Mama Mia Pizza

Battle of the Amsterdam Neighborhoods – Week 1 (of 2)

Coursera Capstone Project for the IBM Data Science Certification
Wiljo Meijnhout

January 2020

Introduction & Business Problem

Introduction and background

This report is created as the final assignment for the Applied Data Science course from Coursera which results in an IBM Data Science Certification. The business problem and names I've defined, as part of the final assignment, is fictional but the data used to solve the problem is real data and can solve a real live business problem without any issues.

Business Problem

In Rome, Italy a famous franchise company called 'Mama Mia Pizza' want to spread their wings across the European capital cities, starting with Amsterdam in the Netherlands. Amsterdam City, according to Wikipedia has a population of 866,737 and is very popular amongst tourist from all over the world. According to DW.COM in 2018 19.000.000 tourists have visited Amsterdam. All these tourists have to stay of course in hotels, but as we all know also in one of the many Airbnb houses/rooms.

My client wants to '***Start a Mama Mia Pizza Restaurant in a neighborhood in Amsterdam, crowded of Airbnb hosts but where the restaurant market is less saturated compared to other neighborhoods'***.

In contrast to tourists who stay in a hotel an 'Airbnb tourist' have their dinner not at home but eat it i.e. in a restaurant. Pizza is a world famous-, well known, and an accessible dish. According to the Pizzajoint.com approximately 3 billion!! pizzas are sold in only the U.S. each year.

Data

Data Gathering

In order to solve my client's business problem a desk study was performed to gather the available data and to decide the Data Science approach. Airbnb itself does not provide any data at all according to their company policies, so that was a dead end. But the website insideairbnb.com does provide all the information we need. insideairbnb.com does scrape the Airbnb website in a smart way and represents the scraped data visually on their website and also make the data available via downloads in several formats.

Airbnb data

insideairbnb.com/amsterdam/ shows almost 20.000 Airbnb hosts across Amsterdam. (see fig.1). For computation power performance but also visuality, 20.000 Airbnb hosts in a dataset is too much and not needed at all to solve the business

problem. Below is explained how we decreased this number to approximately 7000.

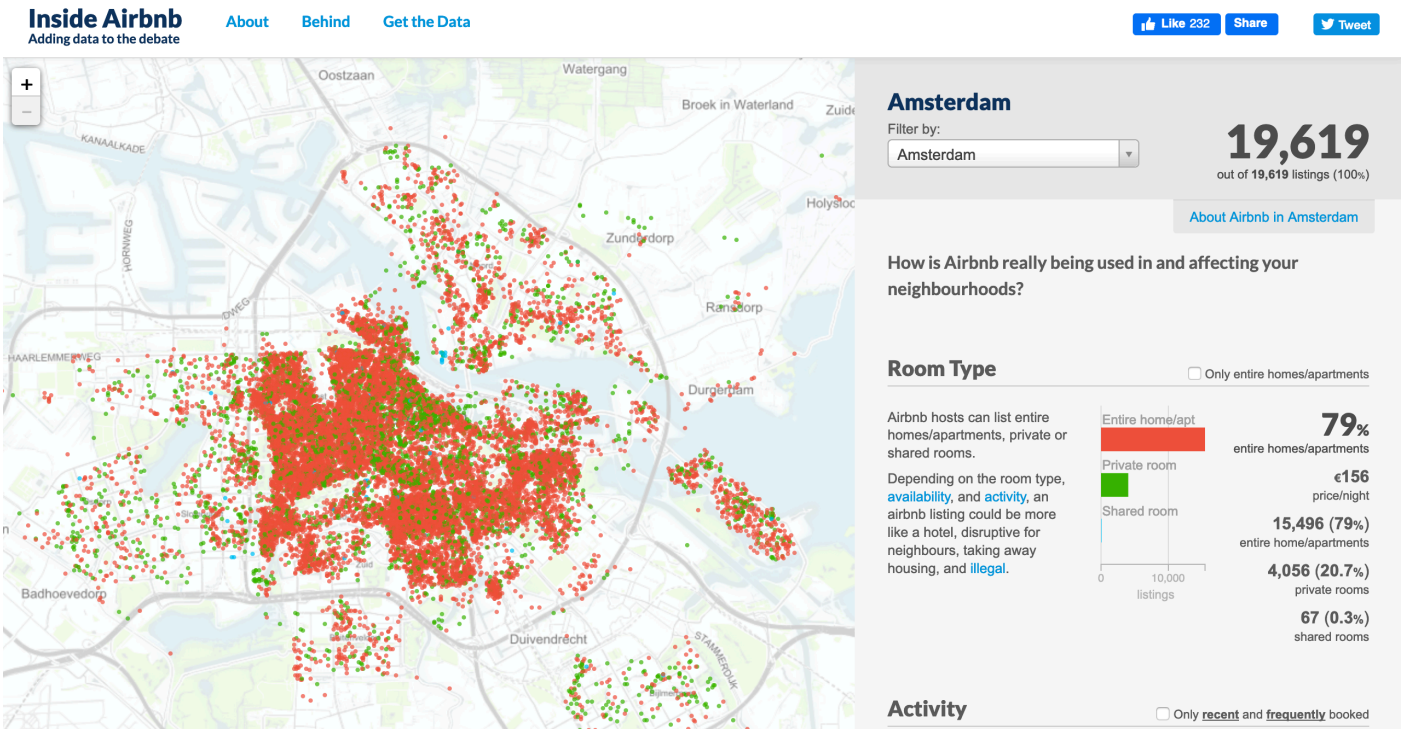


Figure 1 Screenshot inside Airbnb - Amsterdam

Geographical data of Amsterdam

Furthermore, neighborhood geographical data from the Amsterdam neighborhoods were needed. This data is available via the Amsterdam government data website: maps.amsterdam.nl. The dataset is downloaded as a geojson file for easy displaying on a Choropleth map.

Restaurant data from FourSquare

To get an overview of currently settled restaurants in Amsterdam a 'nearby venues' lookup for every Amsterdam neighborhood was extracted from the FourSquare database using a developer account. foursquare.com/city-guide is a website to find the best places to eat, drink, shop, or visit in any city in the world. Access over 75 million short tips from local experts.

Approach / Data cleaning

The Notebook can be found on my Github repository. https://github.com/WiljoM/Coursera_Capstone

In some cases, the maps are not visible via Github. In that case I suggest you to use this document.

Airbnb Data

Starting with the Airbnb data from InsideAirbnb I found almost 20.000 records. Due to performance- and visual reasons this needs to be cleaned up:

2404 hosts didn't have rating at all, meaning they were active. Let's remove them: (see fig.2)

```
In [7]: # Check for empty cells
        airbnb_data.isnull().sum()
```

```
Out[7]: id                                    0
        name                                 34
        host_id                               0
        host_name                           158
        neighbourhood_group               20025
        neighbourhood                         0
        latitude                              0
        longitude                             0
        room_type                             0
        price                                 0
        minimum_nights                        0
        number_of_reviews                     0
        last_review                        2404
        reviews_per_month                  2404
        calculated_host_listings_count        0
        availability_365                      0
        dtype: int64
```

Drop 2404 rows where never has been a review, assuming these are no active Airbnb hosts

Figure 2 2404 non active airbnb hosts

The dataset was still to large. Let's focus on the Price per night and see if we can clean up the outliers. (see fig.3)

Dataset is still too large and needs to be decreased

```
In [10]: # What about the distribution of 'price' ?

         # np.histogram returns 2 values
         count, bin_edges = np.histogram(airbnb_data['price'])

         print(count) # frequency count
         print(bin_edges) # bin ranges, default = 10 bins
```

```
[17565    48     1     1     0     2     0     0     1     3]
[   0.  900. 1800. 2700. 3600. 4500. 5400. 6300. 7200. 8100. 9000.]
```

```
In [11]: # 17000+ records left, so let's remove al the outliers above 400 euro per night
         indexNames = airbnb_data[airbnb_data['price'] > 400 ].index
         airbnb_data.drop(indexNames , inplace=True)
         airbnb_data.shape
```

```
Out[11]: (17253, 16)
```

Figure 3 Dataset still too large

We need to dive further into the price and check the distribution of it. Therefor a distribution plot was made (see fig.4). The majority is between 100 and 150 euro per night. Let's focus on the Airbnb hosts between 100 and 150 euro per night to decrease the dataset to a workable level of approximately 7000 records.

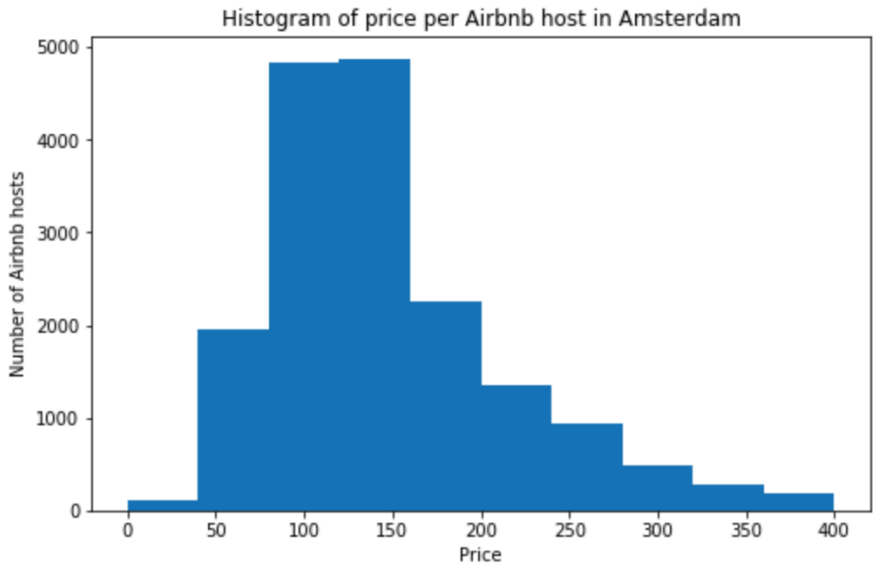
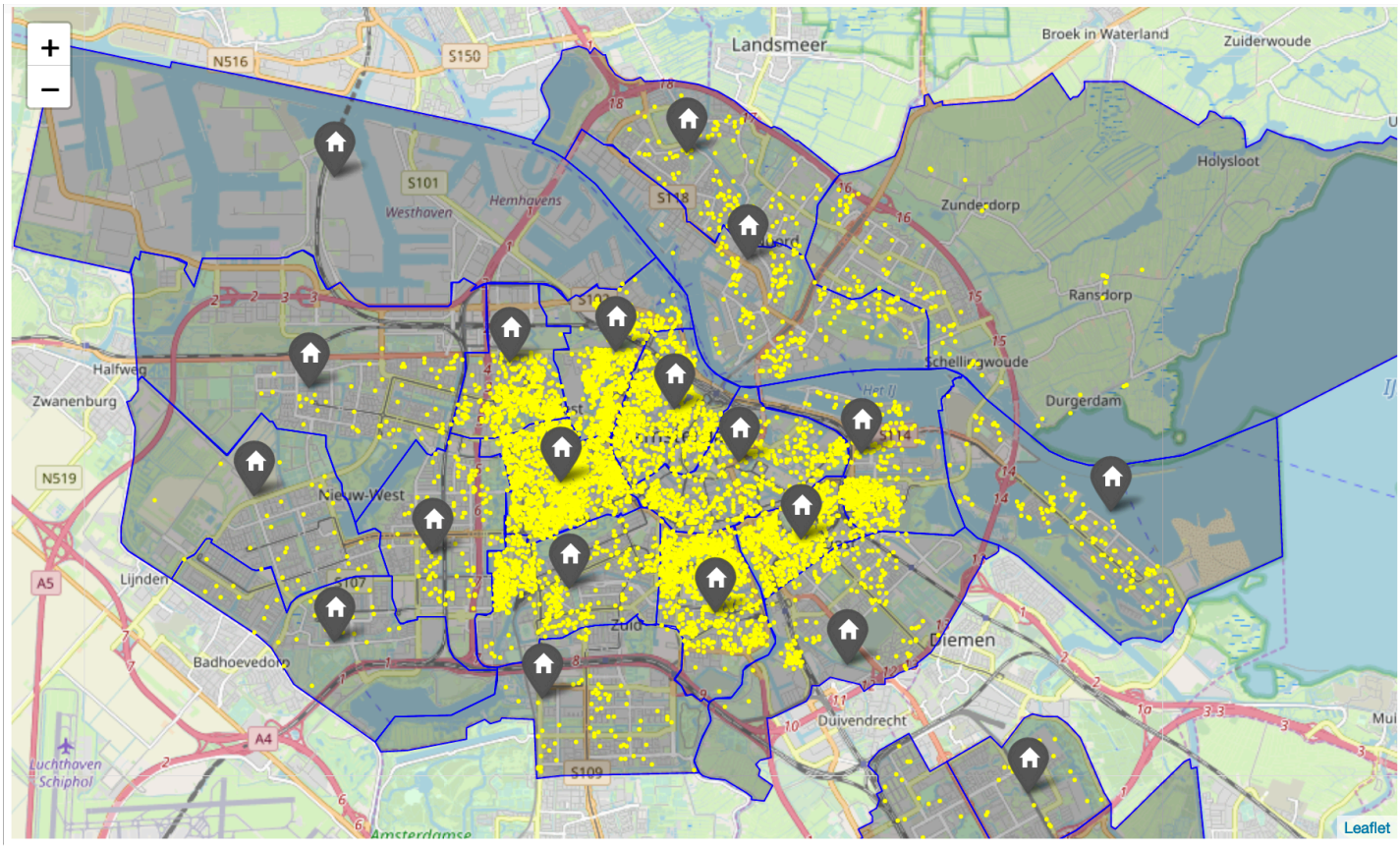


Figure 4 Histogram of price per Airbnb host in Amsterdam

Together with the geojson file with the neighborhood boundaries a map was created to plot all the ~7000 Airbnb hosts related to the neighborhoods where they reside. (see fig.5)



The yellow dots represent the Airbnb hosts, and the 'house' markers represent the Amsterdam Neigborhoods or Areas (Gebieden in Dutch)

Figure 5 Airbnb hosts plotted on the Amsterdam neighborhoods

FourSquare data cleaning

The export of the Nearby Venues from the FourSquare database consists of 1889 venues (see fig.6)

```
In [24]: # How many venues were returned?
         print('{} venues were returned by Foursquare.'.format(amsterdam_venues.shape[0]))
         amsterdam_venues.head()

         1889 venues were returned by Foursquare.
```

Figure 6 Number of venues returned from FourSquare

These 1889 are all of the nearby venues, but we're interested in restaurants only. Before displaying the restaurants on a map, a filter was written to extract only the Restaurants in Amsterdam. (see fig.7 and 8)

Create an array of all restaurants

In [28]:
```
array = []
search = 'Restaurant'
for i in amsterdam_venues.venue_category :
    if search in i:
        array.append(i)

# and add Pizza place because this is also a restaurant
array.append('Pizza Place')
```

Figure 7 Filter out all restaurants

Create a new dataset of all restaurants in Amsterdam

In [29]:
```
amsterdam_restaurants = amsterdam_venues.loc[amsterdam_venues['venue_category'].isin(array)]
amsterdam_restaurants.head()
```

Out[29]:

	neighborhood	neighborhood_latitude	neighborhood_longitude	venue	venue_latitude	venue_longitude	venue_category
4	Westpoort	52.411465	4.807319	KFC	52.427470	4.820170	Fast Food Restaurant
8	Westpoort	52.411465	4.807319	McDonald's	52.427396	4.820760	Fast Food Restaurant
19	Bijlmer-Oost	52.319564	4.976832	De Smeltkroes	52.322755	4.974752	South American Restaurant
22	Bijlmer-Oost	52.319564	4.976832	Pasta di Mamma	52.314779	4.955087	Italian Restaurant
24	Bijlmer-Oost	52.319564	4.976832	Margherita Tutta La Vita!	52.329054	4.955773	Pizza Place

Figure 8 New dataset with only restaurants in Amsterdam

And a plot of all restaurants together with the Airbnb hosts in the Amsterdam neighborhoods (see. fig.9)

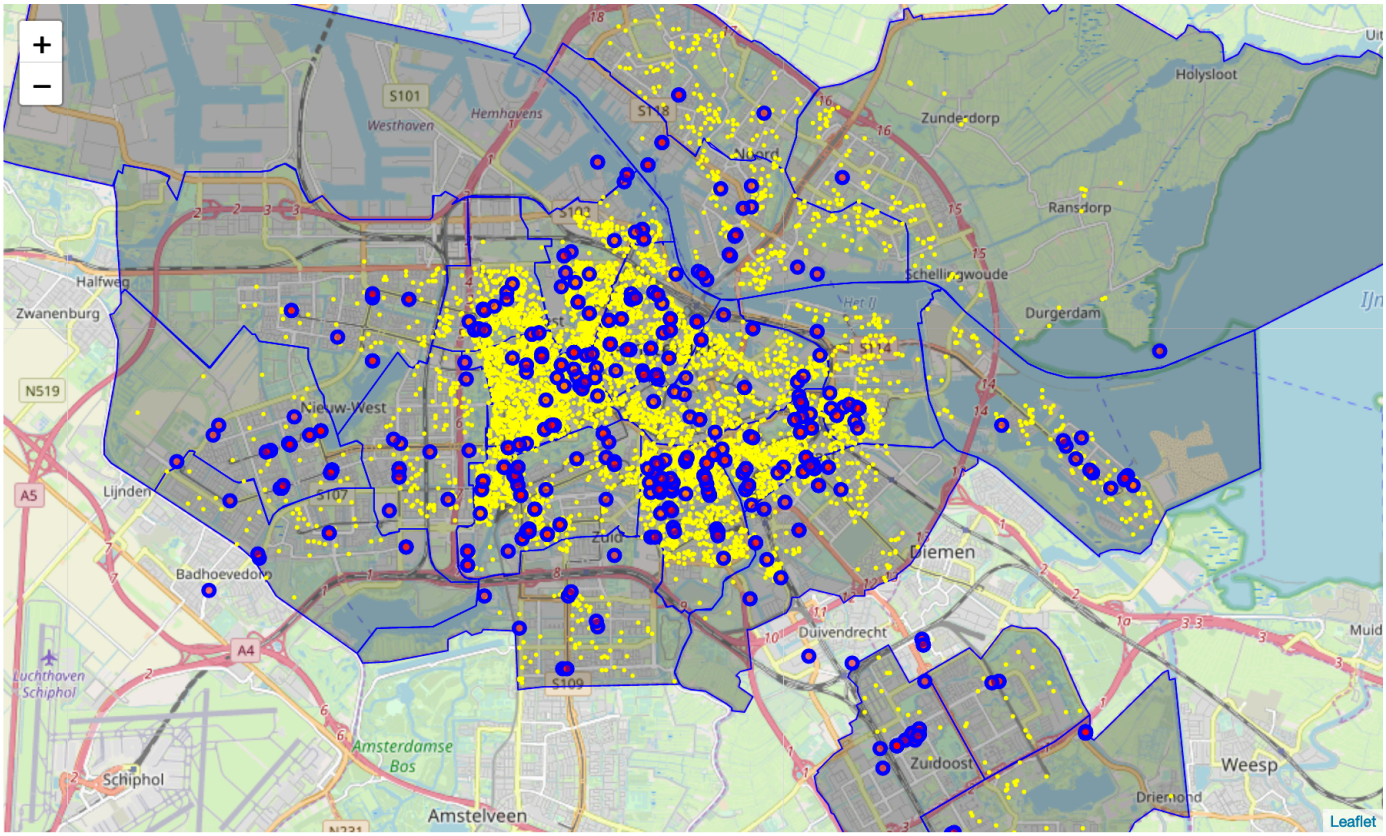


Figure 9 All restaurants plot on the map. Airbnb host in yellow, restaurants in blue

Sources

https://en.wikipedia.org/wiki/Amsterdam

https://www.dw.com/en/how-amsterdam-is-fighting-mass-tourism/a-47806959

https://www.thepizzajoint.com/pizzafacts.html

http://data.insideairbnb.com/the-netherlands/north-holland/amsterdam/2019-12-07/visualisations/listings.csv

https://maps.amsterdam.nl/open_geodata/?k=204

https://foursquare.com/city-guide