**Segmenting University Cities around the Globe**

Diego Saenz

February 02, 2020

1. **Introduction**

This report is going to study and analyze the cities that are ranked best for university students. For many students and agencies, it is quiet troublesome to analyze which city is best for them to pursue a mayor not only on the degrees offered by the university's programs, but to the lifestyle the university students want to have around them. Therefore, it is of mayor interest to cluster the best university cities around the globe in order to provide an overview and a list of similar options to students and agencies.

Using the list of cities created by the **QS World University Rankings**, which can be found at [QS Best Student Cities](https://www.topuniversities.com/city-rankings/2019), we are going to observe similar cities based on the kind of businesses and venues they have around **10km** from their *central business district*. This area of each city gives similar chances to see what kind of venues are agglomerated within the business center which can be helpful to most students that want to work and live in the same city they study their degree.

On another note, the QS rankings were derived from scores on in eight out of 12 categories. Four categories are mandatory, while institutions must choose the remaining four optional categories:

* Teaching
* Employability
* Research
* Internationalization
* Facilities
* Online/Distance Learning
* Arts & Culture
* Innovation
* Inclusiveness
* Social Responsibility
* Subject Ranking
* Program Strength

1. **Data Acquisition and Cleaning**

Using the webpage of QS Ranking, we web scraped the list of 120 cities in the 2019 ranking with their country and rank information using the BeautifulSoup library in order to scrap the information from a json file. Then we obtain the latitude and longitude of each city using the Geopy library and creating addresses for each city with the name of the city and their country in a for loop. We had to clean some of the data because of specific names such as China (Mainland) to China for the Geopy library to find the correct addresses.



Next, using the FourSquare API we search for the most venues possible within the 10 miles radius of each city. As we have the basic plan on FourSquare we were able to find 2500 venues for each city that already give us a pretty good idea of the general culture within the cities. This information was transformed in order to encode the proportion of each kind of venues so that we can later cluster cities having the most similar cultures for entertainment, outdoor spaces and restaurants.

Since every city has a variety of venues available and we can only make a call to 2500 of them for each city we chose the top 10 most common venues to represent each city.

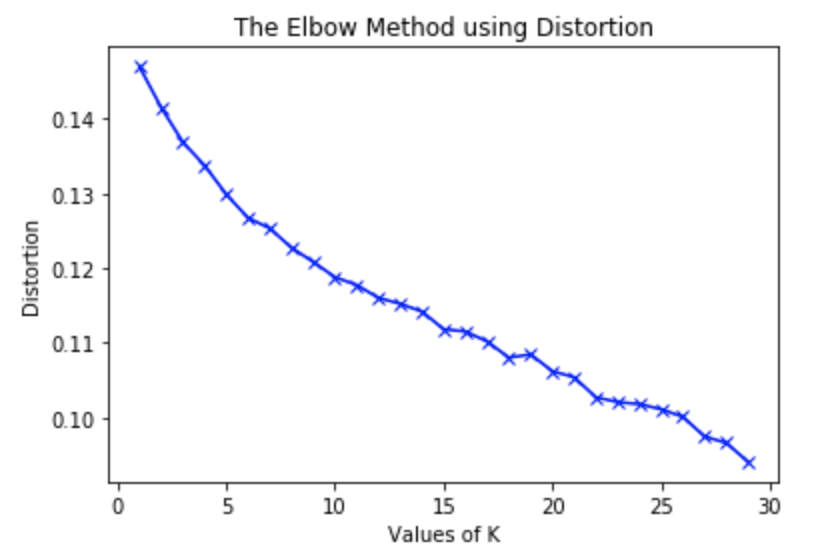


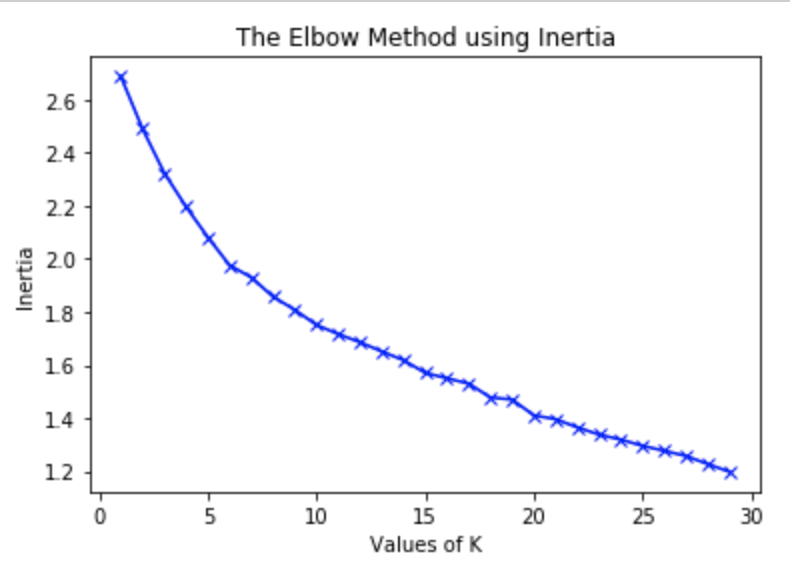
Now we are ready to look for the best number of clusters to group the cities based on information of their most common places.

1. **Exploratory Data Analysis**

Using the Skylearn library we explore the best number of groups of cities based on their venues information.

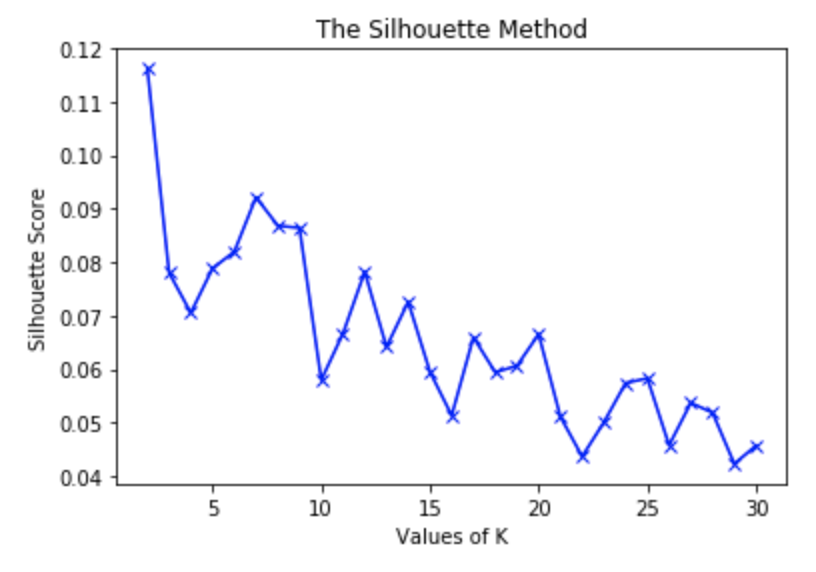
First, we applied the Elbow Method to find the best number of clusters by looking at how compact they are and at the percentage of variance explained as a function of the number of clusters. We looked at the range from 1 to 30 clusters using Distortion from each k of clusters as well as Inertia:





The two graphs above show a turning point at around 5 to 10 clusters where there appears to have compact clusters represented by both metrics. This result needed to be conclusive using the Silhouette Method where we look at how well each datapoint was cluster and how similar it its to its neighbors within each cluster and how different they are to other datapoints in other clusters.

This method shows that we best achieve segmentation using 7 clusters to separate the datapoints for each city:

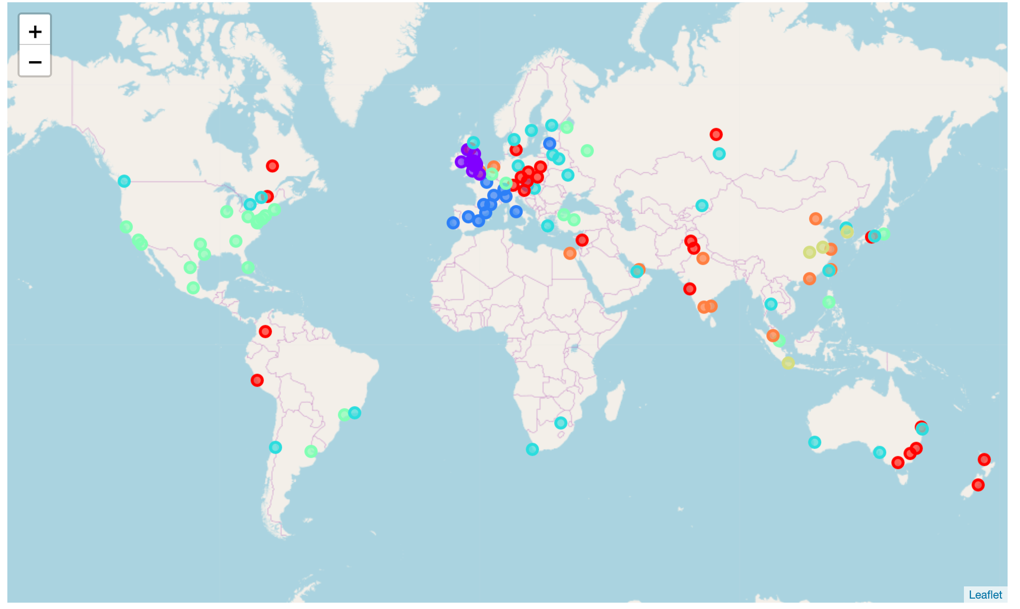


1. **Clustering**

After having come to the conclusion of using only 7 clusters to group the cities, we applied the k-means clustering using the information of the 10 most common venues for each city.

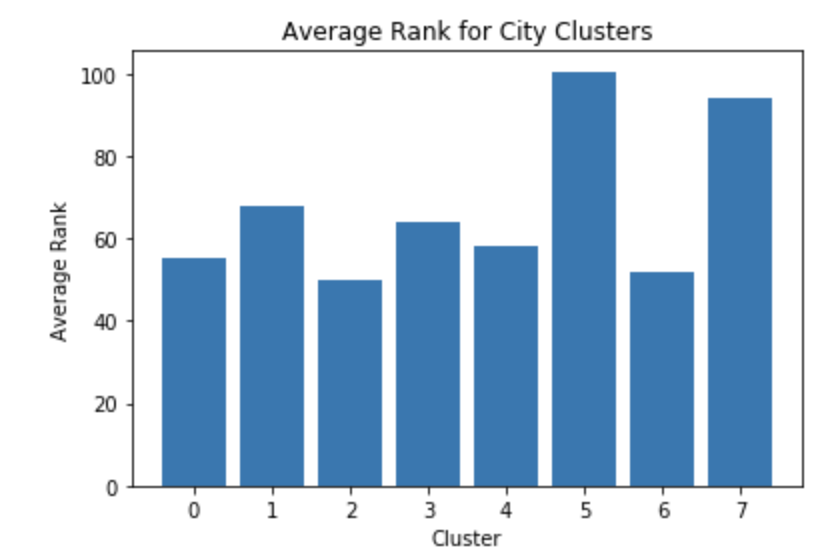


In order to convey this information better we used the Folium library to show how each cluster is spread around the globe:



We can already find some interesting clusters when we observe how British cities tend to be very similar to each other in terms of venues available within 10km radius of their downtown. American and Western European cities tend to have a similar pattern to their neighbor cities as well. Eastern European, Australian and some South American and Asian cities present similar patterns of venues available to the public.

Lets look at how each cluster ranks in the QS Ranking for Best University Cities:



Some insights are:

* Clusters 2 and 6 have the lowest average rank in the **QS Ranking**; this clusters correspond to Western European and Asian Cities.
* Cluster 2 has **Hotels and Plazas** the most around their inner 10km radius.
* The most common venues in cities within cluster 6 are **Hotels**.
* Cluster 0 has cities from **South America, Eastern Europe and Oceania**, making it the most diverse group where **Cafés** are the most common venues in their inner cities.
* Cluster 1 are mostly **British cities** and unsurprisingly the most common venues are **Pubs**.
* Cluster 3 have **Coffee Shops** as their most common venues and are cities located from **Canada** all the way across to **Australia**.

1. **Conclusions**

This report has shown the similarities between cities from the QS Ranking and has provided some interesting insights from the cultures of each city from the perspective of venues available to the public. For any prospective university student and agency, this information could help decide what kind of destination is most suitable and which countries have similar cultures.