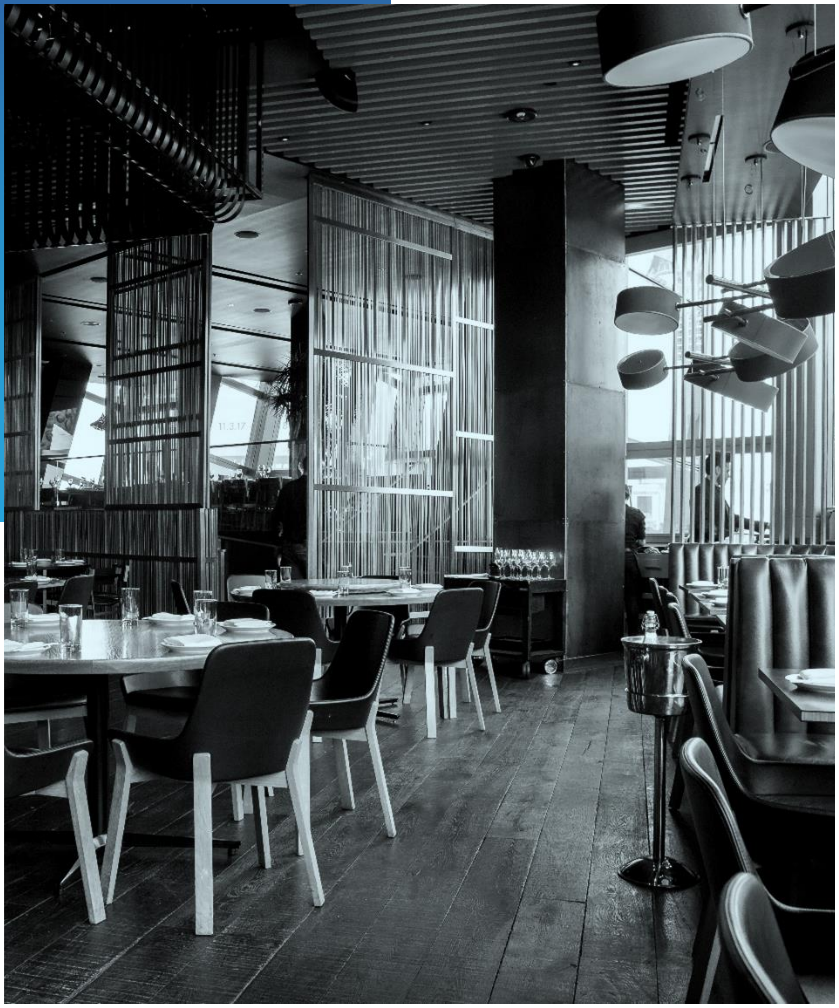


# THE BEST NEIGHBOURHOOD TO OPEN

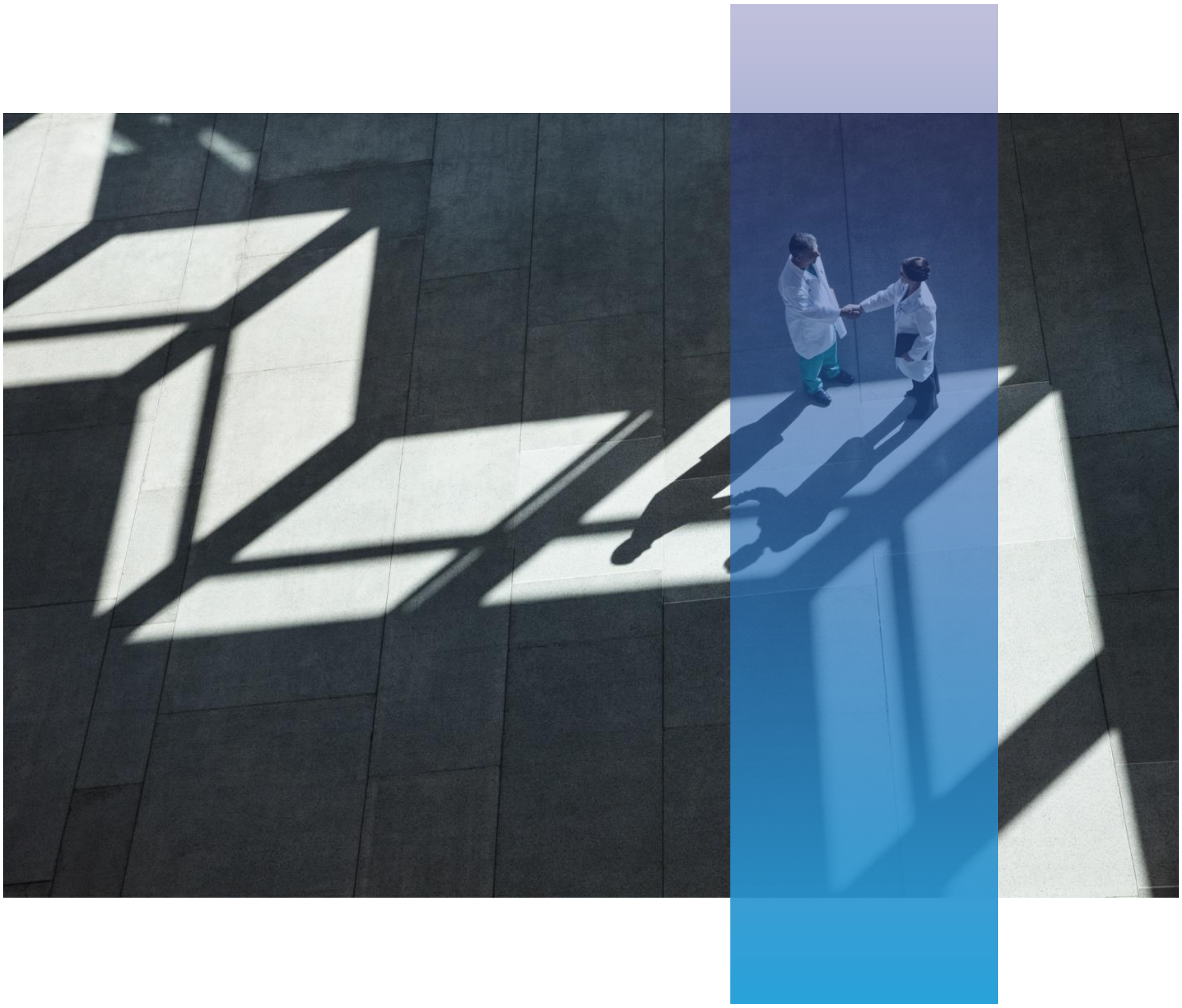
# MY RESTAURANT

**Capstone Project**

August 2020

Nachiketa N. Kurhade

# Table of Contents



## Where Should I Open My Restaurant?

Many entrepreneurs face this problem while opening their first restaurant or while opening a new branch of their restaurant. Hands down, setting up a new restaurant is costly indeed, and if it fails to return profits, the investment becomes a disaster. On the other hand, a restaurant in profit is a business success. So, if you want to put your money in the right place, you better do your homework, you better have the formula for success!!

In this report we aim to determine the best neighborhood in the city of Toronto to open up a new restaurant which will require minimum initial cost and will result in maximum profit.

To solve this problem, we need to answer two questions:

1. Which neighborhoods are performing better in terms of restaurant business?
2. What are the factors that are making these neighborhoods perform better?

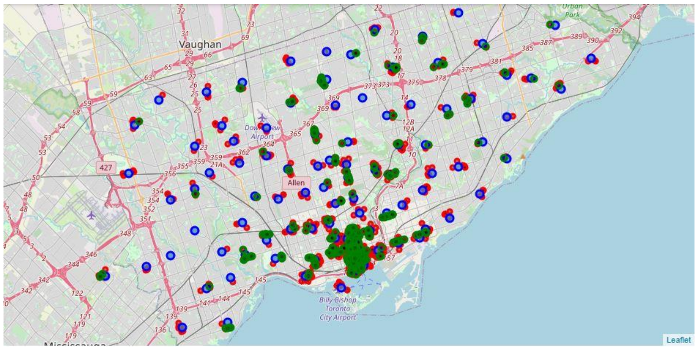
# Data

We will be breaking down this analysis in three steps

1. Collect data about neighborhoods in Toronto.
2. Define performance parameters and rank the neighborhoods by their performance.
3. Find the neighborhood characteristics that support your performance results.

For this analysis, the data was collected using Foursquare API, Open Data Portal – Toronto City, IBM Cognitive Class Labs, and Wikipedia.

- Primary neighborhood data (contains basic information about neighborhoods) was collected using the Wikipedia page that contains information about 'boroughs', 'neighborhoods', and their 'postal codes' for the city of Toronto.
  Using data provided by IBM Cognitive Class Labs, we assigned 'co-ordinates' to these neighborhoods, thus preparing our primary neighborhood database.
- Secondary neighborhood data (contains primary neighborhood database with venue data from each neighborhood) was collected using Foursquare API.

  Regular Foursquare API calls were made to collect information such as top 100 venues, their coordinates and venue category.
  Radius was set to 500 meters.
  Venue return Limit was set to 100
  A quick folium plot (*blue markers for neighborhood locations, red markers for the venues returned and green markers for the restaurants)* showed that some places had neighborhood centers are really close. Hence, we had to limit radius to 500 meters as increasing radius might have resulted in duplicate venues.
  Secondary data had too many 'Venue Categories', and many of these values were actually the same value written in a different way, so it was cleaned up to remove such duplicates.
- Tertiary database was created by separating restaurants from our venues data frame (Secondary data). As we don't have an actual turnover data of each of these restaurants, we needed to set up some other criteria to analyze how well they're performing.
  Using the data returned by Premium Foursquare API calls which returned venue 'tips', we considered using two parameters, 'Rating' and 'Check Ins', to determine their performance. As both of these parameters are vital and directly proportional to restaurant's performance, we will set up another parameter 'Score' which is the product of both.
  The 'performance' of a neighborhood was defined as the score of the respective neighborhood per restaurant. Therefore, the 'performance' parameter won't get biased like 'Score' as a neighborhood 'Score' can get more just because there are a greater number of restaurants in the first place. The definition of 'performance' corrects this problem and thus even if there are a smaller number of restaurants in a particular neighborhood but if they 'Score' exceptionally, the corresponding 'performance' will be more and the respective neighborhood will rank higher.

This is how we answered our first question i.e. **Which neighborhoods are performing better in terms of restaurant business?**
Ranking neighborhoods in ascending order of 'performance' should give us an idea about the neighborhoods whose restaurants are receiving most recommendations and are being frequently visited by people.

- Going through some reports on factors to consider while setting up a new restaurant, we realized there are more factors which may contribute to our 'performance'. This kind of data can-not be collected using Foursquare API. Therefore, Extra Database was collected from 'Open Data Portal – City of Toronto'. This extra data included information about

  1. *Festivals and events*
  2. *Neighborhood profiles*
  3. *Parking lot facilities*
  4. *Places of interest and Attractions*
  5. *Public art work assets*
  6. *Culture*
  7. *Economics*
  8. *Recreation*
  9. *Safety*
  10. *Transportation*

  Out of these I am excluding 'Festivals and events', 'Recreation' and 'Public art work assets' as given a close look they seem much less relevant compared to others.
  Also, we took liberty to collect the data from 'Neighborhood profiles', 'Culture', 'Economics', 'Safety', and 'Transportation' into a single dataset 'overall' for easy handling.
  This data was used to check the common claims made by above mentioned reports by using machine learning algorithms like regression.
  Also, in Canada, the Neighborhoods as defined by postal data are different than the ones defined by the city administration, later being based on population census.
  This induced one complication, i.e., neighborhoods in this new data don't match to the ones we have been using till now. To solve this, we identified the new neighborhoods by their postal codes, since these boundaries overlap, the most frequent postal code had to be taken. This column served as key to connect this extra database with other databases we had been developing so far.
- Logically speaking, if we categorize venues from Secondary Database, we should be able to find some 'Venue Categories' here which have significant impact on our 'performance'. But as our input set is categorical, we needed to convert it using one hot encoding first, then we performed the usual multiple linear regression analysis.
  By performing the analysis mentioned in last two points we were able to answer our second question i.e. **What are the factors that are making these neighborhoods perform better?**

# Exploratory Data Analysis and Results

So far, we have four databases at hand

1. Primary Database – Contains basic information about the neighborhoods of the city of Toronto, and their geographic locations.
2. Secondary Database – Contains basic information about top 100 venues under 500m of the specified neighborhoods.
3. Tertiary Database – Contains details about the “restaurant” category venues from the Secondary Database.
4. Extra Database – Contains additional data about city neighborhoods.

To answer our first question, i.e. **Which neighborhoods are performing better in terms of restaurant business?** We needed to rank neighborhoods as discussed in the data section. For that we set up a ranking criterion ‘performance’. Few things to recall:

For a Restaurant:

**‘Score’ = ‘Rating’ * ‘Check Ins’**

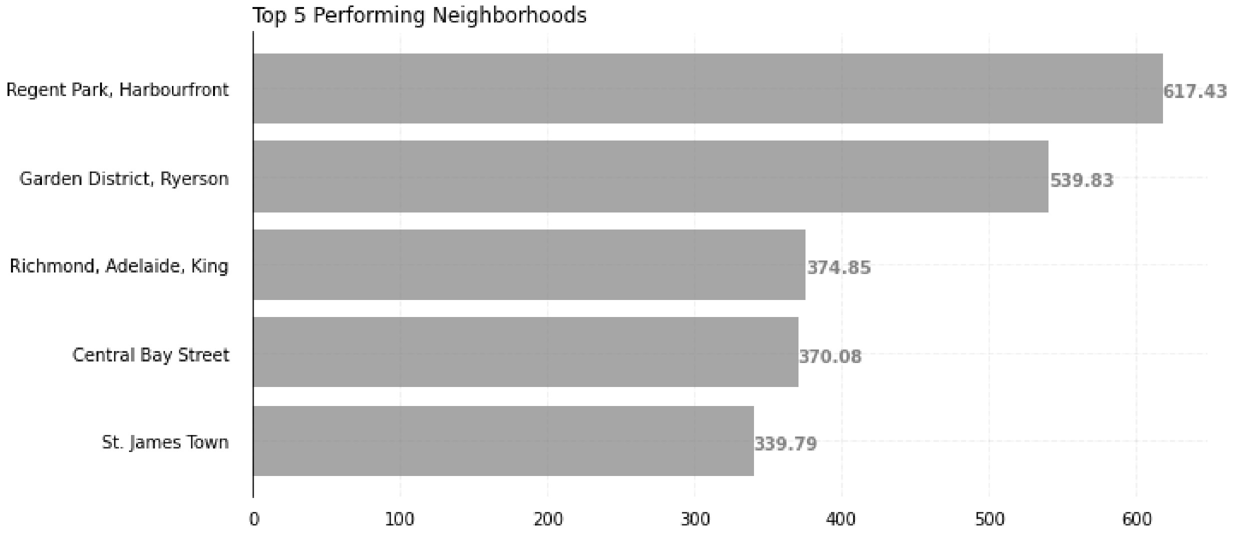
Now of course for some restaurants, absence or very less value of either data lowered the score drastically, which was justified as it meant that the restaurant concerned didn’t have much online presence, which will affect its revenue as people usually check these two things online to see if it’s good enough.

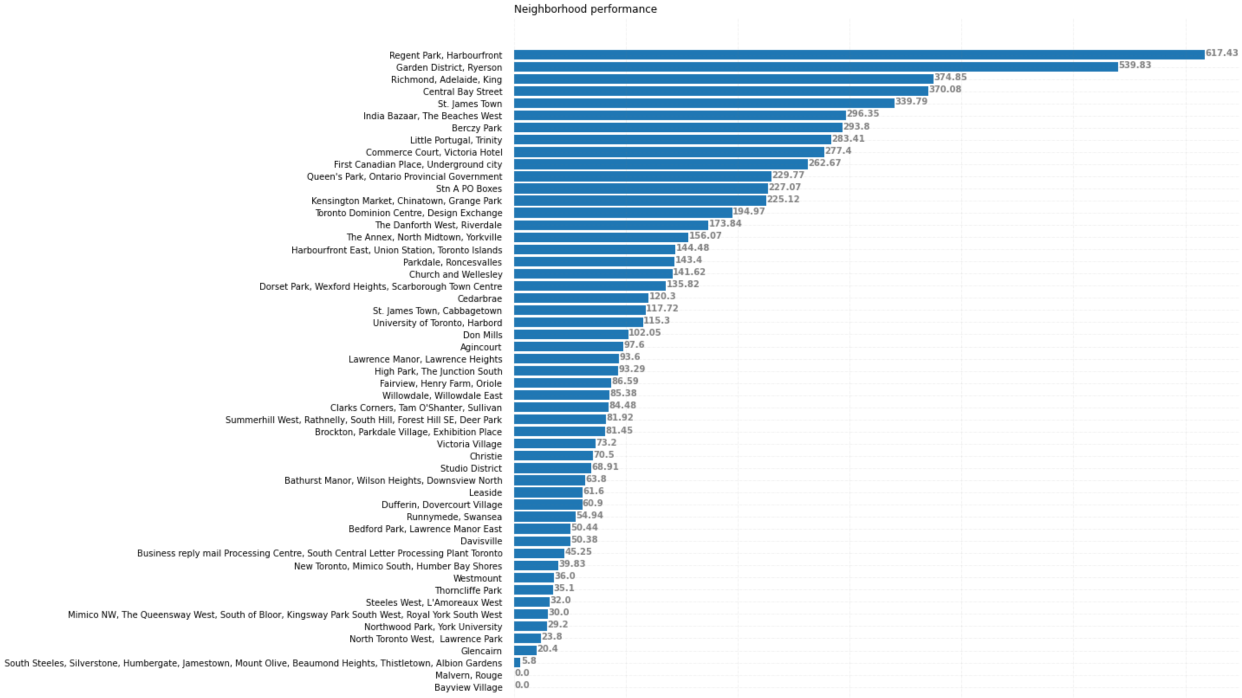
For a Neighborhood:

**‘performance’ = Total Score / Total No. of Restaurants**

The ‘performance’ of a neighborhood hence didn’t only depend on the number of restaurants it had; it also took the ‘Score’ of the restaurants in concerned neighborhood into account. Thus, it became possible for the neighborhoods having less restaurants to outrank others based on their score.

Ranking neighborhoods based on their performance is summarized in following bar charts:

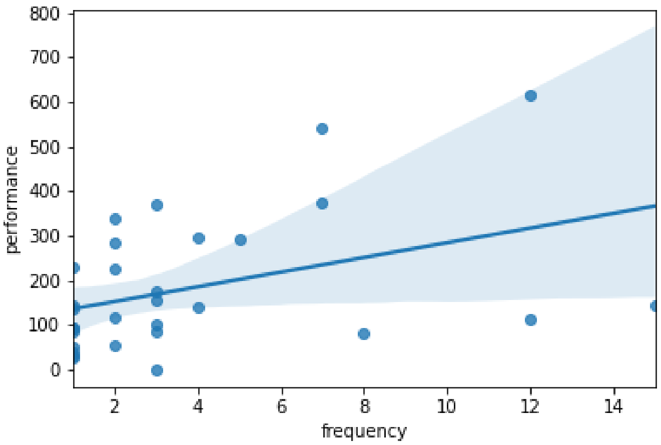




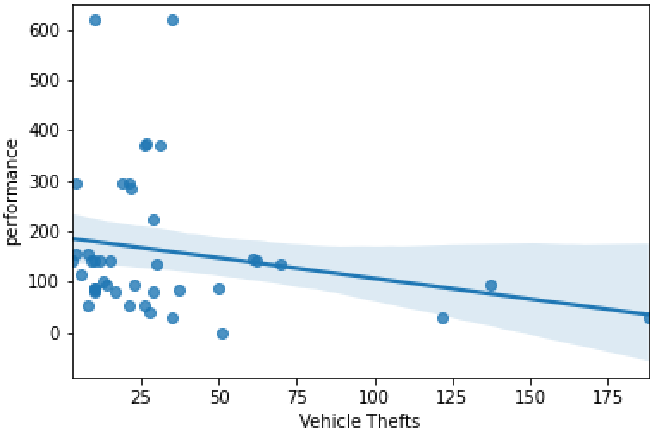
This way we answered our first question, we now had the list of neighborhoods which are performing better than others according to the criterion we set up. Now we need to know why they are performing better, i.e. our second question, **what are the factors that are making these neighborhoods perform better?**

To answer this, we started exploring our Extra Database and it was needed to find out which of its features are having correlation with our 'performance' and if yes, how strong it is?

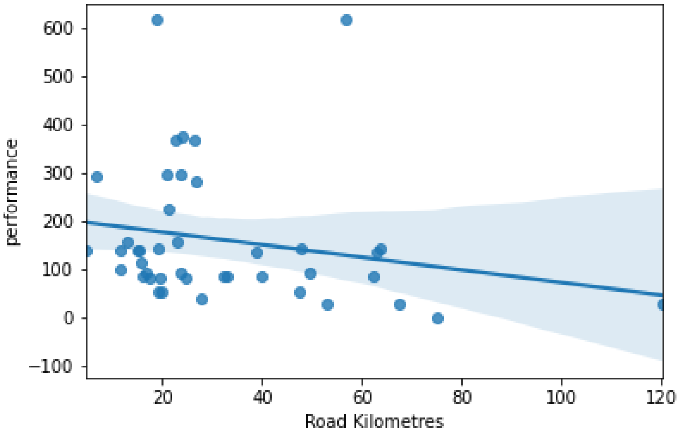
1. **Tourist Destinations:** One can assume that more the number of tourist sites in the area means more customers, and hence greater performance, our scatter plot here showed that yes, having more number of tourist sites does indeed improve your neighborhood 'performance' but presence of few lower performing neighborhoods despite having high number of tourist sites suggested this shouldn't always be taken for granted. With a closer look at data, we could tell that these under-performing high-frequency neighborhoods had a greater number of restaurants, hence implied more competition. *Visitors visiting a certain tourist place also varies greatly, hence with a more detailed dataset we could refine our results even more.* But with the data available, we can only say that the No. of Tourist Destinations in the neighborhood has a weak positive correlation with the 'performance' of the neighborhood.



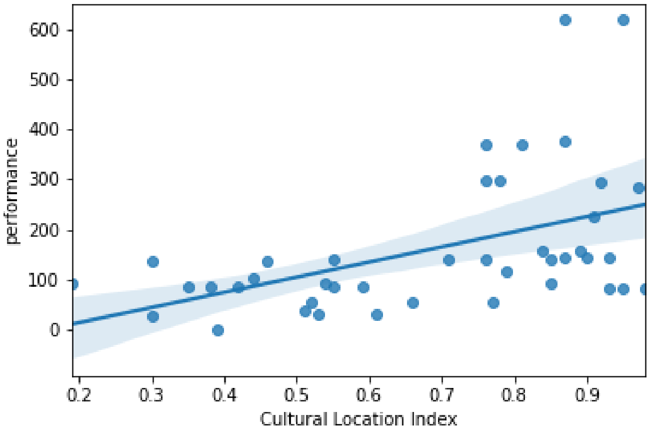
2. **Vehicle Thefts:** Having a safer neighborhood for your parked vehicle is really important for any restaurant customer. Thus, we expected a decline in performance with increase in vehicle thefts, and we weren't disappointed. Higher 'performance' neighborhoods were seen to lie in the areas having lesser vehicle theft cases reported compared to the rest. Although it was a weak negative correlation, we considered it for our analysis.



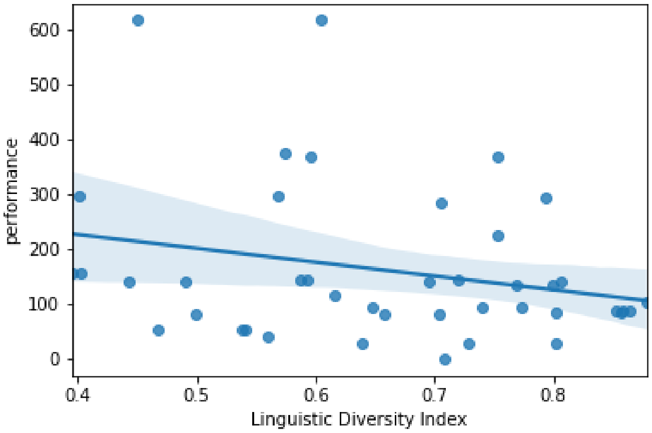
3. **Road Kilometers:** This was an unexpected result for us, as a neighborhood having more road kilometers should have indicated more walkability and hence more places to open businesses, possibly even a market. Thus, we expected a positive correlation, instead we got a weak negative correlation. Seems like people want it simple.



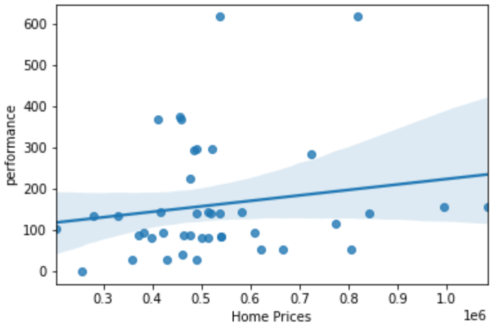
4. **Cultural Location Index:** As a cultural location index usually includes share of artists and cultural workers working in a neighborhood and the number of cultural facilities in the same, we expected more performance for neighborhoods with more cultural index, and the trend appears to be exponentially increasing, only problem, it scatters more and more as the cultural index approaches one, making it a weak positive correlation. A regression line with the scatter plot is shown in the adjoining figure.

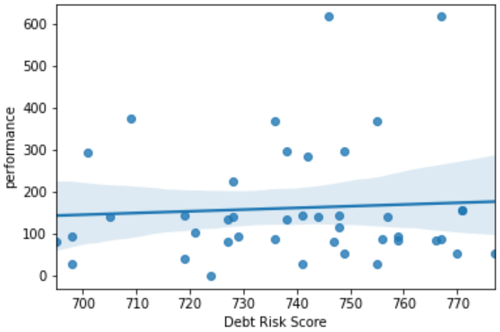


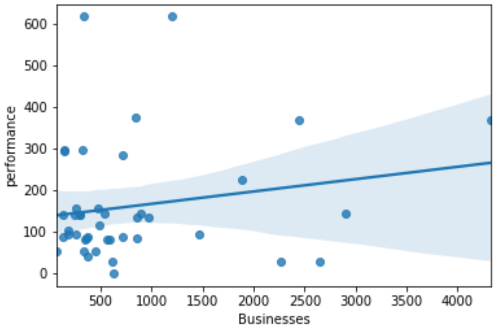
5. **Linguistic Diversity Index:** Opposed to the behavior of the Cultural Location Index, with increasing Linguistic Diversity index the Performance seemed to exponentially decrease, thus showing lower performance for neighborhoods having higher linguistic diversity. This unusual trend also seemed to be converging with increasing value of the index.

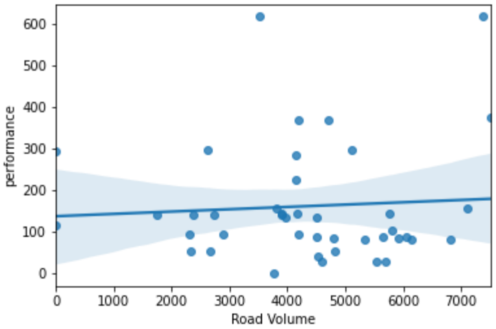


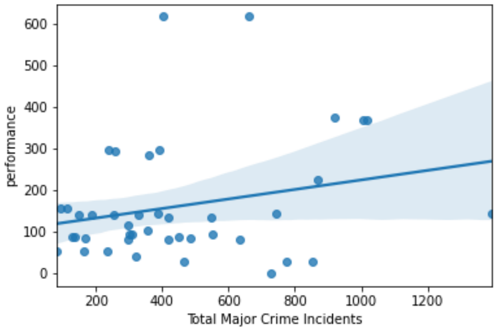
6. **Others:** The remaining features didn't show any significant correlation with our 'performance' hence were omitted from our analysis. Either there was no correlation or the points were to scattered to get even considered as a weak correlation. Following scatter plots will give you an idea about their relationship with 'performance'.

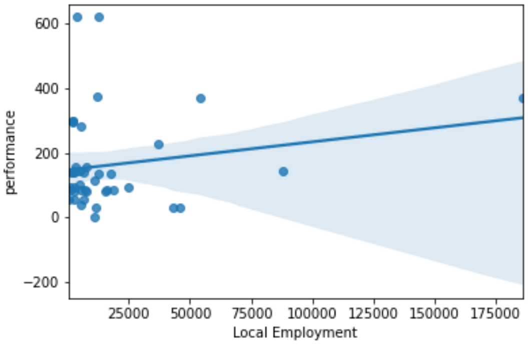












The absence of correlation in features like Housing Prices and Local Debt Score with performance may be due to the fact that both of these quantities don't generally affect the running 'Score' of the restaurants, but they can affect initial set up cost for the new restaurant. Since our problem statement required us to analyze the 'performance' of the neighborhoods, we restricted our analysis to the first five features.

We used the following machine learning tools:

1. Multiple linear regression to fit the data for 'No. of Tourist Destinations', 'No. of Vehicle Thefts' and the 'Road Kilometers' in a neighborhood against 'performance'.
2. Non linear regression to curve fit an exponentially increasing curve with the data from 'Cultural Location Index' against 'performance'.
3. Simple linear regression to fit the data from 'Linguistic Diversity Index' against 'performance'.

The following results were obtained:

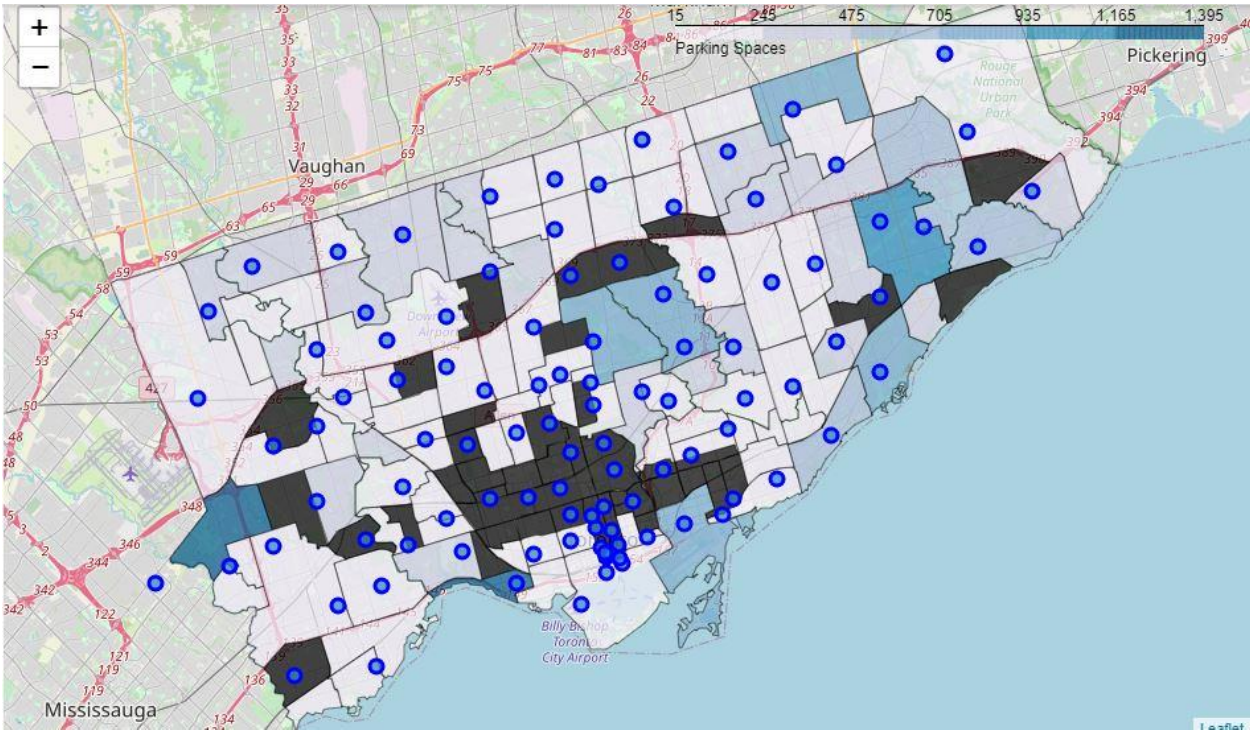
| Features | Regression Results | | |
|---|---|---|---|
| • No. of Tourist Destinations<br>• No. of Vehicle Thefts<br>• Road Kilometers | 0.30<br>R2 Score | (0.74)<br>RSS | |
| • Cultural Location Index | 0.11<br>R2 Score | (0.20)<br>MSE | (0.09)<br>MAE |
| • Linguistic Diversity Index | 0.35<br>R2 Score | (0.61)<br>RSS | |

'Parking lot facilities' was one of the features of our Extra Data, while analyzing it, we noticed that it lacked information about many neighborhoods, 40 to be precise. Hence, exploring relationships between parking lot facilities and 'performance' was likely to result in an erroneous outcome. Also, parking lot facilities is more of factor that affects the initial set up cost of a new restaurant than its running 'Score' as every established restaurant is likely to have its own parking arrangement of some sorts.

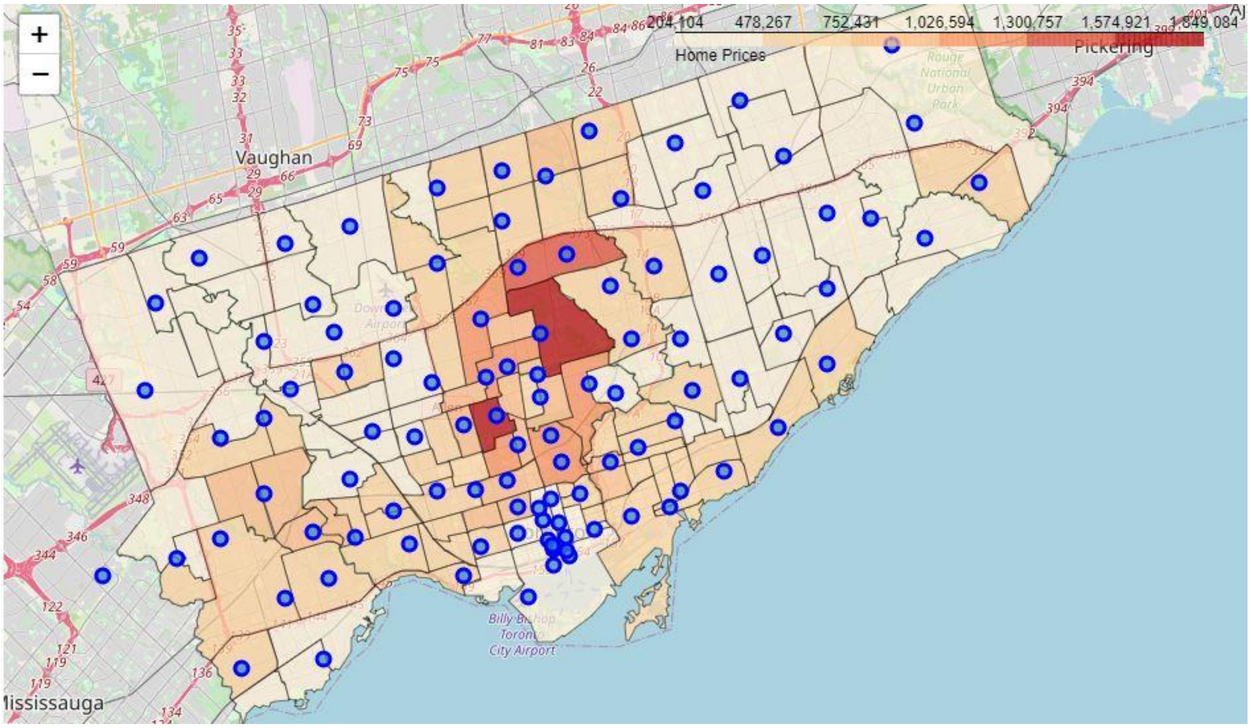
Choropleth plots of 'Parking lot facilities' and 'House Prices' however, presented us with an insight regarding the neighborhoods which had the potential to reduce the initial set-up costs for a new restaurant. Close public parking and lower housing prices usually mean that one will have to invest less to arrange them for the new restaurant.

The shapes for plotting these maps were taken from the 'Open Data Portal – City of Toronto' which divides the city of Toronto into neighborhoods in a different way (population based). Therefore, these borders do not correspond to the neighborhoods we have been studying so far, we have indicated our 'in use' neighborhoods with blue markers for easier interpretation.

Following Choropleth Plots were obtained:



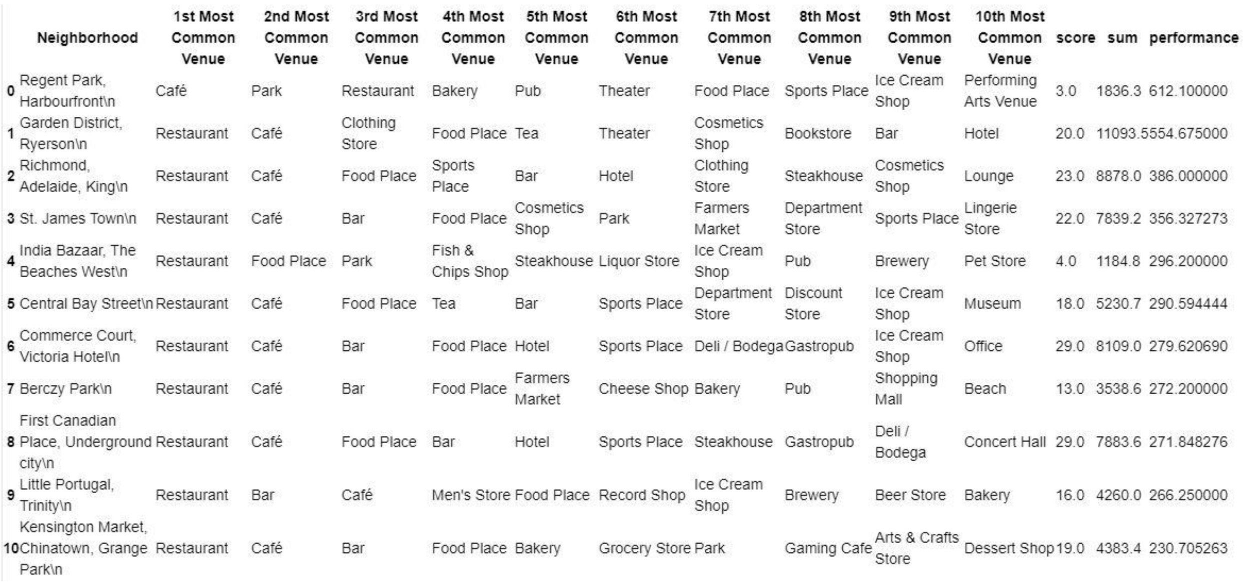
*Parking Spaces In Toronto*
*(Black shaded neighborhoods indicate absense of data)*



*Housing Prices in Toronto*

Now we moved on to our next dataset, the biggest one yet, our Secondary Database. It contained the top 100 venues returned within 500m redius of the neighborhood centers provided by our Primary Dataset. This dataset included one important feature i.e. 'Venues Category'. Theoretically speaking, using multiple linear regression as our analysis tool, we should be able to find the categories that have the most significant impact on our performance.

Taking a look at top 10 most common venue categories of each of these neighborhoods, we realized that the venue categories having a stronger positive correlation with our 'performance' would be the ones that are most frequent, and appear selectively in highest performing rows. Appearing selectively in lowest performing rows would result in a stronger negative correlation with our 'performance'. If some 'Venue Category' is appearing everywhere, then regardless of its appearing frequency it wouldn't have much impact on the 'performance'.



| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue | score | sum | performance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Regent Park, Harbourfront\n | Café | Park | Restaurant | Bakery | Pub | Theater | Food Place | Sports Place | Ice Cream Shop | Performing Arts Venue | 3.0 | 1836.3 | 612.100000 |
| 1 | Garden District, Ryerson\n | Restaurant | Café | Clothing Store | Food Place | Tea | Theater | Cosmetics Shop | Bookstore | Bar | Hotel | 20.0 | 11093.5 | 554.675000 |
| 2 | Richmond, Adelaide, King\n | Restaurant | Café | Food Place | Sports Place | Bar | Hotel | Clothing Store | Steakhouse | Cosmetics Shop | Lounge | 23.0 | 8878.0 | 386.000000 |
| 3 | St. James Town\n | Restaurant | Café | Bar | Food Place | Cosmetics Shop | Park | Farmers Market | Department Store | Sports Place | Lingerie Store | 22.0 | 7839.2 | 356.327273 |
| 4 | India Bazaar, The Beaches West\n | Restaurant | Food Place | Park | Fish & Chips Shop | Steakhouse | Liquor Store | Ice Cream Shop | Pub | Brewery | Pet Store | 4.0 | 1184.8 | 296.200000 |
| 5 | Central Bay Street\n | Restaurant | Café | Food Place | Tea | Bar | Sports Place | Department Store | Discount Store | Ice Cream Shop | Museum | 18.0 | 5230.7 | 290.594444 |
| 6 | Commerce Court, Victoria Hotel\n | Restaurant | Café | Bar | Food Place | Hotel | Sports Place | Deli / Bodega | Gastropub | Ice Cream Shop | Office | 29.0 | 8109.0 | 279.620690 |
| 7 | Berczy Park\n | Restaurant | Café | Bar | Food Place | Farmers Market | Cheese Shop | Bakery | Pub | Shopping Mall | Beach | 13.0 | 3538.6 | 272.200000 |
| 8 | First Canadian Place, Underground city\n | Restaurant | Café | Food Place | Bar | Hotel | Sports Place | Steakhouse | Gastropub | Deli / Bodega | Concert Hall | 29.0 | 7883.6 | 271.848276 |
| 9 | Little Portugal, Trinity\n | Restaurant | Bar | Café | Men's Store | Food Place | Record Shop | Ice Cream Shop | Brewery | Beer Store | Bakery | 16.0 | 4260.0 | 266.250000 |
| 10 | Kensington Market, Chinatown, Grange Park\n | Restaurant | Café | Bar | Food Place | Bakery | Grocery Store | Park | Gaming Cafe | Arts & Crafts Store | Dessert Shop | 19.0 | 4383.4 | 230.705263 |

*Top 10 Neighborhoods and their Top 10 Most Common Venue Categories*

Our x_set here was categorical, hence we needed to convert it using 'one hot encoding', grouping by neighborhoods and taking average of these columns returned our x_set. We fitted this data with 'performance' as our y_set using multiple linear regression.

The results we got were better than the analysis we have done so far:
Multiple Linear Regression:

| | | |
|---|---|---|
| x_set: Venue Categories | RSS: | 0.98 |
| y_set: performance | Variance Score: | 0.43 |

This result suggested the need to refine our x_set to reduce the unnecessary noise induced by other redundant categories. Thus the venue categories were ranked according to their regression coefficients and the multiple linear regression analysis was once again used against y_set, but this time our x_set included only 20 categories i.e. the top 10 categories having the highest coefficients which suggested that they had the most positive correlation with 'performance', and

last 10 categories having the lowest coefficients (negative) which suggested that among all Venue Categories these 10 had the most negative correlation with ‘performance’.
As expected, this resulted in even better performance of our model:

| | | |
|---|---|---|
| x_set: Top 20 Venue Categories | RSS: | 1.09 |
| y_set: performance | Variance Score: | 0.51 |

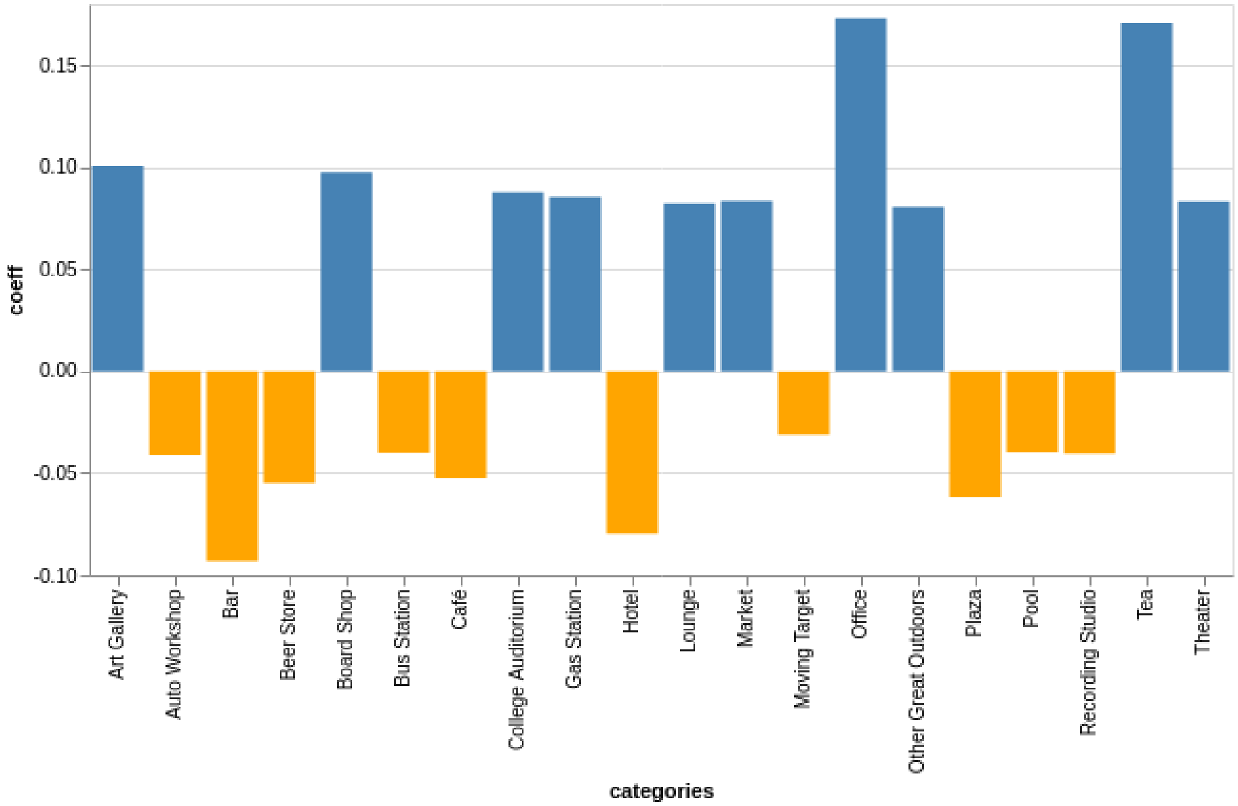


Figure above gave us a rough idea about the importance of these venue categories in determining the ‘performance’ of the neighborhood. The coefficients though, should be based on the first multiple regression model as this second one, although giving a better result, was based on assumption that these were the only venue categories in the neighborhood. Hence, the coefficients of the second regression model tend to provide a less reliable insight.

| Rank | Venue Categories<br>That increase ‘performance’ | Venue Categories<br>That decrease ‘performance’ |
|---|---|---|
| 1 | Office | Bar |
| 2 | Tea | Hotel |
| 3 | Art Gallery | Plaza |
| 4 | Board Shop | Beer Store |
| 5 | College Auditorium | Café |
| 6 | Gas Station | Auto Workshop |
| 7 | Market | Recording Studio |
| 8 | Theater | Bus Station |
| 9 | Lounge | Pool |
| 10 | Other Great Outdoors | Moving Target |

And this result seemed logical too!!

A lot of this was starting to make sense now, like one usually wants to grab lunch or dinner after the office hours, tea places don't usually serve meals, so instead of competing with your restaurant their presence will complement you! How you ask?
Say, “Me and my significant other went to this amazing tea place yesterday and we noticed a new Italian restaurant in the area, it looks nice, maybe we should go there sometime”
The same logic goes for rest of these places, these places are the ones where people usually spend their evenings and holidays. These are the entertainment centers, these aren’t food places, so the obvious conclusion, people are going to be hungry, and having your restaurant in proximity of these areas means only one thing, PROFIT!!

One look and you will know it, most of the ‘performance’ decreasing venue categories are your competition. Someone who has been to a bar or cafe, or maybe he's staying at the hotel or plaza, they have one thing in common, these all places serve food already, they all are going to have their tummies full. In simple terms, that means NO BUSINESS for you!!

With this, we think we found answer to our second question i.e. **What are the factors that are making these neighborhoods perform better?**

# Conclusion

Through the course of this study, we were searching for the best neighborhood to open a new restaurant in the city of Toronto. We decided to analyze the performance of current restaurants operating in Toronto and use the insights obtained to decide the criterions for the “best” neighborhood.
We thus needed to answer two questions:

1. Which neighborhoods are performing better in terms of restaurant business?
2. What are the factors that are making these neighborhoods perform better?

We answered the first question in terms of the ‘performance’ of the neighborhoods, we took two features indicating public response to any restaurant ‘Rating’ and ‘Check Ins’ and we calculated ‘Score’ of the restaurant based on these two features. This enabled us to rank the neighborhoods in terms of their ‘performance’ and thus we were able to answer our first question.

Now to answer the second question, we were needed to find out the factors that were having impact on our performance. For the first part of our analysis, we analyzed the features of the ‘Extra Data’ we collected for ‘performance’ using regression as our machine learning tool and we found out that having more:

- Tourist Destinations and,
- Cultural Location Index

Improved the ‘performance’ of respective neighborhood while having more:

- Vehicle Thefts,
- Road Length and,
- Linguistic Diversity Index

Reduced the ‘performance’ of respective neighborhood. While some of these factors were unexpected, one should take at least the first three into account for a better neighborhood ‘performance’ as a Tourism based and a Safe neighborhood will always attract customers to your restaurants.

For the second part of our analysis, we analyzed the ‘Venue Category’ of the top 100 venues within 500m radius from the neighborhood centers as returned by the Foursquare API which formed our secondary data. The aim was to find out the presence of which ‘Venue Category’ affects the ‘performance’, how much and, in what way? We used multiple linear regression here as our machine learning technique and found out that having certain ‘Venue Category’ in your neighborhood improved your ‘performance’ while the presence of certain ‘Venue Category’ took a toll upon our ‘performance’. This result will be of great help to identify the favorable neighborhoods to open a new restaurant and it will also help in comparing them in a quantitative manner. With this result, we were able to answer our second question.

Our analysis also returned another helpful insight, we, with available data, were able to roughly rank neighborhoods to minimize the initial expenses spent on ‘Parking lot facilities’ and ‘Housing Prices’. However, as we didn’t have the complete data, this result should only be applied after using the above ones to add a supplementary layer of decision parameters.

As the data returned by these data portals is dynamic, we believe that with these results, we should be able to use above mentioned methodologies to find the best neighborhood to open a new restaurant in the city of Toronto at any time. Investors and Entrepreneurs can use this analysis to make sure they are investing in the right place. Such analysis will prove to be our key to success!!

# Further Directions

The data we analyzed in this study was of socio-economic kind. It tends to have a number of uncertainties and sometimes even lacks sufficient data. The maximum R2 score we were able to obtain with these models was 0.51, that leaves out a significant variance that could not be predicted by the models in this study. Furthermore, all datasets used in this study are dynamic, some of them, like Foursquare database update daily. Thus, it was observed that the results tend to update each day by a slight difference. This is a ‘pro’ as well as a ‘con’ to our study, as our results update over time, we should be able to predict the best neighborhood at any time in future, on the other hand, after sufficient time, the results we have derived today, may be different than those derived that time. This study was conducted on the databases available on 22nd August 2020. One thing we can do here is to study how these results vary over time and thus try to improve their reliability or predict their pattern.

There was this another difficulty, the neighborhoods determined by the Canadian Postal Service are different than the ones determined by Toronto city’s administration. This resulted in several difficulties while performing data cleaning and data analysis, as without a unique primary key, relating and joining databases became impossible. Furthermore, the Foursquare API doesn’t recognize these new neighborhoods as places, hence there was no firm criterion to relate these two neighborhood lists either. As systems get updated, we will be able to eliminate this problem in a better and a standard way.

With digitalization and development of projects like ‘Open Data Portal’, we will soon be able to widen our scope of the research and thus deliver better and reliable models for any city in the world.

## References

1. Toronto - Wikipedia
2. Open Data Portal – City of Toronto
3. Foursquare API
4. Blogs
5. Geospatial Data