Data

For my project, I will be using the 128 major cities mentioned in the introduction located in the United States and Canada. These were recorded by infoplease and reflect the largest metro areas by population in these two countries. The data included latitudes and longitudes. However, using the infoplease data, the latitudes and longitudes were rounded to the nearest degree leading to large inaccuracies of location when integrating the Foursquare data. For this reason, I took the 128 cities and used Microsoft Excel and the geography tools in the data tab to create accurate latitudes and longitudes for each city. This excel file is used in my included Jupyter Notebook and is also included in my Github repository for easy access of any interested reader.

Once I had the data of the 128 cities with latitude and longitude, I was able to integrate the Foursquare data. Like the previous project, I then located 100 businesses in the radius of the city and broke them down by industry to create scores for the proportion of businesses in each category. This serves as my data which I will be using for clustering as it will identify cities which have similar industry composition. I also ranked the most popular business types of each city for reference.

The data I used is included in the initial steps of my final Juptyer Notebook. I also included a data preparation notebook which is linked in my project submission so that the reader can follow my initial steps in preparation of the clustering analysis. The link is also included below:

https://github.com/AndrewKJacobs2/Coursera_Capstone/blob/master/CourseraFinalProject_DataPrep.ipynb