Homework 1

Name: Yu Hsiang Tseng

1) Data visualization: flights at ABIA

I want to focus on answering the following questions:

1. What is the best time of day to fly to minimize delays, and does this change by airlines?

2. What is the best time of year to fly to minimize delays, and does this change by destination?

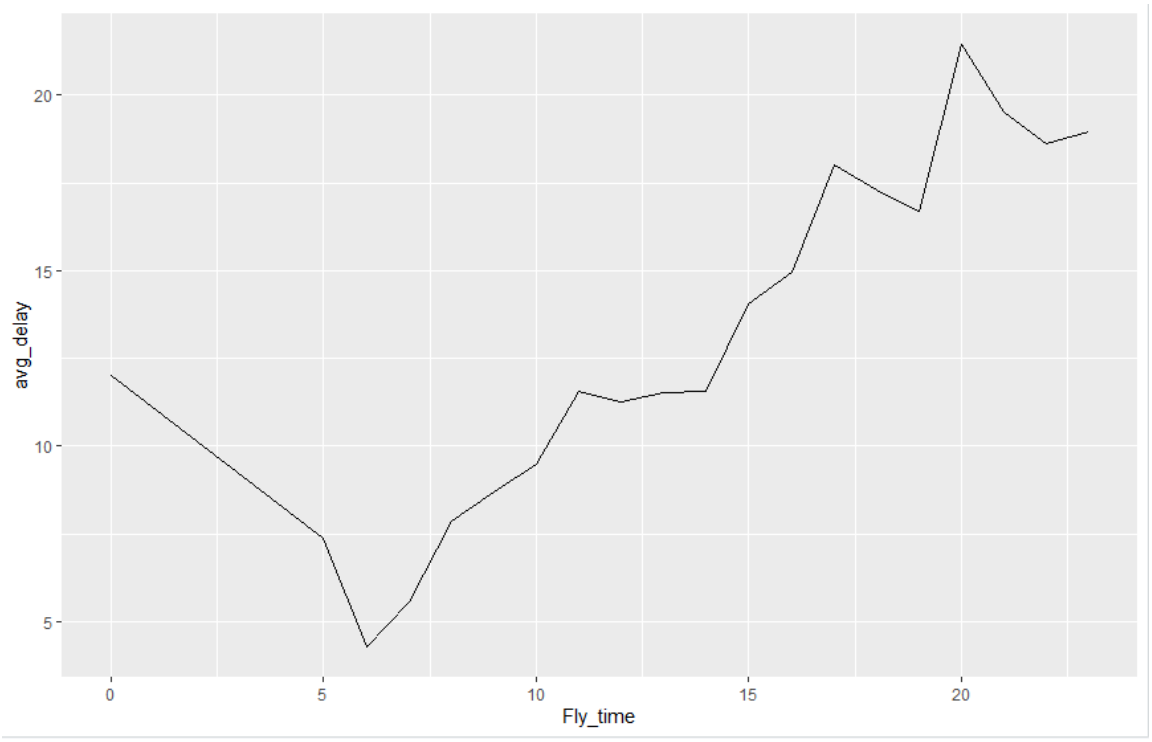
For these two questions, I face a hard problem, how do I define delay? Is the early arrival counted delay also? Or only the late arrival counted as a delay. In my opinion, most people only care about whether they can arrive on time. Even though they get to the destination so early, they won't argue about it. Similarly, even though they departed late, they would still be happy if they can arrive on time. In addition, most people would arrive at the airport early for check-in, early departure won't have a problem. In conclusion, I would only focus on late arrivals, which could make people's schedule disturbed.

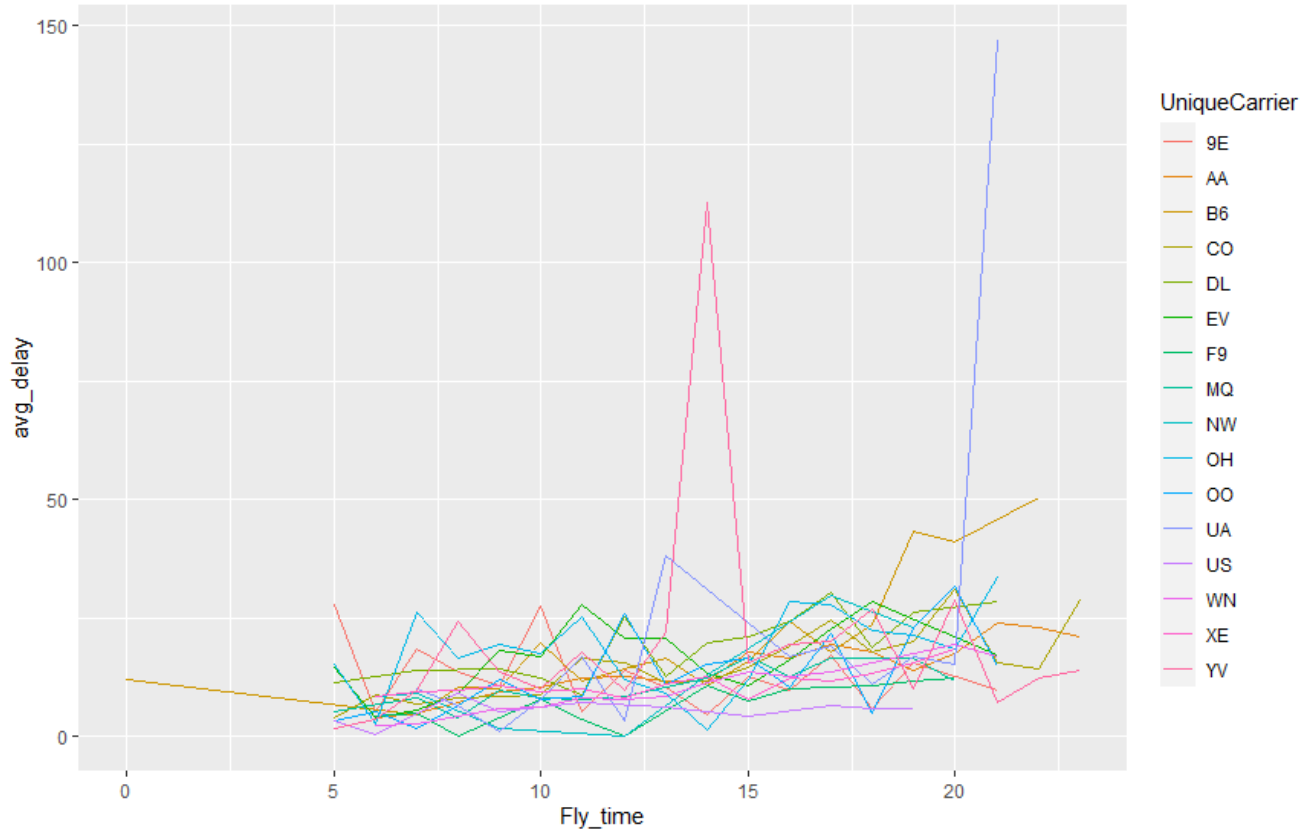
In the ABIA dataset, I would only count the positive number of ArrDelay as delay and the other would be counted on time so the delay_mins of on time would be zero.

Besides, "the best time to fly" should be a specific time or month, for the first question, since the departure time would be so trivial, I distribute them in 24 periods to make the analysis clear. Variable "Fly_time" is defined as 0 when the departure time is between 0 AM and 1 AM.

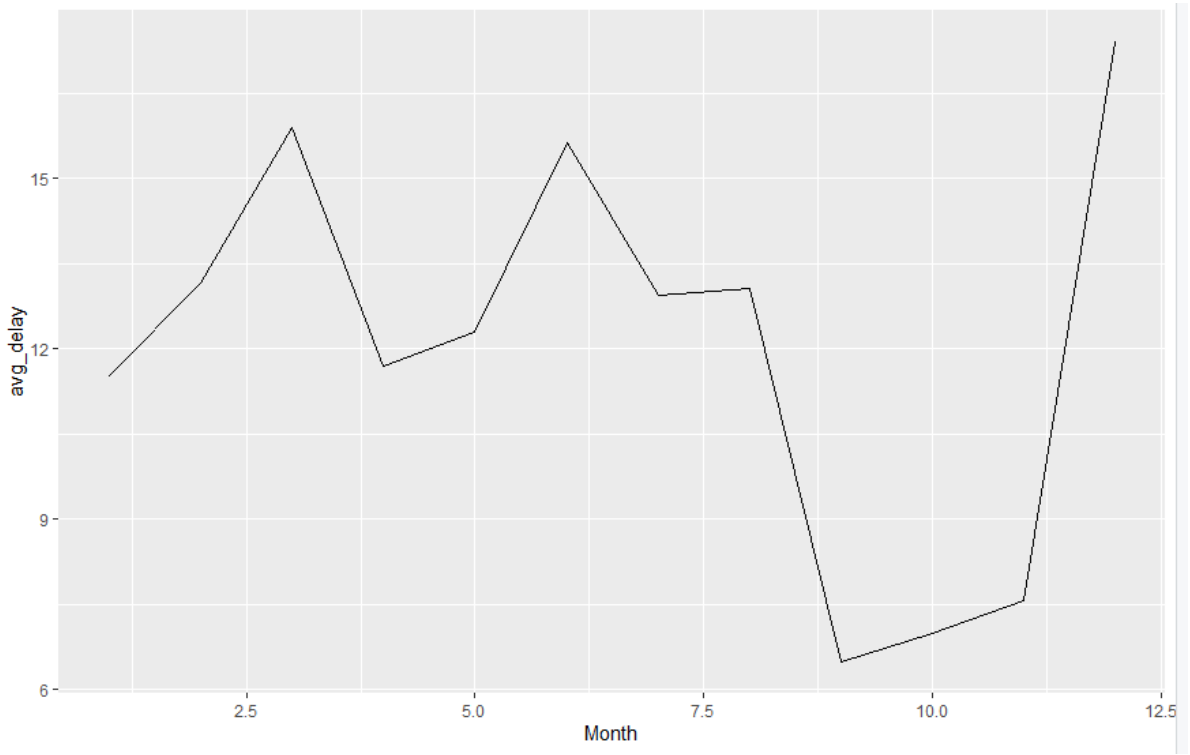
Also, there is a problem with question 2, how to define a popular destination. By observing the data by "xtab", I find that showing numbers over 2000 are only 9 destinations, which can be considered popular destinations.

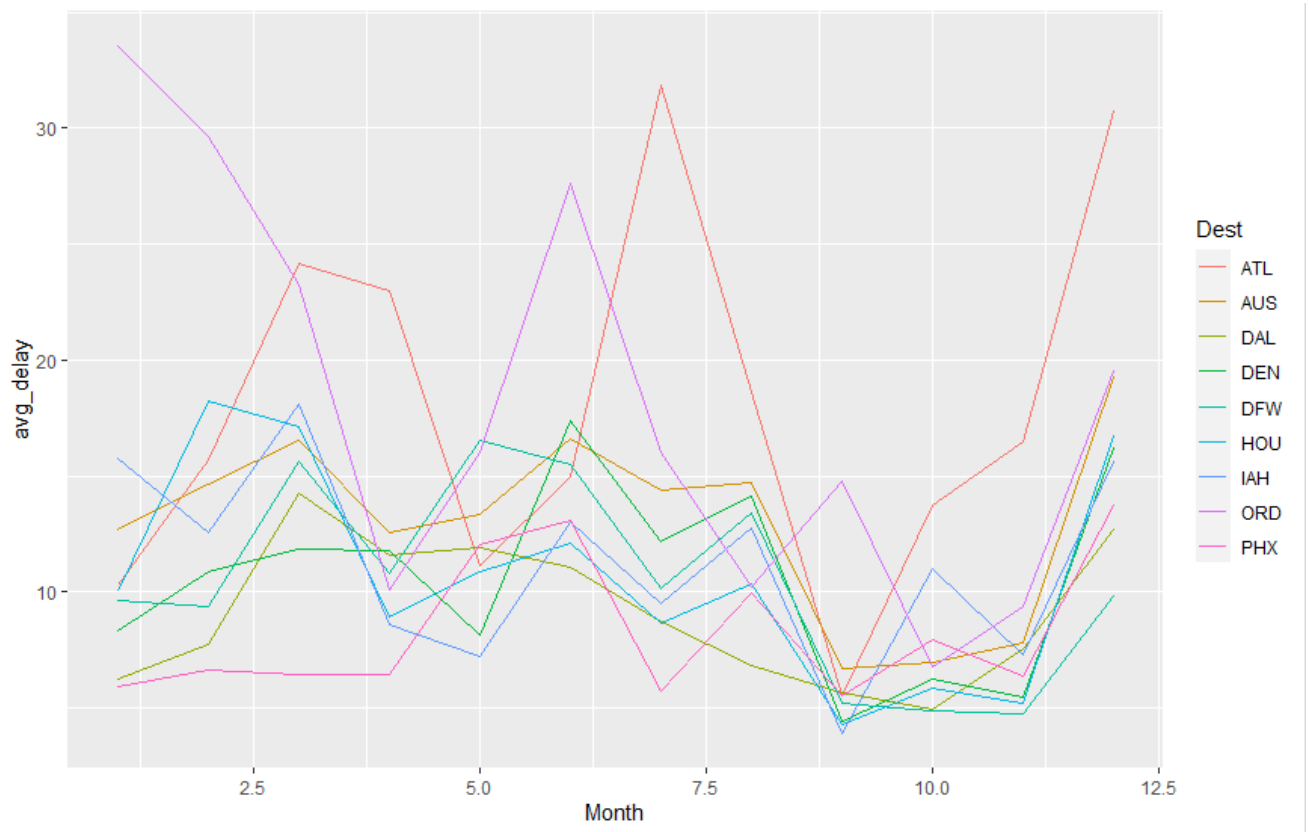
The first two graphs are to answer the first question, and the last two are for the second question.





From the first plots, we can say the period with the lowest average delay is 6 AM to 7 AM for all flights. However, this does change by the airline. From the second graph, we could find that every airline has its best period to fly. Not all colors of the line have the bottom at 6 AM to 7 AM.





From these two graphs, September is the best time of year to fly to minimize delays. According to the last graph, we can find September is the best time of year to fly to minimize delays for almost all of the popular destinations except for ORD.

2) Wrangling the Olympics

A) What is the 95th percentile of heights for female competitors across all Athletics events (i.e., track and field)? Note that sport is the broad sport (e.g. Athletics) whereas the event is the specific event (e.g. 100-meter sprint).

Answer:

95th percentile of heights for female competitors across all Athletics events is 183 cm.

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?

Answer:

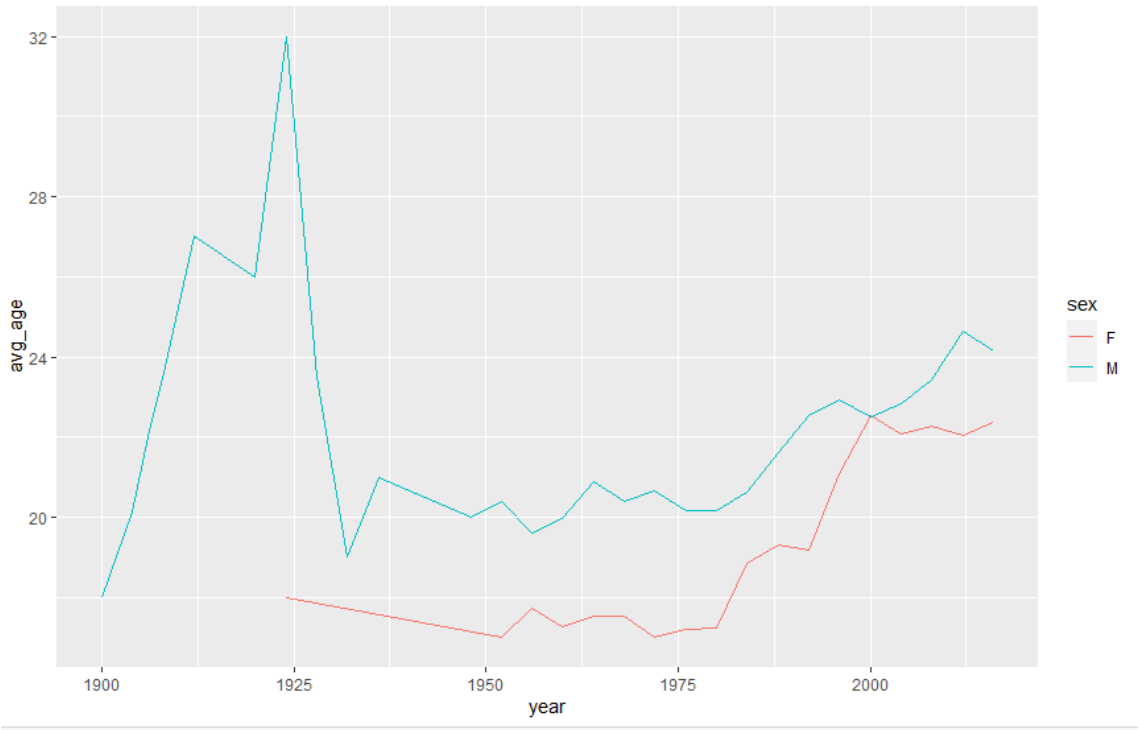
The single women's event that had the greatest variability in competitor's heights across the entire history of the Olympics is Rowing Women's Coxed Fours, which has a 10.9 standard deviation.

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.

Answer:

These are part of the data frame and the according graph:

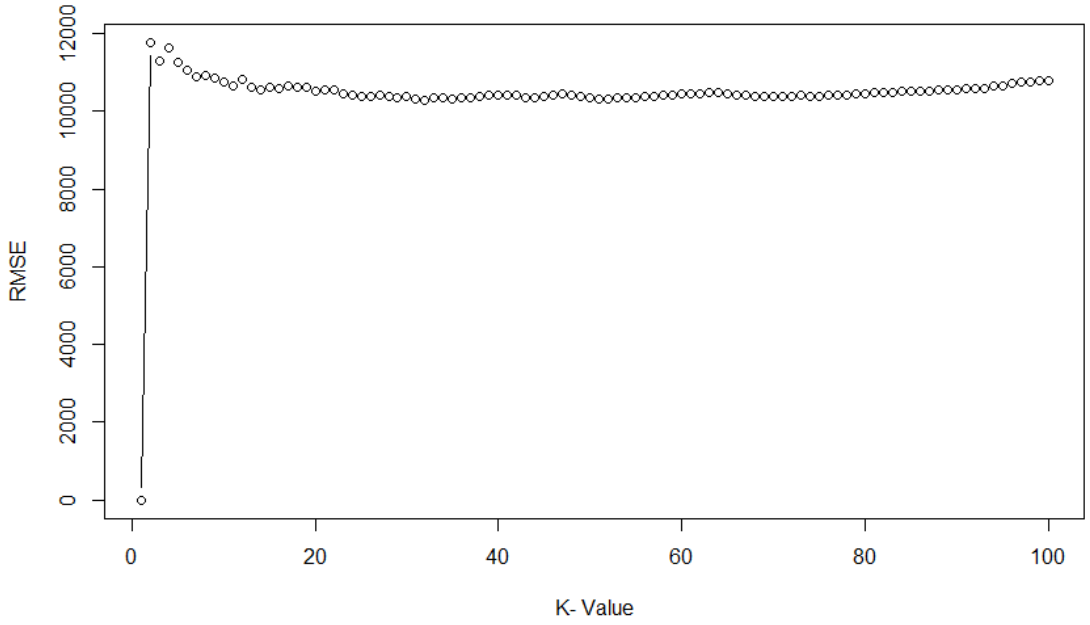
	Year	Sex	Average age
1	1900	M	18
2	1904	M	20.1
3	1906	M	22
4	1908	M	23.5
5	1912	M	27



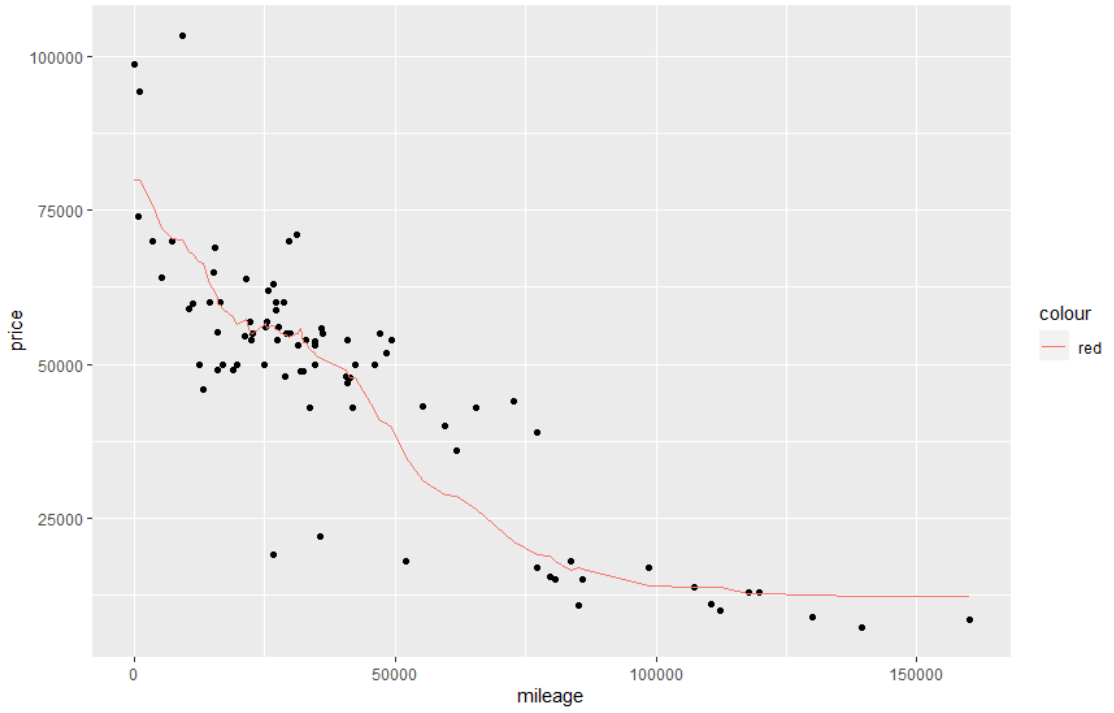
From this plot, we can first find that the average age of Olympic swimmers increased first until 1925, and then suddenly decreased until 1950. After that, the average age has a stable increase until 2016. In addition, even though there are not any female data until 1925, the trend after 1925 looks very similar between female and male swimmers, both of them first go down then go up. However, the average age of male swimmers is larger than that of female swimmers each year.

3) K-nearest neighbors: cars

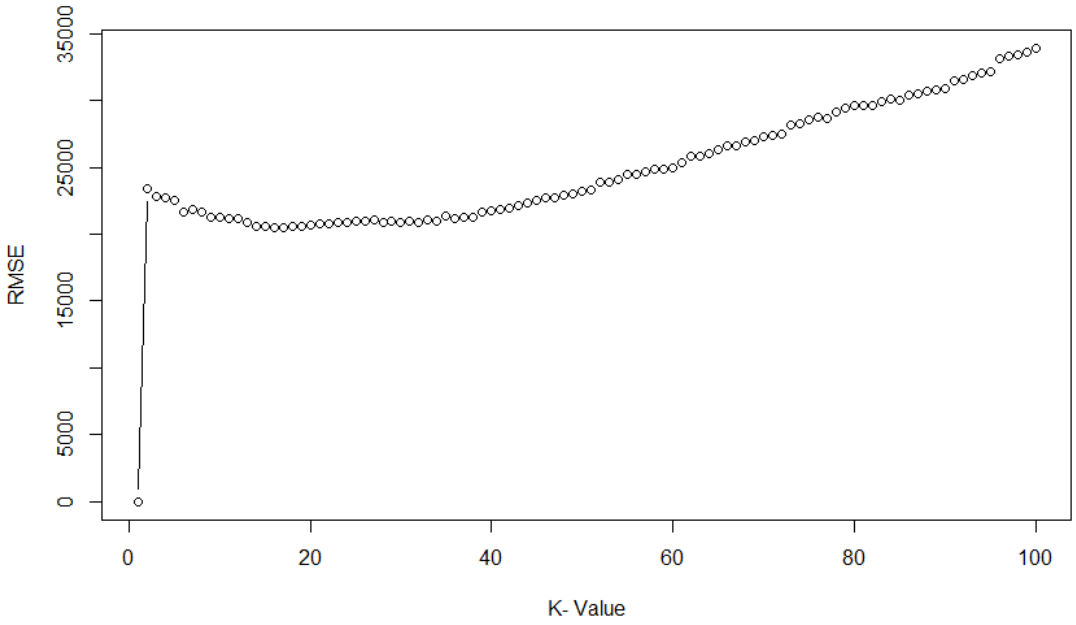
First, I used K-nearest neighbors model for K=2-100 with trim level=350', and find the according root mean-squared error (RMSE) for each value of K. Then, I make a plot of RMSE versus K and find that the optimal value of K equals 32.



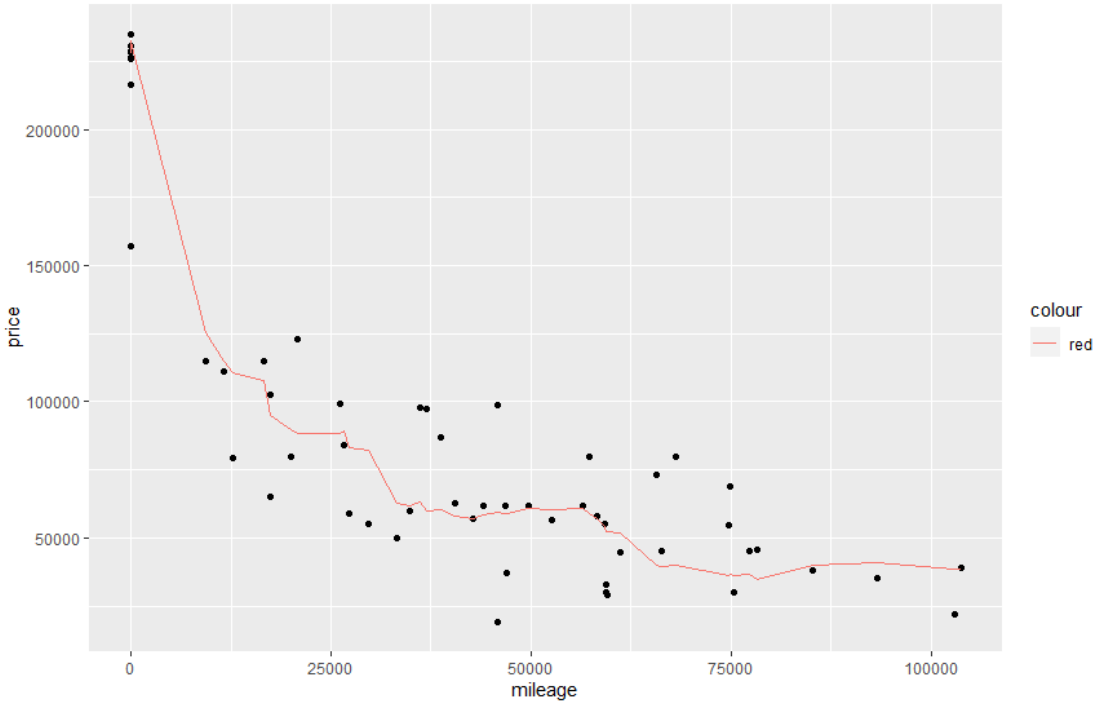
Then, for the optimal value of K equals 32, I make a plot of the fitted model. The red line is the prediction and the black dots are actual prices.



Second, I do the same thing for 65 AMG's. I also plot the RMSE versus K and find that the optimal value of K equals 16.



Then, for the optimal value of K equals 16, I make a plot of the fitted model. The red line is the prediction and the black dots are actual prices.



Question:

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

Answer:

In this case, the trim level = 350's yields a larger optimal value of K. It might suggest that for these two training data sets, 350's has a comparing lower price variation than 65 AMG's, so it could use more neighbors to predict the prices precisely.