**Comparing the neighborhoods in Saint Petersburg and Berlin**

**Nikita Gorynin**

**June 16, 2020**

1. **Introduction**

**Business Problem**

People, choosing to live in a different city may want to live in a neighbourhood somewhat familiar to one they currently live in in terms of venues they have around - someone who enjoys living in a quite place with a park and sport complex nearby might want to live in a neighborhood with a similar vibe, yet in another city or even a country.

Thus, it will be nice to have a possibility to compare the neighborhoods of two cities - service, providing such functionality, might be included in a bigger mapping/navigation platform or exist independantly. The present notebook is an example of comparison between the neighborhoods of two cities, namely Saint Petersburg and Berlin - both have a status of cultural and educational capitals of their countries and are similar in metropolitan population and average population density (~6 200 000 inh. and ~4 000 inh./km^2 respectively)

1. **Data**

**Data Sources:**

1. Borough, name, population and area of each neighborhood for both cities, taken from Wikipedia:
   * For Saint Petersburg:[Wiki page for Saint Petersburg](https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D0%BC%D0%B8%D0%BD%D0%B8%D1%81%D1%82%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE-%D1%82%D0%B5%D1%80%D1%80%D0%B8%D1%82%D0%BE%D1%80%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%A1%D0%B0%D0%BD%D0%BA%D1%82-%D0%9F%D0%B5%D1%82%D0%B5%D1%80%D0%B1%D1%83%D1%80%D0%B3%D0%B0)
   * For Berlin:[Wiki page for Berlin](https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin)
2. ArcGis coordinates for points in each neighborhood
3. OpenStreetMap data on borough and city boundaries
4. Foursquare data on venues in each neighborhood, gathered using Foursquare API
   1. **Wikipedia page parsing:**
5. Raw tables from Wikipedia were parsed and transformed to a Pandas dataframe – only Name, Area, Population, and Borough name were kept for both cities:



Figure 1. Raw data table for Saint Petersburg

Neighborhood

Area

Population

Borough

All the unnenessary markers, such as arrows and reference numbers were removed and Density values were calculated to get a Pandas dataframe of the following look:



Figure 2. Cleaned Saint Petersburg Neighborhoods dtaframe

* 1. **Neighborhood coordinates**

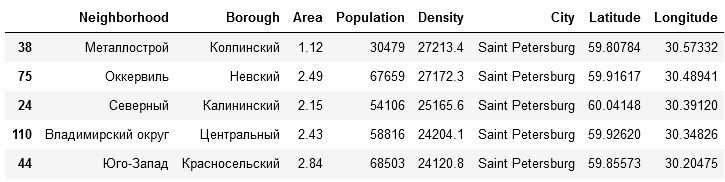
Coordinates of point within each neighborhood were obtained, using the ArcGis API

Figure 3. Saint Petersburg dataframe with neighborhood coordinates

Saint Persburg and Berlin dataframes were than concatenated, to get a full list of neighborhoods in both cities

* 1. **Borough boundaries**

To create a chloropleth map, data on borough boundaries was taken from OpenStreetMaps – Geojson file of separate boroughs for Saint Petersburg and the entire city for Berlin. Dataframe, storing the borough name and density was also created for each city

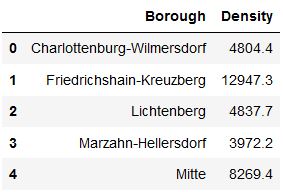


Figure 4. Berlin borough dataframe

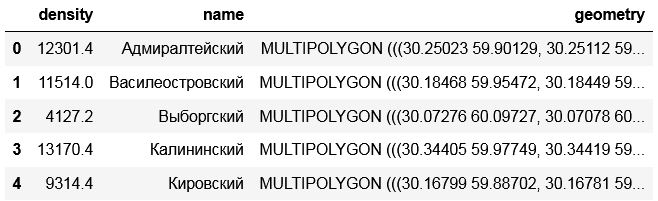
To be able to make an interactive borough layer, additional dataframe, storing the geometry of each borough was also created for both cities:

Figure 5.Interactive layer dataframe for Saint Petersburg

* 1. **Foursquare data**

For each neighborhood, list of venue within 500m radius was obtained using Foursquare API. Foursquare provides number of properties to get, but only venue category was kept. Then the number of venues of each category was counted and a dataframe of top-10 most popular venues for each neighborhood was created



Figure 6. Dataframe with most popular venue types in each neighborhood

1. **Methodology**

Having a geographical, population and venue data, we want to give an advice to a person, willing to move to a similar neighborhood of one city to another. To do that, we will compare the neighborhoods and get a list of most suitable neighborhoods at the end. The entire analysis will be done in several steps:

1. Collecting the data- we already got the list of the neighborhoods in each city with borough names, areas, total population, density and coordinates of each neighborhood. We also colected the geographical boundaries for each borough and got the venue data for each neighborhood using Foursquare API
2. Comparing the neighborhood population density in two cities- we will draw a boxplot and see how the density distributes across the boroughs in each city
3. Comparing the density distribution across the city's area- we will make a map to see where the most populated boroughs reside in each city
4. Clusterisation of neighborhoods- we will apply the Kmeans algorythm to cluster the neighborhoods of both cities into clusters and then show it on a map
5. Combining the density comparison with clusterization results- we will see, how many neighborhoods of each city fall into one cluster with neighborhoods of another city, estimate the number of inhabitants,who live in those neighborhoods and built a final plot, which show appropriate neighborhoods

We have all the data we need for our neighborhoods, let's see some basic statistics on it

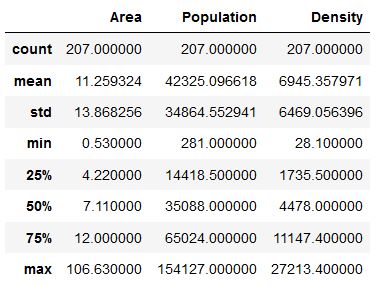


Figure 7. neighborhood list description

There is a huge difference in size and density between the neighborhoods, now let's visualize the data to see which city's neighborhoods are more densely populated As we have a lot of neighborhoods to plot, the barplot of the entire dataframe would very big, so let's draw a boxplot for each borough to see the distribution of neighborhood density in each

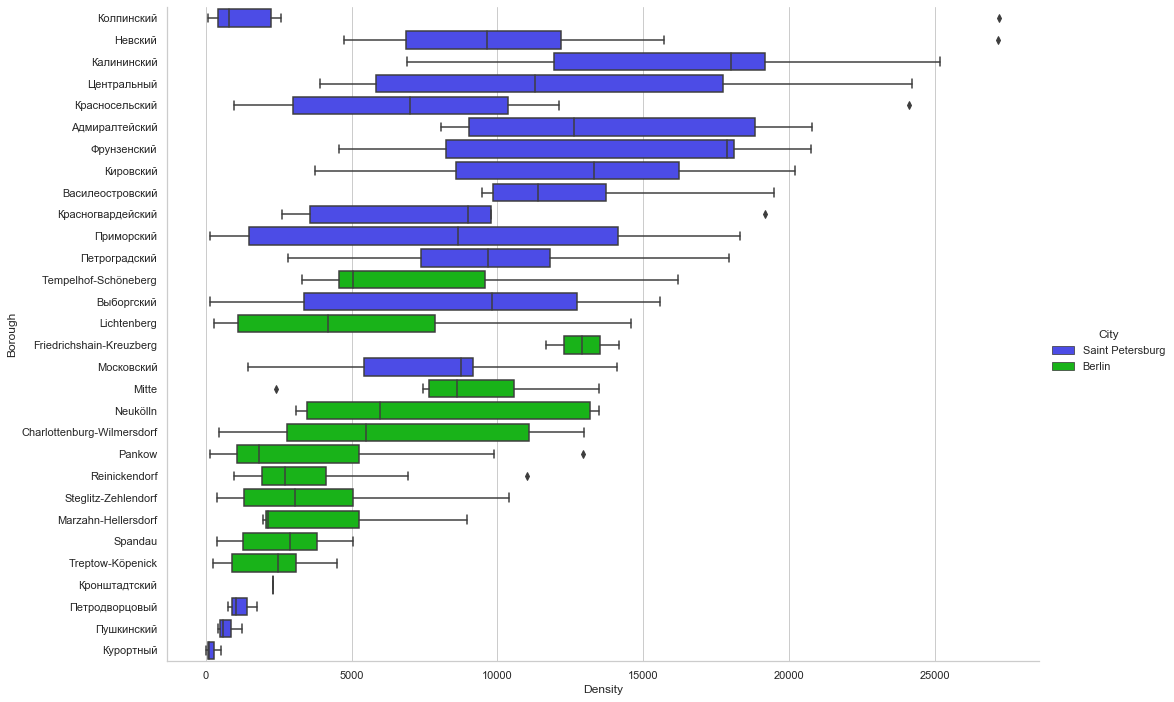


Figure 8. Borough density distribution

Clearly, Saint Peterburg boroughs are mostly more dense than Berlins, yet Saint Peterburg also has four of the least dense boroughs.

So far so good, but it will be even more convinient to show this data on the actual map - the heatmap will be a good representation for density, so let's colour the map according to the population density in each area. As we only have geometry for boroughs, we will use borough densities for coloring the map and add neighborhood densities to the marker labels.

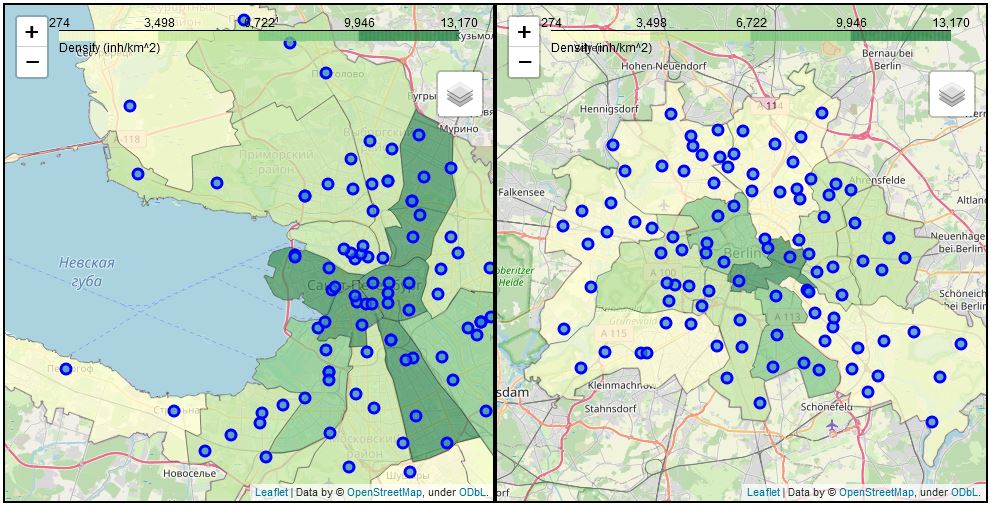


Figure 9. density distribution across the areas of both cities

Not surprisingly, the closer to the city center, the more densely populated the boroughs are, with same logic generally applied to neighborhoods.

We can also see that for the most part, the entire city of Berlin is closer to the Saint Petersburg suburbs in terms of population, with only the Friedrichshain-Kreuzberg being an exception

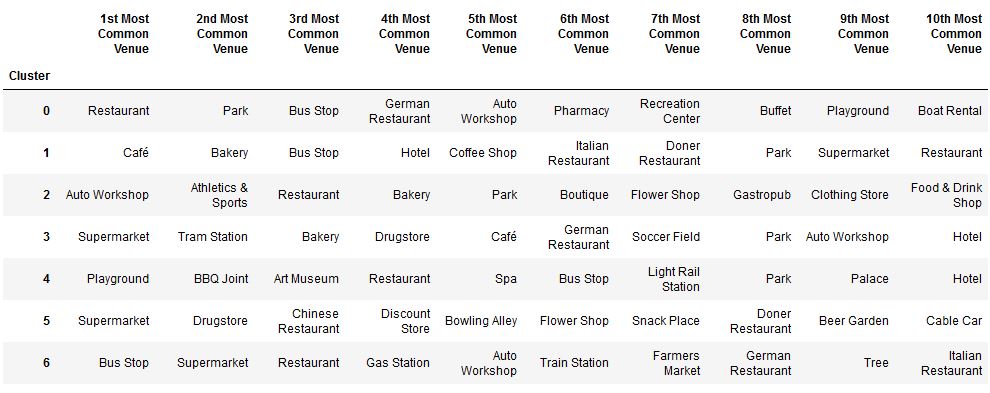
Let's move forward and cluster the neighborhoods by their venues:

Figure 10. Clustered neighborhoods



We succesfully diveded the neighborhoods into clusters, based on the venues they have, let's see what the average neighborhood of each cluster looks like

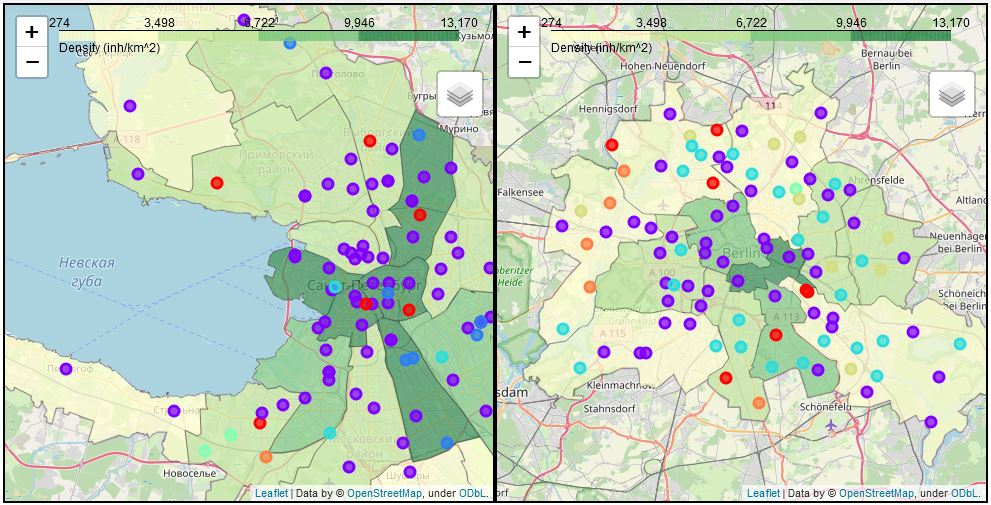
Figure 11. Average neighborhood of each cluster



It's fair to say that clusters 0 and 5 are more of a suburban type, while clusters 1 and 6 have more of arecreational venues

Now let's see how clusters distribute geographicaly - to do that, we will recreate the maps, but this time we will paint the neighborhood markers according to the cluster number

Figure 12. Clustered neighborhoods



There is clearly a big difference in cluster distribution among the cities - some of the clusters are prevalent in Saint Petersburg, but almost absent in Berlin and vise-versa. Let's take a look on a distribution of clusters among the cities to understand it better

There is clearly a big overlap in cluster distribution - despite the majority of clusters being either unique or almost unique to one city, most of the neighborhoods fall into the overlapping clusters:

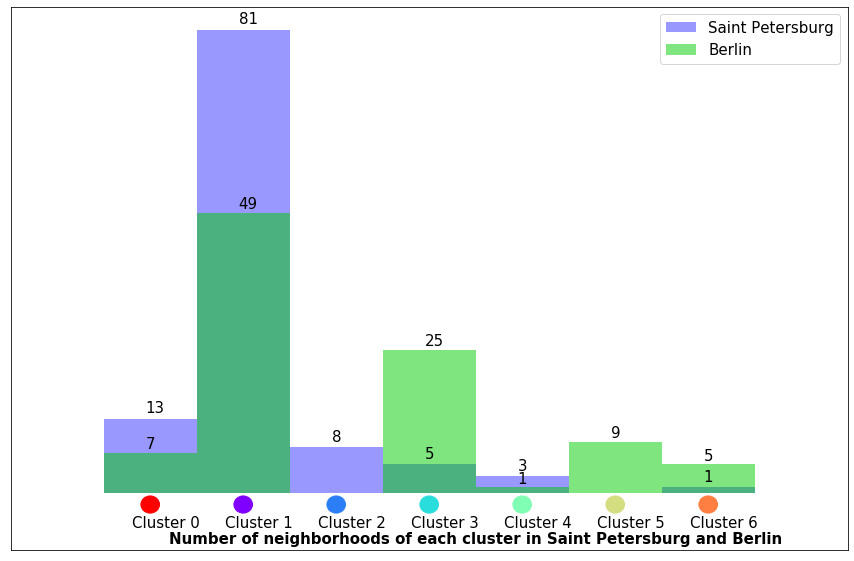


Figure 13. Number of neighborhoods of each cluster in both cities

* 2 clusters ,namely 2 and 5 are unique to it's city
* 2 clusters ,namely 3 and 6 predominantly exist in one city (16.7% overlap)
* 1 cluster (1) is the most popular cluster in both cities

But how much people live in these clusters? let's look at the population of overlapping clusters and find the percentage of people, who can move to a neighborhood of their cluster in a different city:

Initially, we were going to provide a calculated advice for person willing to move from one city to another - as we see now,there are 103 neighborhoods in Saint Petersburg and 87 neighborhoods in Berlin, inhabitants of which can expect to find a similar place in terms of venues. Let's go further and combine the data on clusters with the density comparison to get a final result.

For every overlapping cluster we will only display the neighborhoods for which it's possible the same cluster neighborhood in another city with density difference no more than 25%

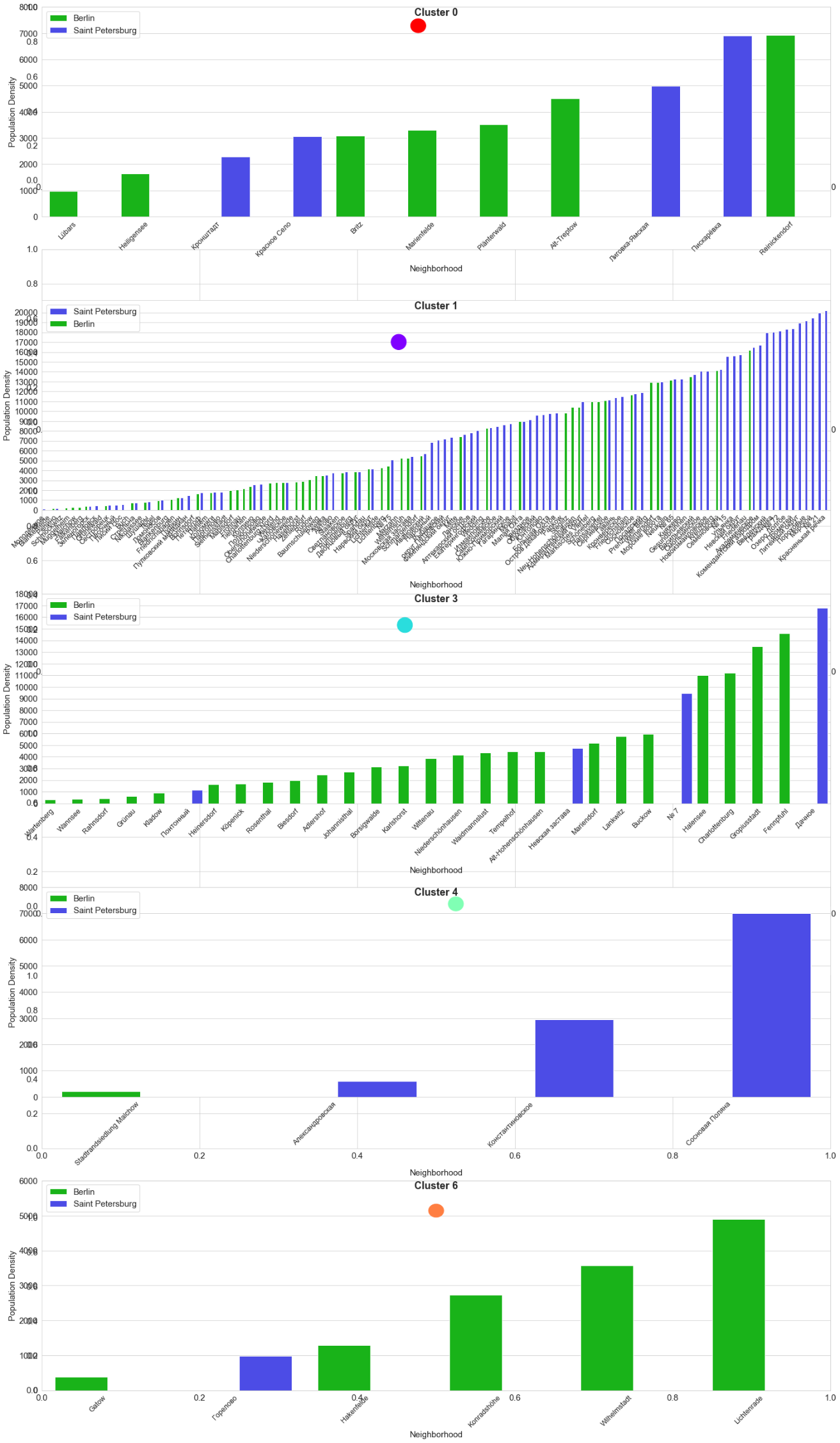


Figure 14. Neighborhoods of each cluster

1. **Results**

Analysis shows, that out of 207 neighborhoods (111 in Saint Petersburg and 96 in Berlin) there are total of 190 neighborhoods, inhabitans of which can find a neighborhood with similar vibe in a different city. In total:

* 92.56% of Saint Petersburg inhabitans can find a neighborhood of same cluster in Berlin
* 91.0% of Berlin inhabitans can find a neighborhood of same cluster in Saint Petersburg:

Interactive map was built to find the most similar neighborhoods:

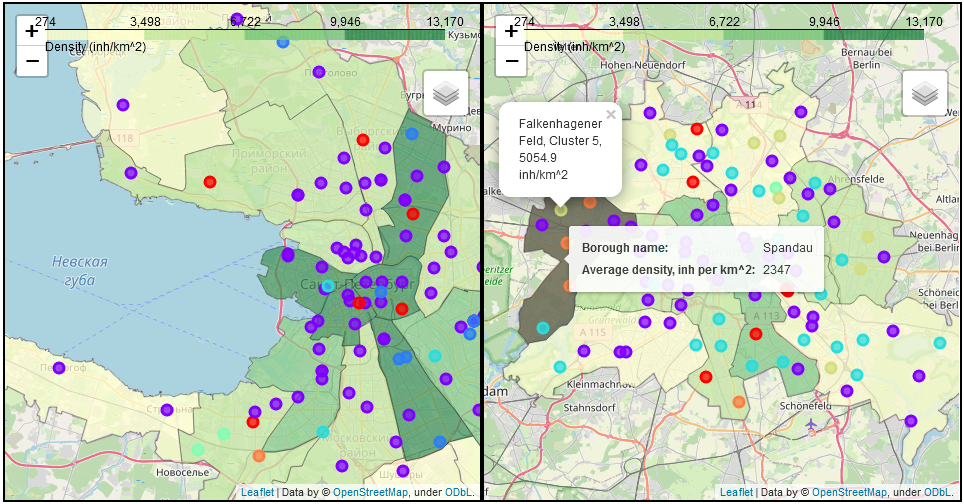


Figure 15. Interactive map, showing similar neighborhoods

For each cluster, which exists in both cities, barplot of neighborhood densities was provided

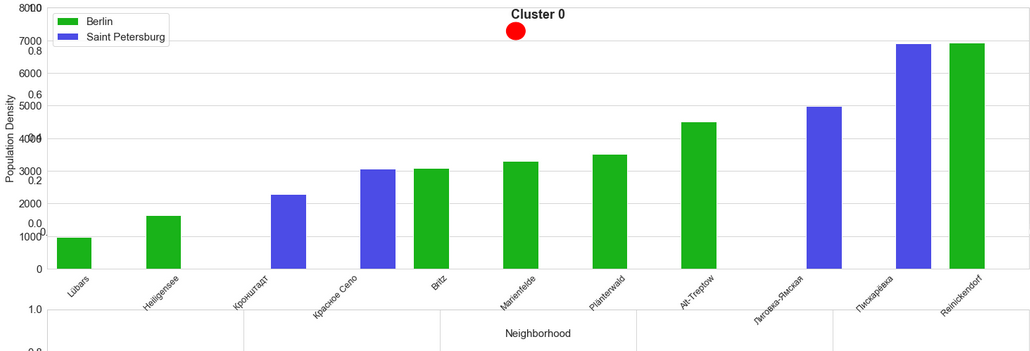


Figure 16. Most similar neighborhoods

1. **Discussion**

As there are different numbers of neighborhoods of each cluster in two cities, moving from one city to another can be easier or harder, depending on the cluster you currently reside:

* It's easier to move from Saint Petersburg to Berlin, if you live in a neighborhood of cluster 3
* It's easier to move from Berlin to Saint Petersburg, if you live in cluster 1

So, majority of people can find a place somewhat similar to one they currently live, but adding restrictions on density difference drops the number of similar neighborhoods - putting the +-25% limit on density difference on average drops the number of suitable neighborhoods to 1-10, depending on the city

1. **Conclusion**

* Advice for people willing to move to another city is provided
* Most of inhabitants in each citie can find a similar place in another one
* It’s easier to move to a similar place from some kinds of neighborhoods while it’s not possible for some of them

However:

* As venues were acquired from a circle, centered on a neighborhood, not all neighborhood venues were analysed
* As heatmap was built using borough data, instead of neighborhood, impression of the density distribution might be distorted
* Clustering results depend on number of clusters initialized, thus demanding additional analysis