IBM Data Science Certificate

Capstone Project

Understanding Neighborhoods

Jordi Capdevila

supervised by

March 3, 2020

Contents

List of Figures

List of Tables

Let's find out with data analytics if the neighborhood you live in shaped you or you build the neighborhood you live in.

1 Introduction

Introduction where you discuss the business problem and who would be interested in this project. Data where you describe the data that will be used to solve the problem and the source of the data

Problem description

Let's say for instance that you're willing to open a store (i.e. franchise) in a city you barely know. Besides from the obvious answer: "Go to the city center!", what else you can do to find the sweetest spot in town? There's a chance that analyzing neighborhoods will give you some more 'hidden' information to better solve the main question: 'Where should I open the store?', or at least is a starting point to your endeavour.

Data description

That's a more obvious question: 'What kind of data do I need to help solve the main question?'. Short answer: 'Well, that's store and neighborhood information'. And the other question: 'Where do I get all that information from?', let's find out:

- Venue information: information provided by Foursquare API and REST API methods to retrieve all sort of details regarding venues.
- Neighborhood information: to get a better understanding of 'How's this neighborhood?'.

  Shape : that solves the question of 'How big or small a neighborhood is', thankfully this information is provided as a GIS file from the Open Data Portal of Toronto. This file provides information from each neighborhood of Toronto, such as: area code, name, latitude and longitude of neighborhood's centroid, area shape and length.

  Demographics : responds to 'Who lives there?', also from the Open Data Portal of Toronto, we'll get information reagarding age, population, dwells and much more for each neighborhood.

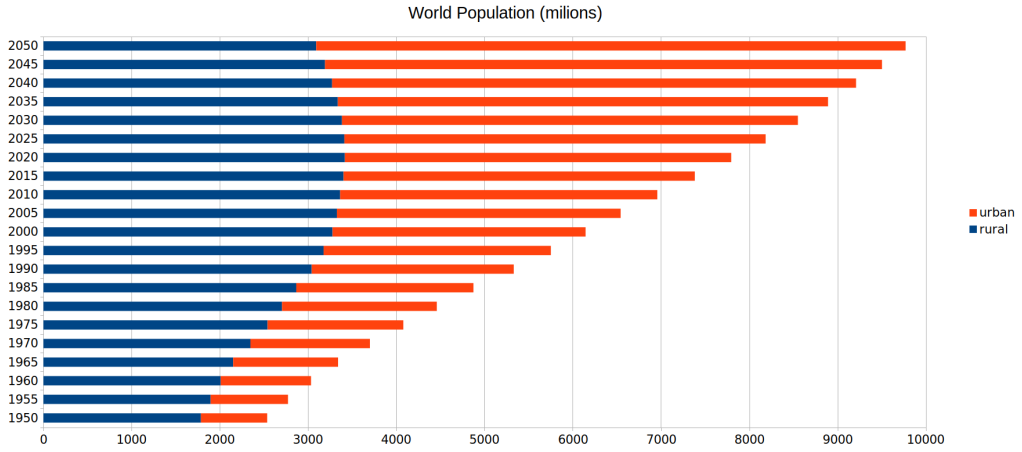

Figure 1: United Nations, Department of Economic and Social Affairs, Population Division (2018). [2]

.

2 Methodolgy

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

Following a previous assignment where we used Foursquare data, to cluster Toronto boroughs, I did a quick search on cities density population [1]. As per my exploratory analysis, I did a top-bottom research to answer the rationale about what I've considered as a problem.

First, let's explore the world population (Figure 1), a sustained growth of 1 2% a year, is expected in the incoming years. Plus, a shifting evidence that rural areas will be less populated in favor of urban areas. Keep in mind that the same observation is valid for developed and less developed regions, as shows Figure 2.

Bottom line, a more densely populated cities is expected in the incoming years. Let's pick Toronto as an exmaple, and find out what trends are expected (Figure 3), a growth over 4 5% in the next 5 years or 300.000 persons will arrive to an already densly populated city. Question here is *'How well is Toronto prepared to absorb almost 60.000 persons a year?'* or from a business point of view *'Where should I open a store to make it work? '*. That is the main point of this report.

To answer the question above, the plan is to cluster Toronto's neighbor-

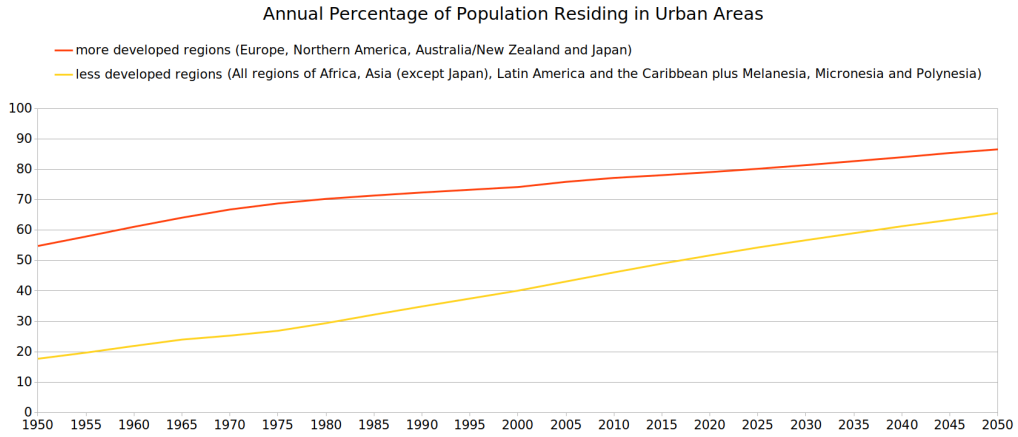

Figure 2: United Nations, Department of Economic and Social Affairs, Population Division (2018). [2]

.

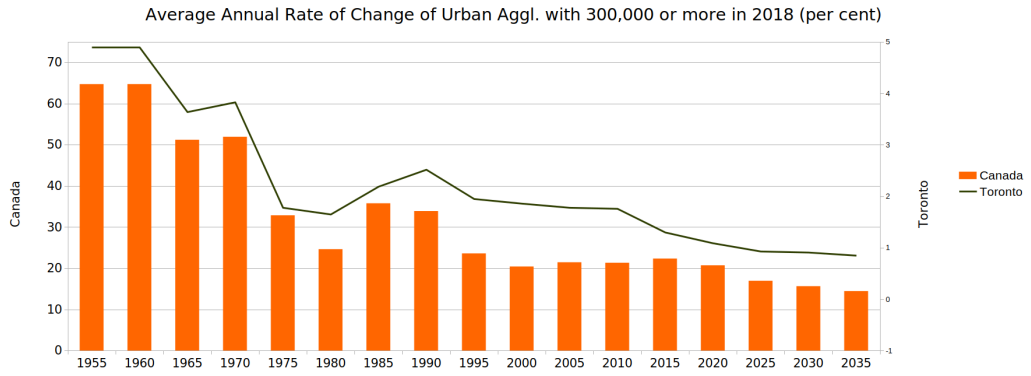

Figure 3: United Nations, Department of Economic and Social Affairs, Population Division (2018). [2]

.

hoods and their neraby venues to get an idea of what sort of businesses shape each neighborhood. And following the main topic, if there's an increase of the population, extrapolate the results to show which businesses are more likely to work.

The main method used to cluster is K-Means, where the centroids are given as longitude and latitude for each neighborhood, and for the number of cluters (k), the *elbow method* will be used to satisfy the distance between the centroids and data points.

3 Results

Results section where you discuss the results.

Let's show the results of the research to solve the main problem under investigation. Figure 4, shows the result of cluster based on data gatherd from Foursquare and Open Data Portal.

Foursquare	
Venue Distance:	in meters, between the centroid and venue
Venue Category:	depicts a more general venue category

Open Data Portal of Toronto	
Area Name:	Name of each 140 neighborhoods in Toronto
Borough Name:	Name of realated borough
Pop. Density:	per neighborhood area
Geometry:	GIS format file with borough's edges

Table 1: Data used to generate dataset for further analysis.

All the information is usable out-of-the-box, but the category field from Foursquare is not, so had to it manually. That means group all venues that fall in the same category, such as: *Restaurants, Shops, Stores, Fast food, Bars, Coffee Shops, Transport, Art&Museum, Tourism, and Sports&Leisure.* To help shape similar neighborhoods also percentiles of venue distance and amount of venues is included in the dataset.

As a final result of the neighborhood cluster, Figure 5 is presented. That will be used in next sections to make assumptions and final conclusion.

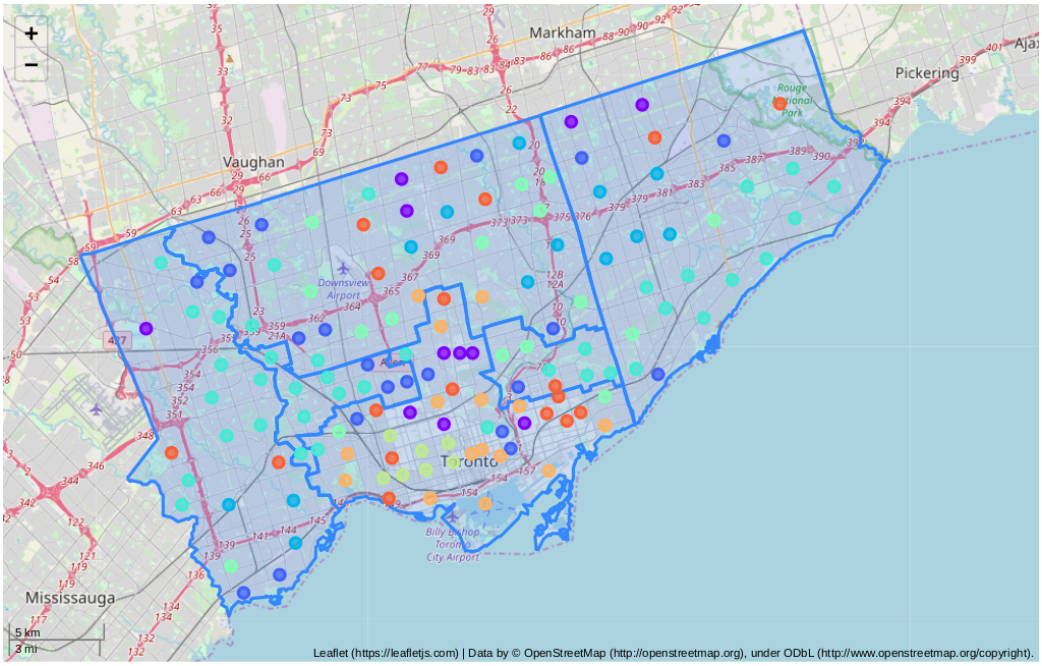

Figure 4: Generated cluster map of Toronto neighborhoods

.

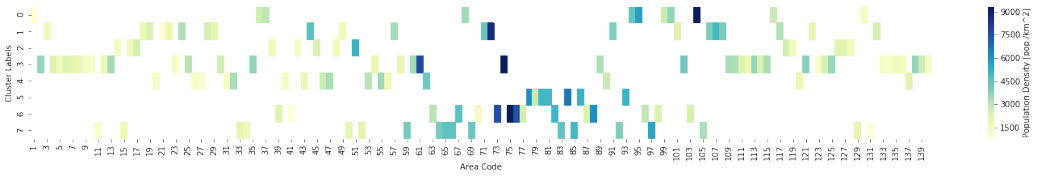

Figure 5: *'DNA'*-sequence of Toronto city neighbourhoods.

4 Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Let me briefly recapitulate about the facts that bring us until here.

- 1. As show in the Introduction, there's expected a sustained growth in the urban areas worldwide in the following years.
- 2. From a business perspective, which venues will be needed to allocate this growth in density population?

Let me start saying that I believe that neighborhoods belong to their neighbors, as such a business is successful if its neighborhood makes it. Behind

this rationale, can neighborhoods be shaped and understood to predict their behaviour? Figure 5 seems that is possible, at least, to get a glimpse of what sort of neighborhood is.

To go further in the analysis, let's plot an histogram (Figure 6) of the neighborhoods and the assigned cluster, or let me called it *behavioural* neighborhood.

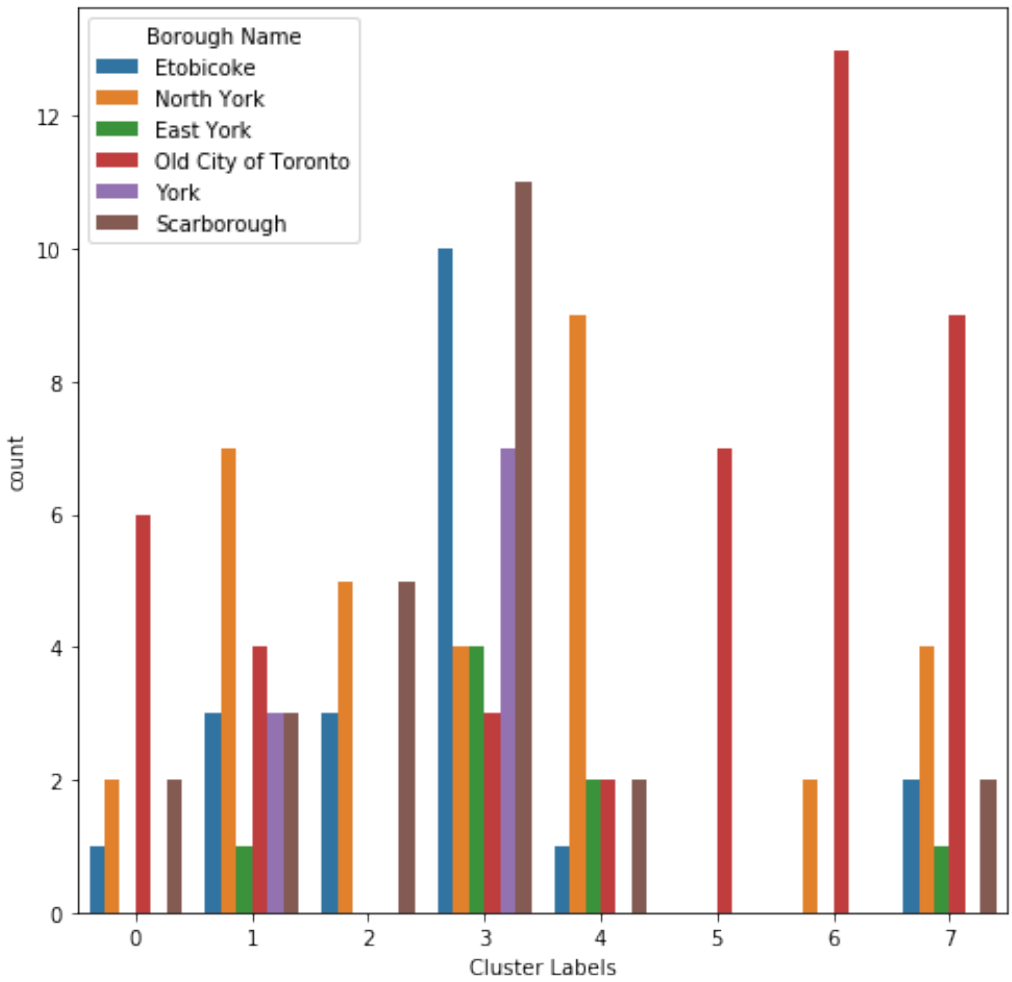

Figure 6: Histogram representing number of hits per borough and their *behavioural* neighborhood.

As a final conclusion, shown in Table 2, two neighborhoods from Cluster=1. With the information above, if there's an increase of the population of *neighborhood=3*, for 6 7% in the next 5 years, will make sense to open a second Gym?. Or what about a fifth clothing store? If these neighborhoods

behave similarly, it's expected similar people settling in to get similar results.

Area Code:	3	18
Population:	10360	11463
1st Most Common Venue:	[Restaurants, 8.0, 0.381]	[Restaurants, 8.0, 0.222]
2nd Most Common Venue:	[Stores, 4.0, 0.19]	[Shops, 6.0, 0.167]
3rd Most Common Venue:	[Shops, 3.0, 0.143]	[Coffee Shops, 5.0, 0.139]
4th Most Common Venue:	[Pharmacies, 2.0, 0.095]	[Fast Food, 4.0, 0.111]
5th Most Common Venue:	[Banks, 2.0, 0.095]	[Parks, 3.0, 0.083]
6th Most Common Venue:	[Sport & Leisure, 1.0, 0.048]	[Stores, 3.0, 0.083]
7th Most Common Venue:	[Fast Food, 1.0, 0.048]	[Sport & Leisure, 2.0, 0.056]

Table 2: Depicted neighborhoods from Cluster=1; in brackets: venue, number of ocurrences and venue percentile.

5 Conclusion

I am convinced that people bring up and shape the neighborhood they live in. As soon as I was ddoing this research project, I realized how important is to work with accurate data. As long as all this report is heavily based on Foursquare, to get a premium account and repeat the same exercise. Especially the distance value as does not seem to be accurate enough in the free/personal account. With that being said, I don't live in Toronto so not really sure if these results makes any sense. At the same time, I'd prefer not to check to much over the internet to keep me neutral. Further analysis could include weather information, economy,... all in all to keep shaping neighborhoods. I'm not sure if this is the best way, but for sure there's one.

References

[1] Migiro, Geoffrey . "The World's Most Densely Populated Cities." WorldAtlas, Nov. 15, 2018, worldatlas.com/articles/the-world-s-most-densely-populated-cities.html.

[2] United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, custom data acquired via website.