HEDONIC HOME PRICE PREDICTION

Anthony Ayebiahwe & Steven Chang

November 6, 2018

INTRODUCTION

In this project, we seek to utilize local intelligence to build a predictive model of home prices in Nashville Tennessee. We aim to use our model to complement or even improve the existing model that Zillow currently has, since its current market predictions are not as accurate as can be. To do this, we gathered data from Nashville's Open Data Portal https://data.nashville.gov/ (https://data.nashville.gov/). This project will help us gain a better understanding on the housing market conditions in Nashville.

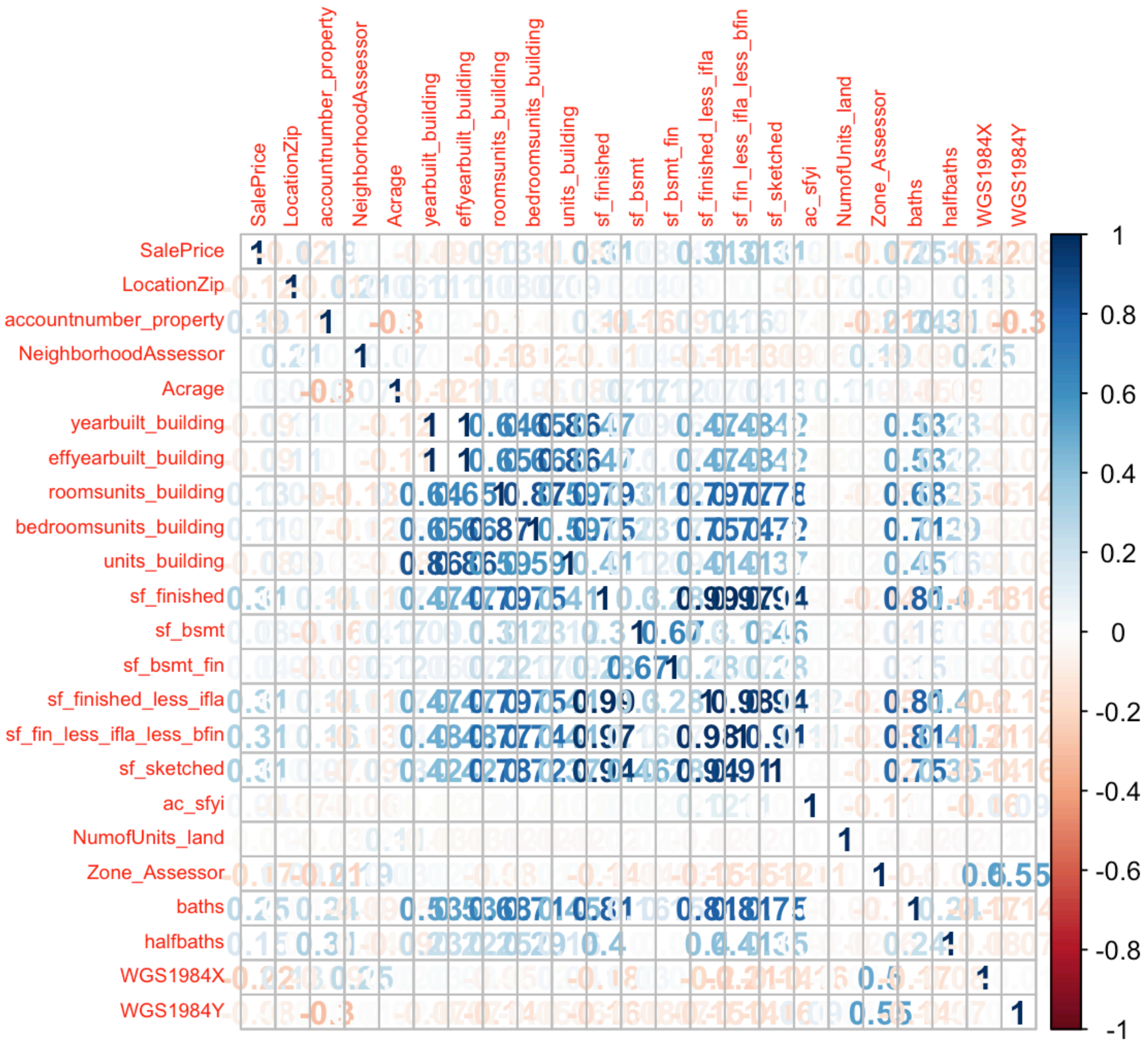
In this particular project, we are confined to OLS regression, making it a challenge to create a model that generates accurate predictions. Moreover, the data required very extensive cleaning and preparing before use. To improve the accuracy of our model, we will utilize feature engineering techniques to craft unique variables that can influence housing prices. For example, we believe that a house in good physical condition is likely to cost more than a house that is more run-down. In this case, we will create a dummy variable for the physical condition of the house, where a value of 0 will represent houses in poor condition and a value of 1 will represent houses in a good condition. Other factors we used include the structure of the frame of the house, the residential type of the house, and the year the house was built. We will then incoprate all the dummy variables into our regression model and find out if each factor truly had an effect on home prices.

From this project, we found that the year in which a house was built is significantly affects housing prices in Nashville. Newer houses tend to cost more than older houses. We also found that houses in good conditions cost more than houses in bad condtions. Moreover, we found that residential condos tend to cost less than single family houses. Finally, internal amenities such as the number of bedrooms, number of bathrooms, and whether or not a house has a basement do not significantly affect housing prices.

DATA GATHERING AND EXPLORATORY ANALYSIS

We gathered data from Nashville's Open Data Portal: https://data.nashville.gov/ (https://data.nashville.gov/), containing 20000 home prices and 57 variables related to each home in the Nashville area. In the data-cleaning process, we removed a few homes that are not located within Nashville itself, and removed all NA values in the dataset. We also filtered for sales prices that are not 0. Finally, we pulled the Nashville basemap and zipcode shapefiles from Nashville's Open Data Portal to map home sale prices and mean absolute error for our predictions.

The summary statistics of our variables, correlation matrix, map of home sale prices across Nashville, and 3 maps of our most interesting independent variables are included below.



This shows the correlation matrix among the variables.

Summary Statistics of All Variables

```
##
## Summary Statistics of All Variables
## ===============================================================================
=====================
## Statistic                      N         Mean        St. Dev.       Min        Pctl(25)
Pctl(75)     Max
## -------------------------------------------------------------------------------
---------------------
## kenID                        5,442    4,966.335     2,880.312       2          2,446.5
7,435.2      10,000
## SalePrice                    5,442   287,345.300   321,992.600      0          135,000
```

```
350,000   6,894,305
## OwnerZip                     5,442   38,565.420   10,741.430     804        37,205
37,215    372,211
## LocationZip                  5,442   37,210.450      4.642     37,201      37,206
37,215     37,221
## CouncilDistrict              5,442     17.434        8.636        1           8
24         34
## CensusBlock                  5,442 37,015,765.000  2,805.065  37,010,105 37,013,202
37,018,102 37,019,600
## accountnumber_property       5,442  141,729.200   71,896.640    19,828    76,727.2
210,866    266,227
## Card                         5,442     1.000        0.000         1           1
1          1
## NeighborhoodAssessor         5,442   3,936.333    1,596.946      107        3,131
4,367      9,336
## Acrage                       5,442     0.236        0.330       0.000       0.000
0.320      8.160
## yearbuilt_building           5,442   1,972.717     28.907      1,790       1,952
2,003      2,018
## effyearbuilt_building        5,442   1,986.601     21.883      1,899       1,970
2,005      2,018
## roomsunits_building          5,442     5.787        1.858         1           5
7          19
## bedroomsunits_building       5,442     2.748        0.860         0           2
3          12
## units_building               5,442     1.000        0.045         0           1
1          4
## sf_finished                  5,442   1,713.695     843.592       348        1,152
2,064      9,466
## sf_ifla                      5,442     0.000        0.000         0           0
0          0
## sf_bsmt                      5,442    227.341      483.320        0           0
0        3,531
## sf_bsmt_fin                  5,442     73.415      241.955        0           0
0        2,600
## sf_finished_less_ifla        5,442   1,706.306     850.586        0         1,150
2,061.5     9,466
## sf_fin_less_ifla_less_bfin 5,442   1,633.385     799.118        0        1,118.2
1,953      9,466
## sf_sketched                  5,442   2,359.725    1,343.007      440       1,403.2
2,995.5     14,068
## ac_sfyi                      5,442     0.996        0.065         0           1
1          1
## NumofUnits_land              5,442     78.322     1,929.763       1           1
1       116,741
## Zone_Assessor                5,442     3.829        2.554         1           2
6          9
## baths                        5,442     1.841        0.826         0           1
2          8
## halfbaths                    5,442     0.345        0.502         0           0
```

```
1           4
## fpla                       5,442     0.000         0.000        0          0
0           0
## WGS1984X                   5,442    -86.767        0.067     -86.923    -86.816
-86.724    -86.599
## WGS1984Y                   5,442     36.140        0.050      36.029     36.101
36.180     36.243
## test                       5,442     0.097         0.296        0          0
0           1
## ------------------------------------------------------------------------------------
---------------------
```

Summary Statistics of Variables with Internal Charateristics

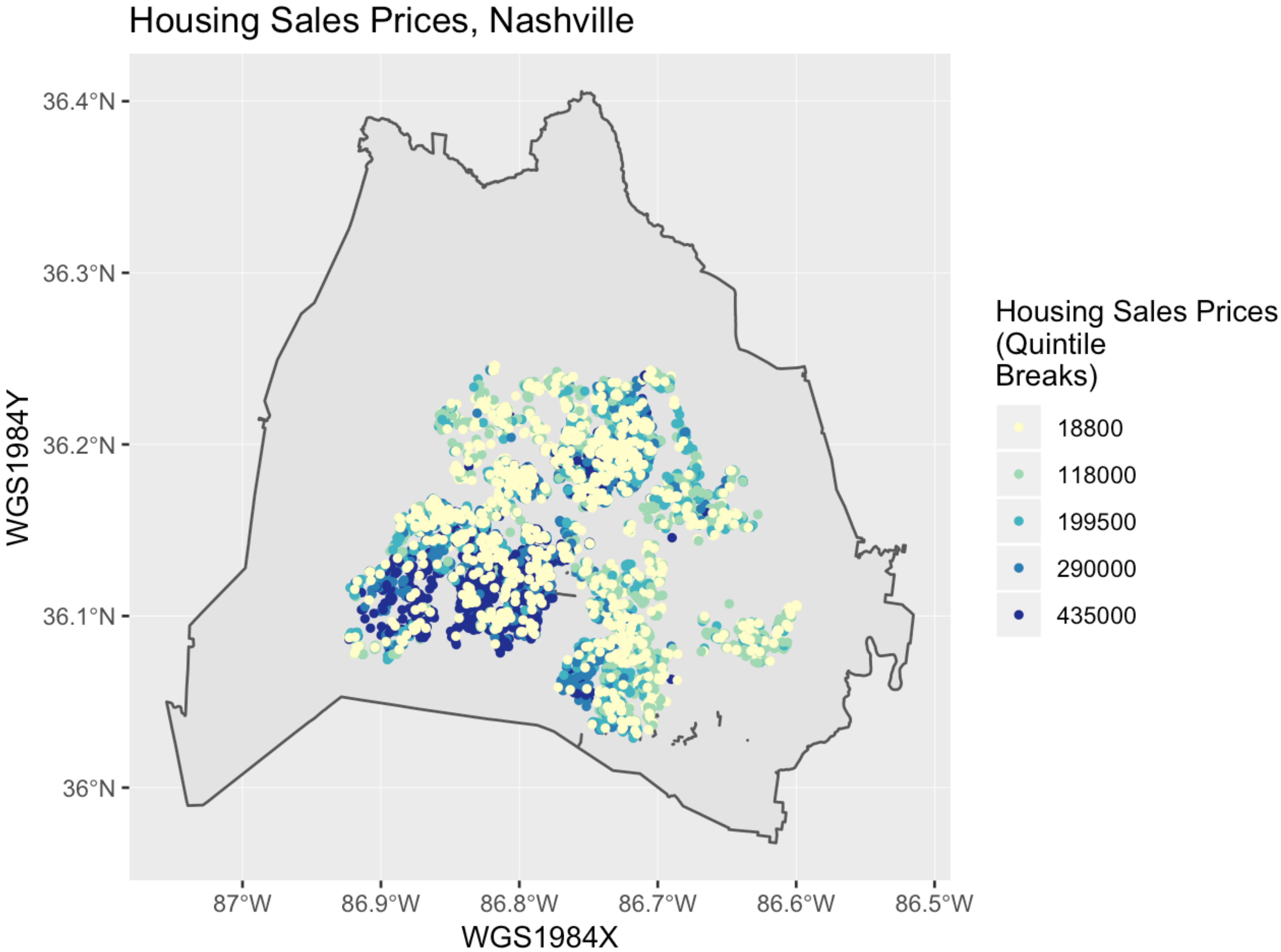
```
##
## Summary Statistics of Variables with Internal Characteristics
## =================================================================================
## Statistic                     N     Mean    St. Dev.  Min Pctl(25) Pctl(75)  Max
## ---------------------------------------------------------------------------------
## roomsunits_building        5,442   5.787     1.858    1     5         7       19
## bedroomsunits_building     5,442   2.748     0.860    0     2         3       12
## sf_finished                5,442 1,713.695  843.592  348  1,152    2,064   9,466
## sf_ifla                    5,442   0.000     0.000    0     0         0       0
## sf_bsmt                    5,442  227.341   483.320   0     0         0     3,531
## sf_bsmt_fin                5,442  73.415    241.955   0     0         0     2,600
## sf_finished_less_ifla      5,442 1,706.306  850.586   0   1,150   2,061.5  9,466
## sf_fin_less_ifla_less_bfin 5,442 1,633.385  799.118   0  1,118.2   1,953   9,466
## sf_sketched                5,442 2,359.725 1,343.007 440 1,403.2  2,995.5  14,068
## ac_sfyi                    5,442   0.996     0.065    0     1         1       1
## baths                      5,442   1.841     0.826    0     1         2       8
## halfbaths                  5,442   0.345     0.502    0     0         1       4
## ---------------------------------------------------------------------------------
```

Summary Statitics of Variables with Amenities/Public Services

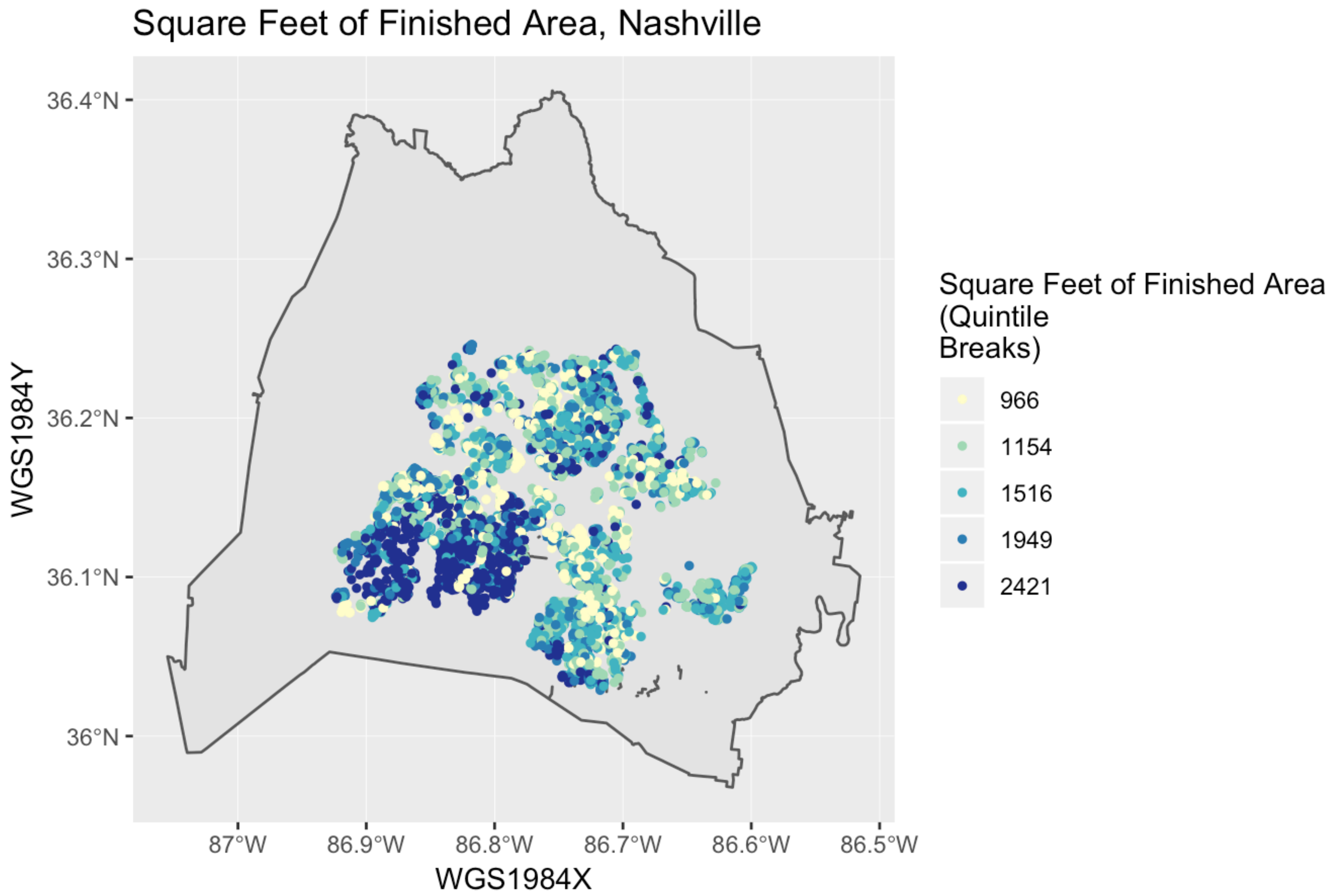
```
##
## Summary Statistics of Variables with Amenities
## ==========================================================
## Statistic   N   Mean  St. Dev. Min Pctl(25) Pctl(75) Max
## ----------------------------------------------------------
## ac_sfyi   5,442 0.996  0.065    0     1        1      1
## ----------------------------------------------------------
```

Summary Statistics of Variables with Spatial Structure

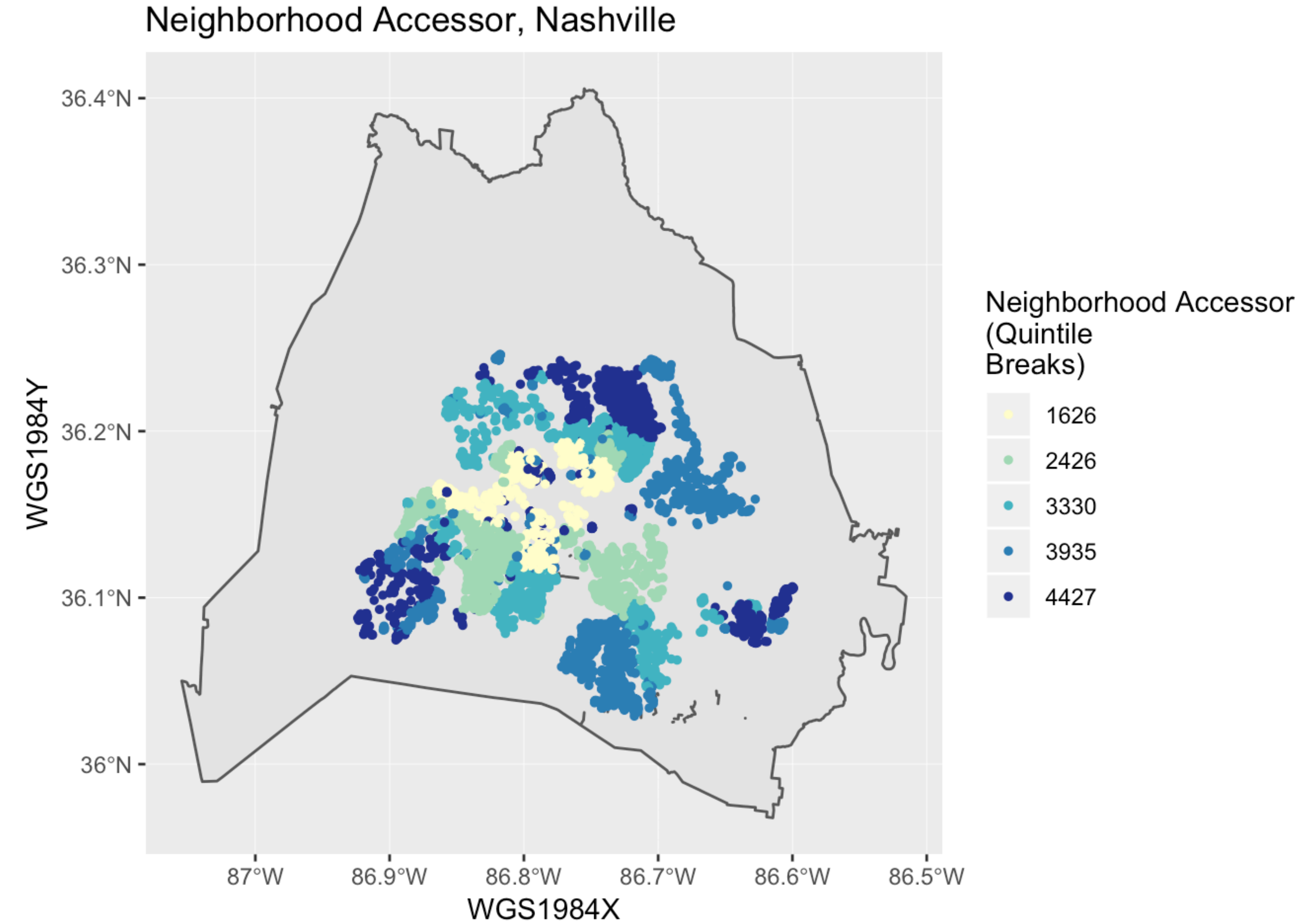
```
##
## Summary Statistics of Variables with Spatial Structure
## ===================================================================================
==================
## Statistic                     N        Mean       St. Dev.     Min      Pctl(25)   Pc
tl(75)     Max
## -----------------------------------------------------------------------------------
------------------
## accountnumber_property   5,442  141,729.200   71,896.640   19,828    76,727.2   21
0,866    266,227
## OwnerZip                 5,442   38,565.420   10,741.430    804       37,205     3
7,215    372,211
## LocationZip              5,442   37,210.450     4.642      37,201     37,206     3
7,215     37,221
## CouncilDistrict          5,442     17.434       8.636        1          8
24          34
## CensusBlock              5,442 37,015,765.000 2,805.065  37,010,105 37,013,202 37,
018,102 37,019,600
## accountnumber_property.1 5,442  141,729.200   71,896.640   19,828    76,727.2   21
0,866    266,227
## Card                     5,442     1.000        0.000        1          1
1          1
## NeighborhoodAssessor     5,442   3,936.333    1,596.946     107       3,131      4
,367      9,336
## Acrage                   5,442     0.236        0.330      0.000      0.000      0
.320      8.160
## yearbuilt_building       5,442   1,972.717     28.907      1,790      1,952      2
,003      2,018
## effyearbuilt_building    5,442   1,986.601     21.883      1,899      1,970      2
,005      2,018
## NumofUnits_land          5,442     78.322    1,929.763      1          1
1       116,741
## Zone_Assessor            5,442     3.829        2.554        1          2
6          9
## fpla                     5,442     0.000        0.000        0          0
0          0
## WGS1984Y                 5,442     36.140       0.050      36.029     36.101     3
6.180     36.243
## -----------------------------------------------------------------------------------
------------------
```



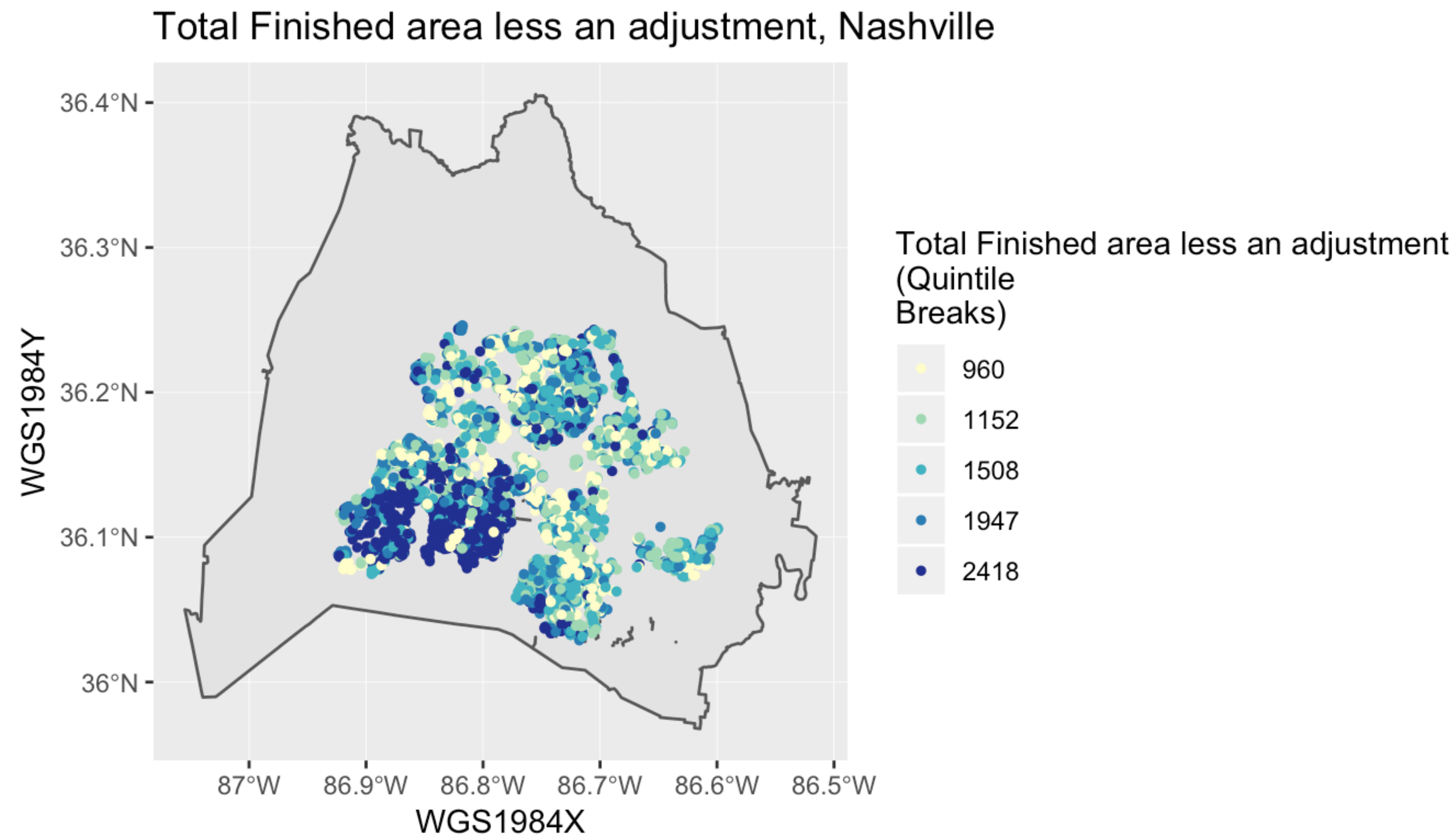
Home sale prices in Nashville range from about 2000 dollars to about 700,000 dollars, with a mean of about 290,000 dollars.



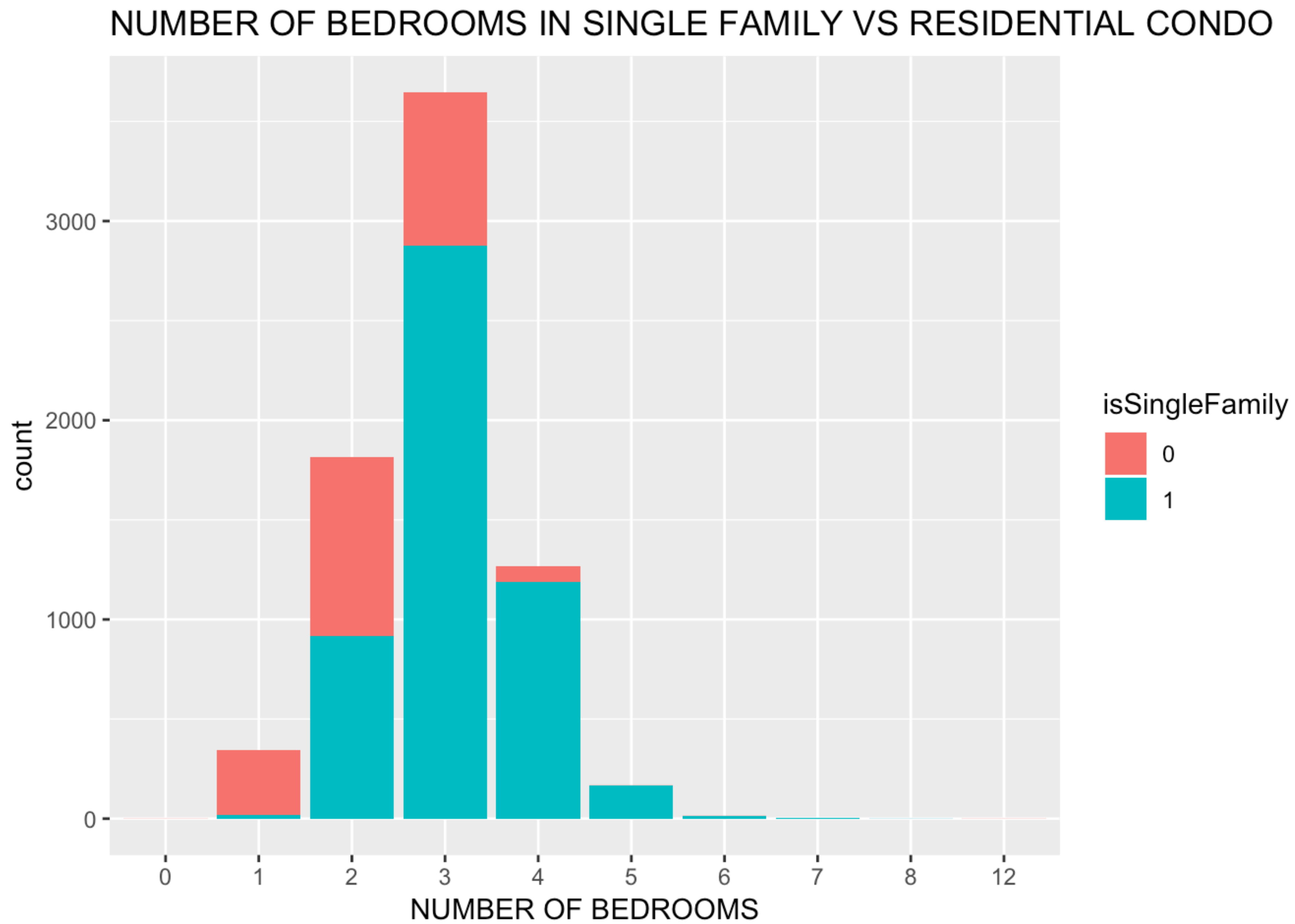
The map above shows the square feet of finished area inside houses across Nashville.



The map above shows the neighborhood accessor across Nashville.



This map shows the total finished area of a building unit across Nashville.



A map representing the number of bedroom in Single and residential condo houses in Nashville. 0 represent residential condos and 1 represents single family homes.

METHODS

In order to mine the data for the most powerful correlations, we used feature-engeering techniques to craft unique features or variables from our data. We hypothesized that there is a qualitative difference in sales price between buildings which are constructed from bricks and those with wooden frames. We also hypothesized that there is a qualitative difference in sales price between buildings which are built more recently (2018) than those that are built a couple years ago. Moreover, we also hypothesized that there is a qualitative difference in sales price between single family homes and residential condos. Finally, we hypothesized that house sales prices are likely to be different based on the number of bathrooms and bedrooms in the house. We started by subsetting variable observations and created a dummy variable that equals 1 for the variables we are using as a base and 0 otherwise. For example, we used a value of 1 for single family homes and a value of 0 for residential condos.

Modeling - In Sample(Training Set) and Prediction Results

Here we built a model with all the data including the dummy variables we created earlier using Nashville data to get regression coefficients. After this, we performed in-sample and out-of-sample predictions by training our data using the sales price that are available to us (test=0), and then randomly split the data to contain 75%

of all observations. We also had a test data set (test=1) of unseen house price data. Finally, we predicted for the unseen housing prices using the training set. The results for this procedure are shown below:

```
##
## Summary Statistics of Training Set
## =======================================================
##                               Dependent variable:
##                           ---------------------------
##                                    SalePrice
## -------------------------------------------------------
## kenID                               -1.663
##                                     (1.632)
##
## LocationZip                      -6,609.127***
##                                   (1,108.616)
##
## CouncilDistrict                  2,357.541***
##                                    (781.754)
##
## CensusBlock                        7.787***
##                                     (2.693)
##
## accountnumber_property             0.799***
##                                     (0.111)
##
## yearbuilt_building               -939.477***
##                                    (248.636)
##
## NeighborhoodAssessor              13.879***
##                                     (3.257)
##
## sf_fin_less_ifla_less_bfin        89.164***
##                                    (13.250)
##
## sf_sketched                       43.098***
##                                     (7.584)
##
## Zone_Assessor                    -6,245.781**
##                                   (2,732.669)
##
## baths                          -191,390.400***
##                                  (40,475.940)
##
## WGS1984X                       -626,247.000***
##                                  (90,749.370)
##
## WGS1984Y                      1,272,306.000***
##                                 (196,080.100)
##
## NUMBER_BATHS1                    199,466.500
```

```
##                                      (163,358.900)
##
## NUMBER_BATHS2                         373,959.800**
##                                      (152,191.200)
##
## NUMBER_BATHS3                        614,151.300***
##                                      (151,338.300)
##
## NUMBER_BATHS4                      1,000,339.000***
##                                      (162,112.800)
##
## NUMBER_BATHS5                        969,136.500***
##                                      (188,597.600)
##
## NUMBER_BATHS6                      1,272,630.000***
##                                      (230,608.900)
##
## NUMBER_BATHS7                      1,586,857.000***
##                                      (297,134.100)
##
## NUMBER_BATHS8
##
##
## effyearbuilt_building                  1,517.081***
##                                        (285.204)
##
## NUM_BEDROOMS2                         70,129.630***
##                                      (21,942.880)
##
## NUM_BEDROOMS3                         70,531.010***
##                                      (24,367.310)
##
## NUM_BEDROOMS4                         69,515.490**
##                                      (29,055.000)
##
## NUM_BEDROOMS5                          69,857.380
##                                      (45,302.450)
##
## NUM_BEDROOMS6                          108,187.900
##                                      (104,658.500)
##
## NUM_BEDROOMS7                          -49,324.330
##                                      (188,024.400)
##
## NUM_BEDROOMS8                         654,196.100**
##                                      (322,504.400)
##
## NUM_BEDROOMS12                        -161,096.900
##                                      (258,631.500)
##
```

```
## isSingleFamily1                          -15,881.170
##                                         (14,886.270)
##
## Constant                               -143,979,123.000
##                                        (116,800,524.000)
##
## -----------------------------------------------------------
## Observations                                 3,689
## R2                                           0.371
## Adjusted R2                                  0.366
## Residual Std. Error                 257,203.700 (df = 3658)
## F Statistic                       71.825*** (df = 30; 3658)
## ===========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```
##   observedSales predictedSales    error absError percentAbsError
## 1             0       853694.7 853694.7 853694.7             Inf
## 2             0       169254.0 169254.0 169254.0             Inf
## 3             0       164526.7 164526.7 164526.7             Inf
## 4             0       726271.0 726271.0 726271.0             Inf
## 5             0       378012.4 378012.4 378012.4             Inf
## 6             0       137008.9 137008.9 137008.9             Inf
```

```
mean(regPred$absError)
```

```
## [1] 373846.9
```

Cross Validation

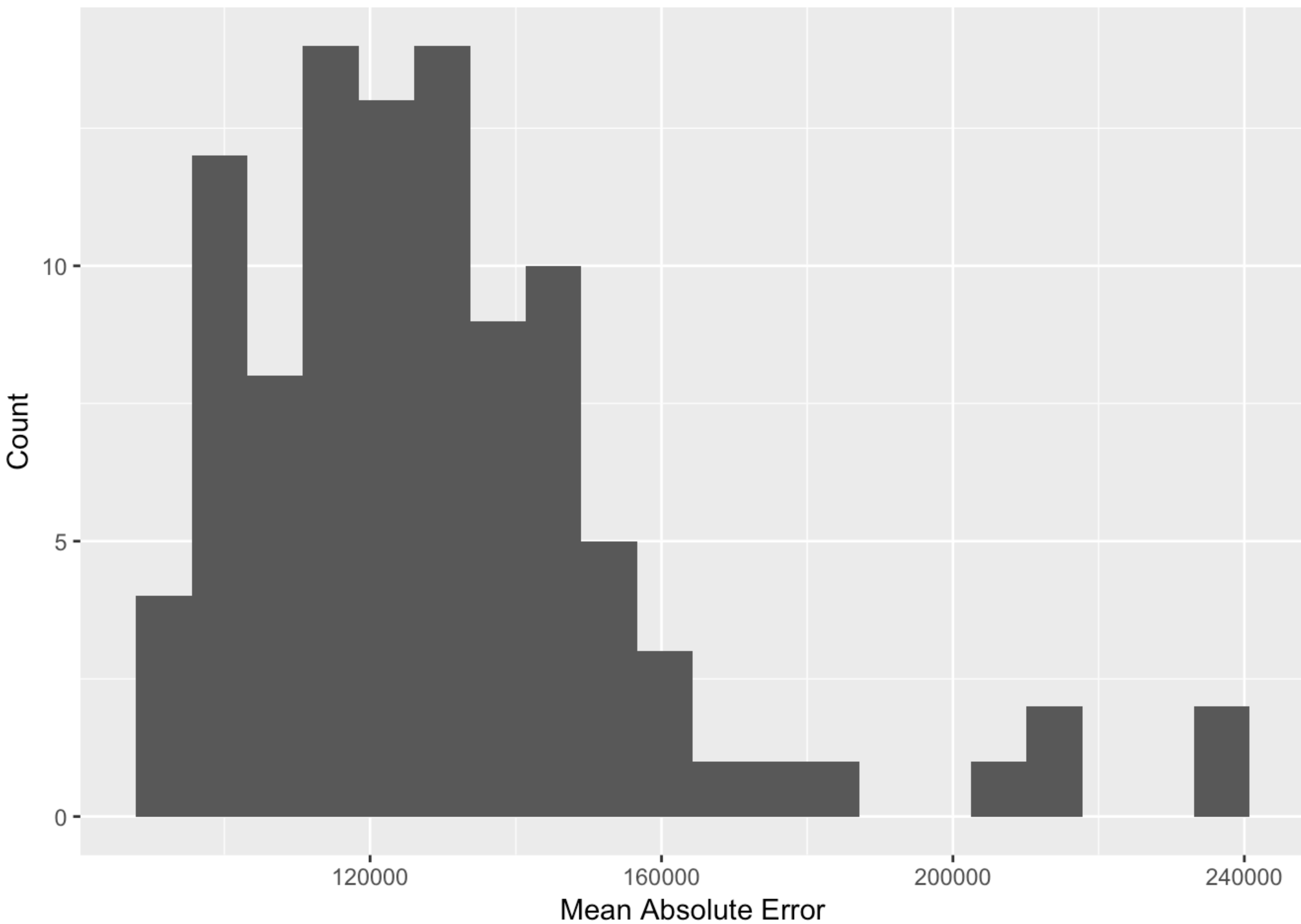
```
##
## % Error: Unrecognized object type.
```

```
## Linear Regression
##
## 5442 samples
##   14 predictor
##
## No pre-processing
## Resampling: Cross-Validated (100 fold)
## Summary of sample sizes: 5387, 5389, 5386, 5387, 5387, 5386, ...
## Resampling results:
##
##   RMSE    Rsquared   MAE
##   232061  0.4029406  130054.3
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
##
## Summary Statistics of lmFit
## ==========================================================================================
===
## Statistic  N     Mean       St. Dev.       Min       Pctl(25)    Pctl(75)       Max
## ------------------------------------------------------------------------------------------
---
## RMSE      100 232,061.000 148,078.800 119,786.000 161,494.600 249,872.900 883,323.
100
## Rsquared  100    0.403       0.175       0.025       0.269       0.533       0.836
## MAE       100 130,054.300 28,796.860  90,497.340  111,055.800 141,837.100 235,679.
900
## ------------------------------------------------------------------------------------------
---
```

```
## [1] 28796.86
```

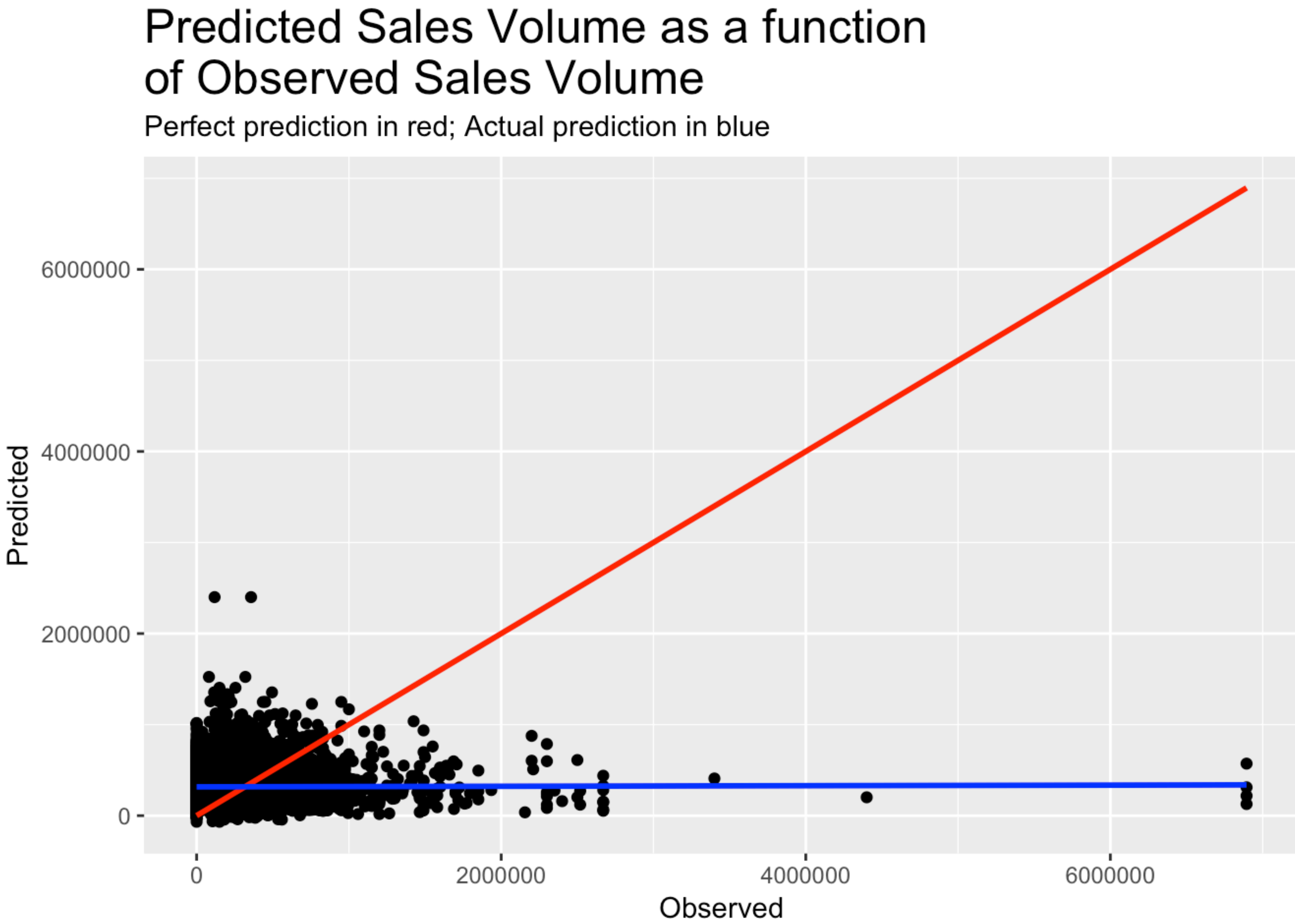
HISTOGRAM OF THE CROSS-VALIDATION



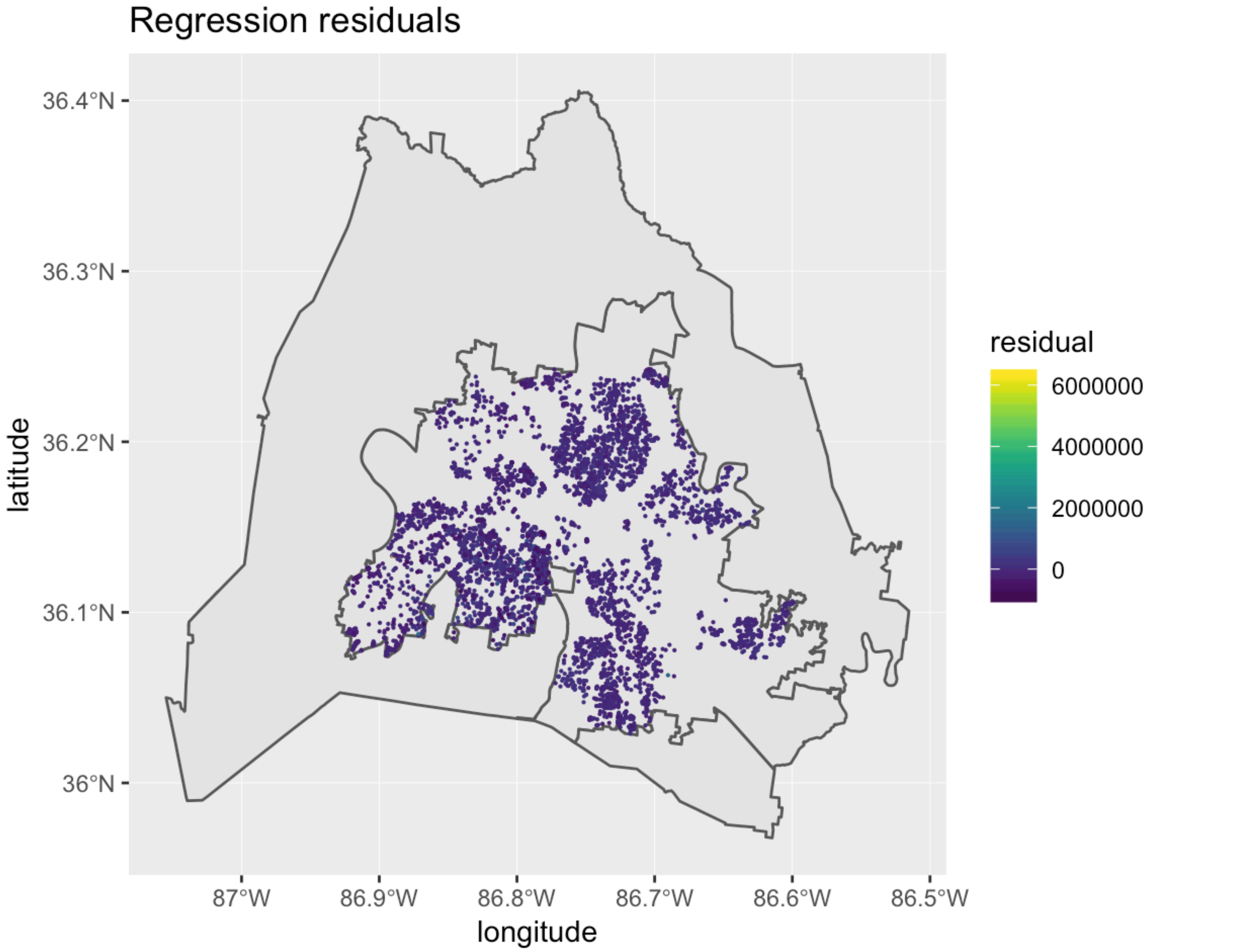
Two of the folds had a MAE that was much different the others. Looks like 5 folds are sufficient in this case. It also seems like much of the mean error is concentrated around 120k. There may be an overfit in the model.

PLOT OF PREDICTED PRICES AS A FUNCTION OF OBSERVED PRICES

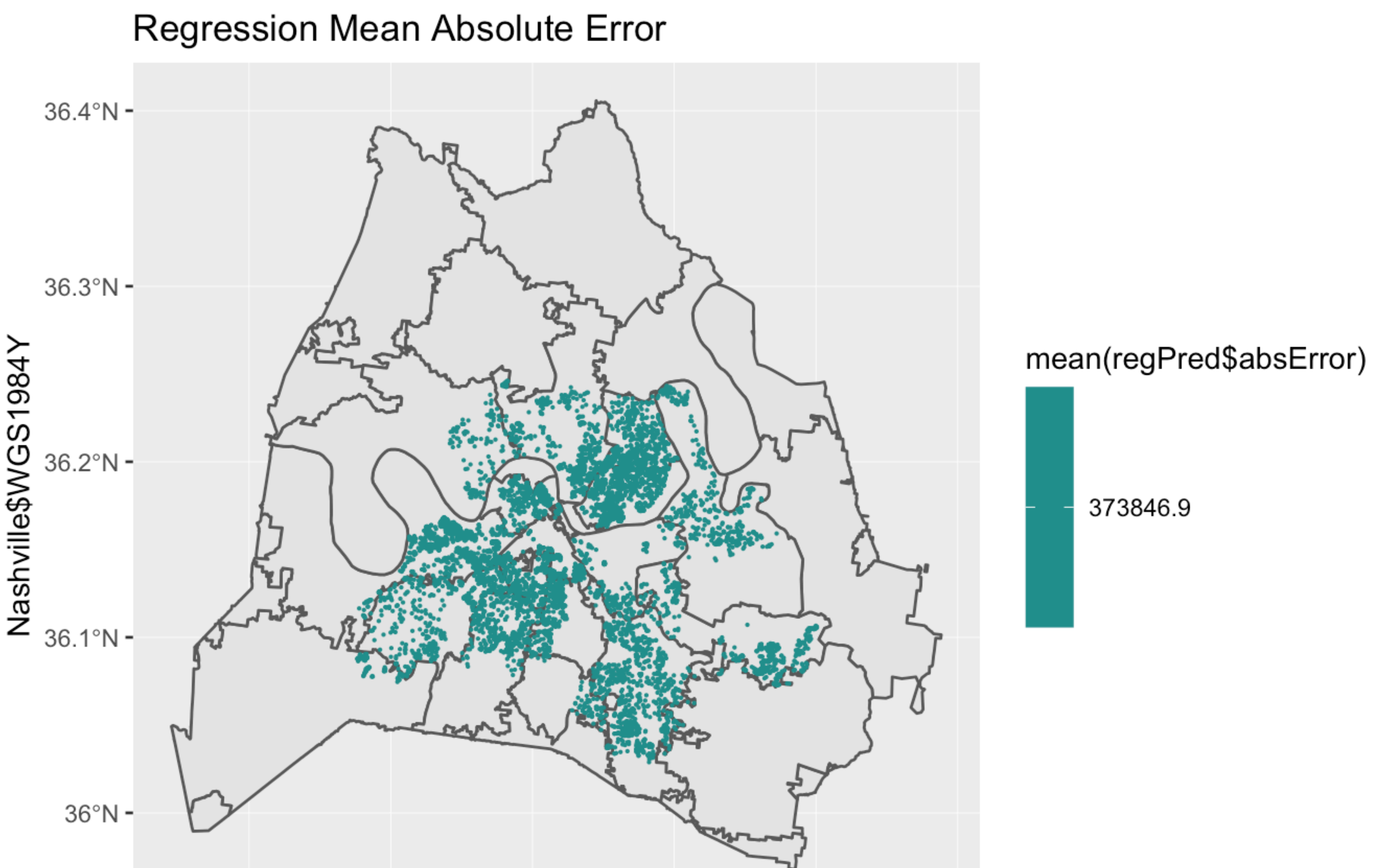
```
## Warning in cbind(Nashville$SalePrice, reg$fitted.values): number of rows of
## result is not a multiple of vector length (arg 2)
```

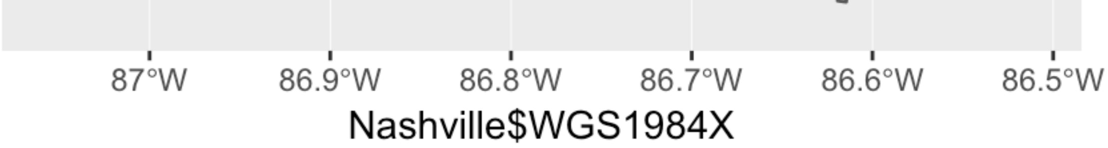


Moran's I and Residuals Mapping for the test set

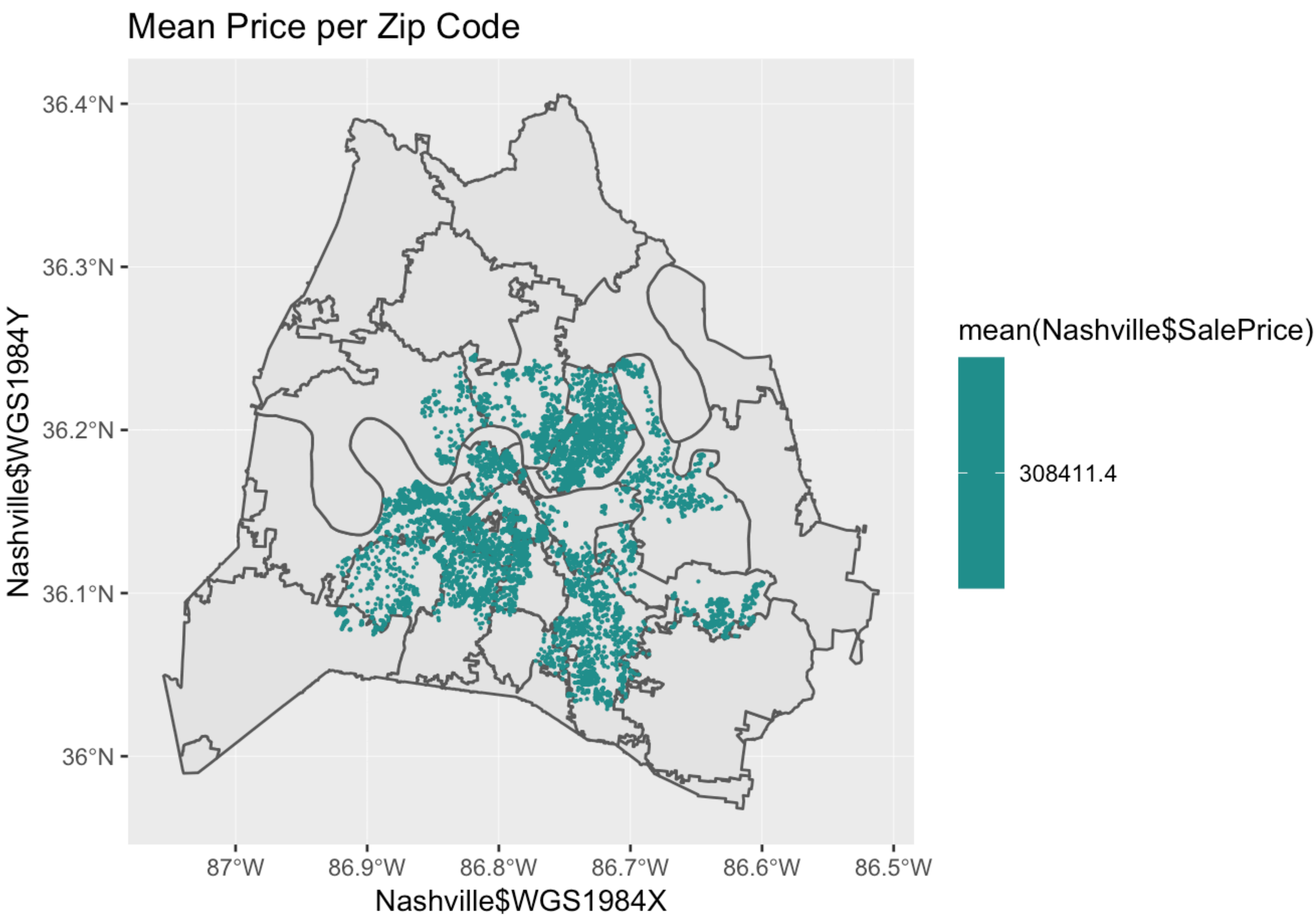


MEAN ABSOLUTE ERROR(MAE) by ZIP CODE





Since the original sales prices of the test was unavailable, we were only able to calculate the mean absolute error of our prediction. We didn't have any available data to compare our predictions to in order to calculate the mean absolute percentage error of our prediction. From above, we witnessed a mean absolute error of about 360,000 dollars.



This map shows the mean home sale prices by ZIP code. The average sales price for houses across all ZIP codes are about 308,000 dollars.

DISCUSSION

One of our most interesting variables was the year in which the house was built. This variable seeks to shows whether the age of a house can have significant effects on its value. Other variables we used include the construction type of the house, the number of bathrooms in the house, the number of bedrooms in the house, and the square footage of finished area of the house. We hypothesized that the construction type of a house (Brick or Frame) can affect the price of the house. Also, the number of bedrooms and bathrooms a house has can also affect its value. For example, a 3 bedroom house is more likely to cost more than a one bedroom house. Furtheremore, a house with more bathroom and a larger furnished area is more likely to cost more than an unfurnished house with fewer bathrooms or bedrooms. Overall, we conclude that our regression model is not very effective at making housing price predictions. There are more robust models that can capture more details in the dataset or offer improved abilities to handle spatial data than OLS. From our maps, higher housing prices are concentrated in the Northeast part of Nashville and the lowest housing prices are scattered throughout Nashville. Our prediction for housing prices in Nashville was off by 360,000 dollars. The model predicted particularly well in the Northeast part of Nashville.

CONCLUSION

We would not recommend our model to Zillow. Despite efforts to increase its generalizability, it is still lacking. Moreover, there are more robust models available that can possibly be used to provide better predictions that are restricted by criteria for this project. In the future, we will continue to improve the accuracy and generalizability of our model.

DATA DICTIONARY

Fields in the Multiple Record and Single Record per Parcel Data

1. kenID - and ID I created.
2. ParcelId_property The Parcel ID; the parcel identificiation string. The ParcelID format is described in the table at the end of this document.
3. UserAccount The same as ParcelId_Property without spaces and punctuation. This field can be used to join to the UserAccount field in the NameAddressLegal data. (Separate data that can be downloaded from the Assessor's FTP site.) It can also be used to join to the SubAreas and YardItems tables.
4. accountnumber_property Internal identification used in the Assessor's office.
5. Card The sequence number of a building or the number of buildings on a parcel. See the Multiple Records per Parcel and Single Records per Parcel topics earlier in this document for an explanation of how the multiple and single record data works.
6. District Identifies which tax Levy area the parcel is in. For example General services district (GSD) or Urban services district (USD). It may be a satellite city like Goodlettsville or Belle Meade, etc.
7. LandUseDescription General land use.
8. LandUseFullDescription Land use.
9. NeighborhoodAssessor The neighborhood code (NBC) is used by the Assessor's office to group similar properties for the purpose of determining property value. It is not associated with common subdivisions or neighborhoods.
10. Acrage Acres of land.
11. Land_Appraisal Value of the land.
12. Improvements_Appraisal Value of the improvements like buildings and yard items.
13. Total_Appraisal Total value of land and improvements.
14. Land_Assessment Assessed value of the land.
15. Improvements_Assessment Assessed value of improvements like buildings and yard items.
16. Total_Assessment Total assessed value of land and improvements.
17. Building_Type Type of building.
18. Story_Height Number of stories of the building.
19. Exterior_Wall Exterior wall type.
20. Grade Rating of building grade.
21. Frame Building frame type.

22. yearbuilt_building Actual year the building was built.
23. avgHtfl_building Average ceiling height in feet for commercial property.
24. roomsunits_building Number of rooms in the building.
25. bedroomsunits_building Number of bedrooms.
26. units_building If multi family like a duplex, the number of units.
27. sf_finished Square feet of finished area.
28. sf_ifla The adjustment amount to Finished area. (When the finish area produced by the computer sketch routine is not precise enough, an adjustment is made. It is implied that it is a negative number.)
29. sf_bsmt Square feet of the basement if any.
30. sf_bsmt_fin Square feet of the basement that is finished.
31. sf_finished_less_ifla Total Finished area less an adjustment, see sf_ifla.
32. sf_fin_less_ifla_less_bfin The effective finished area with any basement finish removed.
33. sf_sketched The gross footage not including the adjustment amount.
34. ac_sfyi Central air. 0 = no central air ; 1 = central air (Residential)
35. Phys_Depreciation Building condition
36. NumofUnits_land Units used for appraisal, may be sq footage or acres or front footage or rental units or sites or others. See the Land_Unit_Type field.
37. Zone_Assessor Zones (jurisdictions) These are 9 large areas of the county used by the appraisal staff to coordinate appraisal teams.
38. Land_Unit_Type Type of units associated with NumofUnits_land.
39. baths Number of baths
40. halfbaths Number of halfbaths
41. HeatingType Method of heating.
42. Fixtures Estimated plumbing fixtures
43. Foundation Type of foundation.
44. notheated This field is no longer supported.
45. fpla This field is no longer supported. Instead you can find fireplaces by looking at the YardItems data.
46. test - the dataset that I will test you on.