CS 6140 Spring 2020: Data Collection Report

Analyzing Chicago Taxi Trips Behaviour via Data Mining

Shaurya Sahai, Abishek Krishnan and Sushmitha Sunkurdi Nataraj

1. How you obtained your data?

Our data was collected from Kaggle [https://www.kaggle.com/chicago/chicago-taxi-trips-bq] which in turn is hosted by Google BigQuery under the table [bigquery-public-data:chicago_taxi_trips.taxi_trips].

This is being collected by the City of Chicago (the regulatory agency) from 2013 to the present. The agency does so through periodic reporting by two major payment processors believed to cover most taxis in Chicago. Currently, it has information about over 100 million Chicago taxi rides.

2. How large is your data?

At present, the table has over 192,029,239 rows (Table size: ~69.6 GB).
The schema of the table is present in Appendix A.

3. In what format are you storing your data? Describe the abstract data type, not just the file format.
We are extracting each row of the dataset into multiple component form, as a key-value pair of attribute and its corresponding value into a python dictionary, we propose this data type as it would lead to effective mining of the dataset.

4. Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?
As the original data is being hosted by Google BigQuery, we didn't have to compress the format as such, but we customize our algorithms in such a way only the rows and columns which are relevant to answer the particular question at hand is being returned from the queries. This way we would be able to handle our data without compromising the quality.

5. How would you simulate similar data?
We would take the required fields from the schema and then store the data as a structure (tuple or dictionary) in our program. Further, for simulation, we can generate data from the assumed structure using Python's random number generation function alongside DateTime, Latitude, Longitude which covers the Chicago area, price and a random string from a set of strings stored for types ex. the company, Payment Type, Pickup location. Modeling the data this way would lead to convenient processing and test if our algorithms perform the mining reliably.

Appendix A
Schema of the table [bigquery-public-data:chicago_taxi_trips.taxi_trips]

unique_key	STRING	REQUIRED	Unique identifier for the trip.
taxi_id	STRING	REQUIRED	A unique identifier for the taxi.
trip_start_timestamp	TIMESTAMP	NULLABLE	When the trip started, rounded to the nearest 15 minutes.
trip_end_timestamp	TIMESTAMP	NULLABLE	When the trip ended, rounded to the nearest 15 minutes.
trip_seconds	INTEGER	NULLABLE	Time of the trip in seconds.
trip_miles	FLOAT	NULLABLE	Distance of the trip in miles.
pickup_census_tract	INTEGER	NULLABLE	The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips.
dropoff_census_tract	INTEGER	NULLABLE	The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips.
pickup_community_area	INTEGER	NULLABLE	The Community Area where the trip began.
dropoff_community_area	INTEGER	NULLABLE	The Community Area where the trip ended.
fare	FLOAT	NULLABLE	The fare for the trip.
tips	FLOAT	NULLABLE	The tip for the trip. Cash tips generally will not be recorded.
tolls	FLOAT	NULLABLE	The tolls for the trip.
extras	FLOAT	NULLABLE	Extra charges for the trip.

trip_total	FLOAT	NULLABLE	Total cost of the trip, the total of the fare, tips, tolls, and extras.
payment_type	STRING	NULLABLE	Type of payment for the trip.
company	STRING	NULLABLE	The taxi company.
pickup_latitude	FLOAT	NULLABLE	The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
pickup_longitude	FLOAT	NULLABLE	The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
pickup_location	STRING	NULLABLE	The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
dropoff_latitude	FLOAT	NULLABLE	The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.
dropoff_longitude	FLOAT	NULLABLE	The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.
dropoff_location	STRING	NULLABLE	The location of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.