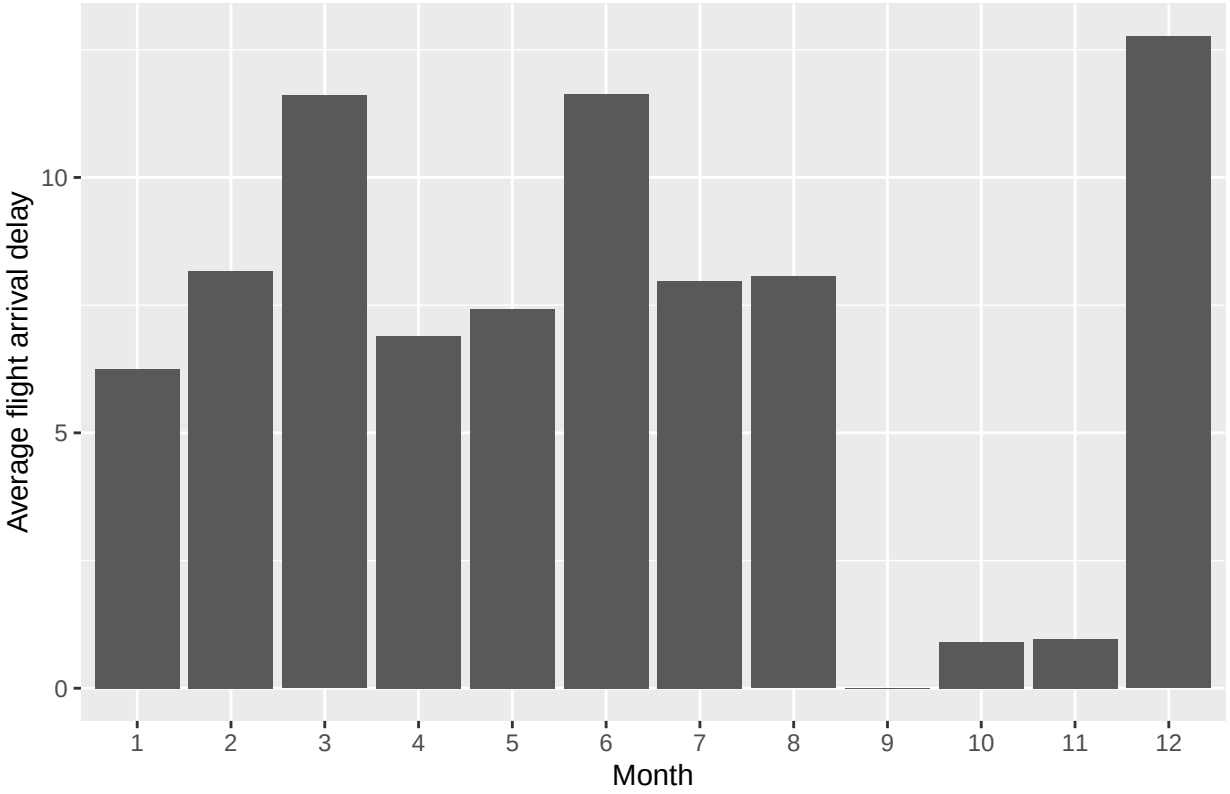
ECO395M_Exercise1

Youngseok Yim (EID: yy9739)

2023-01-30

1) Data visualization: flights at ABIA

Figure 1.1 Average flight delays in various months of the year

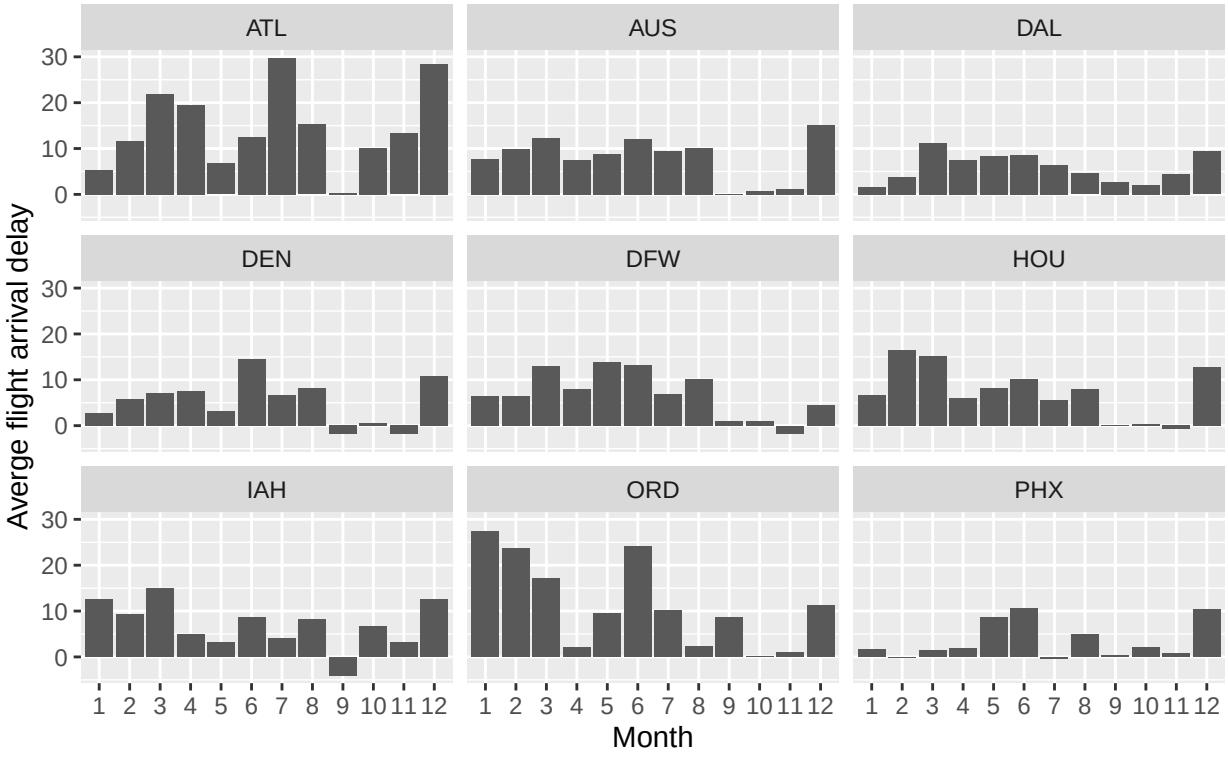


As shown in Figure 1.1, we have months of the year, January(1) through December(12) in X-axis and average arrival delays of the flights in Y-axis. The graph shows the average arrival delays of the flights for each month. We can see that across September(9), October(10) and November(11), the average flight delays are at the lowest with no delays in September. Therefore, we can conclude that the best time of the year to fly when one can avoid delays would be in 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 1.2 Average flight delays in various months of the year faceted by popular destinations



To examine whether this changes by destinations, I have faceted the bar plot by destination. As seen in the Figure 1.2, even considering various destinations, we can see that September, October and November have lower average arrival delays. Thus, we can conclude that this is the best time of the year to fly to minimize delays.

2) Wrangling the Olympics

A) 95th percentile of heights for female competitors across all Athletics events

```
## # A tibble: 2 x 2
##   sex   q95_height
##   <chr>      <dbl>
## 1 F            186
## 2 M            198
```

The 95th percentile of heights for female competitors across all Athletics events is 186 as shown on the table above

B) Which single women's event had the greatest variability in competitor's heights across the entire history of the Olympics, as measured by the standard deviation?

```
## # A tibble: 6 x 2
##   event                                standard_deviation
```

```
##   <chr>                                    <dbl>
## 1 Rowing Women's Coxed Fours               10.9
## 2 Basketball Women's Basketball             9.70
## 3 Rowing Women's Coxed Quadruple Sculls     9.25
## 4 Rowing Women's Coxed Eights               8.74
## 5 Swimming Women's 100 metres Butterfly     8.13
## 6 Volleyball Women's Volleyball             8.10
```

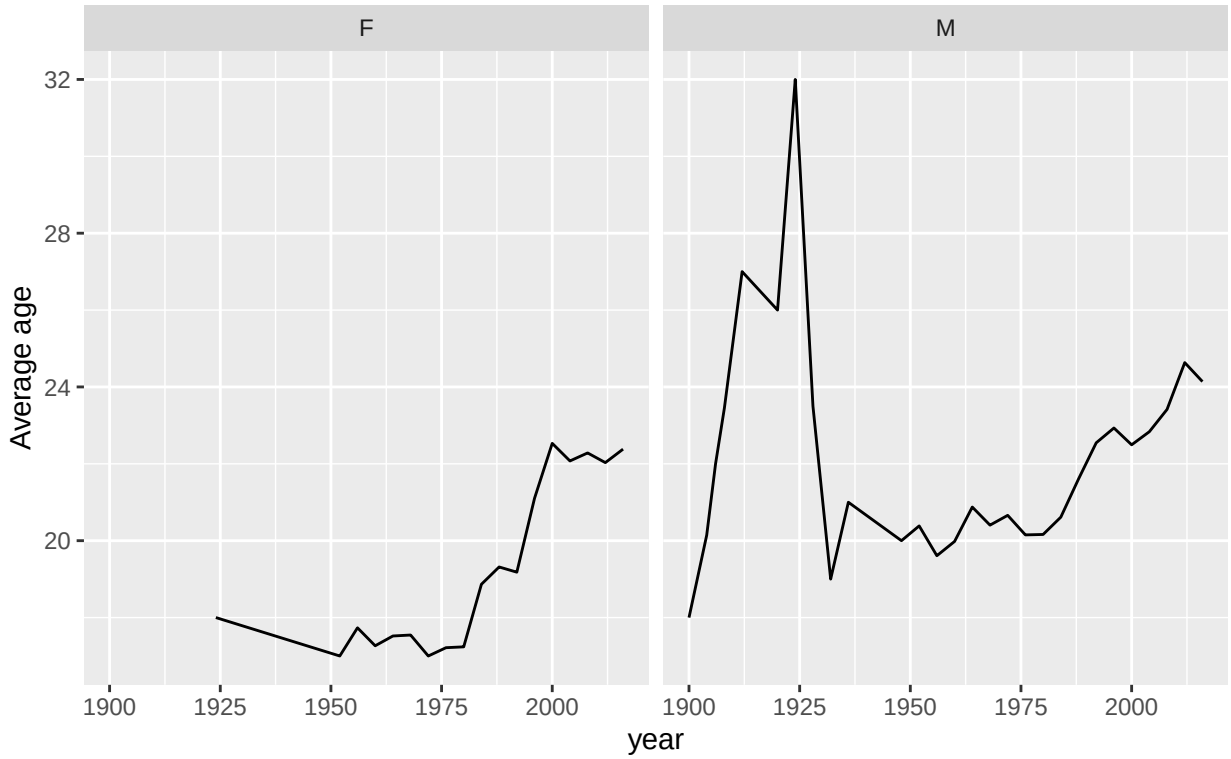
"Rowing Women's Coxed Fours" has the most variability(highest standard deviation) in competitor's height in women.

C) How has the average age of Olympic swimmers changed over time? Does the trend look different for male swimmers relative to female swimmers? Create a data frame that can allow you to visualize these trends over time, then plot the data with a line graph with separate lines for male and female competitors. Give the plot an informative caption answering the two questions just posed.

```
##    year sex  average
## 1  1900   M 18.00000
## 2  1904   M 20.14286
## 3  1906   M 22.00000
## 4  1908   M 23.45455
## 5  1912   M 27.00000
## 6  1920   M 26.00000
## 7  1924   M 32.00000
## 8  1928   M 23.50000
## 9  1932   M 19.00000
## 10 1936   M 21.00000
## 11 1948   M 20.00000
## 12 1952   M 20.38462
## 13 1956   M 19.60870
## 14 1960   M 19.98039
## 15 1964   M 20.87500
## 16 1968   M 20.40449
## 17 1972   M 20.65909
## 18 1976   M 20.15094
## 19 1980   M 20.16176
## 20 1984   M 20.61176
## 21 1988   M 21.60440
## 22 1992   M 22.54651
## 23 1996   M 22.93182
## 24 2000   M 22.49451
## 25 2004   M 22.83516
## 26 2008   M 23.41584
## 27 2012   M 24.63265
## 28 2016   M 24.13978
## 29 1924   F 18.00000
## 30 1952   F 17.00000
## 31 1956   F 17.73333
## 32 1960   F 17.26531
## 33 1964   F 17.52000
## 34 1968   F 17.54545
## 35 1972   F 17.00000
## 36 1976   F 17.21538
## 37 1980   F 17.23810
## 38 1984   F 18.86567
```

```
## 39 1988   F 19.31579
## 40 1992   F 19.18056
## 41 1996   F 21.10345
## 42 2000   F 22.53191
## 43 2004   F 22.07447
## 44 2008   F 22.28283
## 45 2012   F 22.03093
## 46 2016   F 22.38144
```

Figure 3. Average age of Olympic competitors over the years by sex

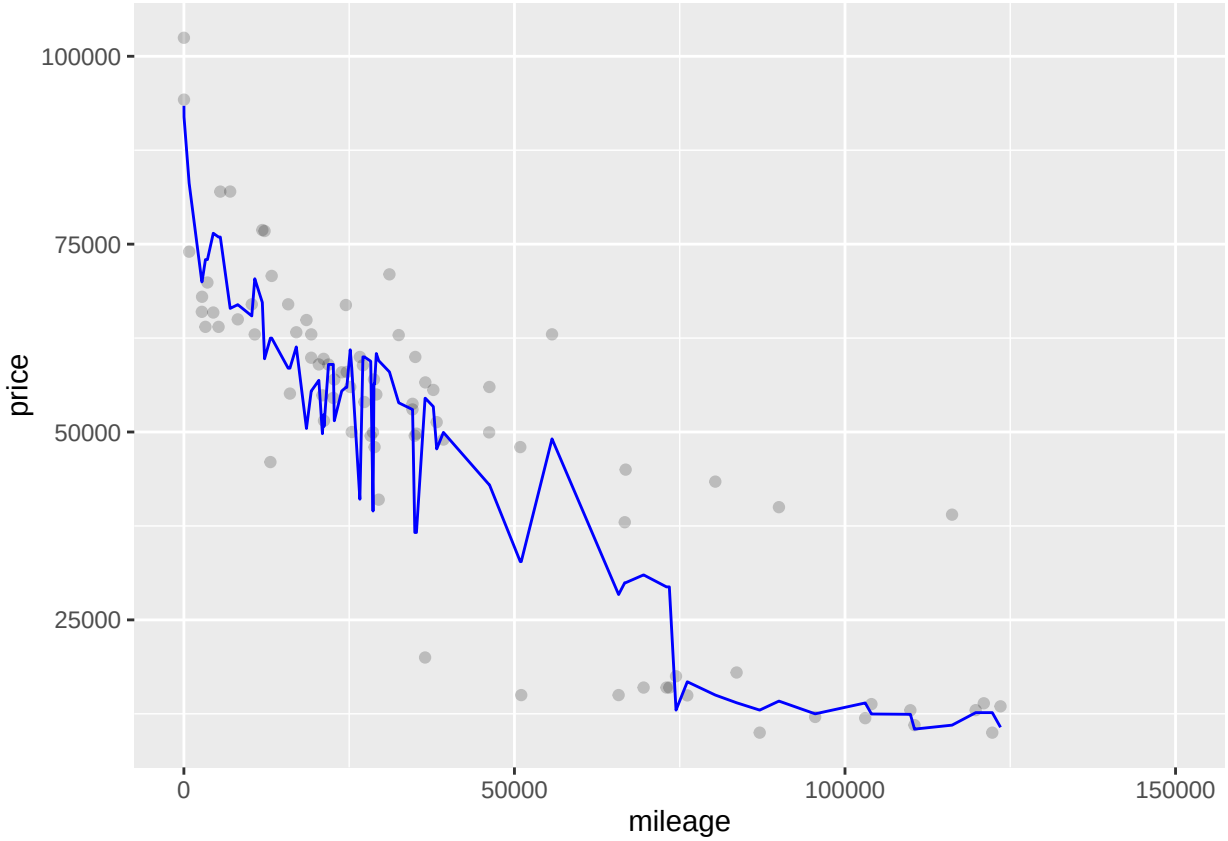


Since 1925, the average age of the Olympic competitors has been steadily increasing across male and female

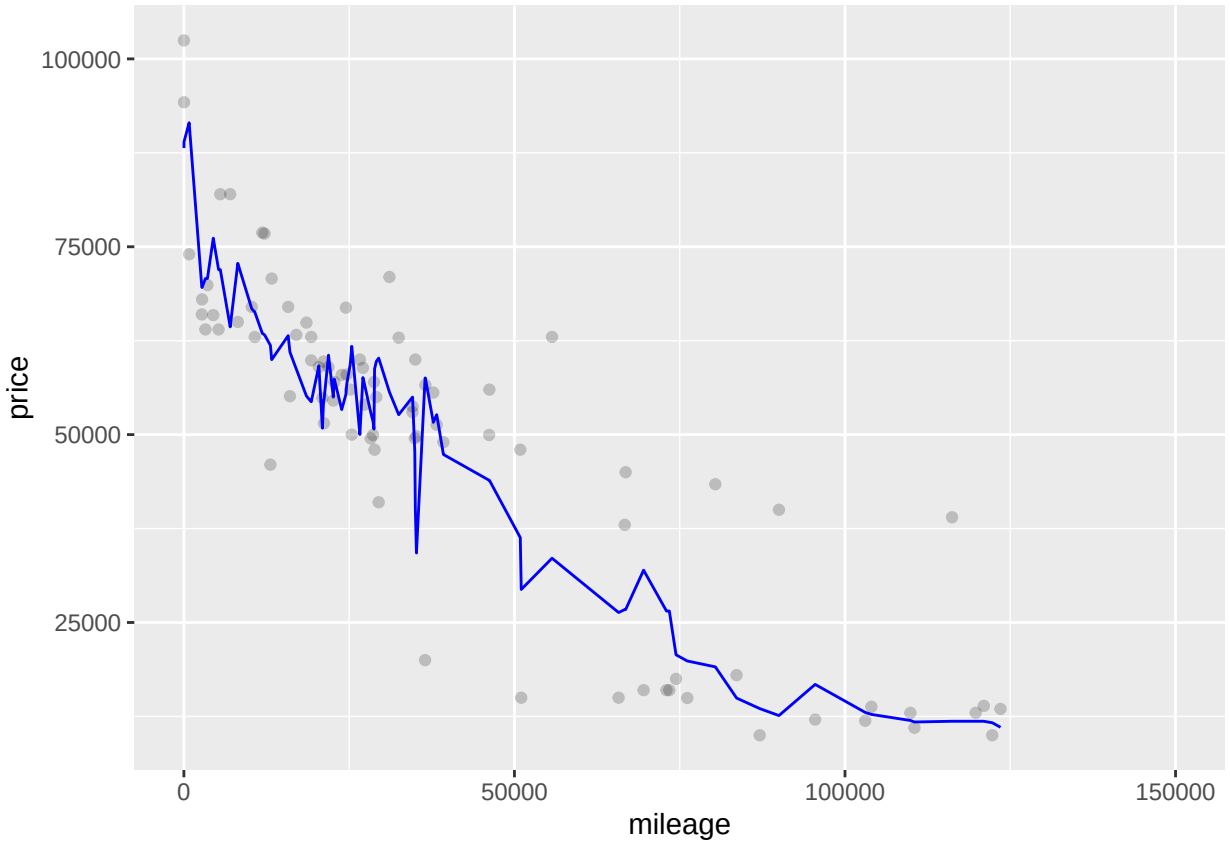
3) K-nearest neighbors: cars

For trim level 350

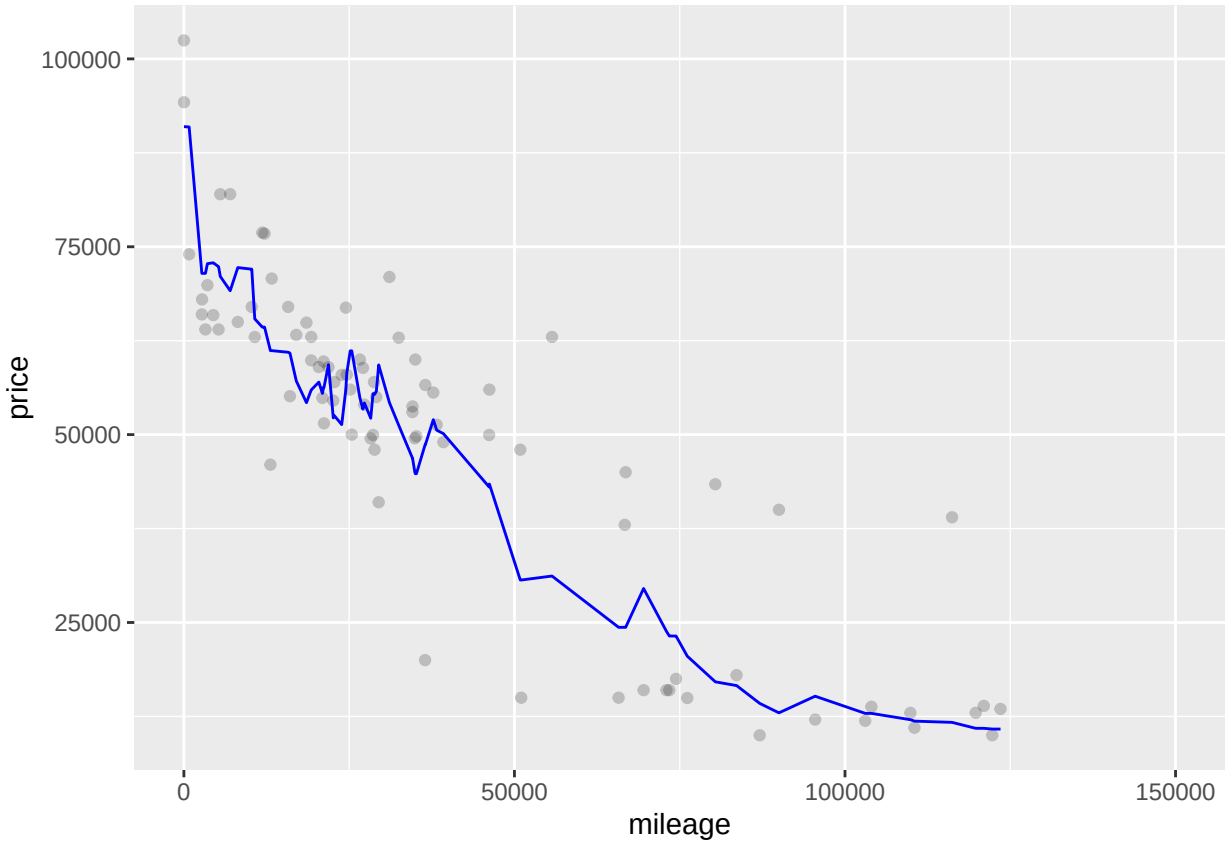
K=2



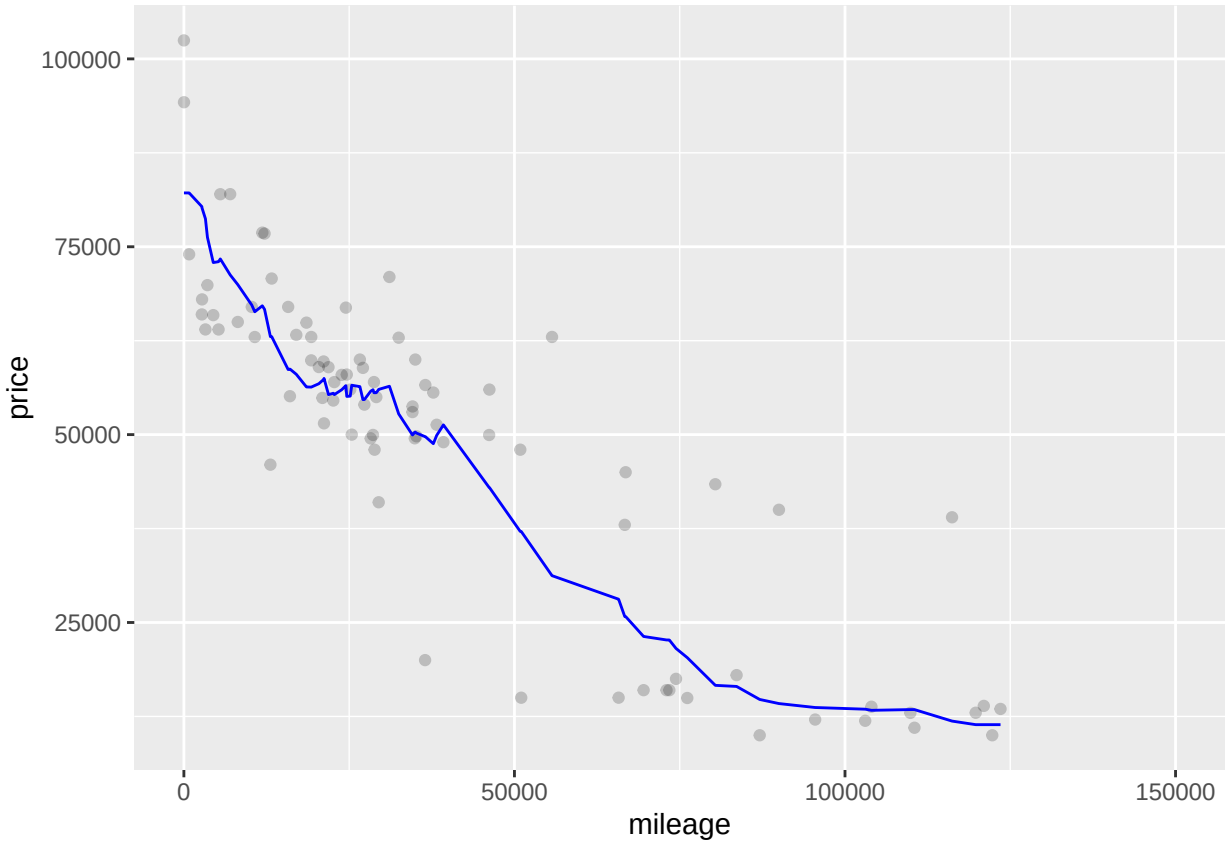
K=5


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

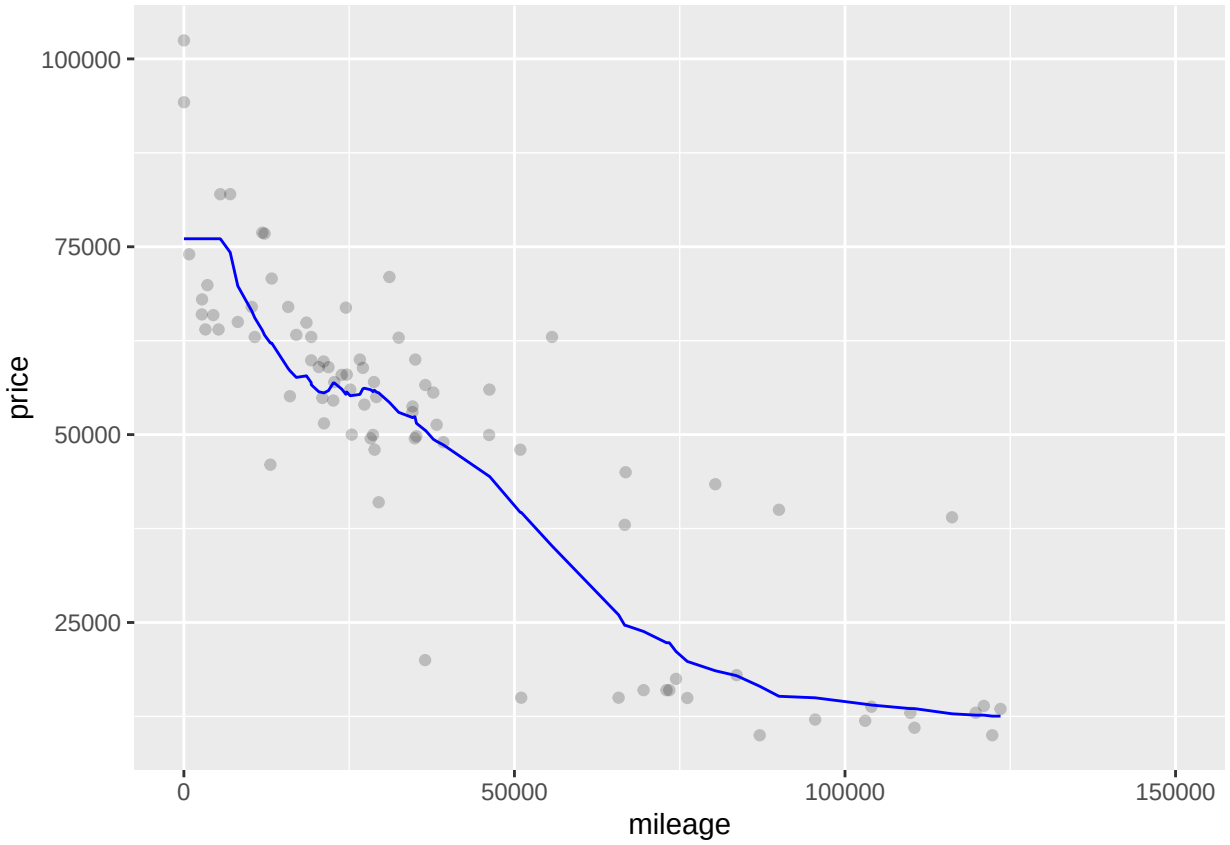
K=10


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

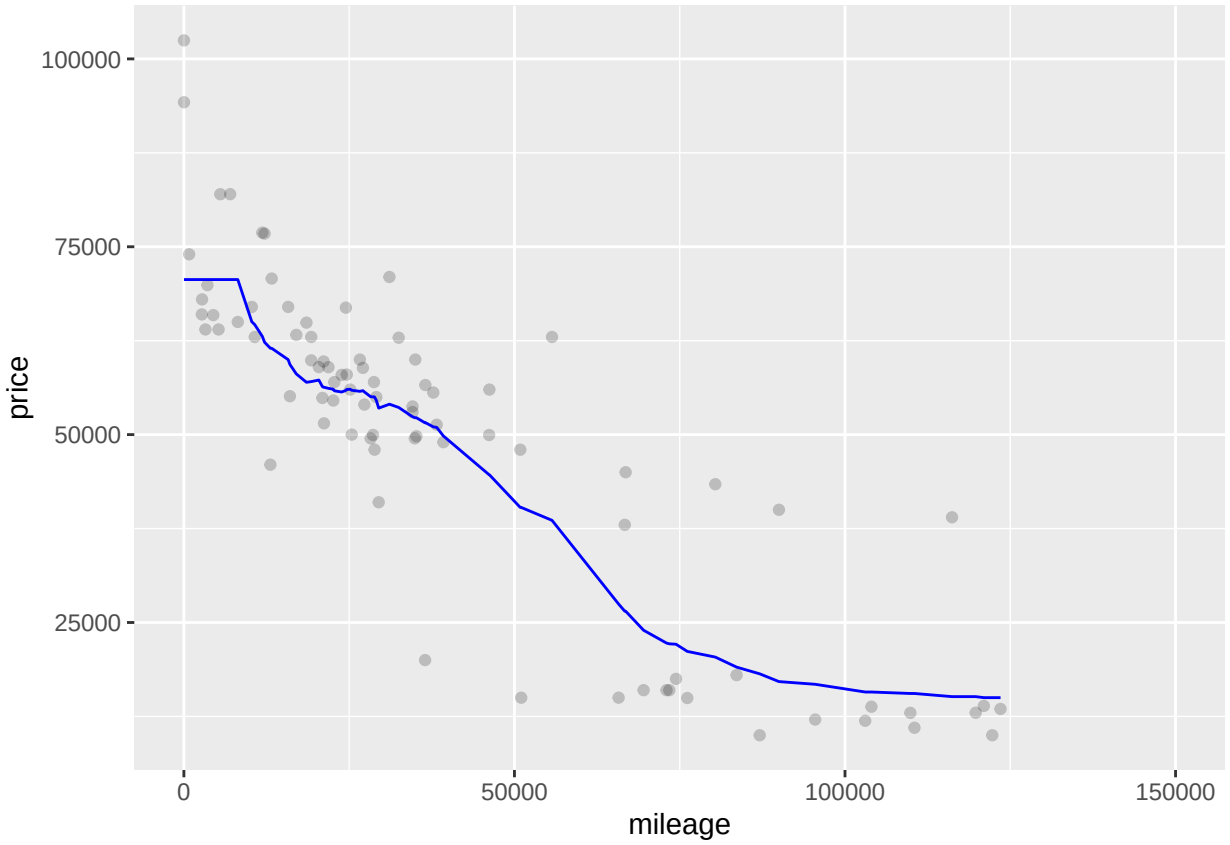
K=25


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

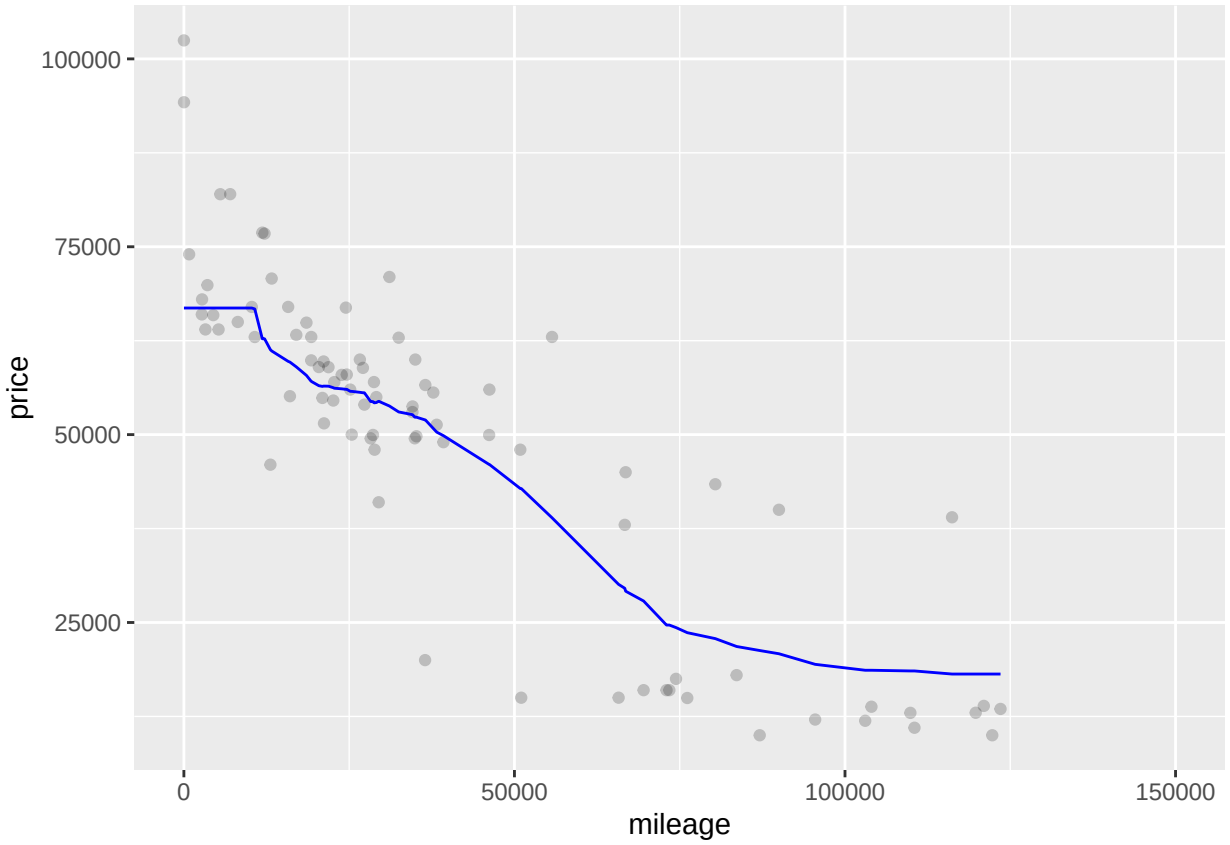
K=50

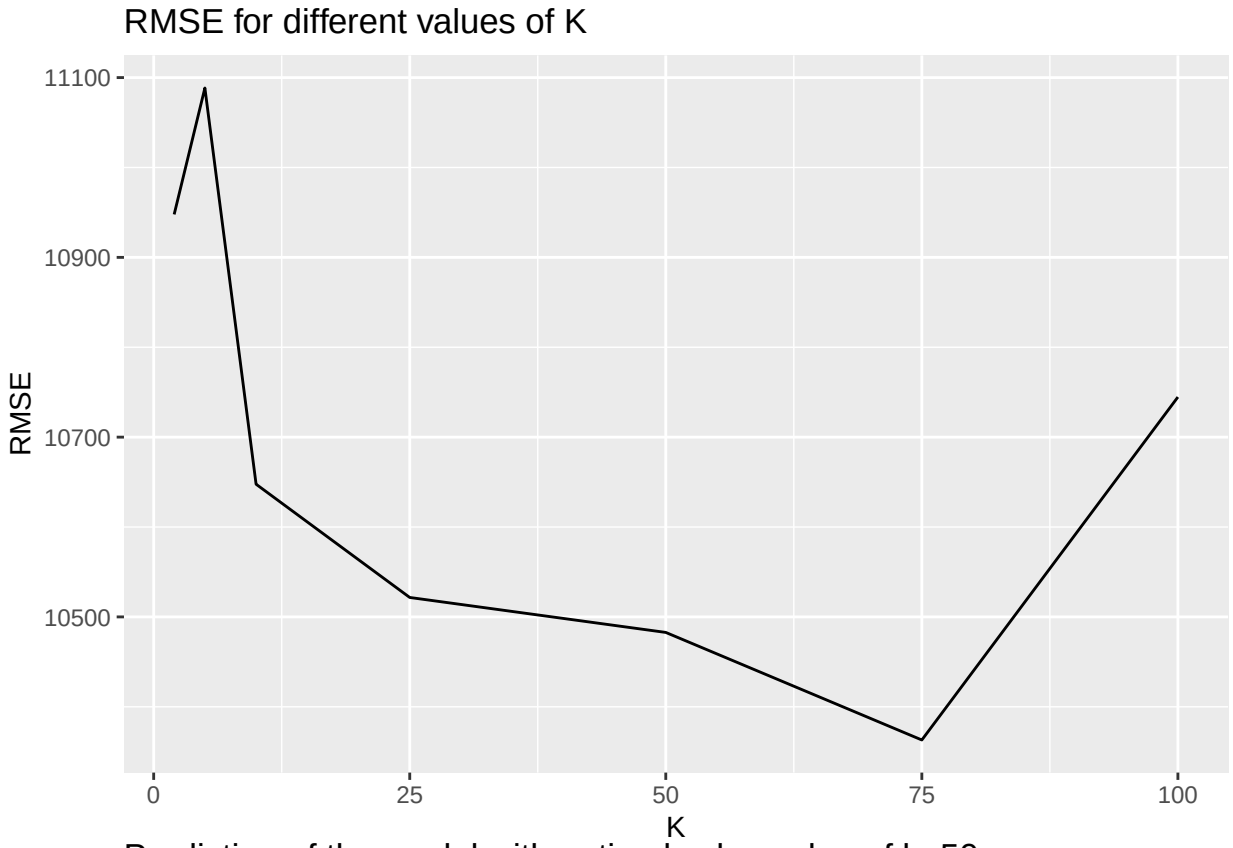

100000
75000
50000
25000
price
0
50000
100000
150000
mileage

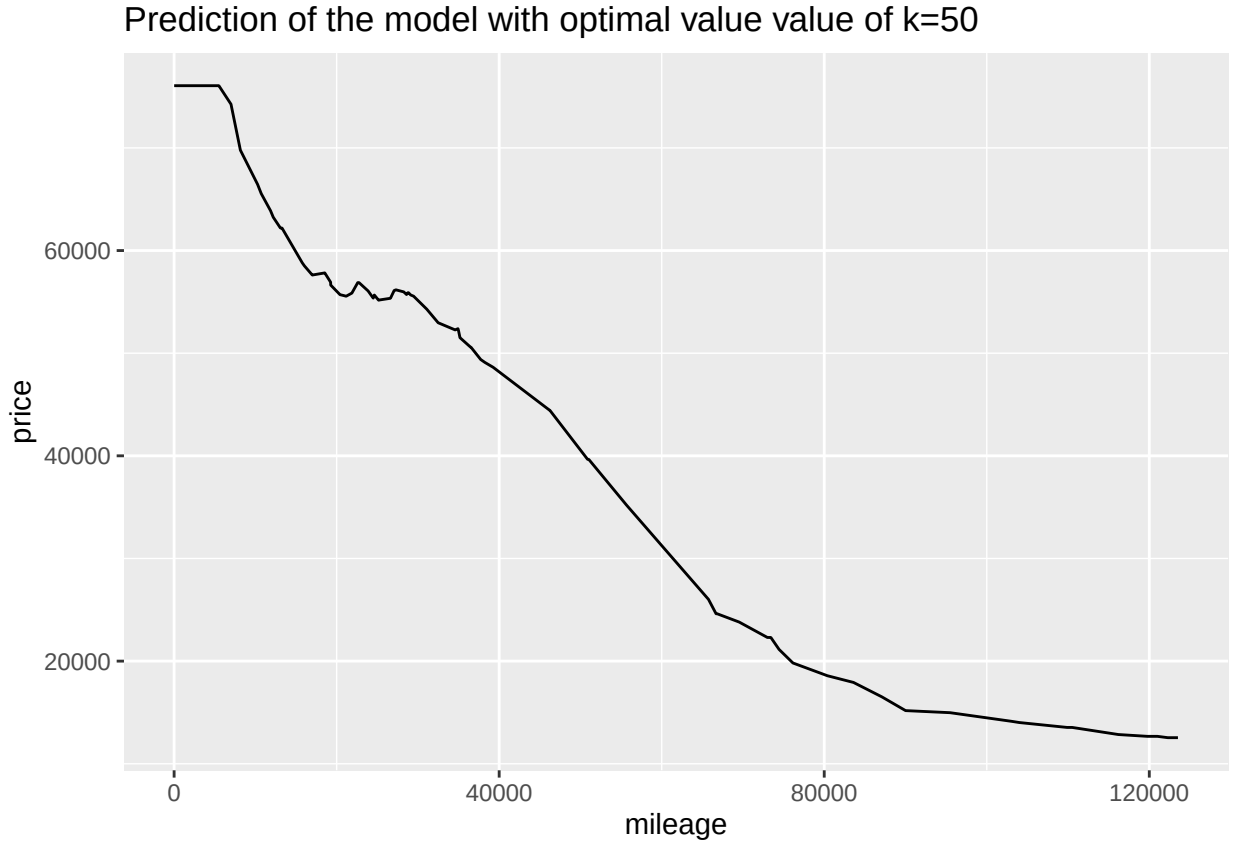
K=75


100000
75000
50000
25000
price
0
50000
100000
150000
mileage

K=100

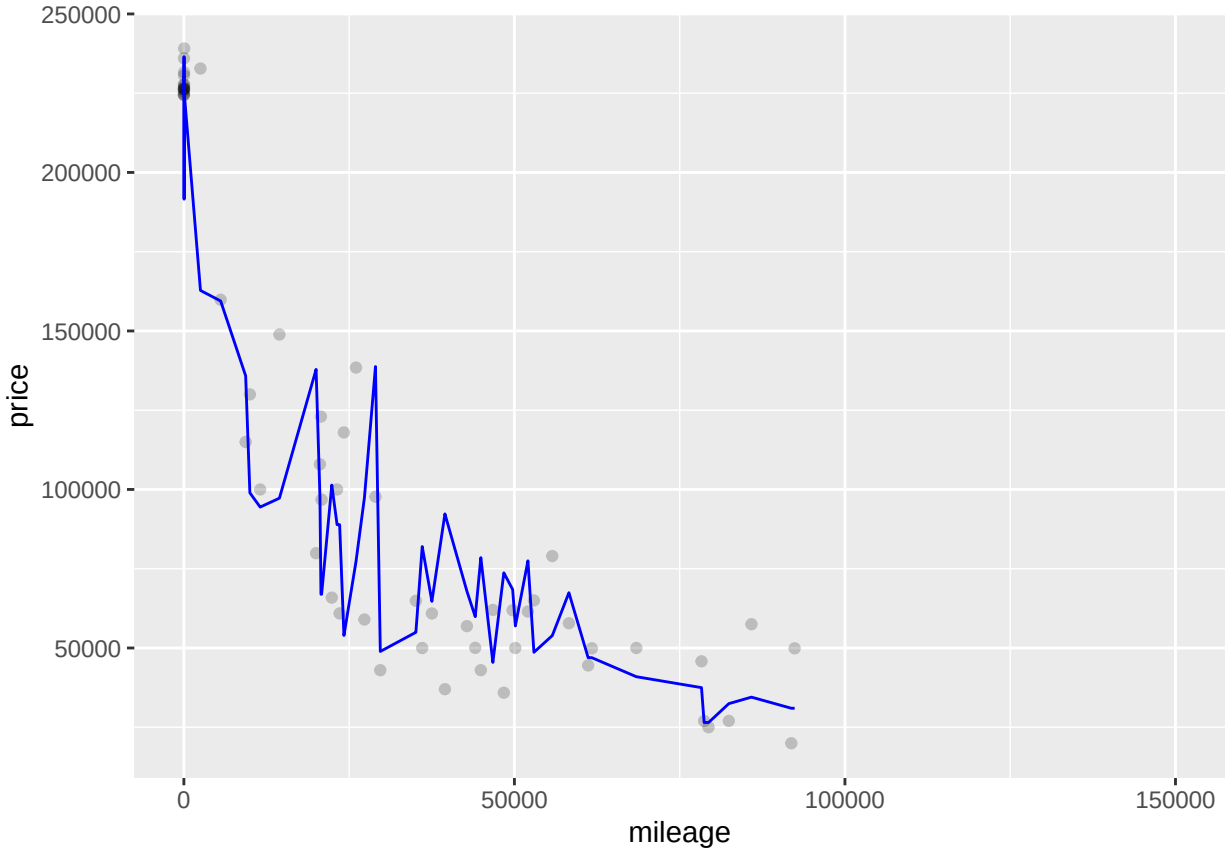

100000
75000
50000
25000
price
0
50000
100000
150000
mileage


RMSE for different values of K
11100
10900
10700
10500
RMSE
0
25
50
75
100
K

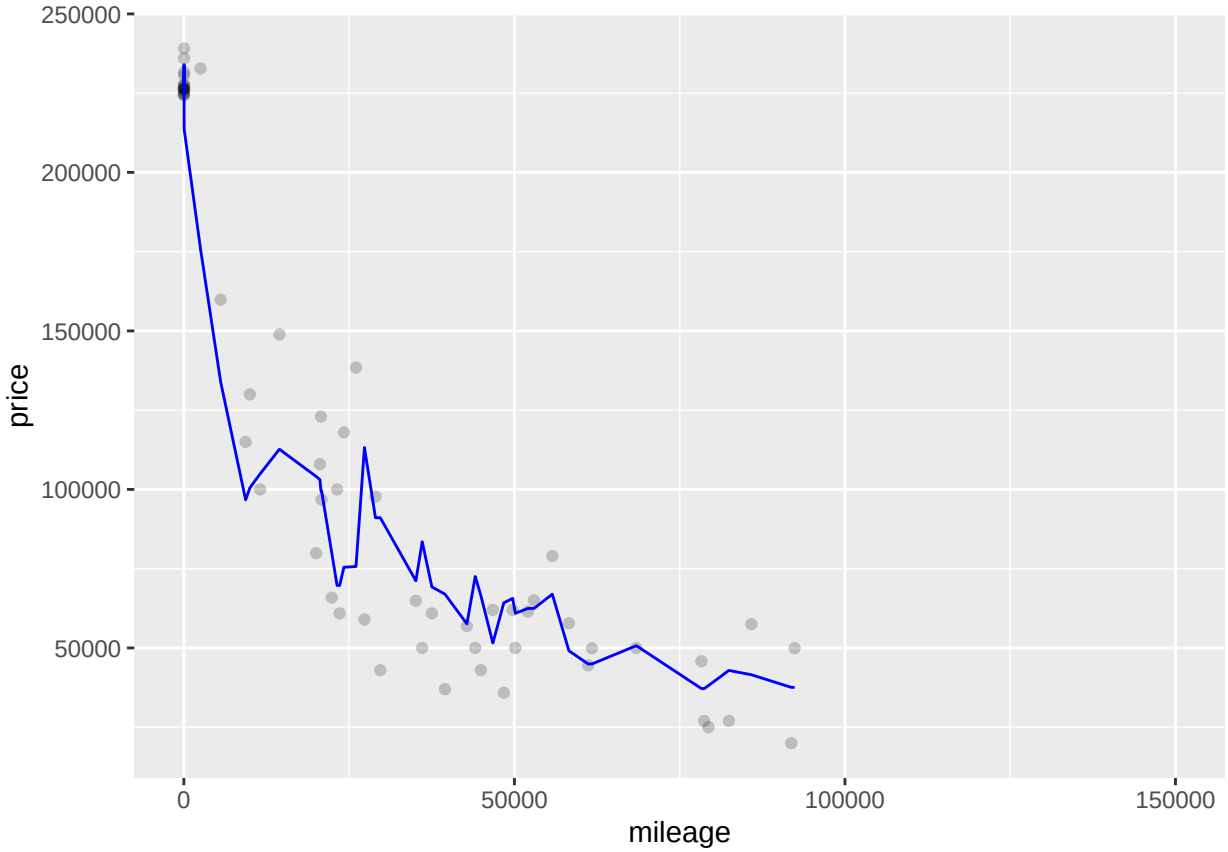

Prediction of the model with optimal value value of k=50
60000
40000
20000
price
0
40000
80000
120000
mileage

For trim level 65AMG

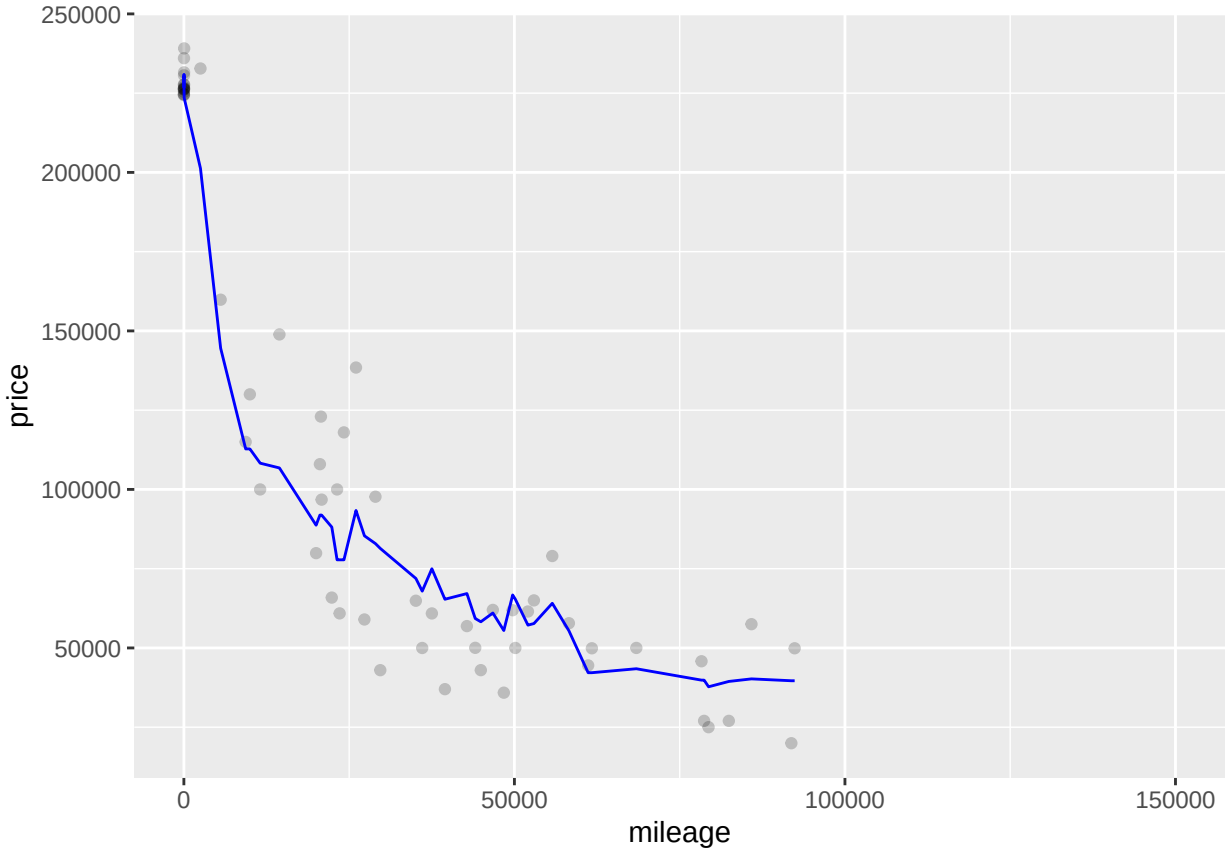
K=2



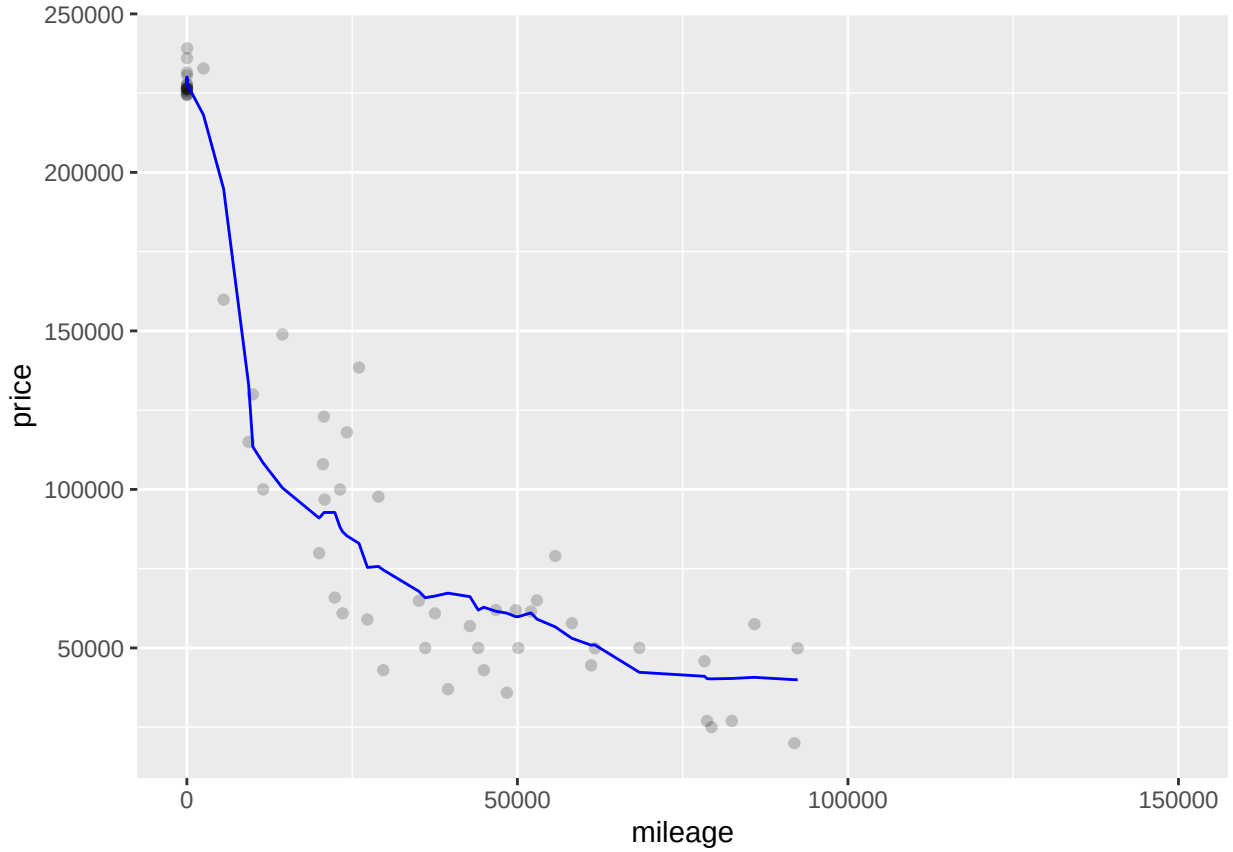
K=5


250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

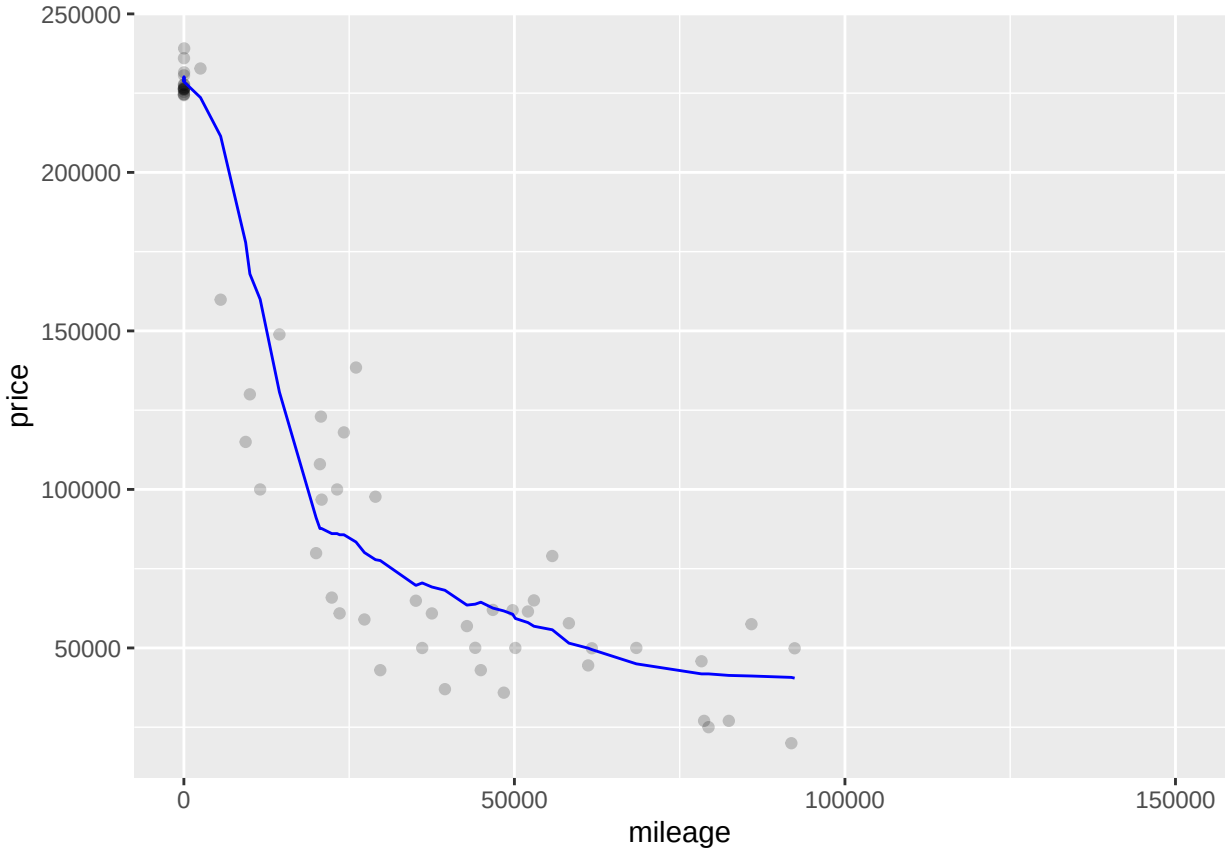
K=10


250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

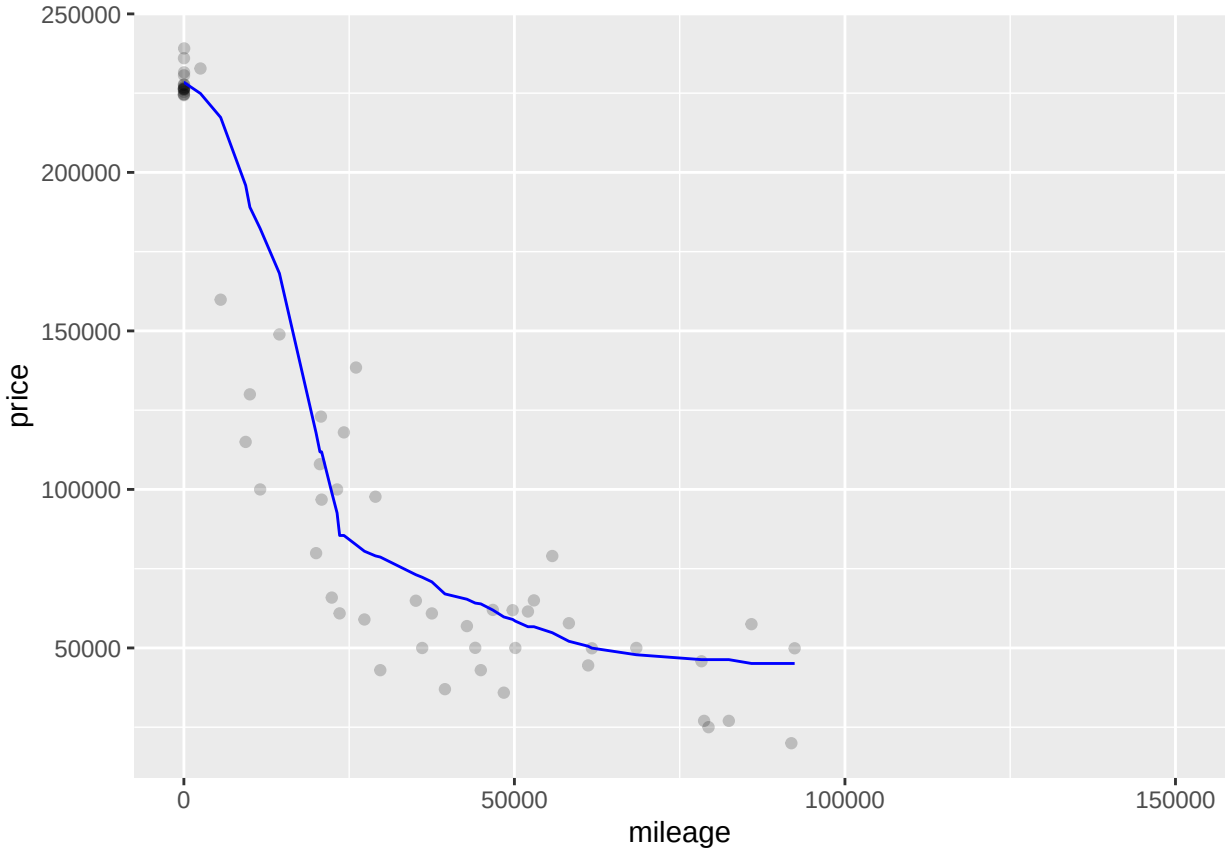
K=25


250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

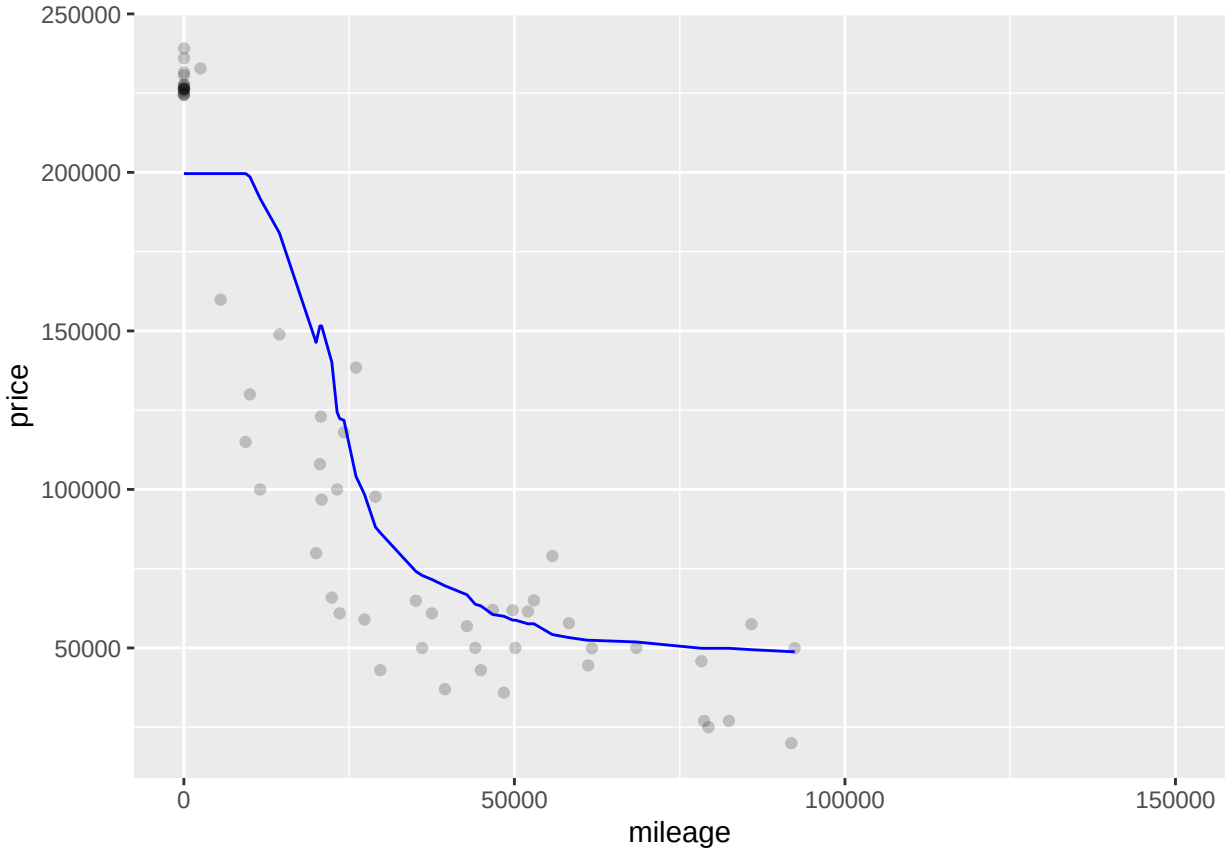
K=50

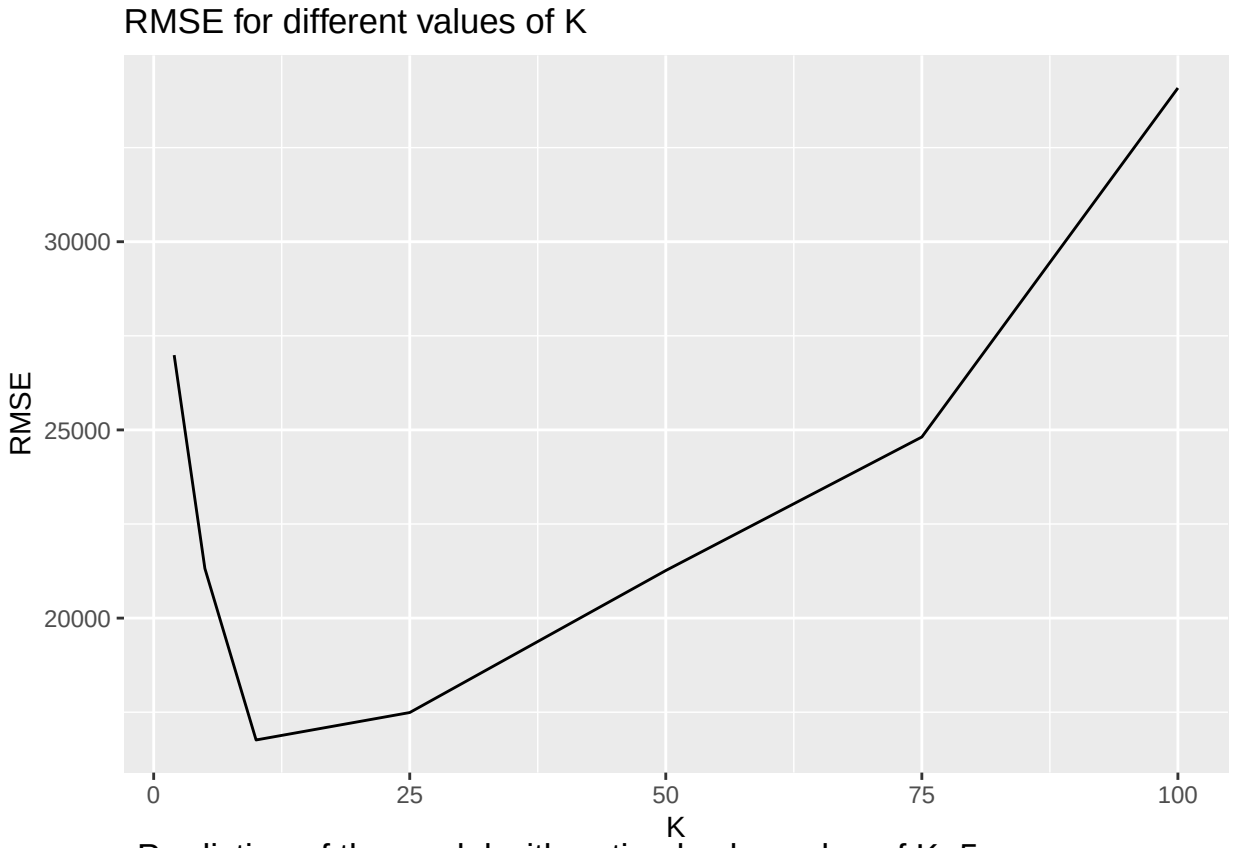

250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

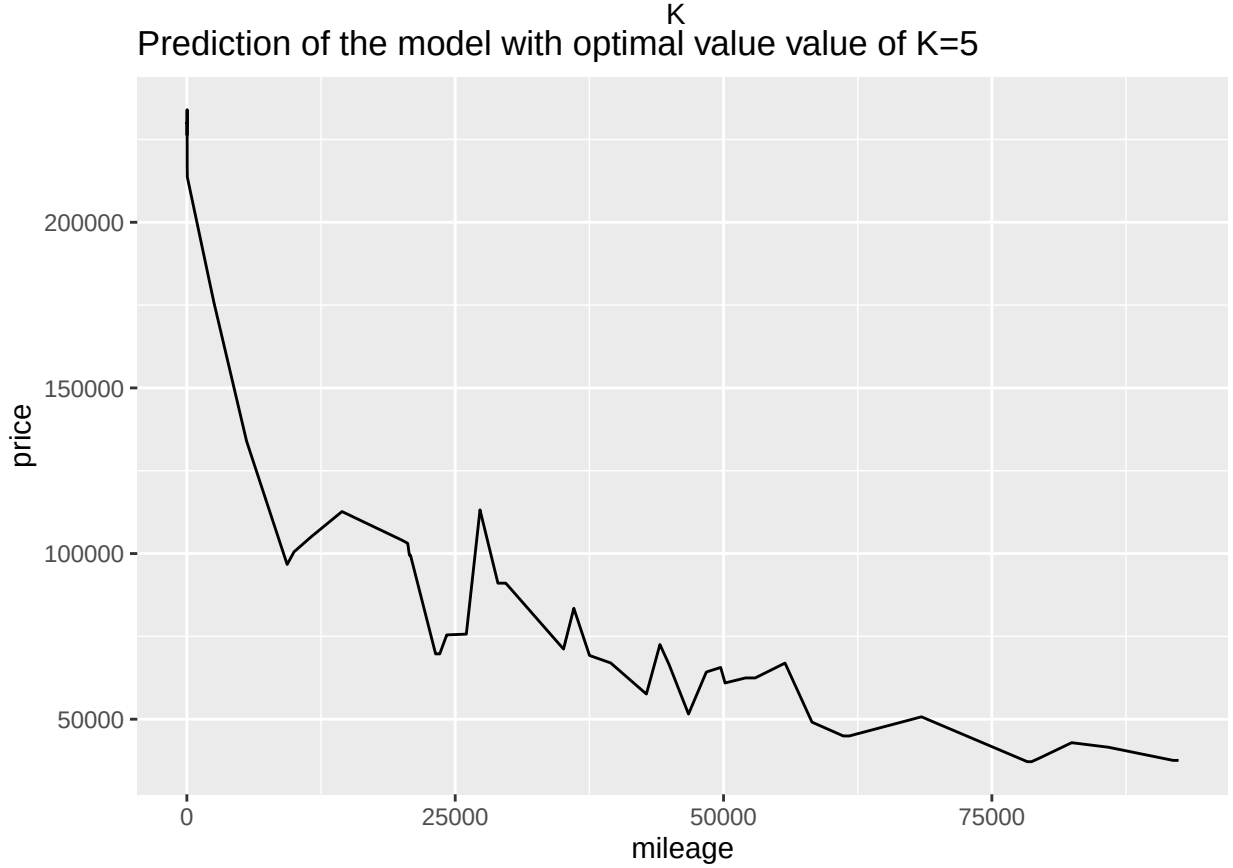
K=75


250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

K=100


250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage





Trim size 350 produces 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.