Impact of Manhattan Congestion Surcharge on For-Hire Vehicles Report 3

Katie Voorhees[1], Rufei sheng[1], Xiaoning He[1], Soham[1], and Sam Manzi[1]

[1]New York University

June 24, 2019

Abstract

Comprehensive understanding of anticipated impacts of transportation pricing policies in an urban system is crucial for both decision-makers and stakeholders. This paper aims to provide a data-driven framework for quantitative evaluation of such impacts with the recently introduced Manhattan congestion surcharge as the case in point. To do so, we will use open data to develop a predictive model to estimate the impact of pricing changes on the demand for public transportation and for-hire-vehicles. Anticipated results include developing a web-based tool to enable voters and policymakers to quantitatively evaluate the impact of proposed congestion pricing changes on transportation modal shift and related urban impacts and stakeholder value.

Introduction and Motivation

Commuters in New York City have a plethora of options. Subways, buses, ferries, private vehicles, taxis, bike-shares and a growing market of ride-sharing platforms are now being met with share programs for electric scooters and bikes. Despite the bevy of choices, much of New York's growing population finds itself regularly frustrated while traversing the city. Subways are experiencing an increasing amount of delays (Pearce, 2018) due to an aging infrastructure. Driving speeds in Manhattan's Central Business District (south of 60th street) have slowed to an average of 7.2 mph, with speeds in the "Midtown core" 30% slower than the rest of the CBD (DOT, 2018). While the subway and bus are still the third and fourth most popular ways to travel in NYC (after private vehicles and walking) (DOT, 2018), public transit ridership has been declining since 2015 (MTA, 2017), exacerbating financial challenges at the MTA. Department of Transportation (DOT) officials attribute the loss in ridership to consumers increasingly choosing For-Hire Vehicles (FHV), particularly during off-peak hours. The challenges facing NYC's complex transportation ecosystem have motivated this research.

New York lawmakers have responded to these challenges with an initiative unprecedented in the United States - congestion pricing. With aims to relieve congestion and recover some diverted revenue, New York State's Department of Taxation and Finance imposed a congestion surcharge on February 2, 2019, aimed at taxis and FHVs(of Taxation & Finance, 2018). The tax adds $2.75 to FHV trips, $2.50 to taxi trips and $.75 to ride-share or pool trips. The tax applies to every trip that begins, ends or passes through Manhattan south of 96th Street. On April 1, 2019, the NY State budget passed, approving a future congestion tax for all vehicles entering Manhattan's CBD below 60th St.(Assembly, 2019). Revenue from the tax is earmarked for MTA capital investments, but this larger scale congestion pricing scheme is not scheduled to be in effect until December 2020. The tax will collect around $11 from both for-hire and private vehicles and will include exemptions for New Yorkers with low incomes and drivers with disabilities.

While transportations advocates eagerly anticipate the revenue from the congestion pricing plan, many drivers, particularly from the outer boroughs and suburban areas, are skeptical of the plan's potential to reduce traffic congestion and lament paying a tax aimed at repairing a public transit system they may not use. Regardless of one's political opinion of the plan, the circumstances emphasize the importance of understanding the effect of pricing changes on people's transportation choices.

This research will use open data to model the dynamics behind transportation mode shift in New York City. We will use this data-driven model to explore the impacts of the initial Feb. 2, 2019 congestion tax and anticipate the mode shift that may be provoked by the December 2020 implementation of larger scale congestion pricing, as well as an alternative scenario in which MTA prices were increased by 30% in order to fund repairs. Below we will discuss the experiences of other world cities that have implemented congestion pricing as well as relevant work on estimating the impact of transportation interventions.

Literature Review

Our project builds upon the previous work done at Center for Urban Science and Progress (CUSP) on the 'Impact of Bike Sharing in New York City' (Sobolevsky, Levitskaya, Chan, Postle, & Kontokosta, 2018) which looks into the balanced baseline scenario based on a transportation choice model to describe projected customer behavior in the absence of the Citi Bike system as well on the 'Prediction of Mode Shift in New York City' (Capstone, 2018) which looks into the disruption of the subway ridership during the morning rush hour due to app-based ride-hailing services(specifically Uber). Tong has observed how the built environment, individual characteristics, and trip-based variables play a significant role in mode choice decisions(Tong, 2015).

It is notoriously difficult to estimate the impact of congestion pricing on traffic outcomes in cities. Congestion policies are often well publicized and discussed prior to implementation, affording users the opportunity to adjust their behavior ahead of time–for instance, altering schedules to avoid the surcharge, or altering routes to bypass the congestion zones altogether. Congestion pricing policies are often accompanied by an expansion of public transit service in or near the pricing zones, which further obfuscates their impact (Gibson & Carnovale, 2015-2016). However, Gibson et al. exploit a natural experiment in the city of Milan to estimate the price elasticity of vehicular volume under such a congestion surcharge policy. In 2012, a municipal court unexpectedly suspended Milan's surcharge pricing in "Area C" before reinstating it several months later. The sudden nature of the ruling precluded the typical measures (increased public transport, altered parking fees, etc.) that accompany changes in congestion pricing. Thus, Milan's unanticipated injunction provides a unique opportunity to estimate the effect of congestion pricing independent of other factors

Gibson et al. find that congestion pricing in Milan's "Area C" reduces vehicle entries into the area by 14.5 percent; in their examination of Stockholm's 2006 congestion charge trial, Jonas Eliasson et al. (Eliasson, Hultkrantz, Nerhagen, & Rosqvist, 2009) find consistent estimates of decreases of 18-22 percent in surcharge moderated traffic zones (dependent on time of day). Gibson et al. also observed a reduction in air pollution of 6 to 7 percent and a net economic benefit of ˜3 billion USD annually. According to Dr. Paolo Beria of Politecnico di Milano, the policy generated 13 million Euros in revenue one year after its implementation(Beria, 2016). They estimate the price elasticity of vehicular entries to be a .3 percent reduction per 1 percent increase in congestion surcharge. In *Downtown Congestion Pricing in Practice,* a study of congestion surcharge practices in Singapore, London, Stockholm, Milan and Gothenberg, Lehe notes a "zero price effect", wherein the majority of the reduction in congestion priced traffic zones can be attributed to the imposition of a non-zero charge.(Lehe, 2019) They note a marked attenuation in traffic abatement as congestion charge increases. Nonetheless, these analyses provides a compelling basis to consider congestion pricing in urban high traffic zones as a legitimate policy initiative.

One of our primary aims is to evaluate the economic impact of transportation mode shift under various pricing scenarios. A highly relevant conclusion of Gibson et al. is that areas with low access to public

transportation are far more sensitive to congestion pricing than are areas with high availability of public transit. Armed with the knowledge that congestion pricing may place a heavier burden on residents who live in transit deserts, policymakers might be better equipped to evaluate the social equity of various pricing scenarios.

In order to see if there was a mode shift and to assess the impact of the policy, we would need to see if the choices people made for commuting would change with the new pricing or not. So, our model has to simulate the transportation choices(Ramming & Scott, 2002) that commuters make based on factors like their wages, the cost of that transportation mode and so on. The discrete choice model has been proven to be an effective method in the past to do transportation choice modeling(Train, 1977). In economics, discrete choice models, or qualitative choice models, describe, explain, and predict choices between two or more discrete alternatives, such as entering or not entering the labor market or choosing between modes of transport. To work with a discrete choice model, there should be a finite set of alternatives which are mutually exclusive and collectively exhaustive. So, considering the modes we wanted to focus on and the available data, we decided to keep 5 modes: Taxis, for-hire vehicles(FHV), shared FHVs, public transportation and walking.

We looked into various types of discrete choice (Yu & Sun, 2012) models to see which one would be a good fit for us. There are 4 typical discrete choice models: Multinomial logit (MNL) model, Nested logit (NL)model, Generalized Extreme Value (GEV)Model and Mixed logit model. Since the choices among taxi, FHVs and shared FHVs are highly correlated with each other, we cannot use(Heiss, 2002) a simple multinomial logit model where all the modes are assumed to be uncorrelated. Also, using a mixed logit model gives too much flexibility as it allows all sorts of correlations but in reality, we just need the model to account for correlation between the taxis and FHVs. So, using a mixed logit is a bit of an overkill. In the nested multinomial logit model, the modes in different nests are assumed to be uncorrelated but, inside the nest, correlation is permissible and so, it seemed like a good fit for what we wanted our model structure to look like. GEV is similar to the nested logit in some aspects and there was no particular motive behind not using it, but with the modes chosen by us, it seemed that the nested logit would be a better option.

Data

The data used in this study can be divided into 2 categories based on their purposes: the first category includes population characteristics data acquired from LEHD and ACS to provide a proxy for the origin-destination commute demand, commute transportation preferences, and income distribution; the second category includes transportation trips data acquired from TLC and Uber API which used to train the model, and Here Map data used to estimate utility of transportation modes considered in the model. We have aggregated all these data sources to the TLC taxi zone level and the final version of the data which is used in the model includes pickup and drop-off locations, commute duration, price, and the wage distribution for that origin-destination pair. A detailed explanation of the data is in Appendix .

Methodology

In order to quantify the impact of the Manhattan Congestion Pricing policy, we have compared the simulated ridership under the new pricing scenario against the simulated baseline scenario. For that purpose, we have first used the LEHD/LODES as a proxy for urban mobility, then trained a model based on choices reported in ACS and actual proportions of taxi, FHV and shared FHV observed in the data, and finally compared the simulated ridership to the prior observed preferences to assess the predicted modal shifts.

Specifically, we have simulated the outgoing commute flow in each taxi zone. The simulation especially studies 6:00am-10:am trip since this period shows high correlation between demand and actual trip amount, and the outgoing commute flow is distributed by destination and wealth (which influence the trade-off

between time and cost of different transportation alternatives). The transportation choice proportions (taxi, FHV, shared FHV, public transportation, and walking) for each origin-destination pair will be predicted using the nested multinomial logit model (NMNL). Our transportation choice model is a two-level nesting structure. 5 transportation alternatives (numbered as $j = 1, 2, 3, 4, 5$) are partitioned into 3 nests N_1 (taxi, FHV, shared FHV), N_2 (Public Transit) and N_3 (walking). The utility score for alternative j is

$$U_j = -\lambda(T_j W + P_j) + \varepsilon_j \tag{1}$$

where T_j and P_j refer to the time taken and the cost for choosing alternative j between the pair of taxi zones in consideration, W is the hourly wage of the commuter which we sample from the wage distribution of the origin taxi zone. Here, the parameter λ determines the shift between each alternative. As opposed to the classical use of utility functions where we are trying to increase its value, we want the alternative which produces the lowest utility score. That is why we have given a negative co-efficient to the deterministic part of the utility $V_j = -\lambda(T_j W + P_j)$. The deterministic part V_j has been used to calculate the probability of each outcome $P(y = j)$; where $y \in 1, 2, 3, 4, 5$ indicates the chosen mode. Within a nest k, the degree of dissimilarity between the transportation alternatives is indicated by the dissimilarity parameter τ_k. The marginal probability of the outcome j is

$$P(y = j) = \frac{e^{\frac{1}{\tau_k} V_j}}{e^{IV_k}} \cdot \frac{e^{\tau_k V_j}}{\sum_m e^{\tau_m IV_m}} \tag{2}$$

with the inclusive value IV_k defined as

$$IV_k = \ln \sum_{l \in N_k} e^{\frac{1}{\tau_k} V_l} \tag{3}$$

The probability for each origin-destination pair for each transportation mode has then be multiplied by the commute population in that OD pair, and has been aggregated by origin. We have used the American Community Survey (ACS) data as the ground-truth on the relative distribution of population's choices to choose the model with the least root mean squared error (RMSE). To find the ideal parameters that result in the model achieving the lowest RMSE, we have performed a Monte Carlo simulation which gave us the values for λ and τ to use in our 3 scenarios. Finally, we have ran the model with these parameters for three scenarios – normal price without congestion, increase in price by \$2.75 (current scenario), and increase in price by \$10 (Surcharge proposed to be implemented in 2020) and compared the outcomes of these.

Risks and Mitigation

One of the risks and limitations in methodology is the inaccuracy of data. We are using the ACS and LODES data as a proxy for transportation demand which does not represent the entire population in an unbiased manner. For example, the ACS data (as well as LEHD) is outdated and does not provide choices between taxi/FHV/shared FHV. This can be mitigated by calculating the proportion from actual data and adjusting the overall fraction of taxi choices by the ratio of total taxi ridership in the year of ACS and taxi+FHV ridership now. Other factors such as the regular commuters are not the entire population or other uncertainties may cause risks such as lack of generalizability or lack of statistical significance of resulting impact. We can try to translate the uncertainty into the model estimates and consider additional data sources that could provide prior estimates for the model parameters and employ Bayesian inference framework to incorporate this information reducing the uncertainty of posterior estimates. Also, we will try to implement a further iteration of the model(time allowing) that will will try to infer the transportation demand based on the actual partially observed ridership instead of focusing on commuters only.

Initial Data Exploration

Taxi & LEHD Data Correlation

In order to understand how good of a transportation demand proxy LEHD is , we explored the 2015 LEHD data together with the 2015 taxi data to find the best correlated hours between the LEHD demand the actual taxi ridership. We aggregated the datasets by each pair of pickup and drop-off taxi zone, and by each hour of the day. We calculated the correlations of the hourly trip counts between the total LODES count, and its correlations between the "expected" taxi ridership (total LODES count multiplying by the percentage of census taxi commuters).

The results show the taxi trips during morning hours (5am-10am) is more correlated with both of the expected taxi ridership (Figure 10) and the total LODES count (Figure 11) compared to other time of the day. A comparison of the two figures reveals that the correlation between actual taxi ridership and the "expected" taxi ridership is much stronger than the correlation between actual ridership and the total LEHD demand. Such results verify our assumption that the morning hours trips are more correlated with the LEHD demands, particularly when paired with ACS data about transportation behaviors. Thus the pickup (origin) taxi zones during the morning hours should be more representative of people's actual residential locations, allowing us to infer information information about those travelers based on ACS data (i.e. income distribution). To accurately utilize regional demographic information in our model, we decided to scope our analyzes within the morning hours time frame.

Taxi Data exploration analysis

To better understand the distribution of trips and to select the appropriate proportion of data for our model use, we analyzed the ridership trends of Taxi by aggregating the total passengers, total rides, average fare and average trip duration for each hour of the day. Results show that the morning rush hours (6–10 am) and evening rush hours (5–7pm) usually has more total rides and total passengers in the weekdays (Figure 12). This trend is less obvious during weekends, however, on the weekends there is a noticeable increase in the midnight trips (Figure 13). Based on the mean taxi fare and mean trip duration, we found the trips around 4 pm tend to be longer and more expensive based on the average fare and average trip duration (Figure 14). We also examined the trips distribution for high passenger volume taxi zones, including Times Sq/Theatre District, Penn Station/Madison Square, Lenox Hill West, Lincoln Square, Upper West Side, and Upper East Side. We found zones that near Central Park (e.g, Lenox Hill, Upper West, and Upper East Site 0have the high passenger's volume all day from 7 am 7 pm, zones such as times Sq/Theatre District, and Penn Station have high volume during evening hours (Figure 15).

FHV Data exploration analysis

To better understand the distribution of FHV and shared FHV trips, we analyzed the ridership trends of For-Hire vehicles by aggregating the total ride counts and the average trip durations for each hour of day and day of the week. Similar to taxi, the hourly rides for both shared FHV and total FHV are both higher during the morning and evening rush hours (Figure 16), but the evening rush hours have slightly more trips compare to morning hours. Trips in the afternoon (3 -4 pm) have longer durations, and trips during midnight are relatively shorter (Figure 17) based on the average trip durations. By comparing the trips count in days of the week, we found Fridays and Saturday have more total trips (Figure 18), and trips during weekends are shorter than the trips during weekdays(Figure 19). Finally, the choropleth maps (Figure 20)show the normalized rides count for each Pick up and drop off Taxi Zones (counts normalized by the area of zones), which we found the zones with high trip amounts(darker zones) are mostly concentrated at lower/middle Manhattan and downtown Brooklyn areas for both Pick up and Drop off locations. Such finding suggests that a large portion of our of our model simulations will be reflecting the trips in these areas, thus we need to be more carefully consider the regional demographic information of these areas as

well as their functionalities (e.g. shopping, parks, etc), to avoid any false assumptions when profiling people choices.

For comparison, the distribution of trip duration using public transportation (subway&bus) is also shown in Figure 21. The average commute time is 4530 seconds (roughly 75 mins).

Initial Model Results

The ultimate goal of this research is to apply our transportation mode shift model in order to quantitatively evaluate the anticipated impacts of Manhattan Congestion Pricing. For our initial evaluation of this impact, we applied our model three pricing scenarios for taxis and FHVs (nest 1) , each reflecting a Congestion Pricing policy scenario:

1. Status quo (no tax added to price)

2. Existing Congestion Pricing tax ($2.75 added to nest 1 trips passing through the Congestion Zone, as trip data was collected before the existing policy went into effect)

3. Proposed Congestion Pricing tax ($10 added to nest 1 trips passing through the Congestion Zone)

We then analyzed the model outputs to determine the anticipated effect of each policy on the NYC transportation as a whole (i.e. total shift in public transportation ridership, walking, and taxi/FHV ridership) as well as the anticipated effect on particular taxi zones.

This table shows the shift in NYC wide transportation choices projected for each policy scenario. Our model predict a reduction in Taxi and (non-shared) FHV trips under both Congestion Pricing Scenarios. We predicted a 1.35% drop (about 39,000 trips) in Taxi or FHV trips under the $2.75 tax and a 4.31% drop (about 124,000 trips) under the $10 tax. The model predicts an increase in Shared FHV trips, Public Transportation trips, and Walking trips.

Scenario	Taxi	FHV	Shared FHV	Public Transportation	Walk
Status Quo Model	16.36	13.18	11.08	46.69	12.69
$2.75 Tax Model	15.59	12.61	11.42	47.54	12.85
$10 Tax Model	13.69	11.54	12.24	49.36	13.17
Ground Truth (ACS Reported Choice)	0.37	0.67	0.26	85.01	13.69

Table 1: This table shows a distribution view of the results of our model simulations, along with our "ground truth" - the commuting transportation mode choices reported in the American Community Survey (ACS). Numbers represent the proportion of total NYC commute trips for each mode predicted under each scenario. Substantial differences from ground truth numbers are discussion in the "Limitations" section of this paper.

Of the 49 origin taxi zones that the model predicts will experience the most mode shift under the $10 Congestion Tax (predicted change is more than 5.42%, which is one standard deviation above the mean predicted change), 45 are in Manhattan and four are in Brooklyn. Figure 1 shades origin taxi zones by the predicted percent decrease in total taxi and FHV ridership under the $10 Congestion Tax scenario.

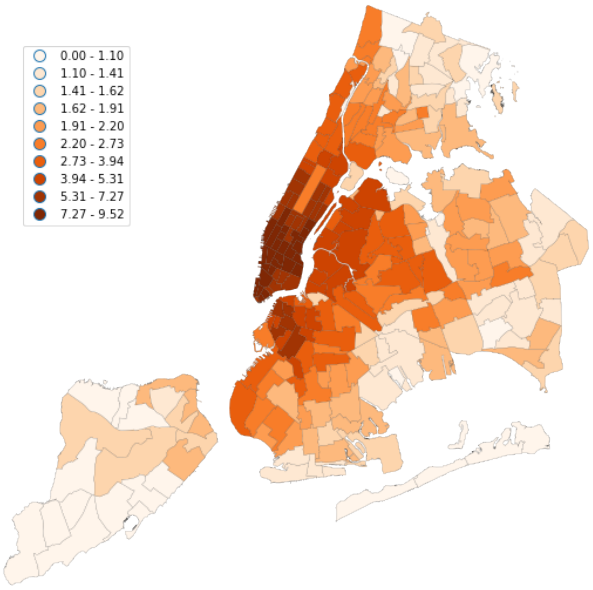

Figure 1: Origin taxi zones are shaded by the predicted percent decrease in total taxi and FHV ridership under the $10 Congestion Tax scenario. Manhattan taxi zones in or adjacent to the congestion zone are predicted to experience a large shift. Commuter in taxi zones in Brooklyn Heights, Carroll Gardens, Downtown Brooklyn and Park Slope are also predicted to decrease taxi and FHV ridership by over 5%.

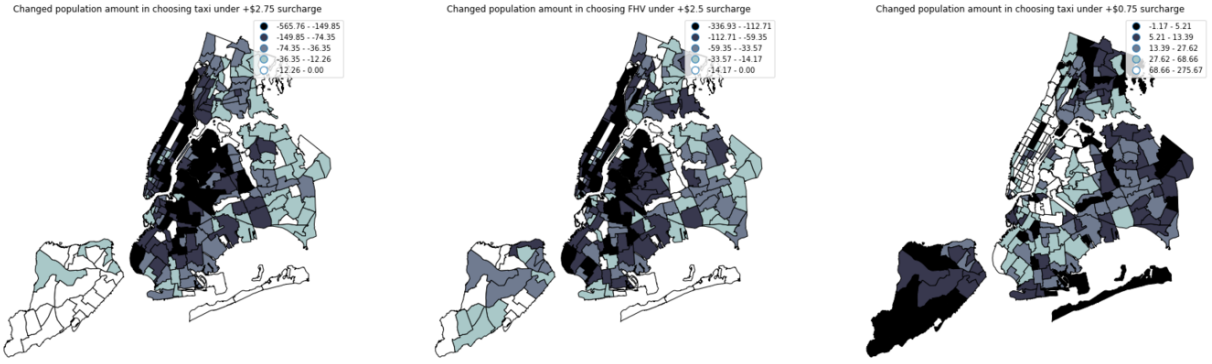

Figure 2: Visualize Changed population in choosing taxi, FHV, and shared FHV under current Manhattan Congestion Surcharge

Figure 2 plots the number of people changed in choosing taxi, FHV, and shared FHV under current Manhattan Surcharge Policy. The darker area indicates more decrease in the number of people, while the lighter area indicates less decrease or an increase in the number of people. The number of people choosing both taxi and FHV showed obviously decreasing (especially in Manhattan area) under surcharge policy, while the number of people choosing shared FHV showed less decreasing and even slightly increasing. This is aligned with our assumption since people tend to avoid too cost methods.

In order to evaluate any potential equity or accessibility concerns with the existing or proposed congestion tax, we will next analyze demographic information for the taxi zones most affected by the policies. We will identify the most affected zones as zones that we anticipate will either pay the most in congestion taxes or experience the most anticipated shift in transportation choices. We will also quantify the anticipated tax revenue under each Congestion Tax policy in order to evaluate the possible public benefit of each plan in terms of increased funding for subway repairs.

Limitations

As seen in the results of the model, it overpredicts the counts for the taxis and FHVs by a huge number as compared to the ground truth data (ACS +LEHD). For example, the ground truth says that only 19,278 people would take non-shared For-Hire Vehicles but the model predicts almost 0.4 million people would take it. One reason behind this is that the ACS data is outdated and so, it is unable to give reliable proportions for the taxi and FHV usage during this so-called 'FHV-era'. So, we are doing parameter tuning based on the ground truth while the ground truth itself might not be very representative of the actual situation and so, the predicted results today might differ by a lot from the ACS+LEHD data.

But even as compared to a typical day of taxi ridership in 2017, only 44,000 people took an Uber that day. So, the lack of reliable ground truth data doesn't seem to explain the results by itself. Another reason behind this huge gap could be that the Utility function does not reflect well enough how a person makes a transportation choice. The utility function would work very well in an ideal world where everyone worked out the economics of their commute on a daily basis but, the reality is that people are just set in their ways and won't make a different choice for each day.

Also, regular commuters don't make up the entire population and the proportions of the taxi and FHV usage for commuters and all users might not be the same.

Updated Research Plan

We updated the research plan (Figure 5) based on our progress. The initial nested logit model design and data processing were completed before the submission of this report. The second model iteration with data update will be finished before July 1, and if necessary the third one will be completed before July 8. An initial website has been created. As we continue to improve our model and analyze our results we will simultaneously update the website and work on our final presentation.

Team Collaboration Statement

All the team members will work together. In order to improve the collaboration efficiency and effectiveness, each domain will be assigned to several people to monitor the progress. Specifically, Xiaoning He and Sam Manzi are responsible for data gathering and cleaning; Rufei Sheng and Soham Mody are responsible for modeling and analysis; Katharine Voorhees is responsible for literature review and report writing.

Figures:

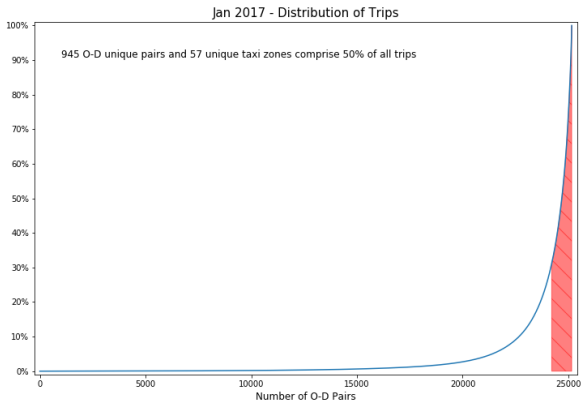

Figure 3: Distribution of taxi trips by origin and destination (taxi zone level). Note that half of the distribution (shaded portion) is comprised of 945 unique O-D pairs and 57 unique taxi zones (see Fig. 2)

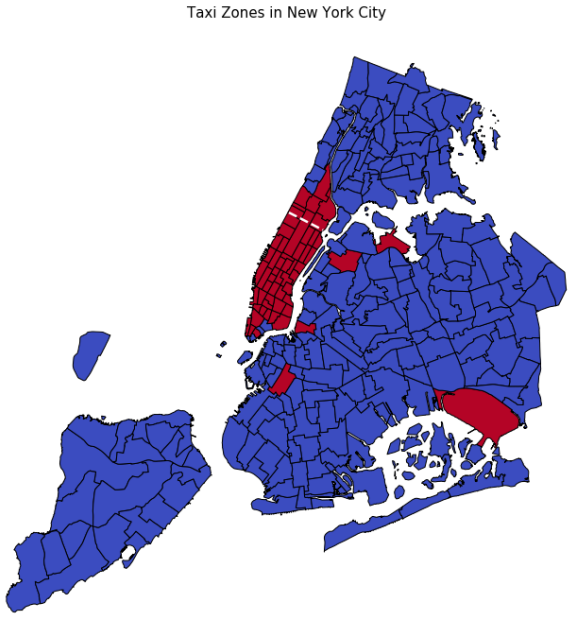

Figure 4: Red taxi zones comprise origins or destinations which constituted 50% of all rides in January of 2017. Zones in Manhattan which lie south of the dashed white line (96th street) are currently subject to congestion pricing.

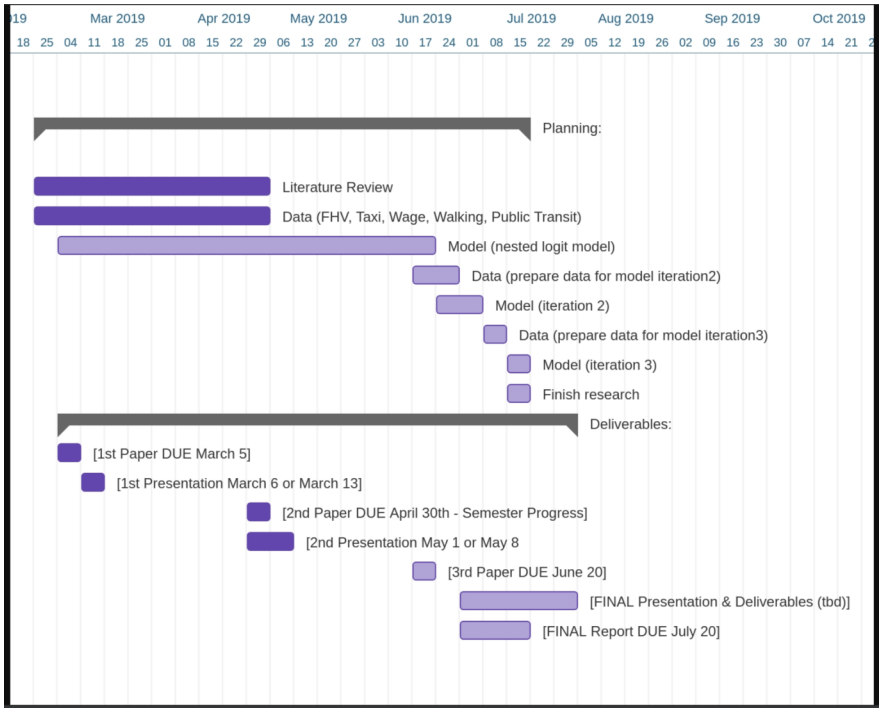

Figure 5: Updated Research Plan

```
FHV.head()
```

	date	PUlocationID	DOlocationID	hour	ride_counts	trip_duation_mean (seconds)	SR_Flag
0	2018-05-01	0.0	0.0	6	4	2865.0	0
1	2018-05-01	0.0	0.0	7	1	1800.0	0
2	2018-05-01	0.0	0.0	8	4	1650.0	0
3	2018-05-01	0.0	0.0	9	3	2400.0	0
4	2018-05-01	0.0	0.0	10	6	3750.0	0

```
FHV_shared.head()
```

	date	PUlocationID	DOlocationID	hour	ride_counts	trip_duation_mean (seconds)	SR_Flag
0	2018-05-01	1.0	142.0	7	1	6144.0	1
1	2018-05-01	3.0	3.0	6	1	180.0	1
2	2018-05-01	3.0	3.0	9	1	92.0	1
3	2018-05-01	3.0	18.0	7	1	1415.0	1
4	2018-05-01	3.0	18.0	8	1	2334.0	1

Figure 6: Samples of For-Hire Vehicles and Shared For-Hire Vehicles data during 5-10 am aggregated on the taxi zone level

```
uber_price.head()
```

	LocationID_pickup	lon_pickup	lat_pickup	LocationID_dropoff	lon_dropoff	lat_dropoff	Pool_Price	Non_shared_Price
1	1	-74.176778	40.689515	2	-73.826141	40.625724	100.0	107.0
2	1	-74.176778	40.689515	3	-73.849479	40.865871	75.5	77.5
3	1	-74.176778	40.689515	4	-73.977024	40.724151	52.0	57.0
4	1	-74.176778	40.689515	5	-74.189938	40.550339	44.5	45.0
5	1	-74.176778	40.689515	6	-74.067786	40.599053	43.5	46.0

Figure 7: Uber pricing data during 8-9 am aggregated on the taxi zone level for shared and non-shared For-Hire Vehicles

duration	price	ODpair	2500	7500	12500	17500	22500	30000	
6.508943	16.647073	3-3	11.412271	16.389811	29.504545	12.840609	18.871367	26.936423	30.5
39.695000	64.000000	3-4	0.250819	0.360216	0.648452	0.282211	0.414755	0.592009	0.6
45.216667	61.500000	3-4	0.250819	0.360216	0.648452	0.282211	0.414755	0.592009	0.6
83.000000	5.500000	3-4	0.250819	0.360216	0.648452	0.282211	0.414755	0.592009	0.6
225.933333	0.000000	3-4	0.250819	0.360216	0.648452	0.282211	0.414755	0.592009	0.6
47.880952	43.157143	3-7	1.504915	2.161294	3.890709	1.693267	2.488532	3.552056	4.0
30.521739	47.000000	3-7	1.504915	2.161294	3.890709	1.693267	2.488532	3.552056	4.0
37.159009	45.000000	3-7	1.504915	2.161294	3.890709	1.693267	2.488532	3.552056	4.0

Figure 8: Aggregated Data (Public Transit mode example)

	taxi_zone	P(mode1)	P(mode2)	P(mode3)	P(mode4)	P(mode5)
0	3	177.762595	73.068204	1.150001	16668.822855	2001.196346
1	4	201.297419	120.885755	61.223127	13318.123341	6279.470358
2	5	49.601278	7.992997	0.306749	8847.276451	142.822526
3	6	164.390442	40.427858	2.346380	22929.859149	2485.976170
4	7	616.150103	162.371679	110.267353	138012.557541	11998.653324

Figure 9: American Community Survey (ACS) as ground-truth, indicating the population for each transportation model based on each origin taxi zone.

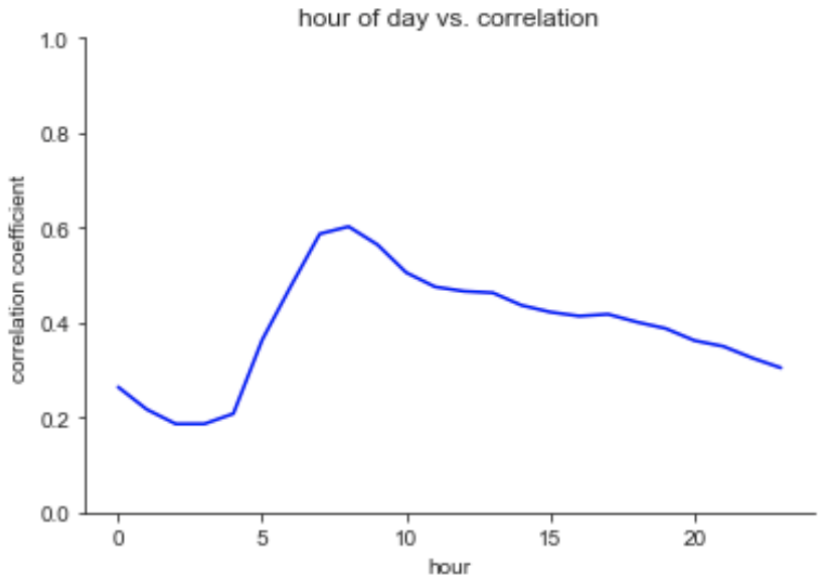

Figure 10: Hourly correlations between actual taxi trip counts for OD pairs and the "expected" taxi ridership for the same pair. "Expected taxi ridership" was calculated as the ratio of people for each OD pair that take a taxi to work on most days (as reported by the ACS) multiplied by the count of commuters for the pair (reported by the LODES data).

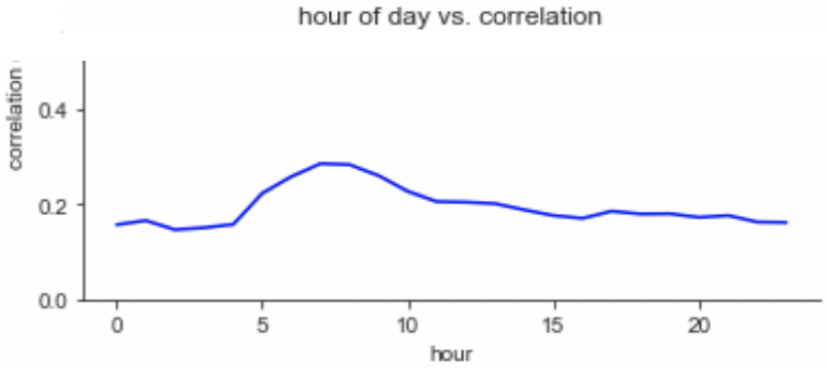

Figure 11: Hourly correlations between actual taxi trip counts for OD pairs and the total LODES count for those pairs.

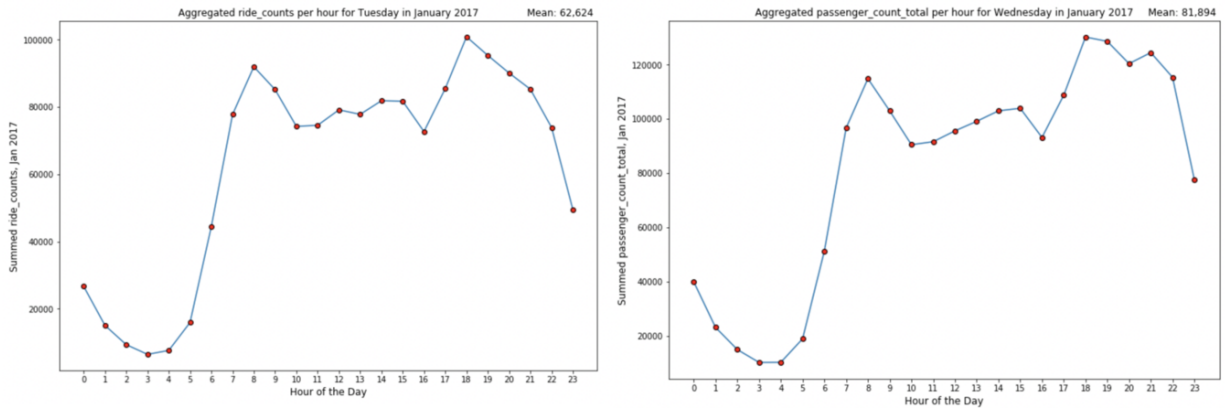

Figure 12: Aggregated Taxi ride counts and passenger counts on weekdays.

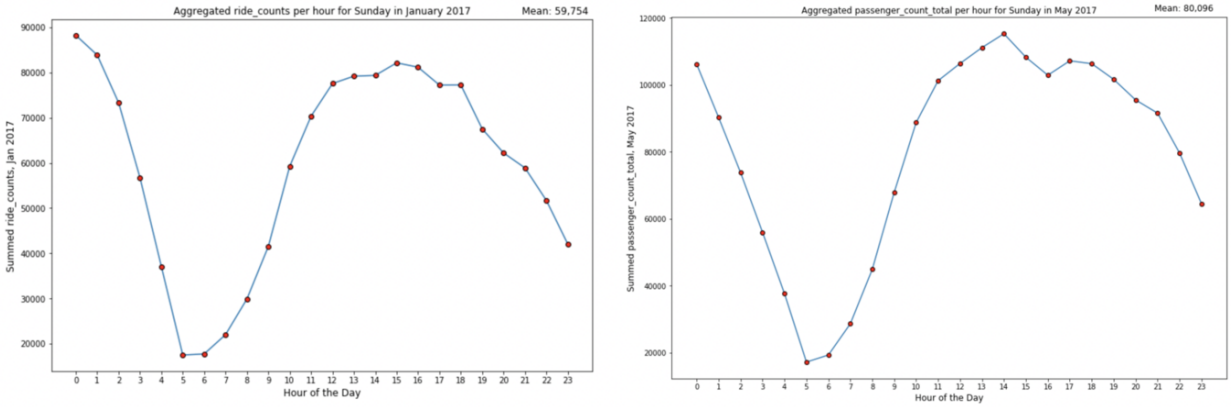

Figure 13: Aggregated Taxi ride counts and passenger counts on weekends.

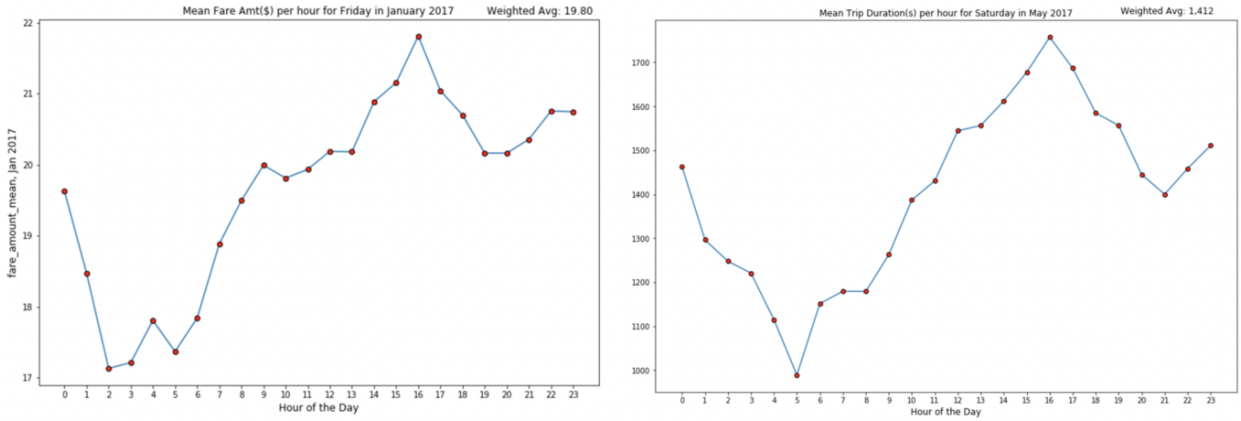

Figure 14: Mean Taxi Fare and Mean Trip Durations per hour.

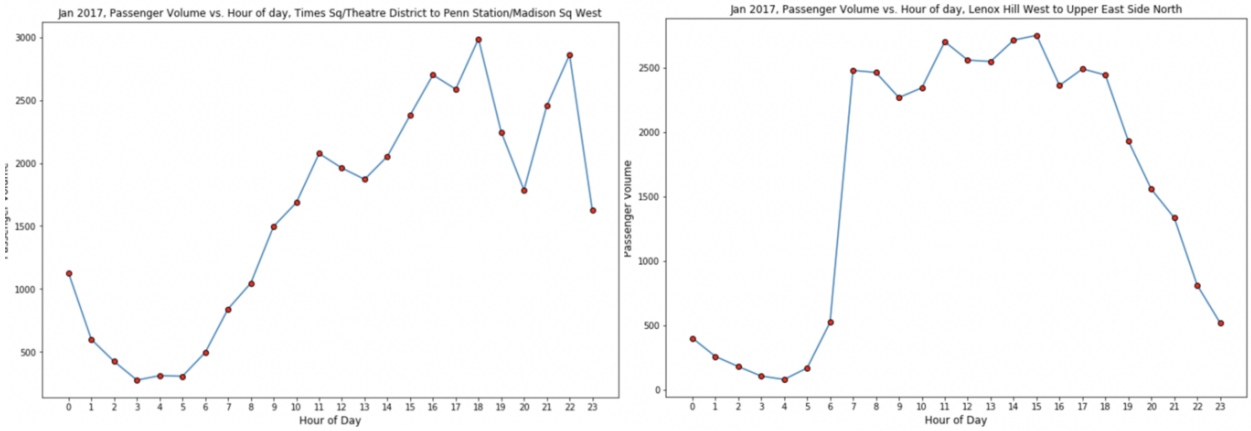

Figure 15: Trips distribution for high passenger volume taxi zones in Jan 2017.

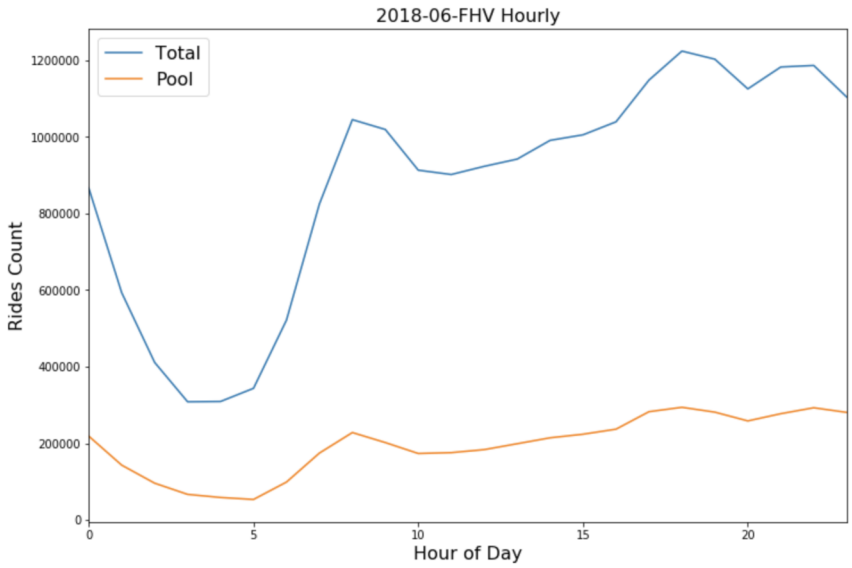

Figure 16: Hourly FHV Trips counts

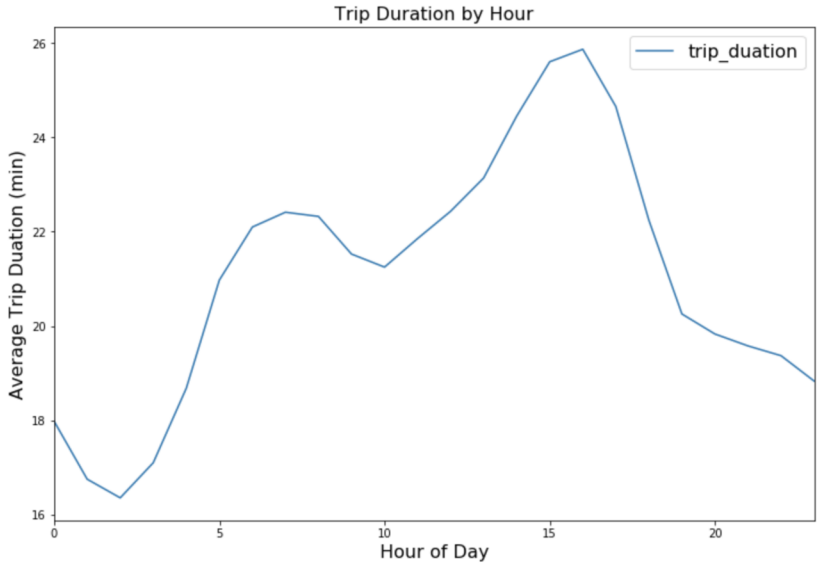

Figure 17: FHV Hourly average Trip Duration

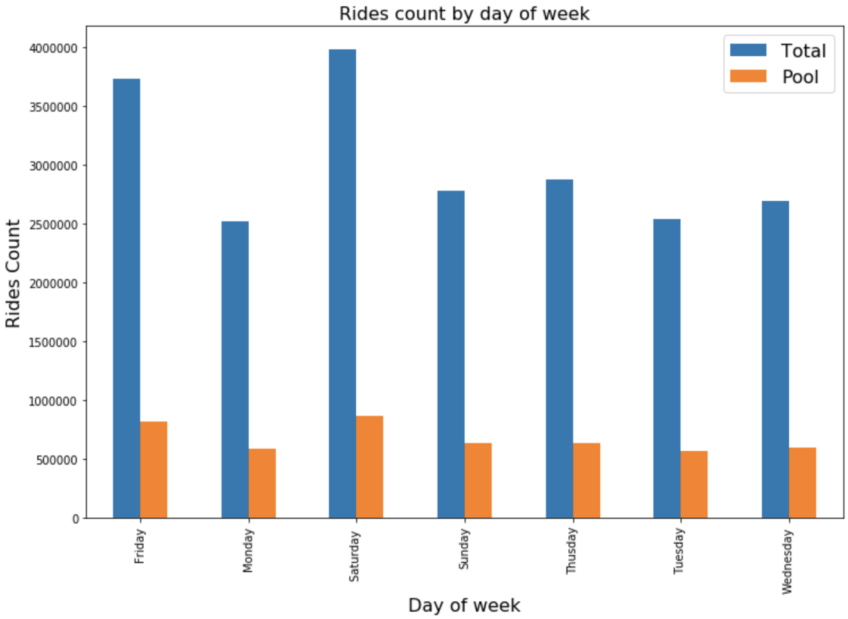

Figure 18: FHV Trips counts by day of week

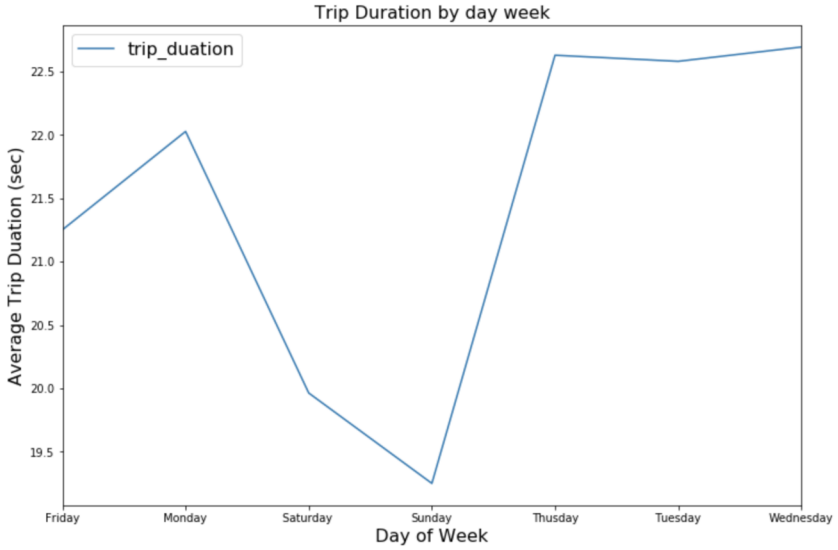

Figure 19: FHV average Trip Duration by day of week

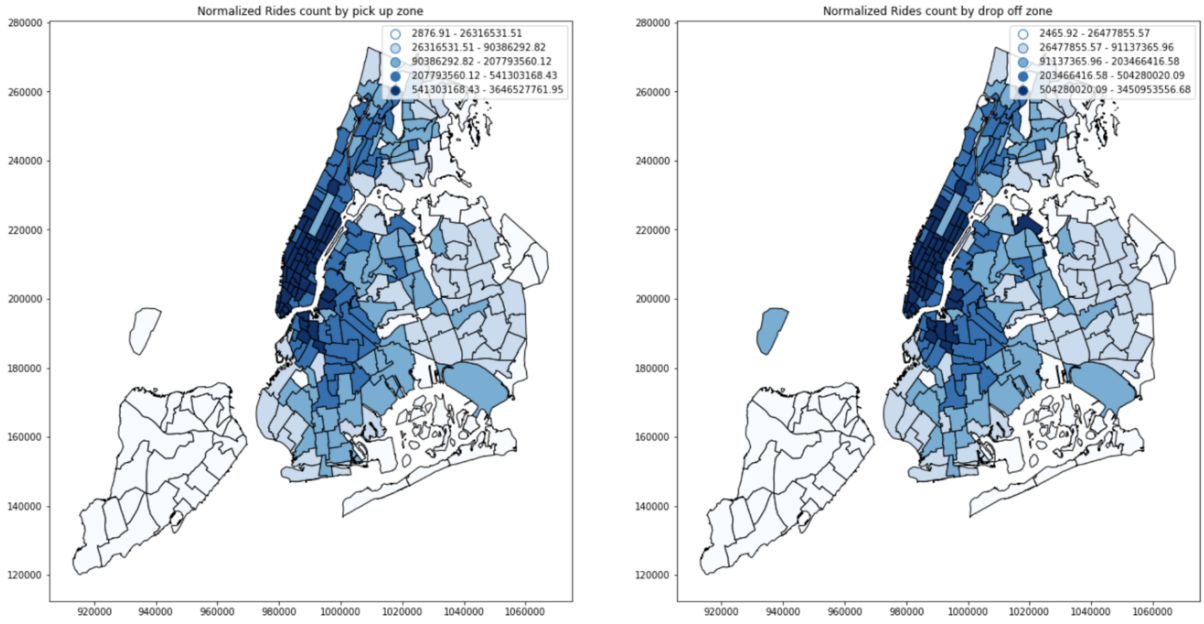

Figure 20: Pick up and drop off density per taxi zone

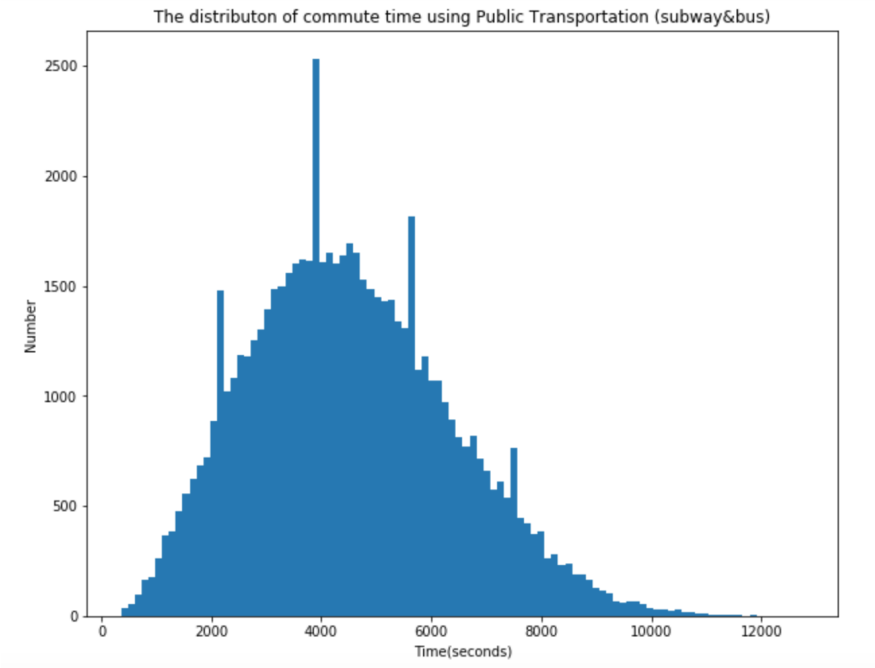

Figure 21: Distribution of trips by commute time

Appendix:

Data:

American Community Survey

American Community Survey (ACS) data was used to reflect reported choices of transportation modes by commuters serving as a ground truth for fitting the model. Census tract level 5-year estimate data was pulled through the Census API in order to generate probabilities of a commuter within each taxi zone of having wages within each Census income bracket. Because the Census provides bucketed rather than continuous income information, the middle of the bracket range was attached to all individuals in that bracket. Commute information from the ACS was used to estimate the percentage of people in each taxi zone that regularly choose each form of transportation for trips to work. Respondents were asked to report which form of transportation they used "most days" for commuting to work. Collectively the ACS data allows us to estimate the probability of any given resident of each origin zone choosing each distinct mode of transportation, as well as predicting that commuter's income (based on the income distribution of the zone).

Trip Data

Taxi

New York City's Taxi and Limousine Commission provides free access to their database of taxi trip data with ride-level granularity. A typical month of data consists of 10 to 15 million rides. For analysis, we have aggregated at the level of origin-destination, hour of day and day of week; we consider mean passenger count, fare amount, tip amount, trip duration, and total ride volume and passenger volume. These data give a sense of the high volume traffic areas in the city, as well as the distribution of trips by time of day. They are crucial to estimate current modal distribution and, when used in conjunction with demographic ACS data, to predict mode shift under various pricing scenarios and within varying demographics.

Preliminary analysis of taxi ride volume in January 2017 shows that 57 taxi zones account for half of all trips (˜4.5 million) (Figure 3). Furthermore, 52 of these taxi zones are in Manhattan, 45 of which are aresubject to congestion pricing (Figure 4). Thus, yellow cab drivers and yellow cab riders are obvious stakeholders when it comes to congestion surcharge policy.

For-Hire-Vehicles (FHV)

For-Hire-Vehicle trip data are provided by TLC consists of individual trip data for different FHV services (ex. Uber, Lyft, etc.). A typical month of data includes 15 to 20 million rides, with around 20% of shared FHV, and 80% Non-shared FHV. For analysis, we have separated the FHV trips to FHV and shared FHV , and aggregated both data at the level of pickup and Drop-off zone, date and hours from 5 am to 10 am. The aggregated data contains attributes of date, pickup location id, drop-off location id, average trip duration (sec), trip counts, and SR_Flag(FHV and shared FHV.) Example is as shown in Figure 6.

Uber API

Uber API provides live pricing of Uber rides at the time of the request. Uber has released a Python library called uber-rides which can be used to request the ride information using a server token private to each user. We need the price of shared and non-shared rides in the time interval of our interest which is 8-9 am. But, there is a limit of 2000 requests per hour for a server token and we needed the information for about 70,000 trips (all combinations of the 263 taxi zones). So, we are performing the data curation over several days at the same time using datetime and Timer functions of Python to schedule this process and also looking into other creative methods to approximate this price instead of getting it for each combination. The ideal thing

would be to use this price from the actual trips which took place as we are using the time taken for the trip from that data. But since that data is not available, this data is collected at the same time(8-9 am) and it gives a good approximation. The curated data on the taxi zone level is shown in Figure 7.

HERE Maps

HERE technology is the company that provides mapping and location data. In order to assess travel time, cost and overall utility of each transportation mode considered in the model given the OD pair, we use HERE REST APIs to gather information such as maps, routing, geocoding, places, positioning, traffic, transit, and weather information. The public transit data can be acquired via specific Public Transit API. Use HTTP GET methods, route information such as the trip time duration, the number of transfer, and the mode for each transfer will be get given departure and arrive location, departure time, and specific mode. The public transit data used in our model is acquired by specifying "mode = subway & bus", while the walking related data using "mode = walking".

Aggregated data

Spatial Join

In order to apply data from the American Community Survey (ACS) and LEHD to this research Census geographies needed to be spatially joined with taxi zones so that we could aggregate the data at the taxi zone level. ACS data was merged with a shapefile of census tract population centroids accessed through NYU's Spatial Data Repository (Repository, 2010), and then spatially joined in GeoPandas to a taxi zone shapefile (Taxi & (TLC), 2019). Similarly, LEHD data was merged with with a shapefile of Census blocks and then spatially joined to the taxi zone shapefile for aggregation.

Modeling dataset

After data processing and aggregation, the final version data for modeling includes pickup and drop-off locations (aggregated on the taxi zone level), commute duration, price, wage distribution. Example is shown in Figure 8.

Methodology:

The probability of mode j given nest k, where j belongs to nest k:

$$P(y = j|y \in N_k) = \frac{e^{\frac{1}{\tau_k}V_j}}{e^{IV_k}} \tag{4}$$

The probability of nest k:

$$P(y \in N_k) = \frac{e^{\tau_k V_j}}{\sum_m e^{\tau_m IV_m}} \tag{5}$$

Conditional probability for mode j, which equals to the probability of mode j given nest k where j belongs to nest k, mutiplies the probability of nest k.

$$P(y = j) = P(y \in N_k) \cdot P(y = j|y \in N_k) \tag{6}$$

References

Assembly, N. Y. S. (2019). *BILL NO A02009C*. Retrieved from https://www.nyassembly.gov/leg/?default_fld=&leg_video=&bn=A02009&term=2019&Summary=Y&Actions=Y&Committee%26nbspVotes=Y&Floor%26nbspVotes=Y&Memo=Y&Text=Y&LFIN=Y

Beria, P. (2016). Effectiveness and monetary impact of Milan's road charge, one year after implementation. *International Journal of Sustainable Transportation*, *Vol. 10*.

Capstone. (2018). Prediction of Mode Shift in New York City. Retrieved from https://cuspcapstones.github.io/Prediction-of-Mode-Shift-in-Cities-Based-on-Trip-Cost-Duration_2018/

DOT. (2018). *New York City Mobility Report*. NYC Department of Transportation. Retrieved from https://www1.nyc.gov/html/dot/downloads/pdf/mobility-report-2018-print.pdf

Eliasson, J., Hultkrantz, L., Nerhagen, L., & Rosqvist, L. S. (2009). The Stockholm congestion – charging trial 2006: Overview of effects. *Transportation Research Part A: Policy and Research*, *Vol. 43*.

Gibson, M., & Carnovale, M. (2015-2016). The effects of road pricing on driver behavior and air pollution. *Department of Economics, Williams College*.

Heiss, F. (2002). *Specification(s) of Nested Logit Models* (MEA discussion paper series No. 02016). Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy. Retrieved from https://ideas.repec.org/p/mea/meawpa/02016.html

Lehe, L. (2019). Downtown congestion pricing in practice. *Transportation Research Part C: Emerging Technologies*, *Vol. 100*.

MTA. (2017). Introduction to Subway Ridership. *MTA.info*. Retrieved from http://web.mta.info/nyct/facts/ridership/

Pearce, A. (2018). How 2 M.T.A. Decisions Pushed the Subway Into Crisis. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2018/05/09/nyregion/subway-crisis-mta-decisions-signals-rules.html

Ramming, & Scott, M. (2002). *Network knowledge and route choice* (PhD thesis). Massachusetts Institute of Technology. Retrieved from https://dspace.mit.edu/handle/1721.1/49797

Repository, N. Y. U. S. D. (2010). *New York City Census Tract Population Centers*. Retrieved from https://geo.nyu.edu/catalog/nyu-2451-34504

Sobolevsky, S., Levitskaya, E., Chan, H., Postle, M., & Kontokosta, C. (2018). Impact Of Bike Sharing In New York City. Retrieved from https://arxiv.org/abs/1808.06606v1

Taxi, & (TLC), L. C. (2019). *NYC Taxi Zones*. NYC Open Data. Retrieved from https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc

Tong, Y. (2015). How New Yorkers Prefer to Take Public Transport? A Comprehensive Analysis Based on 2010-2011 Regional Household Travel Survey.

Train, K. (1977). A validation test of a disaggregate mode choice model. *ScienceDirect*, *Vol. 12*. Retrieved from https://www.sciencedirect.com/science/article/pii/004116477890120X

Yu, L., & Sun, B. (2012). Four types of typical discrete Choice Models: Which are you using? - IEEE Conference Publication. *IEEE*. Retrieved from https://ieeexplore.ieee.org/document/6273550

of Taxation, N. Y. S. D., & Finance. (2018). *Technical Memorandum TSB-M-18(1)CS*. Retrieved from https://www.tax.ny.gov/pdf/memos/cs/m18-1cs.pdf