Uber NYC For-Hire Vehicles Trip Data (2021)

By: Amy W, Jessie E, Alan K, Brayan G, Rodolfo H
Group 4
CIS 4560-01
Github link:
https://github.com/JessieEstrada/NYC_Uber_Trip_Data_2021_Project

Data Background

- The dataset consists of information regarding the different types of taxis in New York City in terms of time, location, mileage, and pay among different ridesharing companies.
    - In this case, the ridesharing companies that are mentioned are Uber, Juno, Via, and Lyft.
- The for-hire vehicle data includes the trip data from high-volume for-hire vehicle bases, community livery bases, luxury limousine bases, and black car bases.
- Uber's trip records are collected as well since it's one of the biggest ride-sharing service providers.

Why is Our Analysis Important?

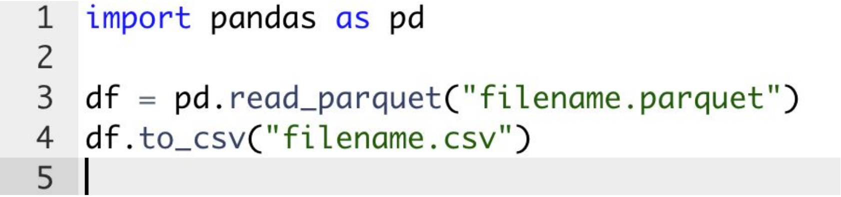
- We know that NYC is a very populated area and there is always a ton of traffic.
- By analyzing the underlying trip patterns in 2021, we can find ways to improve Uber’s ride-sharing services.
- Could help the companies know how to better place pickup locations in NYC.
- Could also help the users make decisions on which ride-sharing company they should rely on.

H/W Experimental Specifications (Data Size, Your cluster version, cluster number of nodes, memory size, CPU speed):

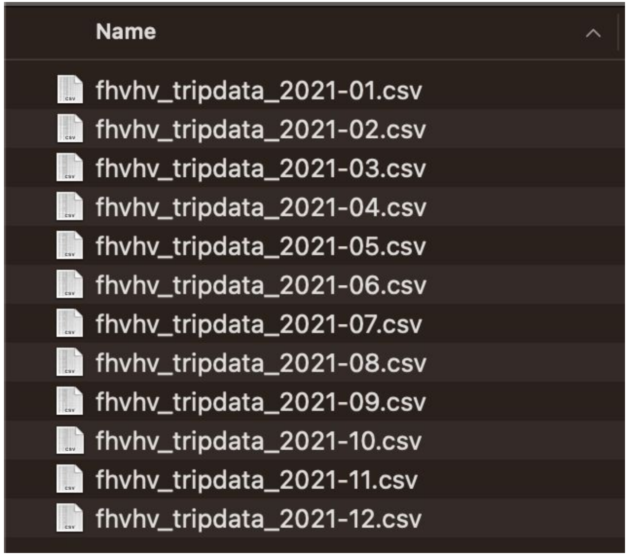
- **Data size:** 5GB
- **Cluster version:** Hadoop 3.1.2
- **Number of nodes:** 5
- **Memory Size:** 400 Gigs
- **CPU Speed:** 2.00GHz
- **Number CPU Cores:** 8

Data Processing & Prep

- Download & Upload Data from source
- Process Data with Python to CSV
- Upload csv data to Dropbox

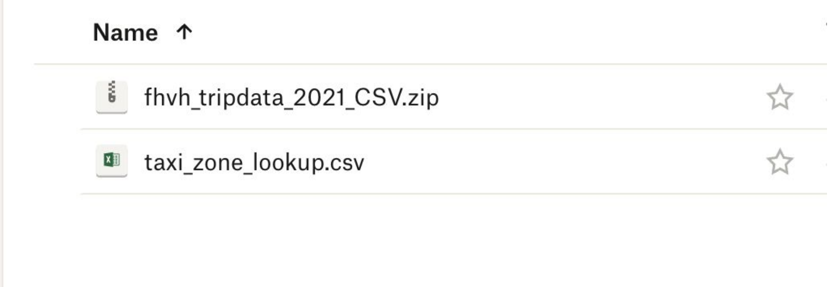


```python
import pandas as pd

df = pd.read_parquet("filename.parquet")
df.to_csv("filename.csv")

```



Name

fhvhv_tripdata_2021-01.csv
fhvhv_tripdata_2021-02.csv
fhvhv_tripdata_2021-03.csv
fhvhv_tripdata_2021-04.csv
fhvhv_tripdata_2021-05.csv
fhvhv_tripdata_2021-06.csv
fhvhv_tripdata_2021-07.csv
fhvhv_tripdata_2021-08.csv
fhvhv_tripdata_2021-09.csv
fhvhv_tripdata_2021-10.csv
fhvhv_tripdata_2021-11.csv
fhvhv_tripdata_2021-12.csv

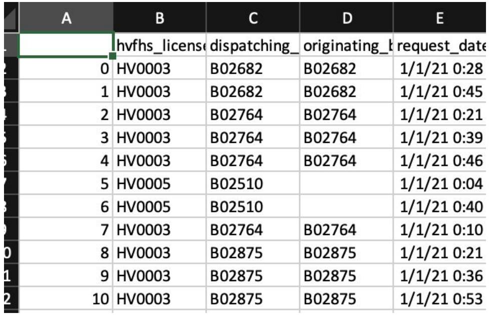


Name ↑

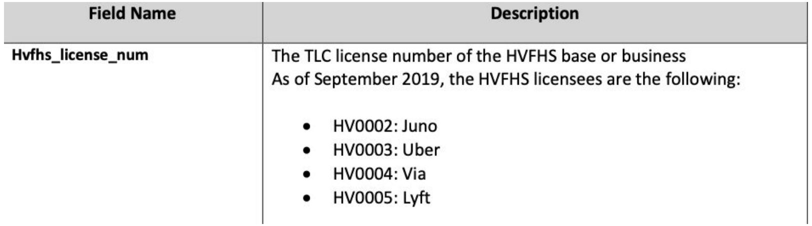
fhvh_tripdata_2021_CSV.zip ☆
taxi_zone_lookup.csv ☆

Data Cleaning

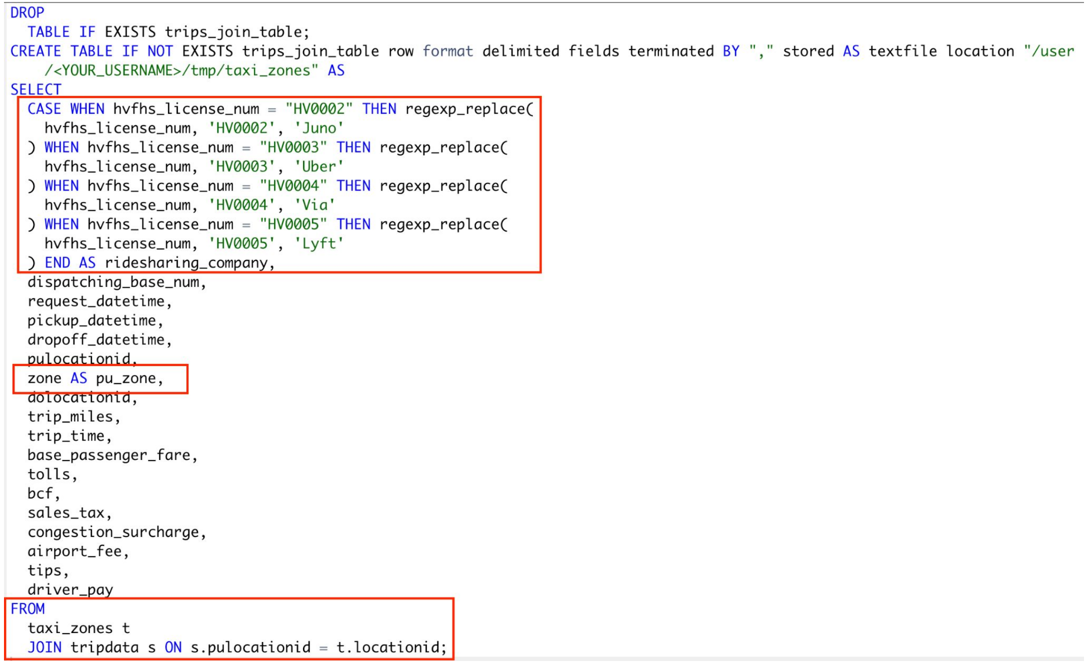
- Replace “hvfhs_license” to actually ridesharing company
- Join Data Table with Taxi Zones



A	B	C	D	E
	hvfhs_licens	dispatching_	originating_	request_date
0	HV0003	B02682	B02682	1/1/21 0:28
1	HV0003	B02682	B02682	1/1/21 0:45
2	HV0003	B02764	B02764	1/1/21 0:21
3	HV0003	B02764	B02764	1/1/21 0:39
4	HV0003	B02764	B02764	1/1/21 0:46
5	HV0005	B02510		1/1/21 0:04
6	HV0005	B02510		1/1/21 0:40
7	HV0003	B02764	B02764	1/1/21 0:10
8	HV0003	B02875	B02875	1/1/21 0:21
9	HV0003	B02875	B02875	1/1/21 0:36
10	HV0003	B02875	B02875	1/1/21 0:53

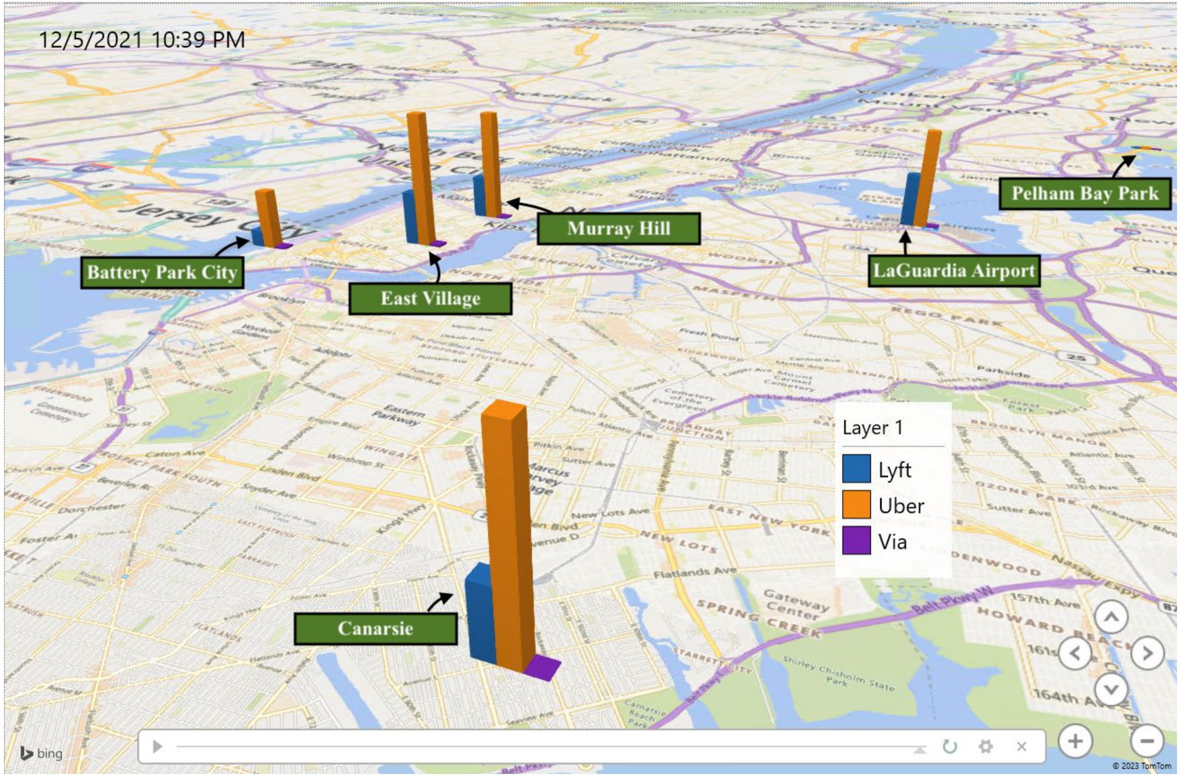


Field Name	Description
Hvfhs_license_num	The TLC license number of the HVFHS base or business As of September 2019, the HVFHS licensees are the following: • HV0002: Juno • HV0003: Uber • HV0004: Via • HV0005: Lyft



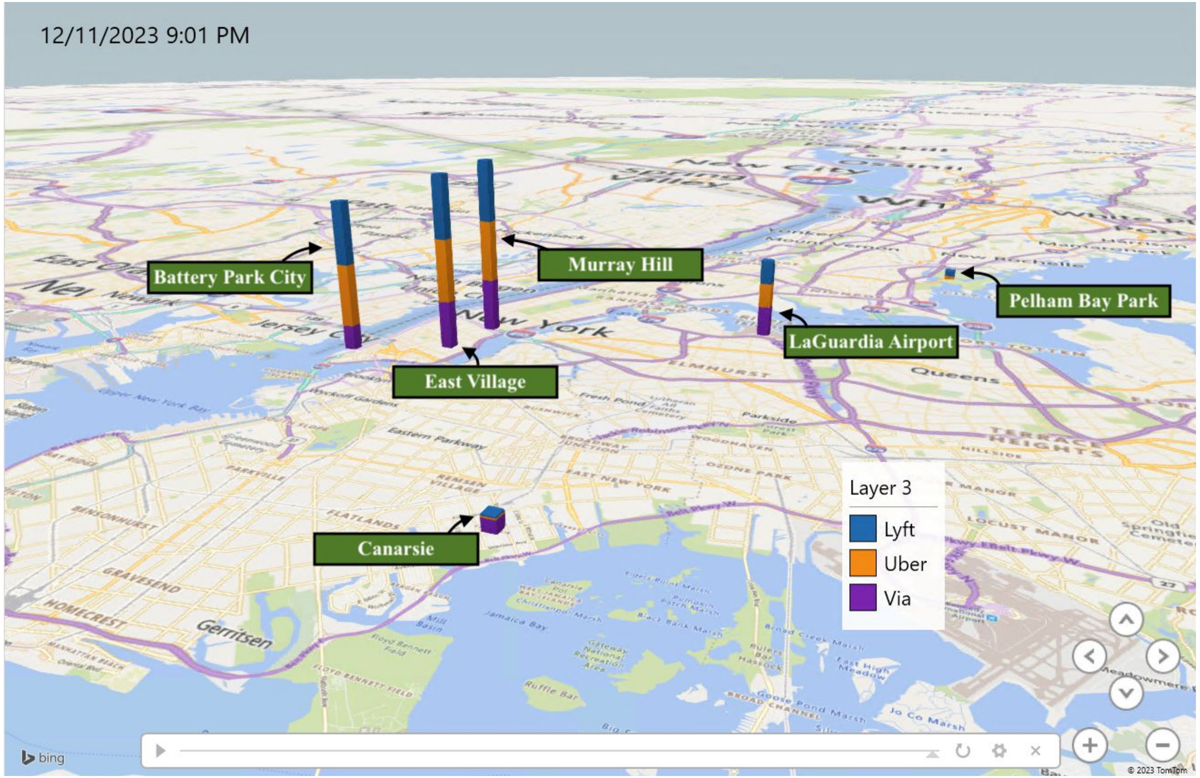
```
DROP
  TABLE IF EXISTS trips_join_table;
CREATE TABLE IF NOT EXISTS trips_join_table row format delimited fields terminated BY "," stored AS textfile location "/user
    /<YOUR_USERNAME>/tmp/taxi_zones" AS
SELECT
  CASE WHEN hvfhs_license_num = "HV0002" THEN regexp_replace(
    hvfhs_license_num, 'HV0002', 'Juno'
  ) WHEN hvfhs_license_num = "HV0003" THEN regexp_replace(
    hvfhs_license_num, 'HV0003', 'Uber'
  ) WHEN hvfhs_license_num = "HV0004" THEN regexp_replace(
    hvfhs_license_num, 'HV0004', 'Via'
  ) WHEN hvfhs_license_num = "HV0005" THEN regexp_replace(
    hvfhs_license_num, 'HV0005', 'Lyft'
  ) END AS ridesharing_company,
  dispatching_base_num,
  request_datetime,
  pickup_datetime,
  dropoff_datetime,
  pulocationid,
  zone AS pu_zone,
  dolocationid,
  trip_miles,
  trip_time,
  base_passenger_fare,
  tolls,
  bcf,
  sales_tax,
  congestion_surcharge,
  airport_fee,
  tips,
  driver_pay
FROM
  taxi_zones t
  JOIN tripdata s ON s.pulocationid = t.locationid;
```

Visualizations (Part 1)



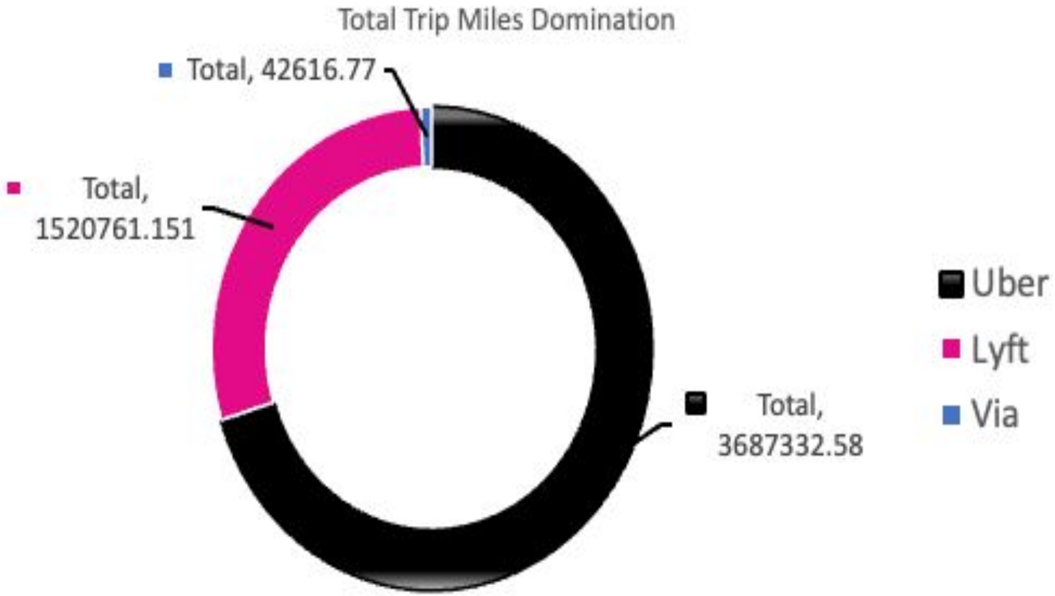
The Visualization here shows a couple of different pick up zones for Uber, Lyft, and Via. As described by the size of each Uber bar, this company is the most popular with the highest amount requests. Lyft comes in at second and Via is the least popular for ride requests. Throughout the days, ride requests increase in the mornings and evenings.

Visualizations (Part 2)



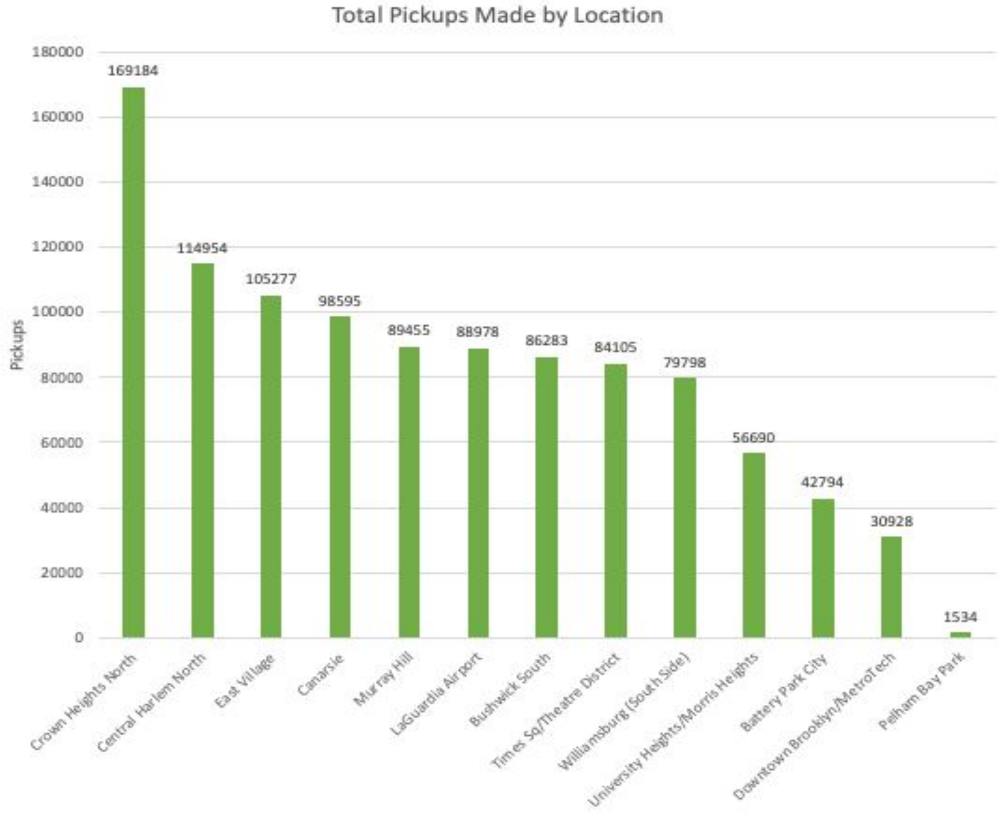
The visualization above shows the average congestion surcharge for each pickup zone, as well as the companies. As seen above, Battery Park City, East, Village and Murray Hill have the highest average congestion surcharge. Throughout the day the surcharge stays fairly stable with Laguardia Airport, Canarsie, and Pelham Bay Park increasing by a minimum.

Visualizations (Part 3)



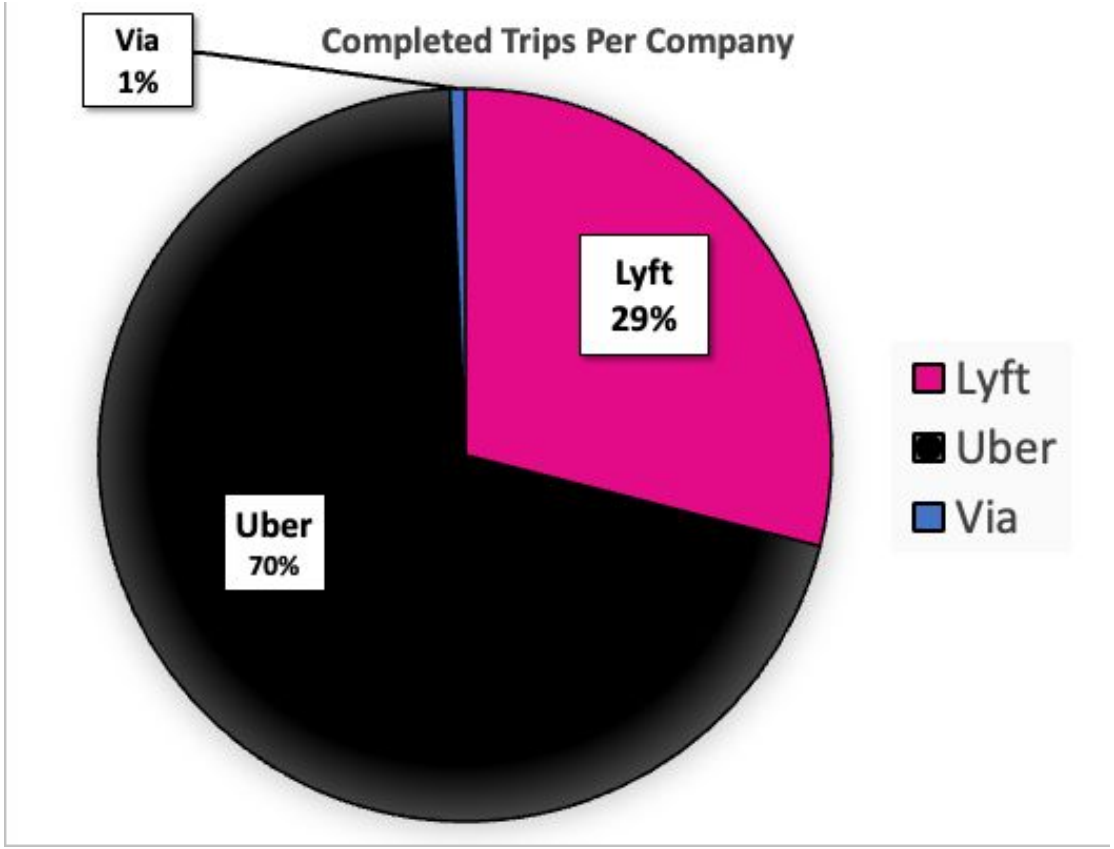
This Visualization showcases the Trip Miles domination between these companies. As we can see, Uber dominates in total trip miles with 3687332.58 of our data being miles from them. Lyft comes in second with 1520761.151and Via with only 42616.77. This demonstrates how Uber dominates these areas and has the most miles traveled with passengers.

Visualizations (Part 4)



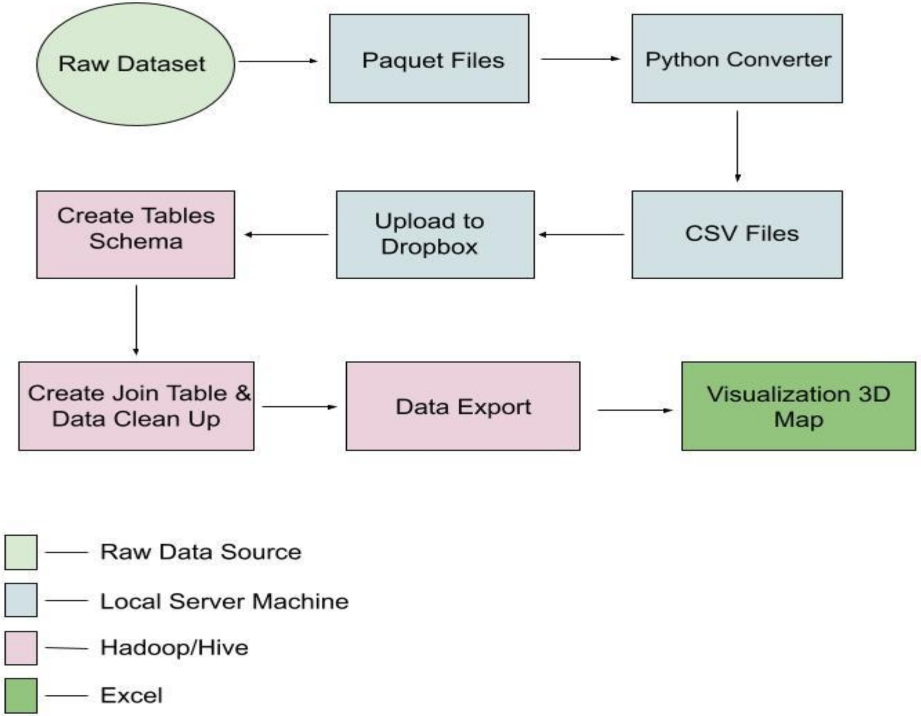
This visualization shows all the different pickup locations and the total amount of pickups done at these locations. As we can see here, in this portion of our data set, Crown Heights North had a total of 169,184 pickups. This supports the idea that this place may be more populated and it is much more visited in comparison to the other locations.

Visualizations (Part 5)



This visualization shows our 3 different companies and their percentage of completed trips. We can see that Uber once again dominated with owning 70% of total trips in our data set. This further supports the idea that Uber is the preferred transportation company. Lyft still does have a decent amount of trips completed, unlike Via with only 1% total completed trips contribution.

Implementation (Flow Chart)



Relevant Work (our differences)

Our dataset contains information about trips taken by different types of for-hire vehicles in New York City, including yellow taxis, green taxis, and high-volume for-hire vehicles like Uber.

We focus on how our approach to exploratory data analysis, our use of weather data in building a predictive model, and our exploration of Uber's user portrait are different from related works. This will help emphasize the unique contributions of our work and distinguish it from other studies that may have used similar datasets.

The three main business goals we want to achieve using this dataset are:

- Exploratory data analysis to identify trip patterns and trends in 2021.
- Building a predictive model to forecast peak footfall based on trip data and weather information.
- Exploring Uber's user demographics in NYC to identify which orders are urgent and which users should be given higher priority.

Conclusion

- Uber is the most popular ride service in New York City compared to Lyft and Via
- Uber has a higher surcharge than Via and Lyft
- Uber has the highest driver pay in comparison to Lyft and Via in New York City
- In New York City, Uber has the most trip miles with 70% compared to Lyft 30% and Via 1%
- The most popular pickup location among the three car ride services was Crown Heights North with 169,184 pickups in 2021