HW1 (teenage_gamble and textile production)

108048110

2022-10-02

Assignment 1

You should present your results in summary format only with R outputs that are required to support your answers.

1. Britain Teenage Gambling data

a.

This this a survey conducted to study teenage gambling behavior in Britain and the data contains 47 rows and 5 columns.

To see what variables we have and what sort of values they take.

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

Table 1: Britain Teenage Gambling Data

Items	Variables	Description
1	Sex	Gender, 0=male, 1=female
2	Status	Socioeconomic status score based on parents' occupation
3	Income	How much does a teen earn pounds per **week**
4	Verbal	Verbal score in words out of 12 correctly defined
5	Gamble	Teenagers spend on gambling pounds per **year**

Expect:

1. **Sex** to be a qualitative nomial variable;
2. **Status** to be Quantitative Discrete variable;
3. **Income** to be Quantitative Continuous variable;
4. **Verbal** to be Quantitative Discrete variable;

5. **Gamble** to be Quantitative Continuous variable.

Get usual univariate summary info

```
##       sex             status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

Error entry check.

- Seems to have no missing values.

```
##     sex            status          income          verbal
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:47        FALSE:47        FALSE:47        FALSE:47
##    gamble
##  Mode :logical
##  FALSE:47
```

Variable1: Sex

According to the data description provided, sex variable is categorized into to classes, **0** and **1** represents **male**** and **female** respectively.

Since I expect sex to be qualitative nomial variable, I designated it as factors and use descriptive labels to presented them.

- There are 28 men and 19 women.
- The researcher have more data from males than from females.

```
##
##   male female
##     28     19
```

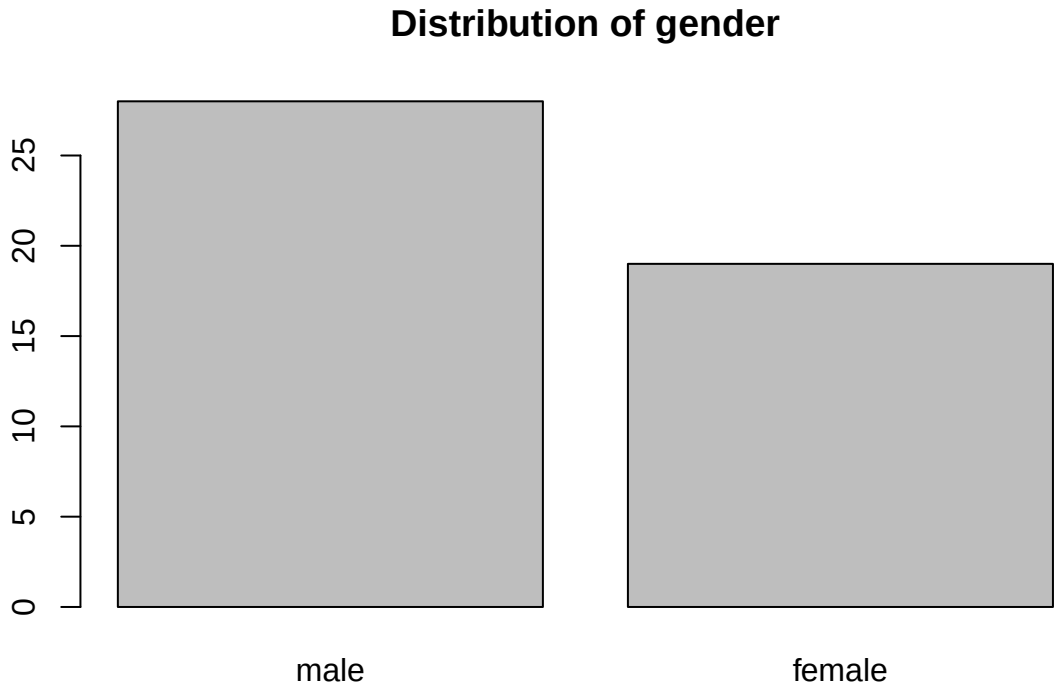
Sex ratio (from the Sociology perspective)

```
## [1] 0.6785714
```

Percentage Distribution of the variable.

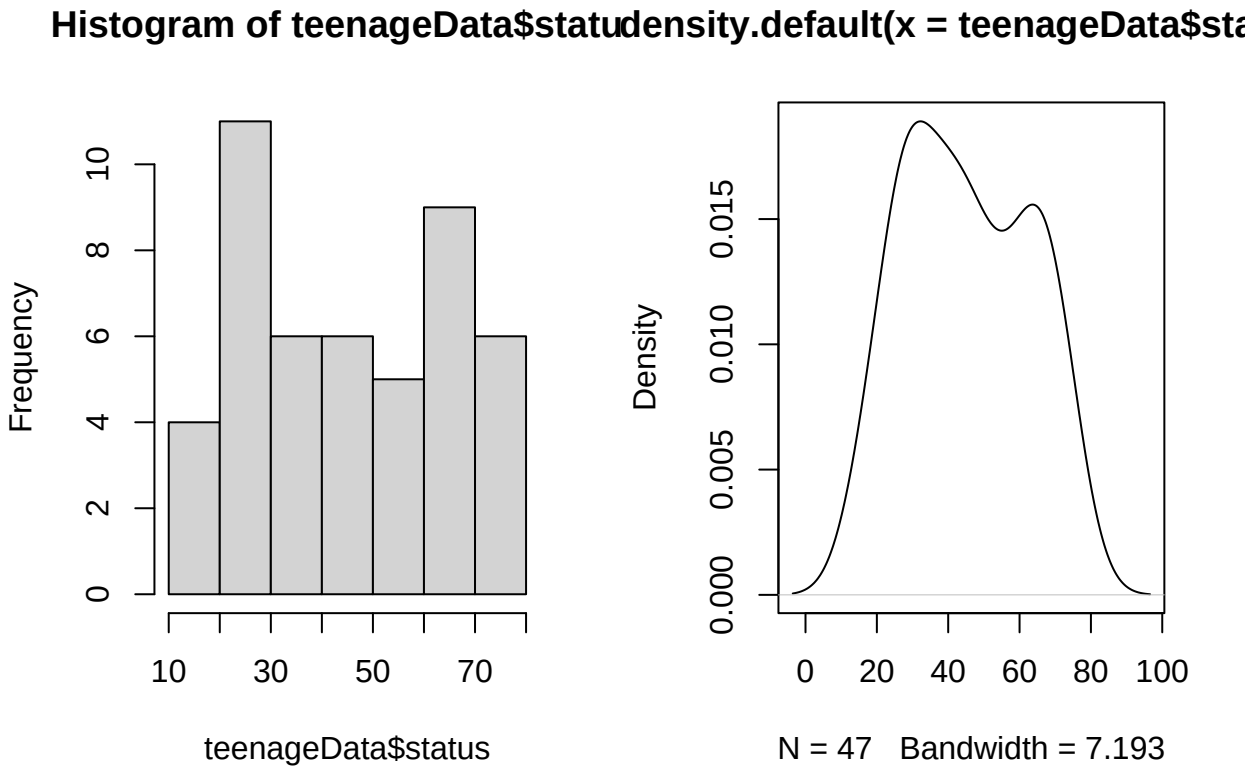
```
##
##      male    female
## 0.5957447 0.4042553
```

Visualization

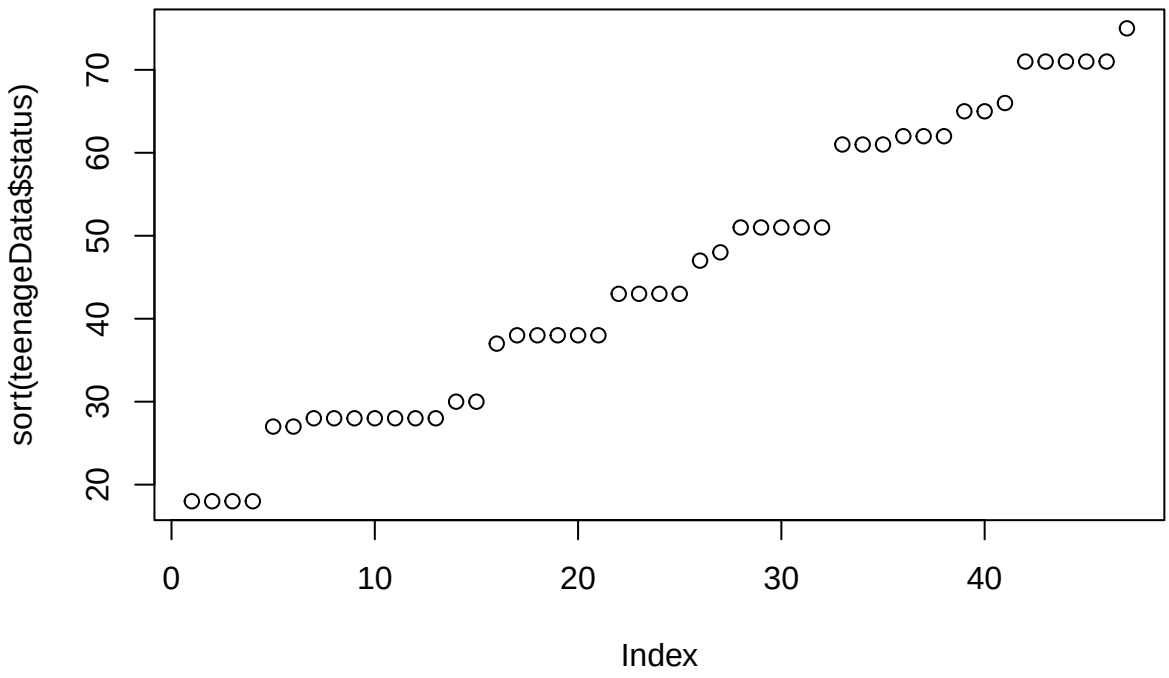


Variable2: Status

Interesting about this variable, since the information about the purpose of the study and how the data were collected are not provided, and the data description only mentioned that variable ***status*** is a score based on a teenager's parents' occupation, I could only comment it by my intuitive perception.



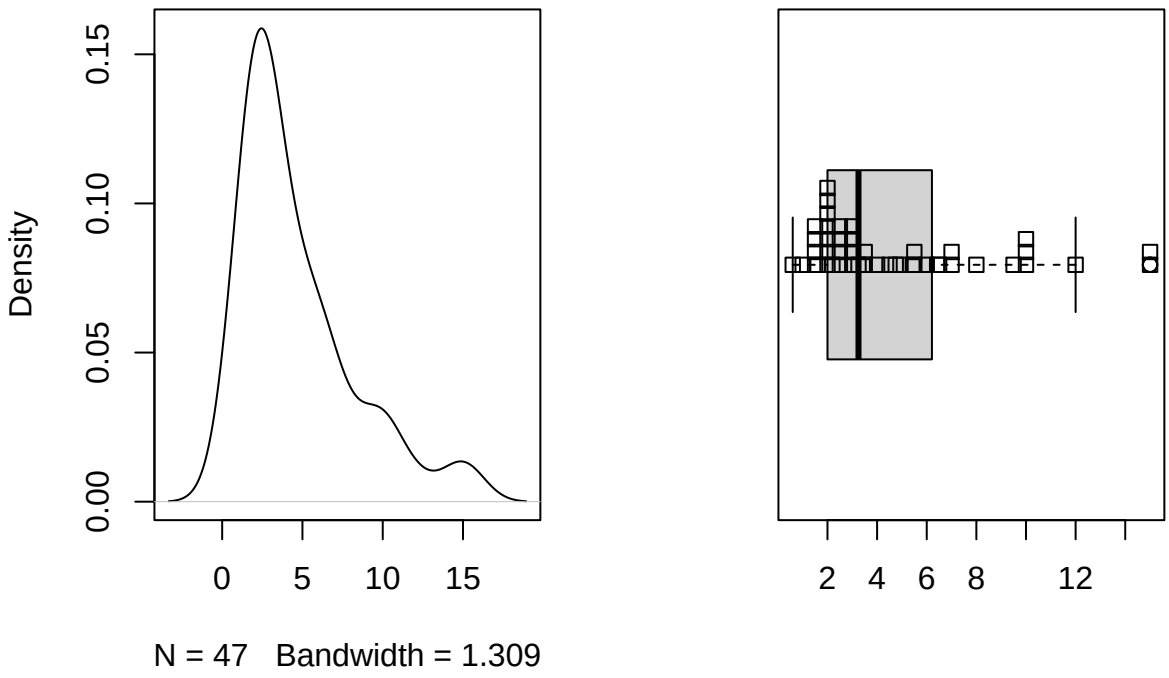
Histogram might be misleading since people can easily manipulate the outcome by setting different bandwidth, thus I would consider a kernel density plot a better choice.



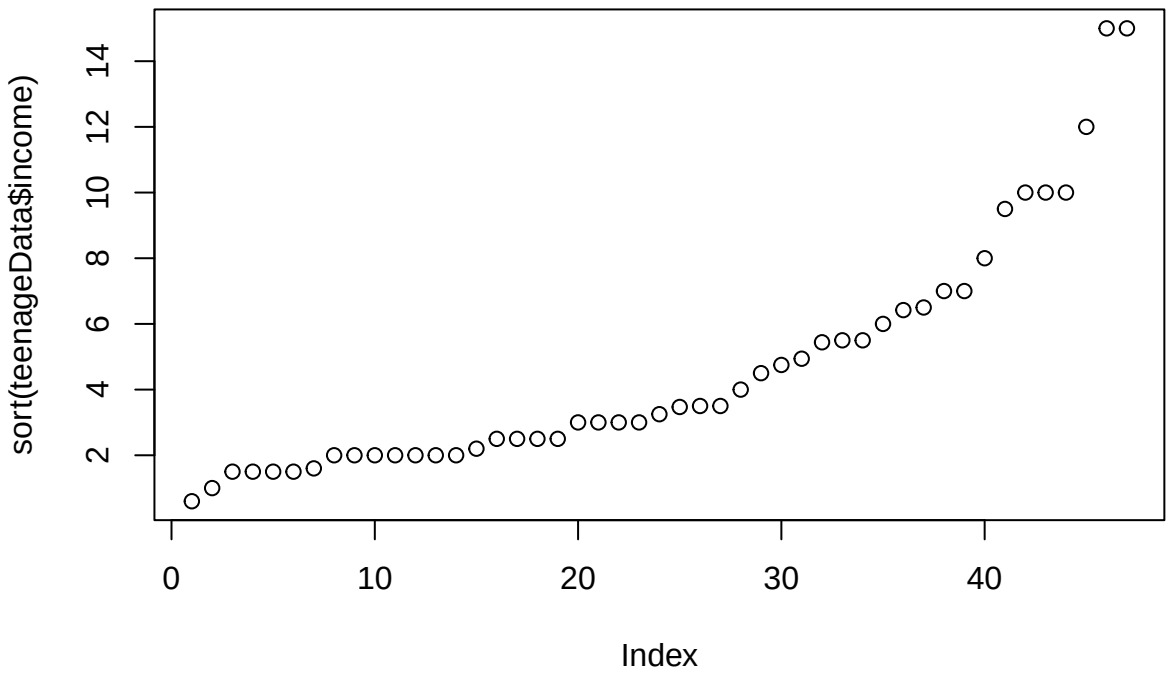
- There is no obvious outliers lied in the plots.

Variable3: Income

Apparently, the phrase "I earn twice more than you do." is hurtful when it is literally true, I would classified the variable as a quantitative ratio variable because the differences between samples have meanings.

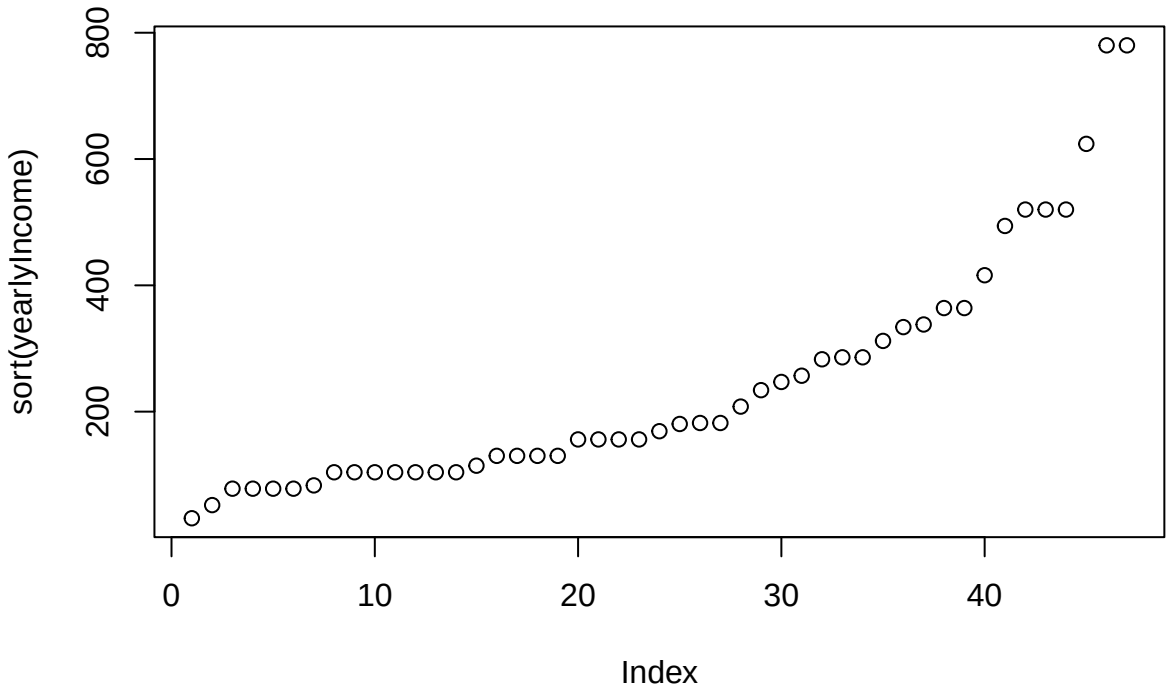


- It seems that someone earned a lot. Lets dig in.

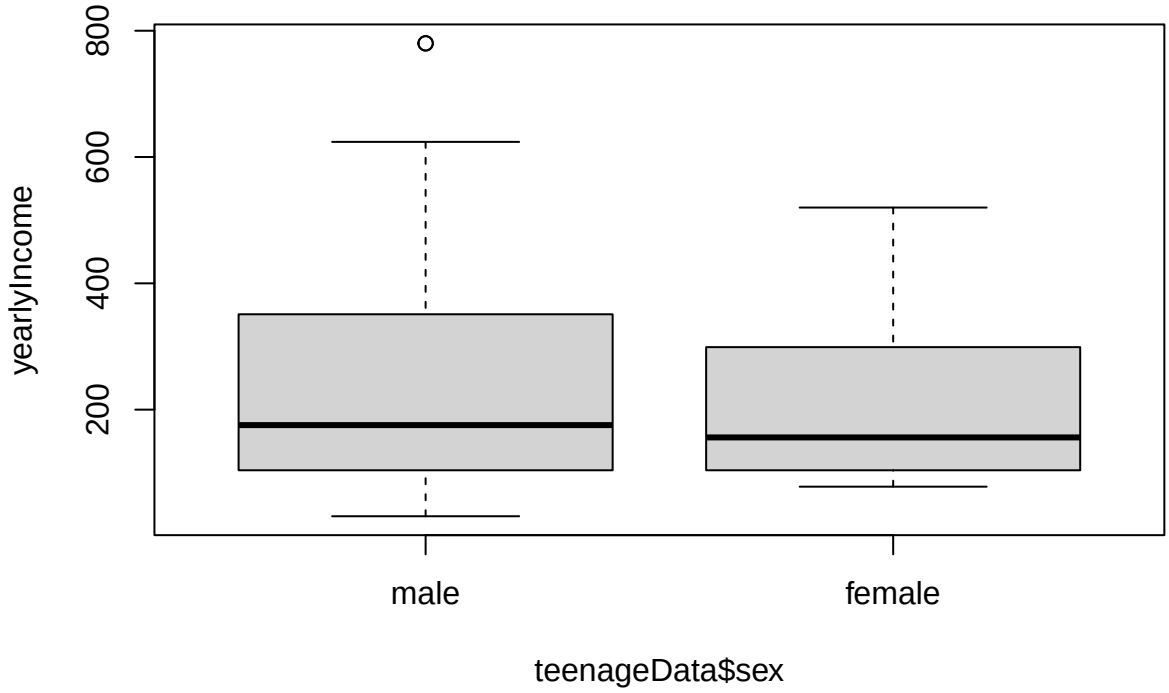


I'd like to analyze the relationships between income and gambling, but two variables are recorded with distinct units. So I unified the units of income (pounds/week) to (pounds/year) by multiplying it by 52. (weeks/year)

- Though it didn't seem much different, the range between them actually are larger.

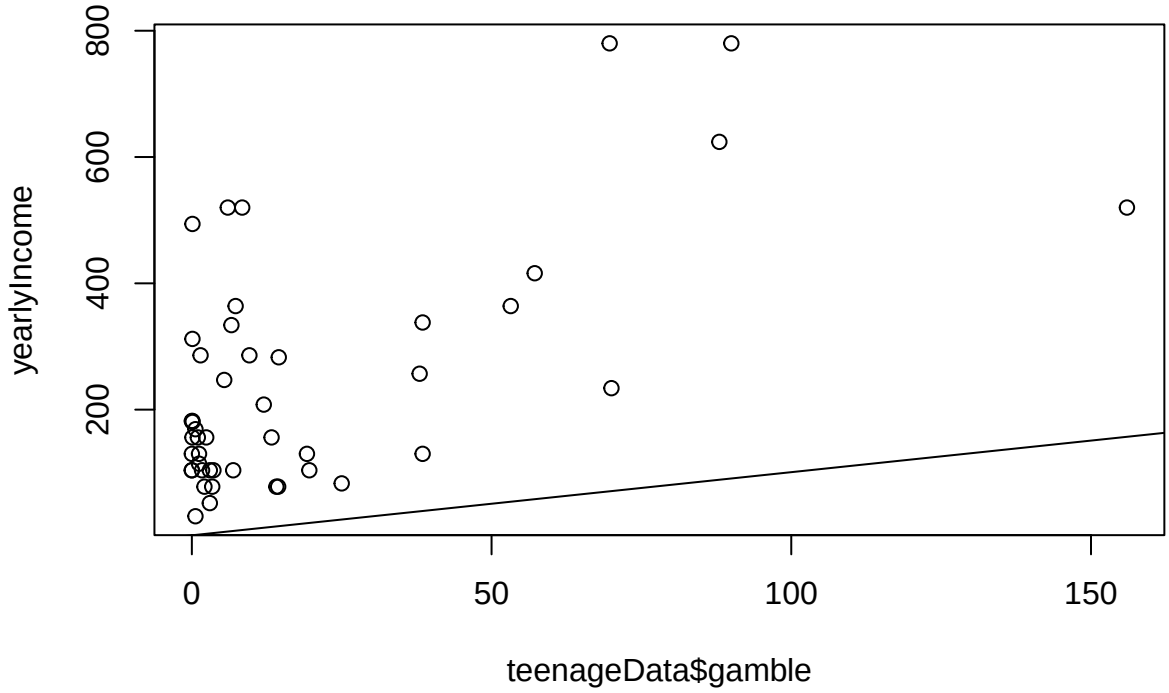


Curious about how gender is related to income, I visualized it and found that in the data, average female income do seem less than male.



And the relationships between gamble and income.

- Seems like these teenagers had good capability of self-controlling. They did not spent all of their money on gambling, good kids.

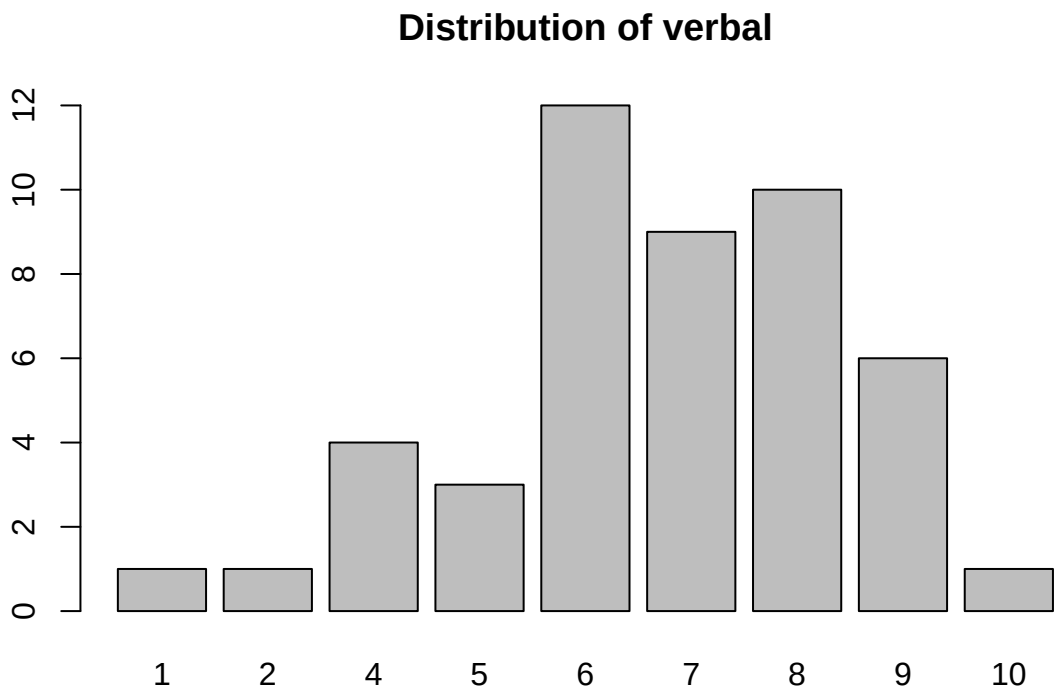


Variable4: Verbal

Since verbal score were calculated by how many words a teenager can correctly define, they are assuredly discrete values.

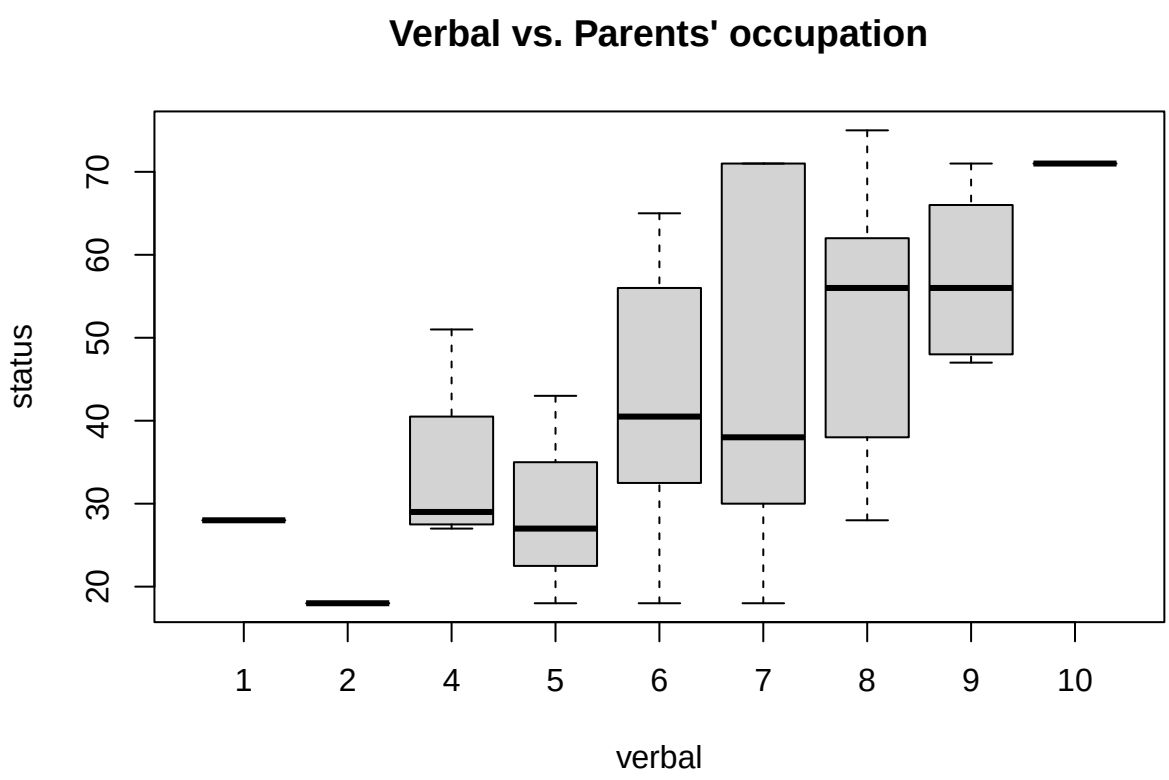
```
##
##  1  2  4  5  6  7  8  9 10
##  1  1  4  3 12  9 10  6  1
```

Designated and gave it descriptive labels.



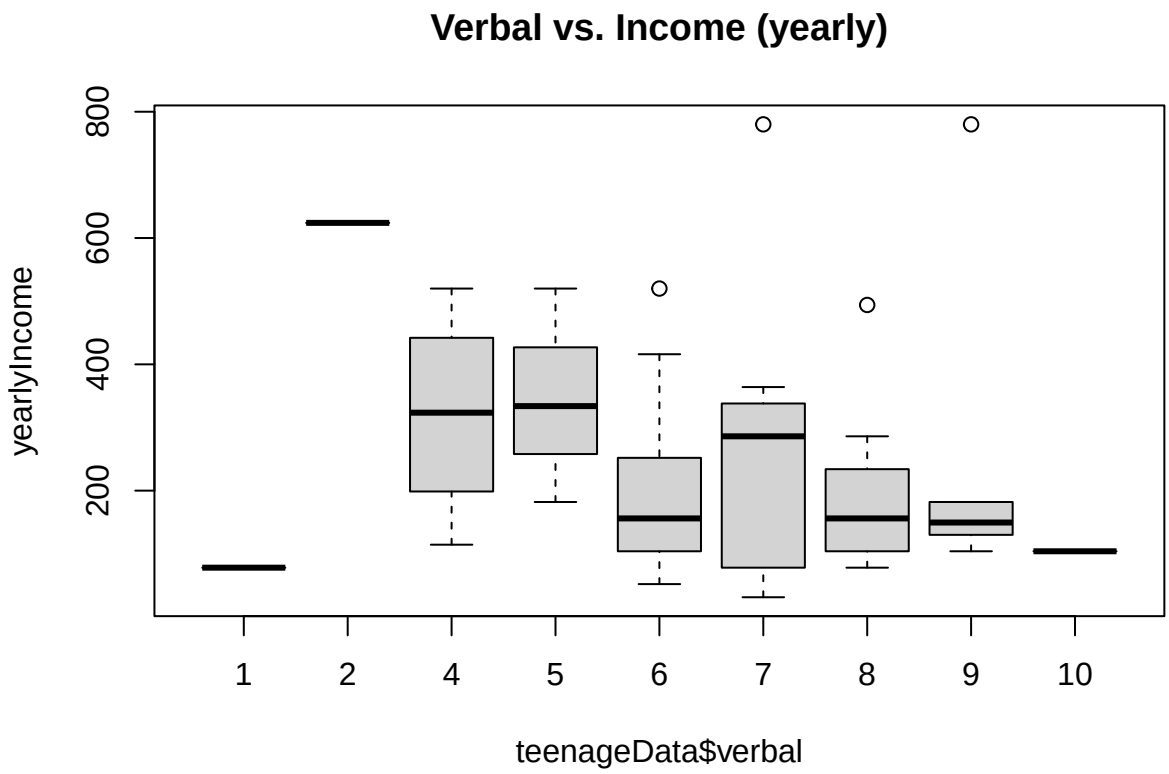
See if knowing more words can bring to higher social status or higher social status means inferring teens recognized more words.

- They might have some association relationships.
- But it seems that higher social status mostly are related to higher verbal scores.



See if knowing more words means that you earn more.

- Weirdly, I expect that students memorizing more vocabularies would result in higher income, but it seems quite the opposite.
- Though the blank on the top left corner may indicate that if a teen wanted a job with better salary, he would need to work harder in school.

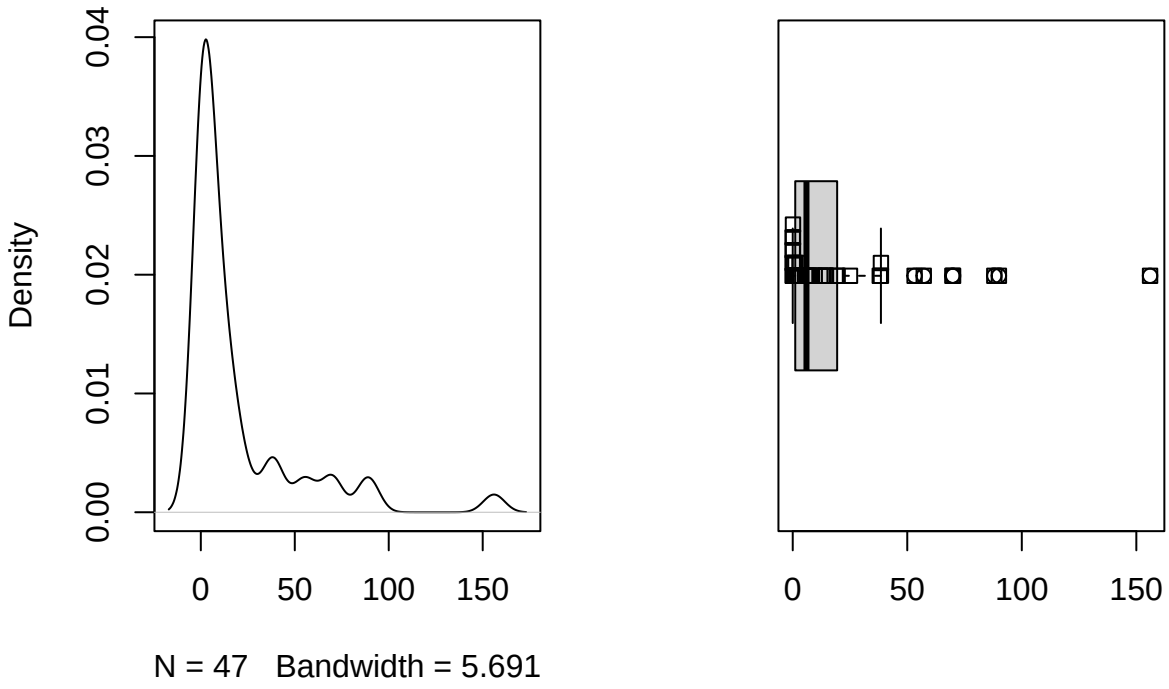


Variable5: Gamble

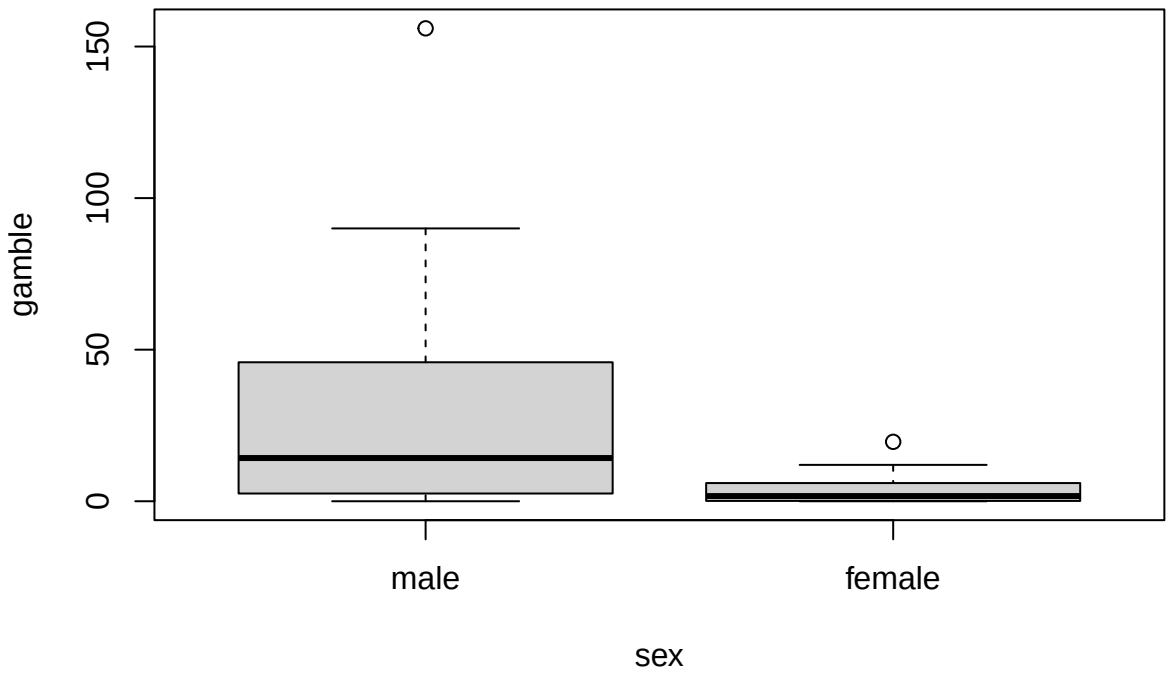
The variable described how much teens spend on gambling.

Similar to income, we can tell the differences between the gambling expenses, thus it is a quantitative ratio variable.

- Several teens showed serious gambling behaviors.

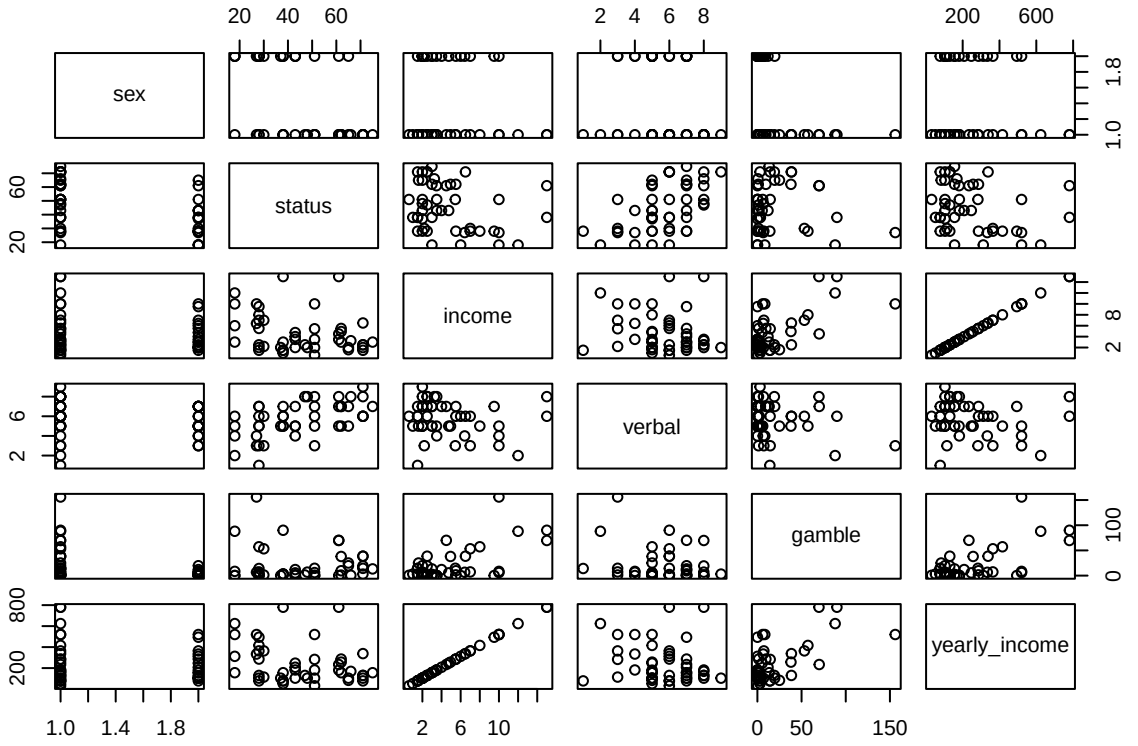


Additionally, I found that on average, men spent more money on gambling than women!



Giving numerical and graphical summaries

```
##      sex         status          income           verbal       gamble
##  male  :28   Min.   :18.00   Min.   : 0.600   6      :12   Min.   :  0.0
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   8      :10   1st Qu.:  1.1
##              Median :43.00   Median : 3.250   7      : 9   Median :  6.0
##              Mean   :45.23   Mean   : 4.642   9      : 6   Mean   : 19.3
##              3rd Qu.:61.50   3rd Qu.: 6.210   4      : 4   3rd Qu.: 19.4
##              Max.   :75.00   Max.   :15.000   5      : 3   Max.   :156.0
##                                               (Other): 3
##  yearly_income
##  Min.   : 31.2
##  1st Qu.:104.0
##  Median :169.0
##  Mean   :241.4
##  3rd Qu.:322.9
##  Max.   :780.0
##
```



b.

I think it is an observational data because the collection of the data seems random and it seems that no researchers are controlling the potential variate that can have an impact on the subject in the study. Since no treatment is imposed, I would consider the data observational.

2. Textile Production data

This is a data about research on textile production.

> Textile production is the process by which fibers, filaments, yarn, and thread - both natural and synthetic - are made.

a.

> To see what variables we have and what sort of values they take.

```
##   press HCHO catalyst temp time
## 1   1.4    8        4  100    1
## 2   2.2    2        4  180    7
## 3   4.6    7        4  180    1
## 4   4.9   10        7  120    5
## 5   4.6    7        4  180    5
## 6   4.7    7        7  180    1
```

Table 2: Textile Production Data

Items	Variables	Description
1	Press	Durable press ratings, it is used for cotton fabrics or textiles with high content of cellulosic fibers.
2	HCHO	formaldehyde concentration, which, in general, are limited to 30 ppm for clothing worn by toddlers.
3	Catalyst Ratio	Catalysts mediate between the reagents in a chemical reaction and control the process leading to the desired end product. When textile material is used as a support for the chemical auxiliaries, the reaction can proceed on a large surface thereby increasing its efficiency. In terms of efficiency, 1% is considered slow mix, 2% is ideal and 3% is a fast mix.
4	Temperature	Curing temperature, measured in °F
5	Time	Curing time

Expect:

1. **Press Ratings** to be a Quantitative Discrete variable;
2. **HCHO** to be a Quantitative Continuous variable;
3. **Catalyst Ratio** to be a Quantitative Continuous variable;
4. **Temp** to be a Quantitative Continuous variable;
5. **Time** to be a Quantitative Continuous variable.

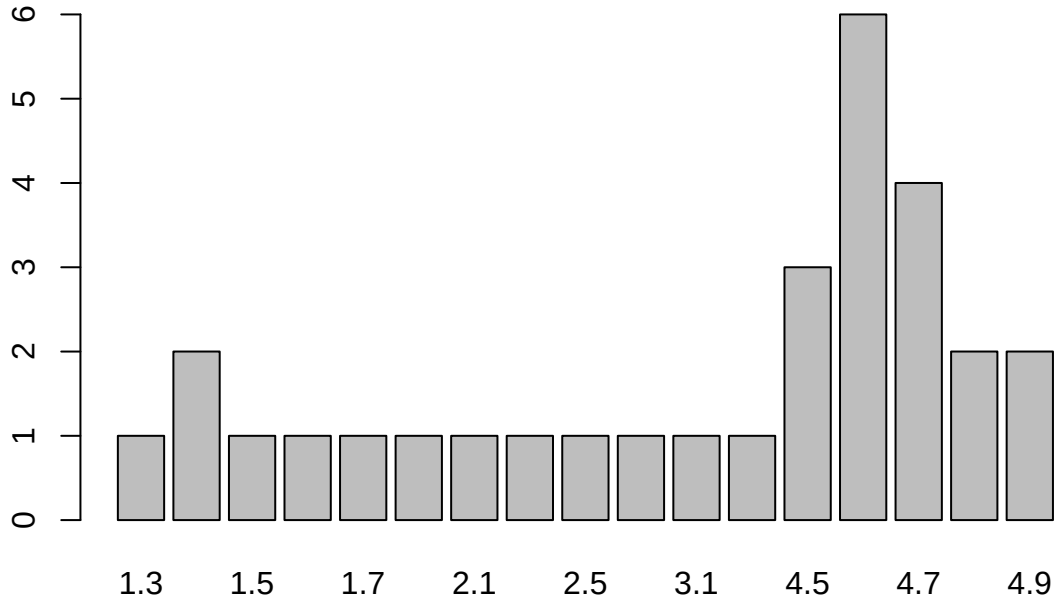
To be able to analyze on the data and its context, we should find out how the data is collected and whether it is representative, but since the information is not available, I would skip this step.

Variable1: press

This variable describes the durable press rating on textiles, intuitively, I believed that ratings are discrete data.

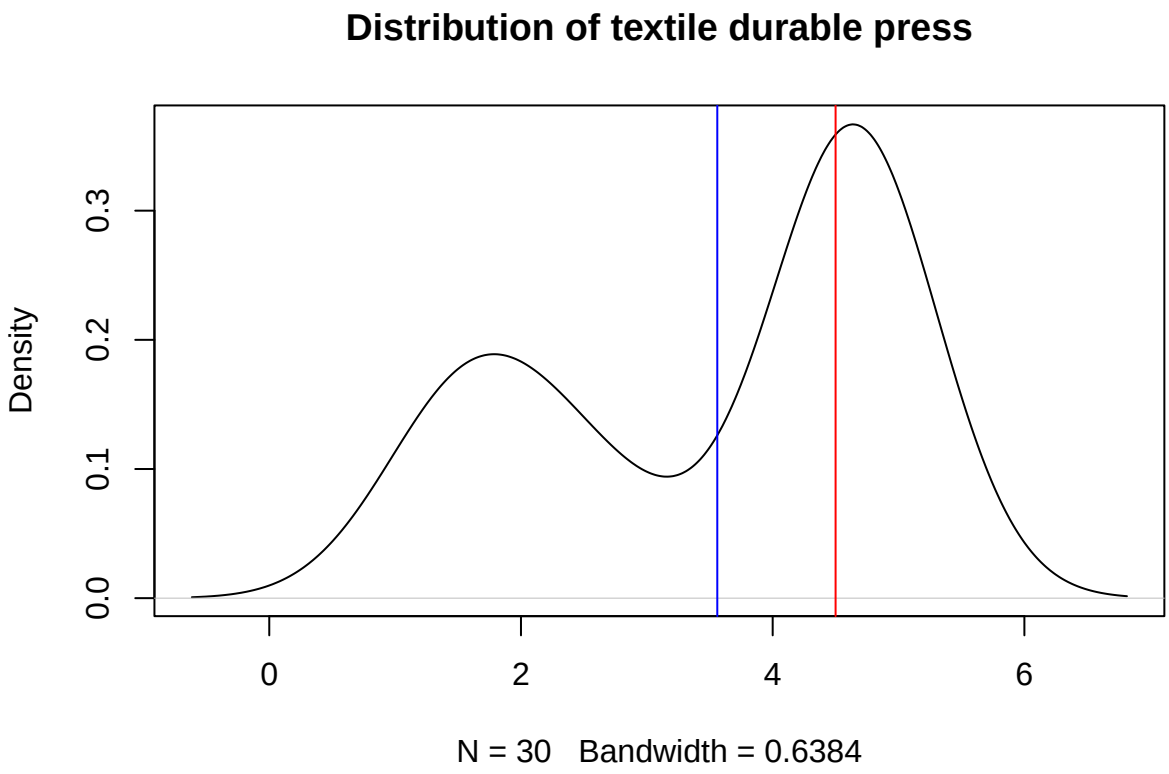
```
##
## 1.3 1.4 1.5 1.6 1.7 1.8 2.1 2.2 2.5 2.6 3.1 4.3 4.5 4.6 4.7 4.8 4.9
##   1   2   1   1   1   1   1   1   1   1   1   1   3   6   4   2   2
```

Visualize

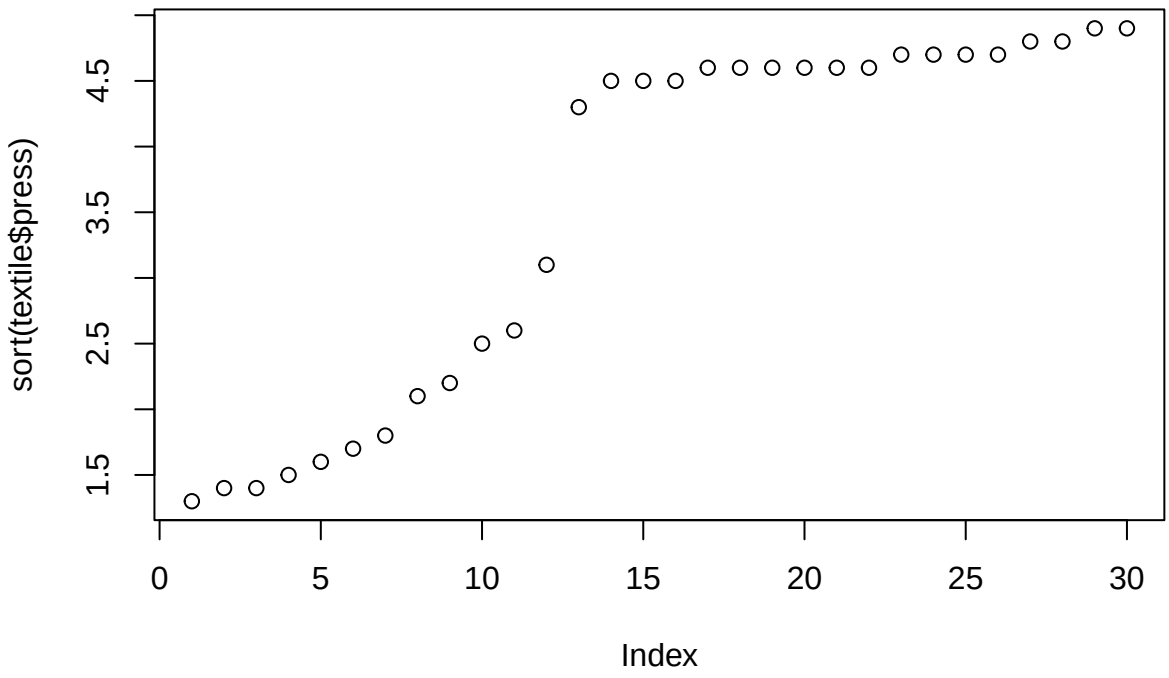


However, I found myself overlooking the possibility that this was an experimental dataset, thus I conjectured that this variable may be continuous and be the *response* of the research.

- So it is better to visualize it by density plots.
- blue: mean; red: median



Something weird happens, There is an obvious break between pressure. See if any of the following analysis on variables can explain this gap.

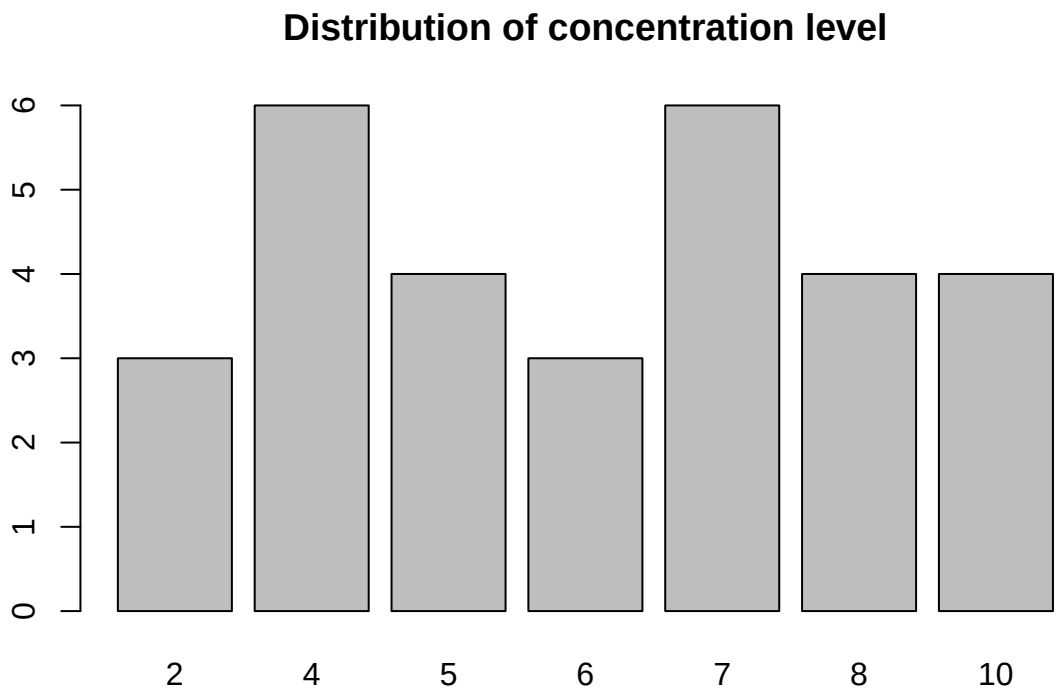


Variable2: HCHO

HCHO is chemical composition, it records the formaldehyde concentration.

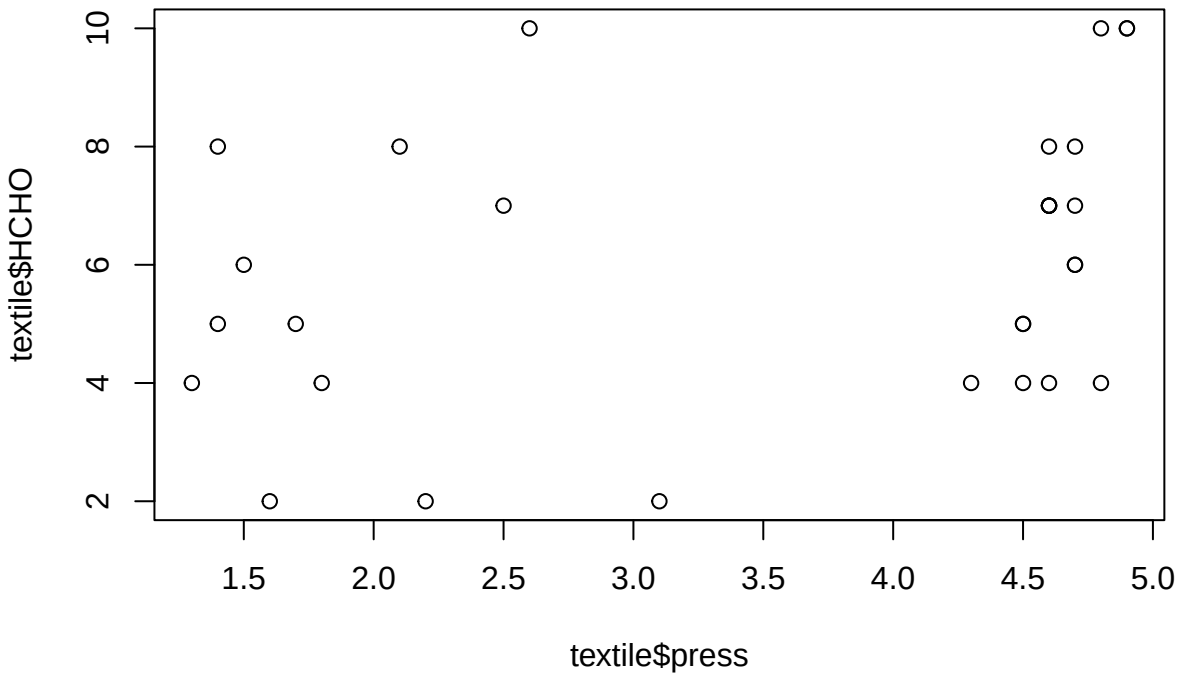
```
##
##  2  4  5  6  7  8 10
##  3  6  4  3  6  4  4
```

> Intuitively, I thought concentration data must be numerical, quantitative ratio data. As the visualization showed, this distribution may be the result of certain control over concentration level.



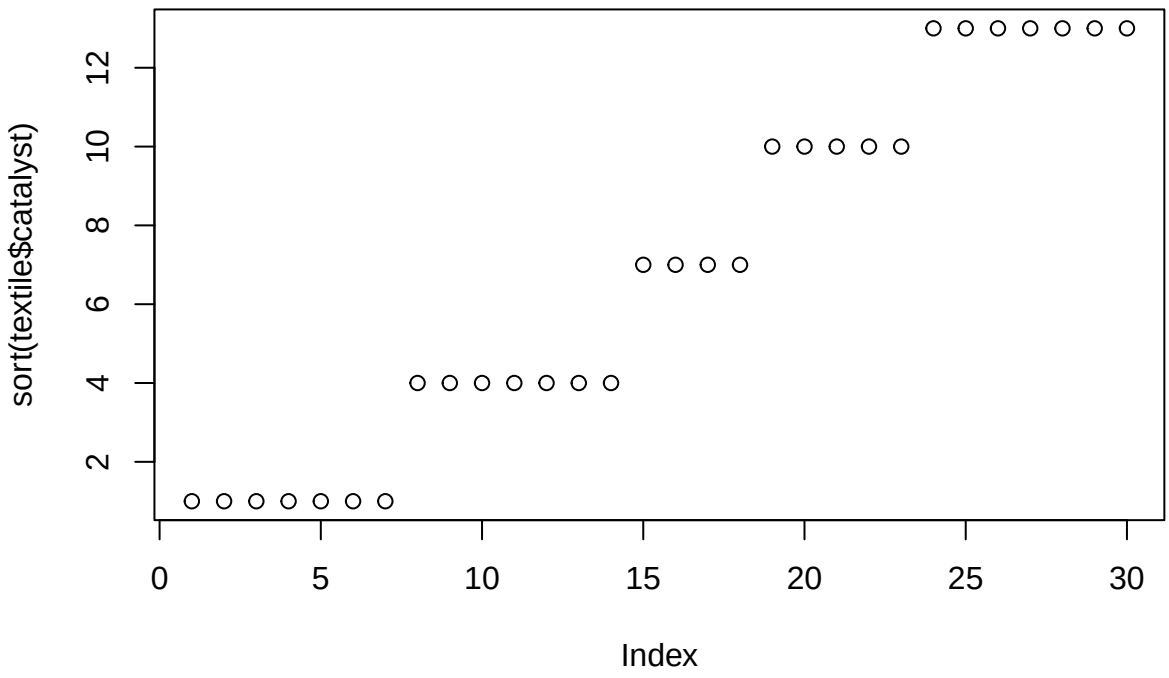
Does HCHO has an impact on the gap between the durable pressure?

- From observation, there is durable pressure values do not lie between 3~4 after HCHO concentration level raises above 2 ppm.



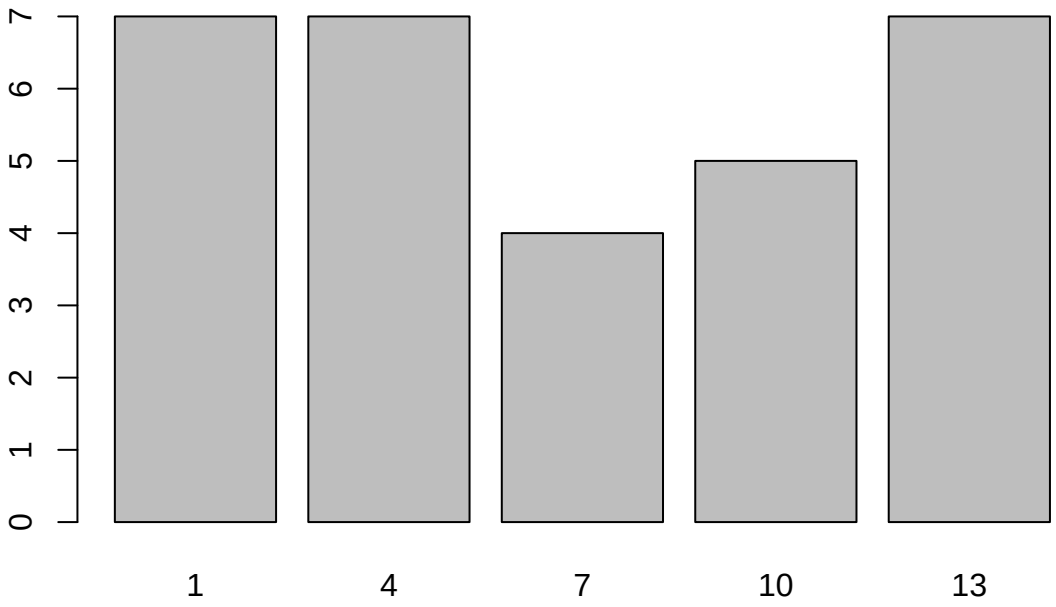
Variable3: Catalyst ratio

Catalyst is used to accelerate the rate of chemical reaction, .



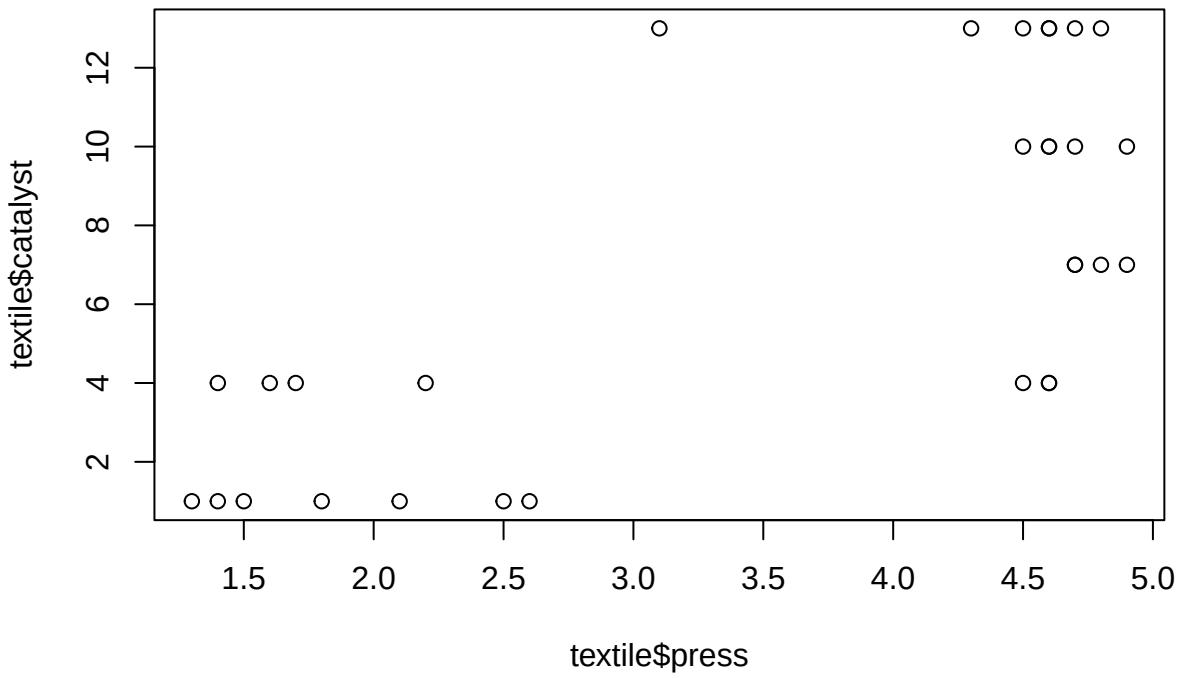
Distribution

```
##
##  1  4  7 10 13
##  7  7  4  5  7
```



I wonder how does the rate of certain chemical reaction has anything to do with durable pressure.

- Surprisingly, if I had more data I would have said it with more certainty, when the catalyst ratio rises above 4, durable pressure of textile suddenly enhanced and sustained the value between 4.5~5 when the catalyst ratio are between 6 to 12. While the value of 3.25 where catalyst ratio is about 12 might be an outlier; however, I would need more data to confirm my suspicion.
- Respond to the previous question, catalyst ratio might be the main reason affecting textile's durable pressure.



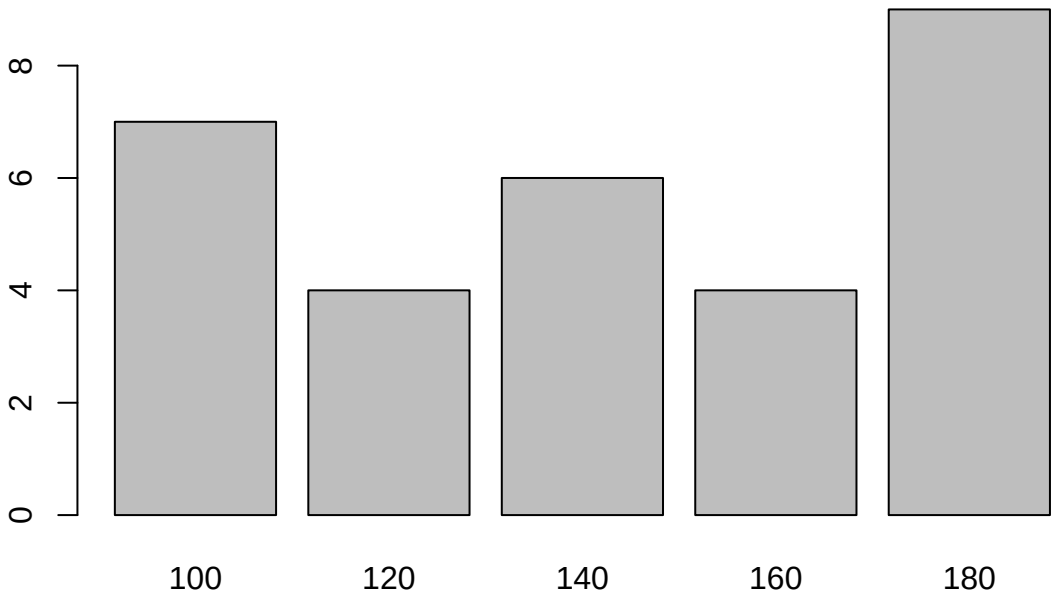
Variable4: Temp

Curing temperature may be related to curing time.

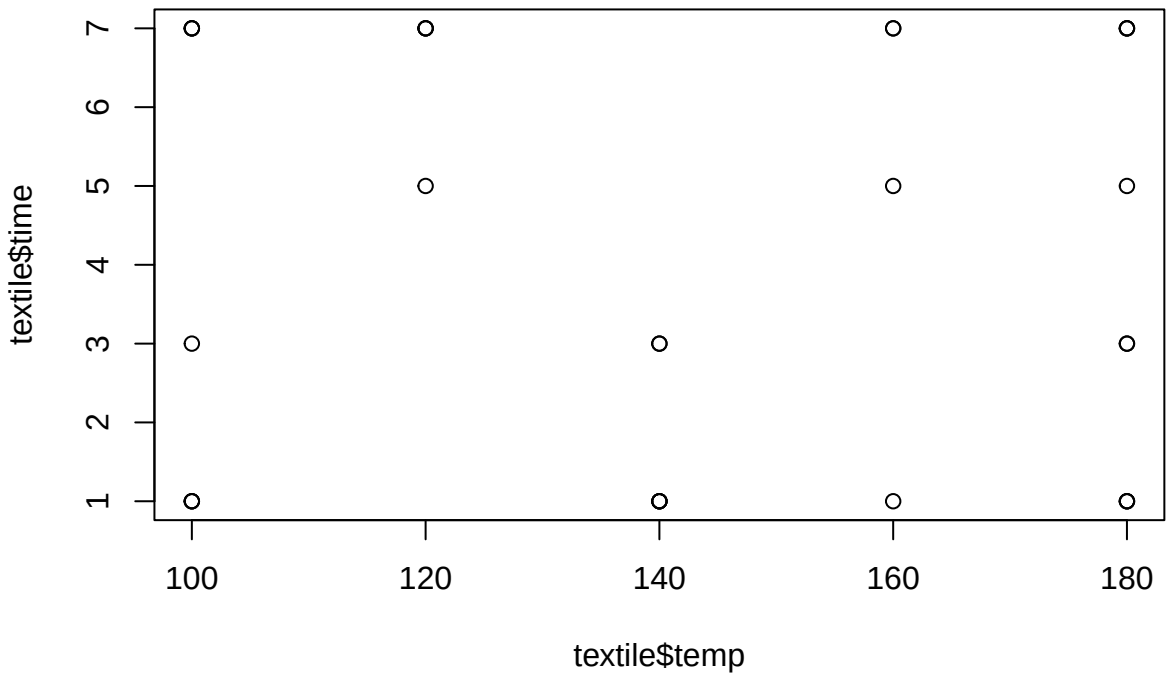
Temperatures range from 100 to 140 and are measured in Fahrenheit. While in Taiwan, we used to measured temperature by Celsius, the data can be modified.

$$Celsius = (Fahrenheit - 32) * 5/9$$

```
##
## 100 120 140 160 180
##   7   4   6   4   9
```

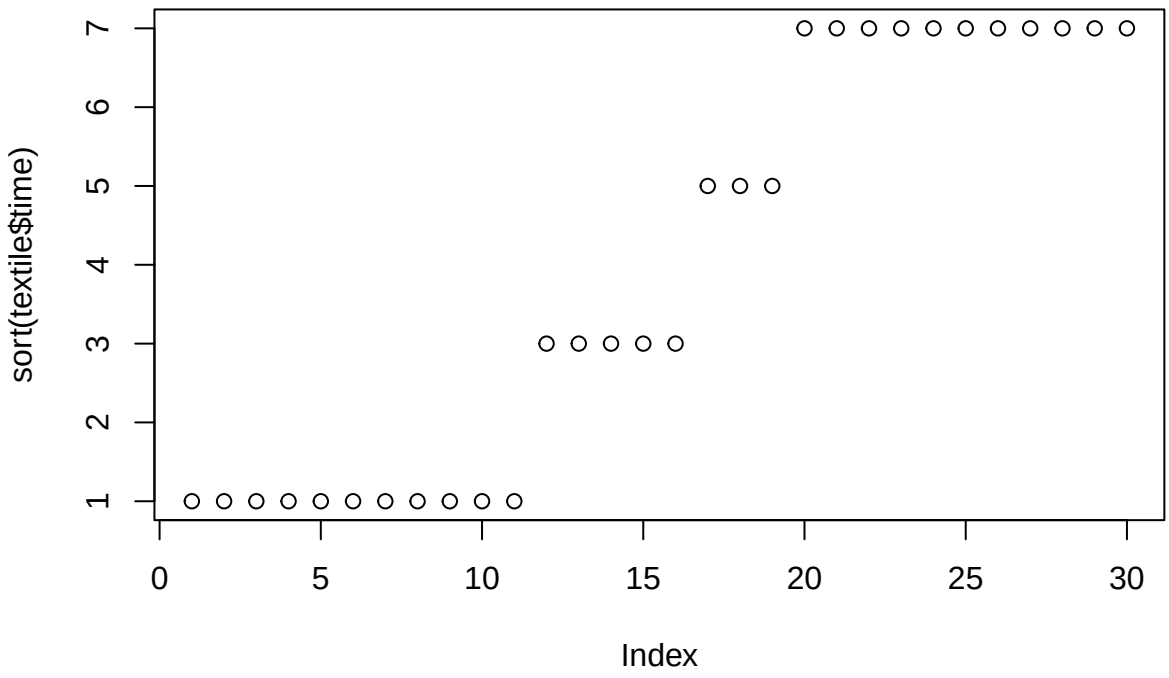


- It seems quite random when I visualize the relationship betwen curing temperature with curing time.

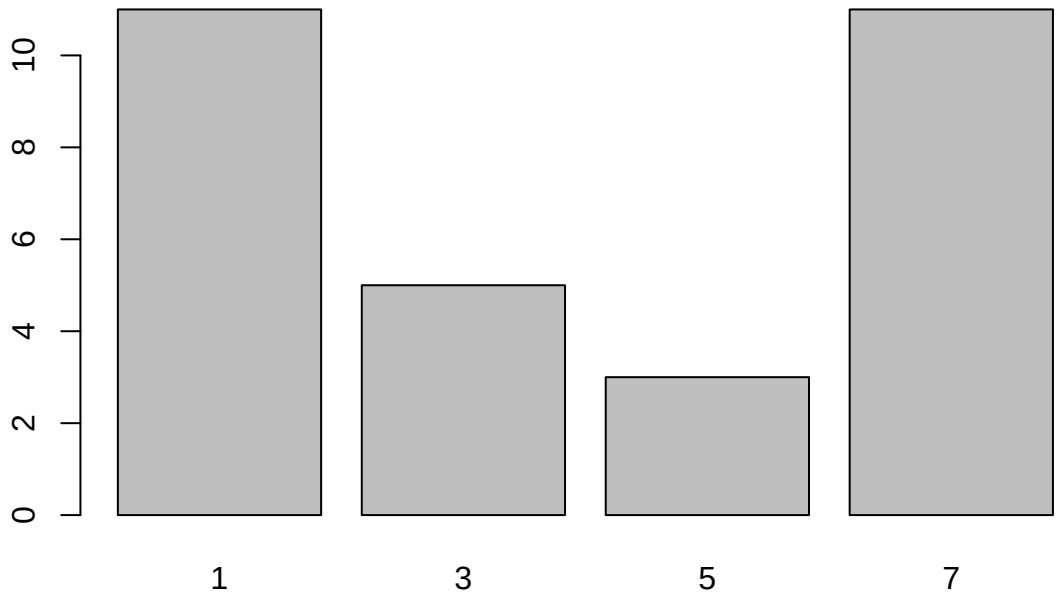


Variable5: Time

hmm... whether the ***time*** variable is measured in minutes, hours or seconds? We can not perform further analysis if the information is not sufficiently provided.

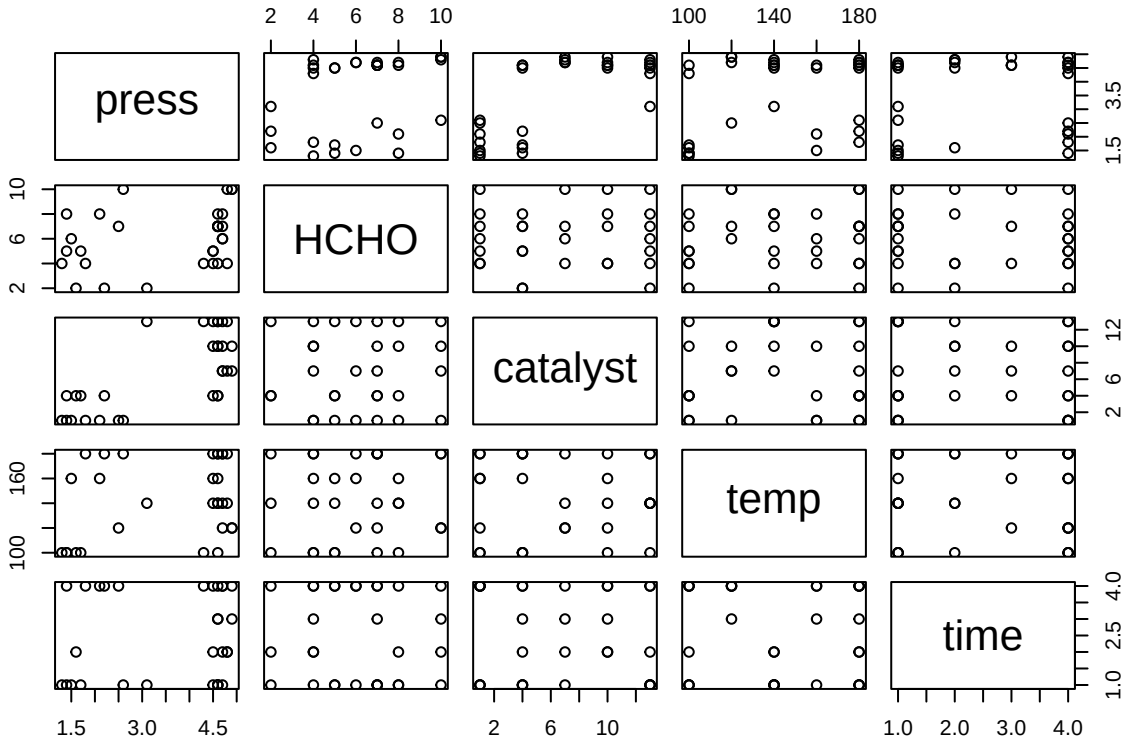



```
##
##  1  3  5  7
## 11  5  3 11
```



Giving numerical and graphical summaries

```
##      press            HCHO          catalyst         temp       time
##  Min.   :1.300   Min.   : 2.000   Min.   : 1.0   Min.   :100.0   1:11
##  1st Qu.:2.125   1st Qu.: 4.000   1st Qu.: 4.0   1st Qu.:120.0   3: 5
##  Median :4.500   Median : 6.000   Median : 7.0   Median :140.0   5: 3
##  Mean   :3.560   Mean   : 6.067   Mean   : 6.8   Mean   :142.7   7:11
##  3rd Qu.:4.675   3rd Qu.: 7.750   3rd Qu.:10.0   3rd Qu.:180.0
##  Max.   :4.900   Max.   :10.000   Max.   :13.0   Max.   :180.0
```



After giving some thought, I think the reason why many of these scatter plots showed such distribution (seems fairly random) is because a rigorous experiment should not include variables that are correlated with each other, the result which aims to identify which element is affecting textile's durable press might be misleading if curing temp is positively related to curing time.

b.

In common sense, pressure rating, concentration level, temperature, or time should all be continuous data, so when I thought it was observational data, it seems really weird and counterintuitive to me then. But after reviewing some materials covered in class and observing patterns, I suddenly realize that this is not observational data!! And therefore, I can firmly state that this must be experimental data, while the research team of this study controlled most of the variables in order to evaluate outcomes under given environments and since the research output should be continuous, I surmised that it may be press, and the rest of the variables are predictors.