2020

Cincinnati Crime Study

CORRELATIONS WITH FOURSQUARE DATA

ALEX M

I. Introduction and Business Problem

Public officials often wonder about and study the factors which may cause crimes to occur in their city. The United States is a prime example of a "melting pot" of many different cultures, religions, values, and ethnicities. As a result, there can be deep pockets of culture even within an individual city.

This is certainly the case with Cincinnati, Ohio. Ohio is a population center of the Midwest United States, having the 7th-ranked total population among all states while only being the 35th-largest by square mileage. While crime is certainly not fairly explained by race, ethnicity, or religion, it can be fairly predicted by behavioral factors. In this short study I examine one such factor: **alcohol use.**

The hypothesis is as follows: the larger the concentration of opportunities to drink alcohol within a certain area, the higher propensity for crimes. If this hypothesis is true, new residents of a city should steer clear of bars, gas stations, and liquor stores when choosing a place to live, because those are the establishments that offer opportunities for people to purchase and drink alcohol. If the hypothesis is not supported by the evidence, then new residents must look to other factors to determine whether crime can be explained. Therefore the key audience of this study is **new residents of Cincinnati, or those looking to move between locations within Cincinnati.**

For this project I will be pulling data using the Foursquare API, and merging it with Cincinnati crime data via the Area Vibes user service. The final model development dataset will be trained via Linear SVM Machine Learning as well as OLS Linear Regression to determine **if crime can be predicted by the presence of alcohol-centered venues.**

II. Data

There are two key sources of data for this project, as follows:

First, venue data will be pulled using the Foursquare API related to alcohol-centered establishments. This includes **bars, gas stations, and liquor stores.** The following queries in Python will be used to pull the data (Figure 1 below):

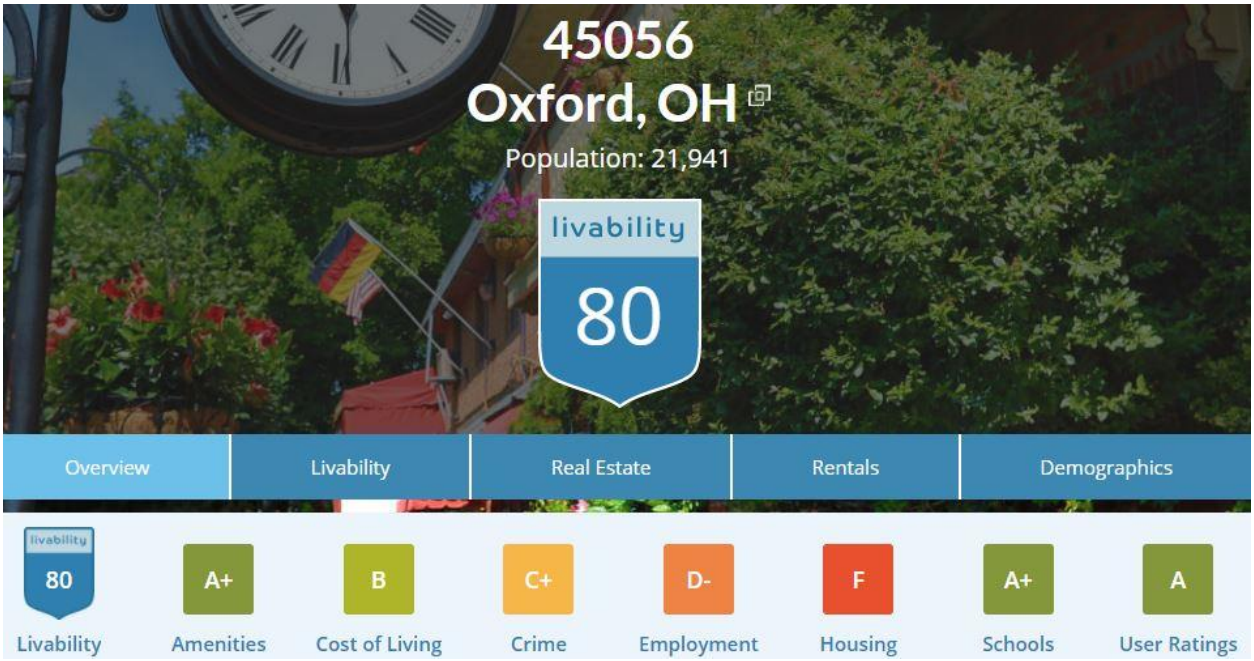
Figure 1: Foursquare Queries

```
[41]: #The first Foursquare query will be Bars within 30 miles of Cinci;
      search_query1 = 'Bar'
      radius = 50000
      print(search_query1 + ' .... OK!')

      Bar .... OK!

[42]: #The second Foursquare query will be Gas Stations within 30 miles of Cinci;
      search_query2 = 'Gas Station'
      radius = 50000
      print(search_query2 + ' .... OK!')

      Gas Station .... OK!

[43]: #The third Foursquare query will be Liquor Stores within 30 miles;
      search_query3 = 'Liquor'
      radius = 50000
      print(search_query3 + ' .... OK!')

      Liquor .... OK!

[44]: #Trying to find more liquor stores via query Spirits;
      search_query4 = 'Spirits'
      radius = 50000
      print(search_query4 + ' .... OK!')

      Spirits .... OK!
```

As can be observed in Figure 1 above, a separate query was utilized for “Bar”, “Gas Station”, “Liquor”, and “Spirits”. The purpose of two separate queries for liquor and spirits is because the name of the liquor store may likely include either key word. Note that the queries return the name and type of establishment as well as the zip code and latitude/longitude of the place. The number of each type of establishment will be aggregated by zip code.

Second, crime data will be pulled from the Area Vibes website. Area Vibes is an online service which helps new residents determine where they should live. Based on the user-entered zip code, Area Vibes provides a Crime Grade on a scale of A+ (best possible grade) to F (worst possible grade). Figure 2 below shows an example of the Area Vibes scorecard.

Figure 2: Cincinnati Crime Data



The data for all zip codes within 30 miles of Cincinnati center was downloaded and saved as a CSV file. The relevant field in the dataset is the Crime score, which is supposedly explained by the presence of Bars, Gas Stations, and Liquor Stores.

III. Methodology

Ideally, we would like to be able to use some type of machine learning algorithm to determine if crime is explained by geographical factors. In this example, Support Vector Machine was used with Polynomial Features for in-sample scoring along with OLS Regression for the interpretability of model coefficients.

First, however, the data had to be cleaned and merged into a final model development dataset. After bars, gas stations, and liquor stores within 30 miles of Cincinnati city center were pulled from Foursquare, all attributes were dropped from the data except for Venue Name, City Name, and Zip Code. Then, each type of venue (bar, gas station, liquor store) was split into a Pandas data frame and manually labeled according to its venue type. Figure 3 below shows an example.

Figure 3: Pandas Data Frames

```
#rename columns
df_bars2.rename(columns={"name": "Venue", "location.postalCode": "Zip","location.city":"City"},inplace=True)
df_gas2.rename(columns={"name": "Venue", "location.postalCode": "Zip","location.city":"City"},inplace=True)
df_liquor2.rename(columns={"name": "Venue", "location.postalCode": "Zip","location.city":"City"},inplace=True)
df_spirits2.rename(columns={"name": "Venue", "location.postalCode": "Zip","location.city":"City"},inplace=True)

df_bars2.dropna(inplace=True)
df_gas2.dropna(inplace=True)
df_liquor2.dropna(inplace=True)
df_spirits2.dropna(inplace=True)

df_bars2['Type']="Bar"
df_gas2['Type']="Gas Station"
df_liquor2['Type']="Liquor Store"
df_spirits2['Type']="Liquor Store"
```

Then, the Area Vibes dataset was brought in via local CSV file. Within the file, all of the zip codes in Cincinnati matching the zip codes from the Foursquare data were pulled in along with their Crime Letter Grade, corresponding Crime Score (on a scale of 1-13 with 1 being the lowest crime and 13 being the highest crime), and area population. An example of the Area Vibes dataset can be observed in Figure 4 below.

Figure 4: Area Vibes Dataset

	Zip	Letter	Crime	Population
0	45202	C+	7	2226
1	41001	A+	1	2637
2	41071	D+	10	15321
3	45219	F	13	16520
4	45209	C+	7	10214

Next, we would like to aggregate the venue data by venue type and zip code, so it can be conveniently merged with the Area Vibes dataset to get the total of each type of venue along with the crime score and zip code. Each of the “bar”, “gas station”, and “liquor store” data frames were aggregated by zip code and then outer merged with the Area Vibes dataset. Figure 5 below shows an aggregation example for the “bar” data frame.

Figure 5: Bars Aggregation

	Zip	Bars
0	41011	4
1	41071	3
2	45202	33
3	45219	2

This process was repeated again for the "gas station" and "liquor store" data frames and then all three of the individual data frames were outer merged together to preserve zip codes. Any resulting NaN values were filled with zero because NaNs after this outer merge methodology represent zero counts of a venue-type zip-code combination. Figure 6 shows the head of the merged venue data.

Figure 6: Merged Venue Data

	Zip	Bars	Gas Stations	Liquor Stores
0	41011	4.0	2.0	12.0
1	41071	3.0	0.0	9.0
2	45202	33.0	0.0	4.0
3	45219	2.0	0.0	1.0
4	41042	0.0	3.0	2.0
5	41048	0.0	1.0	1.0
6	41051	0.0	1.0	2.0
7	41076	0.0	2.0	5.0
8	45005	0.0	1.0	0.0
9	45014	0.0	2.0	0.0

The merged venue data was then inner-merged on the Area Vibes data to achieve the final model dataset. An example can be observed in Figure 7 below.

Figure 7: Final Data

	Zip	Letter	Crime	Population	Bars	Gas Stations	Liquor Stores
0	45202	C+	7	2226	33.0	0.0	4.0
1	41071	D+	10	15321	3.0	0.0	9.0
2	45219	F	13	16520	2.0	0.0	1.0
3	45209	C+	7	10214	0.0	1.0	3.0
4	45206	F	13	298011	0.0	2.0	0.0
5	45203	C+	7	3996	0.0	1.0	0.0
6	45247	B-	6	19068	0.0	2.0	0.0
7	45245	B+	4	7087	0.0	2.0	0.0
8	45230	F	13	11430	0.0	1.0	0.0
9	45220	C+	7	8188	0.0	1.0	0.0
10	45248	A+	1	11495	0.0	1.0	0.0
11	41048	A	2	6095	0.0	1.0	1.0

The final dataset has 48 rows and 8 columns before transformation. Ideally, we would like to be able to explain variation in "Crime" by the counts of "Bars", "Gas Stations", and "Liquor Stores" by Zip code. First, Poly SVM of degree 3 was tried in-sample and then OLS Regression was tried in-sample as well.

IV. Results

The possible crime ratings range from 1 to 13 on an ordinal scale. The total number of bars, gas stations, and restaurants per crime rating is observed in Figure 8 below.

Figure 8: Venue Counts by Crime Rating

	Crime	Zip	Population	Bars	Gas Stations	Liquor Stores
0	1	870336	220252	0.0	8.0	40.0
1	2	88049	9792	0.0	2.0	1.0
2	4	90395	19728	0.0	3.0	0.0
3	5	90255	56422	0.0	2.0	1.0
4	6	90283	39604	0.0	2.0	1.0
5	7	225890	46565	33.0	3.0	9.0
6	10	123188	62377	3.0	3.0	13.0
7	12	90049	97054	0.0	1.0	1.0
8	13	493051	1368525	2.0	12.0	7.0

As can be observed in Figure 8 above, the original hypothesis will not be supported by the data. However, the results still bear an interesting discussion. Not many bars were found by the Foursquare API throughout the city – these venues might not be listed on Foursquare or may have names that are difficult to generalize with keywords. Most of the bars that were found were within Downtown Cincinnati, which is an area with many restaurants and an average crime rating. We will not be able to explain variation in crime ratings by the number of bars because most of the bars found were within a single zip code with average crime.

The "Gas Stations" variable provides perhaps the most hope for its variation explained by crime rating. However, low-crime areas and high-crime areas both had a larger number of gas stations. The "Liquor Stores" variable actually behaved the opposite of what was expected by the original hypothesis, with the lowest-crime areas actually having by far the most liquor stores. For the purpose of discussion, OLS Regression and Poly SVM were fit, as can be observed in Figure 9 below.

Figure 9: Model Results

OLS Regression Results

Dep. Variable:	Crime	R-squared:	0.094
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	1.551
Date:	Mon, 21 Dec 2020	Prob (F-statistic):	0.214
Time:	19:19:34	Log-Likelihood:	-145.57
No. Observations:	49	AIC:	299.1
Df Residuals:	45	BIC:	306.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.2067	1.223	4.258	0.000	2.744	7.670
Bars	0.1312	0.156	0.843	0.404	-0.182	0.445
Gas Stations	1.4177	0.870	1.629	0.110	-0.336	3.171
Liquor Stores	-0.3445	0.411	-0.838	0.406	-1.173	0.484

Omnibus:	18.671	Durbin-Watson:	1.742
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5.438
Skew:	0.500	Prob(JB):	0.0659
Kurtosis:	1.710	Cond. No.	9.64

```
In [90]: clf.score(train_x_poly,model_y)

Out[90]: 0.5714285714285714
```

V. Discussion

As can be observed in Figure 9 above, there were simply too many data limitations to support the original hypothesis that alcohol-related venues can explain crimes. In the OLS Regression results, we can see that none of the three venue types were statistically significant at the 10% level, and the model coefficient for Liquor Stores was actually negative. As expected, Gas Stations showed the most evidence for statistical significance in the positive direction, but the p-value missed weak significance by a couple of points. The Poly SVM score was also poor, showing that even with poly transform the model does not explain much variation in crime. Because most of the bars in the dataset were concentrated within a single zip code, and liquor stores were highly concentrated in low-crime areas, no sound conclusion can be made based on crime rates and alcohol-venue locations. More research and study is required.

VI. Conclusion

No conclusion can be made in favor of Bars, Gas Stations, and Liquor Stores having association with Crime. There are a few reasons for this – almost all of the bars in the dataset were concentrated in a single zip code (Downtown Cincinnati) which has an average crime rating. Additionally, there were almost as many gas stations in low crime areas as in high crime areas, and there were actually many more liquor stores in low crime areas than in high crime areas. Overall, based on the data available, we conclude that a Cincinnati Community's Area Vibes Crime Score is not explained by its vicinity to Bars, Gas Stations, and Liquor Stores. Other behavioral factors must explain the Crime Rating.