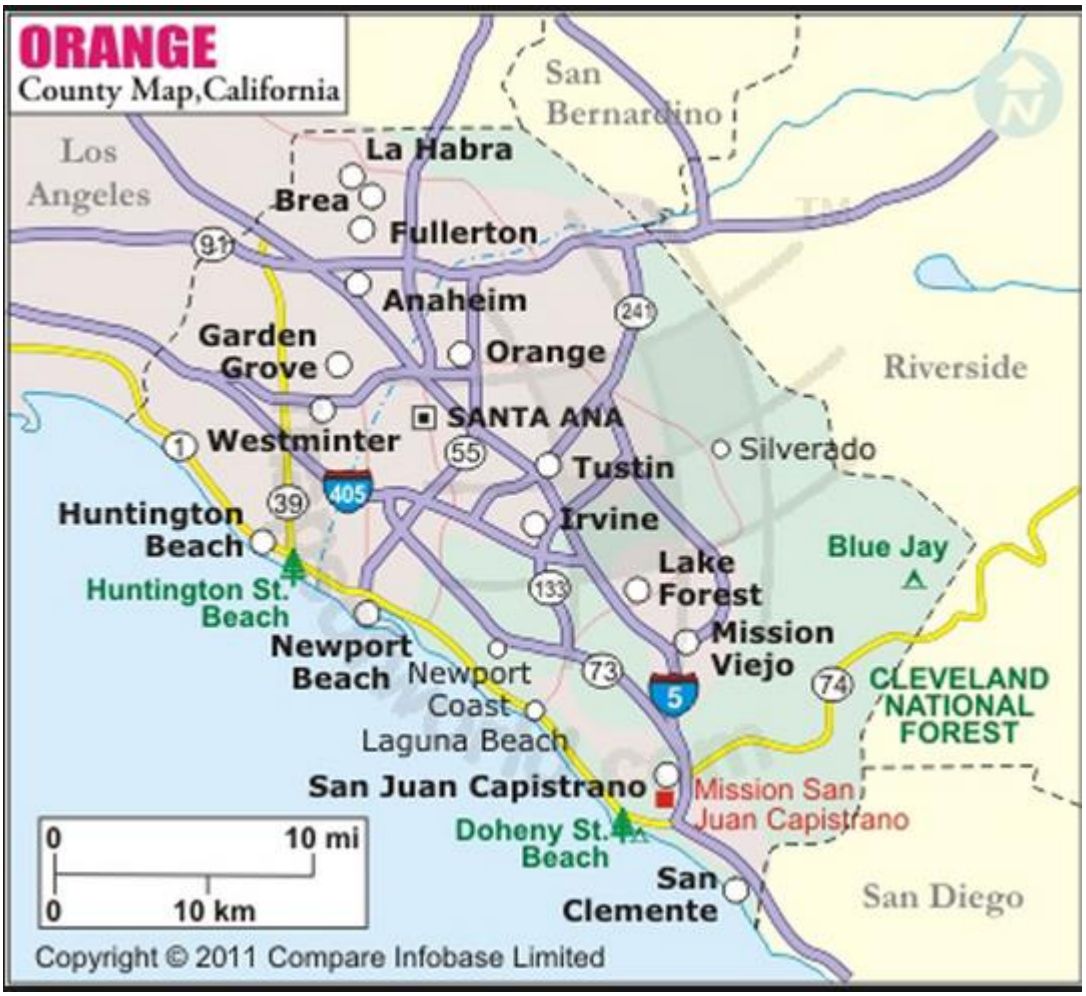
IBM Data Science Professional Certificate

Capstone Project

Researching Crime Rates and Coffee Shops in Orange County, California, United States of America

By: Kelly Anderson – April 2019



Description of the project:

I currently live in Orange County, in Southern California, and the town I live in is named Laguna Beach. There seems to be a lot of sushi restaurants, and Starbucks coffee shops. There also seems to be a lower crime rate in our town, compared to the surrounding cities in Orange County.

My questions that I would like to research, and answer are:

- Does the number of Starbucks and/or Sushi restaurants (per capita), have any correlation to crime rates?
- Of the surrounding cities in Orange County, which are similar to the city that I live in as far as sushi restaurants, Starbucks, and crime rates per capita

Description of the data – Collection Methods and Datasets:

The city crime rates and population data can be found on a Wikipedia page which shows a grid. This data was cited as taken from 2013, so it should be relevant given the timeframe.

https://en.wikipedia.org/wiki/Orange_County,_California

Notes: United States Department of Justice, Federal Bureau of Investigation. Crime in the United States, 2012, Table 8 (California). Retrieved November 14, 2013.

DataGrid:

Cities by population and crime rates [edit]

Cities by population and crime rates [hide]					
City	Population[91]	Violent crimes[91]	Violent crime rate per 1,000 persons	Property crimes[91]	Property crime rate per 1,000 persons
Aliso Viejo	48,999	43	0.88	415	8.47
Anaheim	344,526	1,279	3.71	10,070	29.23
Brea	40,253	74	1.84	1,292	32.10
Buena Park	82,505	206	2.50	2,066	25.04
Costa Mesa	112,635	254	2.26	4,079	36.21
Cypress	48,976	56	1.14	1,018	20.79
Dana Point	34,172	65	1.90	604	17.68
Fountain Valley	56,674	106	1.87	1,469	25.92
Fullerton	138,455	452	3.26	3,937	28.44
Garden Grove	175,079	439	2.51	4,017	22.94
Huntington Beach	194,677	313	1.61	5,470	28.10
Irvine	217,528	110	0.51	3,304	15.19
Laguna Beach	23,283	57	2.45	548	23.54
Laguna Hills	31,090	29	0.93	620	19.94
Laguna Niguel	64,533	47	0.73	764	11.84
Laguna Woods	16,590	4	0.24	148	8.92
La Habra	61,731	147	2.38	1,150	18.63

The numbers of Starbucks and Sushi restaurants will be collected using FourSquare API calls. This will be done by looping through all cities and compiling the unique items that meet the requirements.

Example DataFrame from results from Foursquare API:

```
[269]: dfc.head()
```

[269]:

	id	location.address	location.city	name
0	4b4e6d97f964a5206bed26e3	27020 Alicia Pkwy	Laguna Niguel	Starbucks
1	4ac8ea1cf964a52045bd20e3	23411 Aliso Viejo Pkwy	Aliso Viejo	Starbucks
2	4bca132a511f9521d5f5aec7	27072 La Paz Rd	Aliso Viejo	Starbucks
3	4b3bc69df964a520e97a25e3	28391 Marguerite Pkwy	Mission Viejo	Starbucks
4	4baa360af964a5209e533ae3	25630 Alicia Pkwy	Laguna Hills	Starbucks

```
[150]: dfc.to_csv('Starbucks_out.csv')
```

The datasets can then be merged to create the datapoints needed for the mapping, graphing, and correlation models.

```
[296]: df_comb
```

[296]:

	City	StarBucks_Count	Population	Violent_Crimes	Violent_crime_rate	Property_crimes	Property_crime_rate	latitude	longtitude
0	Aliso Viejo	7	48999	43	0.88	415	8.47	33.576138	-117.725812
1	Anaheim	32	344526	1279	3.71	10070	29.23	33.834752	-117.911732
2	Brea	11	40253	74	1.84	1292	32.10	33.917044	-117.888856
3	Buena Park	7	82505	206	2.50	2066	25.04	33.870413	-117.996217
4	Costa Mesa	21	112635	254	2.26	4079	36.21	33.663339	-117.903317
5	Cypress	5	48976	56	1.14	1018	20.79	33.824824	-118.039937
6	Dana Point	5	34172	65	1.90	604	17.68	33.466972	-117.698108
7	Fountain Valley	7	56674	106	1.87	1469	25.92	33.703815	-117.962735
8	Fullerton	21	138455	452	3.26	3937	28.44	33.870821	-117.929417
9	Garden Grove	13	175079	439	2.51	4017	22.94	33.774629	-117.946372
10	Huntington Beach	22	194677	313	1.61	5470	28.10	33.678334	-118.000017
11	Irvine	34	217528	110	0.51	3304	15.19	33.685697	-117.825982
12	La Habra	6	61731	147	2.38	1150	18.63	33.933016	-117.944777
13	Laguna Beach	3	23283	57	2.45	548	23.54	33.542089	-117.783415
14	Laguna Hills	5	31090	29	0.93	620	19.94	33.594876	-117.688207

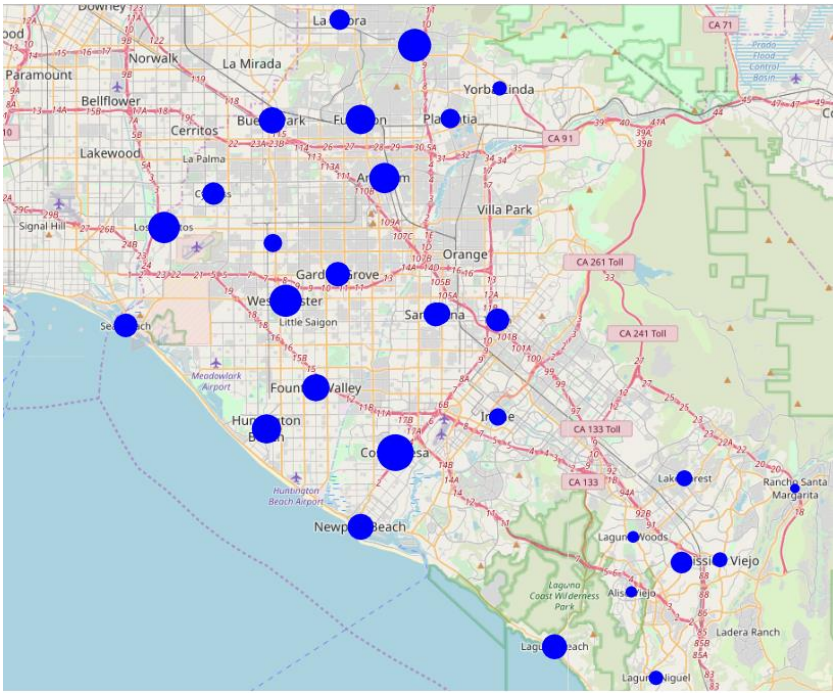
These datasets should provide a enough records to try to draw conclusions as to relationships, as well as classify the cities to find common groups.

Methodology

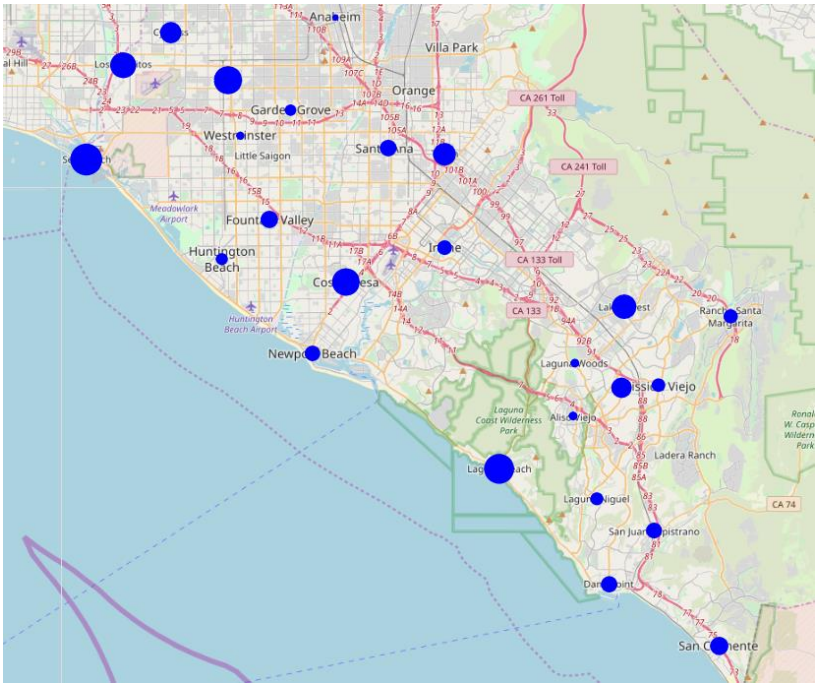
Exploratory data analysis

To take a closer look at the datasets that were generated in the data collection, I used maps to look at the volumes of the cities compared to each other. Below are two maps that show Crime Rates and Restaurants (Starbucks and Sushi) by city using circle markers, with size representing the values:

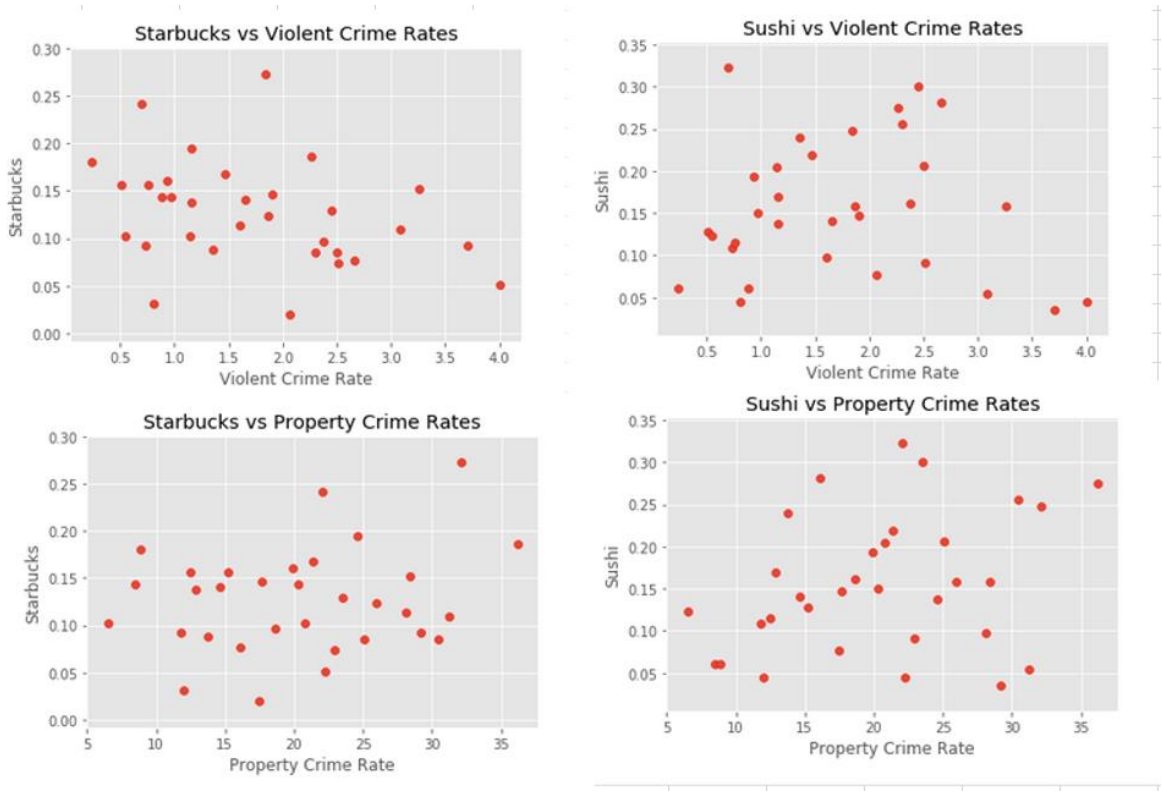
Crime (Property per capita by city):



Starbucks and Sushi (combined values per capita). I noticed some of the tourist and beach cities have a higher per capita amount of values in this map:



I then looked at the ranges of values and plotted the main data points against each other using scatterplot graphs to look at relationships.



As you can see, there does not appear to be any clear relationships between the variables. This is proven in the next section using Simple and Multiple Linear Regression.

Results

Using Simple Linear Regression, we find the evaluation values as below when comparing the following values.

This was done using a 80/20 split for training and test data for the models.

Sushi per capita vs Property Crime Rates

```
Mean absolute error: 6.46
Residual sum of squares (MSE): 68.77
R2-score: -8.45
```

Starbucks per capita vs Property Crime Rates

```
Mean absolute error: 5.48
Residual sum of squares (MSE): 46.52
R2-score: -42.16
```

The Sushi restaurants appear to have a slightly higher correlation, but neither was very accurate.

Multiple Linear Regression

Using both values compared against the Property Crime rates we come up with the following values for the model:

```
Residual sum of squares: 70.04
Variance score: -0.97
```

Since the Variance score is -.97, there does not appear to be a correlation.

Clustering using K-Means

For the second part of the problem, we want to find similar groups of cities based on Starbucks, Sushi, population, and crime rates.

After feeding in the model with the scaled values, we create four clusters.

Now we can create a profile for each group, considering the common characteristics of each cluster.

Group 0- Medium Population, Medium crime, and lots of Starbucks and Sushi

Group 1 - High population, highest crime, Lowest Starbucks and Sushi

Group 2 - Medium population, lowest crime, medium Starbucks and Sushi

Group 3 - Lowest population, Medium crime, Low Starbucks and high Sushi

The characteristics is based on the mean data values of each groups as shown below:

	Population	Violent_crime_rate	Property_crime_rate	Starbucks_PC	Sushi_PC
Labels					
0	62238.0	1.391667	26.043333	0.204258	0.232801
1	212854.5	3.031667	27.035000	0.098625	0.080372
2	79833.7	0.826000	12.322000	0.128471	0.110207
3	48913.7	2.062000	20.938000	0.095249	0.203220

We also wanted to see which cities were closest to the city in which I live, Laguna Beach. They are as shown below:

	City	Population	Violent_crime_rate	Property_crime_rate	Starbucks_PC	Sushi_PC	Labels
3	Buena Park	82505	2.50	25.04	0.084843	0.206048	3
5	Cypress	48976	1.14	20.79	0.102091	0.204182	3
6	Dana Point	34172	1.90	17.68	0.146319	0.146319	3
7	Fountain Valley	56674	1.87	25.92	0.123513	0.158803	3
12	La Habra	61731	2.38	18.63	0.097196	0.161993	3
13	Laguna Beach	23283	2.45	23.54	0.128849	0.300649	3
17	Lake Forest	79166	1.35	13.74	0.088422	0.240002	3
18	Los Alamitos	11728	2.30	30.44	0.085266	0.255798	3
22	Placentia	51778	2.07	17.50	0.019313	0.077253	3
28	Stanton	39124	2.66	16.10	0.076679	0.281157	3

Observations and Recommendations

Overall, there the dependent variable data doesn't appear to be a good indicator of crime rates. I think there a lot of other variables that make the model not very accurate. The cities in the area are very large and diverse. I think for the data to be very good predictor, it would have to be looked at on a much small area for each record. If this was done by neighborhood, I would expect the data to be a much better fit than the model which used data by city.

Conclusion

My conclusion is that at a city-wide level there is not a good correlation between the number of Starbucks and sushi restaurants and the crime rates for those cities.

The clusters do work well though to find similar groups of cities in Orange County based on those factors.