Who rides Uber anyway?

A census-tract level analysis and clustering of ride-shares for the city of Chicago during the era of the pandemic

Shakil Rafi[1] and Arna Nithila[2], *Group 1.*

Written in partial fulfilment
For the requirements of
CSCE 5063: Machine Learning
With
Dr. Ukash Nakarmi

[1] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, 72701
[2] Department of Civil Engineering, University of Arkansas, Fayetteville, AR, 72701

Who rides Uber anyway?
A census-tract level analysis and clustering of ride-shares for the city of Chicago during the era of the pandemic

Shakil Rafi and Arna Nithila

***Abstract*— *The COVID-19 pandemic has led to an unprecedented change in transportation, including shared mobility services. This study attempted to identify the user group of ride-share services by leveraging daily ride-sharing trip data for the year of 2020 associated with other socio-demographic and built-environment attributes of Chicago, Illinois. The study employed K-means clustering for user group segmentation. Results show: i) the cluster with the largest share of census tracts generate lowest average trips which is clearly an impact of the pandemic; ii) The high-income cluster generates short trip and coupled with high population, land-use, and employment density; iii) The low-income cluster generates longer trips coupled with diversity of land-use mx, employment and population density. Results of this study provide insights for policymakers and ride-sharing operators to ensure access to the services among the population irrespective of spatial diversity.*

Keywords—ridesharing, users' clustering, K-means, socio-demographic attributes, built-environment features

I. Introduction and Problem statement

Since the introduction of ride-sharing services nearly a decade ago, the expansion of shared mobility (e.g., Uber, Lyft, and Didi) have significantly impacted travelers' mobility patterns, transportation systems, and societies around the world [1-2] where the pandemic has added a new dimension to their overall usage pattern. Ride-sharing usage studies mostly address the usage patterns based on the built environment, land-use, socio-economic factors [3-6]. Only a handful of research focuses on ride-share user clustering which is crucial for gaining a deeper understanding of ride-sharing's role in the mobility system of urban areas. Soria et al. [7] identified ride-sharing user segments using k-prototype segmentation analysis based on ride-sharing trip data, weather data, transit, and taxi data. However, no study is found to date on identifying user segments based on rideshare trips, socio-economic and built environment attributes.

Therefore, the goal of the research is to gain a holistic understanding of ride-share users' characteristics by grouping them into different segments based on ride-share trips data, socio-economic (median income), and built-environment attributes (population density, employment density, land-use mix) with the help of K-means clustering analysis during the pandemic period.

II. Data and Methodology

A. Data

the TNC database is used which is available at the City of Chicago Data Portal [8]. The authors' downloaded data where pick-ups ranged from Jan 1, 2020, to Dec 31, 2020. This resulted in a CSV of approximately 13GB in size consisting of 49 million rows where each row corresponds to one trip.

Data was first read using only the columns GEOID, Trip Length, and Trip Duration. This resulted in a data frame approximately 49 million rows long and three columns wide. Once read, the data was again cleaned using Pandas' built-in dropna() method, which drastically reduced the size of the data frame down to approximately 29 million rows. It is not lost on the authors that half the data contains NaN's of some sort.

At this point, an exploratory data analysis is conducted by plotting the data of pickups against census tract data obtained from the US Census bureau. See Figure 7 in the Appendix. The data shows a remarkably high degree of variance, with the city center and Chicago O'Hare International Airport (ORD) and Midway International Airport (MDW) showing an extremely high number of pickups while most of the other regions show comparatively paltry levels of pickup. This is expected, ORD and MDW are large international airports, and it provides a "sanity check" for the authors' methods and the algorithms to be used. The final analysis will drop these two census tracts as they are outliers and may unduly bias the data. The second half of the data pertaining to built-environment and demographics is obtained from the Environmental Protection Agency's Smart Locations Database [9]. The total number of variables considered are as follows:

Table 1 Data description

Category	Variables	Unit
Trip Characteristics	Number of Pickups	Number
	Trip Duration	Seconds
	Trip Length	Miles
Socio-demographic Characteristics	Median Income	$
Built Environment	Population Density	People/Acre
	Employment Density	Jobs/Acre
	Land-use Mix	-
	Percentage of zero-car ownership	Percentage
	Distance to nearest transit-stop	Meters

We define land-use mix according to the EPA's definition:

$$D2a_{EpHHm} = \frac{-A}{\ln(N)}$$

Where: $A = \frac{HH}{TotAct}\ln\left(\frac{HH}{TotAct}\right) + \frac{E5_{Ret}}{TotAct}\ln\left(\frac{E5_{Ret}}{TotAct}\right) + \frac{E5_{Off}}{TotAct}\ln\left(\frac{E5_{Off}}{TotAct}\right) + \frac{E5_{Ind}}{TotAct}\ln\left(\frac{E5_{Ind}}{TotAct}\right) + \frac{E5_{Svc}}{TotAct}\ln\left(\frac{E5_{Svc}}{TotAct}\right) + \frac{E5_{Ent}}{TotAct}\ln\left(\frac{E5_{Ent}}{TotAct}\right)$

An exploratory data analysis is done first by describing the data using the built-in Pandas describe() method, the results of which are included in Table 2.

Table 2 Descriptive Statistics of Different Attributes

Numerical variable	Mean	Standard Deviation
Pickups	2.69e4	8.115e4
Trip Miles	6.89	2.64
Trip Seconds	1098.84	230.613
Median Income	84220	54337
Total Population	3679	1832
Population Density	31.174	27.338
Employment Density	11.4696	49.843
Percent Zero Car Ownership	0.2088	0.156
Land Use Diversity	1.535	7.612
Distance from Transit	279.24	112.575

B. Methodology

Dataset pre-processing and transformation

In this stage, the pickup data are first collated according to GEOID, to get mean trip-time and trip-distance. Data from the EPA is also collated but using Excel and using VLOOKUP to combine different instances of the same GEOID, which is then averaged. Instances of zeros are dropped especially as it does not make sense to have zero median income in a census tract, as are instances where the distance to transit is zero, instances where population density as zero and instances where the land-use diversity is zero as the formula clearly shows this cannot be the case. This also cleanly throws out the census tracts with ORD and MDW. The transformation is required for the purpose of data clustering using K-Means in the following stage.

The encoding is performed with the Pandas library. Finally, Pandas DataFrame was used to write the encoded dataset into another CSV file. The PCA approach was then used to reduce the dimensionality of the dataset. The idea underlying PCA, according to Manero, Rimiru, and Otieno [10], is to transform a dataset into a new one comprised of linearly independent variables called Principal Components. Furthermore, each component in the collection seeks to obtain as much variation from the data as feasible. According to Eriksson [11], the Principal Components are obtained by calculating the variance or spread of each variable as well as the correlation between variables in the original dataset. Our dataset consists of 10 features that are complicated and will negatively impact the clustering algorithm performance. Therefore, the number of principal components selected was 2 as shown in Figure 3, that is to plot the data points in a two-dimensional graph.

Using the Elbow Method and computing the Within Cluster Sum of Squares (WCSS) score, the optimal number of clusters was then calculated. Cui [12] stated the WCSS variable or score, which computes the difference between each cluster, with a lower WCSS value indicating more effective clustering. Cui further claimed that when the number of K increases, the WCSS score decreases, and K is chosen on the decline point, which is depicted as an "elbow."

Data clustering

In this stage, K-means clustering algorithm was executed using the Scikit-learn machine learning Python library. Several steps of the algorithm are as follows (Figure 1): (1) Specify the number of clusters based on Elbow method result. (2) Initialize K as the centroids randomly. (3) Calculate the Euclidean distance of every data point from every centroid in the space. (4) Allocate every data point to the nearest centroid based on the calculated distance. (5) Iterate until centroids and data points remain the same and fixed.

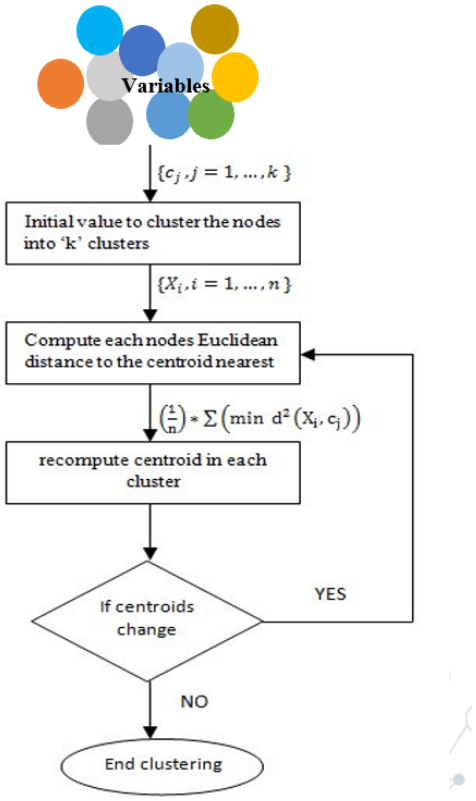

Figure 1 K-Means algorithm

Clusters analysis

The clusters are analyzed to understand the traits in each cluster. This process is executed by visualizing the trends between all the features for every cluster by using parallel co-ordinates plots (Figure 5 and 6). The clusters are analyzed first prior labeling. The analysis involves the following:

- Trip Characteristics analysis covers analyzing number of pickups, trip distance and trip miles in each cluster.
- Socio-Demographic analysis covers analyzing median income, total population and population density in each cluster.

●Built-environment analysis includes identifying the top segment subscribing for every product named in the dataset. This process is intended to label the segments by extracting the unique attributes in every cluster and to distinguish it from other clusters.

III. RESULTS AND DISCUSSION

To understand the relationships among the variables used for K-means clustering, a heat-map is generated (Figure 2). Several interesting trends are observed which will be borne out in the clustering. Median income has a positive correlation with pickups. It may be the case that ride-sharing is more affordable to people with a high income during the pandemic. This finding is consistent with Debnath et al. [13]. Besides, percent of zero ownership of cars in a tract has no association with the number of pickups.

Employment density has a strong correlation with pickup which is consistent with the previous study [13-14]. Indeed, it may be the case that ride-sharing is a function of one's work-life rather than a tool of leisure, at least in the city of Chicago. The primary use-case for rideshares in the city of Chicago seems to be to go to and from work and from one place to another as needed in one's job.

Population density (people/acre) somewhat correlates with more pickups, this is well known in the literature [15-17]. It is very strongly the case that the higher the median income in a tract the smaller the amount of time people spends in ridesharing for each given trip. It seems to indicate that with greater median income one takes shorter trips. There is a mildly negative relation between distance from transit and number of pickups. This might be the case that greater distance from transit corresponds to greater distance from city center which in turn means higher car ownership which in turn means less use of ride-sharing services. This again refers to the discussion above in that a more sophisticated analysis using distance of the centroid of each census tract from the town center could potentially yield a more holistic picture of what is happening, or at-least could yield a potentially large vector when PCA is done.

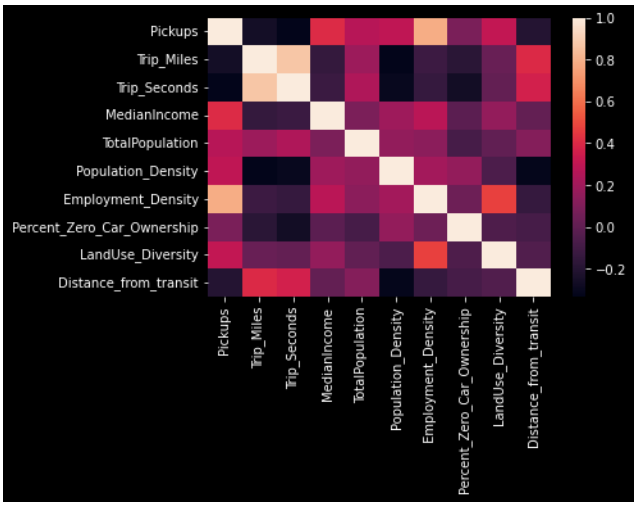

Figure 2 Heat-map of variables

Since all the data is quantitative, the K-means algorithm is tuned to select the optimal number of clusters during the estimation phase for clustering. This is determined by developing models including several clusters ranging from 2 to 10 and calculating the total cost across all observations. The final number of clusters chosen is three, based on interpretability of segmentation variables and guidance from the plot, which shows a clear "elbow" at 3 clusters. An elbow occurs when adding more clusters does not sufficiently improve the objective function. The clustering results are shown in Table 3 along with mean values of the explanatory attributes in each prototype.

A PCA analysis using the built-in Scikit-learn tools is employed to project the data and the centroids down to two dimensions. Projecting the data seems to clearly indicate three clusters with the centroid neatly being in the "middle" of all three clusters (Figure 3).

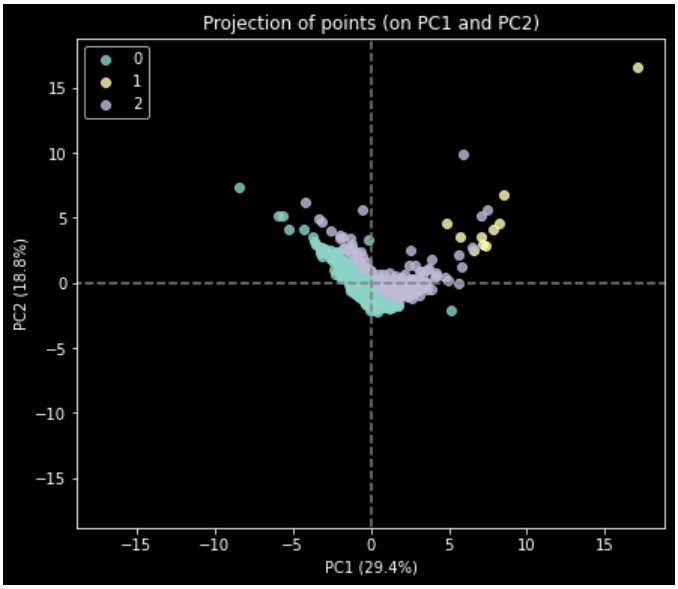

Figure 3 Projection of data to 2D showing three clusters

We now turn to summarize the contours (shown in Figure 5 and 6) of the three clusters. Overall, the analysis did not produce prototypes that were heavily differentiated by percentage of zero car ownership (Table 3). Several observations can be made of the clusters.

Cluster_0 (low ride share usage in low-income areas) is the largest cluster (72% of census tracts) and is characterized by its relatively small number of ride-share trips and longer travel times and distances. This long travel distance, averaging 7.25 mi (Table 3), is coupled with lowest median income compared to other two clusters. In terms of built environment characteristics, this cluster represents lowest population density, and employment density. The distinct nature of Cluster_0 suggests that census tracts with low-income users living in low density areas make small number of rideshare trips compared to other clusters for meeting their longer trip demands.

Cluster_1 (high ride share usage in high-income areas) is the smallest segment. The highest median income cluster goes the shortest distance on rideshares and spends the least amount of time on rideshare trips. Furthermore, Table 3 illustrates that trips in this cluster originated from areas with the highest employment, population densities with highest land-use diversities.

Cluster_2 (moderate ride share usage in middle-income areas) represents a moderate number of rideshare trips generated from areas with mid-ranged diversities. The average trip miles and duration are higher from cluster_1 but less from cluster_0 (Table 3).

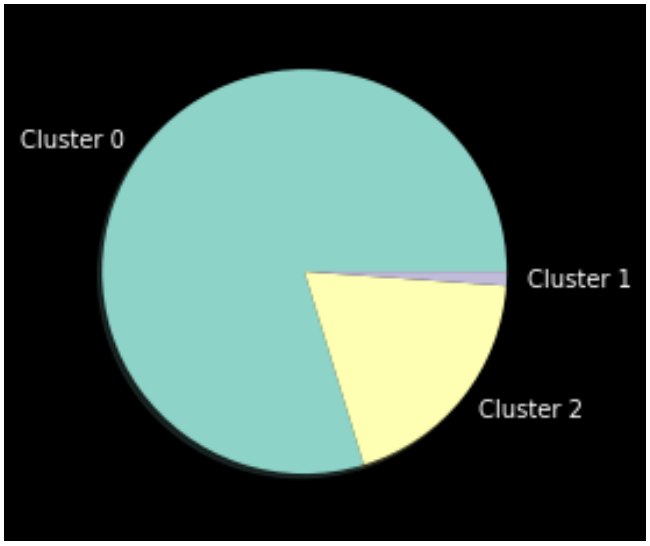

Figure 4 Distribution of different clusters

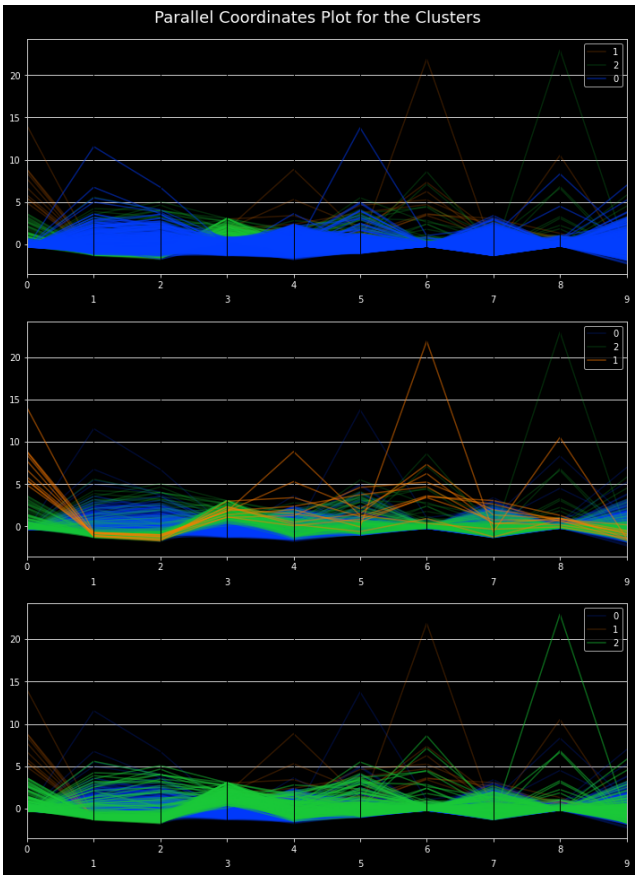

Figure 5 Parallel Plot of the Different Clusters

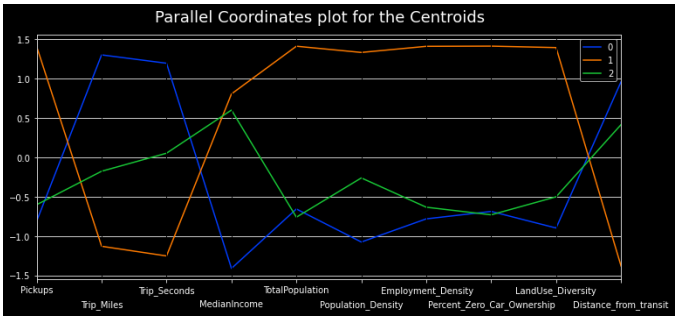

Figure 6 Co-ordinate plot of each cluster

Table 3 Different clusters and their attributes

Cluster	Cluster 0	Cluster 1	Cluster 2
Pickups	8348	644578	68404
Trip Miles	7.25	4.49	5.57
Trip Seconds	1131.59	805.81	979.50
Median Income	61194.45	185783.90	174275.93
Total Population	3672	8360	3436
Population Density	27.76	72.69	42.95
Employment Density	4.39	297.06	24.19
Percent_Zero_Car_Ownership	0.21	0.34	0.21
LandUse_Diversity	0.94	14.44	3.27
Distance_from_transit	285.51	173.05	259.34

From Table 3 and Figure 5 and 6, the out-sized discrepancy of employment density is also noted. The difference between median income of the high and medium income is less than ten percent but the employment density between the high- and medium-income tracts is ten-fold. If anything, this indicates that the variance in employment density for the city of Chicago is large. For instance, that there is less than a 20% difference in the percentage of zero car ownerships between the three clusters. Indeed, Cluster_1 (the highest income cluster) also has the highest percentage of zero car ownerships, indicating that, at least for the city of Chicago, a high income does not necessarily track with owning a car. This tracks well with data already known for the city of New York, where the highest income residents, mostly those from Manhattan tend to own fewer cars than those in less high-income boroughs such as Brooklyn or Bronx, even if by a small margin.

One reason could be that those census tracts with the highest median income are on average the closest to transit. However, this could again be a function of proximity to city center. For instance, closeness to city center means that one is closer to transit stops, and closeness to city center means a higher income simply because higher housing prices tend to self-select for individuals with high median incomes.

This conclusion is again mediated by the fact that in the higher median income cluster the population density is the highest. Exploratory data analysis and intuition shows that it is strongly the case that closeness to city center corresponds highly with population density and with higher median income.

Finally, two more observations warrant special consideration. Cluster 0, the so-called “low income” accounts for the largest segment of the population. While this is called “low” in our analysis this corresponds to what would be called middle-class in demographic analysis, there is a strong middle-class in Chicago. Secondly the youngest median age corresponds to the middle-income group, the slightly higher age group for the highest income group represents a young urban, affluent youth demographic for Chicago. However, the analysis of that is beyond the scope of this report.

The cluster analysis sheds some important insights on the access to rideshare services among the population. For instance, it is found that census tracts with low-income areas generate less trips than the areas with high median income. Past research also found that more than 70% of American with high income use ride-hailing services compared to other income groups [18]. Apart from that, it is also found that percent of zero car ownership has no significant differences within the clusters. Yet it is assumed that people with no car ownership tend to use more ride-share services, but this study shows a different result. Besides, spatial differences in accessibility to ride-share services is also apparent in this study. As evident, areas with high land use diversity and employment diversity generate more trips. All these finding indicates inequity of transportation services both spatially and demographically.

IV. CONCLUSION

This study attempted to conduct a user cluster analysis using machine learning tools. We employed K-means clustering with detailed descriptive analysis to fulfill the goal of the study which is examining characteristics of the ride-share users using their socio-demographic and built environment characteristics and trip attributes during pandemic in the city of Chicago.

The heat-map produced from the features show that the features have high correlation with each other. Like previous studies, median income, population density, employment density, land-use mix show a positive correlation with number of ride-share trips. However, percent of zero car-ownership does not show any correlation with the trip data. Besides, unlike past studies, distant to transit has a slight negative correlation with trips which might be a future revenue for study.

Most importantly, the cluster analysis shows that the ride-share services are more accessible to high-income census tracts compared to the low-income ones. Besides, a spatial difference in ride-share usage pattern is also observed through clustering analysis. This finding raises the question about the fair and equitable distribution of transportation services among the communities.

Apart from those two avenues of research present potential extensions of this study. Firstly, we have focused on a “snapshot” of data for the year 2020. It will be interesting to see change, if any, between the year of the pandemic, 2020, pre-pandemic, 2019, and post-pandemic 2021. A time-series analysis between the low, medium, and high-income tracts to see if the post-pandemic recovery was the same between all three will be interesting. Is it, for example the case that higher income neighborhoods recovered faster or slower than lower-income neighborhoods, and is income the best indicator for recovery, or is it something else, such as employment density? It is also worthwhile to see do a finer-grained analysis of the

yearly data, say at a monthly or weekly scale. This allows us to zoom in in March 2020 and analyze the impact that the global shutdown has had on the different census tracts of Chicago. Is it the case, for example that higher income census tracts had a marked decrease in pickups simply because the nature of the work allowed individuals to work from home than some other forms of employment? Secondly the authors would like to tease out the exact relationship between car-ownership and pickups. It is strongly suspected that this is mediated by a third variable, highly likely distance from city center. It would be interesting to see what distance from city center says about car-ownership, distance from transit and finally about the number of pickups. Much remains to be explored in this front.

This study is not without limitations. User clustering is conducted based on the ride-sharing trips' pick-up data only. Future research should employ origin-destination pair flows rather than just using trip pick-ups. Moreover, the study did not include other socio-demographic variables like median age, gender, marital status, and other built-environment and transit variables. The inclusion of these variables could have a better understanding of users' characteristics.

Overall, the study offers novel findings on the users' characteristics of ridesharing during the pandemic. The findings aid in understanding the importance fair and equitable access of ride-share services among the communities irrespective of spatial differences.

ACKNOWLEDGMENT

We thank Dr. Ukash Nakarmi and our classmates for the thoughtful comments that contributed toward improving our manuscript.

REFERENCES

[1] Schaller, B., 2018. The New Automobility: Lyft, Uber and the Future of American Cities. Brooklyn, NY.

[2] Alemi, F., Circella, G., Handy, S., Mokhtarian, P., 2018a. What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. Travel Behav. Soc. 13, 88–104. https://doi.org/10.1016/J.TBS.2018.06.002.

[3] Lavieri, P.S., Dias, F.F., Juri, N.R., Kuhr, J., Bhat, C.R., 2018. A model of ridesourcing demand generation and distribution. Transp. Res. Rec. https://doi.org/10.1177/ 0361198118756628.

[18] Jiang, J. (2021). More Americans are using ride-hailing apps. https://www. pewresearch.org/fact-tank/2019/01/04/more-americans-are-using-ride-hailing-apps/

[4] Yu, H., Peng, Z., 2019. Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted Poisson regression. J. Transp. Geogr. 1–17 https://doi.org/10.1016/j.jtrangeo.2019.01.004.

[5] Correa, D., Xie, K., Ozbay, K., 2017. Exploring the taxi and uber demands in New York City: an empirical analysis and spatial modeling. In: Transportation Research Board 96th Annual Meeting.

[6] Varone, L.R., 2018. Understanding Spatiotemporal Growth of Ride Source Services In New York City. University of Connecticut.

[7] Soria, J., Chen, Y., Stathopoulos,A., 2020. K-Prototypes Segmentation Analysis on Large-Scale Ridesourcing Trip Data, Transportation Research Board 2020, DOI: 10.1177/0361198120929338

[8] City of Chicago Data Portal, 2021. Transportation Network Providers - Trips.

[9] Smart Location Database. https://www.epa.gov/smartgrowth/smart-location-mapping

[10] Manero, K. M., R. Rimiru, and C. Otieno. 2018. Customer behavior segmentation among mobile service providers in Kenya using K-means algorithm. International Journal of Computer Science Issues (IJCSI) . 15 (5):67–76. doi:10.5281/zenodo.1467663.

[11] Eriksson, L. (2020, August 18). What is principal component analysis (PCA) and how it is used? Umetrics. Retrieved from https://blog.umetrics.com/what-is-principal-component- analysis-pca-and-how-it-is-used

[12] Cui, M. 2020. Introduction to the K-means clustering algorithm based on the elbow method. Accounting, Auditing and Finance 7. doi:10.23977/accaf.2020.010102.

[13] Debnath, K, A., Ogunbureau, A. , Mitra, S,K., (2021). Exploring the ride-share usage pattern during the era of the pandemic, Transportation Research Board 2021

[14] Yang, Z., Franz, M.L., Zhu, S., Mahmoudi, J., Nasri, A., Zhang, L., 2018. Analysis of Washington, DC taxi demand using GPS and land-use data. J. Transp. Geogr. 66, 35–44. https://doi.org/10.1016/j.jtrangeo.2017.10.021.

[15] Lavieri, P.S., Dias, F.F., Juri, N.R., Kuhr, J., Bhat, C.R., 2018. A model of ridesourcing demand generation and distribution. Transp. Res. Rec. https://doi.org/10.1177/0361198118756628.

[16] Yu, H., Peng, Z.-R., 2020. The impacts of built environment on ridesourcing demand: a neighbourhood level analysis in Austin, Texas. Urban Stud. 57, 152–175. https://doi.org/10.1177/0042098019828180.

[17] Brown, A., 2019. Redefining car access: ride-hail travel and use in Los Angeles. J. Am. Plan. Assoc. 85, 83–95. https://doi.org/10.1080/01944363.2019.1603761.

Appendix of Maps

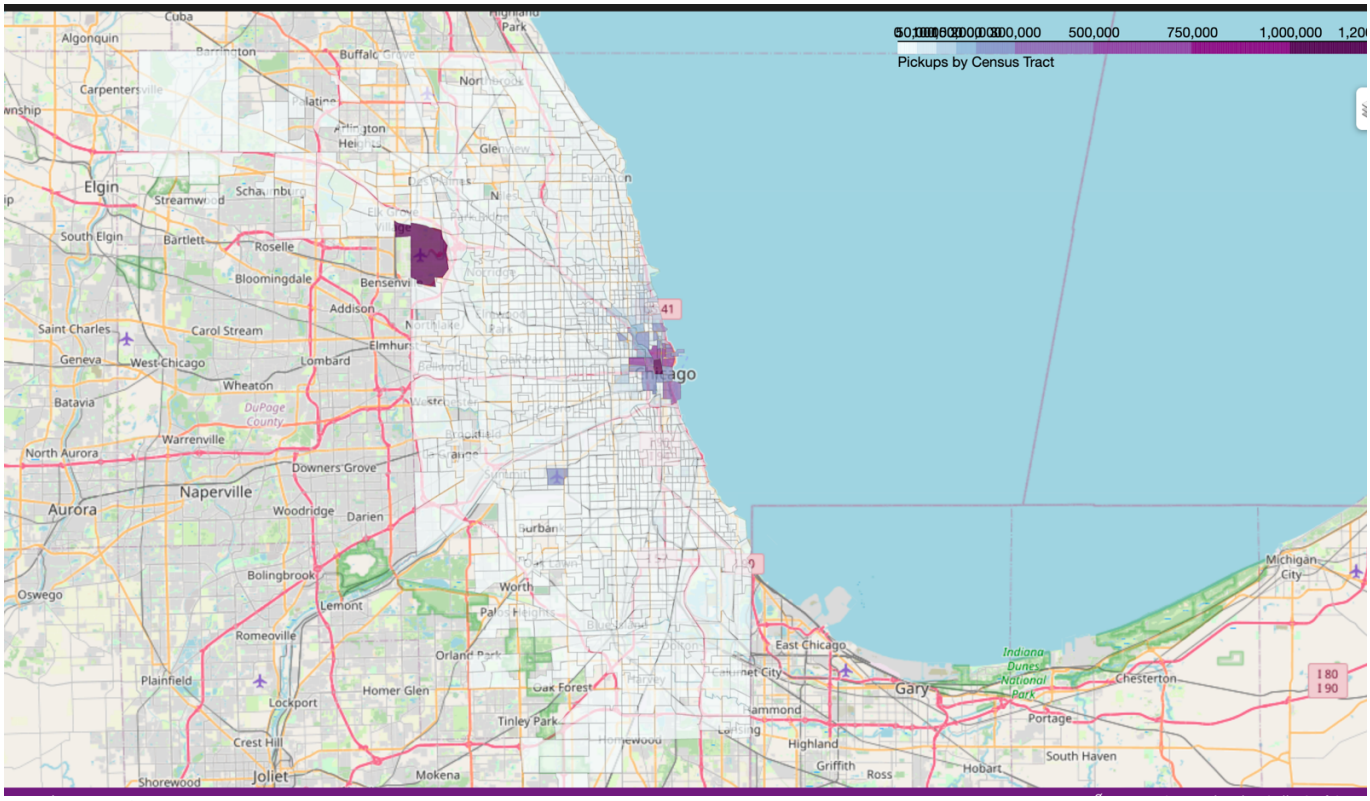

Figure 7 Pickups vs Census Tract

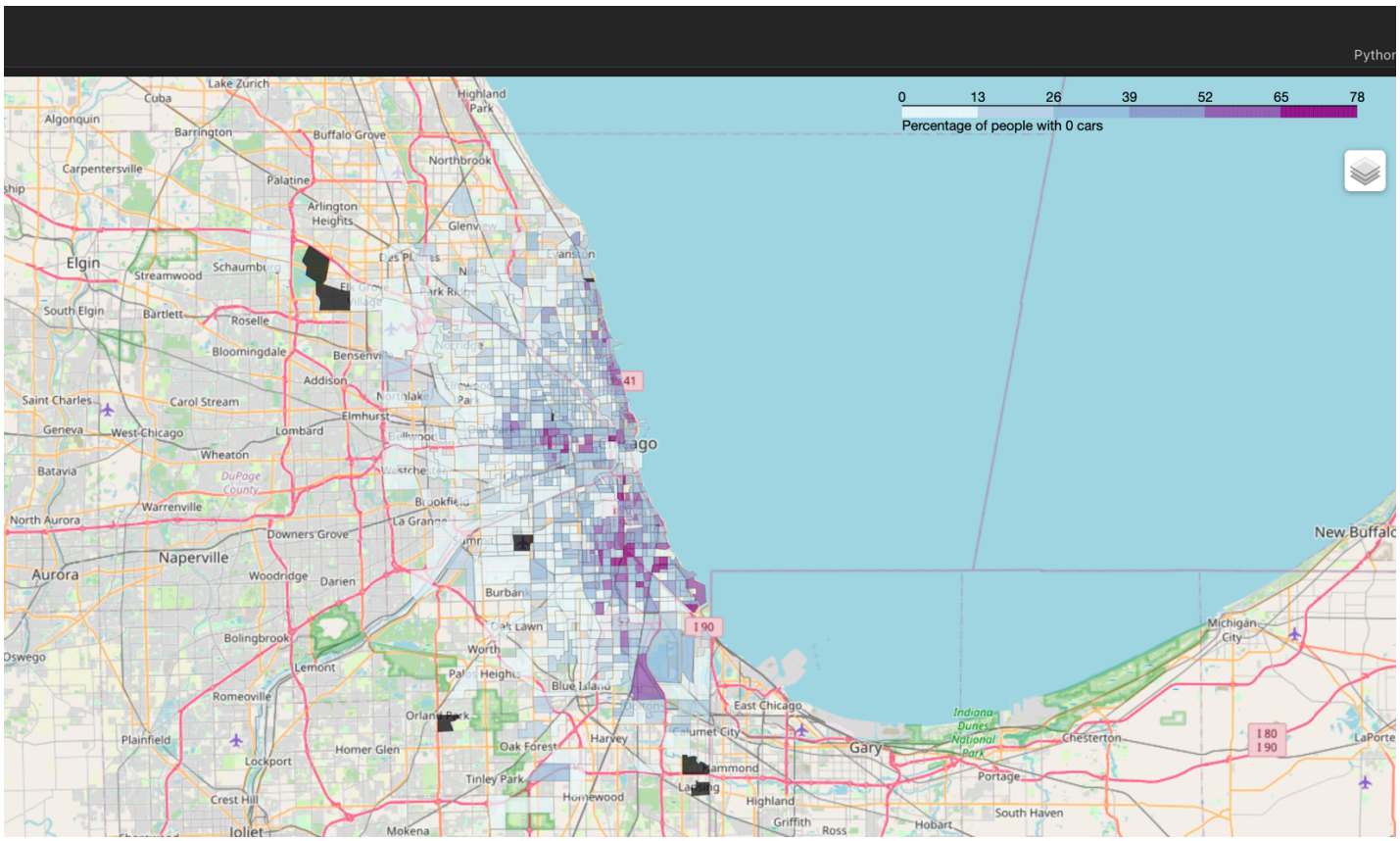

Figure 8 Percentage of zero cars vs census tract

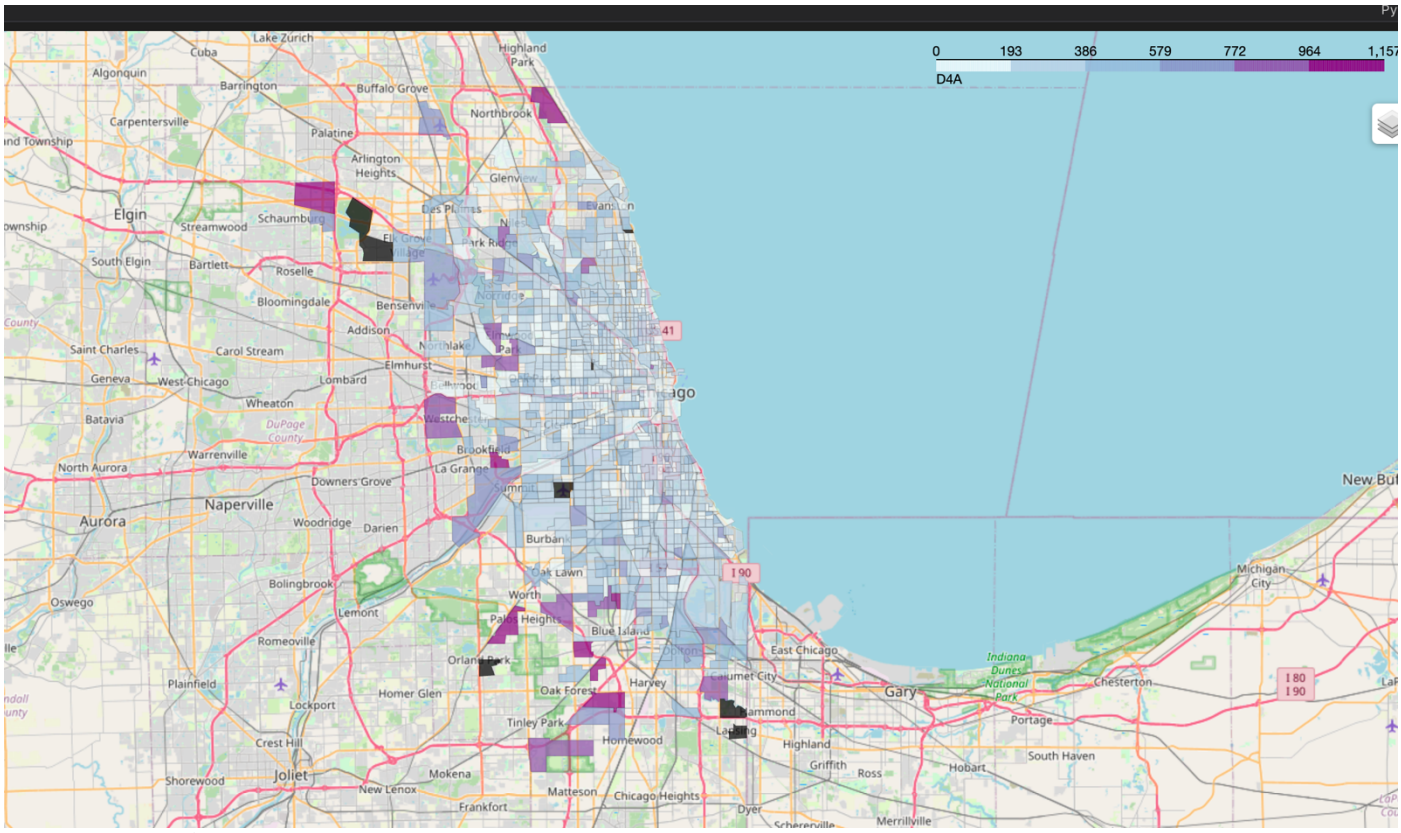

Figure 9 Distance from transit vs census tract

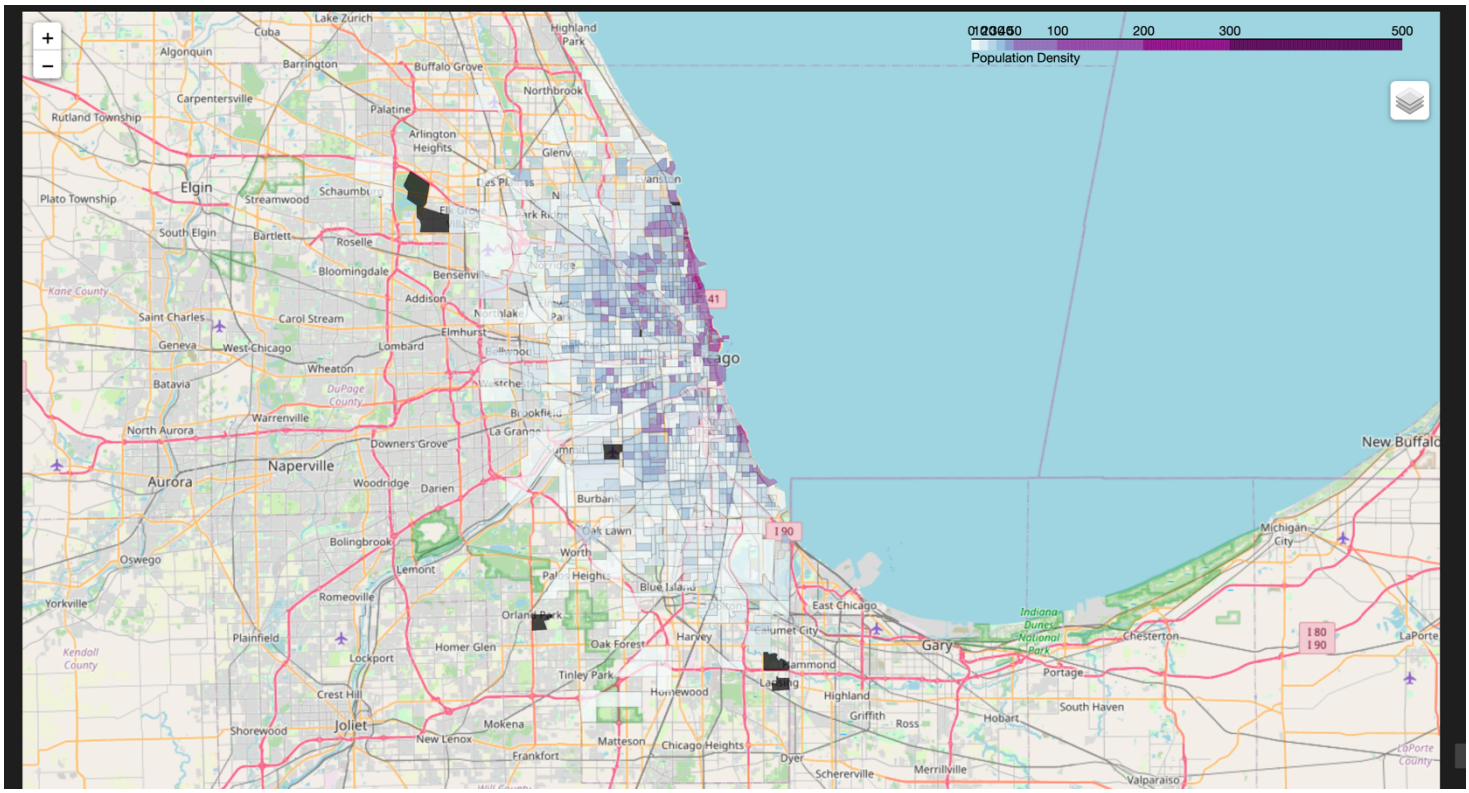

Figure 10 Population density vs census tract