Author: Benjamin Pace
Date: April 21, 2020

Venues as a Predictor of Home Values – Austin, TX:
Using Foursquare, Zillow and Other Data to Explore Home Values in Austin, TX

Introduction:

Austin is a city located in the state of Texas in the United States. It has an approximate population of 964,254 based on 2018 U.S. Census Bureau estimates and has experienced considerable population growth in the past ten years or so.[1] Housing prices in Austin have been some of the fastest growing in the county and, in a recent survey sponsored by Zillow (Jan. 2020), experts predict that this trend will continue.[2]

An interesting question to ask in this case is, are there any neighborhood features (like good schools, low crime, proximity to the city center, or a high density of favorable businesses) that might be correlated with higher housing values? Perhaps we can even design a model to determine approximately how much value (in dollars) features like this add to median housing prices based on real world data. Such a model might be of interest to home owners/buyers or residential real estate investors and could be used to determine how much of an impact changing neighborhood features have on housing values.

In this article I aggregate data containing information about Austin neighborhoods from different sources into a single "master" database. After this, I generate some interesting summary statistics and data visualizations, particularly maps, as a form of exploratory analysis. Finally, I develop a multiple regression model to measure the overall effect these features have on the median housing price and use these results to make some general conclusions about what features are more valuable. That being said, the ultimate goal of this short article is not to provide a full comprehensive statistical analysis of housing prices in Austin. Instead, the aim is to generate interest for the collected data (which contain information on neighborhoods in Austin) and provide a springboard framework for how a more comprehensive and rigorous analysis might be undertaken.

[1] Unites States Census Bureau – Austin city, Texas, available at https://www.census.gov/quickfacts/fact/table/austincitytexas/LND110210.

[2] Olsen, Skylar, "Looking for the Housing Markets Most Likely to Outperform in 2020? Look South," Zillow, January 1, 2020, available at https://www.zillow.com/research/2020-hot-markets-south-26293/; Widner, Cindy, "Austin's housing market expected to outperform national average, says study," Curbed Austin, January 14, 2020, available at https://austin.curbed.com/2020/1/14/21065547/austin-housing-market-real-estate-cost-affordability.

Data:

This article leverages data from several different sources and uses these data to determine the attributes of the various neighborhoods in the Austin area. Below I list each of the data sources used in this article and provide a short summary of each.

Zillow (zillow.com/research/data/): Zillow is an online real estate and rental marketplace and collects a substantial amount of data on housing and rental prices across the Unites States. Zillow also publishes a data set referred to as the "Zillow Home Value Index" which contains estimates on the median home value for 63 different neighborhoods in Austin, TX. For purposes of this article, Zillow's neighborhood level data forms the "base" dataset for the master database constructed.

Foursquare (https://foursquare.com/): Foursquare is a location technology platform that leverages user provided data and allows developers to access an expansive API of location data for use in various projects. In this case, Foursquare "venue" data is used to determine what types of business are located near the neighborhoods of interest.

SpotCrime (https://spotcrime.com/tx/austin/neighborhoods): SpotCrime is crime data aggregator which draws on data from police departments and news reports. SpotCrime contains crime statistics for 64 different neighborhoods in Austin which, to the extent possible, are merged to the Zillow data.

AreaVibes (https://www.areavibes.com/austin-tx/schools/): AreaVibes is a data aggregator and research tool that provides users with a great deal of information on various neighborhoods. This includes information on schools, crime statistics, weather, demographics and cost of living statistics. For purposes of this article, the information AreaVibes collects on school "proficiency scores" is used as a proxy for the quality of schools in the area. AreaVibes collects the information it reports on Austin schools from the 2016 United States Census Bureau American Community Survey (ACS) and from the National Center for Education Statistics.

Other Sources: In addition to the sources listed above, some data like approximate latitude and longitude coordinates are collected from Google Maps and other data is leveraged from specific python packages. To the extent any data is used beyond the four main sources listed above, this is noted in the relevant section of the report.

Methodology:

An important first step when working with new data is to simply review the contents of each data set and start developing an outline for any data merging and exploratory analysis that should be done prior to core analysis (in this case a multiple regression model). To kick off this process, let's take a look at the first 5 rows of each data set. In order to fit the Zillow data onto a single page the preview of the data below only shows a single home value column "2019-01", the full data set contains median home value columns ranging from "1996-04" to "2020-02".

Table 1 – Zillow Data, First 5 Rows

RegionID	RegionName	City	State	Metro	CountyName	SizeRank	2019-01
274772	Northeast Dallas	Dallas	TX	Dallas-Fort Worth-Arlington	Dallas County	1	323484
112345	Maryvale	Phoenix	AZ	Phoenix-Mesa-Scottsdale	Maricopa County	2	178156
192689	Paradise	Las Vegas	NV	Las Vegas-Henderson-Paradise	Clark County	3	267194
270958	Upper West Side	New York	NY	New York-Newark-Jersey City	New York County	4	1279195
118208	South Los Angeles	Los Angeles	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	5	496902

Table 2 – Spot Crime, First 5 Rows

state	city	date	period	neighborhood	total_count
TX	Austin	4/2/2020	last six months	Allandale	99
TX	Austin	4/2/2020	last six months	Barton Hills	73
TX	Austin	4/2/2020	last six months	Bouldin	256
TX	Austin	4/2/2020	last six months	Brentwood	86
TX	Austin	4/2/2020	last six months	Central East Austin	362

Table 3 – AreaVibes School Scores, First 5 Rows

state	city	date	school_name	school_address	district	score
TX	Austin	4/4/2020	Canyon Creek Elementary School	10210 Ember Glen Dr	Round Rock ISD	0.99
TX	Austin	4/4/2020	Forest Trail Elementary School	1203 Loop 360 S	Eanes ISD	0.98
TX	Austin	4/4/2020	Lake Pointe Elementary School	3322 Ranch Rd 620 S	Lake Travis ISD	0.98
TX	Austin	4/4/2020	Laurel Mountain Elementary School	10111 Dk Ranch Rd	Round Rock ISD	0.97
TX	Austin	4/4/2020	Casis Elementary School	2710 Exposition Blvd	Austin ISD	0.97

Right off the bat you can see that we will have to come up with some way to merge these three data sets. Merging the Zillow data and SpotCrime data seems to be the most straightforward, as they both contain "neighborhood" fields. To do this, I've gone ahead and created a bridge based primarily on neighborhood name and, when names could not be matched, used Google searches to approximate matches between neighborhoods. Below are the first 5 rows of the bridge I've made. This will be helpful for merging the two data sets.

Table 4 – Zillow-SpotCrime Bridge, First 5 Rows

zillow_neighborhood	spotcrime_neighborhood
Allandale	Allandale
Barton Creek	n/a
Barton Hills	Barton Hills
Bouldin Creek	Bouldin

A closer review of this bridge shows that 58 out 63 of the Zillow neighborhoods can be matched to the SpotCrime neighborhoods, and notably most can be matched just on name alone.

Merging the AreaVibes school proficiency scores to each neighborhood is going to take a little more creativity. The approach I've chosen in this case is to first geocode each neighborhood and similarly geocode the address for each school.[3] Based on the assigned locations (latitude and longitude coordinates) I am then able to calculate the straight-line distances in miles between each of the neighborhoods and the schools.[4] With this information we can then determine which neighborhood each school is closest to and then, based on this assignment, determine an average school score for each neighborhood.

It is worth mentioning that due to the way school districts draw their borders, there may be some cases where say a house in a neighborhood just on the edge of a school district cannot technically attend the closest available school. Meaning there may be some mismatch between schools and houses in neighborhoods. In this case however, because I am reviewing these data at the neighborhood level, I assume that distance is an acceptable approximation for whether a school score can be assigned to a specific neighborhood.

Another data feature that might be helpful to have for purposes of the regression model is how far each neighborhood is from the city center in miles. This is straightforward to calculate using the same method that was used to calculate distances between the schools and neighborhoods. In this case 30.274630, -97.740365 is used as the location of the city center which corresponds to the capitol building in Austin, TX.

The resulting merged data set now looks like the following, which (for each neighborhood) includes a latitude/longitude, a 2019 average median home price, an average school score, the

[3] This was done using the Nominatim function from the geopy package. In certain cases, a few manual adjustments were made to auto generated data. Specifically, 4 neighborhoods were not geocoded and 2 neighborhoods "Riverside" and "Southeast" appear to have been geocoded incorrectly. Generally, I assume geopy was able to geocode the neighborhoods correctly. Although geocoding data is always tricky, and in a more extensive analysis it would be good to review each location in detail. Geopy – Nominatim, available at https://geopy.readthedocs.io/en/stable/#nominatim.

[4] This was done using the vincenty function also from the geopy package. Geopy – Calculating Distance, available at https://geopy.readthedocs.io/en/stable/#module-geopy.distance.

distances in miles from the city center and the count of crimes which occurred in the last six months for each neighborhood.

Table 5 – Zillow Data Merged Version 1

	region_id	region_nam	city	state	lat	lon	avg_hmval_2019	crime_count	avg_scl_score	center_miles
0	274685	North Austin	Austin	TX	30.4023	-97.7260	313327.5000	418.0000	0.8362	8.8382
1	271391	Franklin Park	Austin	TX	30.1969	-97.7488	224188.2500	255.0000	0.6445	5.3781
2	275057	Pleasant Valley	Austin	TX	30.2310	-97.7158	241565.9167	327.0000	0.5000	3.3454
3	271652	Windsor Park	Austin	TX	30.3135	-97.6911	388482.8333	167.0000	0.5750	3.9826
4	271635	West University	Austin	TX	30.2896	-97.7459	306491.5833	1132.0000	0.8700	1.0795

The data set above is close to the final dataset I ultimately use in the multiple linear regression model; however, the Foursquare data still needs to be merged into the dataset. The Foursquare API has a vast amount of data available to users, but in this case specifically I want to get a listing of which venues like restaurants, coffee shops, and grocery stores are within 5 miles of the chosen center of the neighborhood. The use of a 5-mile radius is somewhat arbitrary, and ideally each radius used would be adjusted to fit the contours of each neighborhood. This is beyond the scope of this article, however, and so in this case I assume a listing of all the venues within a 5-mile radius is sufficient for the model. After working through some data formatting, the data extracted from the Foursquare API looks like the following.

Table 6 – Raw Foursquare Venue Data, First 6 Rows

region_nam	region_lat	region_lon	venue	venue_lat	venue_lon	fq_venue_category
North Austin	30.4023299	-97.7260086	The Domain	30.4006016	-97.7251211	Shopping Plaza
North Austin	30.4023299	-97.7260086	North Italia	30.40233393	-97.72587453	Italian Restaurant
North Austin	30.4023299	-97.7260086	Bakery Lorraine	30.40164809	-97.72268611	Bakery
North Austin	30.4023299	-97.7260086	Starbucks Reserve Bar	30.40262588	-97.72135154	Coffee Shop
North Austin	30.4023299	-97.7260086	Tiffany & Co.	30.403335	-97.724586	Jewelry Store
North Austin	30.4023299	-97.7260086	Sway	30.40067792	-97.72285287	Thai Restaurant

As can be seen in the table above, what has been generated is essentially a list of all of the venues within a 5-mile radius of each neighborhood and a column “region_nam” which gives the neighborhood each venue is assigned to. Looking at the column “Venue Category” we can see that the default Foursquare categories are somewhat erratic. This feature is probably helpful for a user, but for the purposes of the regression model its unlikely that distinguishing between an Italian restaurant and a Thai restaurant will be helpful. As a result, I’ve also made some manual adjustments to the Foursquare output by categorizing each venue into 11 categories. Specifically, I’ve chosen shop/mall, restaurant, bakery/dessert, coffee, grocery store, fitness, hotel, liquor/bar, park/outdoors, entertainment and other as the 11 categories to merge back in to the Zillow data. Other is used as a catch-all category and will not ultimately be used in the regression model due to the varied nature of the venues allocated to this category.

The choice of these 11 categories is again somewhat arbitrary, and a decent amount of additional work could go into simply determining the best types of categorizations for each venue. Given the limited scope of this article I assume that the 11 categories are sufficient for an initial review of the data and preliminary regression model.

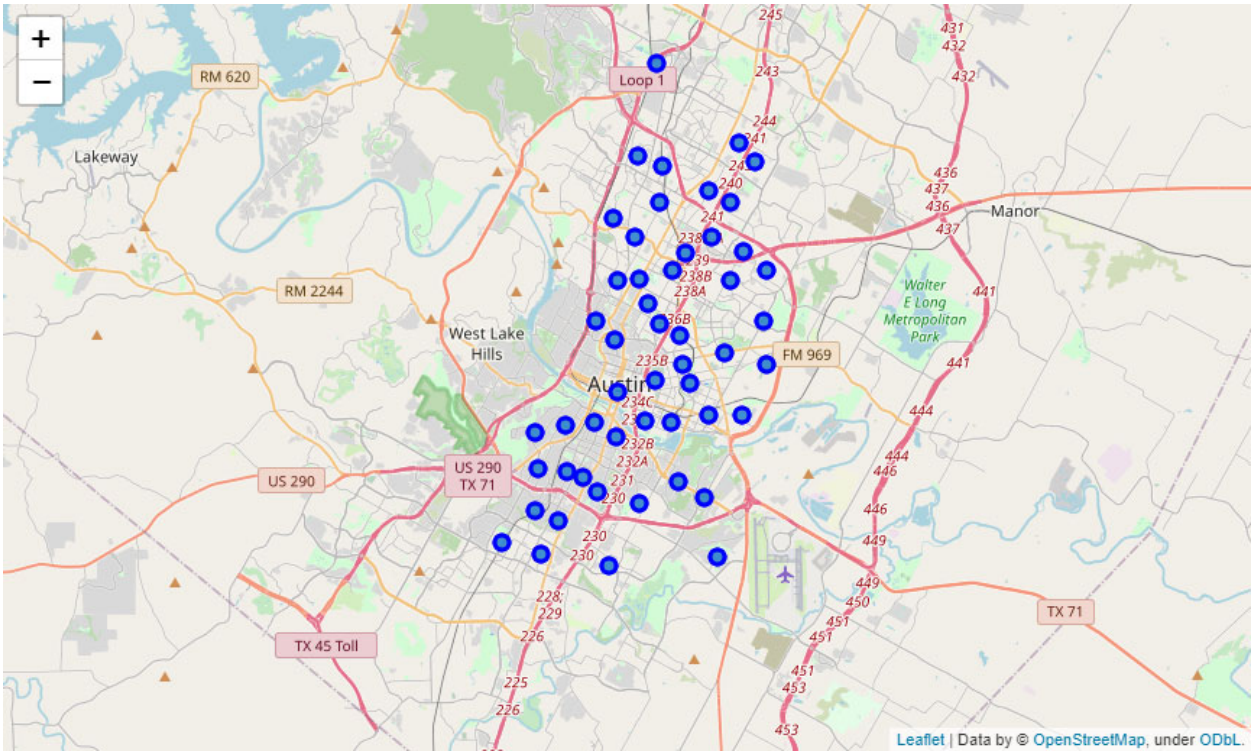
After categorizing each venue, I summarized the count of venues by venue category and neighborhood and then merged these counts back to the Zillow data. Once this is done, any neighborhoods with missing "crime_count" or "avg_scl_score" data are dropped since these will not be used for the analysis.[5] The final "master" database has 20 columns and is too large to display a preview of. The full dataset along with all the code used in the methodology section can be viewed in my Jupyter Notebook which can be accessed at the link provided at the end of the article.

Results – Part 1 Exploratory Analysis:

At this point, it is a good idea to start taking a closer look at what the master dataset looks like. From my perspective the place to start when working with any kind of location-based data is to simply map out each of the locations of interest. This helps to get a sense for the scope of the data and in this case specifically will be a good determination of whether some neighborhoods need to be dropped. The thinking behind potentially dropping neighborhoods, is that neighborhoods that are too close geographically will not exhibit enough variation in their features to be statistically useful. Having variation in the underlying data is important for most statistical models and this is especially true of regression models.

[5] An alternative approach to dropping these entries may be to impute data based on the crime counts and scores of nearby neighborhoods. The way to go about this might be to work through the exploratory analysis and regression model using the current approach and then repeating this with the imputed data. To the extent results varied it would be worth thinking about why and if the imputed data is in fact worth relying on.

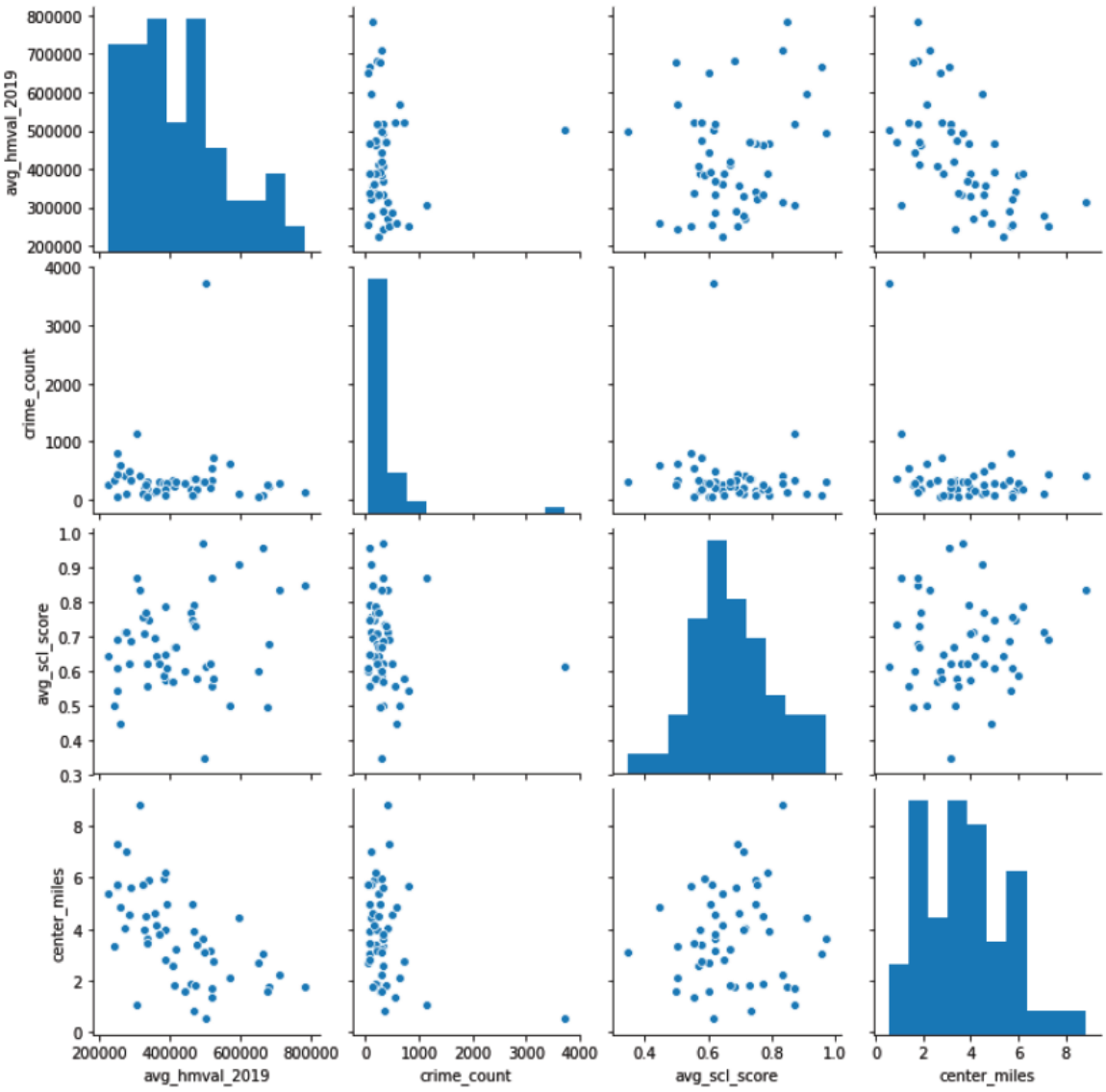
Figure 1 – Map of Austin, TX Zillow Neighborhoods



After reviewing the map above, it seems that the neighborhoods are fairly well spaced and that the greater Austin area is covered fairly well geographically. The next step in this case is to get a better understanding of how the neighborhood features are distributed.

The seaborn python package has an especially helpful function named pairplot which generates a matrix of scatterplots between variables and a histogram of the distribution of each variable. As an initial review, it makes sense to focus on the variables derived from the non-Foursquare sources.

Figure 2 – Pairplot of Non-Foursquare Variables Version 1

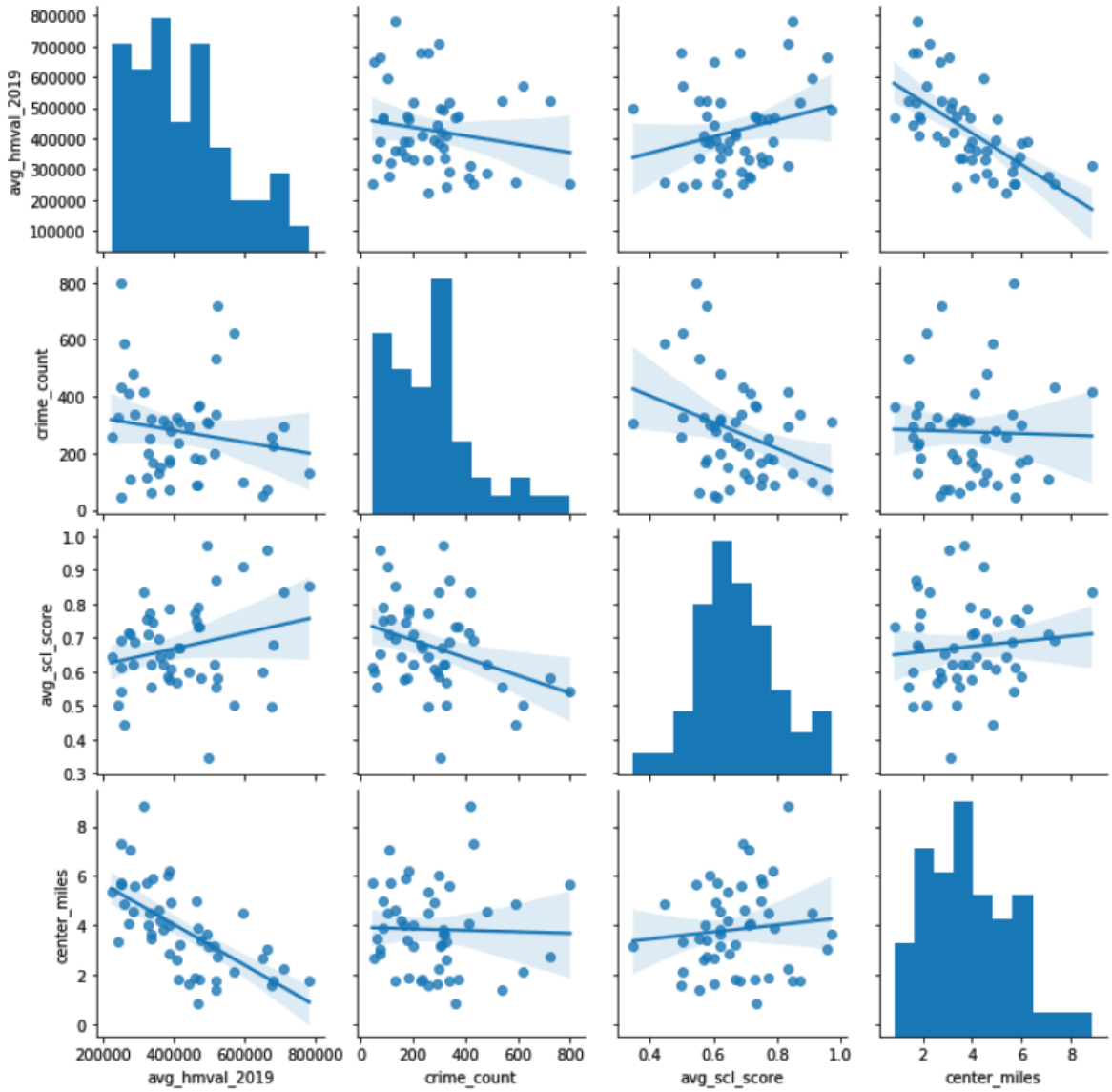


The cross-section of these variables with the “crime_count” variable is interesting. In most cases we can at least an outline of linear trend between the variables, for examples we see a clear linear downward trend between “center_miles” and “avg_hmval_2019” meaning homes tend to be cheaper the further away they are from the city center which makes sense. In contrast, looking at the scatter plots with “crime_count” we see a cluster of points either on the bottom or far left of the graph and a single point on the opposite end of this cluster. This notion is confirmed by looking at the histogram for “crime_count”, which similarly shows a single observation far outside the rest of the data. I’ve identified this specific neighborhood as “Downtown” which has relatively high crime count of 3,726 for the past six months. While not

as visible in the pairplot, there is another neighborhood with a relatively high “crime_count” “West University” with a crime count of 1,132 for the past six months.

Observation like these two can certainly throw off the results of the model and in this case, I’ve assumed that it is prudent to drop these observations. After removing these observations from the data set, I generate a new pairplot and add an additional feature to the plots which shows a simple linear regression line for each cross-section.

Figure 3 – Pairplot of Non-Foursquare Variables Version 2



The “crime_count” scatter plots looks much better now. While there do appear to be a few points on the extreme ends of the spectrum, the data look a lot closer to what we would expect in a typical scatter plot.

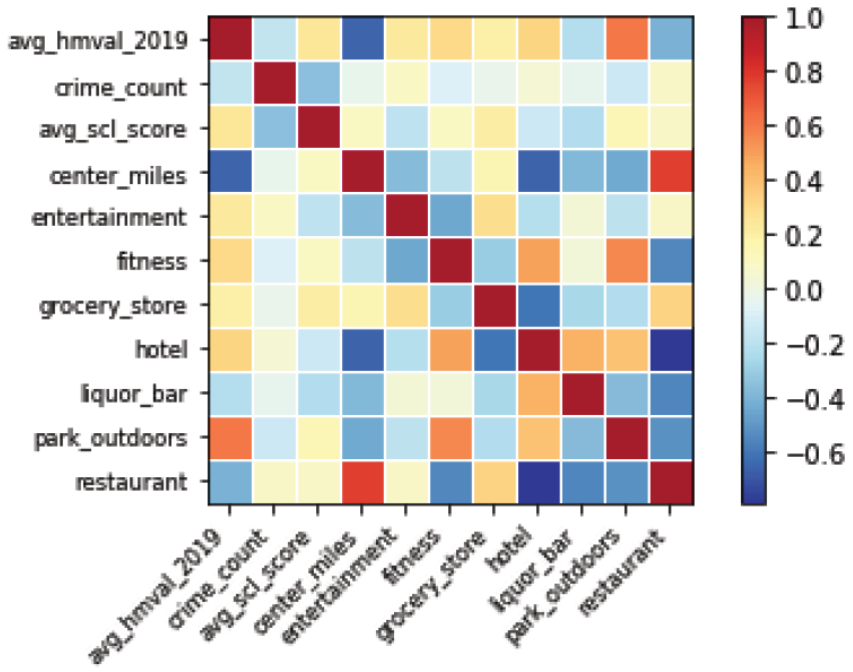
Moving on to the what I've called the Foursquare variables, a good first step in this case is to also start with a pairplot. The particular focus of these plots will be to determine which variables might not have any relationship to the dependent variable "avg_hmval_2019". The pairwise plot is too large to paste into this article, however, a full version is available in the Jupyter Notebook which can be accessed at the link I've provided at the end of the article. The scatterplots show that "bakery_dessert", "coffee" and "shop_mall" have pretty weak relationships with the dependent variable "avg_hmval_2019". Given that this still leaves us with 7 other variables to use for purposes of the regression I assume that dropping these variables is ok for purpose of this article.

That being said, a more rigorous analysis may still decide to keep these variables at least in the first pass of the regression model as there may be some interesting interactions with other variables that we cannot yet see just by looking at the pairplot.

With "bakery_dessert", "coffee" and "shop_mall" dropped from the dataset it is still prudent to check how the remaining variables interact with each other. In this case, I'm primarily attempting to anticipate any issues that could arise from multicollinearity. Multicollinearity occurs when predictor variables are highly correlated, and if present in a model can result in inaccurate regression coefficients.[6] To start this process I've generated a correlation matrix which simply visually represents the correlation between all variables I intend to use in the model. Right off the bat we can see "hotel", "restaurant" and "center_miles" are fairly highly correlated. This makes sense as we would expect most hotels and restaurants to be located near the city center. Unfortunately, this is exactly the kind of predictor correlation that could cause issues in the model.

[6] For a more complete discussion of these issues see the following web page. PennState Eberly College of Science – 10.6 Highly Correlated Predictors, available at https://online.stat.psu.edu/stat462/node/179/.

Figure 4 – Correlation Matrix



Prior to dropping these variables, however, I calculate the Variance Inflation Factors (VIF). The VIF is a more formal approach to identifying multicollinearity among predictor variables. Specifically, the VIF tells us how inflated the variance of the coefficient will be for a given predictor due to substantial correlation with another predictor.[7]

Table 7 – VIF and Updated VIF

	variables	VIF
0	crime_count	1.2851
1	avg_scl_score	1.3572
2	center_miles	7.8698
3	entertainment	2.7630
4	fitness	2.5192
5	grocery_store	2.4495
6	hotel	5.9685
7	liquor_bar	10.9917
8	park_outdoors	10.9596
9	restaurant	17.7936
10	intercept	1353.4742

	variables	VIF
0	crime_count	1.1815
1	avg_scl_score	1.3297
2	center_miles	3.8156
3	entertainment	2.3294
4	fitness	1.6876
5	grocery_store	1.9628
6	hotel	5.4719
7	liquor_bar	1.4044
8	intercept	219.7077

[7] For a more complete discussion of these issues see the following web page. PennState Eberly College of Science – 10.7 - Detecting Multicollinearity Using Variance Inflation Factors, available at https://online.stat.psu.edu/stat462/node/180/.

Looking at the initial VIFs calculated (left side) we can already identify issues with “restaurant”, and “park_outdoors”. Issues in this case refers to VIFs over 10. There are some academic sources (like the PennState webpage previously cited to) which indicate that even a VIF over 5 might be problematic. Given this is not a rigorous statistical study though, I assume VIFs under 10 are ok. After dropping these two variables the updated VIFs calculated (right side) are much better.

One question to ask here might be, why was “liquor_bar” not dropped even with a higher VIF than “park_outdoors”? That is a fair question to ask, and a bit of trial and error was involved here. From the perspective of the model “liquor_bar” is an interesting predictor to include since it is the only other predictor aside from “center_miles” we would expect to be negatively correlated with home value. Therefore, I calculated VIFs only dropping “park_outdoors” and “restaurant” first to see if this would resolve the issue and as shown in the output above it does. I would have liked to have kept “park_outdoors” and “restaurant” as the two have intuitive appeal in a model. However, as shown by the preliminary analysis keeping either of these variables would likely cause statistical issues in the model. A more rigorous analysis might explore various adjustments that could be made either to the data or the model in order to keeps these variables.

Results – Part 2 Regression Model:

At this point we are now ready to run the regression model and, after the model output is generated, run a few diagnostics to check the quality of the model.

Table 8 – Regression Results

OLS Regression Results

Dep. Variable:	avg_hmval_2019	R-squared:	0.814
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	21.84
Date:	Tue, 21 Apr 2020	Prob (F-statistic):	2.50e-12
Time:	10:12:05	Log-Likelihood:	-607.66
No. Observations:	49	AIC:	1233.
Df Residuals:	40	BIC:	1250.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.231e+05	1.38e+05	3.794	0.000	2.44e+05	8.02e+05
crime_count	-134.0896	59.441	-2.256	0.030	-254.224	-13.956
avg_scl_score	1.226e+05	8.42e+04	1.457	0.153	-4.75e+04	2.93e+05
center_miles	-5.204e+04	1.05e+04	-4.952	0.000	-7.33e+04	-3.08e+04
entertainment	5845.3754	7032.899	0.831	0.411	-8368.644	2.01e+04
fitness	1.235e+04	5910.507	2.090	0.043	406.185	2.43e+04
grocery_store	2.01e+04	5769.925	3.483	0.001	8437.185	3.18e+04
hotel	1.15e+04	7131.120	1.613	0.115	-2911.928	2.59e+04
liquor_bar	-1.423e+04	2400.537	-5.929	0.000	-1.91e+04	-9382.243

Omnibus:	2.335	Durbin-Watson:	1.486
Prob(Omnibus):	0.311	Jarque-Bera (JB):	1.528
Skew:	0.181	Prob(JB):	0.466
Kurtosis:	2.215	Cond. No.	5.15e+03

The table above shows the output generated after calculating the regression model using the Ordinary Least Squares (OLS) method.[8] First looking at the adjusted R-squared we see a value of .776, which is decent given that the model is really just a preliminary version. Moving over to the Prob (F-statistic) we see a very small value meaning that the model overall is statistically significant, another good sign. Located under this output we see the coefficient estimates, and all appear to be directionally what we would expect. In other words, higher crime counts, being further away from the city center and a higher amount of liquor stores/bars all result in lower home values. On the other hand, amenities like grocery stores and fitness centers increase home values.

[8] For a more in depth discussion of Ordinary Least Squares see the following e-book which is available for free. Caffo, Brian, Regression Models for Data Science in R, LeanPub, 2019, pp. 17-21, available at https://leanpub.com/regmods.

It is important to note that not all of these predictors are statistically significant and in fact there are a few that are statistically the same value as zero. This does not mean the model is useless. The variables “crime_count”, “center_miles”, “grocery_store”, “fitness” and “liquor_bar” are all statistically significant at the 5% level. In my view, this model may be too preliminary to ask questions like “If a bar opens up near a neighborhood in Austin, by how much will the median home value go down?”. Though a similar model that has gone through more rigorous stress testing probably could. The current model, however, can be useful to answer higher level questions like “Does a fair amount of liquor stores and bars in a neighborhood impact the median home value?”. As can be seen from the results, the answer is yes, a higher density of bars and liquor stores in a neighborhood can potentially drag down home values. Additionally, for someone unfamiliar with the landscape of the Austin residential real estate market this model could potentially be helpful to gain insight into what features home owners and buyers tend to value in neighborhoods.

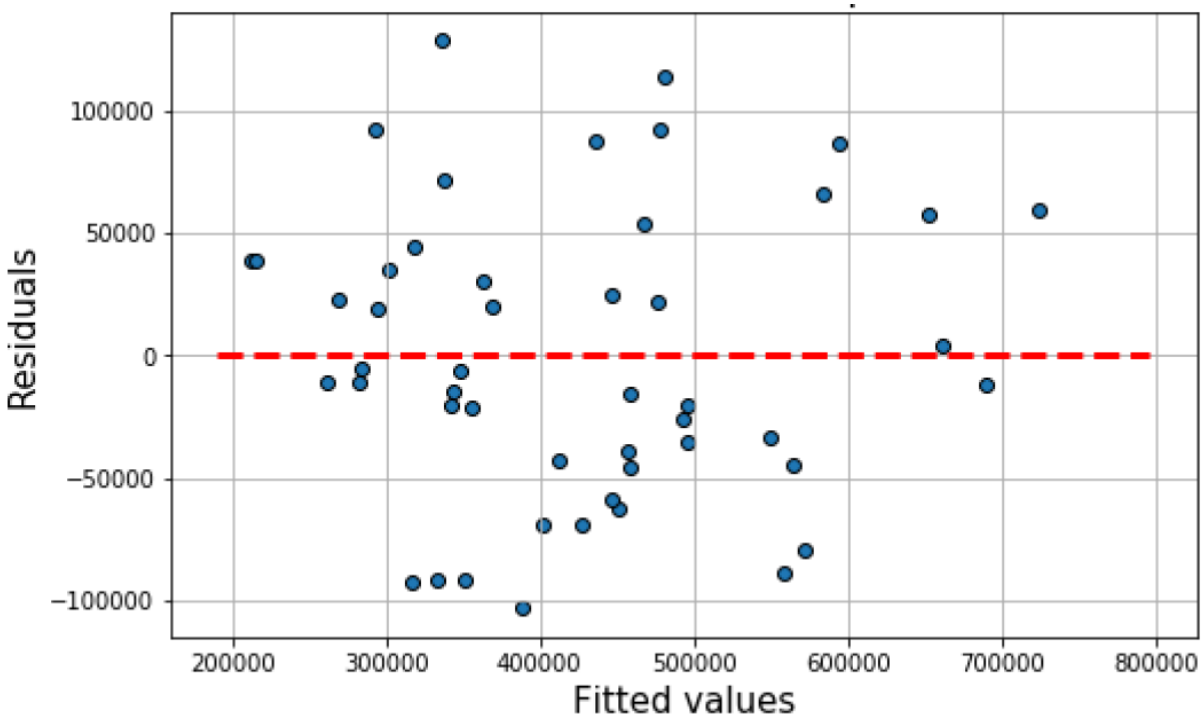
Prior to drawing any firm conclusions though, some diagnostics of the regression model do need to be run primarily to ensure that the core underlying assumptions used by the OLS method are not violated. A more rigorous statistical analysis might generate several different regression models and based on various tests determine which is the most appropriate/interesting to report as the primary model. Generating several models with different variables and assumptions could also act as a form of stress-testing.

While there are a decent amount of regression diagnostics I could run on this model, I focus on four different tests for simplicity and all seem to generate decent results.

The first diagnostic is simply plotting the residuals against the fitted values. As shown in the figure below the residuals are distributed somewhat randomly which is a good indication that the linear assumptions of the model hold. As mentioned in a webpage published by PennState “[the fitted v. residual] plot is used to detect non-linearity, unequal error variances, and outliers.”[9]

[9] PennState Eberly College of Science – 4.2 - Residuals vs. Fits Plot, available at https://online.stat.psu.edu/stat462/node/117/.

Figure 8 – Fitted v. Residuals Plot



Next, I run a Breusch-Pagan test to detect for heteroskedasticity. The previous fitted v. residuals plot also checks for heteroskedasticity, but the benefit of the Breusch-Pagan test is that it defines a formal quantitative cut-off in the form of a p-value. Heteroskedasticity is important to check for when using the OLS method since a model with heteroskedasticity present will have less precise coefficient estimates.[10] Additionally, the p-value for each coefficient could be incorrect in the presence of heteroskedasticity due to the range of the coefficient being less precise. Similar to the fitted v. residuals plot, the Breusch-Pagan test calculates a p-value of .28 meaning no heteroskedasticity is detected. Unlike the other test there is no visual component, but the details of how I calculated this can be seen in the Jupyter Notebook which I provide a link for at the end of the article.

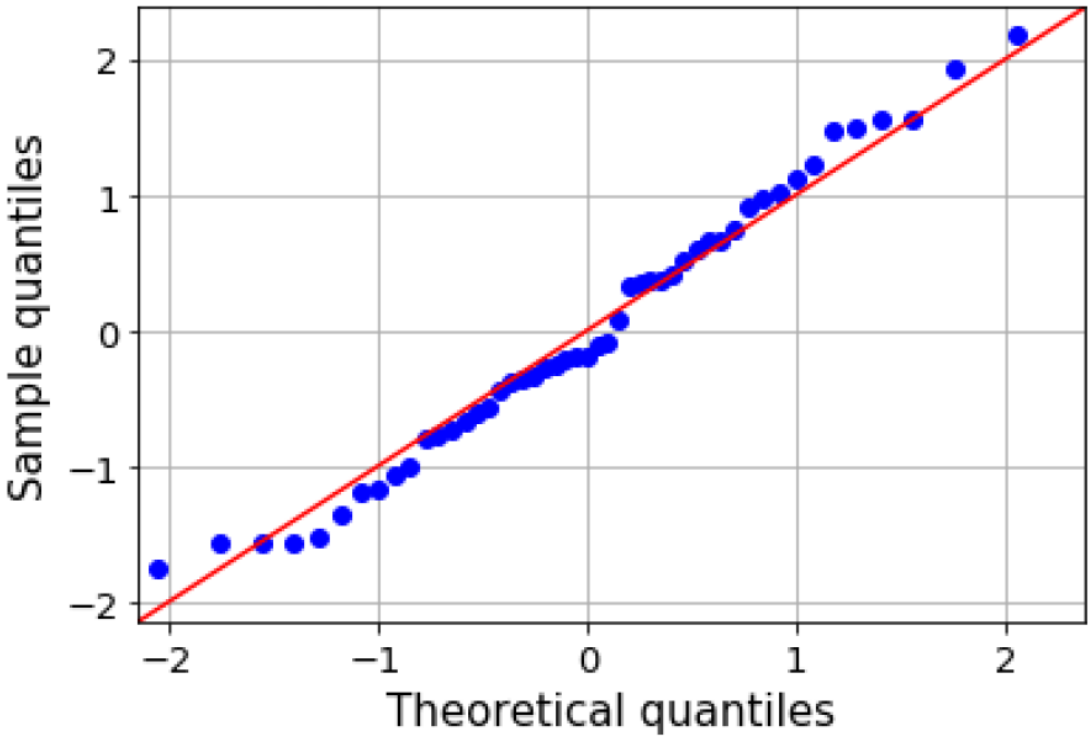
The next test, a QQ-Plot, can be used to determine whether the residuals are normally distributed. The ideal outcome for a QQ-Plot is to see that all points are centered along the 45-degree diagonal line included in the graph. If the QQ-Plot yields unfavorable results, meaning points not centered around the 45-degree line, this may indicate that some or all of the summary statistics such p-values and confidence intervals are unreliable.[11] A common reason

[10] For a good review of heteroskedasticity and the issues associated with it in regression models see the following summary from the University of Notre Dame. Williams, Richard, "Heteroskedasticity," University of Notre Dame, January 10, 2020, available at https://www3.nd.edu/~rwilliam/stats2/l25.pdf.

[11] For a more detailed discussion of QQ-plots and their use see the following webpages. Carnegie Mellon University - Regression Diagnostic Plots, available at

for an unfavorable QQ-plot would be that there is some unknown variable we have not accounted for, in which case the statistics we are estimating would not be as reliable. However, the figure below is favorable and generally indicates that this is not a concern in the current model I have estimated.

Figure 9 – QQ-Plot



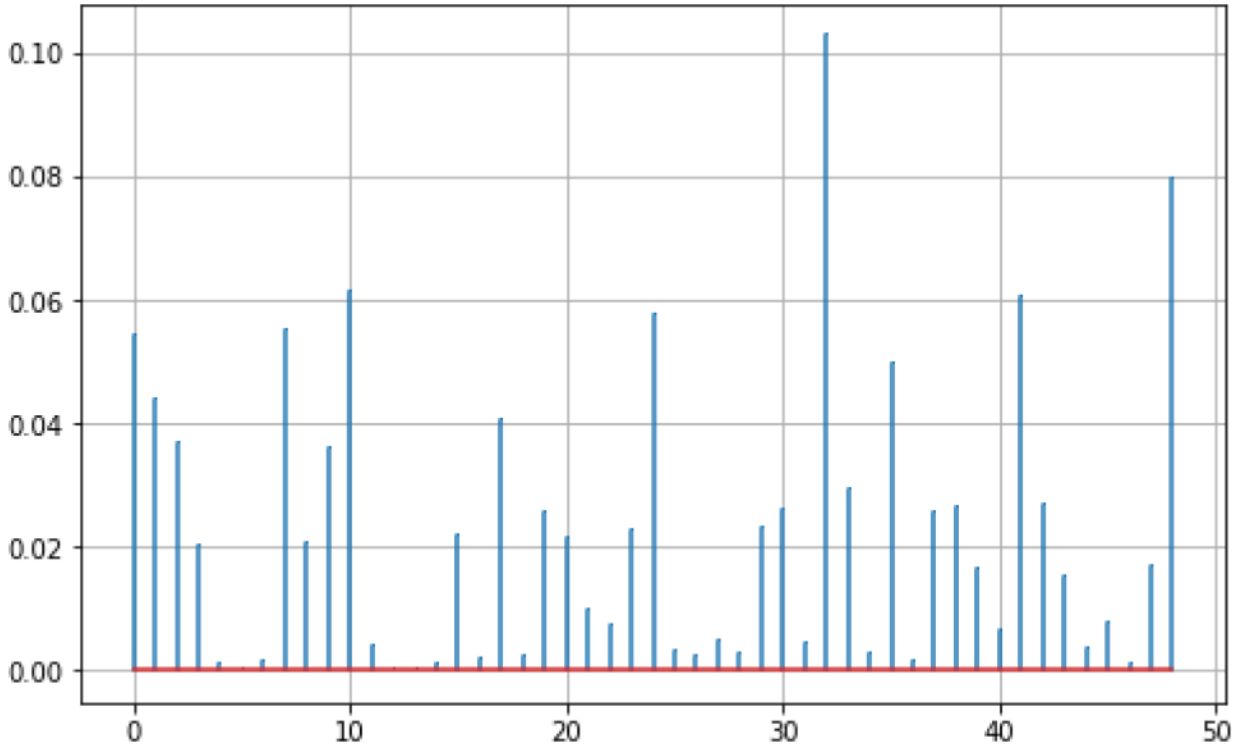
The last diagnostic test I run is a Cook's distance plot, in this case I assume if any of the data points calculated in this graph were over 1 that would indicate some outlier is having a significant effect on the results of our model. It is worth mentioning, that depending on preference, some academic sources indicate that any observations which have a Cook's distance of 4 / (Obs. – Var. – 1) might be of concern.[12] Looking at the graph below, technically there is one point which would be above the more conservative threshold, which in my case is 4 / (49 – 8 – 1) = .1. However, the point is just barely above this threshold and as I keep mentioning the goal here is not an iron clad statistical analysis but more of an initial review. That being said, if I were to leave in the Downtown and West University neighborhoods in the

http://www.contrib.andrew.cmu.edu/~achoulde/94842/homework/regression_diagnostics.html; National Institute of Standards and Technology U.S. Department of Commerce – Quantile-Quantile Plot, available at https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm.

[12] For a discussion on Cook's distance see the following webpage and book. R-Project – Measures of Influence, available at https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html; Bruce, Peter and Bruce, Andrew, Practical Statistics for Data Scientists, O'Reilly, 2017, pdf pg. 270.

model, we would clearly be able to identify these points in the graph below. The issue with leaving a clear outlier in the underlying data of a regression model is that this could bias our results in unexpected ways. Depending on the circumstances arguments can be made for leaving outliers in the model, but for simplicity sake we assume they can be removed in this case.

Figure 10 – Cook's Distance



Discussion and Conclusion:

Despite the regression model being preliminary in nature, the model and the accompanying regression diagnostics show that there are some clear neighborhood features that can impact median home prices in Austin. As noted before, the overall goal of this article was to showcase some of the interesting data available on home values in Austin and also show how we can combine these data across sources. The most interesting result of the model in my view is how significant of an effect the model calculates that liquor stores and bars have on median home prices. I did initially expect some effect, but after working through the analysis this assumption is now more concrete and defined.

This is outside the scope of this article, but it would be interesting to see what the coefficients of this model would look like run with a log-log specification. In fact, results like this may be more intuitive and easier to explain as the coefficients could be interpreted in percentage

terms. In other words, I think it would be interesting to be able to make statements along the lines of “an x% increase in the number of x venues leads to an x% increase/decrease in the median home value of a given neighborhood in Austin, TX.”

The model in its current form, however, does provide some interesting results in the sense that it helps to develop intuition around what venues and neighborhood features matter to home owners and buyers in the Austin area. A more rigorous statistical analysis would likely gather additional data and experiment with more robust model specifications, but hopefully this article has achieved its intended purpose and given a good overview of what kind of data could be used in such a study and provided at least an initial framework for future analysis.