Coursera Capstone Final Report

8/1/2020

Abstract/Executive Summary

In this report, we examine a model for the price of homes in Cook County, IL. In this analysis we examine venue information, restaurants, shops, and attractions attractions using the Foursquare API, and cluster like neighborhoods regarding the amenities that they offer. We develope a data aggregation procedure then consider a standard linear model and a mixed effects model to examine the variability in housing price based on these clusters and features of housing at the zip-code level. The two approachs provide insight about the nature of this clustering. We ultimately determine, that while there is an effect in home prices due to clustering, it may be depended on other underlying factors.

Introduction

Understanding of the behavior of real estate within a metropolitan area is of interest to many stakeholders, i.e. individuals looking for housing, real-estate brokers/relators, businesses looking to relocate to a particular area, investment companies, mortgage lenders/banks, among many more. Many companies, such a Zillow and RedFin, have become a household name for their work in aggregating and predicting housing costs using machine learning techniques. This problem lends its self very naturally to machine learning techniques due to the large number of predictors or features, and high volume of data. These companies also have produced pipelines to collect and analyze data, essentially at the population level. In this report, we shall entertain a different modeling technique that relies on inference, instead of prediction using population level data.

Chicago, IL is the third most populous city in the United States, and the most populous city in the American Midwest. The city is an international hub for finance, industry, education, technology etc... Moreover, it is well known for its diverse cultural scene and various entertainment, dining, artistic, musical activities and venues. Naturally, it is beneficial for the stakeholders, mentioned previously to have a clear understanding of the real estate market in such a large metropolitan area, and understand if and the nature these venues have on real estate pricing.

For those who have taken the Coursera IBM Data Science Capstone course, we observed in the segmentation and clustering of neighborhoods in Toronto, neighborhoods in different geographic locations may cluster together in the venue and amenities that they offer. Regarding real estate valuation, a natural question is to determine if the clusters of neighborhood venues are effect housing prices or cause correlation among housing prices in a cluster. For example, if we have a group of neighborhoods that clusters around a high number of coffee shops and another neighborhood that clusters around a high number of up-scale boutiques, do we expect to see the mean housing price differ between these clusters, or will we see different variability in housing prices in these different clusters? Alternatively, we may consider the clusters as a group effect taken as a factor in a linear regression model. Equivalently, it is not unreasonable to expect that there will exist some degree of correlation between housing prices within each cluster, so a model ought to entertain this possibility. Assessing both the differences in mean prices and differences of price variation within clusters will provide important information about housing costs in these areas. If these clusters are treated as random blocks, then we can partition the unexplained variability as a result of within group variation and reduce the amount of residual variability in the final model, yielding prediction/inference which characterizes reality better. Moreover, when assessing the price of a house, we may be more interested at determining trends related to intrinsic features of the house, like number of square feet and how this effects the price, and not the number of coffee shops within the neighborhood. In the methodology section we will discuss the

difference between the so called "random" effects, what can be considered the clustering, and the "fixed" effects, what can be considered as the number of square feet for example. We shall consider all of Cook County, IL to perform this analysis, as it includes the city of Chicago, and some of the surrounding suburban areas, providing an interesting mix data to be analyzed.

The goal of this report is to understand the variation in housing prices and determine if the clusters generated by venues impact the housing prices as a fixed effect or a random effect. In the methodology section we will discuss the details about ordinary least squares linear regression and mixed effect models and how each method imposes assumptions about the structure of the data and the variability in the response.

Data

Source Information

Several source of data will need to be aggregated in order to answer the questions outlined above, namely we will require: neighborhood information, geographic (coordinate) information, venue information, and real-estate information.

- Neighborhood/community information, for the city of Chicago will be obtained from Wikipedia: https://en.wikipedia.org/wiki/Community_areas_in_Chicago and community/town information for areas outside of the city of Chicago will be obtained from: http://ciclt.net/sn/clt/capitolimpact/gw_ziplist.aspx?ClientCode=capitolimpact&State=il&StName=Illinois&StFIPS=17&FIPS=17031. (Note that within the city neighborhoods are nested in community areas. Community areas will be considered for areas in the City, and towns/villages will be considered for the suburban areas.)
- Geographic coordinate information will be obtained from the geopy.geocoders package.
- Venue information will be obtained using the Foursquare API.
- Real-estate information will be obtained from https://www.realtor.com/research/data which provides aggregated housing information at the zip code level for the entire US. In the future the Zillow API could be integrated to obtain this information at the individual house level, but for the sake of time and simplicity, the aggregated information will be used as a means of exploratory analysis in order to see if there is a significant effect due to blocking by access to like venues. The relator.com data set also provides aggregated zip code level information each month for the past four years, which may enable a longitudinal analysis of housing prices for Cook County.

The cleaning and aggregation process will be described briefly below, but for a full accounting of the process see the notebook for details:

Acquiring, Cleaning, and Merging the Data

Geographic Data

As described above, the data comes from a variety of sources, and will need to be wrangled appropriately. Note that the neighborhoods within the city of Chicago are lumped into the community areas. The is due to the fact that the geocoder that is used is unable to parse the query for various neighborhoods, but returns the expected coordinates when a community area is used. The community areas are used for statistical and planning purposes, so lumping the neighborhoods into community areas should not be a significant concern for the following analysis. Cook County, IL consists of the city of Chicago and various towns or villages in the surrounding suburbs. We treat these town or villages as community areas, however we shall make the distinction between a city community and a suburban community. This is easily done since the city and suburb information come from different data sources.

The `geolocator` package is used for geocoding the coordinates of the communities. The query is particular, and can sometimes yield strange results. We began with the city areas first. Fortunately the library recognized almost all of the official community areas. The web-source for the city areas did not have zip-code information

so, we had to reverse geocode to obtain the zip code. Zip codes are required to match the venue data with the real estate data.

Next the suburban communities were compiled. This was done in the same process as before, but this web-source contained zip code information. Some of the same communities were listed multiple times with different abbreviations e.g. Village and Vlg or Heights and Hts, so these records were removed. Some of the suburban communities were not parsed well by the geocoder and were placed in erroneous locations, like Canada and France, so seven records were dropped from this data. The map below plots the locations of the communities:

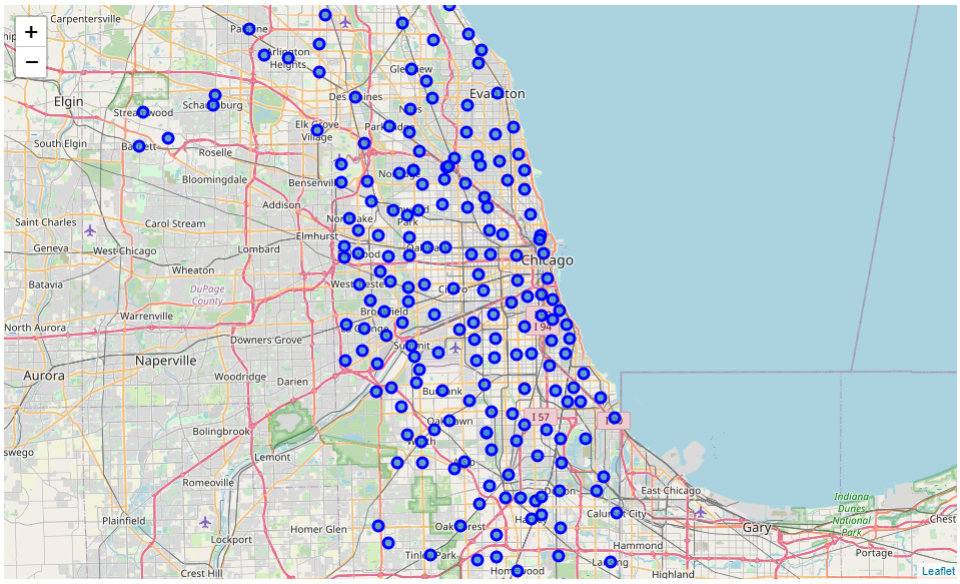

Figure 1: Map of Cook County Communities

With the city and suburban tables ready, they were joined together and retained with only the community name, a distinction of city or suburban, latitude, longitude, and zip-code variables. Two more variables were created describing the distance to the closest community (for use in the Foursquare API) and the distance from the City center of Chicago (as a predictor in the models). The geographic information is now ready for use with the Foursquare API and further analysis.

Venue Data

The Foursquare API is utilized to obtain venue information across Cook County, IL. The main purpose of obtaining the geographic information for each of the communities is to utilize the API effectively. Fortunately one of the labs walked us through this process, so the code for this is adapted for this analysis below. Note that the we only call 200 venues, which should be more an enough for the analysis without exceeding the call limit for the day. Note that this process returns different results on different days due to the nature of the data vendor. Typically, the API returns at least 14,000 venues across the Chicago area. The venues are then encoded using a "one-hot" method, i.e. an indicator variable is generated for each venue then associated with a location so that analysis can be performed. The venues are then grouped by first, second, third, etc... up to tenth most common venue for each community. This aggregation step allows one to investigate the behavior of the venues and how they relate to the clusters.

Real Estate Data

As described earlier, https://www.realtor.com/research/data provides a repository of aggregated real estate data across the USA. This is presented as a `.csv` file, so this makes ones life quite easy. The main disadvantage is that it is aggregated at the zip code level. (Note that the link larger historical inventory file

is provided, so that way the time remains consistent. However, this should be a fine enough scale to capture the overall behavior in real estate prices. This clean data set was loaded into memory then filtered based on the relevant zip codes, given by the geographic information data, and for June 2020. The date field is a numeric variable, and not a date time variable. Since we are not considering time dependence for this analysis, it is not converted to a different type. If a longitudinal or time dependent analysis were performed, then it may be germane to convert to a date-time object.

All three components above will be merged together into a single table after the clustering step.

Methodology

The main idea of this analysis to to examine if venues in the community areas in the city and suburbs of Chicago cluster together, and given this clustering if these clusters directly effect the price and/or if there is correlation in the housing prices. If this correlation does occur, we require a tool that allows for this. Initially, one may consider analyzing the `cook_prop_june2020` data alone as a multiple regression situation, i.e. consider the mean or median home prices as a response the the other features of the data set as predictors. The regression coefficients would yield the price per predictor increase in decrease. One key assumption when performing ordinary least squares regression is that the observations must be independent and identically distributed, that is to say they must be a random sample form a, typically, normal distribution. However, it is often the case that geographic observations tend to be correlated due to grouping by location, which violates the 'iid' assumption of OLS. We require a more flexible framework to account for this.

Mixed effects modeling and REML (restricted maximum likelihood) estimation provides the flexible framework required to do this. In a nutshell, this amounts to specifying a model with fixed effects and random effects. The fixed effects are unknown constants that we are trying to estimate from the data, or the usual regression coefficients on the median square feet from the real estate data for example. A random effect is simply a random variable, which implies we estimate the parameters that define this random variable. The random effects are usually taken to be a blocking variable, account for correlation structure, or account for samples from a representative population. Sometimes the choice is clear, and sometimes people may disagree on the choice of random effects, however this is part of the art of data analysis! For example, in a clinical trial the patient is a random effect and the treatment is a fixed effect. Moreover, the random effect is allowed to take on a nested structure e.g. the individual patient is nested in the hospital. In the case of this report, we wish to see if the house prices in neighborhoods clustered by venues are correlated, and account for this correlation in the model when the fixed effect parameters are estimated. REML estimation is an algorithm that produces unbiased estimates and does not require that the data be balanced, among other useful features.

Instead of using the `scikitlearn` library, a different tool will be used: the `statsmodels` library, which provides ordinary least squares (OLS) estimation and fits mixed effects models. One advantage to this package is the output it produces and the the fact that one enters model formulas in to the regression function instead of inputting an array of features directly. The output also displays the estimated coefficients, t-statistics, confidence intervals, etc... for the regression parameters and other useful measures.

Without delving into too much mathematics (resources are listed in the informal bibliography), let's specify the different models and interpret the parameters in the model. The model for the OLS method is:

$$\text{price}_{ij} = \beta_0 + \sum_{i}^{p} \beta_i \text{feature}_i + \epsilon_{ij} \tag{1}$$

Where β_0 is the overall mean, β_i's are the regression coefficients, feature_i's are the relevant predictors for the model and ϵ are the errors assumed to be distributed normally with variance σ^2. This this the classic multiple regression situation. Note that for the feature variable in this model, the clusters will be taken as factor levels, requiring k additional parameters to be estimated, where k is the number of clusters. These will adjust the intercept of the line. The other feature variables, like square feet, will indicated the unit price per unit feature increase or decrease.

The mixed effects model is similar, but changes the overall structure, namely we have:

$$\text{price}_{ijk} = \beta_0 + \sum_{i}^{p} \beta_i \text{feature}_i + b_k + \epsilon_{ijk} \tag{2}$$

Now, we exclude the clusters as features in the model, but instead assign it to b_k which is a normally distributed random variable with variance $\sigma^2_{\text{cluster}}$, namely the random variation due to the grouping of the clusters. Instead of imposing that the intercepts due to the grouping are directly estimated, we assume that they have random normal distribution. This allows us to understand how the intrinsic features of a house relate to the price, and less upon correlation due to clustering structure.

Analysis

Clustering

With the data prepared from the previous section, we can now cluster the venues together using k-means clustering. We choose 10 clusters, as this value provides enough flexibility for communities to group themselves according to the structure in the data, or we do not want to provide a large constraint on the clustering. Moreover, the number of clusters is not particularly important for the end goal, as we are treating the clusters as not of principle interest to the response. In either model, we wish to control for their effect, or account for the correlation in prices within a cluster. (That being said, we could optimize the number of clusters based on the following heuristic: after the fixed effects are chosen the number of clusters could be varied and the number of clusters that yields the smallest AIC or BIC could be chosen.)The figure below displays the clustering of the venues across Cook County, IL:

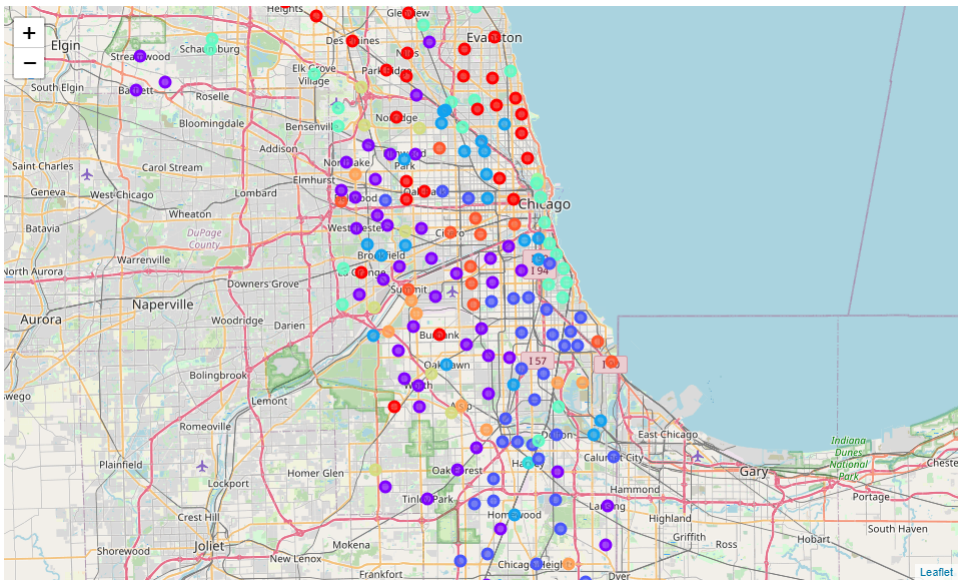

Figure 2: Map of Clustered Community Venues in Cook County, IL

Each color represents one of the 10 clusters returned from the k-means clustering algorithm. The main message that the map communicates is that the clusters are mostly correlated by geography. Let's aggregate venue data into an array where we count the number of venues for the the first, second, third, fourth, and fifth most common venues. We then select only the venues with counts greater than 5, so that we can visualize the proportion of venues among the different clusters.

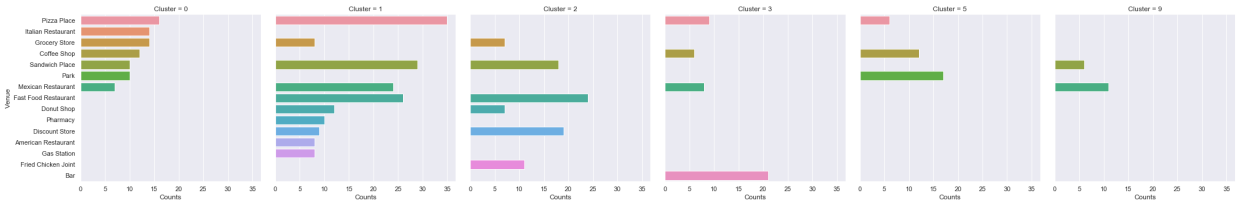

Figure 3: Bar Chart of top Clutering Venues

Note that not all clusters are represented, since they do not have a common venue count greater than 5. This may indicate that these clusters are small, or do not have a common pattern that emerges above this threshold. However from the venues that to meet this threshold level: there are a few things we can glean from this grouping of venues:

- Cluster 0 has quite a few dining options: for the foodie who can't get enough.
- Cluster 1 "has it all" pizza, parks, and various other options: a busy place for a busy-body.
- Cluster 2 has more fast food options and bars: McDonald's is always better when soused.
- Cluster 3 has the night life, with a high number of bars: when Friday bleeds into Saturday.
- Cluster 4 has parks and coffee shops: for the introspective Bohemian.
- Cluster 9 is for sandwich and park lovers: an office workers pleasure.

As discussed before, the particular characteristics of each cluster are not of principal interest, rather we want to examine correlation in housing prices among these clusters or a direct relationship.

Property Data EDA and Feature Selection

Now, lets examine the features of the real-estate information. The median listing price will be chosen as the response, since the median is typically robust to outliers when compared to the mean/average, therefore is a more representative measure of housing prices in each of the zip codes. At this stage we are looking for possible candidates for fixed effects. Collinearity is still an issue, so the features need to be chosen carefully. Notice that there are quite a few features in the real estate data. A glossary of terms is provided on the data source website. Observe that the `_mm` or `_yy` suffix represents the percent change of that measure from the previous month or previous year respectively. The `_mm` feature will be chosen for this analysis, since we are not interested in the effect of time, these variable will be ignored. We will also ignore price per feature ratios, since they provide the same information. Notice that median or average aggregate the same measurement. The median will be chosen since it is robust to outliers. This may be important since large expensive houses could be on the market (like Michael Jordan's was recently) that may effect the aggregated information.

With these observations and assumptions, let's select a feature set and generate a pairs plot looking at the relationships between the variables.

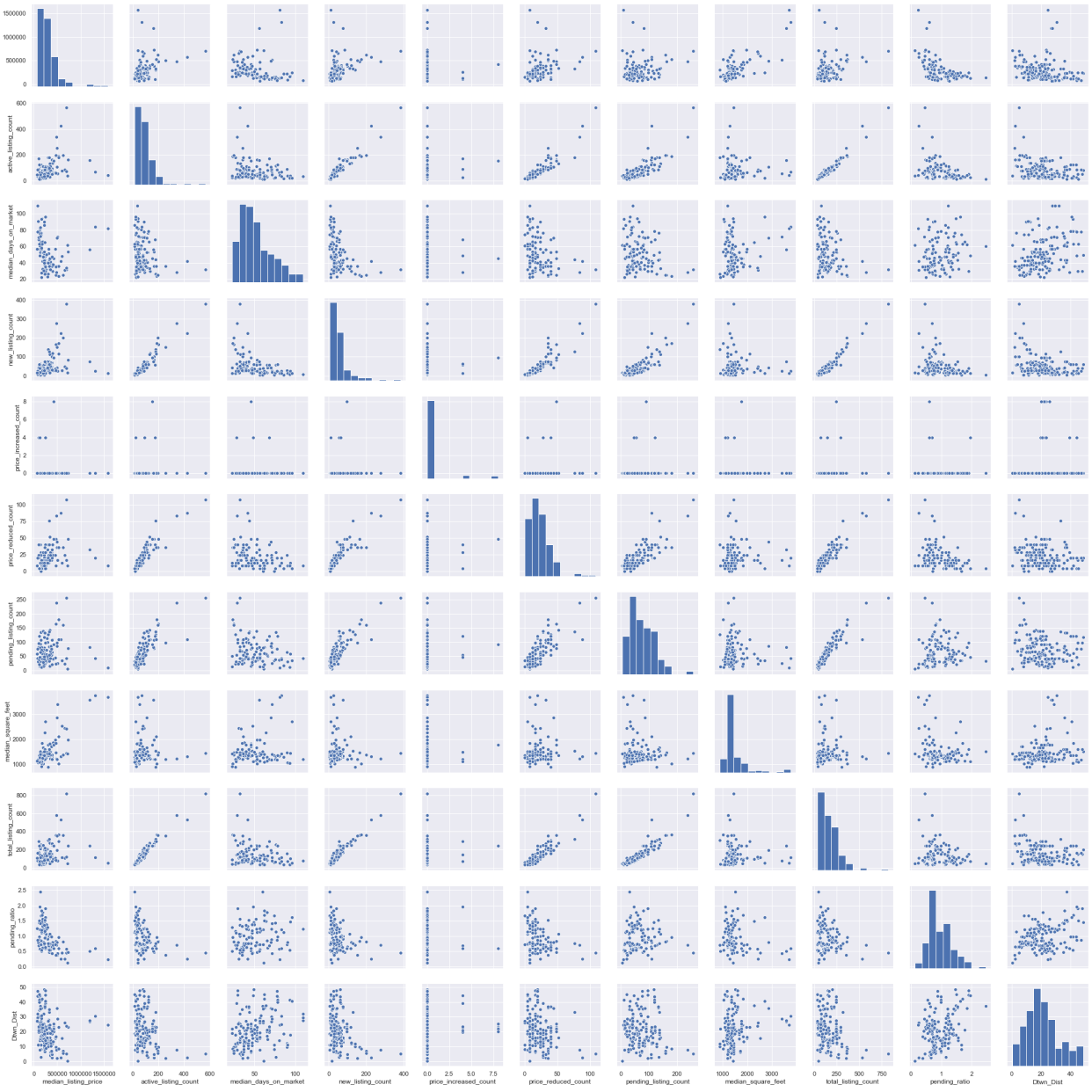

Figure 4: Pairs plot of the relavent variables

Notice that there is a slight issue with the distribution in the response variable (median list price), due to its skewness. This implies is may be germane to consider a power transformation, in this case the cube-root transformation, of the response variable, so that we do not violate the assumption of normality. This is easily assessed with a qq-plot or histogram:

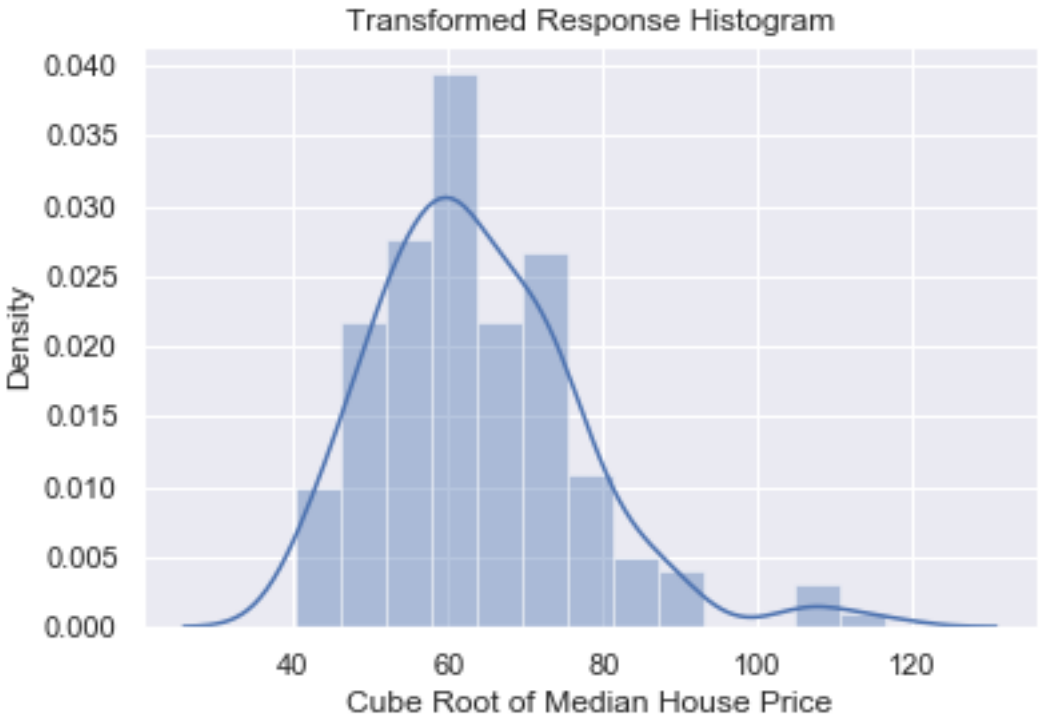

Figure 5: Histogram of Transformed Response

Observe that a cube-root transformation get us close to normality. There are outliers in the data. We can also see that the main features of interest are:

- `median_square_feet` (0.74)
- `pending_ratio` (-0.47)
- `new_listing_count` (0.37)
- `active_listing_count` (0.34)
- `price_reduced_count` (0.33)
- `total_listing_count` (0.26)
- `median_days_on_market` (-0.26)
- `Dtwn_dist` (-0.20)

Where the number in parenthesis indicates the correlation between that particular variable and the response. Notice that the new listing count is strongly correlated with the active listing count, price reduced count, and the total listing count. This makes sense with intuition, since this quartet of values essentially measures the same thing: how may houses are on the market in a particular zip code. Since the new listing count is the most strongly correlated with the median price, it is retained and the others are removed. Thus the new feature set is:

- `median_square_feet` (0.74)
- `pending_ratio` (-0.47)
- `new_listing_count` (0.37)
- `median_days_on_market` (-0.26)
- `Dtwn_dist` (-0.20)

Let's plot the transformed median price versus each of these variables in order of strongest to weakest correlation:

Figure 6: Median Listing Price vs. Median Square Feet

Figure 7: Median Listing Price vs. Pending Ratio

Figure 8: Median Listing Price vs. New Listing Count

Figure 9: Median Listing Price vs. Median Days on Market

Figure 10: Median Listing Price vs. Distance from City Center

Last, we can generate a box-plot of the median listing price versus each cluster:

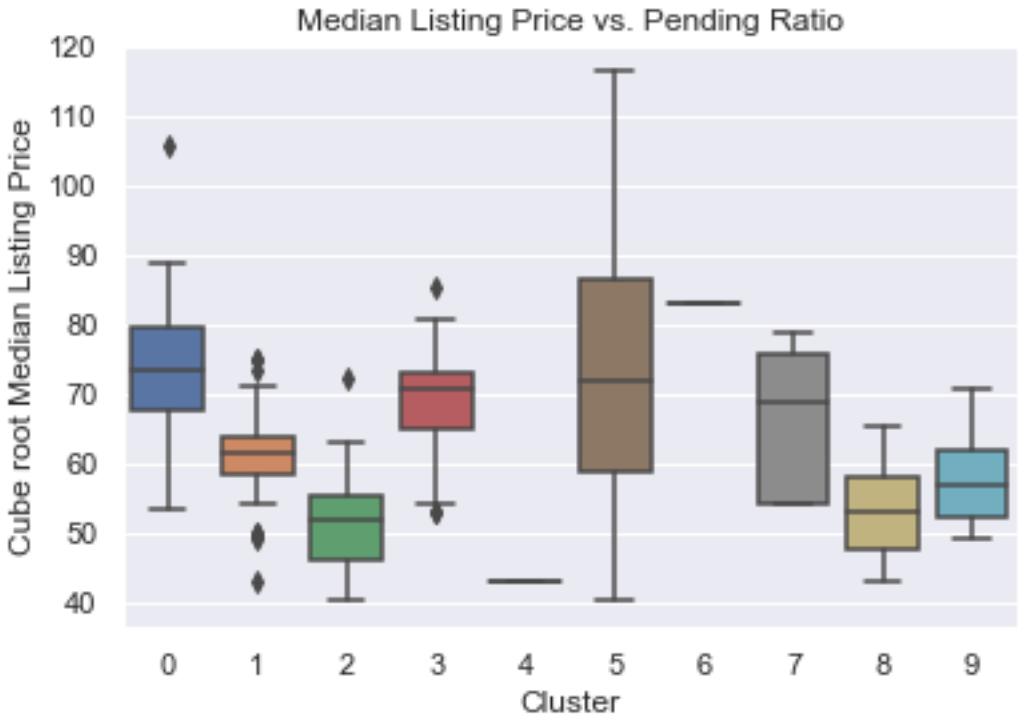

Figure 11: Median Listing Price vs. Cluster Number

In the scatter plots we can observe a dependency between the two variables, and we can see that each cluster tends to form strata in the larger cloud of points. One may anticipate that a model that contains all of these features would perform the best.

Ordinary Least Squares Modeling

To begin, lets consider a linear regression of the features/predictors listed above. We use the `statsmodels` library as it provided an R like output, which makes the process of inference easier to perform, in comparison to the `scikitlearn` library. Another benefit to the `statsmodels` library is that we can fit "R like models" which make the code easier to read, understand, and think critically about, since we write the appropriate formula like we would on paper. Once again, due to the nature of the data, we wish to make inference about the features of the home and how they relate to the price of the home, instead of modeling population level data and performing prediction, in the machine learning sense. In the notebook we entertain seven models, the model that performs the best among these seven, in the sense of largest adjusted R-squared, significance of coefficients, and lowest AIC and BIC, is the one which contains all the variables from the previous section as features, namely a model of the form:

$$\begin{aligned}\text{cube root of price} = \beta_0 + \sum_{k=0}^{9} \text{cluster effect}_k + \beta_1 \text{median square feet} + \\ \beta_2 \text{pending ratio} + \beta_3 \text{new listing count} + \beta_4 \text{median days on market} + \\ \beta_5 \text{distance from downtown} + \epsilon\end{aligned}$$

The results of this regression are displayed:

OLS Regression Results

Dep. Variable:	np.cbrt(median_listing_price)	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.855
Method:	Least Squares	F-statistic:	73.67
Date:	Sat, 01 Aug 2020	Prob (F-statistic):	1.04e-61
Time:	12:35:10	Log-Likelihood:	-517.46
No. Observations:	173	AIC:	1065.
Df Residuals:	158	BIC:	1112.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	51.8721	2.673	19.407	0.000	46.593	57.151
C(Cluster_Labels)[T.1]	-1.4138	1.447	-0.977	0.330	-4.273	1.445
C(Cluster_Labels)[T.2]	-5.9654	1.708	-3.493	0.001	-9.339	-2.592
C(Cluster_Labels)[T.3]	0.9859	1.520	0.648	0.518	-2.017	3.989
C(Cluster_Labels)[T.4]	-2.0938	5.464	-0.383	0.702	-12.886	8.699
C(Cluster_Labels)[T.5]	3.4920	1.560	2.238	0.027	0.410	6.574
C(Cluster_Labels)[T.6]	3.1165	5.313	0.587	0.558	-7.378	13.611
C(Cluster_Labels)[T.7]	-0.7452	2.092	-0.356	0.722	-4.877	3.387
C(Cluster_Labels)[T.8]	-4.7104	2.270	-2.075	0.040	-9.194	-0.227
C(Cluster_Labels)[T.9]	-4.7922	1.945	-2.464	0.015	-8.634	-0.950
median_square_feet	0.0181	0.001	20.509	0.000	0.016	0.020
pending_ratio	-3.1402	1.289	-2.435	0.016	-5.687	-0.593
new_listing_count	0.0571	0.010	5.817	0.000	0.038	0.076
median_days_on_market	-0.1654	0.028	-5.896	0.000	-0.221	-0.110
Dtwn_Dist	-0.2346	0.044	-5.380	0.000	-0.321	-0.149

Omnibus:	10.652	Durbin-Watson:	1.249
Prob(Omnibus):	0.005	Jarque-Bera (JB):	22.697
Skew:	-0.159	Prob(JB):	1.18e-05
Kurtosis:	4.746	Cond. No.	2.27e+04

Figure 12: Print-out from the Regression

Even though ten extra parameters must be estimated, the model that includes the clusters does outperform a model that does not include the cluster labels. Also notice that not all clusters have a significant effect. Note that we do get some warnings about multicollinearity and numerical problems. Ideally we should resale the data for analysis, but we loose interpretablity on coefficients. Let's examine the residuals for this model to check assumptions:

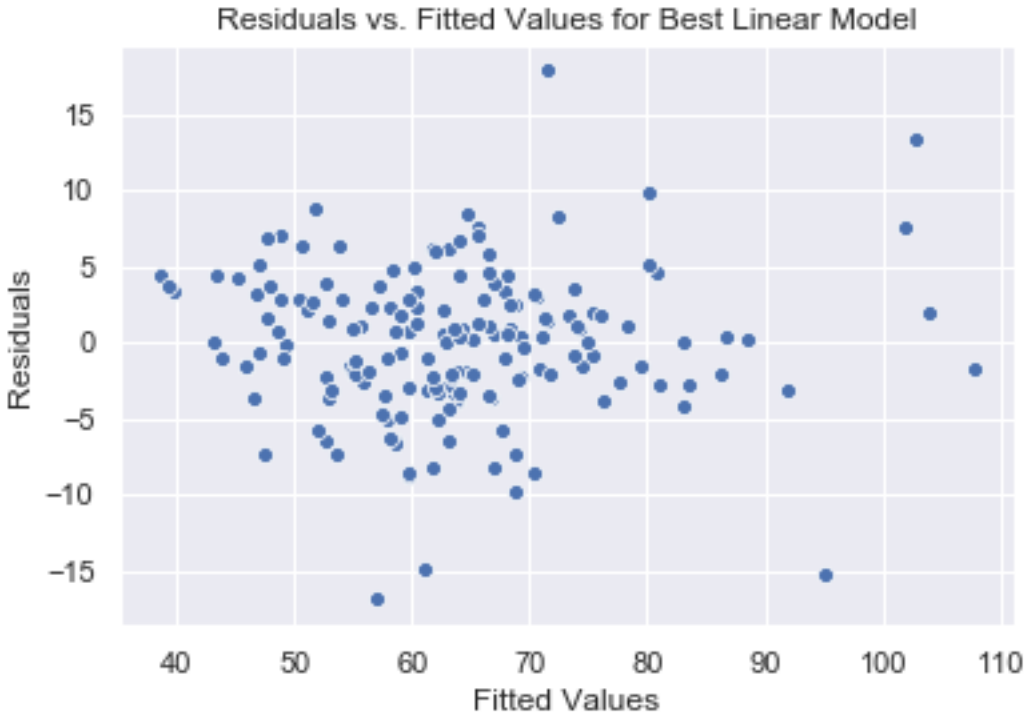

Figure 13: Scatterplot of the Residuals vs. Fitted Values for the OLS Model

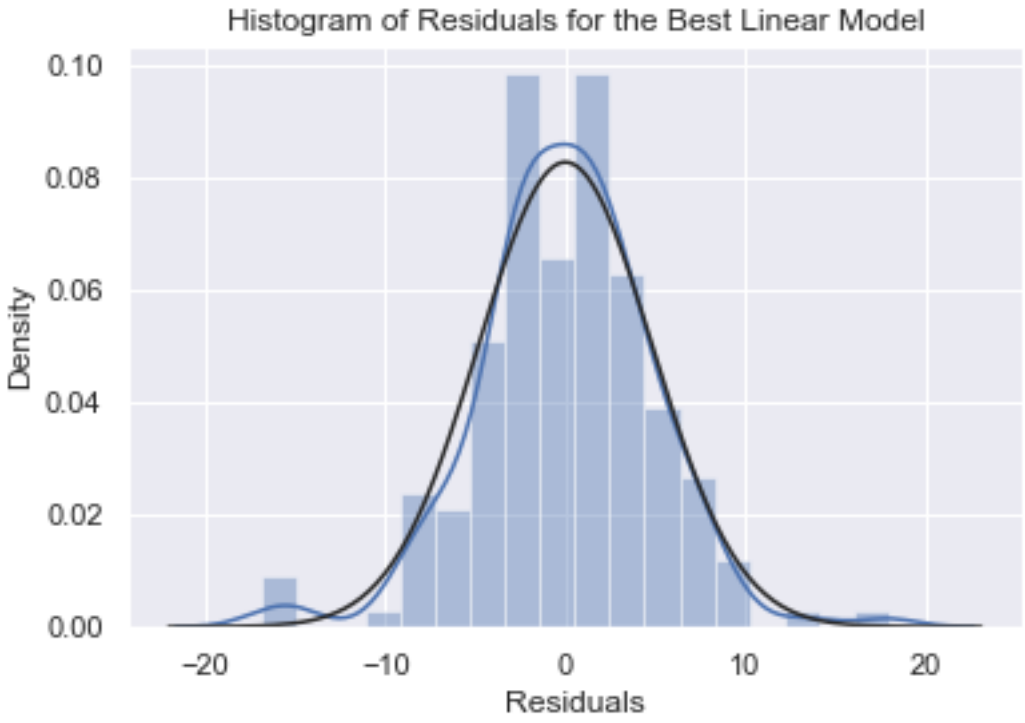

Figure 14: Histogram of the Residuals for the OLS Model

Significant deviations from normally distributed errors do not appear to be an issue. While there are few outliers present, we may expect such behavior due to the nature of the data.

Mixed Effects Modeling

The five predictors previously listed will be studied as fixed effects and the venue clusters will be the random effects for the mixed model analysis. Before fitting fixed effects is often useful to examine an empty model where no fixed effects are present and only the random effect is present. This enables us to calculate the intraclass correlation coefficient which is approximately 0.68, indicating there is correlation within the groups. Several different models were investigated using the mixed effects approach. The model that performed the best is the one that used all features described previously as fixed effects, but allowed a nesting structure in the random effect. If one allows the cluster to be nested in the city or suburban groups, the best among all the models considered here is obtained. The print-out from the function is:

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	np.cbrt(median_listing_price)
No. Observations:	173	Method:	REML
No. Groups:	10	Scale:	24.1146
Min. group size:	1	Likelihood:	-542.9958
Max. group size:	45	Converged:	Yes
Mean group size:	17.3		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	53.410	2.684	19.900	0.000	48.149	58.670
median_square_feet	0.018	0.001	19.211	0.000	0.016	0.020
pending_ratio	-4.016	1.294	-3.104	0.002	-6.552	-1.480
new_listing_count	0.059	0.010	5.754	0.000	0.039	0.079
median_days_on_market	-0.175	0.027	-6.531	0.000	-0.227	-0.122
Dtwn_Dist	-0.308	0.052	-5.970	0.000	-0.410	-0.207
City Var	9.780	1.054				

Model 7 AIC = 1099.9915672961417
Model 7 BIC = 1122.0646084576263

Figure 15: Results from Fitting the Mixed Effects Model

We can check assumptions by examining the residuals from this model:

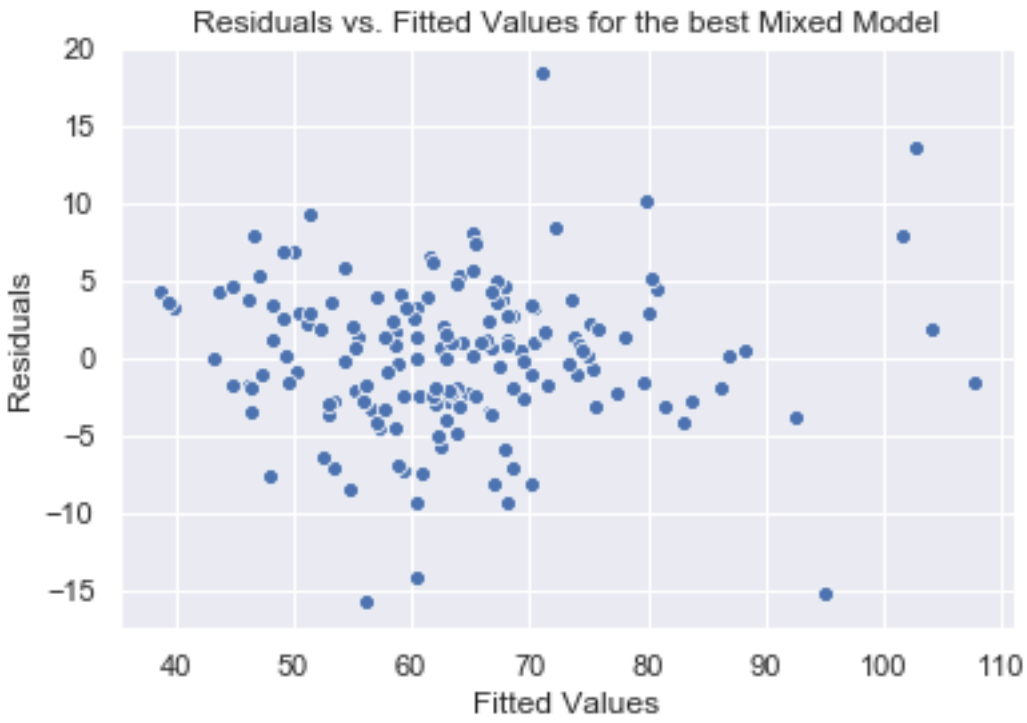

Figure 16: Scatterplot of the Residuals vs. Fitted Values for the Mixed Model

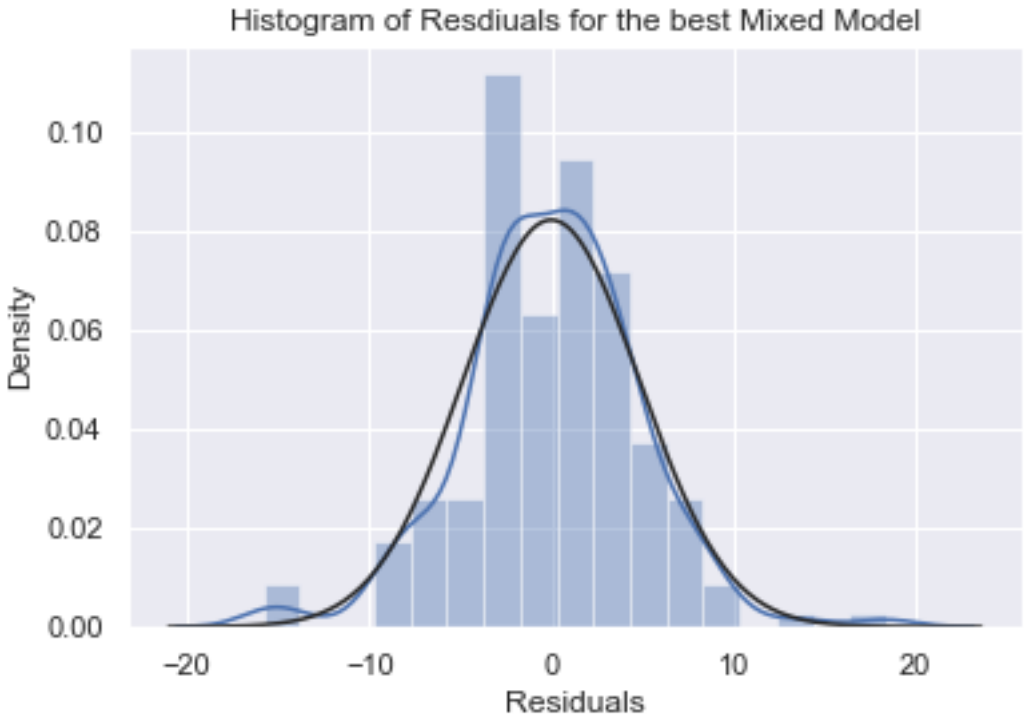

Figure 17: Histogram of the Residuals for the Mixed Model

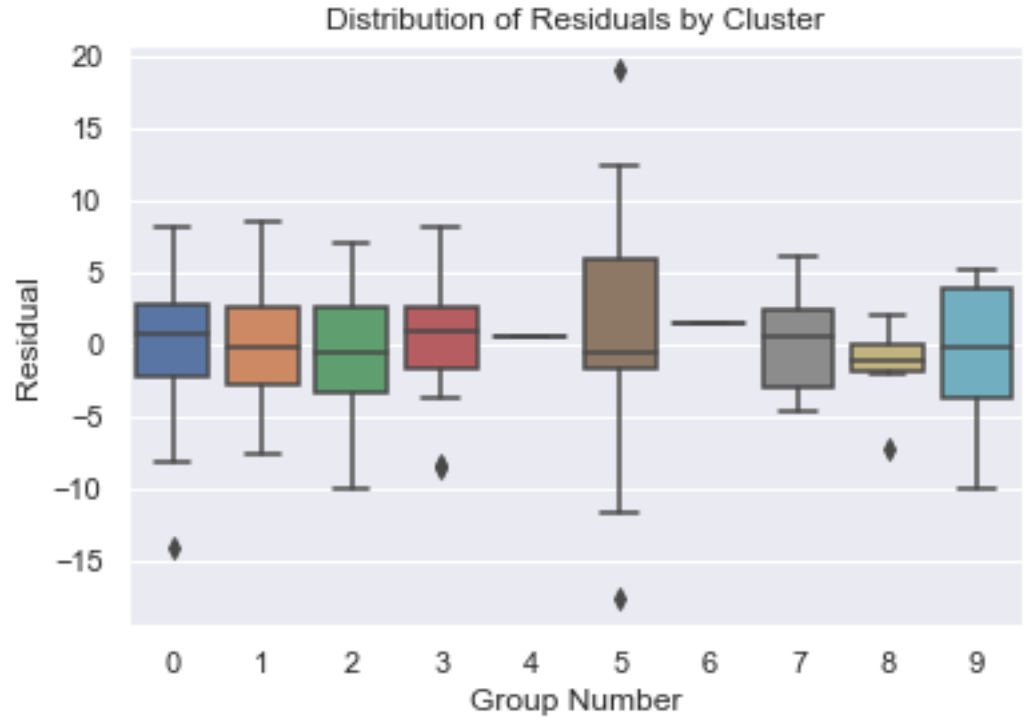

Figure 18: Residuals for each Cluster Group

Noticed that we have similar observations regarding the residuals from the OLS modeling procedure. However, we do appear to violate the assumption that the errors across groups are all the same. This may indicate that the grouping variable can be refined, or that this grouping is not driving the variation in the data.

Comparing the Predictive Accuracy of the Final Models

Let's compare the two models discussed above (considering only the fixed effects) in the accuracy in predicting the July median housing prices. Let's take a random sample of 100 zip codes and predict the house prices for each of the models. Note that Python and the statsmodels library only predict on the fixed effects for the mixed effects model. For the linear model the mean difference in price between the predicted and actual is ≈ -0.40, and for the mixed model is ≈ -0.57. In both instances they are biased lower, which is likely due to a change in price over the month period. Also, the standard deviation in the error of predicting price for the linear model is ≈ 5.08 and for the mixed model is ≈ 5.86. In general this value is larger for the mixed model than the linear model. Both models are able to predict the median home price in a zip code with an error of a few hundred dollars.

In the next section, we will compare and contrast the two models discussed here, in addition to describing problems and possible solutions.

Discussion

To briefly summarize, the analytic approach described in this report was to cluster the venues of Cook County, IL in to like groups then "control" for this grouping, or see what effect, it has on the median housing price across the zip codes of the county. We then can investigate variables intrinsically related to the cost of real estate. We examined a standard linear model and fixed effects model. Let's begin by comparing and interpreting the fitted regression coefficients for the variables related to real estate price:

Feature	Linear Model Coefficients	Mixed Model Coefficients
Intercept	52.261	53.410
Median Square Feet	0.018	0.018
Pending Ratio	-3.140	-4.016
New Listing Count	0.057	0.059
Median Days on Market	-0.165	-0.175
Distance from City Center	-0.235	-0.308

Each of these coefficients is interpreted as the unit change in the cube root of price per unit increase in the feature, e.g. for both models the the cube-root of price increases by \$0.018 per every increase in square foot. All of these coefficients are on the same scale, but the mixed model coefficients are larger in magnitude in general. We can also see that the pending ratio has a strong effect on decreasing a home price. The pending ratio is the ratio of pending listings to the number of active listings e.g. an area with a high number of pending listings and low number of active listings has a high pending ratio and visa versa.

Recall from the OLS regression results the cluster factors. Not all clusters had a significant effect from the overall mean in price, however we do see that clusters 0, 2, 5, and 8 are significantly different (at the $\alpha = 0.05$ level) from the mean. This indicates that these venues have a significant effect on either increasing or decreasing the overall median house price within that cluster.

A natural question at this point is: which method is better? In this analysis, both methods are comparable in terms of performance and inference. The better model is, of course, the model that describes reality better. The main advantage of the OLS model is the interpretabilty of the model and its coefficients, however, we require that each cluster has fixed group effect and do not allow for correlation. The advantage of the mixed effects approach is we can allow for this group effect from clustering to exists, but can allow for correlation among measurements in this group. We can also allow for a hierarchical grouping structure to exist.

There also may be limitations and biases in the data. One major limitation is that the property data is aggregated, at a zip code level, and does not describe the individual house price. There may be additional structure within the zip codes that is not accounted for. The Foursquare API data tends to return restaurant information, which may bias the venue information. In particular it will return more venue information for city areas, which likely have more restaurants per unit area than the suburban areas.

Conclusion

In this report we examined the relationship between the median housing price across the zip codes of Cook County, IL by compiling venue information and and information related to the house and housing market. The venues were clustered using k-means clustering then two regression approaches were considered: a standard ordinary multiple linear regression and a fixed effects approach. Both approaches resulted in similar results and performance characteristics.

The main goal, in addition to predicting the median house price, was to determine the effect of these clusters and examine if the venue clusters impact the price within that cluster. Some of the clusters have a significant overall effect on the price, and some do not. We also observed some evidence to suggest that there is correlation of housing prices among the clusters. So while the clusters do effect the response variable, it may be the case that the clusters is simply a proxy for an underlying variable, like median household income or demographic information. While there may be an effect due to clustering, it is likely not solely attributed to the clustering of venues, but is likely the combination of other factors. If we believe that there is clustering and correlation in home prices due to these different clusters the mixed effects model can be modified to account for this, especially of the clustering is correlated and hierarchical in nature.

"Informal" Bibliography

Libraries used for the analysis:

- `pandas`
- `numpy`
- `folium`
- `json`
- `requests`
- `geopy.geocoders`
- `scipy`
- `vincenty`
- `statsmodels`
- `matplotlib` and `seaborn`

Code for generating the Foursquare API calls and parsing the responses:

- https://www.coursera.org/learn/applied-data-science-capstone/ungradedLti/f0QY7/segmenting-and-clustering-neighborhoods-in-new-york-city

Tutorials on Mixed Effects modeling:

- https://www.statsmodels.org/stable/mixed_linear.html
- https://www.pythonfordatascience.org/mixed-effects-regression-python/