Assignment 1

姓名：邱繼賢

系級：統研碩一

學號：110024516

1. a.

拿到資料後先大致觀察一下資料各項變數的各統計量數值，

```
> summary(data1)
      wage              educ           exper            race              smsa
 Min.   :   50.05   Min.   : 0.00   Min.   :-4.0   Min.   :0.00000   Min.   :0.0000
 1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0   1st Qu.:0.00000   1st Qu.:0.0000
 Median :  522.32   Median :12.00   Median :16.0   Median :0.00000   Median :1.0000
 Mean   :  603.73   Mean   :13.07   Mean   :18.2   Mean   :0.07928   Mean   :0.7435
 3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :18777.20   Max.   :18.00   Max.   :63.0   Max.   :1.00000   Max.   :1.0000
       ne                mw               so               we               pt
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
 Mean   :0.2288   Mean   :0.2438   Mean   :0.3111   Mean   :0.2163   Mean   :0.08965
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
```

可發現以下幾項特徵：

(1) wage 此變數可視為 approximately continuous variable，而此變數有著極大的最大值。

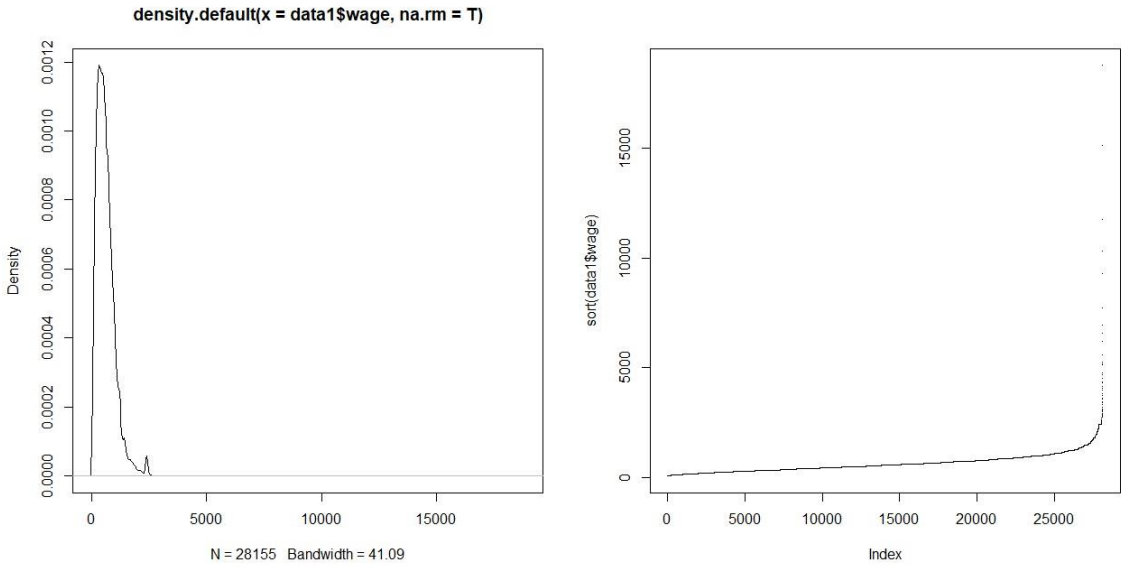
(2) educ 和 exper 為 discrete variable ，exper （工作經驗(年)）中有負數，推測為資料輸入時錯誤，多打一個負號，以下會進行修正。

(3) 剩餘變數皆為以 1 和 0 表示的 categorical variable

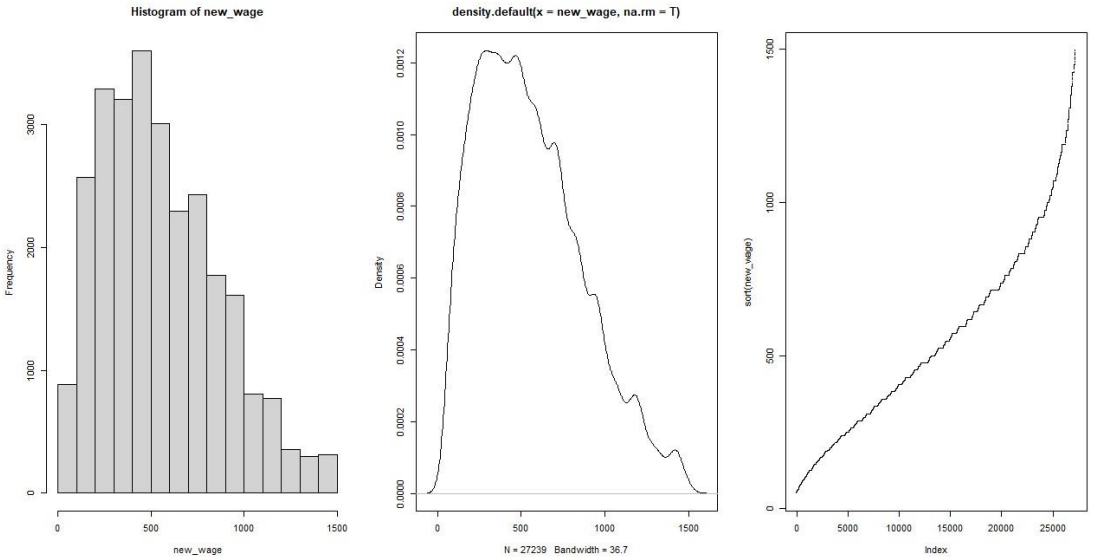
將資料整理過後各統計量呈現如下：

```
> summary(data1)
      wage              educ           exper            race              smsa
 Min.   :   50.05   Min.   : 0.00   Min.   : 0.00   White:25923   Other  : 7223
 1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.00   Black: 2232   in SMSA:20932
 Median :  522.32   Median :12.00   Median :16.00
 Mean   :  603.73   Mean   :13.07   Mean   :18.23
 3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.00
 Max.   :18777.20   Max.   :18.00   Max.   :63.00
       ne               mw               so               we               pt
 Other:21714   Other:21292   Other:19395   Other:22064   No       :25631
 in NE: 6441   in Mw: 6863   in So: 8760   in We: 6091   Part Time: 2524
```

並對 wage 變數進行繪圖，



可觀察因為受到前端極富的極端值影響，圖形呈現為極右偏的現象，將極端值去除後再次繪圖（只取 wage 介於 Q1-1.5*IQR 和 Q3+1.5*IQR 中間的數值），

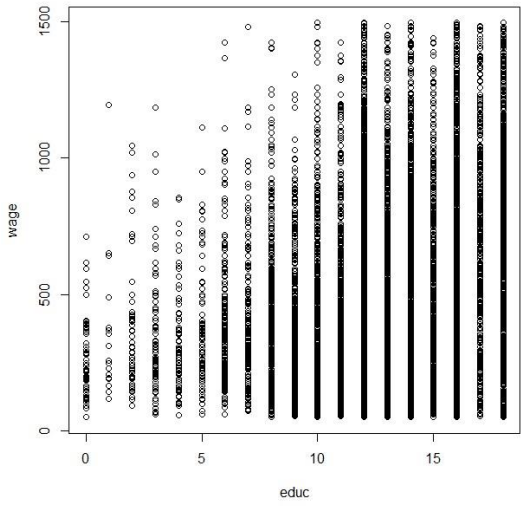


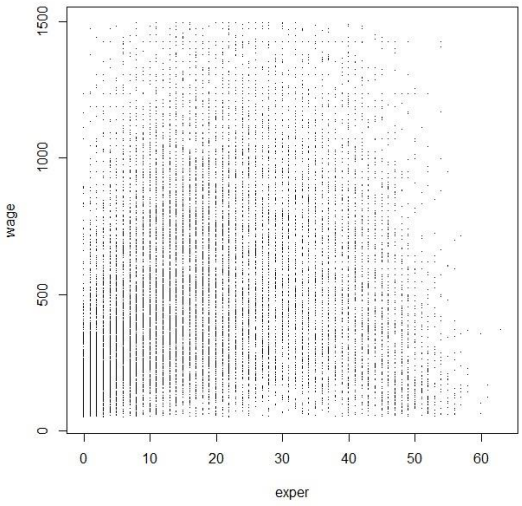
去除極端值後的圖形較為接近常態分配，但依舊有右偏的現象。

接下來將去除極端值後的資料，以 wage 為反應變數，剩下的為解釋變數，進行回歸分析，

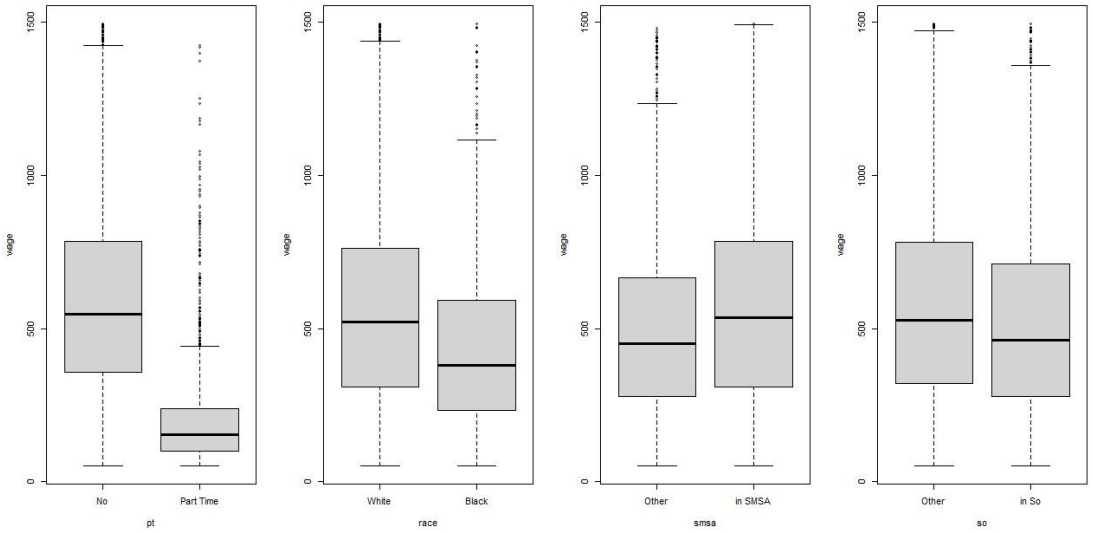
```
Call:
lm(formula = wage ~ educ + exper + race + smsa + ne + mw + so +
    we + pt, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-885.20 -174.47  -25.59  148.10 1193.40

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -160.1214     9.2462 -17.318  < 2e-16 ***
educ          43.8805     0.5701  76.970  < 2e-16 ***
exper          7.8032     0.1248  62.505  < 2e-16 ***
raceBlack    -95.9604     5.7842 -16.590  < 2e-16 ***
smsain SMSA   74.0106     3.5726  20.716  < 2e-16 ***
nein NE        1.4543     4.6924   0.310    0.757
mwin Mw       -6.1763     4.5808  -1.348    0.178
soin So      -33.2752     4.3927  -7.575  3.7e-14 ***
wein We           NA         NA      NA       NA
ptPart Time -348.6836     5.3914 -64.674  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 255.6 on 27230 degrees of freedom
Multiple R-squared:  0.3391,    Adjusted R-squared:  0.3389
F-statistic:  1746 on 8 and 27230 DF,  p-value: < 2.2e-16
```

可發現 we 變數結果皆呈現 NA，這是因為在此筆資料只將全部人分類成五個地區，然而同時也有五個 dummy variable，才會導致此現象，而 educ, exper, race, smsa, so, pt 幾個變數對反應變數皆有顯著貢獻，以下將這些變數皆對 wage 繪圖觀察，





wage, educ, exper 三變數皆為 quantitative variable，以散布圖呈現可看出 wage 有隨著 educ 的上升而提升，但對於 exper 時則圖形上相對不明顯。



pt, race, smsa, so 皆為 categorical variable，故以盒狀圖呈現其對 wage 的影響，可以看出四個變數的變化對 wage 皆有影響，其中更以 pt 和 race 最為明顯。

b.

這筆資料是 observational data，因為這筆資料的每個觀察值皆來自每個人的薪資、學歷、經歷、種族、居住地、工作型態等真實資料，並不是透過實驗去設計出的數據，且此資料的觀測值多達 28155 人，若這全部的數據皆為實驗特別設計的，那將要花費一筆非常驚人的成本，可以推測這筆資料應該是藉由問卷之類的方式，調查並觀察得知。

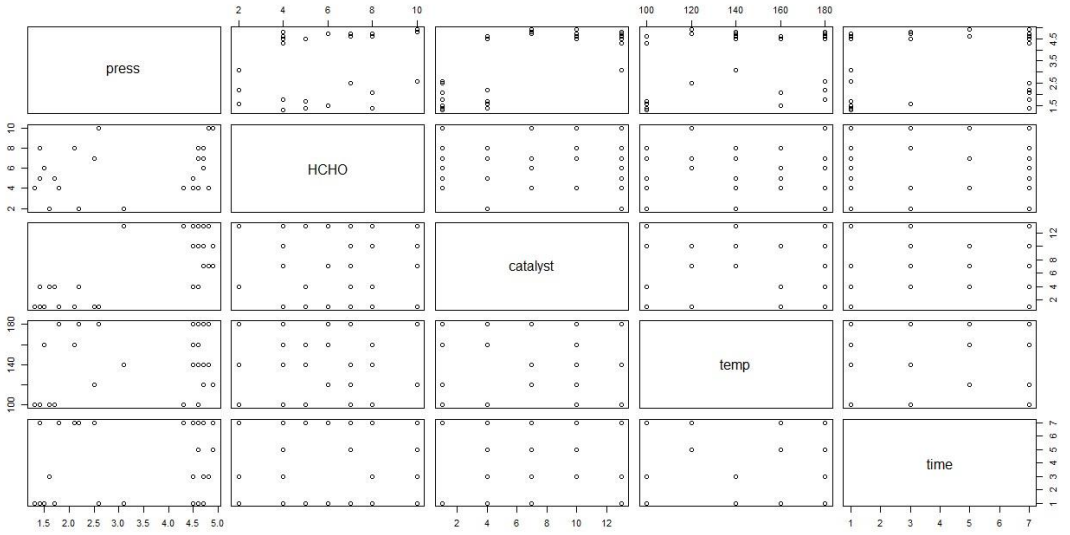
2. a.

一樣拿到資料後大致看一下各變數的各統計量，此筆資料觀測各數相對上一筆來說非常少，只有 30 個，變數也只有 5 個，

```
> summary(data2)
     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
```

此五個變數在定義上皆可視為 continuous variable，但是從獲得的資料上看來，只有 press 變數的數據表示上較接近 continuous variable，其餘四個變數皆更為接近 discrete variable。

將各變數兩兩做散佈圖分析：



可觀察到後面四個變數兩兩之間做圖所畫出的點都非常整齊，但他們個別跟 press 繪圖則無此現象，且可看出 HCHO, catalyst, temp 三變數和 press 皆有正相關趨勢。

以 press 為反應變數，其餘四者為解釋變數，做回歸分析：

```
Call:
lm(formula = press ~ HCHO + catalyst + temp + time, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07876 -0.63939 -0.08531  0.36236  1.65332

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.912212   0.875484  -1.042   0.3074
HCHO         0.160726   0.066166   2.429   0.0227 *
catalyst     0.219783   0.034062   6.452 9.33e-07 ***
temp         0.011226   0.004973   2.257   0.0330 *
time         0.101974   0.058735   1.736   0.0948 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8365 on 25 degrees of freedom
Multiple R-squared:  0.6924,    Adjusted R-squared:  0.6432
F-statistic: 14.07 on 4 and 25 DF,  p-value: 3.845e-06
```

可得到 HCHO, catalyst, temp 三變數對 press 皆有顯著貢獻，其中又以 catalyst 最明顯，與上述觀察圖形所做出的推論一致。

b.

此筆資料為 experimental data，因為這筆資料除了 press 變數之外的其餘四者，在資料數據上都是非常整齊的整數，與日常的生活經驗不符，這很明顯就是做實驗時特別設計出來，而 press 就是反應變數，在其他變數設計並固定好之下所做出來的結果，但其實就算是設計好的實驗數據，也不太可能真的就是這樣固定準確的整數值，因為不管是在測量還是控制變因的時候，都一定還是存在著誤差，但在處理實驗數據的時候，會忽略這些誤差，視為完全準確。