Data science capstone

ERIC MUIJS

Introduction

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.

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.

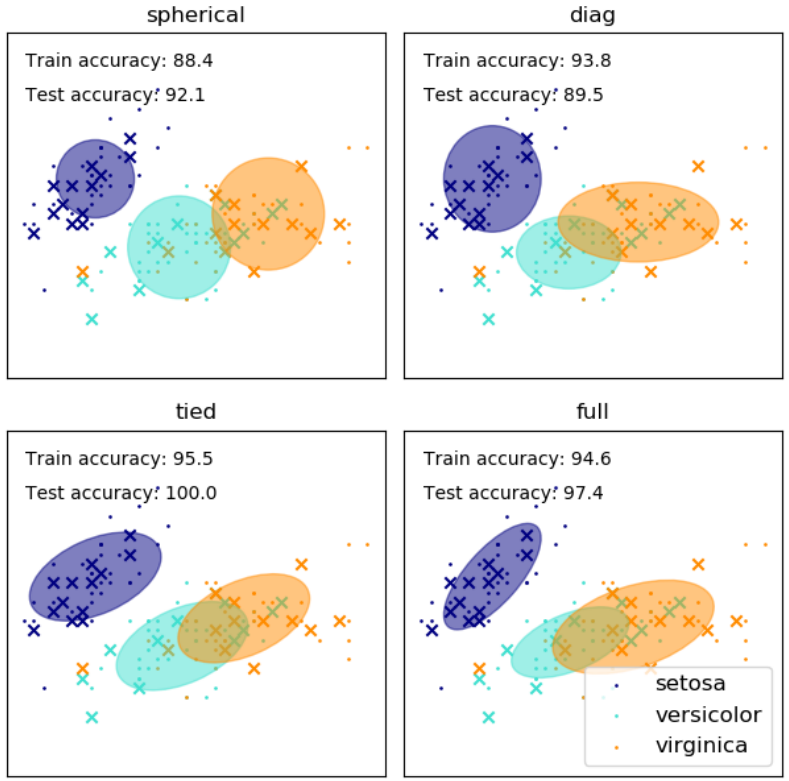
Gaussian Mixture Clustering

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:

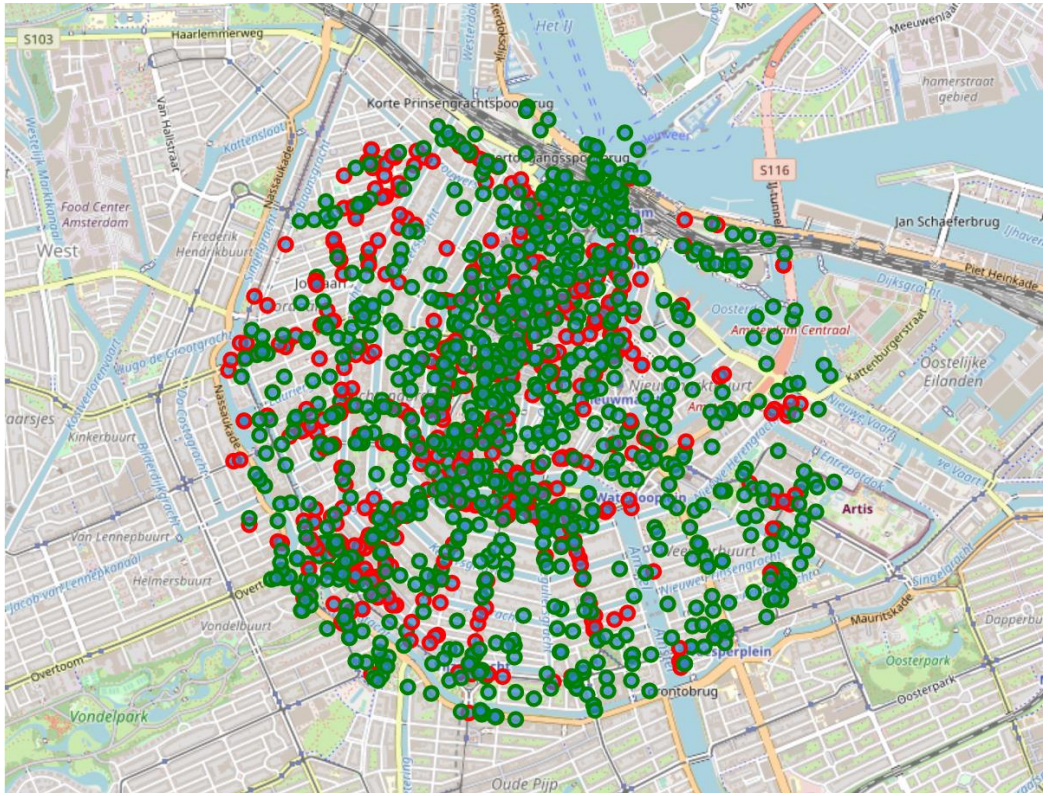
•A mean μ that defines its centre.

•A covariance Σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.

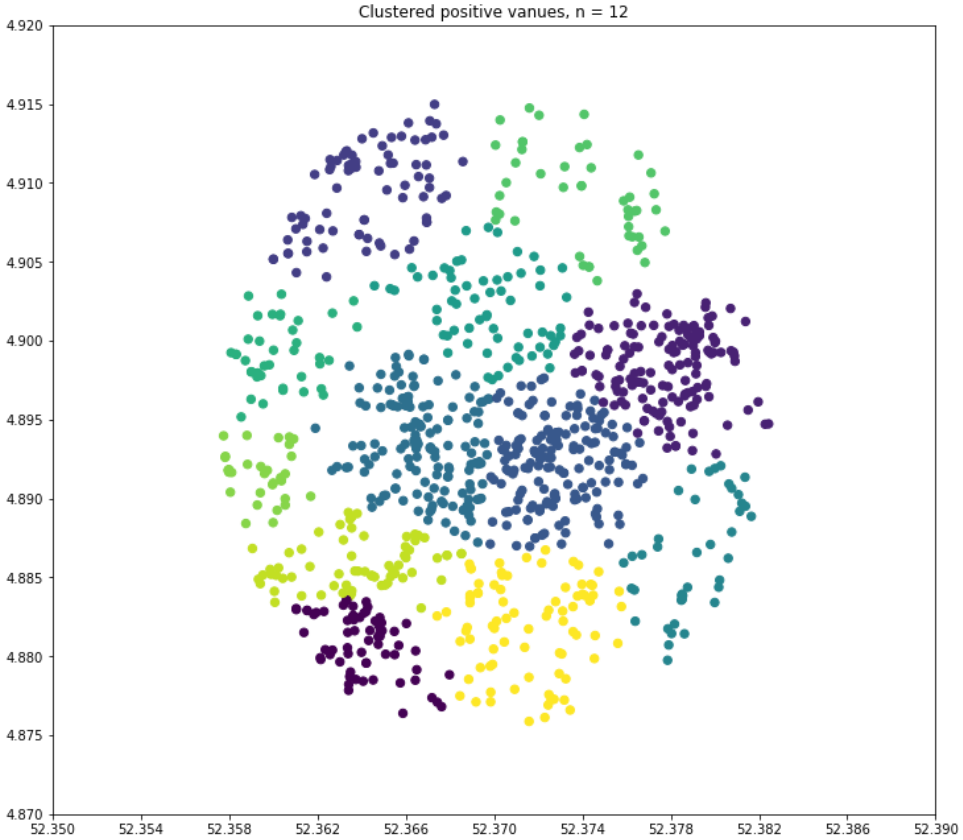
•A mixing probability π that defines how big or small the Gaussian function will be.

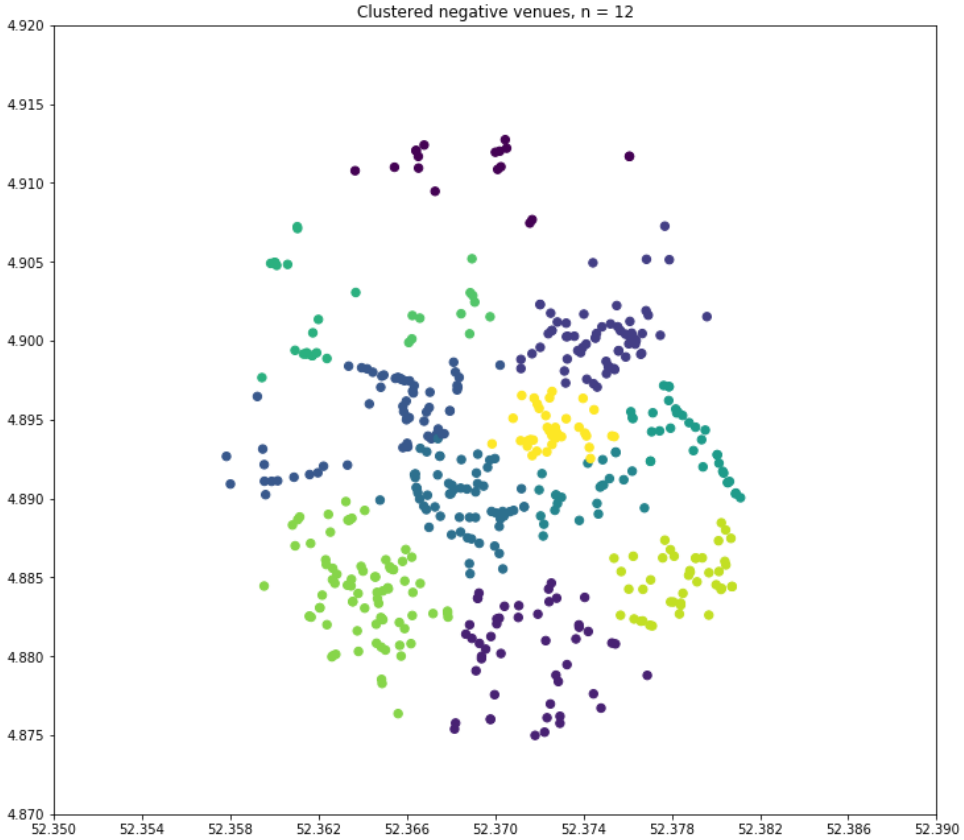


Venues

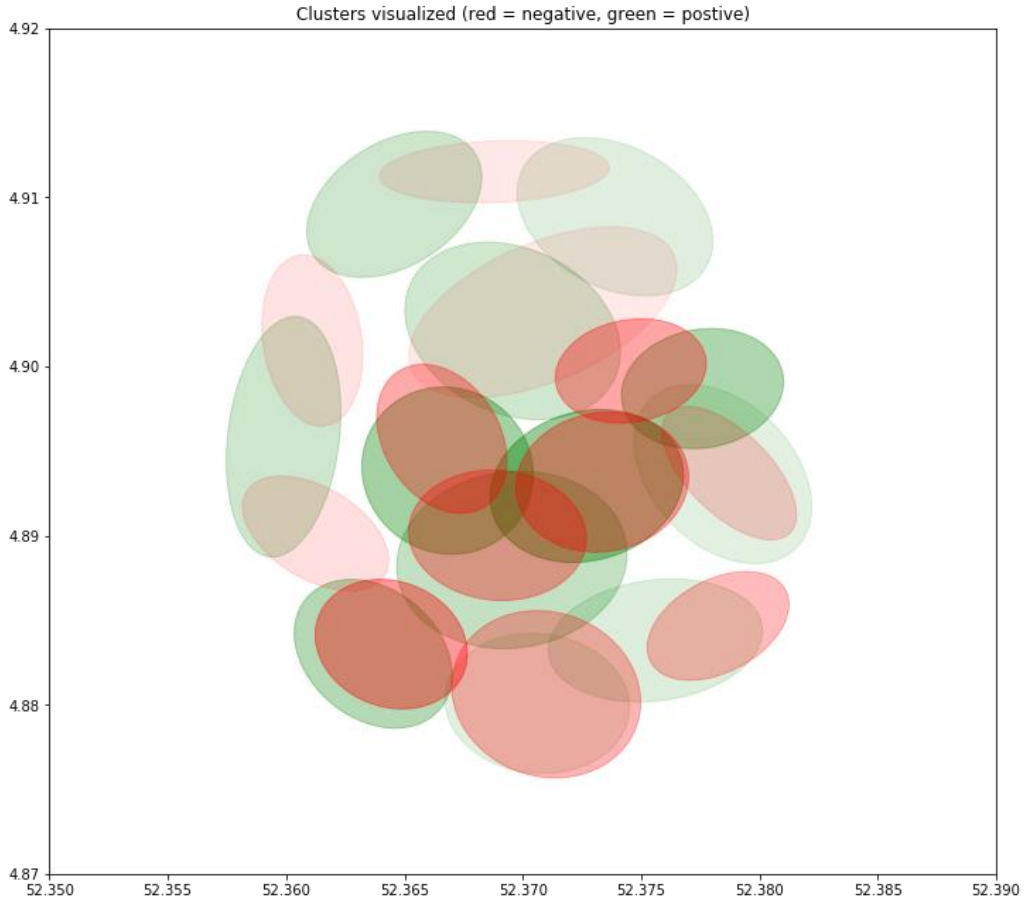

S103
Haarlemmerweg
Van Hallstraat
Korte Prinsengrachtspoorbrug
S116
Jan Schaeferbrug
Piet Heinkade
Food Center Amsterdam
West
Frederik Hendrikbuurt
Kinkerbuurt
Van Lennepbuurt
Helmersbuurt
Vondelbuurt
Vondelpark
Overtoom
Amsterdam Centraal
Oostelijke Eilanden
Artis
Mauritskade
Oosterpark
Oude Pijp

Clustering using GMC





Final result



Conclusion and discussion

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

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

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.