IBM Data Science Professional Certificate

Capstone Project: The Battle of Neighborhoods

Author: Sebastiano Fazzino

August 13th 2020

Introduction

My name is Sebastiano, I'm Italian and I'm 28. I like calling myself a 'World Citizen', as I've been living in four different countries so far, and I'm about to move to a fifth one. I've mostly been working in the customer service industry but I've always been interested in IT, I'm actually an entry level self-taught programmer and my goal for late 2020/beginning of 2021 is to start working in the IT industry at a professional level.

The project.
Although data scientists' work is usually focused on finding solutions to someone else's problems, in this project I'll mostly focus on myself and I'll try to give a proper answer to my question. As I mentioned before, I'm about to move in another country, and this will be Estonia, to be more specific I'll move to Tallinn which is the capital of Estonia. I've been doing some searches about Estonia and I found out so many interesting facts, like that it only has 1.3 million habitants, it is a very digital-oriented country, it is very green, as over 50% of its territory is covered by forests and it has the most start-ups per capita in Europe, even though it is one of the smallest country in the world.

Let's talk about Tallinn.
Tallinn is the capital of Estonia and yet it only has a population of about 430.000. It is located in the northern part of the country and its Old Town is one of the best preserved medieval cities in Europe and is listed as a UNESCO World Heritage Site.

Where in Tallinn should I live?
In this project I'll try to find an answer to my question using the support of data science.
To do so I'll analyze the neighborhoods in Tallinn considering different parameters:

- distance from the city center;
- public transportation;
- shops and facilities in the neighborhood (shopping malls, gym, cafes, restaurants);
- I'd rather living in a neighborhoods with a lot of green and nature;
- number of people living in the neighborhood (I'd rather living in a tranquil neighborhood with no much traffic!

The Data

Data Collection

To realize this project I've extracted information from this Wikipedia page: https://en.wikipedia.org/wiki/Tallinn, this statistics website: https://www.stat.ee/stat-unemployment-rate, Google Maps to help me finding the coordinates for each neighborhood in Tallinn and Foursquare dataset to retrieve information about venues and point of interest in the different neighborhoods. I've also gathered some extra general information from the internet.

Data Preparation

The first step has been analyzing Estonia's unemployment rate and tendency. I've extracted the data needed from the previous mentioned webpage using Pandas pd.read_html function. I've transposed the dataset and I've inserted the columns 'Year' and 'unemployment_rate'.
This is the output:

	Year	unemployment_rate
0	2007	4.6
1	2008	5.5
2	2009	13.5
3	2010	16.7
4	2011	12.3
5	2012	10.0
6	2013	8.6
7	2014	7.4
8	2015	6.2
9	2016	6.8
10	2017	5.8
11	2018	5.4
12	2019	4.4

In the next step I started analyzing Tallinn's neighborhoods. I found the coordinates for each neighborhood using Google Maps and I've created a .csv file containing name of

neighborhood, latitude and longitude. Since Tallinn only has 8 neighborhoods, I've decided to include in the project two suburbs which are relatively close to the city center.
This is the output:

	Neighborhood	Latitude	Longitude
0	Pirita	59.4749	24.8725
1	Pohja	59.4550	24.6894
2	Nomme	59.3807	24.6995
3	Mustamae	59.4010	24.6945
4	Kristiine	59.4164	24.7100
5	Haabersti	59.4267	24.6313
6	Lasnamae	59.4293	24.8352
7	Kesklinn	59.4328	24.7629
8	Viimsi	59.4994	24.8419
9	Peetri	59.3948	24.8118

After that I've used Foursquare dataset to find venues and point of interest around each neighborhood, with a radius of 1 km. I've used the **'get_dummies'** method to convert categorical variables into dummy variables and stored the information in a new dataframe.
After that, grouped by neighborhood and by taking the mean of the frequency of occurrence of each category, I've created a new table with the top 10 venues per neighborhood:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Haabersti	Gym / Fitness Center	Supermarket	Bus Line	Trail	Convenience Store	Bus Station	Eastern European Restaurant	Pharmacy	Pet Store	Bus Stop
1	Kesklinn	Cocktail Bar	Restaurant	Hotel	Café	Coffee Shop	Italian Restaurant	Cosmetics Shop	Gym / Fitness Center	Grocery Store	Gourmet Shop
2	Kristiine	Bistro	Diner	Middle Eastern Restaurant	Burger Joint	Eastern European Restaurant	Clothing Store	Massage Studio	Cupcake Shop	Pizza Place	Playground
3	Lasnamae	Furniture / Home Store	Supermarket	Park	Bowling Alley	Gym	Food & Drink Shop	Food	Fast Food Restaurant	Hotel	Japanese Restaurant
4	Mustamae	Grocery Store	Pizza Place	Convenience Store	Pub	Electronics Store	BBQ Joint	Supermarket	Park	Movie Theater	Bus Station
5	Nomme	Park	Bus Stop	Flower Shop	Restaurant	Train Station	Soccer Field	Diner	Hotel	Coffee Shop	Museum
6	Peetri	Auto Garage	Eastern European Restaurant	Coffee Shop	Paper / Office Supplies Store	Spa	Caucasian Restaurant	Park	Pet Store	Pharmacy	Pizza Place
7	Pirita	Scenic Lookout	Restaurant	Bus Stop	Smoke Shop	River	Botanical Garden	Trail	Garden	Cosmetics Shop	Creperie
8	Pohja	Supermarket	Bus Station	Park	Steakhouse	Bus Stop	Light Rail Station	Boat or Ferry	Liquor Store	Shopping Mall	Basketball Court
9	Viimsi	Supermarket	Café	Bus Station	Italian Restaurant	Burger Joint	Fast Food Restaurant	Soccer Field	Flower Shop	Convenience Store	Bar

Using the data in this table, I proceeded clustering the neighborhoods according on their similarity.

In the last step of the project I've focused on population density per neighborhood. In this step I've obtained the data from Wikipedia. As mention in the introduction, Tallinn only has 8 neighborhoods, so I've decided to include two suburbs in the project: 'Viimsi' and 'Peetri', as they are relatively close to the city centre. Unfortunately, the data for these two suburbs is not present in the table I've imported from Wikipedia, so I have to insert it manually. After cleaning the data, it shows like this:

	District	Population	Area in Km2	Density Population/sq Km
0	Haabersti	45339	22.26	2036.79
1	Kesklinn	63406	30.56	2074.80
2	Kristiine	33202	7.84	4234.95
3	Lasnamae	119542	27.47	4351.73
4	Mustamae	68211	8.09	8431.52
5	Nomme	39540	29.17	1355.50
6	Pirita	18606	18.73	993.38
7	Pohja	60203	15.90	3786.35
8	Peetri	5530	4.60	1202.17
9	Viimsi	2341	3.20	731.56

To conclude, I've plotted this dataframe using a 'pie chart' to show the distribution of Tallinn's population for each neighborhood and I've used a 'horizontal bar chart' to show Tallinn's neighborhoods population density.

Data Analysis and visualization

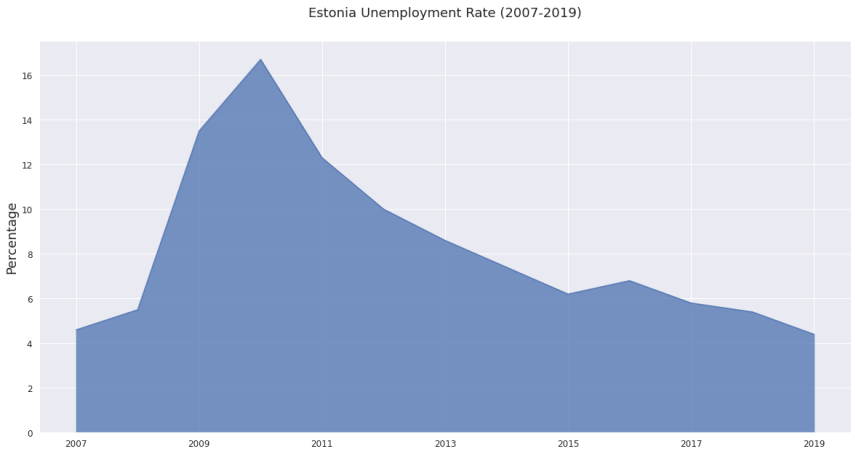
Analyzing 'unemployment_data' dataset we can notice that from 2010 the unemployment rate has decreased significantly. Doing some searches, I found out that on July the 13h 2010 Estonia officially became a European Union member and from December the 31st of the same year, Euro became the new national currency, so I believe that this event and the unemployment rate are tightly related.

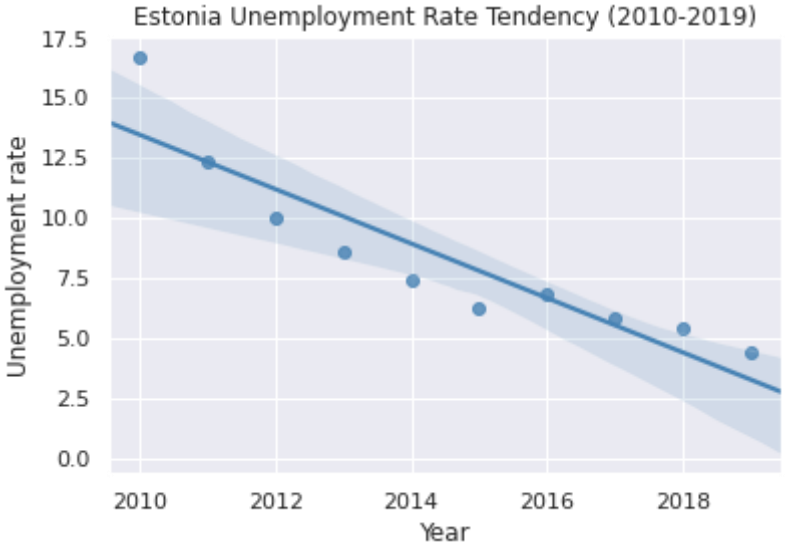
I've decided to plot the data using an area plot and then a regression plot for the data from 2010 to 2019. In the regression plot we can notice that the unemployment rate decreases going ahead with the years from 2010.

We can see that the unemployment rate in Estonia was around 4.5 on 2019 and if it continues following this tendency, should be ever lower on 2020.

This is a very positive and encouraging insight, as one of my priority once in Tallinn will be to find a job asap.

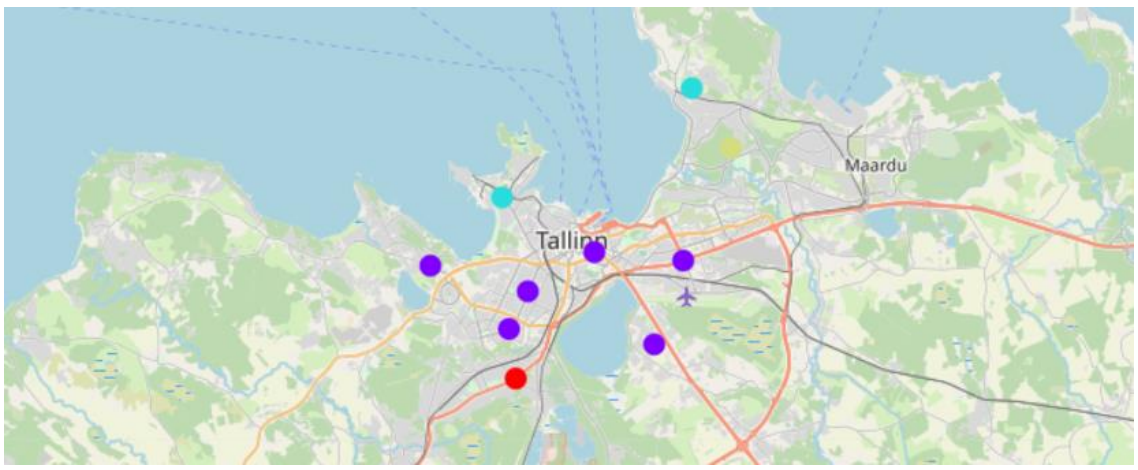
Following we can see the area chart and the regression plot that I've previously mentioned.





In the next step I'll be analyzing the neighborhoods grouped into clusters.

Here's the map of Tallinn with its neighborhoods represented by the colored markers. Every color represents a different cluster.

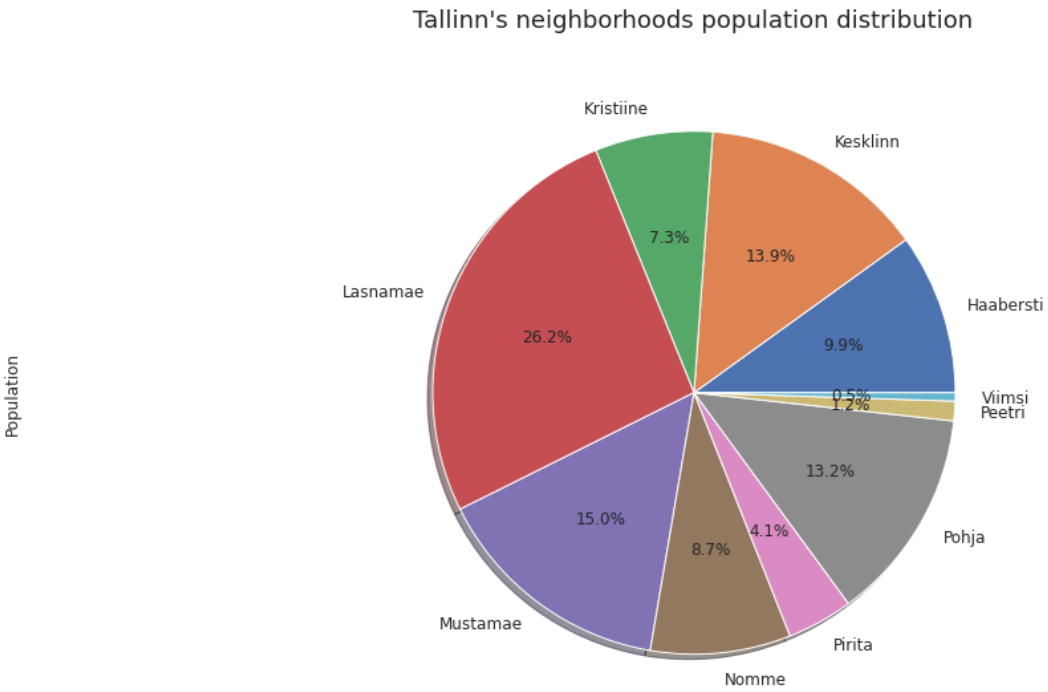


We can notice right away that Pirita and Nomme neighborhoods are the only element in their respective cluster, this makes me understand that they probably don't share much similarities with the other neighborhoods.

Meanwhile Viimsi and Pohja suburbs belong to the same cluster, and it does make sense in a way, since they're both located close to the sea and they have pretty much the same distance from the city center.

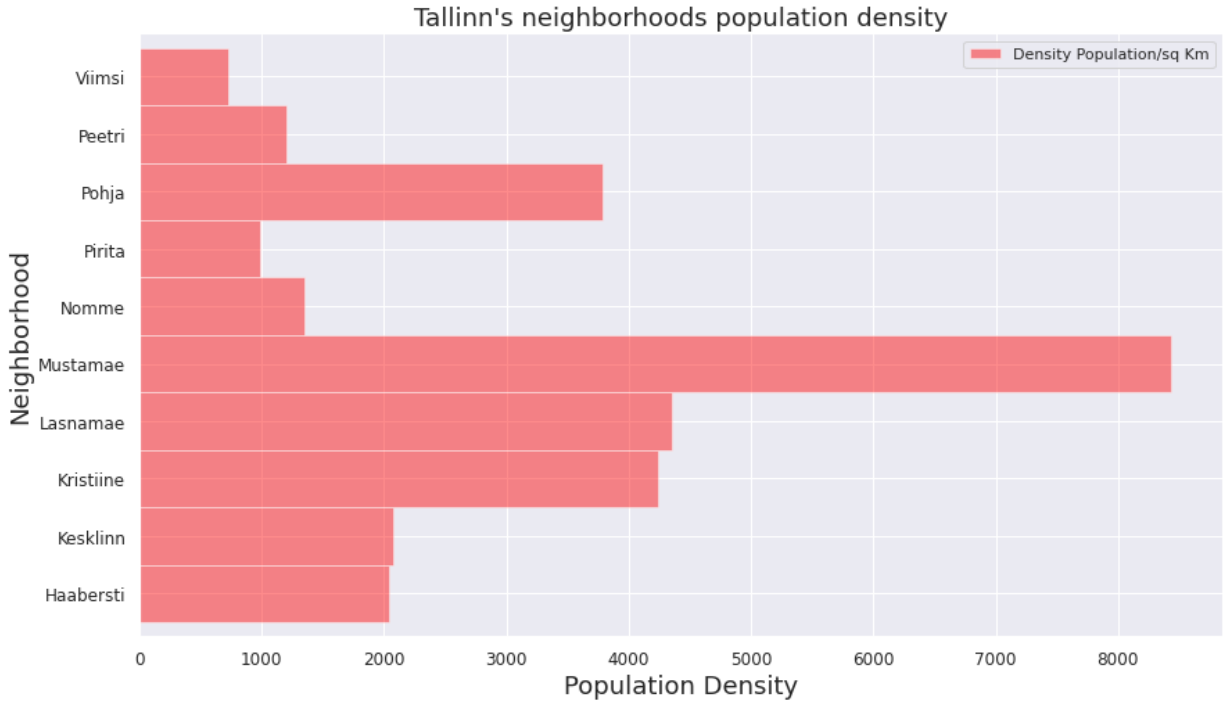
The last cluster, having six elements in it is the biggest. It includes Haabersti, Kesklinn, Kriistine, Lasnamae, Mustamae and Peetri suburbs, which are all situated more or less in the same area.

In the last section of the project I've focused on Tallinn's population distribution and population density per neighborhood.



In this pie chart we can see that 68.3% of Tallinn population lives in 4 of its districts: Mustamae, Lasnamae, Kesklinn and Pohja.

Let's further analyze the population density per km2 in the next chart:



From this bar chart we understand that Mustamae district has the highest population density, followed by Lasnamae, Kriistine and Pohja.
All the other districts seem to be definitely less crowded, having a so much lower population density per km2.

Some extra information

Browsing on the internet, I found out some more useful information about Tallinn, such as:

- living in a Tallinn suburb so much cheaper compared to the city center or a district in proximity of the city center, we're talking of a difference of about 150€ per month for the same kind of accommodation;
- living in the suburb would mean travel a lot with public transportation, but luckily a monthly plan cost around 30€, which is totally reasonable;
- the criminality rate in Estonia, especially in Tallinn is extremely low;
- despite its size, Estonia is one of the world's most technologically advanced countries, which makes it the perfect place where to start a career in IT.

Discussion

Working on this project I had the chance to better know Estonia, the capital, Tallinn and its districts. I have now a better understanding of the population distribution in Tallinn, where every neighborhood is situated and the distance from the city center, what the most common venues and points of interest for each neighborhood are.

I understood that Estonia is a very technology-oriented country, with a very high rate of start-ups per capita and its unemployment rate is very low. All these factors make Estonia one of the best country in Europe where to work as IT professional.

Conclusion

I conclude my project saying that the most suitable neighborhood for me to live in, according to my criteria would be Pirita for the following reasons:

- it only has a population density of around 1000 per km2;
- it is relatively close to the city center but still far enough for be cheaper compared to the central districts;
- it is close to the sea, and specifically 'Pirita Rand' is one of the longest beaches in Tallinn;
- in Pirita's most common venues and point of interest table we can see that it seems to be a very green neighborhood, which is a great plus for me!

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Pirita	Scenic Lookout	Restaurant	Bus Stop	Smoke Shop	River	Botanical Garden	Trail	Garden	Cosmetics Shop	Creperie

Observations

In this study I've been able to get some new knowledge about Estonia, Tallinn and its neighborhoods, and I could also practically work hands on project to put on practice what I've learnt in these nine courses of IBM Data Science Professional Certificate.

It would be interesting to continue working on this project possibly with more data regarding real estate prices in Estonia, or average monthly salary compared to monthly spending or furthermore Estonia's new technologies and tendencies.