Exercise Set 2

Matthew Borelli & Isaac Hulsey & Tristan Robins

3/6/2020

Question 1

On average, the current model for predicting housing price is off by $67000.96. This model is in desperate need of improvement.

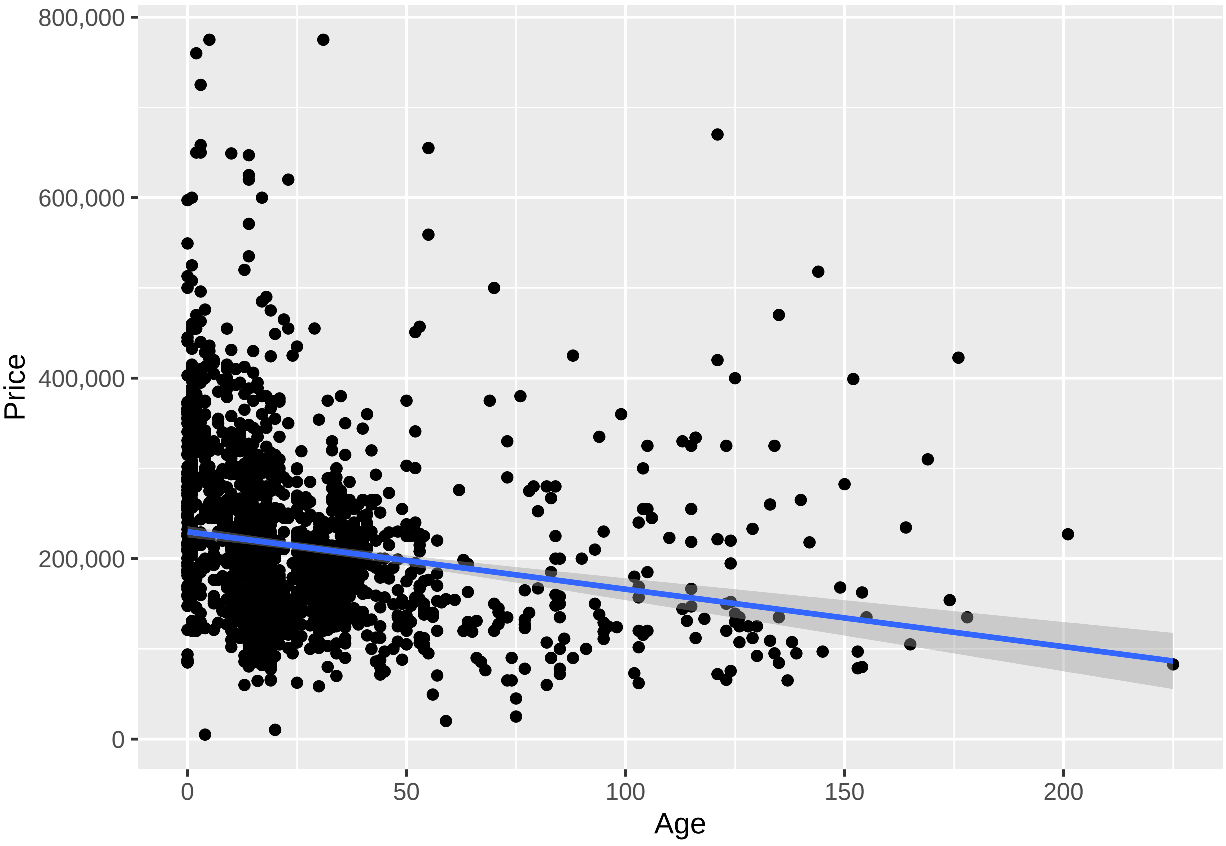
The first thing that comes to mind when looking at the current predictive model is to drop some of the variables. If highly correlatead values are in the regression, it will mess up the predictability of the model.

By dropping age and fireplaces, we get a 0.2% to 0.21% increase in predictability power, or a prediction approximately $137.93 closer to the actual price on average, but we can do much better than this.

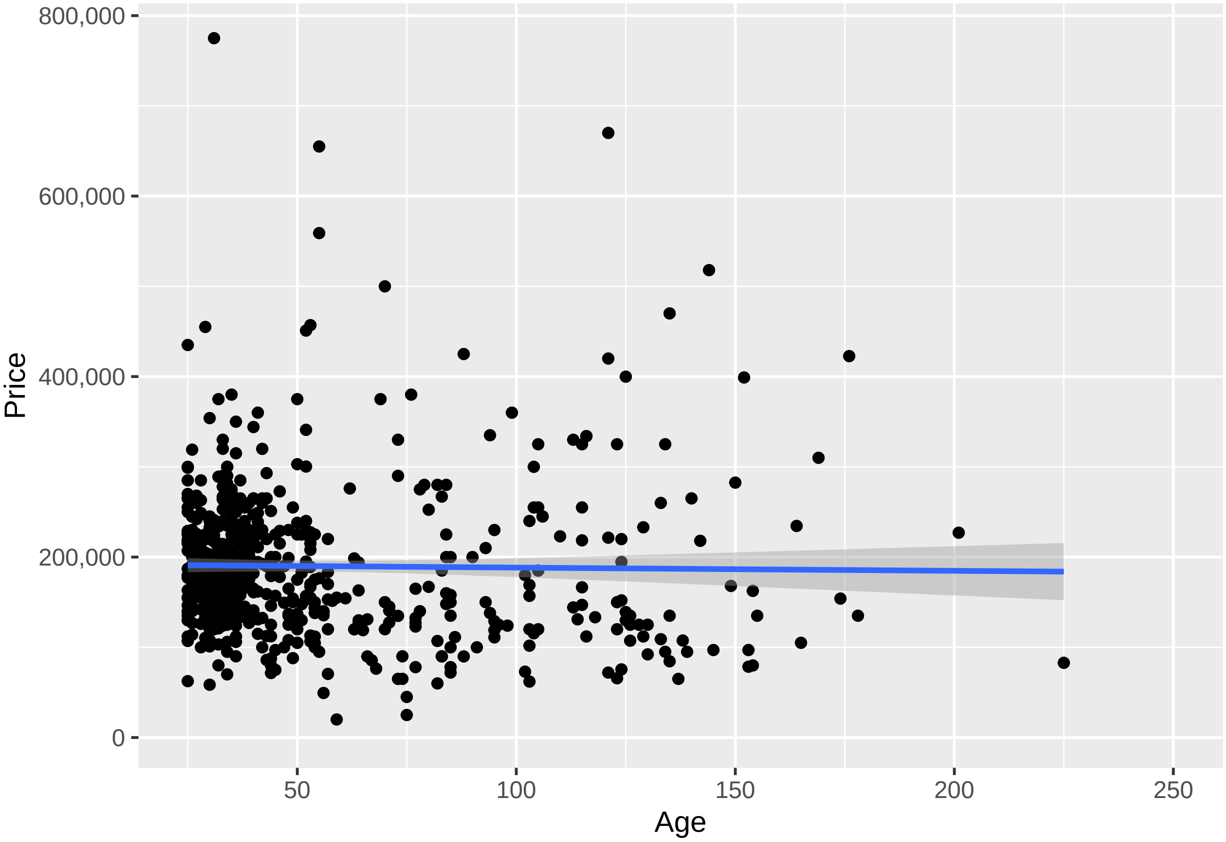
Graphical Analysis

The next thing we want to do is look at trends in the data and see what improvements can be made by adding some variables to the model.

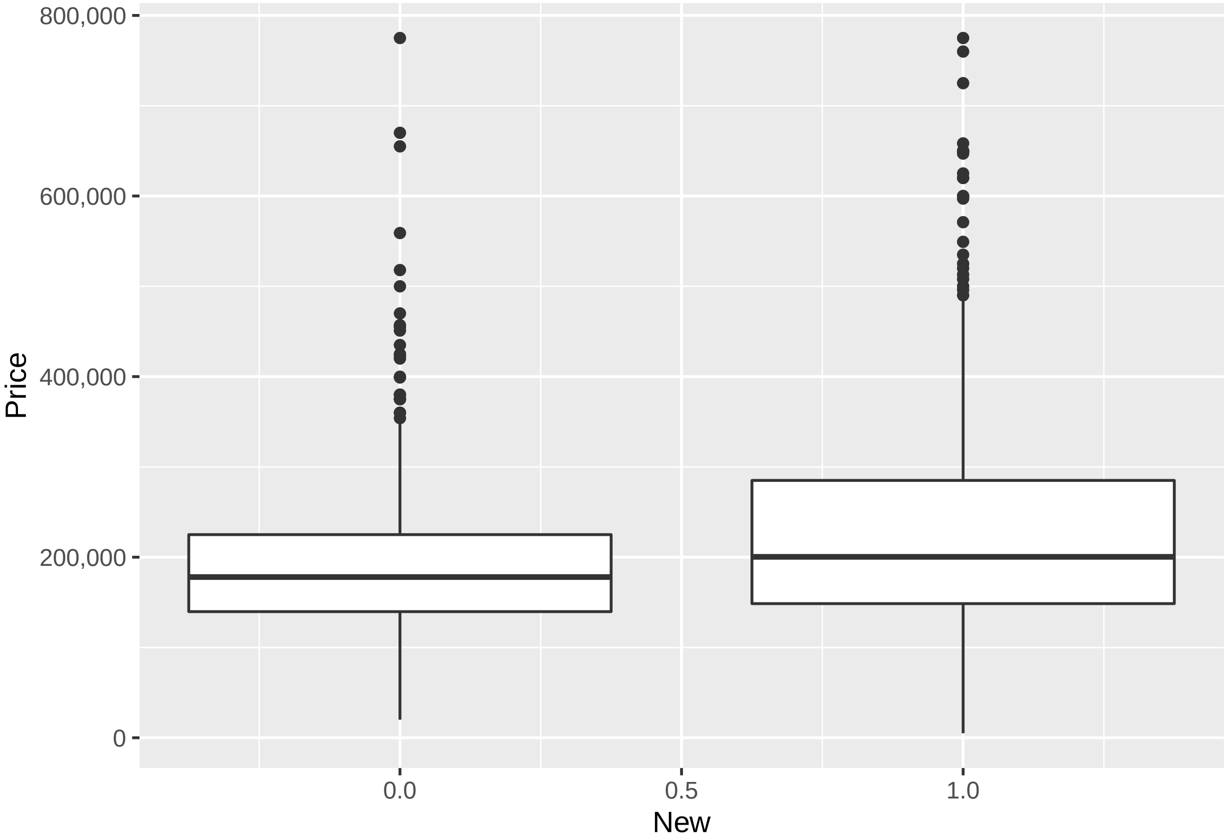
First, we will look at how price of a house changes over the age of a house in years.



The concentration of the data for houses younger than 25, and the higher value for houses before 25 years seems to be creating a downwards bias in the Age variable. I am going to graph this again except censoring houses that are less than 25 years old.

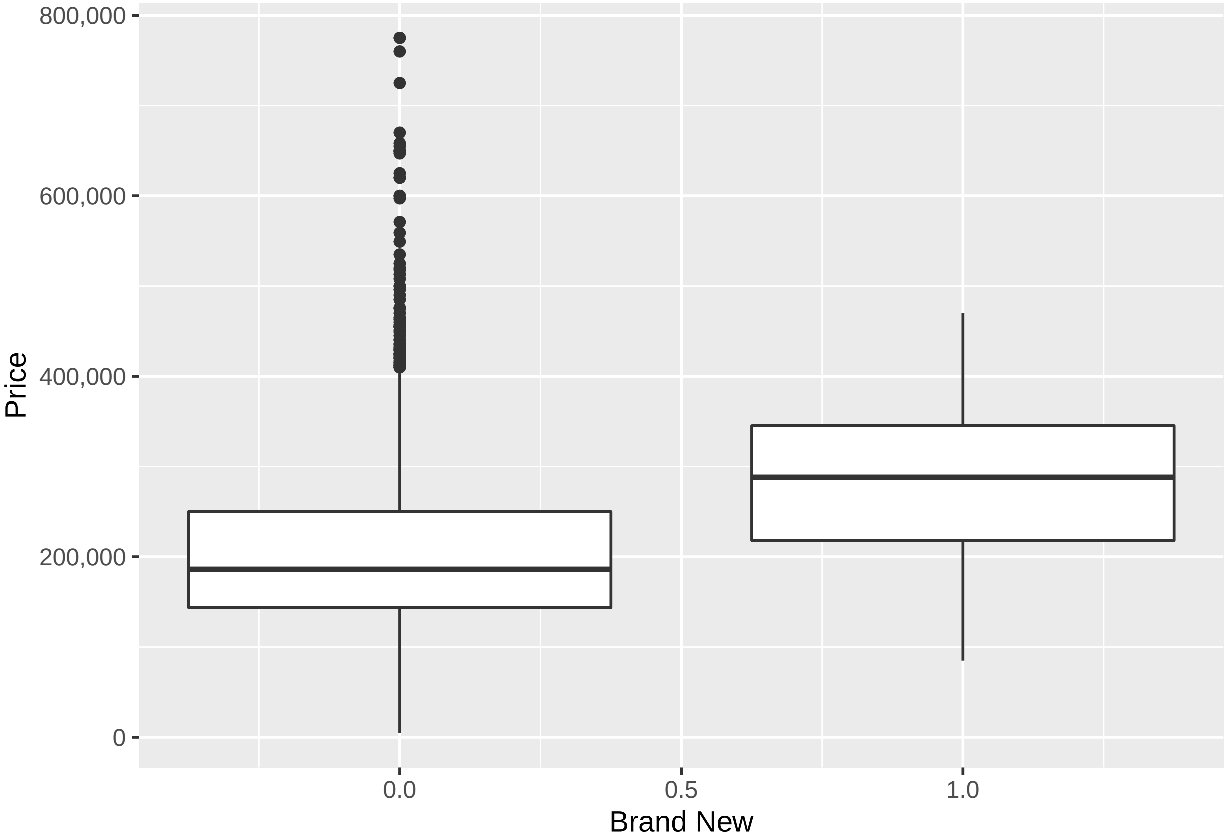


It looks like there is a downward trend for Age right before a house being 25 years old. Kind of like a new house premium that dies off over time and later has no effect on housing prices.



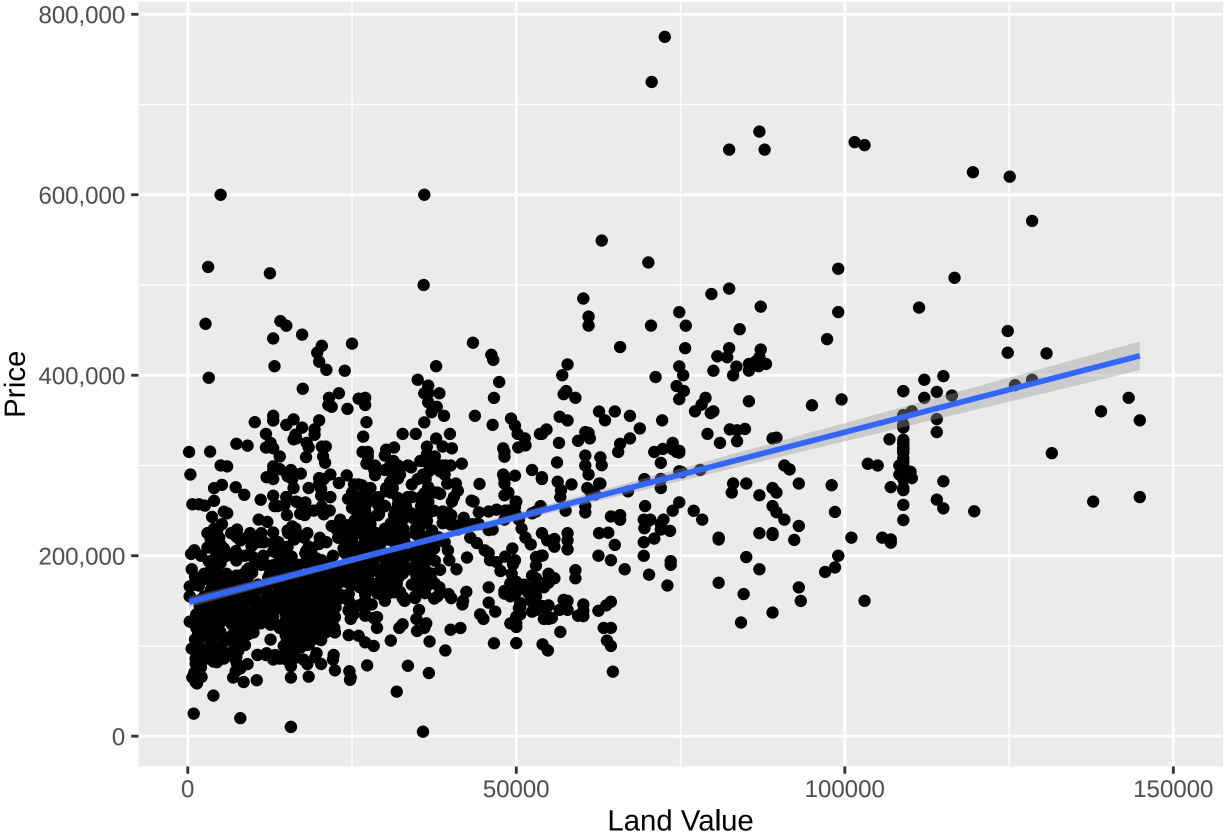
On average, houses that are 25 year of age or younger are worth roughly $25,000 more than houses that are older.

Oddly enough, new houses and a new construction aren not identical in the data, so we need to further transform the data.



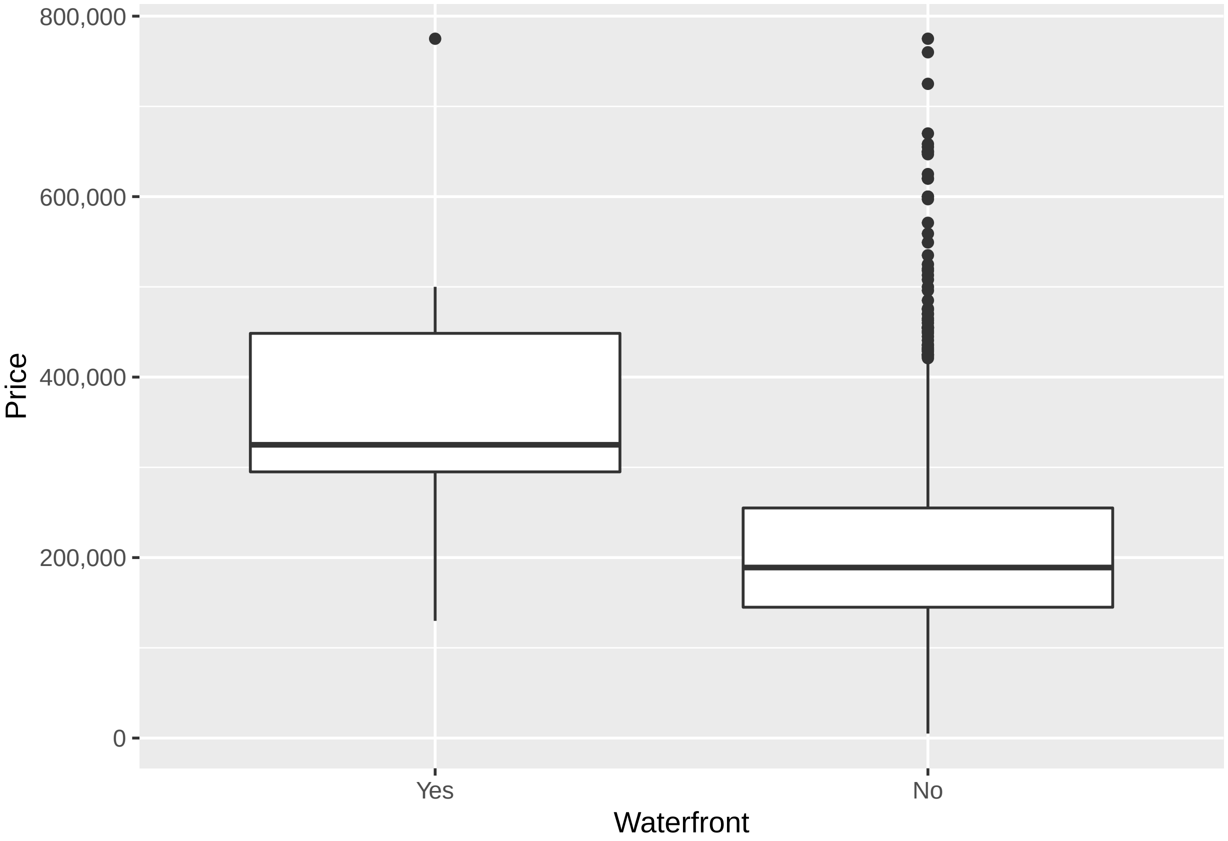
Brand New houses are denoted by 1 and not brand new houses are denoted by 0. The mean of brand new houses is higher by around $100,000 than non-brandnew houses.

Now we will look at landvalue on price



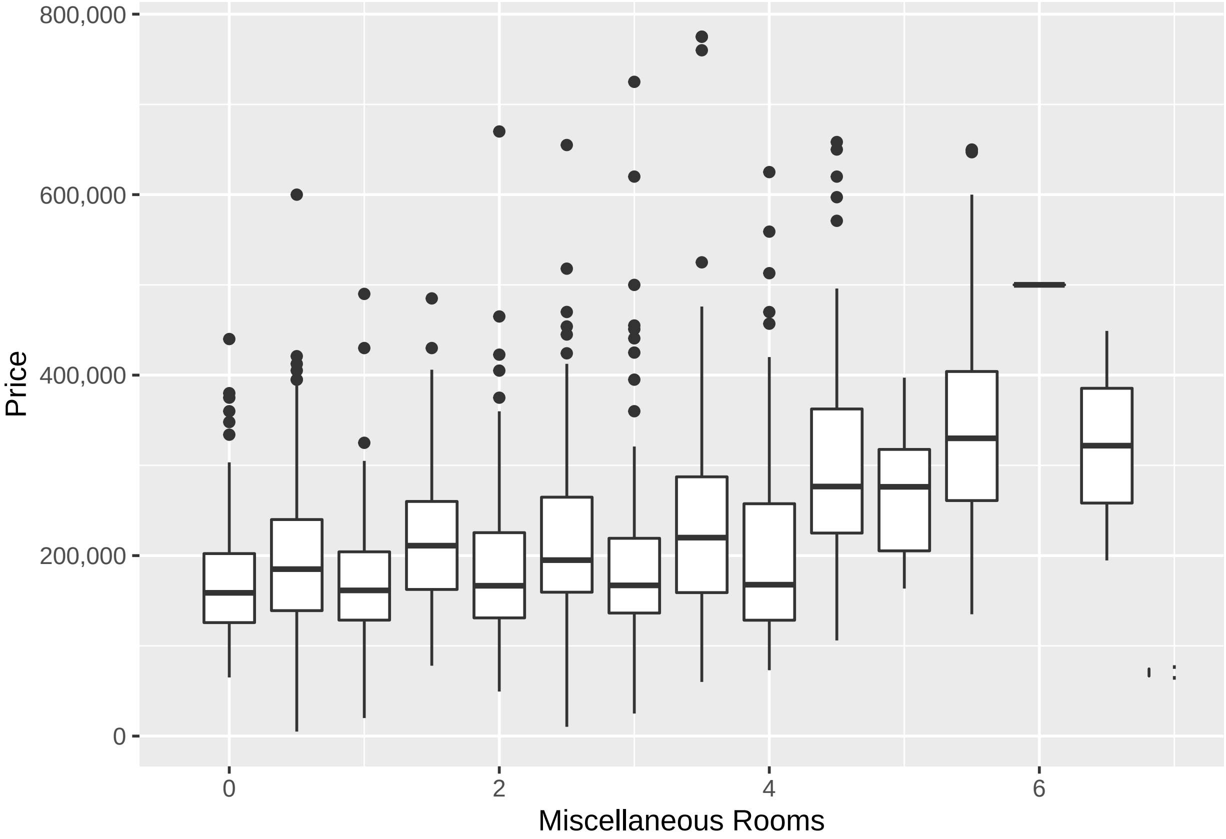
There seems to be an upward relationship between land value and price. This is not being captured in the current predictive model.

We generally assume that waterfront properties have a positive effect on housing price, and that effect is not accounted for in the current model. We will now take a look at a boxplot and see if there is anything evidence behind waterfront adding value to a price of a house.



There is incredibly strong evidence that on average a waterfront property will have a higher price than a non-waterfront property. We will include this in our model.

One final thing that I am interested in looking at is the effect of miscellaneous rooms in house prices. So, I will generate a box plot to look at the possible relationship of an additional miscellaneous room on price.



There seems to be an upward trend in the presence of miscellaneous rooms and price. We will consider adding it to our model.

Model Testing

Now time to test some of these features out. First, we will add waterfront and land value to the new regression model we createad by dropping age and fireplaces as those were the most significant variables and see what sort of jump we see in power of prediction.

We see a 10.93% to 11.03% increase in improvement by adding these two variables from the improved model. This corresponds to roughly a 7340.43$ more accurate guess in the prediction on average.

Using the model just generated as the base model, we will add the variable created for houses 25 years and younger and the variable created for newly constructed houses that were less than a year old at the time of pricing.

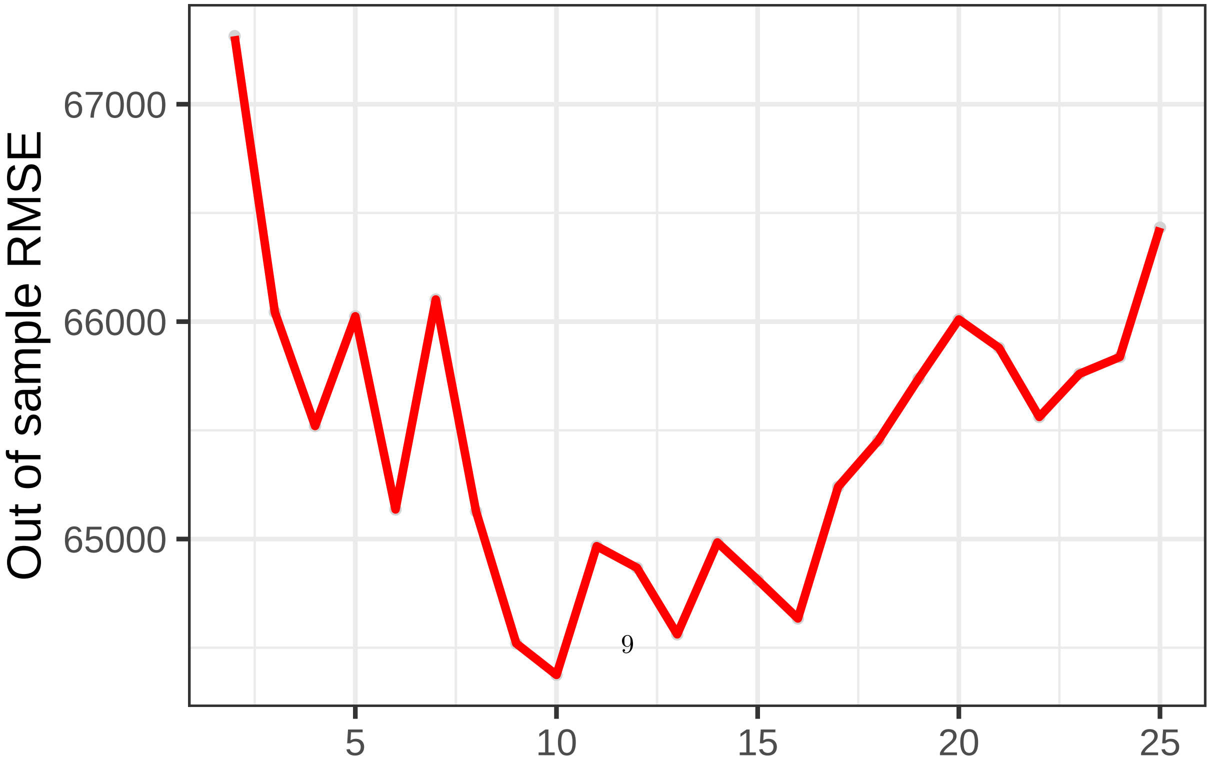
Adding those variables to measure the premium buyers pay for a newer house, we get a 1.63% to 1.64% increase in accuracy which corresponds to on average being $971.45 closer to the actual price.

For the final linear model, I will compare adding miscellaneous rooms to the model and a few interactions and compare it to the new base model.

Adding miscellaneous rooms to the model and a few interactions, we get a 0.68% to 0.7% increase in accuracy which corresponds to on average being $400.61 closer to the actual price.

Looking at our final model and the original base model, we get a 12.45% to 12.55% increase in accuracy which corresponds to on average being $8265.3 closer to the actual price.

Out of curiosity, we will see if we can make a nonparametric model that is better at predicting prices than the best model we generated above.



The above graph shows the relationship between k which is what we use to fine tune this nonparametric model and RMSE which in this case is how far off on average the model is at predicting the house price in dollars. After running lots of these regressions, I found that on average k=10 is the best k for minimizing RMSE.

Conclusion

After running tests on both models, I found that the nonparametric model was on average $60443.07 to $61584.19 off while the linear model was $57147.51 to $58173.02 off. This shows that the best model that I created is still the linear model. While this is an improvement from the base model when I started this anlaysis, there is further room for improvement. If you wish to increase the accuracy of your tax estimations and keep your constituents happy, I strongly suggest collecting more data such as a neighborhood variable because collecting new data generally is the best way to improve model preformance.

Question 2

A Hospital Audit

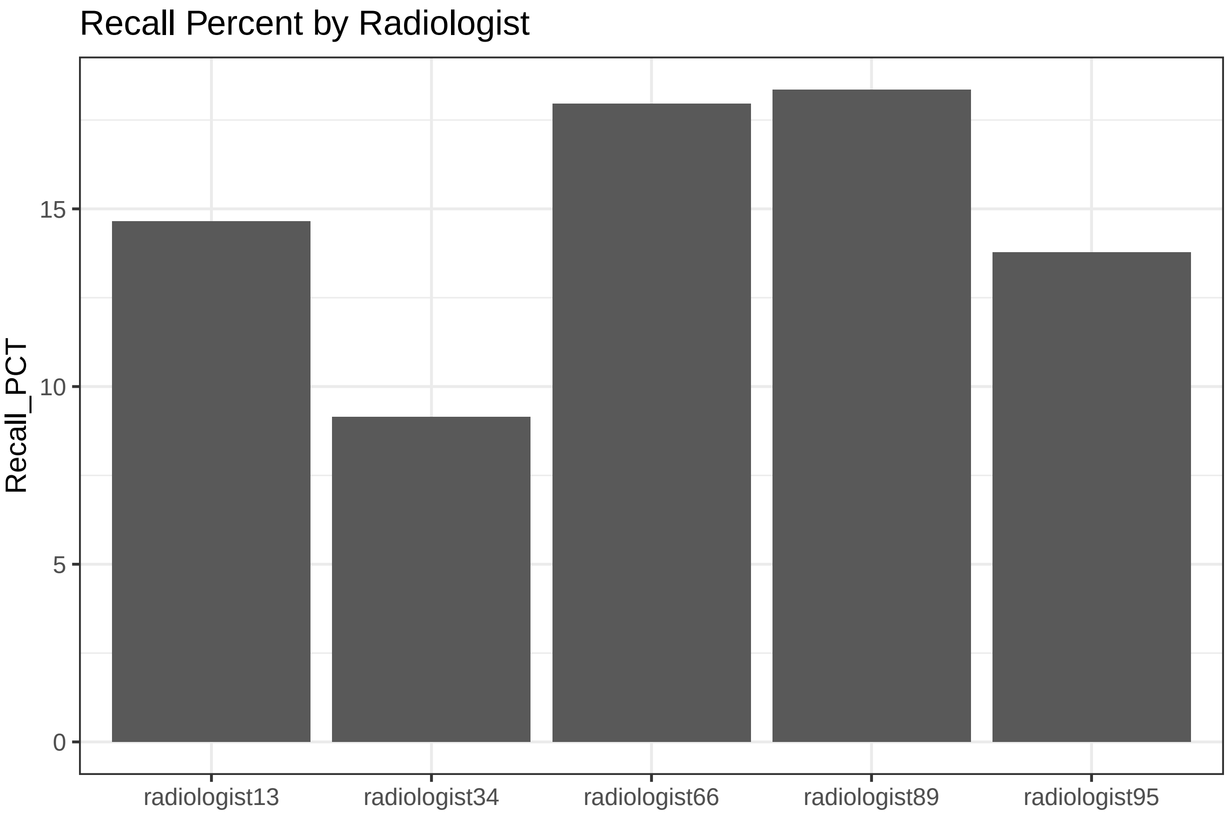
Introduction

The main goal of this statistical analysis is to audit the performance of your radiologists for mammogram screenings. Each doctor wants to limit the amount of false positives and false negatives as much as possible, but mammograms are not perfect. Therefore, it is reasonable to not expect 100% accuracy from your radiologists. However, we can utilize statistical analyses in order to see how efficient the radiologists are at recalling patients. Using the data you provided, approximately 1000 mammogram screenings from five randomly selected radiologists, we want to examing two important questions for you:

- Are some radiologists more clinically conservative than others in recalling patients, holding patient risk factors equal?
- When the radiologists at this hospital interpret a mammogram to make a decision on whether to recall the patient, does the data suggest that they should be weighing some clinical risk factors more heavily than they currently are?

Conservative Radiologists

The naive method for analyzing conservativeness amongst your radiologists would be to look at the raw recall rates. Radiologists with a higher than average recall rate would be considered “conservative”. Below is a bar chart showing the raw recall rates for each of the radiologists in our sample.



Judging from this chart, we would say that radiologists 66 and 89 are more conservative in recalling patients than the other radiologists. However, this simple analysis overlooks an important nuance: Doctors might see different patients in systematic ways that affect their recall rates. The next step we took was to control for patient factors to see if any radiologists are more conservative than others, holding all else fixed. The following is the results from a logit regression model with controls for breat cancer symptoms and breast density classification. All other controls in the data set given had no statistically significant effect on the probability of a recall.

```
##
## =====================================================
##                               Dependent variable:
##                           ---------------------------
##                                     recall
## -----------------------------------------------------
## radiologistradiologist34          -0.521
##                                   (0.326)
##
## radiologistradiologist66           0.361
##                                   (0.275)
##
## radiologistradiologist89          0.474*
##                                   (0.278)
##
## radiologistradiologist95          -0.028
##                                   (0.291)
##
## symptoms                          0.736**
##                                   (0.354)
##
```

```
## densitydensity2                   1.251**
##                                   (0.537)
##
## densitydensity3                  1.523***
##                                   (0.530)
##
## densitydensity4                   1.160**
##                                   (0.587)
##
## Constant                         -3.183***
##                                   (0.554)
##
## ---------------------------------------------------
## Observations                        987
## Log Likelihood                   -402.536
## Akaike Inf. Crit.                 823.072
## ===================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

The logit regression results do not show significant coefficients for any of the radiologists. To note, radiologist 13 is the omitted radiologist. We see no signficanct difference for any of the other radiologists other than radiologist 85, who has a positive (more conserv conservative) difference at the 10% level. It is possible that radiologists 34 and 85 are significantly different, but we can't really be certain that radiologist 34 is less conservative than the others. Overall, since we see that there aren't any radiologists in our sample that are significantally more conservative in recalling patients after using a logit model to hold risk factors constant.

Importance of Clinical Risk Factors

Now, instead of analyzing what might affect the probability of a radiologist recalling a patient, we want to see what factors best predict the likelihood of a patient having breast cancer. To do this, we will use a few different classification methods in order to better understand what risk factors best predict the probability of breast cancer.

Null Models

To start, we want to set a baseline that we can compare following models to. For our baseline, we use a null model which simply predicts the most likely answer for each observation.

```
## Cancer
##   0   1
## 950  37
```

The null model, in this case predicting that each patient does not have cancer, has approximately a 96.25% accuracy rate, or equivalently a 3.75% error rate. These statistics will be one part of our benchmark levels for evaluating the performance of the classificaiton models we will use.

We also want to look at the current performance of radiologists at your hospital, to see if the following statistical models would result in an improvement in recall accuracy.

```
##       Recall
## Cancer   0   1
##      0 824 126
##      1  15  22
```

From this table, we can generate important medical statistics that will help us compare predictions from our model against the actual decisions of radiologists:

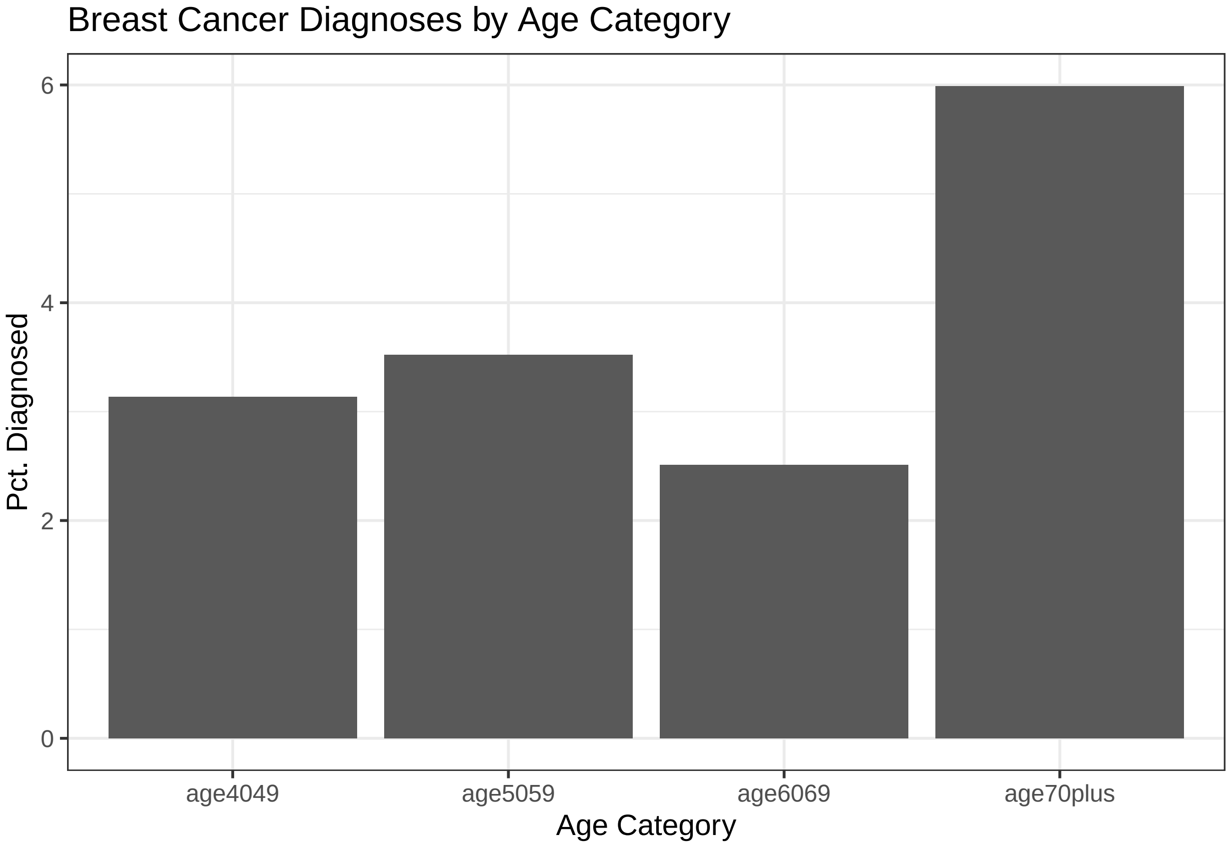
- Accuracy = 85.71%

- Sensitivity = 59.46%
- Specificity = 86.74%
- Positive Predictive Value = 14.86%

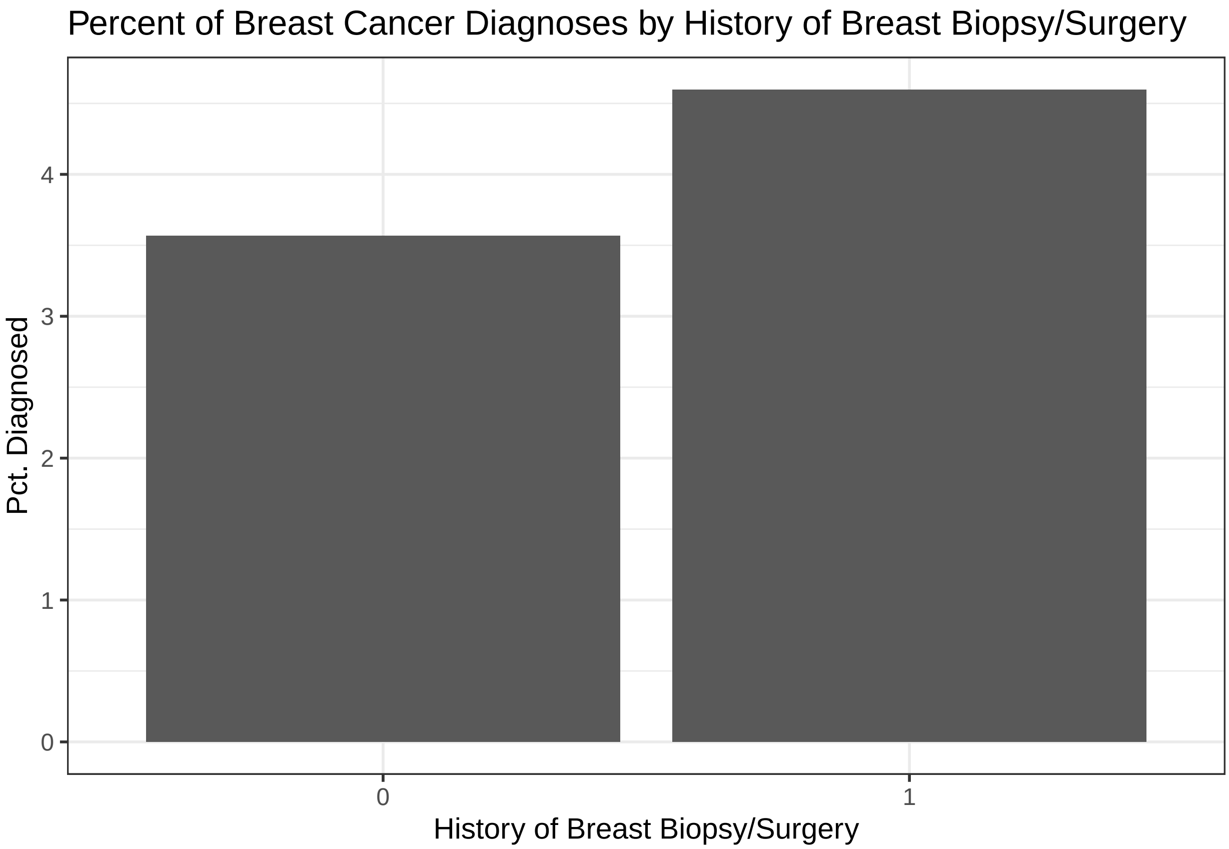
All of these give important information about the recall process, therefore we will analyze all of them for each model. However, the most important your case is the sensitivity, as the consequences of missing a case of breast cancer are worse than incorrectly suspecting that someone does have breast cancer. We will look at all four of these values, but focus on the sensitivity of these models in order to limit the possibility of missing breast cancer patients.

Graphical Analysis

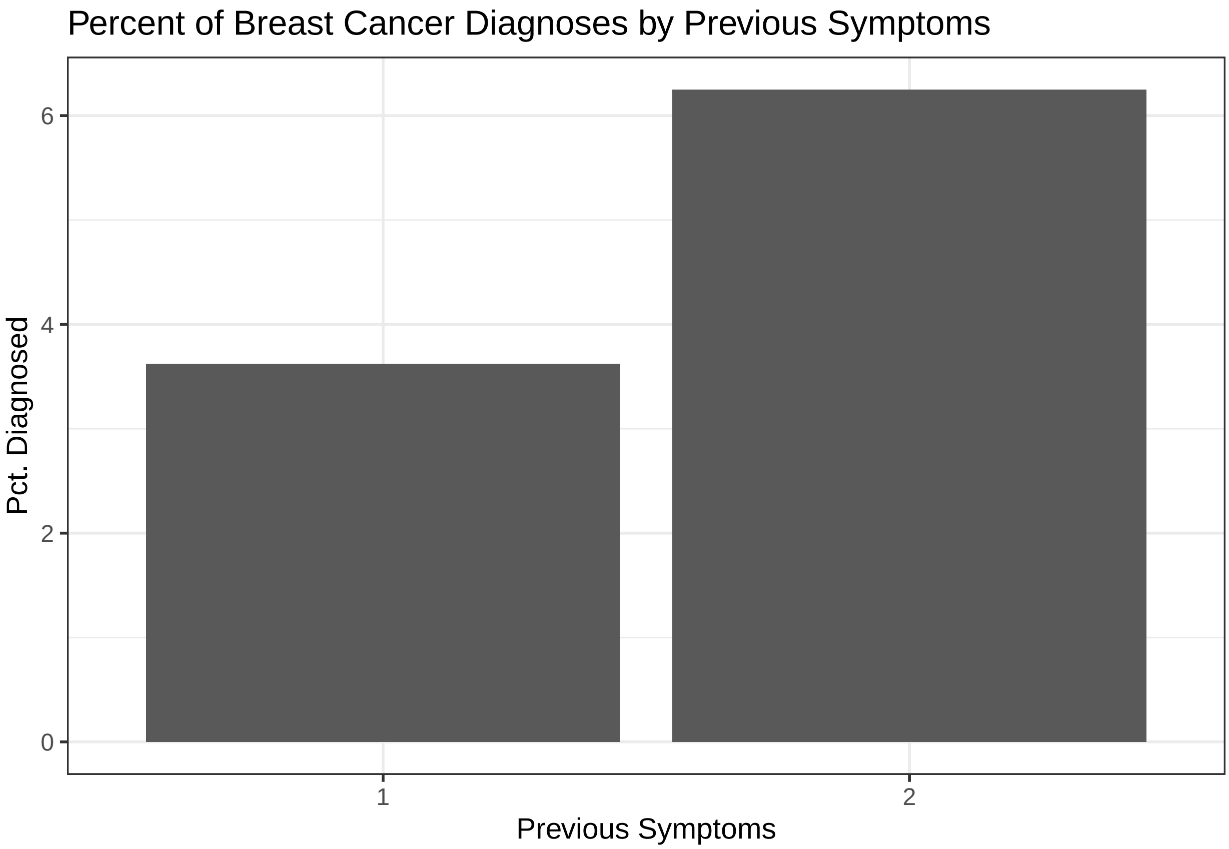
Before deciding on a model, we want to graphically look at the relationships between the risk factors and breast cancer diagnoses. To do this, we will graph simple two-way scatter plots for the risk factors in question.



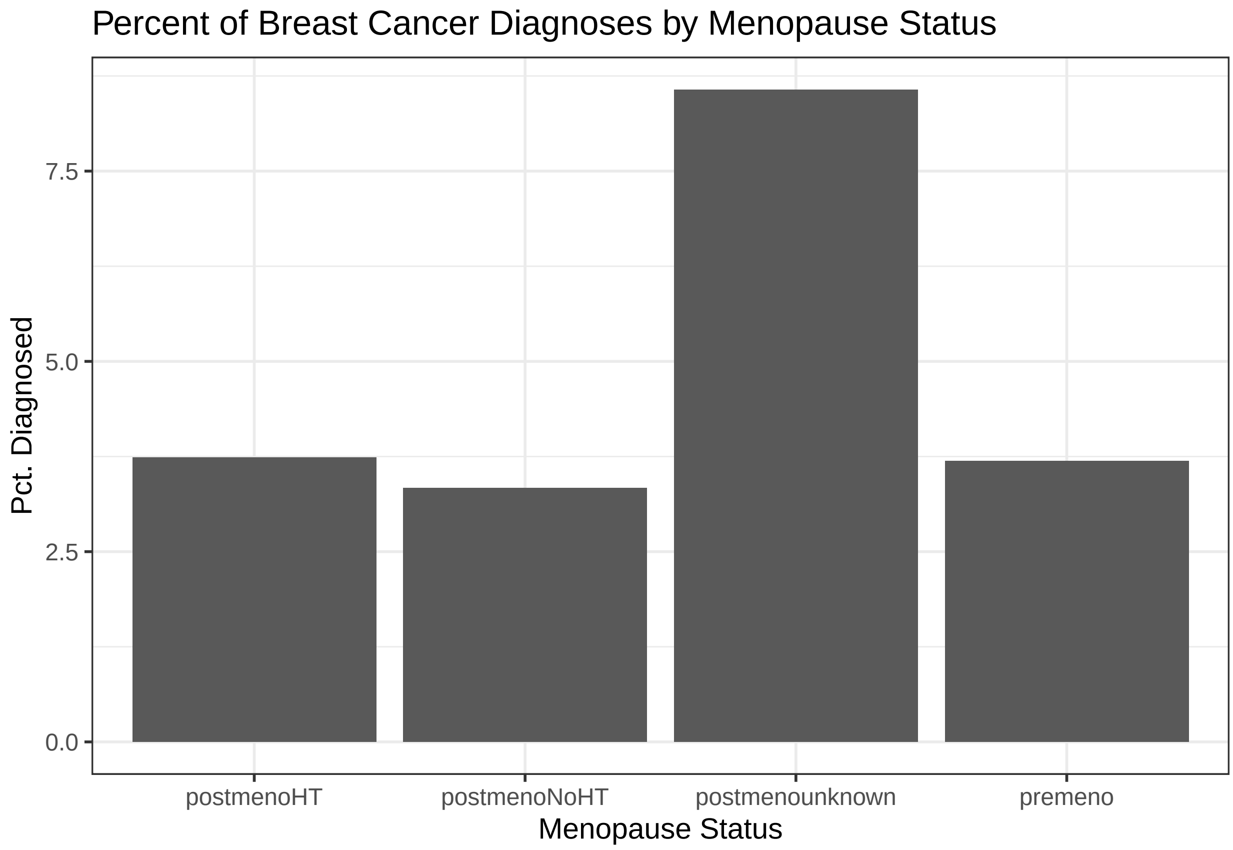
There seems to be a positive trend with age, even though ages 60-69 are less likely to be diagnosed with breast cancer. Patients aged 70+ are almost twice as likely to be diagnosed with breast cancer than any other age category.



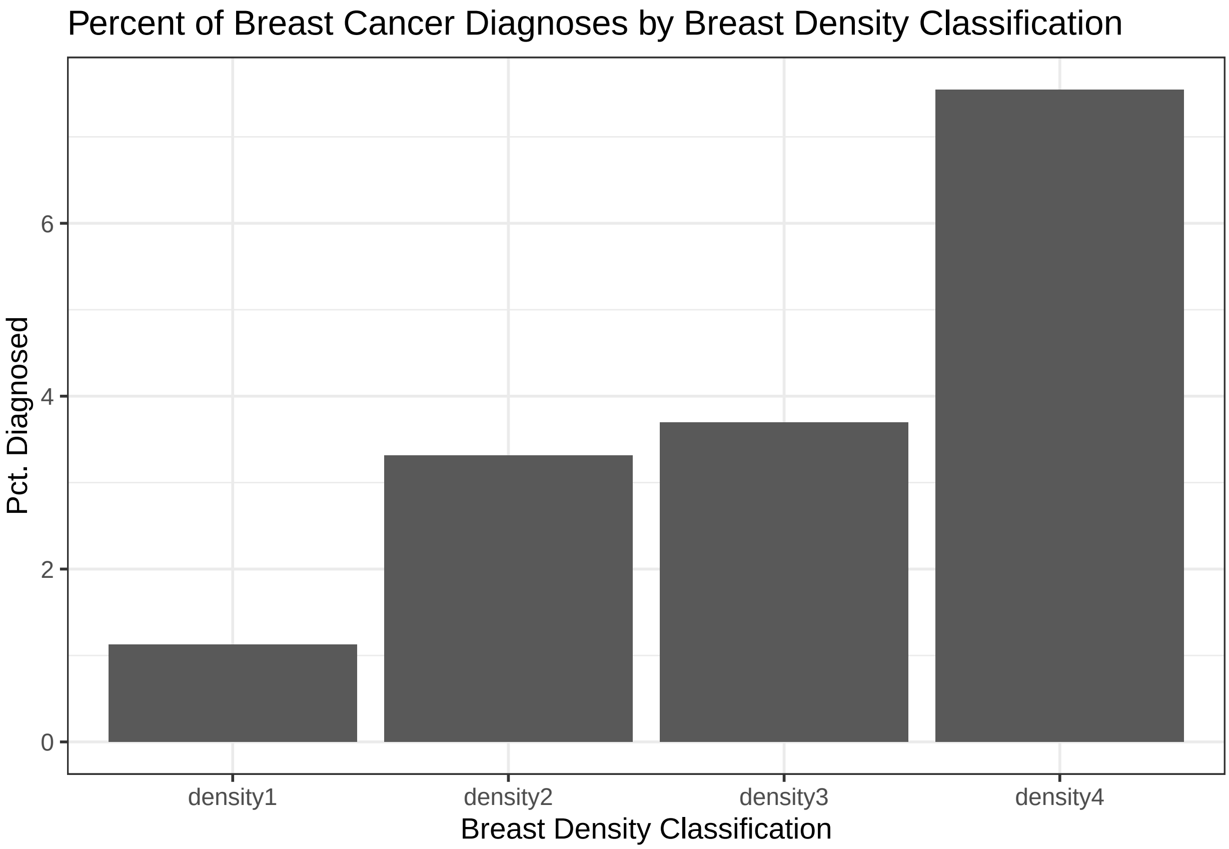
Having a previous breast biopsy or surgery doesn't seem to have too much of an impact on future cancer diagnoses, as there is about a 1% difference in diagnosis rate.



People who have had previous symptoms (coded as 2 here) are diagnosed with breast cancer at a 2% higher rate. This difference might or might not be statistically significant.



Most of the classifications for menopause status are the same, but people who are post menopausal with unknown hormone treatment seem to have a significantly higher diagnosis rate.



People classified as Breast Density 4 (Extremely Dense) seem to have a significantly higher diagnosis rate. On the other hand, people classified as Breast Density 1 (Almost Entirely Fatty) look to have a much lower rate of breast cancer diagnoses.

Importance of Risk Factors

There are two general ways to model classification problems: Linear Probability models (LPM), and K-Nearest Neighbors (KNN). Due to the relative lack of cancer diagnoses and because all of the risk factors are categorical, it makes the most sense to use a LPM instead of a KNN. If given a larger sample of the data, KNN could be a more viable option for modelling breast cancer probabilities.

To start, we will look at the linear probability model with all of the risk factors and no interactions between them. To note, we will use a 5% threshold for the limit at which a doctor should recall a patient for further testing.

```
##    yhat
## y     0   1
##   0 716 234
##   1  20  17
```

- Accuracy = 74.27%
- Sensitivity = 45.95%
- Specificity = 75.37%
- Positive Predictive Value = 6.77%

Our base model has less accuracy, sensitivity, specificity, and positive predictive value than what we currently observe from radiologists. To try and improve the model, we take the two statistically significant variables from the base model, age and density, and create a new LPM to see if we improve any of our measures.

```
##    yhat
## y     0   1
##   0 683 267
##   1  20  17
```

- Accuracy = 70.92%
- Sensitivity = 45.95%
- Specificity = 71.89%
- Positive Predictive Value = 5.99%

This reduced model is even lower in our measures of concern, indicating that we need to consider different models. Particularly, we want to look at interactions between the risk factors, as having multiple high-probability risk factors could be a better indication of the likelihood of a breast cancer diagnosis. For the selection, we will add the recall binary variable to control for potential unobservables that doctors are selecting on when deciding who should be recalled. If there's something that the radiologists are not factoring in, we would expect to see some of the risk factors still exhibit statistical significance since it would improve the decision making process for the radiologists.

```
## 
## =============================================
##                       Dependent variable:    
##                   ---------------------------
##                             cancer           
## ---------------------------------------------
## ageage5059                   0.340           
##                             (0.491)          
##                                              
## ageage6069                   0.218           
##                             (0.605)          
##                                              
## ageage70plus                1.210**          
##                             (0.491)          
##                                              
## densitydensity2              0.780           
##                             (1.074)          
##                                              
## densitydensity3              0.950           
##                             (1.065)          
##                                              
## densitydensity4             2.008*           
##                             (1.119)          
##                                              
## recall                     2.314***          
##                             (0.358)          
##                                              
## Constant                   -5.502***         
##                             (1.108)          
##                                              
## ---------------------------------------------
## Observations                  987            
## Log Likelihood             -131.467          
## Akaike Inf. Crit.           278.934          
## =============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Our final model was the same as the reduced model with the addition of the information on whether a doctor recalled a patient or not. We see that people ages 70+ and people classified with extremely dense breasts have a significant positive relationship with cancer diagnoses, even after controlling for the actual recall decision. This means that if radiologists took more consideration for those two classifications of risk factors, they would likely be more efficient at recalling patients. To show this, we look at one more confusion matrix of predicted probabilities of recalling patients vs their actual diagnosis rates.

```
##    yhat
## y     0   1
##   0 821 129
##   1  14  23
```

- Accuracy = 85.51%
- Sensitivity = 62.16%
- Specificity = 86.42%
- Positive Predictive Value = 15.13%

This final model improves on 3 of our 4 metrics, with specificity being slightly lowered. For sensitivity, which we care about the most, we have a lift of 1.05. Therefore, we recommend that radiologists take act more aggressively in recalling patients over the age of 70 and/or patients with extremely dense breast classifications in order to improve the sensitivity of mammogram testing.

Conclusion

Our audit had two focuses: Evaulating the conservativeness of radiologists in recalling patients, and examining risk factors that might be underaccounted in a radiologist's decision to recall. Our analyses did not show and significant evidence that any of the radiologists were more conservative in issuing recalls. At worst, there are two radiologists who are different at a 10% level, but neither is significantly different than the average in our sample. In this regard, we can say that your radiologists are performing well and you don't particularly need to encourage any behavior change.

We did see that there were some risk factors that weren't being appropriately accounted for in a radiologist's decision on whether to recall a patient or not. In particular, the category of patients above age 70 are more likely to be diagnosed with cancer, holding all else fixed. As well, patients with the breast density classification of "Extremely Dense" are also significantly more likely to be diagnosed with cancer, holding all else fixed. We want to make sure that radiologists still make recall decisions with the same general process, since there might be unobservable reasons that inform their decisions. Therefore, we suggest that radiologists at your hospital should be more conservative in recalling patients who are 70 and older and/or are fit the breast density classification of "Extremely Dense", while otherwise considering their recall decisions in the same way.

Question 3

Predicting how many shares a given article will receive

In order to explian how many shares an article would receive, 100 linear models were conducted using training and testing sets to measure effectiveness.

The main findings of the results of the 100 linear models for the number of shares an article received were:

Confusion Matrices

In sample:

```
##    yhat
## y          0        1
##   0   146.01 15915.56
##   1    64.40 15589.03
```

Out of sample:

```
##    yhat
## y         0       1
##   0   36.71 3983.72
##   1   15.60 3892.97
```

True Positive Rate

```
## [1] 0.9960088
```

False Positive Rate and Specificity

False positive rate:

```
## [1] 0.9908691
```

Specificity:

```
## [1] 0.009130864
```

False Discovery Rate and Precision

False discovery rate:

```
## [1] 0.5057607
```

Precision:

```
## [1] 0.4942393
```

We see that this model has only 1% specificity but has 50% precision. This means that the model predicts many false positives but remains semi-stable in modeling the data. Ideally, specificity should be higher in predicting the outcome.

Prediciting if an article will go viral

To discern whether an article would go viral, a linear probability model was constructed based on the definition of 'viral' as being whether an article received more than 1400 shares. Again, 100 of these models were constructed using training and testing sets.

The main findings of the results of the 100 linear probability models for whether an article went viral were:

Confusion Matrices

In sample:

```
##    yhat
## y         0       1
##   0 9985.33 6088.37
##   1 5759.28 9882.02
```

Out of sample:

```
##    yhat
## y        0      1
##   0 2505.2 1503.1
##   1 1459.8 2460.9
```

True Positive Rate

```
## [1] 0.6276685
```

False Positive Rate and Specificity

False positive rate:

```
## [1] 0.3749969
```

Specificity:

```
## [1] 0.6250031
```

False Discovery Rate and Precision

False discovery rate:

```
## [1] 0.3791877
```

Precision:

```
## [1] 0.6208123
```

Compared to the shares model, specificity has increased drastically. It still predicts some false positives (as will any model since no model is perfect) but predicts far fewer. Additionally, the viral model has higher predicitive power at approximately 62%. This is definite improvement in the model.

Discussion of results

Overall, it is clear that the linear probability model on the likelihood of an article going viral performed significantly better than the linear model of accounting for shares. This is becasue the linear model accounted for any article receiving more than 1400 shares, which happened quite often in the predicted values from the data we used; on the other hand, the linear probability model only had to account for the likelihood that an article would go viral which meant over 50% which happened significantly less, and thus the predicted values were split between, allowing for an accurate representation of how likely an article would go viral. Further, given the metrics of specificity and precision in these models, the linear probability model outperforms the linear model on both accounts. It predicts fewer false positives and also predicts more truly positive outcomes.

As such, it is clear that creating a threshold first and then regressing is the better way to go about this because we are able to create a binary variable and predict likelihood of that binary variable occurring rather than having to predict the whether the metric of interest will surpass a particular threshold.