Applied Data Science Capstone Project

IBM Data Science Professional Certificate Specialization – Coursera

Michelle Sanford

Exploring the Austin Texas Area

Introduction

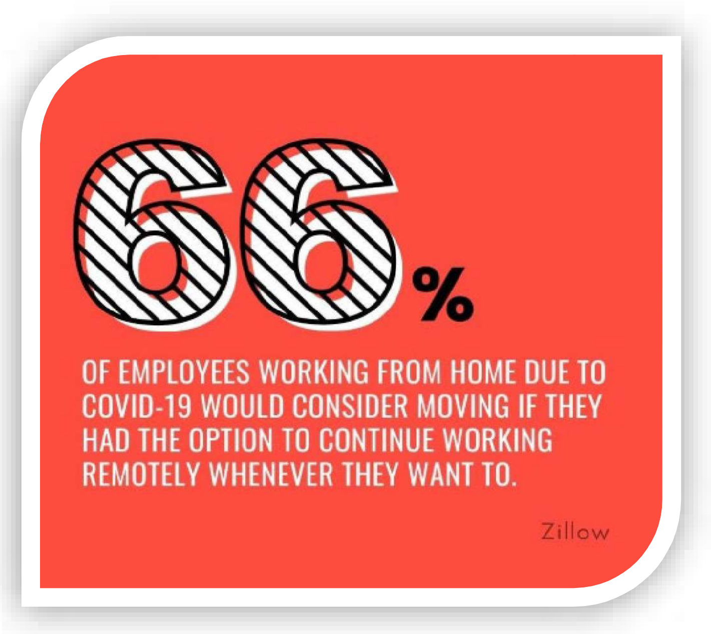
The last few months has forced everyone to adapt in living and working. That has challenged businesses to look at how they can survive the pandemic and adapt. Businesses are reassessing their current fiscal projections and strategies to see how they can adapt and pivot.

Office real estate is being reviewed as one way to significantly reduce operational cost. Employees can live where they would like, and provide value without taking up office space.

This project will explore the Austin Texas area to see if it is a place to call home as the opportunity to work remote and not go to an office everyday could become a growing trend.

The analysis for this project will provide those looking at exploring the Austin Texas area insight into affordability.

“Three quarters of Americans are now working from home because of the coronavirus. Those that were interviewed said they would like to continue if given the option, and two-thirds say they would consider moving if given that flexibility.” http://zillow.mediaroom.com/2020-05-13-A-Rise-in-Remote-Work-Could-Lead-to-a-New-Suburban-Boom


66%
OF EMPLOYEES WORKING FROM HOME DUE TO COVID-19 WOULD CONSIDER MOVING IF THEY HAD THE OPTION TO CONTINUE WORKING REMOTELY WHENEVER THEY WANT TO.
Zillow

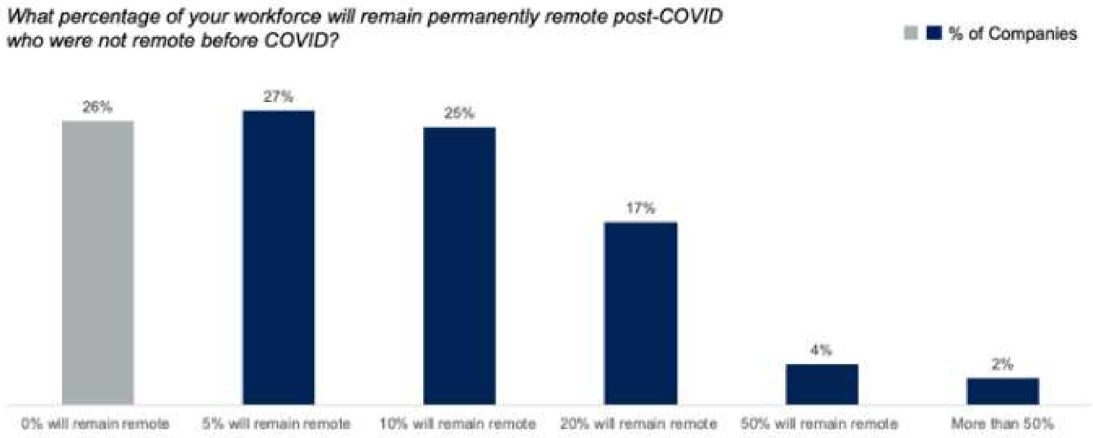
Business Problem:

Is Austin or surrounding area the right place for someone's next move? The workplace is shifting from onsite to remote, offering the opportunity to live wherever the employee chooses. This project will explore the Austin area to see if it is a potential fit.

A Gartner, Inc. survey of 317 CFOs and Finance leaders on March 30, 2020* revealed that 74% will move at least 5% of their previously on-site workforce to permanently remote positions post-COVID 19". Gartner Press Release,

Austin is always a popular choice, but included will be the surrounding area to offer insight into affordability. The comparison, median home prices, and venues will help determine if Austin or one of the surrounding areas is a possible next move.

Figure 1: 74% of Companies Plan to Permanently Shift to More Remote Work Post COVID-19



Source: Gartner (April 2020)

Data

In order to find the necessary data and perform the analysis a number of sources and libraries were used.

- The Foursquare API was used to scrape the surrounding area of Austin Texas for population, median home price, venue categories and the Austin neighborhoods.
- Location data leveraging the categories and details used for analysis can be found on the foursquare site https://developer.foursquare.com/.
- Wikipedia http://www.city-data.com/zipmaps/Austin-Texas.html and OpenCage https://opencagedata.com/ was used to retrieve latitude and longitude, the surrounding areas and populations, which provide the top populated cities

When analyzing the housing prices, sales, median income and neighborhoods, there were several sites used to find the most up to date data as possible.

- Retrieving the details within the uszipcodes led me to a very comprehensive data set on zipmaps http://www.city-data.com/zipmaps/Austin-Texas.html/. This is a great place to get a lot of information and uszips provide a python library to leverage! https://pypi.org/project/uszipcode/.

Libraries

- Numpy to handle data in a vectorized manner.
- Pandas: To create the data frames and analysis.
- Folium: Python visualization library to visualize.
- Scikit Learn: Importing k-means clustering.
- JSON: Library to handle JSON files.
- XML: To separate data from presentation. XML stores data in plain text format.
- Geocoder: To retrieve the location data.
- Matplotlib for plotting.

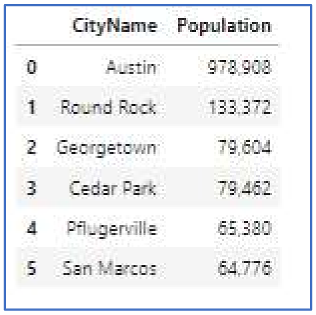
Work Flow

Using the credentials for Foursquare, I leveraged the API, to provide insight into near-by venues. I also used Zillow's API's, USZipcode's library to retrieve and analyze location data.

Methodology

To start the libraries, API's, and credentials were initialized. I wanted to look at Austin and get the top five and then three cities with Austin included to do some basic analysis on population. Interestingly, Georgetown has moved past Cedar park just in the last year.

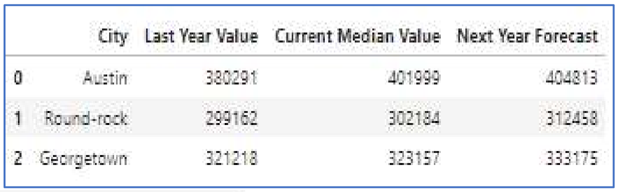
Below are the top five cities, and the top three will be used for the exploration.



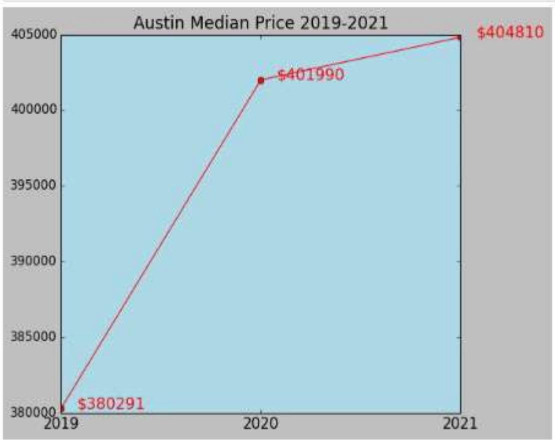
	CityName	Population
0	Austin	978,908
1	Round Rock	133,372
2	Georgetown	79,604
3	Cedar Park	79,462
4	Pflugerville	65,380
5	San Marcos	64,776

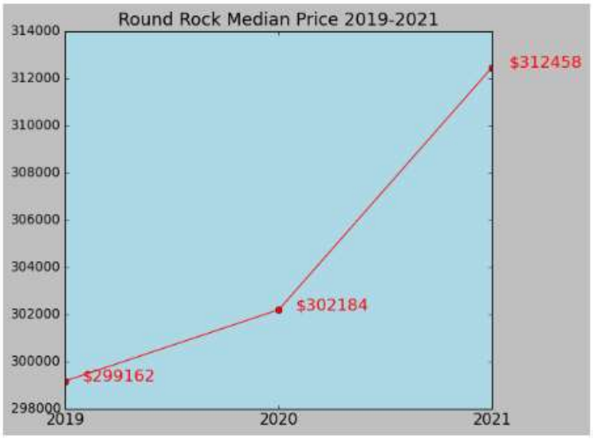
Median Home Price Snapshot

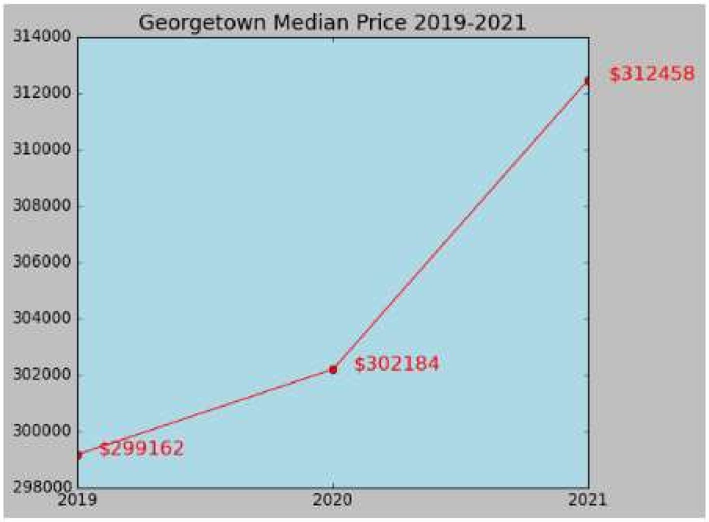
This is where Zillow was leveraged to obtain the necessary housing prices. Based on the analysis, the data shows that all three are expected to have an increase in value for 2021.



	City	Last Year Value	Current Median Value	Next Year Forecast
0	Austin	380291	401999	404813
1	Round-rock	299162	302184	312458
2	Georgetown	321218	323157	333175







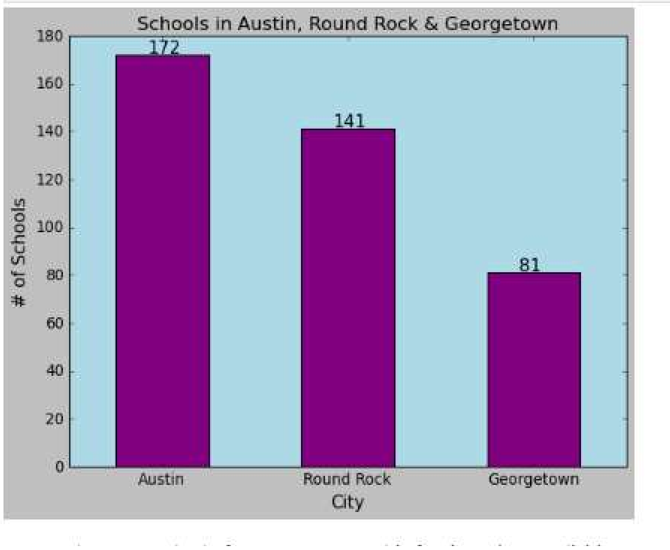
Venue Exploration

Now to explore the schools for parents and college students, as well as food venues across the three cities. Methodology and data analysis was done with FourSquare API's to find a total number of different venues within two categories in foursquare for comparison.

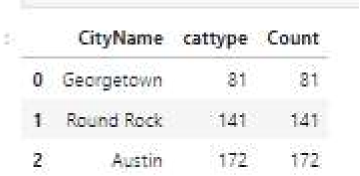
Schools

Schools were inclusive of daycare, elementary through high school, colleges and large universities.

Findings from the data and analysis show the comparison below. We found that Austin had the most number of schools, then Round Rock and Georgetown.



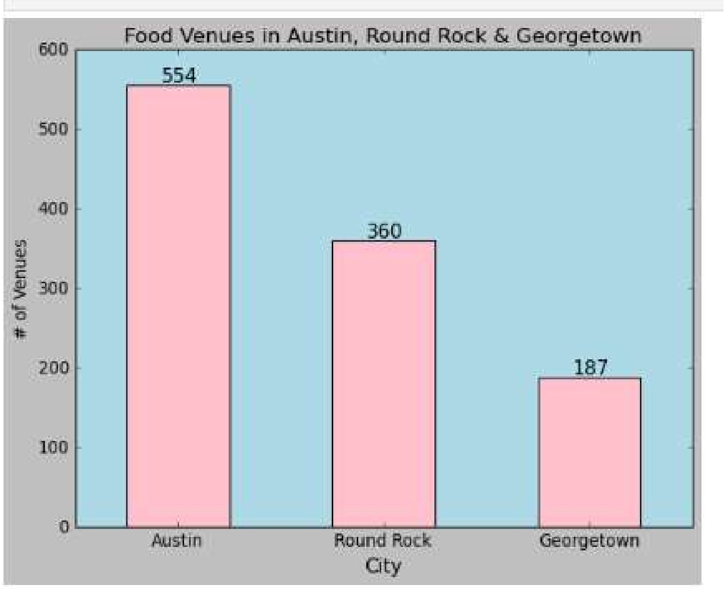
Grouping the data told us the number of each type of school within the different categories, meaning a university, elementary, etc.



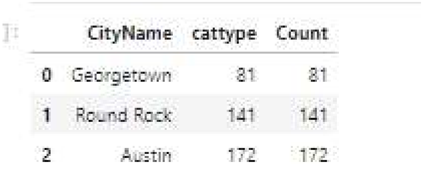
	CityName	cattype	Count
0	Georgetown	81	81
1	Round Rock	141	141
2	Austin	172	172

Food

Grocery stores, access to restaurants and coffee shops were analyzed in the food category. Findings from the data analysis show the comparison below. We found again that Austin had the most number of venues in the food category, Round Rock and then Georgetown.

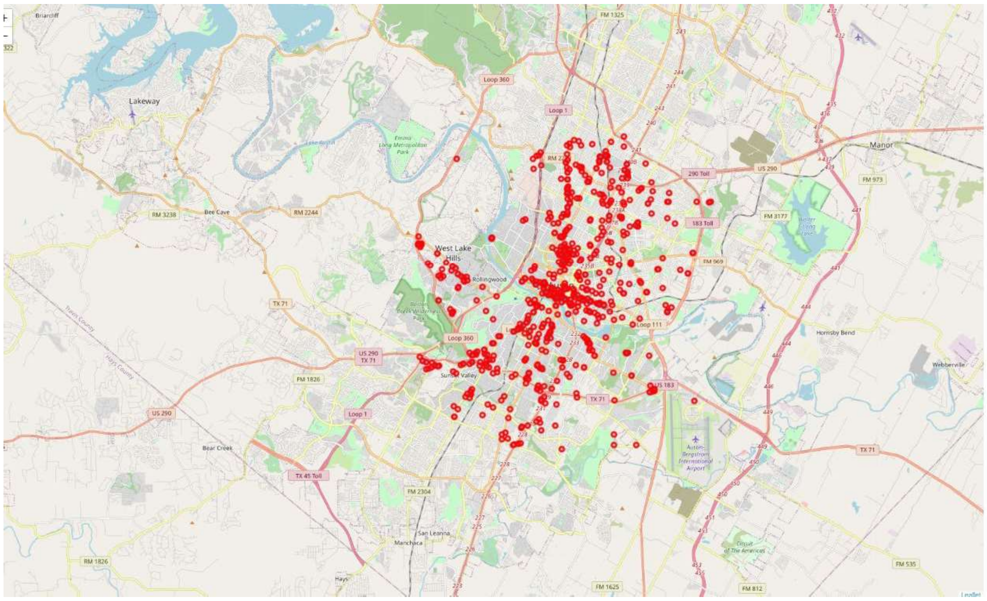


Grouping the data told us the number of each venue within the different food categories, meaning a restaurant, coffee shop, etc.



	CityName	cattype	Count
0	Georgetown	81	81
1	Round Rock	141	141
2	Austin	172	172

Providing the map of these venues in Austin for access to view the different places is also included.



Housing Prices

In order to provide answers to the business problem and insight for the potential Austenite, data from uszipcode.com and a statistical analysis was done.

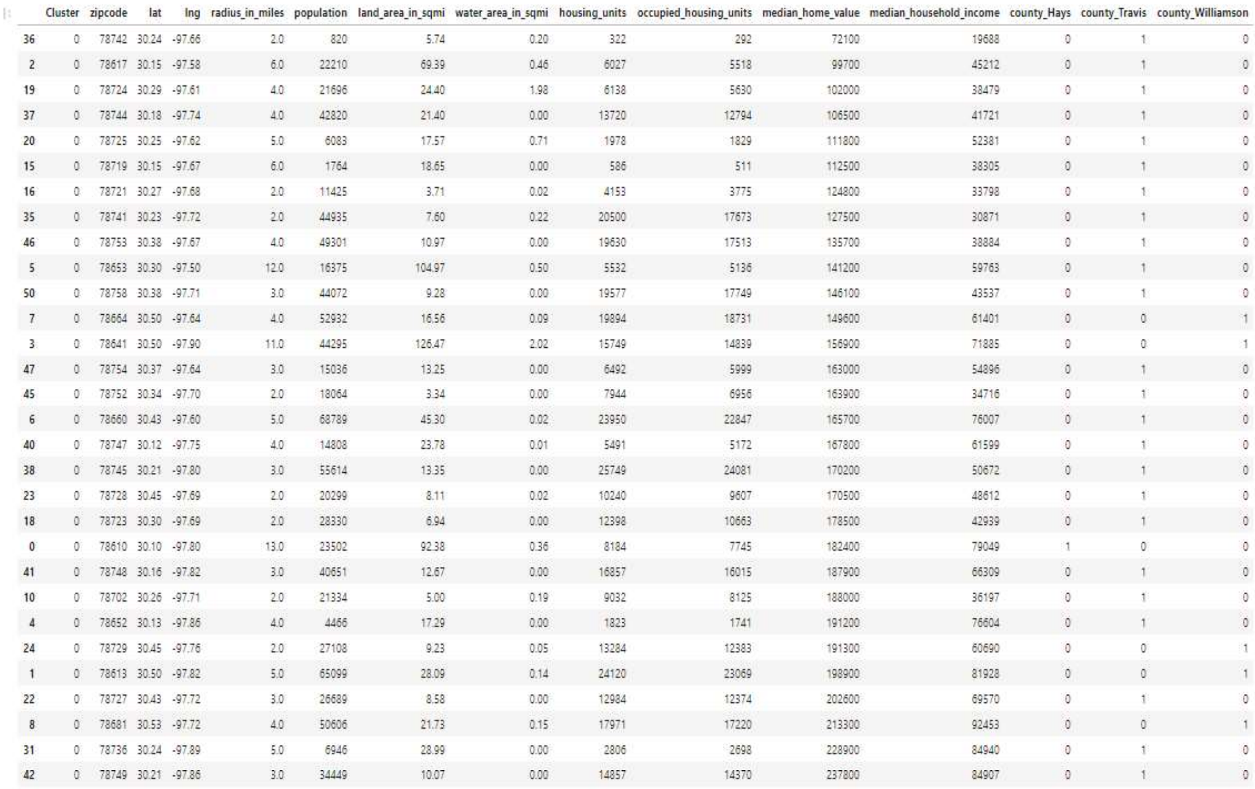
The analysis was to find price ranges for potential home buyers so they will have an idea if the Austin area is a potential place to call home.

- The uszipcode library was used to get details, such as counties, home prices and number of units, area population, and more.
- Several iterations of cleaning up the data were run to build the dataframe and the clustering methodology. Using k-means to analyze location type data sets provided three clusters.

The three clusters show:

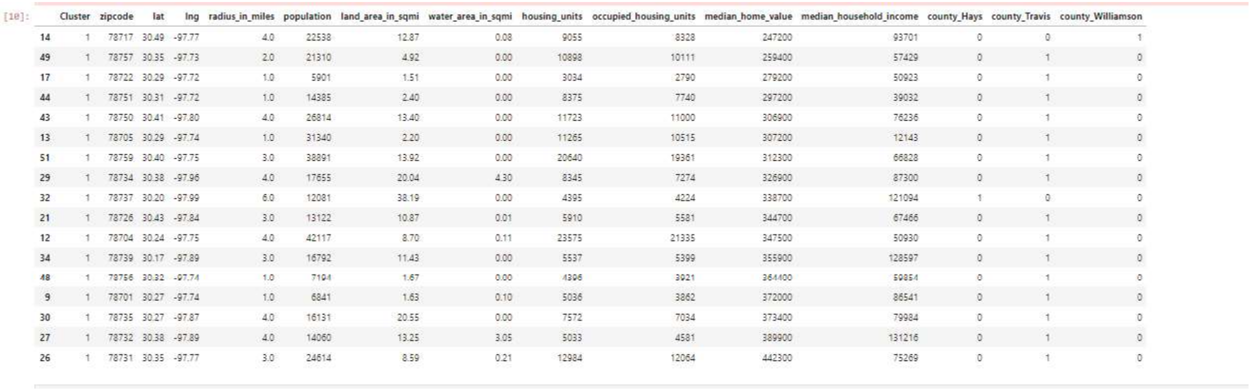
Cluster	Price Range	
	From	To
Cluster one	$ 72,100.00	$ 237,800.00
Cluster two	$ 247,200.00	$ 442,300.00
Cluster three	$ 460,200.00	$ 641,200.00

Cluster one output



	Cluster	zipcode	lat	lng	radius_in_miles	population	land_area_in_sqmi	water_area_in_sqmi	housing_units	occupied_housing_units	median_home_value	median_household_income	county_Hays	county_Travis	county_Williamson
36	0	78742	30.24	-97.66	2.0	820	5.74	0.20	322	292	72100	19688	0	1	0
2	0	78617	30.15	-97.58	6.0	22210	69.39	0.46	6027	5518	99700	45212	0	1	0
19	0	78724	30.29	-97.61	4.0	21696	24.40	1.98	6138	5630	102000	38479	0	1	0
37	0	78744	30.18	-97.74	4.0	42820	21.40	0.00	13720	12794	106500	41721	0	1	0
20	0	78725	30.25	-97.62	5.0	6083	17.57	0.71	1978	1829	111800	52381	0	1	0
15	0	78719	30.15	-97.67	6.0	1764	18.65	0.00	586	511	112500	38305	0	1	0
16	0	78721	30.27	-97.68	2.0	11425	3.71	0.02	4153	3775	124800	33798	0	1	0
35	0	78741	30.23	-97.72	2.0	44935	7.60	0.22	20500	17673	127500	30871	0	1	0
46	0	78753	30.38	-97.67	4.0	49301	10.97	0.00	19630	17513	135700	38884	0	1	0
5	0	78653	30.30	-97.50	12.0	16375	104.97	0.50	5532	5136	141200	59763	0	1	0
50	0	78758	30.38	-97.71	3.0	44072	9.28	0.00	19577	17749	146100	43537	0	1	0
7	0	78664	30.50	-97.64	4.0	52932	16.56	0.09	19894	18731	149600	61401	0	0	1
3	0	78641	30.50	-97.90	11.0	44295	126.47	2.02	15749	14839	156900	71885	0	0	1
47	0	78754	30.37	-97.64	3.0	15036	13.25	0.00	6492	5999	163000	54896	0	1	0
45	0	78752	30.34	-97.70	2.0	18064	3.34	0.00	7944	6956	163900	34716	0	1	0
6	0	78660	30.43	-97.60	5.0	68789	45.30	0.02	23950	22847	165700	76007	0	1	0
40	0	78747	30.12	-97.75	4.0	14808	23.78	0.01	5491	5172	167800	61599	0	1	0
38	0	78745	30.21	-97.80	3.0	55614	13.35	0.00	25749	24081	170200	50672	0	1	0
23	0	78728	30.45	-97.69	2.0	20299	8.11	0.02	10240	9607	170500	48612	0	1	0
18	0	78723	30.30	-97.69	2.0	28330	6.94	0.00	12398	10663	178500	42939	0	1	0
0	0	78610	30.10	-97.80	13.0	23502	92.38	0.36	8184	7745	182400	79049	1	0	0
41	0	78748	30.16	-97.82	3.0	40651	12.67	0.00	16857	16015	187900	66309	0	1	0
10	0	78702	30.26	-97.71	2.0	21334	5.00	0.19	9032	8125	188000	36197	0	1	0
4	0	78652	30.13	-97.86	4.0	4466	17.29	0.00	1823	1741	191200	76604	0	1	0
24	0	78729	30.45	-97.76	2.0	27108	9.23	0.05	13284	12383	191300	60690	0	0	1
1	0	78613	30.50	-97.82	5.0	65099	28.09	0.14	24120	23069	198900	81928	0	0	1
22	0	78727	30.43	-97.72	3.0	26689	8.58	0.00	12984	12374	202600	69570	0	1	0
8	0	78681	30.53	-97.72	4.0	50606	21.73	0.15	17971	17220	213300	92453	0	0	1
31	0	78736	30.24	-97.89	5.0	6946	28.99	0.00	2806	2698	228900	84940	0	1	0
42	0	78749	30.21	-97.86	3.0	34449	10.07	0.00	14857	14370	237800	84907	0	1	0

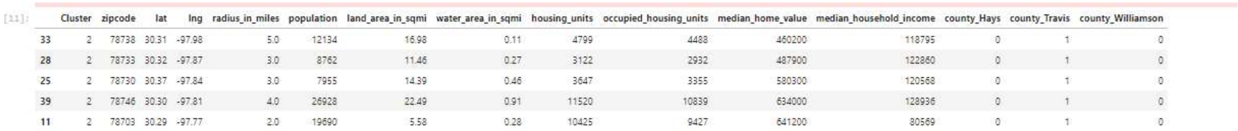
Cluster two output:



[10]:

	Cluster	zipcode	lat	lng	radius_in_miles	population	land_area_in_sqmi	water_area_in_sqmi	housing_units	occupied_housing_units	median_home_value	median_household_income	county_Hays	county_Travis	county_Williamson
14	1	78717	30.49	-97.77	4.0	22538	12.87	0.08	9055	8328	247200	93701	0	0	1
49	1	78757	30.35	-97.73	2.0	21310	4.92	0.00	10898	10111	259400	57429	0	1	0
17	1	78722	30.29	-97.72	1.0	5901	1.51	0.00	3034	2790	279200	50923	0	1	0
44	1	78751	30.31	-97.72	1.0	14385	2.40	0.00	8375	7740	297200	39032	0	1	0
43	1	78750	30.41	-97.80	4.0	26814	13.40	0.00	11723	11000	306900	76236	0	1	0
13	1	78705	30.29	-97.74	1.0	31340	2.20	0.00	11265	10515	307200	12143	0	1	0
51	1	78759	30.40	-97.75	3.0	38891	13.92	0.00	20640	19361	312300	66828	0	1	0
29	1	78734	30.38	-97.96	4.0	17655	20.04	4.30	8345	7274	326900	87300	0	1	0
32	1	78737	30.20	-97.99	6.0	12081	38.19	0.00	4395	4224	338700	121094	1	0	0
21	1	78726	30.43	-97.84	3.0	13122	10.87	0.01	5910	5581	344700	67466	0	1	0
12	1	78704	30.24	-97.75	4.0	42117	8.70	0.11	23575	21335	347500	50930	0	1	0
34	1	78739	30.17	-97.89	3.0	16792	11.43	0.00	5537	5399	355900	128597	0	1	0
48	1	78756	30.32	-97.74	1.0	7194	1.67	0.00	4396	3921	364400	59854	0	1	0
9	1	78701	30.27	-97.74	1.0	6841	1.63	0.10	5036	3862	372000	86541	0	1	0
30	1	78735	30.27	-97.87	4.0	16131	20.55	0.00	7572	7034	373400	79984	0	1	0
27	1	78732	30.38	-97.89	4.0	14060	13.25	3.05	5033	4581	389900	131216	0	1	0
26	1	78731	30.35	-97.77	3.0	24614	8.59	0.21	12984	12064	442300	75269	0	1	0

Cluster three output



[11]:

	Cluster	zipcode	lat	lng	radius_in_miles	population	land_area_in_sqmi	water_area_in_sqmi	housing_units	occupied_housing_units	median_home_value	median_household_income	county_Hays	county_Travis	county_Williamson
33	2	78738	30.31	-97.98	5.0	12134	16.98	0.11	4799	4488	460200	118795	0	1	0
28	2	78733	30.32	-97.87	3.0	8762	11.46	0.27	3122	2932	487900	122860	0	1	0
25	2	78730	30.37	-97.84	3.0	7955	14.39	0.46	3647	3355	580300	120568	0	1	0
39	2	78746	30.30	-97.81	4.0	26928	22.49	0.91	11520	10839	634000	128936	0	1	0
11	2	78703	30.29	-97.77	2.0	19690	5.58	0.28	10425	9427	641200	80569	0	1	0

Results and Conclusion

Based on the above analysis conclusions were able to be made.

- Austin and the surrounding areas, the three cities, Austin, Round Rock, Georgetown were the most populated.
- Austin has a very good pricing range and is a good choice for a home buyer.
  - This result was proven with the number of schools, food venues, home values and forecasting analysis and data.
- The projected forecast for the median home in Austin was clearly proved to be on the rise in the next year.
- The clustering and grouping of the data sets and methodologies noted led to some insight that is supported by the data itself.

It's a great spot to call home and work remotely!

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/b1a946bf-d3a8-4db5-a23d-809df9d7fc1b/view?access_token=72c5d0c93472fbe34b2c2c6e08641e76d2bddcceb9b3b4542e691913cd42dc32

The Capstone.Week.5.ipynb file was published in
the msanford2020/Coursera_Capstone repository. Click #ed1bd77 to view the commit.