Segmenting and Clustering Counties for Foreigners in US

Muhammad Haider Raza

May 5, 2020

1. Introduction

1.1. Background

In the context of this analytical study the foreigners refer to people who hold dual nationalities and U.S. Citizenship. There ancestry originates from a foreign country or foreigners who migrated to and reside in United States. For simplicity we will be focusing on Pakistani Americans. *https://en.wikipedia.org/wiki/Pakistani_Americans [1]*

Typically, when foreigners plan to move to US or even while living in US when they want to relocate to another city/neighborhood/county there preference is to live in a demographic area where nationals from their origin are residing. Apart from that there could be different other factors that influence their decision like distance to workplace, school/college/university for their children, climate, and proximity of amenities.

1.2. Problem

The study here analyzes the available information in this regards and helps take this decision by providing options of similar places (counties) based on an individual preferences. There could be so many decisive factors but here I limit my study as per following scenario:

A Pakistani doctor is looking for counties in US where there are considerable numbers of Pakistani households plus there are good number of job opportunities for him that are near to the potential residence which eventually saves him/her commute time. His/her location of interest could be hospitals, clinics, medical research institutes where he can offer his services. We need to present him all the options and also cluster similar options together in order to help him take an informed decision.

1.3. Interest

Anyone intending to relocate to (within or from outside) US may naturally have this difficulty in first of all knowing all the options and then choosing one among all available options as a place of abode. Using Data Science we can give him an analysis of alike places which could help him choose from the options.

Though there can be a number of options that could help enhance this search but for my segmenting algorithm following are the main variables to keep the scope simple:

- No. of Pakistani households in a county

- The minimum distance/radius in which he is looking for a job can be an input parameter. For example here in this analysis we consider a doctor is looking for number of Hospitals available in a radius of 25 Km.
- No. of hospitals, clinics, medical research institutes within a distance (radius) of the county where considerable number of Pakistanis live. Here the place of interest can vary as per his/her profession. Some options could be:
  - Hospital
  - IT Services
  - Research Station
  - Bank
  - University
  - Law School
  - Medical School
  - Trade School
  - School

2. Data acquisition and cleaning

2.1. Data sources

For purpose of this analysis following 3 data sources are used:

2.1.1. Wikipedia web page: Pakistani Americans

https://en.wikipedia.org/wiki/Pakistani_Americans [1]

This has a table which provides a list of counties with number of Pakistani households, total county population and Concentration of Pakistanis.

Census here are the number of Pakistani Households in the following US Counties.

County	Pakistani Households	County Population	Concentration of Pakistanis
Queens County, New York	15,972	2,296,000	3.5%
Kings County, New York	14,412	2,592,000	2.8%
Cook County, Illinois	12,759	5,241,000	1.2%
Harris County, Texas	11,221	4,337,000	1.3%
Fairfax County, Virginia	7,358	1,131,000	3.3%
Los Angeles County, California	7,025	10,020,000	0.4%
DuPage County, Illinois	4,000	932,126	2.1%
Middlesex County, New Jersey	3,788	828,919	2.3%
Orange County, California	3,658	3,114,000	0.6%
Dallas County, Texas	3,627	2,480,000	0.7%
Hudson County, New Jersey	3,369	660,282	2.6%
Fort Bend County, Texas (Sugar Land)	3,216	652,365	2.5%
Nassau County, New York	3,137	1,352,000	1.2%
Santa Clara County, California	2,824	1,862,000	0.8%
Alameda County, California	2,726	1,579,000	0.9%
Montgomery County, Maryland	2,410	1,017,000	1.2%
Tarrant County, Texas	2,388	1,912,000	0.6%
Miami-Dade County, Florida	2,038	2,617,000	0.4%
Broward County, Florida	1,934	1,839,000	0.5%
Gwinnett County, Georgia	1,712	859,304	1.0%

2.1.2. Geocoder

To get the geographical co-ordinates of each county ie. latitude and longitude I used
geopy.geocoders python library which converts an address (county name in our case) into latitude and longitude values

2.1.3. Foursquare

Foursquare APIs are used to find number of venues around a geographical co-ordinate (lat/long of county). Following were important parameters for this API

- CategoryId
- Radius
- Lat/Long
- Intent
- Client Id & Client secret

2.2. Data Cleaning

Data downloaded or scarped from multiple sources were combined into a single table.

First of all the table scraped from Wikipedia page had following columns:

- County

- Pakistani Households
- County Population
- Concentration of Pakistanis

We excluded and dropped following columns from our algorithm to keep things simple: County Population, Concentration of Pakistanis. These could be considered in future enhancements but since we had a column 'Pakistani Households' that indicates similar information on strength of Pakistanis living in a county, hence we dropped the column 'Concentration of Pakistanis'. There were around 33 counties present in the list. Luckily all columns have values present in all the rows.

Next, we needed lat/long information for each county entry in the wiki table so that we can invoke Foursquare API to get the venues count. For this purpose gecode API was first invoked and the lat/long info was appended to each county resulting in a dataset like:

	County	Pakistani Households	Latitude	Longitude
0	Queens County, New York	15972	40.652493	-73.791421
1	Kings County, New York	14412	40.645310	-73.955023
2	Cook County, Illinois	12759	41.819738	-87.756525
3	Harris County, Texas	11221	29.811977	-95.374125
4	Fairfax County, Virginia	7358	38.815636	-77.283685

	County	Pakistani Households	Latitude	Longitude
0	Queens County, New York	15972	40.652493	-73.791421
1	Kings County, New York	14412	40.645310	-73.955023
2	Cook County, Illinois	12759	41.819738	-87.756525
3	Harris County, Texas	11221	29.811977	-95.374125
4	Fairfax County, Virginia	7358	38.815636	-77.283685

Next we needed a list of venues for each county within a radius of R (25 Kms) in our case. The venues returned should belong to only the category we are interested e.g Hospital in current case. For this Foursquare search API was used:

https://api.foursquare.com/v2/venues/search?intent=browse&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}

While invoking foursquare api for each county we observed that for some of the locations the address info was missing. Though this is optional for my scope of study but the issue was fixed by providing a default value of '-' and hence we were able to print the name and address information for each venue without breaking the code.

Since we needed to have the count of venue only we got that by taking the length of the JSON response Array ["response"]['venues']

Now we had a dataframe like:

```
print(df_final.head())
```

```
                   County  Pakistani Households   Latitude  Longitude  \
0   Queens County, New York                 15972  40.652493 -73.791421
1    Kings County, New York                 14412  40.645310 -73.955023
2     Cook County, Illinois                 12759  41.819738 -87.756525
3      Harris County, Texas                 11221  29.811977 -95.374125
4  Fairfax County, Virginia                  7358  38.815636 -77.283685

   VenueCount
0          30
1          30
2          30
3          30
4          16
```

2.3. Feature Selection

For final data frame we dropped Lat, Long columns as we had the venue count with us. Also the county name was dropped because the K-means clustering algorithm that we will opt can not work with non-numerical values.

Final dataframe with Pearson correlation coefficient was as:

```
pearsoncorr = df.corr(method='pearson')
pearsoncorr
```

	Pakistani Households	VenueCount
Pakistani Households	1.000000	0.006787
VenueCount	0.006787	1.000000

```
df.head()
```

	Pakistani Households	VenueCount
0	15972	30
1	14412	30
2	12759	30
3	11221	30
4	7358	16