Neighborhood selection for relocation

1. Introduction/Background:

A 2019 study by the US Census Bureau tracking the migration within USA showed that approximately 32 million moved to a different apartment or home in 2018. This corresponds to the relocation of approximately 10% of the US population. People move for a multitude of reasons - the top three reasons being bigger/cheaper house, better neighborhood and better job prospects. The latter of these - the job-related relocation industry, is worth approximately $25 billion itself! Relocation expense is a sizeable chunk of the job offer and it is estimated that on an average, firms spend approximately $16 million every year on employee relocation.

While the process of choosing the next apartment or house can feel quite daunting given the different variables at play, it becomes more taxing when it's a job-related relocation. The employers provide the new hires ~2 weeks to accept a job offer and then an additional 4 weeks to report for the new role. To manage the logistics of moving out from a known location to another neighborhood or state in such a short time span can become daunting. Not having enough time to scope out the new location, most of the employees rely on online research for choosing the right neighborhood. As most of this information is available online through different sources, scraping the web for information and using data science techniques to process the data can speed up neighborhood research. This would help the employee use the available time effectively and reduce the strain of the relocation process.

Business Problem:

While neighborhood selection is usually a personal choice, this assignment focusses on the following two factors:

- Safety - as measured by the crime statistics per 1000 people of the city and
- Facilities such as schools, grocery stores, gyms, parks etc.

This project aims to classify neighborhoods based on these factors. To highlight the diversity of neighborhoods, Minnesota state is chosen as it has a mix of big and small cities.

2. Data Sources:

Information about the list of cities and counties in Minnesota and their population is obtained from Wikipedia (here). Crime statistics information was harder to get as the most up to date databases required a paid account. An article listing one such report was used to scrape the crime stats information (here). As crime statistics were not available for all the cities in Minnesota, the data was curtailed to include only those cities with a complete set of information.

The spatial location data (latitude, longitude) was obtained using the Geocoder package and added to the table. Neighborhood facilities within a 1-mile radius were searched for each city using Foursquare API and neighborhoods were classified according to their categories.

This combined dataset was used for clustering neighborhoods by K-means classification method.

3. Methodology:

Data scraped from different websites had to be cleaned up and combined in one table. Inconsistencies in the city names (Saint Paul vs St. Paul) needed to be checked and data was edited so that the tables could be merged together.

Data for all the parameters was not available across different sources and hence only the complete data entries were used for further analysis. This led to a significant reduction in the table size (from 506 entries to 81).

The spatial information (latitude, longitude) was obtained for each city using the geocoder package.

4. Exploratory analysis

i. Cities in Minnesota are concentrated in the southern part of the state

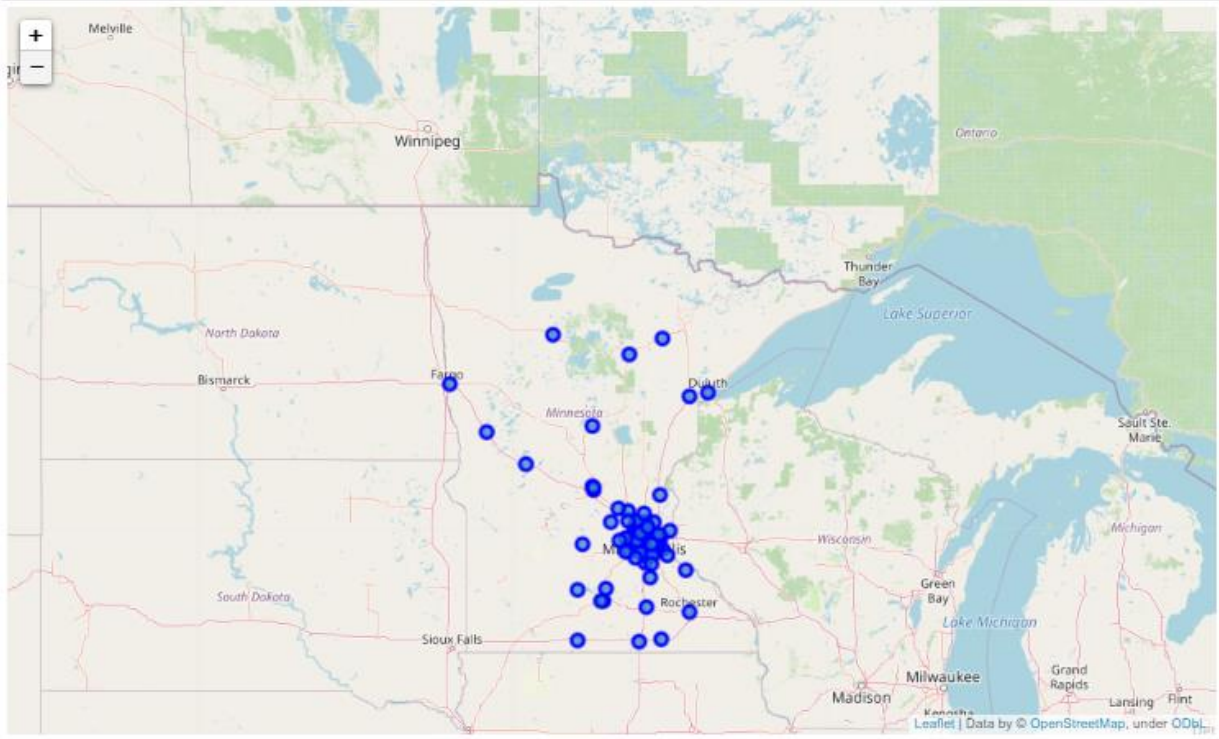

Figure 1: Minnesota Map

ii. Cities in Minnesota are quite diverse - the most populous city of Minneapolis has 429606 people, while the least populous city of Minnetrista has only 8130 people

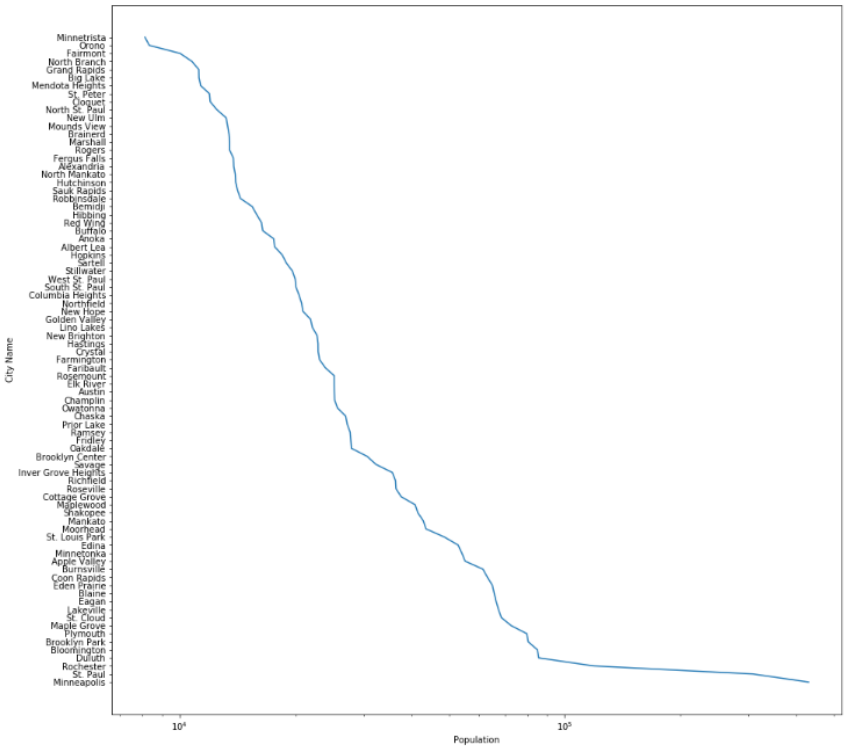

Figure 2: Population plot - Minnesota

iii. Plotting the quartile ranges using boxplots shows some cities have outliers on the higher side

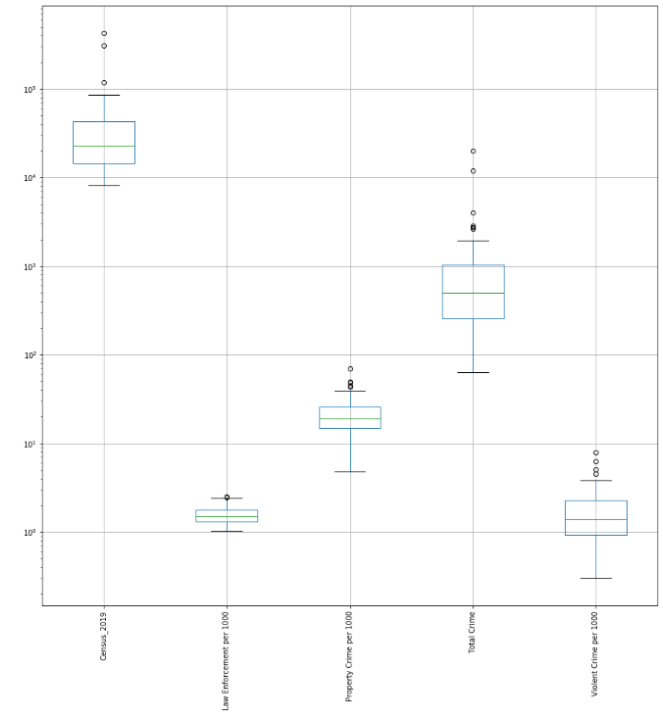

Figure 3: Box Plot - MN

iv. Total Crime as well as Law enforcement per 1000 people follow the population trend exponentially

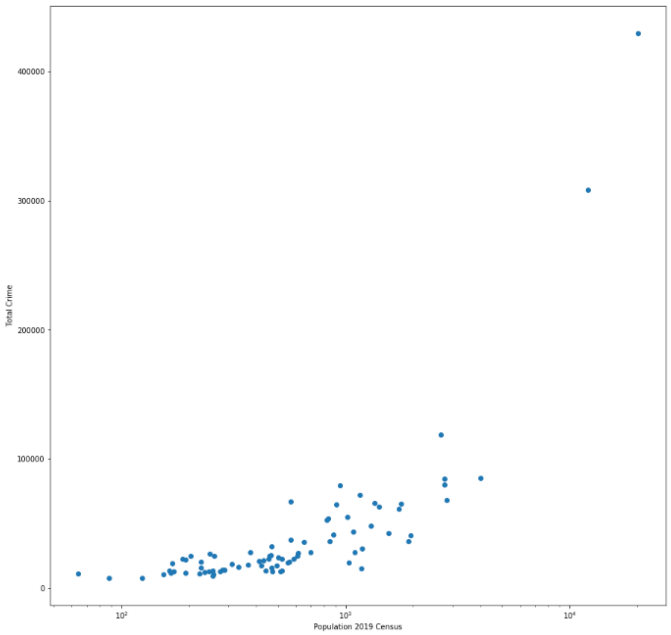

Figure 4: Total Crime vs Population - MN

v. Ratio of Total crime to Law Enforcement personnel gives a rough estimate of the presence of regulatory bodies while negating the bias due to population numbers

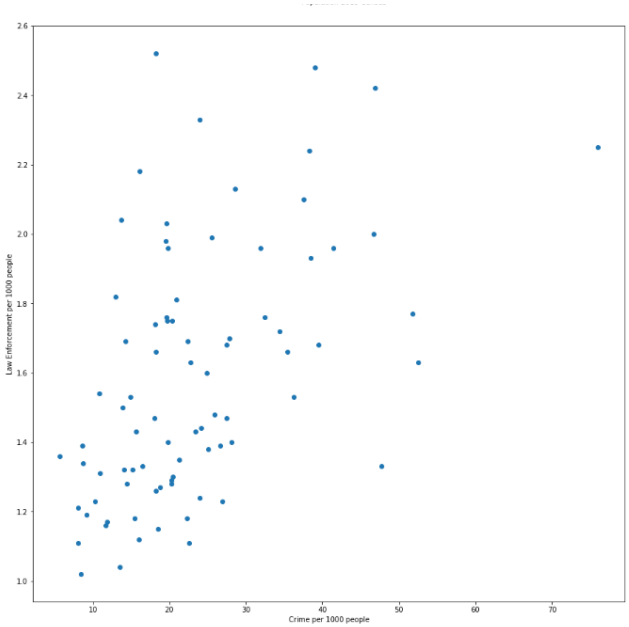

Figure 5: Law Enforcement vs Crime per 1000 people

5. Results:

Neighborhood classification based on crime statistics was used to answer the question of the safest neighborhood in the state. The parameters were normalized according to the maximum value and K-means clustering method was used for classification. It was observed that the wide population range skews the classification results. Some clusters contain only one city.

	Census_2019	Law Enforcement per 1000	Property Crime per 1000	Total Crime	Violent Crime per 1000
Labels					
0	368851.000000	2.450000	35.830000	16084.000000	7.100000
1	16073.647059	1.949412	17.721765	313.705882	1.619412
2	37298.400000	1.436000	23.071000	907.700000	1.731000
3	15434.000000	2.250000	70.450000	1173.000000	5.090000
4	43016.307692	1.816154	37.340769	1735.923077	3.217692
5	32681.357143	1.280000	12.451786	455.892857	0.805714

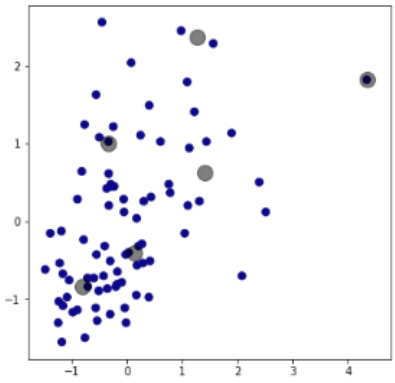

Figure 6: Neighborhood classification by Crime Statistics

Neighborhood facilities were obtained using the Foursquare API. The spatial location was obtained using Geocoder package and added to the data frame. All the venues within a 1 mile radius were collected and tabulated to completely explore a city. In all 248 unique categories were found across all the cities. Venues were rolled up according to the city name and top 5 venues were tabulated for reference.

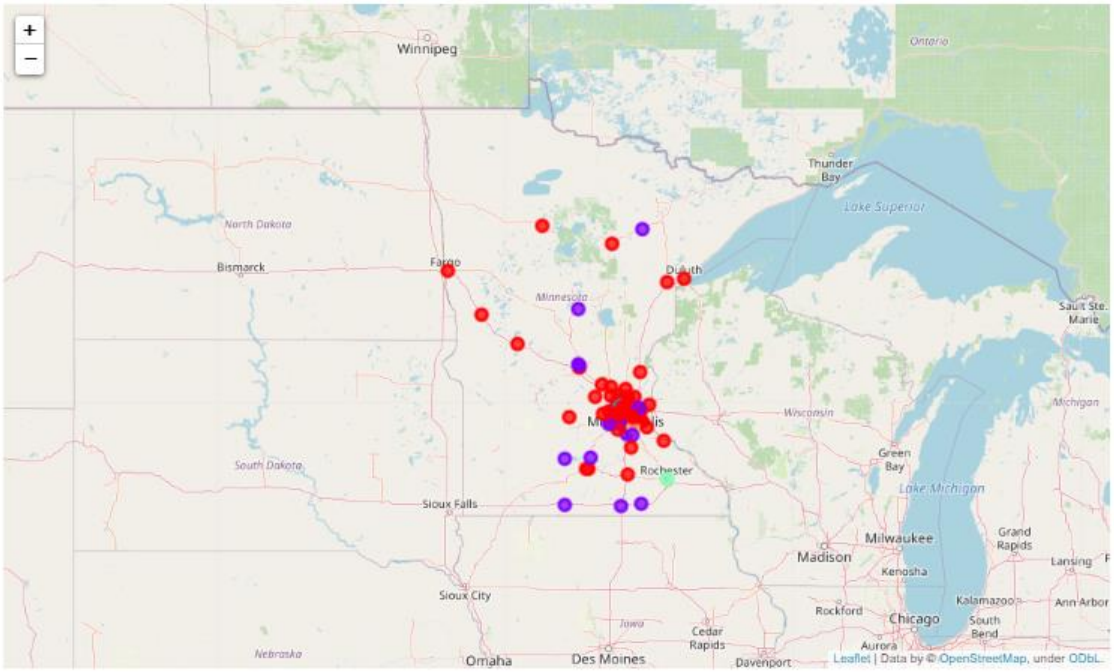

Figure 7: Neighborhood classification - Map

6. Discussion:

Analysis of the cities in Minnesota highlights the potential neighborhoods for migration. The distribution of resources based on the population can be clearly observed. The tradeoff neighborhood facilities and safety score can be observed. K means clustering helps to classify similar neighborhoods across the state.

The logical next step would be to pair this neighborhood information with real estate information to provide a complete picture for relocation. The apartment/ house rental costs and aesthetics would be influenced by the community requirements. Hence one may find that while bigger cities have more vacant units, the rental costs may be too high in a few neighborhoods. Coupling such apartment statistics would therefore bring such trade-offs to the fore.

Another potential improvement is to search for more comprehensive data sources for crime statistics. Adding nationwide ranking to cities might help to provide perspective to such numbers.

7. Conclusions:

This analysis showed the diversity of the facilities available across different cities in the state of Minnesota, USA.