Data Intake Report

Name: Canmert Demir - G2M insight for Cab Invesment firm
Report date: 13.10.2023
Internship Batch: 30 Sept to 30 Dec 2023
Version: 3.11.4.final.0
Data intake by: Data Glacier
Data intake reviewer: Canmert Demir
Data storage location: https://github.com/Canmertdemir/Canmertdemir-DataGlacierWeek2

Tabular data details:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	Cab_Data.csv
Size of the data	19.2 MB

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	Customer_ID.csv
Size of the data	MB

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	Transaction_ID.csv
Size of the data	10.1 MB

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	City.csv
Size of the data	0,000583648681640625 MB

Total number of observations	359392
Total number of files	1
Total number of features	14
Base format of the file	Big_data.csv
Size of the data	41.1 MB

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

Proposed Approach:

The 'Cab,' 'Transaction_ID,' 'City,' and 'Customer_ID' datasets are complete, with no missing values. These data frames have been merged to create a big data set which is called big_data.csv. It is named as df in Pyhton codes in this project . In this big data, there are structural variables that need to be expressed numerically, but they are currently categorical in the data frame. In the study, these variables were converted to numerical variables using the necessary Python methods. It is assumed that these data sets accurately reflect the real world. It is also assumed that the data is current and reliable and accurately represents real-world events.

It is not assumed that the data is suitable and relevant for analysis or processing, and it is not assumed that unnecessary or irrelevant data has been pre-filtered. This aspect was addressed after variable analysis. In other words, these data sets have been made suitable and functional for serving a specific purpose, and they have been formatted to contain no unnecessary data for processing or analysis. Key fields relevant to the business problem were identified and selected accurately, defined according to their intended use.

First, the descriptive statistics of the obtained big data were examined. In the next stage, variable analyses were determined using Python functions based on the results obtained from descriptive statistics. There are no NA values for any variable in the data set. After that, the distributions were examined as the variables were categorized. Statistically insignificant variables were removed from the data set.

In the analysis stage, a Python function for profit calculation was first written, and the results were examined on a per-taxi company basis. Based on the results, by using statmodel.api library statistical model was developed to predict the connection between profit and distance traveled. This statistical model was visualized using the Seaborn and Matplotlib libraries.

In the final stage, to generate solutions for the business problem, five different hypotheses were formulated and analyzed in detail using analytical methods and statistical tools. The results obtained from the study are provided in detail below."

Taxi Usage: There is no relation between Customer’s income and price charged. This also shows that income and taxi usage are not proportional. At the other side it means that income and km travelled are not in relation.

Profit: Yellow cab company has more profit than Pink cab company. İt is easy to see that XYZ firm should make investment to Yellow cab Company.

Attention: Another Features Analysis made in Hypothesis test. That is why this document is not include those results. They are in Pyhton Commends**.** Also hypothesis explanation and result will be given in Week 3 representation with details.