**Coursera Capstone Project**

**Problem Statement**

Governments across the world are coining extraordinary ways to serve its citizens increasingly through smart applications. How governments can innovate may often depend on its physical assets — places of recreation, nature, entertainment, eateries, transportation hubs, and so on. Some governments have access to a greater number and diversity of assets than others, which impacts what they can do to become more lively, loveable, and smarter cities. For example, coffee shops oftentimes are a source of paid, sponsored, or free Wi-Fi. The lack of coffee shops may therefore indicate that a city government should direct funding and attention to establishing public Wi-Fi networks.

As a small contribution to the smart city discourse, I will venture to shed light on the combination and scale of assets required to scale up smart city initiatives and programmes. Possible insights may include: which cities have the greatest potential to scale up smart city initiatives by virtue of their combined assets; what are the various pathways to becoming a smart city; which type of assets contribute most to the development of a strong smart city; and so on. City governments can then devise effective mechanisms to stimulate the creation of meaningful assets, in the form of grants, incentives, regulation, programs, platforms, and exchanges.

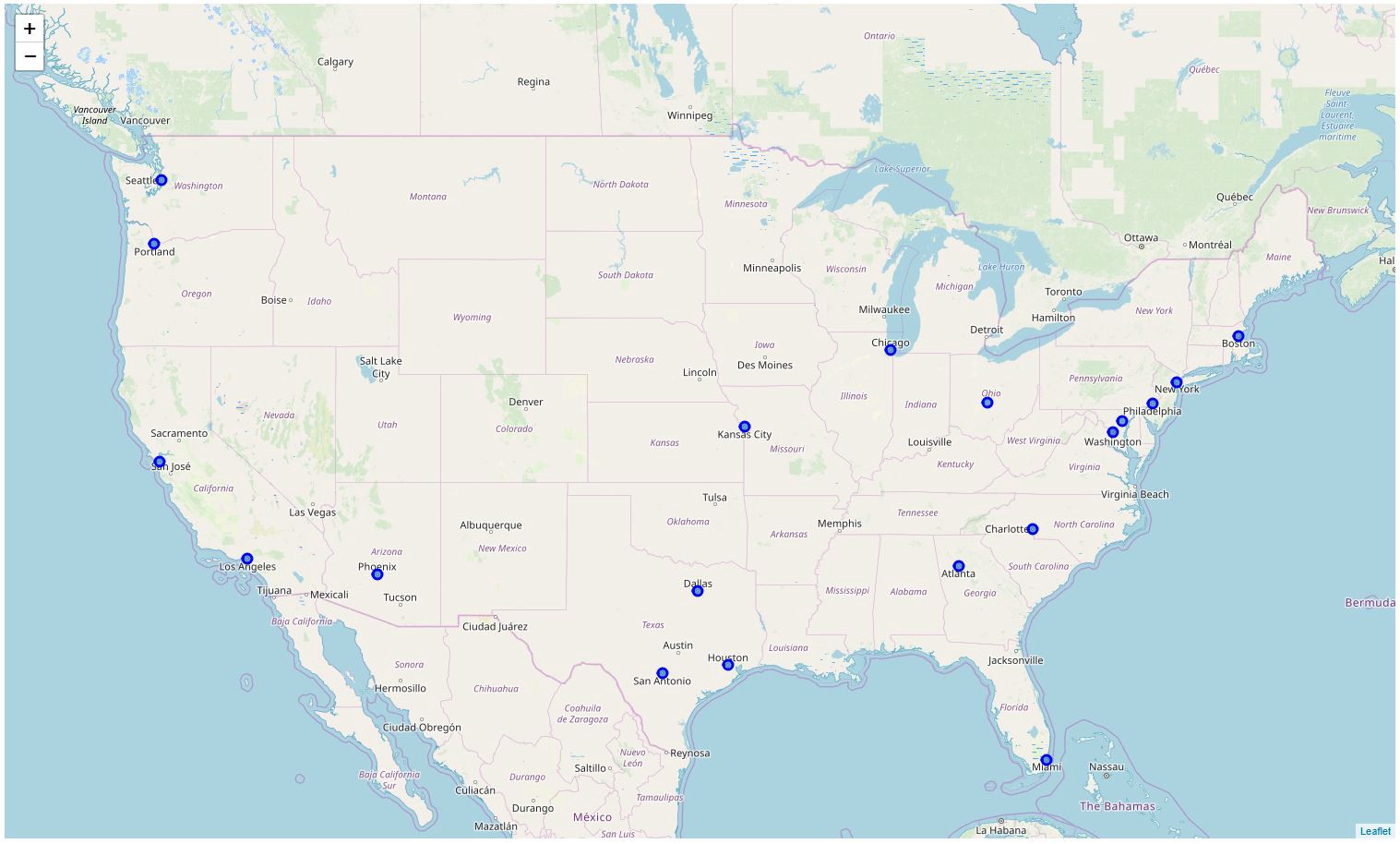
**Foursquare dataset**

The Foursquare data itself will comprise a diversity of venues such as those mentioned as physical assets in the Introduction, including the venue name, venue category, average ratings, and location.

**Methodology**

**Visualizing the cities**

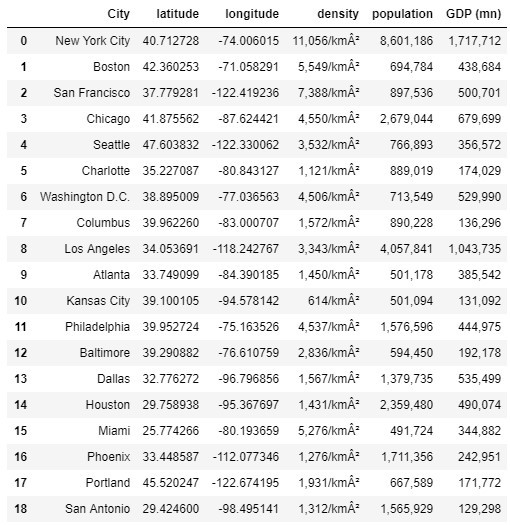
First, the basic city coordinates need to be obtained using *Geolocator* from the *GeoPY* library. Then we are able to plot the following using the *Folium* library.



Cities of evaluated Smart City governments

We can see there is a good spread across the U.S. of the smart cities, which happen to more or less coincide with many major U.S. cities.

The density, population, and metropolitan GDP were scraped from their respective websites using the *BeautifulSoup*and *Ixml*libraries, to come up with the below table.



Raw City dataframe

**Foursquare data**

Because we are looking at venues of the entire city, we need a rather large radius. To illustrate, New York City’s Manhattan alone can stretch for 20km. Thus, a rather large radius of 15km was selected. Foursquare’s API is limited to passing 100 venues per coordinate called, yielding a total of 1,900 venues for our dataset.¹ Of these 1,900 venues, there were 247 categories that were unique. The most common venues can then be determined which will form the basis for clustering. The following table excerpt was produced to give a sense of the types of venues are common in each city.

[1] There are workarounds that could be performed, such as passing through multiple coordinates with minor variations of the original city coordinates or breaking each city down into its composite boroughs and neighborhoods which coordinates could be passed. These techniques could be useful to future studies.



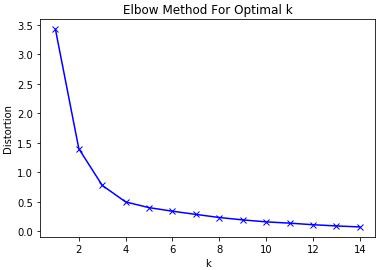
Top 10 venue categories per city

Here we can quickly see that Parks are quite common across each city, but may not be so defining as to nicely trace out clusters.

At this point, the last stages of data preparation and preprocessing were performed. The initial city dataset was max-min normalized to get rid of any potential biases due to the significant differences in figures. The normalized city dataset was merged with the mean occurrence of venues by category in each city. Finally, city, latitude, longitude, and population were removed, with the latter made redundant by density. The clustering dataset then comprised of 19 rows of Density, GDP, and the 247 venue categories.

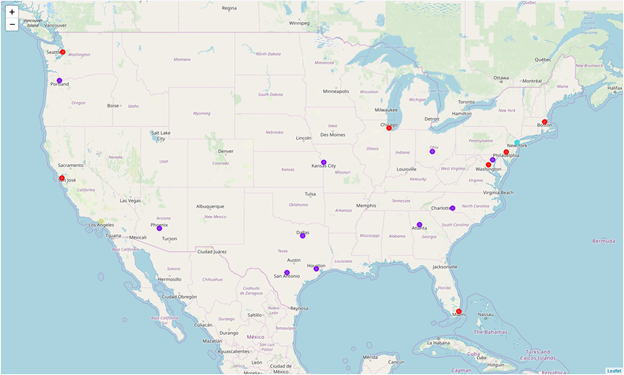
**Clustering**

I used the K-means elbow method to determine the optimal number of clusters, as below.



Optimal k is about k = 3 or k = 4

It seems a toss-up here between k = 3 and k = 4. Because I would like to have some nuance in the clustering, I would prefer to use k = 4. After running the clustering with this k value, the clustering looks rather interesting.



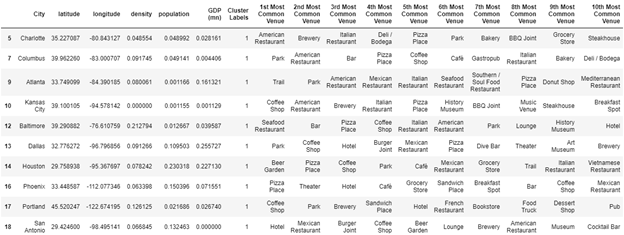
Cities according to cluster

**Results**

The generated clusters shown below had no clear trends in the venue categories, arousing some suspicion as to how they were clustered.



Cluster 0



Cluster 1

https://miro.medium.com/max/780/1*FR48zsGU4AT0BpaOxo1XtQ.png

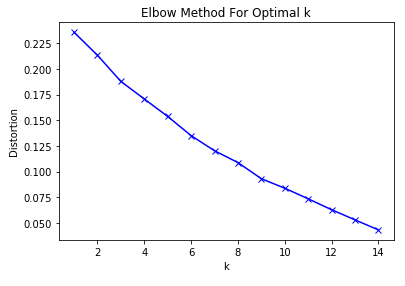
Cluster 2

https://miro.medium.com/max/780/1*7zL3FlkhxdNa1dLVBid0ow.png

Cluster 3

Eyeballing the clusters suggested a bias towards density and GDP (mn). Ironically, these features were meant to be augmenting features, but they ended up being the clusters’ main drivers.

At this point I re-ran the elbow method to find the optimal k value without density and GDP (mn), and the results were quite interesting, if unfortunate.



No optimal k value can be easily seen

Clustering only on the basis of venues produced no discernible k value at all. In that case, 100 venues per city is not a sufficient basis for clustering cities, and alternative approaches may need to be adopted to generate more meaningful content.

**Discussion**

Based on the above, limited insights can be drawn.

* In the first group of clusters, we can see that a city’s density and, to a lesser degree, GDP may have been a factor in their ranking of the Top 50 Smart City Governments. The city governments of more densely populated and somewhat prosperous cities in Cluster 0 were generally ranked higher than those of the relatively more sparsely populated and less prosperous cities in Cluster 1.
* Cluster 0 contains many of the top ranked city governments, which may suggest a pathway for cities of the same cluster. In other words, somewhat densely populated cities such as Philadelphia and Miami are well positioned to quickly ascend the rankings if they draw on the lessons learned by city governments of Boston, San Francisco and others of the same cluster.
* Cluster 1 contains a few shining stars such as the governments of Charlotte and Columbus that can show their relatively more dense and prosperous peers how developing into a smart city can be accomplished through fewer means.
* Some venues are quite common across all cities, Parks in particular. Many city governments pursue parks (in many cases, “digital parks”) as spaces where smart city applications can be implemented.

**Conclusion**

Certainly, a more in-depth look at how cities can be clustered according to venues is needed, the major constraint of which is the API limit. Once overcome, I anticipate a better k-means model can be constructed.

With the information available to us now, however, we can get a rough idea about where further peer-learning of city governments can be prioritized. This study can be further extended by:

* Including the types of smart city applications implemented in each city
* Including cities of other countries
* Including other city data (e.g. crime, weather, poverty level, demographic spread, etc.)