UBER Pickups in NYC

By Cristina Iacob

Goal

optimization of Uber pickups in NYC

> *Can we help Uber optimize the pickups in NYC based on pickups geolocations as well as time, day of the week and/or day of the month?*

Intended audience

Uber resource allocation team, TLC, Uber stakeholders

Data

https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city

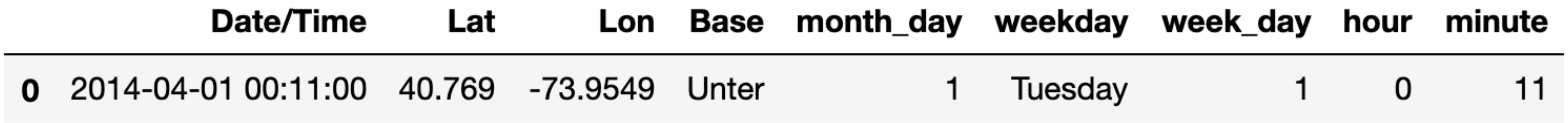
This directory contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015

uber-raw-data-apr14.csv (600, 000 pickups)

sample – 20,000 pickups

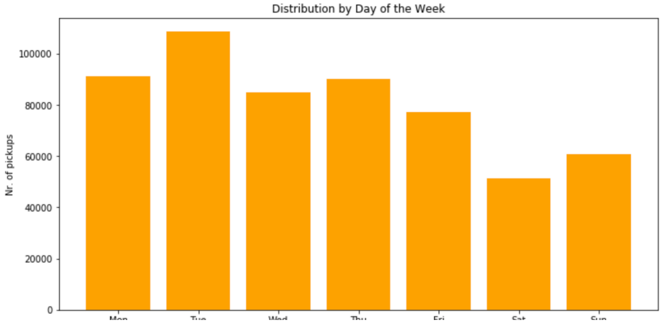
Data Description

- **Date/Time** : The date and time of the Uber pickup
- **Lat** : The latitude of the Uber pickup
- **Lon** : The longitude of the Uber pickup
- **Base** : The TLC base company name/code affiliated with the Uber pickup

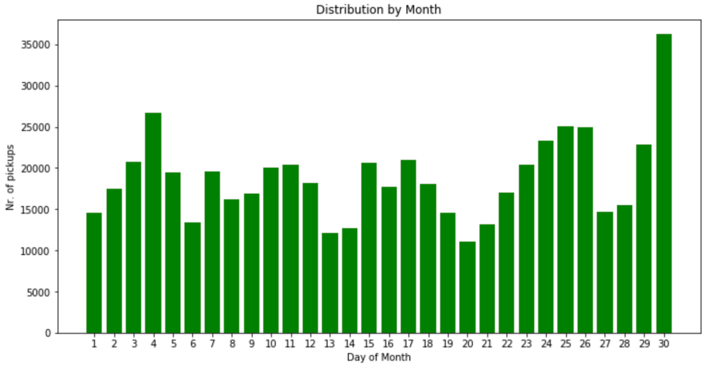


	Date/Time	Lat	Lon	Base	month_day	weekday	week_day	hour	minute
0	2014-04-01 00:11:00	40.769	-73.9549	Unter	1	Tuesday	1	0	11

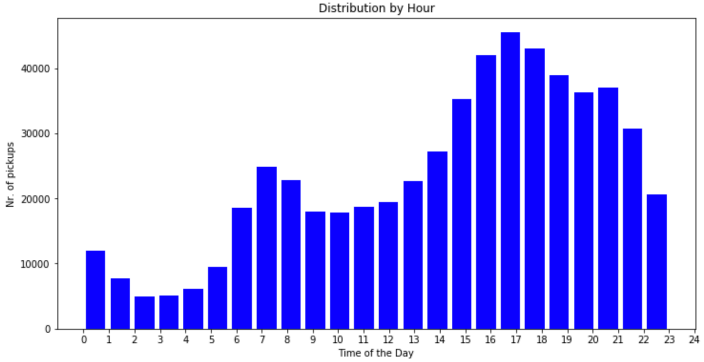
Features Visualization



- During the month of April, the busiest day is Tuesday while on Saturday are the least pickups.

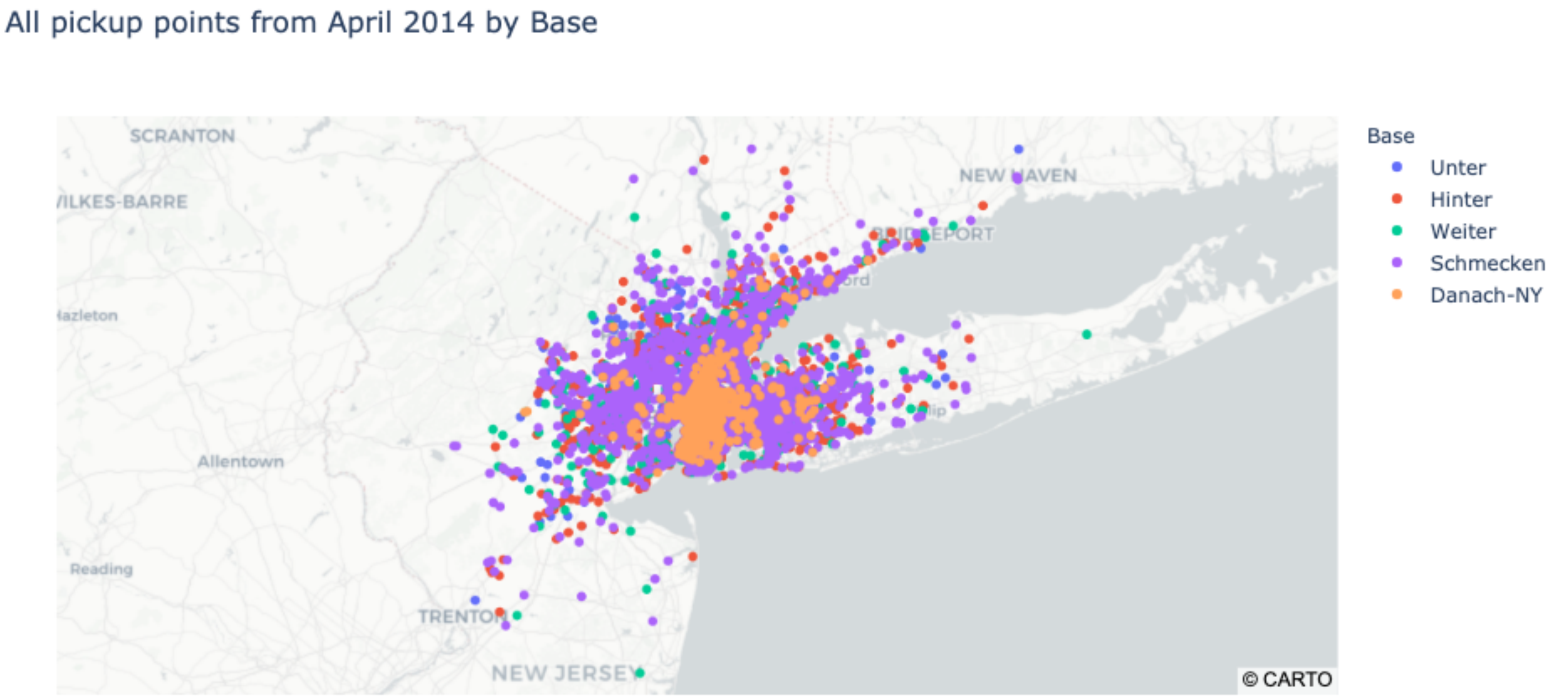


- The most pickups were at the end of April.

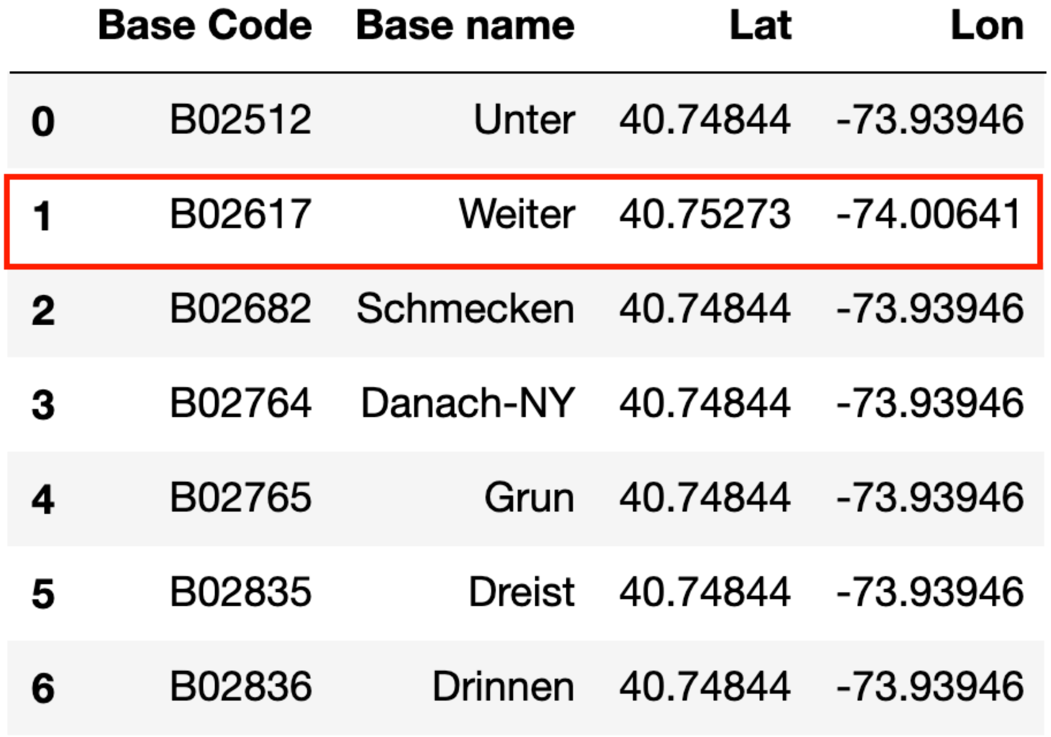


- During the month of April, the most pickups are between 3pm and 10pm.

Pickups Locations by Base



Base Locations



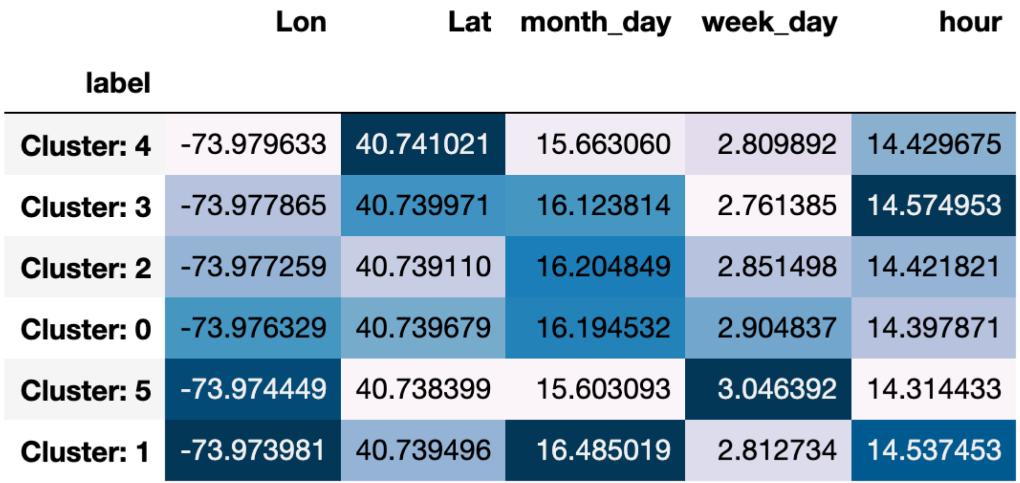
	Base Code	Base name	Lat	Lon
0	B02512	Unter	40.74844	-73.93946
1	B02617	Weiter	40.75273	-74.00641
2	B02682	Schmecken	40.74844	-73.93946
3	B02764	Danach-NY	40.74844	-73.93946
4	B02765	Grun	40.74844	-73.93946
5	B02835	Dreist	40.74844	-73.93946
6	B02836	Drinnen	40.74844	-73.93946

Analysis

K-means clustering is an unsupervised machine learning algorithm that groups data entries into groups, known as clusters, through the calculation of distance between cluster centroids

DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN groups together points that are close to each other based on a distance measurement and a minimum number of points.

Results -- All features --



	Lon	Lat	month_day	week_day	hour
label					
Cluster: 4	-73.979633	40.741021	15.663060	2.809892	14.429675
Cluster: 3	-73.977865	40.739971	16.123814	2.761385	14.574953
Cluster: 2	-73.977259	40.739110	16.204849	2.851498	14.421821
Cluster: 0	-73.976329	40.739679	16.194532	2.904837	14.397871
Cluster: 5	-73.974449	40.738399	15.603093	3.046392	14.314433
Cluster: 1	-73.973981	40.739496	16.485019	2.812734	14.537453

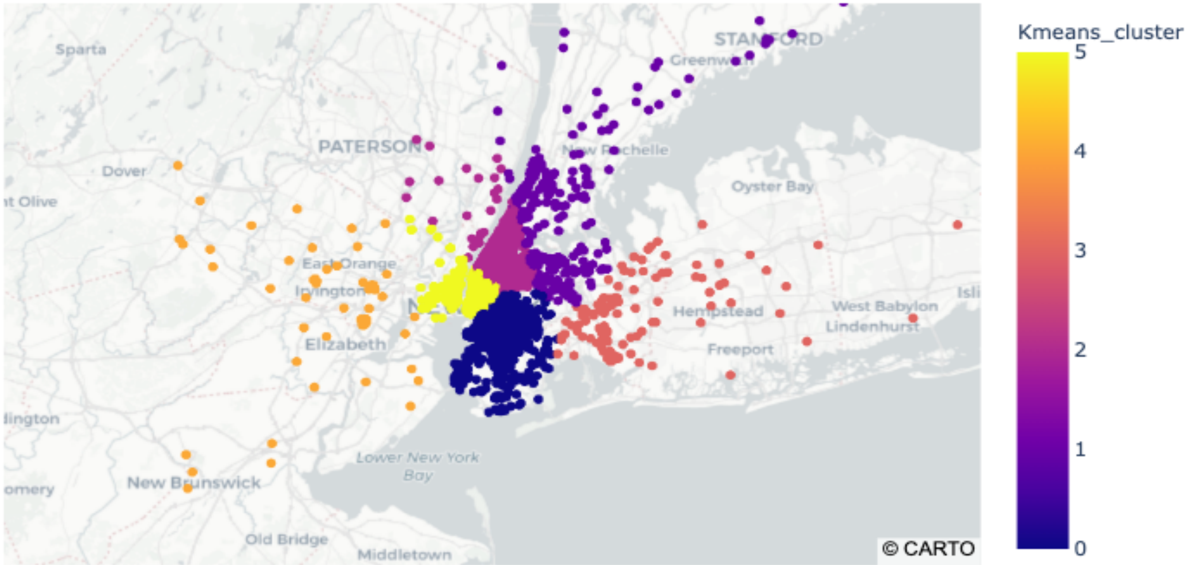
Cluster 3: heavy on pickups during weekeend evenigns, midle of the month

Cluster 0: heavy on pickups at the of the month, light at beginning of week, moderate trafic afternoon

Cluster 4: moderate pickups mid-month, mid-week

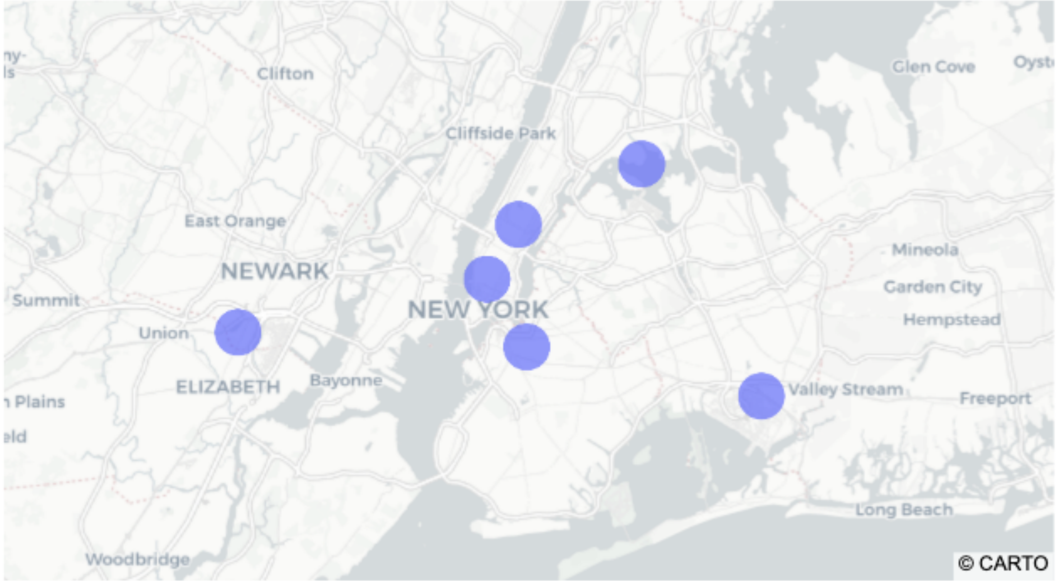
Results -- Geolocation -- KMeans

Clustering sample points from April 2014 with Kmeans k=6



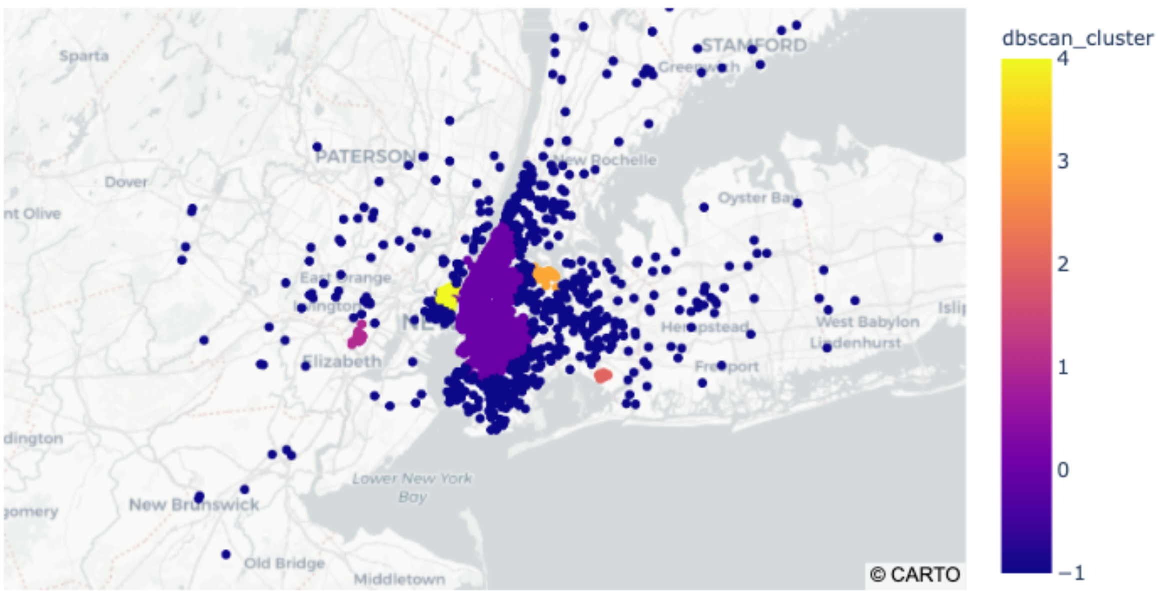
Results -- Centroids – KMeans

Centroids locations

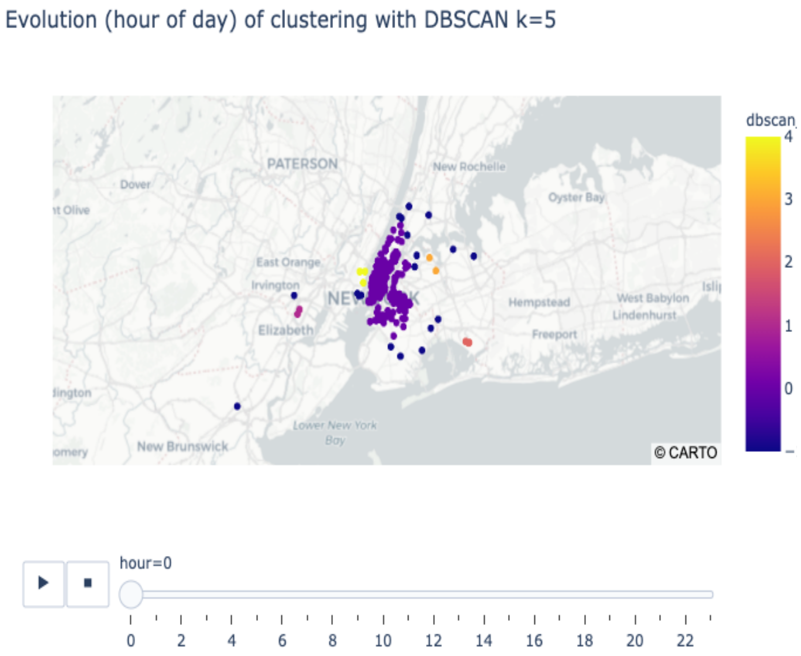


Results -- Geolocation -- DBSCAN

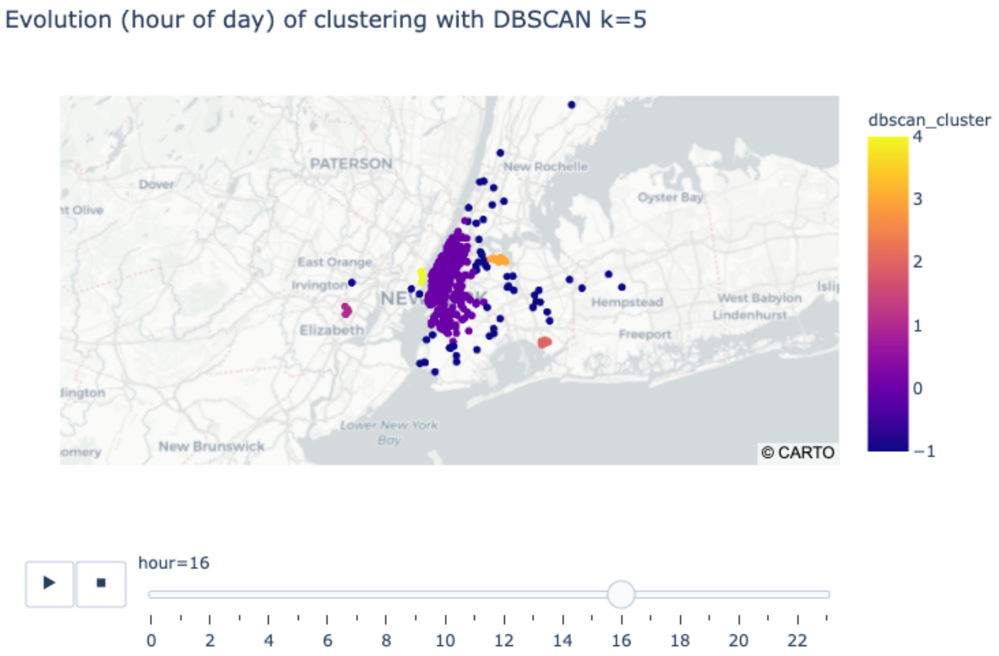
Clustering sample points from April 2014 with DBSCAN k=5



Results (cont'd)



Evolution (hour of day) of clustering with DBSCAN k=5



Evolution (hour of day) of clustering with DBSCAN k=5

Conclusions

Both clustering methods show improved pickups allocation than existing one

DBSCAN is more suitable for geolocation clustering

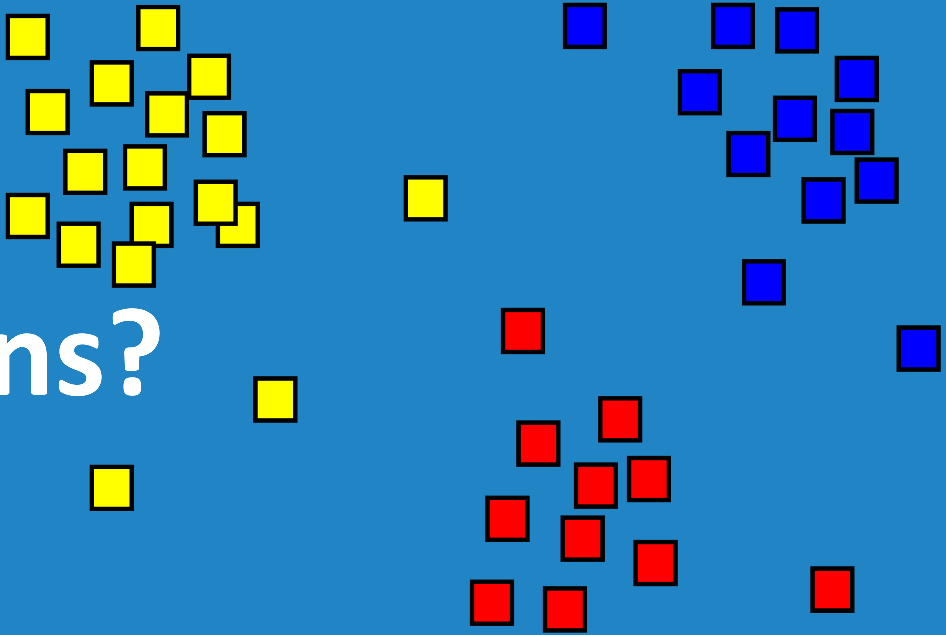
Limitations and Next Step

Limitations: computation power (run at AWS) and time

- better understanding of allocation (TLC)
- fine tunning of DBSCAN parameters
- analyze all data
- more current data
- combine with demographics, hotels, restaurants, and venues as well as weather data

Thanks!

Any questions?



https://github.com/cristina-iacob/uber_pickups_new_york_city

EXTRA RESOURCES

TLC: Taxi and Limousine Commission

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars and luxury limousines), commuter vans, and paratransit vehicles. The Commission's Board consists of nine members, eight of whom are unsalaried Commissioners.

The salaried Chair/ Commissioner presides over regularly scheduled public commission meetings and is the head of the agency, which maintains a staff of approximately 600 TLC employees.

Over 200,000 TLC licensees complete approximately 1,000,000 trips each day. To operate for hire, drivers must first undergo a background check, have a safe driving record, and complete 24 hours of driver training. TLC-licensed vehicles are inspected for safety and emissions at TLC's Woodside Inspection Facility.

TLC Bases Info:

https://data.cityofnewyork.us/Transportation/CURRENT-BASES/eccv-9dzr/data

DBSCAN tunning parameters

- ***eps***: *if the eps value chosen is too small, a large part of the data will not be clustered. It will be considered outliers because don’t satisfy the number of points to create a dense region. On the other hand, if the value that was chosen is too high, clusters will merge and the majority of objects will be in the same cluster. The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable.*
- ***minPoints***: *As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1. Larger values are usually better for data sets with noise and will form more significant clusters. The minimum value for the minPoints must be 3, but the larger the data set, the larger the minPoints value that should be chosen.*