HW1

Ashesh Shrestha

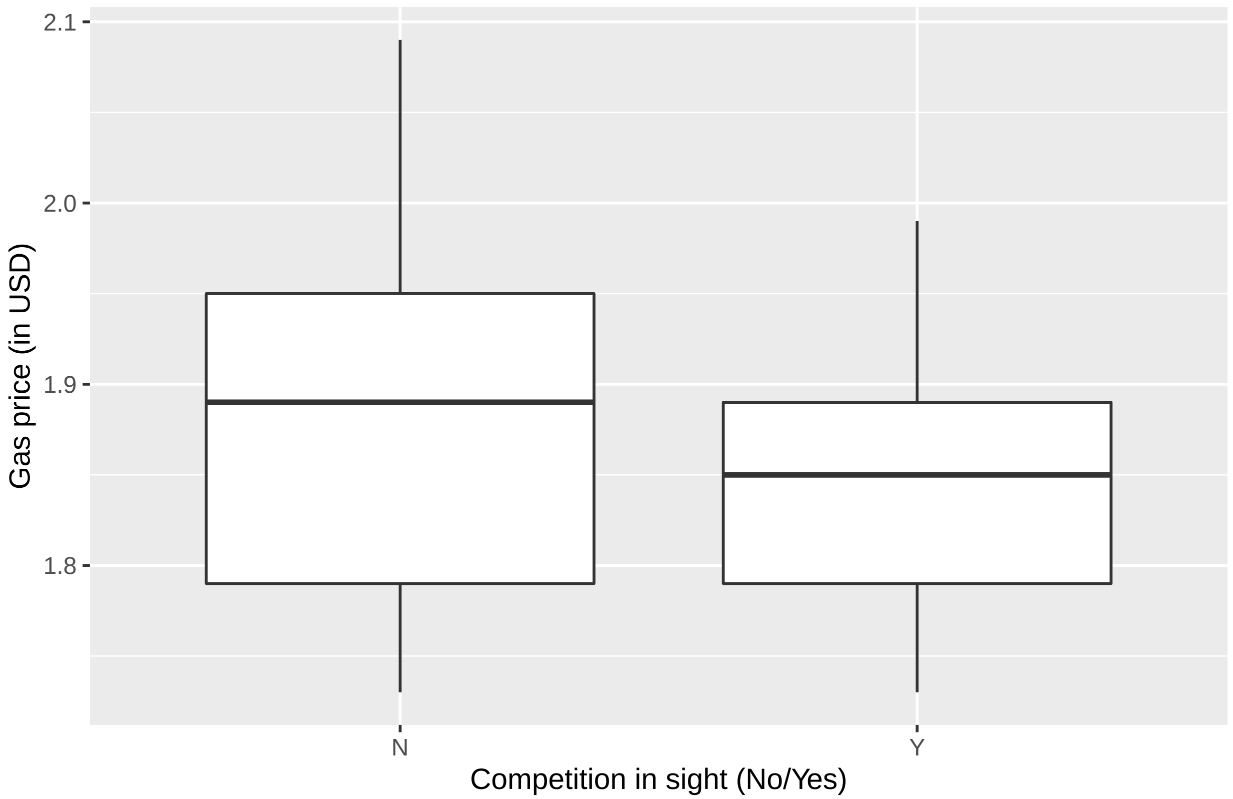
2/6/2021

ECO 395: Exercise 1

1) Data visualization: gas prices

A) Theory: Gas stations charge more if they lack direct competition in sight.

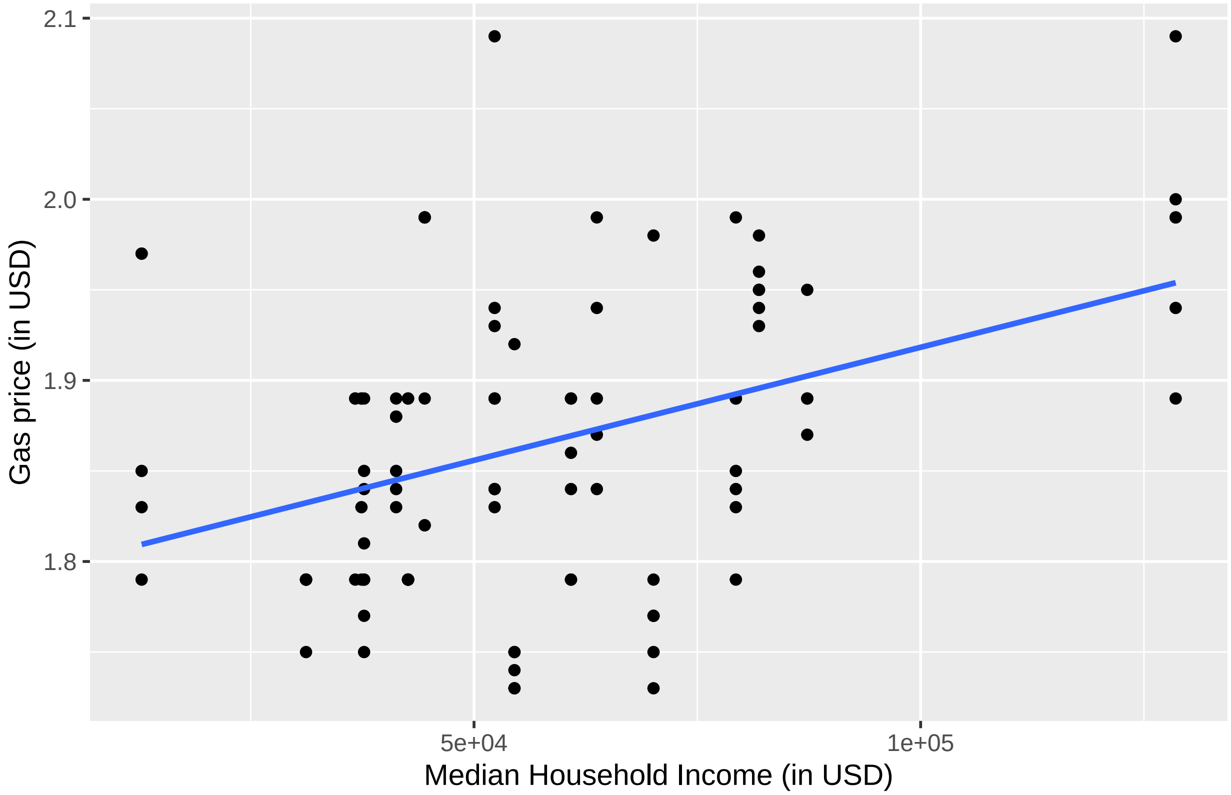
Figure 1.1 Gas stations charge more if they lack direct competition in sight



The theory states that Gas stations charge more if they lack direct competition in sight. In order to test the theory, I have made a boxplot using the data set which contains data from 101 gas stations in the Austin area collected in 2016. We can see that the median price of gas stations which do not have direct competition in sight is higher than the median price of gas station which have direct competition in sight. Therefore, we can conclude that the theory that Gas stations charge more if they lack direct competition in sight is supported by the data.

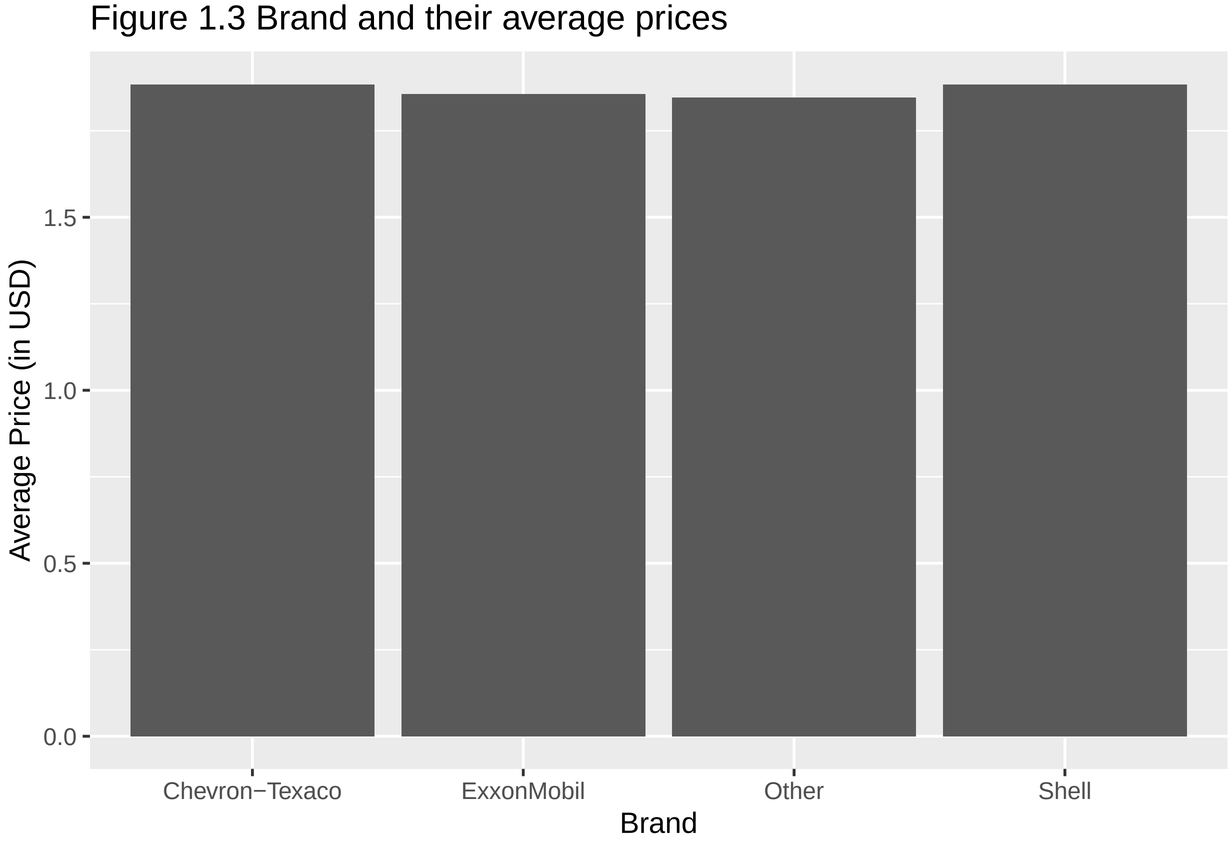
B) Theory: The richer the area, the higher the gas price.

Figure 1.2 The richer the area, the higher the gas price



The theory states that the richer the area, the higher the gas price. In order to test the theory, on the basis of data on gas prices, I have plotted a scatterplot of gas prices of various gas stations against median household income of the people where those gas stations are located. However, as it was difficult to figure out the relationship between median household income and gas prices just by looking the scatterplot, I have fitted a linear regression line to the plot. The positive slope of the linear regression line indicates the positive relationship between median household income and gas price, thus supporting the theory that the richer the area, the higher the gas price.

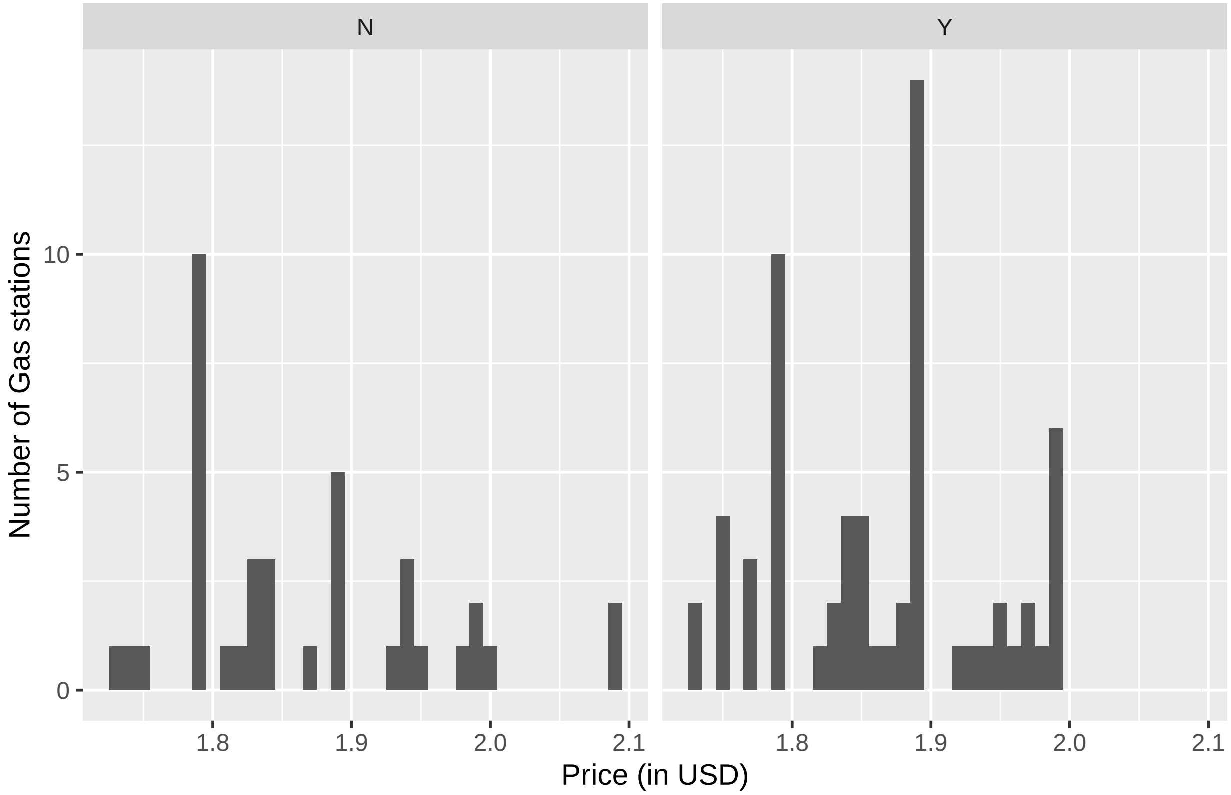
C) Theory: Shell charges more than other brands.



The theory states that Shell charges more than other brands. In order to test the theory, I have made a bar plot showing average price charged by different brands. As seen in the bar plot, averaged price charged by both Shell and Chevron-Texaco are equal and higher that the one charged by ExxonMobil and other brands. Hence, we can reject the theory that Shell charges more than other brands

D) Theory: Gas stations at stoplights charge more.

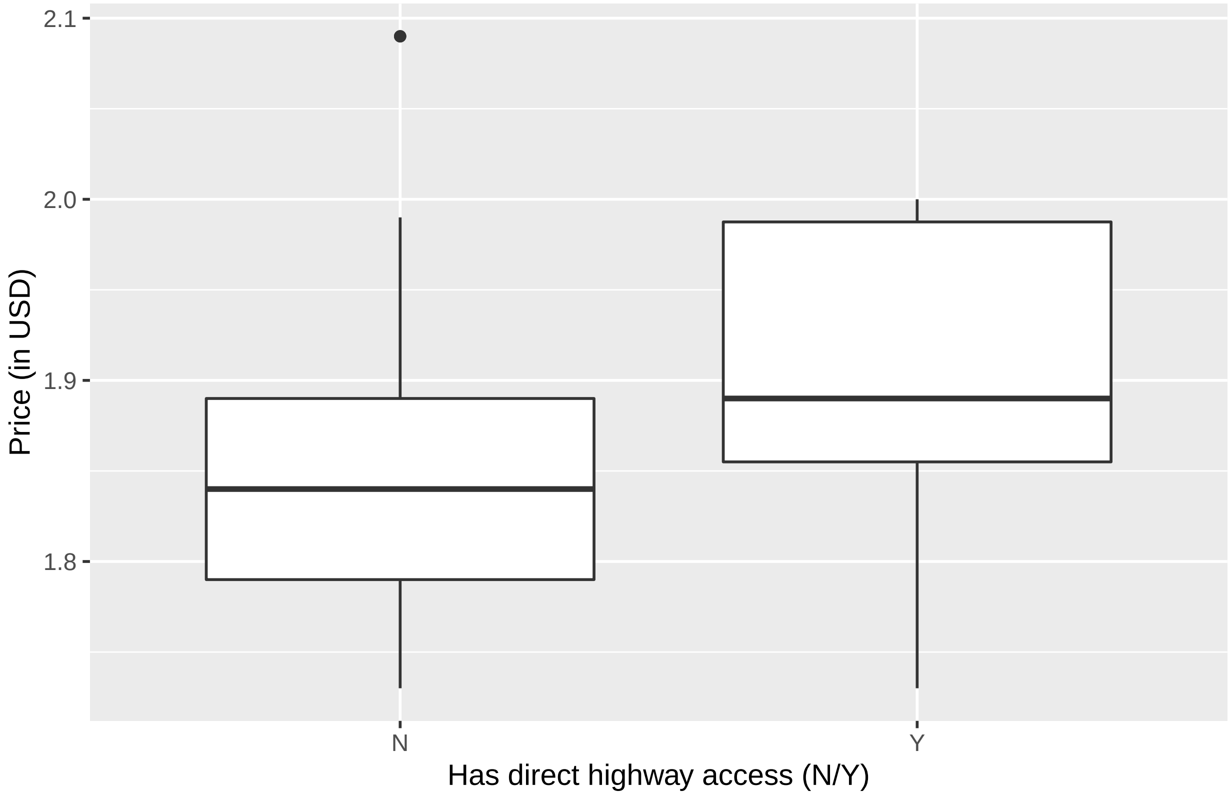
Figure 1.4 Do gas stations at stoplight charge more ?



The theory states that gas stations at stoplights charge more. In order to test the theory I have made a histogram depicting number of gas stations which are charging different prices faceted by whether they are at stoplights or not. As most number of gas stations whether they are at stoplights or not charge between 1.79 USD to 1.89 USD, we cannot support the theory that gas stations at stoplights charge more.

E) Theory: Gas stations with direct highway access charge more

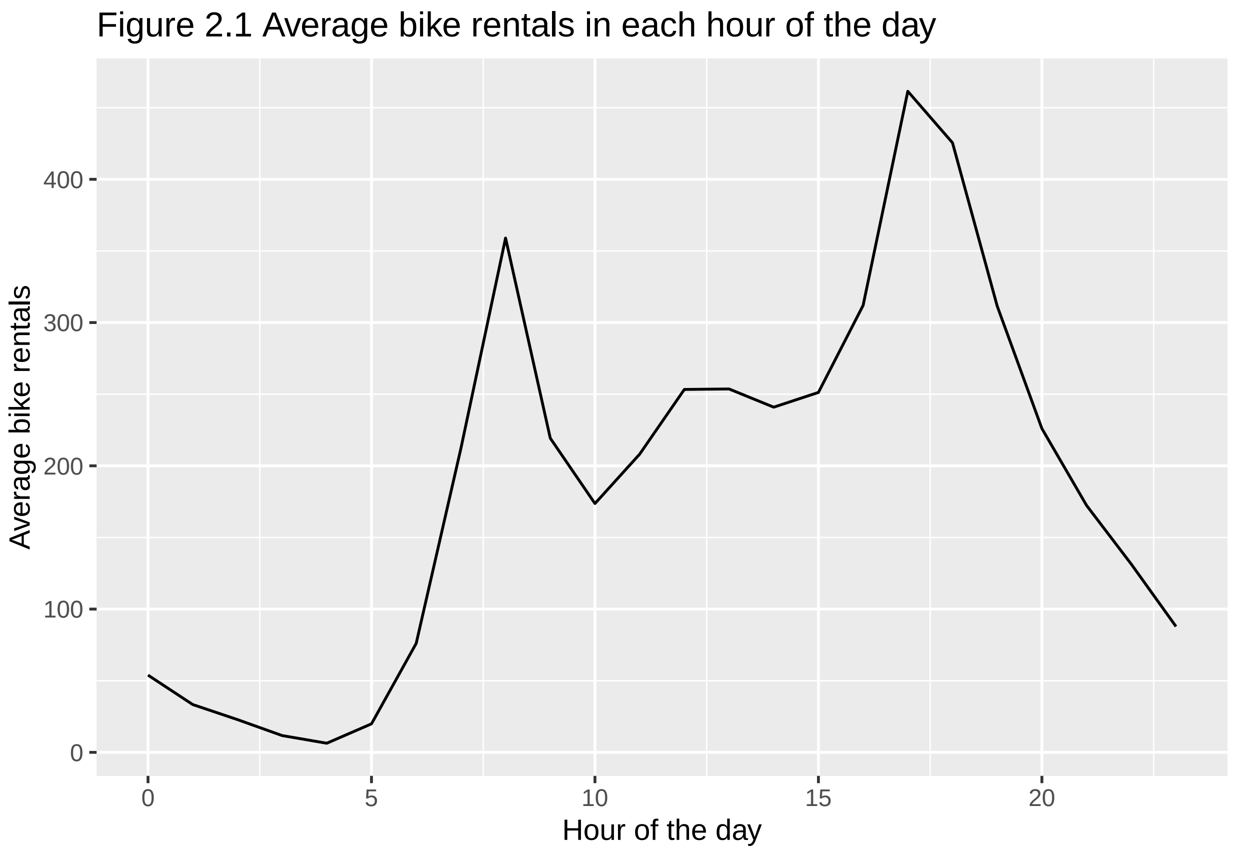
Figure 1.5 Gas stations with direct highway access charge more



The theory states that gas stations with direct highway access charge more. I have made boxplots for comparing median price of stations with direct highway access with the median price of those without direct highway access. As we can see that median price of highways with direct access to highway is higher that without direct access, we can conclude that the data supports the theory.

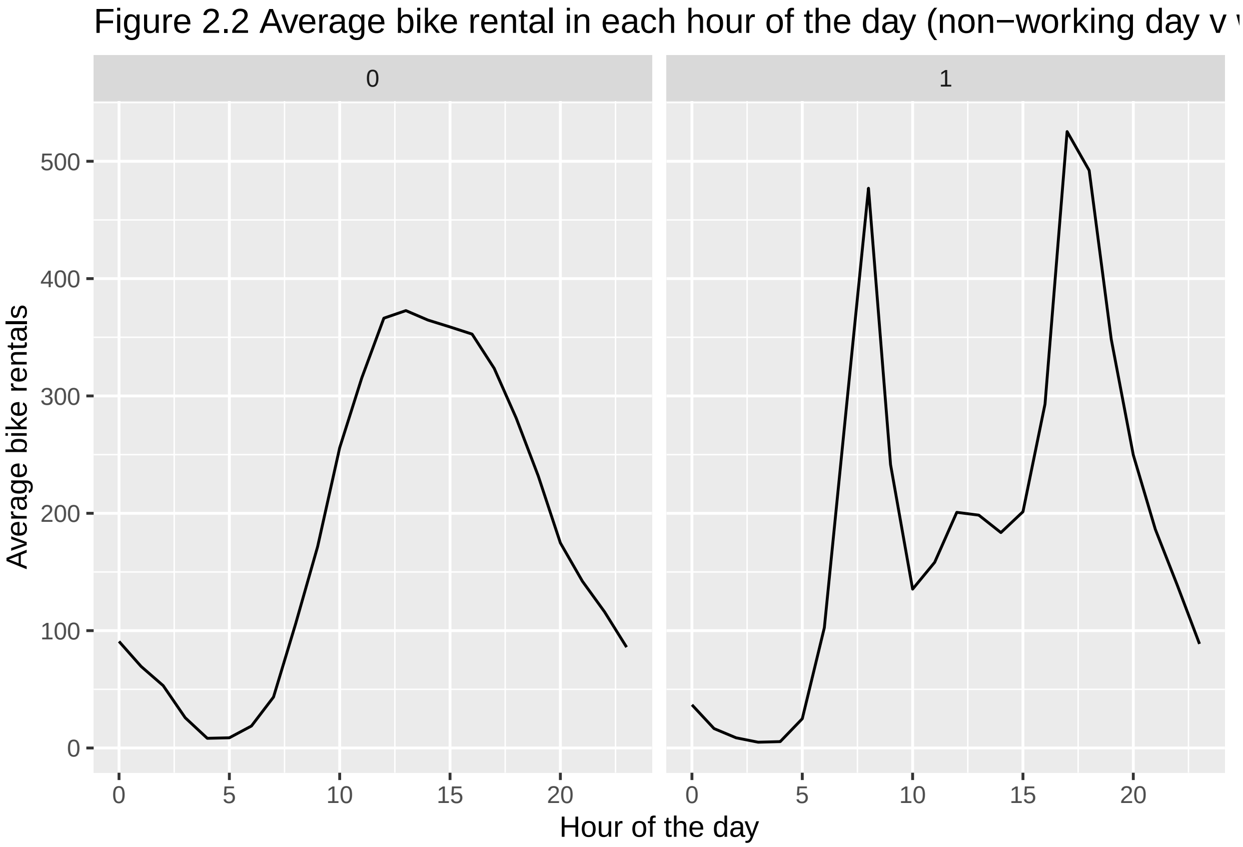
2) Data visualization: a bike share network

Plot A: a line graph showing average bike rentals (total) versus hour of the day (hr).



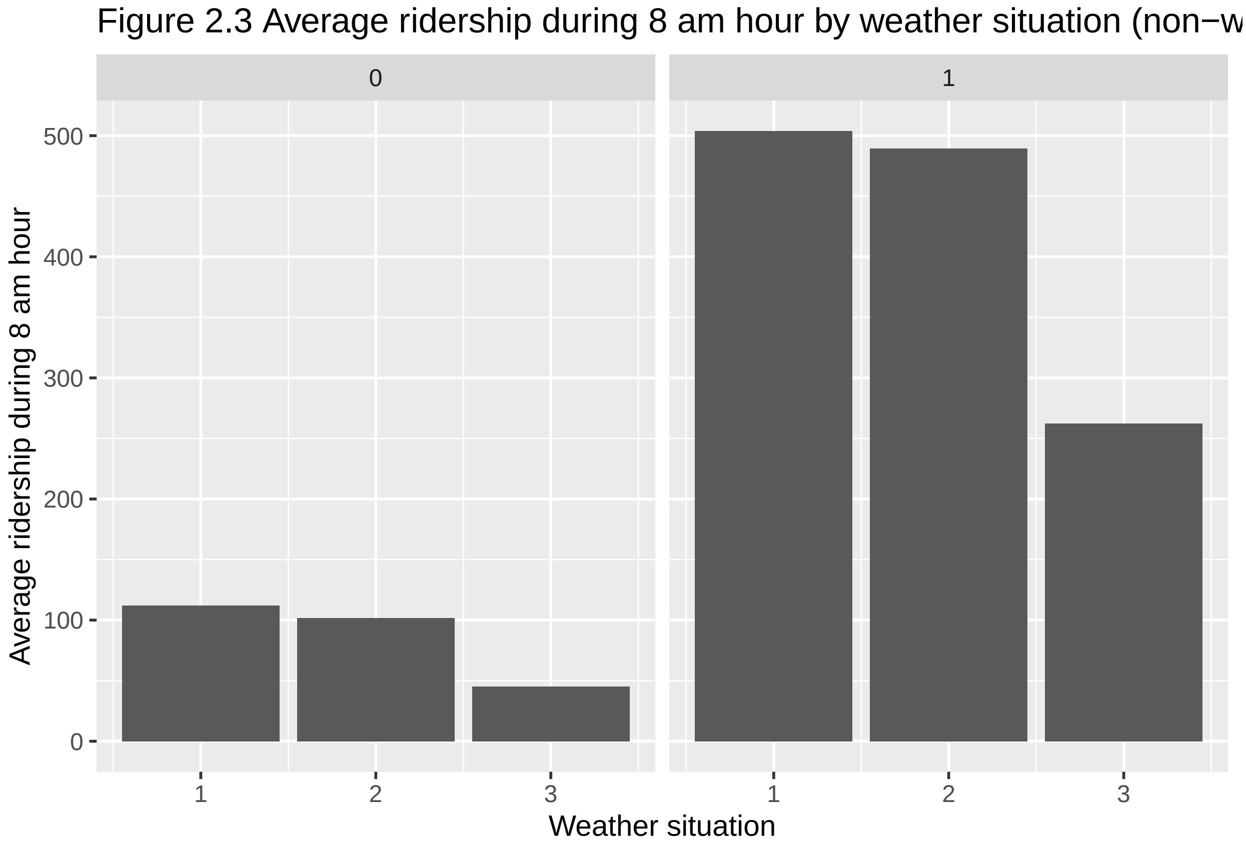
The Figure 2.1 depicts the average bike rentals during each hour of the day using the data set of two-year historical log (2011 and 2012) from Capital Bikeshare system in Washington DC. The X-axis measures the hour of the day, 0 through 23 and the Y-axis measures the average number of bikes rental during those hours. As can be seen in the plot, the average bike rentals declines steadily until the 4 am hour and reaches close to zero before increasing rapidly until 8 am hour. After reaching about 360 at 8 am hour, there is a sudden decline in average bike rentals which continues till 10 am hour. From 10 am hour, even though bike rentals increase, it does not increase by much until 3 pm hour of the day. The average bike rental sharply ascends and reaches its peak amounting to 460 at around 5 pm hour. Starting from 5 pm hour, the average biker rental continuously plunges and reaches about 80 at the end of the day. To conclude, the average ridership during the start of the business hours and end of the business hours are very high, whereas, the ridership between those hours is moderate.

Plot B: a faceted line graph showing average bike rentals versus hour of the day, faceted according to whether it is a working day



The figure 2.2 shows the average bike rentals during each hour of the day faceted according to whether it is a working day or not. The X-axis measures the hour of the day, 0 through 23 and the Y-axis measures the average number of bike rentals during those hours. If we compare average bike rentals during first four hours of the day of non-working day with that of working day, we can see that average bike rentals are plummeting during both kind of days. However, the average bike rentals during non-working days are relatively higher than that in working days. The average bike rentals start rising up starting at 4 am hour both during non-working and working days, but the rise is steeper during working days. During non-working days, the average bike rentals reaches highest to about 370 at 1 pm hour of the day, after which it continuously decline. Whereas, during working days, the average bike rentals escalates starting at 4 am hour until it reaches 470 during 8 am hour. Again, there is a sharp decline at the rate equivalent to the rise. The decline persists until 10 am hour of the day. Then, we can see a significant rise beginning at 3 pm hour which lasts for 2 hours until 5 pm of the day starting from which it continuously decreases till the end of the day. In conclusion, bike ridership during different hours of the day depends on whether it is a working day or a non-working day.

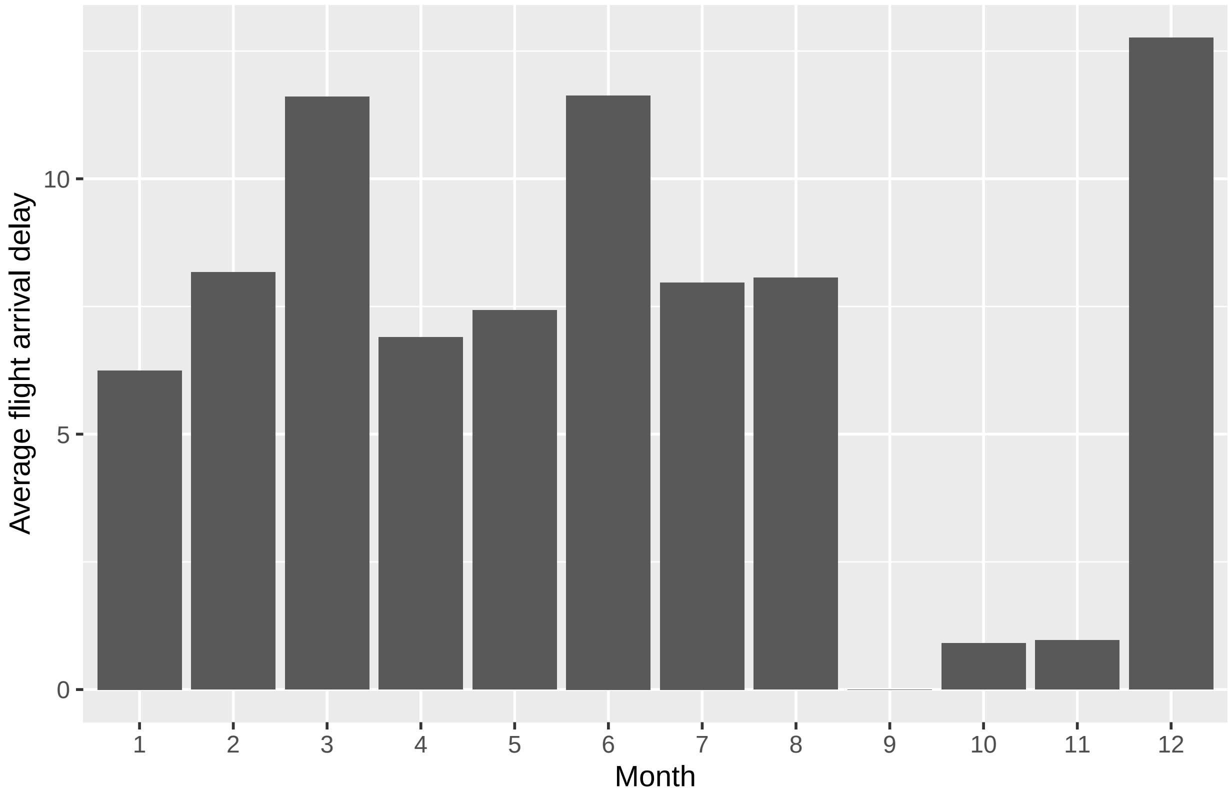
Plot C: a faceted bar plot showing average ridership during the 8 AM hour by weather situation code (weathersit), faceted according to whether it is a working day or not.



The figure 2.3 shows the average ridership during 8 am hour of the day by weather situation faceted according to whether it is a working day or not. The X -axis shows various weather situations; 1= Clear, Few clouds, Partly cloudy, Partly cloudy, 2= Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3= Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4= Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, and the Y-axis measures the average bike ridership during 8 am hour of the day. As seen in the bar plot, regardless whether it is a working day or not, the average ridership declines as weather worsens during 8 am hour. Moreover, the average ridership at 8 am hour of working day outnumbers average ridership at 8 am hour of non-working day, irrespective of the weather condition. In conclusion, both weather situation and type of day (working or non-working), affect bike ridership during 8 am hour of the day. In conclusion, average ridership during the beginning and end of the business hours are very high during working day, while, during other hours of the day, average ridership is higher during non-working days.

3) Data visualization: flights at ABIA

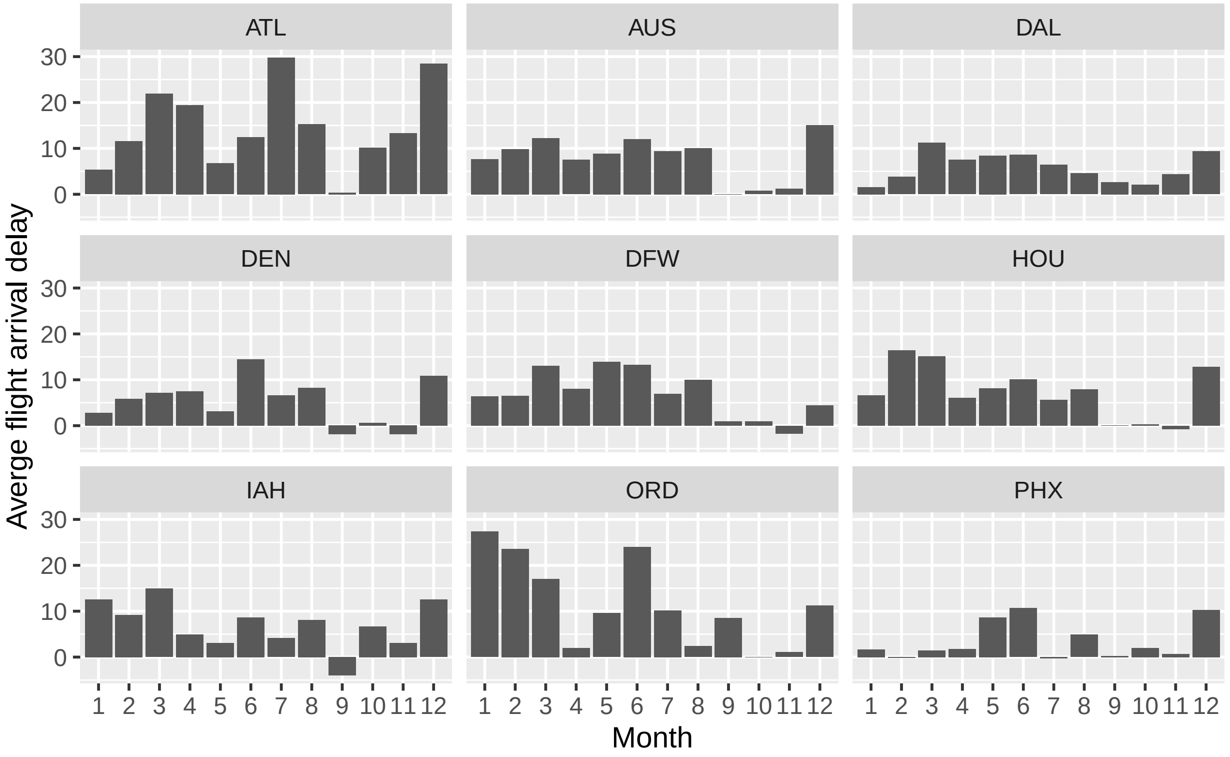
Figure 3.1 Average flight delays in various months of the year



In the Figure 3.1, we have months of the year, 1 through 12 in X-axis and average arrival delays of the flights in Y-axis. The bar plot shows the average arrival delays of the flights by month. We can see that during 9th, 10th and 11th month of the year, that is, September, October and November, the average flight delays are lowest with no delays during the month of September. Hence, we can say that the best time of the year to fly to avoid delays would be during September, October and November

```
## # A tibble: 53 x 2
##    Dest  count
##    <chr> <int>
##  1 AUS   49637
##  2 DAL    5573
##  3 DFW    5506
##  4 IAH    3691
##  5 PHX    2783
##  6 DEN    2673
##  7 ORD    2514
##  8 HOU    2319
##  9 ATL    2252
## 10 LAX    1733
## # ... with 43 more rows
```

Figure 3.2 Average flight delays in various months of the year faceted by popular destinations

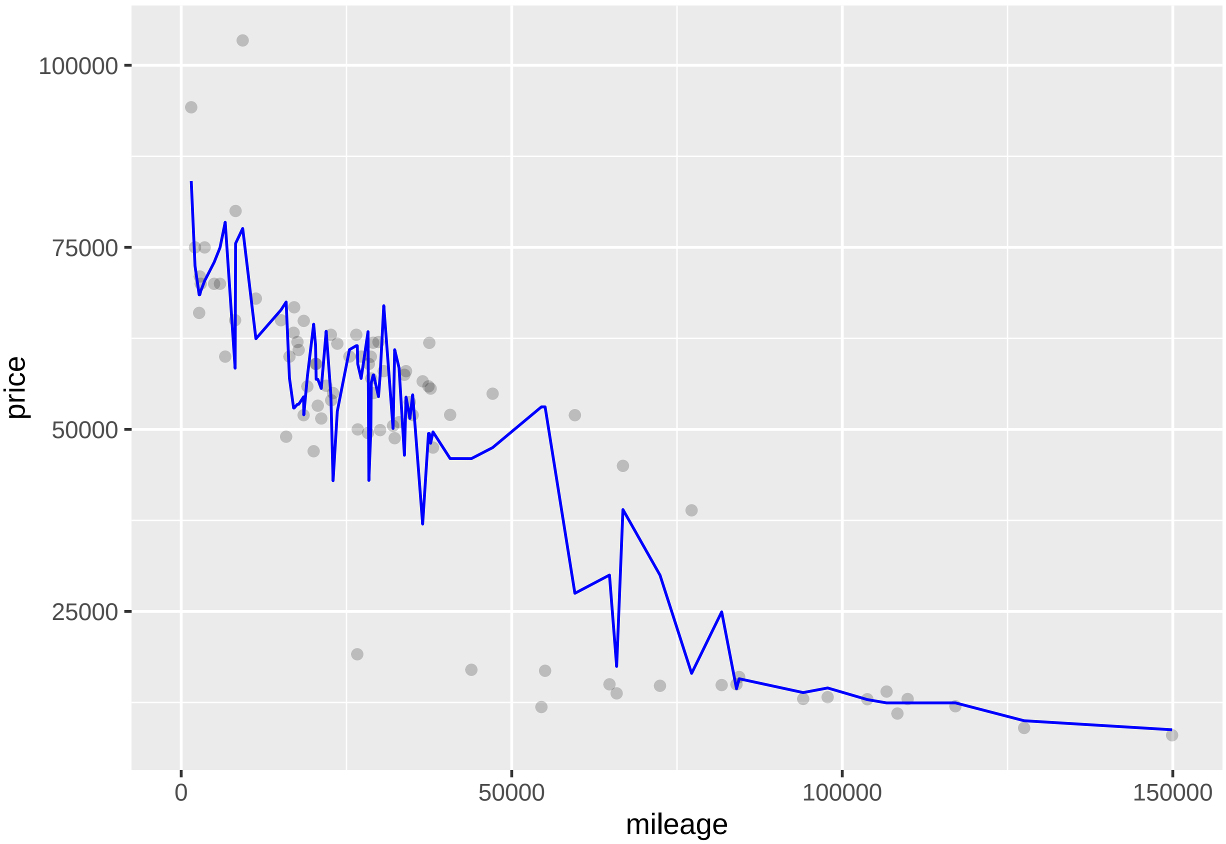


Now, to see whether this changes by destinations, I have faceted the bar plot by destination. As seen in the Figure 3.2, even when considering various destinations, we can see that September, October and November have low average arrival delays. Thus, we can conclude that this is the best time of the year to fly to minimize delays.

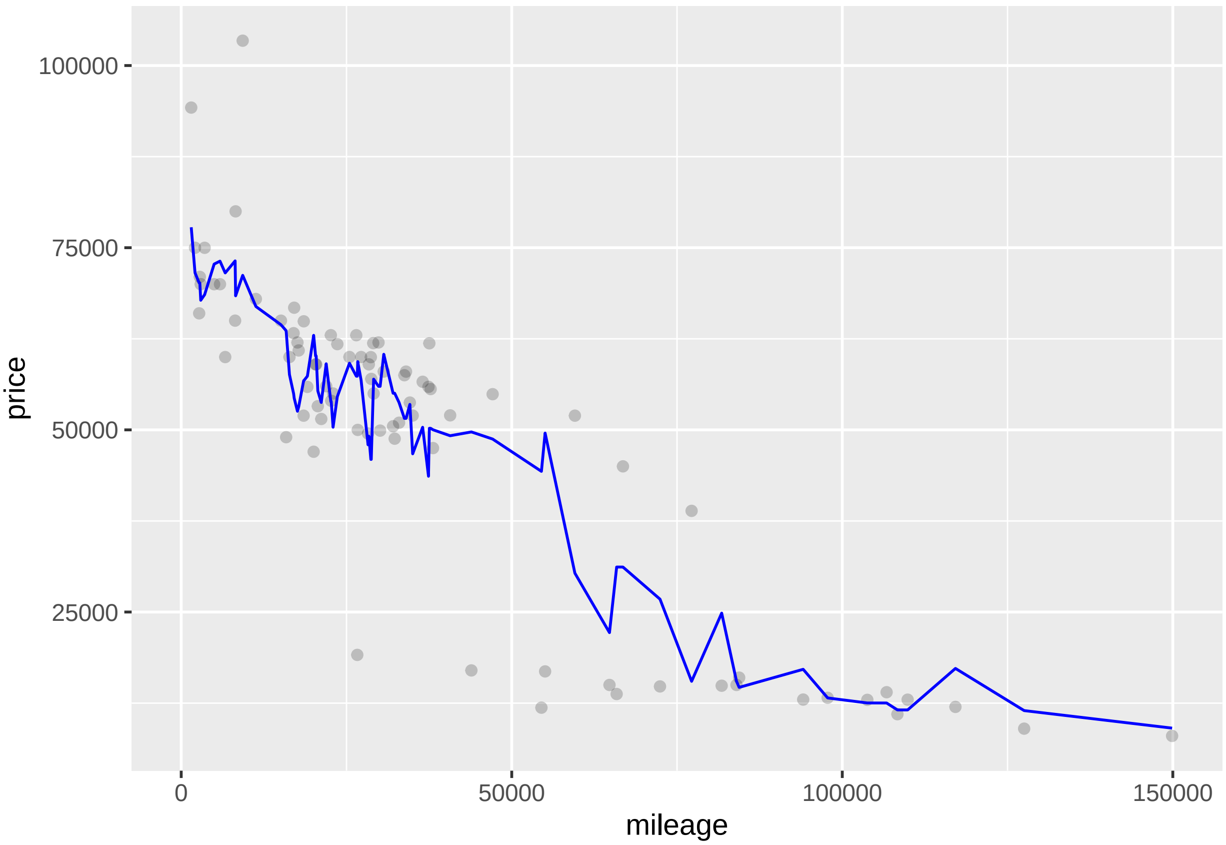
4) K-nearest neighbors

For trim level 350

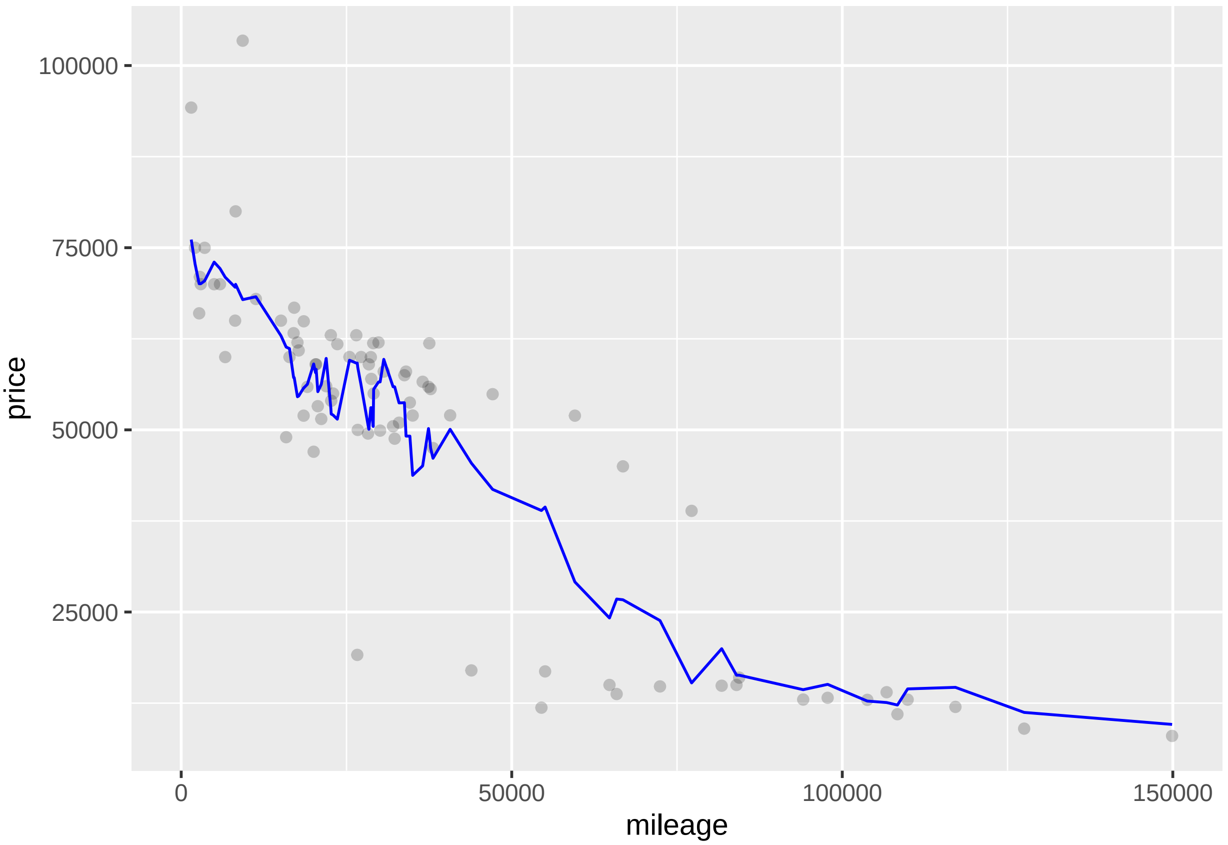
K=2



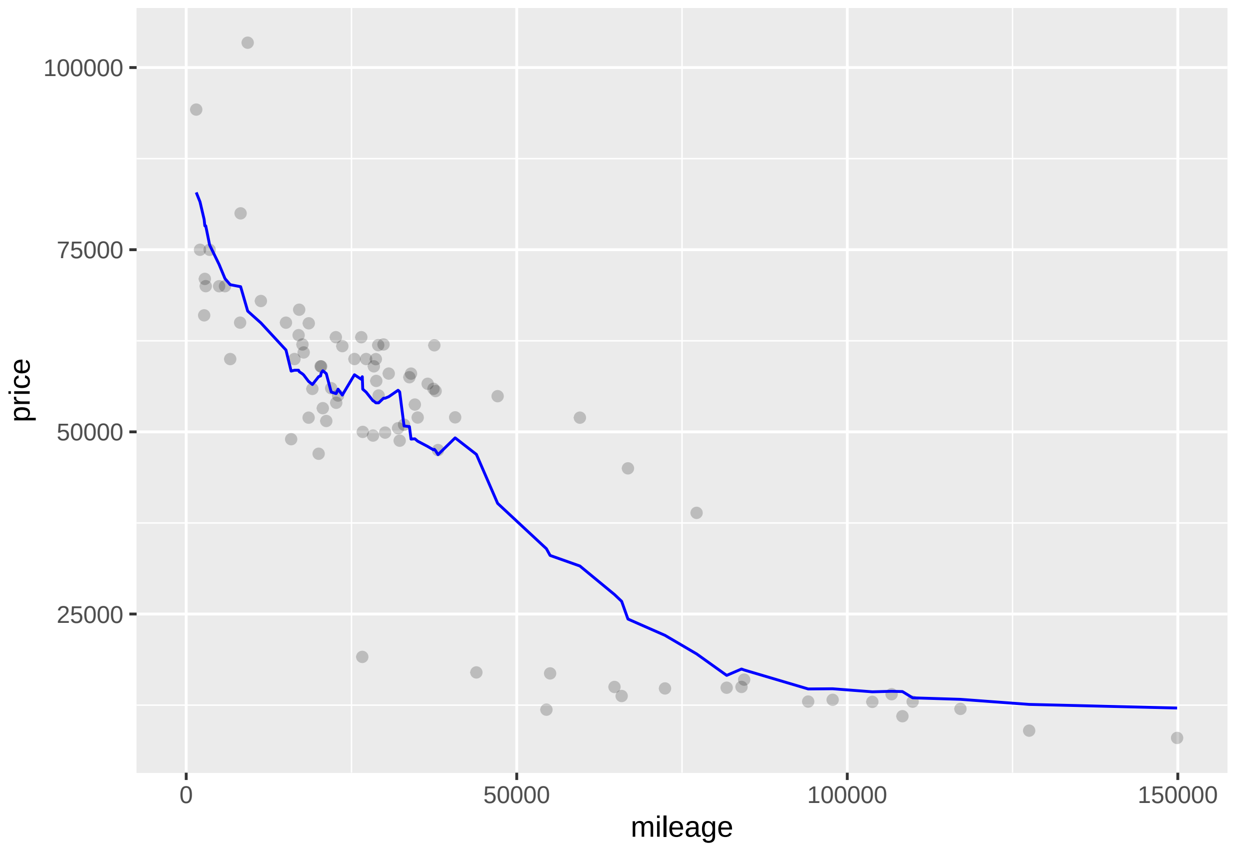
K=5


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

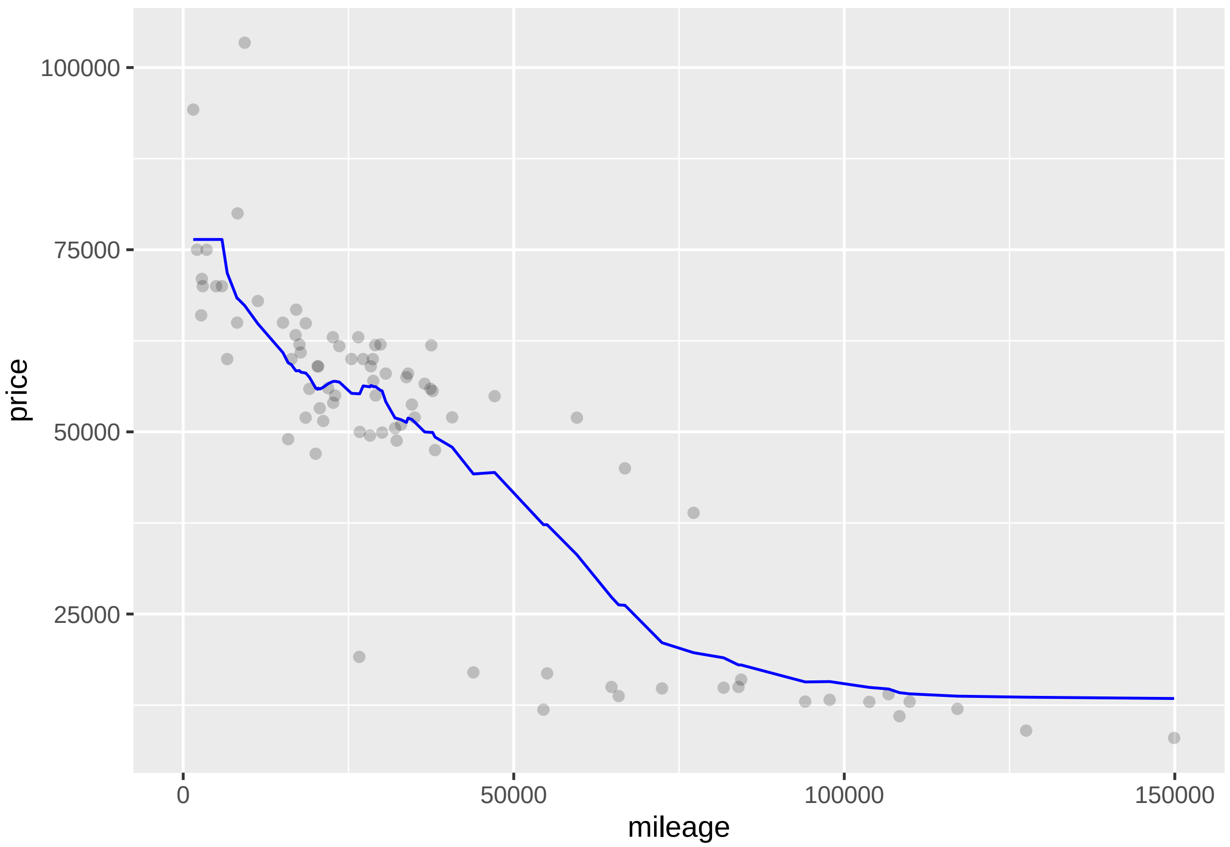
K=10


price
100000
75000
50000
25000
0
50000
100000
150000
mileage

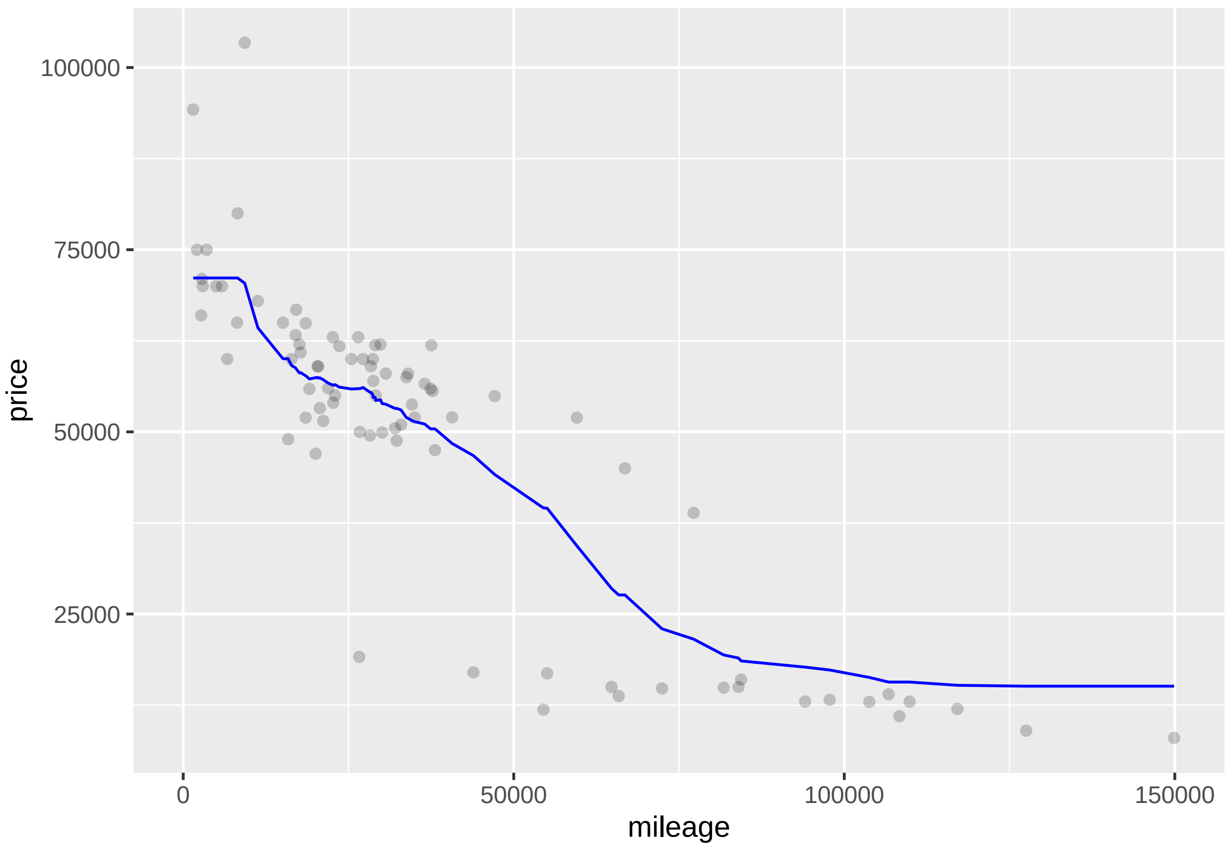
K=25


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

K=50

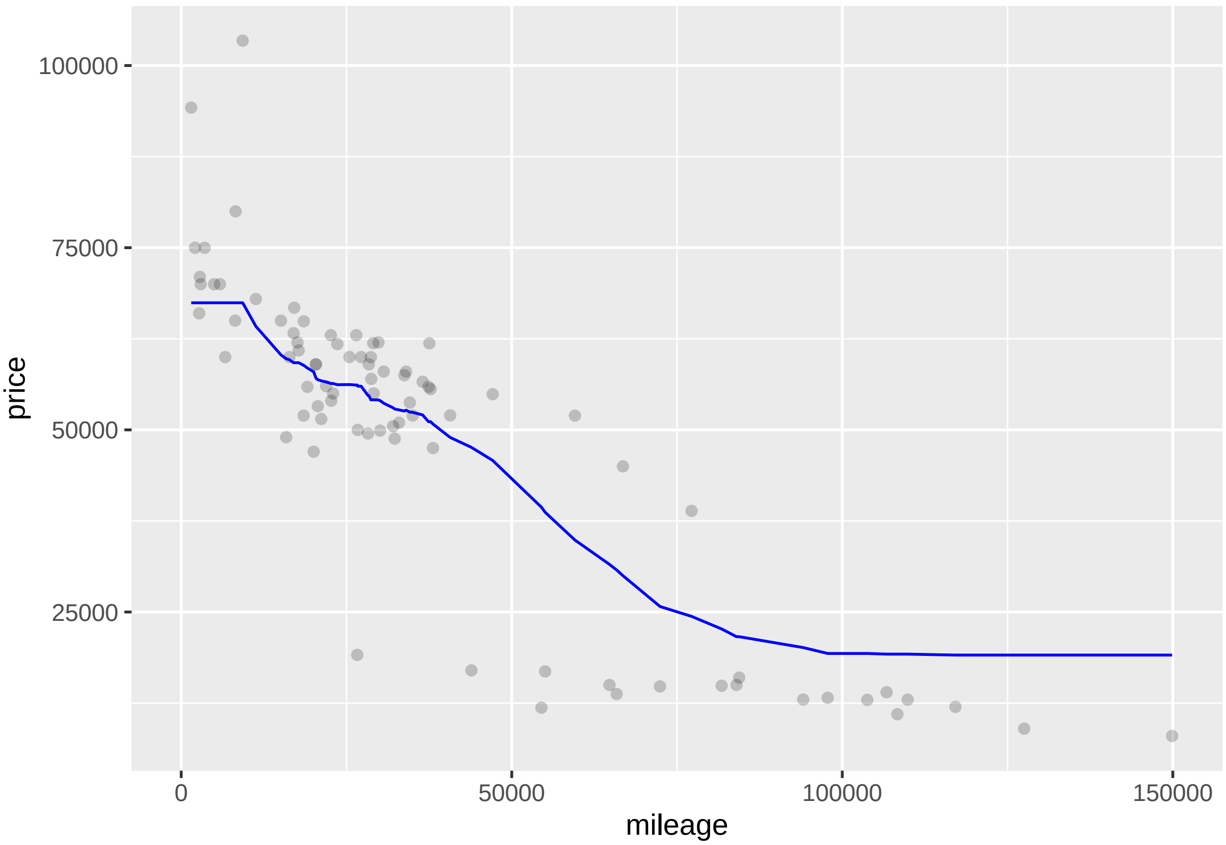

100000
75000
50000
25000
price
0
50000
100000
150000
mileage

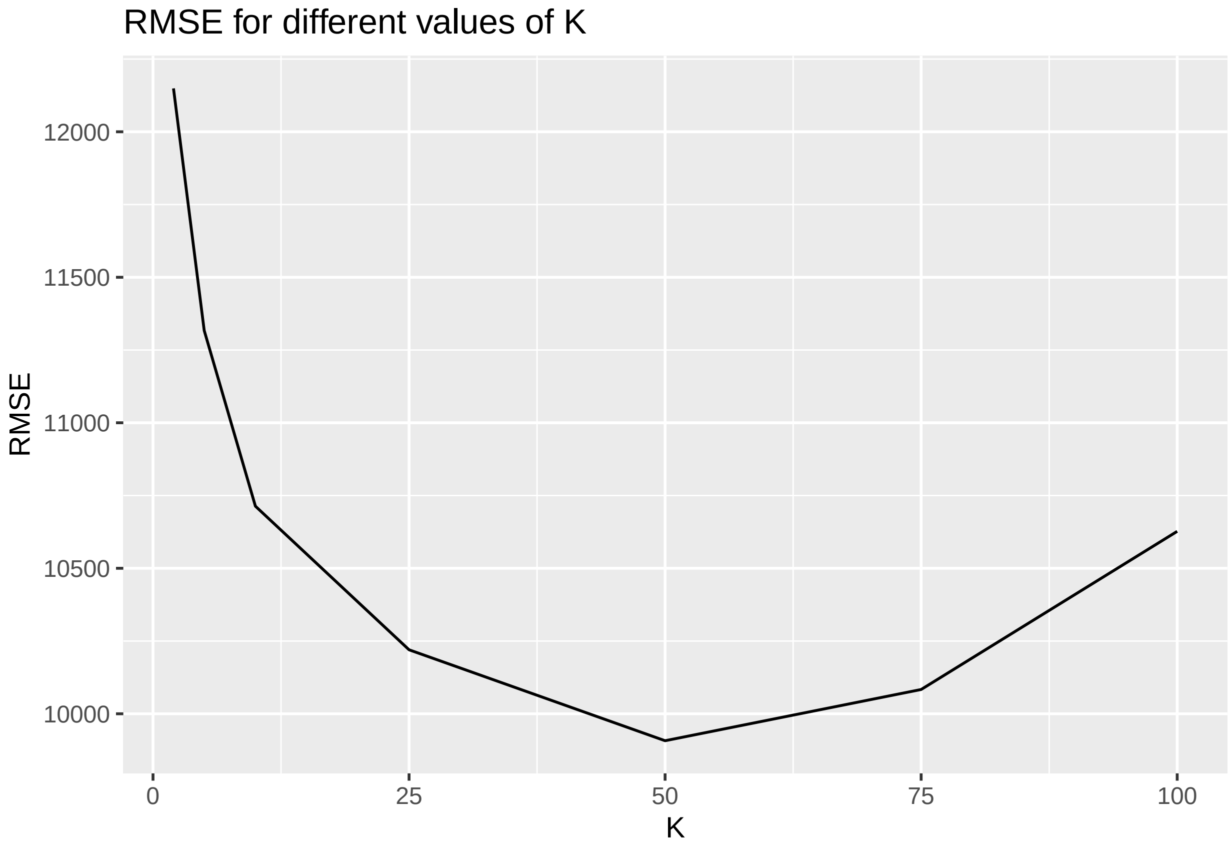
K=75

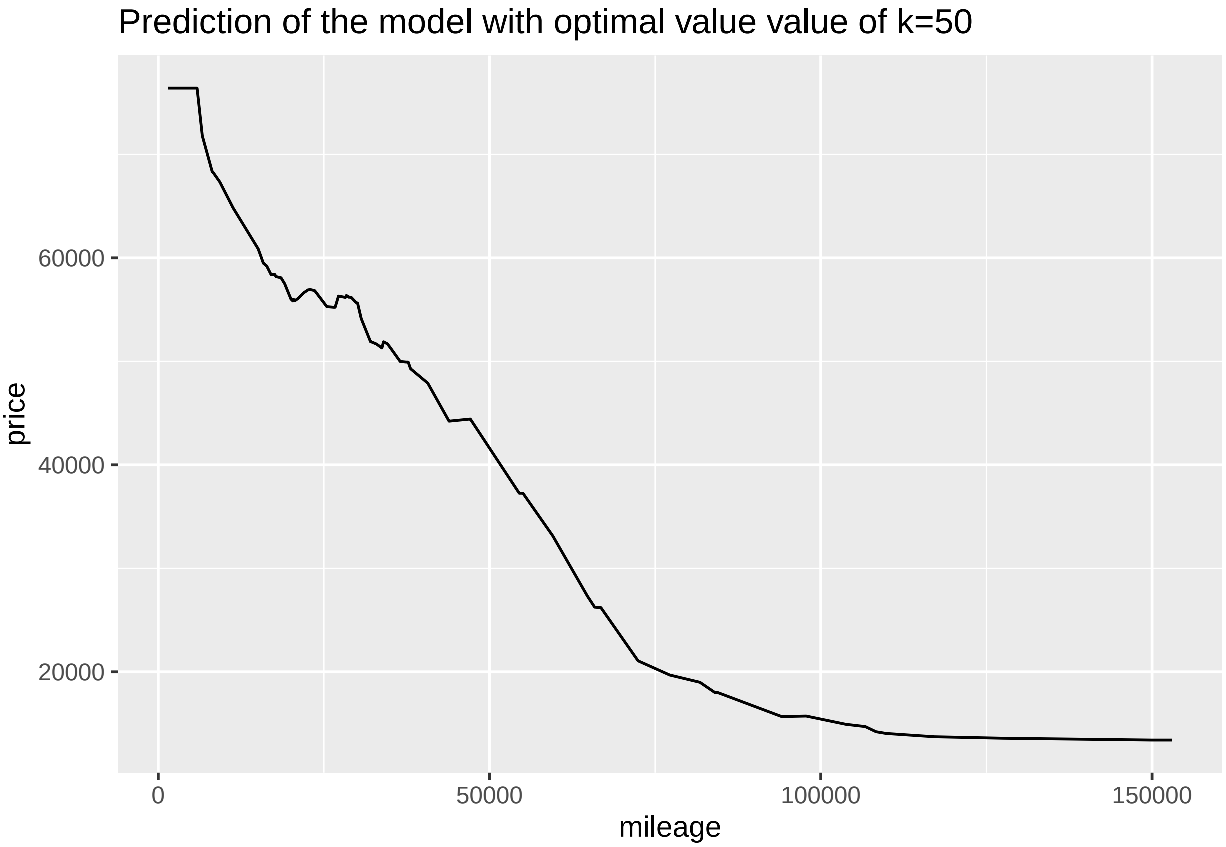


price
100000
75000
50000
25000
0
50000
100000
150000
mileage

K=100

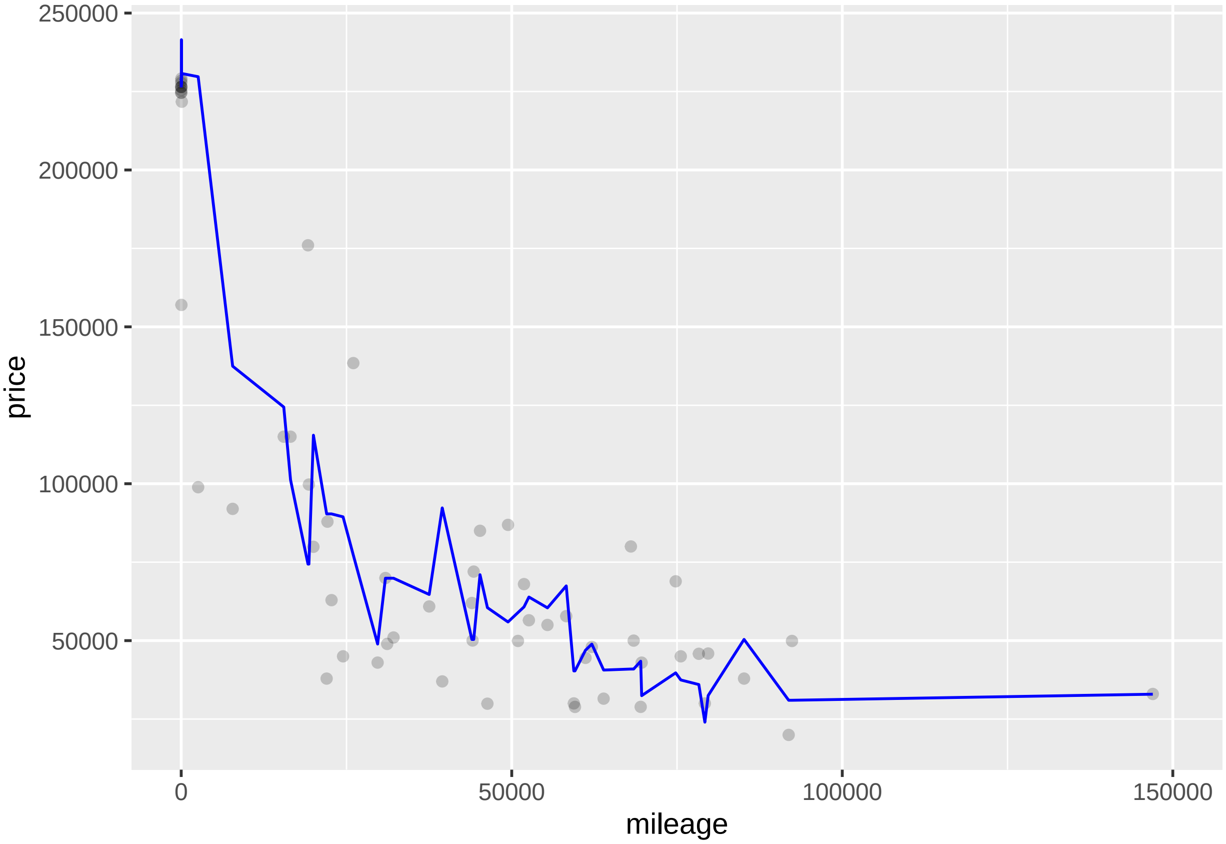

100000
75000
50000
25000
price
0
50000
100000
150000
mileage


RMSE for different values of K
RMSE
12000
11500
11000
10500
10000
K
0
25
50
75
100

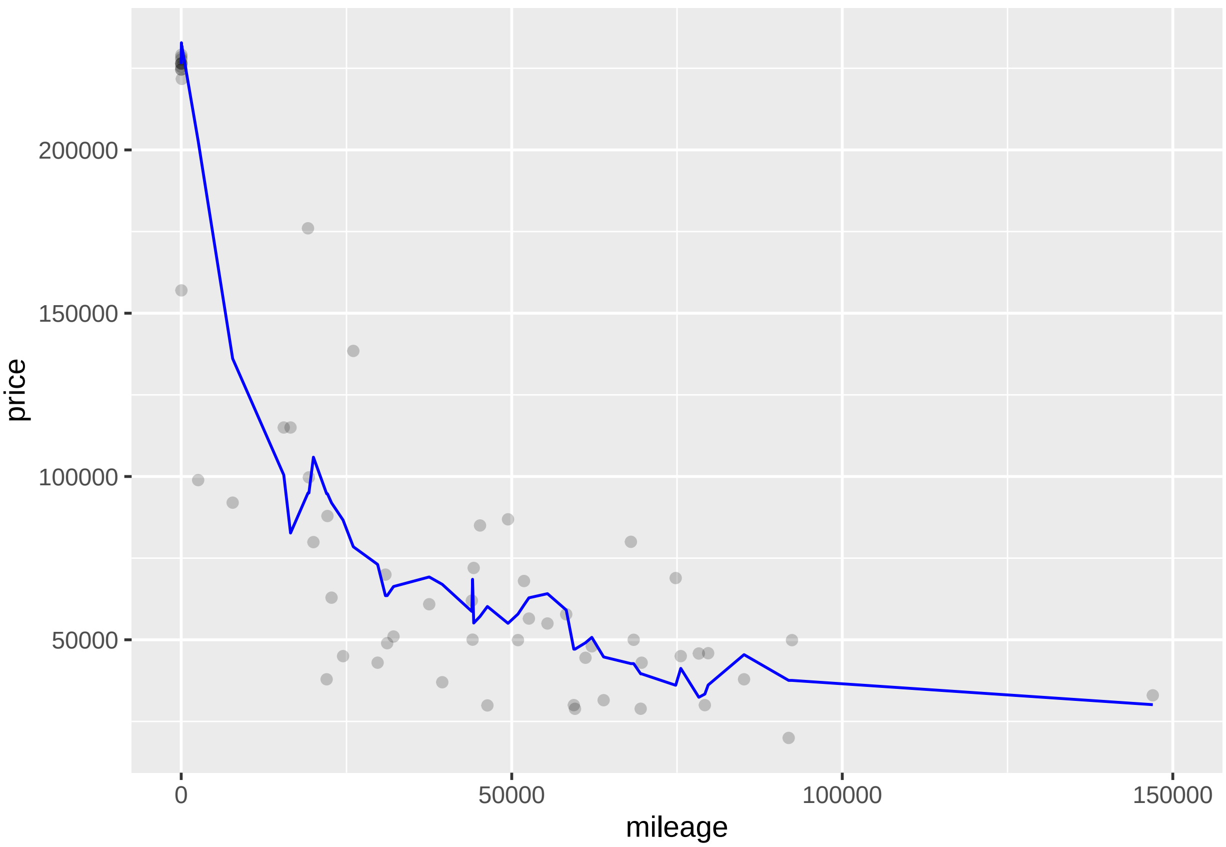

Prediction of the model with optimal value value of k=50
price
60000
40000
20000
0
50000
100000
150000
mileage

For trim level 65AMG

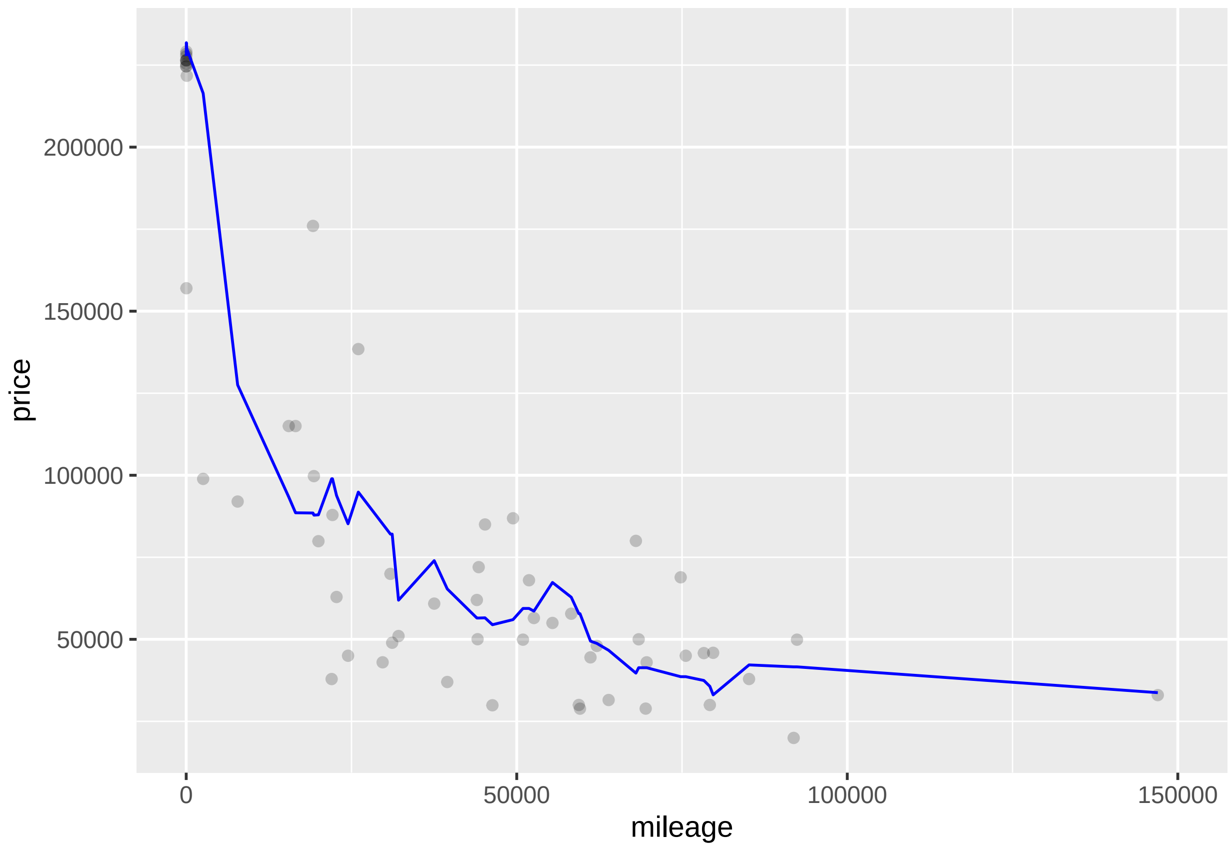
K=2



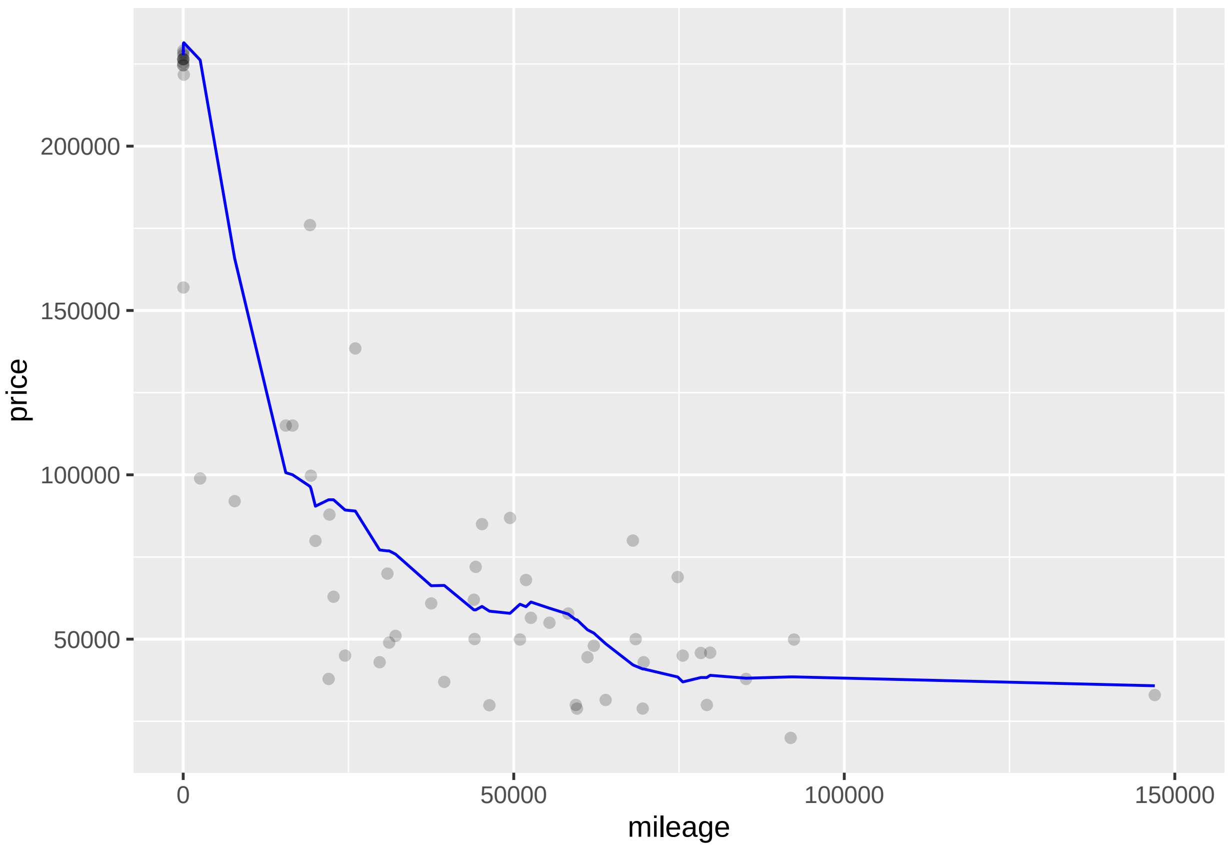
K=5


price
200000
150000
100000
50000
0
50000
100000
150000
mileage

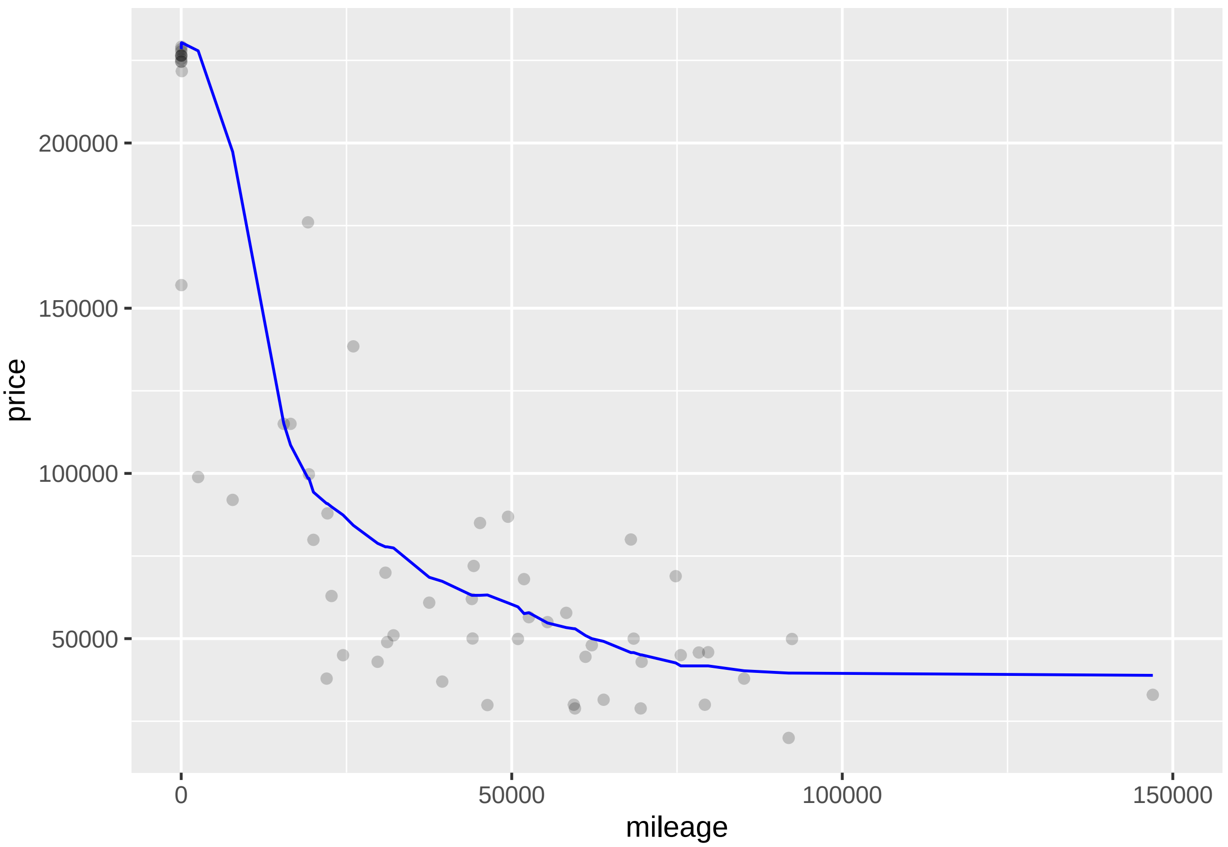
K=10


200000
150000
100000
50000
price
0
50000
100000
150000
mileage

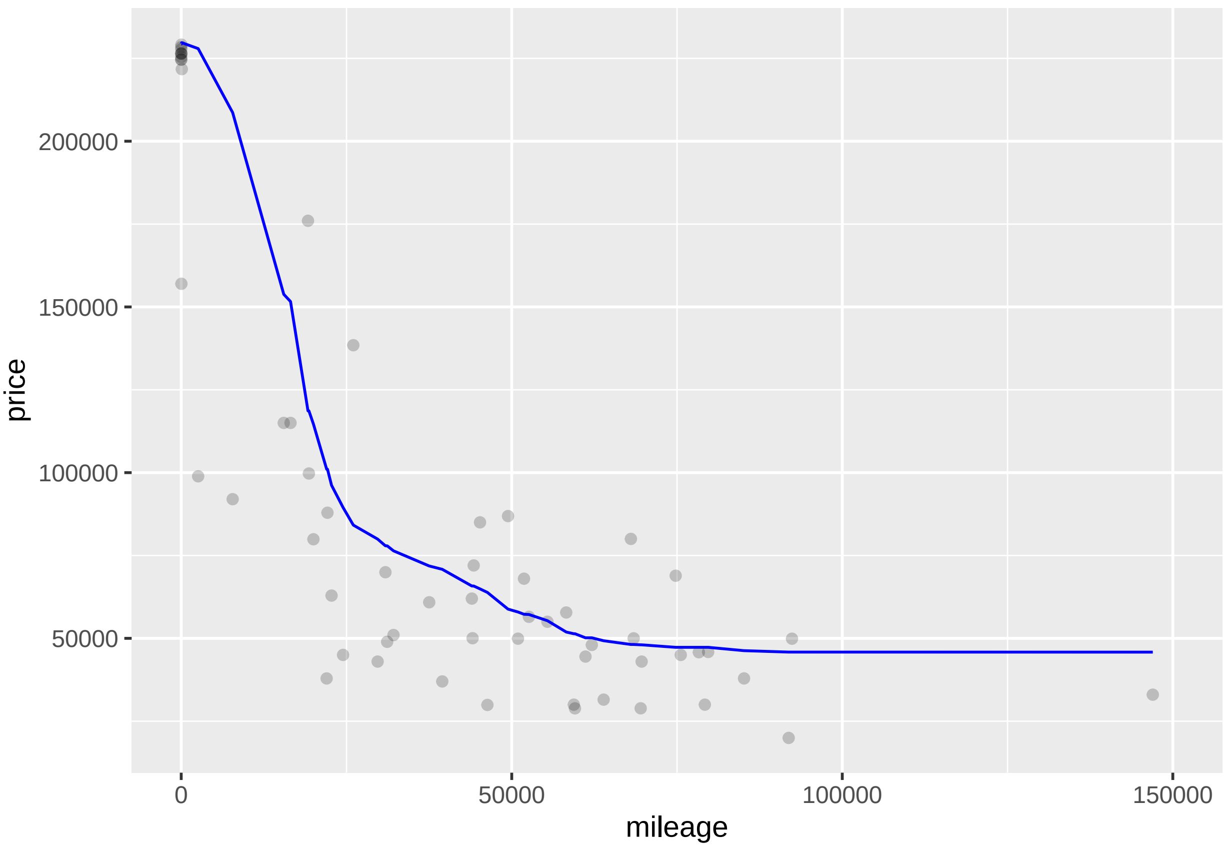
K=25


price
200000
150000
100000
50000
0
50000
100000
150000
mileage

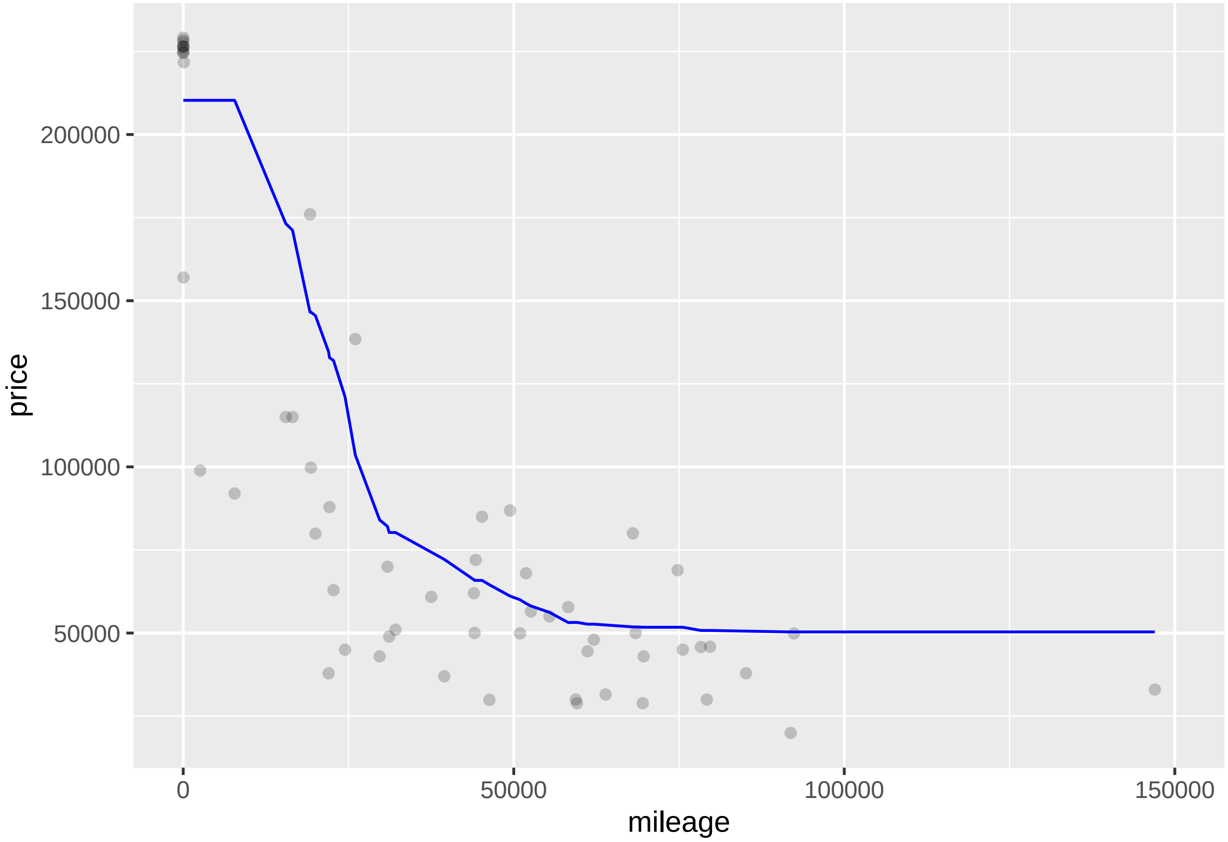
K=50

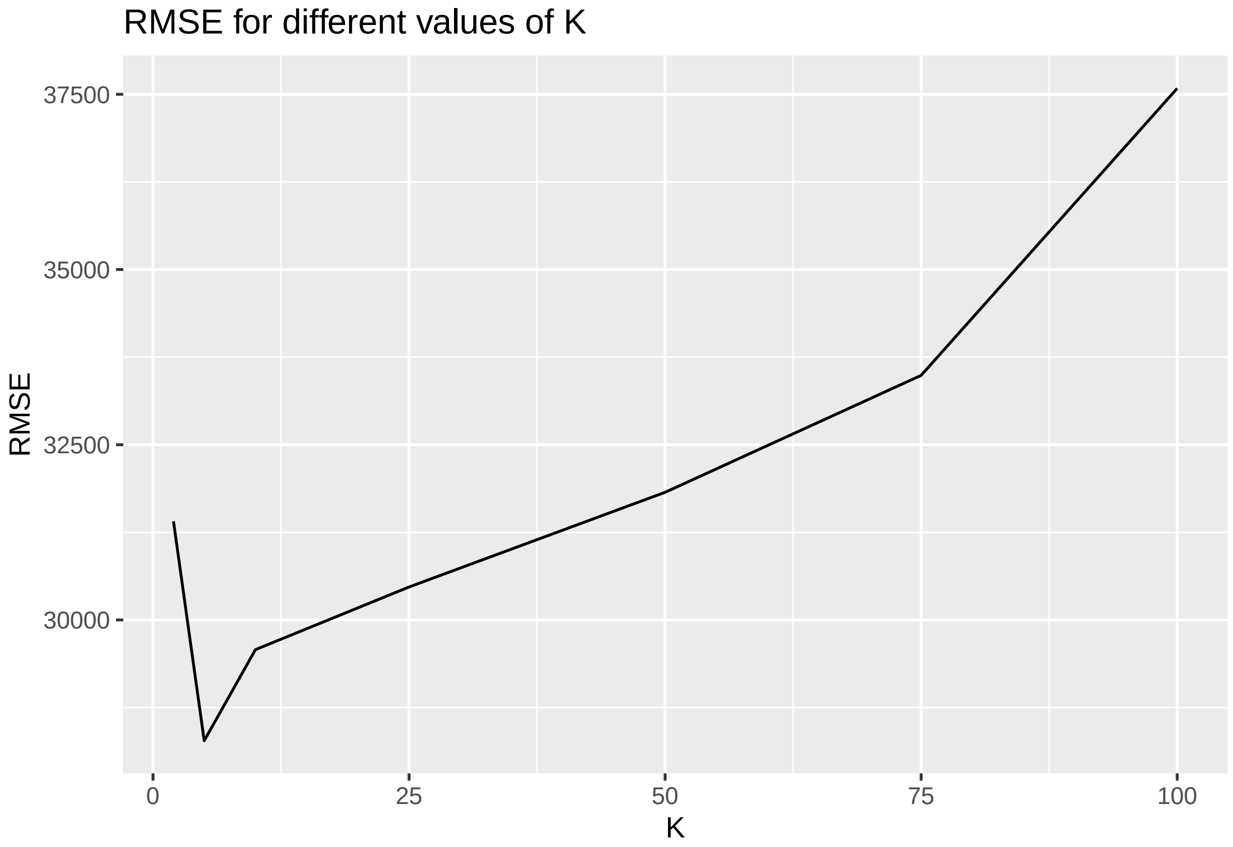

price
200000
150000
100000
50000
0
50000
100000
150000
mileage

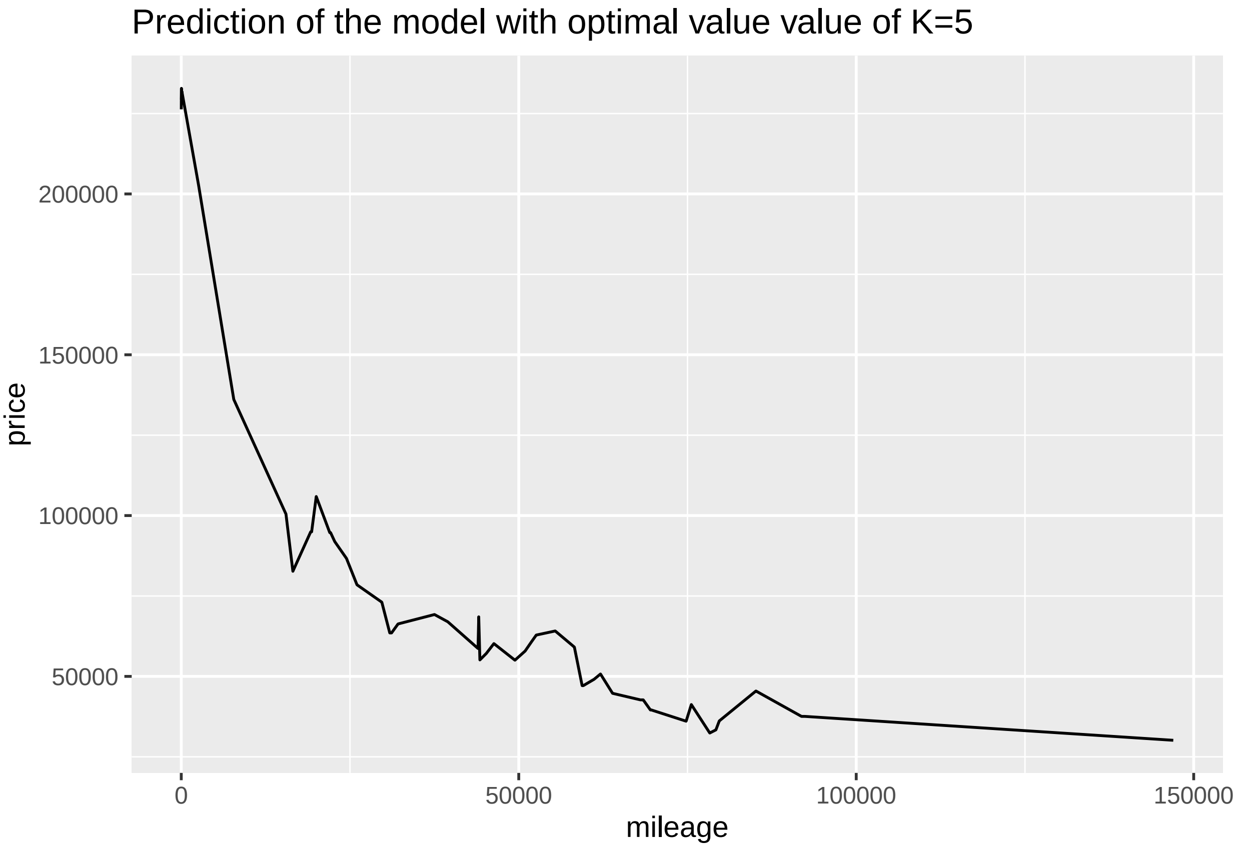
K=75


200000
150000
100000
50000
price
0
50000
100000
150000
mileage

K=100


price
200000
150000
100000
50000
0
50000
100000
150000
mileage


RMSE for different values of K
RMSE
37500
35000
32500
30000
K
0
25
50
75
100



Trim size 350 yields a larger optimal value of K. RMSE differs from one train/test split to another. In this particular random assignment of data into training and testing data in the ratio of 80:20, it so happened that for trim size 350, larger value of K yielded lowest estimate of RMSE.