Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 09/13/2024
Internship Batch: LISUM37
Version:1.0
Data intake by:
Data intake reviewer:
Data storage location: <location URL eg: github, cloud>

Tabular data details:

Cab_Data.csv

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20.2 MB

Tabular data details:

Customer_ID

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1027 KB

Tabular data details:

City

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	759 Bytes

Tabular data details:

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8.58 MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

Before I can do anything, I must first of all ensure that the data is in the right format. This means that I'll have to do data cleaning on all of the csv files.

After downloading the csv files, I needed to convert them into a data frame for me to work with on Pandas, so I used the pd.read_csv method and converted all the files in a data frame.

Once I converted all the files to data frames, I needed to perform a maintenance check to see whether there are any missing data from any of the columns. In this case, no data was missing from the files, so I had to check if there were any duplicate data. From my python code, I confirmed that there were no duplicates any of the four files.

Next, I had to look at what features these data frames had, to get a better idea of what I'm looking at. With that done, I realized I was dealing with a record of customers who had used certain cabs within specific cities. I didn't have the time to convert the recorded data to the mm/dd/yyyy format, but I could tell it had to do with the number of times the cabs has been used throughout the year.

What was important was for me to gather what data type each of these features had. That way, before merging it with the other data frames, I could be convinced which values were numeric and which were categorical.

Once I was sure that the shape of each data frame remained the same, and it was properly filtered , I needed to merge the data. Once I had successfully merged the data, I needed to use visual plots to make sense certain occurrences that were taking place.

First, it seems that New York City is where most cabs are being used frequently. This could be because most of the foreigners have to use taxis instead of Uber when they're coming to New York for the first time. Due to misunderstanding the instructions, I couldn't test this hypothesis on why this was so.

Second, from the visual plots, it seems that people use the Yellow Cab a lot more than the Pink Cab.

Third, from the ages of the users, most of the passengers were from a range of 18 to 65 years old, with the age 20 and 23 as the highest users of about 12000 customers. For some reason, which I don't have the time to describe, people from 41 years and above don't use cabs as much as those between the ages of 18 to 40.

And my final observation was that there were more male customers using the taxis than females. Why this was the case, I couldn't tell.