ECO395M Homework 1

Kimberly Hu, Meilin Li, Yueting Zhang

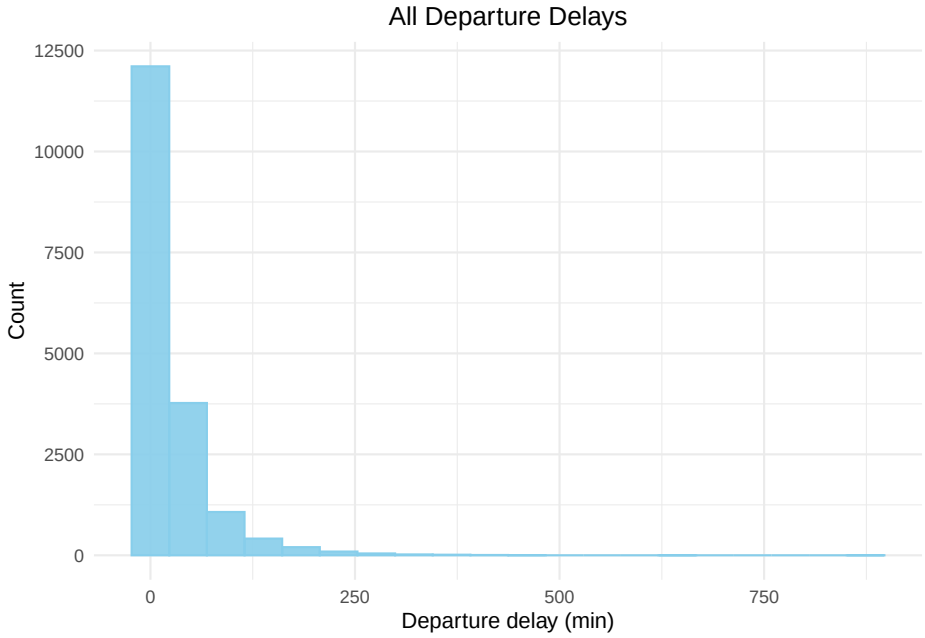
2024-02-05

1. Data visualization: flights at ABIA

Flight delays can often prove to be quite inconvenient. In order to enhance our overall travel experience, it is beneficial to examine the patterns of flight delays. In this question, we focused on flights that departed from the Austin-Bergstrom International Airport (AUS) in 2008, specifically those that departed later than the scheduled time.

The output below shows the summary statistics and distribution of all departure delays in minutes. The range of delays is quite large, with the minimum at 1 minute and the maximum at 875 minutes. Despite the right tail being very long, it is worth noting that most flights delayed no more than half an hour, since the third quartile is only 31.

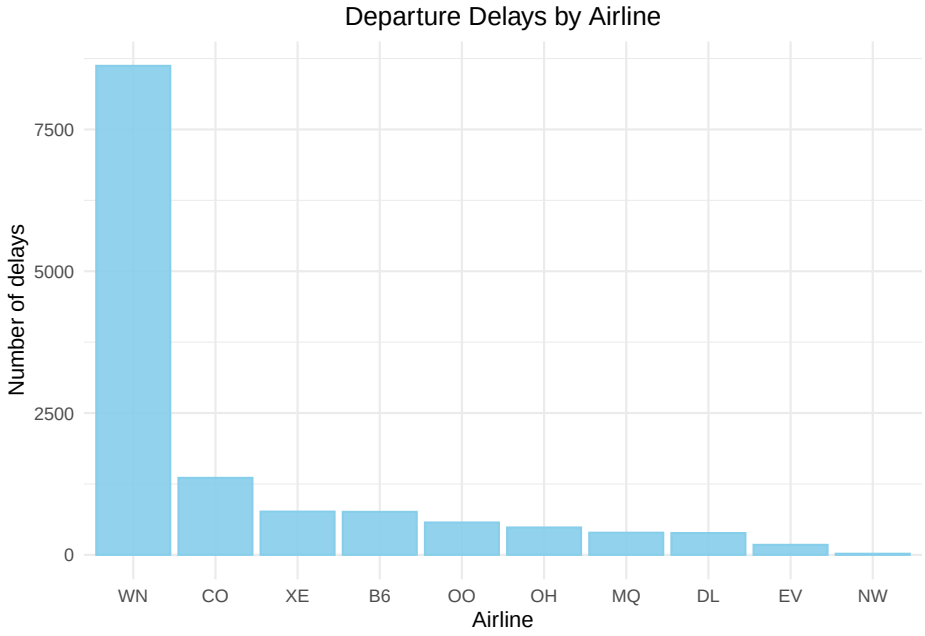
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00   11.00   27.29   31.00  875.00
```

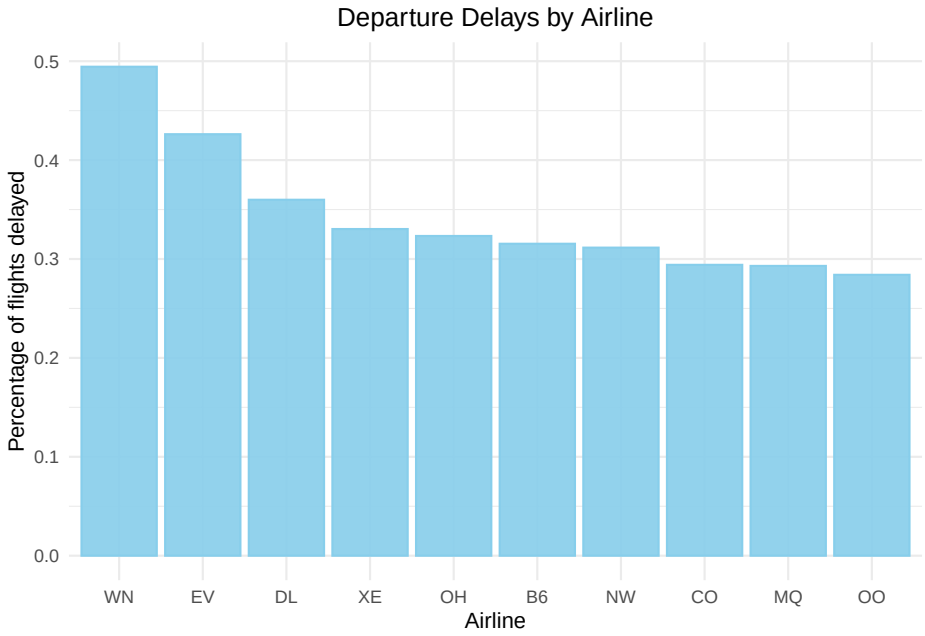


We extracted the top 10 longest delays from the data set. All 10 flights experienced delays longer than 400 minutes, and were spread out throughout the year and hours of the day. Out of the 10 delays, 4 were operated by JetBlue (B6).

Year	Month	Day	Scheduled Departure	Airline	Flight Number	Departure Delay (min)	Origin	Destination
2008	12	27	530	9E	5800	875	AUS	ATL
2008	7	29	1440	B6	1402	665	AUS	FLL
2008	7	25	1122	OO	6202	475	AUS	DEN
2008	10	18	1205	B6	1264	442	AUS	BOS
2008	12	1	1550	B6	1416	442	AUS	MCO
2008	3	18	920	MQ	3364	437	AUS	DAL
2008	8	3	1040	AA	813	417	AUS	LAX
2008	3	18	945	AA	511	413	AUS	DFW
2008	4	24	700	DL	822	412	AUS	ATL
2008	8	11	1850	B6	1401	408	AUS	SFO

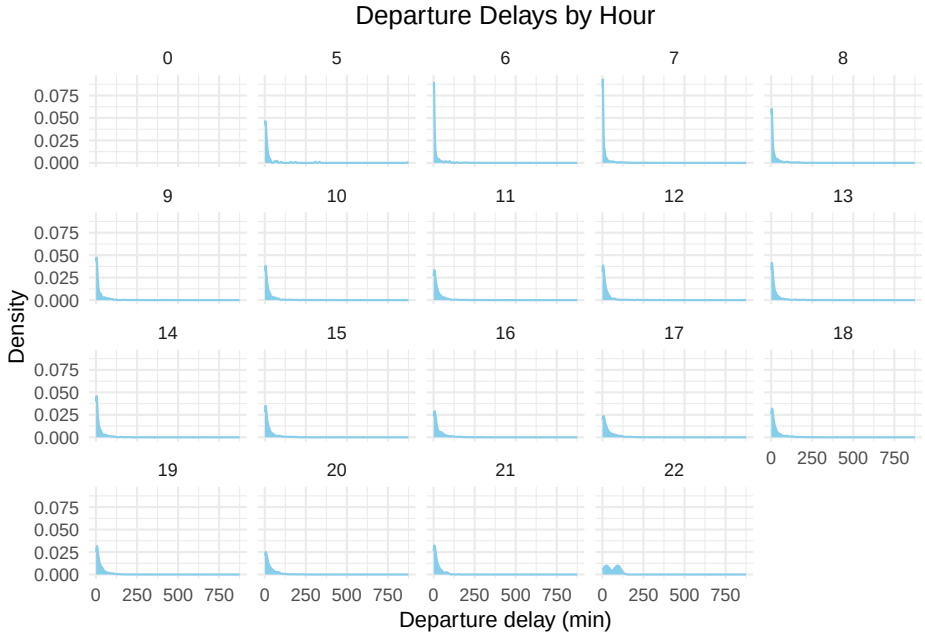
The following bar plots show the top 10 airlines with the most departure delays, in terms of number of delays and percentage of flights delayed. We noticed that there is an overlap of airlines in the two plots. Winair (WN) had significantly more delays than others, both in terms of count and percentage. Following Winair, ExpressJet (EV), Delta (DL), JetSuiteX (XE) and PSA Airlines (OH) had the highest percentage of flights delayed.

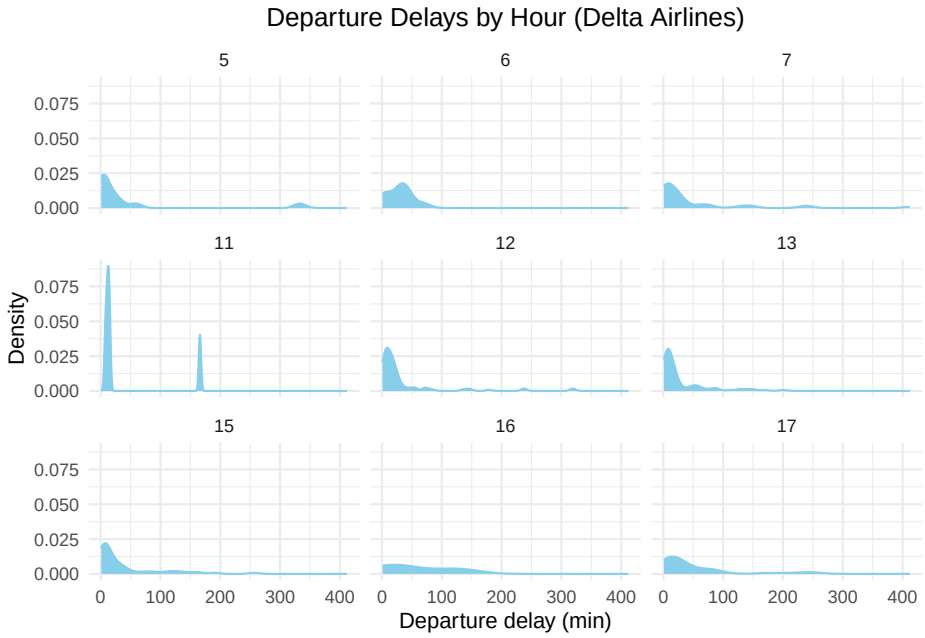




We then investigated whether delay patterns differ by hours of the day. From the density plot for all delayed flights, it seems that flights in the evenings tend to have longer delays than those in the mornings, reflected by thicker right tails.

Since Delta Airlines is a major commercial airline that had a lot of delayed flights, we wanted to know if it had the same patterns as the whole sample. The density plot shows that Delta didn't have flights departing after 6 pm, but it still experienced a lot of delays during the day. One interesting point is that their flights departing from 3 pm to 4 pm had a good number of delays over 2 hours. So it may be a good idea to choose another airlines if we want to leave Austin in the afternoon.





2. Wrangling the Olympics

In this part, we want to investigate some statistical facts about all Olympics medalists from 1896 till recent. This data set contains information including competitors' basic information, the categories of sports they play, the event in which they won the medal, etc.

A) What is the 95th percentile of heights for female competitors across all Athletics events ?

To answer this question, we first filtered the female players who played athletics sports from the original data set, then calculated the 95% quantile of heights. The 95% quantile of heights for female Athletics players is 183 cm.

```
##   percentile_95
## 1           183
```

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?

Standard deviation reflects the variability in data. We calculated the standard deviation of height grouped by each event, then found the event with the highest standard deviation. The event with the greatest variability in players' heights is Rowing Women's Coxed Fours, with a standard deviation of 10.9.

```
## # A tibble: 1 x 3
##   event                        mean    sd
##   <chr>                       <dbl> <dbl>
## 1 Rowing Women's Coxed Fours  173.  10.9
```

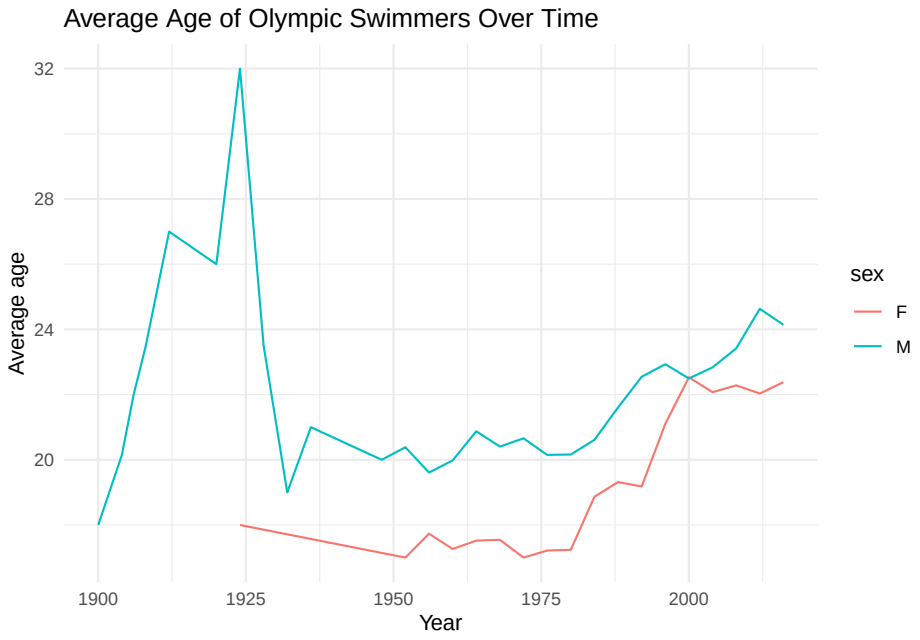
C) How has the average age of Olympic swimmers changed over time? Does the trend look different for male swimmers relative to female swimmers?

In the plot below, the x-axis is the year while the y-axis is the average age of swimmers. The two lines, with the blue one tracking the values of average age of male swimmers and the red one tracking that of female swimmers, tells us about how average age of swimmers changed from 1900s till recent. We can see that generally, the female swimmers are younger than male swimmers.

The average age of male swimmers reached a peak in 1925, which was 32. However, that age dramatically dropped in the following years to as low as below 20. About the growing average age of swimmers before 1925, we suspect that this pattern may be resulted by two reasons: 1) This sport was still young in the Olympics, so not many professional young swimmers were trained as successors. 2) Due to the ongoing wars and the unstable economic environment in the postwar period, young males were not able to professionally prepare for Olympics games. The sudden drop in age of swimmers indicates a turning point in which the older swimmers retired and younger ones joined the game.

The line of female competitors is shorter than that of males, which echoes with the fact that female athletics were not accepted in international sports events until that time. Starting to join Olympics in around 1920s, the average age of female swimmers has always been younger than that of male swimmers except for the year of 2000. In 2000, the average age of male and female competitors intersected at between 22-23 years old.

Generally speaking, the average age of both female and male swimmers slowly grew after 1950, with very slight fluctuations. This indicates that the selection and preparation for international professional swimmers has settled into a steady state.



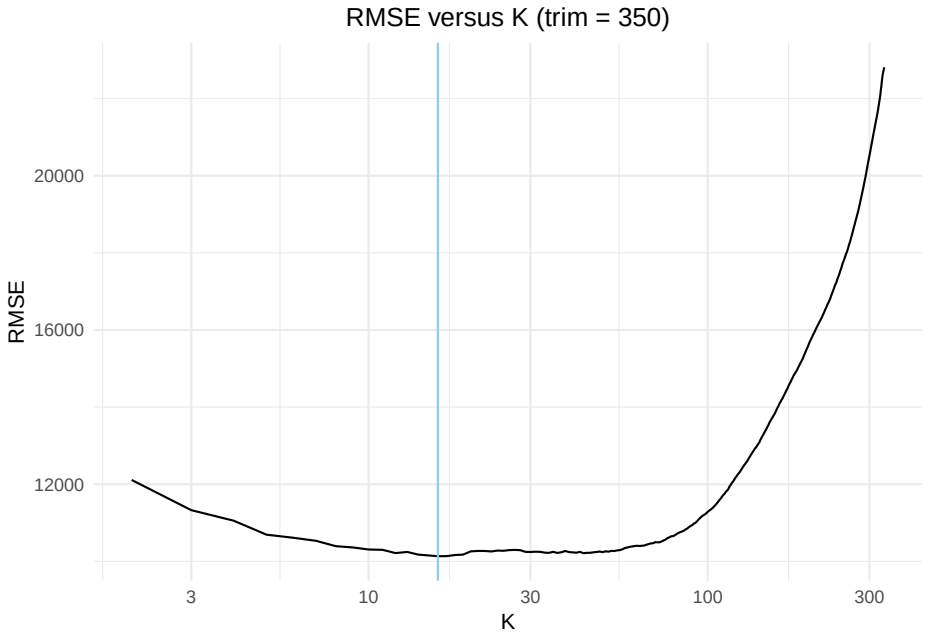
3. K-nearest neighbors: cars

Our goal is to use K-nearest neighbors to build a predictive model for price, given mileage, for cars with two trim levels: 350 and 65 AMG.

Cars with trim level 350

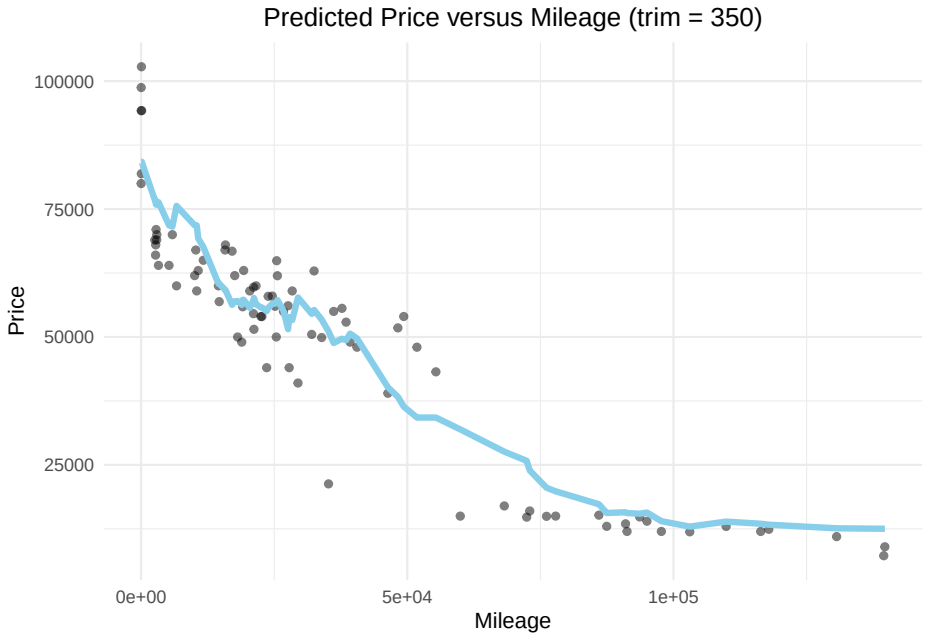
We first filtered cars with trim levels of 350, and split the data into training and testing data. To find the optimal K value for the KNN model, which minimizes RMSE on the training data, we calculated RMSE at different K values. The range of K values tested was from 2 to the number of total observations in the data set. We employed K-fold cross validation to enhance prediction accuracy of RMSE, selecting a K-fold value of 5. The results below show the K value that minimizes mean RMSE and a plot of RMSE versus K.

```
##            k      err  std_err
## result.15 16 10133.81 361.0058
```



The RMSE value changes with different splits of the data. After running the model multiple times, we found that the K value with the minimum RMSE falls in the range of 12 to 20. Typically, models with a lower RMSE (indicating reduced bias) tend to exhibit increased variance. To balance the bias-variance trade-off, we chose the median K = 16 as the optimal K value.

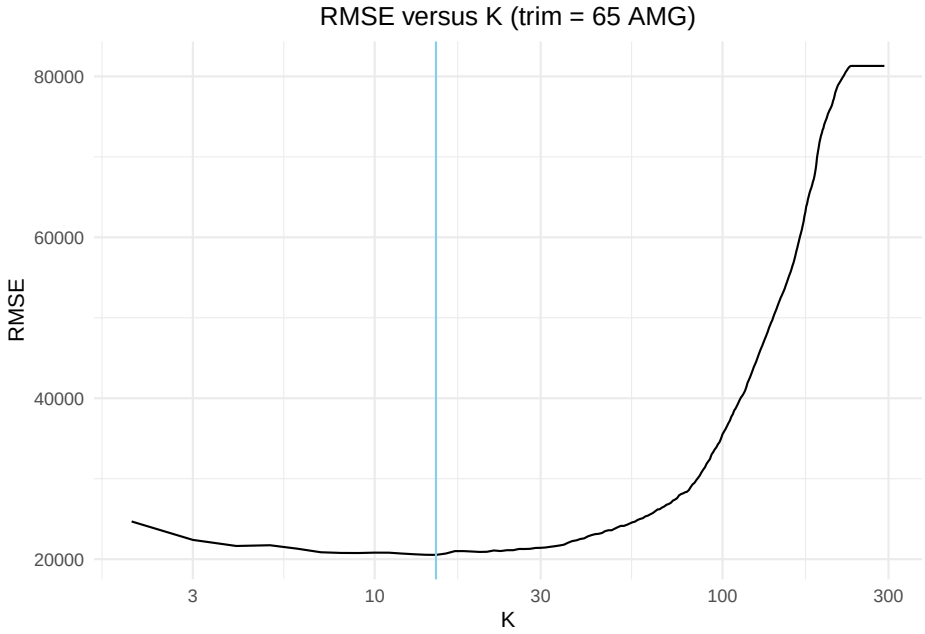
We used KNN model with K = 16 on the testing data to predict price based on mileage. The plot below shows predicted price versus mileage for cars with trim levels 350.



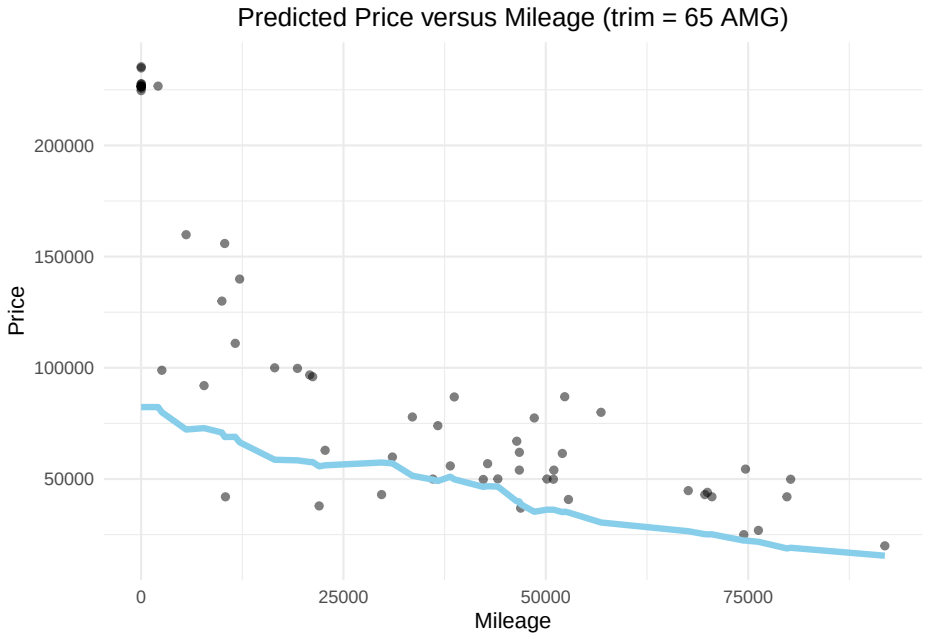
Cars with trim level 65 AMG

We performed similar procedures for cars with trim levels of 65 AMG. The K value with the minimum RMSE falls in the range of 9 to 30. To balance between bias and variance, K = 20 was chosen as the optimal K value.

```
##            k      err  std_err
## result.14 15 20534.75 2610.394
```



The plot below shows predicted price versus mileage for cars with trim levels AMG 65.



In conclusion, the Trim 65 AMG model, with optimal K value of 20, demonstrates a preference for a larger K value compared to the Trim 350 model, which has an optimal K value of 12. This distinction suggests that the Trim 65 AMG model benefits from a slightly higher level of complexity to accurately capture its underlying patterns, likely due to its data distribution or larger sample size.