Stuck in New York City Traffic? NYC Yellow Taxi Cab Rides from January 2022

New York City (NYC) is a mecca for traveling, working, and living. With approximately 60 billion people visiting the city each year, NYC maintains its reputation as a hub for business and pleasure travel. To support this large influx of visitors, NYC has robust transportation infrastructure in place, including yellow taxi cabs. With over 13,500 licensed taxis in the city, NYC is the taxi capital of the US. And with many in NYC forgoing owning cars and the ever-changing landscape of ride-sharing apps, my team and I utilized taxi ride records from January 2022 to learn how yellow taxi cabs were financially performing, distributed throughout the city, and key trip characteristics impacting rides.

Our data, initially sourced from the NYC Taxi and Limousine Commision (TLC), contained ~2.5 million rides for January 2022. It is important to note that during this time, NYC was in its final stages of lifting COVID-19 restrictions on travel. Therefore, we treated our dataset as if it was representative of typical NYC taxi rides throughout the month. Our columns included trip & fare characteristics (trip length, total fare amount, additional chargers), spatial data (origin and destination of trips), payment method information, and temporal data. These columns include those originally in the dataset and those added by team members for easy analysis. With this information, we focused on 4 main areas of information (spatial analysis, cost/payment analysis, temporal trends, and trip characteristics) in order to gain a full understanding of yellow taxicab trips with the intention of better understanding the landscape of taxi rides in NYC.

Data cleaning and analysis were completed using the Python programming language and libraries such as Pandas and Matplotlib. The complete code for each part of our work can be found here: temporal-trends, trip-characteristics, spatial-analysis, and payment/cost-analysis on our github-repo.

Data Cleaning

Before team members could get to the bulk of analysis, it was necessary to modify our original dataset. The largest change was merging our two datasets together. One dataset simply included the records of rides while the other (sourced directly from the NYC TLC website) included the borough and their corresponding TLC Taxi Zone LocationID. To streamline analysis, we merged these two datasets together.

We then dealt with any missing data from our added `Pickup_Borough` and `Dropoff_Borough` columns by identifying LocationIDs missing their corresponding borough and mapping the correct value.

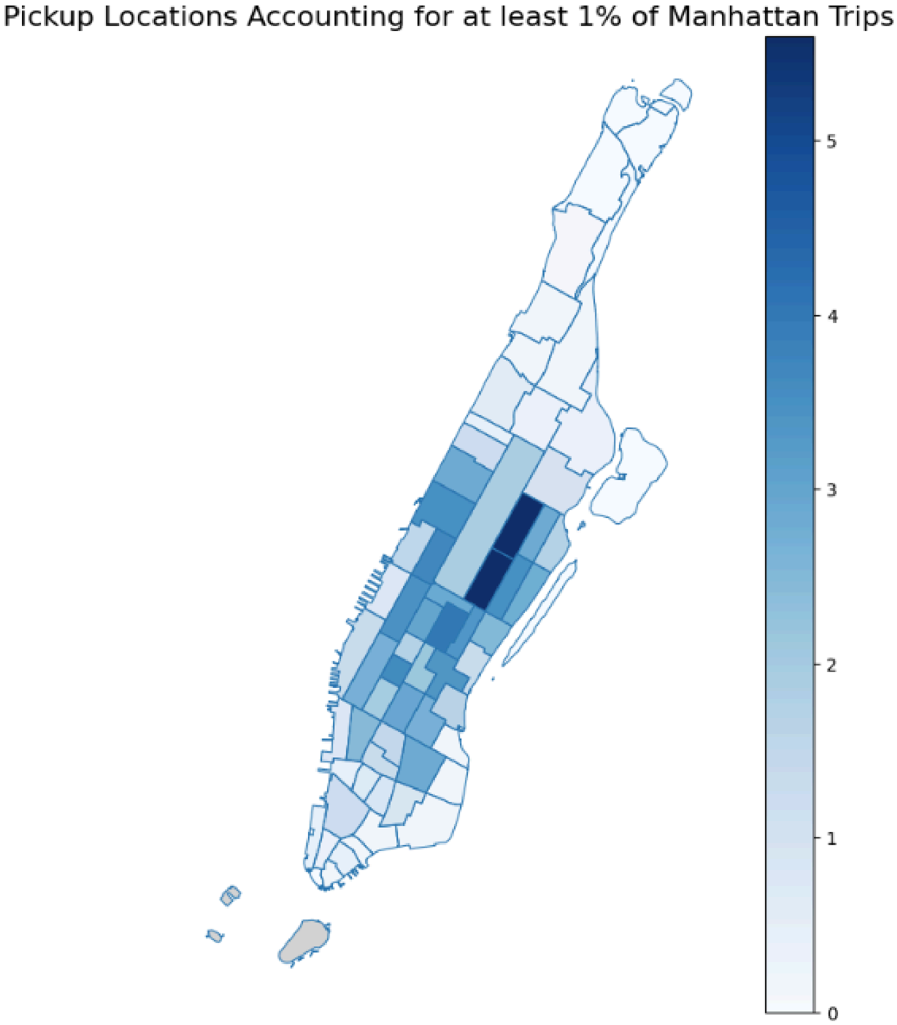
Other data cleaning involved checking columns with expected limited unique values and shifting data types. The code used for these tasks can be found here (link to one of our Github repositories).

Spatial Analysis

Spatial analysis included identifying common origins and destinations for trips to understand the distribution of taxi cabs throughout the city. We started with simply identifying the most common pickup and dropoff boroughs. Calculations revealed that Manhattan was accounting for 90.3% of pickup locations and 89.8% of dropoff locations. A full breakdown of boroughs and their corresponding percentages can be found below:

Pick up Borough	**Drop off Borough**
Manhattan - 90.3%	Manhattan - 89.8%
Queens - 7.72%	Queens - 4.57%
Unknown - 1.21%	Brooklyn - 3.65%
Brooklyn - 0.56%	Unknown - 1.06%
Bronx - 0.16%	Bronx - 0.73%
EWR - 0.01%	EWR - 0.16%
Staten Island - 0.01%	Staten Island - 0.02%

It is important to note the similarities within ride distribution between pickup and drop off locations. This suggests that taxi rides follow fairly predictable routes in terms of boroughs being traveled to and from. This heavy concentration of traffic in Manhattan created a need to zoom in a bit more and understand the exact LocationID for taxi rides. Starting with Manhattan Pickup LocationIDs (PUIDS), we saw that the Upper East Side of Manhattan was seeing a majority of taxi ride origins. Below is a map based on the official TLC taxi zones map of Manhattan indicating where these PUIDS are located. This map and each subsequent one in this section is colored based on how many rides specific LocationIDs are receiving. The darker the blue, the more rides to and from that destination.


Pickup Locations Accounting for at least 1% of Manhattan Trips
5
4
3
2
1
0

LocationID 236 and LocationID 237 are situated right next to central park and include several subway stations, college campuses, hotels and notable attractions such as the Guggenheim Museum and the Smithsonian. These characteristics help to explain why both are accounting for 5.6% of Manhattan origin locations.

The most common dropoff boroughs for Manhattan were also LocationIDs 237 and 236, suggesting that these locations consistently see high volumes of taxi cab traffic.

Trip Length Variations

Our next steps were identifying if and how most common pickup and dropoff locations differed based on trip length. Broken up into very short trips (>1 mile), short trips (1-5 miles), long trips (5-10 miles) and very long trips (10+ miles), we only start to see deviations in taxi ride origins and destinations when we started looking at long trips between 5 and 10 miles.

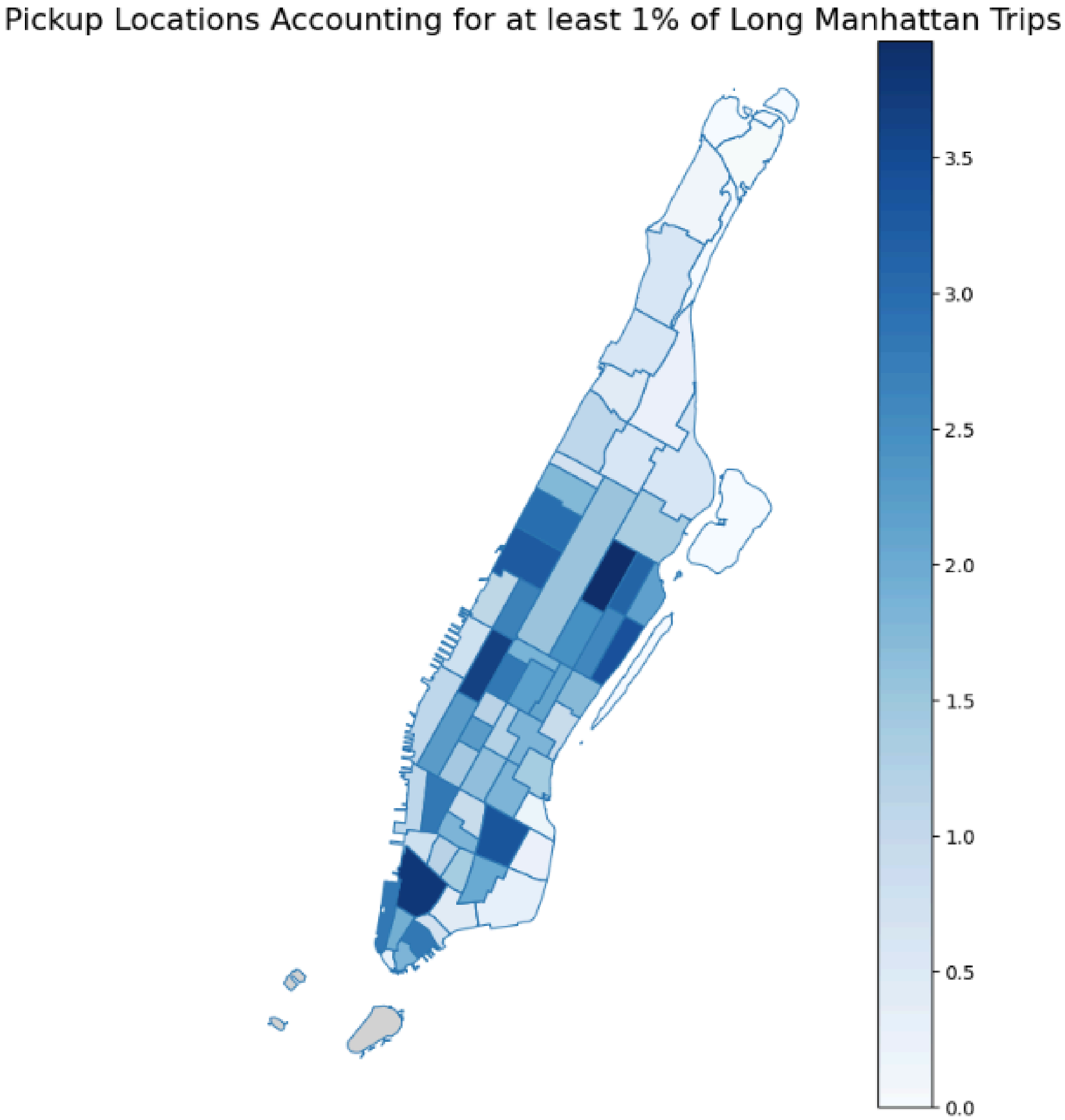
5 - 10 Mile Trips

For our long trips, Manhattan continued to be a hotspot for origins and destinations, accounting for 69.4% of pickups and 61.4% of drop offs. However, Queens also starts to account for some trips, with 25.7% of long trips starting in Queens and 17.9% ending in Queens. A full breakdown of the percentages can be found below:

Pick up Borough	Drop off Borough
Manhattan - 69.4%	Manhattan - 61.4%
Queens - 25.7%	Queens - 17.9%
Unknown - 2.94%	Brooklyn - 16.7%
Brooklyn - 1.65%	Bronx - 2.73%
Bronx - 0.36%	Unknown - 1.32%
EWR - 0%	EWR - 0%
Staten Island - 0%	Staten Island - 0%

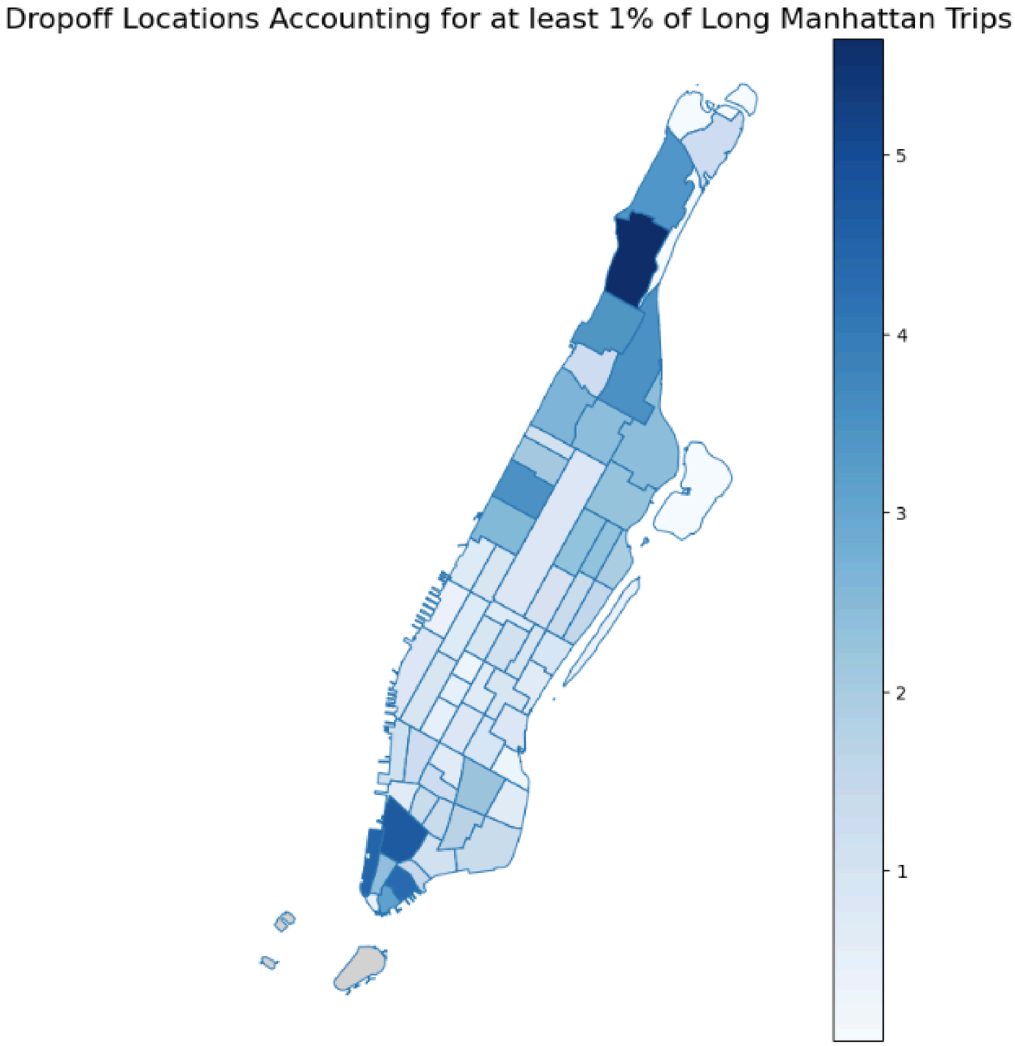
During 5-10 mile trips is when we see deviations from the borough percentages pattern previously apparent. While Manhattan continues to be a hotspot for both origins and destinations, Queens sees an increase in taxi rides coming from this borough. The changes here might suggest that trips between 5-10 miles are more varied in their routes.

Once again starting with Manhattan, the map of ride taxi distribution for PUIDS highlights two neighborhoods.

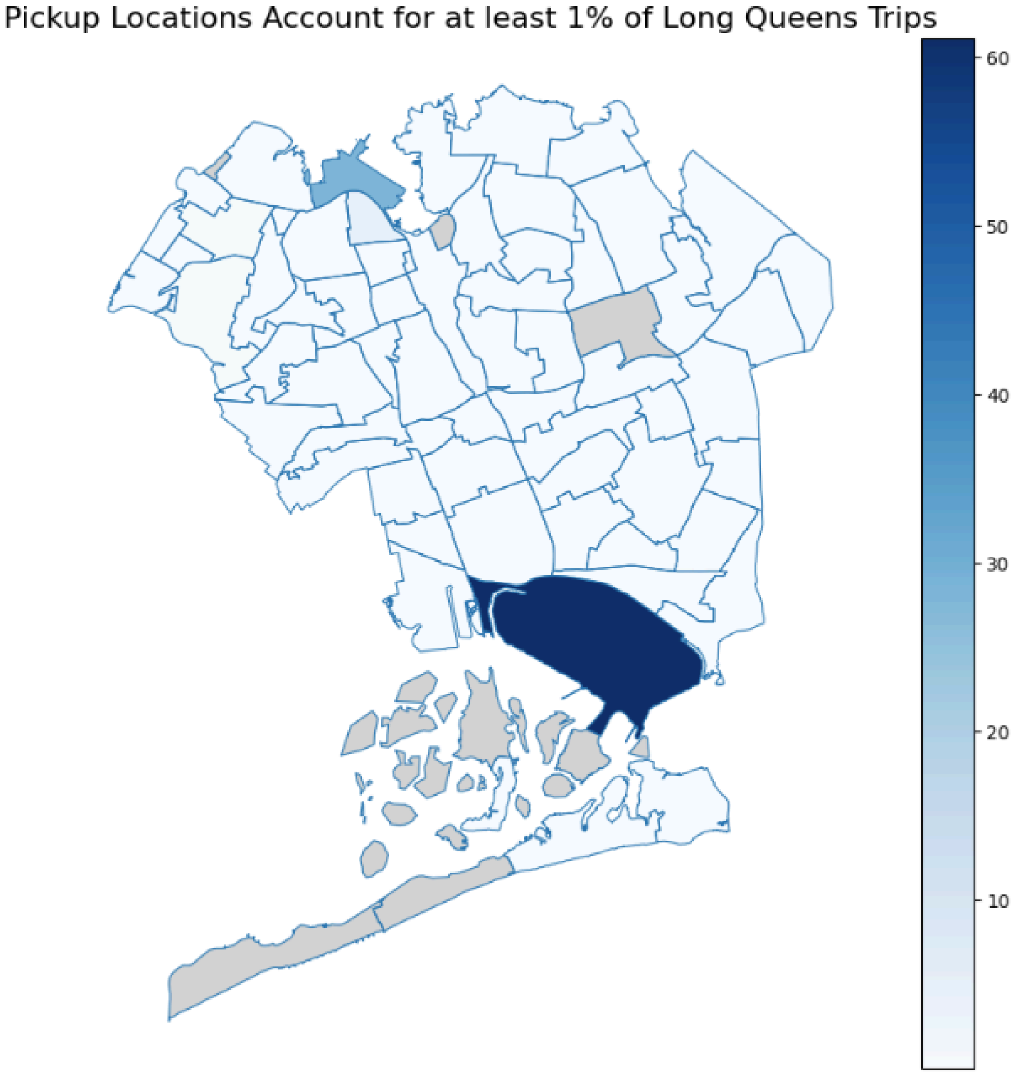

Pickup Locations Accounting for at least 1% of Long Manhattan Trips
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

The first of these is LocationID 231 which is located in the TriBeCa and Hudson Square neighborhood. More of a local hotspot rather than a tourist one, this location included 2 subway stations, hotels, and the Color Factory museum. Next is LocationID 48, which spans the Hell's Kitchen neighborhood and is just west of Times Square. Accounting for 3.66% of rides, there are several hotels in the area, subway stations, and a museum in this location.

As for Manhattan Drop Off LocationIDs (DOLIDS), 5.65% of rides end up in the south side of the Washington Heights neighborhood (LocationID 244). This destination has 6 subway stations within its boundary, some local attractions such as the United Palace Performing Arts Theatre and the Hispanic Society Museum & Library, as well as a hotel and medical center.


Dropoff Locations Accounting for at least 1% of Long Manhattan Trips
5
4
3
2
1

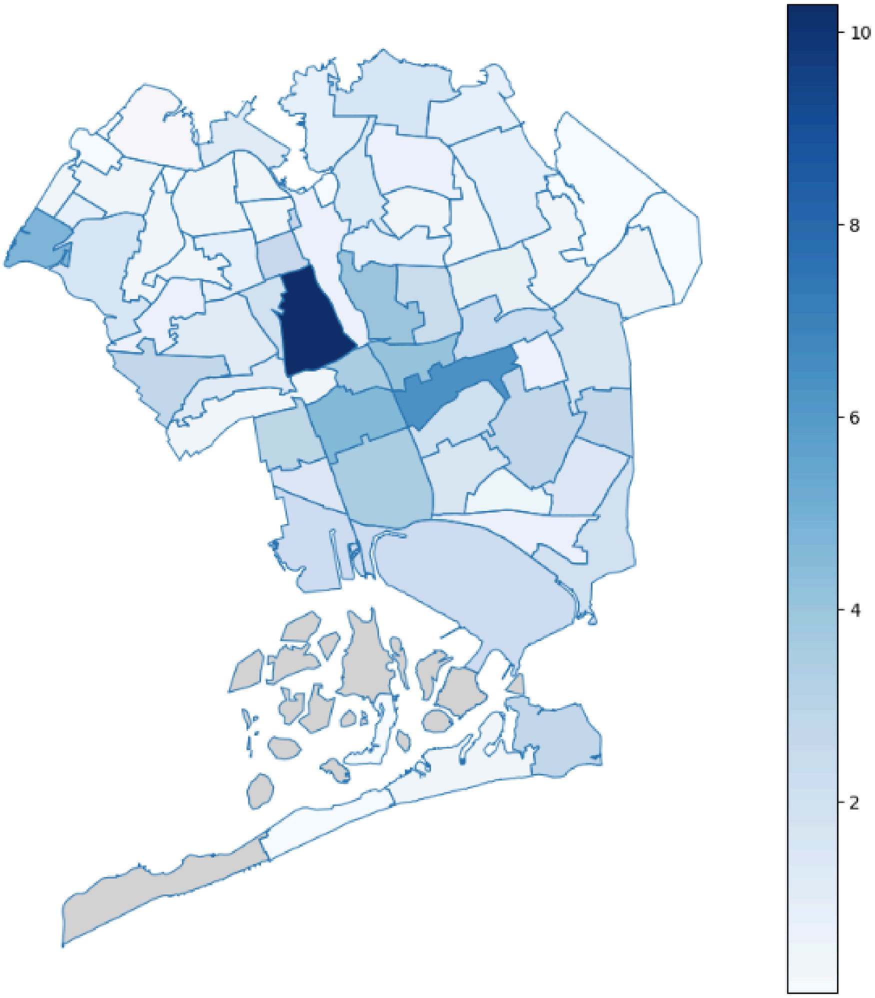
Moving away from Manhattan, we saw that Queens was drawing a significant amount of traffic, with most concentrated in LocationID 132, with 61.1% of trips originating from this borough.


Pickup Locations Account for at least 1% of Long Queens Trips
60
50
40
30
20
10

This location's high number of trips is easily explained by the fact that the PUID spans the JFK International Airport. As a city known for its high tourist population, one of the busiest airports in the U.S is going to draw quite a crowd.

We would like to note the high percentage of taxi rides coming from JFK. Our analysis showed that specific LocationIDs drawn in at max 16-17% of rides for that borough. LocationID 132 breaks that pattern because of the lack of other transit options in this area.

As for where Queens-bound taxis are traveling, we end up at LocationID 95, accounting for 10.3% of rides. Located in the Forest Hills neighborhood, the area has a subway station centrally located, both public school and university campuses, and a hospital in its boundary.


Dropoff Locations Account for at least 1% of Long Queens Trips

During analysis of 5-10 mile trips, the team noticed a pattern for LocationIDs accounting for a large portion of rides. Origins and destinations located within high-density neighborhoods with other transit options, lodging and local attractions were likely to be the busiest locations. This is a trend we will see reinforced when looking at very long trips over 10 miles in length.

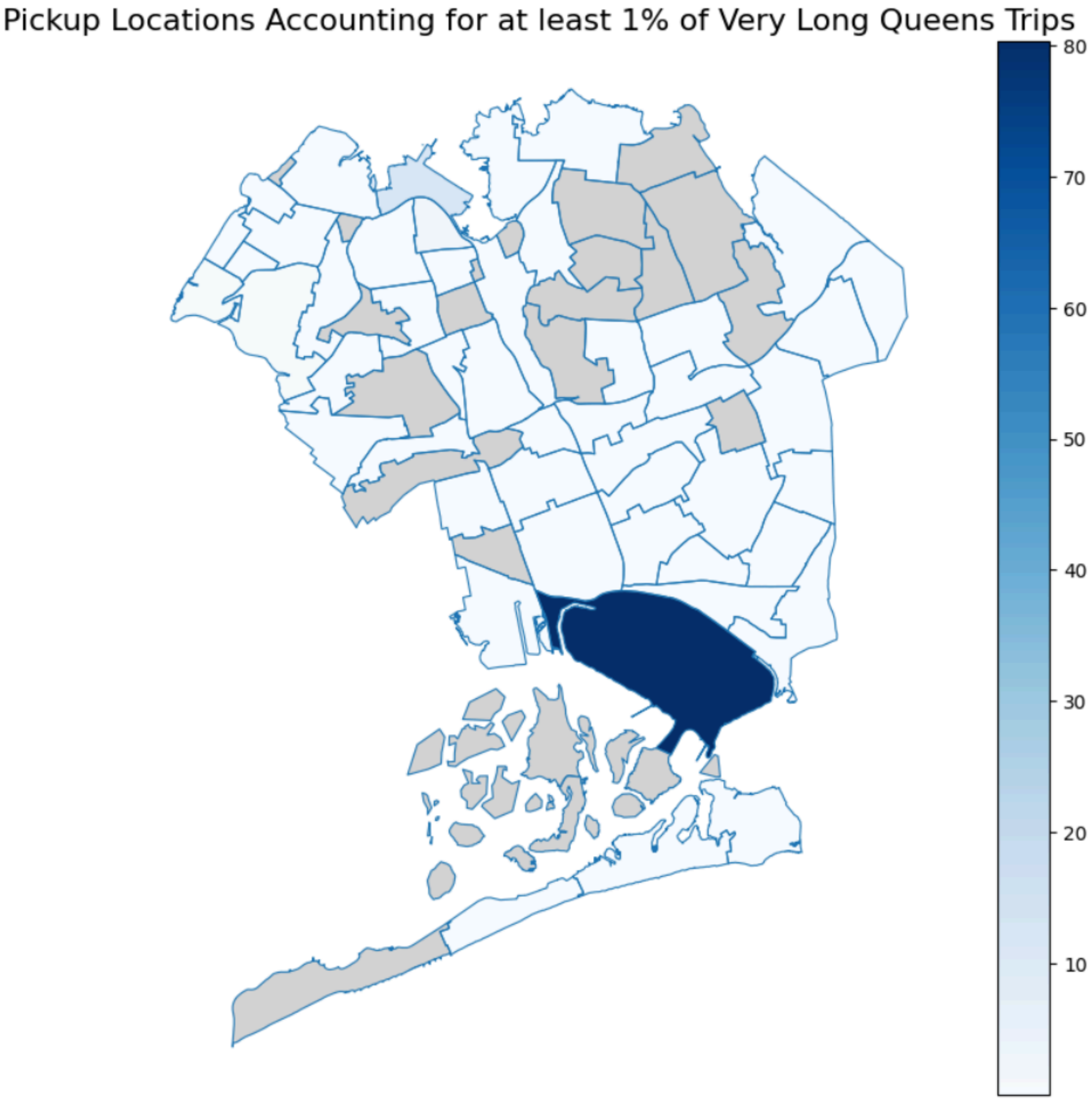
10+ Mile Trips

Finally, we analyzed the trips over 10 miles, which were characterized as very long. A full breakdown of the percentages can be found below but note that the pattern of similarity between common pick up and drop off locations breaks down. This suggests that the trend we saw in long trips continues into very long trips in that these rides are more varied than those traveling less than 10 miles.

Pick up Borough	Drop off Borough
Queens - 69.9%	Manhattan - 44.7%
Manhattan - 25.1%	Queens - 24%
Unknown - 3.66%	Brooklyn - 20.4%
Brooklyn - 0.92%	Bronx- 5%
Bronx - 0.40%	Unknown- 3.65%
Staten Island - 0.05%	EWR - 2.3%
EWR - 0.01%	Staten Island - 0.30%

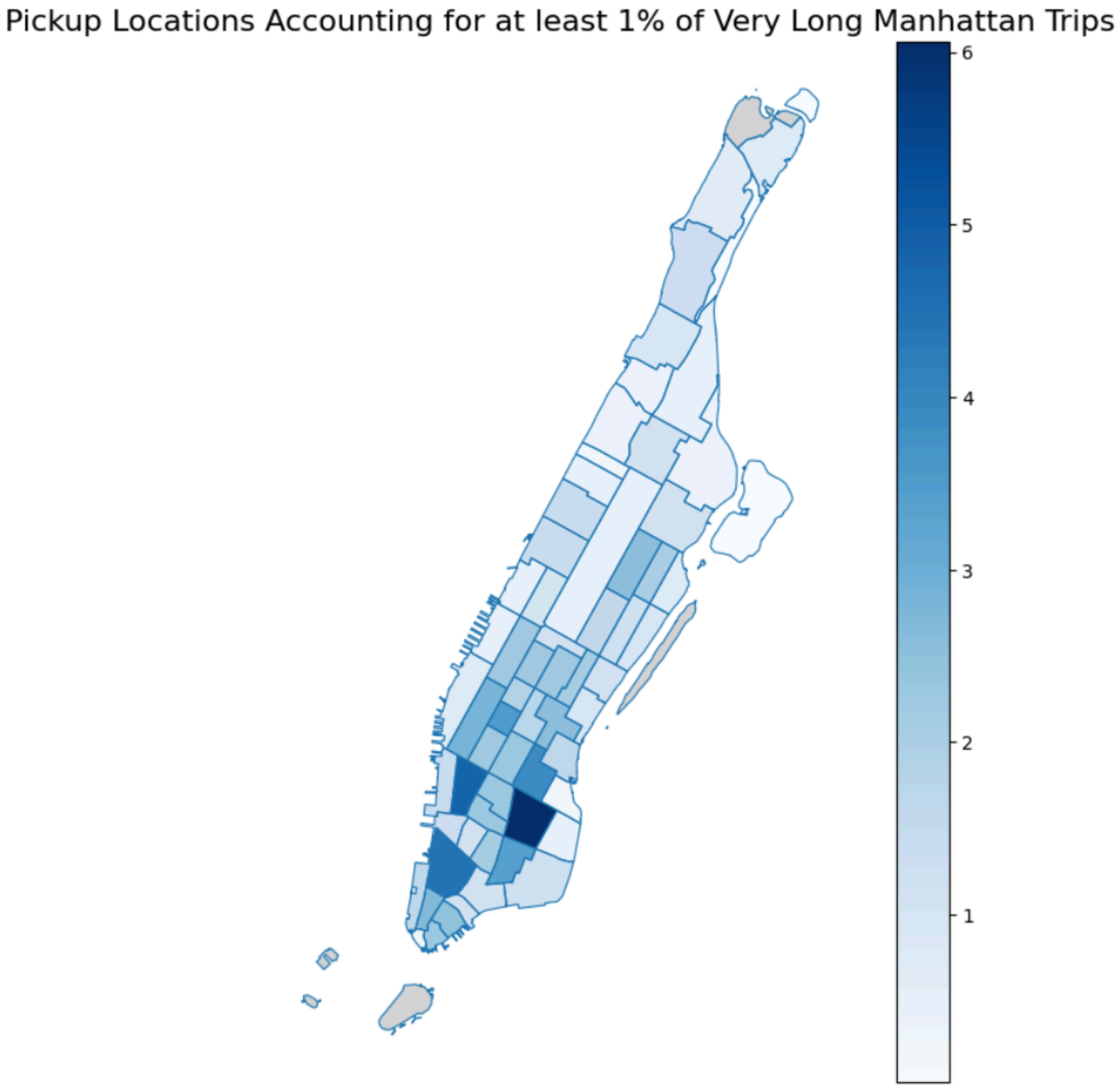
While a clear neighborhood for further analysis is less apparent, we continued to focus on Manhattan and Queens because while they do switch places between pickup and drop off location and have quite different traffic distribution across the two, they remained the top two boroughs for trips.

Starting with Queens, the most common PUID was once again JFK Airport (LocationID 132), accounting for 80.3% of rides.


Pickup Locations Accounting for at least 1% of Very Long Queens Trips
80
70
60
50
40
30
20
10

Rides arriving in Queens, however, are heading for LocationID 138 or LaGuardia Airport. This location represents 13% of rides. We saw previously that JFK Airport was bringing in 60 - 80% of rides so a natural question is why doesn't LaGuardia have the same type of taxi concentration. The simple answer is that LaGuardia is quite smaller than JFK. In fact, it is the smallest out of the three major airports in NYC.

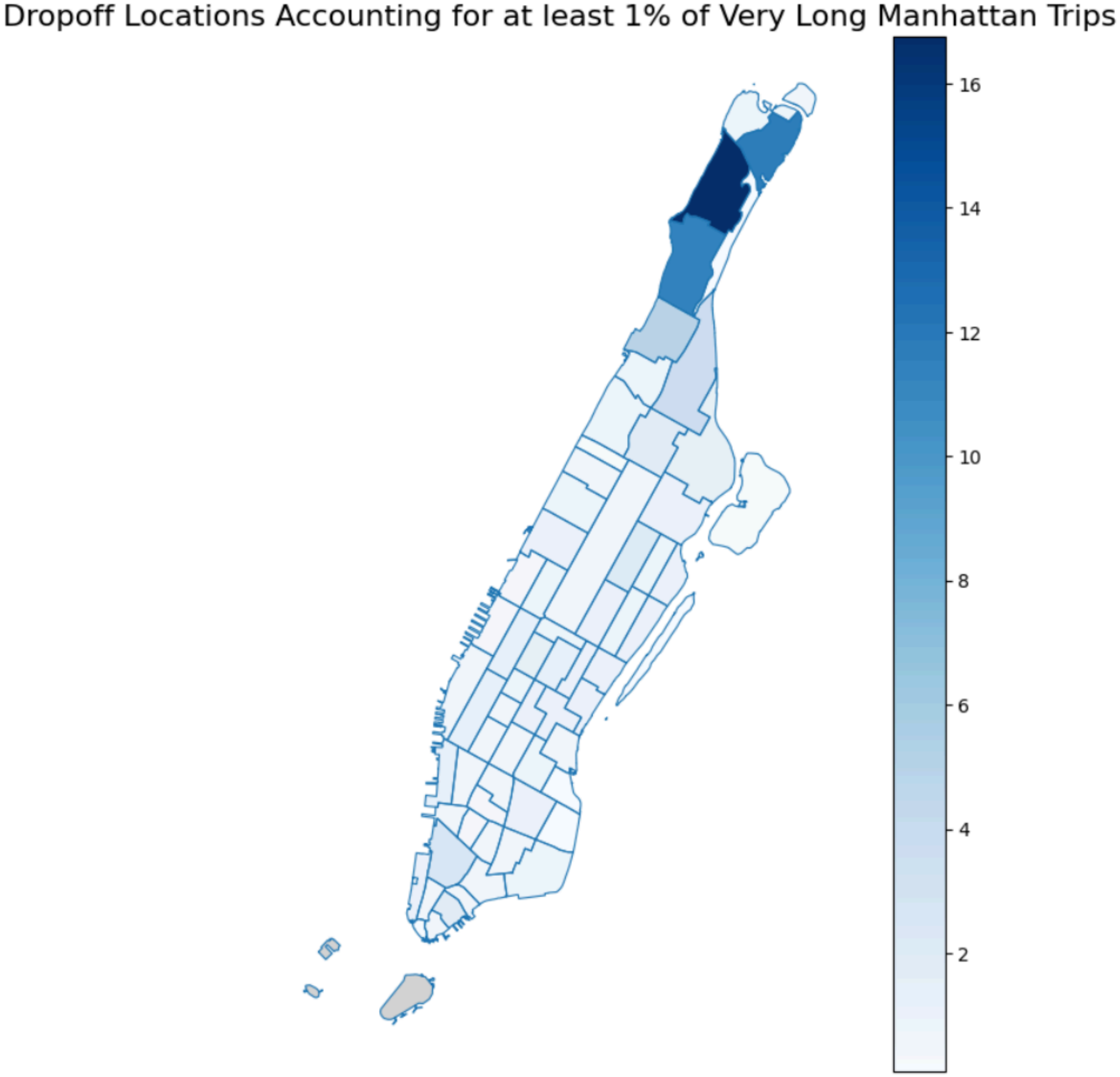
Finally, we return to Manhattan and see that rides continue to cluster on the southern end of the island. However, a new hotspot emerges with LocationID accounting for 6% of rides.



Pickup Locations Accounting for at least 1% of Very Long Manhattan Trips

Spanning the East Village, Alphabet, and Ukrainian Village neighborhoods, LocationID 79 includes 5 subway stations within its boundary as well as the local attraction 6 BC Botanical Garden. Given how LocationID 79 is located in Manhattan, we can easily see how it would be contributing a lot of traffic and rides to the area.

With the DOIDS for very long Manhattan trips, we move towards the northern tip of the island.


Dropoff Locations Accounting for at least 1% of Very Long Manhattan Trips
16
14
12
10
8
6
4
2

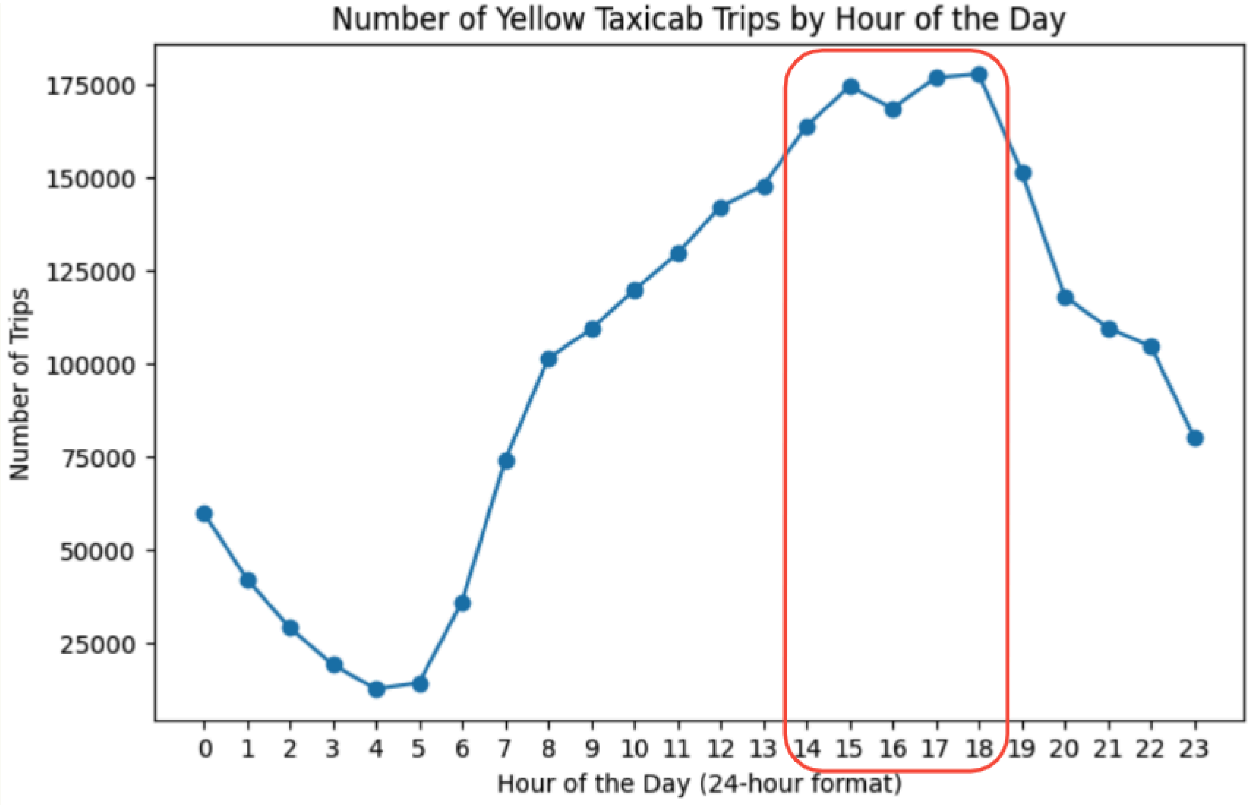
Here drop off LocationID 243 is accounting for 16.8% of rides.

This is the northern part of the Washington Heights neighborhood. Here we have 5 subway stations, several lodging options, a hospital, nursing home, and the Cloisters Art Museum.

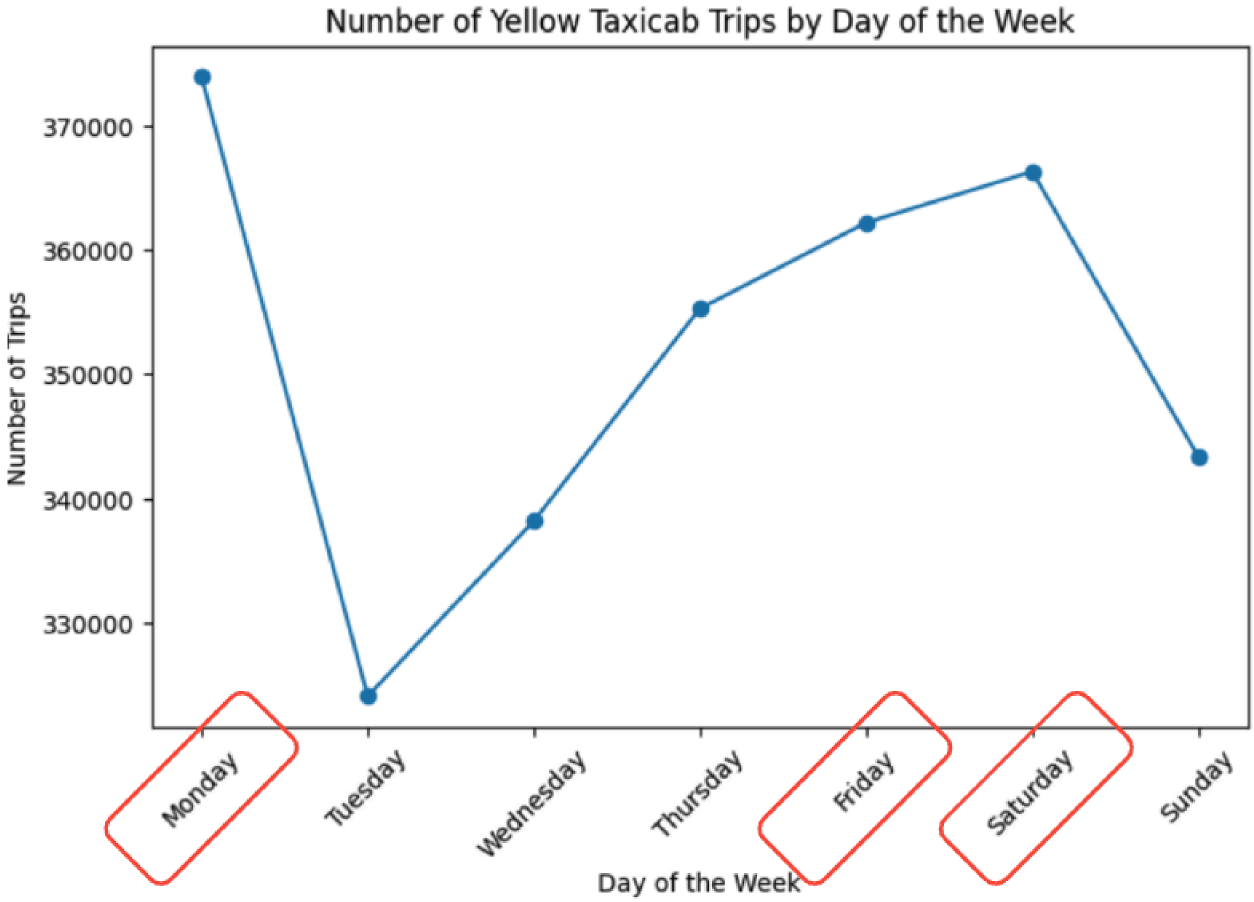
Based on our analysis of where taxi rides are coming and going, we can see that Manhattan and Queens consistently saw a majority of rides and that within these boroughs, locations that span high-density neighborhoods including transit options, local attractions and hotels are likely to be busy. Additionally we saw that JFK Airport accounted for a higher percentage of rides due to the lack of other transit options and high volume of potential taxi customers. What this tells us is that taxi drivers can expect to be traveling to and from Manhattan and Queens and but what exactly is happening in those taxi rides?

Temporal Trends

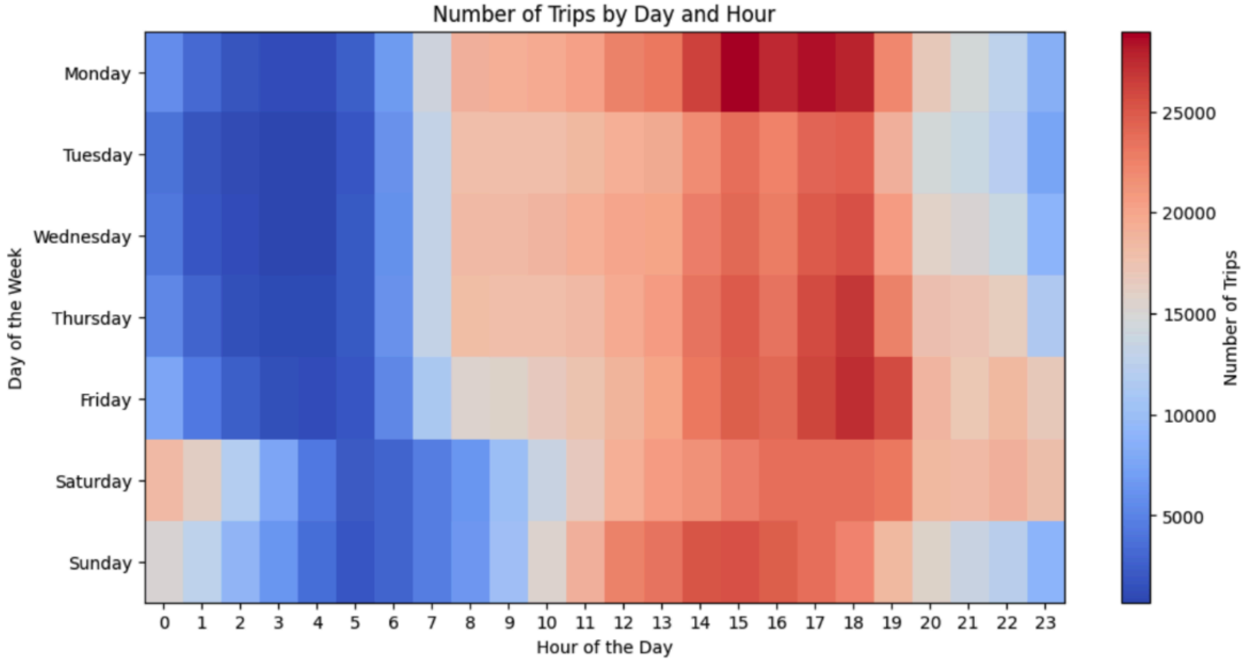
One of the first things we analyzed was taxi demand across different times of the day and week. Identifying peak hours and days provides insight into the city's transportation patterns. Taxi rides follow a predictable pattern throughout the day, peaking in the afternoon between 2 PM - 6 PM. This timeframe aligns with common commuting and errand-running hours.



Breaking it down further, we observed differences in demand across the week. Mondays, Fridays, and Saturdays had the highest ride counts.



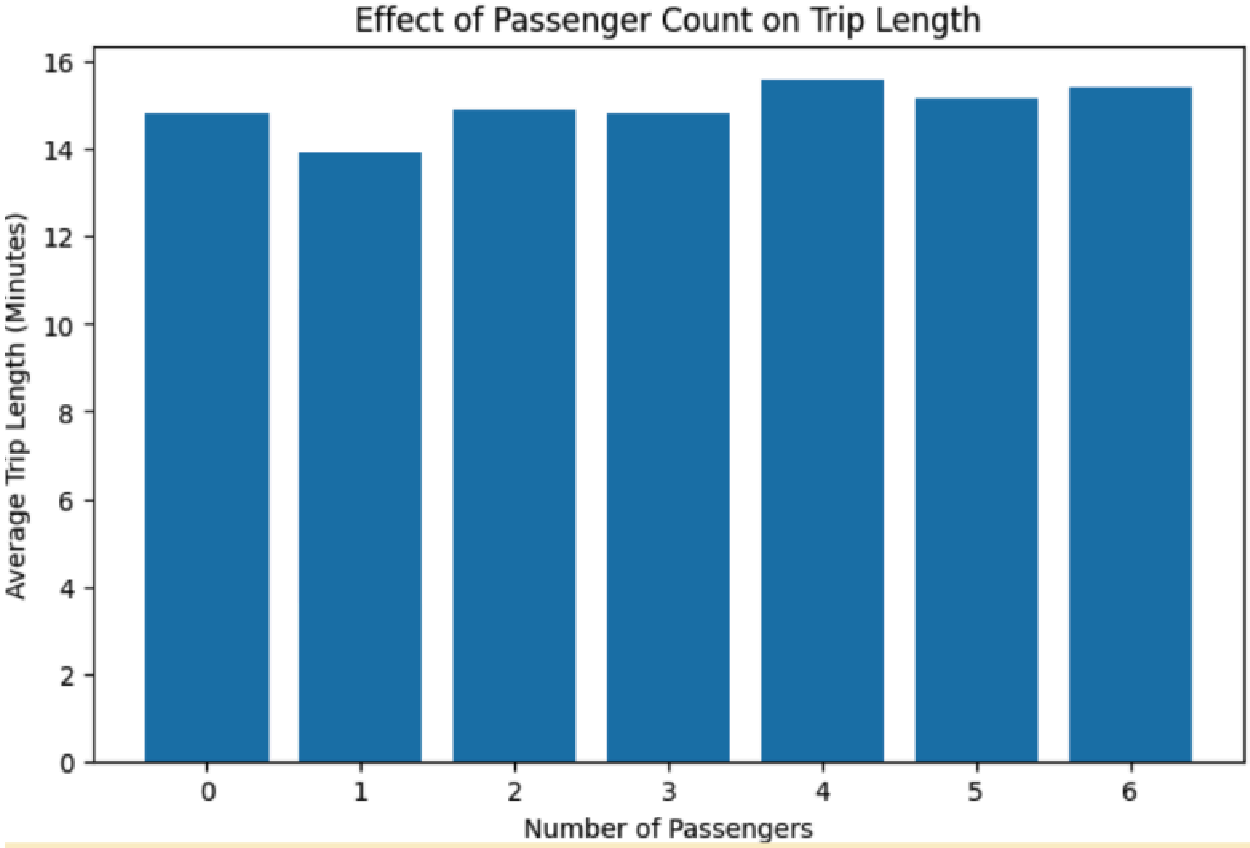
Putting this information together, we can get a higher level look at what traffic hours look like on an average week in NYC.



The heat map above shows us no surprises. Monday, Thursday, and Friday are busiest days, with the busier hours extending further into the night time on Fridays and Saturdays.

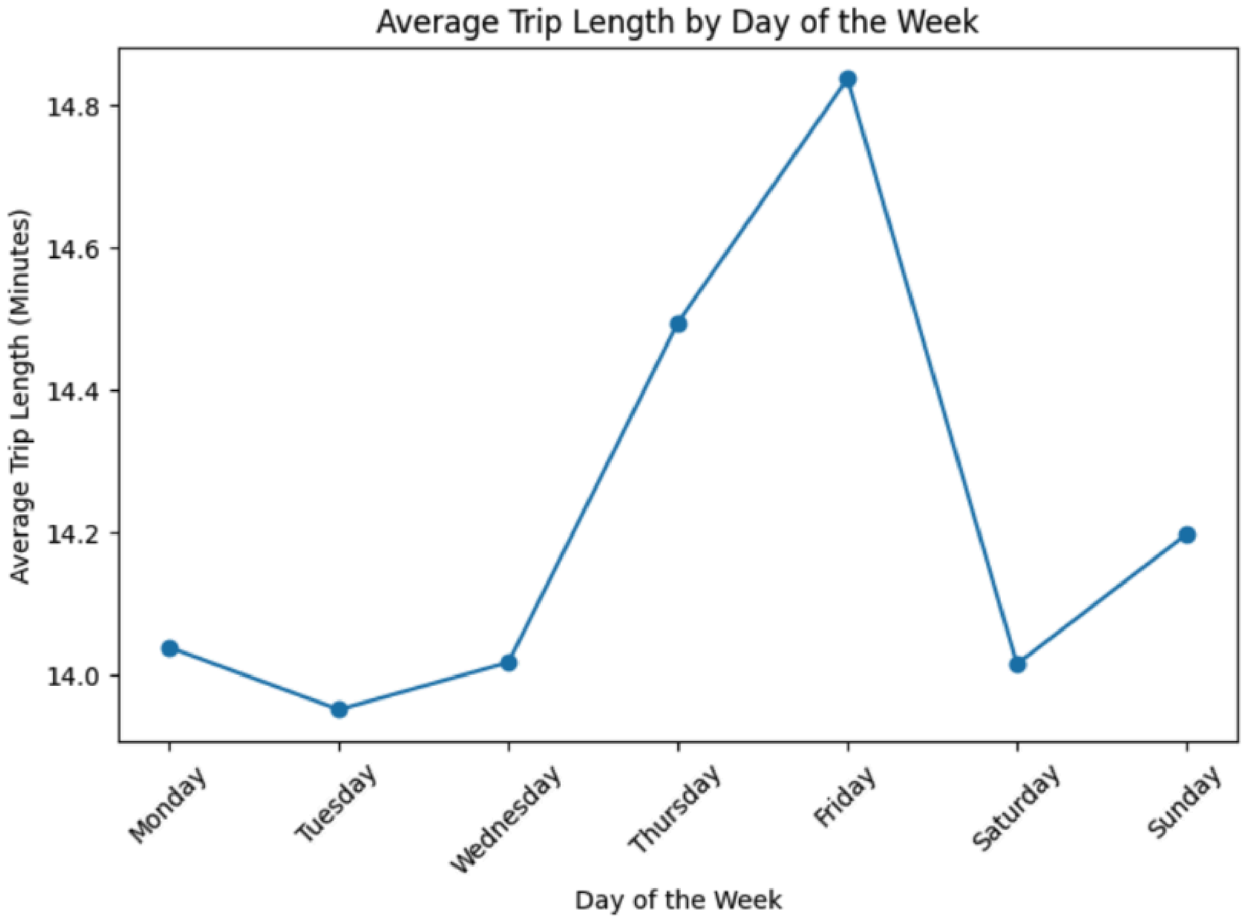
Trip Characteristics

After analyzing the temporal trends, we wanted to get a closer look as to what the taxi trips look like. The first thing we did is see how passenger count influences characteristics. The graph below shows that the passenger count effect has no real effect on trip length.

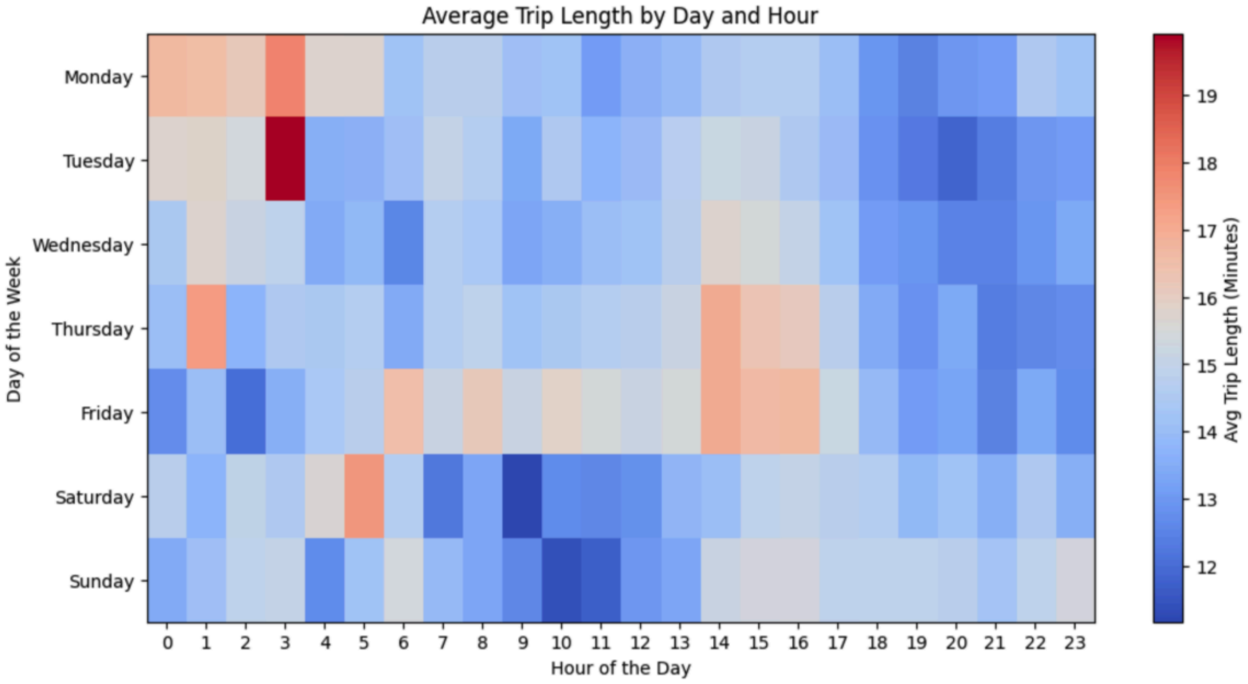


That said, it's worth noting that this investigation lead to some meaningful insights. For example, we realized that there are a few (less than 10) trips with more than 6 passengers. This likely indicates the existence of limousines or some special service offered by one of the taxi vendors (described as vendor 2 in the data).

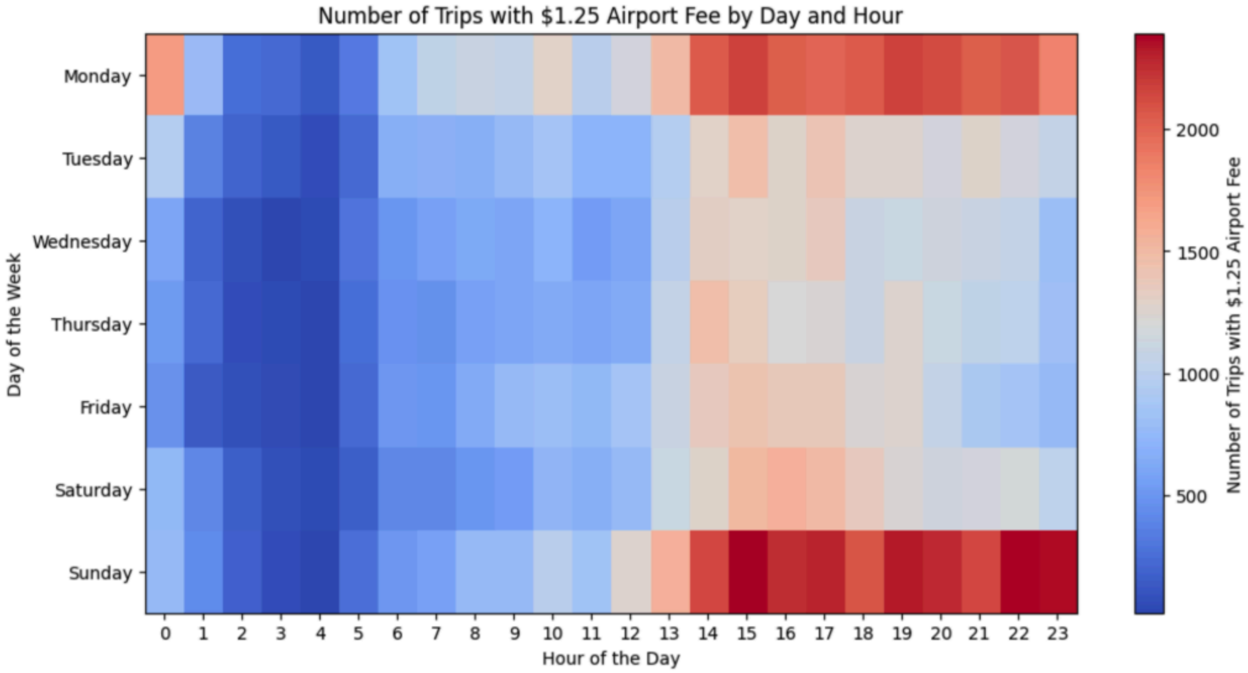
Continuing off of this, we looked to see how the trip length varies throughout the day. As seen in the graph below.



Similar to the heat map shown in the temporal trends section, we can make a graph that combines these two pieces of information.



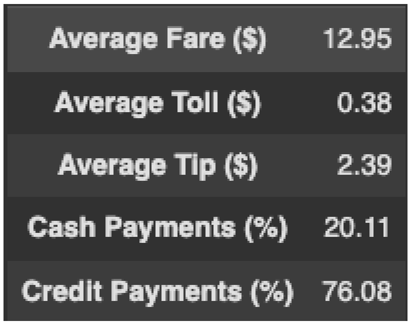
Unlike the previous heat map, this map shows an anomaly. The corner shows how something is causing trips to become lengthier on Mondays and Tuesdays around midnight through 3 am. We attempted to find what is causing these higher trip times, as traffic isn't expected on these hours. None of the characteristics seemed to deviate from the average (fare prices, distances, passenger count, etc.), so we decided to investigate if perhaps airports were the reason for that spike.



However, this graph above shows this isn't the case. Instead, Monday and Sunday evenings contain the most airport trips, leaving the Monday/Tuesday situation a fun mystery.

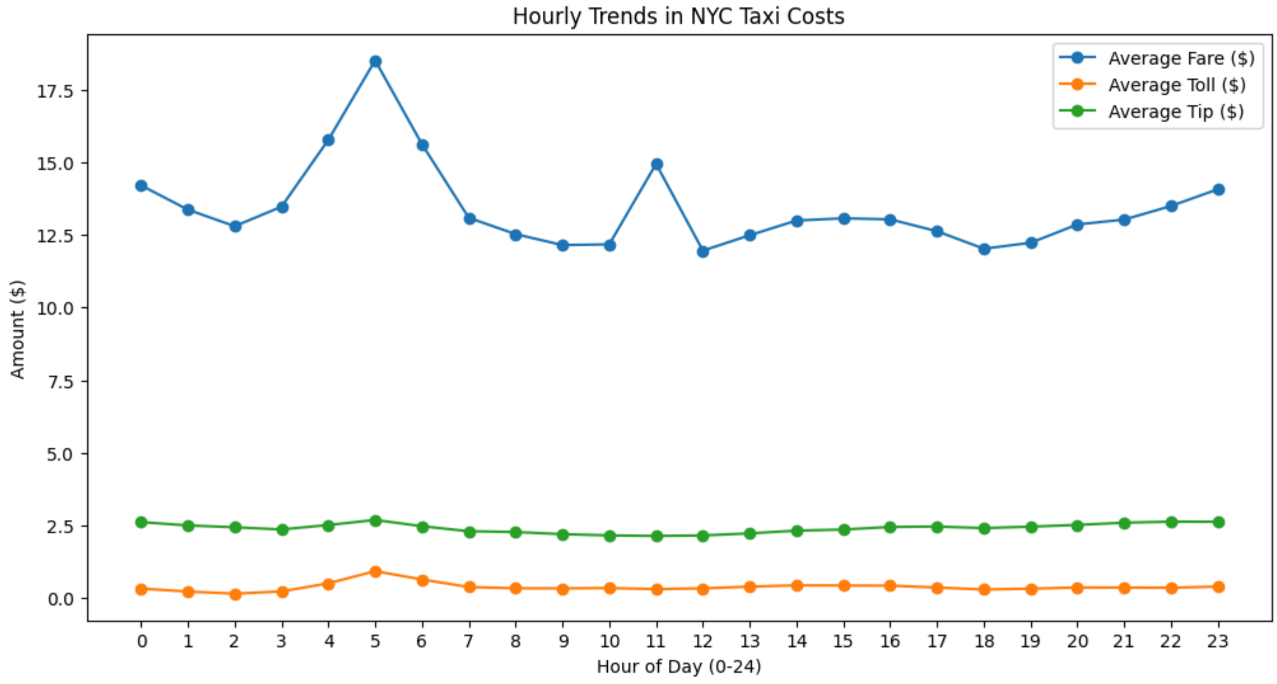
Payment Analysis

With this, what we started with is looking at the variables that seemed important regarding payment analysis. In this case, Average Fare, Average Tip, Average Toll, and the percentage of payment methods.

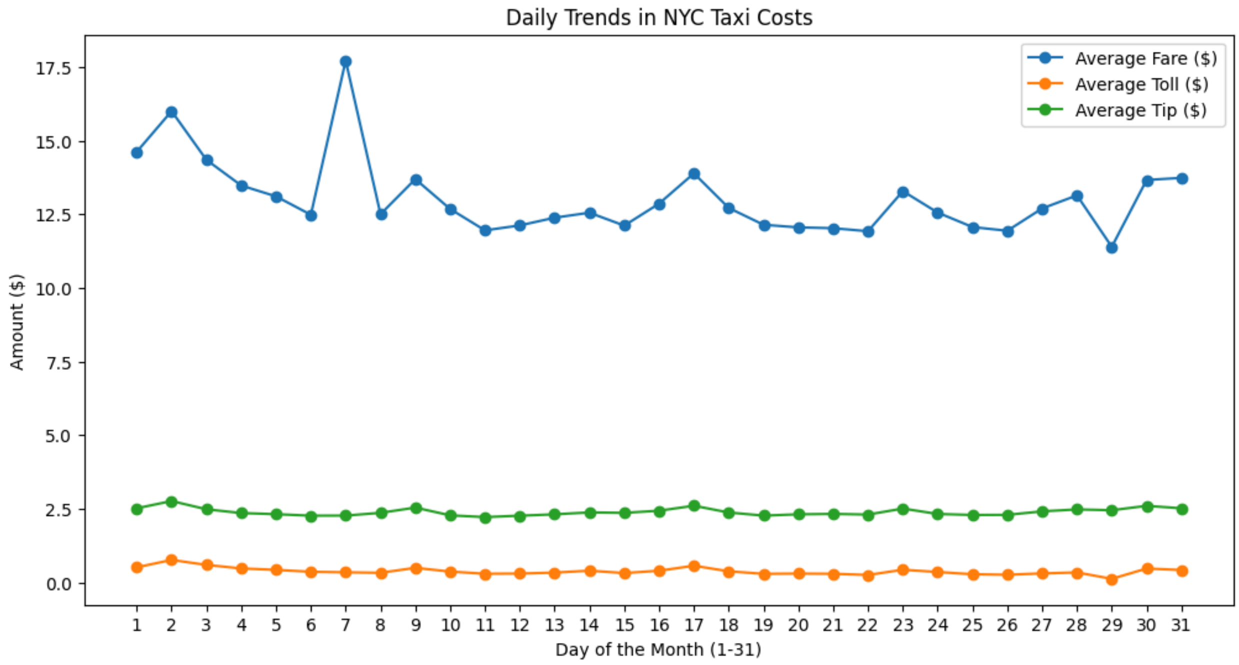


Average Fare ($)	12.95
Average Toll ($)	0.38
Average Tip ($)	2.39
Cash Payments (%)	20.11
Credit Payments (%)	76.08

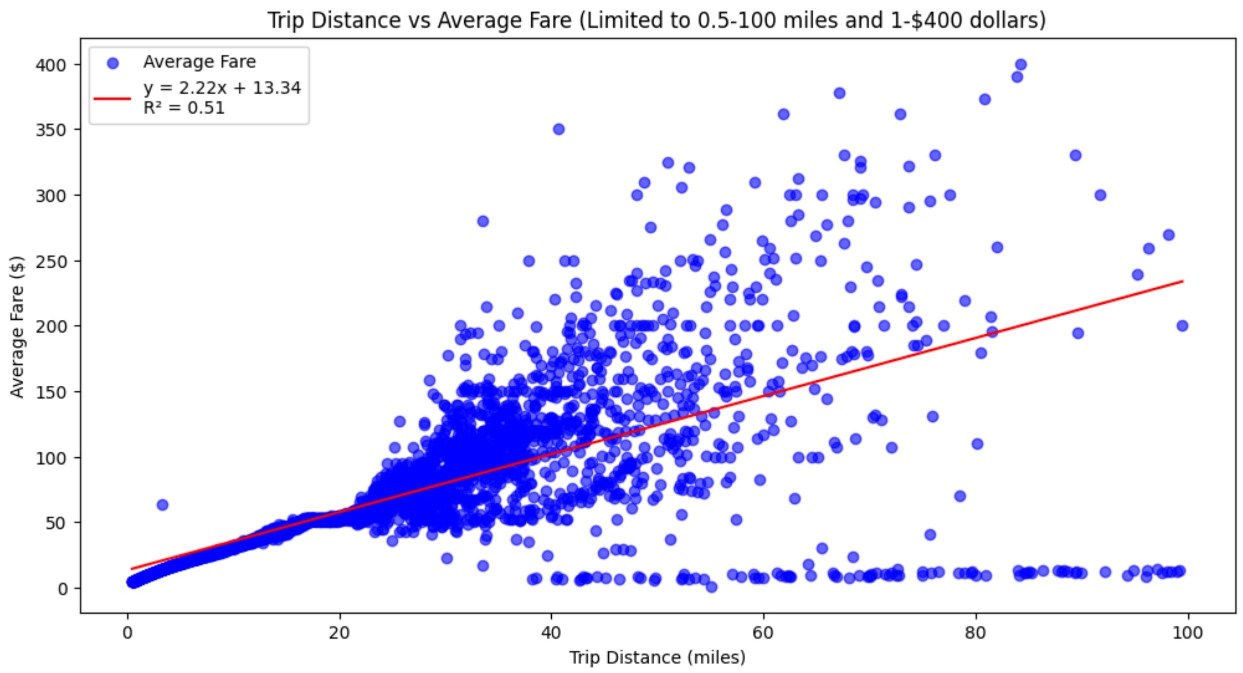
Then we first took these averages that had direct correlation with money and graphed them on an hourly basis. Essentially, what this means is that each hour in the graph represents the average data for that in the whole dataset in that time period.



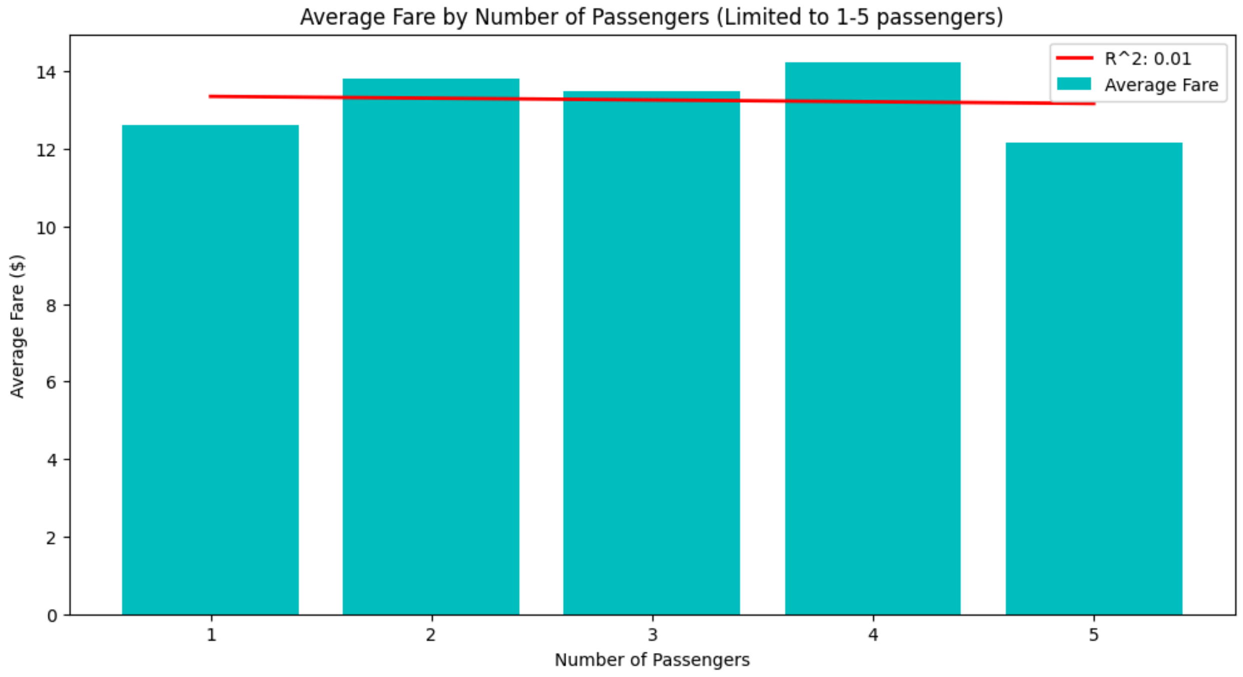
Here, we can see that the toll and tip averages by the hour do not vary that much, therefore we went ahead and focused on the average fare which has the most variation seen. Another thing to note is that there is a large spike in the average fare which can be explained with the next graph where we took the same averages and graphed them by the day of the month. Again, what this means is that each day of the month in the graph represents the average data for that in the whole dataset in that time period.



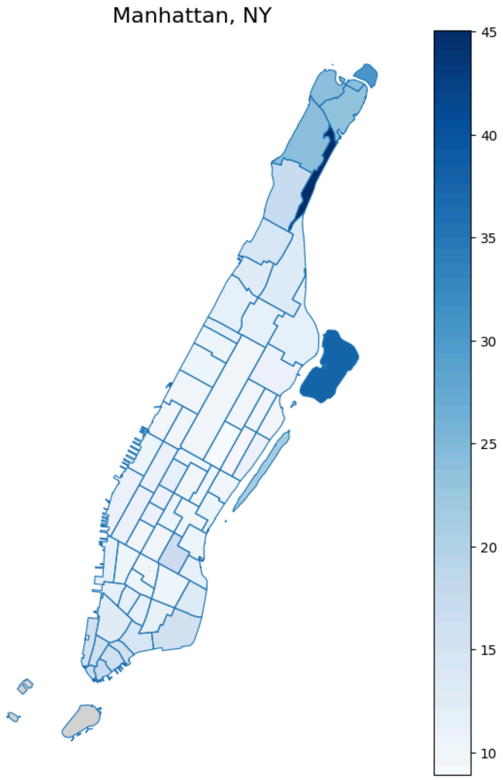
In this graph, we can see that the average toll and tip does not vary that much. However, unsurprisingly the average fare varies a decent amount. Also, as it was brought up earlier, the spike in this graph likely represents an event of some sort. For example a concert would see a high average ridership on the day of and due to that, the average fare would also follow that trend, which is what we are likely seeing here in that spike. However, that does not explain the rest of the variation. To get a better idea of what is causing that, take a look at some things that may influence the data.



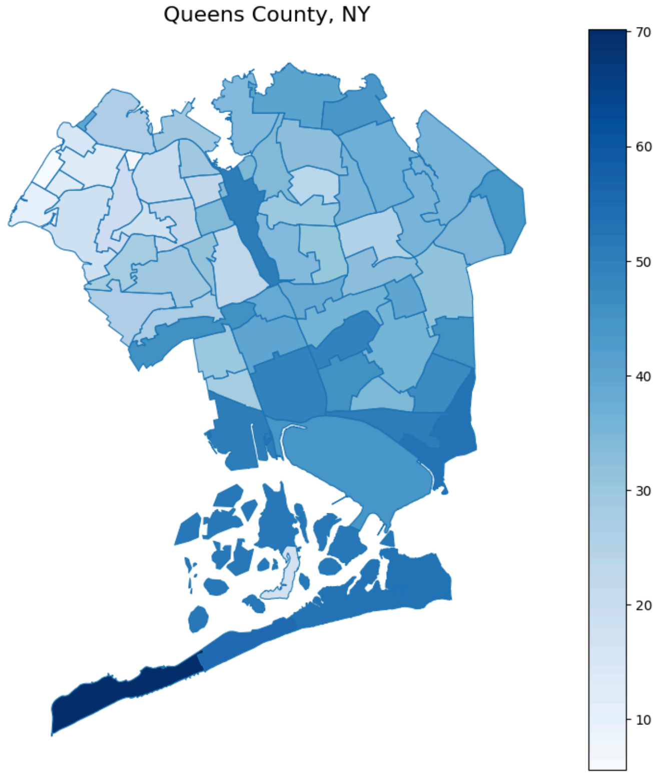
In this graph, we narrowed down the distance to 100 miles to accurately show where most of the data falls. In this case, the New York metro area is about 100 miles north to south and roughly 120 miles west to east. By looking at the points focused on in this area, we can see a trend that suggests **51%** of the variation of the average can be explained by distance traveled. Now let's take a look at another influence that could affect the average fare.



When it comes to passenger count, surprisingly, there is no correlation present when it comes to being an influence on the average fare. So let's keep looking.

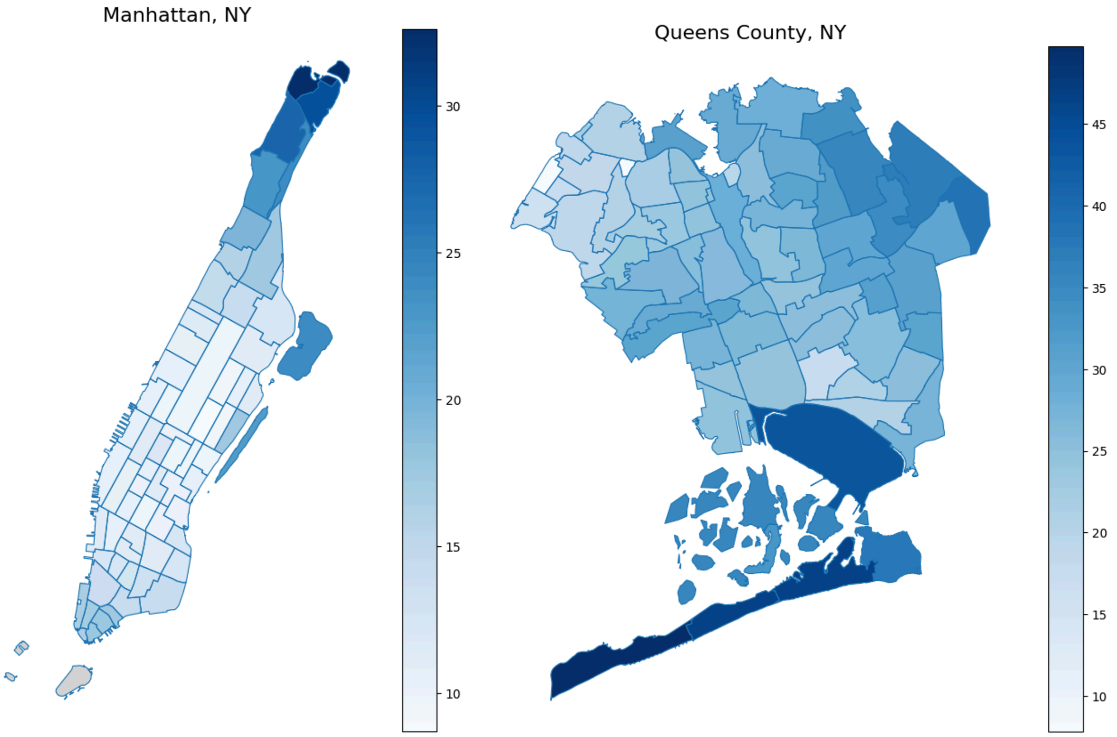


Manhattan, NY



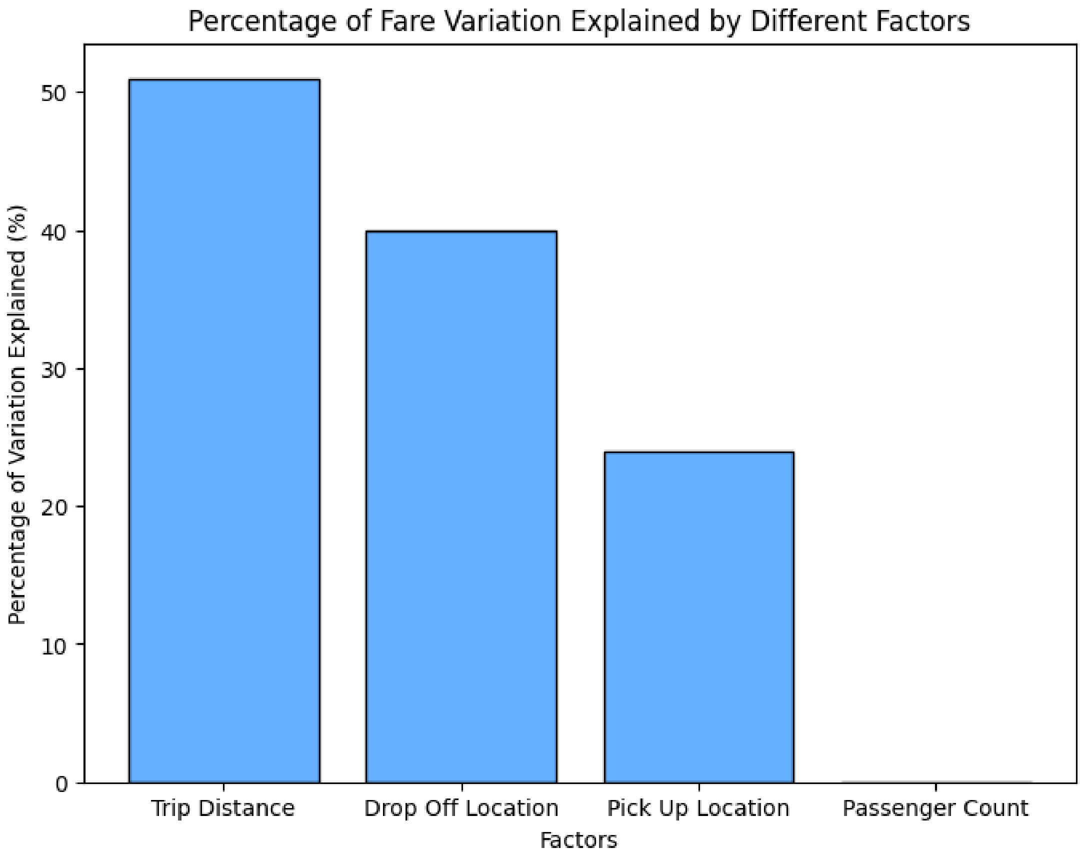
Queens County, NY

When it comes to pick-up a location, there are about 4 places on the map that have a higher average fair. The thing these places all have in common is that they are parks which can explain why the average fair is higher in these spots.


Manhattan, NY

Queens County, NY

When it comes to pick-up locations, unsurprisingly the areas with the highest average fares go up a little more compared to pick-up based locations. It is also important to note that JFK international airport location in Queens County sees higher average fares for drop-off compared to pick-ups. Now, how does this look in the big picture? Looks take a look from a different angle.



Here, we took the r^2 values of each influence and put it in a graph. Obviously it does not add up to 100% due to the nature of this statistical test. However, we can see that 51% of the variation of the average fare can be explained by distance traveled, 39% can be explained by drop location, and 24% can be explained by the pickup location. These are how the influences affect the average fare.

Conclusion

Throughout this article, we have taken you through the varying patterns regarding spatial distribution, temporal trends, and trip characteristics of NYC Yellow Taxi Cab rides. Our work has demonstrated that rides cluster in Manhattan and Queens, peak travel hours are between 2pm and 4pm, typically on Mondays, Fridays, and Saturdays and that key trip characteristics such as number of passengers minimally affects average fare.

With these findings, we’ve shown that the factors that are affecting average daily fare (our only marker for financial standing within the dataset) are primarily trip distance, followed by destination and origin of rides. All together, our analysis has illustrated where taxi drivers and NYC TLC support staff should expect traffic to cluster, when those higher concentrations of taxi rides should occur and how each of these factors will affect the overall financial performance of NYC taxi rides.