IBM Data Science Professional Certificate

Toronto Neighborhoods' Business and their influence on Crime in 2020

Applied Data Science Capstone

Alberto Carrillo Ortega

METHODOLOGY

- *Introduction*
- *Data Sources*
- *Data Cleaning*
- *Preliminary Analysis*
- *Modeling*
- *Conclusions*

Introduction

We aim to gain Insights in the relationship between a given set of business/venues in a neighborhood on the city of Toronto, and the crime rate of that community.

We intend to answer some question such as:

- Is there a correspondence between a predominant type of business and an exceptional high or low criminality?
- Are more diverse neighborhoods less crime prone?
- What characterizes neighborhoods of notable delinquency?

To answer these questions and give some prescriptive claims we ought to arm ourselves with information.

This information will be extracted from two data sources: Toronto Open Data for crime rates in each neighborhood, and Foursquare API for business and landmarks in each region.

Data Acquisition

Two Sources of data:

TORONTO OPEN DATA

- Open data initiative born in 2009 and directed by the government of Toronto City.
- Offers a wide catalog of data sets ranging from festivals and events to Wellbeing Neighborhoods indexes.
- We will be using a dataset referring to neighborhoods crime rates.
- This dataset contains information about assaults, auto theft, robberies and other crimes.
- It’s recorded annually and there is information from 2014 to 2020.

FOURSQUARE PLACES API

- REST interface that allows us to access its services through simple HTTP requests.
- A search-and-discovery service developed by Foursquare Labs Inc.
- The system provides us with an easy-to-use tool for locating nearby places like restaurants, parks, and all kind of activities.
- We will use this API to obtain a set of venues nearby the coordinates of each neighborhood and relate these locales and their type with the crime rate of the neighborhood.

Data Cleaning

The data we gather from our sources needs some work to be usable.
The data we obtain from Toronto Open Data is a perfectly clean dataset.
This is not the case for Foursquare because as it is a REST API, the data we get from it, is in JSON format. We will have to manipulate the HTTP response in order to obtain a useful format.

TORONTO OPEN DATA

	_id	OBJECTID	Neighborhood	Hood_ID	F2020 Population Projection	Assault 2014	Assault 2015	Assault 2016	Assault 2017	...
0	1	1	Yonge-St.Clair	97	14083	16	25	34	25	...

FOURSQUARE PLACES API

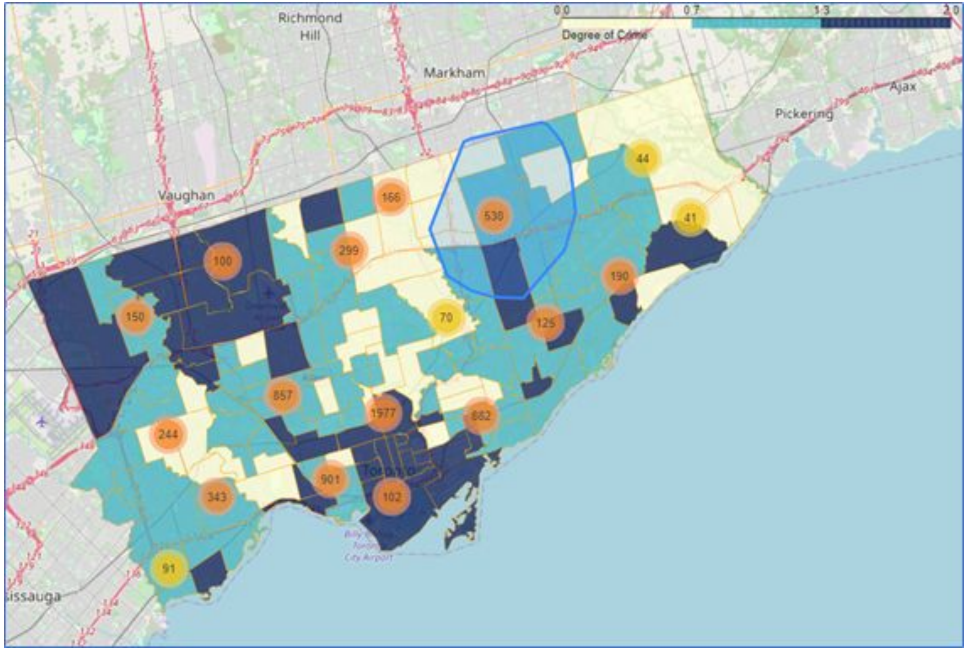
	Neighborhood	Venue	Venue Category	Venue Latitude	Venue Longitude
0	Yonge-St.Clair	Daeco Sushi	Sushi Restaurant	43.687838	-79.395652

RESULTING DATA SET

	Neighborhood	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Nightlife Spot	Travel & Transport	F2020_Population_Projection	SumCrimes	SumCrimes Cat
0	Yonge-St.Clair	38	10	11	0	3	1	14083	319.534220	Low

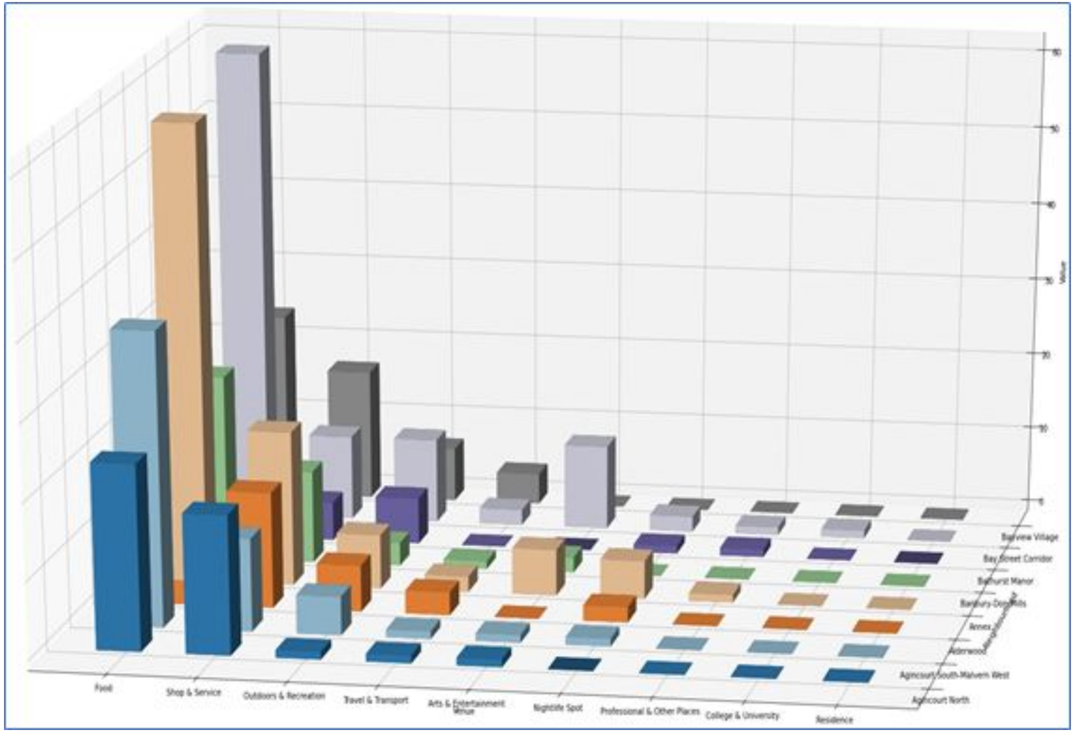
Preliminary Analysis

Representation of the Data



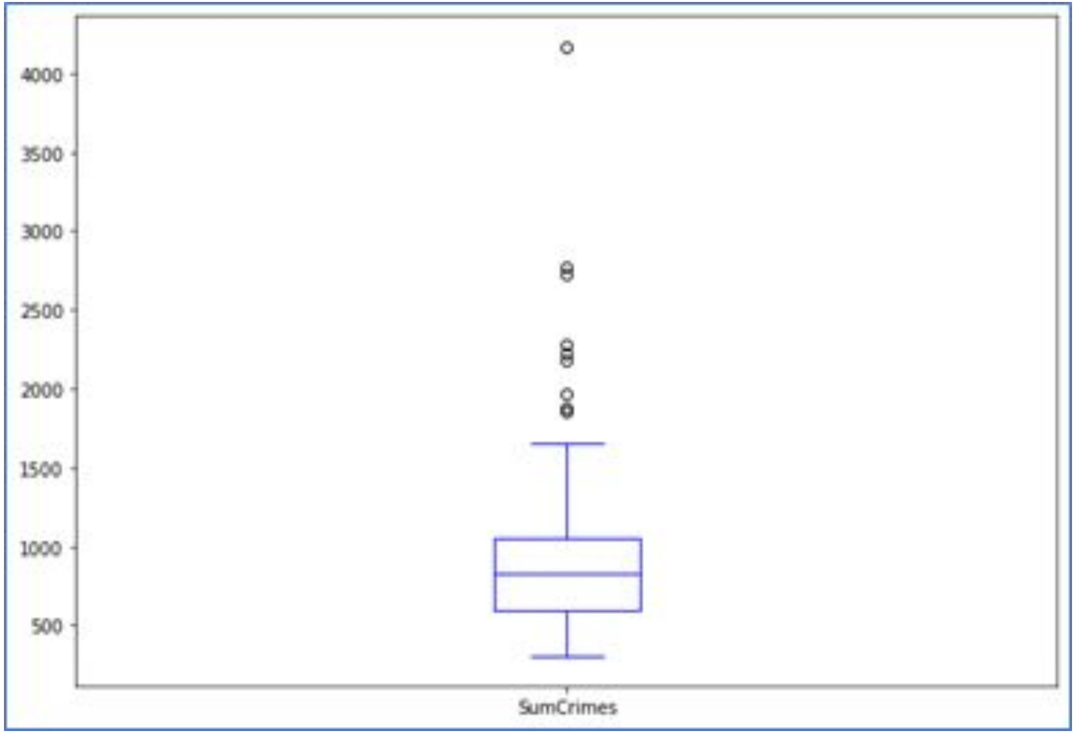
- Each neighborhood is colored by its crime rate. The darker the color, the higher the crime.
- Venues are displayed around the city, each venue relates to a neighborhood and is of a type.

Venues´ Categories Distribution



- Most Venues are of the Food type.
- The distribution of the rest of types is mostly the same across all the other neighborhoods.
- The possible reason for this distribution might be the origin of the data. Foursquare API is for commercial use.

Neighborhoods´ Crime Rates Distribution



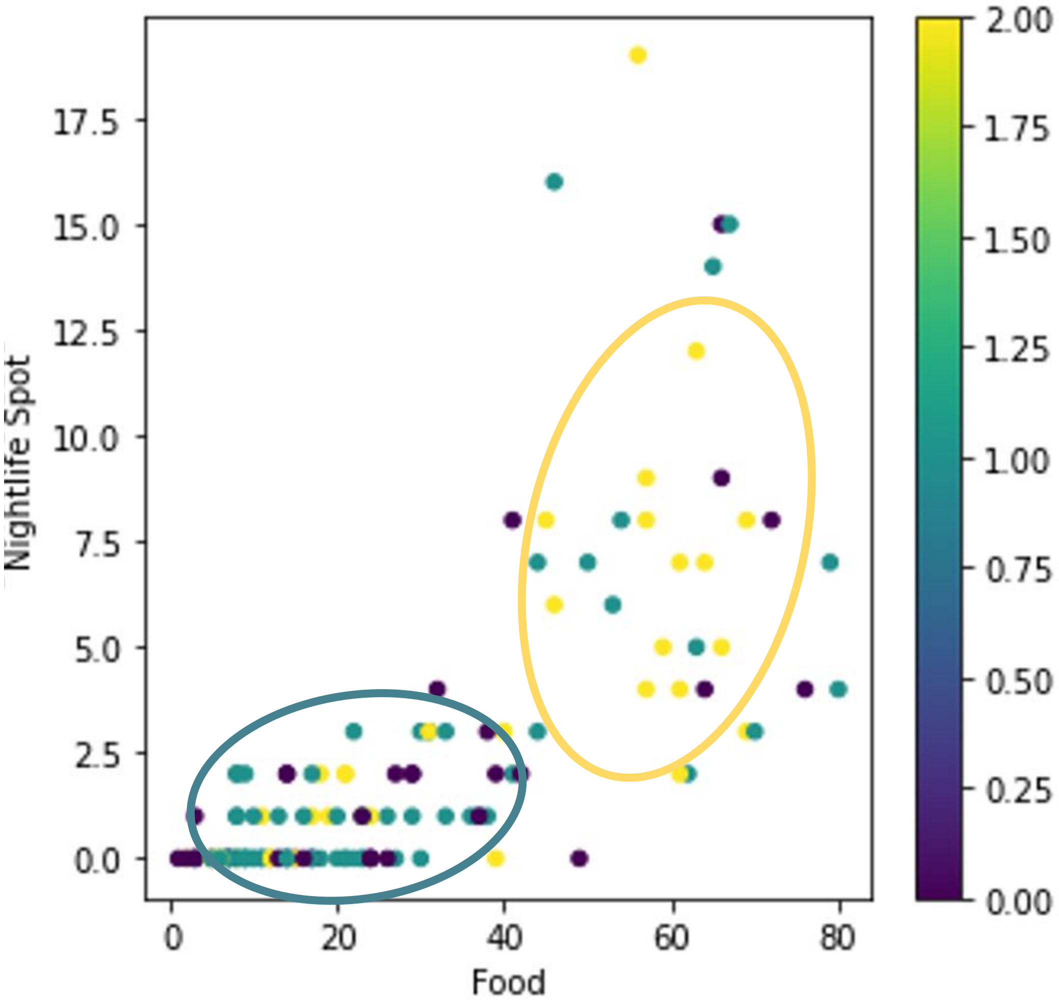
- There are some outliers in the upper degrees of crime rate.

Variables Against each other Colored by Crime

- By drawing one variable against the other and coloring the result on the category of crime rate we can display a lot of information in a small space.
- We are looking for possible regression lines or similar occurrences.
- The lack of diagonal lines in these graphs means that there is not a strong linear correlation. However, we can appreciate some clustering between the variables.

Variables Against each other Colored by Crime -Clustering-

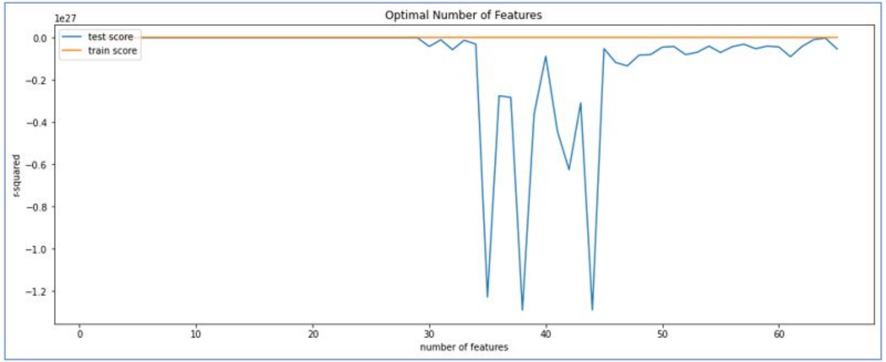
When drawing the number of Food venues against the Nightlife ones we can observe that if the number of both types are high, the majority of the crime rates are high, and the opposite is also true.



Predictive Modeling

A Regression Approach

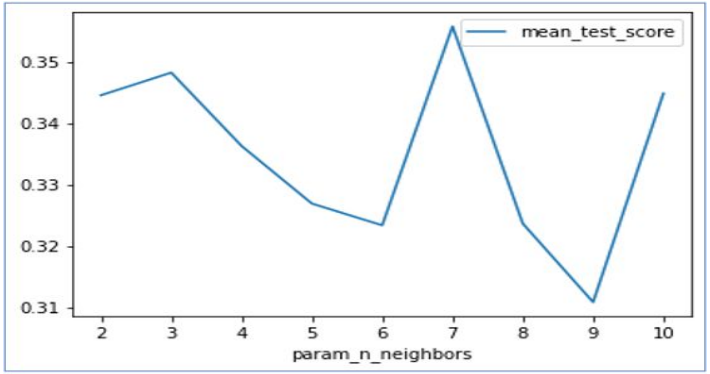
	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_n_features_to_select	...	mean_test_score	std_test_score	rank_test_score
0	0.111	0.013	0.000823	0.000021	1	...	-1,57E+04	3,09E+04	1
1	0.100	0.001	0.000804	0.000003	2	...	-1,01E+05	3,36E+05	2
2	0.100	0.002	0.000808	0.000008	3	...	-3,75E+05	8,09E+05	3
3	0.098	0.001	0.000805	0.000014	4	...	-8,11E+05	1,50E+06	4
4	0.100	0.005	0.000806	0.000007	5	...	-1,35E+06	2,34E+06	5



- Even though we have already stated that there is not a strong linear correlation between the variables, there might still exist a polynomial one.
- We need to transform the data from first degree to second degree. This can be archived through polynomial elevation.
- After processing our dataset our originals ten variables are now 66. To solve this issue, we make use of recursive feature elimination (RFE) and a grid search.
- The parameter mean_test_score is the average score of each fold in our grid search, and is the most important value. It says how well our model generalizes.
- In this case the values are near 0 which for our scoring method R^2 means that our model can't explain the data. It seems like the data is not appropriate for a regression model of second degree.

A Classification Approach
K-Neighbors Classification

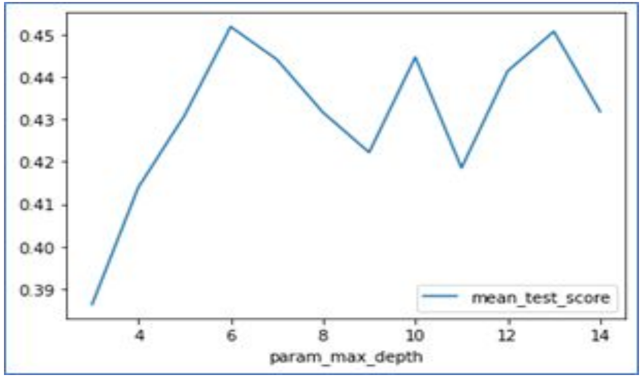
	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_n_neighbors	...	mean_test_score	std_test_score	rank_test_score
5	0.001612	0.000042	0.004027	0.001721	7	...	0.355768	0.060544	1
1	0.001778	0.000358	0.002986	0.000190	3	...	0.348230	0.095085	2
8	0.001614	0.000028	0.003186	0.000296	10	...	0.344836	0.028137	3
0	0.001995	0.000883	0.002898	0.000201	2	...	0.344541	0.091309	4
2	0.001630	0.000034	0.002997	0.000021	4	...	0.336271	0.056821	5



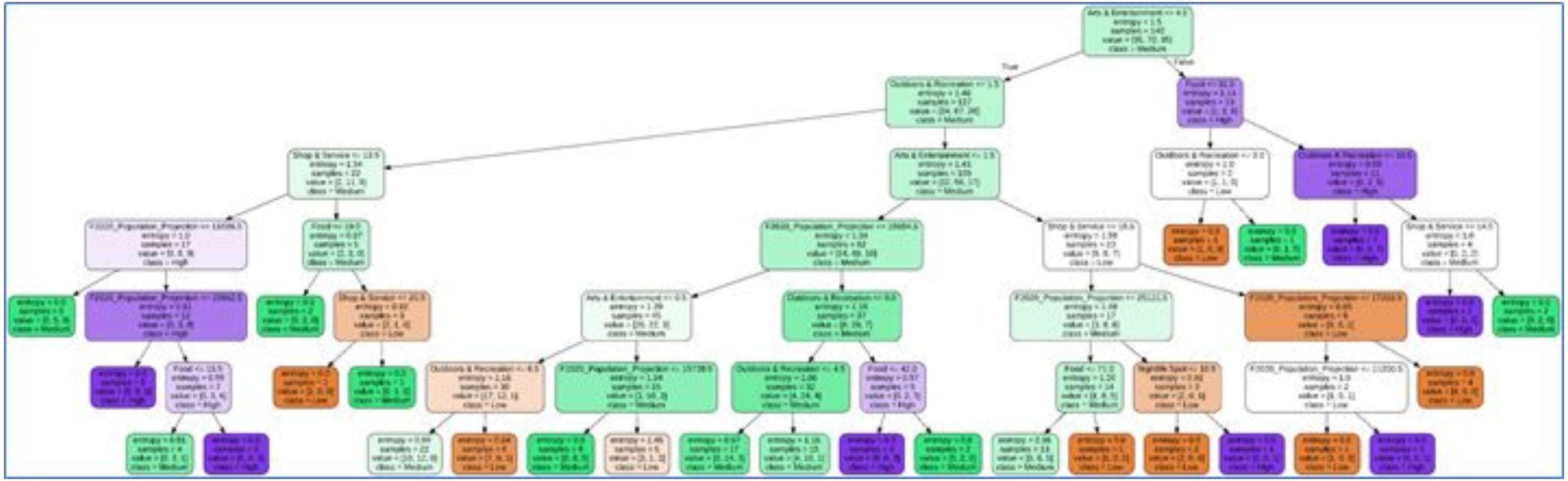
- This time we want to classify the neighborhoods. This is an easier endeavor as classification a neighborhood is akin to give it a value but instead, we are assigning it in an interval.
- There are plenty of classification models. For this problem we will choose K-neighbors because of the previous clustering we saw in the preliminary analysis
- To find the optimal value for K we can make use of the grid search. We will search for K between 2 and 11.
- The search returns that the optimal number for K is 7, and such model has a F1 score of 0.5824 and Jaccard Index of 0.4232.
- These scores are not near their maximum value, but they are much higher than the values we got from our linear regressor.

A Classification Approach
Decision Tree Classification

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_max_depth	...	mean_test_score	std_test_score	rank_test_score
3	0.002335	0.000448	0.002373	0.001077	6	...	0.451819	0.056028	1
10	0.002073	0.000029	0.001378	0.000004	13	...	0.450706	0.040228	2
7	0.002097	0.000046	0.001424	0.000029	10	...	0.444654	0.027192	3
4	0.002183	0.000236	0.001418	0.000013	7	...	0.444108	0.042790	4
9	0.002086	0.000045	0.001420	0.000013	12	...	0.441384	0.038887	5



- The decision of employing a decision tree in this problem comes from their good performance in the majority of classification problems, and the possibility of direct interpretation of their structure.
- Classification trees also have a hyperparameter. Their max deep. A low value won't be able to classify effectively and a high value overfits to our training set.
- To set this parameter we can make use of the grid search. We will search for values between 3 and 15.
- The optimal value for deep is 6, with a mean score of 0.4518. Other scoring returns values of 0.8087 for F1 score and of 0.68254 for Jaccard index.
- This means that there is a correlation between the neighborhoods' venues and their criminality.
- The lower the value of deep the easier to read the tree is, so it's good that the best value is six and not a higher one.



One of the pros of using decision trees is that they are interpretable by non-experts.

For example. from our tree structure we can conclude that:

- 75% of neighborhoods have two or less locales of "Arts & Entertainment" type. The tree also tells us that, of those which do have five or more, 9/13 have high crime rates.
- We can also say that low population usually influences the neighborhood so it doesn't have high crime rates. (This doesn't mean the crime rates are necessarily low).
- Etc.

A Classification Approach
Decision Tree Classification

Conclusions

Conclusions

After some work, we came to the conclusion that there is some degree of clustering between the neighborhoods.

However, even though there is a measure of similitude between same type, this correlation is not very strong.

We also found out that our data is unfit at all for a linear/polynomial regression model.

And finally, thanks to decision tree classification we came to some answer to the questions we formulated at the start.

- Is there a correspondence between a predominant type of business and an exceptional high or low criminality?
  - After the analysis we cannot say with certainty that there is a definitive factor that defines high or low crime rates.
- Are more diverse neighborhoods less crime prone?
  - We cannot conclude that with the data that we gathered. It is notable to say that most low crime neighborhoods have low amounts of Outdoors & Recreation and Arts & Entertainment venues.
- What characterizes neighborhoods of notable delinquency?
  - With the current data at our disposal in this analysis is hard to differentiate between low and medium crime neighborhoods. However, contrary to what could believe, high population neighborhoods don't correlate with high crime rates neighborhoods.