Battle of Neighborhoods – Capstone Project

Siva Jayashankar

March 15, 2019

1. Introduction

1.1 Background

Buying a home is a big investment. This capstone project targets prospective retirees, who are looking for cities in US, that has excellent weather year round, low crime rate and a great neighborhood. San Diego, California tops the list. This Project will help ease the home search process in San Diego County.

1.2 Problem

There isn't a single repository to obtain the crime rate by zip code, for the San Diego county.

1.3 Interest

Though this project is for retirees to find a good neighborhood, anyone looking to buy a home in the San Diego county will find the analysis useful.

2. Data sources and Description

There can be several criteria when it comes to buying a home. This project is focused on finding neighborhoods in San Diego with relatively low crime rate.

2.1 Data sources

The first and foremost criteria for a prospective home buyer is to make sure the neighborhood has a low crime rate. First dataset is from San Diego Association of Governments (SandAG), to help identify the crimes reported in the San Diego county.

- Crime Data set for San Diego Neighborhood
  http://www.sandag.org/programs/public_safety/arjis/CrimeData/crimedata.zip

Knowing the number of crimes in the area will not depict true picture, unless it is related to the number of people who live there. Getting the population of San Diego by zip code will help us achieve that. Number of crimes reported divided by the population times 100,000 will tell how many crimes occur per 100,000 people.

- Population of San Diego by zip code
  https://www.zip-codes.com/city/ca-san-diego.asp

To validate the correctness of the Crime data set pertaining to San Diego, a listing of all zip codes that fall under San Diego county is needed.

- Zip codes of cities in San Diego county can be obtained from San Diego court portal http://www.sdcourt.ca.gov/portal/page?_pageid=55,1524259&_dad=portal&_schema=PORTAL

For any Home buyer, setting a budget for their home purchase is vital. Knowing the Home sale prices from Zillow, a popular web site for home sales statistics from all cities and joining that with the zip codes in San Diego county will give the home prices specific to San Diego county .

- Home sales from Zillow
  https://s3-us-west2.amazonaws.com/econresearch/Reports/Core/RDC_InventoryCoreMetrics_Zip_sfh.csv

- Latitude and Longitude of zip code
  Latitude and longitude of zip code is needed to find the venues specific to an area.
  https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/

- Four square venues API will help in retrieving venues around the neighborhood of San Diego for the specific latitude & longitude
  _https://developer.foursquare.com/docs/api/venues/details

2.2 Data cleaning

- Crime Data set for San Diego County is imported into a pandas data frame object

```
# How does the data look like
sdcrimedata.head(5)
```

	CM_LEGEND	agency	Charge_Description_Orig	activityDate	BLOCK_ADDRESS	ZipCode	community
0	DRUGS/ALCOHOL VIOLATIONS	EL CAJON	DRUNK IN PUBLIC: ALCOHOL, DRUGS, COMBO OR TOLU...	10/9/2018 20:25:00	1200 W BLOCK MAIN STREET	92020.0	EL CAJON
1	DRUGS/ALCOHOL VIOLATIONS	EL CAJON	POSS CONTROLLED SUBS PARAPHERNALIA (M)	10/10/2018 8:53:11	JAMACHA ROAD / SHADOW VISTA WAY	92020.0	EL CAJON
2	DRUGS/ALCOHOL VIOLATIONS	EL CAJON	DISORDERLY CONDUCT: ALCOHOL	10/10/2018 0:40:00	E MAIN STREET / N MOLLISON AVENUE	NaN	EL CAJON
3	DRUGS/ALCOHOL VIOLATIONS	EL CAJON	DRUNK IN PUBLIC: ALCOHOL, DRUGS, COMBO OR TOLU...	10/11/2018 1:51:00	200 BLOCK LINCOLN AVENUE	92020.0	EL CAJON
4	DRUGS/ALCOHOL VIOLATIONS	EL CAJON	POSSESS CONTROLLED SUBSTANCE (M)	10/9/2018 18:00:00	400 BLOCK BROADWAY	92021.0	EL CAJON

1. Rows which have missing values in zip code column was dropped from the data frame
2. Zip code is converted to type int

3. To confirm data set obtained is only that of zip codes from San Diego county, it was joined with another source, San Diego court portal. Data was scraped from the San Diego county web site to obtain zip codes.

- Home sale prices for all zip codes was imported into a pandas data frame from Zillows web site
    1. Missing values filled with '0' and converted to type int to match the data type for joining with home sales data

```
In [47]: # Fill missing values in Zipcode with 0 and convert to type int
allzipcodeprice = allzipcodeprice.fillna('0')
allzipcodeprice['ZipCode'] = allzipcodeprice['ZipCode'].astype(int)
allzipcodeprice['Avg Listing Price'] = allzipcodeprice['Avg Listing Price'].astype(int)
```

    2. Home sales data frame is joined with crime data frame to get home prices only for zip codes in San Diego county

```
In [48]: #merge Crime by Zip with all zip code with inner join
sd_zip_crime = pd.merge(sd_zip_crime,allzipcodeprice, how='inner')
```

- Population of San Diego
    1. Zip code is renamed to match the column name of crime data set.
    2. Population of San Diego by zip code data set is joined with merged data of crime and home sales. Population data is necessary to get the crime rate.

```
In [66]: #sandiego population by zip code from https://www.zip-codes.com/city/ca-san-diego.asp
sd_pop = pd.read_excel("sd_population.xlsx")
```

```
In [67]: sd_pop.head(5)
```

Out[67]:

	ZIP Code	City	Population
0	91901	Alpine	17403
1	91902	Bonita	17653
2	91903	Alpine	0
3	91905	Boulevard	1700
4	91906	Campo	3627

```
In [68]: sd_pop.rename(columns={'ZIP Code':'ZipCode'}, inplace=True)
```

2.3 Feature selection

In a typical feature selection process, correlation between the target and the features are considered. Correlation between features do not have much impact in the current analysis.

Dataset for the neighborhood analysis needs zip code, crime rate and house values.

```
In [49]: sd_zip_crime.info()`

<class 'pandas.core.frame.DataFrame'>
Int64Index: 56761 entries, 0 to 56760
Data columns (total 11 columns):
CM_LEGEND                56761 non-null object
agency                   56761 non-null object
Charge_Description_Orig  56761 non-null object
activityDate             56761 non-null object
BLOCK_ADDRESS            56761 non-null object
ZipCode                  56761 non-null int32
community                56761 non-null object
Name                     56761 non-null object
Neighborhood             56761 non-null object
ZipName                  56761 non-null object
Avg Listing Price        56761 non-null int32
dtypes: int32(2), object(9)
memory usage: 4.8+ MB

In [52]: # drop columns not needed
sd_zip_crime.drop(['agency','Charge_Description_Orig','activityDate','BLOCK_ADDRESS','community','Name'], axis = 1,inplace=True)
```

- Add latitude & longitude for the zip codes

```
In [94]: # Merge Latitude and Longitude with zip Code
sd_zip_crime_pop_geo = pd.merge(sd_zip_crime_pop,latlong_zip[['ZipCode','Latitude','Longitude']], on='ZipCode', how='inner')

In [97]: sd_zip_crime_pop_geo.head(5)
```

Out[97]:

	ZipName	ZipCode	Avg Listing Price	CrimeCounts	City	Population	crimeRate	Latitude	Longitude
0	Borrego Springs, CA	92004	402984	17	Borrego Springs	3881	438.031435	33.184028	-116.26597
1	Rancho Santa Fe, CA	92067	4877557	43	Rancho Santa Fe	9535	450.970110	33.016492	-117.20264
2	San Diego, CA	92129	847213	269	San Diego	51536	521.965228	32.961014	-117.12510
3	San Diego, CA	92131	1221888	179	San Diego	32787	545.948089	32.918035	-117.08438
4	Julian, CA	92036	745463	19	Julian	3440	552.325581	33.027570	-116.53109

3. Exploratory Data Analysis

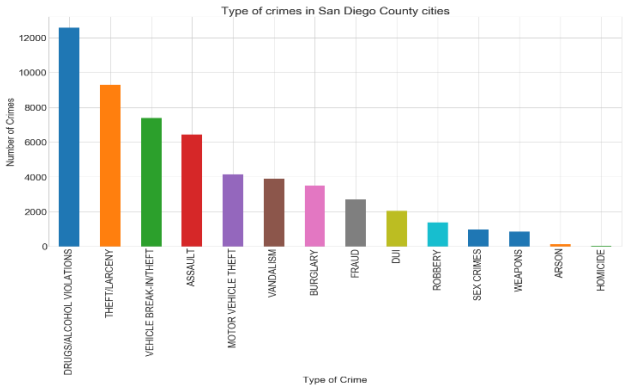


Figure 1- Number of crimes by category in San Diego county

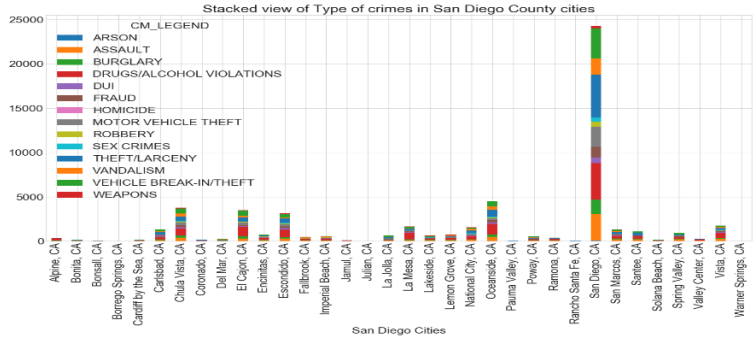


Figure 2 Type of Crimes in each city in San Diego county

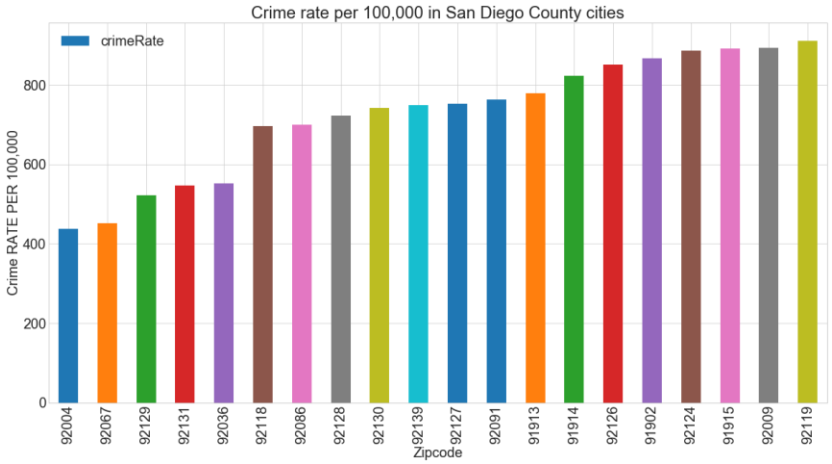


Figure 3 Crime rate per 100,000 by zip code in San Diego county

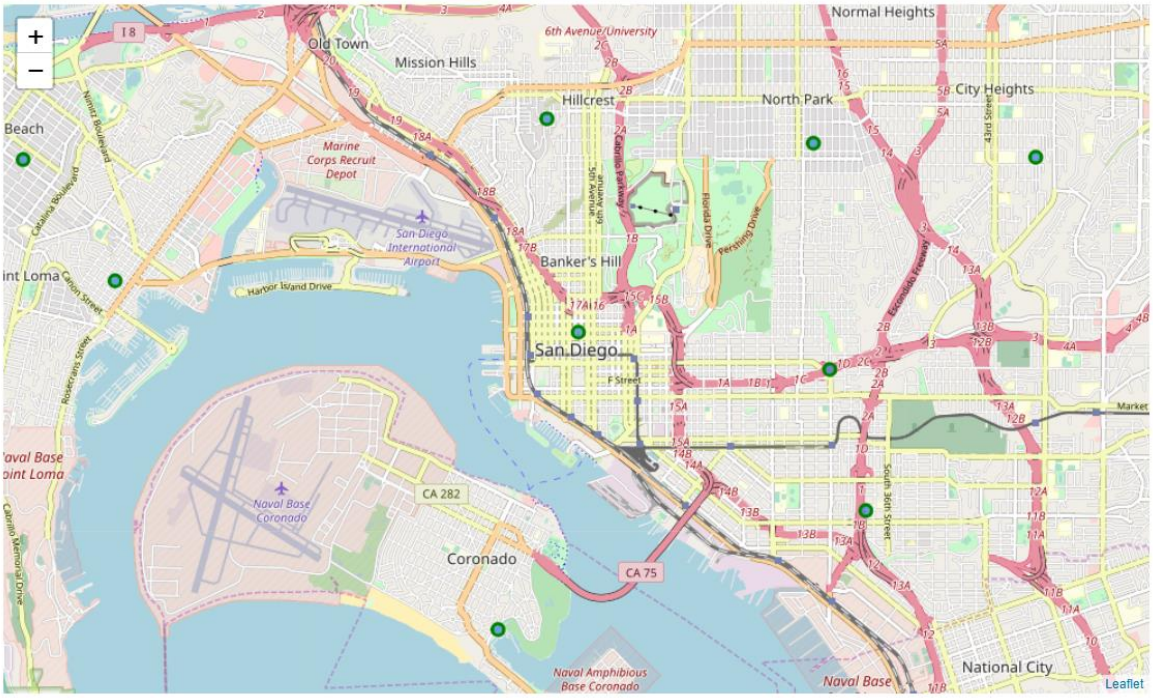
4. Clustering Neighborhood

k-means clustering algorithm is used for cluster the neighborhood of San Diego counties. Four Square REST API is leveraged to find venues.

- Figure lists the top zip codes in San Diego with less crime rate, the average home prices and the common venues

	ZipName	ZipCode	Latitude	Longitude	Avg Listing Price	CrimeCounts	Neighborhood	Population	crimeRate	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	
39	Rancho Santa Fe, CA	92067	33.016492	-117.202640	4877557	43	Rancho Santa Fe	9535	450.970110	0.0	American Restaurant	Restaurant	Grocery Store	T
65	San Diego, CA	92129	32.961014	-117.125100	847213	269	San Diego	51536	521.965228	0.0	Mexican Restaurant	Coffee Shop	Pizza Place	
67	San Diego, CA	92131	32.918035	-117.084380	1221888	179	San Diego	32787	545.948089	0.0	Mexican Restaurant	Coffee Shop	Pizza Place	
12	Coronado, CA	92118	32.682727	-117.174410	3870059	164	Coronado	23575	695.652174	0.0	Seafood Restaurant	Hotel	Ice Cream Shop	R
64	San Diego, CA	92128	32.998855	-117.070540	793956	343	San Diego	47490	722.257317	0.0	Mexican Restaurant	Coffee Shop	Pizza Place	
66	San Diego, CA	92130	32.946776	-117.219180	2051795	363	San Diego	48940	741.724561	0.0	Mexican Restaurant	Coffee Shop	Pizza Place	

Out[102]:



5. Conclusions

This project analyzed zip codes in San Diego county, its crime rate and house values. Used Four square to get the venues in the zip codes. K means algorithm was used to cluster neighborhoods. With varying data sources, the steps used can be used to analyze neighborhoods other states/cities.