Geospatial Clustering of Taxi Rides: An Analysis

Chirasthi Amarasingha
Decision Sciences Department
University of Moratuwa
Colombo, Sri Lanka
amarasinghacu.20@uom.lk

***Abstract*—This study explores the application of clustering algorithms to explore the geospatial and temporal insights of taxi rides. The study addresses the insights revealed through clustering and their implication for decision-makers. Further analysis has been conducted to understand how clusters vary over time and their implications. The study concludes with recommendations for further research and the continued exploration of advanced clustering methods in geospatial analyses.**

***Keywords*—Spatial Clustering, K-Means, OPTICS, Temporal Analysis**

1. Introduction

Most economic activities rely heavily on transportation. It's a service that aids businesses to transport goods and people to move from one place to another. Taxi rides form one of the main transportation modes of urban livelihood due to the complex, busy schedules and economic activities surrounding the area. The significance of taxi rides extends beyond the simple act of commuting. It holds valuable insights for city planners, transportation authorities, businesses, and the general public. In understanding the importance of taxi rides, geographical locations form the key to information. As taxis travel, city streets, they generate a spatial trail that can be analyzed to reveal hidden patterns and trends. For this, spatial analysis becomes essential.

Spatial analysis is a broad field that examines spatial and geographical data to uncover insights, patterns, and trends. It can be used in several fields making it a powerful tool for further decision making. It uses various data analysis techniques to discover spatial relationships to make better decisions [1].

Spatial clustering is one such technique that involves grouping data points in a spatial space based on their proximity or similarity. The fundamental idea is to identify concentrations of entities that exhibit similar characteristics in their locations. This helps to identify relationships within spatial data, providing insights into how spatial entities are distributed over space. It is widely used in geography, public health, criminology, and many other fields [2].

Transportation, particularly the taxi industry is one such industry that applies spatial clustering for decision making. For example, taxis in New York City offer over 150 million rides per year [3]. This generates a large volume of data that shows spatial relationships over popular destinations, fare information, and peak hours that can be potentially useful. In this study, taxi ride information in New York City is analyzed by applying clustering algorithms to uncover hidden patterns in taxi rides. Therefore, the primary goal of this study is to conduct geospatial clustering to identify population destinations for rides and to understand how they change over time.

2. Methodology

In order to understand the patterns existing in the clustering of taxi rides, a dataset from a competition hosted by Google Cloud through the Kaggle platform was obtained. The dataset contains around 55 million records of taxi rides provided in New York City from 2009 to 2015. The main features provided were the passenger count, fare, pickup date and time, and the pickup and drop-off coordinates of each taxi ride. An overview of the data is shown in Figure 1: Overview of Data. This dataset was cleaned, transformed, and clustered to obtain the results.

A. Data Preprocessing

The data contains a large number of records of taxi rides. Due to this, a random sample was selected from the dataset to reduce computational time. The sample was selected using stratified random sampling based on the year of the pickup date. As our objective is to understand how clusters change over time, this ensures that a representative sample is selected for the temporal analysis to be conducted.

After selecting the sample, the dataset was explored for further understanding. Firstly, the dataset was tested for missing values. This showed an insignificant number of records against the large dataset. Thus, they were safely removed from it. Secondly, visualizations and metrics were used to gain an understanding of the dataset. This is shown in Figure 2: Summary Statistics.

Upon this examination, several issues were found. The summary statistics show that location coordinates range from -3400 to 3400. This is unrealistic as latitudes only range between -90 and 90 and longitudes range from -180 to 180. Additionally, our analysis exclusively focuses on taxi rides in New York City which has a latitude of 40.730610, and a longitude of -73.935242. To address, this records that fall out of range of these coordinates have been filtered out. Another issue can be observed in the fare amount where some records take negative values and show outliers. They have been fixed by replacement and filtering. It can also be observed that certain records show reversed coordinates where latitudes are shown as longitudes and vice versa. This has been corrected by identifying such records and switching them.

	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	2009-06-15 17:26:21.0000001	4.500	2009-06-15 17:26:21 UTC	-73.844	40.721	-73.842	40.712	1
1	2010-01-05 16:52:16.0000002	16.900	2010-01-05 16:52:16 UTC	-74.016	40.711	-73.979	40.782	1
2	2011-08-18 00:35:00.00000049	5.700	2011-08-18 00:35:00 UTC	-73.983	40.761	-73.991	40.751	2
3	2012-04-21 04:30:42.0000001	7.700	2012-04-21 04:30:42 UTC	-73.987	40.733	-73.992	40.758	1
4	2010-03-09 07:51:00.000000135	5.300	2010-03-09 07:51:00 UTC	-73.968	40.768	-73.957	40.784	1

Figure 1: Overview of Data

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	5000000.000	5000000.000	5000000.000	4999964.000	4999964.000	5000000.000
mean	11.341	-72.507	39.920	-72.507	39.917	1.685
std	9.820	12.810	8.964	12.848	9.487	1.332
min	-100.000	-3426.609	-3488.080	-3412.653	-3488.080	0.000
25%	6.000	-73.992	40.735	-73.991	40.734	1.000
50%	8.500	-73.982	40.753	-73.980	40.753	1.000
75%	12.500	-73.967	40.767	-73.964	40.768	2.000
max	1273.310	3439.426	3310.364	3457.622	3345.917	208.000

Figure 2: Summary Statistics

B. Feature Engineering

In order to generate interesting information, more features have been created using the available information. The study's objective is to cluster taxi rides and understand how they change over time. For this, the temporal aspects are important. Thus, the 'pickup_datetime' column has been parsed and used to create additional features such as year, month, day, date, hour.

In analyzing taxi rides, the distance traveled in each ride provides valuable information for further analysis. Therefore, a distance feature has been created using Haversine distance. It's the angular distance between two points on the surface of a sphere. This can be applied to the surface of Earth as it's approximately spherical. This is a better measure than Euclidean distance as it considers the curvature of Earth. This becomes important when precision is required, as using a flat Earth approximation can lead to significant errors [4]. The distance measure can be used for further analysis in the study. After completing all the above steps, the final dataset is acquired. This is displayed in Figure 3: Overview of Final Dataset. As our analysis is to identify popular hotspots for pickup and drop off, the locations will be separately analyzed. The pickup and drop-off coordinates are displayed separately in Figure 4: Pickup Locations and Figure 5: Drop Off Locations respectively.

C. Cluster Analysis

After preparing the dataset, cluster analysis can be performed. The choice of the clustering method plays a crucial role in the effectiveness of the clustering process. Different algorithms may perform well with different datasets. Therefore, several methods must be used to see their performance with the dataset. This will enable us to identify clustering algorithms that align with the specific characteristics of the dataset to achieve meaningful results. For this study, both K-means and OPTICS are chosen to be used as clustering techniques.

K-means is a simple unsupervised machine learning algorithm that is quite popular. It groups similar data points into a predefined number of clusters based on the Euclidean distance. Each data point is assigned to the nearest cluster centroid. This method is suitable to be used for the problem at hand due to its computational efficiency. The dataset is quite large with about 55 million rows. K-means has a time complexity of O(NTK) which is lower than most algorithms. Therefore, by using K-means, clustering time can be reduced. Additionally, it requires only a few parameters and steps making it fast and scalable. K-means is the primary algorithm used in conducting the taxi ride analysis due to these reasons. However, its limitations must also be acknowledged. It is heavily affected by outliers due to its algorithm using mean to calculate distance from centroids. However, since the dataset has been cleaned to only include data points from New York City, there will be less sensitivity to outliers.

The second clustering algorithm used is the OPTICS algorithm. OPTICS (Ordering Points To Identify the Clustering Structure) is a density based clustering algorithm. It extends the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm by introducing a reachability distance, allowing it to discover clusters with varying densities.

Here, each data point is assigned a reachability distance based on the density of its neighborhood. This distance measures the distance at which a point can be reached while considering the local density. This helps in identifying clusters with varying densities.

This algorithm is particularly useful for the analysis to be conducted. One reason is that it's versatile for spatial datasets with irregularly shaped clusters. The dataset at hand is of taxi rides and clusters may exhibit irregular shapes and varying densities in different geographic regions. The algorithm is also capable of handling large datasets efficiently. It processes data points in an order which removes the need for parameter settings. The algorithm also reveals the hierarchical nature of clusters. This allows for a more detailed exploration of clustering structures.

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	distance	year	day	hour
count	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000	4879544.000
mean	11.332	-73.975	40.751	-73.974	40.751	1.691	2.707	2011.738	15.720	13.511
std	9.712	0.039	0.030	0.038	0.033	1.306	3.949	1.866	8.684	6.516
min	0.010	-74.989	40.034	-74.998	40.006	1.000	0.000	2009.000	1.000	0.000
25%	6.000	-73.992	40.737	-73.992	40.736	1.000	0.853	2010.000	8.000	9.000
50%	8.500	-73.982	40.753	-73.981	40.754	1.000	1.552	2012.000	16.000	14.000
75%	12.500	-73.968	40.768	-73.965	40.768	2.000	2.830	2013.000	23.000	19.000
max	500.000	-72.063	41.923	-72.067	41.998	6.000	213.420	2015.000	31.000	23.000

Figure 3: Overview of Final Dataset

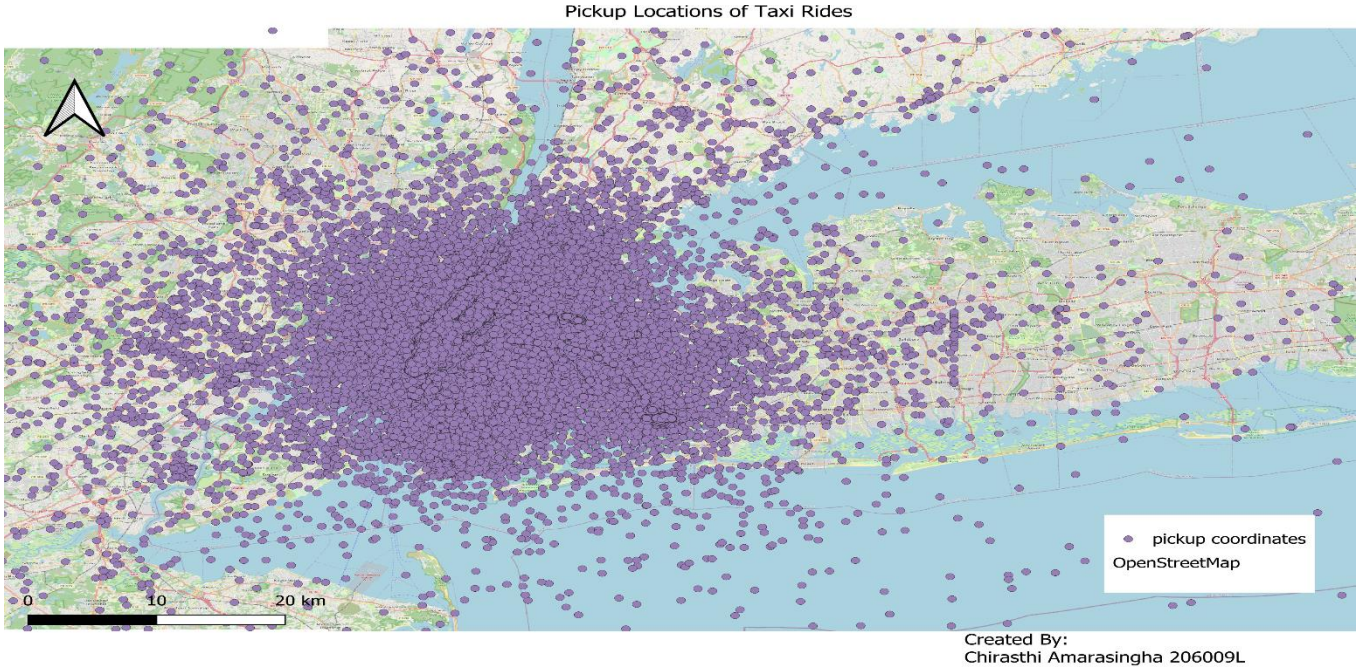

Figure 4: Pickup Locations

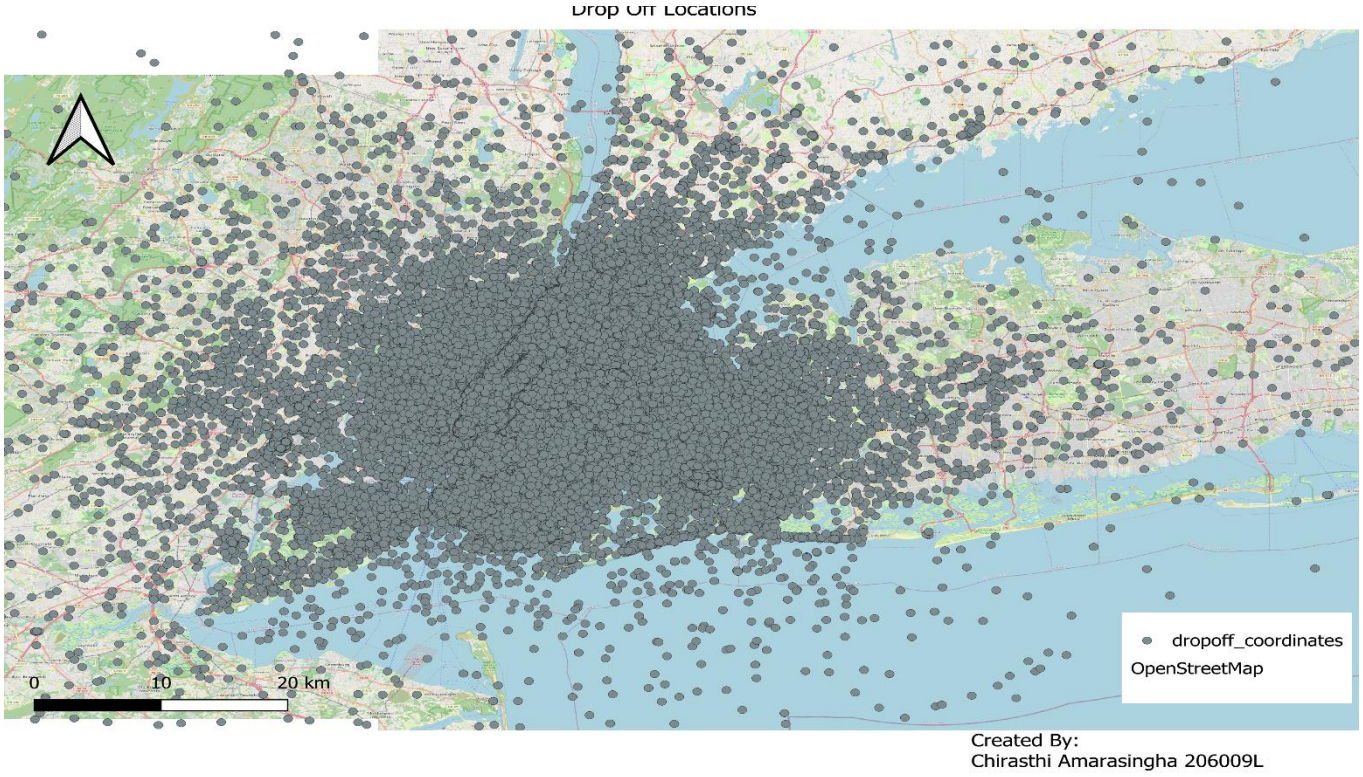

Figure 5: Drop Off Locations

3. Results and Discussion

A. *K-means clustering*

To perform k-means clustering, the number of clusters must be known beforehand. This is one of the limitations of the algorithm. Since the optimum number of clusters is not known beforehand, the elbow method can be used to determine the best 'k' number of clusters for the dataset. The K distance graphs drawn for pickup and drop off locations are displayed in Figure 6: Elbow Chart for Pickup Clustering and Figure 7: Elbow Chart for Drop Off Clustering respectively. It can be observed that 6 clusters and 5 clusters seem optimal for pickup and drop off clustering respectively. After clustering pickup coordinates, Figure 8 was obtained. From this, it can be observed that there are 6 clusters present within the data. When examined closely, the hotspots within the clusters can be observed. Thus, the clusters have been named and shown in Figure 9 and Table 2: Pickup Clusters.

To understand which clusters are the most popular locations for taxi pickups, the density proportion of each cluster was calculated. This will enable us to identify how dense a cluster is and indicate hotspot clusters for taxi rides. The results are shown in Figure 10.

It can be seen that cluster 0 is the most dense out of all the clusters. Cluster 0 contains a network of roads that connects several tunnels and the metro station. This is one reason for the denser cluster. Cluster 4 becomes the second densest cluster with a proportion of 0.28. It overlooks the area of lower Manhattan. It can be seen that cluster 5 and 1 also have a significant density proportion. This is due to the presence of airports within those clusters. People would often choose to use taxi rides instead of other modes when traveling to and from an airport due to their efficiency. The lowest density is seen in cluster 3. This is also reflected in the clustering, where there are sparse data points in cluster 3. Overall, the k-means clustering identifies that the cluster which overlooks the road network and covers several transportation stations is the most popular location to be picked up for taxi rides.

The same process was applied to the clustering of drop off coordinates, to identify taxi hotspots. The clustering results are shown in Figure 12. After clustering, the most popular destinations for drop offs were also examined. These results are shown in Figure 11.

Once again, the results indicate that cluster 0 which covers the road network is the most dense. Clusters 1 and 4 show similar results for the density proportions. These clusters cover different areas of Manhattan which is known to be an urban hotspot due to the economic activities that occur there. Another cluster also shows significant density. This covers the La Guardia Airport and John F. Kennedy Airports. This shows that customers prefer taxi rides when dropping off at the airport. Overall, we can also identify popular destinations for drop offs using this method.

Table 1: Drop Off Clusters

Cluster	Drop Off Locations
0	Lincoln Tunnel, West 53rd Street
1	Williamsburg, Lower Manhattan, Broadway
2	La Guardia Airport, John F. Kennedy Airport
3	Long Island
4	Manhattan, Central Park, Robert F. Kennedy Bridge

Table 2: Pickup Clusters

Cluster	Pickup Locations
0	Lincoln Tunnel, New York Penn Station, Grand Central Terminal
1	John F. Kennedy Airport
2	Manhattan, Park Avenue
3	Long Island
4	Broadway, Holland Tunnel, Lower Manhattan
5	La Guardia Airport

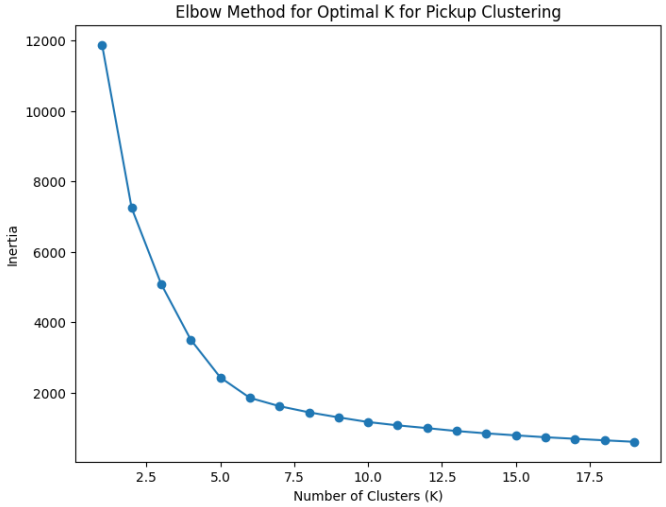

Figure 6: Elbow Chart for Pickup Clustering

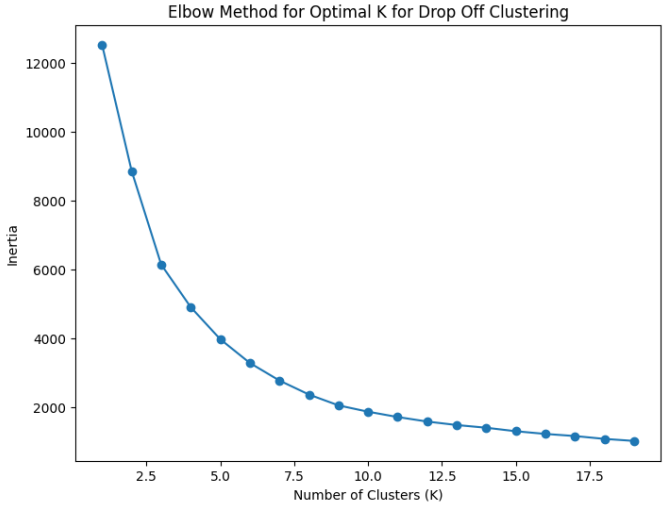

Figure 7: Elbow Chart for Drop Off Clustering

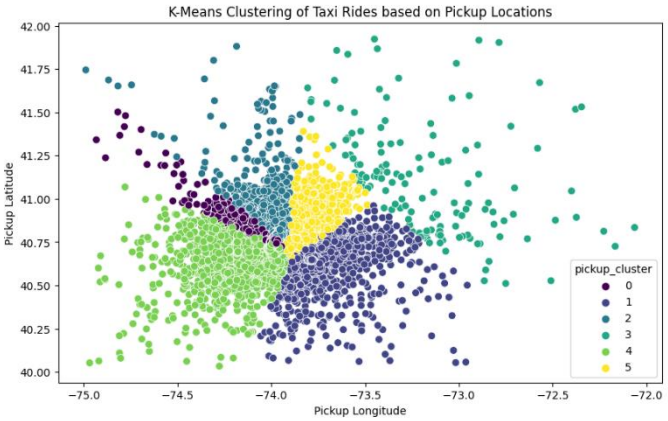

Figure 8: K-Means Clustering for Pickup Location

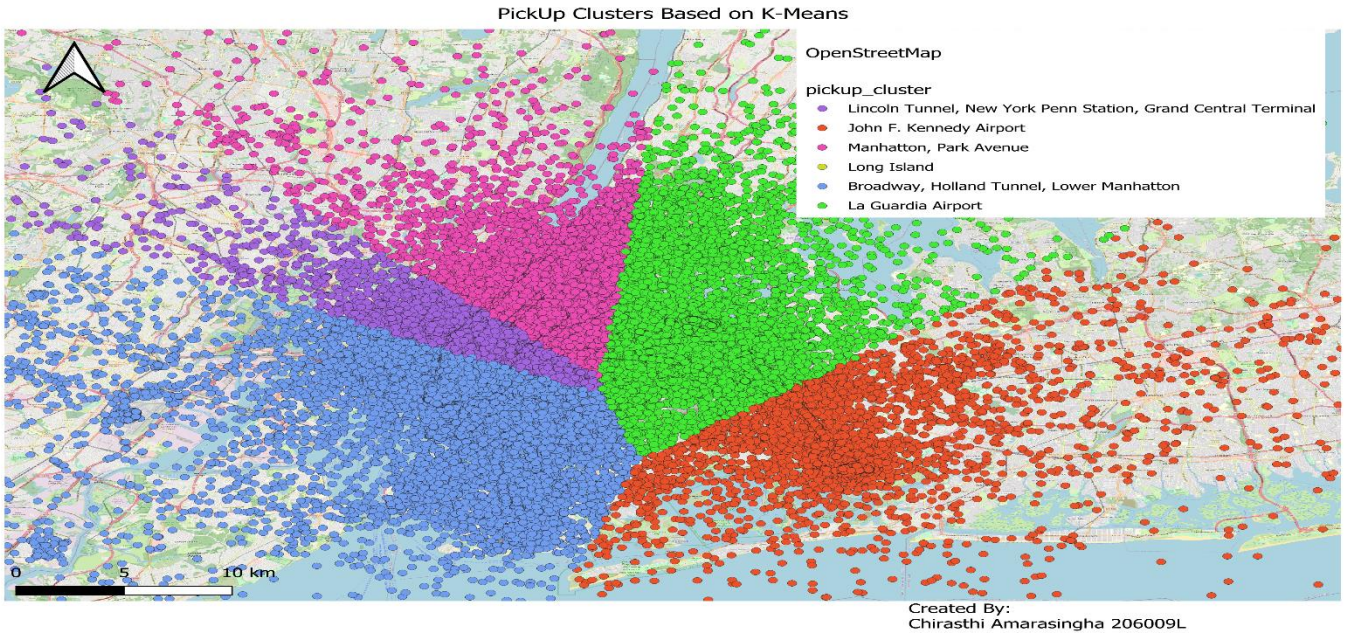

Figure 9: Pickup Clusters Based on K-Means

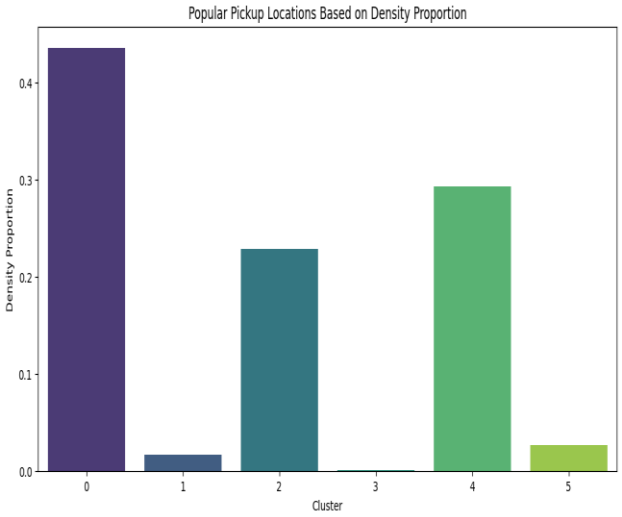

Figure 10: Cluster Popularity Based on Density Proportion

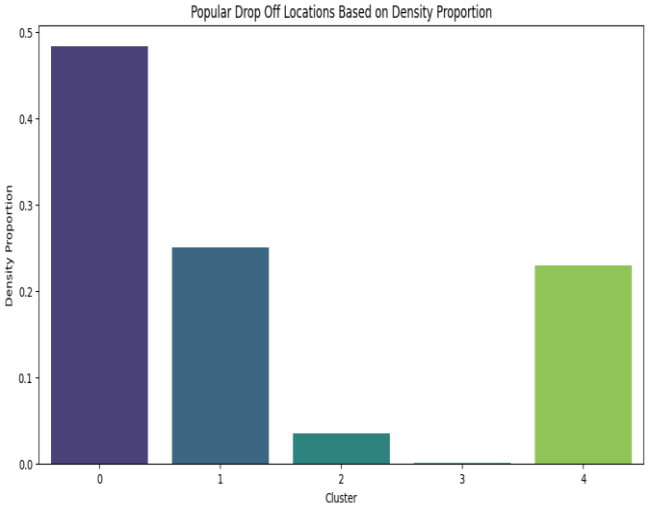

Figure 11: Cluster Popularity Based on Density Proportion

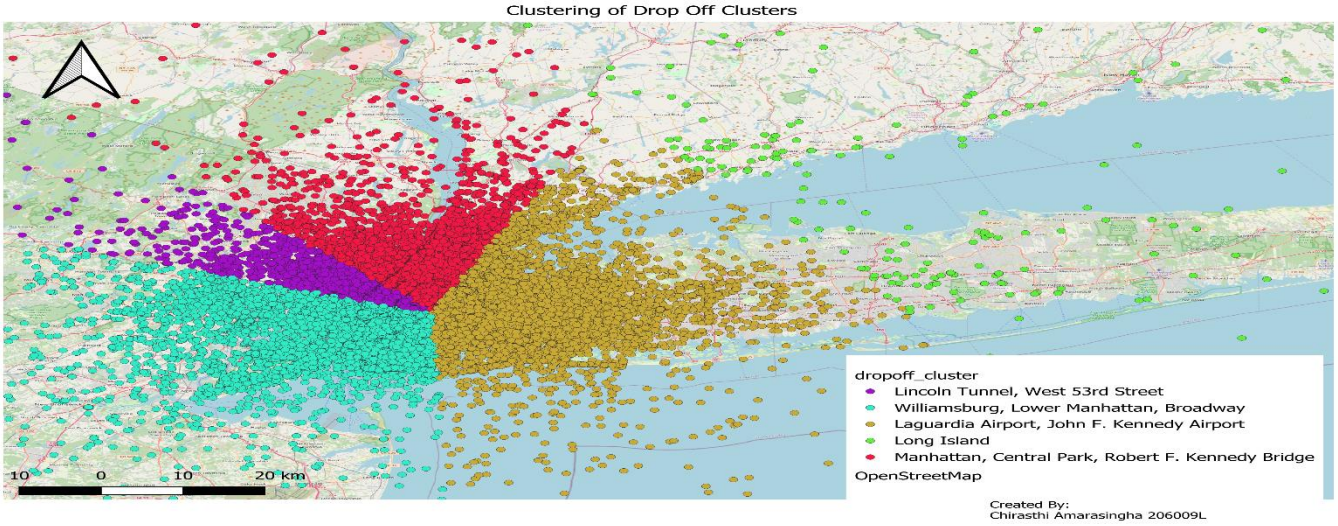

Figure 12: K-Means Clustering of Drop Off Locations

The second objective of our study is to analyze how the clusters behave over time. This requires a temporal analysis of both the pickup and drop off clusters. First, all clusters were analyzed based on the month. The analysis for pickup clusters is shown in Figure 13. It can be seen that most of the clusters have experienced similar demand throughout all the months with little variation. This indicates that there is a lesser seasonality effect for taxi ride pickups. In cluster 0 (Lincoln Tunnel, Metro Station) higher demand is seen during the first half of the year. For cluster 1, (Kennedy Airport) the highest demand was received in May and June with lower demand in the later months. For cluster 2, the highest demand was seen in May with all other months having similar demand. Cluster 3 (Long Island) shows the largest variations. The first 3 months have lesser demand, but the demand has increased from April onwards. The highest demand is seen in June and July. This indicates that there is a seasonality effect in demand in cluster 3. Americans celebrate the 4th of July as their Independence Day. This could be a result of it. Clusters 4 and 5 have similar results where higher demand is seen in May and June.

When analyzing the drop off clusters based on the month, it showed similar results. This is shown in Figure 14. In cluster 0 (Lincoln Tunnel, West 53rd Street) higher demand is seen closer to the beginning of the year. For cluster 1, (Williamsburg, Lower Manhattan, Broadway) the highest demand was received during the first half of the year and a slight decrease following June. In cluster 2 (La Guardia Airport, John F. Kennedy Airport) the highest demand was received in May and a constant pattern in the latter half of the year. Cluster 3 (Long Island) shows the most variations. Starting out normally the demand took a huge dip during February which lasted till a rise in April followed by a huge rise in May. This indicates that there is a seasonality effect for demand in cluster 3. Cluster 4 shows much less variation except for a slight dip in demand in August.

The clusters were also analyzed based on hours. This could derive insights into peak and slump hours for taxi rides. When observing the pickup clusters, it can be seen that there is an upward trend across all clusters where demand has slowly increased gradually over the hours. This is shown in Figure 15. Most of the clusters show increased demand from 6 p.m. to 10 p.m. This provides rich insights into the nightlife of US citizens. New York is a prime location for nighttime activities and is enjoyed by many tourists [5]. Taxi demand has been lower in the early morning hours and increased slightly in late morning and afternoon hours. The highest taxi demand is seen after 4 p.m. This coincides with office leave hours.

When observing the drop off clusters, it can be seen that there is an upward trend across all clusters where demand has slowly increased gradually over the hours similar to pickup clusters. In Figure 16 the lowest demand can be seen in the first 6 hours of the day and this can be seen in every cluster. The peak demand is always created during the last 6 hours of the day except for cluster 2 (La Guardia Airport, John F. Kennedy Airport). This matches with the previous observations made with pickup clusters. Taxi demand has been lower for early morning hours and increased slightly for late morning and afternoon hours as New York is a prime location for nighttime activities and is enjoyed by many tourists [5].

The clusters were also analyzed based on US holidays. For this, a dataset with US Holidays was obtained from the Kaggle Platform. It can be seen that clusters 0, 2, and 4 have had large variations in taxi intensity during holidays while cluster 1,3,5 have had little effect from holidays. Across all clusters high taxi demand can be observed during the 4th of July (Independence Day of USA), Christmas Eve, Easter, Labor Day Weekend, Memorial Day, and Valentine's Day. This is displayed in Figure 17.

. These insights have significant implications for urban transportation planning and management. The delineation of high-density clusters can inform the optimization of taxi fleet deployment, aiding authorities in allocating resources efficiently. Understanding temporal variations enables policymakers to anticipate and address fluctuations in demand, potentially enhancing the overall efficiency and responsiveness of urban transportation systems. Furthermore, the analysis contributes valuable information for urban planners and policymakers aiming to implement data-driven strategies for optimizing transportation services and improving the overall mobility experience in urban environments.

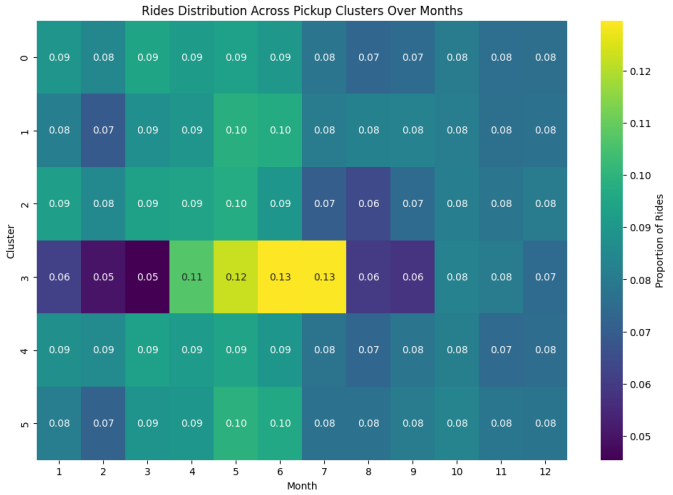

Figure 13: Ride Distribution Across Pickup Clusters Based on Months

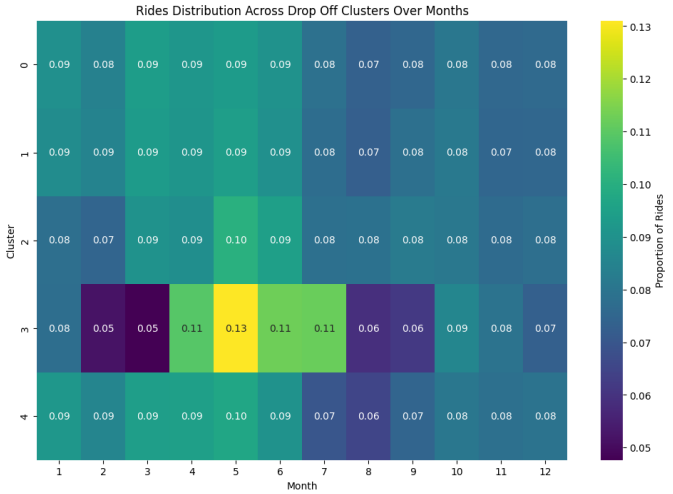

Figure 14: Ride Distribution Across Drop Off Clusters Based on Months

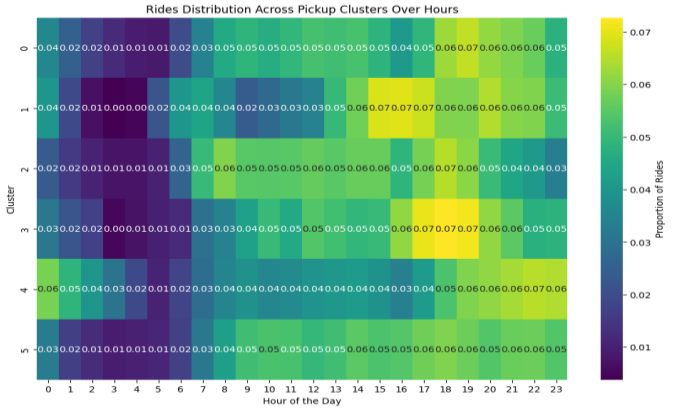

Figure 15: Ride Distribution Across Pickup Clusters Based on Hours

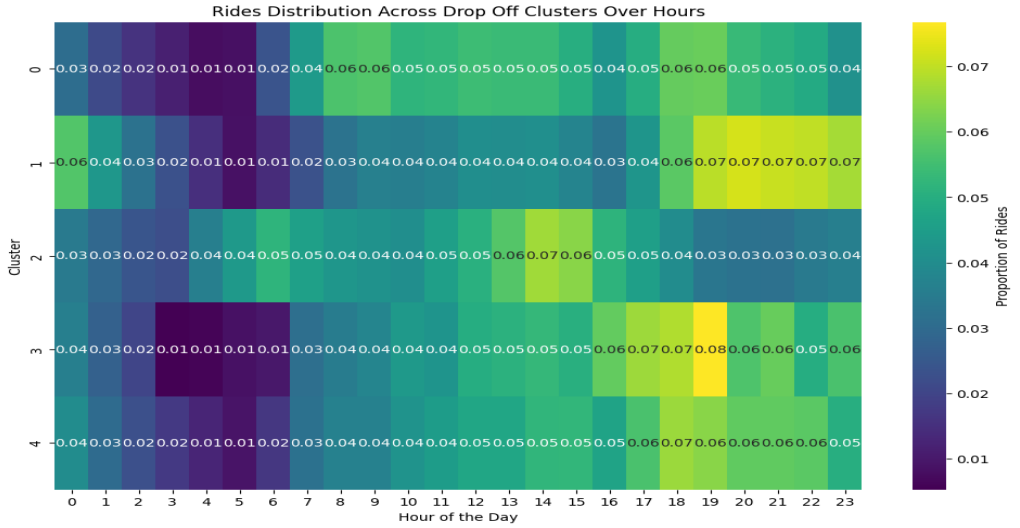

Figure 16: Ride Distribution Across Drop Clusters Based on Hours

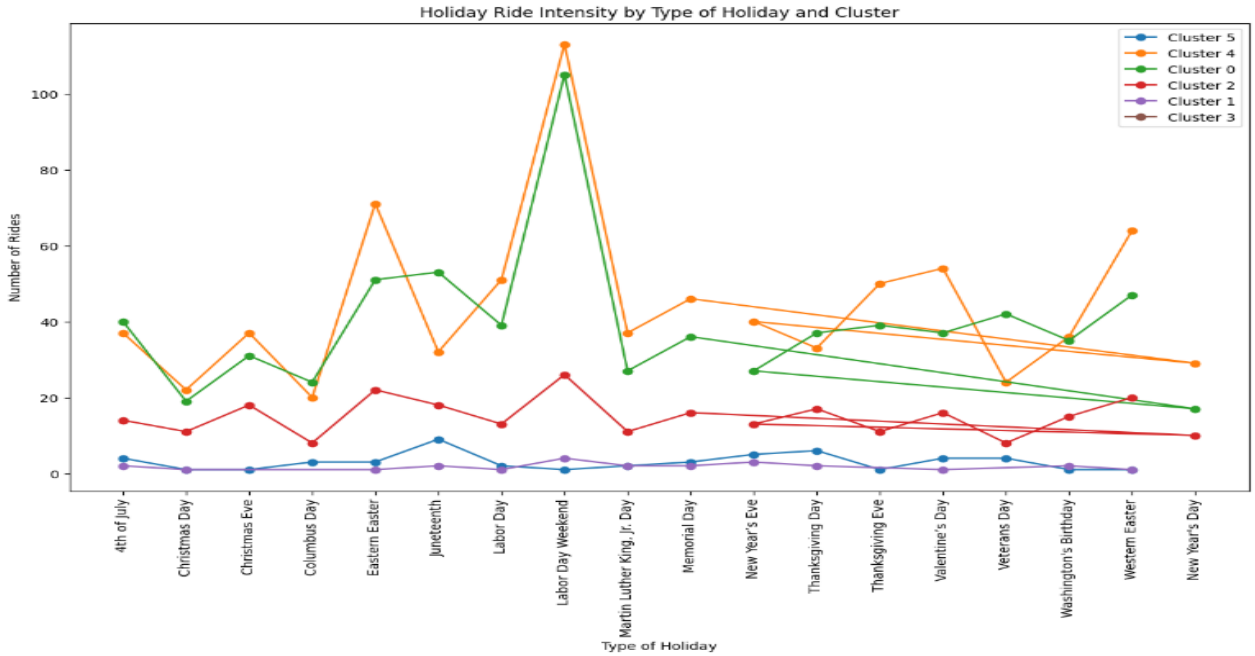

Figure 17: Ride Distribution During Holidays

B. OPTICS Clustering

As previously performed OPTICS clustering was applied to both pickup and drop off locations separately. To identify the best parameters for each cluster analysis, several iterations were run against a silhouette score. The best clusters produced are shown in Figure 18 and Figure 19 respectively. Unlike K-means, OPTICS has been able to identify arbitrarily shaped clusters giving more insights.

Both cluster analyses identify the majority of taxi rides as one cluster while there are several clusters at La Guardia Airport (clusters 0,1, 2) and John F. Kennedy Airport(clusters 3 and 4). This shows that there is a high density of taxi rides at the airports.

Both cluster analyses were examined on a temporal basis to understand how clusters change over time. The pickup cluster analysis based on hours of the day is shown in Figure 20

Cluster -1 shows that taxi rides have remained relatively consistent. Clusters at La Guardia Airport have fewer taxi rides from 12 midnight to 8 a.m. This increases gradually and peaks around 2 p.m. to 6 p.m. This indicates that most flights occur around this time and passengers seek taxi rides from the airport to their desired destination. Clusters at John F. Kennedy Airport have little to no taxi demand from midnight to 6 a.m. The hours from 7 a.m. to 8 a.m. have high demand which has reduced afterward. This demand has picked up again for the evening and night hours.

The drop off clusters were also analyzed based on the hours of the day which is shown in Figure 21. This revealed a different pattern than pickup clusters. Cluster -1 shows a consistent pattern of taxi rides throughout the day. Clusters at La Guardia airport (0 and 1) show that taxi ride drop-offs start around 4 a.m. which increases gradually and peaks at noon till 3 p.m.. Thereafter, the drop offs reduce till it becomes null around 8 p.m. and after. The cluster at John F. Kennedy Airport (cluster 2) shows a similar pattern where taxi drop offs peak during the hours between 1 p.m. to 3 p.m.

These insights could be beneficial to taxi ride companies so that they can plan the timings for taxi rides. Taxi rides for pickups peak at night while it peaks for drop offs in the afternoon. Taxi rides could become more efficient and save many costs by dropping off customers in the afternoon and picking up customers coming from flights at night.

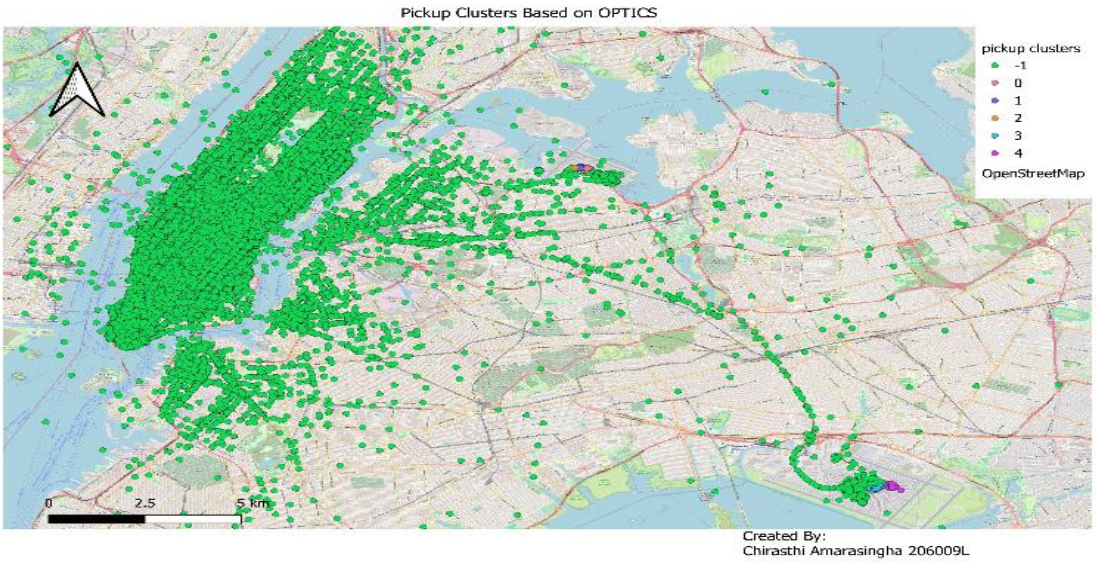

Figure 18: OPTICS Clustering of Pickup Locations

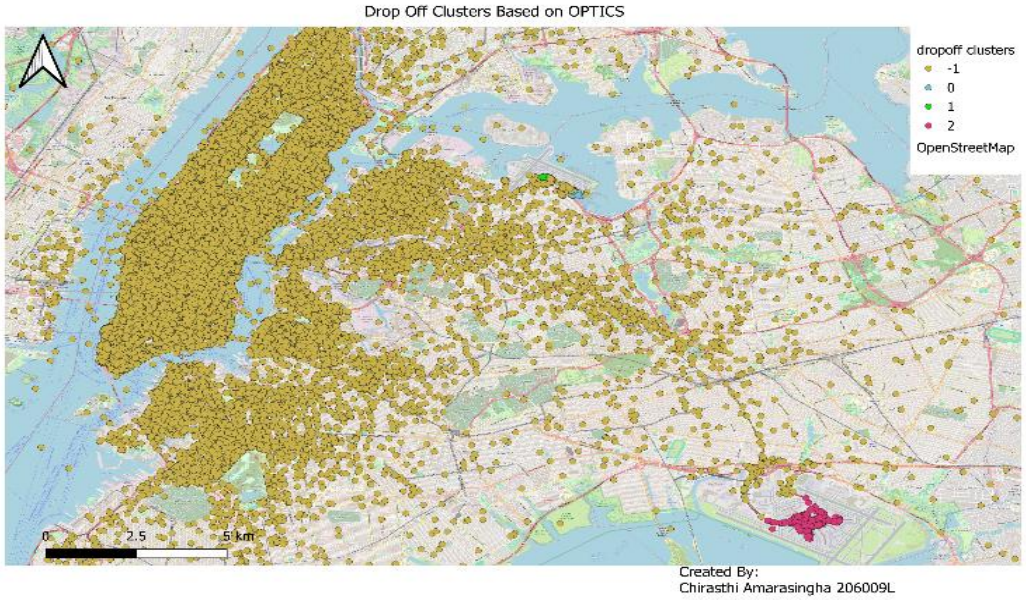

Figure 19: OPTICS Clustering of Drop Off Locations

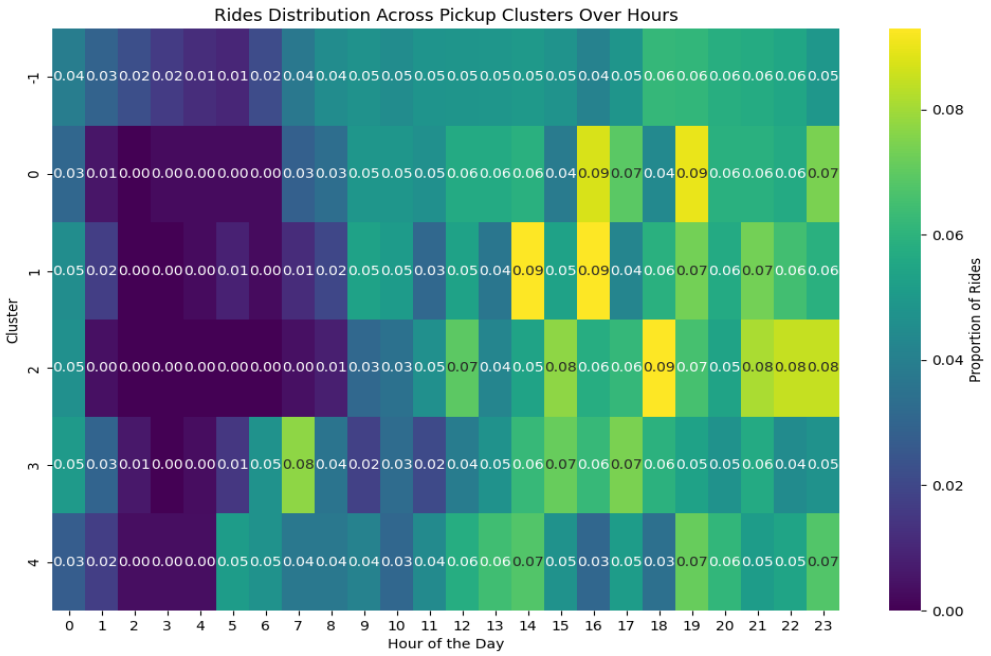

Figure 20: Pickup Cluster Distribution Over Hours

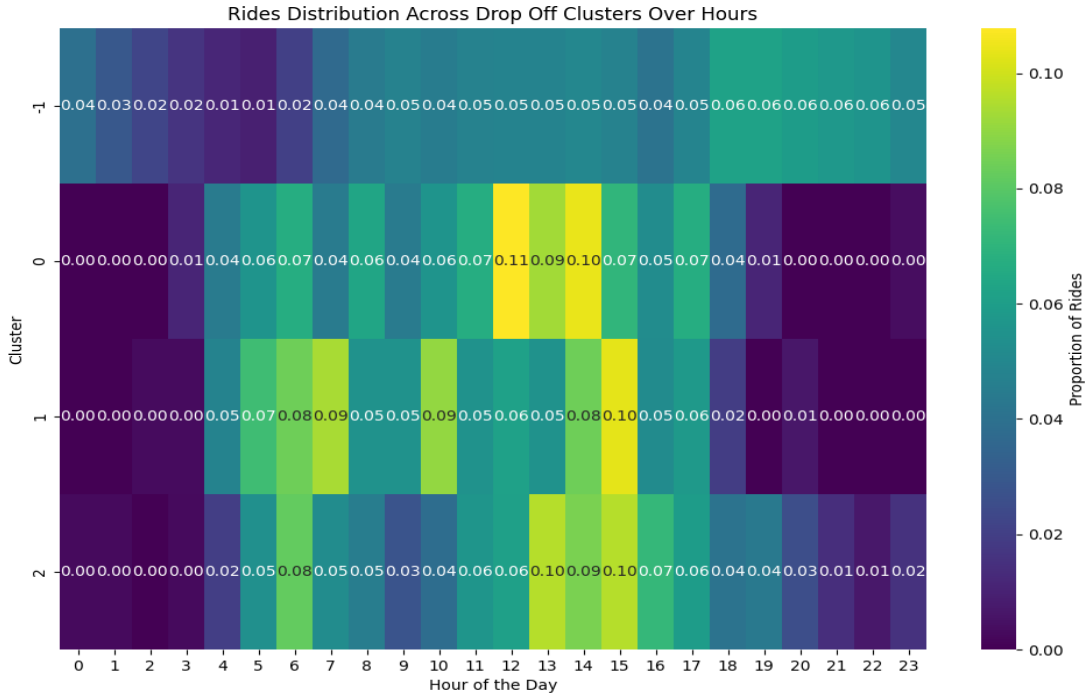

Figure 21: Drop Off Cluster Distribution Over Hours

C. *Limitations*

The clustering analyses conducted so far revealed insightful information on how taxi rides fare in New York City. However, the reliability of these insights is limited by certain factors.

The analyses were conducted with only two clustering algorithms. There could be other techniques that work better with the dataset. Both k-means and OPTICS algorithms have their own limitations. The analysis becomes limited due to these factors. For example, K-Means is highly sensitive to outliers. This could have impacted the clustering results.

Additionally, the clustering was done with a representative sample of the dataset due to the large size of the dataset. Therefore, there could be interesting patterns that were missed during the initial analysis.

4. CONCLUSION

In conclusion, this study delves into the application of clustering algorithms to a large geospatial dataset, particularly focusing on taxi ride data. The primary objective was to identify spatial patterns, uncover clusters, and explore the temporal dynamics of these patterns over time. Various clustering algorithms, including DBSCAN and OPTICS, were discussed and applied to the dataset, considering their strengths and limitations.

The study emphasized the information revealed by the clustering analysis. Through the temporal analysis, it was revealed that there are certain patterns for taxi ride times and months. Certain hours of the day showed higher intensity for taxi rides than others. This could be useful for many stakeholders in their decision-making. For example, taxi companies can better plan their offers. Policymakers could use this information when building road networks, traffic systems, etc.

The integration of external information, such as holidays, enriched the analysis, allowing for a more comprehensive understanding of how clusters evolve over time.

The analysis conducted is limited by several factors such as the choice of algorithm etc. Despite these challenges, clustering algorithms prove valuable for uncovering hidden structures within spatial datasets, providing insights that can inform decision-making processes. Moving forward, the study suggests the need for further research into advanced clustering methods, and a continued effort to bridge the gap between algorithmic outputs and actionable insights in real-world applications of geospatial clustering.

5. REFERENCES

[1] A. T. Murray, "Significance Assessment in the Application of Spatial Analytics," Annals of the American Association of Geographers, vol. 111, no. 6, pp. 1740-1755, 2021.

[2] A. T. Murray, T. H. Grubesic and R. Wei, "Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense," Annals of the Association of American Geographers, vol. 104, no. 6, pp. 1134-1156, 2014.

[3] B. Ata, N. Barjesteh and S. Kumar, "Spatial Pricing: An Empirical Analysis of Taxi Rides in New York City," The University of Chicago Booth School of Business Chicago, 2019.

[4] R. Agramanisti Azdy and F. Darnis, "Use of haversine formula in finding distance between temporary shelter and waste end processing sites," Journal of Physics: Conference Series, vol. 1500, no. 1, 2020.

[5] New York Tourism, "NYC Nightlife," 2023. [Online]. Available: https://www.nyctourism.com/things-to-do/nightlife/. [Accessed 09 12 2023].

[6] QGIS.org, %Y. QGIS Geographic Information System. QGIS Association. http://www.qgis.org