Capstone Final

Assignment

- Introduction where you discuss the business problem and who would be interested in this project.
- Data where you describe the data that will be used to solve the problem and the source of the data.
- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
- Results section where you discuss the results.
- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
- Conclusion section where you conclude the report.

Report

Introduction

As a starting café entrepeneur it is difficult to find a location where to start a café, especially in crowded locations like the centre city of Amsterdam or Rotterdam. In this case study I will investigate using data science and the FourSquare API what a good location could be to open a new café.

Data

The main data source used in this study is the FourSquare API, which contains different kinds of venues (also known as "places of interest"). Primarily, I will use the explore API call to get the following data of the venues:

- Venue name
- Venue location (latitude / longitude)
- Venue category

Methodology

First step is to gather a set of as many venues as possible using the FourSquare API. Based on the category of the venues I will classify them manually as either positive or negative for a café. Using a machine learning algorithm called DBSCAN I will investigate the local densities of both positive and negative classified venues. After the clustering by DBSCAN, local clusters within the city centre can be identified with a relative high density of positive venues while not in a high density of negative classified venues.

Gaussian Mixture Clustering

A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by k ∈ {1,…, K}, where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters:

- A mean μ that defines its centre.
- A covariance Σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
- A mixing probability π that defines how big or small the Gaussian function will be.

Source: Wikpedia.

Further reading: [https://scikit-learn.org/stable/modules/mixture.html (https://scikit-learn.org/stable/modules/mixture.html)](https://scikit-learn.org/stable/modules/mixture.html)

Steps summarized

1. Fetch categories from the API and classify the categories
2. Fetch venues from the API based on their location around the centre of Amsterdam
3. Classify venues positively or negatively based on their category
4. Use the gaussian mixture clustering algorithm to find local positive clusters and negative clusters
5. Model the clusters using elipses to find overlapping and non-overlapping areas to find a location for the café

Results

0. Fetch categories from the API

Below category table shows the id, name and effect of a vanue from a café perspective. effect = 1 means positive, effect = -1 means negative.

In [13]:

```
category_df
```

Out[13]:

	id	name	effect
0	4d4b7104d754a06370d81259	Arts & Entertainment	1
1	4d4b7105d754a06372d81259	College & University	1
2	4d4b7105d754a06373d81259	Event	1
3	4d4b7105d754a06374d81259	Food	-1
4	4d4b7105d754a06376d81259	Nightlife Spot	-1
5	4d4b7105d754a06377d81259	Outdoors & Recreation	1
6	4d4b7105d754a06375d81259	Professional & Other Places	1
7	4e67e38e036454776db1fb3a	Residence	1
8	4d4b7105d754a06378d81259	Shop & Service	1
9	4d4b7105d754a06379d81259	Travel & Transport	1

1. Fetch venues from the API based on their location

Centre of Amsterdam location:

- latitude = 52.370216
- longitude = 4.895168

Below the result of the FourSquare API.

In [15]:

```
all_fetched_venues.head()
```

Out[15]:

	Search Category Id	Search Latitude	Search Longitude	Venue Name	Venue Latitude	Venue Longitude	Ven Distan
0	4d4b7104d754a06370d81259	52.370216	4.895168	De Kleine Komedie	52.367050	4.895938	3
1	4d4b7104d754a06370d81259	52.370216	4.895168	Pathé Tuschinski	52.366620	4.894706	4
2	4d4b7104d754a06370d81259	52.370216	4.895168	Perdu	52.368888	4.896399	1
3	4d4b7104d754a06370d81259	52.370216	4.895168	Frascati	52.370218	4.893686	1
4	4d4b7104d754a06370d81259	52.370216	4.895168	Zaal 1 | Pathé Tuschinski	52.366255	4.894519	4

2. Classify venues positively or negatively based on their category

With a left join on category id the effect of a venue is added.

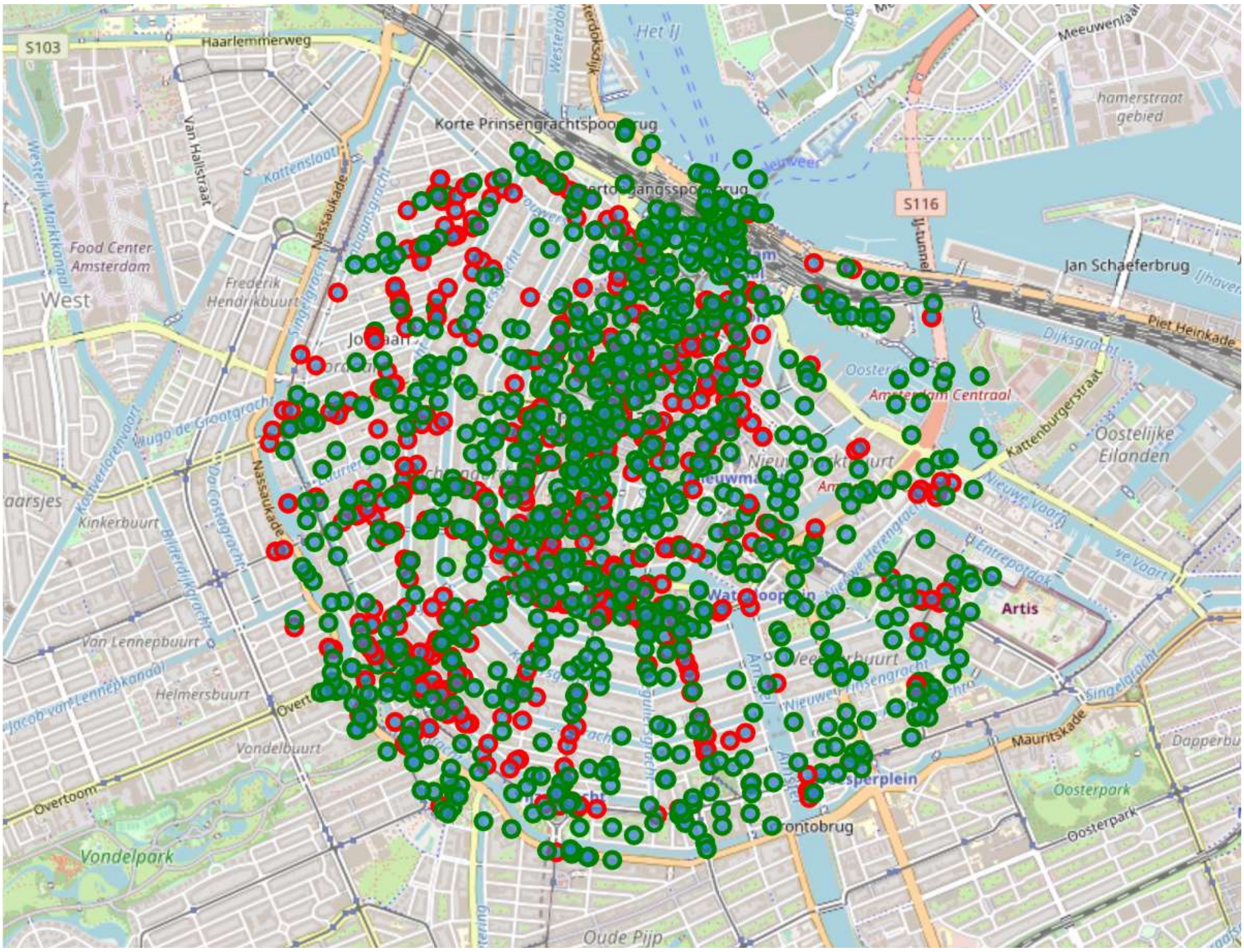
In [17]:

```
all_fetched_venues.head()
```

Out[17]:

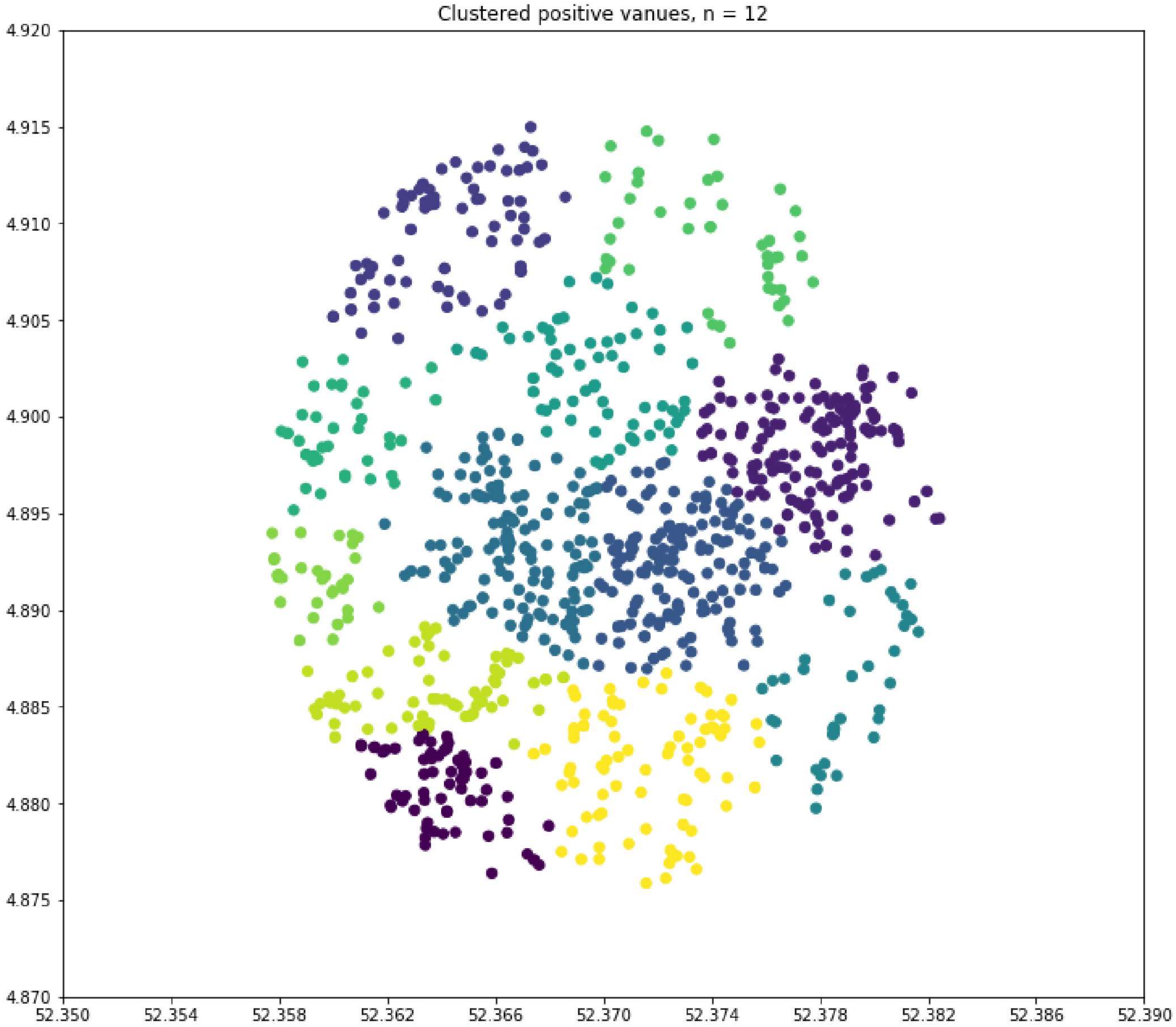
	Search Category Id	Search Latitude	Search Longitude	Venue Name	Venue Latitude	Venue Longitude	Ven Distan
0	4d4b7104d754a06370d81259	52.370216	4.895168	De Kleine Komedie	52.367050	4.895938	3
1	4d4b7104d754a06370d81259	52.370216	4.895168	Pathé Tuschinski	52.366620	4.894706	4
2	4d4b7104d754a06370d81259	52.370216	4.895168	Perdu	52.368888	4.896399	1
3	4d4b7104d754a06370d81259	52.370216	4.895168	Frascati	52.370218	4.893686	1
4	4d4b7104d754a06370d81259	52.370216	4.895168	Zaal 1 | Pathé Tuschinski	52.366255	4.894519	4

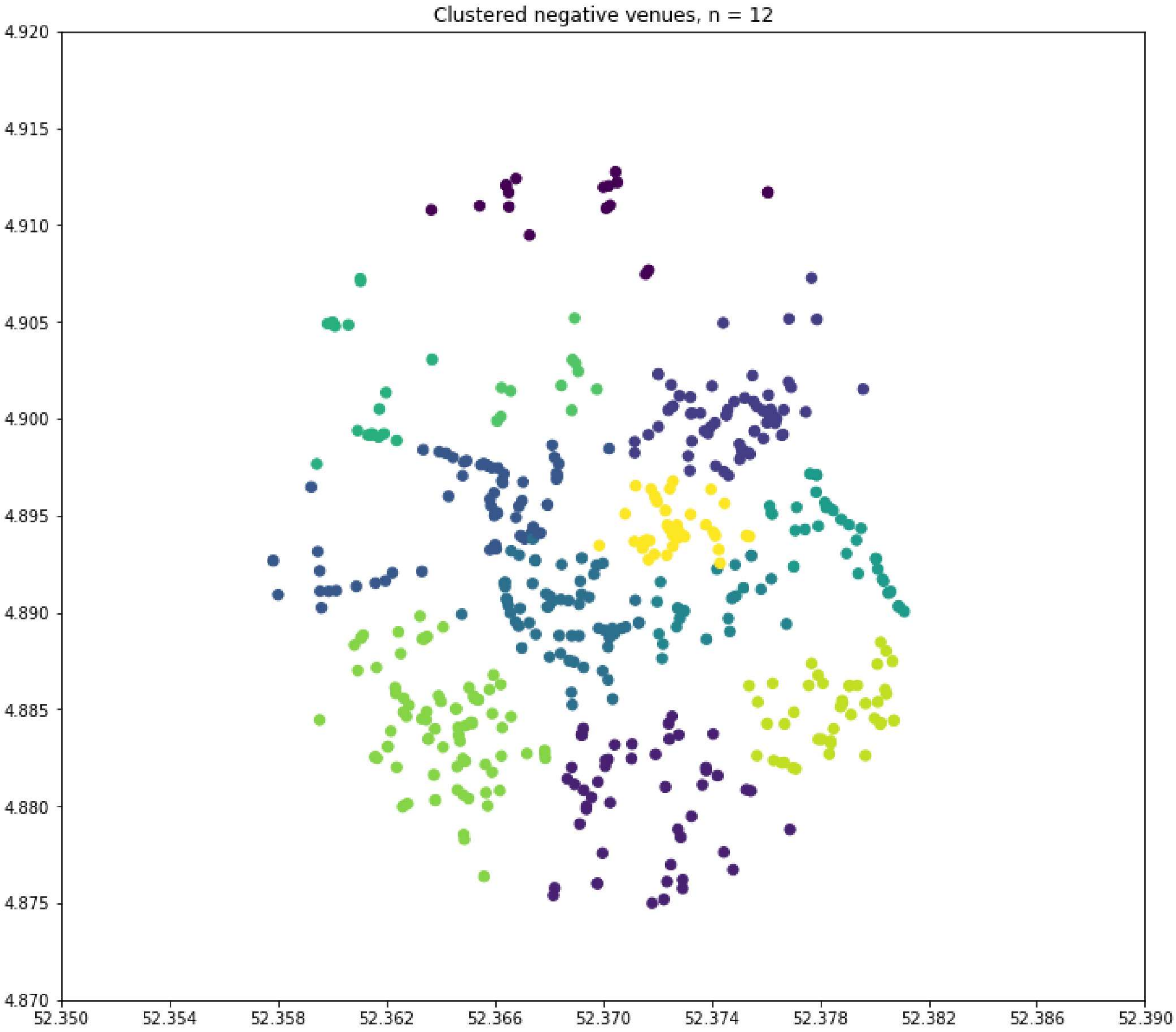
Based on the category we can draw all venues on the map including a colour indicating it's effect.


S103
Haarlemmerweg
Van Hallstraat
Westelijk Marktkanaal
Food Center Amsterdam
West
Frederik Hendrikbuurt
Kattenslootl
Nassaukade
Singelgracht
Korte Prinsengrachtspoorbrug
Het IJ
Westerdok
Westerdoksdijk
Meeuwenlaan
hamerstraat gebied
S116
IJ-tunnel
Jan Schaeferbrug
IJhaven
Piet Heinkade
Dijksgracht
Kattenburgerstraat
Oostelijke Eilanden
Oosterdok
Amsterdam Centraal
Jordaan
Hugo de Grootgracht
Kostverlorenvaart
Da Costagracht
Nieuwmarktbuurt
Nieuwmarkt
Nieuwe Vaart
Entrepotdok
aarsjes
Kinkerbuurt
Bilderdijkgracht
Nieuwe Herengracht
Artis
Waterlooplein
Van Lennepbuurt
Weesperbuurt
Amstel
Nieuwe Prinsengracht
Singelgracht
Jacob van Lennepkanaal
Helmersbuurt
Overtoom
Mauritskade
Dapperbu
Vondelbuurt
Weesperplein
Oosterpark
Torontobrug
Vondelpark
Oude Pijp

3. Use the gaussian mixture clustering algorithm to find local positive clusters and negative clusters

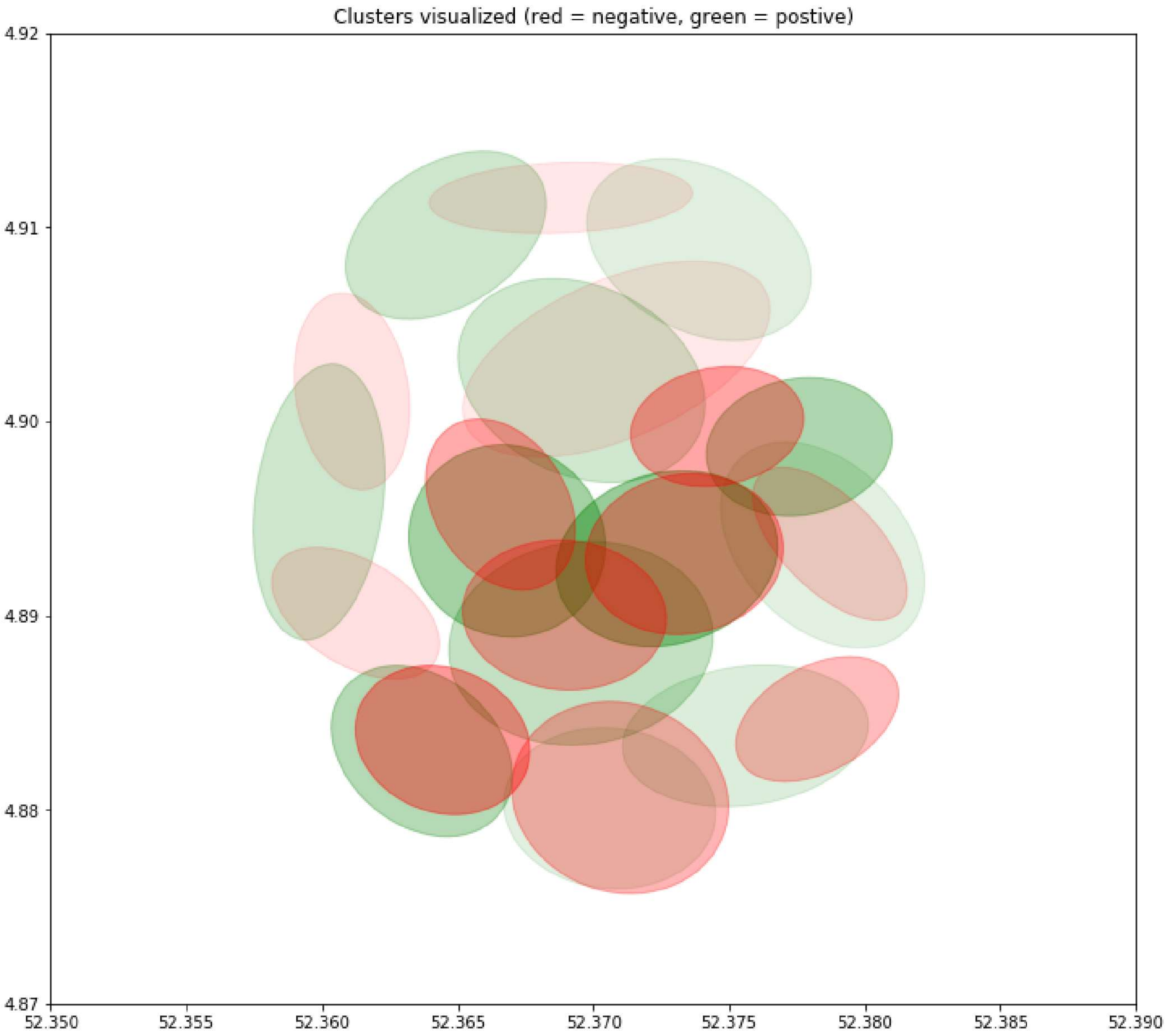
Using the gaussian mixture clustering algorithm the positive venues and negative venues are clustered in 12 clusters.





4. Model the clusters using elipses to find overlapping and non-overlapping areas to find a location for the café

Based on the centroids of the clusters and the local cluster covariance (spread), the venues can be modelled approximately by elipses. By adding a colour based on their effect, we can see clearly how the areas compare. The darker the colour of the elipsoid, the more venues it represents.



Discussion

When interpreting the results I would like to stress the following discussion points which could influence the results:

1. The FourSquare API only provides a limited amount of venues recommended, not all FourSquare venues available
2. Lot's of venues will not be on FourSquare
3. I modelled the venues using elipsoids, which might not represent how venues are actually distributed in the clusters.

Conclusion

Using the FourSquare API in combination with Gaussian mixture model, it is possible to model the venues as cluster area's and find the best areas for starting a café.