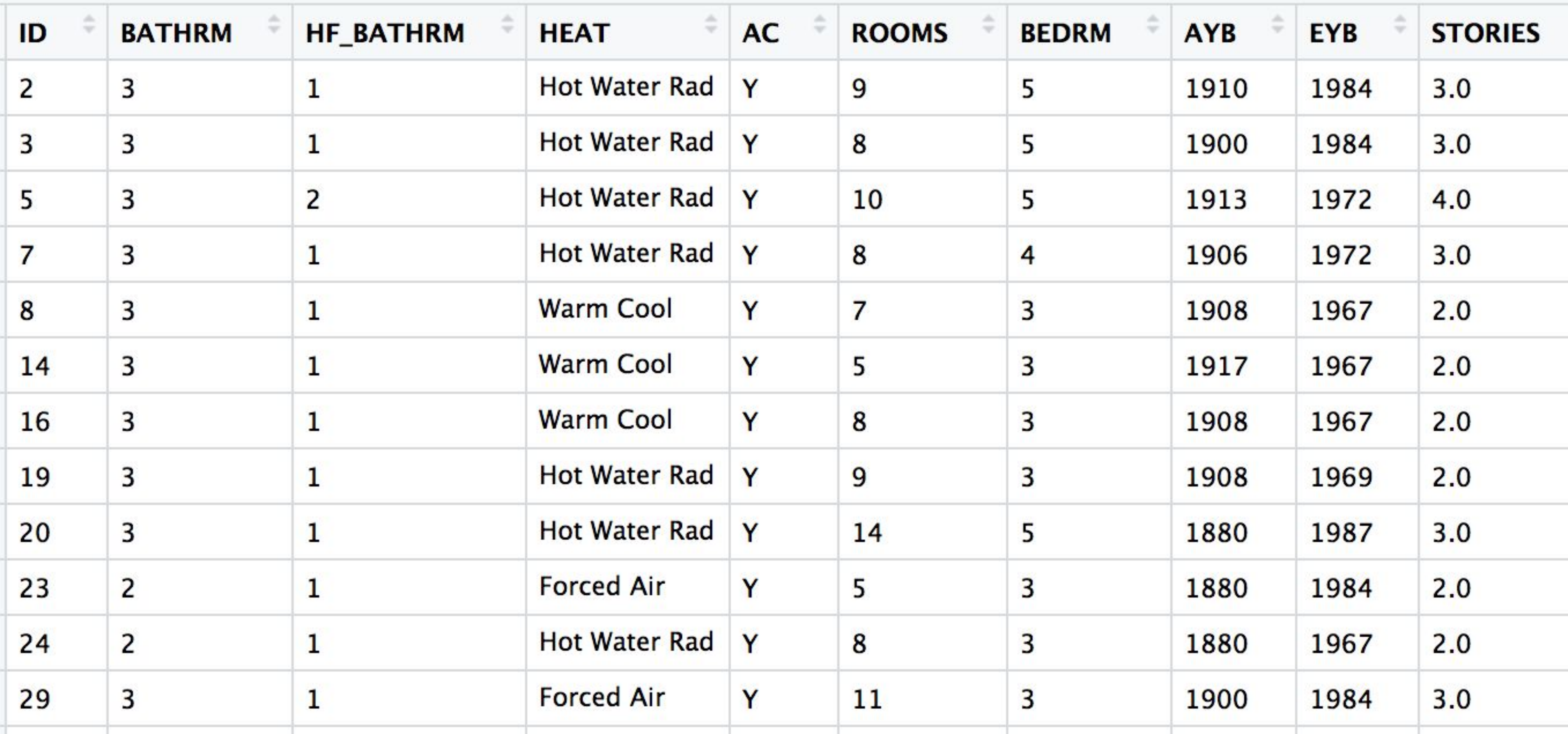
DC Properties Qualification

By Aaron Niecestro & Kingsley Iyawe

Background and Story

- We all go to American University as students
- We all live either in DC or the surrounding states = ?
- This type of regression was not conducted before
- Housing models usually have a price response variable with multiple linear regression
    - Ours has a qualification response variable with logistic regression
- Qualification = paperwork is in order and inspection is passed

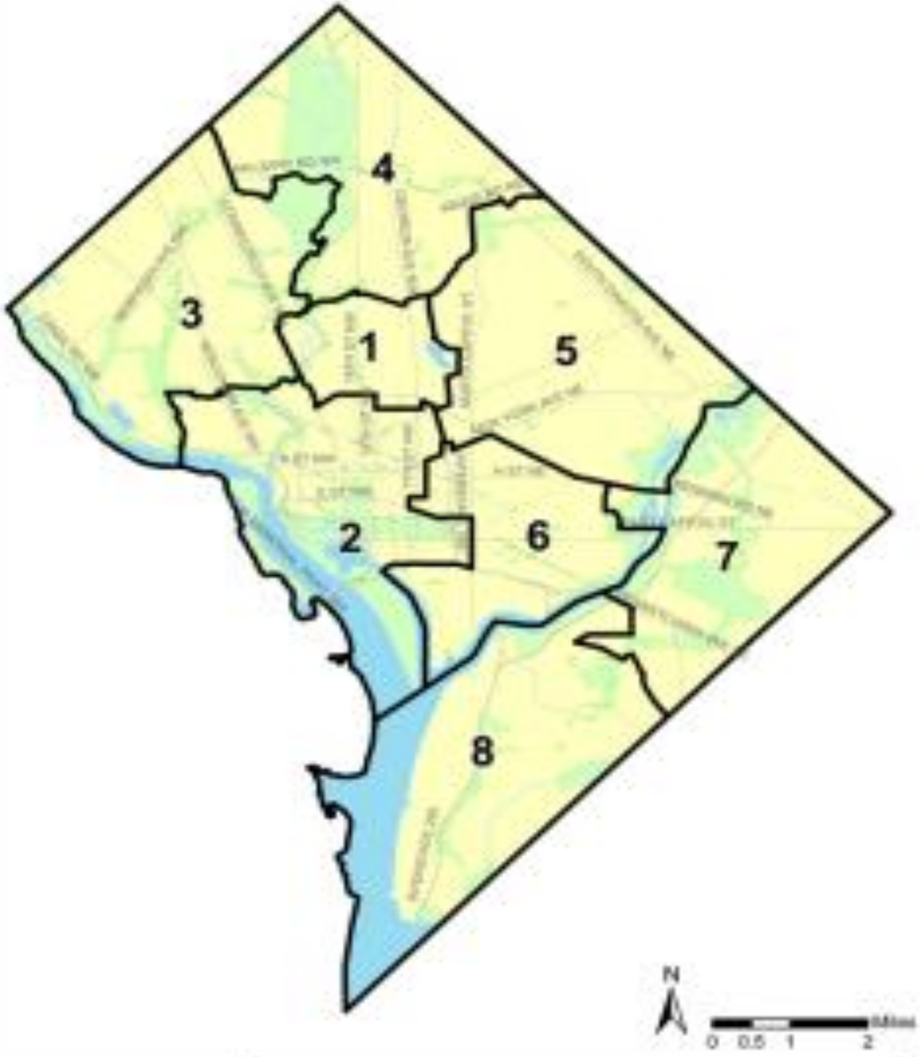
A small Part of the Data Set



ID	BATHRM	HF_BATHRM	HEAT	AC	ROOMS	BEDRM	AYB	EYB	STORIES
2	3	1	Hot Water Rad	Y	9	5	1910	1984	3.0
3	3	1	Hot Water Rad	Y	8	5	1900	1984	3.0
5	3	2	Hot Water Rad	Y	10	5	1913	1972	4.0
7	3	1	Hot Water Rad	Y	8	4	1906	1972	3.0
8	3	1	Warm Cool	Y	7	3	1908	1967	2.0
14	3	1	Warm Cool	Y	5	3	1917	1967	2.0
16	3	1	Warm Cool	Y	8	3	1908	1967	2.0
19	3	1	Hot Water Rad	Y	9	3	1908	1969	2.0
20	3	1	Hot Water Rad	Y	14	5	1880	1987	3.0
23	2	1	Forced Air	Y	5	3	1880	1984	2.0
24	2	1	Hot Water Rad	Y	8	3	1880	1967	2.0
29	3	1	Forced Air	Y	11	3	1900	1984	3.0

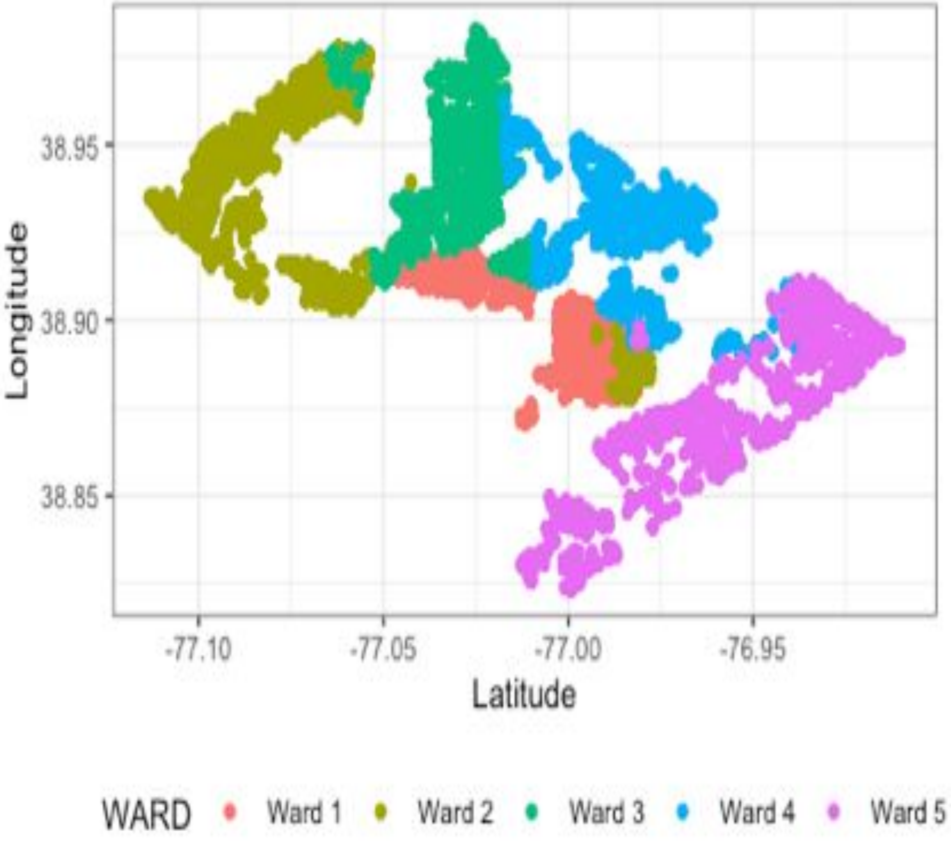
Questions

1. What is Qualification?
2. What qualifies a residential property to be sold on the housing market?
3. Is the property price the most important factor in determining whether a property is qualified to go on the market?
4. Do the realtors even care about whether a property is qualified to sell before listing it or is it all about the money?
5. Are we creating the most optimal regression for modeling properties?
6. Do we follow previous linear regression housing model approaches for predictor variables, or should we come up with our own model and approaches from scratch?
7. **Is money the most important thing? If so how does that define the world?**



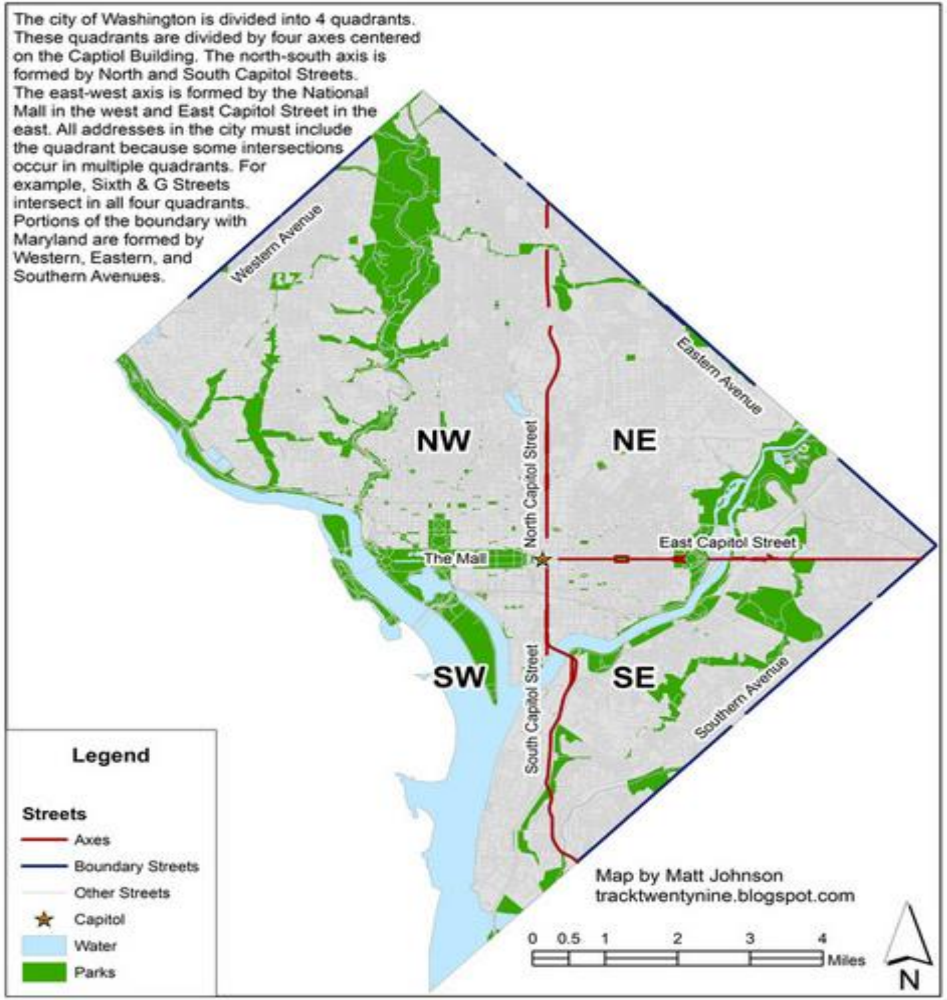
Map of Data by Ward

Originally there were 8 Wards



Data from Kaggle.com

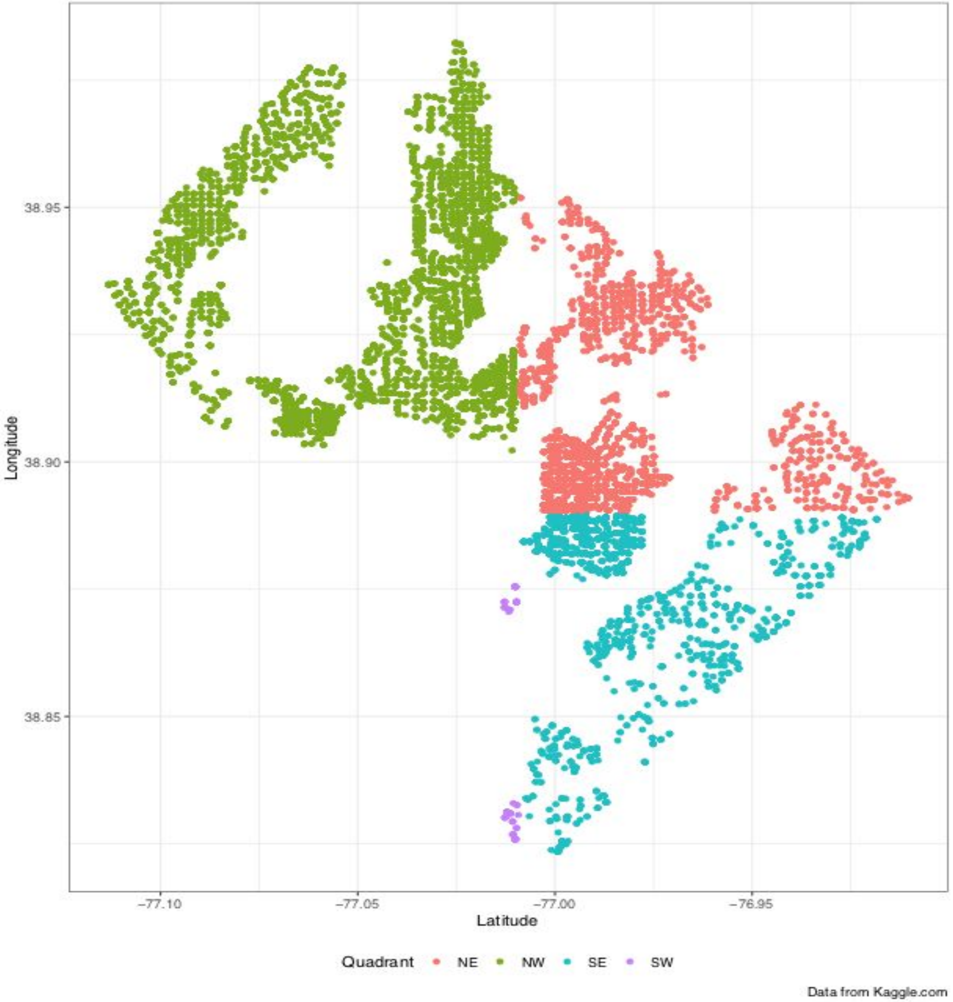
Transportation: Boundaries and Axes



The city of Washington is divided into 4 quadrants. These quadrants are divided by four axes centered on the Captiol Building. The north-south axis is formed by North and South Capitol Streets. The east-west axis is formed by the National Mall in the west and East Capitol Street in the east. All addresses in the city must include the quadrant because some intersections occur in multiple quadrants. For example, Sixth & G Streets intersect in all four quadrants. Portions of the boundary with Maryland are formed by Western, Eastern, and Southern Avenues.

Map by Matt Johnson
tracktwentynine.blogspot.com

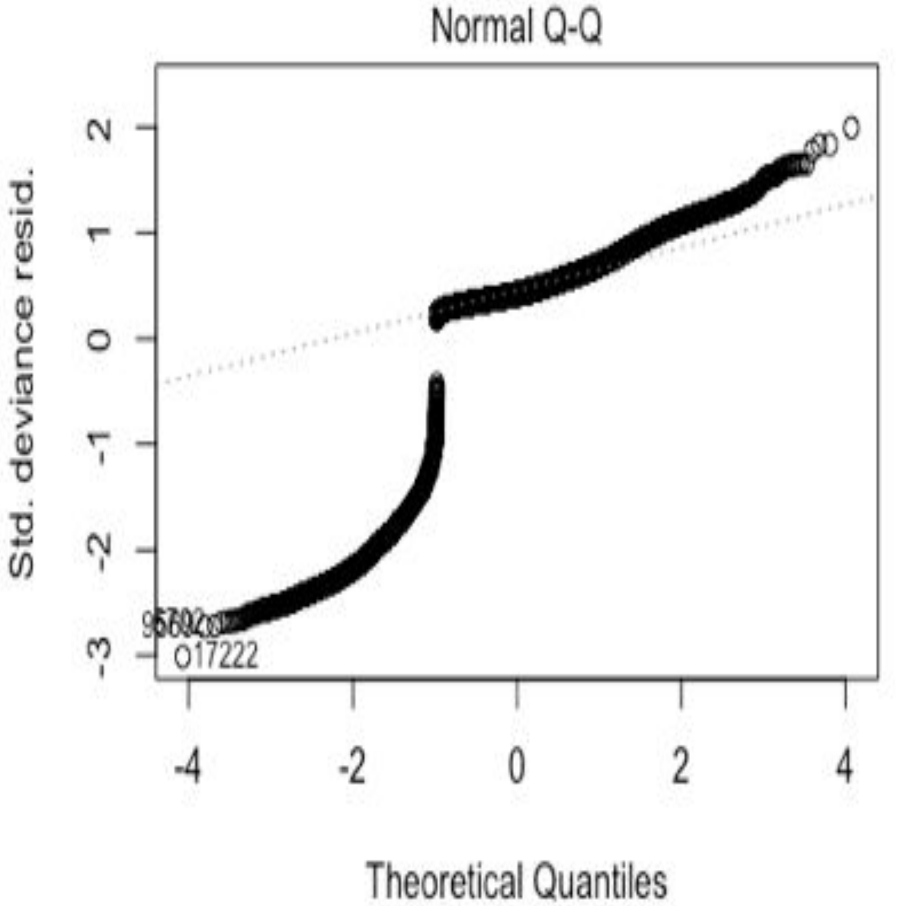
Map of Data By Quadrants
Little to no South–West area

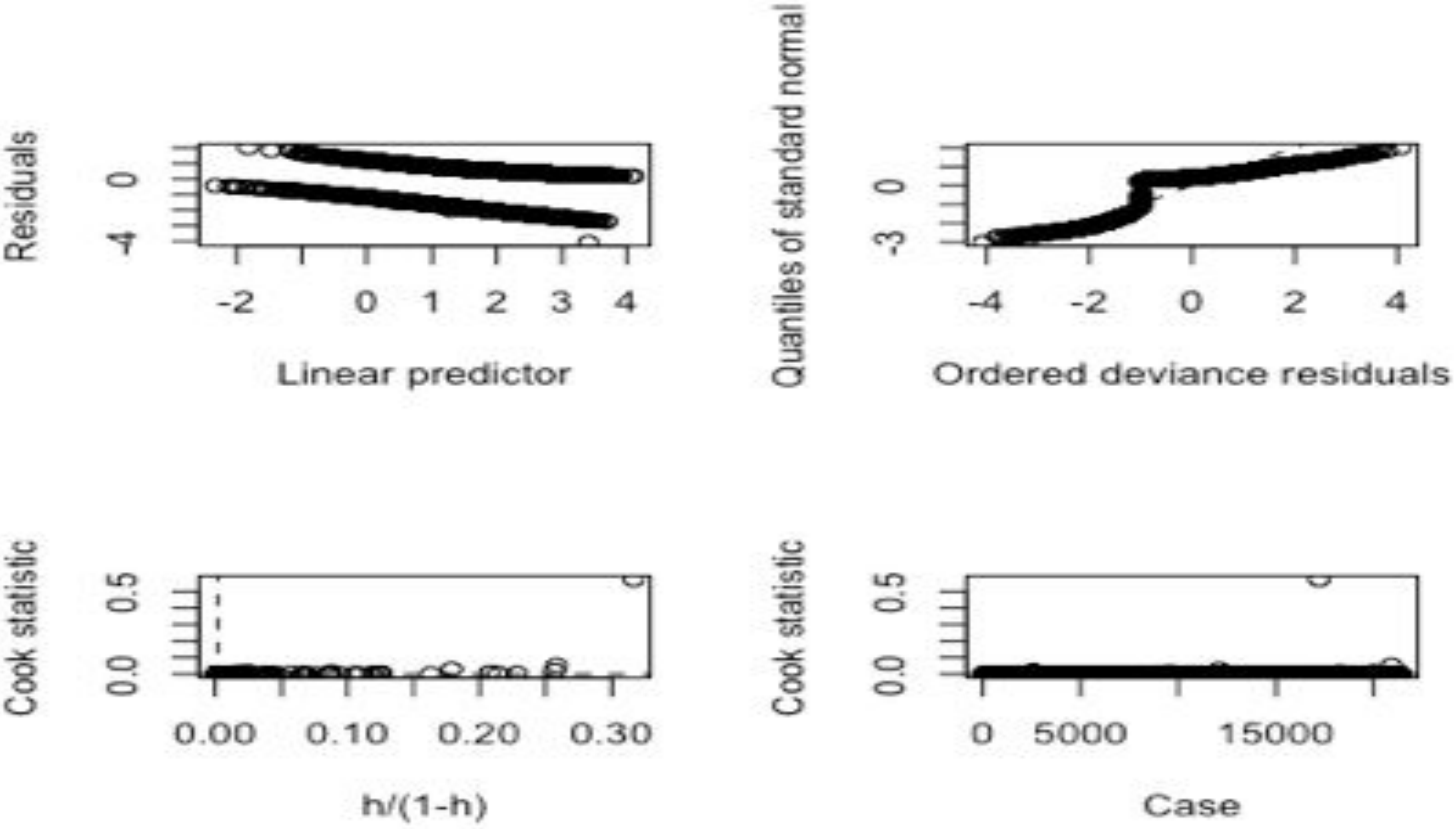


Data from Kaggle.com

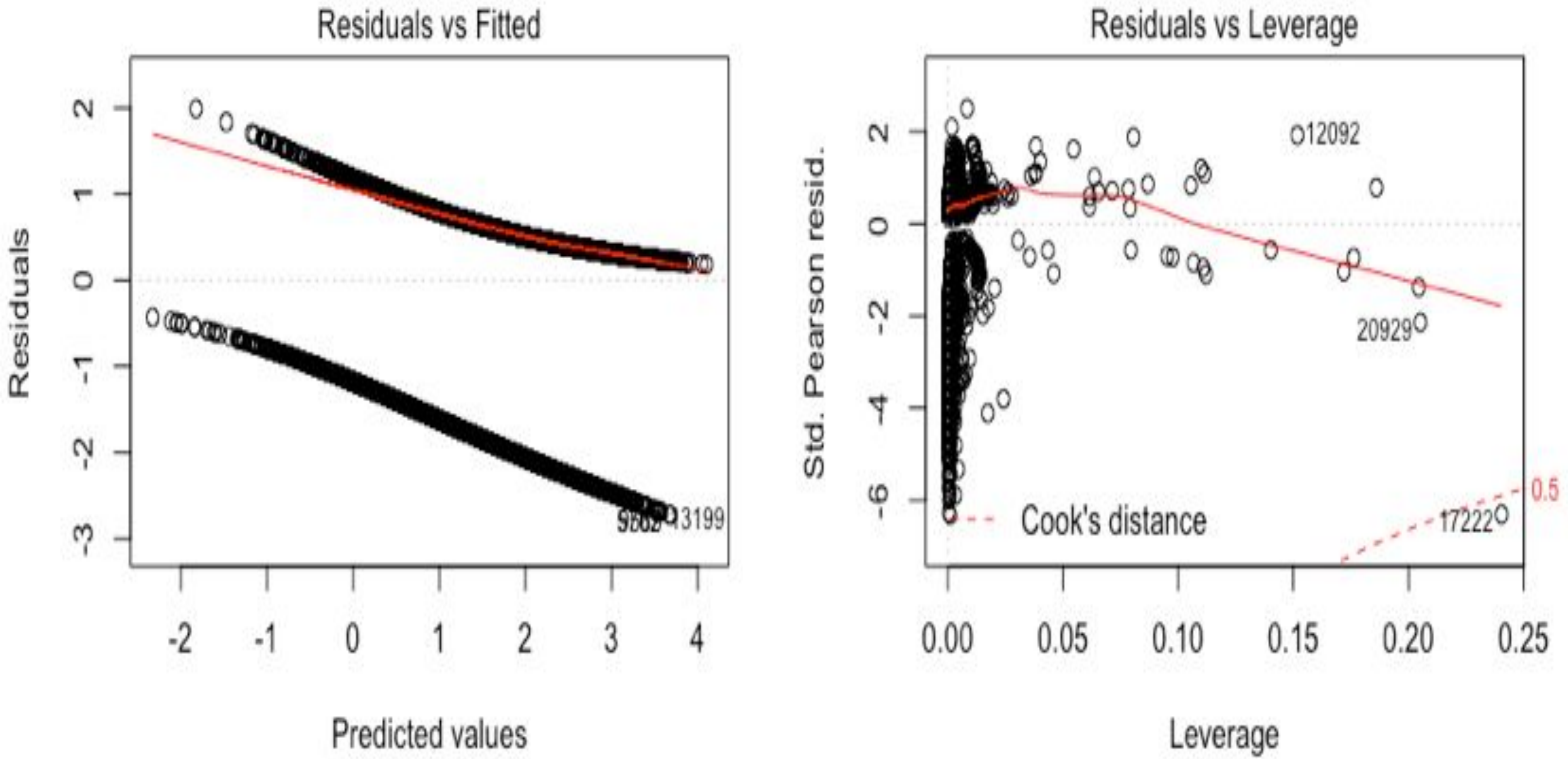
Diagnostic Plots

- Data is not normally distributed
- We definitely have outliers
    - The outliers can be taken out of our model
- This QQ-plot will not follow the standard bell curve when compared with a histogram





Residuals
0
-4
-2
0
1
2
3
4
Linear predictor
Quantiles of standard normal
0
-3
-4
-2
0
2
4
Ordered deviance residuals
Cook statistic
0.5
0.0
0.00
0.10
0.20
0.30
h/(1-h)
Cook statistic
0.5
0.0
0
5000
15000
Case


Residuals vs Fitted
Residuals
2
1
0
-1
-2
-3
9562
13199
-2
-1
0
1
2
3
4
Predicted values
Residuals vs Leverage
Std. Pearson resid.
2
0
-2
-4
-6
12092
20929
17222
Cook's distance
0.5
0.00
0.05
0.10
0.15
0.20
0.25
Leverage

Some issues with model

- Multicollinearity
  - Style and number of floors (STORIES)
  - Half bathrooms and bathrooms
  - Longitude, latitude, quadrant, and wards
  - Interaction variables
- Too many variables
  - Our second model originally had 33 variables
    - Style alone was broken up into 14 variables
- Not all the categories were being added to our model
- **Constantly losing data**

AIC & BIC

- AIC and BIC Analysis from our first model Results
    - Variables Left:
        - PRICE, BATHRM, AC=Yes, ROOMS, BEDRM, Style dummy variables, Condition dummy variables, KITCHENS, Ward 2, Ward 3, Ward 4, Ward 5
    - AIC = 17418, BIC = 17526
- Stepwise AIC Analysis with all 2-way interaction terms Results
    - Variables Left
        - PRICE, AC=Yes, ROOMS, BEDRM, CNDTN dummy variables, WARD 2-5, PRICE*(AC=Yes), PRICE*ROOMS, PRICE*BEDRM, PRICE*CNDTN, PRICE*WARD, CNDTN*WARD
    - AIC = 16960
- Stepwise BIC Analysis with all 2-way interaction terms Results
    - Variables Left:
        - PRICE, AC=Yes, ROOMS, BEDRM, CNDTN dummy variables, WARD dummy variables, PRICE*(AC=Yes), PRICE*ROOMS, PRICE*WARD
    - BIC = 17200

Hypothesis Test Results

Likelihood Ratio Tests:

- Tested if price, AC, and ward were equal to 0 or not
- Tested if the interaction terms were equal to 0 or not

Results:

- We could reject both null hypotheses since the p-value is 0.00

Goodness of Fit Test:

- H0: Our model fits the data
- HA: Our model does not fit the data

Results:

- We can reject the null hypothesis
- Our model doesn’t fit the data

Final Model Equation

Log(**π**/1-**π**) = - 3.158 - 0.000004374*Price + 0.007901*√Price - 0.2907*(AC=Yes) + 0.1821*Rooms + 2.221*(Rooms^0.2) - 0.04816*√BEDRM + 0.1069*(CNDTN=Excellent) - 1.196*(CNDTN=Fair) + 0.2849*(CNDTN=Good) - 1.373*(CNDTN=Poor) + 0.6739(CNDTN=Very Good) – 0.4306*(Ward=2) – 0.2858*(Ward=3) – 0.0515*(Ward=4) + 0.2901(Ward=5) + 0.000001085PRICE*(AC=Yes) + 0.00000002.698*(PRICE*ROOMS) + 0.0000003.897(Price*Ward=2) + 0.0000005486*(Price*Ward=3) + 0.000001052*(Price*Ward=4) – 0.0000003.313*(Price*Ward=5)

AIC: 16748

How to interpret the final model results

Continuous Variables:

- For every additional (A), the odds a property being qualified to sell will (B) of (C)

Dummy Variables:

- If a property has (A), then the odds a property being qualified to sell will (B) of (C)

Interaction Terms

- For every additional (A.1) and if a property has (A.2), the odds a property being qualified to sell will (B) of (C)

Interpretation of the Final Model

Variable (A)	**Change (B)**	**Number (C)**
Dollar in price	Decrease by a factor	0.00000437402
Dollar of the square root of Price	Increase by a factor	1.007932
AC = Yes	Decrease by a factor	0.3373633
One room	Decrease by a factor	0.1997342
Room raised to ⅕ power	Increase by a factor	9.216543
Square root of bedrooms	Decrease by a factor	0.6186622
Condition = Excellent	Increase by a factor	1.112823
Condition = Fair	Decrease by a factor	2.306863
Condition = Good	Increase by a factor	1.329629

Condition = Poor	Decrease by a factor	2.947174
Condition = Very Good	Increase by a factor	1.961874
Ward 2	Decrease by a factor	0.5381802
Ward 3	Decrease by a factor	0.3308263
Ward 4	Decrease by a factor	0.05284919
Ward 5	Increase by a factor	1.336561
Price * AC = Yes	Increase by a factor	1.000001
Price * Rooms	Increase by a factor	1
Price * Ward 2	Increase by a factor	1
Price * Ward 3	Increase by a factor	1.000001
Price * Ward 4	Increase by a factor	1.000001
Price * Ward 5	Decrease by a factor	0.0000003313001

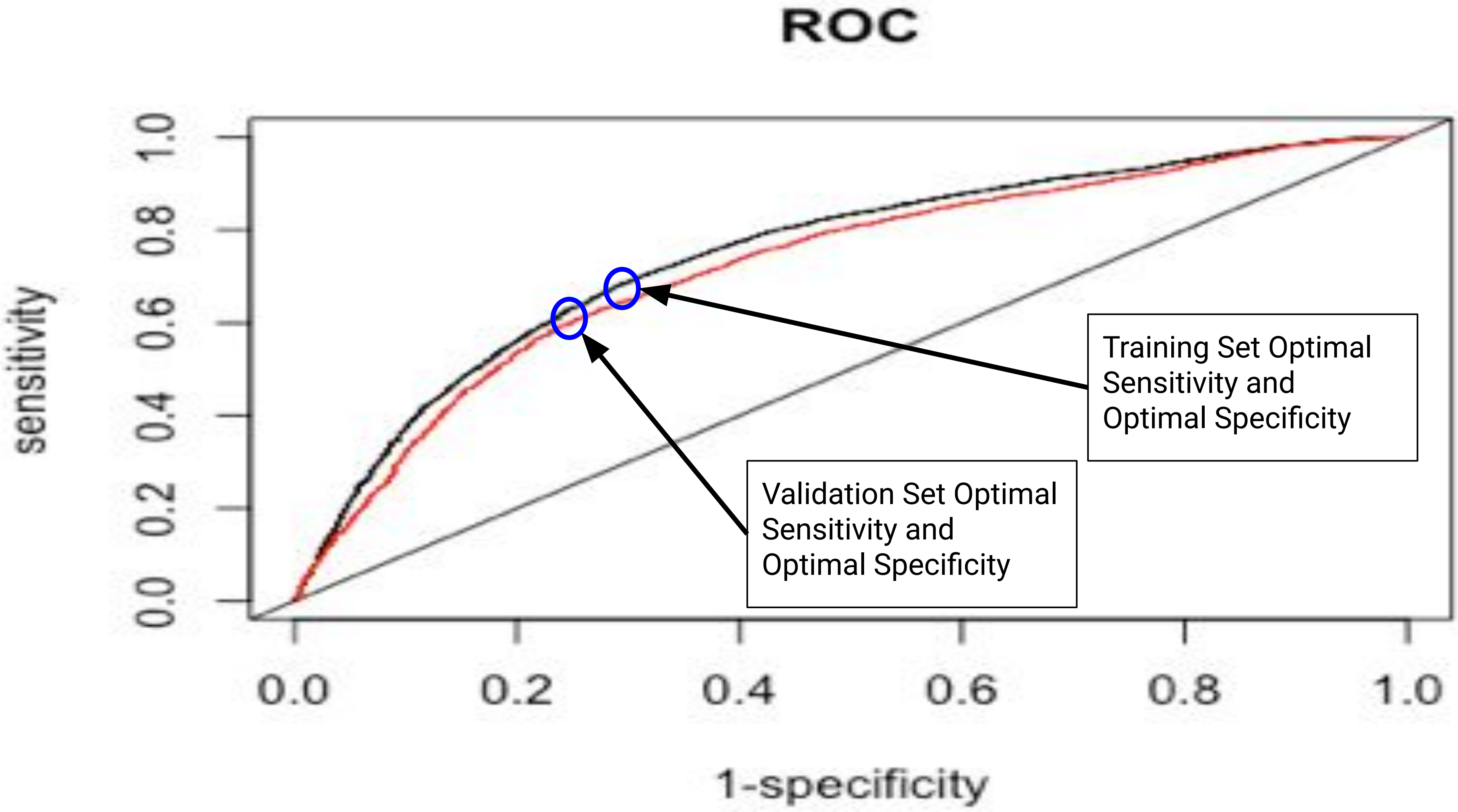
ROC Curve

Training Set:

- Area = 0.7497219
- Cutoff = observation 18551,
    - C = 0.8461571
- Optimal Sensitivity = 0.6789462
- Optimal Specificity = 0.7118789

Validation Set:

- Area = 0.7233407
- Cutoff = Observation 7238
    - C = 0.8782725
- Optimal Sensitivity = 0.595809
- Optimal Specificity = 0.7596588



Training Set Optimal Sensitivity and Optimal Specificity

Validation Set Optimal Sensitivity and Optimal Specificity

Conclusion

- Model could be better
  - Lower AIC & BIC (below 200)
- ROC is good but we would like the area to of been above .85
- Multicollinearity Issues:
  - Still have problems with the interaction terms and the single variables being highly correlated with one another
  - The Price and square root of price are highly correlated with one another
  - The rooms and the rooms^0.2 are highly correlated with one another
- Data is still flawed = Bias is increased

Future Work

- Do more time analysis
- Maybe do some sentiment analysis on the street, neighborhood, and State one lives in
- Try to see if we can hear back on what qualification meant in the dataset
- Add a few more variables
  - Heat, interaction terms of heat and AC
- Collect data from realtors websites and fill in the information ourselves
- Add neighborhood rating, neighborhood review
- Collect data from the surrounding states (West Virginia, Virginia, Maryland)

It's All About the Money

Work Cited

1. McKay, Allie W. “Farmers' Markets vs. Food Deserts: Which Are Winning in DC?” *The Capital's Markets*, 31 July 2014, thecapitalsmarkets.wordpress.com/2014/07/31/farmers-markets-vs-food-deserts-which-is-winning-in-dc/.
2. Johnson, Matt. “Washington's Systemic Streets.” *Greater Greater Washington*, ggwash.org/view/2530/washingtons-systemic-streets.
3. “Money Is The Root Of All Evil Stock Photos and Images.” *Alamy*, www.alamy.com/stock-photo/money-is-the-root-of-all-evil.html.

R Shiny

- https://aaronniecestro.shinyapps.io/DC-Housing/

References

1. “Types of Housing Models and Programs.” *The 519*, www.the519.org/education-training/lgbtq2s-youth-homelessness-in-canada/types-of-housing-models-and-programs.
2. Dobbins, Tim, and John Burke. “Predicting Housing Prices with Linear Regression Using Python, Pandas, and Statsmodels.” *Learn Data Science - Tutorials, Books, Courses, and More*, www.learndatasci.com/tutorials/predicting-housing-prices-linear-regression-using-python-pandas-statsmodels/.
3. Corsini, Kenneth Richard. “STATISTICAL ANALYSIS OF RESIDENTIAL HOUSING PRICES IN AN UP AND DOWN REAL ESTATE MARKET: A GENERAL FRAMEWORK AND STUDY OF COBB COUNTY, GA .” *A Thesis Presented to The Academic Faculty*, Georgia Institute of Technology, Dec. 2009, smartech.gatech.edu/bitstream/handle/1853/31763/Corsini_Kenneth_R_200912_mast.pdf.
4. “Regression Data for Inclusionary Housing Simulation Model | DataSF | City and County of San Francisco.” *San Francisco Data*, data.sfgov.org/Economy-and-Community/Regression-data-for-Inclusionary-Housing-Simulatio/vcwn-f2xk/data.
5. Leonard, Kimberlee. “What Forms Are Needed to Sell a Home by Owner?” *Home Guides | SF Gate*, 29 Dec. 2018, homeguides.sfgate.com/forms-needed-sell-home-owner-7271.html.
6. Leonard, Kimberlee. “What Is the Procedure for Closing a for Sale by Owner House Sale?” *Home Guides | SF Gate*, 15 Dec. 2018, homeguides.sfgate.com/procedure-closing-sale-owner-house-sale-65511.html.