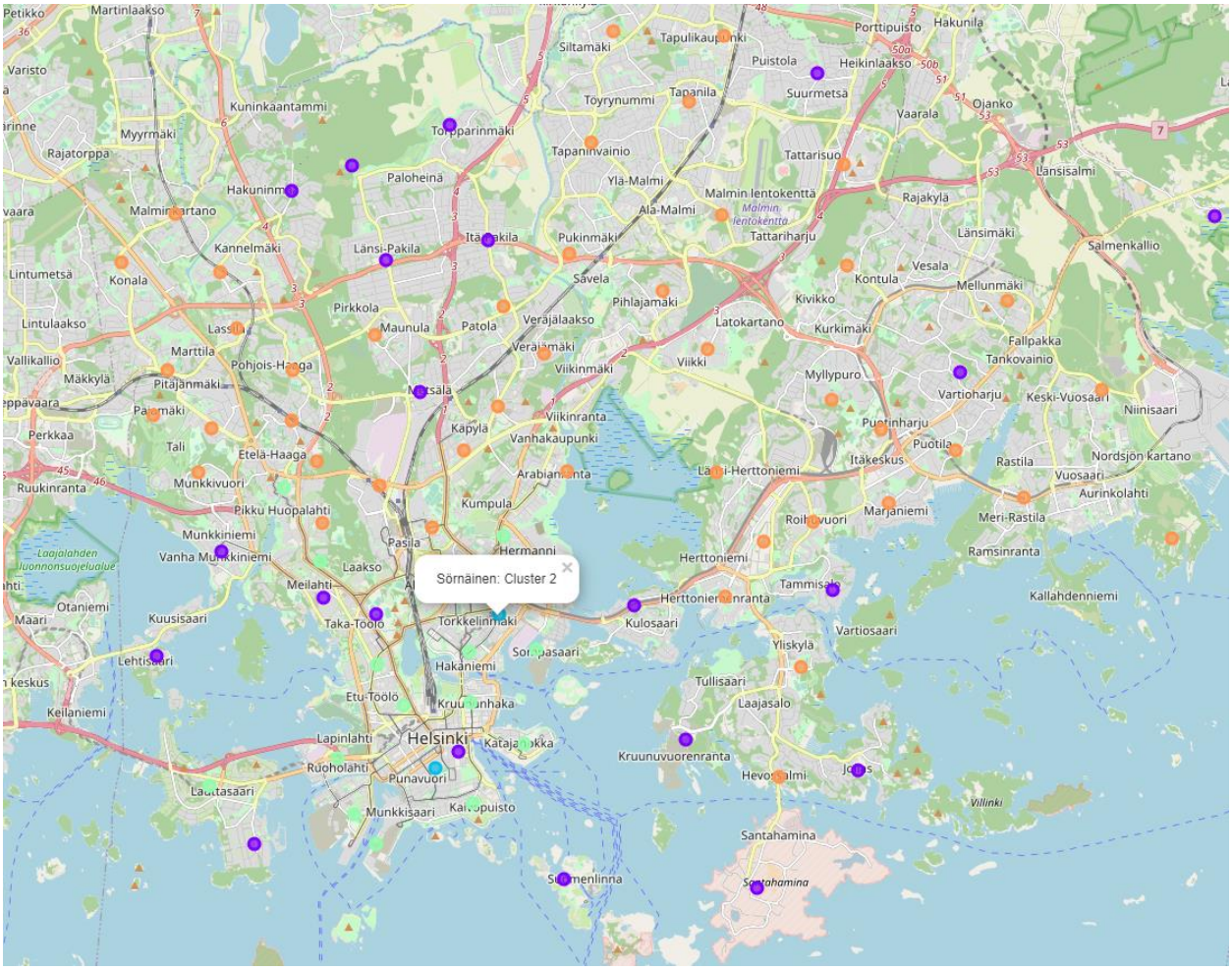
THE BATTLE OF NEIGHBORHOODS: A CLUSTERING APPROACH FOR DETERMINING THE OPTIMAL NEW CAFÉ LOCATION IN HELSINKI

by Young Hun Ji, Ph.D.



A report submitted to Coursera
in partial fulfillment of the requirements for the

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

2021

ABSTRACT

I used *k*-means clustering to determine the optimal location(s) within Helsinki, the capital of Finland, for opening a new café. The criteria used to determine the "optimal" neighborhood cluster was based on several business assumptions. Namely, I assumed that the optimal location has the following neighborhood characteristics:

- High median income among inhabitants
- High population density
- High proportion of people between the ages of 18 to 34
- High concentration of restaurants
- Low concentration of pre-existing cafés

After gathering neighborhood data regarding each of the above (from *Statistics Finland* and *Foursquare*), I segmented the neighborhoods into four clusters. The decision to use $k = 4$ clusters was based on a comparison of *inertia* and *silhouette* scores across a range of *k* values. Finally, I ranked the emerging clusters based on the new café's estimated annual revenue in each cluster. Annual revenue estimates were computed using a formula based on a set of mathematical assumptions regarding each of the features' impact on revenue. According to the results, the neighborhood cluster consisting of *Sörnäinen* and *Punavuori* are the optimal new café locations in Helsinki. Limitations and future directions are discussed.

INTRODUCTION

Background

Finland has the highest per capita consumption of coffee in the world at approximately 12 kg per year (Center for the Promotion of Imports, 2017). This makes Finland a compelling market for coffee. In this study, a hypothetical client wants to open a new café (called *Helsinki R&B Café*) somewhere in Helsinki, the capital of Finland. Additionally, if the café is successful, she would like to open additional cafés in other Helsinki neighborhoods with similar characteristics to the first. The client understands that the location of a café is a critical factor for its success and has gathered some preliminary business insights regarding what determines the "optimal location," based on a review of prior research. However, she lacks information regarding each of the neighborhoods specifically and thus cannot determine which of them fits the optimal location's profile.

Problem Statement

The overarching objective of this project is to determine the optimal Helsinki neighborhood(s) within which to open a new café. In this first phase of the investigation, the goal is to segment and cluster the neighborhoods in Helsinki based on a set of features that are linked

to the new café's likelihood of success. Additionally, the new café's annual revenue in each neighborhood cluster should be estimated (using a set of business and mathematical assumptions), and a recommendation based on the results should be made.

Assumptions: "What Constitutes Optimal Location?"

The café that the client wants to open is modern and relaxed with trendy R&B music and uses organic ingredients. She is targeting the 18-34 age demographic, as the venue is expected to be especially popular among students and workers. Based on a review of prior research, the client has produced a set of business assumptions regarding what constitutes the optimal location of the new café. Table 1 summarizes characteristics that the optimal location should possess. Prior to clustering analysis, five features were selected with respect to each of the following business assumptions.

Table 1
Assumptions Regarding Optimal Location Characteristics

Assumption	Rationale
High median income among inhabitants	Going to a café, especially one that uses organic ingredients, is a luxury activity. As such, the optimal location should accommodate inhabitants with high disposable income.
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 the new café is tailored to fit the tastes of young people between the ages of 18 and 34.
High concentration of restaurants	In Helsinki, people often go to 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 to minimize competition.

METHOD

Data Sources

Two types of data were collected at the neighborhood level: those related to the population (i.e., neighborhood population density, income, age) and those related to venues (i.e., concentration of restaurants and cafés in the neighborhood). First, population-related data was obtained from *Tilastokeskus* **(**also known as *Statistics Finland*): link to homepage.

Tilastokeskus is a Finnish governmental agency that annually publishes online census and other types of data pertaining to Finland at the country, municipal, and neighborhood levels. I obtained data on the following features regarding each Helsinki neighborhood: (a) name and postal code, (b) surface area, (c) number of inhabitants, (d) median income, and (e) proportion of inhabitants according to age brackets via the following link: data published in 2021 (Tilastokeskus, 2021). Specifically, on the "*Choose Variables*" page, I marked the following variables and downloaded the data in .xlsx format (also available in .csv format):

- Surface area
- Inhabitants, total, 2019(HE)
- 0-2 years, 2019 (HE)
- 3-6 years, 2019 (HE)
- 7-12 years, 2019 (HE)
- 13-15 years, 2019 (HE)
- 16-17 years, 2019 (HE)
- 18-19 years, 2019 (HE)
- 20-24 years, 2019 (HE)
- 25-29 years, 2019 (HE)
- 30-34 years, 2019 (HE)
- 35-39 years, 2019 (HE)
  (cont'd)
- 75-79 years, 2019 (HE)
- 80-84 years, 2019 (HE)
- 85 years or over, 2019 (HE)
- Aged 18 or over, total, 2019 (KO)
- Median income of inhabitants, 2019 (HR)

Second, venue-related data was obtained using the *Foursquare API*. I used the API to generate a list of venues within a 500m radius area of each neighborhood's latitude-longitude coordinates, along with each venue's category information (e.g., restaurant, café, shopping mall). The latitude-longitude coordinates of each neighborhood were obtained using *Geocoder*, a geocoding library that converts postal codes into latitude-longitude coordinates. After obtaining the venue data from Foursquare, I counted the number restaurants and cafés within a 500m radius of the neighborhood's central coordinates. I then used those counts as proxies for the concentration of restaurants and pre-existing cafés throughout each neighborhood.

Data Wrangling

I cleaned and analyzed the data using *MySQL* and Python's *pandas* library. First, I imported the population data into MySQL, and then used the *SQL magic extension* to write SQL

queries directly into a Jupyter notebook. Specifically, I ran queries to filter unused variables, relabel columns, and split column values. For example, the raw dataset contained a single column where each field contained both the name and postal code of a given neighborhood, and thus a query was written to split the names and postal codes into two separate columns. After the initial wave of data cleaning in SQL, I saved the population dataset into a *pandas* dataframe. Subsequently, in pandas, I computed several new variables using existing ones (e.g., I computed "population density" using the existing variables "surface area" and "population"). In addition, I collected venue-related data using the *Foursquare API* and appended it to the pandas dataframe.

Features

Five features were selected for clustering analysis, each based on the previously described assumptions regarding what constitutes the optimal new café location. They were (a) median income of neighborhood inhabitants, (b) neighborhood population density, (c) percentage of inhabitants aged 18-34, (d) concentration of restaurants in the neighborhood, and (e) concentration of pre-existing cafés in the neighborhood. Table 2 depicts the first five rows of the feature dataset prior to cluster analysis.

Table 2
First Five Rows of the Feature Dataset

Median Income	Population Density	Percentage aged 18-34	Restaurant Venues	Café Venues
29,706.00	7,791	33.10	11	3
29,816.00	17,246	30.77	18	4
32,894.00	3,688	27.75	8	5
31,718.00	8,540	27.36	8	5
28,478.00	6,889	33.93	0	0

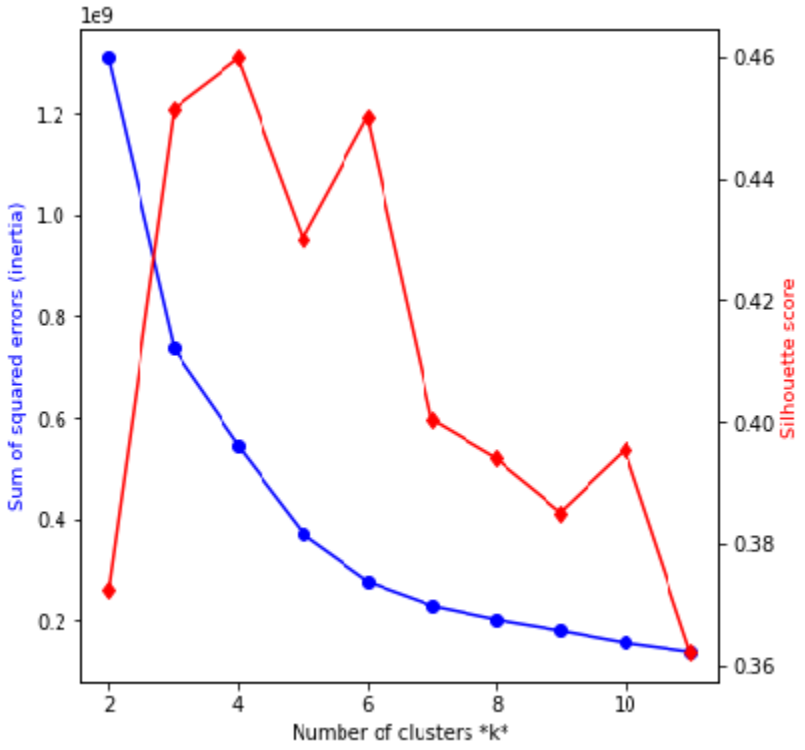
Note: Median income is in euros (€). Population density is the number of people per sq km. Restaurant and café venues are measured in terms of the number of restaurants and cafés within a 500km radius area of a neighborhood's central coordinates.

Clustering Approach

K-means clustering was implemented using the *scikit-learn* library. Prior to clustering, the optimal number of clusters was determined via a comparison of *inertia* and *silhouette* scores across *k* values ranging from 2 to 11 (for this analysis, a *k* value exceeding 11 was deemed too large to be practical). Figure 1 depicts a plot of the inertia and silhouette scores. Based on the

plot, I 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 plot began to "flatten." When $k = 4$, the *inertia* and *silhouette* score are equal to 5.46×10^8 and 0.46, respectively.

Figure 1
A Comparison of Inertia and Silhouette Scores Across a Range of *K*-Values



Estimating Annual Revenue Using Mathematical Assumptions

After clustering, I computed the means of each cluster's feature values and used them to estimate the "annual revenue" for each cluster (i.e., annual revenue of the new café if it were to open within that cluster). Annual revenues were estimated using a formula based on a set of mathematical assumptions regarding each feature's impact on the new café's annual revenue. For example, I assumed that people aged 18-34 are four times more likely than others to visit the new café. Table 3 summarizes the assumptions used to estimate annual revenues. Based on the assumptions set, I devised the following formula to estimate annual revenue:

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

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.

Table 3
Mathematical Assumptions Regarding Each Feature's Impact on Annual Revenue

Assumption	**Example**
The size of the café's target market is more-or-less equal to population density (i.e., the number of neighborhood inhabitants per sq km).	If a neighborhood has 1000 people per sq km, then 1000 is the maximum number of people that the café can reach per year.
If there are NO restaurants or pre-existing cafés nearby, the new café will attract 20% of inhabitants aged 18-34 versus 5% of all other inhabitants.	Assume that a neighborhood has a population density of 1000 and that 50% of them is aged 18-34. The new café will attract 100 customers aged 18-34 (i.e., 20% of its target age base of 500) and 25 customers of all other ages (i.e., 5% of its non-target age base of 500), assuming there are no restaurants or cafés nearby
If there ARE restaurants nearby, the probability of attracting customers, regardless of age, is enhanced by "$\log_{10}$(10 + 'concentration of restaurants')" times. In other words, the number of customers attracted rises logarithmically with respect to increasing concentration of restaurants.	Example: In a neighborhood with 0 restaurants (and 0 cafés), the café has a 20% chance of attracting a person aged 18-34. In contrast, in a neighborhood with 1 restaurant per 500m radius area, the probability rises to 20.8% (i.e., 20% times $\log_{10}$ (10 + 1)). If the concentration of restaurants is 10, 50, or 100, then that probability rises to 26.0%, 35.6%, and 40.8%, respectively.
If there ARE pre-existing cafés nearby, the probability of attracting customers, regardless of age, is divided by "1 + 'concentration of cafés'". In other words, the number of customers attracted falls proportionately with respect to increasing concentration of pre-existing cafés.	In a neighborhood with 0 cafés (and 0 restaurants), the café has a 20% probability of attracting a person aged 18-34. In contrast, in a neighborhood with 1 pre-existing café per 500m radius area, the probability reduced to 10%.
Collectively, the new café's visitors spend 2% of their annual income at the café per year.	Assume that the total number of unique visitors in a year is 100. Additionally, the median income of those visitors is €50,000. The estimated annual revenue of the new café, then, is €100,000 (i.e., 2% of 100 times €50,000).

Finally, the neighborhood clusters were ranked according to annual revenue estimates, and a recommendation was made based on that ranking.

RESULTS

Table 4 summarizes the means, standard deviations, and quartile feature values across Helsinki neighborhoods. On average, the 83 neighborhoods had a median annual income of 26,511.92 €, population density of 4409 per sq km, and people aged 18-34 comprising 25.9% of the inhabitants. Furthermore, there was a mean of 2 restaurants and 1 café per 500m radius area.

Table 4
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

Table 5 summarizes the mean feature values and annual revenue estimates for each of the neighborhood clusters. Cluster 1 had the highest median income at 30,623.95 €, but cluster 2 had the highest population density (i.e., 19,117 people per sq km), percentage of inhabitants aged 18-34 (i.e., 40.22%), and concentration of restaurants and cafés (i.e., 14 and 4 respectively). Moreover, cluster 2 had the highest estimated annual revenue (i.e., 344,007 €), which was computed using the set of mathematical assumptions described previously. These findings suggest that cluster 2 is the optimal location. The charts in Figure 2 compare the mean feature values, estimated number of unique yearly visitors, and annual revenue, across clusters. Finally, Figure 3 depicts a map of Helsinki with the emerging clusters superimposed on top.

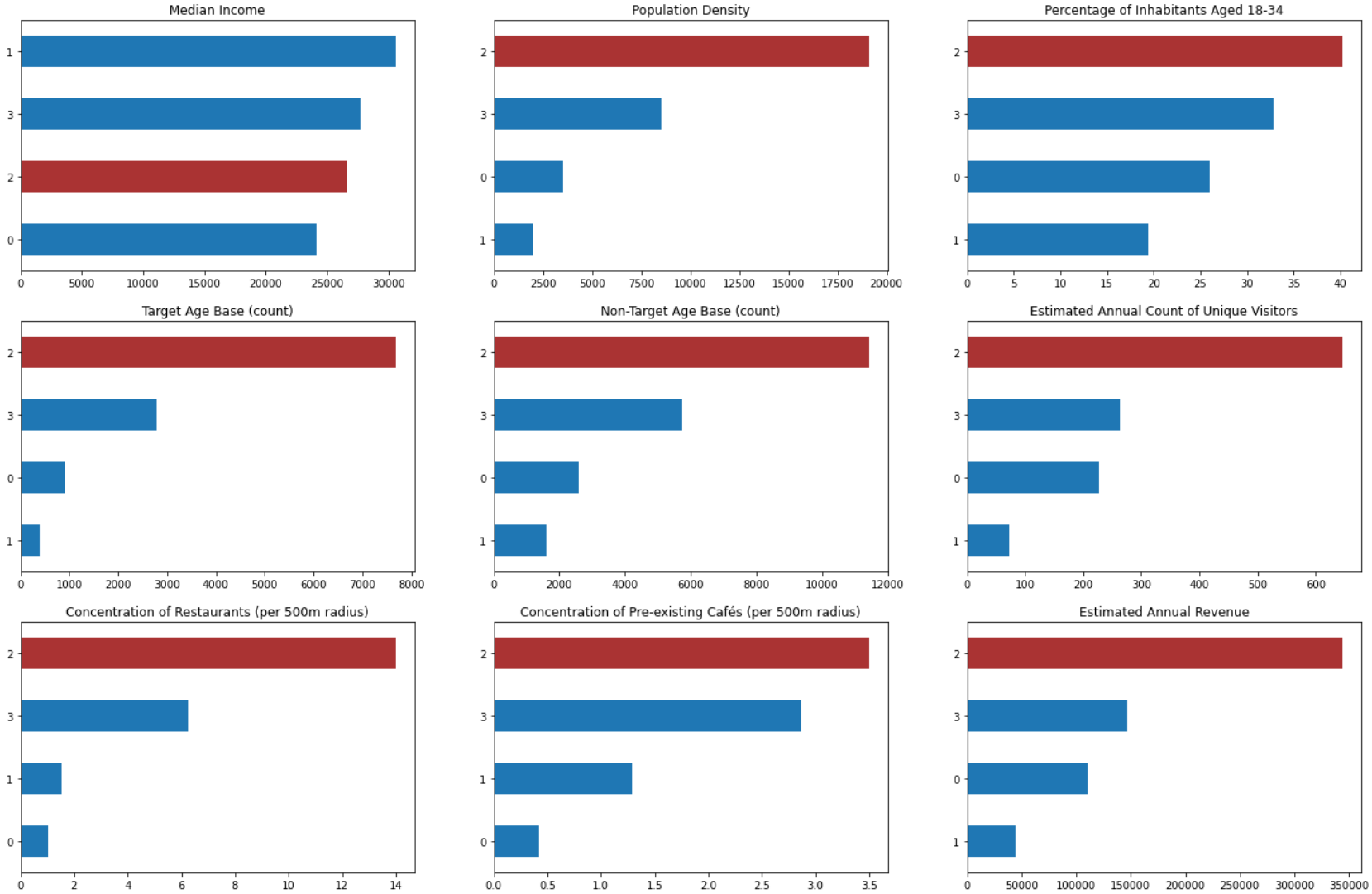
Table 5
Mean Feature Values and Revenue Estimate for Each Neighborhood Cluster

Cluster	Median Income	Population Density	Percentage 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 number of unique visitors the café will attract in one year. Annual revenue is in euros (€).

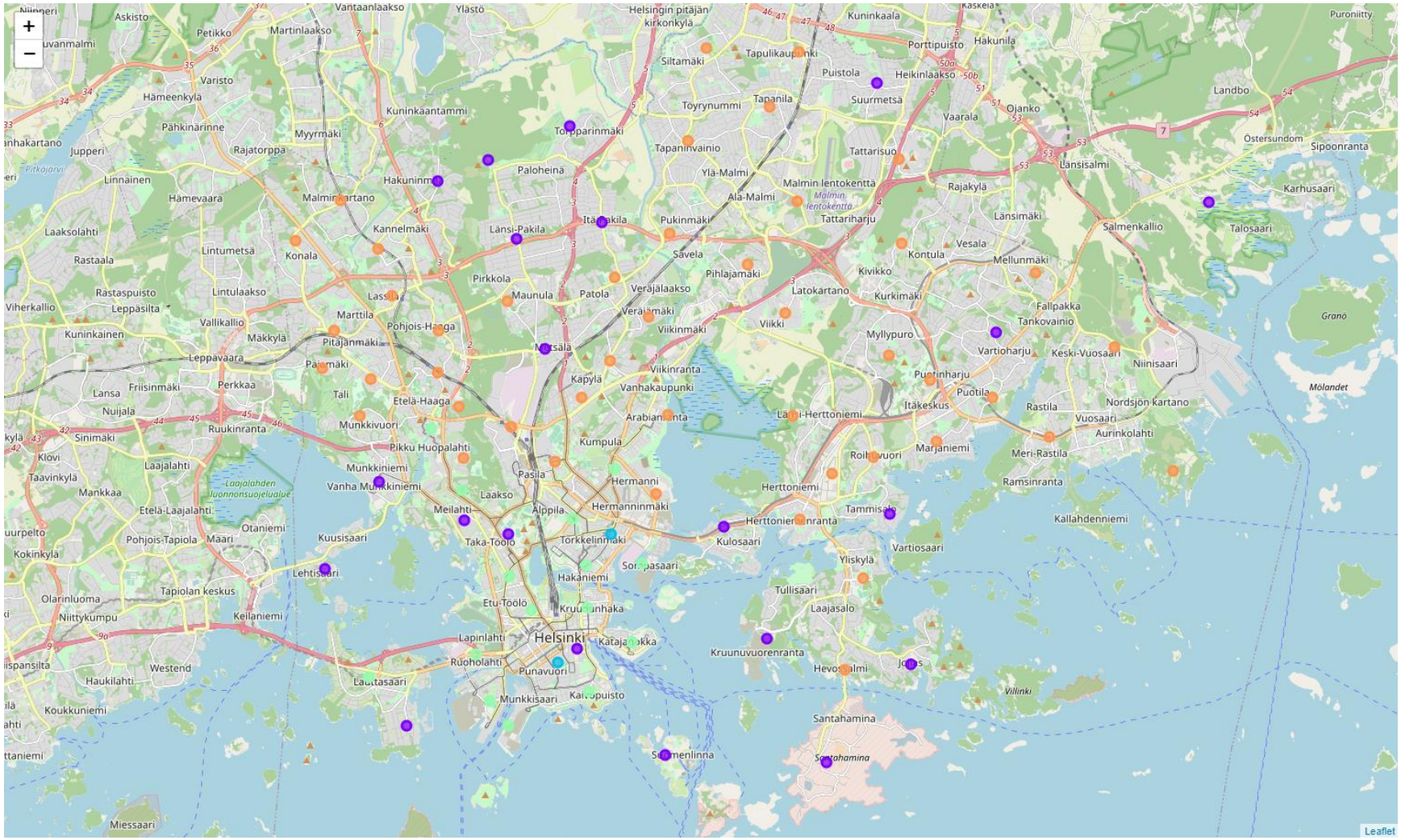
Figure 2

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



Note: "Target age base" refers to the number of inhabitants aged 18-34 that the café can reach in a year (assumed to be equal to density times the proportion of those aged 18-34). In contrast, "Non-target age base" refers to the number of inhabitants *not* aged 18-34 that the café can reach.

Figure 3
A Map of Helsinki with Emerging Neighborhood Clusters Superimposed on Top



DISCUSSION

According to the results, neighborhood cluster 2 was deemed optimal because it yielded the highest estimated annual revenue. Cluster 2, however, did not have the optimal values on *all* of the key feature variables. In fact, Cluster 2 had the following disadvantageous characteristics:

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

These were nonetheless offset by the following advantageous characteristics:

- It had the highest population density.
- It had the highest proportion of inhabitants aged 18-34.
- It had the highest concentration of restaurants per 500m radius area.

The hypothetical client should keep in mind each of the foregoing pros and cons if she decides to open the new café within this neighborhood cluster.

Cluster 2 was comprised of two neighborhoods: *Punavuori* and *Sörnäinen*. Between the two, Sörnäinen had the larger annual revenue estimate of 397,474 € versus Punavuori's 285,637 €. As such, the practical recommendation to the client is to open the new café in Sörnäinen first. Then, if the café is successful, she could consider opening another one in Punavuori.

Limitations and Future Directions

The present analysis represents the "first phase" of what would become a larger examination into the hypothetical client's problem. As such, I acknowledge that there are numerous limitations that should be addressed in subsequent analyses.

First, I chose to cluster the neighborhood based on a total of 5 features only: median income, percentage of inhabitants aged 18-34, population density, concentration of restaurants, and concentration of pre-existing cafés. While this parsimony helped to make data collection and analysis efficient, there are many other factors that could drive a neighborhood's suitability for opening a new café, such as its safety reputation, presence of "R-kioski" (i.e., a popular convenience store chain that sells coffee), accessibility by public transport, neighborhood development plans, and the quality of nearby restaurants and rival cafés. As such, follow-up analyses should aim to collect and examine such additional features.

Second, the concentration of pre-existing restaurants and cafés in each neighborhood was measured using proxies. Specifically, I counted the number of restaurants and cafés within a 500m radius area from each neighborhood's central coordinates, which may not accurately

represent the presence of those venues throughout the entire neighborhood. For example, there may be a neighborhood with a high concentration of cafés in its outskirts despite having only a few in its central coordinates.

Third, I estimated annual revenue in each cluster using a formula based on a set of relatively simple mathematical assumptions using only a few variables. For example, the estimation did not take into account sales from tourists or other patrons outside of the focal neighborhood. For example, in a neighborhood such as, say, Suomenlinna, which is a popular destination for travelers with relatively few inhabitants, coffee sales may be driven largely by tourists and locals from other neighborhoods. As such, the true revenue for the island neighborhood could be substantially higher than what I had estimated.

Finally, I determined the optimal new location solely based on estimated revenues. As such, I did not take into account the variation in costs (e.g., property cost) of opening a new café in the different locations. I recommend that follow-up analyses address this limitation by estimating total costs in addition to revenue.

Conclusion

In this project, I examined the hypothetical client's question of "*What is the optimal new café location in Helsinki?*" Using population-related data and venue-related data gathered from *Tilastokeskus* and *Foursquare*, respectively, I segmented the neighborhoods of Helsinki into four clusters based on features such as population density, income, proportion of young people, and concentration of nearby restaurants and pre-existing cafés. I then ranked the clusters 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. Follow-up examinations should aim to utilize additional data (e.g., cost-related data) using more robust assumptions for greater validity.

REFERENCES

Center for the Promotion of Imports. (2017). Exporting coffee to Finland. https://www.cbi.eu/market-information/coffee/finland

Tilastokeskus. (2021). Data published in 2021. https://pxnet2.stat.fi/PXWeb/pxweb/en/Postinumeroalueittainen_avoin_tieto/