# The Battle of Neighbourhoods

## Final Project

Karl De Ruyck

A report presented in due course of the capstone project from the

IBM Data Science Professional Certificate.


coursera

by IBM at coursera.org

June 25, 2021

# Contents

# Chapter 1

# Introduction

## 1.1 Background

The United States of America constitutes a federation of states, each of which has its own capital city, as identified in Fig. 1.1. These state capitals are not always the most populous city in the state, but are usually considered the location of each state's legislative, judicial, and/or executive authority.

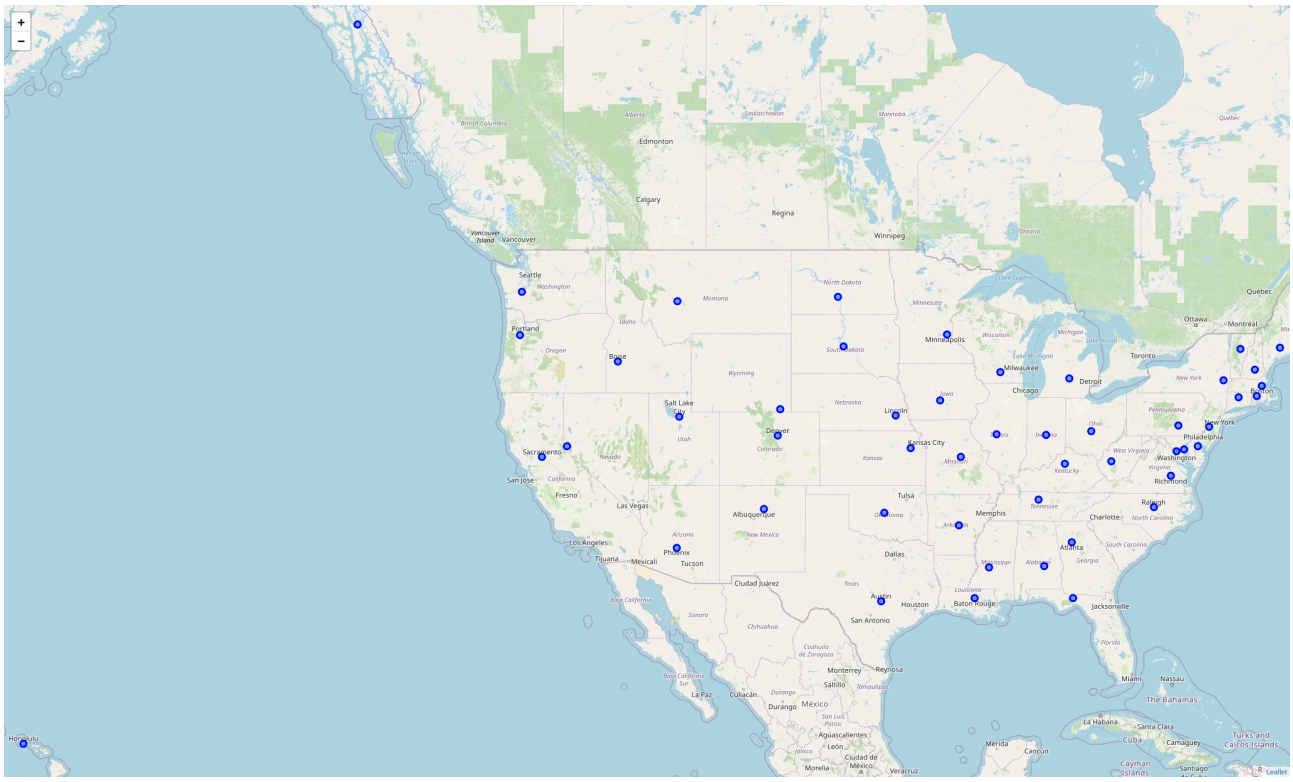


Figure 1.1: Map of the United States of America, with location of state capital cities marked by blue circles.

Considering the intimate relationship between a state's capital city and the interests of that state, it could be assumed that the state capital is representative of the state in even more ways than those previously mentioned. For example, it could be assumed that the culinary specialities of a state may be well represented in its capital city.

On the other hand, modern society is under strong influences toward global homogeneity. Consequently, it could also be assumed that all the capital cities of American states would be very similar in terms of popularly available cuisine.

## 1.2 Business Problem

Which assumption posited above is accurate? Can each state capital *differentiate themselves* by popular types of restaurant, or do people in state capitals *all prefer the same* types of food? Such an investigation could also produce mixed results, where *some* state capitals are similar to each other, while others stand apart.

## 1.3 Interest

Observing these trends would provide guidance to a variety of decisions:

- Restaurateurs may understand where to expand without needing to change their menus to suit the local palate.
- Tourists may be enticed to cities for a cuisine they wouldn't easily find elsewhere.
- Anthropologists may recognize patterns in human movement that link groups of state capitals with similar tastes.
- Alex Aklson may allow me to graduate from this course.

Consequently, it can be demonstrated that this is really essential research, and we cannot escape having to go through with it.

# Chapter 2

# Methodology

## 2.1 Data Source

### 2.1.1 State Capitals and GPS Co-Ordinates

A list of American states and respective capital cities was available on the Wikipedia website, and downloaded from the following URL:

`https://en.wikipedia.org/wiki/List_of_capitals_in_the_United_States`

Each state and the name of its capital city was extracted, assembling a basic data frame. The name of each capital city was then passed to Nominator for lookup on OpenStreetMaps, in order to obtain the GPS co-ordinates of the city. Latitudes and longitudes were added to each city in the data frame.

### 2.1.2 Popularity Contest

The location data for each city was passed to the Foursquare REST API with an 'explore' endpoint, along with keys indicating that the category ID should match 'Food' and that the results should be ranked by popularity. This query returned a JSON object including up to 10 of the most popular restaurants or eateries within a 2 km radius of the geographical centre of each capital city.

The JSON object was parsed to extract the name of each restaurant, its location, and the type of food served there. A new data frame was then created to list each restaurant returned as popular venues in each capital city, and include the aforementioned variables alongside each listing.

## 2.2 Exploratory Analysis

By grouping restaurants according to the type of food served, it was possible to visualize the variety in each city. As shown in Fig 2.1 below, cities where certain cuisines are predominant among their top 10 restaurants can be distinguished from cities where many different cuisines were represented. By observing long bars of a single colour, cities with predominant popular cuisines are identifiable.


Figure 2.1: Stacked bar chart showing culinary heterogeneity in each capital city.

## 2.3 Clustering Capital Cities

### 2.3.1 Data Preparation

Each city was characterized by the types of food served at its top 10 most popular restaurants. This data was then reduced further by counting the number of restaurants serving each of these cuisines, thus enabling a ranking for popular cuisines in each city to be created.

The length of this ranking was experimented with, from the top 3 to the top 10 cuisines in each city. Finally, the top 5 were considered the most representative and the most conducive to effective clustering.

### 2.3.2 k-means Clustering

A machine learning package that implements the $k$-means clustering algorithm was used to address the business problems as stated in Section 1.2. Since clustering is an unsupervised learning problem, this implementation constitutes an appropriate demonstration of machine learning techniques, as applied to solving a practical problem.

The number of clusters $k$ that the 50 capital cities in the dataset were divided amongst was experimented with, from 3 clusters to 10 clusters. Some experiments produced clusters with only one capital city, while some experiments produced clusters with many cities that did not appear to have strong similarities in terms of culinary popularity rankings. Finally, using 8 clusters was found to produce well-populated clusters with distinctive features.

# Chapter 3

# Results

## 3.1 Clusters Identified

After assigning each of the 50 capital cities in the dataset to one of 8 clusters, each cluster was observed to be dominated by certain cuisines, which was the expected outcome of applying the algorithm as previously described. In one cluster of three cities, Food Courts were the most popular type of restaurant. A cluster of six cities was dominated by Mexican restaurants, while another cluster of four cities appeared to prefer Bakeries and Seafood restaurants.

Interestingly, two clusters were heavily dominated by Fast Food as the most popular type of restaurant. However, one of these clusters included a variety of international cuisines amongst its cities' top 5s, while cities in the other cluster listed Breakfast and Steakhouse restaurants among their top 5s.

## 3.2 Geolocation of Clusters

In order to draw further insights from the clustering analysis, the capital cities in each cluster were plotted to a map of the United States of America, with each cluster depicted using a different colour, as shown in Fig. 3.1 on the following page.

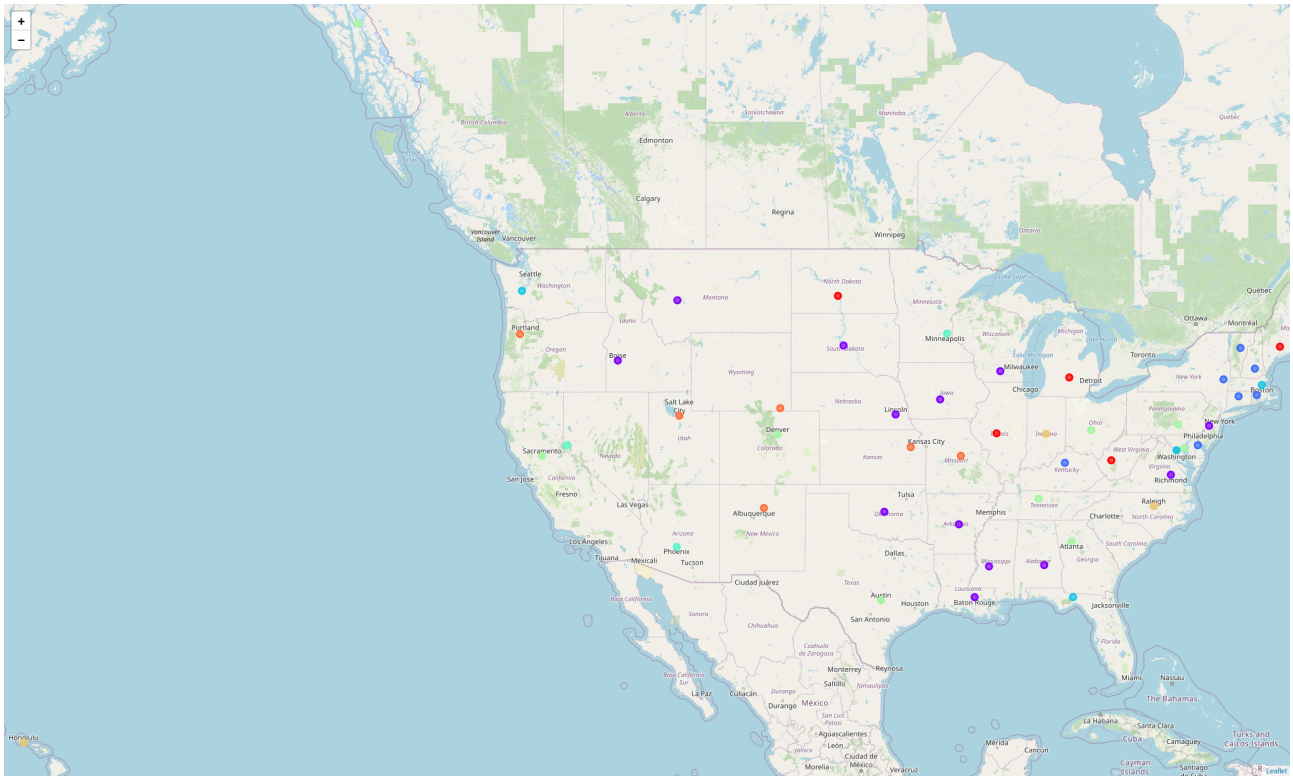


**Figure 3.1:** Map of state capital cities in the United States of America, with locations marked by coloured circles. The colour of each city's circle indicates divisions identified by $k$-means cluster analysis of popular cuisines.

Observing the map in Fig 3.1 allowed identification of geographical regions in the United States of America where capital cities have similarities in terms of popular types of restaurant. For example, an orange cluster appears to identify several cities in the southern central states. A purple cluster indicates that cities above and below the orange cluster appear to have similarities. There is a blue cluster that only represents states to the east, with the majority coming from a small region to the north-east.

Another way of looking at Fig 3.1 shows that most state capitals are surrounded by state capitals with different tastes. That is to say, there is no state that shares all borders with other states that are all in its own cluster. Some states, like Mississippi and Vermont, are bordered by several states whose capitals cluster together, but are also bordered by one state whose capital belongs to a different cluster.

Some clusters appear to be very weakly related to geolocation. For example, a light blue cluster includes three capital cities spread from the top to the bottom of the east coast, and a fourth city far to the north-west. Similarly, a yellow cluster includes the capital cities of North Carolina, Indiana, and Hawaii, which are all quite far from each other. Although each of these three states possesses a coastline, one is on the Pacific Ocean, another on the Atlantic Ocean, while the third is on Lake Michigan!

# Chapter 4

# Discussion

## 4.1 Assumptions and Sources of Bias

Several interesting trends were observed amongst the types of restaurant popular in each capital city. However, there are some important caveats to keep in mind when interpreting the results of the previous chapter.

Firstly, the data source must be considered as a biased representation of restaurant types in the American capital cities. If a restaurant was not listed by Foursquare, it would have been omitted from the analysis, regardless of its popularity. Similarly, if a restaurant's patrons did not check-in to Foursquare when dining, then that restaurant's recorded popularity would be deflated below the actual level.

Another issue potentially arising from the use of Foursquare data relates to reliance of the present analysis on some of Foursquare's own analyses, which were not validated in the course of this work. For example, the categorization of restaurants according to the type of food served there may not represent the types of food that customers actually tend to order there. It could even be possible that the restaurant owner or kitchen staff consider the food they serve as belonging to a different category than how it is listed by Foursquare.

Using a 2 km search radius, centred on the geographical centre of the city, as the criteria for listing restaurants that would represent that city, may also reduce the reliability of this analysis. To begin with, it was assumed that the Nominator package would return geographical co-ordinates that accurately represented the centre of each capital city. However, modern American capital cities can be very large in area, and it is not unreasonable to suspect that major culinary centres or even a few highly popular restaurants may lie outside a 2km radius of the city's geographical centre.

## 4.2 Business Problem

The points raised above notwithstanding, this analysis does offer some insight to the types of food that are popular in each American capital city. Addressing the previously stated business problem directly, this analytical implementation of a machine learning technique embraces the third position: *some* state capitals are similar to each other, while others stand apart.

Furthermore, the results of this analysis allow enumeration of responses to each of the interest groups listed in Section 1.3:

1. The city where a successful restaurant is currently situated may be identified as part of a cluster. This cluster further identifies other cities where an entrepreneurial restaurateur may seek to expand their business without adjusting their menu.
2. Persons seeking different types of restaurants should only have to cross a single state border in order to access a capital city that belongs to a different cluster than the one where they came from. Which state border(s) they could cross to achieve this may be identified from the map in Fig 3.1.
3. The identification of various clusters of capital cities with similar types of popular restaurant has shown that clusters may be but are not always also related by

geolocation. Therefore, an investigative anthropologist may attempt to use these clusterings to explain other trends in human behaviour, or vice versa.

4. The criteria listed by Alex Aklson as required components of this report and the analytical steps that should precede it have all been addressed. Consequently, I remain optimistic regarding potentially imminent fulfilment of this course's assessed requirements.

# Chapter 5

# Conclusions

Several interesting trends were identified by this analysis, and knowledge of these trends has good potential for utility to all parties interested in the initial business case. *Some* state capitals have similar types of popular restaurant, while other capital cities are similar in different ways. Eight clusters were identified by applying the k-means unsupervised clustering machine learning approach. Some of these clusters included cities with geographic similarities, while other clusters did not appear strongly associated with geolocation.

Application of these results should take into account several caveats mentioned in Section 4.1, and future works extending the present body may seek to address and eliminate these, in order to strengthen the reliability of the study. Critical input may also be received from interest groups seeking to apply the results of this analysis, helping shape future analyses toward greater specific utility, as per each group's precise requirements.