Лабораторна робота №2

A.

Виконати чистку dataset

Що було зроблено:

1. Перевірка на пропущені дані. (Їх не було)
2. Зміна типу даних деяких колонок на factor, бо там можливі тільки два значення.
3. Змінили назви деяких колонок на більш зручні.

```
#### (A) ####
# Перевіряємо чи є пропущені дані. Їх немає.
sum(is.na(data))

summary(data)

# Змінюємо тип у деяких даних на factor, оскільки вони мають тільки 2 можливих значення.

data$UNDER_CONSTRUCTION <- as.factor(data$UNDER_CONSTRUCTION)
summary(data$UNDER_CONSTRUCTION)

data$BHK_OR_RK <- as.factor(data$BHK_OR_RK)
summary(data$BHK_OR_RK)

data$READY_TO_MOVE <- as.factor(data$READY_TO_MOVE)
summary(data$READY_TO_MOVE)

data$RESALE <- as.factor(data$RESALE)
summary(data$RESALE)

# Змінюємо назви колонок для зручноті

colnames(data)[colnames(data)=="TARGET.PRICE_IN_LACS."] <-  "PRICE"

colnames(data)[colnames(data)=="BHK_NO."] <-  "BHK_NO"
```

Було побудовано модель, яка показувала залежність ціни на будинок в Індії з п’ятьма різними факторами. А також ми знайшли коефіцієнти за допомогою матричного методу. Результати зійшлись.

```
# Незалежні

x1 <- data$RERA
x2 <- data$BHK_NO
x3 <- data$SQUARE_FT
x4 <- data$LONGITUDE
x5 <- data$LATITUDE

# Будуємо модель з 5-ма незалежними змінними

modAll <- lm(Y ~ x1 + x2 + x3 + x4+ x5)

summary(modAll)
#  (Intercept)  1.274e+02
#  x1           1.018e+02
#  x2           8.501e+01
#  x3           1.389e-04
#  x4          -4.802e+00
#  x5          -1.571e+00

# Рахуємо коефіцієнти матричним методом
X <- cbind(1, x1,x2,x3,x4,x5)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y

beta
#        1.274163e+02
#    x1  1.018060e+02
#    x2  8.500818e+01
#    x3  1.388777e-04
#    x4 -4.801891e+00
#    x5 -1.571043e+00
#    Співпадає
```

B. Побудуйте однофакторні лінійні моделі

```
#### (B) ####

# Будуємо однофакторні лінійні моделі за 5-ма факторами

mod1 <- lm(Y~x1)
summary(mod1)

mod2 <- lm(Y~x2)
summary(mod2)

mod3 <- lm(Y~x3)
summary(mod3)

mod4 <- lm(Y~x4)
summary(mod4)

mod5 <- lm(Y~x5)
summary(mod5)
```

C. Визначте кращу модель за R^2

Оцінено отримані моделі. Судячи з наведених нижче результатів, виявилось, що найкращою моделлю є модель №3 через найбільше значення R^2

№1

```
> mod1 <- lm(Y~x1)
> summary(mod1)

Call:
lm(formula = Y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
 -207.0   -96.6   -67.6   -27.6 29887.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  112.567      4.624   24.34   <2e-16 ***
x1            95.408      8.201   11.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 655.4 on 29449 degrees of freedom
Multiple R-squared:  0.004575,  Adjusted R-squared:  0.004541
F-statistic: 135.3 on 1 and 29449 DF,  p-value: < 2.2e-16
```

№2

```
> mod2 <- lm(Y~x2)
> summary(mod2)

Call:
lm(formula = Y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1605.0   -98.9   -65.0   -10.0 29806.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -57.815     11.028  -5.243 1.59e-07 ***
x2            83.901      4.327  19.391  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 652.7 on 29449 degrees of freedom
Multiple R-squared:  0.01261,   Adjusted R-squared:  0.01257
F-statistic:   376 on 1 and 29449 DF,  p-value: < 2.2e-16
```

№3

```
> mod3 <- lm(Y~x3)
> summary(mod3)

Call:
lm(formula = Y ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-7552.8  -102.3   -78.4   -40.2 13640.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.401e+02  3.504e+00    40.0   <2e-16 ***
x3          1.391e-04  1.843e-06    75.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 601.3 on 29449 degrees of freedom
Multiple R-squared:  0.1622,    Adjusted R-squared:  0.1621
F-statistic:  5700 on 1 and 29449 DF,  p-value: < 2.2e-16
```

№4

```
> mod4 <- lm(Y~x4)
> summary(mod4)

Call:
lm(formula = Y ~ x4)

Residuals:
    Min      1Q  Median      3Q     Max
 -269.3  -105.3   -78.3   -33.4 29829.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 213.0504    13.6789  15.575  < 2e-16 ***
x4           -3.2935     0.6166  -5.342 9.28e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 656.6 on 29449 degrees of freedom
Multiple R-squared:  0.000968,  Adjusted R-squared:  0.000934
F-statistic: 28.53 on 1 and 29449 DF,  p-value: 9.279e-08
```

№5

```
> mod5 <- lm(Y~x5)
> summary(mod5)

Call:
lm(formula = Y ~ x5)

Residuals:
    Min      1Q  Median      3Q     Max
 -345.8  -104.0   -80.3   -37.2 29857.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 225.3832    28.1158   8.016 1.13e-15 ***
x5           -1.0735     0.3625  -2.961  0.00307 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 656.8 on 29449 degrees of freedom
Multiple R-squared:  0.0002977,  Adjusted R-squared:  0.0002637
F-statistic: 8.769 on 1 and 29449 DF,  p-value: 0.003066
```

D. Знайти значення оцінених значення Y тобто $\hat{Y}$ за допомогою математичної моделі, а саме лінійної регресії $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$; А також обчислити $e_i = \hat{y}_i - y_i$;

Функція predict виводить прогнозовані значення моделі.

```
#### (D) ####

Y_hat <- predict(modAll)
Y_hat
e <- Y_hat-Y
e
```

E. Опишіть ваші дії, припущення та висновки.

Підсумовуючи, ми:

- Редагували нашу таблицю, зробивши її більш зручною як для читання, так і для побудови моделей.
- Зробили моделі для того, щоб зробити висновки щодо залежності ціни на житло від п’яти числових змінних з нашої таблиці. Однофакторна модель у якої незалежною змінною була площа житла виявилась найкращою.
- Спрогнозували ціну на житло, використовуючи функцію predict.