Modeling Home Prices

AN INVESTIGATION OF MODELING HOME PRICES ACROSS COOK COUNTY, IL

About Chicago, IL

Chicago is the third most populous city in the United States and largest city in the American Midwest at 234 square miles.

26 miles of lakefront, 8,800 acres of greens space and 600 parks

Art Institute of Chicago houses one of the largest collections of Impressionist/Post-Impressionist painting outside of Paris

250 theatres, 225 music venues, 200 dance companies.

Headquarters to many companies and financial firms

Truly a diverse and vibrant City

Introduction to the Objective

Simply: How much does a home cost?

We would like to investigate the variables that contribute to home cost.

Secondary Question: Do venues (e.g. restaurants, theatres, parks, et…) directly impact the cost of a home?

Nitty-Gritty Question: How do we model this behavior?

- Let’s suppose areas with like venues cluster together
- Using these clusters, will they result in a significant effect in the housing price?
- Do these clusters have structure, e.g. is there correlation in the housing prices within the clusters.
- Do these clusters have a “unique identity” describing the areas?
- Is there a distinction between urban and suburban/do they have and effect?

What data is required?

Neighborhood/community area information:

- Chicago neighborhood information is available on Wikipedia.com, and information on the suburbs is available from ciclt.net.
- Use the geopy.gecoders package to geocode the data.
- Main requirments are: zip code, latitude, longitude, urban/suburban, name.
- Assumption: we treat the community areas in the City of Chicago the same as towns/villages in the suburbs.

Venue Information: use of the Foursquare API

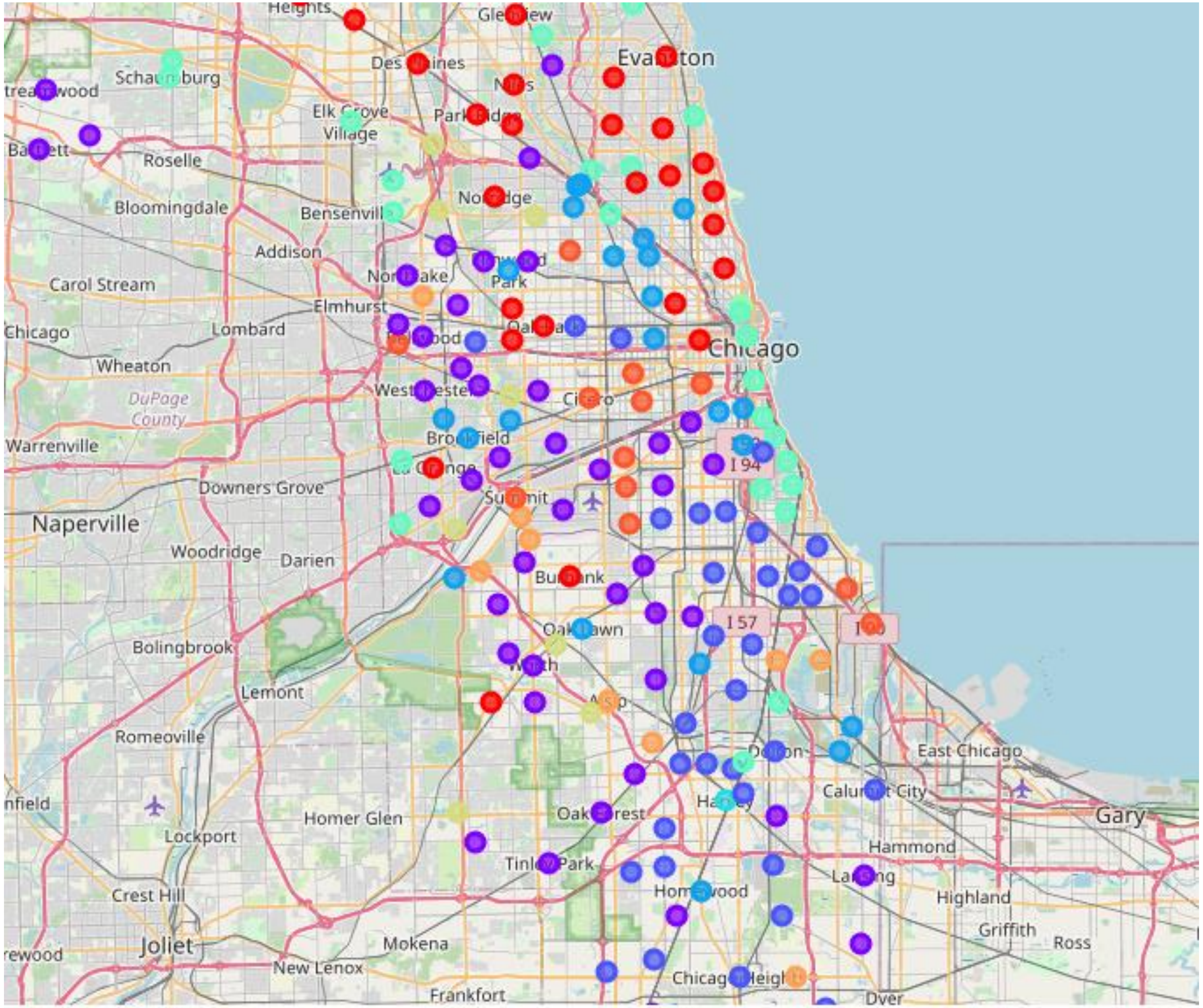
Real-estate information: realator.com provides real estate data for the past four years for all zip-codes across the U.S.

Results of Clustering

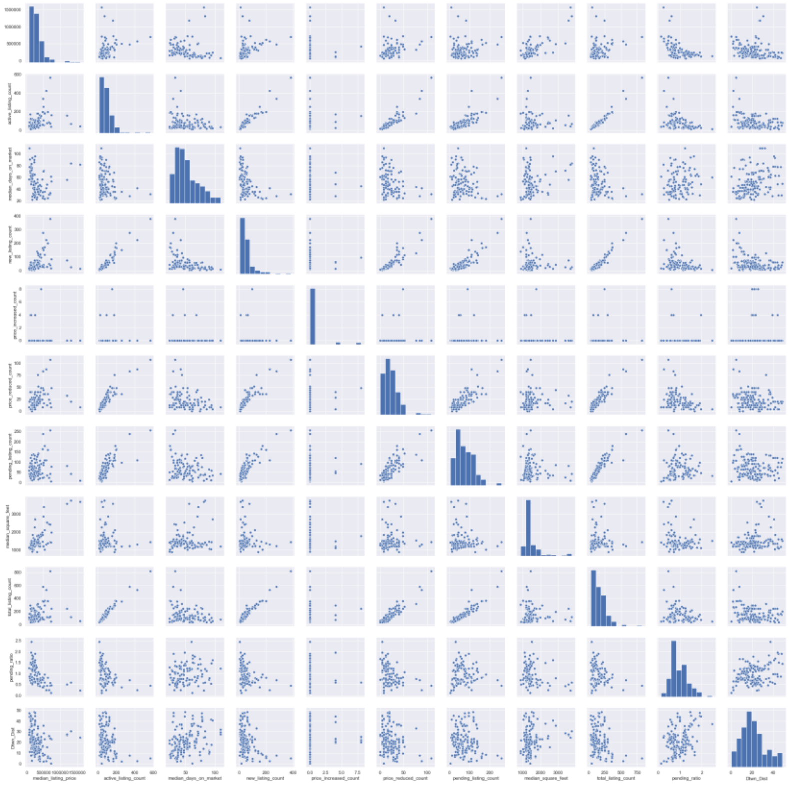
Clusters are mostly correlated by geography.

Not a clear description of the “identity” for each cluster. (Difficult to examine what makes each cluster a cluster based on the venues.)

Limitations with the Foursquare Data?

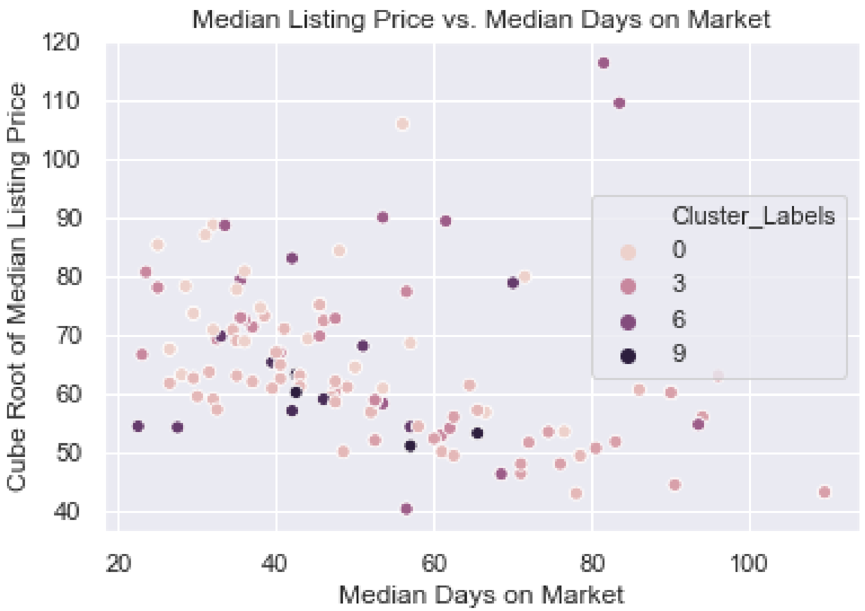


Real Estate Data Pairs Plot



Plots of Candidate Predictors

Plots of Candidate Predictors (2)



Plots of Candidate Predictors (3)

Results of Multiple Regression

```
                            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
==============================================================================
```

Some clusters have a significant effect some do not

The other predictors have significant and very interpretable effects on the listing price.

Residual diagnostics do not indicate anything problematic.

Results of Mixed Effects Modeling

```
                  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
```

Estimate fewer parameters.

Fixed effects are comparable to the previous model (but are larger overall)

The best model was determined to have a nesting structure between city and the cluster.

Some problematic residual diagnostics: residuals not the same across the different groups.

Conclusion and Future Directions

Both models have predictive abilities

Evidence to suggest that there is a group effect, and correlation within those groups.

The clusters may act as a proxy for an underlying variable.

Future Directions include:

- Aggregating more data
- Getting more specific housing data
- Changing the modeling method
- Understanding that there simply may not be an effect due to venue clustering.