

Amsterdam v/s New York

Segmentation And clustering

of

different venues in both

cities

Kumar Shivam | Capstone Project | 20/10/2019

* Introduction

Few cities in Europe pair style and dilapidation like Amsterdam. With a downright raunchy [red light district](https://www.eurocheapo.com/amsterdam/hoods/red-light-district.html) on the one hand and some of the world's very best museums on the other, cosmopolitan Amsterdam takes its contradictions in stride.

New York City, officially the City of New York, historically New Amsterdam, the Mayor, Alderman, and Commonality of the City of New York, and New Orange, byname the Big Apple, city and port located at the mouth of the [Hudson River](https://www.britannica.com/place/Hudson-River), southeastern [New York](https://www.britannica.com/place/New-York-state) state, northeastern U.S. It is the largest and most influential American metropolis, [encompassing](https://www.merriam-webster.com/dictionary/encompassing) Manhattan and Staten islands, the western sections of [Long Island](https://www.britannica.com/place/Long-Island-New-York), and a small portion of the New York state mainland to the north of Manhattan

This project is comparison of different venues in Amsterdam and New York city. The most popular venues in both cities according to the ratings.

Using Foursquare API we can find the popular venues in both the cities. Clustering is applied on the venues of each city.

* Problem

The objective of this capstone project is to analyze different popular venues in both the cities using the machine learning. Segmenting and Clustering is applied for different venues in both the cities.

The solution of this project will give the comparison between both the cities and if anyone want to go for a tour between the two cities, then he/she can visit the popular places in respective neighborhood using the Foursquare API. The map will make visible popular places in both the cities.

* Target Audience

This project will be useful for the people who want to visit these two cities. The popular places in the cities to visit. This project can be useful for companies who deals in tour packages for people. Popular travel and booking websites can use the information from the project for recommendation of different places in these cities.

* Data

In this project we will require data of different neighborhood in New York city and different neighborhood in Amsterdam.

The latitude and longitude of different location will be required.

Data used in the project is obtained from Kaggle. It is data of different neighborhoods in Amsterdam. For New York data there is an New York Airbnb open data consisting of different neighborhood in New York. We need to pre-process the data and use it for the project.

Both the data is being downloaded from Kaggle and Data cleaning is done for both the data.

* Source of data

We will download the dataset from Kaggle and perform some data cleaning of the data. Both the datasets are available on Kaggle.

To download the Dataset for Amsterdam:

<https://www.kaggle.com/dushyantsapre/amsterdam-neighbourhood-and-venues>

To download the dataset for New York city:

<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, but we will particularly focus on top rated venues in both the cities.

In this project many data skills will be used like Web Scrapping (to search the data set), use of API (here Foursquare API and OpenCageGeocode API), Machine learning algorithm like Segmentation and Clustering (K-mean clustering), Data visualization, Map visualization (Folium), Data Wrangling etc. In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

* Data Cleaning

Data Obtained from different sources needs to be cleaned. The New York data need to be cleaned as it contains unnecessary columns that needs to be deleted.

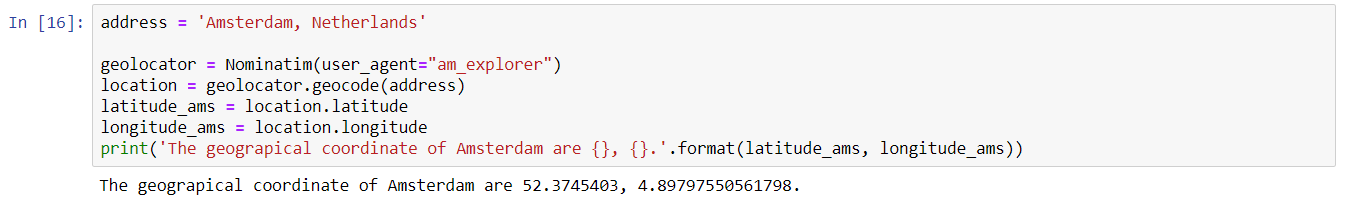
The New York data is obtained from the json file and the data frame is created from the json file and obtained data frame needs to be cleaned

The Amsterdam data is downloaded from the Kaggle and no need to be cleaned as it contains all the necessary information

The obtained dataframe saved as a csv file in the OS. Not much data cleaning is needed in this project

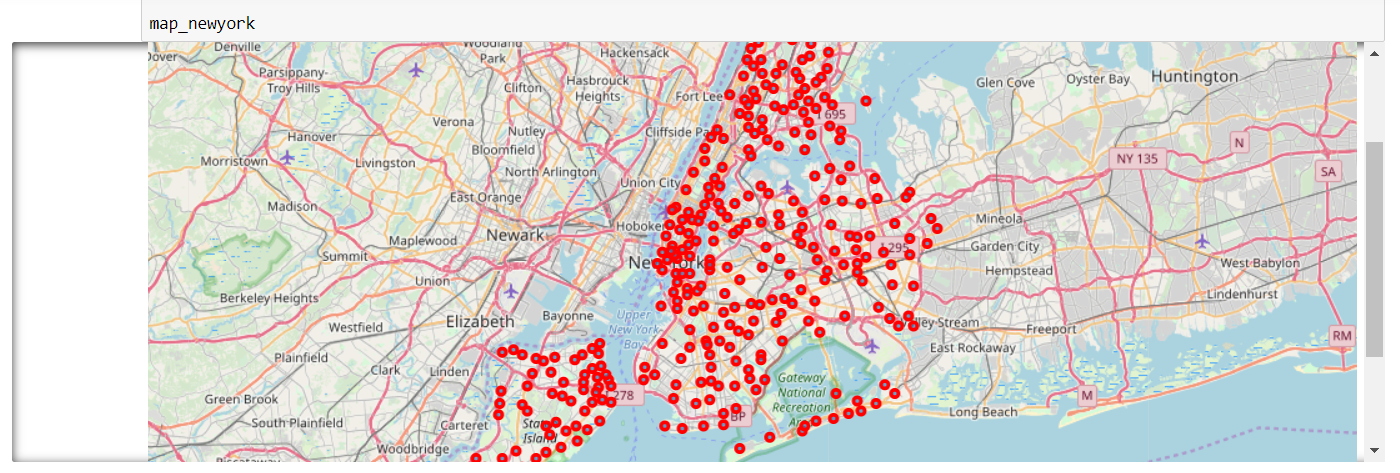
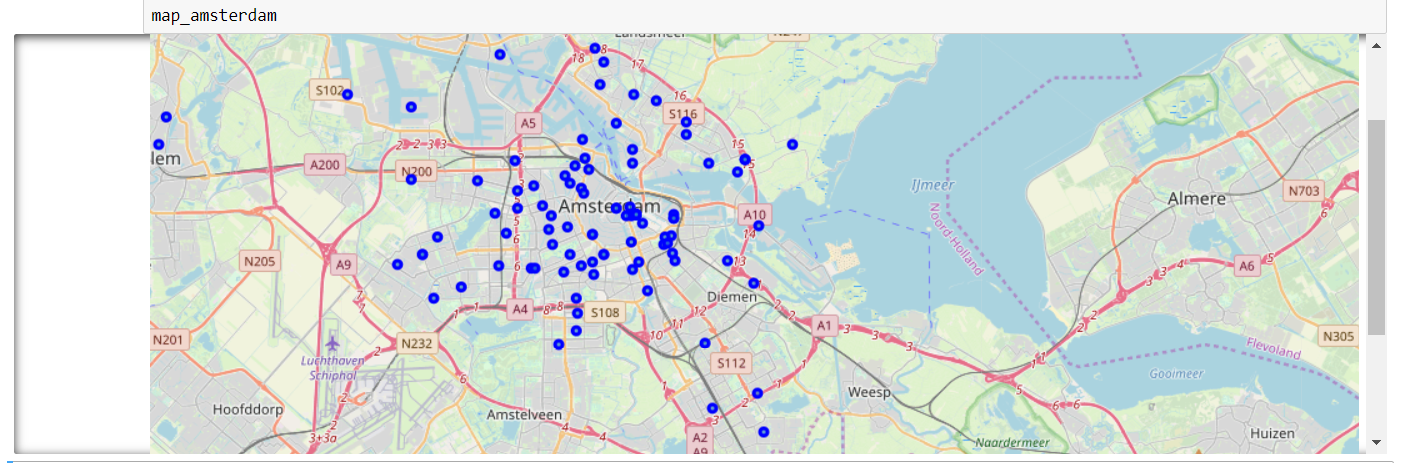
* Exploratory Data analysis

The obtained data needs to marked in the map using folium and the coordinates to both the location needs to be obtained.

Coordinates of the location is obtained using the Geocoder Nominatim.



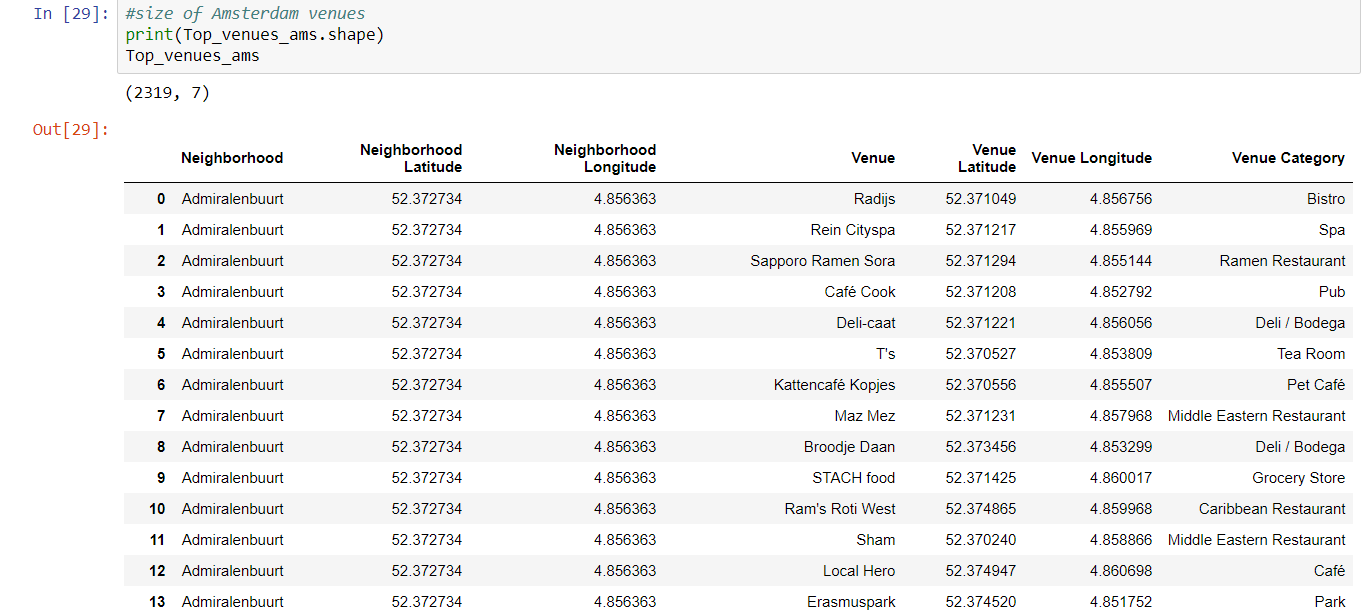
Different venues for both the location is obtained using the Foursquare API. Using Foursquare API different venues in New York and Different venues in Amsterdam is obtained for the corresponding neighborhoods for both the location.

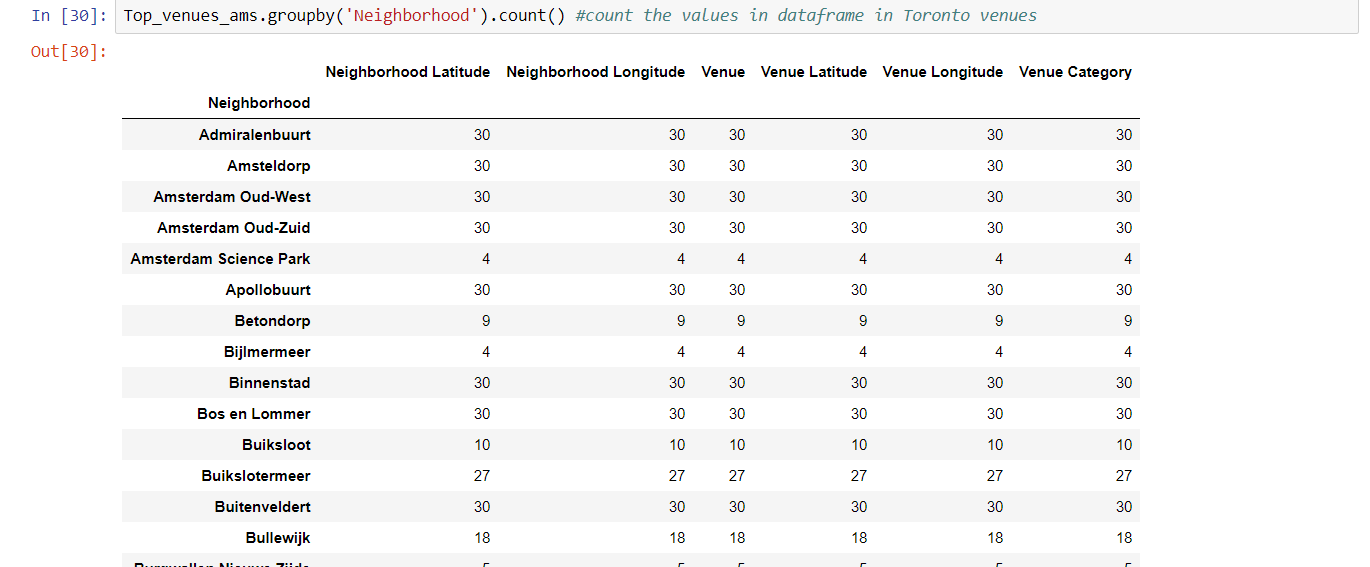
* Different venues are marked in the map using Folium

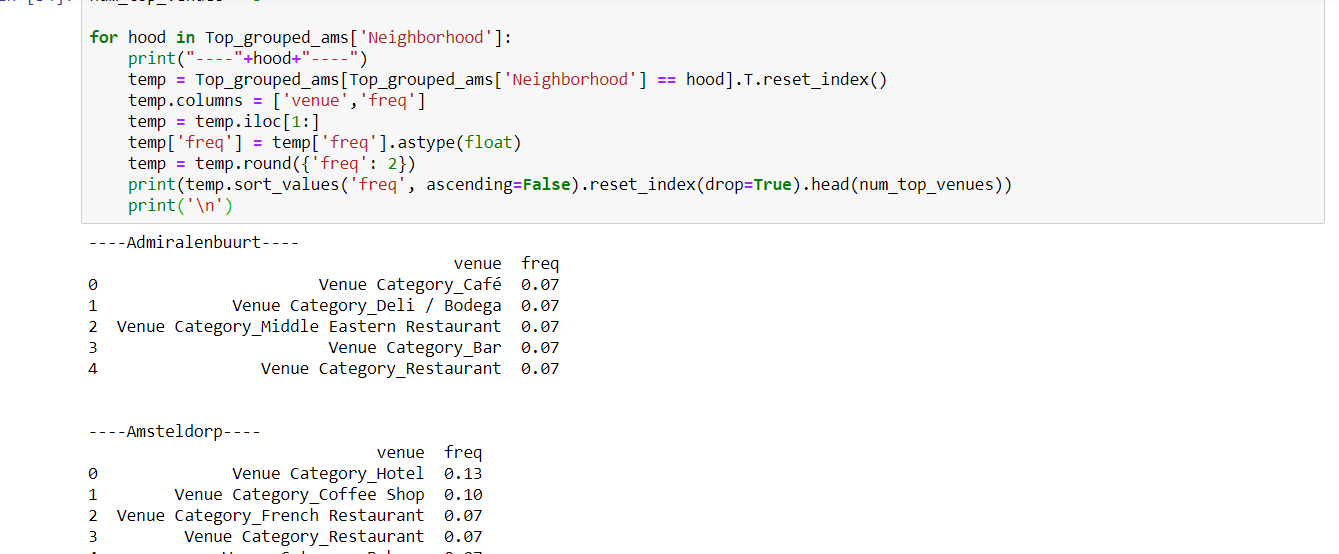
Different location has been marked on the map as shown above.

Find top venues in different location and create a data frame. Top venues are obtained using the foursquare API and the different venue data is merged with the neighborhood data and data is grouped

* The new dataframe is created and the Segmenting and Clustering is applied on the top venues in both the location which are the popular venues in the New York and popular venues in Amsterdam.
* Machine learning algorithm segmenting and clustering is applied in the new data frame using K-Means clustering.
* K-Means clustering with values of k=8 is used for the clustering.
* Then the obtained data is plotted on the map in Folium.

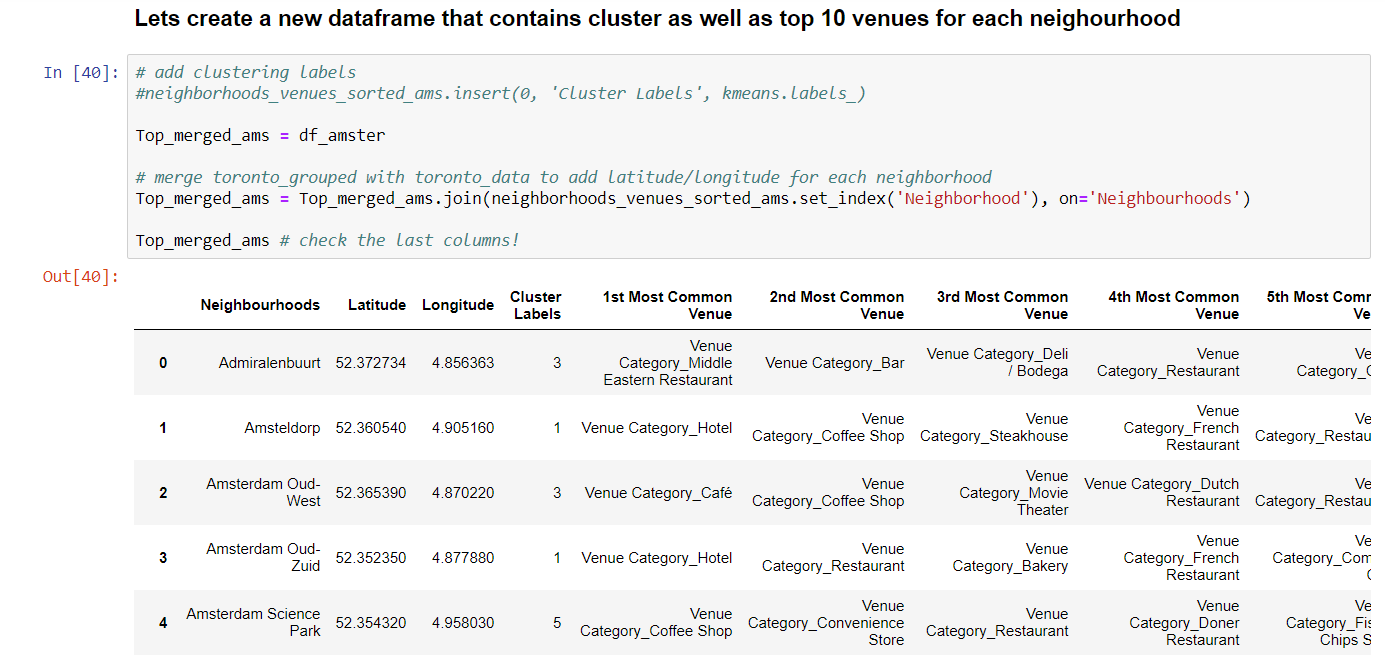


The venues are grouped and counted.

The top venues has been grouped and printed

The top 10 venues

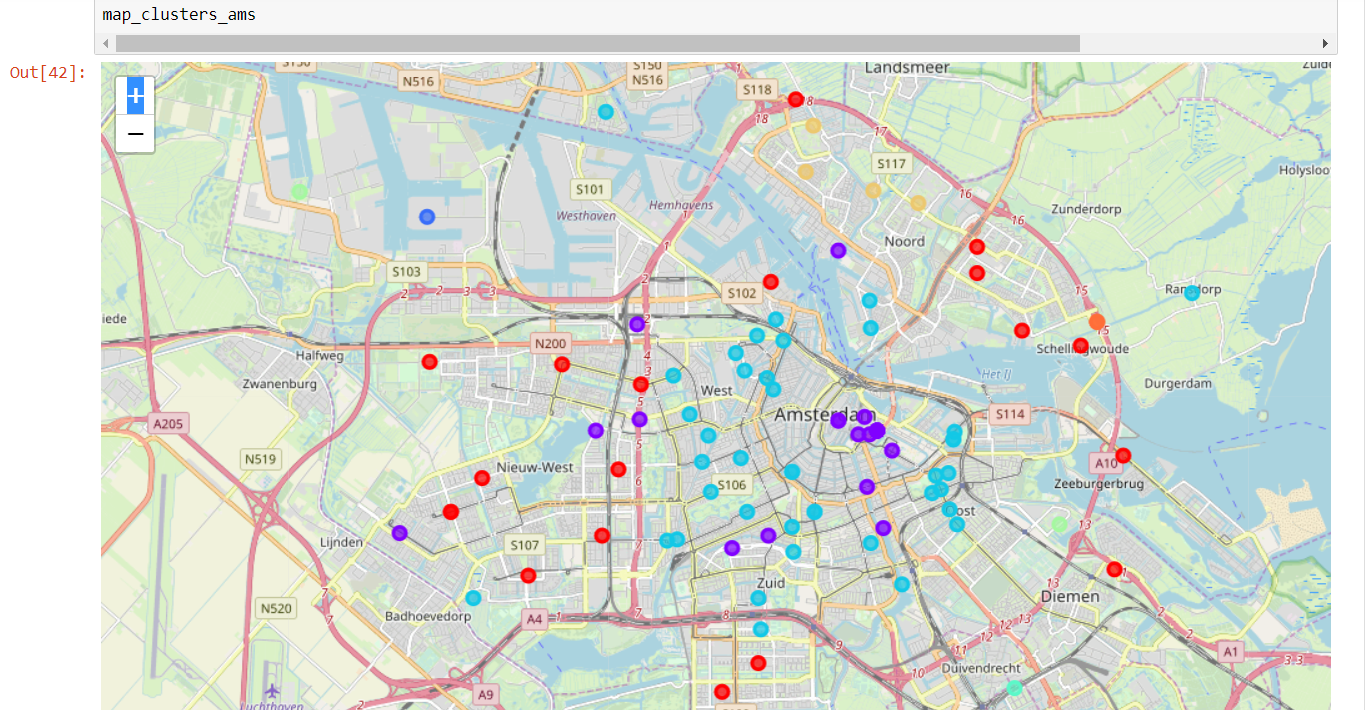
Then K-Means Clustering is applied for the venues. The values of k is taken as 8 and clustering is done

New dataframe is created for the top venues that contains cluster.

* The obtained data is plotted on the map using folium and we can see different cluster.
* Result

The clustering is done for different location and is plotted on the map.

The clustering is done using the K-Means cluster algorithm and the value of k=8.

New data frame is created containing the cluster and the data frame along with the latitude and longitude is marked on the map of the location using folium.

* Discussion Section

Different venues of same type has been clustered and marked on the map for the location so it will be easy to locate same type of venues in the city.

We can use this data for tourism where we can visit similar type of places which tourist like in different location and it will benefit lot to the travel companies and would definitely give boost to their business.

We observed here that there are lot of different venues located in different neighborhood of locations and can be segmented and clustered using K-means cluster algorithm.

* Conclusion

We obtain different types of similar types of venues in different location using K-Means clustering and there are many different venues on both the location.

I was and interesting analysis of the data and got to know about different venues as well during this project.