Selecting Neighborhoods when Moving to Austin, TX

An Exploration of Applied Data Science

Bob Blackard

February 4th, 2021

Contents

Introduction 3
Available Data 3
Exploratory Data Analysis 4
Dataset Cleanup 4
Neighborhood Reporting Areas 5
Neighborhood Venues 6
Housing and Population 8
Families and Age 10
Results 11
Venues 12
Housing and Population 13
Families and Age 13
Discussion 14
Conclusion 15
References 15

Introduction

Austin, TX is an interesting town. I say town, but the Austin Greater Metropolitan area has a population in excess of 2.2 million as of 2019 according to census data found on Wikipedia - Greater Austin. Austin itself is the 11th largest US city, and the 2nd largest US State Capital according to Wikipedia - List of US cities by population.

Austin bills itself as The Live Music Capital of the world (see Austin Relocation Guide), hosts University of Texas at Austin, is one of the very few State Capitols with a large economy unrelated to government business, and ranked number 1 in the Best State Capitals to Live in in 2020. Further, Austin is the heart of the Silicon Hills Wikipedia - Silicon Hills, hosting an incredible number of variety of high-tech companies, providing a wealth of employment opportunities for people in a variety of areas. Austin boasts a wide variety of opportunities for outdoor activity, and also the home of the first purpose-built Formula 1 venue in the United States, Circuit of the Americas, which hosts a number of events and activities throughout the year, not all directly associated with motorsports.

Considering Greater Austin has experienced growth of over 100 people per day for several consecutive years, and is one of the fastest growing cities with over 1 million population (see Austin remains fastest-growing big city in country), it makes an interesting city to investigate as a place to live.

Choosing where to live in a new city can be a daunting task and a variety of criteria may be interesting to people, but not all criteria are important to all people. For example, a young family might be concerned about schools, but that is of less interest to a young single person or a DINK (dual income, no kids) couple. But there are criteria that can lead to finding options for places to live.

The question, then, is: What might be good neighborhoods for a new Austinite to choose to live in?

This investigation uses Data Science tools and techniques to help a prospective new Austinite narrow down area of the city to investigate as places to live. Specifically, Jupyter Notebooks, Python, GeoJSON data, and K-Menas clustering are used to group areas together, allowing the selection of a smaller set of areas to live based on criteria that might be important to a given person.

Available Data

Unfortunately, there isn't an aggregated collection of datasets covering all the Counties and Municipalities that make up Greater Austin. Further, I was unable to find disconnected datasets that could create a unified picture of the area. For example, while West Lake Hills is a very nice area, with well rated schools and a diverse community, it is not part of the City of Austin. Neither are several areas that have chosen not to be incorporated with the city, but are essentially surrounded by it.

Therefore, the investigation will be limited to the Incorporated areas of the City of Austin itself. Unincorporated areas, and interesting, vibrant and growing areas outside the city limits are excluded from analysis.

Austin provides access to a trove of datasets through the Austin Open Data Portal, as well as datasets from Austin's GIS Data and Maps Department, and Demographic Data aggregated by the city. Digging into the data, there are three primary geographic breakdowns for the city. First is what Austin calls Neighborhood Reporting Areas. Each Neighborhood Reporting Area is comprised of several smaller elements such as subdivisions or individual neighborhoods, but The City of Austin uses Neighborhood Reporting Areas as the primary mechanism for planning and reporting.

There are also breakdowns by Census Tract or Census Block Group, and breakdowns by Zip Code. While there is a one-to-many relationship between Counties and Census Tracts, the relationships between Zip Codes and Cencus Tracts, Neighborhoods to Zip Code, or Neighborhoods to Cencus Tracts are many-to-many - meaning one Neighborhood may be related to more than one Zip Codes, and one Zip Code may be related to more than one Neighborhood, for example. Therefor the investigation is limited to the use of Neighborhood Reporting Area.

The specific datasets and sources selected for this analysis are:

- There is a GeoJSON file on the Austin Open Data Portal providing boundary data for each of Austin's 103 Neighborhood Reporting Areas. The latest version of that file is from January 4, 2021.

- The City of Austin also maintains a Demographic Data site that provides two interesting sets of demographic data by Neighborhood Report Area as Excel files - Table I: population, race and ethnicity, housing and density and Table II: household characteristics and age structure.

- Finally, nearby venue data for each neighborhood collected using the Foursquare API will be processed to extract frequencies for different types of venues.

Combining demographic and venue frequency data should result in some interesting analysis opportunities.

Exploratory Data Analysis

Ultimately, clustering to group neighborhoods into related sets will be attempted. An investigative dataset will be built by combining sets of demographic data with frequencies of the nearby venue types reported by Foursqure within and given distance the neighborhood center. Next, this data will be used to perform K-Means Clustering, and the resulting clusters will be analyzed to try and extract key or defining characteristics of each cluster. This information will be presented for use when seeking to determine where one might choose to live if moving to Austin, TX.

Before finally integrating the data, some initial work was done to handle each dataset independently.

In all cases, neighborhood boundary data from the January 4, 2021 GeoJSON file was used to produce a color-coded Choropleth map in Folium, rather than color-coding map markers. By color coding the bounding areas of each neighborhood, it is hoped the content will be more meaningful.

Dataset Cleanup

There are some discrepancies in the datasets that need to be resolved. Specifically, while the neighborhood names between the two Excel tables match, there are some neighborhoods in those tables that are not in the GeoJSON data. Specifically, there is data in one set that is not in another:

In Excel Tables - Not in GeoJSON
Central West Austin
Mueller
NACA

In GeoJSON - Not in Excel Tables
North Lamar Rundberg
RMMA
West Austin Ng

Additionally, there are discrepancies in some of the data that must be mapped in order to correctly match between the sources:

In GeoJSON	In Excel Tables
Allandale	Allendale
Bouldin Creek	Bouldin
Davenport Lake Austin	Davenport--Lake Austin
Georgian Acres	Georgia Acres
McNeil	McNiel
Pecan Springs-Springdale	Pecan Springs--Springdale
St. Johns	St. John
Sweetbriar	Sweet Briar
Triangle State	Triangle-State

Finally, inspection of the various datasets revealed that one Neighborhood Planning Area encompassed nearly nothing more than Austin Bergstrom International Airport, with zero Households, and only two (2) Housing Units. It seems safe to exclude this from the analysis.

Neighborhood Reporting Areas

The GeoJSON data was processed first as it provided a source for latitude and longitude of each neighborhood. While ArcGIS was used to find the latitude and longitude for the City of Austin as a whole, it did not prove a reliable source for the center of the neighborhoods. Using the Python GeoPandas library provided a mechanism of computing the latitude and longitude from the GeoJSON feature data which proved to be quite accurate. Additionally, the feature data provided a source for computing an approximate radius for the geographic region that could be used with the Foursquare API to better focus the venue search feature thereof.

Note: The source GeoJSON file was processed using Python to change the capitalization of the neighborhood names. (Blackard, 2021)

After computing the location and radius data, Folium maps were created to validate the results, one map for the city as a whole, and one for a selected Neighborhood Reporting Area with a complex, disconnected boundary, that being Del Valle East.

While not exact, the location and radius data proved sufficiently close for the purposes of this analysis, and the review provided confidence that the GeoJSON data would provide value.

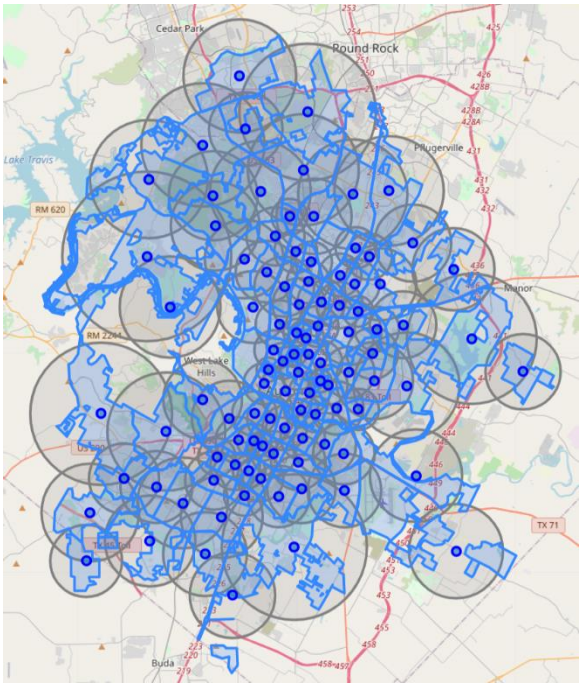


Figure 1 - Austin Neighborhood Map

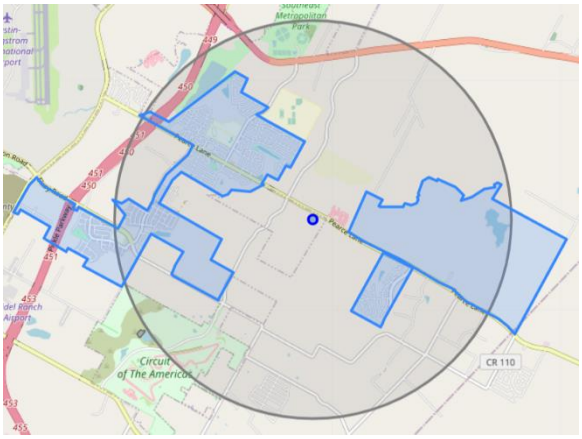


Figure 2- Del Valle East Neighborhood

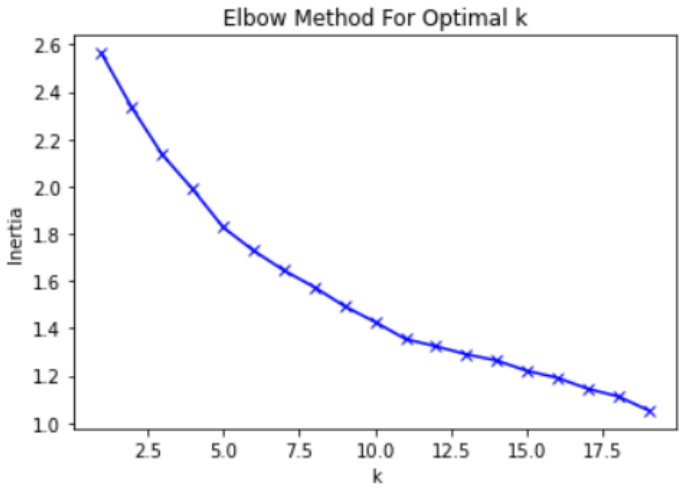
Neighborhood Venues

The Foursquare API was used to collect neighborhood venue data, which was then converted into frequency data that could be used to perform K-Means Clustering. First a subset of the data was used to find the most frequently found venues for a subset of the neighborhoods:

	Neighborhood	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
0	Allandale	Mexican Restaurant	Pizza Place	Coffee Shop	Burger Joint	Gym	Taco Place	Korean Restaurant	Indian Restaurant	Ice Cream Shop	Asian Restaurant
1	Anderson Mill	Park	Mexican Restaurant	Italian Restaurant	Grocery Store	Sushi Restaurant	Burger Joint	Movie Theater	Pizza Place	Bakery	Seafood Restaurant
2	Avery Ranch--Lakeline	Gym	Pizza Place	American Restaurant	Sandwich Place	Spa	Coffee Shop	Park	Pharmacy	Donut Shop	Movie Theater
3	Barton Creek Mall	Trail	Coffee Shop	Pizza Place	Park	Bakery	Sandwich Place	Cosmetics Shop	Intersection	Seafood Restaurant	Scenic Lookout
4	Barton Hills	Coffee Shop	Taco Place	Trail	Ice Cream Shop	Grocery Store	American Restaurant	Cosmetics Shop	Sandwich Place	Bar	Spa

Figure 3- Most Frequent Venues for First Five Neighborhoods

An analysis to find the best k for the K-Means Clustering normally uses what is referred to as “the elbow method” to evaluate the standard deviation and find where leverage ceases being found for increasing values of k based on the tendency of this increase to approach zero. However, for this dataset, K-Means Clustering didn’t converge as rapidly as is normally seen.



As can be seen, this graph doesn't really converge, and doesn't have a clear "elbow". It looks like the elbow might be somewhere between 5 and 10. Taking a look at some other data from the clustering, we can see that the higher k's are not necessarily valuable.

As k increases, there are more single-member clusters, which themselves aren't very useful to making choices between similar neighborhoods. But at k=6, we have just one single-member cluster, meaning we have five clusters useful for making choices.

	K	Inertia	Single-Member Clusters	Multi-Member Clusters	Max Cluster Size
1	1.0	2.563672	0.0	1.0	99.0
2	2.0	2.334243	0.0	2.0	88.0
3	3.0	2.135081	1.0	2.0	75.0
4	4.0	1.989036	2.0	2.0	87.0
5	5.0	1.828109	2.0	3.0	71.0
6	6.0	1.729225	1.0	5.0	59.0
7	7.0	1.645566	4.0	3.0	42.0
8	8.0	1.574904	4.0	4.0	38.0
9	9.0	1.493703	4.0	5.0	33.0
10	10.0	1.427222	4.0	6.0	28.0

Proceeding with the K-Means analysis, and mapping the data with Folium resulted in the following map.

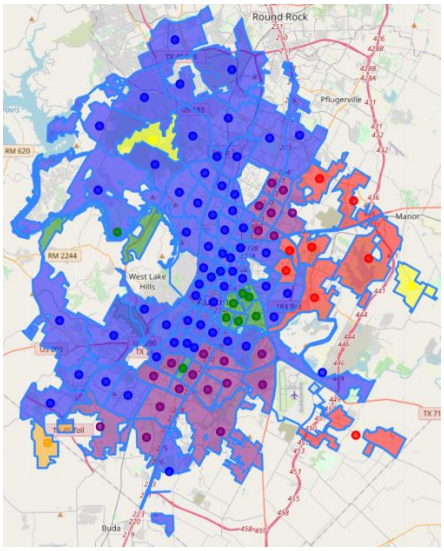


Figure 4 - Colorized Map of Neighborhoods Clustered by Venue

This represents just grouping based on venues, which is sufficient to indicate the usefulness of the dataset, but is not sufficient for out purposes. More data needs to be included in the analysis.

Housing and Population

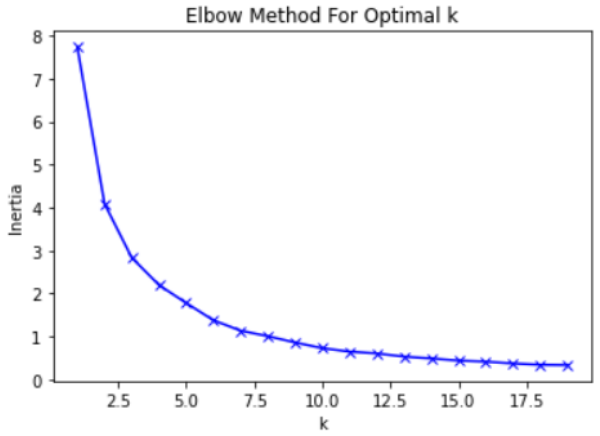
Starting with the source data for Housing, Ethnicity and Population, Pandas was used to collect the key data for this round of processing and the next, with some of the fields converted from counts to percentages. Additionally, the Ethnicity data was dropped as it shouldn't be relevant to picking where to live. The first five rows of the resulting dataframe can be seen here.

	Neighborhood	Total Pop	Total Units	Occupied Housing	Vacant Housing	Owner Occupied	Pop Density	Acres	Pop Density Scaled
0	Allandale	6643	3612	0.903378	0.096622	0.641128	5.096345	1303.483062	0.145573
1	Anderson Mill	28473	11507	0.947597	0.052403	0.672872	4.989661	5706.399565	0.142526
2	Avery Ranch--Lakeline	14785	6108	0.930092	0.069908	0.605351	3.185472	4641.385040	0.090990
3	Barton Creek Mall	5147	2195	0.945786	0.054214	0.637283	2.220364	2318.088291	0.063423
4	Barton Hills	8022	4965	0.929305	0.070695	0.313827	3.936028	2038.095147	0.112430

Having Pandas describe the data provides confidence that the values for Occupied and Vacant Housing, Owner Occupied Housing and a called Population Density would provide good metrics for cluster analysis.

	Total Pop	Total Units	Occupied Housing	Vacant Housing	Owner Occupied	Pop Density	Acres	Pop Density Scaled
count	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000
mean	7674.818182	3460.212121	0.912347	0.087653	0.466428	6.176468	1963.767531	0.176426
std	5894.031890	2665.591343	0.056270	0.056270	0.232560	4.758289	1982.370437	0.135917
min	2.000000	2.000000	0.500000	0.000000	0.000000	0.001113	180.526570	0.000032
25%	3920.500000	1662.000000	0.898362	0.057155	0.323618	2.852885	646.534073	0.081490
50%	5779.000000	2758.000000	0.916963	0.083037	0.454545	5.708519	1230.645954	0.163059
75%	10400.500000	4626.000000	0.942845	0.101638	0.595625	7.936201	2306.333859	0.226691
max	28473.000000	13330.000000	1.000000	0.500000	1.000000	35.008841	9993.018538	1.000000

Note the zero or near zero minimum values and 1 or near 1 maximum values for these fields. Again, an analysis to find the best k results in some lack of clarity, but examining additional data proves useful.



	K	Inertia	Single-Member Clusters	Multi-Member Clusters	Max Cluster Size
1	1.0	7.731252	0.0	1.0	99.0
2	2.0	4.085708	0.0	2.0	59.0
3	3.0	2.831586	0.0	3.0	55.0
4	4.0	2.195541	0.0	4.0	49.0
5	5.0	1.777720	0.0	5.0	36.0
6	6.0	1.376442	1.0	5.0	36.0
7	7.0	1.134374	2.0	5.0	40.0

Again, the cluster results were computed and used to create Folium map, color coded by cluster. Note the better distribution here than was seen for venue analysis.

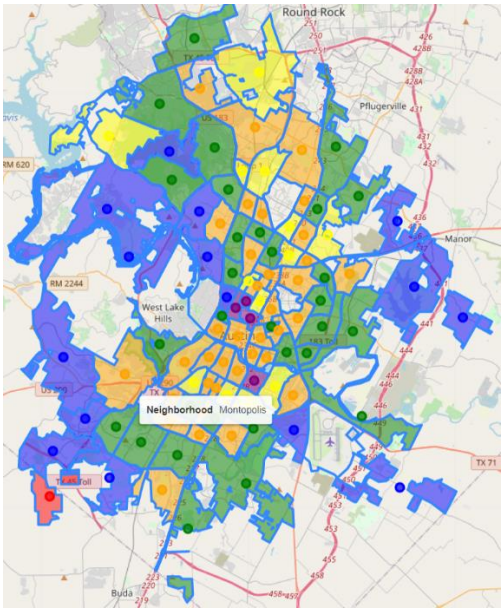


Figure 5 - Colorized Map of Neighborhoods Clustered by Housing and Population Data

As was true with venue analysis, this isn't sufficient on its own for decision making, but proves the data can be used effectively for clustering.

Families and Age

Similar to the approach with Housing, Ethnicity and Population data, the Family and Age data was processed using Pandas, with some data being converted to percentages, and other values being dropped. Further, the age groupings were consolidated as follows:

- Young Children : Pct Age 0 to 4
- School Children : Pct Age 5 to 9, Pct Age 10 to 14, Pct Age 15 to 17
- Young Adult : Pct Age 18 to 19, Pct Age 20 to 24
- Adult : Pct Age 25 to 34, Pct Age 35 to 44, Pct Age 45 to 54, Pct Age 55 to 59
- Senior : Pct Age 60 to 64, Pct Age 65 to 74, Pct Age 75 to 84, Pct Age Over 85

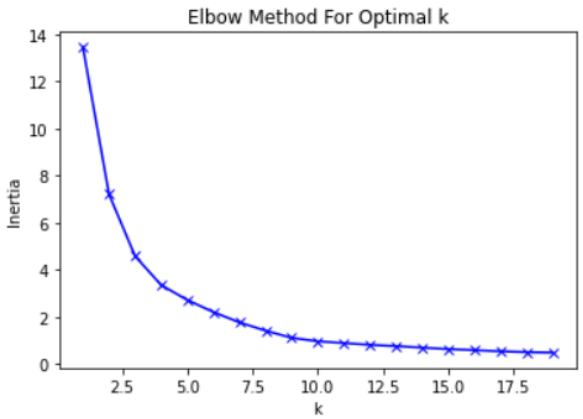
The first five rows of the resulting dataframe are as follows:

	Neighborhood	Total Pop	Total Households	Pop in Households	Household Size	Pop in Group Quarters	Family Households	Non-Family Households	Families with Children	Pct Families with Children	Single Mother Headed Households	Pct Single Mother Headed Households	Young Children
0	Allandale	6643	3263	6621	2.02911	22	0.505670	0.494330	723	0.221575	107	0.0327919	0.063224
1	Anderson Mill	28473	10904	28368	2.60161	105	0.716434	0.283566	4037	0.370231	587	0.0538335	0.062410
2	Avery Ranch--Lakeline	14785	5681	14785	2.60253	0	0.682626	0.317374	2472	0.435135	294	0.0517515	0.111532
3	Barton Creek Mall	5147	2076	5099	2.45617	48	0.654624	0.345376	719	0.346339	93	0.0447977	0.052263
4	Barton Hills	8022	4614	7890	1.71001	132	0.313177	0.686823	556	0.120503	119	0.0257911	0.030666

The data descriptions again provided good metrics for those fields that would be involved in the analysis:

	Total Pop	Total Households	Pop in Households	Pop in Group Quarters	Family Households	Non-Family Households	Families with Children	Pct Families with Children	Single Mother Headed Households	Young Children	School Children	Yo A
count	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.00(
mean	7674.818182	3176.909091	7475.989899	198.828283	0.526937	0.473063	813.404040	0.258255	201.111111	0.069278	0.142569	0.142
std	5894.031890	2500.368243	5891.353920	838.255733	0.193475	0.193475	771.372986	0.132001	183.425005	0.029943	0.066150	0.148
min	2.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00(
25%	3920.500000	1496.000000	3775.500000	2.000000	0.397187	0.340058	327.500000	0.163893	77.500000	0.053614	0.097894	0.084
50%	5779.000000	2521.000000	5701.000000	21.000000	0.537685	0.462315	554.000000	0.241155	142.000000	0.065543	0.140750	0.108
75%	10400.500000	4288.500000	9930.000000	99.000000	0.659942	0.602813	1121.000000	0.338313	289.000000	0.091428	0.189181	0.132
max	28473.000000	12659.000000	28368.000000	7240.000000	1.000000	1.000000	4037.000000	0.595745	871.000000	0.144172	0.333333	0.985

Similar results for finding the best k were seen for this dataset as well.



	K	Inertia	Single-Member Clusters	Multi-Member Clusters	Max Cluster Size
1	1.0	13.459727	0.0	1.0	99.0
2	2.0	7.209846	0.0	2.0	58.0
3	3.0	4.582356	0.0	3.0	51.0
4	4.0	3.352924	0.0	4.0	37.0
5	5.0	2.720518	1.0	4.0	37.0
6	6.0	2.207156	1.0	5.0	40.0
7	7.0	1.767748	1.0	6.0	32.0
8	8.0	1.416871	2.0	6.0	32.0

Similar results were also seen with mapping the clusters, computing the clusters and mapping with Folium.

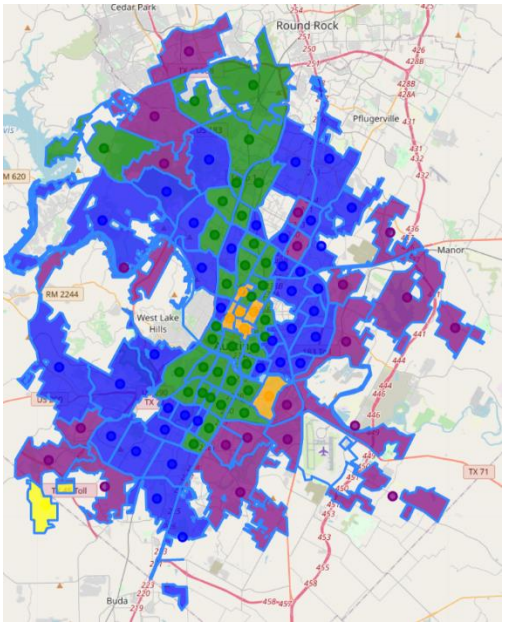


Figure 6 - Colorized Map of Neighborhoods Clustered by Family and Age Data

Results

Having developed confidence that each dataset independently was usable for cluster analysis, the source data was aggregated into a single larger (wider) dataset and that dataset was used to analyze the best k, compute the cluster and create the Folium map.

Having completed clustering with aggregated data, analysis of the characteristics of each cluster can proceed. This is done be applying the cluster labels to the source data and recomputing the key metrics based on aggregated values. This allows the examination of the characteristics for the clusters as a whole.

The resulting colorations by cluster are:

Cluster	Color
0	Blue
1	Orange
2	Green

3	Purple
4	Yellow
5	Red

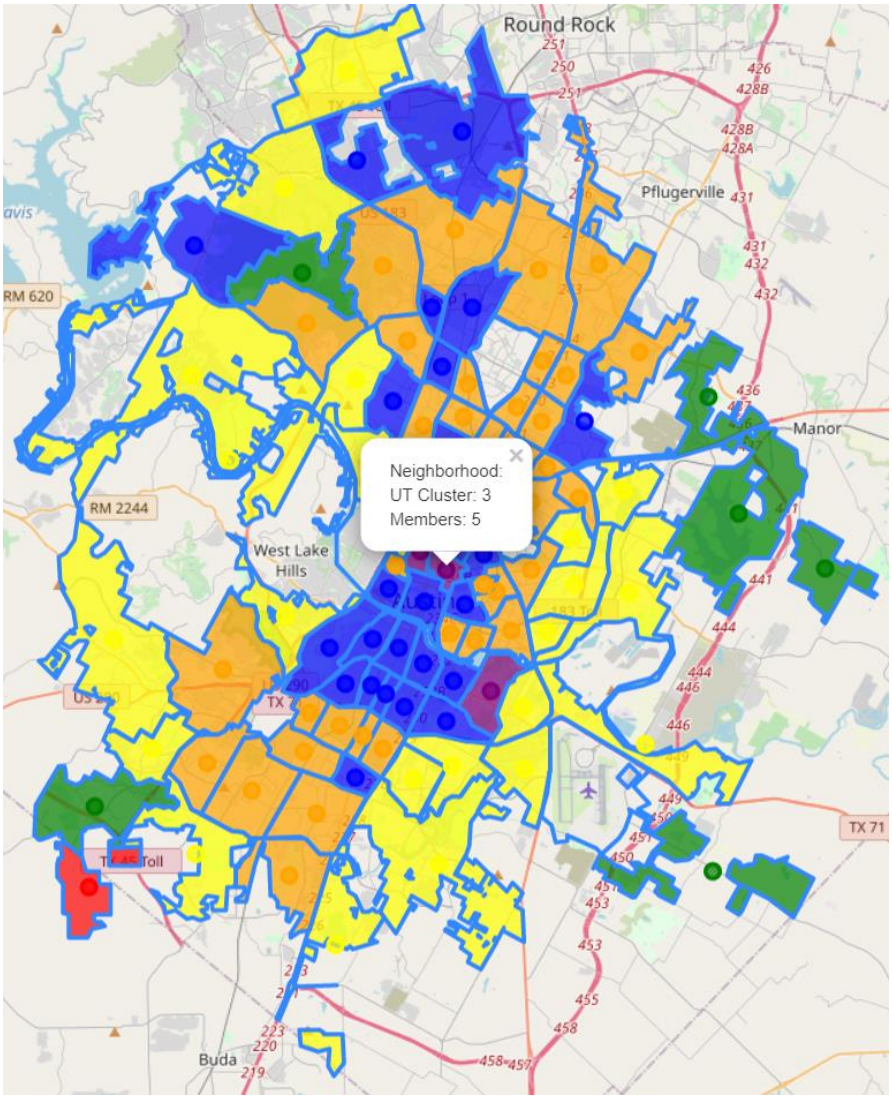


Figure 7 - Colorized Map of Neighborhoods by Aggregated Data

Venues

By mapping each neighborhoods to its cluster and then recomputing the venue frequency data, the most frequent venues for each cluster can be examined.

	Cluster Labels	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
0	0	Coffee Shop	Mexican Restaurant	Taco Place	Sandwich Place	Pizza Place	Food Truck	Park	Burger Joint	Bar	American Restaurant
1	1	Mexican Restaurant	Coffee Shop	Pizza Place	Sandwich Place	Park	Grocery Store	Food Truck	Burger Joint	Fast Food Restaurant	Convenience Store
2	2	Park	Intersection	Convenience Store	Golf Course	Trail	Music Venue	Racetrack	Restaurant	Tennis Court	Pool
3	3	Food Truck	Coffee Shop	Mexican Restaurant	Sandwich Place	Taco Place	Pizza Place	Burger Joint	Park	Bar	Convenience Store
4	4	Mexican Restaurant	Park	Pizza Place	Sandwich Place	Coffee Shop	Convenience Store	Hotel	Fast Food Restaurant	Food Truck	American Restaurant
5	5	Home Service	Furniture / Home Store	Food Truck	Trail	Miscellaneous Shop	Farmers Market	Event Space	Exhibit	Eye Doctor	Fabric Shop

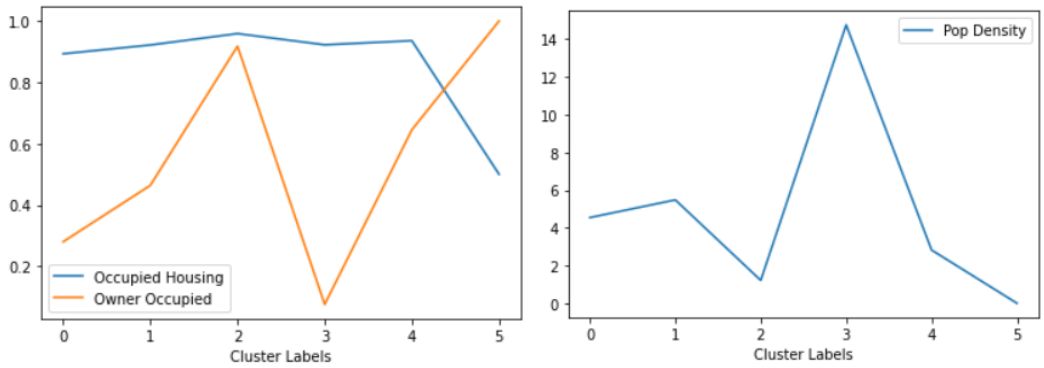
From this clusters 0, 1, 3 and 4 appear reasonably similar, while cluster to has a lot of recreational space, and cluster 5 seems more rural.

Housing and Population

The percentages for Housing and Population data after clustering are as follows.

	Cluster Labels	Occupied Housing	Vacant Housing	Owner Occupied	Pop Density
0	0	0.893131	0.106869	0.279843	4.546549
1	1	0.921766	0.078234	0.463823	5.482517
2	2	0.958952	0.041048	0.917912	1.215120
3	3	0.922387	0.077613	0.075510	14.763077
4	4	0.935747	0.064253	0.645022	2.824604
5	5	0.500000	0.500000	1.000000	0.001113

Note the predominance of Occupied Housing in all but cluster 5, but the wide variety of Owner Occupied housing. Also note that cluster 3 is the highest Population Density, which makes sense being in the core of Austin around the University of Texas. These results are easier to see graphically.

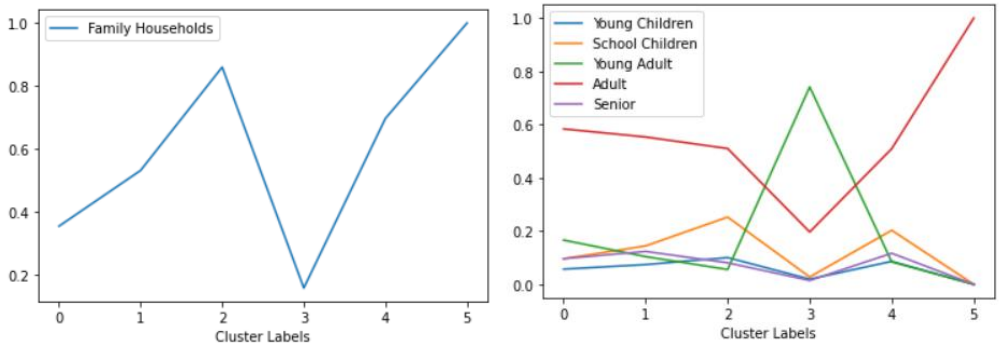


Families and Age

Lastly, the percentages for Family and Age are as follows.

	Cluster Labels	Family Households	Non-Family Households	Young Children	School Children	Young Adult	Adult	Senior
0	0	0.354943	0.645057	0.057217	0.096169	0.166428	0.583691	0.096496
1	1	0.531875	0.468125	0.074254	0.144368	0.104207	0.553420	0.123751
2	2	0.860005	0.139995	0.100538	0.252592	0.055828	0.510010	0.081032
3	3	0.158764	0.841236	0.019617	0.027730	0.741408	0.196703	0.014543
4	4	0.697093	0.302907	0.086431	0.203011	0.084253	0.509518	0.116786
5	5	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000

Similar to the Housing and Population Data, graphs are useful for examination. Note that the Family Households graph does not include Non-Family Households as that is simply an inversion of the former.



Discussion

A review of the results allows for the formulation of some key characteristics of each cluster as follows.

Cluster	Color	Type	Occupied By	Population Density	Families	Children
0	Blue	Residential	Renter	Medium	Mixed	Yes
1	Orange	Residential	Mixed	Medium	Mixed	Yes
2	Green	Residential Recreation	Owner	Low	High	Yes
3	Purple	Residential	Renter	High	Low	No
4	Yellow	Residential	Mixed	Medium Low	High	Yes
5	Red	Rural	Owner	Low	High	No

Clusters 0 and 1 are very similar, with families and children. A closer look at the most common venues suggests that cluster 0 might be a little more ‘casual’, which 1 might be a little more family centric, being the only cluster with Grocery Stores in the top 10 frequently seen venues.

Cluster 2 again seems families with children, but it has a lower population density, and more recreational space in the top 10 venues than other clusters.

Cluster 3 is clearly centered on the University of Texas, with a preponderance of young adults, those most likely to be in college, low owner-occupation, and few families or children.

Cluster 4 is similar to clusters 0 and 1, though with lower population density and a higher percentage of families.

And finally, cluster 5 seems to be an oddity, largely disconnected from the rest of the neighborhoods around Austin.

Conclusion

Data Science is being applied to a wide variety of problems in research and business today. Its tools are powerful, and can uncover significant insights.

The application of these technologies to problems facing everyday people, helping them solve complex problems in a simpler fashion will continue to evolve and the quantity and variety of data available continues to grow.

The application of something as ordinary as deciding what neighborhoods might be best to investigate when relocating to a city provided an opportunity to demonstrate both the power of the tools and techniques of Data Science, but also the opportunities for democratizing information being collected at a more rapid pace than has happen in the history of our world.

References

Blackard, B. (2021). *Selecting Neighborhoods When Moving to Austin, TX.* Retrieved from Jupyter Notebook Viewer: https://nbviewer.jupyter.org/github/blackard/Coursera_Capstone/blob/master/Selecting%20Neighborhoods%20When%20Moving%20to%20Austin%2C%20TX.ipynb