Data Science: Cleansing And Visualization For Beginners

This article is an outcome of a short project undertaken by collaborators, persevering in discovering the ropes of Data Science. The names of all the collaborators are mentioned at the end of this article.

Data science is a fascinating discipline that is both artistic and scientific simultaneously. A project journey in Data Science involves extracting and gathering insightful knowledge from data that can either be structured or unstructured. The entire tour commences with data gathering and ends with exploring the data entirely for deriving business value, during which many procedures are applied systematically. Broadly speaking, the cleansing of the data, selecting the right algorithm to use on the data, and finally devising a machine learning function is the objective in this journey. The machine learning function derived is the outcome of this art that would solve the business problems creatively.

This article focuses exclusively on the Data analysis, cleansing, exploration, and imputation of data. I describe the steps that we undertook in this journey, forming the crux of this article.

Step 1. Import Libraries.

We started by importing the libraries that are needed to preprocess, impute, and render the data. The Python libraries that we used are Numpy, random, re, Matplotlib, Seaborn, and Pandas. Numpy for everything mathematical, random for random numbers, re for regular expression, Pandas for importing and managing the datasets, Matplotlib.pyplot, and Seaborn for drawing figures. These libraries are imported with a shortcut alias as below.

```
import numpy as np
import random
import re
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Step 2. Loading the data set.

We were handed over e-commerce data to load. The dataset was loaded and information is displayed in jupyter.

```
ecom = pd.read_csv('Ecommerce Purchases.csv')
ecom.info()
```

These output is :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
```

Column Non-Null Count Dtype

--- ------ -------------- -----

0 Address 10000 non-null object

1 Lot 10000 non-null object

2 AM or PM 10000 non-null object

3 Browser Info 10000 non-null object

4 Company 10000 non-null object

5 Credit Card 10000 non-null int64

6 CC Exp Date 10000 non-null object

7 CC Security Code 10000 non-null int64

8 CC Provider 10000 non-null object

9 Email 10000 non-null object

10 Job 10000 non-null object

11 IP Address 10000 non-null object

12 Language 10000 non-null object

13 Purchase Price 10000 non-null float64

dtypes: float64(1), int64(2), object(11)

memory usage: 1.1+ MB

<<TBD>>

Step 2. Interpreting and transforming the data set.

This step involved loading the data set and trying to understand what the data set contains. In a real-world scenario, the data information that one starts with

could be either raw or unsuitable for Machine Learning purposes. We had to transform the incoming data suitably.

<<TBD>>

Step 3. Cleansing and imputing the data:

We looked into invalid values in the data set ignorer to impute the dataset with clean data in this step.

<<TBD>>

In our case, we introduced errors purposefully to understand how we can impute the data.

<<TBD>>

Step 4. Exploring and Analysing the data: A cleaned up and structured data is suitable for analyzing and finding exemplars using visualization.

Conclusion

<<TBD>>