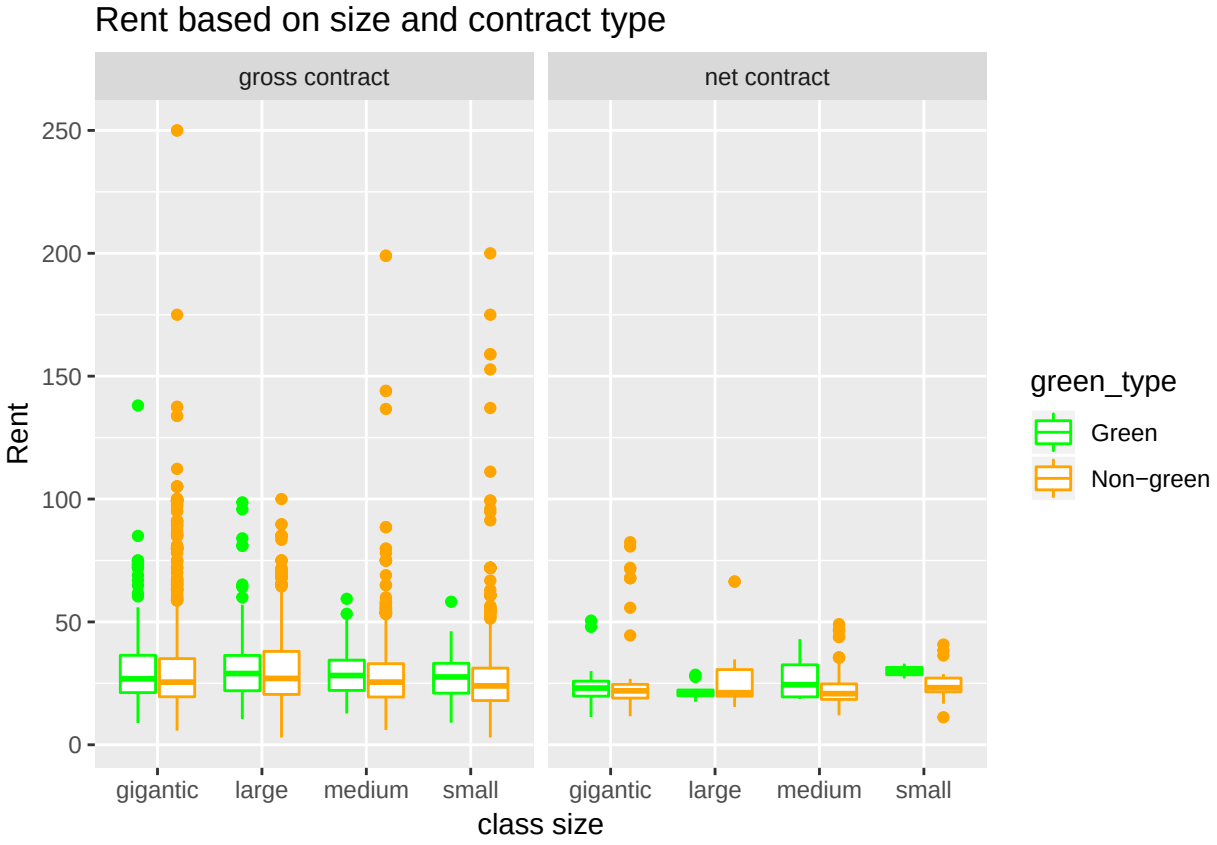
Homework 1

Question 1

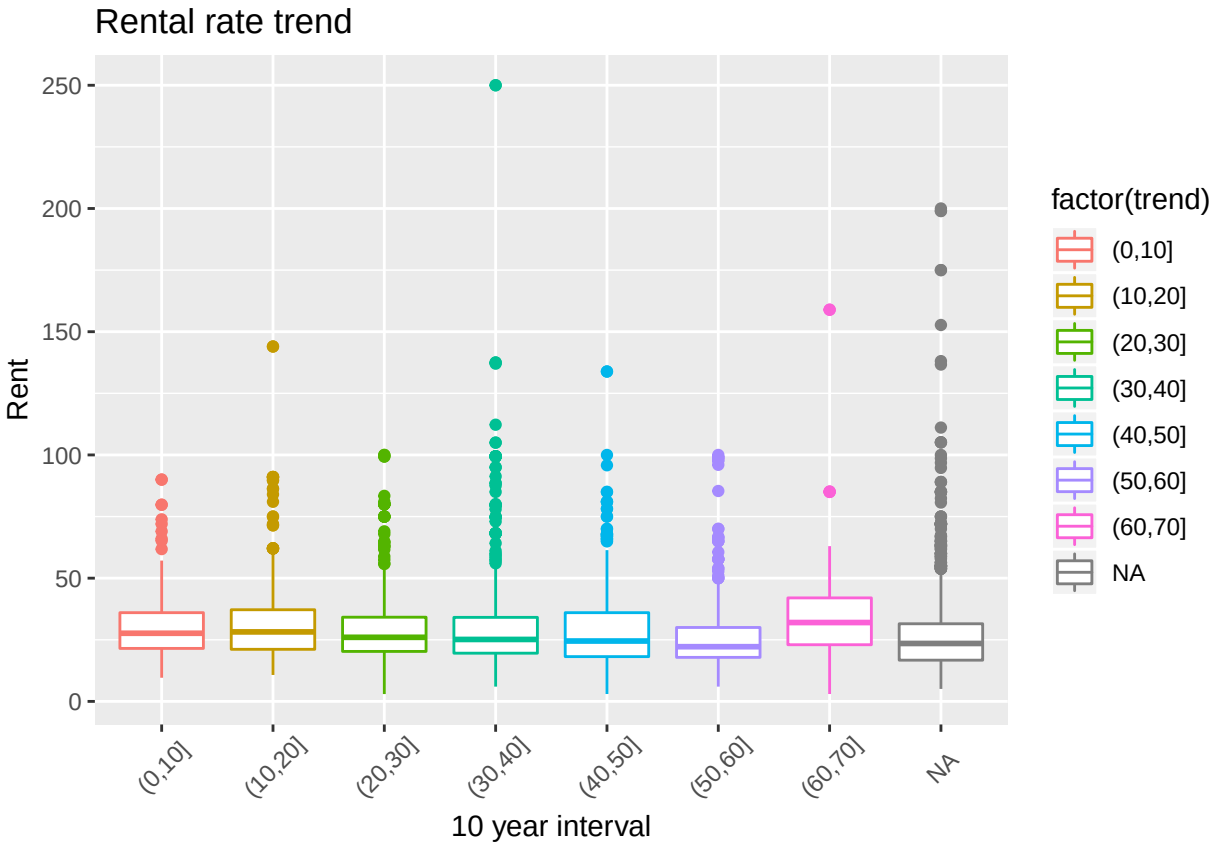
I do not agree with the analysis given. If we look at the distribution of rent according to the size of the properties, we observe that the median rate may not vary that much. However, if we seperate the rent based on the type of contract, either gross contract or net contract, there will be discrepancies across median rental rate based on property size. I will classify the size of property based on four quartiles based on the collecting data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1624   50891  128838  234638  294212 3781045
```



As you can observe from the box plot, the medians rental rate between green versus non-green property for large properties ($128838 < x < 294212$ $feet^2$) do not differ much in net contracts but vary greatly in gross contracts, with green building enjoys around $3.20 in addition. Hence, the rough estimate of difference in rental rate between green and non-green building without factoring in contract type is questionable.

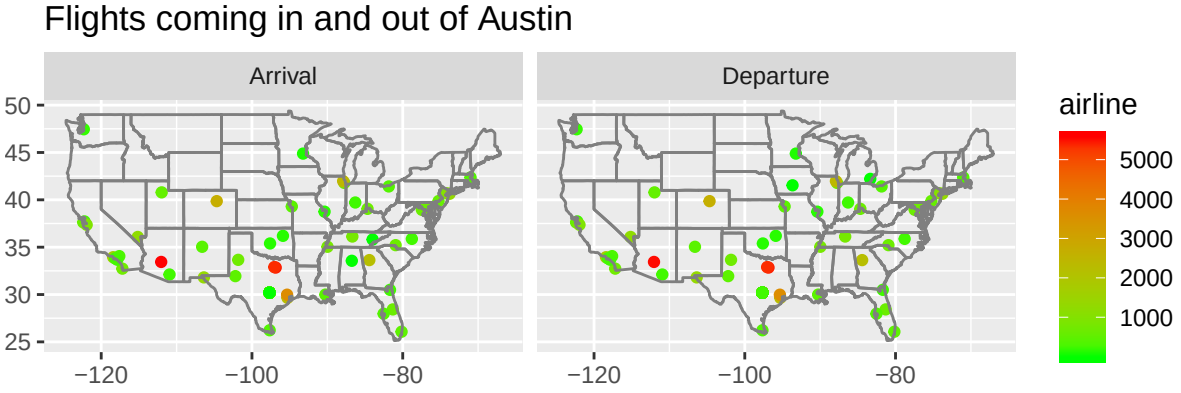
The next questionable assumption is the rent will be constant throughout the life of the building. I will look into the median rent for an interval of every 10 years



Hence, forecasting the rent to stay constant throughout building's lifetime is irrelevant. Therefore, the revenue stream will not be constant, making the profit prediction after cost repcuperation is dubious.

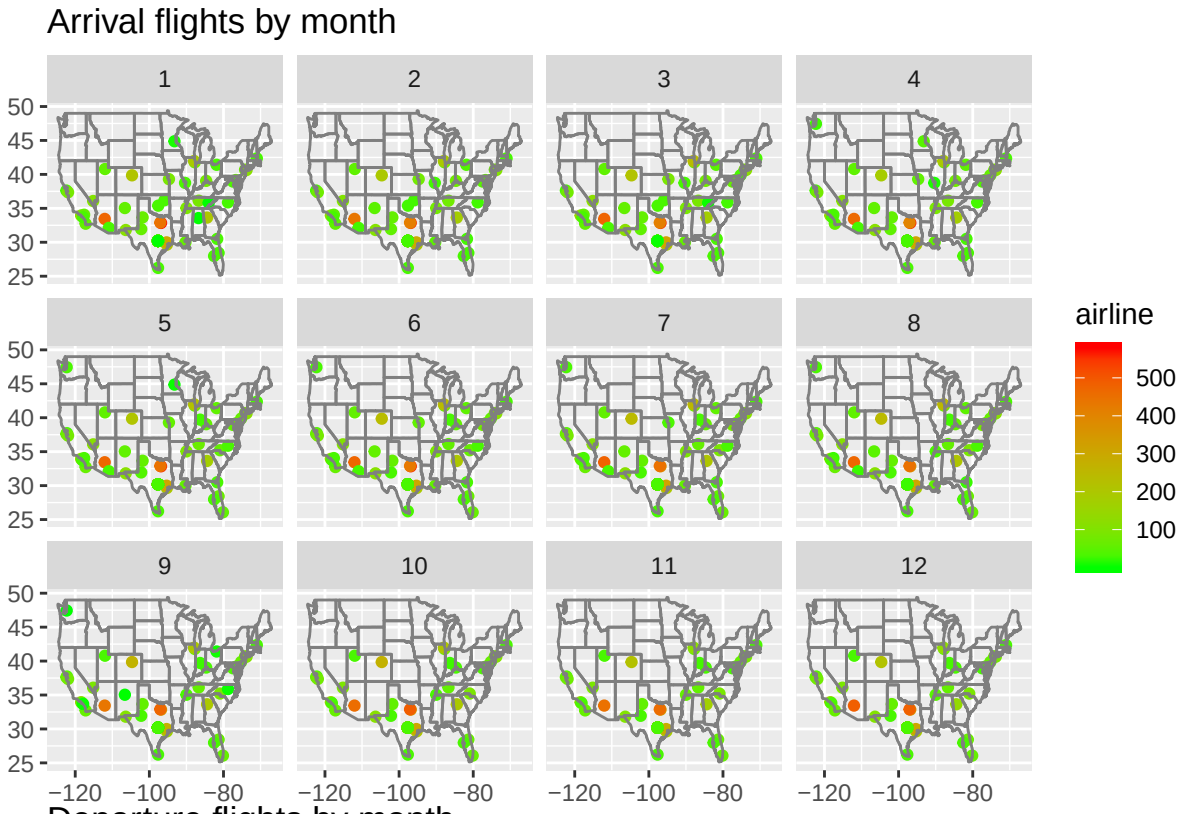
Question 2

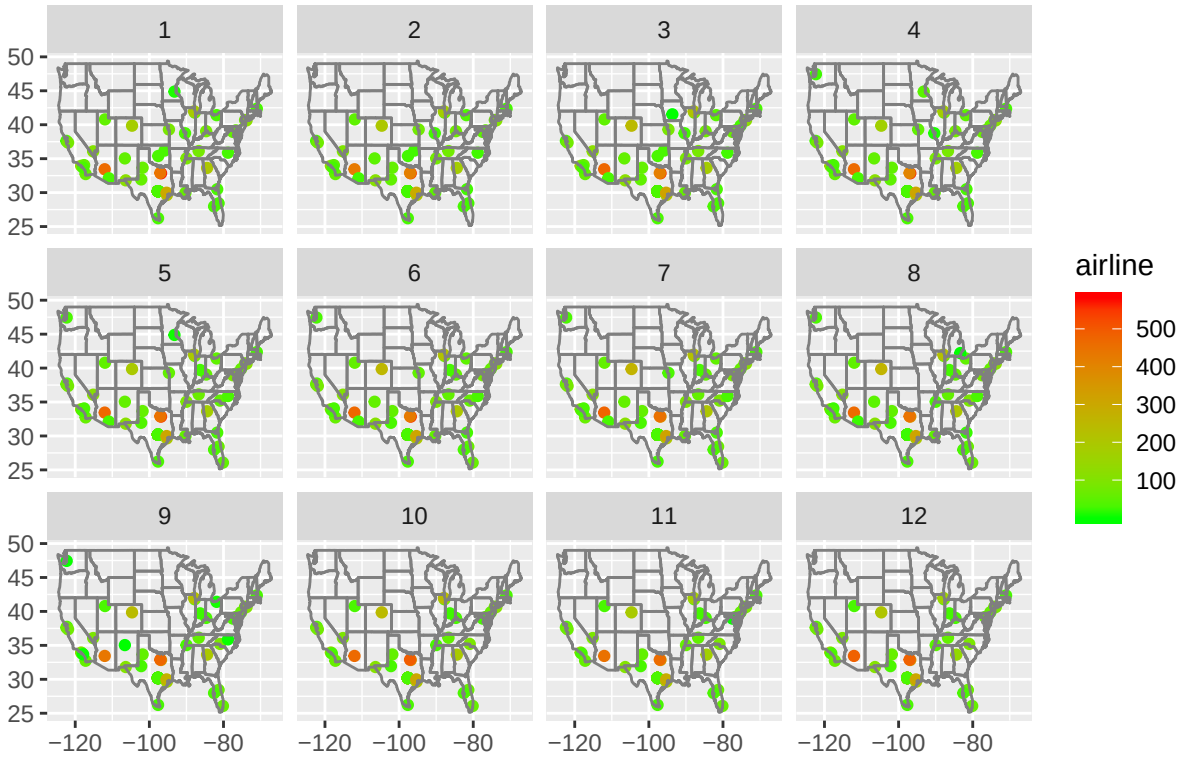
We can look at the volume of flights in and out of Austin to see how the patterns of flights differed state by state



As we observe most of the flights to Austin are coming from Dallas/Fort Worth International, Geogre Bush Intercontinental, Phoenix Sky Harbor and Denver International Airport. Similarly, the departure flights from Austin are coming to the above 5 airports.

To see how the flight pattern changes over the course of the year, let's look at each month in the flight maps of total number of flights arrived to Austin.

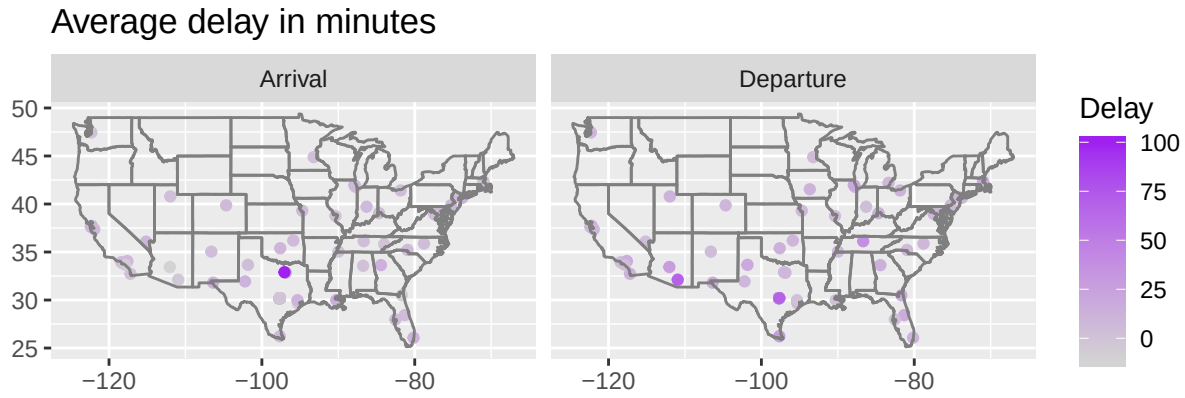




As we observe, the pattern stays constant throughout the year. The top 5 busiest airports that have arrival and departure flights Austin remain the 5 aforementioned airports.

Besides the traffic of flights, we can investigate how badly the delays of airlines occured based on arrival and

departure destinations.

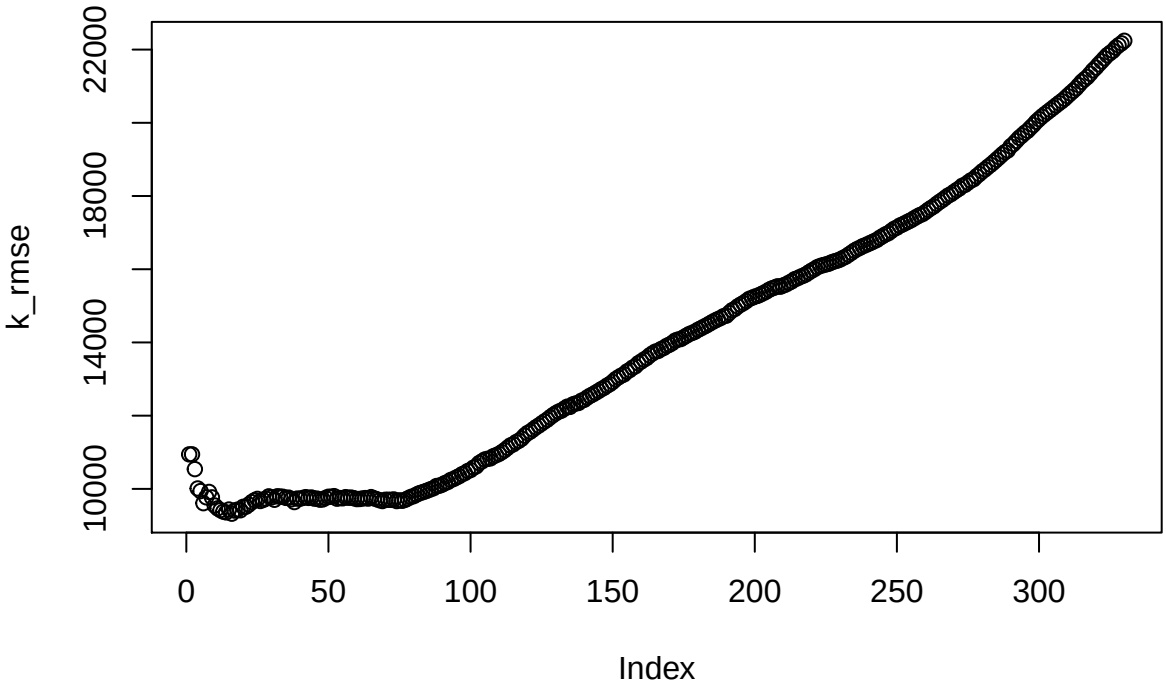


From the data, we obtain that the worst airport to fly to is DFW, as the average delay is highest amongst all other airports and the volume of flights is constantly high throughout the year, both for arrival and departure flights.

Question 3

Following the three steps required for each trim: 1. Split the data into a training and a testing set. 2. Run K-nearest-neighbors, for many different values of K, starting at K=2 and going as high as you need to. For each value of K, fit the model to the training set and make predictions on your test set. 3. Calculate the out-of-sample root mean-squared error (RMSE) for each value of K.

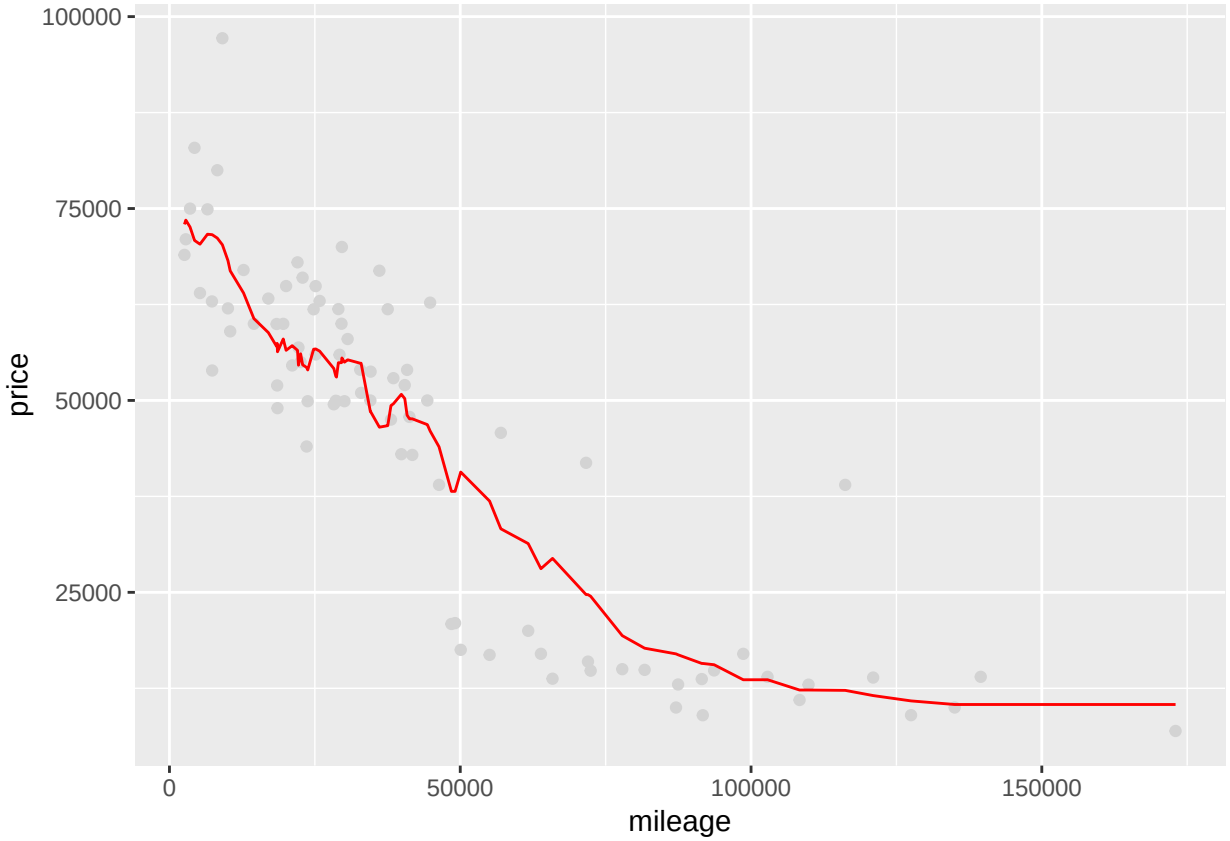
The graph of root mean squared error for prediction with various level of k is:



Hence, the optimal K from running repeated knn regression from k = 2 to maximum possible k is

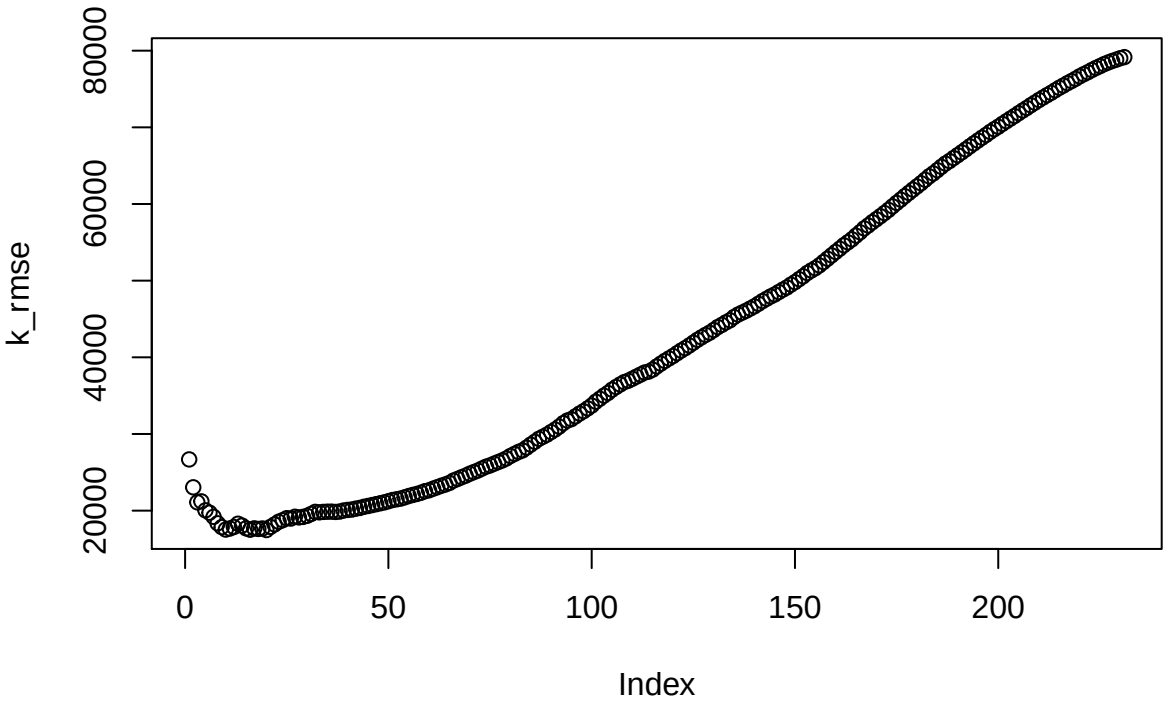
```
## [1] 18
```

The plot for this optimal K = 18 for trim 350 class is as following:



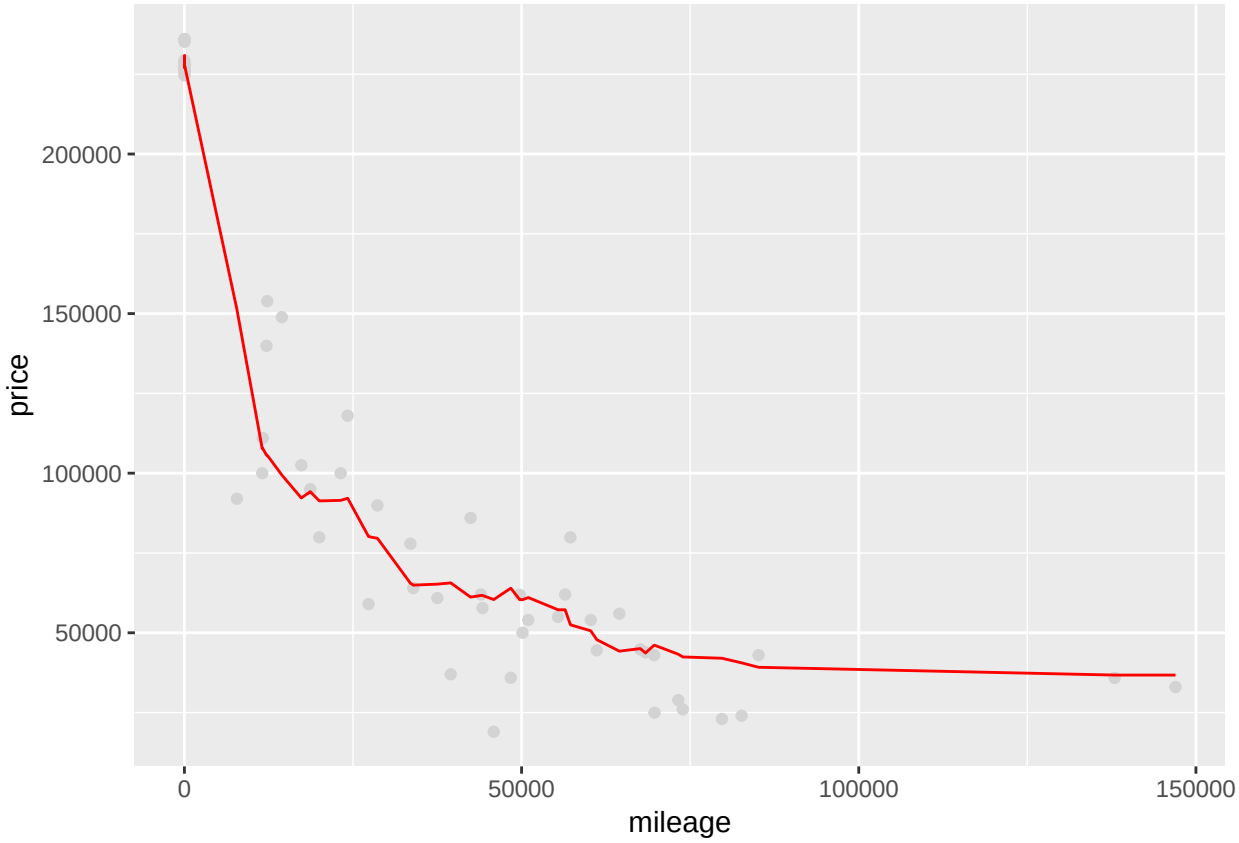
Similarly for 65AMG trim we followed the 3 guided steps

We obtain the graph of K versus root mean square error as follow:



The optimal K for 65AMG trim is

For this optimal K = 24, we have the plot of fitted model:



Since the process of splitting data into training and testing set is random, the optimal K for each trim changes from time to time. For each specific run, it is hard to tell which optimal K going to be larger. However, on average, optimal K for trim 350 class should be smaller than trim 65 AMG since there are more data for 350 class than 65 AMG.