Data Wrangling Report

Dataset:

The dataset is taken from Kaggle. It includes taxi trips for 2016, reported to the City of Chicago in its role as a regulatory agency.

It has more than 10 million observations for each month in the year 2016.

The data includes the following fields:

1. taxi_id – ID assigned to each taxi
2. trip_start_timestamp – date and time when the trip started
3. trip_end_timestamp – date and time when the trip ended
4. trip_seconds – total seconds taken to complete the trip
5. trip_miles – total miles travelled
6. pickup_census_tract – neighbourhood from where the customer was picked up
7. dropoff_census_tract – neighbourhood where the customer was dropped off
8. pickup_community_area – area of pickup
9. dropoff_community_area – area of drop off
10. fare – charge of the taci ride
11. tips – tips given to thr driver
12. tolls – amount paid for tolls
13. extras – extra amount included
14. trip_total – total trip amount
15. payment_type – type of payment made
16. company – company to which the taxi belongs
17. pickup_latitude – latitude from where customer was picked up
18. pickup_longitude – longitude from where customer was picked up
19. dropoff_latitude – latitude where customer was dropped off
20. dropoff_longitude – longitude where customer was dropped off

Data Wrangling:

The dataset was comprehensive with few missing values. It required some cleanup and reformatting. The steps taken are described below.

Columns which were not needed for the analysis were removed for example pickup census tract as this column did not have any values included, due to security purposes.

The dataset imported to Python was stored in the dataframe where NAN values were checked and removed. These transformations were helpful to conduct preliminary exploration and data visualization.

In the taxi trip data, few observations had 0 pickup_latitude, 0 pickup_longitude, 0 dropoff_latitude, 0 dropoff_longitude. Hence, these records were removed as it was of no use having these observations.

After considering all these factors from the data and cleaning up the data, now the data is ready for further analysis.