03/11/2020

Battle of the neighborhoods

# **introduction**

## **Background**

## Food is provided in many forms and is offered in a variety of ways. You can eat it, drink it, dip it, decorate it but ultimately, enjoy it. What makes food so unique is the diversity behind the variety. Different cultures bring out something new and unique which expands out pallet and mind. Some may not like certain types of food due to texture or taste, and that is fine. It’s what makes the variety of food so great as our world grows, culture is brought to life through food. For this example, the information provided will revolve around the Argentinian food culture and how one may use this information to open an Argentinian styled restaurant in New York City, New York, or Toronto, Canada.

## **Problem**

## Since the variety of accessible food and information about restaurants is easy to obtain, the data provided will contribute to locating a city where an Argentinian restaurant can flourish. The project will provide two locations, for both major cities, based off data received.

## **Interest**

## Entrepreneurs, Chefs, Food enthusiast, and tourists would be interested in opening up their own restaurant and/or visiting a solid location filled with wonderful foods.

# **Data Acquisition and Cleaning**

## **Data Sources**

## We would need customers, from a diverse area, in a diverse region. The United States Census Bureau, found [here](http://www.census.gov/), collects data and provides prediction charts to simulate 2019 Census results-- in comparison to 2016. A variety of analysis is also shared but population is specifically used. For Toronto population, I scraped ‘Demographics of Toronto’ in [Wikipedia](https://en.wikipedia.org/wiki/Demographics_of_Toronto). Specifically, table nine through twelve. These tables contained a breakdown of population per Toronto Borough. For latitude and longitude data in Toronto Boroughs, I used ‘List of postal codes of Canada:M’ on [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Using postal codes, I used Google API to get the latitude and longitude information. Four Square API is used to locate three queries per counties. There are six counties searched in New York and nine counties searched in Toronto. Foursquare is also used to receive queried results, of venue locations, and top checked-in venues for requested counties.

## **Data Cleaning**

## For this brief study, we focused on New York City Counties population, specifically in the following regions: New York City as a whole, Bronx County, Kings County (Brooklyn), New York County (Manhattan), Queens County, Richmond County. The population data received by the US Census is an excel document named “QuickFactsFed2020.csv”. After reviewing the data as a whole, I removed all ‘NaN’ values, and all rows not pertaining to “Population estimates, July 1, 2018, (V2018).” Afterwards, all values leftover in the one row was converted to an integer for future use. A separate table was created to present percentage of the total population by borough.

## For the Wikipedia scraping of ‘Demographics of Toronto’, the following tables were used: Toronto & East York, North York, Scarborough, and Etobicoke & York. After creating a DataFrame with the requested data, I removed all columns besides ‘Population’. Cleaning this data included removing all null data, replacing ‘/n’ values with blank spaces, removing comma values as this will not filter through folium maps later, converting all values to float, and adding a total value of all population within the borough row with NumPy. This sequence was applied to all 4 tables scraped.

## For Wikipedia scarping of ‘List of postal codes of Canada:M’, I removed all ‘/n’ values created during the import and verified each row had a value. If the value ‘Not available’ was found, it was removed. After retrieving the latitude and longitude, postal codes were aligned with the respected lat / lng data. Feature headers were renamed from ‘Borough\_x’ and ‘Neighbourhood’ to ‘Borough’ and ‘Neighborhood’. Borough\_y was removed. Foursquare data cleaning included: removing many duplicates by ‘id’, retrieving Latitude and Longitude location data from the location feature column then creating a feature column for lat/lng, retrieving the category of venues and creating a feature column. Lastly, adding cluster label values to separate by borough.

## **Feature Selection**

## After cleaning the data, 20 features for both New York City and Toronto population data. For New York City Foursquare data, features included name, categories, location, latitude, longitude, city, id, and cluster labels. There were a number of issues found when retrieving the Foursquare data. Due to the search revolving around the radius of the given latitude and longitude, many searches came back with duplicates. The geography of New York City made the search radius overlap one another. To combat this, select cities were removed from searches that did not align with the city being searched. To verify, I used folium maps to visualize the results. I modified as needed for each New York City borough.

## Toronto fortunately had 4 results found for our study. The same verification methods were used in the case a location was close to a borough border.

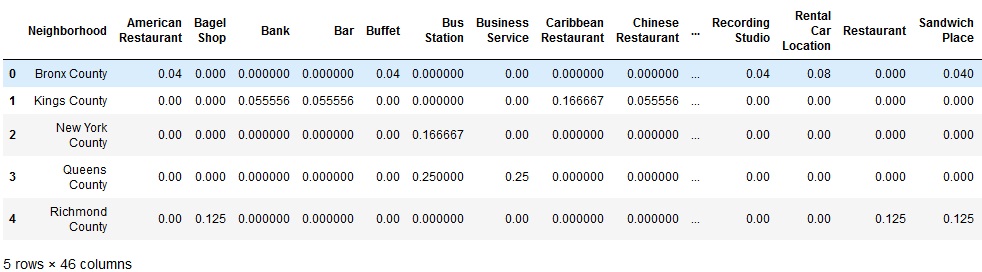
# **Exploratory Data Analysis**

## **Folium maps module**

## The folium maps module was used to verify and explore the regions we were searching. After requesting the latitude and longitude values from Google API and Geocode API, I used those variables to locate New York city boroughs and Toronto postal codes. After plotting, it was found that some tweaking of the radius parameter was needed as New York City radius parameter calls overlapped one another.

## **Foursquare: Top venue check-in's**

## Foursquare top check-in's by venue was created by gathering a limit of 100 venues per borough/neighborhood searched, via the latitude and longitude variables. The category and venue name were extracted and added to a DataFrame also containing the neighborhood name and lat/lng. Venues were then grouped by neighborhood but displayed by the count value. One hot encoding was used next to prove a 1 or 0 on the category of the venue as shown below



## The mean of all values in the finalized encoding was then displayed in a separate DataFrame. Finally, the data was converted to the frequency of the mean per county, as shown below, to provide the top 10 most common venue check-ins per neighborhood.



## The same method was applied to both New York City and Toronto boroughs. This information is used to locate the popularity of restaurants or cafes in each borough. The amount of potential foot traffic, and in this case four square check-ins, per borough. Lastly, the information was used to identify the diverse selection and popularity of existing restaurants and/or cafes in each borough.

# **Predictive Modeling**

## **US Census Predictions**

The U.S. Census Bureau prediction models were used to calculate what the population would be from 2016 through 2018. As the Census is taken every 4 years, various sources of data were collected by the bureau and was models to closely predict outcomes of population.

“The U.S. Census Bureau conducts a variety of research projects that use administrative records linked to census and survey data. Some of these projects generate new social and economic statistics, while others investigate ways to use linked data to improve sampling frames, survey measurement, and reduce respondent burden.

Additionally, U.S. Census Bureau collaborates with other stakeholders, including federal and state partners, to support the need for using statistical research to conduct evidence-driven program evaluation. Collaborations with researchers from academic and other institutions have led to the exploration of new avenues of research.” (Census.gov, Data Linkage Infrastructure)

Unfortunately, predictive and regression models were not used in this quick study but future research will be conducted and explained below under ‘6. Future Direction.’

# **Conclusion**

## **Final thoughts**

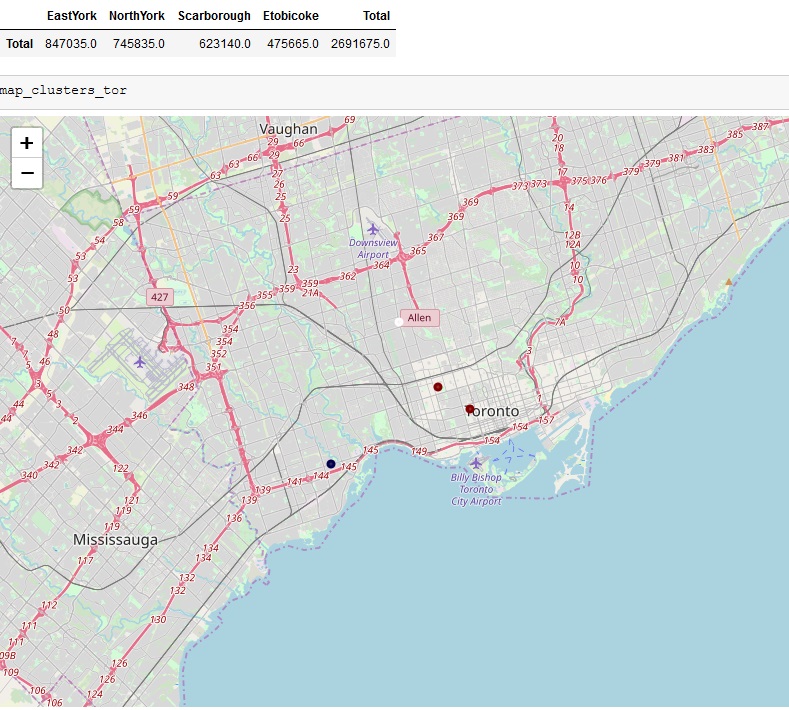
The purpose of this project was to identify and recommend two neighborhood cities within two diversely popular counties. The results would aid prospective restaurant entrepreneurs to open an Argentinian styled restaurant/cafe. Using the United States and Canadian Census, we located highly populated and culturally diverse boroughs within two major cities. Google API was used to find the latitude and longitude of the selected boroughs for each city. Foursquare API gave us access to further analysis on by calculating and providing the top 10 most common venues per borough; which provided information on competitors and consumer trends in the area. Foursquare also provided information such as restaurant name, latitude, longitude, and category name of the Argentinian restaurant queries we requested.

We then clustered, by major county, the results of the queries and provided a visual map representing each result by its respected borough. This map will help aid future investors and entrepreneurs in locating a future location and can be used to find competitor locations to aim for competitor clients or avoid locations all together.

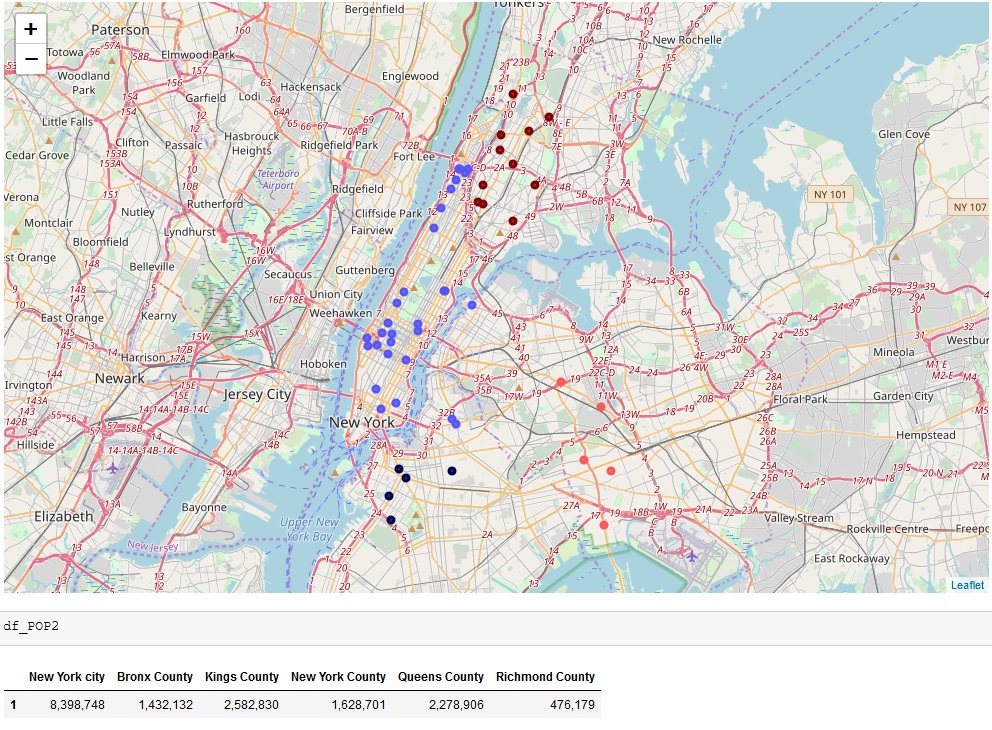
Finally, we pinpointed two boroughs within New York City and Toronto which aligned with our purpose in providing solid locations for a future Argentinian restaurant. Each location was chosen based off low or no amounts of current Argentinian restaurants in a 6000m radius along with the highest population in selected borough.

**5.2 Conclusion**

Toronto & East York make up a good percentage of Toronto's population, at 847,035. The amount of check-in's on the top 10 most common venues that happen to be restaurants/cafes also landed on Toronto & East York. The diversity found in the search leads me to believe an Argentinian restaurant would be a competitive location in the East York & Toronto city region. Most notably in the East York region as there are no Argentinian style restaurants as shown below. North York would be the next best location as there are no Argentinian locations found during our foursquare searches.



Kings county, New York has a population of 2,582,830 while having five Argentinian styled restaurants. Kings county is provided three of the top 10 most common venues relating to food. The variety range between Caribbean, Chinese, and Mexican restaurants.

This result, along with continued population growth, allows us to recommend Kings County as a good county for future Argentinian Restaurants. Other Counties in the area have more than 5 results with lower population count. Kings and Queens county lead with the most population in New York City.

# **Future Direction**

## **What else can I provide?**

## As this is my first project, many ideas came to light throughout the course of the research. Although time was an issue in creating and providing the results to all ideas, I wanted to list what future possibilities could be added to solidify the prediction. Ideas can also alter the prediction, so we will leave that possibility open.

## The U.S. Census data provided information around regional economic wages and ethnicities. The Canadian Census also has the same information. Other information such as business restaurant trends, profits, and economic growth could be factored into picking a precise location for a future restaurant. In the future, I plan on updated this project to reflect those features and apply different scientific models in hopes to further solidify a restaurant location.