IDVE Assignment 2

Ilona Smith - 2471308
Luca von Mayer - 2427051
Tumi Jourdan - 2180153
Tao Yuan - 2332155

[1]*School of Computer Science and Applied Mathematics, University of the Witwatersrand*

Abstract

This report details the process of applying cleaning techniques and data exploration on the NYC Uber dataset given for IDVE Assignement 2. Observations are provided in plots, descriptive statistics and explanations for relevant sections.

1 Data Cleaning

- Trip duration
- Speed
- Longitude and Latitude

1.1 Speed and Trip duration

Both of these were cleaned using IQR, with a scaled range by 12. Even though longitude and latitude do the bulk of the cleaning, these are important for later questions and data analysis. Some trip speeds that fell in the lower bound were valid (due to high traffic) and so the upper bound was set by IQR, but the lower bound set to zero to clean any negative values.

1.2 Longitude and Latitude

This was cleaned by applying IQR to the dataset. The correct range for the IQR cleaning is determined by applying from 0 to 20 and inspecting each graph to see when the most data is cleaned with the least amount of important data loss (see figure 1. The range scalar we chose was 12.

1.3 Distance

Distance cleaning did not improve the data much as it is encoded in speed, trip duration, and the co-ordinates which have already been cleaned.

1.4 Others

An attempt was made to use shape map as a bounds for the data, but it was far too aggressive and removed all data in the Pennsylvania state.

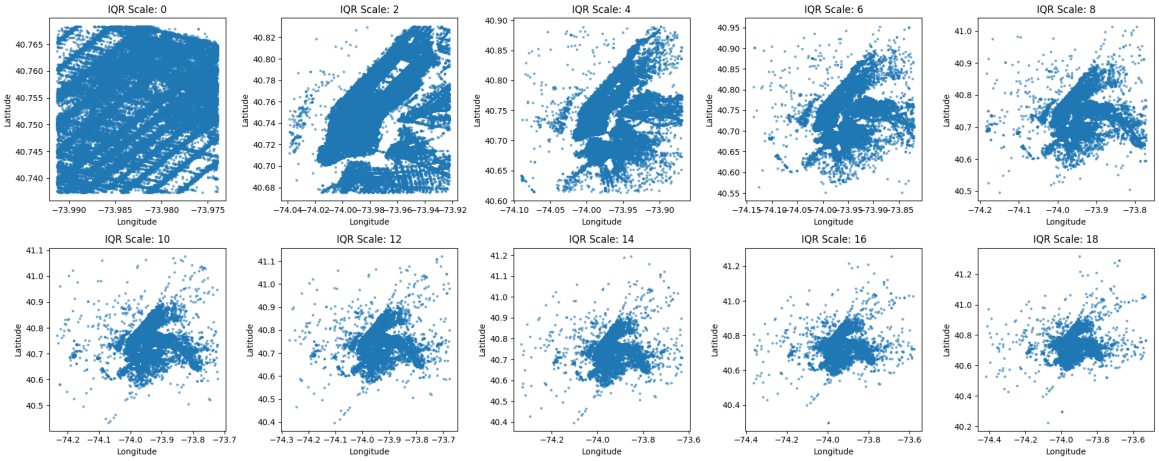


Figure 1: Added Features

2 Feature Generation

Additional derived features were designed to assist in the visualization of the data. These included

- Distance of trip - calculated as a straight line between two locations using the given start and end coordinates
- Day of week - derived using the date of the trip
- Average speed of trip - derived by taking the distance divided by time

	trip_distance_km	day_of_week	average_speed_kmh
0	1.497580	Monday	11.848984
1	1.804374	Sunday	9.797504
2	6.381090	Tuesday	10.815406
3	1.484566	Wednesday	12.457894
4	1.187842	Saturday	9.830418

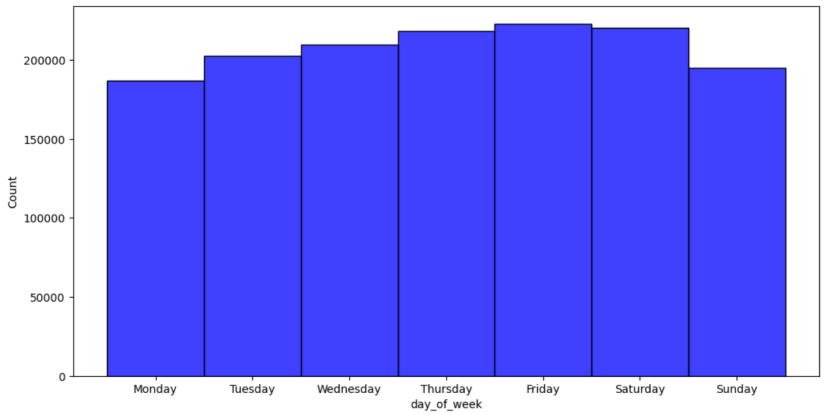
Figure 2: Added Features

3 Time-Based

In this section we explore the time series of the data. In particular we are looking for times when there are frequent trips and how this affects the speed of the trip.

3.1 Most popular day of the week

Friday and Saturday are the most popular days of the week, being roughly equal total trips after data cleaning.



day_of_week	
Friday	223101
Saturday	220393
Thursday	218180
Wednesday	209796
Tuesday	202376
Sunday	194933
Monday	187098

Figure 3: Popularity of each day of the week

3.2 Most popular hour of each day

During weekdays it appears that total trips begin increasing up around 7 AM until 9AM which corresponds to times people are expected to travel to work. In the afternoon roughly 6PM - 10PM a sharp increase is seen as this corresponds to times people finish their workday and commute home. During Friday an extended period of popularity for evening commute can be observed. This is due to the end of the week and people going for social gatherings.

During the weekends there is a shift showing high demand in late night hours (likely as people to night activities and home from them) and a later commute time and in lower frequency in the mornings. This may be due to the later wake up hours after retiring late and the lack of a need to wake up for work in the morning.

Sunday shows to be a low frequency in all hours of the day compared to other days. This indicates most people tend to not travel much on this day, either spending time with family or resting and preparing for the following work week. The only exception is the early hours of Sunday, likely people returning home after Saturday night activities.

Most popular hours:

- Monday: 6 PM
- Tuesday: 6 PM
- Wednesday: 7 PM
- Thursday: 9 PM
- Friday 7 PM
- Saturday: 11 PM
- Sunday 12 AM

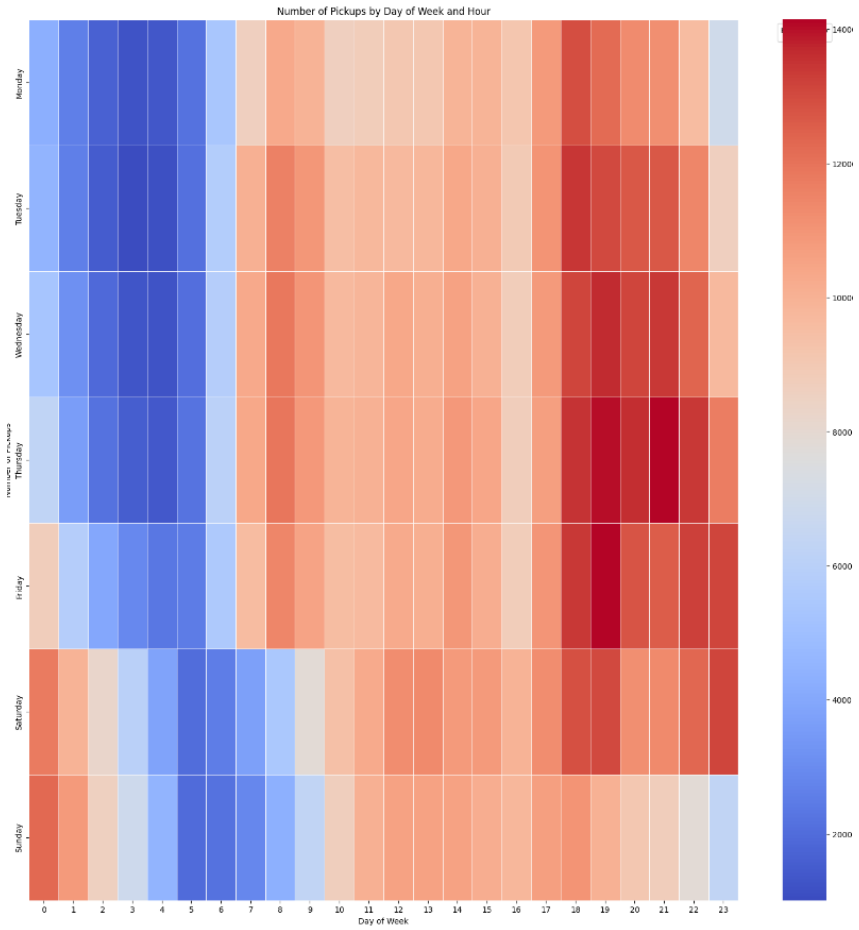


Figure 4: Heat map showing total trips by hour

3.3 Weekends vs Weekdays

To find the differences between weekends and weekdays the following factors were considered:

- Average duration per trip: Weekdays consisted of slightly longer trips on average. Possibly due to the higher times spent in traffic.
- Average distance per trip: Weekend is slightly higher distances than weekdays. This may be due to passengers living close to their, resulting in a shorter travel distance during the week
- Average speed per trip: Weekends have a higher average speed than weekdays. Likely due to lesser traffic.
- Total trips: Weekdays have a far higher number of people commuting than weekends. This is due to the need to travel to work during weekend and possibly preferring to stay at home during weekends.

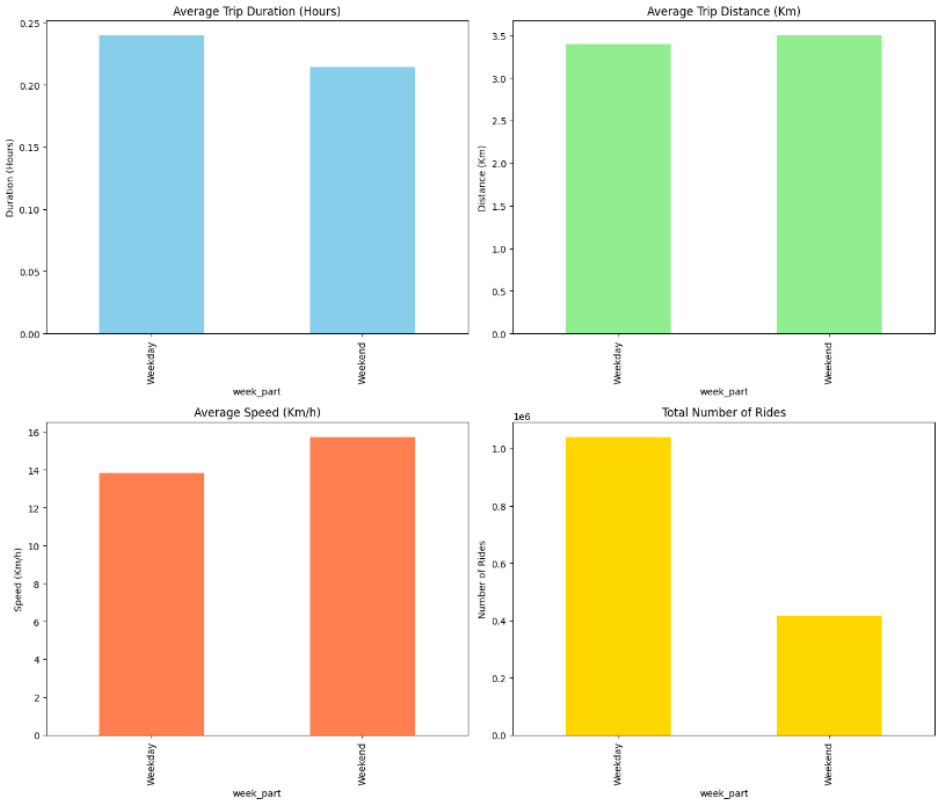


Figure 5: Weekend vs Weekday

3.4 Holidays

For each of the holidays, the data is filtered to include only entries where the date of the pickup_datetime is the same as the date of the holiday in 2016. The day of the week for the holiday is then identified using day_of_week. To establish a baseline for comparison, average hourly trip data is calculated for all other occurrences of that same weekday throughout the year, excluding the holiday itself. Major holidays are expected to result in deviations from the standard commute due to shifts in interest. The following holidays were selected to be investigated.

- St. Patrick's Day (Thursday 17 March 2016)
- Easter Sunday (Sunday, 27 March 2016)
- Memorial Day (Monday 30 May 2016)
- Valentine's Day (Sunday 14 February 2016)
- Martin Luther King Day (Monday 18 January 2016)

The following metrics were investigated to be compared to the same day on average

- Hourly trip distribution
- Total trips
- Peak hour

The patterns of taxi trips on major holidays show varying degrees of change compared to the average for their respective weekdays, with some holidays displaying significant differences while others remain largely similar to typical patterns.

No large variations were observed during St. Patrick's Day or Easter. In contrast, the other three holidays display more noticeable variations. Martin Luther King Day shows lower trip volumes during typical morning commute to work times compared to an average Monday, but sees an increase in afternoon trips. Valentine's Day, occurring on a Sunday, exhibits a substantial surge in afternoon and evening trips, likely due to dinner dates and celebrations. Memorial Day demonstrates the most dramatic departure from the norm, with significantly fewer trips throughout the day from 5am compared to an average Monday.

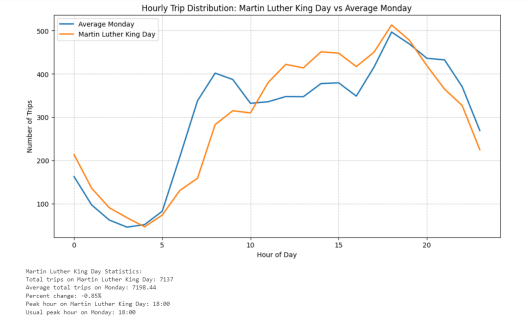


Figure 6: MLK trip distribution

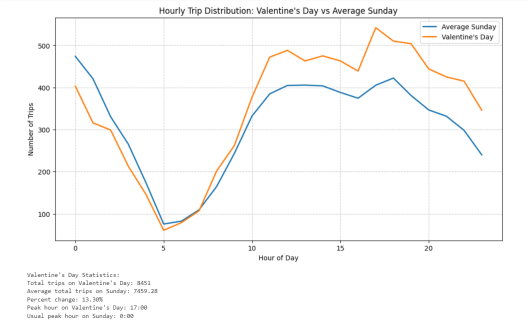


Figure 7: Valentines trip distribution

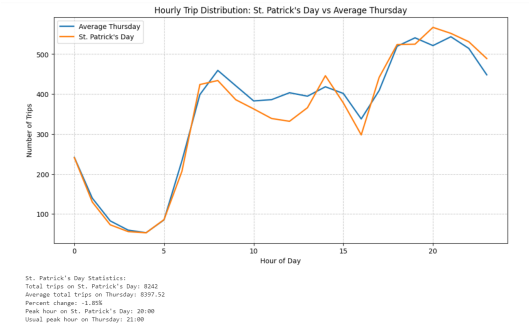


Figure 8: St. Patrick's Day

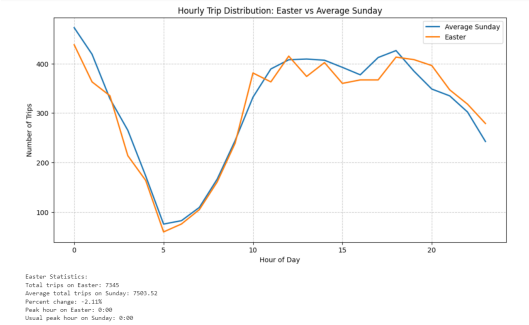


Figure 9: Easter Sunday

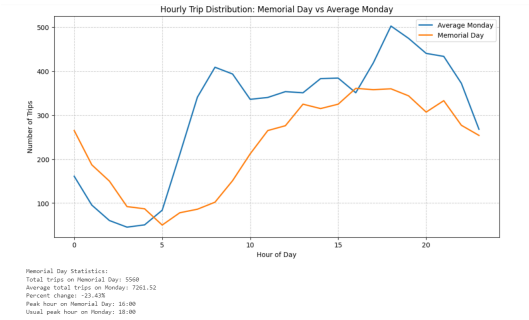


Figure 10: Memorial Day

3.5 Average speed

The average speed is calculated based on the time taken of the trip and the distance travelled. It increases until a maximum of 19km/h at 5 AM then sharply decreases between 6-8AM, likely due to increased traffic. It remains low until 3PM with slight increases until 6PM before slowly climbing up towards 15 km/h at 11PM. This speed makes sense according to https://www.nyc.gov/html/dot/downloads/pdf/mobility-report-2018-screen-optimized.pdf

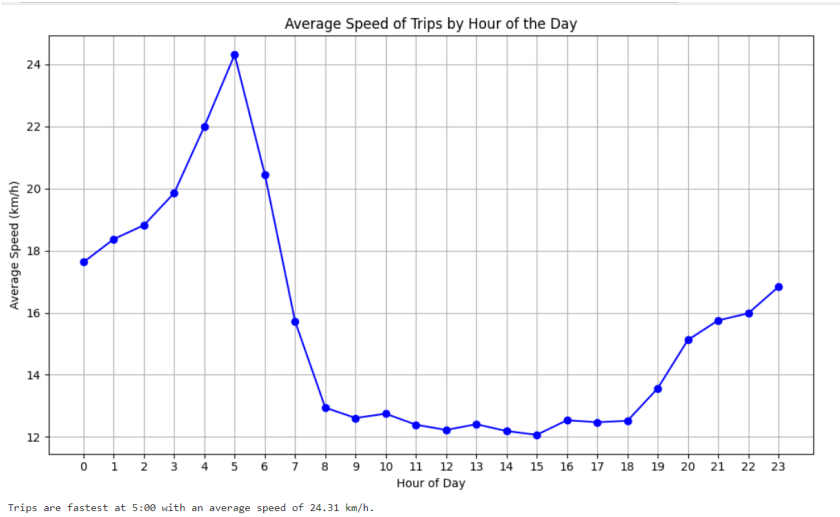


Figure 11: Average speeds throughout the day

4 Location Clusters

4.1 Heatmaps

- During the week , mid town east sees the most traffic, and this traffic is during the evening when people are going home, indicating a business district. This aligns with reality as midtown east encompasses one of the biggest business districts. See Figure 14.
- The Weehawken area across the Hudson also exhibits high pickup rates in the morning , but low in the evening indicating this is a residential area

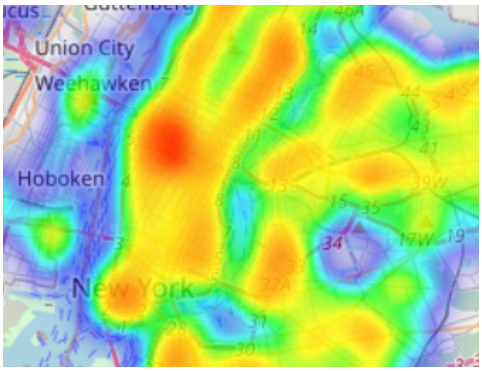


Figure 12: Morning

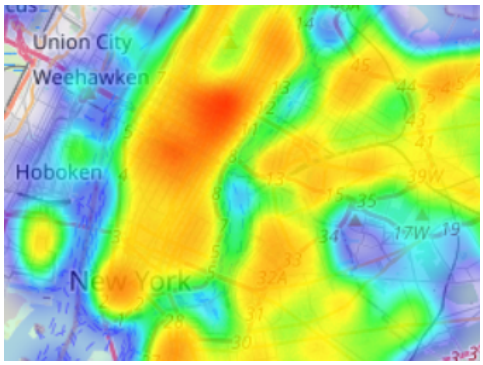


Figure 13: Evening

Figure 14: Side by side of morning and evening

4.2 Hotspots

Generating meaningful clusters required very high hyperparameters due to the high density of pickups. If the minimum and range were too low the data would display all of Manhattan as a single cluster, which is not meaningful or helpful.

Setting the minimum cars to 500 with a max distance of 150m created 14 clusters. Referring to figure 15 the clusters are mostly in the southern half of Manhattan, with a few clusters in the airports : JFK and LaGuardia.

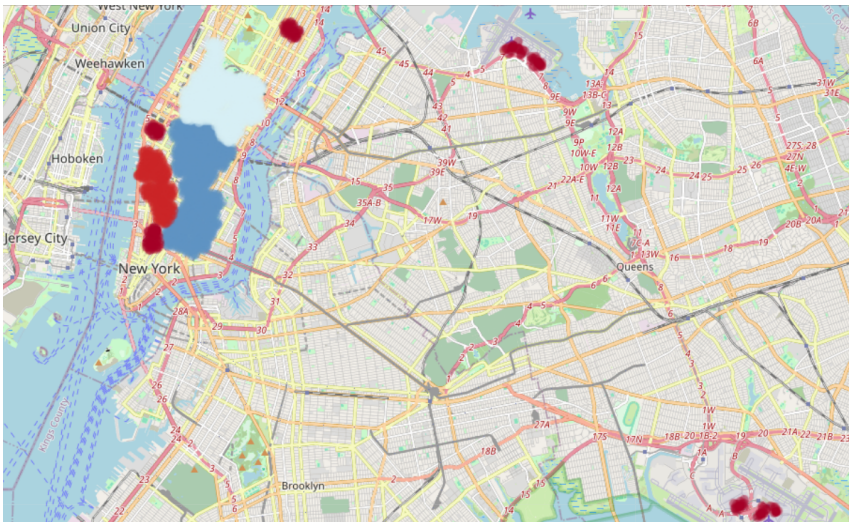


Figure 15: Hotspots

5 Airports

First, the coordinates for key locations (JFK Airport, Empire State Building, and Newark Airport) are defined. A function utilising the provided haversine formula is created to determine if a trip's pickup and dropoff point is within a specified radius of these locations. The data is then filtered to isolate trips originating near the Empire State Building and ending near either JFK or Newark Airport. This creates two separate datasets for each airport route.

For each route, a function plots the average travel times by hour of the day. It processes the data by extracting the hour from the pickup time and calculates the mean trip duration for each hour. This information is visualized in a line plot, showing how travel times vary throughout the day. The function also calculates and prints summary statistics: the overall average travel time, median travel time, and the total number of trips analyzed for each route.

Average travel time to JFK: 0.8 hours

Average travel time to Newark: 0.63 hours

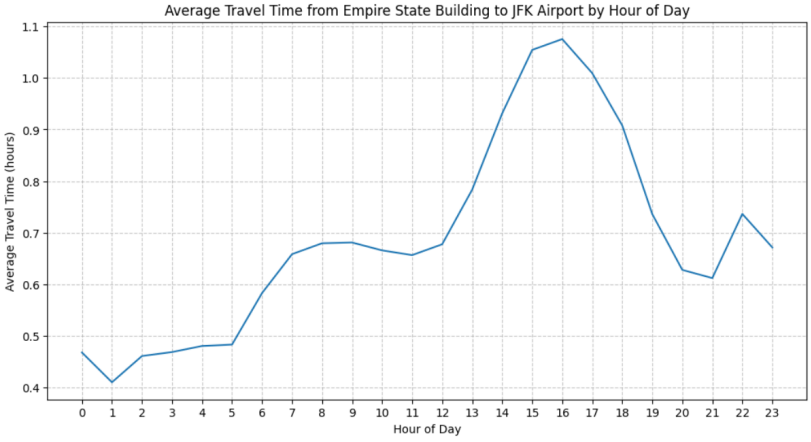


Figure 16: Empire State Building to JFK Airport

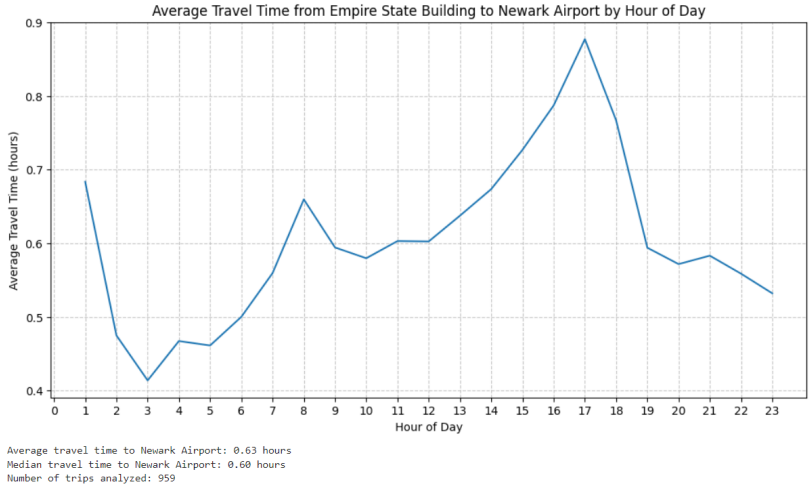


Figure 17: Empire State Building to Newark Airport

On average the travel time from the Empire State Building to Newark Airport is shorter than travelling to JFK airport. The longest average travel times take place in the afternoon with the peak travel time to JFK airport being at 4pm, and 5pm to travel to Newark airport.

A radius of 1.11km around the center of the locations was used. This is significantly large enough to cover the Empire state building pickups, and dropoffs in and around the airport while accounting for passengers that wish to get on/off a small distance away from high traffic areas.

6 Boroughs

6.1 Pickup and Drop-off Neighbourhoods

Two fields were added to the dataset being the pickup and dropoff locations. The code loads NYC neighborhood boundaries as a GeoDataFrame, converts taxi pickup and dropoff locations into geographical points, and performs spatial joins to assign each pickup and drop-off location to a neighborhood based on whether the point falls within a neighborhood's boundary.

	pickup_neighborhood	dropoff_neighborhood
0	Lincoln Square	Upper East Side-Carnegie Hill
1	Murray Hill-Kips Bay	West Village
2	Midtown-Midtown South	Battery Park City-Lower Manhattan
3	SoHo-TriBeCa-Civic Center-Little Italy	Battery Park City-Lower Manhattan
4	Upper West Side	Upper West Side

Figure 18: Pickup and Drop-off Neighbourhoods

6.2 Pickup and Drop-off Chloropeth

The code counts the number of taxi pickups and drop-offs for each neighborhood, adds these counts to the NYC neighborhood GeoDataFrame, and creates two choropleth maps to visualize the distribution of pickups and dropoffs across the city. It also prints summary statistics, including total pickups/dropoffs and the neighborhoods with the highest activity.

We notice that majority of the pickups and dropoff are happening within Manhattan with the Midtown South area having the most pickups and dropoffs. Outside of Manhattan the areas with the most pickups and dropoffs are LaGuardia Airport and JFK Airport.

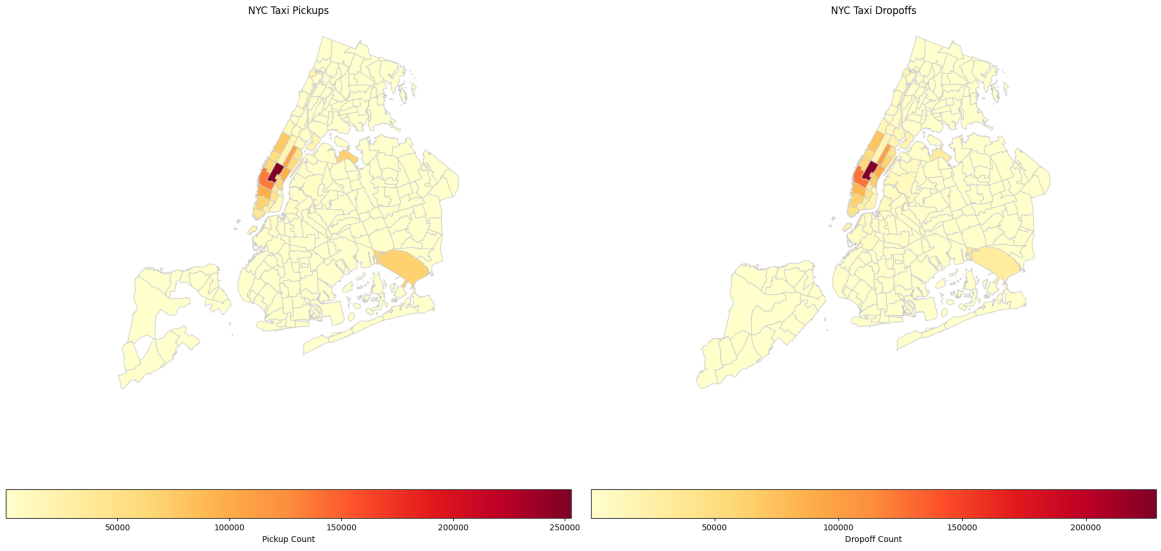


Figure 19: Pickup and Drop-off Chloropeth

6.3 Quietest Neighbourhoods

The code filters the taxi trip data for rides that occurred between midnight and 5 AM, creates GeoDataFrames for both the pickup and dropoff locations, and then performs spatial joins to map these points to NYC neighborhoods. It counts the night pickups and dropoffs for each neighborhood, calculates total activity (pickups + dropoffs), sorts the data, and displays the 10 quietest neighborhoods between midnight and 5AM.

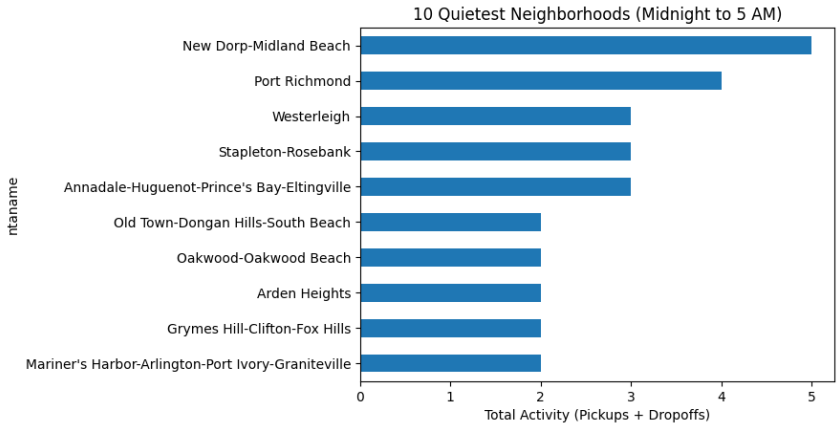


Figure 20: Quietest Neighbourhoods

6.4 Busiest Neighbourhoods

This uses the same total activity as calculated above but sorts the data in descending order to produce the busiest neighbourhoods.

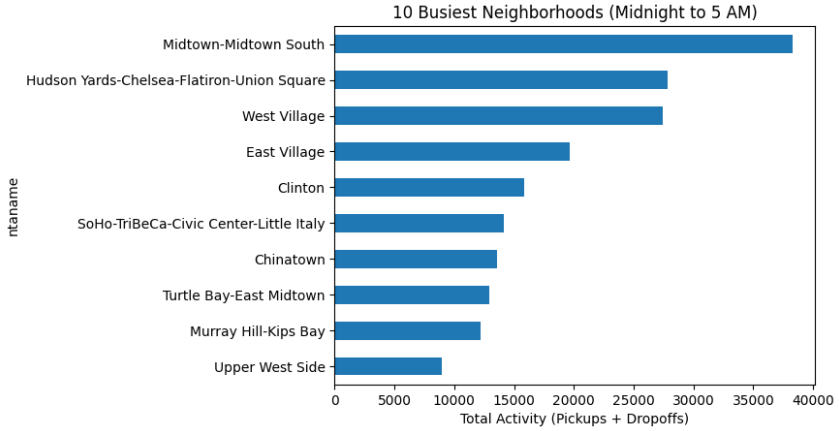


Figure 21: Busiest Neighbourhoods