Comparing Helsinki neighborhoods

Capstone project for ”Applied Data Science”

Background / Business problem

- This study tries to shed light on this question: What kind of areas there are in Helsinki, Finland, in terms of available services and key socio-economic data?
- The results of the study are interesting for example to people who are planning moving to Finland. There is a constant lack of skilled ICT professionals in Helsinki area, for example.
- The analysis will be conducted by postal code area. These divide Helsinki region to 84 different areas, on average 2.2 square kilometers.

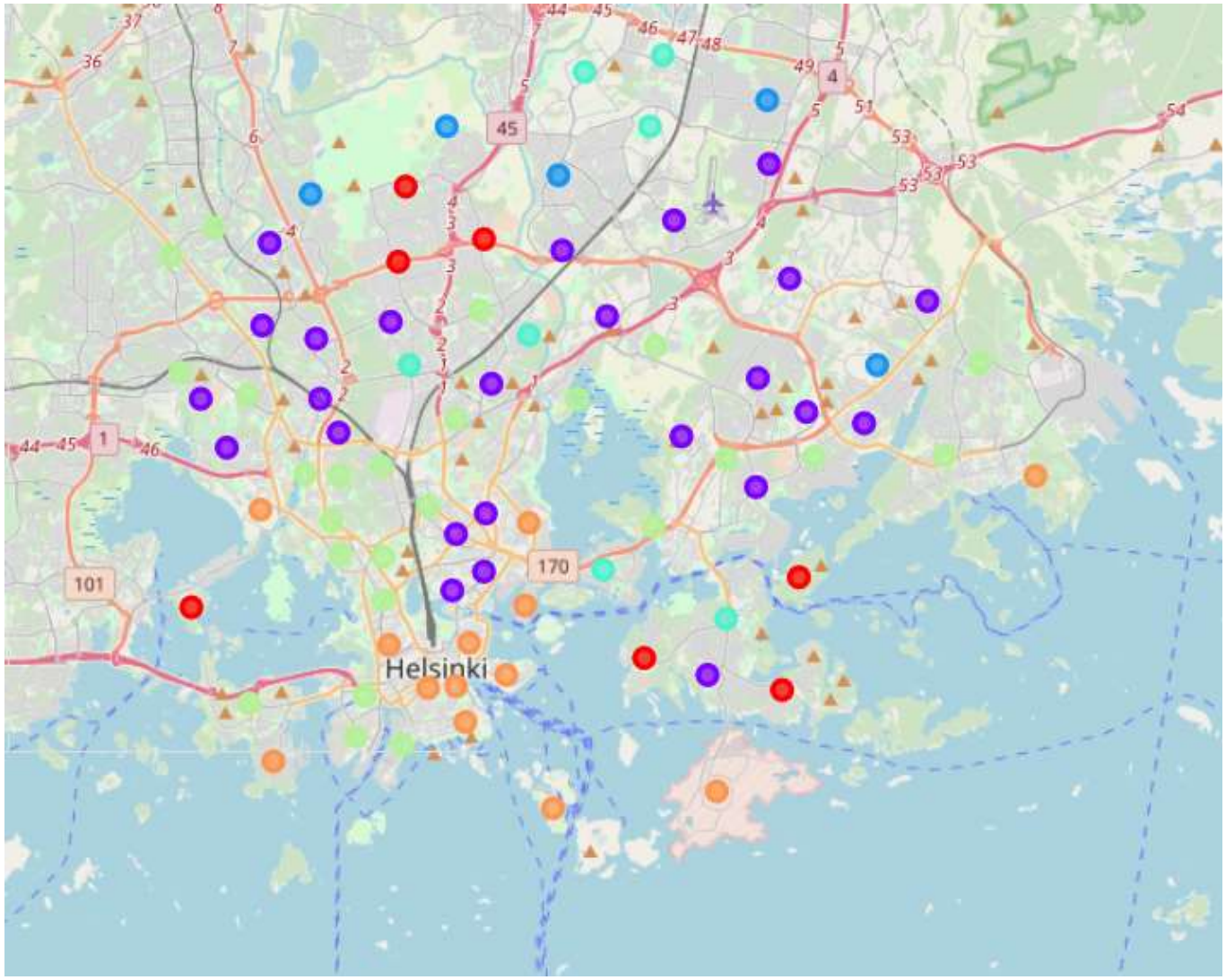
Data sources used

- Postal code data from Helsinki Region Infoshare (open data)
- Various Statistics Finland per-postal-code open data sets
  - Occupational status of residents
  - Median income of residents
  - Housing information (average size of houses/apartments, relate share of houses vs. apartments)
- Coordinates of postal codes (Bing Maps REST service, free developer account needed)
- Service offering in each area, based on Foursquare venue data (free developer account needed)

Methodology

- All handling was done in Jupyter Notebook within Watson Studio in IBM public cloud
- Various Python libraries were used in processing: Numpy, Pandas, Folium, scikit-learn, Matplotlib, geocoder
- Basically, data was fetched, processed, normalized to 0..1 and put into a ”master DataFrame” indexed by postal codes
  - Some discrepancies in data set coverage were found; 2 out of 84 postal codes were dropped, due to lack of data in some dataset
  - The end results was a 82 x 260 data matrix, 246 data points being Foursquare venue categories
- Finally, K-means clustering was done to create 6 clusters of Helsinki areas

Results 1/2



Results 2/2

- Cluster 0 (Red): Large houses, very well-to do people, not that much services
- Cluster 1 (Purple): Small apartments, medium-to-low income, basic services
- Cluster 2 (Light blue): Large houses, medium income, basic services
- Cluster 3 (Cyan): Half apartments, half houses, medium plus income, decent services
- Cluster 4 (Lime): Apartments, medium income, good services
- Cluster 5 (Orange): Large apartments, well-to-do people, good services

Discussion

- The results are credible based on everyday experience
- There are some strong correlations between the variables – we could perhaps have been OK with some less data
- Quality of Foursquare data in Helsinki – especially in the suburbs – is quite low
- Some further data could perhaps make a more fruitful analysis
  - Skip Foursquare data, replace with Statistics Finland data
  - Add some data on native languages / countries of origin of residents?

Conclusion

- The data set used did what was expected of it.
- At all stages, one has to check…
  - That the data you got was OK
  - That the data set covers all of your requirements
- When you live on the borders of Europe, you cannot trust on big US data providers to have proper data on your area!