Data Intake Report

Name: XYZ Cab Investment EDA
Report date: 2021-03-17
Internship Batch: LISP01
Version: 1.0
Data intake by: Juan Carlos Gutiérrez
Data intake reviewer:
Data storage location: https://github.com/jaycee-ds/XYZ-investment

Tabular data details:

Total number of observations	359,392
Total number of files	1 (CabData)
Total number of features	7
Base format of the file	.csv
Size of the data	19.2 MB
Total number of observations	20
Total number of files	1 (cityData)
Total number of features	3
Base format of the file	.csv
Size of the data	608 bytes
Total number of observations	49,171
Total number of files	1 (customersData)
Total number of features	4
Base format of the file	.csv
Size of the data	1.5 MB
Total number of observations	440,098
Total number of files	1 (transactionsData)
Total number of features	3
Base format of the file	.csv
Size of the data	10.1 MB
Total number of observations	36
Total number of files	1 (holidays)
Total number of features	2
Base format of the file	Pandas DataFrame
Size of the data	704 bytes
Total number of observations	20
Total number of files	1 (overseasTourists)
Total number of features	3
Base format of the file	Pandas DataFrame
Size of the data	640 bytes
Total number of observations	20
Total number of files	1 (unemploymentRate)

Total number of features	3
Base format of the file	Pandas DataFrame
Size of the data	640 bytes
Total number of observations	20
Total number of files	1 (averageTemperature)
Total number of features	12
Base format of the file	Pandas DataFrame
Size of the data	2 KB
Total number of observations	359,392
Total number of files	1 (df, final dataframe)
Total number of features	19
Base format of the file	Pandas DataFrame
Size of the data	52.1 MB

Our focus on this dataset was on transactions, even if the same customer was on rides over and over again (and there are many). Every single transaction may be different (and actually they are) regarding km travelled, cost, price charged and profit made. That is why duplicated customers were not a problem to deal with. Obviously, possible duplicated transactions or duplicated customers in their own tables (duplicated transactions ID in transactionsData, or duplicated customers ID in customersData) might have been a problem, but we did not find any duplicate.

Main processing tasks were the following:

- Convert integer dates into correct formatted dates (spanning from 2016-01-02 to 2018-12-31).
- Created Profit column computed from price charged and cost of trip.
- Updated populations since there seemed to be some imbalances according to census data.

The final data frame (used for EDA) is comprised of (data from external sources are in blue):

- Cab data
- City data
- Customers' data
- Transactions' data
- US federal holidays
- Unemployment rate
- Average temperature in the city
- Overseas visitors