

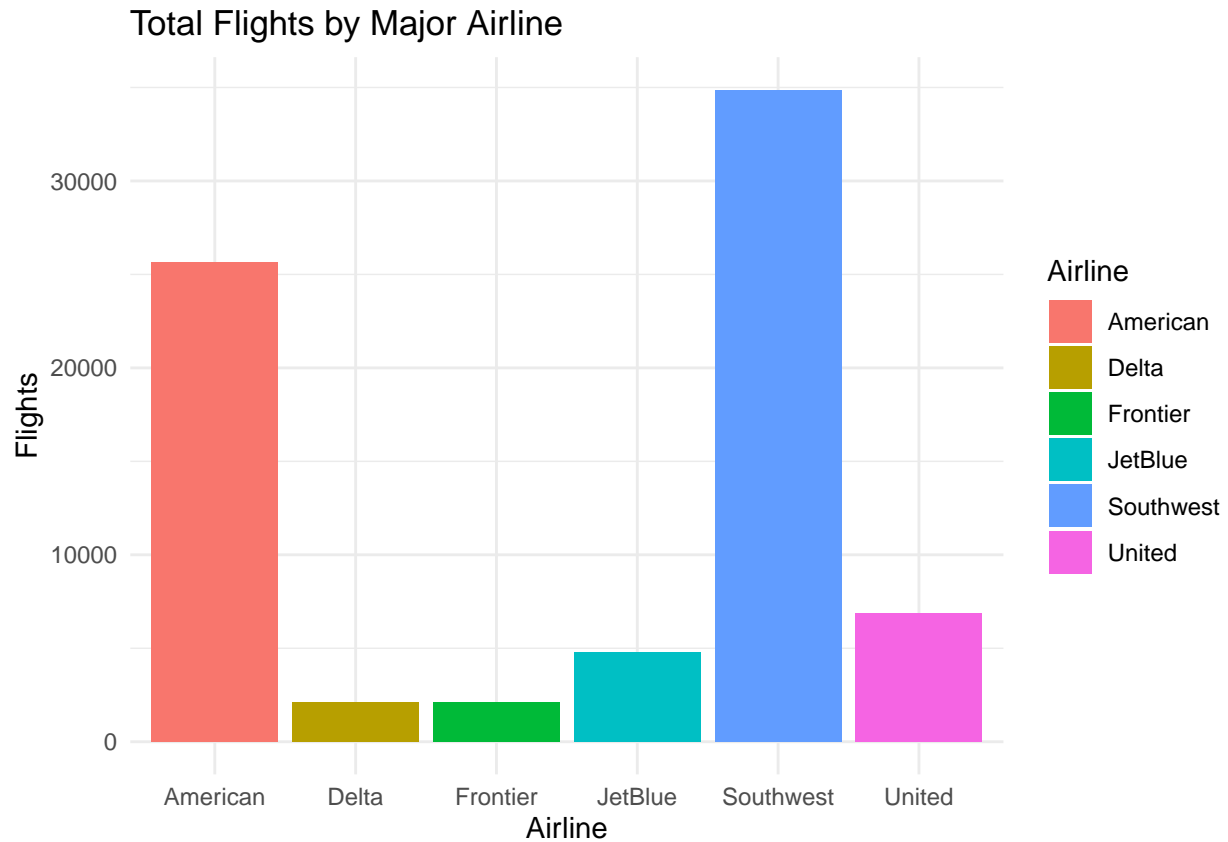
ECO 395M: StatLearning Exercise 1

Joseph Williams, Aahil Navroz, Suqian Qi

2024-02-04

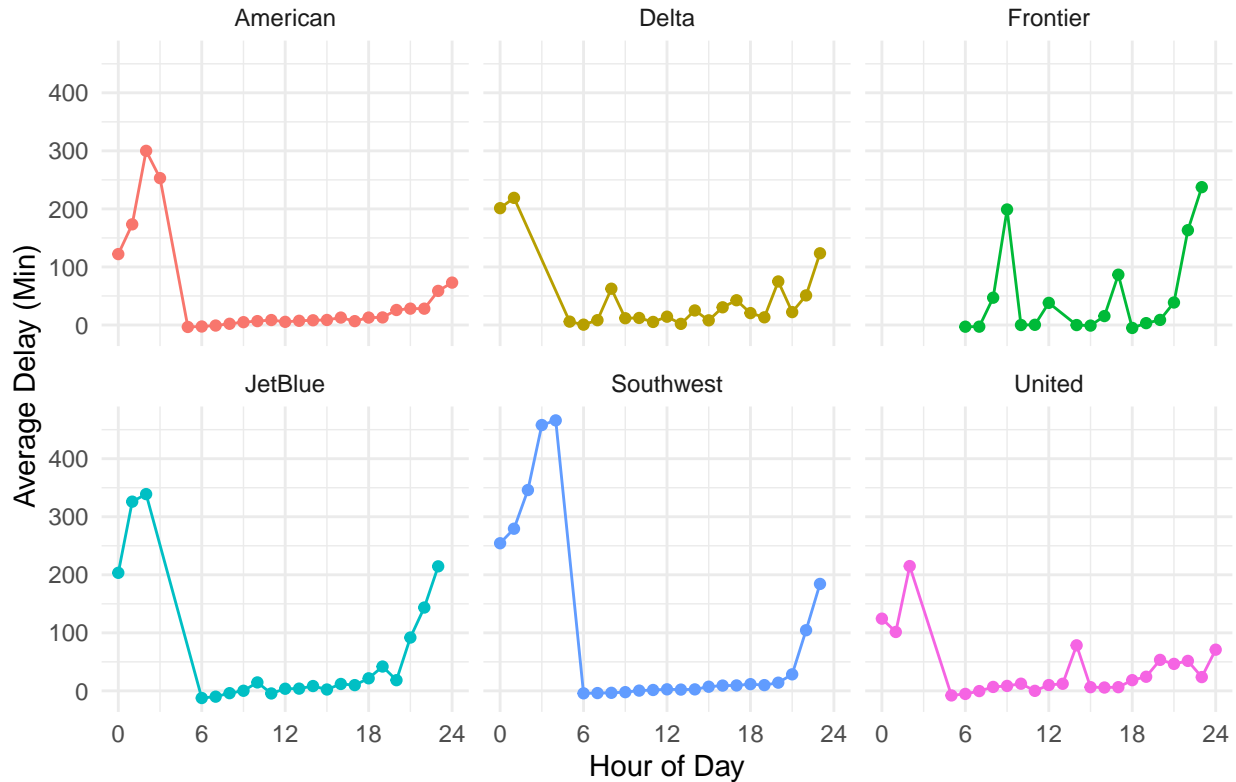
1) Data visualization: flights at ABIA

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.



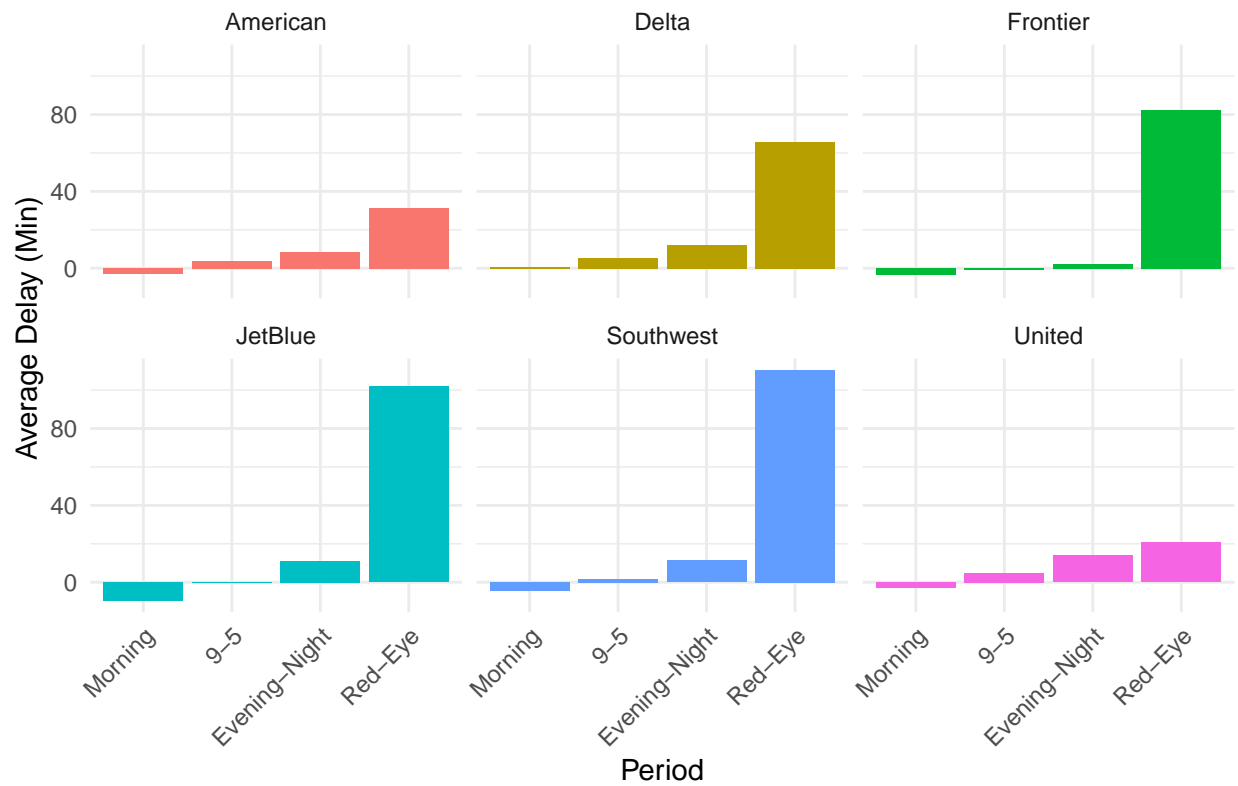
Looks like American and Southwest are king... but can they handle the volume!? Lets look at arrival delays for each company. Given these are arrival delays for flights coming into and out of Austin, overall it will be a fine measure for the timeliness of the airline.

Average Arrival Delays by Airline and Hour



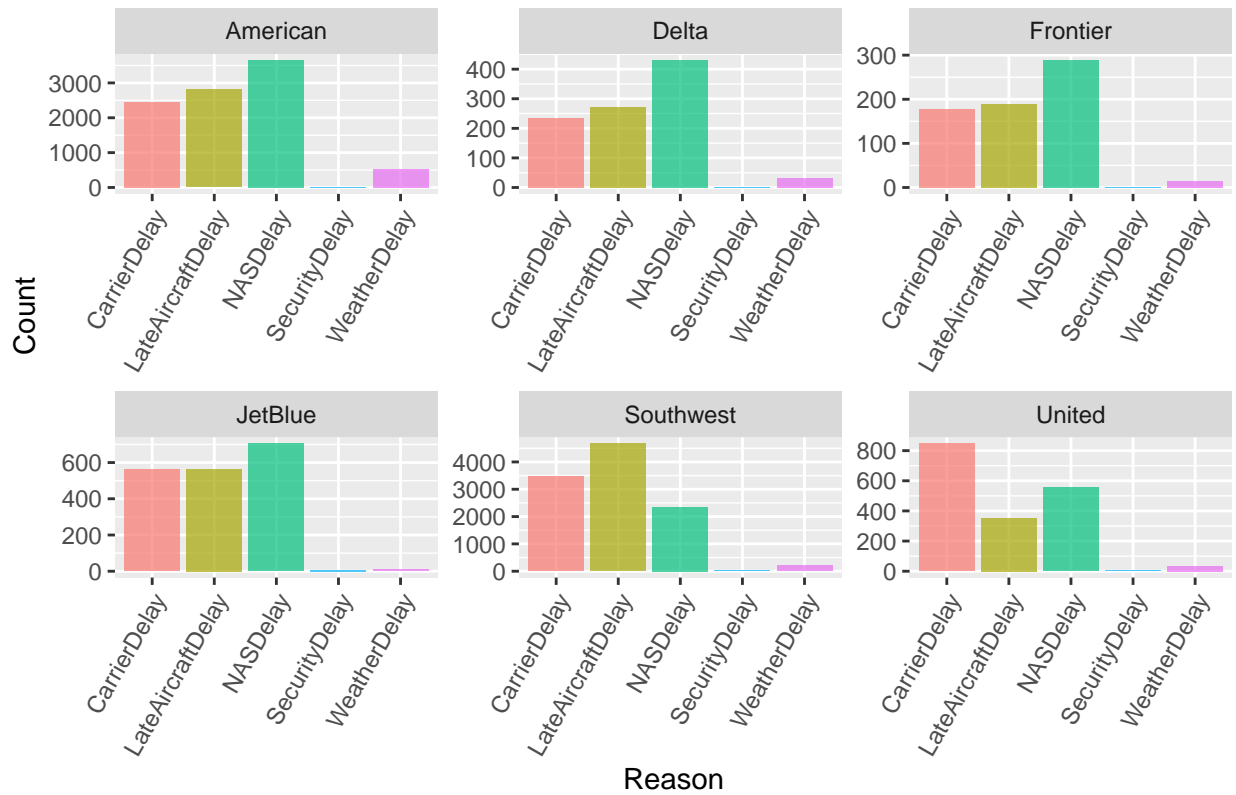
Okay, we're seeing some detail here. Seems like most companies experience their delays before the hour of 6am. Since we're looking for rules of thumb. Lets classify into 'Red Eye', 'Early Morning', '9-5' and 'Evening-Night', and see if we can quantify delay times over periods, rather than specific times. 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. Lets also subtract WeatherDelay, SecurityDelay, and NASDelay from ArrivalDelay since those aren't related to the airline.

Average Arrival Delays by Airline and Period

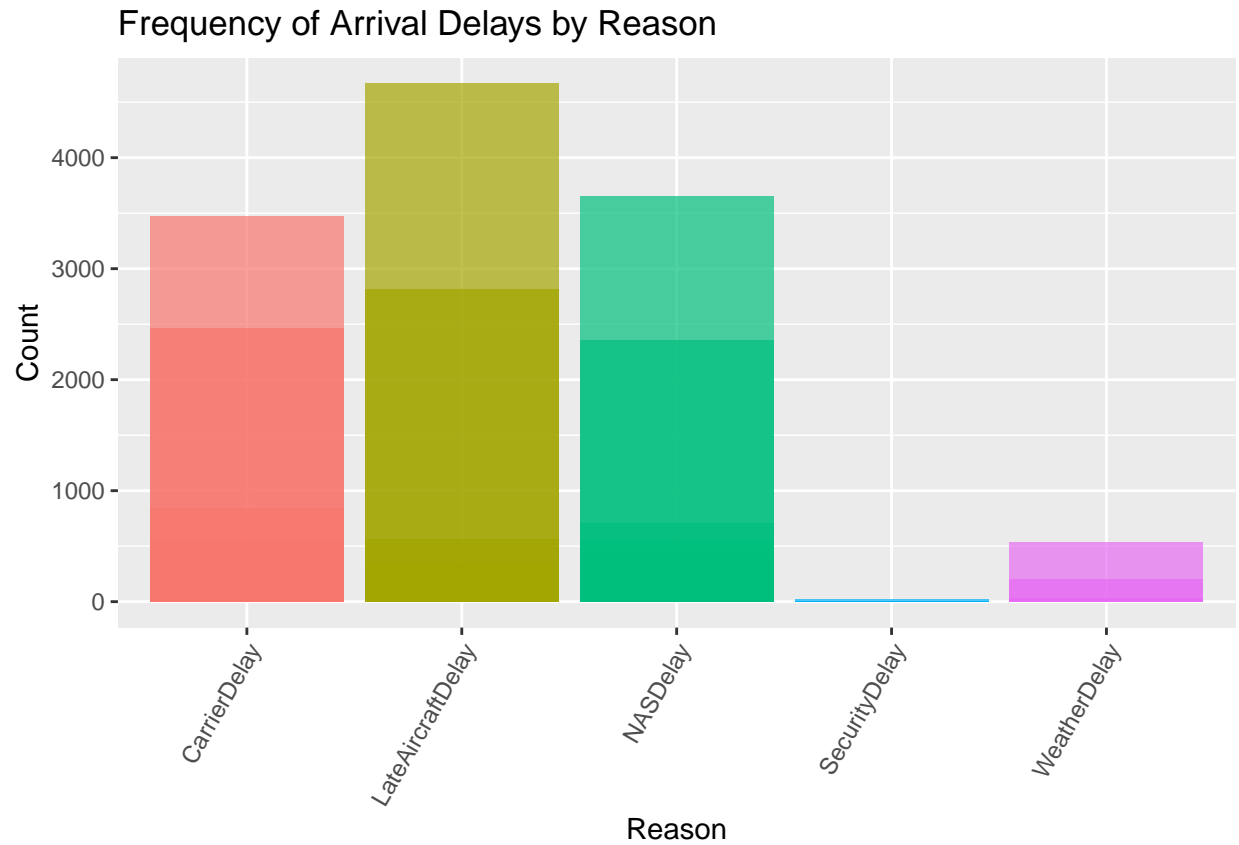


Okay now this is more useful! Let's see, if cost difference is minimal and its important that my flight goes smoothly, I am brand-indifferent for Morning and 9-5 flights, prefer Frontier for night flights, and will only fly United for red-eyes. You're welcome, 2009 travelers!

Frequency of Arrival Delays by Reason and Airline

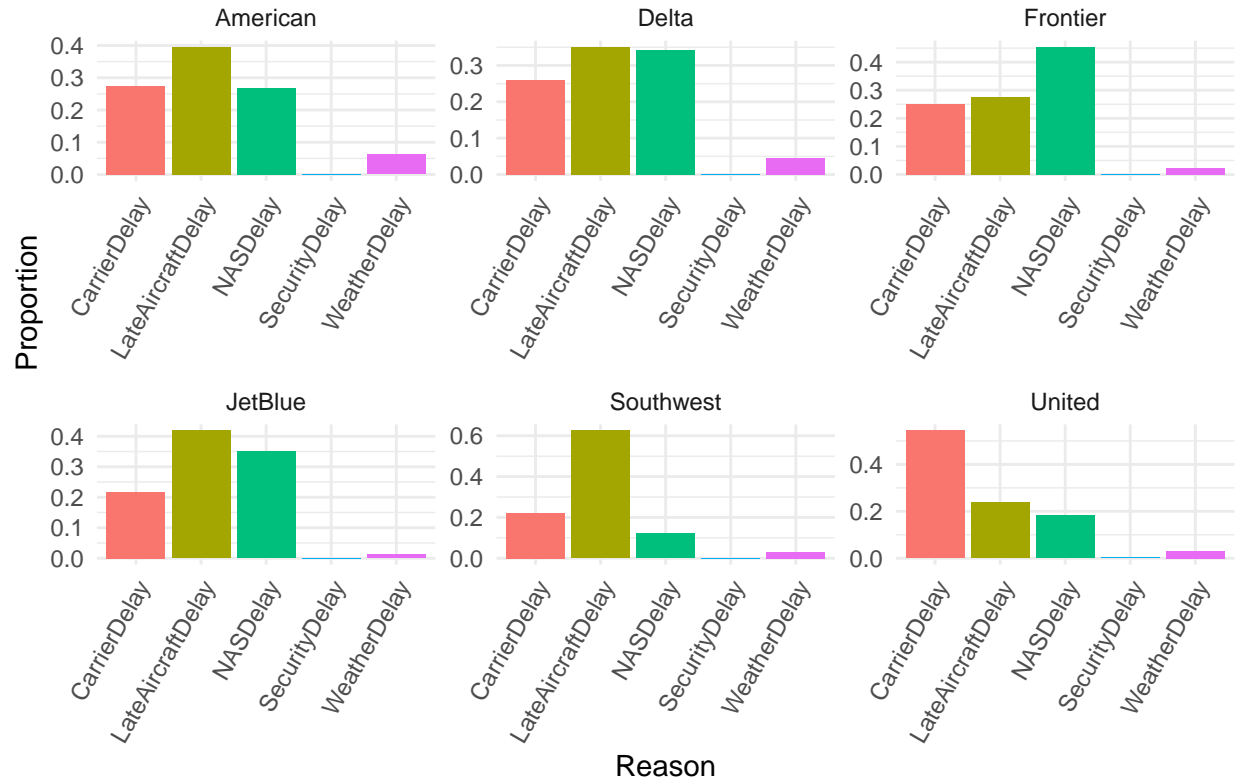


Through the plot we can see for American, Delta, Frontier and JetBlue, the most common reason for delay is NASDelay. For Southwest, the most common reason is LateAircraftDelay and for United it's CarrierDelay.

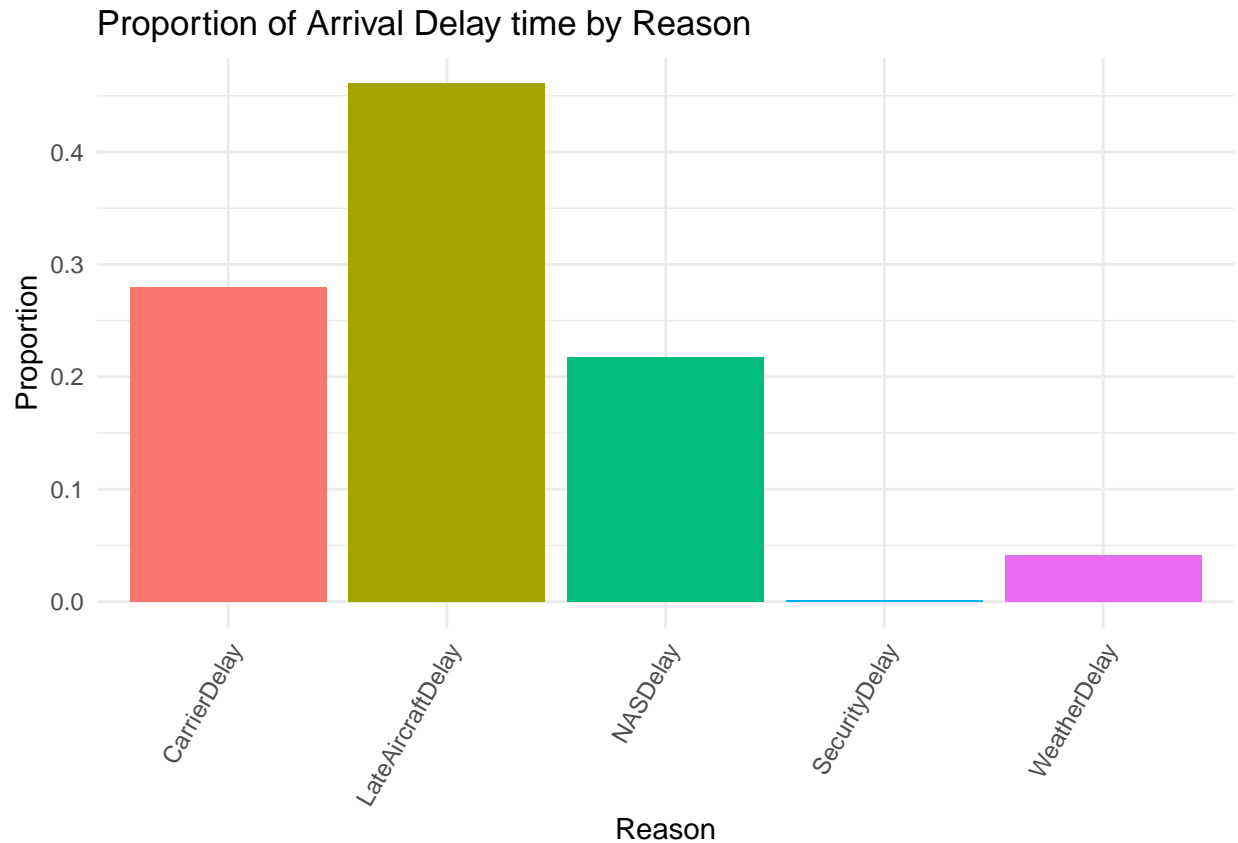


However, considering the far ahead flight amount of Southwest Airlines, the overall most common reason for delay is LateAircraftDelay. Now, let's see the proportion of delay time for each reason whether corresponding to their frequency.

Proportion of Arrival Delay time by Reason and Airline



We can see for American, Delta and JetBlue, the LateAircraftDelay takes the most part of delay instead of NASDelay while the others remains the same.



With no surprise, LateAircraftDealy is the most common reason for all flights. However, all these conclusion may not be valid, since about 85% of the data are missing for delay reason. All our research are based on the rest 15% of the overall data.

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
```

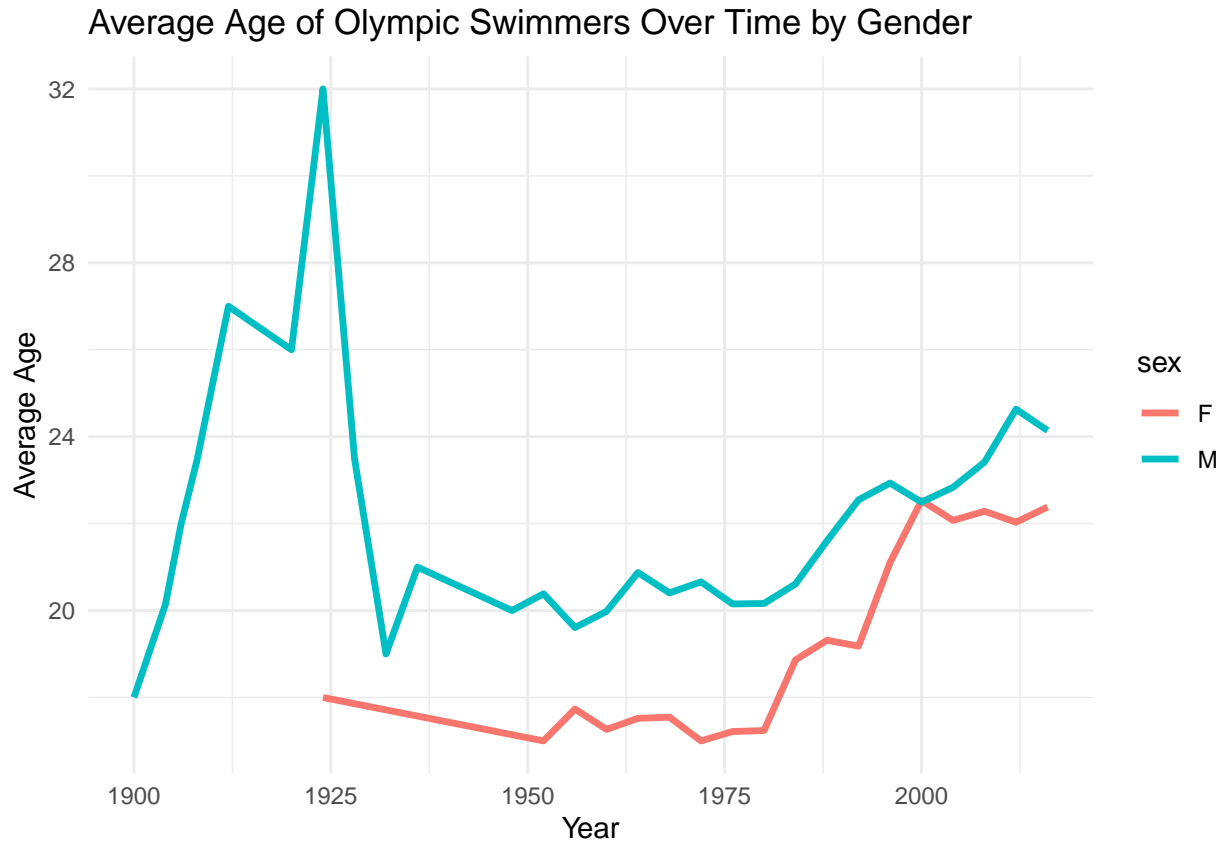
The 95th percentile of heights for female competitors across all Athletics events is 183.

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?

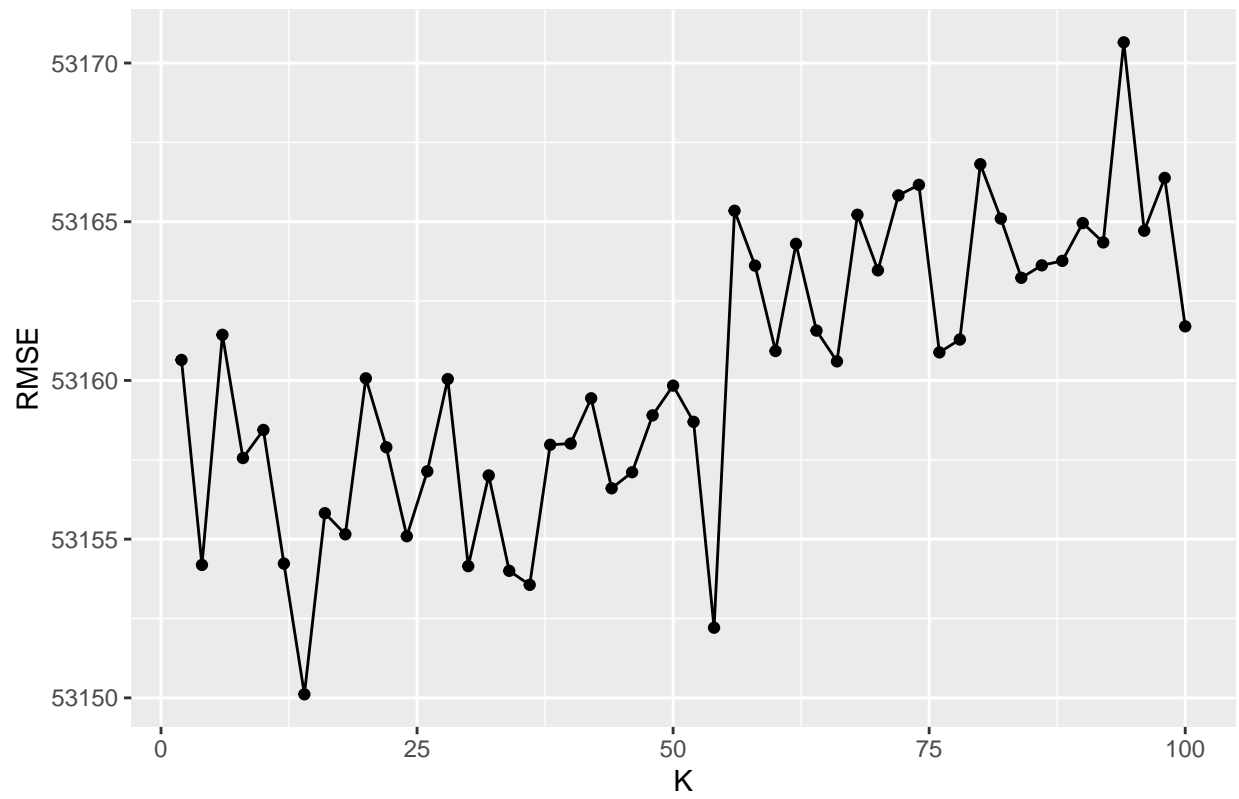


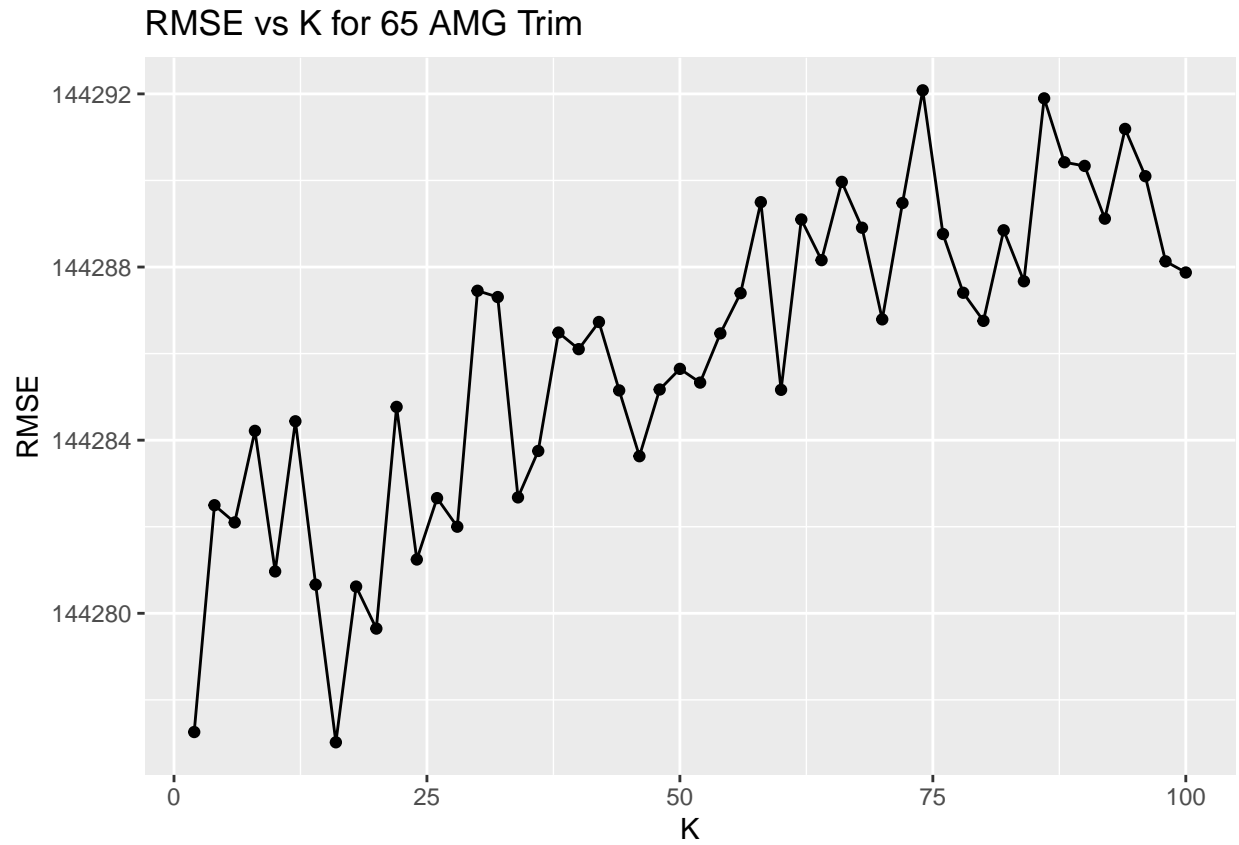
Average age by gender has increased since 1920s for both type's of athletes but began to increase more rapidly for women leading up to year 2000. Suggesting women are competing at the highest level of sport later in their lives, on average. This likely a result of increased popularity of woman's sports,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

3) K-nearest neighbors: cars

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.)

RMSE vs K for 350 Trim





We have shown the RMSE for different K values vary from 2 to 60 of both trims. For trim 350, the best K value is 14 and for trim 65 AMG, the best K value is 16.

