Data Intake Report

Name: G2M Cab Investment Project
Report date: 20-Jun-2021
Internship Batch: LISUM01
Version: 1.0
Data intake by: Muadh Faizan
Data intake reviewer: Muadh Faizan
Data storage location: https://github.com/MuadhF/DG-G2M-Investment/tree/main/DataSets

Tabular data details: Weather

Total number of observations	47759
Total number of files	4
Total number of features	16
Base format of the file	.csv file
Size of the data	7 MB

- The initial dataset was taken from https://www.kaggle.com/sobhanmoosavi/us-weather-events which was a large dataset which was almost 700MB with all cities in US with a total of around 6 million rows. I had to separately filter out the cities and keep only the data from 2016 to 2018 so the dataset was reduced to around 47k rows.
- Each day had multiple weather records, so I ended up only selecting the first weather record for each day. This is not ideal but due to time restrictions I had to resort to this basic approach. Eg. if a day was reported to be rainy in the morning and snowy in the afternoon, only rainy was assigned to the day.
- There were many days with no weather details, so I assumed that those days had normal weather conditions (not rainy or foggy etc.).

Tabular data details: Cab_Data

Total number of observations	359392
Total number of files	4
Total number of features	7
Base format of the file	.csv file
Size of the data	21.7 MB

Proposed Approach:

- I assumed that the reason there were separate columns for charged and cost, is that the company applied some taxes, mark ups etc to the cost. So I assumed that the difference in the variables was the profit per transaction.

Tabular data details: Customer_ID

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

- I was not able to do any deduplication procedures unfortunately, as I had not heard or learnt of this until it was too late. Therefore I will be sure to do it in the next projects and assignments.

Tabular data details: Transaction_ID

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	.csv file
Size of the data	8.7 MB