Predicting Crime Levels in San Francisco Based on Foursquare Data

Colin Howarth
June, 2019

Problem

Summary

There are a variety of factors that may contribute to the crime level in different neighborhoods of a city. This project aims to identify whether Foursquare data can be used to predict crime levels in a neighborhood, and whether particular types of Foursquare location are particularly predictive of crime.

Audience

The output of this project may be of interest to city employees who are involved in planning for future zoning and development. It may also be of interest to police officers for planning staffing levels or for choosing locations for future police stations.

Data

Data Sources

Two primary data sources will be used:

1. The Foursquare Search API, which allows retrieval of a categorize list of venues for a given location.
2. San Francisco Police Department (SFPD) incident reports from 2018 onwards, available from the city's open data site.

Data Retrieval and Cleansing

SFPD Incident Reports

The SFPD data includes a lot of different columns. For example:

	Incident Datetime	Incident Date	Incident Time	Incident Year	Incident Day of Week	Report Datetime	Row ID	Incident ID	Incident Number	CAD Number	...	Longitude	point	SF Find Neighborhoods	Current Police Districts	Current Supervisor Districts	Analysis Neighborhoods	Zones as of 2018-06-05	OWED Public Spaces	Ma
0	2018/12/02 12:45:00 AM	2018/12/02	00:45	2018	Sunday	2018/12/02 01:56:00 AM	74374327130	743743	180908554	183360210.0	...	-122.404795	(37.78490829943, -122.40479506276)	32.0	5.0	10.0	8.0	NaN	NaN	
1	2018/12/01 08:30:00 PM	2018/12/01	20:30	2018	Saturday	2018/12/01 09:18:00 PM	74370071000	743700	180908112	183353564.0	...	-122.408036	(37.786409612811, -122.408036237445)	19.0	6.0	3.0	36.0	NaN	NaN	
2	2019/03/18 02:01:00 PM	2019/03/18	14:01	2019	Monday	2019/03/18 02:21:00 PM	78164004134	781640	190194129	190772267.0	...	-122.406699	(37.756833733806, -122.406699002688)	53.0	3.0	2.0	20.0	3.0	NaN	
3	2019/03/20 08:00:00 AM	2019/03/20	08:00	2019	Wednesday	2019/03/20 02:06:00 PM	78169706244	781697	190199583	190792201.0	...	-122.404865	(37.78400661242, -122.404864795177)	32.0	1.0	10.0	34.0	NaN	NaN	
4	2019/03/12 01:30:00 PM	2019/03/12	13:30	2019	Tuesday	2019/03/15 06:02:00 PM	78154706372	781547	196055103	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

The SFPD data was retrieved and cleaned by:

- Removing un-needed columns (such as the incident ID and date-time of each incident)
- Removing incident reports that don't have latitude and longitude included.
- Removing incident report types that don't relate to a crime (such as "Lost Property" and "Non-Criminal")

The city of San Francisco was then divided into a 20x20 grid, and each incident mapped to one of the grid squares. The total number of crime reports of each type were then calculated for each grid square.

Foursquare Venues Lists

For each grid square, the Foursquare API was called to retrieve the list of venues in that grid square across each of Foursquare's ten top level categories. The count of the venues of each type was thus calculated for each grid square.

Combined Dataset

The two datasets were then combined to produce a single table listing the number of incident reports (by type) and the number of Foursquare venues (by type) for each grid square.

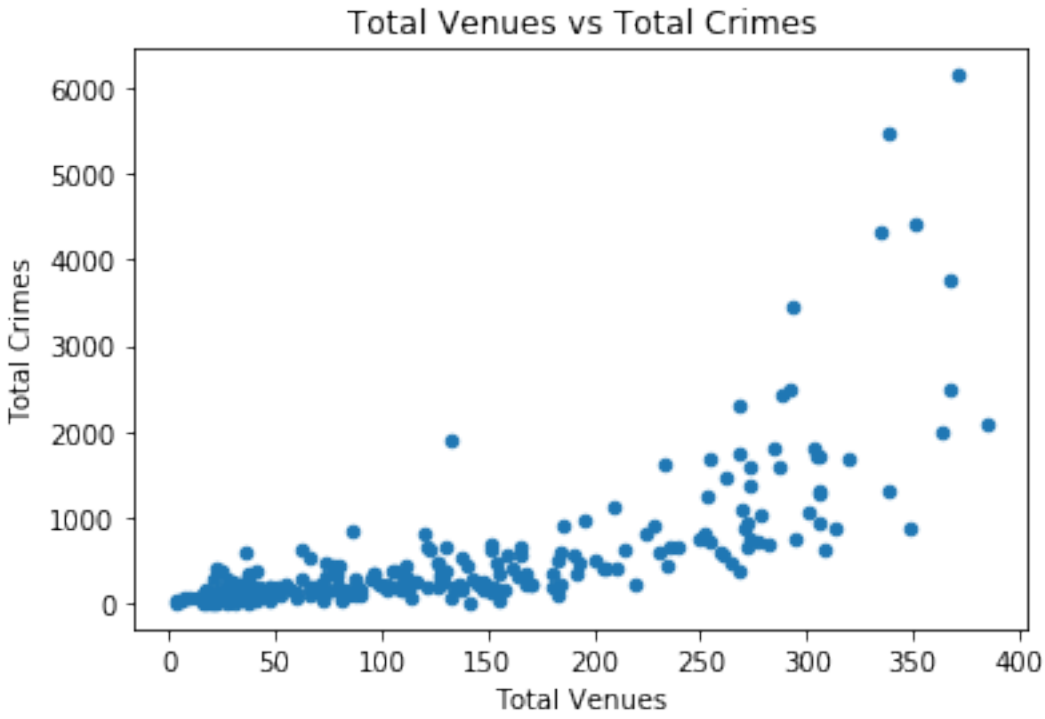
The combined dataset:

	LatLongBucket	Arts and Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	...	Robbery	Sex Offense	Stolen Property
0	0_1	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	...	0.0	0.0	1.0
1	0_3	1.0	0.0	0.0	3.0	0.0	7.0	3.0	0.0	1.0	...	2.0	0.0	1.0
2	0_5	1.0	0.0	0.0	5.0	3.0	11.0	8.0	7.0	8.0	...	6.0	0.0	1.0
3	0_6	3.0	0.0	0.0	1.0	1.0	5.0	5.0	1.0	6.0	...	10.0	0.0	1.0
4	0_7	0.0	1.0	0.0	16.0	3.0	10.0	13.0	2.0	37.0	...	13.0	1.0	0.0

Methodology

Exploring Correlation Between Factors

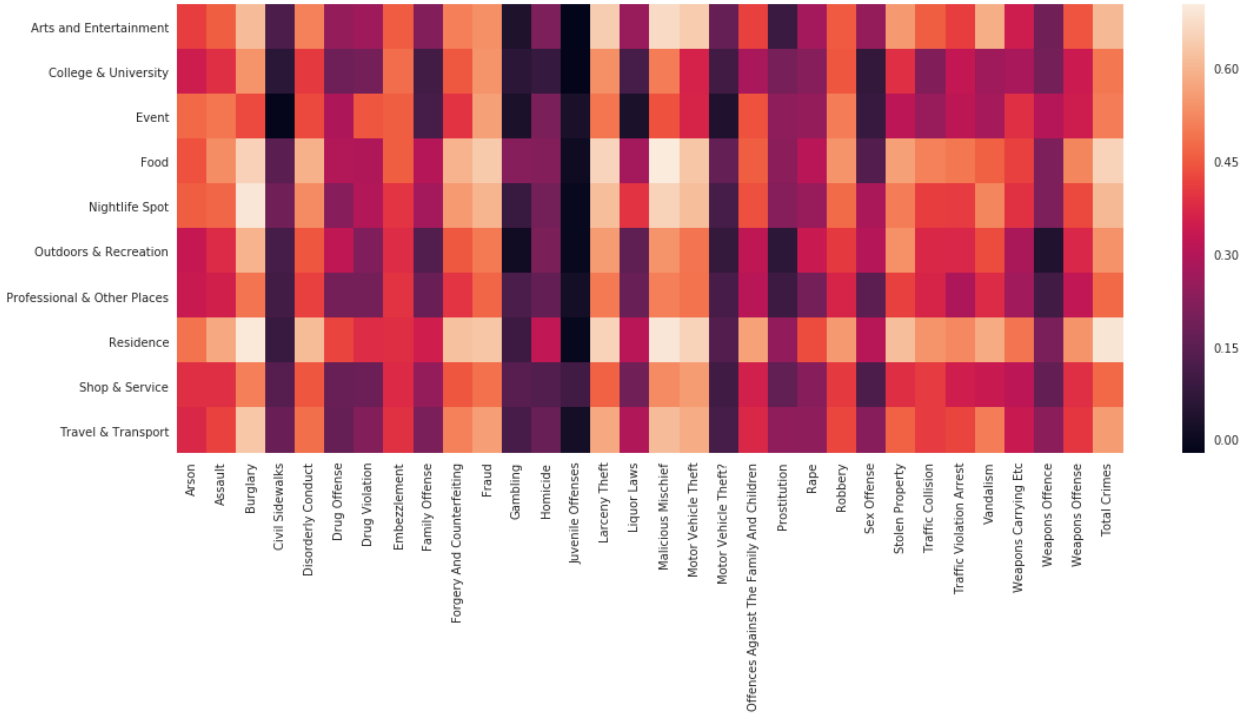
Drawing a scatter plot of the total number of venues against the number of crimes in each grid square, we can see that there does appear to be some relationship between the two.



The number of crimes rises as the total number of venues increases.

Relationship of Venue Type to Crime Numbers

Examining a heatmap of the correlation between different venue types and different types of crime incident, we can see that the relationship between the two numbers is much stronger for certain types of venue and certain types of crime.



This aligns with what one might expect intuitively – for example the number of burglaries is correlated with the number of residence-type venues.

In general, we can also note that the total number of crimes has a differing level of correlation with different types of venue.

Venue Type	Correlation to Total Crimes (R-squared)
Arts and Entertainment	**0.608578**
College & University	0.497027
Event	0.506816
Food	**0.656209**
Nightlife Spot	**0.608501**
Outdoors & Recreation	0.541738
Professional & Other Places	0.474160
Residence	**0.690965**
Shop & Service	0.475970
Travel & Transport	0.558462

Evaluating Prediction Models

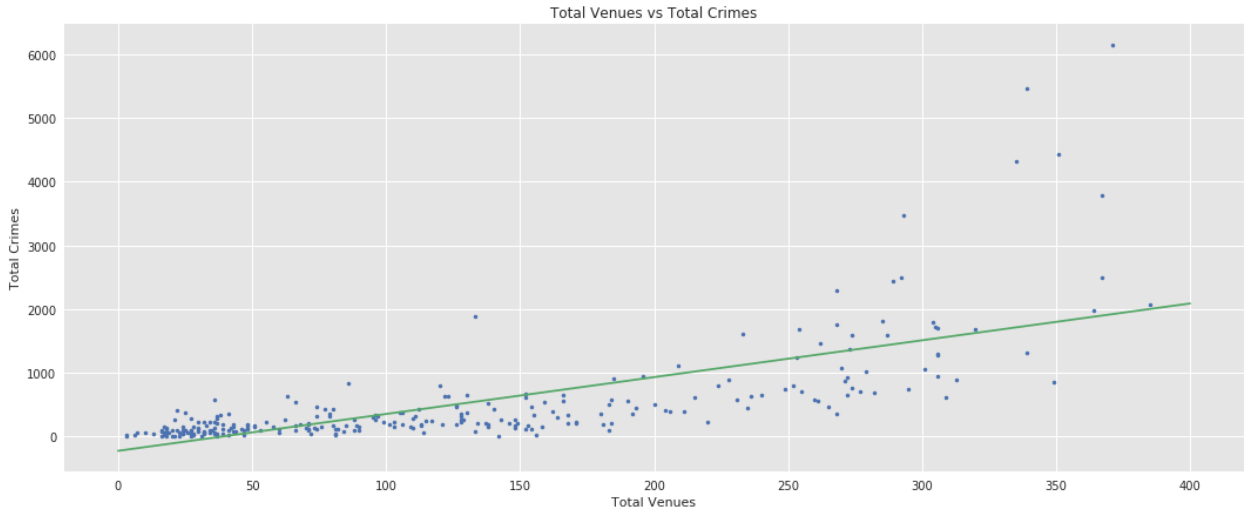
Three different types of model were evaluated - linear regression, multiple linear regression and polynomial fit.

Linear Regression Model

A single linear regression model was generated, taking Total Venues as an input and predicting Total Crimes as an output.

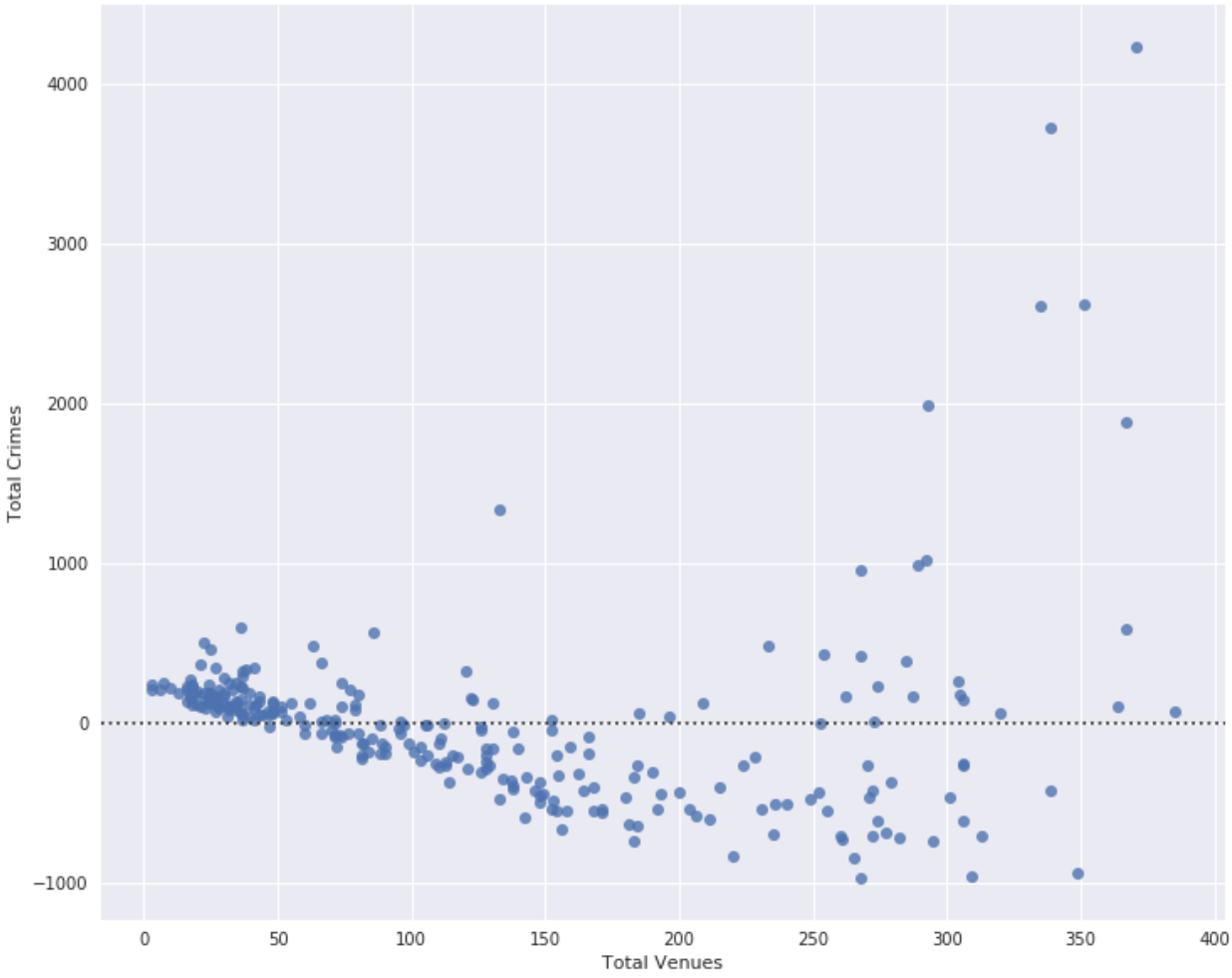
The generated model:
Total Crimes = 5.781 * Total Venues - 224.1

Visualizing this:



We can see that the model is likely to be weaker as the number of venues gets particularly high.

This is emphasized looking at a residual plot of the same data.



We can see that the data points are not evenly distributed above and below the line, meaning that a simple linear model is likely not the best fit.

The R-squared for this model was calculated to be 0.502, meaning that it can explain approximately 50% of the variation in the output.

Multiple Linear Regression Model

Reviewing the heatmap in the previous section, it was established that the following categories have the strongest correlation with the total number of crimes: 'Arts and Entertainment', 'Food', 'Nightlife Spot' and 'Residence'.

A multiple linear regression model was created based on these factors.

The model is as follows:
Total Crimes = 8.7062263 * 'Arts and Entertainment' + 16.31905148 * 'Food' - 4.58273384 * Nightlife Spot' + 29.26167973 * 'Residence'

Looking at the distribution plot for this model, we see that there is broadly a good fit of prediction to outcome, but the model is likely to underpredict at smaller counts.

The R-squared for this model is 0.564095529376.

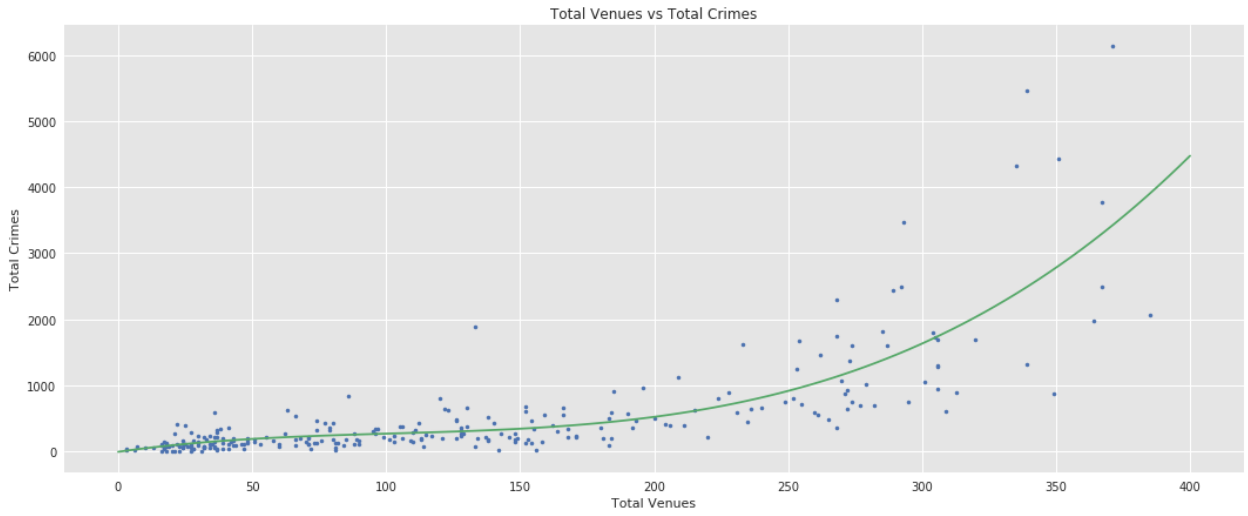
Polynomial Fit

A polynomial fit model was generated based on the Total Venues.

The model is:
Total Crimes = 0.0001464 x^3 - 0.04499 x^2 + 5.789 x - 11.4

Where x = the Total Venues for a grid square.

Plotting the datapoints against the predictions for the model, we can see that this model does a better job as the number of venues (and crimes) increases. There is however still significant noise visible.



The R-squared for this model is 0.661064573049.

Results

Comparing the R-squared for each model:

Model	R-Square
Single Linear Regression	0.502358210859
Multiple Linear Regression	0.564095529376
Polynomial Fit	0.661064573049

We can see that the polynomial fit is providing the strongest match to the data.

Discussion

Overall, there is certainly correlation between the number of venues of different types. A simple model based on the total venues is able to provide a good degree of fit with the data.

Future work could explore in more detail the relationships between different types of venue and the different types of crime incident. One approach might be to classify neighborhoods based on their mix of venue types, and to attempt to generate separate models based on each neighborhood type. For example, one might expect to see higher numbers of burglaries in primarily residential neighborhoods.

Future work could also explore the different venue types in more fine-grained detail. For example, the number of bars in a neighborhood might be a strong predictor for the number of disorderly conduct incidents that are reported.

Another avenue for future exploration would be to look at other Foursquare data types – for example, the number of checkins.

Conclusion

Foursquare data is able to provide a reasonable prediction of the number of crimes in different neighborhoods of the city of San Francisco.