# ECO 395 Homework 1

Yefu Chen

2/3/2021

## QUESTION 1 Data visualization: gas prices

People have a lot of pet theories about what explains the variation in prices between gas stations. Here are several such theories below. Which of these theories seem plausible, and which are unsupported by data? Take each theory one by one and assess the evidence for or against the theory using the suggested plot in parentheses.

### A) Gas stations charge more if they lack direct competition in sight (boxplot).

According to the claim, the stations with competitors should have a lower price compared to those without. The Figure 1-1 is consistent with this claim. The mean gasoline price of stations with competitors does be lower than those without competitors.

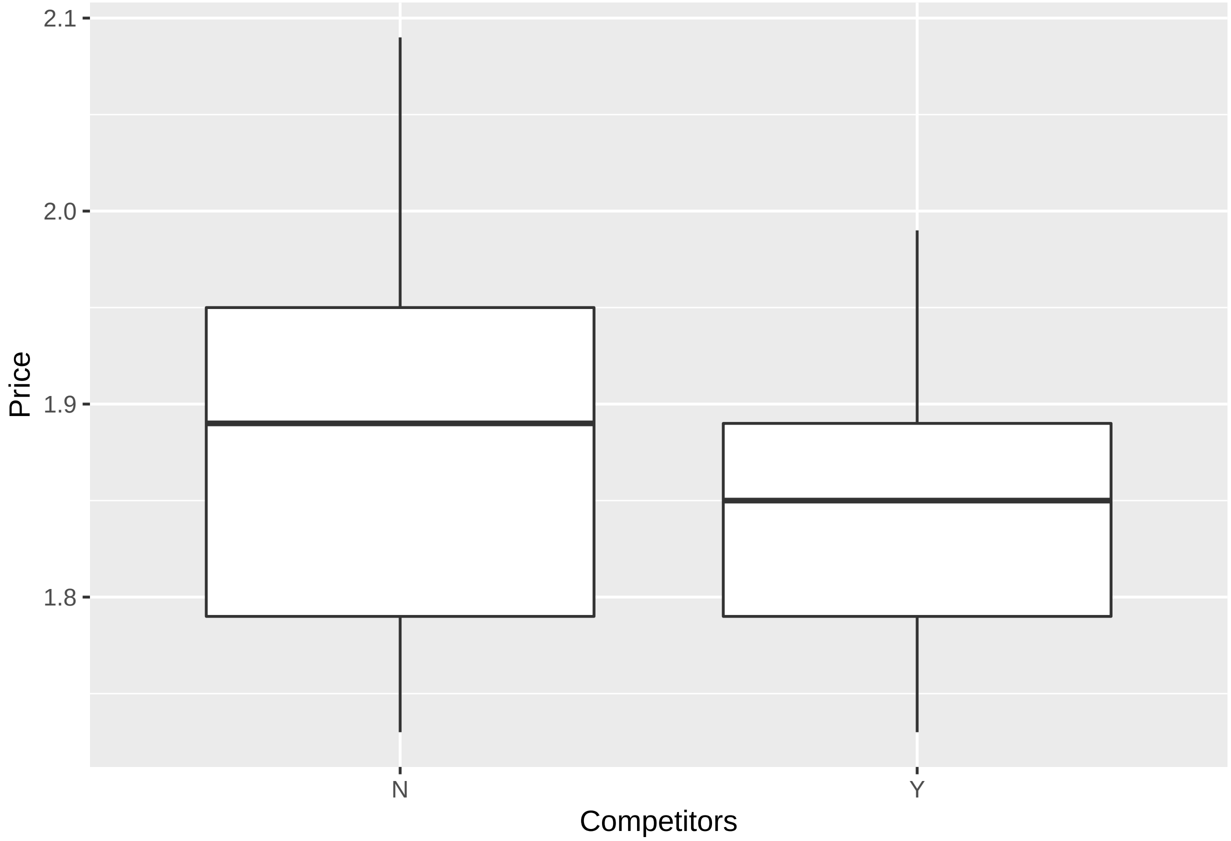


Figure 1-1.

**B) The richer the area, the higher the gas price (scatter plot).**

According to the claim, the price of stations in areas with higher income should be higher. The Figure 1-2 is consistent with this claim. With income rising, the price of gasoline is also increasing.

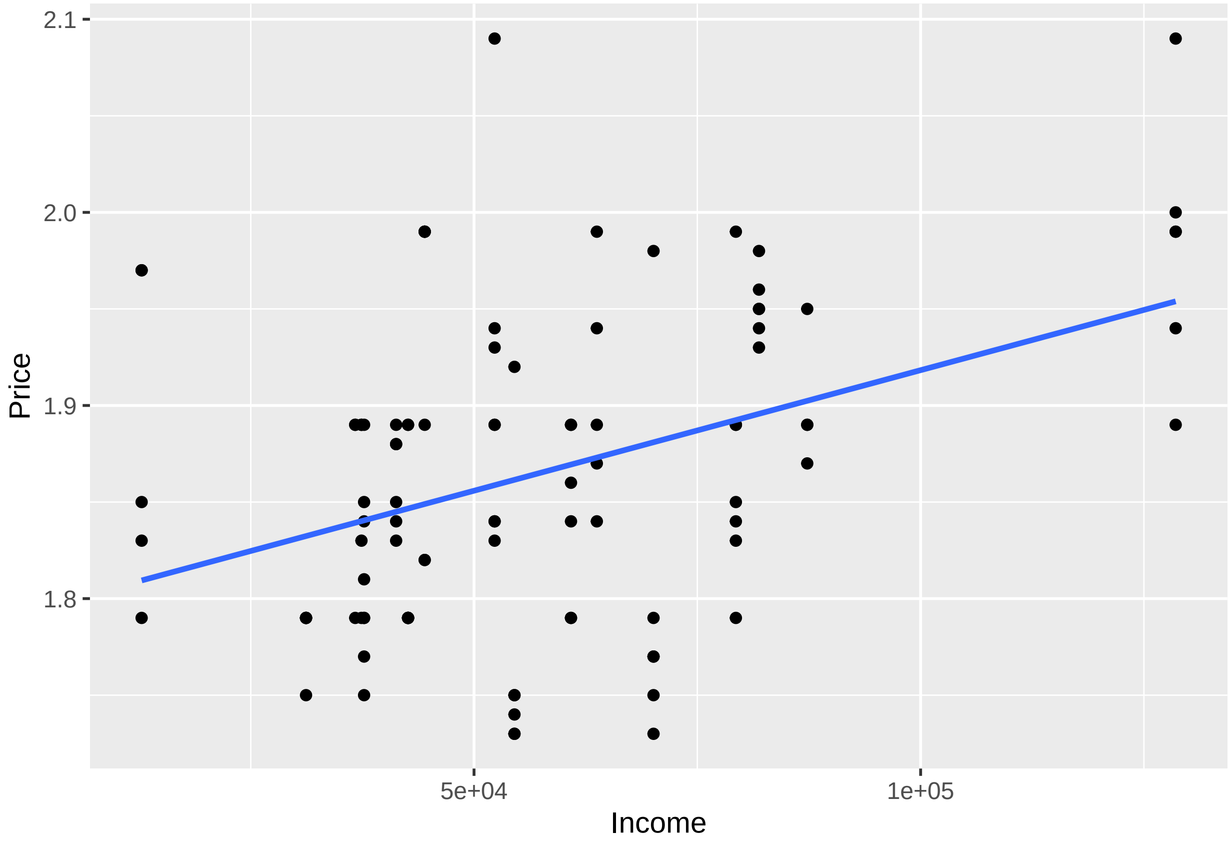


Figure 1-2.

**C) Shell charges more than other brands (bar plot).**

According to the claim, the price of Shell gasoline should be the most expensive. The Figure 1-3 is inconsistent with this claim. The price of Shell gasoline is high but lower than “other brands”.

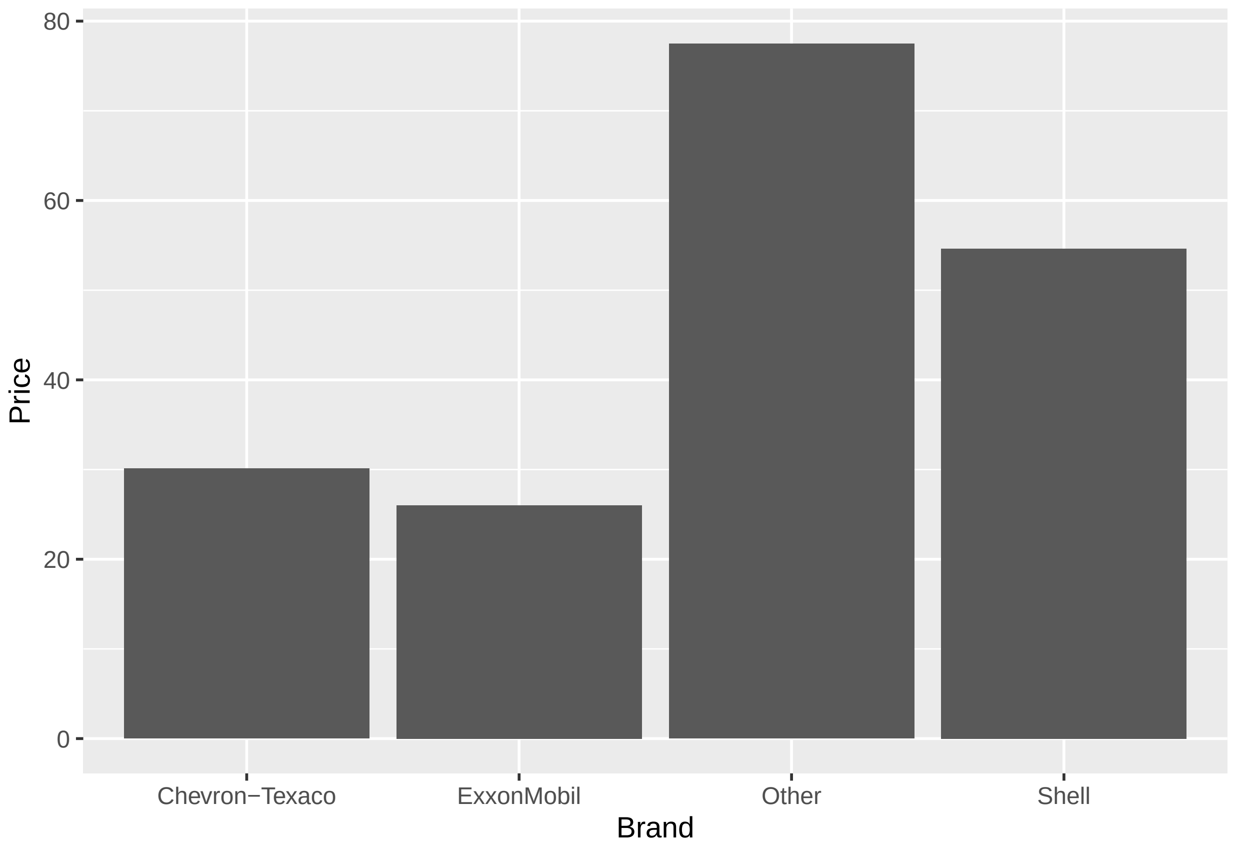
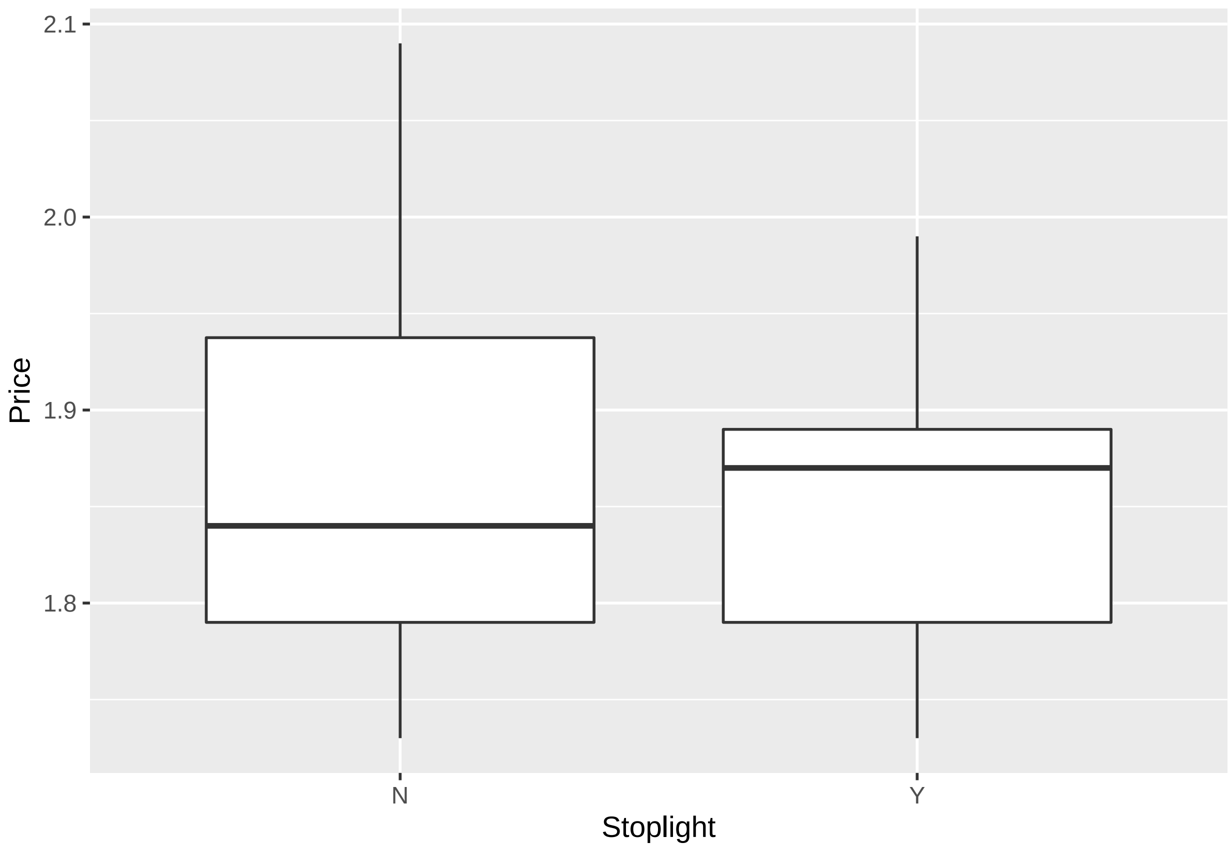
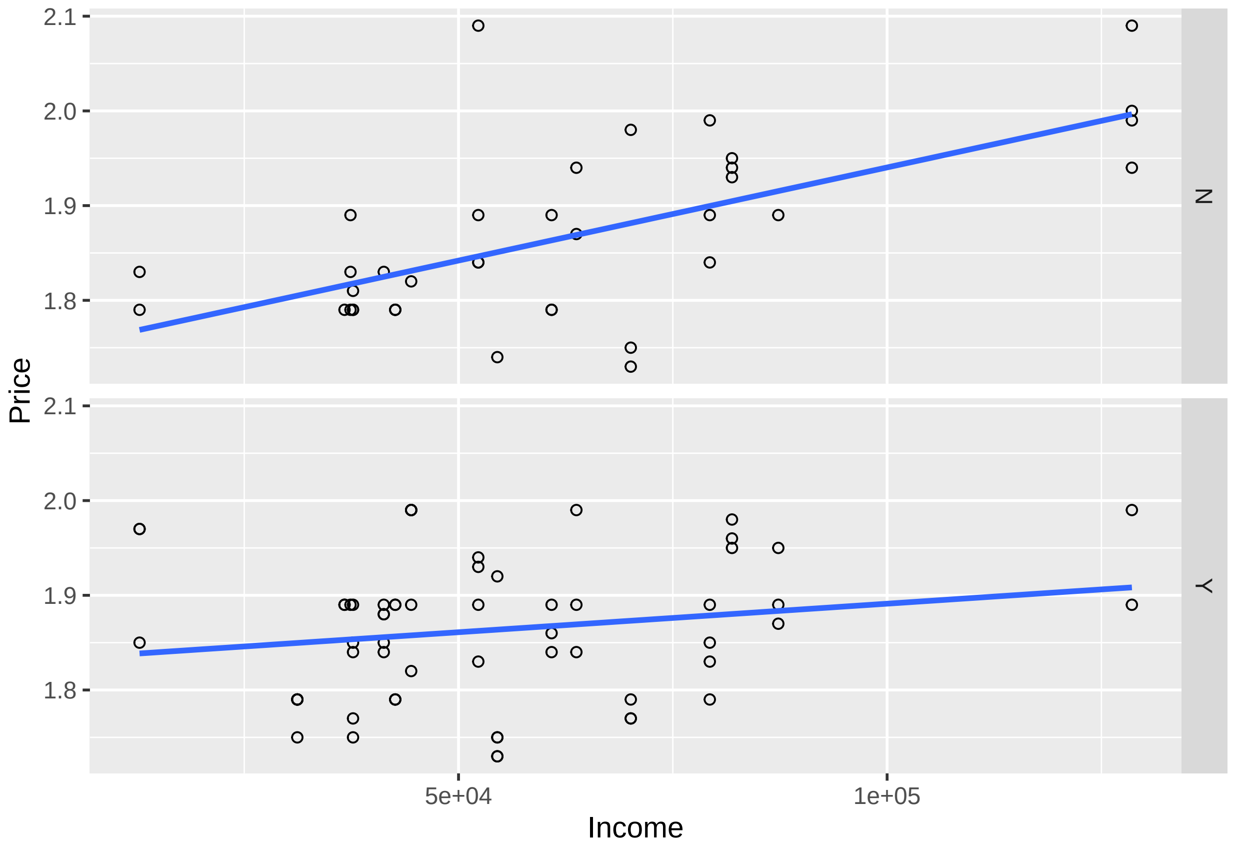


Figure 1-3.

**D) Gas stations at stoplights charge more (faceted histogram).**

According to the claim, gas stations at stoplights charge more. The first figure supports this claim. The average price of gas stations at stoplights is more than those far away. However, if we control the effect of income, shown as the second figure, the effect of neighborhood income on gasoline price is larger in gas stations near stoplights. Hence, this claim needs further analysis to be supported or rejected.

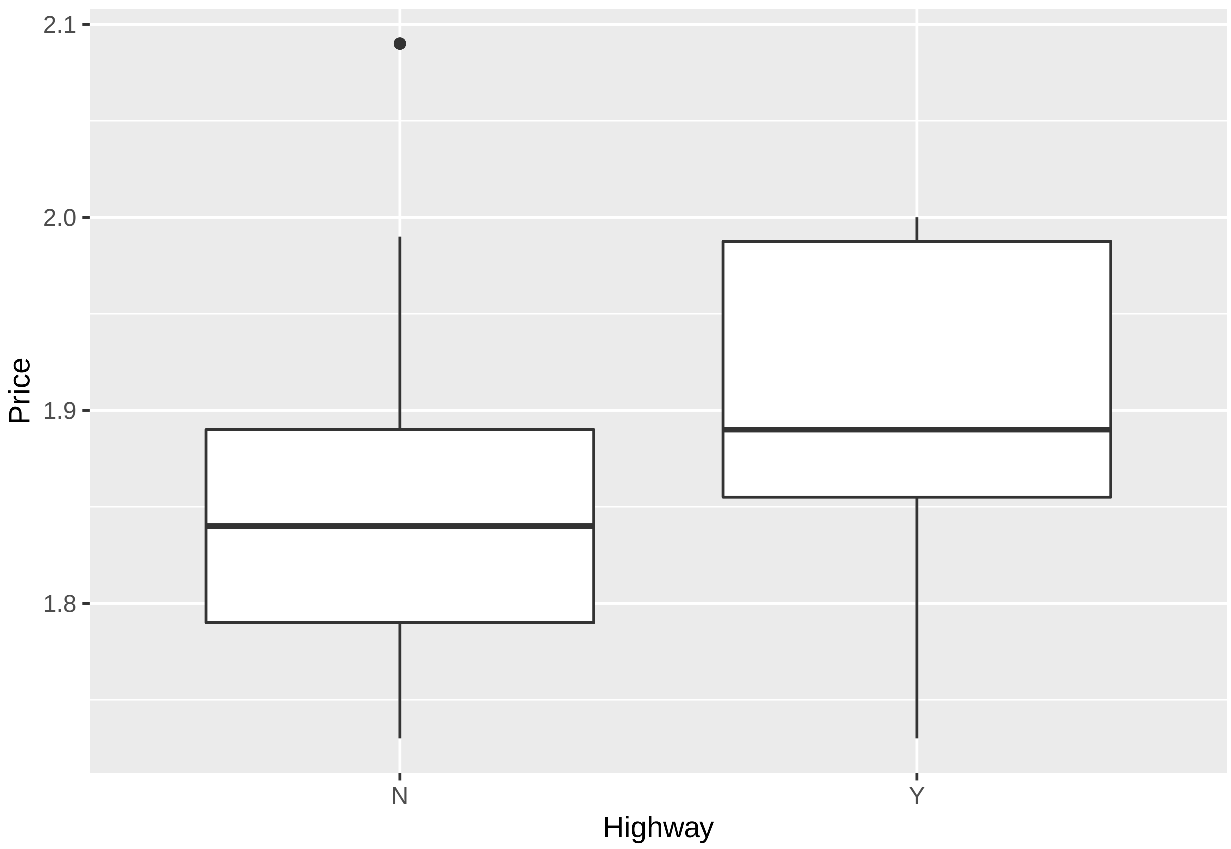

2.1
2.0
1.9
1.8
Price
N
Y
Stoplight

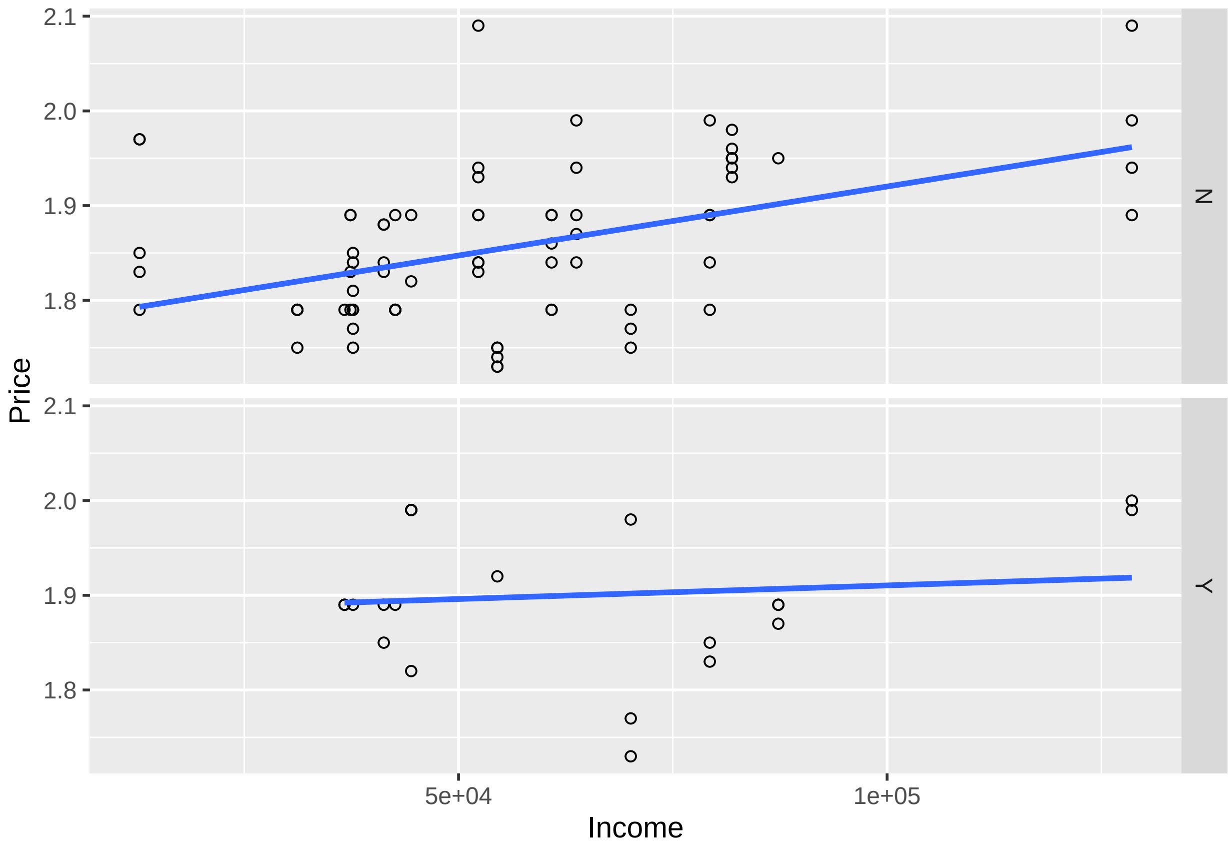


Figures 1-4.

**E) Gas stations with direct highway access charge more (your choice of plot).**

According to the claim, gas stations with direct highway access charge more. The first figure supports this claim. The average price of gas stations near highway access is more than those far away. However, if we control the effect of income, shown as the second figure, the effect of neighborhood income on gasoline price is larger in gas stations near highway entrances. Hence, this claim needs further analysis to be supported or rejected.


2.1
2.0
1.9
1.8
Price
N
Y
Highway



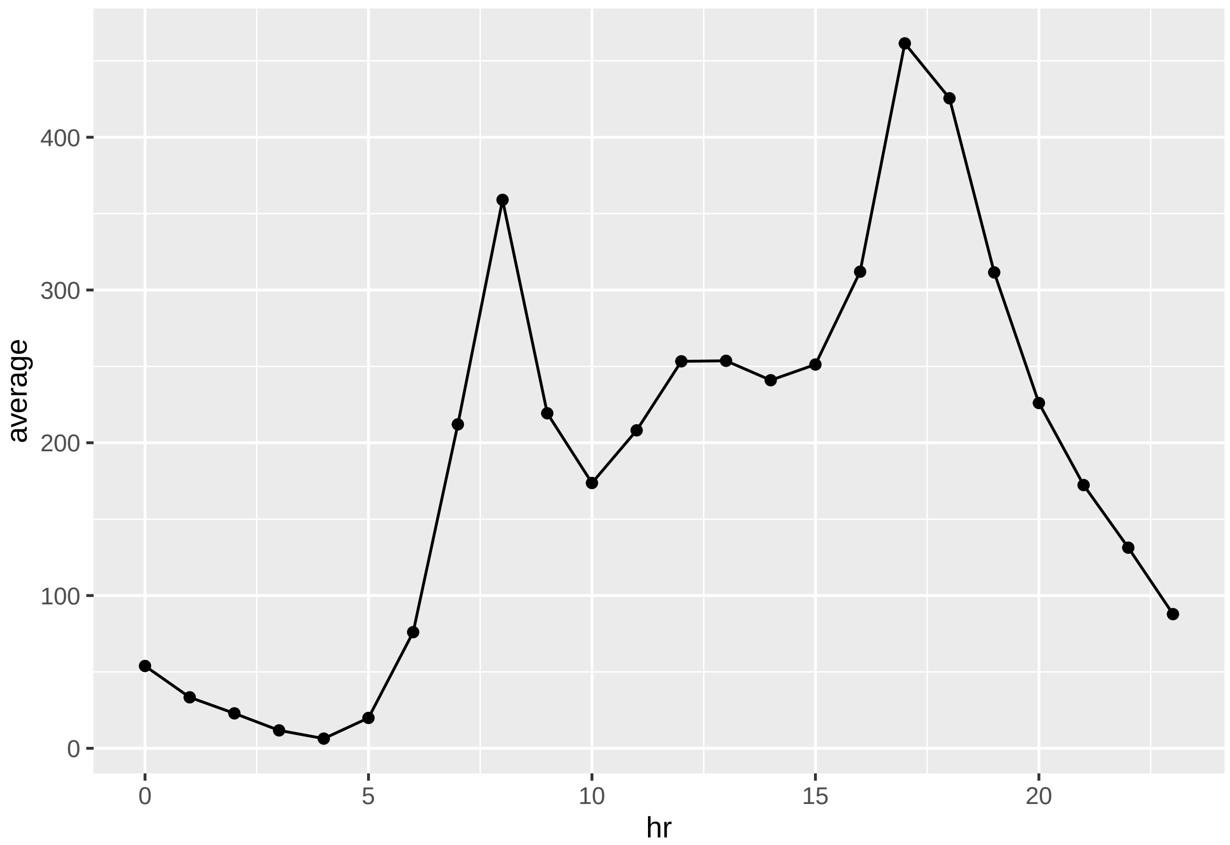
Figures 1-5.

## QUESTION 2 Data visualization: a bike share network

Your task in this problem is to prepare three figures.

### Plot A: a line graph showing average bike rentals (total) versus hour of the day (hr).

The following Figures 2-1 presents the average bike rentals by hours. It indicates that there are two peak hours. The morning peak hour happens at 8 am and the afternoon peak hour happens at 5 pm. Most of the bike rentals gather between 6 am to 8 pm.

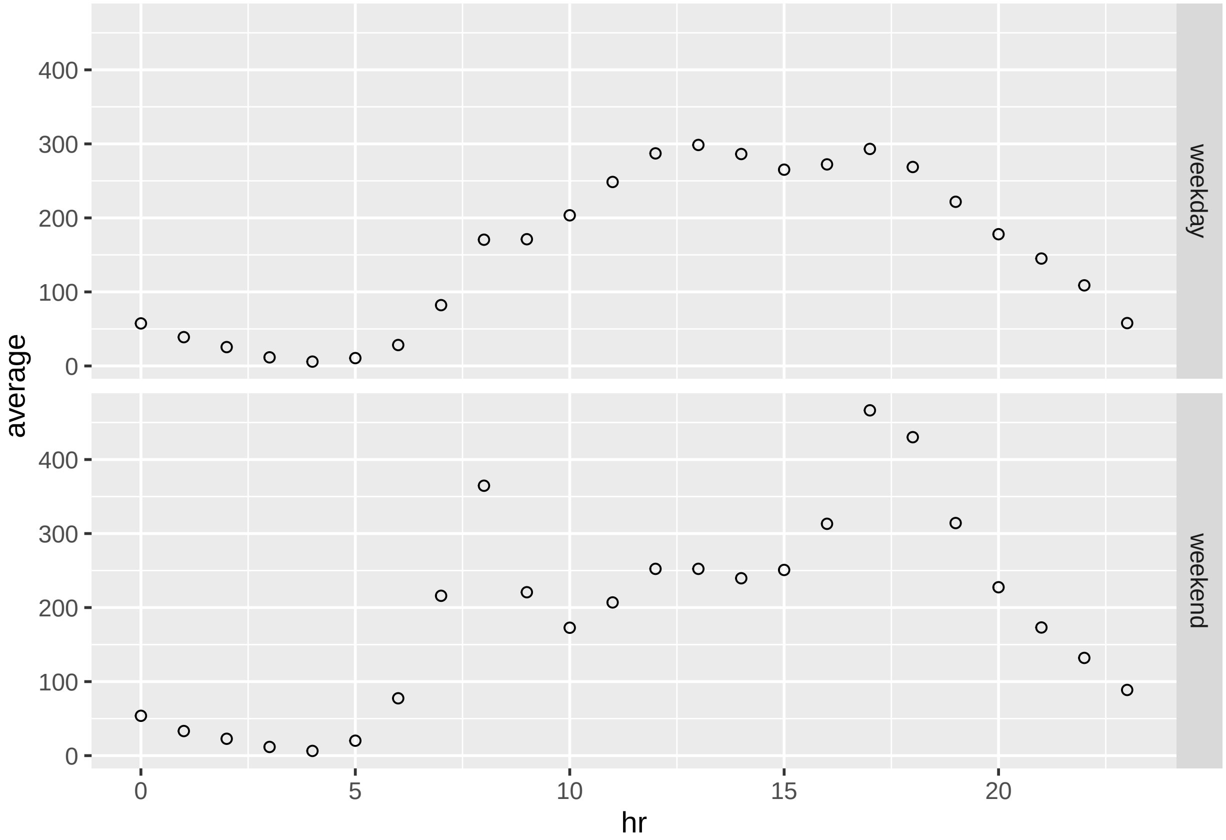


Figures 2-1.

**Plot B: a faceted line graph showing average bike rentals versus hour of the day, faceted according to whether it is a working day (workingday).**

The following Figures 2-2 shows average bike rentals versus the hour of the day, faceted according to whether it is a working day. It indicates that the "peak" hours are not obvious on weekdays while significant during weekends.

There are two peak hours during weeknds. The morning peak hour happens at 8 am and the afternoon peak hour happens at 5 pm. Most of the bike rentals gather between 6 am to 8 pm. Also, it seems that there can be more usages during the weekend, which indicates that riders may use the shared bikes for entertainment or recreation instead of commuting.

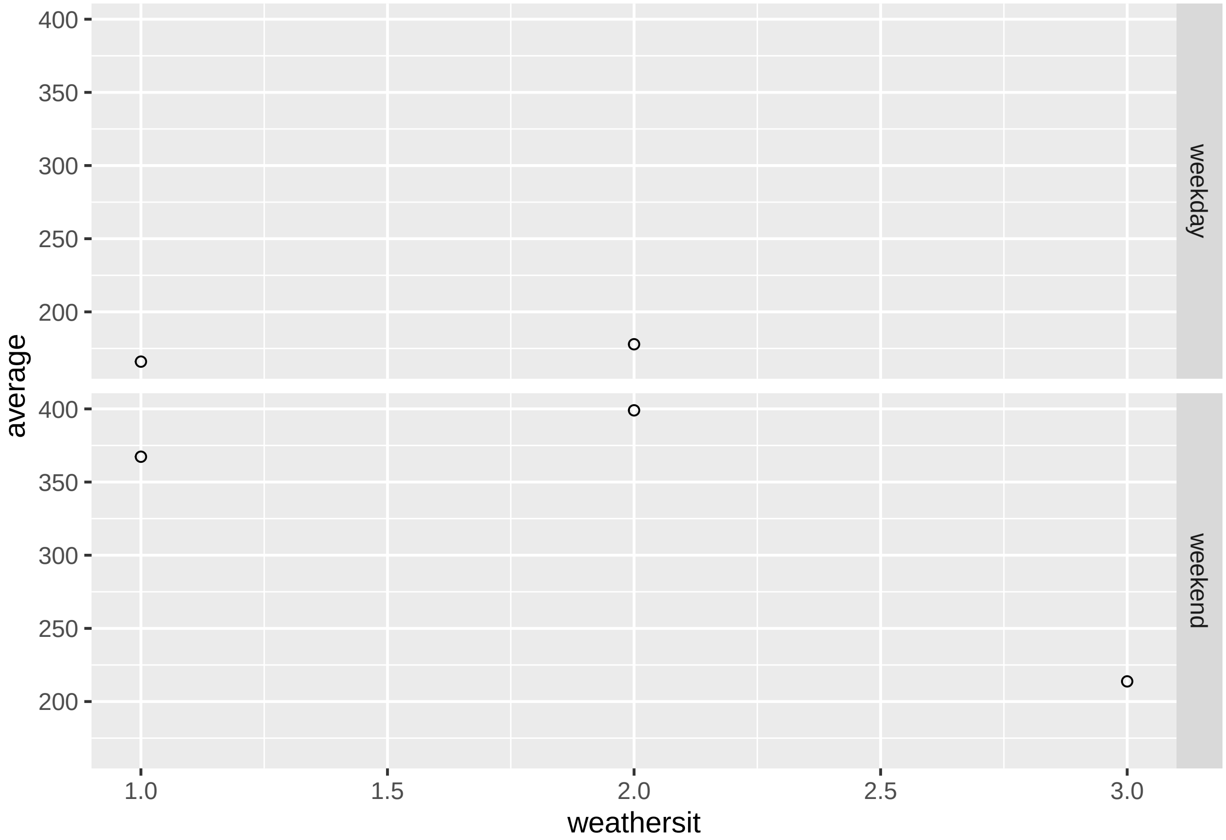


Figures 2-2.

**Plot C: a faceted bar plot showing average ridership during the 8 AM hour by weather situation code (weathersit), faceted according to whether it is a working day or not. Note: remember you can focus on a specific subset of rows of a data set using filter.**

Interestingly, it seems that there are only two weather conditions during the weekdays and 3 conditions during weekends. During the weekdays, the weather conditions include type 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) and type 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist). During the weekends, the weather conditions not only include conditions during weekdays but also type 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

Also, considering the average rentals, in both conditions during weekdays, the rentals are small, while it seems that bad cloudy cannot stop riders to ride a bicycle at 8 am during the weekend, but snow and rain can.



Figures 2-3.

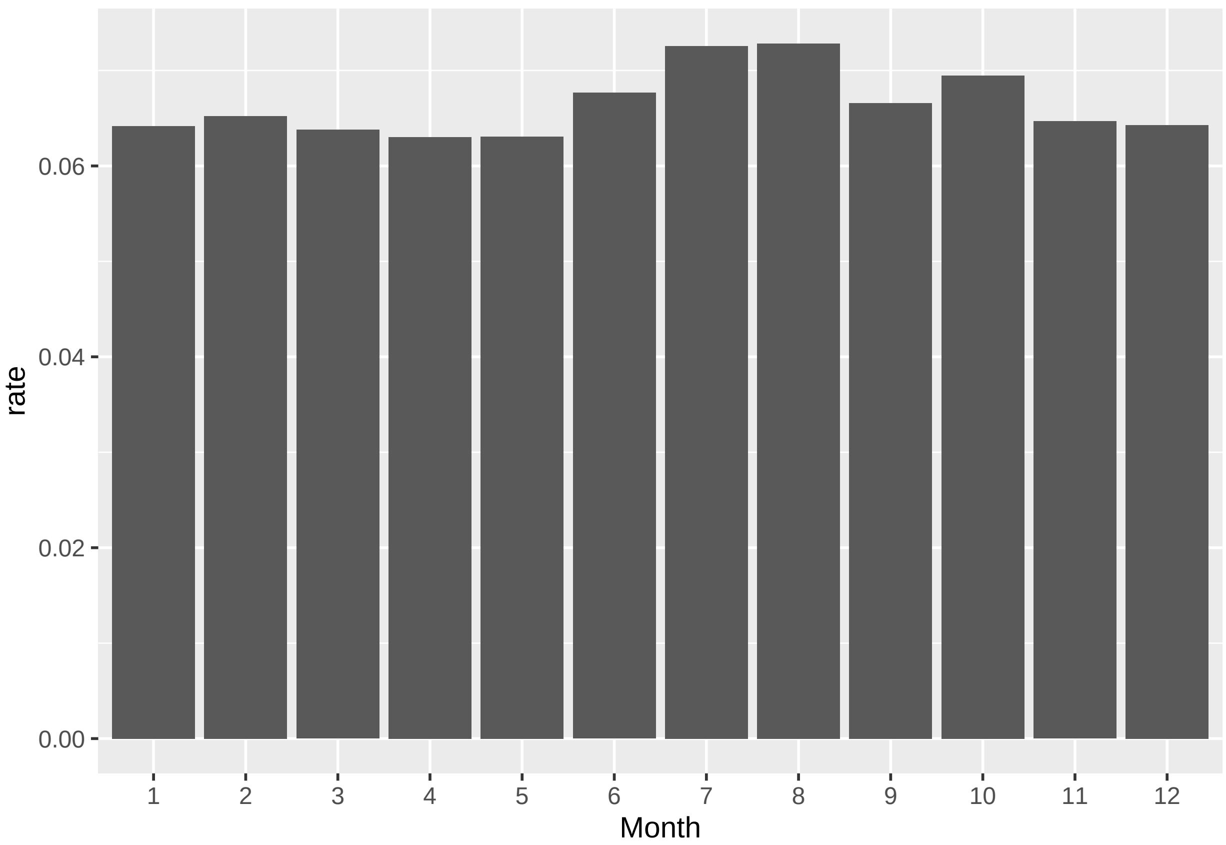
## QUESTION 3 Data visualization: flights at ABIA

Your task is to create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. You should annotate your figure(s), of course, but strive to make them as easy to understand as possible at a quick glance. (A single figure shouldn't need many, many paragraphs to convey its meaning.)

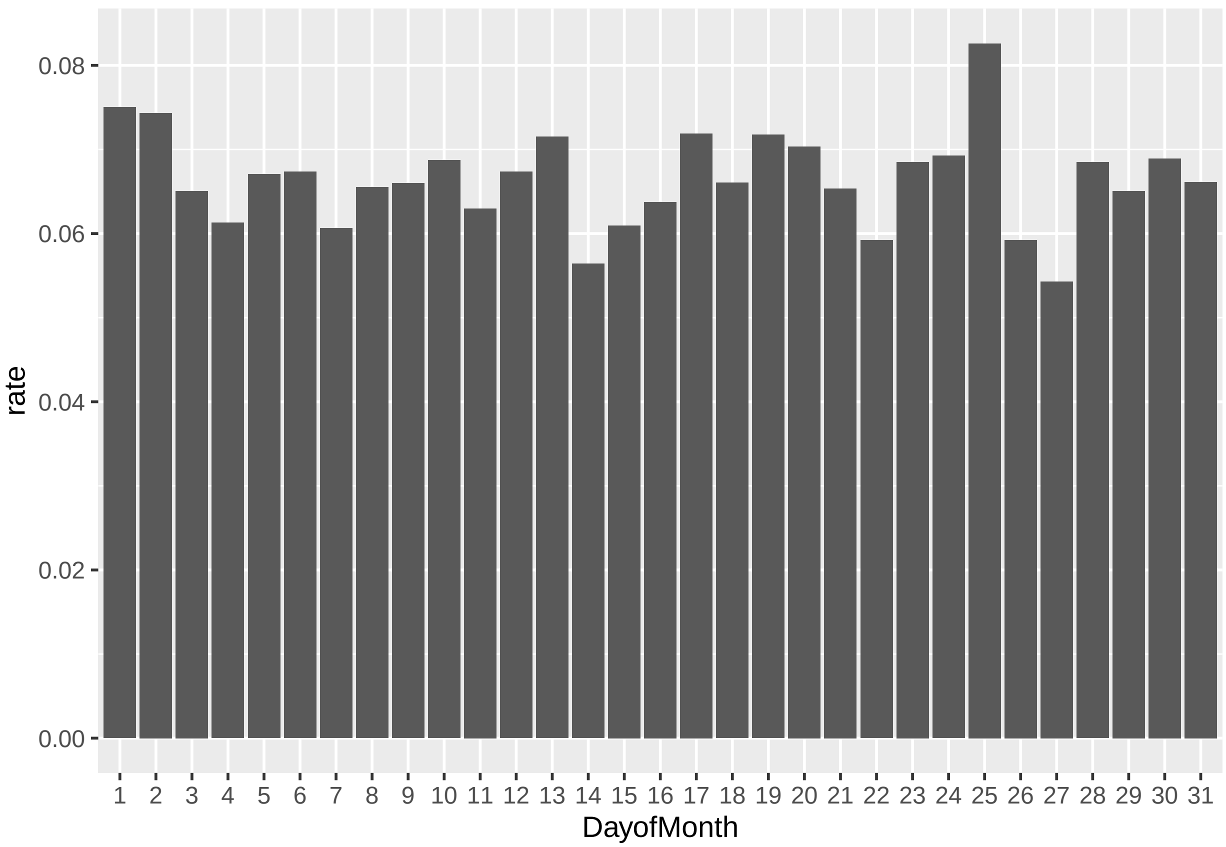
**My reserach question is when is the best to fly from/to austin (with least delay).**

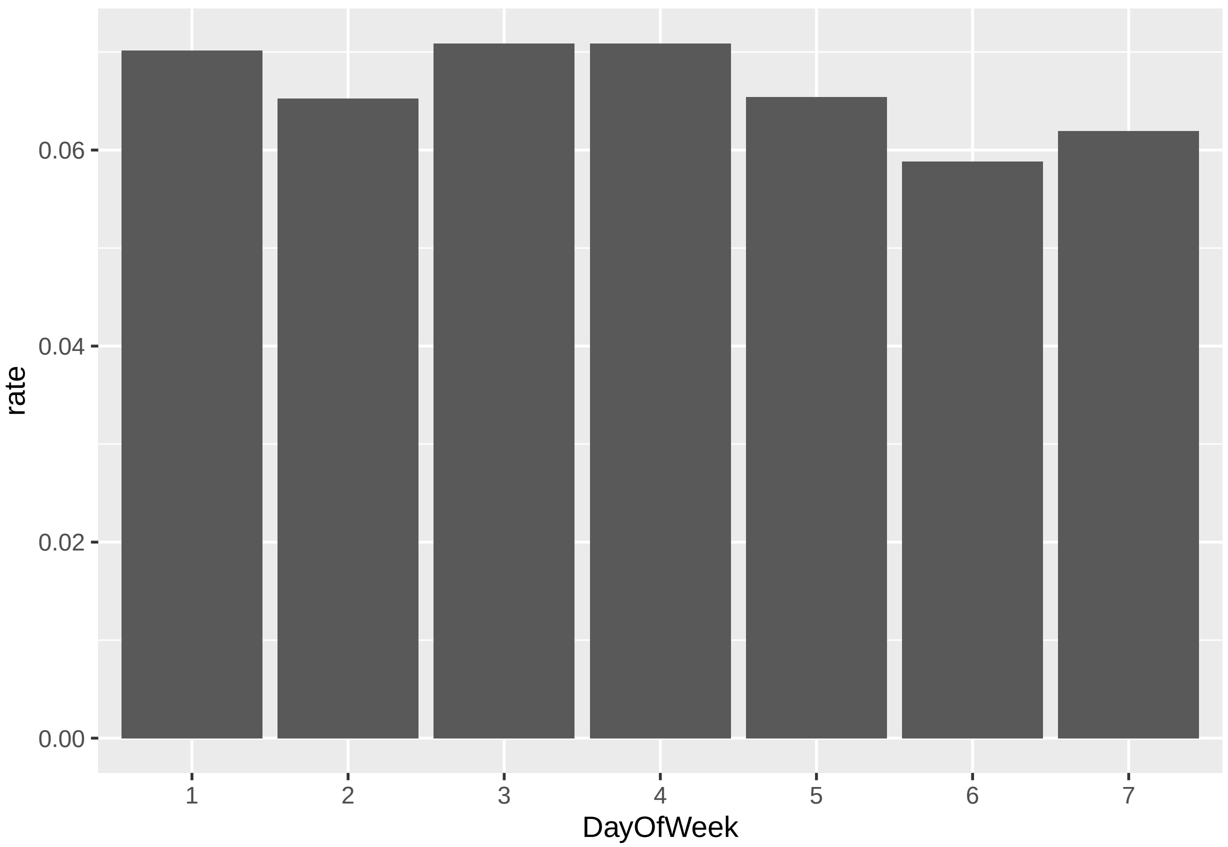
We cannot simply sum the number of delayed flight to find the "worst" day since the number of flight by days is various. We have to nominate the number of delayed fight as the rates of delay.

The following Figures 3-1 present the rate of delay by months, days, and weekdays. Focusing on the first figure, we may refuse to vist Austin airport in summer break, including June, July, and Auguts. Every 25th of months is also the date we want to avoid to go to the Austin airport. Monday, Wednesday, and Thursday are three dates when rates of delay are high.



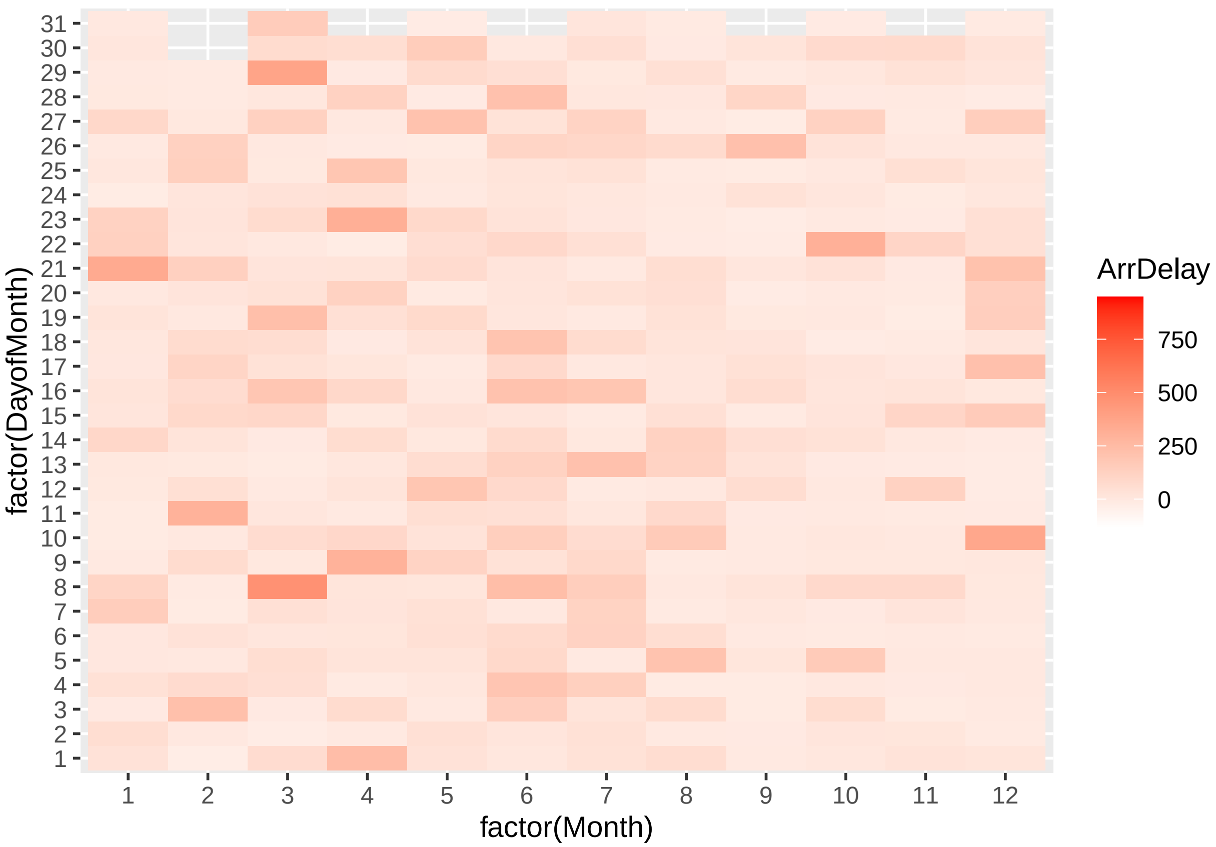
0.06
0.04
0.02
0.00
rate
1
2
3
4
5
6
7
8
9
10
11
12
Month

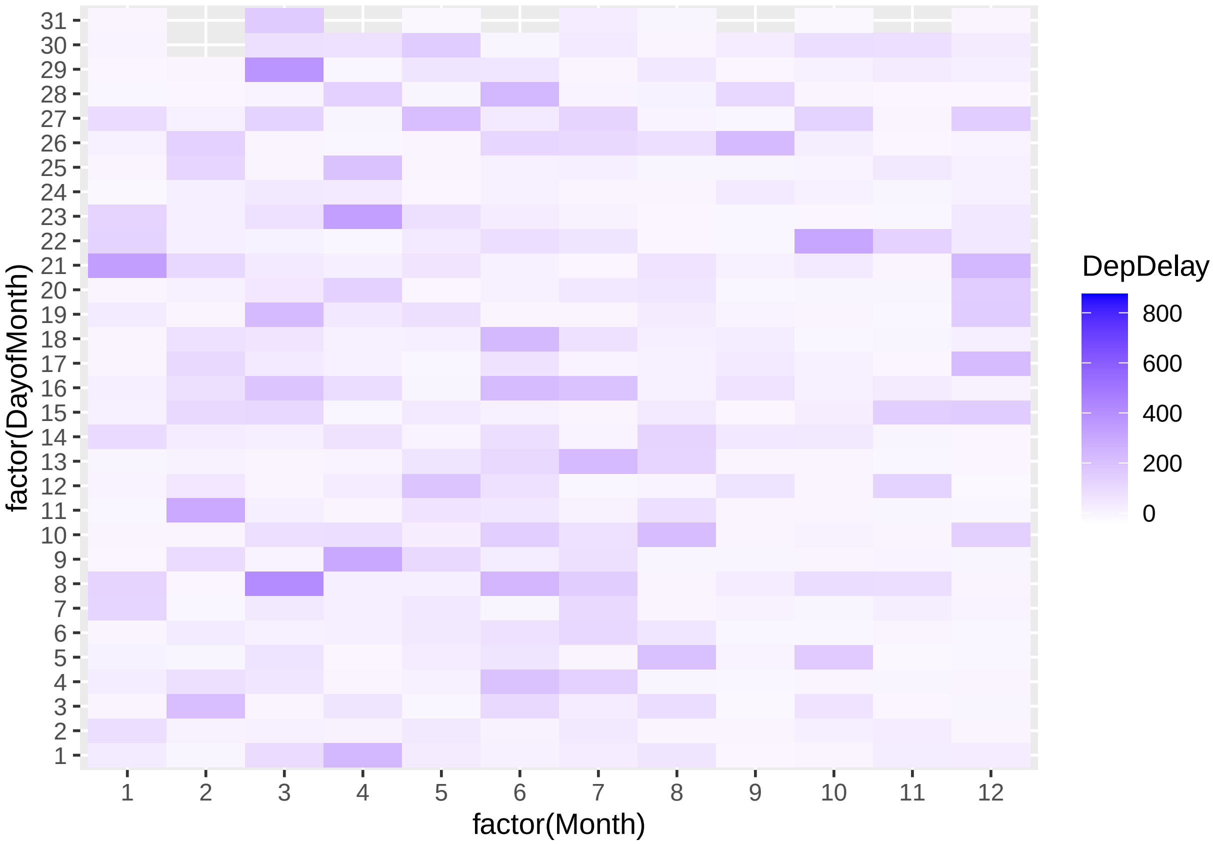

0.08
0.06
0.04
0.02
0.00
rate
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
DayofMonth



Figures 3-1.

The following Figures 3-2 present the average arrival delay and departure delay by date. The first figure presents the arrival delay. It indicates that some dates, such as March 8 and 29, may not be a good chance to visit Austin by airplane. The second figure presents some dates, such as March 8 and 29 still, which may not be a good chance to leave Austin by airplane. Also, the similarity indicates that if a day with more arrival delay, it is a high possibility of more departure delay.


31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
factor(DayofMonth)
1
2
3
4
5
6
7
8
9
10
11
12
factor(Month)
ArrDelay
750
500
250
0



Figures 3-2.

## QUESTION 4 K-nearest neighbors

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

*Conclusion: the model performs best for 350 and 65 AMG when the number of neighbors is 20.*

Figure 4-1 presents the testing process of 350 trim. I chose the number of neighbors as 1, 3, 5, 20, and 30. Results indicated that the best parameter is 20 (with the least RMSE). When the number of neighbors is less than 20, more neighbors lead to more accurate models. In contrast, situations change when neighbors are larger than 20.

```
##    Length     Class      Mode
##     29466 character character
```

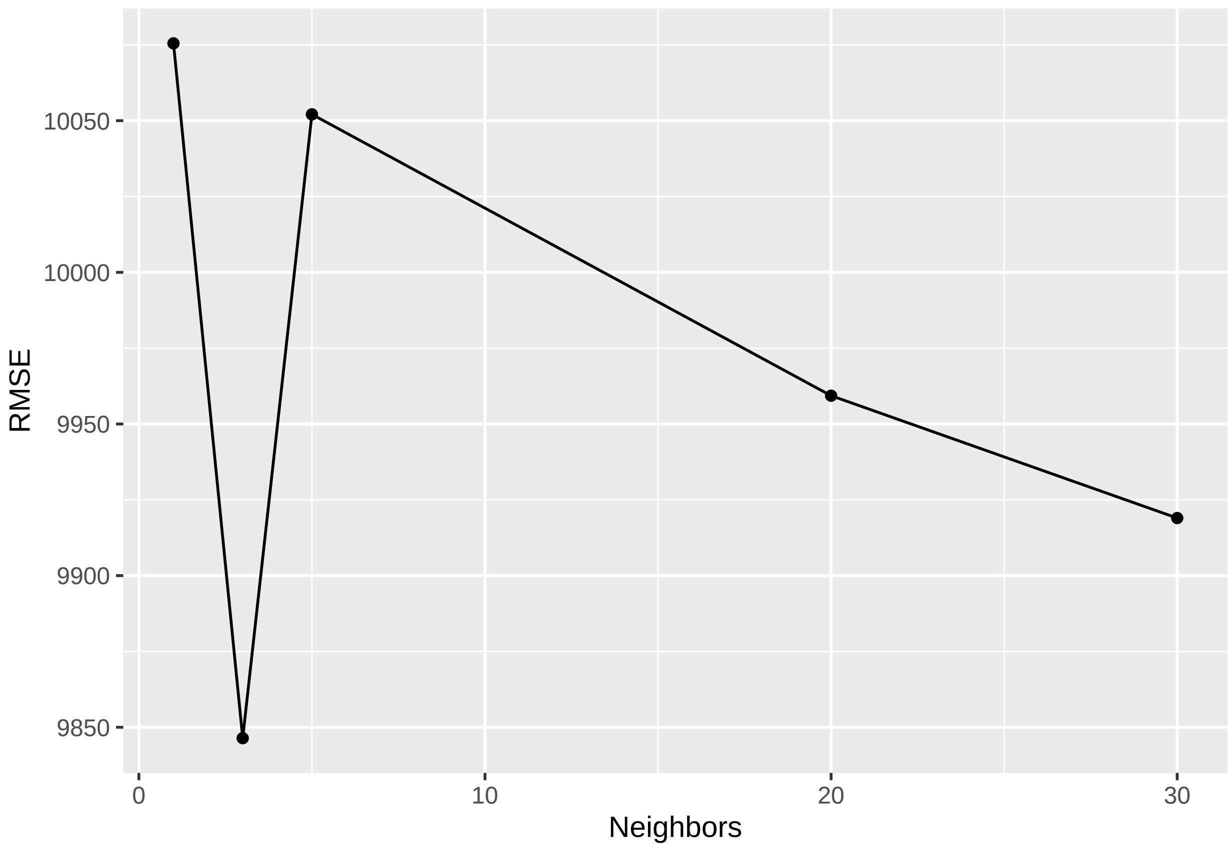


Figure 4-1.

Figure 4-2 presents the testing process of 65 AMG trim. I chose the number of neighbors as 1, 3, 5, 20, and 30. Results indicated that the best parameter is 20 (with the least RMSE). When the number of neighbors is less than 20, more neighbors lead to more accurate models. In contrast, situations change when neighbors are larger than 20.

```
##    Length     Class      Mode
##     29466 character character
```

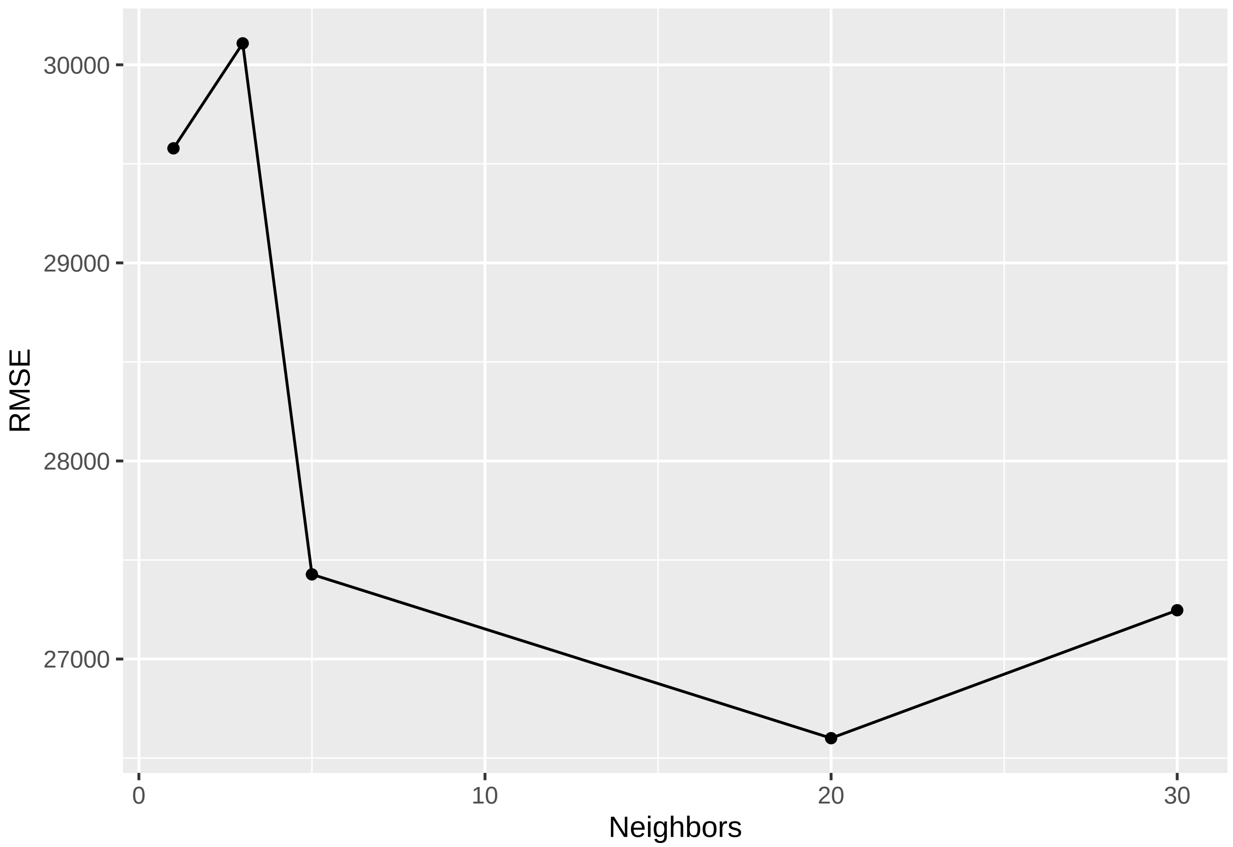


Figure 4-2.