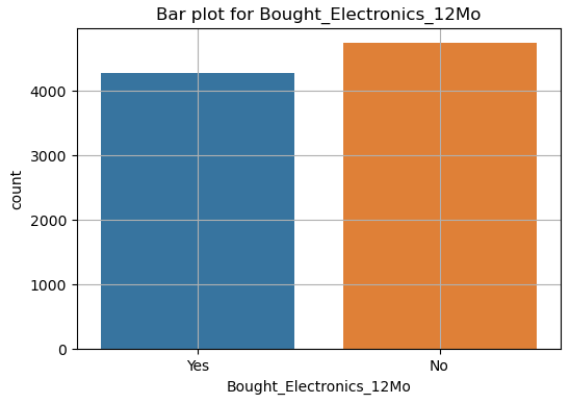
Description automatically generatedA graph with blue squares and orange bars

Description automatically generated

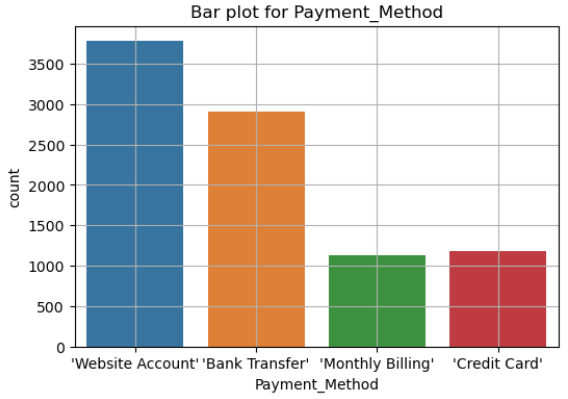
The activity on the website, most of the people are seldom, this is a negative thing. You should do something intereseting to fix that problem.

As we can see the browsed electronics most of then are yes, so you should do some good offers for this to make the website activity frequenet or regular visited.

A bar graph with blue and orange squares

Description automatically generated

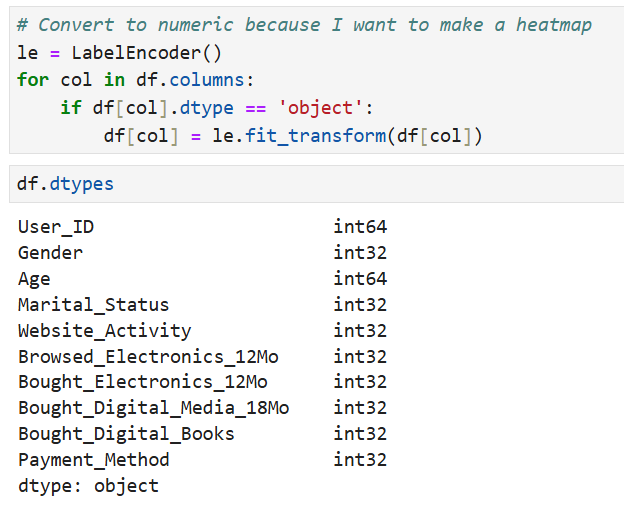
As we can see the bought of the electronics, there are a lot of people did not buy anything of them. So, as I said previously you should do offers on the electronics. In the second plot this is improve for my words before, we say most of the people bought digital thing because it’s cheaper, and he don’t have to wait for it, immediately will reach the user.

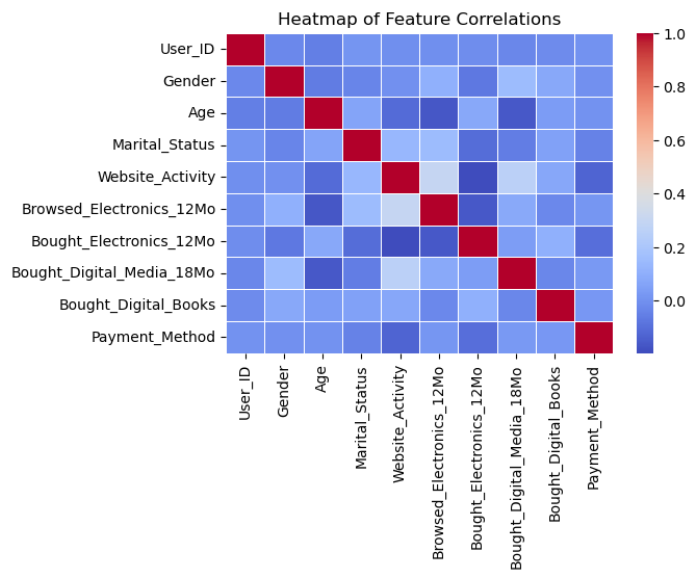


A bar graph with blue and orange squares

Description automatically generated

Most of the people on our website don’t read books at all, but most of them as we said browse the electronics. Most of the people use the website account and bank transactions to pay for their goods in general.

***A3:***

First, I used Label Encoder to convert the nominal features into numeric features, primarily to generate a heatmap and observe any correlations between the features.

However, as we can see from the heatmap, there are no strong correlations between the features, unfortunately.

A graph of a distribution of age

Description automatically generated

I plotted the 'Age' feature to identify the most common life stage among our website users, as our marketing approach differs for each life stage.

A graph of a scatter plot

Description automatically generated

We also created a scatter plot comparing ages and payment methods to understand how different age groups prefer to pay. We plotted this before but this another improves to the results.

A graph with lines and a rectangle

Description automatically generatedA diagram of blue and red dots

Description automatically generated

We used scatter plots and box plots to identify outliers. The outliers the ages that higher than 70 and the ages that approximately lower than 20.

***B1:*** We used the preprocessing technique to minimize the range of the `Age` values (normalize) to make the data balance and easy to understand.

#### **B2:** We categorize the Age features into 5 features as you will see in the csv file.Teenager 1-16, Young 17-35, Mid\_Age 36-55, Mature 56-70, Old 70+

***B3:*** We did this step in the Label Encoder step. 😁

All of these steps stored in an csv file.