Homework1

黃晨溦、廖偉傑

2022-10-02

1.a

Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

- sex: 0=male, 1=female
- socio-economic status: a score based on parents' occupation
- income: in pounds per week
- verbal score: words out of 12 correctly defined
- expenditure on gambling: pounds per year

1.a.Ans

```
attach(data1)
str(data1)
```

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```
data1$sex[which(sex == 0)] <- "male"
data1$sex[which(sex == 1)] <- "female"
table(sex=sex);table(status=status);table(income = income);table(verbal = verbal);table(gamble=gamble)
```

```
## sex
##  0  1
## 28 19
```

```
## status
## 18 27 28 30 37 38 43 47 48 51 61 62 65 66 71 75
##  4  2  7  2  1  5  4  1  1  5  3  3  2  1  5  1
```

```
## income
##  0.6    1  1.5  1.6    2  2.2  2.5    3 3.25 3.47  3.5    4  4.5 4.75 4.94 5.44
##    1    1    4    1    7    1    4    4    1    1    2    1    1    1    1    1
##  5.5    6 6.42  6.5    7    8  9.5   10   12   15
##    2    1    1    1    2    1    1    3    1    2

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

## gamble
##    0  0.1  0.6    1  1.2 1.45  1.7  2.1  2.4    3  3.4  3.6  5.4    6  6.6  6.9
##    4    5    2    1    2    1    1    1    1    2    1    1    1    1    1    1
##  7.3  8.4  9.6   12 13.3 14.1 14.4 14.5 19.2 19.6   25   38 38.5 53.2 57.2 69.7
##    1    1    1    1    1    1    1    1    1    1    1    1    2    1    1    1
##   70   88   90  156
##    1    1    1    1
```

```
detach(data1)
```

根據題目，先將類別變數 sex 從數值標記轉變成類別名稱標記方便識別，其餘變數皆為連續型。透過 str 可知這筆資料包含 47 個觀測值以及 5 個變數。

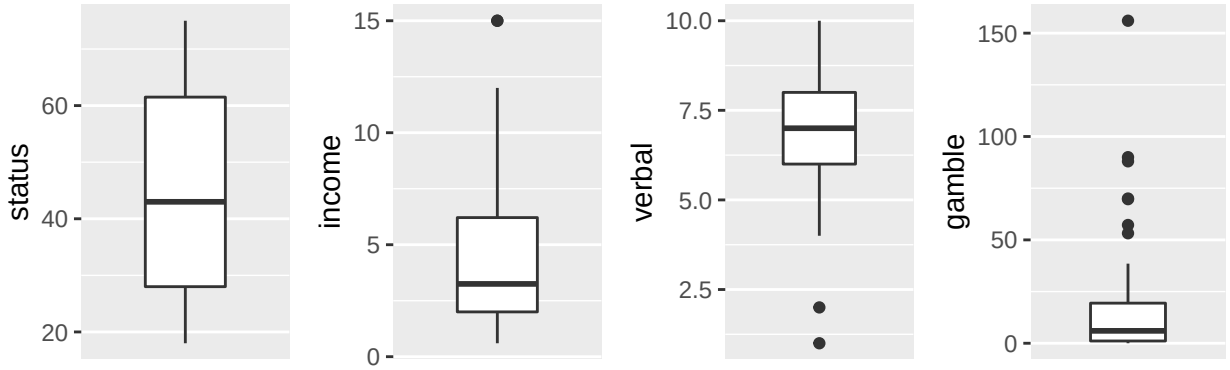
透過 table 觀察數值特性：

- 資料並無任何缺失值，且所有變數數值皆大於正，並無不合理處。
- 各變數都有符合題目敘述多樣化的數值。

```
summary(data1)
```

```
##      sex                status          income           verbal
##  Length:47          Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  Class :character   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Mode  :character   Median :43.00   Median : 3.250   Median : 7.00
##                     Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##                     3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##                     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
```

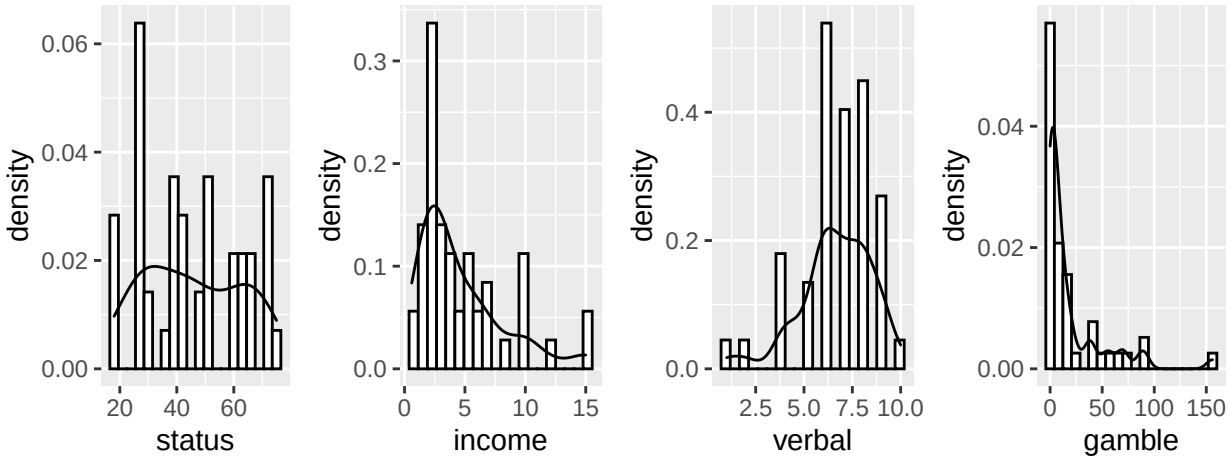
```
p1 <- ggplot(data1, aes(y = status)) + geom_boxplot() + scale_x_discrete()
p2 <- ggplot(data1, aes(y = income)) + geom_boxplot() + scale_x_discrete()
p3 <- ggplot(data1, aes(y = verbal)) + geom_boxplot() + scale_x_discrete()
p4 <- ggplot(data1, aes(y = gamble)) + geom_boxplot() + scale_x_discrete()
gridExtra::grid.arrange(p1,p2,p3,p4, ncol = 4)
```



透過 summary 和 boxplot 觀察連續型變數敘述統計量：

- 所有變數數值皆大於正，符合題目並無不合理處。
- Expenditure on gambling (gamble) 的數據分佈不均，中位數為 6 較平均 19.3 小許多，再綜合對 Q3 的觀察，可發現 gamble 的數值範圍很廣。在大於中位數的區段，gamble 數值之間的差異非常大。
- Socio-economic status (status)、income、verbal score (verbal) 的數據分佈較為平均。

```
myhist <- function(x){
  ggplot(data1, aes_string(x = x)) +
  geom_histogram(aes(y = ..density..), bins = 20, colour = 1, fill = "white") +
  geom_density() + xlab(x)
}
gridExtra::grid.arrange(myhist("status"),myhist("income"),
                        myhist("verbal"),myhist("gamble"), ncol = 4)
```



透過 histogram 觀察連續型變數分佈：

- income 和 gamble 的分佈右偏，其中又以 gamble 極端值較多。
- status 數值的分佈平均，略呈雙峰。
- verbal 數值的分佈平均，呈單峰，形似常態。

```
ggpairs(data1, aes(color = sex, alpha = 0.5))
```

透過 scatter plot 和相關係數分別觀察散佈特性與相關性：

- 考慮反應變數 gamble 與其他的解釋變數關係：
  - gamble 在兩性區分（sex）下分佈有明顯不同，所反應出的資訊為男性每年在博弈上的花費比女性高，與傳統認為男性較愛從事賭博行為的想法相符。
  - gamble 與變數 status 相關性極小，但在性別區分下，與 status 的相關係數兩者為異號，這樣的情況下可能是由 status 跟 sex 的交互作用所導致，將在解釋變數間做討論。
  - gamble 與變數 income 有正相關，其中如果又以性別區分後，發現正相關主要由男性族群貢獻，意即男性的收入高低與花費在賭博上的金額有明顯正相關性，而女性的收入高低不影響其對賭博的花費，同樣為 sex 與 status 兩變數對於 gamble 的交互作用。
- 考慮解釋變數之間相關性：
  - status 在 sex 的區分下，明顯分佈不同，以至於在 histogram 的呈現上有雙峰的趨勢，而對於解釋變數上所造成交互作用現象，其背後原因可能為對於抽樣的人群特性，例如在博弈活動區域中抽樣，對於抽到花費博弈金額較高的男性較多為家庭社經地位較高的對象，而對於家庭社經地位高的女性，其或許不會花費這金額在博弈上，但對於資料背後的資訊，可能還有其他原因造成這樣的現象。
  - 其他解釋變數之間無明顯相關性。

b. Is this observational or experimental data? Explain your reasoning.
此資料調查 teenage gambling，但人員並沒有人為控制 gambling 以外其餘變因的變動，或對調查進行其他人為干預，所以這應該是 observational data。

2.a

Do some short numerical and graphical summaries of the data, commenting on any features that you find interesting.

- press: durable press rating
- HCHO: formaldehyde concentration
- catalyst: catalyst ration
- temp: curing temperature
- time: curing time

2.a.Ans

```
str(data2)
```

```
## 'data.frame':    30 obs. of  5 variables:
##  $ press   : num  1.4 2.2 4.6 4.9 4.6 4.7 4.6 4.5 4.8 1.4 ...
##  $ HCHO    : int  8 2 7 10 7 7 7 5 4 5 ...
##  $ catalyst: int  4 4 4 7 4 7 13 4 7 1 ...
##  $ temp    : int  100 180 180 120 180 180 140 160 140 100 ...
##  $ time    : int  1 7 1 5 5 1 1 7 3 7 ...
```

```
attach(data2)
table(press=press);table(HCHO=HCHO);table(catalyst);table(temp);table(time)
```

```
## press
## 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

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

## catalyst
##  1  4  7 10 13
##  7  7  4  5  7

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

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

```
detach(data2)
```

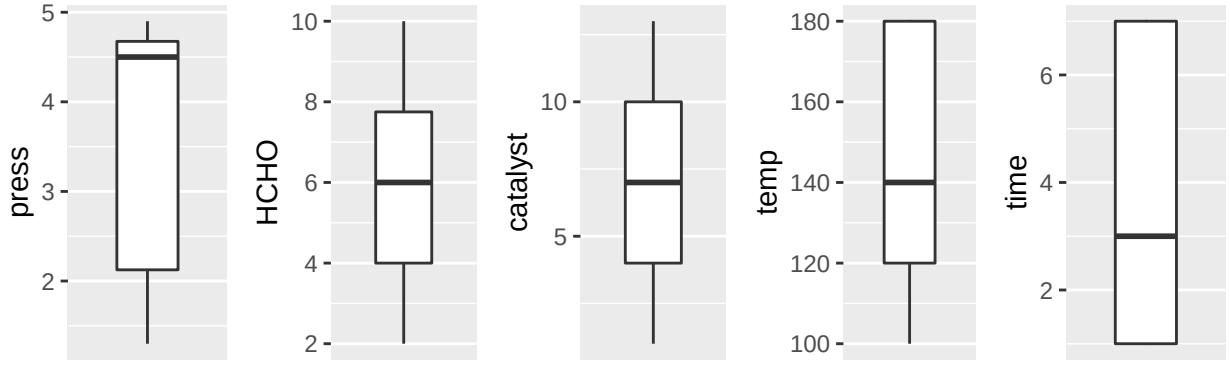
根據題目，所有變數皆為連續型。透過 str 可知這筆資料包含 30 個觀測值以及 5 個變數。

透過 table 觀察數值特性：
- 資料並無任何缺失值，且所有變數數值皆大於正，符合題目並無不合理處。
- Durable press rating (press) 數值的多樣性較高。
- HCHO concentration (HCHO)、catalyst ration (catalyst)、Curing temperature (temp) 和 Curing time (time) 變數的數值都非常特定，沒有小數點。這違反了他們作為連續型變數的特性，代表他們應當是作為實驗因子受人員操控。

```
summary(data2)
```

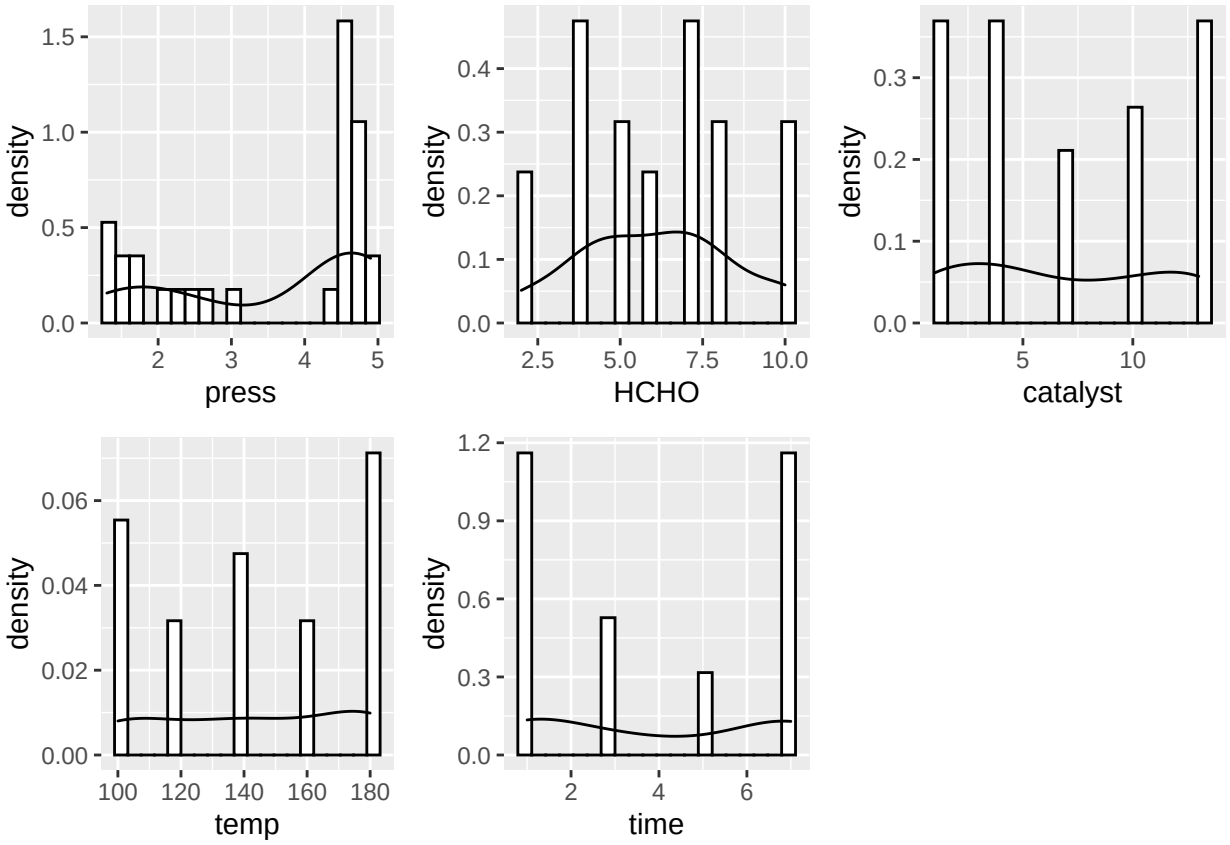
```
##      press            HCHO          catalyst         temp      
##  Min.   :1.300   Min.   : 2.000   Min.   : 1.0   Min.   :100.0  
##  1st Qu.:2.125   1st Qu.: 4.000   1st Qu.: 4.0   1st Qu.:120.0  
##  Median :4.500   Median : 6.000   Median : 7.0   Median :140.0  
##  Mean   :3.560   Mean 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

```
p1 <- ggplot(data2, aes(y = press)) + geom_boxplot() + scale_x_discrete()
p2 <- ggplot(data2, aes(y = HCHO)) + geom_boxplot() + scale_x_discrete()
p3 <- ggplot(data2, aes(y = catalyst)) + geom_boxplot() + scale_x_discrete()
p4 <- ggplot(data2, aes(y = temp)) + geom_boxplot() + scale_x_discrete()
p5 <- ggplot(data2, aes(y = time)) + geom_boxplot() + scale_x_discrete()
gridExtra::grid.arrange(p1,p2,p3,p4,p5, ncol = 5)
```



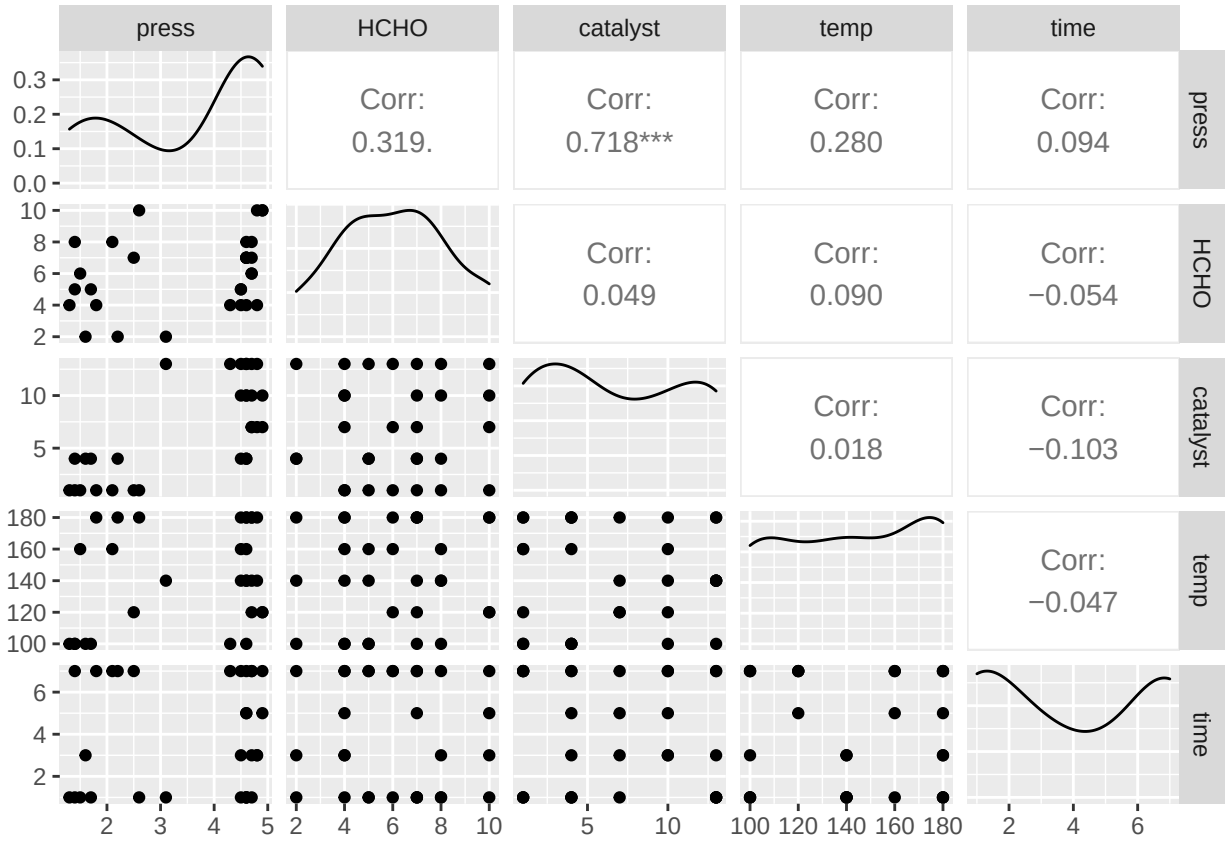
透過 summary 和 boxplot 觀察敘述統計量：
- 所有變數數值皆大於正，並無不合理處。
- press 的數據分佈不均，中位數為 4.5 較平均 3.56 大許多，再綜合對 Q3 和最大值數值的觀察，可發現 press 大多落在 4 到 5 之間。
- HCHO 和 catalyst 的盒鬚圖上下大致對稱，數據分佈較為平均。
- temp 的第三分位數和最大值重疊，表示有四分之一 temp 數值設定都是 180。• time 的第三分位數和最大值重疊，表示有四分之一 time 數值設定都是 7；而第一分位數和最小值重疊，表示另有四分之一 time 數值設定都是 1。

```
myhist <- function(x){
  ggplot(data2, aes_string(x = x)) +
  geom_histogram(aes(y = ..density..), bins = 20, colour = 1, fill = "white") +
  geom_density() + xlab(x)
}
gridExtra::grid.arrange(myhist("press"),myhist("HCHO"),myhist("catalyst"),
                        myhist("temp"),myhist("time"), ncol = 3)
```



透過 histogram 觀察分佈：• press 的數據分布不均，分別集中分佈在 2 以下和 4 到 5 之間。• HCHO 、catalyst 、temp 在個數值的分佈平均，整體並沒有和特定分佈相像。• time 分別集中分佈在 1 和 7 左右。• 從 histogram 裡得到的觀察，與 summary 得到的觀察相符

```
ggpairs(data2, aes(binwidth = 0.5))
```



透過 scatter plot 和相關係數分別觀察散佈特性與相關性：

- 考慮解釋變數之間相關性：
  - Overall：四個解釋變數之間無明顯相關性，解釋變數兩兩之間因設定成特定數值而導致的不自然整齊排列，在散佈圖上看尤其明顯；在此可確認這些變數的數值並非來自帶隨機性的實驗結果，這也符合前面的觀察。
  - Temp vs time 和 catalyst vs time：一般來說，temp 代表的材料硬化反應時需維持之硬化溫度（curing temperature，或稱養護溫度）和 time 代表的硬化時間（curing time）呈負相關。而 catalyst 催化劑也會縮短反應時間，也應和 time 呈負相關。但在這筆資料裡，遠本預期會有負相關的解釋變數之間的相關係數數值都特別低。綜合前一點，這應是實驗設計下的結果，不能貿然拿來推論變數間的特性。
- 考慮反應變數 press 與其他的解釋變數關係：
  - press 和 HCHO、catalyst、temp、time：四個解釋變數，都分別具有正相關性代表四個解釋變數都能一定程度的解釋 press。而正相關性由高到低為 catalyst > HCHO > temp > time，其中只有 catalyst 具顯著性，在 catalyst 沒有和其他變數有共線性的情況下，代表 catalyst 會是解釋 press 時相對重要的變數。
  - press vs catalyst：catalyst 催化劑並不會影響生成物多寡，僅用來加速化學反應進行。不過和增加溫度不同的是，催化劑是透過降低活化能促進反應進行，意即直接降低化學反應所需的能量門檻，而增加溫度並不會像催化劑改變化學反應的路徑。在四個解釋變數沒有共線性，且 time 對 press 或 temp 對 press 相關係數也不高的情況下，我們可能可以推測，比起反應速度，反應路徑是影響 press 相對重要因素之一。不過實際情況還需要了解確切反應，以及不同實驗條件下，可能有所不同的 side reaction 才能確認。

- 另外值得注意的是，從 press 和各個解釋變數之間的散佈圖上也可以看到數據點大多集中在兩側，雖然 HCHO、catalyst 是集中於左下角和右上角的趨勢，但如果要對線性關係作更嚴謹的推論，可能需要再多做一些真實實驗，以取得中間段的數據點。不過，也有可能分散的兩群資料本就代表不同情況，可注意是否有其他重要類別參數被忽略。

2.b

Is this observational or experimental data? Explain your reasoning.

2.b.Ans

資料裡除了 durable press rating，其他應當在數值上同樣擁有的彈性的變數卻有不自然的規律分布，他們更像是作為實驗條件被設定在特定值以使收到的 data 能刻意包含到各種不同情況。這種人為設計下收集的 data，應是 experimental data。

Except for the durable press rating, other variables in this data were deliberately set at specific values to represent different experimental conditions. The data collected under this artificial design should be an experimental data.