



- IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Using k-Means clustering to find the best area to set up a pizzeria in Toronto

Table of Contents

Pizzeria set up

Data Collection and Cleaning

Data Analysis

K-Means Clustering

Result Analysis

Conclusion and Recommendations



Pizzeria Business Set-up

This project looks at a new pizzeria opening up in the Toronto City. The management of the pizzeria wants to get a strategic location that will be profitable once the pizzeria opens up. To arrive at this, there is need for guidance to choose the best possible location to set up to ensure that the business is profitable.



Problem Statement

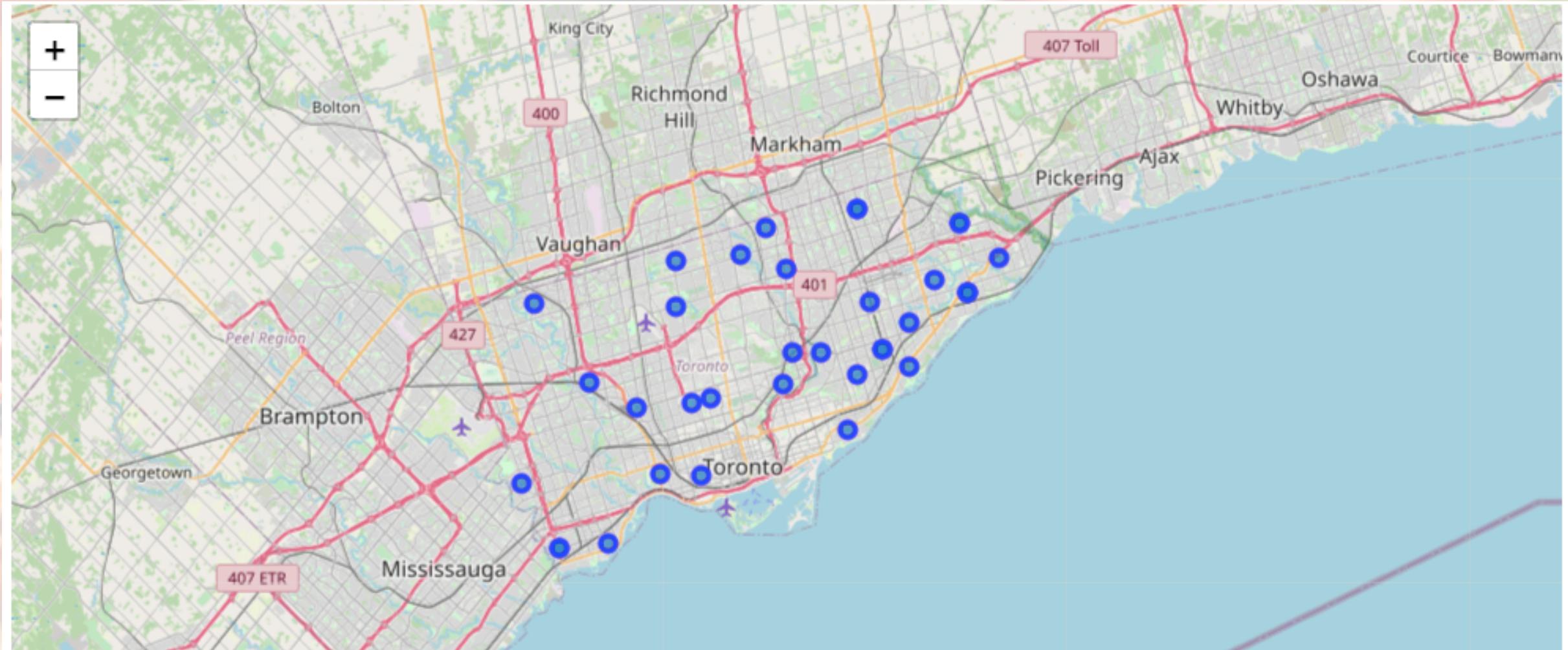
To find a strategic location to set up a pizzeria in Toronto City leading to good sales.

Data Collection and Cleaning

- 1 Web scrape Toronto neighborhood data from Wikipedia.
- 2 Find the number of nearby restaurants using Foursquare API.
- 3 Household income and neighborhood population size data extracted from Toronto Wellbeing.
- 4 Household income, neighborhood population size and nearby restaurants are all merged into a single data frame.
- 5 The clean data contains 3 features (household income, population size and restaurant count)

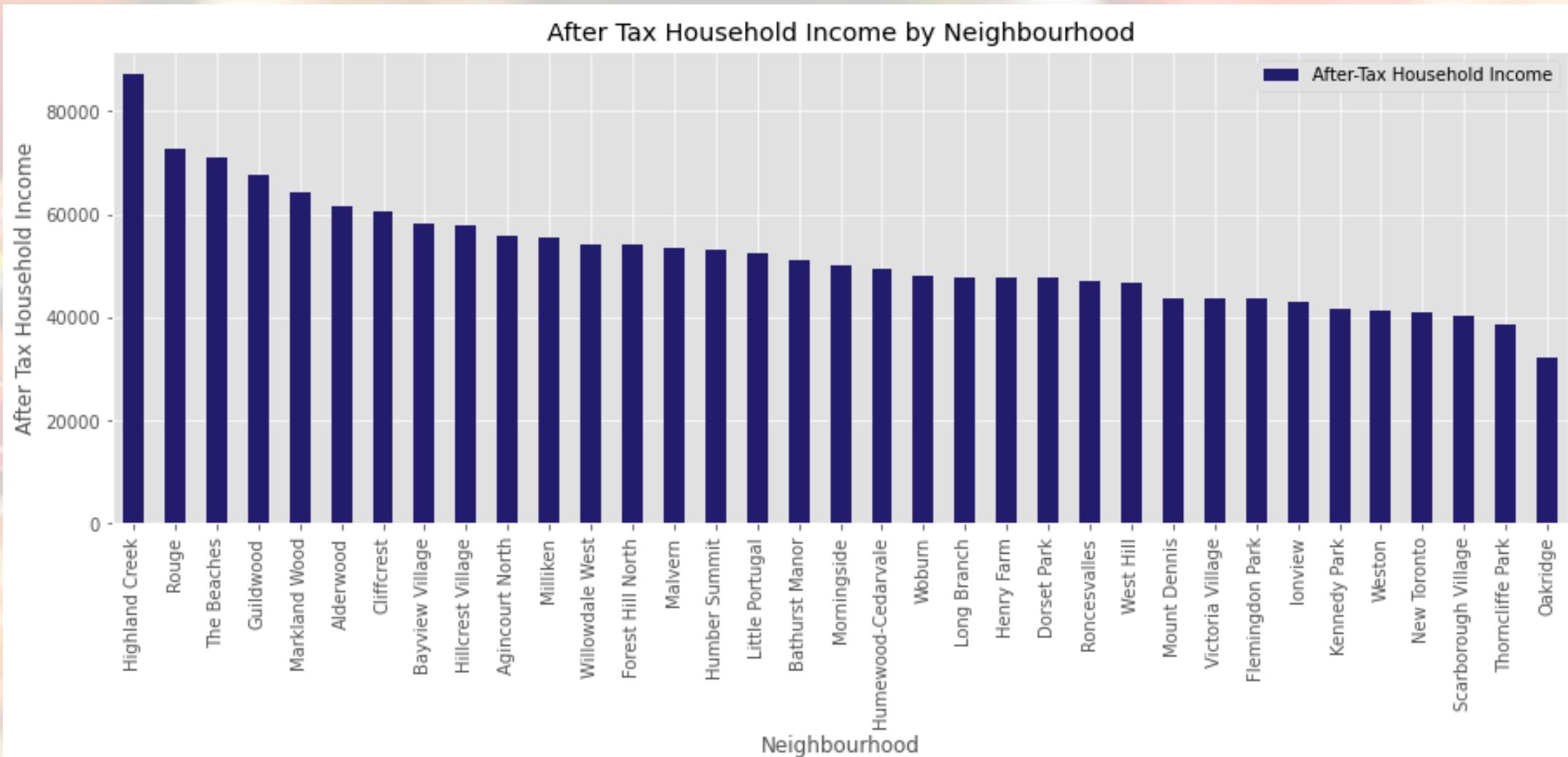
Data is collected and cleaned through these processes.

Data Analysis : using Folium mapping



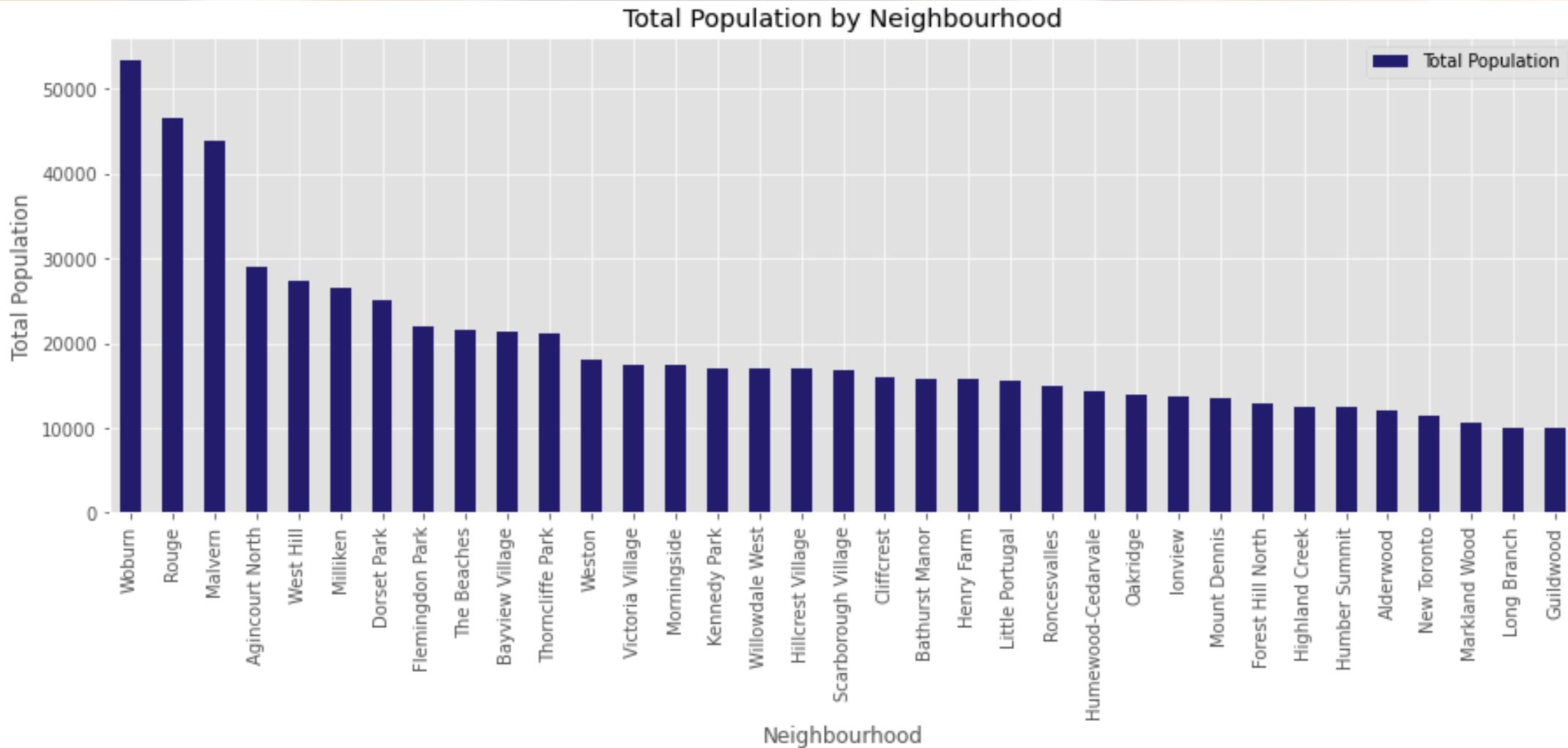
A folium map showing Toronto Neighborhoods.

Data Analysis : Household Income by Neighborhood



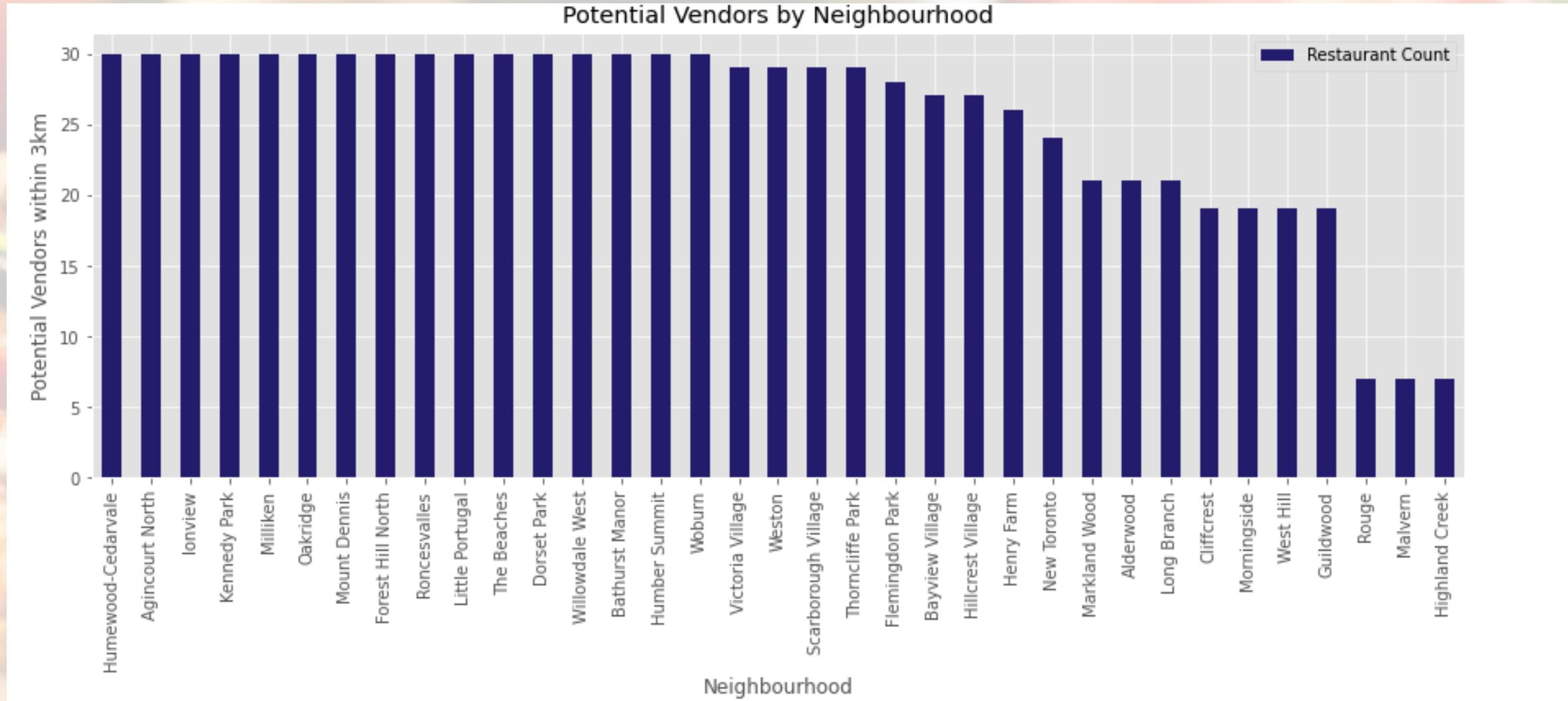
The graph for Toronto Neighborhoods showing household income after tax by the neighborhood .

Data Analysis : Population by Neighborhood



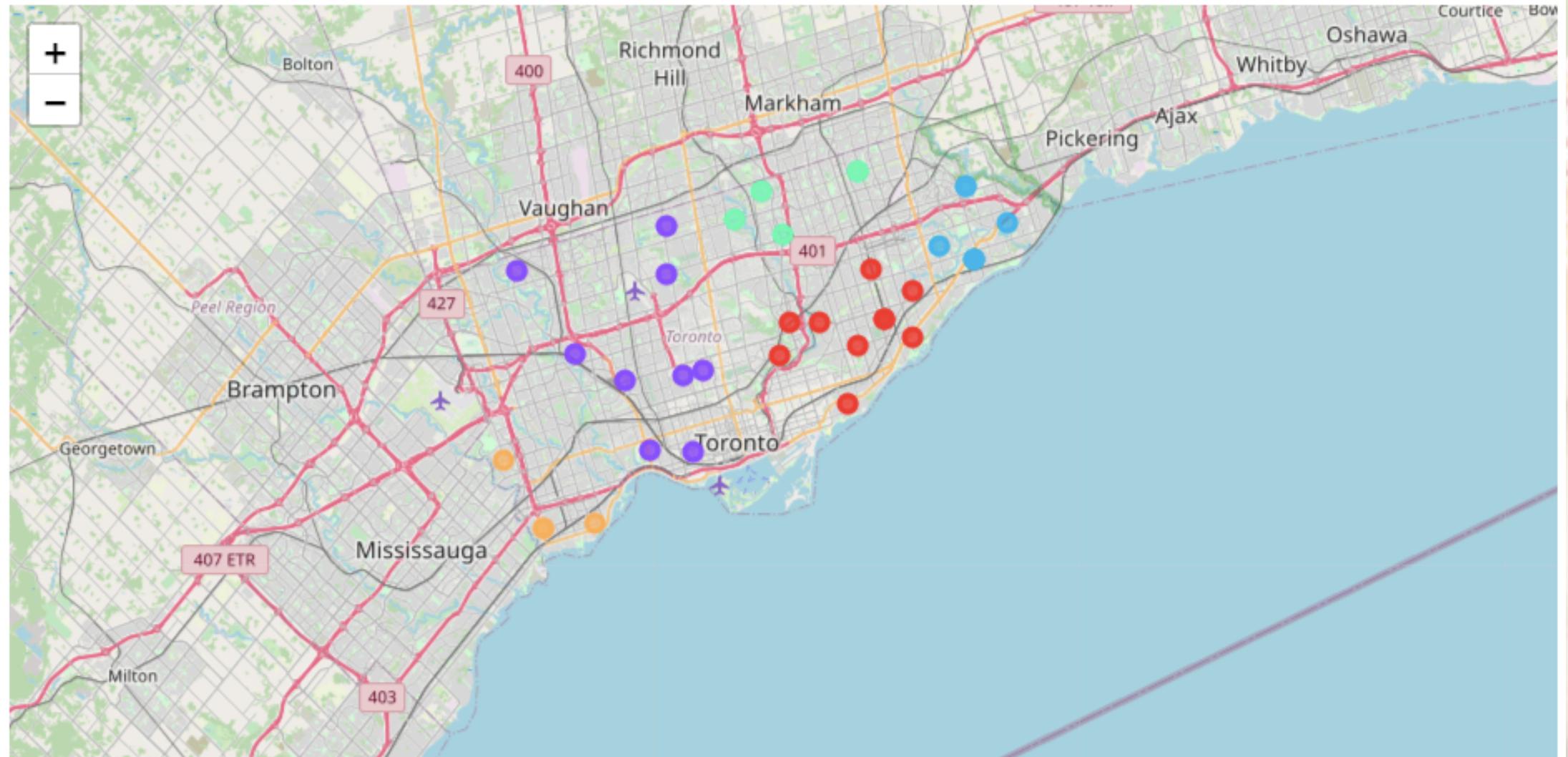
The graph for Toronto Neighborhoods showing total population by the neighborhood .

Data Analysis : Vendors by Neighborhood



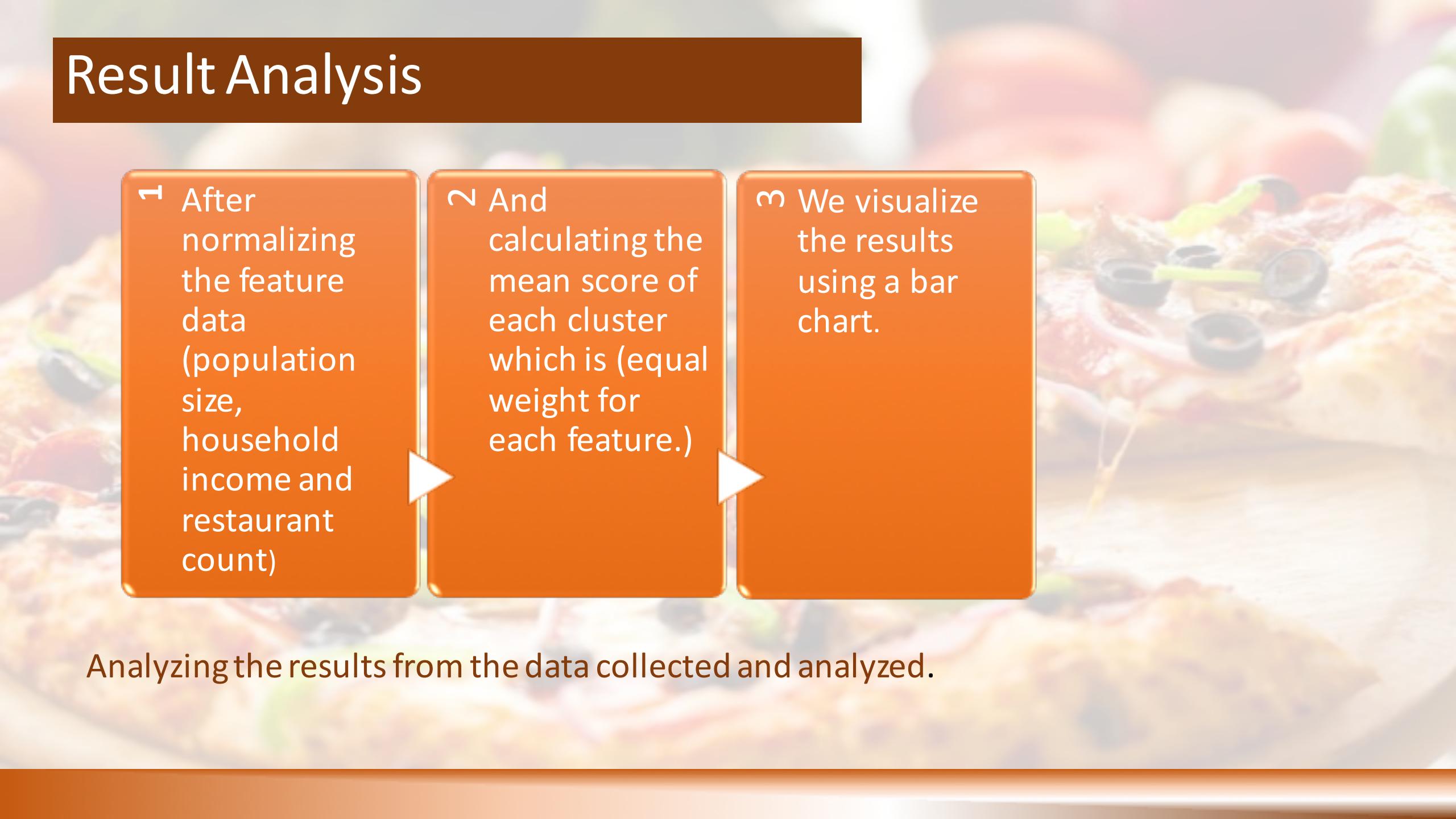
The graph of Toronto Neighborhoods showing vendors by their neighborhood .

K-Means Clustering



K-Means returns a map showing **FIVE** clusters

Result Analysis

- 
- After normalizing the feature data (population size, household income and restaurant count)
 - And calculating the mean score of each cluster which is (equal weight for each feature.)
 - We visualize the results using a bar chart.

Analyzing the results from the data collected and analyzed.

Conclusion and Recommendations

Results:

Looking at the results on the graph on the right hand side, we can see that cluster(neighborhood) 3 is the best location to set up the pizzeria. Cluster 2 is also a good fit since both those clusters have a score higher than 0.5.

Recommendations:

1. The Model shows that certain locations in Toronto have more restaurants than others. This however shouldn't be the only variable considered before setting up the pizzeria. It would be best to consider population in areas where there is high number of restaurants.
2. Population and average household income after tax used in this model are also limited. The cost of food on average should also be looked into as setting the wrong price couldn't lead to losses and not profits as majority of consumers might be priced out.
3. Another factor to be considered could be cost of property or rent. Too high rent means decrease in profits which reduces success of the model.

