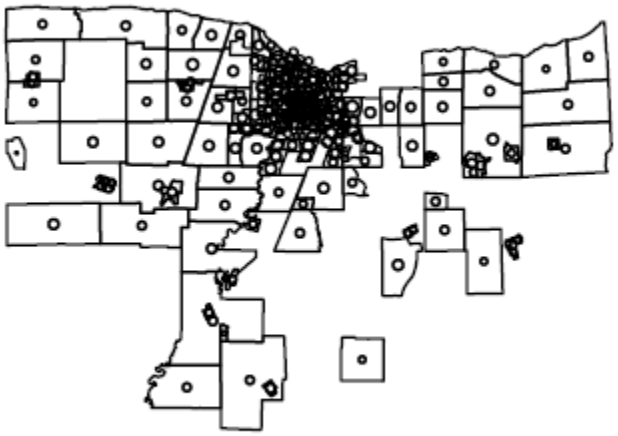


# Rochester Repair

Milstein Junior Year Project

May 3, 2024

Ming DeMers

# Abstract

How does one quantify a city? How can one define the streets and neighborhoods that make up a city? This project focuses on answering those questions for the mid-sized city of Rochester, NY. By collecting and creating a dataset, running a machine learning clustering algorithm, and investigating a number of indicators, we discover a way to classify and quantify a city with a holistic outcome variable. We find that the city has three "levels of synergy." That is, the value of a house/street/neighborhood, and the area's fortitude to decay. For example, a level 1 synergy indicates an area where housing is valued higher than the city median, and there are greater amenities such as schools, fire departments, parks, and trees. These levels distinctly follow a target shape, notably with one boundary being defined almost perfectly by the Inner Loop. In total, then, we see that a city can be defined by a number of observations: from trees to taxes, from crime rate to census data.

# Introduction

As the proud home of giants like Kodak and Xerox, pioneers of history like Frederick Douglass and Susan B. Anthony, and destinations like the Erie Canal and Genesee Valley Park, Rochester was once the center of contemporary America. Today, this Rust Belt city is the third-most impoverished in the country. It reports one of the highest crime rates per capita and is among the lowest in graduation rates. How did Rochester resign to this fate?

We present a data science report on the city of Rochester, NY. Our goal is to quantify the city and define the streets and neighborhoods that make it up. To do this, we collected and created a dataset, ran a machine learning

clustering algorithm, and investigated a number of indicators. We find that the city has three "levels of synergy":

1. Value of a house/street/neighborhood
2. Fortitude to decay
3. Presence of Amenities

# Methods

Foremost, we aim to find indicators of interest in our dataset. Thus, creating a robust and clean dataset is paramount. The data collected for this project was sourced from the Rochester Open Data Portal. The portal contains The Rochester Open Data Portal is invaluable for gleaning more data insight about the city. It contains a wealth of data on a variety of topics, including crime, education, housing, and transportation. This data can be used to track trends, identify problems, and develop solutions. Data was also sourced from the 2020 Census, for the same aforementioned reasons.

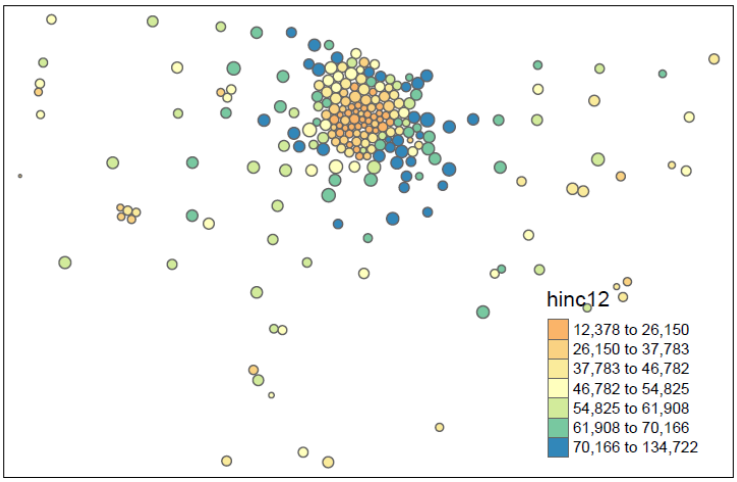
We utilize the R programming language, as well as a number of packages to extend the language's capabilities. Chiefly, we use tidyverse for data wrangling and plotting, ggmaps for choropleths, plotly for interactive visualizations, and tidymodels for machine learning models.

# Results

## Clustering

We first aim to conduct an unsupervised machine learning clustering of the data. We first look at income of the city via a dorling cartogram of income

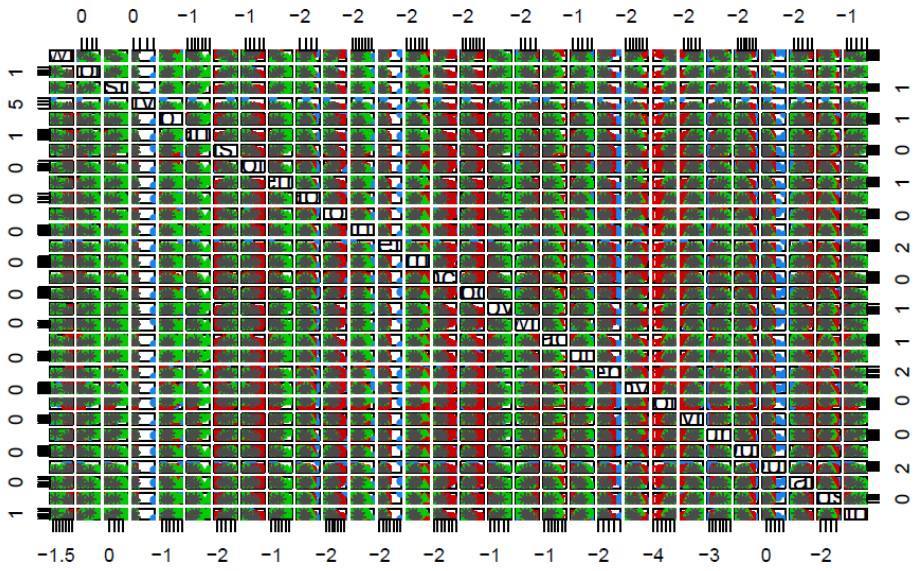
levels in the city of Rochester. The resulting cartogram perhaps gives us a sneak peek into the results to come.



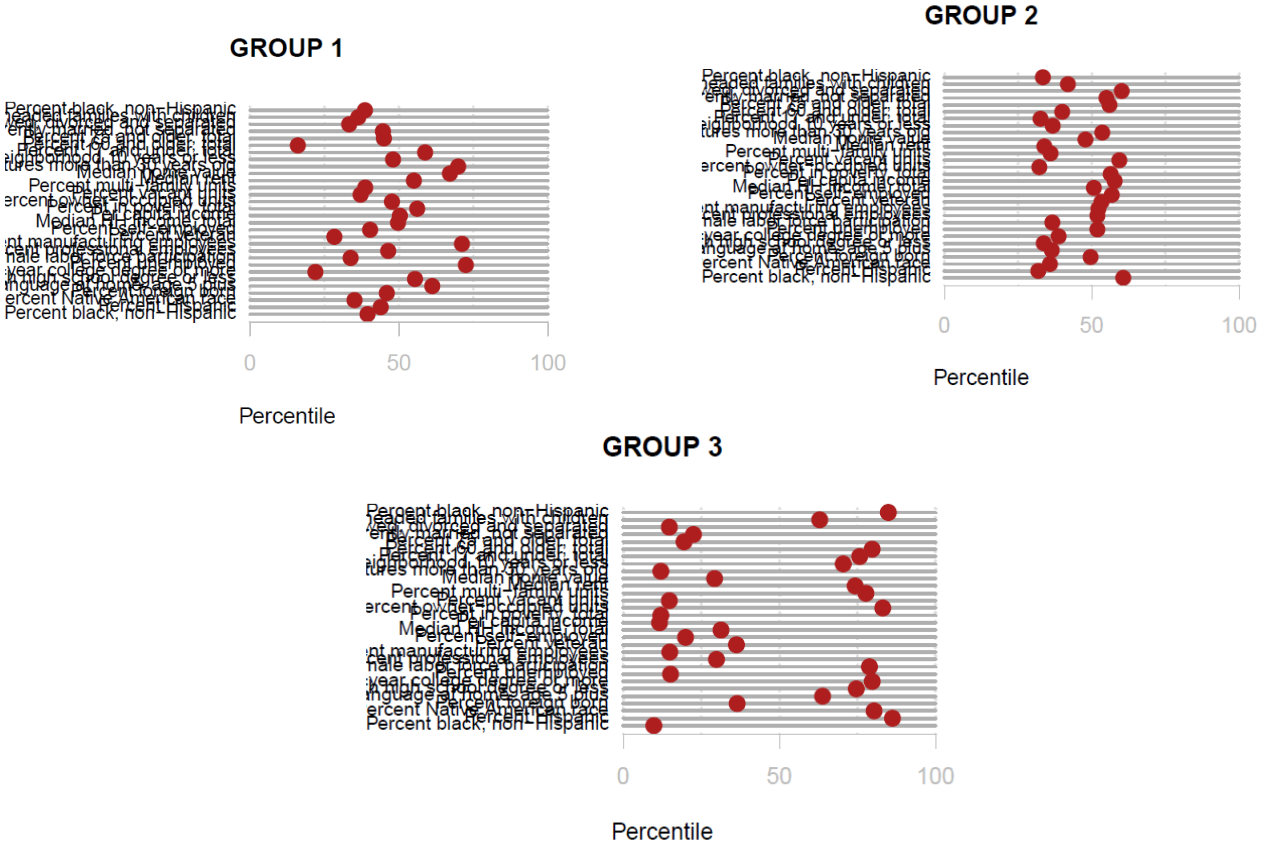

*Dorling cartogram of the income levels in Rochester*

The resulting map shows a generally large concentration of lower-income individuals, (12,000 - 35,000) towards the inner city, with levels of 50,000 and above increasing as the points go further away from the city. The highest concentration of high earners seems to be located in the lower right. One can generally pick out three major areas of income: from the inner city bubble to a yellowish ring around it, then a green-blue surrounding the ring. Beyond the city and within the county are income levels falling around the middle of the scale.

Without census data, we keep only indicators of potential interest, including demographics such as the percent white/Hispanic/non-Hispanic, the percent with a 4-year college degree, the percent of houses 30+ years old, the percent widowed/ or divorced, etc. In total, we have 16 indicators, create a frequency heatmap, and extract three potential “clusters.”



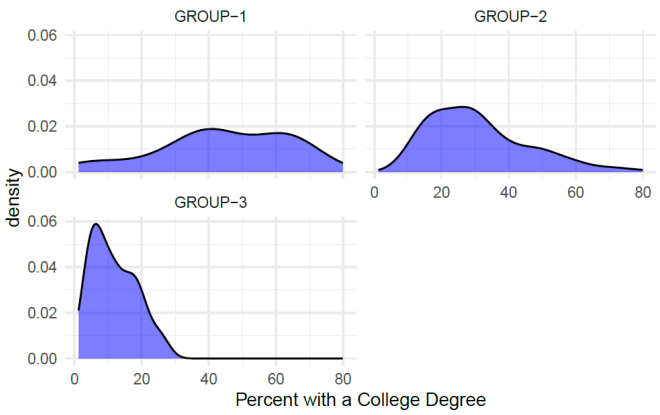
*Frequency heat-map of the 16 indicators*




*Three clusters and their indicators*

From the three groups, we see distinct differences between them. Group 3 stands out the most as having the most extreme levels. Within this third group, we see it's in the 85th percentile for black non-Hispanic, with a high percentage of family units, low college attainment, and low median income, among others. Group 2 seems to have a tighter spread, where its indicators are closer to the 50th percentile: it's around 40th black and non-Hispanic, median income is in the 50th percentile, around 60th for multi-family units, and middle degree attainment. Finally, group 1 swings in the other direction: low percentile for black and non-Hispanic, with higher percentiles of income, degree attainment, and property value, than compared to the previous groups.

For example, we can closely investigate college attainment for the groups.

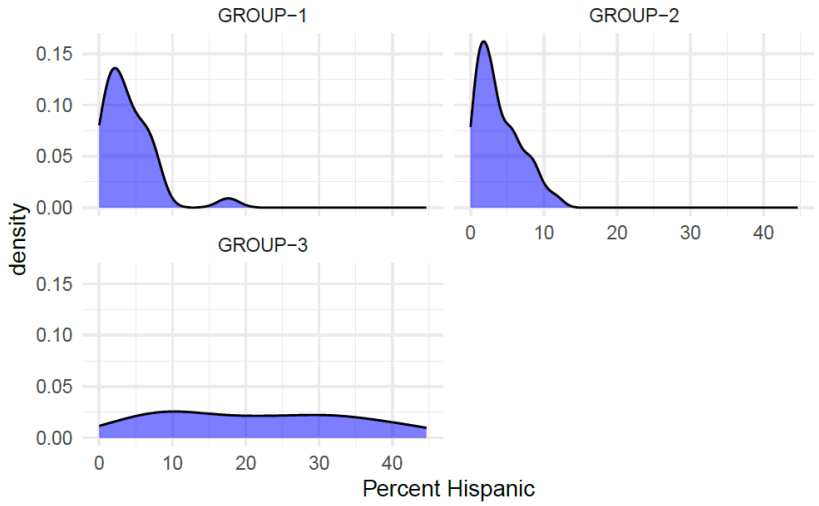



*College degree attainment among the clusters*

We see that among the clusters, the first has a relatively flat curve, with a somewhat bimodal distribution, showing that there's concentration around the 40% and 70% marks; it shows that this group has a high degree of attainment. In group 3, we see a concentration under 30%, showing a much lower attainment, with 20% of the cluster having 0% attainment. The second group is somewhere in the middle. The unimodal distribution

peaks at 15%, but still represents areas with 80% attainment, to 0% attainment.

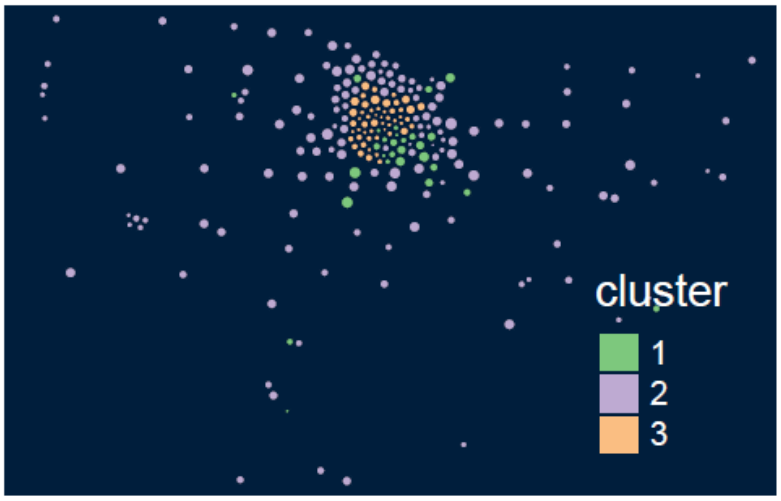
The same analysis can be conducted on Hispanic population in the clusters.



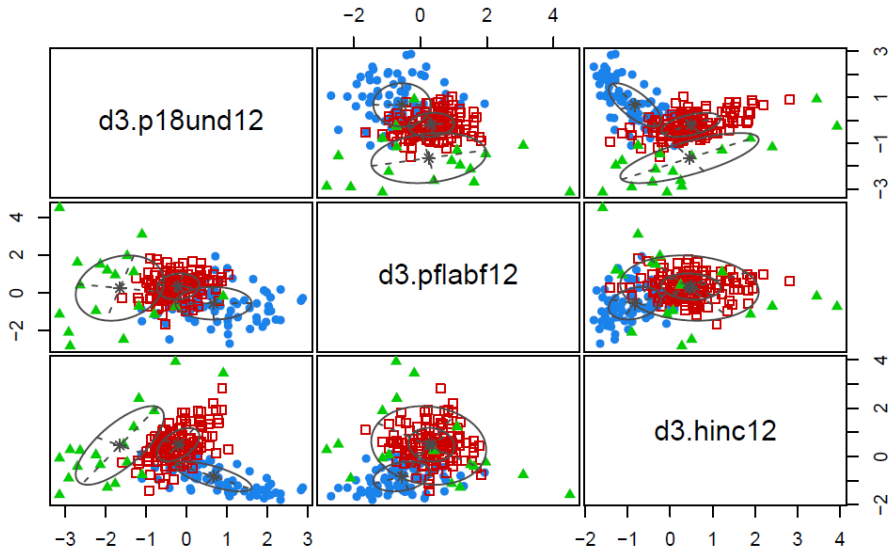

*Percent Hispanics among the clusters*

Briefly, it is quite clear that groups 1 and 2 have a very low Hispanic population, compared to group 3, which has an even distribution but also the highest percentages.

Finally, we may plot these clusters and their data points onto a choropleth.




*Choropleth of the three clusters.*



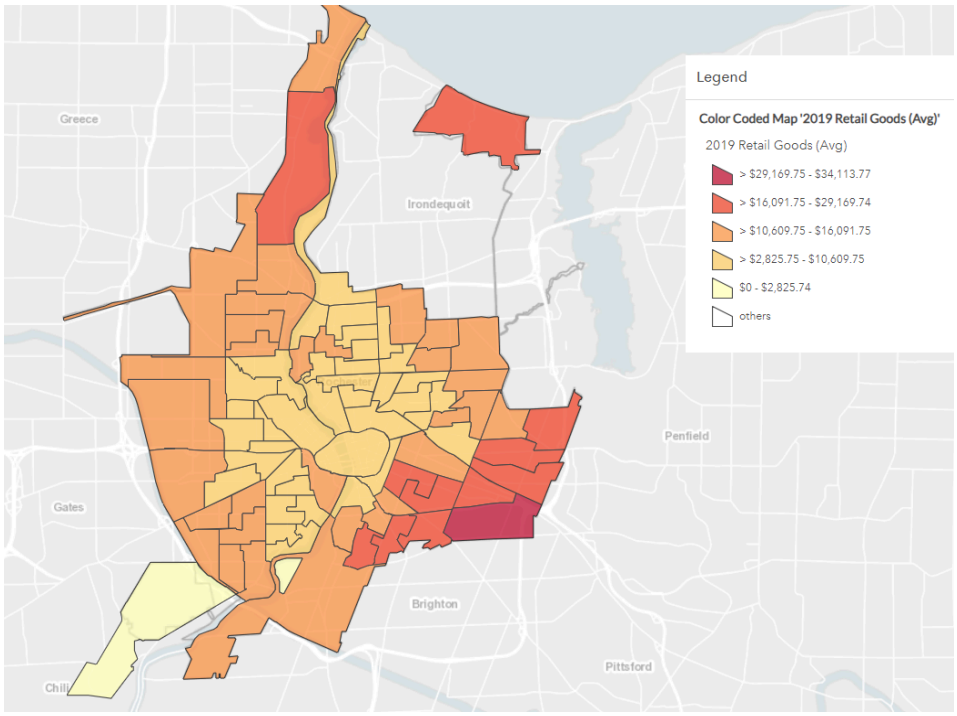

*Cluster array with different indicators selected.*

We see our three clusters are tightly arranged in a “target” pattern, with group 3 in the very center of the city, followed by group 2 encompassing the third group, and representing the most areas, including places not colloquially considered the city. Group 1 is a small pocket in the south-east of the inner city, as well as a few infiltrating spots of group 2.

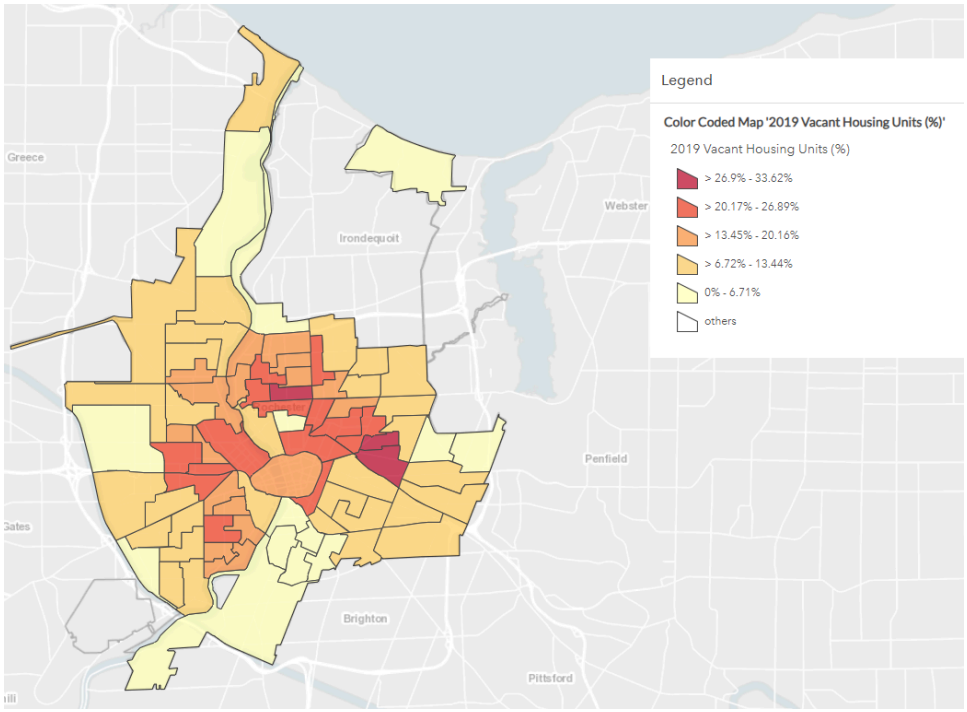
# Indicators

The following are graphs that plot various indicators and their binned values over a choropleth of Rochester. They reinforce the three-cluster results on an individual basis.

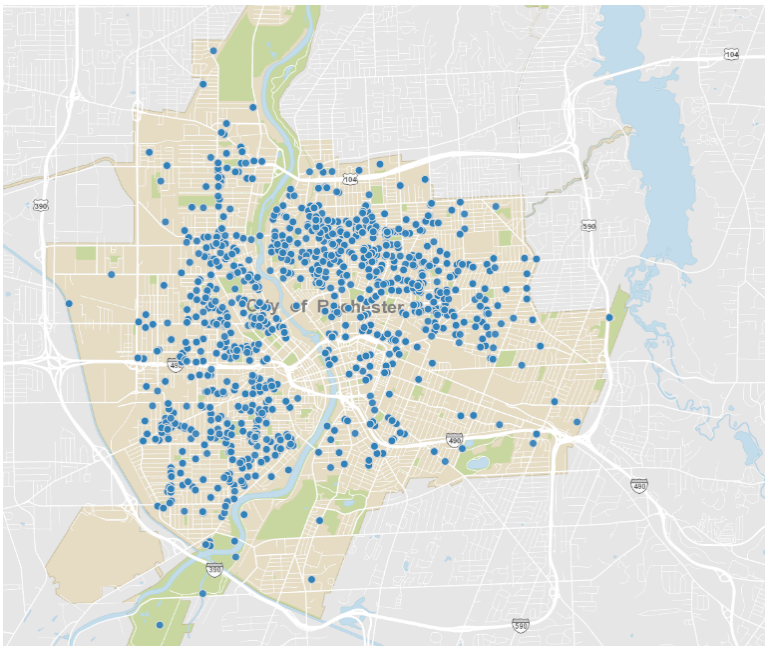
## A. Percent Money Spent Shopping



B. Percent Vacant Housing




C. Homicides since 2010



# Conclusion

Conclusion

In this paper, we have presented a data science-based analysis of the city of Rochester, NY. We have identified three levels of synergy that define the city's neighborhoods: the value of a house/street/neighborhood, fortitude to decay, and the presence of amenities. We have also shown that the city can be classified into three distinct clusters based on these indicators. Our findings provide a valuable resource for understanding the city of Rochester and for developing policies and interventions to address its challenges.

Our work has a number of implications for the city of Rochester. First, it provides a comprehensive understanding of the city's neighborhoods and their unique characteristics. This information can be used to target interventions and services to the areas that need them most. Second, our findings can help to inform the city's land use and zoning policies. By understanding the factors that contribute to a neighborhood's success, the city can make decisions that will promote economic development and improve the quality of life for its residents.

Finally, our work can help to raise awareness of the challenges facing Rochester and the need for action. By understanding the data behind the city's problems, we can begin to develop solutions that will make a difference.