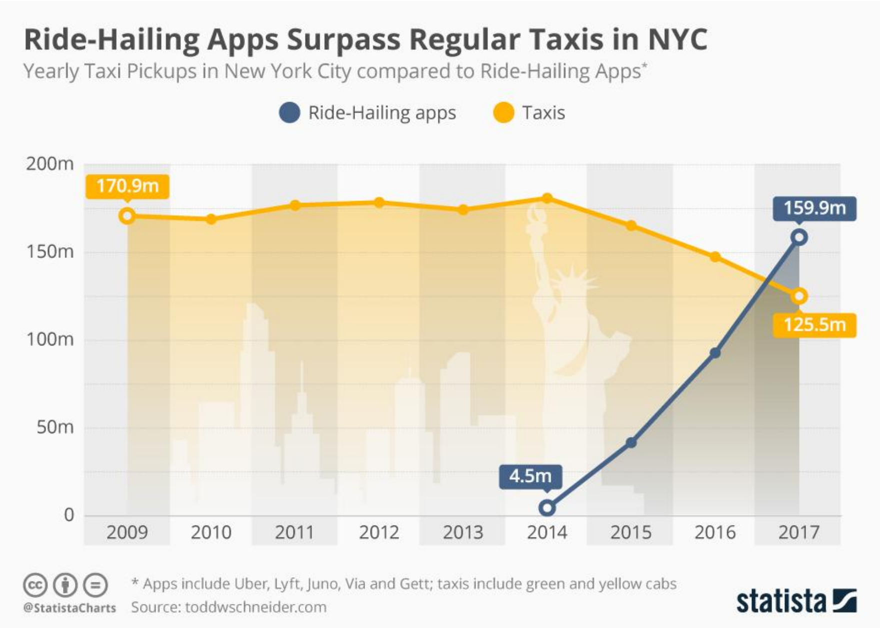
Predicting Taxi Fare Prices in NYC

David Estoque
11/30/2018

NYC Mobility Statistics

- Taxis are losing market share to Uber and Lyft [1]
    - Still important function of NYC Mobility
    - 13,000 Taxicab medallions in NYC
        - 2nd in US- Chicago about 6,000
-



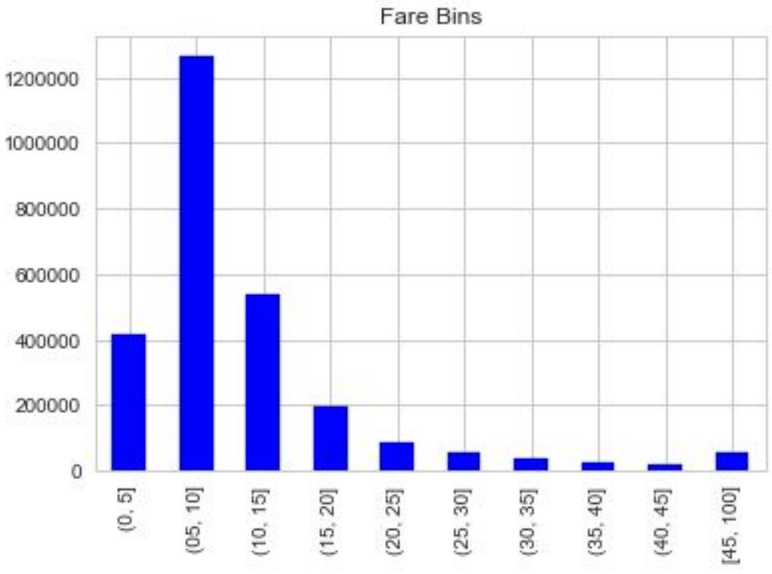
About the Data

	key	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	2015-01-27 13:08:24.0000002	2015-01-27 13:08:24 UTC	-73.973	40.764	-73.981	40.744	1
1	2015-01-27 13:08:24.0000003	2015-01-27 13:08:24 UTC	-73.987	40.719	-73.999	40.739	1
2	2011-10-08 11:53:44.0000002	2011-10-08 11:53:44 UTC	-73.983	40.751	-73.980	40.746	1
3	2012-12-01 21:12:12.0000002	2012-12-01 21:12:12 UTC	-73.981	40.768	-73.990	40.752	1
4	2012-12-01 21:12:12.0000003	2012-12-01 21:12:12 UTC	-73.966	40.790	-73.989	40.744	1

- Obtained using NYC OpenData
- Over 55 million rows of cab rides
  - Shortened to 2.75 million
- Cab ride data obtained from 2009-2015
- Average fare amount was $11.34

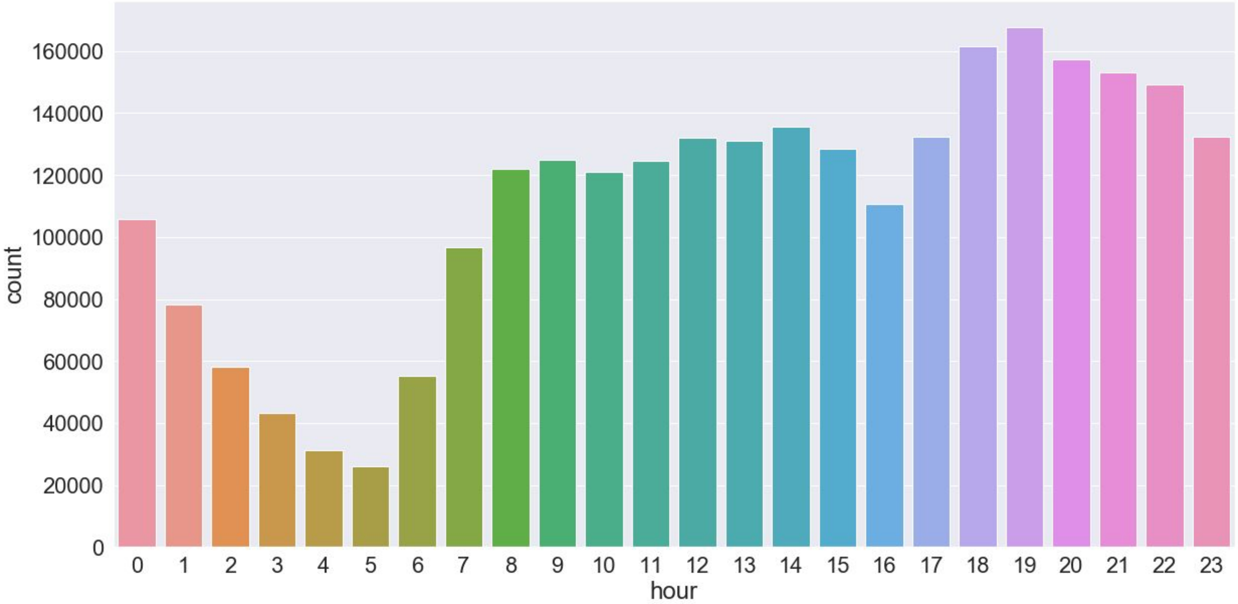
	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	2749978.000	2749978.000	2749978.000	2749978.000	2749978.000	2749978.000
mean	11.340	-72.517	39.926	-72.517	39.921	1.684
std	9.828	13.153	8.513	12.808	10.155	1.325
min	-62.000	-3426.609	-3488.080	-3408.430	-3488.080	0.000
25%	6.000	-73.992	40.735	-73.991	40.734	1.000
50%	8.500	-73.982	40.753	-73.980	40.753	1.000
75%	12.500	-73.967	40.767	-73.964	40.768	2.000
max	1273.310	3439.426	2912.465	3414.307	3345.917	208.000

Fare Distribution

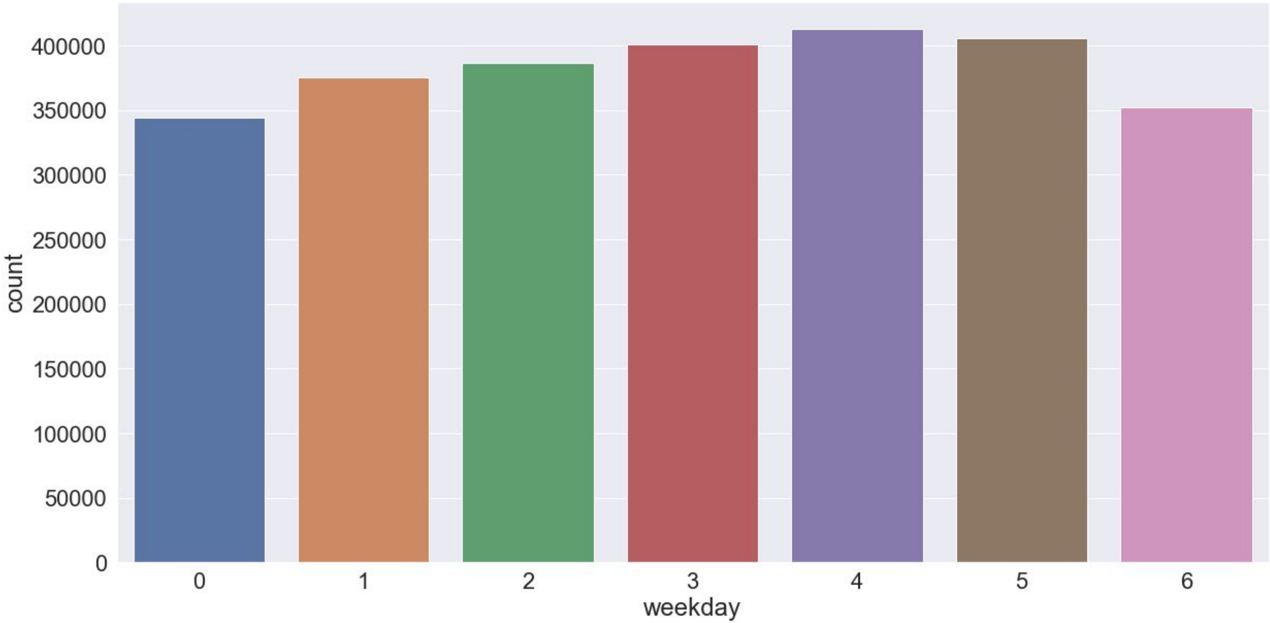


- Preprocessing for fare price
  - Included negative values
  - Max = $1,273
  - Set dataset max to $100
- Rides to airport
  - Base fare of $45

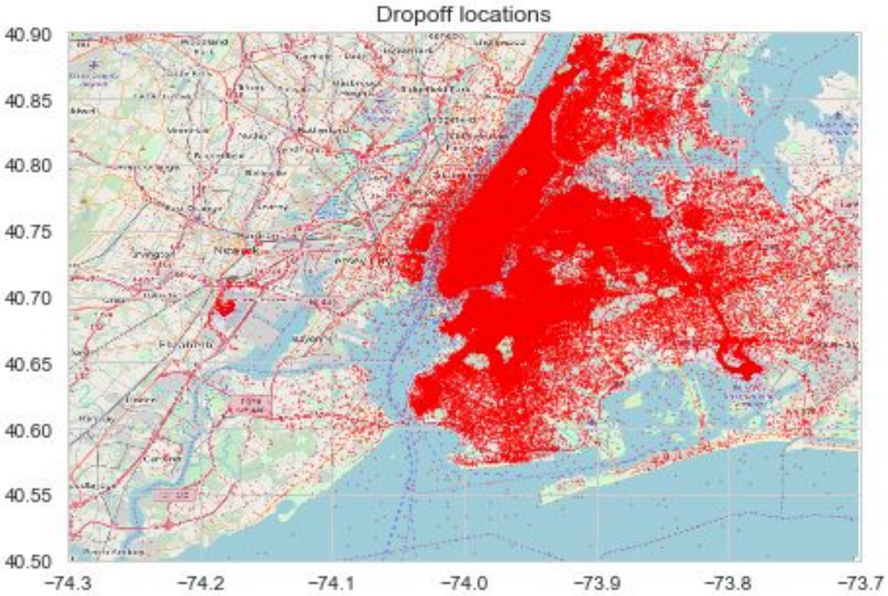
Taxi Cab Rides by Hour



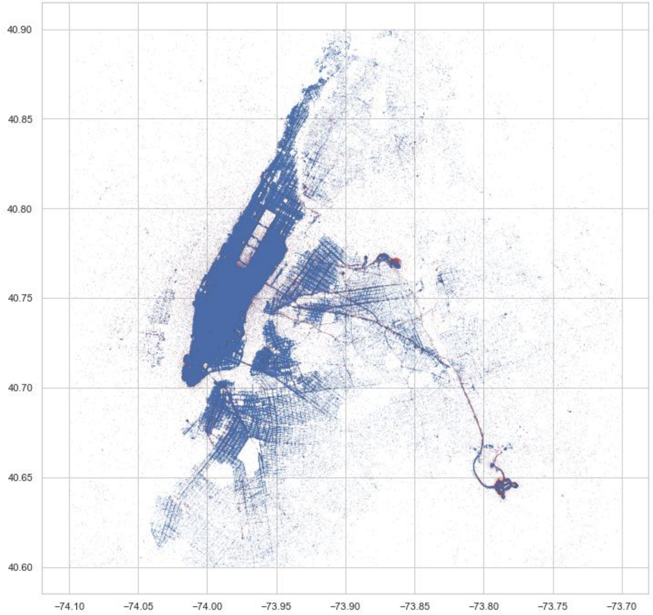
Taxi Cab Rides by Day



Taxi Pickup/dropoff Locations



NYC Pick Ups



40.90
40.85
40.80
40.75
40.70
40.65
40.60
−74.10
−74.05
−74.00
−73.95
−73.90
−73.85
−73.80
−73.75
−73.70

Data Processing

	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	month	day	hour	distance_traveled
1045136	-73.976	40.752	-73.975	40.742	1	2013	10	26	7	0.011
342264	-73.993	40.748	-74.006	40.731	2	2011	9	16	19	0.021
2138657	-73.981	40.748	-73.989	40.737	2	2009	11	14	12	0.014
1480376	-73.951	40.810	-73.956	40.818	1	2012	1	28	21	0.008
1570444	-73.975	40.760	-73.993	40.768	1	2011	9	23	23	0.019

- Calculated Euclidean distance with coordinate data
- Stripped “pickup_datetime” to hour, day, weekday, month, and, year
    - Hot encoded weekday
- Training Set shape
    - 1881700 rows
- Test Set Shape
    - 806443

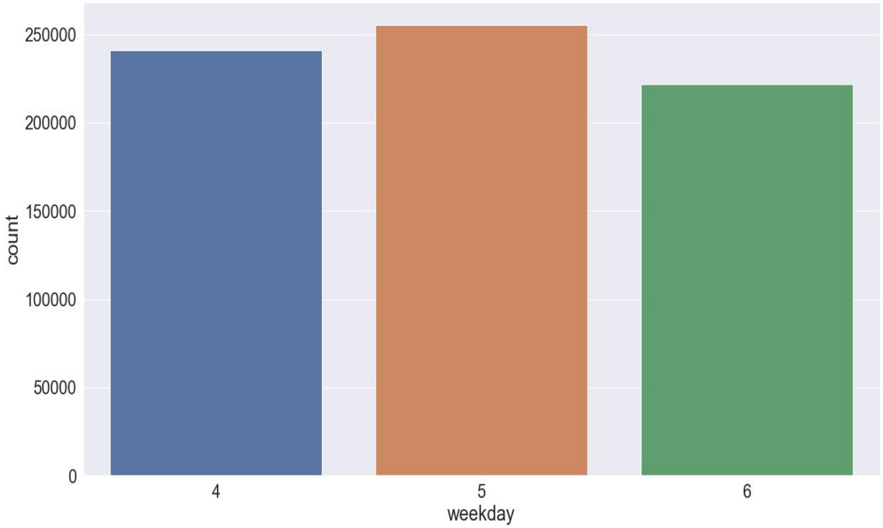
Key Objectives

- What are the characteristics of clusters in the data?
- Predict Tax Fare price
  - Use prediction as a part of Mobility as a Service (Maas)
  - Enhance mobility through Maas

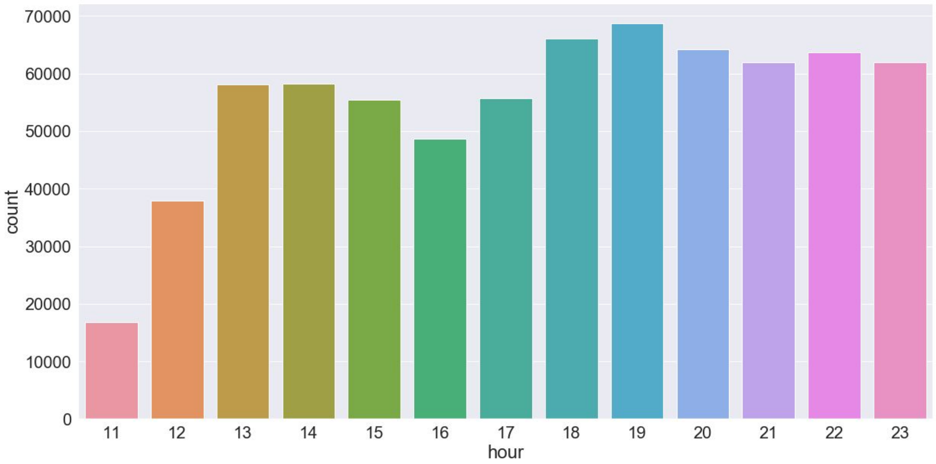
Cluster on Entire Dataset

	clusters_all_set	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	hour	weekday	distance_traveled
0	0	11.150	-73.975	40.751	-73.976	40.751	17.666	4.973	0.033
1	1	11.345	-73.974	40.752	-73.974	40.752	8.979	0.973	0.033
2	2	11.391	-73.977	40.749	-73.973	40.750	6.029	4.406	0.035
3	3	11.199	-73.975	40.751	-73.975	40.752	18.807	1.678	0.033

Cluster 1

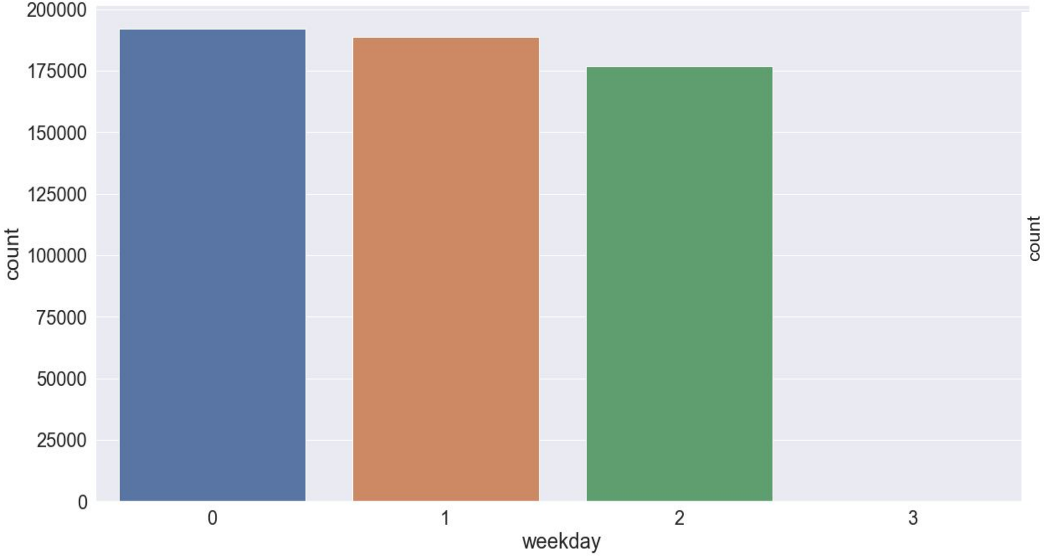


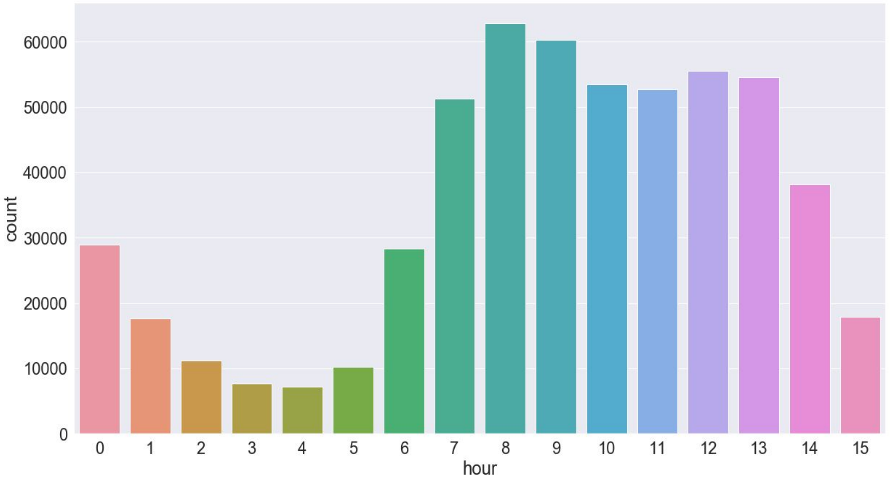
250000
200000
150000
100000
50000
0
count
4
5
6
weekday



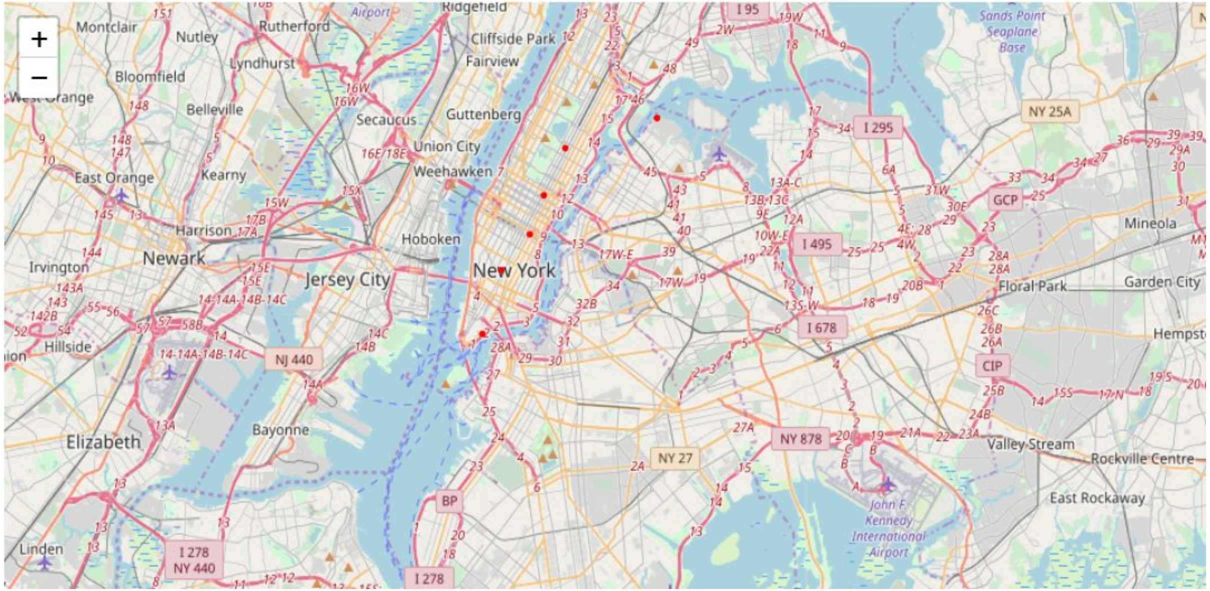
70000
60000
50000
40000
30000
20000
10000
0
count
11
12
13
14
15
16
17
18
19
20
21
22
23
hour

Cluster 2


count
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
weekday


count
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
hour

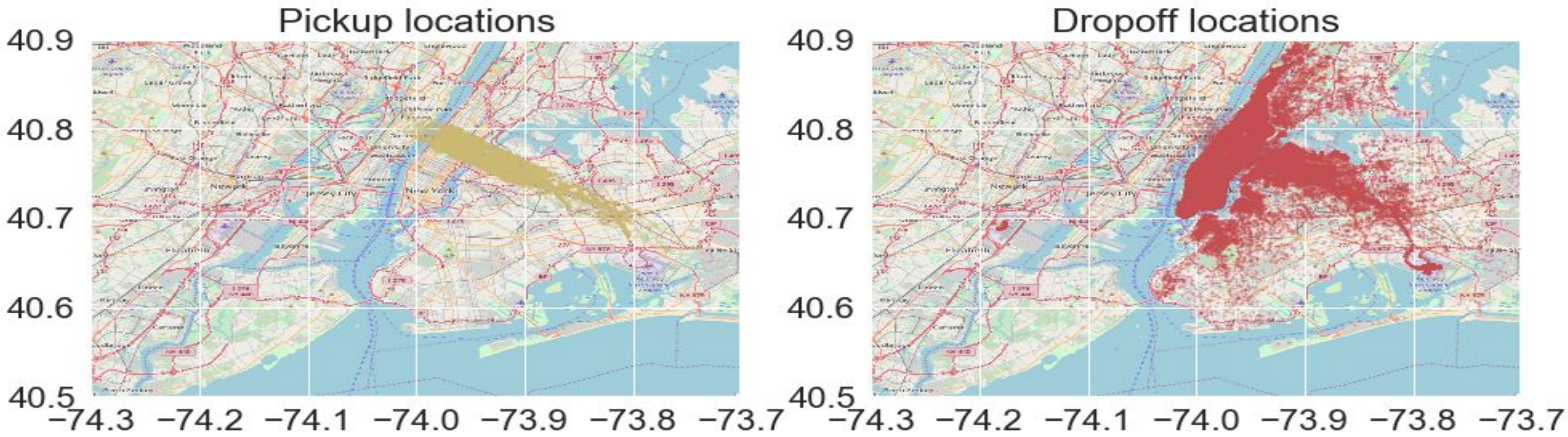
Geoclustering



Geocluster

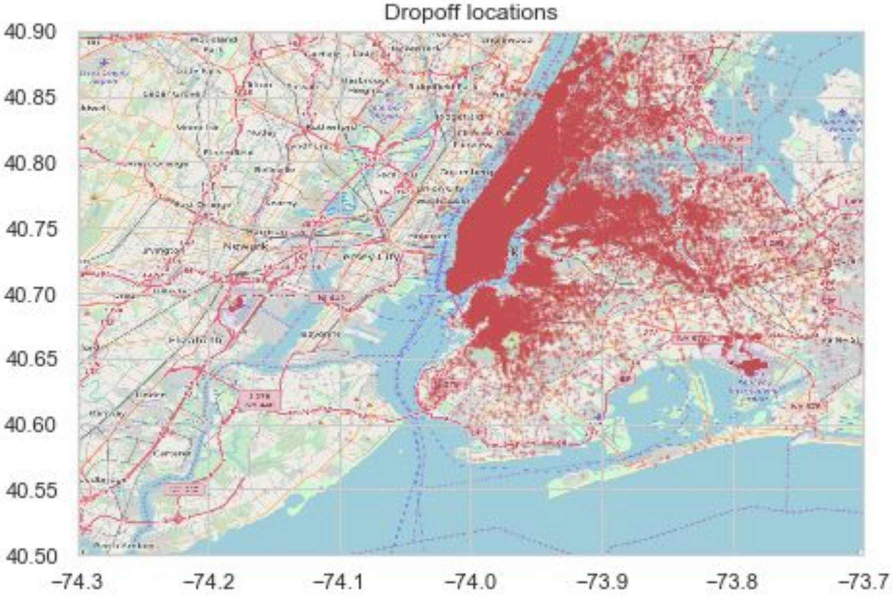
	clusters_loc	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	month	hour	weekday	distance_traveled
0	0	9.995	-73.959	40.779	-73.966	40.767	6.252	13.459	2.985	0.029
1	1	10.253	-73.994	40.731	-73.983	40.739	6.260	12.855	3.300	0.030
2	2	10.809	-73.971	40.760	-73.972	40.757	6.263	14.044	2.892	0.032
3	3	13.307	-74.004	40.706	-73.984	40.729	6.315	13.508	3.100	0.042
4	4	20.661	-73.909	40.790	-73.956	40.767	6.342	13.927	2.966	0.071
5	5	11.359	-73.979	40.745	-73.977	40.748	6.270	13.549	2.991	0.034

Geocluster 0



- Fare amount $9.95
    - $2 below mean

Geo Cluster 4



- Fare Amount $20.66
    - $9 above mean

Day of Week and Hour Clusters

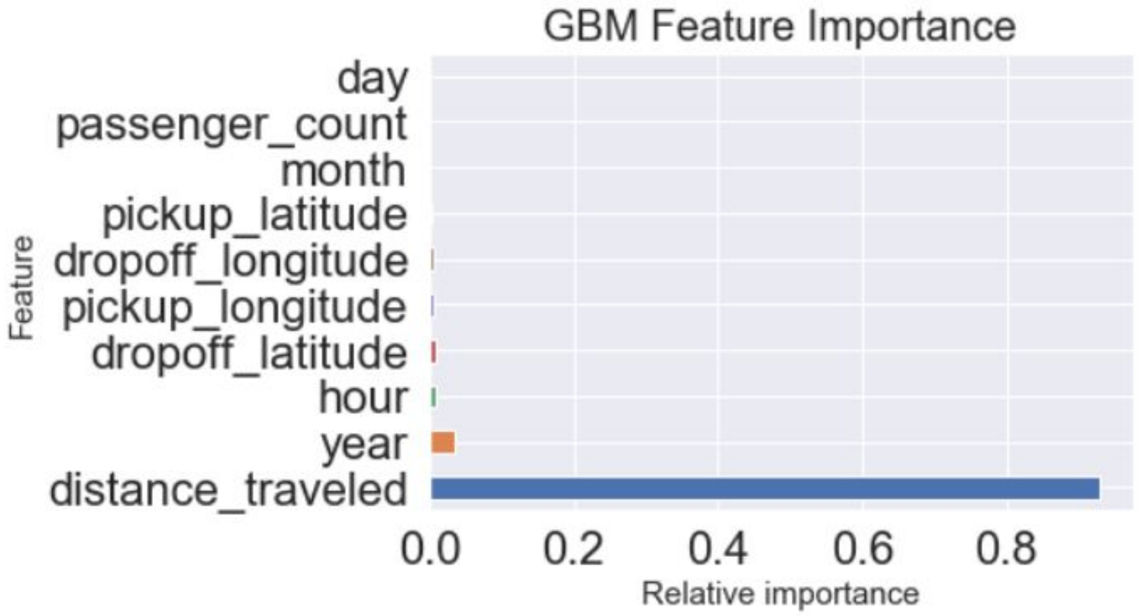
- Day of Week Cluster
  - Simply split clusters into days of the week as expected
- Hour Cluster Created 3 nonsense clusters
  - 7 pm
  - 11 am
  - 2 am

Taxi Fare Prediction

Model Comparison

Model	RMSE Test	RMSE Train	Mean Absolute Percentage Error	Variance
Baseline	9.33	NA	NA	NA
Linear Regression	4.239	4.233	78.53%	-0.006
Gradient Boosting	3.71	3.7	83.76%	-0.009
Random Forest Regressor	3.52	1.49	81.17%	-2.026
XGBoost	3.35	3.27	82.19%	-0.084

Feature Importance



Conclusions

- Clustering worked fairly well
  - Geoclustering was able to pinpoint popular destinations
  - Time cluster developed clusters at odd hours
    - Perhaps investigate another method to cluster time of day
  - Cluster by time of day just clustered by day as expected
- A Taxi Fare prediction application could be useful for **consumers**
  - Budget and plan their trip
  - Compare prices with Uber
- Taxi Cab **Owners** could utilize fare prediction as well
  - Deploy drivers at optimum times to reduce costs
  - Allow taxi cab companies to adjust fares in regards to surge pricing

Conclusions

- **Developers**
  - Taxi fare prediction could be useful to to developers of MaaS (Mobility as a Service)
    - Combine transportation services from public and private transportation providers through a unified gateway that creates and manages the trip
    - Users pay for with a single account. Users can pay per trip or a monthly fee for a limited distance.
    - The key concept behind MaaS is to offer travelers mobility solutions based on their travel needs. [2]
      - i.e. , Getting there cheaper or faster

Conclusions

- Next steps
    - Split train and test before running clusters
        - Then use clusters to predict!
    - Generate additional features to include popular destinations
    - Combine other datasets to increase exactness
            - Weather data
            - Combine information from other sources to develop MaaS App
                - Public transit data
                - Regional public transit
                - Uber
                - Airline data
                - Traffic data

Sources

[1] - https://www.statista.com/chart/13480/ride-hailing-apps-surpass-regular-taxis-in-nyc/

[2] - https://en.wikipedia.org/wiki/Mobility_as_a_service