ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

TIỂU LUẬN CUỐI KÌ

MÔ HÌNH HÓA THỐNG KÊ

Giảng viên hướng dẫn: **TS. Nguyễn Thị Mộng Ngọc**

Nhóm thực hiện: **Nhóm 4**

Học viên:	**Phan Thị Thùy An**	MSHV: 20C29002
	Đinh Thị Nữ	MSHV: 20C29013
	Lý Phi Long	MSHV: 20C29028
	Đặng Khánh Thi	MSHV: 20C29038

TP. Hồ Chí Minh – Tháng 04, 2021

Mục lục

Chương 1

Dữ liệu tự chọn

- Tên "đề tài", nguồn gốc của dữ liệu, giới thiệu các biến.
- Mô hình chọn được; phân tích kết quả
- Đưa ra những phương pháp/phân tích khác có thể giúp cho kết quả tốt hơn.
- Kết luận.

1.1 Dữ liệu 1: Mô hình hồi quy đa biến

1.2 Dữ liệu 2: Hồi quy thành phần chính

Chương 2

Dữ liệu có sẵn

- Chọn mô hình phù hợp nhất giải thích biến phụ thuộc với từng bộ dữ liệu.
- Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó.
- Nói rõ ý nghĩa của mô hình đã chọn.

2.1 Dữ liệu 1

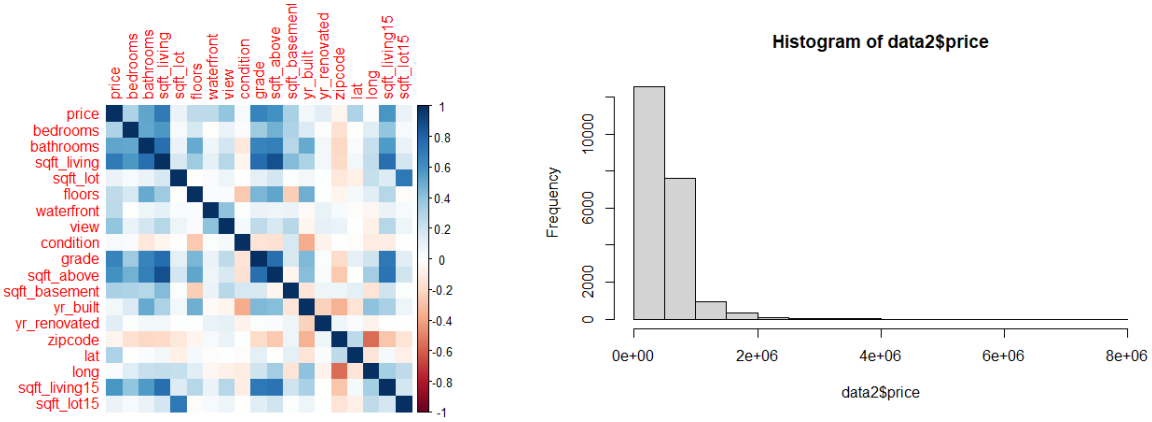
2.2 Dữ liệu 2

Bộ dữ liệu ghi lại lịch sử về những ngôi nhà được bán từ 5/2014 đến 5/2015 ở quận King, bang Washington, Hoa Kỳ. Bộ dữ liệu bao gồm 21613 quan trắc, gồm 21 biến.

Tìm hiểu dữ liệu

```
> mydata <-read.csv("data2.csv")
> head(mydata)
           id       date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
1 7129300520 10/13/2014  221900        3      1.00        1180     5650      1          0    0
2 6414100192  12/9/2014  538000        3      2.25        2570     7242      2          0    0
3 5631500400  2/25/2015  180000        2      1.00         770    10000      1          0    0
4 2487200875  12/9/2014  604000        4      3.00        1960     5000      1          0    0
5 1954400510  2/18/2015  510000        3      2.00        1680     8080      1          0    0
6 7237550310  5/12/2014 1230000        4      4.50        5420   101930      1          0    0
  condition grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat     long
1         3     7       1180             0     1955            0   98178 47.5112 -122.257
2         3     7       2170           400     1951         1991   98125 47.7210 -122.319
3         3     6        770             0     1933            0   98028 47.7379 -122.233
4         5     7       1050           910     1965            0   98136 47.5208 -122.393
5         3     8       1680             0     1987            0   98074 47.6168 -122.045
6         3    11       3890          1530     2001            0   98053 47.6561 -122.005
  sqft_living15 sqft_lot15
1          1340       5650
2          1690       7639
3          2720       8062
4          1360       5000
5          1800       7503
6          4760     101930
```

(a) Một số quan trắc đầu tiên



(b) Hệ số tương quan giữa các biến

(c) Phân bố của biến phụ thuộc

Hình 2.1: Một số quan sát ban đầu của bộ dữ liệu

Bộ dữ liệu cung cấp gồm 21 biến, trong đó biến **id** và **date** sẽ được loại bỏ khỏi dữ liệu trước khi tiến hành phân tích, vì nhóm em nghĩ các biến này chỉ để ghi lại chỉ số và thời gian mua bán, không có ý nghĩa thống kê.

```
> # Create full model
> mod_full_1 = lm(price ~ ., data2) #full model
> summary(mod_full_1)

Call:
lm(formula = price ~ ., data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1291631   -99089    -9569    77778  4330096

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.564e+06  2.933e+06   2.238  0.02523 *
bedrooms      -3.556e+04  1.901e+03 -18.707  < 2e-16 ***
bathrooms      4.128e+04  3.268e+03  12.632  < 2e-16 ***
sqft_living    1.496e+02  4.397e+00  34.033  < 2e-16 ***
sqft_lot       1.289e-01  4.792e-02   2.690  0.00714 **
floors         6.474e+03  3.602e+03   1.797  0.07229 .
waterfront     5.833e+05  1.736e+04  33.593  < 2e-16 ***
view           5.278e+04  2.141e+03  24.652  < 2e-16 ***
condition      2.679e+04  2.353e+03  11.387  < 2e-16 ***
grade          9.701e+04  2.161e+03  44.894  < 2e-16 ***
sqft_above     3.129e+01  4.361e+00   7.174 7.53e-13 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.628e+03  7.272e+01 -36.135  < 2e-16 ***
yr_renovated   1.983e+01  3.656e+00   5.425 5.87e-08 ***
zipcode       -5.819e+02  3.299e+01 -17.635  < 2e-16 ***
lat            6.022e+05  1.074e+04  56.071  < 2e-16 ***
long          -2.156e+05  1.316e+04 -16.385  < 2e-16 ***
sqft_living15  2.116e+01  3.451e+00   6.131 8.88e-10 ***
sqft_lot15    -3.907e-01  7.334e-02  -5.327 1.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201300 on 21579 degrees of freedom
Multiple R-squared:  0.7001,    Adjusted R-squared:  0.6999
F-statistic:  2964 on 17 and 21579 DF,  p-value: < 2.2e-16
```

Hình 2.2: Mô hình hồi quy đầy đủ ban đầu

Phân tích, chọn mô hình

∗ **Phương pháp chọn: Stepwise - lùi; tiêu chuẩn chọn: BIC**.

Bộ dữ liệu (sau khi loại bỏ id và date) có 18 biến giải thích, do đó nhóm em chọn phương pháp lùi (**stepwise - backward**) cho bộ dữ liệu này. Trong mô hình hồi quy đầy đủ (Hình 2.2), đa số các biến giải thích đều có ý nghĩa thống kê, do đó tiến hành phương pháp lùi (loại biến dần dần) sẽ tiết kiệm thời gian hơn so với các phương pháp còn lại.

Kết luận

```
> mod_BIC_1$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors +
    waterfront + view + condition + grade + sqft_above + sqft_basement +
    yr_built + yr_renovated + zipcode + lat + long + sqft_living15 +
    sqft_lot15

Final Model:
price ~ bedrooms + bathrooms + sqft_living + waterfront + view +
    condition + grade + sqft_above + yr_built + yr_renovated +
    zipcode + lat + long + sqft_living15 + sqft_lot15


             Step Df     Deviance Resid. Df   Resid. Dev      AIC
1                                     21579 8.739836e+14 527659.9
2 - sqft_basement  0            0     21579 8.739836e+14 527659.9
3        - floors  1 130842954998     21580 8.741144e+14 527653.1
4      - sqft_lot  1 284802633671     21581 8.743992e+14 527650.2
```

Hình 2.3: Mô hình khi chọn bằng tiêu chuẩn BIC

2.3 Dữ liệu 3

2.4 Dữ liệu 4