**Austin, Texas Venue Clusters**

Emmitt Luna

March 26, 2019

1. **Introduction**
   1. **Background and Problem**

I lived in the city of Austin, Texas for 4 years. I personally saw how the city was increasing in different types of businesses. HoIver, Austin is a small city which can make it difficult to place new businesses in already crowded locations like downtown. The crowdedness in a single area is not the only concern, but that single area can also have similar types of businesses.

* 1. **Analytic Approach**

In this report I will cluster similar venues around the different zip codes in Austin. New business owners should then be able to use this report to strategically place their new business in Austin. With the help of Foursquare, I should be able to: see the amount of businesses in an area, the types of businesses, and the most common venues of an area.

1. **Data acquisition and cleaning**
   1. **Data Requirements**

The data will be similar to the data used in the IBM course, which consists of: Zip Code, Longitude, Latitude, Venue, and Venue Category.

I need to acquire Austin’s zip code data and the coordinates of each zip code. I will then use the zip code data in order to pull the list of venues within that zip code. I will also pull the type of venue, the category, which help with the clustering algorithm.

* 1. **Data Collection**

I will collect data from <https://www.zip-codes.com/m/city/tx-austin.asp> which has the zip codes for Austin. I will also collect the latitude and longitude from this Ibsite with the use of a function.

**Foursquare** will then provide me with the venue data around the zip codes.

The Latitude and Longitude data will be used by the library **Folium** so a map and the markers can be created.

The data frame will look like below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Index** | **ZipCode** | **Latitude** | **Longitude** | **1st Common Venue** | **2nd Common Venue** | **...** | **10th Common Venue** |

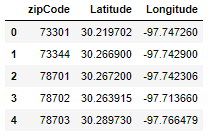
1. **Exploratory Data Analysis**
   1. **Data understanding**

Our data is pretty straightforward in the beginning; however, I will need to get rid of any categorical values and unnecessary data which will not be needed when clustering the data.

I have chosen to use all of the zip codes in Austin instead of only choosing a part of Austin because Austin does not have many zip codes.

Table 1. Our first data frame after pulling the information from the Website

<https://www.zip-codes.com/m/city/tx-austin.asp> using df.**head()**



* 1. **Data Preparation**

Since some parts of Austin varies in size, I decided to keep the radius of the search to about half a mile, around 804.672 meters. This way there can be enough data for each zip code. The data frame will remain simple in the beginning whenever I am pulling the venue information.

TABLE IMAGE

I will then remove the categorical data and change it to integers by using one-hot encoding. Using one hot encoding will help us count the amount of times a venue type appears in a zip code.

A good idea to take on the data is to get the mean, instead of count of each area since some areas will be more crowded than others. This way I can easily compare each area.

TABLE IMAGE

1. **Predictive Modeling**
   1. **Modeling**

I will be using the same algorithm as in our lab, which is KMeans. We are dealing with unlabeled data and we are also clustering to find similarities in the Austin City.

* 1. **Evaluation**

1. **Conclusions**
2. **Future Directions**