STAT 444/844 Winter 2024 Final Project Report

Bruce Liu

UW ID: 20843361

Contents

1 Introduction

The objective of this project is to predict the prices houses in a certain geographical area. Throughout this report, we will explore the dataset in further detail. We will compare regression models to see which regression model performs the best. We will discuss how to interpret the models that we derived and the conclusions that we can reach by interpreting said models. This report is split into three sections. The first is some exploratory data analysis on the dataset itself. The second is the statistical analysis of the data through the models we have selected. The final section is the conclusion to the report.

2 Data description

To begin, the data set provided consists of many variables which can be roughly grouped into these categories:

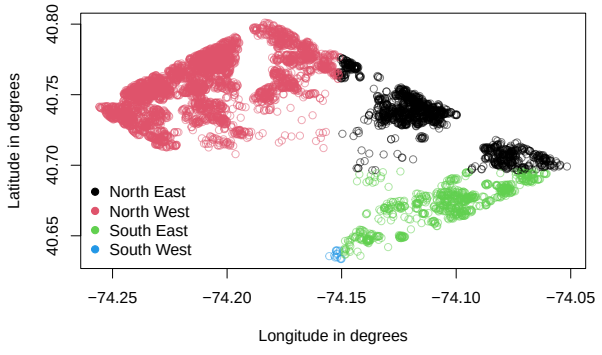
- Geography: latitude, longitude, neighbourhood, quadrant, ward, land area, gross building area
- Time: the earliest time the main portion of the building was built, the year the structure was remodelled, the year an improvement was built, date of sale
- Quality: condition, grade
- Features: type of exterior wall, type of interior wall, type of roof, type of heating, style of the home, stories, whether the house has air conditioning
- Rooms: number of bathrooms, half bathrooms, kitchens, fireplaces, bedrooms

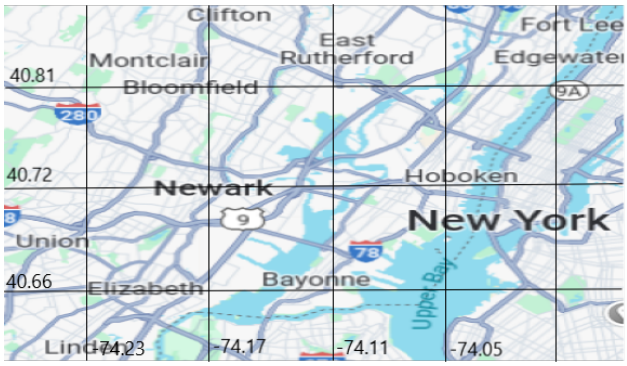
The response variate, as detailed from the introduction, is the price of the house. Let us take a look at the five categories that we have split the variables into.

2.1 Descriptive Analysis

2.1.1 Geography

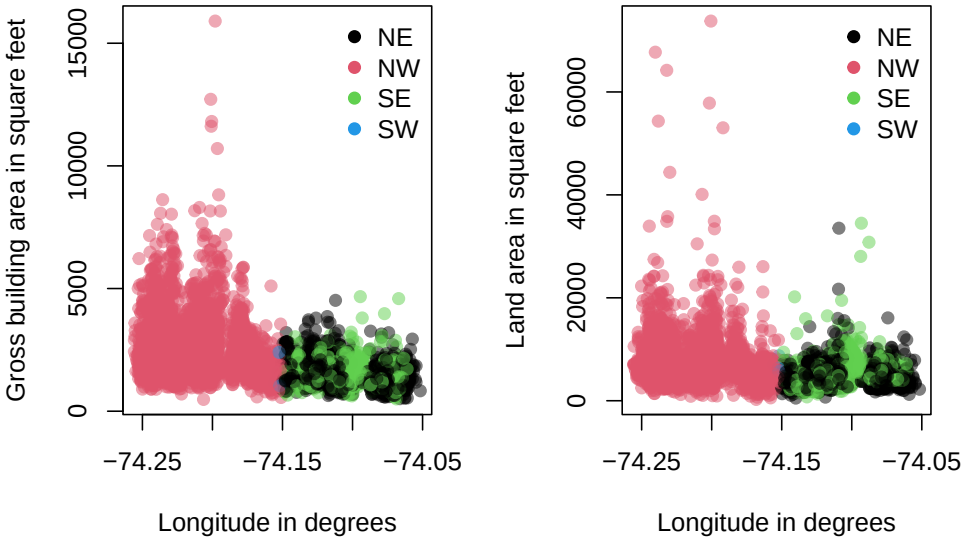
With this dataset, it may be helpful to get an idea of where the data is coming from. Let us begin with a simple plot showing the locations of the houses. We will colour the plots based on which quadrant the house belongs to.





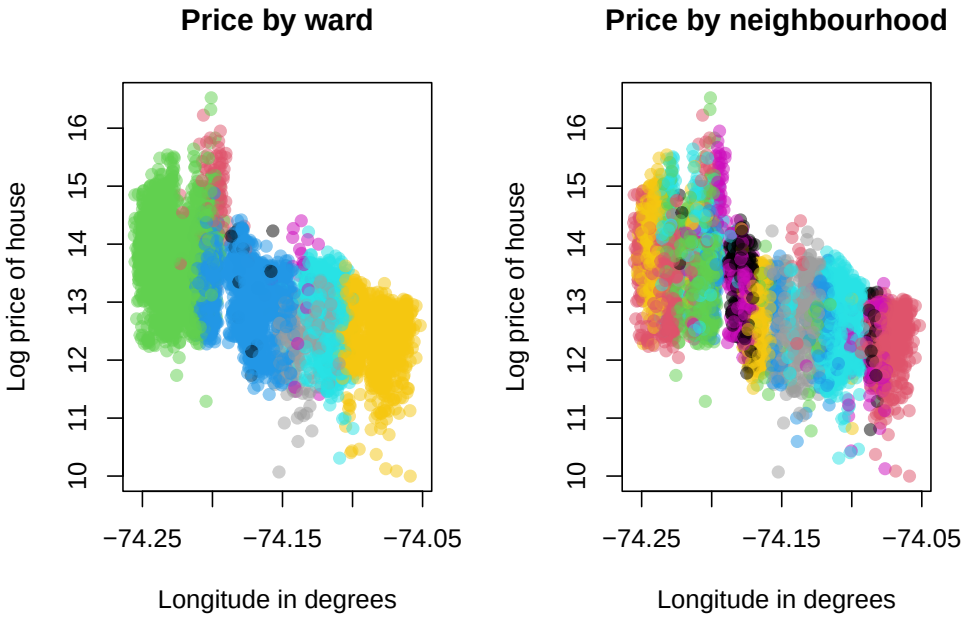
We can see that the locations of the homes are clustered in four distinct areas. With the coordinates of longitude and latitude for the dataset known, we realize that the houses are within the area surrounding New York City.

Having covered the location of the data, we can start to see if there are any interesting patterns that emerge from this data. Let us take a look at plots of gross building area of a house and the land area of a house compared to where it is situated.



From these plots, we can see that the houses situated in the northwest quadrant of the map are the houses that have the largest areas. Similarly, we can see that the houses with the smallest area is the houses in the southwest quadrant of the map.

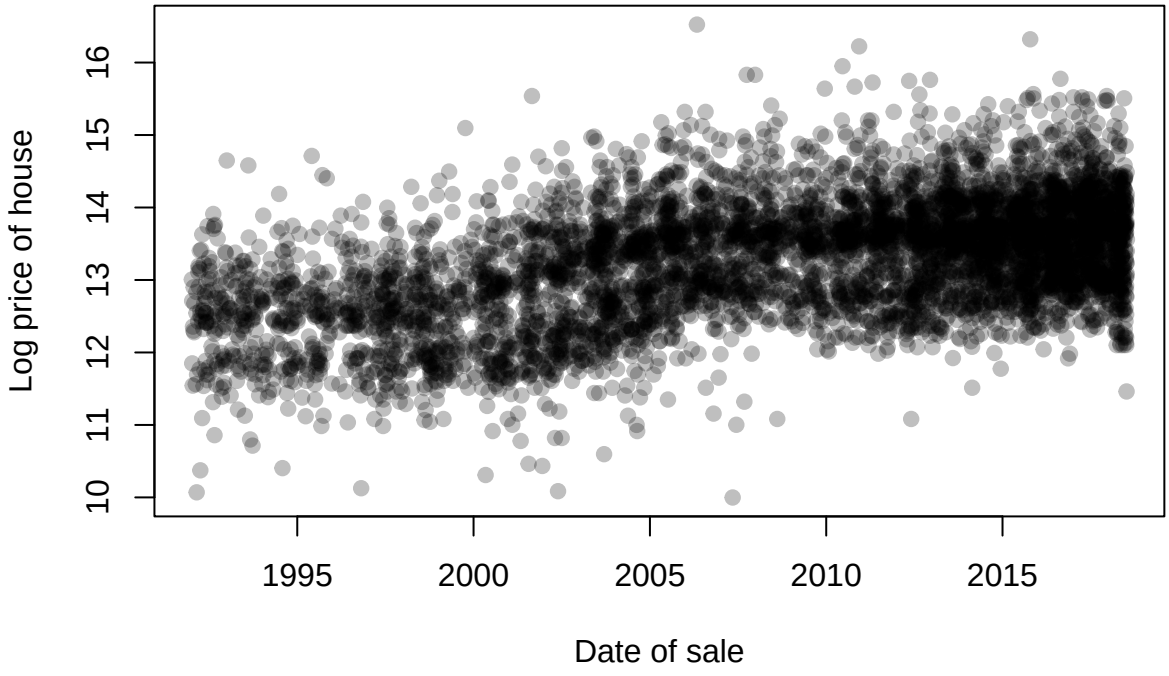
Perhaps we should take a look at how the wards and neighborhoods may influence the prices of the buildings.



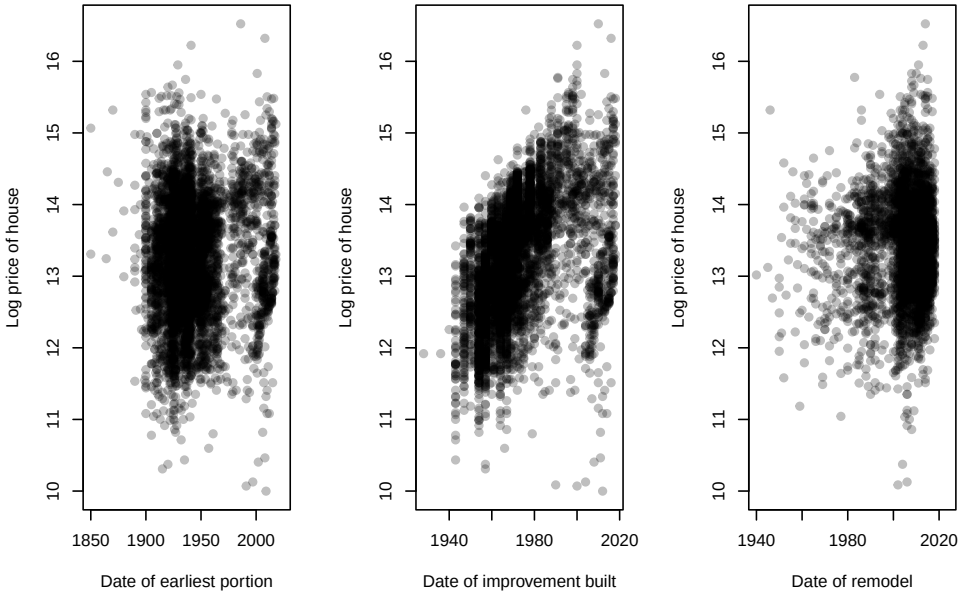
We see that the neighbourhood and the ward that the house is situated in has an influence into the price of the house. Comparing the prices to the map of the houses, we see that houses that are situated in the western half of the map tend to be more expensive than the houses situated in the eastern half of the map.

2.1.2 Time

Here, we look to see if there is any trend of house prices as time goes on. We might expect that as the sale date moves forward from the 1990s, the prices of houses increases. Let us see if this intuition is reasonable.



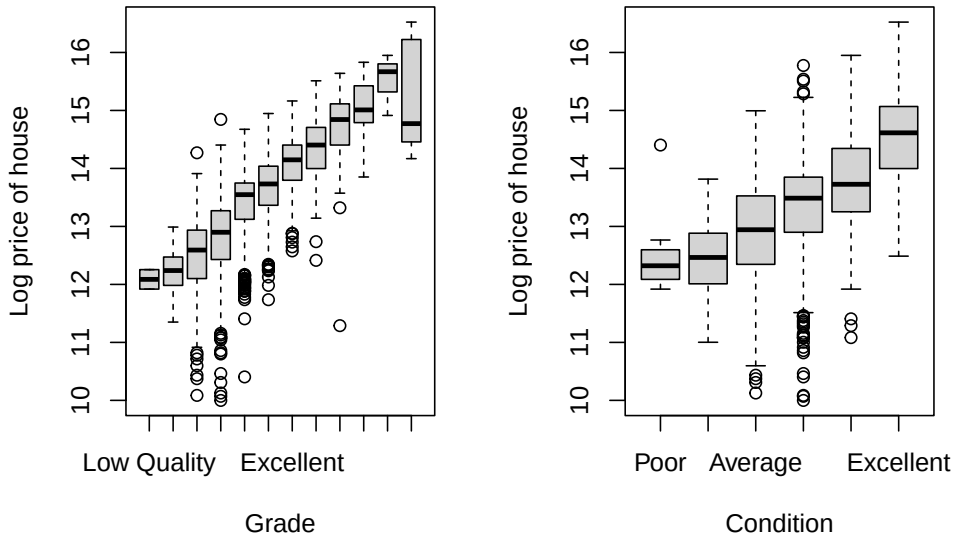
When interpreting this graph, we must be cognizant that we are dealing with a log scale in price. This means that an increase of one unit in the log price scale is approximately a 2.7 fold increase of the actual price. We see that the price of the houses has generally increased as the sale date moves forward from the 1990s. It decreased around 2008 and this is explained due to the Great Recession in the United States during that time. The trend has since moved up again. Our intuition is reasonable.



Looking at the other time based variables, it seems that the newer the improvement that was built, the more expensive the house is. There is no clear trend with the other time based variables.

2.1.3 Quality

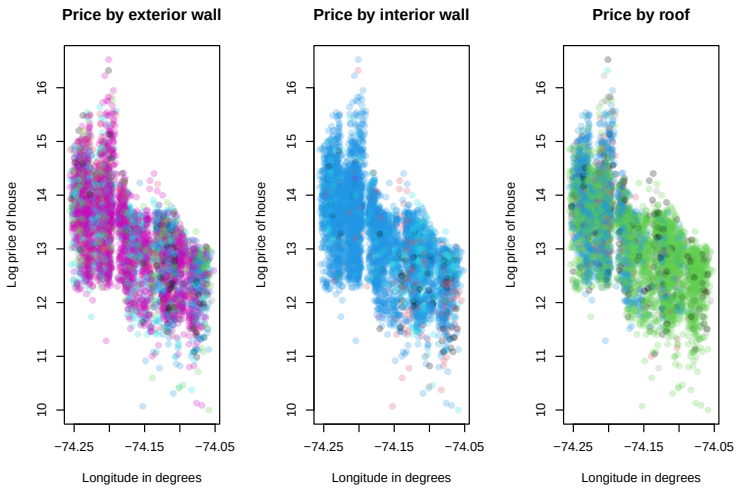
Here, we look to see if there is any trend of house prices based on the quality of the house. We might expect that a house with better quality is more expensive.



These two boxplot graphs show us that, in general, the better the quality of the house, the more expensive the house is. Our intuition would seem to be reasonable.

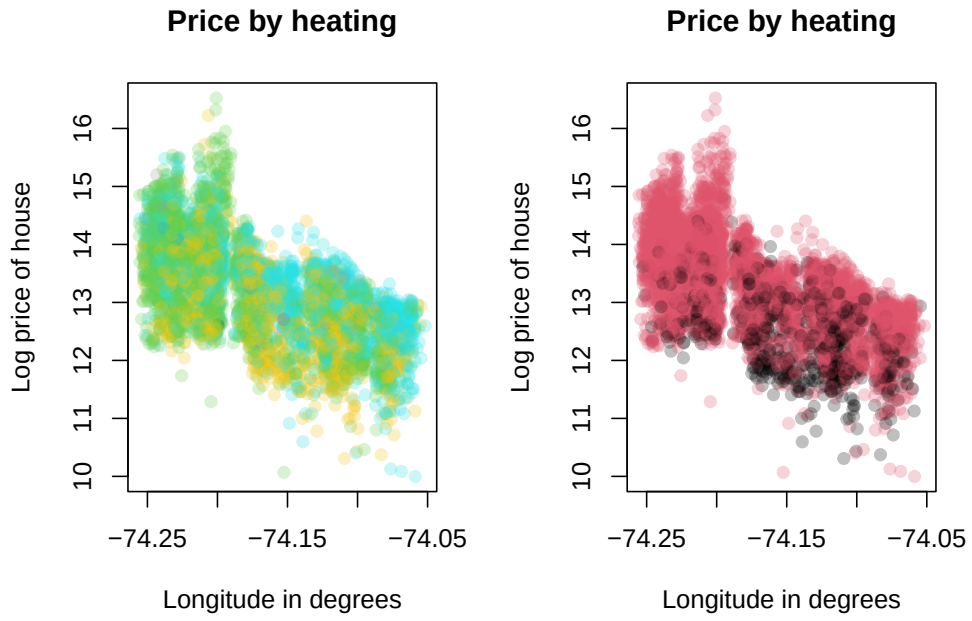
2.1.4 Features

For this section of the variables, we might want to ask whether the type of materials used for the house influences the price of the house.



From these three plots, it seems that the materials the roof is made of might influence price. We see that the houses with roofs labelled in blue are generally more expensive than the roofs labelled in green. The other two variables do not seem to show any sort of trend. The data seems to be too noisy to reach any reasonable conclusions for the type of material in the exterior and interior walls.

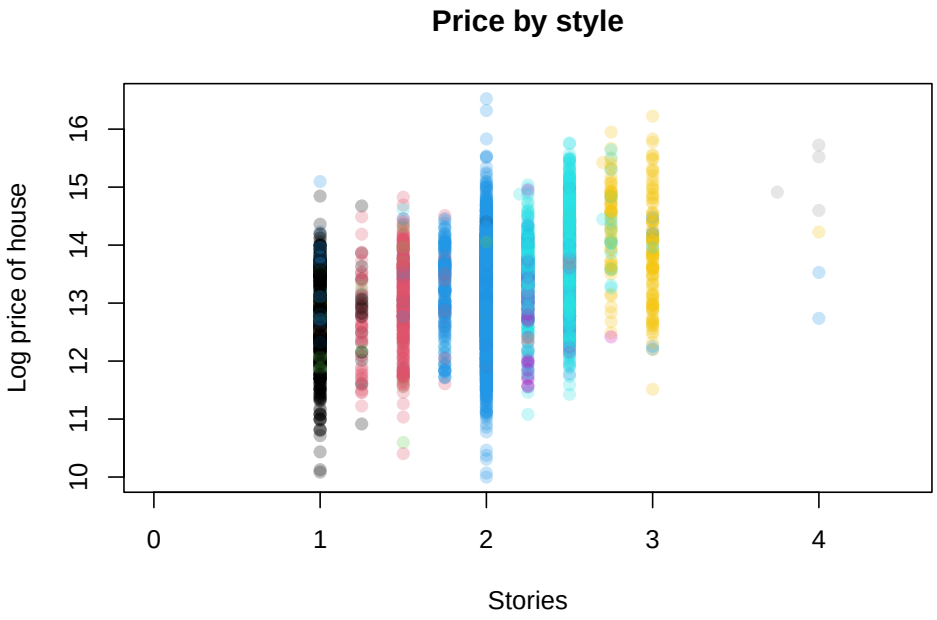
We may also want to explore whether the type of heating in the house may influence the price of the house.



From the two graphs about heating, it seems that while the type of heating in the house may not influence the price of the house, houses with air conditioning seem to be generally more expensive than houses without air conditioning.

Finally, we check to see if the style of the house influences the price of the house. Here, the style refers

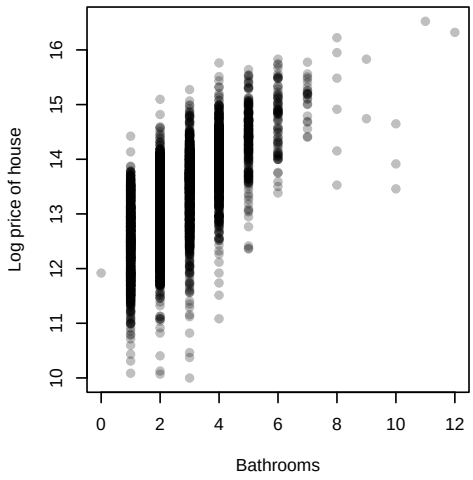
not to the architectural style of the house (e.g. Victorian, Art Deco) but rather the structure of the house (e.g. type of house based on how many stories it has).

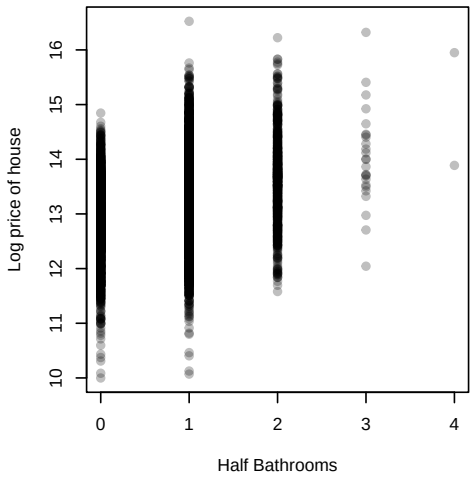


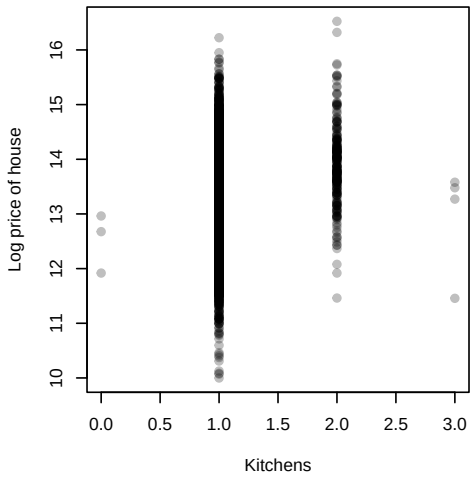
We see from this graph that, in general, the more stories a house has, the more expensive the house is. The colouring is based on the style and we see that the styles are usually associated with the number of stories in a house.

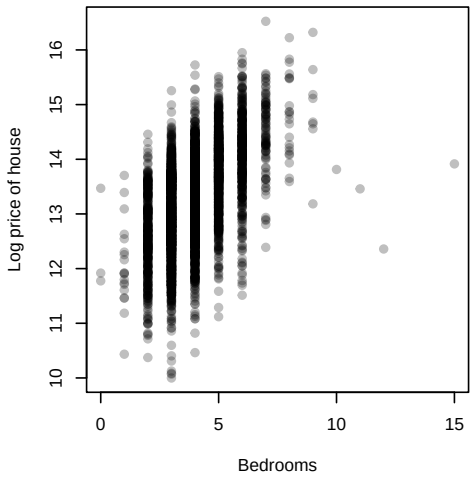
2.1.5 Rooms

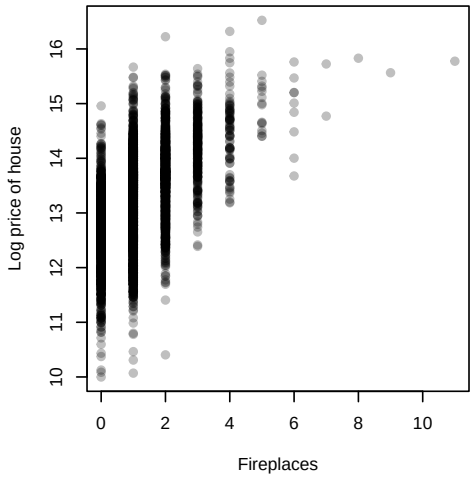
This last section is based on how many listed rooms a certain house has. We might want to ask whether the number of listed rooms a house has influences the price of the house.

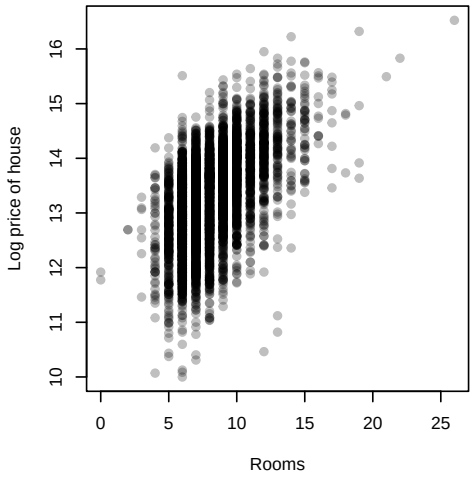

Log price of house
10
11
12
13
14
15
16
0
2
4
6
8
10
12
Bathrooms


Log price of house
10
11
12
13
14
15
16
0
1
2
3
4
Half Bathrooms


Log price of house
10
11
12
13
14
15
16
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Kitchens


Log price of house
10
11
12
13
14
15
16
0
5
10
15
Bedrooms


Log price of house
10
11
12
13
14
15
16
0
2
4
6
8
10
Fireplaces


Log price of house
10
11
12
13
14
15
16
0
5
10
15
20
25
Rooms

From the graphs above, it would seem that having more rooms listed in a house increases the price of the house.

In summary, we see that many of the variables in the dataset have some sort of significance into determining the price of a house. The next section of this report will explore topics such as variable importance and sensitivity to outliers. Alongside these concepts, we will construct a smoothing based model and a random forest based model which of the two models performs the best in this kind of problem.

3 Statistical analysis

3.1 The Models for Each Method

3.1.1 Smoothing

The final model is:

```
fit.sm <- mgcv::gam(price^(1/15)~s(proximity, by = ward, k = 26)+s(saledate,ayb,eyb)
                            +te(gba, landarea)+s(grade)+s(rooms,bedrm,bathrm)+s(fireplaces, k = 5)
                            +s(as.numeric(heat),by = ac)+s(hf_bathrm, k = 4)+s(cndtn, k = 6)
                            +roof +extwall+intwall+s(nbhd)
                            , data=datSmooth)
```

3.1.2 Random Forest

The final model is:

```
fit.rf <- ranger::ranger(price^(1/15) ~ saledate + nbhd + longitude + proximity +
                           gba + grade + bathrm + eyb + ward + cndtn + latitude,
                         data = datRF, mtry = 7, respect.unordered.factors = TRUE,
                         min.bucket = 1, min.node.size = 3, max.depth = 42)
```

3.2 Comparison summary

Model	Run Time (s)	Training Time	CV Error (APSE)
Smoothing Model	191.11	~10 hours	223731.4
Random Forest Model	5.92	~2 hours	103260.7

3.3 Predictive accuracy

We see that the predictive accuracy for the random forest is much higher than the predictive accuracy for the smoothing model. As we will discuss further down in this section, there may be some reasons for this discrepancy. The main metric for this cross-validation error is the average prediction squared error (APSE). For response variates like the prices of houses, the APSE will naturally be large. Imagine that the predictions are at most 1000 dollars away from the true price of the house, the APSE will still be very large because the APSE is defined as $\frac{\sum_{i=1}^{n}(\hat{y}-y)^2}{n}$ and a difference of 1000 is 1 million since $1000^2 = 1000000$. The general rule of thumb is that the lower the cross-validation error, the better the model is at predicting the true values of the dataset. As we can see from the table, since the APSE of the smoothing model is twice as high as the APSE of the random forest model, the random forest model is the better model in this instance.

3.4 Computational complexity and runtime

The runtime of the random forest model is low. This is in part, due to the `ranger` implementation of the random forest model as `ranger` is intended to be a fast implementation of random forests. The random forest is an average over a certain number of decision trees. For the default in `ranger`, it is 500. Each decision tree randomizes the split points. From there, the decision tree splits the data and repeats until a certain threshhold is obtained. This may be through the maximum depth of the forest like in `max.depth` or through the minimum number of observations in a leaf node like in `min.bucket`.

Meanwhile, the runtime of the smoothing model can be quite high as we see from the table. The `gam` method is the generalized additive model. There are a few unique functions that can be used in `gam`. The two that used in the model are `s()` and `te()`. `s()` is the smoothing spline function while `te()` is the tensor product function. While the tensor product is relatively inexpensive, the smoothing spline can be computationally expensive. What also makes the smoothing model runtime long is the metric it uses to solve the smoothing

problem. `gam` uses the generalized cross-validation criterion (GCV). As detailed in the smoothing project, GCV takes much more time to process than the fast Restricted Maximum Likelihood (fREML) does as is used in `bam`. `gam` trades off speed for accuracy and so in getting a more accurate smoothing model, it must run for longer.

3.5 Ease of use/model building

We begin this comparison of model building with the random forests model. We see that the random forests model is easy to build and easy to use. This is due to the flexibility of decision trees in general. For categorical variables like the kind of material used for the roof, we may simply factor the variable for use in the random forests model. Once the variable has been factored, all we have to do is tune the parameters for the random forest. In the `ranger` documentation, there are four main parameters to tune: `mtry`, `max.depth`, `min.bucket`, `min.node.size`. `mtry` is the number of variables to split at in each node, `max.depth` is the maximum depth of the tree, `min.bucket` is the minimum size of the leaf, and `min.node.size` is the minimum size of a node that can be split. `respect.unordered.factors` is an argument that orders unordered factors by the mean response as highlighted in the documentation for `ranger`. The documentation recommends using it as this process is not computationally expensive. While there are four main parameters to tune, for our dataset, we only needed to tune `mtry` and `max.depth` as the optimal parameters for the other two variables were the default parameters set by the `ranger` function. A thorough parameter search can be done through a nested for loop.

In comparison to the random forests model, the smoothing model using splines is more finicky to use. This finicky manner is a trade off. In the random forests model, there is very little to adjust in terms of parameters. In the smoothing model, particularly with the `mgcv` implementation, there are more options in terms of what smoothing methods to apply to the variables. As mentioned in the previous section, there are the smoothing spline and the tensor product functions. There is also the interaction function `ti()`. For the smoothing spline function `s()`, variable used must be of a numeric kind. For dealing with categorical variables that have a linear trend (e.g. grade), ordering the categorical variable so that the smoothing spline function accurately models the variable can make it harder to use. From the documentation of `mgcv`, we see that the interaction function `ti()` is a "variant designed to be used as interaction terms when the main effects (and any lower order interactions) are present." This makes it not as obvious to know whether to use `ti()` or not. In terms of parameter tuning, the only tuning that is really done for the smoothing model is the basis dimension `k`. If there are not enough basis dimensions in the variable, then tuning the dimension `k` is required as is if there are too many basis dimensions for the variable.

All in all, the random forest model is easier to build and easier to use, especially with the `ranger` implementation. It is relatively easy to find the optimal parameters through brute force with a random forest. You can not do the same with a smoothing model.

3.6 Interpretation

One of the advantages of a generalized additive model like the smoothing model we use is the ease of interpretation. Reading the formula given inside the best smoothing model, we see that it is the sum of the following:

- the proximity to the most expensive house by which ward the house is in
- the combination of the date of sale, earliest time the main portion of the building was built, and year an improvement was built
- the tensor product between the gross building area and the land area of the houses plot
- the grade
- the combination of rooms, bedrooms and bathroom
- the number of fireplaces in a house
- the heating of the house based on whether or not the house has air conditioning
- the number of half-bathrooms in a house
- the condition of a house
- the neighbourhood where the house is located at

- the type of roof, exterior, and interior wall

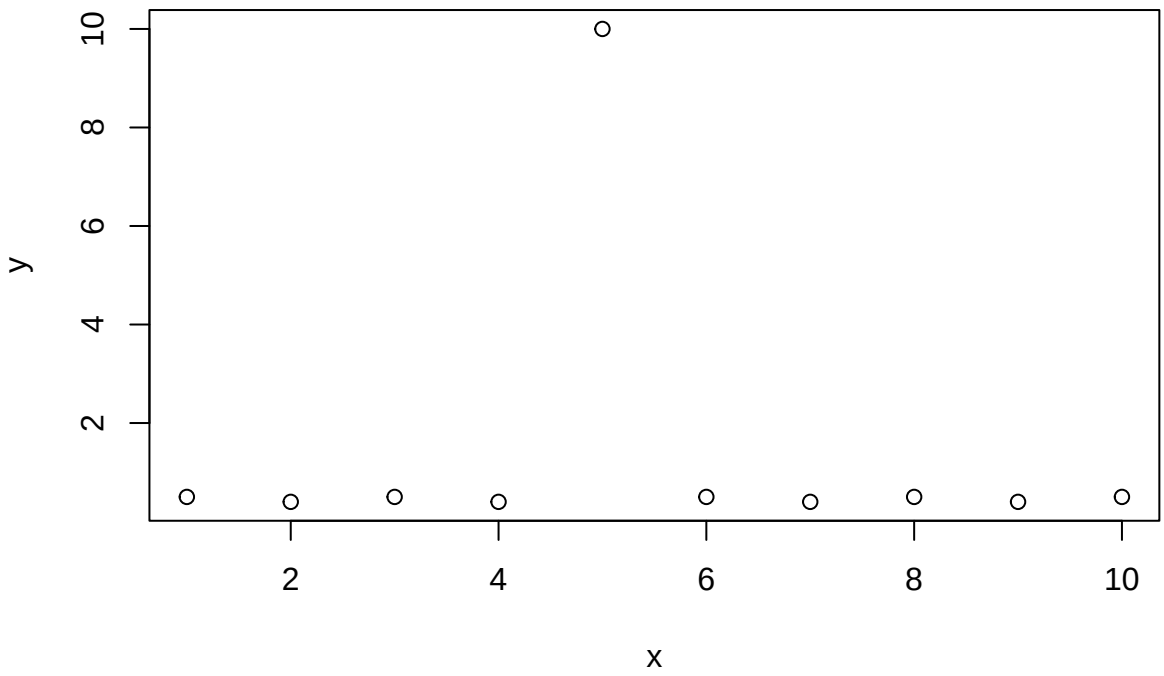
We can not say the same when it comes to the random forest model. We can say that the explanatory variables we used to produce the random forest model are:

- the date of sale and the year an improvement was built
- the neighbourhood, longitude , ward, latitude, and proximity to the most expensive house
- the gross building area, grade, condition, and number of bathrooms

However, we can not really say how these variables are related to each other. We can only see how important they are to the fitting of the model. In a sense, the random forest is kind of like a black box. We can see the output of the model, we can feed the data and parameters to the model, but we can not really see how it determines which model is the best model.

3.7 Sensitivity to outliers

Due to the random forest aggregating the values from the number of decision trees in the forest, the random forest model is generally less sensitive to outliers. The same can not be said about the smoothing model. Imagine the following plot,



Imagine that the values are pointwise. A smoothing model would smooth out the points into a curve and may not potentially capture the data properly whereas the random forest model would be able to capure the pointwise nature of the data well through implementations like the case-specific random forest.

3.8 Insights

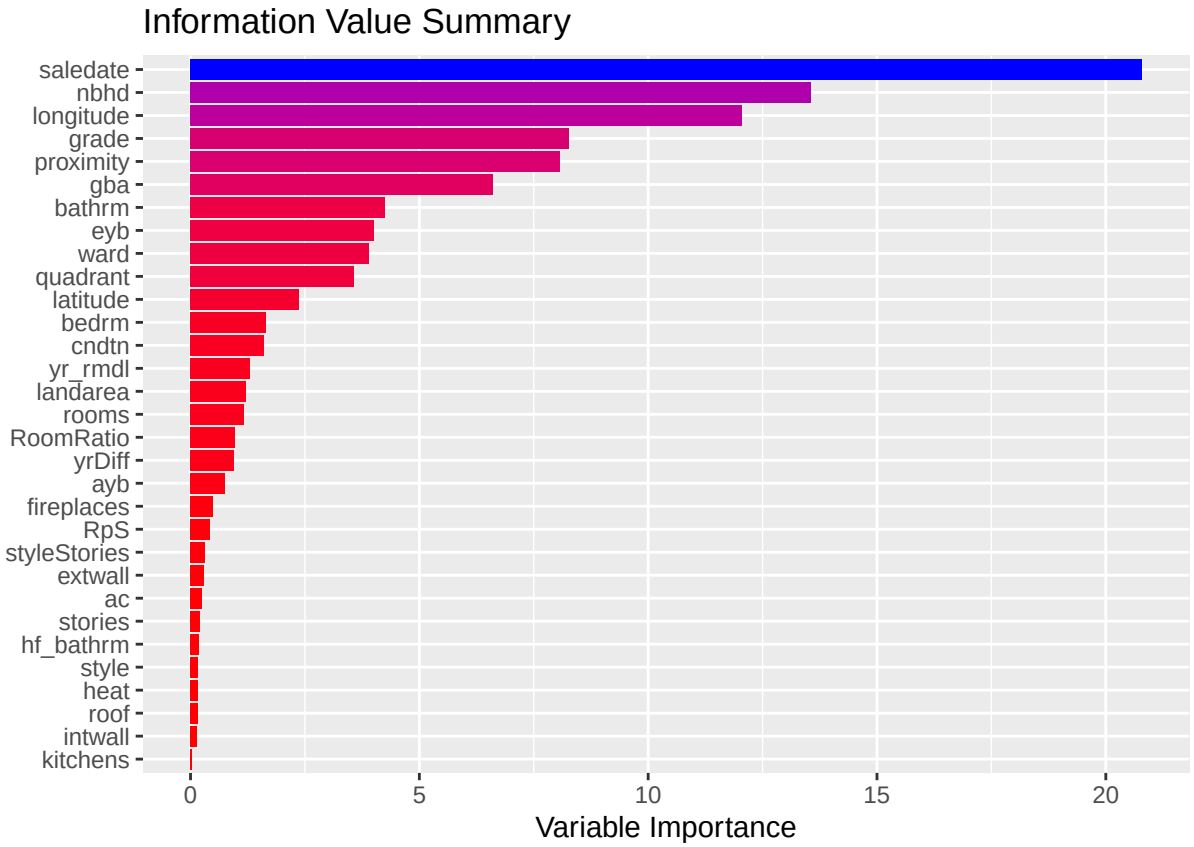
3.8.1 Important interactions

Let us briefly touch upon the interaction between the gross building area and the land area of a house.

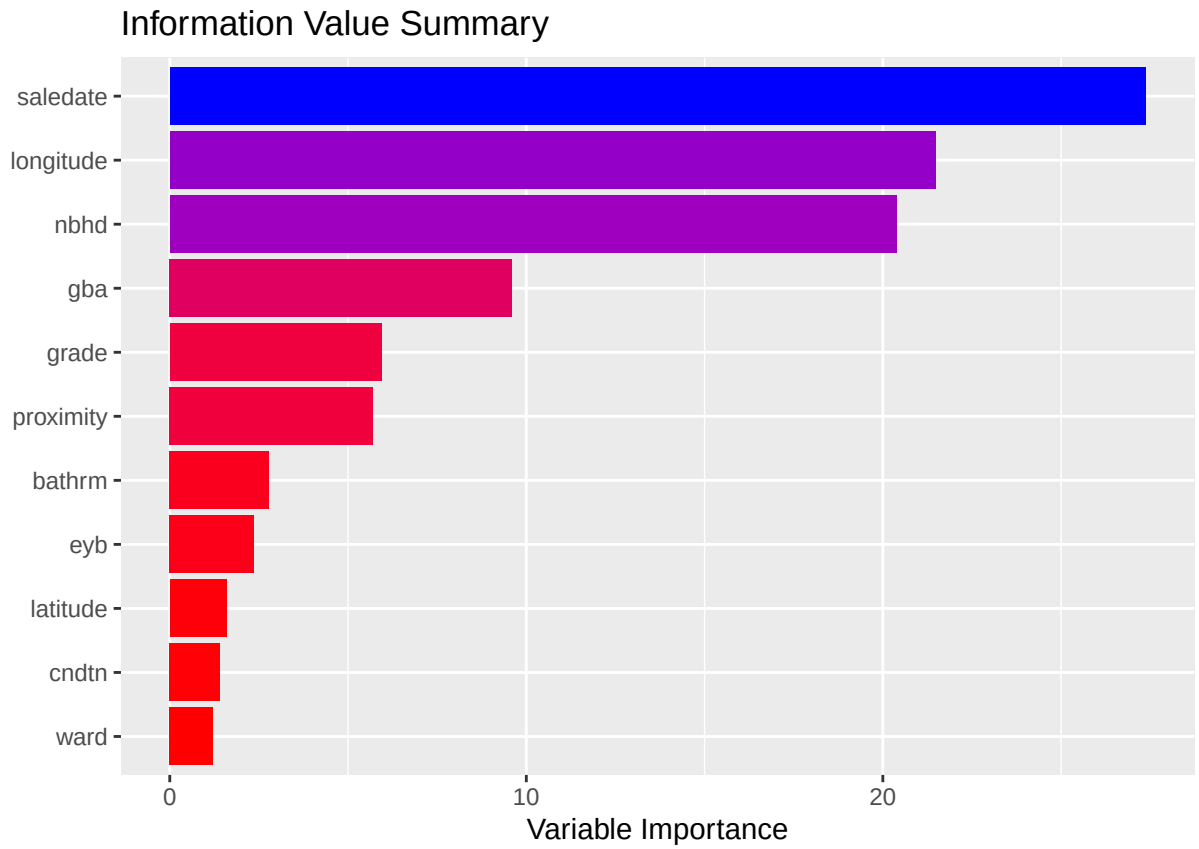
As we can see, there is a clear relationship between the gross building area and the land area. The relationship is that the more gross building area a house has, the larger the land area a house has. This is also compounded by the graphs from section one showing that houses with larger areas are also pricier.

3.8.2 Variable importance

It may be useful to see which variables end up being important for this problem. From the random forest model, `ranger` has a built in value called `variable.importance`. Let us see what the random forest model considers as important.



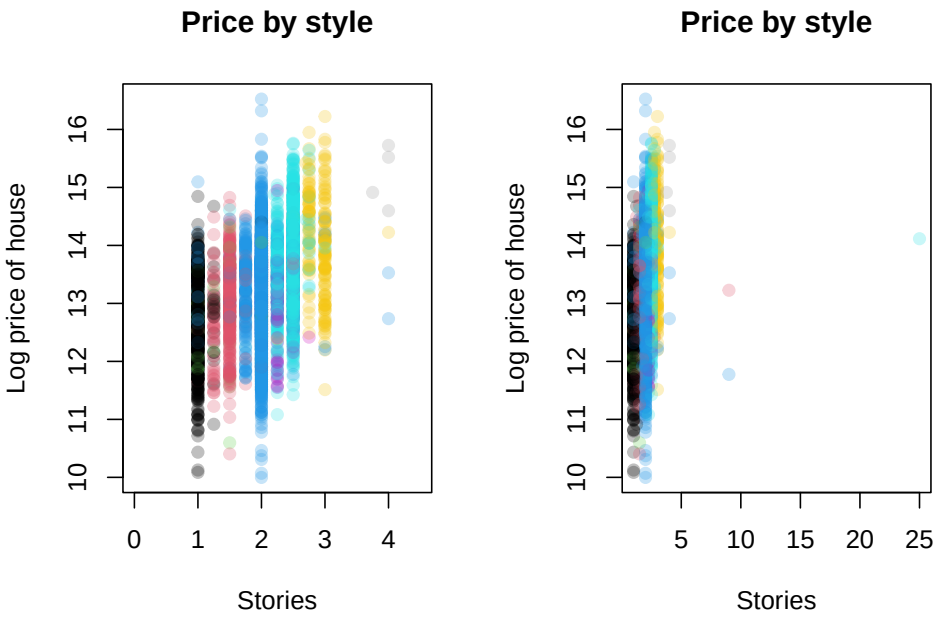
We see from a model that uses all the variables that the date of sale is the most important variable in the problem. Barring that, the neighbourhood that the house is situated in, the quality of the house, the gross building area, and the longitudinal position of the house are the most important variables in this problem. Let us compare it to the best model we have for the random forest.



We see that in the smaller model, there is more importance placed on the date of sale, longitudinal position, and neighbourhood of the houses.

3.8.3 Outliers

In the first section of the report, we provided some plots to show what the data looks like from some relations like the date of sale versus the price of the house. What some of the graphs omitted, were outliers among the data that we wanted to present. Take for example, the plot of the number of stories in a house compared to the price of the house. A look at the plot for the full dataset reveals the following:



There are three houses that seems to have more than 5 stories. Let's take a closer look at these houses.

	rooms	ayb	yr_rmdl	eyb	stories	saledate	style	landarea	quadrant
296	11	1939	1975	1960	25	2013-09-15	2.5 Story Fin	7398	NW
3435	7	1966	NA	1969	9	2018-04-19	Default	4655	NW
4220	0	1905	NA	1943	9	2002-05-10	2 Story	2496	SE

It seems that the only thing that is similar between these three houses is that they had their last new part of their homes constructed before 1980. However, take a look at house #4220. It says that there are 0 rooms in this house. That too is puzzling. How could a house have 0 listed rooms? Taking a look at the houses with 0 listed rooms we have,

	bathrm	rooms	bedrm	ayb	yr_rmdl	eyb	stories	saledate	style	landarea	quadrant
457	0	0	0	1941	NA	1928	1	2006-01-06	1 Story	3011	NE
4220	1	0	0	1905	NA	1943	9	2002-05-10	2 Story	2496	SE

We see that the two houses with 0 listed rooms were built before the 1950s. In fact, house #457 is a peculiar house.

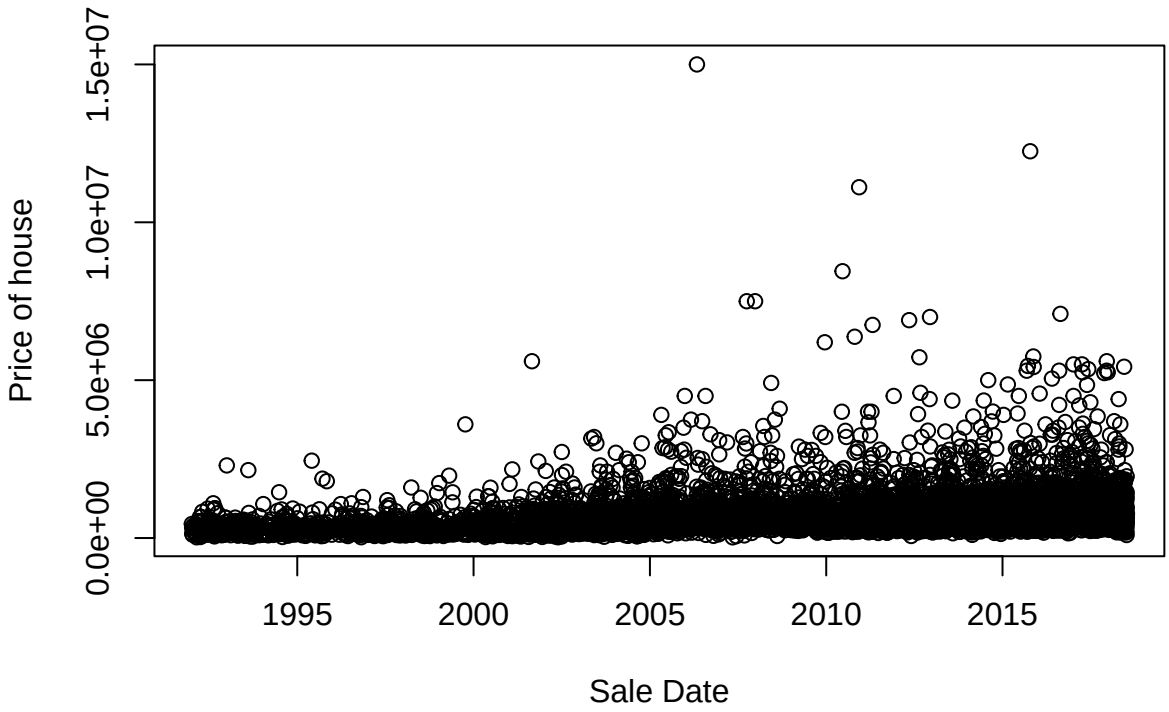
	bathrm	hf_bathrm	heat	ac	rooms	bedrm	ayb	yr_rmdl	eyb	stories
457	0	0	No Data	N	0	0	1941	NA	1928	1

	saledate	price	gba	style	grade	cndtn	extwall	roof	intwall	kitchens
457	2006-01-06	150000	640	1 Story	Low Quality	Poor	Common Brick	Comp Shingle	Lt Concrete	0

	fireplaces	landarea	latitude	longitude	nbhd	ward	quadrant
457	0	3011	40.70116	-74.07629	B9	Ward 7	NE

It would seem that this house has no rooms because this house would reasonably be a house that has been abandoned. Evidence for it being abandoned can be pointed to the quality of house where the grade of the house is Low Quality and the condition of the house is Poor. It seems that the house has been gutted as there are no listed rooms. This also seems to suggest that the sale of this house is for someone who intends to potential use this house as a fixer-upper or to demolish the house to build a new house on the property. As we can see, there is comparatively more land area than gross building area.

Another place to look for outliers is in the sale date versus price plot.



We see that there is a few clear outliers. Let us take a look at these houses.

	bathrm	hf_bathrm	heat	ac	rooms	bedrm	ayb	yr_rmdl	eyb	stories
2176	11	1	Warm Cool	Y	26	7	1986	2014	2010	2
4271	8	2	Hot Water Rad	Y	14	8	1941	2013	2000	3
4887	12	3	Warm Cool	Y	19	9	2008	NA	2016	2

	saledate	price	gba	style	grade	cndtn	extwall	roof	intwall
2176	2006-05-03	15000000	11616	2 Story	Exceptional-D	Excellent	Stone	Neopren	Hardwood
4271	2010-12-08	11111111	6937	3 Story	Exceptional-D	Excellent	Common Brick	Metal-Sms	Hardwood
4887	2015-10-15	12250000	12713	2 Story	Exceptional-D	Excellent	Stone/Stucco	Concrete Tile	Ceramic Tile

	kitchens	fireplaces	landarea	latitude	longitude	nbhd	ward	quadrant
2176	2	5	23541	40.72645	-74.20089	D8	Ward 3	NW
4271	1	2	12749	40.71982	-74.20646	C8	Ward 2	NW
4887	2	4	16525	40.72580	-74.20122	D8	Ward 3	NW

A few immediate observations come to mind. These three houses are newly built as they have been remodelled or renovated within the last 10 years. They are large houses. The main giveaway is that they have large gross building areas and large land areas. They also have many bedrooms, bathrooms etc. They are situated very closely to one another. Finally, they are the only houses that sold for more than 10 million dollars. How much of an influence do these outlier prices have? We will look at it in two ways: the average price of a house in the year that it sold, and the neighbourhood the house is situated in.

First, the year in which the house is sold. We have the following:

```
mean(dat$price[which(years == 2006)]) # Average house sale price in 2006
```

```
## [1] 905594.5
```

```
difference2006
```

```
## [1] 75757.58
```

```
mean(dat$price[which(years == 2010)]) # Average house sale price in 2010
```

```
## [1] 976253.9
```

```
difference2010
```

```
## [1] 48947.63
```

```
mean(dat$price[which(years == 2015)]) # Average house sale price in 2015
```

```
## [1] 1001936
```

```
difference2015
```

```
## [1] 29951.1
```

We see that the house sold in 2006 increased the average price of a house sold in 2006 by 75,758 dollars. Similarly, we see that the house sold in 2010 increased the average price of a house sold in 2010 by 48,948 dollars and the house sold in 2015 increased the average price of a house sold in 2015 by 29,951 dollars. Now, we turn to the neighbourhoods that the houses are in.

```
mean(dat$price[which(dat$nbhd == "D8")])
```

```
## [1] 2921212
```

```
# Average house sale price in neighbourhood D8
differenceD8
```

```
## [1] 648809.5
```

```r
mean(dat$price[which(dat$nbhd == "C8")])
```

```
## [1] 3276698
```

```r
# Average house sale price in neighbourhood C8
differenceC8
```

```
## [1] 308642
```

From looking at the neighbourhoods the houses are situated in, the average price of a house without the outliers is still well over 2 million dollars. What this seems to suggest is that the houses that the most expensive are the most expensive because the neighbourhoods that they reside in have some of the most expensive houses on average. It may also seem to suggest that those neighbourhoods are receiving active development as the houses have been newly renovated.

3.8.4 Exploratory Data Analysis (EDA)

Using this dataset, what meaningful questions you can answer?

For a problem like this, there are many possible perspectives to take when answering potential questions that arise from this dataset. Nonetheless, let us try to attempt to answer some of these potential questions.

Potential Question 1: I am a person in my late 20s. If I were to move to New York City for my job, where should I live that does not break my bank?

Potential Answer: In general, try to to find places to live that are closer to the city. You may find that the apartments will be quite small. However, the trade off is that you are closer to the city centre and in general, the housing is cheaper towards the East end of New York City.

Potential Question 2: I am looking to renovate my home and potentially flip it if I decide to move out. What should I do to increase the value of my home?

Potential Answer: In general, try to modernize any part of the house that is old. If your house has no active air conditioning unit, invest the money into installing one for the house. Does the house look worn down? Make sure to renovate the house to modern standards. The impressions of how good the house looks makes a big impact into the price of a house.

Potential Question 3: I am looking to build a new house on this plot of land. What would you recommend as the best house to build?

Potential Answer: In general, aim for a two and a half story house. It seems to be the sweet spot for how many stories a house should have. There should be around 2-4 bathrooms in the house. Build only one kitchen as it seems that there is no need for a second one. Build three to five bedrooms. Install a fireplace in the house as well. One seems to be enough for the house to increase in value. If you are concerned about the materials to use for the house, it seems that the kind of material you use for the interior wall is important so keep that in mind.

4 Conclusions

To conclude, we see that for a problem like this, there are many approaches to predicting the prices of the houses in, for our dataset, the New York City area. It all involves tradeoffs. We see that while random forests can be good for being flexible to drastic changes in data, it cannot explain the relationships between the variables. The converse is true for the smoothing model. We can explain the relationships between the variables well with a smoothing model, but it can be suceptible to outliers. We see that location and time are two of the most important factors when it comes to the price of a house. We see that the prices for houses are increasing year after year aside from the Great Recession of 2008.