4/1/2021

Capstone Project

The battle Of Neighborhoods

Evert Johannes Woudstra
EJ.WOUDSTRA@GMAIL.COM

Table of Contents

***

To My Loving Wife Nicoletta and My sweet Daughter Alisia

***

1 Introduction

This project is the final assignment of the IBM Professional Data science Certificate. This certificate consists of a series of 9 courses in which data science skills, including Data Science Methodology, data mining and analysis with python has been studied and applied.
The main objectives for this project are the following:

- leverage location data provided by Foursquare and the Data of Amsterdam website
- applying data science skills in machine learning and data visualization

The results will be presented in this report as well as in a presentation and a Jupyter notebook published on Github.

1.1 Context of the project

Amsterdam is the capital city of the Netherlands and is famous for its multicultural identity. With 872.922 inhabitants from 177 different nationalities, it belongs to one of the most divers cities of the world. The city has the largest number of bars and pubs of any city in the Netherlands, although the number is declining over the last decade. Small traditional pubs are transitioning to more fancy venues targeting the market of “hipsters” and the modern millennial customer seeking for experiences and high-quality food.
The opposite trend is visible in the category of restaurants. Since 2010 the number of restaurants is increasing not only limited to the old city center but also surrounding neighborhoods. New forms like food markets as “Rollende Keukens” and concepts restaurants are being developed over the last few years. The growing popularity of Amsterdam worldwide and the annually increasing number of tourists indicates good prospects for opening a new restaurant concept.

1.2 Business problem

Opening a restaurant in a city like Amsterdam is a complex project. Although a good food concept can be a profitable and fulfilling business, reality is that many restaurants fail during their first year, mainly due to a lack of planning and having a sound business plan.
In the business plan many topics like funding, choosing a concept, advertising and logistics and choosing an optimal location has to be carefully considered.
The scope of this project is addressing the question what the optimal location could be for opening an Italian restaurant in the city of Amsterdam with a concept of choice.

1.3 Audience and stakeholders

Because of the business problem and the approach in using machine learnings this project could be potentially interesting to the following groups

- Entrepreneurs who want to start a food business in Amsterdam
- Brokers to advise people on buying properties based on their business plan

The insights found in this project can be of great help for people wanting to start their own restaurant in Amsterdam. Furthermore, brokers helping their clients in finding available venue and advise them on the neighborhoods.

2 Data description

To find an answer to the business problem we will make use of the following two data sources:

- Geographical data provided by the Amsterdam Data Website
- Location data provided by Foursquare

In this chapter these sources are further introduced and explained.

2.1 Data sources

In this chapter data courses are defined and described. Also, some introduction has been made addressing the possibilities and difficulties handling those courses.

2.1.1 Location data neighborhoods

The Amsterdam data website (https://maps.amsterdam.nl) has all the geographical data this project needs. The geographical data of the neighborhoods is downloadable in different formats like .csv and Geo Json.
The advantage of the Geo Json format, which is used in this project, is its precise definition of each neighborhood where the borders are defined by a polygon (fig 1) consisting of al the latitudes and longitudes.

```
In [15]: amsterdam_data.geometry.head(1)

Out[15]: 0    POLYGON ((4.90326 52.37658, 4.90298 52.37668, ...
         Name: geometry, dtype: geometry
```

Fig 1. Example of a Geo Json polygon object.

This list of location coordinates describes the borders of each neighborhood as shown in figure 2.

```
In [8]: p_list = amsterdam_data.geometry[0]
        p_list
Out[8]:
```

Fig 2. Plot of a polygon entry in the geojson dataset amsterdam_data.

2.1.2 Venue Data: The Foursquare database

The second source which is used in this project is the Foursquare database.
Foursquare (https://foursquare.com) is a company which build and maintains a massive dataset of accurate location data. This data is freely available using the RESTful API.
After registration on their website on can do a limited number of queries with different search parameters so called “endpoints” and collect a dataset containing all the interesting venues a location has to offer. Figure 3 shows what a RESTfull API url would looks like:

```
# create the API request URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'
```

Fig 3. Example of quiry on Foursquare database.

As show in figure 3, one of the limitations of this API is the search method “radius” around a location of interest.
Since neighborhoods in old cities like Amsterdam tends to have irregular forms and sizes there is a need for a more specific definition of the search area.
If the radius has been defined to small, interesting venues could be missed in the neighborhood. On the other side, choosing a to large radius will result in duplicate venues and venues assigned to the wrong neighborhood.

Therefore, a method has to be developed finding most of the important venues in each neighborhood while preventing duplicates or venues wrongly assigned.

2.1.3 Venue Data: Feature selection

To retrieve only venues interesting for this project the following parameters has been user for the query.

Latitude and Longitude: This is the geographical location of the search point
Radius: radius within venue information will retrieved.
CategoryId: Here a sting of categories can be given to include in the search

The categories of interest are:

- 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

After performing the query, the above-mentioned features will be selected, cleaned, venues hot encoded and put into a data frame (fig 4).

	Neighborhood	Longitude	Latitude	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
0	Burgwallen-Oude Zijde	4.896919	52.372084	44.0	23.0	29.0	64.0	28.0
1	Burgwallen-Nieuwe Zijde	4.895055	52.374195	69.0	37.0	17.0	46.0	81.0
2	Grachtengordel-West	4.887844	52.373349	29.0	8.0	13.0	23.0	32.0
3	Grachtengordel-Zuid	4.892044	52.365930	39.0	29.0	31.0	35.0	27.0
4	Nieuwmarkt/Lastage	4.904592	52.371729	31.0	11.0	14.0	16.0	33.0
...	...	...	...	...	...	...	...	...
94	Bijlmer Oost (E,G,K)	4.982963	52.324360	4.0	0.0	3.0	4.0	2.0
95	Nellestein	4.999182	52.307183	1.0	0.0	4.0	0.0	5.0
96	Holendrecht/Reigersbos	4.976040	52.293713	5.0	1.0	3.0	2.0	4.0
97	Gein	4.994089	52.296346	1.0	0.0	0.0	1.0	4.0
98	Driemond	5.011249	52.311792	0.0	0.0	0.0	0.0	0.0

Figure 4: Final data frame containing the features to be analyzed

2.2 Data cleaning

An essential step in Data Science is cleaning of data before using in any analysis.
When data does have flows or is not in the correct format for the tools to be used the results can be sub optimal and results inconclusive.
In this chapter the most important data cleanings will be mentioned and explained.

2.2.1 Removing unnecessary features

After performing the query on the foursquare data base, it appeared there where contaminations of venues types outside of categories provided (fig 5).

Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category id
Bakery	1	1	1	1	1	1	1	1
Bar	3	3	3	3	3	3	3	3
Bed & Breakfast	2	2	2	2	2	2	2	2
Bistro	1	1	1	1	1	1	1	1
Brasserie	6	6	6	6	6	6	6	6
Burrito Place	1	1	1	1	1	1	1	1
Café	1	1	1	1	1	1	1	1
Chocolate Shop	1	1	1	1	1	1	1	1

Figure 5: Found venues outside of the predefined categories to be removed.

For example, the query in the restaurant category yielded also venues like “Hotel”, “Bar” and grocery shop.
After a short investigation of the unwanted categories, python techniques for data frame dropping have been used to get rid of unwanted venues.

2.2.2 Removing duplicate values

One of the disadvantages of the Foursquare database is the lack of flexibility in defining the search area. A search can only be made by a circular of rectangular area.
Due to overlap of the search areas for small crowded neighborhoods in de city center, there will be a huge number of duplicated venues which have to be removed.

2.2.3 Assigning venues to correct neighborhood

As introduced in the previous chapter. Searching neighborhoods varying in shape and size by an API request with a fixed radius can result in either missing many venues or having duplicated values spread over multiple neighborhoods. To overcome this problem an optimal radius has to be chosen and the dataset will be cleaned of duplicates. Finally, each venue assignment will be checked based on its location and the definition of the neighborhood polygon.

3 Methodology

This chapter will describe the methods used for analysis, statistical testing of the machine learning model.

3.1 Exploratory Analysis: Assigning venues to the correct neighborhood

The below figure 6 shows a close up of the neighborhoods in the city center of Amsterdam.
The blue lines are representing the official borders of the neighborhoods and the green dots the calculated centers.

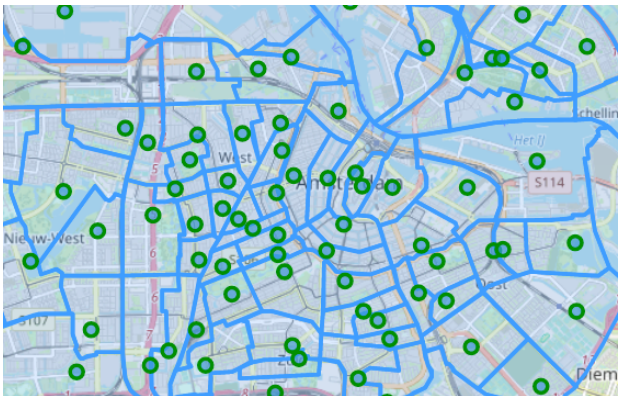

Figure 6: Neighborhoods and location centers

As one easily can see, the neighborhoods differ in size and shape and searching venues using a radius will likely result in duplicate findings and wrongly assigned venues.
To illustrate this in more detail we will review the venue assignment for the neighborhood "Burgwallen-Nieuwe Zijde" in more detail.

After the data has been acquired from the foursquare database using a radius of 500.
The data has been cleaned from duplicates, pre-processed and put into a test data frame.
In figure 7 the found venues for "Burgwallen-Nieuwe Zijde" are plot as red dots for Italian restaurants and blue dots for general restaurant. Notice that the venues are also located outside of "Burgwallen-Nieuwe Zijde" and are spread over at least 4 different neighborhoods.
Conclusion: many venues are wrongfully assigned and this has to be corrected.

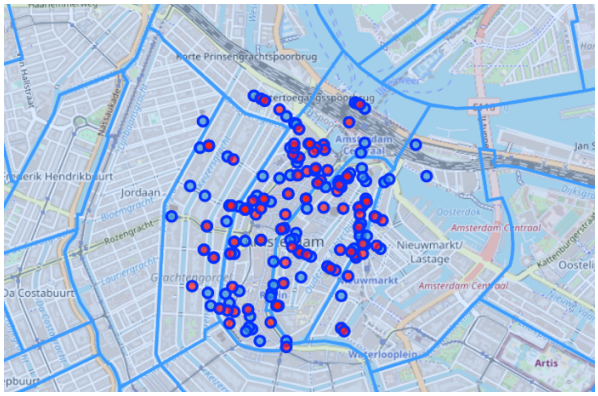

Figure 7: venues assigned to "Burgwallen-Nieuwe Zijde"

To achieve good venue assignment, a function has been designed using Python Geopandas library. With this library an object (neighborhood) can be defined as a polygon marking the borders by geographical coordinates.

The created function loops over a list of neighborhoods and checks all available venues for their latitude and longitude to verify if they fall inside of the defined polygon area.
After passing all the neighborhoods, a short quality check takes place to see if all venues have a neighborhood assigned.
After applying this function, we plot again the venues assigned to the neighborhood “Burgwallen-Nieuwe Zijde” (figure 8) to see if any improvement has been made.

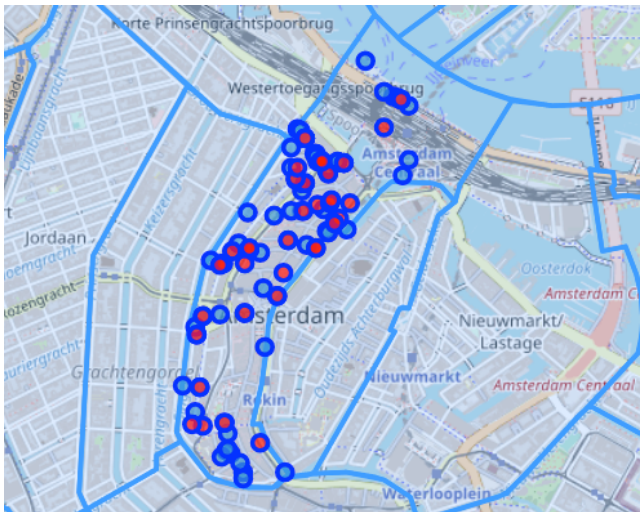

Figure 8: venues assigned to “Burgwallen-Nieuwe Zijde” after reassigning

As figure 8 shows, all assigned venues “Burgwallen-Nieuwe Zijde” are now located within the defined borders. This is good news since we want to evaluate and cluster the neighborhoods based on their own characteristics.

3.2 From venue description to feature creation: One hot Encoding

Categorical variables like venue category (fig. 9) and venue names are difficult to analyze using machine learning techniques.
Therefore, an effort has been made to convert those categorical values into numerical ones.
One of the most applied methods to do this is called “one hot encoding” by assigning a zero or one depending of state of the variable.

	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

Figure 9: data frame containing categorical values for Venue and Venue Category.

After hot encoding the different categories (example fig 10) the data frame is grouped by neighborhood summing al found venues within each venue category.
This process has been repeated for all made foursquare queries to distillate the 5 features of interest.

	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

Figure 10: snapshot of one hot encoded data frame for the restaurants query

The one hot encoding and grouping of all neighborhoods leads to the wanted 5 features (fig. 11):

- All restaurants: Find all venues to assess the indirect competition
- Italian Restaurants: Find all venues to assess the direct competition
- Arts and Entertainment: Attractive venues for tourists \ possible business partner
- Nightlife: Indicator for popularity by the locals and tourists
- Travel and Transport: Indicator how reachable a potential neighborhood is

	Neighborhood	Longitude	Latitude	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
0	Burgwallen-Oude Zijde	4.896919	52.372084	44.0	23.0	29.0	64.0	28.0
1	Burgwallen-Nieuwe Zijde	4.895055	52.374195	69.0	37.0	17.0	46.0	81.0
2	Grachtengordel-West	4.887844	52.373349	29.0	8.0	13.0	23.0	32.0
3	Grachtengordel-Zuid	4.892044	52.365930	39.0	29.0	31.0	35.0	27.0
4	Nieuwmarkt/Lastage	4.904592	52.371729	31.0	11.0	14.0	16.0	33.0
...	...	...	...	...	...	...	...	...
94	Bijlmer Oost (E,G,K)	4.982963	52.324360	4.0	0.0	3.0	4.0	2.0
95	Nellestein	4.999182	52.307183	1.0	0.0	4.0	0.0	5.0
96	Holendrecht/Reigersbos	4.976040	52.293713	5.0	1.0	3.0	2.0	4.0
97	Gein	4.994089	52.296346	1.0	0.0	0.0	1.0	4.0
98	Driemond	5.011249	52.311792	0.0	0.0	0.0	0.0	0.0

Figure 11: data frame containing the features of interest to perform final analysis.

3.3 Machine learning: k-means clustering

The machine learning technique applied in this project is K-Means Clustering.
This chapter contains a brief explanation what K-Means Clustering is and why is has been applied. Furthermore, the specific application on the dataset will be discussed as well as parameter choice and the validation of the model.

3.3.1 What is k-means and why is the method of choice

K-Means Clustering is a machine learning technique for calculating the similarity and dissimilarity of points in a given dataset. It is a form of unsupervised machine learning being capable to handle unlabeled data which is the case of our feature set.
By calculating the distances between data points clusters are formed by two optimizations.

Create clusters where distance of points within the clusters are minimized and the distance between de clusters is maximized.
Since we are trying to separate the neighborhood based on unlabeled features and 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.

K-means is known for the risk of over fitting at a large number for k and many iterations. This might lead to an extreme separation of datapoints but a bad result in clustering and will be examined in the following chapters.

3.3.2 Normalizing the data

Before performing k-means clustering on the data set the features have to be normalized.
The K-Means algorithm is sensitive for unscaled data. Since (Euclidean) distances will be calculated we have to make sure every feature will have the same impact on the calculation of distances. The most commonly used normalization method is Min\Max scaling (eq 1).

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

Applying the above equation on the feature data will result in a scaling between 0 and 1 where 0 = min value and 1 is max value. Figure 12 contains an example of the 5 features sets normalized by Min\Max method.

	0	1	2	3	4
0	0.523810	0.621622	0.690476	1.000000	0.345679
1	0.821429	1.000000	0.404762	0.718750	1.000000
2	0.345238	0.216216	0.309524	0.359375	0.395062
3	0.464286	0.783784	0.738095	0.546875	0.333333
4	0.369048	0.297297	0.333333	0.250000	0.407407

Fig. 12: Normalization of the features.

3.3.3 Parameter selection: number of K

The most critical and part of the K-Means clustering is determining the best number of k.
K naturally will be within the following to extreme cases
When K is chosen 1 than there will be only one cluster containing all datapoints.

Cluster distance is maximized (no other cluster available) but there is no separation of data points.
When K is chosen very large (for example 99 as the number of neighborhoods), every cluster might contain 1 datapoint. This is also not useful since in this case there is no clustering at all.
To find an optimal value of K an iterative approach has been chosen. The clustering algorithm has been applied on the data set using different values of k and 'quality' of the clustering examined using the commonly used Elbow method and Silhouette score.
in the next chapter the results of these examinations will be shown and discussed.

3.3.4 Statistical testing: Elbow method

In the elbow method multiple runs of the K-means algorithm are executed over a range of values for K. In our case the range for k is 1-10, which is commonly used. For each run the inertia score is calculated on the datapoints after clustering.
Inertial is the sum of the distances of data points to their closest cluster center. An increasing number of K will result in a decrease in Inertia.
In figure 13, the calculated inertia is plot against the number of clusters.
There is a slight elbow change of steepness visible at k= 4 to 5, which is the commonly advised optimal point for K.

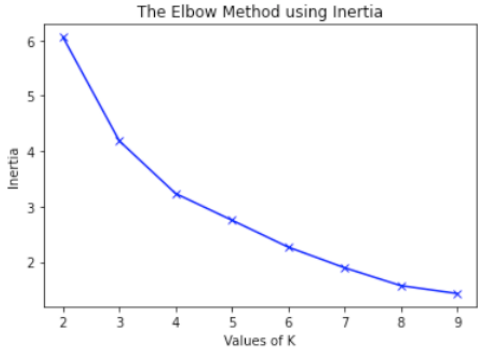

Figure 13: Results of application of the elbow method on the dataset.

3.3.5 Statistical testing: silhouette score

The Silhouette method measures how similar a datapoint is to its own cluster(cohesion) and to other clusters (separation). The optimal value should be the higest value in the range +1 and -1.
Figure 14 shows the results of the silhouette test plotted against the number of K.
There is a small local maximum visible at k= 5.

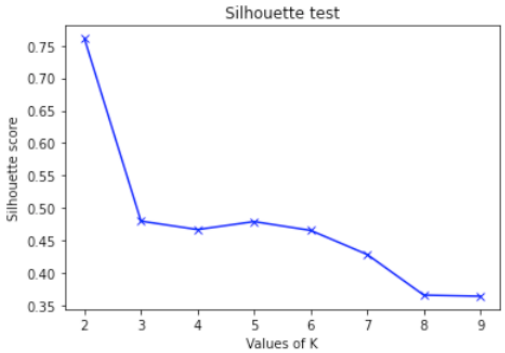

Figure 14: application of the silhouette test on the data set. Local maximum at k = 5

Conclusion: Evaluating the results of the elbow and silhouette test, the optimal value for K is 5.
The next chapter contains the results of the final K-means clustering run using the found value for K.

4 Results

This chapter covers the results achieved in this project. In figure 15 the neighborhoods are plotted into a folium map. The colors of the marker points are corresponding with the calculated labels of the clustering.

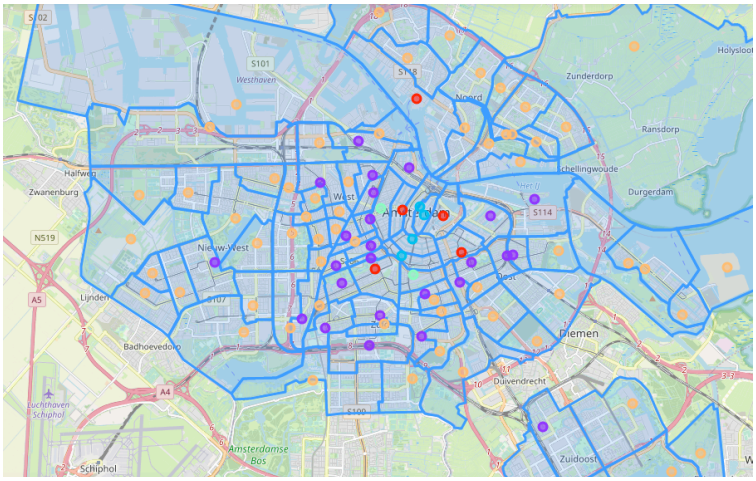

Figure 15: clustering visualized by plotting in folium map.

Clusters are assigned by following colors:

- Purple: Cluster 1, containing 25 neighborhoods (for more details see also figure 16)
- Dark blue: Cluster 2, containing 4 neighborhoods (for more details see also figure 17)
- Light blue: Cluster 3, containing 2 neighborhoods (for more details see also figure 18)
- Orange: Cluster 4, containing 62 neighborhoods (for more details see also figure 19)
- Red: cluster 5, containing 6 neighborhoods (for more details see also figure 20)

Now we have visualized the clusters in a map, the cluster data has to be reviewed in more detail to address the characteristics and form conclusions about the possibilities to start a Italian restaurant.

Cluster 1: These are neighborhoods mainly located outside of the touristic citycenter and therefore less interesting to starting a restaurant for tourist and dayvisitors.

	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

Figure 16: snapshot of contents of cluster 1 (purple marker points)

Best reachable are: *Vondelbuurt* and *Zuidas* which have relatively good transport possibilities and low number of italian restaurants. Possibly interesting neighborhoods to start a restaurant targeting the local inhabitants of Amsterdam

Cluster 2: The most popular neighborhoods of the city. Plenty entertainment venues and fairly saturated with restaurants. This cluster is full of restaurants art and nightlife venues. Because its central position it is the best reachable area of the city in terms of public transport.
This is a hard cluster to start a restaurant because of huge competition and therefore not advisable.

	Neighborhood	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
0	Burgwallen-Oude Zijde	44.0	23.0	29.0	64.0	28.0
1	Burgwallen-Nieuwe Zijde	69.0	37.0	17.0	46.0	81.0
3	Grachtengordel-Zuid	39.0	29.0	31.0	35.0	27.0
7	De Weteringschans	56.0	28.0	33.0	64.0	50.0

Figure 17: snapshot of contents of cluster 2 (dark blue marker points)

Cluster 3: Very "Hipster" and popular neighborhoods, reachability great and there is a lot to do in the sense of nightlife. Unfortunately crowded with restaurants as well as plenty Italian restaurants. Difficult cluster to start new business because of intensive competition and therefore not advisable.

	Neighborhood	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
6	Jordaan	84.0	22.0	20.0	47.0	31.0
47	Oude Pijp	76.0	13.0	7.0	50.0	26.0

Figure 18: snapshot of contents of cluster 3 (light blue marker points)

Cluster 4: Neighborhoods outside of the city center with low reachability by public transport. Low restaurant count but unfortunately also low cultural and nightlife possibilities.
Unpopular areas for going out, therefore not advisable.

	Neighborhood	Restaurants	Italian restaurants	Arts\entertainment	Nightlife venues	Trasport venues
10	Westelijk Havengebied	3.0	0.0	3.0	2.0	2.0
11	Houthavens	3.0	0.0	1.0	2.0	10.0
14	Centrale Markt	2.0	0.0	1.0	4.0	4.0
18	Van Lennepbuurt	12.0	1.0	4.0	6.0	4.0
22	Sloterdijk	1.0	0.0	2.0	5.0	7.0
...	...	...	...	...	...	...
94	Bijlmer Oost (E,G,K)	4.0	0.0	3.0	4.0	2.0
95	Nellestein	1.0	0.0	4.0	0.0	5.0
96	Holendrecht/Reigersbos	5.0	1.0	3.0	2.0	4.0
97	Gein	1.0	0.0	0.0	1.0	4.0
98	Driemond	0.0	0.0	0.0	0.0	0.0

62 rows × 6 columns

Figure 19: snapshot of contents of cluster 4 (orange marker points)

Cluster 5: Most interesting cluster. All is in balance since located close to city center there are many options for public transport and many venues in the category's arts\entertainment and nightlife. Most interesting neighborhoods are: 'Museum kwartier 'and 'Weesper buurt'. These are popular areas with a good restaurant \ Italian restaurant ratio. Transportability is good and due to the many cultural venues and nightlife to be considered as location to start a new Italian restaurant.

	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

Figure 20: snapshot of contents of cluster 5 (red marker points)

5 Discussion

This chapter will briefly cover the observations made during the project and present some recommendation and future directions.

5.1 Observations

During the project we worked with the Foursquare Database and applied machine learnings as K-Means clustering. Let's review the most prominent observations

Foursquare API: Retrieving data using the Foursquare database might seem straight forward at first use.
After carefully reviewing the data, there are some pitfalls to take into account.

- Using categories can focus your search but still allow other unwanted categories to contaminate the data and therefore still needs to be cleaned.
- Duplicate venues have to be deleted and remaining venues can be reassigned which greatly improves accuracy.
- The free account gives limited possibilities to perform queries, but saving results and export them can save some time and frustration.
- Overall database is rich of valuable data and Foursquare has proven to be reliable because of the implementation of the RESTFull API.

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 choose and determine greatly the outcome and precision of the clustering.
- Other parameters might be investigated on their effect on clustering
- Evaluate independence of features is still an area to be considered for future development

5.2 Recommandations

As presented in the results and conclusion section, some neighborhoods have been found to be potential interesting for starting a restaurant.
The following recommendations can be made in further addressing the business question.

- Use foursquare data to review specific venues inside the targeted neighborhood on ratings popularity during traffic hours
- Profile potential customers in the neighborhood; what could be their spending budged?
- Review of menus \ concepts of most popular competition and asses if there are opportunities to diversify or fill a gap in the portfolio of the neighborhood

6 Conclusion

During the final stage of the IBM Data Science certificate course machine learning has proven to be an effective tool for discovering insights from large amounts of data.
The Foursquare database is an interesting data source but has its limitations when it comes to data precision.
This precision can be enhanced by using additional data cleaning and especially using Geopandas object for assigning venues to the correct neighborhoods.

Opening a restaurant is a complex job and finding a promising area to start can be of great help. Using K-means clustering we have been able to group 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).

These neighborhoods are potentially interesting starting an Italian restaurant because of good transportability, relatively low restaurant competition and enough close to popular venues which could promote enough traffic of potential customers.