# A Census-Tract Based Analysis of Rideshare using clustering for the City of Chicago in the era of pandemic.

*Shakil Ahmed Rafi*, Department of Mathematical Sciences, University of Arkansas, sarafi@uark.edu
*Arna Nishita Nithila*, Department of Civil Engineering, University of Arkansas, annithil@uark.edu

## Objectives

The authors take inspiration from Soria, Chen & Stathoupolous 2020[1] in trying to use K-means clustering to segment ride-sharing data.

Unlike the previous authors the present authors seek to segment ridesharing data using K-means using built-environment and demographic features for the year 2020, the sample size is also much larger.

It is suspected that the segmentation will fall strongly along economic lines.

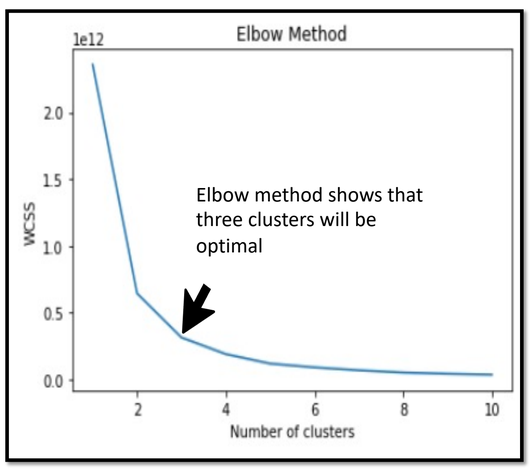
## Tools, Methods, & Data Sources

Since all the data is quantitative the authors seek to use K-means clustering to classify the data. They also use the standard "elbow method" to find out the number of Ks to use in the clustering. The "elbow method" finds out the MSSE for each value of K and plots them on a graph. The optimal K is the first instance where one observes a large change in slope from very negative to slightly negative. The authors observe it at K=3.

The authors also use the built in K-Means algorithm in Scikit-learn.

Data for pickups was first downloaded from the City of Chicago Data Portal[2] as a csv file which was read with Pandas, and from Smart Location Database by EPA[3].

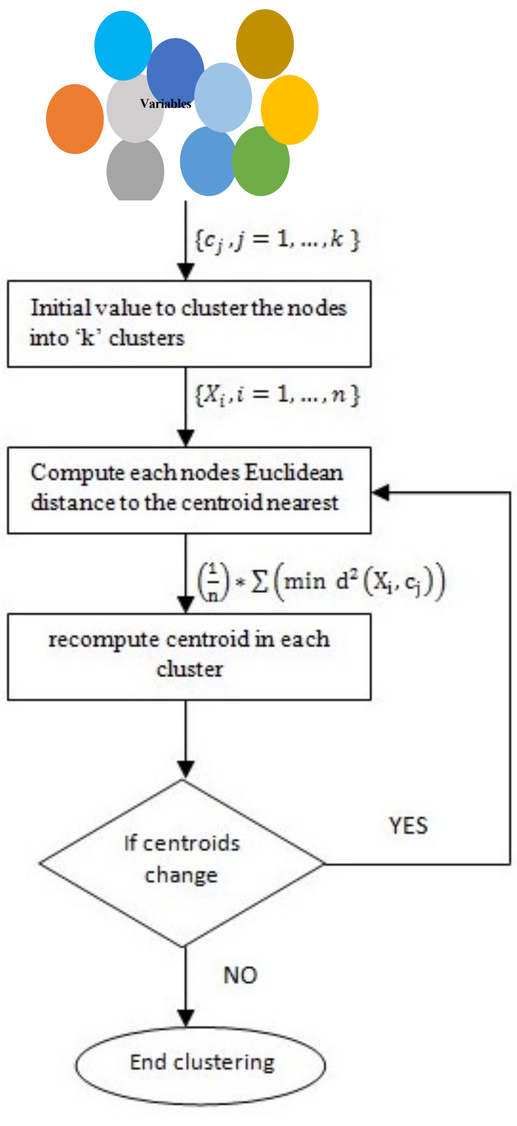
The data was cleaned by dropping NaN's, and data from the Smart Databases was collated using Pandas pd.groupby.mean().




A preliminary test-run showed a marked elbow in WCSS error at K=3, where WCSS is defined as:

$$WCSS(k) = \sum_{j=1}^{k} \sum_{x_{i \in cluster}} ||x_i - x^*|\,|^2$$

Whereas K-means is defined as[4]:



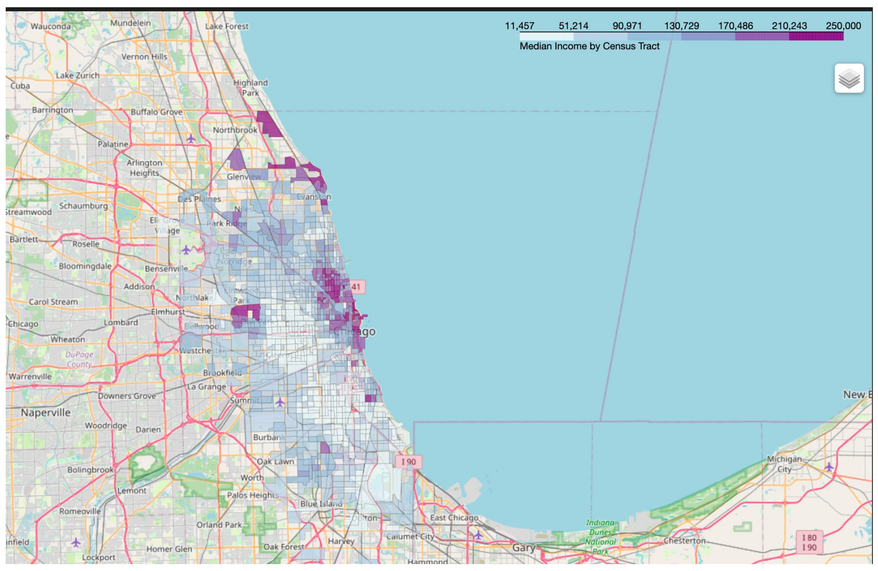

## Exploratory Data Analysis

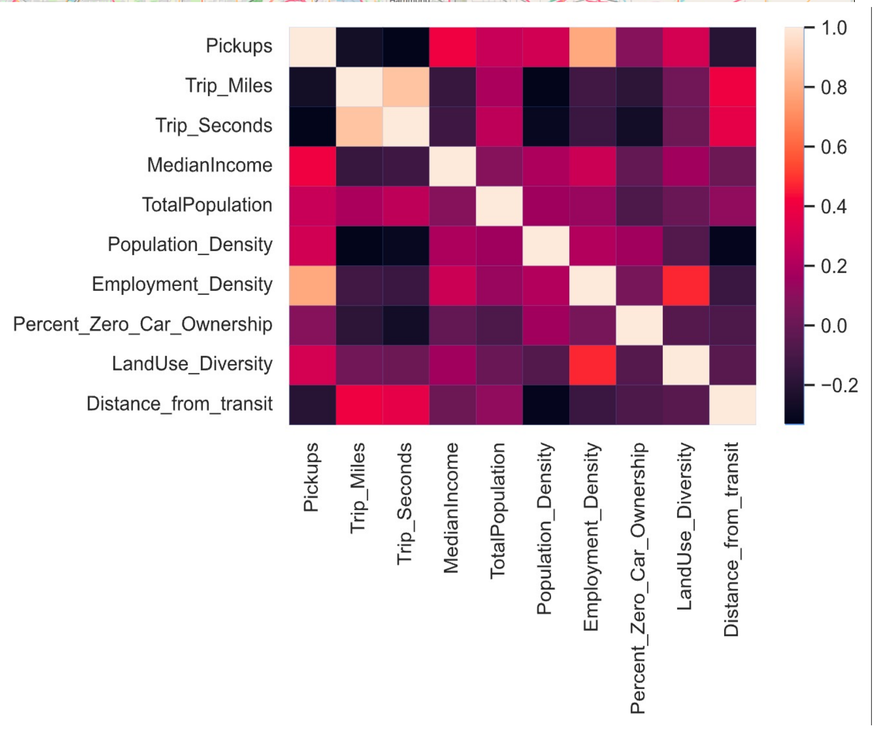
The authors first perform some exploratory data analysis. Choropleth maps were drawn for different variables against the census tract information using the Folium library. GeoPandas was used to store the geometries for the census tracts. We show one of these, median incomes which are of particular interest to the authors.

The distribution of median income to be expected, the neighborhoods of Lake View and Wrigleyville are well known to be affluent. Melrose Park is also known to be rather affluent and as well there are pockets of high-median income in the north.

We notice several interesting trends in the heat-map:

- Median Income correlates somewhat strongly negatively with pickups. It may be the case that ride-sharing is a rich person's game.
- Employment density is rather highly correlated with pickup.
- Ownership of cars in a tract is rarely an indicator of the number of pickups
- Population density (people/acre) somewhat correlates with more pickups
- It is very strongly the case that the higher the median income in a tract the smaller the amount of time people spend in ride-sharing for each given trip. It seems to indicate that with greater median income one takes shorter trips, perhaps within the same block.
- There is a mildly negative relation between distance from transit and number of pickups. This presents a future avenue of research. The hypothesis the authors propose is 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. Further research will bear this out, the authors hope.




## Results

Notice the three clusters with their centroid plotted (normalized) on the right. For exposition purposes the authors name the centroids:

**0. "Middle Median Income"** cluster, with a median income of around $191,000 in green.
**1. "Low Median Income"** cluster with a median income of around $108,000 in blue.
**2. "High Income"** cluster with a median income of around $49,000 in orange.

**Their characteristics are as follows:**

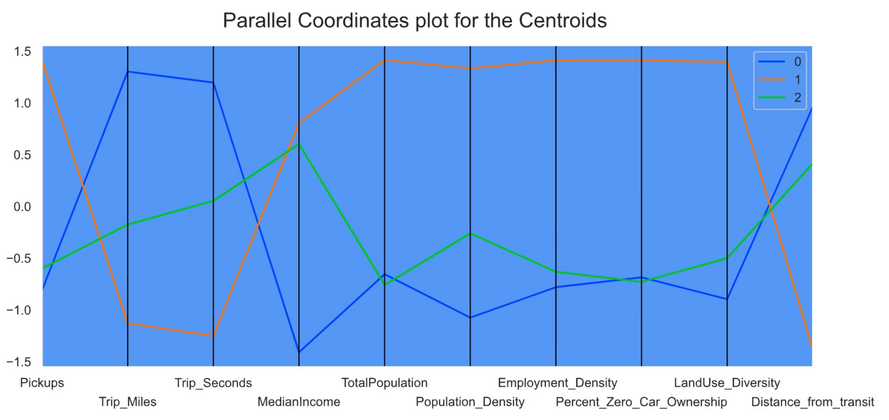
There is the **high-income group** consisting of people around median income $185k, living close to transit who around travel 3.86 miles, and where 33% don't own cars.

There is the **medium income group** consisting of people around median income $174k, living a bit further to transit who around travel 4.7 miles, and where 21% don't own cars.

There is the **low-income group** consisting of people around median income $61k, living close to transit who around travel 5.36 miles, and where 21% don't own cars.

Notice also the line plot of the three centroids, and the bar-plot of the percentage of people in each cluster. We realize that the most common is cluster **0** followed by cluster **1** and followed by cluster **2**.

To verify the result, the authors do principal component analysis. They notice that the data very clearly can be split into three parts. They then have these plotted onto two dimensions from nine dimensions to two principal dimensions and the authors see a clear demarcation into three clusters.

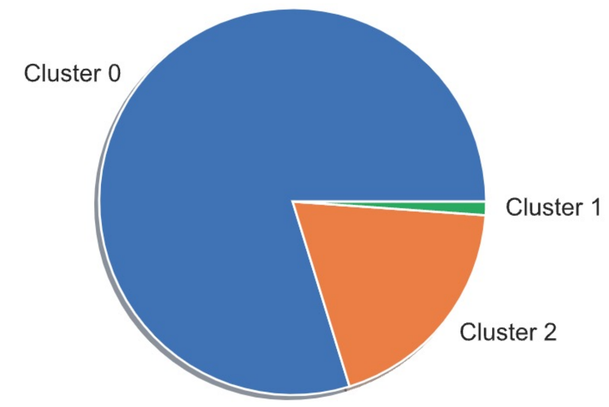



| | Pickups | Trip_Miles | Trip_Seconds | MedianIncome | TotalPopulation | Population_Density | Employment_Density | Percent_Zero_Car_Ownership | LandUse_Diversity | Distance_from_transit |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 8348.44992947817 | 7.248445411863000 | 1131.58803383858 | 61194.454160789900 | 3672.187588152330 | 27.764838912593800 | 4.3926292647701000 | 0.20788289459379400 | 0.9366700061480960 | 285.5142184159800 |
| **1** | 644577.6 | 4.490997429025810 | 805.8118093923320 | 185783.90000000000 | 8360.400000000000 | 72.68572882800000 | 297.063445252 | 0.33921535720000000 | 14.4384502384 | 173.04891667000000 |
| **2** | 68403.8705882353 | 5.572115030600570 | 979.5037149422680 | 174275.92941176500 | 3435.9764705882400 | 42.95103079969410 | 24.18524445075290 | 0.20517524539411800 | 3.272948279152940 | 259.3384166682350 |

The authors notice that percentage of zero car ownership is relatively the same throughout the three clusters. Indeed, the defining factor seems to be employment density and median income, indeed to the extent that one can say that pick-ups have predictors it seems to be land-use diversity, median income and employment density.

Somewhat counter-intuitively, higher income people cluster people tend to spend less time on ride-sharing and go less distance. Their percentage of zero car ownership is higher as well.

Interestingly the bar chart of percentages in each cluster shows a rather strong "middle class" with median incomes around $108k. Finally, low-median income tracts are on average further from the nearest transit stops although not by much compared to the other clusters.



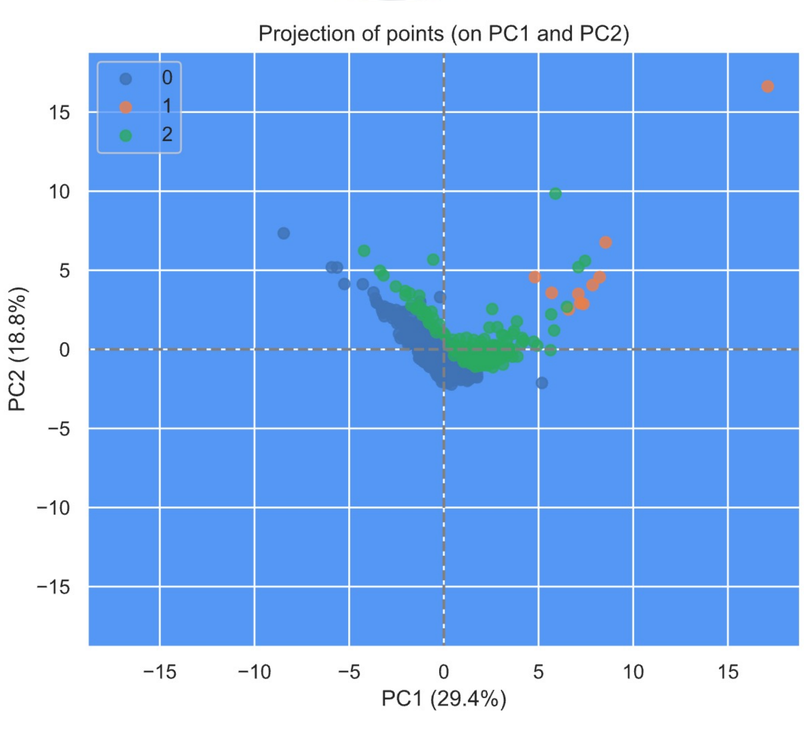

## Further Research

While at present the authors have focused on a "snapshot", it will be interesting to do a time-series analysis of the data. This is especially true given that the authors are working with 2020. It is expected that sharp decrease in ride-sharing is expected around mid-March 2020, due to the pandemic.

The authors expect the ride-sharing decline to be concentrated around tracts with low to medium median income, but again that presents another fruitful avenue of research.

It would also be worthwhile to see the recovery of ride-sharing post-pandemic (2021-onwards) and to see whether higher median-income tracts bounced back to 2019 levels faster than low-medium tracts.

The surprising conclusion that car ownership has little to do with number of pickups also deserves some special attention. In other words, is it the case that the relation between car-ownership and number of pickups is mediated by a third variable, perhaps by distance from city-center? Much remains to be explored.




## References

1. Soria, J., Chen, Y., & Stathopoulos, A. (2020). K-Prototypes Segmentation Analysis on Large-Scale Ridesourcing Trip Data. Transportation Research Record, 2674(9), 383–394. https://doi.org/10.1177/0361198120929338
2. City of Chicago Data Portal, 2021. Transportation Network Providers - Trips. https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p/data.
3. Smart Location Database, Environmental Protection Agency. https://www.epa.gov/smartgrowth/smart-location-mapping
4. K-means clustering: Framework Source: https://medium.datadriveninvestor.com/k-means-clustering-b89d349e98e6