



Venue-Based Clustering using Foursquare API

COUSERA APPLIED CAPSTONE PROJECT REPORT

NISHANT ZOPE

Table of Contents

EXECUTIVE SUMMARY	2
INTRODUCTION	2
LITERATURE REVIEW.....	3
METHODOLOGY	3
Data Sources.....	3
Data Collection	4
Data Preprocessing and Cleaning.....	4
RESULTS	4
Exploratory analysis.....	4
Modeling	6
DISCUSSION.....	7
Observations	7
Further improvements.....	8
CONCLUSION.....	8
ACKNOWLEDGEMENTS AND REFERENCES	8
APPENDIX.....	8

Executive Summary

In this project, we performed venue-based clustering analysis of all the capital cities of the United States to find similarity between the cities based on the mix of venues that each city has. We used States dataset consisting of latitude and longitudes of cities along with venue dataset obtained using Foursquare API. We performed K-Means clustering with 10 clusters. The model was able to depict similarity between cities that lied in same cluster but failed to justify cities lying in different clusters. Hence, it was observed that this approach of qualitatively analyzing cities to find similarities between them is very feasible, but needs highly refined and granular dataset and use of relatively sophisticated clustering algorithms like density based clustering.

Introduction

Background

Consider two cities, Las Vegas and Atlantic City, which are quite different in terms of landscape, weather, economic, political and cultural aspects. However, the two cities still have similarities in terms of the venues that each these has to offer to its visitors and residents such as – casinos, pubs, hotels and restaurants, etc. This type of characterization of cities could be named as Venue-Based similarity which focuses on finding similarities between locations based on venues and amenities they offer.

Whenever someone is looking for moving to a new city or buying or renting a house in another city, they are highly interested in cities which satisfy their criteria of venues and amenities available. For instance, if a person is nature lover, he might want to move to a city that has significant number of parks and gardens. While several researches comparing cities based on economic and geographic aspects between cities have been extremely useful, a venue-based similarity clustering is relatively less explored. Utilizing the similarity of venues between cities could be helpful for people in making critical decisions.

Problem Statement

Hence, we hypothesize that venue-based clustering is a way of categorizing cities by the ensemble of venues it offers. Hence, we aim to find similarity between cities based on the venues and amenities it offers. For limiting the scope of this project, we would apply this hypothesis on the capital cities in each state of the United States.

Target Audience

Our target audience would be people who are looking to move to a different city or planning to buy or rent a property in another city in the United States.

Literature Review

A lot of research has been conducted on folksonomic characterization of cities based on types of amenities located within them¹. A similar research was conducted Daniel, Justin and Tae in their research paper named 'Exploring venue-based city-to-city similarity measures'¹. This analysis is based on the idea of this research to further analyze the qualitative comparison of cities.

Methodology

Data Requirement

To solve the above problem, we would use the data about capital cities in each state of the United States and top 100 venues within 10-mile radius of these city centers. Specifically, we would need 2 different datasets for this analysis:

1. Dataset with geographical coordinates of all the capital cities in the United States.
2. Dataset with top 100 venues within 10-miles of radius of each city obtained using Foursquare API.

Data Sources

For the data for capital cities of each state of the U.S along with geographical coordinates, we used STATES directory [made available by John Burkardt from The Department of Scientific Computing at Florida State University]², which contains datasets with information on U.S states. The dataset was obtained using 2 different files:

state_capitals_II - Contains latitude and longitude of the capitals of the states, in alphabetical order by state name, including the District of Columbia, Puerto Rico, and the US.

state_name - Contains full names of the states, in alphabetical order by state name, including the District of Columbia, Puerto Rico, and the US.

The files contain a single line of data for each state. The lines are given in alphabetical order of the full state name. Each line begins with a two-letter postal code identifier for the state, whose alphabetic order is "almost" consistent with that of the full state names

For the venues, the data was collected using Foursquare API which is a widely used location based social network. The US city data acquired was used to explore the nearby venues using the Explore functionality of Foursquare API. Top 100 venues in 5 km radius for each city were extracted which contain data in below format/columns:

Venue – Name of the venue.

Venue Latitude – Latitude of the venue address.

Venue Longitude – Longitude of the venue address.

Venue Category – Type of venue.

Data Collection

The US city data was collected by manually downloading the 2 files **state_capitals_II** and **state_name** in csv format. While the Venue data was obtained using the explore function of Foursquare API, which was received in json format which was converted to pandas dataframe.

Data Preprocessing and Cleaning

After the datasets were collected separately, they were cleaned and restructured for modeling. The 2 files **state_capitals_II** and **state_name** were joined using State as the key to create one dataframe named **df** which becomes the base dataset for our analysis. The venue dataset obtained from Foursquare.com was grouped by **City** and venue categories were one-hot-encoded to convert categorical **Venue Category** column into numeric columns (each binary numeric column representing one venue category).

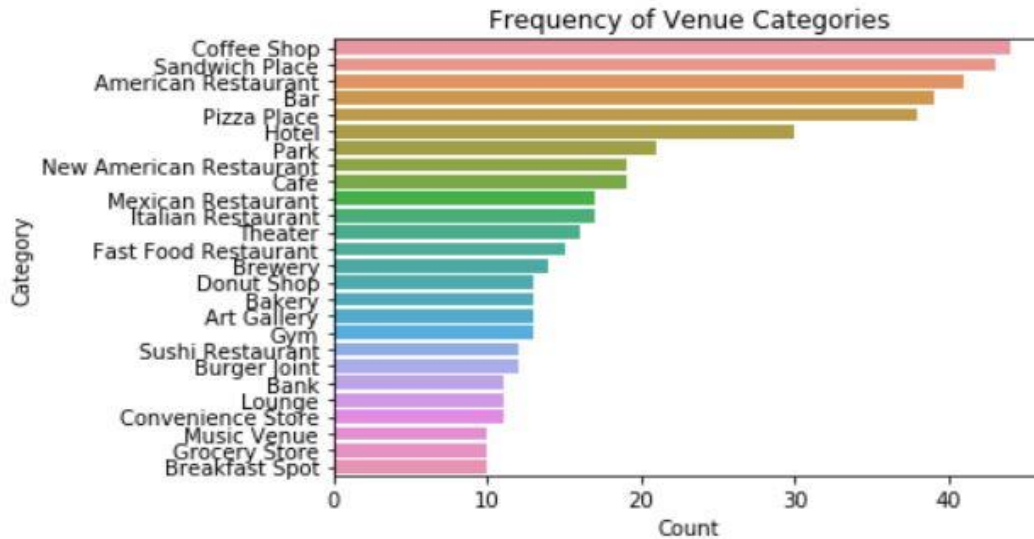
Since, the dataset was obtained in structured format and had no missing or incorrect values, minimal cleaning efforts were needed. Below is the screenshot of final dataset used for clustering:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Montgomery	32.36	-86.28	Shashy's Bakery & Fine Foods	32.362289	-86.283226	Bakery
1	Montgomery	32.36	-86.28	Martin's Restaurant	32.357262	-86.282862	Fried Chicken Joint
2	Montgomery	32.36	-86.28	Subway	32.357502	-86.283664	Sandwich Place
3	Montgomery	32.36	-86.28	Havana Dreamin'	32.358157	-86.281660	Smoke Shop
4	Montgomery	32.36	-86.28	Jimmy John's	32.358197	-86.282985	Sandwich Place

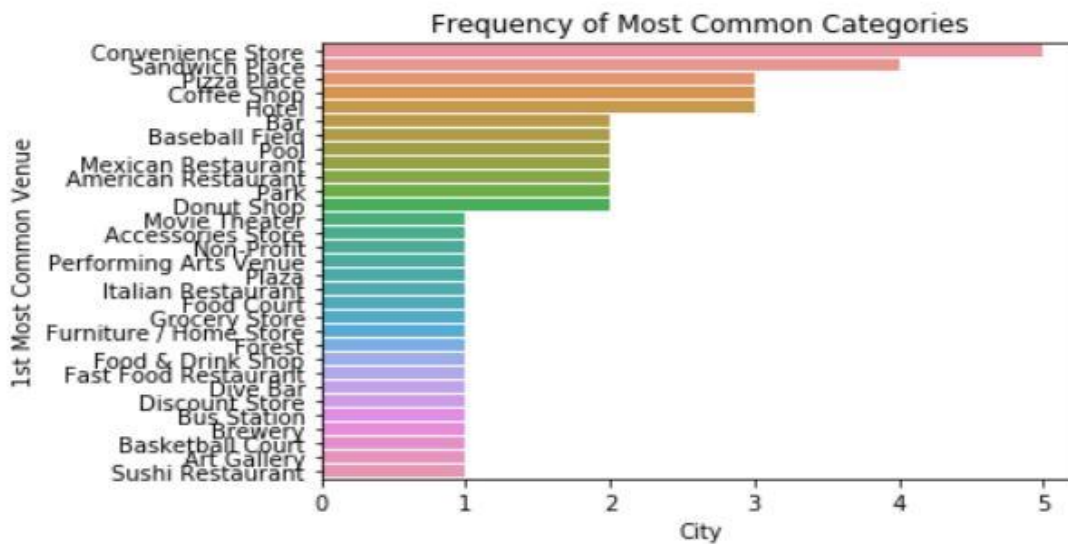
Results

Exploratory analysis

The exploratory analysis of the dataset was mostly focused on determining the distribution on Venue Categories and the frequency of occurrence of Most Common Venues across the cities. Below plot shows that Food related venues like Coffee Shops, Restaurant and Sandwich places are more common in these cities followed by Hotels and Parks. However, it is noticed that the categorization of the venues is not accurate for this clustering approach as there are various categories which fall under one category. For example, American Restaurant and New American Restaurants must be included in one category – Restaurant.



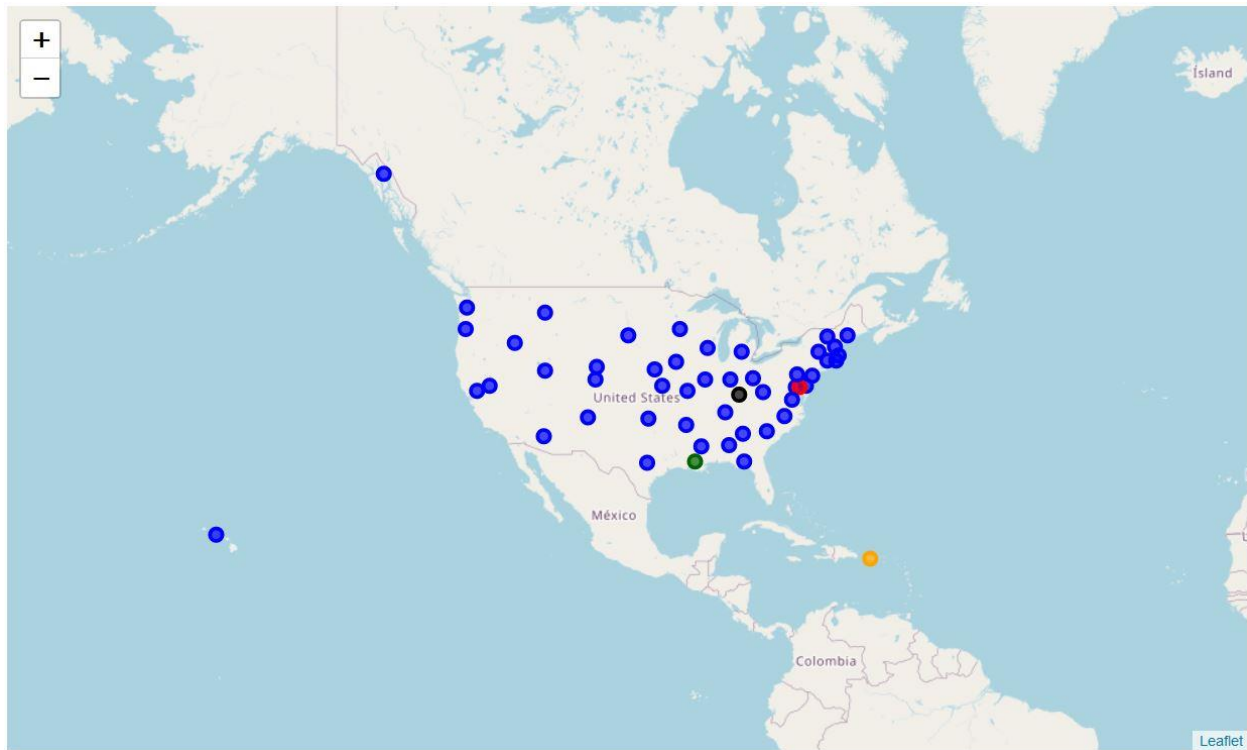
Below plot shows that most of the capital cities in the US have Convenience Store and Eateries like Sandwich place, coffee shops and pizza places.



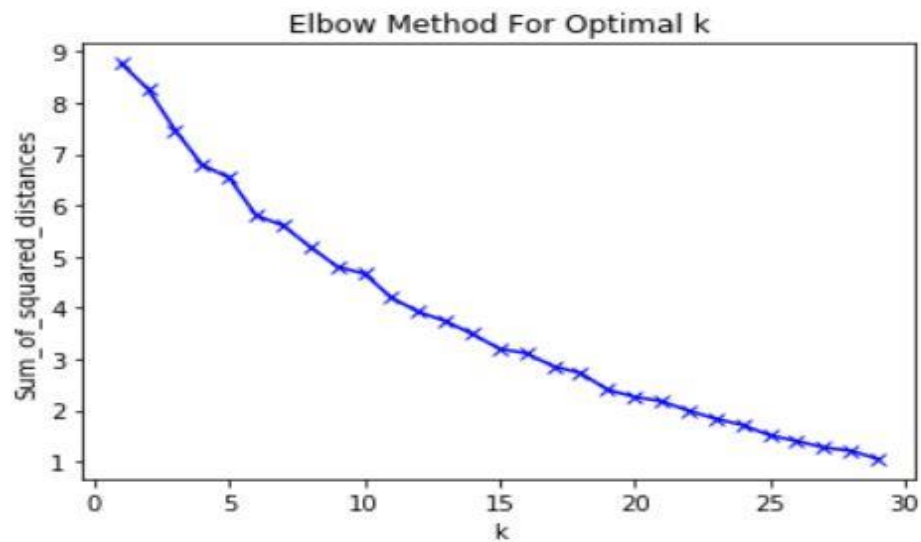
The above analysis of Most Common Categories also gives an initial estimate of clusters of $K = 5$ would be a good starting point.

Modeling

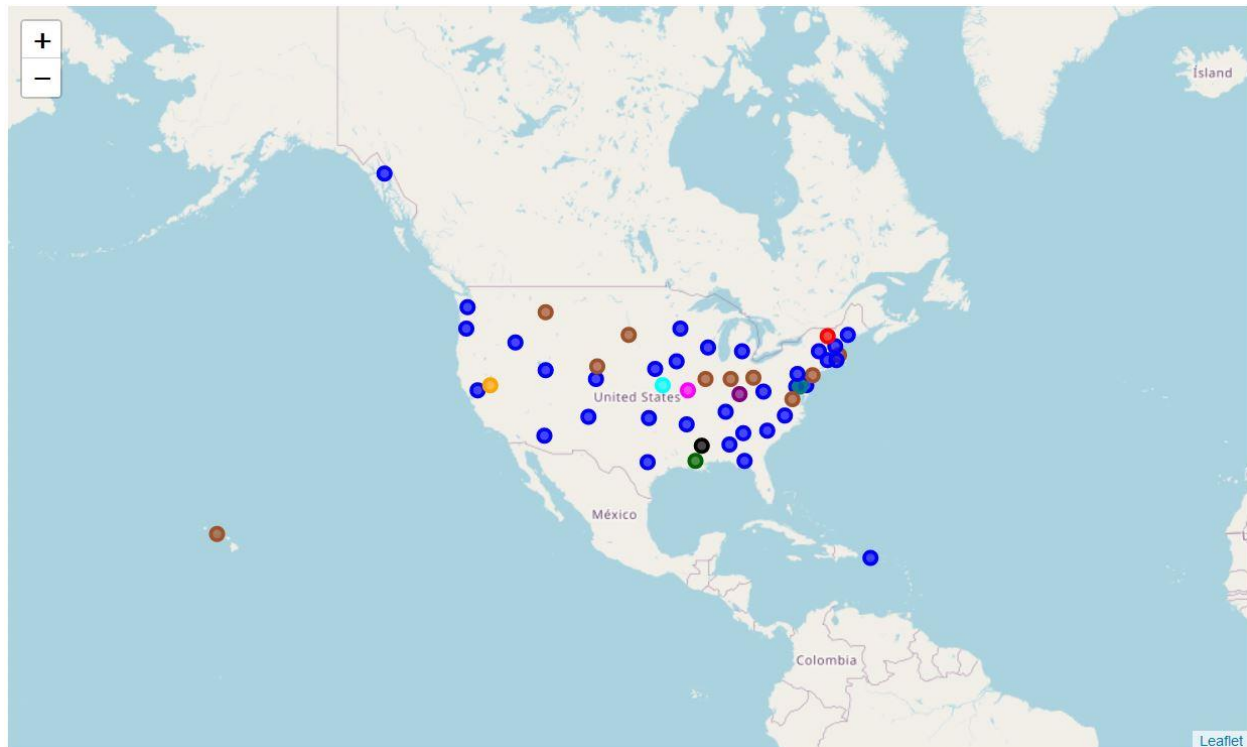
K-Means clustering algorithm was used for modeling with initial number of clusters as 5, which we picked after exploratory analysis of the dataset. Below are the clusters obtained with $K = 5$:



Clearly, the number of clusters seems incorrect as one cluster contains 99% of the data points. Hence, to find the optimal number of clusters for this problem, we used Elbow method to determine the value of K , which was obtained by comparing the error for each value of K from 1 to 25. Below is the plot:



From above plot, it was observed although 5 was a good starting point, the distance between clusters is getting lesser with increasing value of K. Hence, we need to find an elbow for higher value of K. However, increasing K is not the only solution we have to make sure that the clustering is justified with the K selected. For example, selecting K greater than 20 would mean that in best case most of the clusters would contain 2 or 3 cities and in worst case most of the clusters would be empty or would contain only one city which is not practical. Hence, we select K = 10 as there is elbow at this value and these many number of clusters seems justifiable. Below are the clusters obtained using K = 10 which looks far better than previous clustering:



Discussion

Observations

From above cluster analysis, we see that cluster 1 and cluster 6 has cities that look similar when we compare the venues which shows that the approach though is naive still results in clusters with similar data points or cities. For example, Cluster 1 which has cities like Albany, Denver and Atlanta have Restaurants as most common venues, while Cluster 6 consists of cities like Helena and Columbus where Parks are the most common venues. However, only these 2 clusters consist multiple data points or cities and rest of the clusters are formed with only one city.

Further improvements

Above observations imply that this clustering approach is not sufficient enough for determining venue-based similarities between cities. There could be multiple reasons behind this. Some of them are:

1. The Venue Categories obtained from Foursquare database are overlapping which makes it difficult for clustering algorithm to find similar cluster points. For instance, there are sub-classes of restaurants like Mexican Restaurant and American Restaurant which must be combined into one single category of Restaurants, because the choice of restaurant would vary person to person in the city.
2. More sophisticated algorithms could be used like Density based Clustering to create more accurate clustering.
3. More data could be used for such complex analysis as the current analysis is only based on top 100 venues within 10 miles of radius. Increasing venues and radius might also impact the efficacy of this clustering algorithm.

Conclusion



This approach of venue-based clustering seems a good way of finding similarities between different cities across the United States, however, this approach needs more granular and refined datasets along with more sophisticated algorithms to provide accurate clustering results.

Acknowledgements and References

1. <https://www.sas.upenn.edu/~danielpr/files/cities13urbcomp.pdf>
2. <https://people.sc.fsu.edu/~jburkardt/datasets/states/states.html>

Appendix

Below are the US Capital datasets:

US Capitals with Latitude and Longitude	 state_capitals_ll.csv
US Capital Cities	 state_capitals_name.csv