Bài toán: Nghiên cứu về sự ảnh hưởng tỷ lệ tiêu thụ năng lượng của 1 căn hộ (dependent variable, Y) và các biến độc lập là diện tích căn hộ (X_1), số phòng ngủ (X_2), và khoảng cách từ trung tâm thành phố (X_3).

B1: Nhập dữ liệu + Kiểm tra dữ liệu

Dữ liệu được tham khảo trên trang Kaggle: link

Ta lưu dưới dạng csv rồi đọc nó trên CSV bằng lệnh:

```
data <- read.csv("energy.csv", header = T, sep = ",")
data[1:20,] # Hien thi 20 hang dau tien cua bo du lieu
```

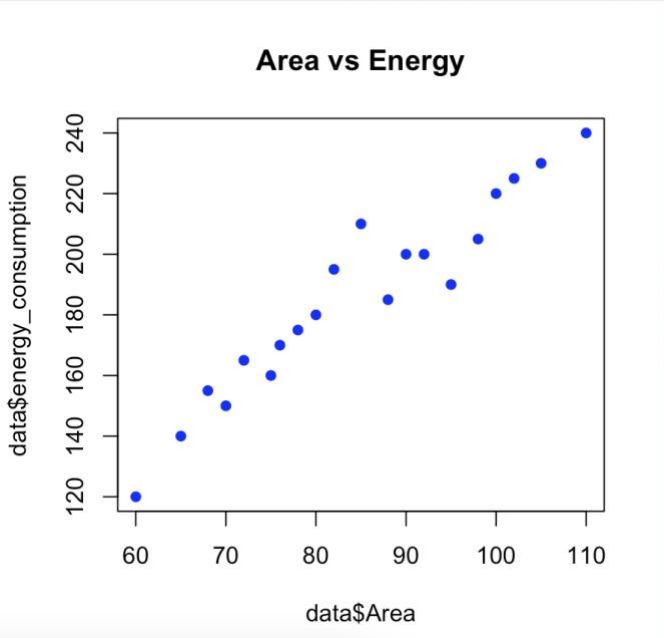
Kiểm tra dữ liệu:

Area	num_bedroom	distance	energy_consumption
70	3	5	150
90	4	3	200
80	2	7	180
60	3	10	120
100	4	2	220
75	2	8	160
95	3	4	190
85	4	6	210
65	2	9	140
110	3	1	240
78	4	5	175
88	2	7	185
72	3	6	165
98	4	3	205
68	2	8	155
105	3	2	230
82	4	4	195
92	2	6	200

B2: Kiểm tra sự tương quan của biến phụ thuộc đối với từng biến ngẫu nhiên

2.1 Sự tương quan giữa "Area" với "Energy_consumption"

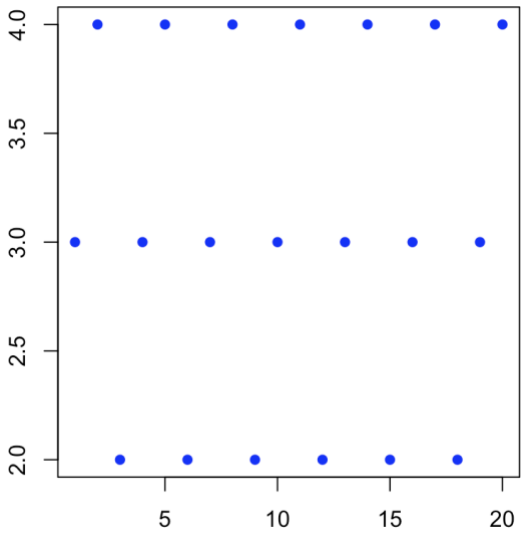
```
plot(data$Area, data$energy_consumption, pch = 16, col = 'blue', main = "Area
vs Energy")
```



Từ biểu đồ trên ta có thể thấy diện tích (Area) và khả năng tiêu thụ năng lượng (Energy consumption) có mối quan hệ tương quan tuyến tính thuận, tức là diện tích căn hộ càng lớn thì khả năng tiêu thụ điện càng cao. Điều này cũng khá đúng trong thực tế

2.2 Sự tương quan giữa "num_bedroom" với "Energy_consumption"

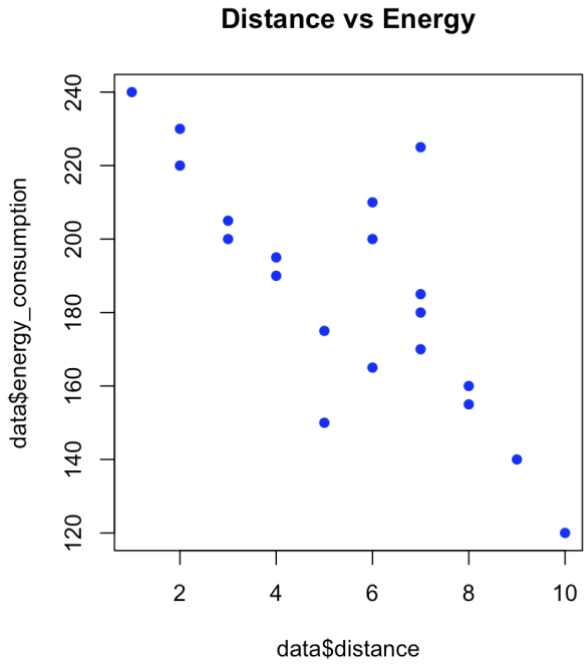
```
plot(data$num_bedroom, data$energy_consumption, pch = 16, col = 'blue', main
= "Number of bedroom vs Energy")
```



Như vậy số lượng phòng ngủ và tỷ lệ tiêu thụ năng lượng có sự tương quan khá rời rạc, trên thực tế ta cũng không thấy quá nhiều sự liên quan giữa 2 biến này.

2.3 Sự tương quan giữa "Distance" với "Energy_consumption"

```
plot(data$distance, data$energy_consumption, pch = 16, col = 'blue', main = "
Distance vs Energy")
```



Từ biểu đồ trên ta nhận thấy rằng khoảng cách đến trung tâm thành phố. (Distance) và tỷ lệ tiêu thụ năng lượng (Energy consumption) có mối quan hệ tương quan tuyến tính nghịch, tức là căn hộ nào càng gần trung tâm thành phố thì sẽ tiêu thụ càng nhiều năng lượng, điều này cũng khá đúng trong thực tế vì có thể do các hoạt động kinh doanh, hay chỉ đơn giản là nhu cầu về sử dụng năng lượng của họ cao hơn.

B3: Xây dựng mô hình

Ta Dùng hàm summary để tóm tắt cho đối tượng R, như một mô hình hồi quy tuyến (linear regression model), một ma trận, hoặc một vector.:

```
model.energy <- lm(energy_consumption ~ Area + num_bedroom + distance, data = data)

summary(model2.marketing)
```

Kết quả:

```
Call:
lm(formula = energy_consumption ~ Area + num_bedroom + distance,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.9389  -2.0431  -0.5874   3.9369  19.3816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.48960   30.16045   0.049    0.961
Area         2.03372    0.23949   8.492 2.54e-07 ***
num_bedroom  4.21297    3.04881   1.382    0.186
distance    -0.09815    1.50389  -0.065    0.949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 16 degrees of freedom
Multiple R-squared:  0.928,     Adjusted R-squared:  0.9145
F-statistic: 68.74 on 3 and 16 DF,  p-value: 2.332e-09
```

Như vậy, mô hình hồi quy bội về ảnh hưởng của chi phí quảng cáo theo các hình thứ khác nhau (Youtube, Facebook) lên doanh thu được cho bởi:

```
energy_consumption = 1.48960 + 2.03372×Area + 4.21297 ×num_bedroom - 0.09815 × distance
```

Giải thích:

1. **Residuals (Dư thừa):**
    - Min: -16.9389
    - 1Q (Quarter 1): -2.0431
    - Median: -0.5874
    - 3Q (Quarter 3): 3.9369
    - Max: 19.3816

Dữ liệu residuals cho biết sự chênh lệch giữa giá trị thực tế và giá trị được dự đoán của energy_consumption. Dữ liệu này có thể được sử dụng để kiểm tra sự phân phối của dư thừa và xác định liệu mô hình có hiệu suất tốt hay không.

2. **Coefficients (Hệ số):**
    - Intercept (Hệ số chệch): 1.48960, với độ lệch chuẩn (Std. Error) là 30.16045 và t-value là 0.049. Tuy nhiên, p-value (Pr(>|t|)) là 0.961, tức là không có bằng chứng thống kê để bác bỏ giả thuyết H0: Intercept = 0.
    - Area: Hệ số là 2.03372, độ lệch chuẩn là 0.23949, t-value là 8.492, và p-value rất thấp (2.54e-07), cho thấy có bằng chứng thống kê mạnh mẽ rằng biến Area có ảnh hưởng đáng kể đến energy_consumption.
    - num_bedroom: Hệ số là 4.21297, nhưng p-value là 0.186, vượt quá ngưỡng 0.05. Điều này ngụ ý rằng không có bằng chứng thống kê đủ để kết luận rằng num_bedroom ảnh hưởng đến energy_consumption.
    - distance: Hệ số là -0.09815, với p-value là 0.949, tức là không có bằng chứng thống kê để kết luận rằng distance ảnh hưởng đến energy_consumption.

1. **B0 (hệ số chệch - Intercept):** Là giá trị của energy_consumption khi tất cả các biến độc lập (Area, num_bedroom, distance) đều bằng 0. Trong trường hợp này, B0 là 1.48960, tức là khi tất cả các biến đầu vào đều bằng 0, giá trị dự đoán của energy_consumption là 1.48960.
2. **B1 (hệ số của biến Area):** Cho biết mức độ tăng/giảm của energy_consumption khi biến Area tăng lên một đơn vị, giữ nguyên các giá trị của các biến khác. Trong trường hợp này, B1 là 2.03372, nghĩa là khi diện tích (Area) tăng lên một đơn vị, giá trị dự đoán của energy_consumption sẽ tăng thêm 2.03372.

3. **B2 (hệ số của biến num_bedroom):** Tương tự như B1, nhưng áp dụng cho biến num_bedroom. Trong trường hợp này, B2 là 4.21297, cho biết mức độ tăng/giảm của energy_consumption khi số phòng ngủ (num_bedroom) tăng lên một đơn vị.
4. **B3 (hệ số của biến distance):** Là hệ số của biến distance, cho biết ảnh hưởng của biến này đối với energy_consumption. Trong trường hợp này, B3 là -0.09815, tức là khi khoảng cách (distance) tăng lên một đơn vị, giá trị dự đoán của energy_consumption sẽ giảm đi 0.09815.

Đề tìm khoảng tin cậy cho các hệ số hồi quy, ta sử dụng hàm **confint()**:

```
confint(model.enrgy)
```

Kết quả:

```
                2.5 %    97.5 %
(Intercept) -62.447701 65.426892
Area          1.526021  2.541410
num_bedroom  -2.250219 10.676166
distance     -3.286264  3.089959
```

Khoảng tin cậy 95% cho các hệ số hồi quy cho bởi:

$$-62.447701 \leq \beta 0 \leq 65.426892,$$

$$1.526021 \leq \beta 1 \leq 2.541410,$$

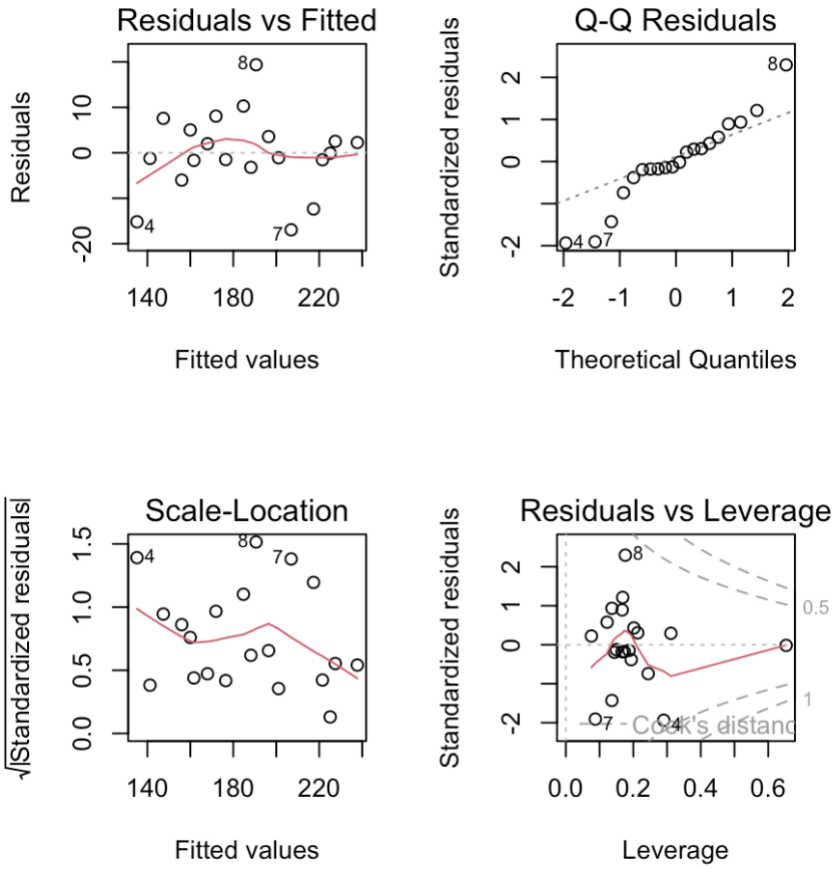
$$-2.250219 \leq \beta 2 \leq 10.676166,$$

$$-3.286264 \leq \beta 2 \leq 3.089959.$$

B4: Kiểm tra các giả định của mô hình

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình:

```
par(mfrow = c(2, 2))
plot(model.energy)
```



- Đồ thị thứ 1 (*Residuals vs Fitted*) vẽ các giá trị dự báo với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu (giả định 1) và tính đồng nhất của các phương sai sai số (giả định 3). Nếu như giả định về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định thứ 3 (phương sai đồng nhất) thì các điểm thặng dư phải phân tán đều nhau xung quanh đường thẳng $y=0$.
- Đồ thị thứ 2 (*Normal Q-Q*) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa.
- Đồ thị thứ 3 (*Scale - Location*) vẽ căn bậc hai của các giá trị thặng dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định thứ 3 (phương sai của các sai số là hằng số). Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả định thứ 3 được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả định thứ 3 bị vi phạm.
- Đồ thị thứ 4 (*Residuals vs Leverage*) cho phép xác định những điểm có ảnh hưởng cao (*influential observations*), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét (Cook's distance), và có một số điểm vượt qua đường thẳng khoảng cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao. Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao.

- Đồ thị thứ 1 (*Residuals vs Fitted*) vẽ các giá trị dự báo với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu (giả định 1) và tính đồng nhất của các phương sai sai số (giả định 3). Nếu như giả định về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định thứ 3 (phương sai đồng nhất) thì các điểm thặng dự phải phân tán đều nhau xung quanh đường thẳng $y=0$.
- Đồ thị thứ 2 (*Normal Q-Q*) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa.
- Đồ thị thứ 3 (*Scale - Location*) vẽ căn bậc hai của các giá trị thặng dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định thứ 3 (phương sai của các sai số là hằng số). Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả định thứ 3 được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả định thứ 3 bị vi phạm.
- Đồ thị thứ 4 (*Residuals vs Leverage*) cho phép xác định những điểm có ảnh hưởng cao (*influential observations*), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét (Cook's distance), và có một số điểm vượt qua đường thẳng khoảng cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao. Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao.

Nhận xét:

- Đồ thị *Normal Q-Q* cho thấy giả định sai số có phân phối chuẩn được thỏa mãn.
- Đồ thị thứ 1 (*Residuals vs Fitted*) cho thấy giả định về tính tuyến tính của dữ liệu hơi bị vi phạm, ta có thể thấy rằng sự vi phạm này bởi vì mối quan hệ giữa 3 biến ngẫu nhiên là phi tuyến tính.
- Đồ thị thứ 1 và thứ 3 (*Scale - Location*) cho ta thấy rằng giả định về tính đồng nhất của phương sai cũng hơi bị vi phạm. Tuy nhiên, ta cũng thấy này sự vi phạm này tương đối nhỏ và có thể chấp nhận được.
- Đồ thị thứ tư chỉ ra có các quan trắc thứ 6, 36 và 131 có thể là các điểm có ảnh hưởng cao trong bộ dữ liệu.