Data Intake Report

Name: cab case study
Report date: 20-06-2021
Internship Batch:
Version:1.0
Data intake by: Almudena Zhou Ramírez López
Data intake reviewer:

Cab dataset

Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

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	20.1 MB

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

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	1.00 MB

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

Weather dataset

Data storage location:
https://kilthub.cmu.edu/articles/dataset/Compiled_daily_temperature_and_precipitation_data_for_the_U_S_cities/7890488

Tabular data details:

Total number of observations	461
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	43.2 KB

All city files:

Total number of observations	46386-55152
Total number of files	210
Total number of features	4
Base format of the file	.csv
Size of the data	1.17-1.57 MB

Holidays dataset

Data storage location: https://data.world/sudipta/us-federal-holidays-2011-2020

Tabular data details:

Total number of observations	100
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	2.77 KB

Proposed Approach:

- Mention approach of dedup validation (identification): In the city weather dataset, there are 461 observations and only 210 cities so the csv ID which were duplicates were removed, and only were kept the name of the city. I have considered joined the cab dataset between them and deleted the columns which were shared by them.
- Mention your assumptions (if you assume any other thing for data quality analysis): I have not taken in consideration the file of the cities (population, users) since I have not found it could give extra value to the purpose we had.