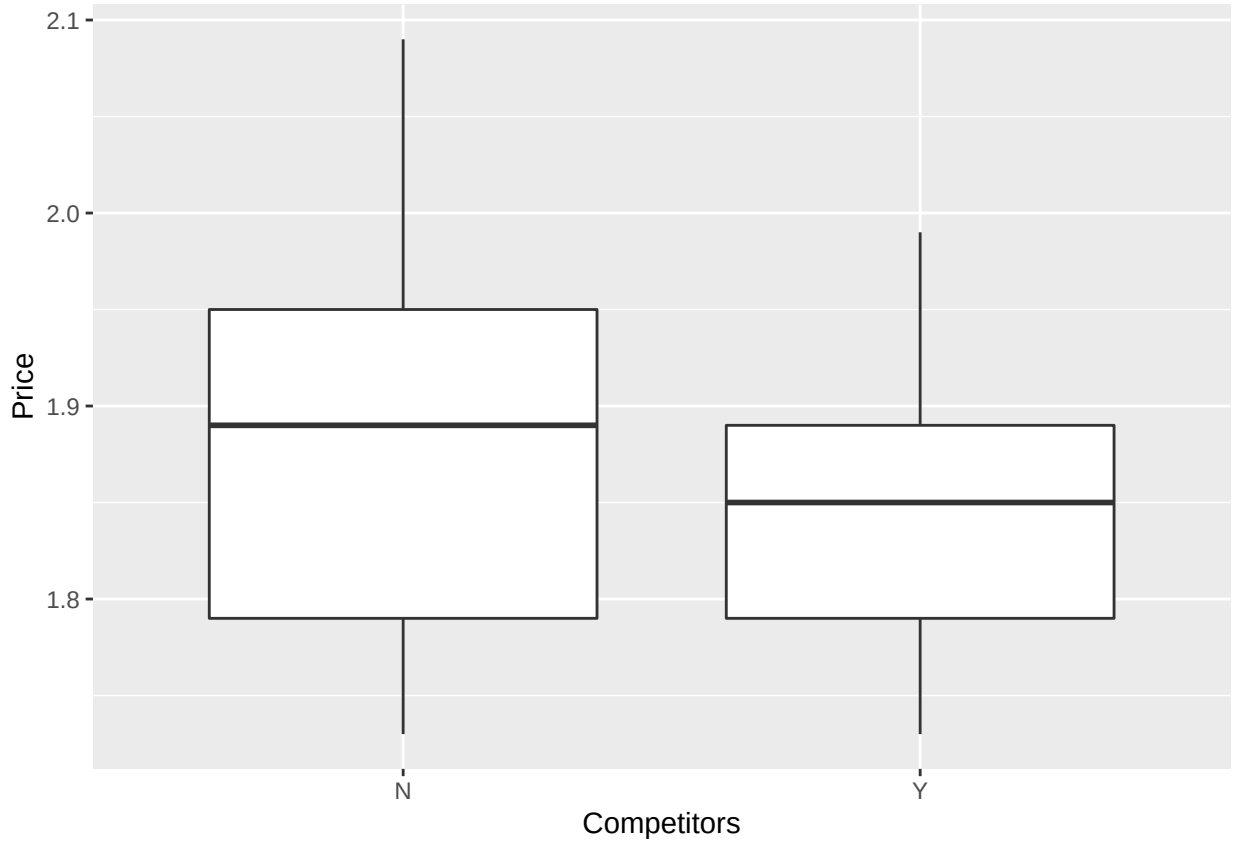
Exercise 1

Zhiqian Chen, Yi Zeng, Qihang Liang

2/6/2021

Gas prices

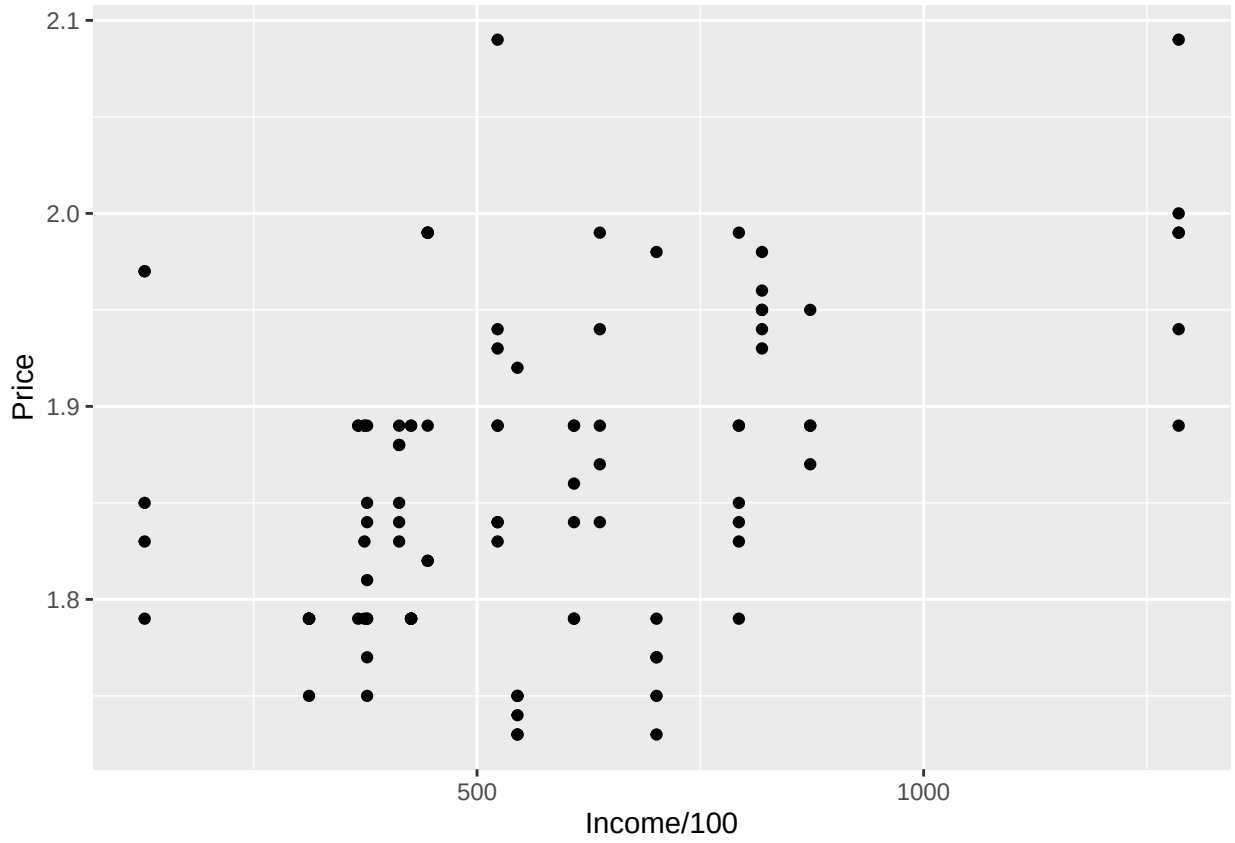
1.A



Claim:The theory says that gas stations charge more if they lack direct competition in sight. In my opinion, I think the theory is reasonable. If there is a competition between two gas stations, price is the best strategy. In order to attract customers, some gas stations will reduce prices. Therefore, when gas stations lack direct competition, they tend to charge more.

Conclusion:According to the plot, it shows that the prices of the gas stations with competitor is higher than without competitor. The theory is supported by the data.

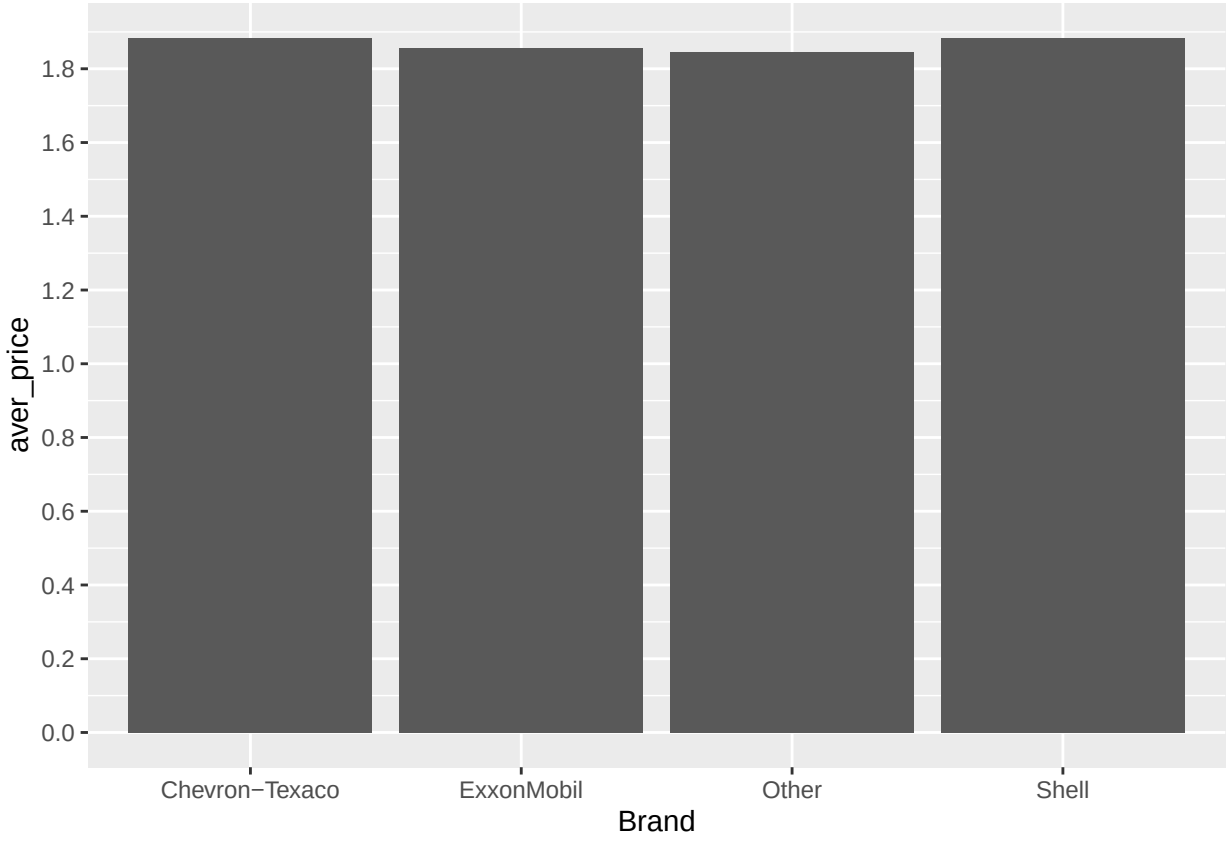
1.B



Claim: This theory says that there will be higher gas price in the richer area. I think this theory is reasonable since residents in richer areas will have higher incomes, so their consumption power will be stronger. Higher gas prices will not have much impact on the sales.

Conclusion: In the plot, x aes is the income level and the y aes is the price of gas, here, the income is large, in order to be clearly displayed in the plot, divide the income by 100. According to the distribution of the point in the plot, when the income is low, the distribution of most points tends to 0, which means low prices. When the income increases, the points are distributed outwards, which means high prices. So, the theory is supported by the data.

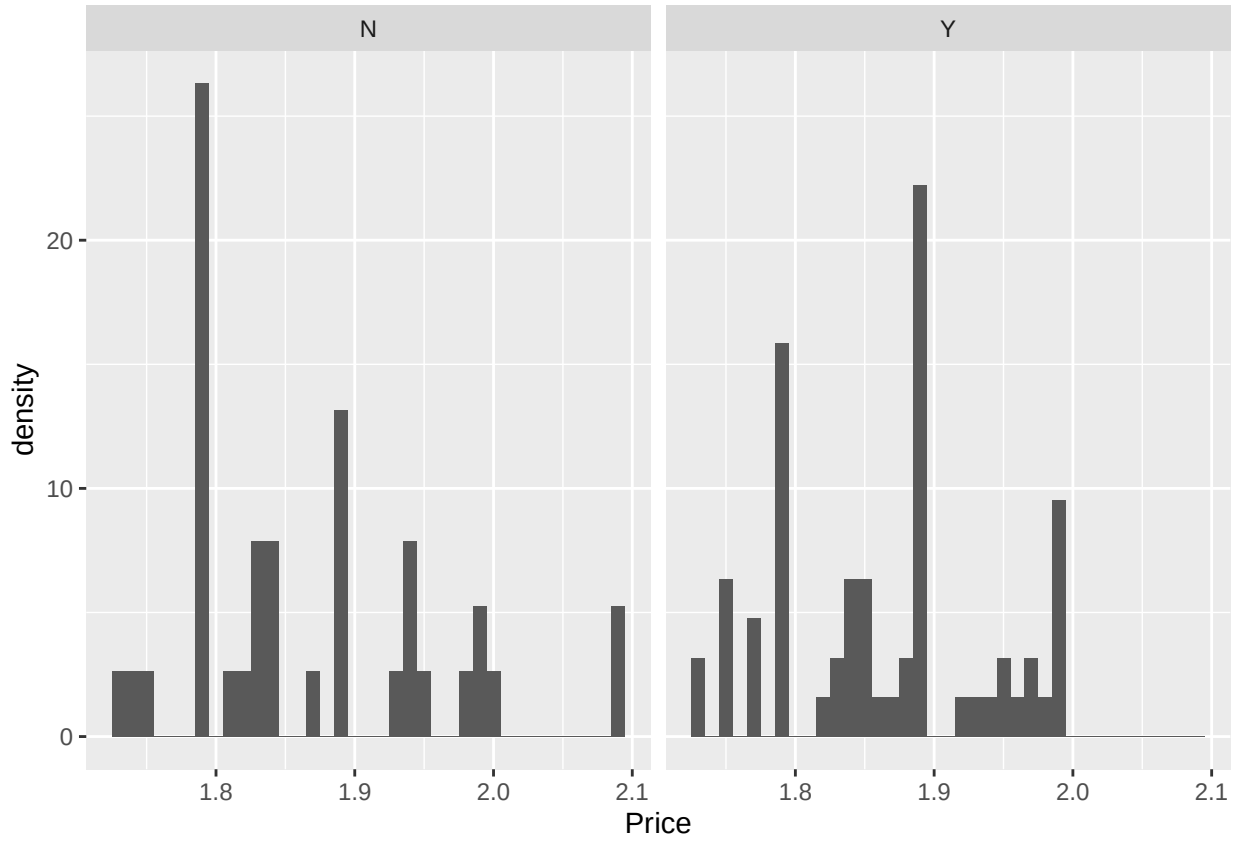
1.C



Claim: This theory says that Shell charges more than other brands. The data shows the price of each gas station of different brands, so if we want to compare the price of each brand, we need to compute the average price of each brand for comparison.

Conclusion: The bar plot shows that Shell's average gas prices is a little higher than other brand, so the theory is supported by the data.

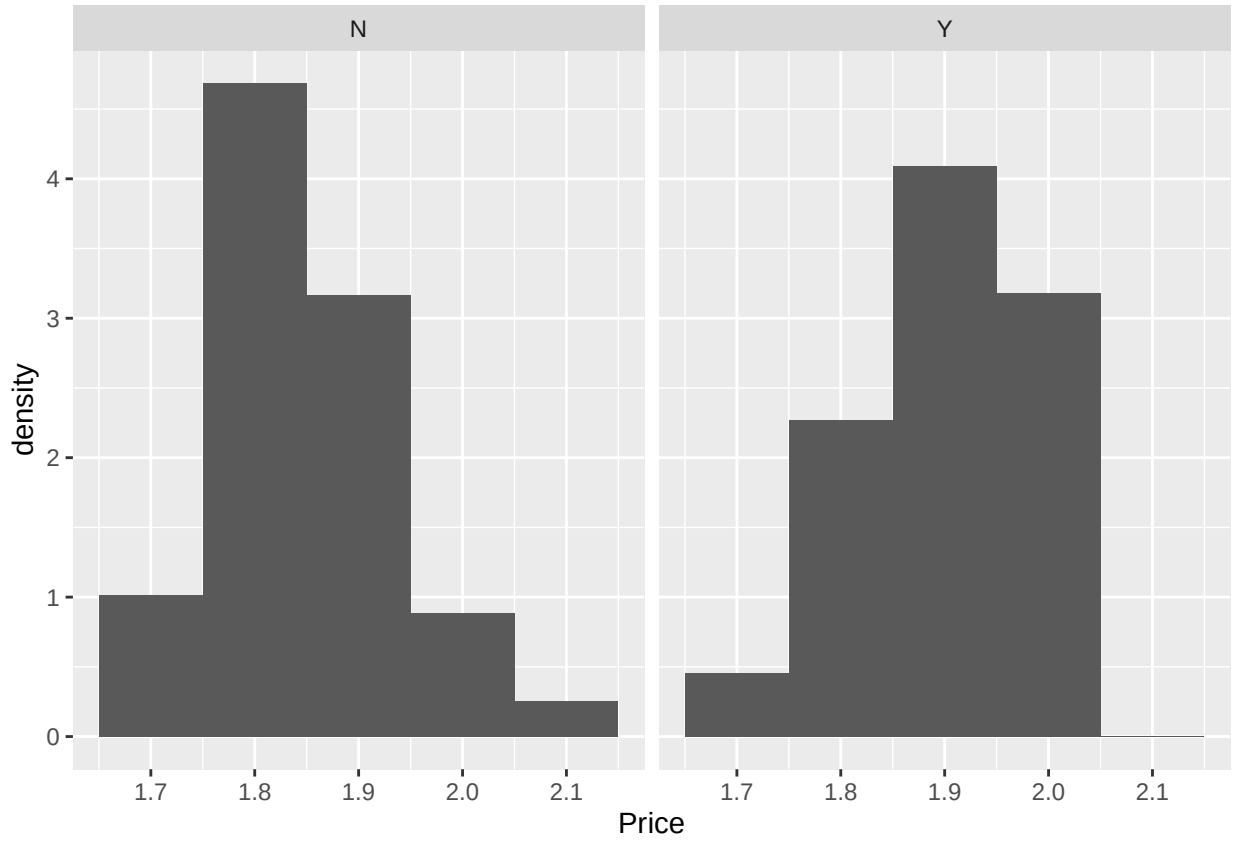
1.D



Claim: The theory says if there is a spotlight in front of the gas station, the gas station will charge more. This theory may be reasonable. When there are stoplight in front of the gas station, cars may enter the gas station while waiting for the spotlight. At the same time, customers may notice the gas station while waiting for the stoplight. If the customers are driving on the road, they may focus most of their attention on the traffic conditions rather than the gas stations. Therefore, gas stations near the stoplight can easily attract customers' attention. Also, the gas stations near the stoplight has good location, the rent may be high, so they may charge more.

Conclusion: In the faceted histogram, the density of price of gas stations do not in front of stoplight is distributed most around \$1.75 to \$1.85, and the density of price of gas stations in front of stoplight is distributed most around \$1.85 to \$1.95. So the plot tell us that gas stations in front of stoplight charge more, the theory is supported by data.

1.E

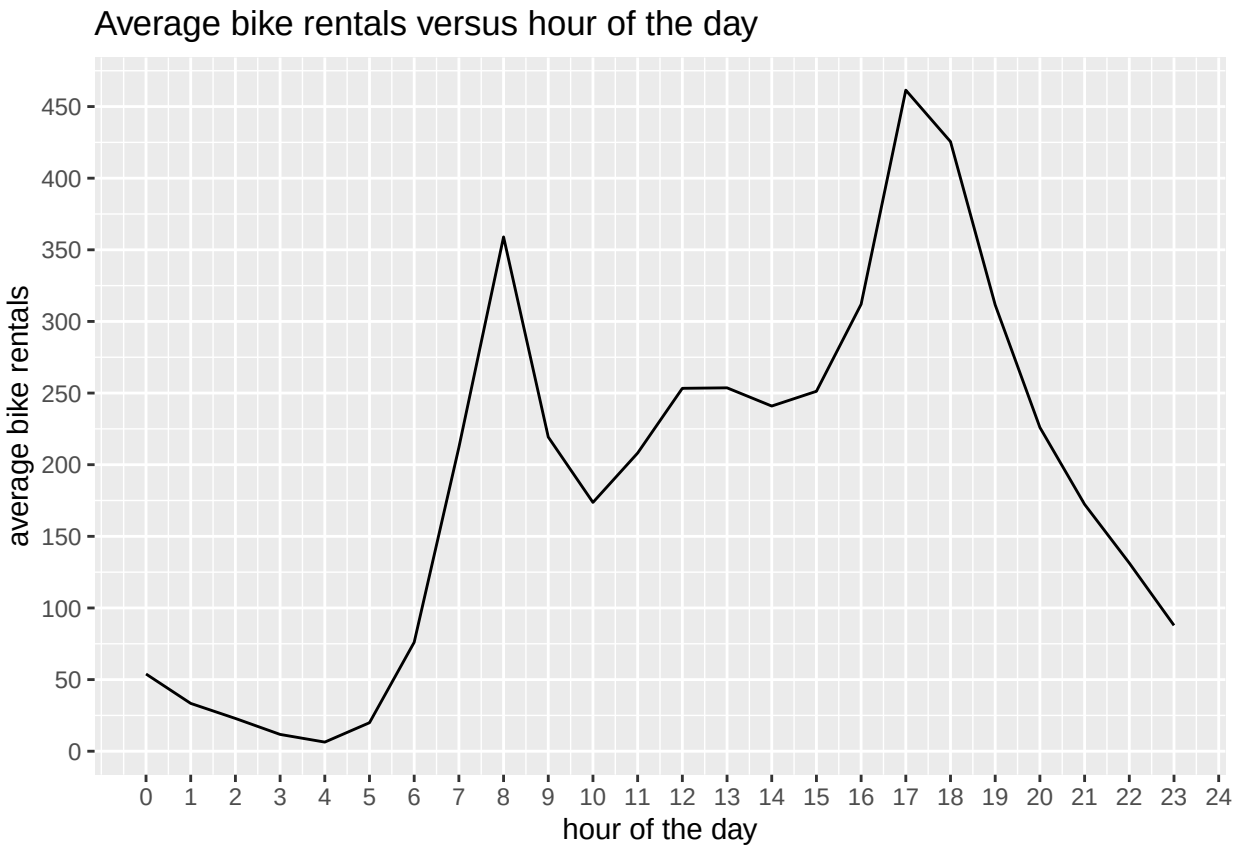


Claim: This theory says that gas stations with direct highway access charge more. This theory is reasonable since there will be many customers at gas stations that have direct access to the highway. The traffic volume on highway is much greater than on ordinary highways. At the same time, when people travel long distances, gas stations are not only a place for gas, but also a place to rest. Therefore, gas stations that can directly access the highway often charge more.

Conclusion:In the faceted histogram, the density of price of gas stations cannot directly access the highway is distributed most around \$1.75 to \$1.95, and the density of price of gas stations can directly access the highway is distributed most around \$1.85 to \$2.05. So the plot tell us that gas stations can directly access the highway charge more, the theory is supported by data.

A bike share network

2 plot A

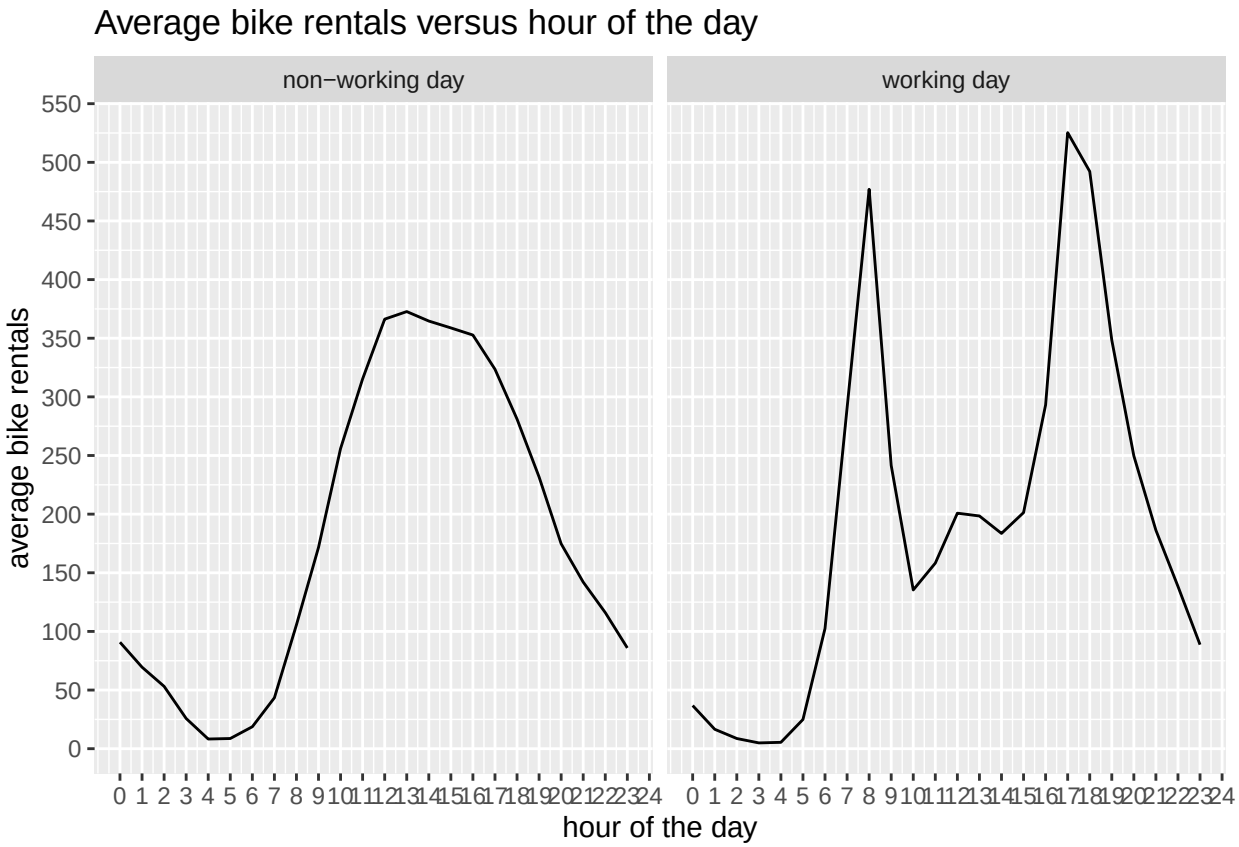


Annotation: This graph showing average bike rentals versus hour of the day.

From this graph, we can know that the average bike rentals from 0 o'clock to 4 o'clock decrease from about 50 to about 5. The average bike rentals from 4 to 8 have increased from about 5 to about 355. From 8:00 to 17:00, the average bike rentals fluctuates, and by 17:00, the average bike rentals reach a peak of about 460. After 17:00, the average number of bike rentals began to decrease and dropped to about 80 at 23:00.

This plot is reliable. Because in the early morning and evening, most people stay at home. Therefore, people's demand for renting bikes decreases, and the average bike rentals will drop. In the morning, people are going to work or school go play outside, and their demand for bikes increases, leading to an increase in average bike rentals. In addition, when people leave school or get off work at 17:00, people's demand for bikes increases again, so the average bike rentals increase.

plot B

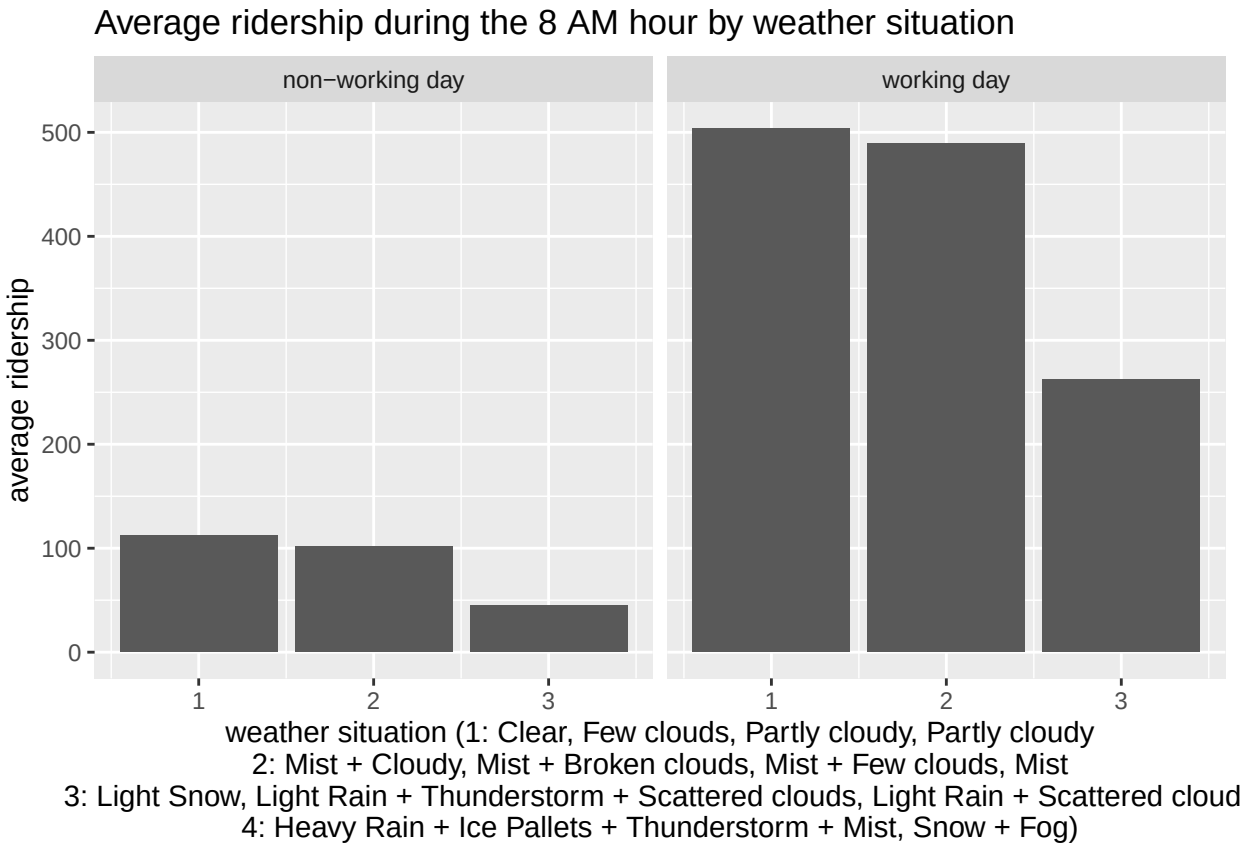


Annotation: This picture shows two different lines. One is the average bike rentals versus the hour of the day during the working day. The other one is the average bike rentals versus the hour of the day during holidays or weekends.

For the weekday line, its trend is very similar to that of plot A, reaching its lowest value at four o'clock and its highest value at 17 o'clock. For the line that is not a working day, it is in a downward trend from 0 o'clock to 4 o'clock, from 4 o'clock to 13 o'clock in an upward trend, and from 13 to 23 o'clock it is in a downward trend again.

This plot is reliable. For the workday line, people start to go to work at 6 o'clock and return home from getting off work between 16:00 and 17:00. At this time, people's demand for bikes reached its peak, and the average bike rentals also reached its peak. For the line that is not a working day, people go out to play at 7 o'clock and go home between 3 o'clock and 4 o'clock, and the average bike rentals reach the highest point.

plot C



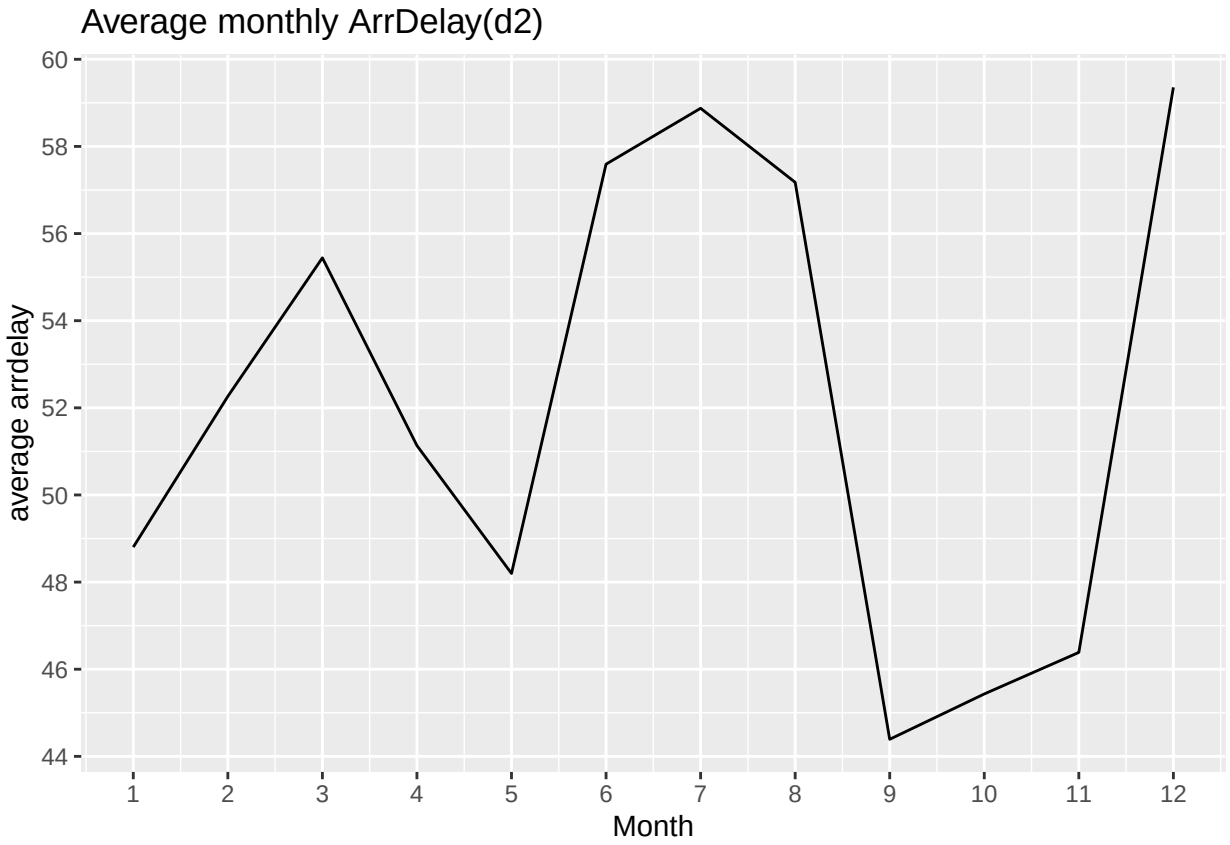
Annotation: This bar plot showing average ridership during the 8 AM hour by weather situation, faceted according to whether it is a working day or not.(1: Clear, Few clouds, Partly cloudy, Partly cloudy. 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist. 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)

From this plot, we can see that the average ridership on working days at 8 o'clock is much higher than the average ridership on non-working days. In good weather conditions, the average ridership on working days can reach around 500, but it can only reach around 100 on non-working days. In addition, from this plot, we can see that the weather has a great influence on ridership. In poor weather conditions, such as Light Snow, Light Rain, Thunderstorm, Scattered clouds, Light Rain, or Scattered clouds, the average people The ridership is reduced by half. In very bad weather conditions, such as Heavy Rain, Ice Pallets, Thunderstorm, Mist, Snow, or Fog, no one rents a bike to travel.

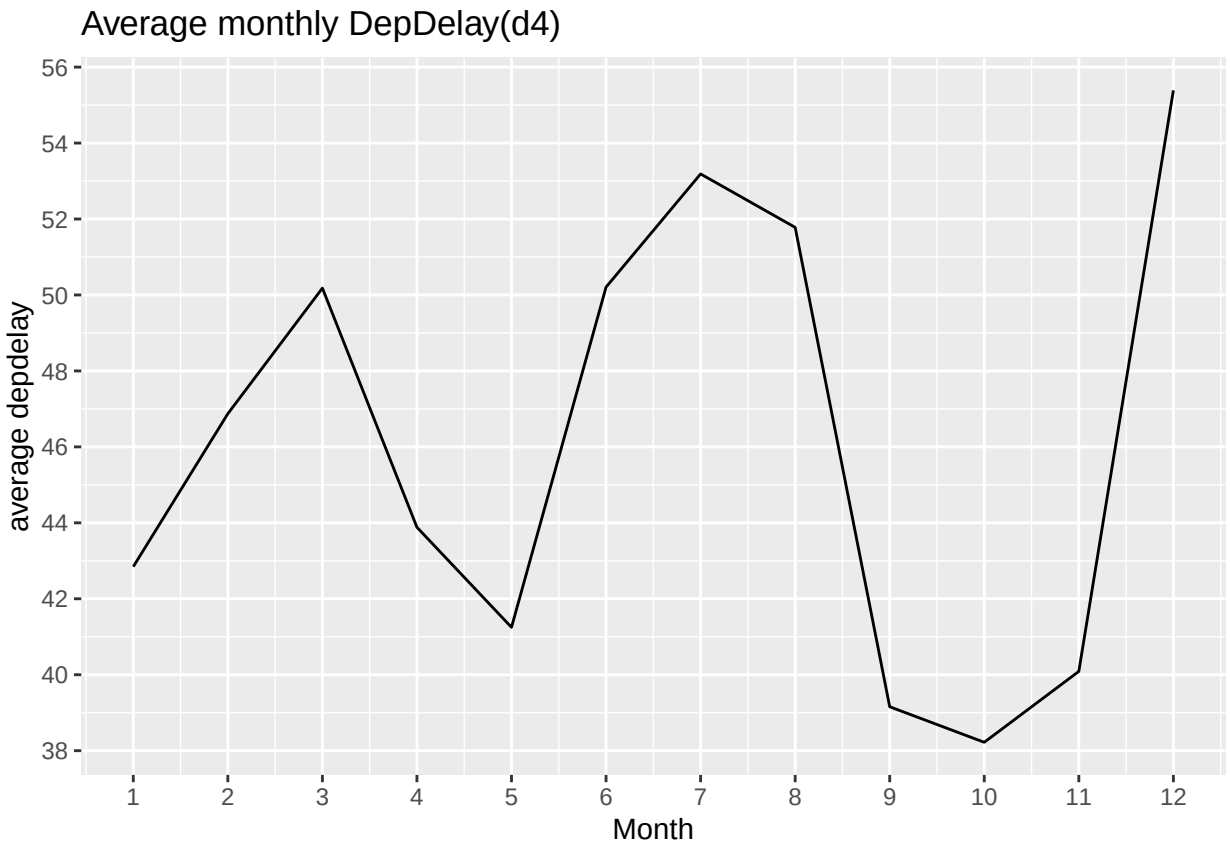
3.Flights at ABIA

```
## # A tibble: 12 x 2
##    Month average_arrdelay
##  * <int>            <dbl>
##  1     1             48.8
##  2     2             52.3
##  3     3             55.4
##  4     4             51.1
##  5     5             48.2
```

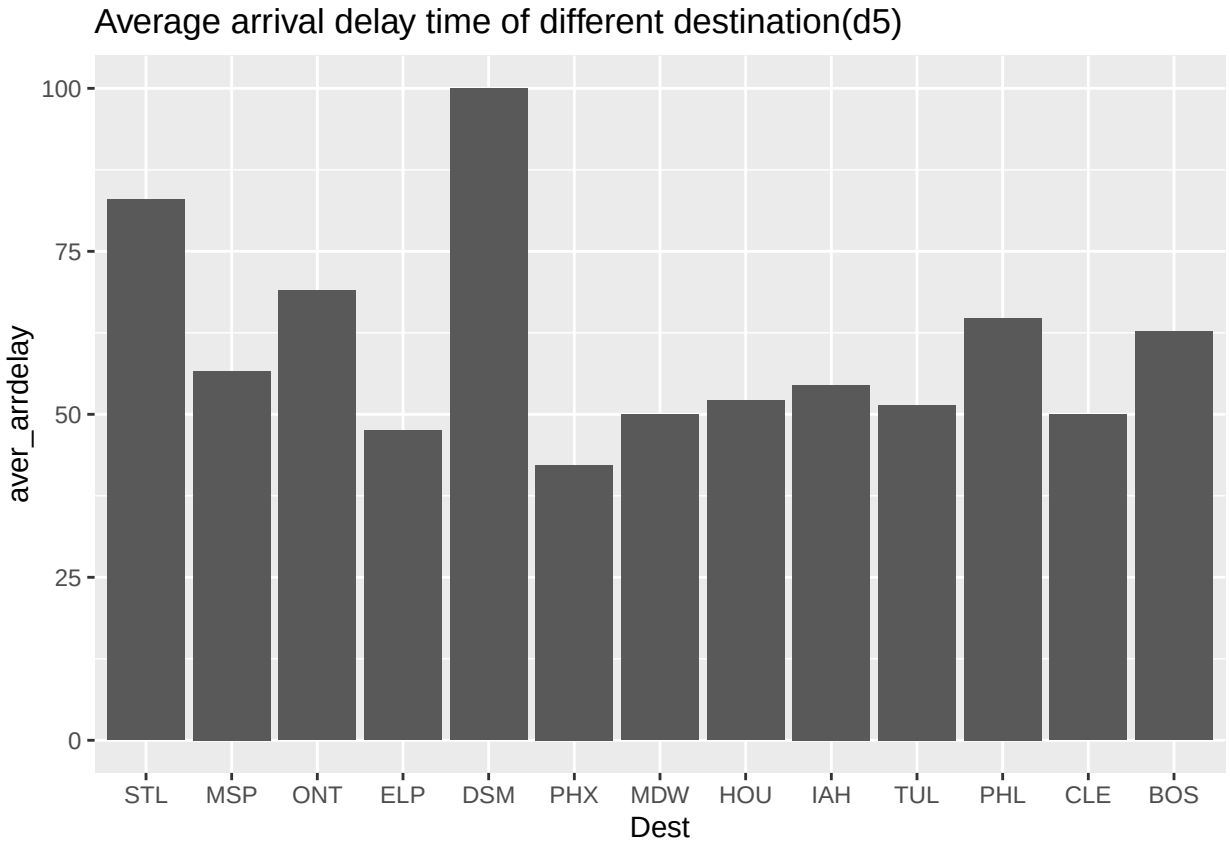
```
##  6     6             57.6
##  7     7             58.9
##  8     8             57.2
##  9     9             44.4
## 10    10             45.4
## 11    11             46.4
## 12    12             59.4
```



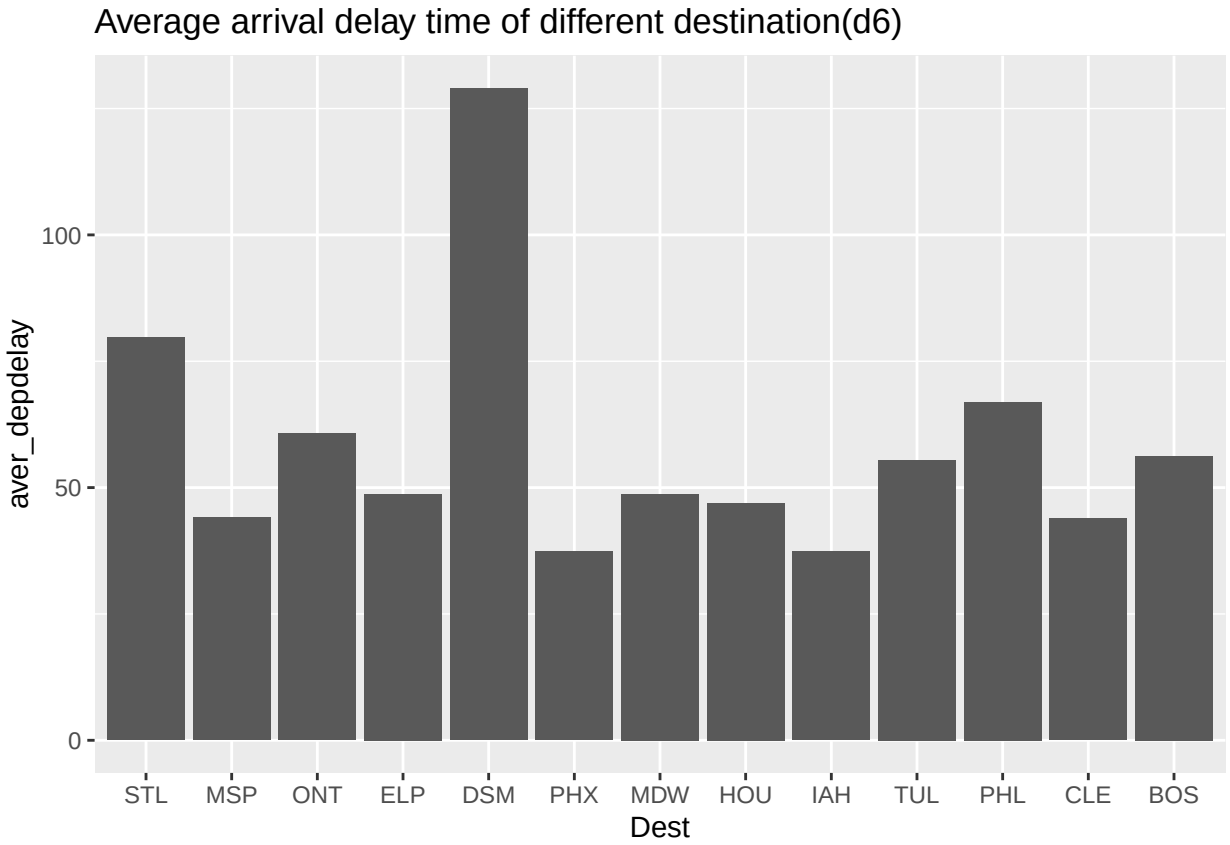

```
## # A tibble: 12 x 2
##    Month average_depdelay
##  * <int>             <dbl>
##  1     1              42.8
##  2     2              46.9
##  3     3              50.2
##  4     4              43.9
##  5     5              41.2
##  6     6              50.2
##  7     7              53.2
##  8     8              51.8
##  9     9              39.2
## 10    10              38.2
## 11    11              40.1
## 12    12              55.4
```



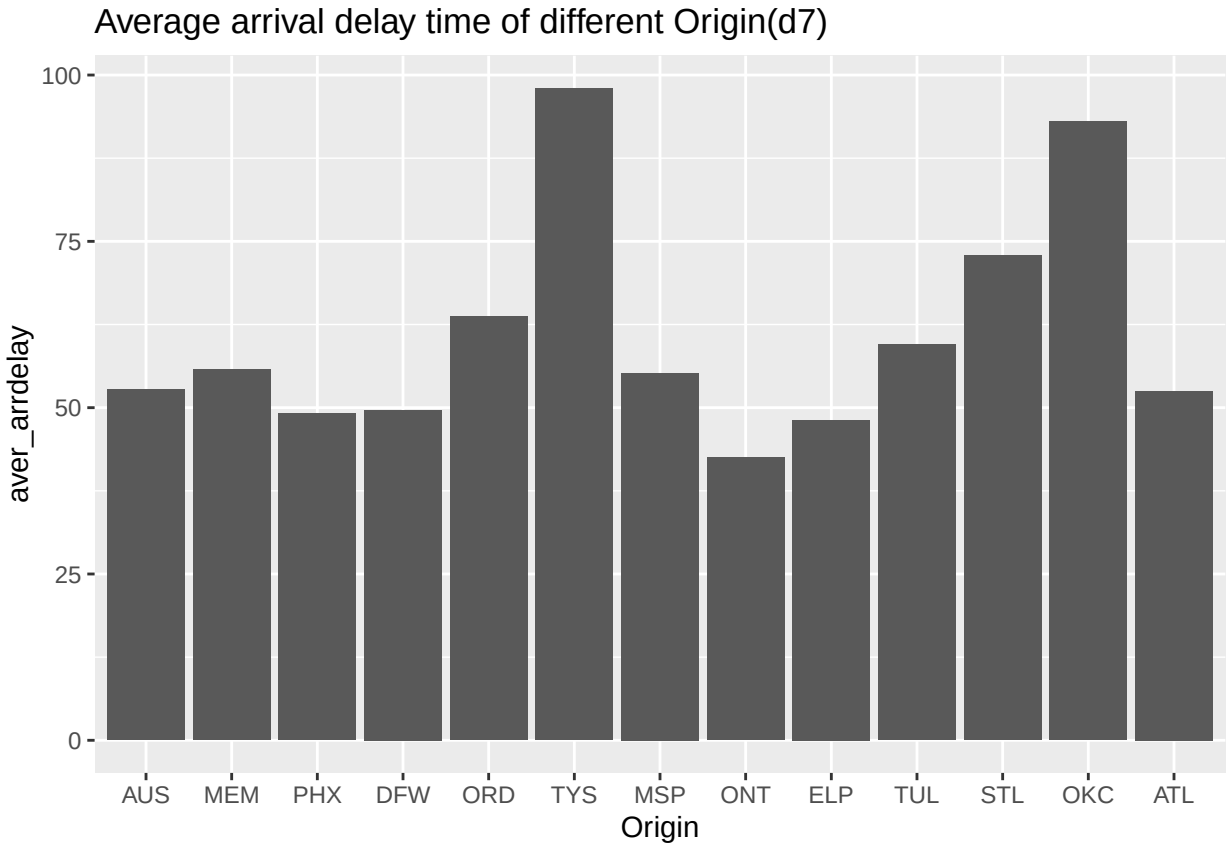

```
## Warning: Removed 38 rows containing missing values (position_stack).
```



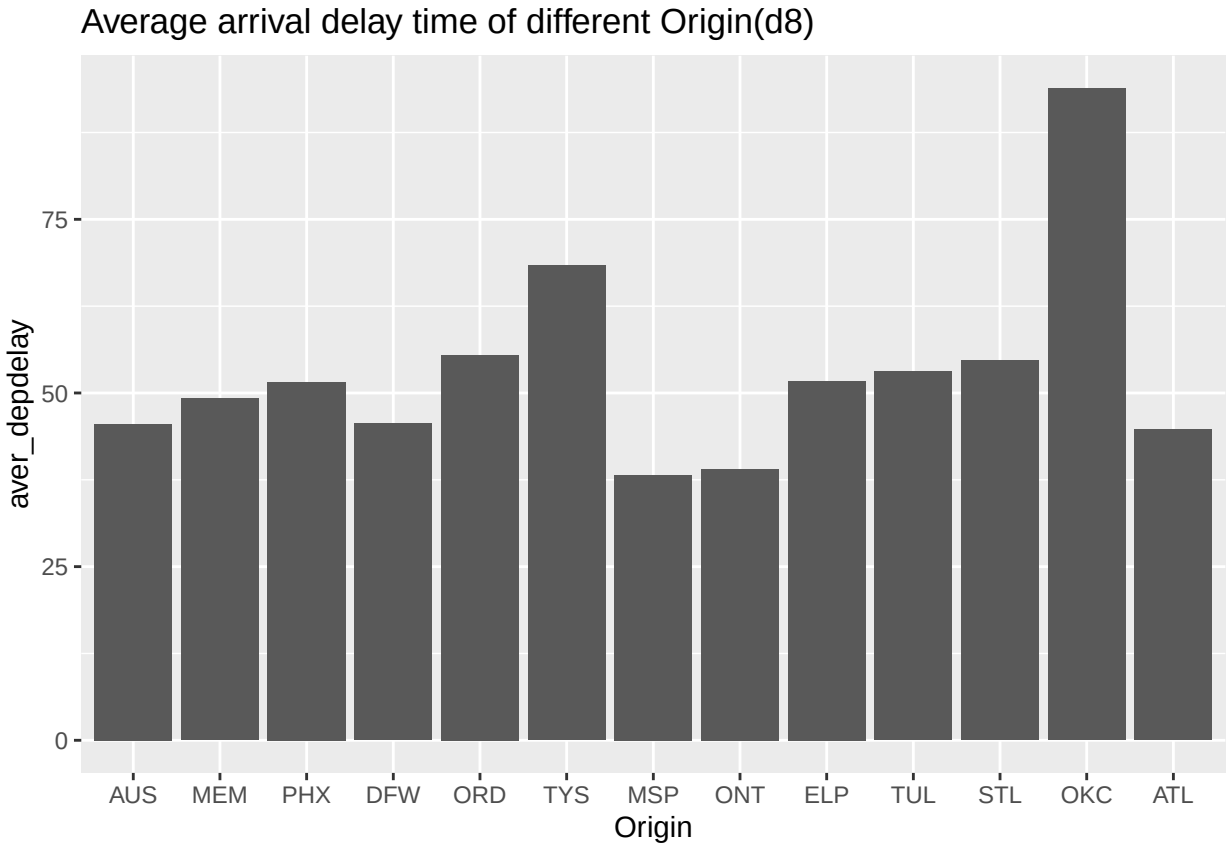

```
## Warning: Removed 38 rows containing missing values (position_stack).
```




```
## Warning: Removed 39 rows containing missing values (position_stack).
```




```
## Warning: Removed 39 rows containing missing values (position_stack).
```



From this figure (d2), it can be seen that September, October, and November are the months with relatively small average arrival delays, which are about 44.5 minutes, 45.5 minutes and 46.5 minutes respectively. Among them, September is the shortest arrival delay time. month. However, July and December are the months when the average arrival delay time is relatively large, about 59 minutes and 59.5 minutes respectively.

From this figure (d4), it can be seen that September, October, and November are the months with relatively few average departure delays, which are about 39 minutes, 38 minutes and 40 minutes, respectively. Among them, October has the least average departure delay time. month. However, July and December are the two months with the highest average departure delay, which is about 53 minutes and 55.5 minutes respectively.

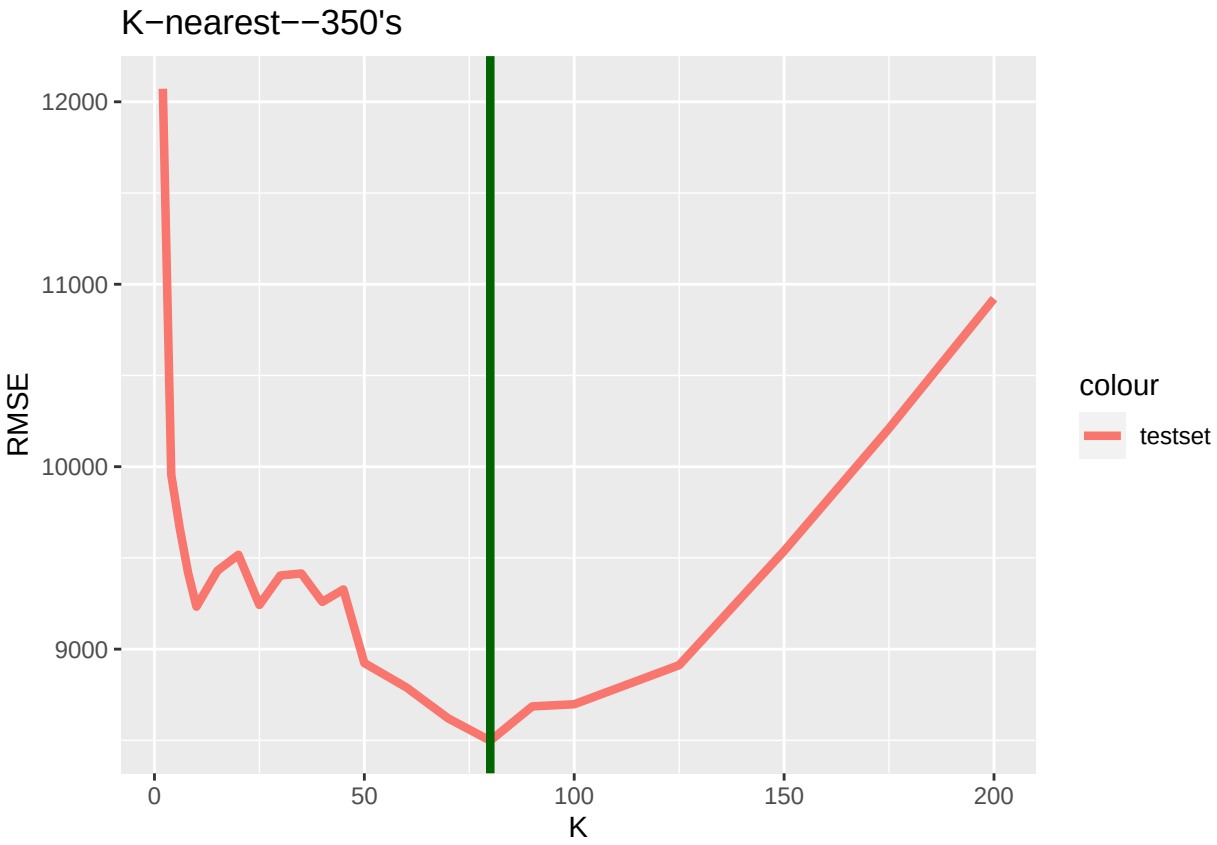
From these two figures (d2,d4), we can infer that the months with the most chance of minimizing the delay time of the year are September and October.

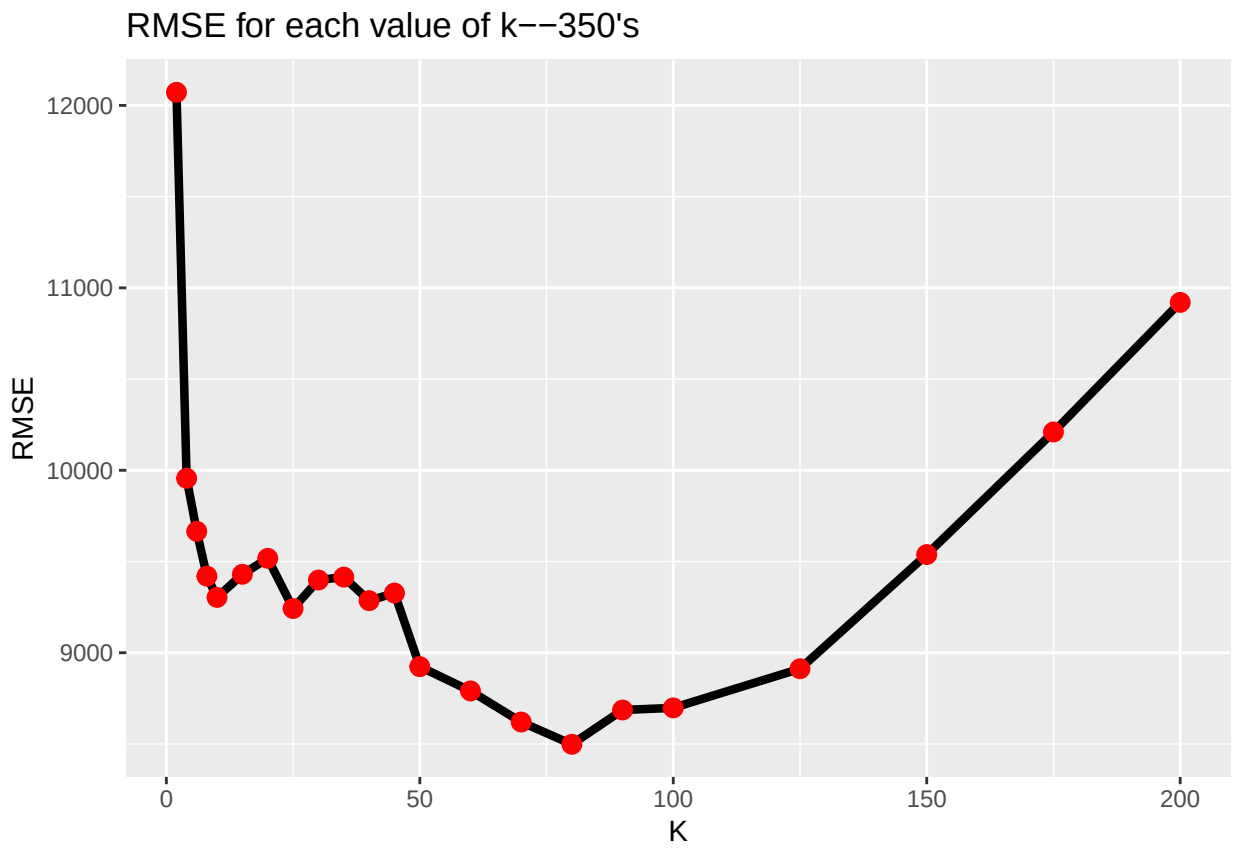
This chart is reliable, because in September and October, the weather will be more stable and sunny throughout the year, which gives the flight a greater chance of taking off and arriving on time. In addition, September and October are the release months of the annual financial statements of most companies in the United States. Most of the workers are at work. This will be part of the reason for the decrease in the number of tourists, and this reason will lead to fewer flights and fewer flights. The impact of air traffic control on flight departure and arrival on time.

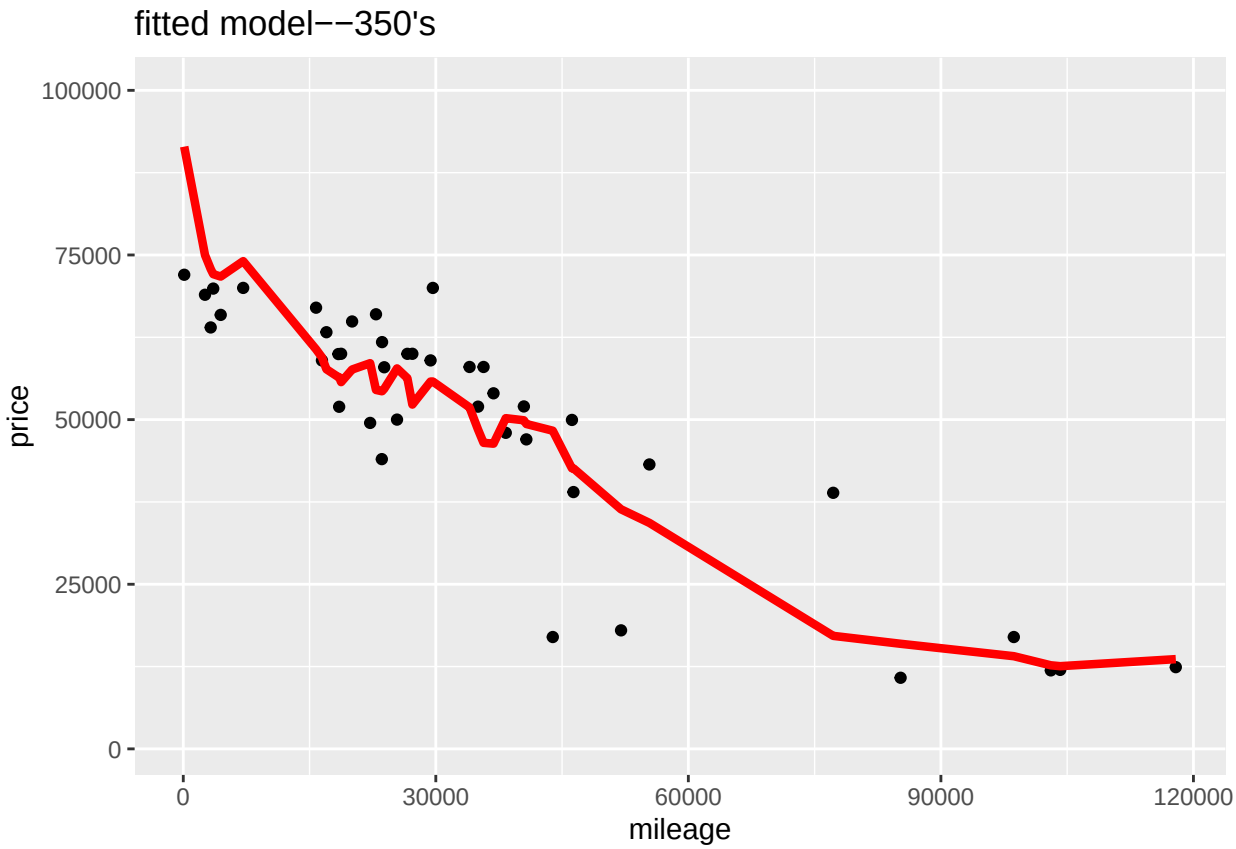
From these four figures (d5, d6, d7, d8), it can be seen that if the origin or destination is different, this will cause the average arrival delay time and the average departure delay time to be different. These four figures are relatively reasonable, because the population of different places is different, which may cause a big difference in the number of people traveling. In addition, different regions have different economic levels and number of attractions. Where the economy is high, the number of flights will be much more, and at the same time, where there are more scenic spots, there will be more flights, which is likely to cause flight delays caused by too many flights.

4 K-nearest neighbors

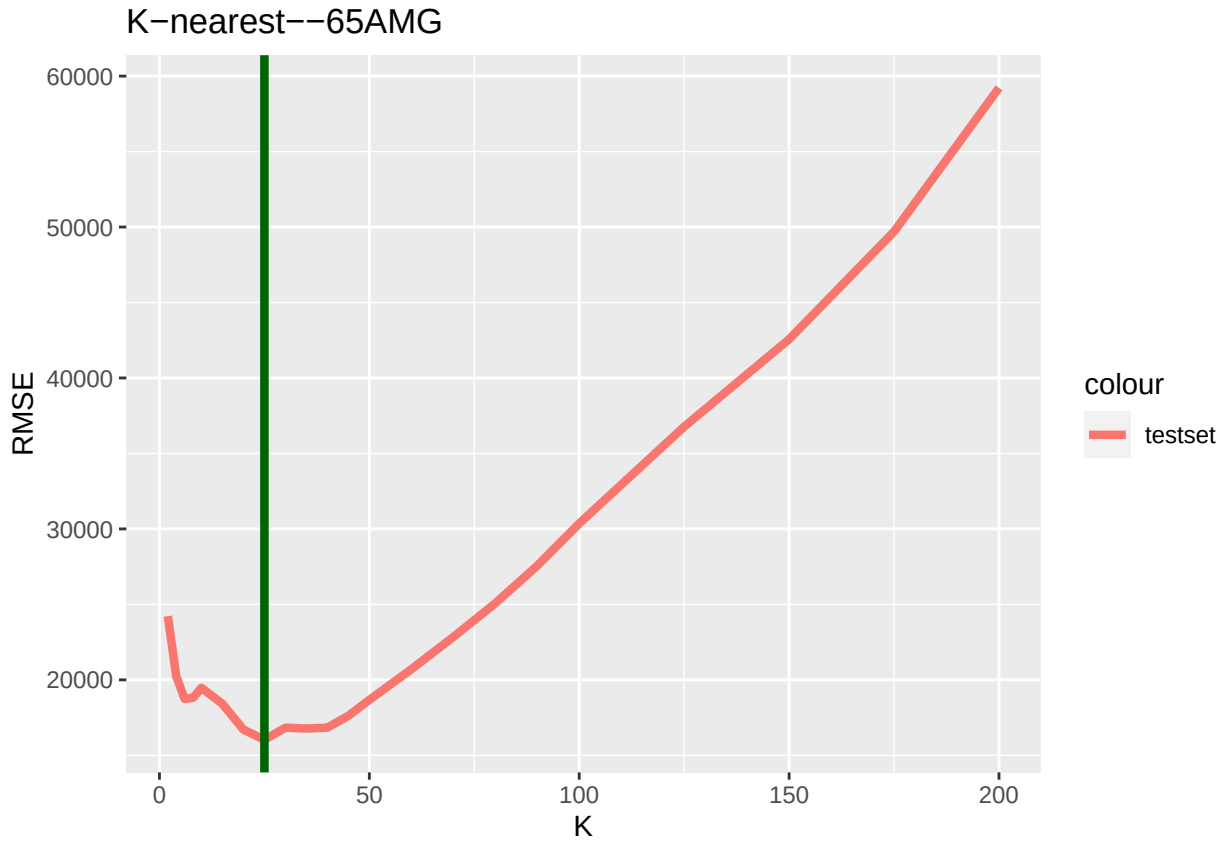
350's

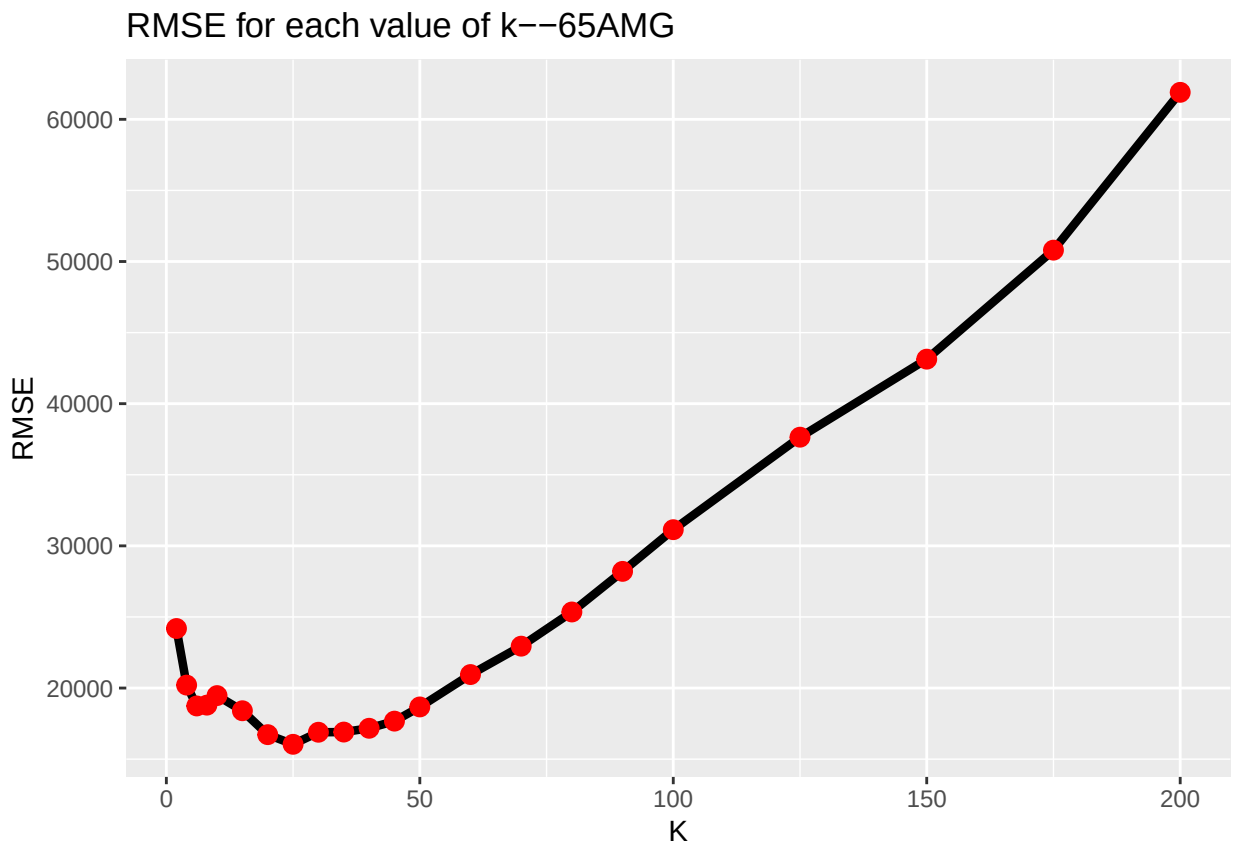


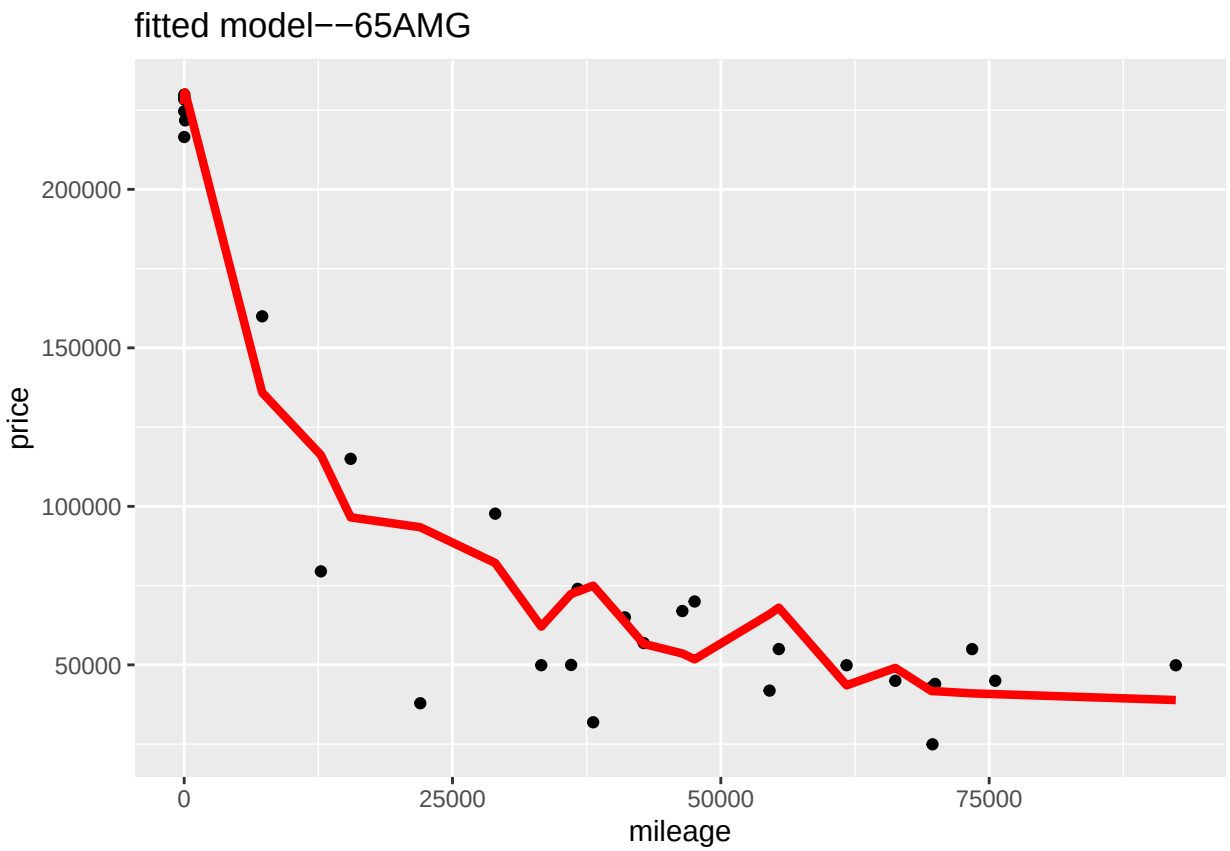

RMSE for each value of k−−350's
12000
11000
10000
9000
RMSE
0
50
100
150
200
K


fitted model−−350's
price
100000
75000
50000
25000
0
0
30000
60000
90000
120000
mileage

65 AMG


K−nearest−−65AMG
60000
50000
40000
30000
20000
RMSE
0
50
100
150
200
K
colour
testset


RMSE for each value of k−−65AMG
60000
50000
40000
30000
20000
RMSE
0
50
100
150
200
K



k_best350

```
## [1] 80
```

k_best65AMG

```
## [1] 25
```

350's yields a larger optimal value of k. We think S-class 350 yields a larger optimal value of K. In the plot of RMSE for each value of K, we can find the best K. The S-class 350 has higher K than S-class 63 AMG. The S-class 63 AMG has lower K value, so it has more chance of memorizing random noise. If there is a random noise, it will cause more bias for prediction. For higher K, it has lower variance and less chance of memorizing random noise. Therefore, For S-class 350 dataset, it has more data points, so it's optimal K can be larger to decrease bias.