ECO 395 Homework 1:Chengkan Tao

chengkan_tao

```r
devtools::install_github(c('rstudio/rmarkdown', 'yihui/tinytex'))
tinytex::install_tinytex()
```

Q1

```r
ggplot(GasPrices, aes(y = Price, x = Competitors)) + geom_boxplot()
```

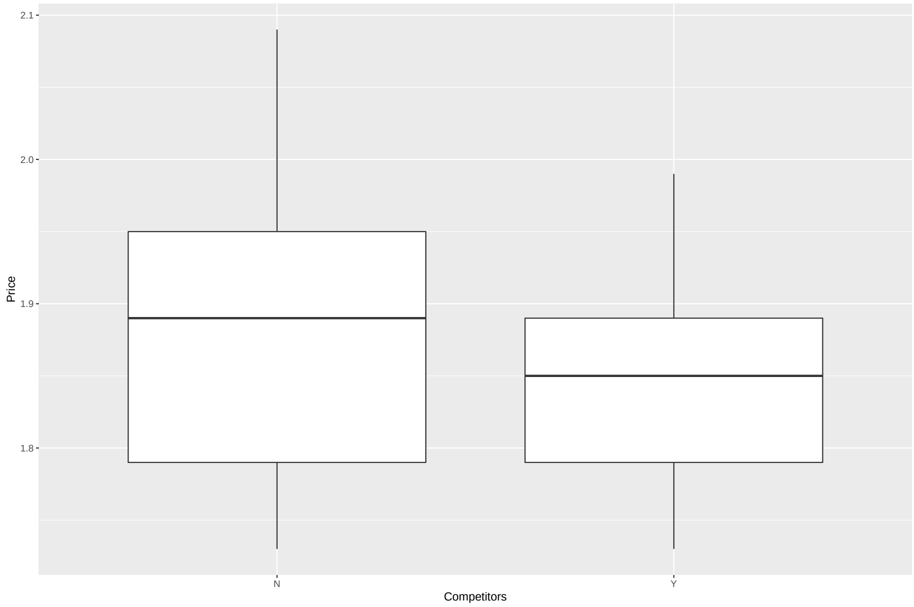

图 1: Q1(1)

The figure shows that price of gas stations with competitor in sight is lower than price of gas stations without competitor in sight on average. From the figure, the median of price of gas stations without competitors in sight is roughly equal to upper quartile of price of gas stations with competitors. The upper quartile and maximum of price of gas stations without competitors are much higher than those who have competitors in sight.

```
ggplot(GasPrices, aes(x = Income, y = Price)) + geom_point()
```

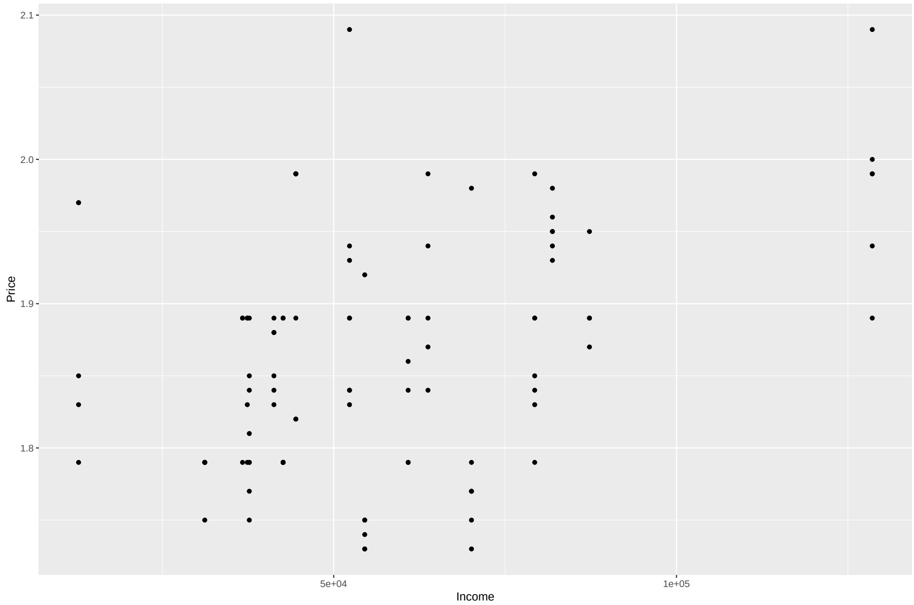

图 2: Q1(2)

The richer the area is, the lower the gas price is. In figure, 5 points in the top right-hand corner shows that gas price is high in areas with median income of 10000. And most points spread over left side of figure correspond to lower prices.

```
Brand_price=aggregate(GasPrices$Price,list(GasPrices$Brand),mean)
ggplot(Brand_price, mapping = aes(x = Brand_price[,1], y = Brand_price[,2])) +
  geom_bar(stat = "identity")+
  geom_text(aes(label = Brand_price[,2]), vjust = 1.5, colour = "White")
```

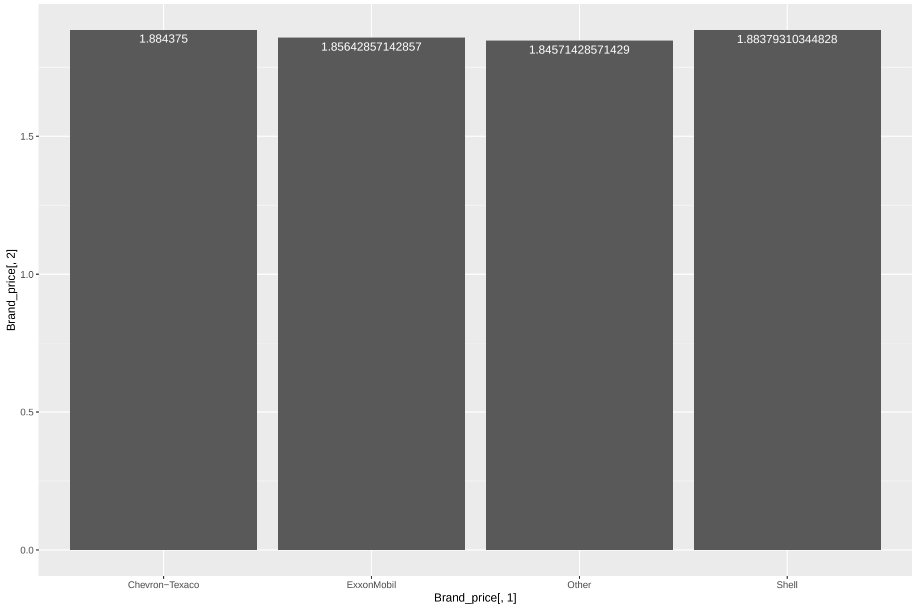

图 3: Q1(3)

Shell doesn’t charge more than all other brands. In figure, Shell charges more than Other and ExxonMobil on average. Shell charges less than Chevron-Texaco on average.

```
ggplot(data=GasPrices)+
  geom_histogram(mapping = aes(GasPrices$Price))+
  facet_wrap(~Stoplight)
```

Many gas stations at stoplights charge more than gas stations without stoplights. In figure, mode of prices in gas stations at stoplights is nearly 1.9. Mode of prices in gas stations without stoplights is nearly 1.8. Nearly half of prices in gas stations at stoplights range from about 1.8 to 1.9

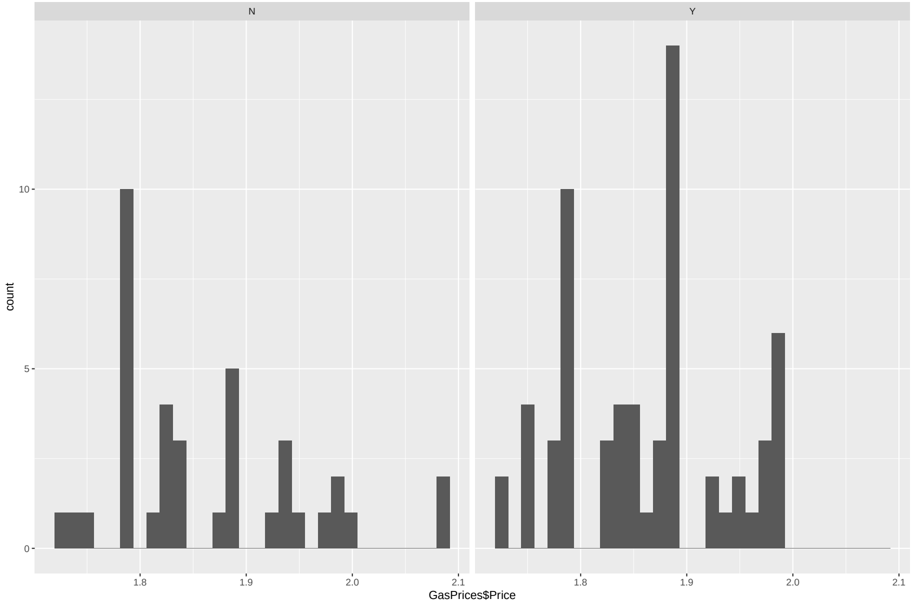

图 4: Q1(4)

```
ggplot(GasPrices, aes(y = Price, x = Highway)) + geom_boxplot()
```

Gas stations with direct highway access charge more Price of gas stations with direct highway access is higher than price of gas stations without highway access on average. In figure, the median of price of gas stations with direct highway access is roughly equal to upper quartile of price of gas stations without highway access. The upper quartile and maximum of price of gas stations without highway access are lower than those who have highway access.

Q2

```
plot(y=average_bike_rentals$x,x=average_bike_rentals$Group.1)
lines(y=average_bike_rentals$x,x=average_bike_rentals$Group.1)
```

In figure, we can find relationship between hour of a day and bike rentals.

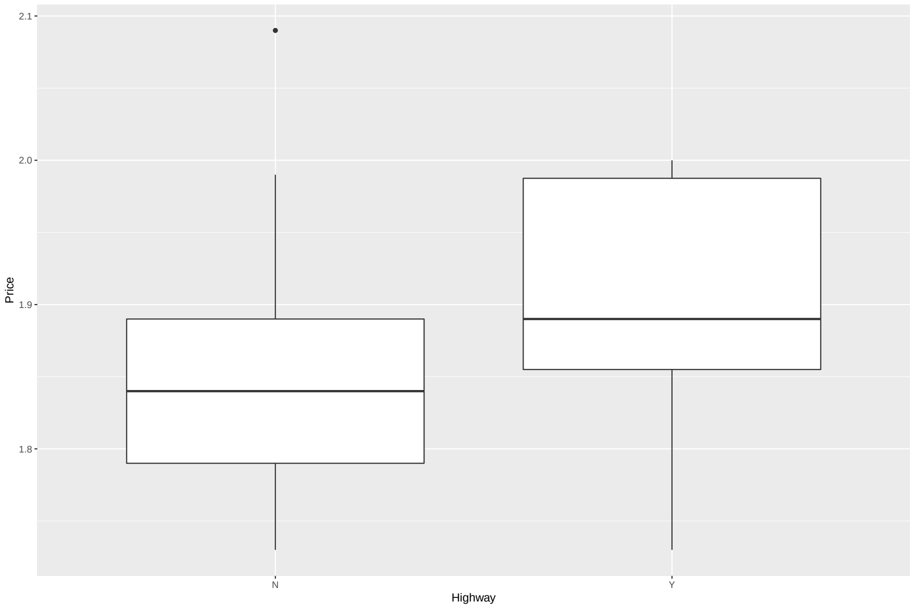
2.1
2.0
1.9
1.8
Price
N
Y
Highway

图 5: Q1(5)

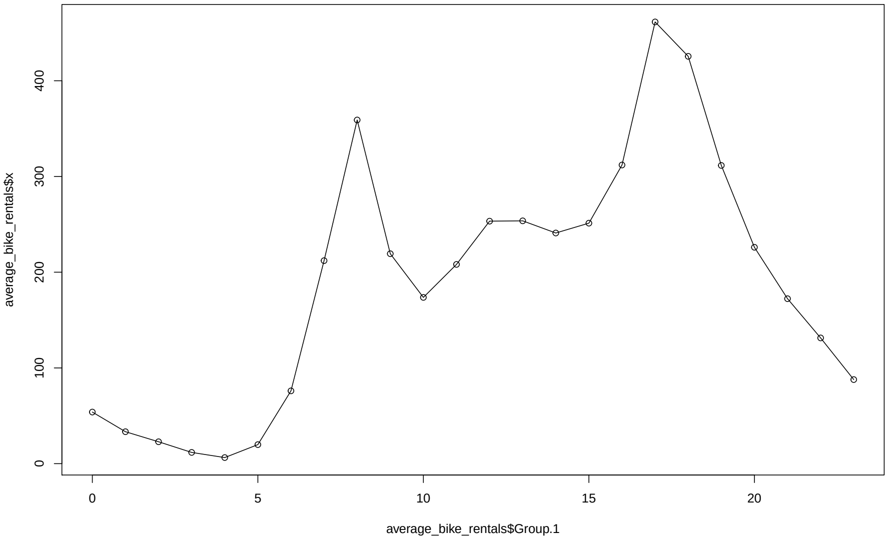
0
100
200
300
400
average_bike_rentals$x
0
5
10
15
20
average_bike_rentals$Group.1

图 6: Q2(1)

The bike rental is sampled every hour. The y-axis is average bike rental at each time point, and x-axis shows 24 time points. At night, bike-sharing rental demand is low on average. After 12 o'clock, it's under 100. The demand rises obviously from about 50 to nearly 65_AMG in the morning rush hour. After the peak, rental demand remains around 250. The maximum appears at 17 p.m.

```r
ggplot(total.hr) +
  geom_line(aes(x=hr, y=total.average)) +
  facet_wrap(~workingday)
```

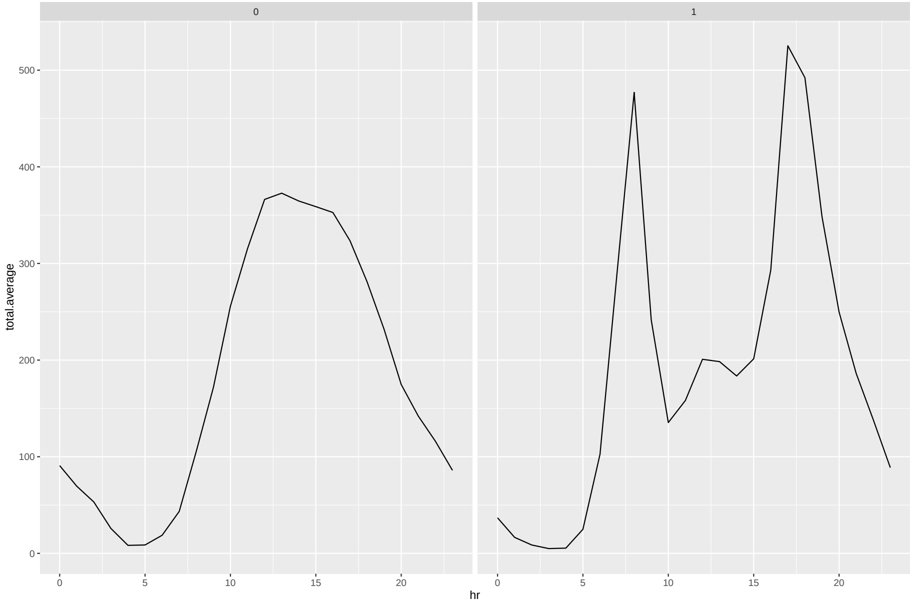

图 7: Q2(2)

The figure shows bike-sharing rental demand every hour, faceted according to whether it is a working day. The x-axis and y-axis are the same as before. Right picture shows the relationship on weekends and left one shows relationship on working days. On working days, two peaks mentioned before, morning rush hour (6:00-9:00) and evening rush hour (16:00-19:00), are shown in the picture. At noon, the demand ranges from 150 to 200.

On weekends, two peaks disappear. The average bike rental grows from morning to noon and then drops as time goes on.

```r
ggplot(total.weathersit) +
  geom_line(aes(x=weathersit, y=total.average)) +
  facet_wrap(~workingday)
```

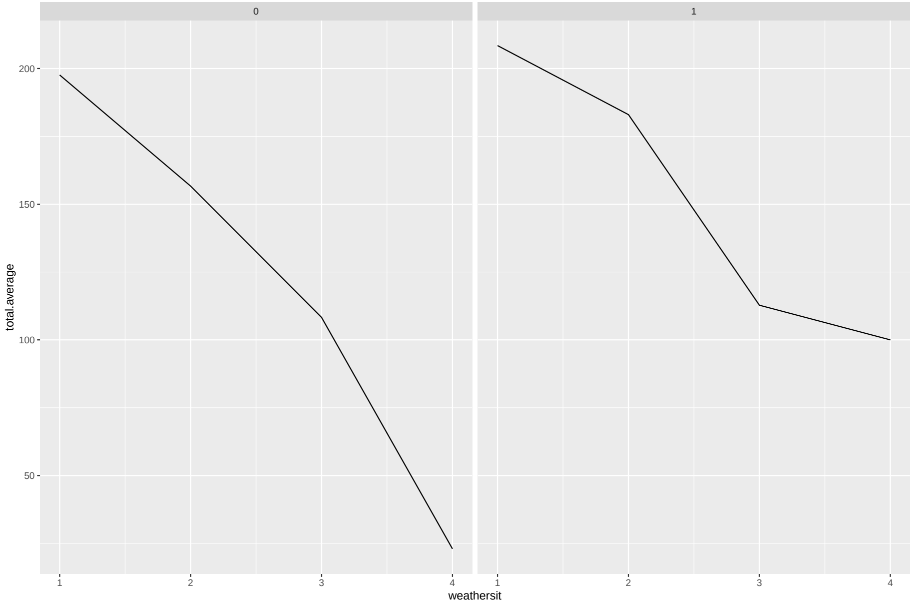

图 8: Q2(3)

The figure shows average ridership in different weather during the 8 a.m. hour, faceted according to whether it is a working day. The x-axis shows 4 kinds of weather. 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 The y-axis shows average bike rentals. In figure, the worse the weather, the less bikes are rented during the 8 a.m. hour on both working days and weekends. The demand on working days is still higher than demand on weekends in different weather.

#Q3

```
ggplot(DepDelay.UniqueCarrier,aes(x=UniqueCarrier, y=DEPdelay.average)) +
  geom_boxplot()
```

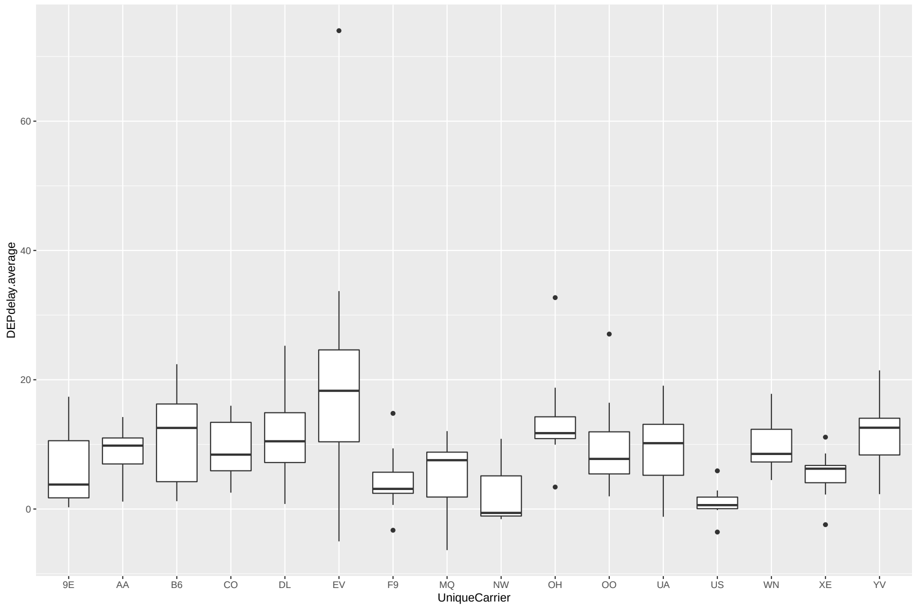

图 9: Q3(DepDelay)

```
ggplot(ArrDelay.UniqueCarrier,aes(x=UniqueCarrier, y=ARRdelay.average)) +
  geom_boxplot()
```

The figure shows average departure delay in minutes of each unique carrier per month. The x-axis shows all unique carriers, and y-axis shows average departure delay per month. The average departure delay of EV is the longest because its upper quartile, mean, and maximum are the highest. Relatively speaking, NW and US have the least average departure delay.

The second figure shows average arrival delay of each unique carrier per month. The x-axis shows all unique carriers, and y-axis shows average arrival delay per month. The average arrival delay of EV and OH is the longest. 9E and US are relatively on time.

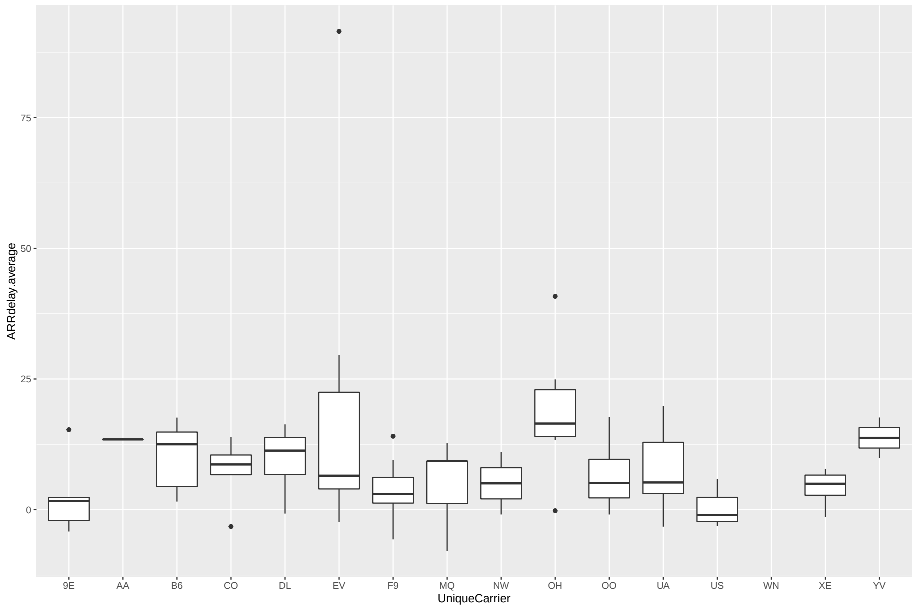

图 10: Q3(ArrDelay)

#Q4

```
## [1] 11668.64
```

```
## [1] 467.2826
```

```
ggplot(sclass_grid) +
  geom_point(aes(x=k, y=err)) +
  geom_errorbar(aes(x=k, ymin = err-std_err, ymax = err+std_err)) +
  scale_x_log10()
```

```
## [1] 24756.22
```

```
## [1] 1511.418
```

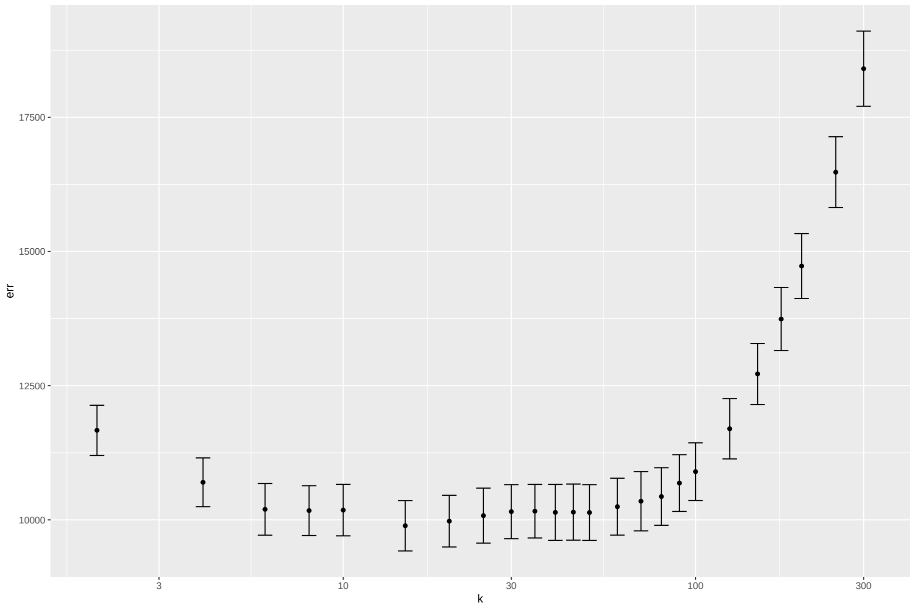

图 11: Q4(sclass_350)

```r
ggplot(sclass_grid) +
  geom_point(aes(x=k, y=err)) +
  geom_errorbar(aes(x=k, ymin = err-std_err, ymax = err+std_err)) +
  scale_x_log10()
```

K = c(2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, 300) The second picture shows estimated out-of-sample root mean-squared error(cross-validated error rate) for each value of through K-fold cross validation in sclass-350. When K=15, mean of RMSE is the lowest.

The third picture shows estimated out-of-sample root mean-squared error(cross-validated error rate) for each value of through K-fold cross validation in sclass-350. When K=25, mean of RMSE is the lowest.

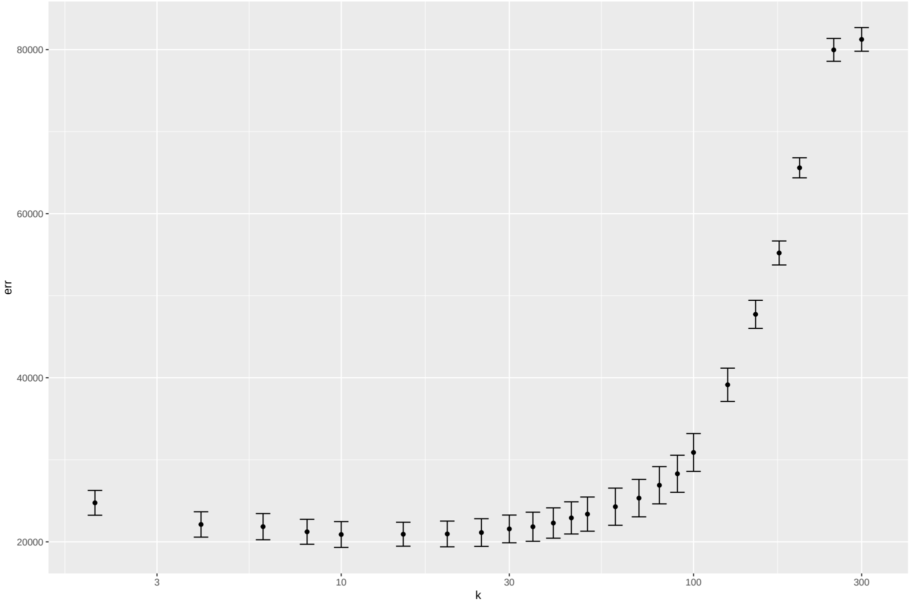

图 12: Q4(65AMG)