Stat 444 – Function Estimation

An Analysis of Washington D.C. Housing Prices

Jack Swiatoschik

April 18, 2020

Introduction

What follows is a discussion of a data exploration and modelling exercise leveraging Washington D.C. residential property dataset. This dataset is current up to July 2018. The goal of this this analysis is generating a model to predict the price of the single-family home within Washington D.C. using the success metric of Root Mean Squared Log Error. Other goals were understanding the factors which best influence the price of a home and identifying other interesting learnings.

Washington D.C. is the capital of the United States of America making it an important city in international affairs, and a popular tourist destination. The population is approximately 700,000. Being a federal capital, the government is a major employer of the city which is also home to unions, lobbying firms and non-profits. Washington D.C. is among the most expensive cities in the U.S. and it's GDP per capita is 3 times that of the next highest state, which we expect to inflate property values. Due to building height restrictions inside the city, many major companies are headquartered just outside the city in cities in Maryland and Virginia. While this broader Washington D.C metropolitan area contains sections of these states and as well as parts of West Virginia, the dataset considered for analysis only considers homes residing inside the District of Columbia.

Dataset Description

The Data set contains 29 variables (28 Predictors, and 1 Response) and a preliminary round of data processing has already taken place. An outline of the variables is given below as well as how they are presented in the dataset:

- Price/ PRICE (Response Variable)
- Number of Bathrooms/ BATHRM
- Number of Half Bathrooms/HF_BATHRM
- Type of Heating in the Home/ HEAT
- Whether Air Conditioning is in the house/ AC
- Number of Rooms/ ROOMS
- Number of Bedrooms/ BEDRM
- The earliest time the main portion of the building was built/ AYB
- Year structure was remodeled/YR_RMDL
- The year an improvement was built more recent than actual year built/ EYB
- Number stories in primary dwelling/ STORIES Date of most recent sale/ SALEDATE
- Gross building area in square feet/ GBA
- Best description of style of the home (bi-level, 3 stories etc.)/ STYLE
- Grade of the home/ GRADE
- Condition of home (good, very good etc.)/ CNDTN Type of Exterior wall/ EXTWALL
- Type of Roof/ ROOF Type of Interior wall/ INTWALL
- Number of Kitchens/ KITCHENS
- Number of Fireplaces/ FIREPLACES
- Land area of property in square feet/ LANDAREA
- Zip code/ ZIPCODE
- Latitude/ LATITUDE
- Longitude/ LONGITUDE
- Assessment Neighbourhood (Ex: Foggy Bottom)/ ASSESSMENT_NBHD
- Assessment Sub-Neighbourhood/ ASSESSMENT_SUBNBHD
- Ward out of 8 Wards in the District/ WARD
- Quadrant of the District (DC is split into 4 either NW,SW,NE,SE)/ QUADRANT

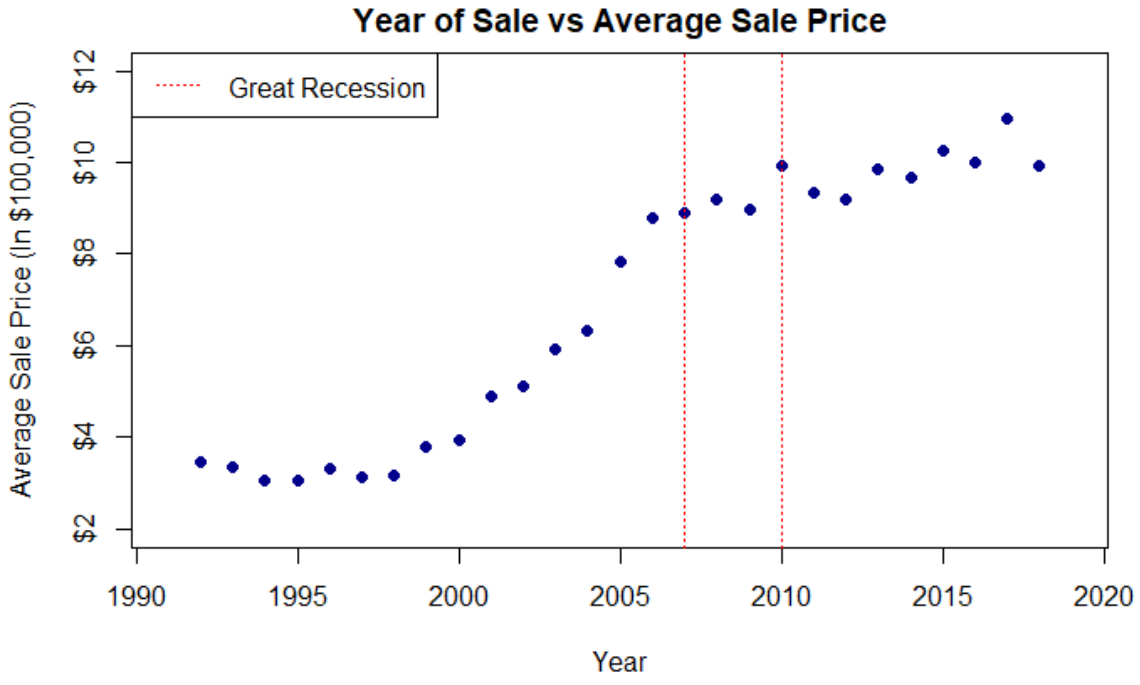
Data Exploration

Sales Over time

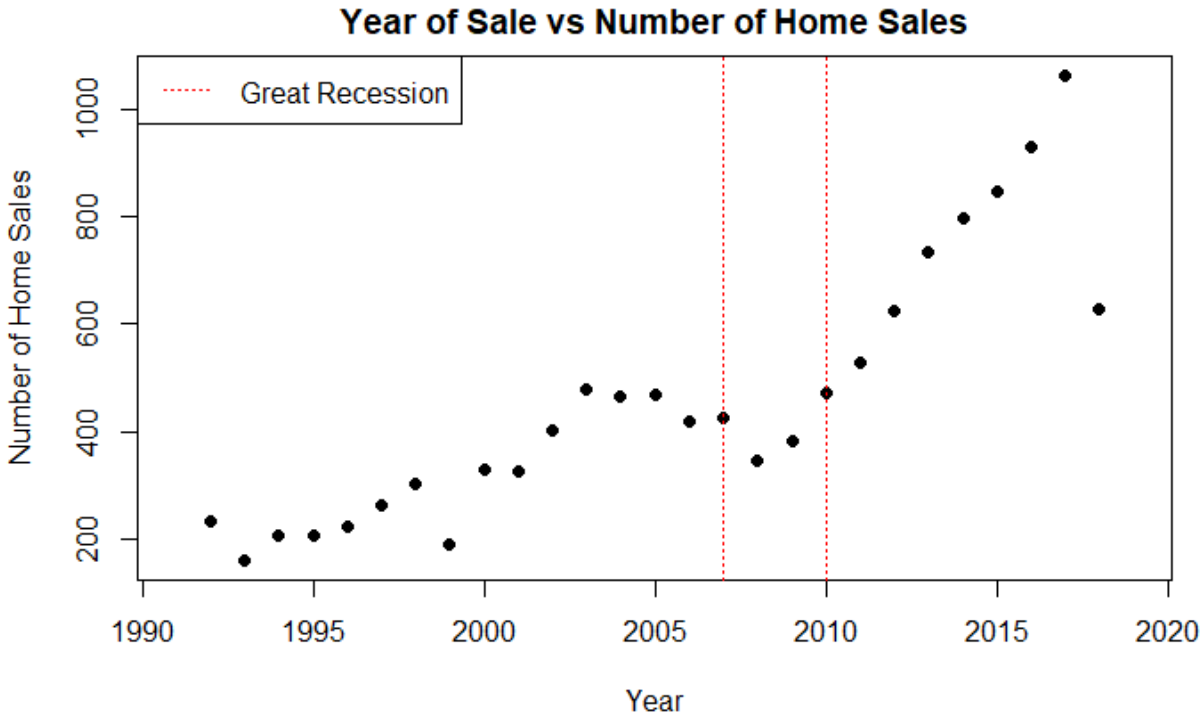
The below graph explores how average sale prices have changed over time. Prices were consistent between 1992 to 1999, increased almost $500,000 between 2000 and 2009, and have been slowly increasing since. The increase is possible due to the housing bubble in 2000's but surprisingly there is no sharp crash between 2007-2009 as was seen in other areas of the country. As the local economy is stabilized by the presence of the recession-immune federal government, it would make sense why the economy and housing market was not particulately impacted as hard as other areas. However, it is interesting how the region still felt the increase in house price during the bubble. What is particularly interesting, is the nearby regions of the D.C. Metro area were not as insulated from the crash.

https://www.nvar.com/realtors/news/re-view-magazine/article/sep-oct-2017/2017-09-10-market-metrics-home-sales-prices-continue-toskyrocket

This would suggest that as the private sector businesses located outside Washington D.C. flourished, it had an upward impact on the local housing market, including Washington itself. But as the Great Recession occurred, Washington D.C. itself was able to stay insulated enough as an economy and the housing market was able to stay fairly flat.

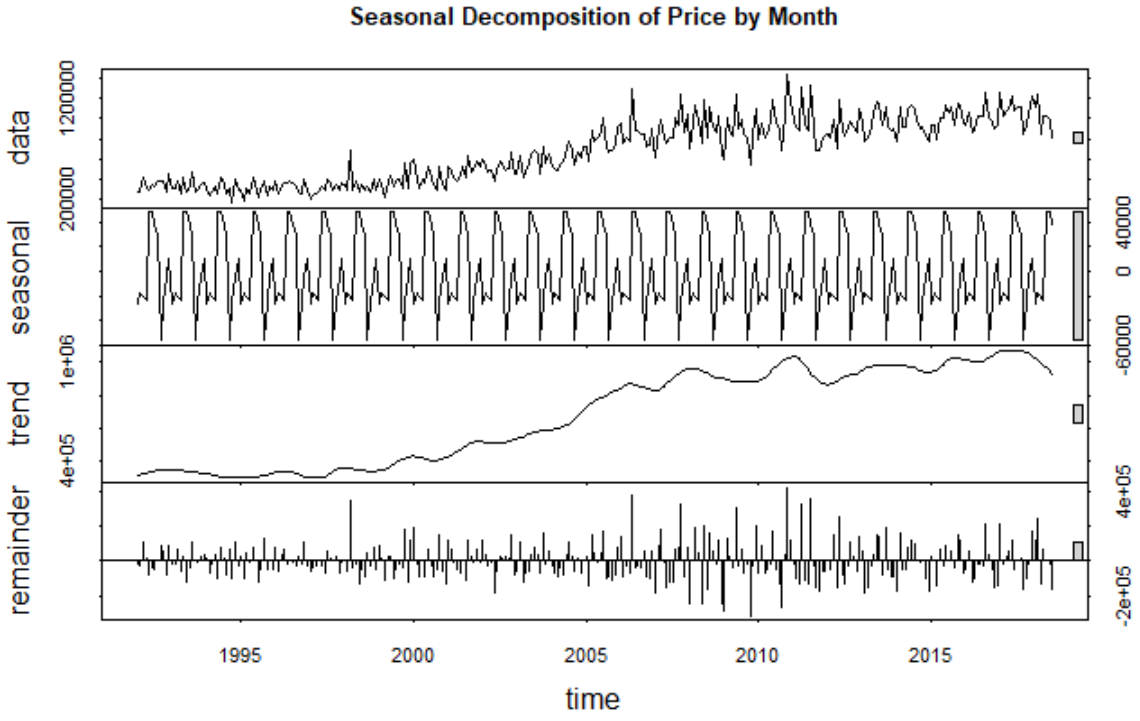


The below graph explores how the number of home sales per year has changed over time in the below graph. A clear outlier exists in 2018, and this stems from the fact that the data is only current up to July 2018. However, after these 7 months, 2018 is on pace to be higher than 2017, so the trend seen from 2008-2017 can be expected to continue in this year. There does not seem to be any correlation between the Number of Sales data, and the Average Price Data. From 2000-2010, there is a 'hump' in the number of sales, but the prices skyrocket. In this dataset, a dip can be seen in 2007-2009 which corresponds with the Great Recession. After 2010, prices flatline, while number of sales increases. It is possible, that the increased number of sales would flood supply and flatline prices. A potential concern illustrated here, is the there are farm more observations from recent years as compared to previous years. This could potentially be a data availability issue, and there are more sales especially during the 1990s which are simply not included in this dataset.



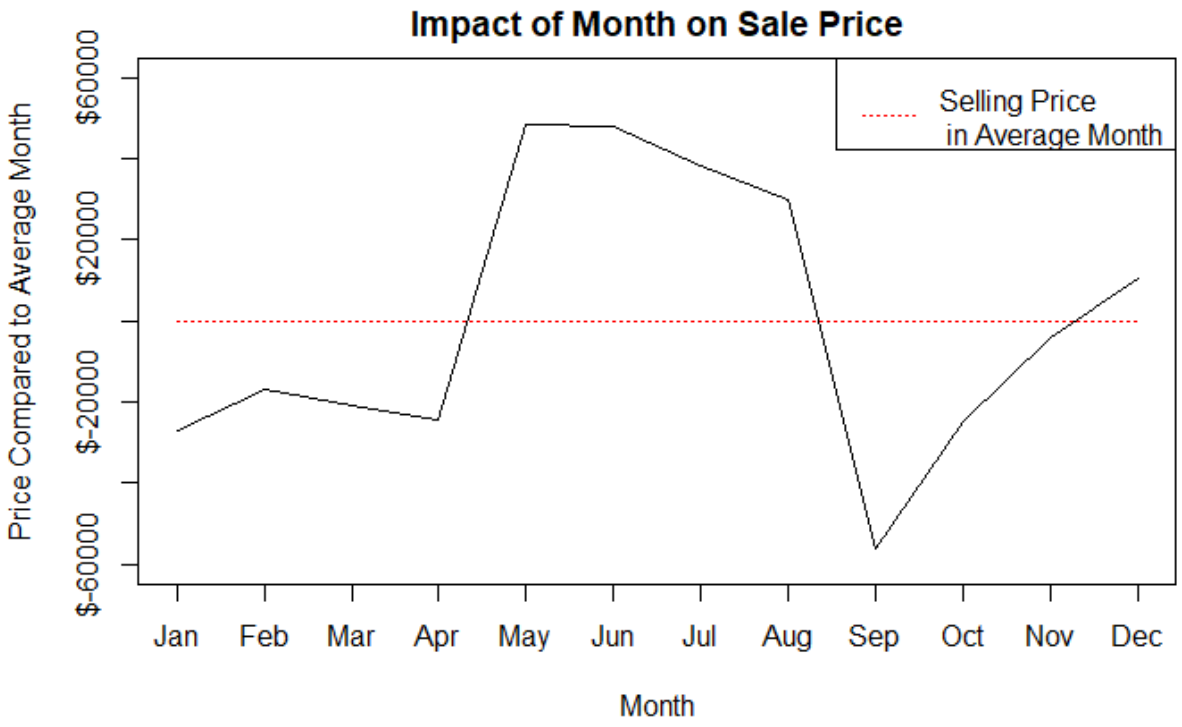
Seasonally Decomposing Sale Price

A next step would be to see how price and number sales change month to month. Simply averaging by month is not as appropriate as year, since a house sale in January 1992 would be grouped with a house in January 2005, and if there were more sales in specific months in specific years they interpretation of results could be skewed by the long term year-to-year trend. It then makes sense to seasonally decompose our data to get that insight. We begin by looking at prices.

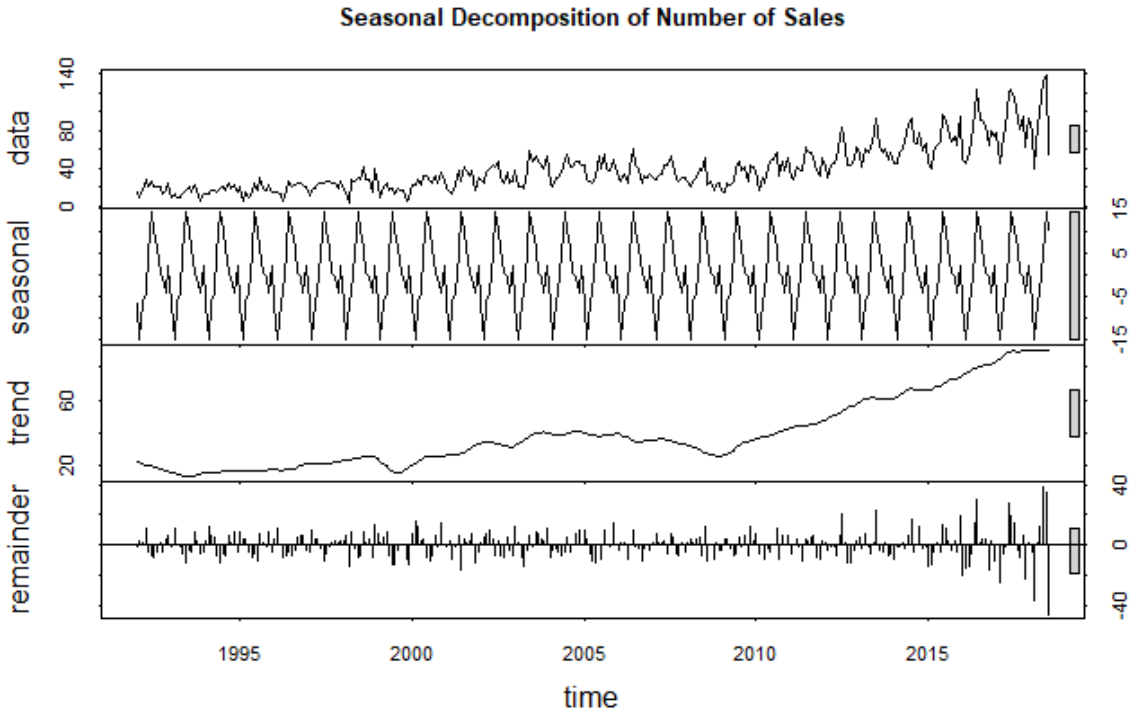


As the bars on the right-hand side are of equal size in terms of dollars, year-to-year changes have a larger impact than month-to-month price changes. The error has a large influence meaning simply using month and year is not a terrific predictor. Nevertheless, it is interesting too see which months see homes sell for more.

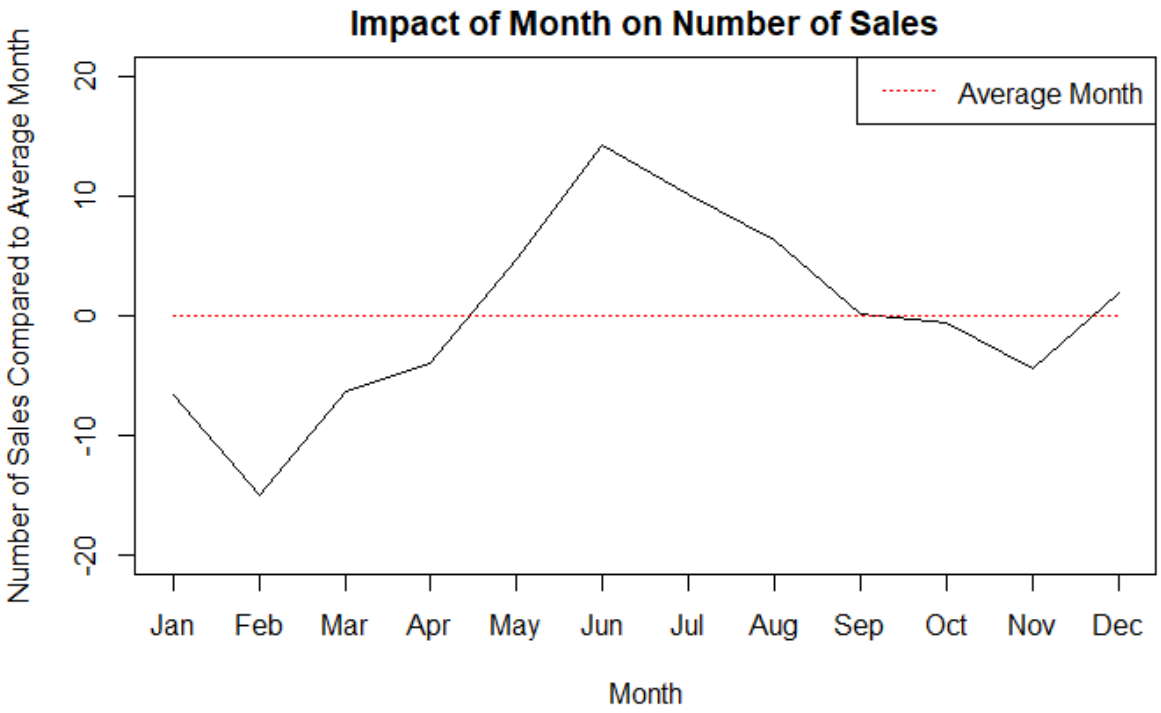
Houses sell higher than average during the summer months (May-August), as well as in December. September is the worst month to sell in. A potential reasoning would be that parents would be reluctant to move during the school year, and especially at the beginning of the school year in September. This would reduce the list of potential buyers and cause fewer bidding wars which would reduce prices. Another possible explanation is that moving is easier in the summer months due to better weather, and people are more inclined to deal with the logistics involved when weather is favourable. However, because month to month seasonality does not have a massive impact on prices as shown by the seasonal decomposition, these insights should be discounted.



Seasonally Decomposing Sale Price



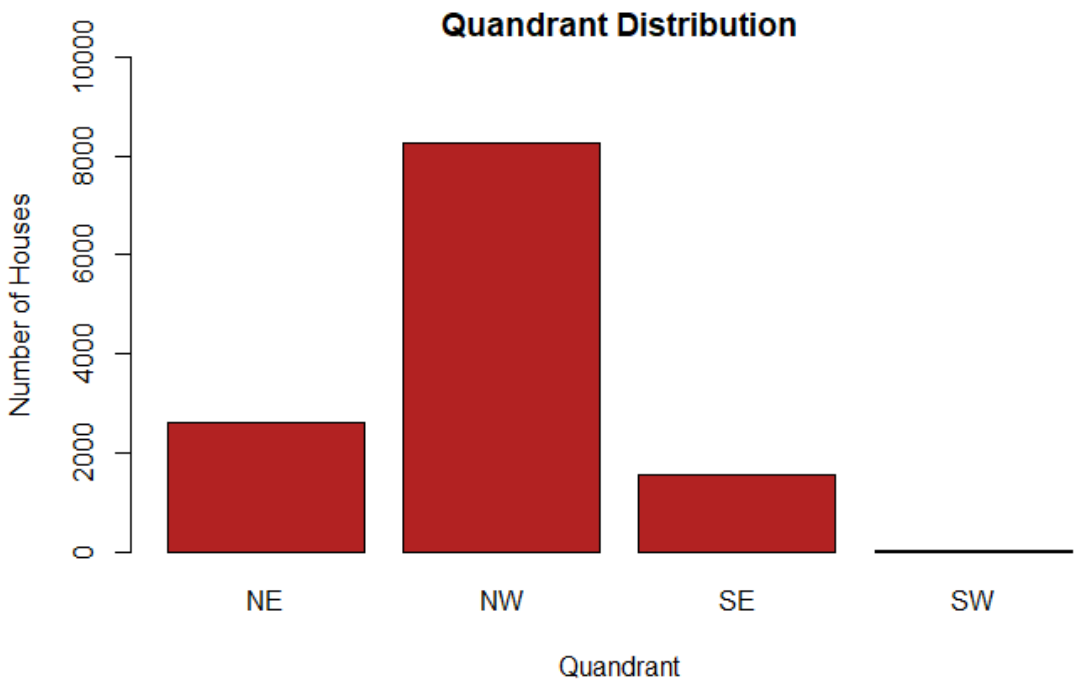
The bars on the right-hand side are of equal size in terms of number of house sales. Again, as in the seasonal decomposition of prices, year-to-year changes have a larger impact than month-to-month price changes. The error has a large influence meaning simply using month and year is not a terrific predictor. Nevertheless, it is interesting too see which months see more home sales.



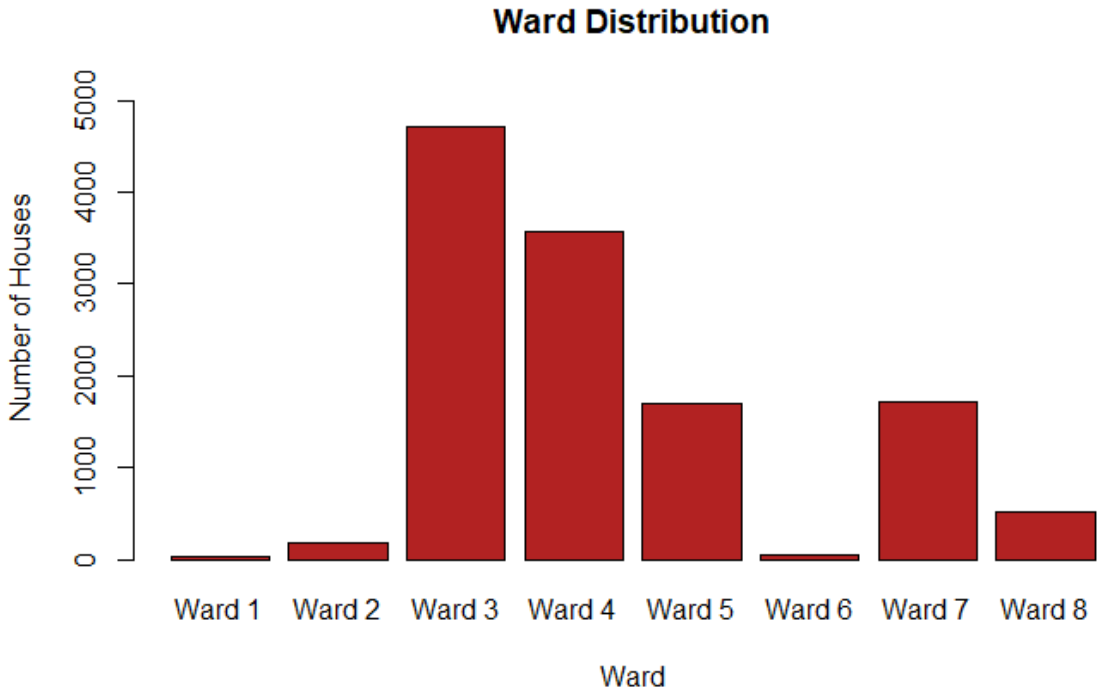
The monthly trend is similar for number of sales as it is to the price. The summer months (May-August) as well as December see a larger number of house sales. This suggests that more people are interest in buying houses in the summer, and because of this more people are interested in selling hence more sales. An interesting difference is that September is an average month in terms of number of homes sold, but well below average in the average price.

Seasonally Decomposing Sale Price

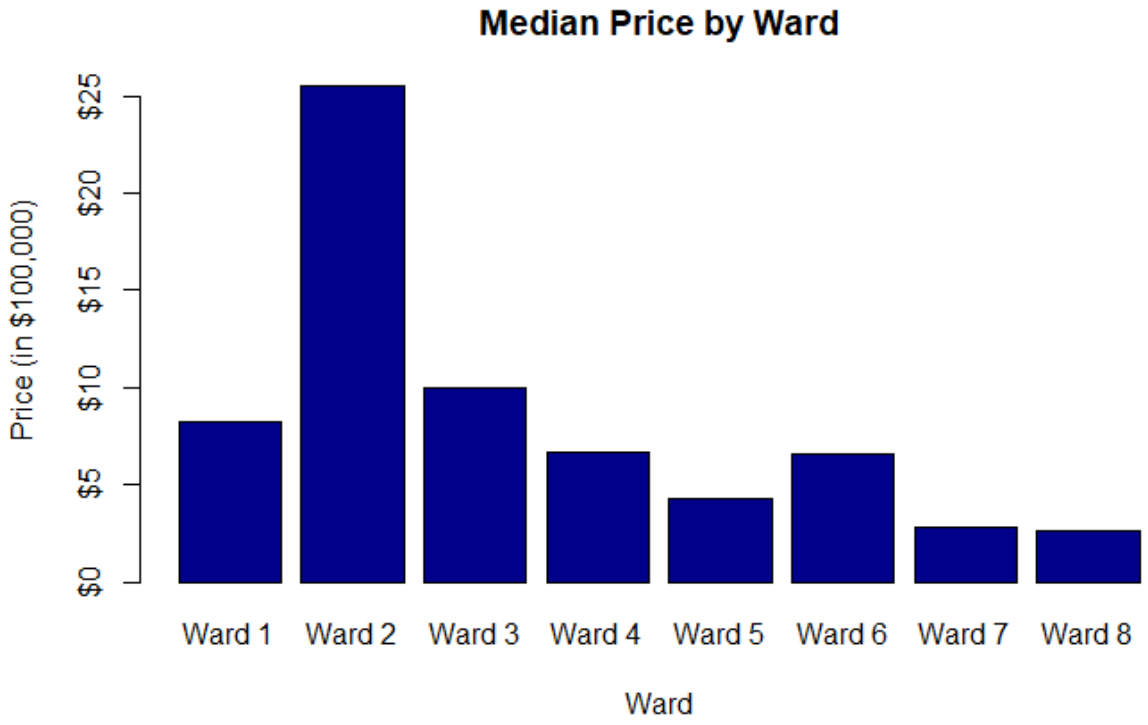
Assessing the quadrant locations of the houses, they appear to be representative of Washington D.C. and not skewed. The large majority (8245) are in the North West Quadrant, and the fewest (31) are in the South West Quadrant. This is unsurprising. The South West Quadrant is the smallest of the 4 quadrants and contains mostly of non-residential buildings such as the Jefferson Memorial, the Southern Part of the National Mall, and the Joint Base Anacostia–Bolling military base. Only the Southern most point in Bellevue contains large sections of residential land. The South East quadrant contains the second least number of houses. This is the second smallest quadrant and contains Capital Hill, Navy Yard and Hill East, non-residential districts, but contains more residential sections further South and East. The Northwest Quadrant contains by far the most homes. It is the largest by area and contains the majority of the residential sections of Washington DC.



Conducting a similar analysis on wards, the findings are continued. There are very few houses from Wards, 1,2 and 6. These Wards being those which contain and surround the centre of the city. Wards 3 and 4 are in the North West and have massive residential areas, Ward is in the North East and has mostly residential area as well, and Wards 7 and 8 cover the eastern tip and South/South East respectively.

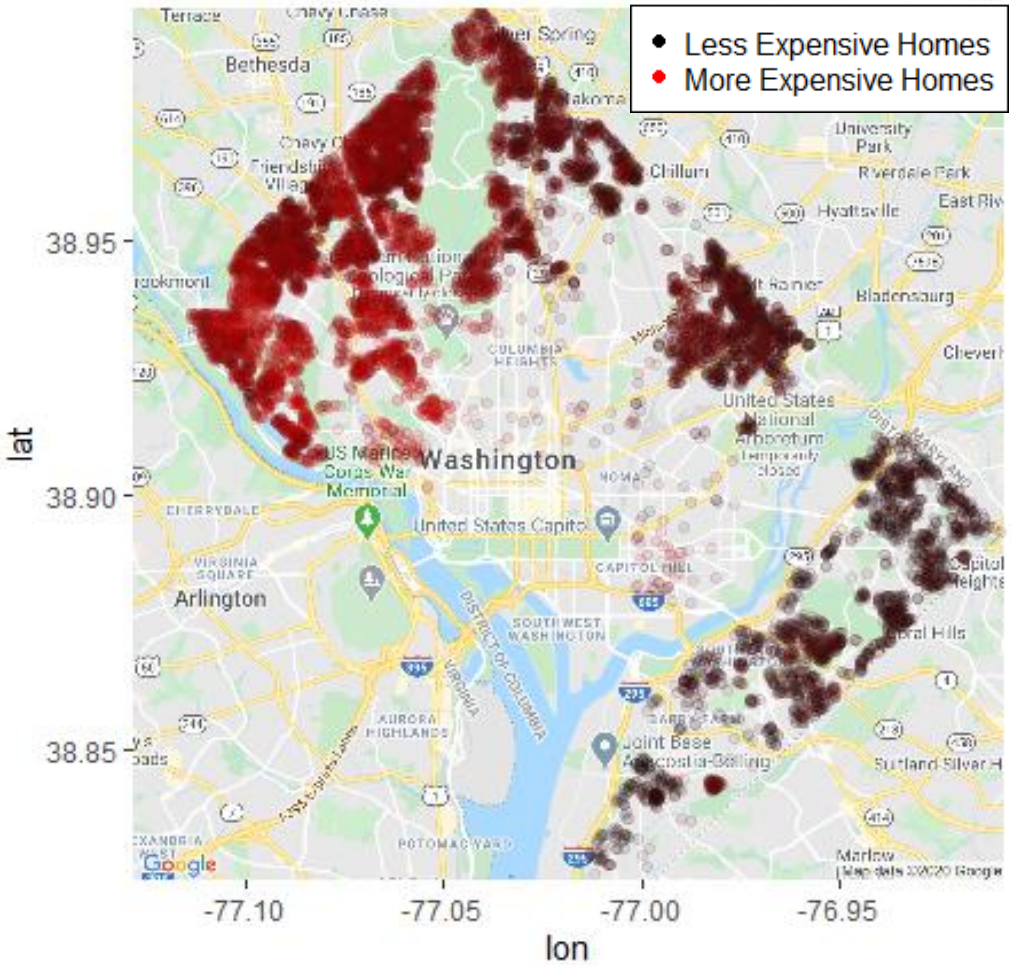


While having among the smallest supply, Ward 2 has the highest median price at approximately $2.5 Million. This is not surprising as ward 2 is located right in the heart of downtown and contains the White House, Capital Building, and National Mall. These factors would make it a desirable place to live, and the minimal area for residential homes would create an upward force on prices.



These insights can be better illustrated on a map of DC. Where the darker colours correspond to the cheaper homes, and red color corresponds to more expensive. The South and East (Wards 7 and 8) have many homes but they are cheaper. There is another cluster in the North East (Ward 5) which is mostly dark with some red. The Northwest clearly has the largest number of homes, and this area is also sum of the most red. Houses get more expensive (redder) as you get closer to downtown.

Map of D.C. Homes by Price



Size of Homes

Assessing the distribution of building size, most homes are between 1500 and 2500 square feet with a median of 1857. This is right in line with the national median of 1600. The fact that D.C. is above average is slightly surprising considering it is an urban city and would be expected to have more tightly packed smaller homes compared to the rest of the country which would include larger homes in rural areas of the country.

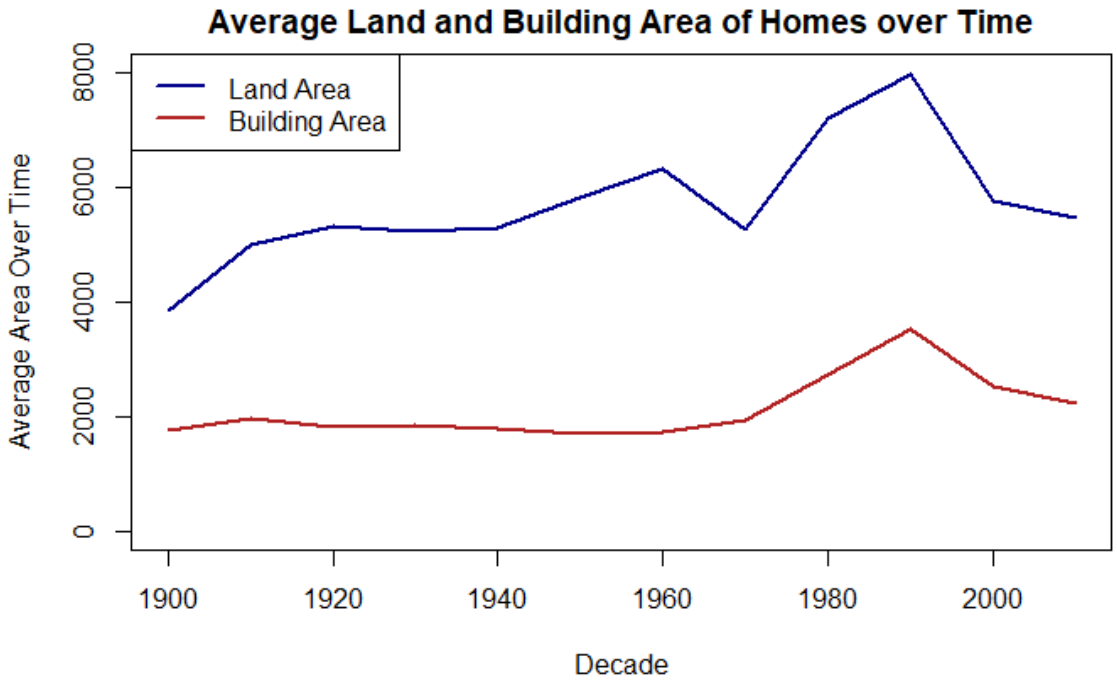
https://www.theatlantic.com/family/archive/2019/09/american-houses-big/597811/

Distribution of Gross Building Area of Washington D.C. Homes				
1% Percentile	25% Percentile	50% Percentile	75% Percentile	99% Percentile
804	1440	1857	2512	5982

Tracking how land and building area have changed over time yields interesting insights. For the analysis `AYB` (The earliest time the main portion of the building was built) is utilized to track how the shape of D.C. homes have changed over time. From the 1960s to 1990s, both land area and house area increase, land from 6000 to 8000 square feet, and building area from 2000 to 3500 square feet. Since the 1990s, both land area and home size has been decreasing albeit at differing rates.

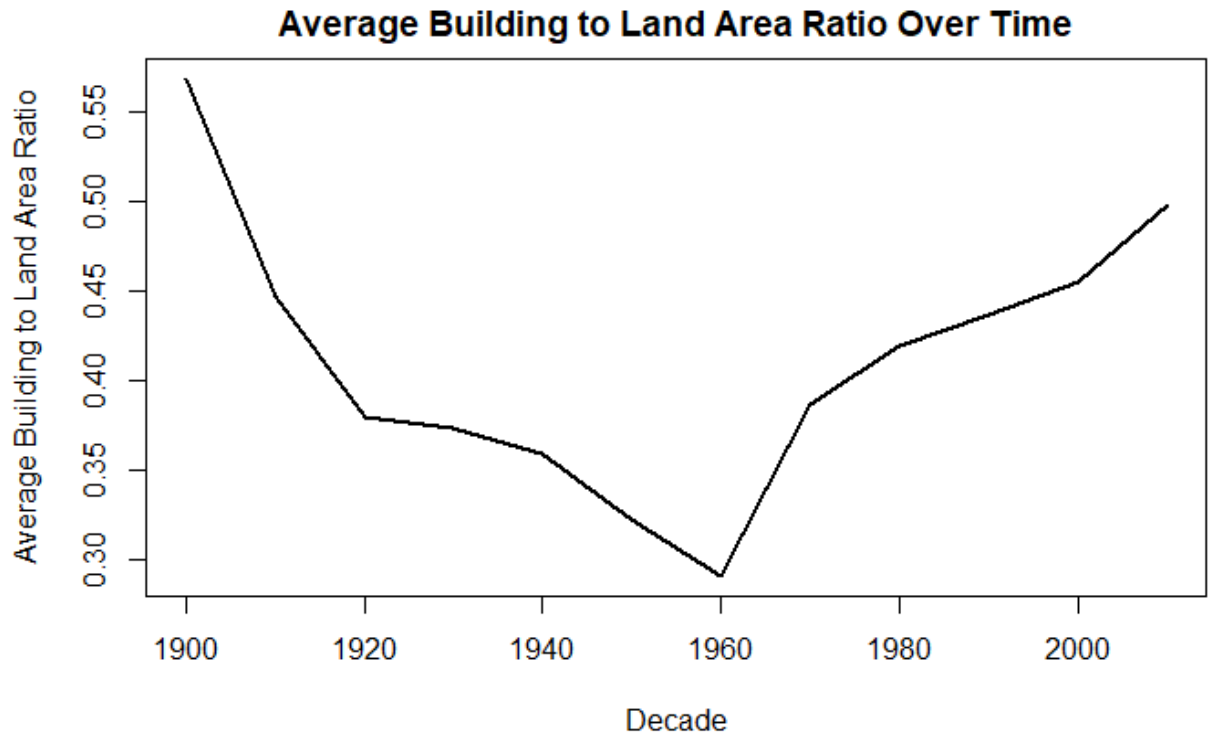
Interestingly, from the 1960s to 2000s, Washington D.C. also witnessed a decline in population, and having the population density decrease over this time would make it possible for the increase in the size of land area and building size of houses.

https://www.dcpolicycenter.org/publications/regional-population-density-since-1970/



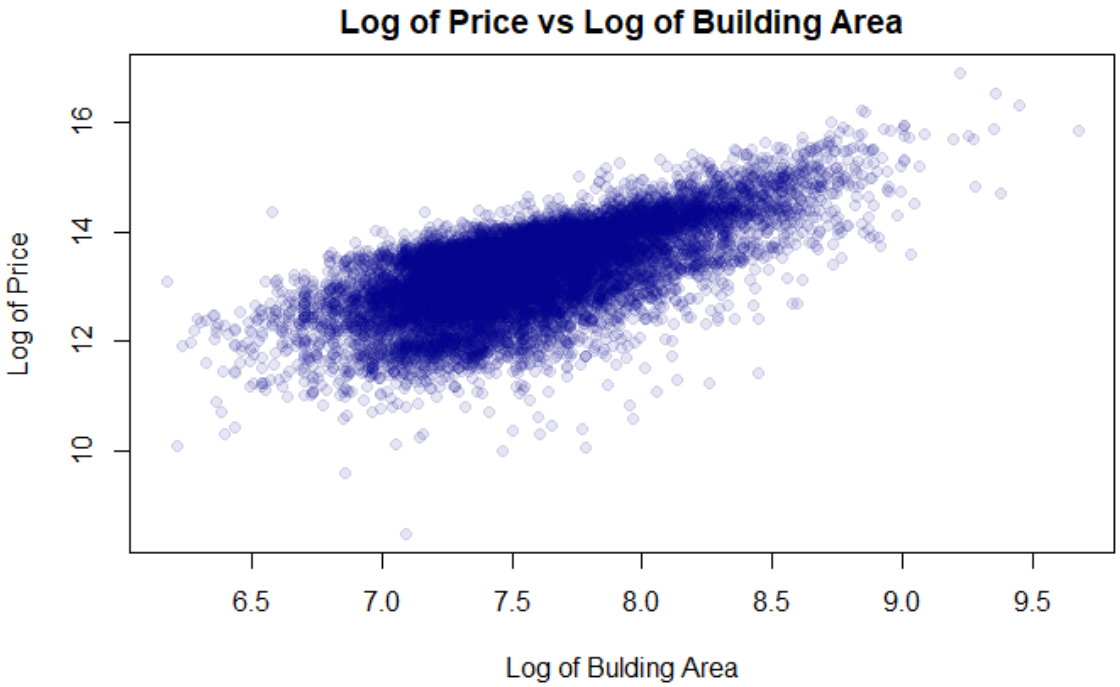
A follow up would be to see how the rate of building area and land area has changed over the decades. Since the 1960s, the home's share of area has been increasing as new builds try to optimize housing area on smaller lots. This is a trend which has been seen nationally.

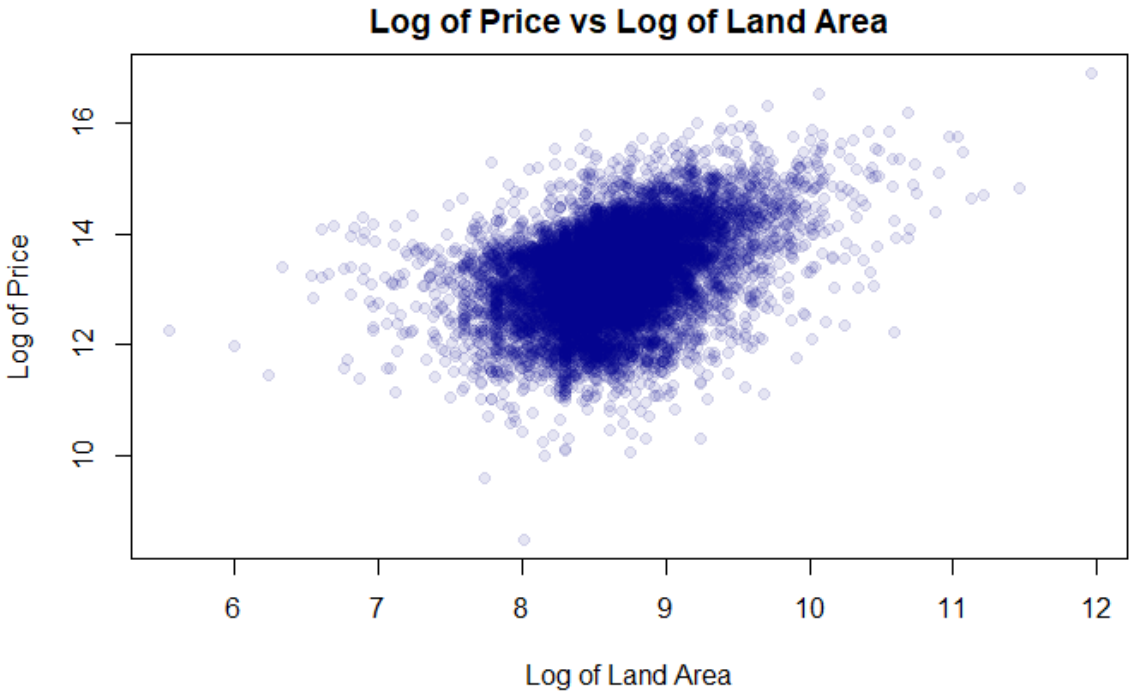
https://www.theatlantic.com/business/archive/2016/07/lawns-census-bigger-homes-smaller-lots/489590/



Impact of Home Size on Price

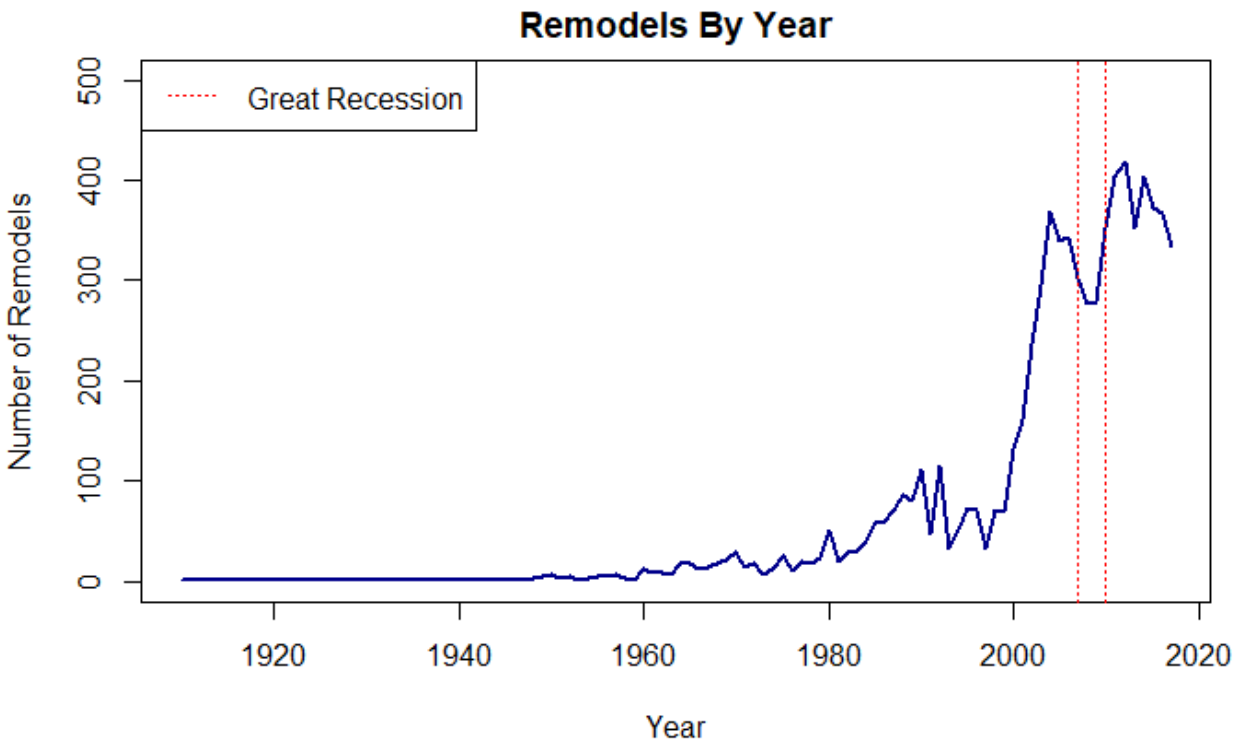
Investigating the impact of building and land area on price, there is a weak correlation between the log transform of land area and the log of price, but a strong correlation between the log of the building area and the log of the price.


Log of Price vs Log of Building Area
Log of Price
10
12
14
16
6.5
7.0
7.5
8.0
8.5
9.0
9.5
Log of Bulding Area

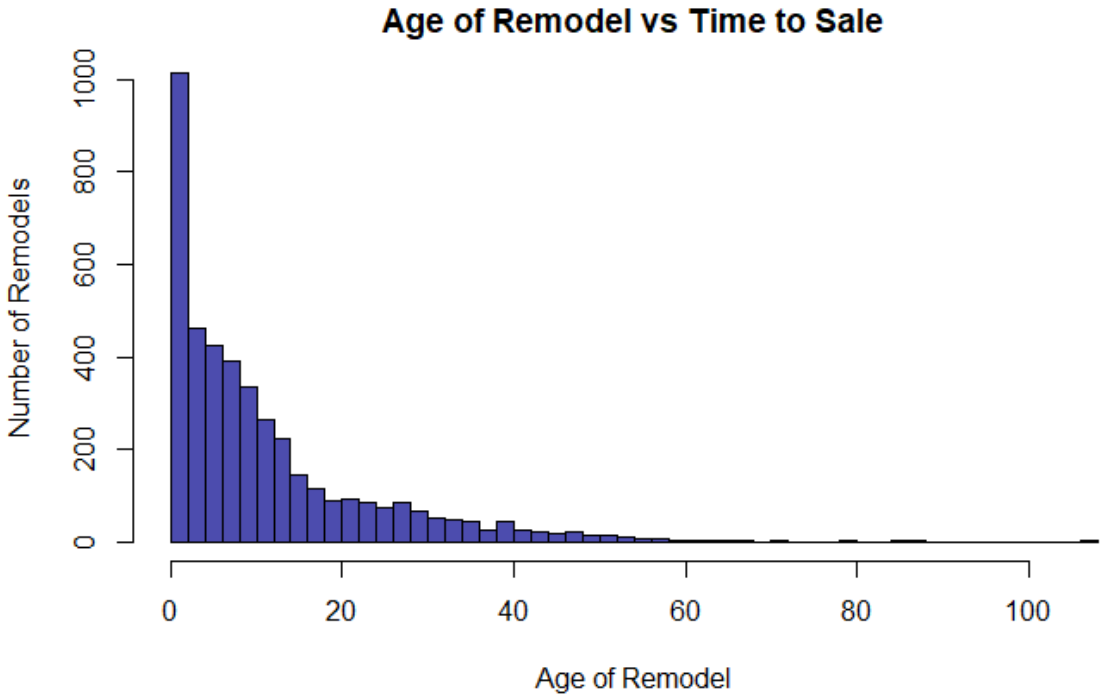

Log of Price vs Log of Land Area
Log of Price
10
12
14
16
6
7
8
9
10
11
12
Log of Land Area

Analysis of House Remodels

Assessing the number of remodelling's by year, there are 3 distinct trends. Firstly, a sharp and sudden increase in the number of remodels beginning in 2000 and continuing until 2007. Between 2007 and 2009 the number of remodels declines almost 25% , but begins to increase again after this. This is of interest because while the Great recession did not seem to impact the price of homes greatly in Washington D.C. it appears to have hindered appetite for a major, expensive remodel.

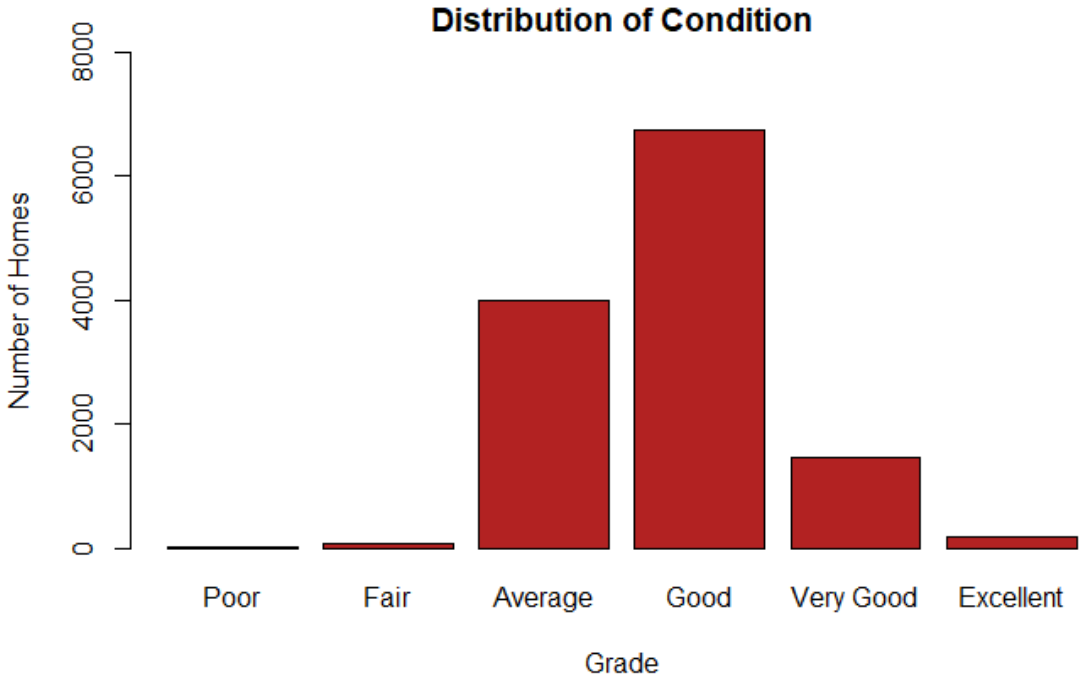


Finally, remodelled homes are typically sold very soon after the remodel, 0-2 years is the most common time between a remodel and a sale of home.

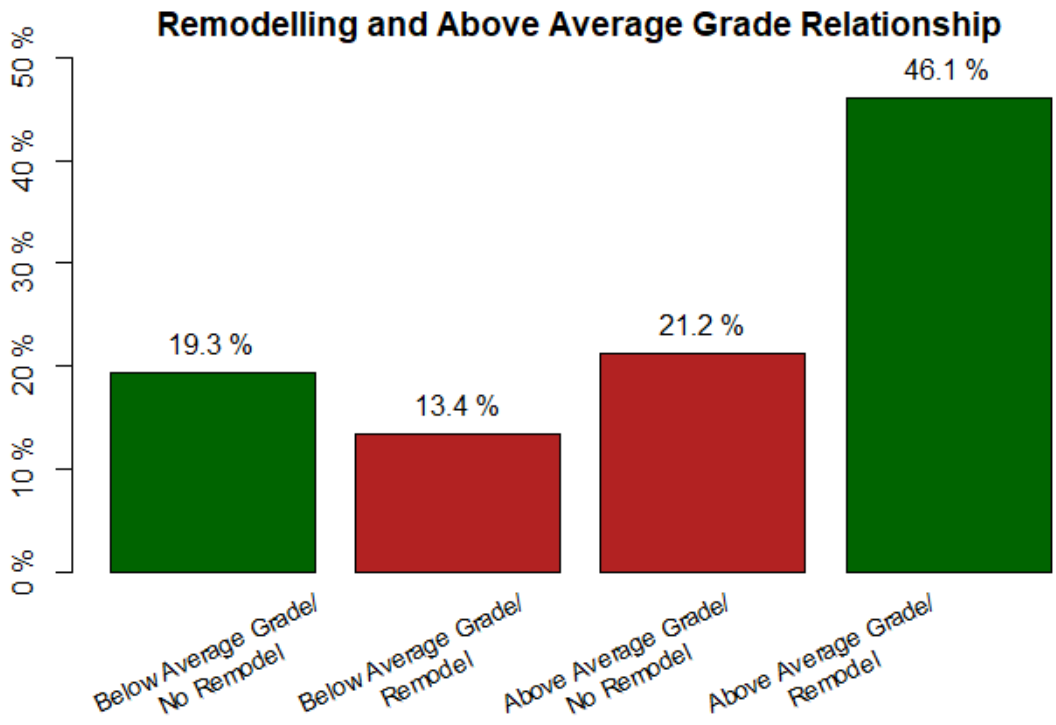


Analysis of Condition

Assessing condition grades given out, most homes are given a grade of Average or Good. Very few are given Excellent, Poor or Fair grades. A follow up would be to assess how remodels impact the grade a home is give.



Plotting the interaction between remodelling and receiving an above average condition shows a strong relation. One predicts the other at a rate of approximately 66%.



Pre-Processing

Missing Data

Column	# Missing Values		Column	# Missing Values
LANDAREA	0		HEAT	0
ZIPCODE	0		AC	0
LATITUDE	0		ROOMS	0
LONGITUDE	0		BEDRM	0
ASSESSMENT_NBHD	0		AYB	28
ASSESSMENT_SUBNBHD	2994		YR_RMDL	5049
WARD	0		EYB	0
GRADE	0		STORIES	7
CNDTN	0		SALEDATE	0
EXTWALL	0		PRICE	0
ROOF	0		GBA	0
INTWALL	0		STYLE	0
KITCHENS	1		YEAR	0
FIREPLACES	0		YEAR1	0
BATHRM	0		AYB2	28
HF_BATHRM	0		EYB2	0
QUADRANT	60		YR_RMDL	5049
Id	10002		EYB	0
Usage	0		STORIES	7
Age	0		SALEDATE	0
Col	1		PRICE	0
cnd_clust	0		GBA	0
rmd	0		STYLE	0

After exploring the data, some steps should be taken prior to modelling with the data. Most of the columns do not have any missing data, however those which do should be treated. Additionally, some light feature engineering is suggested based on the type of data in each column. The following sections describe how missing data is treated for different levels, and what transformations we take of explanatory variables.

Treatment of Missing Data

- QUADRANT: replace with new class names "Unknown".
- KITCHENS: Replace with the mean number of KITCHENS in the non-missing data
- AYB: Replace with the mean value of AYB in the non-missing data
- STORIES: Replace with the mean value of STORIES in the non-missing data
- ASSESSMENT_SUBNBHD: Replace all classes not found in training set with "Missing", this will cover both Missing data and levels not found in the training data.

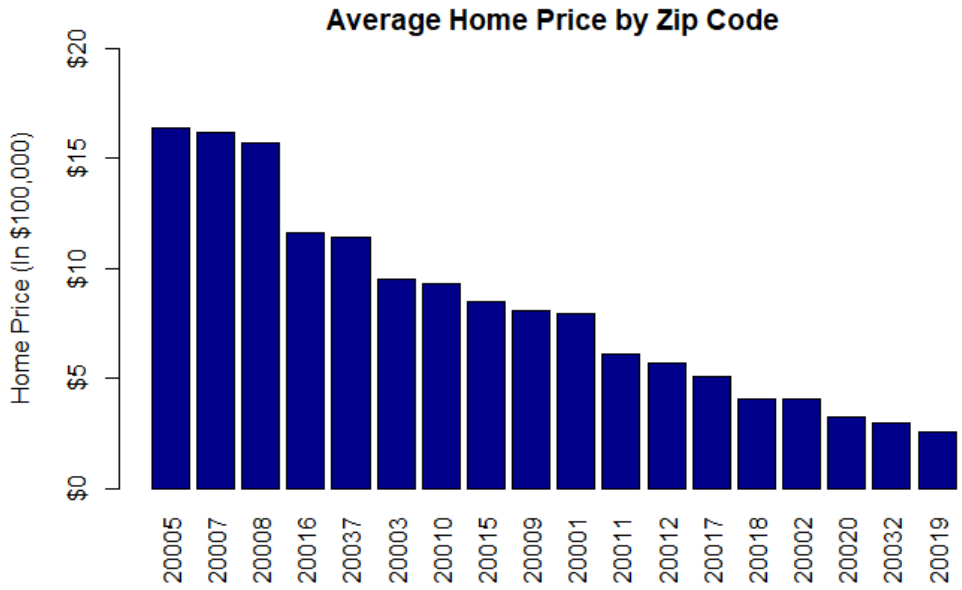
Transformations

- PRICE: A log transformation of the response variate PRICE
- GBA: Log transformation
- LANDAREA: Log transformation
- SALEDATE: Extract year and month
- age_of_ren: Which is YR_RMDL-year or EYB-year if YR_RMDL is not available
- FIRE: Convert to 3 levels, whether the value is 0,1,2 or >2
- Lat1: Indicator variable whether Lat is greater 38.905 (This roughly partitions the data in to the NW/NE and SW/SE Quadrants, see map of D.C. previously)
- HOTMONTH: Grouped Variable whether the month was April - August, December or other
- GRADECUT: Group GRADE levels with similar prices into 4 groups
- roof1: Group roof levels with similar prices into 4 groups
- AGE: YEAR – AYB
- AGE_RENO: YEAR-EYB

Motivation for Transformations

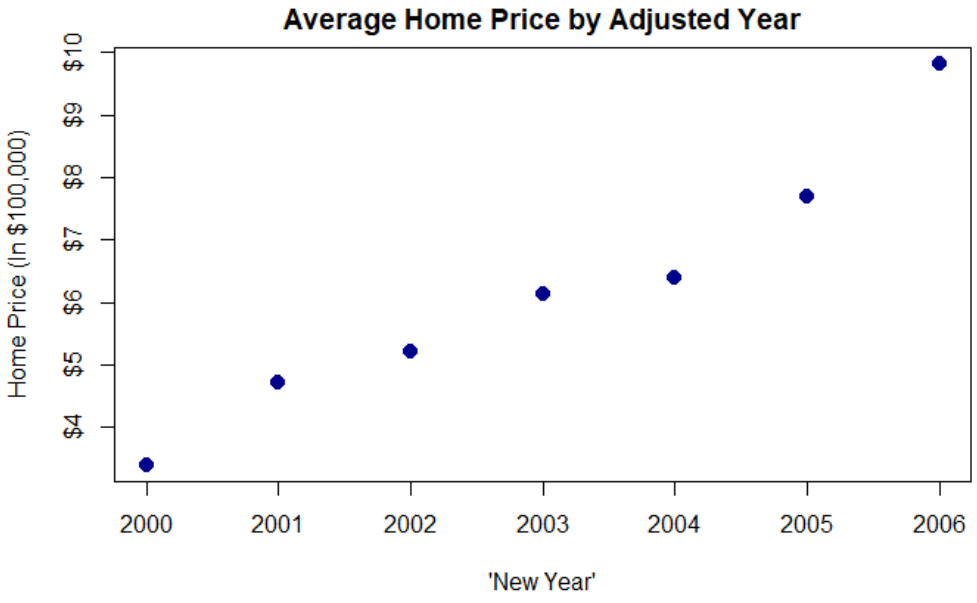
Zip Code

Assessing the mean house price by Zip Code, a new variable GOODZIP is created as an indicator whether the home is in the top 5 zip codes by average home price.



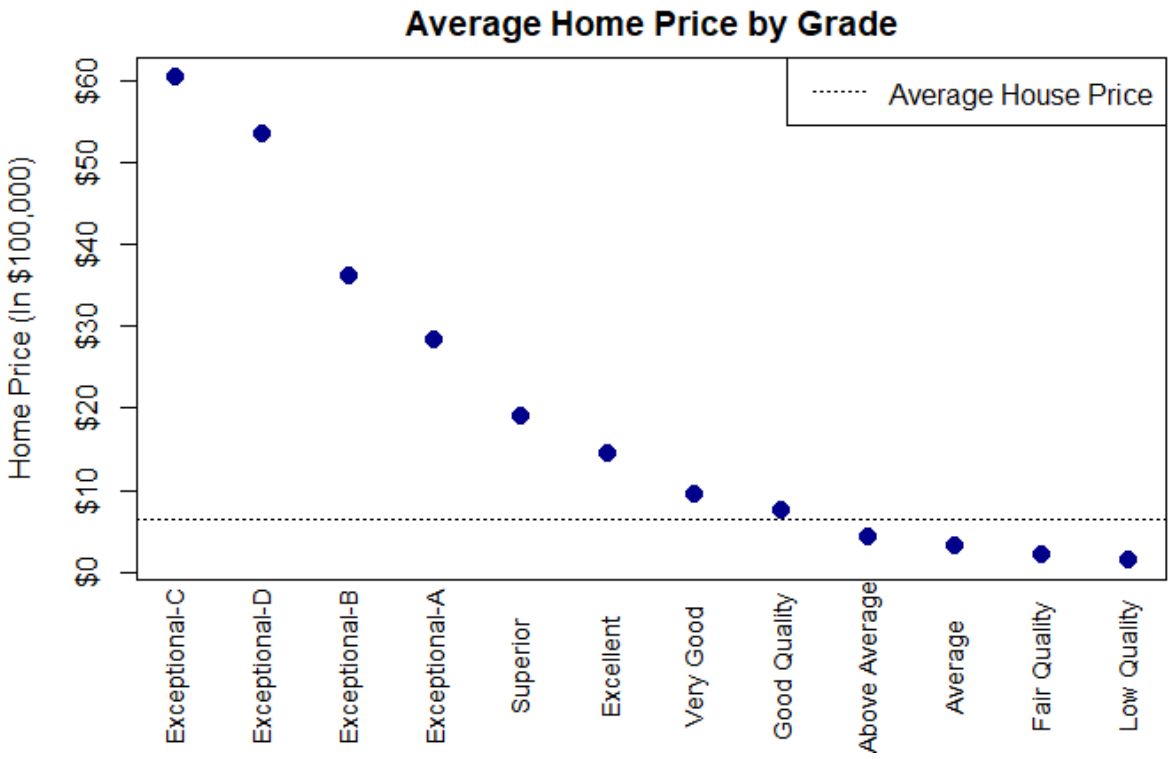
Year

Recalling the plot of home prices over time, prices were flat after 2006 and before 2001. To create a more linear relationship, the variable YEAR1 is created. This is a variable which is the same value (2000) for all years before 2001. The value is the year of sale if the home was sold between 2001 and 2005 and 2006 if the home was sold after 2005.



Grade

Grouping on Grade given to a home, we few consolidations are made. A highest group with Exceptional (A/B/C/D) group 3, a second with Superior and Excellent group 1, and a final group with Low Quality, Fair Quality, Average, and Above Average group 2. The middle group is kept as a control. These values will be stored as factor in the variable GRADECUT.



Safety Checks on Data Processing

Prior to any model building, some safety checks are added into the data cleaning and transformation procedure. For both numeric and factor variables it is possible that missing data that is not present now may appear in the future. There are also many factor/character variables in our data. It is possible that on current or future test data, new factors may appear, or this field may be no longer reported.

For all character values, there typically exists a "Default" value. For example, the feature 'Heat' has a level "No Data". For all character values, if a factor exists for this variable that is not in the list of "Non-Default" options, we adjust it to be this level, since this level has not been seen in training prior. Note, this is also an easy way to adjust for NA values which may appear as well. After making this adjustment to any training and test data. One-hot encoding is applied these values to create useful easy features for machine learning models to assess.

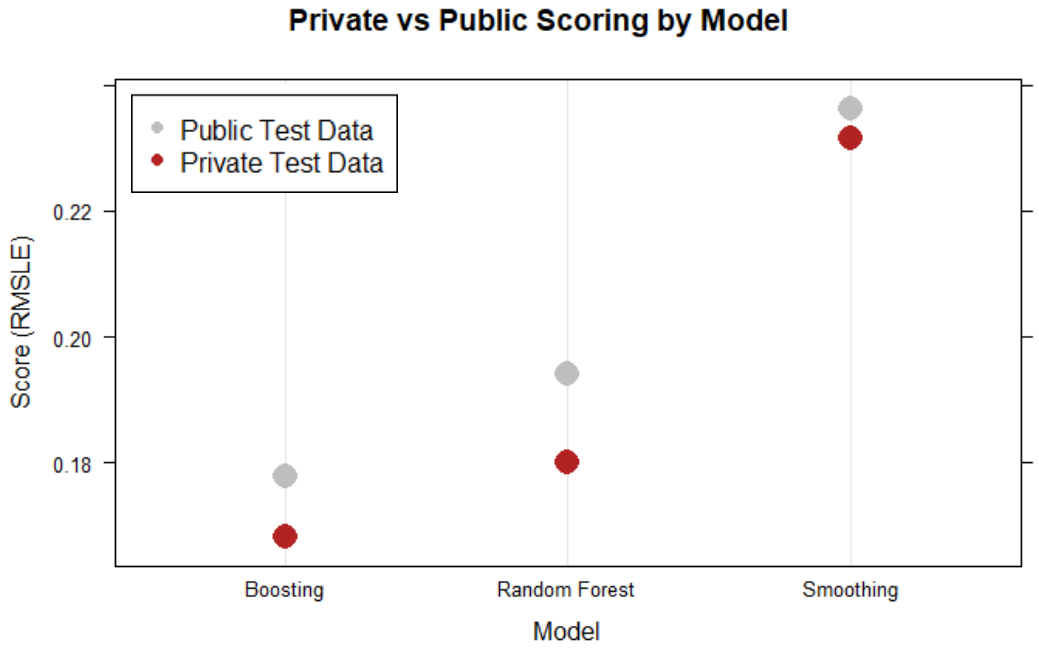
Variables given this procedure include:

- Heat
- Grade
- Condition
- Exterior Wall
- Roof
- Interior Wall
- ASSESSMENT Neighbourhood
- Quadrant
- Zipcode (While this is a numeric variable but should be treated as a factor)

For numeric data, the check of removing all missing values and mean imputing them, means a full variable by variable manual data cleaning will not be required for future data.

Statistical Analysis

Thee different Statistical learning models were applied to this dataset, both to understand which model has the best predictive power as well as to interpret which features have the strongest predictive power on the price of a D.C. home. The models chosen were Smoothed regression, Random Forest, and Gradient Boosted Trees. Our target metric for this analysis is Root Mean Squared Log Error, meaning running models on the log transform of the Price is the best way to get strong results.

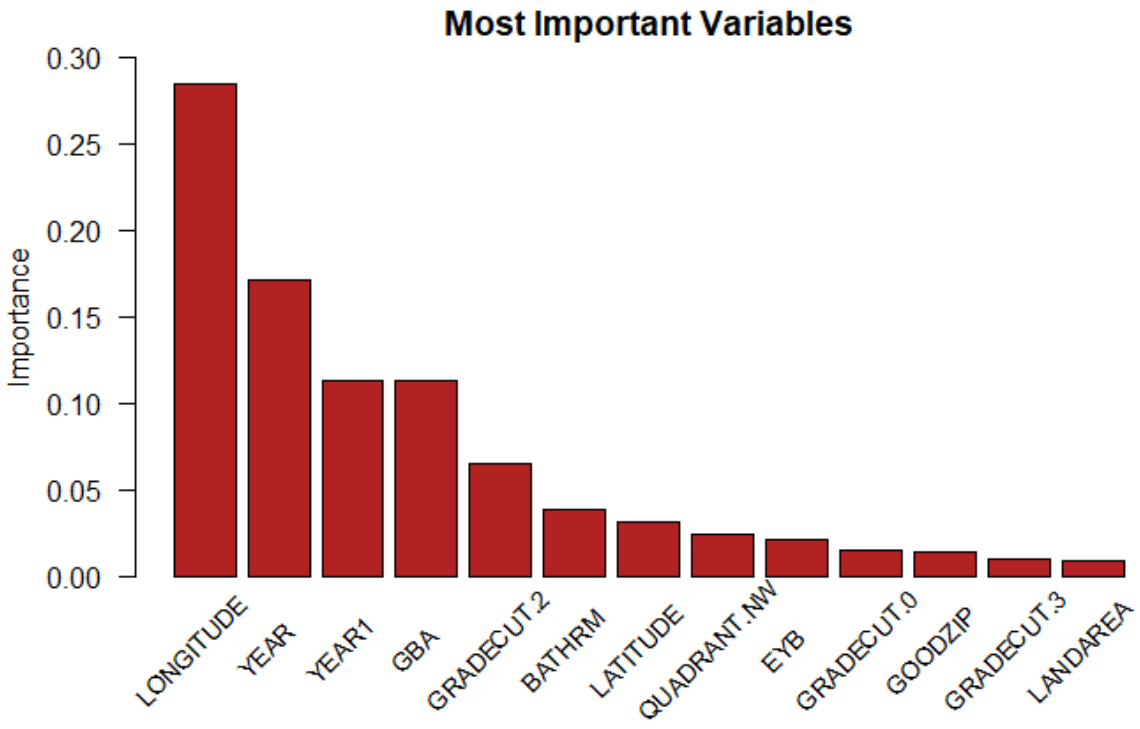


There are multiple interesting insights from this chart, which motivate further analysis. Private scores are higher than the public score for all 3 models. Boosting yields the best results and smoothing yields the highest error.

Variable Importance

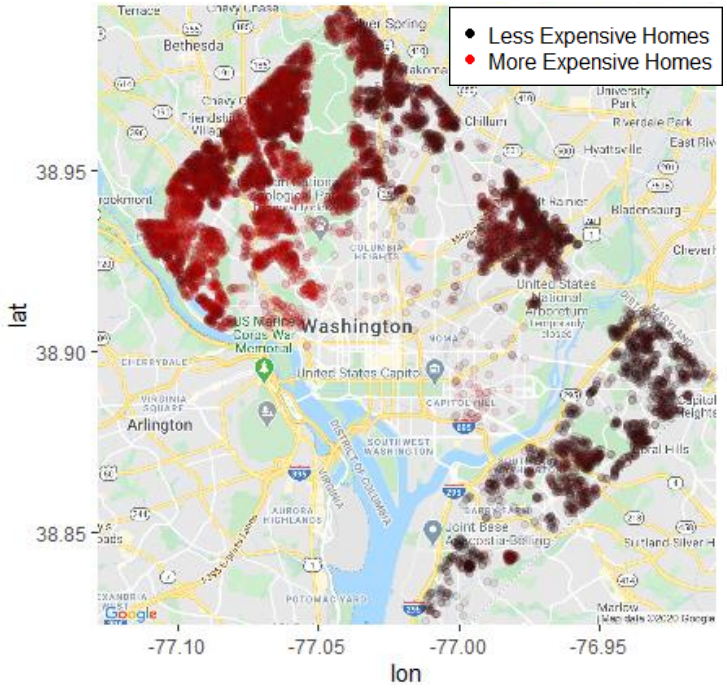
To understand just what exactly these models were learning we determine the importance of each variable to the model. First, the importance for each variable are obtained then re-weighted such that the sum of the importance of all the features for each model is 1. For each variable we take a weighted sum of the importance across each other models. In this sum the test error is used.

Combining the variable importance from the random forest and boosted trees models and weighting their importance, the top 13 variables and show a clear and distinct patter.



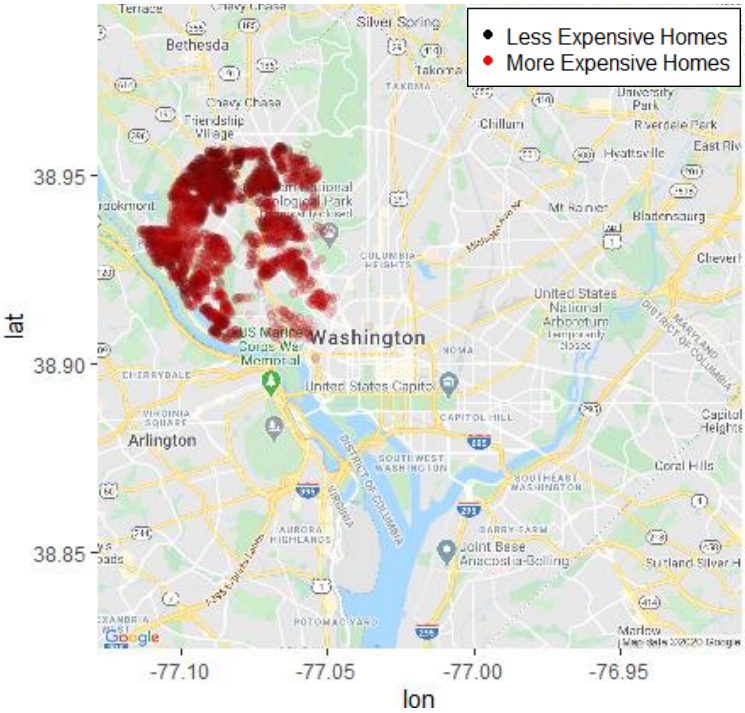
The factors which best predict the price of homes is Location, Time and Size. Recalling our map of the District with price overlay, there was an obvious pattern. The Northwest showed much more expensive homes then in the South and East. The model interpreted this through using Latitude and Longitude to learn this. Unsurprisingly, longitude was more important as an inspection shows a vertical line (at around long = -77.015) best partitions the red and black dots better than a single horizontal line.

Map of D.C. Homes by Price

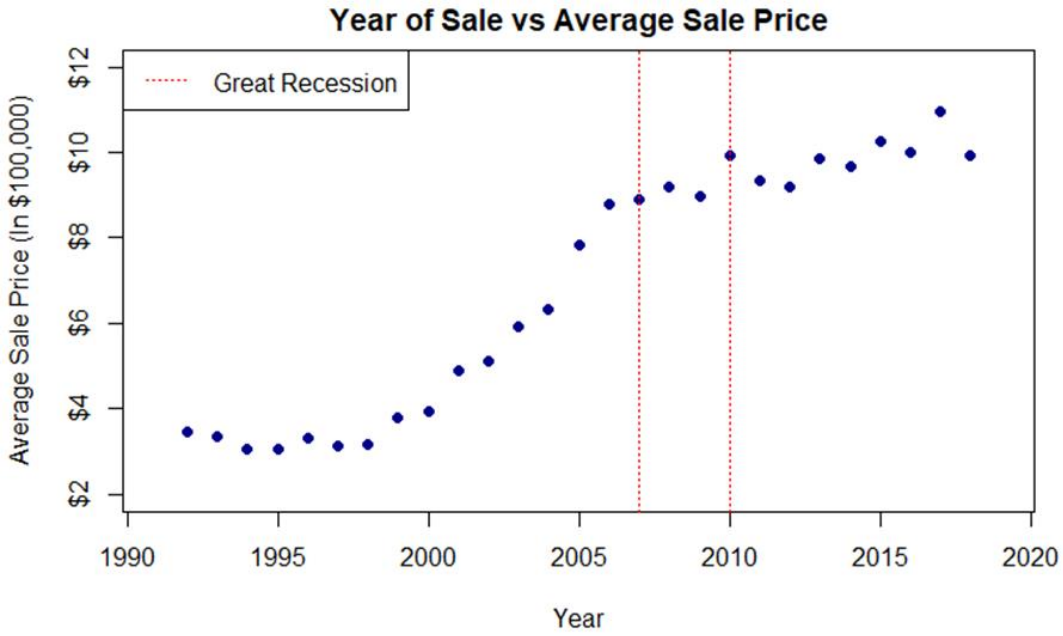


The variable good zip was an indicator variable on whether the home was in zip code 20007,20008,20016,20037 or 20005. Assessing the average price in these zip codes they were strictly higher than the rest. Plotting only these homes, using the color coding these zip codes clearly have some of the reddest dots and most expensive homes.

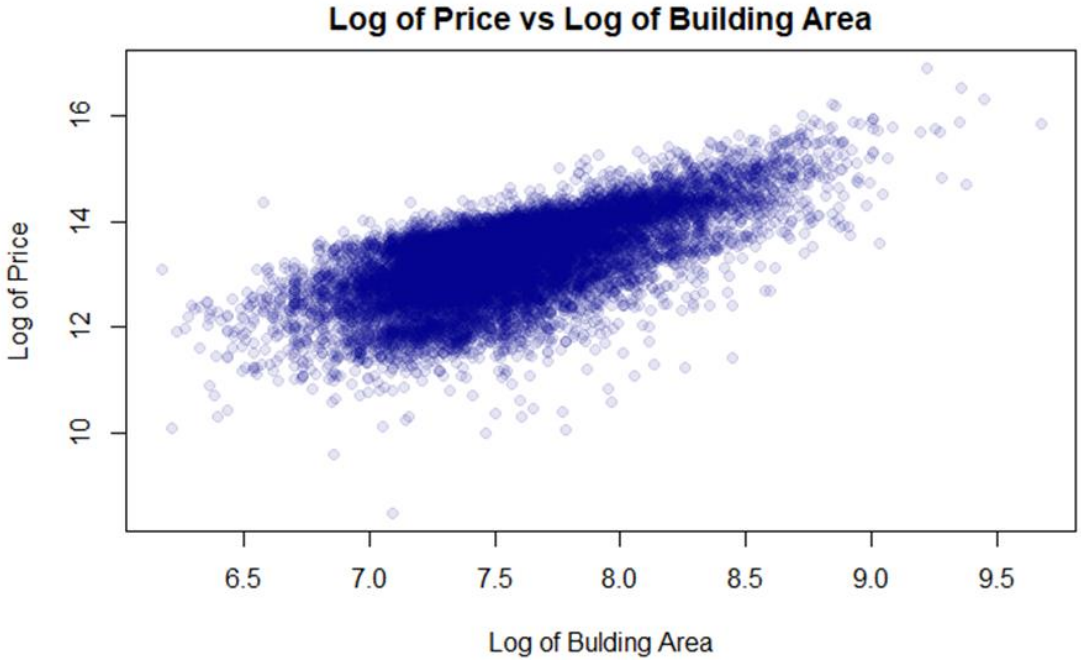
Map of D.C. Homes in Expensive Zip Codes



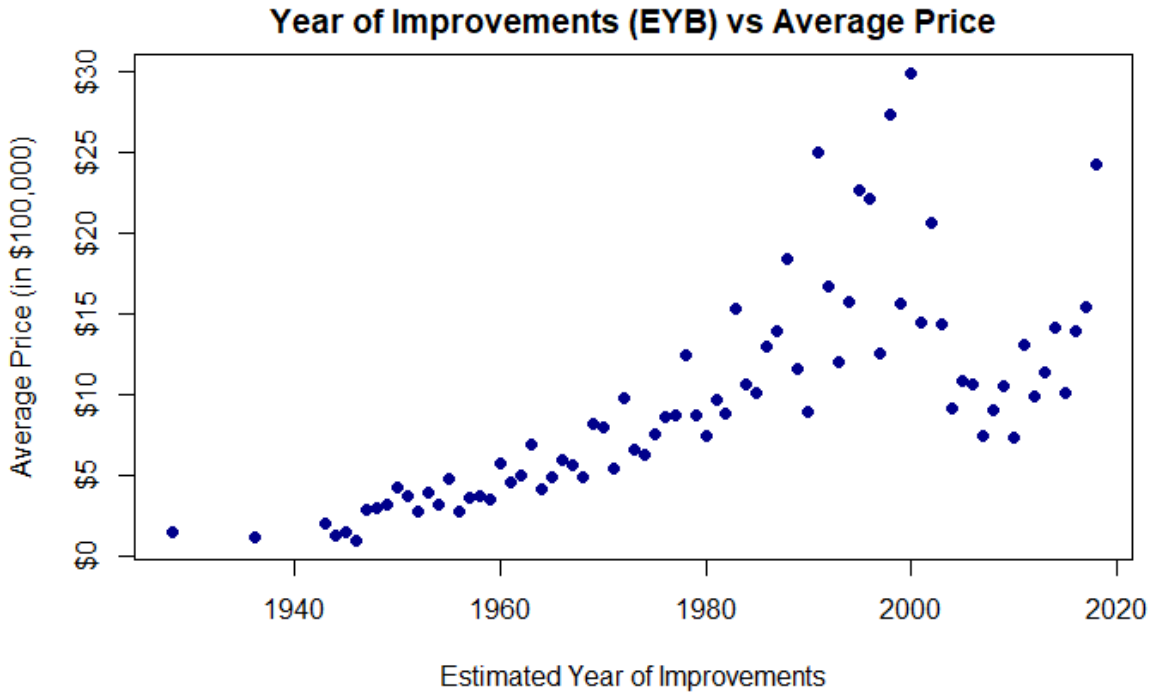
The second most important factor in impacting home prices is timing. As seen in the exploratory analysis, there has been a strong increase in the price of DC homes, especially since 2006. The further in the past the home was sold homes were greatly cheaper and the magnitude of the difference makes it unsurprising this is a critical factor.



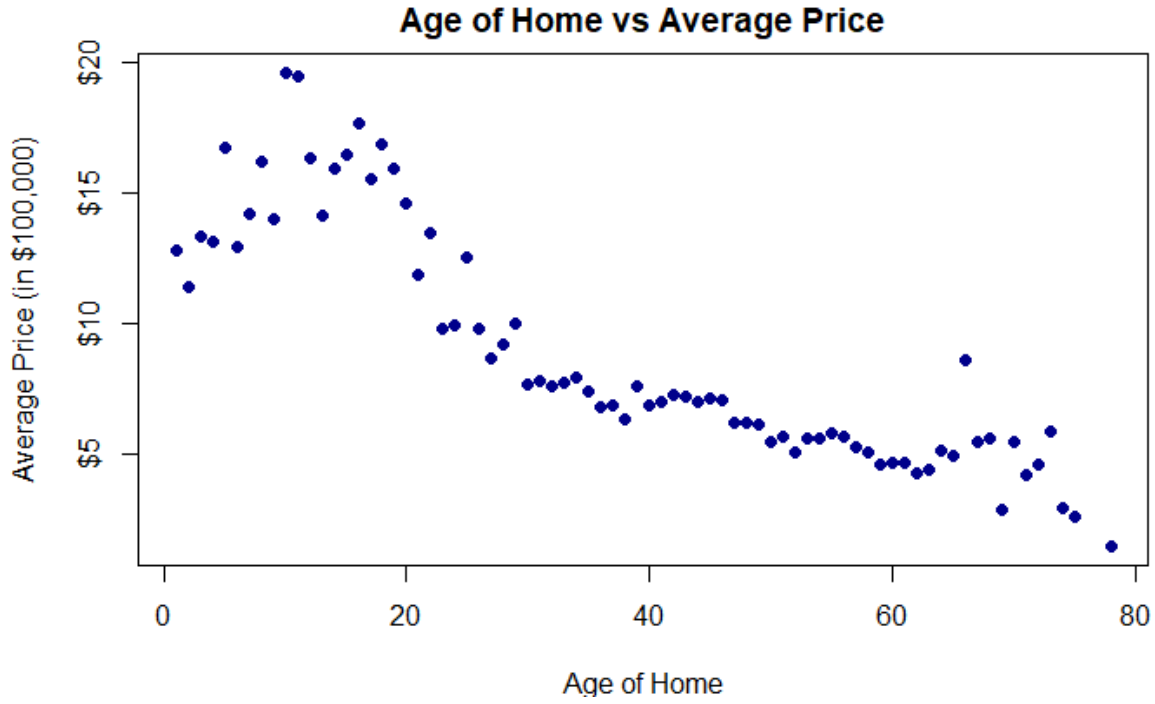
The final most important factor is land size. As seen in our exploratory analysis, Gross Building Area and Land area has strong and weak correlations when log transformations were taken and seeing them is not unsurprisingly.



Year of improvement is one of the lesser important variables. Plotting average sale price by year of improvement, it appears improvements made more recently have a stronger impact on price. Homes prices have increased greatly over time. This trend follows the overall price increase over time suggesting that newer homes sell for more money.



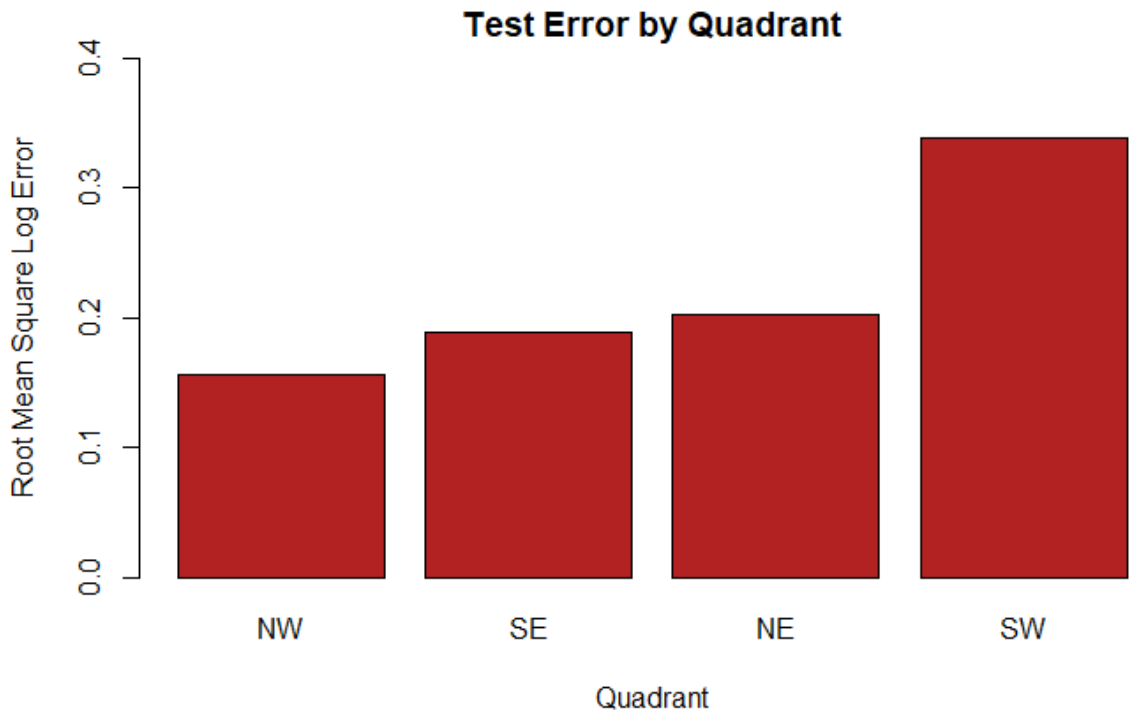
A plot of Age of the home (Year-EYB) confirms this. This suggests that newer built homes sell for more money.



Other Insights

Impact of Imbalanced Data

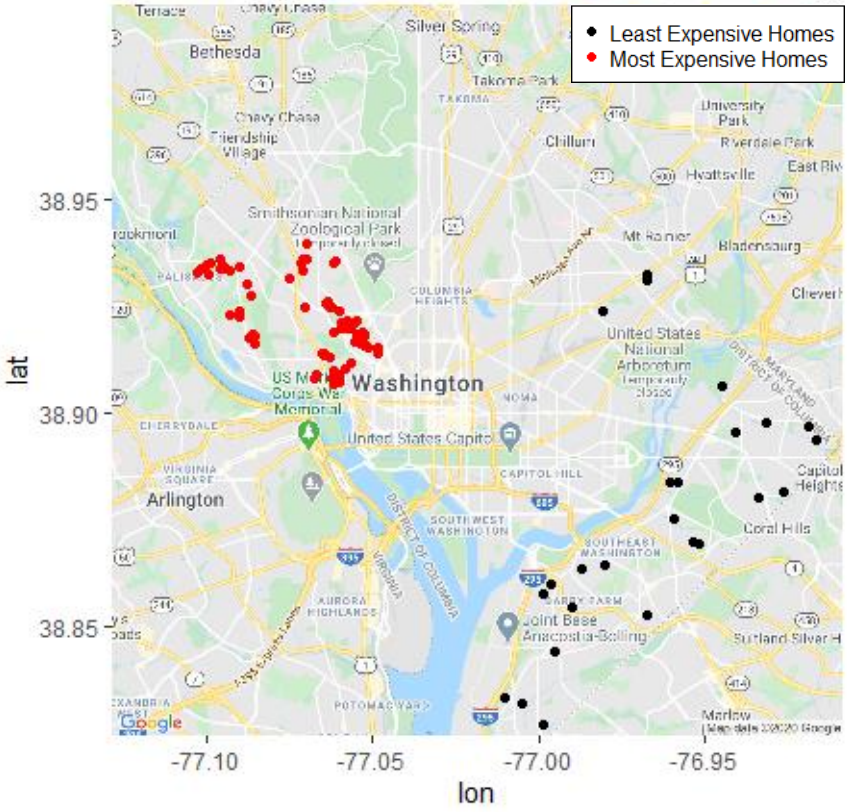
During exploratory work, it was discovered that most of the homes in the dataset were from the Northwest Quadrant. Plotting test error against quadrant, the Northwest performed the best and the Southwest the worst, and the other two Quadrants were roughly equal. This suggests we do not have enough data to accurately predict home prices in the Southwest Quadrant and more data from the Northeast and Southeast would improve model accuracy. However, as discussed earlier this is difficult to acquire due to the distribution of residential land in Washington D.C. This is a limitation to the model as it is bias to the Northwest Quadrant.



Most Valuable Homes

Bellow, a plot assesses where D.C.'s most expensive and least homes are. The most expensive homes (>$4,500,000) are all grouped in the Georgetown area just North West of the White house. The least expensive homes (<$50,000) are all on the outskirts of the city on the south and easternmost points of the city.

Map of Most and Least Expensive Homes in DC



Conclusion

From the analysis, it is clear the factors that one would expect to impact the price of a home are among those which impact the prices of homes in Washington D.C. the most, those being location, size and when the sale took place. The analysis shows that land and building size is important as well as a home being in desirable neighbourhoods with proximity to the city center. The importance of time of sale also shows that there have been tremendous increases in home values in Washington D.C. particularly in between 2006 to 2018.