


2019

# Investigating the effect crime has on Uber and Yellow Taxi pickups in NYC

Kebing Li  
Colby College

Follow this and additional works at: <https://digitalcommons.colby.edu/honorstheses>

 Part of the [Econometrics Commons](#), [Economic Theory Commons](#), [Industrial Organization Commons](#), [Labor Economics Commons](#), and the [Regional Economics Commons](#)

Colby College theses are protected by copyright. They may be viewed or downloaded from this site for the purposes of research and scholarship. Reproduction or distribution for commercial purposes is prohibited without written permission of the author.

## Recommended Citation

Li, Kebing, "Investigating the effect crime has on Uber and Yellow Taxi pickups in NYC" (2019).  
*Honors Theses*. Paper 924.  
<https://digitalcommons.colby.edu/honorstheses/924>

This Honors Thesis (Open Access) is brought to you for free and open access by the Student Research at Digital Commons @ Colby. It has been accepted for inclusion in Honors Theses by an authorized administrator of Digital Commons @ Colby. For more information, please contact [mfkelly@colby.edu](mailto:mfkelly@colby.edu).

**Investigating the effect crime has on Uber and Yellow Taxi pickups in NYC**

Kebing Li

Honors Thesis

Economics Department, Colby College

May, 2019

## **Abstract**

**How to manage the relationship between Uber and the local taxi industry has been a long-lasting and hot topic for most of the major cities around the world. Whether Uber is stealing money from and undermining the local taxi drivers or it is beneficial for public transportation has no certain conclusions. In this paper, we focus on the city of New York, where both Uber and traditional Yellow Taxi play important roles in public transportation and city culture in general, and we are trying to investigate the factors that are going to affect Uber and Yellow Taxi pickups in New York City. Among many socio-economic factors, we especially want to see what role crime would play in this setting since Uber has claimed that unlike yellow taxi drivers, Uber drivers do not have geographical discrimination based on the number of crime occurs in a given neighborhood. We use the data from April to September in 2014 for both Uber and Yellow Taxi pickups, and socio-economic data in 2014 by census tract to develop several econometric models treating each census tract as an individual observation.**

## 1. Introduction

On August 8<sup>th</sup>, 2018, New York City became the first US city to pass legislation restricting the number of ride-hailing vehicles operating on its roads.<sup>1</sup> In addition to limiting the number of ride-hailing services, the city's Taxi & Limousine Commission can also set minimum pay standards for drivers of ride-hailing services to make up the difference between the pay floor and a taxi driver's hourly earnings, in an attempt to rebalance the supply side of the taxi industry. These new series of regulations, therefore, have become the most recent addition by the NYC government to the hot and long-lasting debate in nearly all major cities around the world with respect to ride-sharing services. The concern is that ride-hailing services like Uber and Lyft may affect the balance of the public transportation system, something often heavily regulated.

Although the passing of new legislation seems to mitigate some concerns, both the government and the corporate sides are still closely looking at this issue and thinking about their next steps. It is clear that both are able to list several pros and cons of having massive ride-hailing vehicles on the roads, but neither of them actually has a clear view on how to strike a balance between the traditional New York City Yellow Cab and the new, often

---

1

<https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3479666&GUID=01C67FF7-C56D-474A-BA53-E83A23173FA7&Options=ID%7cText%7c&Search=838>

consumer-favored ride-hailing vehicles. Therefore, the space leaves for us, as researchers, to investigate is plenty.

On the government side, Uber and other ride-hailing vehicle services are extremely hard to control since their drivers serve more like contractors rather than employees, which allows their parent companies, such as Uber and Lyft, to differ from traditional taxi service companies in things like insurance coverage or their willingness to take responsibility for their drivers' behaviors. The explosion in the number of ride-hailing drivers also has several negative impacts on the local economy. According to the Wall Street Journal, the number of app-based, for-hire vehicles jumped from 25,000 in 2015 to over 80,000 in 2018.<sup>2</sup> The oversupply of the ride-hailing services led to more than 40% of cars traversing the city without having any passengers, making the already-awful Manhattan congestion even worse.<sup>3</sup> In addition to adding more pressure to the city government, the extremely imbalanced supply and demand structure in taxi markets resulted in huge decreases of disposable income for traditional taxi drivers, who had to face serious competition from their competitors offering cheaper and often more convenient services. By report, six taxi drivers committed suicide within seven months because of their economic

---

<sup>2</sup> <https://www.wsj.com/articles/new-york-city-council-votes-to-cap-uber-and-lyft-1533759263>

<sup>3</sup> <https://www.bloomberg.com/news/articles/2018-08-20/why-tensions-between-uber-and-cities-peaked-in-nyc-quicktake>

desperation and financial pressure, creating a large-scale public outcry that spurred the city council to act on this issue.<sup>4</sup>

On the other side of the story, however, large ride-hailing firms, argue that besides offering cheaper and convenient public transportation to the consumers and providing more employment opportunities, they are especially beneficial for people in those “illed-served” areas. Even though the firm didn’t explicitly explain the meaning of the term “illed-served”, people interpreted it as areas typically associated with higher crime rates, lower income levels and lower educational level. As the core and strongest argument provided by Uber, if it can be proved valid, then it would be hard for the regulators to implement further and stricter regulations since the ride-hailing service firms could argue that their contributions help against regional discrimination. This consideration is the main motivation of this paper, which aims to evaluate the validity of Uber’s argument. Hence, we use data from 2014 to study whether regional differences in NYC can actually affect Uber and Yellow Taxi pickups.

## **2. Previous Literature**

The Uber-related problems, or in general, the ride-hailing-services-related problems have rarely been studied in academia environment not only because this is a recently developed

---

4

<https://www.nbcnews.com/news/us-news/sixth-new-york-city-cab-driver-dies-suicide-after-struggling-n883886>

industry but also because most ride-hailing service providers had a history of trying to maneuver around regulations and therefore did not disclose their data. In recent years, as the Uber-NYC tensions mounted, the government has been requesting the release of data from Uber, but only very limited data sources were available to the general public. One study conducted by the Office of Mayor of NYC in 2016 investigated the relationship between ride-hailing service companies and other public transportation offered in NYC and forecasted the future of those for-hire vehicles in NYC.<sup>5</sup> However, looking back then, we could see a huge underestimation of the growth rate of the for-hire vehicle industry and an overly optimistic attitude towards its impact on the city. In another research, Schwieterman and Livingston evaluated the monetary and nonmonetary tradeoffs of Transport Network Companies and Transit Services in Chicago.<sup>6</sup> They provide several recommendations to embrace the ride-hailing companies as part of the public transportation system and suggest cooperation of public sector and private sector.

The results of my research could fill an important gap in the literature by quantifying and testing Uber's argument. Broadly, this can be thought of as an experiment in studying the industrial organization of markets under pressure from both traditional players and innovative players.

---

<sup>5</sup> <https://www1.nyc.gov/assets/operations/downloads/pdf/For-Hire-Vehicle-Transportation-Study.pdf>

<sup>6</sup> [https://www.researchgate.net/publication/325262585\\_Uber\\_Economics\\_Evaluating\\_the\\_Monetary\\_and\\_Nonmonetary\\_Tradeoffs\\_of\\_TNC\\_and\\_Transit\\_Service\\_in\\_Chicago\\_Illinois](https://www.researchgate.net/publication/325262585_Uber_Economics_Evaluating_the_Monetary_and_Nonmonetary_Tradeoffs_of_TNC_and_Transit_Service_in_Chicago_Illinois)

### 3. Data

#### [Data Description]

The main data we use consists of three parts: the Uber data, the Yellow Cab data, and the Census data in New York City that includes the number of crimes, socio-economic factors including education level, income level, and total population, and demographic information including racial composition and number of people by different age groups.

Since Uber has been the leading ride-hailing firm for years, we choose to use Uber as the industry representative for the private sector players, i.e. ride-hailing firms, in our model. We acquire our Uber dataset from Kaggle.com<sup>7</sup>, an open source data science website that usually publicizes datasets for data scientists to investigate. The Uber data provided is relatively simple containing only the coordinates (latitude, longitude) of every Uber trip pickup location and the time, which is between April 2014 and September 2014. This dataset is the only one, and therefore, the most prevalent Uber trip data online since Uber only provided a sample of their data to the public in order to protect their users' privacy and safety.

The Yellow Taxi dataset is from the NYC Taxi & Limousine Commission<sup>8</sup> and it contains the pickup coordinates of every trip, the time, and some taxi-specific information. To align with our Uber data, we only look at the period from April to September in 2014. Since

---

<sup>7</sup> <https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city>

<sup>8</sup> [http://www.nyc.gov/html/tlc/html/about/trip\\_record\\_data.shtml](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)



Yellow Taxi has a much larger market share than Green Taxi, we choose to only use them to represent the public sector players.<sup>9</sup>

The third part of our data involves Census data including the number of crime occurrences, socio-economic variables, and demographic information.<sup>10</sup> This part is key to determine whether geographical differences would have significantly different impacts on Uber and yellow taxi pickups. The variables here not only help us validate the “illed-served” area argument by Uber but also contain factors like average commute time to work and the number of people who are residents for at least one year, which we think would be relevant to our study.

#### [Data Implementation & Visualization]

To have a direct understanding of the data, we believe that we should create some heatmaps since the question we are interested in is highly related to geographical differences, and heatmaps can highlight some features of the data that we can hardly capture otherwise.

[Figure 1] and [Figure 2] are two heatmaps for Uber and Yellow Cab pickups in the NYC area, in which redder means more pickups happened in that region, and [Figure 3] is a heatmap for the crime data, in which redder means more crimes happened in that region. For the two pickup heatmaps, we use 20-meter times 20-meter grids as the smallest unit. We count the number of pickups by Uber or Yellow Cab within that unit and color the unit according to the

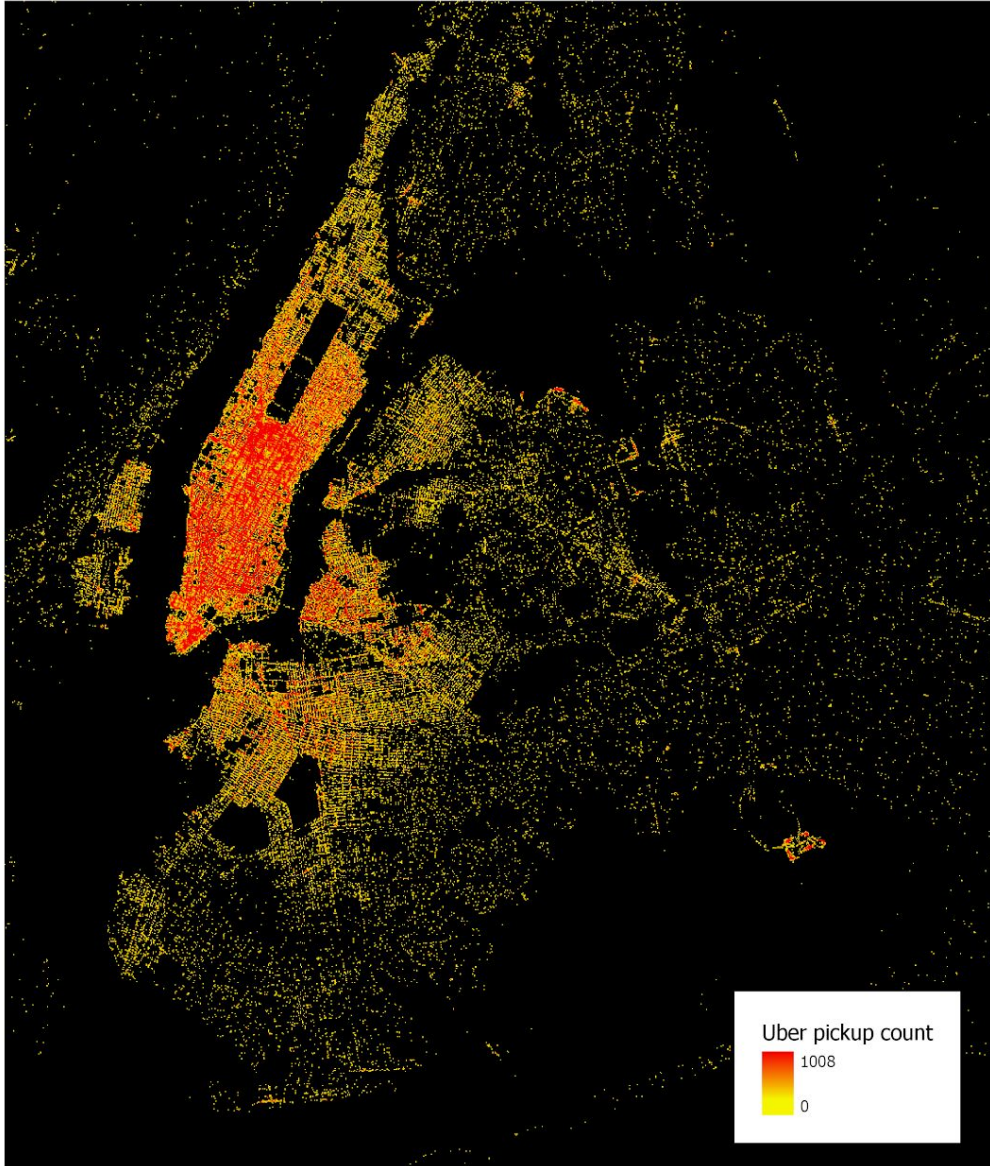
---

<sup>9</sup> <http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/>

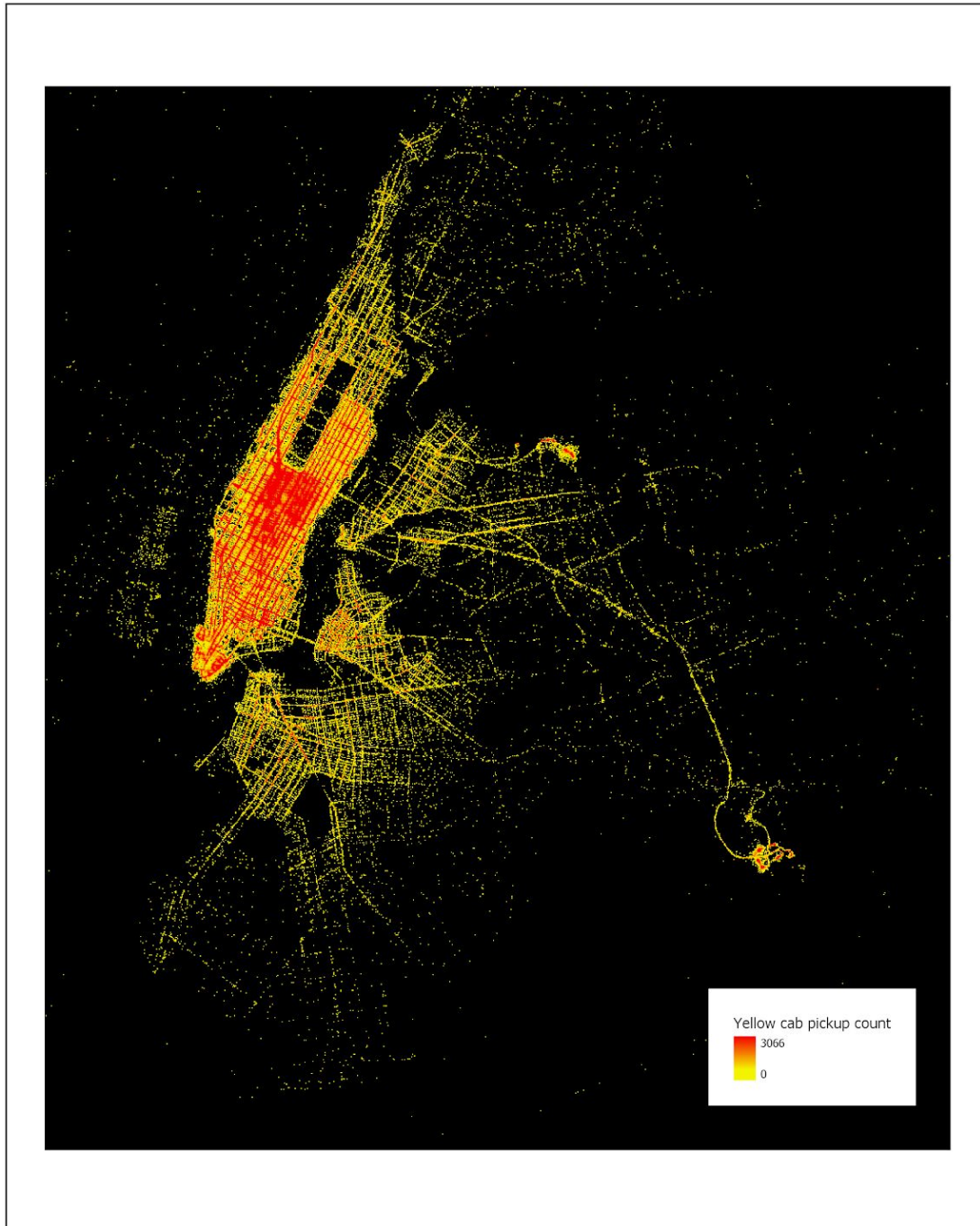
<sup>10</sup> <https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u>

level of pickups. However, one should be aware that the scale of the two heatmaps are different, which means that a really red point in Uber's heatmap may not contain more pickups than a yellow point in Yellow Cab's heatmap. For [Figure 3] we use a 100-meter times 100-meter grid since the number of crimes are far less than the number of pickups. Also, instead of having the exact location where the crime happened, our crime data only recorded the location where the criminals were arrested. This is going to be a limitation for our research and future research can work on improve this.

[Figure 1]

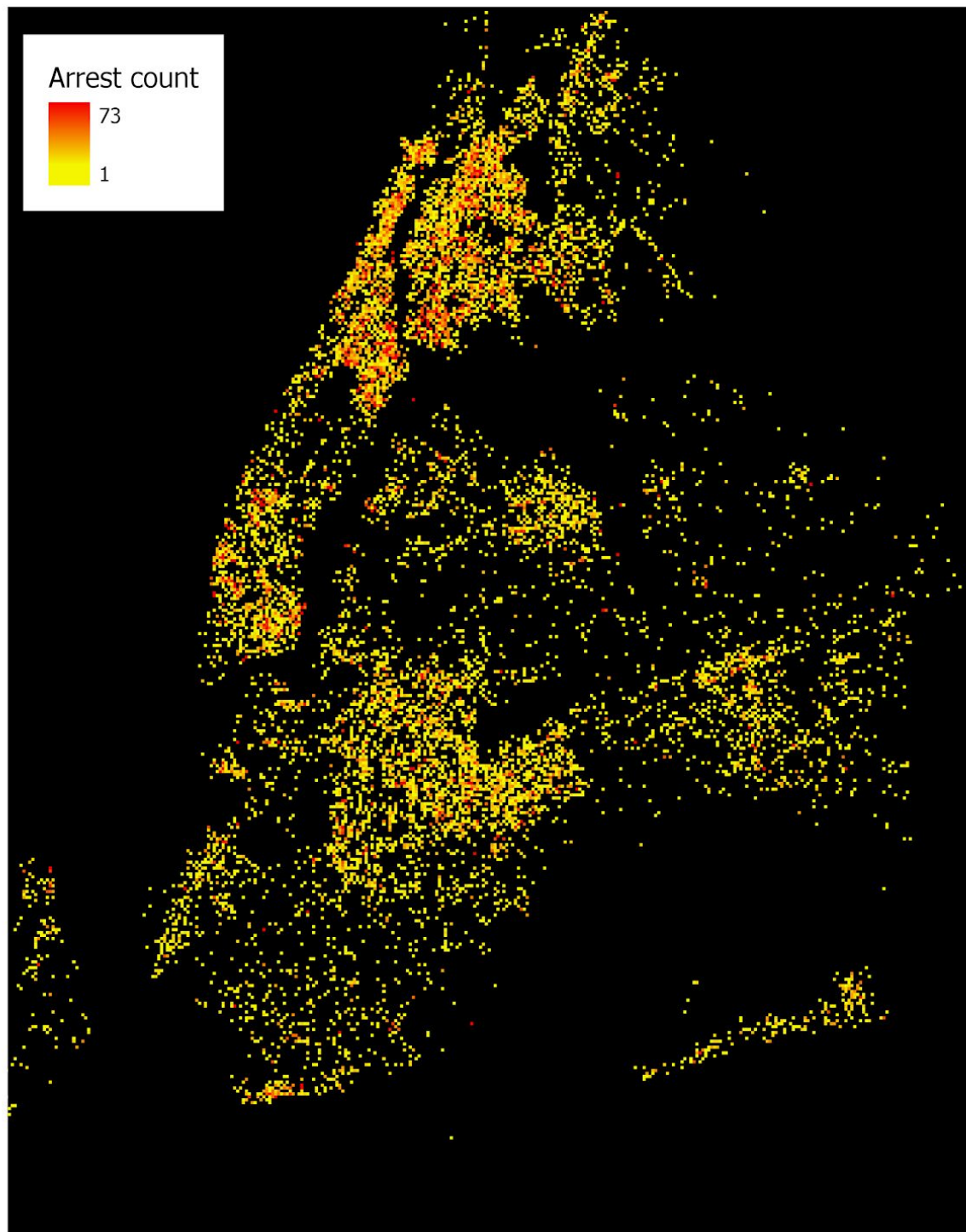


[Figure 2]



From the two heatmaps, we can see those yellow cab pickups are more concentrated in the midtown Manhattan area while Uber serves apparently a much larger area. Also, if we look carefully at the colors in these two maps, we can see that most “red” areas from the Yellow Cab map are at the midtown Manhattan area but Uber drivers seem more likely to be more evenly spread out with also many “red” areas outside Manhattan. These two observations motivate the investigations into the reasons for such geographical differences.

[Figure 3]



As a much smaller dataset compared to the size of the other two, the crime data heatmap is shown by [Figure 3]. We notice that even though midtown Manhattan is one of the areas suffering many crimes, other places like Harlem, the Bronx, and Brookline also have high crime occurrences. Combining the two heatmaps for Uber and Yellow Cab, we can see that Uber's argument seems to make sense intuitively, as Uber appears to be in places relating to high crime occurrences more often than Yellow Cab appears to be. However, the conclusion that Uber is more accessible in high crime areas can be due to the fact that we do not consider confounding variables, such as education level, income level, population, etc. Therefore, to fully understand the story behind these data, we have to use some econometric models.

#### [Methodology]

These three figures above give us a basic direction for our empirical analysis. The key idea of the analysis is to treat each census tract in the NYC region as an observation, and build an econometric model upon these census tracts. In our scenario, each census tract will have a variable recording the number of Uber pickups within that region, a variable recording the number of Yellow Cab pickups within that region, a variable recording the occurrences of crimes within that region, and many other census-tract-level variables, which includes total population, median household income, the number of people who achieve bachelor degrees or higher, the number of people older than 65, average commute time to work, and the number of people living

in this place for at least one year. By construction, we have to use the count data model to estimate our data since all of our dependent variables are non-negative counts.

[Summary Statistics]

	Mean	Standard Deviation
# Uber Pickups	336.74	1096.93
# Yellow Taxi Pickups	1049.44	3907.58
# Total Pickups	1386.17	4937.99
# Total Crimes	12.32	17.66
Total Population	4070.13	2227.35
# Above 65	551.43	427.57
# Bachelor or above	1053.47	427.57
Median Household Income	63542.45	32262.33
Residents Mobility	4018.17	2199.47
Average Commute to Work	41.60	7.11

The above table summarizes the key variables in our research. For example, a mean of 336.74 for the number of Uber pickups means that on average for a given census tract it will have 336.74 pickups by Uber drivers.

#### **4. Empirical Model & Analysis**



[Model Selection]

a.) Standard Poisson Count Model

To deal with the count data, we believe that Poisson regression is a good start since the Poisson Distribution is a popular count model to estimate the number of times an event occurs in an interval of time or space. In our case, the dependent variables, the number of pickups by Uber and the number of pickups by Yellow Cab, both satisfy this condition, so we decide to start with implementing a standard Poisson regression model with all the variables mentioned above.

A Poisson Count Data Model will have the form:

$$\log(E(Y_{it}|x)) = \alpha + \beta x$$

where  $Y_{it}$  represents the number of pickups at census tract  $i$  in period  $t$  and  $\alpha + \beta x$  is the linear predictor including the total number of crime happen in census tract  $i$ , and other socio-economic factors describing census tract  $i$ .

The regression outputs are shown in [Figure 4] and [Figure 5] respectively for Uber data and Yellow Cab data.

Predictors	dataSUber			dataSUber			dataSUber			dataSUber			dataSUber			dataSUber			
	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	
(Intercept)	5.550756	5.547797 – 5.553715	<0.001	5.097234	5.092757 – 5.101711	<0.001	2.960411	2.953602 – 2.967220	<0.001	2.946294	2.939448 – 2.953140	<0.001	4.785177	4.775318 – 4.795035	<0.001	10.491322	10.472288 – 10.510356	<0.001	10.402383 – 10.441987
dataSTotal Crime	0.016739	0.016659 – 0.016820	<0.001	0.013021	0.012952 – 0.013110	<0.001	0.022319	0.022242 – 0.022395	<0.001	0.021948	0.021870 – 0.022027	<0.001	0.021165	0.021093 – 0.021238	<0.001	0.013436	0.013357 – 0.013514	<0.001	0.013157 – 0.013316
dataSTotal Pop				0.000112	0.000112 – 0.000113	<0.001	0.000112	0.000111 – 0.000113	<0.001	0.000140	0.000139 – 0.000142	<0.001	-0.000377	-0.000380 – -0.000374	<0.001	-0.000175	-0.000178 – -0.000173	<0.001	-0.001500 – -0.001411
dataSMedian Household Income							0.000024	0.000024 – 0.000024	<0.001	0.000024	0.000024 – 0.000024	<0.001	0.000012	0.000012 – 0.000012	<0.001	0.000003	0.000003 – 0.000003	<0.001	0.000003 – 0.000003
dataSAbove 65							-0.000170	-0.000176 – -0.000164	<0.001	-0.000694	-0.000700 – -0.000687	<0.001	-0.000471	-0.000478 – -0.000464	<0.001	-0.000492	-0.000499 – -0.000485	<0.001	-0.000499 – -0.000485
dataSBachelor												0.000967	0.000963 – 0.000971	<0.001	0.000540	0.000537 – 0.000544	<0.001	0.000518 – 0.000526	
dataSAvg Com To Work															<0.001	-0.142858	-0.143304 – -0.142412	<0.001	-0.142563 – -0.141662
dataSResidents Mobility																0.001309	0.001264 – 0.001355	<0.001	0.001264 – 0.001355
Observations	1957			1957			1957			1957			1957			1957			1957
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000

[Figure 4]

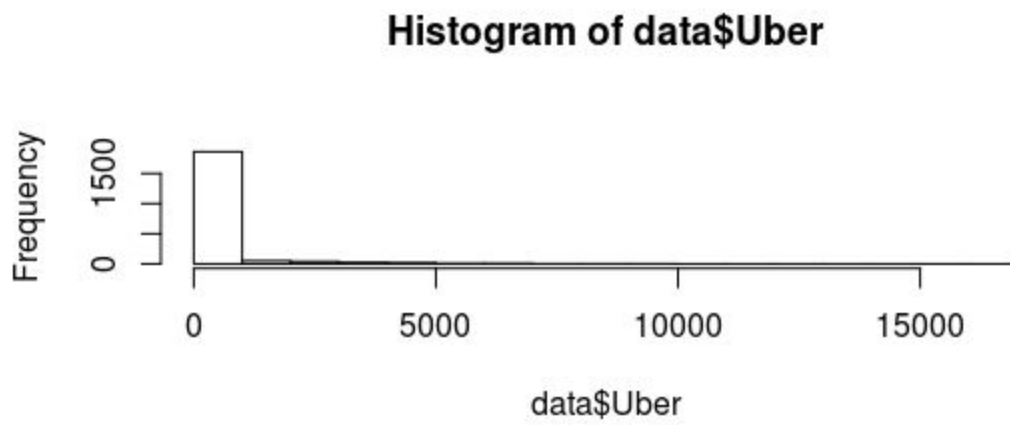
Predictors	data\$Yellow			data\$Yellow			data\$Yellow			data\$Yellow			data\$Yellow			data\$Yellow					
	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p			
(Intercept)	6.652324	6.650643 – 6.654005	<0.001	6.065252	6.062813 – 6.067692	<0.001	3.701047	3.697101 – 3.704992	<0.001	3.711794	3.707822 – 3.715767	<0.001	5.813822	5.808008 – 5.819635	<0.001	12.935002	12.924052 – 12.945951	<0.001	12.873923	12.862823 – 12.885022	<0.001
data\$Total Crime	0.018328	0.018285 – 0.018371	<0.001	0.013927	0.013879 – 0.013974	<0.001	0.023588	0.023548 – 0.023629	<0.001	0.023759	0.023718 – 0.023800	<0.001	0.022320	0.022281 – 0.022358	<0.001	0.013672	0.013631 – 0.013714	<0.001	0.013482	0.013440 – 0.013524	<0.001
data\$Total Pop				0.000140	0.000140 – 0.000141	<0.001	0.000150	0.000150 – 0.000151	<0.001	0.000138	0.000137 – 0.000139	<0.001	-0.000508	-0.000510 – -0.000507	<0.001	-0.000217	-0.000219 – -0.000216	<0.001	-0.001450	-0.001473 – -0.001426	<0.001
data\$Median Household Income							0.000026	0.000026 – 0.000026	<0.001	0.000026	0.000026 – 0.000026	<0.001	0.000012	0.000012 – 0.000012	<0.001	0.000001	0.000001 – 0.000001	<0.001	0.000002	0.000002 – 0.000002	<0.001
data\$Above 65										0.000065	0.000062 – 0.000068	<0.001	-0.000394	-0.000397 – -0.000390	<0.001	-0.000133	-0.000136 – -0.000129	<0.001	-0.000151	-0.000155 – -0.000148	<0.001
data\$Bachelor										0.001138	0.001135 – 0.001140	<0.001	0.000583	0.000581 – 0.000585	<0.001	0.000565	0.000563 – 0.000567	<0.001	0.000565	0.000563 – 0.000567	<0.001
data\$Avg Com To Work													-0.188506	-0.188785 – -0.188227	<0.001	-0.187853	-0.188134 – -0.187572	<0.001			<0.001
data\$Residents Mobility																0.001261	0.001238 – 0.001285	<0.001			<0.001

[Figure 5]

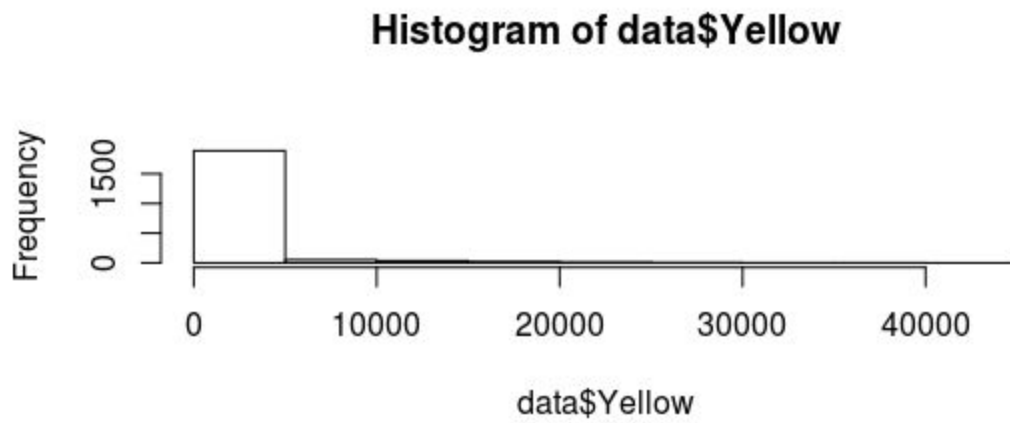
To our surprise, the outputs for both Uber and Yellow Cab seem to have no big differences in terms of magnitude and relationship. From the above two tables, we can see that all the variables we include are statistically significant for the two types of transportation. If we look at the results more closely, we can see that the coefficients for the variable Total Crime are always positive for Uber and Yellow Cab, meaning that a higher number of crimes in the census tract relates to a higher number of pickups. This result seems unintuitive at the beginning since we would suppose that people in high-crime-occurrence areas should be less likely to want to take Uber or Yellow Cab. However, the positive relationship may suggest that people in those high-crime-occurrence areas are more likely to take taxis or for-hire vehicles because they would not like to take other kinds of public transportations due to the higher chances of encountering crime events when taking a subway or a bus. Another important feature from the result indicates that there's a negative relationship between the total population and the number of pickups by both Uber and Yellow Cab. This result may also suggest we have an omitted variable bias that makes the coefficient of Total Population greater than its true value.

b) Zero-inflated Poisson Model

[Figure 6]



[Figure 7]



By looking at the distribution of both Uber and Yellow Cab pickups histograms from [Figure 6] and [Figure 7], we can see that both of them have many zero values and an overdispersion problem. Therefore, in this case, the standard Poisson Model may not be a good estimation to capture all the effects, so we want to use a zero-inflated model instead, in which the regression treats zero values and non-zero values differently and runs a two-step regression for the variables specified. To use the zero-inflated model, we can account for the fact that many values are zero and solve the overdispersion problem. However, the zero-inflated function in R is really sensitive to multicollinearity, and hence many of the variables we are interested cannot be put into it. To deal with this problem, we come up with two solutions: 1) manually mimic the zero-inflated Poisson model by classifying the values of the dependent variables into zeroes and positive values and run two regressions on both Uber and Yellow Cab --- a standard Poisson regression on the positive values and a probit regression on the whole dataset, and 2) use the zero-inflated model in R to run an alternative model by replacing the variables with high VIF (Variance Inflation Factor) values by some other similar measurements.

1) The results from the standard Poisson regression on only positive values are shown by [Figure 8] and [Figure 9] for Uber and Yellow Cab respectively.

Predictors	Log-Mean	CI	positive Uber pickups\$Uber	p	Log-Mean	CI	positive Uber pickups\$Uber	p	Log-Mean	CI	positive Uber pickups\$Uber	p	Log-Mean	CI	positive Uber pickups\$Uber	p	Log-Mean	CI	positive Uber pickups\$Uber	p
(Intercept)	5.556608	5.553647 – 5.559568	<0.001	5.105132	5.100649 – 5.109615	<0.001	2.962217 – 2.975846	<0.001	2.955013	2.948161 – 2.961864	<0.001	4.781775 – 4.801488	<0.001	10.487018	10.467984 – 10.506052	<0.001	10.418194	10.398843 – 10.437546	<0.001	
positive Uber pickups\$Total Crime	0.016669	0.016589 – 0.016750	<0.001	0.012971	0.012883 – 0.013060	<0.001	0.022287 0.022363	<0.001	0.021913	0.021835 – 0.021992	<0.001	0.021145 0.021217	<0.001	0.013428	0.013350 – 0.013507	<0.001	0.013229	0.013150 – 0.013309	<0.001	
positive Uber pickups\$Total Pop					0.000112	0.000111 – 0.000113	0.000112 0.000112	<0.001	0.000140	0.000139 – 0.000141	<0.001	-0.000377 -0.000374	<0.001	-0.000175	-0.000178 – -0.000173	<0.001	-0.001456	-0.001501 – -0.001412	<0.001	
positive Uber pickups\$Median Household Income							0.000024 0.000024	<0.001	0.000024	0.000024 – 0.000024	<0.001	0.000012 0.000012	<0.001	0.000003	0.000003 – 0.000003	<0.001	0.000003	0.000003 – 0.000003	<0.001	
positive Uber pickups\$Above 65								-0.000171	-0.000177 – -0.000165	<0.001	-0.000694 -0.000687	<0.001	-0.000472	-0.000478 – -0.000465	<0.001	-0.000492	-0.000499 – -0.000486	<0.001		
positive Uber pickups\$Bachelor										0.000966	0.000962 – 0.000970	<0.001	0.000540	0.000536 – 0.000543	<0.001	0.000522	0.000518 – 0.000525	<0.001		
positive Uber pickups\$Avg Com To Work													-0.142669	-0.143115 – -0.142223	<0.001	-0.141921	-0.142371 – -0.141470	<0.001		
positive Uber pickups\$Residents Mobility																0.001311	0.001265 – 0.001357	<0.001		
Observations	1947			1947			1947		1947		1947		1947		1947		1947		1947	
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	1.000 / 1.000			1.000 / 1.000			1.000 / 1.000		1.000 / 1.000		1.000 / 1.000		1.000 / 1.000		1.000 / 1.000		1.000 / 1.000		1.000 / 1.000	

[Figure 8]

Predictors	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p	Log-Mean	CI	p
(Intercept)	6.809548	6.807850 – 6.811246	<0.001	6.254278	6.251773 – 6.256783	<0.001	3.952043	3.948031 – 3.956056	<0.001	3.966800	3.962758 – 3.970842	<0.001	6.003927	5.998114 – 6.009741	<0.001	13.014286	13.003139 – 13.025434	<0.001	12.952463	12.941157 – 12.963770	<0.001
positive Yellow pickups\$Total Crme	0.016703	0.016659 – 0.016747	<0.001	0.012798	0.012750 – 0.012846	<0.001	0.022643	0.022602 – 0.022684	<0.001	0.022855	0.022814 – 0.022897	<0.001	0.021705	0.021667 – 0.021744	<0.001	0.013289	0.013248 – 0.013331	<0.001	0.013110	0.013068 – 0.013153	<0.001
positive Yellow pickups\$Total Pop				0.000130	0.000129 – 0.000130	<0.001	0.000138	0.000137 – 0.000138	<0.001	0.000122	0.000121 – 0.000123	<0.001	-0.000506	-0.000508 – -0.000505	<0.001	-0.000222	-0.000223 – -0.000220	<0.001	-0.001371	-0.001394 – -0.001347	<0.001
positive Yellow pickups\$Median Household Income							0.000025	0.000025 – 0.000025	<0.001	0.000025	0.000025 – 0.000025	<0.001	0.000012	0.000012 – 0.000012	<0.001	0.000001	0.000001 – 0.000001	<0.001	0.000001	0.000001 – 0.000001	<0.001
positive Yellow pickups\$Above 65							0.000083	0.000080 – 0.000086	<0.001	-0.000357	-0.000361 – -0.000353	<0.001	-0.000124	-0.000128 – -0.000121	<0.001	-0.000142	-0.000146 – -0.000139	<0.001	-0.000146	-0.000146 – -0.000139	<0.001
positive Yellow pickups\$Bachelor										0.001107	0.001105 – 0.001109	<0.001	0.000576	0.000573 – 0.000578	<0.001	0.000560	0.000558 – 0.000562	<0.001	0.000560	0.000558 – 0.000562	<0.001
positive Yellow pickups\$Avg Com To Work													-0.186701	-0.186984 – -0.186418	<0.001	-0.185955	-0.186240 – -0.185669	<0.001			
positive Yellow pickups\$Residents Mobility																0.001175	0.001151 – 0.001199	<0.001			
Observations	1699			1699			1699			1699			1699			1699			1699		
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000			1.000 / 1.000		

[Figure 9]



From the outputs, we can see that the general pattern of these results is the same as the results from [Figure 4] and [Figure 5], indicating that if we only look at the positive pickup values, then the relationship between the dependent variable and the independent variables will remain the same as if we just do a standard Poisson on the whole dataset. Therefore, since the positive values do not affect the standard Poisson model, we are interested in seeing what if we run a Probit model on the whole dataset when we only care about whether there is any pickup (variable zero = 0) or no pickup at all (variable zero = 1) in a given census tract. In this way, we can tell whether the zero values should be treated differently in the model. The results are shown by [Figure 10] and [Figure 11] for Uber and Yellow Cab respectively.

Predictors	Log-Odds	dataSuber zero CI	p	Log-Odds	dataSuber zero CI	p	Log-Odds	dataSuber zero CI	p	Log-Odds	dataSuber zero CI	p	Log-Odds	dataSuber zero CI	p	Log-Odds	dataSuber zero CI	p
(Intercept)	-2.436687	-2.708200 – -2.165175	<0.001	-2.083386	-2.598992 – -1.567779	<0.001	-1.690565	-2.581936 – -0.799193	<0.001	-1.543783	-2.460906 – -0.626661	0.001	-1.945882	-2.996480 – -0.895285	<0.001	-3.728551	-5.898208 – -1.558894	0.001
dataSTotal Crime	-0.015488	-0.040835 – 0.009860	0.231	-0.010513	-0.036711 – 0.015684	0.432	-0.015228	-0.044696 – 0.014239	0.311	-0.022445	-0.055488 – 0.016598	0.183	-0.023212	-0.056404 – 0.009980	0.170	-0.027092	-0.064621 – 0.010437	0.157
dataSTotal Pop				-0.000116	-0.000274 – 0.000042	0.150	-0.000126	-0.000294 – 0.000042	0.141	0.000042	-0.000164 – 0.000249	0.687	0.000135	-0.000094 – 0.000364	0.247	0.000177	-0.000074 – 0.000427	0.168
dataSTMedian Household Income							-0.000005	-0.000015 – 0.000005	0.294	-0.000005	-0.000015 – 0.000005	0.314	0.000002	-0.000010 – 0.000014	0.771	0.000001	-0.000012 – 0.000015	0.841
dataSAbove 65							-0.001768	-0.003480 – -0.000056	0.043	-0.001349	-0.003213 – 0.000515	0.156	-0.001901	-0.004092 – 0.000291	0.089	-0.001902	-0.004127 – 0.000324	0.094
dataSTBachelor							-0.000780	-0.001803 – 0.000243	0.135	-0.000618	-0.001748 – 0.000511	0.283	-0.000618	-0.001748 – 0.000511	0.283	-0.000618	-0.001749 – 0.000512	0.284
dataSAvg Com To Work													0.040575	-0.002027 – 0.083178	0.062	0.040553	-0.002597 – 0.083704	0.065
dataSTResidents Mobility																0.000027	-0.007627 – 0.007681	0.995
Observations	1957			1957			1957			1957			1957			1957		
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	0.001 / 0.018			0.002 / 0.038			0.003 / 0.048			0.006 / 0.090			0.007 / 0.115			0.009 / 0.149		

[Figure 10]

	data\$yellow zero			data\$yellow zero			data\$yellow zero			data\$yellow zero			data\$yellow zero			data\$yellow zero		
<i>Predictors</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>	<i>Log-Odds</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.922776	-1.012684 - -0.832869	<0.001	-0.332581	-0.585007 - -0.080155	0.010	-0.321350	-0.573508 - -0.069191	0.012	-1.130238	-1.42662 - -0.837813	<0.001	-3.052442	-3.750114 - -2.354770	<0.001	-3.229833	-3.934809 - -2.524856	<0.001
data\$Total Crime	-0.020670	-0.027596 - -0.013744	<0.001	-0.014268	-0.021693 - -0.006842	<0.001	-0.012834	-0.020307 - -0.005361	0.001	-0.016828	-0.024807 - -0.008849	<0.001	-0.016142	-0.024533 - -0.007750	<0.001	-0.016174	-0.024542 - -0.007806	<0.001
data\$Total Pop				-0.000150	-0.000197 - -0.000103	<0.001	-0.000194	-0.000261 - -0.000127	<0.001	-0.000017	-0.000093 - -0.000060	0.665	-0.000019	-0.000099 - -0.000060	0.635	0.003019	0.000849 - 0.005189	0.006
data\$Median Household Income				-0.000002	-0.000004 - -0.000001	0.204	-0.000002	-0.000004 - -0.000001	0.145	0.000011	0.000007 - 0.000015	<0.001	0.000012	0.000009 - 0.000016	<0.001	0.000013	0.000009 - 0.000017	<0.001
data\$Above 65							0.000303	-0.000008 - -0.000614	0.057	0.001025	0.000666 - 0.001384	<0.001	0.000915	0.000537 - 0.001292	<0.001	0.001062	0.000661 - 0.001462	<0.001
data\$Bachelor										-0.001316	-0.001596 - -0.001036	<0.001	-0.001211	-0.001505 - -0.000916	<0.001	-0.001233	-0.001527 - -0.000939	<0.001
data\$Avg Com To Work										0.041707	0.028474 - 0.054940	<0.001	0.045006	0.031711 - 0.058300	<0.001			
data\$Residents Mobility																-0.003098	-0.005311 - -0.000884	0.006
Observations	1957			1957			1957			1957			1957			1957		
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	0.026 / 0.048			0.050 / 0.092			0.052 / 0.095			0.108 / 0.200			0.127 / 0.235			0.131 / 0.241		

[Figure 11]

The above results highlight an important feature, that is the number of crime occurrence is always statistically significant in predicting the chance of having no yellow taxi pickups but never statistically significant in predicting the chance of having no Uber pickups. Although we cannot simply conclude that yellow cab drivers indeed have a geographical discrimination by this outcome, we can at least say that for census tracts with zero pickups, the reason for not having any Uber pickup is more random while the reason for not having any yellow cab pickup is associated with high crime occurrences, income level, education level, and the number of elderly people.

2) To avoid the multicollinearity problem that prevents us from using the zero-inflated Poisson model, we have to exclude several variables that have high VIF values which includes the number of people who have a bachelor's degree or above, median household income, and the number of people above 65. However, since these socio-economic factors are what we want to focus on, we decide not to replace them with other variables and therefore we will not use the zero-inflated model provided by R.

c) Pooling the data with a Uber dummy

To focus on the effect of crime occurrence, we use an alternative way to build up our model, in which we pool the data by duplicating it and add an Uber dummy. Hence, we can include an interaction term of the total number of crimes and the Uber dummy into our equation,

which can directly capture the effect of crime occurrences on both Uber and Yellow Cab pickups. Since we see from part (b) that zero values have significant impacts on our model, we decide to also include two parts of analysis after pooling the data --- a standard Poisson regression on positive-only values and a Probit model to capture the difference between zero pickup and positive pickups. [Figure 12] presents the regression result of the positive-only standard Poisson regression.



From the above table, we can see that if we only take into account the total number of crimes, the interaction term, and the Uber dummy variable, the interaction term is negative but not statistically significant. However, if we start to add other variables into the model, the interaction term begins to become statistically significant while remain being negative. This negative coefficient is somehow surprising because it means that being Uber negatively affects the strength of the positive relationship between the number of crime occurrences and the number of pickups, but the magnitude of this negative coefficient is not large enough to flip the positive relationship to negative. In other words, for a given census tract, the predicted number of pickups by Uber is less than the predicted number of pickups by the Yellow Cab, and as the number of crime occurrences increase, the difference between the pickups will get larger.

[Figure 13] shows the Probit model we build for the pooled data.

Predictors	rep data\$zero			rep data\$zero			rep data\$zero			rep data\$zero			rep data\$zero			rep data\$zero					
	Log Odds	CI	p	Log Odds	CI	p	Log Odds	CI	p	Log Odds	CI	p	Log Odds	CI	p	Log Odds	CI	p			
(Intercept)	-0.922778	-1.012686 – -0.832870	<0.001	-0.461269	-0.625543 – -0.296996	<0.001	-0.318888	-0.563894 – -0.073882	0.011	-0.313079	-0.557649 – -0.068509	0.012	-1.068971	-1.365517 – -0.777426	<0.001	-2.940940	-3.599307 – -2.282572	<0.001	-3.090495	-3.754544 – -2.426446	<0.001
rep data\$Total Crime	-0.020670	-0.027596 – -0.013743	<0.001	-0.013108	-0.020097 – -0.006118	<0.001	-0.014579	-0.021985 – -0.007172	<0.001	-0.013649	-0.021132 – -0.006166	<0.001	-0.017832	-0.025783 – -0.009880	<0.001	-0.017307	-0.025677 – -0.008936	<0.001	-0.017311	-0.025657 – -0.008964	<0.001
rep data\$Uber X Total Crime	0.005182	-0.021095 – 0.031459	0.699	0.003580	-0.022780 – 0.029939	0.790	0.003350	-0.023653 – 0.030352	0.808	0.003251	-0.023418 – 0.029920	0.811	0.002662	-0.020354 – 0.031677	0.670	0.005311	-0.022630 – 0.033253	0.709	0.005454	-0.021976 – 0.032884	0.697
rep data\$Uber Dummy	-1.513909	-1.799921 – -1.227898	<0.001	-1.538126	-1.825725 – -1.250526	<0.001	-1.539316	-1.829644 – -1.248988	<0.001	-1.536686	-1.825300 – -1.248071	<0.001	-1.635150	-1.930675 – -1.339625	<0.001	-1.680716	-1.985738 – -1.375694	<0.001	-1.686957	-1.991060 – -1.382854	<0.001
rep data\$Total Pop				-0.000146	-0.000191 – -0.000102	<0.001	-0.000148	-0.000193 – -0.000103	<0.001	-0.000176	-0.000240 – -0.000112	<0.001	-0.000004	-0.000077 – 0.000068	0.911	-0.000003	-0.000078 – 0.000075	0.942	0.002733	0.000661 – 0.004804	0.010
rep data\$Median Household Income							-0.000002	-0.000004 – 0.000001	0.125	-0.000002	-0.000005 – 0.000000	0.098	0.000010	0.000007 – 0.000014	<0.001	0.000012	0.000008 – 0.000015	<0.001	0.000012	0.000008 – 0.000016	<0.001
rep data\$Above 65							0.000196	-0.000107 – 0.000499	0.205	0.000889	0.000542 – 0.001235	<0.001	0.000769	0.000403 – 0.001135	<0.001	0.000769	0.000403 – 0.001135	<0.001	0.000891	0.000505 – 0.001278	<0.001
rep data\$Bachelor										-0.001272	-0.001540 – -0.001004	<0.001	-0.001172	-0.001456 – -0.000889	<0.001	-0.001172	-0.001456 – -0.000889	<0.001	-0.001190	-0.001473 – -0.000907	<0.001
rep data\$Avg Com To Work										0.040469	0.028030 – 0.052908	<0.001	0.043304	0.030817 – 0.055791	<0.001	0.040469	0.028030 – 0.052908	<0.001	0.043304	0.030817 – 0.055791	<0.001
rep data\$Residents Mobility																-0.002789	-0.004902 – -0.000675	0.010			0.010
Observations	3914			3914			3914			3914			3914			3914			3914		
Cox & Snell's R <sup>2</sup> / Nagelkerke's R <sup>2</sup>	0.087 / 0.222			0.099 / 0.251			0.100 / 0.253			0.127 / 0.323			0.137 / 0.348			0.137 / 0.348			0.138 / 0.352		

[Figure 13]



From this table above, we find that all the coefficients for the interaction term become positive and not significant when we distinguish zero pickups and positive pickups. This positive association between the interaction term and the dependent variable is as what we expect, but the insignificance of the coefficients shows that the number of crimes, in general, has no significantly different impacts on getting zero Uber or Yellow taxi pickups.

#### d) OLS Regression with density

After doing several Poisson regressions, we want to go back to the simplest model, which is the OLS regression. Even though by performing an OLS regression, we ignore some of the data generating processes, it is still useful to include it because it does a better job when adding interaction terms to the model. The following results are shown by [Figure 14], [Figure 15], and [Figure 16] are OLS regression results for Uber pickups, Yellow Cab pickups, and Pooled dataset with an Uber dummy and an interaction term combining crimes and Uber dummy together respectively. For the OLS regression, to make it easier to interpret, instead of using the counts (number of pickups and number of crime occurrences), we divide them by the total population at each census tract to make these variables continuous.

Predictors	Estimates	data\$Uber Den CI	p	Estimates	data\$Uber Den CI	p	Estimates	data\$Uber Den CI	p	Estimates	data\$Uber Den CI	p	Estimates	data\$Uber Den CI	p
(Intercept)	-0.183006	-0.232007 – -0.134006	<0.001	-0.849342	-0.943043 – -0.755641	<0.001	-0.842130	-0.948834 – -0.735426	<0.001	-0.922325	-1.040127 – -0.804522	<0.001	-0.080922	-0.404747 – -0.242902	0.624
data\$Crime Den	97.599607	89.518842 – 105.680372	<0.001	106.420007	98.743803 – 114.096210	<0.001	106.281955	98.542115 – 114.021795	<0.001	106.507857	98.783927 – 114.231787	<0.001	102.820953	95.040091 – 110.601814	<0.001
data\$Median Household Income				0.000010	0.000009 – 0.000011	<0.001	0.000010	0.000009 – 0.000011	<0.001	0.000012	0.000010 – 0.000013	<0.001	0.000011	0.000009 – 0.000012	<0.001
data\$Above 65							-0.000013	-0.000106 – 0.000080	0.782	0.000124	-0.000002 – 0.000251	0.055	0.000183	0.000055 – 0.000310	0.005
data\$Bachelor										-0.000086	-0.000140 – -0.000032	0.002	-0.000143	-0.000201 – -0.000086	<0.001
data\$Avg Com To Work													-0.018207	-0.024741 – -0.011672	<0.001
data\$Residents Mobility															
Observations	1957			1957			1957			1957			1957		
R <sup>2</sup> / adjusted R <sup>2</sup>	0.223 / 0.222			0.313 / 0.312			0.316 / 0.315			0.327 / 0.325			0.327 / 0.325		

[Figure 14]

Predictors	dataSYellow Den			dataSYellow Den			dataSYellow Den			dataSYellow Den			dataSYellow Den			dataSYellow Den		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	-0.893005	-1.074355 – -0.711655	<0.001	-3.425131	-3.770680 – -3.079582	<0.001	-3.480019	-3.873493 – -3.086544	<0.001	-0.852446	-2.045633 – -0.340740	0.162	-0.739971	-1.992365 – 0.512423	0.247			
data\$Crime Den	413.373915	383.467234 – 443.280597	<0.001	446.892095	418.583964 – 475.200226	<0.001	447.942700	419.401722 – 476.483679	<0.001	435.858483	407.188604 – 464.528362	<0.001	436.282168	407.571730 – 464.992605	<0.001			
data\$Median Household Income				0.000038	0.000034 – 0.000043	<0.001	0.000038	0.000034 – 0.000043	<0.001	0.000042	0.000037 – 0.000048	<0.001	0.000042	0.000035 – 0.000048	<0.001			
data\$ABove 65							0.000100	-0.000242 – 0.000442	0.567	0.000724	0.000258 – 0.001190	0.002	0.000932	0.000463 – 0.001402	<0.001	0.001017	0.000467 – 0.001568	<0.001
data\$Bachelor										-0.000391	-0.000390 – -0.000192	<0.001	-0.000594	-0.000806 – -0.000382	<0.001	-0.000561	-0.000800 – -0.000322	<0.001
data\$Avg Com To Work													-0.004746	-0.088822 – -0.040670	<0.001	-0.064787	-0.088868 – -0.040706	<0.001
data\$Residents Mobility																-0.000035	-0.000152 – 0.000083	0.562
Observations	1957			1957			1957			1957			1957			1957		
R <sup>2</sup> / adjusted R <sup>2</sup>	0.273 / 0.273			0.362 / 0.361			0.362 / 0.361			0.367 / 0.365			0.376 / 0.374			0.376 / 0.374		

[Figure 15]

Predictors	rep data\$Pickup Den			rep data\$Pickup Den			rep data\$Pickup Den			rep data\$Pickup Den			rep data\$Pickup Den			rep data\$Pickup Den		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	-0.89005	-1.025837 – -0.760172	<0.001	-2.492236	-2.695185 – -2.289286	<0.001	-2.516074	-2.741953 – -2.290194	<0.001	-2.738468	-2.984182 – -2.492754	<0.001	-0.821684	-1.458528 – -0.184840	0.011	-0.757460	-1.425244 – -0.089677	0.026
rep data\$Crime Den	413.373915	391.468341 – 435.279489	<0.001	434.543205	413.548825 – 455.537584	<0.001	434.999482	413.917469 – 456.081495	<0.001	435.025949	414.592640 – 456.659237	<0.001	427.226872	406.141879 – 448.311865	<0.001	427.468795	406.368609 – 448.568982	<0.001
rep data\$Crime Den X Uber	-315.774308	-346.753468 – -284.795148	<0.001	-315.774308	-345.316085 – -286.232532	<0.001	-315.774308	-345.319024 – -286.229592	<0.001	-315.774308	-345.248331 – -286.300385	<0.001	-315.774308	-345.099192 – -286.449424	<0.001	-315.774308	-345.101470 – -286.447146	<0.001
rep data\$Uber Dummy	0.709998	0.522145 – 0.897852	<0.001	0.709998	0.530861 – 0.889136	<0.001	0.709998	0.530843 – 0.889154	<0.001	0.709998	0.531272 – 0.888724	<0.001	0.709998	0.532176 – 0.887821	<0.001	0.709998	0.532162 – 0.887835	<0.001
rep data\$Median Household Income				0.000024	0.000022 – 0.000027	<0.001	0.000024	0.000022 – 0.000026	<0.001	0.000028	0.000025 – 0.000031	<0.001	0.000027	0.000024 – 0.000030	<0.001	0.000026	0.000023 – 0.000030	<0.001
rep data\$Above 65							0.000043	-0.000137 – 0.000224	0.637	0.000424	0.000178 – 0.000670	0.001	0.000558	0.000310 – 0.000806	<0.001	0.000606	0.000315 – 0.000897	<0.001
rep data\$Bachelor										-0.000239	-0.000344 – -0.000135	<0.001	-0.000369	-0.000480 – -0.000257	<0.001	-0.000350	-0.000476 – -0.000223	<0.001
rep data\$Avg Com To Work													-0.041476	-0.054201 – -0.028752	<0.001	-0.041500	-0.054226 – -0.028774	<0.001
rep data\$Residents Mobility																-0.000020	-0.000082 – 0.000042	0.531
Observations	3914			3914			3914			3914			3914			3914		
R <sup>2</sup> / adjusted R <sup>2</sup>	0.271 / 0.271			0.338 / 0.337			0.341 / 0.340			0.341 / 0.340			0.348 / 0.347			0.348 / 0.347		

[Figure 16]

We can see that the general pattern of the results from OLS regression stay unchanged as what we found before when using the Poisson model. That is, a) there is a positive relationship between the crime density and the Uber/Yellow Cab pickup density, b) there is a negative and significant relationship between the interaction term and the pickup density, which suggests that being Uber in fact negatively affects the strength of the positive relationship between crime density and pickup density, but the magnitude of this coefficient is also not large enough to flip the positive relationship to negative.

#### e) Other Potential models

Throughout our research, we have also attempted to use several other econometric models, such as Poisson Pseudo Maximum Likelihood (ppml) and Fractional Multinomial Logit (fmlogit) to best estimate our data. However, the ppml model requires a distance measure which we do not have in our dataset and the fmlogit model doesn't provide us the results we are looking for, so we do not include them in our discussion.

## **5. Qualification & Conclusion**

We can conclude two important findings from our research. First, no matter which model we choose to use, the total number of crime occurrences (or the crime density) is positively associated with both the number of pickups by Uber (or Uber pickup density) and the number of

pickups by Yellow Cab (or Yellow pickup density). This is possible because people are not willing to take public transportation or walking on the street when they are in regions with high crime occurrences. The second observation, at least from our result, shows that Uber and Yellow taxi do not have statistically significant different preferences on pickup locations, which is not what we expect from the heatmaps. However, due to several limitations of our research, our results may not be able to fully capture the whole picture. First, our model may face an omitted variable bias since we do not have the data to account for the availability of public transportation in each census tract. As a substitute for Uber or Yellow Cab, public transportation such as subways or buses may play significant roles in our setting. Second, our data source is limited and probably outdated. In 2014, Uber was only a startup with less than 100k drivers on the road but now it has grown to a giant company with more than 500k registered drivers. Therefore, to use the 2014 data to draw conclusions about the contemporary situation may not be appropriate. However, our research will provide a basic guideline for future researchers to work on this type of problem. Third, future research can expand the use of the dataset from a cross-sectional dataset to panel data and take into account the time effects as well. Fourth, if data allows, future research can look at the exact location where crimes happened rather than the arrest location we use.

## **6. Acknowledgement**

I want to thank Tim Hubbard, who is my primary reader and thesis advisor for all the help including model selection, results interpretation, presentation preparation, etc. I also want to thank Lindsey Novak, who is my second reader and helps me a lot with my writings. I want to express my special thanks to Manny Gimond, who works for the GIS Lab and help me with making beautiful maps.