NYC Yellow Cab Demand Analysis

Stephen Cho, Minjee Kim

November 2024

1 Overview

The data for this project was sourced from the New York City Taxi and Limousine Commission (TLC) website, which provides comprehensive information on yellow taxi trips across New York City. The raw dataset includes trip-level details, such as pickup and dropoff timestamps, trip distances, passenger counts, fare information, and location IDs corresponding to the pickup and dropoff zones.

For spatial analysis, we needed to map these location IDs (PULocationID and DOLocationID) to geographic coordinates. This was done by obtaining centroid information for each taxi zone from the NYC Taxi Zones dataset, which is publicly available on the City of New York's open data portal [2]. This allowed us to match the zone IDs from the TLC trip data to their corresponding spatial locations, enabling us to analyze the movement patterns of yellow cabs across the city.

The dataset used for this project covers yellow taxi trips from August 2024, the most recent data available. The original dataset contained over 3 million rows, which was too large for R to handle efficiently. To overcome this limitation, we randomly sampled and batched 1 million rows. We used this 1 million row data analyzed in this project.

Using this cleaned NYC yellow cab data, our goal is to use spatio-temporal clustering to capture the yellow cab demand with underlying motivations, such as yellow cab demand from outer areas into inner, Manhattan areas, during commuting hours and movements into bar area during weekday nigh times.

To do so, we take advantage of temporal bases coefficient clustering using Fourier transform and KNN spatial structure as a spatial smoothing constraint. We use the ClustGeo package [1] for computational efficiency and describe the model set up for justification in the methodology section.

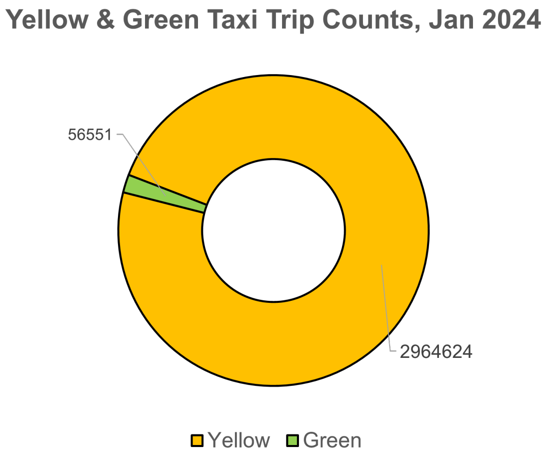
Our results show a clear trend of demand to ride into the city on weekends around 8 10pm and to ride out of the city during weekday nights after dinner time, which were not visible trends from EDA, proving our method to be efficient.

2 Introduction

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's medallion (yellow) taxis, street hail livery (green) taxis, for-hire vehicles (FHVs) and high-volume FHVs, commuter vans, and paratransit vehicles. The TLC cooperates with taxi technology providers (now called technology service providers, or TSPs) to collect trip record information for each taxi and FHV trip completed by licensed drivers and vehicles.

2.1 Taxi Types

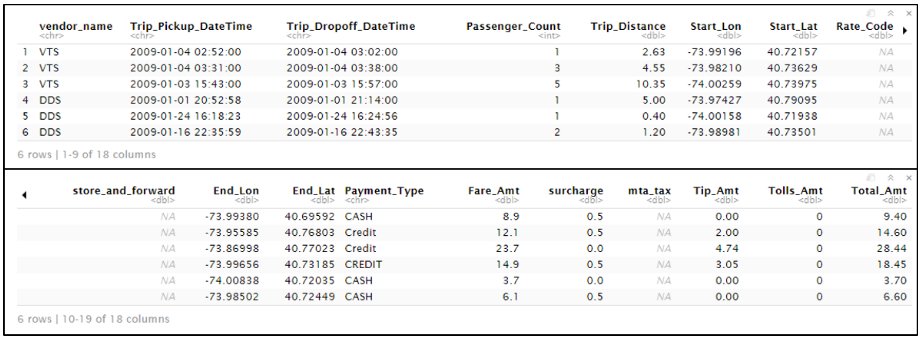
Taxi trip data can be acquired from the TLC website [2], where the trip records are published by vehicle type (yellow/green/FHV/high-volume FHV), each further separated by year and month. Among the four vehicle types, our target is first narrowed down to yellow and green taxis, as they are the "traditional" taxi types that respond to street hails, as well as being incorporated under a more reliable source of data collection, contrary to FHV trip records that rely on corporations such as Uber, Lyft, etc.


Yellow & Green Taxi Trip Counts, Jan 2024

Regarding the two taxi types, we can easily observe that the usage of green, or the "boro" taxis, are very limited compared to yellow taxis; this is mainly due to their specific purpose of serving outer boroughs, which limits vehicles from picking up new passengers within the "yellow zone" of Manhattan, or within airports. This has led to 86 percent plunge in numbers of operating green cabs from 6,500 in 2015 to less than 900 in 2023. [3]

Aside from having very little demand compared to yellow taxis, it is clear that the nature of green taxis does not fit our purpose of analyzing taxi trip patterns across all of New York City. Therefore, the dataset used for this analysis will consist of trip record data of yellow taxis only.

2.2 Yellow Taxi Trip Data



	vendor_name <chr>	Trip_Pickup_DateTime <chr>	Trip_Dropoff_DateTime <chr>	Passenger_Count <int>	Trip_Distance <dbl>	Start_Lon <dbl>	Start_Lat <dbl>	Rate_Code <dbl>
1	VTS	2009-01-04 02:52:00	2009-01-04 03:02:00	1	2.63	-73.99196	40.72157	NA
2	VTS	2009-01-04 03:31:00	2009-01-04 03:38:00	3	4.55	-73.98210	40.73629	NA
3	VTS	2009-01-03 15:43:00	2009-01-03 15:57:00	5	10.35	-74.00259	40.73975	NA
4	DDS	2009-01-01 20:52:58	2009-01-01 21:14:00	1	5.00	-73.97427	40.79095	NA
5	DDS	2009-01-24 16:18:23	2009-01-24 16:24:56	1	0.40	-74.00158	40.71938	NA
6	DDS	2009-01-16 22:35:59	2009-01-16 22:43:35	2	1.20	-73.98981	40.73501	NA

6 rows | 1-9 of 18 columns

store_and_forward <dbl>	End_Lon <dbl>	End_Lat <dbl>	Payment_Type <chr>	Fare_Amt <dbl>	surcharge <dbl>	mta_tax <dbl>	Tip_Amt <dbl>	Tolls_Amt <dbl>	Total_Amt <dbl>
NA	-73.99380	40.69592	CASH	8.9	0.5	NA	0.00	0	9.40
NA	-73.95585	40.76803	Credit	12.1	0.5	NA	2.00	0	14.60
NA	-73.86998	40.77023	Credit	23.7	0.0	NA	4.74	0	28.44
NA	-73.99656	40.73185	CREDIT	14.9	0.5	NA	3.05	0	18.45
NA	-74.00838	40.72035	CASH	3.7	0.0	NA	0.00	0	3.70
NA	-73.98502	40.72449	CASH	6.1	0.5	NA	0.00	0	6.60

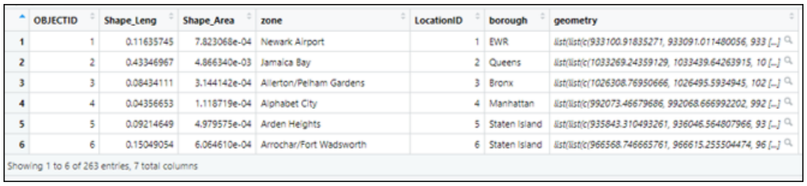
6 rows | 10-19 of 18 columns

Figure 1: Yellow Taxi Data of January 2009

The NYC TLC releases dataset of monthly taxi trip records, which consists of 18 columns roughly 3 million rows in average. Variables provide key details of each trip, such as Pickup/Dropoff times and locations, passenger count, trip distance, cost breakdown (fare, surcharge, tip, tolls, etc.). Our key variables that represent spatial and temporal information are:

- `Trip_Pickup_DateTime, Trip_Dropoff_DateTime`: temporal
- `PU_Longitude/Latitude, DO_Longitude/Latitude`: spatial

A major change has been made to the spatial variable in recent years, however; the TLC no longer provides the exact coordinates of each pickup/dropoff locations (represented as Start/End Lon and Lat columns in figure 1), now replaced by location ID information of the ”taxi zone” that each location falls into. While the exact lon/lat would best serve our purpose, only the records from 2009 and 2010 are not yet updated and available in ”old” format. Since the scope of this analysis is to address recent trends of taxi demand in NYC, working with the ”new” format was inevitable. Therefore, we have decided to utilize taxi zone shape file, also provided by the TLC, to make amends for the missing coordinates information.



	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
1	1	0.11635745	7.823068e-04	Newark Airport	1	EWR	list(list(c(933100.91835271, 933091.011480056, 933 [...]
2	2	0.43346967	4.866340e-03	Jamaica Bay	2	Queens	list(list(c(1033269.24359129, 1033439.64263915, 10 [...]
3	3	0.08434111	3.144142e-04	Allerton/Pelham Gardens	3	Bronx	list(list(c(1026308.76950666, 1026495.5934945, 102 [...]
4	4	0.04356653	1.118719e-04	Alphabet City	4	Manhattan	list(list(c(992073.46679686, 992068.666992202, 992 [...]
5	5	0.09214649	4.979575e-04	Arden Heights	5	Staten Island	list(list(c(935843.310493261, 936046.564807966, 93 [...]
6	6	0.15049054	6.064610e-04	Arrochar/Fort Wadsworth	6	Staten Island	list(list(c(966568.746665761, 966615.255504474, 96 [...]

Showing 1 to 6 of 263 entries, 7 total columns

Figure 2: Overview of taxi zones information

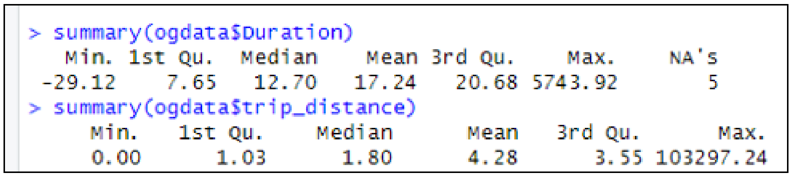
The TLC has divided New York City into 263 taxi zones, which are represented by newly added ”PULocationID”,”DOLocationID” columns in the taxi trip dataset. As the plots under EDA section will also suggest, the city is micro-zoned to a good number of zones. Therefore, we reached to a conclusion that approximating the coordinates of pickup/dropoff location based on taxi zones would serve as an acceptable replacement of exact coordinates. In order to provide coordinates for each row in order to conduct spatial analysis, we utilize taxi zone information to calculate centroid coordinates for each zone and combine it to the taxi trip dataset.

This approach now allows coordinate-based spatial analysis methods by adding approximated coordinates of each pickup/dropoff location to any yellow taxi dataset. We will use the yellow taxi trip records of August 2024, as it is the most recent data published by the TLC at the time of this analysis. While the dataset originally has more than 3 million rows, conducting spatio-temporal analysis for dataset of this size has been an overwhelming task for R and its packages. Therefore, we use 1 million randomly sampled and batched rows for better efficiency.

2.3 Data Cleaning

Observations of this dataset are prone to either human or technical errors, often resulting in missing or invalid values in one or more variables. Therefore, instead of keeping only the ’perfect’ observations, we have decided to focus only on the key variables for data cleaning, in order to keep as many samples as possible for spatio-temporal analysis.

For spatial variables, both ”PULocationID” and ”DOLocationID” had valid inputs for all data points. Temporal variables, however, required data cleaning.



```
> summary(ogdata$Duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 -29.12    7.65   12.70   17.24   20.68 5743.92       5 
> summary(ogdata$trip_distance)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
     0.00      1.03      1.80      4.28      3.55 103297.24 
```

Figure 3: Summary output of Duration and trip distance

”Duration” variable was added by calculating time difference between pickup and dropoff time for each taxi trip. Summary output for this variable (figure 3) indicated missing data (shown by the number of NA’s) and abnormal trip records (extreme or negative values). Trip distance served as another reference to detect and remove extreme values, as some data points had unrealistic records (also shown in figure 3), possibly due to technical errors.

3 Exploratory Data Analysis

3.1 Pickup/Dropoff Counts by Location

Total pickup/dropoff counts by taxi zone can be a simple yet effective representation of taxi demand patterns. Visualization can be done by drawing boundaries with taxi zone .shp file and color representing total pickup counts for each zone:

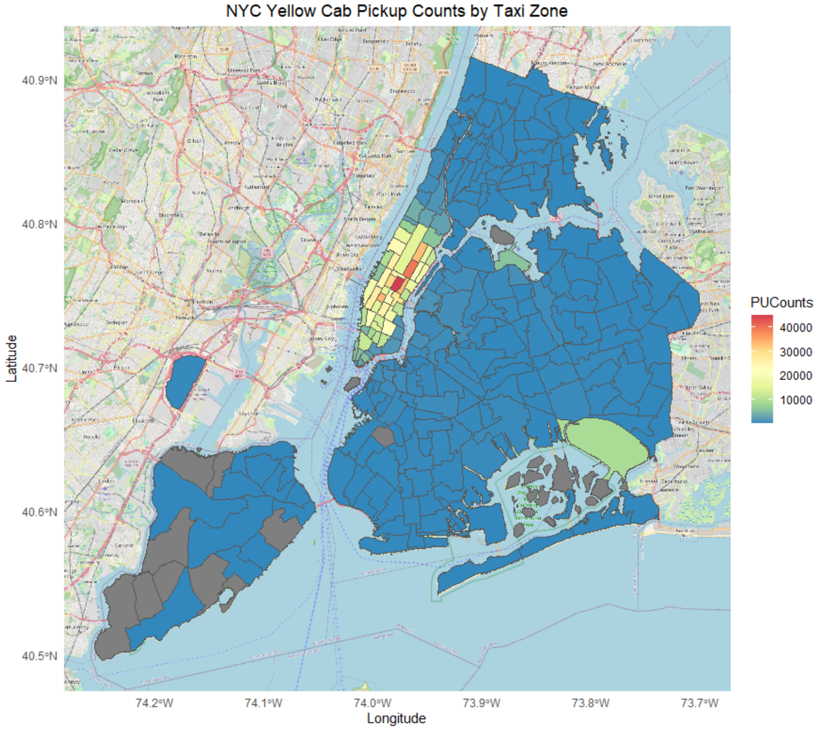


Figure 4: Yellow Taxi Pickup Counts of Aug 2024

Yellow taxi pickups are heavily focused in Manhattan borough, especially midtown Manhattan area. With 'Midtown Center' being the zone with the most pickups of 44892, LaGuardia Airport and JFK Airport are the only two non-Manhattan area with significant volume of pickups.

There are also several taxi zones in gray, due to having no pickup recorded from that zone in this particular dataset. However, it is notable that 'Governor's Island/Ellis Island/Liberty Island' will always have zero pickup counts despite being registered as taxi zones, since these areas can only be accessed by ferry boats.

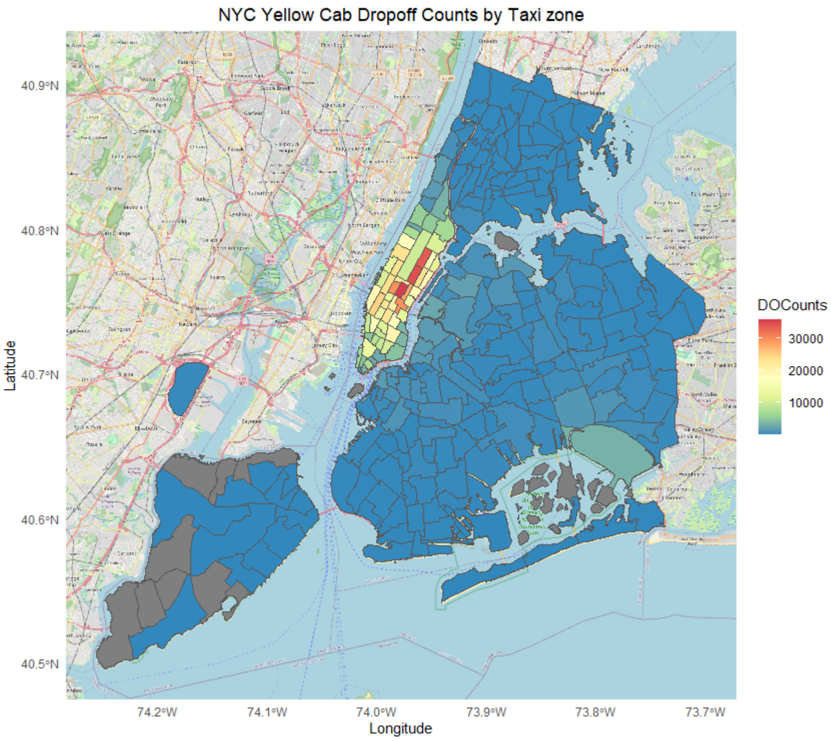


Figure 5: Yellow Taxi Dropoff Counts of Aug 2024

Dropoff counts show similar pattern in Manhattan borough, although more distributed in lesser zones, with Midtown Center boasting the largest counts of dropoffs as well. Outside Manhattan, dropoffs are less concentrated in Airport zones compared to pickups.

3.2 Pickup Counts by Time Periods

Overview of pickup counts based on two types of time period - hour and weekday - is an another great basis of spatio-temporal analysis.

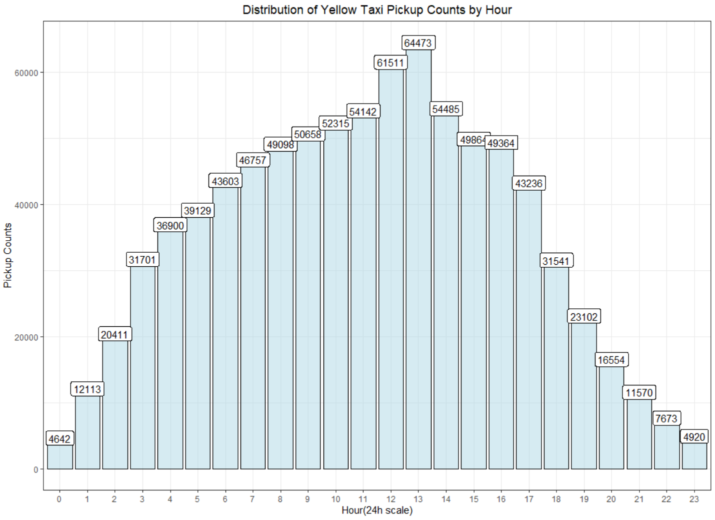


Figure 6: Yellow Taxi Hourly Pickup Counts

Pickup counts between 12PM to 2PM were the most significant for August 2024, while the numbers decline rapidly from 6 PM.

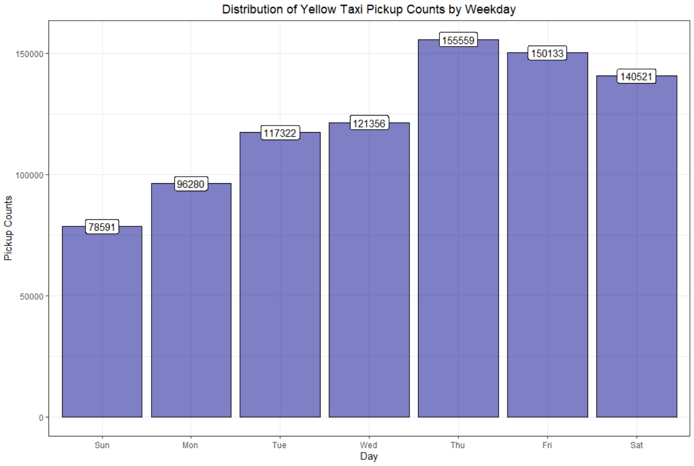


Figure 7: Yellow Taxi Pickup Counts by Weekdays

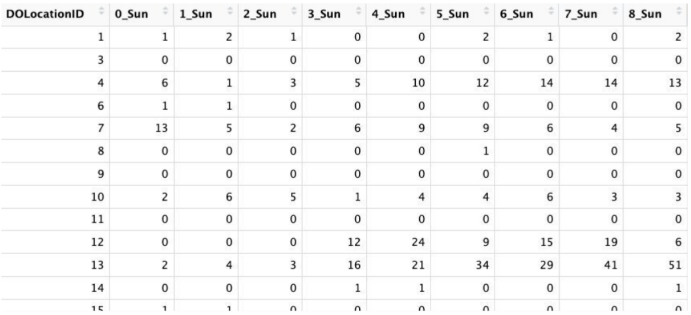
Next, pickups occur the most on Thursdays, and it is notable that there are significantly less pickups on Mondays.

4 Methodology: Spatio-Temporal Clustering

In order to understand the demand for traffic flow based on time, we proceed our analysis by respecting the **Origin-Destination** nature of our data. That is, to understand the patterns of the traffic flow, we perform clustering on pick up information and drop off information separately and match the clusters to examine the flow.

Specifically, each row of the data is a recorded yellow cab trip ride in August of 2024. For each trip, it is recorded where and when the customers were picked up and when and where they were dropped off. Our goal is to use pick up information (zone and time) to cluster each trip into **pick up clusters**, and drop off information (drop off zone and time) to cluster the same trips into **drop off clusters**. Then, have two labels for each trip: pick up clusters and drop off clusters. Using these two cluster information, we can get a much better insight into the traffic flow.

First, the data was wrangled such that the time information only contains the day of the week and hour of the day, grouped by hourly intervals. This was done, because the focus of the analysis is not the exact day of the month or the specific time stamp, but the general demand for ”Friday at 7pm” or ”Sunday at 10am.” After wrangling, for each zone (from 1 to 263), we aggregated the toal number of trips that occurred during a specific time combination, such as ”Sunday at 10am”. Now, the entry for each time and zone can be understood as a ”demand” which is the response for each time t and zone s. Now, the idea is to use cluster each zone into a cluster based on the demand patterns over time, which is aligned with our goal.



DOLocationID	0_Sun	1_Sun	2_Sun	3_Sun	4_Sun	5_Sun	6_Sun	7_Sun	8_Sun
1	1	2	1	0	0	2	1	0	2
3	0	0	0	0	0	0	0	0	0
4	6	1	3	5	10	12	14	14	13
6	1	1	0	0	0	0	0	0	0
7	13	5	2	6	9	9	6	4	5
8	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0
10	2	6	5	1	4	4	6	3	3
11	0	0	0	0	0	0	0	0	0
12	0	0	0	12	24	9	15	19	6
13	2	4	3	16	21	34	29	41	51
14	0	0	0	1	1	0	0	0	1
15	1	1	0	0	0	0	0	0	0

Figure 8: LocationID with Time Combinations

4.1 Model Components

The general spatio-temporal as discussed in lecture is set up as $Y_t(s)$ response, taxi demand, in our case, at spatial location s and time t, given by:

$$Y_t(s) = M_t(s)\beta + w(s,t) + \epsilon_t(s),$$

where:

- $Y_t(s)$: Observed demand (pickup/dropoff counts) at spatial location s and time t.
- $M_t(s)\beta$: Covariates ($M_t(s)$) and their associated coefficients (β), which are not considered in this setup.
- $w(s,t)$: Spatio-temporal random effect capturing patterns over time and space.
- $\epsilon_t(s)$: Error term capturing random noise.

It was also discussed during the lecture that we can use the temporal basis functions $f_1(t), \ldots, f_m(t)$, at each location to model the spatio-temporal random effect $w(s,t)$ such that:

$$w(s,t) = \sum_{i=1}^{m} f_i(t)\psi_i(s),$$

where:

- $f_i(t)$: Temporal basis functions like wavelets that describe temporal patterns.
- $\psi_i(s)$: Spatially varying coefficients that quantify the influence of the i-th temporal basis function at spatial location s.
- m: Number of temporal basis functions.

In this analysis, we omit the $M_t(s)\beta$ as we are not incorporating any covariates other than the space and time information to cluster each zone. Instead, we focus on representing $w(s,t)$ using temporal basis functions and spatially varying coefficients.

4.2 Our Approach: Spatio-Temporal Model with Spatial Constraints

The observed pickup demand $Y(s,t)$ at spatial location s and time t is modeled as:

$$Y(s,t) = f_s(t) + w_s + \epsilon(s,t),$$

where:

- $Y(s,t)$: Observed demand (pickup/dropoff counts) at location s and time t.
- $f_s(t)$: Temporal demand pattern at location s, capturing the temporal variability such as daily or weekly cycles.
- w_s: Spatial constraint, representing the inherent spatial connectivity.
- $\epsilon(s,t)$: Error term capturing random noise or unmodeled variability.

We set up our model this way to connect the idea with the ClustGeo modeling technique [1] which we will discuss in section 4.3.

4.2.1 Temporal Demand Patterns $f_s(t)$

Examine the taxi demand at zone 88 in Manhattan:

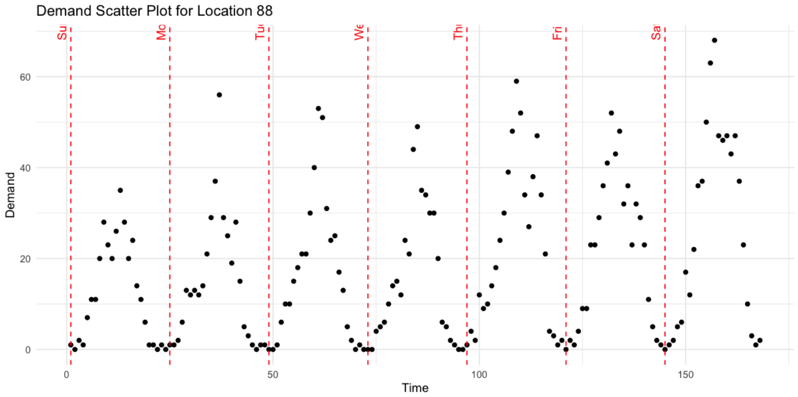


Figure 9: Weekly Demand Patterns at Zone 88: Manhattan

There is a clear cyclic trend of demand over time. The vertical dotted lines are the day of the week, and the cycle of demand correspond directly with the day of the week. Examine a different location, zone 33 from Brookly:

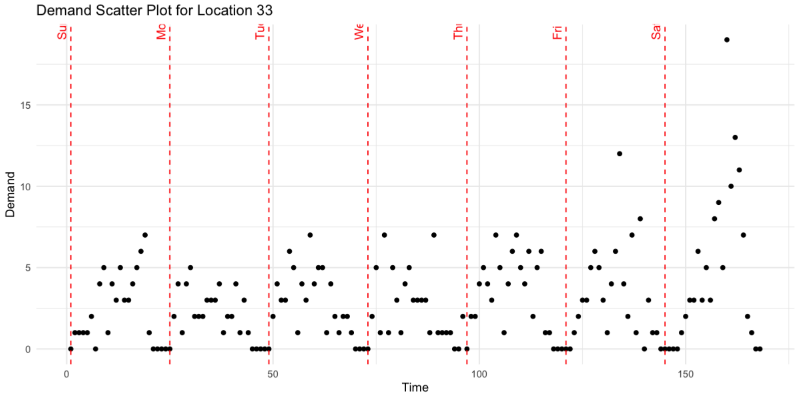


Figure 10: Weekly Demand Patterns at Zone 33: Brooklyn

While the peaks are noticeably smaller, the cycles aligne exactly with the day of the week, similar to zone 88. The temporal demand $f_s(t)$ at location s and time t can be expressed using Fourier transform as:

$$f_s(t) = \beta_{s,0} + \sum_{k=1}^{m} \beta_{s,k} \cos(2\pi k t) + \sum_{k=1}^{m} \gamma_{s,k} \sin(2\pi k t),$$

where:

- $\beta_{s,0}$: The baseline demand at location s.
- $\beta_{s,k}$: Coefficients of the cosine components.
- $\gamma_{s,k}$: Coefficients of the sine components.
- m: The number of frequency components (basis functions) used in the model.

The Fourier Transform is very useful for picking up on the periodic patterns by decomposing a signal $x(t)$ into a sum of sine and cosine waves at different frequencies [4]. Using Euler's formula:

$$e^{-i2\pi ft/N} = \cos\left(\frac{2\pi ft}{N}\right) - i\sin\left(\frac{2\pi ft}{N}\right),$$

the Fourier Transform can be rewritten as:

$$X(f) = \sum_{t=0}^{N-1} x(t)\cdot\cos\left(\frac{2\pi ft}{N}\right) - i\sum_{t=0}^{N-1} x(t)\cdot\sin\left(\frac{2\pi ft}{N}\right).$$

- **Real Part (Re$(X(f))$):** Represents the amplitude of the cosine component at frequency f:

$$\mathrm{Re}(X(f)) = \sum_{t=0}^{N-1} x(t)\cdot\cos\left(\frac{2\pi ft}{N}\right).$$

- **Imaginary Part (Im$(X(f))$):** Represents the amplitude of the sine component at frequency f:

$$\mathrm{Im}(X(f)) = -\sum_{t=0}^{N-1} x(t)\cdot\sin\left(\frac{2\pi ft}{N}\right).$$

The total contribution of frequency f is represented by the magnitude of the Fourier coefficient:

$$|X(f)| = \sqrt{\mathrm{Re}(X(f))^2 + \mathrm{Im}(X(f))^2}.$$

The coefficients of the fourier transform, then correspond with the amplitude of the frequency. Taking the zone 88 as an example:

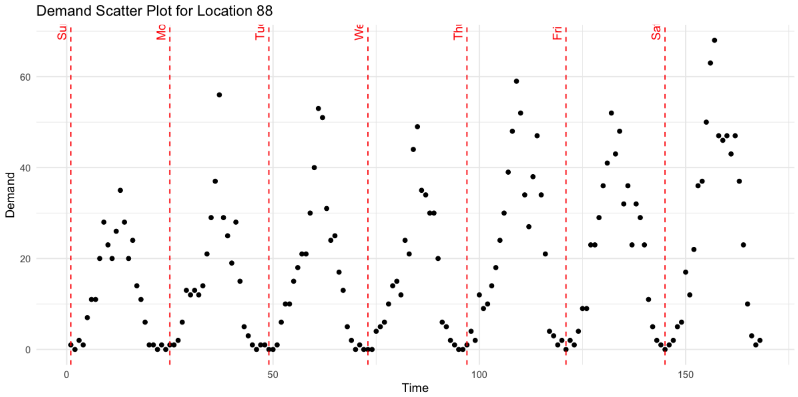


Figure 11: Weekly Demand Patterns at Zone 88

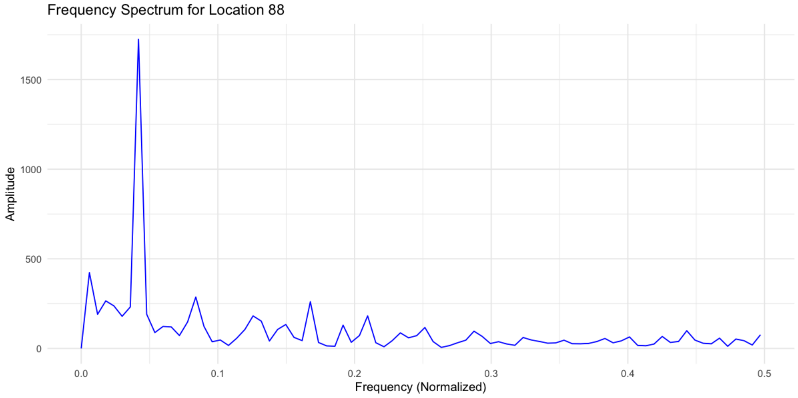


Figure 12: Frequency Spectrum of Zone 88 based on the Weekly Time Series

The highest amplitude indicates a frequency of 1/7 which corresponds to a period of 7 days. That is, the coefficients of the fourier transformed temporal bases are well suited and efficient to describe the patterns of demand at each location.

These temporal coefficients created for each zone are encoded into D_0, which is later used for the ClustGeo package for computation (more on this later).

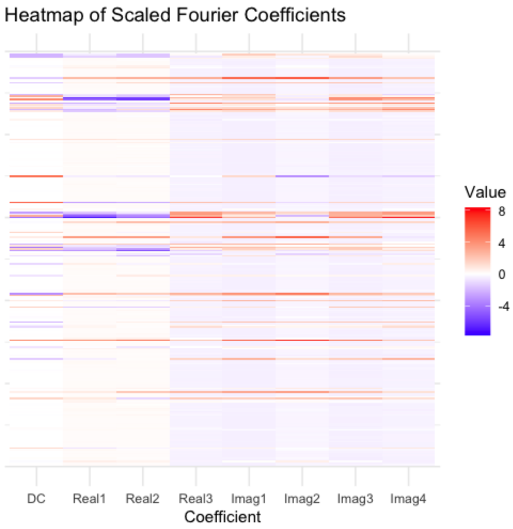


Figure 13: Fourier Coefficients Heatmap

4.2.2 Spatial Smoothing Term: w_s

The spatial effect w_s represents the spatial connectivity between locations s. It is added as a constraint to the temporal bases function $f_s(t)$ to model the $Y(s,t)$ to ensure that clusters reflect both temporal patterns and spatial proximity.

The spatial connectivity w_s is derived from an adjacency structure, using k-nearest neighbors (kNN) graph with k=5 neighbors, which defines which locations are spatially connected. The adjacency structure is encoded in an adjacency matrix W, where:

$$w_{ij} = \begin{cases} 1 & \text{if locations } s_i \text{ and } s_j \text{ are spatially connected,} \\ 0 & \text{otherwise.} \end{cases}$$

This spatial connectivity matrix W is then used to construct a spatial dissimilarity matrix D_1 that penalizes clusters splitting spatially adjacent locations.

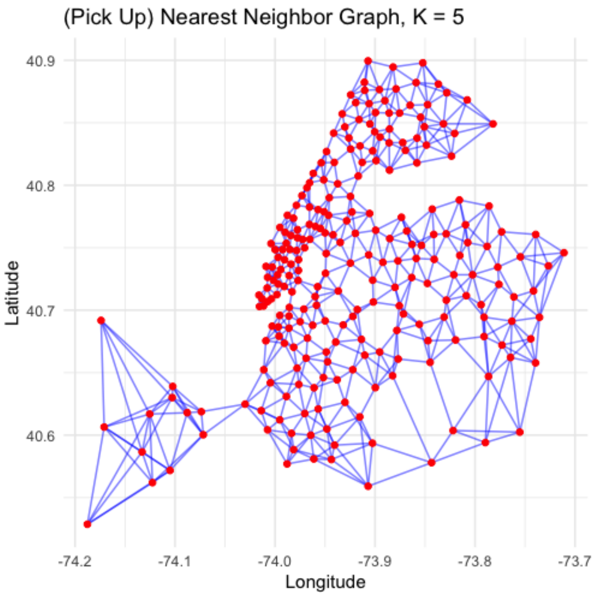


Figure 14: KNN Graph with 5 neighbors

4.3 Implementation using ClustGeo

For clustering NYC taxi zones based on temporal demands, the ClustGeo package was used, leveraging both temporal patterns and spatial continuity. The framework integrates temporal demand features derived from Fourier coefficients and spatial constraints from a kNN graph.

- Feature-Based Dissimilarity (D_0): Derived from the temporal basis coefficients $\psi_i(s)$, D_0 measures the dissimilarity of periodic demand patterns across locations.
- Spatial Dissimilarity (D_1): Constructed from the adjacency matrix W of the kNN graph, D_1 enforces spatial continuity by penalizing splits in spatially connected locations.

The combined dissimilarity is defined as:

$$\Delta_\alpha = (1-\alpha)D_0 + \alpha D_1,$$

where $\alpha \in [0,1]$ controls the trade-off between temporal similarity and spatial continuity.

4.3.1 Choosing α

The parameter α determines the amount of constraint D_1 we will place on clustering so that when $\alpha = 0$, the clustering is driven entirely by temporal patterns (D_0), and when $\alpha = 1$, the clustering prioritizes spatial continuity (D_1).

α can be chosen using the `choicealpha()` function in ClustGeo, which evaluates the quality of clustering for different values of α.

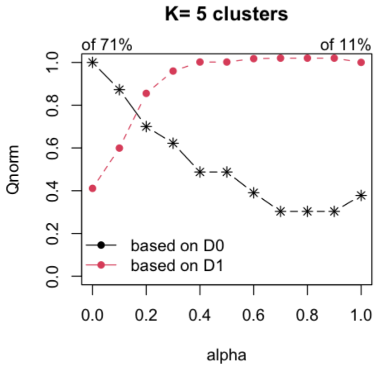


Figure 15: Pick Up Alpha

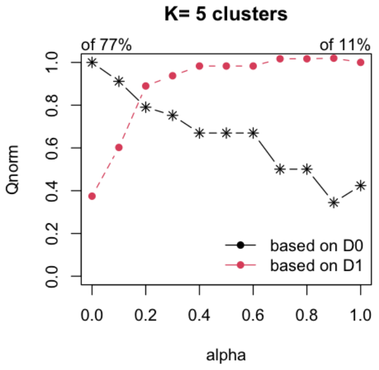


Figure 16: Drop Off Alpha

We have chosen $\alpha = 0.2$ for both pick up and drop off clustering to ensure a meaningful compromise between temporal patterns and spatial coherence.

5 Result

5.1 Pick-Up Spatio-Temporal Clustering

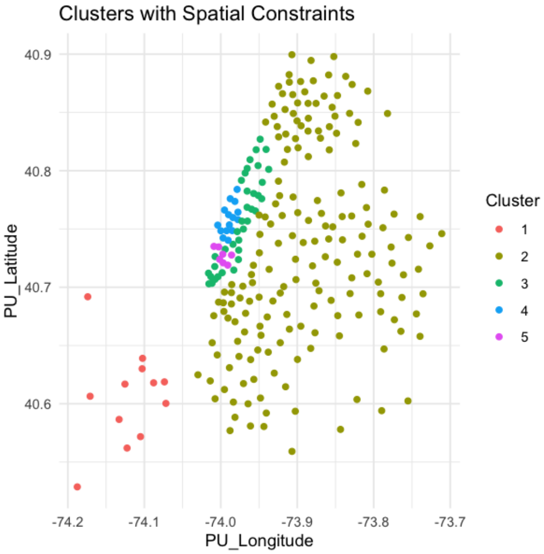


Figure 17: Clustered Pick Up Zones

Notice that the clusters that were created based on temporal demand clustered three areas within Manhattan as different groups and yet combined a lot of the zones around Manhattan into one group.

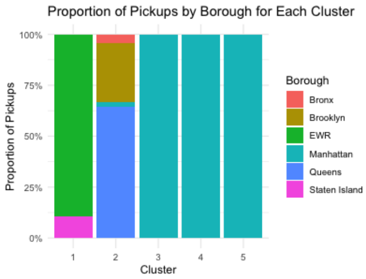


Figure 18: Clustered Pick Up Zones by Borough

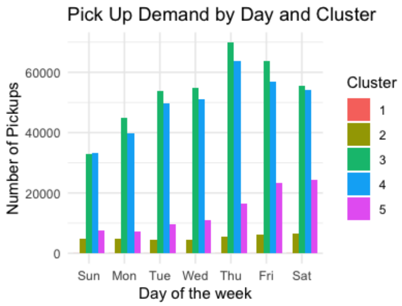


Figure 19: PU Clusters: Hour

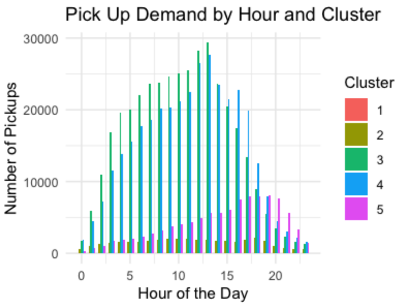


Figure 20: PU Clusters: Day

Here, it is difficult to see the time trends within each cluster, because most of the demand came from the zones within Manhattan, and the three clusters in Manhattan have very similar time trends.

5.2 Drop-Off Spatio-Temporal Clustering

Surprisingly, the drop off clustering resulted in the same type of clusters where most of the demand for drop off came from within Manhattan.

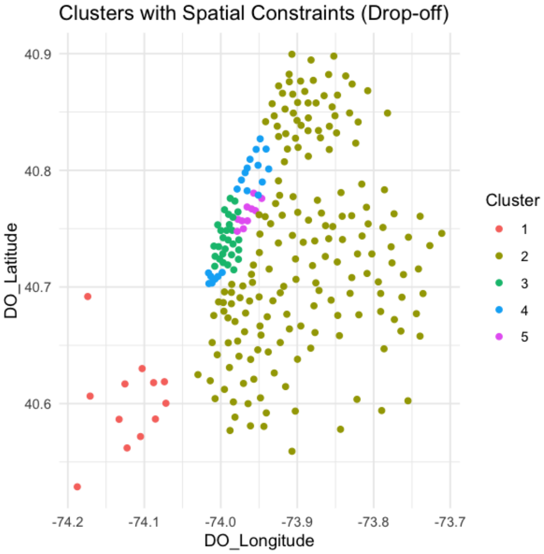


Figure 21: Clustered Drop Off Zones

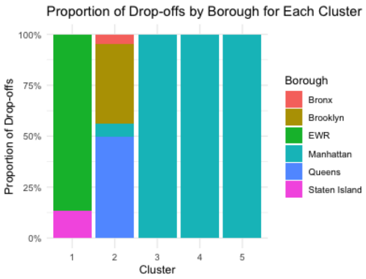


Figure 22: Clustered Drop Off Zones by Borough

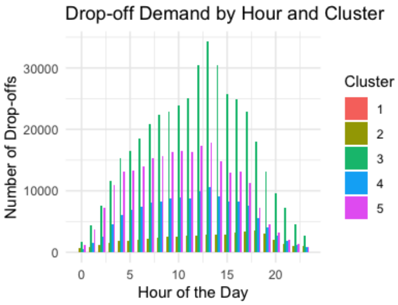


Figure 23: DO Clusters: Hour

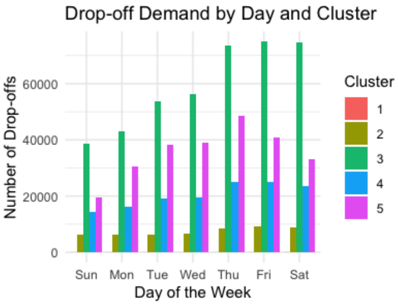


Figure 24: DO Clusters: Day

Here again, the time trends are difficult to see within the clusters of Manhattan, but there seems to be more trend clustered than the pick up clusters.

6 Discussion

6.1 Traffic Flow

Since the five clusters based on pick up and drop off had similar spatial characteristics, it is easy to examine the traffic flow by matching the clusters.

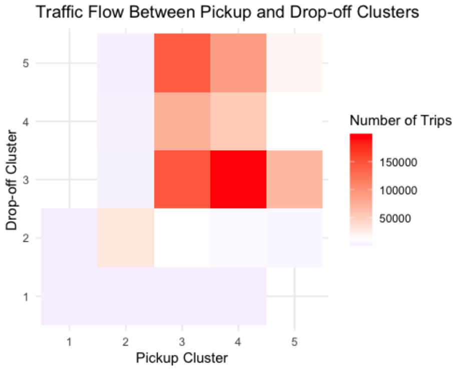


Figure 25: Matched Clusters

Clearly, the highest demands come from inner Manhattan rides. Since our goal is to capture the traffic flow, we can examine the **into Manhattan**, **out of Manhattan**, and **within Manhattan** traffic flows based on the pick up and drop off clusters to exmaine any time patterns.

6.1.1 Into Manhattan

To examine the traffic flow into Manhattan, we can match the pick up clusters 1, 2 with drop off clusters 3, 4, 5:

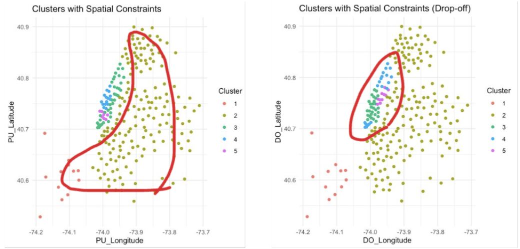


Figure 26: Into City Traffic Flow

Based on the traffic flow into the Manhattan area, we see the highest demand around dinner time on Saturday, which could be the traffic into the bar area on the weekend.

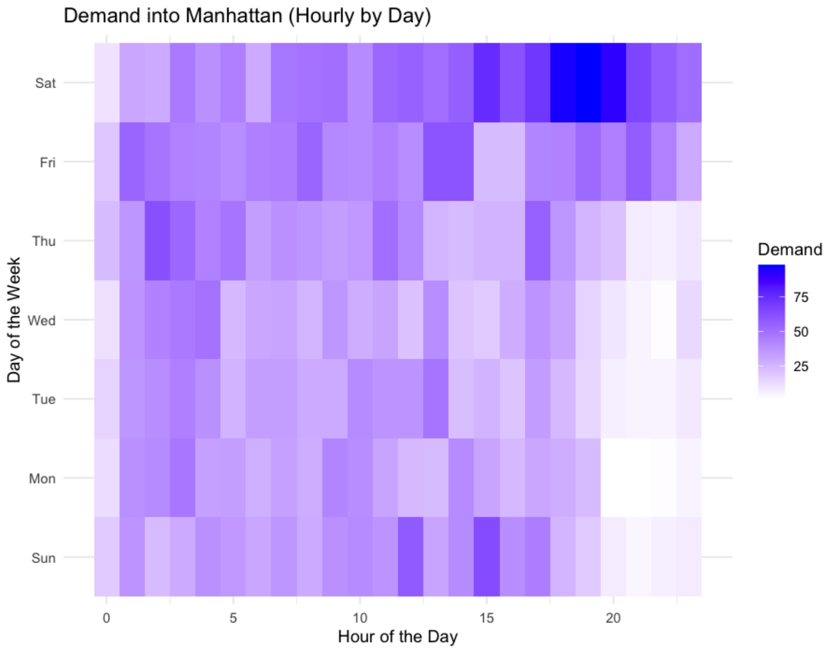


Figure 27: Into Manhattan Demand Time Heatmap

6.1.2 Out of Manhattan

To examine the traffic flow into Manhattan, we can match the pick up clusters 3,4, 5 with drop off clusters 1,2 :

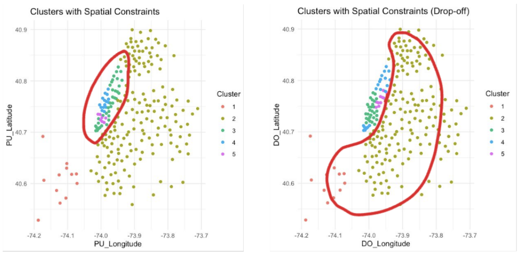


Figure 28: Out of City Traffic Flow

Similarly, the time trends can be examined:

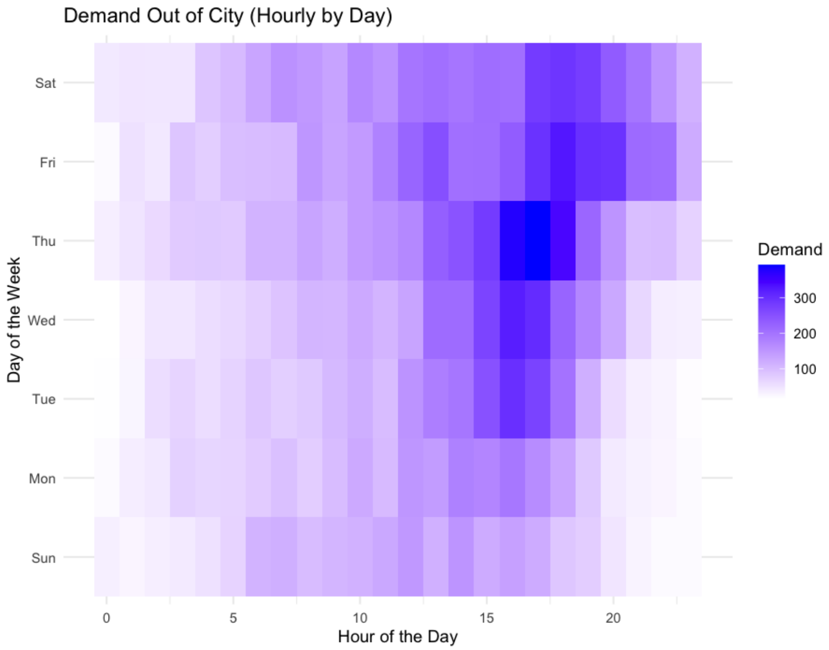


Figure 29: Out of Manhattan Demand Time Heatmap

Interestingly, the out of Manhattan demand is the highest around dinner time on week days, which could explain the traffic out of Manhattan after working hours, commuting out of Manhattan.

6.1.3 Within Manhattan

Finally, to examine the traffic flow within Manhattan, we can match the pick up clusters 3,4, 5 with drop off clusters 3,4,5:

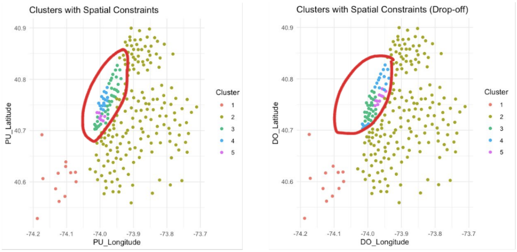


Figure 30: Within City Traffic Flow

The within city traffic demand shows the highest demand on Thursday around dinner time. This corresponds with our initial examination in the exploratory data analysis, where we saw the highest demand for pick up to be on Thursdays around dinner time. Since most of the trips occurred within Manhattan, it makes sense that the overall highest demand aligns perfectly with the highest demand within Manhattan.

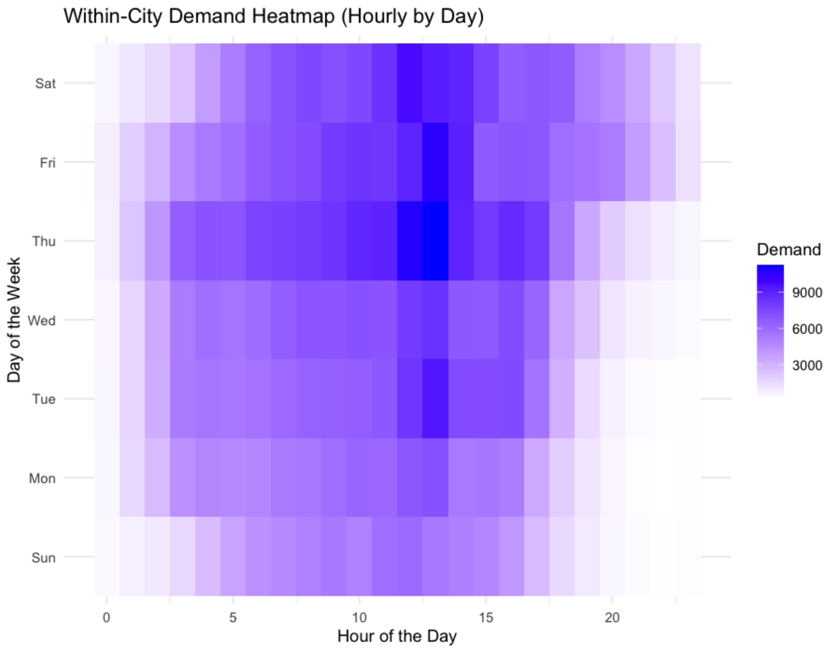
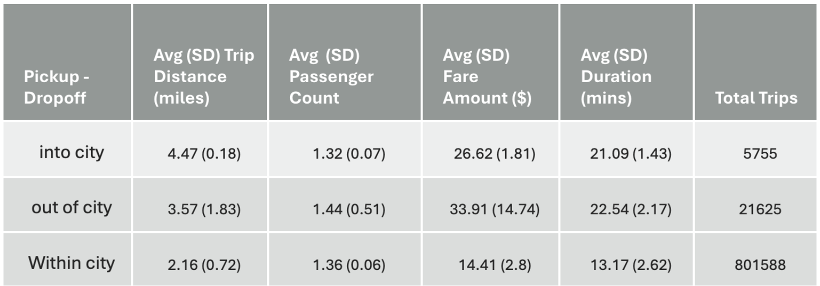


Figure 31: Within Manhattan Demand Time Heatmap

6.2 Conclusion

We examined the traffic flow into, out of, and within Manhattan to understand the patterns of yellow cab demand in NYC. Using spatio-temporal clustering, we revealed the temporal patterns of demand to ride into the city and out of the city, which was not immediately visible based on our initial exploratory data analysis.



Pickup - Dropoff	Avg (SD) Trip Distance (miles)	Avg (SD) Passenger Count	Avg (SD) Fare Amount ($)	Avg (SD) Duration (mins)	Total Trips
into city	4.47 (0.18)	1.32 (0.07)	26.62 (1.81)	21.09 (1.43)	5755
out of city	3.57 (1.83)	1.44 (0.51)	33.91 (14.74)	22.54 (2.17)	21625
Within city	2.16 (0.72)	1.36 (0.06)	14.41 (2.8)	13.17 (2.62)	801588

Figure 32: Summary Table based on Traffic Flow

Our summary table highlights that the demand for ride within the city was the highest, which explains why the traffic demand was not immediately visible in the inital examination. Furthermore, the summary shows that the average fare was the most expensive going out of the city, and within the city, the fares were relatively cheap because of the shorter travel distance.

References

[1] Marie Chavent et al. "ClustGeo: An R package for hierarchical clustering with spatial constraints". In: *arXiv preprint arXiv:1707.03897* (2017). DOI: 10.48550/arXiv.1707.03897. URL: https://arxiv.org/abs/1707.03897.

[2] City of New York. *NYC Taxi Zones.* https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc. Accessed: 2024-11-12.

[3] NBC New York. *Green Cabs Are Being Phased Out — Here's What Will Replace Them.* https://www.nbcnewyork.com/news/local/green-cabs-are-being-phased-out-heres-what-will-replace-them/4302496/. Accessed: 2024-11-12.

[4] Gilbert Strang. *Fourier Transform.* Accessed: November 19, 2024. 2018. URL: https://math.mit.edu/~gs/cse/websections/cse41.pdf.