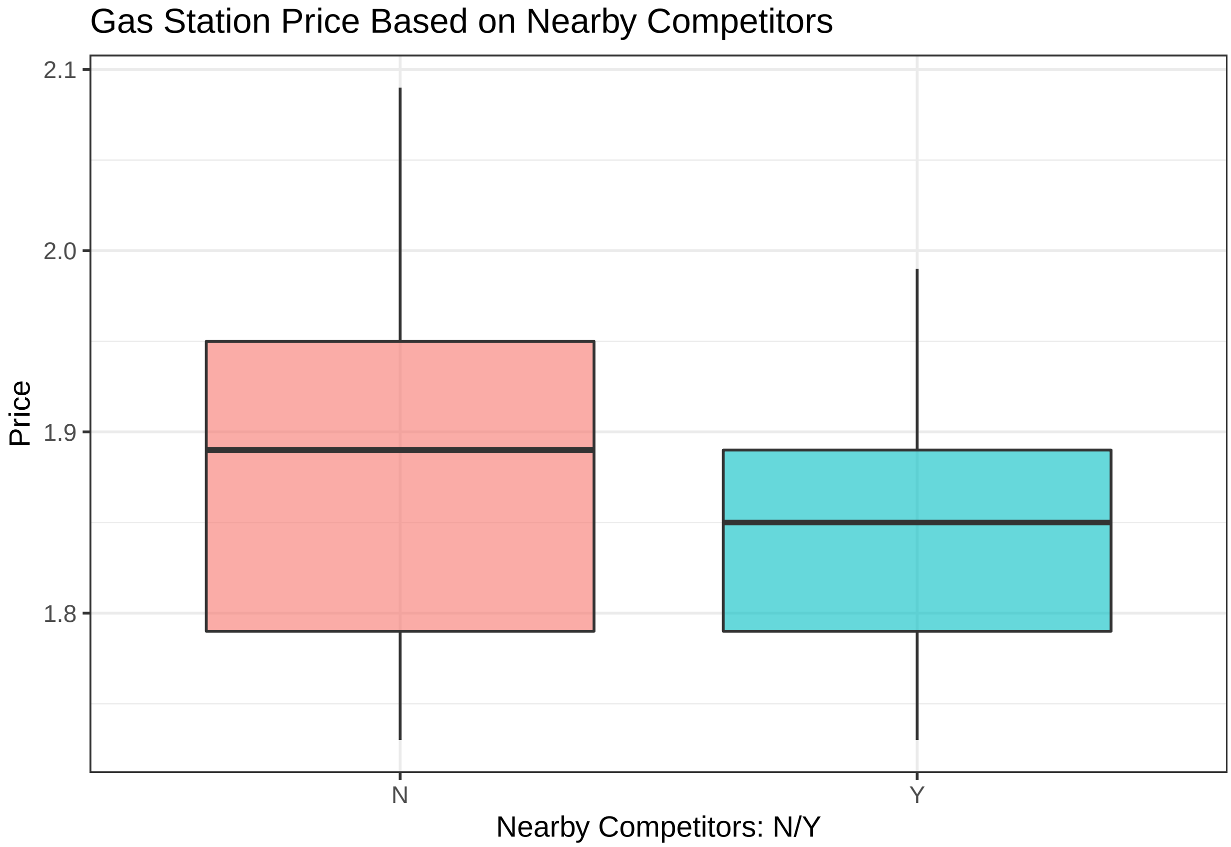
Exercises #1 by Gaetano Dona-Jehan & Jordan Fox

Question 1) Data visualization: Austin-area Gas Prices

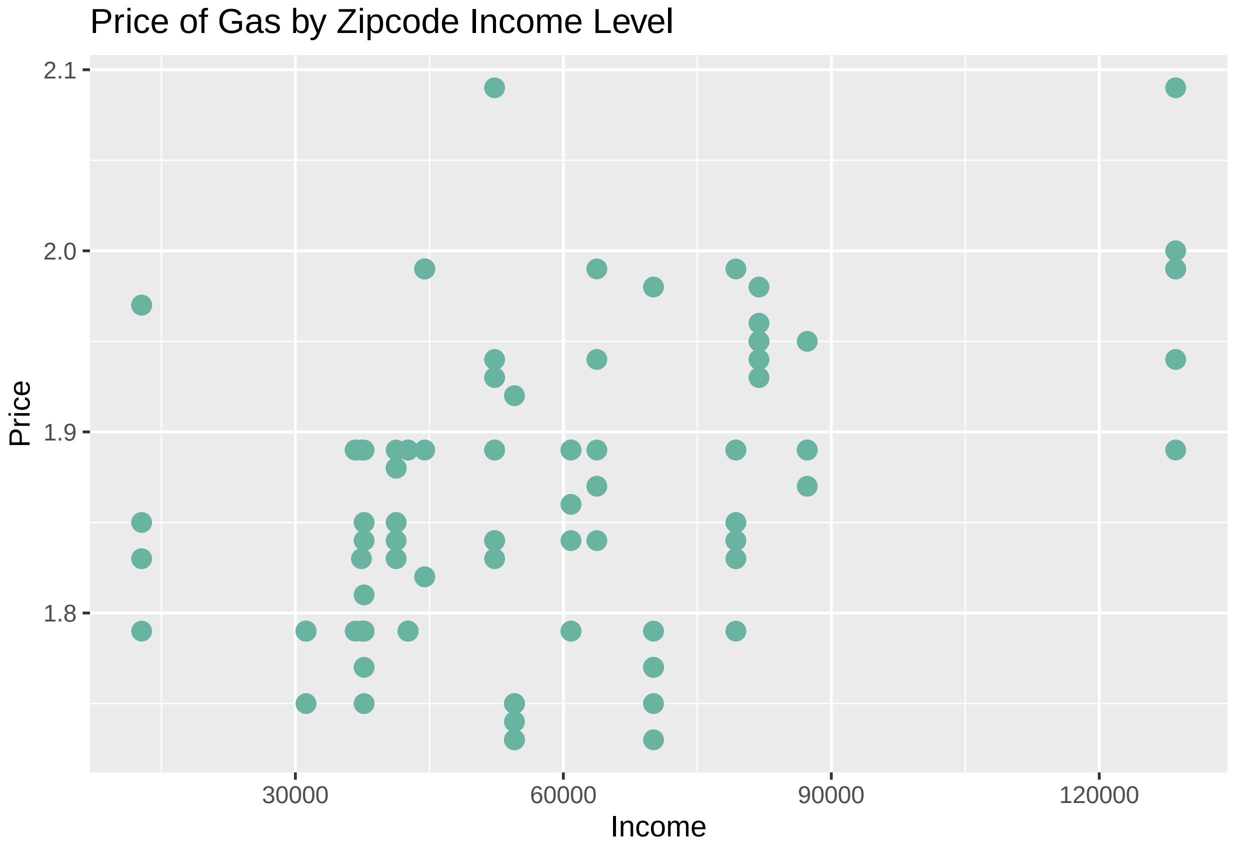
The first problem contains a number of data visualization exercises with data on gas stations in the greater Austin area. Data contained in the data set include the price offered at the station, the brand of the seller, and a number of categorical variables which capture the various characteristics of each station. These will be used to assess several theories on how these characteristics impact the price offered at the pump.

Part A: Do Gas Stations Charge More if They Lack Competition?



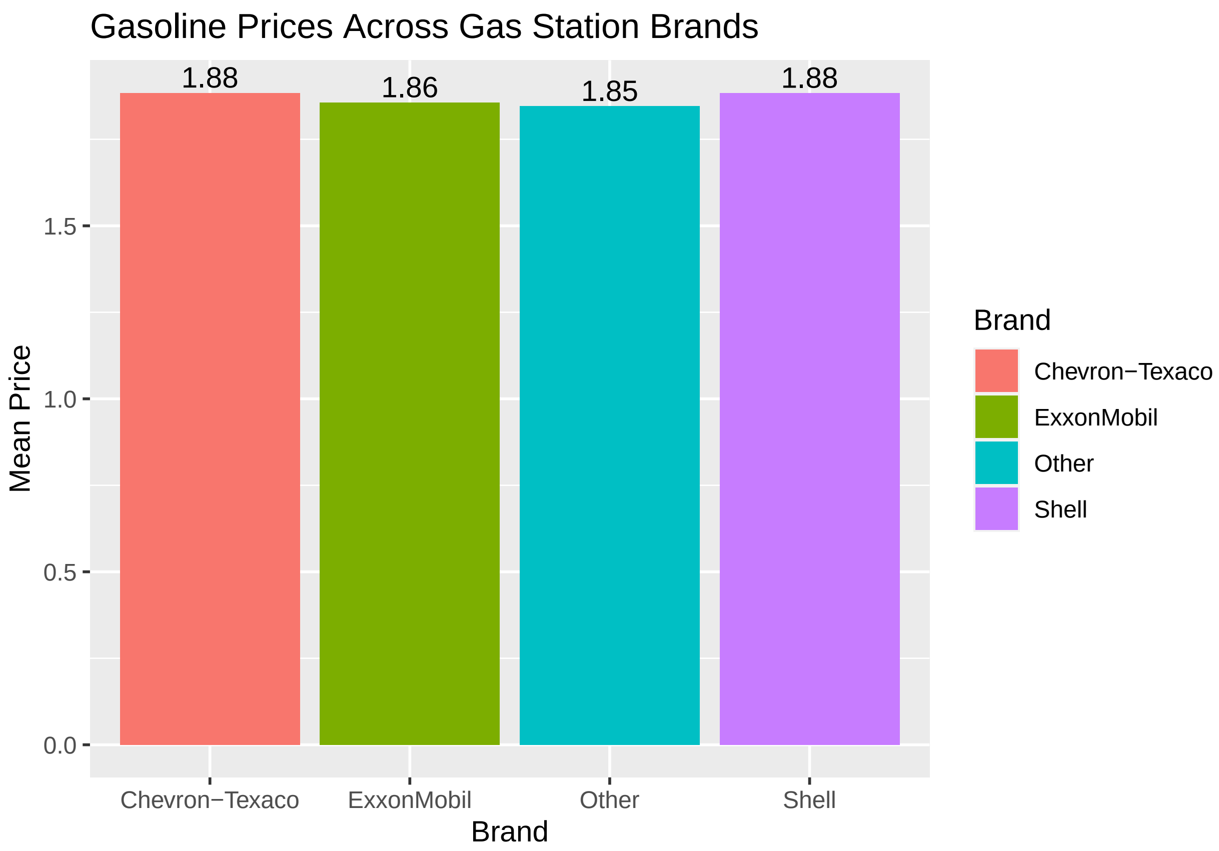
Based the boxplots, we can see that there is indeed evidence supporting the claim that a lack of direct competition leads to gas stations charging higher prices. The black bar in the middle of box plots represents the median price, which is higher for gas stations that do not have competitors in sight. Those that do not have a nearby competitor charge about $1.88, whereas those with one tend to charge about $1.85. Furthermore, the interquantile range for the gas stations lacking competition is greater than that of those with competition, suggesting that there is a greater variation in prices for the competitive gas stations. What is interesting to note is that both competing and non-competing gas stations have the same range of prices in the 1st quartile.

Part B: Do Richer Areas Have Higher Gas Prices?



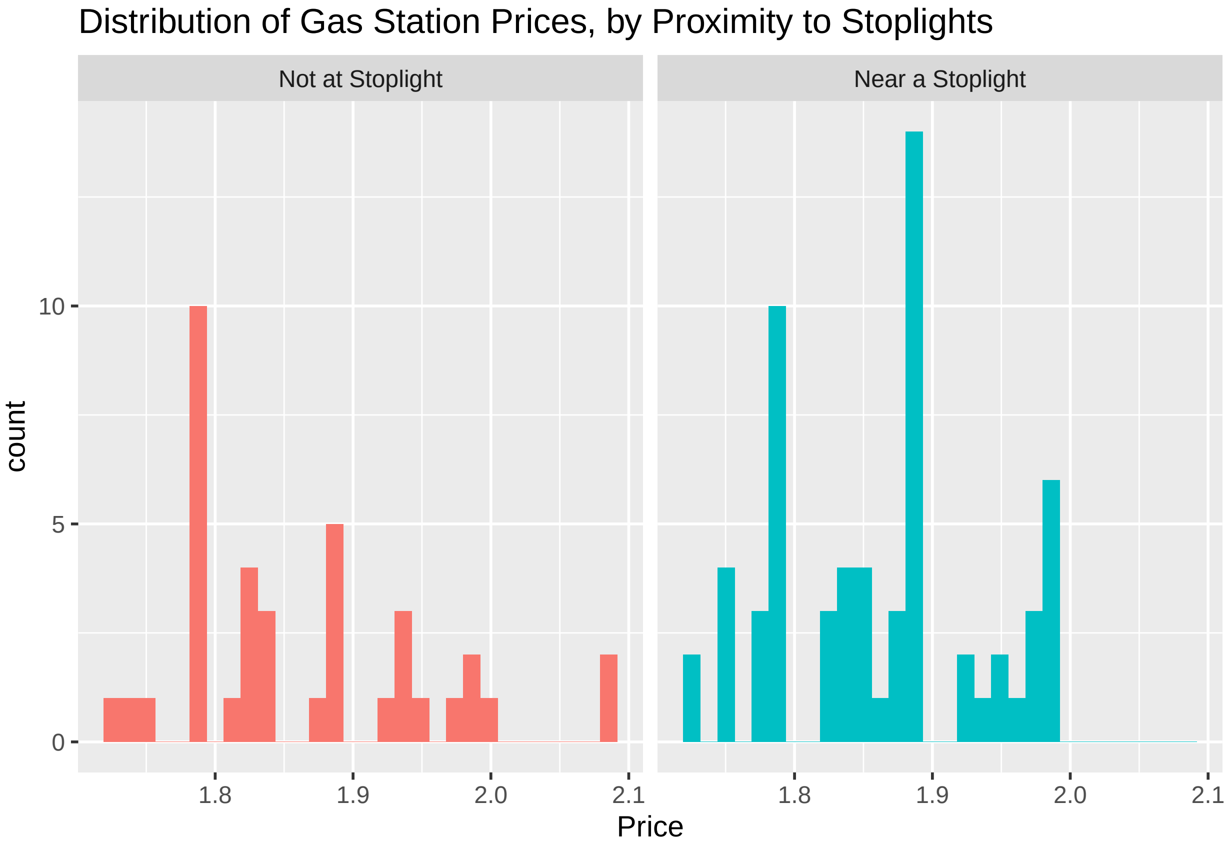
The above scatterplot does seems to suggest a positive correlation between the median household income of the zipcode of the station and the price of gasoline. However, it also seems to suggest that there is an increase in the price of gasoline in lower income areas as can be seen at the $10000 - $15 000 range. This could be due to two things. The first is that these might simply be outliers. The second, and the most probable reason, is that these gas stations have higher prices because the income elasticity for the citizens that live in these areas are significantly lower. This could be due to the fact that these citizens are more likely to wait until the “empty fuel tank” signal turns on before getting gas due to budgetary constraints, and thus preventing them from traveling to a cheaper gas station. However, this is simply a conjecture and would need to be studied further.

Part C: Does Shell Charge More Than Other Brands?



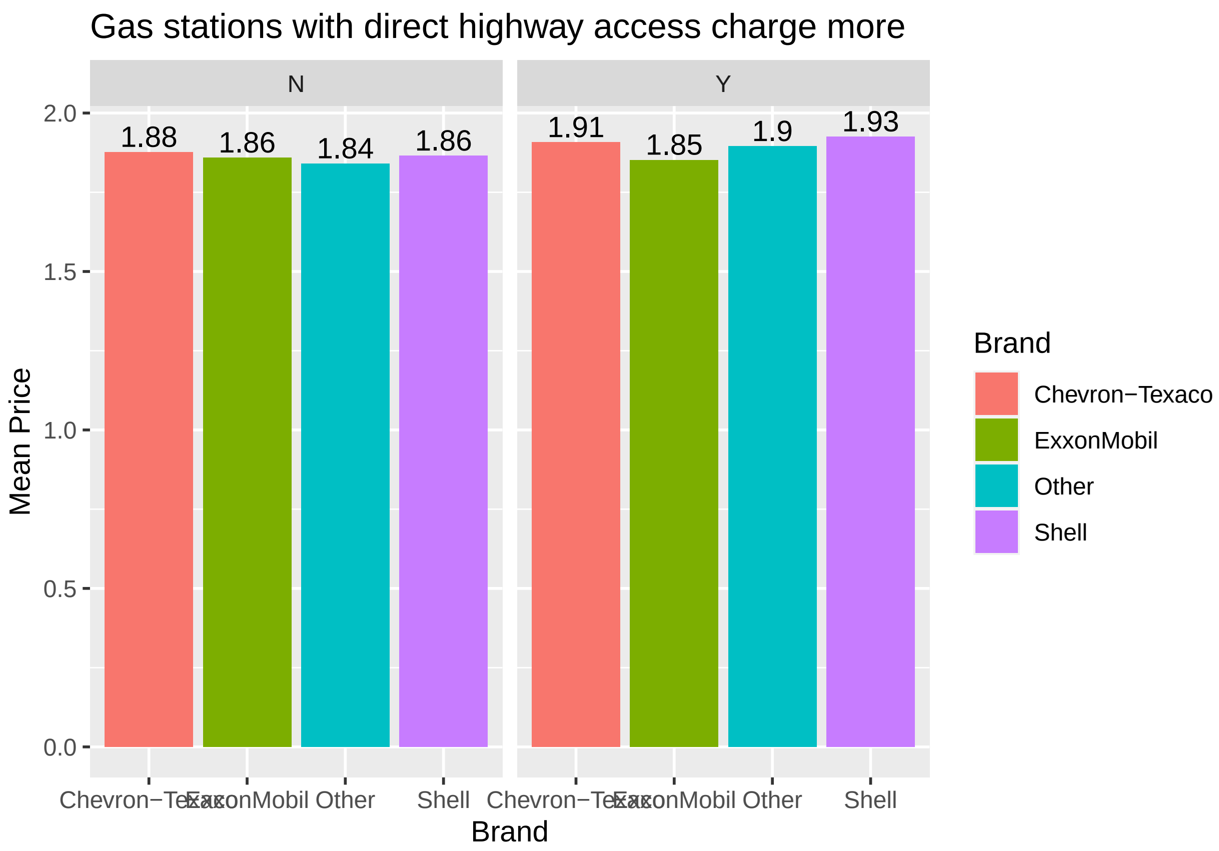
It would seem like Shell does not charge more when we compare the average gas price between brands. As we can see, the average price of gasoline for Chevron-Texaco is the same as Shell. ExxonMobil and "other" charge only a few cents less than Shell on average.

Part D: Do Gas Stations at Stoplights Charge More?



There is modest evidence that gas stations located at stoplights do charge higher prices than those not at stoplights. First, the distribution of prices for non-stoplight stations has a slight rightward skew, with the modal price of just under $1.80 per gallon. The distribution for stoplight stations is more normally distributed and centered around just below $1.90. It also has a greater concentration of prices in the range of $1.85- $2.00. While there are a few observations in the non-stoplight distribution that are above the highest price found in the stoplight distribution, this is not enough evidence that non-stoplight stations charge more than their counterparts.

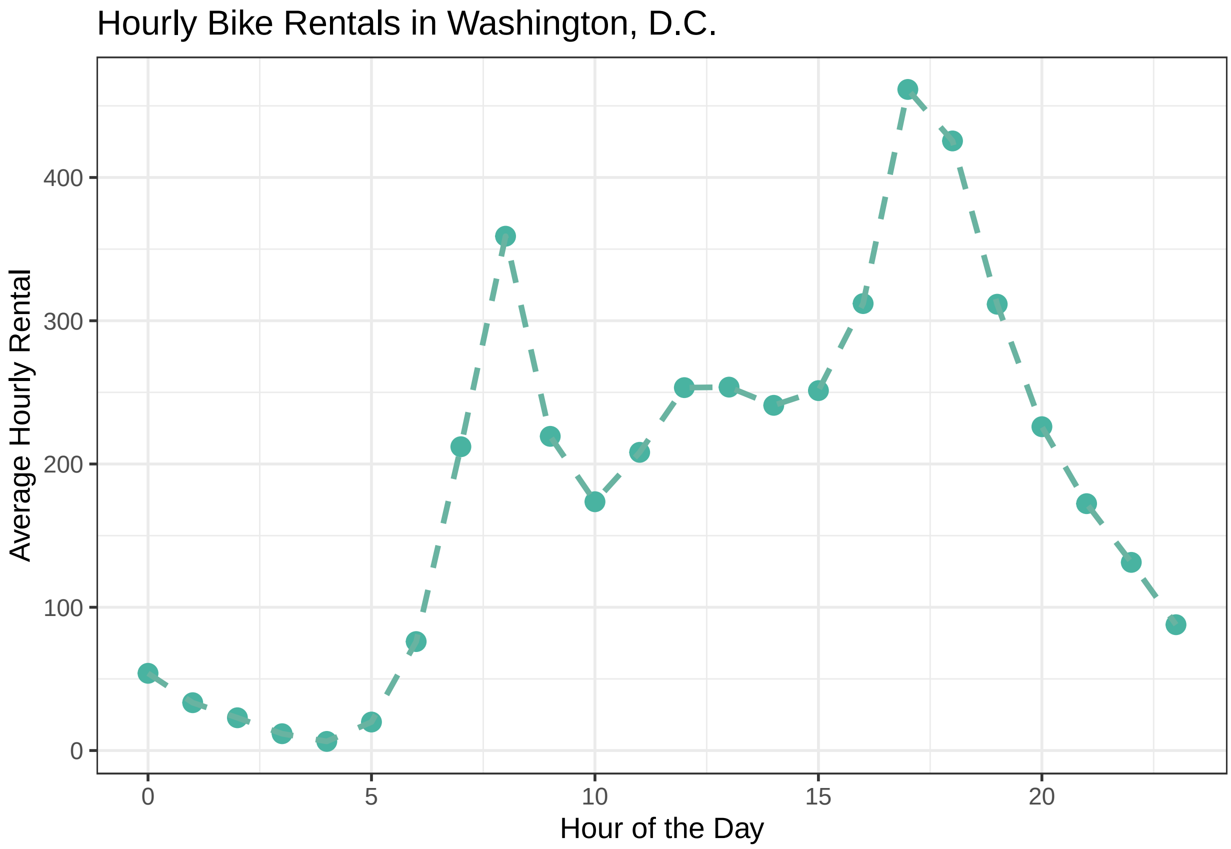
Part E: Do Gas Stations with Direct Highway Access Charge More?



The evidence supports the claim that gas stations with direct highway access charge more, with the exception of ExxonMobil. As we can see from the graphs, Chevron,the other gas stations and Shell have increased their average prices by $0.03, $0.05 and $0.07 respectively. This could possibly be evidence of collusion, or alternatively, price wars between gas stations, which tend to cluster together.

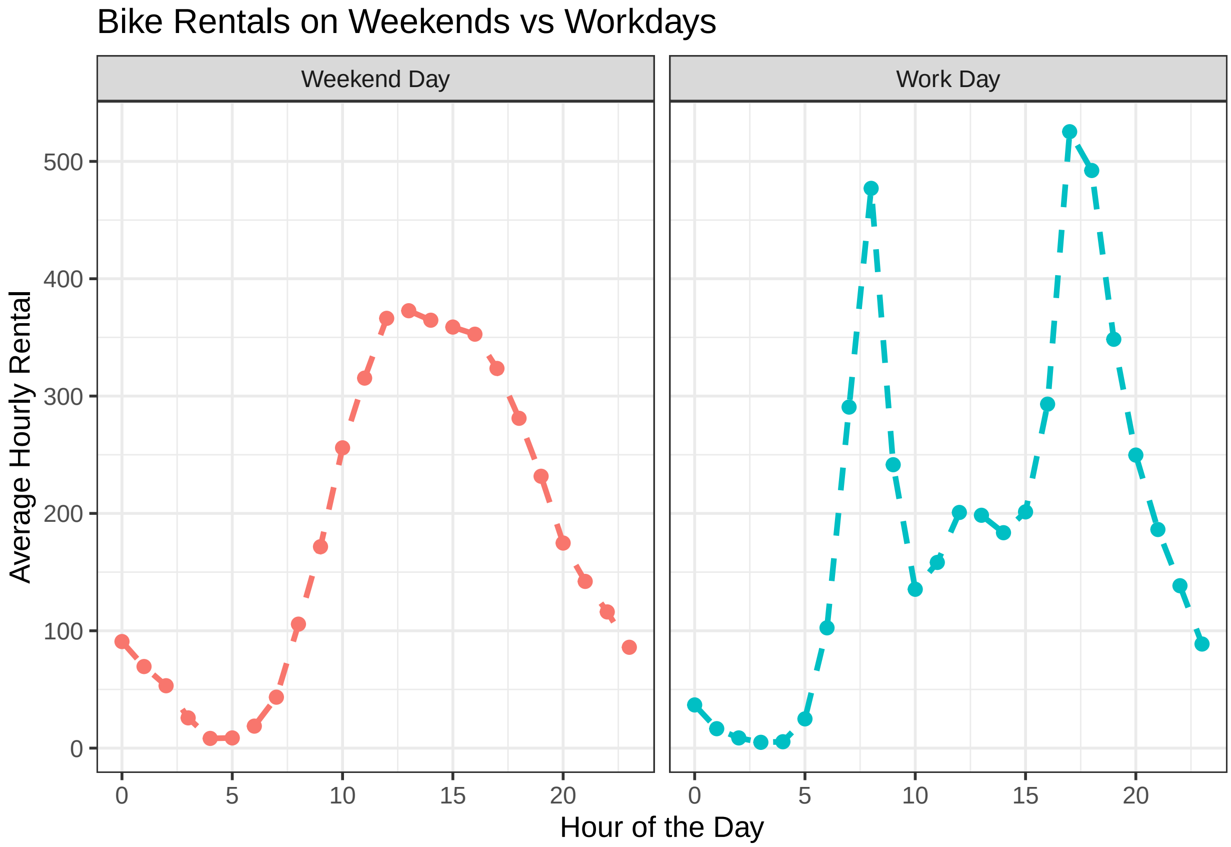
Question #2: Bike Rental Data from Washington, DC

Average Bike Rentals Thoughout the Day



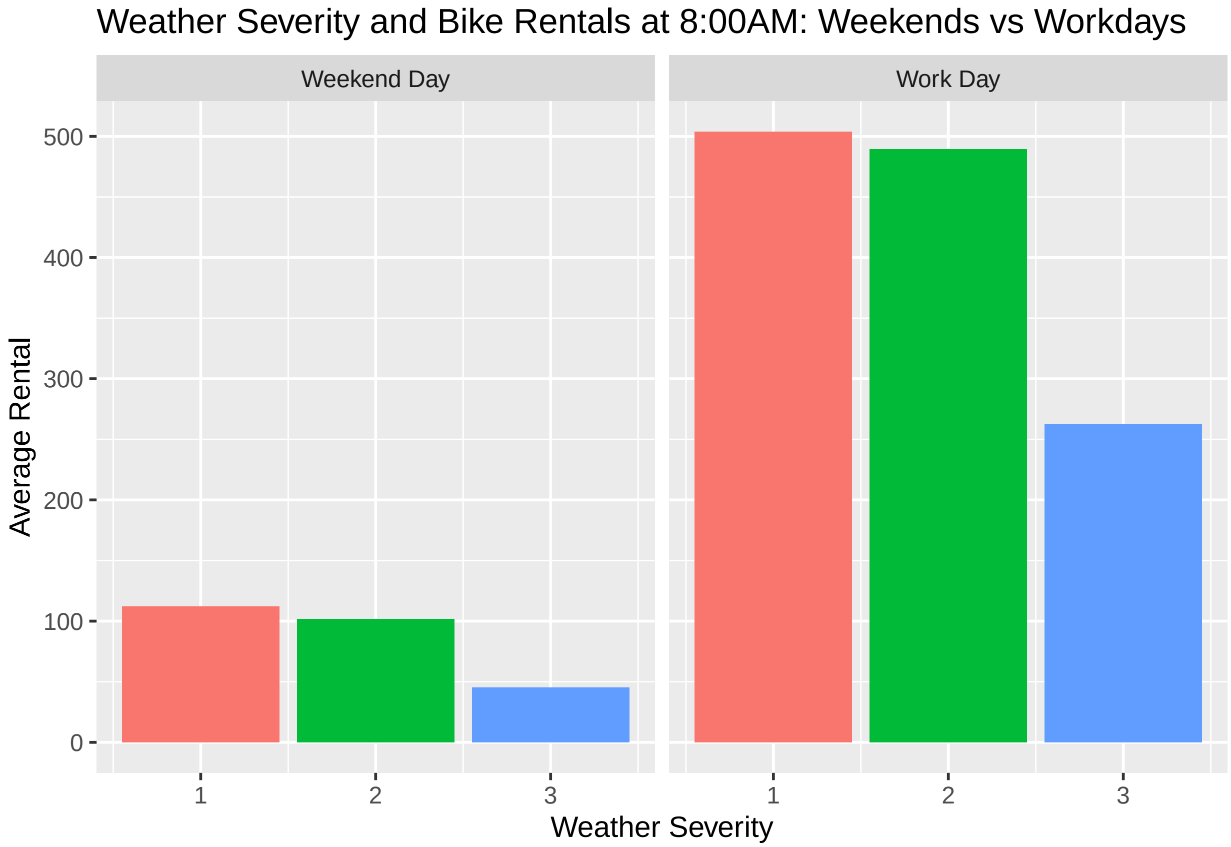
There are two spikes in bike rentals, with a slight uptick around the lunch hour. The spikes correspond to the 8:00AM and 5:00PM rush hours, respectively, and the uptick represents increased demand from lunchtime. Rentals slowly taper off into the evening.

Average Bike Rentals Thoughout the Day, Weekend vs Weekday



Here, we see two patterns emerging; what we observe for work days is like what we saw in the previous graphic, with the three upticks in bike rentals throughout the day representing rush hour, the lunch hour, and the 5PM rushes, respectively. For weekend days, we see that utilization tends to increase rapidly between 6AM and 12PM, and peaks at 2PM before gradually tapering off. Interestingly, the rental demand decreases far more sharply after 5PM on workdays, whereas they seem to slowly taper off on weekend days.

Average Bike Rental at 8AM by Weather Type

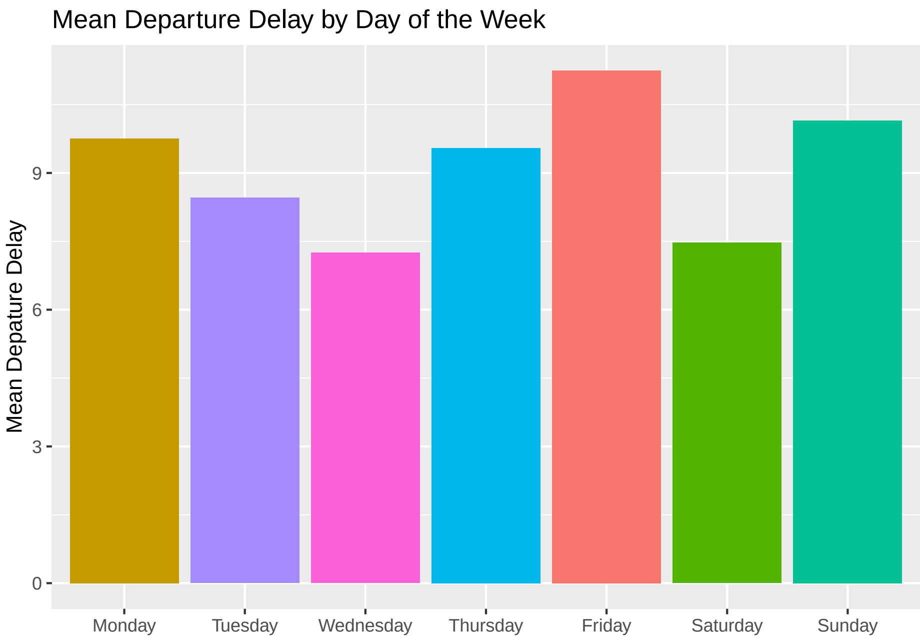


Here, we see that bike rental demand at 8:00AM is far greater on work days than it is on weekend days. This isn't surprising, as 8:00AM might be a bit early for a bike ride for most people.

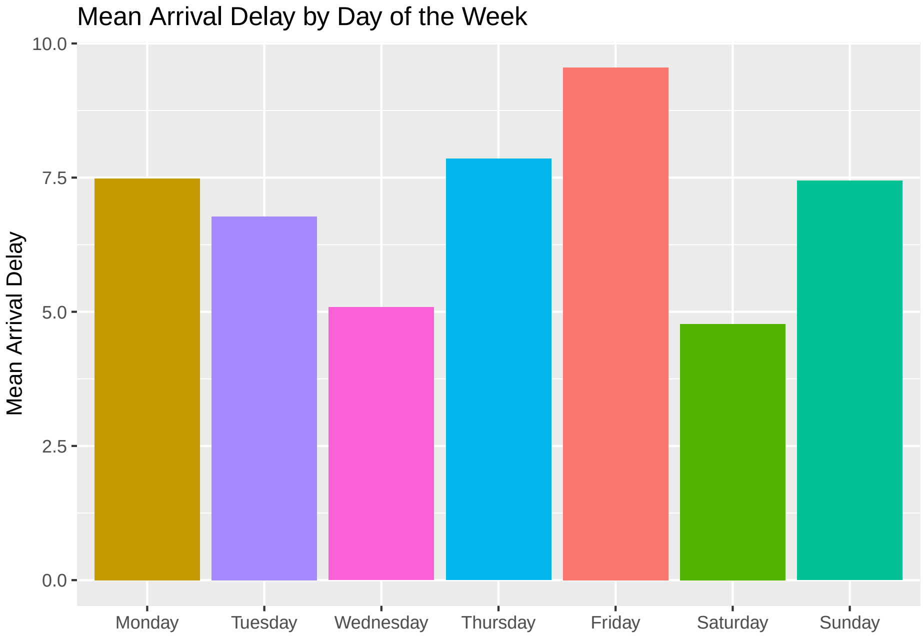
Question # 3: ABIA Flight Data: When to Fly and Which Carriers to Avoid to Minimize Cancellations and Delays

The goal of these figures is to help readers discern when to fly and who to fly with in order to minimize cancellations, arrival delays, and departure delays. First, we plot the average number of cancellations and delays for days of the week and months of the year. Then, we plot cancellations and delays by airline carrier using the UniqueCarrier code provided in the data.

First, the average departure delay by day of the week:

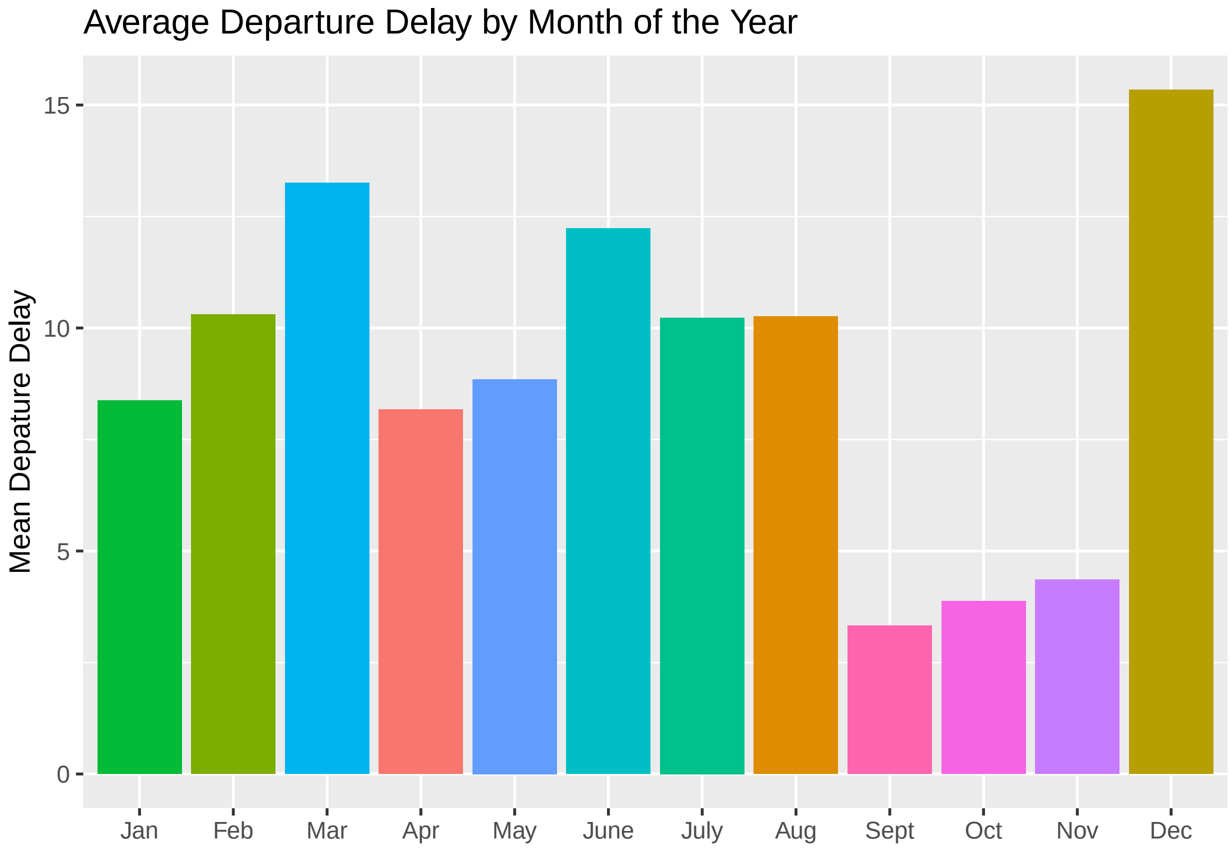


Not surprisingly, Fridays and Sundays have the greatest departure delays. This could be the result of people tending to fly out on a Friday, and return the following Sunday. We can see that delays are also higher on Mondays as well, which could be reflective of travelers embarking on work trips.

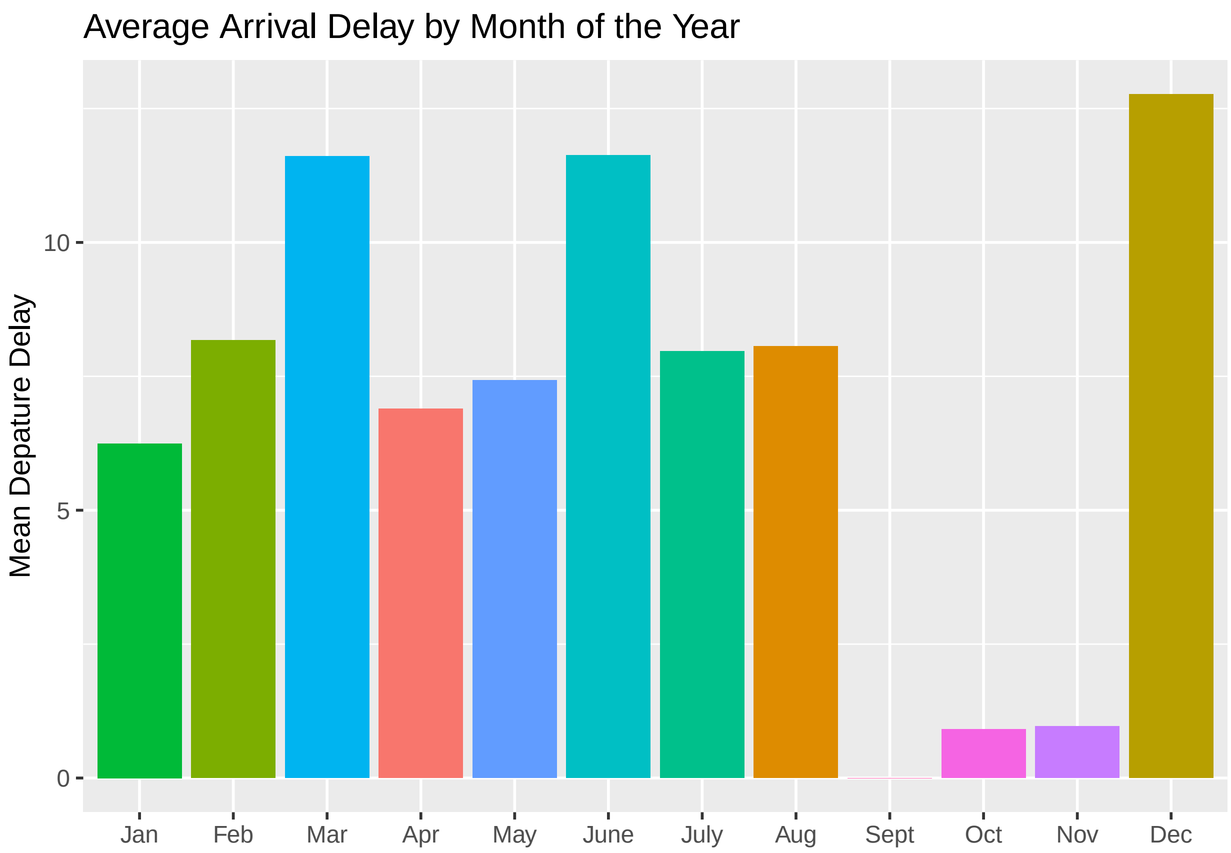


As with departure delays, arrival delays appear to spike on Fridays and Sundays.

Next, we plot the average departure and arrival delay times by month of the year:

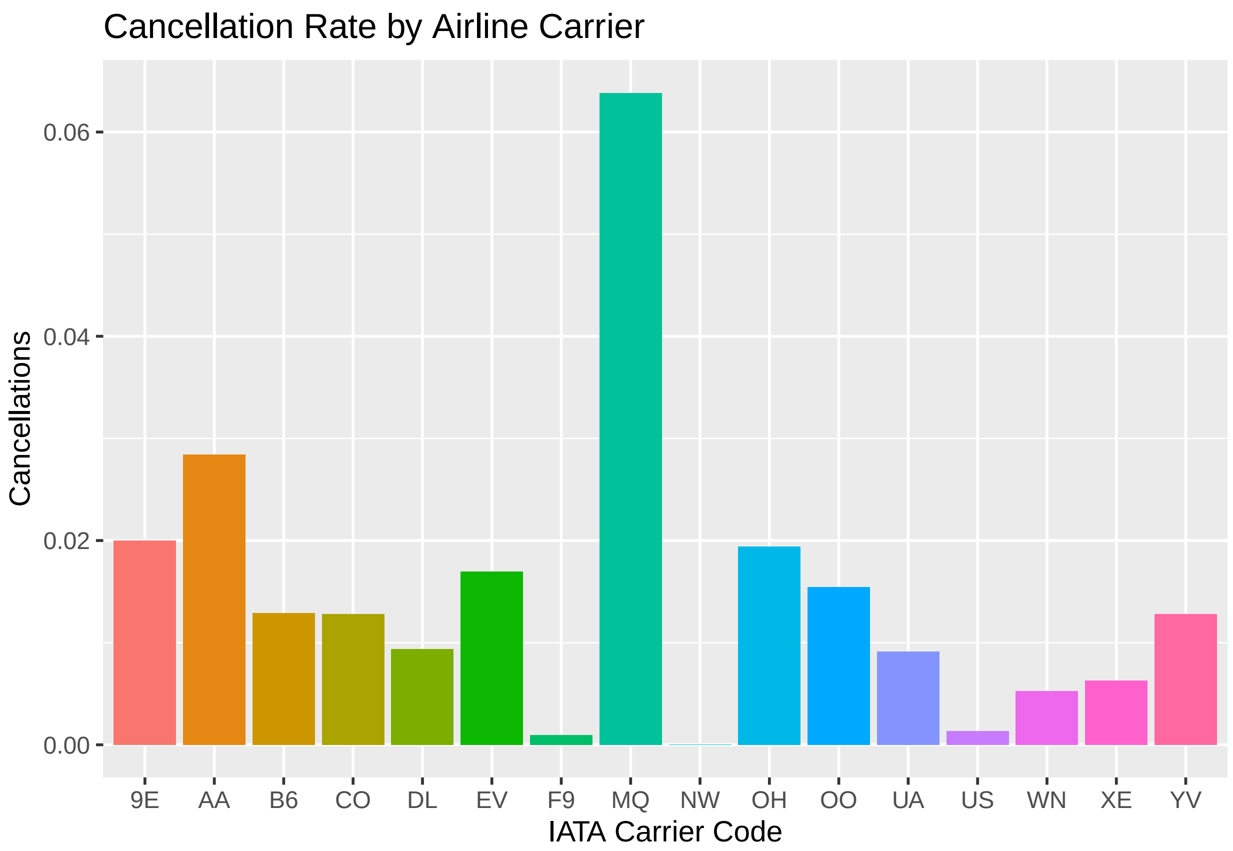


Clearly, the worst time of the year to travel in terms of departure delays is December due to the holidays. After that appears to be March, which could be due to people traveling for Spring Break. Next would be June, the beginning of summer vacation for most of the United States. The best time of the year to travel to minimize departure delays would be between the end of Summer and the beginning of the holidays.

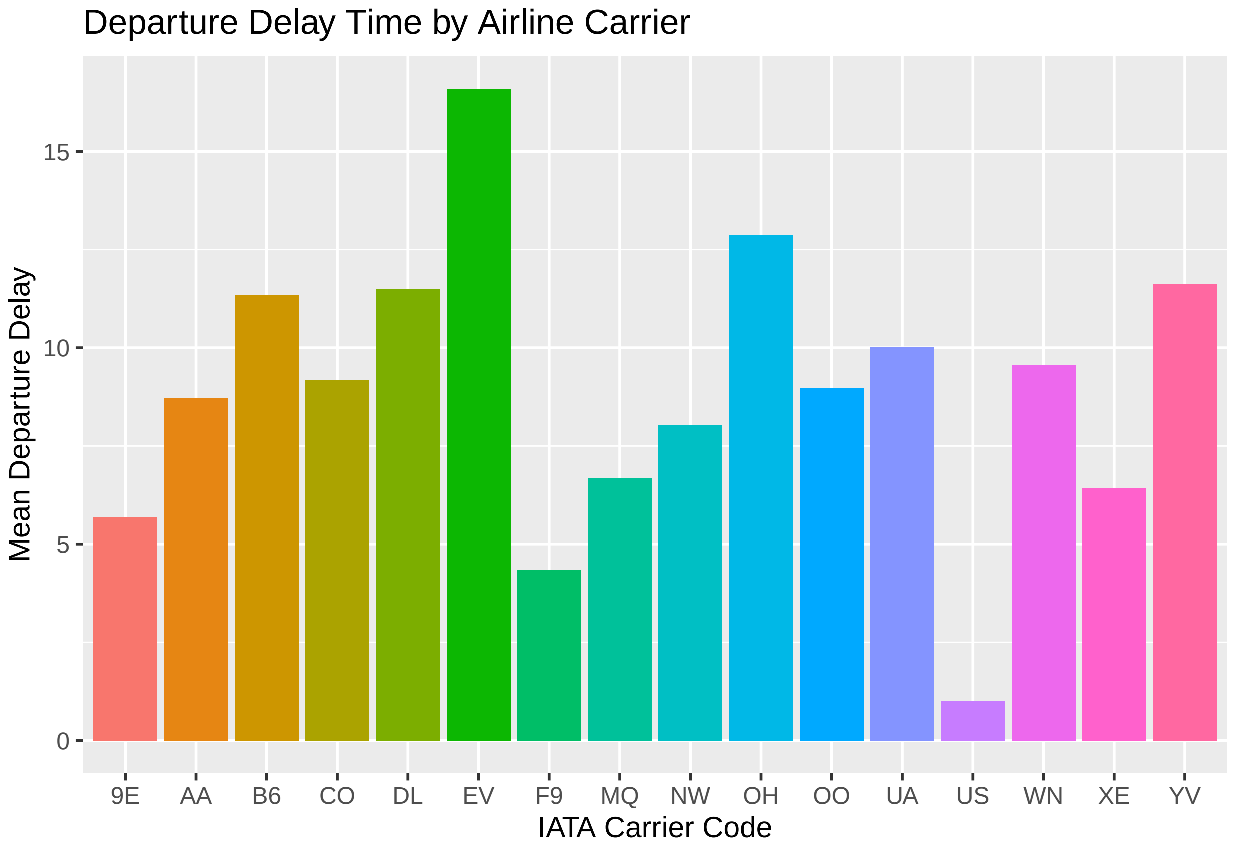


As with our previous monthly figure, the worst times to travel when considering for arrival delays is December. March and June also have significantly higher delay times as well. September has virtually no delay, and October and November also have disproportionately low delay times.

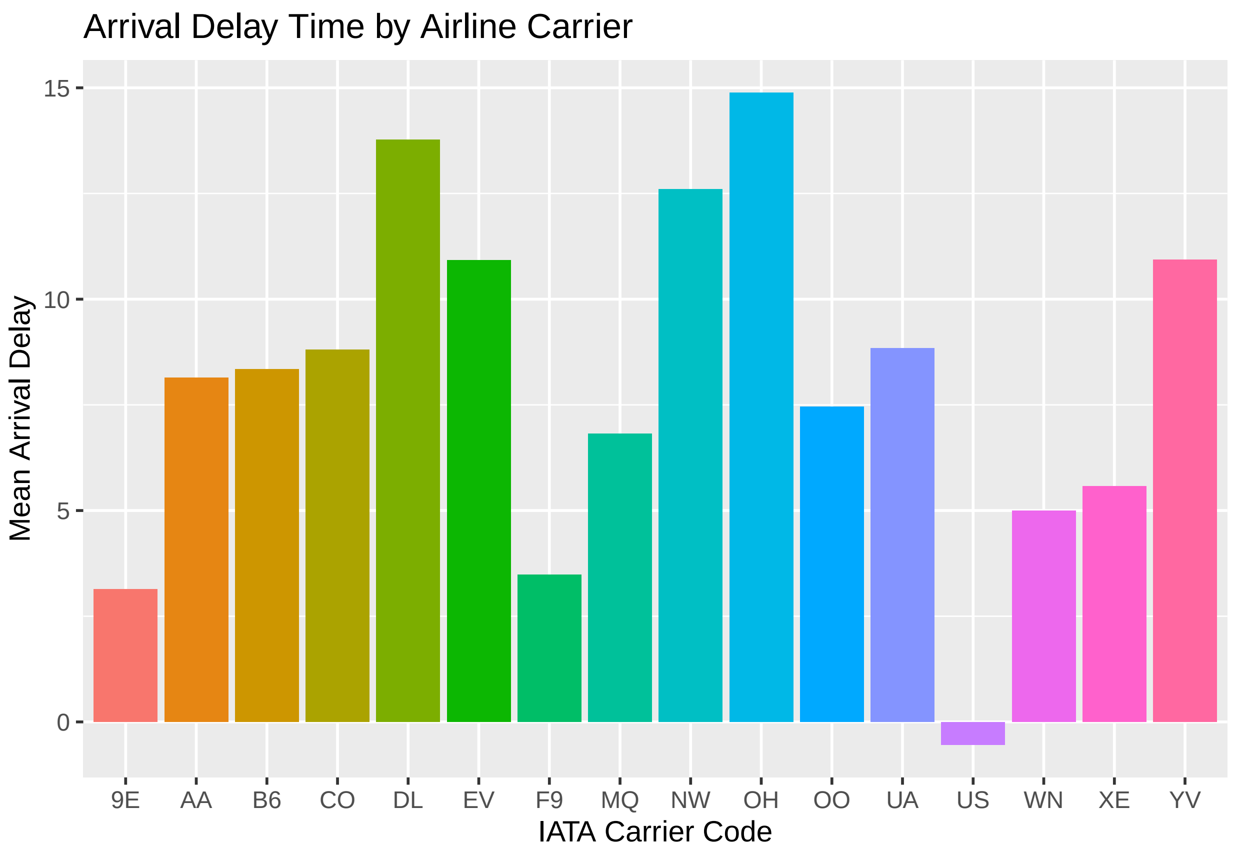
Now, we plot total cancellations by carrier to see which carriers are the worst offenders.



It would appear that Envoy Air (MQ) has the highest rate of cancellations by far. We can see that it has more than twice the number of cancellations as American Airlines (AA), which has the second-highest number of cancellations. PSA Airlines(OH) comes in third.



In terms of departure delays, it seems that ExpressJet(EV) has the highest average departure delay time. The second-highest average departure delay time belongs to PSA Airlines(OH) with several contenders tying for third-highest departure time: JetBlue(B6), Delta(DL), and Mesa(YV). The shortest departure delay times belonged to US Airways(US), Endeavor Air(9E), and Frontier(F9).



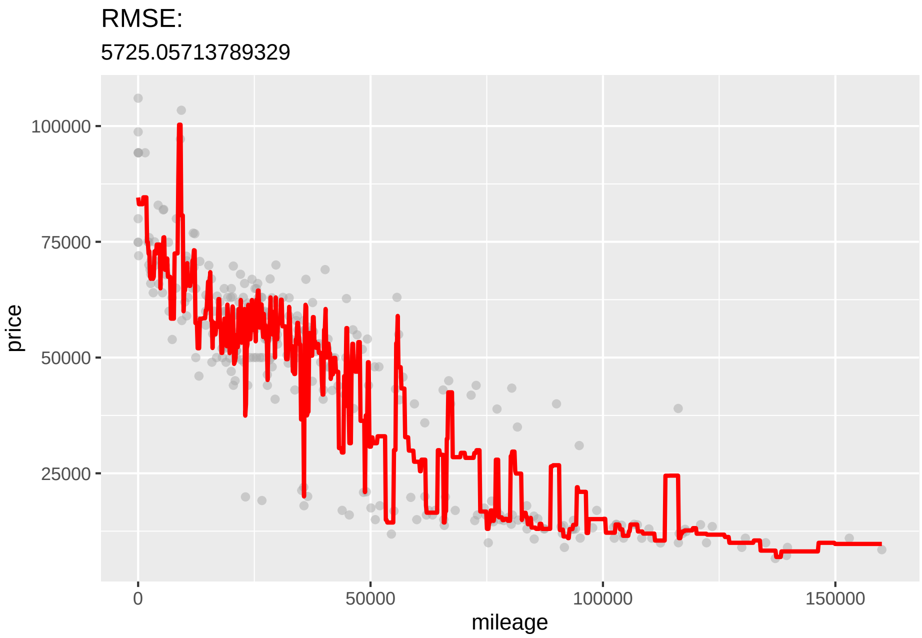
Arrival delays tells a similar story as before, but with some differences. First, the worst offender in terms of arrival delays is PSA Airlines(OH), with an average wait time of about 15 minutes. Next are Delta(DL), Northwest(NW), and Mesa Airlines(YV) with wait times of just over 10 minutes. The carriers that performed best in this category were United(US), Endeavor Air(9E), and Frontier(F9).

Question 4) K-Nearest Neighbors

For this exercise, we are using the sclass.csv dataset containing data on over 29,000 Mercedes S Class vehicles. For the purpose of this exercise, we are only focusing on two trims (similar to a sub-model designation): the 350 and the 65 AMG.

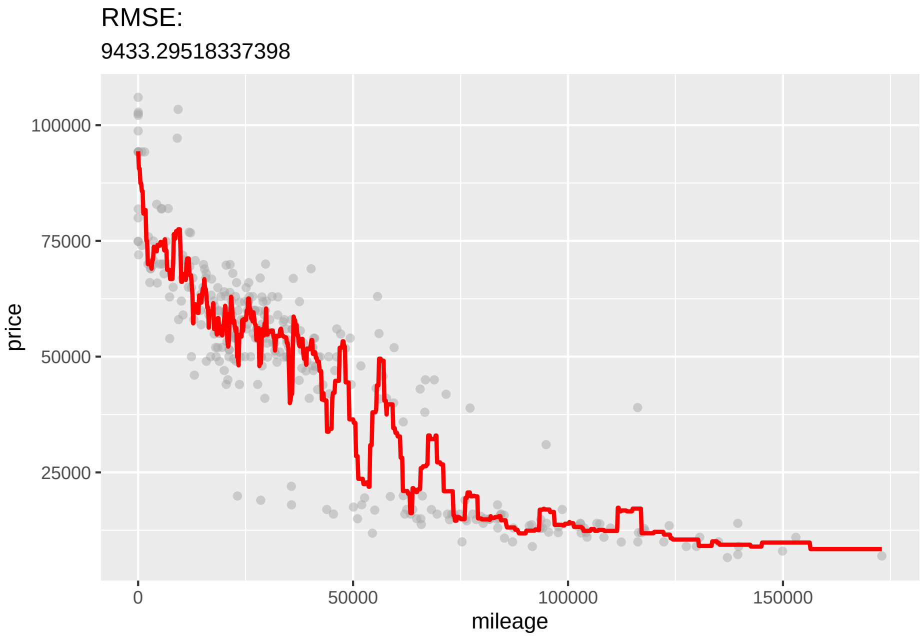
The goal is to build a predictive model of the price of a vehicle, conditional on its mileage. To do so, we'll use k-nearest neighbors, a non-parametric learning algorithm that predicts $f(x^*)$ by taking an average of the k x_i values closest to x^*. The idea is to find the optimal value of k to make predictions on $f(x)$. The following pages show the performance of the model over a range of k, followed by the selection of the optimal k to minimize our prediction errors..

k = 2

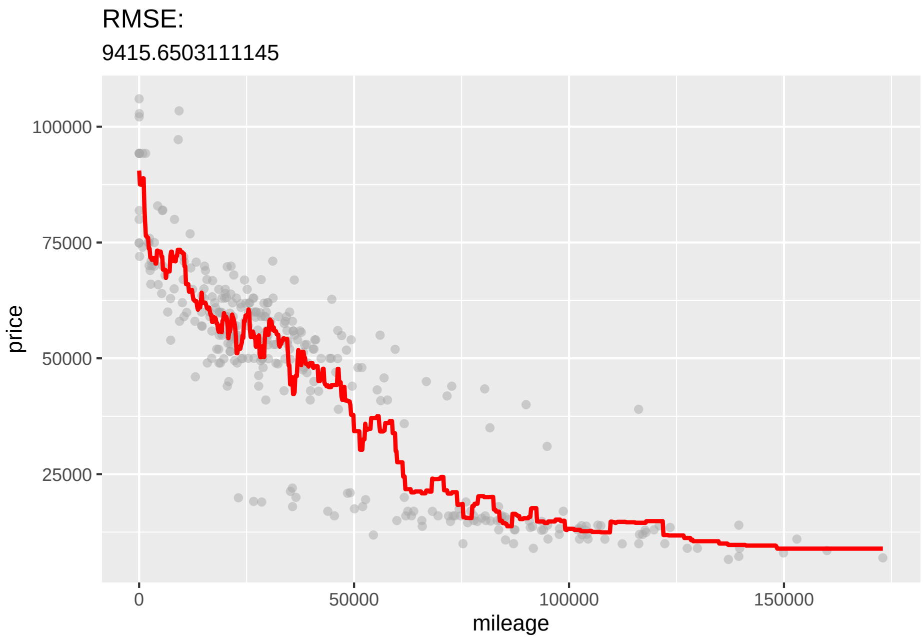

RMSE:
5725.05713789329

Above is the scatterplot representing the selected variables from the training set and is overlayed with the predicted curve for the Class 350 at k=2. There's a high degree of variability due to k being so low, so this level of k is not optimal for making predictions on price. As we continue to increase k, we should witness a gradual smoothing of the plotted curve, as well as a gradual decrease–followed by an increase–in RMSE.

K =5:

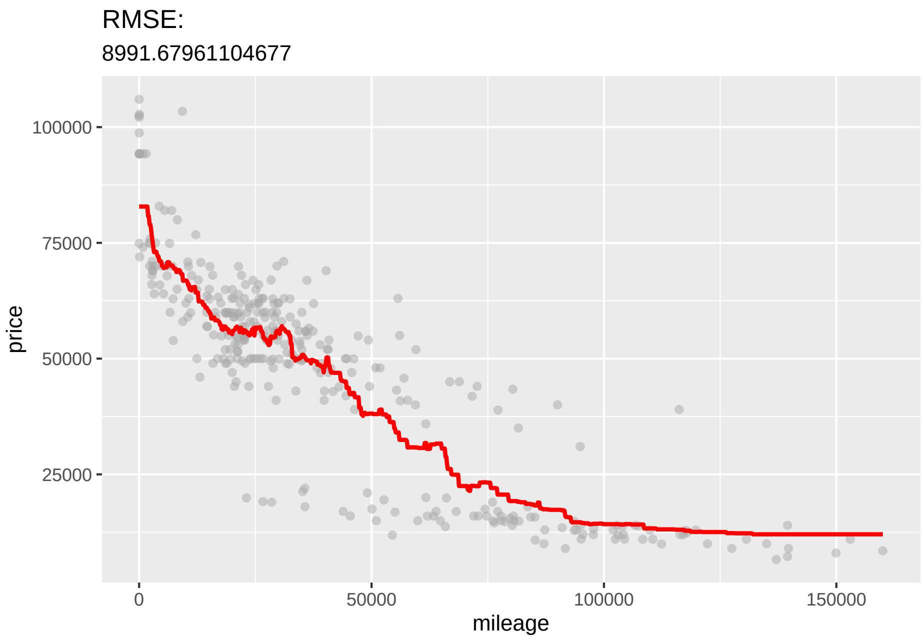

RMSE:
9433.29518337398

K = 10

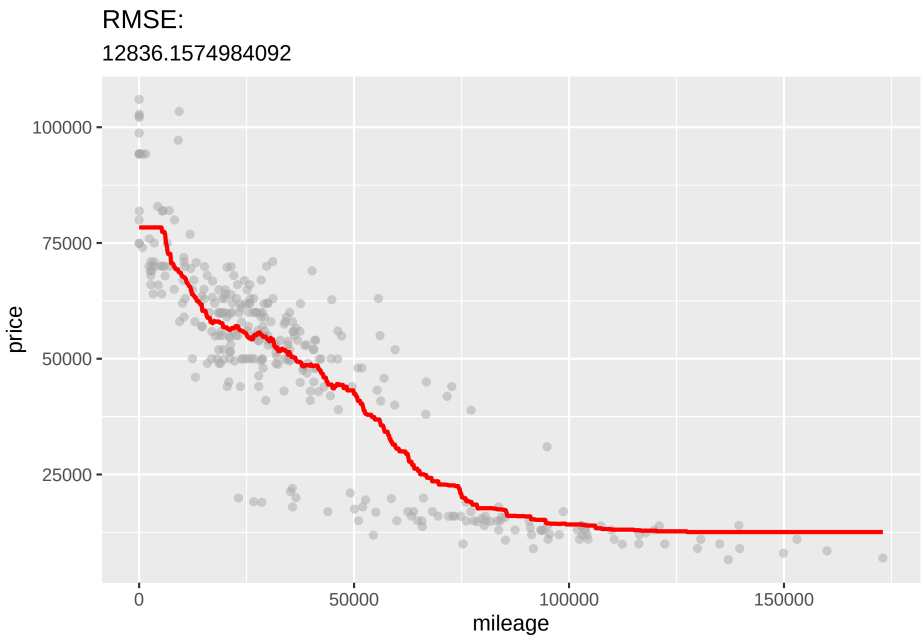


RMSE:
9415.6503111145

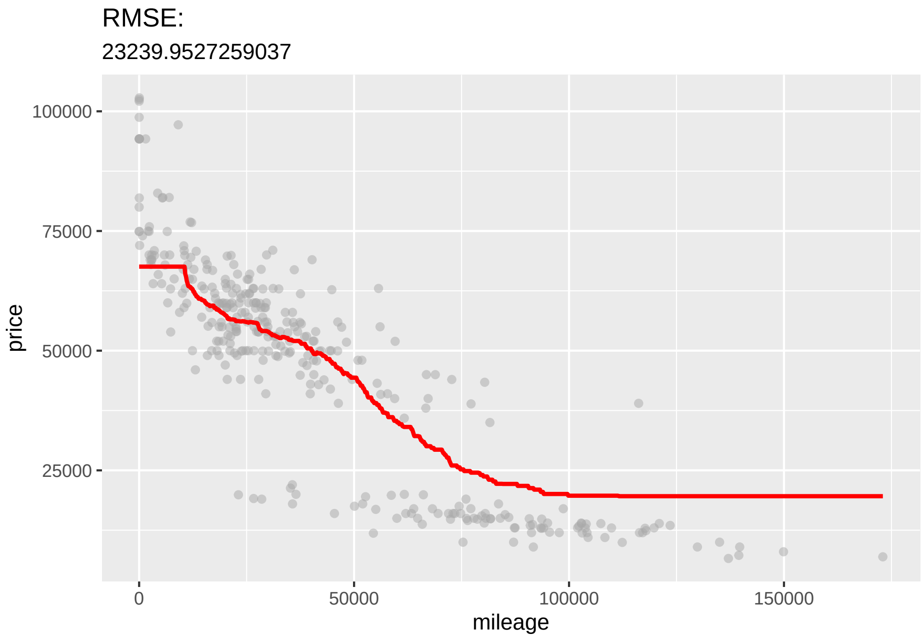
k = 25



RMSE:
8991.67961104677

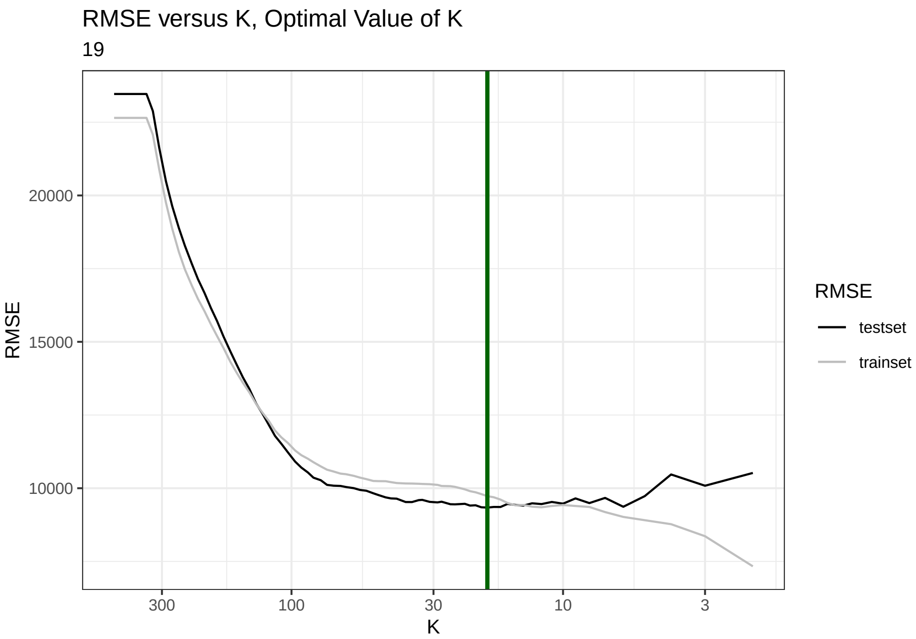


At k=50, the RMSE has seen another significant jump, if we look closely at when the mileage is equal to zero, we can see that our predictions are becoming worse near the endpoints of our data. This is consistent with what we know about how KNN makes predictions: because there are less observations at the endpoints of the range, predictions for values in these regions tend to be biased.



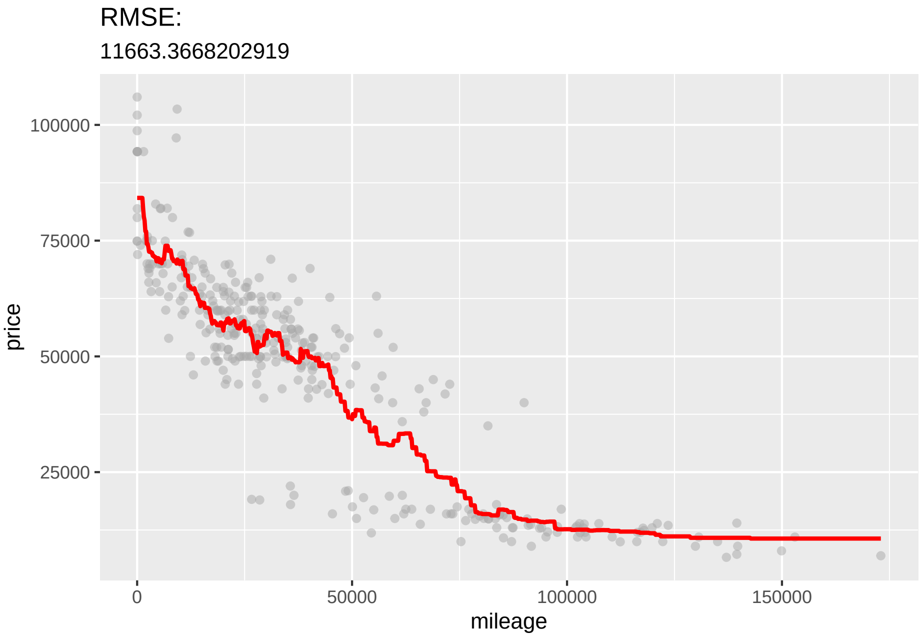
The RMSE has again increased sharply. At this level of k, we can see that the predictions for price past a mileage of about 75,000 are severely biased upward. It isn't worth increasing k past this point, because we know that our predictions will become increasingly biased, eventually fitting a curve that is the average value of price for all observations in the data set.

To get an idea of what the optimal k is, we can plot the RMSE versus K, comparing the performance of our model between the training and test sets. We want to pick a k for our model that gets the lowest RMSE when applied to the test set:



As we can see from the graph above, the optimal value of K is 19. This value is determined by finding the point where the RMSE is lowest in the training set. At K=19, the RMSE for the test set is visually below the trainset, indicating that our model performs better on our testing set at this level of k. We can see that this would actually persist for a range of k, just beyond the point of k = 100. However, as we saw previously, at this level of k, the amount of bias takes away predictive power for our model.

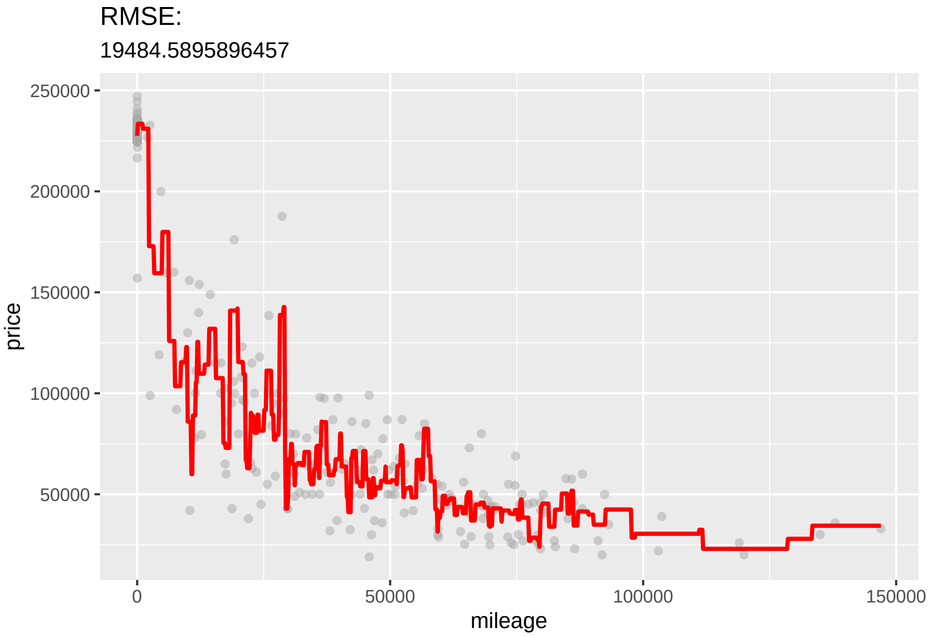
Optimal K:



Here, we have the plot for the optimal level of k for the 350 trim, which was found to be 19.

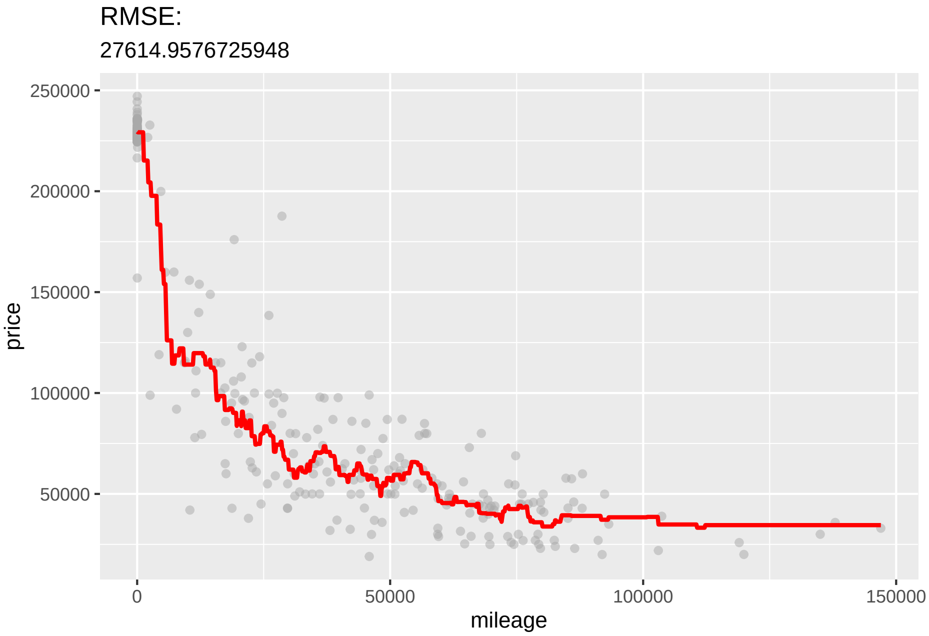
Now, we'll repeat the process above but for the other subsection trim category, 65 AMG. As before, we'll start at a low number of k.

k = 2

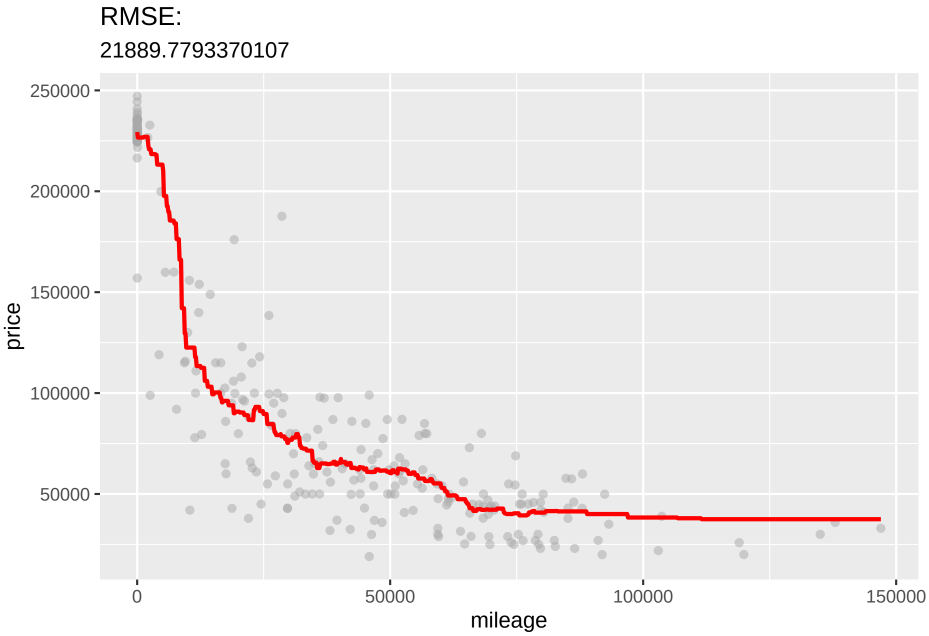


Again, this level of k produces a figure with a lot of volatility.

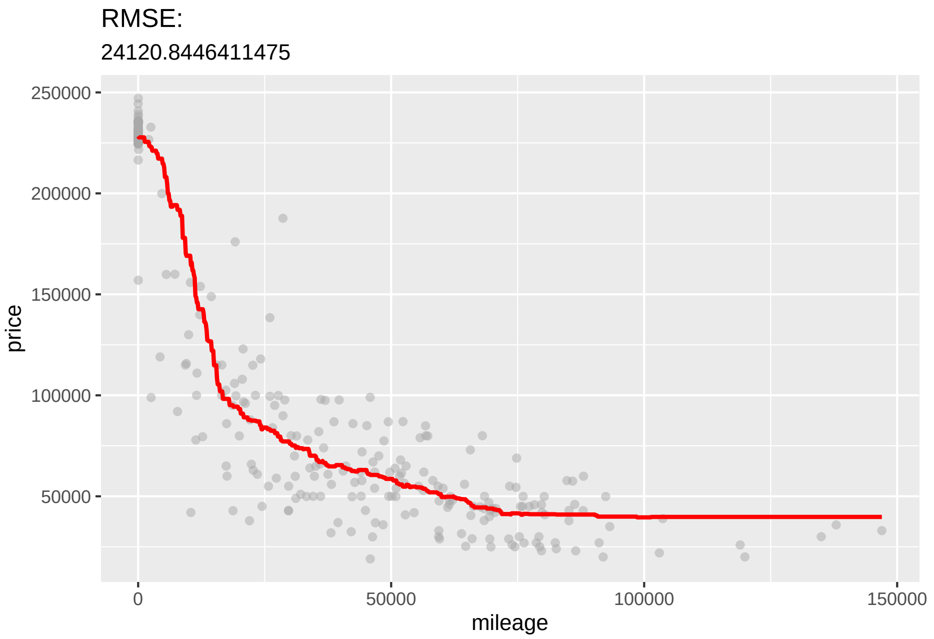
k = 10


RMSE:
27614.9576725948

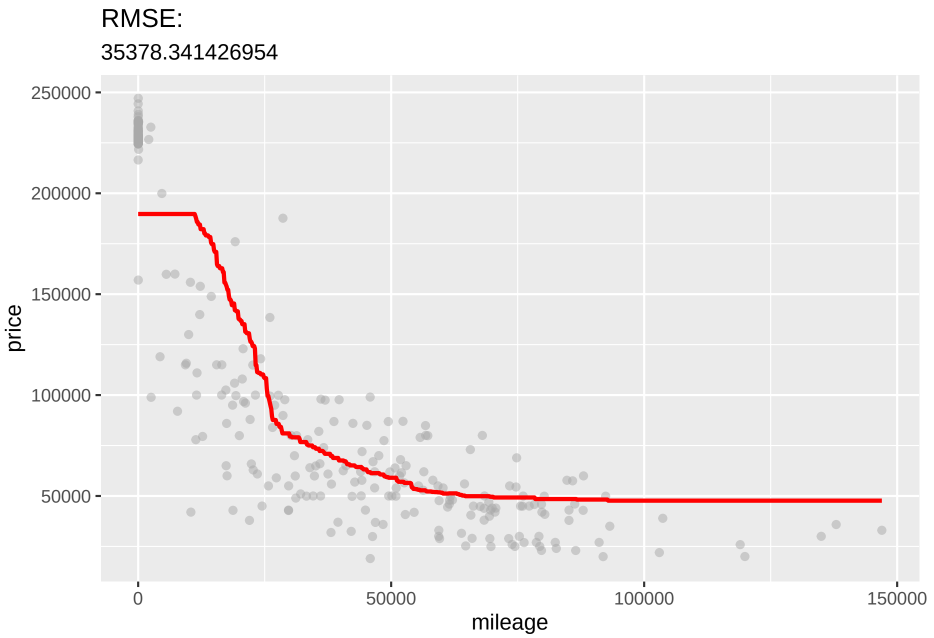
k = 25


RMSE:
21889.7793370107

k = 50

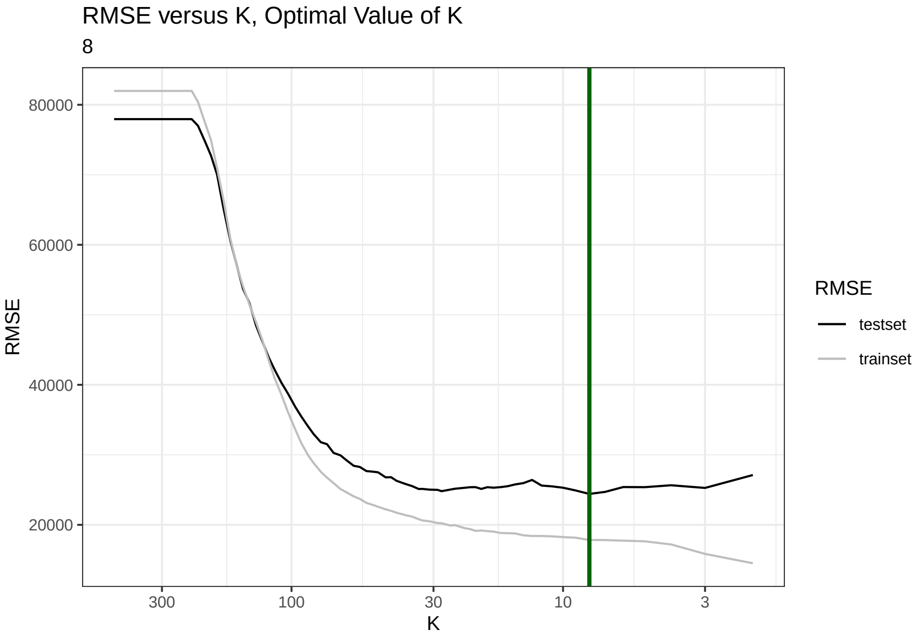


k = 100



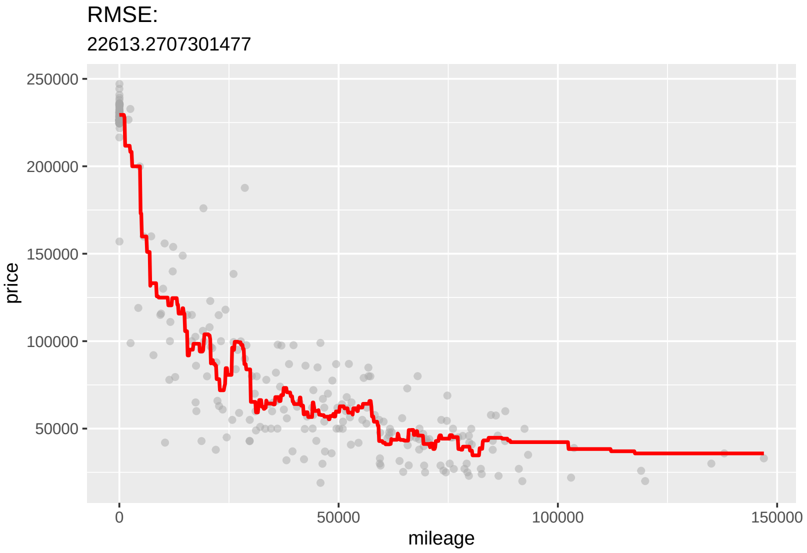
Here, the bias in the prediction is visually apparent for values near the endpoints for the range of mileage. We can see that the price for a vehicle with over 50,000 miles is going to tend to be biased upward.

Fitting RMSE vs K



The optimal K value for the 65 AMG trim is only 8, significantly lower than what we have found for the Class 350. In our opinion, this has to do with the differences in distribution of prices between the two groups. For example, when we look at the scatter plot for the Class 350, we can see that there is this sort of gap in the price range, seeming to create two different groups of cars. Whereas on the other hand, when we look at the 65 AMG data point distribution, there is a more "continuous" flow of the data points. Consequently, this means that in order to bridge the gap in prices for the Class 350, it has to take the average of more data points, as it has to make up for the lack of information.

Optimal K:



"'