A VISUAL ANALYSIS OF COMPETITION BETWEEN GREEN AND YELLOW NEW YORK TAXIS

By Matthew Tregear

Motivation, data and research questions

In this report I use Visual Analytics to inform competition analysis between green and yellow taxis in New York.

One major consideration when undertaking a competition analysis for a merger between two firms are to what extent firms compete closely depending on the geographical area. More specifically, in the hypothetical context of green and yellow taxis being run by separate firms in New York City, an initial analysis of overall market shares indicates that green taxi journeys comprise only around 6% in the city – suggesting that green taxis only apply a minimal constraint on yellow taxis. Yet this does not tell us whether certain routes or pickup locations are dominated by green or yellow taxis – or more generally how competition between the two taxi types varies spatially.

To look at this, in this report, I use Visual Analytics to investigate:

- How the shares of green and yellow taxi journeys vary across major pickup locations?
- How the shares of green and yellow taxis vary across major origin destination routes?

I do this using an origin destination dataset for a week in July 2015 for the journeys of yellow and green taxis obtained from the New York Taxi and Limousine Commission. This contains the pickup and dropoff location for each taxi journey.

I believe this data is suitable for my analysis as it allows me to look at the overlap between green and yellow taxis by identifying similar pickup locations and routes they take. I also believe this dataset is particularly well suited for a hypothetical case study of competition analysis because, unlike yellow taxis, green New York City taxis are not allowed to pick up passengers below East 96th and West 110th Street in Manhattan and the two airports in New York City, and were introduced to provide more taxi coverage to New York City's less heavily trafficked areas. As a result, we have reason to suspect that green taxi behavior will vary from yellow taxi behaviour because green taxis may be less keen to collect fares for routes where they are less likely to have a return pickup fare – so, in this regard, the two taxi types may arguably have different behaviour.

Finally, to narrow the scope of my analysis, I restrict my dataset to taxi journeys with more than 15km in distance.

Tasks and approach

Whether there are greater or smaller proportions of green and yellow taxi journeys at major pickup locations**?**

I will need to undertake three analytical tasks to answer this.

1) I will need to visualize the distribution of all pickup locations in a symbol plot. This will provide me with an understanding of the main areas where green and yellow taxi pickup locations overlap.

2) I will need to categorize individual taxi journey pickup locations into groups based on their special similarity using density based clustering.

   Initially, I plan to density based cluster (DBSCAN cluster) pickup locations into similar groups based on their spatial coordinates, depending on how densely populated an area is by taxi pickup locations (ie. how closely pickup locations are located to others?). This should provide an overview of the main taxi pickup location areas, but it may not account for the fact that taxis are likely to search for more targeted pickup areas in more densely populated areas, where different clustering would be better in less dense and more dense areas.

   To better account for differing population densities I will therefore undertake further density clustering in two ways:

   First, I will apply additional density clustering only on larger densely populated areas that are not well separated, using a smaller distance parameter in my clustering algorithm. This will allow me to cluster into smaller groups in these more densely populated areas while, at the same time, keeping larger groups in less densely populated areas.

   Second, I will undertake density based clustering on the centres of the density clustered groups to better cluster pickup locations in less densely populated areas. This will not only allow me to cluster pickup locations in the outer regions of New York better, but it will correct cases where locations have been clustered under a lower distance parameter than they should be because they are too closely located to other more densely populated pickup location areas.

   I will evaluate the success of my clustering by visualizing the ‘residuals’ – the distance between individual pickup locations and their corresponding cluster centres.

3) I will need to visualize the proportions of green and yellow taxi locations in each major pickup location cluster. I will do this by focusing on only those pickup location clusters where more than 100 taxi journeys have occurred over the week. I will then apply a diverging colour scheme that maps the proportion of green taxi journeys to each of these pickup locations on a symbol plot. This will show the proportions of green taxi journeys in each pickup location cluster.

Whether there are greater or smaller proportions of green and yellow taxis on major origin destination routes?

I will need to undertake two additional tasks to answer this:

1) I will need to categorize individual taxi journey dropoff locations into groups based on their special similarity using density based clustering. To do this, I will use the same methodology as I plan to use to cluster pickup taxi locations above.

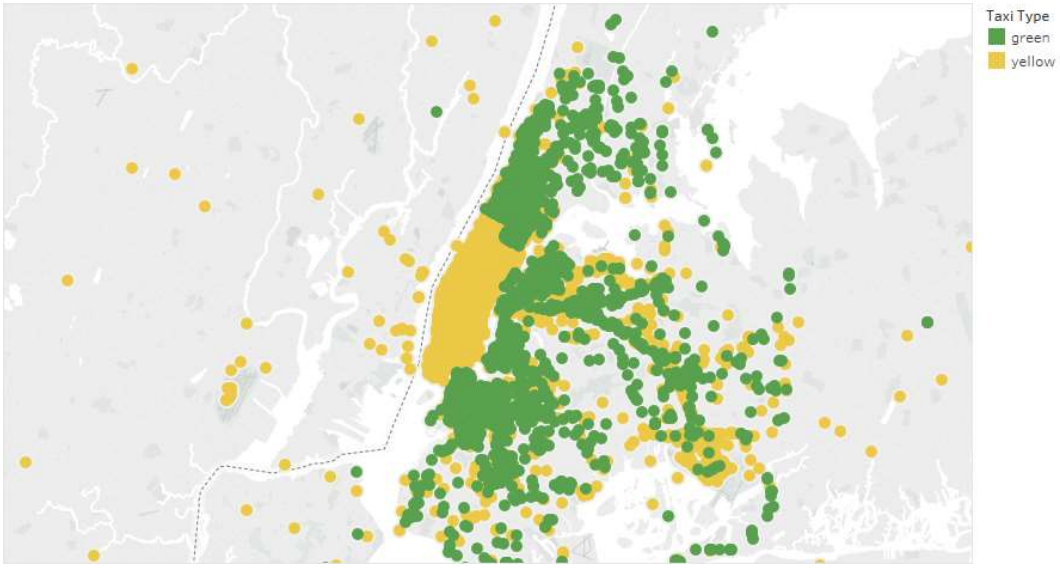
2) I will need to visualize the proportions of green and yellow taxi journeys on major routes (those with more than 50 journeys). I will do this by visualizing the origin destination pairs for the centres of each pickup cluster/ dropoff cluster combination in a spider map. This will show the proportions of green taxi journeys for each pickup cluster/dropoff cluster combination using a diverging colour scheme– and thus the proportions for each major route.

Analytical steps

ANALYTICAL TASK 1: Looking at the distribution of green and yellow taxi pickup locations

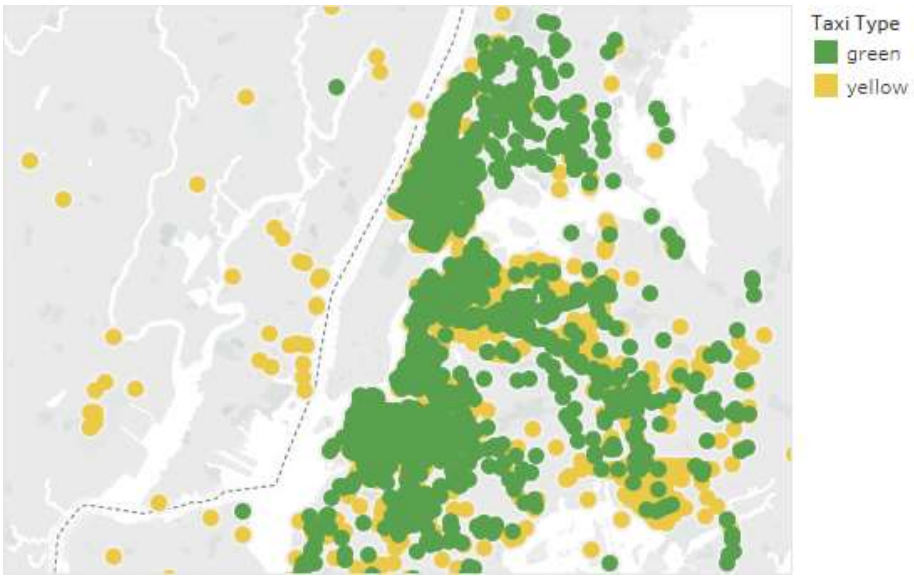
I first look at how yellow and green taxi pickup location are spatially distributed within a symbol plot focused on New York. Green taxi pickup locations exist in many of the same locations but, as expected, they do not pickup fares in South Manhattan.

Figure 1: Distribution of green and yellow taxi pickup locations



I remove the South Manhattan pickup locations as we know that green taxis are prevented from picking up fares here. I decide not to do the same for airport pickups as they are less easy to isolate.

Figure 2: Distribution of green and yellow taxi pickup locations (excluding South Manhattan)

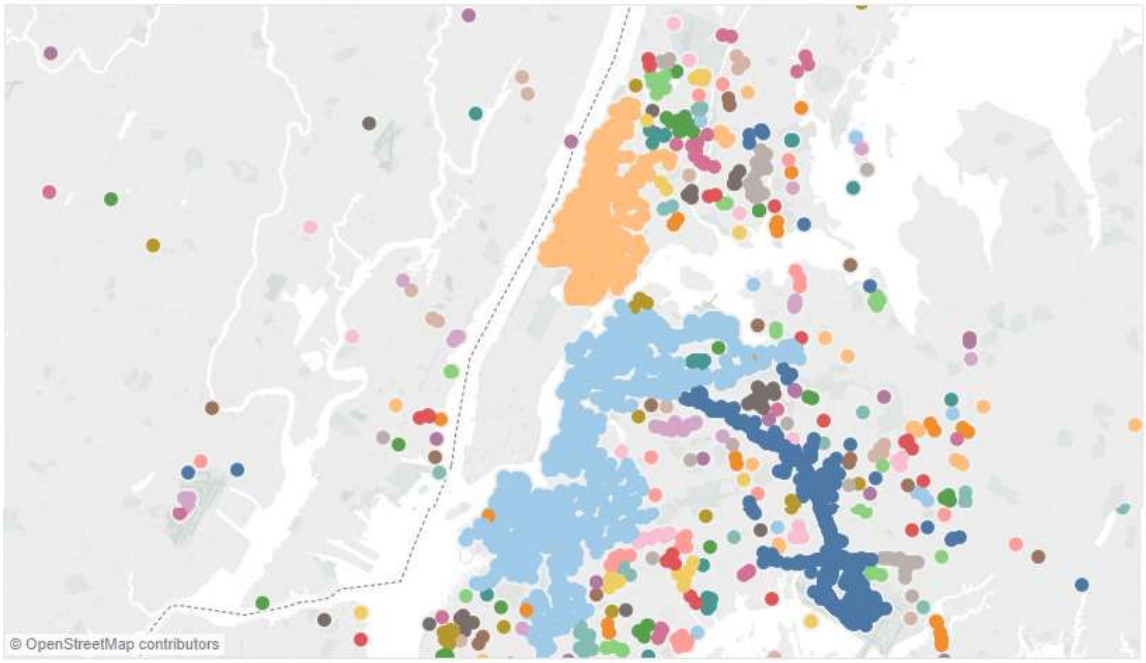


This reduces the dataset but, purely based on the symbol plot, it is still unclear how green and yellow taxi pickup locations should be grouped.

ANALYTICAL TASK 2: Identify groups of similar pickup locations used by green and yellow taxis

I first apply an initial DBSCAN clustering with a distance parameter of 500m[1]. The algorithm clusters locations in the outer areas of New York rather well but performs poorly in more densely populated central areas. There are three large groups (a yellow, a dark blue and a light blue one) where taxi pickup locations are badly separated.

Figure 3: Initial clustering of green and yellow taxi pickup locations across New York (excluding the South Manhattan Area) using a distance parameter of 500m



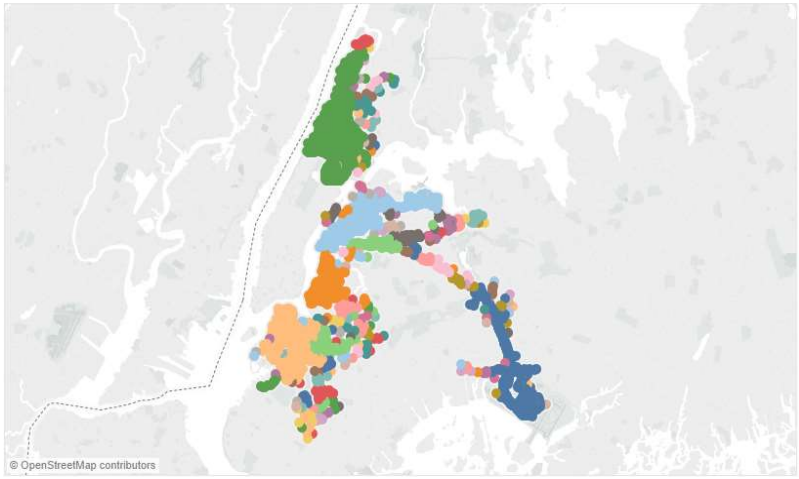
Reducing the distance parameter for my clustering would allow me to identify smaller clustered groups in the visualization. But this comes at the expense of losing the well-separated groupings under a greater distance parameter.

So, rather than simply changing the cluster parameters, I first partition my dataset into two sections: one including poorly separated large clusters and another with the remaining clusters of pickup locations where density clustering has worked well.

I then perform additional clustering on the partition with poorly separated clusters using a lower distance radius parameter of 300m. This splits the poorly separated clusters into smaller groups but leaves the well separated areas unchanged.

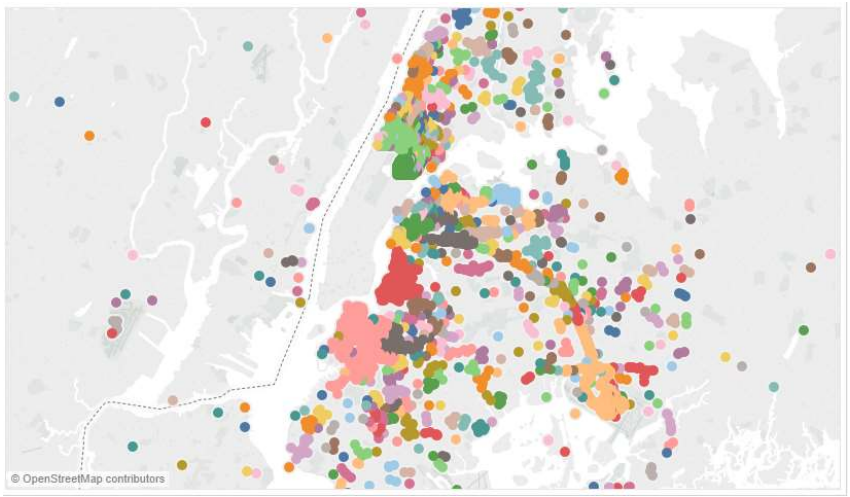
[1] And k=1 nearest neighbours.

Figure 4: Additional clustering of green and yellow taxi pickup locations using a distance parameter of 300m



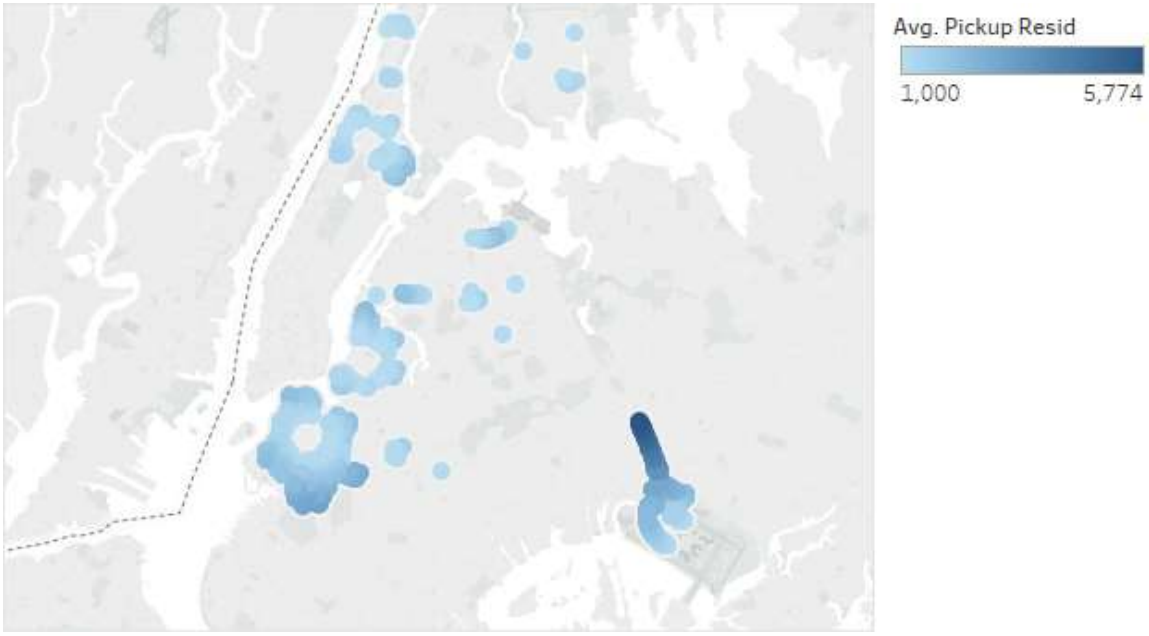
I find that three large problematic areas remain (the dark green, light blue and dark blue groupings). I repeat the process of filtering and selectively clustering badly separated large groups a further 2 times (using distance parameters 200m and 130m as shown in the appendices) until all clusters are well separated. I combine the results of all the clusters into a single visualization of pickup locations.

Figure 5: Combined clusters of green and yellow taxi pickup locations with varying distance parameters



The results of my clustering seem mixed. No badly separated large groups remain. Even amongst locations with the greatest residuals, all except for ones in the JFK area are within 3km of their cluster centres.

Figure 6: Residuals for the distance between yellow and green taxi pickup locations greater than 1km

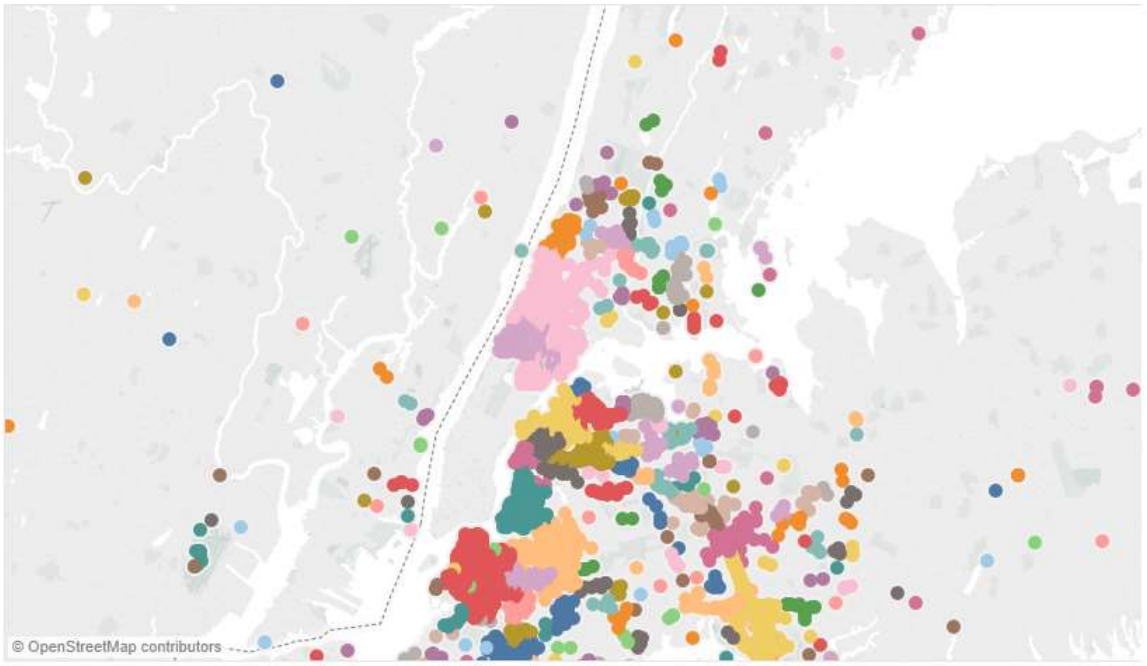


That said, there are many '0' residuals located close to each other that should probably be clustered together.

Figure 7: 0km residuals yellow and green taxi pickup locations

To address these pickup locations that have been overly separated, I cluster on the central points for each cluster (using a greater distance parameter of 750m). This results in the final visualization of each pickup cluster.

Figure 8: Final pickup clusters for yellow and green taxi pickup locations after clustering the centre of each pickup cluster



The pickup locations with 0 residuals are now further away, so the clustering now performs better in less dense areas.

Figure 10: Final 0km residuals for yellow and green taxi pickup locations

ANALYTICAL TASK 3: Visualizing the market shares of green and yellow taxi in each major pickup location group.

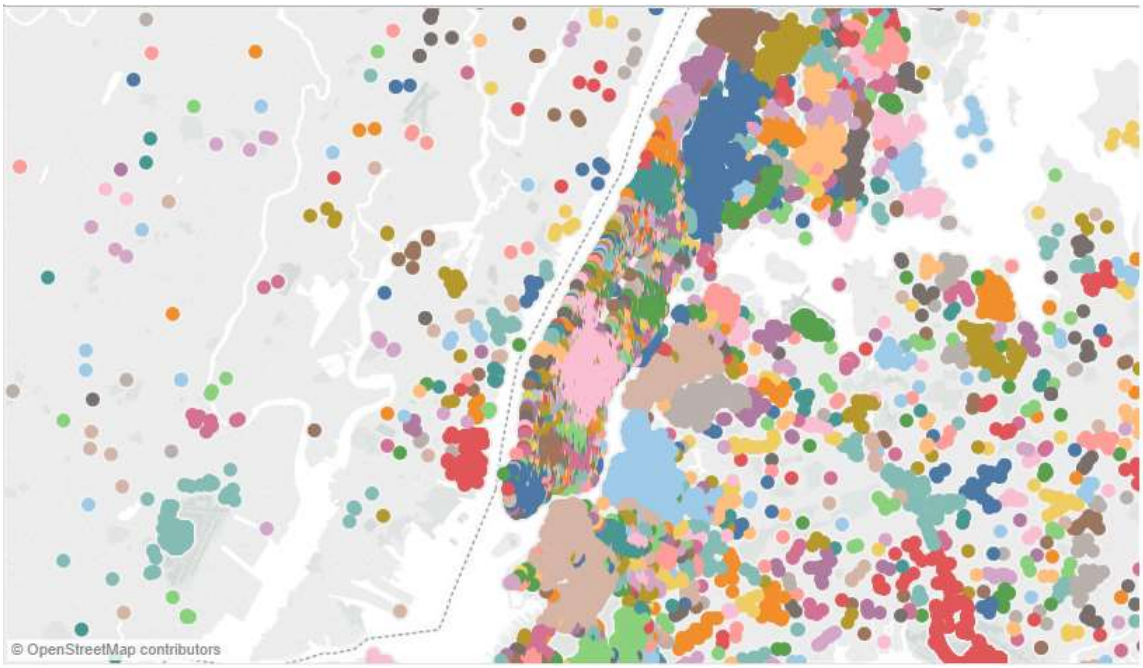
To identify major pickup locations, I visualize the centres of the pickup location clusters with more than 100 journeys. I then apply a green and yellow diverging colour scheme to show the average proportion of green taxi rides at each major pickup location.

Figure 11: Proportions of green and yellow taxis at major pickup locations

ANALYTICAL TASK 4: Identifying groups of similar dropoff locations used by green and yellow taxis

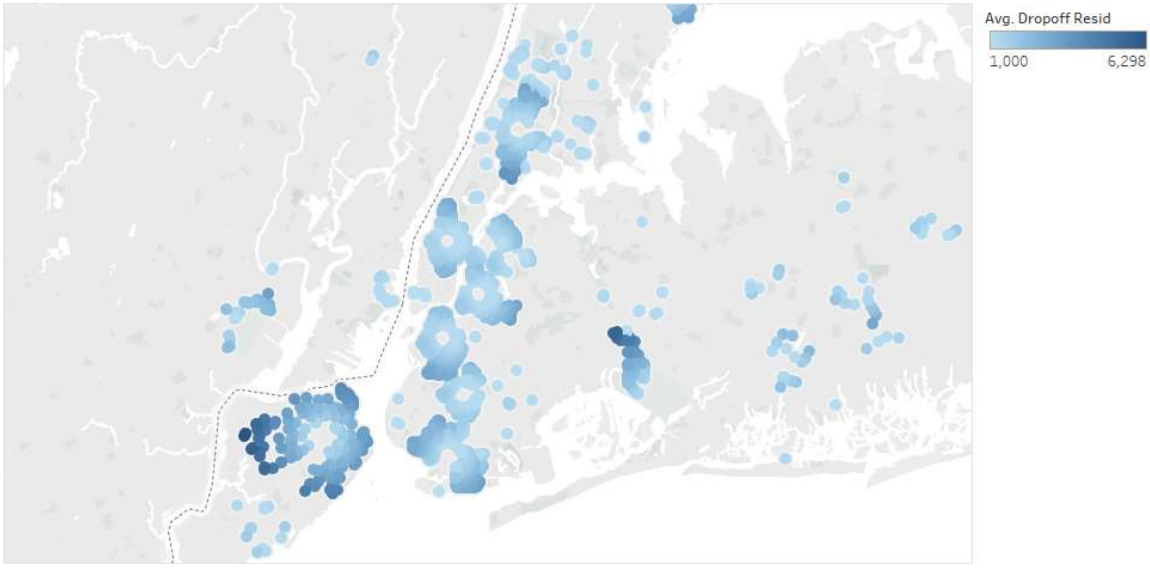
To identify groups of similar dropoff locations, I follow the same methodology as the clustering of pickup locations. An initial DBSCAN clustering of dropoff locations (using a 1km distance parameter with 1 nearest neighbor) separates dropoff locations in the outer regions well but leaves a very large group in Manhattan. Then repeating the same approach as I use to group pickup locations, I undertake clustering a further 6 times using distance parameters of 400m 270m, 200m, 150m, 100m and 50m respectively for successively filtered clusters (as shown in the appendices). This results in the following cluster visualization of all dropoff clusters.

Figure 12: Combined clusters of green and yellow taxi dropoff locations with varying distance parameters



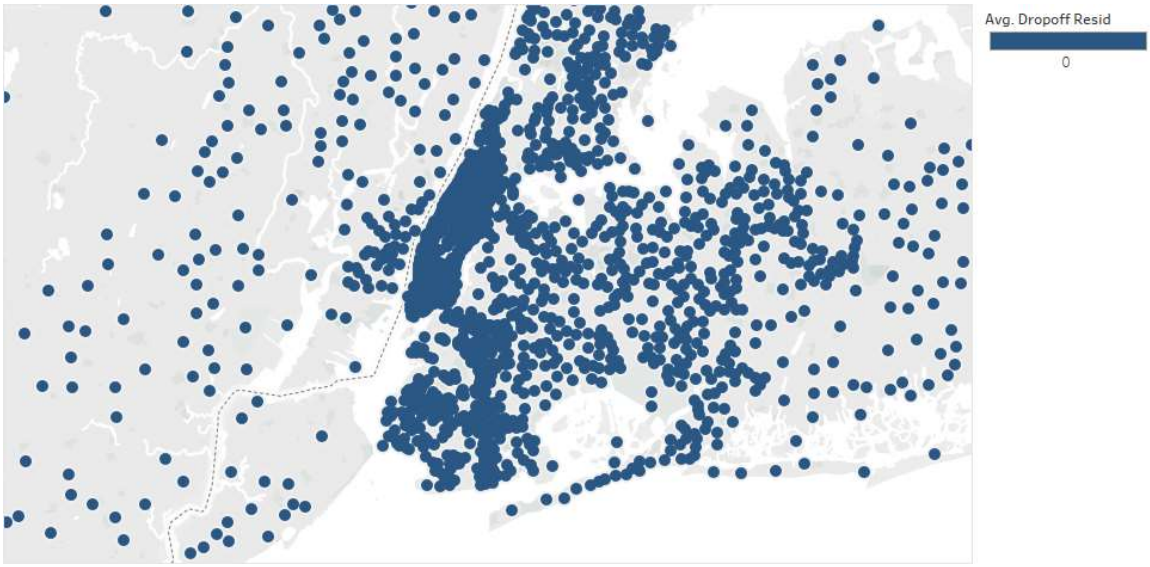
The residuals suggest that each point is grouped to within around 6km of a cluster centre. This is reasonable for our dropoff clustering performance given that we are only interested in whether customers are taken to vaguely the same locations for dropoff centres to identify common routes (and not the exact dropoff location itself) – so we do not need such fine groupings as for pickup clustering.

Figure 13: Residuals for yellow and green taxi dropoff locations greater than 1km



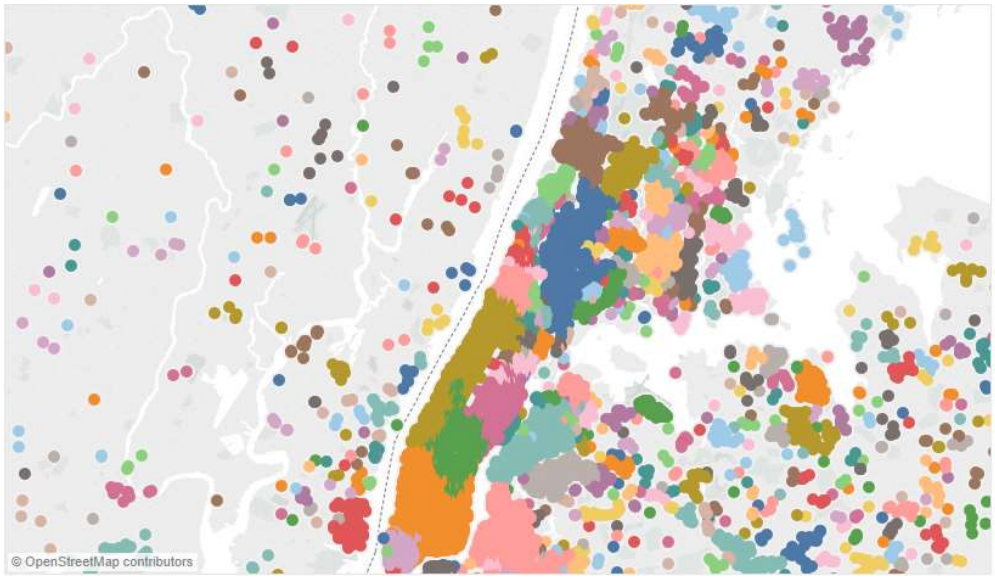
Still, there are also a larger number of dropoff locations with 0 residuals which are located close to each other that should probably be clustered together.

Figure 14: 0km residuals for yellow and green taxi dropoff locations



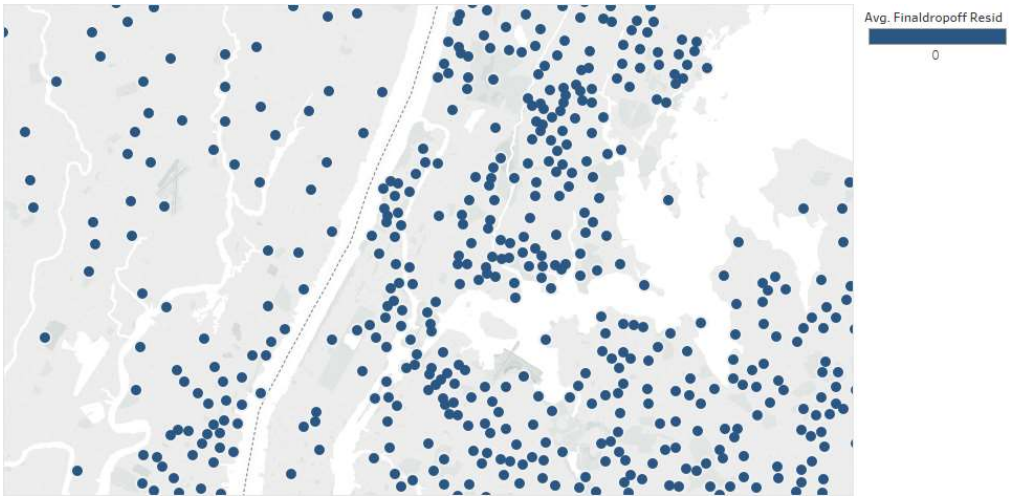
As a final step, I cluster the central points for each cluster using a distance parameter of 300m to provide better clustering performance in less dense areas. This provides the final visualizations below for each dropoff location cluster.

Figure 15: Final pickup clusters for yellow and green taxi dropoff locations after clustering the centres of each dropoff cluster



This final plot now has a better performance in less dense areas as the 0 dropoff location residuals are now well separated.

Figure 16: Final 0km residuals for yellow and green taxi dropoff locations



ANALYTICAL TASK 5: Looking at the market shares for common routes taken by green and yellow taxis.

To visualize the proportions of green taxis on major routes, I create a spider map displaying the origin destination for each pickup and dropoff cluster centre combination.

Figure 17: Proportions of green and yellow taxis on major origin destination routes

Findings

My visual analytics analysis of pickup locations has two main findings.

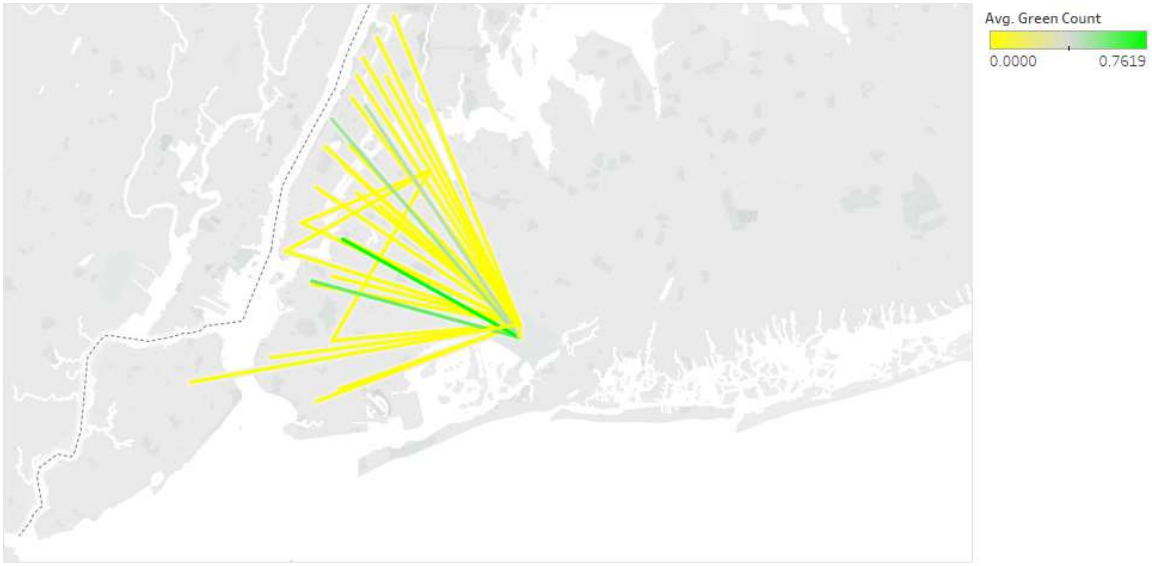
Looking at the symbol plot for the taxi journeys at the major pickup location groups (with more than 100 taxi fares) I find that, as expected, yellow taxis dominate the two airport pickup locations with more than a share of 99.5% of fares due to the restrictions on green taxis. La Guardia airport is the yellow spot in the North East of the symbol plot and JFK airport is the yellow spot in the South East. However, more interestingly, when these airport pickup location clusters are removed from the dataset, I find that the share of green taxis is increased dramatically from around 6% to 55%. So, excluding the restricted pickup areas for green taxis, the market share between yellow and green taxis is roughly even.

I also find that green taxis dominate pickups at Coney Island Beach, the bright green spot in the South of the map, and the Bedford Avenue metro in Williamsburg, the bright green spot closest to the centre of the map (with a more than 75% share of taxi fares at each location). But, at the remaining 7 pickup location clusters, the share of green taxis is more evenly distributed between 65% and 50%. These pickup locations are in: Columbia University and Harlem in North Manhattan; Long Island City High School, Queensboro Plaza station and a street in Astoria; The New York Transit Museum in the Brooklyn heights; and Jamaica station.

Figure 18: Proportions of green and yellow taxis at major pickup locations

For my analysis of the major taxi routes (those with more than 50 journeys each), I find that all routes are journeys to and from airports (JFK or La Guardia). Unsurprisingly most of these routes are dominated by yellow taxis. But, even though they are prevented from pickup airport fares on return, there are still some airport routes heavily used by green taxis. There is one green taxi dominated route from Northern Williamsburg where green taxis make up 75% of all taxi fares on the route. There are also three routes with even proportions of green and yellow taxi journeys (where the share of green taxis lies between around 45% and 60%), where taxis compete more closely.

Figure 19: Proportions of green and yellow taxis on major origin destination routes



Critical reflection

Overall, my analysis shows that outside the restricted pickup areas of the airport and South Manhattan green and yellow taxis compete evenly except for a few select pickup locations (such as the Coney Island area pickup area). This suggests that outside the green taxi restricted areas, the two types of taxi are close substitutes for each other. Most notably, my analysis shows the pickup restriction on green taxis at airports does not deter them from dominating certain routes to drop off passengers at the airport, even though they will not be able to collect a return fare when leaving the airport. This indicates that, if green taxis were and yellow taxis were run by different firms, there would be cause for regulatory concern if these firms tried to merge -given that they compete so closely in areas outside the green taxi restricted pickup areas.

The results of my analysis, however, are contingent on the clustering correctly identifying the appropriate geographical level at which taxis search for customers. For example, green

and yellow taxis' fare search patterns may vary at a higher geographic level where, say for example, green taxis are less likely to search in the North rather than the South, East or West than yellow taxis. If this were the case, then my analysis may not capture these differences as it is more locally targeted at more specific clusters.

My analysis also has several other caveats:

a) First, a critical factor in analyzing origin destination data is the data generating process that creates the origin and the destination. In this study, it is reasonable to assume that a taxi's fare search pattern will largely be responsible for the pickup location, but it is less clear whether a driver will have much control over the choice of destination. For example, for the airport routes I have identified, it may be that taxis search for passengers seeking a specific journey. Or, alternatively, taxis could advertise online or over the phone for specific taxi transit journeys (such as trips to the airport). However, it may also be the case that at certain pickup locations there may just happen to be many airport taxi fares. In this the case, my analysis of routes may be an indicator that a large proportion of taxi journeys more than 15km are taxi rides – and not evidence of taxi behaviour.

b) Second, the data is taken from a single week during the summer of 2015 across New York. One obvious issue with this approach is that, if you were to look at other times of year for this analysis, one may find differences due to seasonal changes. For example, the Coney Island pickup location will likely be less popular in the Winter months as fewer people use the beach. But there is also a more general drawback. The fact that the data can only provide a snapshot of taxi behaviour may mean that it may not provide insights or for New York taxi competition in the future as taxi behaviour may change. Similarly, I may not be able to generalize the analysis of taxi behaviour to other cities – as the selective competition on routes and at pickup locations are likely to heavily depend on the specific geography of New York.

c) Third, it may be the case that both taxi types are used in even proportions in certain locations, but they do not compete closely because passengers use them at different times. If there are concerns that green and yellow taxis are used at different times of day – this may be a useful extension for the analysis. Or alternatively that analysis could be adjusted so that it uses spatio-temporal clustering (to identify taxi hotspots at certain times of day) rather than simply relying on spatial clustering.

In terms of the applicability of my analysis, my analysis is, of course, of minimal direct value since it is a hypothetical exercise: green taxis and yellow taxis are not necessarily licensed by separate operators and do not represent separate firms. Still, as a case study, it provides a useful framework for how similar analysis could be applied to mergers of transport firms in different contexts. For example, one of the key factors in the clearance of the Streetcar &

Zipcar merger (two car clubs) in 2010 by the Competition Commission[2] was to what extent the two firms competed more closely at certain pickup locations.

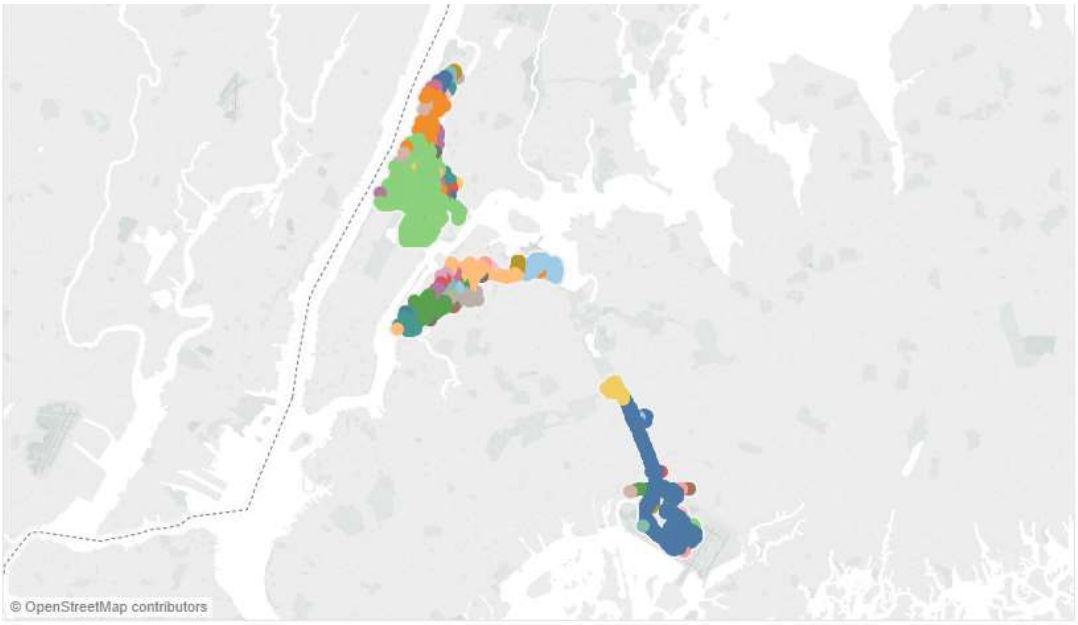
More specifically, both these two companies (and another two firms Hertz and City Car Club) provided a network of readily accessible vehicles, parked in local areas, for hourly or daily rental on a commercial basis across London. So, an important question in this merger was: to what extent Zipcar and Streetcar competed more closely than the other two car clubs? If the Commission had requested an origin destination dataset for each for the four companies' customers' car journeys, using their Statutory Powers[3], they could have undertaken a similar analysis as the one I outlined for green and yellow taxis. This would have provided the Commission with a better understanding of the extent to which the merging car clubs competed more closely than the other firms at a local level.

A similar analysis may also be more generally applied to the domain of logistics, where a transport company, such as a delivery firm, where after a merger or more generally may want to better optimize the vehicles flows on their delivery routes.

Appendices

Part I: Additional clustering of pickup locations

Figure 20: Additional clustering of green and yellow taxi pickup locations in large badly separated areas using a distance parameter of 200m



[2] https://assets.publishing.service.gov.uk/media/551946fc40f0b61404000076/final_report.pdf
[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/212286/CMA2con_-_Mergers_Consultation_Document_and_Guidance_FINAL__11JUL13_.pdf

Figure 21: Additional clustering of green and yellow taxi pickup locations in large badly separated areas using a distance parameter of 130m

Part II: Additional clustering of dropoff locations

Figure 22: Initial clustering of dropoff locations (using a 1km distance parameter)

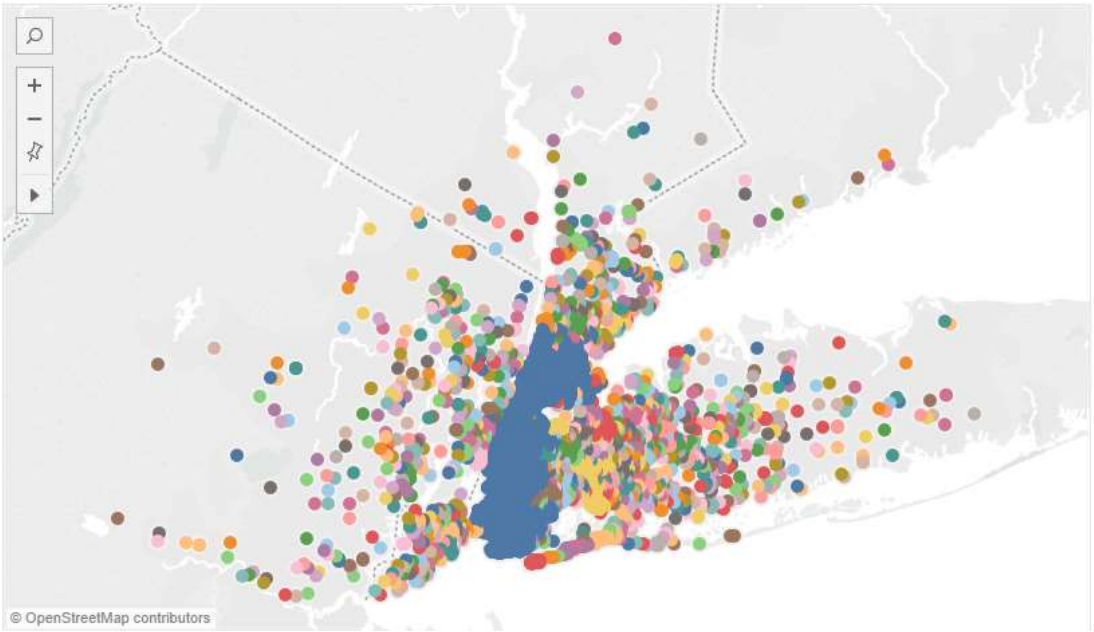


Figure 23: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 400m

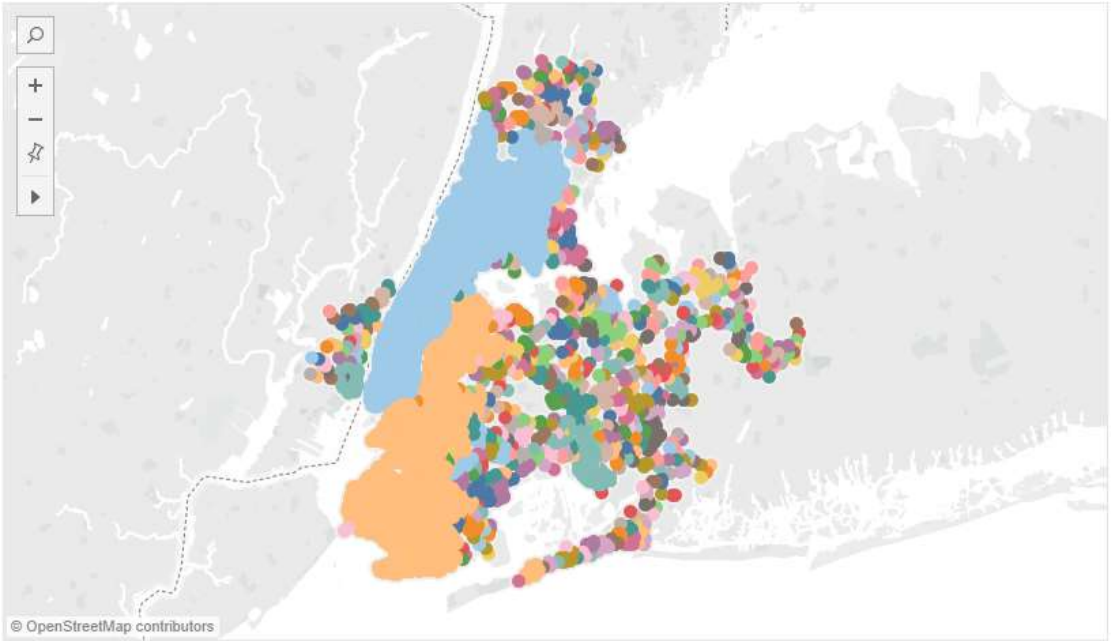


Figure 24: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 270m

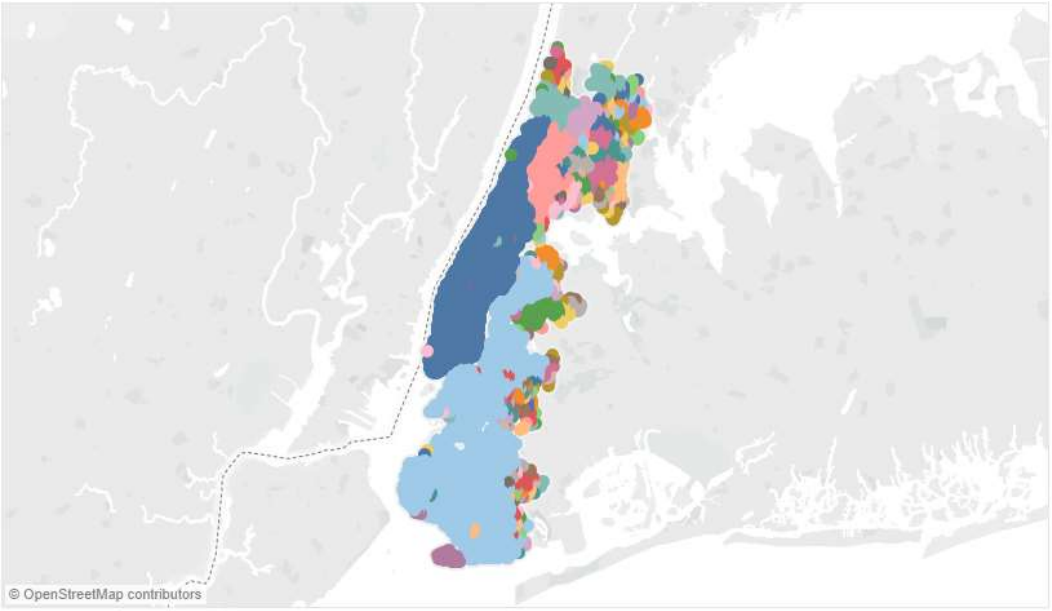


Figure 25: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 200m

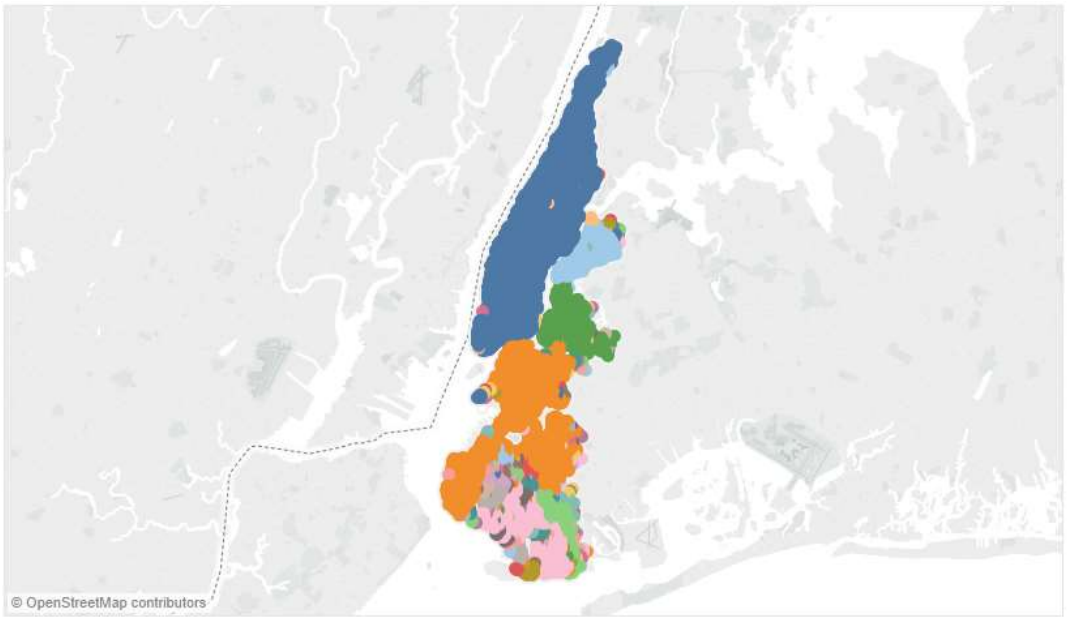


Figure 26: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 150m

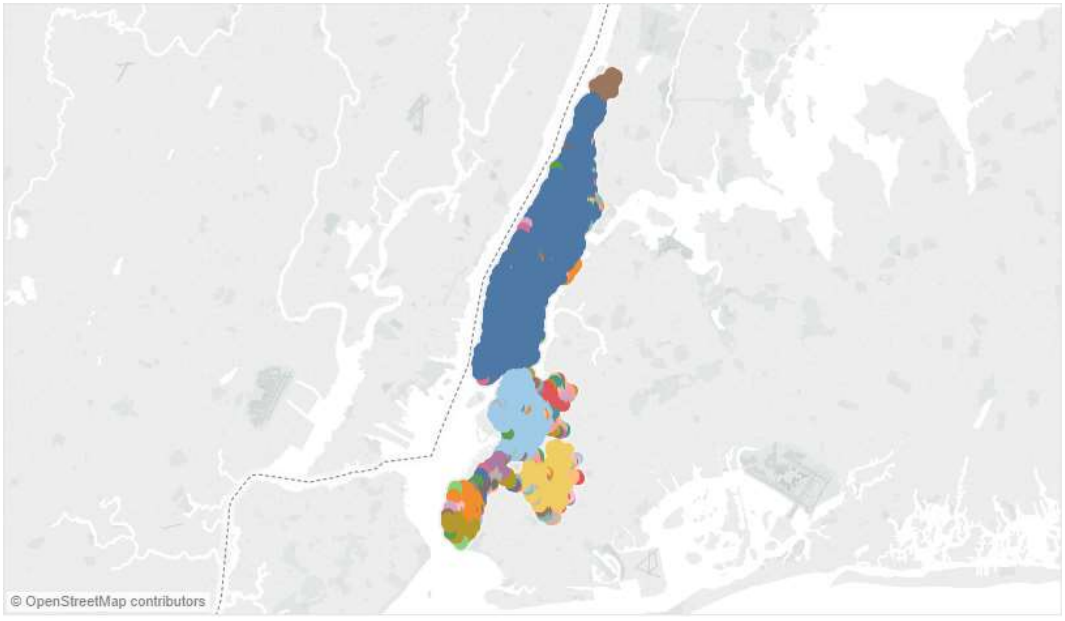


Figure 27: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 100m

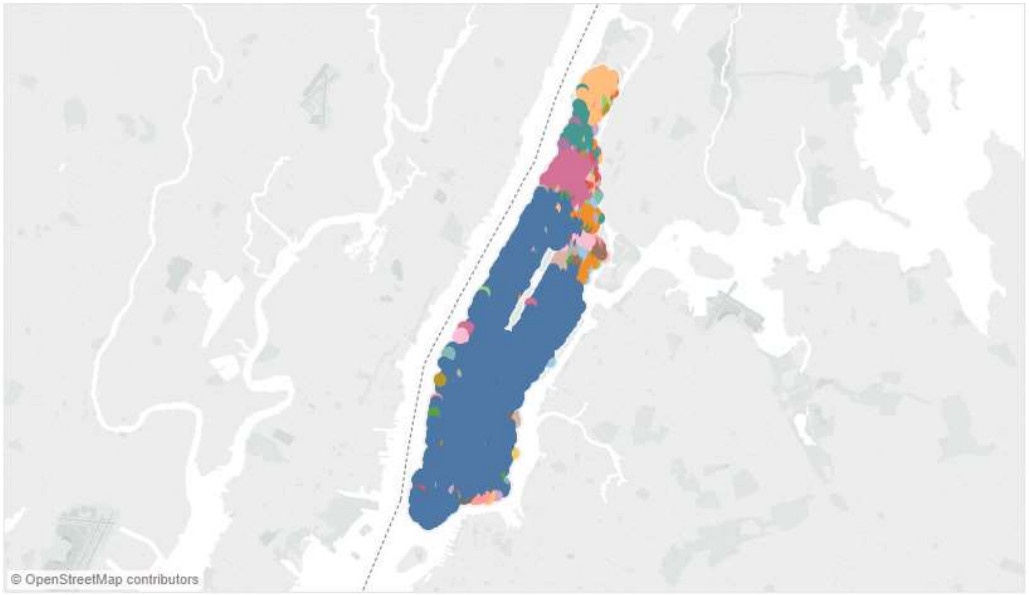


Figure 28: Additional clustering of green and yellow taxi dropoff locations in large badly separated areas using a distance parameter of 50m