HWDiamond Price

高嘉妤、柯堯珹、吳承恩、趙友誠

2024-11-20

Table of contents

0. 資料簡介 **1**

1.Data Preprocessing **2**

2.Data visualization for exploratory data analysis **4**

3.Construct a predictive model for price **8**

CCA 8

Price model 10

```
library(readr)
library(psych)
library(Hmisc)
library(DataExplorer)
library(ggplot2)
library(MASS)
library(car)
library(stargazer)
Diamonds_Prices2022 <- read_csv("Diamonds Prices2022.csv")
data <- Diamonds_Prices2022
```

0. 資料簡介

Dimension of the Data : ***53943 samples x 11 columns***

Variables	Explanation	remark
carat	克拉(重量)	連續變數(公克)
cut	切工	類別變數, Fair,Good,Ideal,Premium,Very Good
color	顏色	類別變數, D,E,F,G,H,I,J 無色(D~F),近乎無色(G~J)
clarity	淨度	類別變數, IF: 內部無暇,VVS1: 極輕微瑕,VS1: 輕微內含物 1,VS2: 輕微內含物 2,SI1: 微內含物 1,SI2: 微內含物 2,I1: 內含物
depth	深度	連續變數 (mm)
table	檯面尺寸	連續變數
price	價格	連續變數
x	鑽石的長	連續變數 (mm)
y	鑽石的寬	連續變數 (mm)
z	鑽石的高	連續變數 (mm)

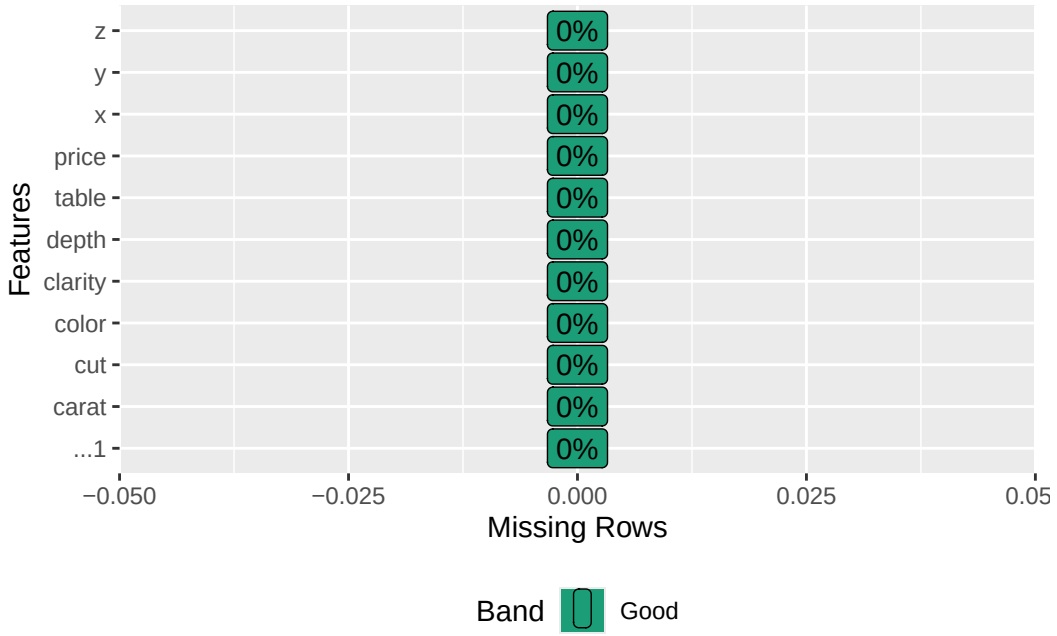
1.Data Preprocessing

```
describe(Diamonds_Prices2022)
```

```
Diamonds_Prices2022

 11  Variables      53943  Observations
--------------------------------------------------------------------------------
...1
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0    53943        1    26972    17981     2698     5395
     .25      .50      .75      .90      .95
   13486    26972    40458    48549    51246

lowest :     1     2     3     4     5, highest: 53939 53940 53941 53942 53943
--------------------------------------------------------------------------------
carat
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      273    0.999   0.7979   0.5122     0.30     0.31
     .25      .50      .75      .90      .95
    0.40     0.70     1.04     1.51     1.70

lowest : 0.2  0.21 0.22 0.23 0.24, highest: 4    4.01 4.13 4.5  5.01
--------------------------------------------------------------------------------
cut
       n  missing distinct
   53943        0        5

Value           Fair      Good     Ideal   Premium Very Good
Frequency       1610      4906     21551     13793     12083
Proportion     0.030     0.091     0.400     0.256     0.224
--------------------------------------------------------------------------------
color
       n  missing distinct
   53943        0        7

Value          D     E     F     G     H     I     J
Frequency   6775  9799  9543 11292  8304  5422  2808
Proportion 0.126 0.182 0.177 0.209 0.154 0.101 0.052
--------------------------------------------------------------------------------
clarity
       n  missing distinct
   53943        0        8

Value         I1    IF   SI1   SI2   VS1   VS2  VVS1  VVS2
Frequency    741  1790 13067  9194  8171 12259  3655  5066
Proportion 0.014 0.033 0.242 0.170 0.151 0.227 0.068 0.094
--------------------------------------------------------------------------------
depth
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      184    0.999    61.75    1.515     59.3     60.0
     .25      .50      .75      .90      .95
    61.0     61.8     62.5     63.3     63.8
```

```
lowest : 43   44   50.8 51   52.2, highest: 72.2 72.9 73.6 78.2 79
--------------------------------------------------------------------------------
table
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      127     0.98    57.46    2.448       54       55
     .25      .50      .75      .90      .95
      56       57       59       60       61

lowest : 43   44   49   50   50.1, highest: 71   73   76   79   95
--------------------------------------------------------------------------------
price
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0    11602        1     3933     4012      544      646
     .25      .50      .75      .90      .95
     950     2401     5324     9821    13107

lowest :   326   327   334   335   336, highest: 18803 18804 18806 18818 18823
--------------------------------------------------------------------------------
x
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      554        1    5.731    1.276     4.29     4.36
     .25      .50      .75      .90      .95
    4.71     5.70     6.54     7.31     7.66

lowest : 0     3.73  3.74  3.76  3.77 , highest: 10.01 10.02 10.14 10.23 10.74
--------------------------------------------------------------------------------
y
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      552        1    5.735    1.269     4.30     4.36
     .25      .50      .75      .90      .95
    4.72     5.71     6.54     7.30     7.65

lowest : 0     3.68  3.71  3.72  3.73 , highest: 10.1  10.16 10.54 31.8  58.9
--------------------------------------------------------------------------------
z
       n  missing distinct     Info     Mean      Gmd      .05      .10
   53943        0      375        1    3.539   0.7901     2.65     2.69
     .25      .50      .75      .90      .95
    2.91     3.53     4.04     4.52     4.73

lowest : 0    1.07 1.41 1.53 2.06, highest: 6.43 6.72 6.98 8.06 31.8
--------------------------------------------------------------------------------
```

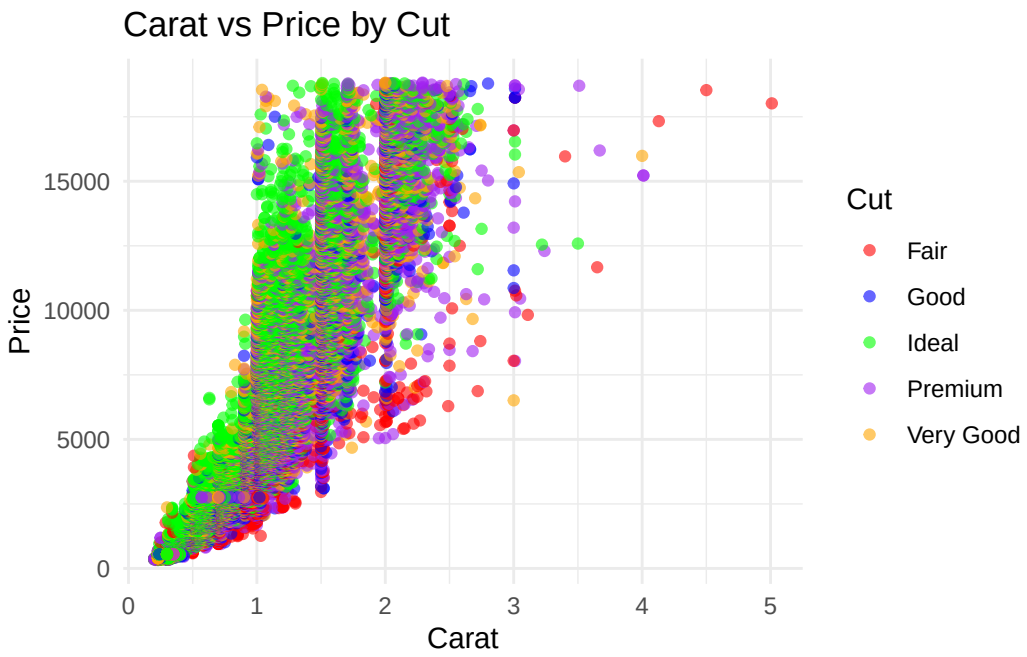
```
DataExplorer::plot_missing(Diamonds_Prices2022)
```



此資料集中未出現缺失值

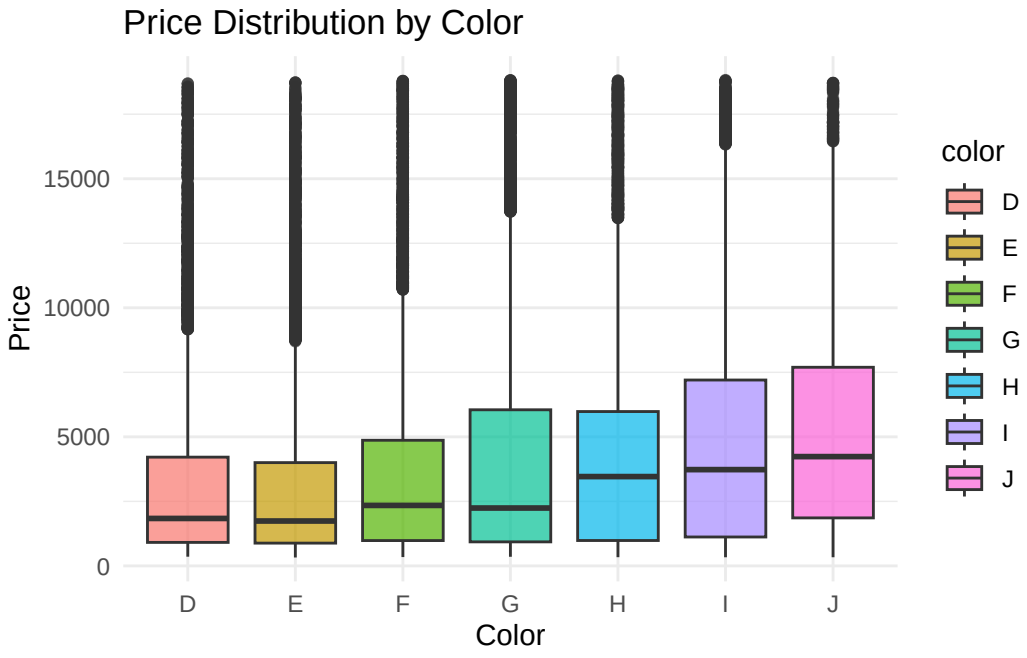
2.Data visualization for exploratory data analysis

```
# 克拉對價格（加切工）
ggplot(data, aes(x = carat, y = price, color = factor(cut))) +
  geom_point(alpha = 0.6) +
  labs(title = "Carat vs Price by Cut",
       x = "Carat",
       y = "Price",
       color = "Cut") +
  scale_color_manual(values = c( "Fair" = "red", "Good" = "blue", "Ideal" = "green", "Premium" = "purple
  theme_minimal()
```



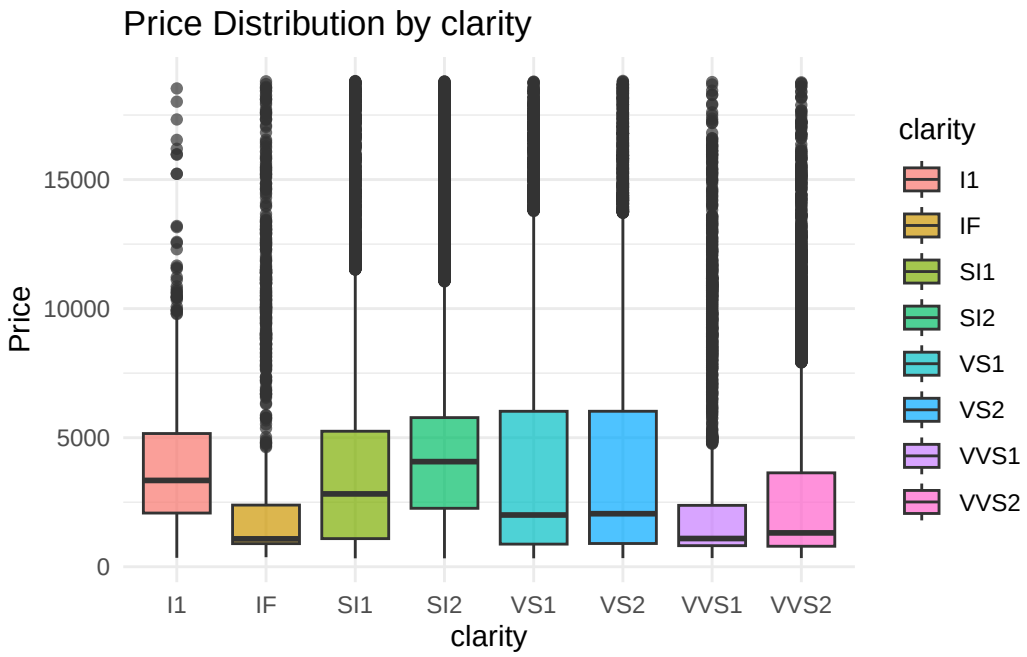
從克拉對價格的圖中可發現大致上越重的鑽石價格越高

```
# 顏色對價格圖

ggplot(data, aes(x = color, y = price, fill = factor(color))) +
    geom_boxplot(alpha = 0.7) +
    labs(title = "Price Distribution by Color",
         x = "Color",
         y = "Price",
         fill = "color") +
    theme_minimal()
```



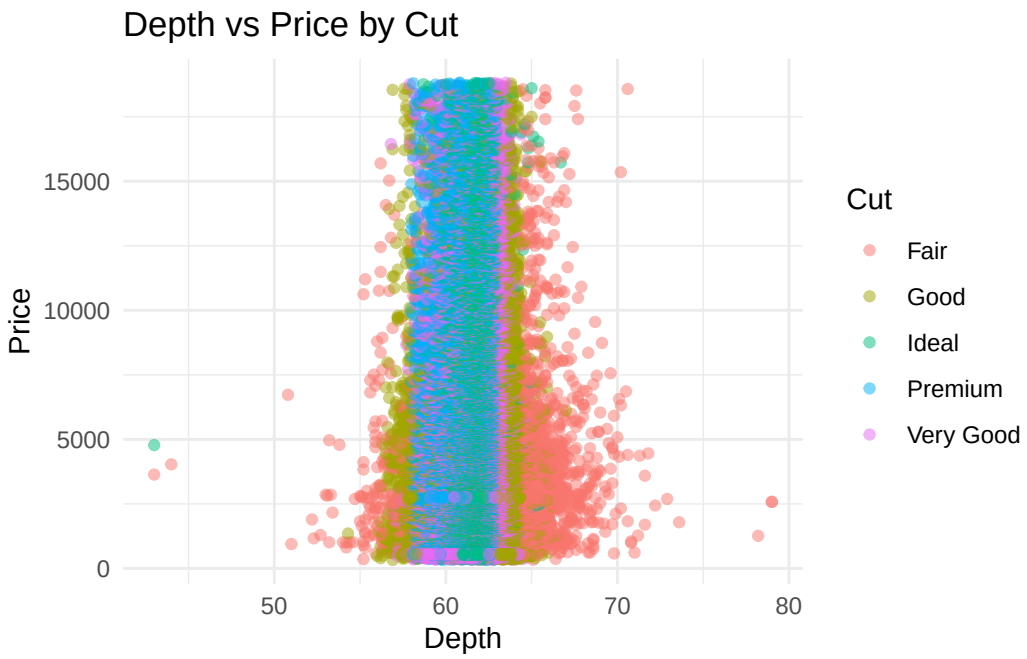
從顏色對價格的圖中可發現當分類越靠近接近無色時價格越高

```
# 淨度對價格

ggplot(data, aes(x = clarity, y = price, fill = factor(clarity))) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Price Distribution by clarity",
       x = "clarity",
       y = "Price",
       fill = "clarity") +
  theme_minimal()
```



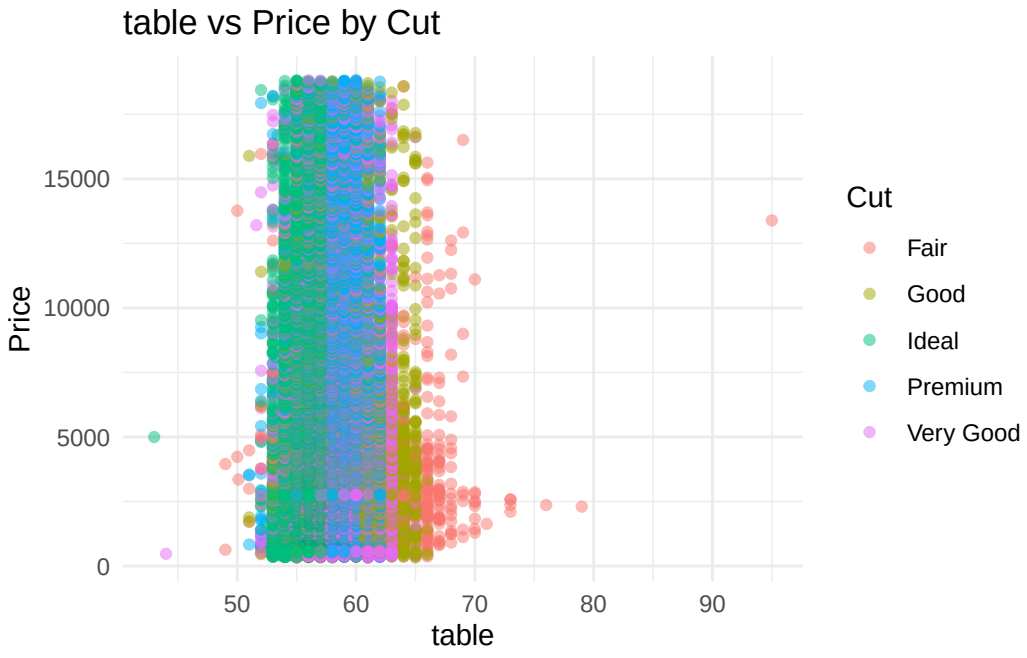
從淨度對價格的圖可發現單一淨度指標對價格並沒有直接關連, 高淨度的鑽石未必會有高價格

```
# 深度對價格
ggplot(data, aes(x = depth, y = price, color =factor(cut))) +
  geom_point(alpha = 0.5) +
  labs(title = "Depth vs Price by Cut",
       x = "Depth",
       y = "Price",
       color = "Cut") +
  theme_minimal()
```



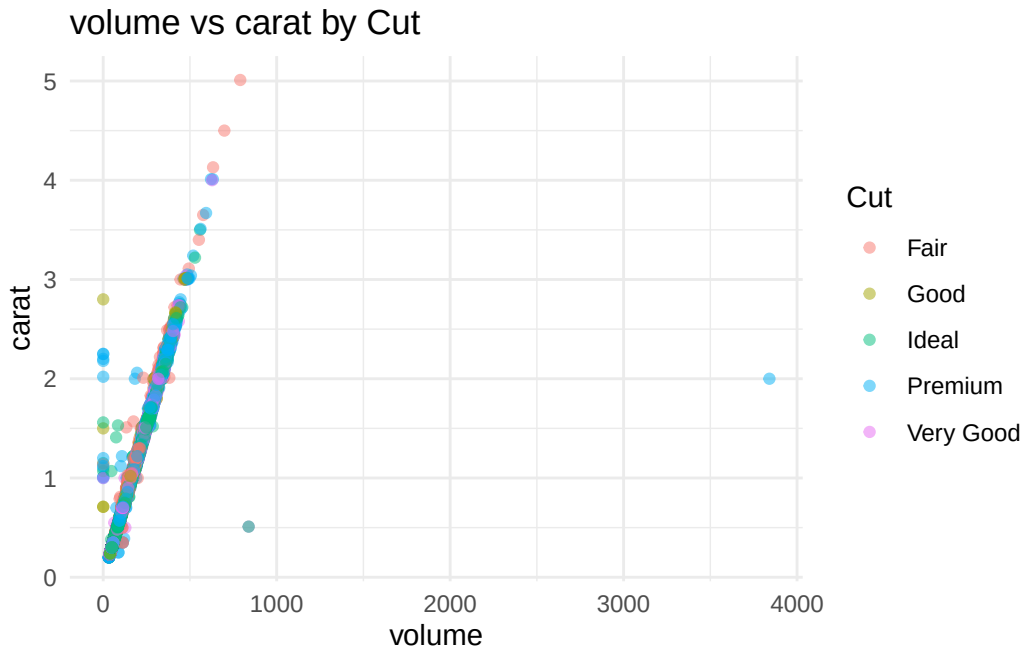
從深度對價格的圖中可發現深度和價格沒有相關，且深度大多集中於 60 附近，推測是因為深度比例在此區間能切割出最明亮的鑽石

```
# 檯面尺寸對價格
ggplot(data, aes(x = table, y = price, color =factor(cut))) +
  geom_point(alpha = 0.5) +
  labs(title = "table vs Price by Cut",
       x = "table",
       y = "Price",
       color = "Cut") +
  theme_minimal()
```



從檯面尺寸對價格圖中可發現檯面尺寸對價格沒有相關, 且檯面尺寸約集中在 56~62 之間, 推測也是在這個區間中能切割出最好的鑽石

```
# 體積對重量 (體積 =x*y*z)
data$volume <- data$x * data$y * data$z
ggplot(data, aes(x = volume, y = carat, color =factor(cut))) +
  geom_point(alpha = 0.5) +
  labs(title = "volume vs carat by Cut",
       x = "volume",
       y = "carat",
       color = "Cut") +
  theme_minimal()
```



3.Construct a predictive model for price

```
# 定義類別順序
levelcut <- c("Fair", "Good", "Ideal", "Premium", "Very Good")
levelcolor <- c("D","E","F","G","H","I","J")
levelclarity <- c("I1","SI2","SI1","VS2","VS1","VVS2","VVS1","IF")

# 使用 match 進行編碼
data$cut <- match(data$cut,levelcut)
data$color <- match(data$color,levelcolor)
data$clarity <- match(data$clarity,levelclarity)
```

CCA

```
# 欲分析幾何特性 vs. 做工及價格之間的關係
# 選擇兩組變數
X <- data[, c("carat", "color", "clarity", "volume")]
Y <- data[, c("price", "cut", "depth", "table")]
```

```r
cca <- cancor(X,Y)
print(cca)
```

```
$cor
[1] 0.95139791 0.21122911 0.05451472 0.01018712

$xcoef
                 [,1]          [,2]          [,3]          [,4]
carat   -9.801127e-03 -0.0190491190  0.0252977415 -0.0263102116
color    3.574531e-04 -0.0005972323  0.0015054342  0.0020970765
clarity -5.878458e-04  0.0023280863  0.0013122705 -0.0006956651
volume  -7.854915e-07  0.0001289852 -0.0001582564  0.0001516935

$ycoef
               [,1]          [,2]          [,3]          [,4]
price -1.071610e-06  1.700518e-07  9.164083e-08 -9.077403e-09
cut    9.156390e-05  1.926362e-04 -1.889023e-04  4.284068e-03
depth -9.371534e-05 -2.323161e-03  2.063865e-03  7.065729e-04
table -1.020052e-04 -1.746995e-03 -1.043162e-03 -1.598028e-04

$xcenter
      carat       color     clarity      volume
  0.7979347   3.5941271   4.0509797 129.8485391

$ycenter
      price         cut       depth       table
3932.734294    3.553047   61.749322   57.457251
```

```r
cca$cor
```

```
[1] 0.95139791 0.21122911 0.05451472 0.01018712
```

```r
cca$xcoef;cca$ycoef
```

```
                 [,1]          [,2]          [,3]          [,4]
carat   -9.801127e-03 -0.0190491190  0.0252977415 -0.0263102116
color    3.574531e-04 -0.0005972323  0.0015054342  0.0020970765
clarity -5.878458e-04  0.0023280863  0.0013122705 -0.0006956651
volume  -7.854915e-07  0.0001289852 -0.0001582564  0.0001516935
```

```
               [,1]          [,2]          [,3]          [,4]
price -1.071610e-06  1.700518e-07  9.164083e-08 -9.077403e-09
cut    9.156390e-05  1.926362e-04 -1.889023e-04  4.284068e-03
depth -9.371534e-05 -2.323161e-03  2.063865e-03  7.065729e-04
table -1.020052e-04 -1.746995e-03 -1.043162e-03 -1.598028e-04
```

欲分析幾何特性 vs. 做工及價格之間的關係

第一典型相關變數: 最大典型相關係數為 0.9513, 第一典型變數主要由 carat 和 table 貢獻組成

第二典型相關變數: 最大典型相關係數為 0.2112(相關性低)

```r
X_loadinds <- cor(X,as.matrix(X) %*% cca$xcoef)
Y_loadinds <- cor(Y,as.matrix(Y) %*% cca$ycoef)
X_loadinds;Y_loadinds
```

```
              [,1]        [,2]       [,3]       [,4]
carat   -0.9724210 -0.19165693 -0.0265517  0.1302304
```

```
color   -0.1830604 -0.15797158  0.6017737  0.7611848
clarity  0.1643141  0.82509439  0.5154062 -0.1630230
volume  -0.9508616 -0.07424404 -0.1590087  0.2550811
```

```
              [,1]        [,2]         [,3]        [,4]
price -0.998417597  0.05235614  0.006992934 0.019293413
cut   -0.019609238  0.06612605 -0.256512293 0.964076899
depth -0.009214484 -0.51541247  0.854714376 0.061060730
table -0.166643740 -0.65107003 -0.740498238 0.000185882
```

第一典型變數主要受 carat(-),volume(-) 和 price(-) 影響

第二典型變數主要受 clarity(+),depth(-) 和 table(-) 影響

Price model

```r
model <- lm(price ~ carat + cut + color + clarity + depth + table  + x + y + z, data = data)

# 使用 stepAIC 進行變數選擇
step_model <- stepAIC(model, direction = "both")
```

```
Start:  AIC=766611.1
price ~ carat + cut + color + clarity + depth + table + x + y +
    z

          Df  Sum of Sq        RSS    AIC
- y        1 2.3244e+06 8.0124e+10 766611
- z        1 2.9176e+06 8.0125e+10 766611
<none>                  8.0122e+10 766611
- cut      1 3.7035e+08 8.0492e+10 766858
- depth    1 7.6453e+08 8.0886e+10 767121
- x        1 8.4887e+08 8.0971e+10 767178
- table    1 9.0254e+08 8.1024e+10 767213
- color    1 1.4456e+10 9.4578e+10 775557
- clarity  1 3.1005e+10 1.1113e+11 784255
- carat    1 6.3464e+10 1.4359e+11 798079

Step:  AIC=766610.6
price ~ carat + cut + color + clarity + depth + table + x + z

          Df  Sum of Sq        RSS    AIC
- z        1 2.2983e+06 8.0126e+10 766610
<none>                  8.0124e+10 766611
+ y        1 2.3244e+06 8.0122e+10 766611
- cut      1 3.7191e+08 8.0496e+10 766858
- depth    1 7.7868e+08 8.0903e+10 767130
- table    1 9.0725e+08 8.1031e+10 767216
- x        1 1.0512e+09 8.1175e+10 767312
- color    1 1.4457e+10 9.4581e+10 775557
- clarity  1 3.1013e+10 1.1114e+11 784258
- carat    1 6.3549e+10 1.4367e+11 798109

Step:  AIC=766610.2
price ~ carat + cut + color + clarity + depth + table + x
```

```
          Df  Sum of Sq        RSS    AIC
<none>                  8.0126e+10 766610
+ z        1 2.2983e+06 8.0124e+10 766611
+ y        1 1.7051e+06 8.0125e+10 766611
- cut      1 3.7093e+08 8.0497e+10 766857
- table    1 9.0577e+08 8.1032e+10 767215
- depth    1 1.0378e+09 8.1164e+10 767302
- x        1 2.2159e+09 8.2342e+10 768080
- color    1 1.4455e+10 9.4582e+10 775555
- clarity  1 3.1012e+10 1.1114e+11 784257
- carat    1 6.3548e+10 1.4367e+11 798108
```

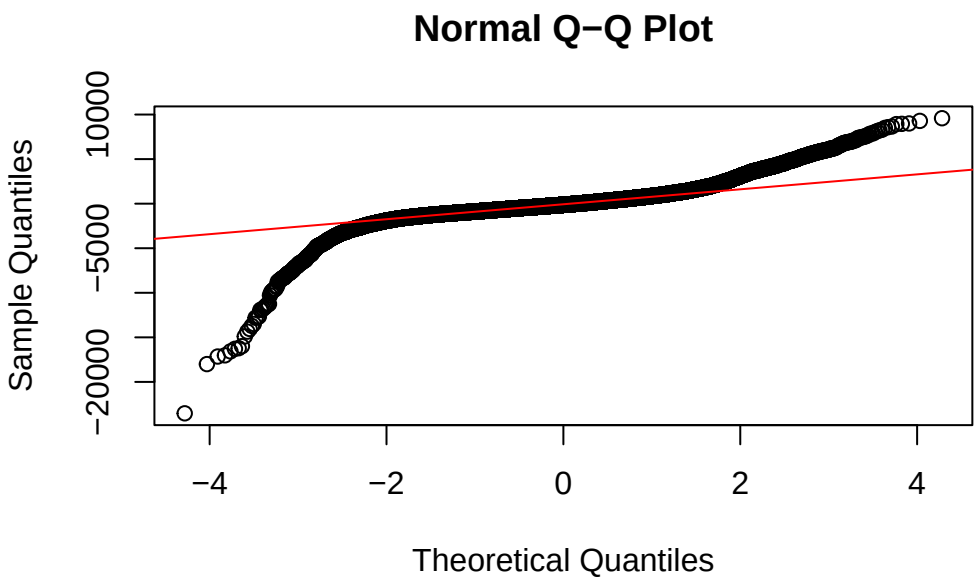
```
summary(step_model)
```

```

Call:
lm(formula = price ~ carat + cut + color + clarity + depth +
    table + x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-23527.3   -633.6   -129.1    498.4   9588.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9237.931    353.706   26.12   <2e-16 ***
carat       10736.731     51.913  206.82   <2e-16 ***
cut            82.687      5.233   15.80   <2e-16 ***
color        -322.071      3.265  -98.64   <2e-16 ***
clarity       508.019      3.516  144.48   <2e-16 ***
depth        -106.726      4.038  -26.43   <2e-16 ***
table         -62.768      2.542  -24.69   <2e-16 ***
x            -849.719     22.001  -38.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1219 on 53935 degrees of freedom
Multiple R-squared:  0.9067,    Adjusted R-squared:  0.9067
F-statistic: 7.485e+04 on 7 and 53935 DF,  p-value: < 2.2e-16
```

```
vif(step_model)
```

```
    carat       cut     color   clarity     depth     table         x
21.984883  1.050132  1.120086  1.217850  1.215152  1.171580 22.115606
```

```
qqnorm(resid(step_model))
qqline(resid(step_model), col = "red")
```




```
model2 <- lm(price ~  carat +cut + color + clarity + depth + table    , data = data)
summary(model2)
```

```
Call:
lm(formula = price ~ carat + cut + color + clarity + depth +
    table, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-19674.8   -695.2   -170.1    557.7   9324.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3474.299    325.082   10.69   <2e-16 ***
carat       8791.876     12.786  687.64   <2e-16 ***
cut           87.202      5.303   16.44   <2e-16 ***
color       -317.578      3.308  -96.01   <2e-16 ***
clarity      528.008      3.526  149.77   <2e-16 ***
depth        -69.524      3.975  -17.49   <2e-16 ***
table        -62.154      2.577  -24.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1236 on 53936 degrees of freedom
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.9041
F-statistic: 8.473e+04 on 6 and 53936 DF,  p-value: < 2.2e-16
```

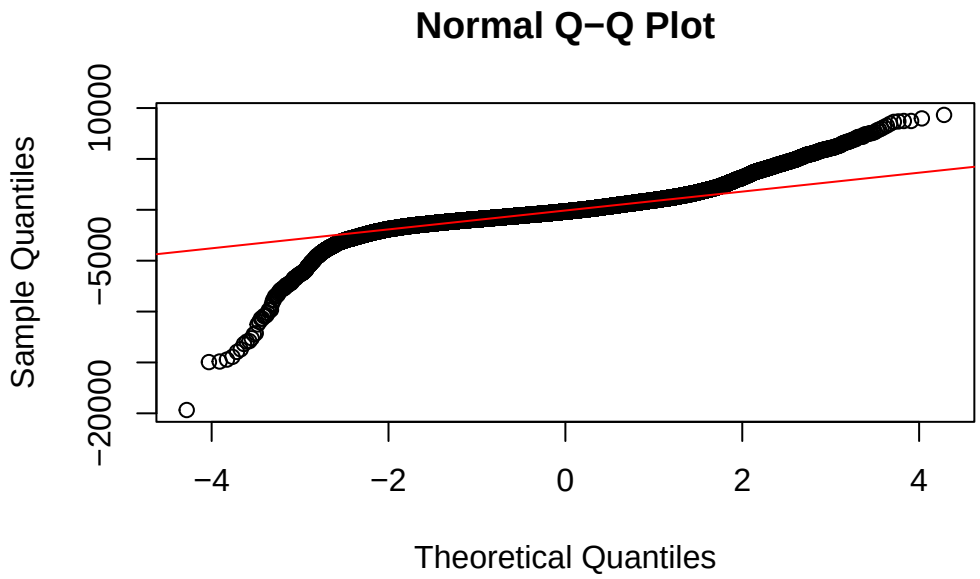
```
vif(model2)
```

```
   carat      cut    color  clarity    depth    table
1.297723 1.049608 1.118664 1.191464 1.146007 1.171534
```

```r
mean(resid(model2))
```

```
[1] 8.167887e-13
```

```r
qqnorm(resid(model2))
qqline(resid(model2), col = "red")
```




```r
model2 <- lm(price ~  carat +cut + color + clarity + depth + table    , data = data)
summary(model2)
```

```
Call:
lm(formula = price ~ carat + cut + color + clarity + depth +
    table, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-19674.8   -695.2   -170.1    557.7   9324.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3474.299    325.082   10.69   <2e-16 ***
carat       8791.876     12.786  687.64   <2e-16 ***
cut           87.202      5.303   16.44   <2e-16 ***
color       -317.578      3.308  -96.01   <2e-16 ***
clarity      528.008      3.526  149.77   <2e-16 ***
depth        -69.524      3.975  -17.49   <2e-16 ***
table        -62.154      2.577  -24.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1236 on 53936 degrees of freedom
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.9041
```

```
F-statistic: 8.473e+04 on 6 and 53936 DF,  p-value: < 2.2e-16
```

```
vif(model2)
```

```
   carat      cut    color  clarity    depth    table 
1.297723 1.049608 1.118664 1.191464 1.146007 1.171534 
```

由於 step_model 選取的模型中, 經由 VIF 檢查有兩個變數 (carat 和 x) 出現多重共線性, 因此剔除 x 改成 model2 而 model2 的 R-squared = 0.9041