CAPSTONE PROJECT - Benefit Zones for Chain of Indian Restaurants

Prakirth Govardhanam

Applied Data Science Capstone by IBM/Coursera

CONTENTS

Introduction

- Background
- Aim

Data

Methodology

- Hypothesis
- Assumptions

Results & Discussion

- Data Preparation
- Exploratory Data Analysis

Conclusion

INTRODUCTION

BACKGROUND

Generally, Businesses, especially Chain Stores, thrive when the possibility of risk/competition is either low or close to none (Not Applicable to Start-Ups). This perspective (unlike Mark Zuckerberg's strategy to monopolize Social Media) signifies tact and subtlety and not eliminating competition but circumventing the need to compete when possible.

However, competition for every Business is as essential as work-outs for a healthy human body and mind. Hence, I conclude by quoting Jack Ma (Co-founder of Alibaba Group).

"Forget about Your Competitors. Focus on Your Customers"

INTRODUCTION

AIM

In this project, I focus on the possibility of establishing a chain of Indian restaurants in Helsinki, capital city of Finland (one of the Nordic countries). I try to find possible-beneficial locations within the Neighbourhoods (Districts) of Helsinki for establishing a chain of Indian restaurants. The primary conditions for the project are as follows:

1. Locate *Popularity Centre* (Major Assumption 1) in the District – for attention
2. Identify Districts with:
   - Absence of Indian restaurants – for profitable Business
   - Presence of Indian restaurants – for moderate competition to evolve Business model
3. Presence of Indian restaurants within Popular venues – to circumvent extreme competition

DATA

Data sources used to determine Districts within the city of Helsinki are provided by:

- Wikipedia – for labels of the Districts of Helsinki
- ***geopy*** (Python package) – for coordinates of the Districts of Helsinki
- **Foursquare** (API) – for popular venues, restaurants and their respective geospatial data

METHODOLOGY

HYPOTHESIS

Although, Indian restaurants are specific to Indians, the variety of Indian cuisine is enjoyed delightfully by all the Nordic countries. Hence, the Customer base for Indian cuisine in the Nordics is awaiting the arrival of native-authentic restaurants for Business.

(PS: Market research for establishing the above hypothesis is out-of-scope for this project)

METHODOLOGY

ASSUMPTIONS

Major Assumption

1. Popularity Centre = the centroid (mean position) of the most popular venues from the Top10 Venue Categories (by frequency of occurence) in each District will be considered as the "Popularity Centre" within every District

2. In fact, more Indian Restaurants exist than explored Indian Restaurants using FourSquare API. Since the project is based on "**using FourSquare API for implementation of the Idea**" we will assume the following:

   "*explored Indian Restaurants*" == "*existing Indian Restaurants*"

Clarification #1: Popular Venues from Top10 Venue Categories were ideally planned to be filtered by Ratings of Venues. Unfortunately, I have a Sandbox account & Ratings of Venues at the scale I need would be possible only with Premium accounts

METHODOLOGY

ASSUMPTIONS

Minor Assumption

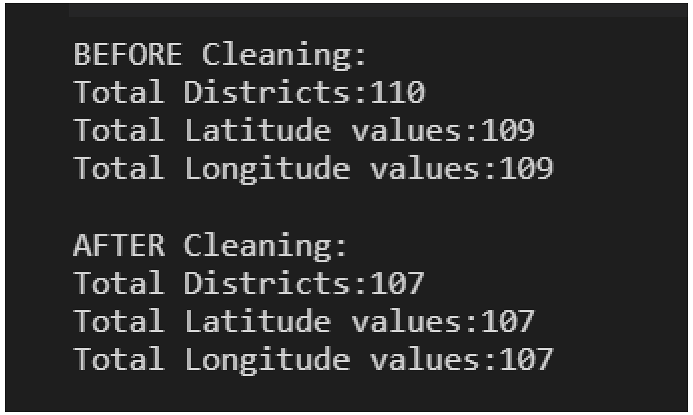
1. From the source of District labels, Swedish-names of Finland could be confused with Swedish-names of Sweden in the FourSquare API (as observed during analysis). Hence, we will extract and work only with Finnish-names of the Districts.
2. Since, *explored Indian Restaurants* are very less in number, we will consider both Venue Categories, ***Indian Restaurant*** and ***Himalayan Restaurant*** as *'Indian Restaurant'*

RESULTS & DISCUSSION

DATA PREPARATION

Data Preparation involves: (*refer code for details*)

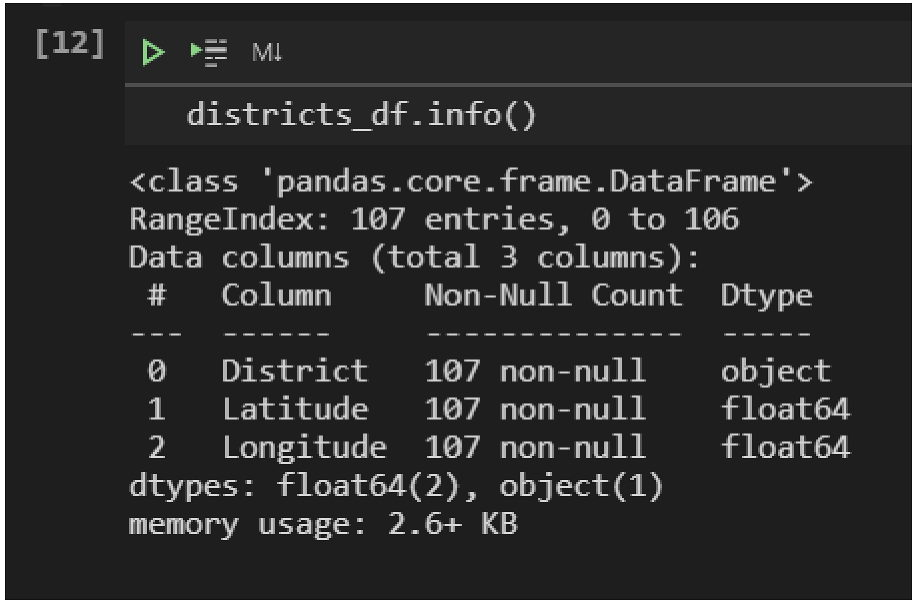
Data Extraction & Data Cleaning



```
BEFORE Cleaning:
Total Districts:110
Total Latitude values:109
Total Longitude values:109

AFTER Cleaning:
Total Districts:107
Total Latitude values:107
Total Longitude values:107
```

Data Organizing



```
districts_df.info()
```

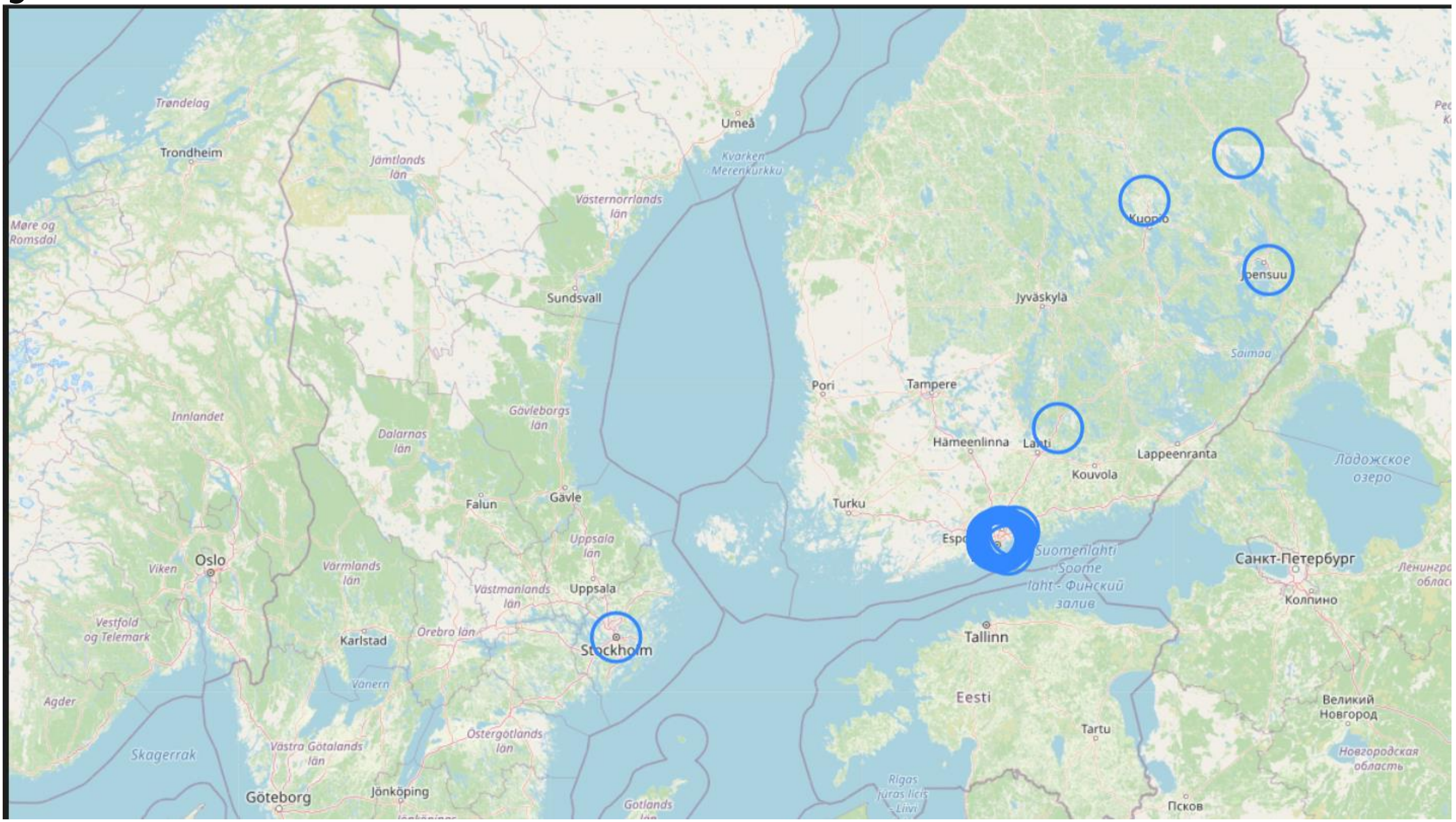
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   District   107 non-null    object 
 1   Latitude   107 non-null    float64
 2   Longitude  107 non-null    float64
dtypes: float64(2), object(1)
memory usage: 2.6+ KB
```

RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis – involves a combination of both, Data Analysis & Data Visualization

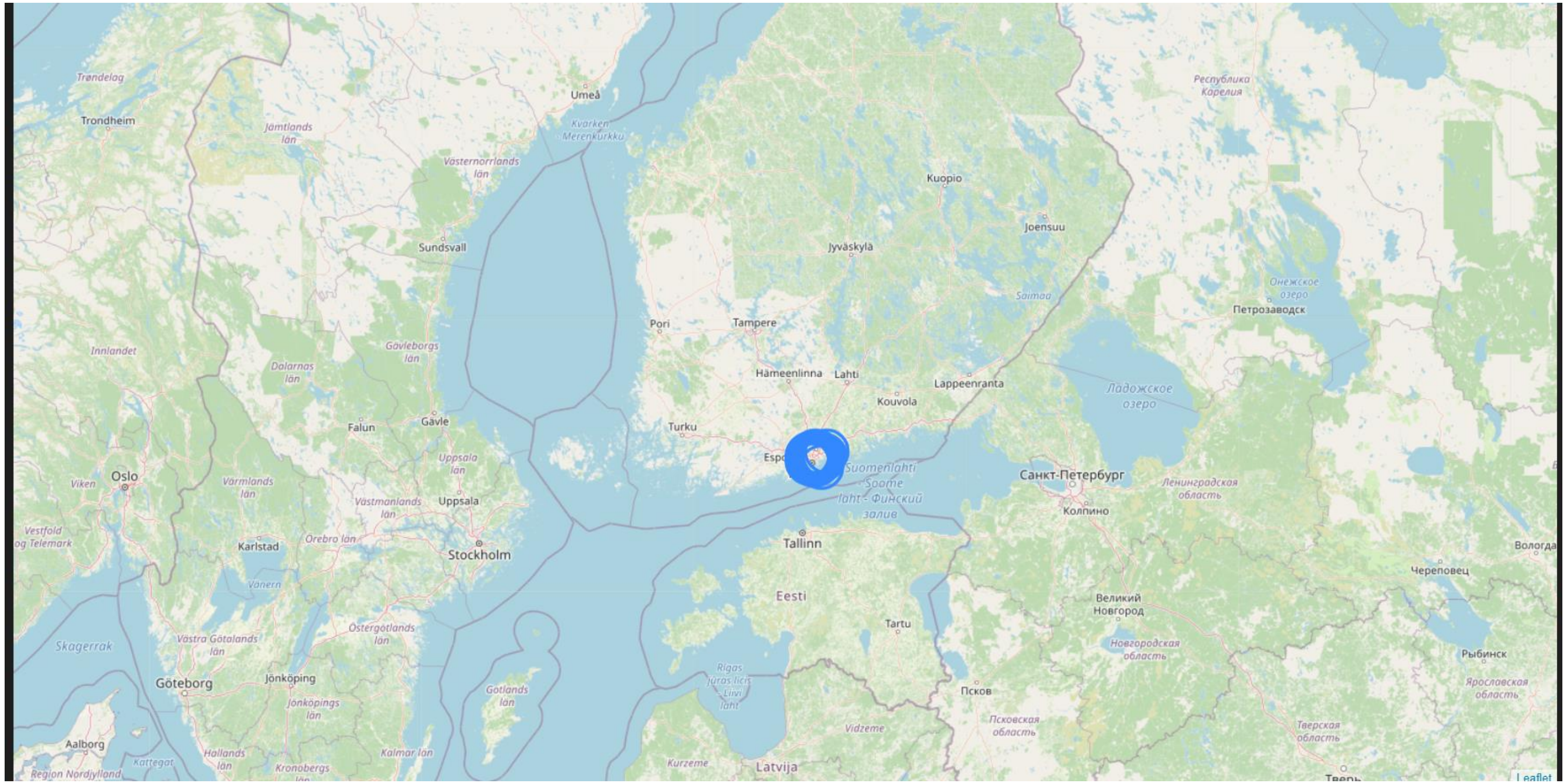
- Five Districts ('Vanhakaupunki', 'Siltasaari', 'Reijola', 'Vironniemi', 'Koivusaari') were identifed as irrelevant Districts, which was probably due to overlapping names of the Districts across regions using *folium* (python library)



RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

Districts and their coordinates identified as irrelevant were omitted from the final data as well, leading to 102 Districts in total.



RESULTS & DISCUSSION
EXPLORATORY DATA ANALYSIS

Clarification #3: Venues data acquired from FourSquare API through function, *getNearbyVenues*(), differs from day-to-day, probably due to satellite movement and communication differences. Reproduciblity cannot be expected in the Top10 Venue Categories on different days of execution. Hence, the values such as the following might differ:

- Total Districts with venues,
- Total venues,
- Total Venue Categories &
- Total Indian venues

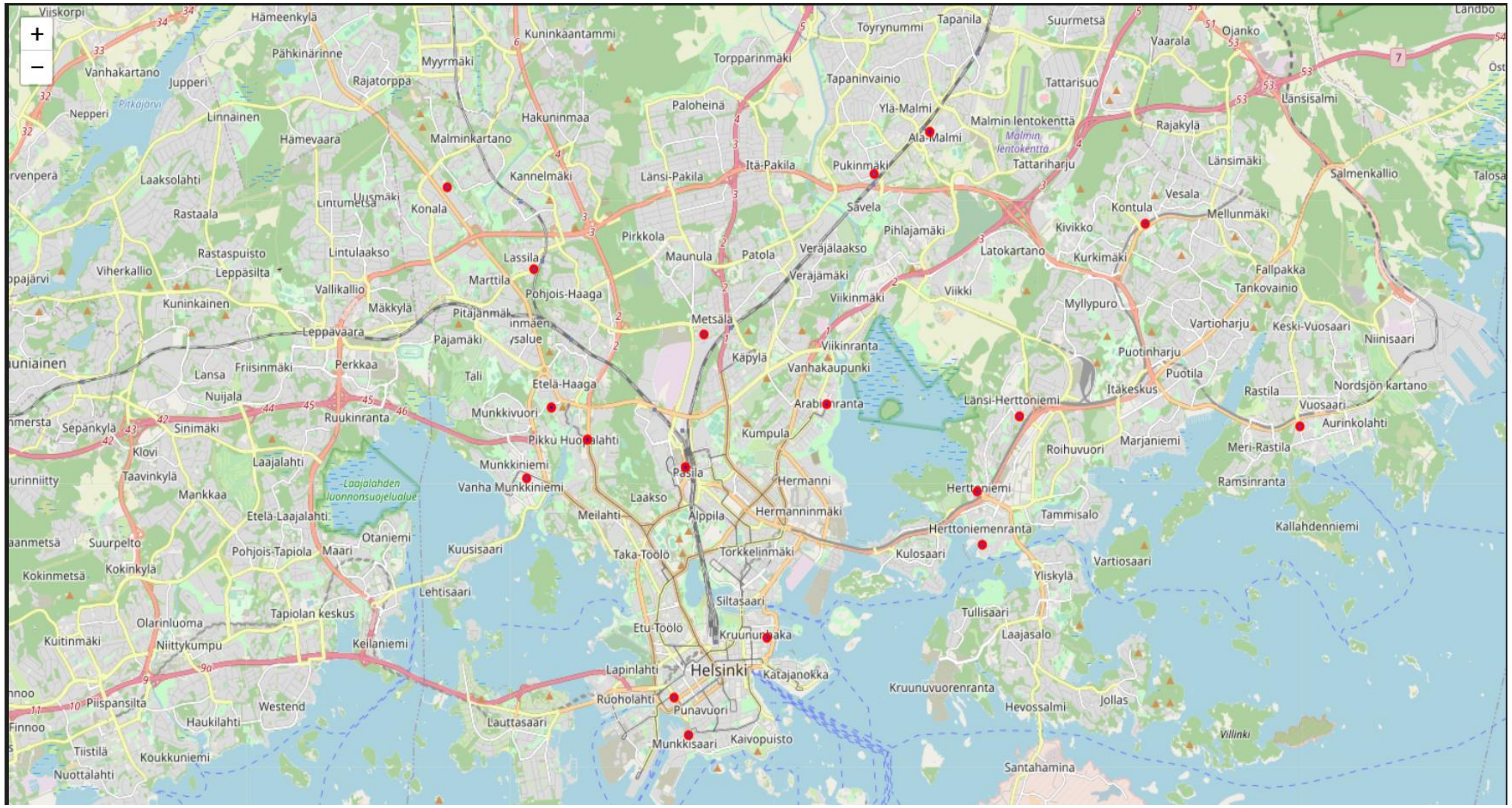
However, the code relies on the analysis and not on the values. Hence, hassle-free.

(***With regard to Clarification #3, date of run:*** *December 30th, 2020*)

RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

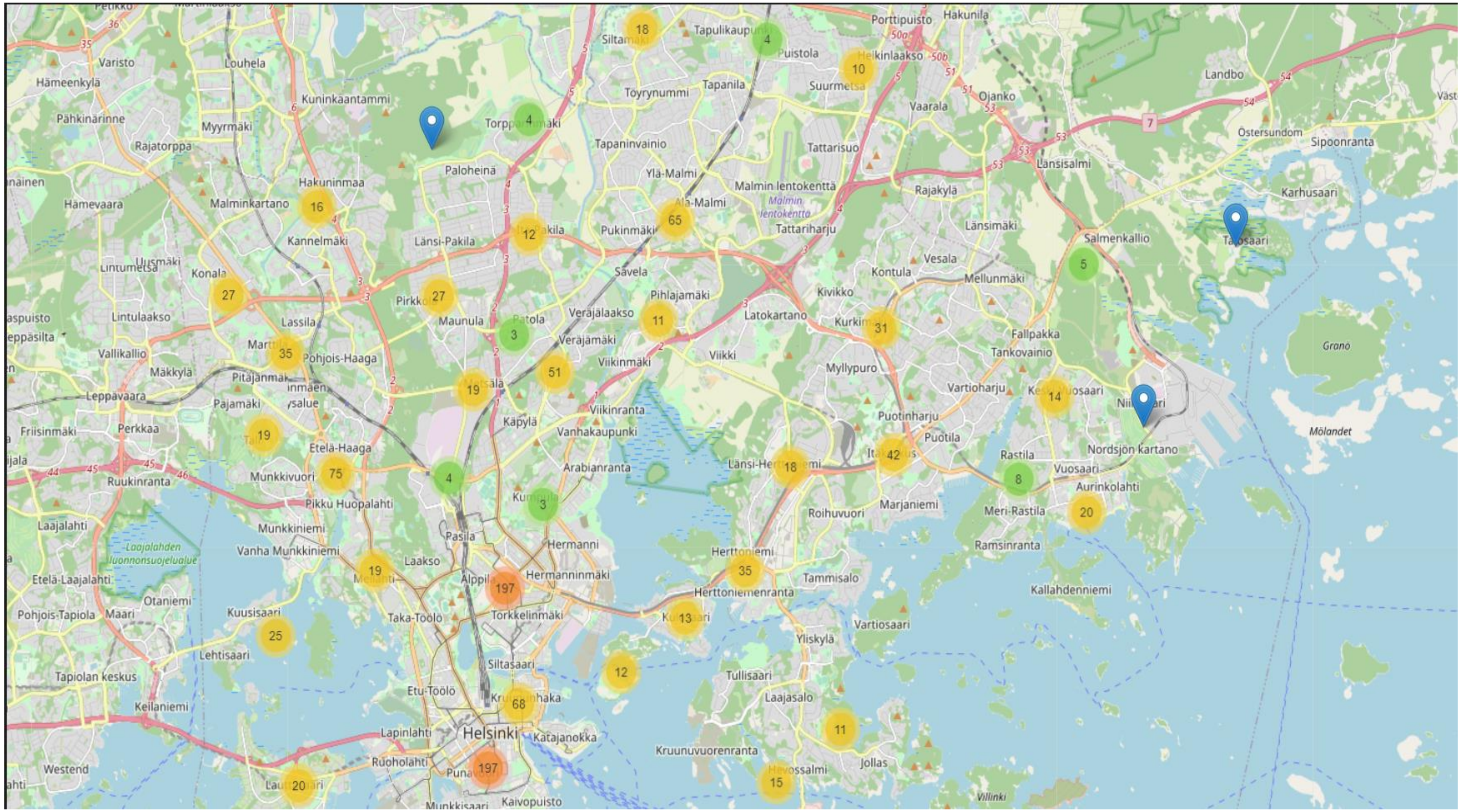
- A total of **1771 venues, 256 unique venue categories,** across **100 Districts** in Helsinki were extracted.
- Indian restaurants in Helsinki Districts, are identified as per Major Assumption 2.
- There are **22 Districts with Indian restaurants** across Helsinki and are visualized as Red Zones (Major-Competition Zones) using *folium*.



RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

- The venues were analysed for extracting popular venues based on the Top10 venue categories (by frequency of occurrence across Districts).
- Popular venues dataframe was reconstructed with detailed information about venues and Districts to identify popular venues and Districts by using geospatial data of the venues.
- Popular venues were visualized using *ClusterMap* in *folium*.



RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

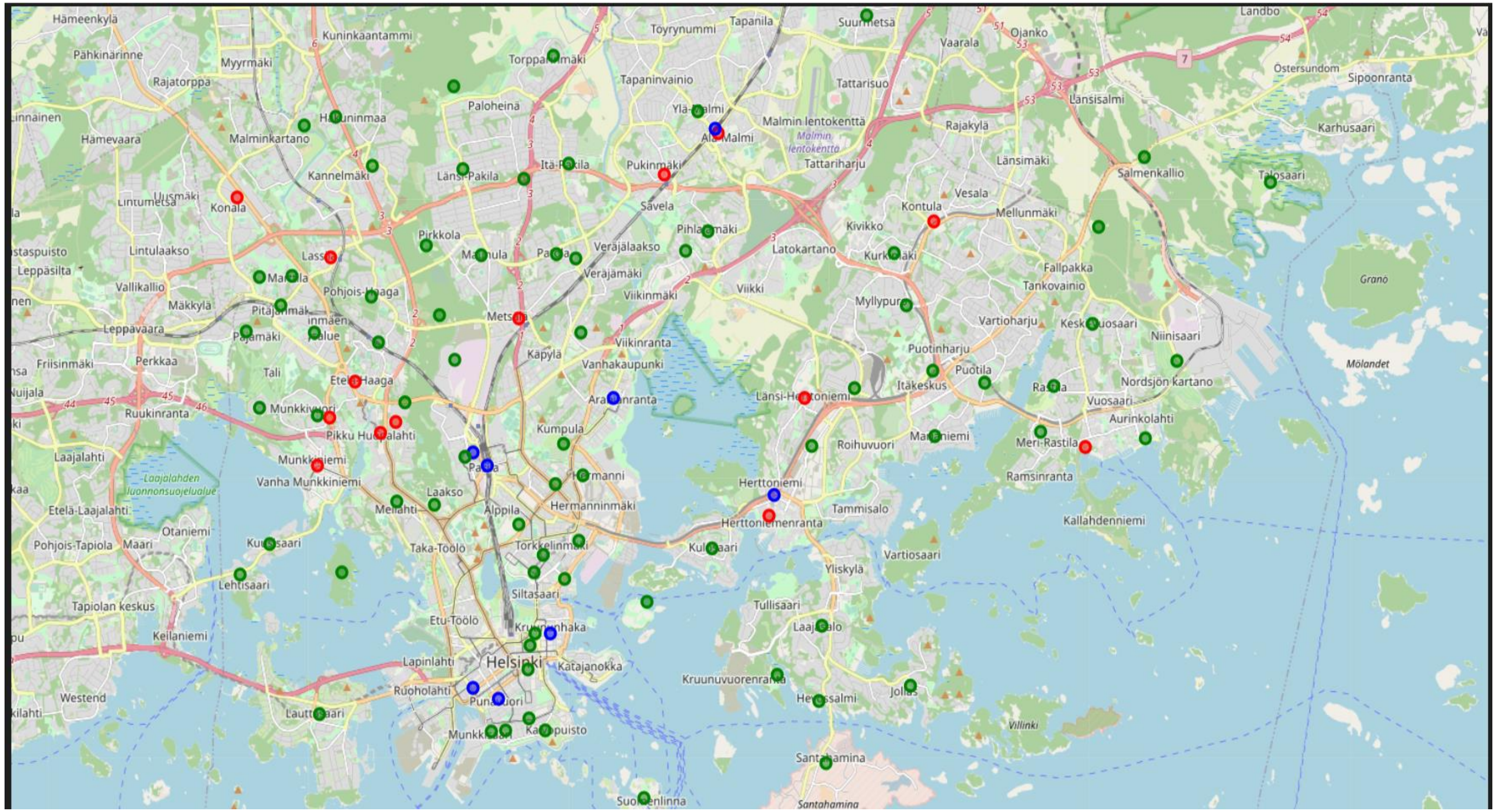
Total number of Districts were identified based on the presence/absence of Indian restaurants in the Districts as follows:

- Districts **WITH** Indian Restaurants: 22
- Districts **WITH** Indian Restaurants **in Top10 venue categories:** 14
- Districts **WITH** Indian Restaurants **NOT in Top10 venue categories:** 8
- Districts **WITHOUT** Indian Restaurant: 78

RESULTS & DISCUSSION

EXPLORATORY DATA ANALYSIS

- *Popularity Centres* (coordinates) were calculated from popular venues dataframe as the mean of latitudes and longitudes of venues grouped by Districts. (*Centroid is also known as mean, by definition, when a cluster is identified which is a District, in our case*).
- *Popularity Centres* were visualized as Helsinki city map using *folium*.



CONCLUSION

- Based on the analysis, the visualized *Popularity Centers* are tabulated as different levels of competition for establishing a chain of Indian Restaurants across Helsinki.
- Green Zones would be ideally considered to be the **"Benefit-Zones"** for the Chain of Indian restaurants.

Zone	**Competition**	**Indian Restaurant in District**	**Indian Restaurant in Top10 venue categories**
RED	Major	Yes	Yes
BLUE	Minor	Yes	No
GREEN	Low/None	No	No

DECLARATION

All the analysis and the assumptions are based on the data provided by the FourSquare API. Hence, I conclusively declare that these analysis could only be as accurate as the data extracted from the FourSquare API.