GE5219: Spatial programming
Final Group Project Report

Exploratory Data Analysis of New York City Yellow Taxi Data

Chen Xinyu A0198779W

Ho Shi Yun A0225536E

Lyu Wenling A0224892X

Catalogue

Introduction

Over the last decade, due to the marked growth of population along with jobs and tourists, roadways in New York City (NYC) have become more congested than ever especially within downtown areas in Manhattan. Aside from general public transportation, private vehicles and hiring taxis are two other popular transportation modes among residents in NYC. The Mobility Report of New York City (2019) shows that even though overall number of cars entering Manhattan CBD continue to decrease, empty For-Hire taxis circling the area brought no relief to traffic congestion.

Another issue which compounds the problem of traffic congestion is that demand for ride hailing taxis in NYC exhibits high variability in trips recorded and spatial patterns exhibited (Safikhani et al., 2020). A cursory glance at the Yellow taxi trip records will reveal that there are around 11 million trips recorded for the month of June in 2016 alone. Indeed, such sheer magnitude of data and variability of demand for taxis within NYC points towards the dire need for the application of data analytics and machine learning to draw sufficient taxi fleets to where it is needed the most and alleviate the problem of traffic congestion.

As such, this paper will attempt to conduct exploratory data analysis of New York taxi data and characterize spatial and temporal attributes of New York Yellow Taxi data using a combination of summary statistics and clustering algorithms. Our analysis details our attempt at quantifying spatiotemporal trends of high taxi demand in NYC and thereby identifying accurate spatial pattern of Yellow Taxi trips through exploration of the clustering algorithms K-means, DBSCAN and HDBSCAN.

As the big data nature of such phenomena calls for a data analytics approach to understand the characteristics of this issue, we will follow a variation of the OSEMN (Obtain, Scrub, Explore, Model, Interpret) model by Mason and Wiggins (2010) to guide processing of our data and address the following research questions:

a) What regions have the most pickups and drop-offs?

b) What are the characteristics of traffic flows?

c) What are the differences between short and long-distance trips?

d) What is the temporal pattern of Yellow Taxi data?

e) Can clustering algorithms to identify a more accurate spatial pattern of Yellow taxi trips?

Study Area

Our study area spans across the entirety of New York City (40.730610, -73.935242) as shown in Fig. 1. As we are keen on capturing the spatiotemporal attributes of demand for taxis, our study area will include the five boroughs of NYC: Bronx, Manhattan, Queens, Brooklyn and Staten Island. With about 264 taxi zones within the study area, data on taxi will hence be revealing of the vast and dynamic nature of demand for taxis in between boroughs.

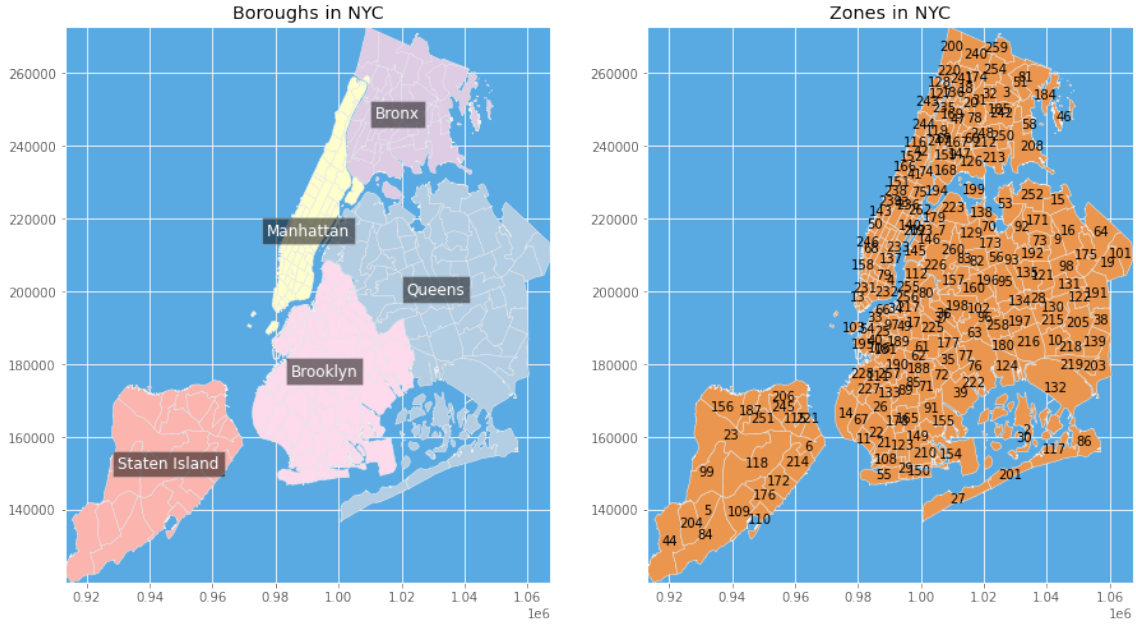

Figure 1. Study Area of New York City

Fig. 2 shows the land use distribution in New York City, where Zhang et al. (2017) posit that it might strongly correlate with the traffic patterns. We can see that land use is not evenly distributed in every district. The five boroughs of New York City all have their own positioning and different functionality (Techspo, n.d.). Bronx is in the northernmost of New York City, and it is a highly residential borough. Brooklyn, on the other hand, is home to a diverse and growing retail market, entertainment and residential areas.

As the borough with the highest population density, Manhattan is the largest county-level economy in the state and employs 61.6% of the City's workers. Queens is one of the most active industrial areas in the five boroughs while Staten Island has more than 12,300 acres of park area and around 4,000 acres of waterfront.

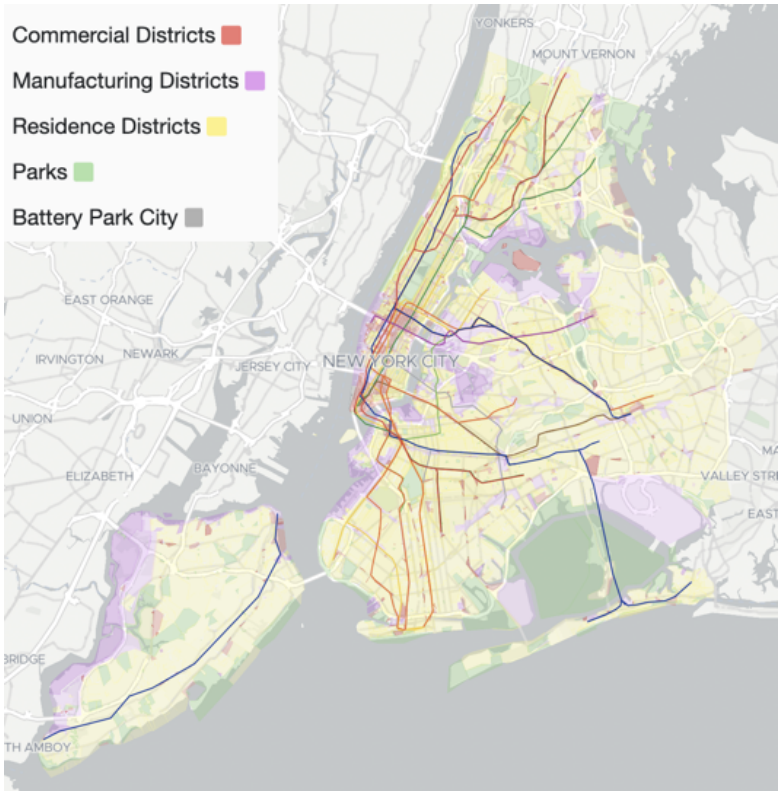

Figure 2. Land Use Map of New York City

Data

Data source

Our exploratory analysis entails the use of the Yellow Taxi Trip Records of NYC in June 2016. Our choice of obtaining relatively dated data (as part of first step in OSEMN framework) is primarily driven by the lack of precise geolocation data released by NYC Taxi and Limousine Comission (TLC) in light of privacy concerns. As we are concerned with precision of existing algorithms to characterize taxi demand and trends, aggregated data is therefore not our prime choice for analysis.

Whilst relatively dated, this data contains information about 11,135,470 Yellow taxi trip records, which will prove to be a sizeable amount of data to unveil interesting insights into the dynamics of taxi traffic within New York. A summary of the dataset used for our study is in Table 1:

Table 1. Dataset used for our study

S/N	Dataset Name	Dataset Description	Data Source	Data Format	Last updated
1	Yellow Taxi Trip Records	Pickup_longtitude (longititude where meter was engaged) Pickup_latitude (latitude where meter was engaged) Dropoff_longitude (longitude where meter was disengaged) Dropoff_latude (latitude where meter was disengaged) Trip_distance (elapsed trip distance in miles)	NYC Taxi and Limousine Commission (TLC)	CSV	June 2016

Preliminary data visualization

Given the extensive records and data evident within our NYC Yellow Trip Records, process of data cleaning (i.e. scrubbing of OSEMN) will require visualisation of data to eliminate anomalies in data. As we are guided by motivations to ascertain true geographic context of the area, this dictates the need for such big data to be examined relative to its geographic location and satellite imagery.

As such, we adopted the use of Kepler in visualizing such enormous dataset. A series of maps were generated: (a) map depicting taxi flow across NYC (Fig. 3); (b) map illustrating distribution of pickup and drop-off points in NYC (Fig. 4a and 4b); and (c) illustration of pickup and drop-off points in NYC against satellite imagery (Fig. 5a and 5b).

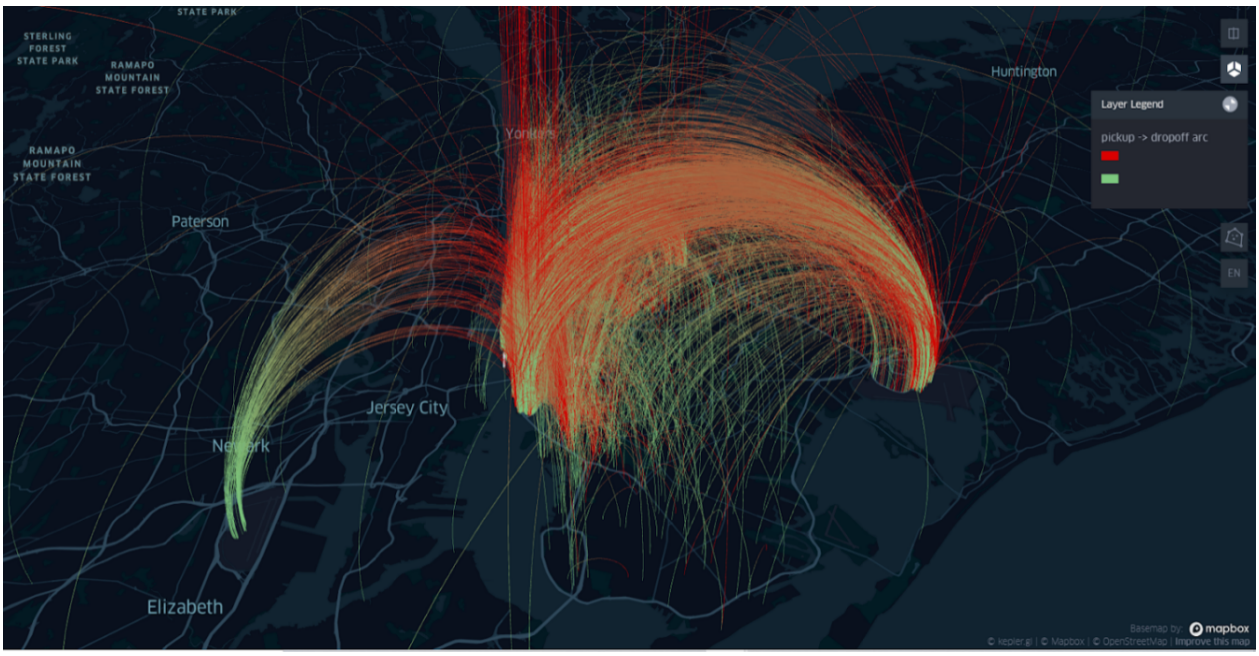

Figure 3. Data Visualization of taxi flows on Kepler

As evident from Fig. 3, visualisation on Kepler offers an instant revelation of inaccurate logging of some coordinates given that it is almost impossible for taxis within New York to travel to other countries within a day.

Fig. 4a and 4b show the spatial distribution of pickup and drop-off points within New York. Both pick-ups and drop-offs largely exhibit similar spatial pattern where taxis are

concentrated on Manhattan over other regions and points being distributed in varying densities.

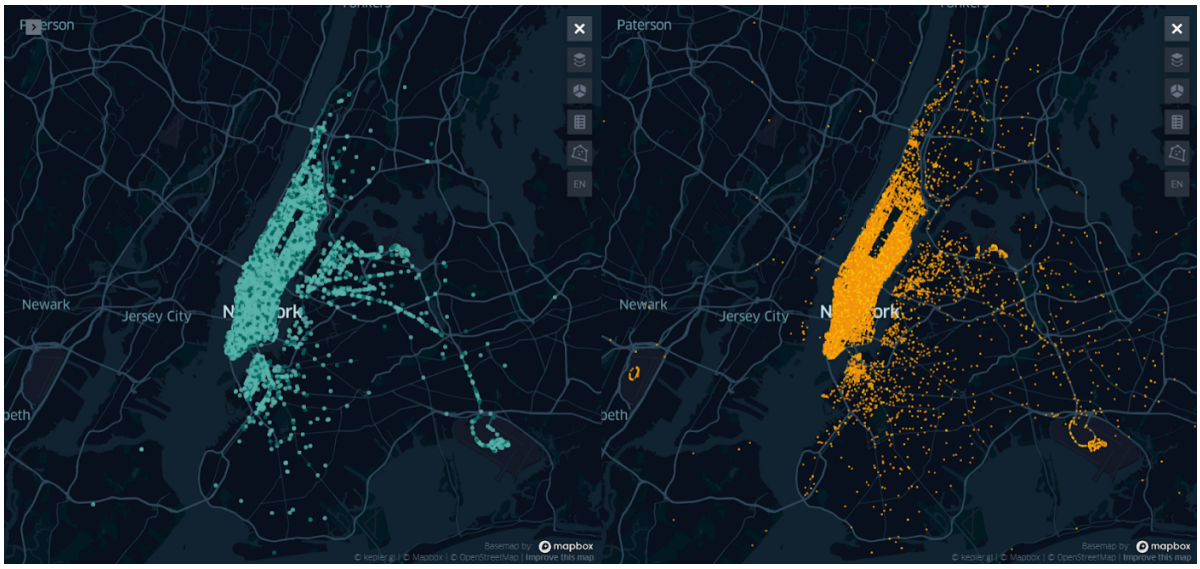

Figure 4a. Pick-up points across NYC

Figure 4b. Drop-off points across NYC

Fig. 5a and 5b show the same spatial distribution of points albeit with satellite imagery. Within Manhattan, Central Park has little to no pick-ups and drop-offs and inner-city areas (such as Harlem) exhibit significantly less dense distribution in comparison to the south of Central Park.

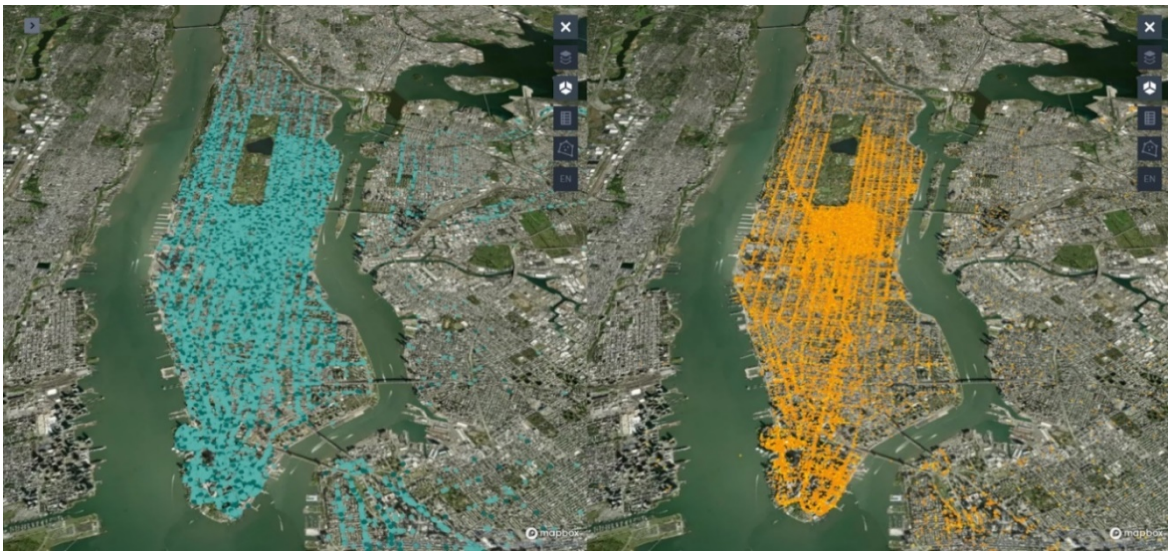

Figure 5a. Pick-up points across NYC in comparison to satellite imagery

Figure 5b. Drop-off points across NYC in comparison to satellite imagery

Methodology

Framework

Given that our attempt at analysing taxi data is an exploratory approach, our methodology will be primarily driven towards processes which can visualize spatiotemporal attributes of data and identify spatial clusters where taxi demand is relatively higher. A summarized workflow of the methodology adopted for our analysis can be seen in Fig. 6:

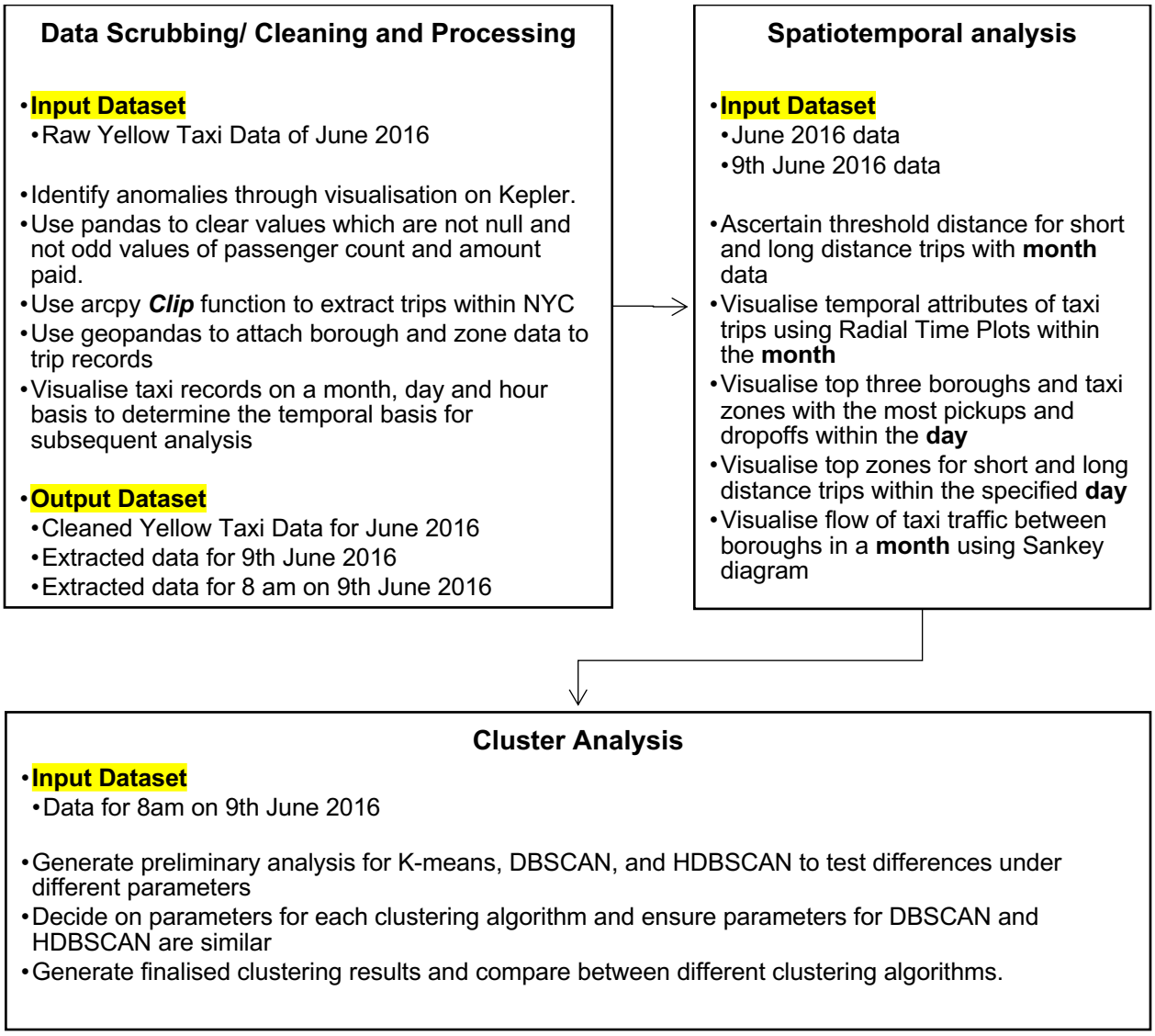

Figure 6. Workflow adopted for our study

Given that our research objective is to conduct exploratory analysis of areas of higher taxi demand within NYC, we decided to use clustering algorithms K-means, DBSCAN and HDBSCAN and compare its applicability to understanding the spatial nature of taxi demand.

K-means

Proposed by MacQueen (1967), the objective criterion used in k-means clustering method is a squared-error function defined by this equation:

$$E = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} |x - m_i|^2$$

where X represents the point in space representing the given object, and m_i is the mean of the cluster C_i. (MacQueen, 1967)

Han et al. (2001) provides an excellent preface as to how the iterative relocation algorithm works. According to them, K-means is a partitioning clustering method where points are divided into clusters which share similarities. The centroid (or the mean) of objects in the cluster is denoted as cluster centre. The K, in this case, is determined by running a series of k determined by us and thereby using silhouette analysis to determine the optimal number of clusters. Known to be relatively scalable and efficient in processing large datasets because of the computational complexity of the algorithm, this clustering method will be compared against the density clustering methods DBSCAN and HDBSCAN in our analysis.

DBSCAN

Proposed by Ester et al. (1996), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm which grows regions with sufficiently high density into clusters, and is proficient in discovering clusters of arbitrary shape in spatial databases. Under DBSCAN, clustering result depends on two parameters: *ε* (*Eps*) and *minPts.* ε is the maximum radius of the neighborhood while *minPts* denotes minimum points in neighborhood for a point to be a core point (Tang et al., 2021). As shown in Fig. 7, if a point has more than the specified number of points (MinPts) within the Eps, it will be defined as a core point. A border point is a point that has fewer points in Eps than MinPts

but is in the vicinity of a core point. Outliers (noise) points are the points that are neither core points nor border points.

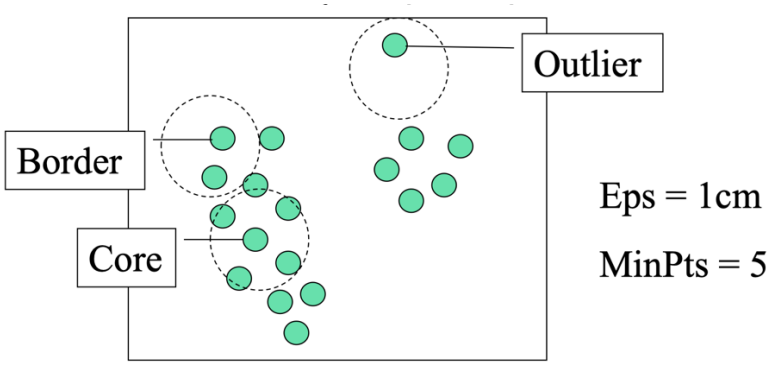

Figure 7. An example of DBSCAN (Eps = 1cm, MinPts = 5)

HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is a clustering algorithm proposed by Campello et al. (2013). HDBSCAN searches high density regions separated by low density regions in the space of the input data set. A cluster stability metric and "mutual reachability distance" are used to determine the meaning of "high density" for a given data set (Melvin et al., 2018). The mutual reachability distance is defined as follows (McInnes et al., 2016):

$$d_{\text{mreach}-k}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}$$

where *d(a,b)* is the original metric distance between a and b.

There are three parameters in HDBSCAN (McInnes et al., 2016): 1) min_cluster_size: the smallest size grouping that you wish to consider a cluster; 2) min_samples: a measure of how conservative the clusering will be, where a larger number repesents a more conservative clustering – more points will be declared as noise, and clusters will be restricted to more dense areas; 3) cluster_selection_epsilon: a parameter that ensure clusters below the given threshold are not split up any further.

The choice of two density-based clustering algorithms within comparison with K-means is primarily due to varied density of points distributed across NYC. HDBSCAN is noted to be more focused on high density areas while DBSCAN may sometimes create clusters even with relatively low density. It will therefore be interesting to draw comparisons with

such clustering methods with such nuanced differences and sensitivity of clustering algorithms.

Results

Spatiotemporal patterns

a. Regions with highest pick-ups and drop-offs

We first started with analysis from a broader scope at the borough level. Our results show parallels in patterns of boroughs with most pickups and dropoffs. Unsurprisingly, Manhattan, which has the highest population density in NYC, is the busiest borough. Bronx and Staten Island, on the other hand, with large areas of parks and waterfront, are the least busy boroughs (Fig. 8).

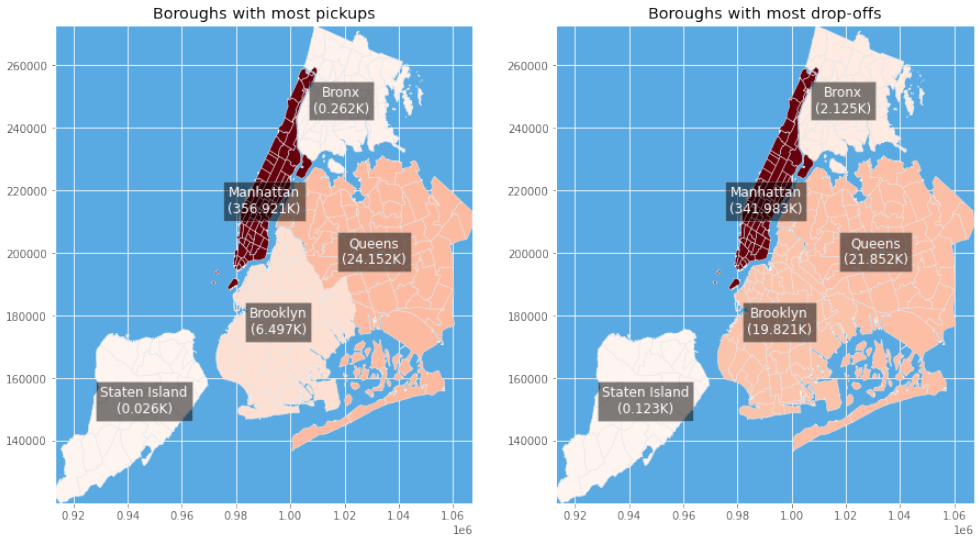

Figure 8. Boroughs with the most pick-ups and drop-offs

On a more specific level, top three zones for pickups are found to be at Midtown Centre, Upper East Side North, and Upper East Side South in Manhattan. These areas are also the top three zones for drop-offs as shown in Table 2 and Table 3:

Table 2: Taxi Zones with the most pickups

LocationID	Taxi Zone	Borough	Pickup Count
161.0	Midtown Centre	Manhattan	14146
236.0	Upper East Side North	Manhattan	16321
237.0	Upper East Side South	Manhattan	17482

Table 3: Taxi Zones with the most drop-offs

LocationID	Taxi Zone	Borough	Dropoff Count
237.0	Upper East Side South	Manhattan	17482
236.0	Upper East Side North	Manhattan	16321
161.0	Midtown Centre	Manhattan	14146

Top three taxi zones generated from our analysis are Midtown Centre, Upper East Side North, and Upper East Side South for both pickup and drop-off points as shown in Fig. 9. Notably, the taxi zone of JFK airport in Queens is also a region with relatively high demand for taxis.

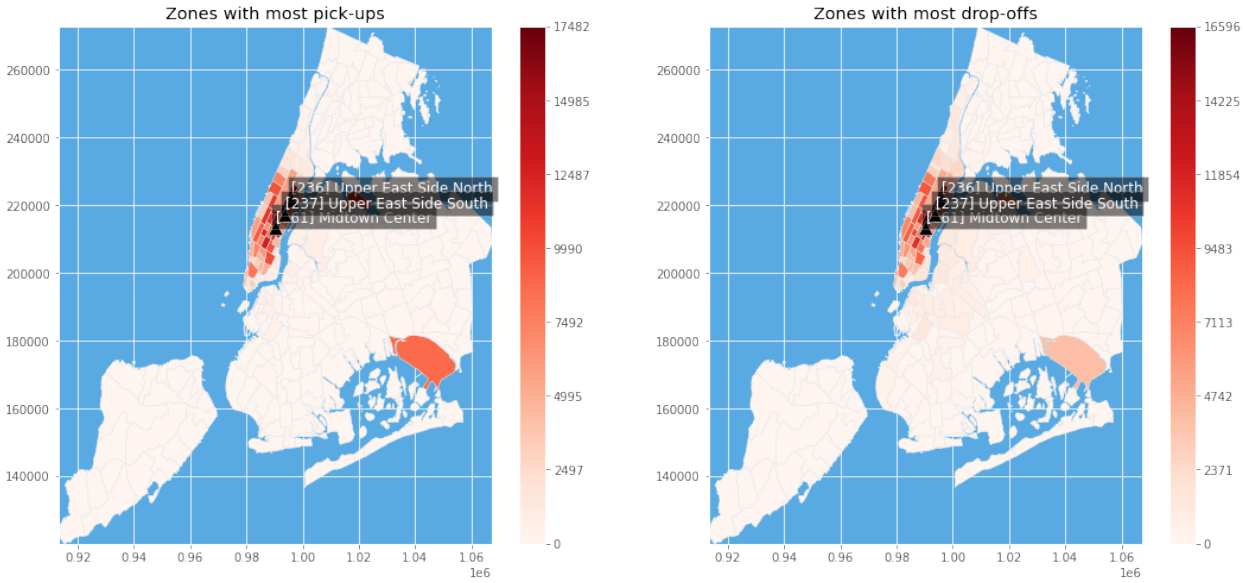

Figure 9. Taxi Zones with the most pick-ups and drop-offs

Compared with the land use map of New York City (Fig. 10), we can see that large area of commercial districts coincide with dense road networks in the taxi zones with most pickups and drop-offs. This indicates that commercial areas with high density road networks and airport area might be the busiest regions for Yellow taxis, while other land uses such as industrial area, park area and residential area might have relatively less demand for taxis.

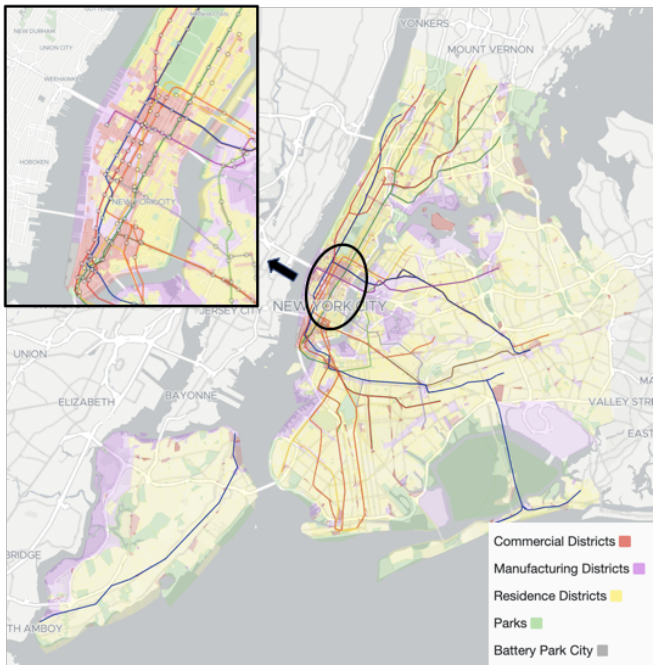

Figure 10. Land use map of New York City

c. Characteristics of taxi flow

Fig. 11 shows the Sankey diagram of taxi flow among different broughs. In the diagram, the thicker the nodes, the greater number of taxis travelling to and forth the zones. It shows that most taxis travel within Manhattan. Flows between other regions are considerably smaller, with flow between Manhattan and Queens, and Queens to Manhattan being the next largest flow of taxis.

Figure 11. Sankey Diagram illustrating flow between taxi boroughs

d. Ascertaining characteristics of short and long-distance trips

To determine the temporal characteristics of taxi data at different spatial scales, we have to first define short and long-distance trips. As shown on Fig. 12, distribution of trip distance in the histogram is rather skewed, with most trips happening within the first 40 miles.

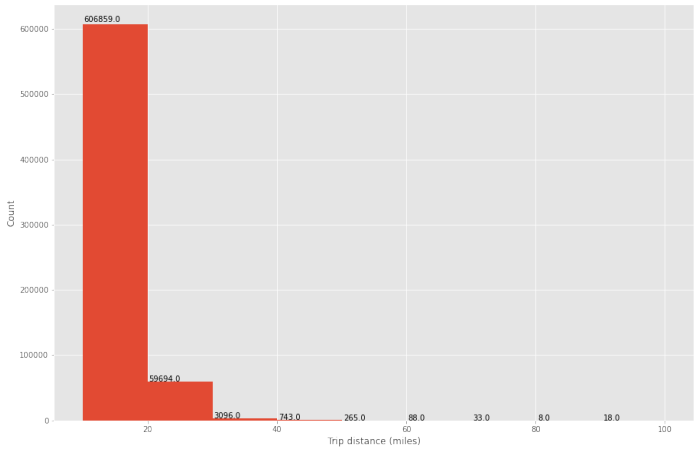

Figure 12. Distribution of trip distance for taxi rides (as visualised on Python using matplotlib)

Upon closer scrutiny of the histogram at a smaller scale (as shown in Fig. 13), most trips are made between 1 mile and 24 miles. As such, we determined short distance trips as less than 24 miles, and long-distance trips as 24 miles and above.

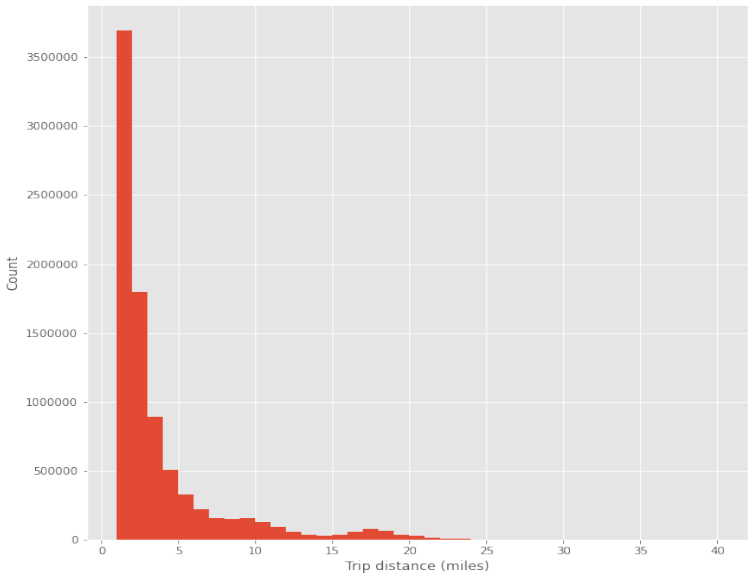

Figure 13. Distribution of trip distance for taxi rides of less than 40 miles (as visulised on Python using matplotlib)

Spatial Attributes of short and long-distance trips

Fig. 14 shows that top three taxi zones with the most pickups and drop-offs, long and short distance trips have no evident differences. In particular, taxi zones within Manhattan and JFK Airport are the most popular pickup and drop-off zone in NYC.

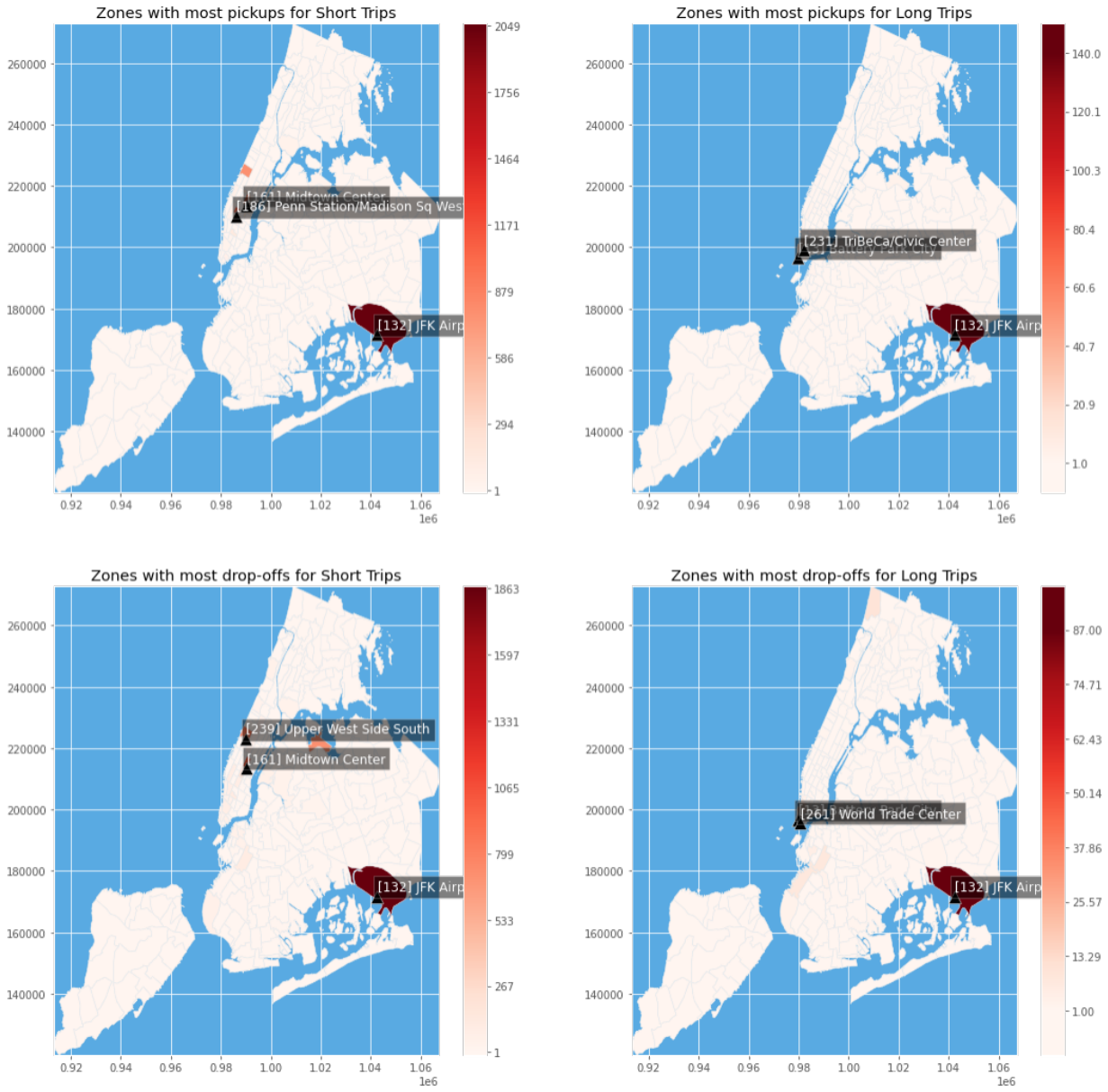

Figure 14. Short and long-distance trips with the most pickups and drop-offs

Temporal Attributes of short and long-distance trips

In order to determine the temporal characteristics of taxi data, a month's worth of data is chosen to recognise the peak hours where pickups and drop-offs are the highest.

For long distance trips, pickup time peak hours are between 130 to 230pm and 930 to 1130 pm while drop-off peak hours are between 130 to 330pm, 530 to 730 pm and 1030pm and 1230 am. For short distance trips, both pickup time and drop-off peak hours are within the same time frame from 530 to 730 pm. As we can see from the blue radial plot, long distance trips have more pronounced peak hours (Fig. 15). Such temporal patterns can point towards the possibility of nature and motivation for travel using taxis as transportation modes, where shorter distances may be motivated by the need to travel for commercial activities and longer distance trips may primarily be driven by the need to travel to airports.

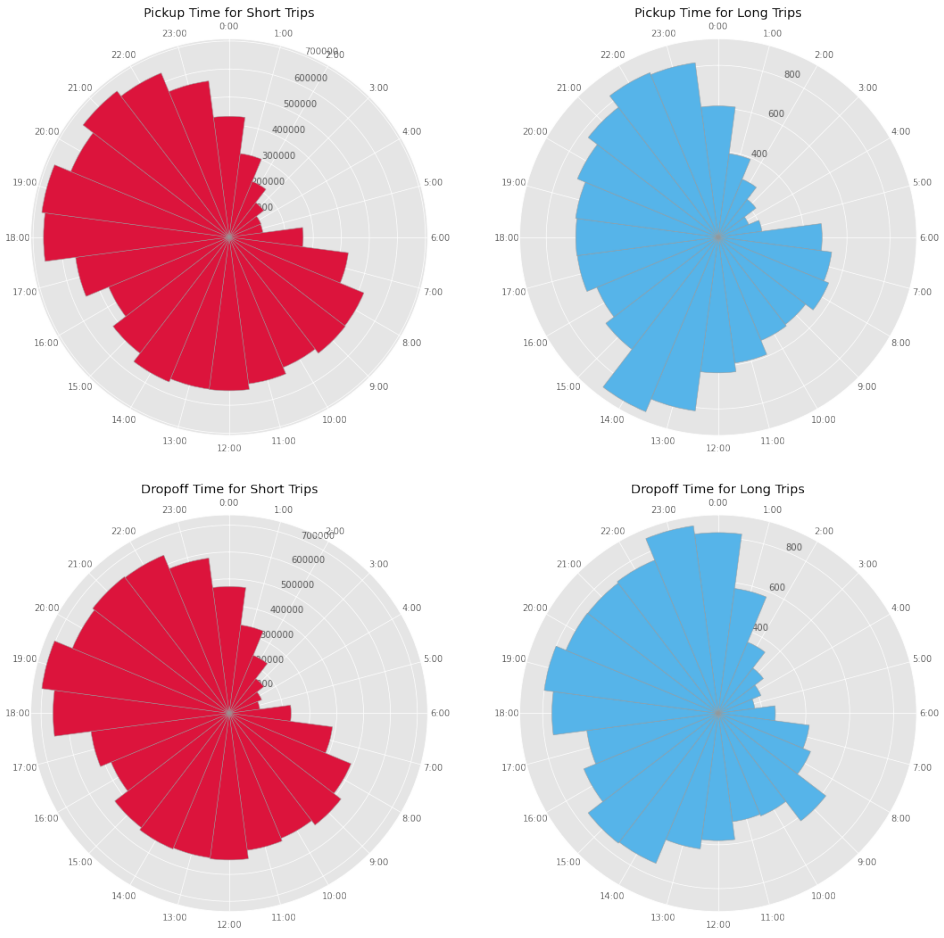

Figure 15. Radial Plots for short and long-distance trips in June 2016

e. Temporal pattern of Yellow Taxi data

Figure 16 and 17 depict the temporal pattern of yellow taxi data. Of all days within the month of June 2016, Thursday seems to be the day with the most pickups and drop-offs (Fig. 16). For the analysis of hours within one day, Pickups seem to be highest during 7-9 am in the morning, and 7-10 pm in the evening. Dropoffs seem to exhibit similar patterns, with 8-10 am in the morning being peak periods and 7-10 pm as the evening peak periods (Fig. 17).

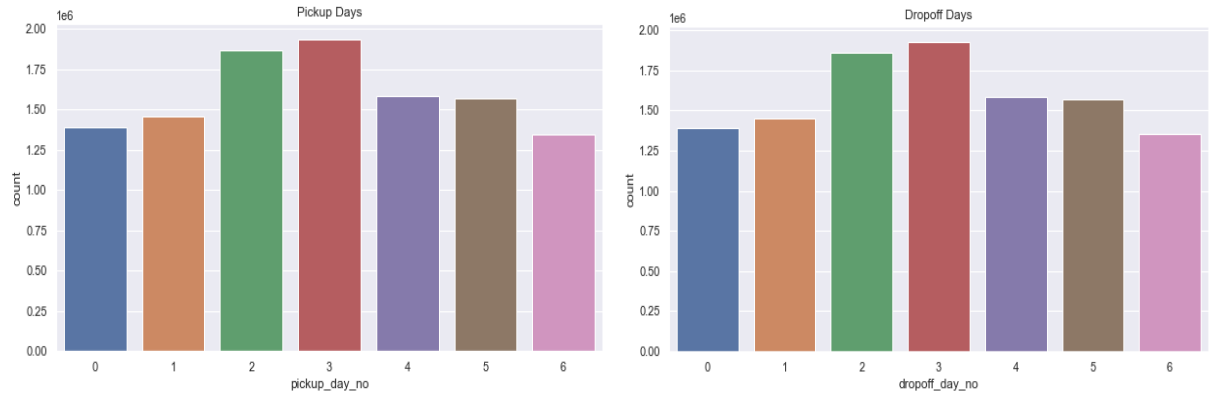

Figure 16. Counts of yellow taxi records on different days of a week in June 2016

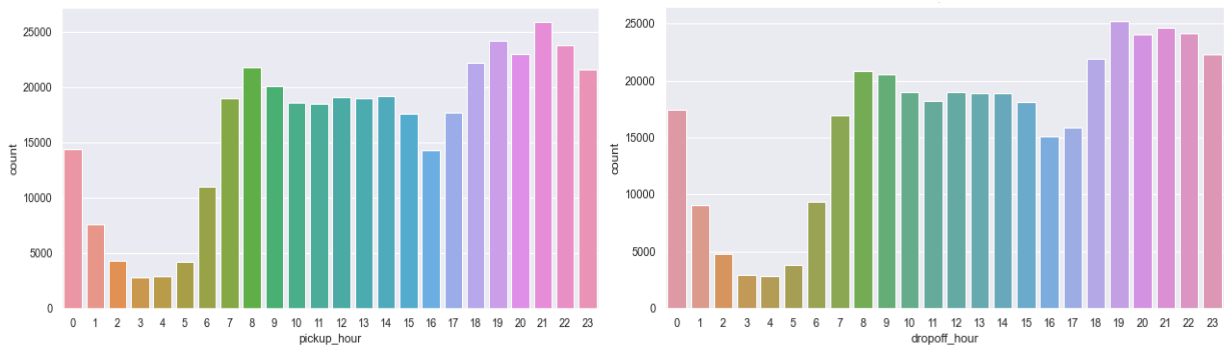

Figure 17. Counts of yellow taxi records in different hours on June 9th, 2016

Clustering Algorithms

a. K-means

Multiple attempts were made to produce outputs of K-means clustering. Such clustering result has been evaluated with Silhouette analysis, wherein the coefficient varies between -1 and 1; where values closer to 1 suggests that data points are well clustered and values closer to -1 suggest that data points are badly clustered (Wang et al., 2017).

At first, we have decided on the use of K=100 as there are multiple taxi zones within NYC and a large number such as this could perhaps reflect the key zones where demand is relatively higher. The clustering result is as shown in Fig. 18, with a silhouette score of ~0.408.

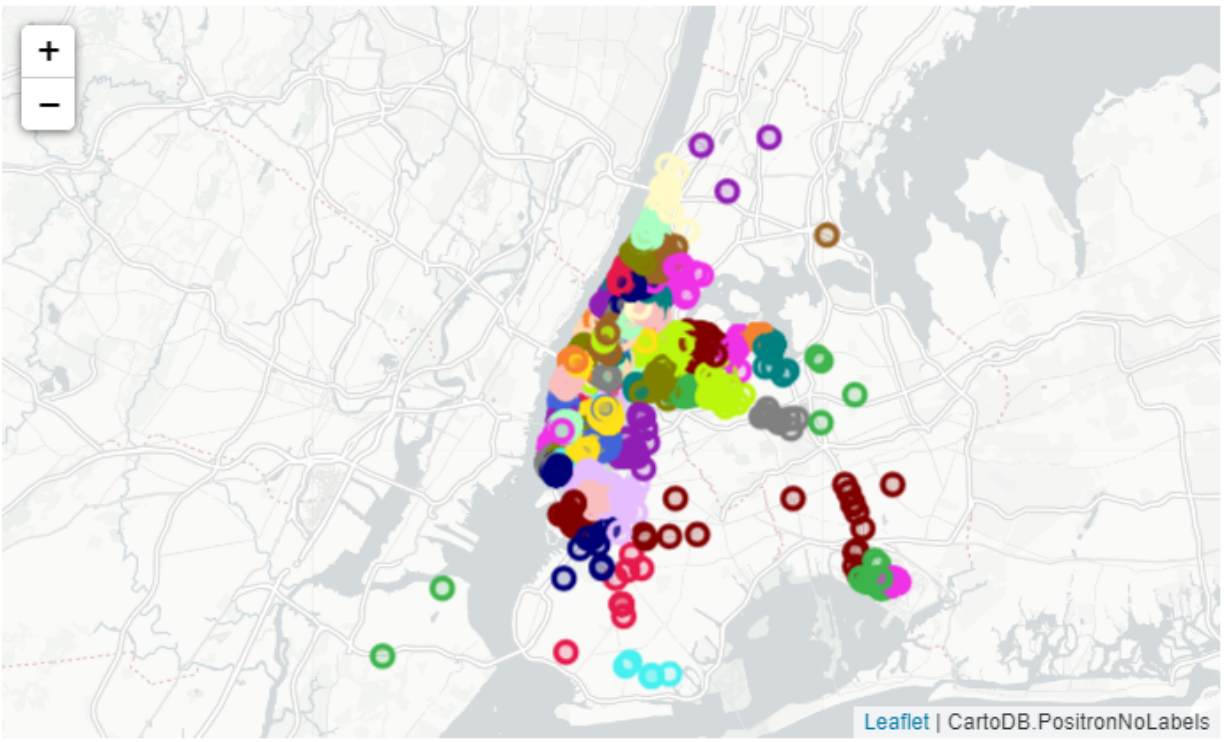

Figure 18. Results of K-means clustering when K = 100

The result shows a hundred clusters being generated, with noticeable points (such as the inner-city areas of Manhattan and West New Brighton to the southwest of Manhattan) being grouped as the same cluster despite being located rather far apart. In essence, under K = 100 has shown noise data grouped as the same cluster despite being relative far apart.

To ensure that comparison is valid, we also used Silhouette analysis to determine the best K value. We found out that K=2 scores the best amidst all other values with a silhouette score of 0.747.

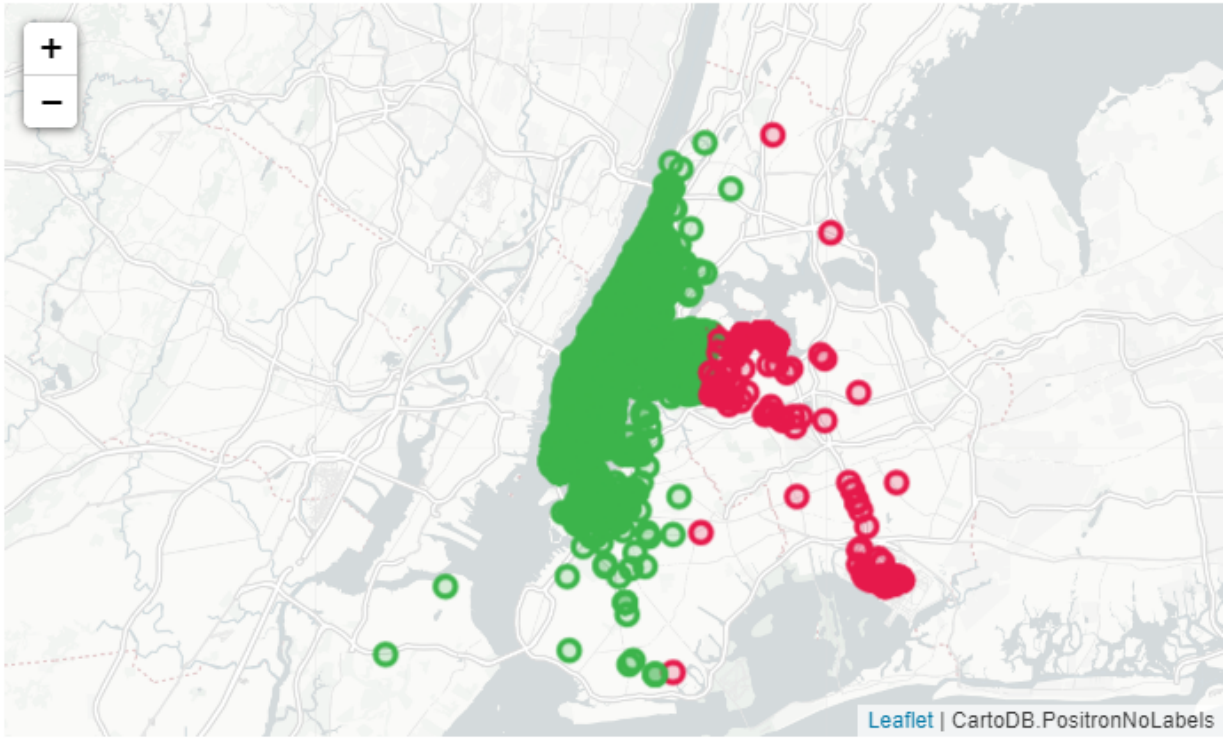

Figure 19. Results of K-means clustering when K = 2

However, the generated clustering result (as shown in Fig. 19) is insufficient to reflect our objective of meeting specific areas where demand is high, with noticeable noise data clustered as main clusters and the cluster size being too big to make any significant interpretations of zones of higher taxi demand.

b. DBSCAN

Parameters for DBSCAN have been decided to be epsilon=0.01, and samples = 30. In our case, this clustering algorithm is repeated multiple times and 30 samples are chosen as we feel that it is an appropriate sum to reflect sufficient demand for taxis within an area. The result (Fig. 20) shows that there are three clusters in the scale of NYC, including the biggest one covering the whole Manhattan and part of Brooklyn and Queens and other two small ones in Queens. Grey circles represent noise data which are not regarded as a cluster under DBSCAN, and therefore suggesting efficiency of the algorithm in ignoring noise data.

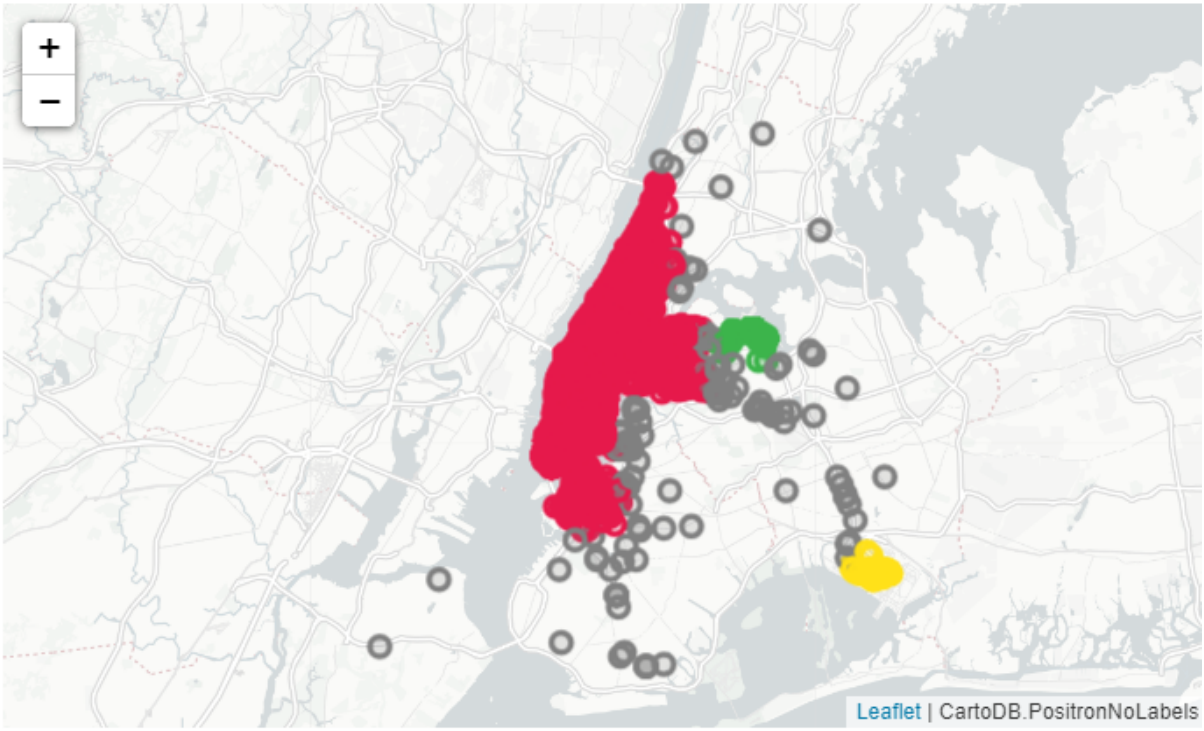

Figure 20. Results of DBSCAN clustering

c. HDBSCAN

Our choice of parameters is inherited from our DBSCAN analysis earlier to reflect the intrinsic differences in clustering between the two density-based clustering algorithms. After attempting multiple times, we decided on the parameters where epsilon=0.01, minimum samples = 30, and minimum cluster size=60.

The clustering result by using HDBSCAN shows five apparent clusters in the scale of NYC, which seems to be more reasonable than the ones using DBSCAN (Fig. 21).

Cluster in Manhattan is further separated from other clusters in Brooklyn and Queens. Four of the five clusters are close to each other around the area of Manhattan, and one cluster is in the area of JFK airport in Queens. Similarly, HDBSCAN parallels results with DBSCAN in its efficiency in filtering out noise data under its clustering.

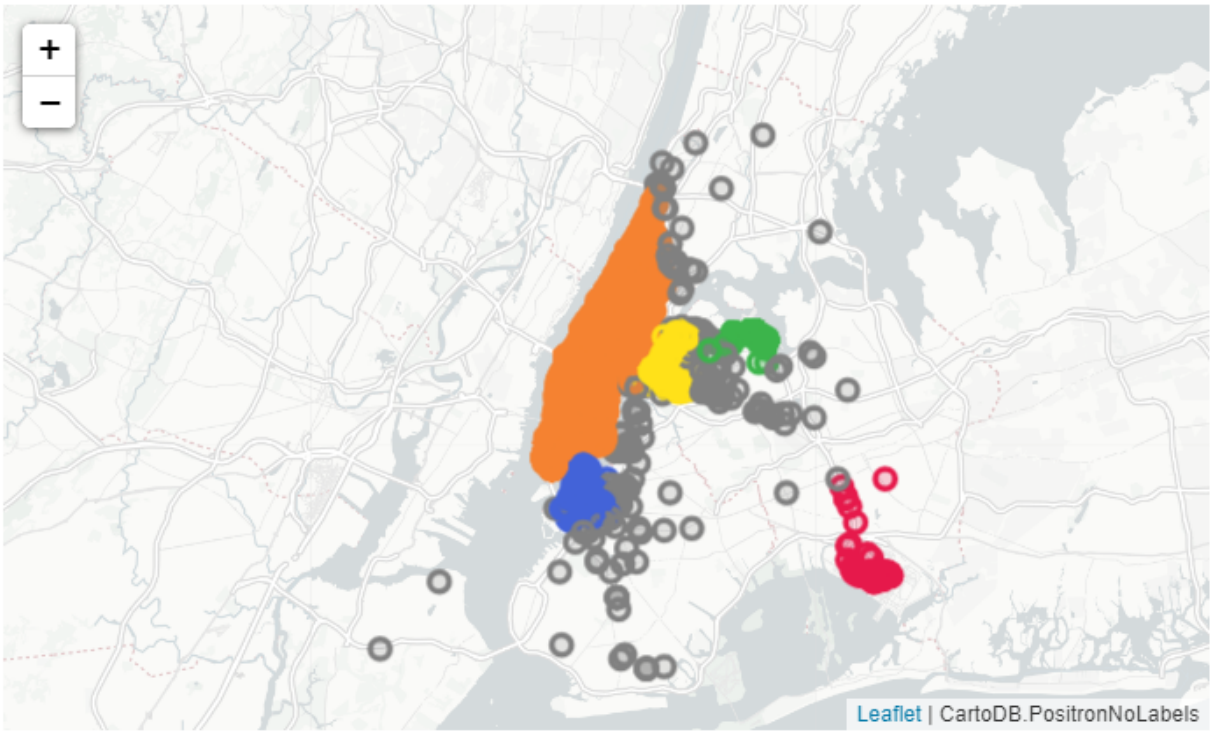

Figure 21. Results of HDBSCAN clustering

Discussion

Spatiotemporal attributes

As noted in earlier sections, there are distinctive spatial and temporal patterns across Yellow taxi trip records in NYC.

Traffic is characterized by massive flows within Manhattan, and between Manhattan and Queens and Brooklyn. This affirms the findings of NYC TLC and Department of Transport (2019), where most taxi traffic accounts for large presence within Manhattan core. On a zonal scale, Manhattan, Astoria, John F Kennedy Airport, East Elmhurst and Downtown Brooklyn are regions where pickups are highest.

Too, our study also shows that most trips (97.1%) are made within 24 miles with most passengers taking the taxi alone. Such characteristics of ridership suggests that passenger capacity within may not be fully optimized. Alisoltani et al.'s (2021) study shows that dynamic ride-sharing can ease stress on network traffic in densely populated cities when accumulated demand for taxis is high. More, therefore, could be looked to determine the possibility of ridesharing as a viable solution to address the problem of inadequate supply to meet taxi demand and traffic congestion within NYC.

In particular, short and long-distance trips exhibit different temporality and spatial characteristics. Temporal peaks for hailing cabs and drop-offs differ, where more peak periods are noted for long distance trips. Whilst spatial attributes of highest pickups and drop-offs may differ according to the choice of temporal scale, our comparison against the land-use map of New York City points towards the possibility of commercial, entertainment zones and airports being areas where pickups and drop-offs seem to be high. More could be done in future research to study its correlation.

Cluster Analysis

Subjective evaluation of the clustering results generated shows that HDBSCAN could best fit our purpose of identifying zones of high taxi demand at the spatial scale of a city. In short, comparison of our clustering results reflects that K-means clustering method fails

to detect noise data and reflect clusters with arbitrary shape as it produces clusters which are more spherical and Gaussian ball in nature. DBSCAN, on the other hand, creates clusters even when density is not particularly significant. Visual interpretation between the few clustering algorithms shows that HDBSCAN is perhaps more suited for our research objective with its capability of focusing clustering on high density areas.

In the following section, we will: (a) critique on different clustering methods in relation to our context, (b) discuss temporal considerations for analysing clusters within NYC; and (c) include computational efficiency considerations within our evaluation of different clustering methods.

a. Critique on Clustering Methods adopted

A method which 'uses iterative reallocation to improve clustering quality from an initial solution', K-means has a proclivity to find clusters which are 'spherical shape and similar in size' and are 'more useful for applications like facility allocation' where 'objective is to … minimise the sum of distances from the data objects to their cluster centers' (Han et al., 2001). Within the context of our study, point data seems to be arranged in an arbitrary shape, outlining the entirety of Manhattan rather than spherical shape which K-means adopts. This therefore suggests that such implications of the K-means clustering method may not necessary be appropriate for the distribution of points in NYC.

Despite such inability in classifying clusters of arbitrary shapes, K-means can be used as a preliminary form of analysis to generate number of dimensions upon which other clustering methods can be used subsequently. Alisoltani et al. (2020) used modified K-means clustering method to identify two main clusters before applying Dynamic Taxi Dispatching Algorithm (DTaD) within each main dimension/ cluster. Other uses include using K-means to cluster according to features which influence individual driver's profitability (e.g. driving distance, duration and income) and using map matching method and grid cell-based density method to group trips in Wuhan, China, with DBSCAN being used at the end to study spatiotemporal patterns of long stopping spots for taxis (Naji et al., 2017). K-means may also be used as a complementary approach for the use of

checkerboard clustering where they can regroup and refine regular spatiotemporal clusters and interaction patterns of taxi data (Liu et al., 2021).

DBSCAN marks a point of departure from the aforementioned partitioning algorithm, where the density-based algorithm uses 'density of data points with a region to discover clusters' (Han et al., 2001). Han et al. (2001) note that despite its affordance in identifying clusters of arbitrary shapes in noisy data (which is characteristic of the shape of data points in NYC), it is sensitive to the input parameters c and M in Pts. Our attempts at determining the most appropriate parameters have pointed at the need for us to run the algorithm multiple times with different permutations of parameters. In big datasets such as NYC taxi data, such trial-and-error approach may represent significant drop in efficiency with such high dimension clustering. However, applications of DBSCAN can be seen in Tang et al.'s study of clustering taxi pick-up and drop-off points and identification of urban hotspots of taxi demand due to its aforementioned ability in grouping data into arbitrary shape and sizes (as quoted Li et. al. 2021).

Conversely, HDBSCAN has been posited to be ideal for exploratory data analysis given its characteristics of being 'a density based algorithm with a small number of intuitive parameters and few assumptions about data distribution' (McInnes & Healy, 2017). An extension built upon theoretical development of DBSCAN, this algorithm inherits the benefits of DBSCAN and Hierarchal Clustering, thereby being ideal in scenarios of discovering spatial clusters with arbitrary shapes with efficiency, filtering out noise data at higher velocities, and handling clusters which vary in densities (Li et al., 2021). Such affordance, therefore, is noted by Li et al. (2021) to be well fitted for large quantities of GPS trajectory points within context of differing shape, sizes and numbers of urban hotspots.

b. Temporal scale considerations for clustering

Determining areas where taxi demand is high will require due care in choice of the temporal scale. While our original intention was to use a month's worth of data to identify significant clusters, our preliminary exploration on Kepler has proved that the magnitude

of pickup and drop-off points within a month may prove to be difficult in discerning perceptible clusters of high demand.

Scaling down of temporal scale to the hour with the highest demand within the day of highest number of trip records of the month has shown to be effective in elucidating the nuanced differences in clustering methods.

c. Computational Efficiency considerations for clustering

Computational complexity and efficiency, too, should be brought into consideration of weighing the use of appropriate clustering algorithm for NYC.

K-means utilizes O (n. K. T) (where n is the number of samples, K is the K value chosen and T is the number of iterations) (Scikitlearn, n.d.) and its computing such equation could be significantly faster than DBSCAN and HDBSCAN.

DBSCAN, on the other hand, is not memory efficient because it creates full pair-wise similarity matrix which will consume n^2. HDBSCAN, too, uses $O(N^2)$, but can be made more efficient with an accelerated algorithm proposed by McInnes & Healy (2017), where they've used Dual Tree Boruka for Euclidean Minimum Spending Trees to revise over the use of $O(N^2)$ to cut computational time.

With accelerated algorithms utilized in Python packages ScikitLearn and HDBSCAN, we noticed no significant differences in computational time required to generate the three clustering algorithms.

As such, while originally proposed complexity of algorithms may hinder computational time especially for large datasets, our experience with the variations and modifications proposed (within Python packages ScikitLearn and HDBSCAN) have proved to be computationally efficient for our temporal scale.

Further Studies

Future studies which strive to analyse the dynamics of taxi within NYC could consider extending work within several directions.

We posit that future attempts at clustering may consider amalgamating spatiotemporal and other aspects within their work. Li et al. (2021), for instance, used ST-HDBSCAN in

their examination of relationship between interaction between points in within time dimension and finding urban hotspots in relation to their spatiotemporal similarity. Chang et al. (2010), on the other hand, proposes a multi-component clustering model of K-means, agglomerative hierarchical clustering and DBSCAN to rank requests for taxi according to time, location, and weather context (as quoted Naji et al., 2017).

Conversely, algorithmic efficiency in computing spatial clusters may also be considered with the extensive range of real time data provided. We attempted to scale down our study to smaller temporal scale while others have attempted to scale down demand nodes within network into geographically dense clusters or clusters which reflects spatiotemporal dependencies of supply of taxis (Alisoltani et al., 2020).

Alternatively, other directions may include modifying the existing DBSCAN algorithm and innovating a hotspot recommendation model which advises on the best precise hotspots for drivers to find passengers (Mu & Dai, 2019). While prevalent literature has been vast in proposing a myriad of clustering algorithms to examine the issue of taxi traffic for large datasets such as NYC, more attention ought to be diverted too to the comparison of the efficiency and accuracy of such algorithms in identifying demand clusters.

Conclusion

In sum, our attempt at exploring the dataset for NYC Yellow Taxi Data has illustrated distinct spatiotemporal trends with a proclivity towards trips within certain regions over others. Our attempt at exploring unsupervised machine learning methods, particular partitioning and density based clustering methods, has shown that DBSCAN may be more apt at identifying regions of higher demand with its ability to account for varying densities and outliers. Comparison with existing literature has also shown that further forays into utilising cluster analysis to identify zones of high taxi demand may encompass a multi component cluster analysis and due consideration of computational efficiency when handling large datasets.

References

Alisoltani, N., Leclercq, L., & Zargayouna, M. (2021). Can dynamic ride-sharing reduce traffic congestion? *Transportation Research Part B: Methodological*, *145*, 212–246. https://doi.org/10.1016/j.trb.2021.01.004

Alisoltani, N., Zargayouna, M., & Leclercq, L. (2020). A Sequential Clustering Method for the Taxi Dispatching Problem Considering Traffic Dynamics. *IEEE Intelligent Transportation Systems Magazine*. https://doi.org/10.1109/MITS.2020.3014444

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining*, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Han, J., Kamber, M., & Tung, A. (2001). Spatial clustering methods in data mining: A survey. *Data Mining and Knowledge Discovery - DATAMINE*.

Li, F., Shi, W., & Zhang, H. (2021). A two-phase clustering approach for urban hotspot detection with spatiotemporal and network constraints. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *PP*, 1–1. https://doi.org/10.1109/JSTARS.2021.3068308

Liu, Q., Zheng, X.-Q., Stanley, H., Xiao, F., & Liu, W. (2021). A Spatio-Temporal Co-Clustering Framework for Discovering Mobility Patterns: A Study of Manhattan Taxi Data. *IEEE Access*, *9*, 34338–34351. https://doi.org/10.1109/ACCESS.2021.3052795

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297.

Mason, H., & Wiggins, C. (2010). *A Taxonomy of Data Science*. http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

McInnes, L., & Healy, J. (2017). Accelerated Hierarchical Density Clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42. https://doi.org/10.1109/ICDMW.2017.12

McInnes, L., Healy, J., & Astels, S. (2016). *How HDBSCAN Works—Hdbscan 0.8.1 documentation*. The Hdbscan Clustering Library. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Melvin, R. L., Xiao, J., Godwin, R. C., Berenhaut, K. S., & Salsbury, F. R. (2018). Visualizing correlated motion with HDBSCAN clustering. *Protein Science*, *27*(1), 62–75. https://doi.org/10.1002/pro.3268

Mu, B., & Dai, M. (2019). Recommend Taxi Pick-up Hotspots Based on Density-based Clustering. *2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET)*, 176–181. https://doi.org/10.1109/CCET48361.2019.8989132

Naji, H. A. H., Wu, C., & Zhang, H. (2017). Understanding the Impact of Human Mobility Patterns on Taxi Drivers' Profitability Using Clustering Techniques: A Case Study in Wuhan, China. *Information*, *8*(2), 67. http://dx.doi.org.libproxy1.nus.edu.sg/10.3390/info8020067

New York City Taxi and Limousine Comission and Department of Transport. (2019). *Improving Efficiency and Managing Growth in New York's For-Hire Vehicle Sector* (p. 35) [Technical Report]. New York City Taxi and Limousine Comission and Department of Transport. https://www1.nyc.gov/assets/tlc/downloads/pdf/fhv_congestion_study_report.pdf

Safikhani, A., Kamga, C., Mudigonda, S., Faghih, S. S., & Moghimi, B. (2020). Spatio-temporal modeling of yellow taxi demands in New York City using generalized STAR

models. *International Journal of Forecasting*, *36*(3), 1138–1148. https://doi.org/10.1016/j.ijforecast.2018.10.001

Scikitlearn. (n.d.). *sklearn.cluster.KMeans—Scikit-learn 0.24.1 documentation*. Retrieved 21 April 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Tang, J., Bi, W., Liu, F., & Zhang, W. (2021). Exploring urban travel patterns using density-based clustering with multi-attributes from large-scaled vehicle trajectories. *Physica A: Statistical Mechanics and Its Applications*, *561*, 125301. https://doi.org/10.1016/j.physa.2020.125301

Techspo. (n.d.). *New York City Boroughs*. Retrieved 18 April 2020, from https://techsponyc.com/new-york-city-boroughs/

Wang, F., Franco-Penya, H.-H., Kelleher, J., Pugh, J., & Ross, R. (2017, July 20). *An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity*. https://doi.org/10.1007/978-3-319-62416-7_21

Zhang, T., Sun, L., Yao, L., & Rong, J. (2017). Impact Analysis of Land Use on Traffic Congestion Using Real-Time Traffic and POI. *Journal of Advanced Transportation*, *2017*, e7164790. https://doi.org/10.1155/2017/7164790