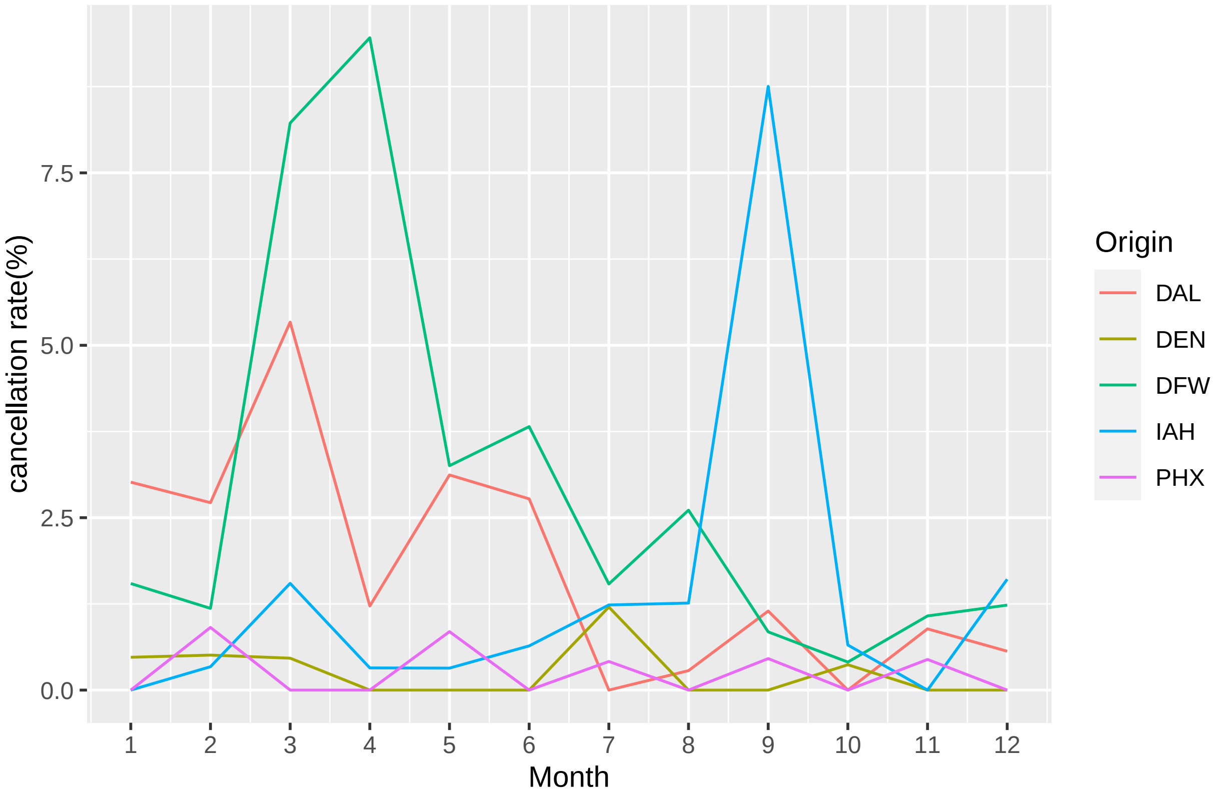
ECO395M Homework1

Evan Aldrich, Chenxin Zhu, Somin Lee

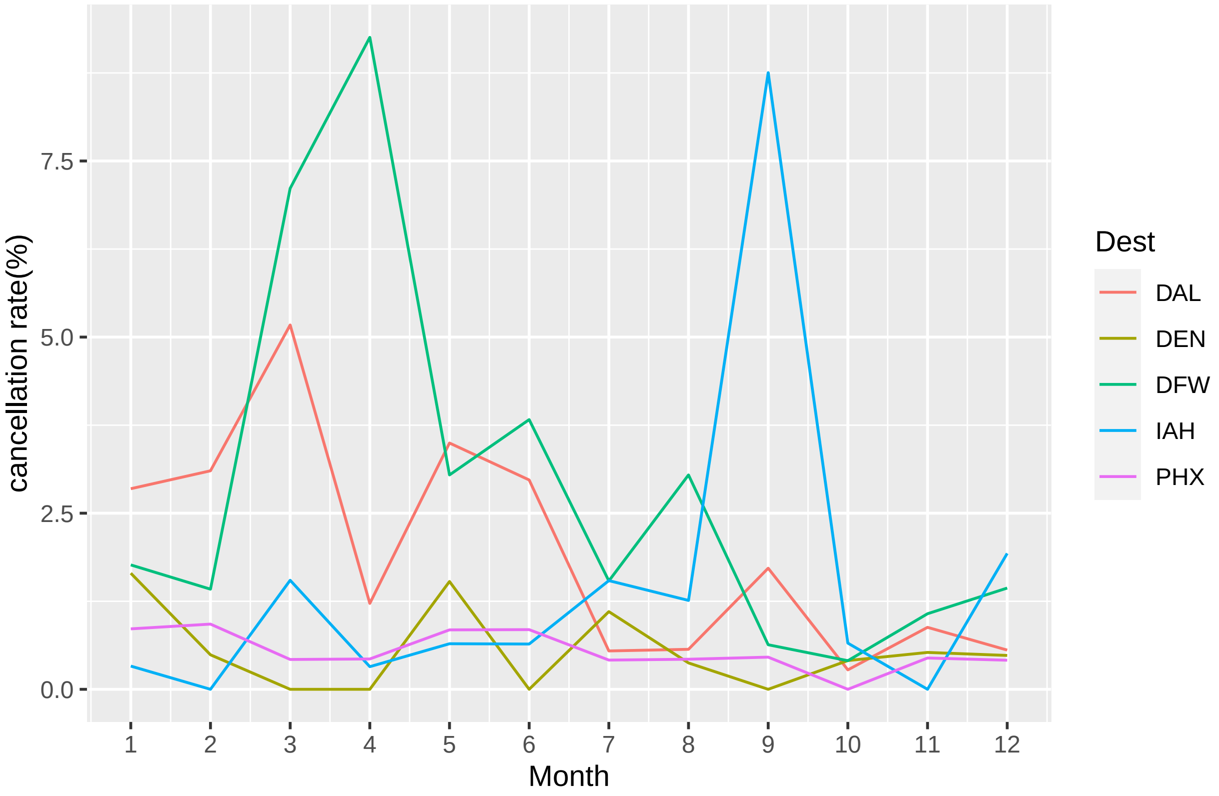
2024-01-20

Question 1: Data visualization: flights at ABIA

Montly cancellation rate of top 5 flight to AUS

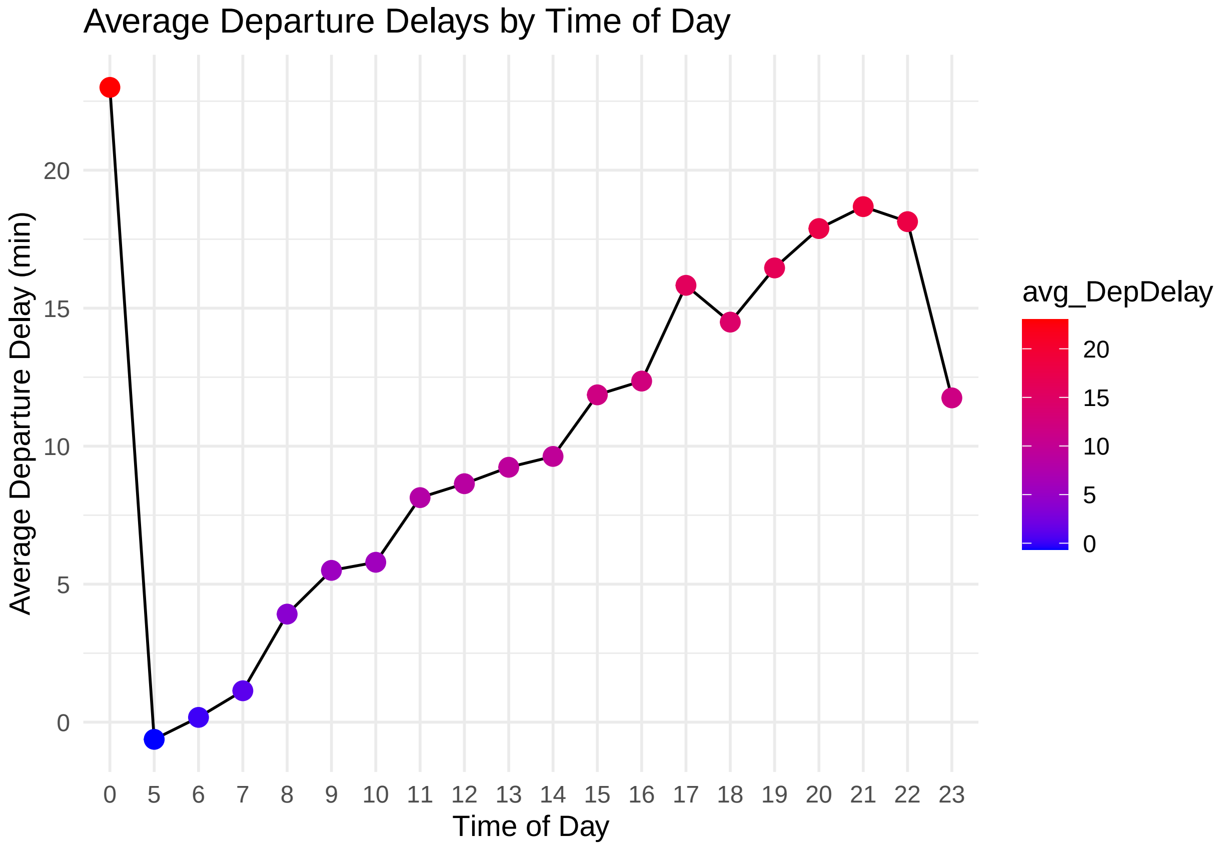


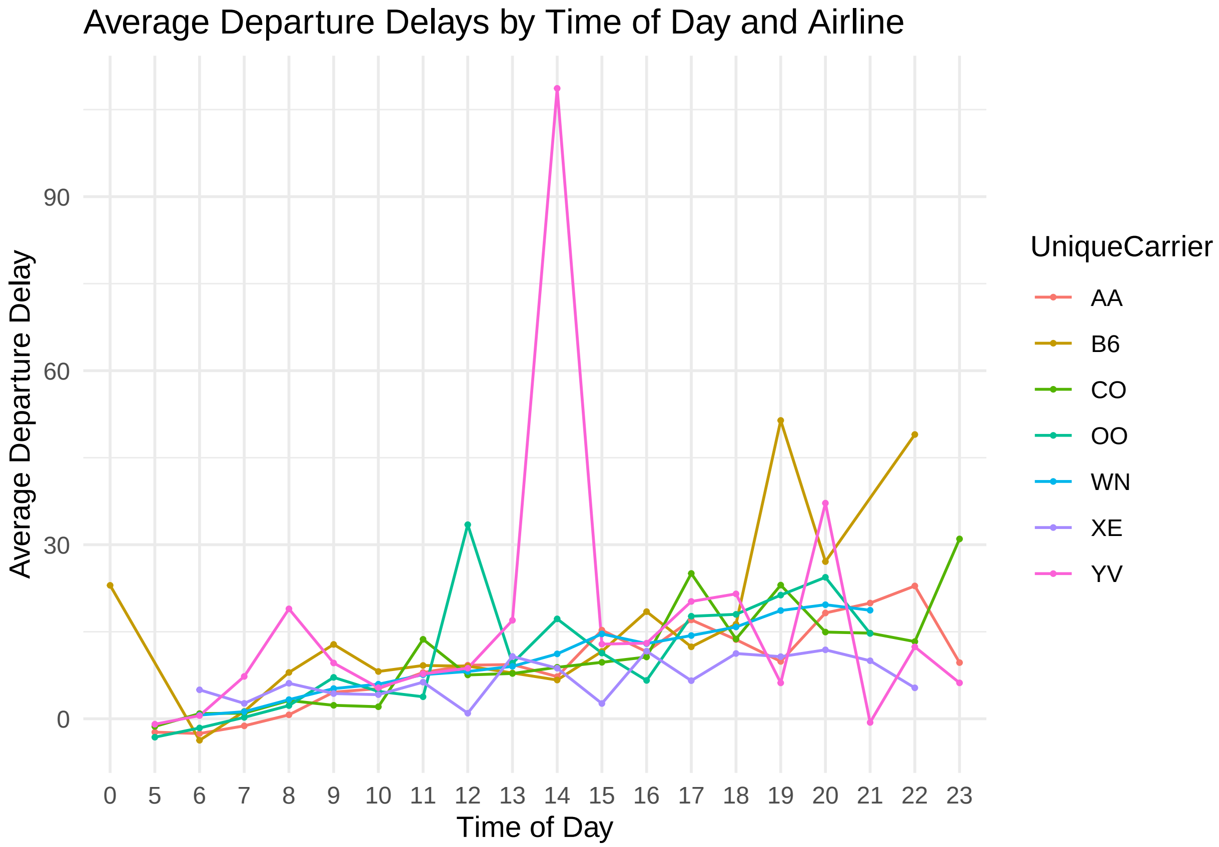
Montly cancellation rate of top 5 flight from AUS



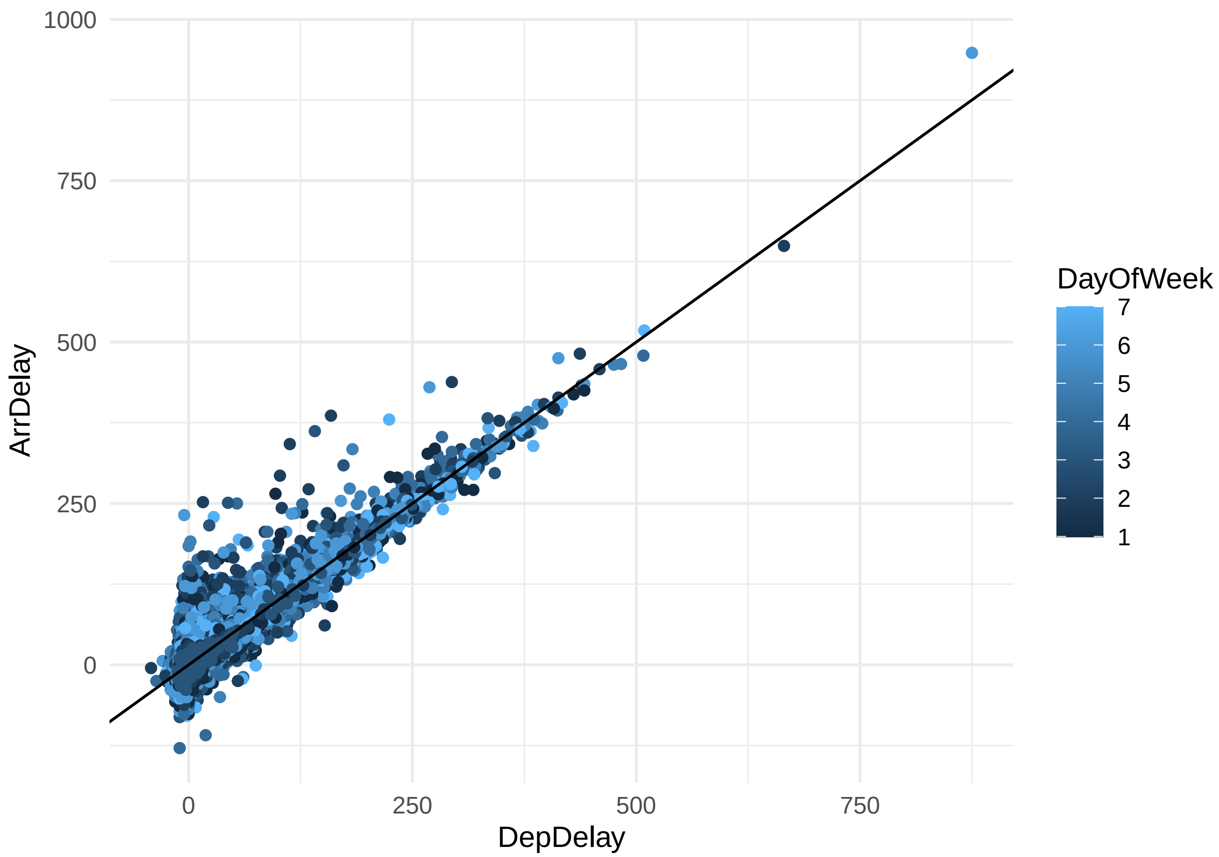
The two graphs above showcase the cancellation rates to and from Austin from Austin's 5 most common flight destinations. We see very similar patterns in the to and from graphs. What we can see is the months that have the highest cancellation rates are the months of March and April as well as well as August and September. This could be explained by the large events taking place in Austin in those months, in early March is South by Southwest, the massive media conference and festival, and in August and September is the massive music festival Austin City Limits. Later in the year we have the smallest cancellation rates. So we see that potentially major events alter the flights across the year but when is the best time of the day to fly, that's observed below.

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





The first of the two graphs above showcases that when we aggregate across all airlines, the average departure delay is much lower in the early mornings. The later in the day we look, and into the evening, the later we can expect an airline to be departing compared to its normally scheduled time. The second graph above splits up the data to showcase the top 7 airlines that fly from Austin and how their departure delays change over the course of the day. Interestingly, we have an outlier at 2pm. Assuming the data is true, this is likely due to airline YV being a local airline based in Phoenix, Arizona and during the early to mid afternoon, it may be particularly hot in Arizona and planes depart late due to whether conditions. Again, we can see a positive trend showing that the later in the day we leave, the later we can expect the departure delay to be. Our Hypothesis is that if a plane departs late it likely arrives late and thus delays add up over the day. This is observed in the next graph.



As we can see above, most of the points lie on or above the line that equates departure delay to arrival delay, there are not very many points that lie well below the line. This gives evidence for the trend we see in the previous graphs that when planes depart late from one location, they may arrive even later in the next location, so throughout the day, we can observe an increase in departure delays.

Question 2: Wrangling the Olympics

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

```
## [1] 183
```

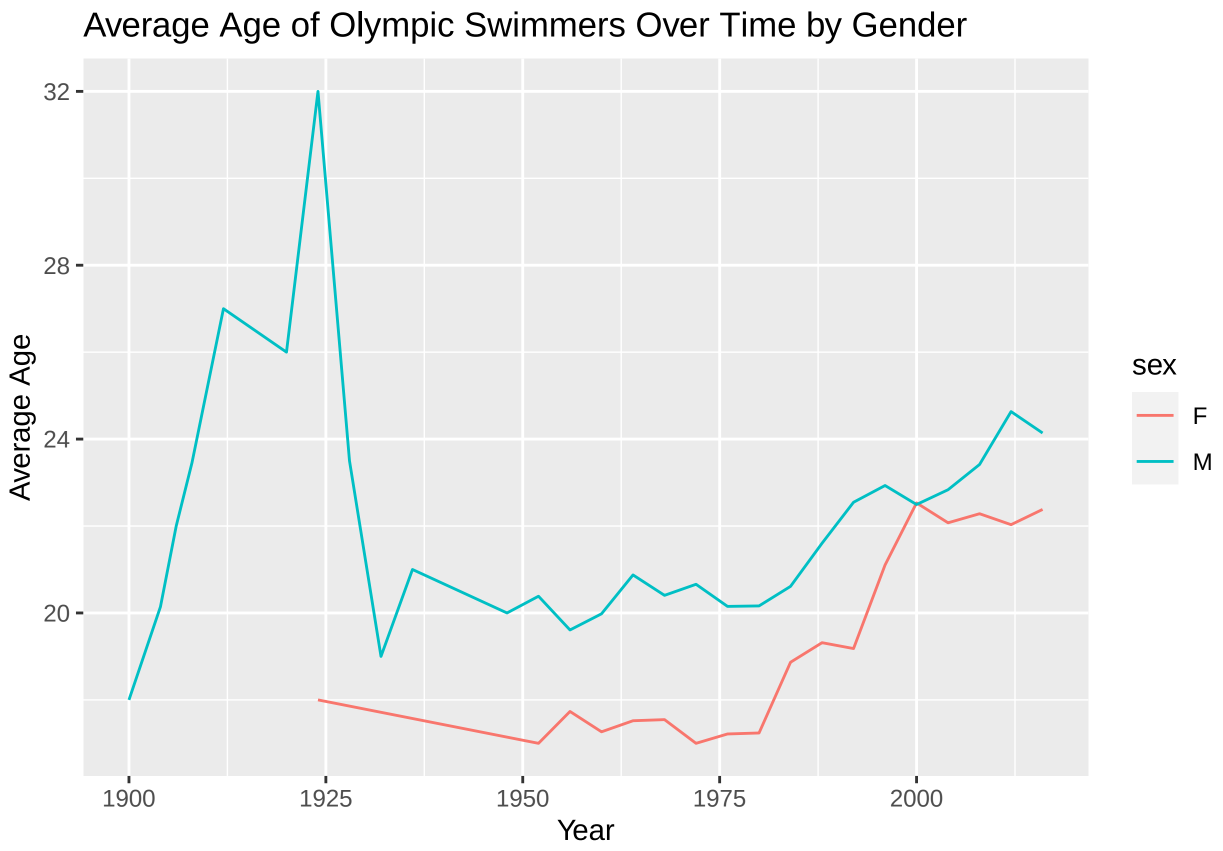
Across all events under the sport "Athletics" the 95th percentile for female competitors is 183cm or just about 6 feet tall.

Part 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?

```
## # A tibble: 10 x 2
##    event                                  std_dev_height
##    <chr>                                           <dbl>
##  1 Rowing Women's Coxed Fours                      10.9 
##  2 Basketball Women's Basketball                    9.70
##  3 Rowing Women's Coxed Quadruple Sculls            9.25
##  4 Rowing Women’s Coxed Eights                     8.74
##  5 Swimming Women’s 100 metres Butterfly           8.13
##  6 Volleyball Women’s Volleyball                   8.10
##  7 Gymnastics Women’s Uneven Bars                  8.02
##  8 Shooting Women’s Double Trap                    7.83
##  9 Cycling Women’s Keirin                          7.76
## 10 Swimming Women’s 400 metres Freestyle           7.62
```

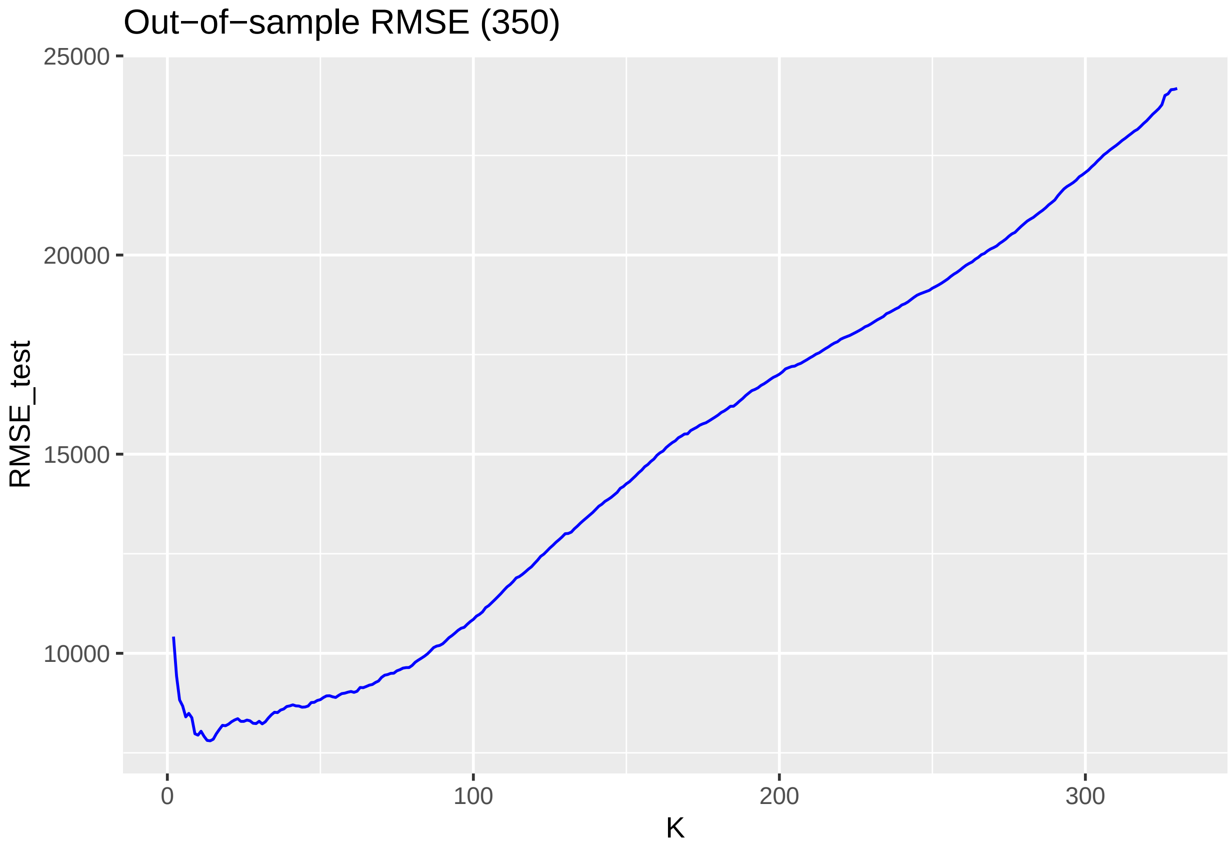
Rowing Women’s Coxed Fours has the highest variability of all women’s Olympic events measured in the data set. This could be a result of the event only taking place 4 times in the Olympics. Additionally, in rowing, for the most part, the height of the coxswain does not impact the rowing ability of the rowers in the boat (in this case 4) so they may be shorter than the rowers increasing the variability in height of those on the boat. This in-part may explain why 3 of the top 4 events with the most varied heights are rowing events with a coxswain.

Part 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.



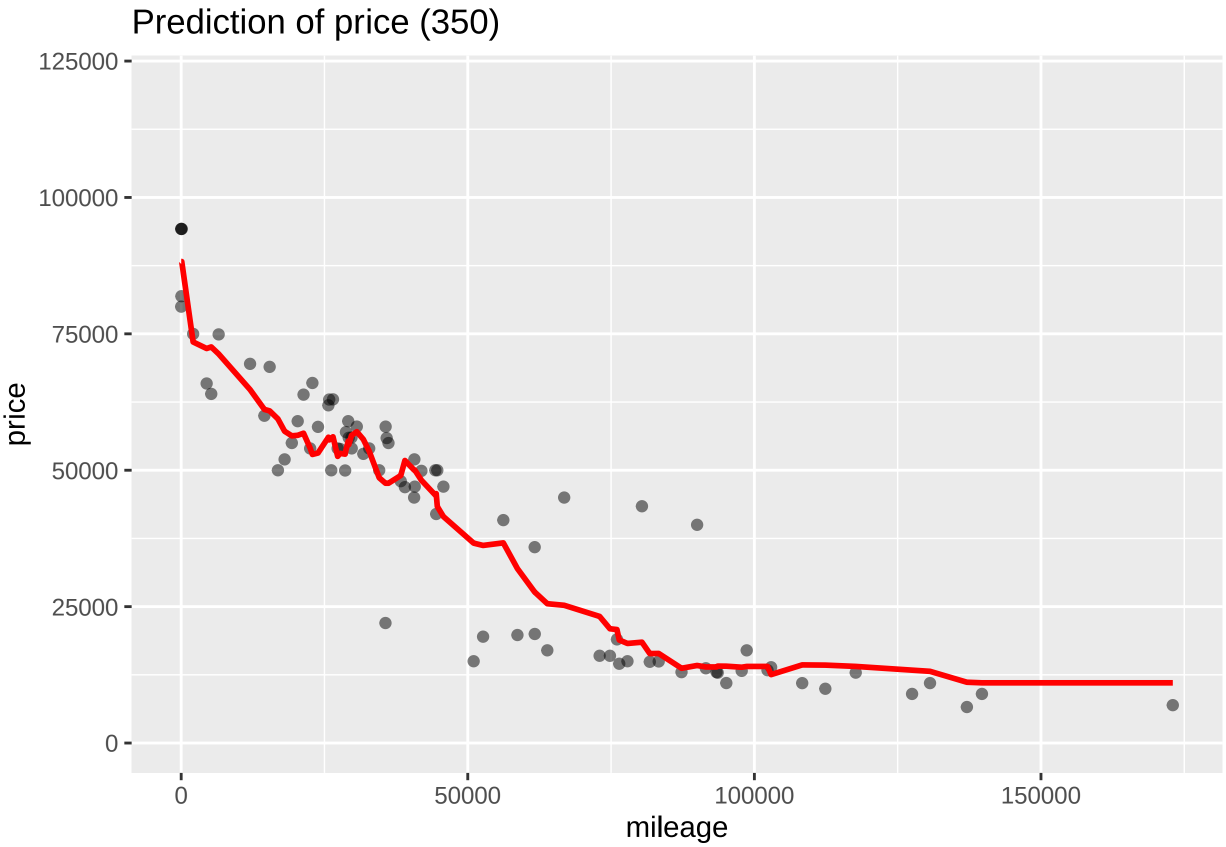
In the graph we see that there is a massive spike in average age in the 1910s and early 1920s. This could likely be a result of the young athletic men serving in World War 1 in the 1910s and not being able to compete in the Olympics. After about 1924 after woman’s swimming becomes a top 50 Olympic sport by participants both gender groups fall slightly until the 1950 which likely marks the end of the young athletes involvement in World War 2. Then from 1950, both men and woman gradually increase in age until the end of the data set. For The entire time frame of the data, the average men’s age is always at least as high as the women's average age. The men's average is strictly greater than the women's average for each of the Olympics except 2000. So other than this consistent higher average, both men and women's average age for simming follow abut the same trend.

Question 3: K-nearest neighbors: cars

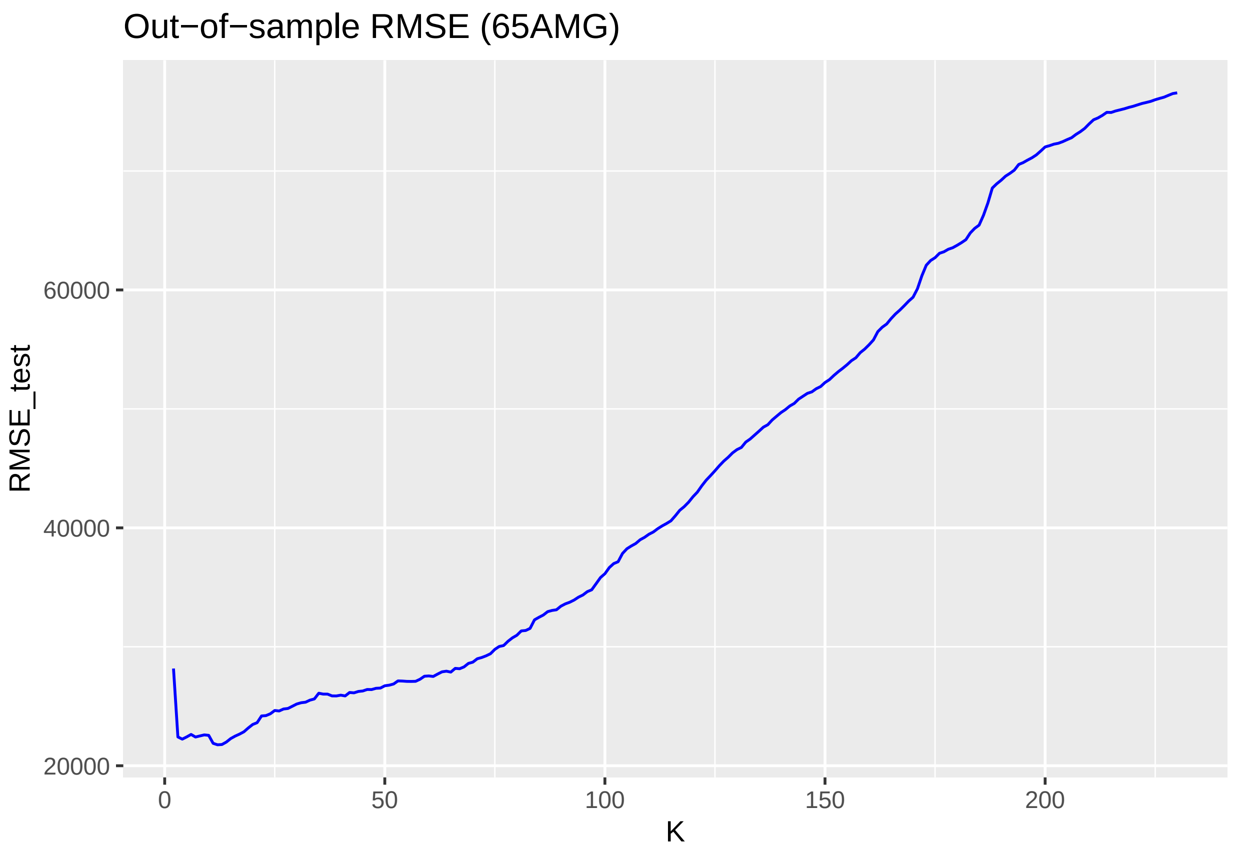


```
## [1] 14
```

```
## [1] 7801.907
```

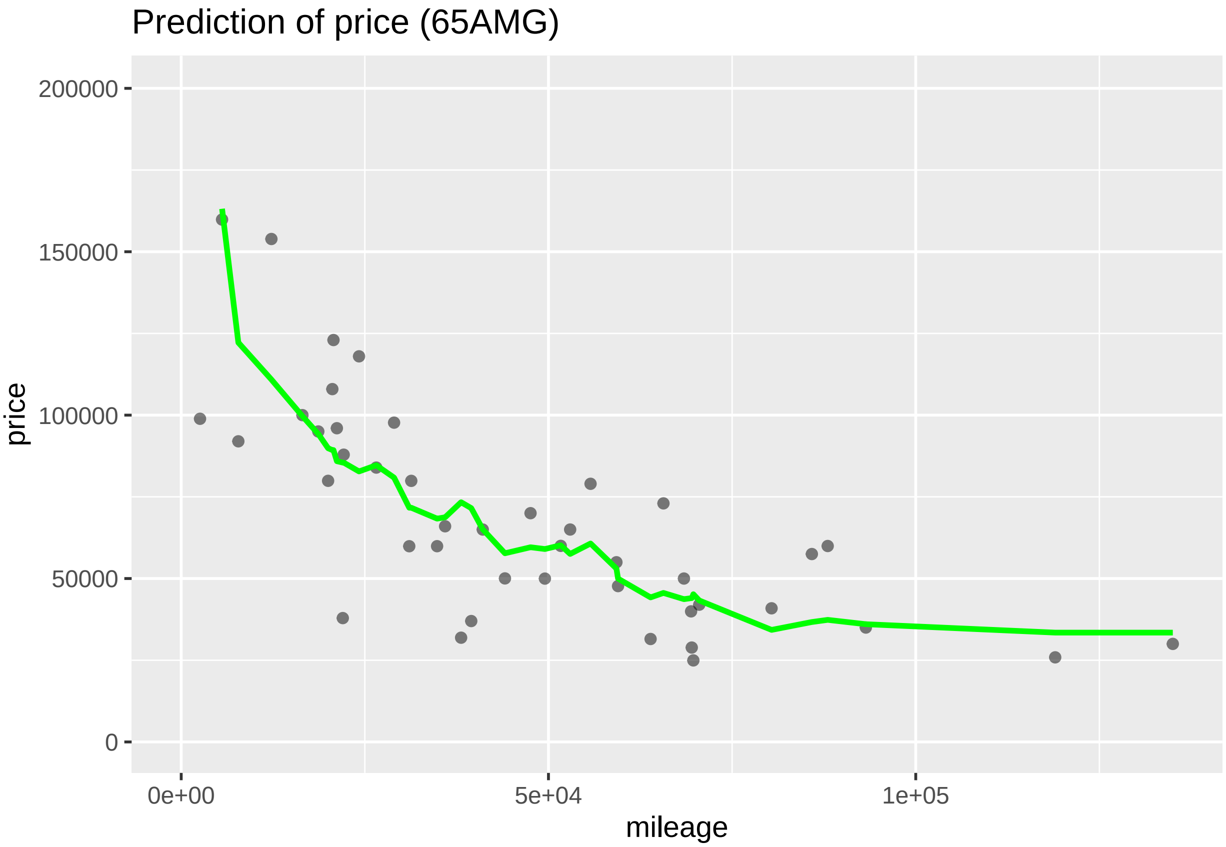


The First plot showcases the out-of-sample root mean-squared error (RMSE) for each value of K for the 350 AMG Trim. The first number after the plot if the optimal value of K where the RMSE bottoms out in the plot, and the second number is the Out-of-Sample RMSE that is achieved at the optimal k value. The second plot shows the predicted line for the optimal k-nearest neighbors in the out of sample data. The same process was conducted on the 65 AMG trim below.



```
## [1] 12
```

```
## [1] 21761.33
```



Although the optimal K may change each time the test is run depending on the train-test split, it appears that more times than not, the optimal value of K is higher for the 350 AMG trim. This is likely due to the fact that there is much more data points for the 350 AMG trim so taking more neighbors would not change the prediction as much also, the prices values are closer to each other in the 350 MAG where prices range from around 75,000 to about 20,000 dollars where the 65 AMG trim ranges from around 150,000 to about 25,000 dollars. So again, taking more data points in the k-nearest-neighbors for the 65 trim would likly increase variability of the predicted price.