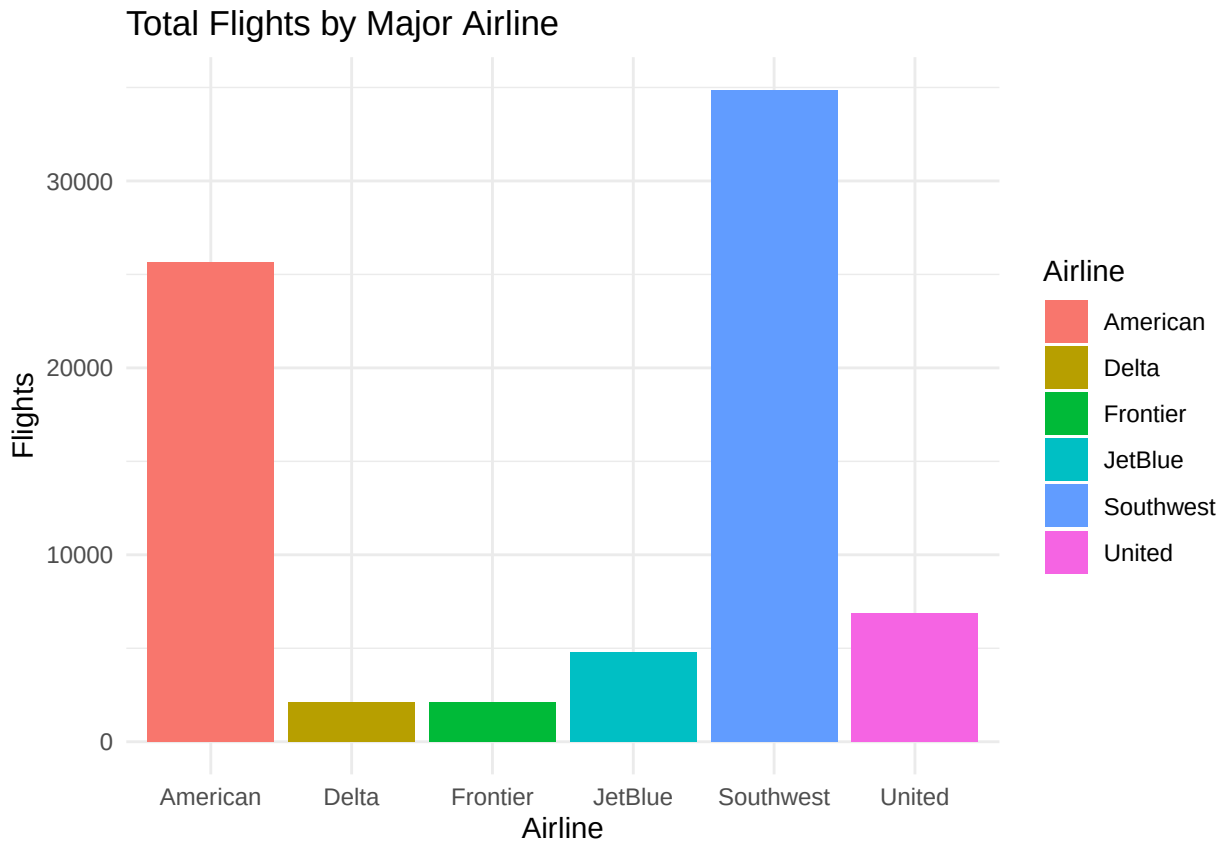
ECO 395M: StatLearning Exercise 1

Joseph Williams, Aahil Navroz, Suqian Qi

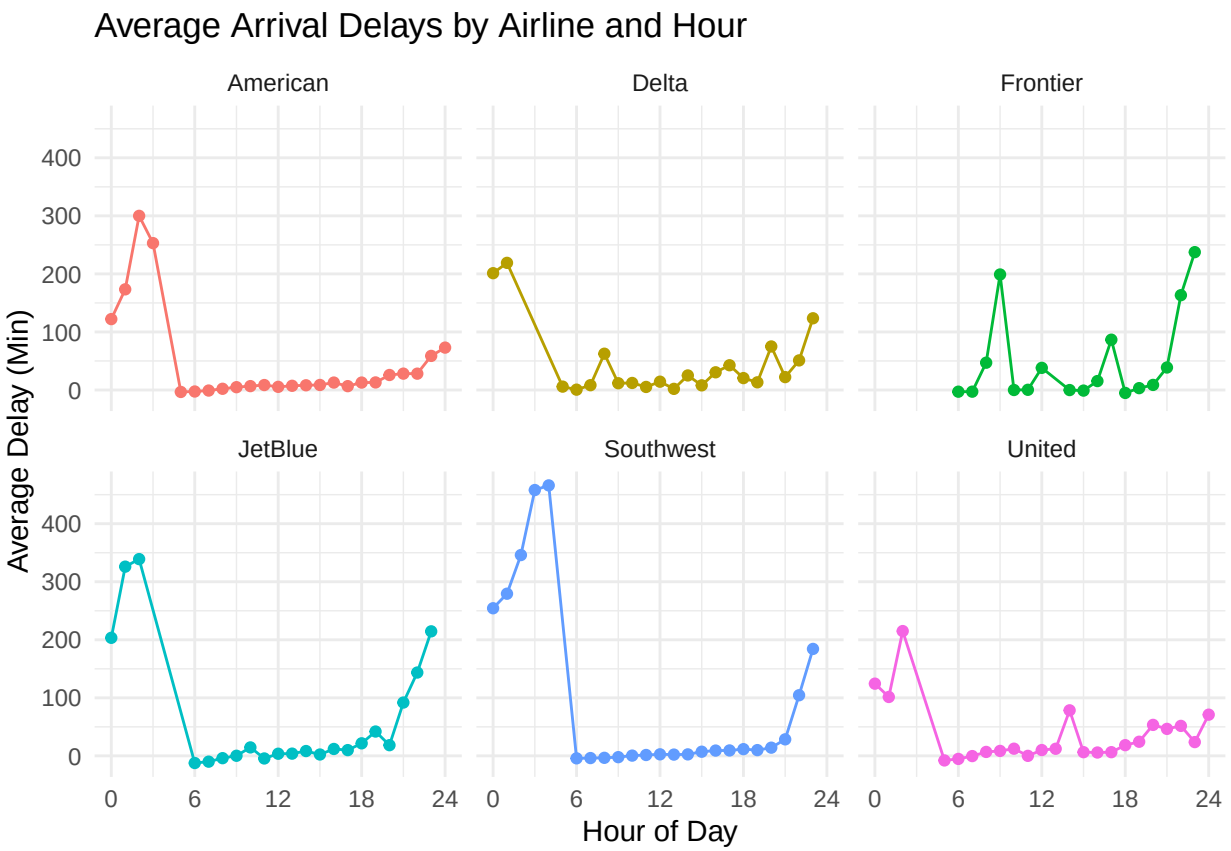
2024-02-04

1) Data visualization: flights at ABIA

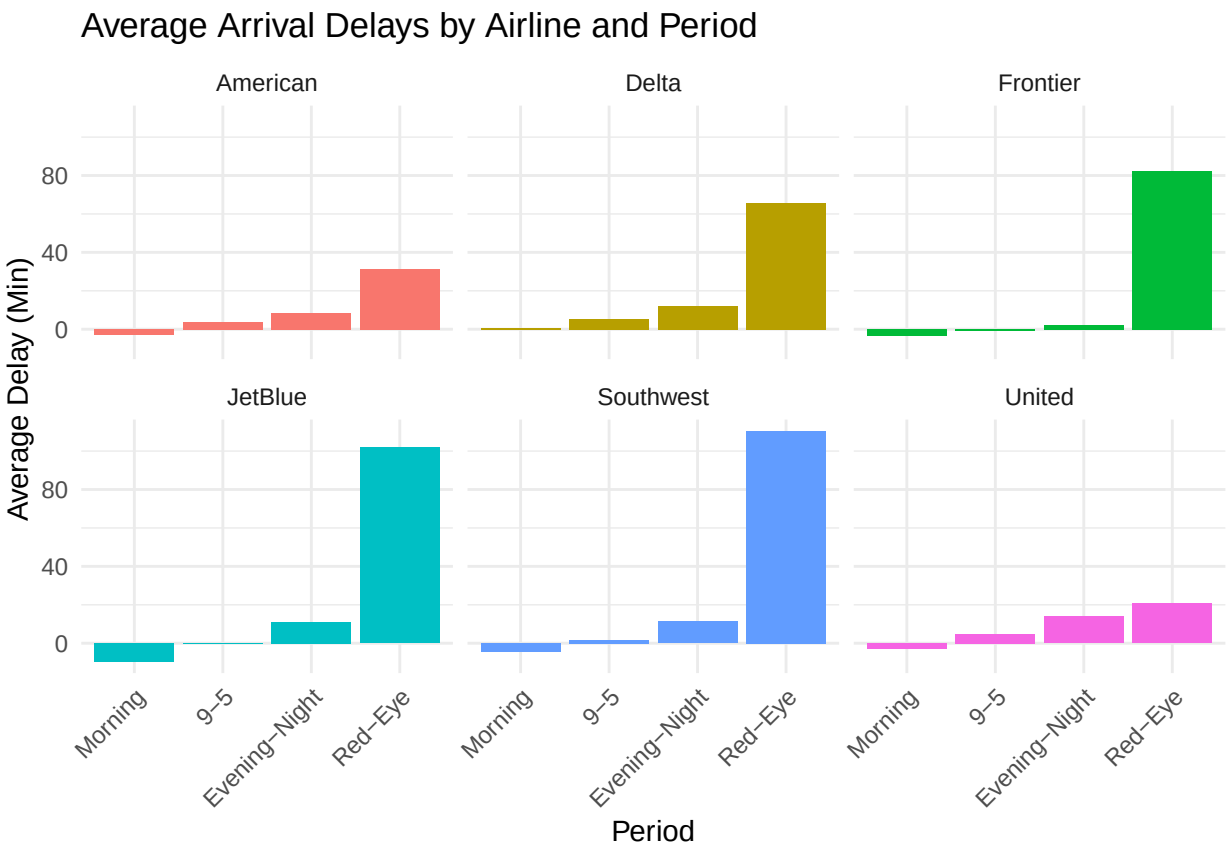
For this question we wanted to help flyers (say fliers in 2009) build intuition, or 'rules of thumb', that they can use when choosing between airline companies. To begin we want the data to correspond to recognizable names, so we mapped UniqueCarrier to airline brands or their parent brands. Here's a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



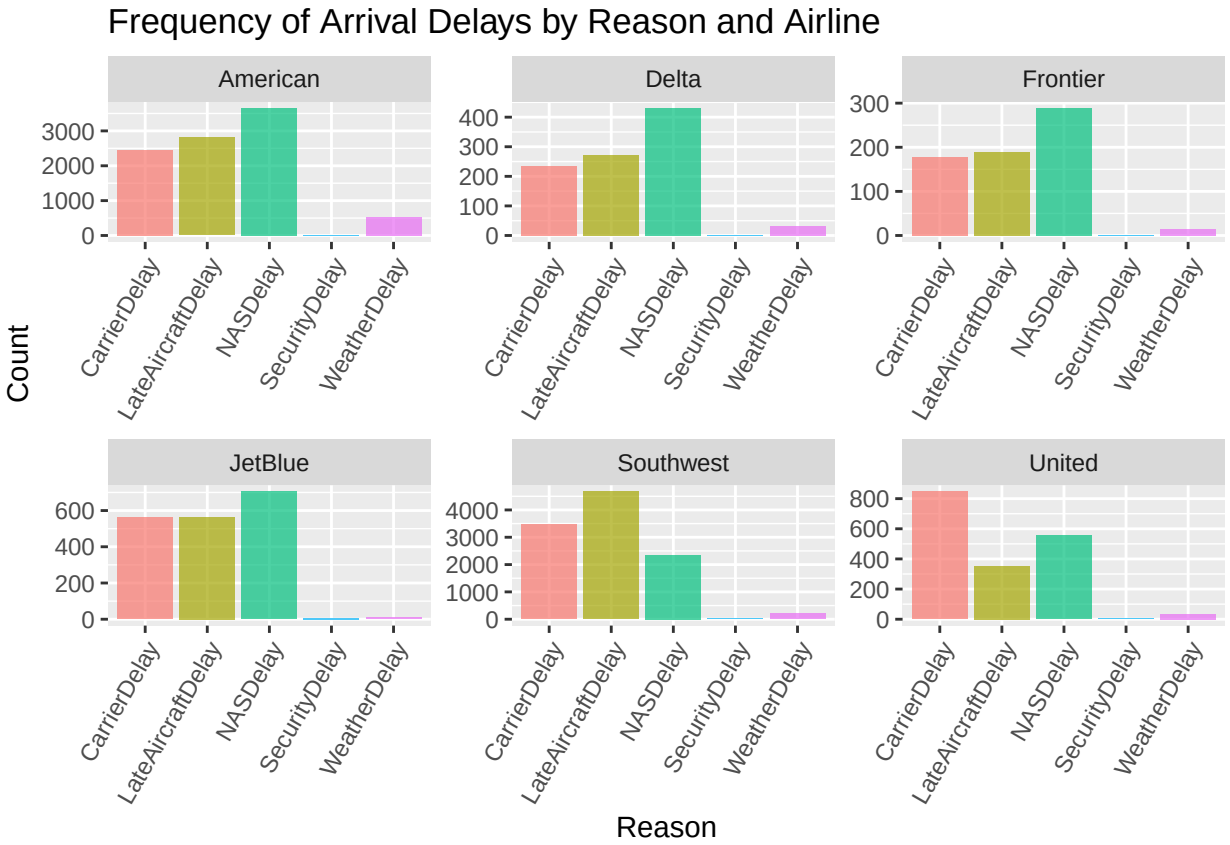
Looks like American and Southwest are king... but can they handle this type of volume. Lets look at arrival delays for each company. Given these are arrival delays for flights coming into and out of Austin, overall it will be a fine measure for the timeliness of the airline.



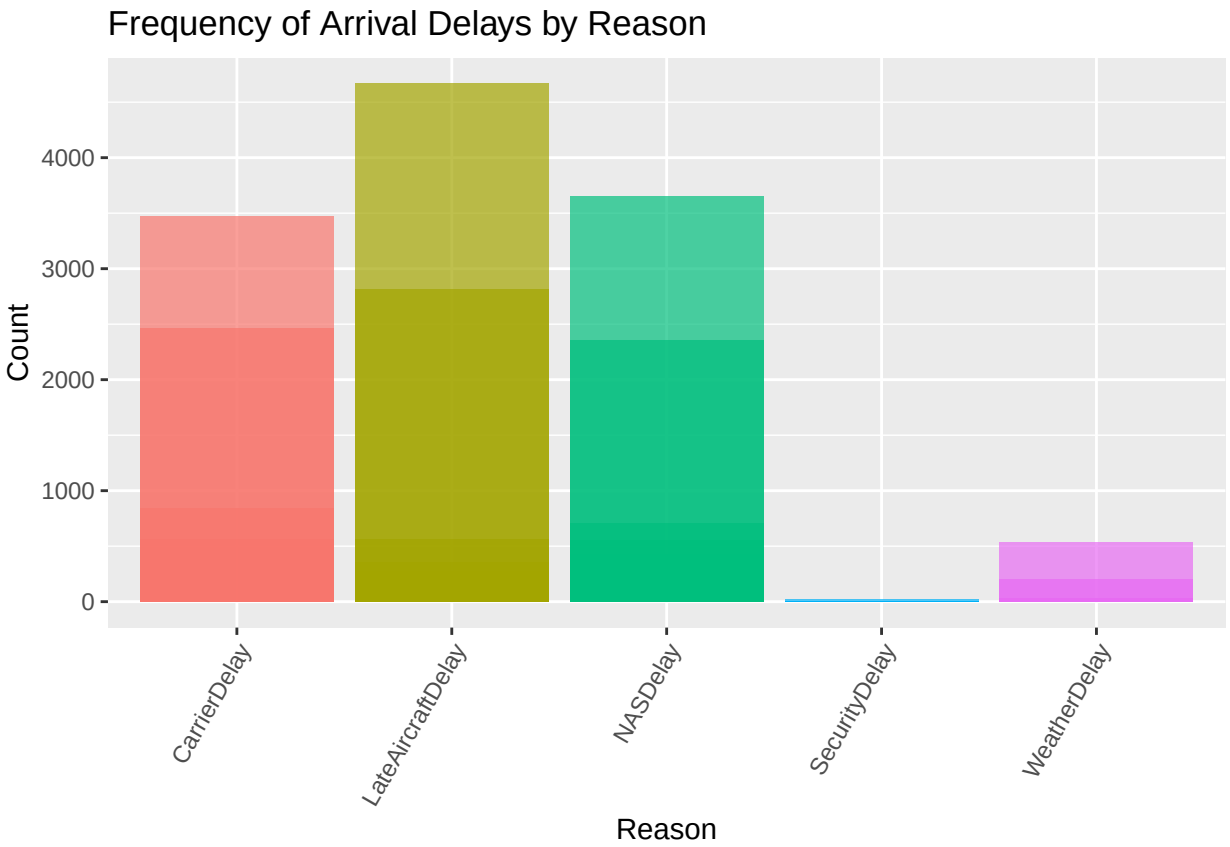
Okay, we're seeing some detail here. Seems like most companies experience their delays before the hour of 6am. Since we're looking for rules of thumb. Lets classify into 'Red Eye' 'Early Morning', '9-5' and 'Evening-Night', and see if we can quantify delay times over periods, rather than specific times.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



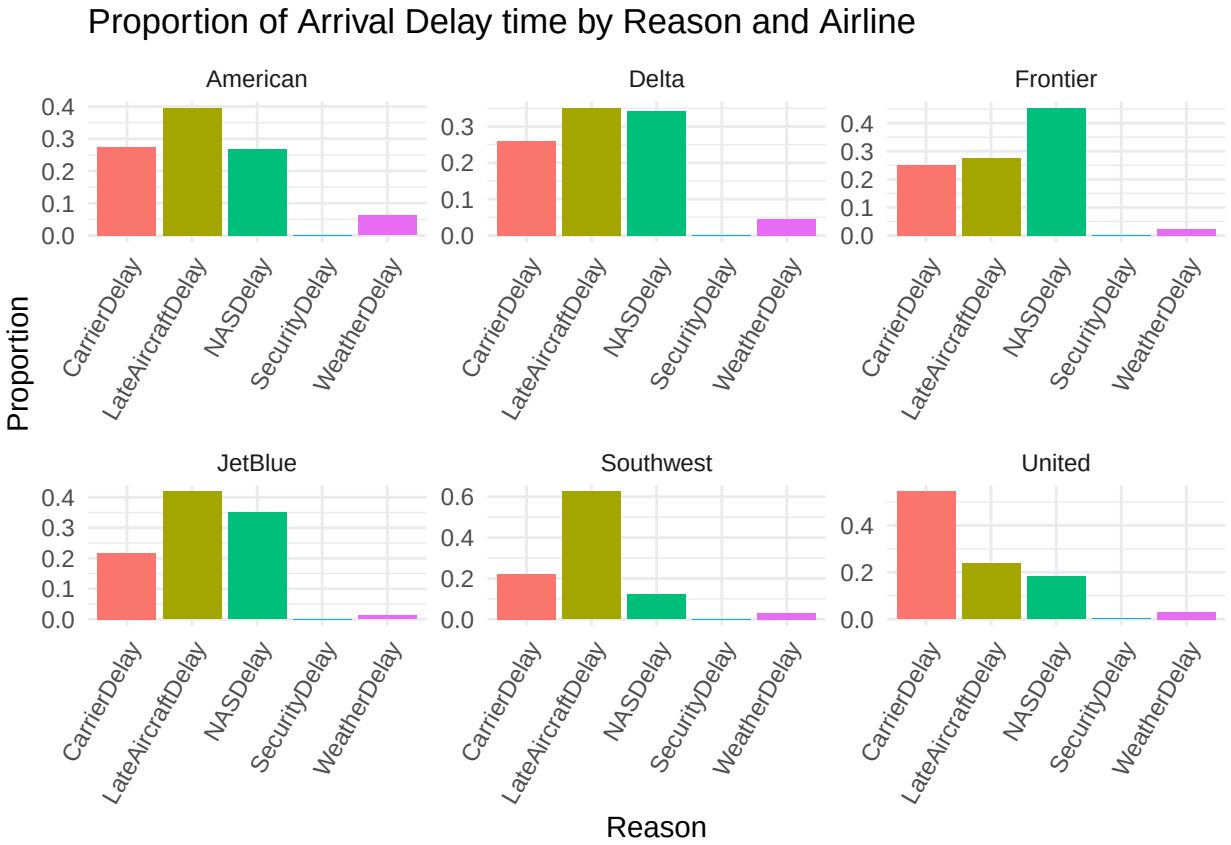
Okay now this is more useful! These average delay times seem too high, though. Lets check out the data and see if we can remove some outliers for more useful information.



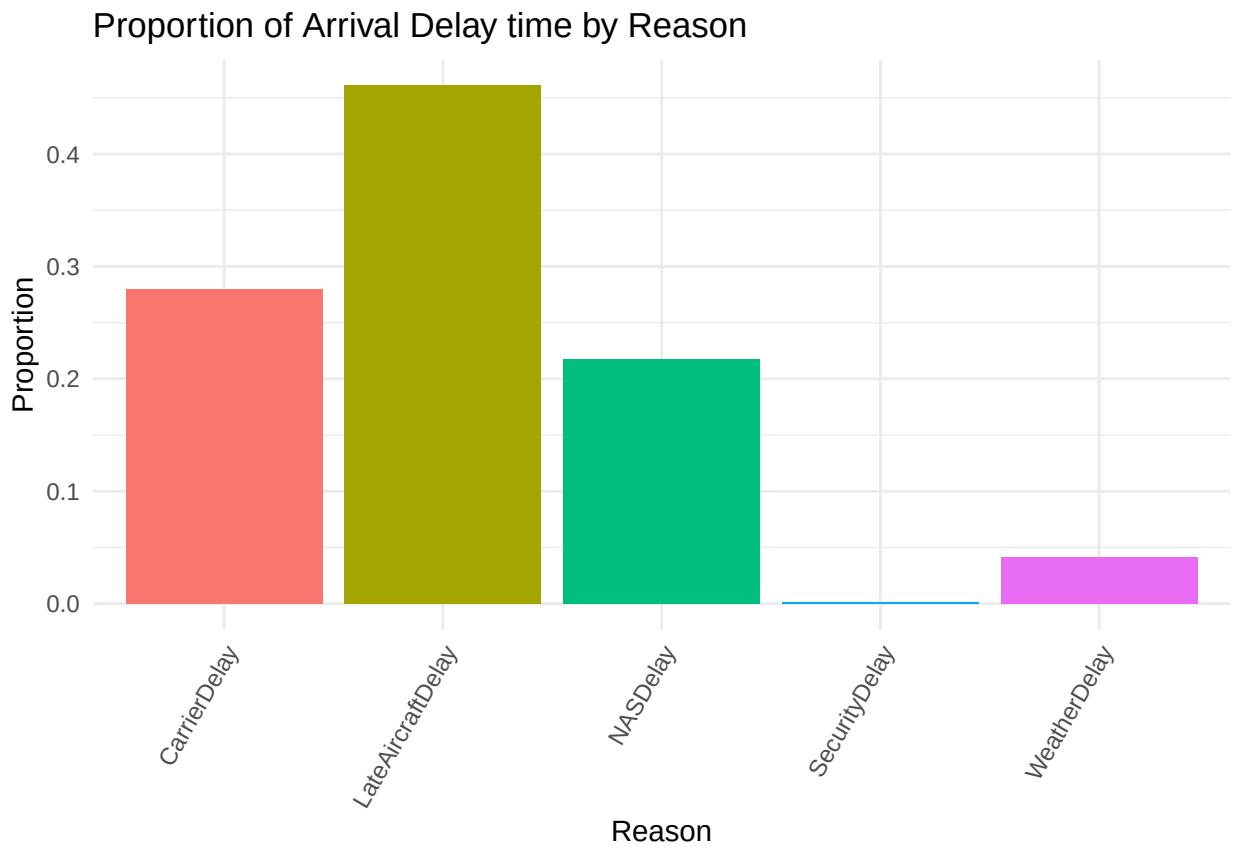
Through the plot we can see for American, Delta, Frontier and JetBlue, the most common reason for delay is NASDelay. For Southwest, the most common reason is LateAircraftDelay and for United it's CarrierDelay.



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.



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

The data in olympics_top20.csv contains information on every Olympic medalist in the top 20 sports by participant count, all the way back to 1896. Use these data to answer the following questions. (The names of the columns should be self-explanatory.)

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 `event` is the specific event (e.g. 100 meter sprint).

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?

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.

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

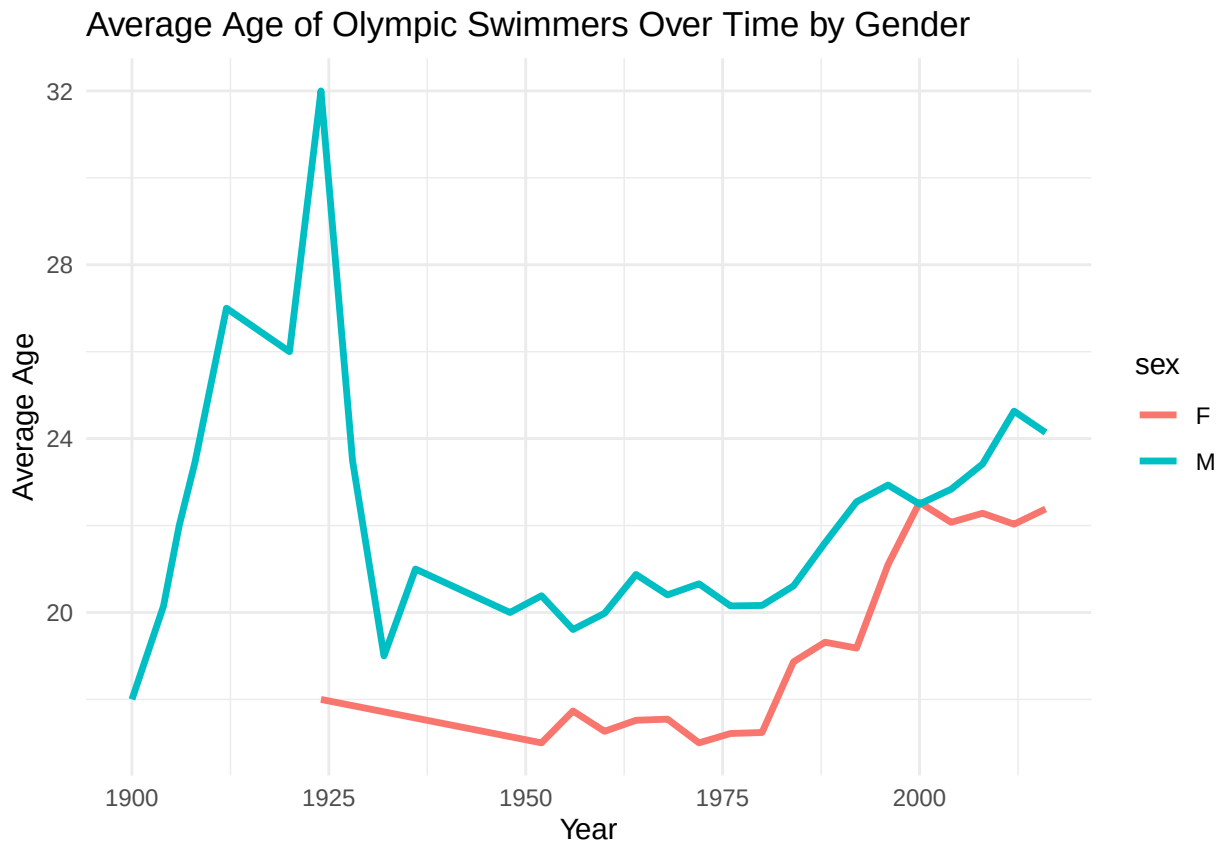
A) The 95th percentile of heights for female competitors across all Athletics events is 183.

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

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

B) 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) Generally, after the female participate in Olympics, the average age of swimmers has a upward trend over years. Before that, the male average age quickly reached a peak at 1924 and then fell down.



3) K-nearest neighbors: cars

The data in sclass.csv contains data on over 29,000 Mercedes S Class vehicles—essentially every such car in this class that was advertised on the secondary automobile market during 2014. For websites like Cars.com or Truecar that aim to provide market-based pricing information to consumers, the Mercedes S class is a notoriously difficult case. There is a huge range of sub-models that are all labeled "S Class,"" from large luxury sedans to high-performance sports cars; one sub-category of S class has even served as the safety car in Formula 1 Races. Moreover, individual submodels involve cars with many different features. This extreme diversity—unusual for a single model of car—makes it difficult to provide accurate pricing predictions to consumers.

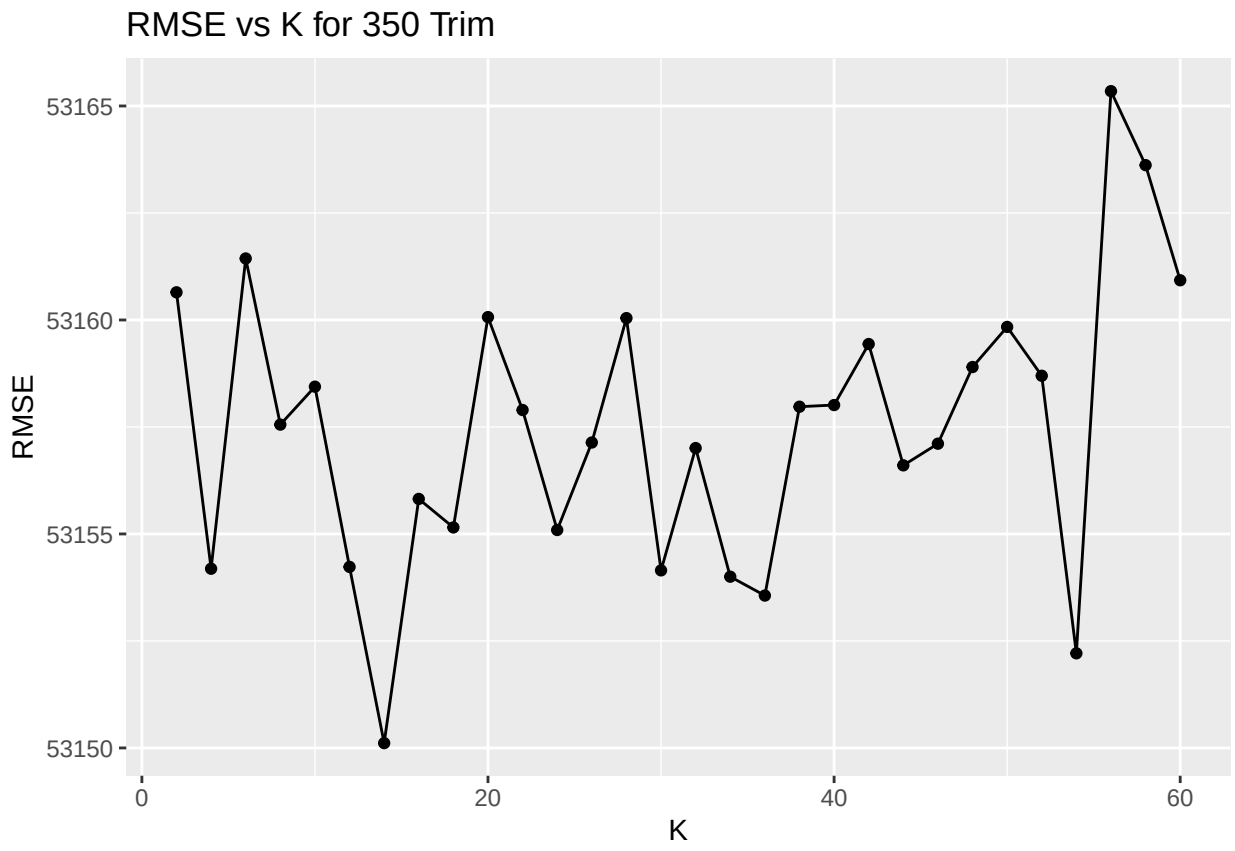
We'll revisit this data set later in the semester when we've got a larger toolkit for building predictive models. For now, let's focus on three variables in particular: - trim: categorical variable for car's trim level, e.g. 350, 63 AMG, etc. The trim is like a sub-model designation.
- mileage: mileage on the car - price: the sales price in dollars of the car

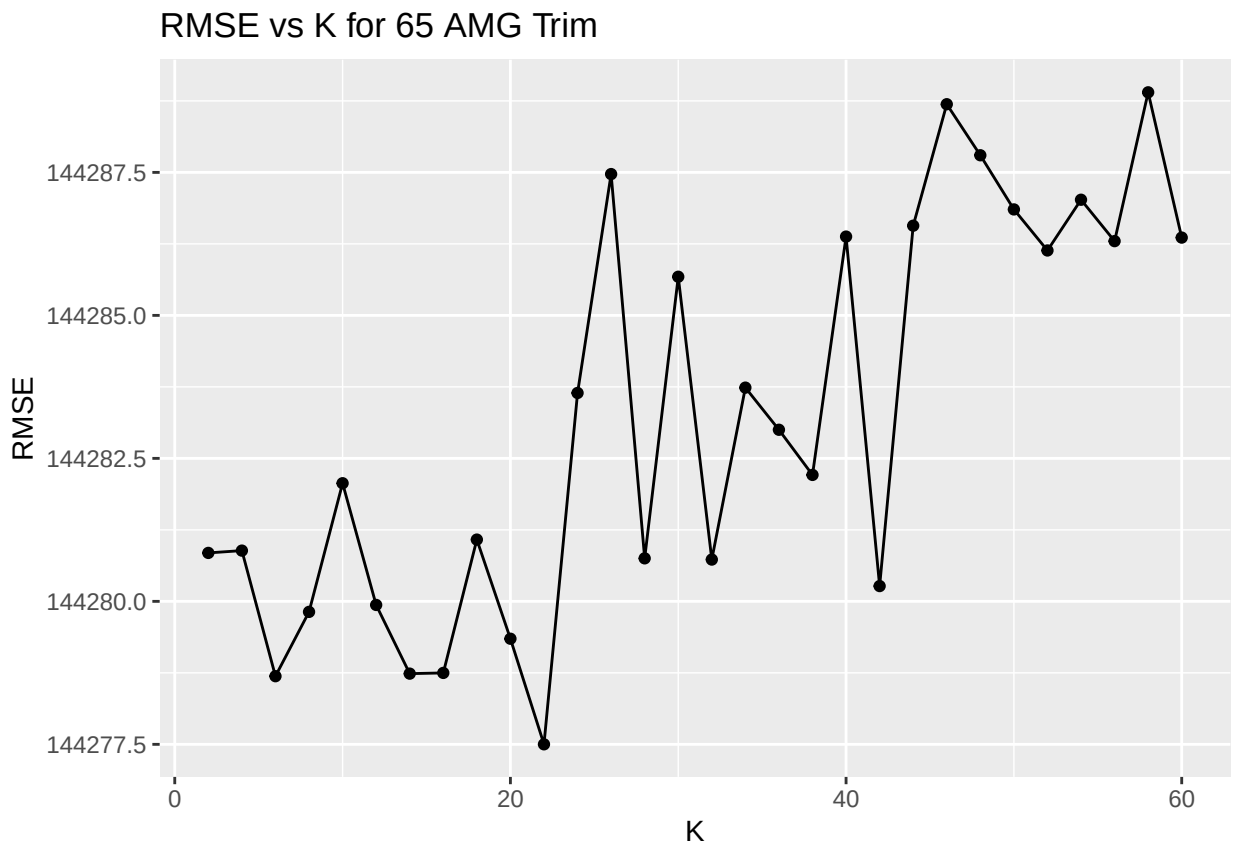
Your goal is to use K-nearest neighbors to build a predictive model for price, given mileage, separately for each of two trim levels: 350 and 65 AMG. (There are lots of other trim levels that you'll be ignoring for this question.) That is, you'll be treating the 350's and the 65 AMG's as two separate data sets. (Recall the `filter` command.)

For each of these two trim levels: 1) Split the data into a training and a testing set.
2) Run K-nearest-neighbors, for many different values of K, starting at K=2 and going as high as you need to. For each value of K, fit the model to the training set and make predictions on your test set. 3) Calculate the out-of-sample root mean-squared error (RMSE) for each value of K.

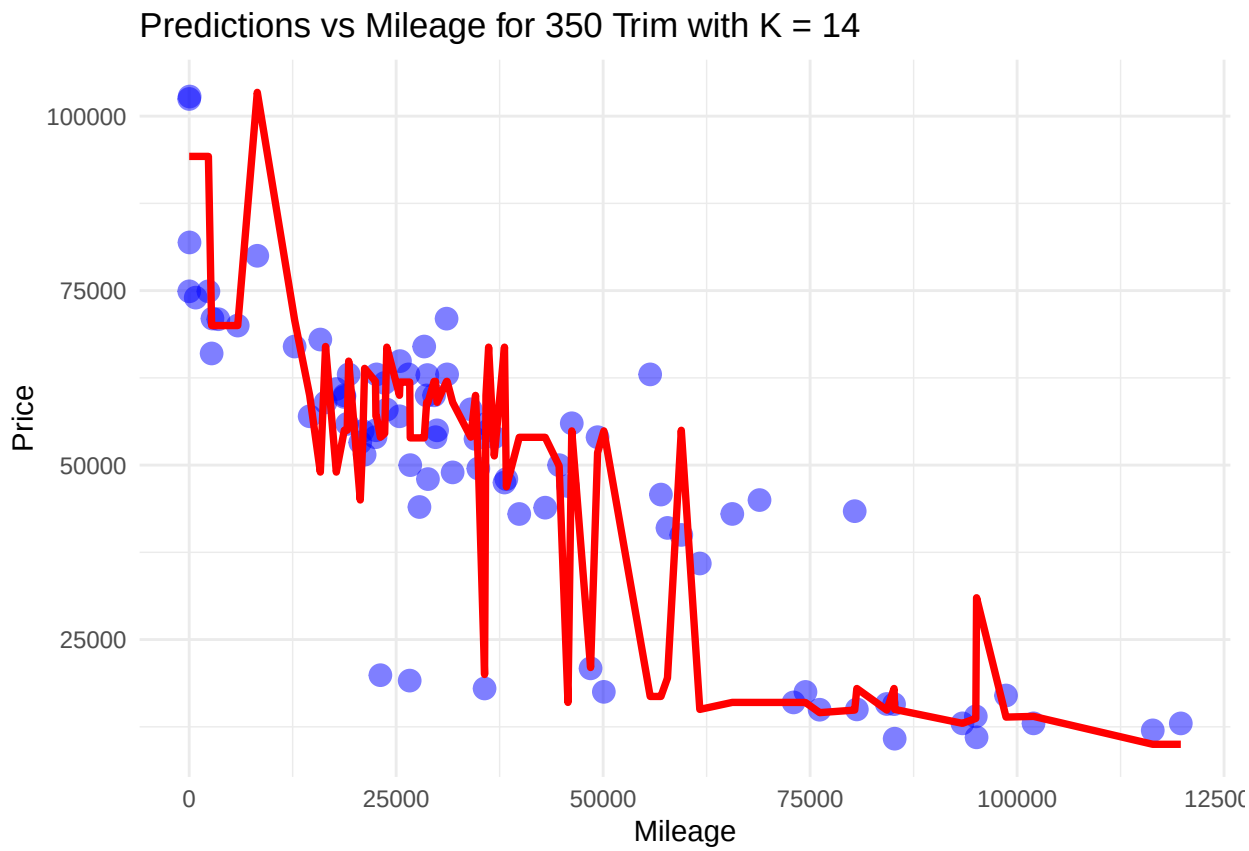
For each trim, make a plot of RMSE versus K, so that we can see where it bottoms out. Then for the optimal value of K, show a plot of the fitted model, i.e. predictions vs. x. (Again, separately for each of the two trim levels.)

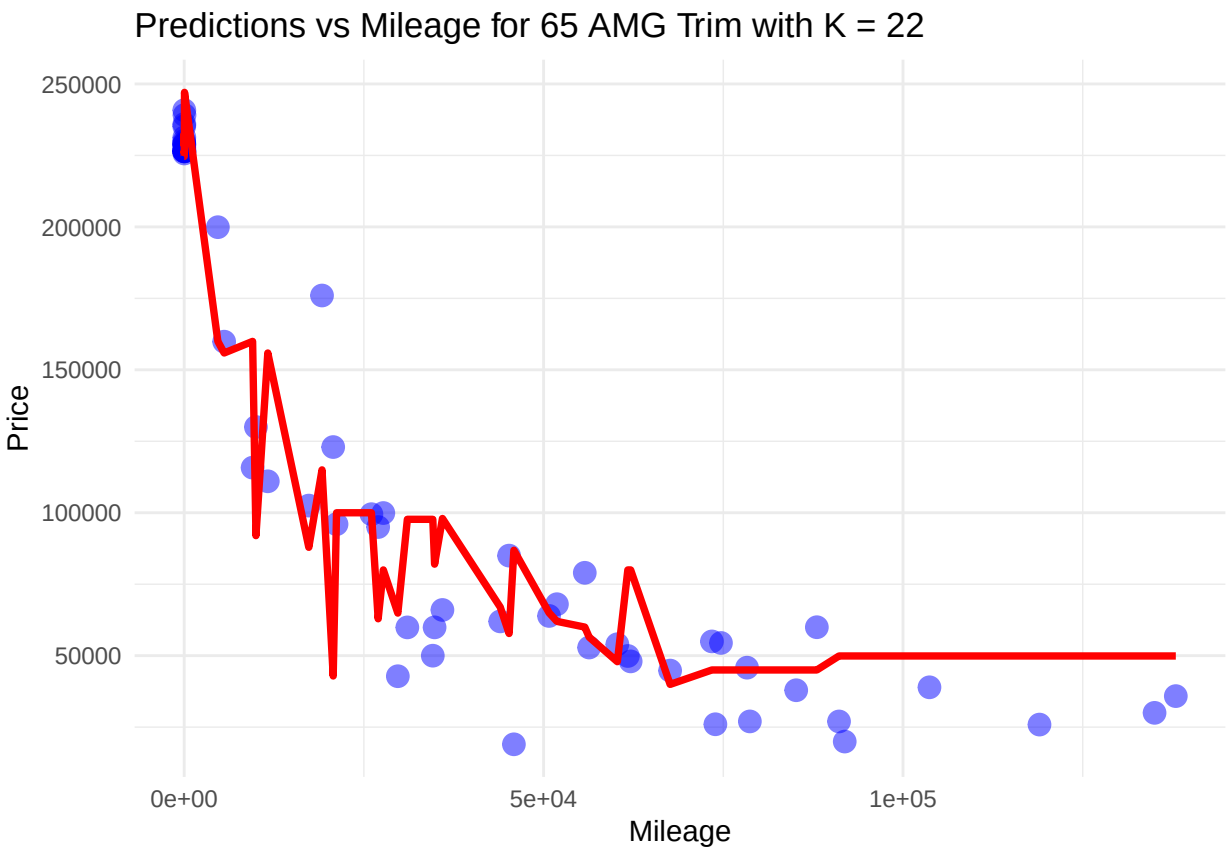
Which trim yields a larger optimal value of K? Why do you think this is?





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 22.


Predictions vs Mileage for 350 Trim with K = 14
Price
100000
75000
50000
25000
0
25000
50000
75000
100000
12500
Mileage


Predictions vs Mileage for 65 AMG Trim with K = 22
250000
200000
150000
100000
50000
Price
0e+00
5e+04
1e+05
Mileage