Capstone Project

The battle Of Neighborhoods

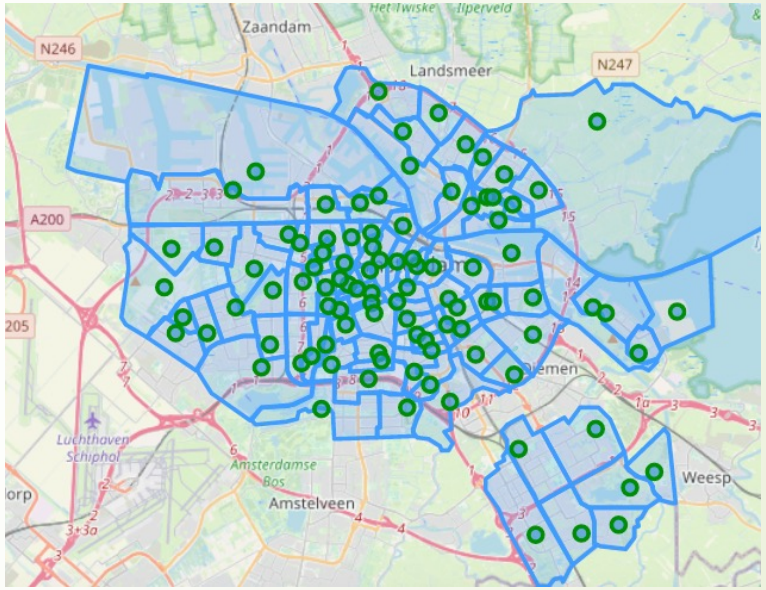
By: Evert Johannes Woudstra

Introduction: IBM certificate

- Final assignment of the IBM Professional Data science Certificate
  - Series of 9 Courses
  - Data science skills studied and applied under which:
    - Data Science Methodology
    - Data mining
    - Analysis and visualization with python
    - Artificial intelligence, Machine learning
- Learning objectives
  - Leverage location data provided by Foursquare and the Data of Amsterdam website
  - Applying data science skills in machine learning and data visualization

Introduction: Context of the project

- Amsterdam, What a city!
  - 872.922 inhabitants from 177 different nationalities
  - Most bars and pubs of any city in the Netherlands
  - Trend since 2010: number of restaurants increasing
  - Amsterdam increasingly popular among tourists
- Business problem
  - Where to open a new restaurant
- Stakeholder \ audience
  - Entrepreneurs
  - Brokers for properties
  - Data Analysts interested in Geo Pandas



Data Description: Data sources

- Location data neighborhoods
  - The Amsterdam data website (https://maps.amsterdam.nl)
  - Data format in GeoJson
  - Polygon objects describing the borders of each neighborhood

- Venue Data: The Foursquare database
  - Foursquare (https://foursquare.com) build and maintains a massive dataset of accurate location data
  - Foursquare uses the RESTful API
  - Search can only be made using radius or squared region which will present difficulties finding venues in he above example of neighborhoods

Data Description: feature selection

- Foursquare queries will be made using the following categories
  - Food: Find all venues for venues to assess the indirect competition
  - Italian Restaurants: Find all venues for venues to assess the direct competition
  - Arts and Entertainment: attractive venues for tourists \ possible business partner
  - Nightlife: indicated for popularity for locals and tourists
  - Travel and Transport: indication how reachable a potential neighborhood is
- The query results will be cleaned preprocessed and ordered to achieve a dataset ready to be analysed using machine learnings

Methodology: assiging venues

- Venue assignment by foursquare using a search radius of 500 m results in:
  - Presence of duplicates
  - Wrongly assigned venues

- Venue assignment done by GeoPandas
  - delivers a much more precise result
  - all venues are located within the defined neighbourhood borders

Methodology: One hot encoding

- Why we need one hot encoding
  - Machine learning algorithms won't work nicely with categorical values
  - Conversion of categorical values to numerical ones

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant
0	Amstel III/Bullewijk	0	0	0	0	2	0
1	Apollobuurt	0	0	0	0	0	0
2	Banne Buiksloot	0	0	0	0	0	0
3	Bedrijventerrein Sloterdijk	0	0	0	0	0	0
4	Bijlmer Centrum (D,F,H)	0	0	0	0	2	0

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category	Venue Category id
0	Burgwallen-Oude Zijde	52.372084	4.896919	Bridges Restaurant	4b8a4fdaf964a520076832e3	52.370818	4.895087	Seafood Restaurant	4bf58dd8d48988d1ce941735
2	Burgwallen-Oude Zijde	52.372084	4.896919	Eetcafé Bern	4a26ff69f964a520857f1fe3	52.372575	4.900645	Swiss Restaurant	4bf58dd8d48988d158941735
3	Burgwallen-Oude Zijde	52.372084	4.896919	Kaagman & Kortekaas	55e5ff02498e9eb3a234f62c	52.374878	4.892455	French Restaurant	4bf58dd8d48988d10c941735
4	Burgwallen-Oude Zijde	52.372084	4.896919	La Zoccola del Pacioccone	5648c396498edda4852d4c23	52.375297	4.893965	Italian Restaurant	4bf58dd8d48988d110941735
5	Burgwallen-Oude Zijde	52.372084	4.896919	The White Room	5718cbbd498efa35585946b1	52.373178	4.894687	French Restaurant	4bf58dd8d48988d10c941735

Machine learning: k-means clustering

- What is k-means
  - technique for calculating the similarity and dissimilarity of points in a given dataset
  - unsupervised machine learning being capable to handle unlabeled data
- Why we want to use it
  - separate the neighborhood based on unlabeled features
  - finding and those similar neighborhoods which will have the ideal feature values for starting a restaurant
- Those ideal feature values are:
  - Total number of restaurants: low, meaning less indirect competition
  - Total number of Italian restaurants: low, meaning less direct competition
  - Arts\entertainment venues: high, opportunities to form business partners and attractive for tourists
  - Nightlife: High, opportunities to form business partners attractive for the target customer being in the area.
  - Transport venues: high, for a restaurant to become a success the neighborhood needs to be reachable.

Methodology: Normalization

- Why do we need to normalize?
  - The K-Means algorithm is sensitive for unscaled data.
  - Assure all features will have an equal impact on the clustering
- Which method ?
  - The most commonly used normalization method is Min\Max scaling (eq 1).
  - Applying the eq. 1 on the feature data will result in a scaling between 0 and 1 where:
    - 0 = min value
    - 1 = max value

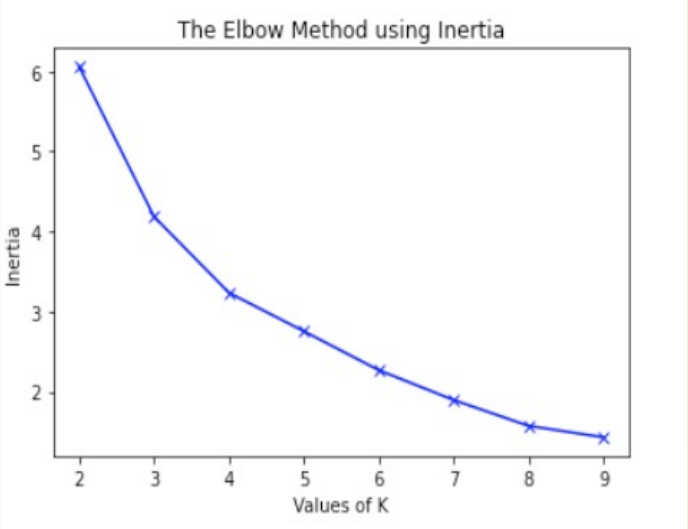
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{(eq. 1)}$$

Methodology: parameter selection

- K-means clustering most powerful parameter: value for K
- Which K is optimal? IT will be between two extremes:
  - K= 1: there will be only one cluster containing all datapoints.
    - Cluster distance is maximized (no other cluster available)
    - No separation of data points.
  - K= max (99 neighborhoods)
    - Inter cluster separation minimised (one datapoint per cluster
    - No clustering at all; every cluster contains one data point just like the dataframe
  - Optimal K? Iterate over a range of possible K's and determine the 'clusering quality using Elbow method and Silhouette score

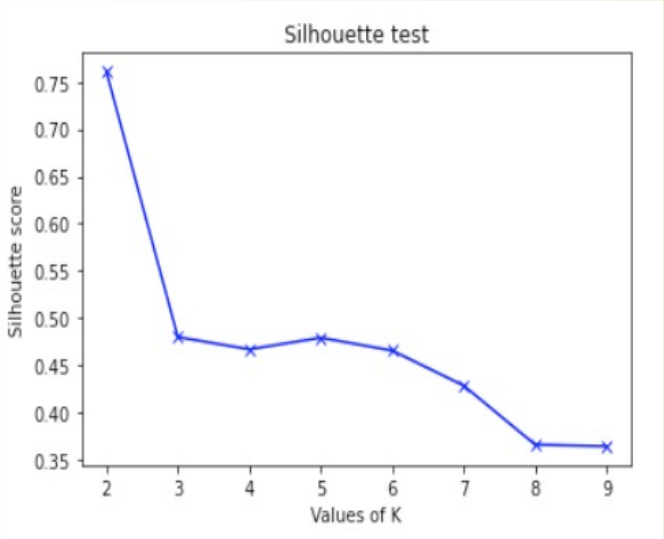
Statistical testing: Elbow method

- elbow method: multiple runs of the K - means algorithm over a range of values for K = 1-10
- Calculate inertial
  - Inertial is the sum of the distances of data points to their closest cluster center
- Find the elbow point in the graph
  - Elobow point is where the graph suddenly change in steepness



Statistical testing: Silhouette test

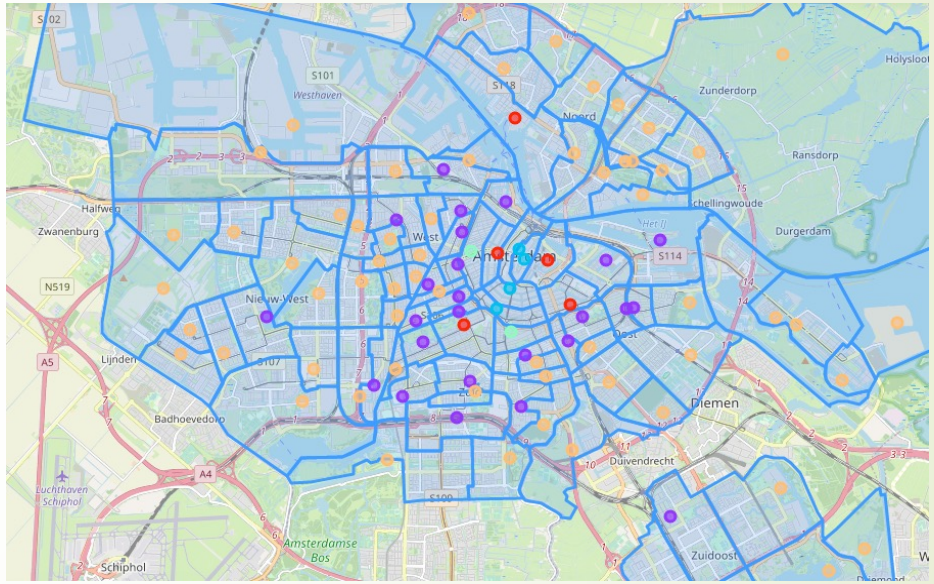
- The Silhouette method:
  - measures how similar a datapoint is to:
    - own cluster (cohesion)
    - other clusters (separation)
- Optimal value?
  - Find local maximum in graph
  - Values will be between -1 and 1
- Which value of k? 5 could be a good choice based on both methods



Results: clustering the neighbordhoods

Clusters are assigned by following colors:

- Purple: Cluster 1, containing 25 neighborhoods
- Dark blue: Cluster 2, containing 4 neighborhoods
- Light blue: Cluster 3, containing 2 neighborhoods
- Orange: Cluster 4, containing 62 neighborhoods
- Red: cluster 5, containing 6 neighborhoods



Results: promising clusters

- **Cluster 1:** These are neighborhoods
  - mainly located outside of the touristic citycenter
  - less interesting to starting a restaurant for tourist and dayvisitors.
  - have relatively good transport possibilities
  - low number of italian restaurants.
  - *'Vondelbuurt'* and *'Zuidas' to be considered*
- **Cluster 5:** All is in balance
  - located close to city center
  - many options for public transport
  - many venues in the category's arts\entertainment and nightlife.
  - 'Museum kwartier 'and 'Weesper buurt' to be considered

	Neighborhood	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
5	Haarlemmerbuurt	30.0	3.0	6.0	16.0	18.0
9	Oostelijke Eilanden/Kadijken	25.0	4.0	7.0	13.0	17.0
12	Spaarndammer- en Zeeheldenbuurt	31.0	4.0	10.0	19.0	8.0
13	Staatsliedenbuurt	18.0	6.0	2.0	8.0	8.0
15	Frederik Hendrikbuurt	15.0	2.0	0.0	8.0	10.0
16	Da Costabuurt	31.0	8.0	0.0	7.0	7.0
17	Kinkerbuurt	36.0	4.0	4.0	11.0	9.0
19	Helmersbuurt	24.0	3.0	4.0	8.0	21.0
20	Overtoomse Sluis	30.0	4.0	1.0	13.0	8.0
21	Vondelbuurt	12.0	3.0	5.0	6.0	23.0
23	Landlust	26.0	2.0	3.0	6.0	16.0
37	Osdorp-Oost	15.0	1.0	8.0	6.0	10.0
46	Zuidas	37.0	5.0	8.0	14.0	25.0

	Neighborhood	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
2	Grachtengordel-West	29.0	8.0	13.0	23.0	32.0
4	Nieuwmarkt/Lastage	31.0	11.0	14.0	16.0	33.0
8	Weesperbuurt/Plantage	27.0	7.0	42.0	14.0	33.0
53	Museumkwartier	50.0	11.0	27.0	21.0	49.0
88	Noordelijke IJ-oevers West	17.0	0.0	15.0	11.0	23.0
92	Amstel III/Bullewijk	17.0	0.0	23.0	20.0	6.0

Discussion: observations Foursquare

Foursquare API: Retrieving data using the Foursquare database might seem straight forward at first use.

Observations and Pitfalls

- Using categories can focus your search but still allow other unwanted categories
- Duplicate venues have to be deleted
- venues need to be reassigned which greatly improves accuracy.
- The free account : limited possibilities to perform queries,
- Overall database is rich of valuable data and Foursquare has proven to be reliable

Discussion: K-means clustering

Clustering with K-Means: A popular way to cluster data and the used library Sklearn works very well. Some Observations can be mentioned.

- value for K is still one of the hardest parameters to
- Other parameters might be investigated on their effect on clustering
- Evaluate independence of features is still an area to be considered for future development

Discussion: recommandations

The following recommendations can be made in further addressing the business question.

- review specific venues inside the targeted neighborhood on ratings and popularity during traffic hours
- Profile potential customers in the neighborhood; what could be their spending budged?
- Review of menus \ concepts of most popular competition
  - Diversify
  - Find gaps in portfolio

Conclusion

- Mlearning has proven to be an effective tool for discovering insights from large amounts of data.
- The Foursquare database
  - interesting data source but has its limitations when it comes to data precision
  - This precision can be enhanced by using Geopandas
- Using K-means clustering we have been able to cluster the neighborhoods based on their common features.

The following neighborhoods have been highlighted:

- *Vondelbuurt* and *Zuidas (cluster 1)*
- *Museum kwartier* and *Weesper buurt* (Cluster 5).

Thank you all for your attention