# Exercise 1

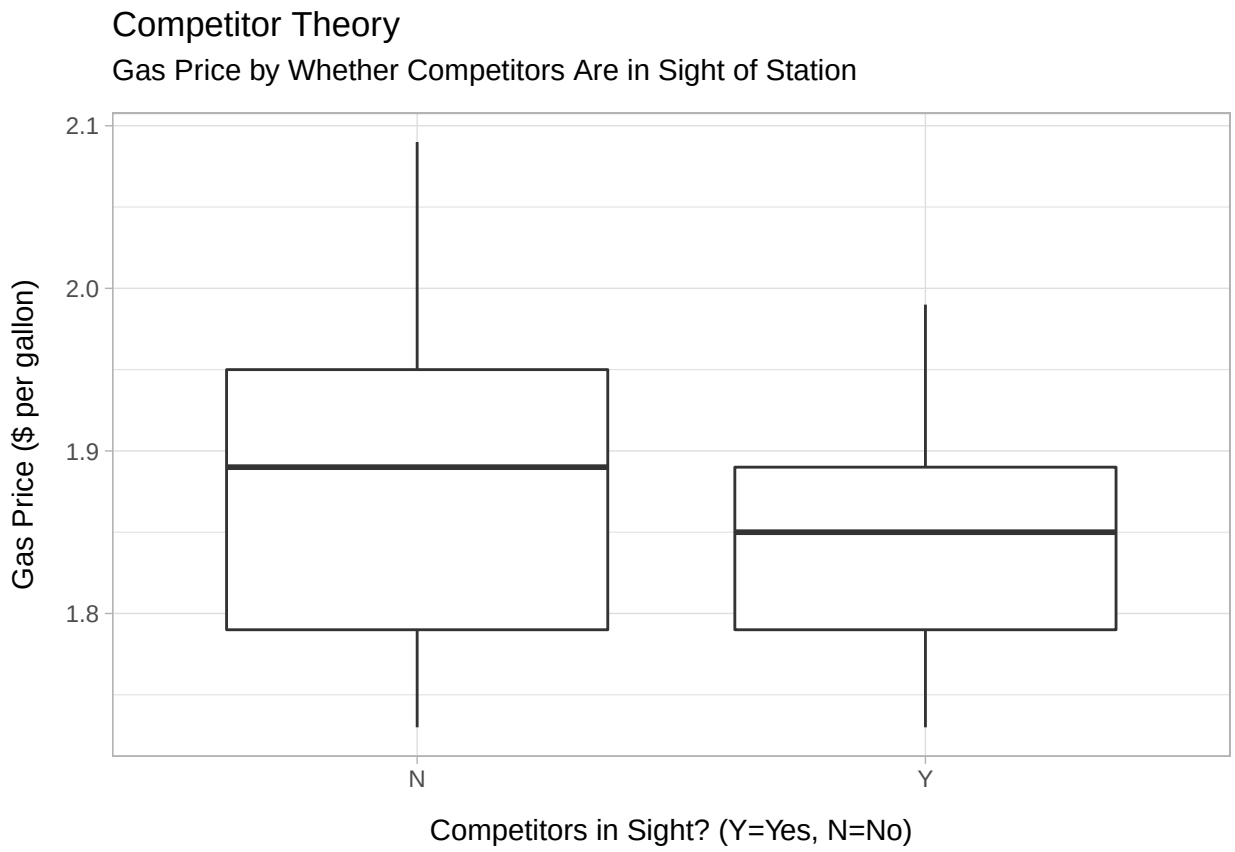
Robert Toto

2/08/2021

## 1) Gas Prices

In this problem, we assess five possible theories about consumer gas prices in Austin, TX using 2016 price data from 101 gas stations in the city. The theories consider five different factors that may drive gas prices: proximity of competitors, income of local neighborhood, brand name, proximity of a stop light, and proximity to a highway.

### A) Competitor Theory: Gas stations charge more if there is no competition in sight

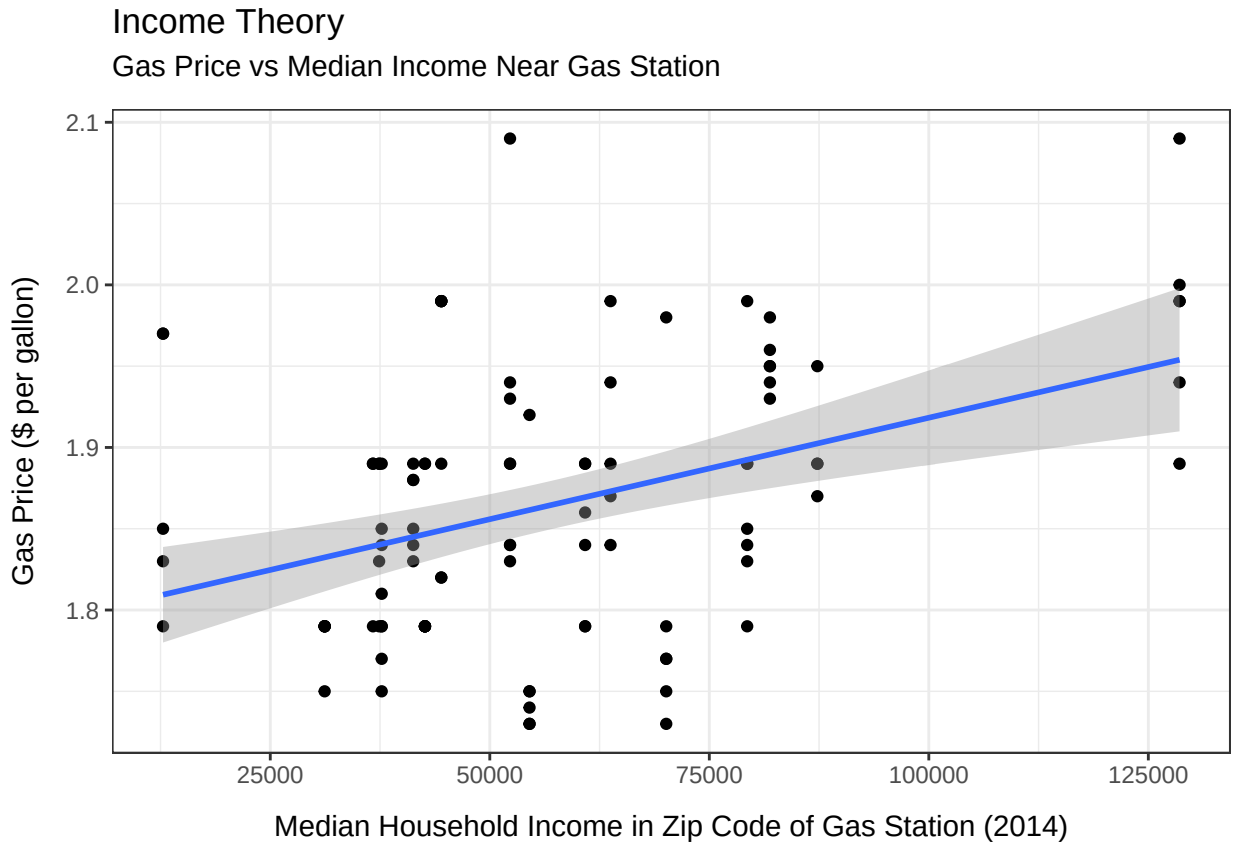



If a driver has more than one option for a gas purchase, she will minimize her expenses by choosing a station with lower prices, all else held equal. If the driver can physically view more than one gas station, the cost of optimizing (minimizing) expenses drops essentially to zero because she does not need to research alternatives to a single station. She has full information on more than one from physically viewing them.

Gas stations themselves would understand this and lower their prices to compete with nearby (viewable) competitors. However, if no competitor is nearby, a station has a localized monopoly and will charge a greater mark-up. The boxplot below attempts to validate this monopoly pricing theory by comparing the distributions of prices set by stations when a competitor is within view to the distribution when no competitor is visible.

The boxplot shows a clear visual difference in the two distributions. When a competitor is **not** in sight, the median price a gas station charges is $0.04 higher per. The boxplot further demonstrates that the upper limit of pricing is much greater for the local monopolies (no competitors in sight). For these stations, the top 25% of prices fall between $1.95 and $2.09, whereas the top 25% of prices for stations with visible competition is entirely beneath $2.00. One can see that monopoly pricing has a longer upper range, higher maximum, and overall wider distribution, demonstrating that local monopolies have much greater pricing power and price-setting flexibility. The boxplot certainly lends visual credibility to the competitor theory. One can expect to pay lower prices when there is direct competition.

**B) Income Theory: The richer the area, the higher the gas price**

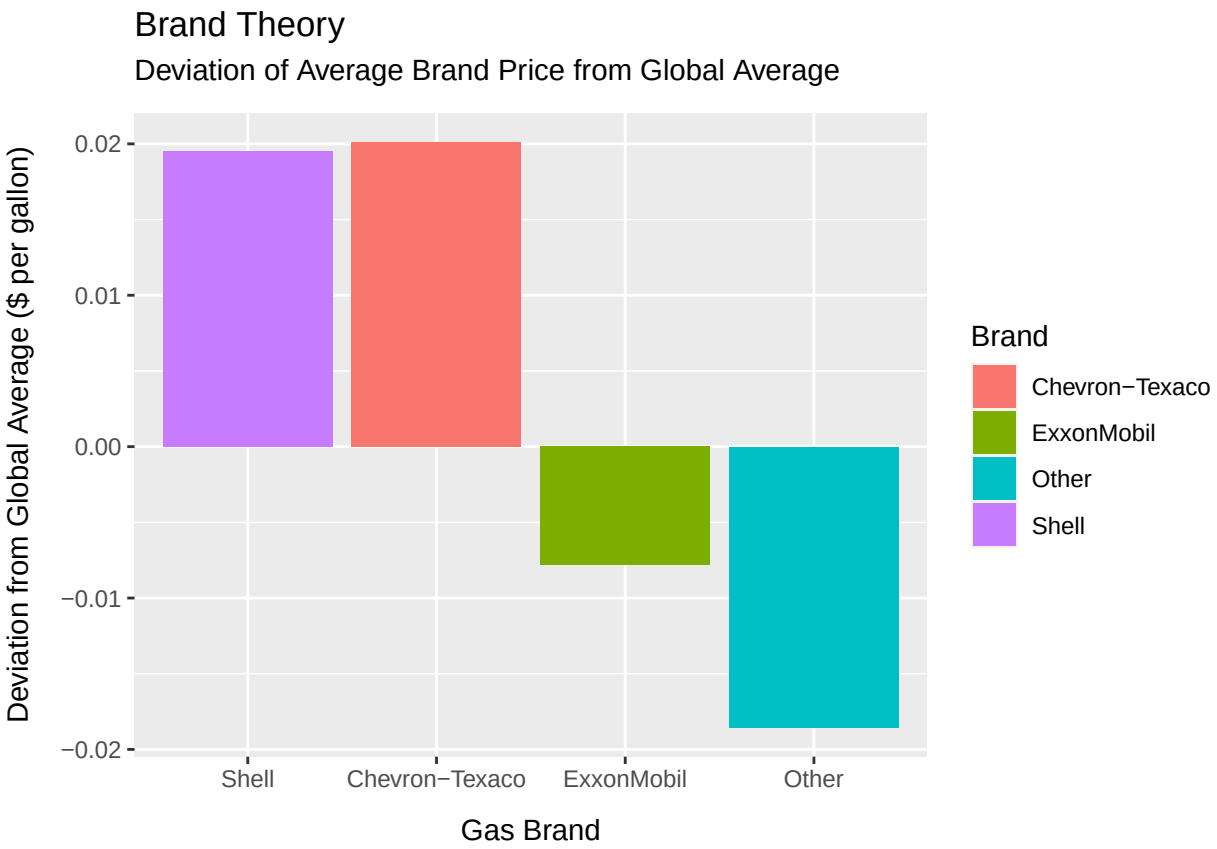



In theory, a neighborhood with higher incomes would have higher local gas station prices because people with greater expendable income have a higher market willingness-to-pay for basic goods, like gasoline, since these goods constitute such a small portion of their overall income. Conversely, low-income gas stations would have lower mark-ups and overall prices because people with lower incomes are more sensitive to the prices of basic goods, like gasoline, since these goods constitute a large portion of their overall income and since they may substitute away from some of them. For instance, if a gas station charges a very high price in a low-income neighborhood, local residents are likely to substitute away from personal transportation to public transportation–a cheaper option.

The scatterplot weakly demonstrates that, as the income of a neighborhood rises, so do the gas prices. But

this would likely not be true if not for the extreme values past the $125,000 income level. If these values were removed, the regression line would likely be almost totally flat, indicating no relationship. The data at the very high range of incomes is nonetheless interesting. For neighborhoods with median incomes greater than $125,000, the average gas price is never below $1.85 per gallon, which is about the median price for all average prices from neighborhoods with median incomes under $100,000. Interestingly, at incomes below $25,000, there is no relationship. Of the four data-points in this range, average gas prices can rise above $1.95 per gallon, well above the median of the overall data. It would be interesting to gather more data on gas prices exclusively in lowest-income neighborhoods to see if this lack of trend holds. Overall, this scatterplot indicates that there is no definite relationship between gas price and neighborhood income.

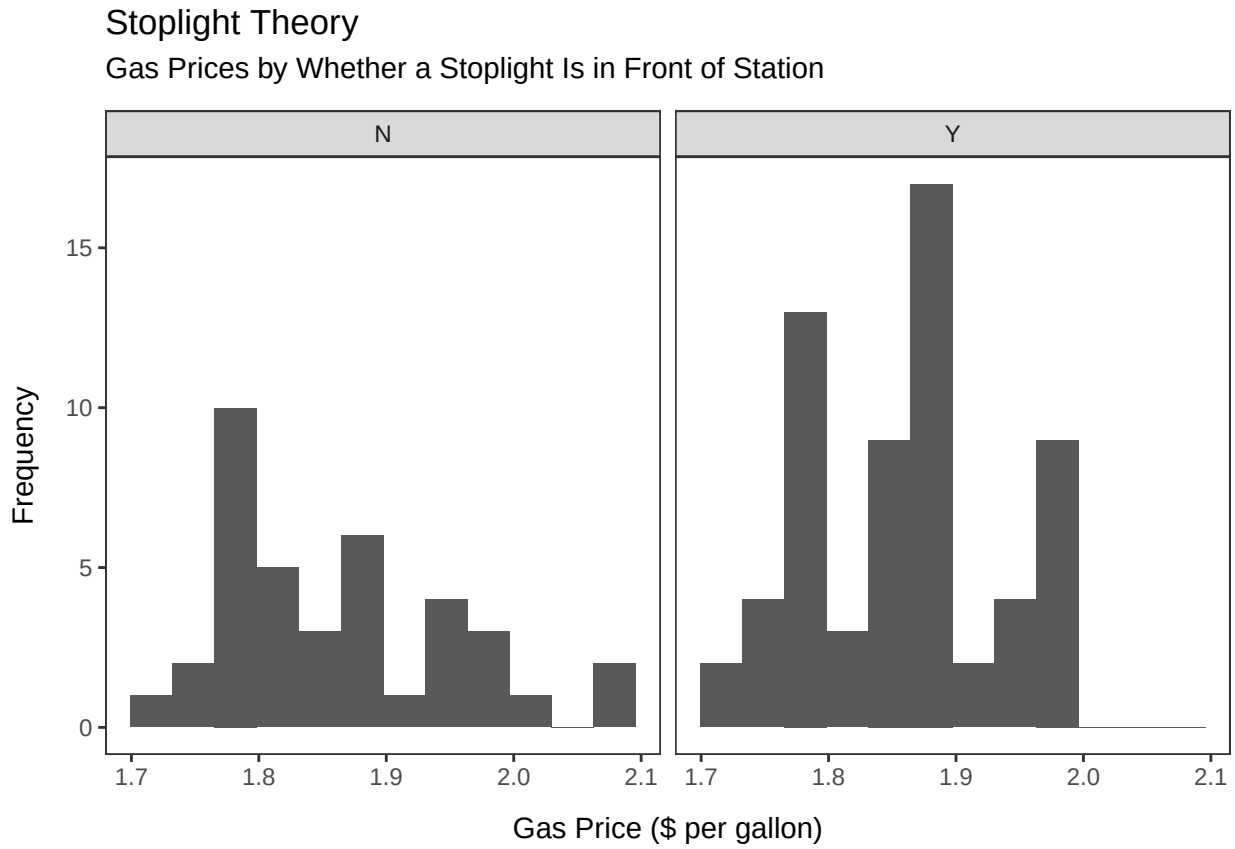
**C) Brand Theory: Shell charges more than other brands**




The initial theory claims that Shell-owned gas stations charge more for gasoline than other brands. One way to check this would be to calculate the average $ per gallon charged by each brand in the data. However, due to high competition of retail gasoline, these average values are extremely similar and graphically difficult to discern. A better way to test this claim is to calculate the "global average"–the average of all station prices in the dataset–and compare each brand category's average to the global average.

The bar graph shows the brand-level average gas price deviation from the global average gas price. Chevron-Texaco and Shell charge more than the global average. All other brands charge less. Comparing Chevron-Texaco to Shell, it can be discerned that Chevron-Texaco charges slightly more than the global average than Shell does. This indicates that Shell does not charge more than other brands. Rather, Chevron-Texaco charges the highest prices, with Shell as a close second.

**D) Stoplight Theory: Gas stations at stoplights charge more**

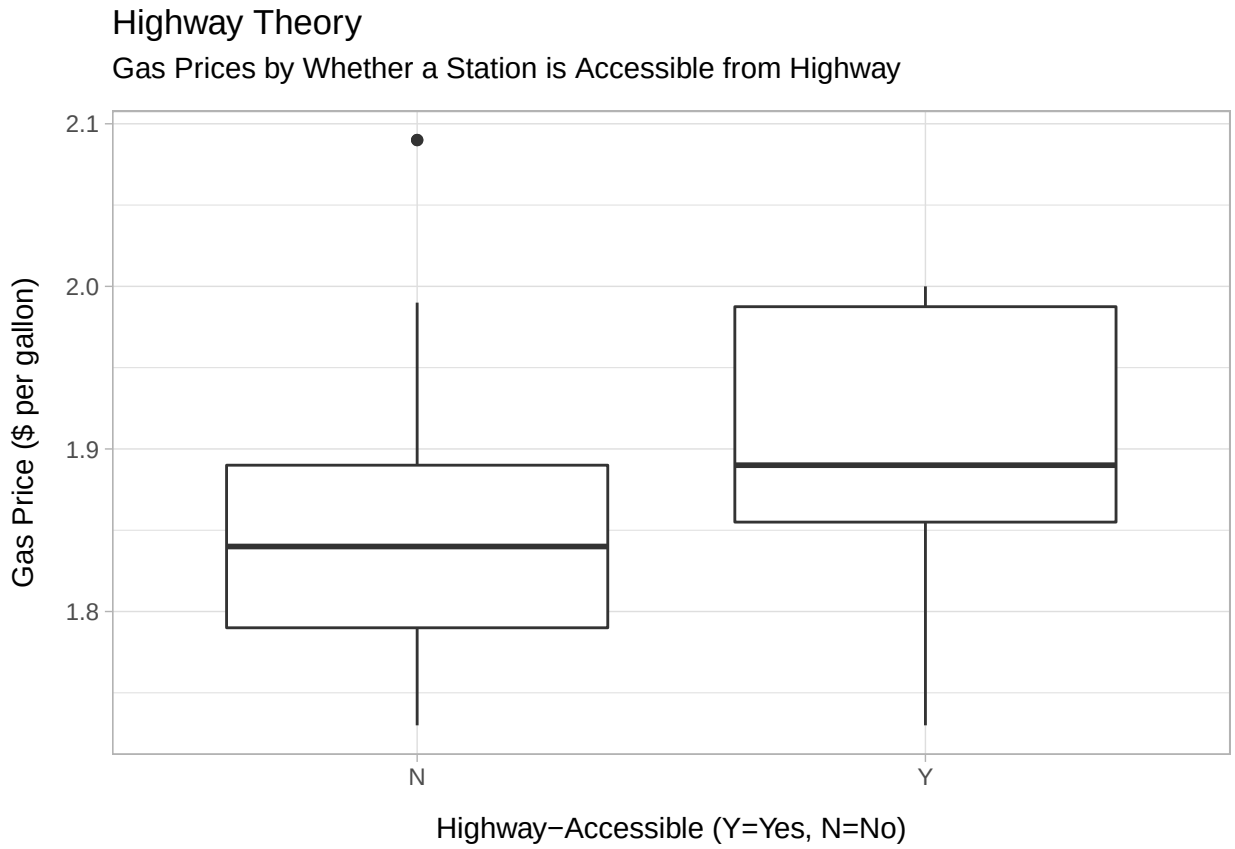



```
## # A tibble: 2 x 2
##   Stoplight mean_price2
## * <chr>           <dbl>
## 1 N                1.87
## 2 Y                1.86
```

The stoplight theory claims that gas stations charge higher prices when a stoplight is in front of the station. The reasoning is that if people are forced to stop near a gas station, they are more likely to make a snap decision to get gas since they are already stopped.

The faceted histograms compare price distributions between stations with and without a stoplight nearby. The graphics show that gas stations without a stoplight generally have lower prices (rightward skew). The stations that do have a stoplight have a higher frequency of prices close to $1.90 per gallon and above. So, graphically, the story is that stoplights are associated with higher prices. However, when we look at just the average prices of these groups in the accompanying tibble, we can see that average prices for non-stoplight gas stations are one cent higher than the average for stoplight gas stations. This tells the opposite story of the histograms, demonstrating the importance of data visualization. The reason the mean for non-stoplight stations is slightly higher is that there are generally less data points in this group, so it is easier for high prices to sway the average higher. Overall, the prices are very close between these groups.

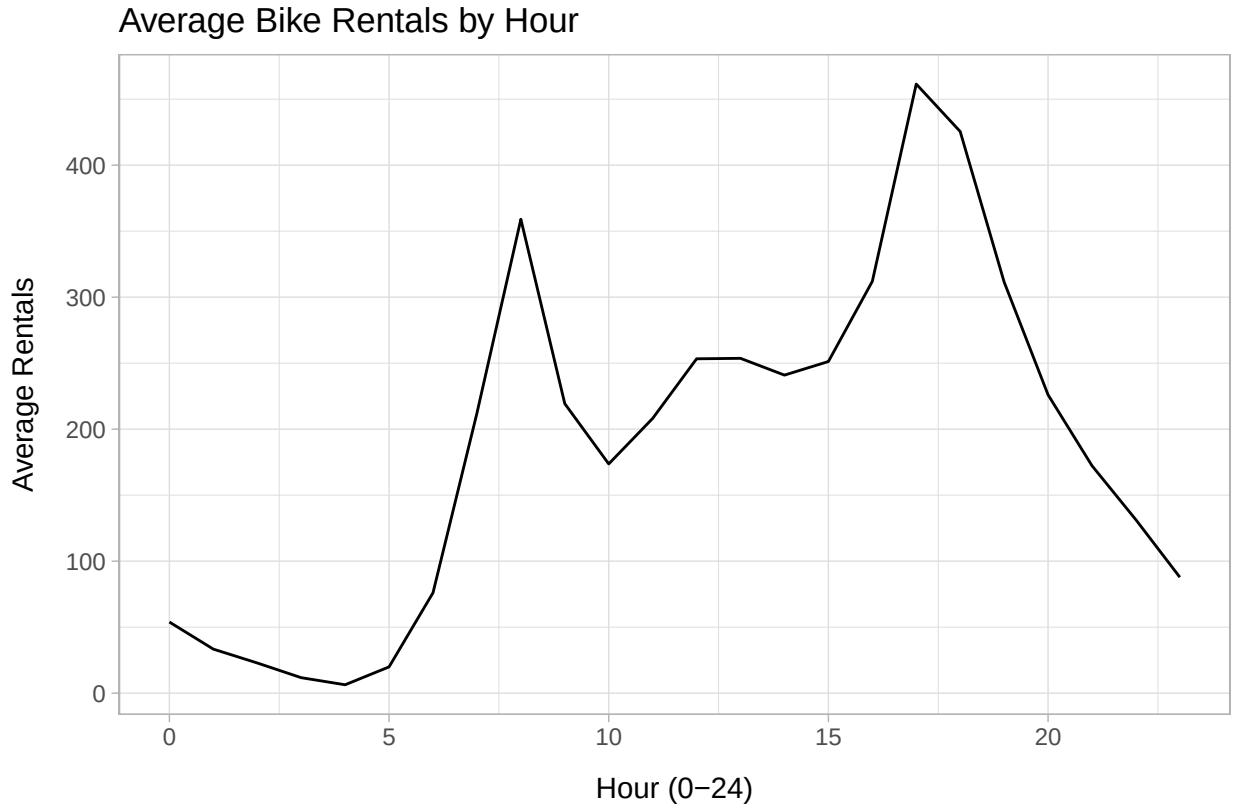
**E) Highway Theory: Gas stations with direct highway access charge more**




In theory, a gas station close to a highway (or highway entrance ramp) could charge higher prices due to higher demand from good accessibility, compared to a station not near accessible from a highway.

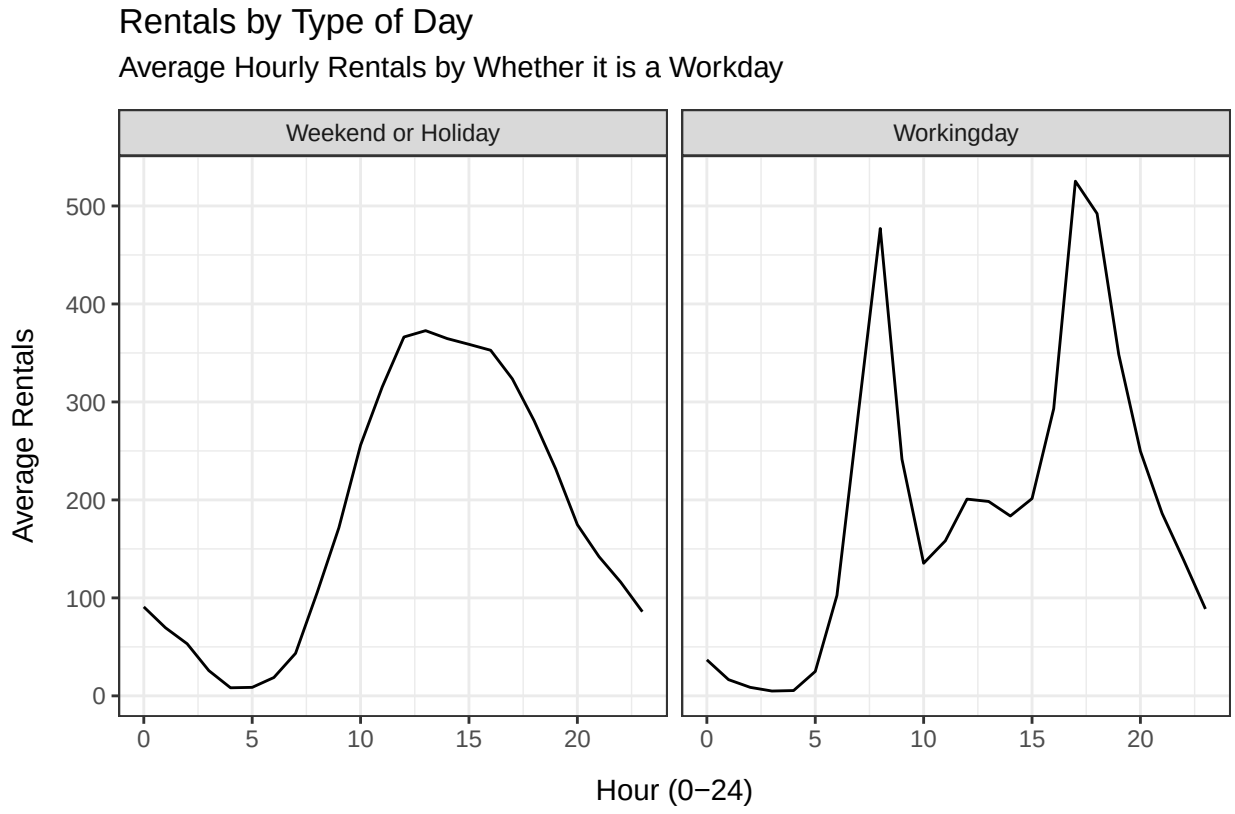
The boxplots demonstrate a large difference in price between these two groups. Stations accessible via highway have an average price of about $1.89 per gallon, while other stations have a lower average price of only about $1.84 per gallon. Furthermore, the distribution of highway station prices between the second and third quartiles is higher than (and has very little overlap with) the same quartile range of non-highway stations. It may be safely said that the highway theory has some merit, although the direction of causality remains unclear. What is clear, however, is that stations accessible via highway generally have higher prices.

## 2) Bike Share Network



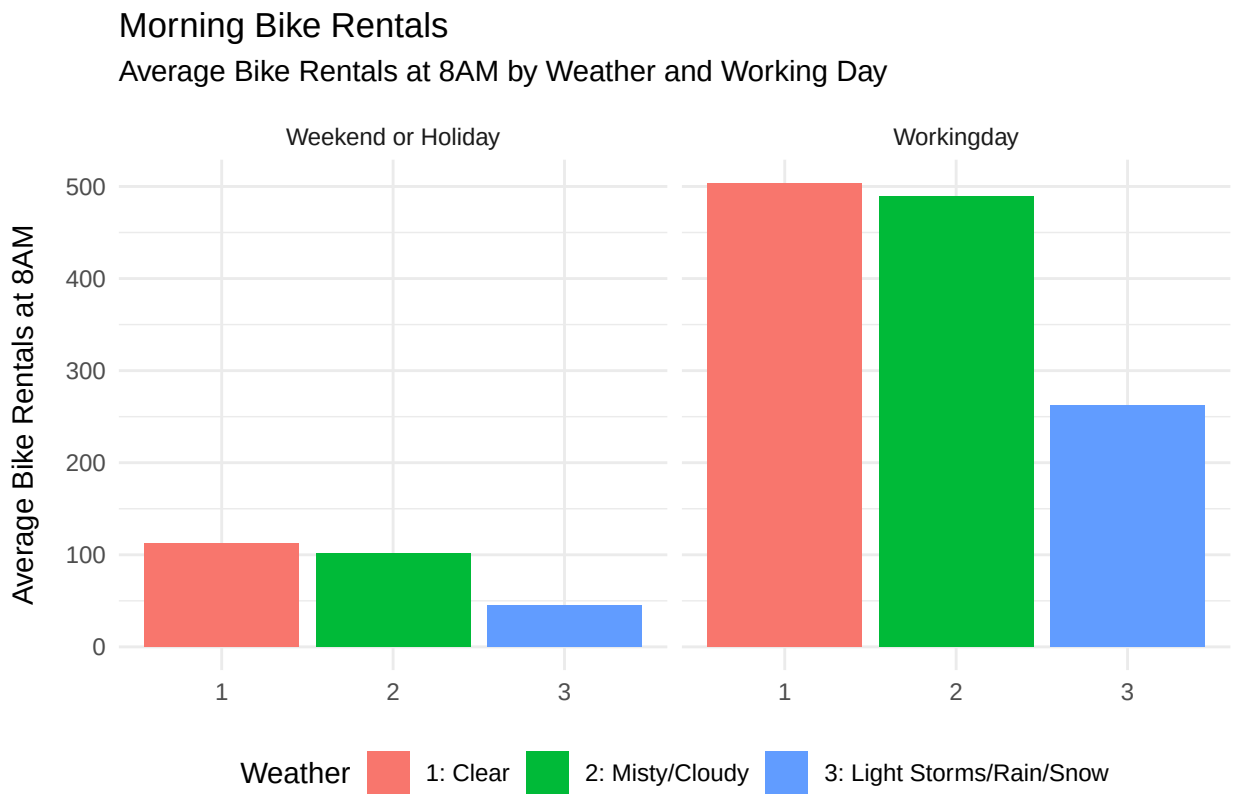

The plot shows the average bikes rented each hour in the Capital Bike Share network of recreational bicycles in Washington, DC in 2011 and 2012. The x-axis shows the hour of the day from 0 to 24, and the y-axis shows the average number of bikes rented for a particular hour of the day. From hour 17 (5:00PM) to hour 4 (4:00AM), bike rentals decline, indicating rentals generally decline during night-time. More notably, there are two distinct peaks in rental demand: a peak at 8:00AM and then a large peak at 5:00PM.

This implies that rental bikes are being used as a means for commuting to work in D.C. because these peak hours are the same times of day that people travel to and from work. (When I lived there in 2019, this was certainly true, with many people renting electric scooters to commute as well.)




This visualization adds a needed level of nuance to the initial graph. Since do not travel to work every day (weekends and holidays), it is necessary to show average rentals by workdays and non-workdays. The line graph on the left shows average bike rentals by hour for weekends and holidays. This tells a very different story from the line graph of workday rentals. On weekends and holidays, the 8:00AM and 5:00PM peaks from the initial graphs disappear, and an a single, less severe peak emerges. On weekends and holidays, people seem to be renting bikes mainly in the late morning and through the late afternoon.

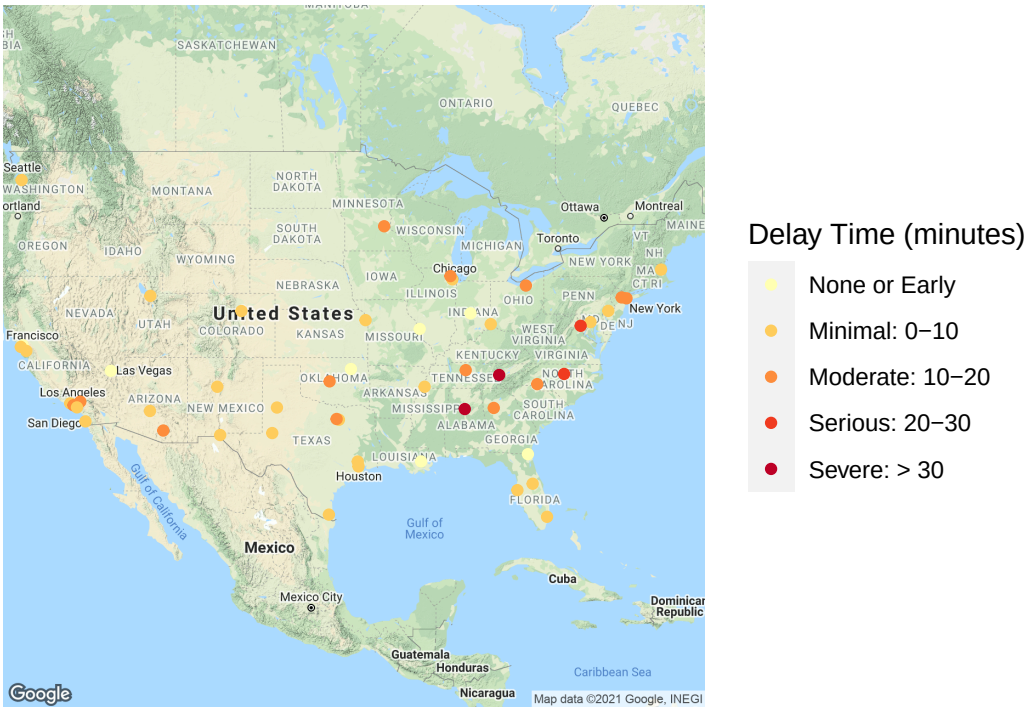
Overall, this pairing demonstrates that bike rental patterns are very different for workdays and non-workdays. On workdays, people use bikes to commute to and from work, whereas on non-workdays, people use bikes for recreation.




The final visualization adds a further level of analysis: weather. Since biking is an outdoor activity, it is expected to fall during bad weather. This visualization specifically compares average bike rentals **only** for 8:00AM for workdays and non-workdays and for different weather categories. The red bar represents very good, clear weather; the green bar is misty or cloudy weather; and the blue bar is lightly storming, snowing, or raining weather. The plots show that for both types of day, average bike rentals decrease as the weather gets worse.

Also of note is that bike rentals are generally much higher for workingdays. This is because, as seen in the previous graphics, people use bike rentals to go to work on workdays. On non-workdays, very few people rent bikes as early as 8:00AM. Generally, the timing of peak bike rentals depends on whether the bikes are for recreational or work-related use, and people will ride bikes less (for either purpose) as the weather becomes worse.

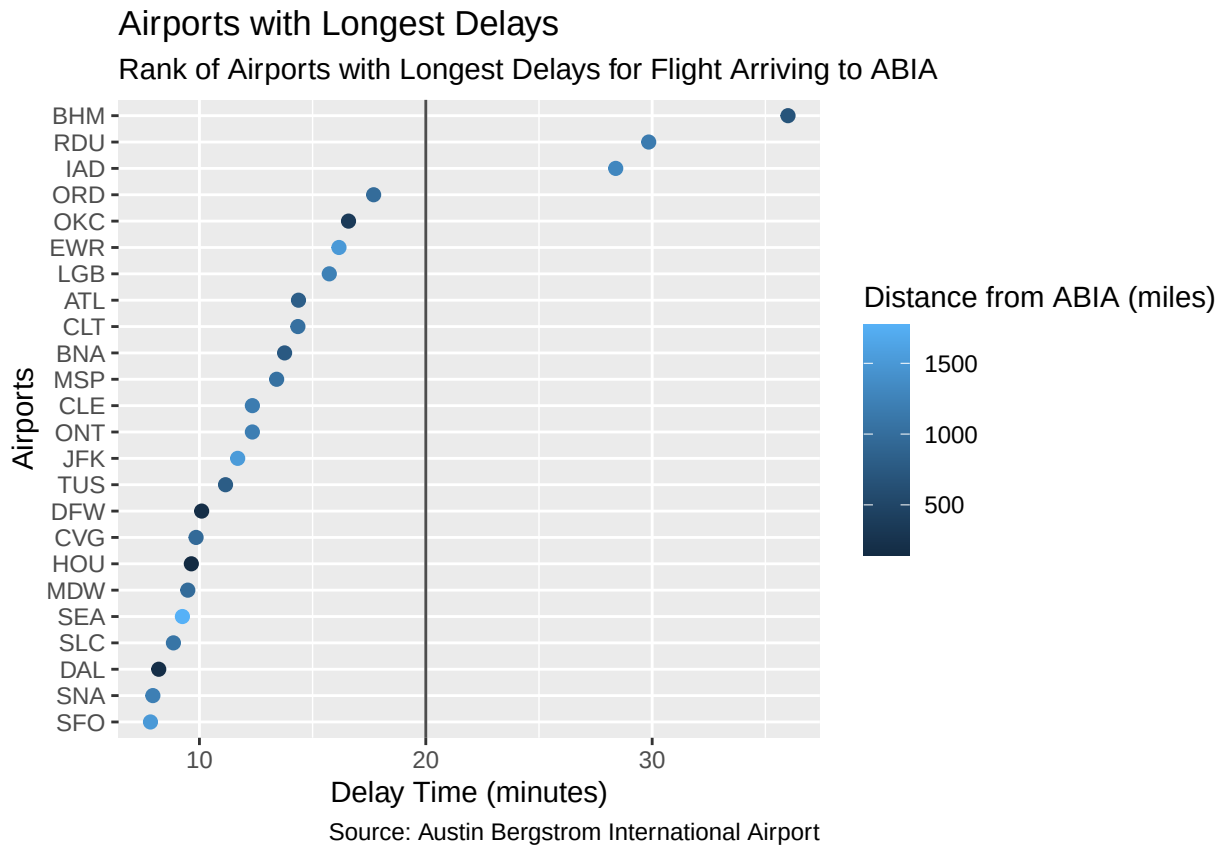
## 3) Delays for Flights to Austin Bergstrom International Airport




Source: Austin Bergstrom International Airport; Google

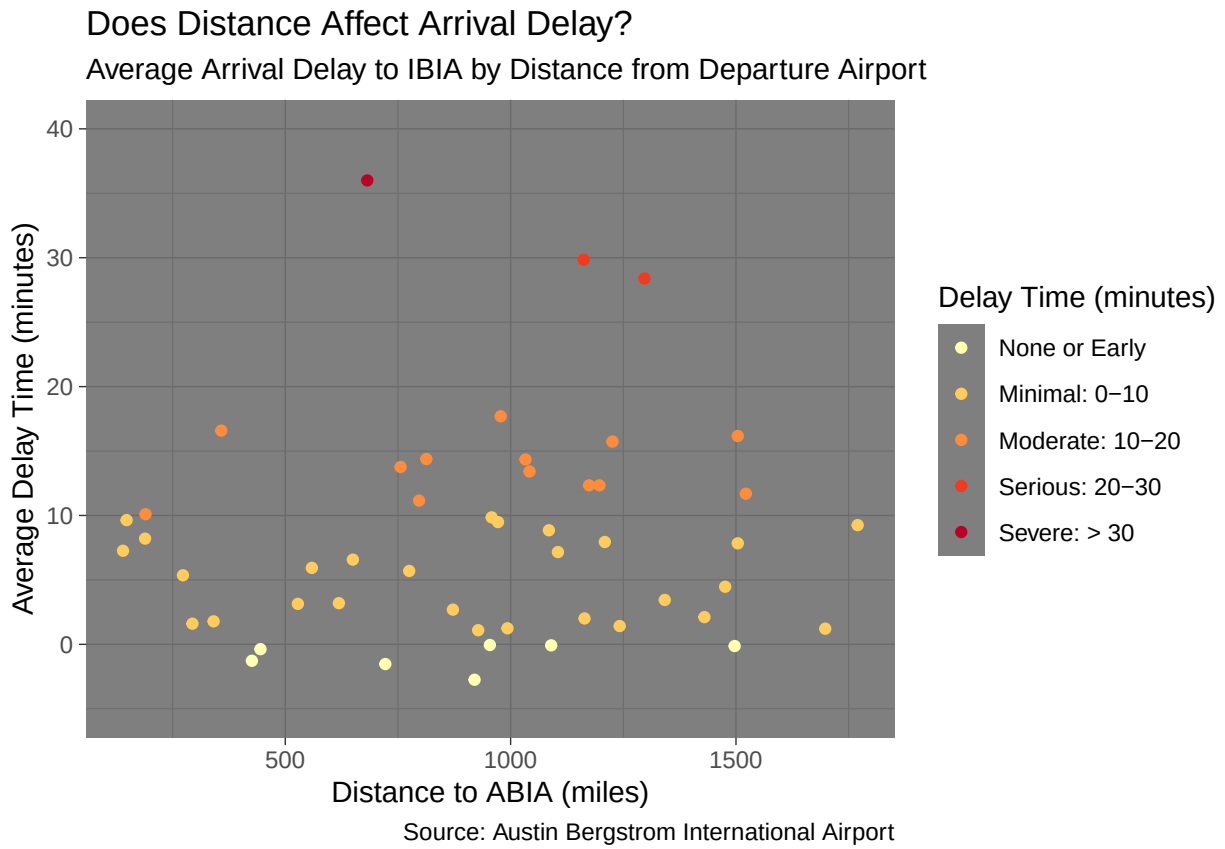
This shows a map of the United States, with colored points representing airports with flights leaving for Austin Bergstrom International airport (AUS or ABIA) in Austin, TX. The underlying data is the average flight delay for these inbound flights to Austin. Each point is an average delay time in minutes for all flights going from the airport on the map to AUS. Delay times have been ranked into discrete categories shown in the legend. As color goes from yellow to dark red, the average flight delay from an airport increases. For instance, the airport in Las Vegas has very low average delays for flights going to Austin, whereas airports nearby in Los Angeles have higher delays on average.

Overall, this map and the following visualizations and analysis attempt to show whether distance is a major driver of flight delay. As seen on the graph, distance does not seem to be a major cause of delay since there are airports with very high average delays (e.g. Alabama) that are much closer to Austin than airports with lower delays (e.g. Seattle).




This graphic gives the average arrival delay by Airport for flights inbound to Austin Bergstrom International Airport. The data show that most average arrival delay is below 20 minutes for most airports, which shows that most of the flights incoming to Austin are well-organized and without complication. Interestingly, there does not appear to be a relationship between arrival time and distance. For instance, the average delay for flights from Dallas (DFW) to Austin is greater than for flights coming from San Francisco (SFO) even though Dallas is only 189 miles away, while San Francisco is over 1,000 miles away. Furthermore, there appear to be outlier airports that suffer from much greater arrival delays–the three airports at the top of the list for which the average delay is well beyond 20 minutes. These airports are Dulles International Airport in Virginia (IAD), Raleigh-Durham International Airport in North Carolina (RDU), and Birmingham-Shuttlesworth Airport in Alabama (BHM).

Notably, BHM is by far the closest of these three outliers but has the greatest average delay (about 37 minutes per flight). Overall, the driver of delays seems to have far more to do with airport management or weather than distance. One could surmise that SFO is very well-managed, while BHM is poorly-managed. Alternatively, it is possible that the airline is responsible. Most nonstop flights from SFO to AUS are carried by United Airlines, whereas most nonstop flights from BHM are carried by American Airlines. Looking just at Dallas, there are two airports, DFW and DAL. Flights from DFW, which are all American Airlines flights have a 10 minute average delay, whereas flights from DAL, which are all Southwest Airlines flights have half the delay (5 minutes). So while the distance is effectively the same, the main difference driving the delay variation seems to be airline. Some airlines may be better at optimizing flight schedules than others, more effectively minimizing delays.

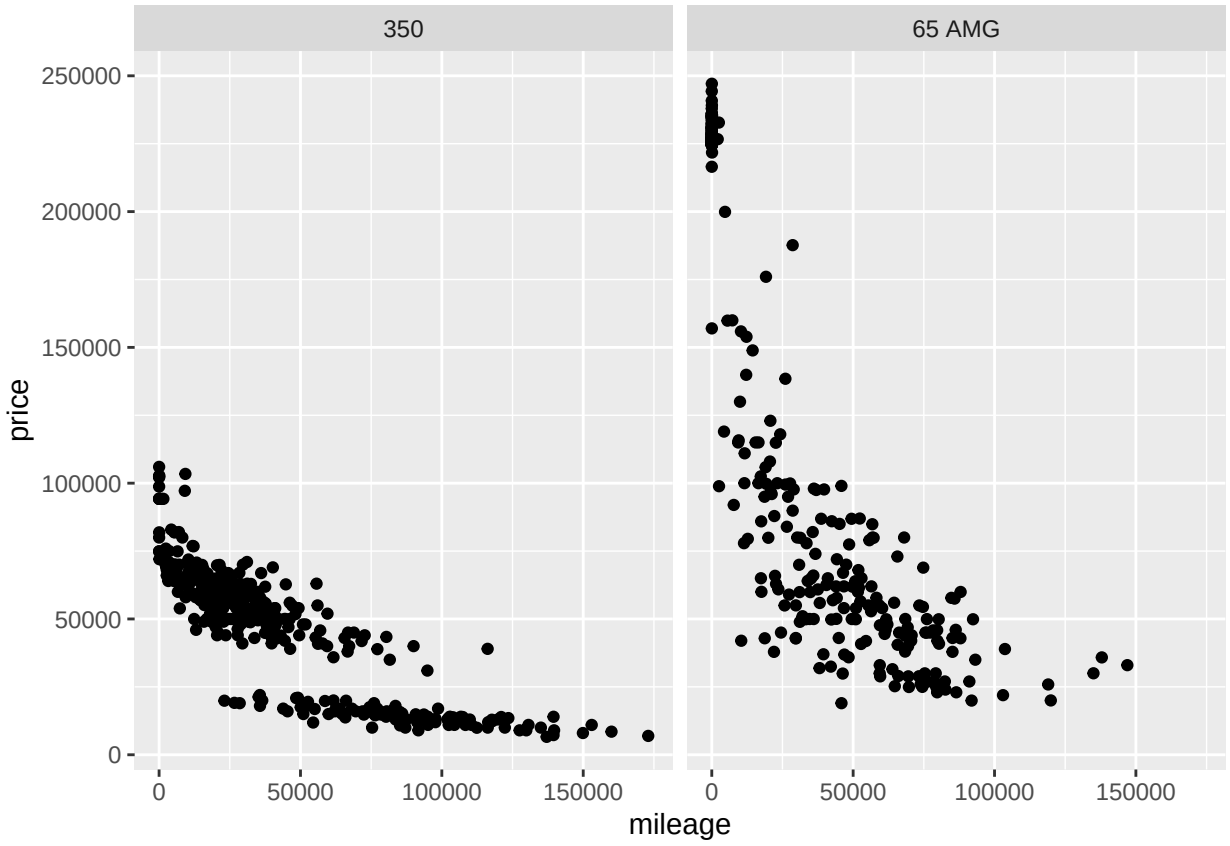



This scatterplot confirms that there is no relationship between flight distance and average delay, discussed in the previous graph. Most airports have only a minimal delay, and these airports are homogenously scattered by distance from 0 to 1,500 miles from Austin. This is strong evidence that distance is not a major driver of delay. The same homogenous characterization holds for airports with moderate delays, which are scattered evenly across the distance axis. In fact, the only airport with severe delays on average, Birmingham-Shuttlesworth Airport in Alabama (BHM), is below the median distance for all airports. Along with the previous graph, this is good confirmation that factors other than distance are the most important for delays, including airport and airline management and optimization.

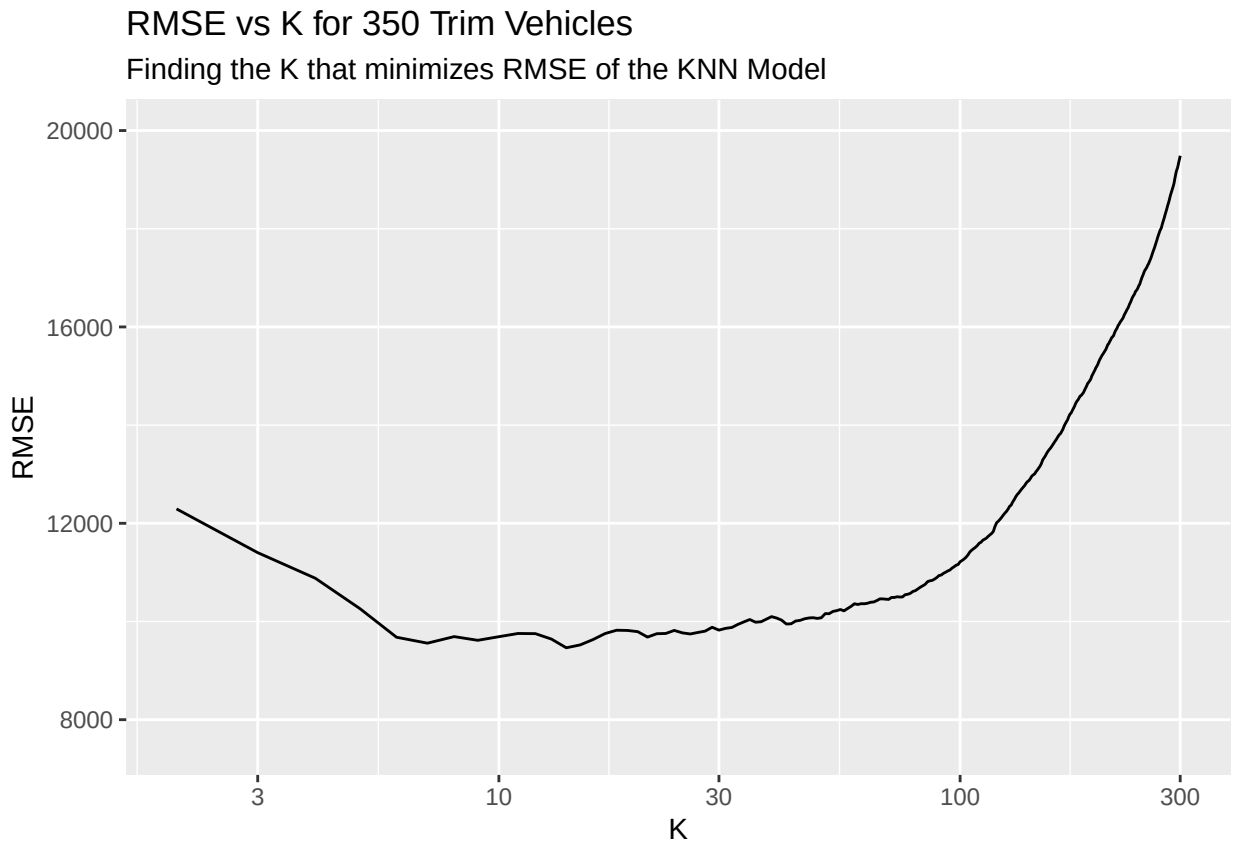
Since different airports are hubs for different airlines, carriers are not homogenously distributed across all flight paths incoming to Austin. As we saw by looking at Dallas, American Airlines seems to be worse at minimizing delays than Southwest Airlines for flights from Austin to Dallas. This is likely because the incoming flights to Dallas that will carry the flight to Austin are coming from different original cities in both cases. Possibly the origin city has more delays into Dallas for the American Airlines schedule than the Southwest Airlines schedule.

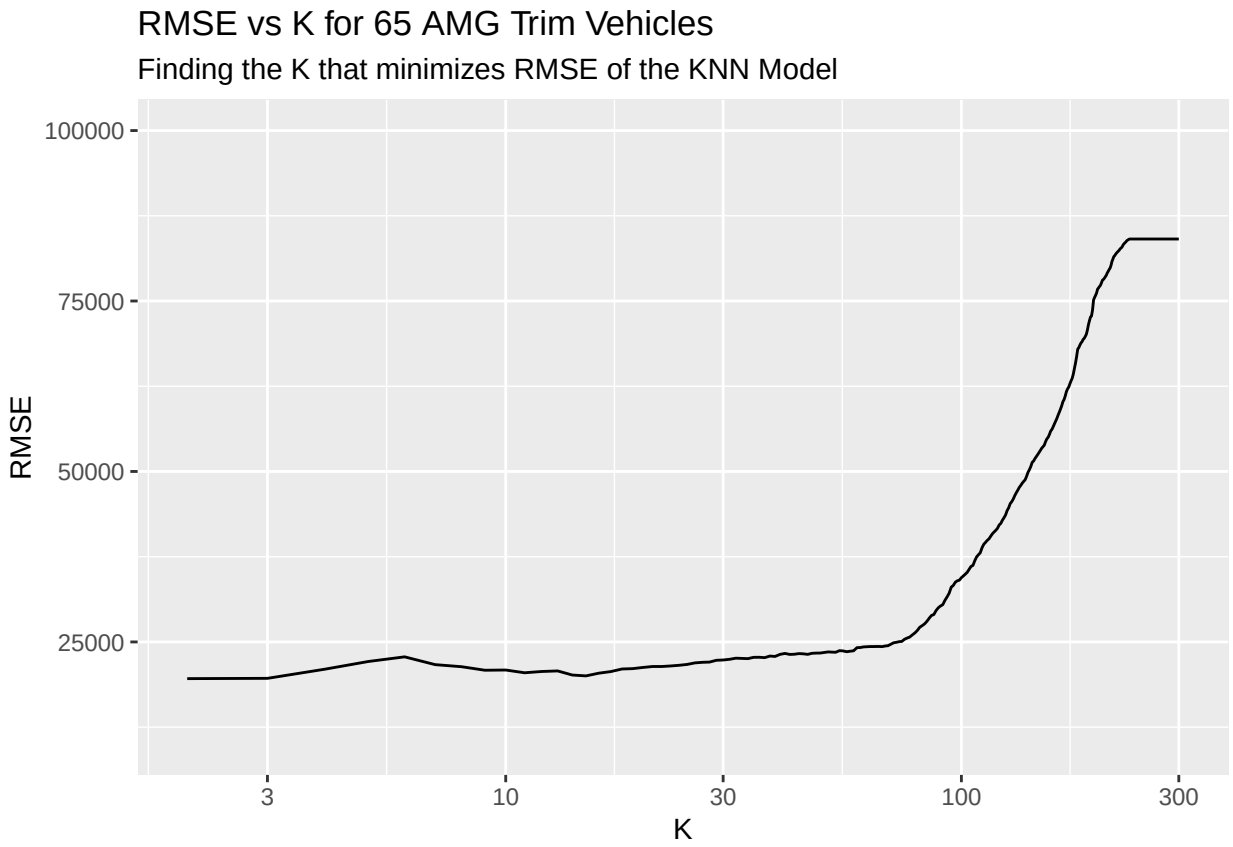
## 4) K-Nearest Neighbors Modeling

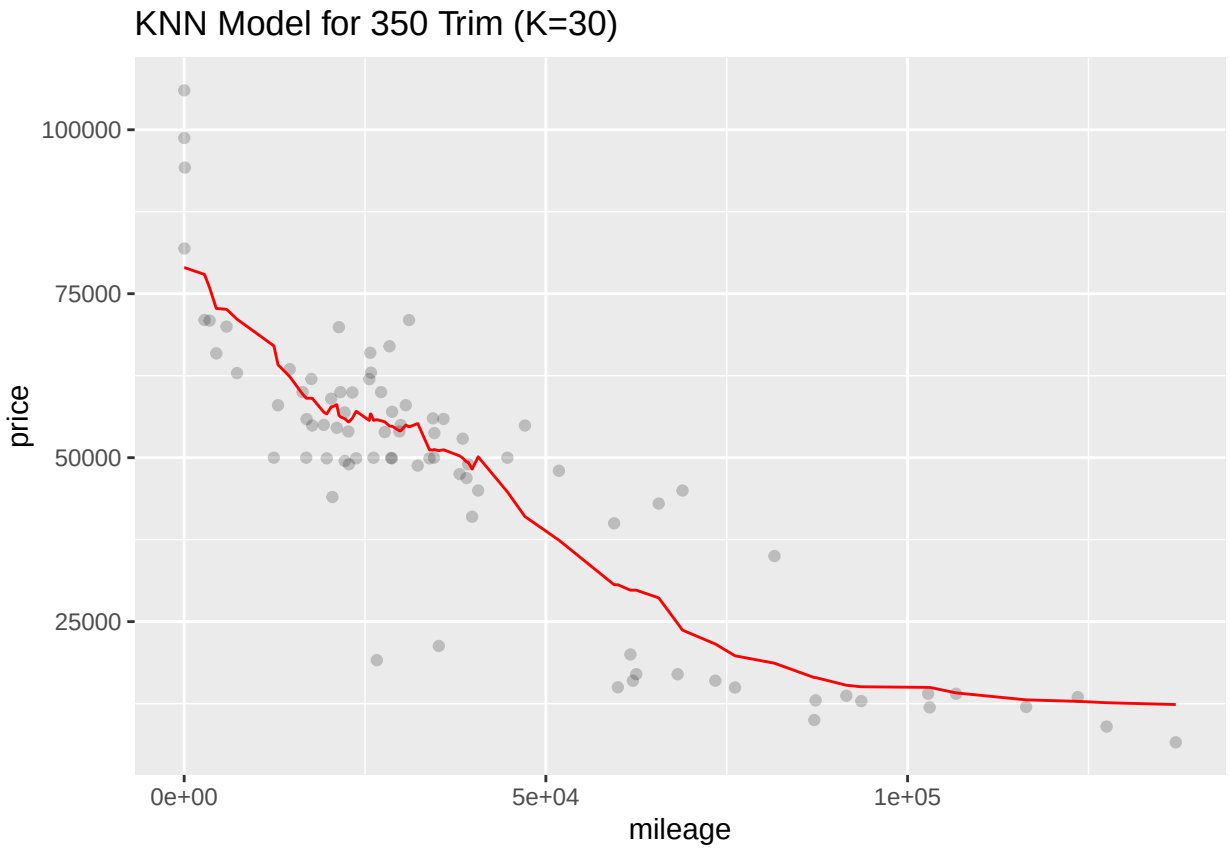
The following analysis fits a non-parametric model to car prices for Mercedes S-Class trim submodels (350 and 65 AMG) using the train-test split method. The modeling error is analyzed across a range of K-values for the K-Nearest Neighbor nonparametric model. The optimal K-value is chosen where the error (RMSE) is visually seen to be minimized across the possible K-values for the predictive model. In best-practice, a cross-validation method for choosing the optimal K-value should be used. In this case, only a single train/test split (80/20) is used, which means the RMSE error metric for the possible models is subject to high variance, making the optimal K-value more opaque.

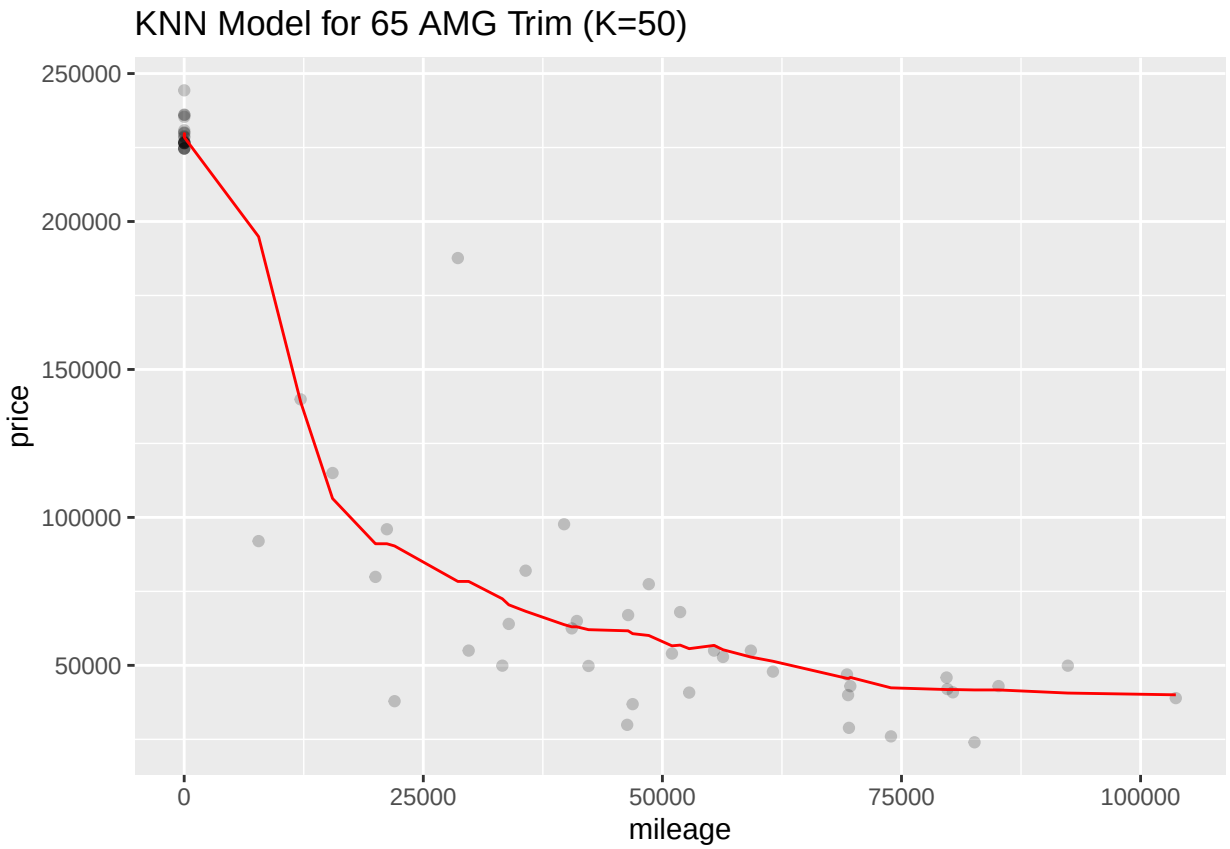



Initially, comparing the raw data of prices and mileage between the 350 trim and 65 AMG trim submodels of the S-Class vehicles, it is evident that the variability in price is much greater for 65 AMG trim cars. In fact, 65 AMG cars can have prices higher than $200,000, while 350 trim cars rarely go above $100,000. Clearly the 65 AMG trim submodel contains a wide variety of car types. The 350 trim submodels are visually far more concentrated. Interestingly, the 350 trim cars seem to fall into two unseen categories due to some latent variable causing two distinct clusters at different price levels.


RMSE vs K for 350 Trim Vehicles
Finding the K that minimizes RMSE of the KNN Model
20000
16000
12000
8000
RMSE
3
10
30
100
300
K


RMSE vs K for 65 AMG Trim Vehicles
Finding the K that minimizes RMSE of the KNN Model
100000
75000
50000
25000
RMSE
3
10
30
100
300
K


KNN Model for 350 Trim (K=30)
price
100000
75000
50000
25000
0e+00
5e+04
1e+05
mileage




The 65 AMG trim submodels seem to yield a larger optimal K than the 350 trim submodels. I suspect this is because the 65 trim data has more noise, as shown in the initial plots of original trim data. Since the 65 trim data has more noise, a lower K-value will be more sensitive to this variability and generate a high-variance prediction model as a result. The 350 trim data is more concentrated (low-variability), so a lower K-value does not risk picking up noise. The lower K-value for the 350 K-Nearest Neighbors model means the model is less smooth than the 65 AMG K-Nearest Neighbors model. This is evident when comparing the smoothness of the red prediction lines in the two plots.

Choosing the optimal K-value in each case remains tricky, however, because the optimal K can vary with the training-test split samples. Each time a new split is conducted for each trim type above, producing a graph of RMSE vs K-value, a slightly different low-point emerges (lowest RMSE across K-values). After running many splits, I generally picked up that the optimal K tended to be lower for the 350 trim (roughly a K of 10 to 50), whereas the optimal K tended to be higher for the 65 trim (roughly a K of 30 to 100).

Therefore, I have chosen an optimal K of 30 for the 350 trim data and an optimal K of 50 for the 65 trim data. The variability I saw in the optimal K's is due to the variance in the RMSE value. Each single train-test split produces a different RMSE for a particular K, so there is a need to understand the RMSE variance when finding the RMSE-minimizing K-value for KNN. Because I only used a single training-test split, I was unable to reduce this variance in RMSE. This is why Cross-Validation so importance. In K-Fold Cross-Validation, you divide the data into non-overlapping groups and generate a cross-validated error rate that reduces RMSE variance. This gives you a more precise RMSE estimate that you can use to choose the RMSE-minimizing K-value during KNN. For this reason, Cross-Validation is best practice when using train-test split models and, if applied here, would allow me to better decide on the optimal K for each set of trim data because I would reduce the variance of the RMSE.