Utilising geo-spatial data to aid a new accounting company open a new office in the city of Athens, Greece.

Christos Nikolaos Zacharopoulos

Motivation & Research Question

Motivation (Case-study)

A new accounting company (*Th-squared*) seeks to open a business office in the capital of Greece, Athens. In order to succeed, *Th-squared* needs to eliminate the local competition of the existing accountants. Due to a certain flexibility offered by the Greek constitution, there is a certain overlap of the services that can be offered by accounting offices with legal/lawyer offices. Thus, legal offices are also a part of the competition pool. Lastly, an integral part of the accounting company, is the interaction with the local banks due to the nature of the provided services.

Research Question

Where should the business office of *Th-squared* be located?

Methods

Data Fetching

We utilize geographical data along with other associated attributes (e.g: total number of ratings) using the GOOGLE PLACES API to find the optimal location for the new business office of *Th-squared*. We use the data of the competitors (accounting offices and lawyers) as well as the local banks to find the optimal location for the new business office. The code used for this analysis is generic and can be used to fetch data from multiple cities.

preprocessing

The preprocessing of the data can be divided into two main sections:
online preprocessing (feature selection and extraction during data fetching) and offline preprocessing (outlier detection and feature selection).

Preprocessing

Online preprocessing

The following steps were taken to ensure that we capture and downlad the entirety of the available data.

1. The input to the Google-API was given in both *English* and *Greek*.
2. All possible combinations of the target profession. were given as input to the API fetcher (e.g: "Accountant in Athens", "Accountant Athens", "Accounting office Athens", "Accountant-Athens", etc.)
3. The text API entries were curated (space stripping and capitalization was removed).
4. All the not-found entries were set to NaN values for compatibility with the sklearn framework.
5. The duplicates were excluded and the data were stored as .csv files.

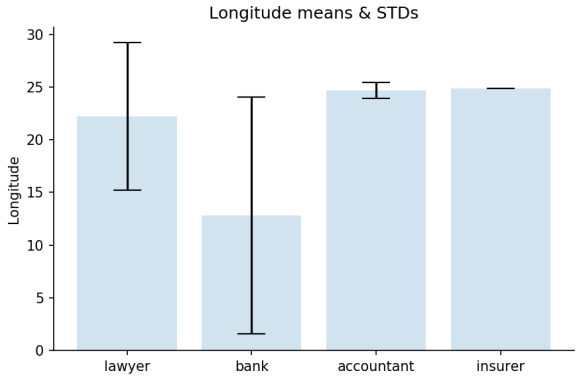
Preprocessing

Offline preprocessing

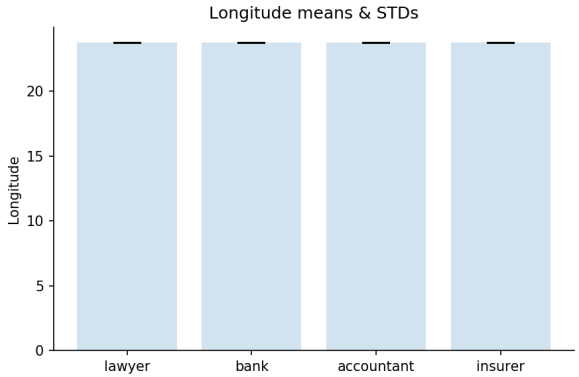
During the preliminary exploratory data analysis (EDA) we observed high standard deviation values (std) with respect to the longitude values of specific professions. This was a clear indication that the dataset contained outliers and required further curation. This was indeed a result of the fact that the fetching algorithm would fetch data from the vicinity of the IP from which the search was launched.
This issue was addressed using the following two approaches:

1. Removal of common coordinates detected in multiple Greek cities (Thessaloniki, Xanthi, Heraklion)
2. After step (1) was implemented, the longitude of the data were sorted per profession.. This allowed the analyst to identify a rapid increase in the longitude values indicating a complete change of geographic location. Using visual inspection, the analyst set a longitude threshold that allowed for removal of the outlying values.

Preprocessing



(a) Longitude means per profession. prior to outlier removal.



(b) Longitude means per profession. after the deviant values were excluded from the dataset.

Figure: Outlier detection and removal on the geospatial values of the professions.

EDA-feature extraction

This part of the analysis essentially focused on answering the following question:

Can we use the user-ratings as a reliable factor?

In other words, are the ratings an important feature? To answer this, we plotted the distribution of ratings per profession. for the city of interest (Athens). To ease the comparison amongst profession.s we plot the density of the distributions using a Gaussian kernel in figure 2. We observe a uni-modal distribution centered around 0. The center of the uni-modal distribution indicates that the ratings cannot be considered as a feature of interest and he uniformity of the distributions indicates that this is common amongst all professions.

EDA-feature extraction

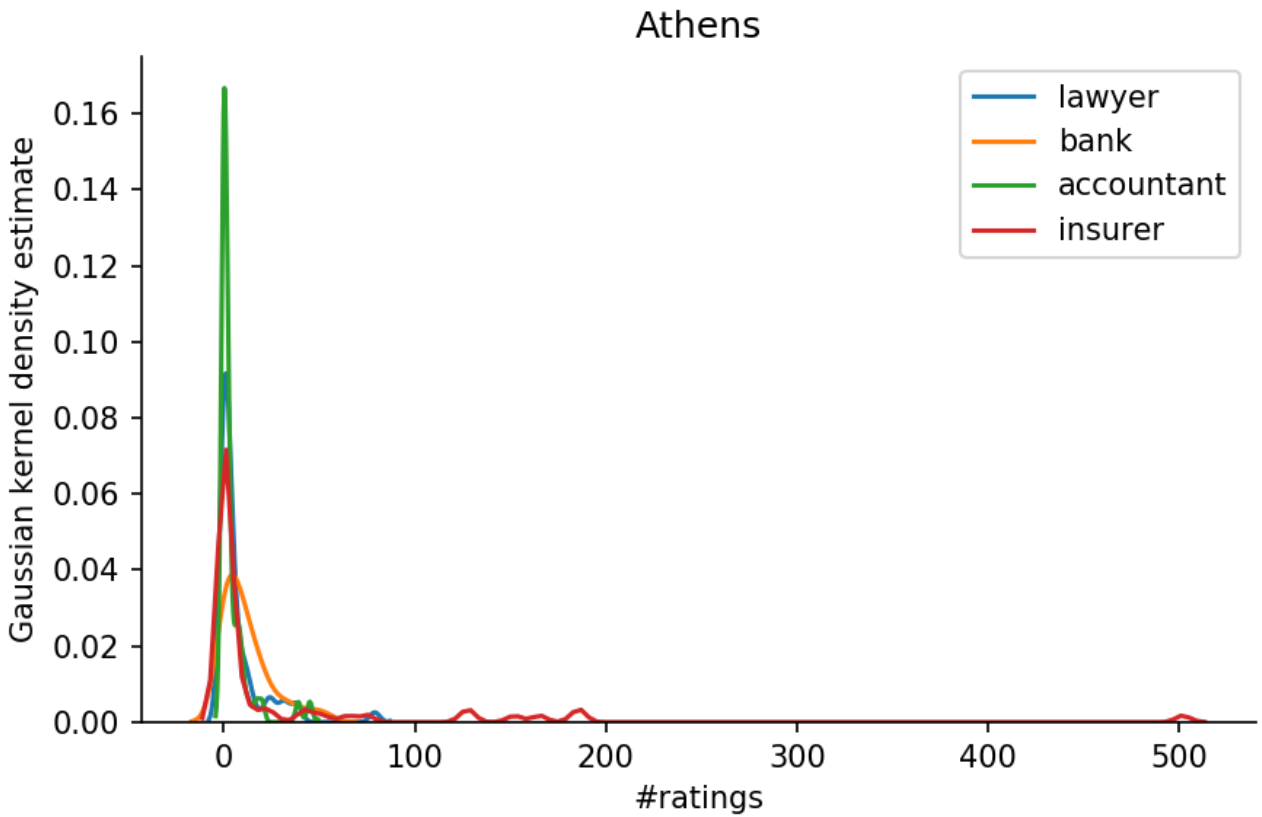


Figure: 2 Distribution of total number of ratings per profession. for the city of Athens. To facilitate comparison we only show the Gaussian kernel density estimate. We can observe a uni-modal distribution centered at 0 for all profession.s. Therefore, since the users did not rate in numbers these profession.s, the actuall rating cannot be used as a feature of interest and are thus discarded from the analysis.

Unsupervised Clustering

k-means clustering - Silhouette Method

The latitude and longitude coordinates for every office of each category were extracted and plotted in a scatter-plot (figure 3). **The goal of the analysis is to suggest a location using an unsupervised clustering approach using this dataset.** To that end, we applied a K-MEANS CLUSTERING per profession category and identified the corresponding centroid(s) per category. To identify the optimal number of k_s per office category, we utilised the *Silhouette Method*. The results can be observed in figure 4.We observe a high-density, unimodal distribution for the category of lawyers, a bimodal distribution for the category of banks and multimodal distributions for the categories of the insurers and the accountants.

Unsupervised Clustering

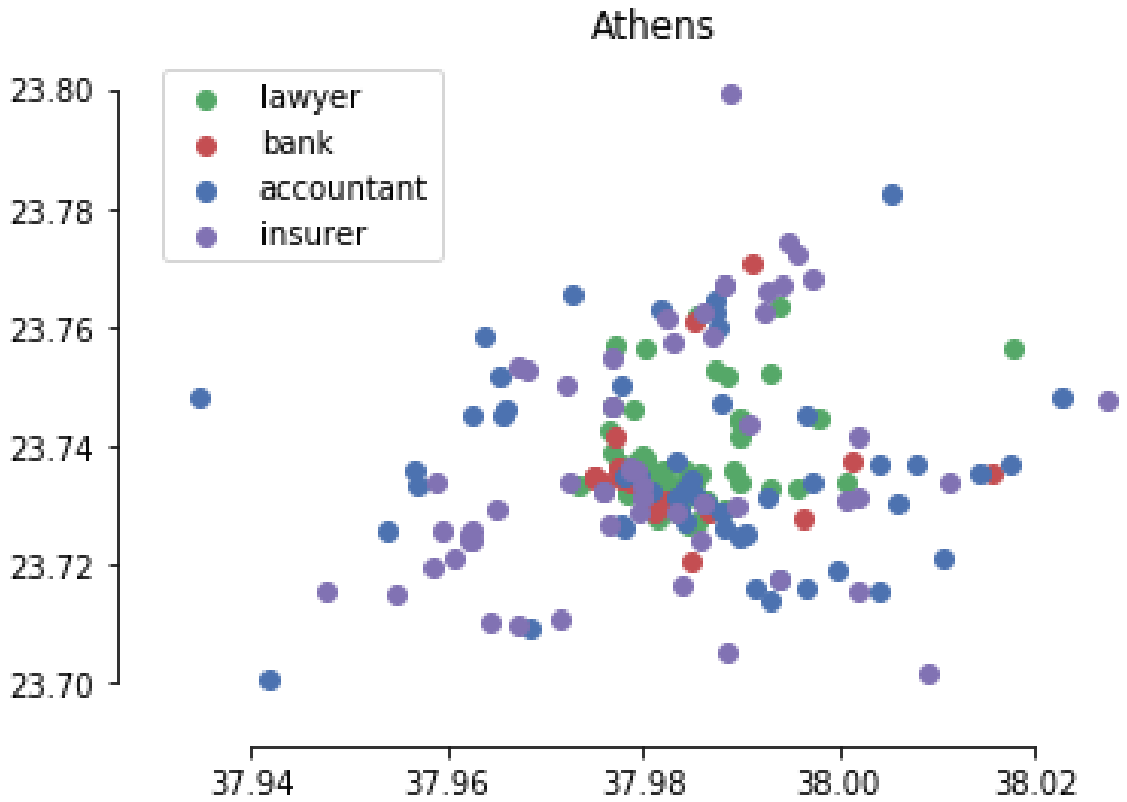


Figure: 3 Geo-spatial representation of the offices per company-category for the city of Athens. The scatter plot indicates longitude-latitude values. We can observe a rather spread-out spatial distribution for the insurer category, whereas the banks seem to be more concentrated in the city-center. These values will be used to identify the optimal region for the new office of *Th-squared*.

Unsupervised Clustering

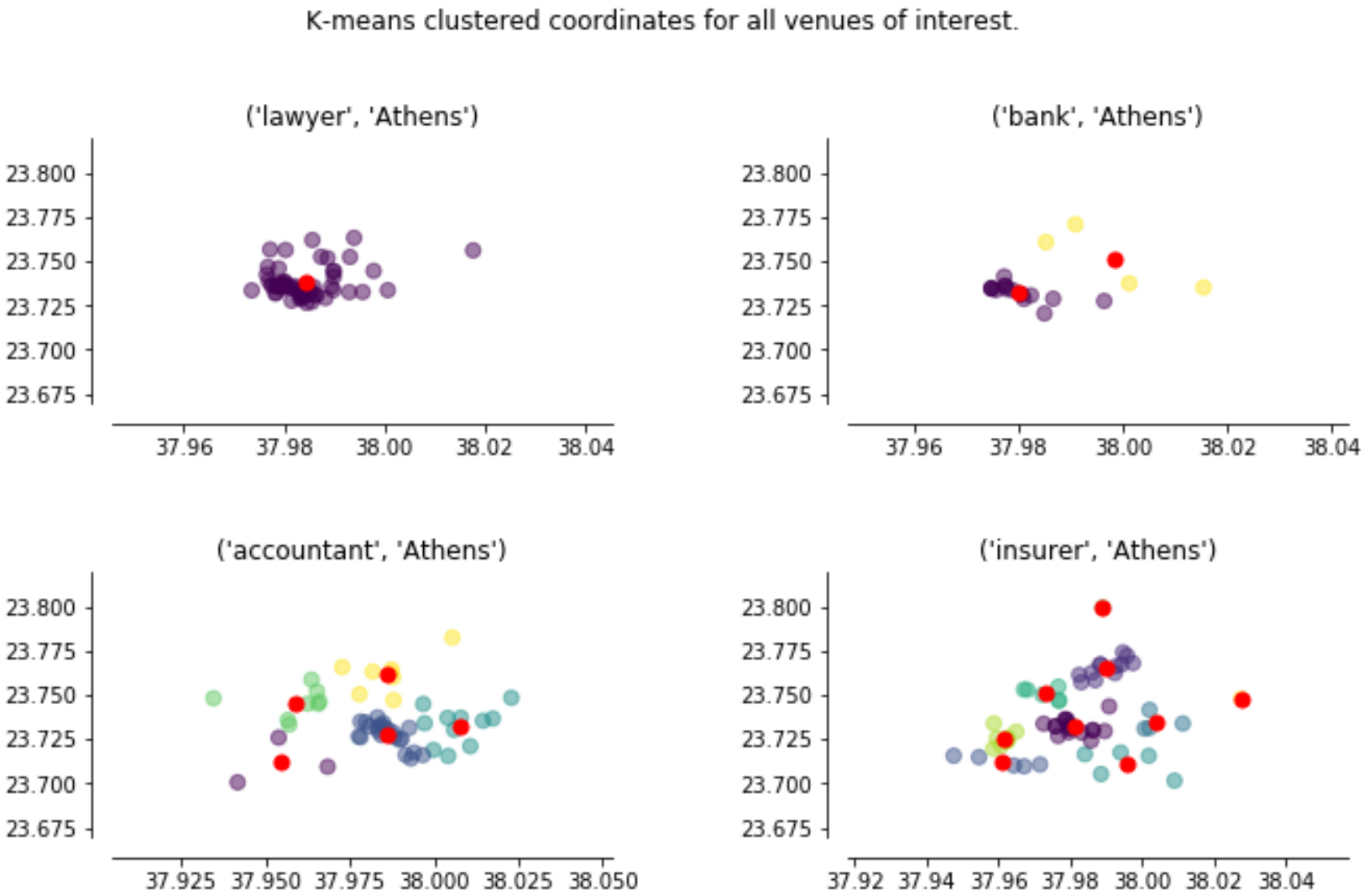


Figure: 4: K-means clustering of the geo-spatial data per office category.

Results

A *synergetic-antagonistic* view of the results.

At the final step of the analysis, we plotted the clustering centroids of figure 4 into a common plot to identify the optimal location for the new office. To facilitate the analysis, we further reduce the dimensionality of the dataset by splitting the professions into *synergetic* and *antagonistic* with the only synergetic element being the bank locations. This version of the dataset can be observed in figure 6. The location can be selected via visual inspection under the assumption that ***the new position must minimize the distance from a synergetic node while maximizing the distance from every other antagonistic node within the radius of this synergetic node.*** An example can be seen in 6 where a candidate starting point is selected as the mean distance of the three closest synergetic elements.

Results

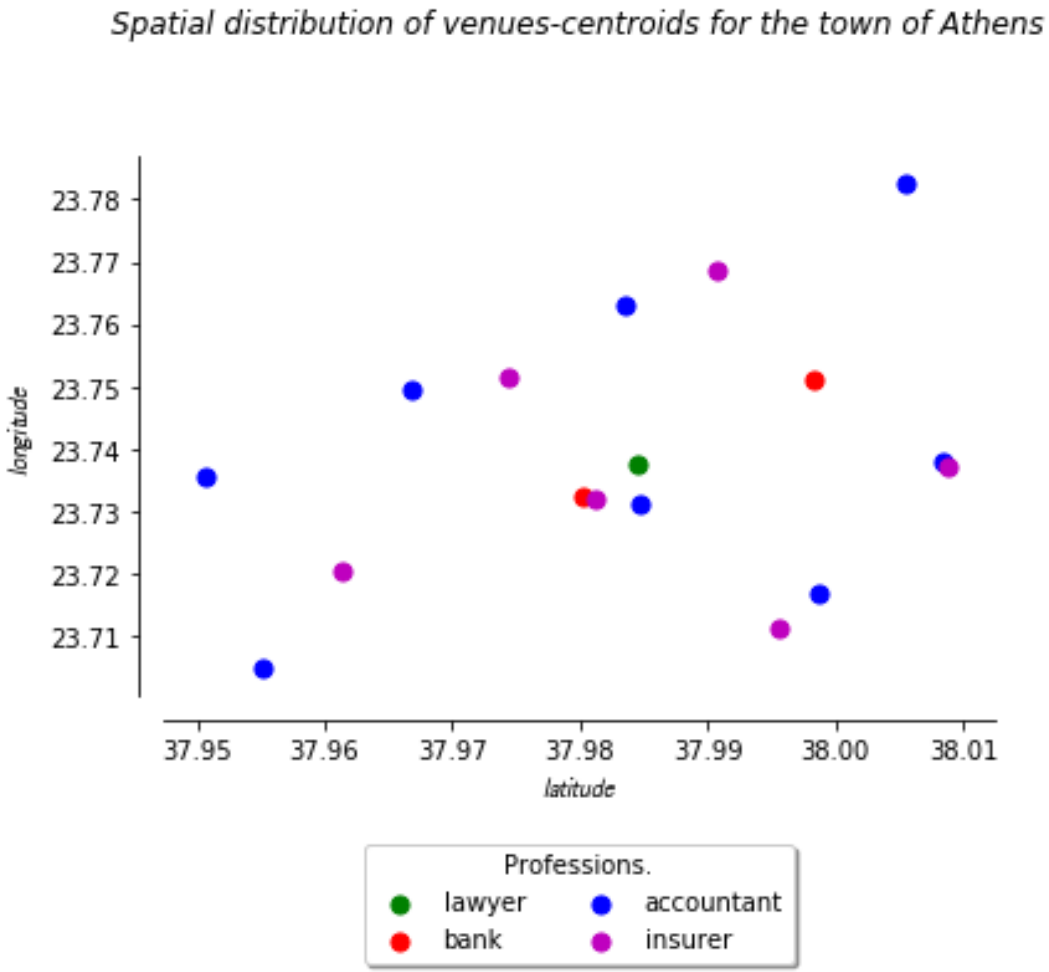


Figure: The k-means centroids per profession are plotted under common latitude and longitude values. At this stage, we have not yet split the data into synergetic and antagonistic.

Results

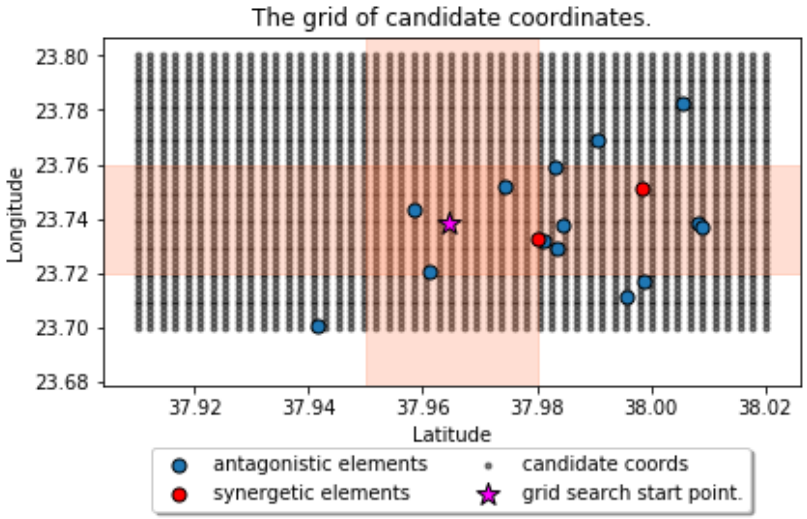


Figure: The k-means centroids per profession are plotted under common latitude and longitude values. As a final step of the dimensionality reduction process, we divide the data into synergetic and antagonistic. Corporate can then focus into a synergetic node (i.e: banks) and select a new location within a pre-selected radius centered around this node. The rationale of this selection should be the following: The new location must have a minimal distance from the synergetic node, while maximizing it's distance from all the other antagonistic elements that fall into this radius. A candidate starting search point is given as the mean latitude-longitude value of the three closest antagonistic nodes of a synergetic node (star marker).

Discussion

Future directions

In this case study we utilised geo-spatial data to suggest a location for a new office of the *Th-squared* company. In this approach, we split the data into synergetic (banks) and antagonistic (lawyers, accountants, insurers) elements and suggested a new candidate location under a *minimization of distance framework*. There are two elements that this approach did not take into account and can be altered in future releases.

- The first would be to assigning weights in both the antagonistic and synergetic elements and perform the last-step of the analysis in a weighted manner.
- The second would be a completely automatized way of selecting the new location via an implementation of a simple minimization algorithm.

Nevertheless, we think that this case study suffices to illustrate the proposed method and allow for multiple implementations of the method in various cities and/or scenarios.