Exercises 2

Stephenson Gokingco, Akash Thakkar, Caroline Hao, James Cornejo

3/13/2020

1) KNN practice

The data in sclass.csv contains data on over 29,000 Mercedes S Class vehicles—essentially every such car in this class that was advertised on the secondary automobile market during 2014. For websites like Cars.com or Truecar that aim to provide market-based pricing information to consumers, the Mercedes S class is a notoriously difficult case. There is a huge range of sub-models that are all labeled "S Class,"" from large luxury sedans to high-performance sports cars. Moreover, individual submodels involve cars with many different features. This extreme diversity—unusual for a single model of car—makes it difficult to provide accurate pricing predictions to consumers.

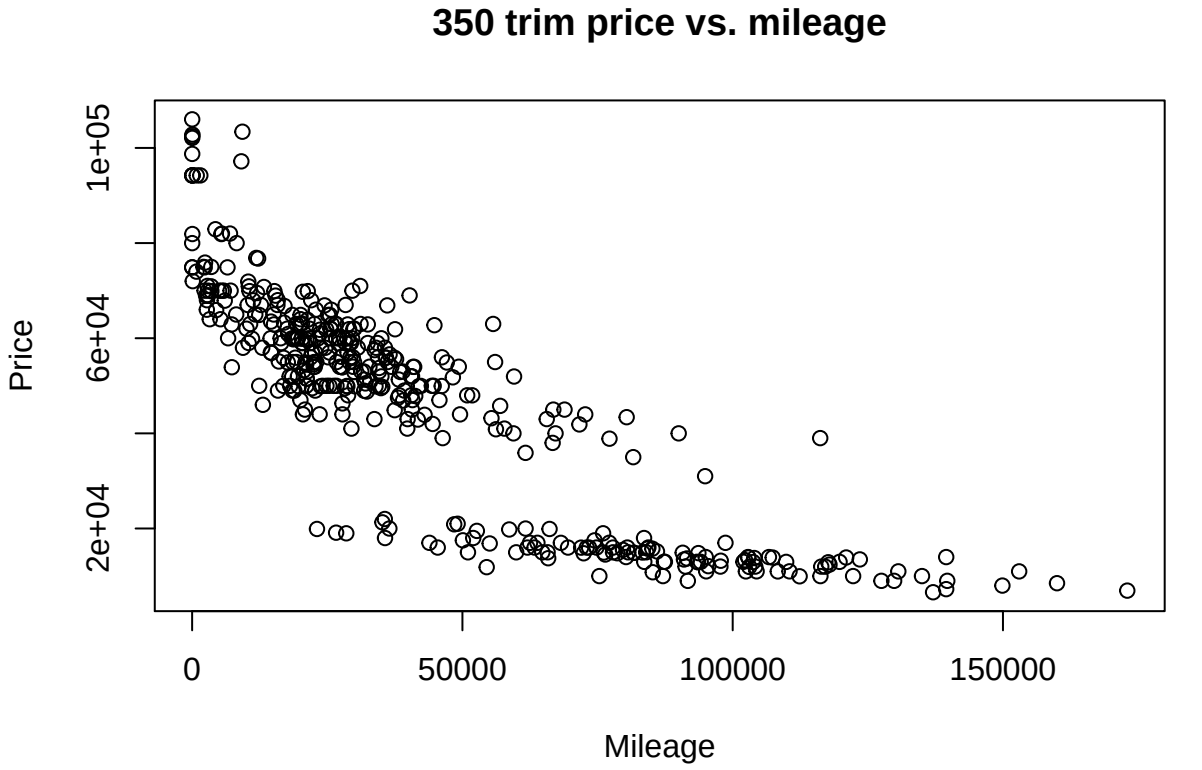
For this report, we will be focusing on three variables in particular: trim (categorical variable for car's trim level, e.g. 350, 63 AMG, etc. The trim is like a sub-model designation.), mileage (mileage on the car), and price (the sales price in dollars of the car).

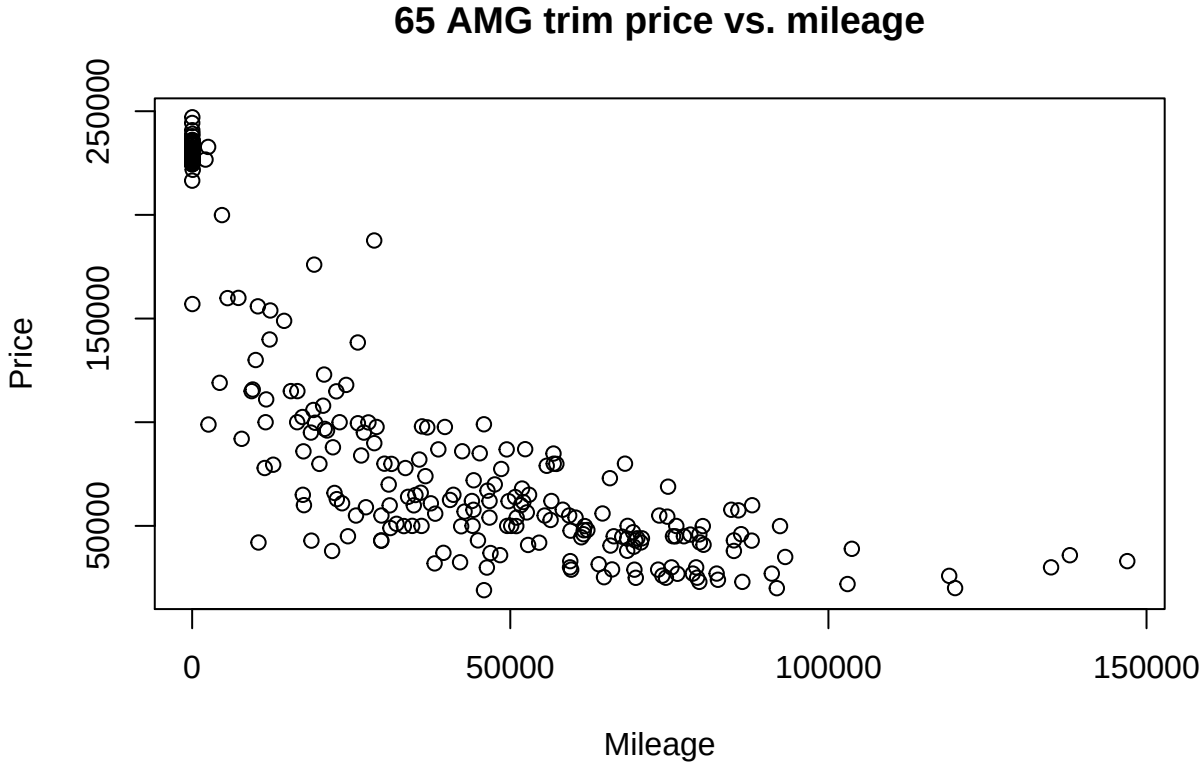
We will use K-nearest neighbors to build a predictive model for price, given mileage, separately for each of two trim levels: 350 and 65 AMG. That is, we'll be treating the 350's and the 65 AMG's as two separate data sets.

First, let's split the 350 and 65 AMG's into two separate datasets:

```
sclass350 = subset(sclass, trim == '350')
sclass65AMG = subset(sclass, trim == '65 AMG')
```

Next, let's just take a look at the price vs. mileage for each trim level. This will give us a better picture of the high-level differences between the two datasets.



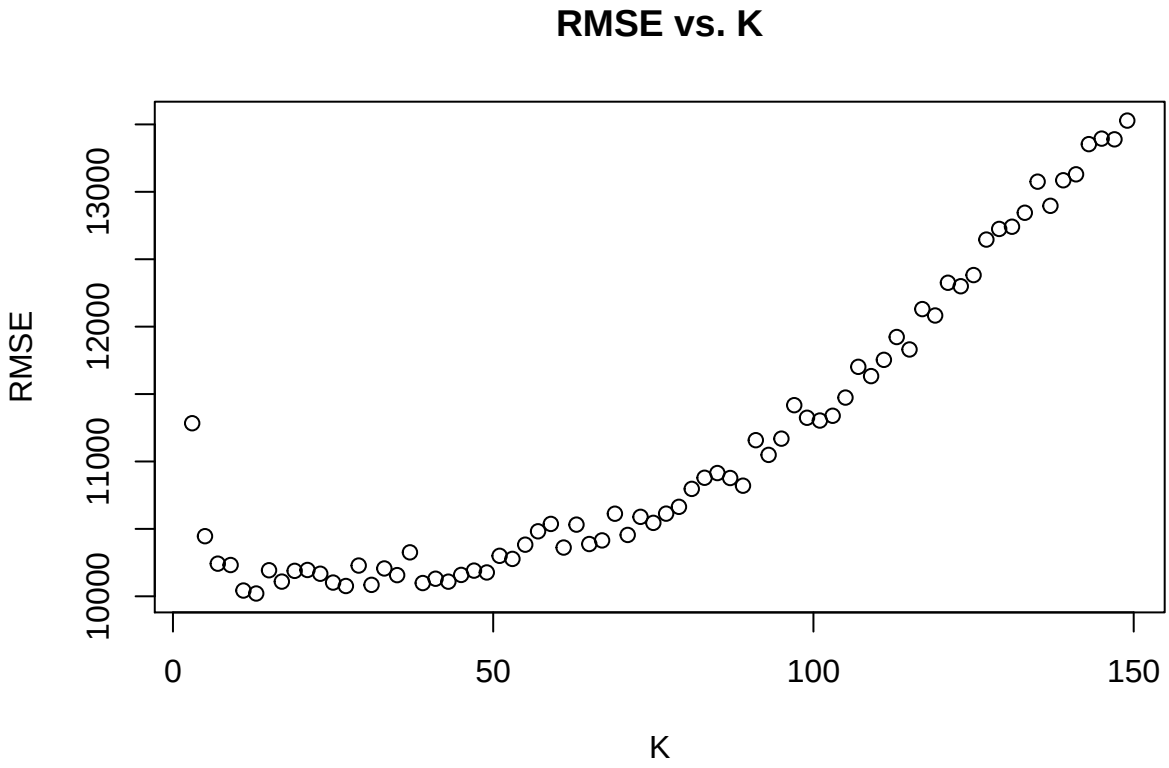


Let's start the actual KNN. First, we will be splitting the data into a training and testing set for each of the two trim levels. We'll start with the 350.

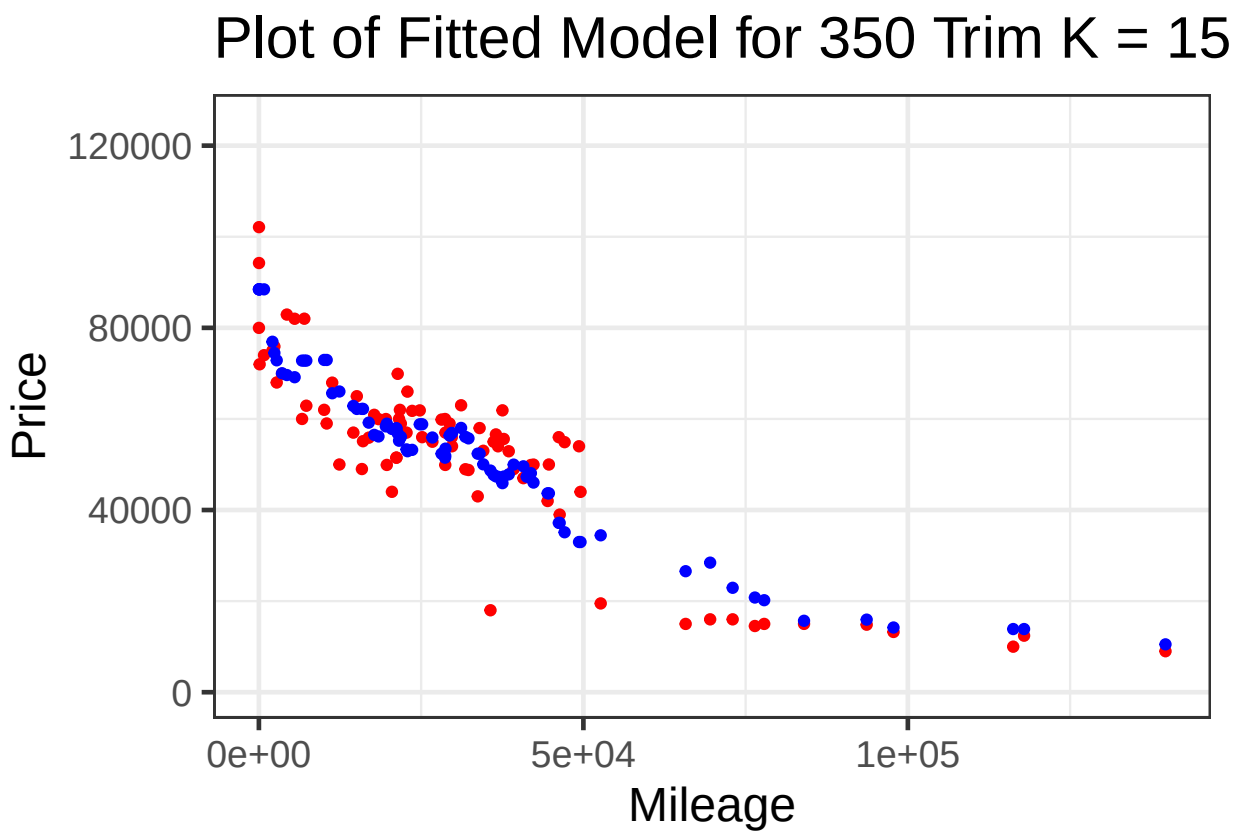
```
# Train-test split
N = nrow(sclass350)
N_train = floor(0.8*N)
```

```
N_test = N - N_train
# Randomly sample a set of data points to include in the training set
train_ind = sample.int(N, N_train, replace=FALSE)
# Define the training and testing set
D_train = sclass350[train_ind,]
D_test = sclass350[-train_ind,]
# Now, separate the training and testing sets into features (X) and outcome (y)
X_train = select(D_train, mileage)
y_train = select(D_train, price)
X_test = select(D_test, mileage)
y_test = select(D_test, price)
```

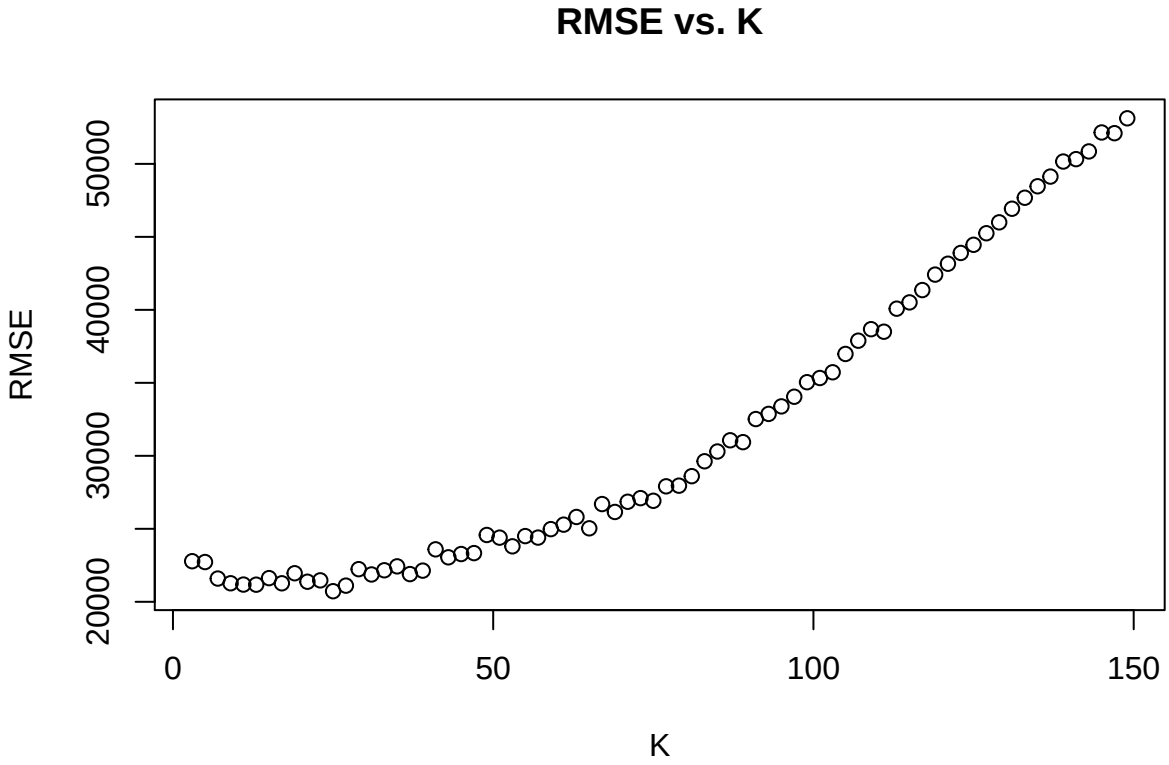
We then run K-nearest-neighbors for many different values of K, starting at K=2 and going as high as we need to. For each value of K, we fit the model to the training set and make predictions on our test set. We calculate the out-of-sample root mean-squared error (RMSE) for each value of K and made a plot of RMSE vs. K.



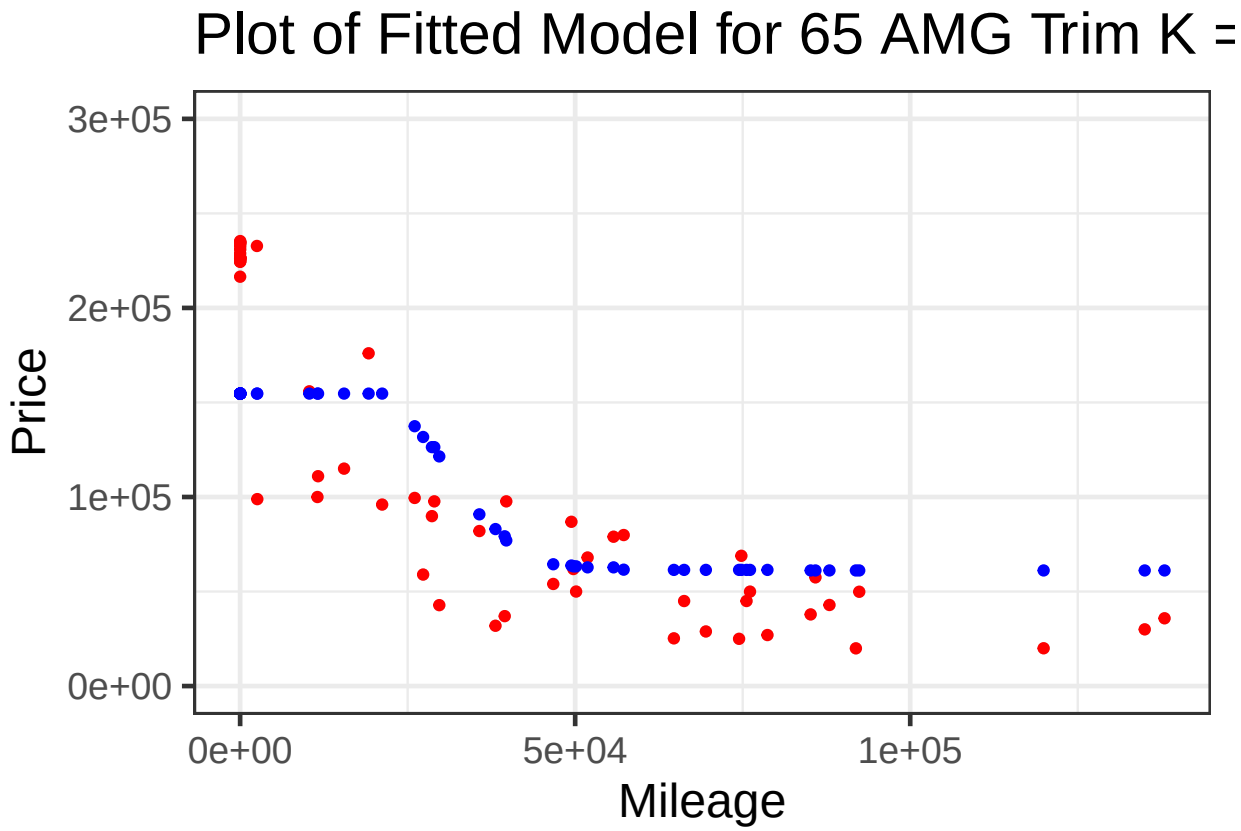
Based on the RMSE vs. K plot, it seems that the optimal value of K is 15. Let's show a plot of the fitted model using K = 15.



Following a very similar procedure with the 65 AMG trim, we get the following RMSE vs. K plot:



For the 65 AMG trim, the optimal value of K is 19. Let's see what the plot of the fitted model looks like for K = 19.



Based on our findings, it seems that the 65 AMG trim yields a slightly larger optimal value of K than the 350 trim ($K = 19 > K = 15$). If we take a look at the first scatterplots of price vs. mileage for both trim levels, we can see that for the 350 trim, there's areas of high-density, low variation in terms of where the points are located. For the 65 AMG trim, we don't see much density of points. In other words, the 65 AMG plot "needs" the higher K to account for the higher variation of values and minimize RMSE.

2) Predicting House Prices in Saratoga, New York

Using data on the prices and characteristics of houses in Saratoga, New York, we seek to create a predictive model of house prices in the city. It is imporant to note that in our model, we will be excluding land value (which is an observed house characteristic in the data). This is because land value likely would bring endogenieity issues in the model.

For example, suppose there is an area with expensive houses and an area with non-expensive houses. The area with expensive houses would generate greater property tax revenue. Tax revenue from property taxes help fund public amenities and services such as schools and education. Thus, schools in the expensive housing area would likely receive more funding than schools in the non-expensive housing area. Since school funding is likely a strong positive correlate with school quality, the schools in the expensive housing areas will likely be better. Having better schools in an area would then likely increase the land value in that area. Hence, to avoid these issues of endogeneity, we will not include land value as a variable in our model. Asides from land value, the rest of the house characteristic variables will be included in the model.

To begin creating the model, we will clear the working environment and load the SaratogaHouse data and the needed libraries. We will also generate or recode dummy variables for the existing categorical variables in the data.

To determine the exact model specifications, we will use forward and step-wise selection. We will use the

higher performing of the both selection methods. To begin the step-wise selection process, we need to start with a base model. We will use the following:

```
lm_medium = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
               fireplaces + bathrooms + rooms + heating_hotair + heating_steam + fuel_oil + fuel_gas +
                 centralAir, data=SaratogaHouses)
```

The step-wise and forward selection models both will likely include unintuitive interactions among the house characteristic variables. However, because our objective is model improvement from this baseline, we are willing to sacrifice interpretability to improve performance. To test if the selection models do make improvements from the base model and to determine which selection model is better, we will split the data into a training and test set. After fitting the three models (baseline, forward selection, and step-wise selection) to the training data, we will predict the testing data and calculate the out-of-sample root mean squared error (RMSE). To reduce the effect of the random sampling, we will iterate this process 250 times.

The below table averages the RMSE values for each model for the 250 iterations. Note, V1 refers to the base model from above, V2 refers to the forward selection model, and V3 refers to the step-wise selection model. From the table, we conclude that the step-wise selection model is the highest performing of the three options.

```
##       V1       V2       V3
## 66862.78 64387.55 62531.57
```

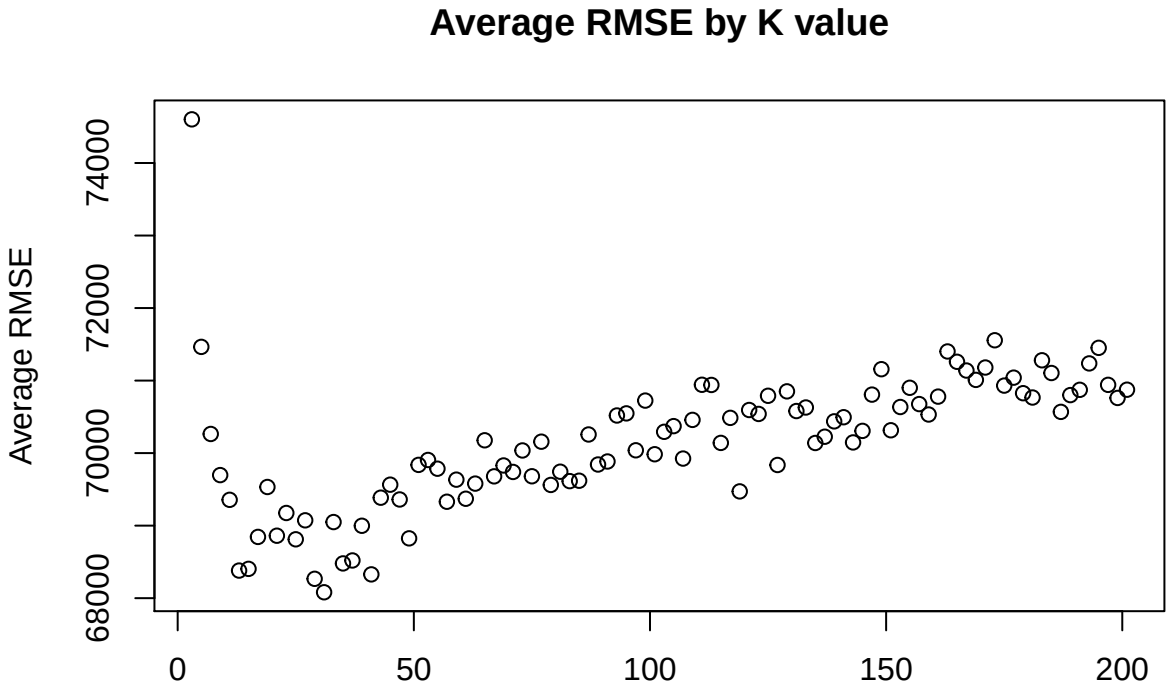
The following shows the different combinations of feature variables used in the step-wise model, as well as their regression coefficients. As stated above, this model is not intuitively interpretable. However, our primary objective is prediction and thus we are willing to sacrifice readability for model improvement. For example, the largest regression coeffient is for the waterfront indicator variable and the largest interaction regression coefficient is for the lotsize and waterfront interaction. However, this is difficult to interpret because the interaction variable more than offsets the positive effect of being a waterfront property.

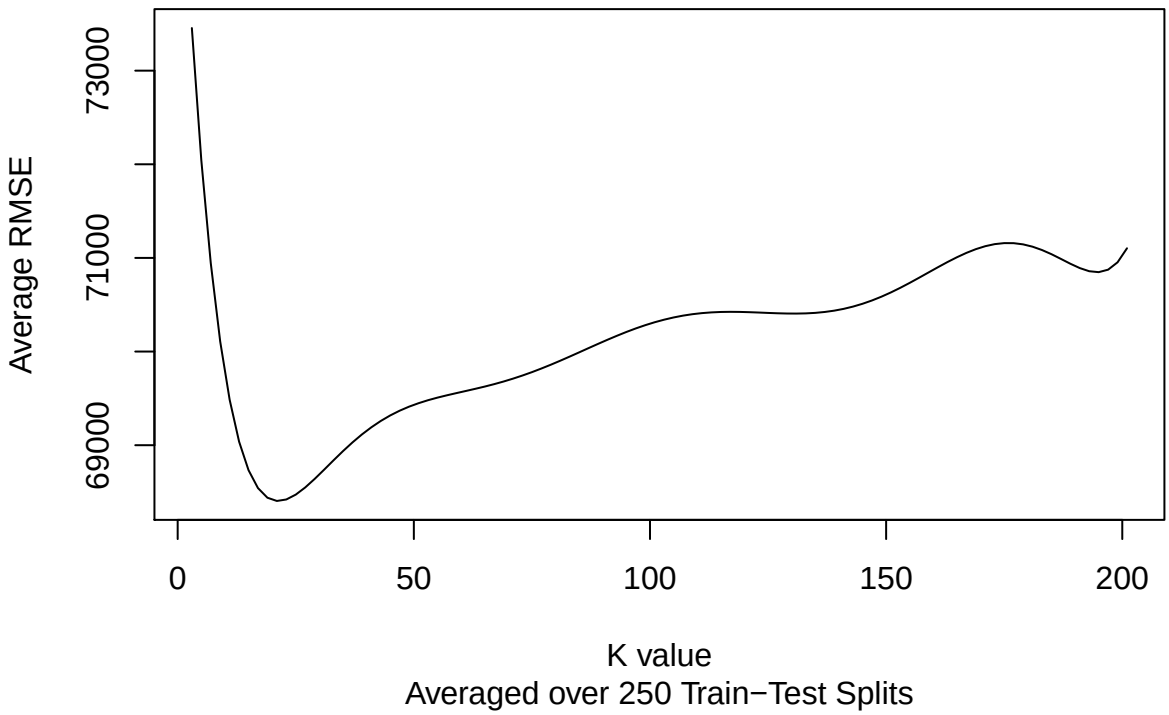
```
## lm(formula = price ~ lotSize + age + livingArea + pctCollege +
##     bedrooms + fireplaces + bathrooms + rooms + heating_hotair +
##     heating_steam + fuel_oil + fuel_gas + centralAir + waterfront +
##     newConstruction + sewer_pubcom + livingArea:centralAir +
##     age:pctCollege + bathrooms:heating_hotair + livingArea:fuel_oil +
##     pctCollege:fireplaces + livingArea:fireplaces + bedrooms:fireplaces +
##     lotSize:waterfront + fireplaces:waterfront + age:heating_steam +
##     fuel_oil:centralAir + age:centralAir + bathrooms:waterfront +
##     pctCollege:fuel_gas + rooms:heating_hotair + bedrooms:fuel_gas +
##     pctCollege:bathrooms + centralAir:newConstruction + age:sewer_pubcom +
##     heating_hotair:sewer_pubcom + livingArea:rooms + lotSize:fireplaces +
##     bedrooms:sewer_pubcom + bedrooms:fuel_oil + fuel_oil:waterfront +
##     fuel_oil:sewer_pubcom + fireplaces:fuel_gas + bathrooms:heating_steam,
##     data = SaratogaHouses)
```

```
##                (Intercept)                    lotSize
##               9.992777e+04               1.568830e+04
##                        age                  livingArea
##              -2.035139e+03               5.275222e+01
##                 pctCollege                    bedrooms
##              -1.452864e+03               2.222435e+04
##                 fireplaces                   bathrooms
##               9.254184e+04              -3.734018e+04
##                      rooms              heating_hotair
##              -3.187453e+03              -7.985037e+04
##              heating_steam                    fuel_oil
##              -4.070324e+04               1.258795e+05
##                   fuel_gas                  centralAir
```

```
##               3.710314e+04              -2.149834e+04
##                 waterfront           newConstruction
##               1.378903e+05              -1.155567e+04
##               sewer_pubcom     livingArea:centralAir
##               1.726395e+04               2.721738e+01
##            age:pctCollege  bathrooms:heating_hotair
##               3.331562e+01               2.322832e+04
##       livingArea:fuel_oil     pctCollege:fireplaces
##              -4.419572e+01              -1.294826e+03
##     livingArea:fireplaces       bedrooms:fireplaces
##               1.921132e+01              -1.348635e+04
##        lotSize:waterfront     fireplaces:waterfront
##              -1.952458e+05               9.393521e+04
##         age:heating_steam       fuel_oil:centralAir
##               2.646387e+02               3.626339e+04
##            age:centralAir      bathrooms:waterfront
##              -7.915473e+02               4.927191e+04
##       pctCollege:fuel_gas      rooms:heating_hotair
##               6.515554e+02               4.733448e+03
##         bedrooms:fuel_gas      pctCollege:bathrooms
##              -1.801361e+04               7.218728e+02
##  centralAir:newConstruction          age:sewer_pubcom
##              -3.090325e+04               4.332923e+02
## heating_hotair:sewer_pubcom          livingArea:rooms
##               1.697266e+04               1.716436e+00
##        lotSize:fireplaces     bedrooms:sewer_pubcom
##              -7.652053e+03              -1.101925e+04
##         bedrooms:fuel_oil       fuel_oil:waterfront
##              -1.923210e+04              -5.643228e+04
##     fuel_oil:sewer_pubcom       fireplaces:fuel_gas
##               1.855462e+04              -1.109002e+04
##    bathrooms:heating_steam
##               1.558160e+04
```

We will now consider using a K-nearest-neighbors model to see if it can further improve our prediction performance. To motivate our choice in K, we will loop over different K values to see what range of K values have the lowest average RMSE values for 250 iterations of training and testing splits. We will first plot the scatter plot of these averages and then plot a line of best fit of an eighth degree polynomial function. The purpose of this high degree polynomial function is to flexibly show the general relationship between the average RMSE and the K values.





Based on the figures above, K=21 is a reasonable choice that has a low RMSE. Following this, we will compare the average RMSE value based on the 250 iteractions with the RMSE values presented previously. The below table shows these averages. Based on the table, the KNN model performs worse than the three prior models.

```
##  median model.V1 forward model.V2    step model.V3   knn21 model.V4
##         66862.78         64387.55         62531.57         68519.94
```

Based on our analysis, we conclude that the step-wise selected model is the best approach to predicting house prices. It improved upon the baseline model and outperformed the forward selection model. The K nearest neighbors model with a K value of 21 performed worse than the baseline model as well as the selection-based models. It is important to note that the selection models do suffer from low interpretability. However, they significantly improve model performance. In the context of forming predicted market values for properties and calculating their respective property tax liabilities, model performance is of highest importance.