# Hotspot Pickup and Dropoff Region Analysis on NYC Taxi Dataset

CMPUT 697 Project Report

Arif Hasnat
Department of Computing Science
University of Alberta
Email: hasnat@ualberta.ca

**Abstract**

Transportation is one of the most vital services in large cities. Taxis play an important role as a transportation alternative in these large cities. It is very important to analyze the hotspot regions for pickup and dropoff in order to provide better taxi services and to integrate it effectively and efficiently in transportation system. In this project, we analyzed the NYC Taxi dataset provided by the city of New York to find hotspot pickup and dropoff locations where trips are more dense and their relationship with the different periods of weekday and weekend using density based clustering algorithm HDBSCAN.

## 1 Introduction

Transportation is one of the most vital services in large cities. Taxis play an important role as a transportation alternative in these large cities. But there are many misconceptions in TLC (Taxi and Limousine Commission) of the New York city, that how the taxi services should be conducted in the city that too based on certain assumptions. Some of these assumptions are about most pickups, time, duration, distance, airport timings etc. In order to plan effectively to meet the transportation demand of people and to control the traffic flow efficiently in large cities like New York, it is very important to analyze the information available on taxi service to have better understanding. The NYC Taxi dataset [1] provides the relating information such as where taxis are used, when taxis are used and factors which tend public to use taxi as divergent to other modes of transportation.

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

The goal of this project is to analyze this dataset to find hotspot pickup and dropoff locations where trips are more dense and their relationship with the different periods of weekday and weekend using density based unsupervised clustering algorithm, HDBSCAN[3]. With the knowledge about these regions and their relationship with time, planning and distributing taxi service throughout the city will be more effective.

To achieve the goal, we analyzed dense regions for pickup and dropoff of yellow taxi trips in the city of New York with respect to the different time periods of weekday and weekend. We divided trip records of New York city Yellow taxi from September 2015 to weekday trips, weekend trips and trips in five different periods of the day. Then we clustered those trips using HDBSCAN. After that, we used Silhouette Coefficient, Calinski-Harabasz and Davies-Bouldin Index to evaluate our clustering results. Then, we compared the clustering results with DBSCAN[5] and OPTICS[1]. Finally, We compared our results with trip records from different months to show that the results are consistent throughout the years.

## 2 Related Work

The data made publicly available by the New York City (NYC) Taxi and Limousine Commission (TLC)[1] has been the subject of a number of studies.

Patel and Chandan [8] focused on the technical solutions needed to work with the NYC taxi data. Since 2014, there were around 180 million taxi rides, and the authors saw the need for the use of big data tools to analyze the data in a reasonable amount of time. Using MapReduce programming, they conducted analysis to recommend the top driver based on the most distance travelled, most fare collected, most time travelled and most efficient driver. They also appraised the most pickup locations and same for the dropoff locations expending pickup latitude, pickup longitude, dropoff latitude and dropoff longitude. Furthermore, they used Hive to quantify and analyse the total pickups and dropoffs by time of day based on location. Finally, they analysed the fare to get the drivers both gross and net revenue with the help of another Hadoop ecosystem technology called as Pig. The authors were mostly concerned with evaluating the technical aspects of working with such a large amount of data efficiently. Our work is partially related to them as we also analyzed pickup and dropoff locations but using density based clustering algorithms.

Xiong et al. [12] presented an analysis on the taxi trajectory data in Wuhan, China. They clustered the extracted taxi passenger's pickup locations to produce passenger hotspots using DBSCAN clustering algorithm. Then they visualized the dynamic of taxi moving trajectory using interactive animation. In our project, we analyzed the hotspots in New York city using HDBSCAN clustering algorithm.

Ibrahim and Shafiq [6] focused on a data-driven approach to enhance the traffic in Porto city

by delivering a set of recommendations based on trajectory analysis. They extracted insights on a high granularity level where taxi trips are processed based on the day and hour they started. The authors applied HDBSCAN and Random Swap (RS) clustering for both origin and destination points and visualized the taxi trips heatmap per weekday and hour of the day in Porto city to provide useful guidelines for taxi drivers, passengers, and transportation authorities. Our work shares same clustering algorithm HDBSCAN with them.

Stoyanovich et al. [10] analyzed TLC yellow cab data spanning July 2015 through June 2016, to identify hotspots - the most popular origins and destinations. Theses regions point to the lack of convenient public transportation options, and popular routes that can be used to motivate ride-sharing solutions or addition of a bus route. They found some interesting facts. Such as La Guardia participates in many more taxi trips than JFK, despite being a much smaller airport. Another insight is that the single most frequent route is between Penn Station and Grand Central — two train station in Midtown Manhattan. Both fndings can be explained by the lack of convenient public transportation options that connect these out-of-town transportation hubs: Unlike JFK, La Guardia is not reachable by subway from Midtown Manhattan. It takes 2 trains (with a connection at the very busy Times Square station) and 20 minutes to travel between Grand Central and Penn Station by subway. They used a system called Portal, which implements efficient representations and principled analysis methods for evolving graphs in their analysis. Our work is most similar to them. But their approach is different from us as we used clustering algorithms to find hotspots.

# 3 Methodology

## 3.1 Dataset

Taxi and Limousine Commission (TLC) of the city of New York provides trip records for the yellow and green taxi and the For-Hire Vehicle (FHV) operating in the city. The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts[1]. Each data set display these information for a single month. The locations for pickup and dropoff is the target feature for our analysis. The Travel Time Data Collection Handbook suggests that the four time elements for consideration can be month, day of week, day type, and time of day [11]. Therefore, we also need to consider the times for pickup and dropoff to analyze the dataset according to those time elements.

The yellow taxi trip records comprised trips occurred from 2009 to 2019. But due to privacy issues, trip records since July 2016 use Location ID, each corresponding to a bounded zone of neighborhood, instead of longitudes and latitudes. Because of that our analysis on geographical

coordinates (longitude, latitude) is based on date before July 2016. According to preliminary exploration of the data, the distributions of the trip locations and time are similar in each month. Therefore, in order to improve computation efficiency, we primarily analyzed the trip records of yellow taxi from September 2015 and then showed that the results are consistent for other months too.

The following figure shows a snapshot of the trip records with the relevant attributes to our analysis from September 2015:

| tpep_pickup_datetime | tpep_dropoff_datetime | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude |
|---|---|---|---|---|---|
| 2015-09-01 0:05 | 2015-09-01 0:31 | -73.79135132 | 40.64669037 | -73.85743713 | 40.84826279 |
| 2015-09-01 0:05 | 2015-09-01 0:07 | -73.97893524 | 40.75285339 | -73.9860611 | 40.7553978 |
| 2015-09-01 0:05 | 2015-09-01 0:16 | -73.9908905 | 40.72397232 | -74.00955963 | 40.72891617 |
| 2015-09-01 0:05 | 2015-09-01 0:05 | -73.93265533 | 40.80376816 | 0 | 0 |
| 2015-09-01 0:05 | 2015-09-01 0:30 | -73.98777771 | 40.73819351 | -73.94475555 | 40.82816696 |
| 2015-09-01 0:05 | 2015-09-01 0:12 | -73.99025726 | 40.73728943 | -73.99508667 | 40.74485016 |
| 2015-09-01 0:05 | 2015-09-01 0:08 | -73.94365692 | 40.82059479 | -73.93763733 | 40.82869339 |
| 2015-09-01 0:05 | 2015-09-01 0:10 | -74.00408936 | 40.75182724 | -74.00380707 | 40.74201965 |
| 2015-09-01 0:05 | 2015-09-01 0:32 | -73.98703003 | 40.76649094 | -73.91083527 | 40.77112961 |
| 2015-09-01 0:05 | 2015-09-01 0:19 | -73.99694061 | 40.72531509 | -73.95201874 | 40.69283676 |
| 2015-09-01 0:05 | 2015-09-01 0:22 | -73.98669434 | 40.73014832 | -73.9811554 | 40.67548752 |
| 2015-09-01 0:05 | 2015-09-01 0:22 | -73.99407196 | 40.73230743 | -73.98139191 | 40.76856232 |
| 2015-09-01 0:06 | 2015-09-01 0:08 | -73.97113037 | 40.7553215 | -73.96334076 | 40.76594162 |
| 2015-09-01 0:06 | 2015-09-01 0:15 | -73.98622894 | 40.73055267 | -74.00285339 | 40.73085022 |
| 2015-09-01 0:06 | 2015-09-01 0:10 | -73.98223114 | 40.77463913 | -73.96777344 | 40.80054092 |
| 2015-09-01 0:06 | 2015-09-01 0:15 | -73.86343384 | 40.76905441 | -73.92523956 | 40.75936508 |

Figure 1: Sample data with relevant attributes only

## 3.2 Data Preprocessing

We performed the following steps to preprocess the data provided by the TLC of New York city:

- Read monthly trip records from CSV file. For example, September 2015 trip records from "yellow_tripdata_2015-09.csv" file.
- Select only a subset of the attributes that are relevant to the analysis. The relevant attributes are pickup_datetime, dropoff_datetime, pickup_longitude, pickup_latitude, dropoff_longitude and dropoff_latitude.
- Parse pickup_datetime and dropoff_datetime which are strings of "%Y-%m-%d %H:%M" format in the given data.
- Remove all the trip records containing 0 as value for any of the selected attributes.

- Impose boundary on the longitude and latitude of the trip records where longitude is greater than -74.257159 and less than -73.699215, and latitude is greater than 40.495992 and less than 40.915568. Remove all the trip records that are outside of this boundary.
- Select trip records of a random weekday. Weekdays are Monday, Tuesday, Wednesday, Thursday and Friday.
- Select trip records of a random weekend day. Weekend days are Saturday and Sunday.
- Divide weekday trip records and weekend day trip records into 5 different periods of the day. The periods are 0:00 to 6:00, 6:00 to 10:00, 10:00 to 15:00, 15:00 to 19:00 and 19:00 to 24:00 clock time of a day in 24-hour clock notation.

As we can see from the figure 1, there are some trip records that contain 0 as value for some of the attributes which does not make any sense. Therefore, we removed these type of trip records from our consideration. Also, some trip records contain unrealistic value for longitude and latitude in New York city perspective. We are interested in dense regions rather than these outliers or noisy trip records for our analysis. That's why, we imposed boundary on the longitude and latitude according to New York City Borough Boundary[2] which is "West -74.257159, East -73.699215, North 40.915568 and South 40.495992".

## 3.3 Clustering

We explored three different density based clustering algorithms- DBSCAN, HDBSCAN and OPTICS. We used automatic cluster extraction methods for HDBSCAN and OPTICS. We found the clustering results similar for all these algorithms. But DBSCAN and HDBSCAN is much faster than OPTICS. Also, it is very difficult to set parameters for DBSCAN perfectly for all the trip records of different time periods.That's why, we mainly used HDBSCAN[3] as clustering algorithm in our analysis. Clustering of the preprocessed trip records contains the following steps:

- Choose suitable value for parameters- min_samples and min_cluster_size of HDBSCAN clustering algorithm.
- Use columns- pickup_longitude, pickup_latitude, dropoff_longitude and dropoff_latitude as the features for the trip data points and Euclidean distance metric to measure similarities between data points.

[2]https://www1.nyc.gov/assets/planning/download/pdf/data-maps/open-data/nybb_metadata.pdf?ver=19d
[3] https://pypi.org/project/hdbscan/

- Apply HDBSCAN on weekday data, weekend day data and data of different periods of the weekday and the weekend day
- From each clustering result, select top 7 clusters according to the descending order of the number of trips they contain.

## 3.4 Visualization

We drew locations (longitude, latitude) on New York city map using customized style and Scatter Mapbox[4] for better visualization of data and results. We performed following steps to visualize the trip records and the clustering results:

- Plot pickup and dropoff locations of the trip records on map
- Plot clustering results using only dropoff locations of trip records
- Plot pickup and dropoff locations of selected top 7 clusters from the clustering results on map

.

## 3.5 Validation

As the ground truth clusters are not known, we needed to use External validation indices to evaluate our clustering results. DBCV[7] provides better validation of clustering result from density based clustering algorithms. As we clustered about 100k to 300k trips for our analysis, DBCV[5] continued to run for hours to validate the clustering results. Therefore, we use Silhouette Coefficient[9], Calinski-Harabasz Index[2] and Davies-Bouldin Index[4]. They provide better run time for such large number of data. A brief description of the indices we used as follows:

- **Approximate Silhouette Coefficient**: The Silhouette Coefficient[6] for a set of samples is given as the mean of the Silhouette Coefficient for each sample. To improve performance and decrease the run time, we calculated an approximate Silhouette score by iterating the Silhouette Coefficient calculation 5 times using random samples of size 1000 and then taking the average of those scores. A higher Silhouette Coefficient score relates to a model with better defined clusters. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.

[4]https://www.mapbox.com/maps/
[5] https://github.com/christopherjenness/DBCV
[6] https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

- **Calinski-Harabasz Index**: The index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared). A higher Calinski-Harabasz score[7] relates to a model with better defined clusters.

- **Davies-Bouldin Index**: This index signifies the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves. A lower Davies-Bouldin index[8] relates to a model with better separation between the clusters. Zero is the lowest possible score. Values closer to zero indicate a better partition.

# 4 Experimental Evaluation

## 4.1 Data Selection

For our analysis, we randomly selected Wednesday (September 02, 2015) as our weekday and Sunday (September 06, 2015) as our weekend day. Then we further divided all the trips in those days into five periods of the day. The following table represents trip distribution in different periods of our selected weekday and weekend day:

| **Weekday Time Periods** | **Number of Trips** |
|---|---|
| 0:00 to 6:00 | 26717 |
| 6:00 to 10:00 | 64259 |
| 10:00 to 15:00 | 89931 |
| 15:00 to 19:00 | 73658 |
| 19:00 to 24:00 | 107977 |
| Total | 362542 |

| **Weekend Day Time Periods** | **Number of Trips** |
|---|---|
| 0:00 to 6:00 | 61810 |
| 6:00 to 10:00 | 22420 |
| 10:00 to 15:00 | 72403 |
| 15:00 to 19:00 | 64139 |
| 19:00 to 24:00 | 75598 |
| Total | 296370 |

Table 1: Weekday and Weekend Day Trip Distribution

## 4.2 Clustering Results

For clustering, we selected min_samples = 7 according to the heuristic (2*d - 1) as we used four features for the data points. Furthermore, We selected min_cluster_size = 100. Because with

[7] https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index
[8] https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index

smaller cluster size, HDBSCAN returns many small sized cluster and we are interested in clusters with more data points. The following two tables represent summary of clustering results and validation scores on weekday and weekend day data:

| Time Periods | Trip Counts | Number of Clusters | Number of Noise points | Approx. Silhouette Score | Calinski-Harabasz Score | Davies-Bouldin Score |
|---|---|---|---|---|---|---|
| Whole Day | 362542 | 43 | 25589 | 0.9370 | 547939163.27 | 0.0371 |
| 0:00 to 6:00 | 26717 | 3 | 287 | 0.9242 | 430598.94 | 0.1837 |
| 6:00 to 10:00 | 64259 | 14 | 7022 | 0.9488 | 49084636.38 | 0.0537 |
| 10:00 to 15:00 | 89931 | 14 | 5112 | 0.9526 | 44132931.59 | 0.0544 |
| 15:00 to 19:00 | 73658 | 13 | 5007 | 0.9471 | 34968014.45 | 0.0562 |
| 19:00 to 24:00 | 107977 | 21 | 15877 | 0.9565 | 115270883.62 | 0.0467 |

Table 2: Weekday Clustering Results and Validation Scores Summary

| Time Periods | Trip Counts | Number of Clusters | Number of Noise points | Approx. Silhouette Score | Calinski-Harabasz Score | Davies-Bouldin Score |
|---|---|---|---|---|---|---|
| Whole Day | 296370 | 32 | 32818 | 0.9402 | 243273771.11 | 0.0466 |
| 0:00 to 6:00 | 61810 | 13 | 11968 | 0.9379 | 21553244.41 | 0.0709 |
| 6:00 to 10:00 | 22420 | 7 | 2989 | 0.9374 | 6803233.38 | 0.0807 |
| 10:00 to 15:00 | 72403 | 14 | 4751 | 0.9423 | 32403097.86 | 0.0616 |
| 15:00 to 19:00 | 64139 | 13 | 6534 | 0.9440 | 29769467.98 | 0.0620 |
| 19:00 to 24:00 | 75598 | 17 | 9157 | 0.9412 | 54848502.71 | 0.0559 |

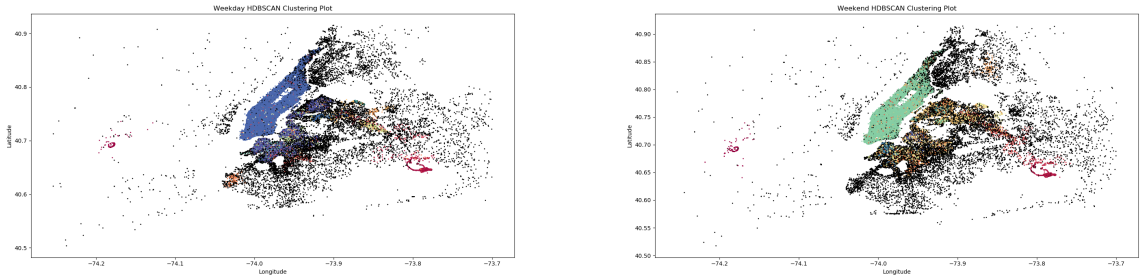
Table 3: Weekend Day Clustering Results and Validation Scores Summary

According to the table 2 and 3, Approximate Silhouette scores for all the clustering results are within the range of 0.92 to 0.96 which are closer to 1 and indicate good clustering. Calinski-Harabasz scores are also very high. Davies-Bouldin scores are within 0.03 to 0.09 which are closer to 0. All these scores indicate good clustering results. These three scores are worst for the clustering results of trips within 0:00 to 6:00 time period of weekday which indicate worst clustering compared to other results. This also can be validated by looking at the clustering result shown in figure 5(a).

In the following sections, we showed clustering results by HDBSCAN and most important

clusters for our analysis from the top 7 selected clusters for different time periods. We also discussed the results in detail. Some clusters like trips between Manhattan and the two airports are common for all the results. Thus, they are not shown in most of the figures.

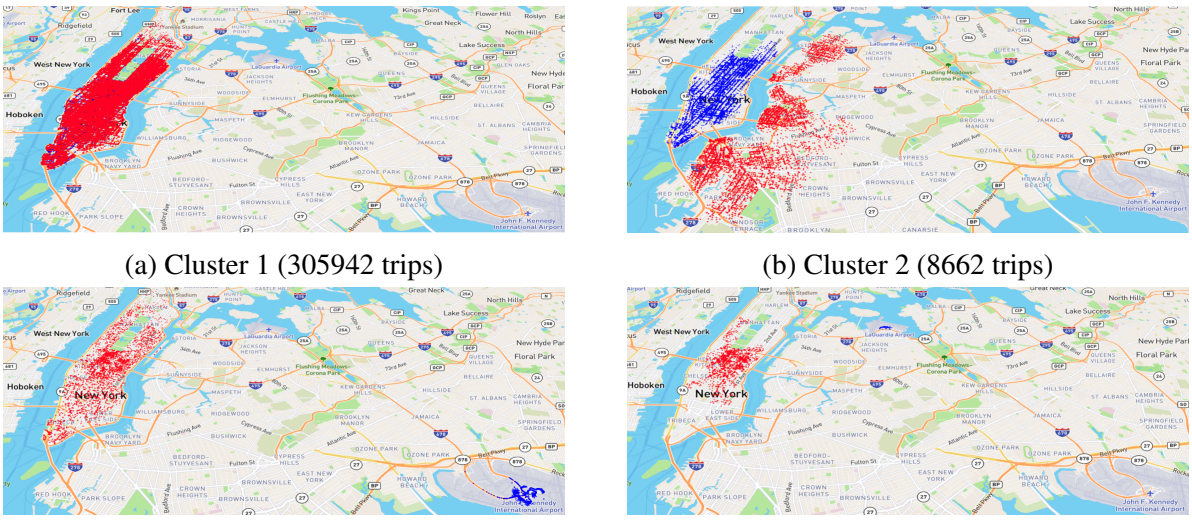
### 4.2.1 Whole Day Data




(a) Weekday (362542 trips)

(b) Weekend (296370 trips)

Figure 2: Clustering plot (only dropoff locations) on whole weekday and weekend data. Black points are noise.




(a) Cluster 1 (305942 trips)

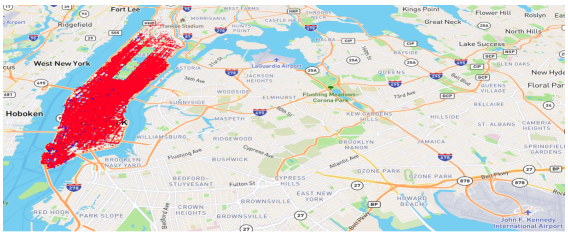
(b) Cluster 2 (8662 trips)

(c) Cluster 3 (4165 trips)

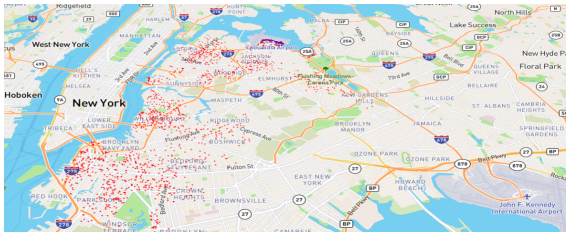
(d) Cluster 5 (2268 trips)

Figure 3: Top clusters on whole weekday data (pickup in blue and dropoff in red)

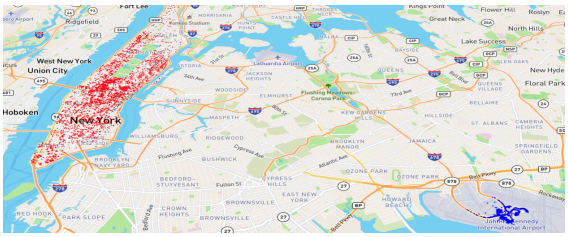
According to figure 3 and 4, pickups and dropoffs between Manhattan and the two airports, JFK and LaGuardia are the common hotspots for both weekday and weekend. But there are a significant number of trips (8662) from Manhattan to some areas of Brooklyn and Queens across the East River bridges on weekday. On the contrary, there are some trips (1668) from LGA Airport to all over Brooklyn and Queens on weekend.



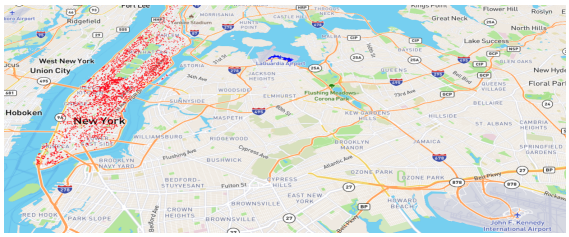
(a) Cluster 1 (236942 trips)



(b) Cluster 7 (1668 trips)



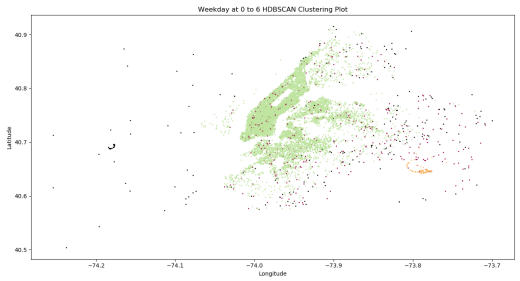
(c) Cluster 2 (3908 trips)



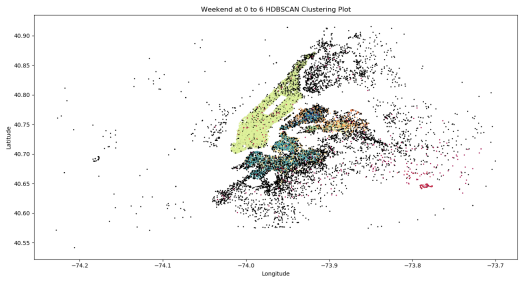
(d) Cluster 3 (3454 trips)

Figure 4: Top clusters on whole weekend data (pickup in blue and dropoff in red)

### 4.2.2 0:00 to 6:00

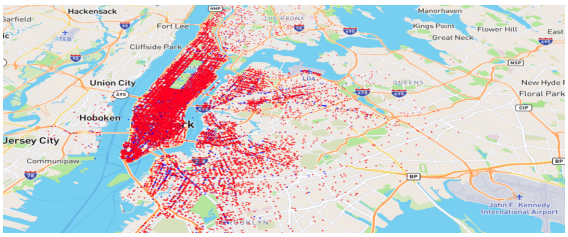



(a) Weekday (26717 trips)

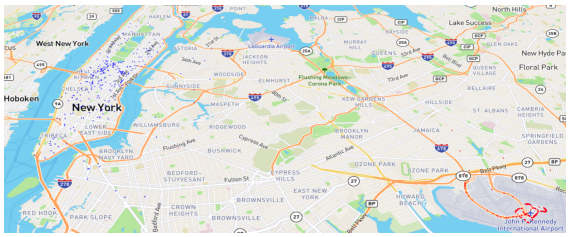



(b) Weekend (61810 trips)

Figure 5: Clustering plot (only dropoff locations) from 0:00 to 6:00 on weekday and weekend

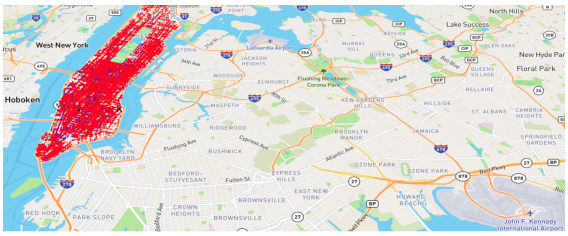


(a) Cluster 1 (25413 trips)

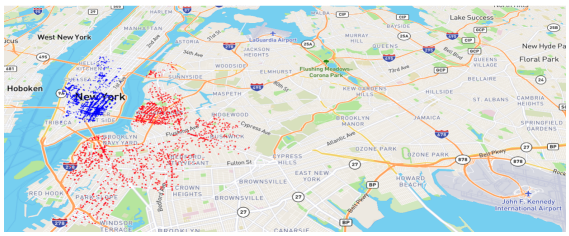


(b) Cluster 3 (335 trips)

Figure 6: Top clusters from 0:00 to 6:00 on weekday (pickup in blue and dropoff in red)




(a) Cluster 1 (43618 trips)

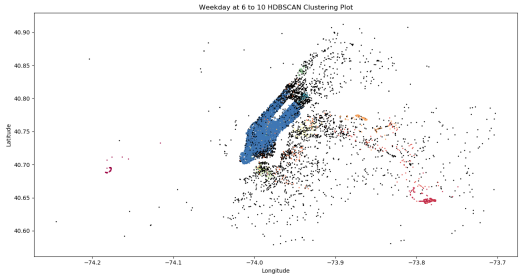



(b) Cluster 2 (2265 trips)

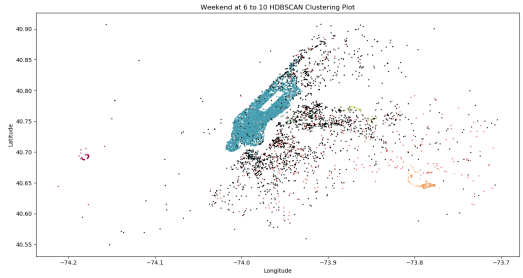
Figure 7: Top clusters from 0:00 to 6:00 on weekend (pickup in blue and dropoff in red)

Table 1 indicates that the number of trips from 0:00 to 6:00 is much less for weekday (26717 trips) compared to weekend (61810 trips). According to figure 6 and 7, most of the trips from 0:00 to 6:00 on weekday is from Manhattan to all over the New York city. There are also some trips from Manhattan to JFK Airport whereas most of the trips on weekend are within Manhattan. Beside that, a significant number of trips are from Manhattan to some areas of Brooklyn and Queens across the East River bridges on weekend 0:00 to 6:00.

### 4.2.3 6:00 to 10:00

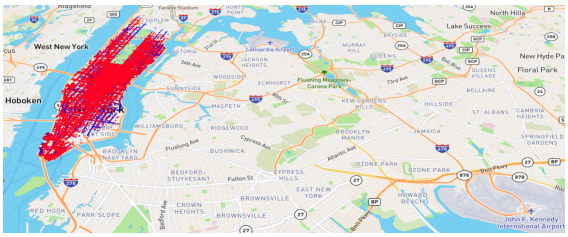



(a) Weekday (64259 trips)

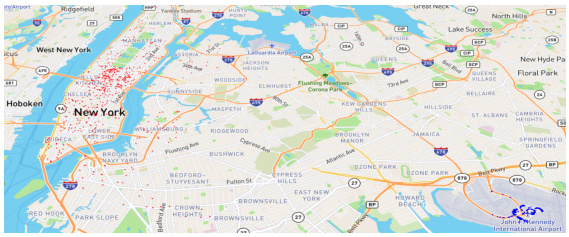



(b) Weekend (22420 trips)

Figure 8: Clustering plot (only dropoff locations) from 6:00 to 10:00 on weekday and weekend

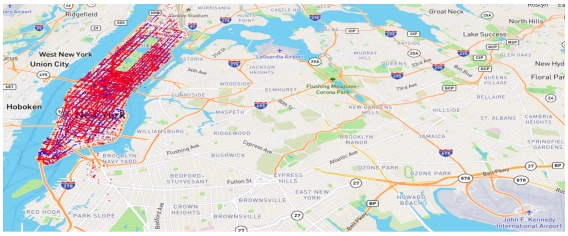



(a) Cluster 1 (52833 trips)




(b) Cluster 3 (674 trips)

Figure 9: Top clusters from 6:00 to 10:00 on weekday (pickup in blue and dropoff in red)

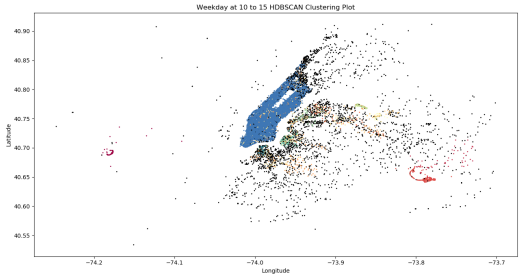


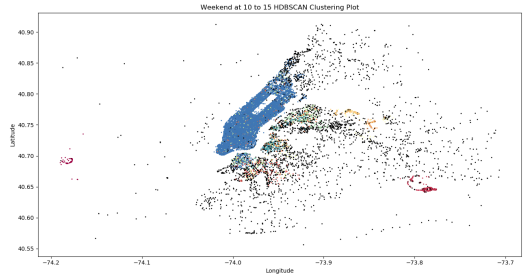

(a) Cluster 1 (16975 trips) (b) Cluster 2 (864 trips)

Figure 10: Top clusters from 6:00 to 10:00 on weekend (pickup in blue and dropoff in red)

Table 1 indicates that the number of trips from 6:00 to 10:00 is much less for weekend (22420 trips) compared to weekday (64259 trips). According to figure 9 and 10, Manhattan is much busier on weekday 6:00 to 10:00 (52833 trips) than on weekend (16975 trips). Also, there are 674 trips from JFK Airport to Manhattan on weekday whereas there are 864 trips from JFK Airport to all over the New York city on weekend.

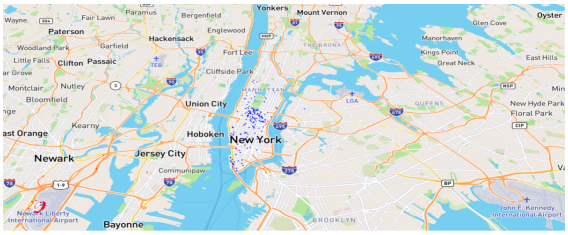
### 4.2.4 10:00 to 15:00

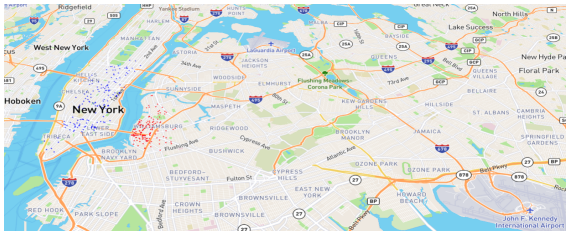



(a) Weekday (89931 trips) (b) Weekend (72403 trips)

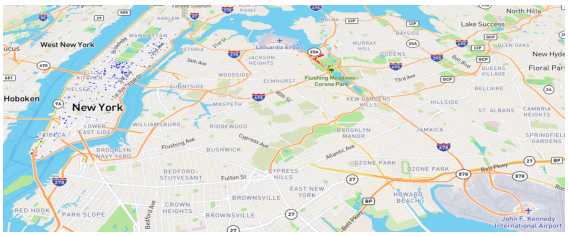
Figure 11: Clustering plot (only dropoff locations) from 10:00 to 15:00 on weekday and weekend



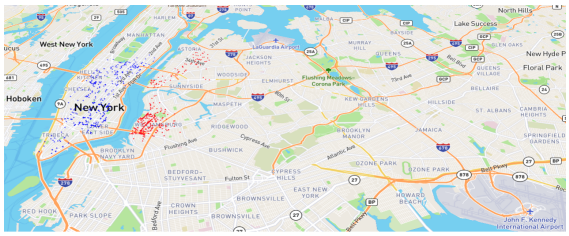




(a) Cluster 6 (184 trips) (b) Cluster 7 (158 trips)

Figure 12: Top clusters from 10:00 to 15:00 on weekday (pickup in blue and dropoff in red)




(a) Cluster 7 (249 trips)

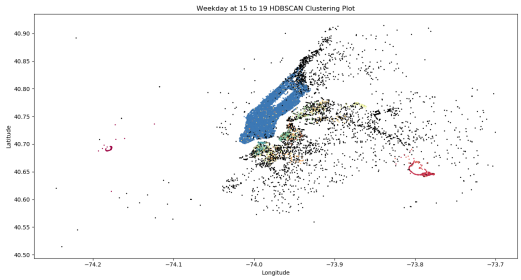



(b) Cluster 6 (402 trips)

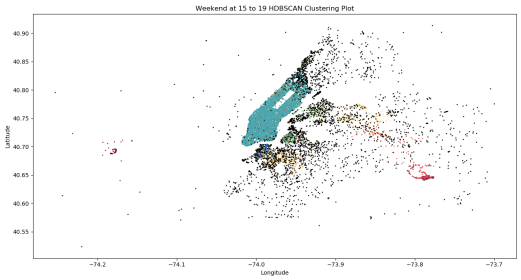
Figure 13: Top clusters from 10:00 to 15:00 on weekend (pickup in blue and dropoff in red)

According to figure 12 and 13, there are some trips from Manhattan to Williamsburg on both weekday and weekend from 10:00 to 15:00. Also, there are some trips from Manhattan to Newark Liberty Airport on weekday whereas there are 249 trips from Manhattan to near Queens Zoo and Queens Museum on weekend.

### 4.2.5 15:00 to 19:00

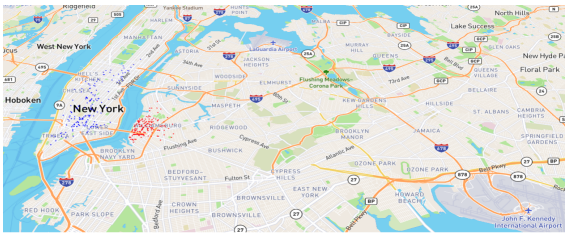



(a) Weekday (73658 trips)

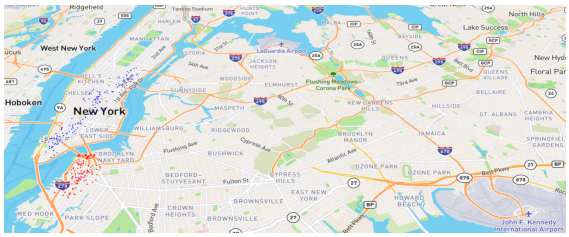



(b) Weekend (64139 trips)

Figure 14: Clustering plot (only dropoff locations) from 15:00 to 19:00 on weekday and weekend

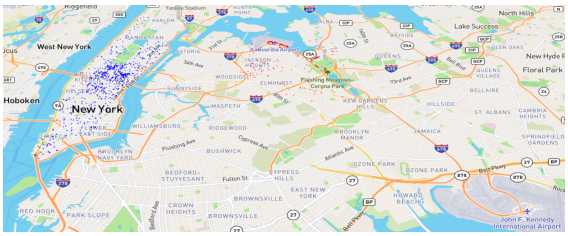



(a) Cluster 6 (173 trips)

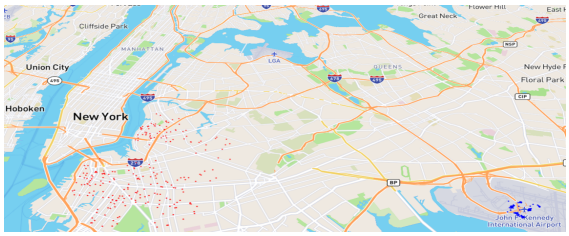



(b) Cluster 7 (151 trips)

Figure 15: Top clusters from 15:00 to 19:00 on weekday (pickup in blue and dropoff in red)




(a) Cluster 4 (770 trips)

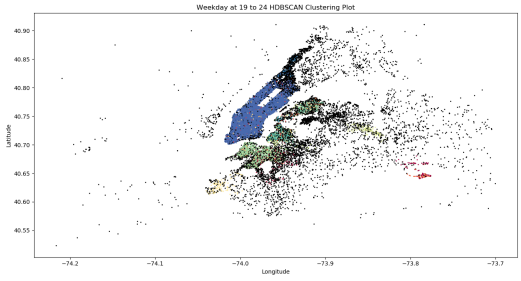



(b) Cluster 6 (261 trips)

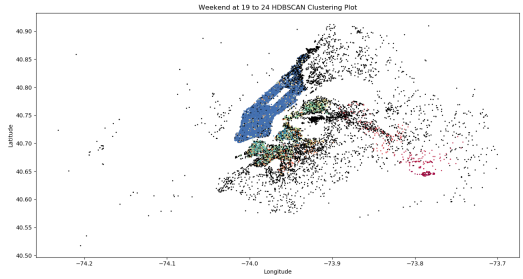
Figure 16: Top clusters from 15:00 to 19:00 on weekend (pickup in blue and dropoff in red)

According to figure 15 and 16, there are some trips from Manhattan to Williamsburg and Brooklyn Heights on weekday 15:00 to 19:00 whereas there are some trips from Manhattan to near Queens Zoo and Museum on weekend. Also, some trips are from JFK airport to Brooklyn on weekend.

### 4.2.6 19:00 to 24:00

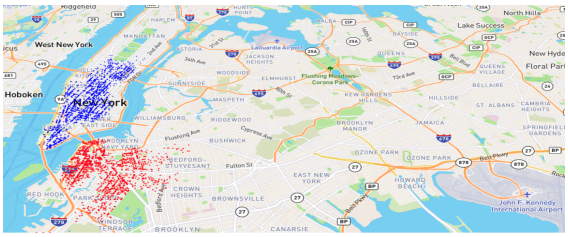



(a) Weekday (107977 trips)

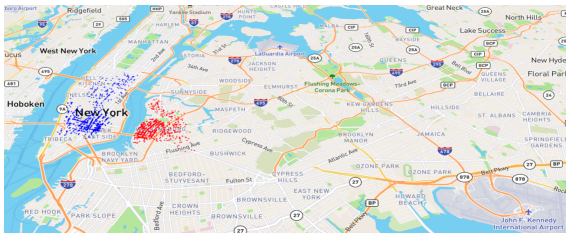



(b) Weekend (75598 trips)

Figure 17: Clustering plot (only dropoff locations) from 19:00 to 24:00 on weekday and weekend

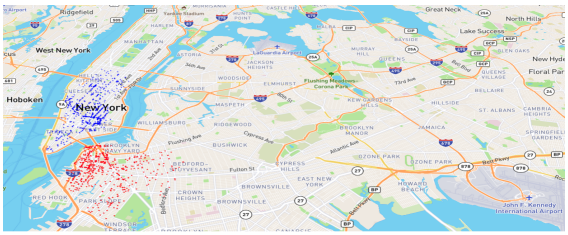



(a) Cluster 2 (2012 trips)

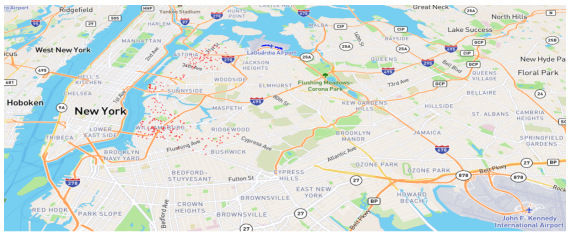



(b) Cluster 4 (864 trips)

Figure 18: Top clusters from 19:00 to 24:00 on weekday (pickup in blue and dropoff in red)




(a) Cluster 4 (724 trips)



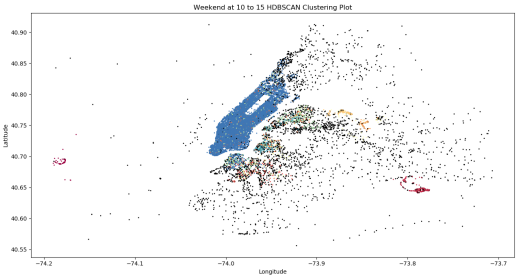

(b) Cluster 7 (271 trips)

Figure 19: Top clusters from 19:00 to 24:00 on weekend (pickup in blue and dropoff in red)

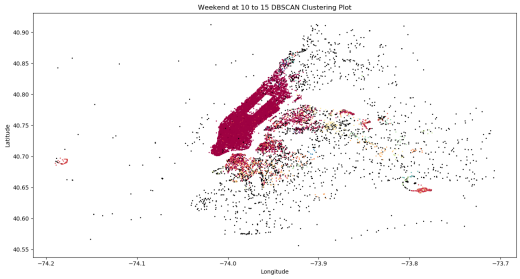
According to figure 18 and 19, there are some trips from Manhattan to Williamsburg and Brooklyn Heights on both weekday and weekend 19:00 to 24:00. But, trips are significantly more on weekday in those areas whereas there are some trips from LGA airport to some areas of Brooklyn and Queens on weekend.

## 4.3 Comparison with DBSCAN and OPTICS

Our clustering results obtained from HDBSCAN are almost same to the clustering results from other density based clustering algorithms such as DBSCAN and OPTICS. We used min_samples = 7 and eps = 0.01 for DBSCAN[9]. For OPTICS[10] with cluster extraction method Xi, we used min_samples = 7, min_cluster_size = 100, max_eps = 10 and xi = 0.05. Distance metric was Euclidean for all the algorithms. The following figures show a comparison among clustering results of trips within 10:00 to 15:00 on weekend by HDBSCAN, DBSCAN and OPTICS.



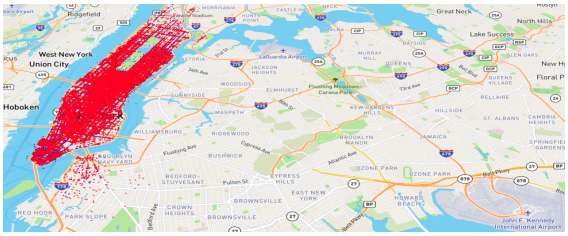

(a) HDBSCAN



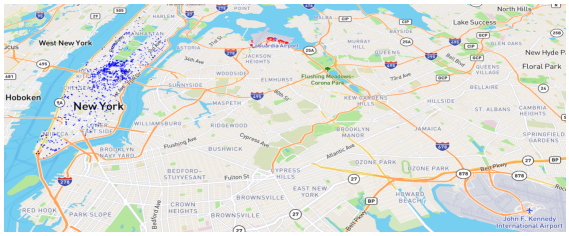

(b) DBSCAN

Figure 20: Clustering plot (only dropoff locations) from 10:00 to 15:00 on weekend. Black points are noise.

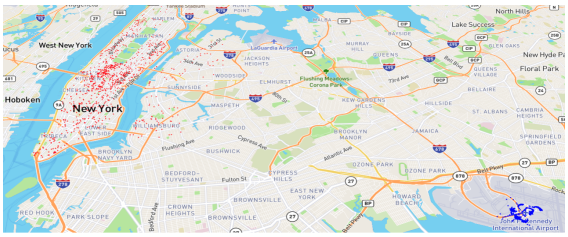
[9]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
[10]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html



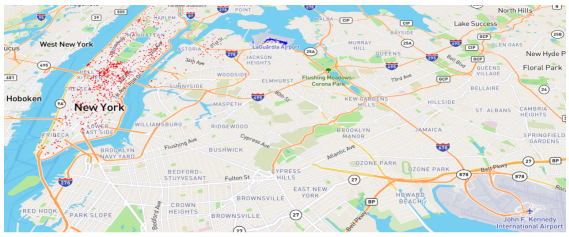
(a) Cluster 1 (61894 trips)



(b) Cluster 2 (1149 trips)

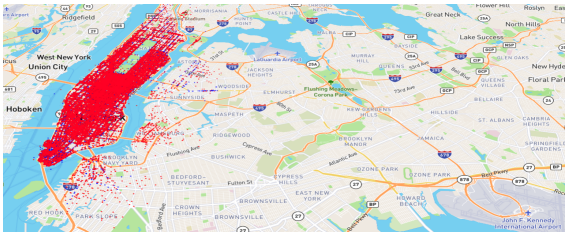


(c) Cluster 3 (1061 trips)

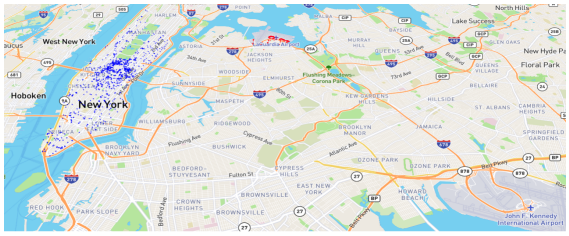


(d) Cluster 4 (1050 trips)

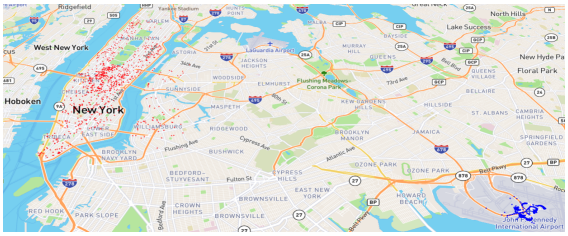
Figure 21: Top clusters using HDBSCAN from 10:00 to 15:00 on weekend (pickup in blue and dropoff in red)



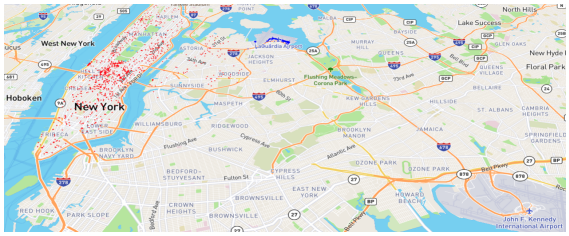
(a) Cluster 1 (63084 trips)



(b) Cluster 2 (1123 trips)



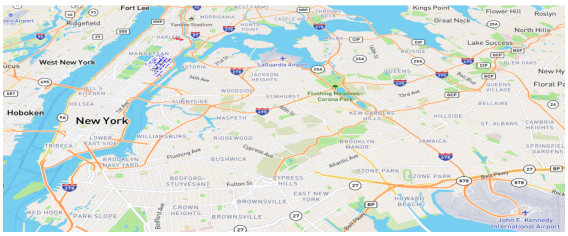
(c) Cluster 4 (1029 trips)



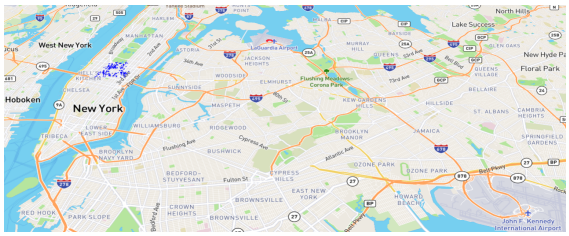
(d) Cluster 3 (1119 trips)

Figure 22: Top clusters using DBSCAN from 10:00 to 15:00 on weekend (pickup in blue and dropoff in red)

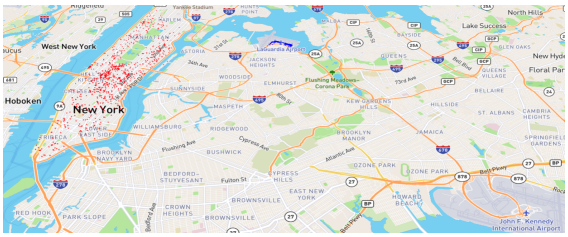
From figure 21, 22 and 23, we can see that top clusters returned by DBSCAN and HDBSCAN are almost same. Clusters returned by these two are also similar to OPTICS except cluster 1. Instead of that big cluster 1 from DBSCAN and HDBSCAN, OPTICS returned some small clusters like cluster 5 for that region. The parameter settings for OPTICS might cause that difference in the clustering result.



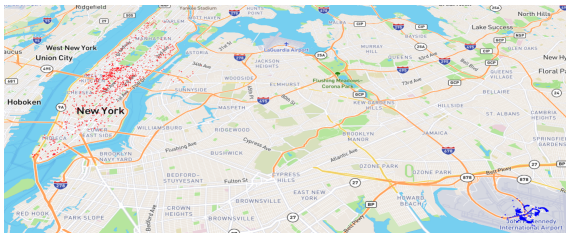
(a) Cluster 5 (145 trips)



(b) Cluster 4 (246 trips)



(c) Cluster 2 (955 trips)

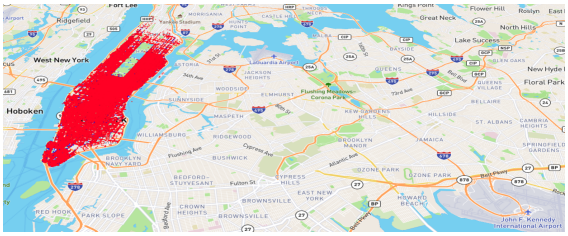


(d) Cluster 1 (966 trips)

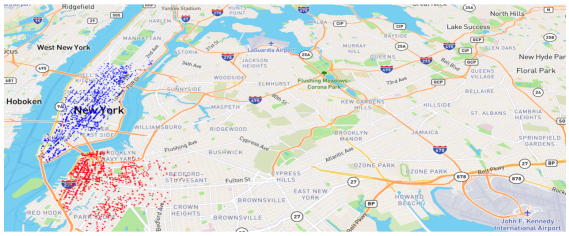
Figure 23: Top clusters using OPTICS from 10:00 to 15:00 on weekend (pickup in blue and dropoff in red)

## 4.4 Comparison with other Data

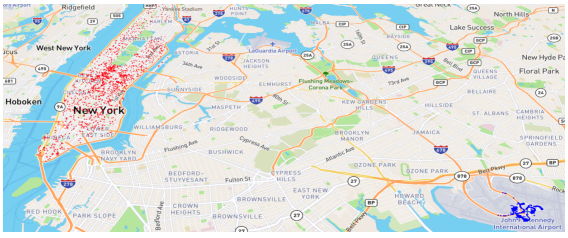
Our obtained clustering results are consistent throughout the months of the years. To show that, we selected Wednesday (January 06, 2016) as our weekday and Sunday (January 03, 2016) as our weekend day. The following two figures show the clusters on January 2016 data which are similar to the clusters presented in section 4.2.1.



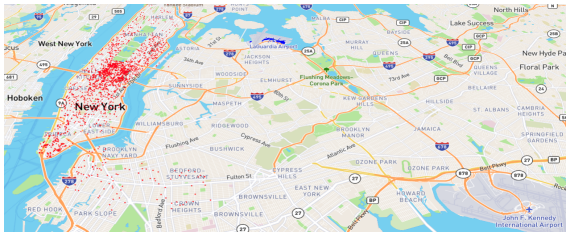
(a) Cluster 1 (130584 trips)



(b) Cluster 4 (1435 trips)

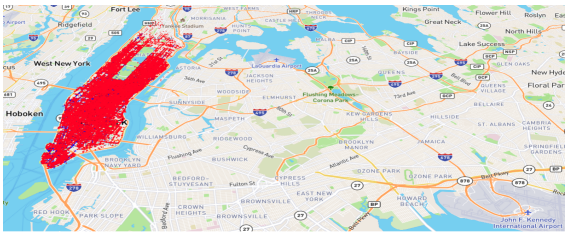


(c) Cluster 3 (1639 trips)

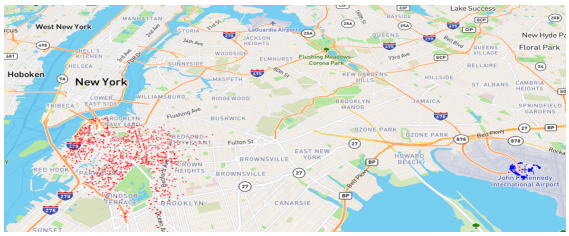


(d) Cluster 2 (2683 trips)

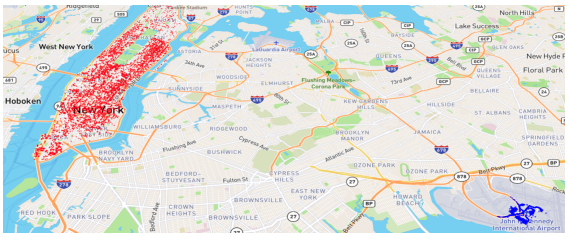
Figure 24: Top clusters on weekday data January 2016 (pickup in blue and dropoff in red)



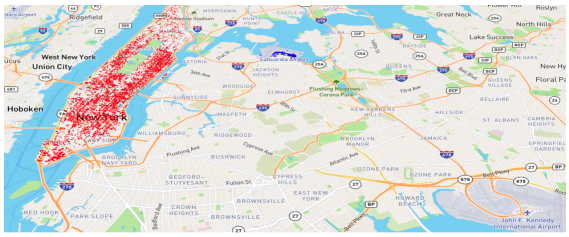
(a) Cluster 1 (234360 trips)



(b) Cluster 7 (1109 trips)



(c) Cluster 3 (5813 trips)



(d) Cluster 2 (7064 trips)

Figure 25: Top clusters on weekend data January 2016 (pickup in blue and dropoff in red)

Comparing figure 24 with 3 and 25 with 4, we can see that weekday and weekend clusters in January 2016 are similar to the previously presented weekday and weekend clusters in section 4.2.1 with some negligible dissimilarities in clusters and the number of trips. For example, cluster 7 on weekend September 2015 is from LGA and for weekend January 2016, it is from JFK. But both are to the other parts (excluding Manhattan) of the city.

## 4.5 Findings

In summary, our findings from analyzing the pickup and dropoff regions are as follows:

- The largest hotspot region for both weekday and weekend, and all the time periods of the day is Manhattan. About 80% of the total trips are within Manhattan. After that, trips between LaGuardia Airport and Manhattan, and also between JFK Airport and Manhattan are in large numbers for any day.
- Other hotspot regions for pickup and dropoff are routes from Manhattan to Newark Liberty Airport, Manhattan to Brooklyn (mainly Brooklyn Heights and Williamsburg), Manhattan to Queens (mainly Hunters Point and Astoria), Manhattan to Queens (near Queens Zoo, Museum and Flushing Meadows Corona Park).
- On weekend, there are a large number of trips from LaGuardia Airport and JFK Airport to various parts of the city (mostly in Brooklyn and Queens) whereas trips from those two airports are mostly to Manhattan on weekday.

- There are a very few trips from Manhattan to Newark Liberty Airport compared to JFK and LaGuardia Airport.
- On weekday, there are a large number of trips from Manhattan to Brooklyn Heights, Williamsburg, Hunters Point and Astoria across the East River bridges whereas there are some trips from Manhattan to Queens Zoo, Queens Museum and Flushing Meadows Corona Park on weekend.
- 0:00 to 6:00 is the time period with lowest number of trips on weekday and trips count from 6:00 to 10:00 is the lowest on weekend. The number of trips are highest from 19:00 to 24:00 for any day.
- Trips from Manhattan to Brooklyn Heights, Williamsburg, Hunters Point and Astoria across the East River bridges are highest in numbers from 19:00 to 24:00 on weekday whereas trips count on those routes are highest from 0:00 to 6:00 on weekend.

# 5 Conclusion

In this project, we analyzed dense regions for pickup and dropoff of yellow taxi trips in the city of New York with respect to the different time periods of weekday and weekend. We divided trip records from September 2015 to weekday trips, weekend trips and trips in five different periods of the day. Then we clustered those trips using HDBSCAN and evaluated the clustering results by calculating Silhouette Coefficient, Calinski-Harabasz and Davies-Bouldin Index. Finally, We compared clustering results with DBSCAN and OPTICS, and with other data to show that the results are consistent.

We found Manhattan, LaGuardia Airport and JFK Airport as the hotspot regions with highest number of pickups and dropoffs. There are also a large number of trips from Manhattan to WilliamsBurg, Brooklyn Heights, Hunters Point, Astoria, Newark Liberty Airport, Queens Zoo and Queens Museum. Furthermore, most of the trips are between Manhattan and the two Airports JFK and LGA on weekday whereas trips are from those Airports to not only Manhattan but also to other parts of the city on weekend which indicates differences between hotspot regions in weekday and weekend. Hotspot regions are also different throughout the day. For example, the number of trips from Manhattan to Brooklyn Heights, Williamsburg, Hunters Point and Astoria across the East River bridges are highest from 19:00 to 24:00 on weekday but from 0:00 to 6:00 on weekend.

It is expected that our analysis will contribute to plan the transportation system in large cities like New York more efficiently and effectively. It will also help to meet the demand of transportation as well as to control the traffic flow of large cities for better service.

# References


[1] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 49–60, New York, NY, USA, 1999. ACM.

[2] Tadeusz Caliński and Harabasz JA. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01 1974.

[3] Ricardo Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. volume 7819, pages 160–172, 04 2013.

[4] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[6] Rami Ibrahim and M. Omair Shafiq. Detecting taxi movements using random swap clustering and sequential pattern mining. *Journal of Big Data*, 6, 05 2019.

[7] Davoud Moulavi, Pablo A Jaskowiak, Ricardo Campello, Arthur Zimek, and Joerg Sander. Density-based clustering validation. 04 2014.

[8] Umang Patel and Anil Chandan. Nyc taxi trip and fare data analytics using bigdata. 10 2015.

[9] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[10] Julia Stoyanovich, Matthew Gilbride, and Vera Zaychik Moffitt. Zooming in on nyc taxi data with portal, 2017.

[11] S M Turner, W L Eisele, R J Benz, and D J Holdener. Travel time data collection handbook, washington, dc: Office of highway information management, federal highway administration, u.s. dept. of transportation, 1998.

[12] H. Xiong, L. Chen, and Zhipeng Gui. A web-based platform for visualizing spatiotemporal dynamics of big taxi data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W7:1407–1412, 09 2017.