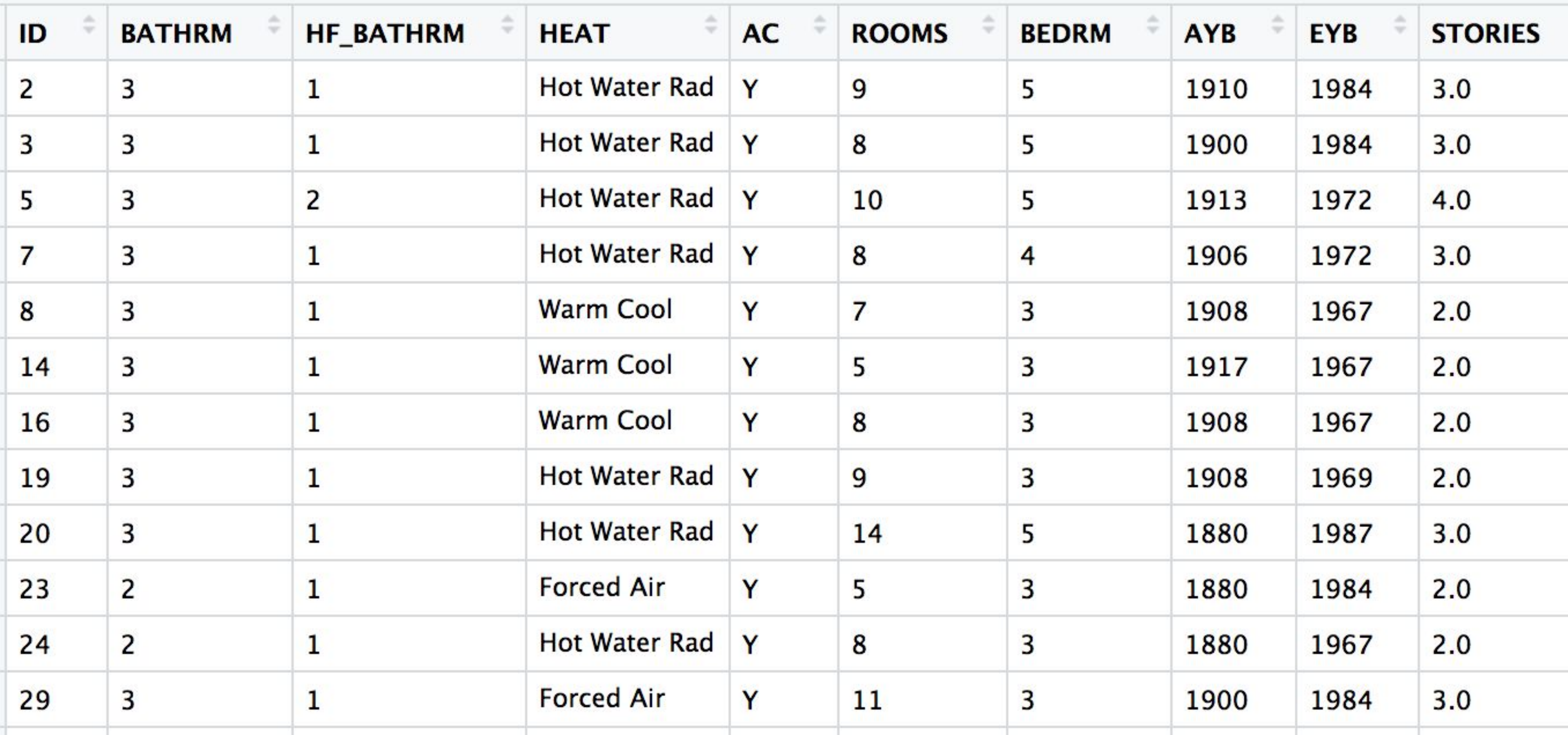
DC Properties Qualification using Logistic Regression

By Aaron Niecestro

Background and Story

- We all are American University students
- We all live either in DC or the surrounding states?
    - I apologize in advance if I have this wrong
- This type of regression for property models was not conducted or analyzed previously or at least what I researched
- 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

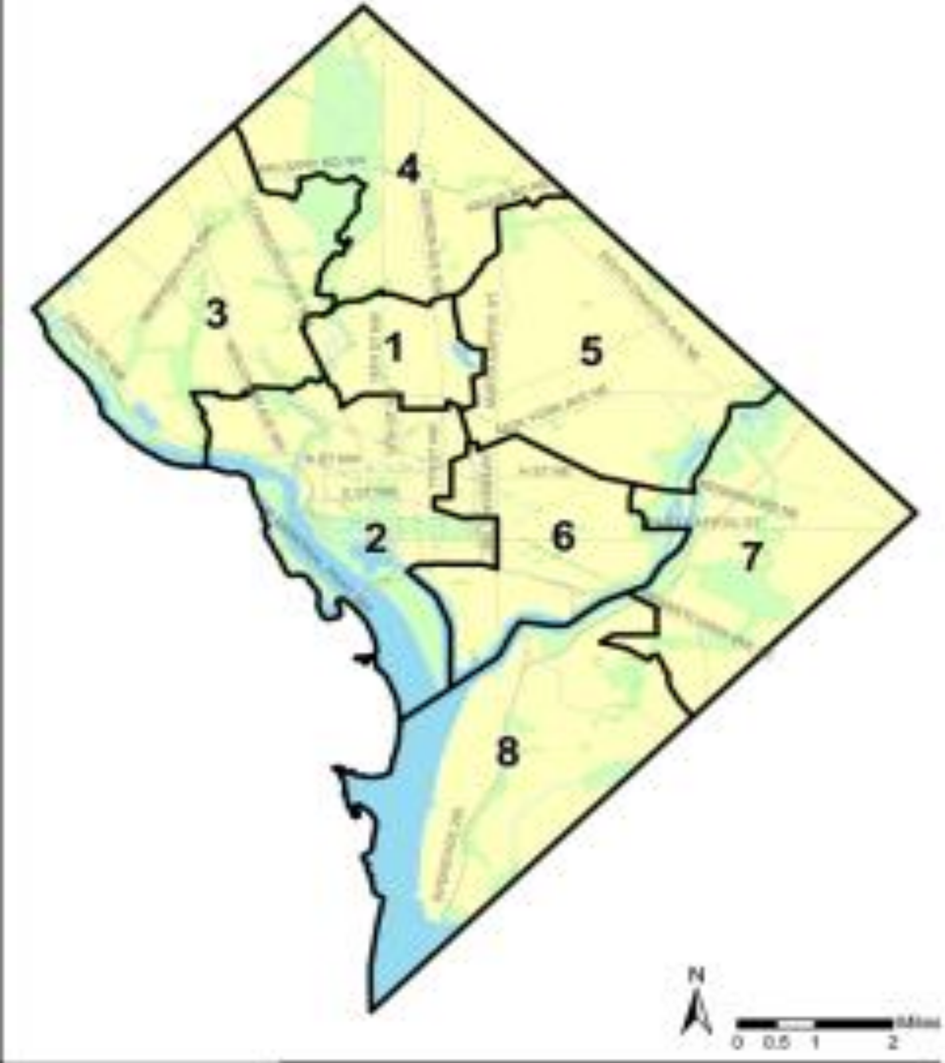
A Sneak Peak of the Data Set

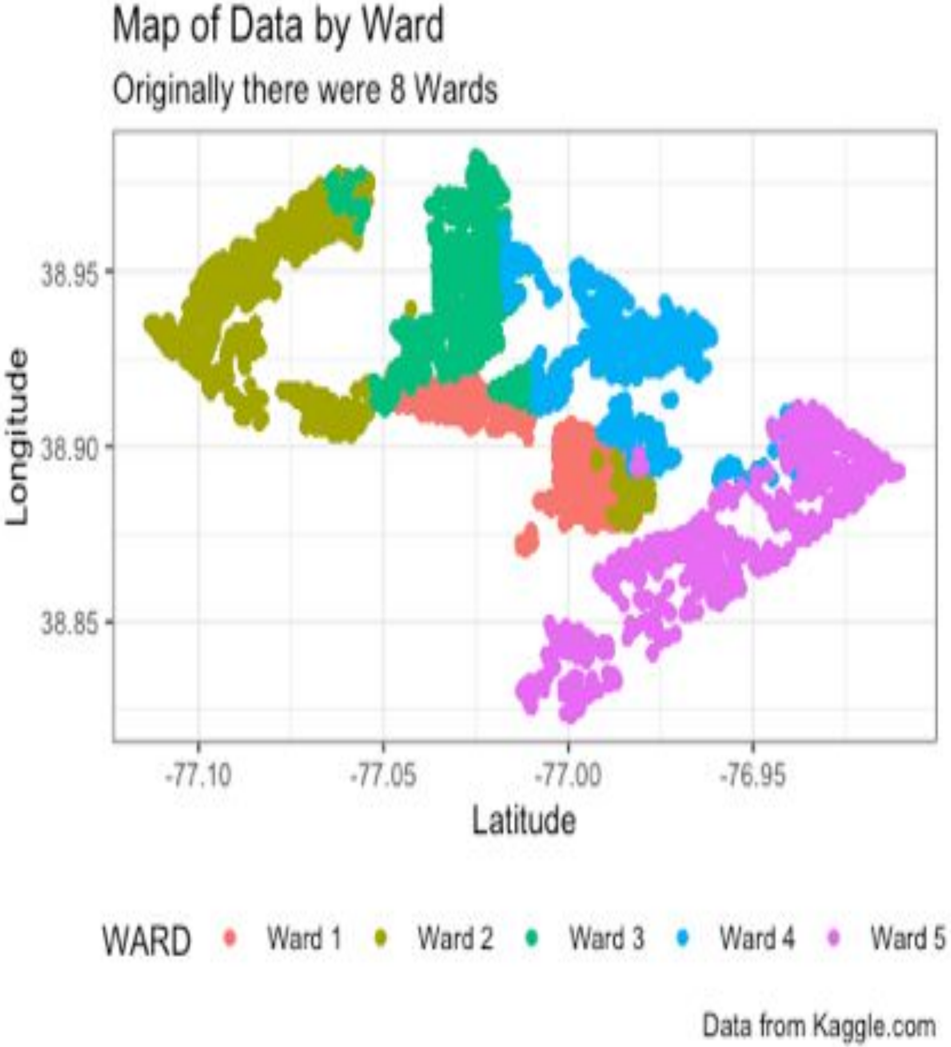


ID	BATHRM	HF_BATHRM	HEAT	AC	ROOMS	BEDRM	AYB	EYB	STORIES
2	3	1	Hot Water Rad	Y	9	5	1910	1984	3.0
3	3	1	Hot Water Rad	Y	8	5	1900	1984	3.0
5	3	2	Hot Water Rad	Y	10	5	1913	1972	4.0
7	3	1	Hot Water Rad	Y	8	4	1906	1972	3.0
8	3	1	Warm Cool	Y	7	3	1908	1967	2.0
14	3	1	Warm Cool	Y	5	3	1917	1967	2.0
16	3	1	Warm 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 does Qualification variable mean?
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?**

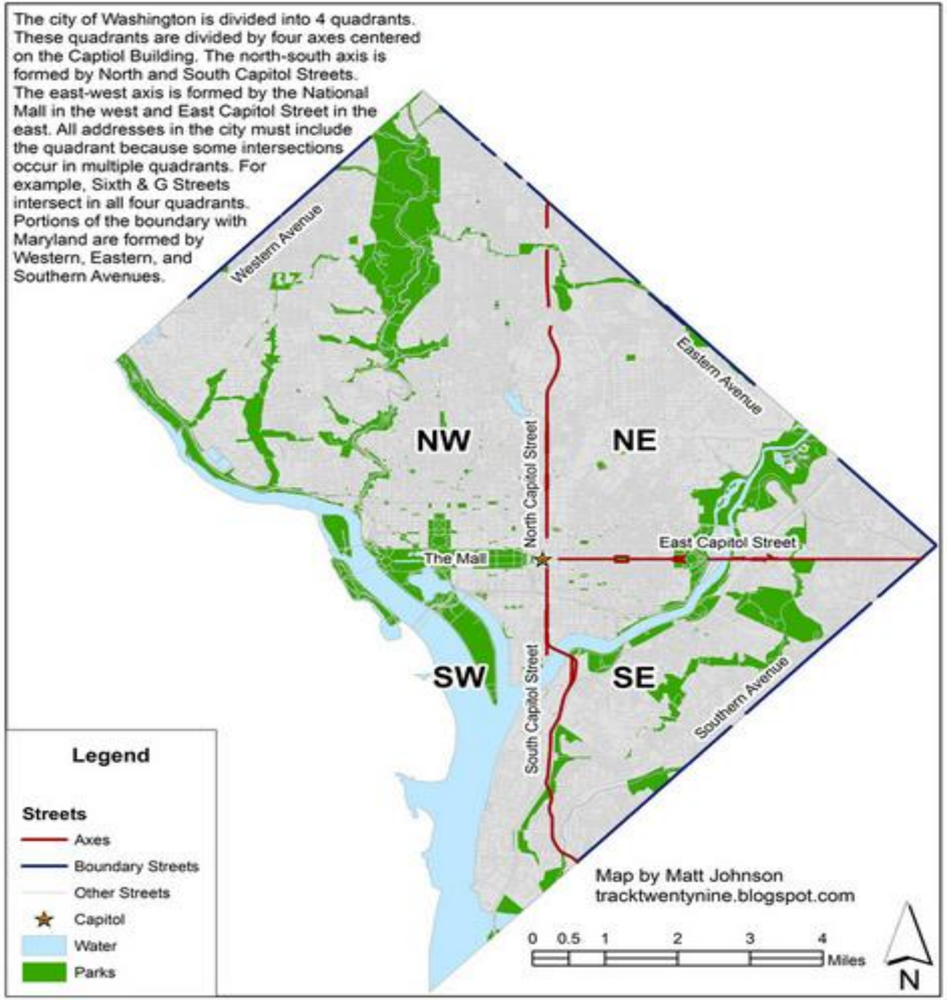



Map of Data by Ward

Originally there were 8 Wards

Data from Kaggle.com

Transportation: Boundaries and Axes

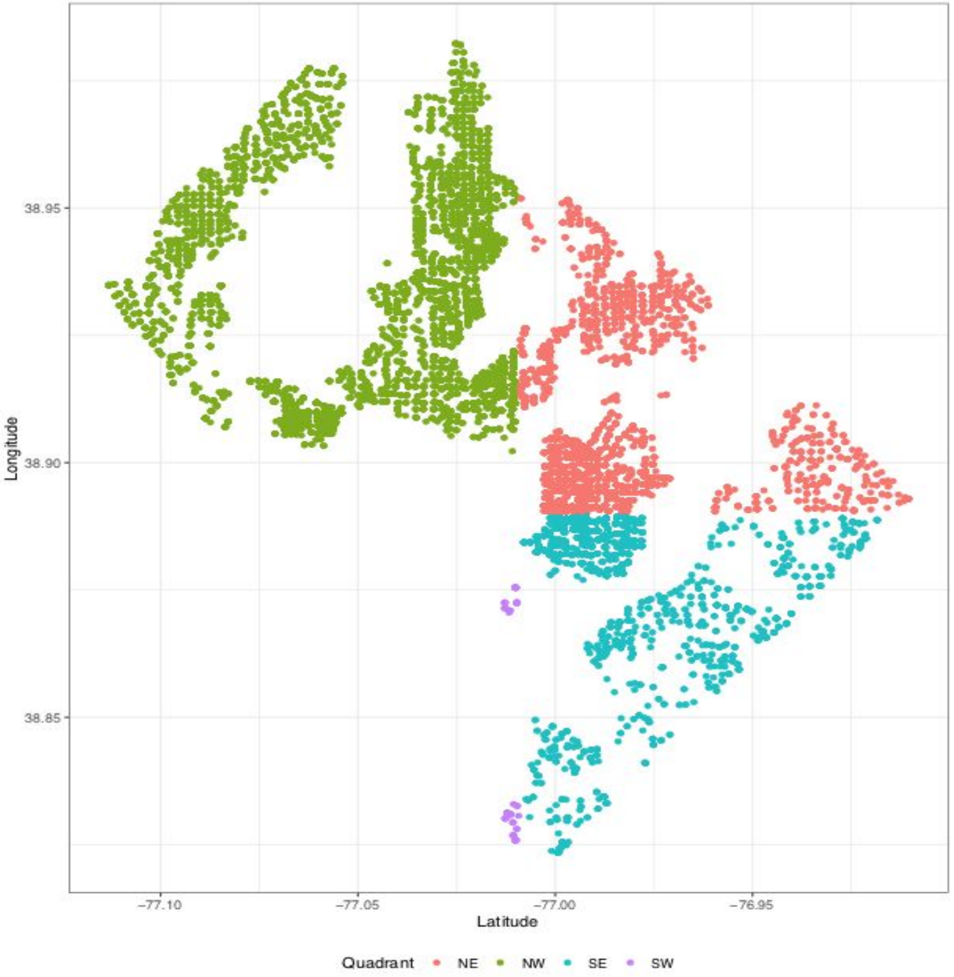


The city of Washington is divided into 4 quadrants. These quadrants are divided by four axes centered on the Captiol Building. The north-south axis is formed by North and South Capitol Streets. The east-west axis is formed by the National Mall in the west and East Capitol Street in the east. All addresses 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

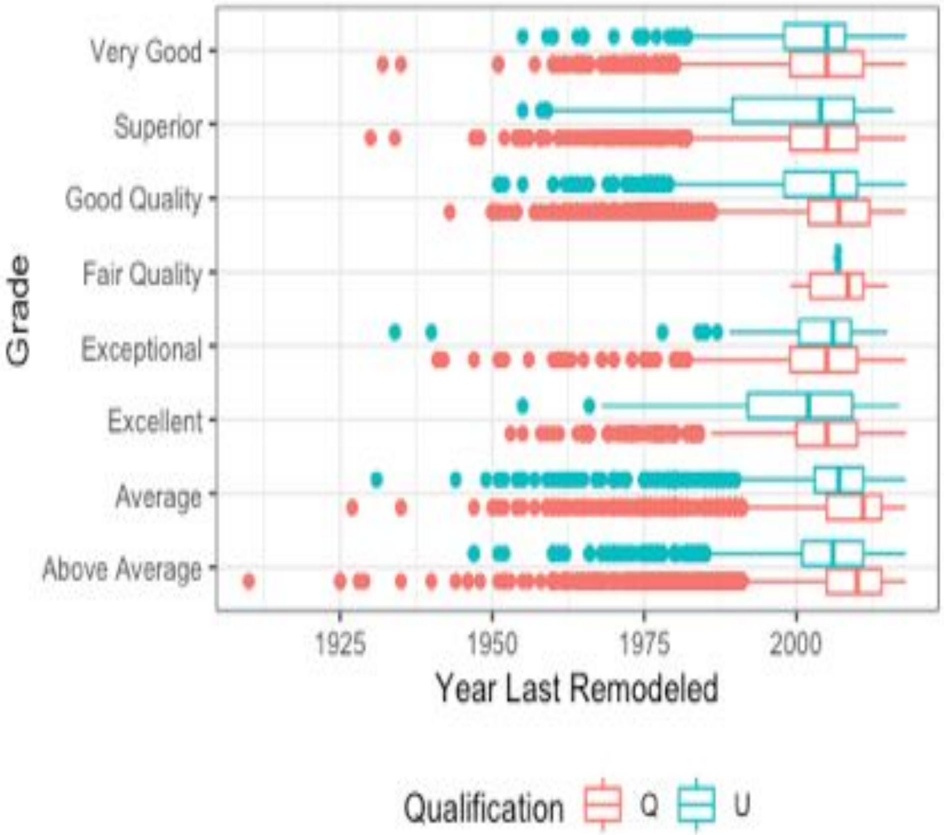
Little to no South−West area



Data from Kaggle.com

Price based on Qualifications and Grade

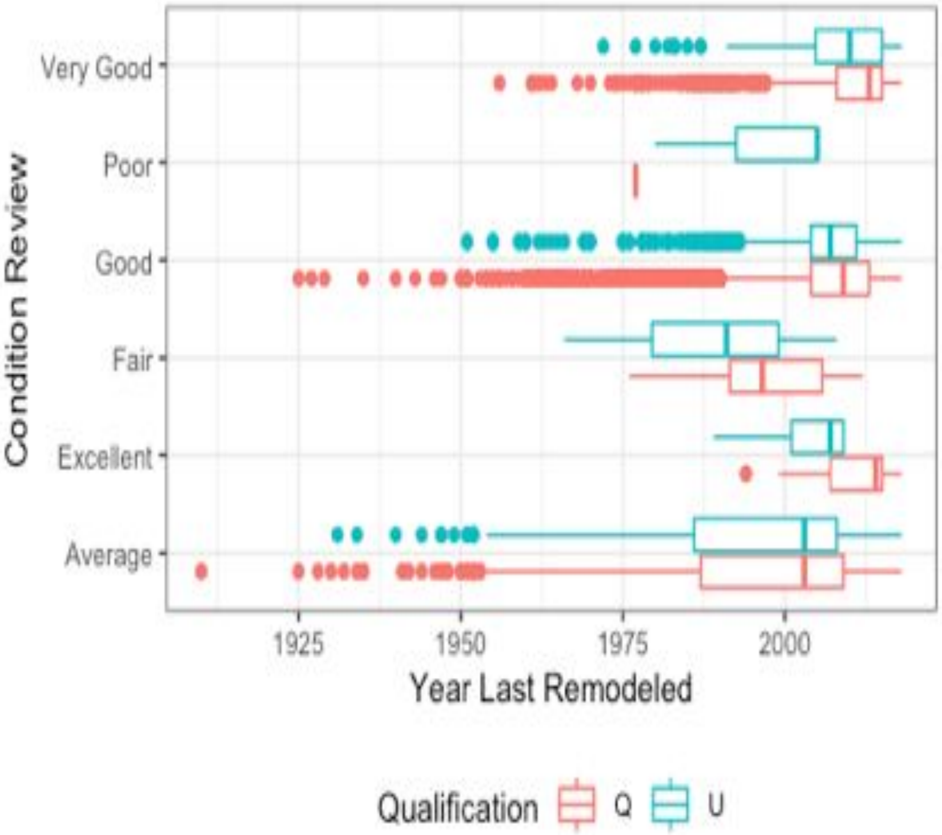
Ward 1 to 3 are the most expensive



Data from Kaggle.com

Price based on Qualifications and Condition

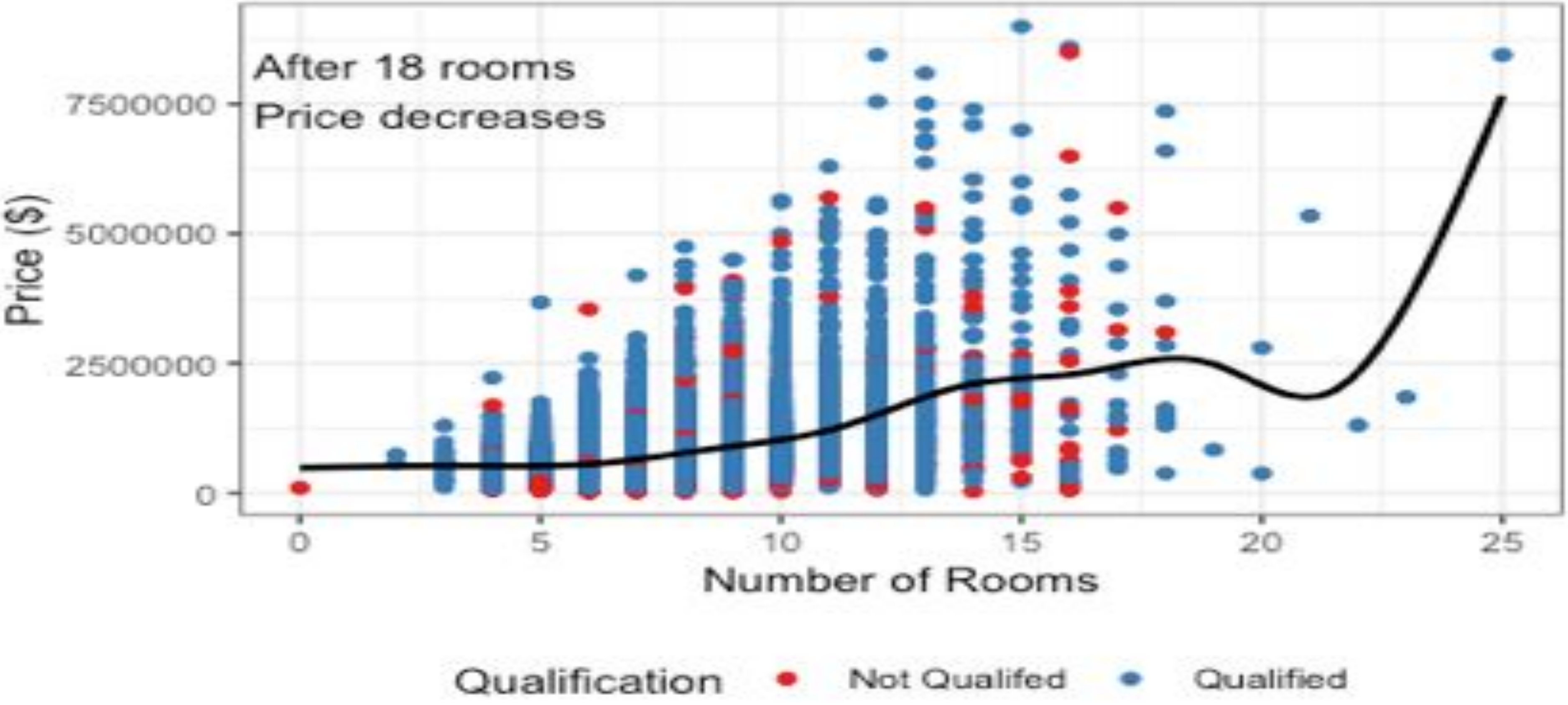
Ward 1 to 3 are the most expensive



Data from Kaggle.com

Price Increases as Rooms Increases

A Majority of Qualified Places to live are less then 1 Million Do

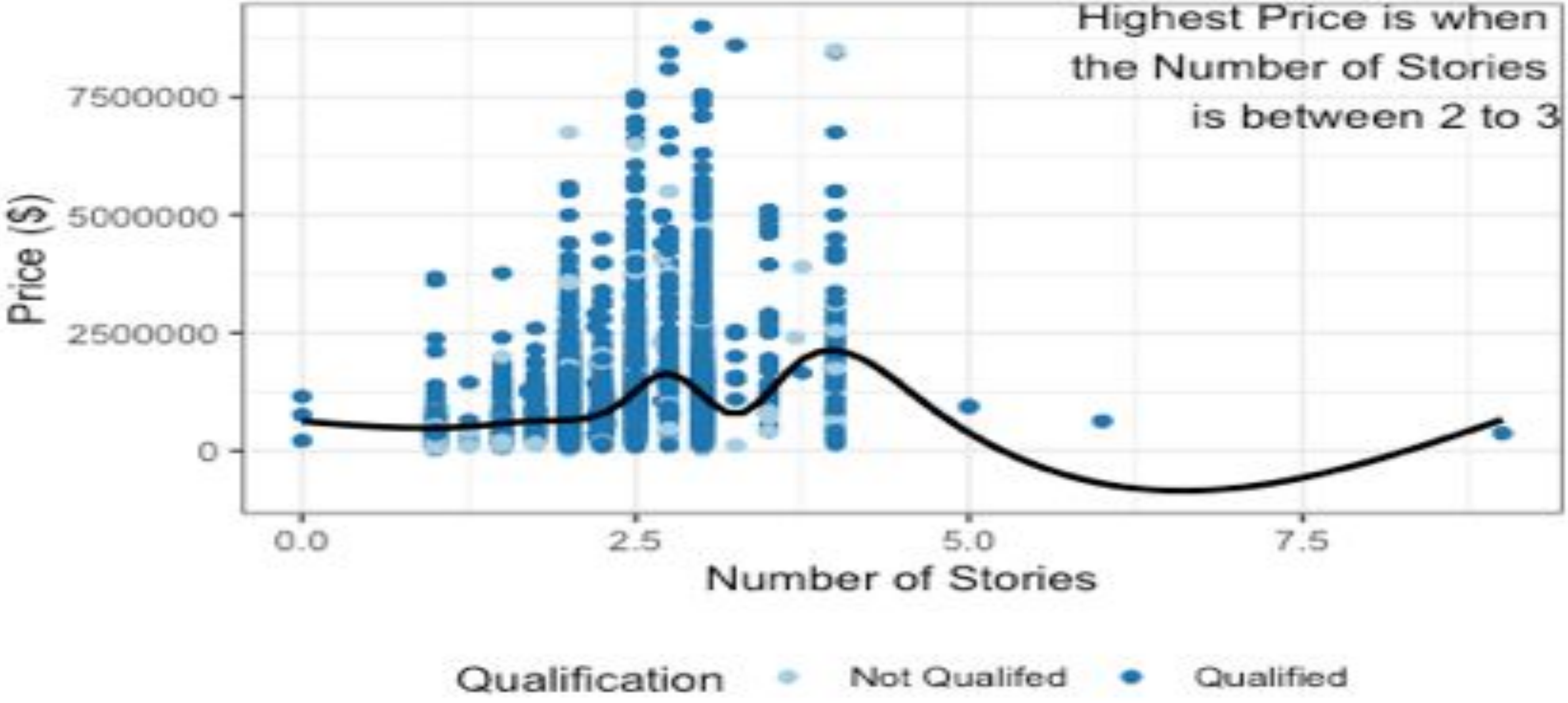


After 18 rooms
Price decreases

Data from Kaggle.com

Stories Increase & Dramatic Decrease with Price

Having between 1 and 5 Stories = Higher Price

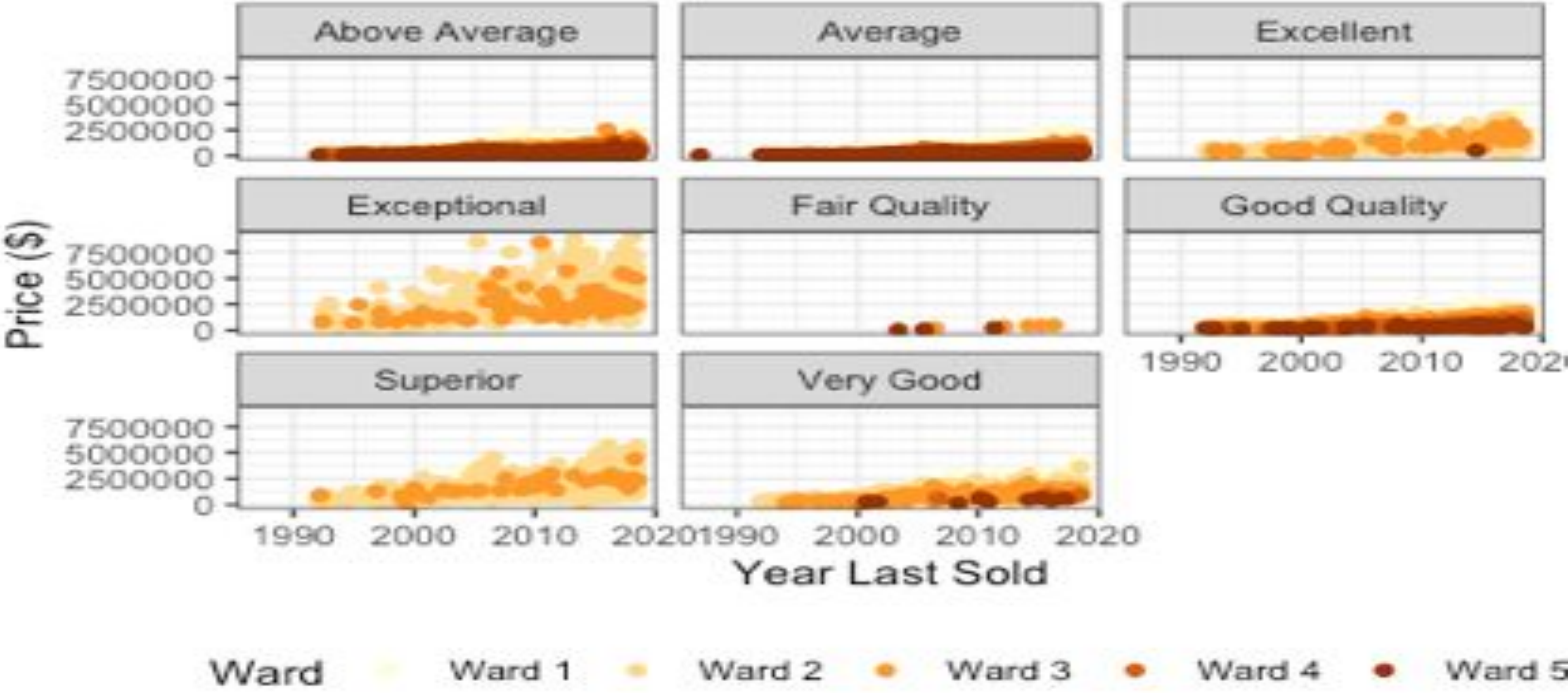


Highest Price is when the Number of Stories is between 2 to 3

Data from Kaggle.com

Stacking of Ward Areas over Years by Grade

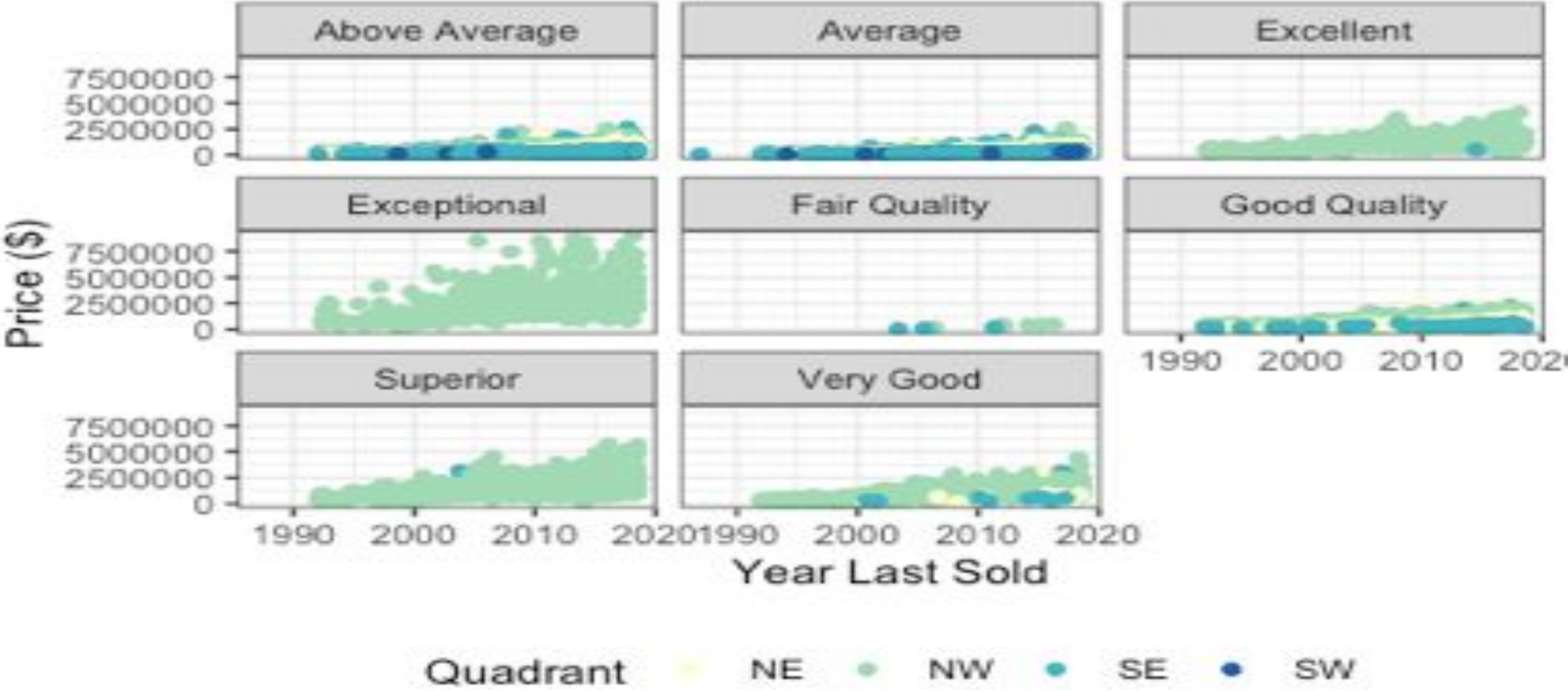
The lower the ward number you are in the higher the price will



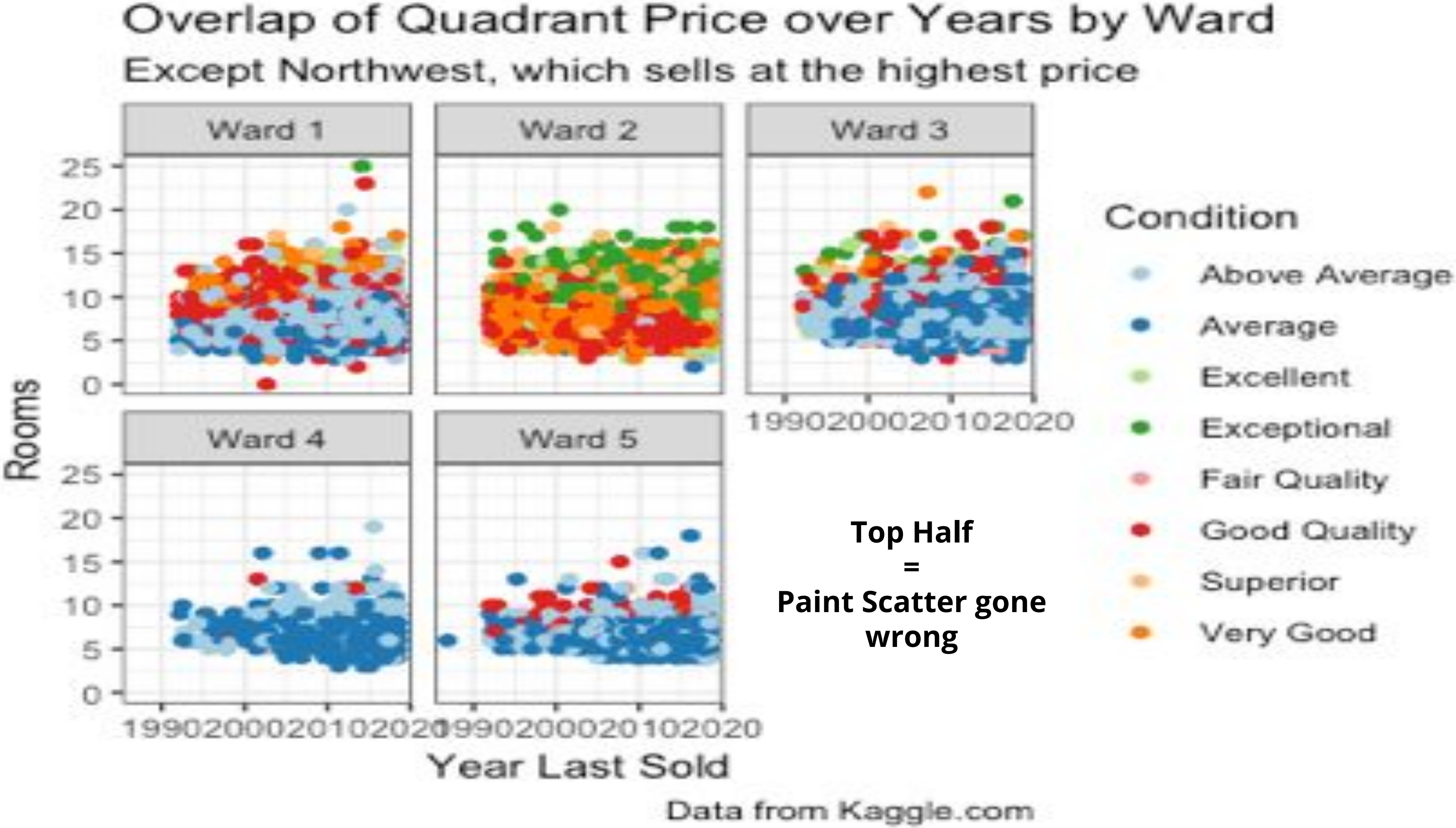
Data from Kaggle.com

Overlap of Quadrant Price over Years by Grade

Except Northwest, which sells at the highest price



Data from Kaggle.com



Top Half
=
Paint Scatter gone wrong

Final Model Equation

$$
\begin{aligned}
&\text{Log}(\boldsymbol{\pi}/1-\boldsymbol{\pi}) = -\ 3.158 - 0.000004374*\text{Price} + 0.007901*\sqrt{}\text{Price} - 0.2907*(\text{AC=Yes}) + \\
&0.1821*\text{Rooms} + 2.221*(\text{Rooms}\hat{}0.2) - 0.04816*\sqrt{}\text{BEDRM} + \\
&0.1069*(\text{CNDTN=Excellent}) - 1.196*(\text{CNDTN=Fair}) + 0.2849*(\text{CNDTN=Good}) - \\
&1.373*(\text{CNDTN=Poor}) + 0.6739(\text{CNDTN=Very Good}) - 0.4306*(\text{Ward=2}) - \\
&0.2858*(\text{Ward=3}) - 0.0515*(\text{Ward=4}) + 0.2901(\text{Ward=5}) + \\
&0.000001085\text{PRICE}*(\text{AC=Yes}) + 0.00000002.698*(\text{PRICE*ROOMS}) + \\
&0.0000003.897(\text{Price*Ward=2}) + 0.0000005486*(\text{Price*Ward=3}) + \\
&0.000001052*(\text{Price*Ward=4}) - 0.0000003.313*(\text{Price*Ward=5})
\end{aligned}
$$

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)
    - Broken up each relationship so one can see it individually but when you interpret it with categorical variables have to add single continuous variable number as well to get final outcome

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

Answers

1. Paperwork, loans to bank are approved, and inspection is passed
2. Paperwork, can add for the property, money
3. Yes, property price the most important factor in determining whether a property is qualified to go on the market
4. Yes, but not our definition of qualification since the data shows otherwise
5. Maybe, it is very hard to say
6. We combined some of the previous linear regression housing model approaches for predictor variables and tried to create a new model using what we thought was important
7. Sadly, it seems money is the most important thing (my belief from this study anyway)

Conclusion

- Model could be better
- Still have some 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
- Still had a lot of bias in our model
  - Wards & Qualification
  - Variable choice
  - Data was only from Washington DC
- **Tried to make the Best out of the Worst**

Further Work and Analysis

- More time and location analysis
- Some sentiment analysis on the street and neighborhood
- 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
- Think about creating and collecting data from realtors websites or add information to the existing dataset
- Find and add neighborhood rating & neighborhood review
- Collect data from the surrounding states (West Virginia, Virginia, Maryland)

100
500
500

Work Cited

1. McKay, Allie W. “Farmers' Markets vs. Food Deserts: Which Are Winning in DC?” *The Capital 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.

Questions?