Homework 2

106307030 財管四 廖偉博

Q1

[20pts] a. 生成一筆資料：

$X_i = a + \varepsilon, i = 1, \dots, 20$

- a 為0~10 任意數字。$\varepsilon \sim N(0, 2)$
- 注意：X必須在0~11內。

Ans:

詳見後方程式碼

```
> y
 [1] 10.2993431  0.2530285  7.8253714  1.0328434  8.0877121  8.6233766  2.7870334  5.7396099  2.0653691
[10]  2.2152110  7.4146825  9.1865351  7.1735514  0.9105971  0.7907242  9.0423911  9.7440828  9.8078056
[19]  9.4195485  2.3070893
```

[20pts] b. Cauchy(θ, 1) 的密度函數，取log後一次微分如下，請寫出此function

$$f(\theta) = -2 \sum_{i=1}^{n} \frac{\theta - x_i}{\{1 + (\theta - x_i)^2\}}$$

Ans:詳見後方程式碼

[10pts] c. 代入a生成的資料至b的function，並令$\theta = 0.3$。

Ans:

詳見後方程式碼

```
> #1.c
> theta = 0.3
> cauchy(theta, y)
[1] 8.656367
```

Q2

建立[哈佛大學]地區房屋價格的迴歸預測模型，找出是什麼因子影響不同房型的房價。

[10pts] a. 根據 Build_year，建立一個新類別變數 year_type，1899 年以前的房子為” centennial” ，1900~1959 年為” old” ，1960 年以上為” new” 。

Ans:

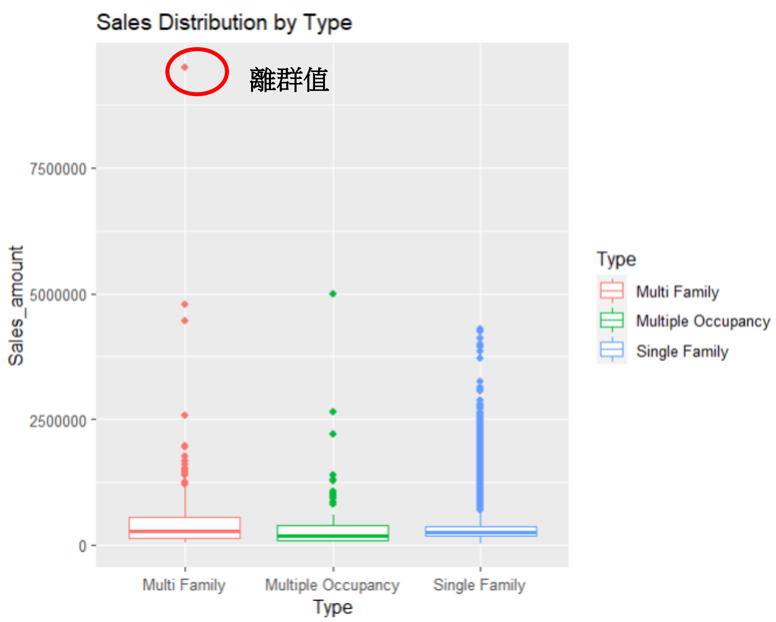
詳見後方程式碼

	Record	Sale_amount	Sale_date	Beds	Baths	Sqft_home	Sqft_lot	Type	Build_year	Town	University	year_type
386	386	385000	2016/5/6	3	3.00	1400	5488.0	Multi Family	1919	Ann Arbor, MI	University of Michigan	old
432	432	510000	2016/4/4	8	4.00	3554	12196.8	Multi Family	1910	Ann Arbor, MI	University of Michigan	old
683	683	72100	2016/5/26	2	1.50	1226	623.0	Multi Family	1980	Athens, GA	University of Georgia	new
689	689	66000	2016/5/3	5	3.00	2580	1198.0	Multi Family	1983	Athens, GA	University of Georgia	new
720	720	70000	2016/5/12	3	2.00	2188	24393.6	Multi Family	1983	Athens, GA	University of Georgia	new
808	808	110000	2016/5/12	2	2.50	1168	1962.0	Multi Family	2006	Athens, GA	University of Georgia	new
855	855	145600	2016/4/26	8	8.00	4256	37461.6	Multi Family	1983	Athens, GA	University of Georgia	new
859	859	725000	2016/4/1	4	3.00	1504	2264.0	Multi Family	1953	Berkeley, CA	University of California Berkeley	old
871	871	889000	2016/4/18	5	3.00	2325	3542.0	Multi Family	1903	Berkeley, CA	University of California Berkeley	old
879	879	860000	2016/2/4	5	3.00	2092	3998.0	Multi Family	1950	Berkeley, CA	University of California Berkeley	old
881	881	1250000	2016/7/8	4	3.00	2584	7500.0	Multi Family	1905	Berkeley, CA	University of California Berkeley	old
883	883	1125000	2016/5/27	3	4.00	2789	4880.0	Multi Family	1920	Berkeley, CA	University of California Berkeley	old
888	888	620000	2016/2/25	4	2.00	1789	4500.0	Multi Family	1944	Berkeley, CA	University of California Berkeley	old
895	895	1100000	2016/5/4	5	3.00	1800	5400.0	Multi Family	1921	Berkeley, CA	University of California Berkeley	old
896	896	1005000	2016/6/10	6	4.00	2277	7479.0	Multi Family	1898	Berkeley, CA	University of California Berkeley	centennial
898	898	1125000	2016/4/22	3	2.00	1643	4050.0	Multi Family	1906	Berkeley, CA	University of California Berkeley	old
907	907	1000000	2016/3/31	4	3.00	3100	5400.0	Multi Family	1900	Berkeley, CA	University of California Berkeley	old

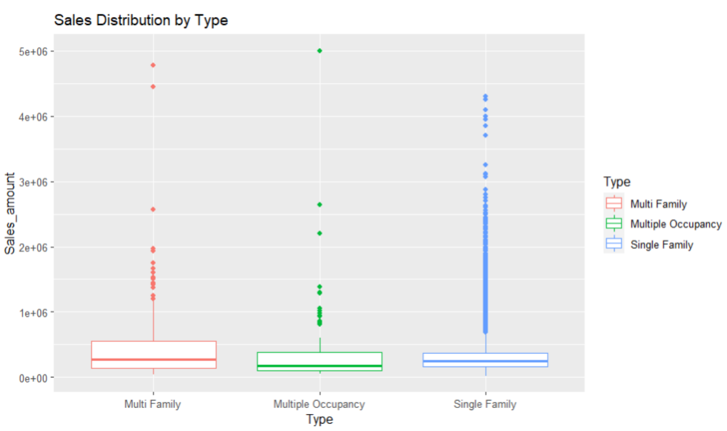
[40pts] b. 決定好你的最佳配適模型後，總結你的發現並根據解釋變數預測房屋價格。

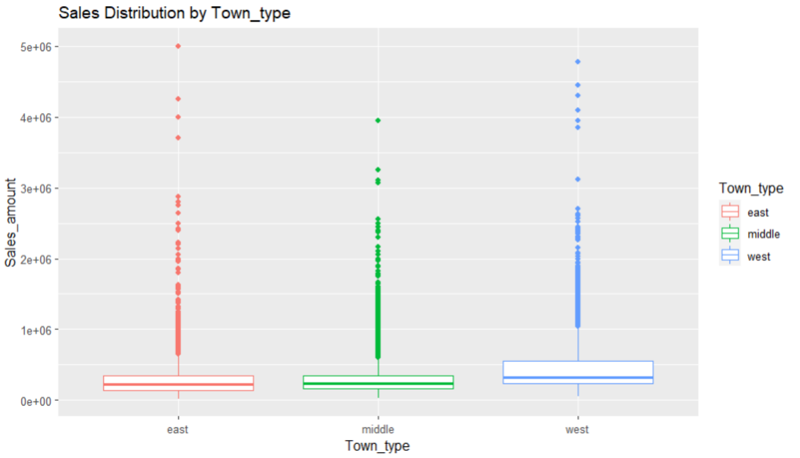
Ans:

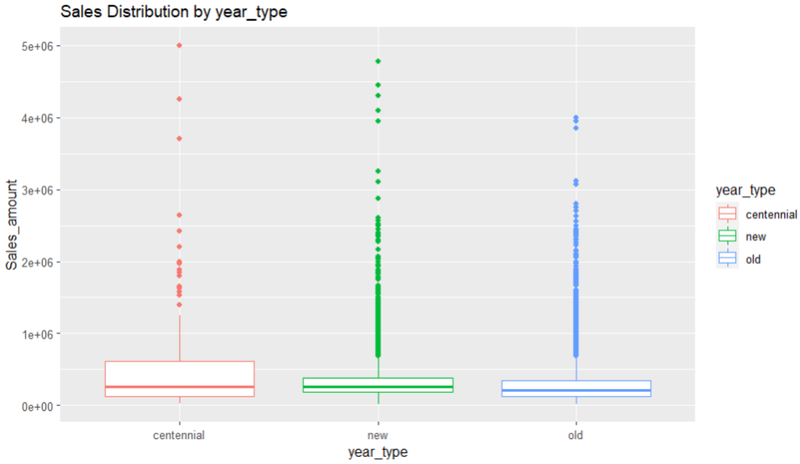
1.先確定你要預測的對象 y 以及可能放進去的多個 x 變數



發現有一個明顯的離群值，將其刪除







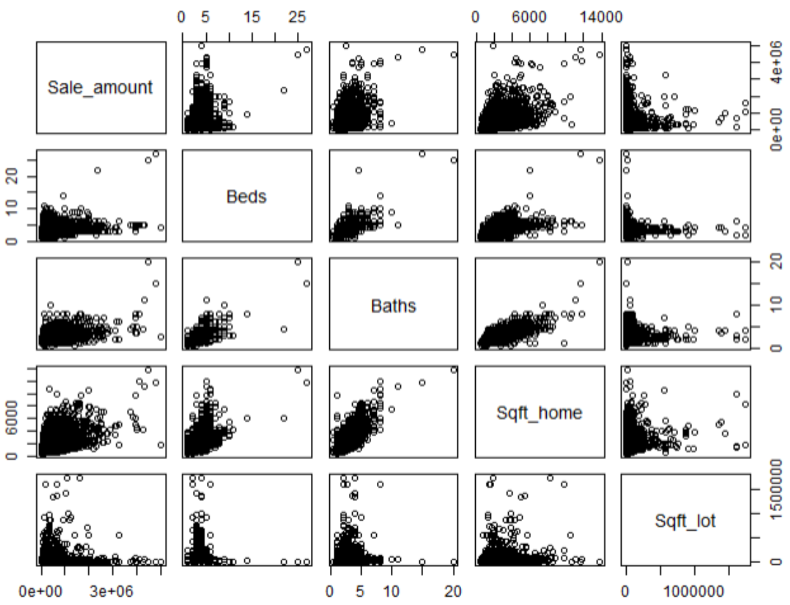
2.處理資料的遺漏值

因為此 dataset 沒有遺漏值，所以跳過此步驟。

3.可以先檢視連續型 x 變數之間的相關性，以及連續型 x 對 y 變數的相關性，看是否要在這步就篩選變數，或做變數變換

=>我先檢測連續型變數與 Sale_amount 間的關係，由下圖可看出 Sqft_lot 與 Sale_amount 的相關性相當低，因此先不把 Sqft_lot 放入模型。

散佈圖：



相關係數：

	Sale_amount	Beds	Baths	Sqft_home	Sqft_lot
Sale_amount	1.0000000	0.31039615	0.4541136	0.5174500	0.11630919
Beds	0.3103961	1.00000000	0.5872812	0.6039610	0.03065125
Baths	0.4541136	0.58728121	1.0000000	0.7592398	0.12205319
Sqft_home	0.5174500	0.60396101	0.7592398	1.0000000	0.18627041
Sqft_lot	0.1163092	0.03065125	0.1220532	0.1862704	1.00000000

4.類別型的 x 變數需要先轉成 dummy variable

=>我將房型、年份類型和所在地區轉換為 dummy variable

(1) 房型

	TypeMulti Family	TypeMultiple Occupancy
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0

(2)年份類型

	Typecentennial	Typenew
1	0	1
2	0	1
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	0	1
9	0	1
10	0	1
11	0	0
12	0	1
13	0	1
14	0	1
15	0	1
16	0	1
17	0	1
18	0	1
19	0	1
20	0	1

(3)所在地區

	Typeeast	Typemiddle
1	0	1
2	0	1
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	0	1
9	0	1
10	0	1
11	0	1
12	0	1
13	0	1
14	0	1
15	0	1
16	0	1
17	0	1
18	0	1
19	0	1
20	0	1

5.跑迴歸模型，檢視模型各項指標、是否有符合常態假設、做離群值的偵測，幫助我們篩選變數及樣本或其他處理

利用 AIC backward 法

```
Call:
lm(formula = Sale_amount ~ Beds + Baths + Sqft_home + Typeeast + 
    Typemiddle + Typecentennial + Typenew, data = train)

Coefficients:
   (Intercept)            Beds           Baths       Sqft_home        Typeeast  
      191851.7        -10310.7         88378.2           135.8       -238609.5  
    Typemiddle  Typecentennial         Typenew  
     -230376.6        221058.2       -170041.1  
```

利用 AIC forward 法

```
Call:
lm(formula = Sale_amount ~ Sqft_home + Typenew + Baths + Typemiddle + 
    Typeeast + Typecentennial + Beds, data = train)

Coefficients:
   (Intercept)       Sqft_home         Typenew           Baths      Typemiddle  
      191851.7           135.8       -170041.1         88378.2       -230376.6  
      Typeeast  Typecentennial            Beds  
     -238609.5        221058.2        -10310.7  
```

利用 BIC backward 法

```
Call:
lm(formula = Sale_amount ~ Beds + Baths + Sqft_home + Typeeast +
    Typemiddle + Typecentennial + Typenew, data = train)

Coefficients:
   (Intercept)            Beds           Baths       Sqft_home        Typeeast
      191851.7        -10310.7         88378.2           135.8       -238609.5
    Typemiddle  Typecentennial         Typenew
     -230376.6        221058.2       -170041.1
```

皆得出同樣的模型

因此設定模型一為：Sale_amount =191851.7 – 10310.7Beds + 88378.2Baths + 135.8Sqft_home – 238609.5Typeeast – 230376.6Typemiddle + 221058.2Typecentennial – 170041.1Typenew

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.919e+05  1.270e+04  15.109  < 2e-16 ***
Beds           -1.031e+04  3.878e+03  -2.659  0.00786 ** 
Baths           8.838e+04  5.174e+03  17.082  < 2e-16 ***
Sqft_home       1.358e+02  4.769e+00  28.463  < 2e-16 ***
Typeeast       -2.386e+05  8.701e+03 -27.422  < 2e-16 ***
Typemiddle     -2.304e+05  8.467e+03 -27.208  < 2e-16 ***
Typecentennial  2.211e+05  2.386e+04   9.266  < 2e-16 ***
Typenew        -1.700e+05  7.402e+03 -22.974  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 277800 on 7992 degrees of freedom
Multiple R-squared:  0.3853,	Adjusted R-squared:  0.3848 
F-statistic: 715.7 on 7 and 7992 DF,  p-value: < 2.2e-16
```

再觀察步驟三的散佈圖，發現 Beds 和 Sqft_home 有線性關係，因此再加入 Beds 和 Sqft_home 的交互作用項觀察

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.857e+05  1.664e+04  17.169  < 2e-16 ***
Beds           -3.836e+04  5.036e+03  -7.616 2.91e-14 ***
Baths           8.333e+04  5.183e+03  16.079  < 2e-16 ***
Sqft_home       1.105e+02  5.573e+00  19.820  < 2e-16 ***
Typeeast       -2.344e+05  8.675e+03 -27.017  < 2e-16 ***
Typemiddle     -2.271e+05  8.437e+03 -26.913  < 2e-16 ***
Typecentennial  2.275e+05  2.376e+04   9.577  < 2e-16 ***
Typenew        -1.613e+05  7.436e+03 -21.688  < 2e-16 ***
Beds:Sqft_home  7.509e+00  8.661e-01   8.670  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 276500 on 7991 degrees of freedom
Multiple R-squared:  0.3911,	Adjusted R-squared:  0.3905 
F-statistic: 641.5 on 8 and 7991 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: Sale_amount ~ Beds + Baths + Sqft_home + Typeeast + Typemiddle +
    Typecentennial + Typenew
Model 2: Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home + Typeeast +
    Typemiddle + Typecentennial + Typenew
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   7992 6.1661e+14
2   7991 6.1086e+14  1 5.7468e+12 75.177 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

發現加入 Beds 和 Sqft_home 的交互作用項厚的模型的 R-squared 確實有上升，且做 ANOVA 分析發現複雜的模型比簡單的模型較好，因此加入此交互作用項成為模型二。

模型二：Sale_amount = 285700 - 38360Beds + 83330Baths + 110.5Sqft_home + 7.509Beds:Sqft_home - 234400Typeeast – 227100 Typemiddle + 227500Typecentennial - 161300Typenew

再觀察步驟三的散佈圖，發現 Beds 和 Baths 也有線性關係，因此再加入 Beds 和 Baths 的交互作用項觀察

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.849e+05  1.729e+04  16.484  < 2e-16 ***
Beds           -3.814e+04  5.225e+03  -7.299 3.17e-13 ***
Baths           8.190e+04  1.044e+04   7.842 5.00e-15 ***
Sqft_home       1.124e+02  1.340e+01   8.387  < 2e-16 ***
Typeeast       -2.343e+05  8.677e+03 -27.006  < 2e-16 ***
Typemiddle     -2.270e+05  8.438e+03 -26.907  < 2e-16 ***
Typecentennial  2.275e+05  2.376e+04   9.575  < 2e-16 ***
Typenew        -1.613e+05  7.437e+03 -21.687  < 2e-16 ***
Beds:Sqft_home  7.012e+00  3.257e+00   2.153   0.0314 *
Beds:Baths      3.696e+02  2.331e+03   0.159   0.8741
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 276500 on 7990 degrees of freedom
Multiple R-squared:  0.3911,    Adjusted R-squared:  0.3904
F-statistic: 570.1 on 9 and 7990 DF,  p-value: < 2.2e-16
```

```
> anova(model_1, model_2)
Analysis of Variance Table

Model 1: Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home + Typeeast +
    Typemiddle + Typecentennial + Typenew
Model 2: Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home + Beds:Baths +
    Typeeast + Typemiddle + Typecentennial + Typenew
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   7991 6.1086e+14
2   7990 6.1086e+14  1 1921104754 0.0251 0.8741
```

發現加入 Beds 和 Sqft_home 的交互作用項厚的模型的 R-squared 沒有上升，且做 ANOVA 分析發現簡單的模型比複雜的模型較好，不加入此交互作用項。

再觀察步驟三的散佈圖，發現 Baths 和 Sqft_home 也有線性關係，因此再加入 Baths 和 Sqft_home 的交互作用項觀察

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.831e+05  1.662e+04  17.035  < 2e-16 ***
Beds             -1.335e+04  6.833e+03  -1.953   0.0508 .  
Baths             5.366e+04  7.544e+03   7.112 1.24e-12 ***
Sqft_home         1.010e+02  5.833e+00  17.310  < 2e-16 ***
Typeeast         -2.347e+05  8.660e+03 -27.104  < 2e-16 ***
Typemiddle       -2.263e+05  8.423e+03 -26.865  < 2e-16 ***
Typecentennial    2.295e+05  2.372e+04   9.677  < 2e-16 ***
Typenew          -1.563e+05  7.480e+03 -20.896  < 2e-16 ***
Beds:Sqft_home   -5.151e-01  1.718e+00  -0.300   0.7643    
Baths:Sqft_home   1.155e+01  2.138e+00   5.405 6.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 276000 on 7990 degrees of freedom
Multiple R-squared:  0.3933,	Adjusted R-squared:  0.3926 
F-statistic: 575.5 on 9 and 7990 DF,  p-value: < 2.2e-16
```

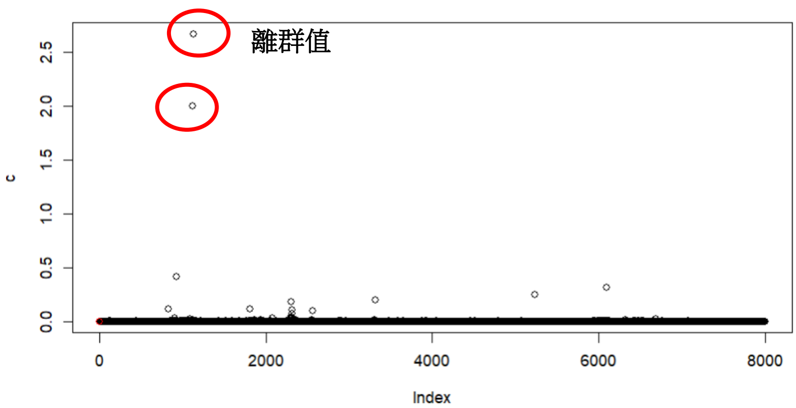
```
Analysis of Variance Table

Model 1: Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home + Typeeast + 
    Typemiddle + Typecentennial + Typenew
Model 2: Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home + Baths:Sqft_home + 
    Typeeast + Typemiddle + Typecentennial + Typenew
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1   7991 6.1086e+14                                    
2   7990 6.0863e+14  1 2.2251e+12 29.211 6.679e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

發現加入 Baths 和 Sqft_home 的交互作用項厚的模型的 R-squared 確實有上升，且做 ANOVA 分析發現複雜的模型比簡單的模型較好，因此加入此交互作用項成為模型三。

模型三：Sale_amount = 283100 - 13350Beds + 53660Baths + 101Sqft_home – 0.5151Beds:Sqft_home + 11.55Baths:Sqft_home -234700Typeeast – 226300 Typemiddle + 229500Typecentennial - 156300Typenew

檢查是否有離群值：



發現有兩筆離群值，將其刪除

最後，將模型一和模型三預測出的結果所算出的 RMSE 去做比較，發現模型三的 AIC、BIC、RMSE 確實都有下降。

```
> AIC(model,k = 2)
[1] 223265.4
> AIC(model_3,k = 2)
[1] 223165.3
```

```
> BIC(model)
[1] 223328.3
> BIC(model_3)
[1] 223242.2
```

```
> RMSE1
[1] 10498731
> RMSE2
[1] 10256357
```

6.利用測試集驗證預測的結果，如果不存在測試集，可以在步驟 4 後切出一部分的觀察值當作測試集

我將前 8000 筆資料設為訓練集，其餘的資料設定為測試集

TypeMulti Family	TypeMultiple Occupancy	Typeeast	Typemiddle	Typecentennial	Typenew	predict2
1	0	1	0	0	1	-11641.624
0	0	1	0	0	1	-6003.110
0	0	1	0	0	1	1543.321
0	0	1	0	0	1	4039.370
0	0	1	0	0	1	6870.214
0	0	1	0	0	1	6870.214
0	0	1	0	0	1	9533.660
0	0	1	0	0	1	12197.106
0	0	1	0	0	1	13861.760
0	0	1	0	0	1	16858.137
0	0	1	0	0	1	17565.981
0	0	1	0	0	1	18189.860
0	0	1	0	0	1	18411.814

得出預測結果。

程式碼：

```
#1.a
y = c()
while(length(y) < 20) {
  a = sample(0:10, 1)
  e = rnorm(1, 0, sqrt(2))
  if(a+e > 0 && a+e <11){
    y = append(y, a+e)
  }
}

y

#1.b
cauchy <- function(theta, x){
  n <- length(x)
  y <- matrix(0,n)
  for(i in c(0:n-1)){
    y[i] <- (theta - x[i]) / (1 + (theta - x[i])^2)
  }
  return(-2 * sum(y))
}

#1.c
theta = 0.3
cauchy(theta, y)

#2.a
install.packages("tidyverse")
library(tidyverse)
year_type <- ifelse (houseprice$Build_year<=1899, "centennial",
```

```
              ifelse(houseprice$Build_year>=1960, "new", "old"))
print(year_type)
houseprice$year_type <- c(year_type)

#2.b
library(broom)
attach(houseprice)

require(ggplot2)

#觀察房型的箱型圖
ggplot(data = houseprice) + geom_boxplot(aes( x= Type, y= Sale_amount, colour = Type)) +
  labs( x = 'Type',
        y = 'Sales_amount',
        title = 'Sales Distribution by Type')

#發現一個明顯的離群值，將其刪除
houseprice <- houseprice %>% filter(Sale_amount < 7500000)
attach(houseprice)

ggplot(data = houseprice) + geom_boxplot(aes( x= Type, y= Sale_amount, colour = Type)) +
  labs( x = 'Type',
        y = 'Sales_amount',
        title = 'Sales Distribution by Type')

#將城市依照西區、中部、東區做分類
Town_type <- ifelse (houseprice$Town %in% c("Tacoma, WA","Corvallis, OR","Eugene, OR","San Luis Obispo, CA",
                    "Claremont, CA","Berkeley, CA","Logan, UT","Bozeman, MT",
                    "Flagstaff, AZ","Tempe, AZ") , "west",
              ifelse (houseprice$Town %in% c("Boulder, CO","Fort Collins, CO","Fargo, ND",
```

```
                                    "Grand Forks, ND","Manhattan, KS","Lincoln, NE",
                                    "Lawrence, KS","College Station, TX","Minneapolis, MN",
                                    "Iowa City, IA","Ames, IA","Columbia, MO","Fayetteville, AR",
                                    "Madison, WI","Champaign-Urbana, IL","Bloomington,
IL"),"middle","east"))

print(Town_type)
houseprice$Town_type <- c(Town_type)

#觀察房子地區的箱型圖
ggplot(data = houseprice) + geom_boxplot(aes( x= Town_type, y= Sale_amount, colour =
Town_type)) +
 labs( x = 'Town_type',
      y = 'Sales_amount',
      title = 'Sales Distribution by Town_type')



#觀察房子年份的箱型圖
ggplot(data = houseprice) + geom_boxplot(aes( x= year_type, y= Sale_amount, colour =
year_type)) +
 labs( x = 'year_type',
      y = 'Sales_amount',
      title = 'Sales Distribution by year_type')

#將 Type 轉為 dummy variables，並試著做迴歸分析
library(dummies)

houseprice$Type = as.factor(as.character(houseprice$Type))
type_df <- data.frame(Type = houseprice$Type)
type_dummies <- dummy.data.frame(type_df)
type_dummies <- type_dummies[,-c(3)]
fit.1 = lm(Sale_amount~.,data = type_dummies)
```

```
summary(fit.1)

#將 Town 轉為 dummy variables，並試著做迴歸分析
houseprice$Town_type = as.factor(as.character(houseprice$Town_type))
town_df <- data.frame(Type = houseprice$Town_type)
town_dummies <- dummy.data.frame(town_df)
town_dummies <- town_dummies[,-c(3)]
fit.2 = lm(Sale_amount~.,data = town_dummies)
summary(fit.2)

##將 year_type 轉為 dummy variables，並試著做迴歸分析
houseprice$year_type = as.factor(as.character(houseprice$year_type))
year_df <- data.frame(Type = houseprice$year_type)
year_dummies <- dummy.data.frame(year_df)
year_dummies <- year_dummies[,-c(3)]
fit.3 = lm(Sale_amount~.,data = year_dummies)
summary(fit.3)

#將 dummy variables 合併到原本的 houseprice 表中
houseprice_final <- cbind(houseprice, type_dummies, town_dummies, year_dummies)
houseprice_final <- houseprice_final[-c(1, 3, 8:13)]
attach(houseprice_final)

#觀察連續型 x 變數之間的關係，以及連續型 x 對 y 變數的相關性

pairs(houseprice_final[,c(1:5)])

library(corrplot)
cor=cor(houseprice_final[,c(1:5)])
cor
```

```
#將 Sqft_lot 變數刪除
houseprice_final <- houseprice_final[-c(5)]
attach(houseprice_final)

#將前 8000 筆資料切為訓練集，其餘為測試集
train=houseprice_final[1:8000,]
test=houseprice_final[8001:10658,]

### stepwise
full <- lm(Sale_amount~.,data=train)
glance(full) %>% select(AIC,BIC)
null <-lm(Sale_amount~1,data=train)

#AIC
step(full, direction="backward")
step(null, scope=list(lower=null, upper=full), direction="forward")

#BIC
step(full, direction="backward", criterion = "BIC")

#做迴歸分析
model <- lm(Sale_amount ~ Beds + Baths + Sqft_home  +
        Typeeast + Typemiddle + Typecentennial + Typenew, data = train)
summary(model)

#加入交互作用項 Beds:Sqft_home
model_1 <- lm(Sale_amount ~ Beds + Baths + Sqft_home +Beds:Sqft_home +
        Typeeast + Typemiddle + Typecentennial + Typenew, data = train)

summary(model_1)
anova(model, model_1)
```

```
#加入交互作用項 Beds:Baths
model_2 <- lm(Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home +
          Beds:Baths + Typeeast + Typemiddle + Typecentennial + Typenew, data = train)
summary(model_2)
anova(model_1, model_2)

#加入交互作用項 Baths:Sqft_home
model_3 <- lm(Sale_amount ~ Beds + Baths + Sqft_home + Beds:Sqft_home +
          Baths:Sqft_home + Typeeast + Typemiddle + Typecentennial + Typenew, data =
train)
summary(model_3)
anova(model_1, model_3)

#發現有兩筆離群值，將其刪除
c= cooks.distance(model_3) #>=1, might be outlier, can remove
plot(c)
which(c>1)
houseprice_final <- houseprice_final[-c(1120,1128),]
attach(houseprice_final)

#做 AIC、BIC、RMSE 測試
train=houseprice_final[1:8000,]
test=houseprice_final[8001:10656,]
AIC(model,k = 2)
AIC(model_3,k = 2)
BIC(model)
BIC(model_3)
predict1 = predict(model,test)
predict2 = predict(model_3,test)
test1 <- cbind(test, predict1)
```

```
test2 <- cbind(test, predict2)

RMSE1=sqrt(mean(sum((test$Sale_amount-predict1)^2)))

RMSE2=sqrt(mean(sum((test$Sale_amount-predict2)^2)))

RMSE1

RMSE2
```