Coursera Capstone

Finding the highest value location for a pet store, within Austin, TX

Aria Raney
April 18th, 2021

1. Introduction

1.1 Background

Austin Texas is known as one of the most pet friendly cities in the United States. While definitive numbers on the number of pets in American households are hard to come by, numerous studies have found that over 50% of households in Austin Texas include pets, making it the third most pet-dense metropolitan area in the country. Given that, it's no surprise that Austin has a thriving market for pet related services. But despite the plethora of options, there are still underserved markets in certain population dense areas that could make a great home for your business.

1.2 Problem

Given Austin's mix of high demand and high competition, determining the ideal location for a new business directed towards pet owners is non trivial. There are few locations within the city with truly no competition, and the uneven distribution of demand means that simply choosing the most isolated location won't guarantee favorable results. Likewise neither will choosing the most population dense location, as these regions tend to already be saturated with existing venues. This project aims to balance all these considerations to determine the computationally ideal location.

1.3 Interest

The city of Austin is rapidly growing, one of the fastest growing metropolises in the country. This fact along with Austin's high density of pet ownership leads to an ever growing market, providing a great many opportunities for either chains or entrepreneurs to open profitable new venues within Austin. This project could provide meaningful insights to any such shareholders looking to explore this opportunity.

2. Data Acquisition and Cleaning

2.1 Data Sources

The ZIP Codes considered Austin Addresses and the population (if any) of residents living within them can be extracted from the ZIP Codes for City of Austin, TX on zip-codes.com

> This dataset contains a list of every zip code considered an Austin address, along with some additional information about each, such as the county it's in and the area codes associated with them. The most important information for our purposes is the population of residents with addresses in this zip code.
>
> This dataset is used to identify the various residential areas within Austin, by filtering out each zip code with a population of 0, for a total of 44 residential Zip codes. The listed population and latitude and longitude coordinates of these zip codes (collected using the Geopy API) will be used for discussing the population distribution of Austin and how it relates to our issue.
>
> For example: the Zip code 78745 has the largest population with 55,614 people living within it, while 78755 contains none at all as it represents a PO box, and can be excluded from our population table.

A list of pet related businesses and their locations can be requested using the Foursquare API, using the Search endpoint and passing the category keys related to our market of interest.

> We use foursquare to create a Database of every venue inside Austin TX that caters to pet owners, such as Pet Stores, Groomers, and Veterinarians. With this database we show that the distribution of pet services shows distinct gaps in areas of high demand.
>
> For example: Austin has over 50 venues that cater specifically to pet owners, but only 2 inside the 78754 ZIP Code (of course, these Venues can serve residents in adjacent ZIP Codes as well. The details of the distribution are discussed in detail during the full analysis).

2.2 Data Cleaning

Many ZIP Codes within the city of Austin did not contain residential addresses, either due to being addresses for P.O. Boxes or for being purely commercial regions. The data web scraped from the ZIP Codes table thus had to be filtered for those with 0 population, in addition to the formatting required to extract the meaningful information from the HTML. I used the BeautifulSoup library to streamline this process.

The Foursquare API call returns a json file with some inconvenient formatting. Simple functions were defined to extract information like the primary category, unique name, and renaming the database columns to match those in the ZIP Code dataframe for future merging.

Finally, a cdist matrix was constructed using the Scipy library, which calculated the distance between each Foursquare Venue and each residential ZIP Code. Since these locations were listed in geographic degrees (Latitude and Longitude values), I used Geopy's geodesic-distance to calculate the true distance in miles.

3. Methodology

3.1 Exploratory Data Analysis

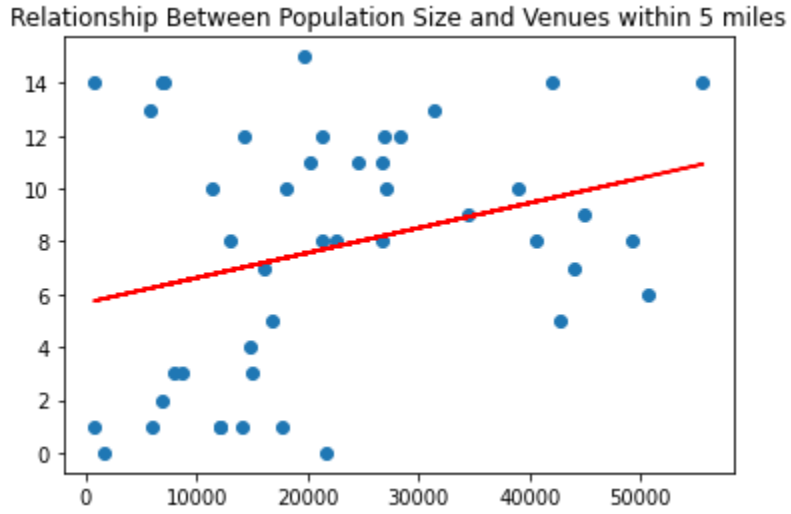
In preparation for calculating the ideal location, we consider how to determine the best existing location.

First, we must consider the maximum distance a consumer is likely to travel. I chose 5 miles, as it is approximately equal to the largest minimum distance from any ZIP Code to the closest existing Venue (excluding outliers).

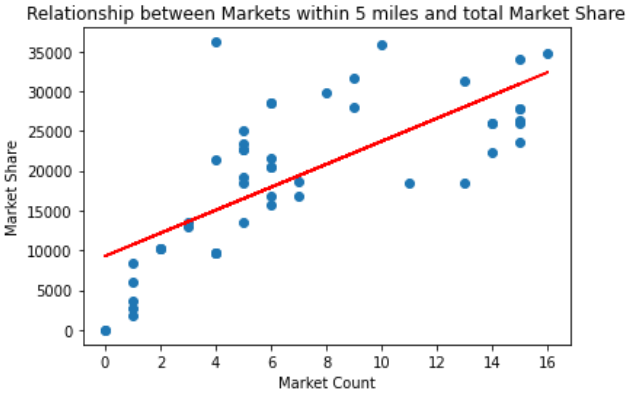
With that, we can use our distance dataframe to create a count of which (and how many) venues fall within the chosen radius of each ZIP Code, and vice versa. With the former we calculate a [Population per Venue] column in the ZIP Code dataframe, and combine it with the latter to create a [Market Share] column in our Venues dataframe.

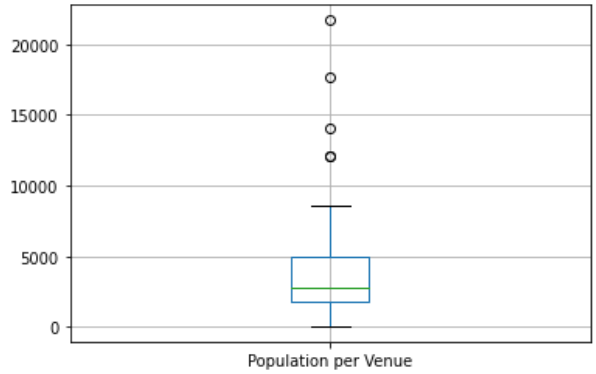
This [Market Share] is the sum of the element-wise multiplication of each ZIP Code's [Population per Venue] with the binary array indicating the ZIP Codes within the market radius. Put plainly, each Venue has a market share equal to its portion of each ZIP Code within 5 miles. Venues within commuting distance of the highest possible population with the least possible competition will have the most success.

3.2 Observed Relationships



I observed a weak linear relationship between ZIP Code population size and number of markets within 5 miles, based on a simple linear regression model. However while ZIP Codes with large resident populations did tend not to have very small numbers of venues within market range, there were still notable exceptions. These ZIP codes could prove promising locations to begin the search.





The relationship between the number of markets and the total share was stronger. Reasonably, given that we see baring outliers the total [Population per Venue] does not show a great deal of variance across ZIP Codes.

This suggests that the ideal location is likely to be in a location with a high number of potential markets within range.

3.3 Grid Search

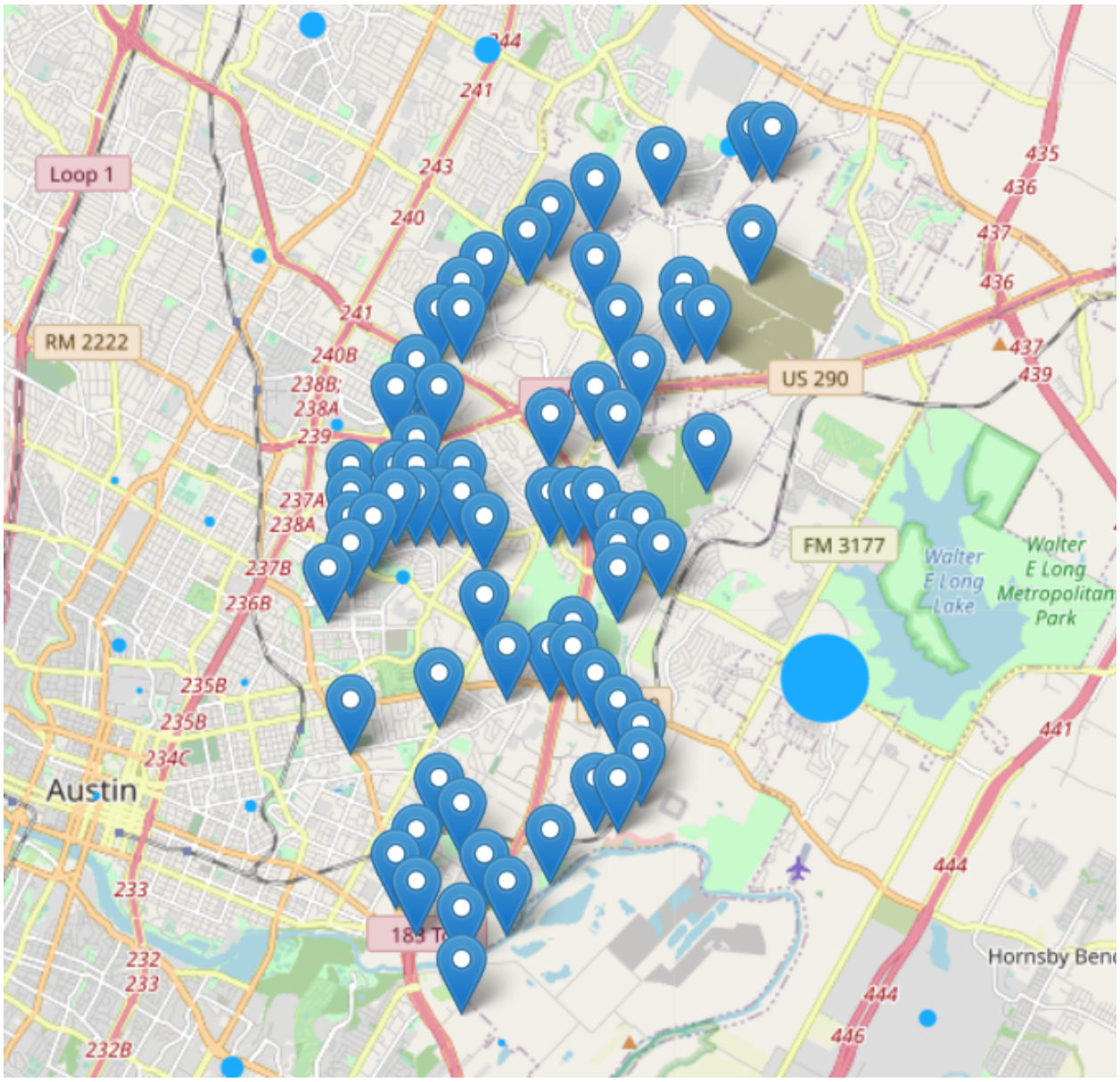
With my methods chosen, I proceeded with determining the proposed locations.

I began by creating a grid using linspace to computer 100 x 100 evenly spaced points across the city. Transforming this grid into a dataframe using the same methods used above, I computed the theoretical Market Share a venue in that location would have had (adding one to the hypothetical Venue Count in each, to account for the venue we'd be adding).

I then compared the resulting grid with the maximum score obtained from existing venues and found a total of 70 unique locations.

4. Results

All 70 locations which showed a higher expected value than existing venues were centered around the 78724 ZIP Code. In addition, these venues were clustered in an arc around the US290 and US183 intersection, a very busy section of the city with many commercial and residential neighborhoods around it.



5. Discussion

78724 was the largest remaining ZIP Code which had no existing venues within the designated radius. It represents a large population of potential consumers who currently

have to drive an uncomfortable distance to meet their pets needs, and based on these results I would suggest building nearer to it. Given the accessibility and prominence of the major intersection in the center of the cluster, it stands to reason that this region is a good place to begin looking for building a new venue.

6. Conclusion

These results represent only an analysis of the upper limit of potential Market Sizes based on population and competition, and many more aspects go in to determining an ideal market. These results however do paint a clear picture for a region of the city for shareholders to focus on for further research.