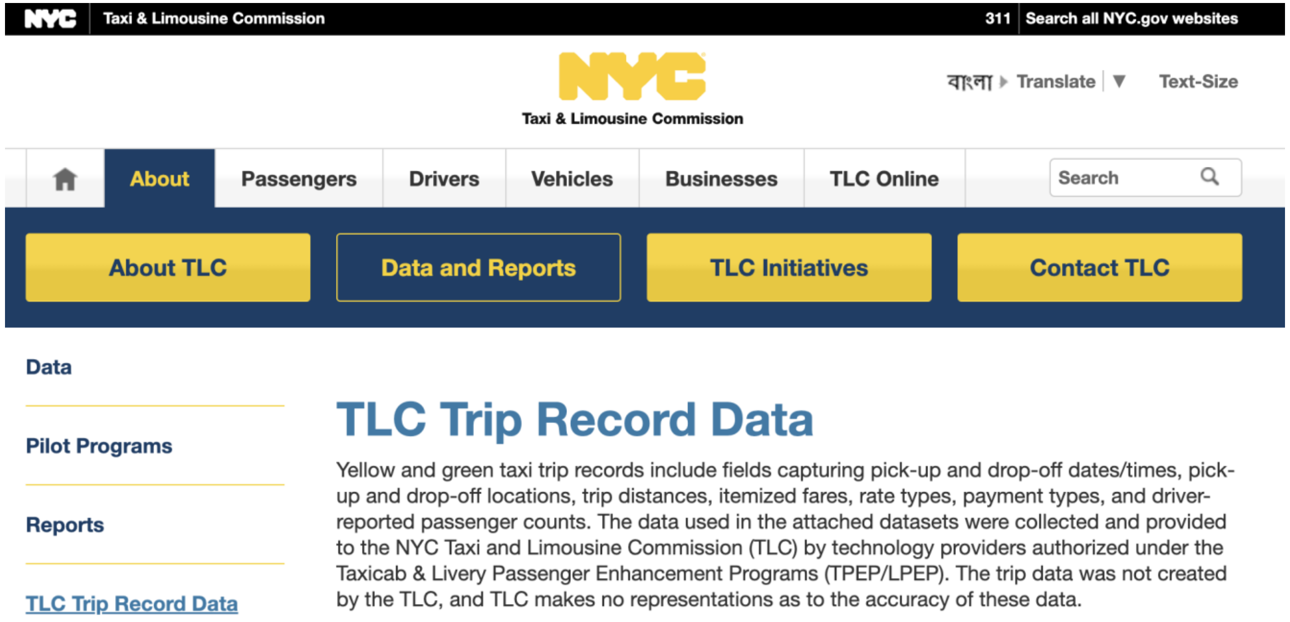
NYC Yellow Cab Demand Analysis

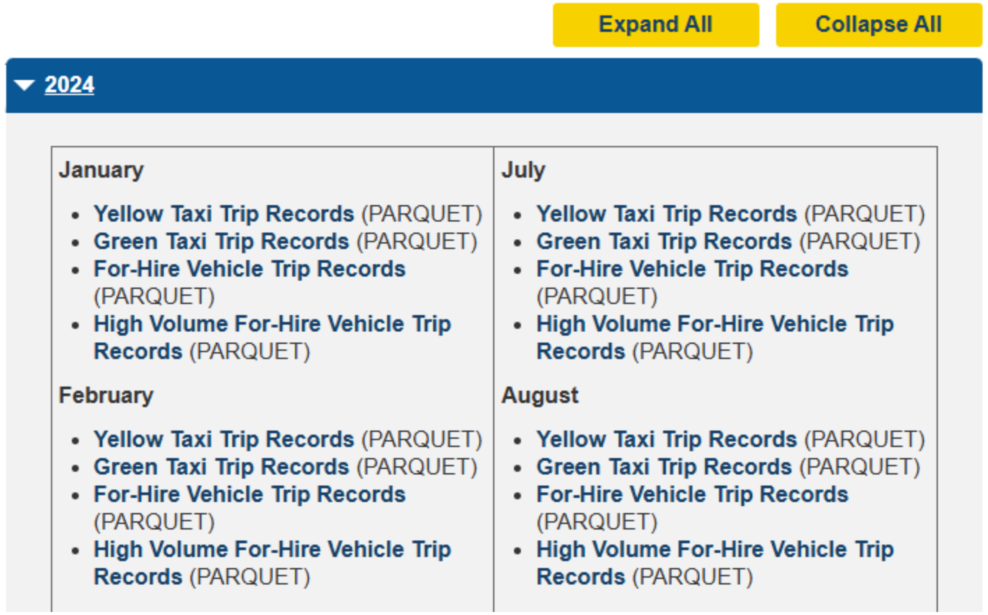
Stephen Cho,

Minjee Kim

Introduction

- Objective: analyzing taxi demand and taxi traffic flow in New York City
- Dataset: Yellow taxi trip record data of Aug 2024 (provided by the NYC TLC)
- Data published on the TLC website, separated by year, month and vehicle type





Vehicle Types

Yellow Taxis

Green Taxis

For-hire Vehicles (FHVs)

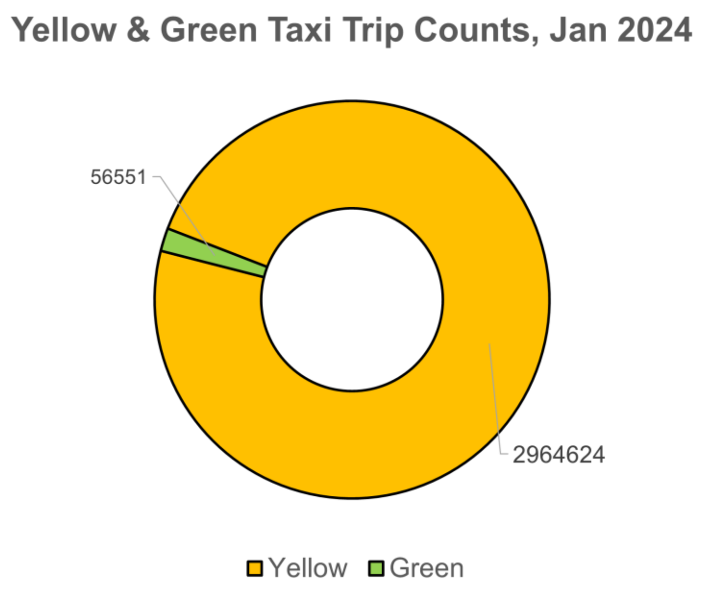
High-volume FHVs

- "Traditional" taxi (respond to street hails)
- More reliable data collection system (collected by TLC-authorized technology providers)

- Vehicles do not respond to street hails
- Data collected & provided by third-party corporations

☑ Target narrowed down to Yellow and Green Taxi trip records

Vehicle Types (continued)



- Green taxi has very small trip counts (2% of total taxi trips)
- Mainly covers outer boroughs (cannot pick up new passengers in "yellow zone")
- **Green Taxi trip record data does not fit our purpose, and is neglectable in size**

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

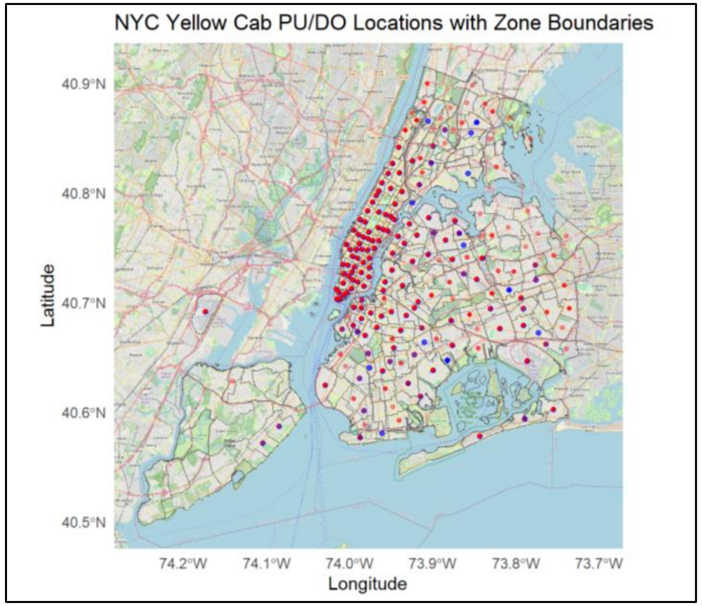
Data Dictionary – Yellow Taxi Trip Records May 11, 2022 Page 1 of 2

This data dictionary describes yellow taxi trip data. For a dictionary describing green taxi data, or a map of the TLC Taxi Zones, please visit http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**

- Raw dataset (old version; left) consists of 18 columns, including key variables representing temporal (Pickup/Dropoff Date & Time) and spatial (Pickup/Dropoff coordinates) information.
- Each row is a yellow taxi trip record
- The TLC has replaced pickup/dropoff location details with "taxi zone" ID information for records since 2011
- Our goal is to analyze recent taxi demand patterns; need to work with the new format by generating (approximate) coordinates to perform spatial analysis

Yellow Taxi Data (continued)



	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

PULocationID <int>	DOLocationID <int>	PU_Longitude <dbl>	PU_Latitude <dbl>	DO_Longitude <dbl>	DO_Latitude <dbl>
237	161	-73.96563	40.76862	-73.97770	40.75803
100	186	-73.98879	40.75351	-73.99244	40.74850
161	114	-73.97770	40.75803	-73.99738	40.72834
100	13	-73.98879	40.75351	-74.01608	40.71204
75	75	-73.94575	40.79001	-73.94575	40.79001
163	162	-73.97757	40.76442	-73.97236	40.75669

- The TLC also provides taxi zone details; great asset for calculating centroid coordinates and visualization
- NYC is divided into <u>263 taxi zones</u>; centroid coordinates are acceptable alternative for exact coordinates
- Columns have shape & geometric information, zone name, location ID, borough name
- *PU_Longitude/Latitude*, *DO_Longitude/Latitude* columns, each working as a pair, are mutated and merged to the Yellow Taxi Trip Dataset with *PULocationID/DOLocationID* used as reference

☑ <u>All rows now have coordinates of pickup/dropoff locations</u>

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
```

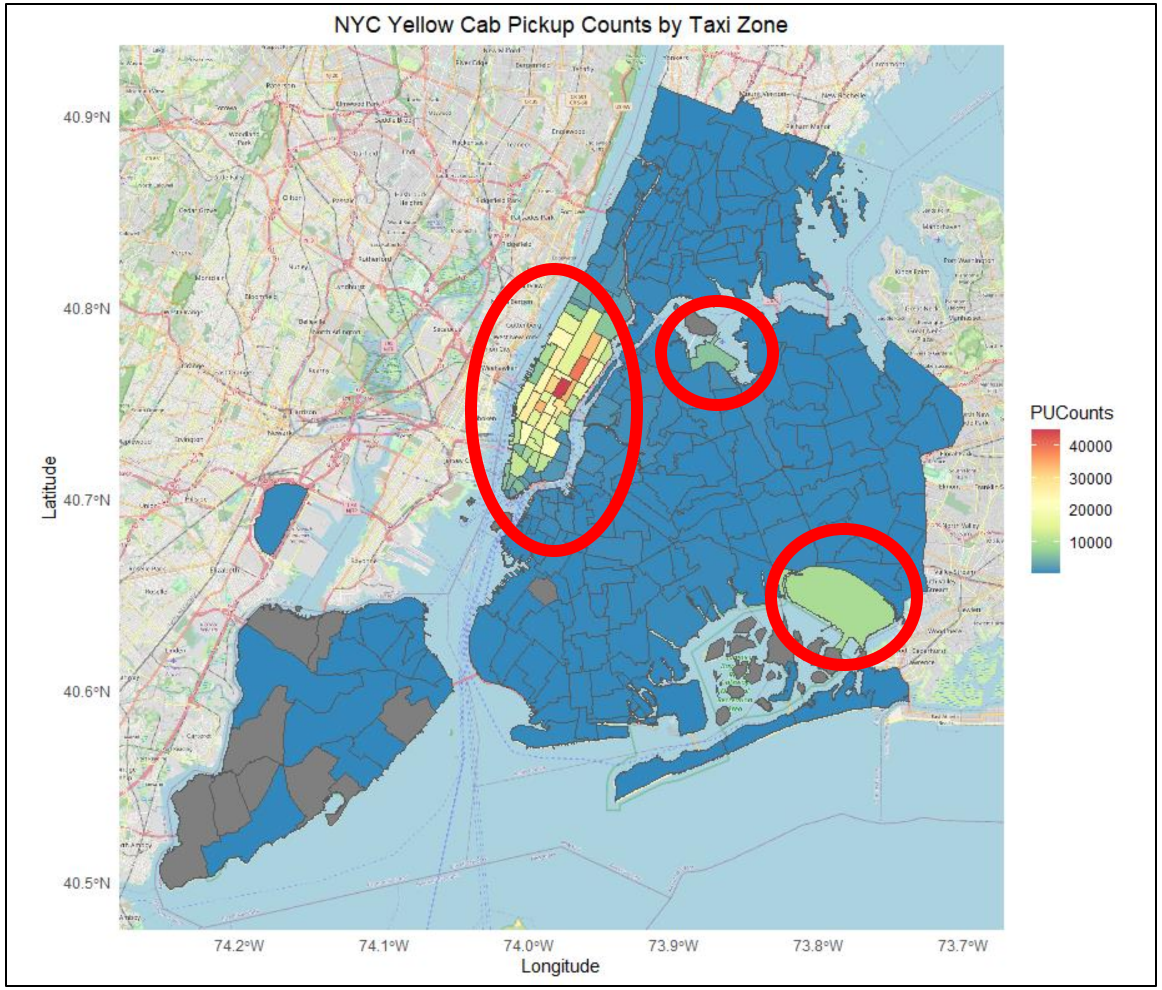
- Original dataset has 3 million rows; used 1 million randomly extracted samples for efficiency
- Most variables often have missing/unusual values; only considered spatial & temporal variables for cleaning
- Spatial variables (PULocationID, DOLocationID) are intact for all rows; temporal variables more vulnerable
- "*Duration*": Gap between pickup and dropoff time (new variable); negative or extreme values removed
- "*trip_distance*": Extreme values removed (by IQR method)
- 859762 rows remain after data cleaning

Exploratory Data Analysis

1. Pickup/Dropoff Counts by Location

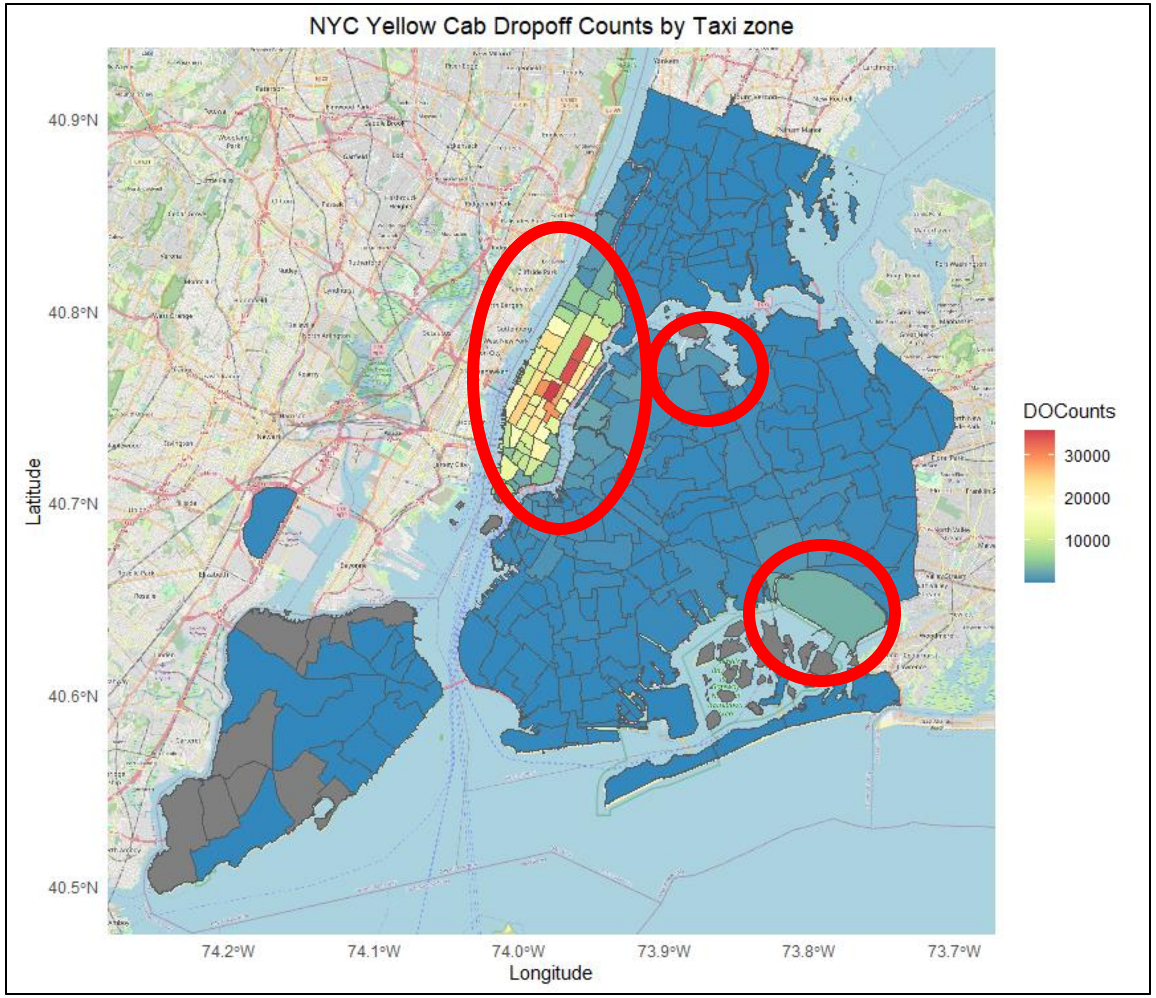
2. Pickup Counts by Time Periods

Pickup Counts by Taxi Zone



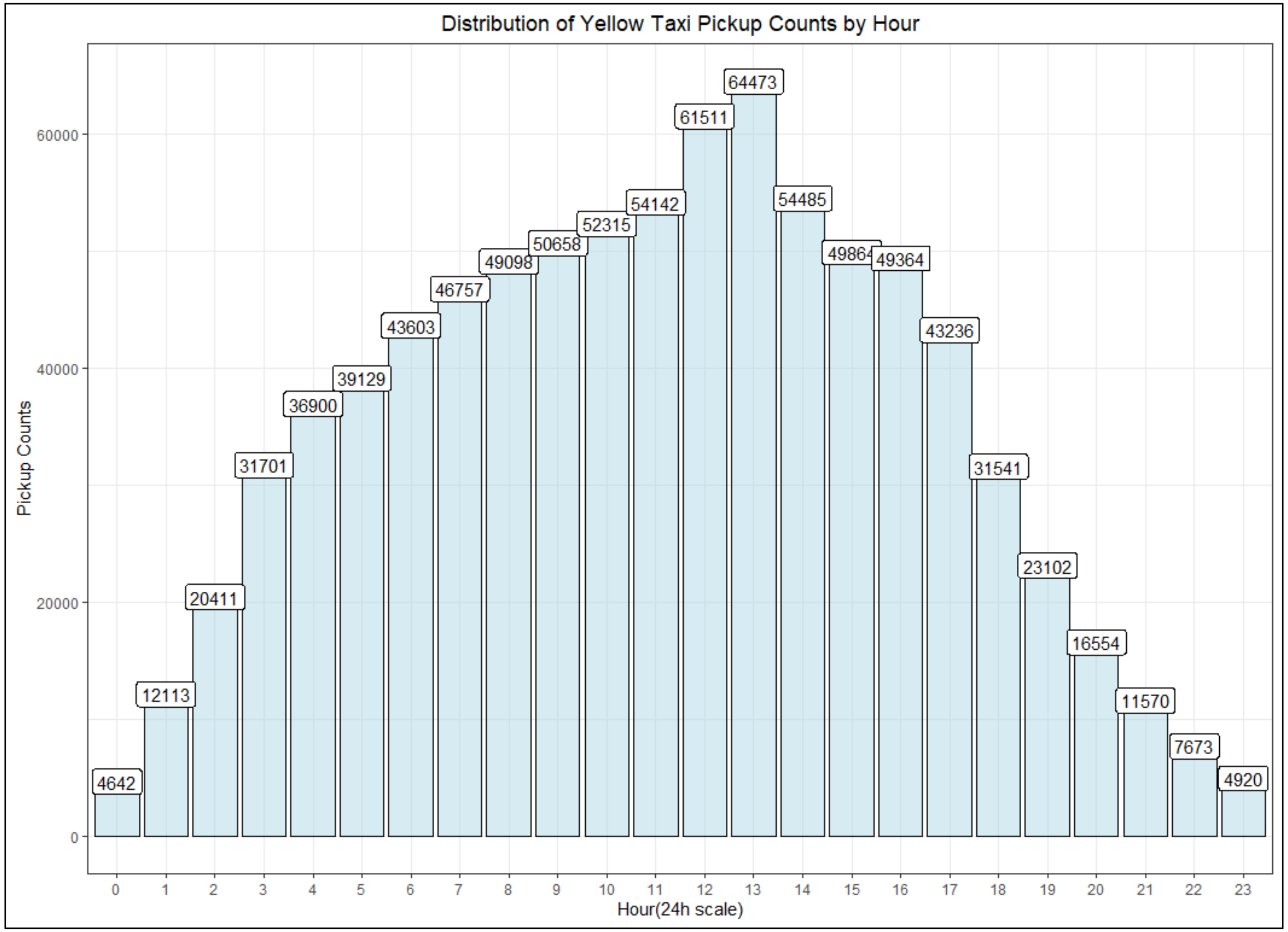
- Pickups are heavily focused in Manhattan borough, especially midtown Manhattan area
- 'Midtown Center' has the most pickups of 44892
- LaGuardia Airport and JFK Airport are the only two non-Manhattan area with significant volume of pickups
- Taxi zones in gray have no pickup recorded
- "Governor’s Island/Ellis Island/Liberty Island" always have zero pickup counts since these areas can only be accessed by ferry boats

Dropoff Counts by Taxi Zone



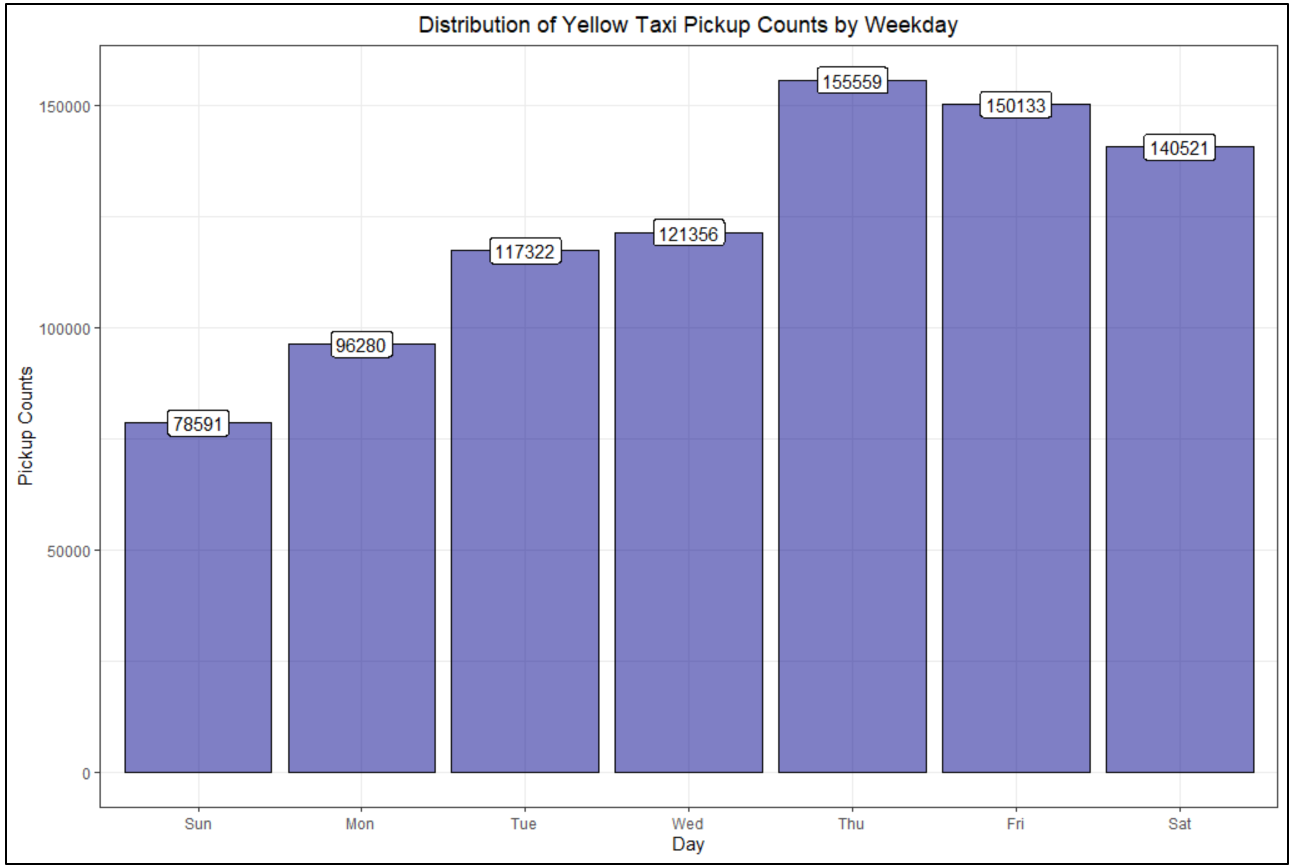
- Dropoffs are also heavily focused in midtown Manhattan area, although slightly more spread out to other zones
- 'Midtown Center' also has the most dropoffs of 35758
- Outside Manhattan, dropoffs are less concentrated in LaGuardia Airport and JFK Airport
- Gray zones exist for dropoff counts as well

Pickup Counts by Hour



- 12 - 1PM has most pickup counts
- Pickup counts decline rapidly from 5 PM

Pickup Counts by Weekday



- Thursday has the most pickup counts
- Significantly less pickups on Mondays

Objective

- To understand the general traffic flow based on demands
- To capture the traffic flow from outer areas into the city during commuting hours and movements into the bar area during night times

Work Flow

Pick-Up Demand Clustering → **Drop-Off Demand Clustering** → **Match Clusters**

- Pickup Location
- Pickup Time

PULocationID	Pick Up Cluster
1	1
3	2
4	3
6	1
7	2
8	2

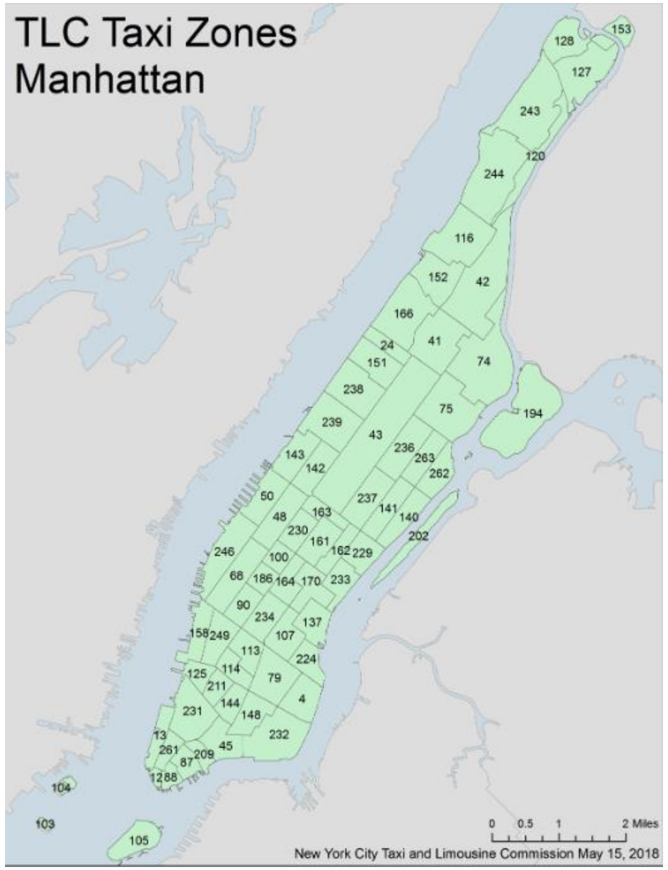
- Drop off Location
- Drop off Time

DOLocationID	Drop Off Cluster
1	1
3	2
4	3
6	1
7	2
8	2

Pick Up Cluster	Drop Off Cluster	Demand
4	3	199527
3	3	146144
3	5	142982
4	5	88819
3	4	75719
5	3	66672

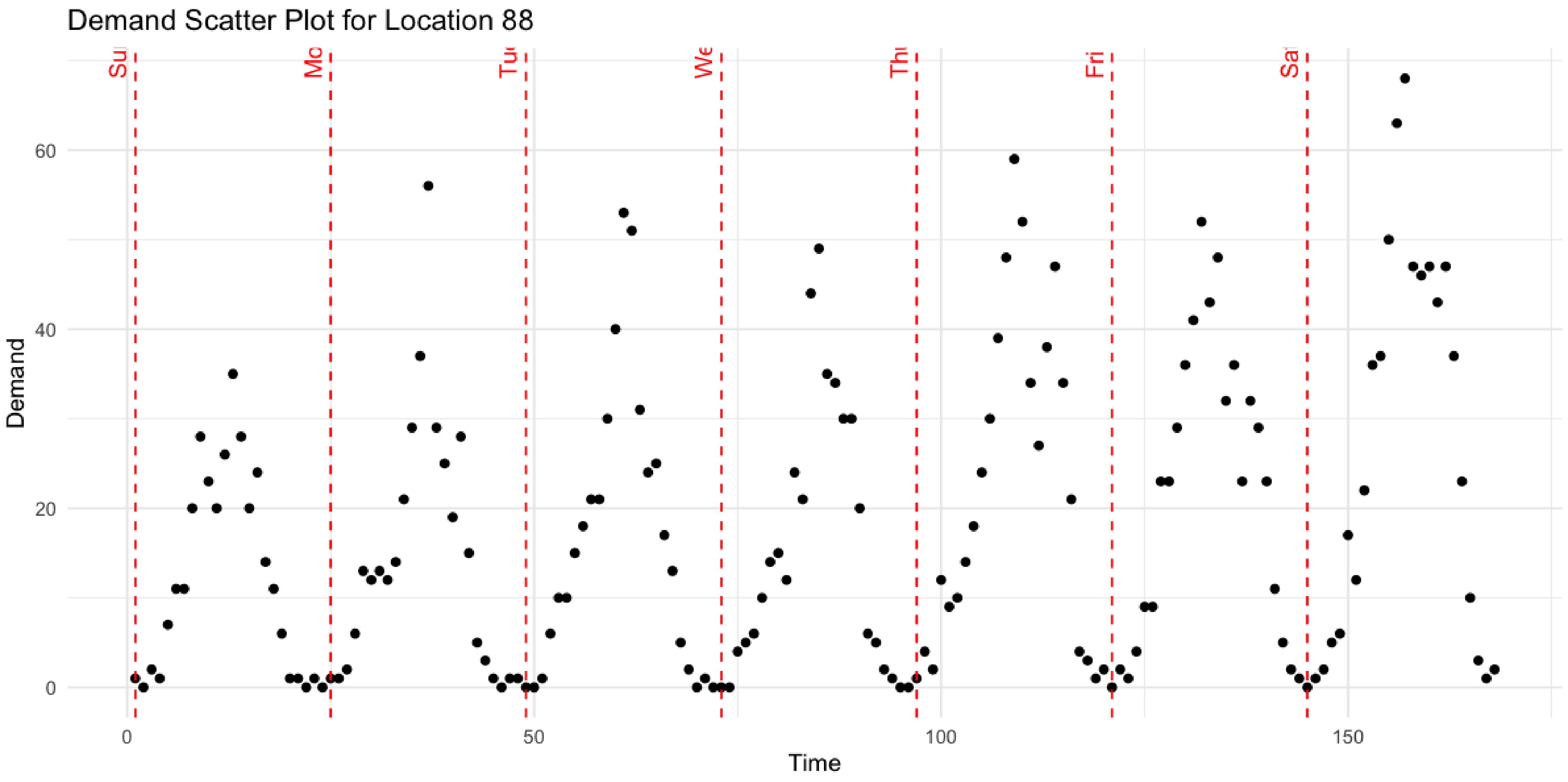
Preparing the Data

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



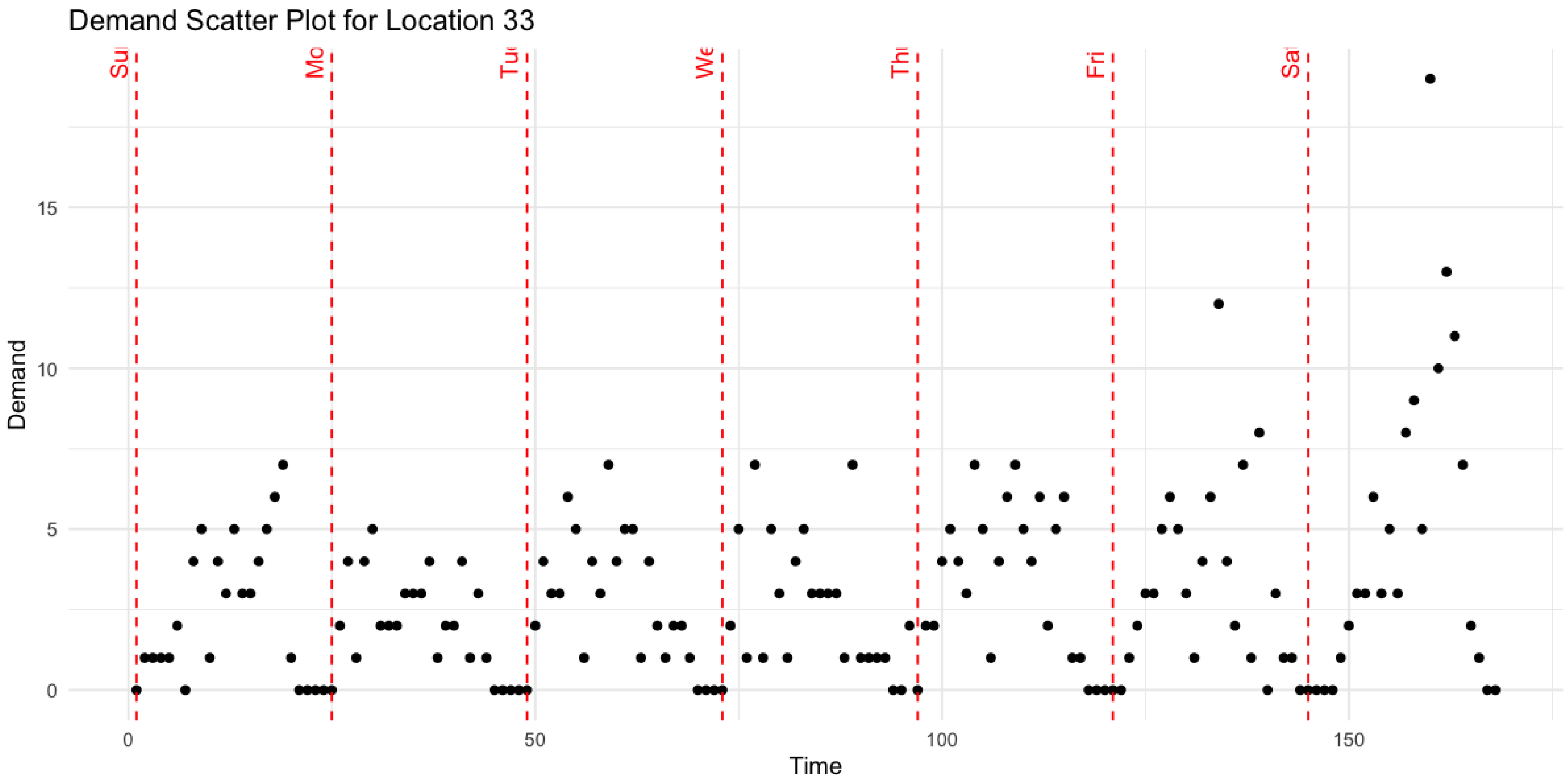
Cyclic Time Behavior

Zone 88 in Manhattan

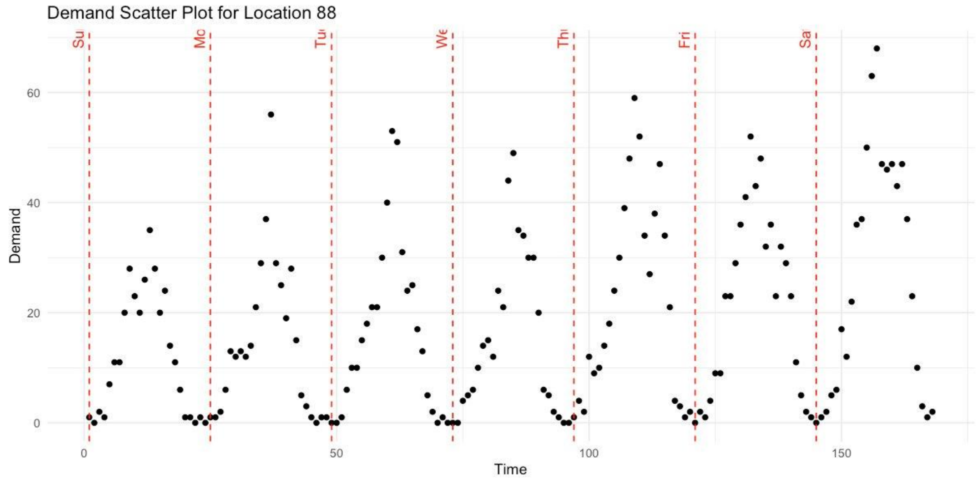


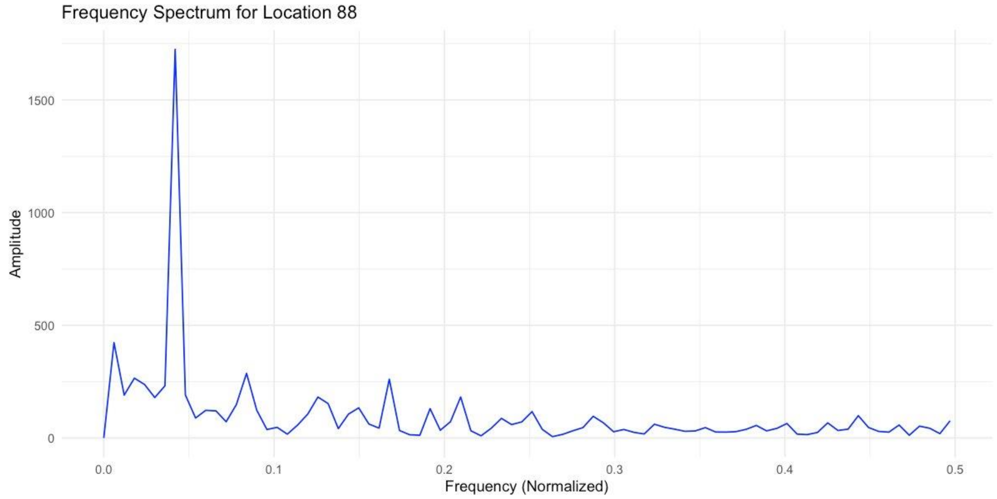
Cyclic Time Behavior

Zone 33 in Brooklyn



Fourier Transform on Time Series





(Lecture 7 – SpaceTime-Discrete)

A multi-resolution wavelet decomposition of a function $f_s(t)$ is an expression of the following form:

$$f_s(t) \approx \beta_{s,00}\phi_{00}(t) + \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{j-1}} \beta_{s,j,k}\psi_{j,k}(t)$$

$\beta_{s,00}$ is the scaling coefficient. The wavelets $\psi_{j,k}(t)$ are generated from a single wavelet $\psi(t)$, the so-called mother wavelet, by scaling and translation. The form of basis functions are known.

Fourier transform is a special case when $\psi(t) = e^{-2i\pi t}$

The temporal demand fs(t) can be represented by Fourier coefficients:

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

Our Model

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

$Y(s,t)$: Observed demand (pickup 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.

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

(Lecture 7 – SpaceTime-Discrete)

$$Y(\mathrm{s},t) = \mathrm{M}(\mathrm{s},t)'\boldsymbol{\beta} + w(\mathrm{s},t) + \epsilon(\mathrm{s},t)$$

for $\mathrm{s} \in D$ and $t \in [0, T]$.

$\mathrm{M}(\mathrm{s},t)$ are local space-time covariate vectors

$\boldsymbol{\beta}$ is an associated coefficient vector

$w(\mathrm{s},t)$: spatial temporal random effect.

ϵ's are pure error terms.

Use temporal basis $f_1(t), \cdots, f_m(t)$: $w(\mathrm{s},t) = \sum_{i=1}^{m} f_i(t)\psi_i(\mathrm{s})$
we need to estimate spatially varying basis coefficients $\psi_i(\mathrm{s})$ (spatial functional data analysis)

Methodology (Chavent, 2017)

Aggregation measure based on the combined dissimilarity of two points i and j:

$$\delta_\alpha(\{i\},\{j\}) = (1-\alpha)\frac{w_i w_j}{w_i + w_j} d^2_{0,ij} + \alpha \frac{w_i w_j}{w_i + w_j} d^2_{1,ij}.$$

Squared dissimilarity based on features

Squared dissimilarity based on spatial relationship

In matrix form:

$$\Delta_\alpha = (1-\alpha)\Delta_0 + \alpha\Delta_1.$$

- **D0:** captures how different two locations are based on their feature patterns
- **D1:** captures how geographically unconnected two locations are based on spatial adjacency
- When alpha = 0, the feature coefficients are clustered without any spatial smoothing

Standard Ward's Method: $I(\mathcal{C}_k) = \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \frac{w_i w_j}{2\mu_k} d^2_{ij}$

$$I_\alpha(\mathcal{C}^\alpha_k) = (1-\alpha) \sum_{i \in \mathcal{C}^\alpha_k} \sum_{j \in \mathcal{C}^\alpha_k} \frac{w_i w_j}{2\mu^\alpha_k} d^2_{0,ij} + \alpha \sum_{i \in \mathcal{C}^\alpha_k} \sum_{j \in \mathcal{C}^\alpha_k} \frac{w_i w_j}{2\mu^\alpha_k} d^2_{1,ij},$$

Methodology (Chavent, 2017)

Model Component	ClustGeo Component	Description
$f_s(t)$: Temporal demand patterns	Feature-based dissimilarity D_0	Temporal patterns (Fourier coefficients) are used to compute D_0, capturing pairwise dissimilarities in temporal behavior across locations.
w_s: Spatial effect	Spatial dissimilarity D_1	Spatial relationships (from adjacency or proximity) are encoded in D_1, penalizing clusters that split geographically connected locations.
$\epsilon(s,t)$: Random noise	Not explicitly modeled	ClustGeo assumes that noise is minor compared to the signal in D_0 and D_1.
Combined effects	Combined dissimilarity Δ_α	$\Delta_\alpha = (1-\alpha)D_0 + \alpha D_1$ balances temporal and spatial effects in the clustering process.

Using ClustGeo in R

Hierarchical clustering with soft contiguity constraint.

The function `hclustgeo` implements a Ward-like hierarchical clustering algorithm with soft contiguity constraint. The main arguments of the function are:

- a matrix `D0` with the dissimilarities in the "feature space" (here socio-economic variables for instance).
- a matrix `D1` with the dissimilarities in the "constraint" space (here a matrix of geographical dissimilarities).
- a mixing parameter `alpha` between 0 an 1. The mixing parameter sets the importance of the constraint in the clustering procedure.
- a scaling parameter `scale` with a logical value. If `TRUE` the dissimilarity matrices `D0` and `D1` are scaled between 0 and 1 (that is divided by their maximum value).

The function `choicealpha` implements a procedure to help the user in the choice of a suitable value of the mixing parameter `alpha`.

Both `hclustgeo` and `choicealpha` can be combined to find a partition of the $n = 303$ French municipalities including geographical contiguity constraint. The two steps of the procedure are :

1. Find partition in K clusters of the 303 municipalities using the dissimilarity matrix `D0`. The clusters of this partition are homogeneous on the socio-economic variables and no contiguity constraint is used.
2. Choose a mixing parameter `alpha` in order to increases the geographical cohesion of the clusters (using the dissimilarity matrix `D1`) without deteriorating too much the homogeneity on the socio-economic variables.

https://cran.r-project.org/web/packages/ClustGeo/vignettes/intro_ClustGeo.html

ClustGeo in R

D0 "feature": time coefficients
D1 "constraint": geographical dissimilarities

Using ClustGeo:

1. Compute the Features
2. Compute the Spatial Constraints
3. Pick alpha
4. Cluster using hgeoclust()

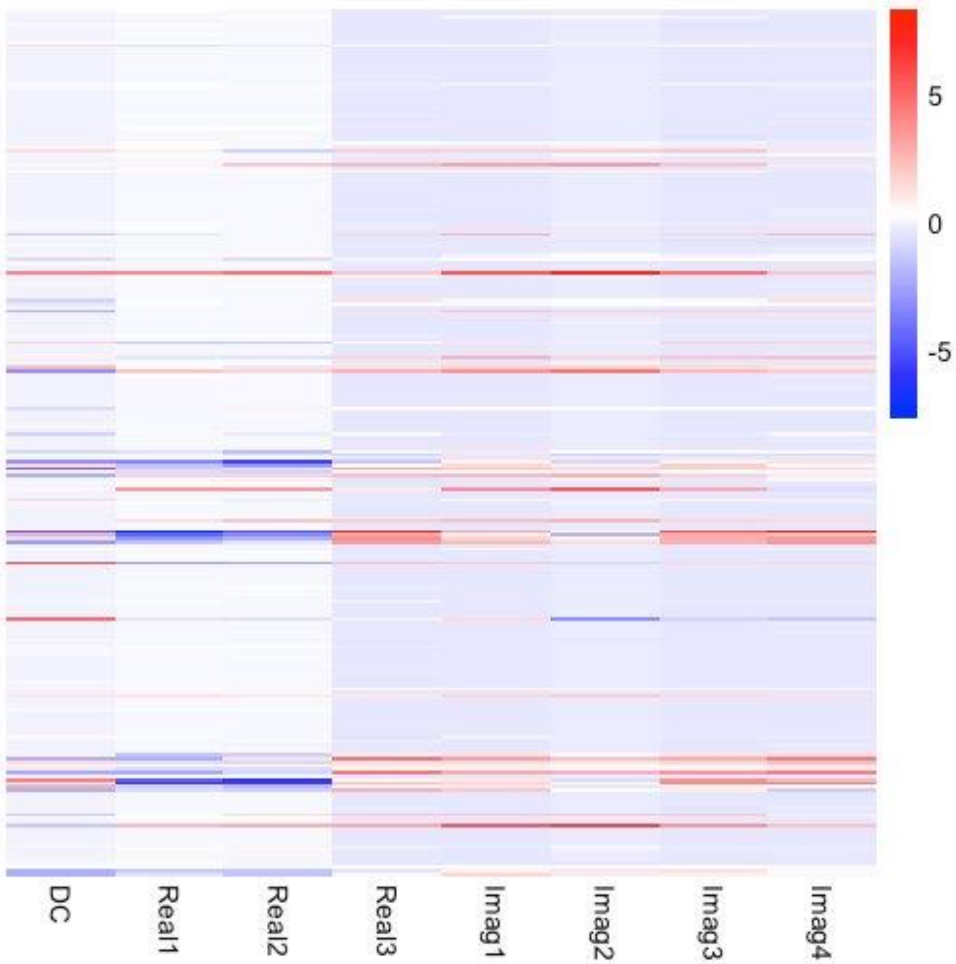
Fourier Transform on Time Series

- Beneficial for capturing cyclical behaviors
- Steps
  1. Prepare the data
  2. Center the demand (helps the analysis focus on the deviation from the baseline)
  3. Use fft() in R to decompose the time series
  4. Extract the first four of real and imaginary components

⇩

Calculate the pairwise distance between the rows of scaled temporal data (used dist() function)

Heatmap of the scaled temporal coefficients

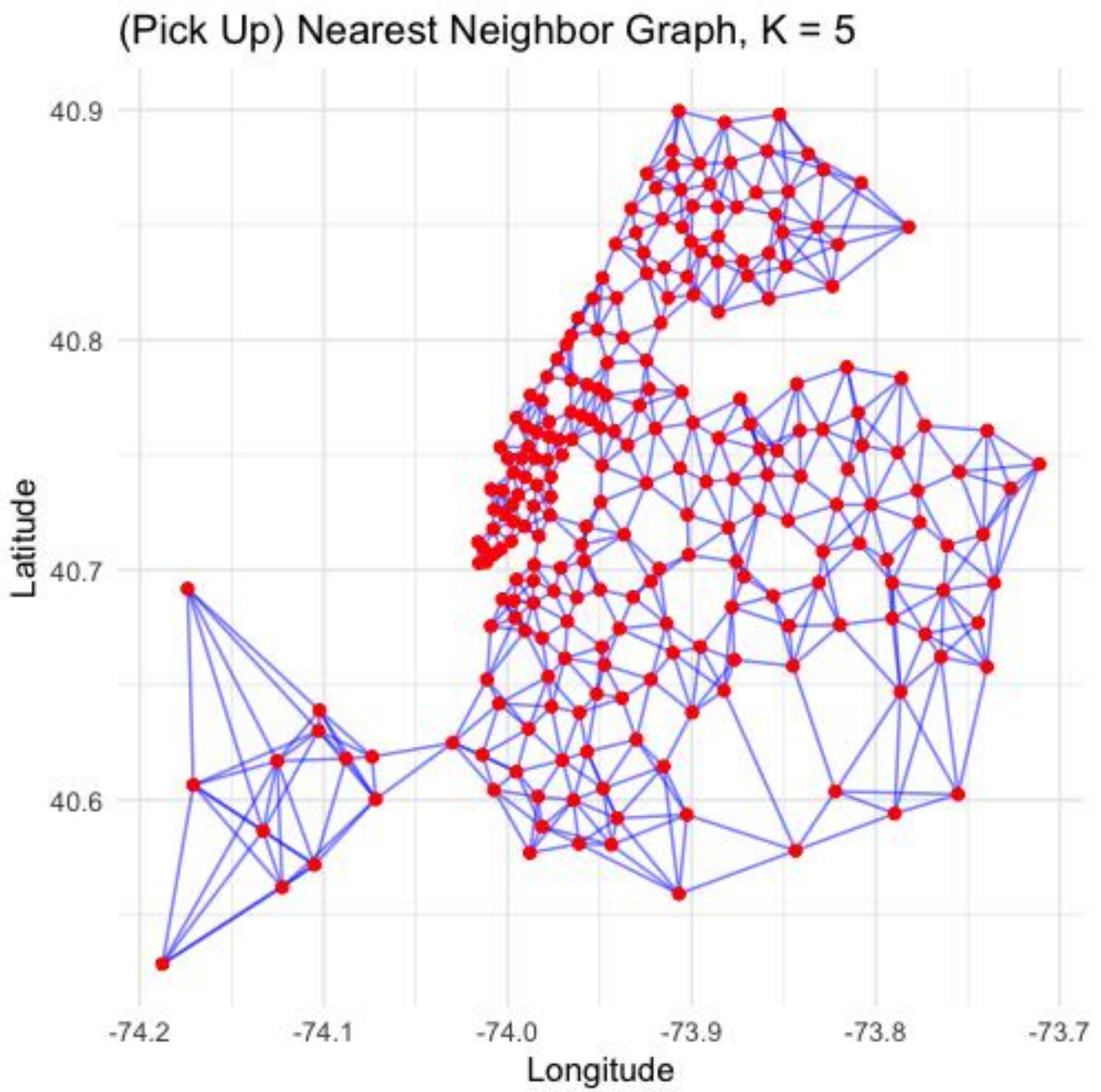


Spatial Constraint

The spatial connectivity w_s is derived from k-nearest neighbor graph to define spatial connectivity.

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

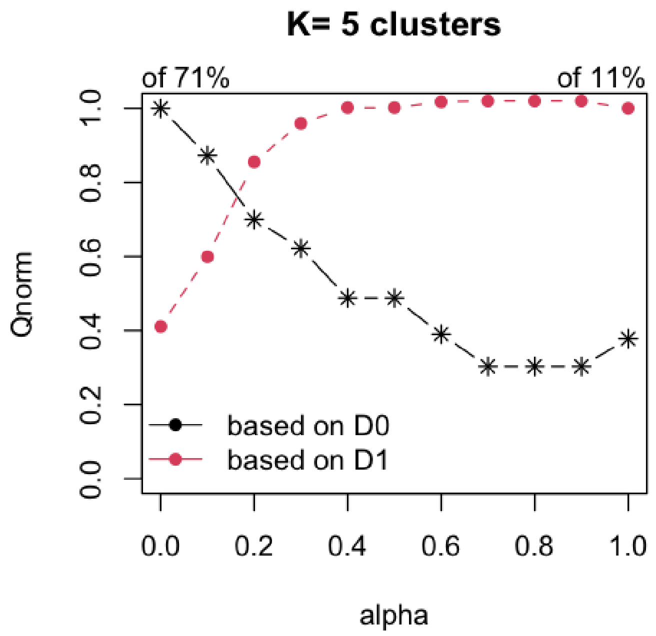
Calculate the spatial dissimilarity matrix as.dist(1-adj_matrix)



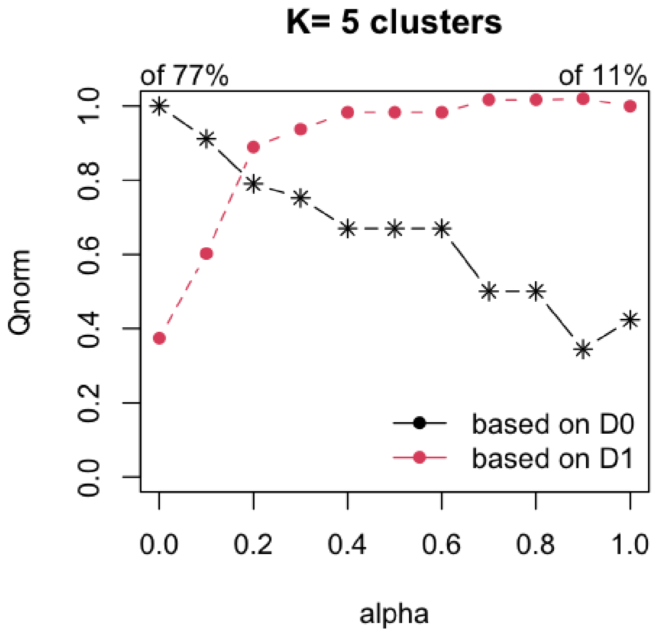
Picking the best alpha

- Q0: Temporal homogeneity
- Q1: Spatial contiguity

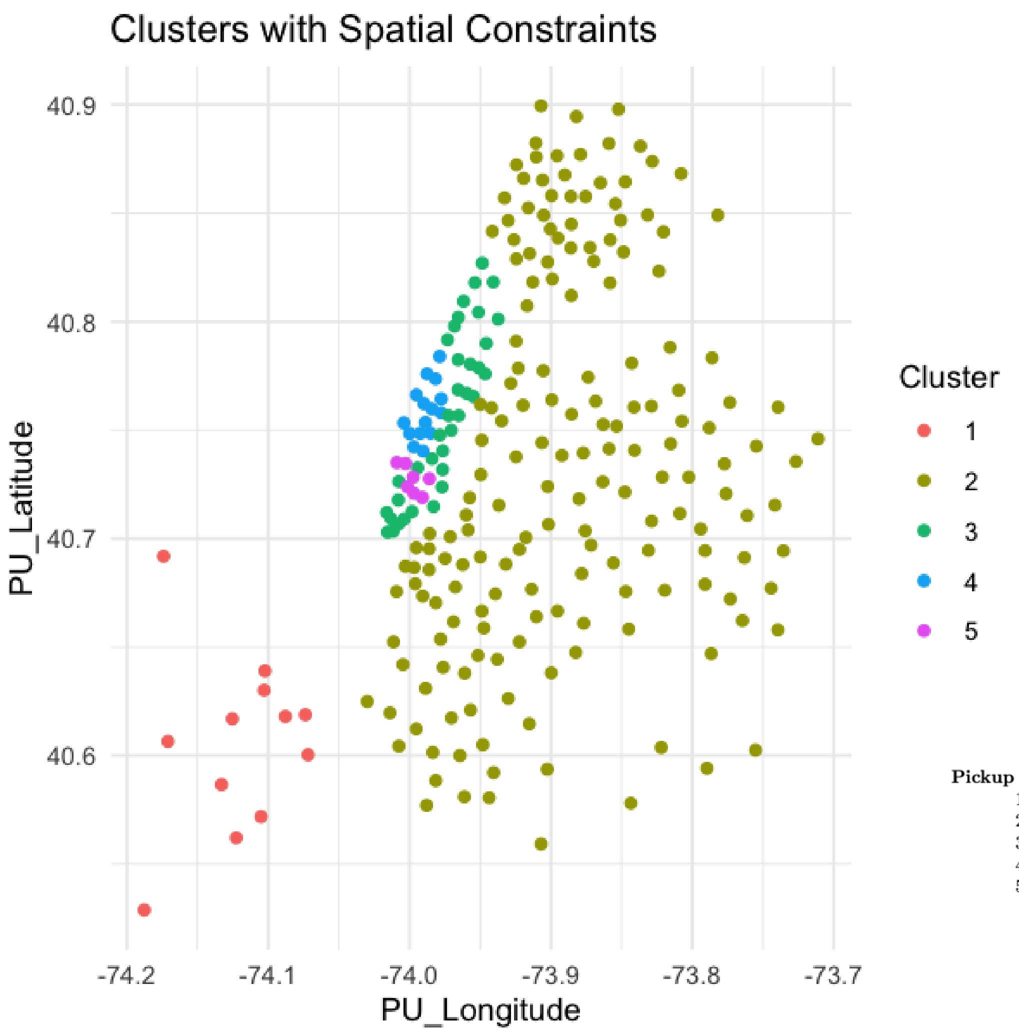
Pick up Demand

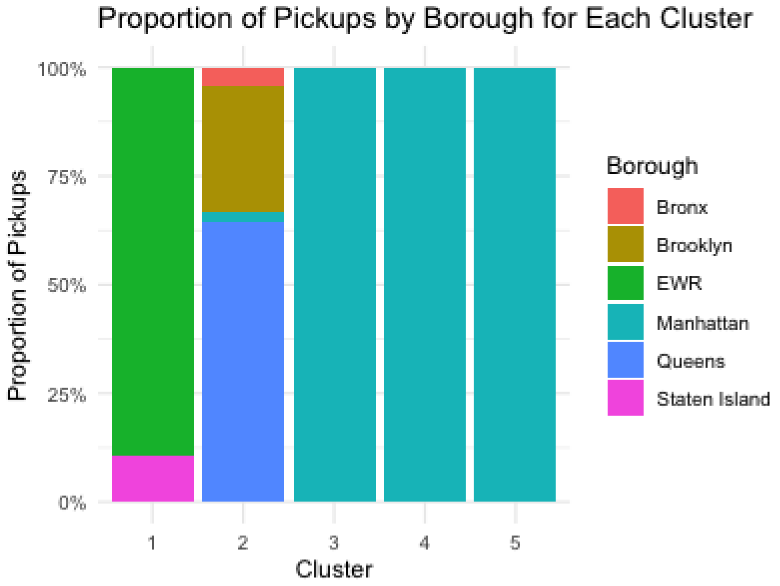


Drop off Demand



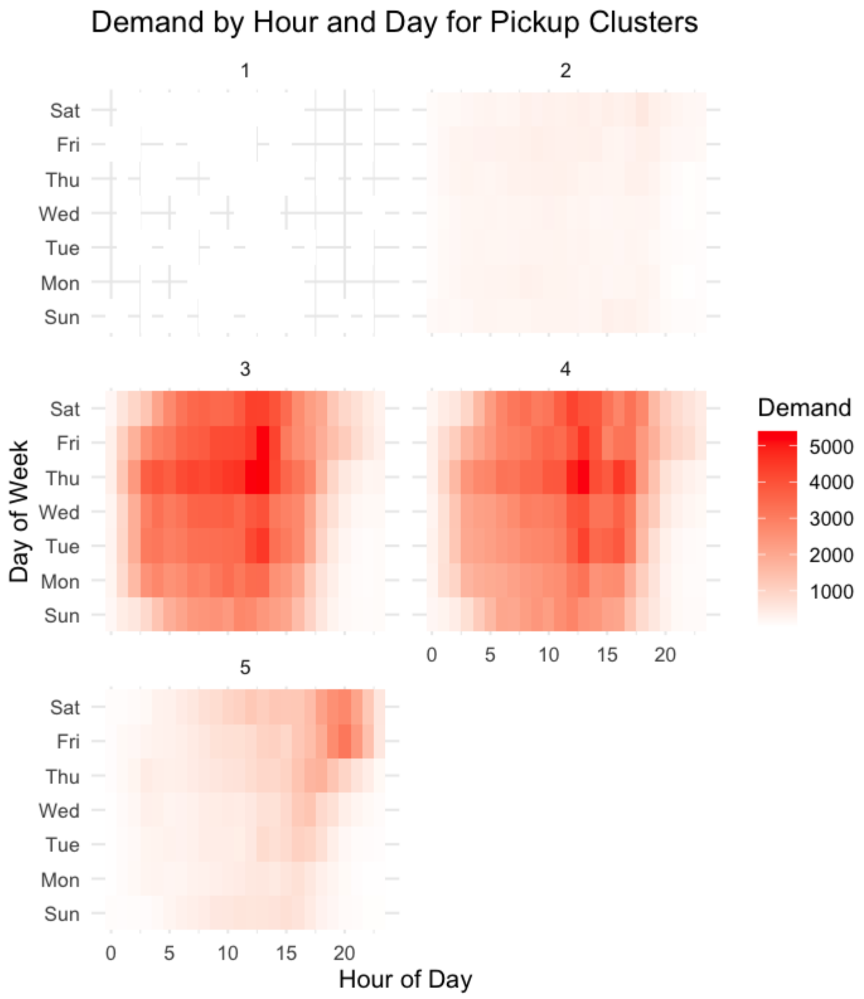
Pick up Demand by Clusters

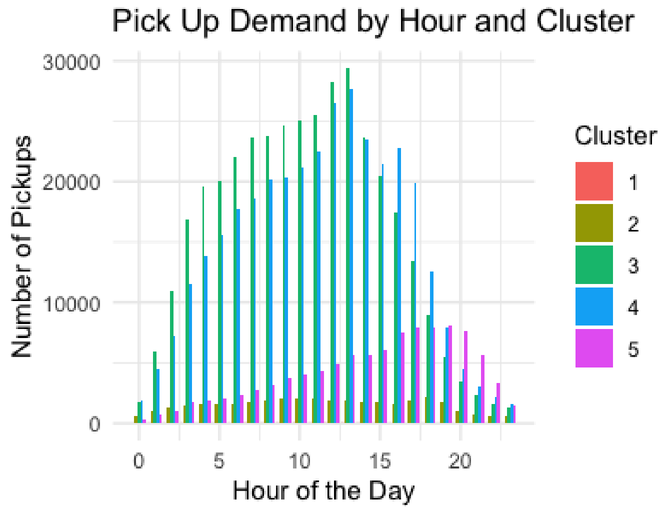


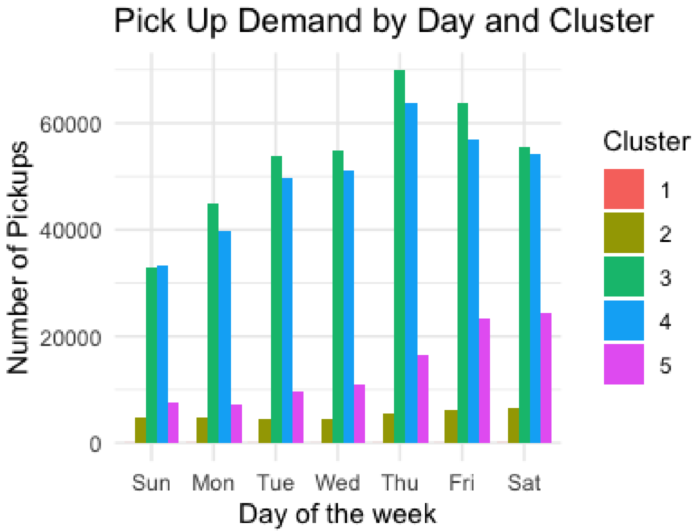


Pickup Cluster	Avg Trip Distance	Median Trip Distance	Avg Passenger Count	Avg Fare Amount	Avg Duration	Total Trips
1	0.271	0.00	1.791	71.233	0.882	201
2	3.237	3.27	1.373	22.088	14.061	36,348
3	1.915	1.51	1.294	13.358	12.007	375,302
4	1.817	1.49	1.368	13.480	12.838	348,398
5	2.127	1.79	1.410	14.157	12.952	99,513

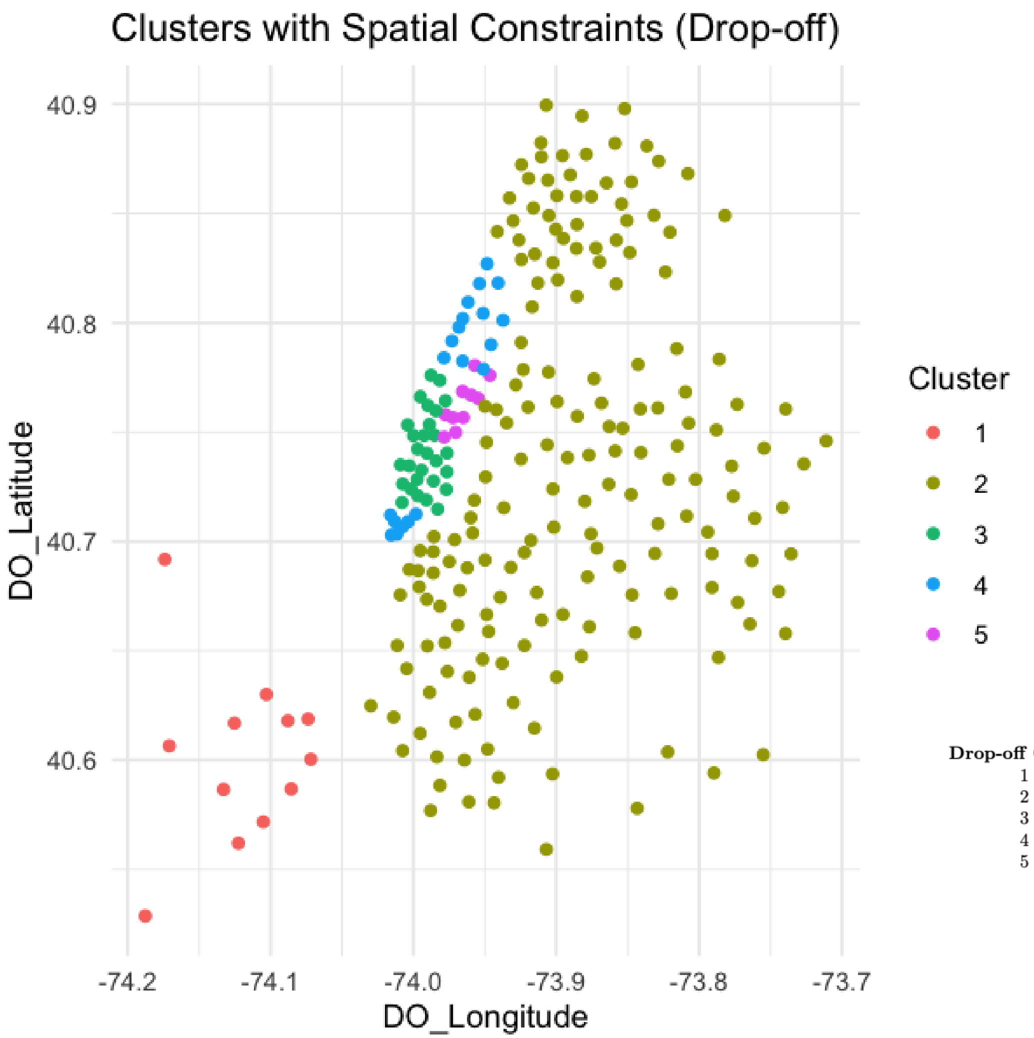
Taxi Pickup Analysis: Time Effect

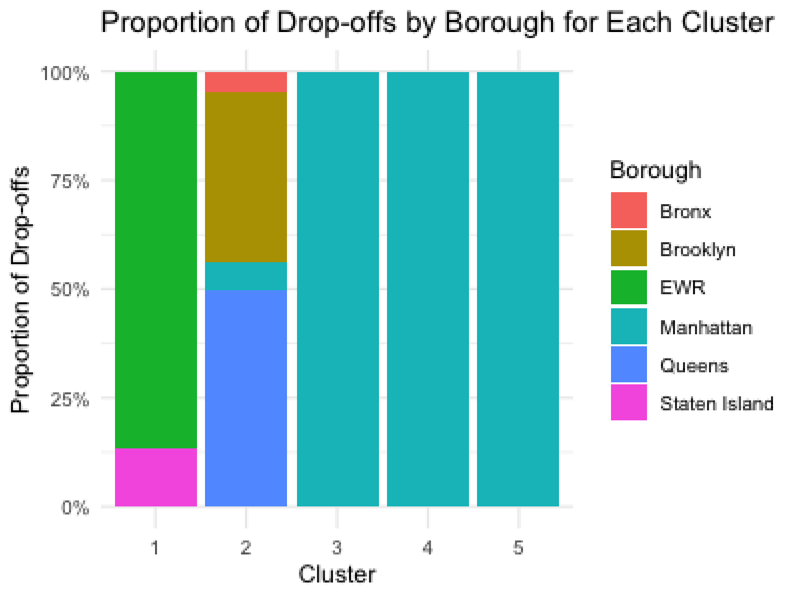






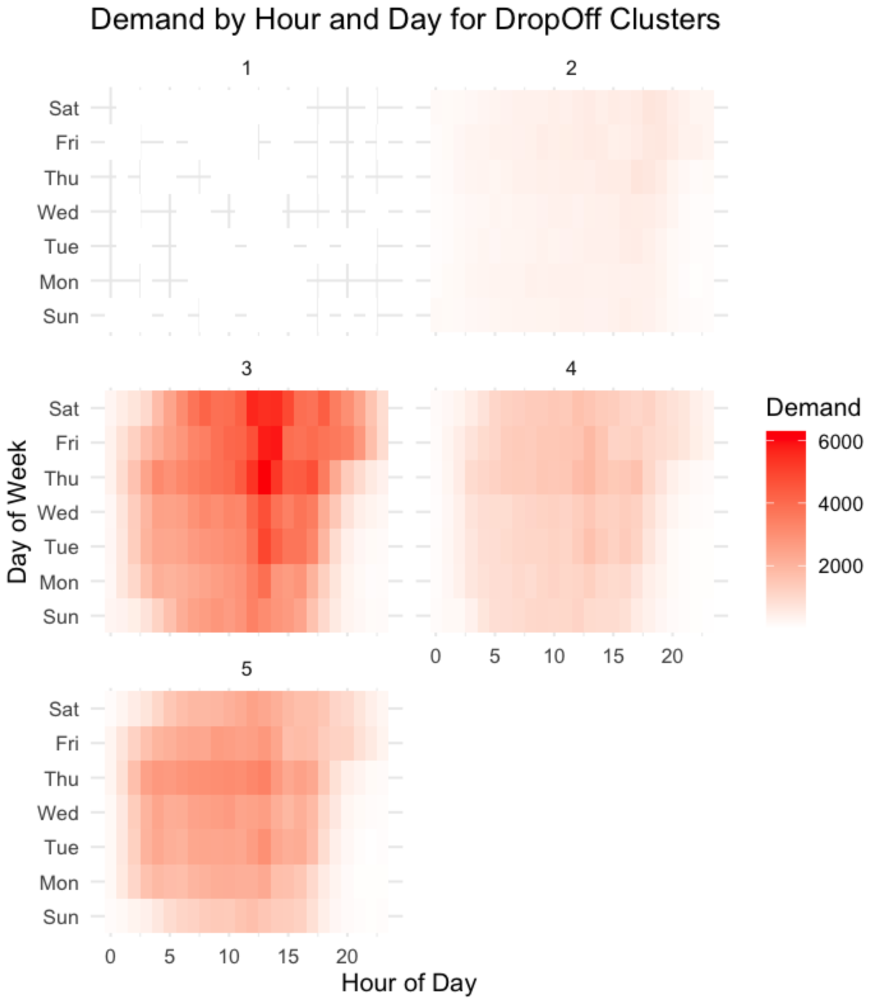
Drop off Demand by Clusters

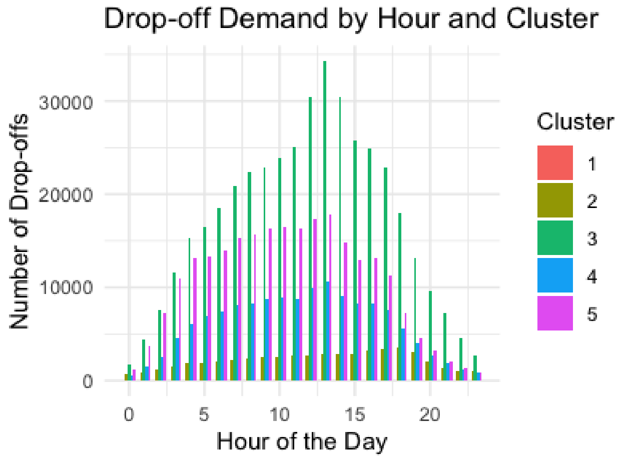


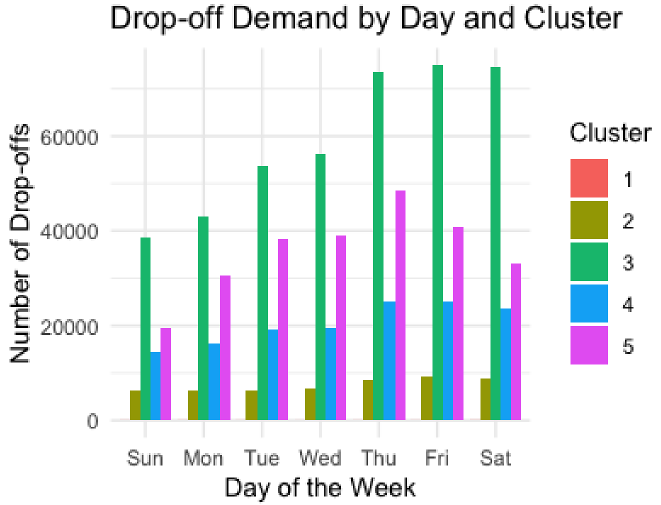


Drop-off Cluster	Avg Trip Distance	Median Trip Distance	Avg Passenger Count	Avg Fare Amount	Avg Duration	Total Trips
1	0.387	0.000	1.761	68.022	1.692	216
2	3.656	3.850	1.341	22.481	16.476	52,203
3	1.761	1.470	1.374	13.325	12.631	414,877
4	2.453	2.000	1.331	15.069	12.932	142,211
5	1.641	1.370	1.292	12.293	11.346	250,255

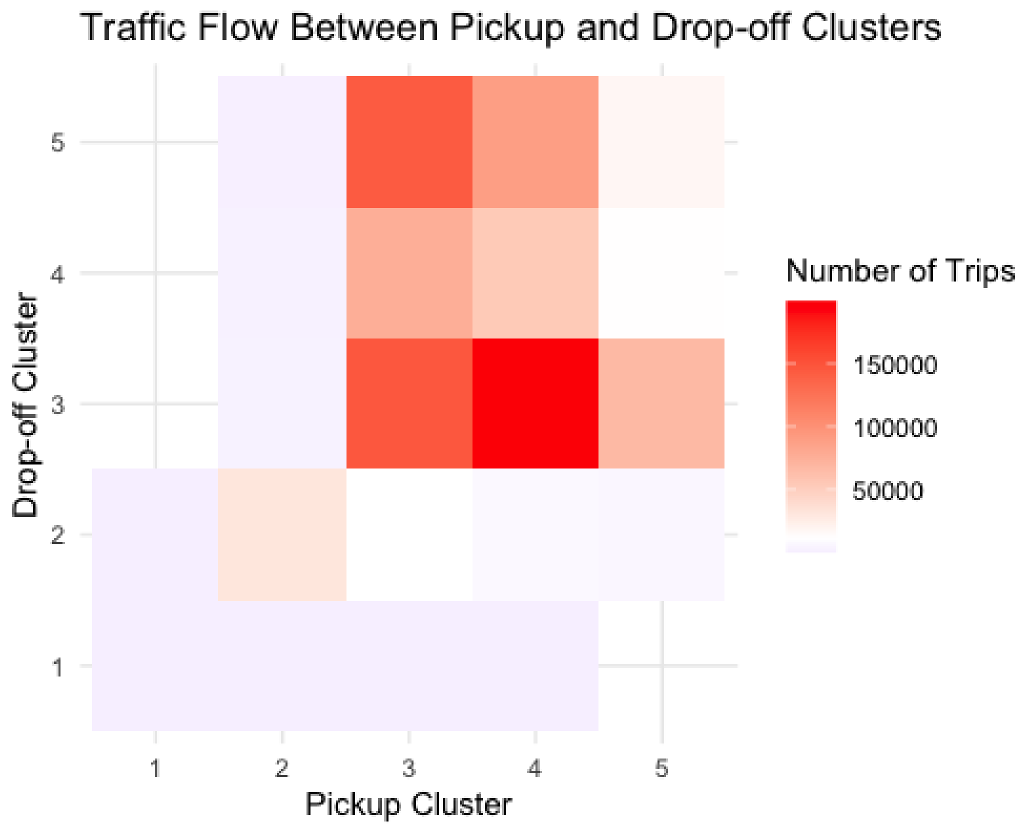
Taxi Dropoff Analysis: Time Effect

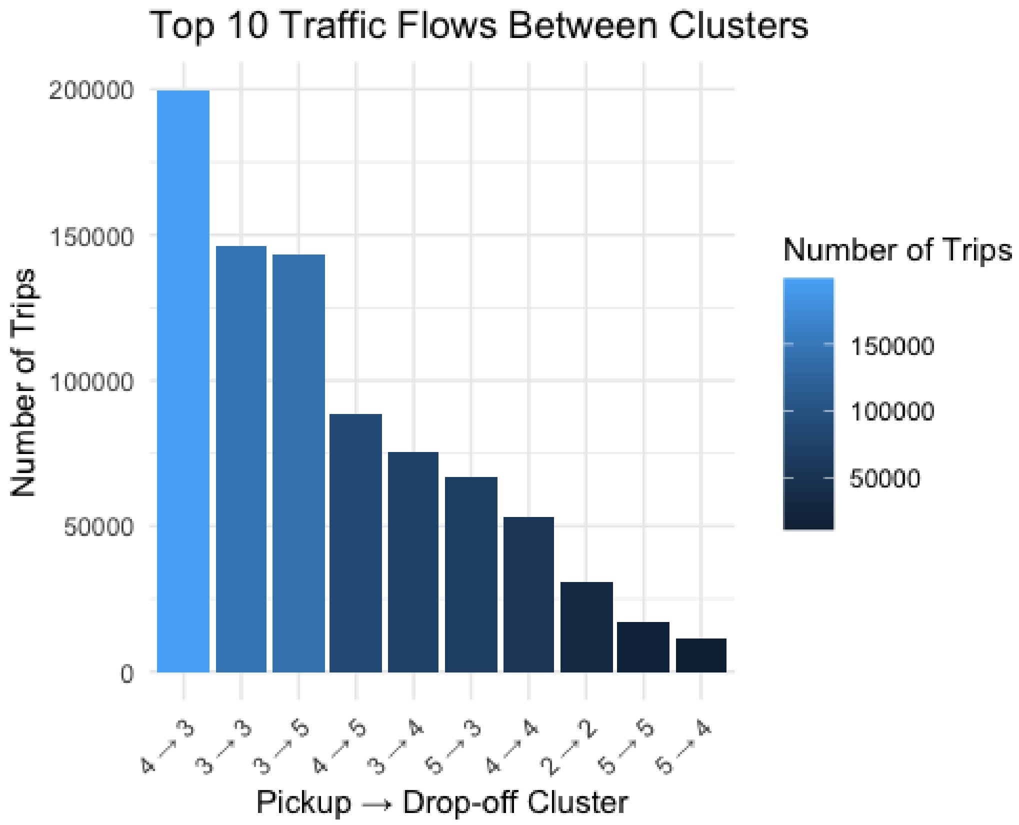




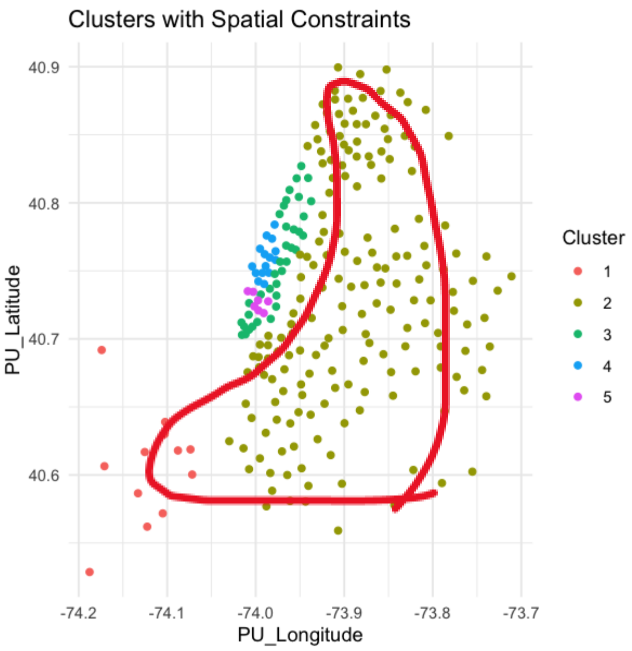


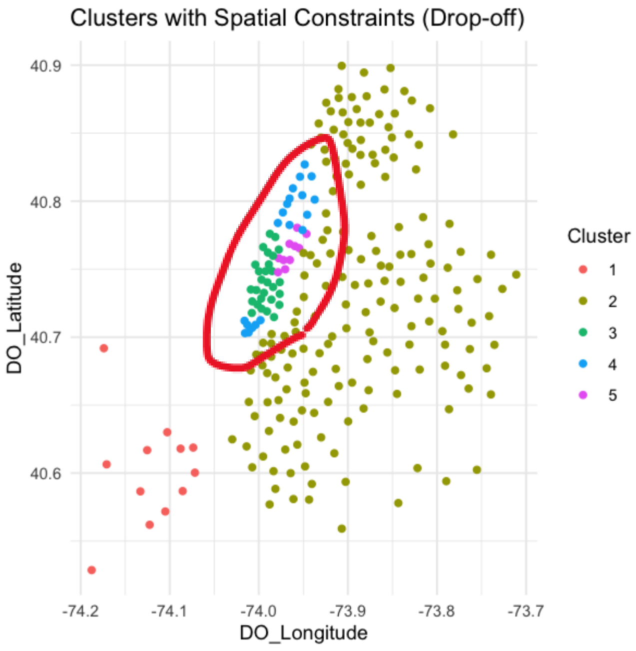
Overall Traffic Flow: Matched Clusters





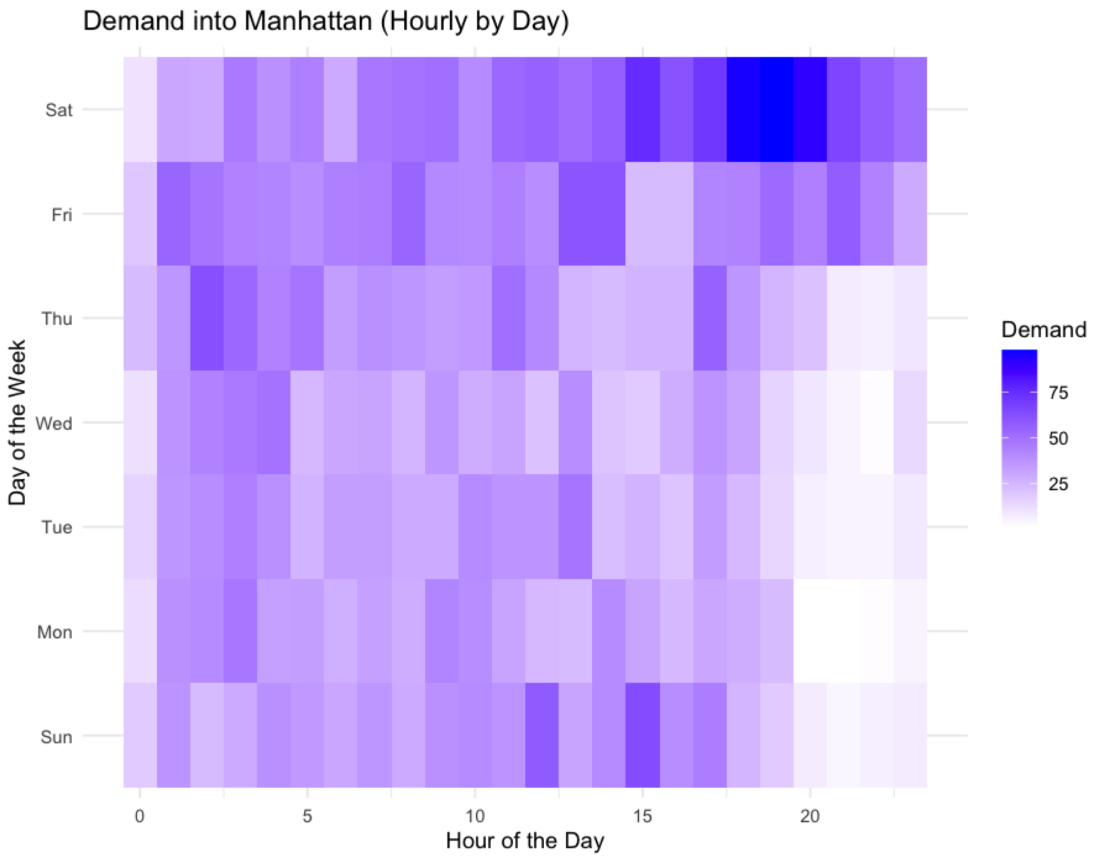
Traffic Flow: Into Manhattan

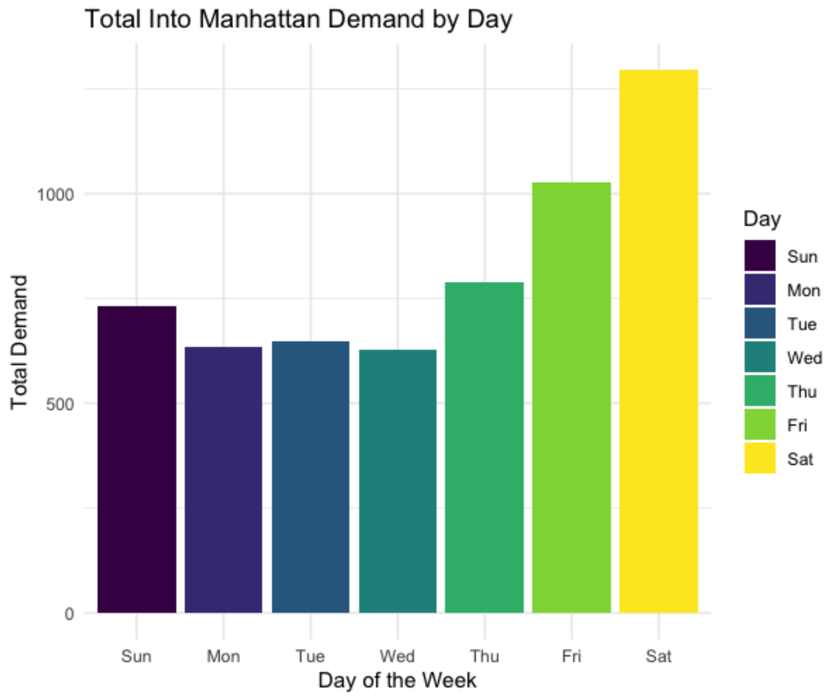




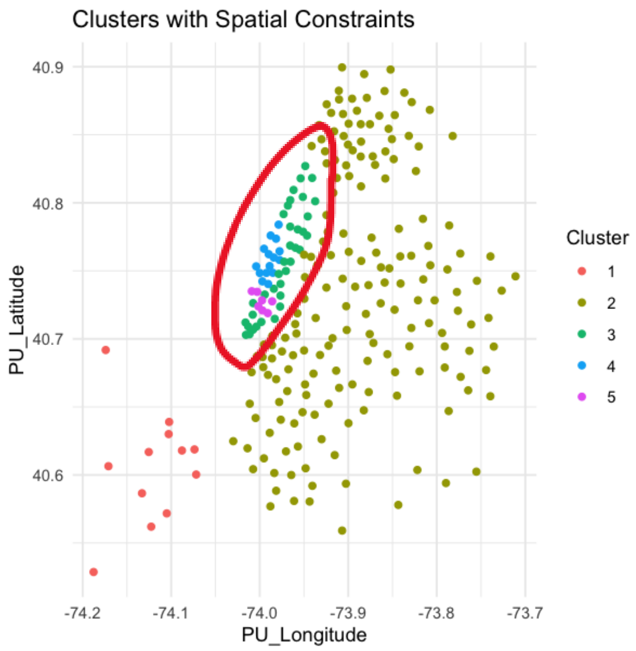
To analyze the traffic flow into Manhattan, we gathered the trips from clusters 1, 2 (pick up) to clusters 3,4,5 (drop off)

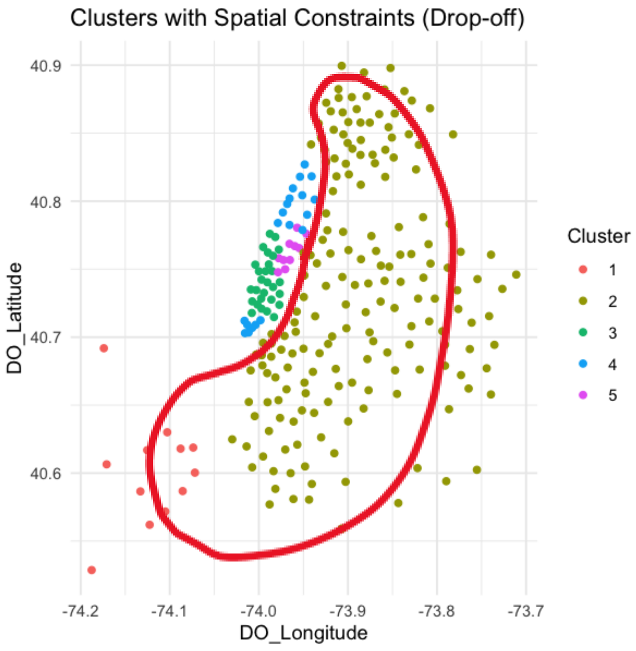
Traffic Flow: Into Manhattan





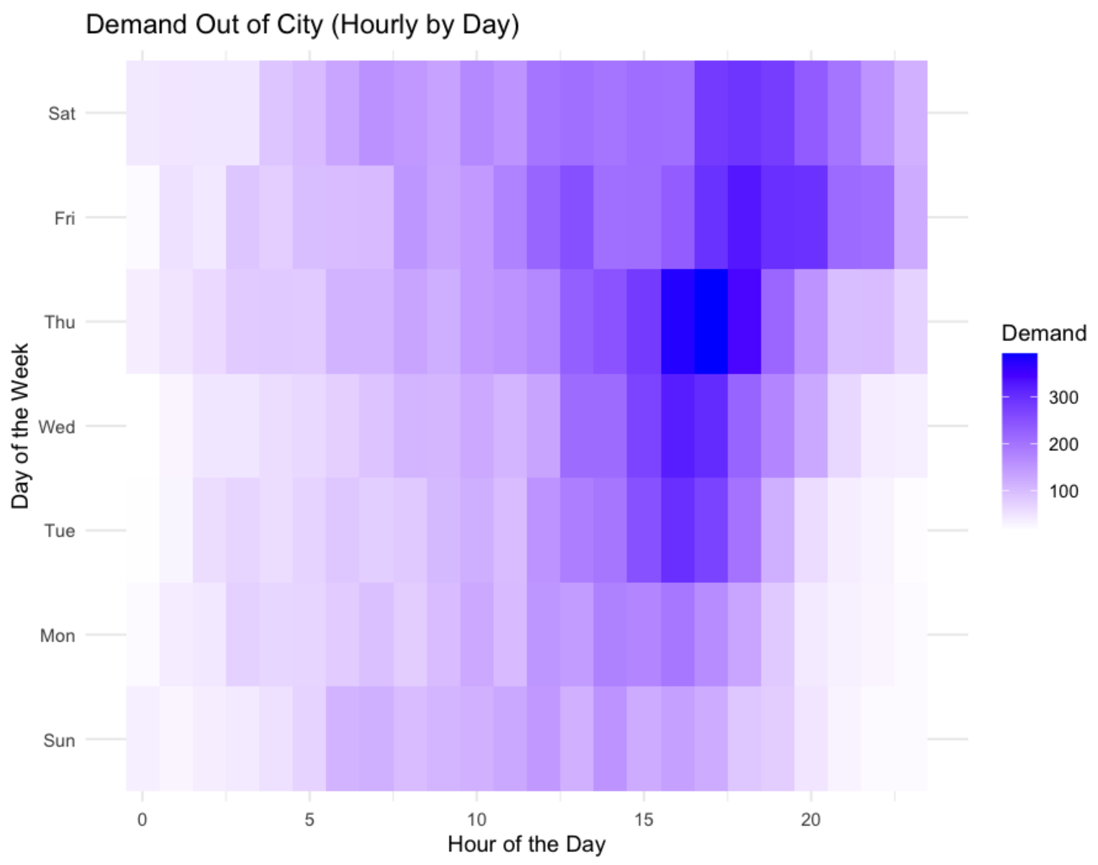
Traffic Flow: Out of Manhattan

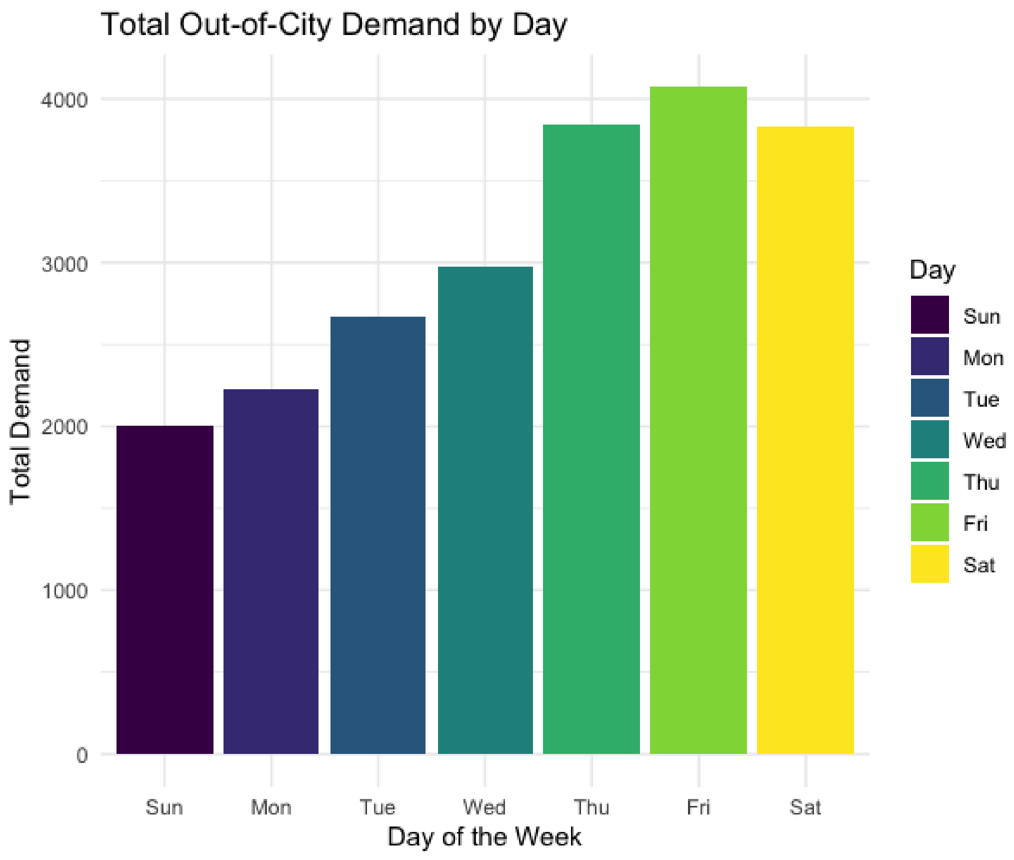




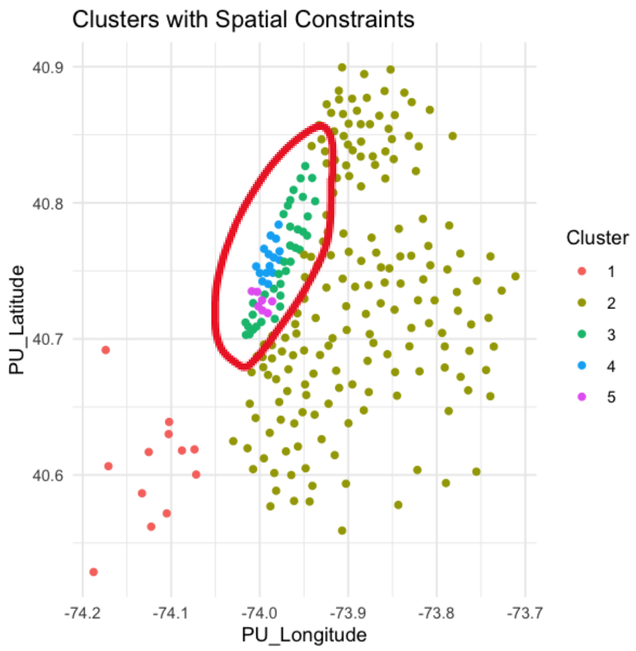
To analyze the traffic flow out of Manhattan, we gathered the trips from <u>clusters 3,4,5</u> (pick up) to <u>clusters 1,2</u> (drop off)

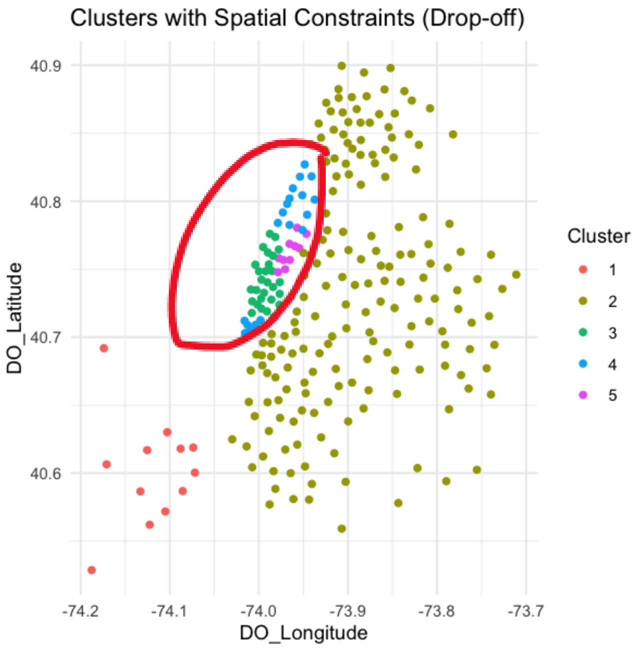
Traffic Flow: Out of Manhattan





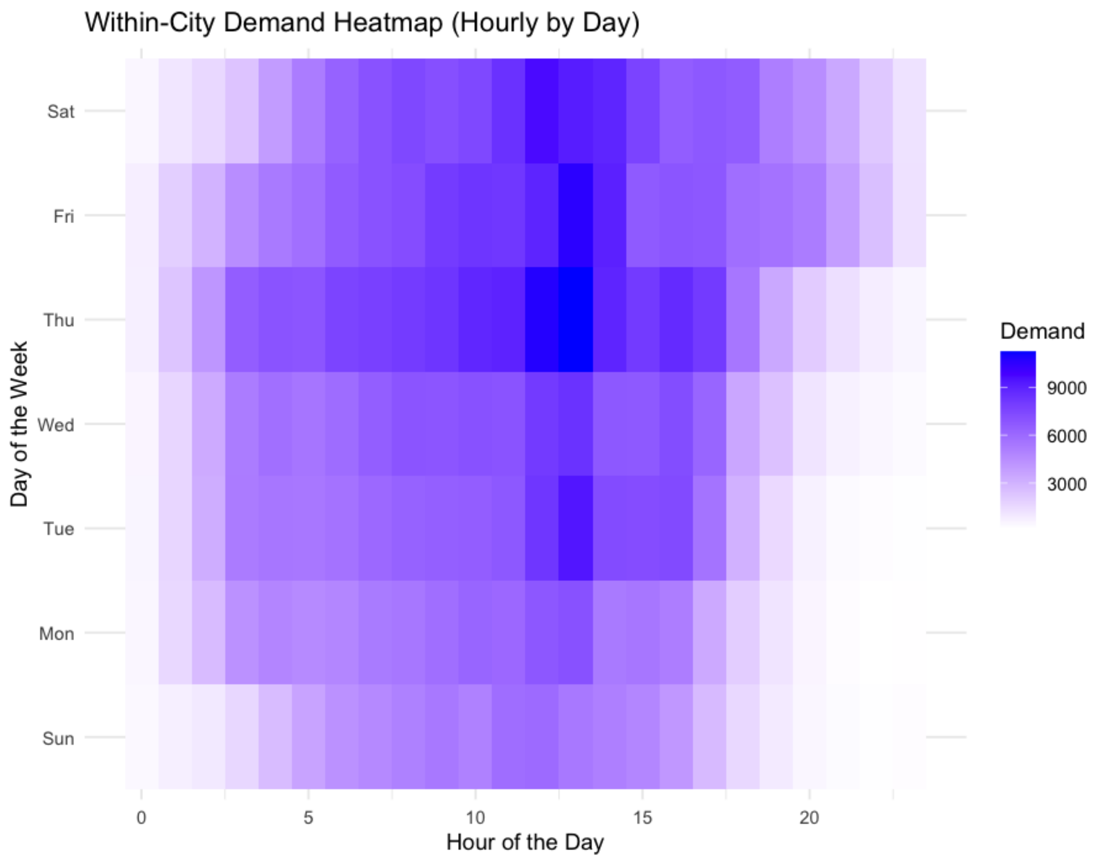
Traffic Flow: within Manhattan

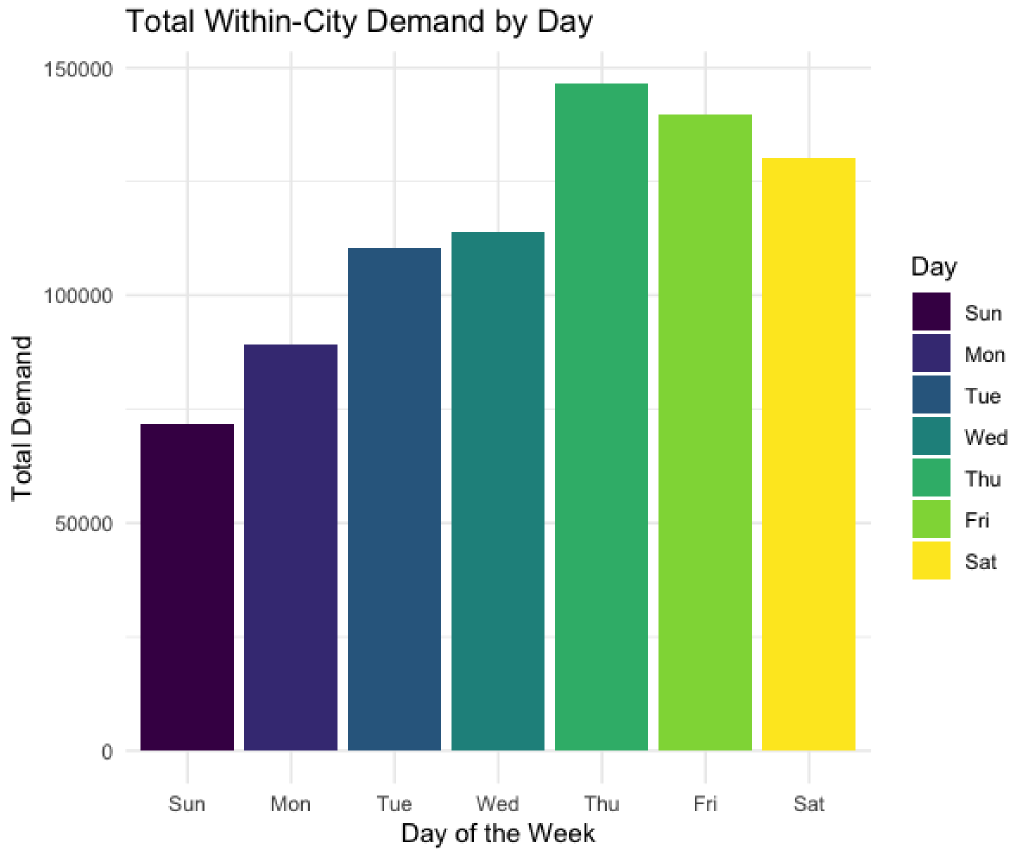




For the traffic flow within Manhattan, we gathered the trips from (pick up) and to (drop off) <u>clusters 3,4,5</u>

Traffic Flow: within Manhattan





Traffic Flow 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

References

- New York City Taxi & Limousine Commission. (2024). *Yellow taxi trip data: August 2024*. Retrieved from https://www.nyc.gov/assets/tlc
- Strang, G. (2018). Fourier Transform. Massachusetts Institute of Technology. Retrieved November 19, 2024, from https://math.mit.edu/~gs/cse/websections/cse41.pdf
- Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2017). ClustGeo: An R package for hierarchical clustering with spatial constraints. *arXiv*. https://doi.org/10.48550/arXiv.1707.03897
- Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2021). *Introduction to ClustGeo*. The Comprehensive R Archive Network. Retrieved from https://cran.r-project.org/web/packages/ClustGeo/vignettes/intro_ClustGeo.html