Coursera IBM Applied Data Science Capstone

Choosing a location for a hookah bar in Austin, TX

Gabriel Alobo

June 25, 2019

Table of Contents

Summary

This project uses data from various sources and a machine learning algorithm to advise a potential small business entrepreneur on the location of a new hookah bar in Austin, Texas. The reasons for why the current time and Austin as a location make for a great opportunity are laid out and the data to be gathered is described and explained.

The decision process behind major steps in the collection, sorting, and processing stages of data handling are explained. The steps and outputs of the modelling part are explained to demonstrate how they bring us closer to decision.

There is a brief discussion of the outcomes and how they are used to arrive at candidate locations in which an entrepreneur can be highly confident that a hookah bar will succeed.

1. Introduction

1.1 Background

Austin, TX is a fast-growing city in many sectors especially technology and startups. The reasons for its growth are access to talent from universities in Austin and nearby, lower tax burden than Silicon Valley and other areas in California, while having similar weather and young population with lots of disposable income. The city will continue to grow for a long time to come and will remain economically strong into the farthest foreseeable future. This would be an excellent time for an entrepreneur to invest in Austin before barriers to entry become as high as they are in Los Angeles, San Francisco, or New York.

A hookah bar would be an excellent investment to make in this city because it is niche, highly profitable, and attracts the type of young multicultural demographic Austin has. A niche business category like this should be the focus of new investors to long-established cities like Austin because older legacy business categories would have been saturated long ago and only yield small profit margins for those who manage to wedge in.

1.2 Problem

After recognizing the potential of a niche business in the city, the next decision becomes how to determine the best location for one. There is a glut of ways to go about choosing one, but with the power of data and machine learning, it is possible to gain unique insight into Austin that will help guide an entrepreneur's decision.

1.3 Interest

The conclusion of this analysis would be very useful to a solo entrepreneur or to a small group of partners with access to only a basic amount of capital who would like to be virtually assured of their success and reap high profit margins.

2. Data requirements, collection, and cleaning

2.1 Data requirements

The bulk of the data required here is location data. We will need to decide what criterion to use in segmenting Austin and find corresponding data. The main options are neighborhood and zip code. We will also need detailed data about venues in Austin such as their category and type. Fortunately, FourSquare and Geopy are robust and reliable and can provide all the required data.

2.2 Data collection

I settled on zip code as the segmenting criterion through trial and error. There weren't that many neighborhood values in any sources I found. Having too few neighborhoods for such a large city would not leave enough room for any meaningful insight to emerge.

I needed a source for the zip codes in Austin that was as robust, detailed, and reliable as FourSquare and Geopy. I did several Google searches and came across Opendatasoft, a repository of data from categories as diverse as currency, weather, ridesharing, public transit, etc. I did a search for zip codes in the United States and got thousands of entries. Their interface contains a convenient feature to select by state and so I chose Texas. Thankfully, the many columns of the table included latitude and longitude which I knew I would need later. I was prepared to pass the zip codes values through Geopy to retrieve corresponding coordinates but Opendatasoft eliminated that step. The site enables accessing tables in several formats so I got mine in the ‘csv’ format used readily in Python module ‘Pandas’

FourSquare’s API provided me with a list of all hookah bars in Austin including their location zip codes. It also provided me with venue data of businesses in all zip codes of Austin (up to a chosen threshold). These were done through and API request.

Geopy provided me with coordinates of some points that were not contained in the Opendatasoft table. I needed these points for finding important city points for analysis. I also used Geopy for calculating distances from these key points as were needed to make my conclusions more informed. This data was acquired using Geopy and its sub-modules in functions in the project’s Jupyter notebook.

2.3 Feature selection and data cleaning

I had been taught earlier that the k-means clustering algorithm was one of the neatest for performing the segment-cluster analysis I would be performing on Austin, so I selected my features accordingly.

First, I dropped all redundant values and labels. From the Opensoftdata table, I first dropped all rows where the city name was not Austin (I had all 2,743 zip codes for the state of Texas). I then dropped all columns except zip code, latitude, and longitude. Some of those other columns were city, state (both redundant), and time zone(unnecessary). This left me with 83 rows and 3 columns with the clean zip codes and coordinates of Austin.

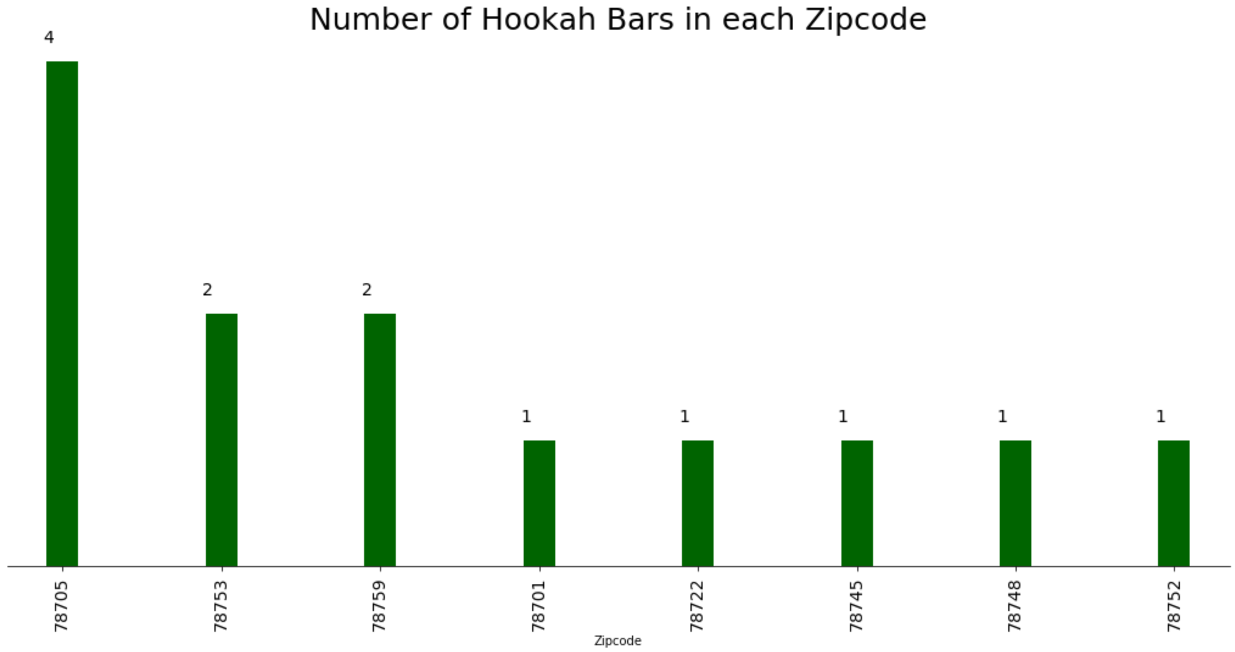
The first FourSquare API request I made was to get a list of hookah bars in Austin. I used the coordinates of the center of Austin as my location value and set a ridiculously wide radius of capture because I knew I would have to drop some rows and columns anyway and this abbreviated my search process. I captured hookah bars as far away as San Antonio but simply dropped all rows where the city wasn’t Austin. I then dropped all columns except name and zip code. Here I found some rows with missing zip code values. I parsed the raw json data and saw that they were indeed missing and that it wasn’t just some error in capturing. I manually searched for the corresponding names on both Yelp and Google Maps and realized the problem: these were locations that had been shut down but remained listed on FourSquare. They may have just been left in there for thoroughness, but I removed those rows and was left with 13 locations for hookah bars in Austin.

The second FourSquare API call was to get businesses in each of the 83 zip codes, up to a limit of 100. I left this data in its raw json format and it needed no cleaning. It would simply be called via function and appended into columns later.

3. Exploratory Data Analysis

Using k-means clustering, I decided that my criterion for choosing the candidate zip codes to build a new bar in would be that they ended up in the same cluster as a certain chosen ideal zip code, and hence had the same underlying characteristics.

I easily chose the ideal zip code by using a simply grouping function in the Jupyter notebook to count and sort the 13 locations from earlier by zip code. At first, I was skeptical that the locations would be too spread out and no insight for a successful location would emerge. Surprisingly, 4 of those bars turned out to located in the same zip code, a welcome observation for a city with 83 zip codes and only 13 bars. This one zip code certainly had whatever factor was needed to have a successful hookah bar.



Were it not for the laws of competition, my analysis could have ended there and I could just advise an entrepreneur to open a location in that zip code and call it a day. Sadly, there is such a thing as saturation, whereby a location is filled with the most of a certain type of business that can operate profitably in it. One small region having 4 hookah bars in a large city with only 13 is to be considered saturated.

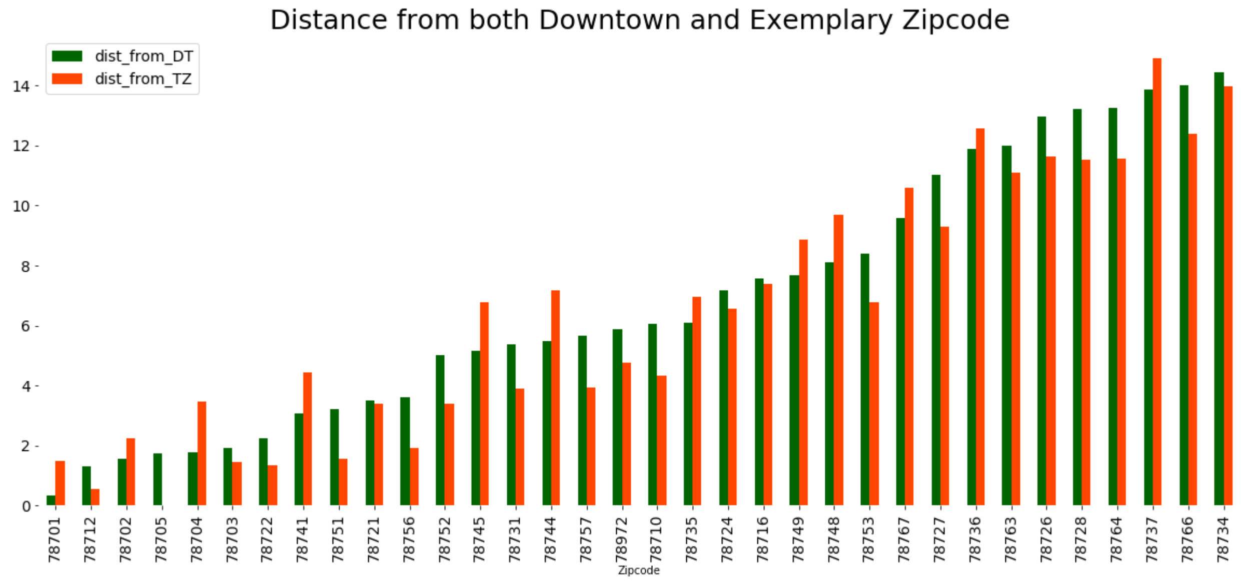
Having located the ideal zip code (78705), I then proceeded to test out how many clusters I would need in the k-means analysis. Too many would give me granular clusters and be worthless for analysis; too few would lead to same problem as the neighborhood segmentation earlier. Through trial and error, I settled on 8 clusters as a good number.

4. Methodology/Modeling

The 15 most popular venues in each of 83 zip codes were appended to the zip code values to create a large table with 83 rows and 16 columns. The k-means clustering algorithm is a classifier that assigns cluster labels based on the user's chosen number, 8 in this case. The 15 most popular venues here are used as the identifying characteristic and the zip codes were split into 8 clusters based on the similarity of which business categories were abundant in them. The underlying assumption is that businesses tend to self-group into eco systems such that their largest customer demographic does not need to travel far to get to other venues that they tend to also visit a lot. An example would be truck stops and hotels tending to be very close to each other in almost any city. The algorithm runs for a short time and fills each of the 8 clusters with zip codes that bear the same eco-system signature.

One important feature that was left out of the characteristics in the clustering algorithm was proximity to city center. I left it out because I wanted the clusters to represent the eco-system signature as purely as possible, with no other factor polluting this signature. However, it is very important for a niche business hoping to earn lots of profit to be close to the city center, which is the center of business and entertainment and is very often the default/naïve destination for tourists and travelers seeking to rapidly and conveniently experience a city. The cluster containing our ideal zip code had 34 zip codes (nearly half the regions) and would need to be narrowed down further. Naturally, proximity to the city center was a latent characteristic waiting to be explored.

I performed the proximity analysis using a function to pass the coordinates of these 34 zip codes through a Geopy sub-module and appended the corresponding distance values to a fresh table. A second distance calculation we made as we needed to be both close to the city center, but far from the saturated ideal zip code. I withered the table of 34 zip codes down by eliminating those that were more than 3 miles away from the city center, but less than 1.5 miles from the saturated area.



5. Results

I ended up with 2 candidate zip codes in which I am highly confident that a new hookah bar in Austin will be successful. My confidence is based on the purity of the data sources (Foursquare, Geopy, and Opendatasoft), and the strength of the k-means clustering algorithm in extracting hidden signatures using characteristics.

The resulting two regions are zip codes: 78702 and 78704.

6. Discussion

Austin is a city of around 1 million people and growing rapidly because of many favorable factors. In particular, this is a great time to invest in the city before every profitable business sector becomes saturated.

Although I am highly confident in the analysis, there were quite a few assumptions and generalizations which had to be made in order to reach the recommendation. A more detailed analysis would involve such information as zoning laws, commercial real estate prices/rents, and most importantly, local ground knowledge of the city of Austin itself. No analysis made from afar, no matter how rigorous, can replace the actual experience of one who has spent a long time around the business and real estate scene of Austin. However, the experienced person armed with these data insights would have an advantage over one without them.

8. Conclusion

The aim of this project is to use available data and insights to help an entrepreneur select a region of Austin, Texas to establish a new hookah bar. We used sourced data from various sources and used several Python modules along with machine learning to extract hidden insights from the data.

The 2 candidate areas extracted from the data are the best places to explore locations for a new hookah bar. They are close to the center of the city central/business district. They are outside the single region which has a saturated amount of hookah bars, but are very similar to it in terms of what businesses tend to thrive in them. As such they are a mix of great opportunity and reduced competition. I have a high level of confidence that an entrepreneur who established a new venue in one these areas and followed other best practices of business would be very successful.