Grouping major world cities

Using venue data to group major world cities

Maxim

2020

Grouping major world cities

Using venue data to group major world cities

Introduction

When you like to go on a city trip it's always difficult to chose which city will be next. If you like a city how can you then tell what is the next city that you will most likely enjoy to visit. This is not only a problem for travelers, but even more so for computer algorithms that try to make suggestions to travelers. This data analysis will try to get new insights which can be added to existing algorithms of travel websites by grouping the cities based on venues. The purpose is to help those travel websites, like websites for hotel bookings or airplane tickets, further improve there predicting capabilities on which next trip a specific travel is likely to spend money on.

Data

To solve this problem we need to collect data. There is a lot of data available on the internet. Because of limited resources, time and computer power a selection will be made in the data to focus the research.

First major cities are defined as cities with a population of over 500.000. By choosing cities with a population of over 500.000 it will focus the research and make the data frame more manageable for further analysis. The website (https://worldpopulationreview.com/world-cities/) provides a comprehensive list of those cities. This data will be scraped form that website. This provides a data frame that has a list of 1133 cities and there total population.

	Rank	Name	Country	2020 Population	2019 Population	Change
0	1	Tokyo	Japan	37393129	37435191	-0.11%
1	2	Delhi	India	30290936	29399141	3.03%
2	3	Shanghai	China	27058479	26317104	2.82%
3	4	Sao Paulo	Brazil	22043028	21846507	0.90%
4	5	Mexico City	Mexico	21782378	21671908	0.51%

In the data frame that is created only the country, the city and the population in 2020 is useful for this analysis. In preparation of the next step a column for the latitude and one for the longitude of the city center should be added. This gives the following data frame:

	City	Country	Population	latitude	longitude
0	Tokyo	Japan	37393129	NaN	NaN
1	Delhi	India	30290936	NaN	NaN
2	Shanghai	China	27058479	NaN	NaN
3	Sao Paulo	Brazil	22043028	NaN	NaN
4	Mexico City	Mexico	21782378	NaN	NaN

Next the geographic location of each city center is determined. The city center information comes from the geopy package and consists of a latitude and longitude. To get this data the cities from the word-cities dataset are used as input for the geopy package. For example: (Toronto, 43.65, -79.34). After adding al the geographical data there are eight cities without any latitude and longitude. Those are deleted to end up with 1125 cities in the data frame.

```
City          0
Country       0
Population    0
latitude      8
longitude     8
dtype: int64
(1133, 5)
```

```
(1125, 5)

City          0
Country       0
Population    0
latitude      0
longitude     0
dtype: int64
```

Last the Foursquare API is used to provide a list of the top 100 venues for each city. Have meaningful results the venues should be within walking distance of the city center since it is used for tourist purposes. The walking distance is defined as 5km. The geographic location from the geopy package is used as input for the Foursquare API. The results consist of a venue type and it's geographic location. When combining this with te previously found data it gives a data frame with 55642 unique rows.

```
(55642, 7)
```

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Tokyo	35.682839	139.759455	Palace Hotel Tokyo (パレスホテル東京)	35.684644	139.761302	Hotel
1	Tokyo	35.682839	139.759455	Kokyo Gaien (皇居外苑)	35.679928	139.758562	Garden
2	Tokyo	35.682839	139.759455	Wolfgang's Steakhouse	35.679185	139.762134	Steakhouse
3	Tokyo	35.682839	139.759455	Aman Tokyo (アマン東京)	35.685236	139.765401	Hotel
4	Tokyo	35.682839	139.759455	Mitsubishi Ichigokan Museum (三菱一号館美術館)	35.678420	139.763260	Art Museum

Om further analysis it's found that only 417 cities have 100 venues. To keep the data consistent only those cities will be used in the final analysis. Furthermore, if a major city has les then 100 venues registered on Foursquare within a 5km radius of it's city center, it probably won't be a tourist destination.

Venue_Count	City
200	1
100	417
99	1
98	1
97	4

Methodology

The data frame with 417 cities and their top 100 venues is the basis for the analysis. They will be grouped on similarity of venue type. First there will be explored how often each type of venue is found in every city. Then this will be converted in a top 10 venue type for each city. This data can be used to group them. This will be done with K-means analysis. K-means will provide clusters which in turn translate to the different groups of cities.

In the data exploration to get to the top 10 venue type per city, dummy variables are created. First it's counted how often each type of venue is found in each particular city. Then te data is normalized to make every count comparable with other counts.

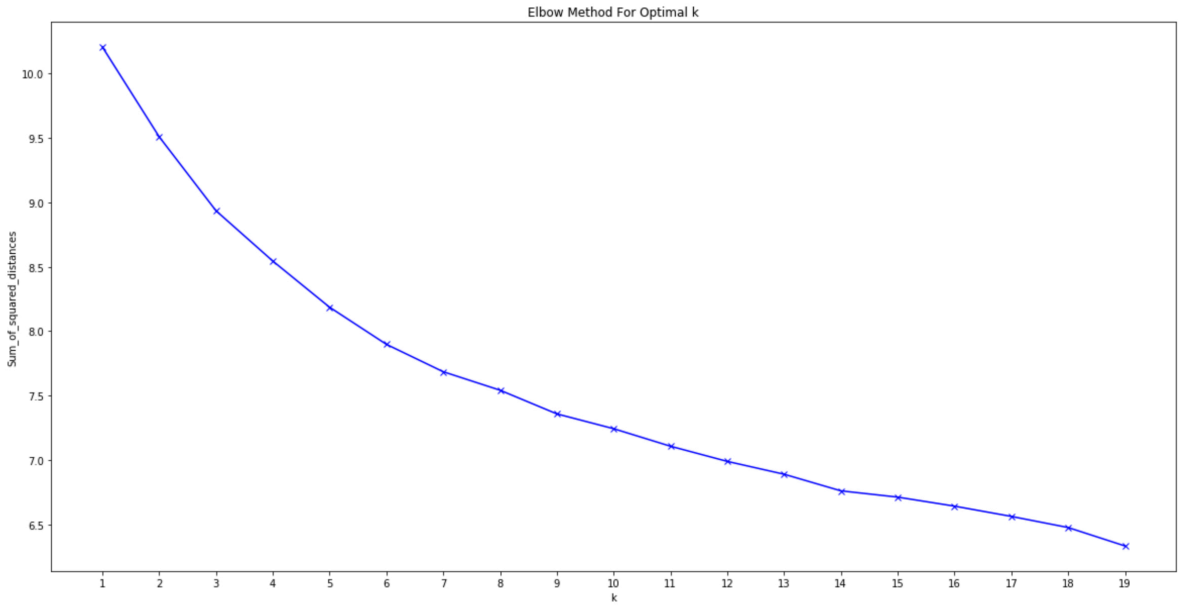
	City	Zoo Exhibit	ATM	Acai House	Accessories Store	Adult Boutique	Adult Education Center	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant
0	Abu Dhabi	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.00	0.0	0.0	0.0	0.0	0.00
1	Abuja	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.03	0.0	0.0	0.0	0.0	0.01
2	Acapulco De Juarez	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.01
3	Accra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.08	0.0	0.0	0.0	0.0	0.01
4	Ad Dammam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.00

The normalized data frame can be used to determine what are the top 10 venue types of every city.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abu Dhabi	Café	Hotel	Middle Eastern Restaurant	Coffee Shop	Chinese Restaurant	Restaurant	Park	Dessert Shop	Asian Restaurant	Beach
1	Abuja	Hotel	Restaurant	Lounge	Fast Food Restaurant	Arcade	Nightclub	Café	Pizza Place	BBQ Joint	Movie Theater
2	Acapulco De Juarez	Seafood Restaurant	Mexican Restaurant	Taco Place	Restaurant	Beach	Coffee Shop	Food Truck	Italian Restaurant	Burger Joint	Ice Cream Shop
3	Accra	Hotel	African Restaurant	Fast Food Restaurant	Bar	Cocktail Bar	Shopping Mall	Bakery	Pizza Place	Nightclub	Restaurant
4	Ad Dammam	Coffee Shop	Bakery	Juice Bar	Ice Cream Shop	Donut Shop	Café	Supermarket	Dessert Shop	Restaurant	Middle Eastern Restaurant

On this data the K-means analysis is performed. In K-means it's necessary to input the number of cluster for the analysis to work. It then will output the best fit for that total number of clusters. The fit of the model is measured as the sum of the squared distance. The better the model is, the lower the squared distance will be. With every extra cluster the squared distance will lower. At some point dis doesn't make sense anymore. The elbow method is used to determine what number of clusters gives the best fit. For the elbow method a plot is made with

the sum of squared distance on the y-axis and the number of clusters on the x-axes. The point where there is a significant decrease in slope of the graph gives the best trade between fit of the model and number of clusters. for this analysis the most noticeable decrease is at 14 clusters.



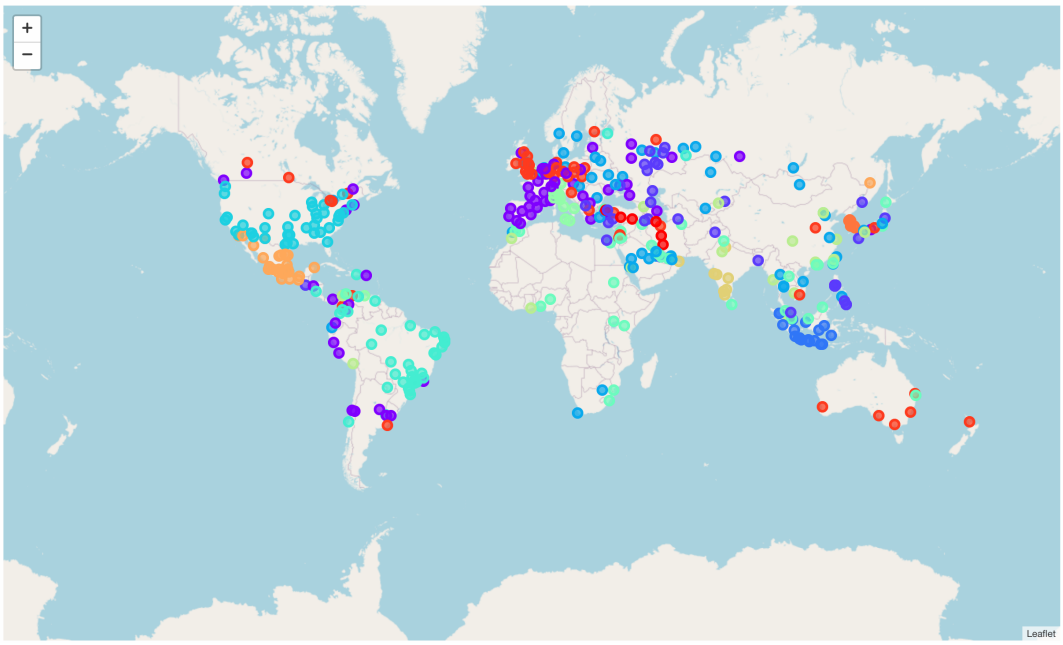
The K-means analysis is done with a input of 14 clusters. This gives a model that makes 14 different groups of cities. This data is combined with the original data frame to give an overview of al 417 cities, their top 10 venue types and their cluster. The final data frame wil further be discussed in the next chapter.

	City	Country	Population	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Tokyo	Japan	37393129	35.682839	139.759455	7.0	Hotel	Coffee Shop	Japanese Restaurant	Soba Restaurant	Café	Ramen Restaurant	Japanese Curry Restaurant	Hobby Shop
1	Delhi	India	30290936	28.651718	77.221939	10.0	Indian Restaurant	Hotel	Café	Lounge	Snack Place	Bakery	History Museum	Breakfast Spot
2	Shanghai	China	27058479	31.232276	121.469207	9.0	Hotel	Coffee Shop	French Restaurant	Spa	Shopping Mall	Dumpling Restaurant	Italian Restaurant	Lounge
3	Sao Paulo	Brazil	22043028	-23.550651	-46.633382	6.0	Italian Restaurant	Ice Cream Shop	Pizza Place	Theater	Bakery	Brazilian Restaurant	Cosmetics Shop	Art Museum
4	Mexico City	Mexico	21782378	19.432630	-99.133178	11.0	Mexican Restaurant	Bakery	Art Museum	Ice Cream Shop	Plaza	History Museum	Taco Place	Deli / Bodega

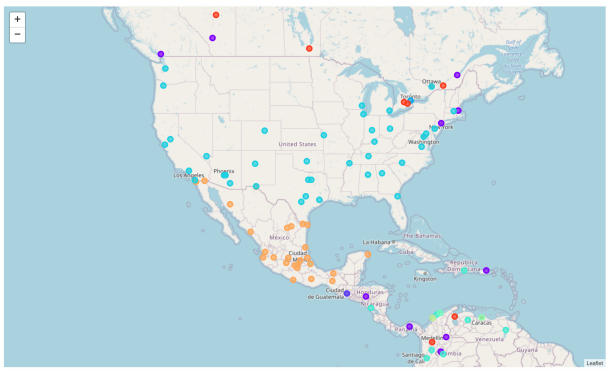
Results

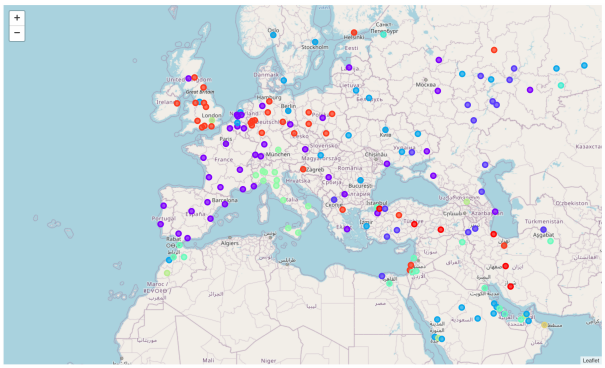
The clustering has resulted in 14 different groups of cities. The groups individual groups are mostly spread on cultural boundaries. This can be shown in a interactive map where each cluster has a different color. A clear distinct color is attributed to Italian style cities. Just like cities in the USA and cities culturally influenced by Spain, Portugal and France for example. Also India has

it's own distinct cluster. And the cultural influence of the UK can also be seen. This is shown in the maps created below.



Cluster Labels	City
13	49
12	8
11	27
10	8
9	21
8	13
7	38
6	41
5	38
4	51
3	21
2	33
1	63
0	6





Discussion

As shown in the results the model is successful in grouping world cities based on cultural links. It clearly outlines some geographic areas with similar cultures like, Indonesia, USA, Mexico, Brazil and Italy. That also means that cities in those groups are likely to be similar.

The model will be able to tell that a customer that enjoyed visiting Rome, that they will probably like Milan. Because of the resources available this model is still limited in helping algorithms decide what city to propose next to a website visitor. To make it more predictive and details cities with less then 500.000

inhabitants should also be included. Next to that it could probably better group the cities if more venues per city are incorporated. To make this happen the computing power should be increased just as the access to the Foursquare API.

This model can be seen as a good exploratory model that tells it is indeed possible to group cities based on the venue data from Foursquare. Further modeling and research is advised to further improve the scope of the model. In the new model a bigger group of cities should be used, by also including cities with less then 500.000 inhabitants. Also a professional account for de Foursquare API is advised to be able to collect more data from it.

Conclusion

In conclusion the model is successful in grouping the major world cities. It groups them on cultural similarities which is also of interest for tourist algorithms. The model can be seen as a good first exploratory model which can be incorporated in existing algorithms. If time and resources are available there is room for further improvement of the model.