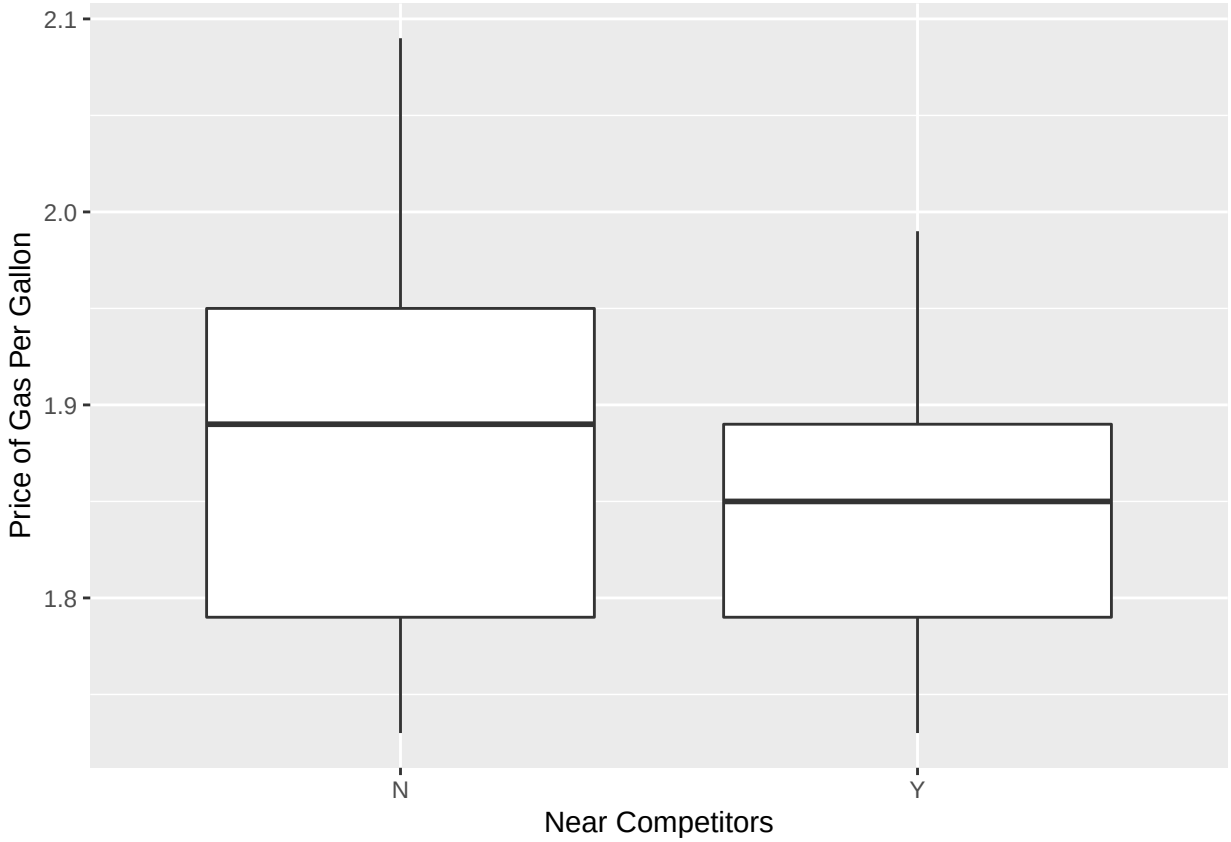
DataMining-HW1

Joey Herrera

1/28/2021

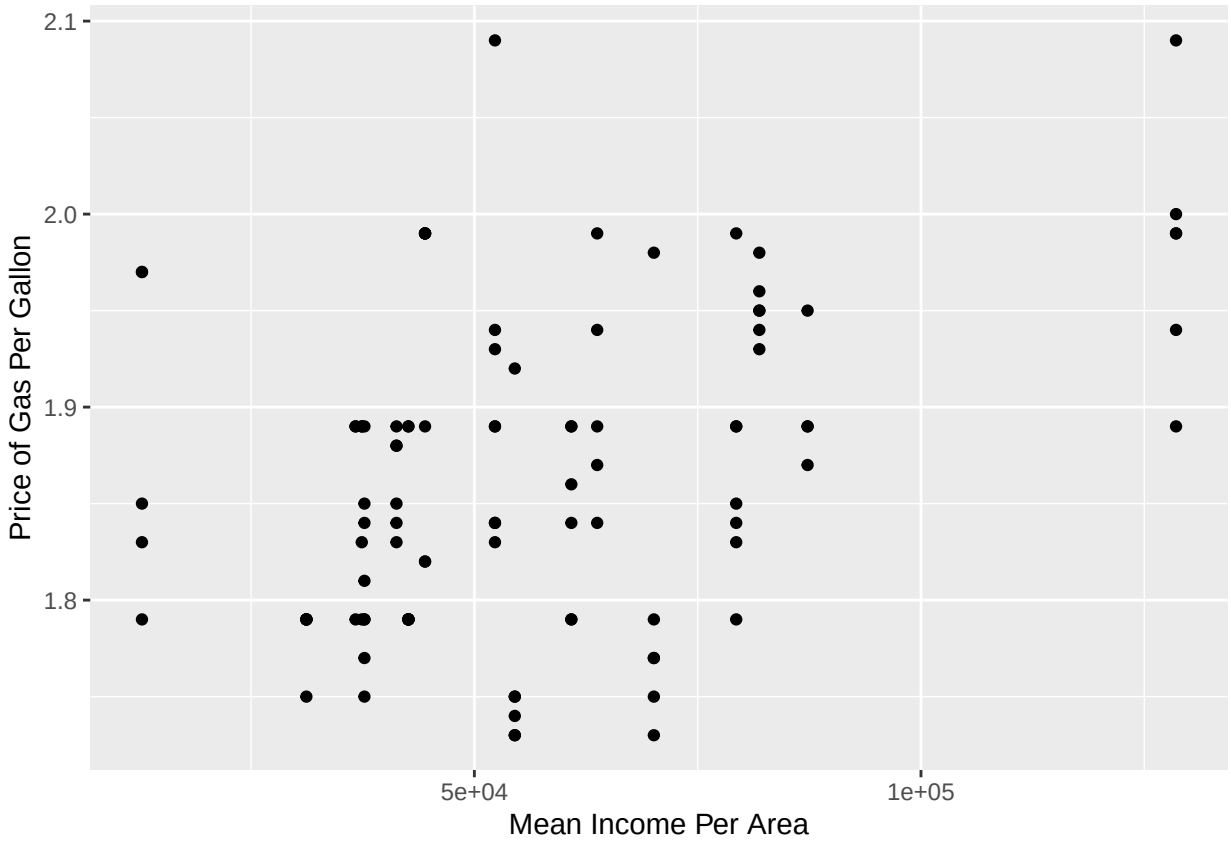
Question 1

1A



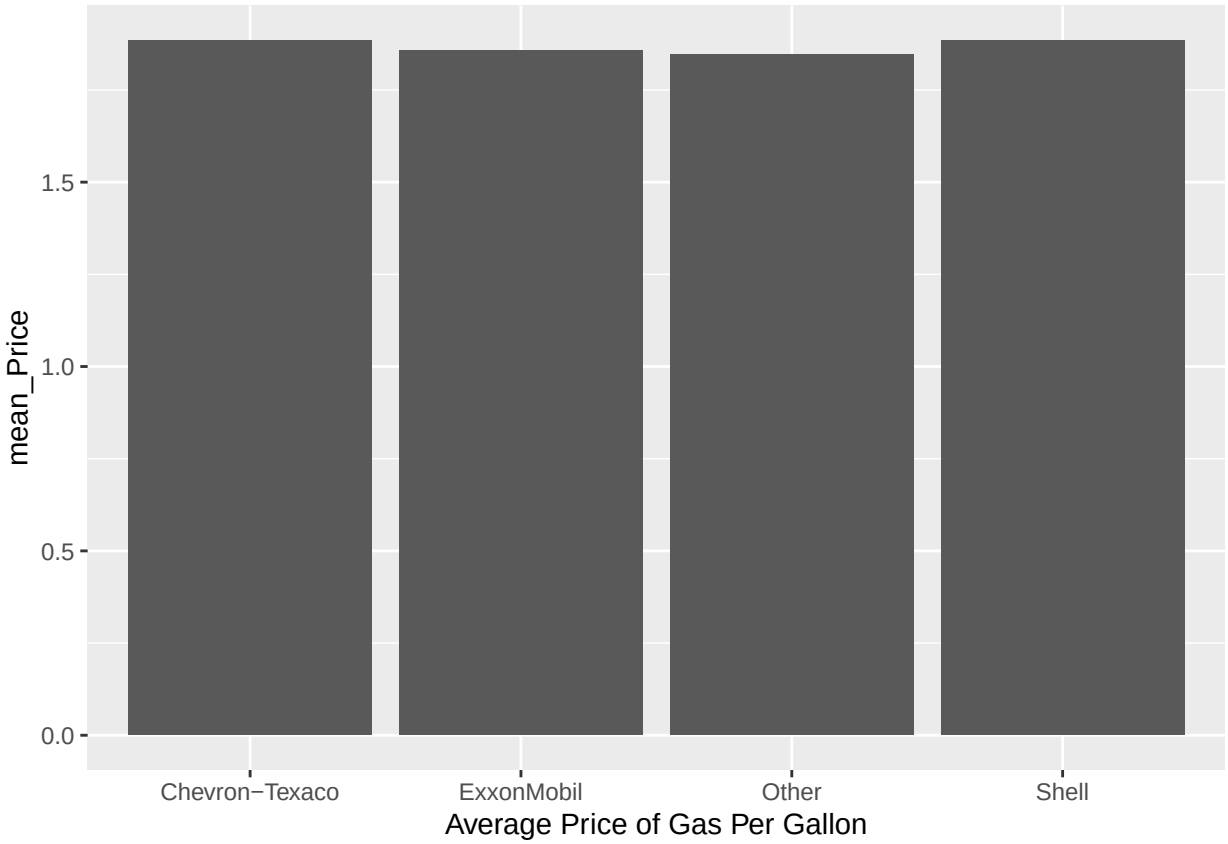
The theory claiming as stations charge more if they lack direct competition in sight seems unplausible. I believe that gas stations will charge less for gas if they are near competitors because both gas stations are fighting for the same customers. As a result, they would try to lower gas prices below the competition to attract more customers. The boxplots show that gas stations tend to charge more money when competitors are not in sight. The first quartile is equivlent for gas stations pricing efforts regardless of if competitors are in sight. The mean and third quartile are greater for gas stations who are not near competitors than gas stations who are near gas stations.

1B



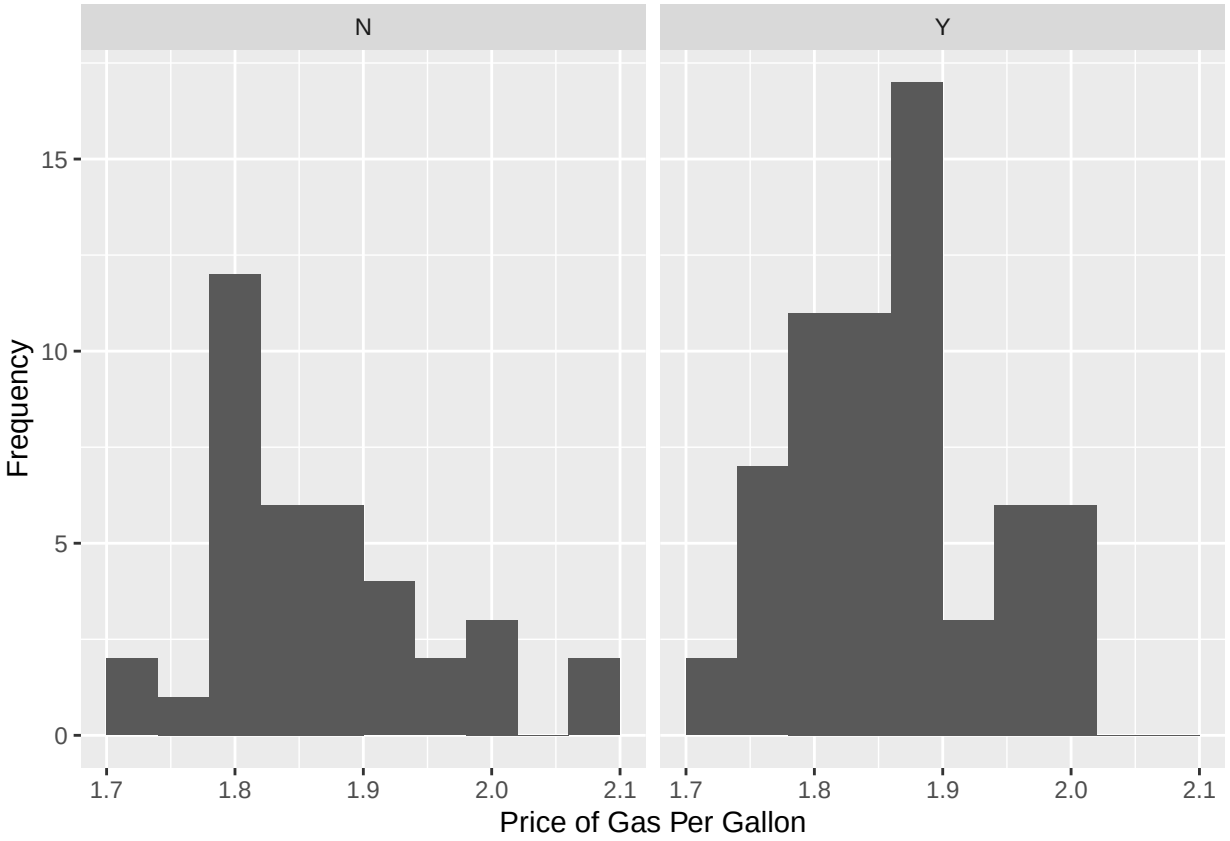
The premise that richer areas charge more for gas has merit because real estate is more expensive in wealthy neighborhoods. Gas stations in wealthy areas would need to charge more to pay rent and keep their profit margins at an optimal level. The scatterplot suggests that there is little evidence to support the claim richer areas charge higher gas prices. Each area, regardless of the median income, seems to have varying prices charged for gas. In low-income areas gas prices can be similar to high-income areas at almost two dollars per gallon.Moderate-income areas see similar gas prices to the majority of low-income areas between 1.80 and 1.90 dollars.

1C



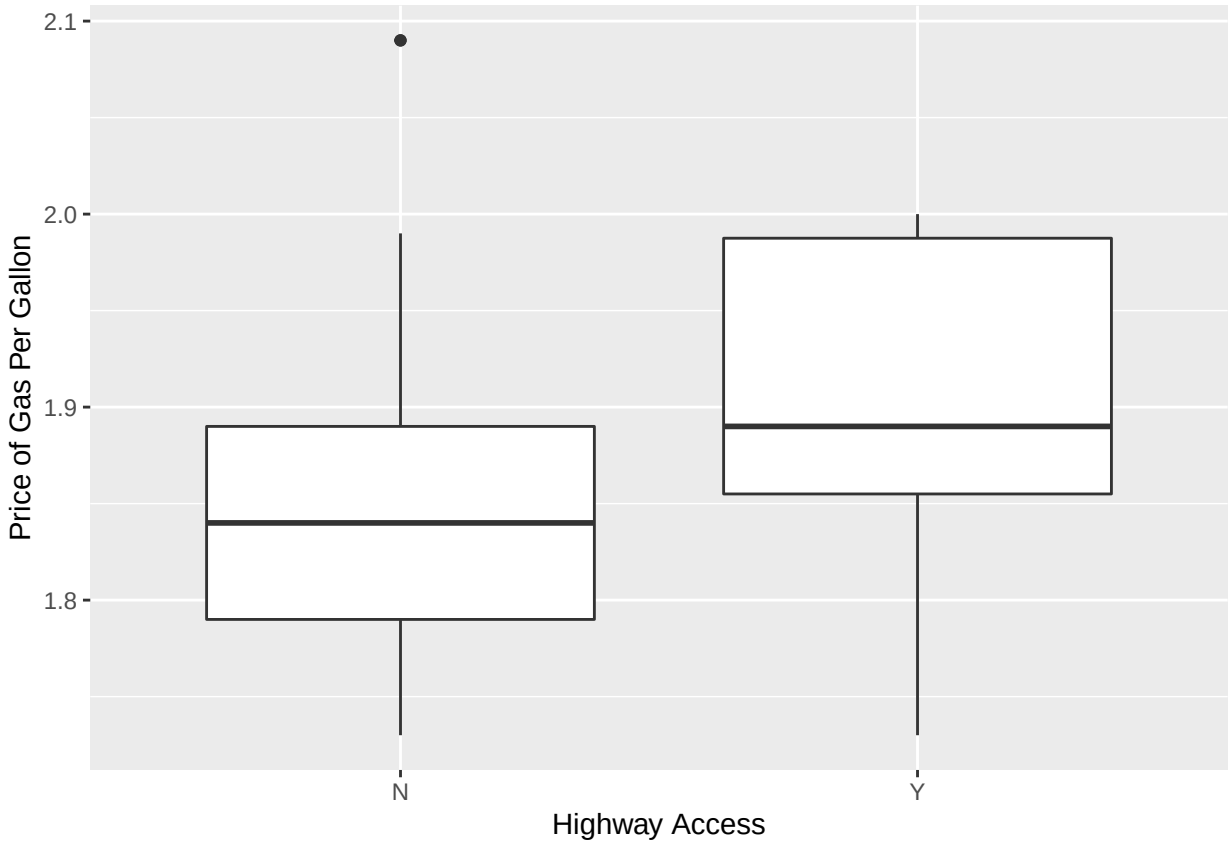
The theory that Shell charges more for the same type of gas seems unplausible (assuming ceteris paribus) because customers would not buy gas from Shell unless no other gas stations were in driving distance. This theory is unsupported by the data. The bar plots indicate that Shell charges comparable prices to every other gas station for a gallon of gas on average. Shell and Cheveron-Texaco seem to charge identical prices and the rest of the gas stations charge slightly less on average.

1D



Gas stations that are near stoplights charging more money seems plausible because the increase in price can be seen as a convieenence fee. There are fewer gas stations around stoplights than there are gas stations near stoplights. The histogram for gas stations not near stoplights have a higher proportion of gas stations charging over $1.80 per gallon than gas stations close to stoplights. The majority of gas stations close to stoplights charge less than 1.80 dollars per gallon. In conclusion, the figure does not support the claim that gas stations near stoplights charge more for gas than gas stations not near a stoplight.

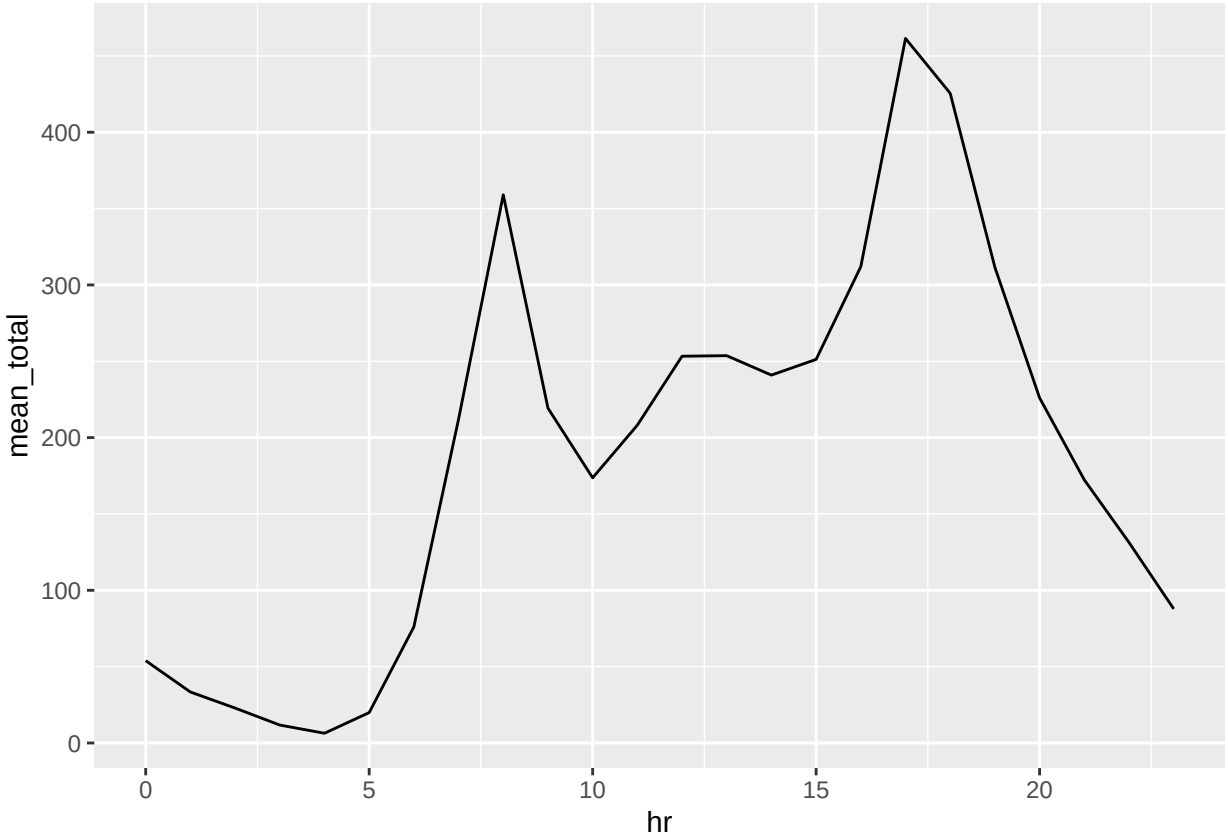
1E



I believe that gas stations with highway access charge more than other gas stations because of convenience. The boxplot for gas stations without highway access has a first, second, and third quartile that are lower than gas stations with highway access. Gas stations with highway access have a third quartile that is approximately 10 cents greater than other gas stations. The mean price of gas per gallon at gas stations with highway access is about 5 cents greater than other gas stations. In conclusion, the boxplots support the claim that gas stations with highway access charge more than other gas stations.

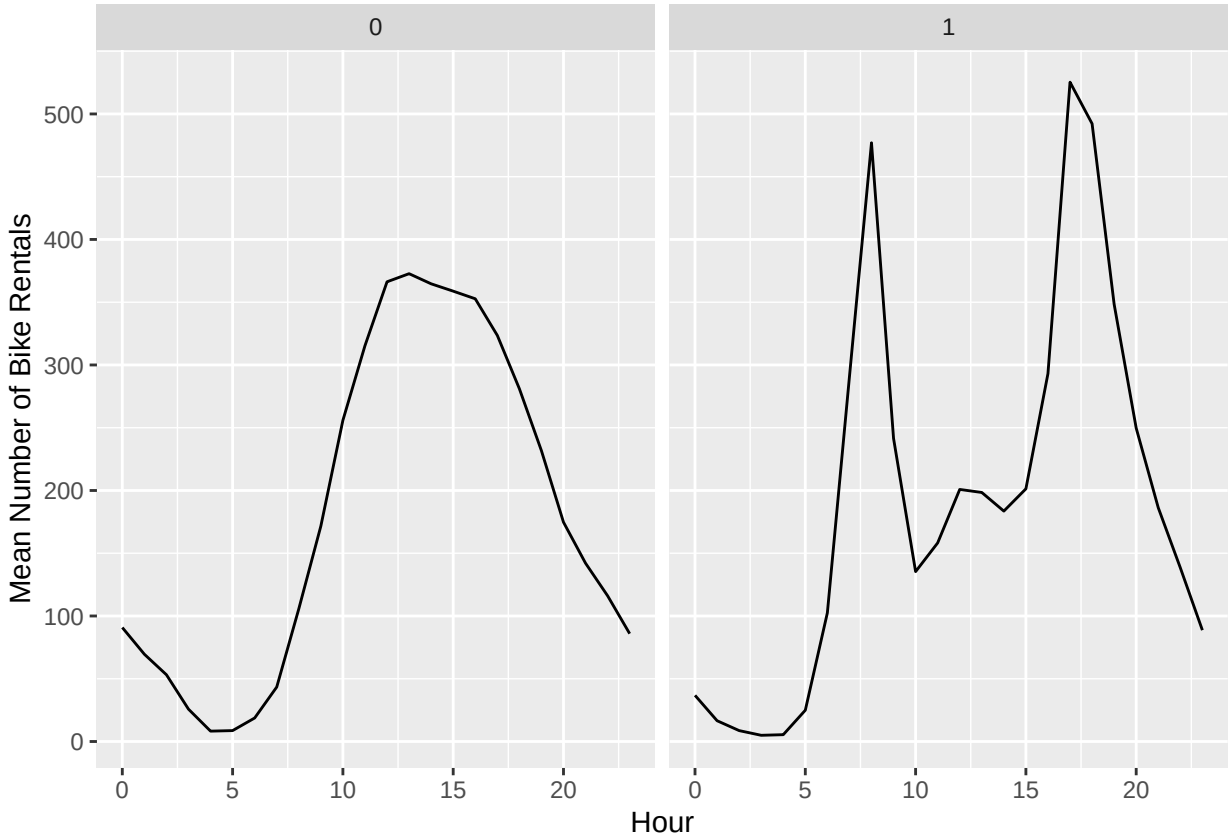
Question 2

2A



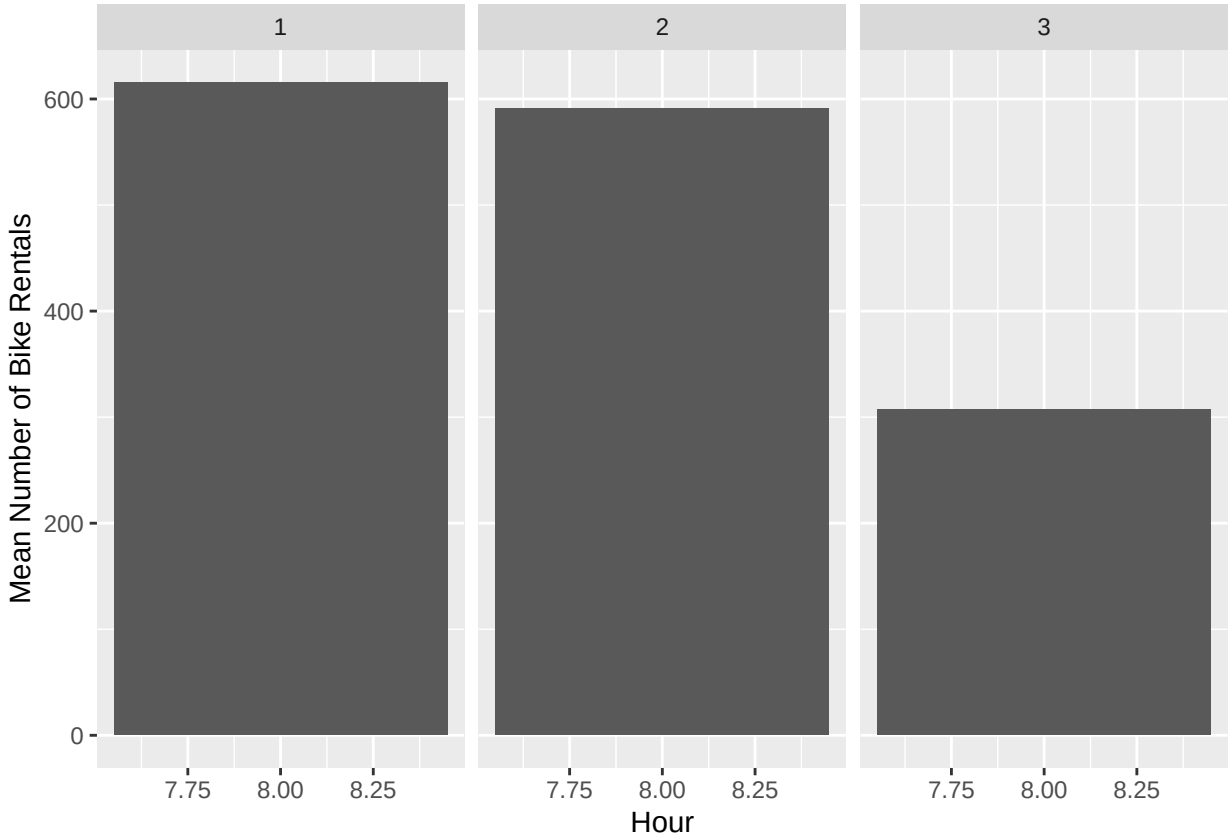
This line graph shows the average bike rentals from hour to hour throughout the day. The average number of bike rentals is graphed on the y-axis. The x-axis shows the hours of the day starting with 0 and ending with 23. This line graph illustrates that the most popular hours for bike rentals on average are 8 and 17.

2B



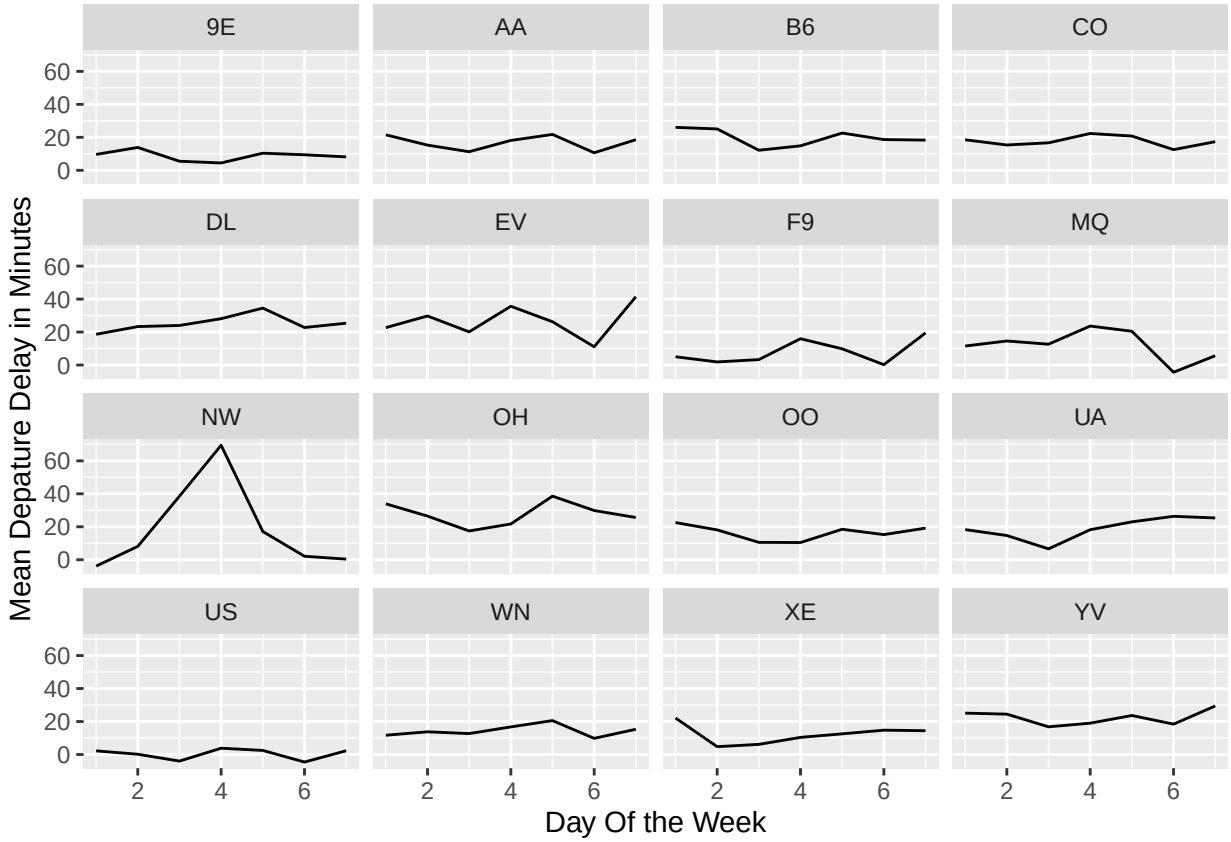
The two line graphs depict the average number of bike rentals per hour during days in the work week and on the weekend. The line graphs plot hour of day the x-axis, starting with 0 and ending with 23. The y-axis plots the average number of bike rentals. These line graphs illustrate that most rentals during the work week occur before and after the typical work day, and that bike rental numbers spike during mid-day on weekends.

2C



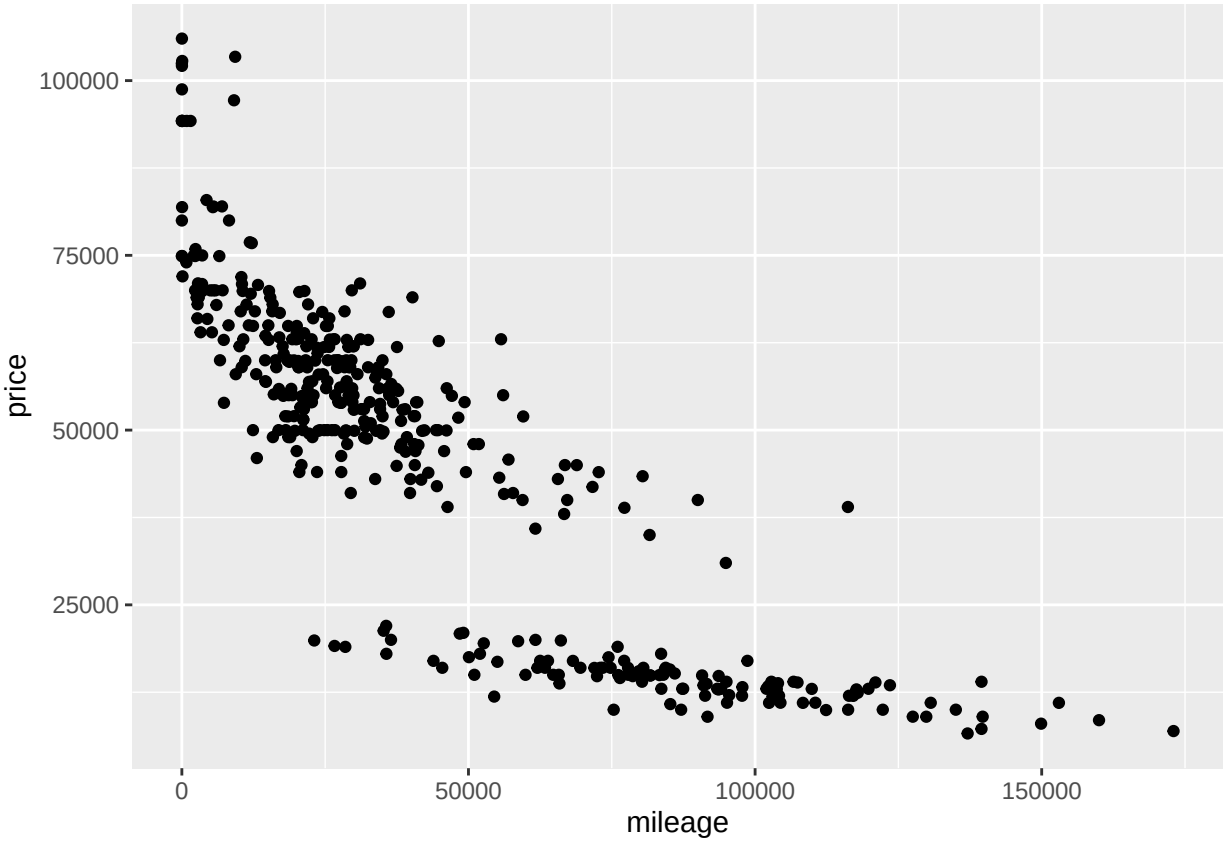
The plot above illustrates the average total of bike rentals per weekday on the y-axis. The x-axis depicts the hour of the day, and each barplot represents a different weather situation. When weather situation equals 1 there is nice weather. When weather situation equals two there is mist and clouds. Finally, weather situation equals three when there is light snow, light rain, or a thunderstorm. The barplots illustrate that on average most bike rentals at 8 am on a workday with clear weather, followed closely by misting weather and there is a substancial decrease in bike rentals when there is significant percipitation.

Question 3



The plots above depict the average delay time on the y-axis and the day of the week on the x-axis, where 1 equals Monday, 2 equals Tuesday, etc. Each line graph depicts a different airline. Most of the airlines have small differences in departure delays regardless of the day of the week. The only airlines of interest are EV, MQ, and NW. If you fly NW you should avoid flying on a Tuesday, Wednesday, or Thursday. If you are flying on EV or MQ you should also avoid mid-week days. You should also avoid flying EV on Sunday.

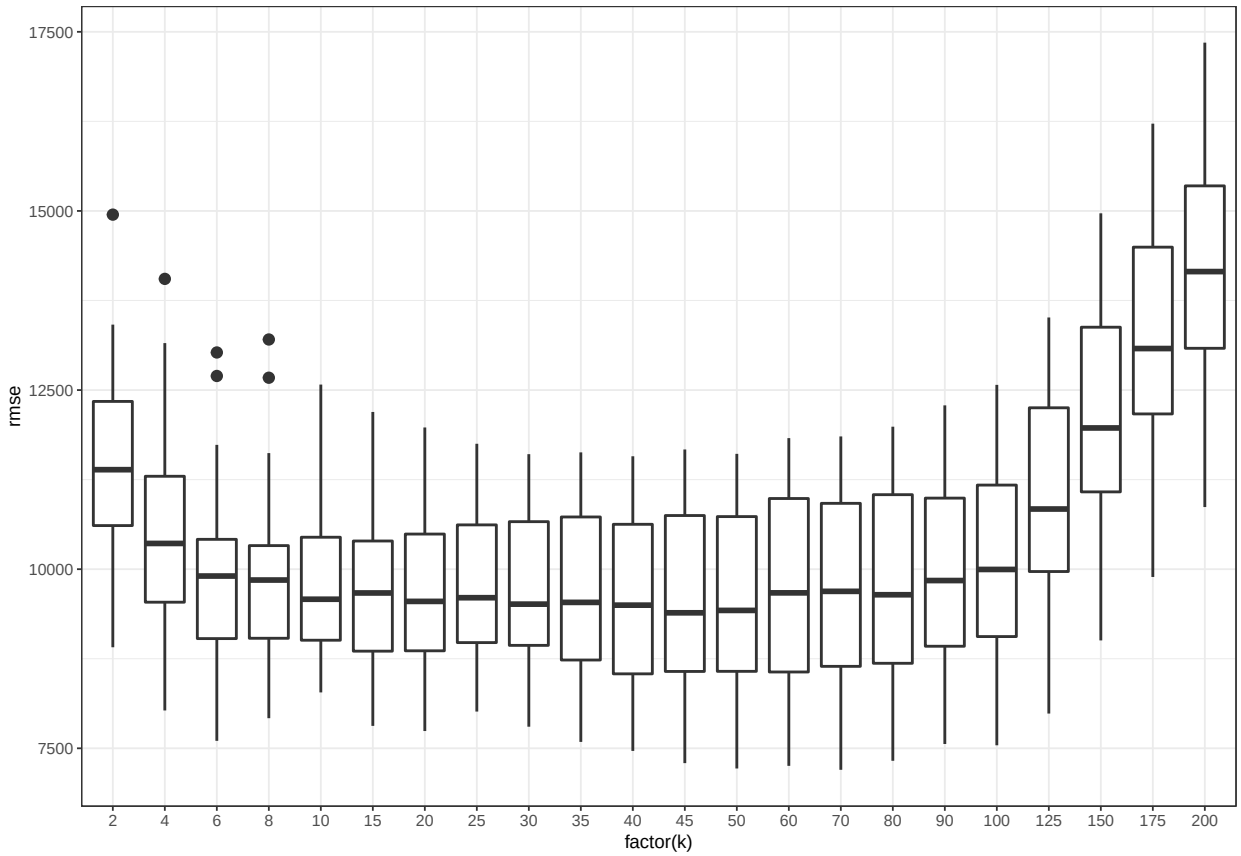
Question 4

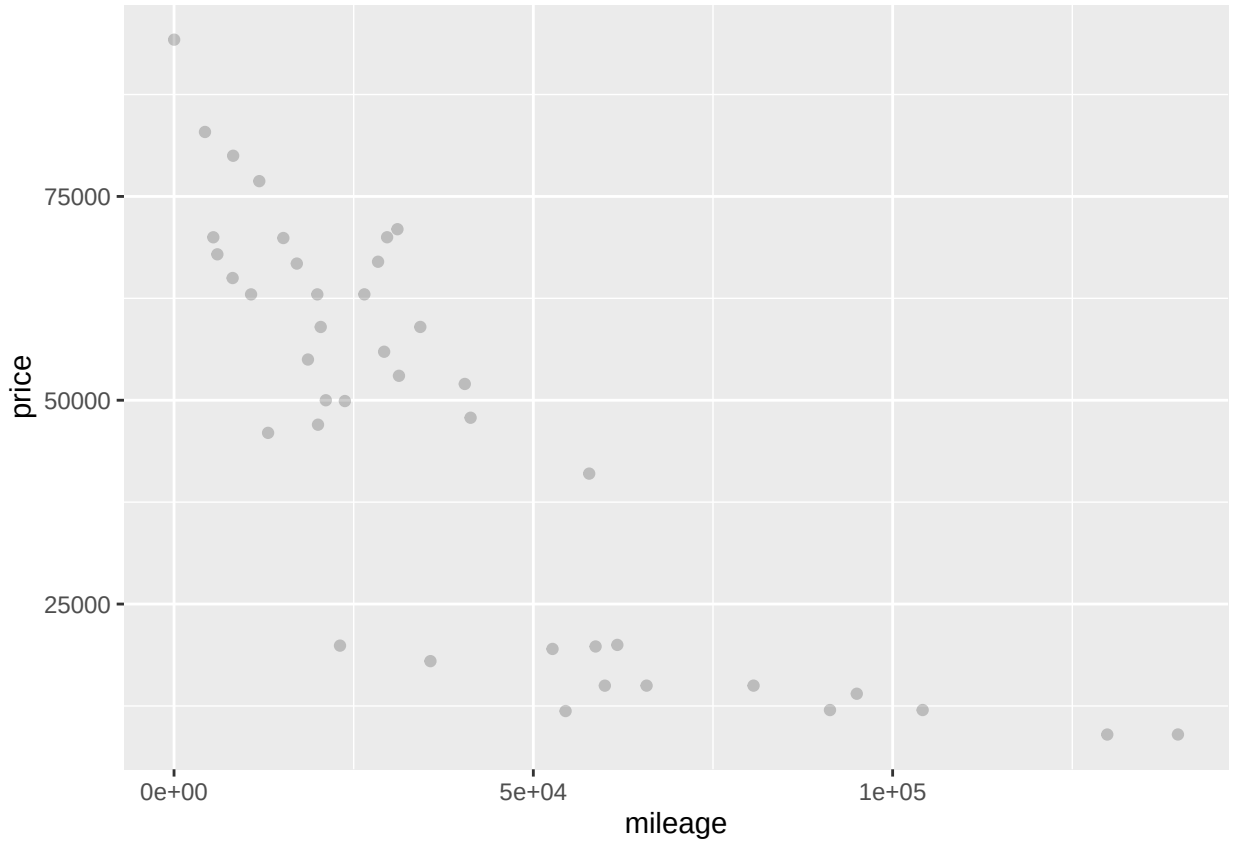


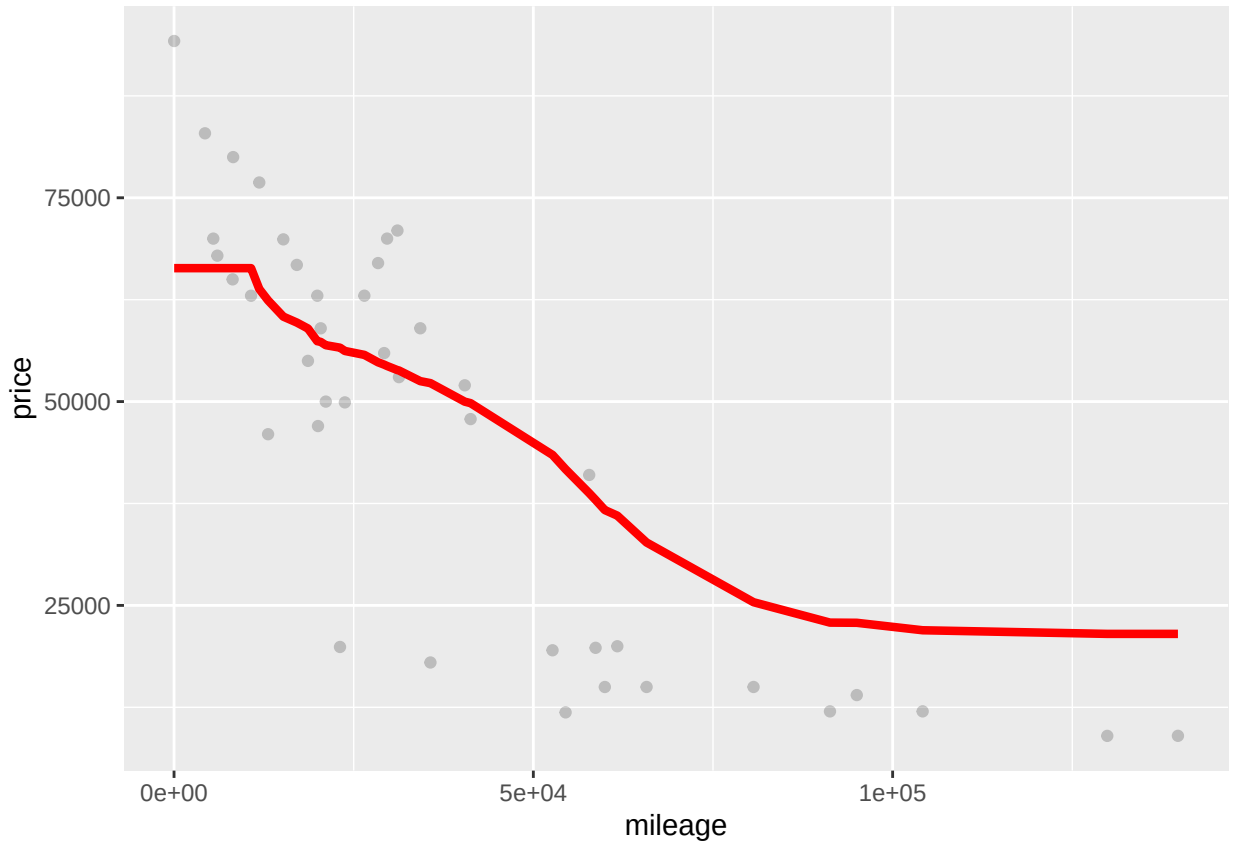
RMSE out for the Mercedes 350 at each K:

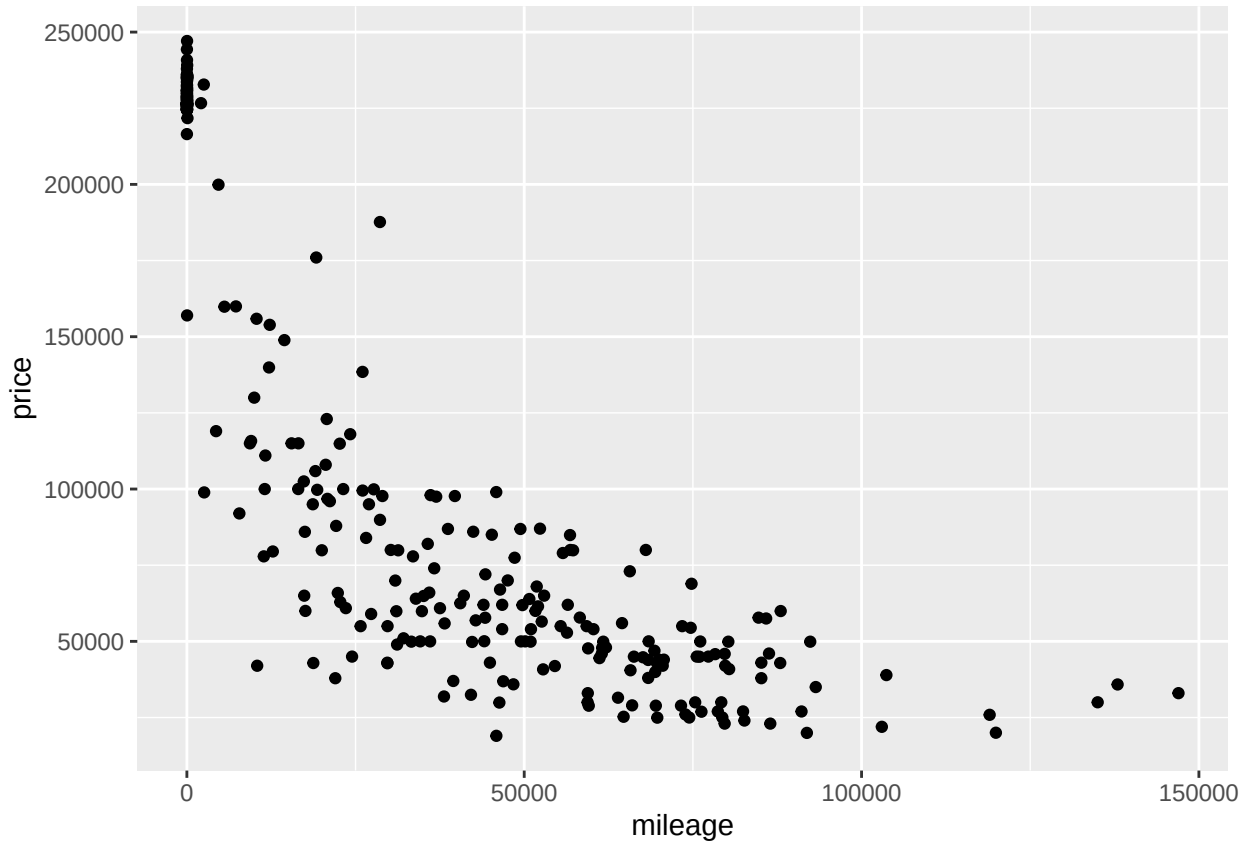
2 12906.974
4 11947.642
6 11639.215
8 11867.496
10 12065.810
15 11633.114
20 11526.127
25 11205.392
30 11208.124
35 11185.184
40 10961.306
45 10573.645
50 10540.959
60 10502.542
70 10211.668
80 9959.282
90 9805.222
100 9653.767
125 9412.362
150 9821.775
175 10273.357
200 10945.053

The optimal level of K for the Mercedes 350 is 125, where the RMSE equals 9412.362.


17500
15000
12500
10000
7500
rmse
2
4
6
8
10
15
20
25
30
35
40
45
50
60
70
80
90
100
125
150
175
200
factor(k)


75000
50000
25000
price
0e+00
5e+04
1e+05
mileage

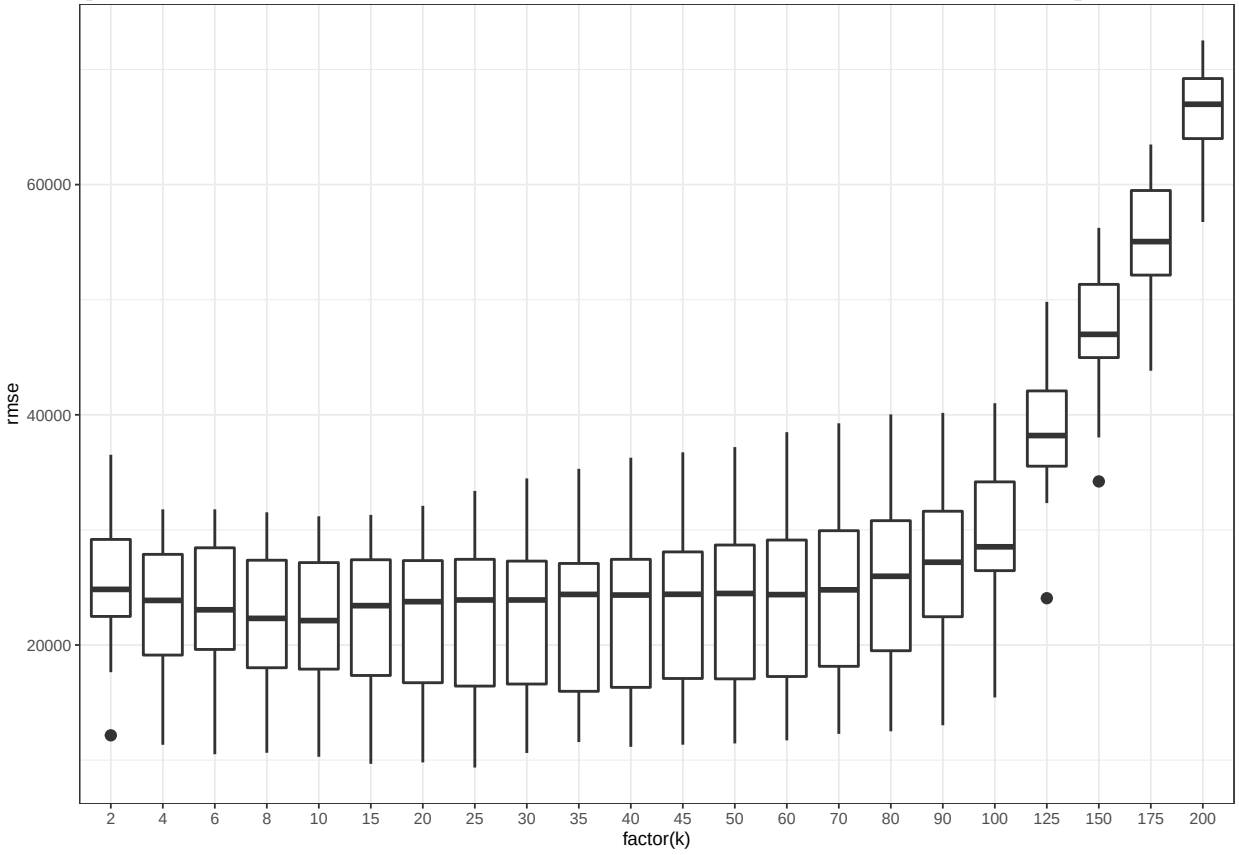

75000
50000
25000
price
0e+00
5e+04
1e+05
mileage

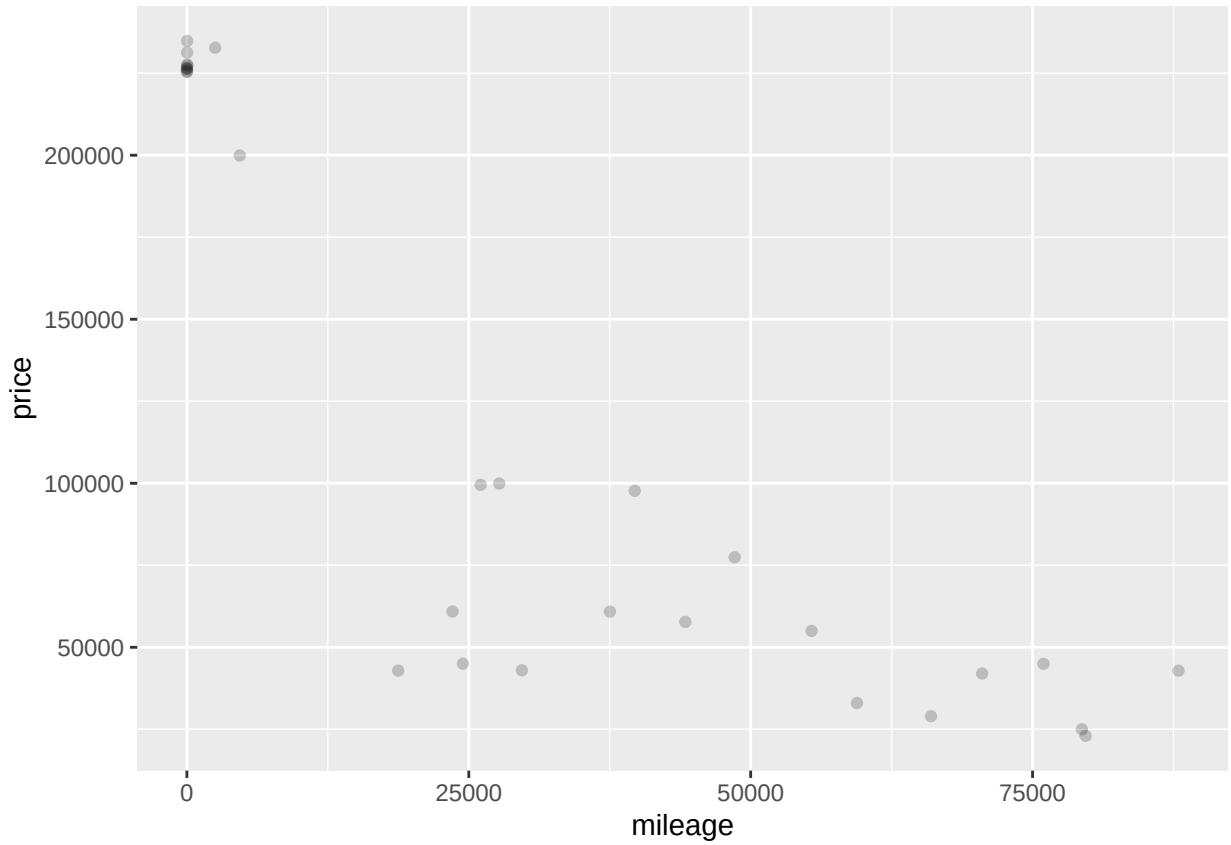

250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

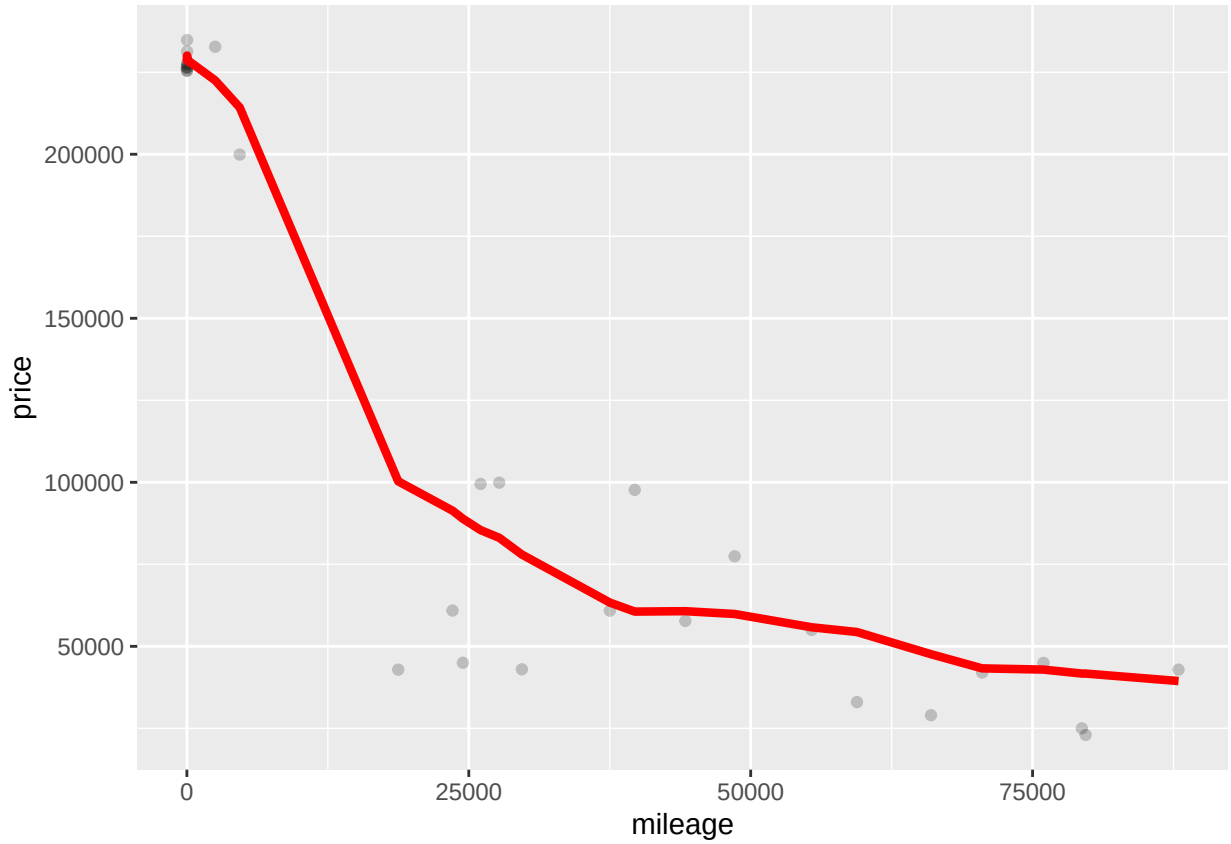
RMSE out for the Mercedes 65 AMG at each K:

2 20002.32 4 15621.97
6 18073.07
8 16423.41
10 15961.65
15 14328.33
20 13961.60
25 14105.38
30 13511.87
35 12698.51
40 12043.71
45 12085.55
50 12451.23
60 12844.28
70 13750.54
80 14387.01
90 16083.07
100 20224.12
125 33209.58
150 44704.92 175 53524.28
200 65526.76

The optimal level of K for the Mercedes 65 AMG is 40, where the RMSE equals 12043.71.







RMSE out for the Mercedes 65 AMG when k is 40 equals 12043.71.

The Mercedes S-class 350 has the larger optimal value of K. This is because there are more obersvations of the 350 trim in the S-class dataset. More observations per trim allows the optimal value of K to be higher before significantly biasing the model.