Getting to know Austin, TX

Introduction

Austin is a vibrant city attracting many newcomers with opportunities in various professional venues - from technology to music to education and beyond. Relocating to a new place is a complex decision as one has to consider many factors and finding objective information on many of them is a challenge.

My project attempts to address this issue by creating profiles of the different areas of the greater Austin area integrating social and crime data to help newcomers target their new home search to areas that would best fit their lifestyle.

The results of this effort can be of interest to constituencies beyond potential new homeowners. These include real estate agents, corporate relocation specialists, city planners and law enforcement agencies.

Data

Data from different sources was be combined to create clusters of ZIP codes with similar characteristics.

Specifically, a list of all ZIP codes, along with the corresponding city, was be scraped from the BestPlaces website. Social data (e.g., number and type of entertainment venue in each ZIP code) was pulled from FourSquare. Geocoder will be used to obtain the coordinates of each ZIP code. The crime data was gathered from the Austin Police Department (APD) crime statistics reports.

Methodology

Data preparation

After each data set was pulled it, it was examined for any data quality issues. For example, once the crime data was read from the PDF file, most columns had to be converted from object type (string values) into numeric. In addition, several columns contained commas to separate the thousands. Thus, those formatting characters were removed before changing the data type. The last row containing column totals was removed. Since the values in all columns except for the ZIP codes, represent numbers of crimes reported to the APD, all NA values were replaced with zeros. The detailed crime data was combined in two categories – violent and property crimes. The existing aggregate data (total indexed, nonindexed and combined) were retained unchanged.

To better understand the character of the different parts of the city, up to the 200 most popular venues within 1000 meters of the center of each ZIP code were pulled. The frequency distribution of the venue categories by ZIP code was calculated and the top 10 categories were used to create the ZIP code profiles. Descriptive statistics for the number of venues per ZIP code were obtained as well.

The ZIP codes were grouped using k-means clustering. The number of clusters to retain was determined by the Elbow method (see Figure 1 below). There were several possible values of k ranging from 4 to 13. 8 clusters were retained to maintain some ability to define the clusters. This machine learning technique allows for efficient data-driven approach to determining the number of profiles without any *a priori* knowledge or constraint.

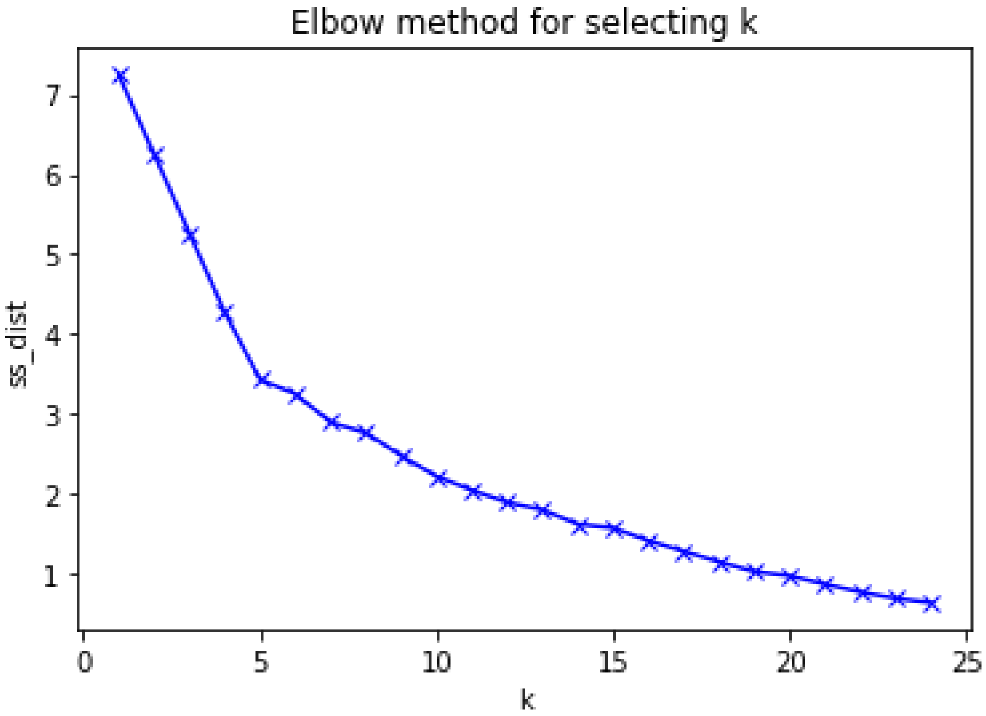


Figure 1. Model inertia for a range of number of clusters

Cluster summary statistics and ZIP code composition were analyzed as well.

Results

88 ZIP codes were obtained from the greater Austin area. Venue data was obtained for all of them. Details on 1,309 venues were obtained and further analyzed. Crime data for 48 of the ZIP codes were obtained from APD.

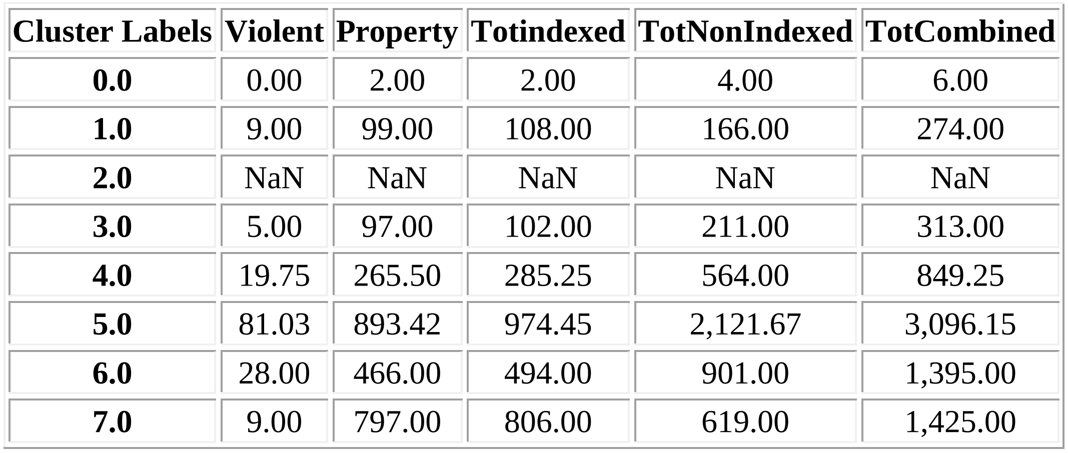
The number of ZIP codes per cluster are presented in Table 1 below:

Table 1. ZIP codes by cluster

Cluster Labels	***ZIPcode***
0.0	1
1.0	1
2.0	1
3.0	1
4.0	6
5.0	35
6.0	1
7.0	1

The average number of crimes per cluster is presented in Table 2 below. The table suggests that there are 3 types of cluster – high, medium and low crime rates. It is interesting to note that the grouping works very well for property and total combined crimes but not so well on the lower end of the Violent crimes category.

Table 2. Average number of crimes per cluster



Cluster Labels	Violent	Property	Totindexed	TotNonIndexed	TotCombined
0.0	0.00	2.00	2.00	4.00	6.00
1.0	9.00	99.00	108.00	166.00	274.00
2.0	NaN	NaN	NaN	NaN	NaN
3.0	5.00	97.00	102.00	211.00	313.00
4.0	19.75	265.50	285.25	564.00	849.25
5.0	81.03	893.42	974.45	2,121.67	3,096.15
6.0	28.00	466.00	494.00	901.00	1,395.00
7.0	9.00	797.00	806.00	619.00	1,425.00

Discussion

Looking at the number of ZIP codes per cluster (and the corresponding Folium map from the Jupyter notebook), I have observed two major areas – cluster labels 4 and 5. They contain 6 and 35 ZIP codes respectively. The latter cluster can be described as the older, more established part of the city. The former is located primarily in the NorthWest corner of the greater Austin area. That part of town has been, and still is, under major development. Surprisingly, it is also the less crime prone of the two… and that is by a wide margin across all crime type categories.

Since the crime data comes only from APD (which does not have jurisdiction over all studied ZIP codes) and some crimes are reported directly to other law enforcement agencies, the numbers presented in Table 2 may not present the full picture.

Conclusion

In conclusion, this research is an initial exploration into the type of areas around a bustling city based on the preferred activities of the local residents. The detailed cluster profiles provide even finer-grained view of the metro area making them even more valuable.

The study presents only one way of disaggregating the city. For some newcomers, the crime activity may be a more important decision criterion than the social activities. In this case, clustering based on crime rates may yield more meaningful results.