Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 30th July 2021
Internship Batch: LISUM02

Version: 1.0

Data intake by: Joseph Antony
Data intake reviewer: N/A

Data storage location:
https://github.com/joeanton719/my_DS_projects/tree/main/5.%20G2M%20insight%20for%20Cab%20Investment%20firm

1. **Tabular data details:** full_df.csv (Master Dataset)

Total number of observations	359392 rows
Total number of files	4
Total number of features	23 columns
Base format of the file	CSV
Size of the data	53.58 MB

2. **Tabular data details:** ny_weather.csv

Total number of observations	1096 rows
Total number of files	1
Total number of features	12 columns
Base format of the file	CSV
Size of the data	68 KB

3. **Tabular data details:** US Holiday Dates (2004-2021).csv

Total number of observations	342 rows
Total number of files	1
Total number of features	6 columns
Base format of the file	CSV
Size of the data	16 KB

4. **Tabular data details:** extreme_weather.csv

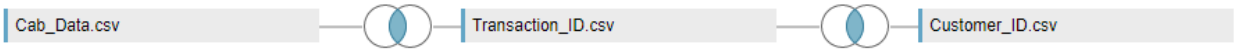
Total number of observations	175721 rows
Total number of files	1
Total number of features	7 columns
Base format of the file	CSV
Size of the data	9.208 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

  a. For the Master Dataset

  First, I created a master dataset by merging the three datasets related to the project (Cab_Data.csv, Customer_ID.csv, Transaction_ID.csv). City.csv dataset was left out as this dataset did not have information regarding the trips. Below is a scheme of how the datasets were merged using inner join.



  Cab_Data dataset was inner joined with Transaction_ID dataset using the 'Transaction_ID' column. Similarly, Transaction_ID dataset was inner joined with Customer_ID dataset using the 'Customer_ID' column.

  Once merged, I extracted both city and state from the 'City' column. Then, I converted the 'Date of Travel' column into a pandas datetime format. The missing values were only found to be in the state column for the cities 'Orange County' and 'Silicon Valley'. As both cities lay in the state of California, the dataset was imputed with 'CA' using pandas 'fillna' function.

  Finally, checking for any duplicated observations returned nothing. After this, the dataset was ready for EDA.

  b. New York Weather Dataset

  The weather datasets for New York were sourced from https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day.

  I was only able to source datasets for New York for all three years.

  The datasets for the year 2016, 2017 and 2018 were on different CSV files. I downloaded the three files and appended the three datasets together row-wise onto one single excel file. After that, I added an extra column called 'state' and inputted 'NY' for all rows. This would make it easier to inner join with the mater dataset using both 'state' and 'date' column to observations for only New York.

After that, I removed columns that was not relevant to the project, keeping only the climate variables such as air temperature, precipitation, etc. When inner joined with the master dataset, no missing values to duplicated rows were found.

c. US Holiday Dates (2004-2021)

US public holiday dataset was sourced from https://data.world/sudipta/us-federal-holidays-2011-2020. The dataset was inner joined with the master dataset using the 'date' column. The joined dataset was then then filtered for observations related to the years from 2016 to 2018.

d. Extreme Weather Dataset

Dataset related to extreme weather conditions for different states of US was sourced from https://www.ncdc.noaa.gov/stormevents/ftp.jsp.

The datasets for the year 2016, 2017 and 2018 were separate. Like done earlier, I appended all three datasets row-wise onto a single excel file. And then I created a new column for state with the state abbreviations to make it easier to 'left join' with the master dataset. I used left join to preserve the trips on the master dataset that didn't have any extreme weather events.

Once joined, the dataset had 200K + missing values on the event column (event column records the type of weather event: Hail, storm, floods, etc.) The empty values were due to being joined with the dates on the master dataset where no event took place (no common dates with the extreme weather dataset). These missing values with the event column were imputed with 'No Event'. The 'MAGNITUDE' column, which records magnitude of certain events were imputed with 0.

It is important to mention that MAGNITUDE had only values for certain weather events, mainly 'Hail', 'Strong Winds', etc. Therefore, those observations other than these weather events were removed prior to joining with the master dataset.

- Mention your assumptions (if you assume any other thing for data quality analysis)

a. 'Cost of trip' variable only involves fuel costs.
b. 'price_charged' only involved the cost of the trip per km, not taking into account the base fares, waiting charged if any, etc.
c. Profit = 0 is assumed to be a loss too.
d. 5th January 2018 is recorded have the highest number of trips for both companies, especially in the state of New York. I haven't succeeded in establishing a cause for such a spike. I assume that this could something to do with the data acquisition process.
e. Finally, the weekly seasonality of daily trips is very different for all three years (2016, 2017, 2018). I assume again these changes does not reflect on the reality and might be introduced into the data during the data acquisition process. I haven't been able to assign any causality to these changes when analyzing the data.