Comparing Kansas City and St. Louis Metros

Matthew Perkins

Contents

1. Introduction

1.1 Background

The Kansas City metropolitan area and St. Louis metropolitan area are two of the largest metro areas in the Midwest. Both are home to many well-established businesses and some businesses seeking to expand. With the cost of real estate being cheaper compared to the east and west coast of the United States, businesses are looking to open locations in one of the metro areas or the other or perhaps both. It is important that businesses know the similarities and differences in the cities within each metro area and between each metro area. This will give businesses an idea of which cities to build their locations in based on a city's characteristics. A company's locations are often a consideration for many employees wanting to work there as well. Additionally, for example, if a company had locations in Kansas City and wanted to open another location in St. Louis, the findings could help find a city of similar or better quality compared to their current location in Kansas City. The findings can benefit startups. Startups can decide where to open their business based on the other types of businesses in each group of cities. Finally, this project can serve the interest of individuals or families looking to move from one metro area to another to find a city like their current city.

1.2 Objective

The problem of this project focuses on finding groups of cities in each metro area that share similar characteristics. These characteristics are the cost of the venues in the city, quality of venues in the city, and types of venues in the city. Overall, those three factors will help assess the quality of life in each city. Considering the constant comparisons made between these two metro areas, the goal is to use data from various sources to determine how similar the cities in each metro area are to one another.

2. Data

2.1 Data sources

The data for this project comes from a multitude of sources. The Cities in each metro area will come from their respective Wikipedia pages (Kansas City's page, St. Louis' page). Coordinates for each city were obtained using the MapQuest Geocoding API. Using those coordinates, the top venues will around each city will be obtained through the Foursquare Places API. Finally, the Yelp API will be used to gather information about price, ratings, and number of people who rated each venue.

2.2 Description of data

The data used for analysis will include venue category, price of the venue, rating of the venue, and number of people who rated the venue. Venue categories are categorical variables and would need to be given dummy values in order to perform a clustering algorithm on the data. Price and number of people who rate the venue will be integers, and rating is a float data type since it can have a decimal place. Prices will range from 1 (least pricey) to 4 (most pricey). Ratings will range from 1 to 5 (with 5 being excellent).

2.3 Purpose of the data

This data will be used for clustering (or grouping) the cities in each metro area. Additionally, it will help assess the essential characteristics of a city's venues. Price is used to measure cost of venues, ratings help asses the quality of the venues, and venue category gives the type of venues available in a city. These venue properties will reflect the expense of living in a city in terms of daily activities in the city, the overall quality of the city, and the diversity of choice for activities in the city. The lower the cost, the higher the quality, and the more unique venue options available are, generally, desirable qualities of a city people want to live in.

2.4 Collecting Data

When the Wikipedia pages were scraped for the cities, a table with columns of city and state was formed for each metro area. Kansas City had 79 cities while St. Louis had 250. Once the coordinates for the cities in each metro area were obtained from MapQuest, these coordinates were merged with their respective data frames. Then the coordinates for each city were passed through the Foursquare API using the explore endpoint to get the popular venues. For Kansas City there were 1394 venues obtained compared to 2374 venues in St. Louis. Finally, all these venues, along with their city's coordinates, were passed into the Yelp API using the search feature to get their price, rating, and number of people who rated the venue.

2.5 Cleaning data

Due to using two different APIs to collect data on venues, some values for price and ratings ended up being not available (NA). For venues with NAs for ratings, they were removed from the data. They were removed, partially, because they made up a small set within each metro areas data. Also, there is no good way to replace these NAs for ratings since ratings are, generally, independent of area and venue category. 178 venues and 69 venues were removed from St. Louis and Kansas City data frames, respectively. Once these venues were removed the NAs for price needed to be fixed. For these NAs, the most frequent price in each city was found. For the venues the most frequent price in the venue's respective city was used to replace the NA. This was done since the price of venues reflects the cost of other venues around them along with price of living. The most frequent price was used based on the probability of venues at each price. Since there were

only a few venues in each city that didn't have a price they are more likely to have the most frequent price appearing in the city.

3. Exploring and analyzing data

3.1 Visualizing data

For each metro area boxplots were created for the prices, ratings, and number of ratings for each venue. (See Figure 1 for Kansas City's boxplots and Figure 2 for St. Louis' boxplots).

Figure 1-Kansas City boxplots

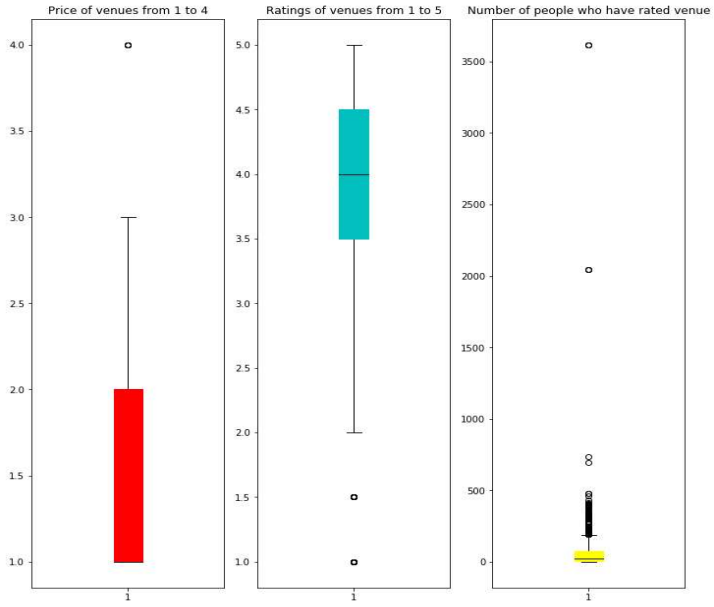
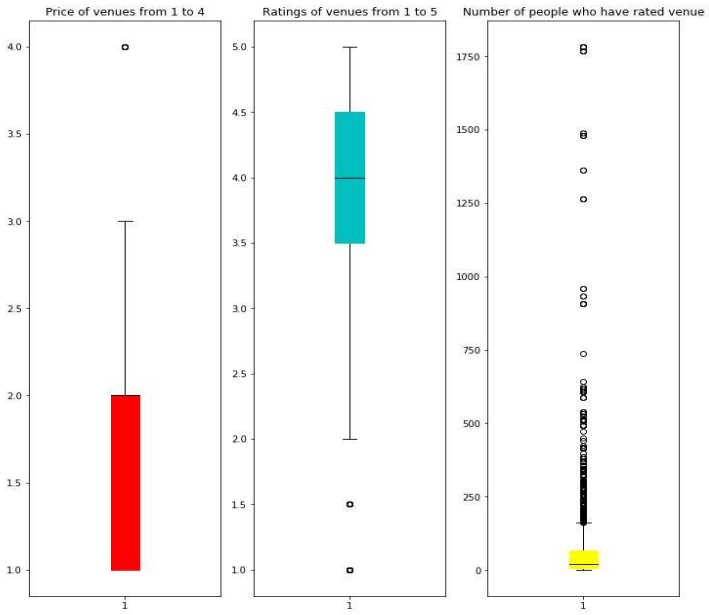
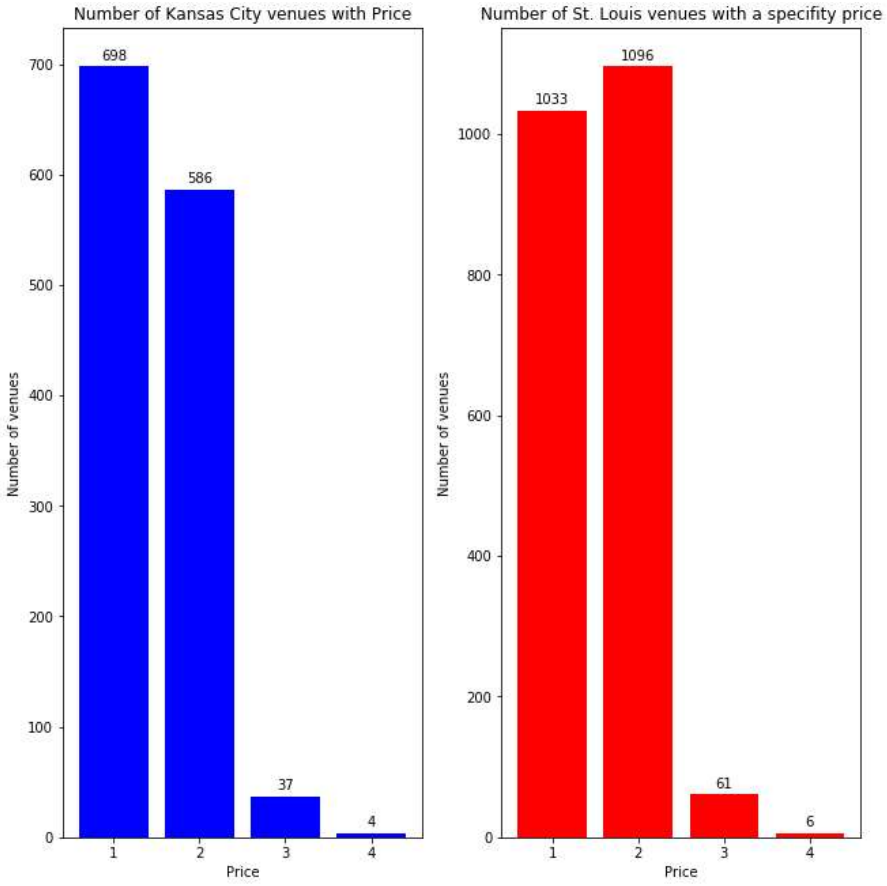


Figure 2-St. Louis boxplots



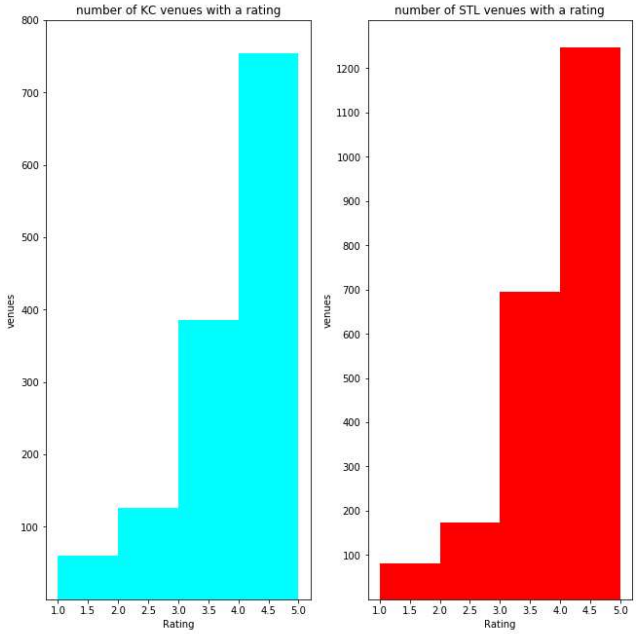
St. Louis and Kansas City have similar distributions for price and ratings. For both venues with price of 4 are outliers and the median is a price at 1. For ratings venues with 1 and 1.5 out of 5 are outliers with a median of around 4 for the ratings. The Price of venues is skewed right in both areas while it ratings are skewed to the left. The boxplots for number of ratings are noticeably different between the two cities. St. Louis has more outliers that have a higher number of ratings. While price has numbers associated with it, the variable is, actually, a categorical variable. Ratings is somewhat like a categorical variable due to yelp only giving ratings by .5 increments. However, ratings will be treated a numeric data type as stated previously. After the boxplots were created, bar graphs were made for price variable in both metro areas to see how many venues fall into each price category. (See figure 3)

Figure 3-Bar graphs for price



As one can see St. Louis has more venues with price 2 than price 1 unlike in Kansas City. It would thus appear that the St. Louis metro area is more expensive to live in with higher prices for venues. Finally, a histogram for the ratings of the venues in each metro area was created. (See Figure 4)

Figure 4-Histogram of ratings

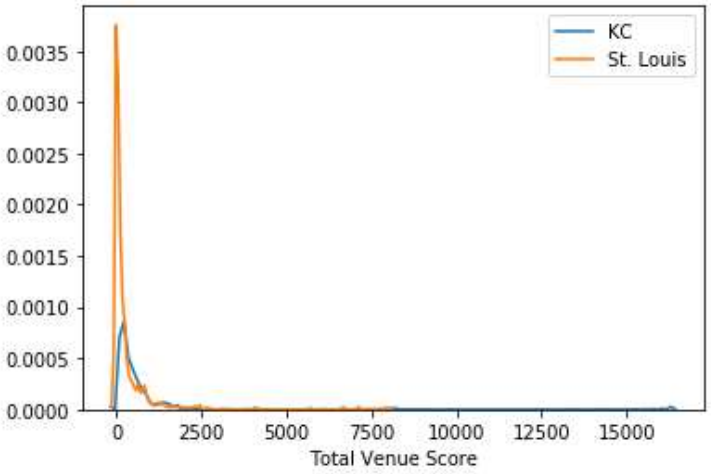


From the histograms one notices there is a larger amount of venues in St. Louis with ratings between 4 to 5, but the proportion of venues in each rating are about the same. The bins include the left endpoint but exclude the right endpoint (except for the bin between 4 and 5 where it includes both endpoints) when determining where to place the venue. Therefore, one could suppose that the quality of venues between the two metro areas is about the same.

3.2 Modifying variables

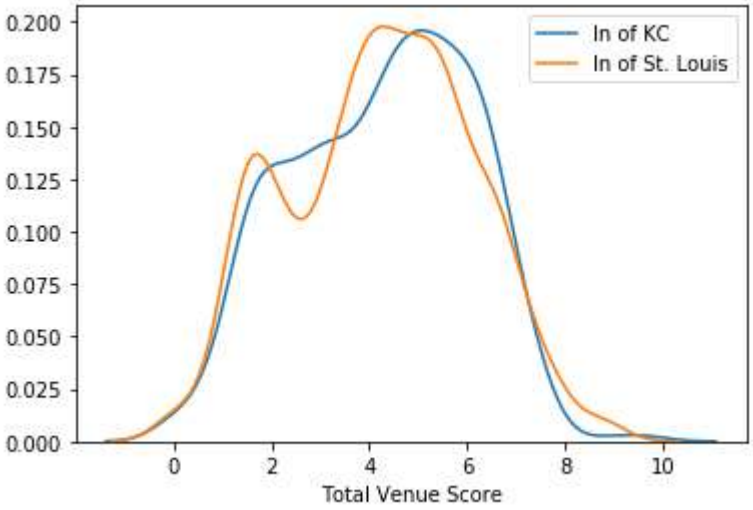
Since the ratings for any venue can be easily skewed if only a few people rate a venue and have a strong opinion one way or the other, a new variable was created from the rating and the number of people who rated the venue. The rating is, approximately, the average of everyone's rating for that venue. Thus, by multiplying the rating by the number of people who rated will give the total score for that venue. This total score can be treated as a more continuous variable and, thus, transformed to remove some of the skewness from the ratings. Note the skewness of the total score in figure 5.

Figure 5-Distribution of total score



To fix the right skewed nature of total venue score, a natural log transformation was performed. The new distribution can be seen in figure 6.

Figure 6-natural log transform distribution



As one can see, the distributions are much more normal with only a slight left skew. The distributions do look somewhat bimodal though. Nonetheless, the transformation gives a significant improvement to the data.

Once this transformation was made the two data frames were combined into one for analysis. Before the data could be put through a machine learning algorithm, the venue categories had to be given dummy values so that any unsupervised learning algorithm can be used. The data was then grouped by city (this made gave 324 rows) where all other columns were averaged. Price and the natural log of total score were both much higher in value then the typical average for each category. Thus, both variables were rescaled by

normalization $[(x-x_{min})/(x_{max}-x_{min})]$ to make both between 0 and 1. Rescaling these variables makes sure the algorithm isn't biased towards those two variables due to their much larger numbers.

4. Grouping the cities

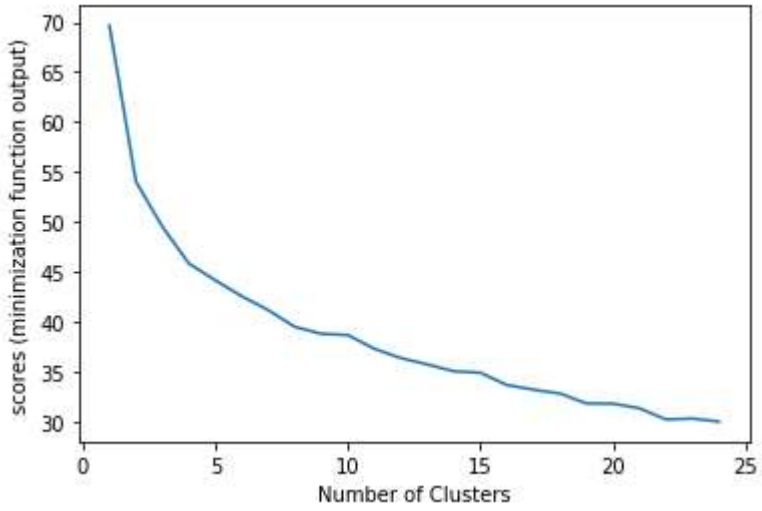
4.1 Algorithm

To group the cities into clusters the KMeans clustering algorithm was used through the scikit-learn package. KMeans clustering algorithm was used for a couple of reasons. First, this algorithm can separate cities into unequal clusters and works well with the size of the dataset. Second, the algorithm is simple to implement which means it won't take as long to compute the clusters.

4.2 Results

The number of clusters chosen was 10 since this seemed to be when the algorithm scores first plateaued and slowly decreased (see Figure 7).

Figure 7-KMeans scores for a particular number of clusters



Clusters 0 through 9 (going with how the program labeled them) had 33, 40, 12, 38, 54, 19, 68, 5, 1, and 54 cities, respectively.

4.3 Observations

Cluster 8 captured the one outlier city in the dataset. This city got only one popular venue which happened to have a high price and lower quality. This city is most likely a small town. The clustering algorithm seemed to have created groups of cities based on their distance from the main downtown city. In general, these clusters can be divided into three groups themselves. One group is for the rural/outskirts portion of cities. The other group includes the suburban cities. The last group involves the cities in or near the downtown portion of the metro areas. Clusters 0, 1,7,8,9 seem to belong to this outskirts group. Clusters 2 and 5 seem to belong to the downtown group. Clusters 3,4, and 6 belong to the

suburbs group. The main difference within each of these groups is price and total venue score (or quality) rather than the venue categories. For example, cluster 4 has low price but high-quality venues, cluster 3 has low-price low-quality venues, and cluster 6 has high-price high-quality venues. The venue category mattered the most for the first set of clusters involving the cities on the outskirts. 160 cities belonged in the suburbs, 133 to the outskirts, and 19 were a part of the downtown group.

5. Potential issues and recommendations

Within this project there are a few parts where caution is needed. The first is in dealing with the NAs. Perhaps, it might be best to keep the NAs and not impute them. Or it might be best to replace price NAs with the mean price of the city the venue is in to keep the same distribution. The amount of price NAs in each city might be important since some cities had 3 or more venues with NAs for price. Additionally, removing venues with NAs in the rating variable could've biased the data. It might be that replacing the NAs with the average or median rating of venues in the same city or of ratings in the same category would give more accurate clusters. Moreover, different ways of combining price, rating, and number of people who rated a venue should be considered. Also, normalizing the data could've impacted the clusters, and it should be investigated as to how much impact normalizing the data had on the clusters. Finally, other clustering algorithms should be used to see which appears to give the more reasonable clusters. Other clustering algorithms weren't explored in this project but, in the future, they should be considered.

6. Conclusion

Overall, Kansas City and St. Louis metropolitan areas are similar to each other. The only noticeable difference is that the city sizes are smaller in St. Louis, and venues are slightly pricier in St. Louis too. Therefore, it was quite easy to see which cities in each metro area belonged in the same group. It turns out that price and quality of venues changes as one moves further away from the main cities in each metro area.