ISyE 6740 – Spring 2025

Final Report

Team Member Name(s): Kevin Hopper

Project Title: Spectral Clustering of Neighborhoods in NYC using Uber Trip Data

Problem Statement

How fascinating would it be to see how people move about the greatest city in the world: New York? Using the New York OpenData table detailing Uber rides in New York City, we will cluster neighborhoods that are highly connected by Uber trips in order to get a glimpse of how people travel across the city at different times of the year.

Assuming the Uber rides to be a proxy for how people move across the city, we will be able to visualize how the city moves through the day. We will identify groups of neighborhoods that are highly connected, propose business ideas to capitalize on the identified demand, and we may even be able to find clusters indicating unique behaviors of the city, informing a savvy uber driver's daily driving patterns. Our method of breaking down these commuting communities will be to spectrally cluster New York City's 263 neighborhoods, defined by taxi zones, using frequency of trips between neighborhoods as the basis of our weights.

Finally, we will compare the clustering over time, seeing if there are differences in the clustering during tourist season vs the low season and during big events.

Data Source

Uber ride data was found on New York City's OpenData website. Each year, NYC's Taxi and Limousine Commission releases a data set detailing information for every single for-hire-vehicle's (FHV's) trip. This data source contains information on pick-up time, ride duration, fare, driver earnings, and importantly for us, pick-up and drop-off location, broken down into taxi zones (263 of them, + 2 to indicate trips out of NYC) that we can construct into a graph with weights based on the number of trips.

This data set can be filtered to Uber, Lyft, etc., but in this analysis, we will only examine Uber, based off of preceding research that Uber is more likely to serve the outer boroughs – this will allow for the graph that we construct to be less Manhattan dominated, and hopefully more interesting to analyze.

Methodology

Exploratory Data Analysis

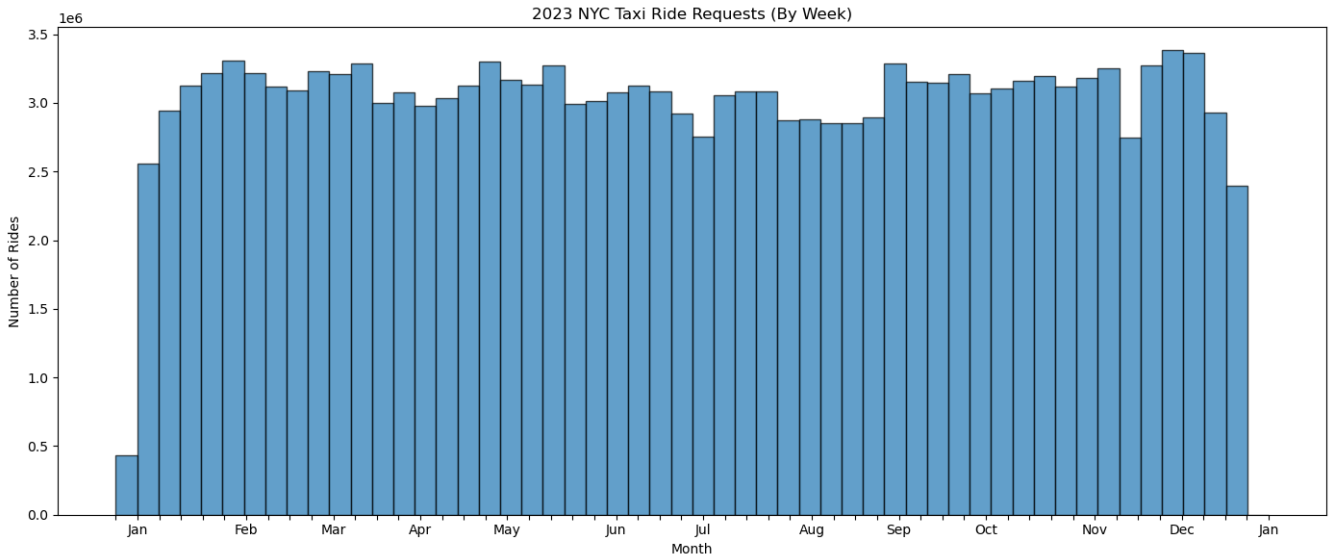
We begin by importing a 26-gigabyte Uber ride dataset into Python. Due to the large file size and limited system memory, we needed to be strategic in how we loaded the data. We minimized memory usage by pulling only the necessary columns and explicitly setting data types during import. This precaution allowed us to load the entire year's worth of data into memory – although even with these optimizations, simply reading the file took about an hour. Ultimately, we extracted over 160 million ride records, focusing on three fields: request_datetime, PULocationID (pick-up location), and DOLocationID (drop-off location).

```
RangeIndex: 160346947 entries, 0 to 160346946
Data columns (total 3 columns):
#   Column            Dtype
---  ------            -----
0   request_datetime  datetime64[ns]
1   PULocationID      uint16
2   DOLocationID      uint16
dtypes: datetime64[ns](1), uint16(2)
memory usage: 1.8 GB
```

Summary of the Uber Data import (1.8 GB of memory!)

With the data loaded, we performed preliminary exploratory data analysis (EDA) to understand ride patterns across the year and differences between different types of days. These insights helped guide our clustering strategy, particularly around identifying periods of concentrated ride activity such as the morning and evening commutes.

When examining ride volume by week over the course of 2023, we find that volume remains relatively steady, averaging around 3 million rides per week. Minor variations include a dip in August and a slight increase in ride activity during the winter months. Weeks with less than seven days (e.g., the first and last week of the year) understandably show lower totals, falling below 2.5 million rides.

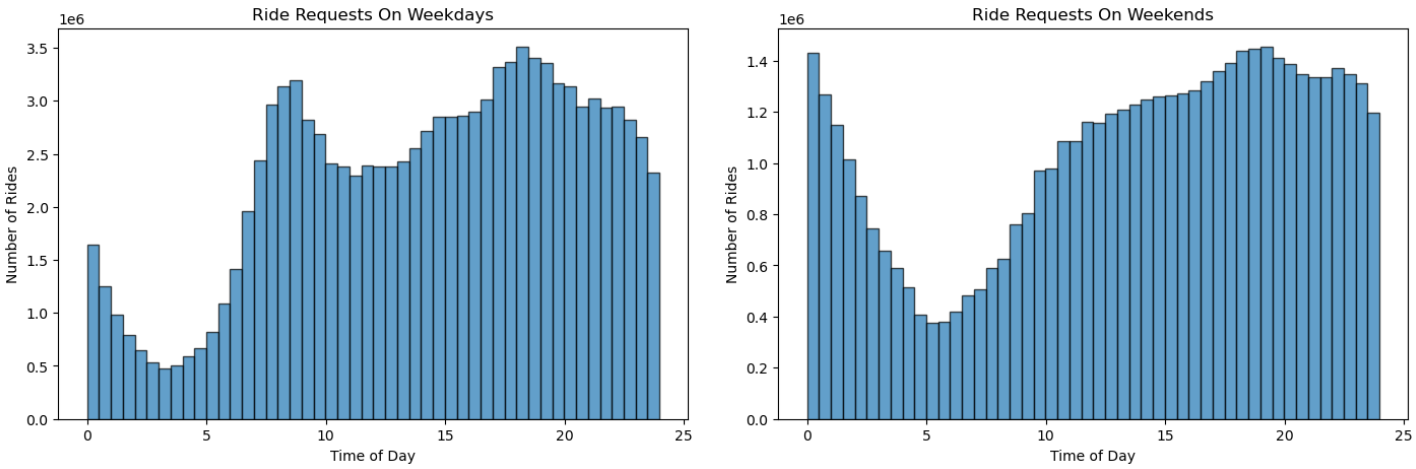


Uber rides over the course of 2023, broken out by week

Breaking rides down by time of day reveals two distinct profiles for weekdays and weekends. On weekdays, consistent with New York's business culture, we observe two strong peaks corresponding to the morning

commute (7–10 AM) and the evening commute (5–7 PM). Ride volume drops sharply during the early morning hours between 2–5 AM.

By contrast, weekend ride patterns are more evenly distributed throughout the day, with a notable increase in late-night rides—reflecting the city's vibrant nightlife. These differences in temporal patterns suggest that New York's mobility network may be structured differently depending on the time of day or day of the week.



Ride Requests on Weekdays vs Weekends, broken out by 30 minute intervals

Identifying Windows of Interest

For our analysis, we aim to compare distinct periods of time to determine whether the clustering of neighborhoods changes from the baseline due to external factors such as time of day, day of week, and major events.

To draw meaningful comparisons, we have selected specific windows of interest that are well-suited to highlight potential differences in neighborhood connectivity. Below, we outline the windows we are examining:

1. Baseline (Full Year 2023):
    a. We begin by establishing a baseline by analyzing ride patterns across all of 2023. This provides a view of how New York City neighborhoods connect on average, without focusing on any particular external influence.
2. The U.S. Open (Event Impact Analysis):
    a. The U.S. Open Tennis Tournament draws nearly 800,000 attendees every summer. We will compare connectivity during the two weeks of the tournament to the two weeks immediately after, in order to determine whether an event of this scale significantly alters neighborhood clustering.
3. Tourism Season (December vs January):
    a. December is typically the peak tourism month in New York, while January sees the lowest tourist traffic. By comparing these two months, we aim to assess whether fluctuations in tourism volume impact the connectivity of the city.

This analysis will allow us to evaluate whether extreme peaks and valleys in demand correspond to changes in how the city's neighborhoods cluster together.

(To see more windows of interest's clusters, please feel free to check out my github, linked here!)

Clustering Methods

We chose spectral clustering to identify groups of neighborhoods based on the structure of the Uber ride dataset, which we modeled as counts of rides between different locations. Spectral clustering is particularly well-suited for uncovering non-linearly separable clusters in complex network structures like this one. By leveraging the connectivity patterns encoded in a graph representation of the data, spectral clustering enables the detection of neighborhood communities that are densely connected internally but more loosely connected externally to other groups.

For each window of interest in our analysis, we followed a consistent methodology:

1. Construct the adjacency matrix

We constructed an adjacency matrix where a link between two neighborhoods was weighted by the number of Uber rides connecting them. Importantly, we counted intra-neighborhood trips – trips where the pick-up and drop-off occurred within the same neighborhood. This decision allowed us to better capture local movement patterns, strengthening the internal weight of clusters. Including these self-loops helped ensure that neighborhoods with strong internal activity – such as many areas in Staten Island, which have dense local traffic but limited connectivity to other boroughs – would still form meaningful clusters. Overall, including intra-neighborhood trips made the clustering process more sensitive to local community structures, allowing spectral clustering to surface groups that are locally cohesive even if they are globally less connected.

2. Apply Spectral Clustering

We applied spectral clustering using scikit-learn's spectral clustering package. To determine the best number of clusters, we iterated through 3 to 20 clusters for each analysis window, selecting the solution that produced the most distinct and intuitive groupings.

3. Visualization

After assigning clusters, we visualized the results in two complementary ways:

- Map View: A geographical map of New York City, where neighborhoods were color-coded based on their cluster assignment. This provided an intuitive, spatial understanding of how clusters formed across the city.
- Network View: A node-and-edge network graph, where each neighborhood is represented by a node, and edges represents ride connections. The same cluster color coding was used, allowing easy comparison between the geographic and network structures.

The map view helped highlight spatial continuity within clusters, while the network view allowed us to observe connection strength and separation between clusters.

Evaluation and Results

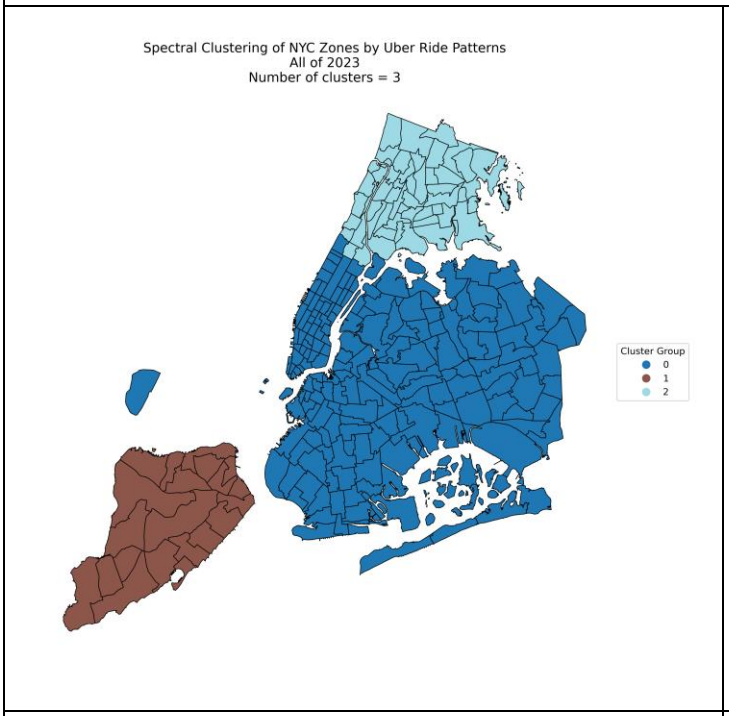
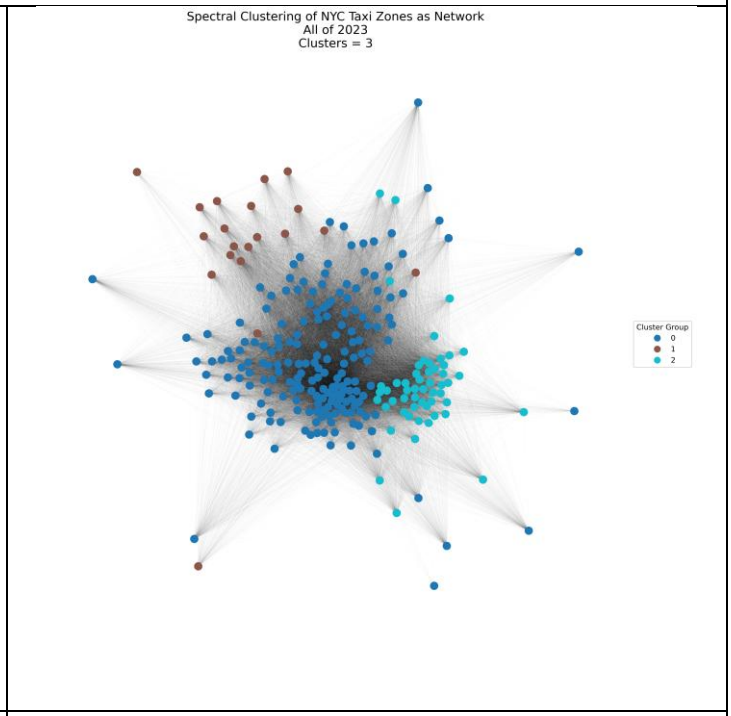
With our methodology established and our visualizations explained, we now turn to analyzing the different windows of interest to see how neighborhood clusters change over time.

Baseline: Full year of 2023

We begin by examining the full year of 2023 as a baseline. This case will help set expectations for what typical clustering patterns look like, and it will also serve as a reference point for interpreting our subsequent results. As a note, for each scenario we analyzed, we generated clusterings ranging from three to twenty clusters. In all future analyses, we will present the number of clusters that produced the most representative and interpretable results. However, for this baseline case, we will examine three versions:

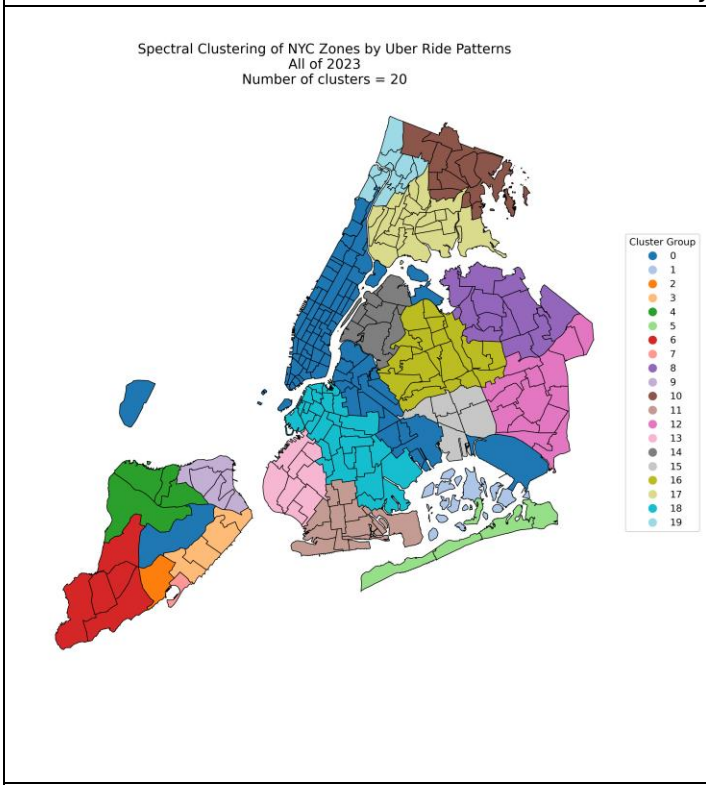
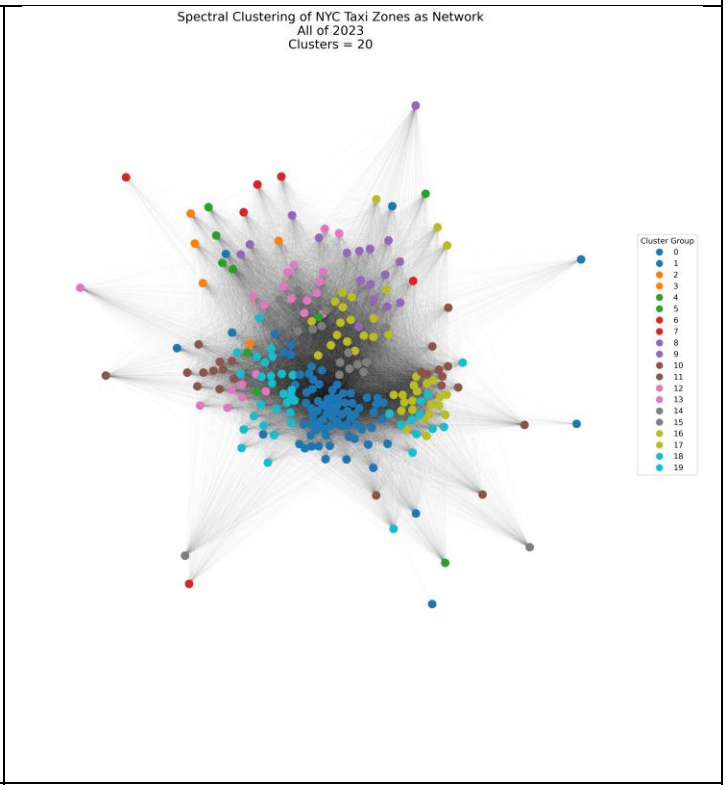
1. A coarse view with three clusters,
2. A fine-grained view with twenty clusters,
3. A balanced view with seven clusters.

This approach will illustrate how the clustering structure changes depending on the level of granularity chosen.

<table>
<tr><th colspan="2">Three Clusters</th></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Neighborhood Map</td><td>Network Graph</td></tr>
</table>

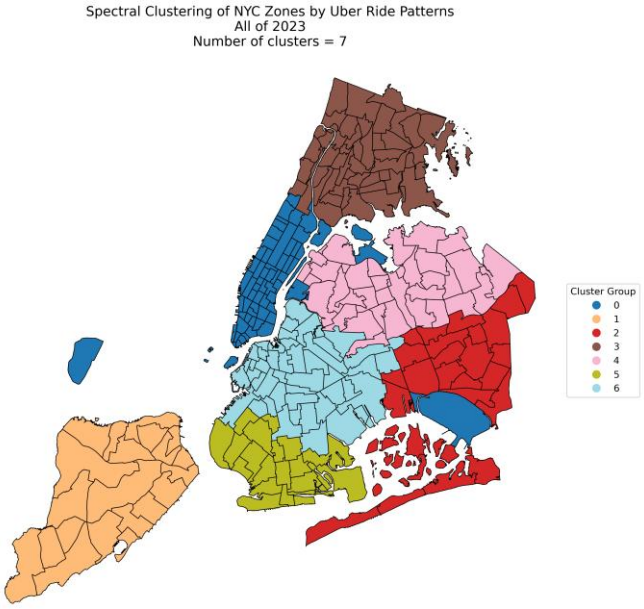
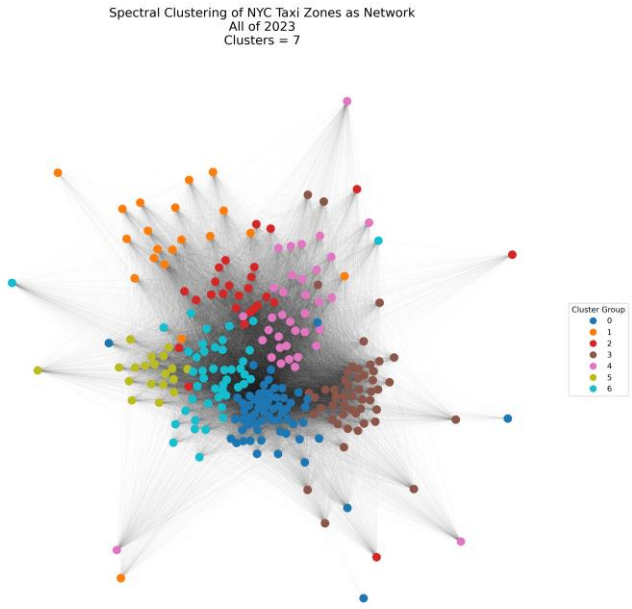
How to interpret this graphic: Each node in the network graph corresponds to a neighborhood shown on the map to the left. The colors are consistent across both visuals – for example, a brown neighborhood on the map (such as Staten Island) matches the brown nodes in the network graph. In the network graph, the distance between nodes (or groups of nodes) reflects the strength of their connection: nodes that are closer together are more strongly connected through ride activity, while nodes that are farther apart are less related, indicating fewer direct rides between them.

This is clearly seen in the graphic above, where the brown cluster (Staten Island) is noticeably separated from the turquoise cluster (the Bronx). This separation aligns with our expectations — we would not anticipate a high volume of Uber rides directly connecting these two boroughs. Let's consider the opposite level of detail: 20 clusters.

Twenty Clusters	
	
Neighborhood Map	Network Graph

In this extreme case of 20 clusters, the map highlights micro-communities very clearly, but at the cost of becoming too granular. For example, Staten Island is split into several small clusters. While the network graph shows that these clusters are all positioned close together — indicating their relatedness — the algorithm still separates them in order to meet the requirement of forming 20 distinct groups. In some cases, this even leads to clusters consisting of a single neighborhood.

On the positive side, this finer breakdown allows us to better see distinctions within Long Island neighborhoods, which might otherwise be hidden in larger groupings. For the majority of our analysis, however, we will select a clustering result somewhere between these two extremes: a number of clusters that provides enough granularity to capture meaningful differences over time, without creating so much complexity that it obscures the larger patterns we want to highlight.

Seven Clusters	
	
Neighborhood Map	Network Graph

Above, the clusters remain clearly defined in the network graph, but now we can also see with greater granularity how different regions are interconnected. Let's walk through the results to set a baseline understanding of how New York neighborhoods connected over the course of 2023, through the lens of Uber rides.

Key Observations:

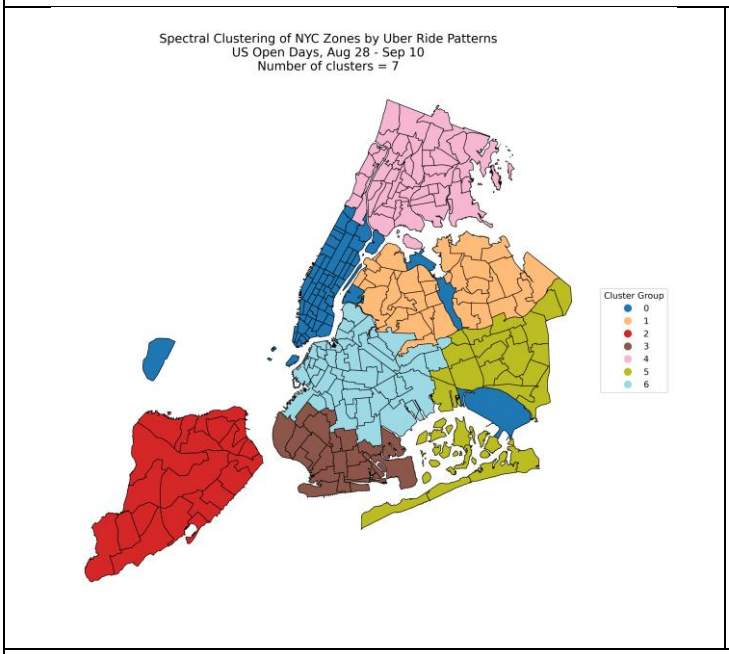
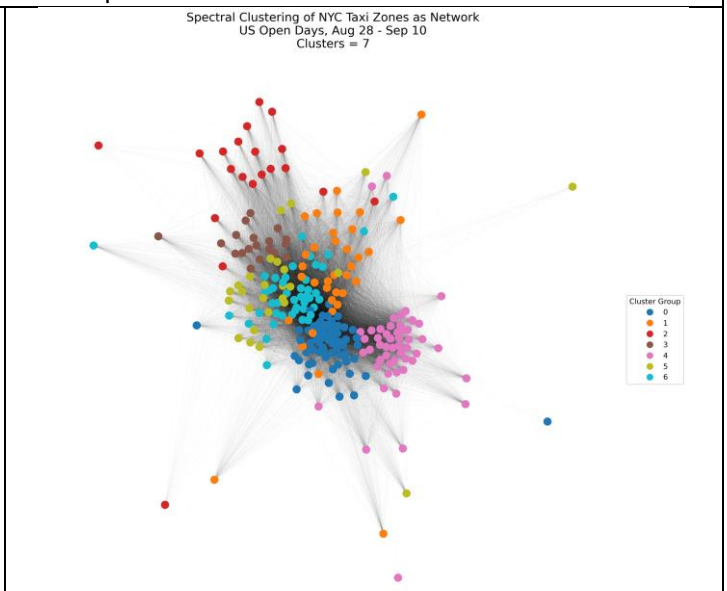
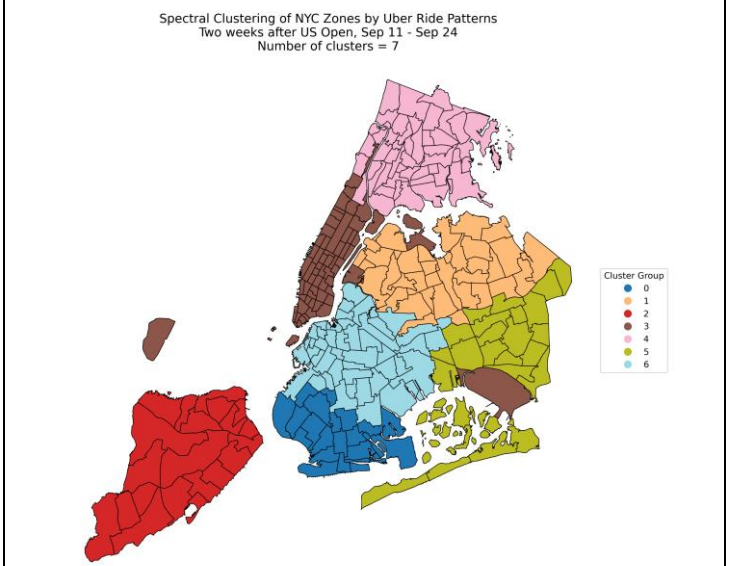
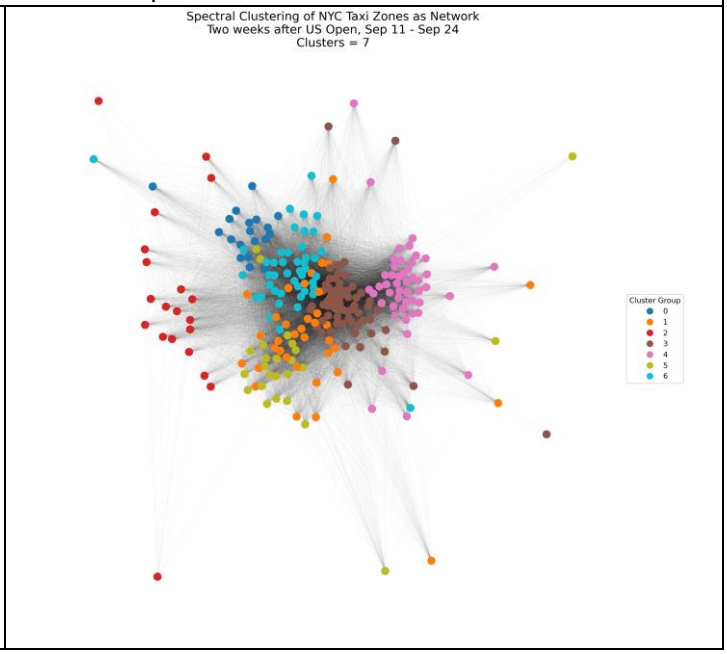
1. Staten Island (Orange) is isolated from the rest of New York.
    a. One consistent theme we will see across our analysis is that Staten Island remains separated from the other boroughs. While it's certainly possible to take an Uber from Staten Island to Manhattan, the ride is long, expensive, and relatively rare, leading to a clear separation in the clustering.
2. The Bronx (Brown) forms a tight, distinct cluster.
    a. Although the Bronx is more connected to the rest of New York than Staten Island, it still forms a tight-knit cluster that is somewhat isolated. While geographically close to Manhattan, the Bronx shows a noticeable gap between itself and nearby Long Island neighborhoods (plotted in pink).
3. Long Island neighborhoods (Yellow, Light Blue, Pink, Red) are fragmented but still linked.
    a. The neighborhoods across Long Island are closely positioned on the network graph but split into several distinct clusters. For example, lower Brooklyn (yellow) is separated from the Astoria/Flushing area (pink), both physically on the map and in the network graph – with North Brooklyn (light blue) sitting between them. This separation makes intuitive sense: rides are more common between closer neighborhoods and less common across longer distances.

4. Airports are strongly connected to Manhattan
    a. Interestingly, three dark blue nodes – JFK (center of the red cluster), LaGuardia (top of the pink cluster), and Newark (isolated to the west in New Jersey) – are grouped with Manhattan despite their physical distance. This is an example of how connectivity, not physical proximity, drives clustering: airports are heavily tied to Manhattan's travel patterns. Throughout our analysis, we will see that the airports consistently cluster with Manhattan.

The U.S. Open (Event Impact Analysis)

The U.S. Open is a major international tennis tournament that drew nearly 800,000 visitors in 2023, making it one of the largest annual events in New York. Given the scale of attendance, we expect to see a shift in the cluster assignment of the specific neighborhood where the tournament takes place, as increased traffic to and from the area could alter its connectivity patterns within the network.

Let’s examine how a major event can influence New Yorkers' transportation patterns. We focus on the two weeks during the U.S. Open and the two weeks immediately after to see if this event causes a change.

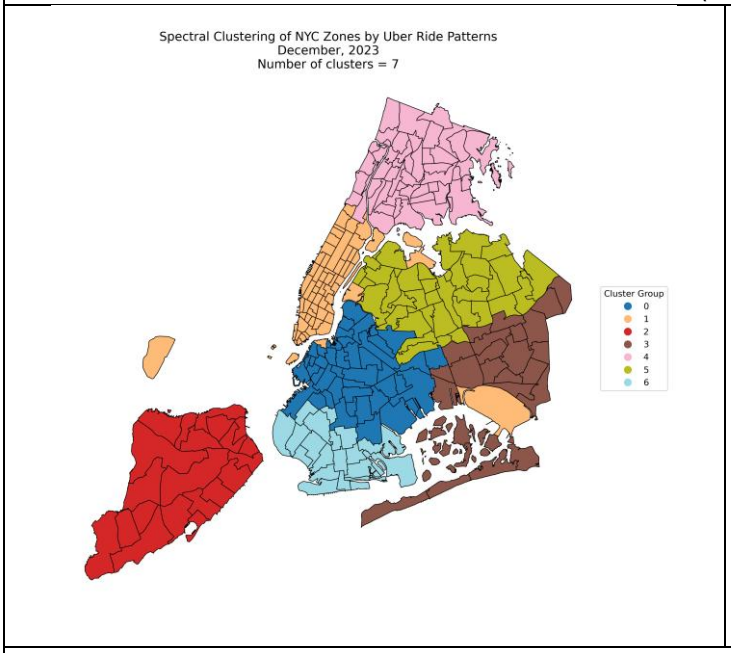
<table>
<tr><td colspan="2">Weeks of the U.S. Open</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td colspan="2">The two Weeks after U.S. Open</td></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>

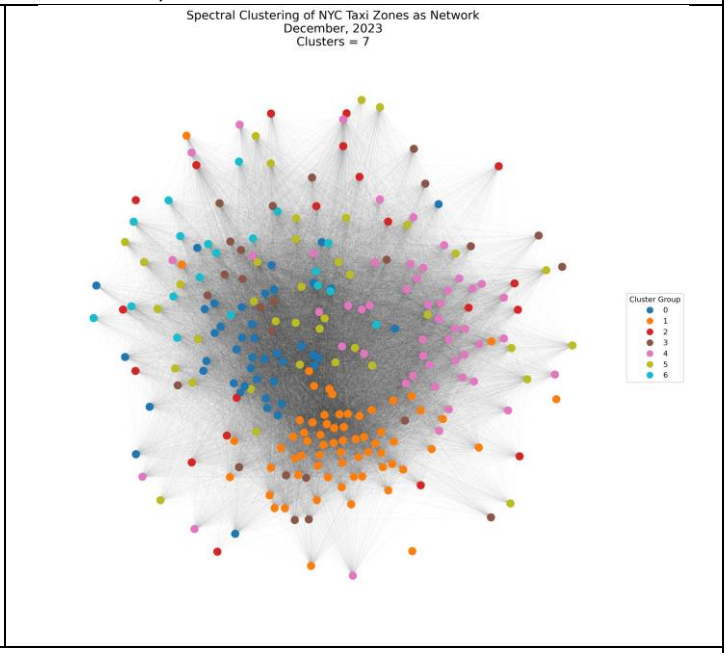
We observe a clear difference between the two maps: during the U.S. Open, the Willets Point neighborhood – home of the tournament – shifts into the same cluster as Manhattan. This indicates a surge in rides between Willets Point and Manhattan neighborhoods, rather than the typical pattern where Willets Point clusters with the nearby Flushing area. This shift highlights how large events can significantly impact transportation patterns, driving new connections between areas that are not typically linked as strongly!

Tourism Season: (December vs January)

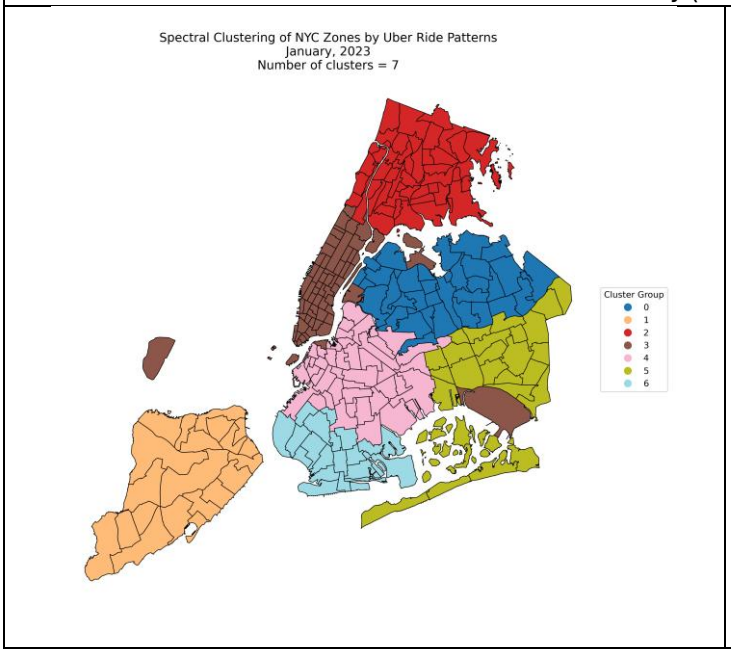
We next examine the peak tourist season (December) versus the low tourism period (January) to explore whether there are notable differences in the network's behavior. While we observe no major shifts in which neighborhoods are assigned to which clusters, there are important differences in the strength of the clusters themselves.

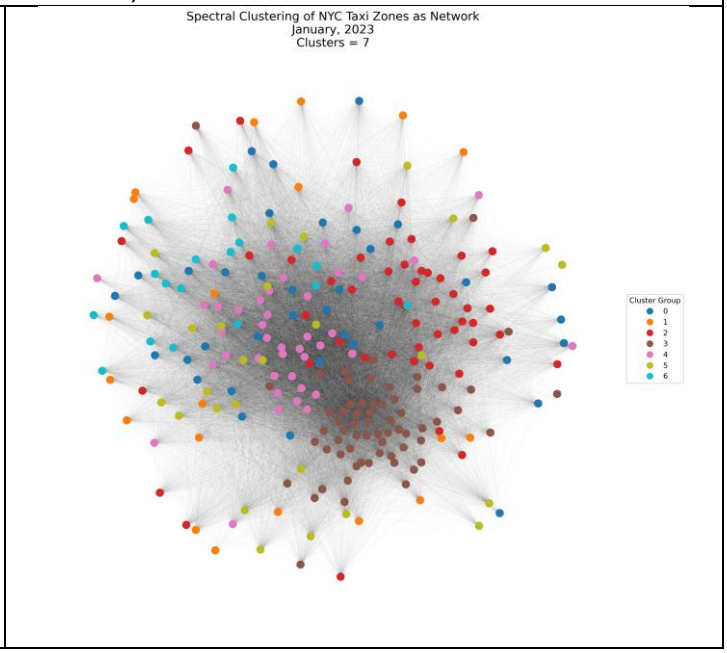
December (Peak Tourism)





January (Low Season)





For this comparison, we focus on the Manhattan cluster (orange in December, brown in January).

In December – during peak tourism – the Manhattan cluster appears more tightly separated from nearby clusters such as the Bronx (pink in December, red in January) and Brooklyn (blue in December, pink in January). This

separation indicates that during the tourist-heavy month of December, Manhattan becomes more internally connected, with a higher proportion of rides happening within Manhattan rather than to or from other boroughs.

This result matches our expectations: tourists are more likely to stay within the core of the city, rather than traveling to outer boroughs like the Bronx or Brooklyn. Thus, there are fewer Uber rides crossing borough boundaries during this period, and more rides happening entirely within Manhattan.

This is a conclusion we could not easily draw from simple ride counts alone – spectral clustering reveals not just where rides happen, but how strongly connected neighborhoods are within clusters, giving us deeper insights into how New York behaves during major tourism seasons.

Conclusions

Through our analysis of over 160 million Uber rides in New York City during 2023, we were able to uncover fascinating insights into how the city's neighborhoods are connected – and how those connections shift in response to major events and seasonal patterns.

Using spectral clustering on neighborhood-to-neighborhood trip volumes, we identified commuting communities that are often intuitive (such as Staten Island's isolation or the Bronx forming its own tight cluster) as well as deeper, less obvious dynamics (like airport neighborhoods clustering with Manhattan). Importantly, we found that the city's connectivity is not static; it adapts based on external factors such as major events and tourism seasonality.

During the U.S. Open, we observed a striking shift where Willets Point – typically clustered with nearby Flushing – moved into the same cluster as Manhattan. This change clearly illustrates how a large, concentrated event can temporarily alter travel patterns, linking neighborhoods that are not usually strongly connected. Similarly, when comparing December (peak tourism season) to January (low season), we found that Manhattan became more internally cohesive during December, reflecting tourists' tendency to stay within the city's core rather than venturing to outer boroughs.

These findings demonstrate the power of network-based clustering to reveal not just where people travel, but how urban mobility networks flex and respond to changes in demand. For businesses, city planners, or even Uber drivers, understanding these patterns can provide a strategic advantage – from optimizing service locations to planning better for surge periods tied to events or seasonal trends.

Overall, our work highlights that New York's transportation patterns are dynamic and event-driven, and that clever analysis of ride data can uncover the living, breathing movement of the city in a way that traditional analysis often cannot.

Sources and References:

538's taxi zone lookup: https://github.com/fivethirtyeight/uber-tlc-foil-response/blob/master/uber-trip-data/taxi-zone-lookup.csv

538's previous research on Uber vs Lyft vs Taxi in NYC: https://fivethirtyeight.com/features/uber-is-serving-new-yorks-outer-boroughs-more-than-taxis-are/

OpenData NYC's FHV data, 2023: https://data.cityofnewyork.us/Transportation/2023-High-Volume-FHV-Trip-Data/u253-aew4/about_data

Kaggle Shapefile of NYC, used to construct GIS visuals: https://www.kaggle.com/datasets/ahmadrezarostamani/nyc-taxi-zone-shapefile?resource=download&select=taxi_zones.shp

NYT's Neighborhood map of NYC: https://www.nytimes.com/interactive/2023/upshot/extremely-detailed-nyc-neighborhood-map.html?unlocked_article_code=1.6kw.kcs8.he_hQaxqP5Vb&smid=re-nytimes