1.13/15

2.14/15

Assigment 1 - Linear model

ID : 111024517 Name : 鄭家豪

2022-10-06

1.(a)

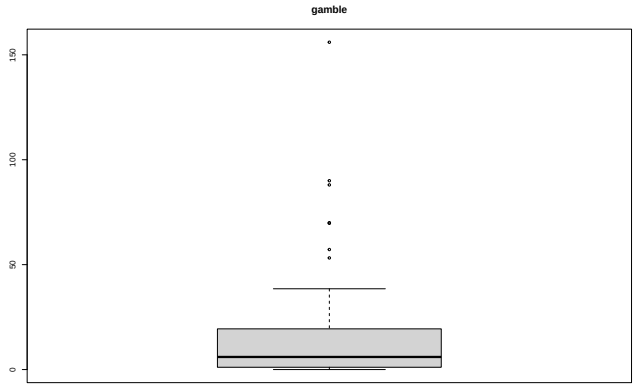
將此筆資料呈現出來，如下。
sex、status和verbal皆為整數。
sex代表性別，沒有量化含意。
status和verbal皆是可量化的變數，值越大其對應的含意越高。

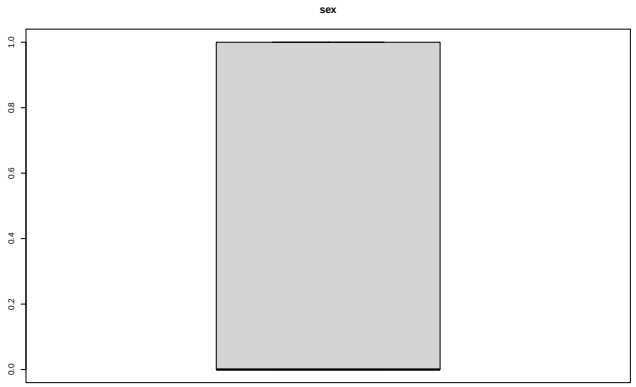
```
##    sex status income verbal gamble
## 1    1     51   2.00      8   0.00
## 2    1     28   2.50      8   0.00
## 3    1     37   2.00      6   0.00
## 4    1     28   7.00      4   7.30
## 5    1     65   2.00      8  19.60
## 6    1     61   3.47      6   0.10
## 7    1     28   5.50      7   1.45
## 8    1     27   6.42      5   6.60
## 9    1     43   2.00      6   1.70
## 10   1     18   6.00      7   0.10
## 11   1     18   3.00      6   0.10
## 12   1     43   4.75      6   5.40
## 13   1     30   2.20      4   1.20
## 14   1     28   2.00      6   3.60
## 15   1     38   3.00      6   2.40
## 16   1     38   1.50      8   3.40
## 17   1     28   9.50      8   0.10
## 18   1     18  10.00      5   8.40
## 19   1     43   4.00      8  12.00
## 20   0     51   3.50      9   0.00
## 21   0     62   3.00      8   1.00
## 22   0     47   2.50      9   1.20
## 23   0     43   3.50      5   0.10
## 24   0     27  10.00      4 156.00
## 25   0     71   6.50      7  38.50
## 26   0     38   1.50      7   2.10
## 27   0     51   5.44      4  14.50
## 28   0     38   1.00      6   3.00
## 29   0     51   0.60      7   0.60
## 30   0     62   5.50      8   9.60
## 31   0     18  12.00      2  88.00
## 32   0     30   7.00      7  53.20
## 33   0     38  15.00      7  90.00
## 34   0     71   2.00     10   3.00
## 35   0     28   1.50      1  14.10
## 36   0     61   4.50      8  70.00
```

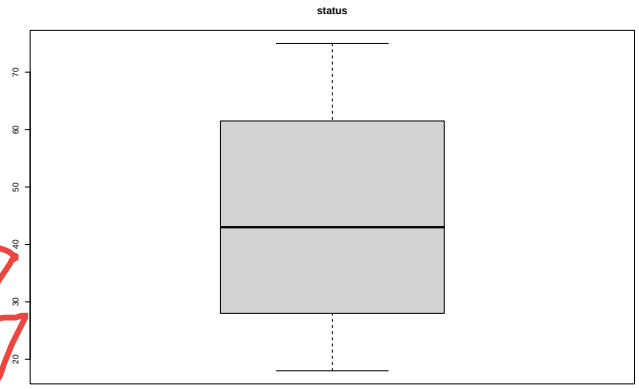
```
## 37   0     71   2.50      7  38.50
## 38   0     28   8.00      6  57.20
## 39   0     51  10.00      6   6.00
## 40   0     65   1.60      6  25.00
## 41   0     48   2.00      9   6.90
## 42   0     61  15.00      9  69.70
## 43   0     75   3.00      8  13.30
## 44   0     66   3.25      9   0.60
## 45   0     62   4.94      6  38.00
## 46   0     71   1.50      7  14.40
## 47   0     71   2.50      9  19.20
```

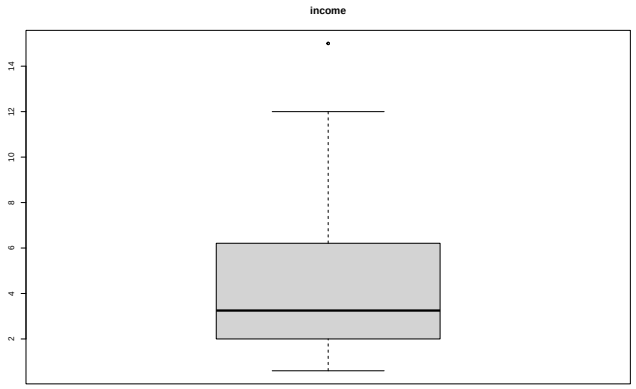
以下是藉由summary()這個指令對此筆資料所出來的結果。
sex代表性別男女，因此其分位數不具備任何含意。
status的數據分佈於[28,61.5]之間相當均勻。
income的數據分佈於[0.6,3.25]之間相當密集。
verbal的數據分佈於[6,8]之間相當均勻。
gamble的中位數與最小值頗相近，代表數據集中在[0,6]之間相當密集。

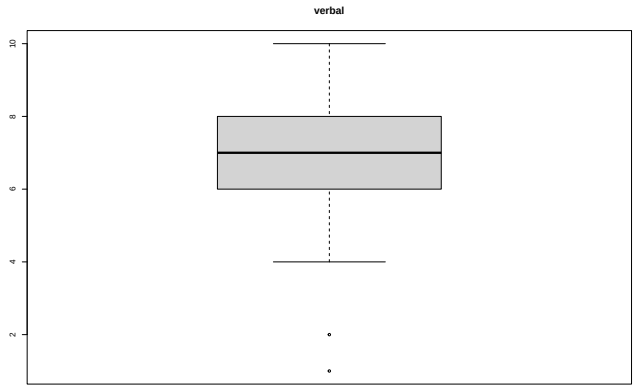
```
##       sex            status          income           verbal          gamble      
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0  
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1  
##  Median :0.0000   Median :43.00   Median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



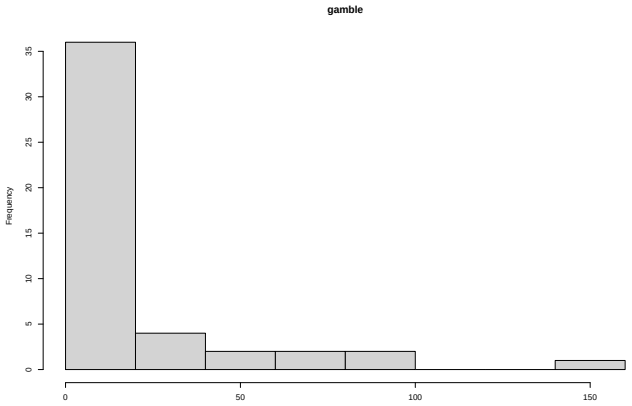


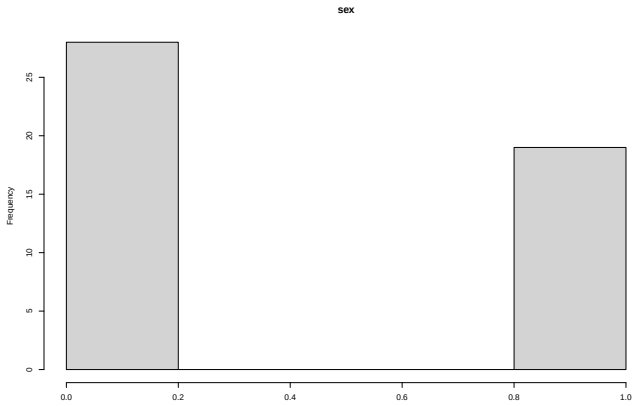


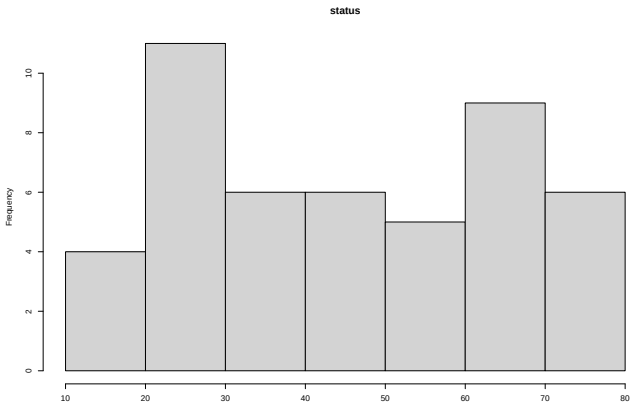


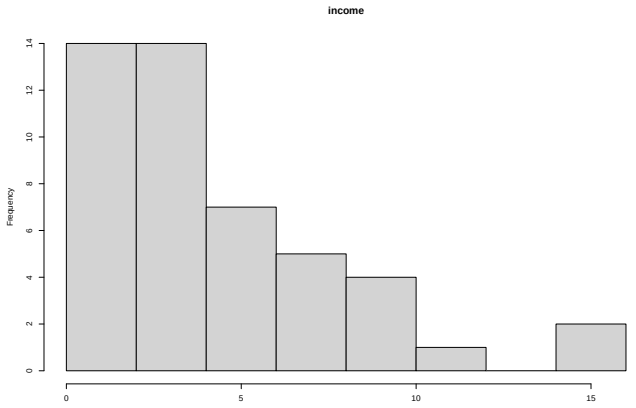


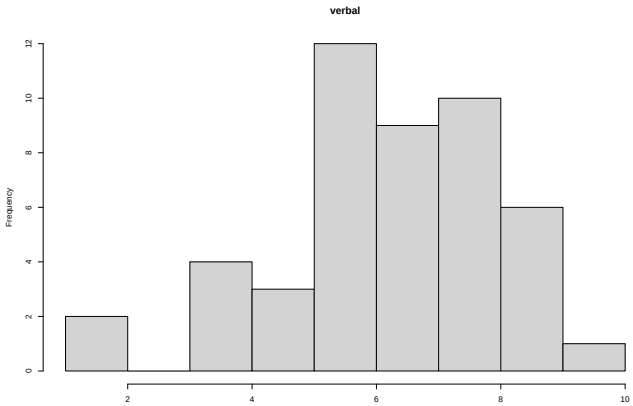
由gamble的boxplot來看，數據分佈於最小值至中間值之間相當密集，中間值至最大值的數據分佈比較分散。
sex 不具量化含意，故其boxplot沒辦法觀察出甚麼內容。
由status的boxplot來看，數據分佈於[28,61.5]之間相當均勻。
由income的boxplot來看，數據分佈於[0.6,3.25]相當密集。
由verbal的boxplot來看，數據分佈於[6,8]之間相當均勻
以上由boxplot觀察的結果與前面summary所得到的結論一致。



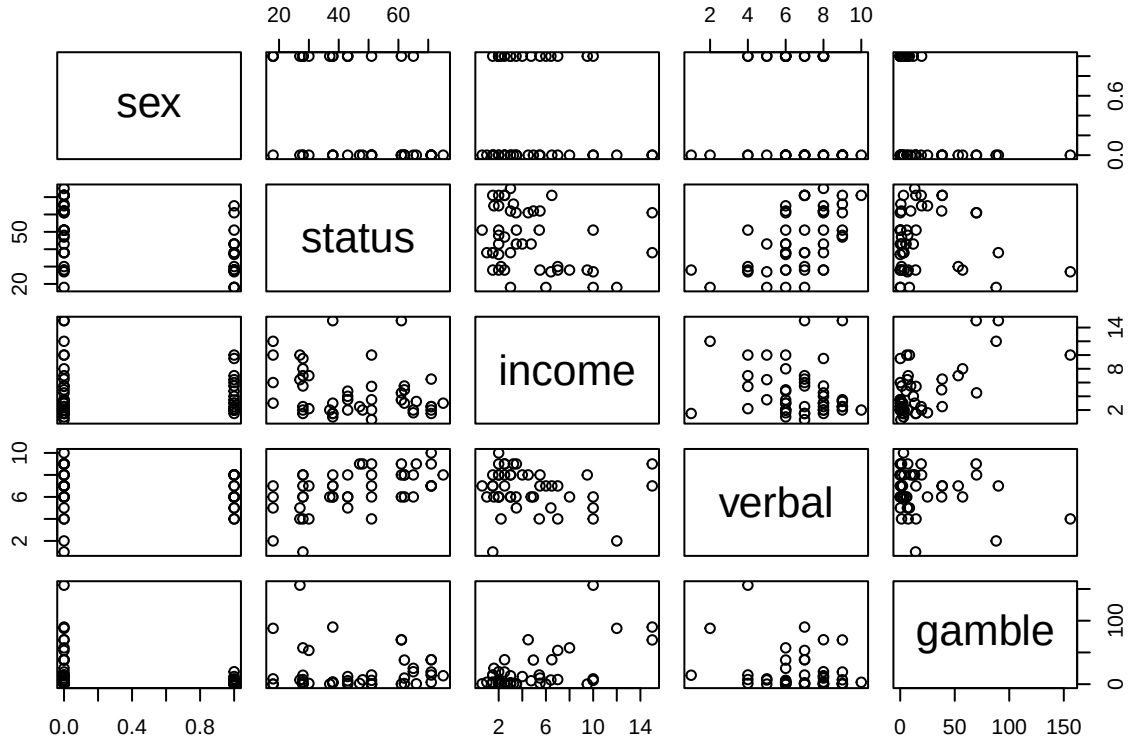








由以上直方圖來看，gamble很顯然是正偏態分佈(skewed to right)，因此數據集中於左方。
此筆資料，性別是男生的數量比較多。
其他變數的直方圖，所觀察的現象，與前面summary和boxplot所得到的結論一致。



觀察sex與gamble的圖，可以發現男生的賭博支出普遍高於女生。
然後sex與其他變數之間的圖，分佈都蠻均勻的。
status與income的圖，和status與verbal的圖，都有蠻明顯的正相關。
gamble與status、income和verbal，似乎有負相關的趨勢。

1.(b)

此筆數據應視為observational data，因為此筆資料比較像是分析各個變數與賭博支出的關係，不像是為了釐清哪些因素能使實驗產生不同結果，而控制變數。

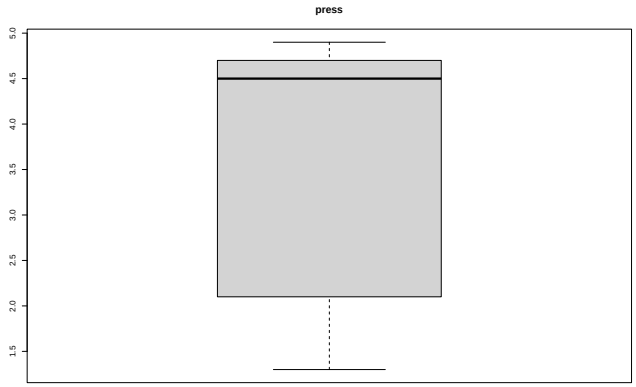
2.(a)

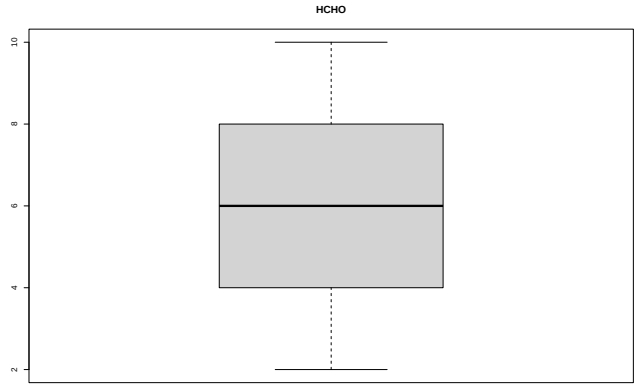
將此筆資料呈現出來，如下。
可發現這些變數:HCHO、catalyst、temp和time，取值皆為整數，而且只有幾種可能值。
如:temp的值只有100、120、140、160、180。
另外可發現除了HCHO和press外，其他變數的取值具有等差數列的性質。

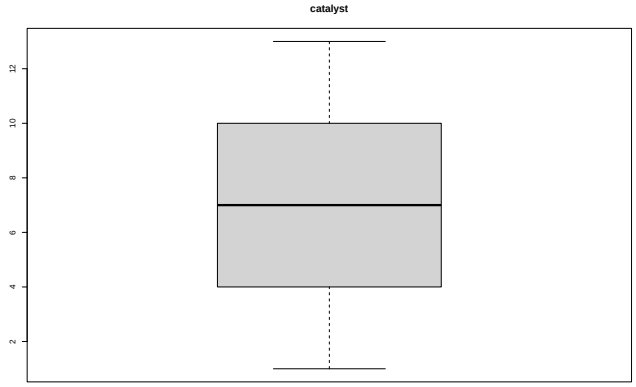
```
##    press HCHO catalyst temp time
## 1    1.4    8        4  100    1
## 2    2.2    2        4  180    7
## 3    4.6    7        4  180    1
## 4    4.9   10        7  120    5
## 5    4.6    7        4  180    5
## 6    4.7    7        7  180    1
## 7    4.6    7       13  140    1
## 8    4.5    5        4  160    7
## 9    4.8    4        7  140    3
## 10   1.4    5        1  100    7
## 11   4.7    8       10  140    3
## 12   1.6    2        4  100    3
## 13   4.5    4       10  180    3
## 14   4.7    6        7  120    7
## 15   4.8   10       13  180    3
## 16   4.6    4       10  160    5
## 17   4.3    4       13  100    7
## 18   4.9   10       10  120    7
## 19   1.7    5        4  100    1
## 20   4.6    8       13  140    1
## 21   2.6   10        1  180    1
## 22   3.1    2       13  140    1
## 23   4.7    6       13  180    7
## 24   2.5    7        1  120    7
## 25   4.5    5       13  140    1
## 26   2.1    8        1  160    7
## 27   1.8    4        1  180    7
## 28   1.5    6        1  160    1
## 29   1.3    4        1  100    1
## 30   4.6    7       10  100    7
```

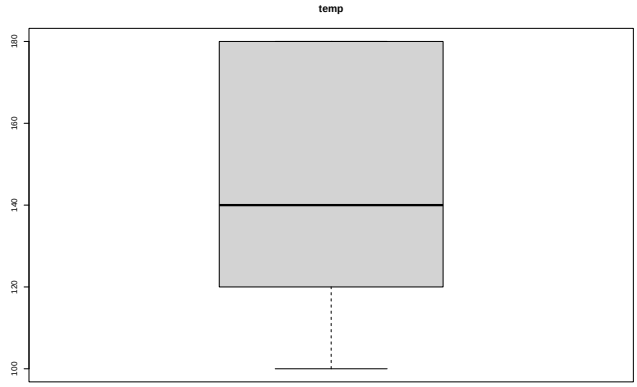
對此筆資料，使用summary()這個指令，結果如下。
press的中位數與最大值頗接近，代表一半的數據集中在[4.5,4.9]之間，且一半的數據分散在[1.3,4.5]之間。
HCHO和catalyst的數據分佈有對稱性的可能。
temp的第三四分位數與最大值相同，代表有四分之一的數據集中在180度(溫度單位)。
time分別各有四分之一的數據集中在1和7，

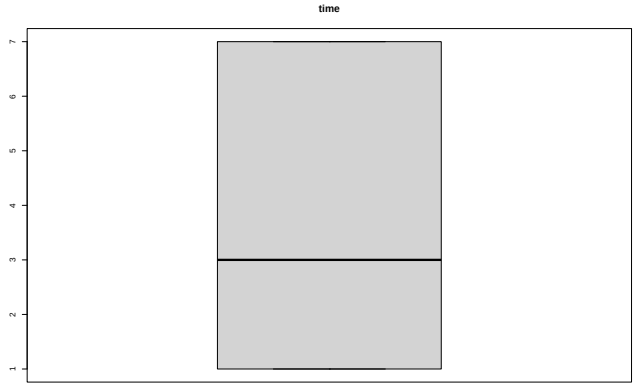
```
##      press            HCHO          catalyst         temp           time      
##  Min.   :1.300   Min.   : 2.000   Min.   : 1.0   Min.   :100.0   Min.   :1.000  
##  1st Qu.:2.125   1st Qu.: 4.000   1st Qu.: 4.0   1st Qu.:120.0   1st Qu.:1.000  
##  Median :4.500   Median : 6.000   Median : 7.0   Median :140.0   Median :3.000  
##  Mean   :3.560   Mean   : 6.067   Mean   : 6.8   Mean   :142.7   Mean   :3.933  
##  3rd Qu.:4.675   3rd Qu.: 7.750   3rd Qu.:10.0   3rd Qu.:180.0   3rd Qu.:7.000  
##  Max.   :4.900   Max.   :10.000   Max.   :13.0   Max.   :180.0   Max.   :7.000
```



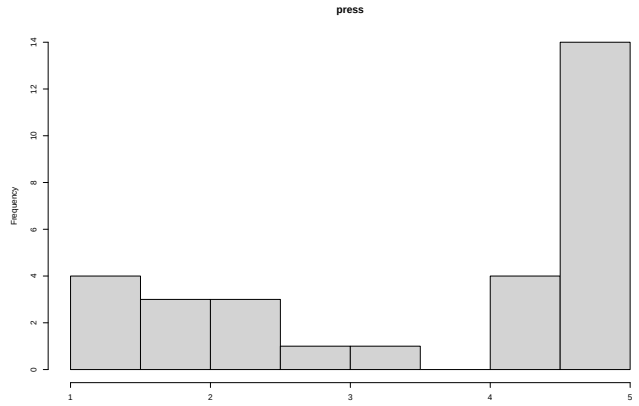


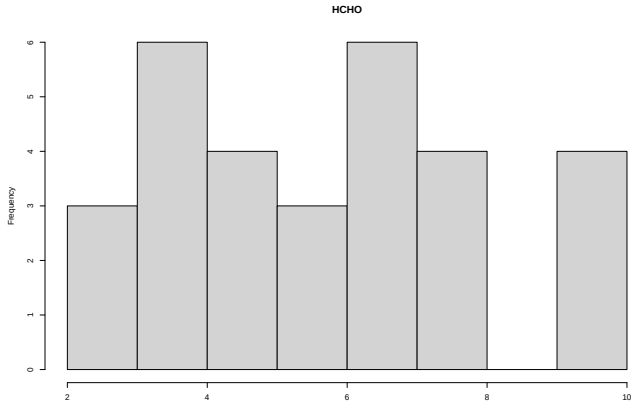


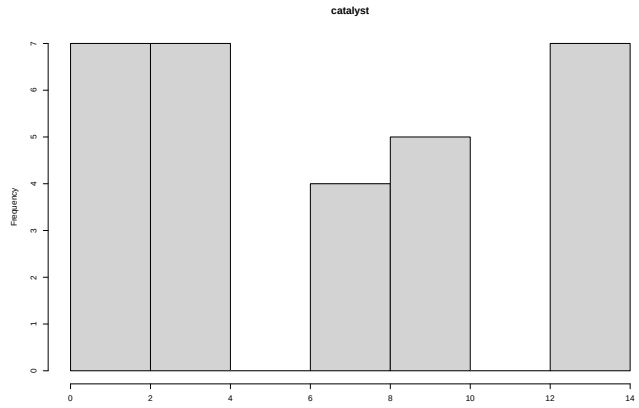


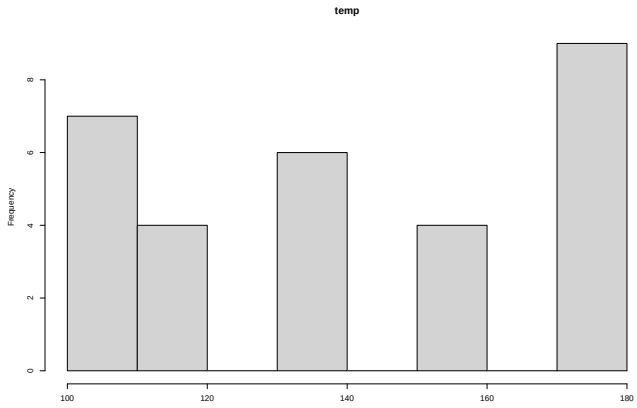


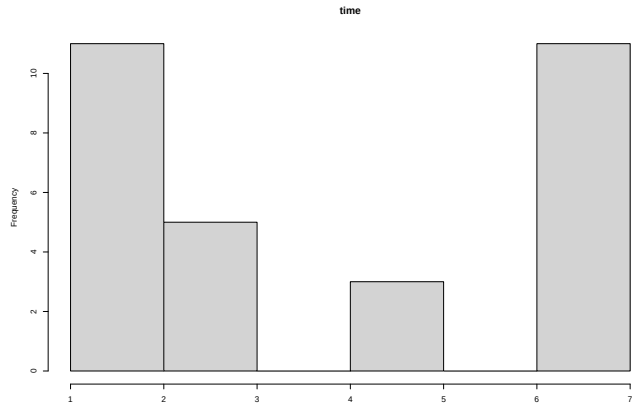
由press的boxplot來看，中位數與第三四分位數蠻接近的，且中位數與第一四分位數相差蠻大的，代表一半的數據集中在[4.5,4.9]之間，且一半的數據分散在[1.3,4.5]之間。
由HCHO和catalyst的boxplot來看，有分布對稱的可能性。
由temp的boxplot來看，最大值和第三四分位數相同。
由time的boxplot來看，最大值等於第三四分位數且最小值等於第一四分位數。
以上由boxplot觀察的結果與前面summary所得到的結論一致。



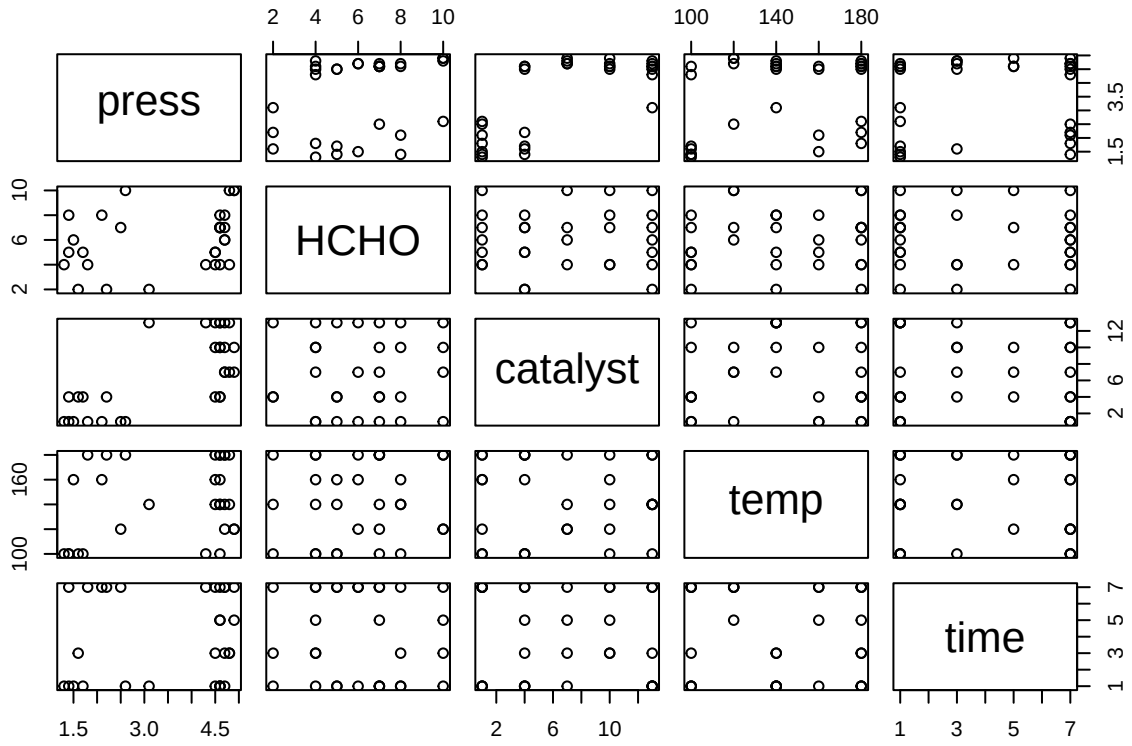








由上方直方圖來看，press之數據集中在[4.5,5]，與summary和boxplot所觀察的結論一致。
前述討論的對稱性，由直方圖觀察出，不支持HCHO和catalyst分布對稱的可能性。
觀察不出temp的第三四分位數多少，然後time的數據明顯集中在1和7。



觀察press對catalyst的圖，由於catalyst分布相當集中在[1,4]和[12,13]區間(由直方圖觀察出來)，
對應之區間的數據相當密集，可看出兩者變數呈正相關。
除了catalyst，其他變數對press的散佈圖，分佈較為分散，無法看出明顯的相關性。
除了press，其他變數相對應的散佈圖，分佈也較為分散，因此無法看出明顯的相關性。
press外的變數，彼此之間的散佈圖可觀察出規則性，可能是因為這些變數取值只有幾種可能，才會導致這樣的結果。

2.(b)

此筆數據應視為experimental data。
因為除了press，其他變數的取值都只有幾種可能，
其目的應該是為了釐清哪些因素能使實驗產生不同結果而形成的概念。