# CAPSTONE PROJECT REPORT : BATTLE OF THE CITIES

## INTRODUCTION

An enthusiastic entrepreneur wants to open up a bookstore in a populated city in Finland. Finland, a hidden gem tucked into the far up North, known for its clean air and water, beautiful forests, amazing blue lakes and its education. Bookstores are vanishing these days due to online books or eBooks but some people still prefer reading from an actual book, some folks are still not very familiar or one could say not comfortable with using the internet to read books, magazines or newspapers. Some people like the feel of reading from an actual book, sit by the window and have a hot beverage along with it. This entrepreneur wants to bring back this same feel of reading from a book, wants to open a bookstore with an in house cafe, a place to sit so the customers can enjoy their book after their purchase and also a kids' corner to enhance their knowledge as well, but the entrepreneur faces a problem in choosing a suitable city among all the beautiful cities in Finland. This project would be aiming at solving this very problem of finding appropriate city.

Use of this project:

This project can be used by anyone who intends to

- open a bookstore or,
- open any other retail business or,
- a chain of bookstores in different cities or,
- a chain of any other retail business in different cities in Finland.

# DATA

For this project a lot of data would be required, in order to find a suitable city in Finland one needs to know the cities in Finland for that the names of the cities in Finland would be required, also to pinpoint the location of the cities on a map , the latitude and longitude of the respective cities would also be needed this data can be found on the internet to be specific on a website called: Simple maps, Finland Cities Database.This website provides a table consisting of the names of the city and their latitudes and longitudes in the form of a CSV file which can later be transformed into a pandas dataframe.

Also for this project, data regarding the most common places visited in the cities would be needed as well. This data helps in finding out what the city is famous for or what the people in that particular city prefer to do for leisure, which is what we are interested in for this project. To get this data, Foursquare API is used.

Foursquare is a social location service that allows users to explore the world around them. The Foursquare API allows application developers to extend the platform in interesting ways. Developers can build location management tools, custom search engines, and even games and other tools that interact with the Foursquare API. For example, one could build a geolocation game that allows players to also check in to Foursquare locations as a natural by-product of normal gameplay. In this case the Foursquare API, would be used to get the names, id, latitude and longitude of the respective venues in and around the cities of Finland.

Packages and libraries used in this project to work with the data are:

- numpy,
- pandas
- random
- requests
- matplotlib

- nominatim
- json_normalize
- Folium
- Kmeans
- geopy
- geocoder

# METHODOLOGY

In this section the use of data for data analysis and the methods used to obtain the required results are discussed.

**Step 1.** Install and import all the necessary libraries and packages.

**Step 2.** Download the cities data from the website into a CSV file and then convert this CSV file into a dataframe using pandas. Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Then the dataframe is  cleaned, checked for any missing values and errors, this is also known as data pre-processing. *Data pre-processing* is a data mining technique that involves transforming raw data into an understandable format. Once that is done then our main data  for the cities is ready, in a dataframe called fincity_df.

| | city | Latitude | Longitude | Region(Finnish) | capital | population |
|---|---|---|---|---|---|---|
| **0** | Helsinki | 60.1756 | 24.9342 | Uusimaa | primary | 642045.0 |
| **1** | Espoo | 60.2100 | 24.6600 | Uusimaa | minor | 269802.0 |
| **2** | Tampere | 61.4981 | 23.7608 | Pirkanmaa | admin | 225118.0 |
| **3** | Vantaa | 60.3000 | 25.0333 | Uusimaa | minor | 214605.0 |
| **4** | Oulu | 65.0142 | 25.4719 | Pohjois-Pohjanmaa | admin | 200526.0 |

**Step 3.** Get a visual of the cities in Finland. The cities  from the database fincity_df, are plotted on to the map of Finland using Folium. Folium is a powerful Python library that helps you create several types of Leaflet maps. It makes it easy to visualize data that's been manipulated in Python on an interactive Leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing Vincent/Vega visualizations as

markers on the map. This is shown in the below picture.
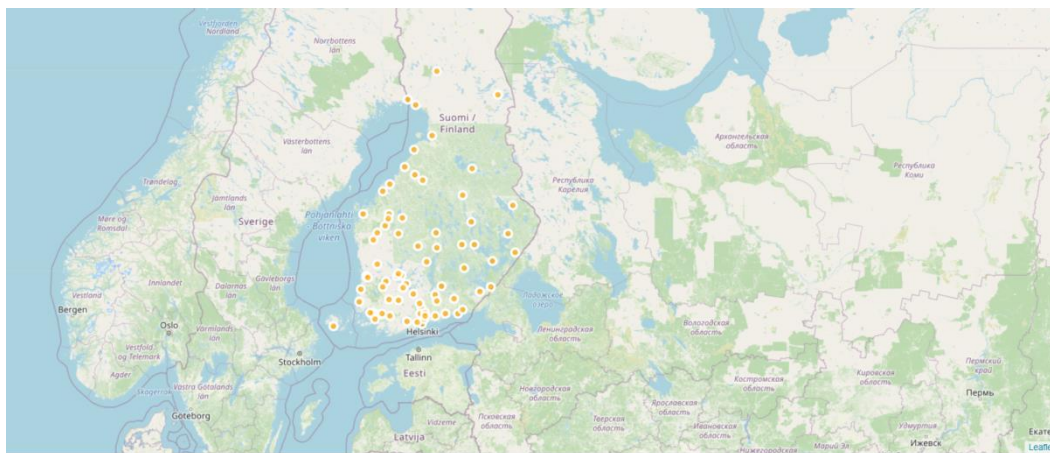


The above map shows cities in Finland.

**Step 4.** For the second dataframe , the Foursquare API is used to retrieve the data containing the venue names and their latitude and longitude locations. To gain access to the Foursquare API, credentials are paced like the client ID and the client secret. When the Foursquare API is used, the data obtained is in the JSON format, JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate, to overcome this problem a function is created called *getNearbyVenues* to clean the data in JSON. This function extracts only the relevant information and then converts it into a dataframe when called.

This gives us our second dataframe called fin_venues, that contains the city, its latitude and longitudes and the names of the venues their Id's and the latitudes and longitudes of each venue.

Size of the new venues dataframe  (3391, 8)

| | city | Lat | Long | Venue | Venue Id | Venue Lat | Venue Long | Venue Category |
|---|---|---|---|---|---|---|---|---|
| 0 | Helsinki | 60.1756 | 24.9342 | Arkadia Oy International Bookshop | 4bc08b95461576b0d6417a32 | 60.173369 | 24.929330 | Bookstore |
| 1 | Helsinki | 60.1756 | 24.9342 | Taidehalli | 4adcdb23f964a520dc6021e3 | 60.172127 | 24.931014 | Art Gallery |
| 2 | Helsinki | 60.1756 | 24.9342 | Sinisen huvilan kahvila | 4be302eb63609c74cfd51bff | 60.181305 | 24.937043 | Café |
| 3 | Helsinki | 60.1756 | 24.9342 | Cafe Rouge | 5555afa9498efb7ce749253c | 60.168711 | 24.933027 | Middle Eastern Restaurant |
| 4 | Helsinki | 60.1756 | 24.9342 | Buongiorno Cafe & Restaurant | 51289893e4b0386981d9e120 | 60.175304 | 24.919294 | Café |

**Step 5.** Analyzing each city's venues. One hot encoding is used to group all the different venues according to their category for each city. One hot encoding refers to splitting the column which contains numerical *categorical data* to many columns depending on the number of categories present in that

column. Each column contains "0" or "1" corresponding to which column it has been placed.

This results in a dataframe called fin_onehot with all categories as the columns and the cities placed in each row showing 1 or 0, 1 if that particular category of venue is in or around the city and 0 if not.

**Step 6.** A function called *return_most_common_venues* is created to sort the venues in descending order

**Step 7.** Another dataframe called f_ venues is created that contains the top 10 most common venues for each city.

**Step 8.** Use of customer segmentation specifically, KMeans algorithm. Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics. One of the algorithms that can be used for customer segmentation is K-Means clustering.

K-Means can group data only unsupervised based on the similarity of customers to each other. It is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data. The objective of K-Means is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters. In our case we use the KMeans algorithm to divide the data into 4 clusters and then finally create one last dataframe that contains all the data, the name of the city, the latitude and longitude of the city, the top 10 most common venues of that city and the cluster number that the city belongs to

**Step 9.** Finally map the clusters on the map of Finland using Folium.
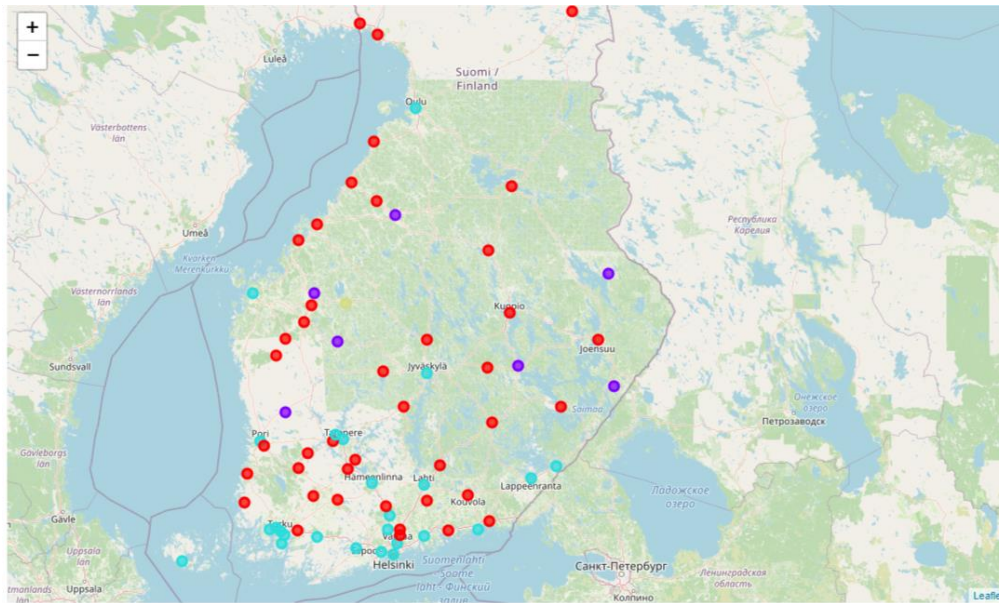


**Step 10.** Print all the different clusters separately.

# RESULTS:

The above analysis results in 4 clusters. As shown in the map

- **Cluster 1**:  Shows venues fit for family leisure like parks , theatres, bookstore, hotels, etc.

Cluster 1 :Leisure and Shopping

```
findf.loc[findf['Cluster Labels'] == 0,findf.columns[[0] + list(range(5, findf.shape[1]))]]
```

| | city | population | Cluster Labels | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | Kouvola | 85855.0 | 0 | Supermarket | Pizza Place | Fast Food Restaurant | Coffee Shop | Sandwich Place | Café | Grocery Store | Gym / Fitness Center | Chinese Restaurant | Pharmacy |
| 11 | Joensuu | 75514.0 | 0 | Supermarket | Café | Gym / Fitness Center | Bar | Fast Food Restaurant | Chinese Restaurant | Scandinavian Restaurant | Sandwich Place | Hotel | Coffee Shop |
| 17 | Mikkeli | 54665.0 | 0 | Supermarket | Restaurant | Café | Fast Food Restaurant | Flea Market | Coffee Shop | Hotel | Movie Theater | Shopping Mall | Grocery Store |
| 26 | Rauma | 39809.0 | 0 | Supermarket | Café | Restaurant | Fast Food Restaurant | Grocery Store | Pizza Place | Hockey Arena | Turkish Restaurant | Kebab Restaurant | Steakhouse |
| 28 | Savonlinna | 35523.0 | 0 | Supermarket | Hotel | Pizza Place | Ski Area | Brewery | Train Station | Scandinavian Restaurant | Seafood Restaurant | Resort | Discount Store |
| 35 | Sastamala | 25220.0 | 0 | Café | Supermarket | Pizza Place | Sandwich Place | Grocery Store | Bakery | Train Station | Chinese Restaurant | Gas Station | Fish Market |
| 36 | Raahe | 25165.0 | 0 | Fast Food Restaurant | Supermarket | Grocery Store | Flea Market | Bakery | Theater | Café | Sandwich Place | Business Service | Chinese Restaurant |
| 38 | Iisalmi | 21945.0 | 0 | Supermarket | Bar | Fast Food Restaurant | Golf Course | Train Station | Gastropub | Café | Grocery Store | Italian Restaurant | Flea Market |
| 39 | Tornio | 21928.0 | 0 | Supermarket | Grocery Store | Smoke Shop | Chinese Restaurant | Hockey Arena | Shopping Mall | Fast Food Restaurant | Clothing Store | Bar | Hotel |
| 40 | Kemi | 21758.0 | 0 | Supermarket | Grocery Store | Hotel | Event Space | Castle | Beer Bar | Gas Station | Ski Area | Boat or Ferry | Train Station |
| 41 | Kurikka | 21734.0 | 0 | Supermarket | Discount Store | Pharmacy | Grocery Store | Restaurant | Hockey Arena | Fast Food Restaurant | Pizza Place | Scandinavian Restaurant | Gas Station |
| 42 | Jämsä | 21542.0 | 0 | Ski Area | Supermarket | Restaurant | Apres Ski Bar | Intersection | Café | Nightclub | Sandwich Place | Athletics & Sports | Shoe Store |
| 43 | Valkeakoski | 21332.0 | 0 | Supermarket | Café | Sandwich Place | Turkish Restaurant | Grocery Store | Harbor / Marina | Bridge | Ski Chalet | Shopping Mall | Chinese Restaurant |

- **Cluster 2**: Shows supermarkets and transport services.

Cluster 2 : Supermarkets and Transportational services

```
findf.loc[findf['Cluster Labels'] == 1,findf.columns[[0] + list(range(5, findf.shape[1]))]]
```

| | city | population | Cluster Labels | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 53 | Kauhava | 16784.0 | 1 | Supermarket | Hotel | Airport | Train Station | Discount Store | Flea Market | Fast Food Restaurant | Filipino Restaurant | Fireworks Store | Fish Market |
| 66 | Lieksa | 11772.0 | 1 | Supermarket | Pizza Place | Discount Store | Train Station | Bakery | Food Service | Food Court | Food Truck | Food & Drink Shop | Food |
| 69 | Nivala | 10876.0 | 1 | Supermarket | Pizza Place | Train Station | Turkish Restaurant | Burger Joint | Fishing Store | Fast Food Restaurant | Filipino Restaurant | Fireworks Store | Fish Market |

- **Cluster 3**: Shows cafe's, bars, pubs and restaurants and all high end leisure venues.

Cluster 3 : Cafe's, Bars, Pubs and Resturants

```
findf.loc[findf['Cluster Labels'] == 2,findf.columns[[0] + list(range(5, findf.shape[1]))]]
```

| | city | population | Cluster Labels | 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 | Helsinki | 642045.0 | 2 | Café | Scandinavian Restaurant | Coffee Shop | Hotel | Pizza Place | Park | Bakery | Indie Movie Theater | Theater | French Restaurant |
| 1 | Espoo | 269802.0 | 2 | Café | Gym / Fitness Center | Beach | Golf Course | Coffee Shop | Gym | Himalayan Restaurant | Pizza Place | Park | Juice Bar |
| 2 | Tampere | 225118.0 | 2 | Café | Gym / Fitness Center | Park | Scenic Lookout | Gastropub | Restaurant | Kebab Restaurant | Bistro | Sauna / Steam Room | Pizza Place |
| 3 | Vantaa | 214605.0 | 2 | Recreation Center | Gym / Fitness Center | Café | Pizza Place | Hotel | Airport Lounge | Coffee Shop | Sushi Restaurant | Sporting Goods Shop | Thai Restaurant |
| 4 | Oulu | 200526.0 | 2 | Café | Pizza Place | Supermarket | Restaurant | Gym / Fitness Center | Fast Food Restaurant | Indian Restaurant | Pub | Pool Hall | Chinese Restaurant |
| 5 | Turku | 187604.0 | 2 | Café | Gym / Fitness Center | Park | Vegetarian / Vegan Restaurant | Scandinavian Restaurant | Gym | Restaurant | Pizza Place | Bar | Beer Bar |
| 6 | Jyväskylä | 137368.0 | 2 | Supermarket | Café | Scandinavian Restaurant | Park | Gym | Grocery Store | Coffee Shop | Gym / Fitness Center | General Entertainment | Music Store |
| 7 | Lahti | 118119.0 | 2 | Café | Supermarket | Restaurant | Ski Area | Bar | Burger Joint | Pizza Place | Beach | Gym | Gym / Fitness Center |
| 8 | Kuopio | 112117.0 | 2 | Supermarket | Grocery Store | Café | Bar | Fast Food Restaurant | Scandinavian Restaurant | Pizza Place | Pub | Hotel | Italian Restaurant |
| 10 | Pori | 85363.0 | 2 | Supermarket | Café | Pizza Place | Grocery Store | Gym / Fitness Center | Fast Food Restaurant | Mexican Restaurant | Park | Bar | Shopping Mall |
| 12 | Lappeenranta | 72875.0 | 2 | Supermarket | Pizza Place | Café | Restaurant | Bar | Grocery Store | Beach | Electronics Store | Clothing Store | Donut Shop |
| 13 | Hämeenlinna | 68011.0 | 2 | Café | Supermarket | Park | History Museum | Golf Course | Scandinavian | Fast Food Restaurant | Pizza Place | Indian Restaurant | Department Store |

- **Cluster 4**: Shows markets like supermarkets and fish market.

Cluster 4 : Markets

```
findf.loc[findf['Cluster Labels'] == 3,findf.columns[[0] + list(range(5, findf.shape[1]))]]
```

| | city | population | Cluster Labels | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 74 | Alajärvi | 10006.0 | 3 | Supermarket | Baseball Field | Gas Station | Flower Shop | Filipino Restaurant | Fireworks Store | Fish Market | Fishing Spot | Fishing Store | Flea Market |

The suitable cluster is  *Cluster 1*.

# Discussion Section:

From the above analysis and resulting clusters we see that cluster 1, shows venues fit for family leisure like parks , theatres, bookstore, hotels, etc. which should be more suitable to open up a bookstore, Cluster 2, shows supermarkets and transport services, Cluster 3, shows cafe's, bars, pubs and restaurants and all high end leisure venues and Cluster 4, shows markets like supermarkets and fish market. Now, as mentioned before the owner wants to open a bookstore where the elderly can read, an in store cafeteria, a kids' corner, a place to sit, so by this description we can assume that the owner wants to cater to a crowd of people who have families, people who are elderly and people who find going to a bookstore to be a type of leisure activity. These qualities are seen the people who reside in the cities in cluster 1 as the most common places visited in these cities are supermarkets, shopping malls, zoo's, cafes, grocery store, theme parks, pizza place, diners, skating rings places that families usually tend to go and if we see closely we find that most of the places are retail businesses as well, which also means that if anyone wants to use this project to open up any other retail business or a chain of retail businesses or bookstores can do so by using the results of this project. Hence cluster 1 fits perfectly.

# Conclusion :

A conclusion can be drawn that cities in cluster 1 are more suitable to open up a bookstore. as the cities in cluster 1 are fairly populated and is advisable for the group of people that the owner plans on targeting. Hence, *Cities in Cluster  are Suitable* for opening a bookstore.

These cities are :

*Kuopio, Kouvola, Joensuu, Seinäjoki, Mikkeli, Kokkola, Järvenpää, Rauma, Kajaani, Savonlinna, Kerava, Nokia, Riihimäki, Sastamala, Raahe, Iisalmi, Tornio, Kemi, Kurikka, Jämsä, Valkeakoski, Hamina, Äänekoski, Heinola, Jakobstad, Pieksämäki, Forssa, Toijala, Loimaa, Orimattila, Kuusamo, Uusikaupunki, Lovisa, Ylivieska, Lapua, Kauhajoki, Ulvila, Kalajoki, Paimio, Huittinen, Keuruu.*