**Project report of segmentation of Virginia cities concerning data science companies and venues**

**Introduction**

According to the publication on virginiabusiness.com in 2019, Virginia is ranked sixth nationwide in net tech employment and 12th in net tech jobs added 2018. This capstone project tries to understand how data science jobs are geographically distributed and examine if there are any different characteristics among different cities concerning the number of data science job posts and venue types. The project analyzes geographical data collected by clustering Virginia cities based on the number of data science job numbers and venue types. The findings of this project may provide insights to the business owners who seek to open new locations in cities with a high density of tech companies as well as unfulfilled commercial needs in that city. Also, this project provides insights for those employees who are in the process of deciding relocation.

**Data description**

This project combines two data sets. The first dataset is the 10000 data science job posting from the USA provided by user JobsPikr on kaggle.com. This data set covers the job listing for data scientists from the United States extracted from popular job boards via JobsPikr. Although the most valuable analysis falls into the Natural Language Processing domain for this textual-data enriched dataset, for this project, the author only uses a part of the dataset and aggregate it with the query results from Foursquare.com database. The second part of the dataset contains the top 100 venues in each city for all listed cities in the first dataset. In this project, the author uses all features from the combined data set and clusters similar cities using the K-Means clustering technique. Each cluster is evaluated based on the top ten most common venues in each city and then aggregated as the common characteristic for that cluster. The author then plots the results on the map layered with a choropleth map showing the number of data science jobs per county.

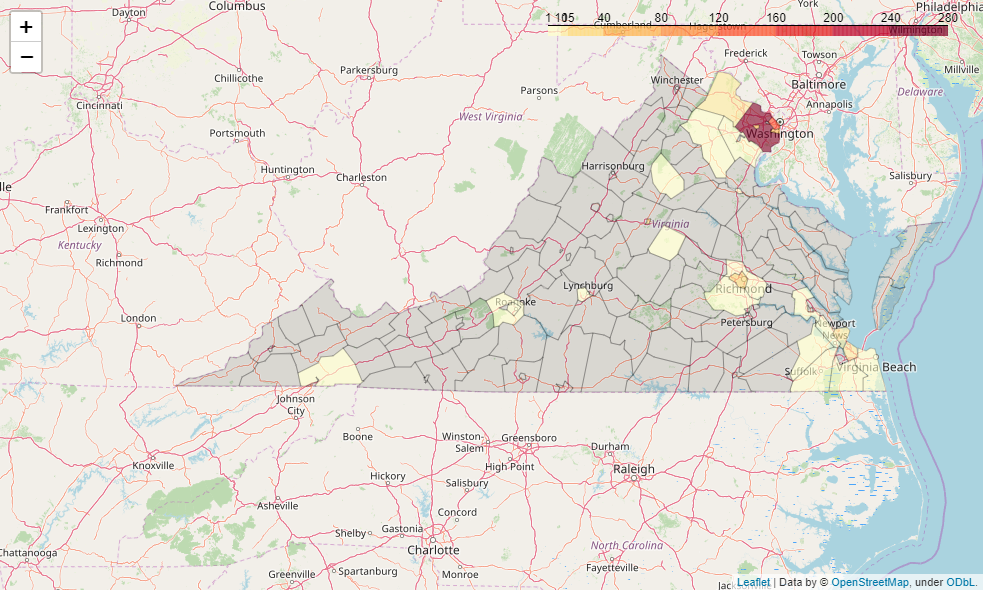
**Methodology**

**Data Exploration**

The primary algorithm used in this project is K-Means clustering. In this project, cities are first selected with the number of job postings from the list of job postings from the JobsPikr dataset. Then the author attaches the top 100 venues to the feature space of corresponding cities. Lastly, all results of clustering are plotted on a map along with the choropleth map showing the density of data science jobs per county. This can help discover both the density distribution of data science jobs in Virginia and the cluster distribution at a count-level perspective. Additionally, the author chooses to use the z-score normalization to scale the resulted dataset prior to clustering.

**K-value selection**

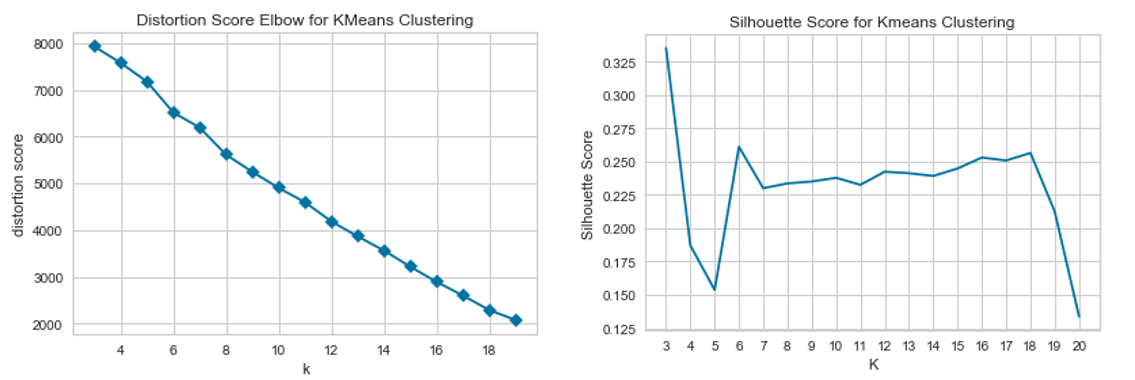
To determine the optimal K for K-Means clustering, the author uses two approaches. First, the minimum value for K is set to be three. This is because, from the data exploration phase, the choropleth map shows that three regions aggregated the most density, namely, Fairfax/Arlington/Alexandria, Richmond, and Virginia Beach. Then the author evaluates the model performance by using the elbow approach and the silhouette score approach. The elbow-point approach calculates the intra-cluster-sum of squared errors (WSS) for different values of K. In a WSS versus K plot, the k value is selected, after which the WSS values first start to diminish. The author also chose to use the silhouette score method as an alternative in case the first approach fails. The elbow-point does not show optimal K value, and the plot is nearly linear during the model evaluation part of the project. In this project, the author uses Euclidian Distance as the metric silhouette score.



The figure shows the choropleth map of data science job density distribution in Virginia.

**Results**

**K-value**

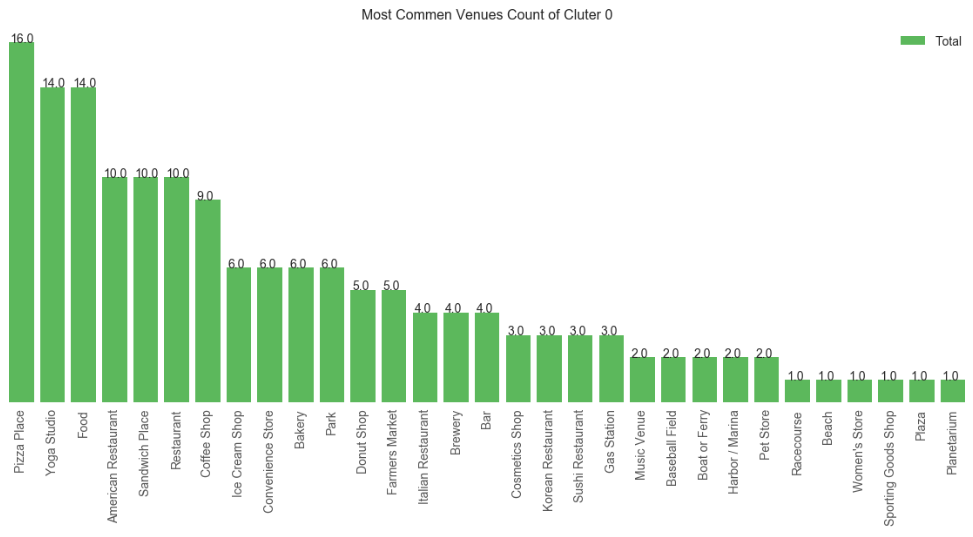


The figure shows the K versus Distortion Score and K versus Silhouette Score plots.

As shown above, the elbow-point approach does not provide the optimal value for K. However, in the K versus Silhouette Score plot, K=3 shows the global optimal outcome. After applying the K-Means model with K=3, the author found there is one cluster that contains the most data points where the other cluster contains only three data points in total.

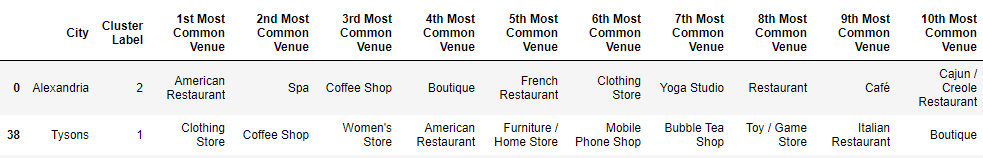
**Cluster Insights**

The first cluster (cluster 0) contains the highest number of data points. The total number of data points in this cluster is 44, whereas the other clusters contain only two distinctive data points in total. The second cluster contains the city Tysons and Tysons Corner, which represents the same geographic location, and the returned venues from the Foursquare dataset are the same. The first cluster covers the most listed cities in Virginia, consisting of 31 different types of venues as the top ten common venues for each corresponding city. The following bar plot shows the distribution of each venue type in the first cluster.



The figure shows the characteristics of cluster 0 from the most common to the lease common.

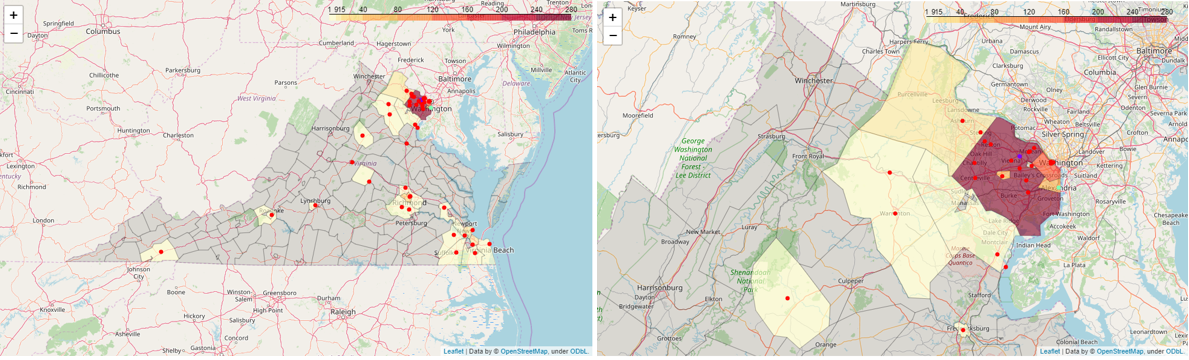
Cluster 0 contains mostly restaurants and other food places with Yoga studio as an exception. It is significantly different for the other two clusters (cluster 1 and cluster 2). By examining the top ten common venues in clusters 1 and 2, the author found that cluster 1 is more leaning towards shopping such as Clothing Store, Women's Store, and Furniture/Home Store, while cluster 2 is a combination of food, shopping, and personal wellness venues.



The figure shows the most common venues in cluster 1 and cluster 2.

**Visualization**

The visualization shows that the three major regions that have the highest Data Science job posts density are the Fairfax/Arlington/Alexandria area, Richmond area, and Virginia Beach area. Among the three, the Fairfax/Arlington/Alexandria area shows the highest density with more than 270 Data Science job posts in total, followed by Arlington with more than 120. Also, it is worth noting that the plot shows that Data Science jobs are geographically aggregated while most of the counties in Virginia do not have any open positions. The two single-datapoint clusters appear in this same region also: Alexandria, an independent city, and Tysons, which is a city in Fairfax.



The figure shows the visualization of the clustering results using Folium. The grey areas are counties that do not have any Data Science job posts

**Discussion**

The first notable finding in this report is that the optimal K value for K-Means Clustering is hard to determine, and the K versus Distortion plot is nearly linear. This shows that the cities that have Data Science job posts are very similar concerning the venue types, therefore not sufficiently separable. When testing the project, the author found that the K-values smaller than the predefined minimum produces a higher Silhouette Score than the selected K-value three. Also, the two single-datapoint clusters appear in the same high-density region. When tuning the clustering model, the author found that as K increases, new cities in this high-density region are clustered into the newly emerged clusters. This could because that the model was capturing noise, however, it may also demonstrate the diversified neighborhoods in this region. Therefore, further analysis can focus on either comparing Data Science job-posting cities and the non-job-posting cities or exploring a smaller dataset targeting only this high-density region.

**Conclusion**

In this project, the author found that Data Science jobs are aggregated in Virginia. The highest-density region is in the Northern Virginia Area, which is adjacent to the D.C. Metro Area. Also, this region shows higher diversity concerning the most common venue types in each city. For business who seeks new opportunities, the food industry in the Virginia area is the most common whereas personal wellness and retail are favored in locations in the Northern Virginia Area.