Exercise1

Zonghao Li

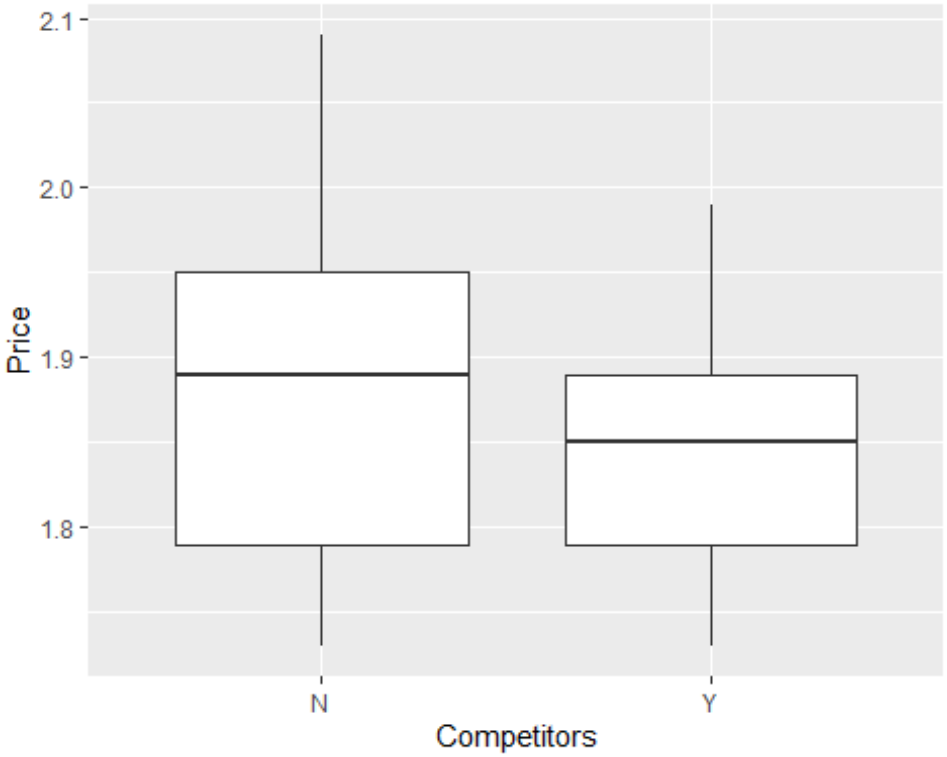
2021/2/8

Group members: Shuheng Huang; Zonghao Li

[Task1: Data visualization: gas prices]

A) 'Gas stations charge more if they lack direct competition in sight.'

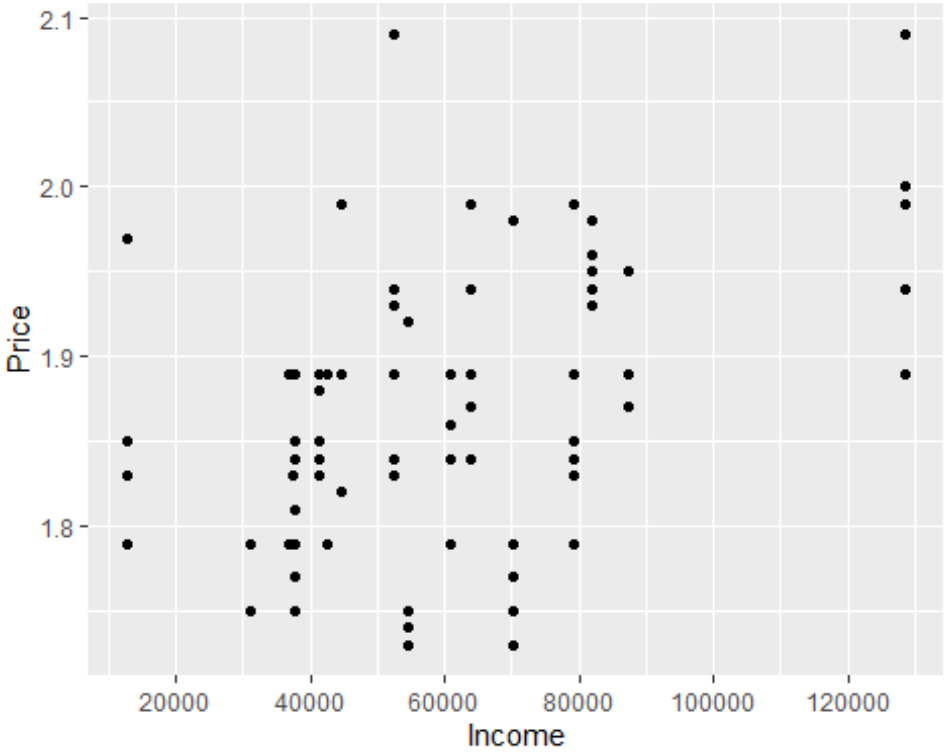
This statement looks meaningful since if there is no competitor, this gas station becomes the only station that provide gas and people living around there have no alternative. By contrast, if this station has several competitors, it may reduce its gas price to attract more customers.



Conclusion: Yes. From the boxplot, we can learn that in the 'No' group, the range is wider and the median is larger than 'Yes' group, so Gas stations charge more if they lack direct competition in sight.

B) 'The richer the area, the higher the gas price .'

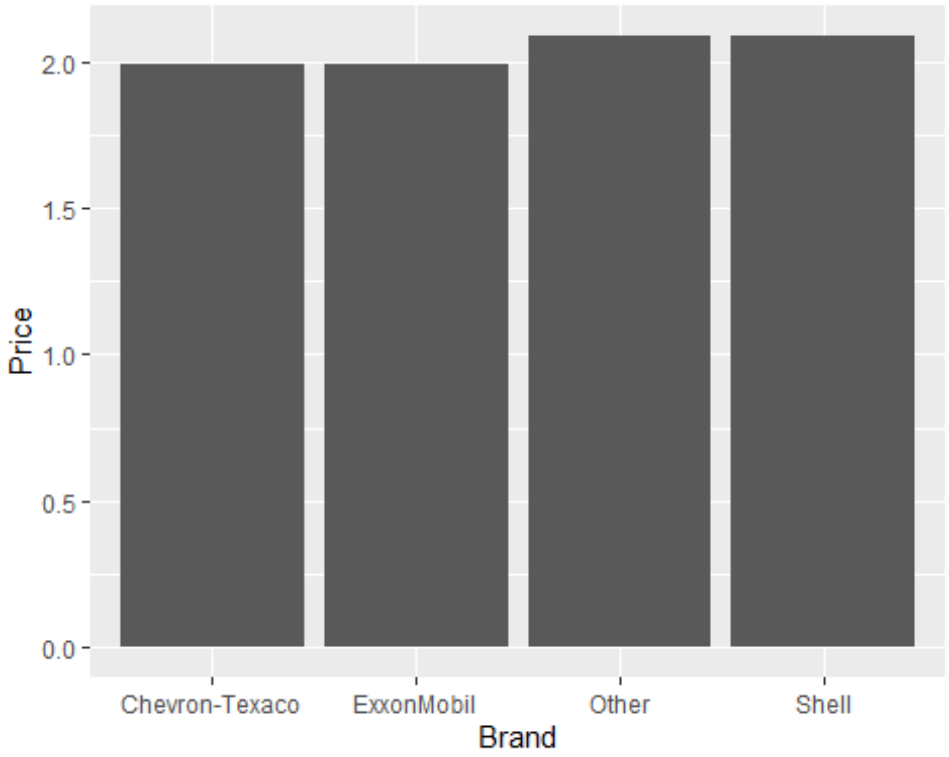
There is some truth to this statement because the richer the area, the higher the living standard of people living in that area. Then the price of goods in that area may be increase, as will the gas price.



Conclusion: Yes. Although the scatter plot shows a weak trend, there is a positive trend between income and price in general. This can show that the richer the area, the higher the gas price.

C) ‘Shell charges more than other brands.’

The brand named Shell may be more high-end than others in that area. If so, this brand will charge more than other brands.



Conclusion: This is not sure. From the bar plot, It seems that ‘Shell’ and ‘Other’ brands charge nearly equally to each other.

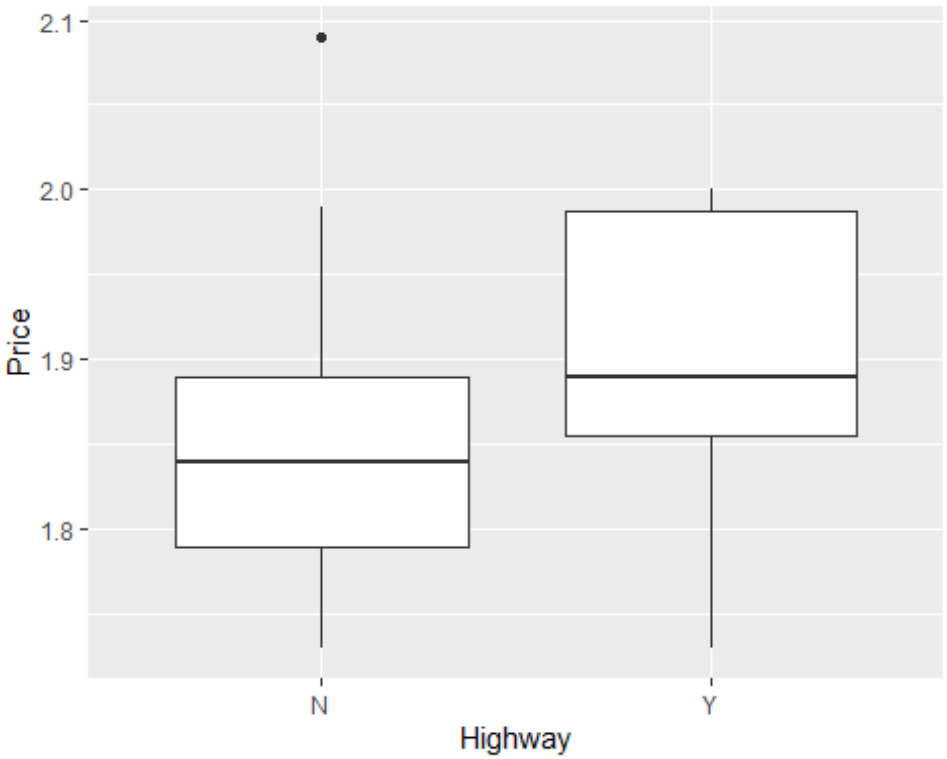
D) ‘Gas stations at stoplights charge more.’

This statement is not very plausible because when the cars stop at the stoplight, it provides more opportunities for them to add gas in the station nearby. And as the consumption goes up, the price of gas will go down to some extent.

Conclusion: The data of ‘Yes’ group focus on the prices prior to 2.0, while there exists ‘No’ group’s data more than price=2.0, so the stations not at stoplights charge more than those at stoplights.

E) ‘Gas stations with direct highway access charge more.’

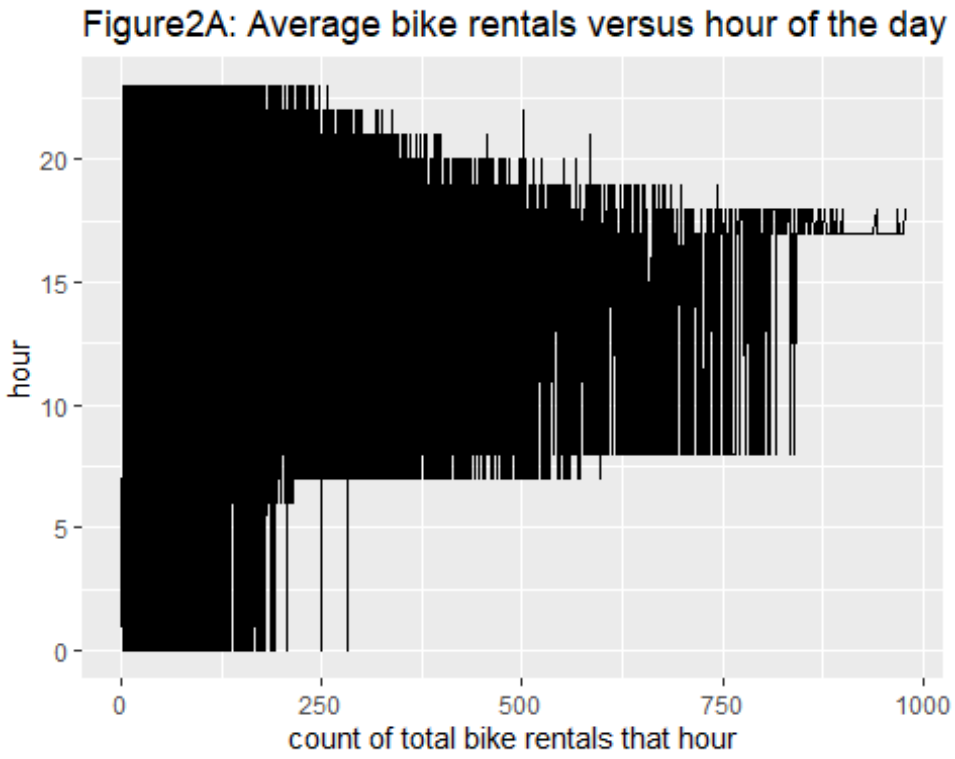
This statement is not very plausible since it provides more opportunities for the cars to add gas when there is direct access from highway to the gas station. And as the consumption goes up, the price of gas will go down to some extent.



Conclusion: From the boxplot, the range of quantiles of group ‘yes’ is higher and larger than the group ‘no’, so the gas stations with direct highway access will charge more than those without direct highway access, which does not correspond to our explanation above.

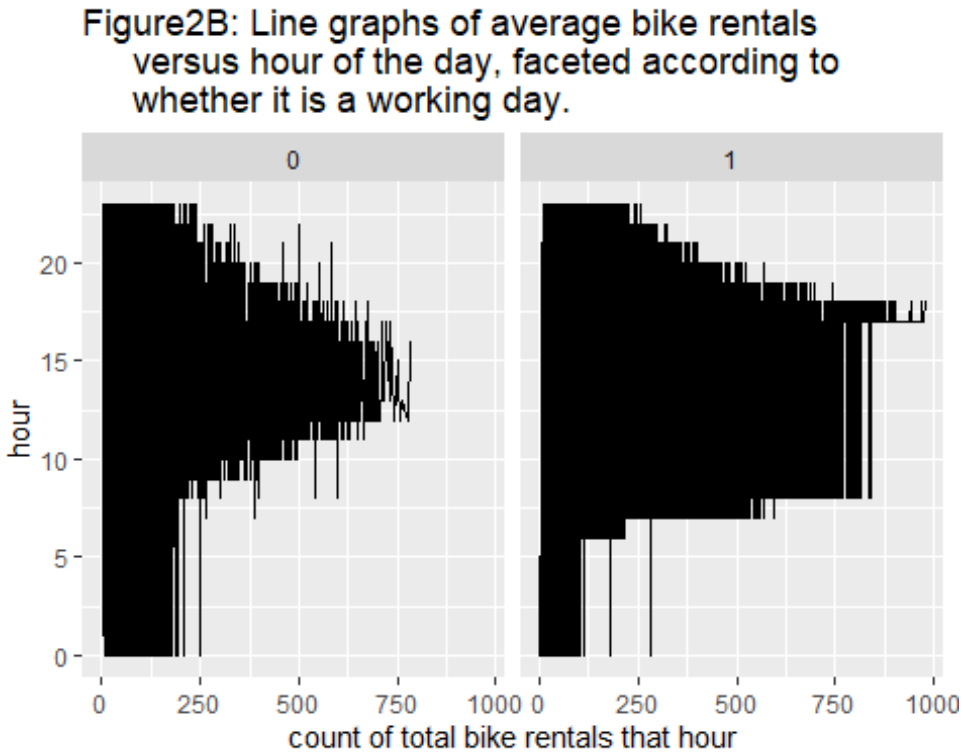
[Task2: Data visualization: a bike share network]

Plot A



Conclusion: On average, the number of bike rentals is high between 8 and 18 o'clock, especially around 17 o'clock.

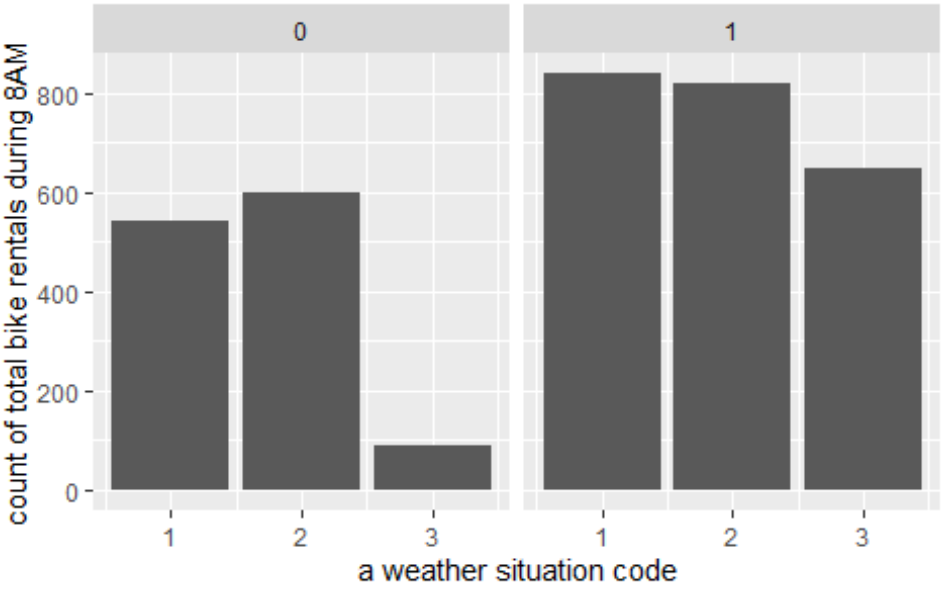
Plot B



Conclusion: When it is working day, the average number of bike rentals is large from day to night. During the weekend or holiday, the number of bike rentals from 0 to 5 in the middle of the night is more than on weekdays.(Or perhaps people on weekdays are more active at night.)

Plot C

Figure2C: Faceted bar plot of average ridership during the 8AM by weather situation code, faceted according to whether it is a working day or not.

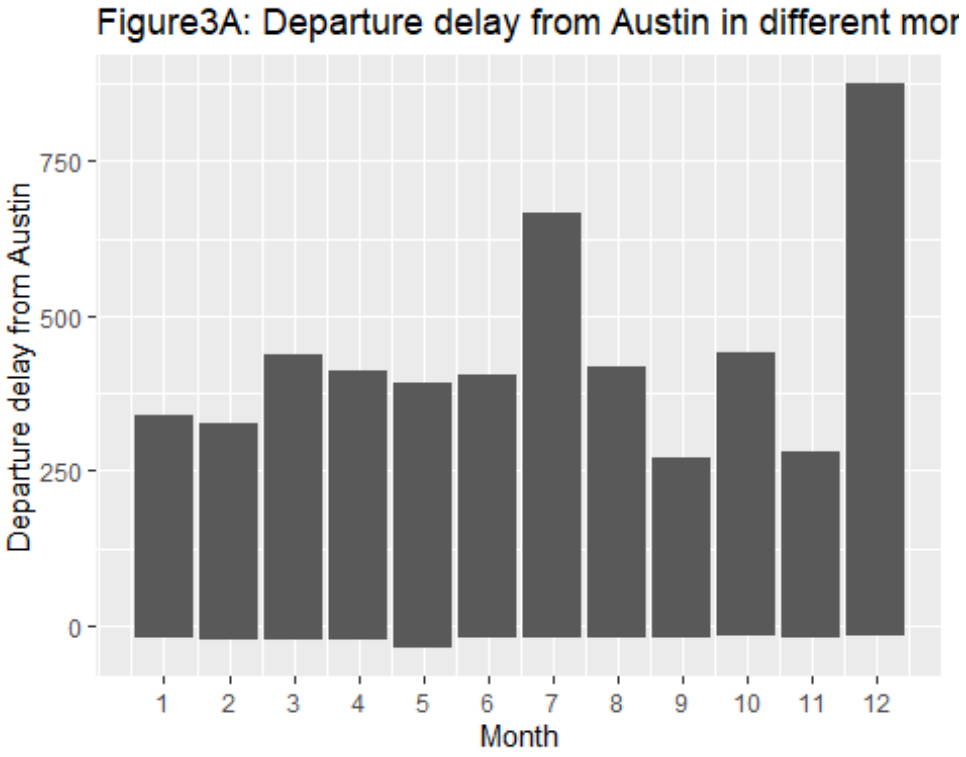


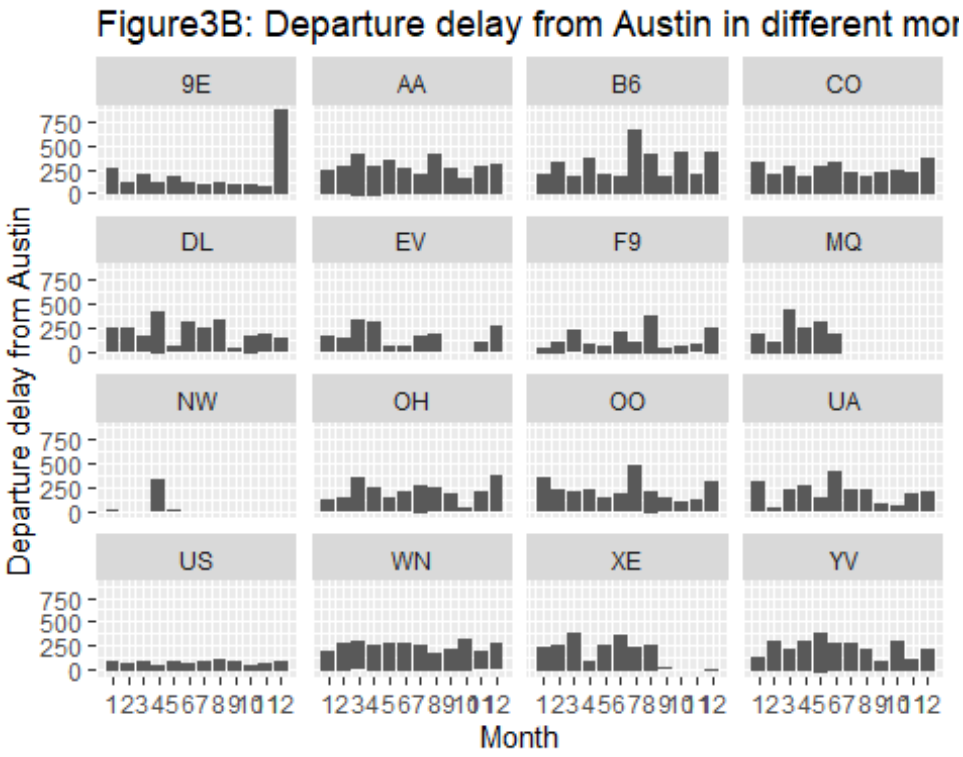
Conclusion: At 8am on a weekday, the number of rental bikes the first two comfortable weathers(like clear, few clouds, mist+clouds, etc) is similar to each other and is more than 800 on average, while the number in the third serious weather(like light snow, light rain+thunderstorm) is a little fewer, but also more than 600. At 8am on weekends or holidays, the number of rental bikes is still relatively similar in the first two more comfortable weathers, but is lower than weekdays, with only 500 to 600 cars respectively. However the number of cars rented in the more serious weather situation is much smaller than the first two, at just under 100. This means that people may prefer to stay at home during bad weather on holidays.

[Task3: Data visualization: flights at ABIA]

In the beginning, we divide the database of ABIA into 2 groups which are filtered by departure from or arrival in Austin, and then classify them by unique carrier code, in order to find the least departure delay time of airlines from or to Austin and which carrier's airlines delay the least in general.

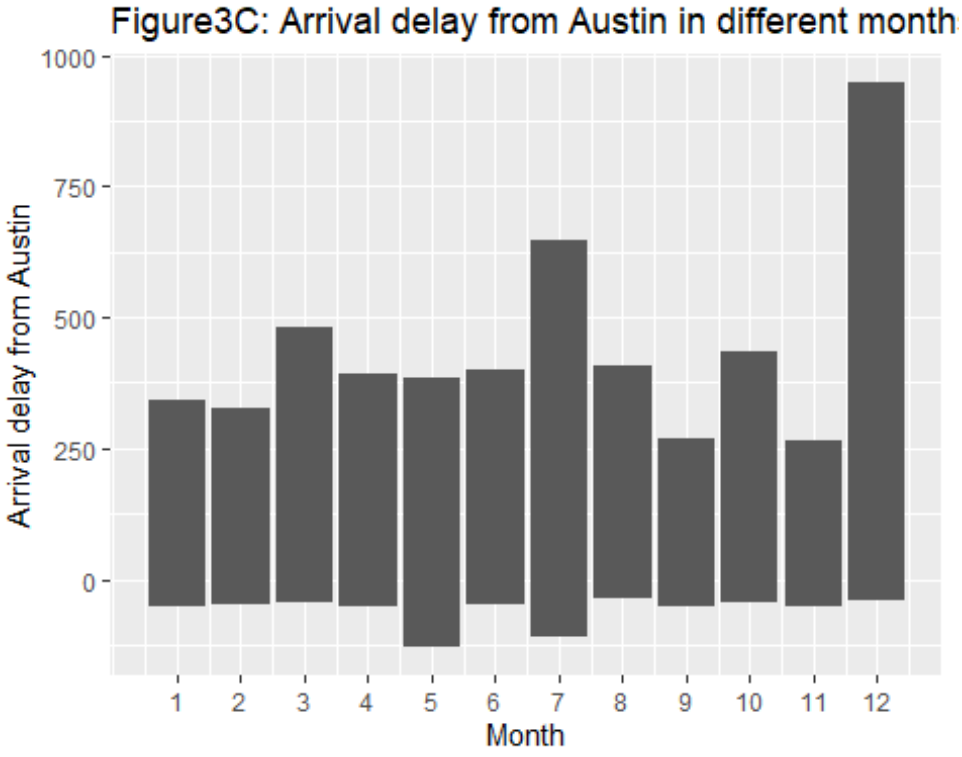
After this we use database of AustinDep to make a barplot of time of departure delay (Austin is the origin).

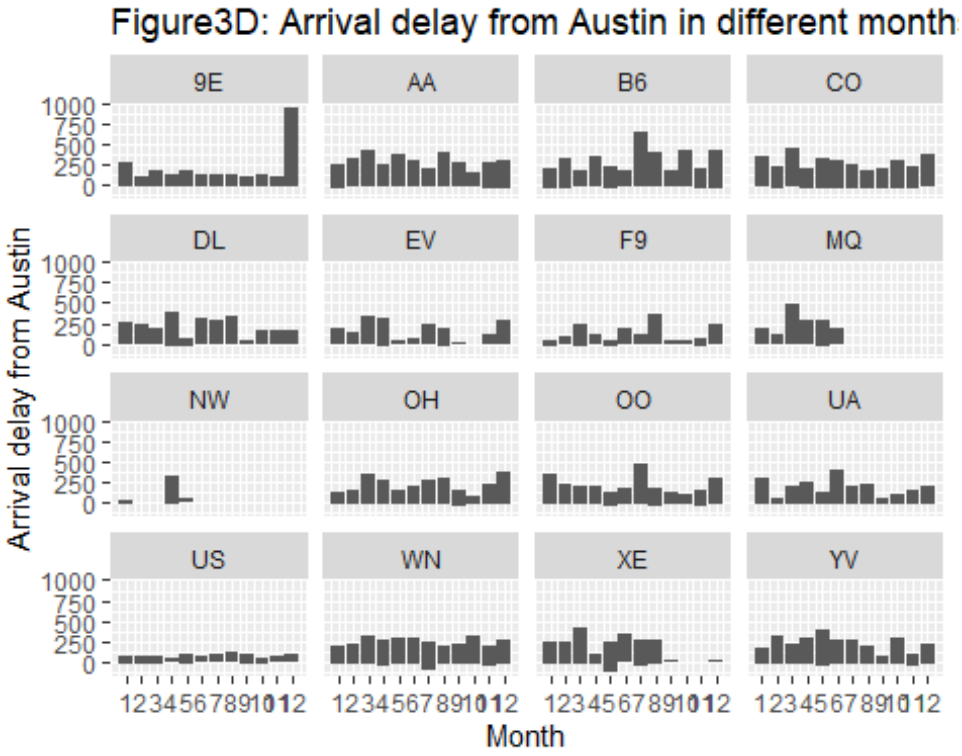




From Figure3A, we can learn that the departure delay time of airlines from Austin to other places showed the least in September. From Figure3B, US carrier's airlines delayed less than other carriers in general and apart from April, the time of departure delay of NW carrier's airlines was almost 0 in 2008.

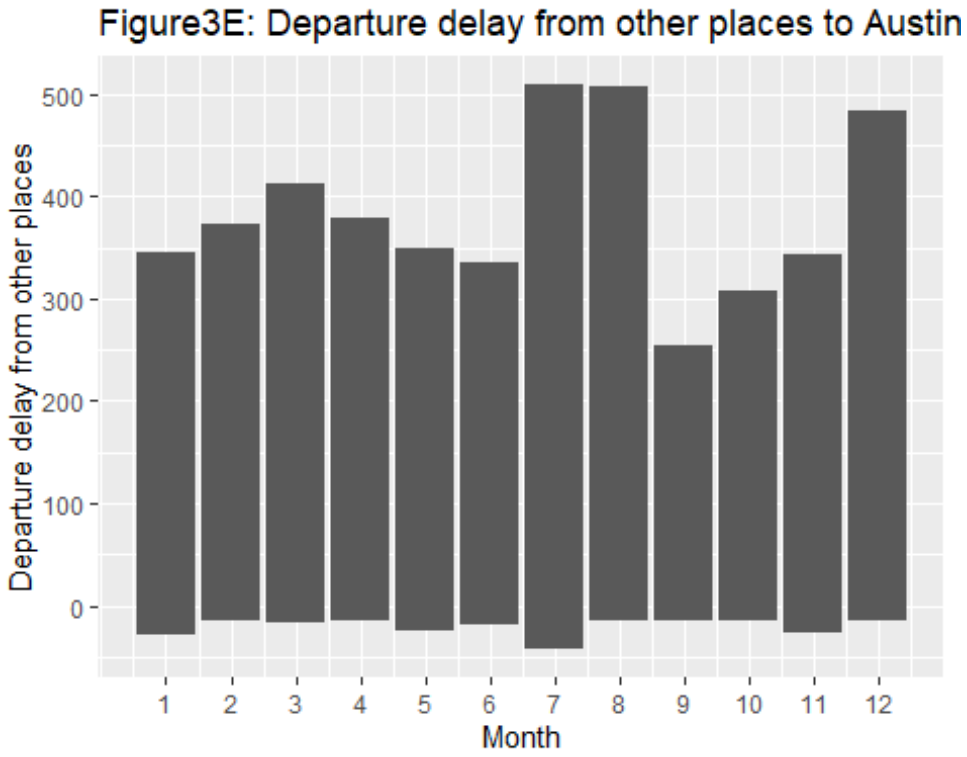
Then we use database of AustinDep to make barplots of time of arrival delay (Austin is the origin).

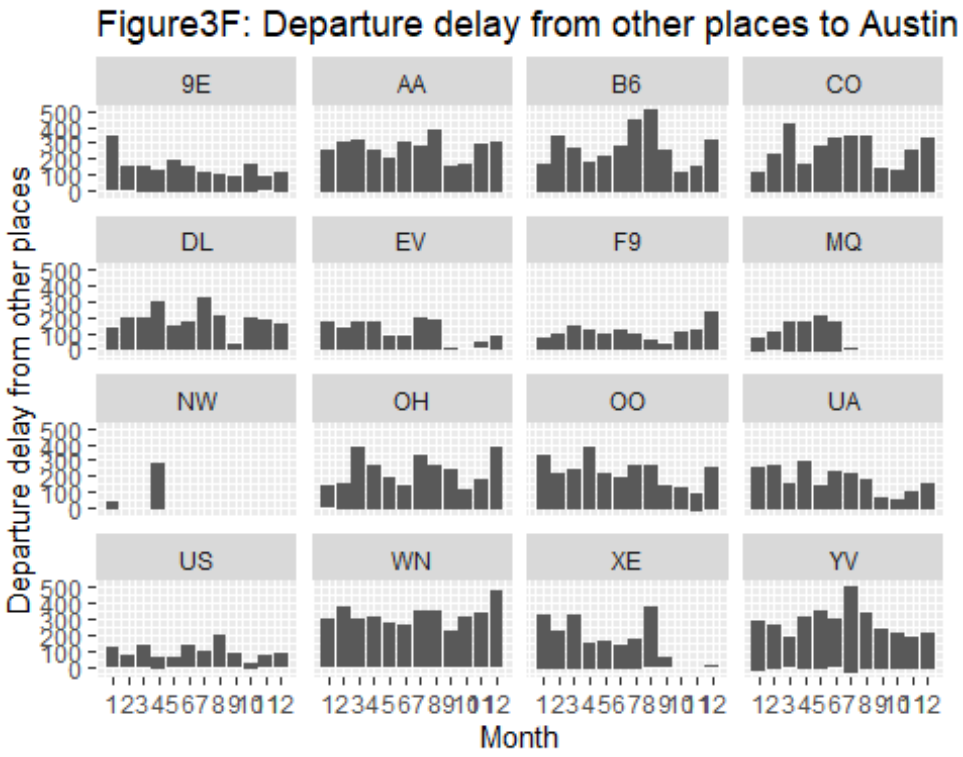




From Figure3C we can learn that the arrival in other places from Austin delayed less in September and November than other months. Similar to the first group, from Figure3D, the US and NW carrier delayed less than other carriers.

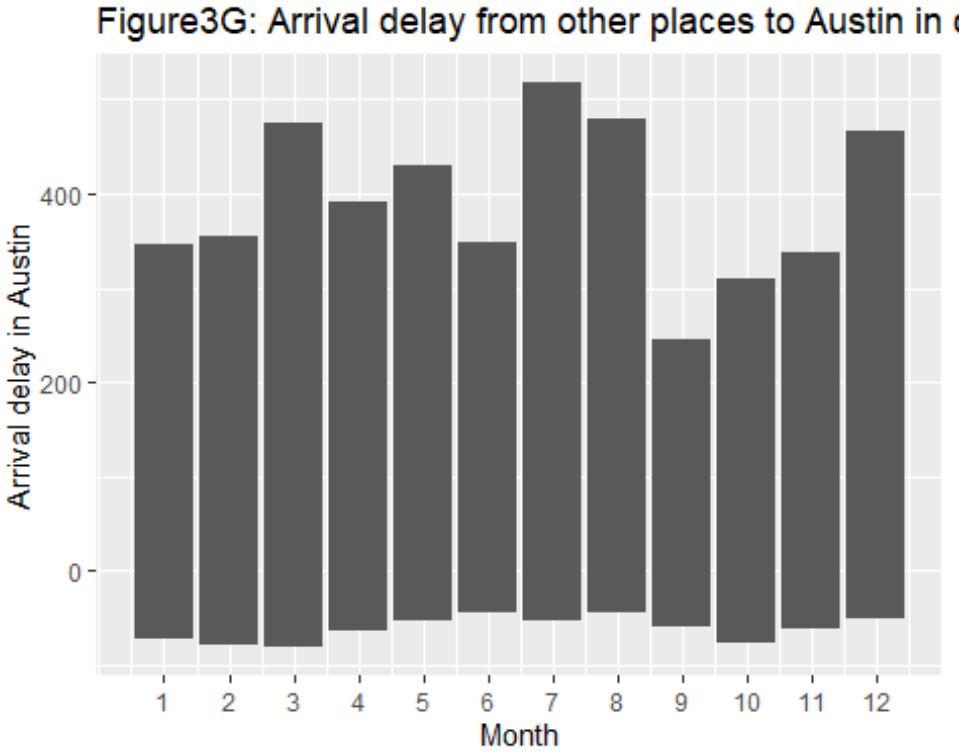
Use AustinDep to make a barplot of time of departure delay (Austin is the destination).

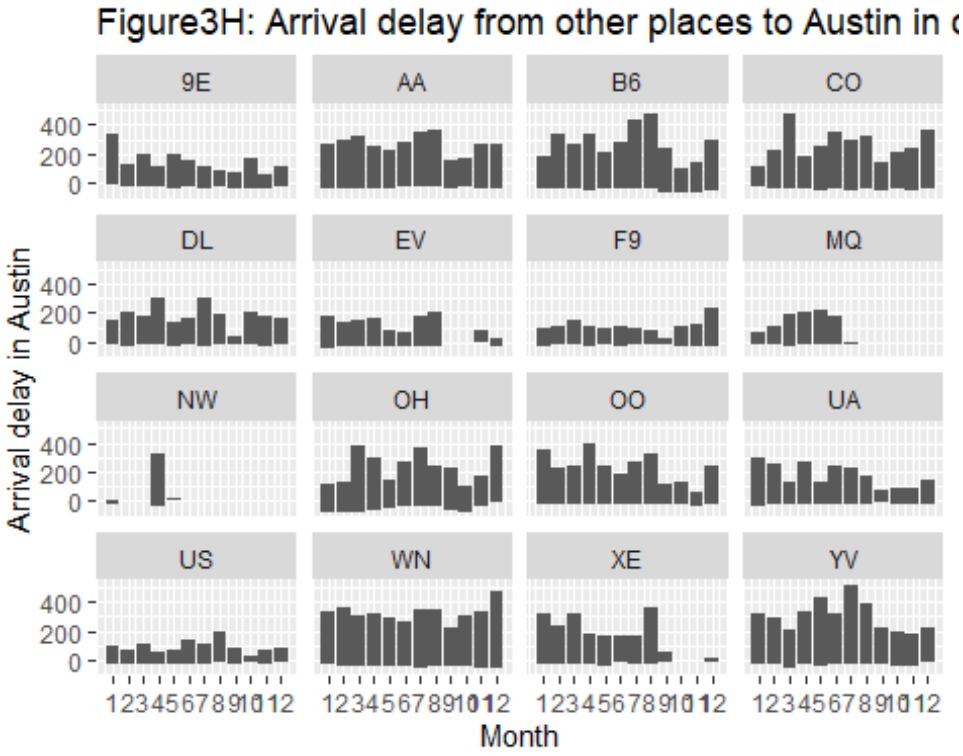




From Figure3E, the arrival in other places from Austin delayed less in September and November than other months. From Figure3F, NW, US and F9 carrier delayed less than other carriers in 2008.

Then use dataset of AustinDep to make a barplot of time of arrival delay (Austin is the destination).





From Figure3G, the arrival in other places from Austin still delayed less in September than other months. From Figure3H, the NW carrier delayed less than other carriers in 2008, and US, F9 and MQ were close to each other.

Next, in general, make barplots of time of departure delay in 2008 considering all airlines.

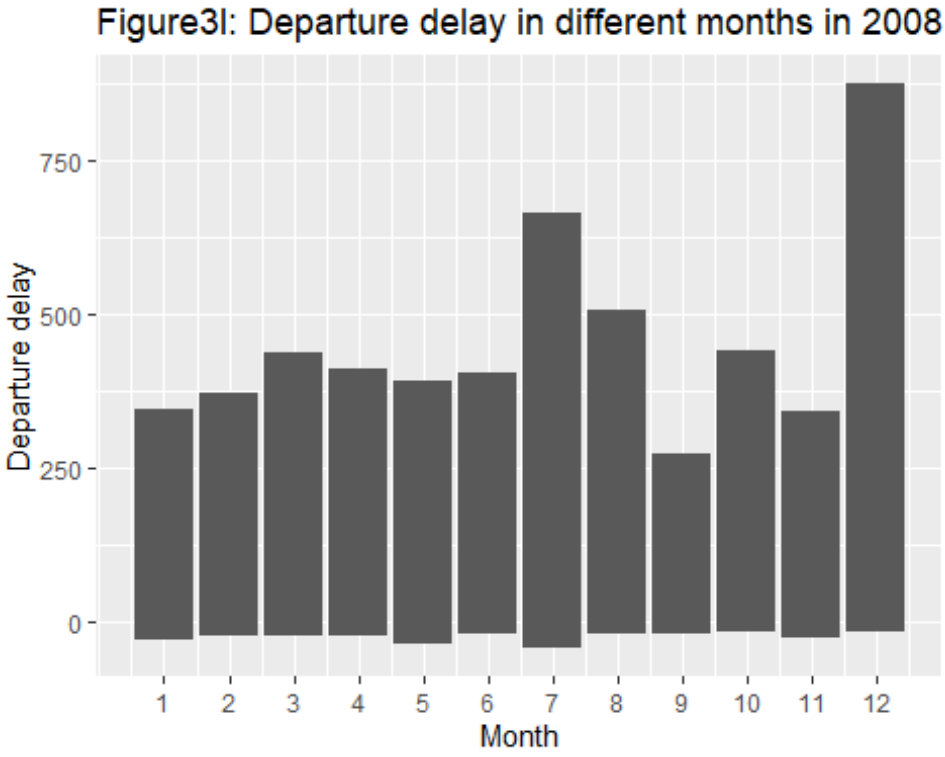
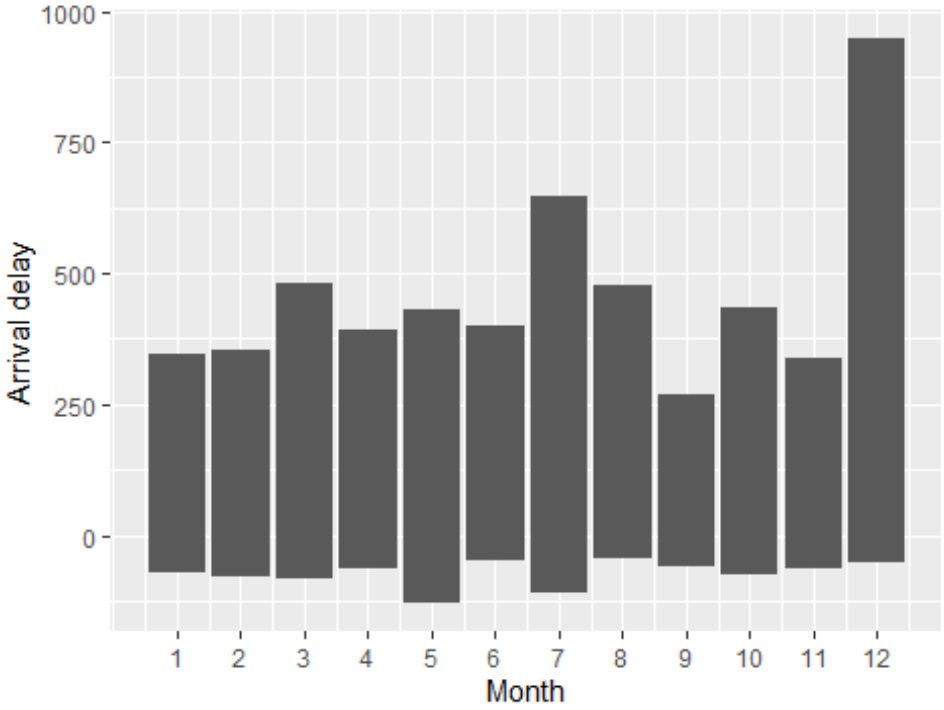
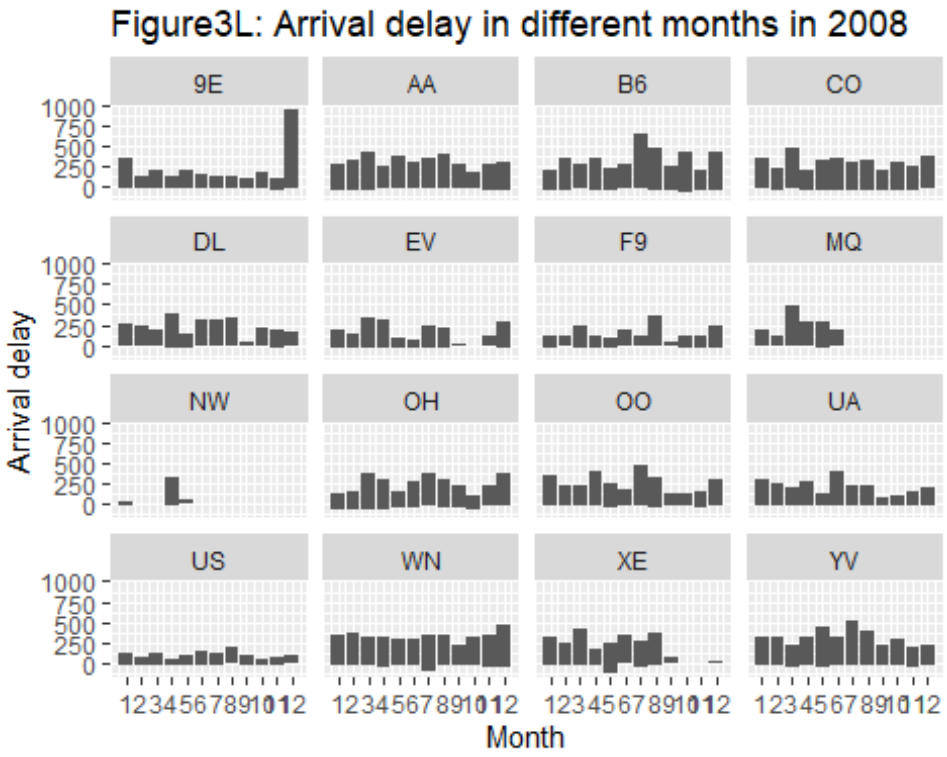


Figure3J: Departure delay in different months in 2008

Make barplots of time of arrival delay in 2008. (Consider all airlines)

Figure3K: Arrival delay in different months in 2008



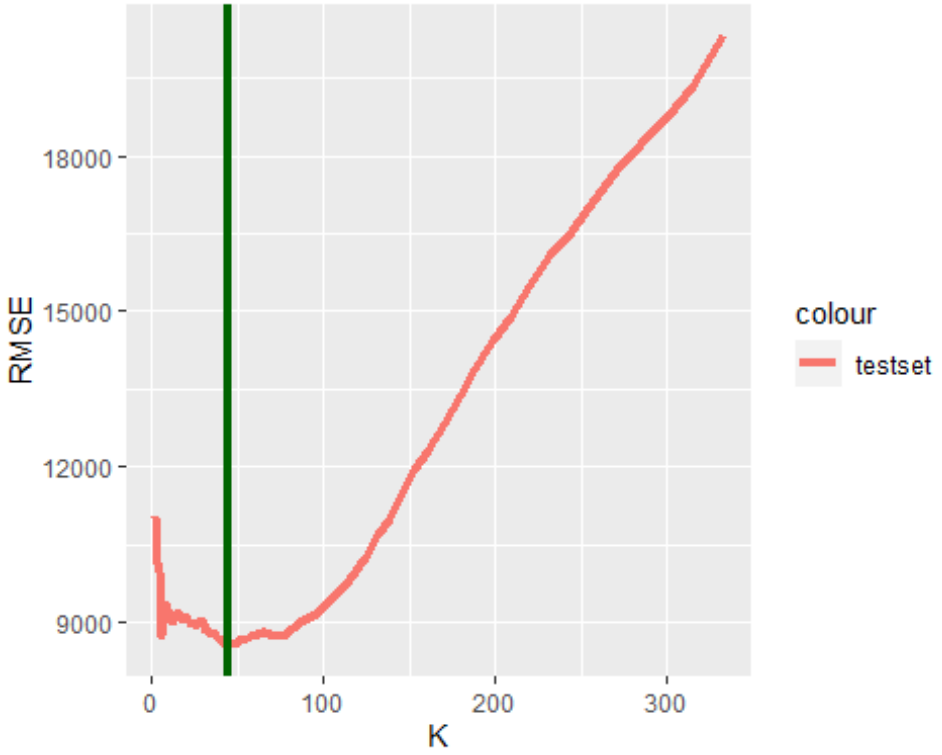


When considering all airlines, we can conclude that airlines delayed the least in September. Among these carriers, airlines of NW and US delayed the least on averge. Thus from the government's point of view, NW and US airlines can be supported more. From the carriers' point of view, it is reasonable to lower ticket prices to some extent in September, or even in November, in order to attract more passengers with lower delay rates.

[Task4: K-nearest neighbors]

It is difficult to provide accurate pricing predictions to consumers due to the unusual characteristic of a single model of car. In this task, our 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.

Firstly, focus on first trim level: 350.

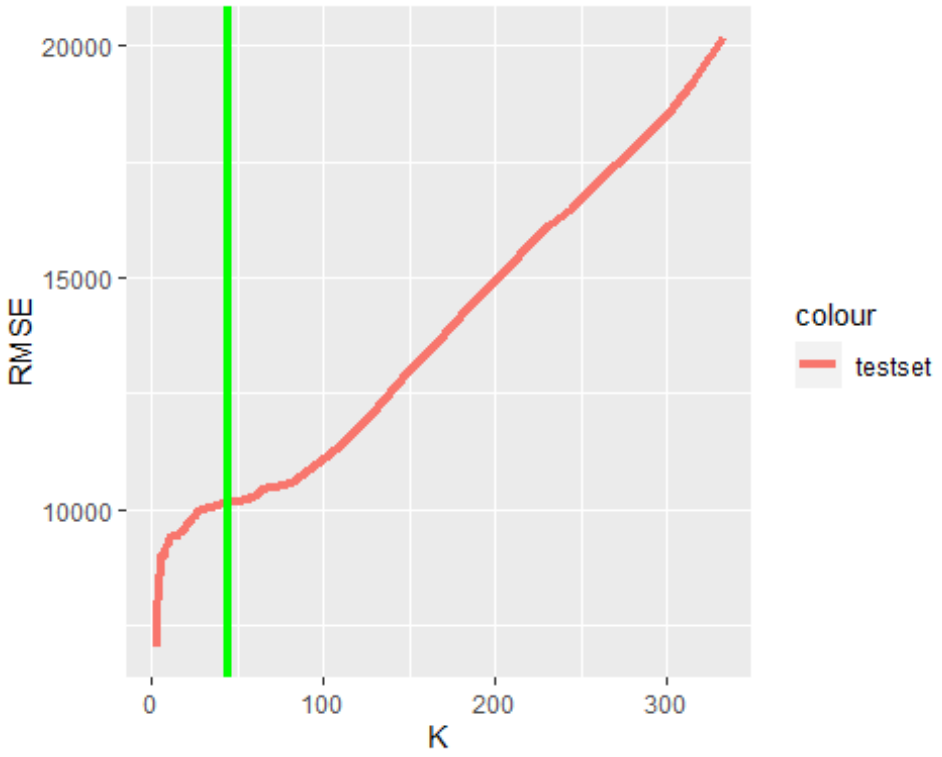



```
## [1] 44
```

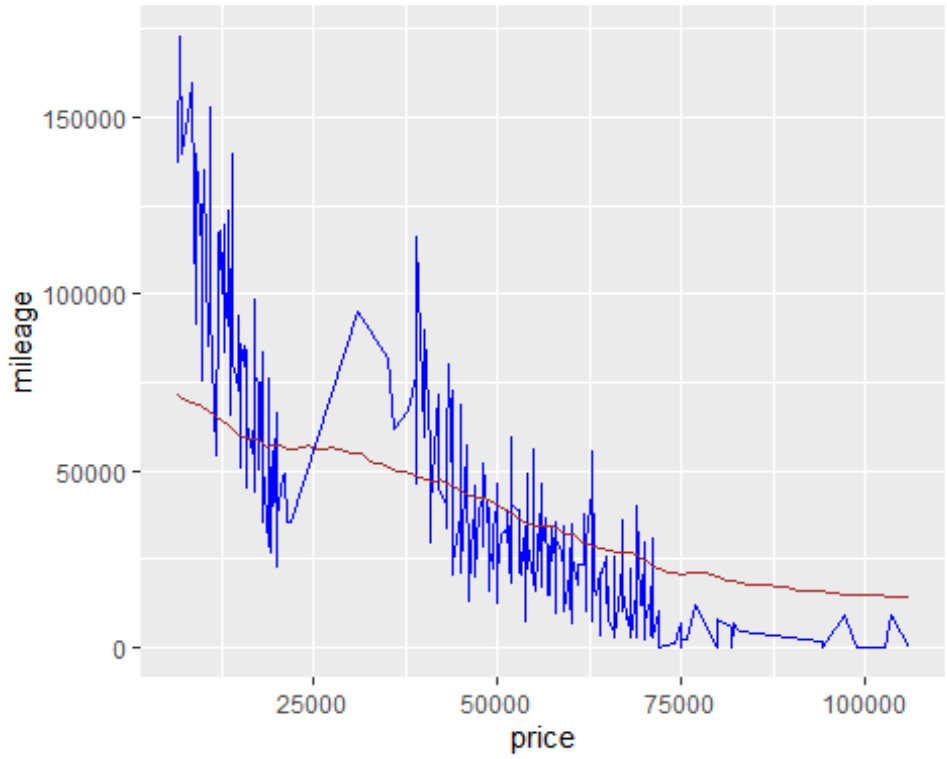
This graph above tells us the relation between RMSE versus K in this model.

Then train the model of 350 trim and calcuate RMSE on the test set.

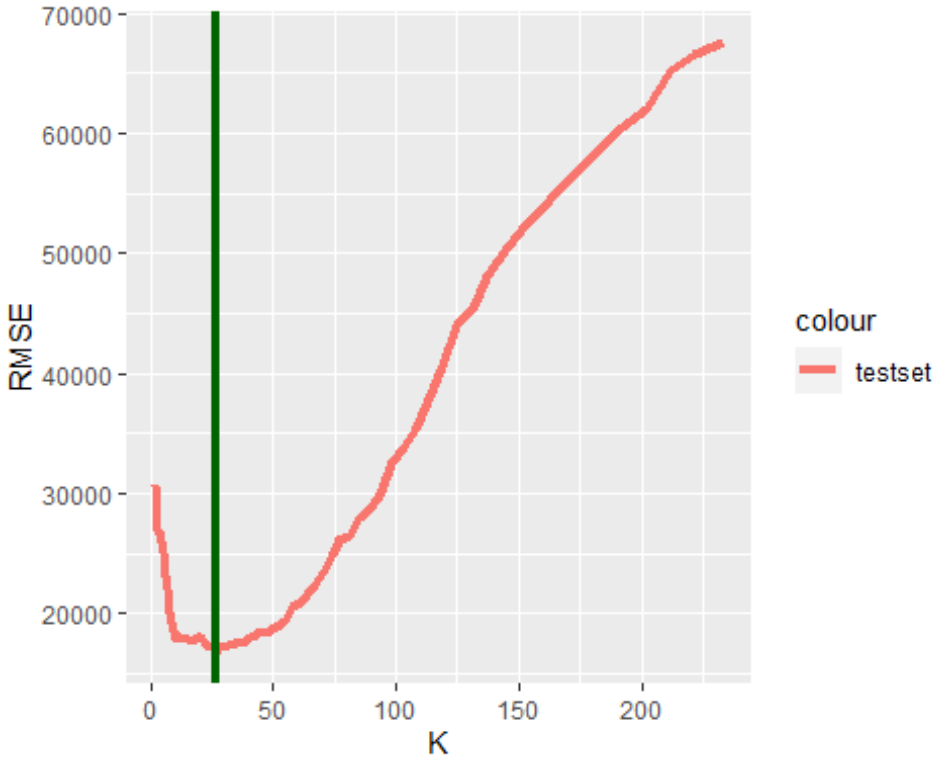
```
## [1] 9122.012
```



Then we fit the model to the training set and make predictions on the test set.



Next, focus on first trim level: 65 AMG.

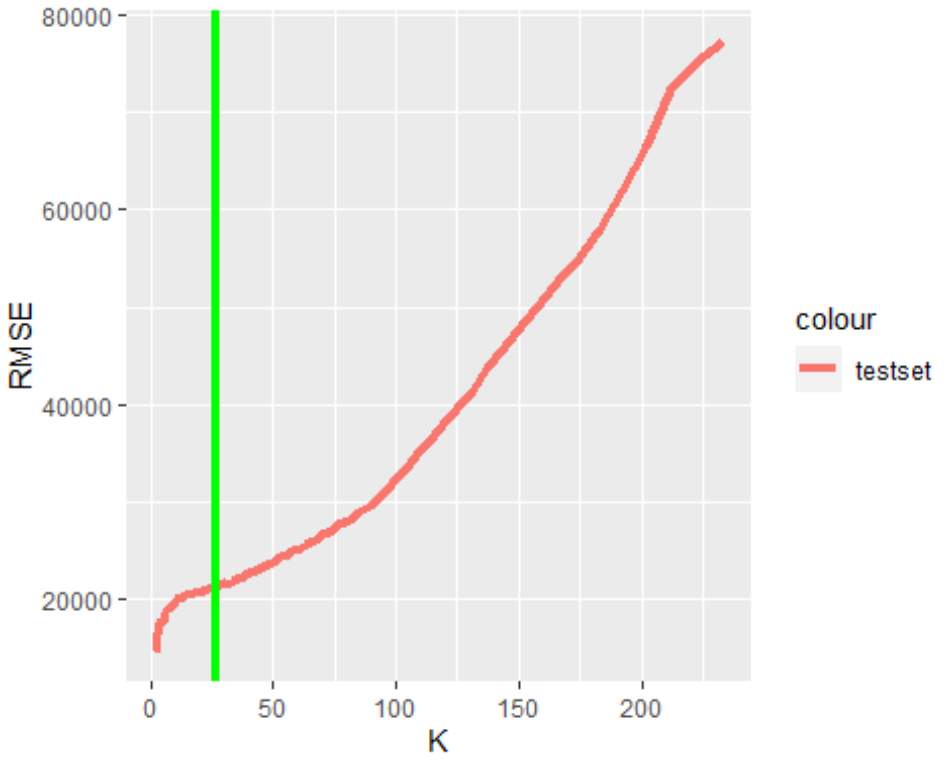



```
## [1] 27
```

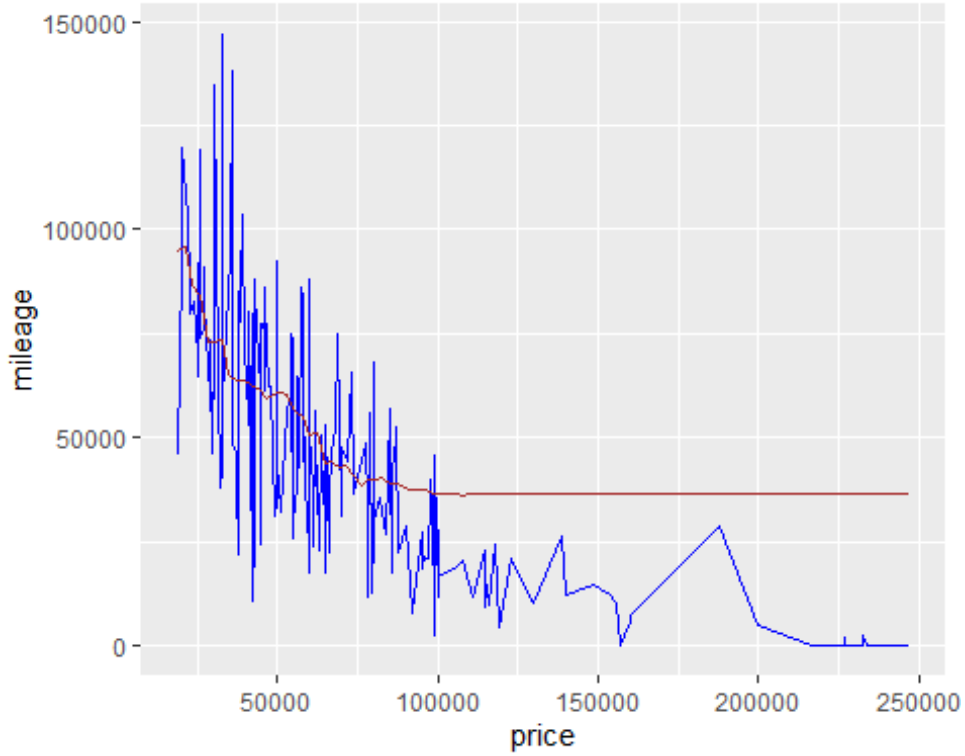
This graph above tells us the relation between RMSE versus K in this model.

Train the model of 65 AMG trim and calcuate RMSE on the test set.

```
## [1] 18072.45
```



Then we fit the model to the training set and make predictions on the test set.



In conclusion, the 350 trim yields a larger optimal value of K, which equals to 42. And under the optimal value of K, the RMSE of 350 trim is smaller than RMSE of 65 AMG trim. To find the reason for that, we can look at the two graphs of relation between price and mileage. We can see that more points is far away from x in the 350 trim graph than that in the 65 AMG trim graph, especially before the field of price=10000. This means that the 350 trim model has high bias since these far-away points bias the prediction, so our K value is a bit larger in the 350 trim graph.