Moving to a Town that’s Right for Me

Determining the Similarity between Cambridge and Three Cities in Texas

Amanda Aparicio

May 2020

Introduction.

Hi, my name is Amanda.

I was born on a small farm in the middle-of-nowhere, Texas. We had ducks, chickens, a miniature donkey, goats, even a pig at some point. Imagine my cultural shock when I was accepted into MIT for undergrad and had to make the move from my small town in Texas to Cambridge, MA. Initially, it was challenging adjusting to the lifestyle in Cambridge, but I eventually came to love the area and even spent 9 years out of my young adult life in Cambridge.

Recently, I decided to make the move back to Texas to be closer to family and to open up housing opportunities since Cambridge is quite expensive. While I am excited to be back in Texas, I want to find an area that has similar features to Cambridge.

Enter this project: my goal is to analyze the venue data for 3 major cities in Texas (specifically, Houston, Austin, and San Antonio) in order to compare them to Cambridge, MA. Selfishly, this analysis will mainly be of interest to me but might also serve others who are in a similar situation to myself decide which city is the most similar to Cambridge, MA.

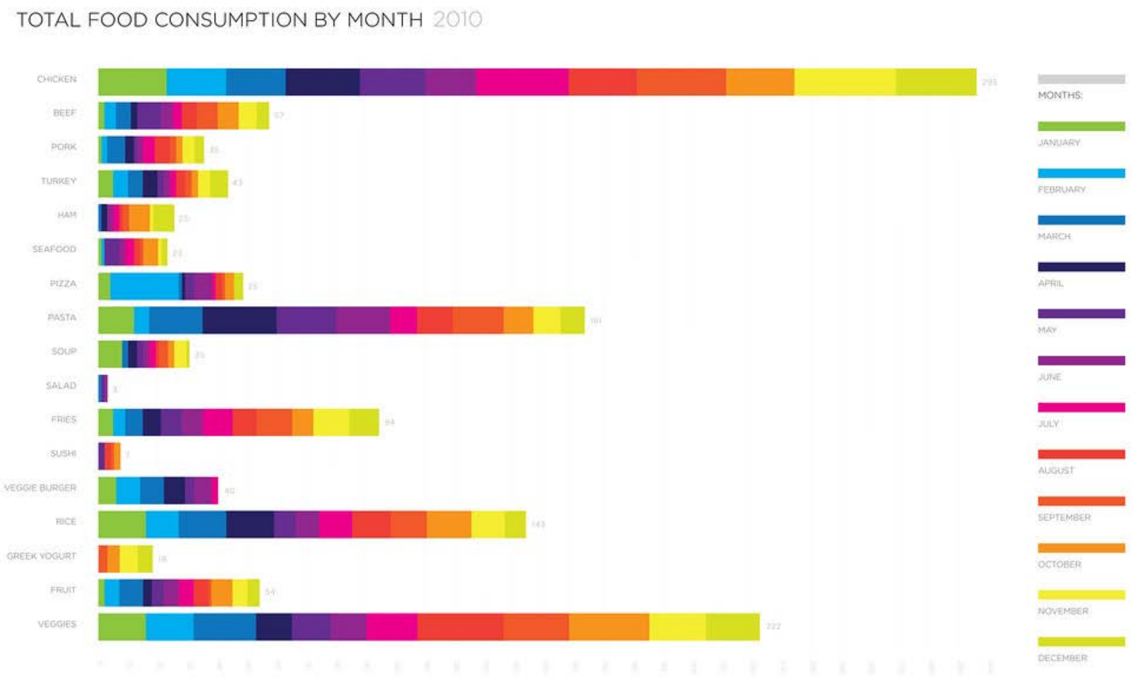
Data.

1. Data Plan

In order to reach the goal describe above, we need to develop a method for comparing two cities and then gather the appropriate data to make the comparison. While there are a plethora of factors we could include in this analysis, we will limit the scope of this project to comparing the venue composition of neighborhoods of each city. The steps of our analysis include:

1. Segment each city into its respective neighborhoods (Data section)
2. Gather venue data for each segment (Data section)
3. Perform *k*-means clustering on all neighborhoods of all cities (Methodology section)
4. Visualize the created clusters (Methodology section)
5. Use the visualization to determine similarity between the cities (Results section)

For our last step in our analysis, we want to be able to look at a horizontal, bar chart for each of the clusters created through *k*-means, stacked by city for each bar. For example, we want to create a visual like shown below, but replace the food names with cluster labels and the month names with our city names.



Source: https://www.flickr.com/photos/laurenmanning/5659583878

By looking at the composition of each cluster, we can better understand how various parts of our 4 different cities are similar and different from each other which will inform our decision about where to relocate to.

Therefore, the corresponding data we will need is:

- neighborhood information for each city (i.e. how does each city organize their neighborhoods, how many neighborhoods are there, and what are their names)
- latitude and longitude data for all the neighborhoods for the cities of Cambridge, Austin, Houston, and San Antonio
- venue information (location and type) for each neighborhood

Below, we will describe how we collected our data for each of the above bulleted points.

2. Data Source of Neighborhood Information

Each of our four cities organizes their neighborhoods in very different ways. For example, Houston not only has neighborhoods but also *super-neighborhoods* which are larger areas comprised of similar neighborhoods. On the other hand, Cambridge has thirteen, neatly-defined neighborhoods. As a result, gathering the neighborhood information for each city took some detective work.

For Houston, we treated the *super-neighborhoods* as regular neighborhoods to make our analysis more manageable (Houston is a very large city with 88 *super-neighborhoods* and thus even more regular neighborhoods!). The data on Houston's neighborhoods was sourced using the Wikipedia article called "List of Houston neighborhoods".

Luckily, Wikipedia also has an article on the neighborhoods of Austin, so we used the neighborhood information listed on the "List of Austin neighborhoods" Wikipedia page.

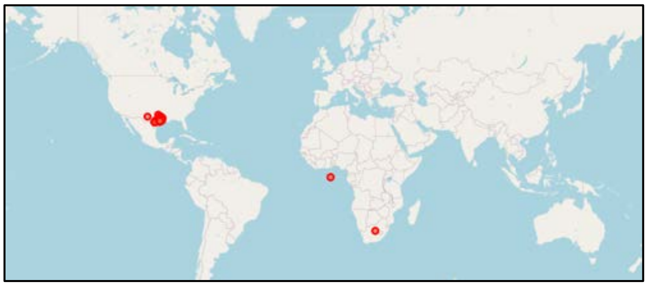
Cambridge's website has a map which lists all of its neighborhoods; therefore, the source of our Cambridge neighborhood information came from the Neighborhood Map of Cambridge.

Unfortunately, the city of San Antonio did not have information readily available about a definitive list of its neighborhoods. Consequently, we used a Wikipedia article called "Neighborhoods and districts of San Antonio" to create a list of neighborhoods.

3. Data Source of Neighborhood Latitude and Longitude Coordinates

Once we had the names of our neighborhoods, we collected data on the latitude and longitude of the centers of our neighborhoods using a combination of python's *geopy* package and Google's search feature. After collecting the latitude and longitude data, we verified the positions of our neighborhoods by plotting using python's *folium* package.

The *geopy* package struggled in some cases to return accurate latitude and longitude data. For example, after feeding the neighborhood names from Houston into *geopy* and plotting, our map of Houston's neighborhoods looked like the following:



Consequently, for Houston, San Antonio, a handful of Austin neighborhoods, and two Cambridge neighborhoods, we used Google's search feature to source the GPS data of the neighborhoods.

4. Data Source of Venue Information

For the composition of venues of each neighborhood, we utilized the FourSquare API. We fed the API latitude and longitude coordinates of each of our neighborhoods and requested data on 100 venues in a radius of 1600 meters (about a mile) from the given GPS coordinates. Notice that this means we made two simplifying assumptions:

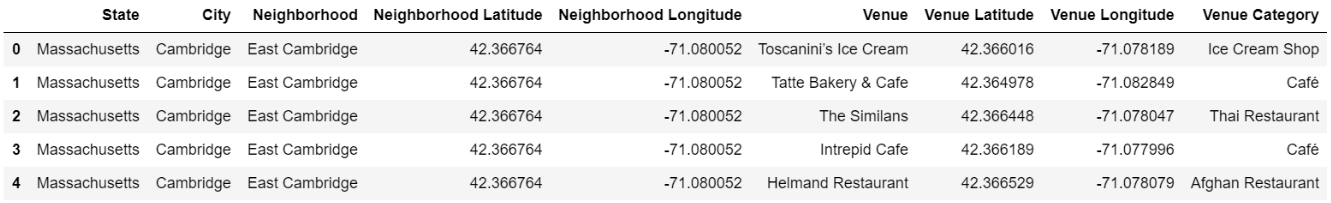
1. 100 venues is sufficient to characterize our neighborhoods
2. Each neighborhood is shaped like a circle with a radius of 1600 meters

The first assumption seems reasonable – 100 venues is more than sufficient to characterize a neighborhood.

The second assumption, however, is a little more problematic. Just glancing at the maps of our 4 cities reveals that the distances between each neighborhood varies greatly, with some neighborhoods being less than 1600 meters away from each other and others being way more than 1600 meters away from each other. Consequently, in our treatment of neighborhoods as perfect circles of radius 1600 meters, some of our neighborhoods are going to overlap while other neighborhoods will be missing some venue data altogether. However, we argue that we can still gain some insight into the comparisons between neighborhoods even with this simplifying measure put into place and will continue with our analysis regardless.

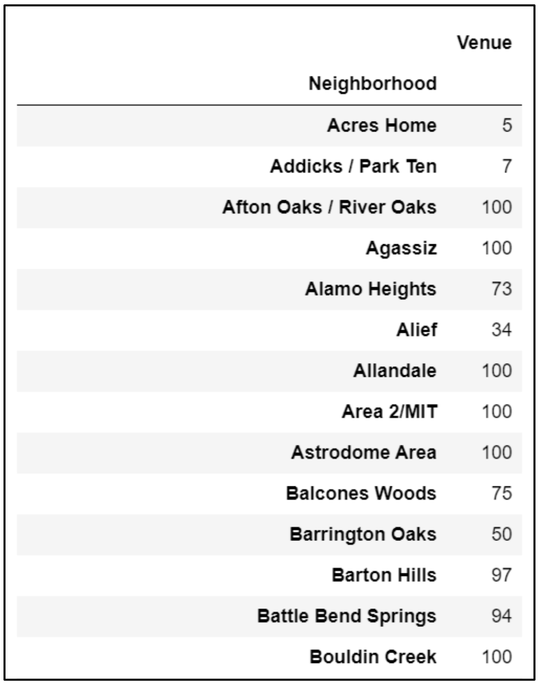
Methodology.

Please note: all screenshots of dataframes are only a portion of a much larger dataframe. Once we finished collecting our data, our dataframe looked as shown below:



	State	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Massachusetts	Cambridge	East Cambridge	42.366764	-71.080052	Toscanini's Ice Cream	42.366016	-71.078189	Ice Cream Shop
1	Massachusetts	Cambridge	East Cambridge	42.366764	-71.080052	Tatte Bakery & Cafe	42.364978	-71.082849	Café
2	Massachusetts	Cambridge	East Cambridge	42.366764	-71.080052	The Similans	42.366448	-71.078047	Thai Restaurant
3	Massachusetts	Cambridge	East Cambridge	42.366764	-71.080052	Intrepid Cafe	42.366189	-71.077996	Café
4	Massachusetts	Cambridge	East Cambridge	42.366764	-71.080052	Helmand Restaurant	42.366529	-71.078079	Afghan Restaurant

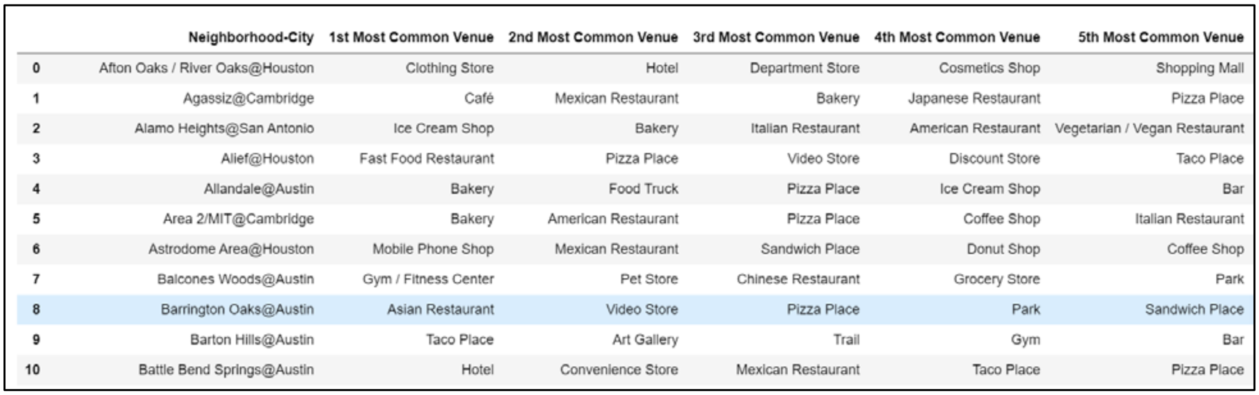
First, we added all venues together for each neighborhood to see how many venues we had per neighborhood:



	Venue
Neighborhood	
Acres Home	5
Addicks / Park Ten	7
Afton Oaks / River Oaks	100
Agassiz	100
Alamo Heights	73
Alief	34
Allandale	100
Area 2/MIT	100
Astrodome Area	100
Balcones Woods	75
Barrington Oaks	50
Barton Hills	97
Battle Bend Springs	94
Bouldin Creek	100

Notice how some of the neighborhoods have very few venues (i.e. Acres Home has five venues, Addicks/Park Ten has seven neighborhoods, etc.). We decided that a neighborhood needed a minimum of ten venues in order to characterize it properly since we are considering the top five types of venues for characterization. Consequently, we dropped all neighborhoods with less than ten venues returned.

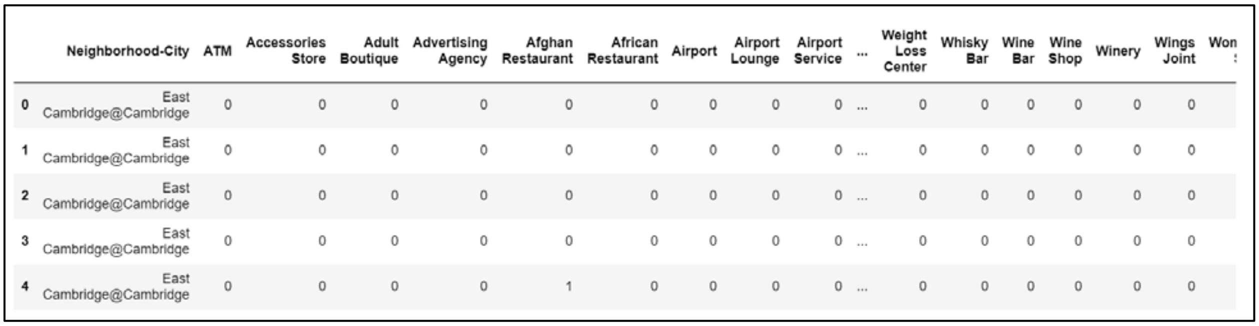
We then grouped the type of venues by neighborhood to see what the five most common types of venues were in each neighborhood. We found the following:



	Neighborhood-City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Afton Oaks / River Oaks@Houston	Clothing Store	Hotel	Department Store	Cosmetics Shop	Shopping Mall
1	Agassiz@Cambridge	Café	Mexican Restaurant	Bakery	Japanese Restaurant	Pizza Place
2	Alamo Heights@San Antonio	Ice Cream Shop	Bakery	Italian Restaurant	American Restaurant	Vegetarian / Vegan Restaurant
3	Alief@Houston	Fast Food Restaurant	Pizza Place	Video Store	Discount Store	Taco Place
4	Allandale@Austin	Bakery	Food Truck	Pizza Place	Ice Cream Shop	Bar
5	Area 2/MIT@Cambridge	Bakery	American Restaurant	Pizza Place	Coffee Shop	Italian Restaurant
6	Astrodome Area@Houston	Mobile Phone Shop	Mexican Restaurant	Sandwich Place	Donut Shop	Coffee Shop
7	Balcones Woods@Austin	Gym / Fitness Center	Pet Store	Chinese Restaurant	Grocery Store	Park
8	Barrington Oaks@Austin	Asian Restaurant	Video Store	Pizza Place	Park	Sandwich Place
9	Barton Hills@Austin	Taco Place	Art Gallery	Trail	Gym	Bar
10	Battle Bend Springs@Austin	Hotel	Convenience Store	Mexican Restaurant	Taco Place	Pizza Place

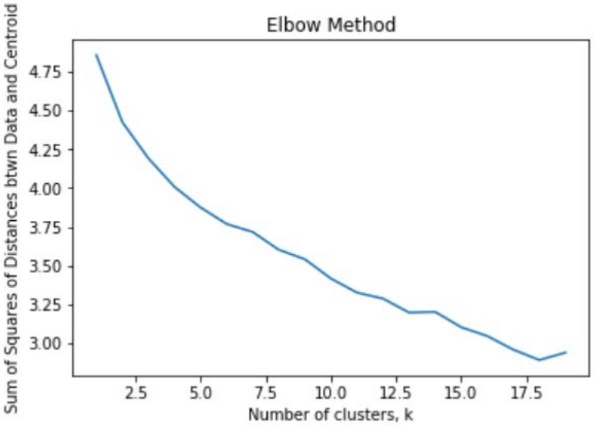
The above dataframe will help us later on to understand the results of our analysis.

We then took each neighborhood and performed a onehot encoding of the types of venues to turn the categorical data into numerical data for our *k*-means modelling later on:

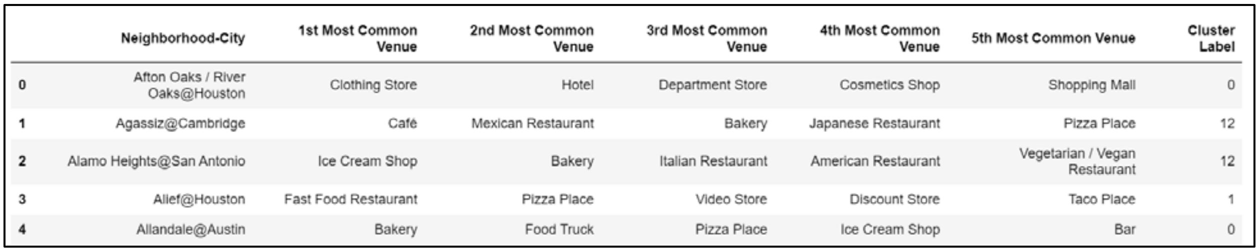


	Neighborhood-City	ATM	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	...	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Won
0	East Cambridge@Cambridge	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	East Cambridge@Cambridge	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	East Cambridge@Cambridge	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	East Cambridge@Cambridge	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	East Cambridge@Cambridge	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	

Next, we took our onehot encoding data and fed it into a *k*-means model to create our clusters. In order to determine what value of *k* to use in our model, we utilized the elbow method and returned the following graph:



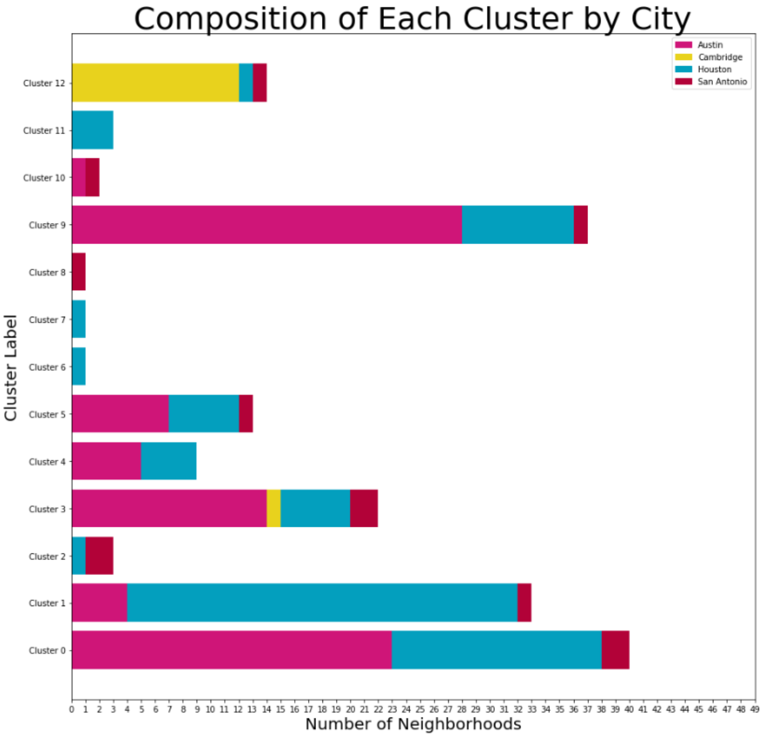
As can be seen from the graph, our first “elbow” appears at *k=13*, so we used this value in our final model. The *k*-means model returned cluster labels which we then added to our dataframe for the top five most common types of venues:



	Neighborhood-City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster Label
0	Afton Oaks / River Oaks@Houston	Clothing Store	Hotel	Department Store	Cosmetics Shop	Shopping Mall	0
1	Agassiz@Cambridge	Café	Mexican Restaurant	Bakery	Japanese Restaurant	Pizza Place	12
2	Alamo Heights@San Antonio	Ice Cream Shop	Bakery	Italian Restaurant	American Restaurant	Vegetarian / Vegan Restaurant	12
3	Alief@Houston	Fast Food Restaurant	Pizza Place	Video Store	Discount Store	Taco Place	1
4	Allandale@Austin	Bakery	Food Truck	Pizza Place	Ice Cream Shop	Bar	0

Let’s pause: what did we just accomplish and why? We broke each of our cities up into their respective neighborhoods, collected data on 10 – 100 venues for each neighborhood, grouped those venues by venue type, and then performed a *k*-means algorithm on those groupings. Our algorithm returned a label for each neighborhood based on its analysis. This label tells us that neighborhoods bearing the same cluster label are similar to each other in their venue composition. For example, we might infer from the dataframe shown above that cluster 12 is comprised of neighborhoods that have many restaurants/places to eat. We can then look at the neighborhoods in the clusters to see which parts of Houston, San Antonio, and Austin are similar to which parts of Cambridge.

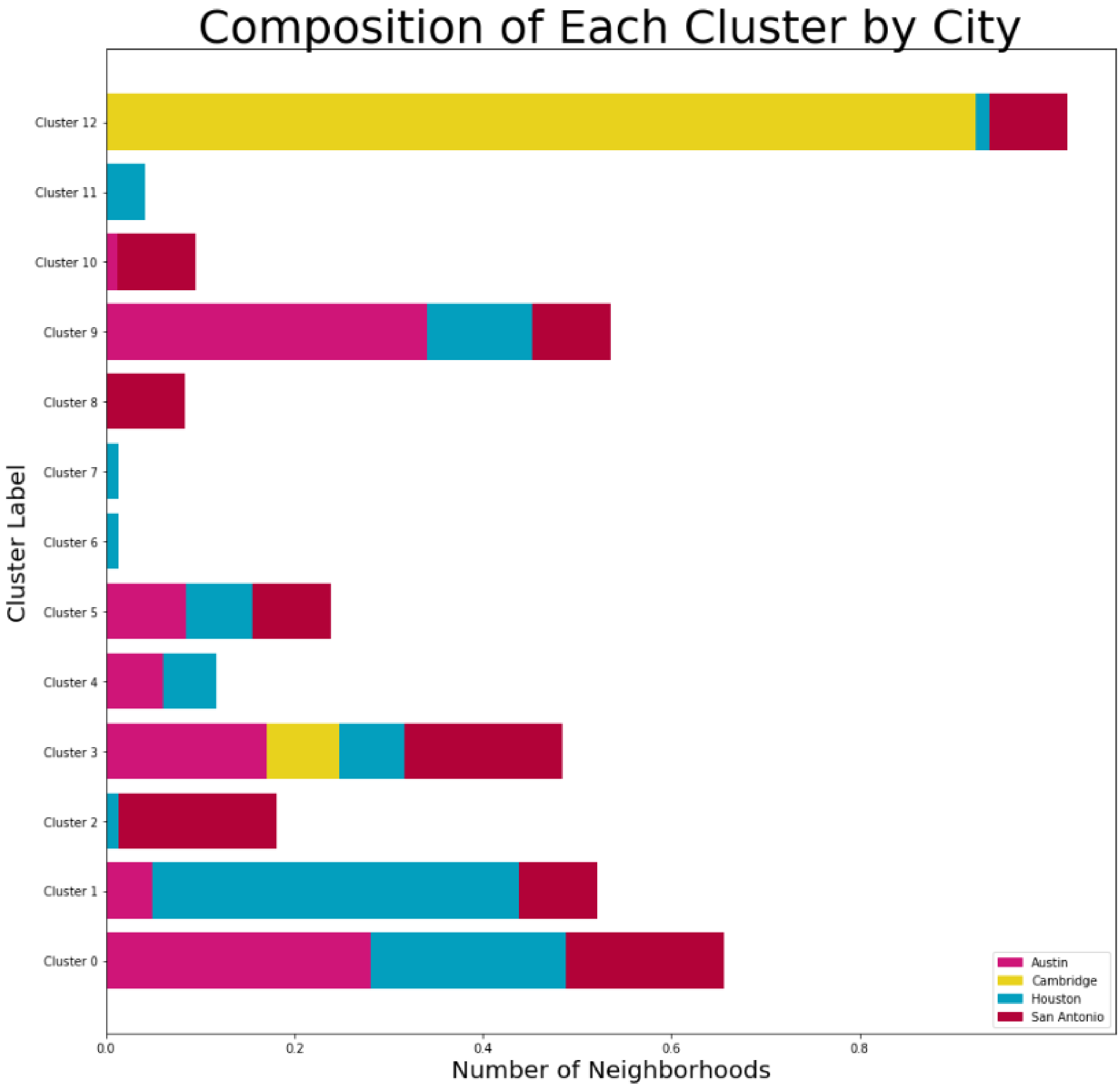
Lastly, we created a stacked, horizontal bar chart of our cluster labels with the stacks created by the city of the neighborhood. We found the following:



Notice that Houston and Austin dominate the graph. This is because our data is not normalized. Because Houston and Austin have so many more neighborhoods compared to San Antonio and Cambridge, overall they will dominate the composition of many clusters. To remedy this problem, we normalized our data by dividing by the sum of the neighborhoods for each city. Our results are discussed in the next section.

Results.

After normalization, our graph looks like the following:



Notice that Cambridge only appears in 2 clusters (cluster 3 and cluster 12) and, in fact, dominates cluster 12. This would seem to imply that neighborhoods in Cambridge are, for the most part, similar to each other but not similar to neighborhoods in any of our three cities. There is some evidence in our chart that suggests this may also hold true for Austin, Houston, and San Antonio (but to a lesser degree). For example, clusters 6, 7, and 11 are entirely made up of Houston neighborhoods while clusters 2, 8, and 10 are dominated by San Antonio. This would imply that these neighborhoods from these cities are similar to each other, but not so similar to neighborhoods in other cities.

Interestingly, Austin decisively dominates only one cluster (cluster 9) compared to three clusters for San Antonio and 4 clusters for Houston. Consequently, one might think that Austin makes an appearance in more clusters than Houston or San Antonio (the neighborhoods of Austin should be more spread out since they only group around one cluster). However, this is not the case, and Austin actually occurs with less frequency among the clusters than both San Antonio and Houston. This implies that the neighborhoods of Austin fall into a smaller number of categories when characterizing them (as opposed to Houston which has a very diverse set of neighborhoods as shown by the number of clusters it appears in on our graph).

Discussion.

So what does this mean in terms of choosing Austin, San Antonio, or Houston? It seems like our efforts to pick a city that is similar to Cambridge, MA out of the three listed is not a fruitful endeavor (at least, trying to decide based off neighborhoods and their venues doesn't seem to be helpful).

However, our efforts aren't completely wasted. Notice that clusters 12 and 3 include neighborhoods for Austin, San Antonio, and Houston. So while we may not be able to pick a city that is similar to Cambridge, what we *can* do is find the areas in Austin, San Antonio, and Houston that are similar to areas of Cambridge. These areas are listed below:

Areas Similar to Cambridge, MA from San Antonio, Austin, or Houston, TX

City	Neighborhood
Austin	Balcones Woods
	Barrington Oak
	Cherry Creek
	Clarksville Historic District
	Estates of Brentwood
	Kincheonville
	Maple Run
	Mueller
	Northwest Hills
	Onion Creek
	Sendera
	Tarrytown
	West Line Historic District
	Woodstone Village
Houston	Braeswood
	Central Northwest
	Clear Lake
	Washington Avenue Coalition / Memorial Park

	Westbury
	Greater Heights
San Antonio	Alamo Heights
	North Central
	East Side

Conclusion.

In our report, we showed how we could potentially use neighborhood venue data in order to compare different cities. Future work on this project could include more clearly defining the boundaries of our neighborhoods (vs. utilizing circles) in order to improve the accuracy of our analysis. Additionally, we could also increase the limit on the number of venues we pull into our database using the FourSquare API.

I greatly enjoyed working on this project, and I can't wait to explore the neighborhoods of Houston, Austin, and San Antonio that are similar to some of the neighborhoods in Cambridge. Happy city exploring!