Data Mining Homework 1

Suyu Liu

2023-01-27

1) Data visualization: flight at ABIA

- What is the best time of day to fly to minimize delays, and does this change by airline?

The figure 1 takes the scheduled departure time of the flight as the abscissa and the average departure delay as the ordinate, showing the average departure delay of the plane every hour of day in 2008. For the departure time, 5am to 6am is the best period with the least average flight departure delay time. According to the figure1, flights during this time even departed earlier.

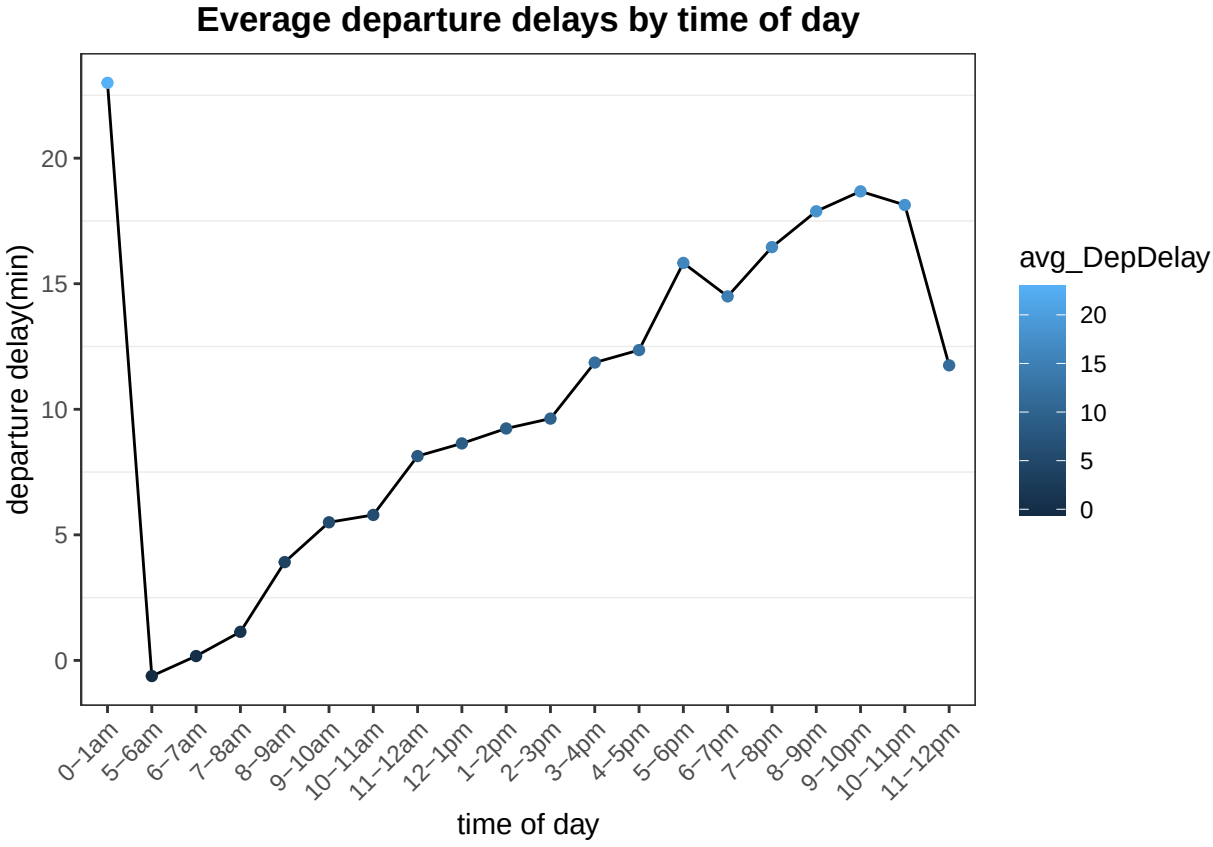


Figure 1: Line graph of the flight departure delay of the time of day

- Does this change by airline?

The figure 2 shows the average departure delay time per day by airline. Upon Figure 2, the situation of each airline is very different. The best time of day to fly to minimize delays changes by airline. US airline is the most punctual airline. UA's flight at 8:00 p.m. is even delayed by an average of more than 150 minutes.

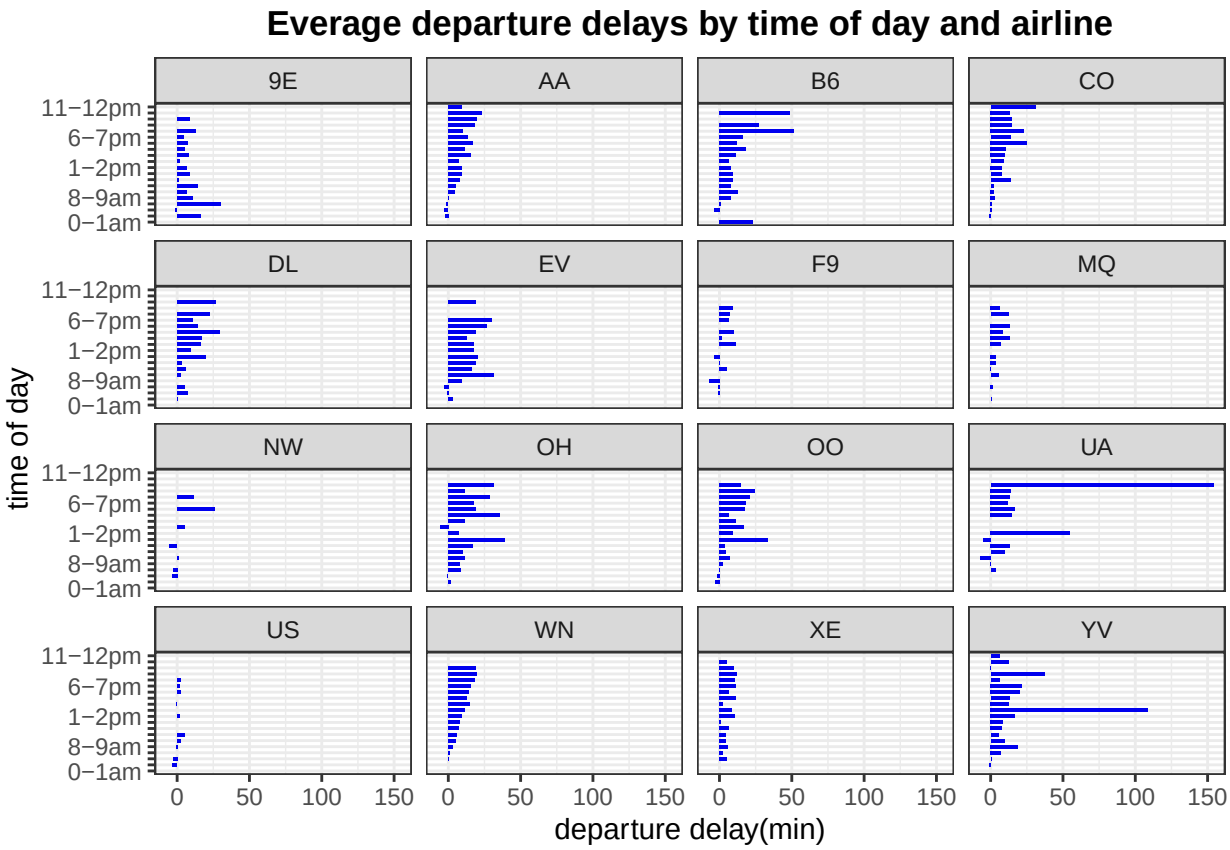


Figure 2: Bar graph of the flight departure delay by the time of day and different airlines

2) Wrangling the Olympics

A) The 95% of heights for female competitors across all athletics events is 183.00.

```
##   q95_heights
## 1         183
```

B) Swimming Women's 100 metres Butterfly was the single women's event with the greatest variability in competitor's heights. The following table shows the top 7 women's events, both team and individual, in which the heights of female competitors vary the most.

event	sd_femaleheights
Rowing Women's Coxed Fours	10.865490
Basketball Women's Basketball	9.700255
Rowing Women's Coxed Quadruple Sculls	9.246396
Rowing Women's Coxed Eights	8.741931
Swimming Women's 100 metres Butterfly	8.134398
Volleyball Women's Volleyball	8.101521
Gymnastics Women's Uneven Bars	8.015942

C) The average age of Olympic swimmers

- How has the average age of Olympic swimmers changed over time?

Upon figure 1, the average age of swimmers has fluctuated over time around the age of 20. From 1900 to 1912, the average age increased a lot. From 1912 to around 1931, the average age dropped again from over 25 to under 20. From then until 1975, the average age remained around 20. Since 1975, the age of the players has increased over time.

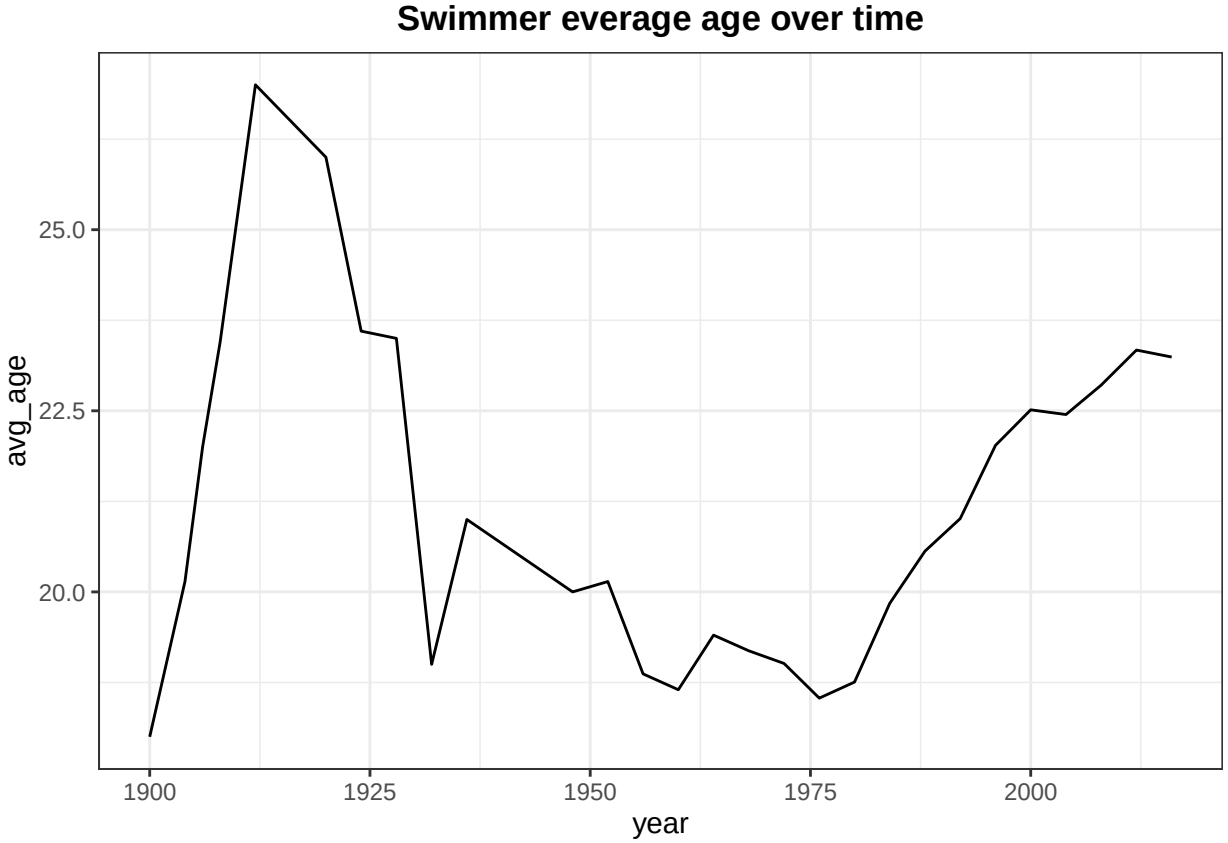


Figure 1: Line graph of average age of Olympic swimmers changed over time

- Does the trend look different for male swimmers relative to female swimmers? The average age of both male and female swimmers has increased over time since 1950. However, since 1925, the average age of female swimmers has been lower than that of male swimmers.

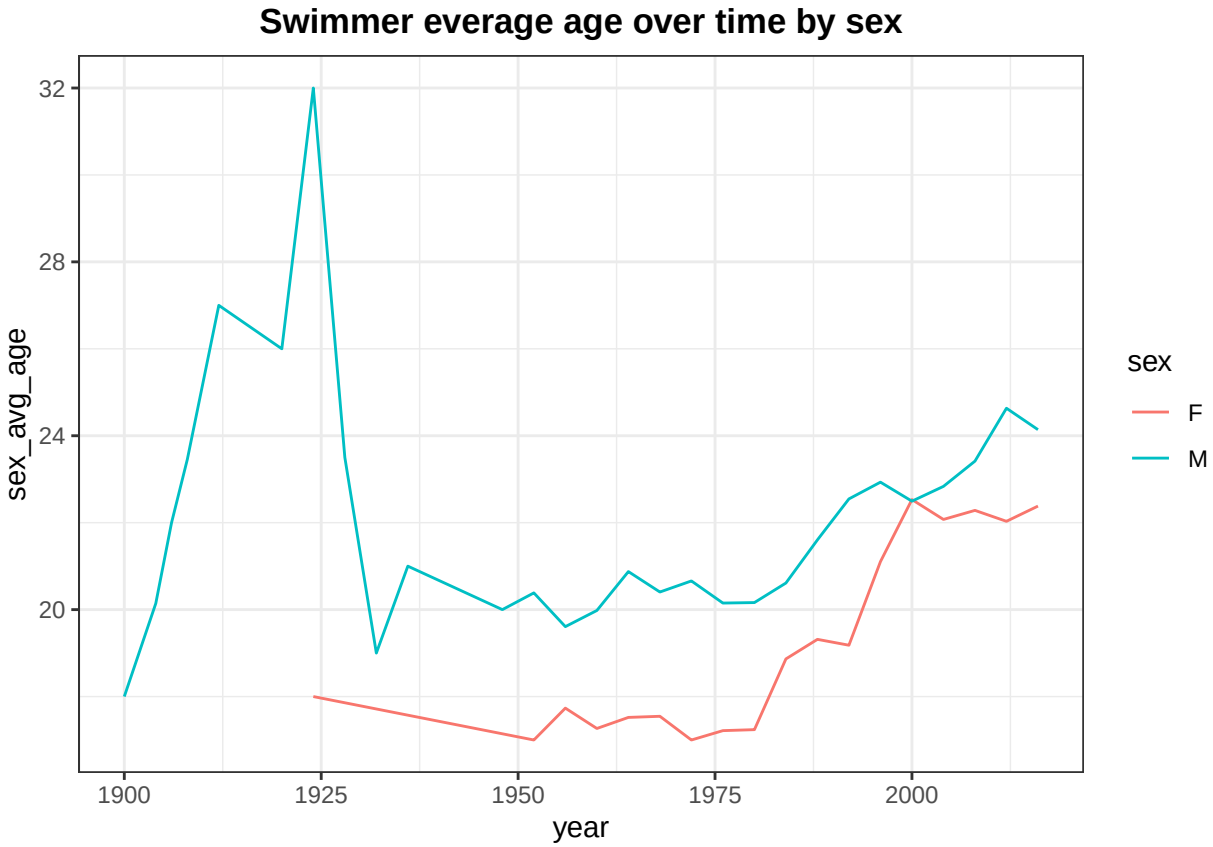


Figure 2: Line graph of average age of Olympic swimmers changed over time by sex

3) K-nearest neighbors: cars

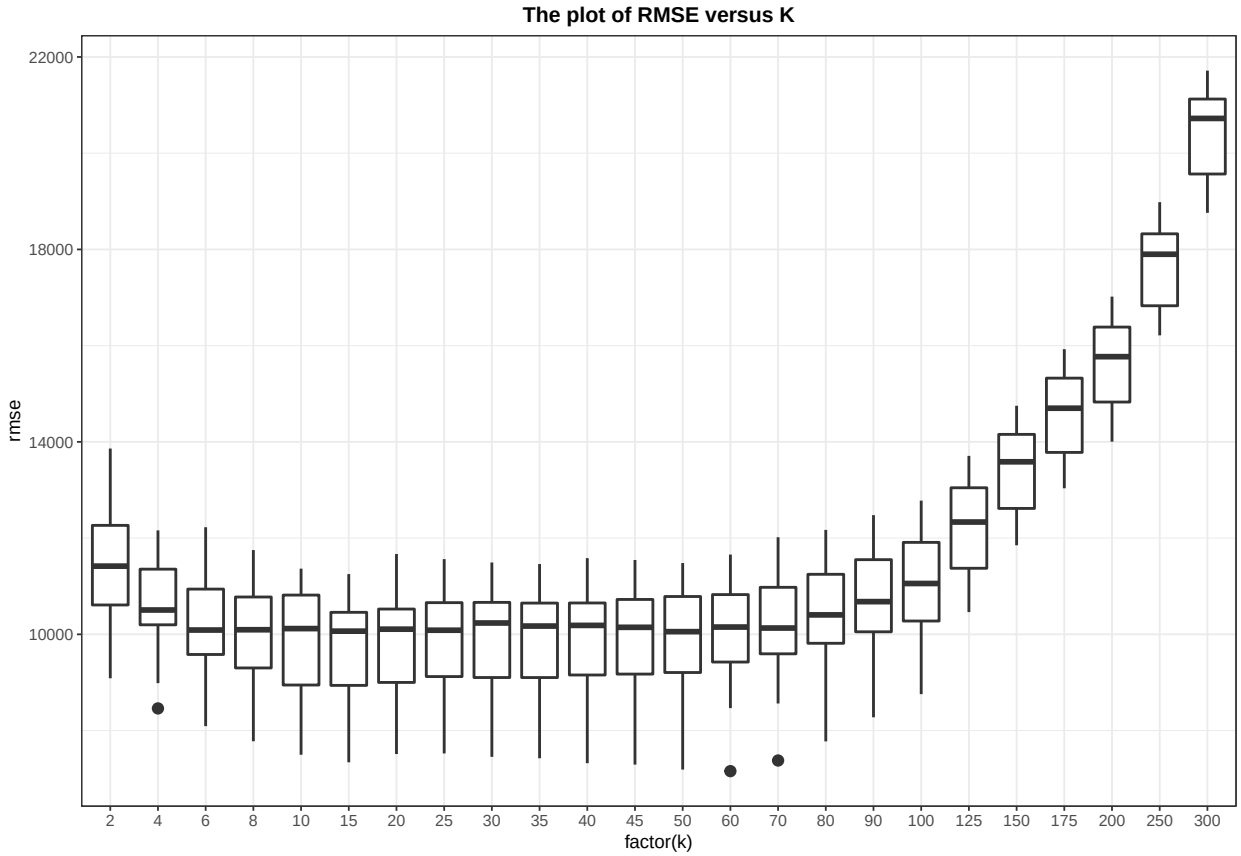
- Two Methods: I split the data under 0.8 probability and also use the K-fold cross validation to get optimal K.

K-fold cross validation:

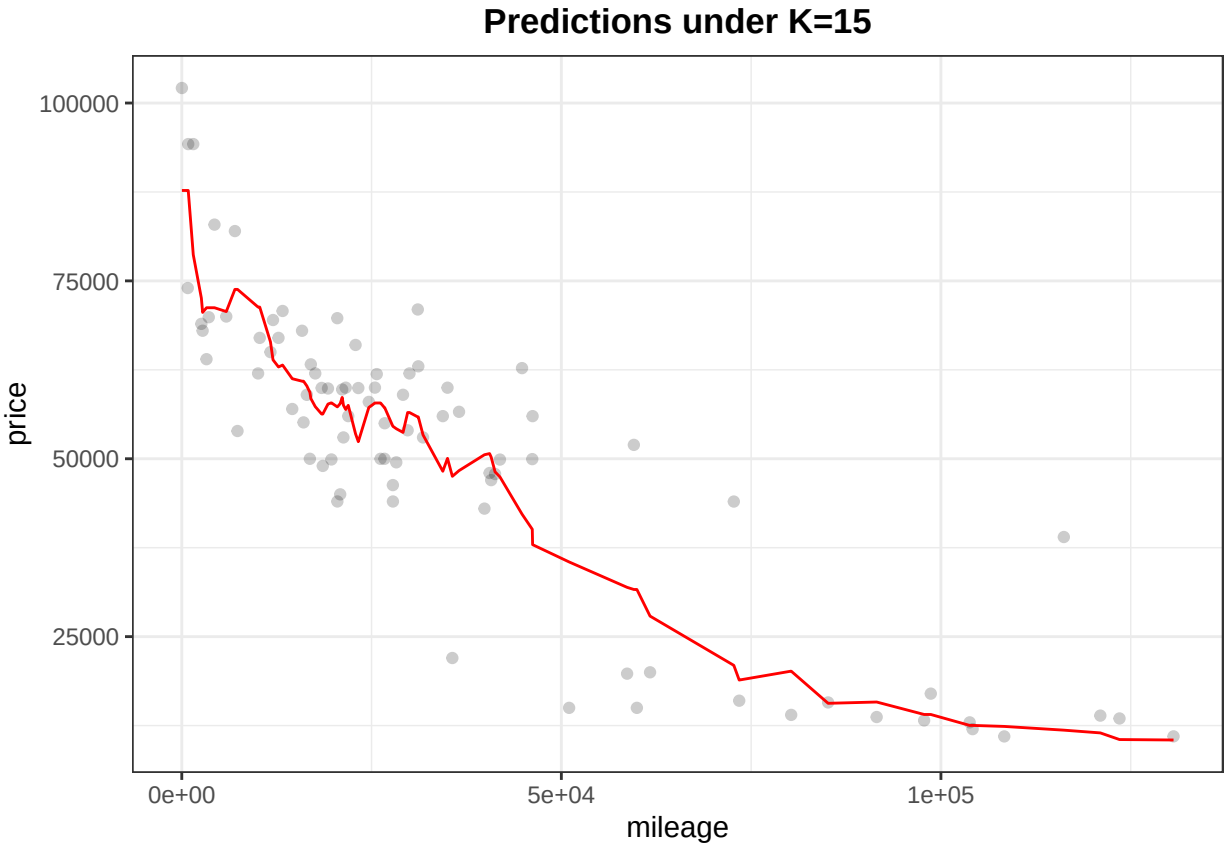
1. Split the data into a training and a testing set with 5 folds randomly.
2. RMSE and prediction.
3. Find the minimum RMSE under different K values(1-300 & 1-233).
4. Plot the fitted model with optimal K.

Trim 350

- Method 1: Split the data under 0.8 probability

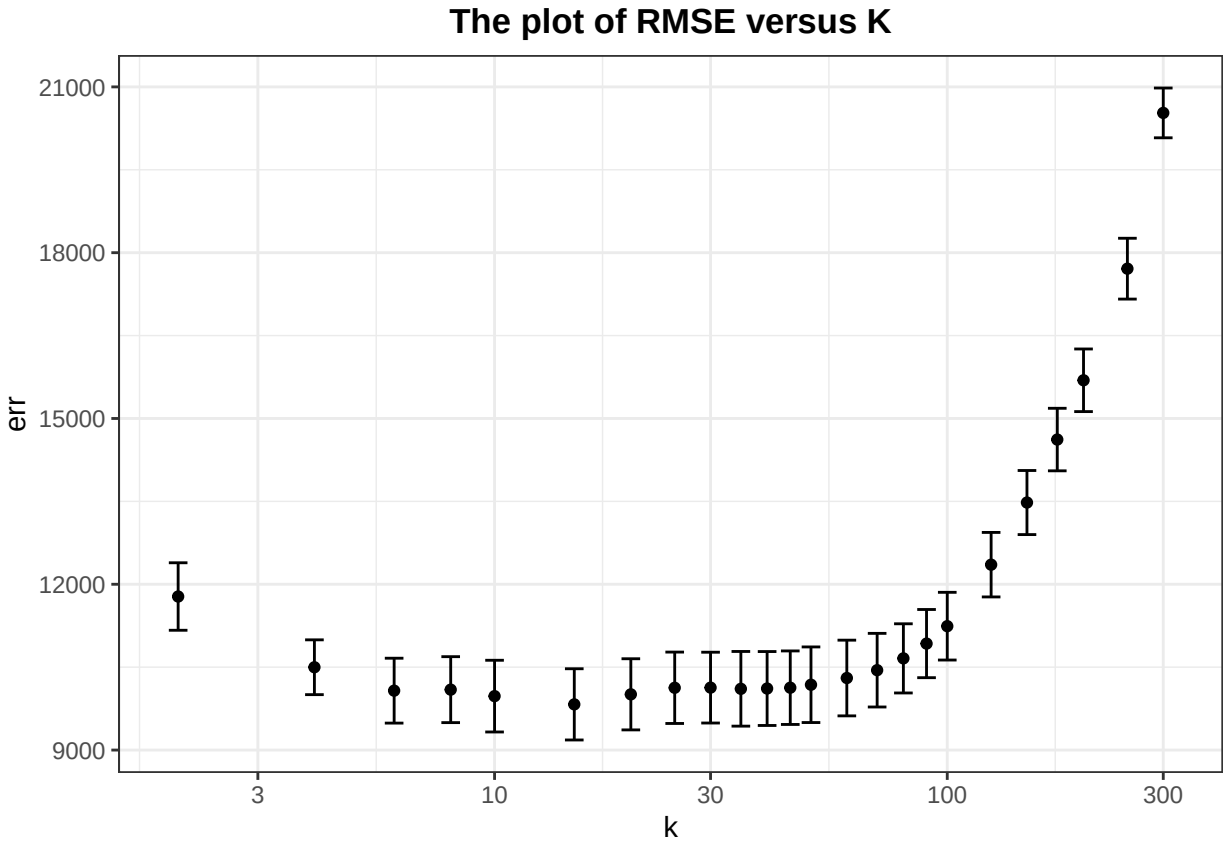


The optimal k is 15, which has the minimum RMSE. The following is the plot of prediction.

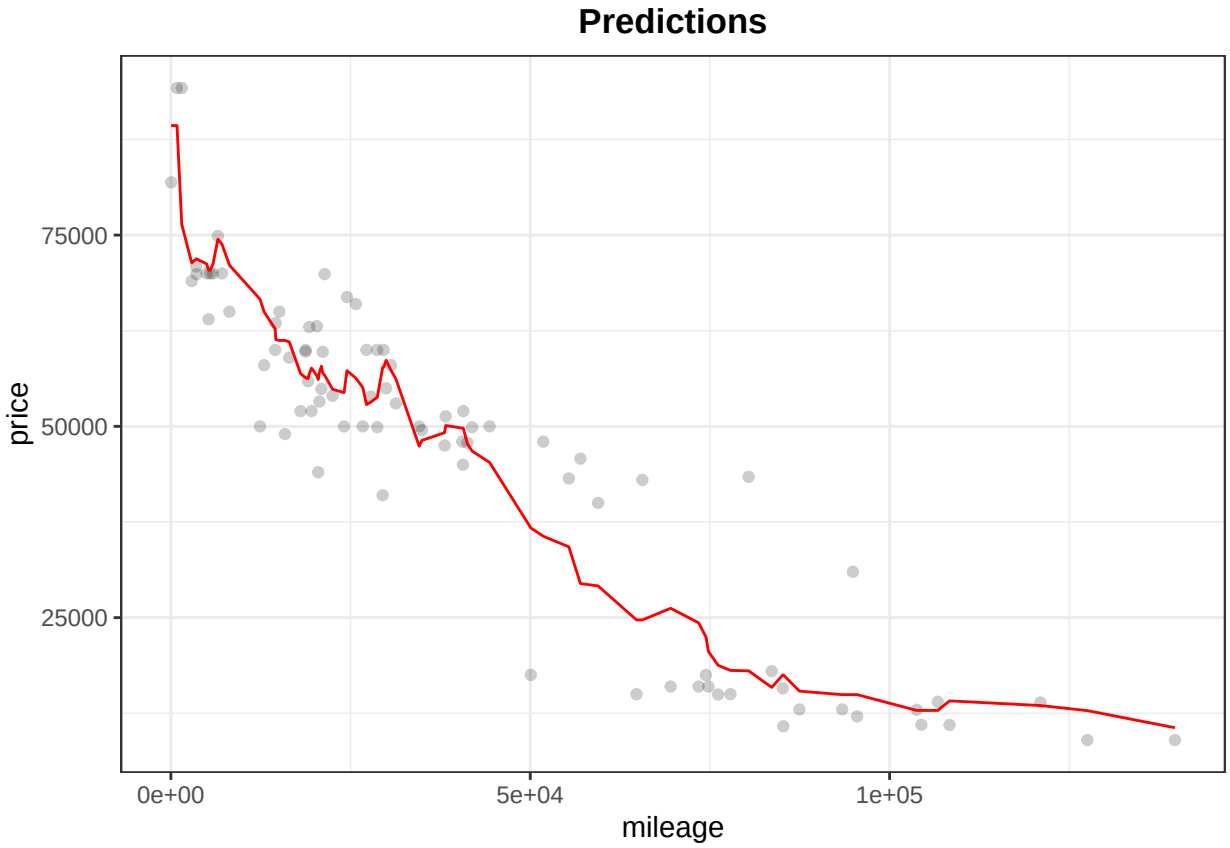


- Method 2: K-fold cross validation
- The plot of RMSE versus K. The following is the minimum RMSE, corresponding to k is 15.

```
##            k      err std_err
## result.6 15 9826.577  644.57
```

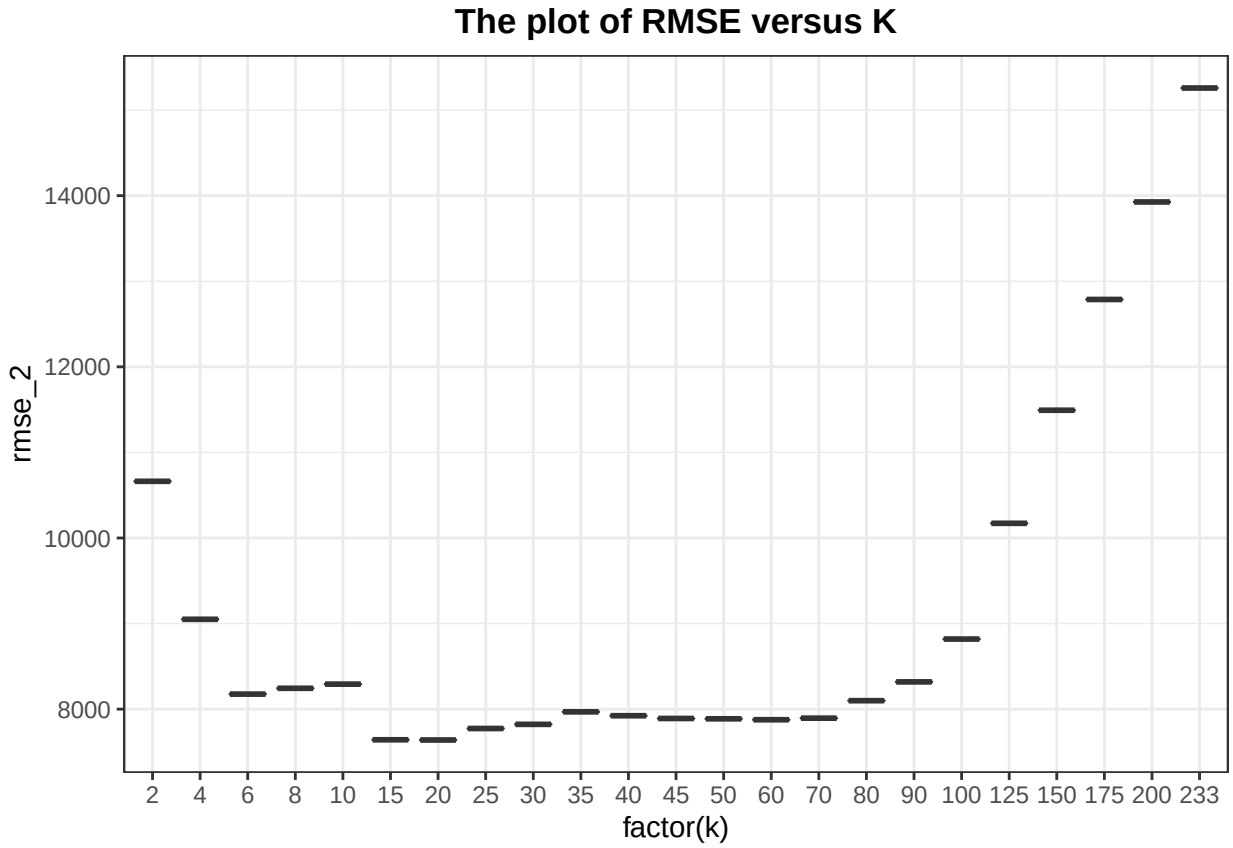


- For the optimal value of K with minimum rmse, show a plot of the fitted model–predictions vs. x.

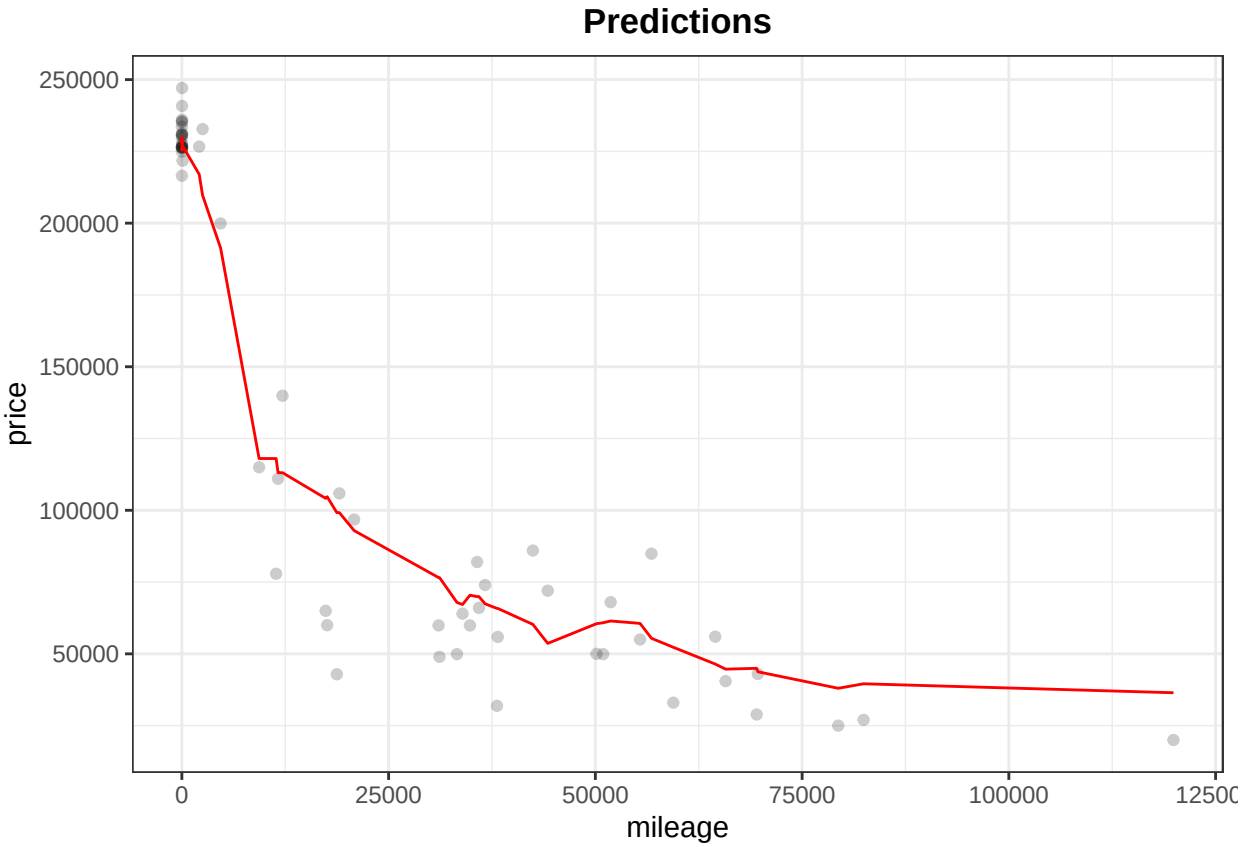


Trim: 65AMG

- Method 1: Split the data under 0.8 probability

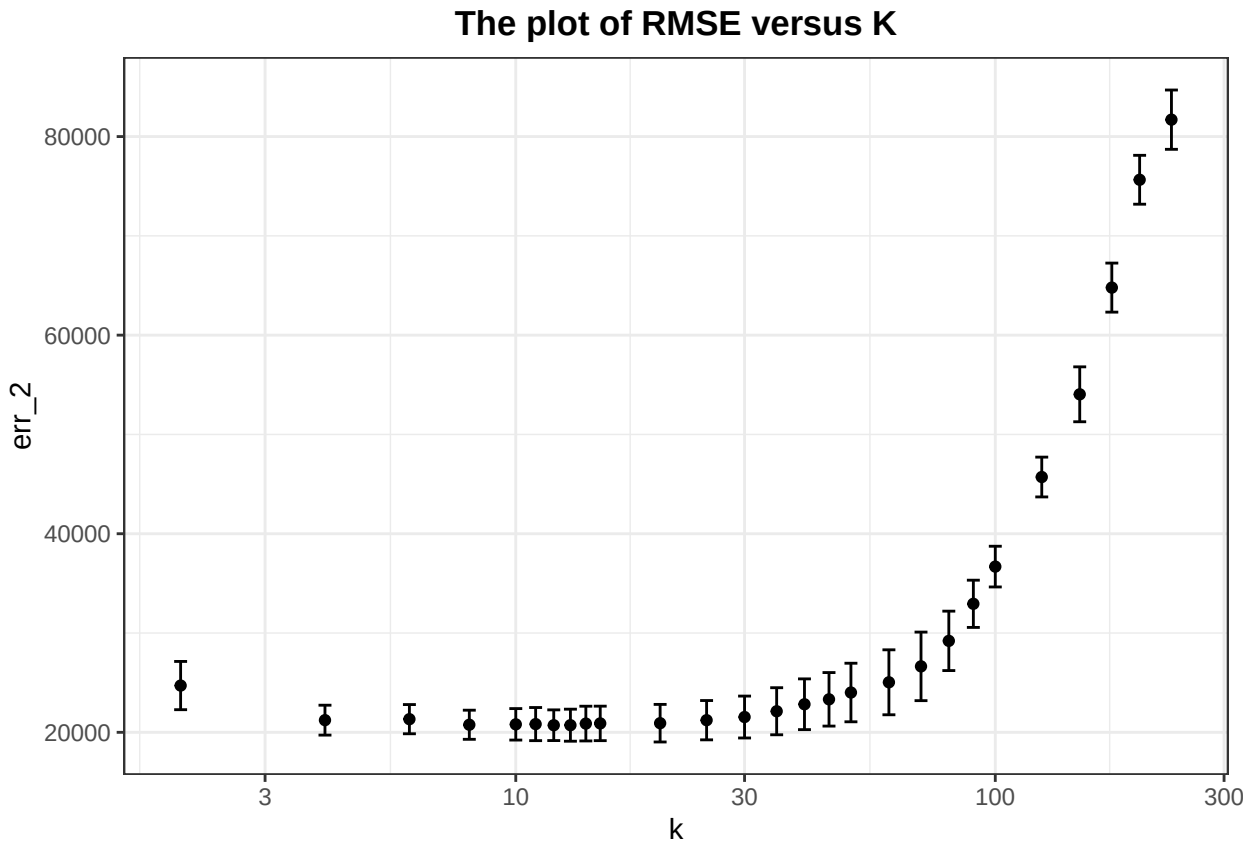


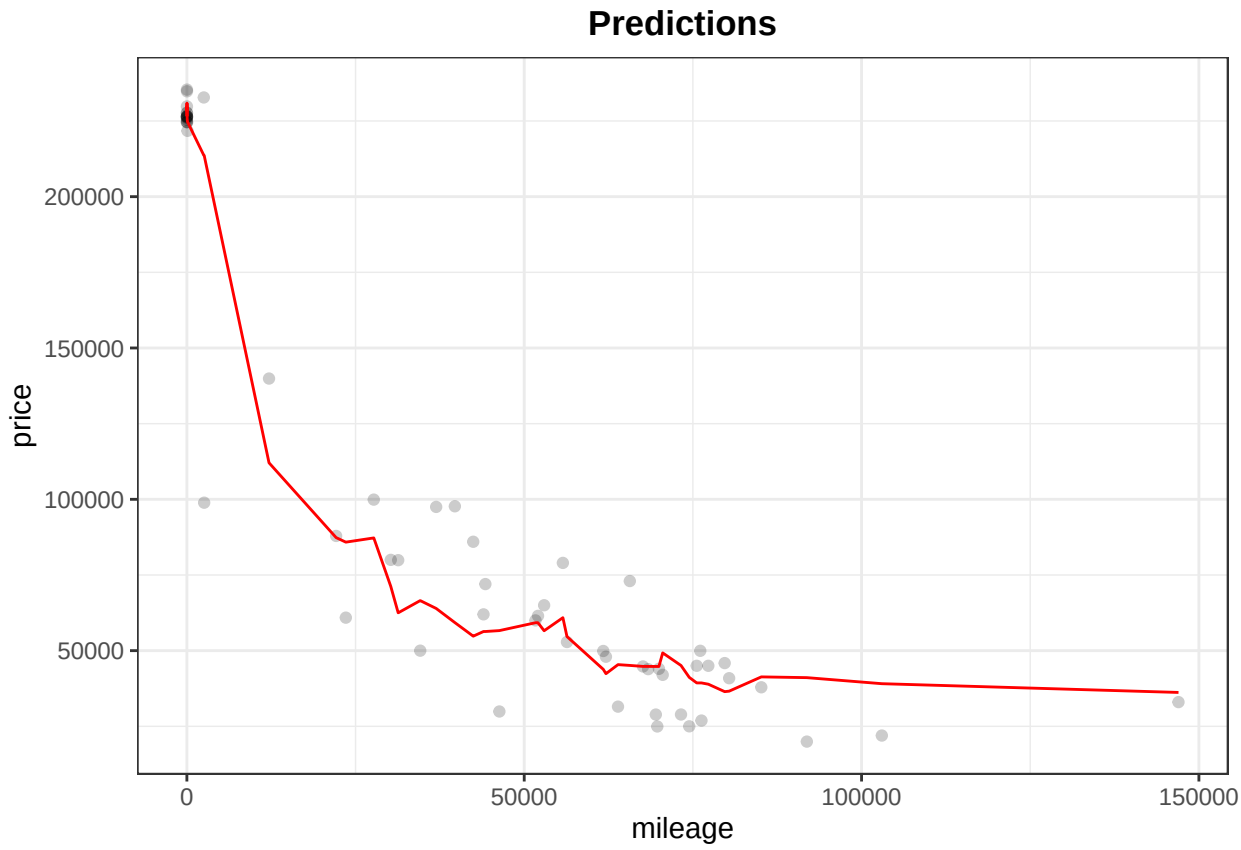
The optimal k is 15, which has the minimum RMSE. The following is the plot of prediction.



- Method 2: K-fold cross validation
- The plot of RMSE versus K. The following is the minimum RMSE, corresponding to k is 12.

```
##            k   err_2 std_err_2
## result.7 12 20715.2  1546.157
```


The plot of RMSE versus K
80000
60000
40000
20000
err_2
3
10
30
100
300
k


Predictions
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

Which trim yields a larger optimal value of K? Why do you think this is?

When using the k-cross folds validation(an more accurate method than the method 1), the 350(K=15) yields a large optimal value of K than 65AMG(K=12). Because the 350's(416) number of observations is larger than the 65AMG's(292). For a better balance between bias and variance, the 350 with larger observations needs a larger K.