Project 2024: DIT863- Statistical Methods for Data Science

David Ward
University of Gothenburg

Contents

1 Introduction

Before starting with the Analysis I want to mention one of the biggest mistakes I made while working on this project. Initially, I was brainstorming about what kind of data I wanted to look into. I went from climate change, to travelling and cargo transport to Swedish criminality records. Every time I found a viable topic I wrote down endless ideas about what kind of statements I want to test, which algorithms i want to implement and so on. Each of these topics failed due to me not being able to find the necessary data I needed for these ideas to work. Especially, the cross referencing part seemed extremely tough to find as you always want the structure of the data sets to be somewhat similar. I changed my approach to first finding a data set before having too many ideas and ended up chosing based on the quality and ability to work with the data set. The dataset mainly has four expectations to fulfil.

- Relatability: I wanted to choose a data set which i can relate to in some kind of way. I have never been to New York but everybody has heard about the crazy traffic and can understand up to a certain degree how the price of a cab ride is determined. This is the main reason why I also wanted to stay away from a more medical data set as my knowledge in that field is particularly limited.
- Interest: I wanted to feel some interest regarding the data set and thinking about the long ongoing fights between taxi drivers and uber drivers which were banned in Germany for a considerable time peaked my interest. From now on I want to call all non-taxi rideshare related transportation as for-hire vehicles or short FHV. This additionally opened up comparisons between the two data sets statistically investigating price or tip differences.
- Opportunity: How many analysis ideas come to my mind? How limited am I to few usable covariates? Many numerical fields which allow lots of analysis. Especially the location tags could be used extensively.
- Operation: How managable is the data set? Secondly, Lastly, the data set needed to be convienient up to a certain degree to handle. My nightmare unveiled in a travel data set which was an excel document formatted with multi-rowed column headers while sometimes having data dispersions in the same column. It was a disaster unravelling and I capitulated to the data mainly due to that reason. On first impression the NYC dataset seems extremely well structured, build for a database instead of visual analysis while offering very little missing or questionable values.

After an extensive search i ended up looking into the New York City Cab data set and ideas started to form immediately.

For this Analysis I would like to thing about myself as somebody who is interested engaging into the market of people transportation in New York City. This data set can provide insights into estimated demands, prices, market gaps estimated arrivals or response rates (for FHV only) and much more. What are cab companies doing right, what are FHVs doing better? Lets explore and start of by having a detailed look into our variables.

2 The Data Set

We mentioned earlier that we consider two largely similar data sets regarding rides of taxis and FHV. The Taxi data refers to the Yellow Cab NYC commpany while the FHV are split between Uber and Lyft data. Both data samples are from January 2023. After excluding some of the less important or just not used for our purposes variables we end up with:

Category	Variables	Additional Comments
time	Pick-up / Drop-off	date and time format
location	PU-LocationID / DO-LocationID	Locations in Section B
trip	distance	given in miles
payment	type / fare / tip	only card tips accounted for
surcharges	airport / congestion / rush-hour / overnight	

Table 1: Available Variables for the Taxi Data Set

Category	Variables	Additional Comments
provider	name	Uber/Lyft
time	Requested/ Driver Arrival / Pick-up / Drop-off	date and time format
location	PU-LocationID / DO-LocationID	Locations in Section B
trip	distance / time	distance in miles
payment	type / fare / tip	only card tips accounted
surcharges	airport / congestion / rush-hour / overnight / total	
Shared ride	requested / shared	binary variables

Table 2: Available Variables for the FHV Data Set

2.1 Pick-up/Drop-off time

Firstly, let us look at the time-related variables. These include the pickup-time and the dropoff-time as well as the request time and the time the driver arrived at the pick-up location for the FHV data set.

One of the first questions we should ask ourselves is how many drivers we need to employ and more importantly how many drivers do we need at different times of the day. To answer this question we will look at the hourly distribution of the number of rides displayed in the first row of Figure **??**. The first column represents the taxi data while the second one the FHV. Not surprisingly the peak is observed around 6pm each day with most people getting of work during that time while other people might meet up with friends, do some sports or get dinner. The only difference between the two distributions is the second peak at around 8am for FHV. This most likely results from people commuting to work. While in a hurry before work it would seem more convenient ordering a pick-up via an app instead of calling a taxi number or standing on the street waving at taxis as the movies suggest.

One of the most important aspects in this first analysis seems to be work Therefore, it might be significant estimating utilization rates based on the fact if it is a workday or not. The second and third column of Figure 1 plots the same variables but separated based on Workdays (second row) and weekends or holidays (third row). The earlier mentioned rush-hour peaks can be observed ever stronger considering only workday data while the weekend histogram appears more balanced throughout the day. Night rides increase drastically on the weekends which doesn't seem surprising. These effects would most likely be even more apparent if we would consider Friday evening as a weekend value while considering Sunday evening a weekday time.

Generally speaking it seems like taxis rates are fluctuating more based on the time while FHV rides have a higher dispersion. This might be due to the easier accessibility of the FHV-related apps especially during odd hours.

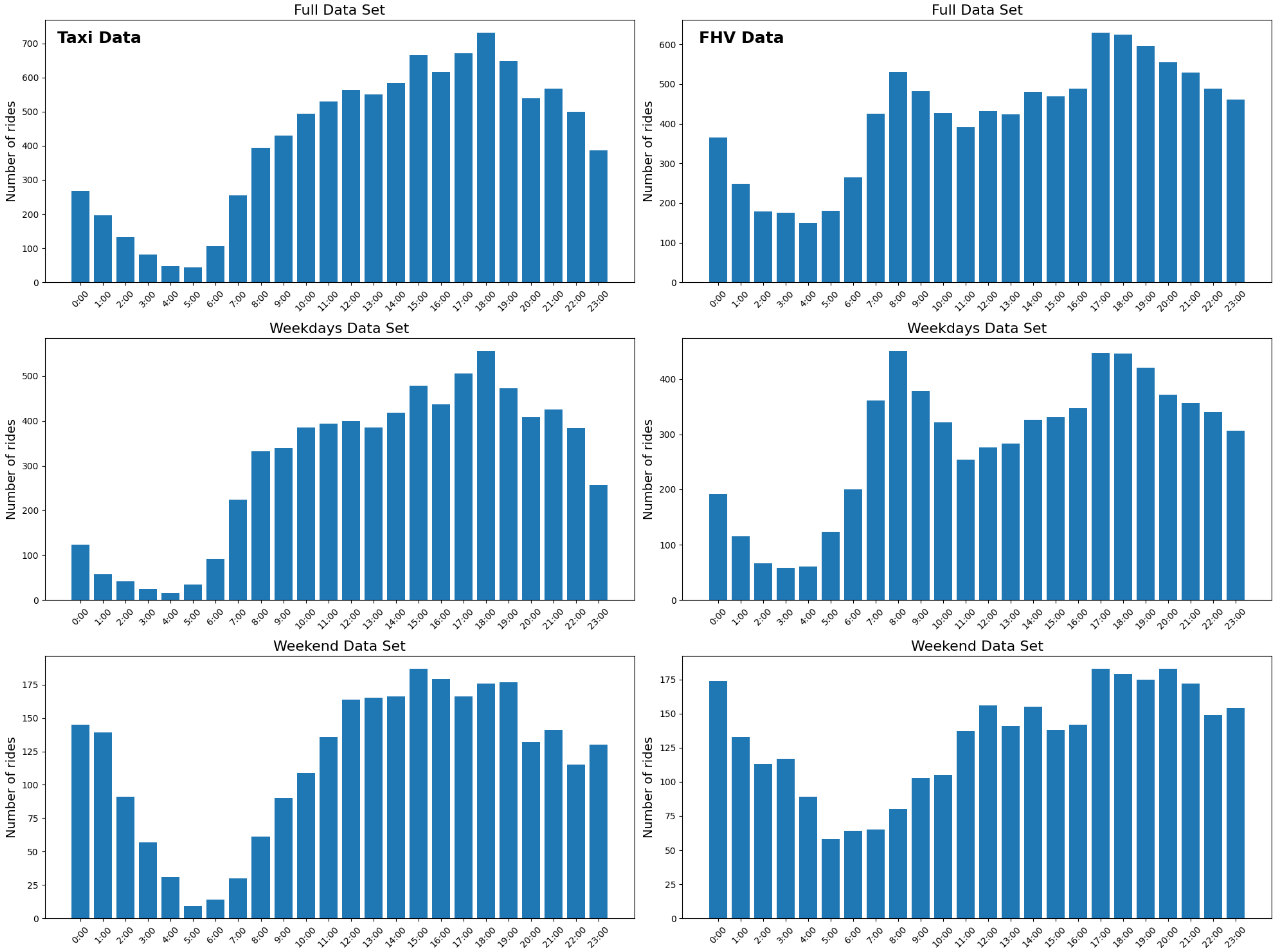


Figure 1: Number of Transportation Services used based on the Hour of the Day separated by Taxi and FHV Services and Weekdays and Weekends or Holidays

This is one of the most essential pieces of information regarding the fact of how many people we should have working during a specific time of the day. On the workdays we have to adjust to the rush-hours and have more drivers ready while on the weekend demand for rides during the night will be rising. For a more detailed analysis we could split into each day of the week. Especially, the difference of Friday evening and Sunday morning could be relevant as well as separating Saturday and Sundays as these might offer quite different utilization patterns. For our entry analysis we have obtain a good overview about demand times.

2.2 Response Time

It is nice to know at which times most people want to use rideshare services but this is no rocketsience to end up with the results above. Considerably more important would be knowing when the demand of people is not met by enough drivers or the next available driver is just far away. Therefore, we do want to analyze the time it takes between the request send by the customer until the driver arrived at the location. Again we can split that up based on the time of the day.

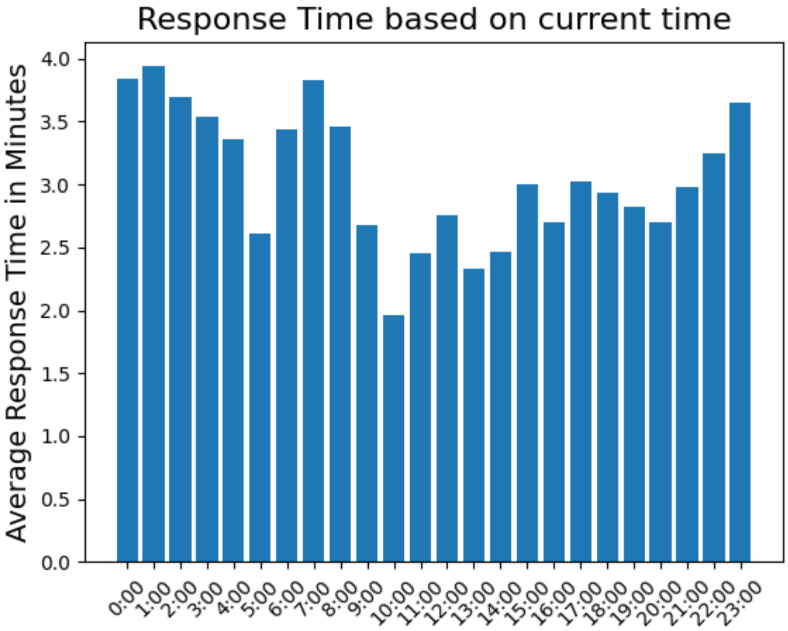


Figure 2: Average Response Time based on the current hour of the day

These results do not tell us too much especially looking at the standard deviation of the means computed in Figure 2 ranging between 2 and 5. This makes these numbers and most definitely the differentiations at which hour the request time is counted insignificant. Maybe the location will tell us more about response times.

2.3 Location

Pick-up and Drop-of Locations have an ID value corresponding to them. A detailed map with all possible locations can be found in Figure 11. Additionally, we can consult a given location lookup table for specific locations.

Following up on the earlier mentioned thought we could see which districts have the highest response time. Visualizing this is quite difficult as there are 239 different locations in our data sample. In Table 3 we display the five highest averaging location response times as well as the lowest ones.

Interestingly enough Location 146 has a negative mean for the response rate. This might seem unreasonable but remembering back that our response time is computed by the drivers arrival minus the guest request it simply means that the driver must have been there before the customer. This is not unlikely at bigger train station or airports. On further inspection i did not find any massively obvious location at this north western area of queens. More importantly we do see that our data gets distributed to far leading to locations having only two pick-ups, which is not a representative number. Therefore we will only split the locations into the five major regions: Manhattan, Brooklyn, Bronx, Queens and Staten Island.

PU Location ID	Mean	Standard Deviation	Count
146	-0.60	9.13	18
192	0.18	8.59	12
193	0.84	7.01	11
66	1.05	6.45	18
118	1.42	0.96	5
⋮	⋮	⋮	⋮
201	6.86	1.24	2
194	7.00	1.46	2
12	7.84	6.49	2
109	8.01	7.10	4
34	10.60	17.44	8

Table 3: The five Locations with the highest and lowest Response Averages

Before we do that there is one additional thing that we can look at more effectively on this finer grade. One of the most important things we need to consider will be surpluses of cars in certain areas. Due to a variety of reasons it could be the case that, assuming there are two locations A and B, many people take a taxi from A to B but nearly nobody takes a taxi from B to A. This would ultimately end up with a surplus of drivers at location B with no customers. We can consider both taxi and FHV data again and look at the areas having the biggest difference between pick-up and drop-off numbers.

Taxi Data				FHV Data			
Location ID	DO	PU	Δ	Location ID	DO	PU	Δ
132	107	450	-343	265	0	411	-411
138	108	319	-211	1	0	55	-55
161	365	503	-138	170	89	117	-28
186	200	327	-127	37	97	125	-28
249	151	235	-84	132	186	213	-27
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
232	52	6	46	113	74	44	30
238	263	217	46	246	129	97	32
50	109	60	49	249	116	81	35
246	207	157	50	230	135	94	41
74	77	26	51	234	134	93	41

Table 4: The five Location with the highest difference between drop-off and pick-up numbers for taxi and FHV data respectively.

The first eye-catching values are the two locations with ID 265 and ID 1 having no drop offs. Consulting the lookup table we figure out that Location 265 refers to unknown values outside of New York City while the location having the ID number 1 is the Newark airport. This might be either due to some laws banning drop-offs right at the airport or maybe rides with end outside of New York City are not included in the dataset. Interestingly, the taxi data has exactly the opposite effect on those two locations with interactions at locations Newark Airport or outside New York City being almost exclusively drop offs. We will look into this in more detail in Section 3.2.

2.4 Trip duration and time

Let us consider the variables that describe the duration of the ride. Namely those two are the trip duration and the distance- I do want to include the fare additionally as I believe that we will be able to predict the price based on those two factors. But firstly lets look into the distribution for the fare. This is the fair excluding all surcharges or additional costs or tips.

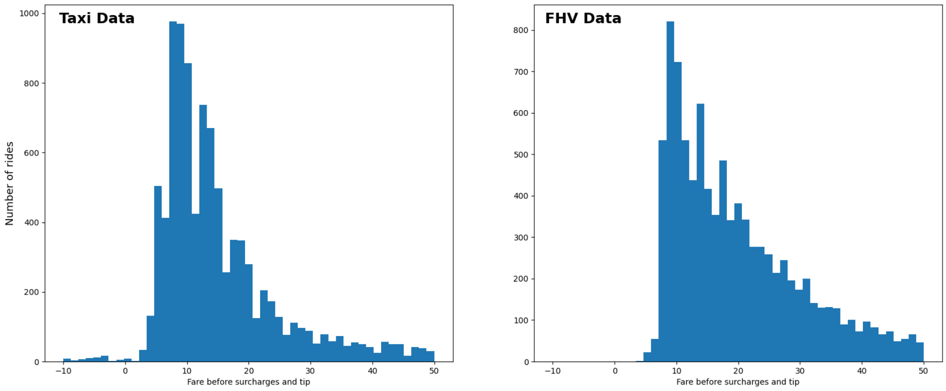


Figure 3: Number of Transportation Services provided by Fare before Tip and Surcharges

The price distributions plotted in Figure 3 seem to follow a gamma distribution which we can confirm by using a QQ-plot. However, the taxi data set seems to have negative values for the initial fare. This seems odd but might be explained by the usage of vouchers, special offers or just errors of data collection. We will disregard those values as we cannot find a logical explanations on how to predict negative fares.

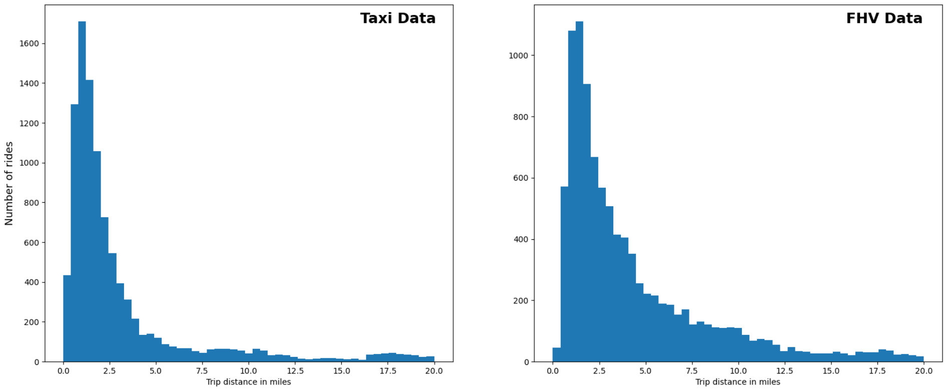


Figure 4: Number of Transportation Services provided by the distance covered

The distance diplayed in Figure 4 also seems to be largely gamma distributed with additional two peaks at around 10 and 17.5 miles. These are correspondent to the distance to New York's three airports with La Guardia Airport trips having an estimate of around 10 miles, JFK Airport trips around 13.5 miles and Newark Airport around 18 miles.

The time of the ride seems to be very similarly distributed and does not show any signs of additional peaks or irregularities. (MISSING QQ PLOT)

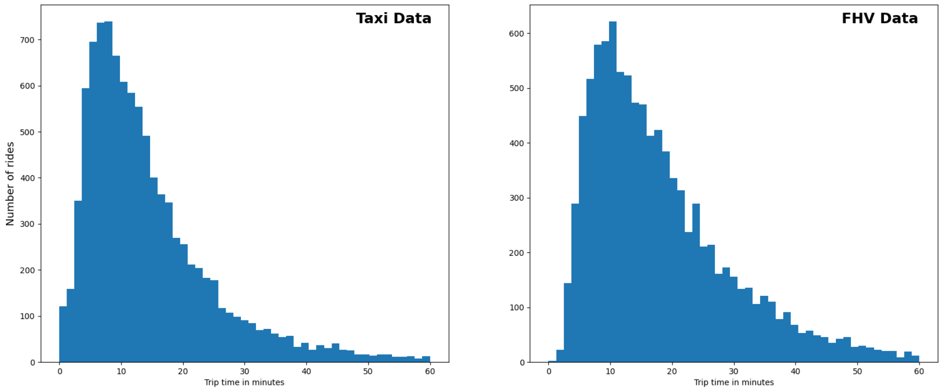


Figure 5: Number of Transportation Services provided by the Time of the Trip in Minutes

Now that we have seen the distribution for fare, distance and time we would like to check if we can predict the price based on the time and distance of the trip. Firstly we can simply check the correlation matrix between the three variables to see if potentially the 'taxi clock' only counts the distance or time.

Taxi	miles	time	fare
miles	1	0.50	0.81
time	0.50	1	0.45
fare	0.81	0.45	1

Table 5: Correlation Matrix for the Taxi Data Set

FHV	miles	time	fare
miles	1	0.79	0.87
time	0.79	1	0.79
fare	0.87	0.79	1

Table 6: Correlation Matrix for the FHV Data Set

It seems that that the distance is the more important influence of the fare of a ride which is logically sound as you shouldn't have to pay much more if the traffic is bad. Interestingly enough the covariance between time and price is significantly worse for taxis compared to the FHV data. This seems surprising as we thought that the price is fixed beforehand for an Uber for example while a traditional taxi uses the taxi meter.

To evaluate if the missing bit of information about the price is in the time we will construct a small linear model that predicts the price based upon the miles and time of the trip for both the taxi and FHV data. The model formula is

$$fare = \beta_0 + \beta_{miles} x_{miles} + \beta_{time} x_{time}.$$

The results are displayed in Table 7. According to this regression it seems like that taxis are quite a bit more expensive. We will evaluate that statistically in **??**. Additionally a base price of 6.14$ is assumed for a taxi while 5.35$ for the FHVs. The price per milage increase seems significantly larger for taxis, however, we should consider that the time seems largely irrelevant for the taxis while it seems to at least have minor influence for the FHV. These estimated beta values for miles and time display the price increase per unit. Taking a transportation inside of New York will surely drive under 60 miles per hour making a minute pass faster and therefore $\hat{\beta}_{time}$ more relevant.

Data Set	$\hat{\beta}_0$	$\hat{\beta}_{miles}$	$\hat{\beta}_{time}$	R^2	MSE
Taxi	6.14	3.47	0.05	0.54	112.36
FHV	5.35	2.39	0.38	0.71	88.03
Taxi fixed	6.42	3.60	0.04	0.86	34.29
FHV fixed	6.13	1.97	0.41	0.75	46.96

Table 7: Linear Regression Results for Covariates Trip Time and Distance to predict the Base Fare

Now for the full model evaluation we observe that the model seems to be doing quite bad for the strong correlation we already have with the miles variable. Upon further investigation, the standard error values seem extremely high with an estimated error of 112$ for the Taxi data set and 88$ for the FHV data set. If we look into the fare values we find quite a few negative values for the taxi rides which made predictions incredibly hard. When filtering out extreme values ($0 <$ fare < 100) the resulting models looks for the taxi rides significantly improved and results are displayed in Table 7 under fixed models. The correlation matrix also improved and is shown alongside the almost unchanged "fixed" FHV correlation matrix in Table 8 and Table 9.

Taxi	miles	time	fare
miles	1	0.50	0.91
time	0.50	1	0.50
fare	0.91	0.50	1

Table 8: Correlation Matrix for the adjusted Taxi Data Set

FHV	miles	time	fare
miles	1	0.79	0.86
time	0.79	1	0.79
fare	0.86	0.79	1

Table 9: Correlation Matrix for the adjusted FHV Data Set

For taxi rides the fare is almost purely dependent on the distance of the trip with a correlation of over 90%. The FHV-Algorithm seems to be more complicated which we might uncover further in Section 4.

2.5 Surcharges and Tips

There are quite a lot of different surcharges and we don't want to go into any details but an overview can be found in Table 10. We identified that some surcharges are always applied which will mean we will mostly disregard them for future analysis. Others only appear rarely. Between the two data set were two significant differences with congestion surcharges having higher occurrences in the taxi data set while tolls appeared more often in the FHV data. Lastly, the rush hour and overnight surcharge are mentioned in a single variable making it hard to differentiate between the two.

The Tip amounts are displayed in Figure 6 following a similar distribution observed for distance and fare measures. For additional Analysis we have to keep in mind that only cash tips are accounted for meaning we will disregard any cash payments from further tip related analysis. This means that most likely tip values are underestimated in the data set.

Surcharge	Values	Percentage of rides	Which Data Set
Tolls	6.94	10.07%	both
Black Car Fund	several	100%	FHV
Airport	1.75(Taxi), 2.5(FHV)	7.8%	both
Sales Tax	several	100%	FHV
Rush Hour	several	57.30%	Taxi
Overnight	several	57.30%	Taxi
Congestion	2.5(Taxi), 2.75(FHV)	64.18%	both
Improvement	1.0(Taxi)	100%	Taxi
MTA Tax	0.5(Taxi)	98.94%	Taxi

Table 10: Different types of Surcharges and their Occurrence

3 Hypothesis Testing

During the ealier chapters we made statements based on our domain knowledge or based on a visual conclusion. however, we would like to statistically evaluate these statements which can be archived using a statistical test, more specifically a hypothesis test. At the end of this section we want to know if taxis are in fact more expensive than FHV. Also we would like to know if there is a connection between the surpluses of cars between FHV and taxis and lastly if people having to pay tolls are less prone to tip.

3.1 Are taxi transports more expensive than FHVs?

Beginning with the question if taxis are more expensive we need to agree on a handful of conditions. Firstly, we will exclude any extreme values as our goal is to predict an ordinary taxi ride. Excluding those makes sense as negative values will alter our computed means significantly while higher distances tend to have lower per mile ratios and not representing an ordinary taxi ride. We want to consider the price including all surcharges as transports should not be considered cheaper just because they hide more of their price inside the surcharges. Remember, we did observe FHV having higher surcharges on first impression. We will compute the price per mile as this is far more important than the actual full price paid if we compare a short ride with a long one. This is of course still not perfect as shorted rides tend to be more more expensive even on a per mile ration.

Therefore we are assuming our random sample passengers to be $X_1, ..., X_{n_x}$ for the taxi data set and $Y_1, ..., Y_{n_y}$ for the FHV data set. We assume both these samples to be independently, identically gamma distributed and additionally between the two samples to be independent. Our parameter of interest are the means (μ_x, μ_y) of the final fare which are normally distributed based on the sample we pulled according to the Central Limit Theorem. We want to evaluate the null hypothesis

$$H_0 : \mu_x - \mu_y = 0 \quad \textbf{against} \quad H_1 : \mu_x - \mu_y \neq 0.$$

We use a two sided hypothesis test as we might end up with the result that taxis are cheaper than FHVs. All these assumptions point towards a two-sample t-test which we can evaluate on a significance level of 95%. Our test statistic is

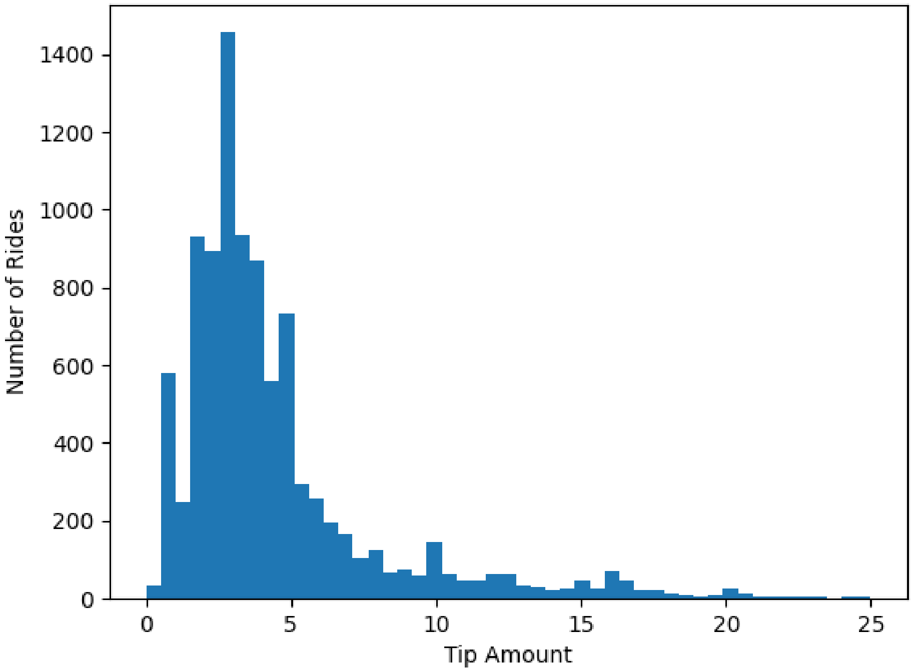


Figure 6: Distribution of the Tip Amount of both Taxi and FHV data

$$t_{obs} = \frac{\mu_x - \mu_y}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim t_{df} \quad \textbf{where} \quad df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2/(n_y - 1)}.$$

The term s^2 stands for the known sample standard deviation. Computing the means we can almost consider the test unnecessary as the mean of the per-mile price of FHV sits at around 7.98 while for taxis it's at 12.44. Unsurprisingly, our t_{obs} is extremely high with around 59.3 rejecting the null hypothesis which a certainty of almost one hundred percent as visible in Figure 7 where i did not include the t_{obs} value as it is too far off the distribution center.

Is it actually the case that taxis are around 50% more expensive than FHV? We did choose a relatively large sample with slighly below ten thousand observations each. Additionally, we dealt with unreasonably large or small numbers. Could it be that taxi rides are usually shorter than FHV rides? Partly, yes because looking back at the distribution histogram in Figure 4 we do observe that taxi rides tend to be slimmer along the axis for the shorter rides. Additonally, with the clearer peaks observed at the airport related rides we do assume transportation to the airport to be more expensive. This is already reasonable with the airport surcharge in place. Could another reason be that people in FHv have the opportunity for a share ride which would obviously result in a lower cost for the company and therefore price for the customer? Not really, as checking the number of shared rides unveils that only around 1.5% of FHV transportations are shared.

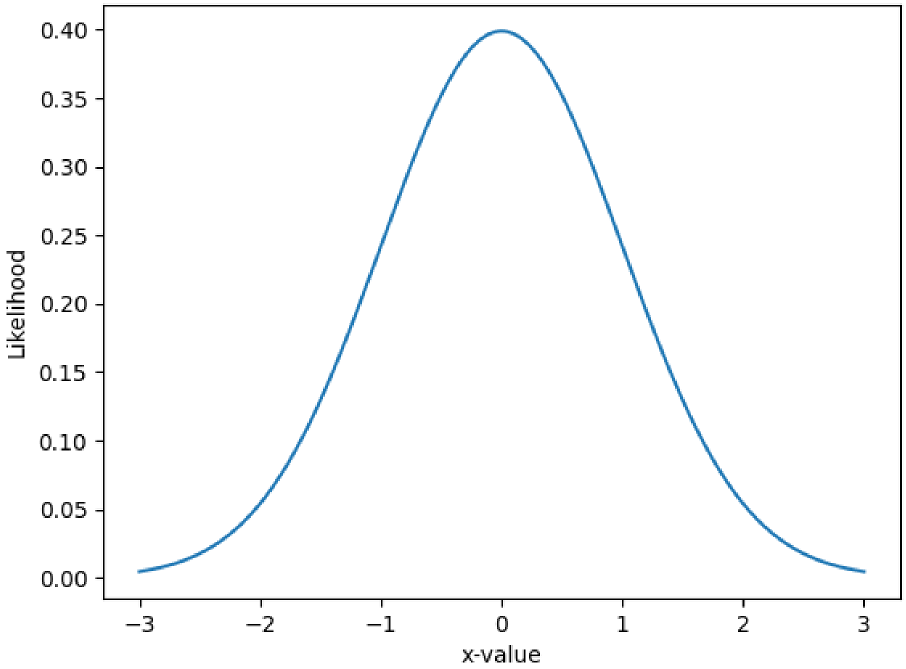


Figure 7: T distribution for around 24.050 degrees of freedom.

3.2 Does one of the data sets create more surplus of drivers in certain locations?

Returning to the car surplus issue mentioned in Table 4 we do want to statistically evaluate if there is any connection between the two data sets regarding the drop off and pick up surpluses. Our idea is to use a McNemar's test between the two groups of surplus of drivers and lack of drivers while the 'experiment' is the switch between the data sets.

Our random variables for both data sets will the a binary variable indicating if the corresponding location is experiencing a surplus or lack of drivers. For the taxi data set we have $X_i \in \{s, l\}$ for $i \in$ LocationIDs while for the fhv data set $Y_i \in \{s, l\}$ for $i \in$ LocationIDs. We want to disregard all Locations which are in neither of the two groups and will consider all Locations with a difference in drop-offs and pick-ups below 10 as insignificant. This means that an observation that has a surplus in cars for taxis while it is balanced in regards to FHVs is deemed insignificant. This seems logical as we want to observe if car surpluses are equalized by the other data set or actually could be meaningful for our entry into the market. Further on we do not want to give any meaning to a Location having a single car surplus in the time of a full month. The resulting Table is displayed in Table **??**

-	$X_i = s$	$X_i = l$
$Y_i = s$	11	10
$Y_i = l$	9	4

Table 11: Number of Locations which have a significant Surplus or Lack of Drivers based on the Number of Pick-ups and Drop-offs and their Movement between the Data Sets.

With these relatively small numbers we can apply an exact McNemars's test and estimate the discordance as

$$\hat{\rho} = \frac{\min(n_{sl}, n_{ls})}{n_{sl} + n_{ls}}.$$

This estimate returns the lower frequentistic probability of the movement of a location having a surplus in the taxi data set to a lack in the fhv data set and vice versa given that a change is happening between the data sets. In more mathematical terms it estimated the discordance

$$\rho = \min(P(X_i = s, Y_i = l | X_i \neq Y_i), P(X_i = l, Y_i = s | X_i \neq Y_i)).$$

The null hypothesis will be is our estimated discordance parameter $\frac{1}{2}$. That would suggest that the data set does not have any effect on the number of surpluses or lack of cars. We will come back later to discuss the results of this test. For now let us evaluate the null hypothesis

$$H_0 : \rho = \frac{1}{2} \quad \textbf{against} \quad H_1 : \rho \neq \frac{1}{2}.$$

We will evaluate this based upon a binomial distribution under the assumption that H_0 is true. The used test statistic s_{obs} will simply be the minimum of the two movement related values n_sl and n_ls.

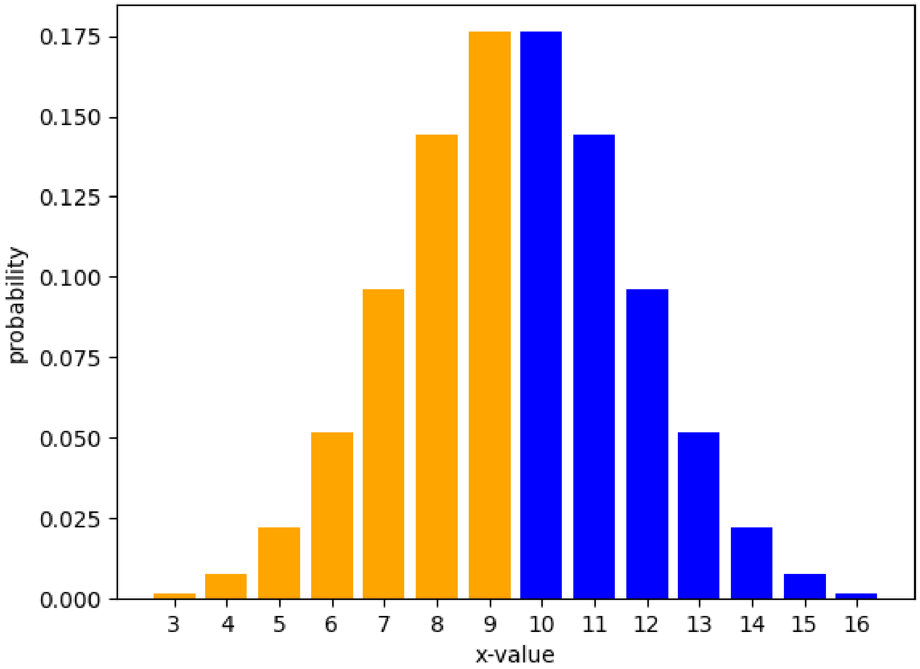


Figure 8: Binomial distribution with $n = 19$ and $\rho = 0.5$ split along the observed s_{obs} value 9.

Figure 8 displays that our observed test statistic value is set exactly in the middle of the binomial distribution. Therefore with a p-value of exactly 0.5 we cannot reject the null hypothesis and assume that the probability of a surplus given that this location is either a surplus or a lack is the same in both data sets.

We can compare the numbers of movement to the locations remaining in the same group. With locations having a surplus in the taxi data having roughly the same frequentistic probability of being a surplus or lack area in the FHV data set. On the other hand, however, locations which are an area lacking taxi cars are 2.5 times more likely being a surplus area for FHVs. This means that many of the created unbalanced areas are equalled out by the other transport provider. This seems to be the more important take away from this analysis.

As we did only consider unbalanced locations with a difference of more than ten cars we need to investigate further to reveal the importance of each of the imbalances to deliver meaningful results. Additionally, there are more companies involved in transportation services in New York (e.g. Green Cabs) which could create or balance out asymmetries of drop-off and pick-up rates.

3.3 Do people tend to include more tips if they don't have to pay surcharges?

For our last hypothesis we want to focus our attention towards the tipping behaviour of the customer. This analysis would be more interesting from a drivers perspective but could yield inequality issues regarding the wages of each driver. Let us assume that the tipping behaviour depends on the amount of surcharges paid. We do want drivers to mostly remain in the same areas as with an increase of local knowledge they would be at lower risks for accidents, traffic jam avoidance and many other thing yielding positive reviews. This would make a driver less likely wanting to pick-up customers at the airport as the estimated tips could be significantly less based on the airport fee paid.

We do want to take a slightly different approach here and assume the mean of the given tip based on the full data set and compare it to a fraction of the data set having to pay more than a certain threshold of surcharges. For the full data set we will still exclude negative fares and remove all cash payments of the taxi data set due to the reason of the data not accounting for any cash tips. This results in an average tip of 2.4177 USD. With this constant value we can construct a one-sample t-test. The null hypothesis is

$$H_0 : \mu = 2.4177 \quad \text{against} \quad H_1 : \mu \neq 2.4177.$$

We filtered the both data sets additionally according to only include congestion surcharged individuals and compute the sample mean μ and sample standard deviation s. With that we can construct our test statistic

$$t_{obs} = \frac{\mu - 2.4177}{s\sqrt{n}} \sim t_{n-1}$$

The calculated $\mu = 3.0751$ and with this big of a sample size (over 11.000 entries) this difference is again deemed extremely significant. The observed t-value is with about 19.8 still extremely far off the cluster of the distribution with an expected probability of H_0 being true according to the sample of almost zero percent. The resulting null distribution is extremely similar to Figure 7 because the t-distribution asymptotically resembles a normal standard normal distribution for high degrees of freedom.

But importantly that we rejected the null hypothesis does not mean that our initial statement is true. A quick look at the computed means suggest the opposite behaviour, that people tend to tip more if they have to pay additional surcharges. Other surcharges confirmed this claim as numbers were even more extreme for values such as the airport surcharge or tolls. However, those seem more reasonable considering that maybe tourists want to loose their last cash on the way to the airport or tolls are only for people who desperately need to get somewhere fast.

4 Machine Learning

Not that we are familiar with the data sets and have statistically answered questions regarding useful bits of information we can start implementing algorithms which will help us engaging in the market. Our final vision is a program that can assist and answer many of the important questions to meet customer needs, deliver reliable information and propose business decision advice. A giant helping tool that would combine a variety off important information and would support us in making the step of entering the market. There are many problems which could be approached with Machine Learning Algorithms but I want to mention a few that we will focus on.

- Predict driver allocations based on trip clusters - We will use K-means to find clusters which have high amounts of traffic between them.
- Predict basic fares for before a ride - We will use linear regression to predict a fare based on the expected miles between two locations.
- Predict arrival times before the trip - Based on the time of the day, response time and other factors we can predict the estimated arrival time.

Further analysis could tackle problems such as the estimated amount of fares to pay, predict the traffic conditions to further improve the predicted arrival time or an estimate of how likely a shared ride will be if requested.

4.1 Predicting driver allocations based on clusters

Concerning the first question mentioned earlier we want to solve the problem of driver allocations. Specifically, that means based upon the obtained frequencies between locations we want to construct a clustering that groups locations together that have many transportations between them while keeping the number of rides between the clusters low.

To utilize K-means for this process we need to decide on a few approaches as we don't have a representation of our desired frequencies yet. Firstly, we will only consider 'important' locations meaning that they have at least fifteen drop-offs or pick-ups from or to a single other location. Let us denote the set of important locations as I. The basic idea is that each location is represented by a vector

$$v_i = [n_{i1}, n_{i2}, ..., n_{i|I|}] \in R^{|I|} \text{ where } n_{ij} = \text{ number of rides from location i to j.}$$

Each location is now represented by a vector that splits all rides starting in this location to the drop-off spots. We are hoping that connected by many rides neighborhood will have a similar pattern as the same popular destinations should be reachable. If this would be successful we would have driver zones which we could use to deploy drivers with a relatively high probability that this Random Walk (or rather drive in this example) would stay within that cluster.

For evaluation purposes and the best choice for k we print the best inertia obtained with ten initializations for each k in Figure 9. This choice is not an easy one to make but values such as 5,7 or 11 could be promising.

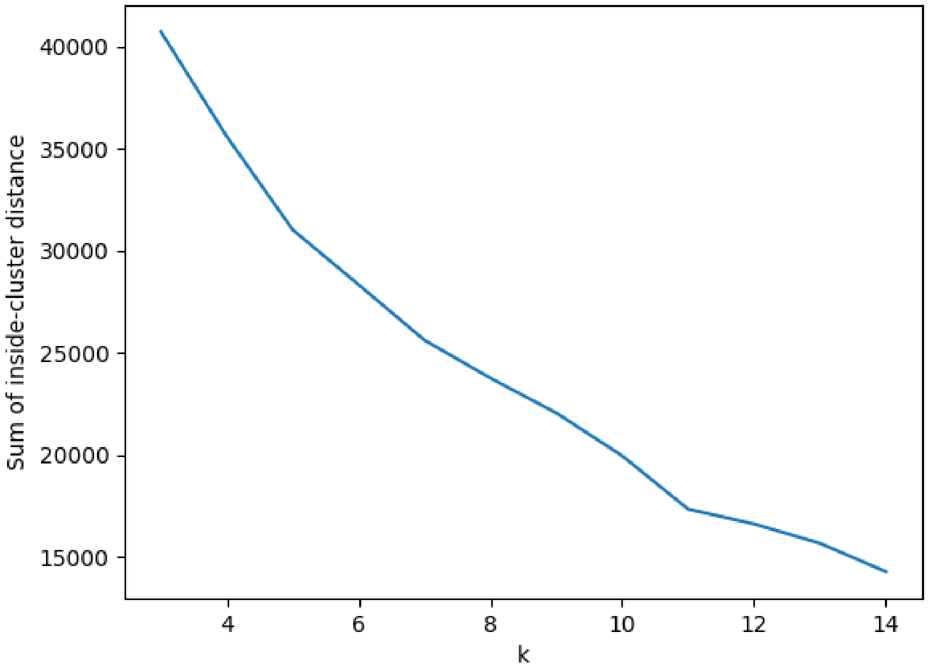


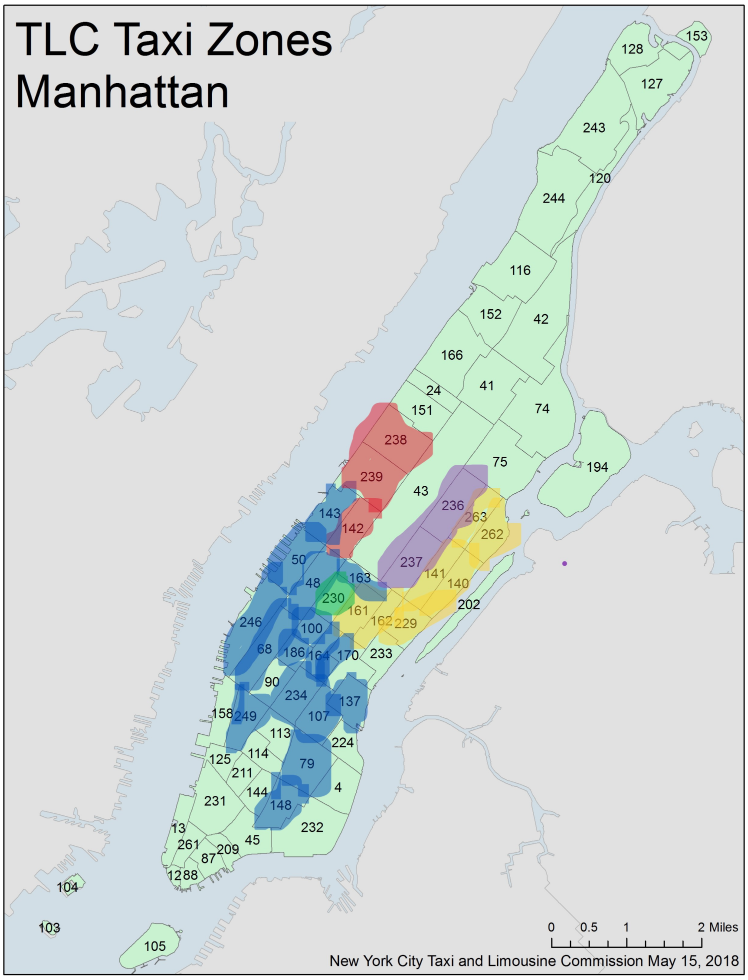
Figure 9: Inertia displayed for each different choice of k using ten random initiations.

The labels refering to the five different clusters might be especially interesting regarding the five districs of New York that the data set is constructed upon. Is the solution as simple to just distribute the locations largely based upon their district? The short answer is no, with the definition for important districts Staten Island and the Bronx completely disappear from our data and while Queen and Brooklyn only have a single label among them, Manhattan is widely split up which is displayed in Figure 10. This does not look too bad with the interesting single outlier in green that occupies a single cluster for himself. Upon further investigation it seems like that quite a lit of trips to the airport start from that location which would explain the fundamentally different structure it has compared to the surrounding areas and why it takes a single cluster for itself.

This does seem like a promising start with evaluations of different important location definitions or varying number of cluster revealing more angles to this analysis. For now we will be satisfied with these results.

4.2 Predicting prices for a transportation

Recall that the basic fare price largely depends on the distance covered (and the time for FHVs). Therfore, we want to build a linear model that predicts the response price based upon



New York City Taxi and Limousine Commission May 15, 2018

Figure 10: K-means results for 5 clusters for the important locations in Manhattan.

the distance and time. However, the values for distance covered and time are fixed after the ride as for a prediction we firstlyneed to predict those as well. The distance can mostly be predicted based upon the location while the time can depend on a multitude of factors such as the time and day, the current traffic conditions and of course the predicted distance. We will consider that question in more detail in 4.3.

For now let us concern ourselfs with the prediction of the distance. When the request for a ride appears we will know the pick-up and drop-off location and the simplest approach would be to take the mean of all earlier occured rides between those two locations. Predicting the distance values like that we can predict the cost for the taxi set with a simple linear model after our analysis done in Section 2.4.

When predicting the distance we obtain the results displayed in Table 4.2

Model	p	MSE	R^2
Simple Linear Regression	1(2)	61.97	0.803

Table 12: Simple Linear Regression Model Evaluation with the predicted distance on the taxi data set

This model can be improved further by including additional taxes, the predicted time values, the utilization rates and much more.

4.3 Predicting arrival times for transportations

missing due to time issues :(

5 Conclusion and Pitfalls

We did work on a variety of tasks regarding the New York City Cab data set. We started off with some basic analysis and figured out that one of the most important factor, regarding the utilization throughout the week, is work. We searched for locations with large discrepancies between drop-off and pick-up numbers which could be useful regarding future improvements of driver distribution. We constructed a small linear Model to evaluate the correlation between distance or time and the fare ob the ride before evaluating all different types of surcharges. Afterwards, we statistically tackled resulted that taxi rides are on average more expensive than FHV based on the price per mile. We also observed that car surpluses are not more common in either of the data sets but that most of the driver lacks or surpluses are fed by the other providers. Lastly, we concluded by figuring out that usually people tip more if there are additional surcharges. We split Manhattan up into five zones based upon similar drop-off location pattern and created a small linear prediction of the price that still needs some work.

There are a multitude of potential problems that still need handling and might be issues in the current analysis. For example there seems to be multiple fraudulent entries especially in the taxi data set which would need to be detected and removed. The hypothesis testing had the issue of to big of a sample making even the smallest differences extremely meaningful.

A Apology

I want to apologize for the most likely awful spelling and sentences but I didn't manage to effectively manage two exams and projects after Christmas. Cheers!

B Location ID Maps

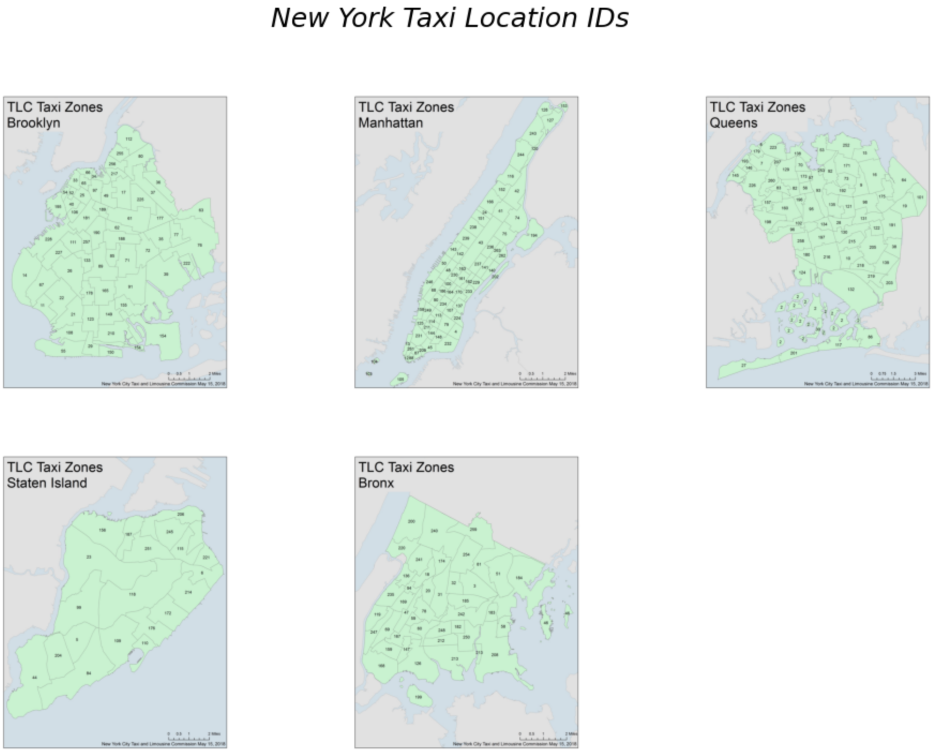


Figure 11: Maps of New York districts with corresponding Location IDs