

Capstone Project: Toronto Real Estate Development

# Vlad Dumitrescu

[**Executive Summary**](#_momqpgiu24rm)2

[**Goals**](#_3at9u9s4e0vp)3

**Data** 3

Census Data 3

Geospatial Data 3

**Methods** 4

Cleaning & Preprocessing 4

Normalization 4

K-Means Clustering 4

**Mapping** 4

[**Conclusion**](#_kkkd1nddgdpx)5

[**Appendix**](#_lwnzwvfcgjw0)6

# 

# Executive Summary

### Toronto has long been one of the top cities in North America. Recent studies have shown that the city is near the top when it comes to the well-being and happiness of its citizens. When the population is happy, business tends to boom, and this is exactly what has been happening with Toronto’s real estate market. This project will focus on helping the investors of a major condominium real-estate developer who are planning on expanding their enterprise to Toronto. The group, however, has no prior knowledge of Toronto's neighbourhoods, and demographic patterns. Our team has been approached by the investors with the task of finding the optimal location in Toronto where they should focus on constructing their new luxury-condominium project.

# Goals

### The goals for this project are simple, to create a comprehensive model of Toronto which can give insight to investors of where the most optimum location of a condominium project should be. If done properly, the project should be successful and should be surrounded by the right kinds of people. Not only does this model apply to our current clients, but it can also be a blueprint for all future developments in the city.

# Data

## **Census Data**

### The data used in this study was retrieved from the city of Toronto data repository and include two major datasets. The first, data from the 2016 national census in Canada which includes information on approximately 3000+ attributes of the 140 neighbourhoods in Toronto (Figure 1). This data was highly comprehensive as it was developed for the city, and by using various tools such as R and Python, I was able to narrow down the attributes to include only the census data that was applicable to the topic of this study. Variables such as Age, Employment, Income, Education, and Population Size, were all chosen as appropriate determinants of a neighbourhoods profile and through preprocessing, I was able to create a new dataset which you can see on my notebook submission.

## **Geospatial Data**

Additionally, the city of Toronto had also included geospatial data on the same set of neighbourhoods in GEOJSON and CSV formats (Figure 2). I chose to use the CSV format because it followed the same guidelines as the census data was merging the two datasets together proved easy. The geospatial data was needed in order to create a map of the final model after clustering.

# Methods

## **Cleaning & Preprocessing**

The first step when dealing with this data was to prepare it for the upcoming modeling. The data included certain columns and characteristics that needed removal, and overall this step was most important. It is also important to note that the original data that I was able to find on the city of Toronto’s website had been used for many studies and had already been thoroughly managed, however, for this particular study, a specific set of guidelines had to be set (Figure 3).

## **Normalization**

Once the data had been properly looked over and cleaned it began to be clear that in order to build a model on the different attributes I had chosen, I needed to eliminate any bias from the data. Normalizing the numeric attributes in question was a way to keep the data within the same range without losing any information in the process (Figure 4 & 5). After normalized, a simple sum of the five numeric attributes gave each neighbourhood a standardized score that represented our client’s interest in the neighbourhood as a potential location for the new condominium project.

## **K-Means Clustering**

# My original plan was to use the total scores derived from the normalization stage to cluster the neighbourhoods. After many iterations, 3 k-clusters were chosen as the optimal number of clusters and K-Means clustering was ran on our data (Figure 6).

# Mapping

After clustering, I was able to use the folium package that we had learned about throughout the course to create a map of Toronto. I plotted the top 20 neighbourhoods by total neighbourhood profile score on the map and a few patterns began to emerge (Figure 7). As the final output, this map is a clear representation of the makeup of Toronto’s neighbourhoods and thereby a useful tool for real estate developers.

# Conclusion

After modelling and the creation of the final piece of our deliverable, it was clear that certain patterns exist in the landscape of Toronto’s neighbourhoods. Areas with a younger population tended to have more work class citizens and then, in turn, more disposable income. However, it was interesting to note that certain areas did not fit this stereotype and exceeded expectations. Rouge Hill in Scarborough scored high and was included in the top 20 as well as other suburban neighbourhoods in which you wouldn’t expect to find luxury condominiums. This can be chalked up to the fact that there is a lot of potential for high-rise condos in these areas. All in all, our team would stand by the study we have completed and in the insights we have delivered to our clients.

# Appendix

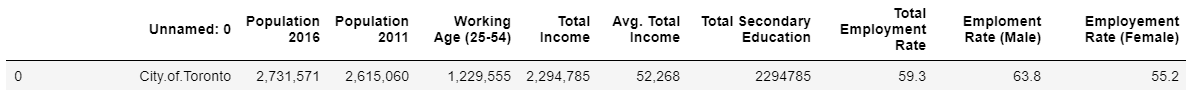


Figure 1 (2016 Census Data: Cleaned)



Figure 2 (Geospatial Data: Raw)



Figure 3 (Geospatial Data: Initial Cleaning)



Figure 4 (Normalizing numeric columns)

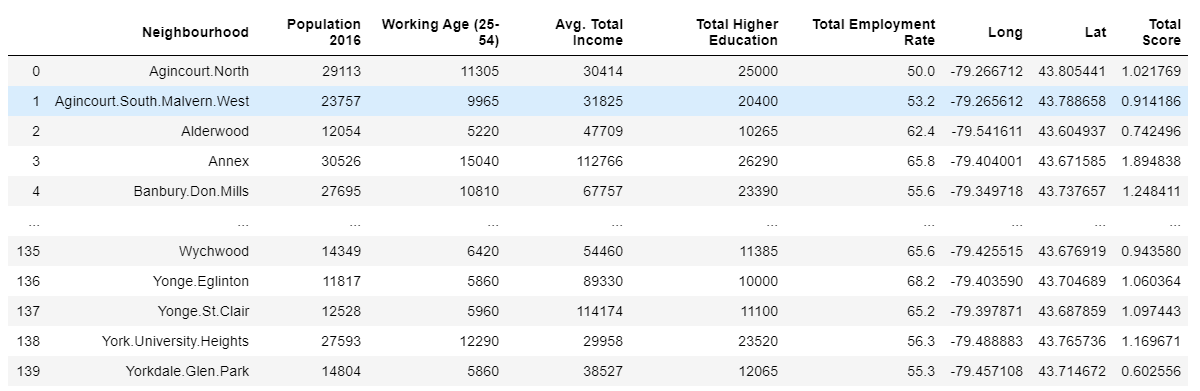


Figure 5 (New data including total score column)



Figure 6 (K-Means Clustering & Corresponding data with ‘Cluster Labels’ Column)

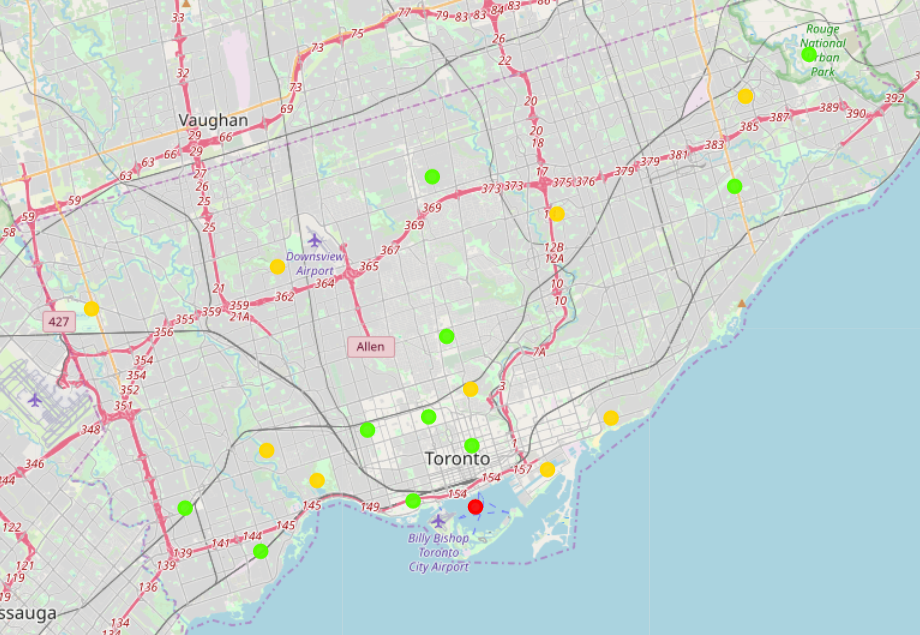


Figure 7 (Map of Toronto with Clustered Neighbourhoods)