Finding similar neighborhoods in Raleigh, NC using K-Means clustering

Anton Bezuglov, Ph.D.

Abstract—

An average American moves 11.4 times in their lifetime, according to U.S. Census bureau. In case of cross country relocation, people sometimes have little knowledge about the new residence and its surroundings. Some might prefer neighborhoods with younger families and more entertainment, others – more established neighborhoods, where the majority of residents have paid off their houses. To make things worse, insufficient and anecdotal information may affect people decision. One way to mitigate the problem is to find neighborhood(s) similar to that of the persons' previous residence. This project focuses on finding neighborhoods in Raleigh, NC that are similar to Murraywood neighborhood in Columbia, SC. The work provides a list of recommended neighborhoods, their location map, and other recommendations. The accompanying code is available at github.com

Keywords—Neighborhood,K-Means clustering,Data Analysis, Web scraping

1 Introduction

An average American moves 11.4 times in their lifetime, according to U.S. Census bureau. In case of cross country relocation, people sometimes have little knowledge about the new residence and its surroundings. Some might prefer neighborhoods with younger families and more entertainment, others – more established neighborhoods, where the majority of residents have paid off their houses. To make things worse, there is lack of well structured information (not the data) and anecdotal evidences may affect people decision.

One way to mitigate the problem is to find neighborhood(s) similar to that of the persons' previous residence. To summarize, the objective of this project is to help people that are planning to move with finding the best fit neighborhood at the new residence.

The author also has personal interest in the project as we are currently planning a cross-country move to Raleigh, North Carolina. For the 'ideal match', we used one of the authors' previous residences – Murraywood, Columbia, South Carolina. Even though these cities and neighborhoods appear in the project, the approach will work for practically any state in the United States.

The general approach to the problem is to cluster neighborhoods in Raleigh and identify what cluster Murraywood would be in. This cluster will contain similar neighborhoods that should be recommended.

- *Anton Bezuglov is Assoc. Professor of Data Analytics at Buena Vista University*
  E-mail: abezuglov@gmail.com

2 Data

2.1 Socioeconomic data

Previously (Coursera classes), the exclusive source of neighborhood data was Foursquare, which will also be used later. Here, however we also collect and analyze socioeconomic data on neighborhood residents. Web resource http://www.city-data.com provides this information at neighborhood level. The following attributes for each neighborhood is collected:

- Neighborhood name
- Area (sq. miles)
- Population, males population, females population
- Average household size

- Average number of cars in apartments and houses
- Medium age of residents (males and females separately)
- Medium household income
- Medium rent, mortgage payments, percent mortgages
- Percent of residents born in this state, in another U.S. state, foreign born residents
- Percent married, never married females and never married males over age of 15
- Percent of people not speaking English well
- Percent of single mother households

For each neighborhood URL, this data is scraped from www.city-data.com with Python code.

2.2 Venues

The venues information is obtained through Foursquare API. The maximum of 100 venues was selected withing 500 meters proximity from the geographic coordinates of the neighborhood.

GeoPy library returns geographic coordinates by neighborhood name.

2.3 List of neighborhoods

City-data.com also contains a list of neighborhoods by each state in the U.S. To obtain the list of neighborhoods in Raleigh, the code first scrapes everything in North Carolina and then saves only neighborhoods pertaining to Raleigh (those containing "Raleigh, NC" as a part of their name).

Further processing was needed for GeoPy, since it could not resolve coordinates for names like "Anderson Heights in Raleigh, NC". Those needed to be transformed to "Anderson Heights, Raleigh, NC".

2.4 Web Scraping

The first step of the scraping is to make a dataset containing neighborhood names and their city-data.com URLs.

At the next step, the socioeconomic data was scraped for each neighborhood by its URL. The code used random waits of 5-20 seconds after each request so that the server is not upset.

Then, GeoPy converted names to coordinates and those were passed to FourSquare. FourSquare allowed up to 100K interactions every 24 hours, so no random timeouts were necessary there. However, the amount of GeoPy requests exceeded the quota *after* the scraping has completed.

At the last state, both datasets joined and saved for further use.

2.5 Data Examples

Below is one example of socioeconomic data for Anderson Heights neighborhood in Raleigh:

- Area: 0.37 sq. miles
- Population: 964
- Average household size: 2.8 people
- Average number of cars in apartments: 1.9
- Average number of cars in houses: 2.3
- Male residents: 485
- Median age females: 39.2 years
- Median age males: 45.1 years
- Median household income: $161,702
- Median rent $1,125
- Percent residents born in another U.S. state: 62.0%
- Percent families with children: 32.5%

Anderson Heights neighborhood is fairly large (964 people) with primarily U.S. born residents with above average household income. Below are a few venues close to the neighborhood:

- Jersey Mike's Subs, Sandwich Place: (34.075, -81.178)
- Howie's Pizza, Pizza Place: (34.071, -81.175)
- Chapala, Mexican Restaurant: (34.075, -81.178)

3 Methodology

This section focuses on cleaning, preparation of previously collected data, exploratory analysis, and machine learning.

3.1 Cleaning and Preparation

Several data fields in the scraped data needed prior processing before the analysis. In most cases, it involved removing extra symbols like "$" in monetary data or "%" in percentages and converting the fields to numbers.

Occasional not-a-number (NaN) data in the dataset were either removed or filled with a mean for the corresponding field. Whenever the number of males living in a neighborhood was missing, a half of the neighborhood population was used instead.

3.2 Exploratory Analysis

At this stage, each field in the dataset was analyzed to understand the distribution of the data, detect outliers, etc.

For instance, Figure 1 shows a histogram of median age of neighborhood residents. It demonstrates that the data looks 'normal', i.e. does not contain negative numbers or numbers exceeding one hundred years.

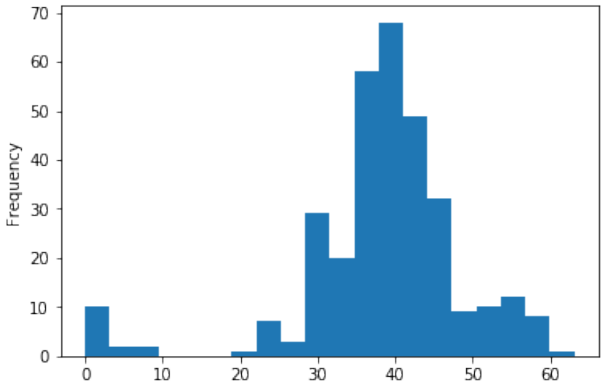

Figure 1: Histogram of Median Age of Residents

Contrary to this, Figure 2 demonstrates one or more cases, where the percentage of single mother households exceeds 100%. This data were later removed from the data set.

Finally for this stage, the exploratory analysis was done to find dependencies between data fields, using pairplot in Figure 3.

3.3 Machine Learning

At this stage we have a data set containing several hundred neighborhoods, each of which has approximately twenty fields containing information about its residents and venues. The core idea is to assign the neighborhoods to clusters according to these data. So that one cluster may contain neighborhoods with more parks and wealthier residents, and so on. Naturally, the cluster containing Murraywood will also contain neighborhoods similar to it and so these neighborhoods must be recommended to the user.

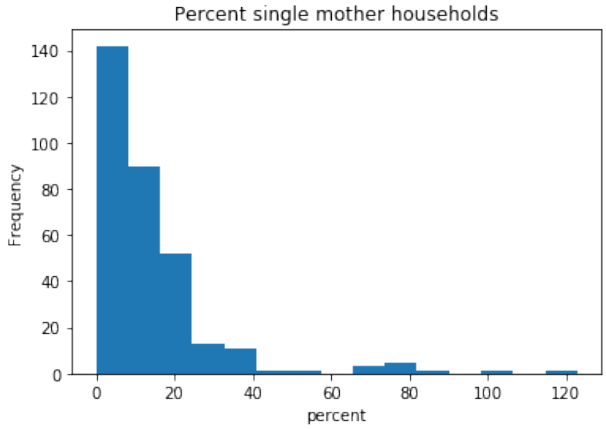

Figure 2: Histogram of Median Age of Residents

The clustering is performed with K-Means algorithm. Elbow method (see Figure 4) allows to find the optimum number of clusters, where the average distance between data points and their cluster centers stops decreasing rapidly. Here, the optimum number of clusters is around 15.

4 Results

After clustering was complete, we grouped the neighborhoods by their cluster, computed the mean of their fields and sorted by median household income in decreasing order. This revealed that the Murraywood cluster was in the middle of the table at rank 8.

When compared to other clusters, a few other interesting properties of Murraywood revealed:

- Murraywood has the second highest number of Parks (0.167), the first is cluster #3
- Murraywood has the second lowest number of units with mortgage (64%), the first is cluster #10
- Murraywood is fairly close to convenience stores (5th closest)

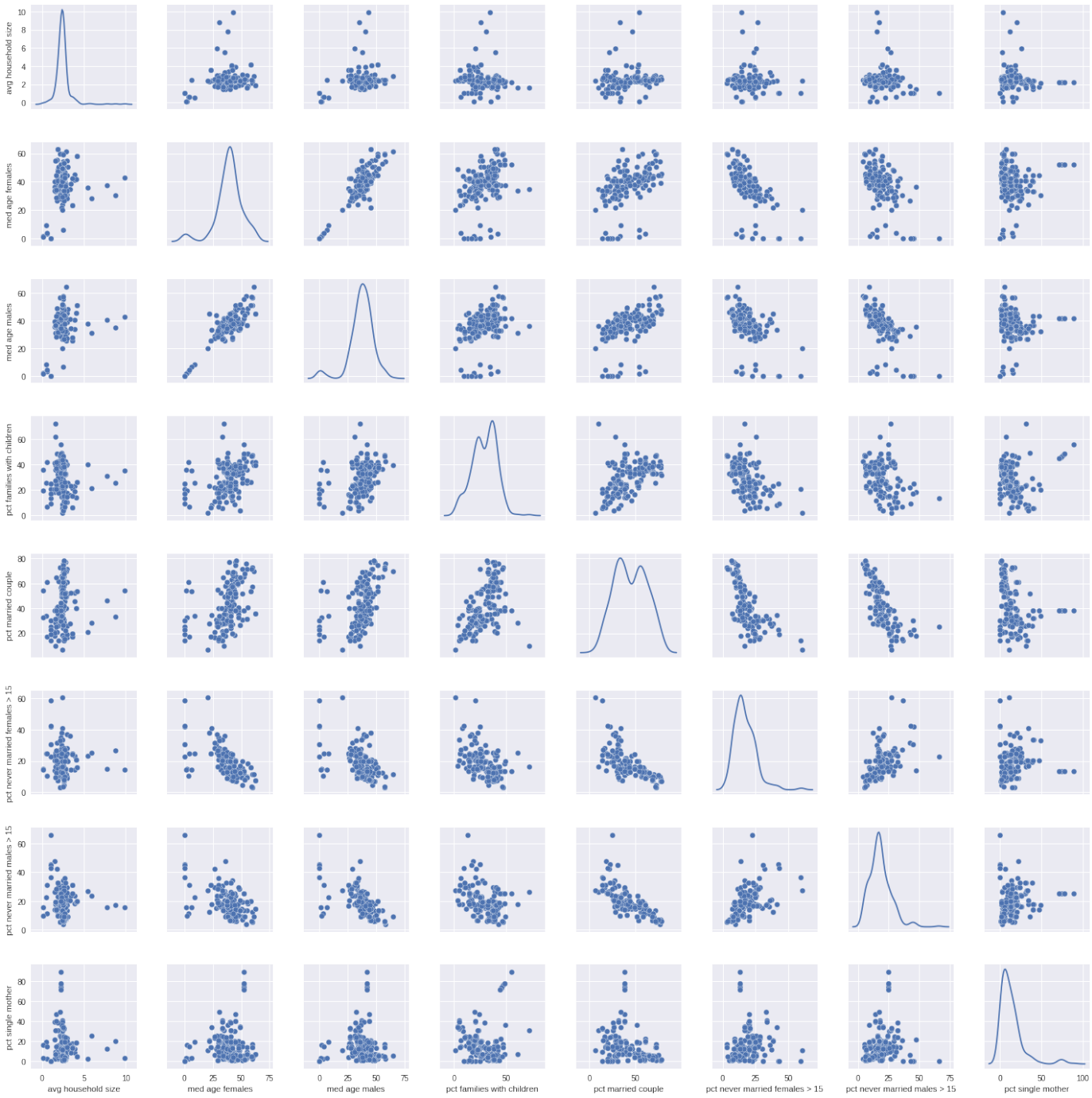

Figure 3: Pairplot of Socioeconomic data

In summary, Murraywood seems to be more established (few units with mortgage) and close to parks. Based on these factors, it was also worth to consider clusters #3 and #10 previously mentioned. The first has more parks, while still not too many units with mortgage. The second has fewer units with mortgage and still some proximity to the parks.

4.1 Murraywood cluster

The Murraywood cluster contains the following neighborhoods:

Alexander Place Townhomes, Anderson Forest, Arbor at Harrington Grove, Blenheim Place, Cambridge, Cameron Village, Cedar Hills Estates, Chadleigh Pointe, Claridge, College Crest, Colony Woods, Eden Forest, Evans Mill, Glenanneve Place, Grove Park, Jordan, Lyon Park,

Figure 4: Finding the best number of clusters

North Park, Northglen, Northwood Acres, Parliament Pointe at Greystone, Ponderosa, Roanoke Park, Scarsdale, Sloans Garden Lane, Springstone, Stafford Townhomes, Stone Quarter Greystone, and Stonehaven.

4.2 Second priority neighborhoods

Below is the list of neighborhoods from other two clusters that are similar to Murraywood:

Anderson Heights, Budleigh, Country Club Villas, Highland Forest, Kenly Court, Lake Boone Place, Lakestone, Lassiter, Starview, and Sulgrave Manor

Figure 5 provides a map of the recommended neighborhoods. In the map, the Murraywood cluster is shown with red and the other two clusters in yellow and orange.

5 Discussion

Based on the analysis above, there are two groups of neighborhoods to consider. The first group is a closer match to Murraywood and it generally includes neighborhoods in North West Raleigh located from downtown to I-540 and approximately 1-1.5 miles NW of I-540. This group most closely matches Murraywood by multiple attributes.

The second group of neighborhoods has a substantially higher median household income. However, it is still fairly close to Murraywood in terms of proximity to parks and the low number of units with mortgage. This group is also located in NW Raleigh area, but it is within I-440 highway. If the higher median income is not an issue, these neighborhoods should be recommended as well.

Considering the location of both groups (North – North-West Raleigh), other neighborhoods in the area close to I-540 may also be recommended. These can make the third priority group of neighborhoods.

6 Conclusion

The project analyzes neighborhood similarity based on two data sources: socioeconomic data, obtained from city-data.com and venues data from Foursquare. The socioeconomic data web scraping code is general enough to pull data for neighborhoods in other cities as necessary. In fact, some neighborhoods in Raleigh area do belong to other cities like Durham, Chapel Hill, Cary, and so it would be nice to include these into the analysis in the future work.

The scraped data needed heavy cleaning as it contained nonexistent neighborhoods, missing neighborhoods, or neighborhoods consisting of a single household. Some socioeconomic data were not trustful either, such as neighborhoods with average number of vehicles exceeding 10 (apartments?).

Even though the venues data was more accurate, we wish Foursquare had returned more venues as the data were very sparse and difficult to use for clustering. In future work, it may also be useful to include data pertaining to schools, universities, and such.

The recommendation can also be improved further by including other data such as police and nuisance reports, accidents data, and other.

7 Appendix

The code is available at Github: https://github.com/abezuglov/coursera_applied_ds

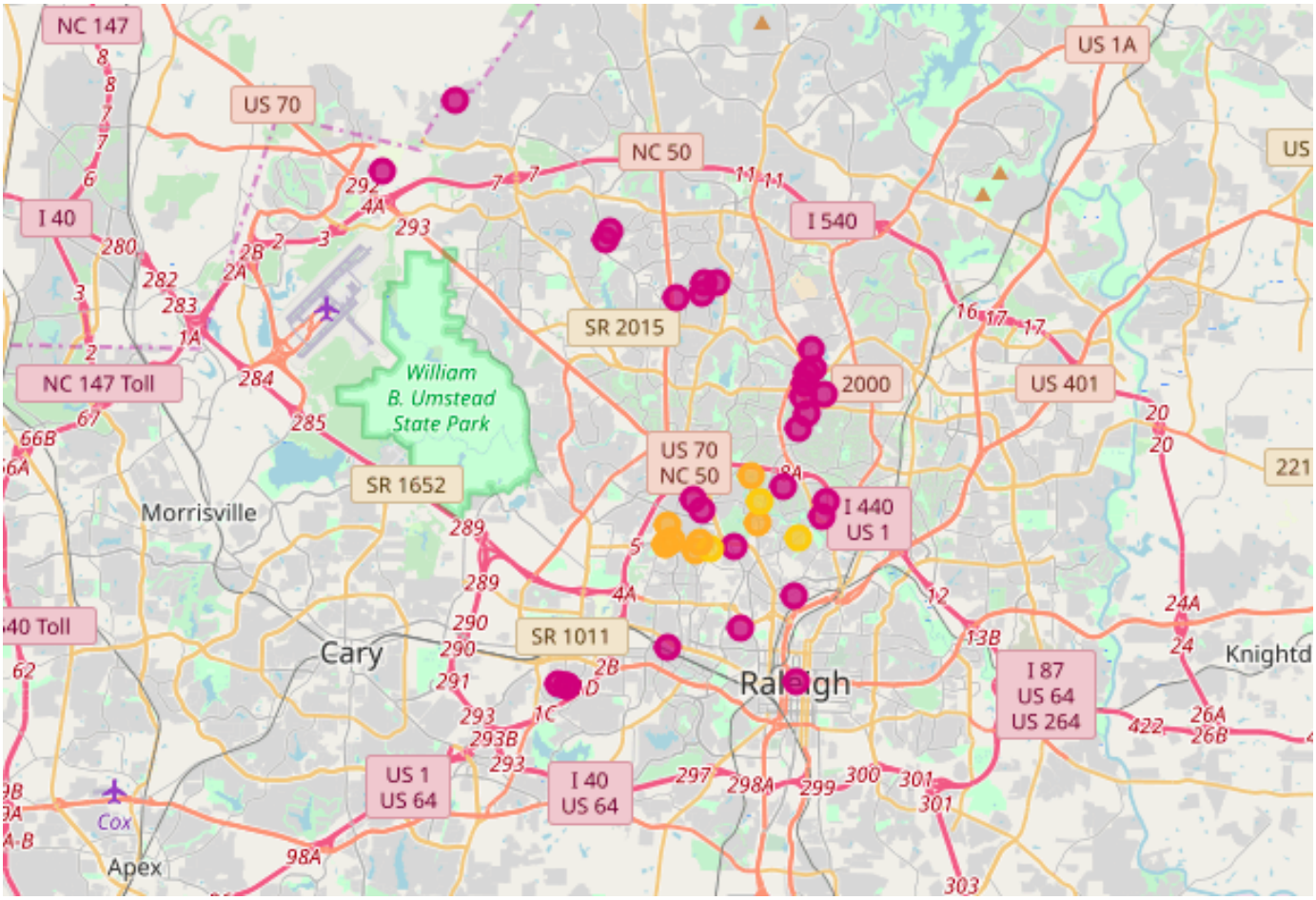

Figure 5: Recommended neighborhoods in Raleigh. Murraywood cluster – red, Clusters 3 & 10 (second priority) – yellow and light orange