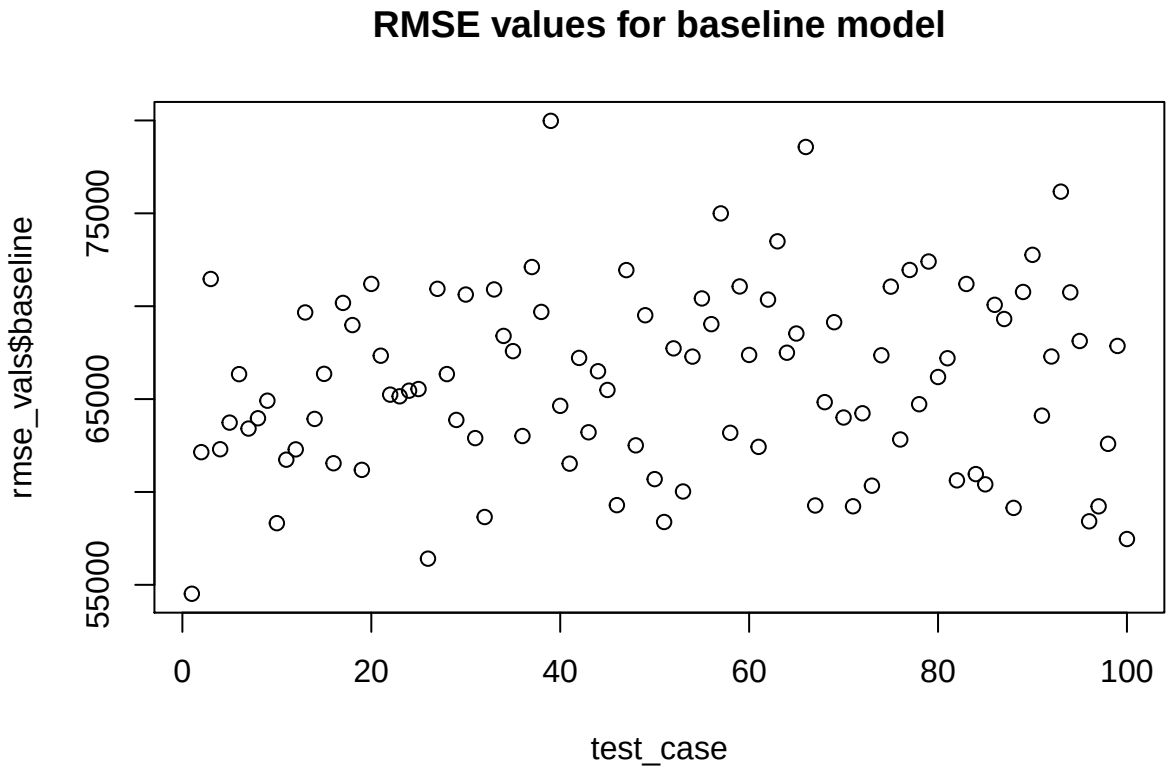
Exercise 2

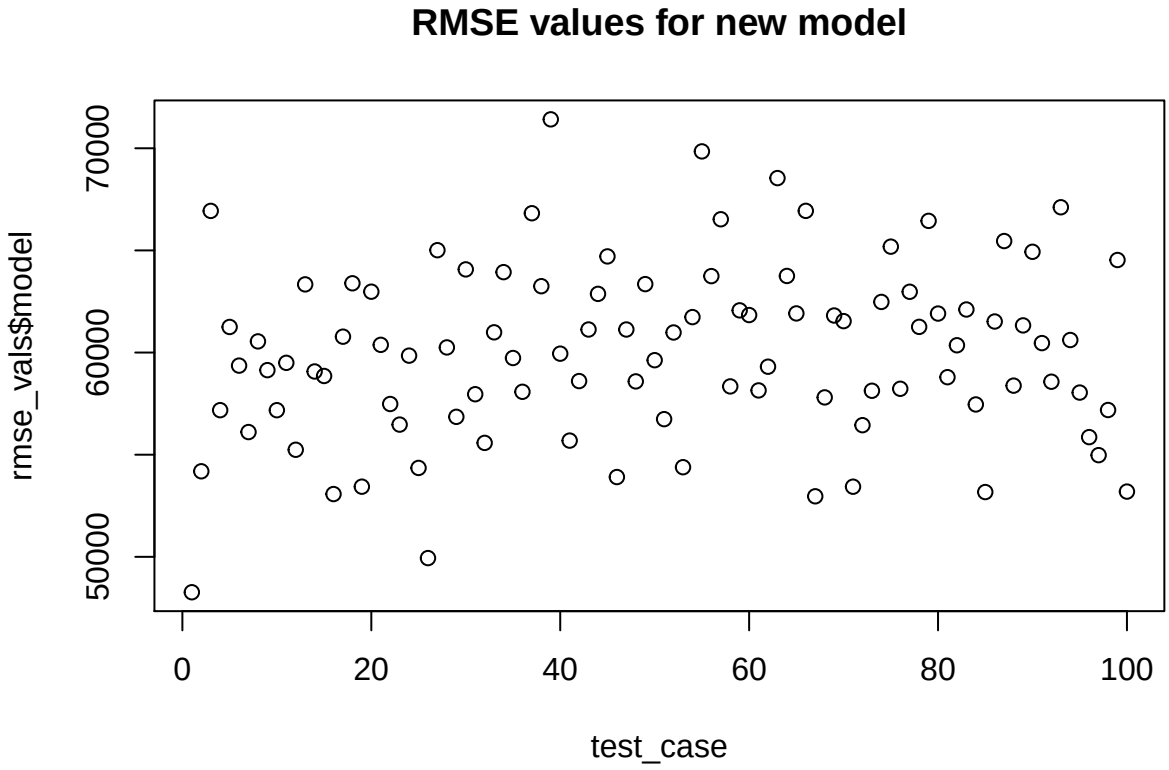
Note: Assistance with code from Linh Nguyen on Q1 and Q3.

Question 1

In order to create a strong pricing model, many factors were considered. Ultimately, I was able to create a better model than the previous baseline. This was specifically possible due to interaction variables of lot size*age, rooms*bedrooms, and rooms*bathrooms. These interactions were especially important, as they combined some of the most important variables to create a more thorough understanding of price. Age certainly could erode price regardless of lot size, and people are interested in the ratio of rooms to bedrooms or bathrooms in order to know how much living space is available. Below are the plots of RMSE for the baseline and the new model, simulated 100 times to create 100 observations.

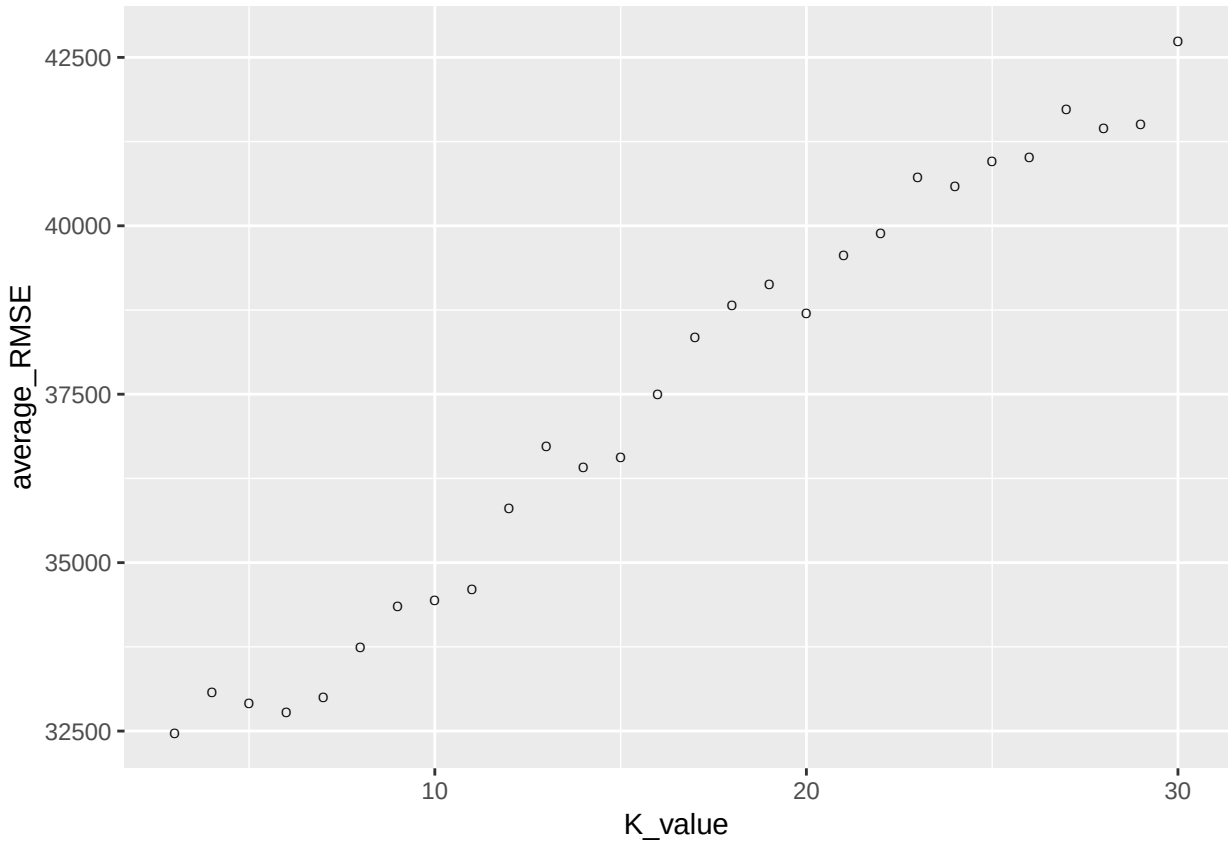



```
## integer(0)
```




```
## integer(0)
```

The baseline model has RMSE that ranges from approximately 60,000 to 75,000, while the new model has RMSE that ranges from approximately 50,000 to 70,000. So, on average, RMSE for the new model is lower. Hence I recommend moving to my new model in order to generate more accurate market values for housing. Furthermore, I was able to create a KNN model that continued to decrease RMSE. This is shown below.



With the KNN regression, RMSE was lowered to approximately 32500 when using a small K value. To conclude, I see strong improvement within the new model, and recommend the company switches to it immediately to improve performance.

Question 2

In order to test how conservative the doctors are, it is crucial that we hold the risk factors constant for each doctor. So, I decided to hold each of the factors in the following states: age between 40-49, breast density at level 3, pre-menopausal, no symptoms, and no family history of breast cancer. I chose these specific states in order to create an environment where it was a feasible for all of the doctors to have the data, as they were all relatively common traits. This fixed nature of the data allowed us to view how conservative each would be in a typical situation. The results are as follows:

```
##
##  0  1
## 15  5
```

```
## [1] 0.25
```

Given the fixed variables, radiologist #13 recalls 25% of patients.

```
##
##  0  1
## 14  2
```

```
## [1] 0.125
```

Given the fixed variables, radiologist #34 recalls 12.5% of patients.

```
##
```

```
##  0  1
## 14  4
```

```
## [1] 0.2222222
```

Given the fixed variables, radiologist #66 recalls 22.22% of patients.

```
##
##  0  1
## 17  4
```

```
## [1] 0.1904762
```

Given the fixed variables, radiologist #89 recalls 19.05% of patients.

```
##
##  0  1
## 14  3
```

```
## [1] 0.1764706
```

Lastly, given the fixed variables, radiologist #95 recalls 17.65% of patients.

We can conclude from these results an ordering of conservativeness among the 5 doctors. From highest to lowest: #13, #66, #89, #95, #34. This is just one variation of many potential samples of fixed variables, but it is a good indicator of general conservativeness of the 5 doctors.

As for the question of if the doctors should be weighing certain clinical factors more heavily than they currently are, the answer is a resounding yes. By using a linear probability model of recall predicting cancer, I was able to deduce that given that the patient is recalled, there is approximately a 14.8% chance of them having cancer, displayed here:

```
## [1] 0.1486486
```

This gave me a baseline for how often a doctor recalled. Next, I created 5 separate linear probability models predicting cancer given recall and one of the following: family history, age, symptoms, menopause testing, and breast density classification. The results are as follows:

```
## [1] 0.1548047
```

Given recall and family history, the odds of the patient having cancer was approximately 15.48%.

```
## [1] 0.1793339
```

Given recall and age, the odds of the patient having cancer was approximately 17.93%.

```
## [1] 0.1602172
```

Given recall and symptoms, the odds of the patient having cancer was approximately 16.02%.

```
## [1] 0.1837585
```

Given recall and menopause results, the odds of the patient having cancer was approximately 18.38%.

```
## [1] 0.2049679
```

Given recall and breast density classification, the odds of the patient having cancer was approximately 20.5%.

Overall, the doctors certainly are not expected to be perfect. However, the results from age, menopause testing, and breast density classification seem troubling, specifically the last two. The doctors should focus on these three areas more when making a decision about recalling a patient.

Question 3

First, lets create a condition for virality in the data, and make a table out of it.

```
##
##     0     1
## 20082 19562
```

Using this table, we can create the frequency at which the null works:

```
nullrate
```

```
## [1] 0.5065584
```

So, we would hope to clear this when creating a model. Now, I will attempt to create two types of models: one which uses a threshold to account for virality and one which directly predicts the viral nature using my variable, "viral". Both models will be linear regressions with very similar predictors. The only difference will be what they are predicting: one predicts "shares" and the other predicts how viral ("viral").

Let us begin with model 1. Here is the confusion matrix:

```
##    yhat
## y      0    1
##   0   56 3933
##   1   17 3923
```

Here is the error rate:

```
## [1] 0.4981713
```

Here is the true positive rate:

```
## [1] 0.9956853
```

Here is the false positive rate:

```
## [1] 0.5006365
```

Let us move on to model 2. Here is the confusion matrix:

```
##    yhat
## y      0    1
##   0 2622 1367
##   1 1385 2555
```

Here is the error rate:

```
## [1] 0.3470803
```

Here is the true positive rate:

```
TPR(matrix2)
```

```
## [1] 0.6484772
```

Here is the false positive rate:

```
## [1] 0.3485467
```

So, it is clear model 2 is superior here, as model 1 is not even better than the null model. In this instance, using the threshold first and regressing second seems dominant. This might be true because it is easier to directly predict something (state of virality) than to do it in a roundabout 1 such as in model 1.