HWDiamond Price

高嘉妤、柯堯珹、吳承恩、趙友誠

11/22/24

Table of contents

0. 資料簡介 **1**

1.Data Preprocessing **2**

2.Data visualization for exploratory data analysis **3**

3.Construct a predictive model for price **6**

CCA 6

Price model 9

Kaggle URL: https://www.kaggle.com/datasets/enashed/diamond-prices/discussion/547512

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

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)

Variables	Explanation	remark
y	鑽石的寬	連續變數 (mm)
z	鑽石的高	連續變數 (mm)

1.Data Preprocessing

```
latex(describe(data, title="Diamond Price Dataset"), title="", file="")
```

data
11 Variables 53943 Observations

V1

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
53943	0	53943	1	26972	17981	2698	5395	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	.25	.50	.75	.90	.95
53943	0	273	0.999	0.7979	0.5122	0.30	0.31	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	.25	.50	.75	.90	.95
53943	0	184	0.999	61.75	1.515	59.3	60.0	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	.25	.50	.75	.90	.95
53943	0	127	0.98	57.46	2.448	54	55	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	.25	.50	.75	.90	.95
53943	0	11602	1	3933	4012	544	646	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	.25	.50	.75	.90	.95
53943	0	554	1	5.731	1.276	4.29	4.36	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	.25	.50	.75	.90	.95
53943	0	552	1	5.735	1.269	4.30	4.36	4.72	5.71	6.54	7.30	7.65

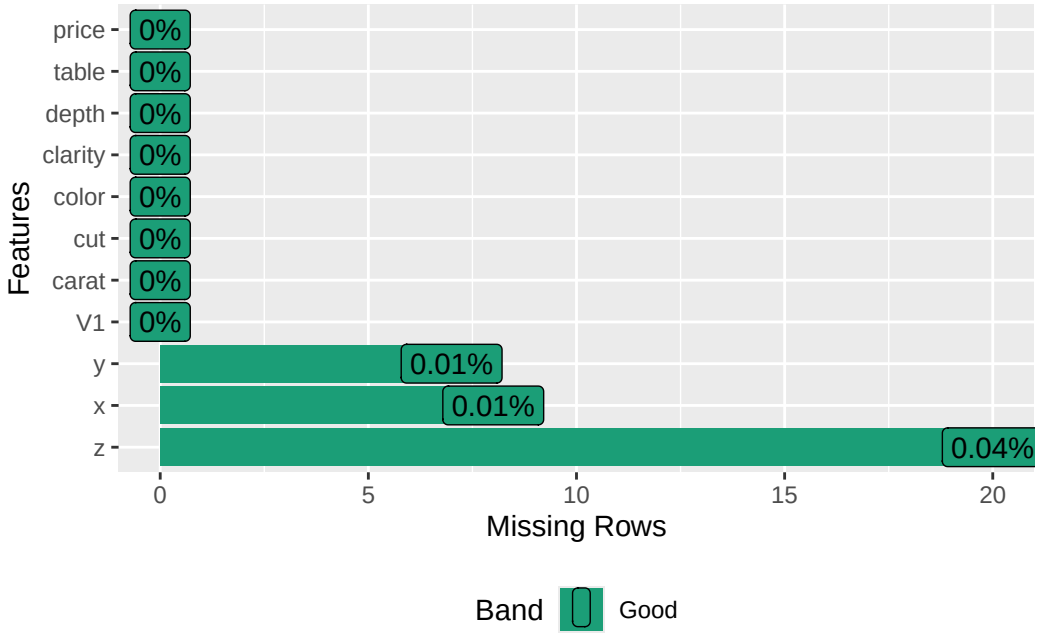
```
lowest : 0     3.68  3.71  3.72  3.73 , highest: 10.1  10.16 10.54 31.8  58.9
```

z

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
53943	0	375	1	3.539	0.7901	2.65	2.69	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
```

```r
data$x[data$x==0] <- NA
data$y[data$y==0] <- NA
data$z[data$z==0] <- NA
DataExplorer::plot_missing(data)
```



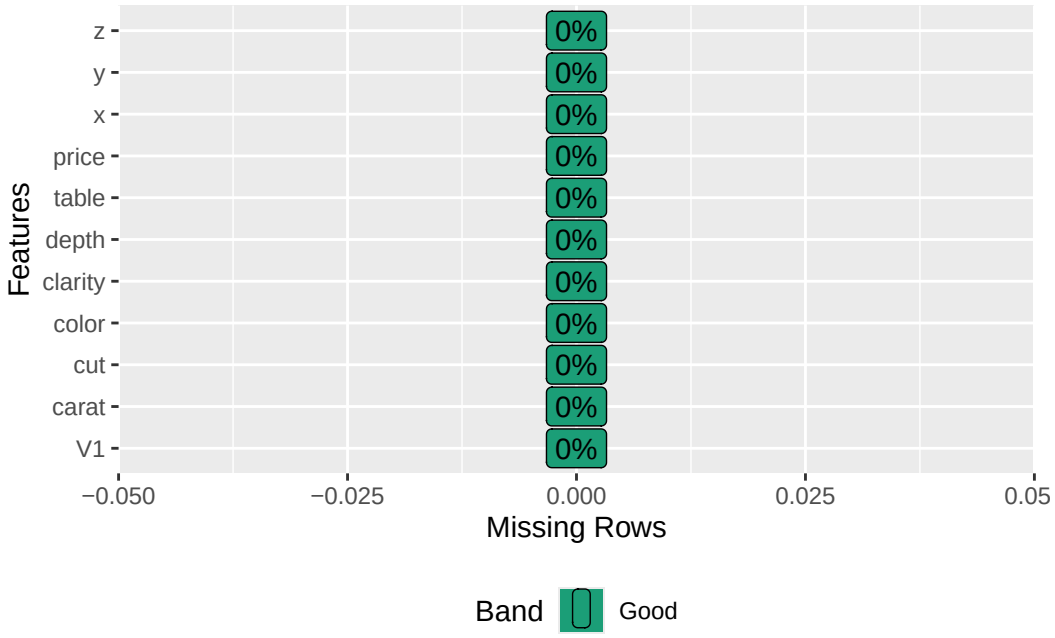

```r
data <- data[!(is.na(data$x) | is.na(data$y) | is.na(data$z)),]
DataExplorer::plot_missing(data)
```

此資料集中有 20 筆資料的長或寬或高為 0，這是不合理的量測數值，因此將其作為遺失值處理。

又因資料僅 20 筆，故刪除。

2.Data visualization for exploratory data analysis

```r
# 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" = "pur
  theme_minimal()
```

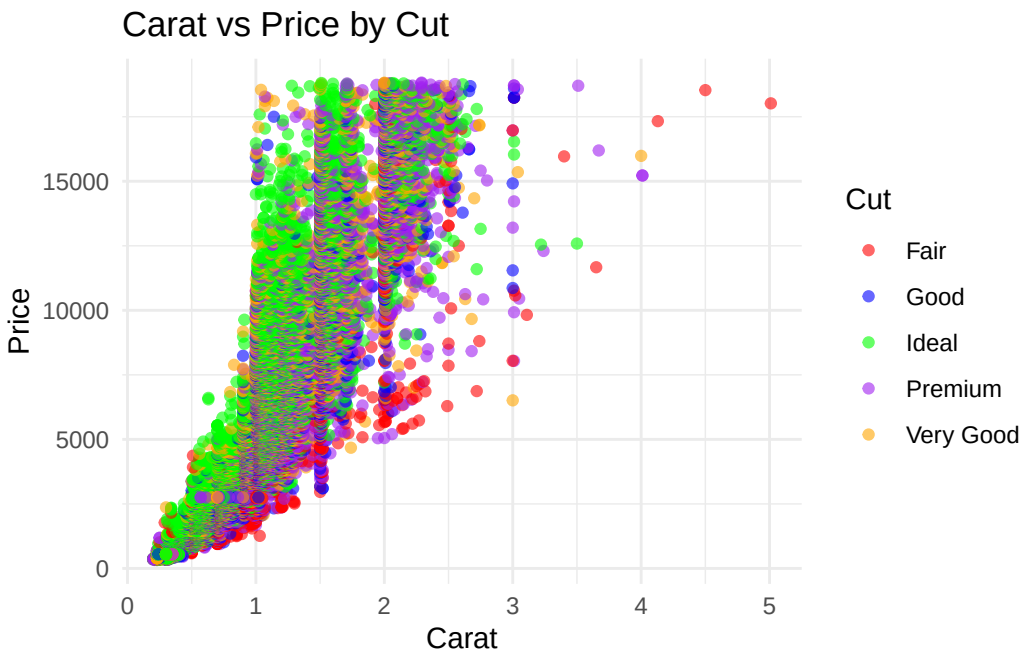
從克拉對價格的圖中可發現大致上越重的鑽石價格越高

```
# 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



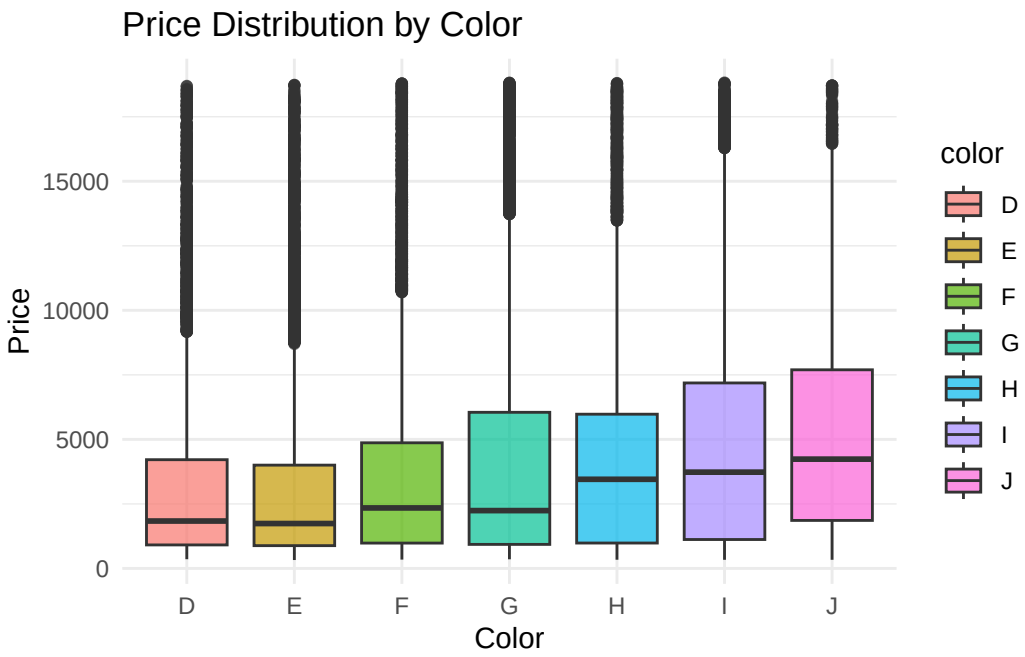

```
# 深度對價格
ggplot(data, aes(x = depth, y = price, color =factor(cut))) +
  geom_point(alpha = 0.5) +
  labs(title = "Depth vs Price by Cut",
       x = "Depth",
       y = "Price",
       color = "Cut") +
  theme_minimal()
```

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()
```

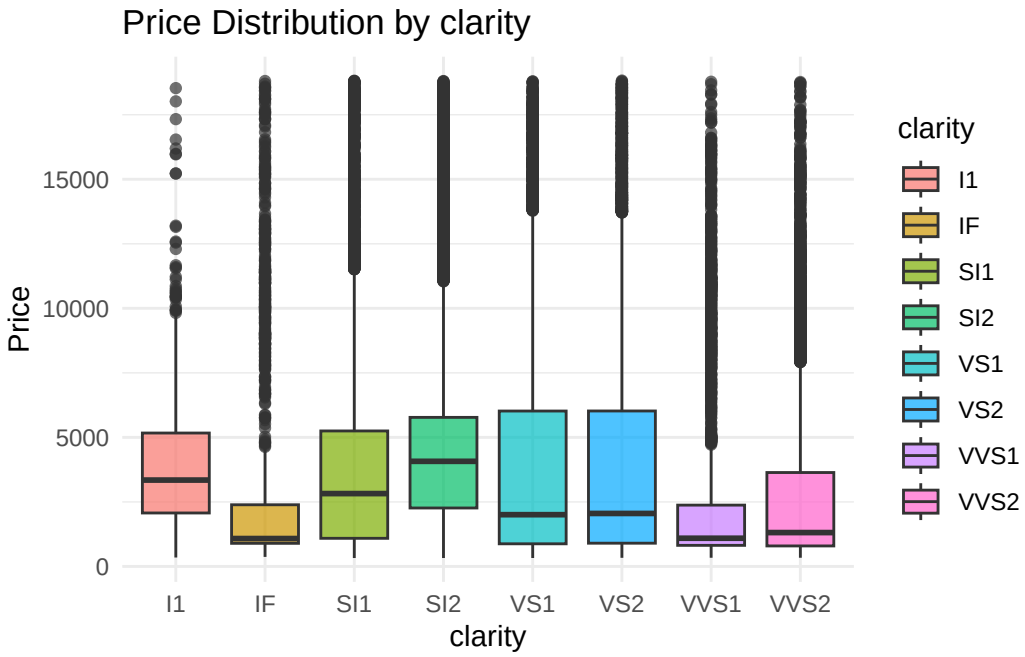
3.Construct a predictive model for price

```
# 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")]

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

```
$cor
[1] 0.95138646 0.21406636 0.05580885 0.00958912
```




```
$xcoef
                [,1]          [,2]          [,3]          [,4]
carat   -9.782523e-03 -0.0209243423  0.0266436152 -0.0266250552
color    3.575074e-04 -0.0005933112  0.0014501269  0.0021370516
clarity -5.878241e-04  0.0022925087  0.0013710461 -0.0007033020
volume  -9.360728e-07  0.0001401503 -0.0001657836  0.0001531097

$ycoef
               [,1]          [,2]          [,3]          [,4]
price -1.072401e-06  1.692041e-07  9.347422e-08 -9.412873e-09
cut    9.137155e-05  1.872188e-04 -1.583651e-04  4.286575e-03
depth -9.380682e-05 -2.346677e-03  2.044904e-03  6.890120e-04
table -1.018839e-04 -1.736365e-03 -1.062866e-03 -1.566226e-04

$xcenter
      carat       color     clarity      volume
  0.7976932   3.5939581   4.0514623 129.8966998

$ycenter
      price         cut       depth       table
3930.927879    3.553122   61.749432   57.456902
```

```
cca$cor
```

```
[1] 0.95138646 0.21406636 0.05580885 0.00958912
```

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

```
                [,1]          [,2]          [,3]          [,4]
carat   -9.782523e-03 -0.0209243423  0.0266436152 -0.0266250552
```


```
color    3.575074e-04 -0.0005933112  0.0014501269  0.0021370516
clarity -5.878241e-04  0.0022925087  0.0013710461 -0.0007033020
volume  -9.360728e-07  0.0001401503 -0.0001657836  0.0001531097
```

```
               [,1]          [,2]          [,3]          [,4]
price -1.072401e-06  1.692041e-07  9.347422e-08 -9.412873e-09
cut    9.137155e-05  1.872188e-04 -1.583651e-04  4.286575e-03
depth -9.380682e-05 -2.346677e-03  2.044904e-03  6.890120e-04
table -1.018839e-04 -1.736365e-03 -1.062866e-03 -1.566226e-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)

```
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   -0.9724303 -0.18934867 -0.03251382  0.1321716
color   -0.1829844 -0.15802158  0.58347916  0.7753051
clarity  0.1642906  0.81271752  0.53429466 -0.1644021
volume  -0.9528425 -0.07286613 -0.15970412  0.2475407
```

```
              [,1]        [,2]        [,3]        [,4]
price -0.998421874  0.05219380  0.00759783 0.019283209
cut   -0.019690093  0.06683905 -0.24962881 0.965831403
depth -0.009170247 -0.52445980  0.84964025 0.054491046
table -0.166668018 -0.64330017 -0.74724181 0.004042751
```

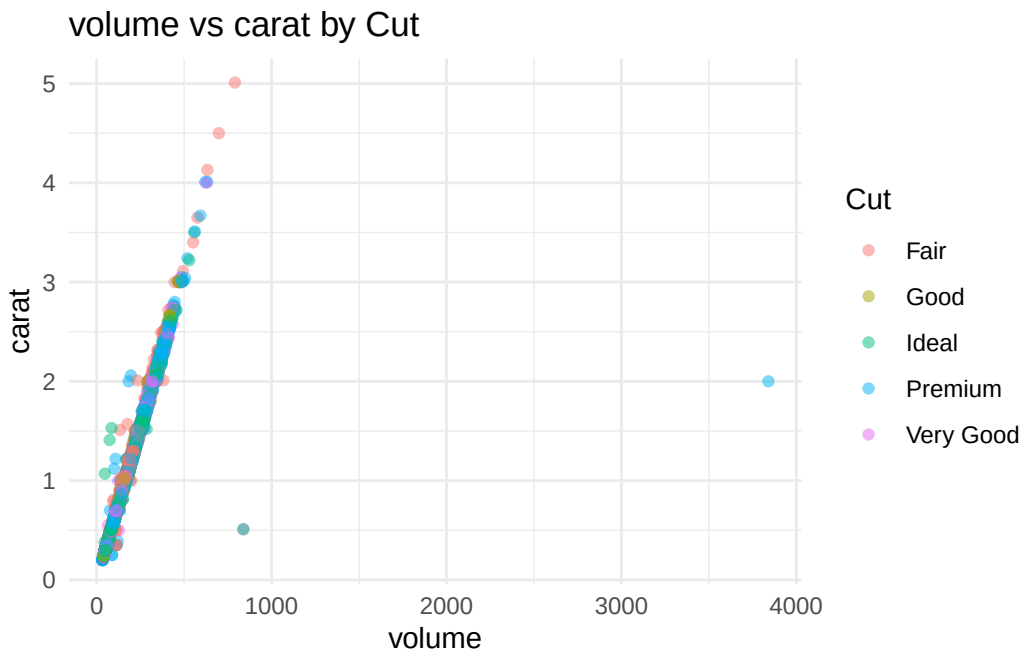
第一典型變數主要受 carat(-),volume(-) 和 price(-) 影響

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

Price model

```
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",trace = 0)
summary(step_model)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-23909.0   -629.0   -125.2    491.8   9616.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9334.311    384.720  24.263  < 2e-16 ***
carat       10938.619     55.147 198.355  < 2e-16 ***
cut            81.592      5.229  15.603  < 2e-16 ***
color        -322.614      3.261 -98.929  < 2e-16 ***
clarity       505.707      3.517 143.791  < 2e-16 ***
depth        -102.704      4.696 -21.873  < 2e-16 ***
table         -62.548      2.542 -24.609  < 2e-16 ***
x            -898.746     37.579 -23.916  < 2e-16 ***
y              42.151     20.933   2.014  0.04405 *
z            -130.809     40.821  -3.204  0.00135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
Residual standard error: 1217 on 53913 degrees of freedom
Multiple R-squared:  0.9068,    Adjusted R-squared:  0.9068
F-statistic: 5.831e+04 on 9 and 53913 DF,  p-value: < 2.2e-16
```

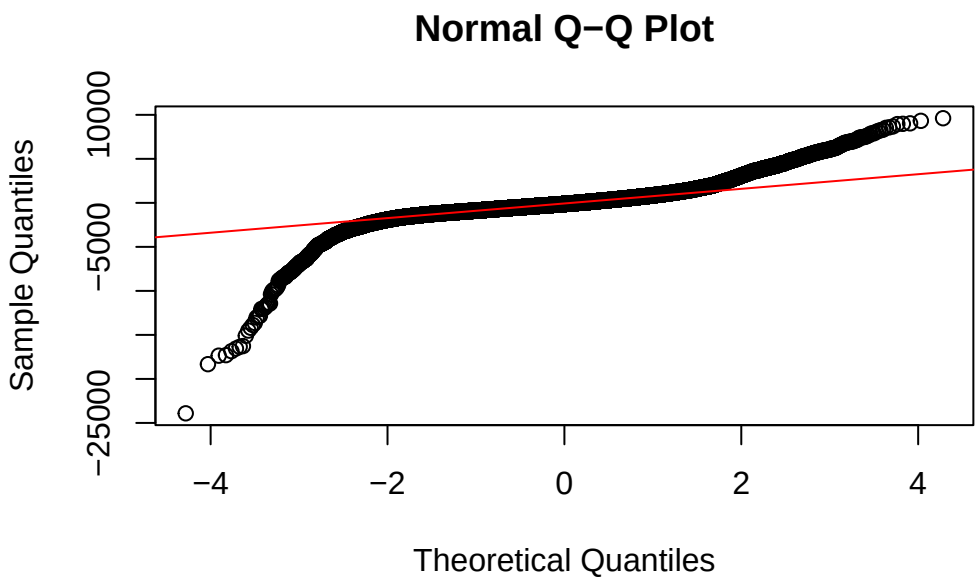
```r
vif(step_model)
```

```
    carat       cut     color   clarity     depth     table         x         y
24.851054  1.051424  1.120489  1.221400  1.646674  1.173853 64.418164 20.733890
        z
29.937704
```

```r
qqnorm(resid(step_model))
qqline(resid(step_model), 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
-19671.5   -695.0   -170.1    557.7   9326.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3485.030    325.056   10.72   <2e-16 ***
carat       8790.663     12.788  687.41   <2e-16 ***
cut           86.974      5.302   16.40   <2e-16 ***
```




```
color       -317.395      3.307  -95.99   <2e-16 ***
clarity      527.590      3.525  149.67   <2e-16 ***
depth        -69.741      3.975  -17.54   <2e-16 ***
table        -62.059      2.577  -24.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1235 on 53916 degrees of freedom
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.904
F-statistic: 8.468e+04 on 6 and 53916 DF,  p-value: < 2.2e-16
```

```
vif(model2)
```

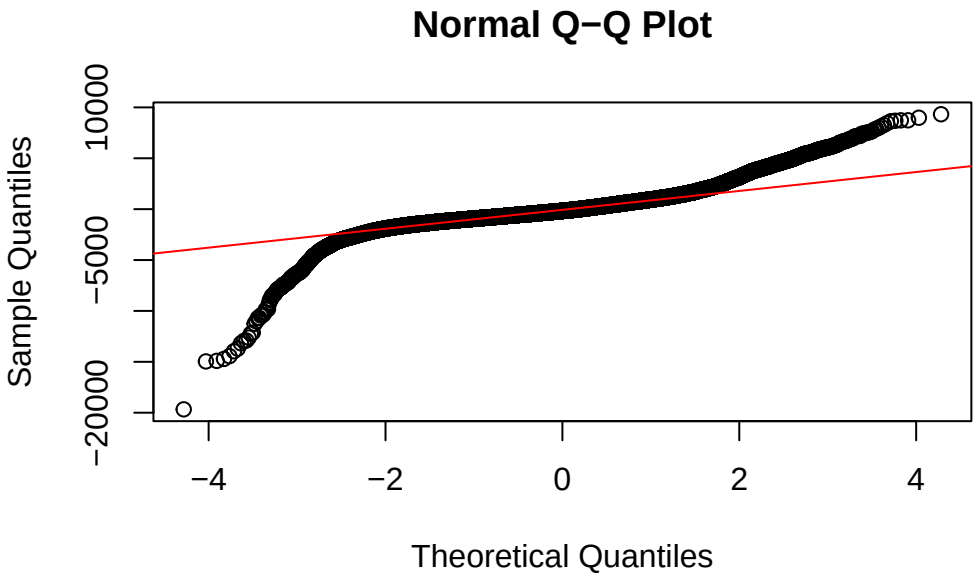
```
   carat      cut    color  clarity    depth    table
1.297621 1.049716 1.118656 1.191416 1.145987 1.171587
```

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

```
[1] -1.515882e-12
```

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

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


Normal Q−Q Plot
Sample Quantiles
10000
−5000
−20000
−4
−2
0
2
4
Theoretical Quantiles