Analyzing Chicago Taxi Trips Behaviour

Shaurya Sahai, Abishek Krishnan, and Sushmitha Sunkurdi Nataraj
School of Computing
University of Utah

INTRODUCTION

Chicago Sun-Times says, “There are about seven thousand licensed cabs operating within the city limits.” Our project focuses on taxi trips within the city of Chicago, in this project; we aimed to find interesting insights about one of the significant components of any trip - the average speed and try to answer questions like-

“If I travel at 5:30 PM on a Monday, what would be my expected average speed?”

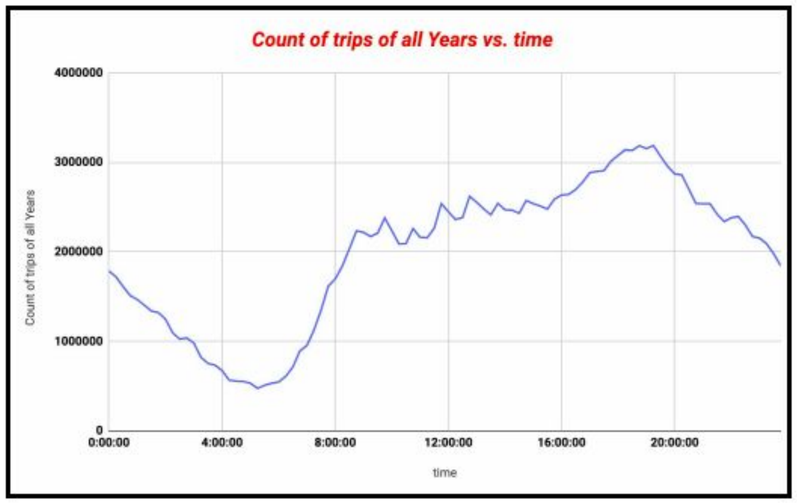
Or

“What can I expect my average speed to be while I take a ride on Wednesday at 4 PM in Chicago downtown?”

DATA

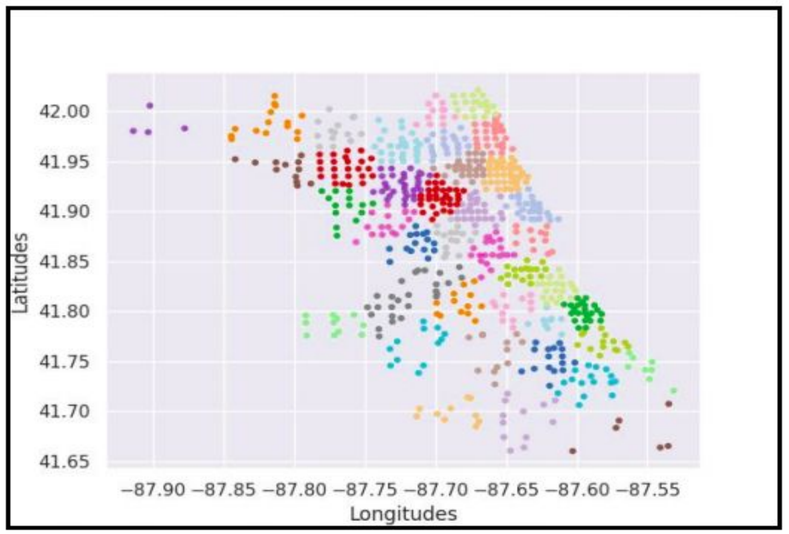
- The data was collected from Kaggle, which in turn is hosted by Google BigQuery.
- The data is still being collected by the City of Chicago (the regulatory agency) from 2013 and is continuously updated.

DATA EXPLORATION



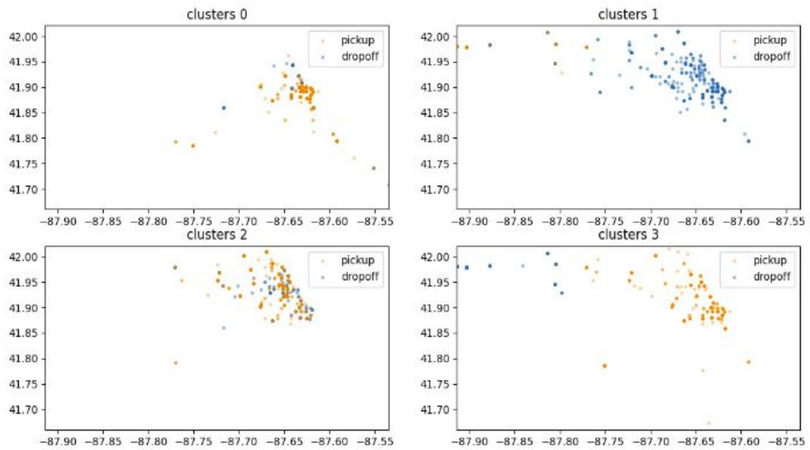
The figure shows that peak time for booking a taxi is between 5PM - 8PM and off peak hours between 3 am - 6 am.We also calculated the top 10 busiest rides during 2019.

CLUSTERING



kmeans++ clustering of pickup locations

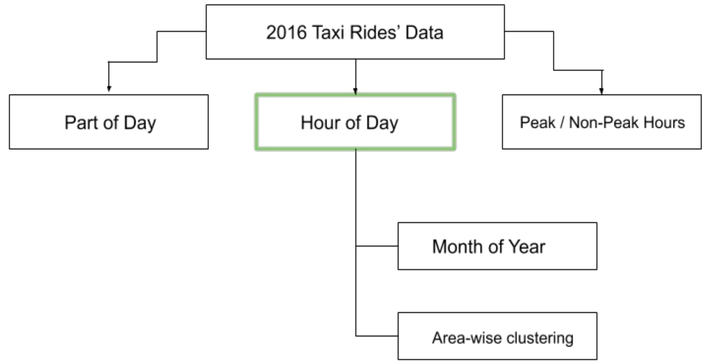
We found a driver can cover around 2.5 miles in 10 minutes with an avg speed of 15.07 miles/hr. Using this and K-means++, we clustered the pickup locations into 40 meaningful clusters. We also used DB-SCAN clustering to concretize the above.



We improved our metrics by clustering on 4 dimensions. [Latitudes, Longitudes] of pickup and dropoff locations. [Resulted in 4 clusters]

REGRESSION

We used regression to model the average speed predictions. We had multiple contributing factors in finding the average speed. We employed a step-by-step approach to finding the most efficient model.



RESULTS

	KEY IDEA	MEAN ABSOLUTE ERROR	MEAN SQUARED ERROR	ROOT MEAN SQUARED ERROR
1	PART OF THE DAY	9.4302	334.6	18.292
2	HOUR OF THE DAY	5.123	46.77	6.839
3	PEAK AND NON PEAK HOURS	9.361	333.3	18.256
4	MONTH OF YEAR	10.349	402.28	19.211
5	AREA WISE CLUSTERING			
	CLUSTER 1	12.741	12295	70.01
	CLUSTER 2	11.392	8771	62.65
	CLUSTER 3	11.341	8748	62.59
	CLUSTER 4	11.357	677.02	26.01

LEARNINGS AND CONCLUSION

- From our experiments, we learned that the unclustered data classified by the hour of the day works best in predicting the average speed of the ride. (Improvements: To Include other factors while modelling)