New Pizzeria in Toronto Data Science Capstone

Introduction

In Toronto, Canada there is a ever expanding demand for new restaurants as new tech companies roll in and residents demands new food options. As one of the fastest growing cities in the Canada there is no doubt a market for a new restaurant despite the already established food industry. Every famous chain from Canada usually has some start in Toronto, but from all the 82 chains started in Canada there are only a handful of Pizzerias. This is why any new restaurant looking to get a hot start should start there get raving reviews from the hordes of foodies that explore the city in search of new food spots. One of the weakest points for food in Toronto, however, is its Pizza scene with few truly amazing pizza places to choose from. This is where we must capitalize on the markets lack of exposure to Pizza before anyone else does to claim the spot as the top Pizza restaurant in Canada.

Business Problem

Location, location, location. As much as people have begun to drive more because of COVID-19 the location is vital for success. The new restaurant must not be too far from the food scene to be a burden for customers, but it can't be washed away buy competing restaurant nearby if there is too much competition. To find the perfect balance there must be a balance in the amount of neighboring restaurant that will surround the potential location for this game changing restaurant. The best way to figure that out is to see how many amenities there are nearby to attract local and visiting Toronto tourists. From there we can see how many restaurants are in the surrounding area to be able to choose an exact spot for the new Pizzeria. With an estimated 1.8 Million new jobs opening within the restaurant industry it’s easy to see why now the perfect time to open a new restaurant.

Data/Methodology

i.) Data Collection

To best understand where we should place our new pizzeria, we must have the data to back up the placement of the restaurant. This is where we will use data extracted from Wikipedia from a page where all the postal codes are listed. This data will read and sorted into a pandas data frame using BeautifulSoup package. Once getting this data we will sort it into columns “Postal Code”, “Borough”, and “Neighborhood” which will stay consistent throughout the entire notebook. If the Borough is not assigned, it will not be listed and ignored when transferred.

ii.) Data sorting

Now that all the data has been transferred from Wikipedia, we can clean the data from the pandas data frame called “data”. The data will all be formatted into a format that easier to read. Any instance where we see the data not contain a neighborhood name where it “not assigned” it will be replaced with the Borough name. Before we can truly explore our data, we must call the geospatial data to collect latitude and longitudes of each postal code to best understand what each neighborhood looks like. To do this we will use the csv “Geospatial Data” to collect the information that contains the locations of our postal codes. Once collected we now have two data frames, one with the neighborhood names and one with the locations of each postal code. Due to this similarity we can match it using that to match neighborhoods with corresponding latitudes and longitude data.

iii.) Data Exploration

Calling the foursquare API will give the exact information over the neighborhoods that we need to understand what is going on in each neighborhood. First though we will define our location and more importantly the search criteria, which in this case will be “Italian Restaurant” and anything within 10000 meters of downtown Toronto. After inputting the corresponding client information, we can login into the API. Once obtaining the information in a JSON format we can turn that information into a data frame which will be

more useful to the digest all the new information. After some sorting and cleaning we are left with the names, categories, and locations of nearby Italian restaurants.

Our first map will visualize where the local Italian restaurants are located from the epicenter of Toronto. Here we can understand what parts of downtown Toronto contain the greatest number of restaurants. To better understand what is going on in each part of Toronto, we will look to sort out what each neighborhood has to offer by listing all nearby venues. Using a For loop we will find all venues for every neighborhood that we listed earlier in our data frame. Grouping our values by neighborhood will make it easier to then cluster our data further down. This make it easier to visualize what each neighborhood has to offer.

iv.) Machine Learning

In order to figure out what each neighborhood has to offer we will use a method called one hot encoding which will turn our categories into numerical values. For every venue type we will see if each neighborhood as that type of venue, and in a data, frame called "Toronto_grouped" we will find the mean compared to each type to find the frequency. We will be looking for the top 5-10 venues in each neighborhood.

v.) Data Analysis

Now that all of our information over the neighborhood's specifics are listed, we can understand what each neighborhood has to offer. We can use a knn algorithm to group our venues so that we can best visualize where exactly in Toronto is best known for having things to do. We get our data frames with neighborhoods, locations, venues, and clusters and use two different maps to understand where the venues are in relation to the Italian restaurants. Using Folium maps we can now see the results of our clustering. Below that we can then use tables to see what each cluster has to offer.

Discussion/Results

Overall, there are several tools that can be used to understand the layout of Toronto's food and entertainment scene. The tools used for this capstone may not be the best for every scenario, but they provide users with good visualization and easy to understand steps to sort data. This showed us that extracting data from online sources is important to best understand relevant problems as we move into a more data dependent world. The results from this capstone helped a small business navigate the busy streets of Toronto without once having to step outside. This becomes more and more important as we become more globalized but require information over places we may have never been. Sources like Foursquare API allow users to gather information over locations around the world and empower use to keep discovering.

Conclusion

In conclusion we see that downtown Toronto is the best place to put our restaurant, more specifically near the Toronto union station as a lot of venues that attract food traffic will low through that area. For a restaurant, having great visibility by anyone is the most important thing. In the end, this course taught us to us data to tell a story, and for this pizza shop, this story is just about to begin because of data science methods.