Exercise 2

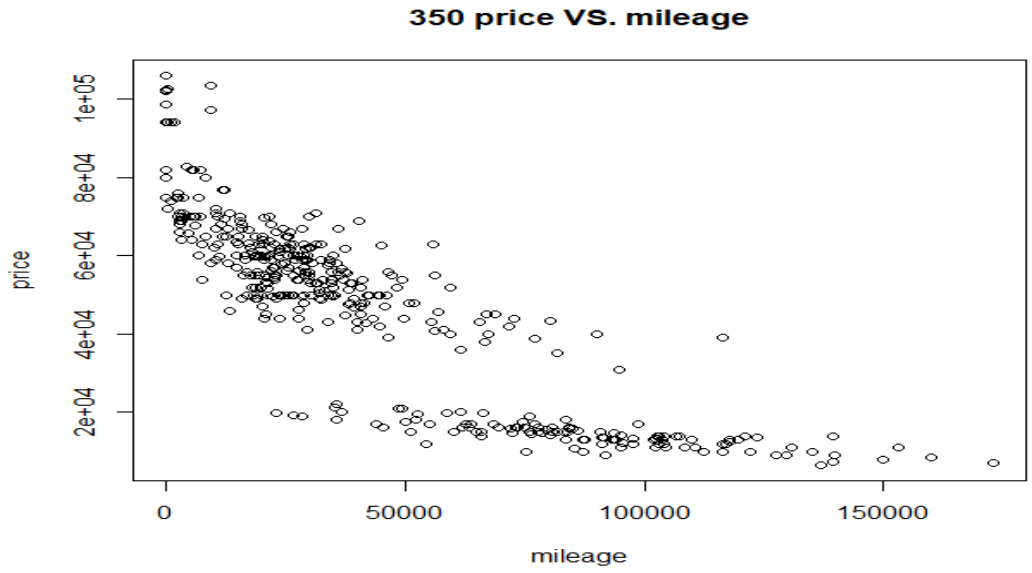
Jaeho Jang (jj36386)

1) KNN Practice

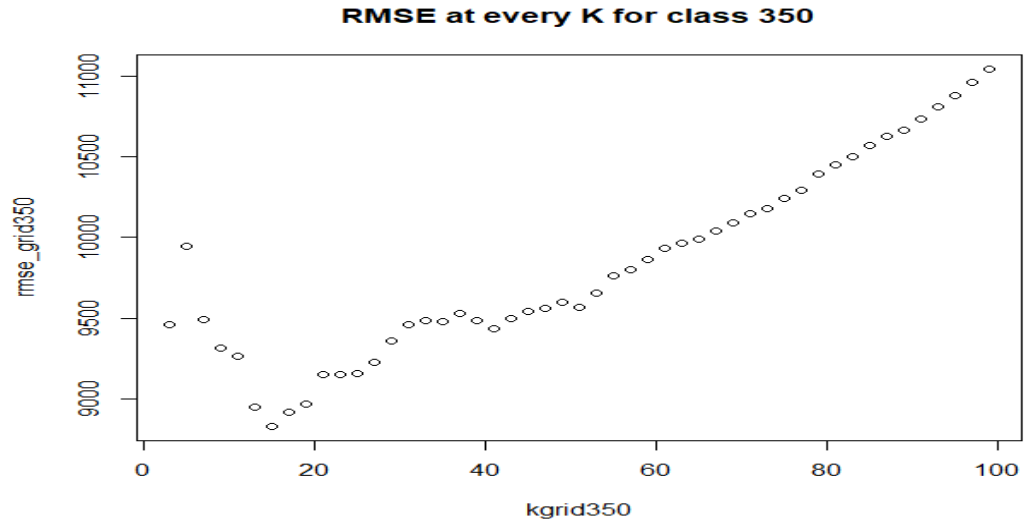
With the data provided regarding 29,000 Mercedes S Class vehicles, our goal is to construct a predictive model, given mileage, for two trim levels: 350 and 65 AMG. After creating a subset for the two levels, according models can be constructed divided into 2 segments.

- The 350 Class

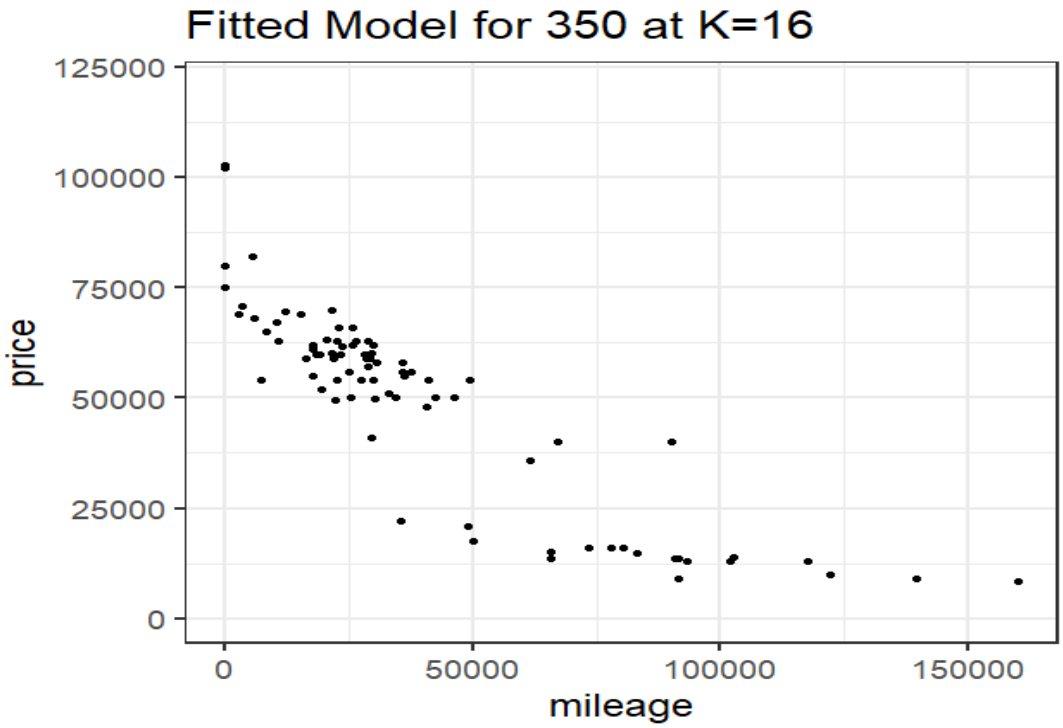
First, we can discover the relationship between the price and mileage easily by simply constructing a plot.



It seems to be that . After, We can derive the KNN by splitting data into training and testing sets. By deriving the KNN, we can accordingly calculate the RMSE for each K value and construct a plot.

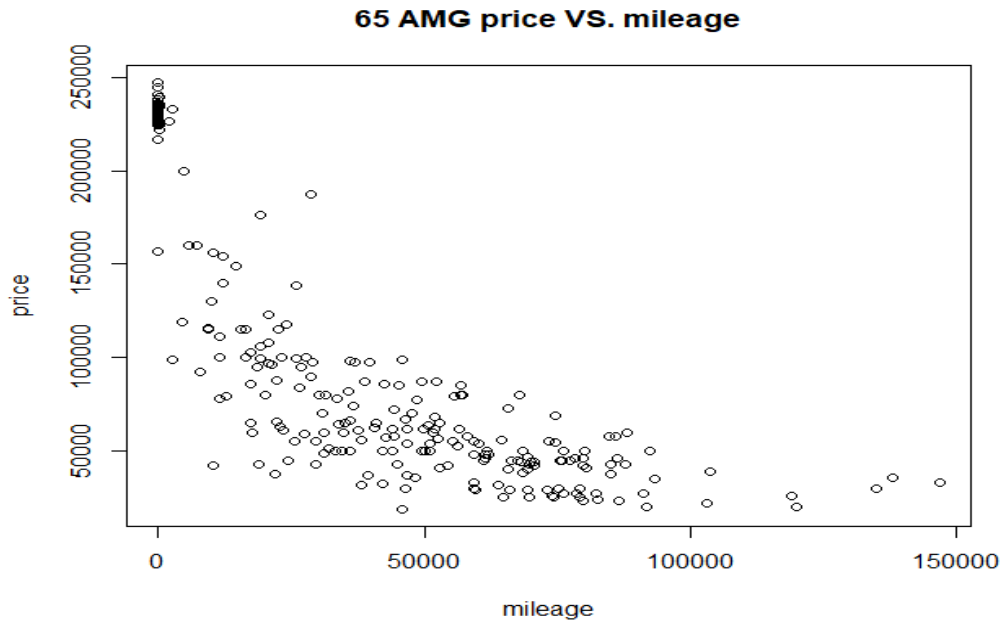


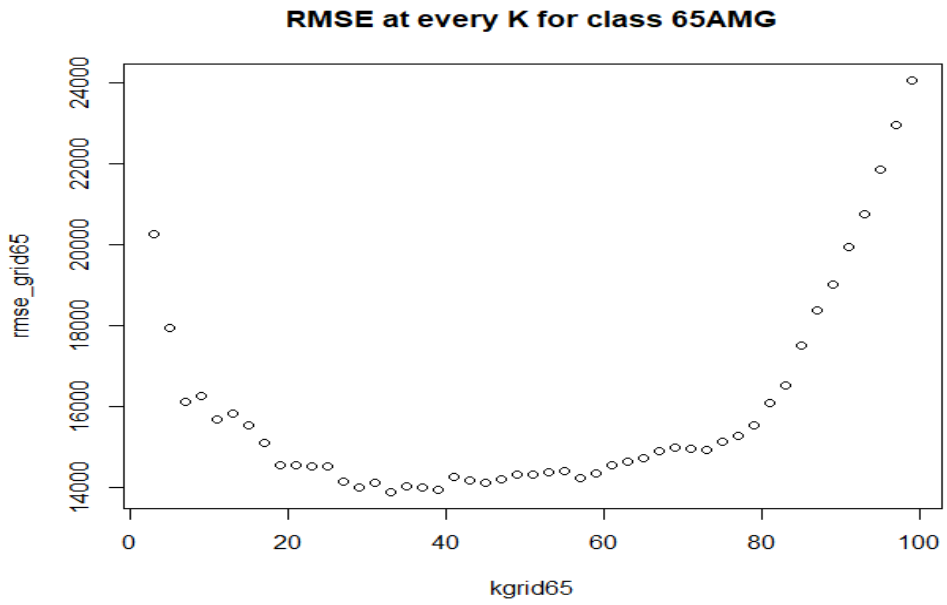
According to the plot, our optimal K value – the x value of lowest point on the plot – is 16. From the K value of 16, we can construct a fitted price vs. mileage model to compare it to the original plot.

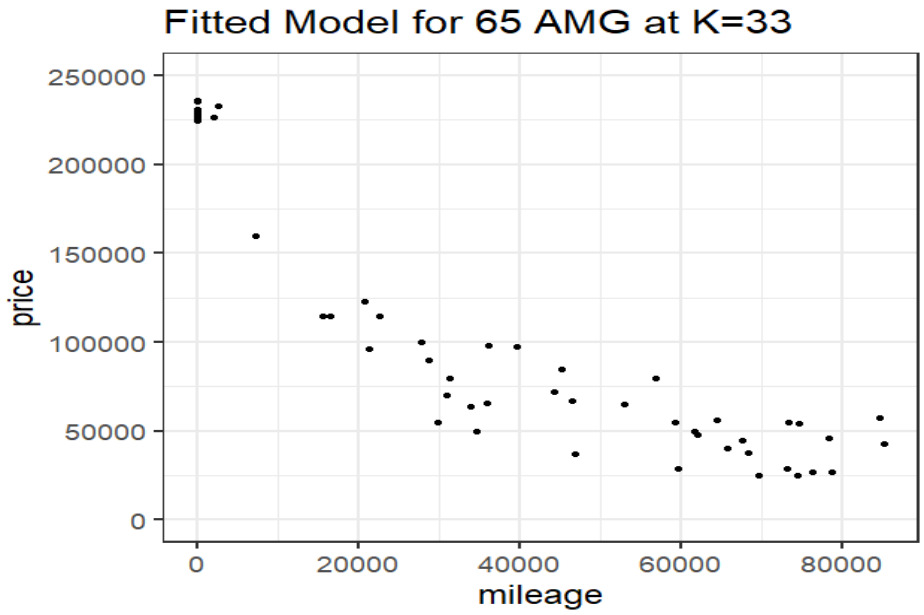


Consequently, repeat the same process for the 65 AMG Class.

- 65 AMG Class







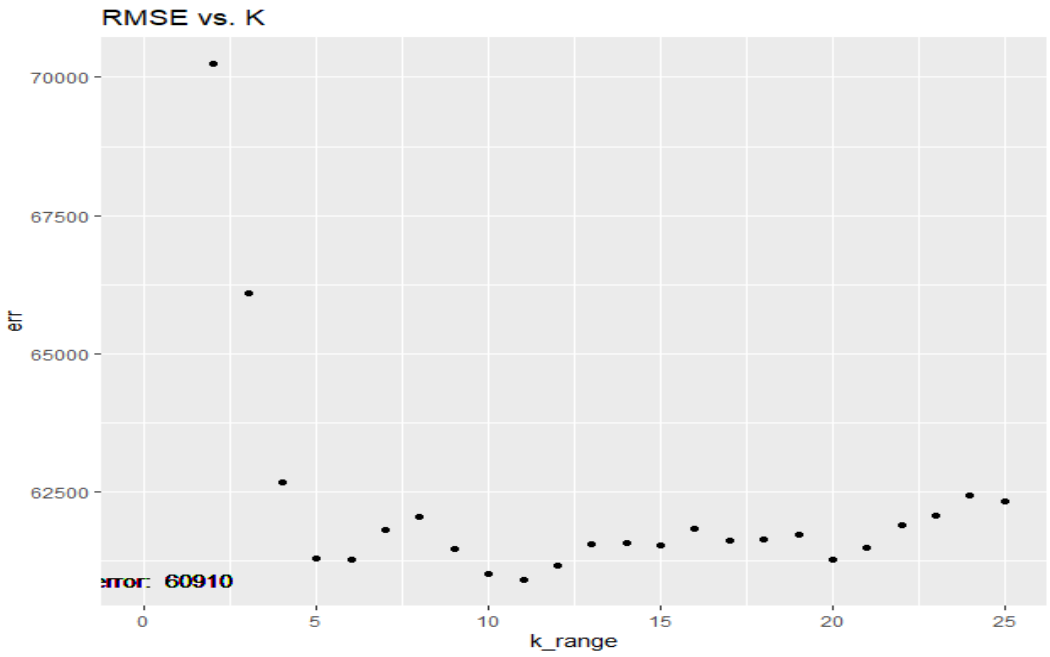
In conclusion, by looking at the price vs. mileage plots, it can be told that both 350 class and 65 AMG class maintain a higher price when the mileage is low - highly concentrated with low variation. According to the RMSE vs. K plots, it looks like 65 AMG yields a higher optimal value of K than that of the 350 class. In conclusion, this means that the 65 AMG class must have a higher K value to minimize RMSE, and therefore, deal with prices that vary more greatly.

2) Saratoga Houses

Assessing the existing KNN model, we've come to name it the "medium" model as it does not perform efficiently enough to predict market value of houses in Saratoga. Accordingly, we have come to generate a more detailed plot that covers a broader and accurate image of the relationship of house characteristics in Saratoga and their price. In fact, we've come to include all the features in the data set to cover any blind spots that the "medium model" did not cover. When the data analysis was performed, our new model seemed to perform much better with the average error of 60152.08, which was much lower that that of the medium model, 65551.57.

V1	V2
65551.57	60152.08

After we came to construct a more efficient linear model, we realized it was worth to construct a KNN model and see if it was any more efficient.



However, the optimal K value existing at K = 10, the average error appeared to be higher than that of the new linear model (60910 > 60152.08). This may be because price is more sensitive to some characteristics than other characteristics. In conclusion, it is recommended that the city utilizes the either of the new model instead of the existing "medium" model.

3) Online News

As there are two different approaches to construct models – regression (regress first and threshold later) and classification (threshold first and regress later) – I decided to construct two different models from each approach to compare and decide which was to predict the virality of articles better.

- Regression; Linear Modeling

```
"Confusion Matrix:"

"11.87   28.38"

"10.48   29.27"

"Overall Error Rate: 0.48575"

"True Positive Rate: 0.737047857285925"

"False Positive Rate: 0.704524891555638"
```

- Classification; Logistic Modeling

```
"Confusion Matrix: "

"37.29   3.2"

"2.08   37.43"

"Overall Error Rate: 0.066"

"True Positive Rate: 0.947334701347267"

"False Positive Rate: 0.0785610450027036"
```

As shown from data above, the logistic model shows a much better result; error rate more 40% lower, greater true positive rate, and lower false positive rate. In conclusion, this means that it is better to regress first and threshold after, as when models are regressed first, variables - such as ‘average token length’ – that do not account to virality of the article compared to other variables may be put into account and bias the regression.