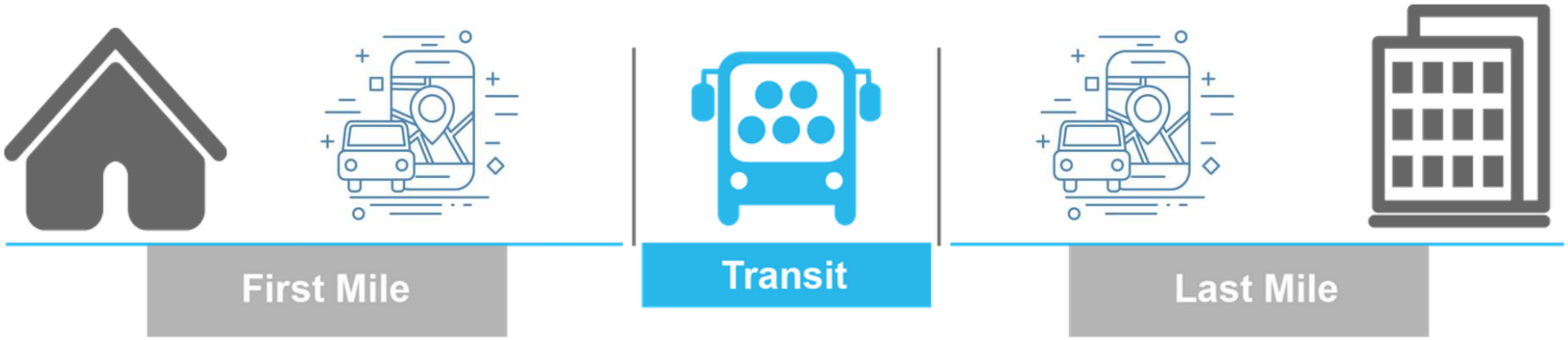
Modeling First-Mile Ride-Hailing Needs and Carpool Likelihood in Chicago, IL



CPLN 505 Final Project
Ruqi Chen, Qi Si, Jiamin Tan
4/25/2021

Ride-Hailing and First/Last-Mile Problems

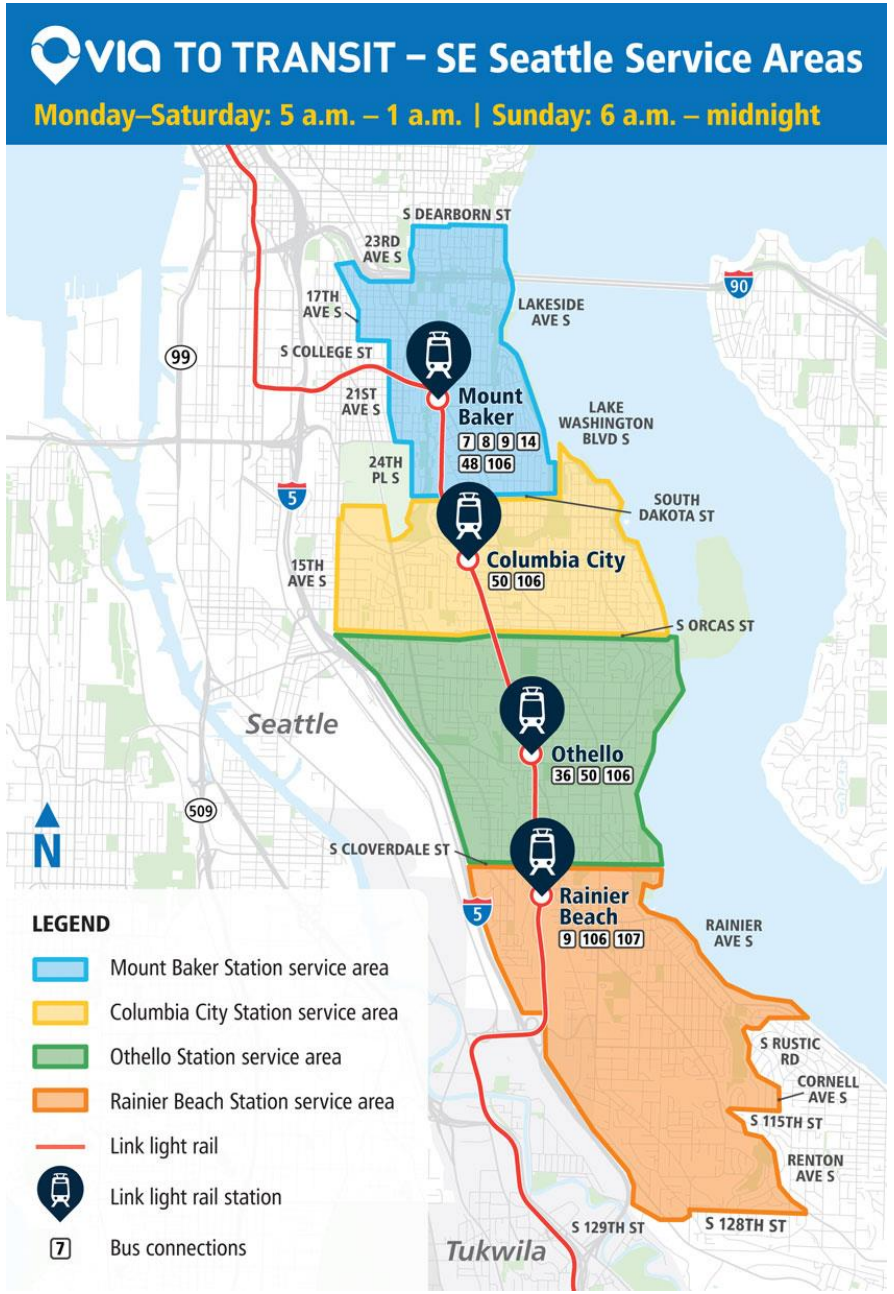
- In Philadelphia, 2016
  - SEPTA partnered with Uber
    - "…Uber rides will be discounted by 40 percent to-and-from 11 suburban Regional Rail stations …with a maximum discount of $10 per ride" (SEPTA website 2016).

Source: iseptaphilly.com

Ride-Hailing and First/Last-Mile Problems

- In Seattle, 2019
  - King County Metro and Via to Transit
    - "Via to Transit is a pilot, ***on-demand service***… connects riders to and from three transit hubs in southeast Seattle and Tukwila".
    - "Rides will be ***shared*** with other Metro customers, while assuring you get to your destination promptly"(King County website).

Source: kingcounty.gov



Source: kingcounty.gov

Research Questions

Research Question 1

What socioeconomic factors influence the number of first-mile ride-hailing trips originated at places near major transit stations?

Research Question 2

What socioeconomic factors influence people's willingness to carpool in first-mile ride-hailing trips?

Ride-Hailing Service Data: *Transportation Network Providers – Trips*, Chicago Data Portal.

- Original Dataset
  - Nov 1, 2018 – April 2, 2021 (last updated)
  - 186 million rows with 21 variables; each row is a trip
    - Spatial resolution: census tract
    - Temporal resolution: rounded to every 15 minutes

- **Data Used in Our Study**
  - Monday, Dec 3, 2018. from 6 am to 9 am.
  - 29359 rows without NAs.

Ride-Hailing Service Data:
Transportation Network Providers – Trips, Chicago Data Portal.

- A glance of the data

trip_start_timestamp	trip_end_timestamp	trip_seconds	trip_miles	pickup_census_tract
2018-12-03 07:45:00	2018-12-03 08:00:00	597	1.4106094	17031081403
2018-12-03 06:45:00	2018-12-03 07:00:00	1089	5.3045963	17031062400
2018-12-03 08:45:00	2018-12-03 09:00:00	776	1.5944449	17031320100
2018-12-03 08:00:00	2018-12-03 08:15:00	989	1.9820785	17031281900
2018-12-03 08:00:00	2018-12-03 08:45:00	2439	6.0767964	17031060200
2018-12-03 06:00:00	2018-12-03 07:00:00	3290	15.9732930	17031241500
2018-12-03 08:45:00	2018-12-03 09:00:00	1310	5.2808794	17031242500

- It does NOT contain any socioeconomic information of each rider

Ride-Hailing Service Data:
Transportation Network Providers – Trips, Chicago Data Portal.

- Why Monday, Dec 3, 2018. from 6 am to 9 am?
    - The date (Dec 3, 2018) was arbitrarily selected…
    - **However**, …
    - Most riders should depart from **home** between 6am and 9 am on a Monday.
    - Therefore, Socioeconomic characteristics of these riders can then be described by ***census data*** of their origin census tracts (home).

Socioeconomics Data:
2018 American Community Survey (ACS) 5-Year Estimate

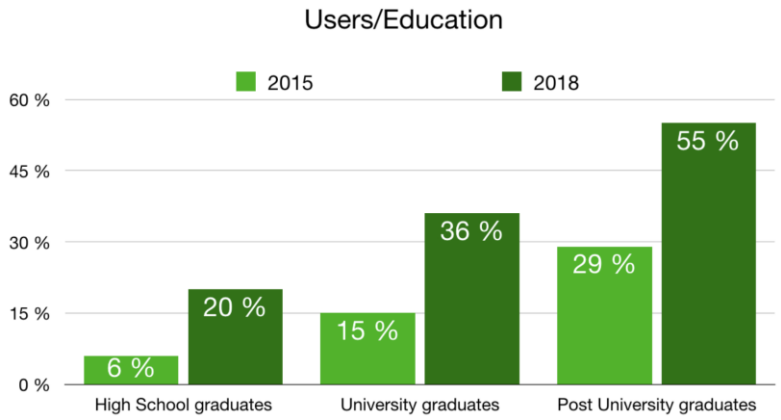
- Spatial Resolution
  - Census Tract
- Factors/Variables Considered

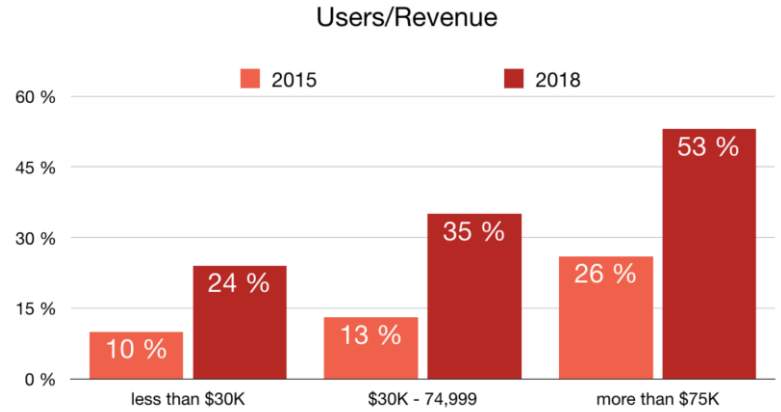
Demographics	Housing	Commuting	Other
Age	Rent	Commuting Time	Educational Attainment
Race	Ownership	Carpooled Amount	Household Vehicle Availability
Gender	Mortgage		Marital Status
Income			
Population Density			

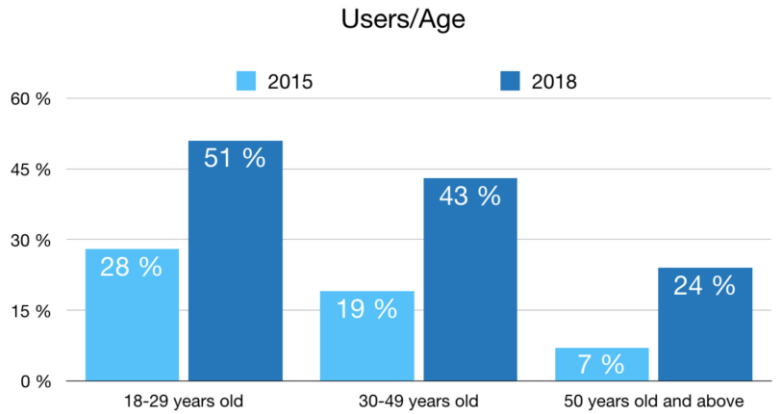
Some Literature Reviewed

- Market research: who are the customers of ride-hailing applications? By Pew Research Center, 2018
- Factors Influencing Willingness to Pool in Ride-Hailing Trips By Hou Yi, et al, 2020

Some Researches on Users of Ride-Hailing Services





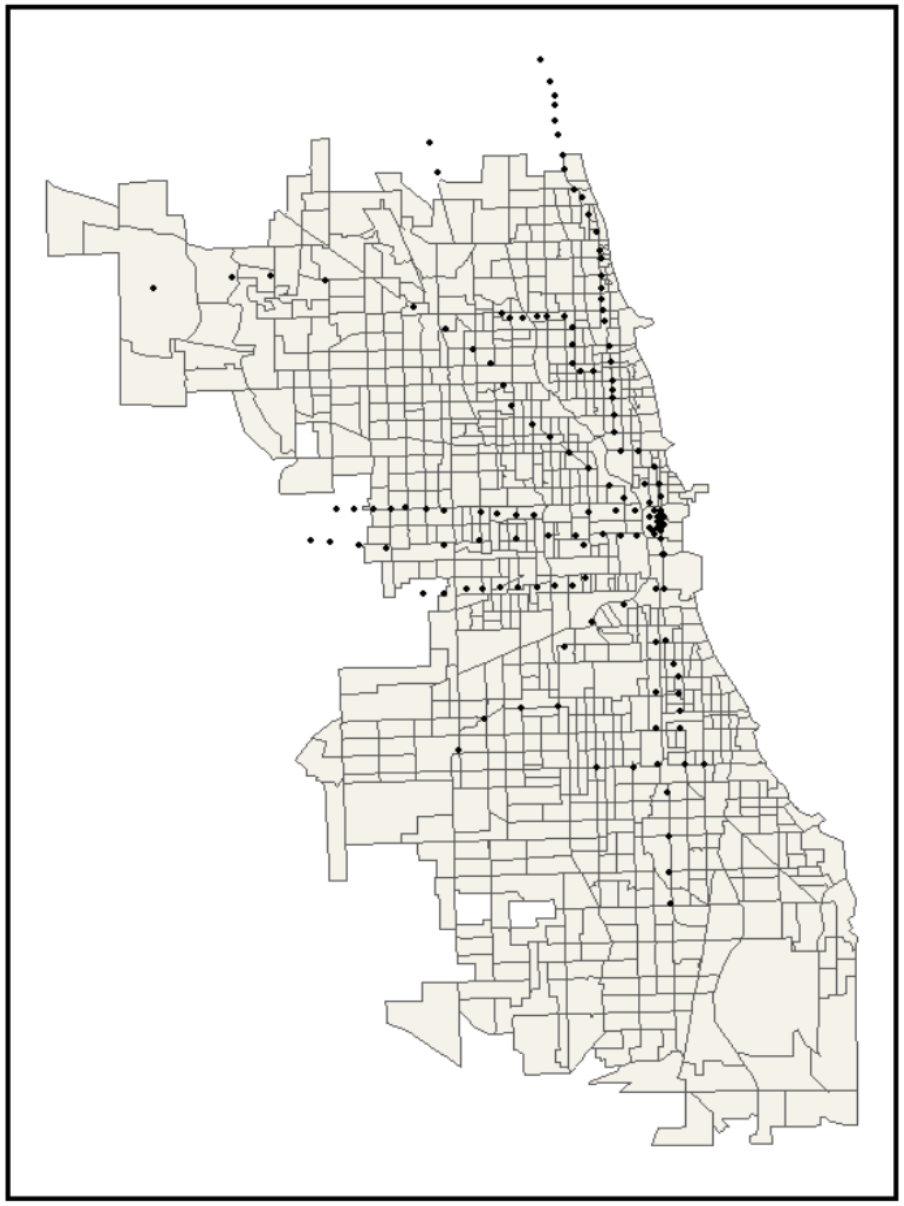


Source: Pew Research Center 2018

Other Relevant Data:

- Administrative Boundary of the City of Chicago
  - 1 Polygon
- Chicago Transit Authority (CTA) Train Stations
  - 144 Points
- CTA Bus Stops
  - 10847 Points

CTA Train Stations and Chicago Census Tracts



Find Potential First Mile Trips:

- The ride-hailing service dataset does NOT indicate which trips are first mile trips
- We made the following assumptions:
  - A trip will be identified as a first mile trip
    - If its destination is in a census tract accessible to at least one CTA train station

      AND
    - If the trip is shorter than 1.5 miles

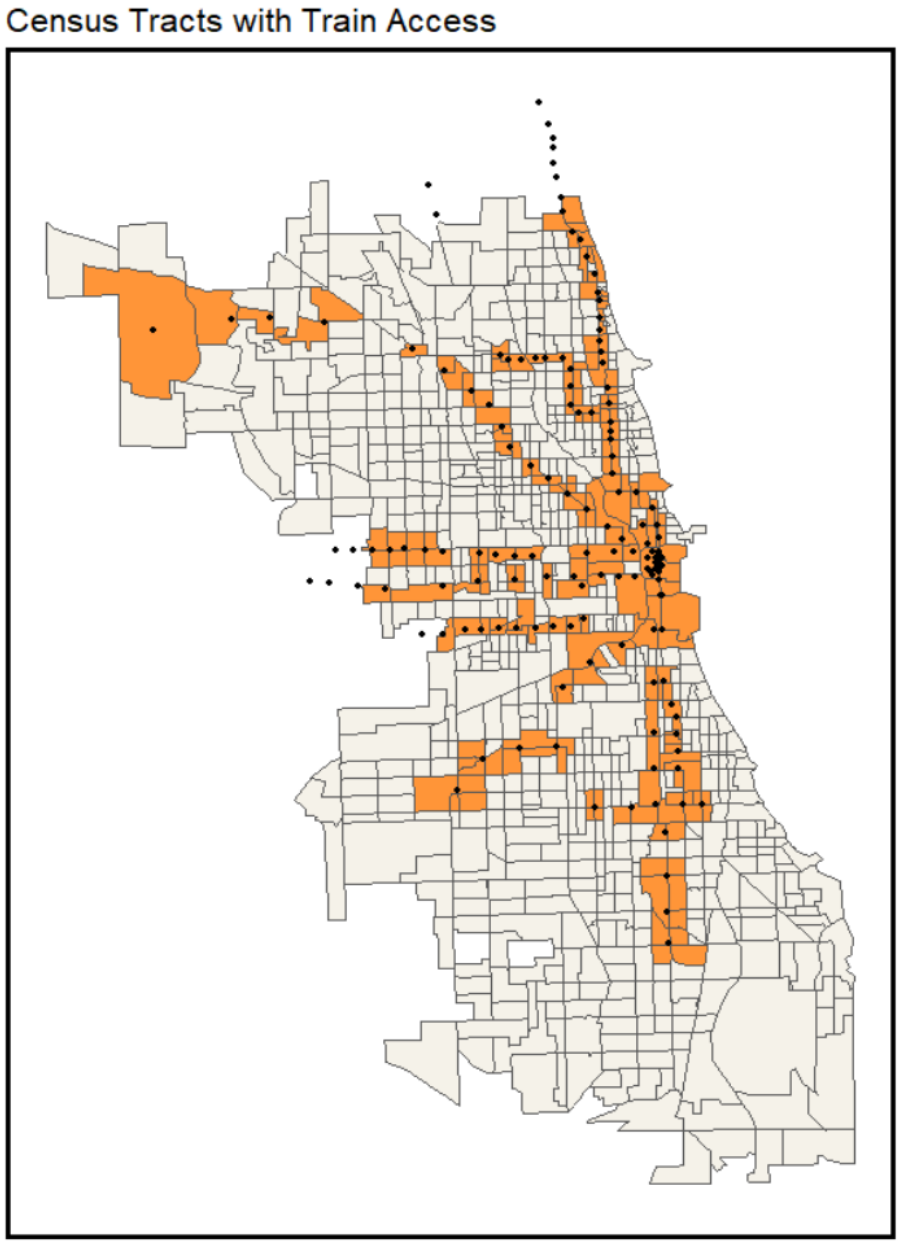
Data Processing:

Step 1:

Find census tracts accessible to CTA trains stations

Method:

Select any census tracts touched by a 100-meter buffer from each station



Data Processing:

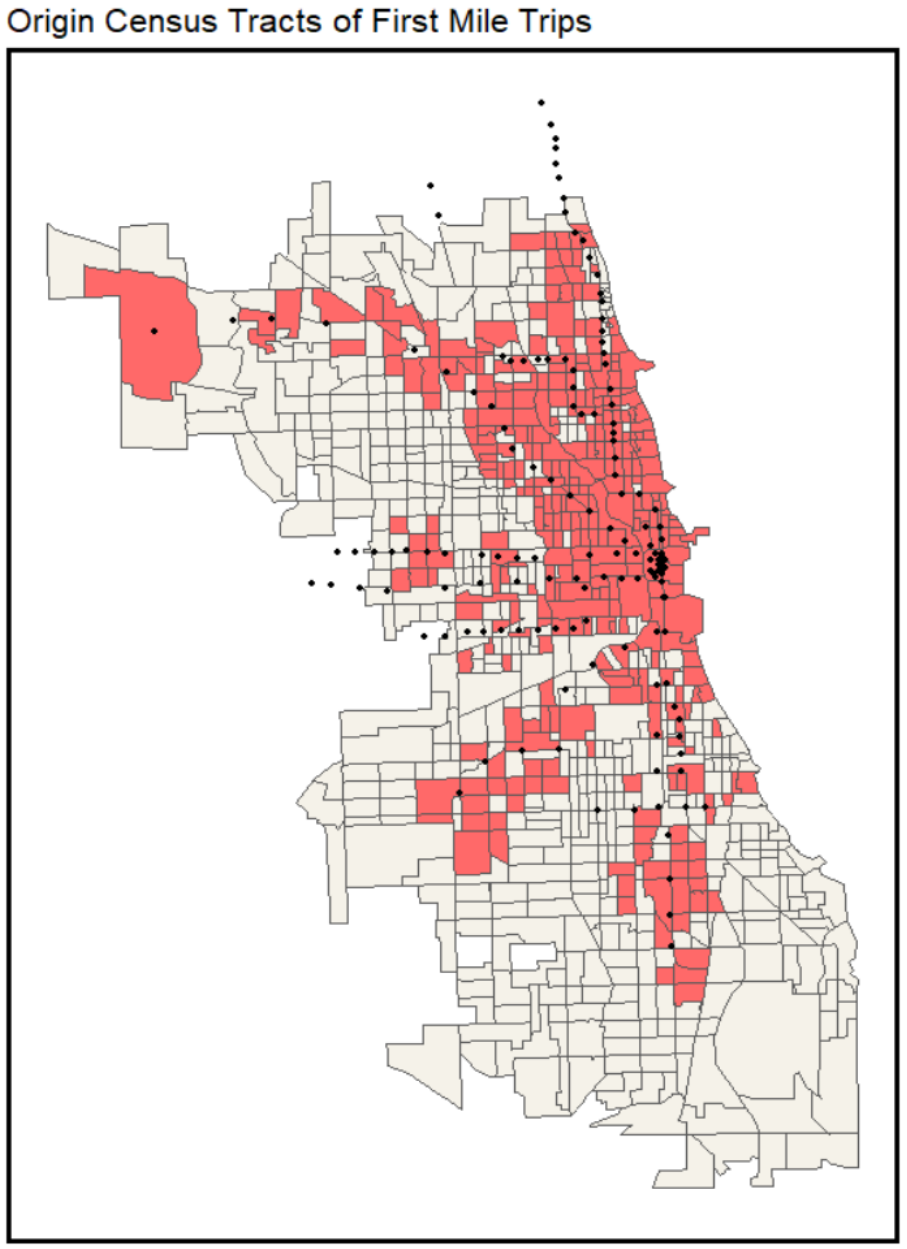
Step 2:

Find the origin census tracts of first-mile trips

Method:

1) Select trips shorter than 1.5 miles (network distance) and dropped off at station tracts

2) Retrieve their origin census tracts



Data Processing:

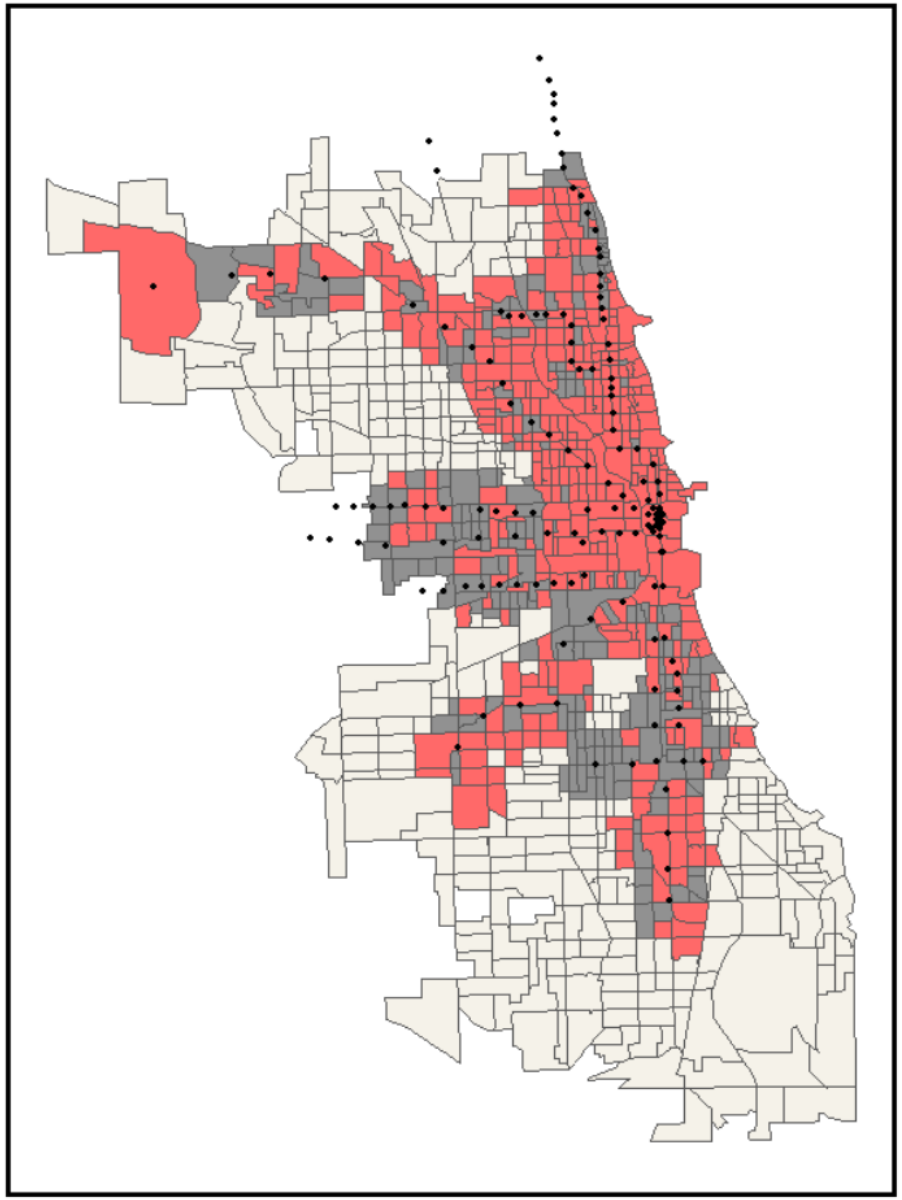
Step 3:

Find the census tracts which are close to train stations but never generated any first-mile trip using ride-hail services

Method:

Select all census tracts whose centroids locates within 1-mile buffers from stations

Census Tracts near Train Stations with Zero Ride-Hail First Mile Trip



Data Processing:

The data has census variables…

Each row is a census tract

	GEOID	TotalPop	MedRent	MedIncome
1	17031010100	4522	874	62177
2	17031010201	7039	1023	47411
3	17031010202	2852	973	51719
4	17031010300	6650	976	66875
5	17031010400	5153	1010	59861

Step 4:

Generate data to be used in regressions

Method:

1) Merge first-mile trip data and census data based on census tracts' IDs;

2) Combine census tracts had ride-hail trips and census tracts had no ride-hail trip.

… and ride-hail service variables

carpool	totalcost	pickup_num	carpool_ratio
0	24.50	3	0.00000000
2	56.67	8	0.25000000
0	12.50	2	0.00000000
1	5.00	1	1.00000000
0	0.00	0	0.00000000

The total number of first-mile ride-hail trips

The ratio of carpooled first-mile ride-hail trips

The final dataset for regression has 528 observations (census tracts)

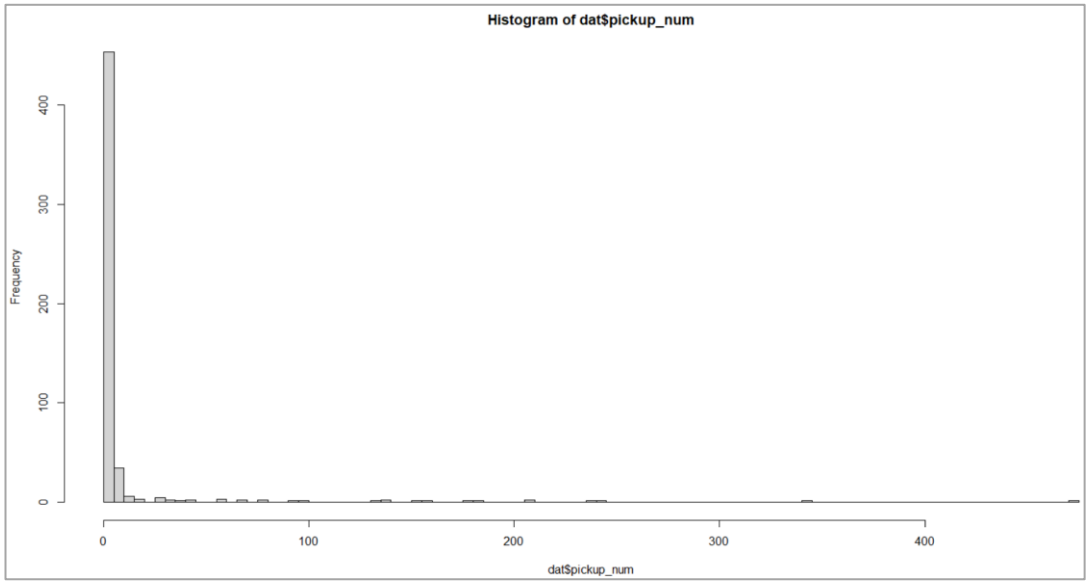
Research Question 1

What socioeconomic factors influence the number of first-mile ride-hailing trips originated at places near major transit stations?

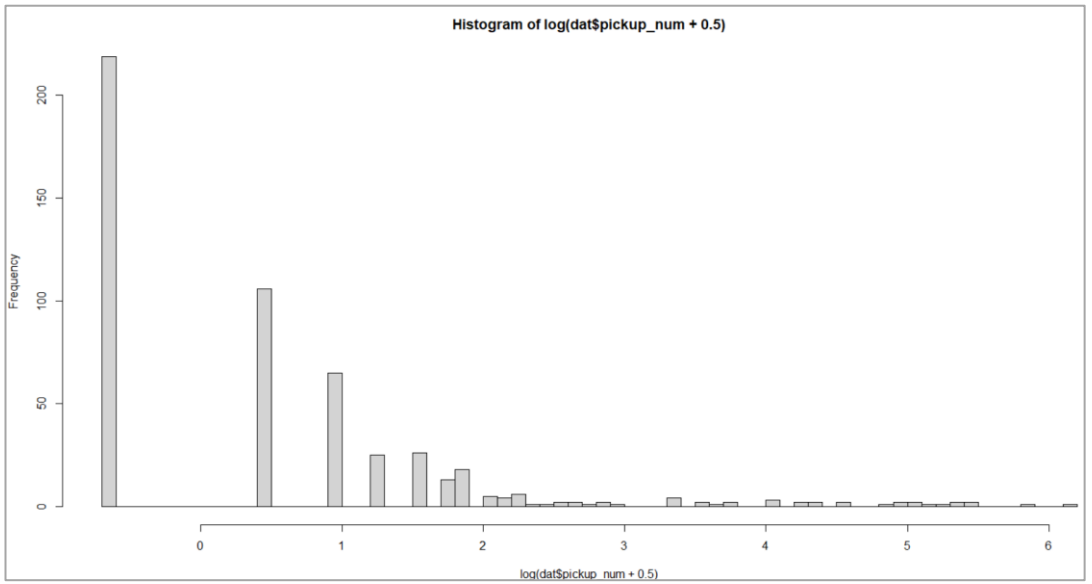
To Answer Research Question 1: Multivariate Linear Regression?

- Intuition: It will be simple...

- However, …
  - The dependent variable is extremely right-skewed
    - NOT normally distributed even after a log transformation
  - The number of trips is count data
    - a census tract cannot have 3.5 ride-hail trips



Distribution of number of trips



Distribution of log number of trips

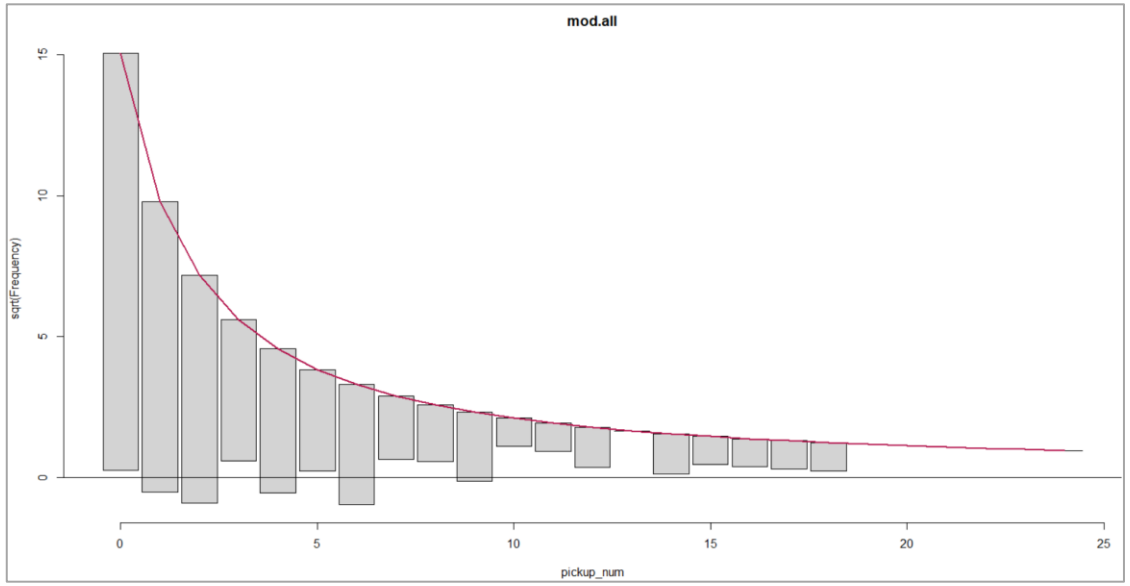
To Answer Research Question 1: Negative Binomial Regression

- Model 1: full model
  - Use all relevant census variables
- Model 2: trimmed model
  - Only use significant variables from the full model
- A likelihood ratio test showed…
  - NO significant difference between the two models

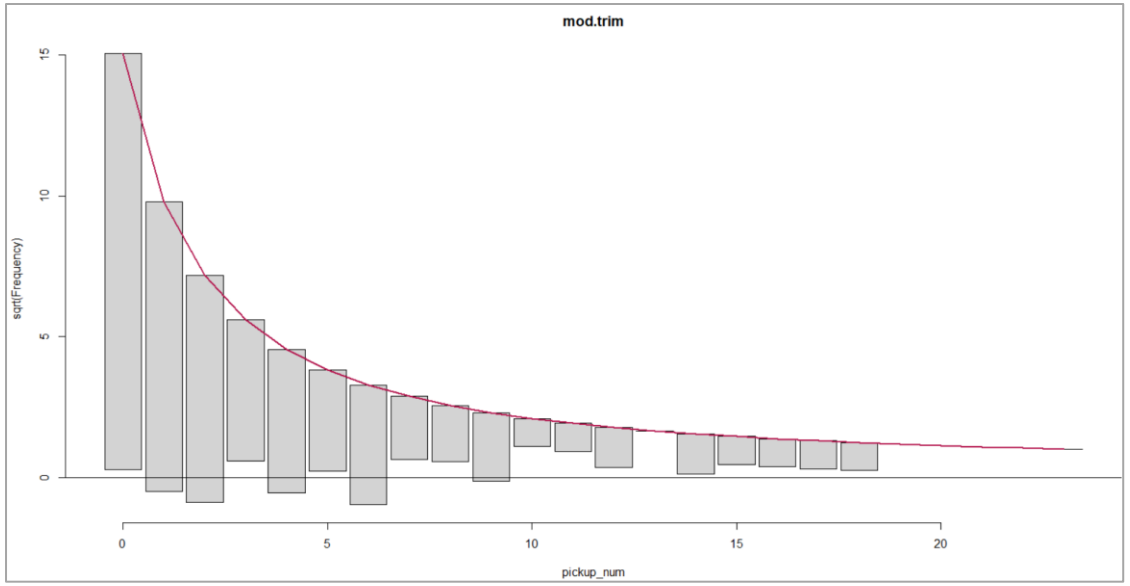
	Dependent variable:	
	pickup_num	
	(1)	(2)
MedRent	0.001***	0.001***
	(0.0003)	(0.0002)
MedIncome	-0.00000	
	(0.00000)	
MedAge	-0.011	
	(0.020)	
CollegeRate	3.962***	4.749***
	(1.236)	(0.721)
white_percent	-0.793	
	(0.669)	
AfricanAmerican_percent	-0.866	
	(0.578)	
as.factor(maleTofemale)1	-0.015	
	(0.127)	
Commute0_29mins_percent	5.761***	6.089***
	(0.567)	(0.506)
carpooled	0.002***	0.002***
	(0.001)	(0.001)
HH_ownership_rate	-2.086*	
	(1.136)	
NoVehicle_rate	2.204***	2.432***
	(0.809)	(0.480)
busstop_den	2,167.453	
	(5,846.572)	
Pop_Den	-9.199	
	(9.969)	
divorce_rate	3.750	
	(2.412)	
single_Rate	-0.149	
	(1.110)	
child_rate	-0.457	
	(0.452)	
mortgage_Rate	2.536*	
	(1.509)	
Constant	-3.488**	-4.844***
	(1.408)	(0.324)
Observations	528	528
Log Likelihood	-1,022.867	-1,028.995
theta	0.833*** (0.085)	0.795*** (0.080)
Akaike Inf. Crit.	2,081.734	2,069.990
Note:		*p<0.1; **p<0.05;** p<0.01

To Answer Research Question 1: Negative Binomial Regression

- Rootograms of the two models
  - Very similar results
- Picked Model 2 (trimmed model) since it used fewer variables



Rootogram of the full model



Rootogram of the trimmed model

To Answer Research Question 1: Negative Binomial Regression

- Interpretation Example (MedRent)
  - For a one-unit increase in a census tract's median rent, the expected log count of first-mile ride-hailing trips in this census tract will increase by 0.001, holding other conditions constant.

- Variables worth considering for ride-hailing demand:
  - Median rent,
  - % of people have bachelor's degrees,
  - % of commuters whose commuting time is less than 30 minutes,
  - number of people carpooled to work,
  - % of households do not own a car

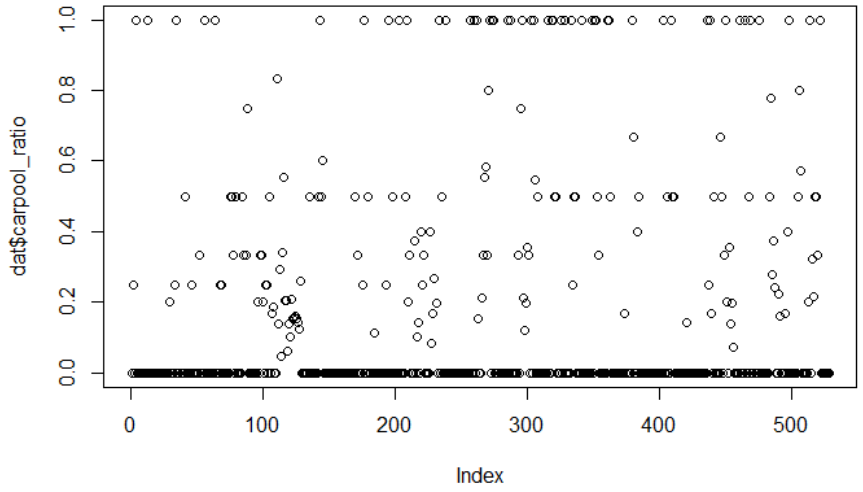
	Dependent variable:
	pickup_num
MedRent	0.001***
	(0.0002)
CollegeRate	4.749***
	(0.721)
Commute0_29mins_percent	6.089***
	(0.506)
carpooled	0.002***
	(0.001)
NoVehicle_rate	2.432***
	(0.480)
Constant	-4.844***
	(0.324)
Observations	528
Log Likelihood	-1,028.995
theta	0.795*** (0.080)
Akaike Inf. Crit.	2,069.990
Note:	*p<0.1; **p<0.05;** p<0.01*

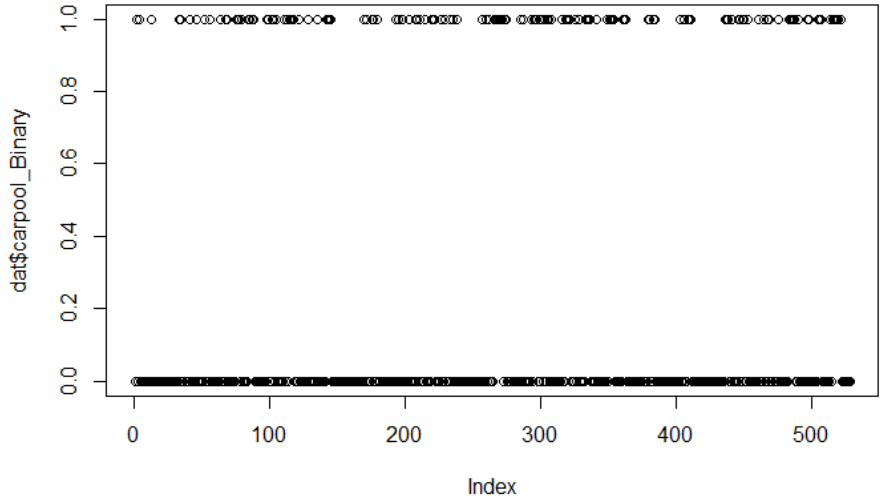
Research Question 2

What socioeconomic factors influence people’s willingness to carpool in first-mile ride-hailing trips?

To Answer Research Question 2: Binomial Logistic Regression

- The ride-hailing trips data documented whether a trip was authorized to be shared.
- Carpooled ride-hailing ratio in each census tract
  - number of authorized shared trips/number of total trips





To Answer Research Question 2: Binomial Logistic Regression

Step 1: Create a binary variable indicate carpool ratio
(carpool ratio < 0.2 -> 0, >0.2 -> 1)

Step 2: Create training and testing samples.
(75% training sample & 25% testing sample)

Step 3: Identify meaningful variables.

Step 4: Regression on all the possible variables.

Step 5: Regression on independent variables selected by stepwise function.

Step 6: Run ROC(C-statistics) to test the fitness of the model.

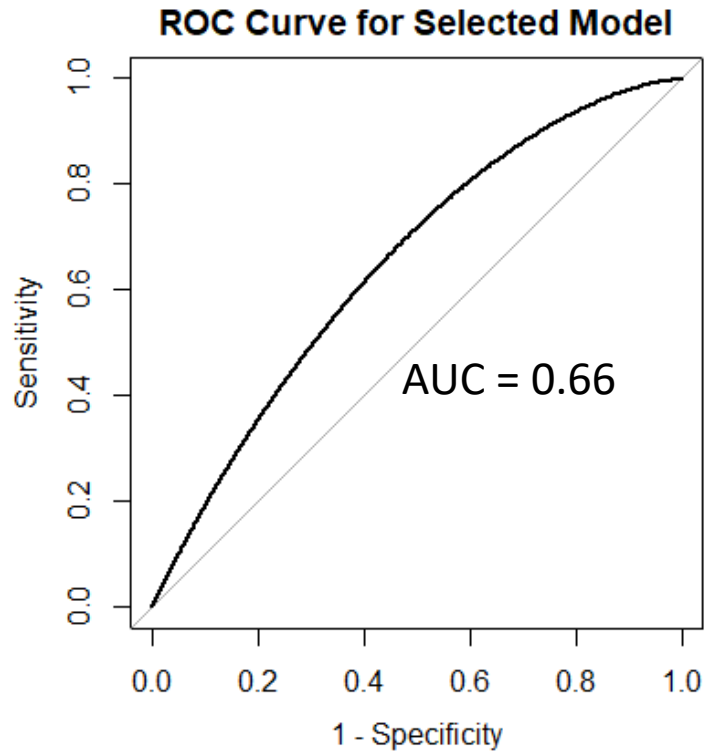
	Dependent variable:	
	carpool_Binary	
	(1)	(2)
TotalPop	0.00004	
	(0.0001)	
CollegeRate	2.483	4.120***
	(2.156)	(1.594)
Commute0_29mins_percent	3.616***	3.574***
	(1.073)	(0.990)
HH_ownership_rate	-1.395	
	(2.123)	
NoVehicle_rate	-0.843	
	(1.595)	
busstop_den	-9,020.633	
	(11,539.110)	
MedRent	-0.0003	
	(0.001)	
MedAge	-0.011	
	(0.038)	
MedIncome	-0.00001	-0.00001**
	(0.00001)	(0.00000)
carpooled	0.002	0.003**
	(0.002)	(0.001)
child_rate	-1.175	
	(0.901)	
single_Rate	1.628	2.079*
	(1.986)	(1.146)
mortgage_Rate	1.195	
	(2.775)	
Constant	-1.672	-3.952***
	(2.725)	(0.800)
Observations	369	369
Log Likelihood	-203.232	-205.022
Akaike Inf. Crit.	434.465	422.043
Note:	*p<0.1;* ***p<0.05;*** p<0.01	

To Answer Research Question 2: Binomial Logistic Regression

- Findings/Interpretations
- Variables lead to increase in probability:
  - Commute0_29mins_percent
  - College Rate
  - Carpooled (private vehicle)
  - Single Rate (unmarried/total population)
- Variables lead to decrease in probability :
  - Median Income

	Dependent variable:	
	carpool_Binary	
	(1)	(2)
TotalPop	0.00004	
	(0.0001)	
CollegeRate	2.483	4.120***
	(2.156)	(1.594)
Commute0_29mins_percent	3.616***	3.574***
	(1.073)	(0.990)
HH_ownership_rate	-1.395	
	(2.123)	
NoVehicle_rate	-0.843	
	(1.595)	
busstop_den	-9,020.633	
	(11,539.110)	
MedRent	-0.0003	
	(0.001)	
MedAge	-0.011	
	(0.038)	
MedIncome	-0.00001	-0.00001**
	(0.00001)	(0.00000)
carpooled	0.002	0.003**
	(0.002)	(0.001)
child_rate	-1.175	
	(0.901)	
single_Rate	1.628	2.079*
	(1.986)	(1.146)
mortgage_Rate	1.195	
	(2.775)	
Constant	-1.672	-3.952***
	(2.725)	(0.800)
Observations	369	369
Log Likelihood	-203.232	-205.022
Akaike Inf. Crit.	434.465	422.043
Note:	*p<0.1;* ***p<0.05;*** p<0.01	

To Answer Research Question 2: Binomial Logistic Regression



Confusion Matrix

		Actual Outcome	
		0	1
Predicted Outcome	0	114	0
	1	4	1

Accuracy Rate = 0.73

Suggestions for Future Researches

- Find better ways to identify first-mile ride-hail trips
- Survey riders' socioeconomic characteristics rather than relying on census tracts
- Use larger sample size from multiple days
- Consider other factors such as distances and existing bus routes