MUSA508-Leah-Palak

Code

Leah Shapiro & Palak Agarwal

October 16th, 2020

Introduction

Zillow publishes "Zestimate" valuations for homes across the U.S. using an algorithm that draws on data from county and tax assessor records, multiple listing services, brokerages, and homeowner submissions. This project sought to improve Zillow's housing market predictions for Miami and Miami Beach.

This endeavor was challenging for several reasons. First, we relied exlusively on ordinary least squares linear modeling. A more sophisted algorithm, such as random forest, may produce a better model. Second, the project used only open source data. This restriction was particularly challenging given the lack of open source data available for Miami Beach as compared to Miami.

We conceptualized our model using the hedonic model, which predicts home prices by summing the value of its constituent parts. Our model includes three types of variables: internal characteristics of houses, amenities/public services, and spatial structures.

Despite our hardwork and dedication to the project, we failed to create a successful model. With a mean absolute error of over $350,000 in our model, we do not anticipate securing a contract with Zillow anytime soon.

Set Up

Code

Themes, Palettes, and Quantile Break Functions

Code

Additional Functions

Many of the features in this model were engineered using two functions: **Nearest neighbor** and **Multiple ring buffer.** The nearest neighbor function finds the average distance from the measuring feature to the measured feature. The function requires 3 input variables - the dependent feature to measure from, the features to measure to, and the number of features. For example, a nearest neighbor feature could measure the average distance from each house to its three closest public parks.

The multiple ring buffer creates concentric rings around point features. This can be used, for example, to determine how many houses are within a .5 mile radius of a metro stop.

Code

Data Wrangling

Data

Internal Characteristics

First we looked at the internal characteristics of the houses. We considered features associated with the number of bedrooms, number of bathrooms, living square feet, actual square feet, year built, effective year built, stories, pools, jacuzzis, fences, patios, docks, and elevators. For number of bedrooms and year built, we tested both discrete and categorical variables in our model.

Amenities/Public services

Proximity to public services and amenities can add value to houses. To get this data we made use of Open Street map as well as other open data portals. We considered the following public services and amenities when building our model:
* **Transit Stops**
* **Restaurants, cafes and bars**
* **Schools**
* **School Attendance Areas** : This information was not available for Miami Beach
* **Parks**
* **Places of worship**
* **Car Parks**
* **Work Centers**
* **Hospitals**

Spatial Structures

In creating our model, we experimented with spatial features at various scales:
* **Neighborhood**: We were not able to find a neighborhood shapefile for Miami Beach, so we use Miami Beach's municipality borders as a proxy. One house was initially not assigned a neighborhood, so we imputed the value "Haynesworth" based on the house's location on the map.
* **Zip Code**
* **City**
* **Shoreline**
* **Sale Price of nearest 5 houses**
* **Median Rent within Census Tract**
* **Median Income within Census Tract**: The mean median income was used for missing values
* **Racial Composition within Census Tract**: Like many cities within the U.S., Miami is highly segregated.

Variable Descriptions

The following variables were ultimately included in our model.

Numeric Internal Features

Internal_Characteristics	Description	Mean	Median	Max	Min	Standard_Deviation
LivingSqFt	Living Sq Ft	2011.431059	1632.00	18006	288	1369.2473426
EffectiveYearBuilt	Age adjusted for Renovation or neglect	1973.362832	1975.00	2019	1905	26.5641355
SalePrice	Last sale price	730835.912075	332500.00	27750000	0	1822987.7886706
ActualSqFt	Updated Living SqFt	2367.747645	1884.00	20192	388	1732.3170408
LotSize	Size of the Lot	7657.781347	6693.75	80664	1250	4401.4024951
Bath	Number of Bathrooms	2.109620	2.00	12	0	1.2986293
Stories	Number of Stories	1.206680	1.00	4	0	0.4420643
Bed	Number of Bedrooms	3.035398	3.00	13	0	1.0919479

Summary Statistics of Internal Characteristics
Table 1.1

Categorical Internal Features

Amenities	Description	Count
Docks	Presence of a dock	684
Luxury Pool	Presence of a luxury pool	7
Whirlpool	Presence of a whirlpool	40
Elevators	Presence of an elevator	364

Count of Categorical Internal Features
Table 1.2

Amenities	Description	Count
Fences	Presence of fences	3
Pool Type 1	Presence of a 8ft pool	7
Pool Type 2	Presence of a 2-4ft pool	11
Pool Type 3	Presence of a 3-6ft pool	2
Pool Type 4	Presence of a 3-8ft pool	67

Count of Categorical Internal Features
Table 1.2

Ammenities/Public Services

Amenities	Description	Mean	Median	Max	Min	Standard_Deviation
Dist.Metro	Distance to the closest Metro stop	2.1307082	1.416819	7.054678	0.0372621	1.7508759
Dist.Restaurants	Average distance to the 5 closest restaurants	0.6615725	0.626169	2.112204	0.0323645	0.3658257
Dist.School	Distance to the closest school	1228.0921477	968.478667	7653.552964	41.1676081	1128.9394913
Dist.Worship	Distance to the closest place of worship	2736.3433302	2280.558722	10085.689827	96.9273545	1752.9820224
Dist.Parking	Average distance to two closest parking spots	2299.7288593	1836.084798	10552.669874	75.3127690	1832.8067300
Dist.WorkCenter	Average distance to ten work centers	2309.9128575	1929.020305	7962.497260	388.8800188	1367.7064751
Dist.Hospital	Distance to the closest hospital	7138.1669146	6460.865236	20188.849300	88.8705949	4152.4067817

Summary Statistics of Amenities and public services
Table 1.3

Spatial Structures

Amenities	Description	Mean	Median	Max	Min	Standard_Deviation
Dist.Shore	Distance to coast	-0.8669539	-0.7315004	1.690956	-8.300917	1.614161
Lag_Price	Average price of the 5 closest houses	607303.7567799	343800.0000000	9733000.000000	0.000000	910008.467836
Median_Rent	Median Rent according to the census tract	1133.0380195	1061.0000000	2271.000000	245.000000	428.221642

Summary Statistics of Spatail Structures
Table 1.4

Amenities	Description	Mean	Median	Max	Min	Standard_Deviation
Median_Income	Median Income according to the census tract	59223.6838938	39821.0000000	172750.000000	14699.000000	44963.056660
Pct_white	Percentage of White residents	74.0446888	89.8468787	98.543548	5.753497	29.387571
Pct_Hispanic	Percentage of Hispanic residents	57.7175499	48.0920478	102.200721	8.747937	30.364491

Summary Statistics of Spatail Structures
Table 1.4

Methods

Models can be judged by their accuracy or their generalizability. The accuracy of a model reflects how close predicted values are to the observed values. This can be measured by the adjusted R-squared, which tells us how much of the variation in the dependent variable, house prices, is explained by the independent variables in the model.

FOr the purpose of this project, we attempted to maximize our model's generalizability. A generalizable model is one that can successfully predict on new data. To determine our models' generalizability, we randomly split our data into a training set and a testing set. We then performed a stepwise regression, adding one dependent variable at a time and determining whether the mean absolute error (MAE) of the model went up or down.

Correlation Matrix

The presence of highly correlated, or colinear, variables in a model can lead to unwanted redundancy. We used the correlation matrix below to determine which variables are colinear.

Code

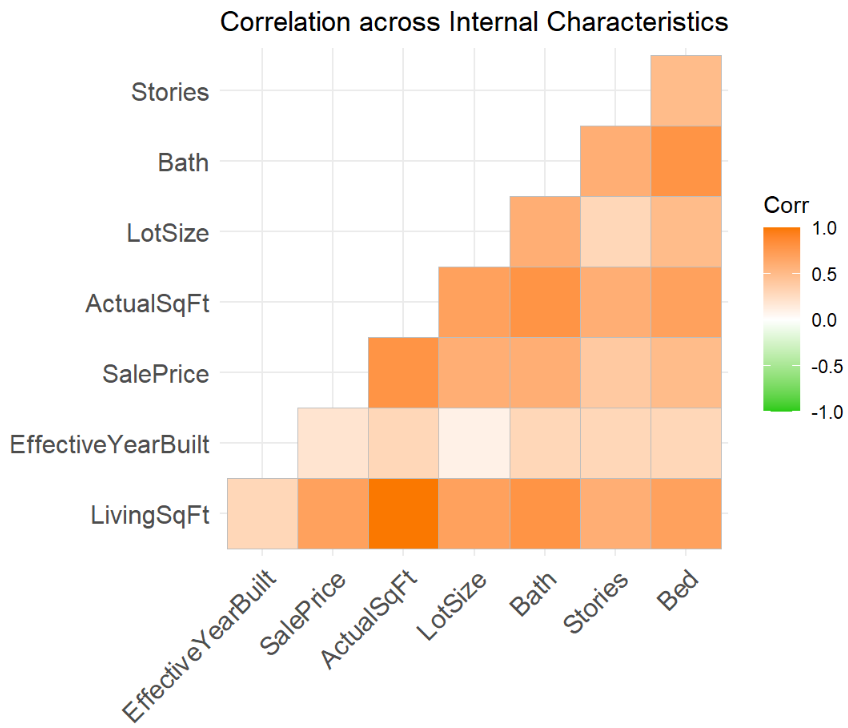


Figure 1.1

Code

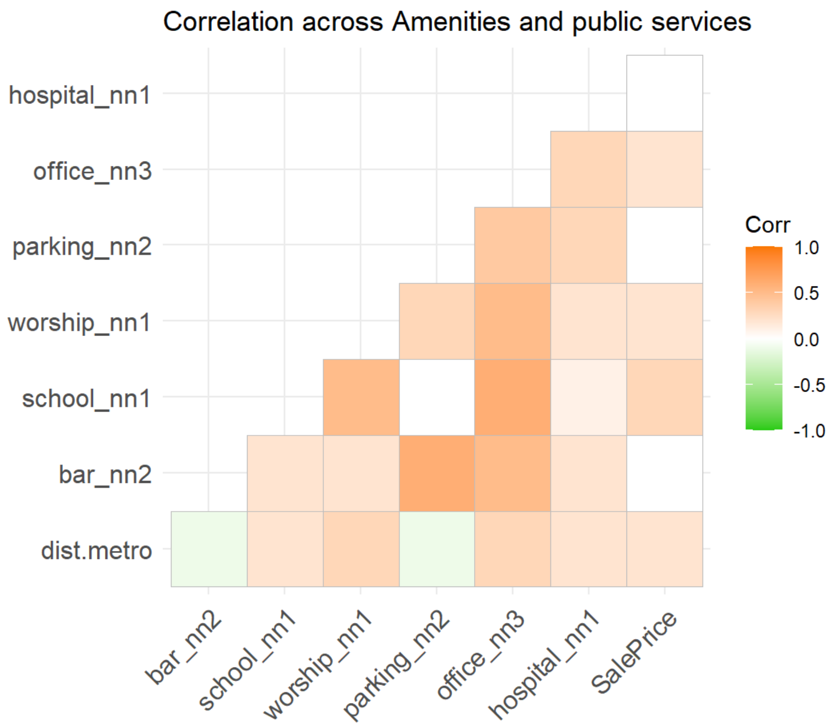


Figure 1.2

Code

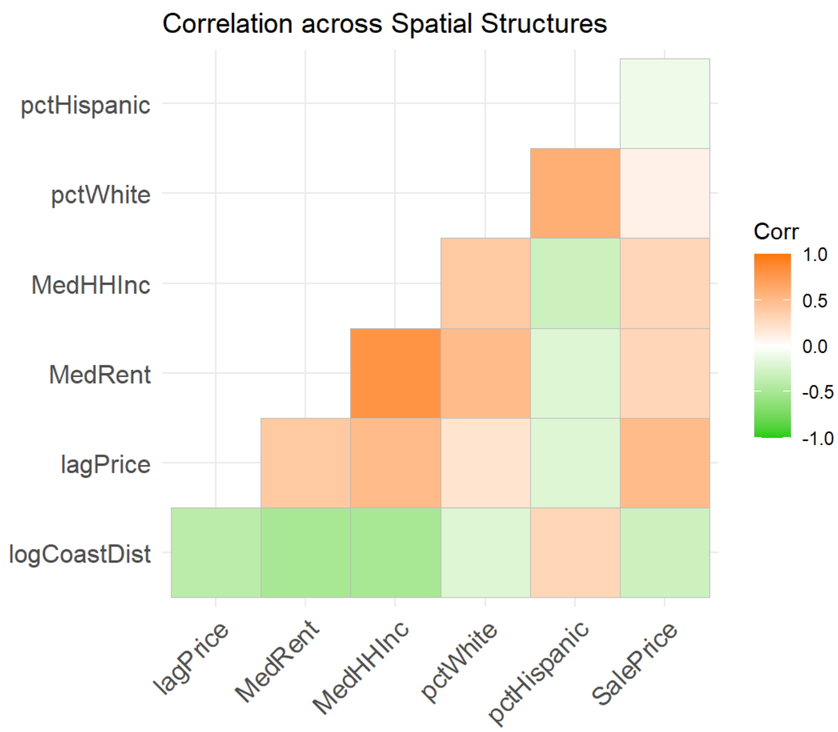


Figure 1.3

Code

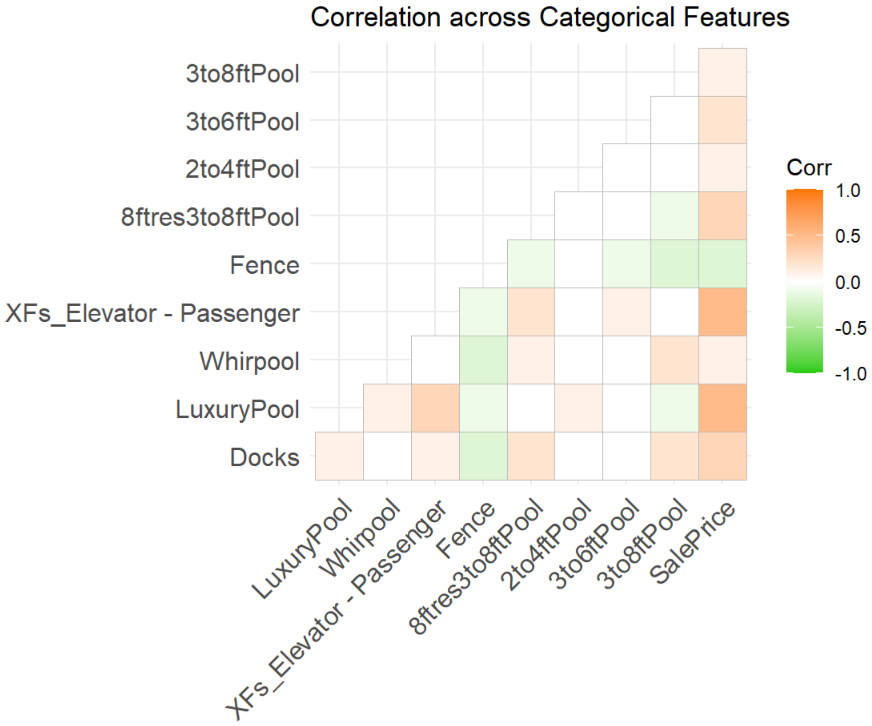
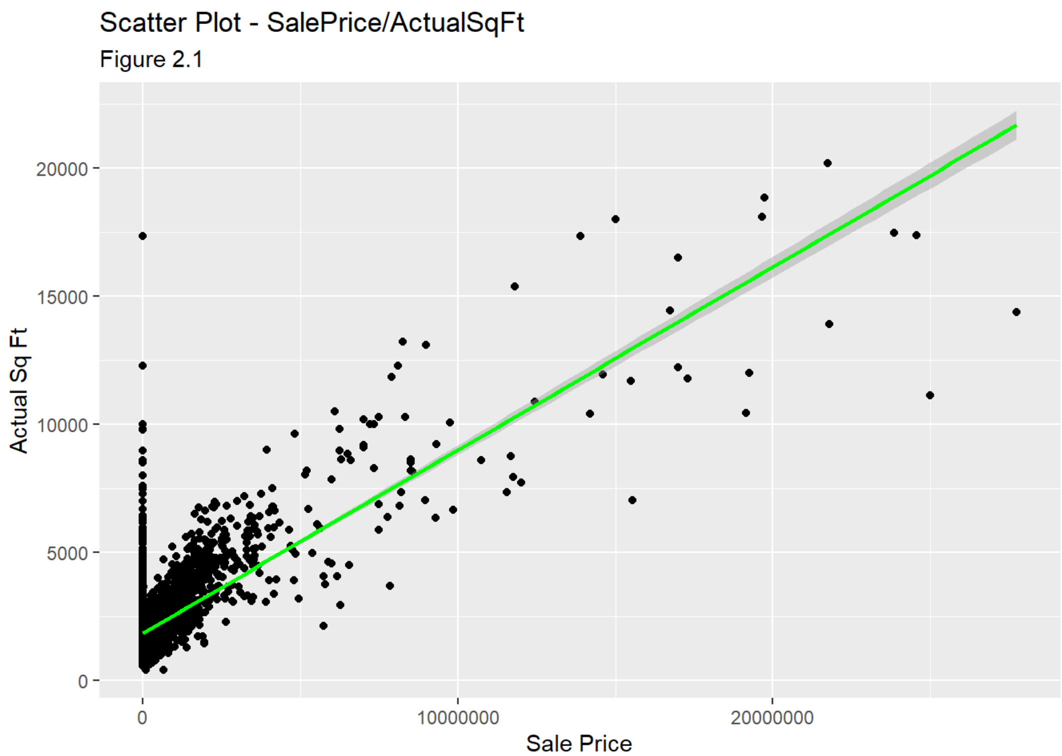


Figure 1.4

Home Price Correlation Scatter Plots

To help us identify which variables might be most important to include in our model, we generated the following scatter plots visualizing the correlation between independent variables and home sale prices.

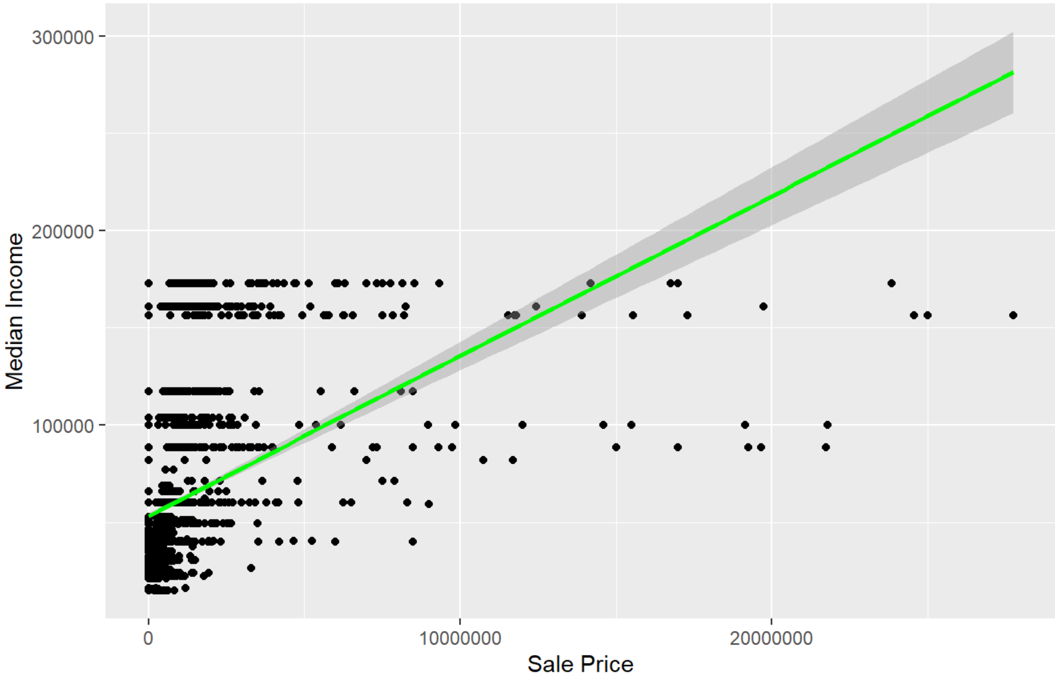
Code



Code

Scatter Plot - SalePrice/Median Income

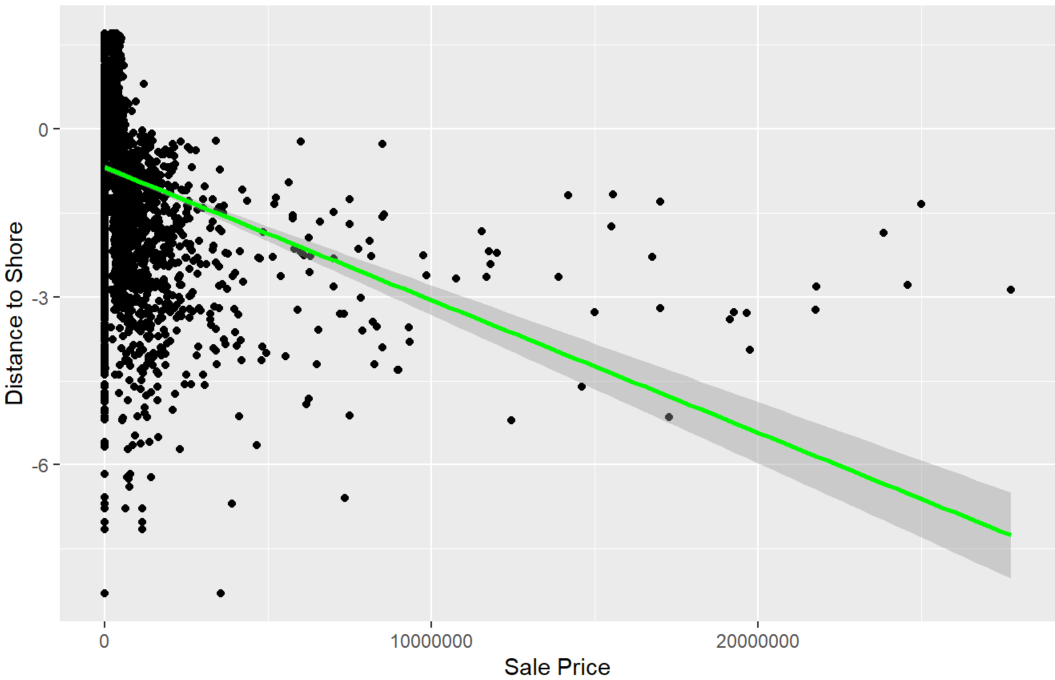
Figure 2.2



Code

Scatter Plot - SalePrice/Shore Distance

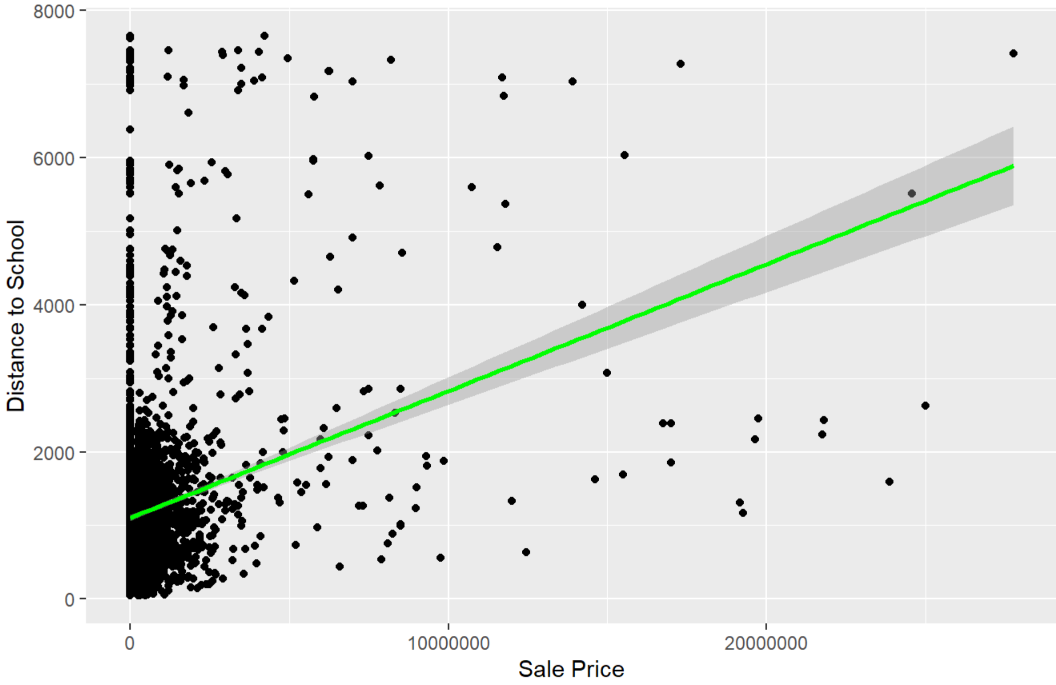
Figure 2.3



Code

Scatter Plot - SalePrice/School

Figure 2.4



Code

Scatter Plot - SalePrice/Place of Worhship

Figure 2.5

Maps

The following map visualizes sale prices from the data set we used to train our model. The clustering of sale prices gave us an idea about certain spatial relationships which we used to identify our independent variables.

Sale Price Map

Code

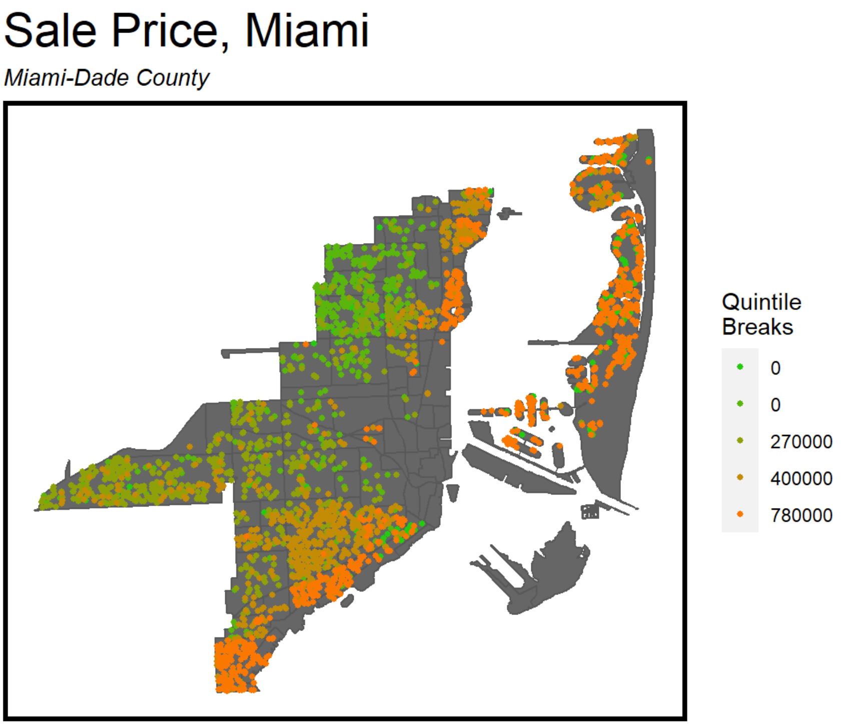


Figure 3.1

Independent Variable Maps

The first two maps represent the racial segregation in Miami and Miami Beach. While we were hesitant to use racial data in our model,including the spatial structure of segregation improved our model's generalizability. The third map shows the location of the metro stops. Although we didn't find a strong correlation between distance to metro stops and home sale prices, we still found that inclusion of this variable improved our model.

Code

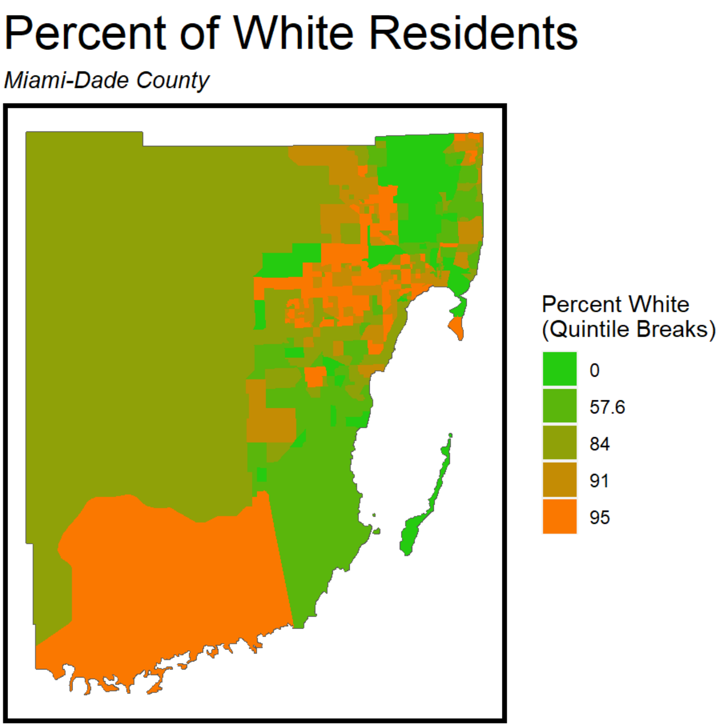


Figure 3.2

Code

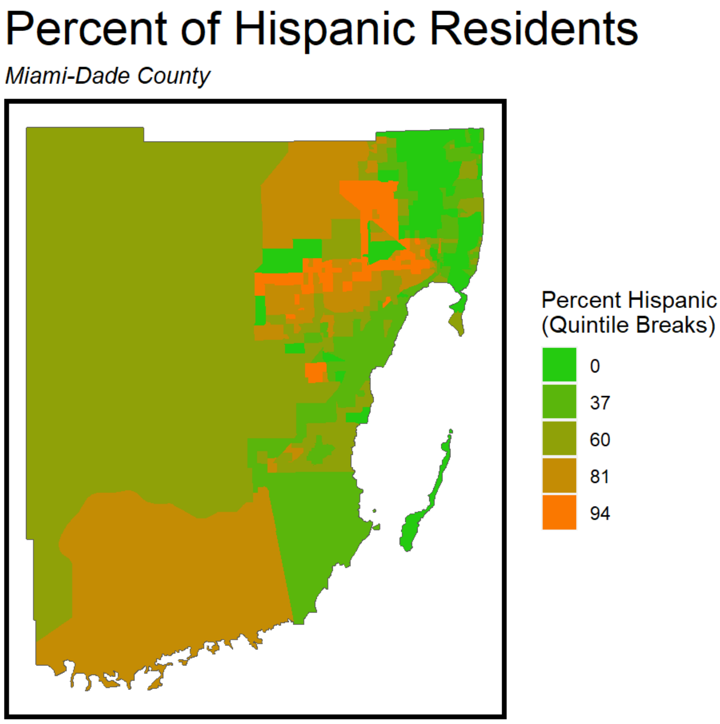


Figure 3.3

Code

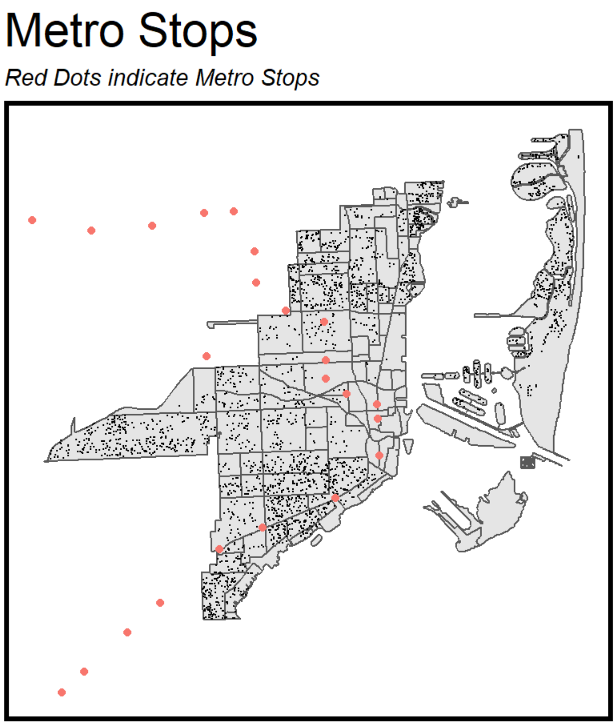


Figure 3.4

Results

OLS Regression

Split dataframe into training and testing datasets

Regression and predicting Saleprice for test dataset

The output below provides a summary of our model. We experimented with different combinations of the independent variable to maximize generalizability as measured by MAE and accuracy measured by adjusted R sqaured.

Observations	1691
Dependent variable	SalePrice
Type	OLS linear regression

F(153,1537)	72.83
R²	0.88
Adj. R²	0.87

	Est.	S.E.	t val.	p
(Intercept)	-4694950.77	2203133.71	-2.13	0.03
ActualSqFt	581.61	32.27	18.02	0.00
Neighbourhood_nameAllapattah Industrial District	62671.24	1649271.41	0.04	0.97
Neighbourhood_nameAuburndale	169662.24	1503488.78	0.11	0.91
Neighbourhood_nameBay Heights	-1360099.30	1392359.23	-0.98	0.33
Neighbourhood_nameBaypoint	-64691.79	1309668.45	-0.05	0.96
Neighbourhood_nameBayside	607408.15	1291305.56	0.47	0.64
Neighbourhood_nameBelle Island	-291922.83	1340222.88	-0.22	0.83
Neighbourhood_nameBelle Meade	519902.91	1286935.13	0.40	0.69
Neighbourhood_nameBelle Meade West	633772.42	1302678.05	0.49	0.63
Neighbourhood_nameBird Grove East	102492.84	1388899.88	0.07	0.94
Neighbourhood_nameBird Grove West	-330691.05	1327849.63	-0.25	0.80
Neighbourhood_nameBiscayne Island	2775760.43	1494410.16	1.86	0.06
Neighbourhood_nameBiscayne Plaza	-589886.21	1516906.37	-0.39	0.70
Neighbourhood_nameBrentwood	784443.80	1721814.89	0.46	0.65
Neighbourhood_nameBuena Vista Heights	872194.56	1671131.59	0.52	0.60
Neighbourhood_nameBuena Vista West	617207.79	1657902.14	0.37	0.71
Neighbourhood_nameCitrus Grove	-84937.74	1427559.69	-0.06	0.95
Neighbourhood_nameCivic Center	-42208.33	1609888.34	-0.03	0.98
Neighbourhood_nameCoral Gate	-210744.99	1438442.99	-0.15	0.88
Neighbourhood_nameCurtis Park	-149778.98	1596102.91	-0.09	0.93
Neighbourhood_nameDouglas Park	-472606.75	1365931.30	-0.35	0.73
Neighbourhood_nameEast Grove	-723089.44	1375626.57	-0.53	0.60
Neighbourhood_nameEast Little Havana	-72248.08	1397254.53	-0.05	0.96
Neighbourhood_nameEdgewater	2541144.93	1817437.39	1.40	0.16
Neighbourhood_nameEdison	710044.60	1670527.06	0.43	0.67
Neighbourhood_nameFair Isle	-61028.11	1391666.46	-0.04	0.97
Neighbourhood_nameFlagami	-263495.30	1530818.27	-0.17	0.86
Neighbourhood_nameFlora Park	-1081377.98	1447365.48	-0.75	0.46
Neighbourhood_nameGrove Center	-544813.94	1377003.22	-0.40	0.69
Neighbourhood_nameHadley Park	-709069.42	1459757.33	-0.49	0.63
Neighbourhood_nameHaynesworth	979964.10	1511769.64	0.65	0.52
Neighbourhood_nameHighland Park	283235.17	1559611.84	0.18	0.86
Neighbourhood_nameHistoric Buena Vista East	937829.57	1688324.20	0.56	0.58
Neighbourhood_nameKing Heights	-1109972.97	1541619.61	-0.72	0.47
Neighbourhood_nameLa Pastorita	407111.05	1471415.66	0.28	0.78

Standard errors: OLS

	Est.	S.E.	t val.	p
Neighbourhood_nameLatin Quarter	51970.14	1446876.87	0.04	0.97
Neighbourhood_nameLe Jeune Gardens	-781291.37	1713141.99	-0.46	0.65
Neighbourhood_nameLemon City/Little Haiti	963262.12	1684441.76	0.57	0.57
Neighbourhood_nameLiberty Square	373258.79	1894792.60	0.20	0.84
Neighbourhood_nameLittle River Central	143575.69	987200.60	0.15	0.88
Neighbourhood_nameLittle River Gardens	813819.31	1200249.06	0.68	0.50
Neighbourhood_nameMelrose	-455789.87	1399971.88	-0.33	0.74
Neighbourhood_nameMiami Avenue	-480748.95	1395067.63	-0.34	0.73
Neighbourhood_nameMIAMI.BEACH.6	1932899.69	1390467.57	1.39	0.16
Neighbourhood_nameMIAMI.BEACH.8	1354216.84	1371972.40	0.99	0.32
Neighbourhood_nameMorningside	301160.95	1298288.66	0.23	0.82
Neighbourhood_nameNorth Grapeland Heights	-329620.29	1558714.50	-0.21	0.83
Neighbourhood_nameNorth Grove	-168663.84	1377474.56	-0.12	0.90
Neighbourhood_nameNorth Sewell Park	-269357.72	1468415.90	-0.18	0.85
Neighbourhood_nameNortheast Overtown	1854856.47	2060875.16	0.90	0.37
Neighbourhood_nameNorthwestern Estates	633298.91	1866122.55	0.34	0.73
Neighbourhood_nameOakland Grove	309602.47	1252049.93	0.25	0.80
Neighbourhood_nameOld San Juan	934683.00	1620252.56	0.58	0.56
Neighbourhood_nameOrange Bowl	-208779.25	1542940.36	-0.14	0.89
Neighbourhood_nameOrchard Villa	-1343027.39	1513083.57	-0.89	0.37
Neighbourhood_namePalm Grove	437880.44	1369119.75	0.32	0.75
Neighbourhood_nameParkdale North	37036.31	1506610.98	0.02	0.98
Neighbourhood_nameParkdale South	-311956.69	1430377.48	-0.22	0.83
Neighbourhood_nameRoads	-488048.85	1417955.74	-0.34	0.73
Neighbourhood_nameSan Marco Island	1147115.98	1472455.89	0.78	0.44
Neighbourhood_nameSanta Clara	-874180.19	1399252.90	-0.62	0.53
Neighbourhood_nameShenandoah North	-155489.99	1431225.17	-0.11	0.91
Neighbourhood_nameShenandoah South	-304102.18	1420008.83	-0.21	0.83
Neighbourhood_nameShorecrest	829155.83	970651.95	0.85	0.39
Neighbourhood_nameSilver Bluff	-352277.87	1419365.11	-0.25	0.80
Neighbourhood_nameSouth Grapeland Heights	-23425.87	1555961.92	-0.02	0.99
Neighbourhood_nameSouth Grove	-814341.76	1396802.60	-0.58	0.56
Neighbourhood_nameSouth Grove Bayside	-720377.98	1392176.92	-0.52	0.60
Neighbourhood_nameSouth Sewell Park	-45787.05	1452779.79	-0.03	0.97
Neighbourhood_nameSpring Garden	-1354225.48	1506502.42	-0.90	0.37
Neighbourhood_nameWest Grapeland Heights	-671565.93	1554238.41	-0.43	0.67
Neighbourhood_nameWest Grove	-333989.87	1303573.05	-0.26	0.80
MedHHInc	5.77	2.30	2.51	0.01
pctWhite	-8902.55	5932.67	-1.50	0.13
pctHispanic	3065.05	5377.12	0.57	0.57

Standard errors: OLS

	Est.	S.E.	t val.	p
worship_nn1	-12.46	27.32	-0.46	0.65
`8ftres3to8ftPool`	-128212.20	126217.22	-1.02	0.31
`2to4ftPool`	-212154.82	498905.74	-0.43	0.67
Whirpool	25346.66	124816.37	0.20	0.84
LuxuryPool	3307161.73	224967.93	14.70	0.00
LotSize	55.46	7.57	7.33	0.00
park_nn4	65.92	56.13	1.17	0.24
bar_nn2	-73014.65	158978.54	-0.46	0.65
hospital_nn1	7.99	16.64	0.48	0.63
`3to6ftPool`	1046079.91	392352.61	2.67	0.01
`3to8ftPool`	-163832.87	68136.80	-2.40	0.02
BedCat1	561709.23	419726.48	1.34	0.18
BedCat2	335956.89	371930.35	0.90	0.37
BedCat3	124312.69	374217.25	0.33	0.74
BedCat4	-143791.52	380893.07	-0.38	0.71
BedCat5	-287284.15	396689.09	-0.72	0.47
BedCat6	-190502.18	422185.12	-0.45	0.65
BedCat7	-614962.50	502825.18	-1.22	0.22
BedCat8+	234267.31	594140.46	0.39	0.69
Docks	862401.79	137684.02	6.26	0.00
lagPrice	-0.00	0.04	-0.03	0.98
parking_nn2	58.85	29.06	2.03	0.04
school_nn1	44.30	40.92	1.08	0.28
ElementarySchoolAuburndale Elementary	-6248.06	606807.46	-0.01	0.99
ElementarySchoolCitrus Grove Elementary	12099.83	519378.96	0.02	0.98
ElementarySchoolCoconut Grove Elementary	-86105.46	330812.89	-0.26	0.79
ElementarySchoolComstock Elementary	-136216.97	837891.65	-0.16	0.87
ElementarySchoolCoral Way K-8 Center	20874.72	402286.54	0.05	0.96
ElementarySchoolDrew, Charles R. K-8 Center	-867133.07	1428662.13	-0.61	0.54
ElementarySchoolDunbar, Paul L. Elementary	254829.51	536348.50	0.48	0.63
ElementarySchoolEdison Park K-8 Center	-786111.17	1085014.35	-0.72	0.47
ElementarySchoolFairlawn Elementary	954474.26	752592.75	1.27	0.20
ElementarySchoolFlagami Elementary	1109593.33	809283.04	1.37	0.17
ElementarySchoolFlagler, Henry M. Elementary	569730.36	740899.95	0.77	0.44
ElementarySchoolHartner, Eneida M. Elementary	-994334.92	1024283.10	-0.97	0.33
ElementarySchoolHolmes Elementary	-882392.43	1364112.87	-0.65	0.52
ElementarySchoolKensington Park Elementary	-195781.58	686437.14	-0.29	0.78
ElementarySchoolKinloch Park Elementary	301889.99	718440.20	0.42	0.67
ElementarySchoolL'ouverture, Toussaint Elementary	-873398.83	1088693.03	-0.80	0.42
ElementarySchoolLiberty City Elementary	-918943.90	1424922.03	-0.64	0.52

Standard errors: OLS

	Est.	S.E.	t val.	p
ElementarySchoolMcCrary, Jr. Jesse J. Elementary	-29261.12	861301.77	-0.03	0.97
ElementarySchoolMelrose Elementary	-134317.73	470810.98	-0.29	0.78
ElementarySchoolMiller, Phyllis R. Elementary	-160395.36	906812.60	-0.18	0.86
ElementarySchoolMorningside K-8 Center	NA	NA	NA	NA
ElementarySchoolOlinda Elementary	516682.45	567684.75	0.91	0.36
ElementarySchoolOrchard Villa Elementary	633315.69	729862.11	0.87	0.39
ElementarySchoolOtherES	48865.08	391049.72	0.12	0.90
ElementarySchoolPharr, Kelsey L. Elementary	204675.33	466071.55	0.44	0.66
ElementarySchoolRiverside Elementary	NA	NA	NA	NA
ElementarySchoolSanta Clara Elementary	147971.47	413149.61	0.36	0.72
ElementarySchoolShadowlawn Elementary	-1033182.94	1063132.31	-0.97	0.33
ElementarySchoolShenandoah Elementary	107176.84	420410.03	0.25	0.80
ElementarySchoolSilver Bluff Elementary	-30.28	385157.70	-0.00	1.00
ElementarySchoolSmith, Lenora Braynon. Elementary	54856.45	573135.85	0.10	0.92
ElementarySchoolTucker, Frances S. Elementary	NA	NA	NA	NA
ElementarySchoolWheatley, Phillis Elementary	-2353559.91	1561665.57	-1.51	0.13
EffectiveYearBuilt	2099.61	897.03	2.34	0.02
Zoning0104 - SINGLE FAM - ANCILIARY UNIT	-16124.44	132290.23	-0.12	0.90
Zoning0800 - SGL FAMILY - 1701-1900 SQ	1308712.09	155296.45	8.43	0.00
Zoning2100 - ESTATES - 15000 SQFT LOT	4391151.35	345795.95	12.70	0.00
Zoning2200 - ESTATES - 25000 SQFT LOT	3581935.68	500672.47	7.15	0.00
Zoning2800 - TOWNHOUSE	555798.21	379303.32	1.47	0.14
Zoning3900 - MULTI-FAMILY - 38-62 U/A	31159.72	187684.64	0.17	0.87
Zoning3901 - GENERAL URBAN 36 U/A LIMITED	81728.61	270469.70	0.30	0.76
Zoning4600 - MULTI-FAMILY - 5 STORY &	210534.21	313451.87	0.67	0.50
Zoning4601 - MULTI-FAMILY - 8 STORY &	-761558.07	1081935.16	-0.70	0.48
Zoning4801 - RESIDENTIAL-LIMITED RETAI	113642.17	491261.54	0.23	0.82
Zoning5700 - DUPLEXES - GENERAL	61029.35	94966.30	0.64	0.52
Zoning6100 - COMMERCIAL - NEIGHBORHOOD	-113964.10	468308.80	-0.24	0.81
Zoning6101 - CEN-PEDESTRIAN ORIENTATIO	-73929.58	446732.35	-0.17	0.87
Zoning6106 - RESIDENTIAL-LIBERAL RETAI	-125921.79	1031455.84	-0.12	0.90
Zoning6107 - RESIDENTIAL-MEDIUM RETAIL	119898.80	339518.21	0.35	0.72
Zoning6110 - COMM/RESIDENTIAL-DESIGN D	102117.09	645460.38	0.16	0.87
Zoning6402 - URBAN CORE 24 STORY/7FLR	NA	NA	NA	NA
Zoning7000 - INDUSTRIAL - GENERAL	59334.58	863879.81	0.07	0.95
Zoning7700 - INDUSTRIAL - RESTRICTED	NA	NA	NA	NA
Bath	57539.12	35865.58	1.60	0.11
Stories	-39318.17	65121.33	-0.60	0.55
logCoastDist	-10580.06	31617.16	-0.33	0.74
Property.CityMiami Beach	NA	NA	NA	NA

Standard errors: OLS

	Est.	S.E.	t val.	p
office_nn3	-10.68	41.62	-0.26	0.80
dist.metro	-449742.43	58091.23	-7.74	0.00
Fence	23679.36	48113.90	0.49	0.62
`XFs_Elevator - Passenger`	994133.10	235472.32	4.22	0.00

Standard errors: OLS

Accuracy - Mean Absolute Error

The following graph shows the distribution of absolute errors in our model. While most of our predictions are clustered at the low end of the graph, there are outliers of over $2,000,000 that are negatively impacting the MAE of the model.

Code

```
## [1] 360841.4
```

Code

```
## [1] 0.5451045
```

Code

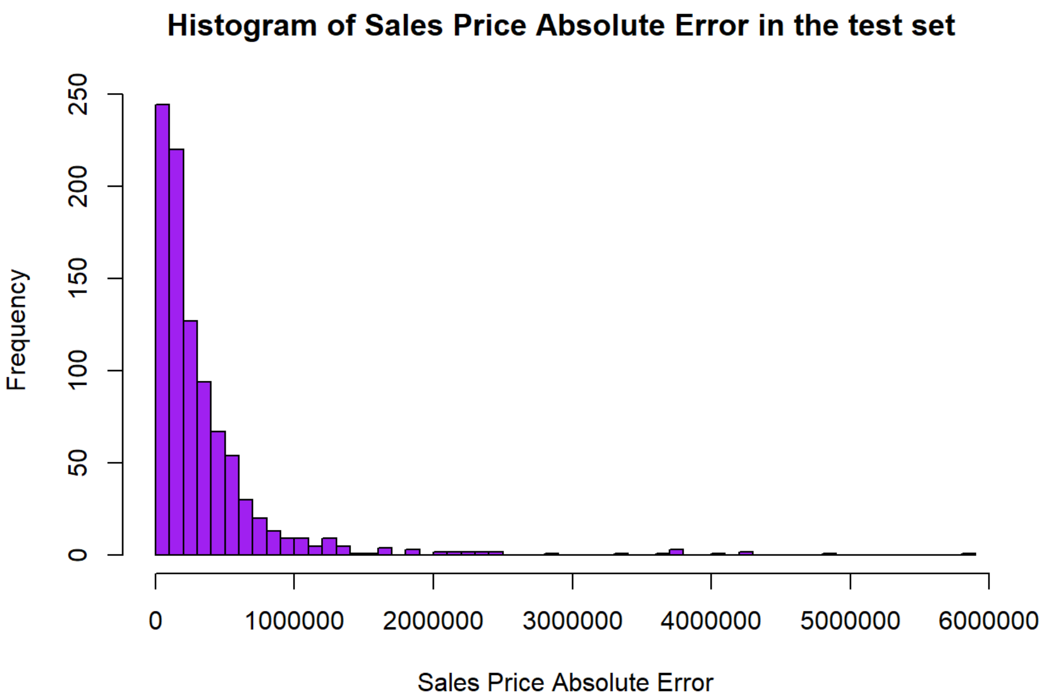


Table of MAE and MAPE

Code

DataSet	Mean_Absolute_Error	Mean_Absolute_Percent_Error
Test set	360841.4	0.5451045

Summary Statistics of Test dataset
Table 2.1

Spatial Correlation of residuals

Code

Residuals

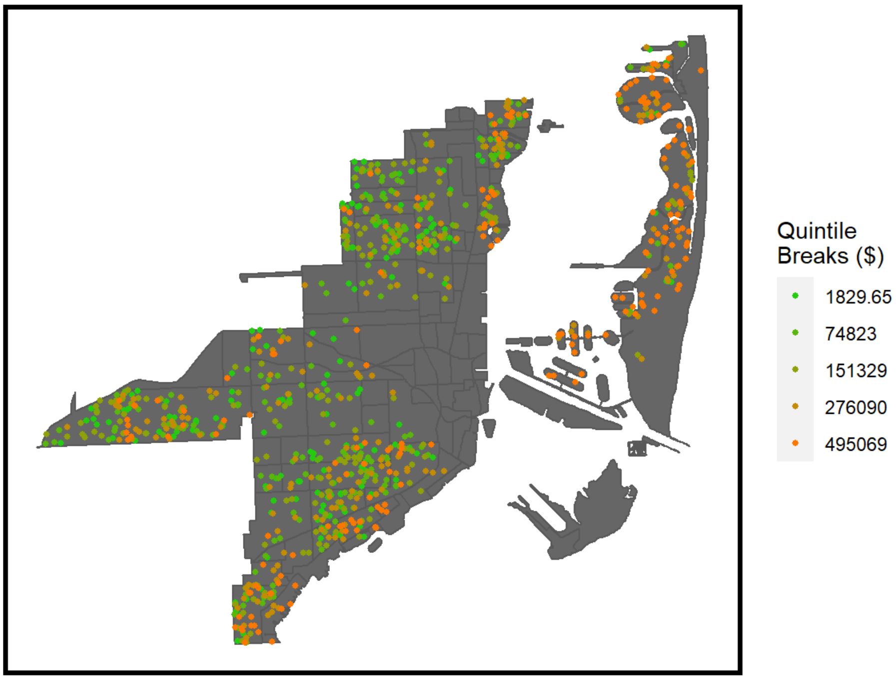


Figure 4.1

Spatial Correlation of error

The first plot shows that as the price of a house increases, the prices of nearby houses also increase. This demonstrates the importance of including spatial features in our model.

The second plot shows that the errors of our model are also spatially clustered. This indicates that our model is missing important spatial features.

Code

Price as a function of the spatial lag of price

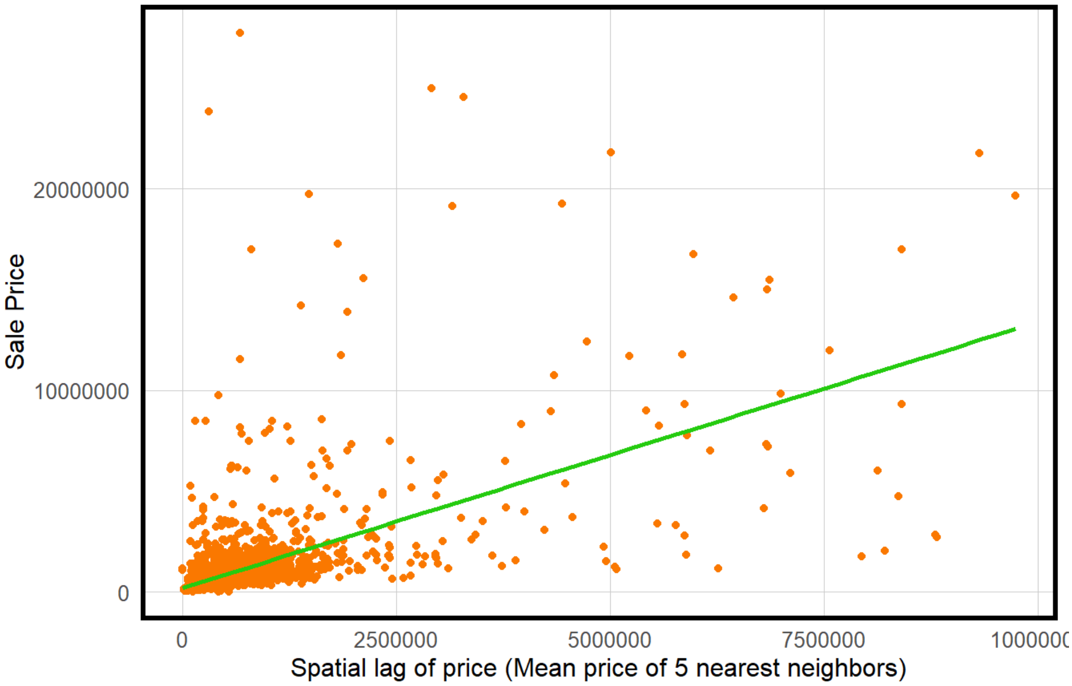


Figure 4.2

Code

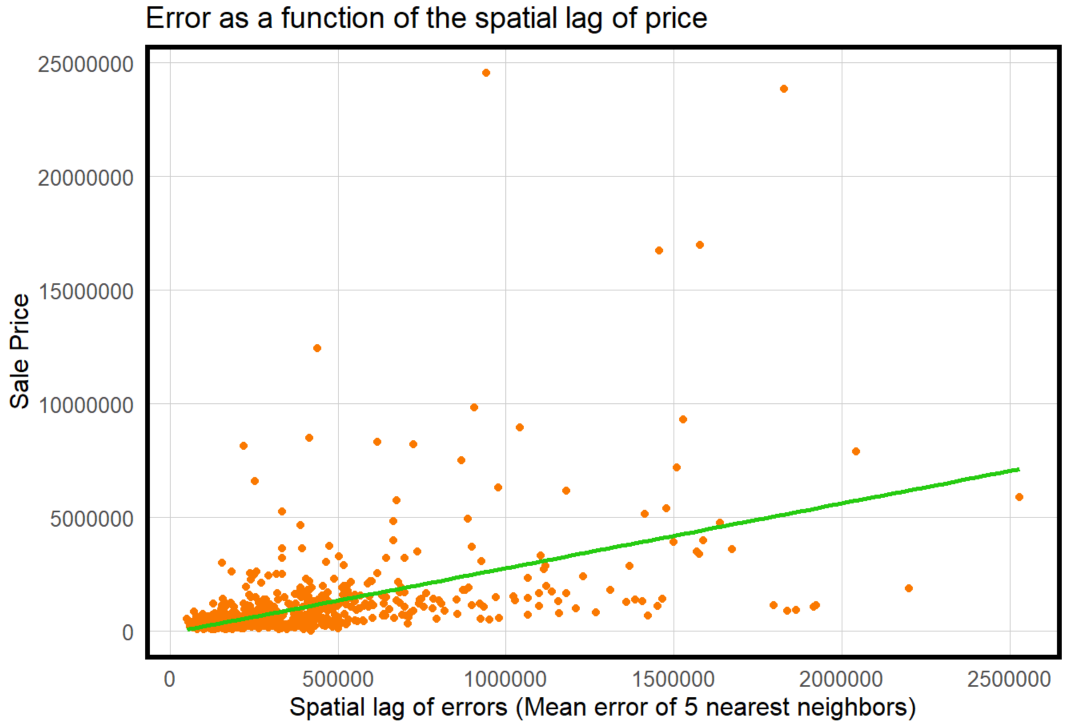


Figure 4.3

Moran I's

Moran's I provides another means of determining whether our errors are spatially correlated. Moran's I here is positive, suggesting positive spatial autocorrelation in our model.

Code

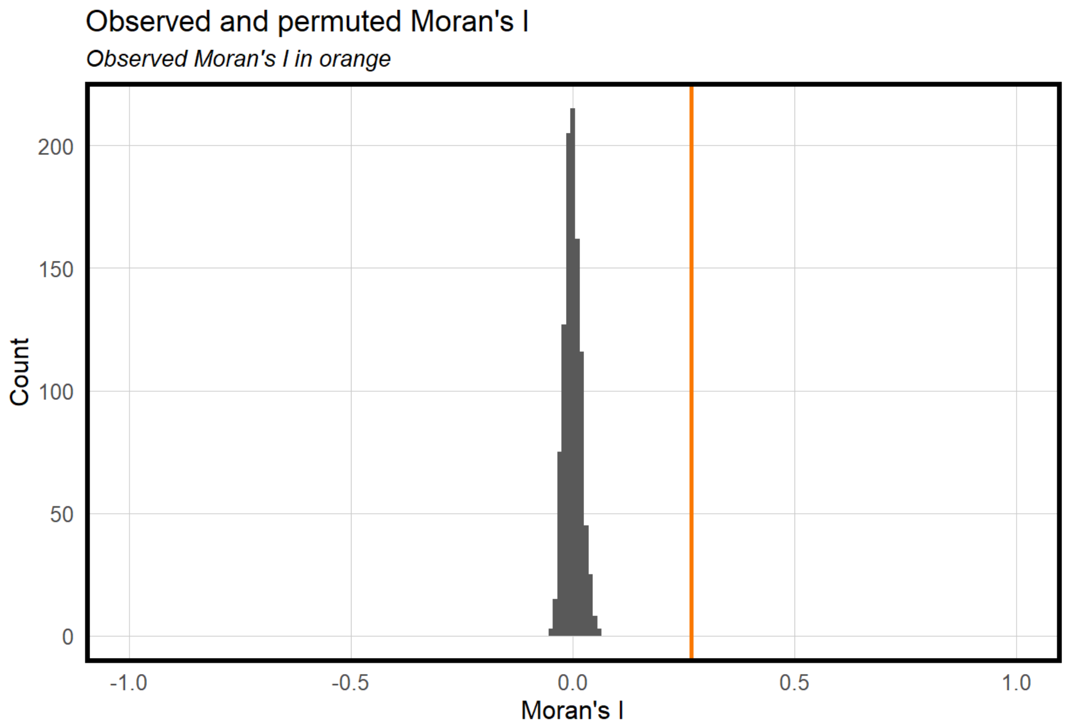


Figure 4.4

Error by neighborhood

To explore the spatial nature of our errors, we calculated the MAE for houses within each neighborhood and for both cities.The first table shows that our model's performance varies across neighborhoods, though it does not perform particularly well in any neighborhood.The lowest MAE is still substantial at $50,365 within the Latin Quarter.

The second table demonstrates that our model performs better on Miami houses than on Miami Beach houses.

Code

Neighbourhood_name	meanPrice	meanPrediction	meanMAE
Fair Isle	1351250.0	2454506.6	1217466.60

Neighbourhood_name	meanPrice	meanPrediction	meanMAE
South Grove Bayside	3718125.0	3242177.4	1157684.29
MIAMI.BEACH.8	2910537.4	2910428.1	964252.82
Palm Grove	911400.0	1656649.2	745249.21
Baypoint	3195000.0	3123451.8	703136.42
Bay Heights	1250875.0	1190346.1	680391.44
MIAMI.BEACH.6	1508058.8	1757176.5	505002.57
East Grove	1822863.2	1407615.5	475958.57
Bird Grove East	538333.3	960711.3	422378.01
West Grove	483825.0	647860.2	419906.88
North Sewell Park	334175.0	524655.6	394472.95
Morningside	940700.0	1034800.8	381091.56
Belle Meade	732694.1	624083.7	375042.00
Shorecrest	599991.3	550017.0	361243.39
South Grove	1311141.7	1309609.2	357220.56
Miami Avenue	1520000.0	1505716.0	340253.51
North Grapeland Heights	301871.4	101556.0	315684.18
Roads	626160.6	614119.8	280137.62
South Sewell Park	360500.0	513606.2	269001.77
Little River Central	247250.0	506597.0	259346.99
La Pastorita	317500.0	562838.9	245338.89
Shenandoah South	458442.9	582039.3	232010.16
West Grapeland Heights	304875.0	527631.7	230667.33
Flagami	335762.3	363393.0	230054.53
Old San Juan	717428.6	795527.6	229612.38
North Grove	961000.0	1041929.6	221931.45
Edison	218891.3	365218.2	220673.24
Santa Clara	223933.3	250264.4	215833.30
Historic Buena Vista East	730000.0	942325.5	212325.46
South Grapeland Heights	310214.3	385532.5	209218.11
Curtis Park	276666.7	477750.7	201084.01
Douglas Park	374892.9	337730.8	199074.82
Parkdale North	458500.0	656431.4	197931.36
Liberty Square	146100.0	122471.9	197864.49
Flora Park	166484.2	183763.8	197786.74

Neighbourhood_name	meanPrice	meanPrediction	meanMAE
Shenandoah North	477625.0	549837.8	187213.02
Buena Vista West	302666.7	362060.7	184281.38
Parkdale South	416600.0	385394.2	180431.27
Bird Grove West	450000.0	269782.4	180217.58
Orchard Villa	234111.1	131382.5	174315.90
Buena Vista Heights	364742.9	521821.7	170733.40
Silver Bluff	492308.6	473923.9	165304.71
East Little Havana	281000.0	217502.5	156060.86
Belle Meade West	367916.7	430345.1	155289.76
Hadley Park	215189.8	185370.7	154069.24
Auburndale	329117.6	355614.6	153613.48
Bayside	629600.0	623811.8	145148.22
Citrus Grove	316378.9	231698.5	140372.64
Coral Gate	440750.0	510352.6	120860.72
Northwestern Estates	186772.7	231155.5	118581.94
King Heights	224250.0	208714.1	117826.34
Melrose	238300.0	163730.5	111800.99
Lemon City/Little Haiti	262428.6	294687.5	111018.73
Allapattah Industrial District	220000.0	111438.5	108561.47
Latin Quarter	255000.0	177690.8	77309.19

Code

Property.City	meanPrice	meanPrediction	meanMAE
Miami Beach	2749441.9	2777960.0	911501.1
Miami	565277.8	579796.4	257418.0

Map of Predicted Values

Code

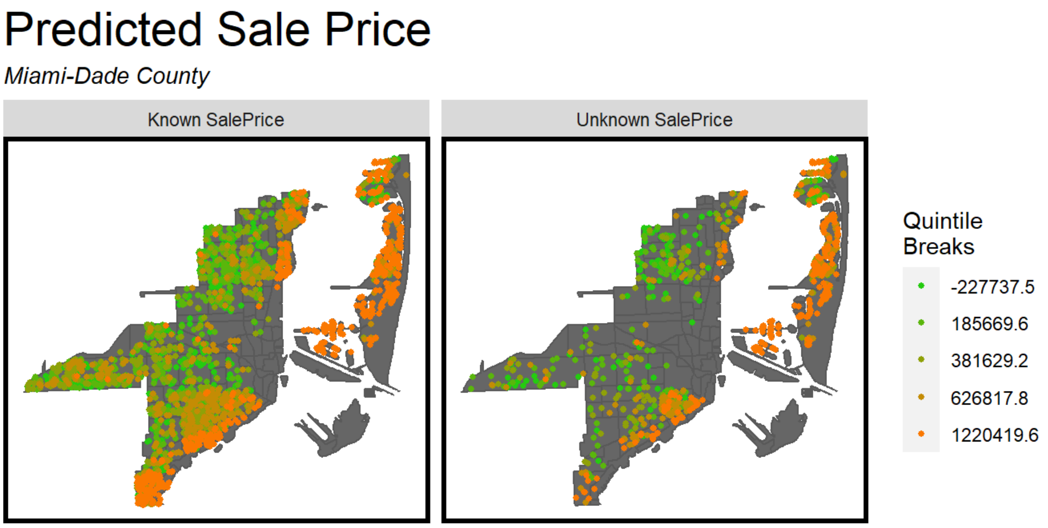


Figure 5.1

Map of MAPE by Neighborhood

Code

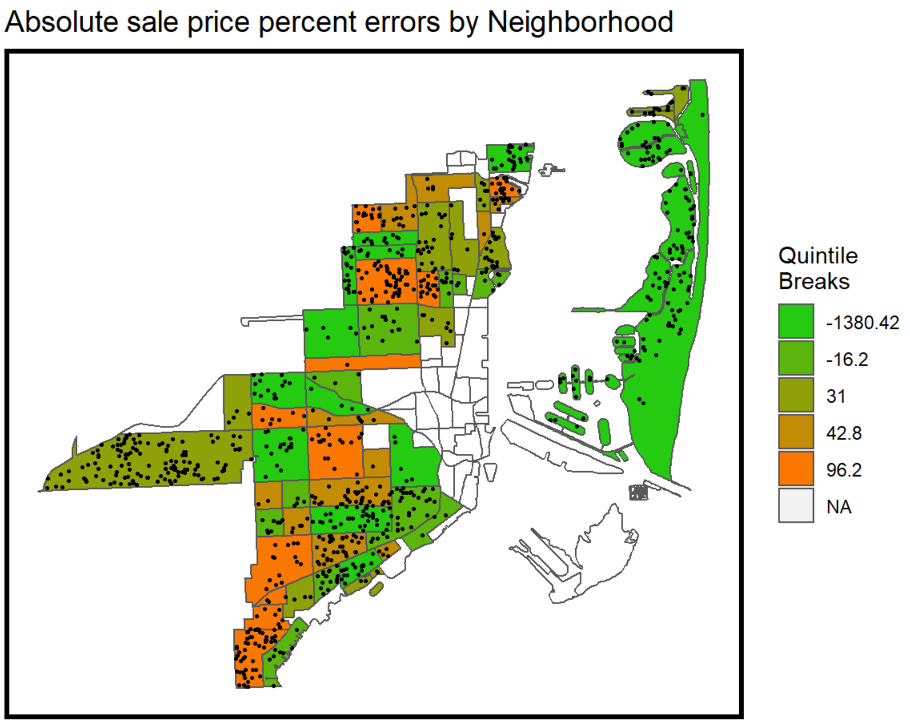
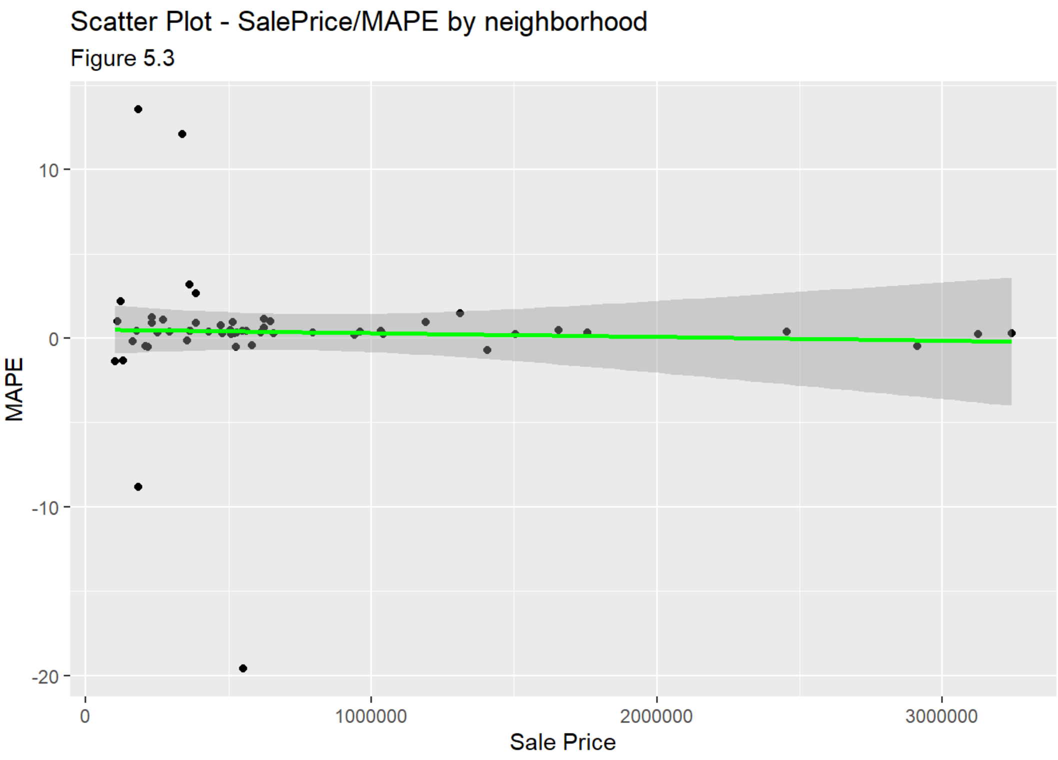


Figure 5.2

Scatterplot of MAPE by Neighborhood

Code



Generalization

The spatial clustering of errors in our model is visualized below.

Code

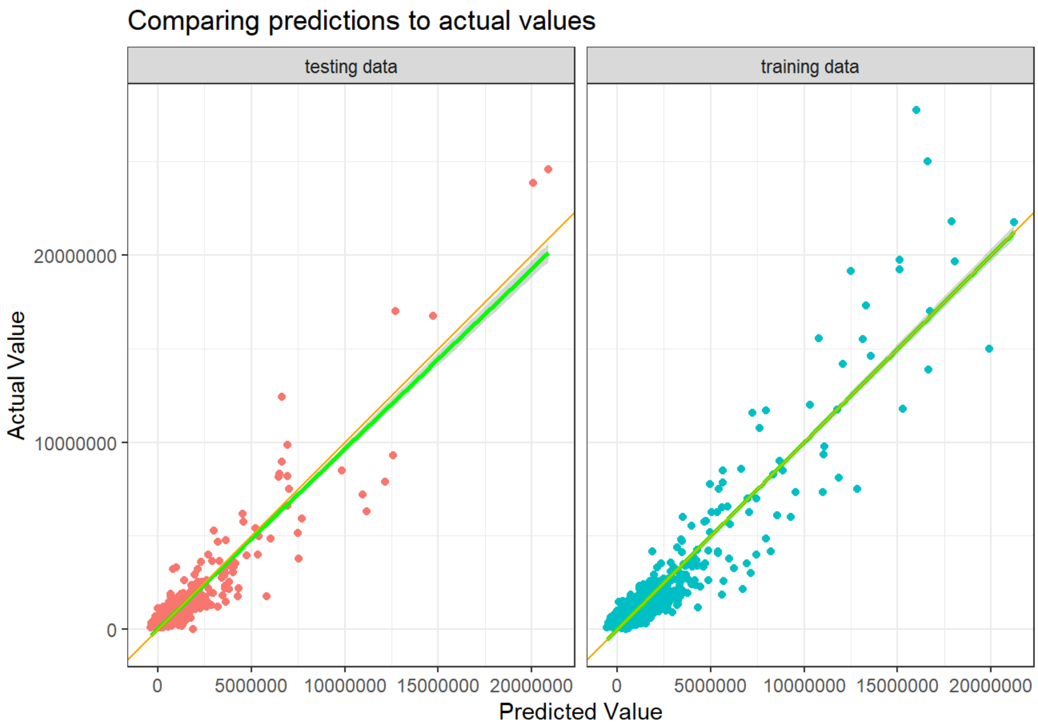


Figure 6.1

Cross Validation

When calculating MAE above, the data was randomly split into a training and a testing set. This method entails a risk that the data will split poorly, generating a misleading MAE. To reduce this risk, we also analyzed our model using cross validation. We divided the data into 100 equal sized subsets, which were further subset into training and testing sets. For each of the 10 subsets, the MAE was calculated. We used this method to find the model with the best average MAE across 100 subsets.

```
## Linear Regression
##
## 2627 samples
##   32 predictor
##
## No pre-processing
## Resampling: Cross-Validated (100 fold)
## Summary of sample sizes: 2602, 2601, 2600, 2600, 2600, 2600, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   731697.3  0.8423277  377179.9
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross Validation Test Results

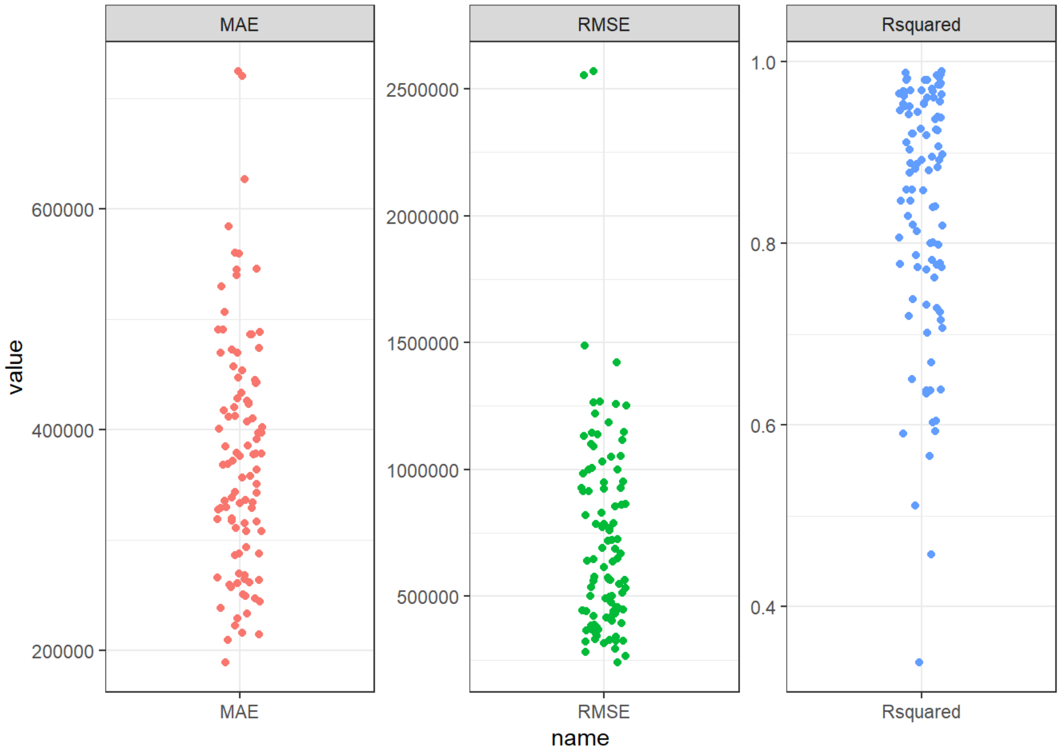
Code

Test	Mean	Max	Min	Standard_Deviation
Cross_Validation	377179.9	724162.3	188802.4	107925.3

Summary Statistics of Cross Validation, k = 100 folds
Table 3.1

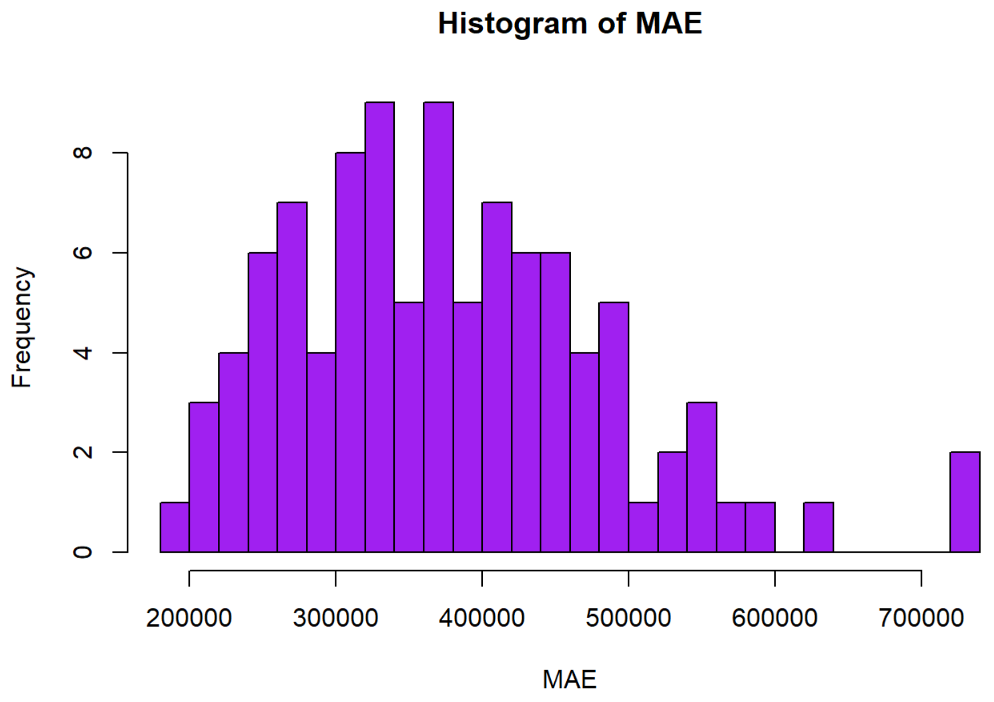
The mean absolute error, root-mean-square-error, and r-squared for each of the 10 subsets is visualized below.

Code



Histogram of MAE

Code



Plot of Predicted Values

Code

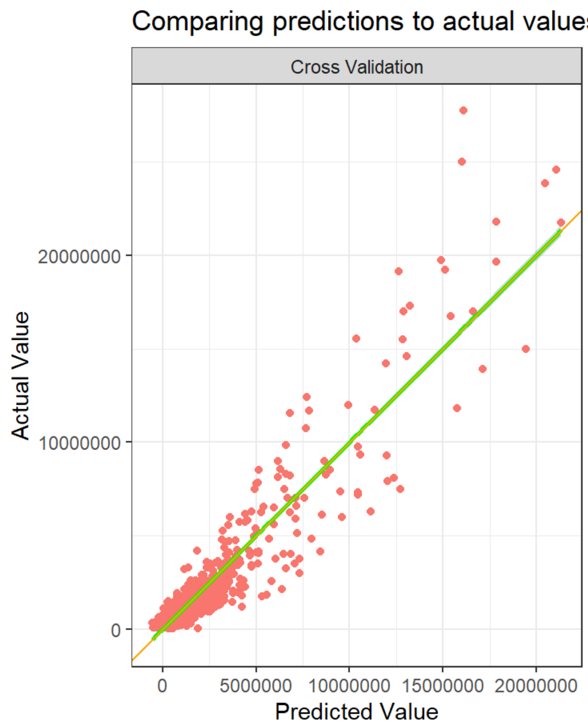


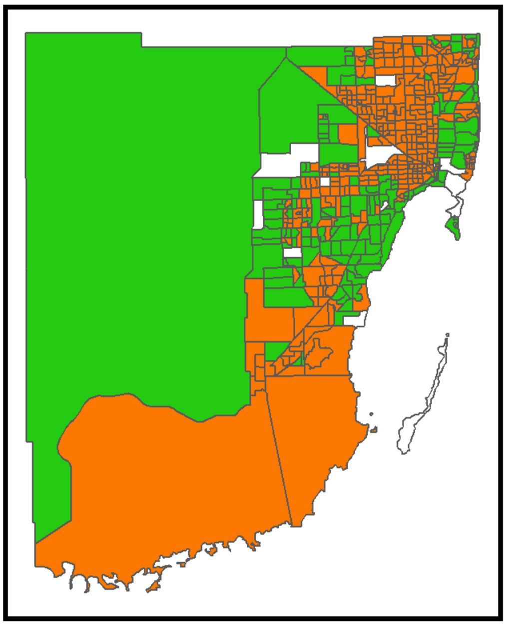
Figure 6.2

Generalizability by Income

To further test the generalizability of the model, we calculated MAPE for below and above average census tracts. The results indicate that our model has larger percentage of absolute error in below average income census tracts when compared to above average income census tracts.

Code

Income Context



Code

Test set MAPE by neighborhood income context

Above Average Income	Below Average Income
28%	65%

Discussion

Unfortunately our analysis suggests that the model is not as effective as we hoped. Our MAE was around $350,000, this magnitude of error would not be acceptable for Zillow's Zestimate. We found that the adjusted square feet of a house is an important predictor for sale price, and we struggled to significantly improve the model beyond that variable. Based on our analysis of the spatial autocorrelations in our model, we are missing important spatial processes. We also suspect that key internal characteristics are missing and would have helped the model if we had access to them.

The lack of open source data for Miami Beach was a key barrier in the development of this model. Our model was unable to successfully predict sale prices for the very expensive houses in Miami Beach. Out of all the variables we used, the negative correlation of the distance to the coast and home sale prices was surprising to us.

Conclusion

We would not recommend our model to Zillow given the large errors in our predictions and the lack of finding uniform data across all the neighborhoods. This model may be improved by additional spatial features.