Helsinki R&B Café

The Battle of the Neighborhoods

A Clustering Approach for Determining the Optimal New Café Location in Helsinki

Agenda

- Problem
- Assumptions
- Method
- Results
- Limitations
- Conclusion

Helsinki R&B Café

Problem

- A hypothetical client wants to open a new café (called *Helsinki R&B Café*), somewhere in Helsinki.
- If the café is successful, she would like to open additional cafés in other Helsinki neighborhoods with similar characteristics to the first.
- She has gathered a set of business insights regarding what constitutes the “optimal” new café location, based on extant research.
- However, she lacks the data and analytics to determine which neighborhoods fit the optimal location profile.
- Problem: Find the optimal new café location.

Helsinki R&B Café

Assumptions: “What Constitutes Optimal Location?”

The client has provided the following business assumptions regarding the optimal location:

Assumption	Rationale
High income among inhabitants	Helsinki R&B café will use high-end organic ingredients. So, the higher the inhabitants’ disposable income the better.
High population density	The higher the population density, the greater number of people the café can reach, ceteris paribus.
High proportion of people aged 18 to 34	The style and atmosphere of Helsinki R&B café will be tailored to fit the tastes of young people between the ages of 18 and 34.
High concentration of restaurants	Many people like to go a café either shortly before or after dining at a "sit-down" restaurant. Hence, the optimal location should have a high concentration of restaurants nearby.
Low concentration of pre-existing cafés	On the other hand, the ideal location should have a low concentration of pre-existing cafés nearby to minimize competition.

Method: Features and Data Sources

Five features were selected for clustering analysis:

- Three were **population-related**:
  1. Median income of neighborhood inhabitants
  2. Neighborhood population density
  3. Percentage of inhabitants aged 18-34
- Two were **venue-related**:
  4. Number of restaurants within a 500m radius area of the neighborhood's central coordinates
  5. Number of cafés within a 500m radius area of the neighborhood's central coordinates

Data sources:

- Population-related data was collected from *Tilastokeskus*, also known as *Statistics Finland* (link to data published in 2021).
- Venue-related data was collected using the *Foursquare* API.

Method: Analytic Approach

- Data wrangling and exploratory data analysis were completed using first *MySQL* and subsequently Python's *pandas* library.
- Clustering analysis was conducted using the *scikit-learn* library.
- The neighborhoods were segmented into *k* = 4 clusters. This number was deemed optimal based on a comparison of *inertia* and *silhouette* scores across *k* values ranging from 2 to 11.
- After clustering, the new café's projected annual revenue was estimated for each neighborhood cluster. This was done using a formula based on mathematical assumptions linking each of the features to projected revenue.
- The neighborhood cluster with the highest projected annual revenue was chosen as the optimal new café location.

Method: Computing Projected Annual Revenue in Each Cluster

The following formula was used to estimate the café's projected annual revenue in each cluster:

$$\textbf{Annual revenue} = \log_{10}(X_4 + 10) / (X_5 + 1) * [0.2\, X_2 X_3 + 0.05\, X_2 (100 - X_3)] * 0.02\, X_1$$

where X_1 through X_5 refer to (1) median income, (2) population density, (3) percentage of inhabitants aged 18-34, (4) concentration of restaurant venues, and (5) concentration of cafés.

The formula above is based on a set of mathematical assumptions linking each of the clustering features to projected revenue:

1. The size of the café's target market is more-or-less equal to X_2 (i.e., the number of neighborhood inhabitants per sq km).
2. If there are NO restaurants or pre-existing cafés nearby, the new café will attract 20% of inhabitants aged 18-34 (i.e., 20% of $X_2 X_3$) versus 5% of all other inhabitants.
3. If there ARE restaurants nearby, the probability of attracting customers, regardless of age, is enhanced by "$\log_{10}(X_4 + 10)$" times. In other words, the number of customers attracted rises logarithmically with respect to increasing concentration of restaurants.
4. If there ARE pre-existing cafés nearby, the probability of attracting customers, regardless of age, is divided by "$X_5 + 1$". In other words, the number of customers attracted falls proportionately with respect to increasing concentration of pre-existing cafés.
5. Collectively, the new café's visitors would spend 2% of their median annual income (i.e., 2% of X_1) at the café per year.

(*please refer to the full report for more details*)

Descriptive Statistics

	Median Income	Population Density	Percentage aged 18-34	Restaurant Venues	Café Venues
Mean	26,511.92	4,408.54	25.88	2.41	1.16
S.D.	3,511.10	3,502.76	7.74	4.14	1.80
Min	19,915.00	58.00	11.01	0.00	0.00
25%	24,056.00	2,194.00	21.05	0.00	0.00
50%	25,897.00	3,595.00	25.09	1.00	0.00
75%	28,691.00	5,395.00	30.91	2.00	2.00
Max	34,641.00	20,988.00	49.66	24.00	8.00

Note: Median income is in euros (€). Population density is the number of people per sq km. Restaurant and café venues are measured in terms of the number of restaurants and cafés within a 500km radius area of a neighborhood's central coordinates.

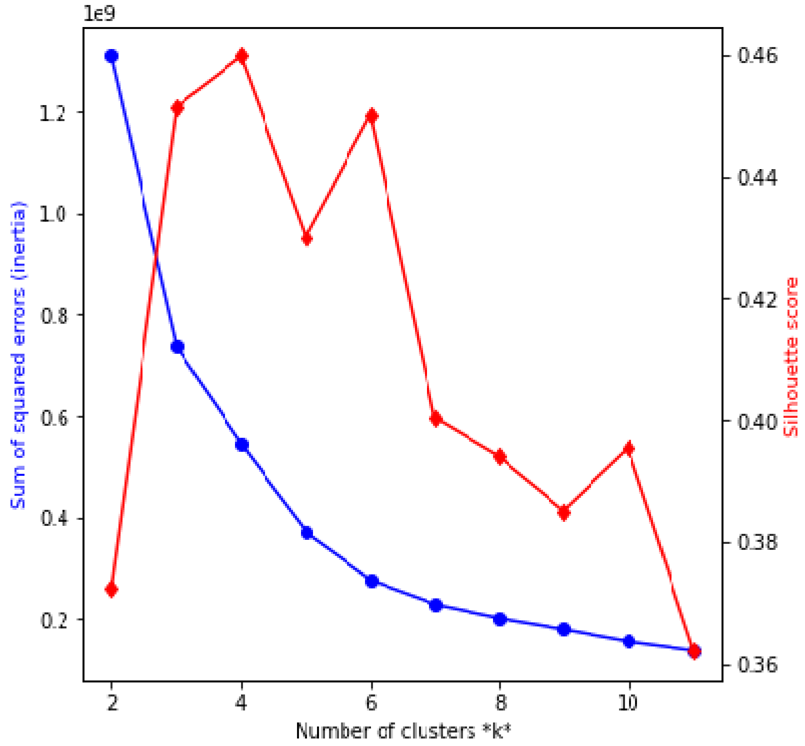
Results

Descriptive Statistics

According to Tilastokeskus's 2019 estimates, the 83 Helsinki neighborhoods had a median annual income of €26,512 and a mean population density of 4,409 inhabitants per sq km. On average, people aged 18-34 comprised 25.9% of the neighborhoods' inhabitants.

According to Foursquare data from April 2021, there was a mean of 2.41 restaurants and 1.16 cafés per 500m radius area from the neighborhoods' central coordinates.

A Comparison of Inertia and Silhouette Scores *K*-Values



Results

Optimal Number of Clusters

Prior to clustering, inertia and silhouette scores were compared across *k* values ranging from 2 to 12.

It was determined that *k* = 4 was the optimal number of clusters, as that was when the silhouette score was at its peak and was also the point at which the inertia scores began to "flatten."

When *k* = 4, the inertia and silhouette score were equal to 5.46 x 10^8 and 0.46, respectively.

Mean Feature Values and Revenue Estimates

Cluster	Median Income	Population Density	% aged 18-34	Restaurant Venues	Café Venues	Total Visitors	Annual Revenue
0	24,173.38	3,512.18	25.96	1.02	0.42	228	110,230
1	30,623.95	1,985.38	19.42	1.52	1.29	72	44,098
2	**26,626.00**	**19,117.00**	**40.22**	**14.00**	**3.50**	**646**	**344,007**
3	27,755.47	8,528.93	32.78	6.27	2.87	264	146,548

Note: “Total visitors” refers to the estimated total number of unique visitors the café will attract in one year. Annual revenue is in euros (€).

Results

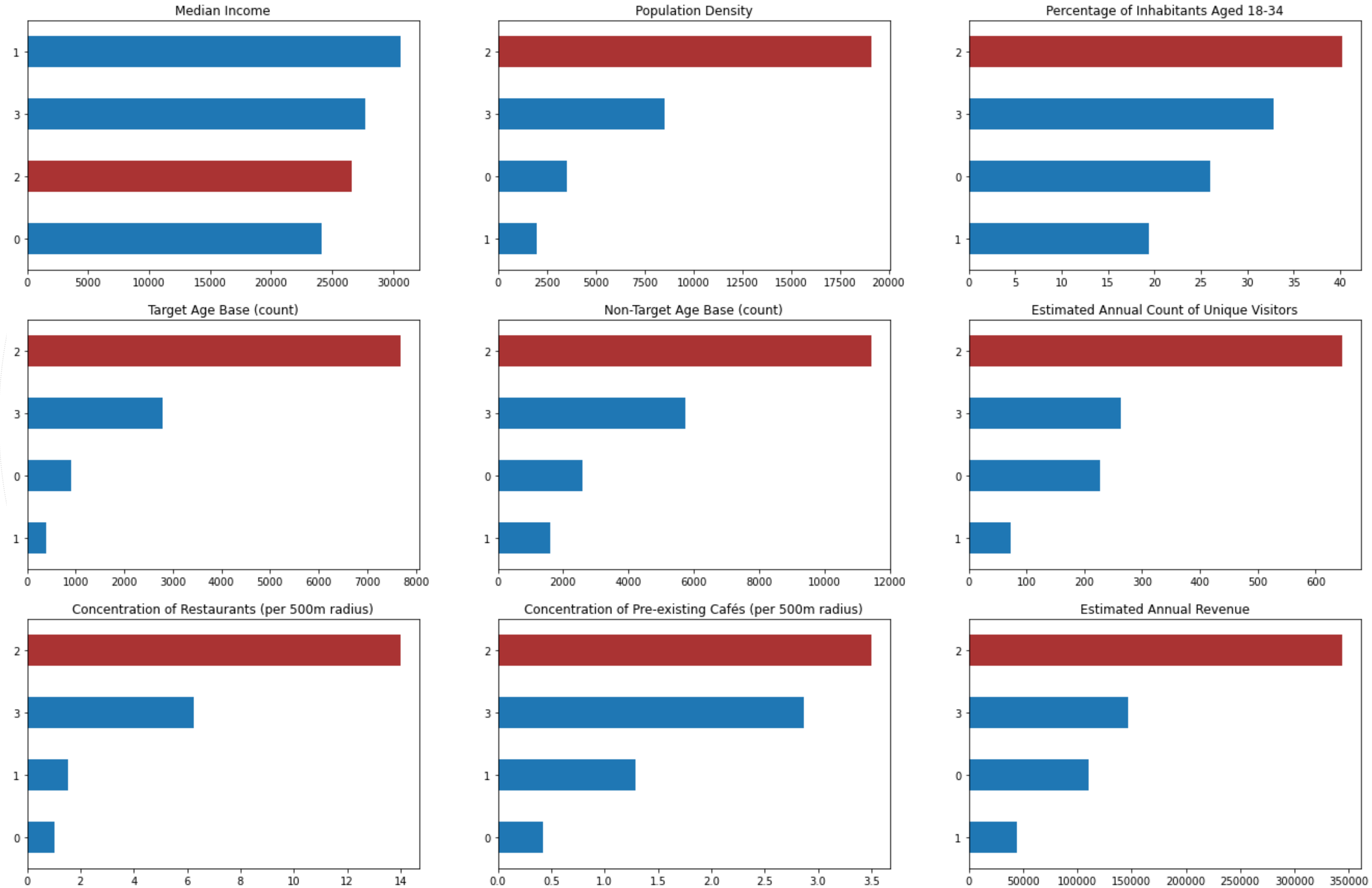
The Optimal Location

Cluster 2, consisting of *Punavuori* and *Sörnäinen,* had the highest estimated annual revenue (i.e., 344,007 €).

Between the two, *Sörnäinen* had the larger annual revenue estimate of 397,474 € (versus *Punavuori*’s 285,637 €).

Recommendation to the client: Consider opening the first Helsinki R&B Café in ***Sörnäinen***. Then, if the café is successful, consider opening an additional café in ***Punavuori***.

Results: Charts Comparing Mean Feature Values, Annual Count of Unique Visitors, and Annual Revenues Across Clusters



Cluster 2 has the following advantages:

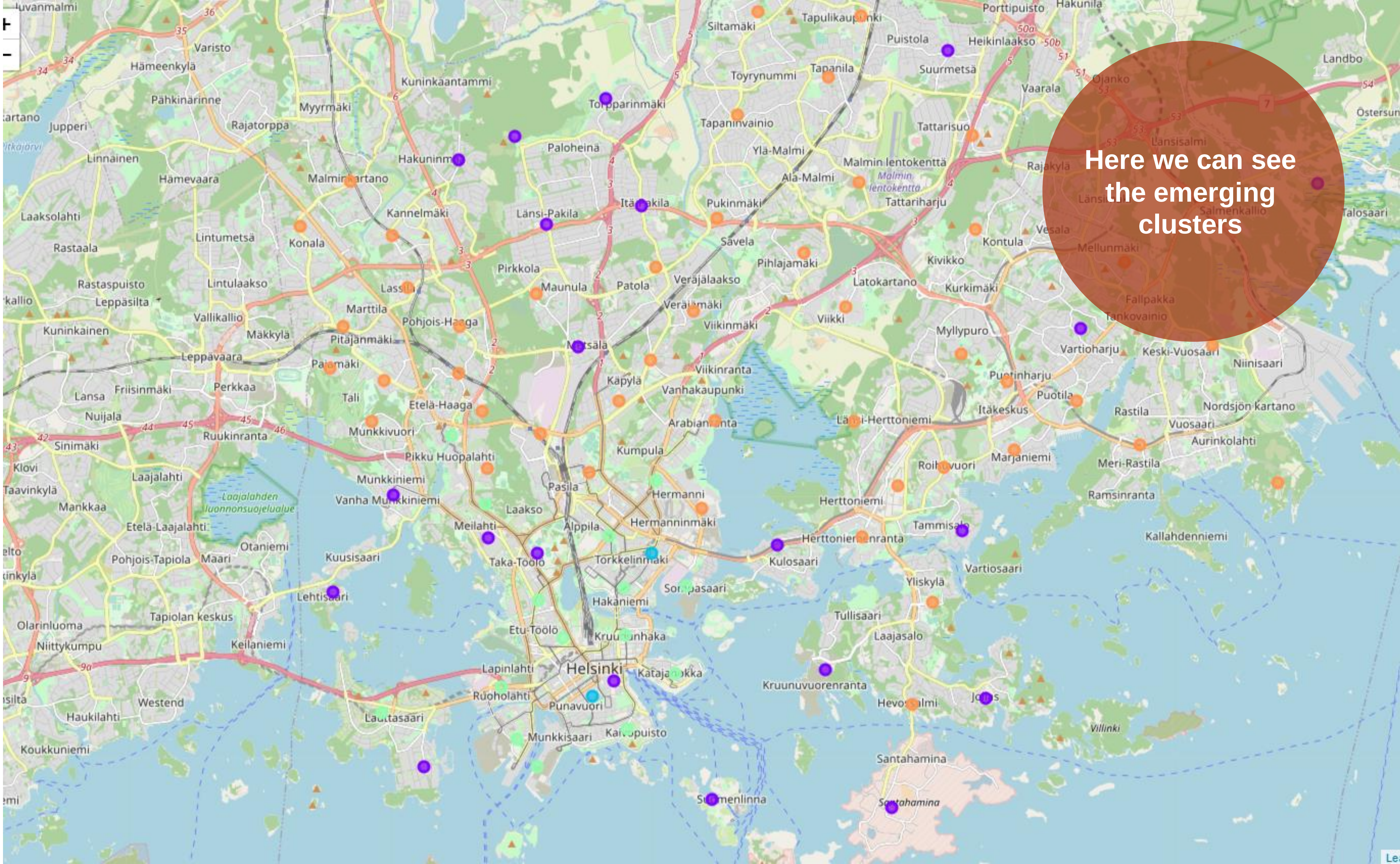
- Highest estimated annual revenue.
- Highest estimated count of annual visitors.
- Highest population density.
- Highest proportion of inhabitants aged 18-34.
- Highest concentration of nearby restaurants.

And the following disadvantages:

- It ranked 3rd (out of 4) in median income.
- It had the highest concentration of pre-existing cafés (i.e., greatest competition).

Note: “Target age base” refers to the estimated number of inhabitants aged 18-34 that the café can reach in a year. In contrast, “non-target age base” refers to the number of inhabitants *not* aged 18-34 that the café can reach.

Helsinki R&B Café

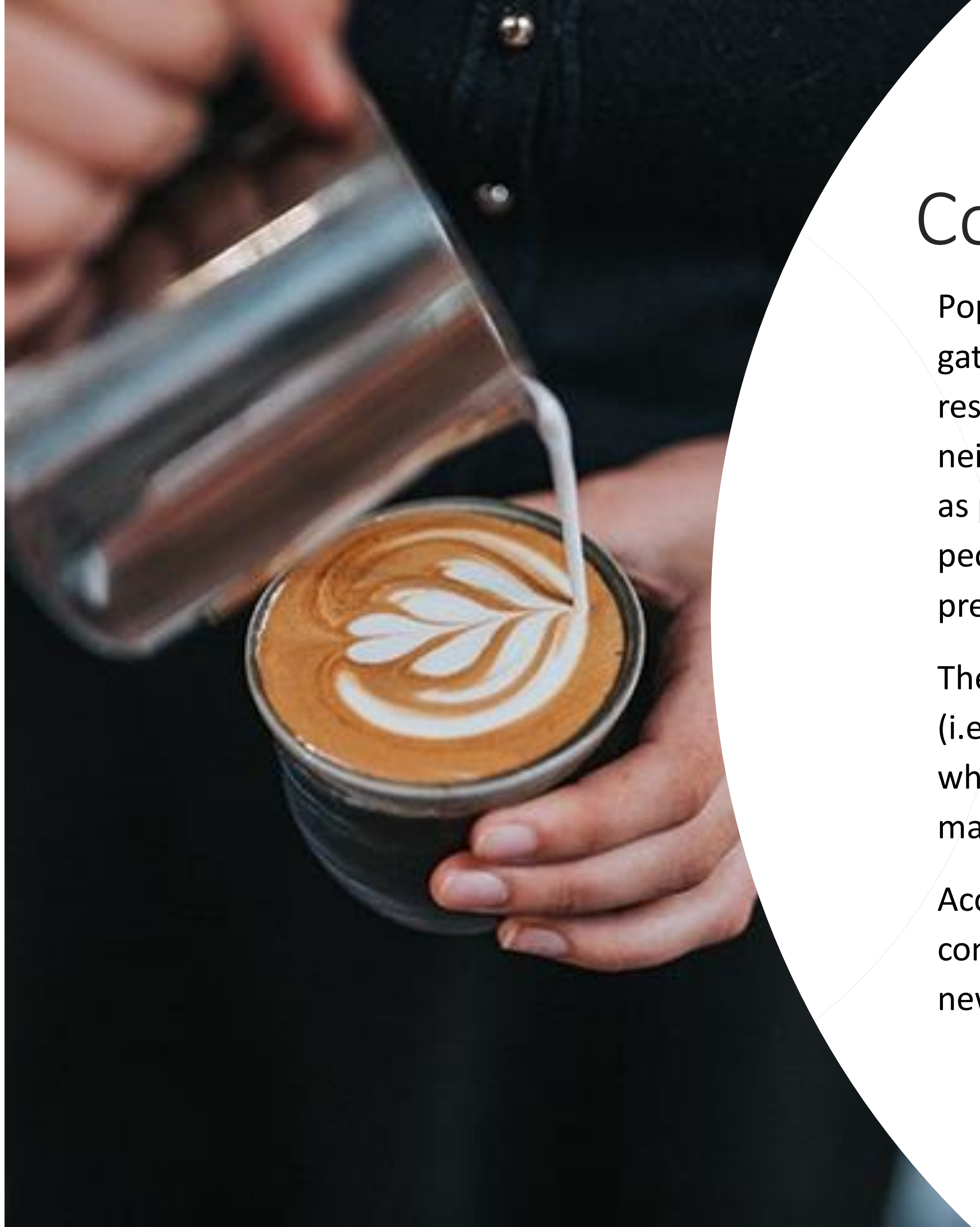

Here we can see the emerging clusters

LIMITATIONS

- The neighborhoods were clustered based on five features only.
- The concentration of pre-existing restaurants and cafés in each neighborhood was measured using proxies (i.e., the number of venues within a 500m radius area from the neighborhoods' central coordinates).
- Annual revenue was estimated using a formula based on mathematical assumptions with few variables.
- The optimal location was determined based on estimated revenue only and thus differences in cost were not considered.

FUTURE DIRECTIONS

- Consider examining additional features such as location's accessibility by public transport, presence of "R-kioski" (i.e., a convenience store chain popular for coffee), and quality of nearby restaurants and rival cafés.
- Consider sales generated not only by neighborhood inhabitants but also tourists and visitors from other neighborhoods.
- Estimate total costs in addition to revenue for determining the optimal location.



Conclusion

Population-related data and venue-related data gathered from *Tilastokeskus* and *Foursquare*, respectively, were used to segment 83 Helsinki neighborhoods into four clusters based on features such as population density, income, proportion of young people, and concentration of nearby restaurants and pre-existing cafés.

The clusters were then ranked based on annual revenue (i.e., if the café were to be opened within that cluster), which was estimated using a set of business and mathematical assumptions.

According to the results, the neighborhood cluster consisting of *Sörnäinen* and *Punavuori* are the optimal new café locations.

THANK YOU

Young Hun Ji

linkedin.com/in/younghunji

github.com/bloonsinthesky