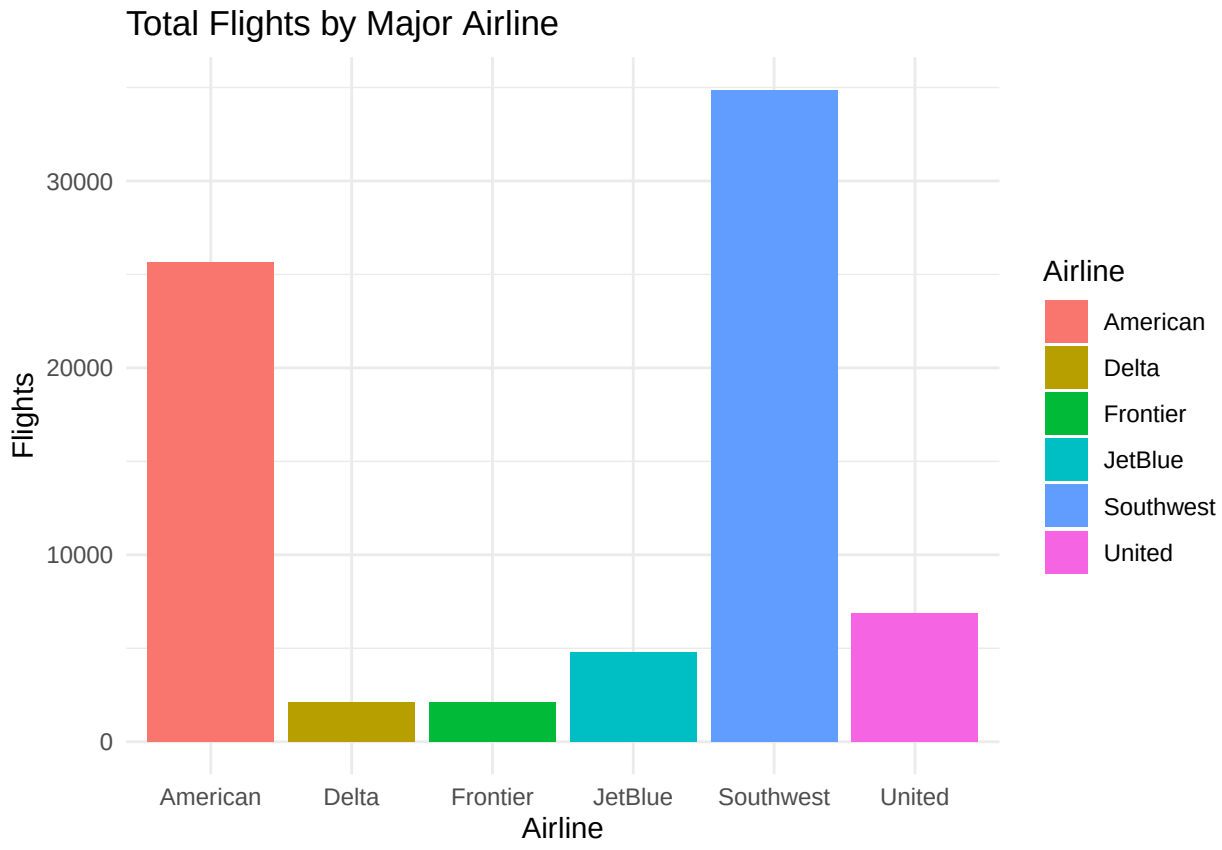
ECO 395M: StatLearning Exercise 1

Joseph Williams, Aahil Navroz, Suqian Qi

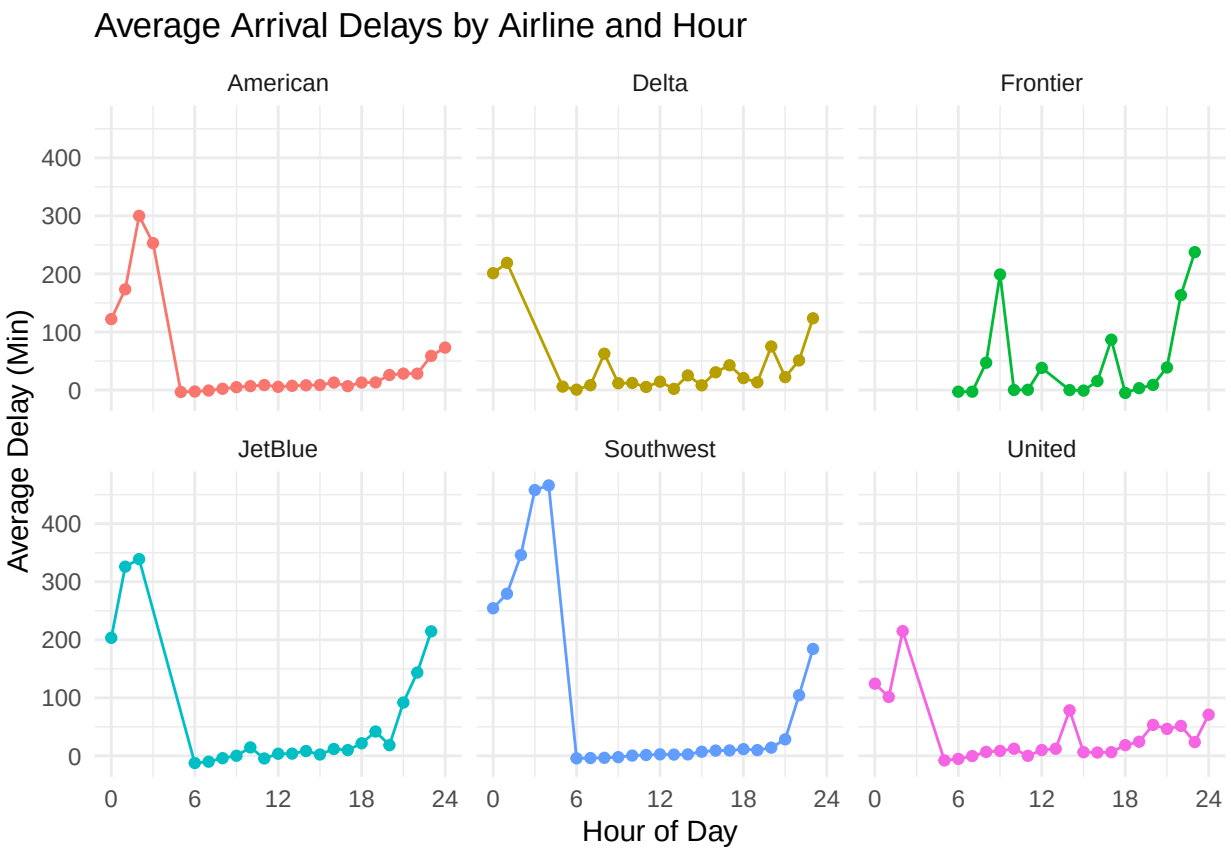
2024-02-07

1) Data visualization: flights at ABIA

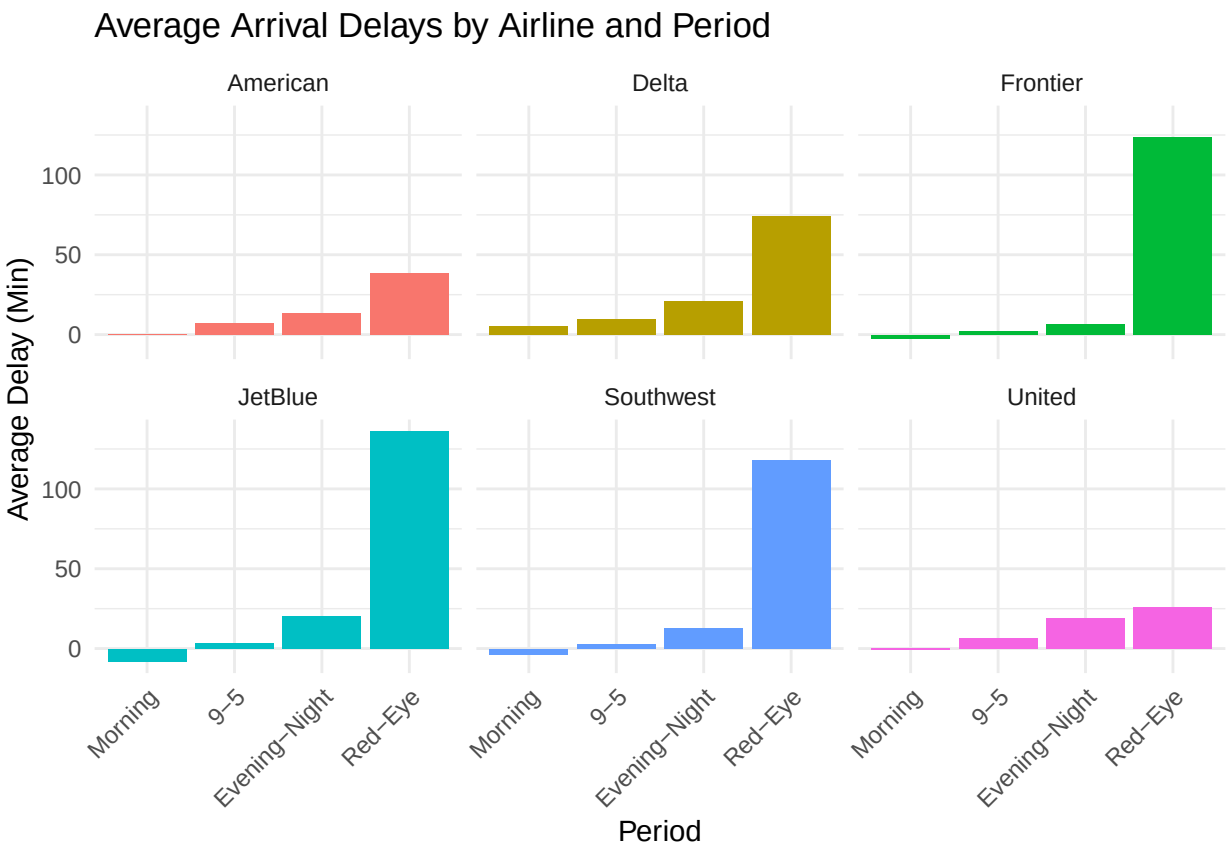
For this question we wanted to help fliers (say fliers in 2009) build intuition or 'rules of thumb' they can use when choosing between airline companies. To begin we want the data to correspond to recognizable names, so we mapped UniqueCarrier to airline brands or their parent brands. Here's a breakdown of which major carriers are running the most flights out of ABIA.



Looks like American and Southwest are king... but can they handle the volume!? Lets look at arrival delays for each company. Given these are arrival delays for flights coming into and out of Austin, overall it will be a fine measure for the timeliness of the airline.

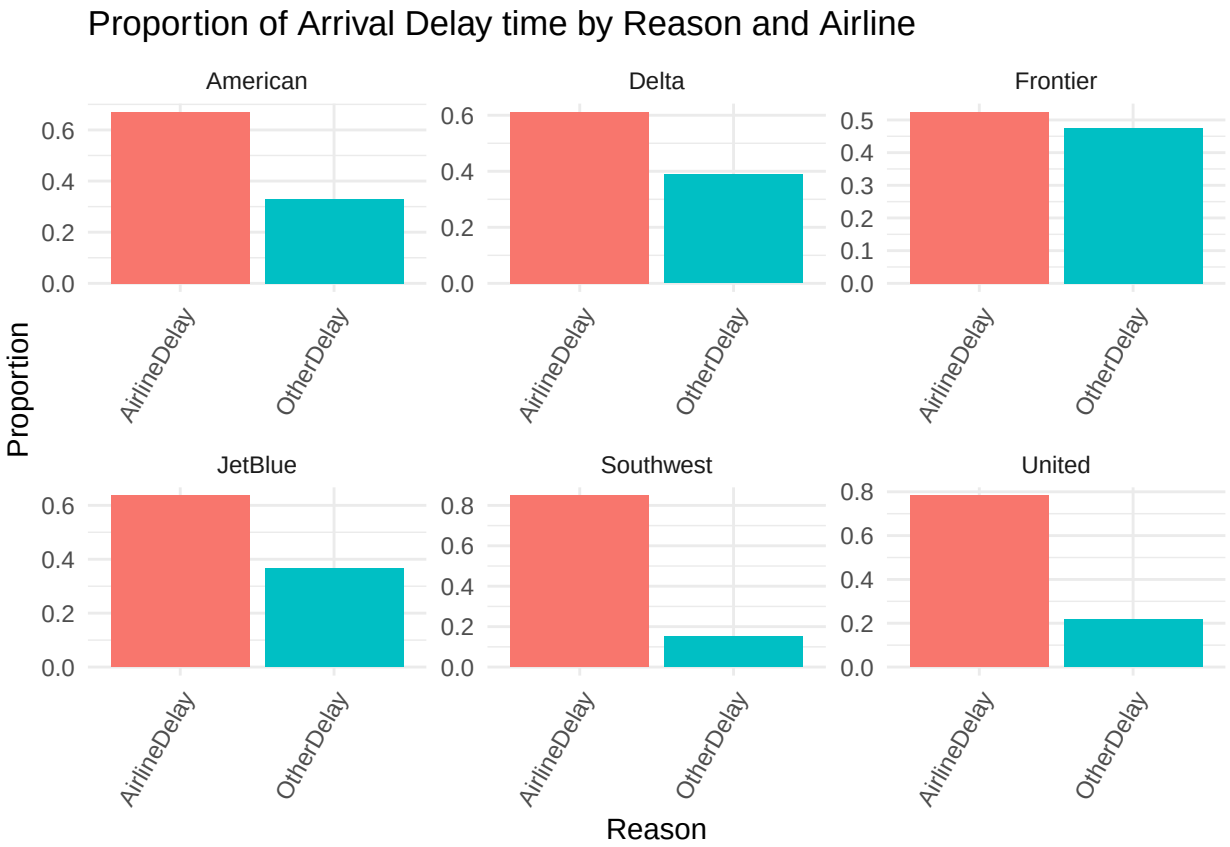


Okay, we're seeing some detail here. Seems like most companies experience their delays before the hour of 6am. Since we're looking for rules of thumb. Lets classify into 'Red Eye' 'Early Morning', '9-5' and 'Evening-Night', and see if we can quantify delay times over periods, rather than specific times. Some of these averages seem too high, too, lets remove observations where ArrDelay is more than 4 hours, since that usually results in a changed flight for me.



Okay now this is more useful! Let's see, if cost difference is minimal and its important that my flight goes smoothly, I am brand-indifferent for Morning and 9-5 flights, prefer Frontier for night flights, and will only fly United or American for red-eyes. Lets dig a little deeper to see where these delays are coming from. Perhaps we can gain even more insight about the airlines.

Here, we summarize arrival delay by delay reason to confirm earlier results. CarrierDelay and LateAircraftDelay are classified as 'AirlineDelay' and WeatherDelay, SecurityDelay, and NASDelay are classified as 'OtherDelay'. We want to see what percentage of the posted delays are actaully the airline's fault!



Of note: Southwest showed high average delay and high 'AirlineDelay' percentage, contrary to their strong reputation. Meanwhile, despite their poor reputation, Frontier shows the lowest AirlineDelay percentage. Just don't use them for red-eyes!

2) Wrangling the Olympics

A) What is the 95th percentile of heights for female competitors across all Athletics events (i.e., track and field)?

```
## 95%
## 183
```

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

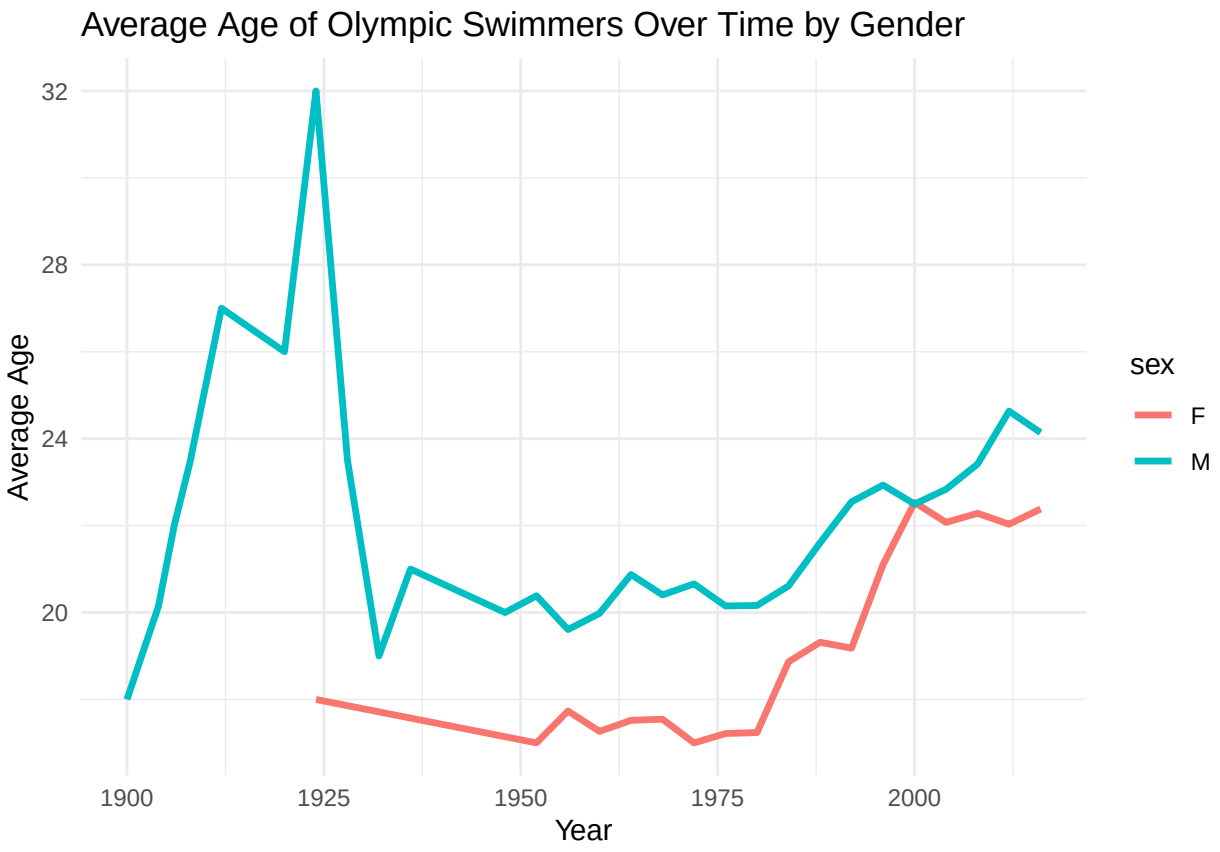
B) Which single women's `event` had the greatest variability in competitor's heights across the entire history of the Olympics, as measured by the standard deviation?

```
## [1] "Rowing Women's Coxed Fours"
```

```
## [1] 10.86549
```

Rowing Women's Coxed Fours had the greatest variability in competitor's heights across the entire history of the Olympics and its corresponding standard deviation is 10.86549.

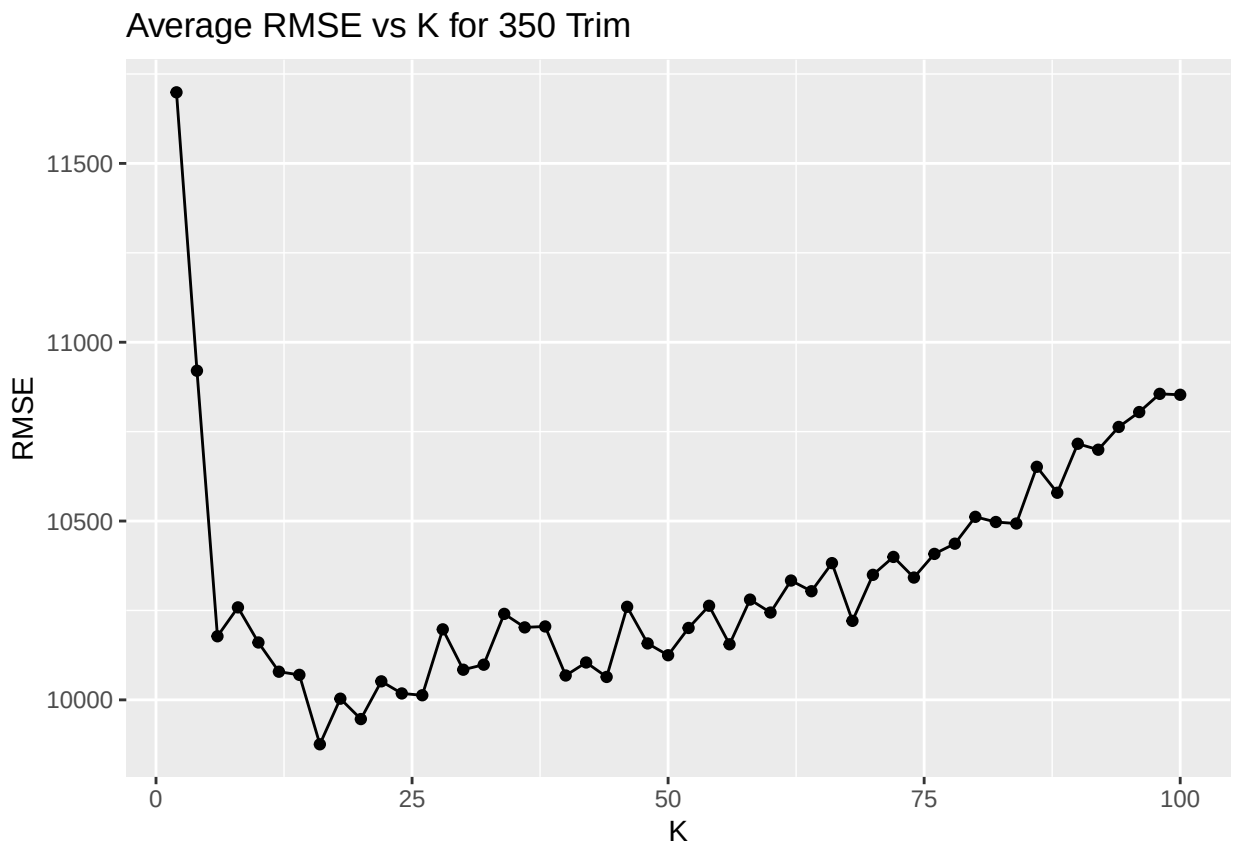
C) How has the average age of Olympic swimmers changed over time? Does the trend look different for male swimmers relative to female swimmers?

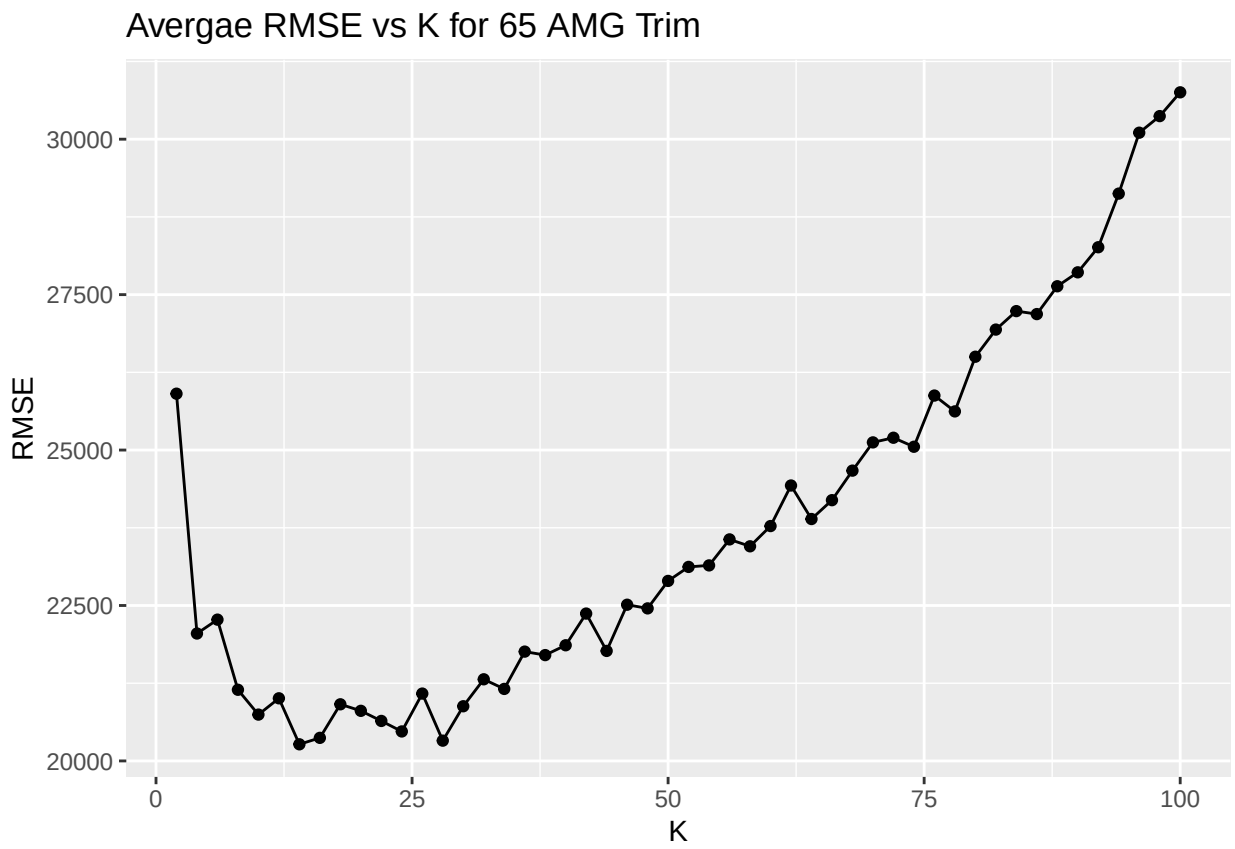


Average age by gender has increased since 1920s for both type's of athletes but began to increase more rapidly for women leading up to year 2000. Suggesting women are competing at the highest level of sport later in their lives, on average. This likely a result of increased popularity of woman's sports, effects of title IX.

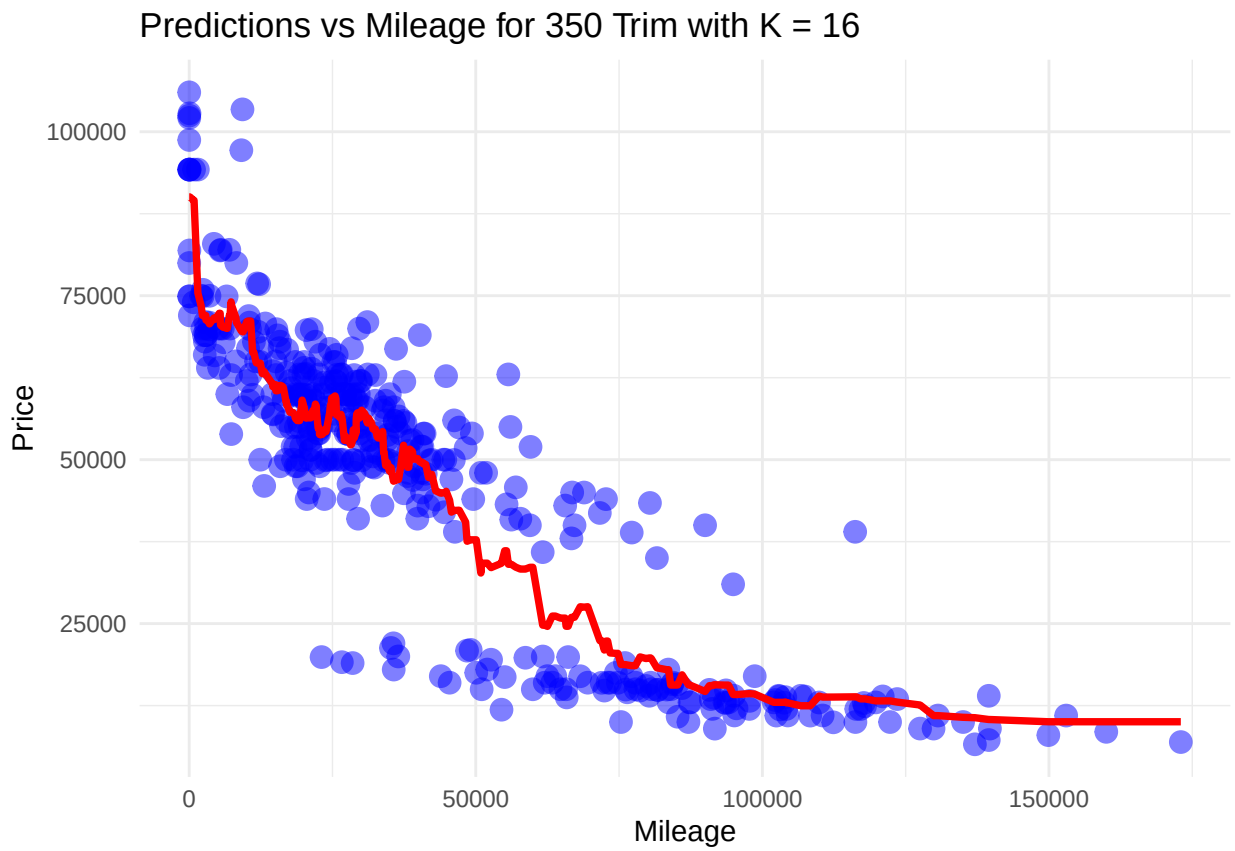
3) K-nearest neighbors: cars

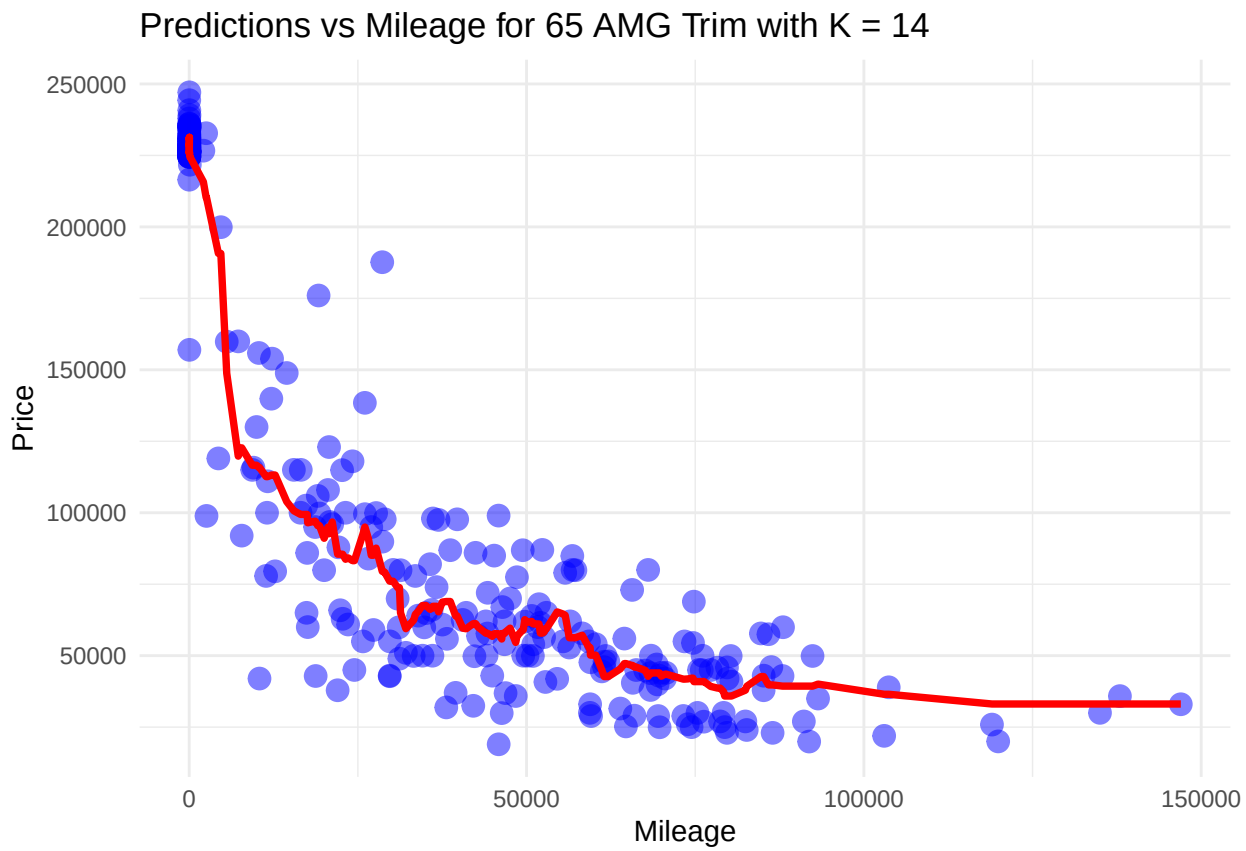
Question: For each trim, make a plot of RMSE versus K, so that we can see where it bottoms out. Then for the optimal value of K, show a plot of the fitted model, i.e. predictions vs. x. (Again, separately for each of the two trim levels.)


Average RMSE vs K for 350 Trim
RMSE
11500
11000
10500
10000
0
25
50
75
100
K



We have shown the RMSE for different K values vary from 10 to 30 for both trims. Using k-fold cross validation with fold = 10 we show an optimal k of 16 for the 350 trim and a k of 14 for the 65_AMG trim.


Predictions vs Mileage for 350 Trim with K = 16
Price
100000
75000
50000
25000
0
50000
100000
150000
Mileage



The higher K of 16 for the 350 trim could be explained by the large gap in the data around the 50,000 mileage mark. A larger K is needed here to account for bias for observations in that range.