08/2021

Toronto Neighborhoods' Business and their influence on Crime in 2020

Applied Data Science Capstone

IBM Data Science Professional Certificate

Alberto Carrillo Ortega
professional.acarrillo@gmail.com

Index

Figures Table

Introduction

In the next few pages, I intent to gain some insight 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.

I aim 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?

Common sense could provide some answers to some of these questions. For example; one might come to the conclusion that nightlife related neighborhoods rank higher in criminality rates. Or that more "family friendly" ones, those which have a bigger diversity and presence of places like parks and schools and less crime related.

However, is essential to contrast these mental constructs with information extracted from real-world data. It might be the case that nightlife zones are more conflictive, but it might also be the case that parks are hot spots of crime. 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 and Preparation

Data Sources

As I said before, we will be drawing our data from two sources:

TORONTO OPEN DATA

An open data initiative born in 2009 and directed by the government of Toronto City. It offers a wide catalog of data sets ranging from festivals and events to Wellbeing Neighborhoods indexes.

More precisely I will be using a dataset referring to neighborhoods crime rates. This data set contains information about assaults, auto theft, robberies and other crimes. It's recorded annually and there is information from 2014 to 2020.

Here is a link to the dataset: Neighborhoods Crime Rates

FOURSQUARE API

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.

Once again, to be more precise, we will be using its Places API. This is but a REST interface that allows us to access its services through simple HTTP requests.

I 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. However, is important to acknowledge that this service has some limitations. The most significant one is that there is a limit to the number of venues that we can obtain given a set of coordinates.

Data Cleaning

Fortunately, the data we obtain from Toronto Open Data is a perfectly clean dataset. The data set of Toronto Crimes Rate is this one:

	_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	…

Figure 1 Toronto Crime Rates Data Frame

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.

This is an example of a response we get when we make a call to Foursquare API.

```
{
   "meta": {
      . . .
   },
   "response": {
      . . .
      },
      "groups": [{
            . . .
                  },
                  "venue": {
                     "id": "4be349d763609c7439e11bff",
                     "name": "Daeco Sushi",
                     "location": {
                        . . .
                        "lat": 43.68783769992881,
                        "lng": -79.39565249242683,
                        . . .
                     },
                     "categories": [{
                           "id": "4bf58dd8d48988d1d2941735",
                           "name": "Sushi Restaurant",
                           . . .
                        }
                     ],
                     "photos": {
                        . . .
                     }
                  },
                  "referralId": "e-0-4be349d763609c7439e11bff-0"
               }, {
                  "reasons": {
                     . . .
                  },
                              . . .
```

Figure 2 JSON Response Example

In the hierarchical JSON we can observe some notable elements. We are looking for venues relating to a location. From the JSON it's easy to extract the venue name, coordinate and category. After calling for each neighborhood we can build a data frame that looks like this:

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

Figure 3 Initial Venue's Data Frame

There are over 350 Venue Categories, and each category only repeats itself in each neighborhood less than 10 times. With this data it would be hard to find any relationship or draw conclusions. Luckily, the categories system follows a hierarchical structure. As so, we can derive a category of a higher order given a specific category. There are still over 100 categories in the second degree of categories, at the highest degree there are 9. Once simplified the venues categorization, the data frame looks like this:

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

Figure 4 Simplified Venue's Data Frame

Once we have a set of venues for each neighborhood we can group then by category and neighborhood and we would get this data frame:

	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Agincourt North	1	0	25	0	1	0	0	14	0

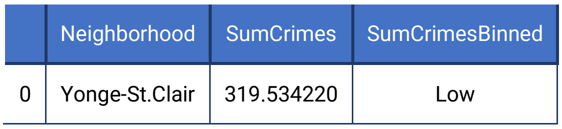
Figure 5 Grouped Venue's Data Frame

Data Preparation and Feature Selection

Once the data frames are ready, is time to merge then and pick the features we wish to use during the rest of the process.

For the Toronto Crime Rates we are only interested in the 2020 data. Therefore, we can get rid of all the data that is not related. Crime on numbers wouldn't be representative of how dangerous a neighborhood is, a larger neighborhood with more population could have more crime due to this bigger citizenry. That doesn't mean that is more conflictive than another with less residents and felonies. The dataset offers us a proportion of crime over the predicted population for that year giving us a tool to compare neighborhoods in an appropriate way.

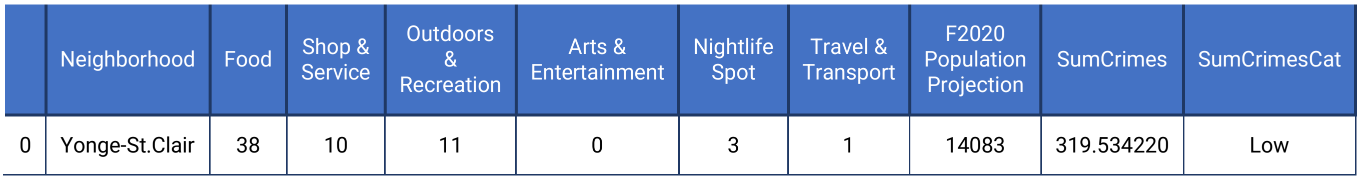
In addition, we will define a derived variable as the summatory of all weighted crime types in a neighborhood. After that we partition the neighborhoods in three categories: Low, Medium and High Crime. The division will be based on the values of the previous summatory, with Low Crime being the first quartile and below, High Crime being above the third quartile, and Medium Crime the space in between.



	Neighborhood	SumCrimes	SumCrimesBinned
0	Yonge-St.Clair	319.534220	Low

Figure 6 Toronto Crime Rates Curated Data Frame

After curating the Crime Rates data frame, we can merge both data frames on the *Neighborhood* key.



	Neighborhood	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Nightlife Spot	Travel & Transport	F2020 Population Projection	SumCrimes	SumCrimesCat
0	Yonge-St.Clair	38	10	11	0	3	1	14083	319.534220	Low

Figure 7 Merged Data Frames

Now we are ready to do an initial approach to analyze the data and gain some insights.

Preliminary Analysis

In this section we will try to do an approach to insight gathering on the data. We will plot their variables against each other in hope of finding or unraveling some relationships.

It's important to get comfortable with the data and to know everything we can about it.

Starting by plotting neighborhoods on a world map we get this image:

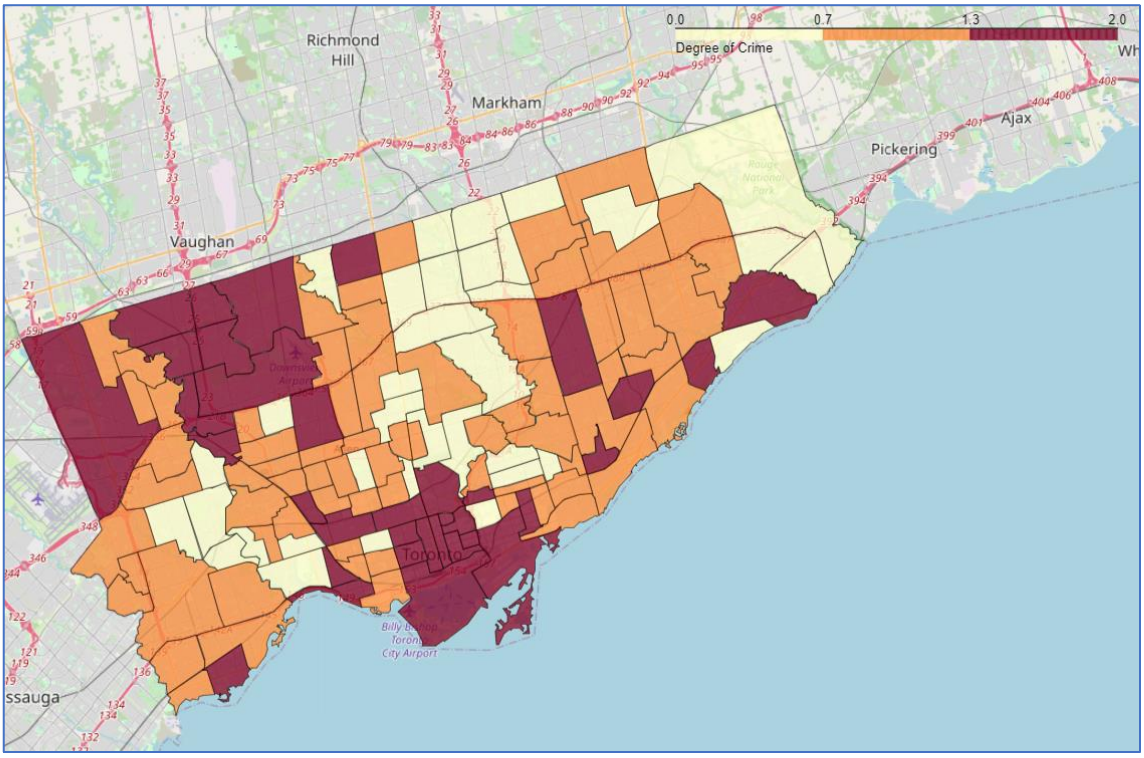


Figure 8 Neighborhoods Crime Rate Map

With crime rates being correlated with the red color, we can discern that for the most part, neighborhoods of the same category (Low, Medium, High) tend to cluster together. This is most notable on the coast and on the north-west corner of the city, where high crime rates are more common.

Now that we know the distribution of the neighborhoods, it would be interesting to see how the venues overlap with them.

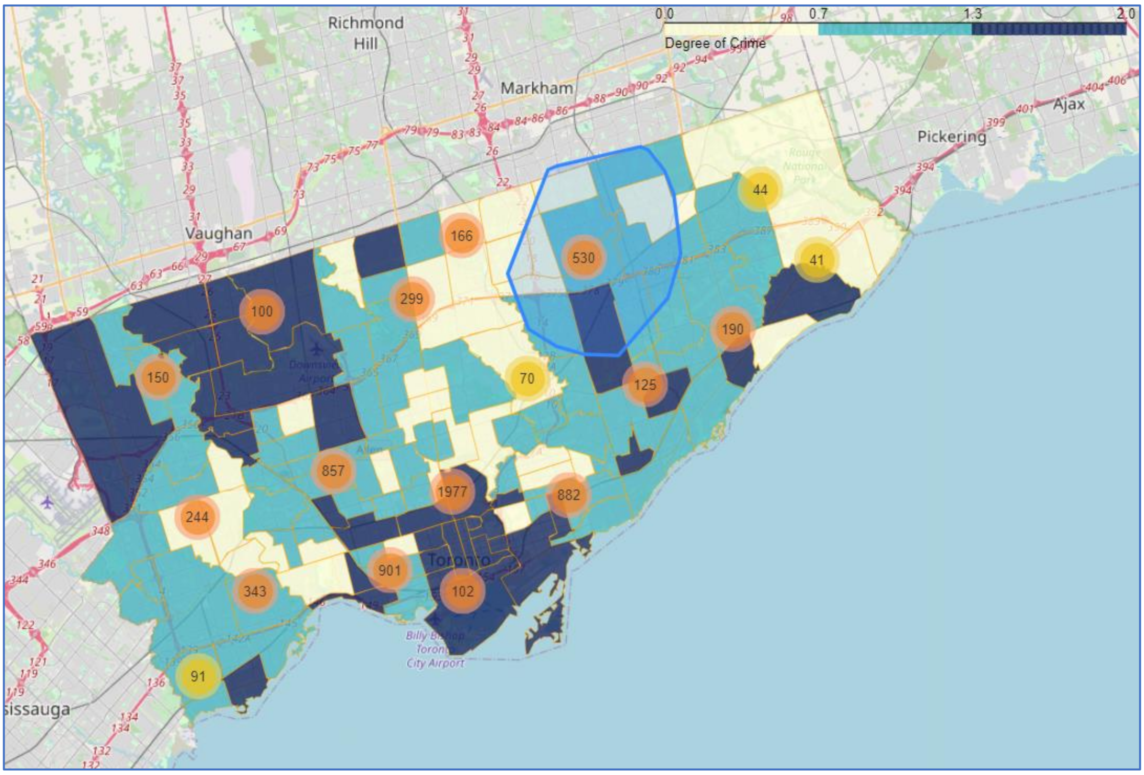


Figure 9 Number of Venues over Neighborhoods Crime Rate Map

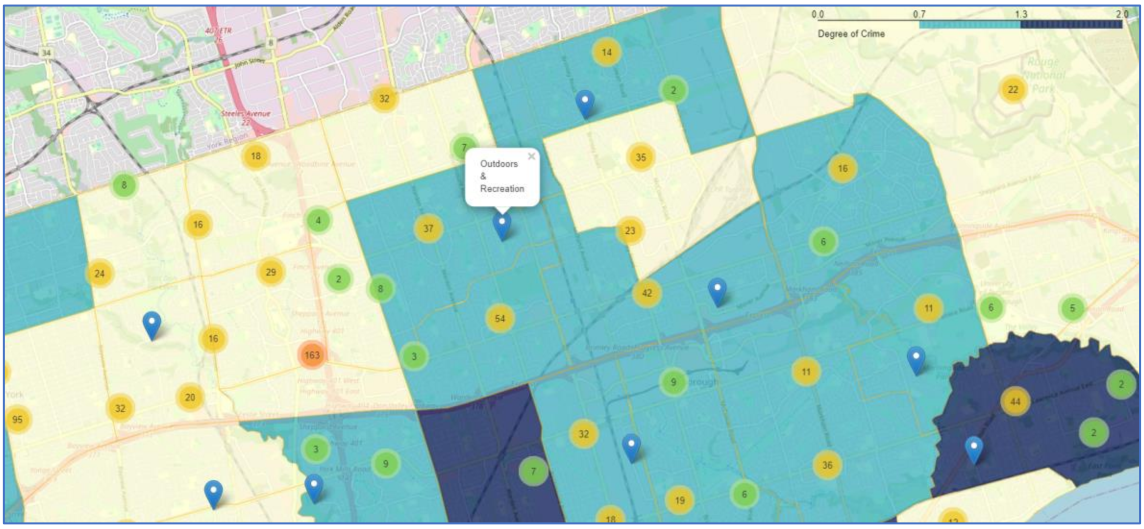


Figure 10 Close Up of Venues over Neighborhoods Crime Rate Map

In these maps, on which one might interact with in the notebook, we can see the distribution of venues by type over the city of Toronto. Thanks to these maps we can better picture the data we are working with, which can prove very valuable.

On the same page, now that we are more accustom to the structure of the city, and how the neighborhoods present some similarity in crime rates with those close by, it would be good to take a similar look at how the business distribute themselves in the neighborhoods, which ones are more common an which are less. We accomplish this through a bar plot, plotting number of venues over categories through several neighborhoods.

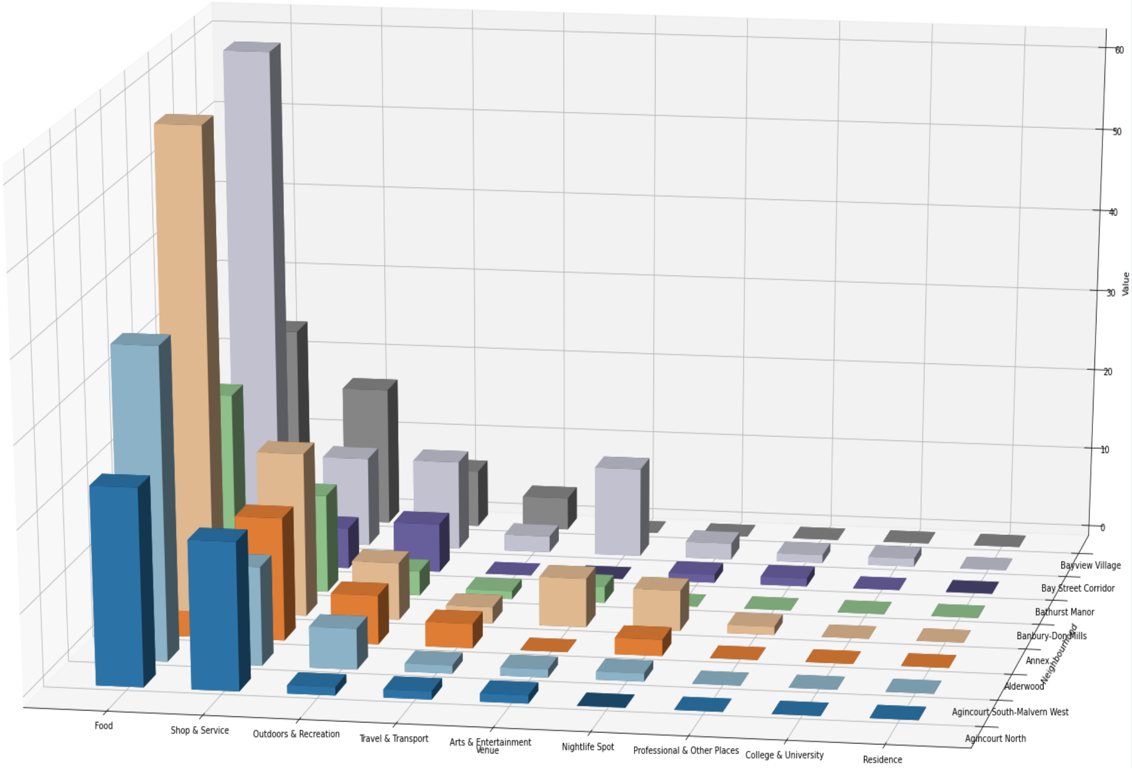


Figure 11 Number of Venues by Type over Several Neighborhoods

The graph shows that the majority of business across the neighborhoods are of the *Food* type, and that the distribution of the rest of types is mostly the same across all the other neighborhoods[i].

A possible reason for this distribution might be the origin of the data. Foursquare API is for commercial use. and the target audience is the average citizen, therefore the API might provide us with biased data directed at this kind of users.

Lastly, we will take a look at the distribution of the crime rates to see if there is something unusual.

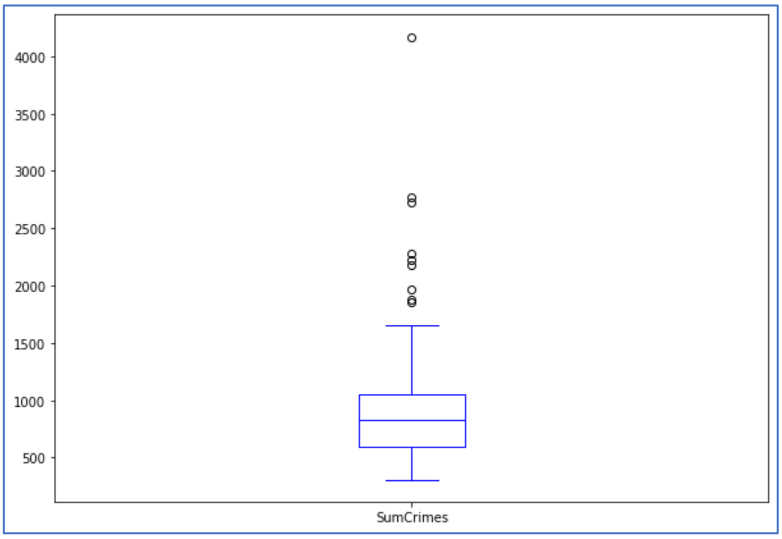


Figure 12 Distribution of Crime Rates

In the box diagram, we can observe that the there is some outliners on the upper degrees of crime rates.

Now we know which landmarks are more common, that neighborhoods close by present similar crime rates and that there is a set of neighborhoods with exceptionally high criminality. We are more familiar with the data and we can approach more complex interactions.

We will seek for some relationship between pairs of variables and the target variable. 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.

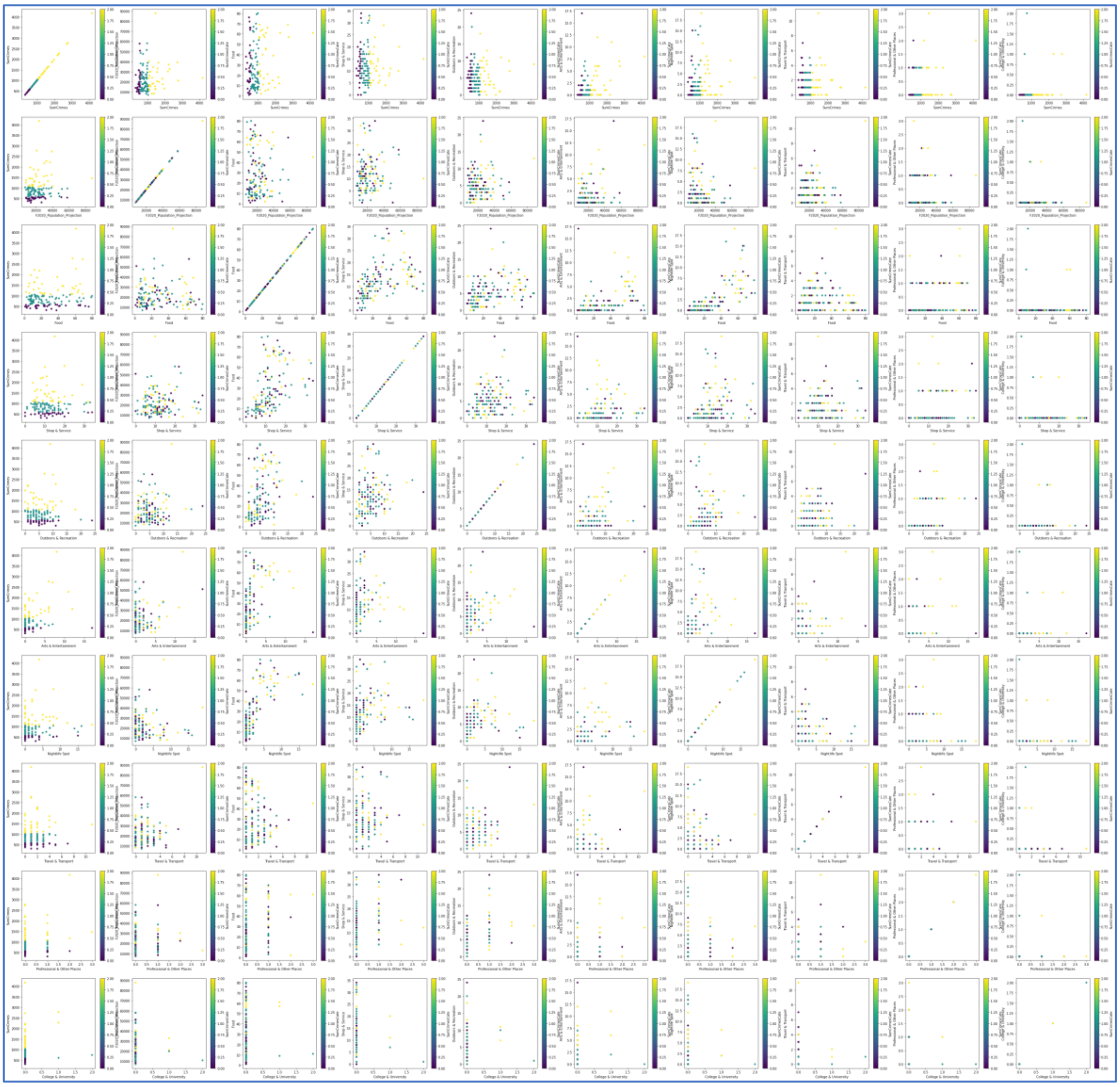


Figure 13 Pairs of Variables Colored by Crime Rates[ii]

We are looking for possible regression lines or similar occurrences. In the first row we have the target variable *SumCrimes* correlated to each other variable in the dataset. The lack of diagonal lines in these graphs means that there is not a strong linear correlation between each variable and the target feature. This can be seen in the diagonal, if it were to be a significant relation, as the diagonal rises so should the color of the points. The

However, we can appreciate some clustering between the variables. For example, 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. At lower amounts of those venues, the crime rates diminish.

After exploring relationship between pairs of attributes, it could prove interesting to examine if there is some other insight to gain with a trio of properties.

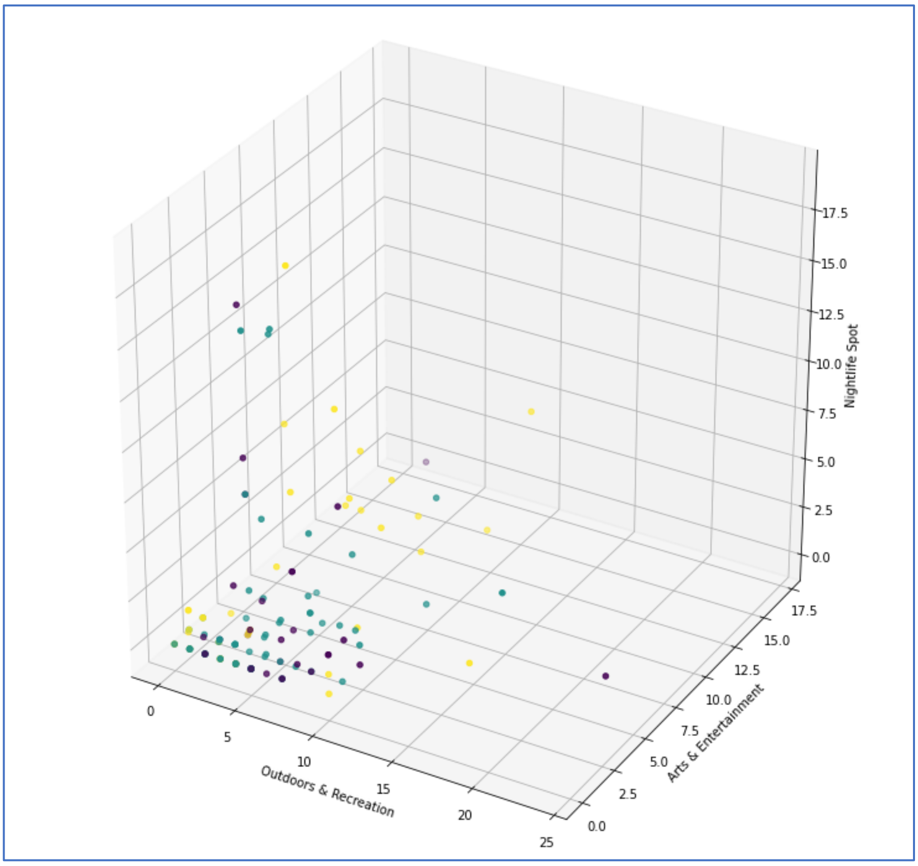


Figure 14 3D Graph of Outdoor, Arts and Nightlife Venues

Examining these three variables, we see once again clustering of low crime areas when the venues are not frequent, and as the quantity of the business rises, the higher crime rates are more common.

Finally, to validate the claim that there is not a strong linear correlation between the variables, we use a correlation matrix.

Correlation Matrix	F2020 Population Projection	Food	Shop & Service	Outdoors & Recreation	Travel & Transport	Arts & Entertainment	Nightlife Spot	Professional & Other Places	College & University	Residence
SumCrimes	0.161	0.201	0.031	-0.016	0.034	0.376	0.110	0.165	0.151	-0.085
SumCrimesCate	0.137	0.128	-0.008	-0.026	-0.018	0.212	0.144	0.057	0.091	-0.085

Figure 15 Correlation Matrix

The correlation matrix corroborates our claim; however, it is still possible that exists a relationship of no linear order.

With this, our initial analysis is concluded and we can approach predictive modeling. We have concluded the existence of clustering and the lack of linear correlation.

Predictive Modeling

This chapter is about the development of predictive models and the extraction of information from withing these models.

Preprocessing

An important part before the development and training of some machine learning model is the previous pre-processing and normalization of the data. If a feature of a dataset is order of magnitude higher that other variable of the same dataset, the algorithm might skew in favor of this bigger property.

To normalize a dataset means to transform the data so the differences between the different properties of an instance are proportional within each feature.

The data before normalization shows like this:

	Neighborhood	F2020 Population Projection	Sum Crimes	Sum Crimes Binned	Sum Crimes Cate	Food	Shop & Service	Outdoors & Recreation	Travel & Transport	Arts & Entertainment	Nightlife Spot	Professional & Other Places	College & University	Residence
0	Yonge-St.Clair	14083	310.030420	Low	0	38	11	11	2	0	3	1	0	0
1	York University Heights	30277	2179.181277	High	2	18	8	1	2	0	2	0	0	0
2	Lansing-Westgate	18146	931.758686	Medium	1	3	5	5	0	0	0	1	0	0
3	Yorkdale-Glen Park	17560	1880.544411	High	2	39	29	2	0	1	0	1	0	0
4	Stonegate/Queensway	27410	669.804890	Medium	1	9	5	4	0	0	0	0	0	0

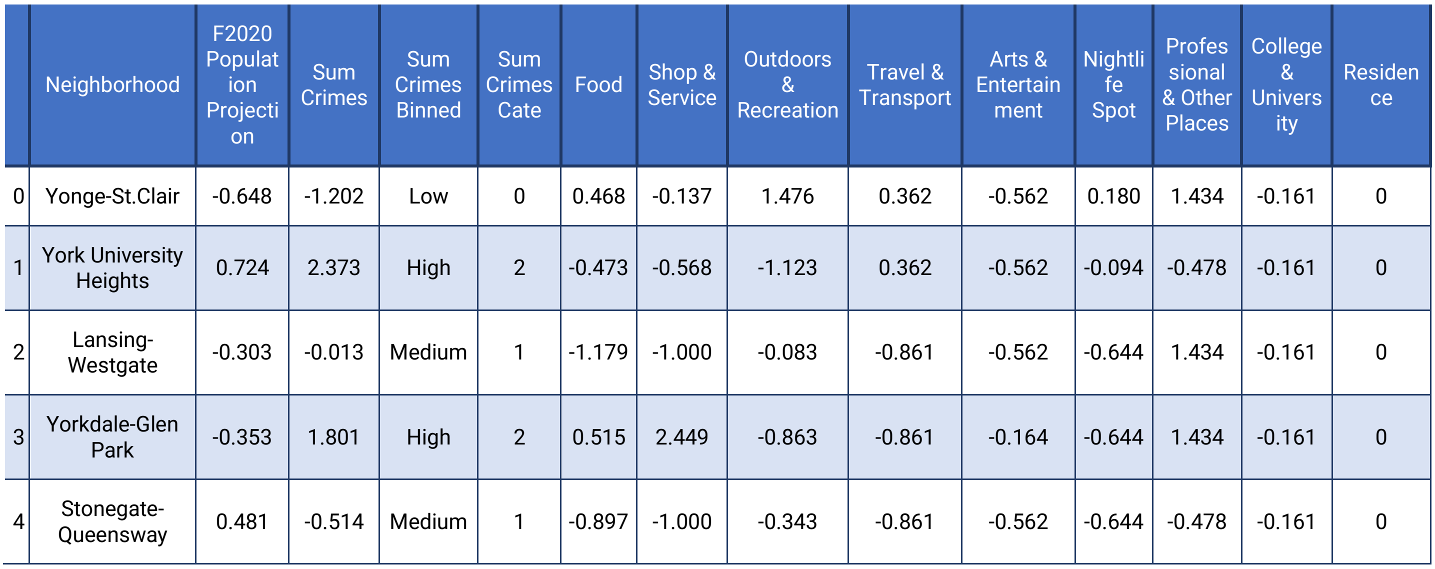
Figure 16 Dataset Before Normalization

As an example of the previously stated, the difference in value between "Yonge/St.Clair" and "York University Heights" Population Projection is orders of magnitude higher than the difference between the two neighborhoods' Food Venues number.

Here is normalization comes into play. There are many forms of normalization. In this case, a Standard Scaler which transforms the data so it has mean 0 and variance of 1 unit, will be applied.

Not all fields need to be normalized. Target and Categorical features don't need to be normalized. That's our case with *SumCrimes* and *SumCrimesBinned*.

Once normalized the dataset looks like this:



	Neighborhood	F2020 Population Projection	Sum Crimes	Sum Crimes Binned	Sum Crimes Cate	Food	Shop & Service	Outdoors & Recreation	Travel & Transport	Arts & Entertainment	Nightlife Spot	Professional & Other Places	College & University	Residence
0	Yonge-St.Clair	-0.648	-1.202	Low	0	0.468	-0.137	1.476	0.362	-0.562	0.180	1.434	-0.161	0
1	York University Heights	0.724	2.373	High	2	-0.473	-0.568	-1.123	0.362	-0.562	-0.094	-0.478	-0.161	0
2	Lansing-Westgate	-0.303	-0.013	Medium	1	-1.179	-1.000	-0.083	-0.861	-0.562	-0.644	1.434	-0.161	0
3	Yorkdale-Glen Park	-0.353	1.801	High	2	0.515	2.449	-0.863	-0.861	-0.164	-0.644	1.434	-0.161	0
4	Stonegate-Queensway	0.481	-0.514	Medium	1	-0.897	-1.000	-0.343	-0.861	-0.562	-0.644	-0.478	-0.161	0

Figure 17 Dataset After Normalization

The gaps between instances' values are still present but is been reduced significantly. After this procedure we can better develop our predictive models.

A Regression Approach

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 will transform the data and use a feature selector to try and develop an accurate regressor of the variable *SumCrimes*. To do so 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 sixty-six. These might be too many variables for the construction of a good model.

To solve this issue, we make use of recursive feature elimination (**RFE**) and a grid search to find to optimal variables.

Recursive feature elimination is used to select a subset of variables from a bigger group. This tool helps us in discerning useful from misleading variables.

Grid search allows us to optimize a hyperparameter for our model. In our case the hyperparameter is the size of the subset that recursive feature elimination returns.

Through these tools we do an extensive approach to optimization of our model. The search will have 5 folds and the method of scoring will be R^2. After the execution of the search, we gather these results:

	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

Figure 18 Regression Model Results

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.

We can see the evolution of the scoring as the number of features to be consider rises.

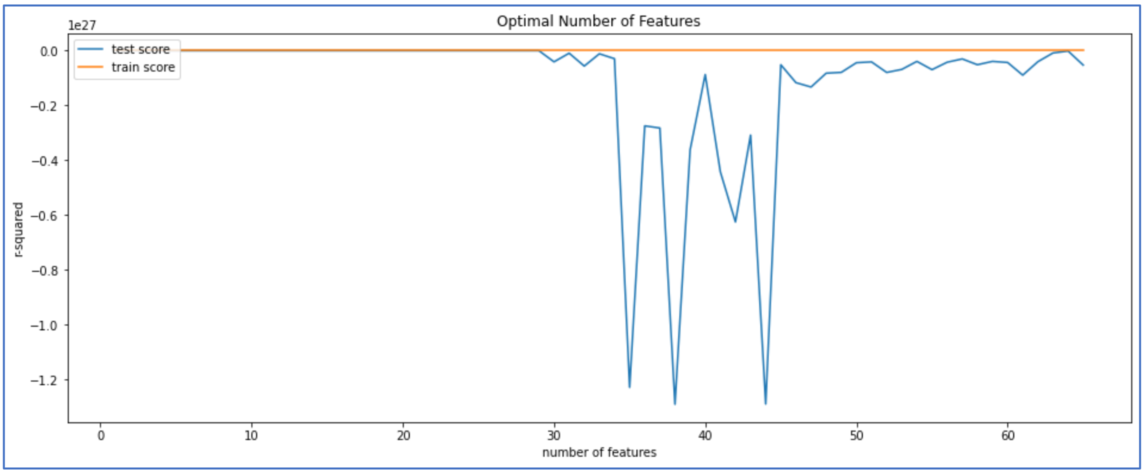


Figure 19 Evolution of Scoring for Regression Model

From the graph we can conclude that after a certain point, the rise in numbers of features correspond with a drop in test score. This is likely because of overfitting.

Even though we haven't been able to develop a good regression model we have confirmed our suspects about the correlation between predictor and target variables. Now we will approach the problem through a Classification model.

A Classification Approach

The categorical variable *SumCrimesCate* will be used to develop a Classification model. Given that this feature is laxer in its criteria perhaps we can still obtain some information.

Previously we tried to do a regression approach to the problem, but this time we want to classify the neighborhoods into low, medium or high crime rates. 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, and a classification tree because of their good adaptivity to most problems.

K-Neighbors Classification

K-Neighbors works by defining a function of distance and comparing a given instance with other instances with this function. Each of these other instances will have a value for the target variable; that is to say it will be low, medium or high income. The prediction will be the mode of these values.

The key parameter in this model is *K* or the number of neighbors to take in consideration when making a prediction. A number too low overfits out data, and a number too high overgeneralizes.

To find the optimal value for *K* we can make use of the grid search like we did in the regression model. We will search for *K* between 2 and 11.

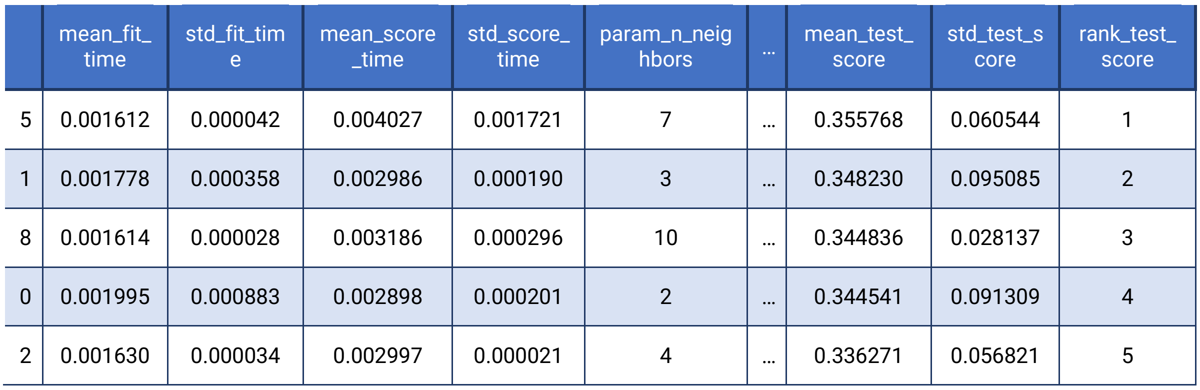
The search returns that the optimal number for K is 7, and such model has a F1 score of 0.5824 and Jaccard Index of 0.4232.

F1 score is the harmonic average of the precision and recall of the model. With precision being the rate of positives identification that were correct and recall the rate of true positives that were identified.

Jaccard index is the similarity between the prediction that the model gave us, and the real data.

These scores are not near their maximum value, but they are much higher than the values we got from our linear regressor. And it confirms that the similitude in the characteristics between neighbors is correlated, to a degree, with a similitude in their crime rate categories.

These are the results of our search:



	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

Figure 20 K-Neighbors Results

The mean test score of the folds in the search was 0.355 for our best model. A significant advance from our linear regressor.

We can see the evolution of the scoring as the *K* value rises in the following graph.

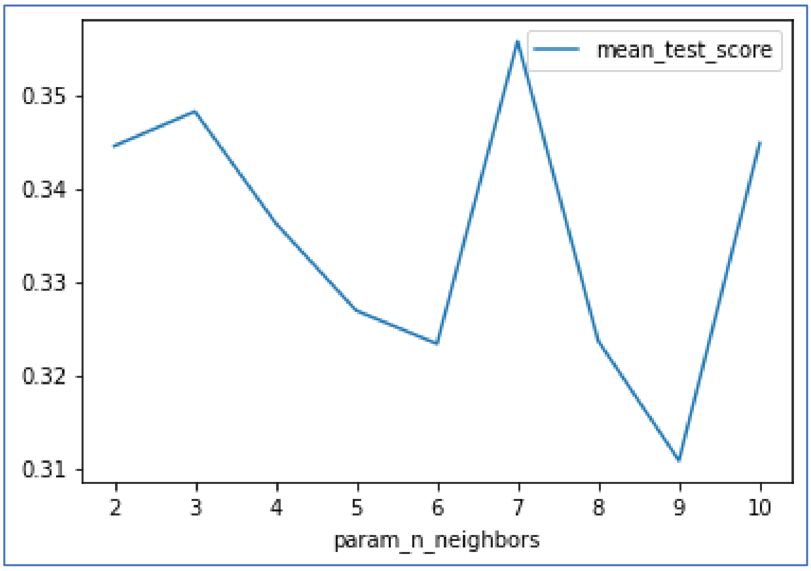


Figure 21 Evolution of K-Neighbors scoring as K value rises

One can observe how the performance peaks at seven neighbors but there is not a constant grow.

DECISION TREE CLASSIFICATION

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.

A decision tree works through nodes, in each node a condition is applied and the instances are divided on this logic statement. The goal of the tree is do devise such node structure that in each possible node the entropy of the instances in that node is minimal.

Whit entropy defined as the dissimilarity of the data. A node on which there is many instances of different criminality classes has high entropy. And a node in which all instances are of the same class, has zero entropy. That is to say, a decision tree is built aiming to classify the instances of our problem based on their more differentiating characteristics.

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 have searched for values between three and fifteen, and these are the results:

	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

Figure 22 Decision Tree Results

With a score of 0.4518, we have improved our previous classification model. Other scoring returns values of 0.8087 for F1 score and of 0.68254 for Jaccard index. These scores are much better than the regression model. Which means that there is a correlation between the neighborhoods' venues and their criminality.

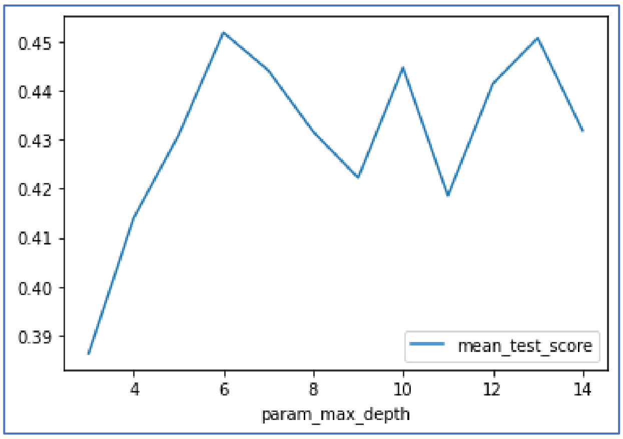


Figure 23 Evolution of Decision Tree scoring as Maximum Deep value rises

The lower the value of deep the easier to read the tree is, so its good that the best value is six and not a higher one. As said before, one of the best things about trees is that you can check their nodes and see how they works. Our decision tree is displayed on the following graph:

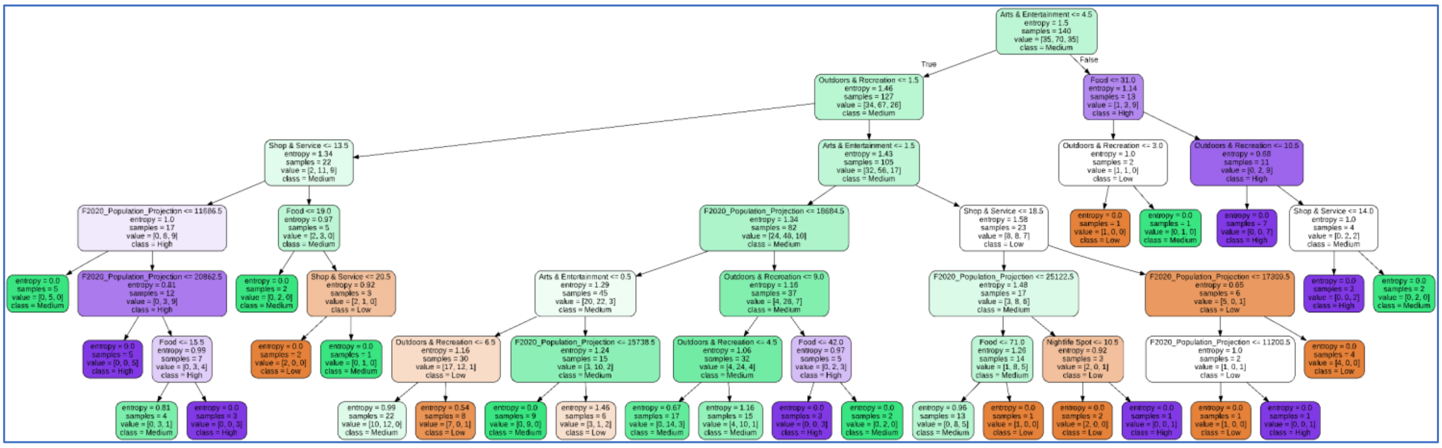

Arts & Entertainment <= 4.5 entropy = 1.5 samples = 140 value = [35, 70, 35] class = Medium
True
False
Outdoors & Recreation <= 1.5 entropy = 1.46 samples = 127 value = [34, 67, 26] class = Medium
Food <= 31.0 entropy = 1.14 samples = 13 value = [1, 3, 9] class = High
Shop & Service <= 13.5 entropy = 1.34 samples = 22 value = [2, 11, 9] class = Medium
Arts & Entertainment <= 1.5 entropy = 1.43 samples = 105 value = [32, 56, 17] class = Medium
Outdoors & Recreation <= 3.0 entropy = 1.0 samples = 2 value = [1, 1, 0] class = Low
Outdoors & Recreation <= 10.5 entropy = 0.68 samples = 11 value = [0, 2, 9] class = High
F2020_Population_Projection <= 11686.5 entropy = 1.0 samples = 17 value = [0, 8, 9] class = High
Food <= 19.0 entropy = 0.97 samples = 5 value = [2, 3, 0] class = Medium
F2020_Population_Projection <= 18684.5 entropy = 1.34 samples = 82 value = [24, 48, 10] class = Medium
Shop & Service <= 18.5 entropy = 1.58 samples = 23 value = [8, 8, 7] class = Low
entropy = 0.0 samples = 1 value = [1, 0, 0] class = Low
entropy = 0.0 samples = 1 value = [0, 1, 0] class = Medium
entropy = 0.0 samples = 7 value = [0, 0, 7] class = High
Shop & Service <= 14.0 entropy = 1.0 samples = 4 value = [0, 2, 2] class = Medium
entropy = 0.0 samples = 5 value = [0, 5, 0] class = Medium
F2020_Population_Projection <= 20862.5 entropy = 0.81 samples = 12 value = [0, 3, 9] class = High
entropy = 0.0 samples = 2 value = [0, 2, 0] class = Medium
Shop & Service <= 20.5 entropy = 0.92 samples = 3 value = [2, 1, 0] class = Low
Arts & Entertainment <= 0.5 entropy = 1.29 samples = 45 value = [20, 22, 3] class = Medium
Outdoors & Recreation <= 9.0 entropy = 1.16 samples = 37 value = [4, 28, 7] class = Medium
F2020_Population_Projection <= 25122.5 entropy = 1.48 samples = 17 value = [3, 8, 8] class = Medium
F2020_Population_Projection <= 17309.5 entropy = 0.65 samples = 6 value = [5, 0, 1] class = Low
entropy = 0.0 samples = 2 value = [0, 0, 2] class = High
entropy = 0.0 samples = 2 value = [0, 2, 0] class = Medium
entropy = 0.0 samples = 5 value = [0, 0, 5] class = High
Food <= 15.5 entropy = 0.99 samples = 7 value = [0, 3, 4] class = High
entropy = 0.0 samples = 2 value = [2, 0, 0] class = Low
entropy = 0.0 samples = 1 value = [0, 1, 0] class = Medium
Outdoors & Recreation <= 6.5 entropy = 1.16 samples = 30 value = [17, 12, 1] class = Low
F2020_Population_Projection <= 15738.5 entropy = 1.24 samples = 15 value = [3, 10, 2] class = Medium
Outdoors & Recreation <= 4.5 entropy = 1.06 samples = 32 value = [4, 24, 4] class = Medium
Food <= 42.0 entropy = 0.97 samples = 5 value = [0, 2, 3] class = High
Food <= 71.0 entropy = 1.26 samples = 14 value = [1, 8, 5] class = Medium
Nightlife Spot <= 10.5 entropy = 0.92 samples = 3 value = [2, 0, 1] class = Low
F2020_Population_Projection <= 11200.5 entropy = 1.0 samples = 2 value = [1, 0, 1] class = Low
entropy = 0.0 samples = 4 value = [4, 0, 0] class = Low
entropy = 0.81 samples = 4 value = [0, 3, 1] class = Medium
entropy = 0.0 samples = 3 value = [0, 0, 3] class = High
entropy = 0.99 samples = 22 value = [10, 12, 0] class = Medium
entropy = 0.54 samples = 8 value = [7, 0, 1] class = Low
entropy = 0.0 samples = 9 value = [0, 9, 0] class = Medium
entropy = 1.46 samples = 6 value = [3, 1, 2] class = Low
entropy = 0.67 samples = 17 value = [0, 14, 3] class = Medium
entropy = 1.16 samples = 15 value = [4, 10, 1] class = Medium
entropy = 0.0 samples = 3 value = [0, 0, 3] class = High
entropy = 0.0 samples = 2 value = [0, 2, 0] class = Medium
entropy = 0.96 samples = 13 value = [0, 8, 5] class = Medium
entropy = 0.0 samples = 1 value = [1, 0, 0] class = Low
entropy = 0.0 samples = 2 value = [2, 0, 0] class = Low
entropy = 0.0 samples = 1 value = [0, 0, 1] class = High
entropy = 0.0 samples = 1 value = [1, 0, 0] class = Low
entropy = 0.0 samples = 1 value = [0, 0, 1] class = High

Figure 24 Decision Tree Structure

For the construction of the tree, we have used unnormalized data because trees work well without the need for the conversion, and are easier to read.

With the help of the tree and the distribution of the data we can start to inference some insights. This is the general distribution of the data:

	F2020 Population Projection	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Nightlife Spot	Travel & Transport	Professional & Other Places	College & University	Residence
count	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
mean	21728.871429	28.057143	11.957143	5.321429	1.414286	2.342857	1.407143	0.250000	0.035714	0.014286
std	11839.460093	21.317325	6.981859	3.859526	2.524657	3.649719	1.639865	0.524576	0.221529	0.119092
min	7130.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	13227.250000	10.750000	7.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	18378.000000	21.500000	11.000000	5.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000
75%	26598.250000	41.250000	15.000000	7.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000
max	87808.000000	80.000000	34.000000	24.000000	17.000000	19.000000	11.000000	3.000000	2.000000	1.000000

Figure 25 Dataset's Variables Distribution

For example: the first node differentiates between neighborhoods with four or less "Arts & Entertainment" venues. But are those many or few? We need to look at the distribution to understand it. With a simple look we know that 75% of neighborhoods have two or less locales of that type. The tree also tells us that, of those which do have five or more, 9/13 have high crime rates. This is valuable information as it makes it apparent that high density of these types of venues leads to higher crime.

From the tree we can conclude that most of the low crime neighborhoods have median or less population projection, under 1º quartile Outdoors & Recreation and below 3º quartile Arts & Entertainment venues.

We can also say low population usually influences the neighborhood so it doesn't have high crime rates. (This doesn't mean the crime rates are necessarily low).

Also is notable that sometimes is difficult to differentiate between high and medium crime neighborhoods.

Conclusions

In the begging we aimed ourselves to find some correlation between crime rates and venues, and get answers to some questions.

After some work, we came to the conclusion that there is some degree of clustering between the neighborhoods. We did this through our preliminary analysis and during the classification approach with the k-neighbors classifier. However, even though there is a measure of similitude between same crime rates neighborhoods, this correlation is not very strong and therefore should be used with caution when trying to make a prediction.

We also found out that our data is unfit at all for a linear/polynomial regression model and that we should stay away from that.

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. Is more likely that there are a set of circumstances that, together, influences the neighborhood in a way or another.

- Are more diverse neighborhoods less crime prone?

We cannot conclude that with the data that we gathered. It is notable to say that usually, 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.

Finally, to conclude this notebook I would like to note that given the requisite of using *FOURSQUARE Places API* to acquire is not without consequence. While the API is great for commercial use, the restrictions that it applies to us are very important, namely that we can only get 50 venues/locales per neighborhood and the randomness of the results it returns, don't allow for a proper analysis of the zone.

Bibliography

[1] Foursquare, n.d. *Foursquare's Places API Documentation.* [Online]
Available at: https://developer.foursquare.com/docs/places-api/

[2] Kanstrén, T., 2200. *A Look at Precision, Recall, and F1-Score.* [Online]
Available at: https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

[3] Pedregosa, 2011. *Scikit-learn: Machine Learning in Python.* [Online]
Available at: https://scikit-learn.org/stable/

[4] Rosebrock, A., n.d. *Jaccard Index.* [Online]
Available at: https://deepai.org/machine-learning-glossary-and-terms/jaccard-index

[5] Sayad, P. S., n.d. *Decision Tree - Classification.* [Online]
Available at: https://www.saedsayad.com/decision_tree.htm

[6] Services, T. P., 2021. *Toronto Neighbourhood Crime Rates.* [Online]
Available at: https://open.toronto.ca/dataset/neighbourhood-crime-rates/

[7] The pandas development team, n.d. *Pandas - Python Data Analysis Library.* [Online]
Available at: https://pandas.pydata.org/

[8] Wikipedia, the free encyclopedia, n.d. *F-score.* [Online]
Available at: https://en.wikipedia.org/wiki/F-score

[9] Wikipedia, the free encyclopedia, n.d. *Jaccard index.* [Online]
Available at: https://en.wikipedia.org/wiki/Jaccard_index

Appendices

[i] Even though the graph only shows a few neighborhoods, we can see that the distribution is the similar across all of them in the data frame.

[ii] Even though the image is small, we can zoom on it to see each graph in more detail.