Data Science and R Final Project

🏠📈🤝

Real Estate Transactions in Seoul

Group 7

Choi Hae Min
Kim Ji Hyun
Son Sung Mo
Supatach Vanichayangkuranont

🏠 OUTLINE 🏠

1. Introduction
2. Visualisations & EDA
3. Modeling
4. Conclusions

1

Introduction

1 Introduction

Why do we need EDA about real estate?

- Stock : Lots of transactions → easy to know present price

- Real estate: Few transactions → hard to know present price

Our processes

- Check variables which influences real estate price

- Find out proper modeling about price

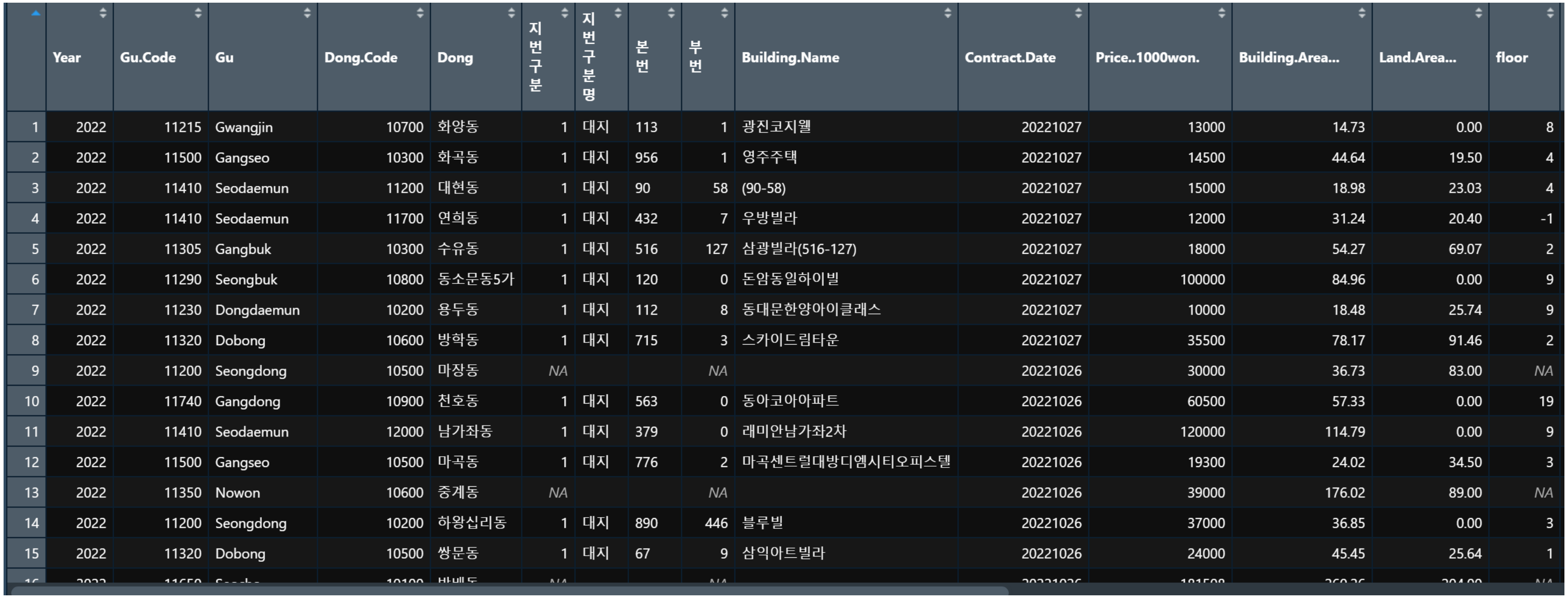
- Make test set, compare predicted and actual values about price

1 Introduction

Introduction about data

What?	Real estate transactions in Seoul (2018 JAN~ 2022 OCT)
Where?	Seoul Open Data
Size?	640,000 obs, 21 variables

How dataset looks like



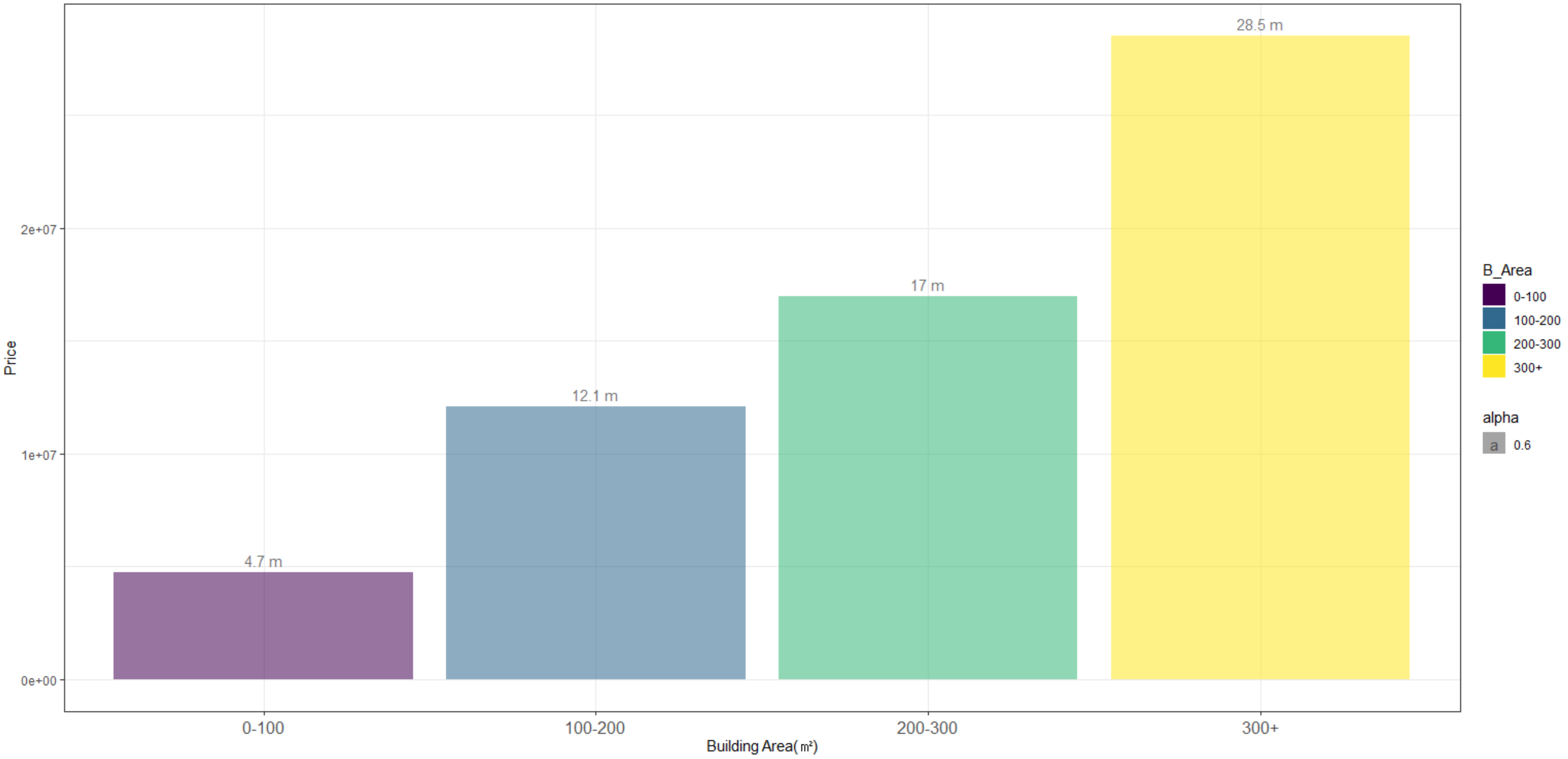
	Year	Gu.Code	Gu	Dong.Code	Dong	지번구분	지번구분명	본번	부번	Building.Name	Contract.Date	Price..1000won.	Building.Area...	Land.Area...	floor
1	2022	11215	Gwangjin	10700	화양동	1	대지	113	1	광진코지웰	20221027	13000	14.73	0.00	8
2	2022	11500	Gangseo	10300	화곡동	1	대지	956	1	영주주택	20221027	14500	44.64	19.50	4
3	2022	11410	Seodaemun	11200	대현동	1	대지	90	58	(90-58)	20221027	15000	18.98	23.03	4
4	2022	11410	Seodaemun	11700	연희동	1	대지	432	7	우방빌라	20221027	12000	31.24	20.40	-1
5	2022	11305	Gangbuk	10300	수유동	1	대지	516	127	삼광빌라(516-127)	20221027	18000	54.27	69.07	2
6	2022	11290	Seongbuk	10800	동소문동5가	1	대지	120	0	돈암동일하이빌	20221027	100000	84.96	0.00	9
7	2022	11230	Dongdaemun	10200	용두동	1	대지	112	8	동대문한양아이클래스	20221027	10000	18.48	25.74	9
8	2022	11320	Dobong	10600	방학동	1	대지	715	3	스카이드림타운	20221027	35500	78.17	91.46	2
9	2022	11200	Seongdong	10500	마장동	NA			NA		20221026	30000	36.73	83.00	NA
10	2022	11740	Gangdong	10900	천호동	1	대지	563	0	동아코아아파트	20221026	60500	57.33	0.00	19
11	2022	11410	Seodaemun	12000	남가좌동	1	대지	379	0	래미안남가좌2차	20221026	120000	114.79	0.00	9
12	2022	11500	Gangseo	10500	마곡동	1	대지	776	2	마곡센트럴대방디엠시티오피스텔	20221026	19300	24.02	34.50	3
13	2022	11350	Nowon	10600	중계동	NA			NA		20221026	39000	176.02	89.00	NA
14	2022	11200	Seongdong	10200	하왕십리동	1	대지	890	446	블루빌	20221026	37000	36.85	0.00	3
15	2022	11320	Dobong	10500	쌍문동	1	대지	67	9	삼익아트빌라	20221026	24000	45.45	25.64	1
16	2022	11650	Seocho	10100	방배동	NA			NA		20221026	181500	260.36	304.00	NA

2

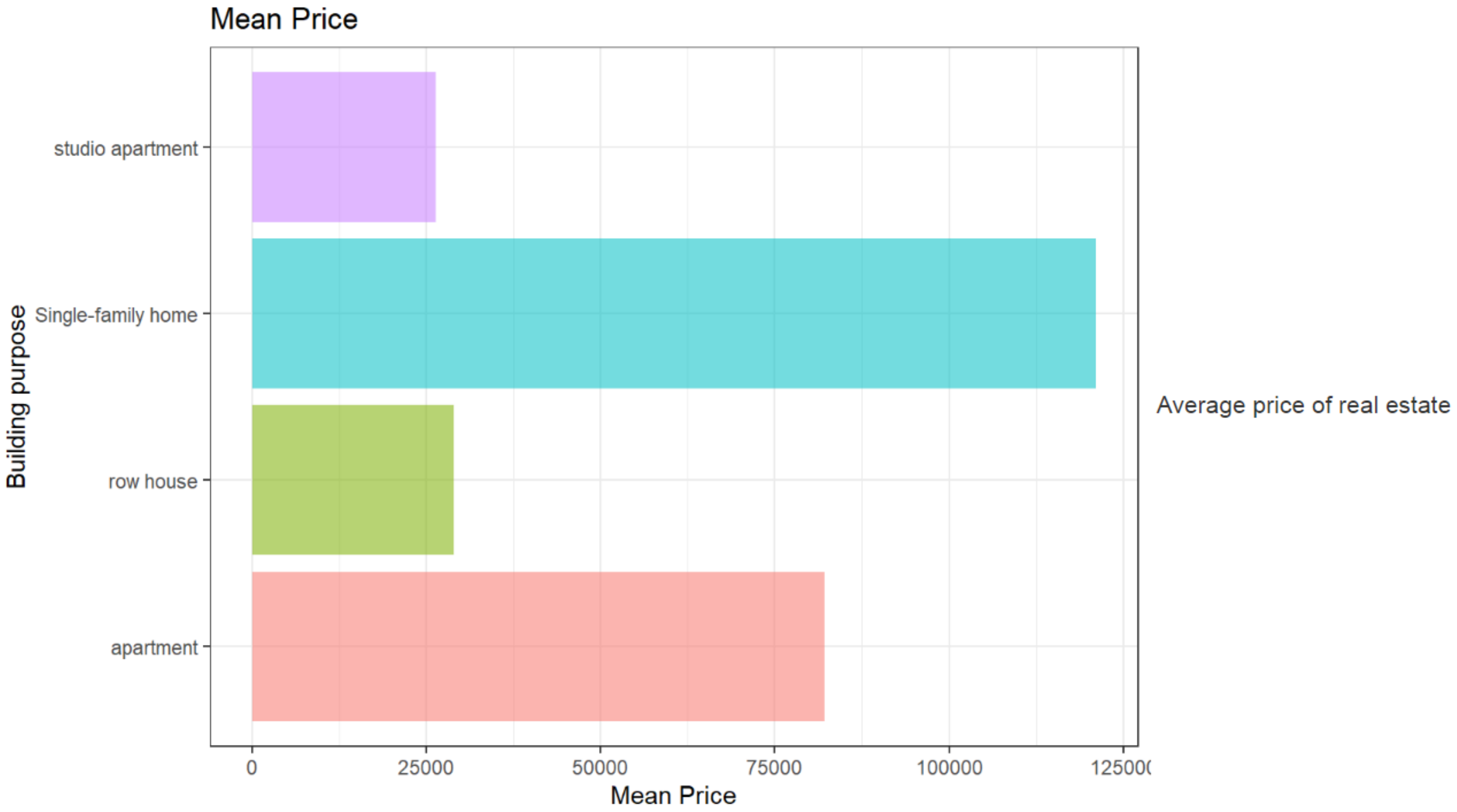
EDA & Visualisations

Average Price by Building Area

Average price of real estate transactions by building area



Average Price by Building Purpose



Price & area by building purpose

- Price: Studio apartment, row house < Apartment < Single-family house
- Area: Studio apartment, row house < Apartment < Single-family house

What if Building purpose → Area → Price ?

Multicollinearity

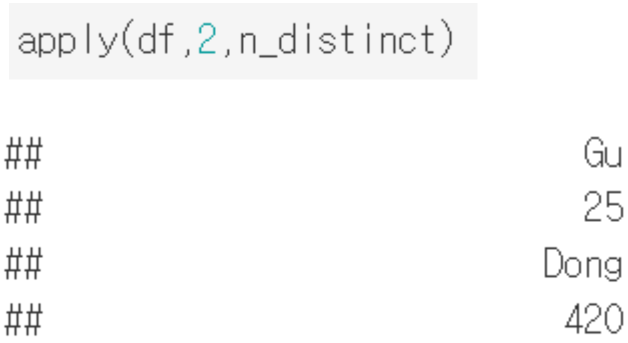
The occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

📍 Multicollinearity can lead to skewed or misleading results in prediction.

2 Visualisations & EDA

Problem of District

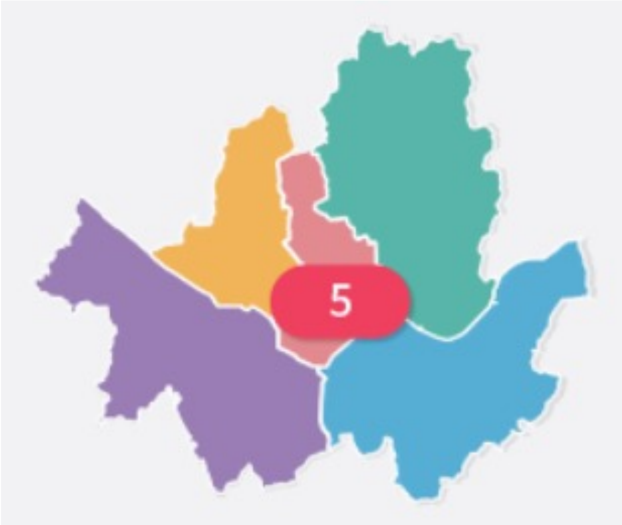
Too many factors in Gu



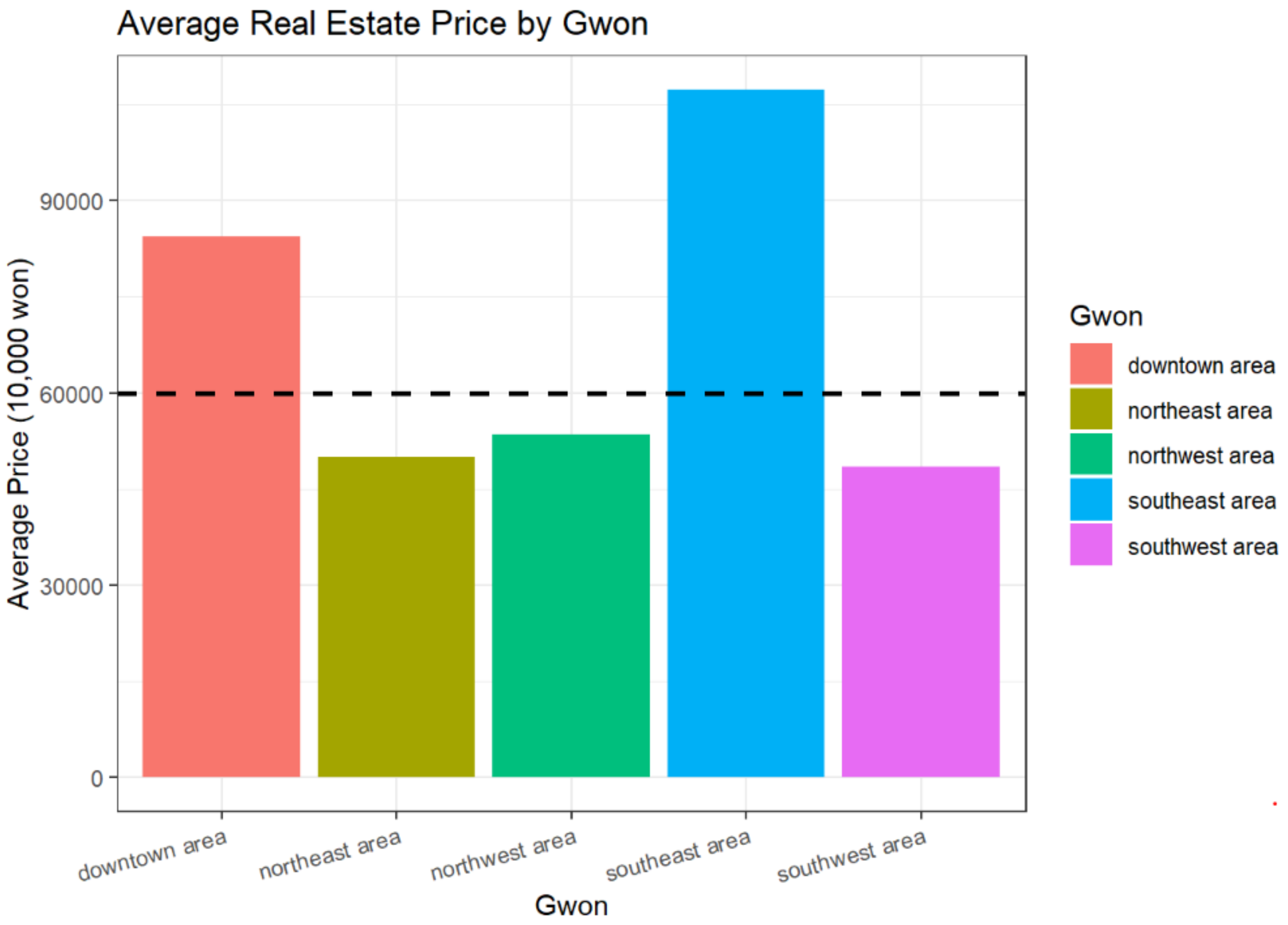
```
apply(df,2,n_distinct)
```

```
##                            Gu
##                            25
##                          Dong
##                           420
```

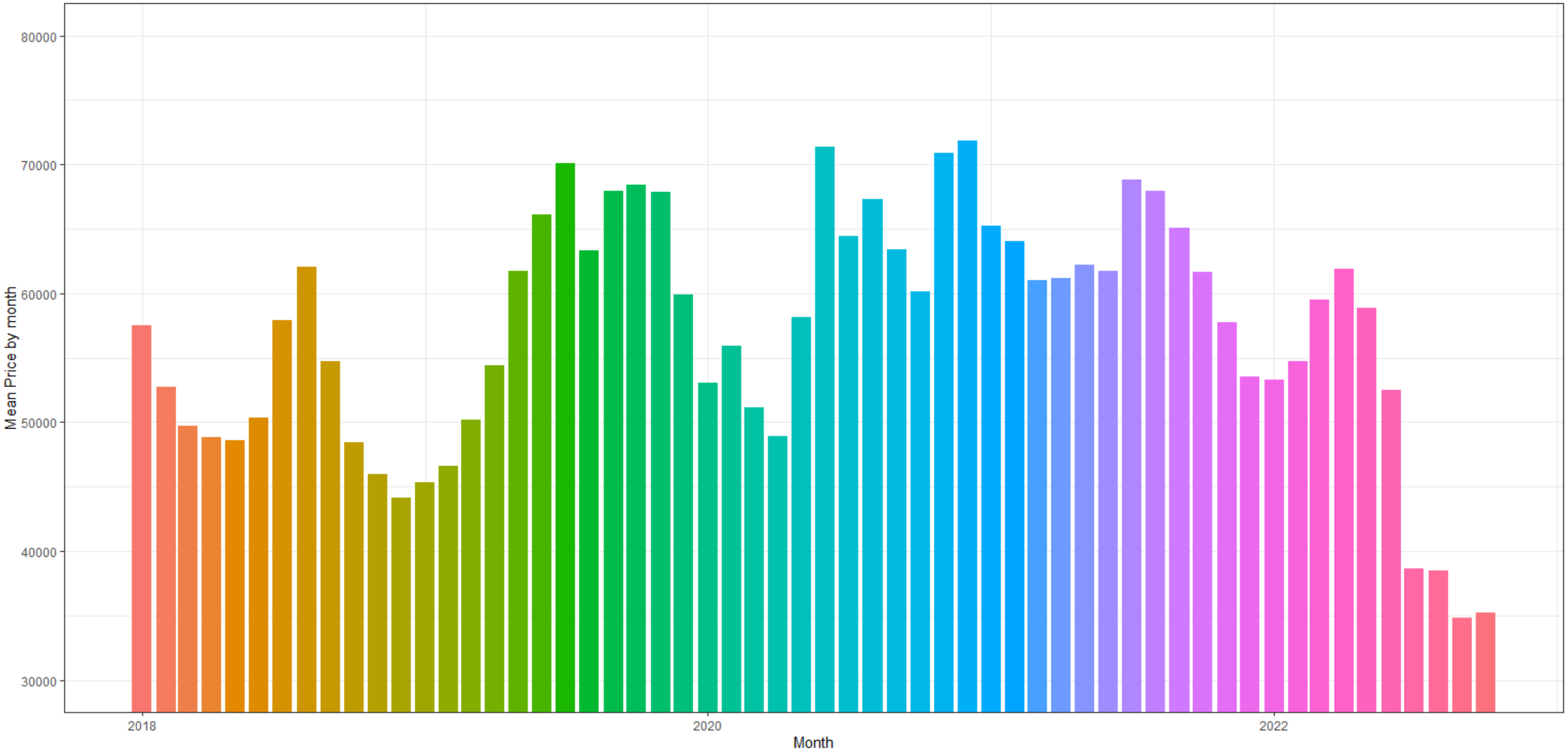
Integration Gu to Gwon



Average Price by Gwon

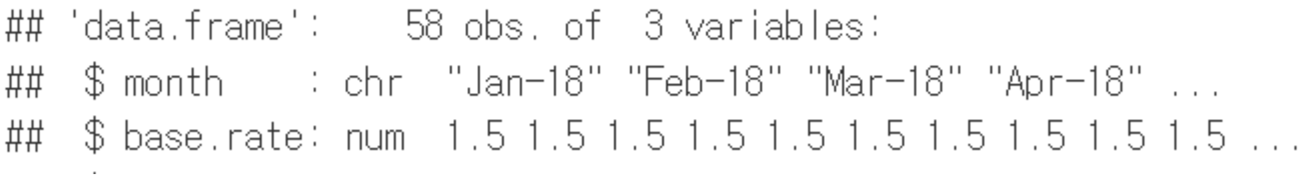


Average Price by time



Adding New Variable: Base Rate

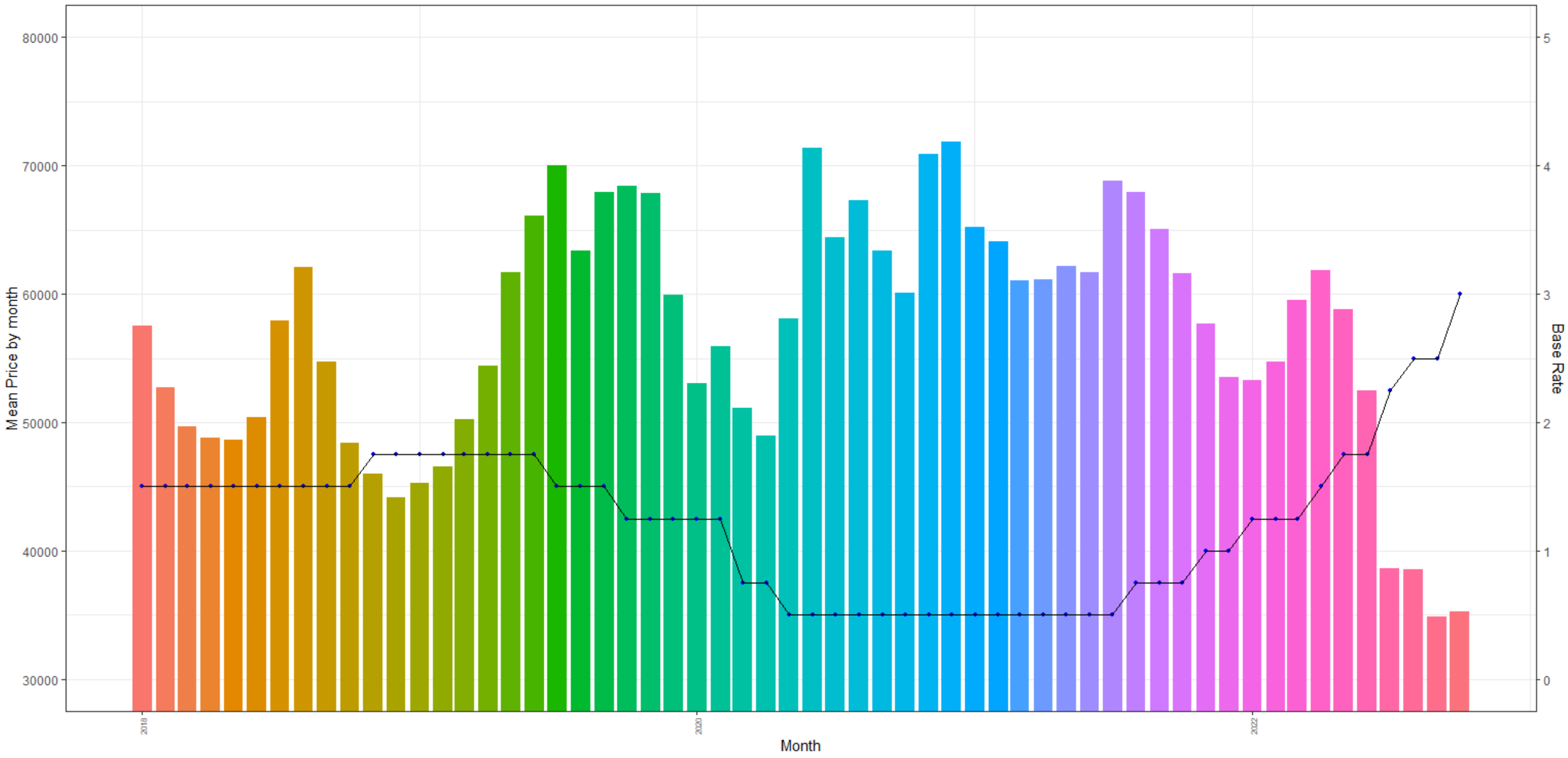
- The interest rate set by Policy
- Known to have (-) correlation with real estate price
- Bank of Korea has dataset about base rate by month



```
## 'data.frame':    58 obs. of  3 variables:
##  $ month    : chr  "Jan-18" "Feb-18" "Mar-18" "Apr-18" ...
##  $ base.rate: num  1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 ...
```

- Both dataset have ym data → inner join

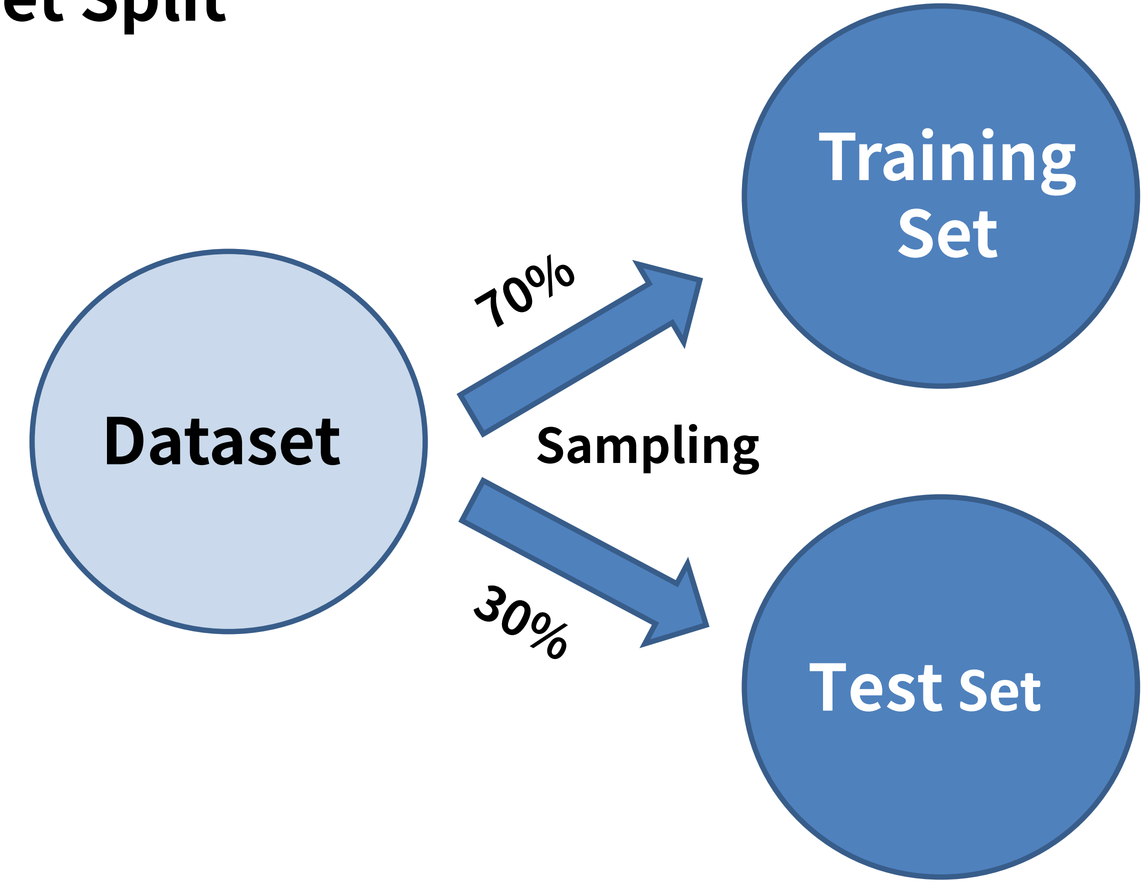
Average Price by Base Rate over time



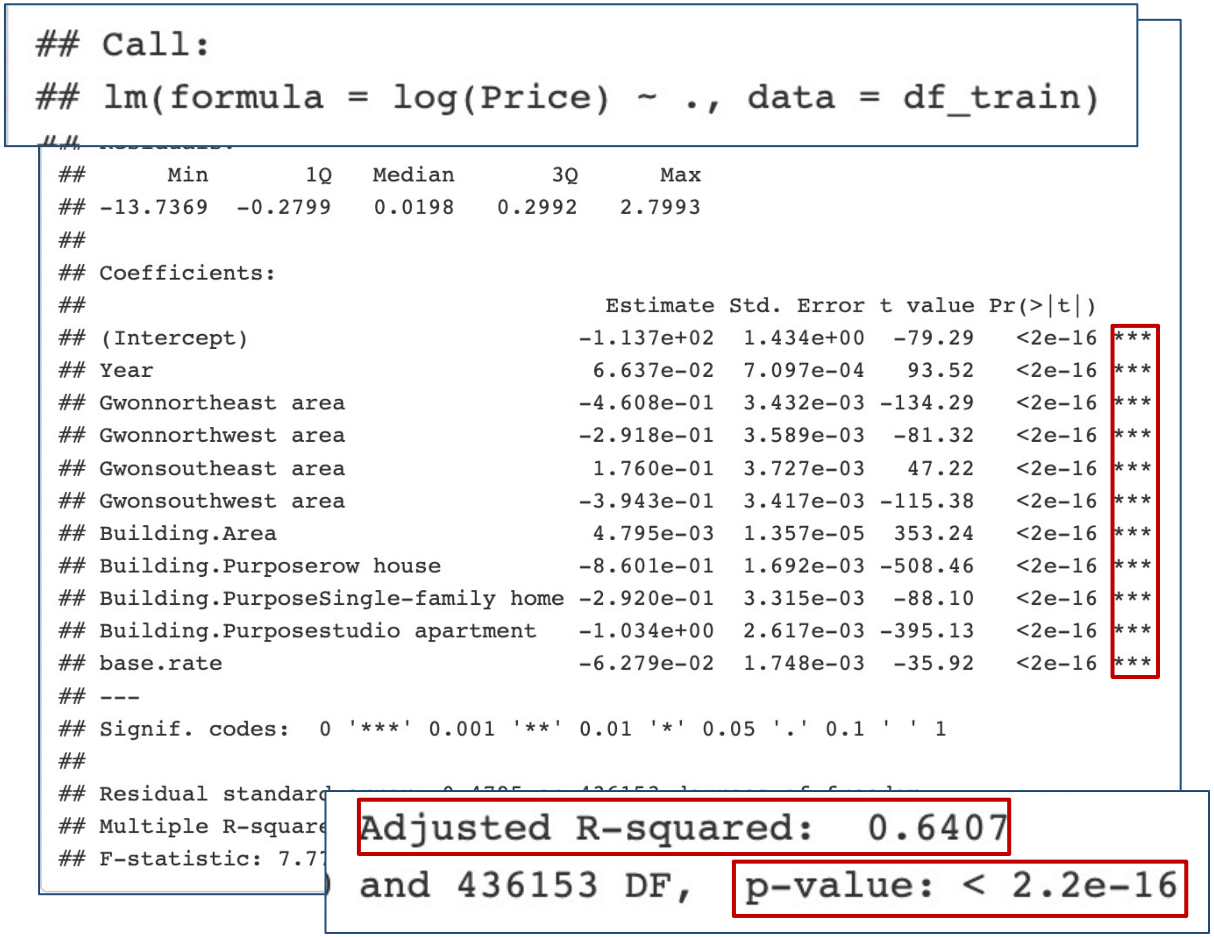
3

Modeling

Dataset Split



First Linear Regression Model (model1)



```
## Call:
## lm(formula = log(Price) ~ ., data = df_train)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -13.7369   -0.2799    0.0198    0.2992    2.7993
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -1.137e+02  1.434e+00  -79.29   <2e-16 ***
## Year                               6.637e-02  7.097e-04   93.52   <2e-16 ***
## Gwonnortheast area                -4.608e-01  3.432e-03 -134.29   <2e-16 ***
## Gwonnorthwest area                -2.918e-01  3.589e-03  -81.32   <2e-16 ***
## Gwonsoutheast area                 1.760e-01  3.727e-03   47.22   <2e-16 ***
## Gwonsouthwest area                -3.943e-01  3.417e-03 -115.38   <2e-16 ***
## Building.Area                      4.795e-03  1.357e-05  353.24   <2e-16 ***
## Building.Purposerow house         -8.601e-01  1.692e-03 -508.46   <2e-16 ***
## Building.PurposeSingle-family home -2.920e-01  3.315e-03  -88.10   <2e-16 ***
## Building.Purposestudio apartment  -1.034e+00  2.617e-03 -395.13   <2e-16 ***
## base.rate                         -6.279e-02  1.748e-03  -35.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard
## Multiple R-square
## F-statistic: 7.77
```

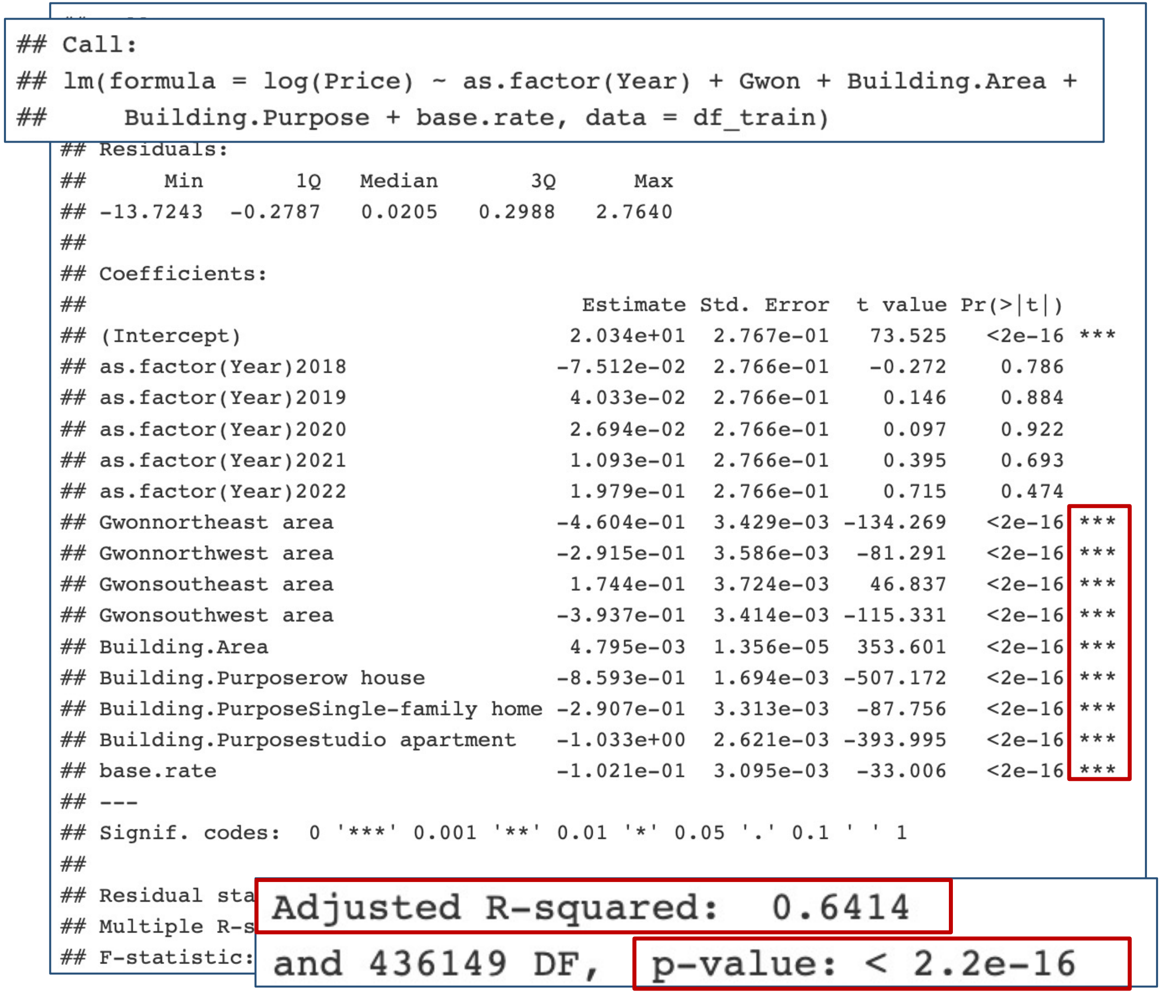
```
Adjusted R-squared:  0.6407
) and 436153 DF,  p-value: < 2.2e-16
```

Dependent Variable (y)
: log(Price)

Independent Variables (x1, x2, ⋯)
: Year, Gwon, Building Area, Building Purpose, Base Rate

💡 Year (2018~2022) variable is categorical, rather than numerical one.

Second Linear Regression Model (model2)



```
## Call:
## lm(formula = log(Price) ~ as.factor(Year) + Gwon + Building.Area +
##     Building.Purpose + base.rate, data = df_train)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7243  -0.2787   0.0205   0.2988   2.7640
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       2.034e+01  2.767e-01   73.525   <2e-16 ***
## as.factor(Year)2018              -7.512e-02  2.766e-01   -0.272    0.786
## as.factor(Year)2019               4.033e-02  2.766e-01    0.146    0.884
## as.factor(Year)2020               2.694e-02  2.766e-01    0.097    0.922
## as.factor(Year)2021               1.093e-01  2.766e-01    0.395    0.693
## as.factor(Year)2022               1.979e-01  2.766e-01    0.715    0.474
## Gwonnortheast area               -4.604e-01  3.429e-03 -134.269   <2e-16 ***
## Gwonnorthwest area               -2.915e-01  3.586e-03  -81.291   <2e-16 ***
## Gwonsoutheast area                1.744e-01  3.724e-03   46.837   <2e-16 ***
## Gwonsouthwest area               -3.937e-01  3.414e-03 -115.331   <2e-16 ***
## Building.Area                     4.795e-03  1.356e-05  353.601   <2e-16 ***
## Building.Purposerow house        -8.593e-01  1.694e-03 -507.172   <2e-16 ***
## Building.PurposeSingle-family home -2.907e-01  3.313e-03  -87.756   <2e-16 ***
## Building.Purposestudio apartment -1.033e+00  2.621e-03 -393.995   <2e-16 ***
## base.rate                        -1.021e-01  3.095e-03  -33.006   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual sta
## Multiple R-s
## F-statistic:
```

```
Adjusted R-squared:  0.6414
and 436149 DF,  p-value: < 2.2e-16
```

Dependent Variable (y)
: log(Price)

Independent Variables (x1, x2, ⋯)
: as.factor(Year), Gwon, Building Area, Building Purpose, Base Rate

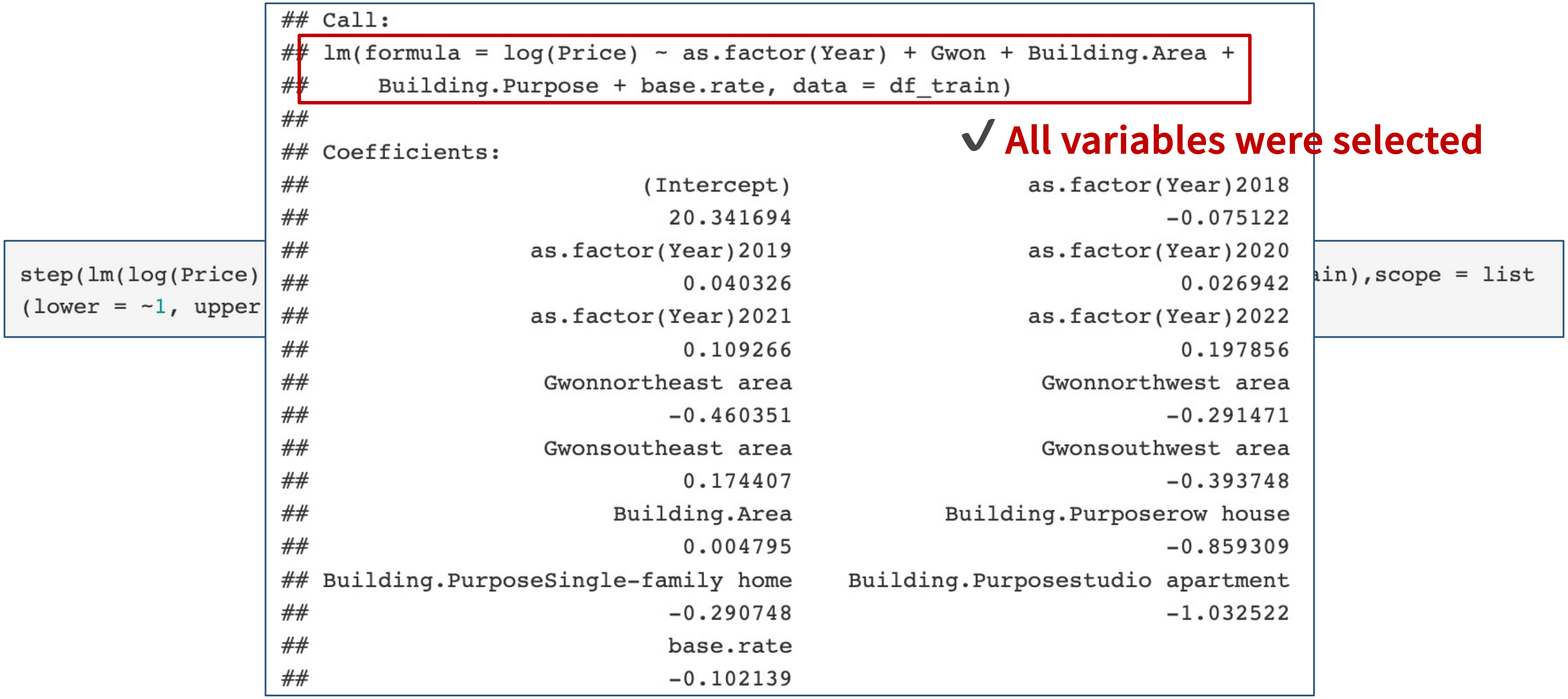
💡 Adjusted R-squared increased,
but the Year variable as a factor was not significant.

💡Variable Selection Method

3 Modeling

Variable Selection (Stepwise Model Selection)

: Taking regression with a number of predictors and then dropping or adding one at a time based on the criteria of model improvement until finding the "best" model



```
step(lm(log(Price)                                                ain),scope = list
(lower = ~1, upper
```

```
## Call:
## lm(formula = log(Price) ~ as.factor(Year) + Gwon + Building.Area +
##     Building.Purpose + base.rate, data = df_train)
##
## Coefficients:
##                       (Intercept)              as.factor(Year)2018
##                         20.341694                        -0.075122
##               as.factor(Year)2019              as.factor(Year)2020
##                          0.040326                         0.026942
##               as.factor(Year)2021              as.factor(Year)2022
##                          0.109266                         0.197856
##                Gwonnortheast area               Gwonnorthwest area
##                         -0.460351                        -0.291471
##                Gwonsoutheast area               Gwonsouthwest area
##                          0.174407                        -0.393748
##                     Building.Area            Building.Purposerow house
##                          0.004795                        -0.859309
## Building.PurposeSingle-family home    Building.Purposestudio apartment
##                         -0.290748                        -1.032522
##                         base.rate
##                         -0.102139
```

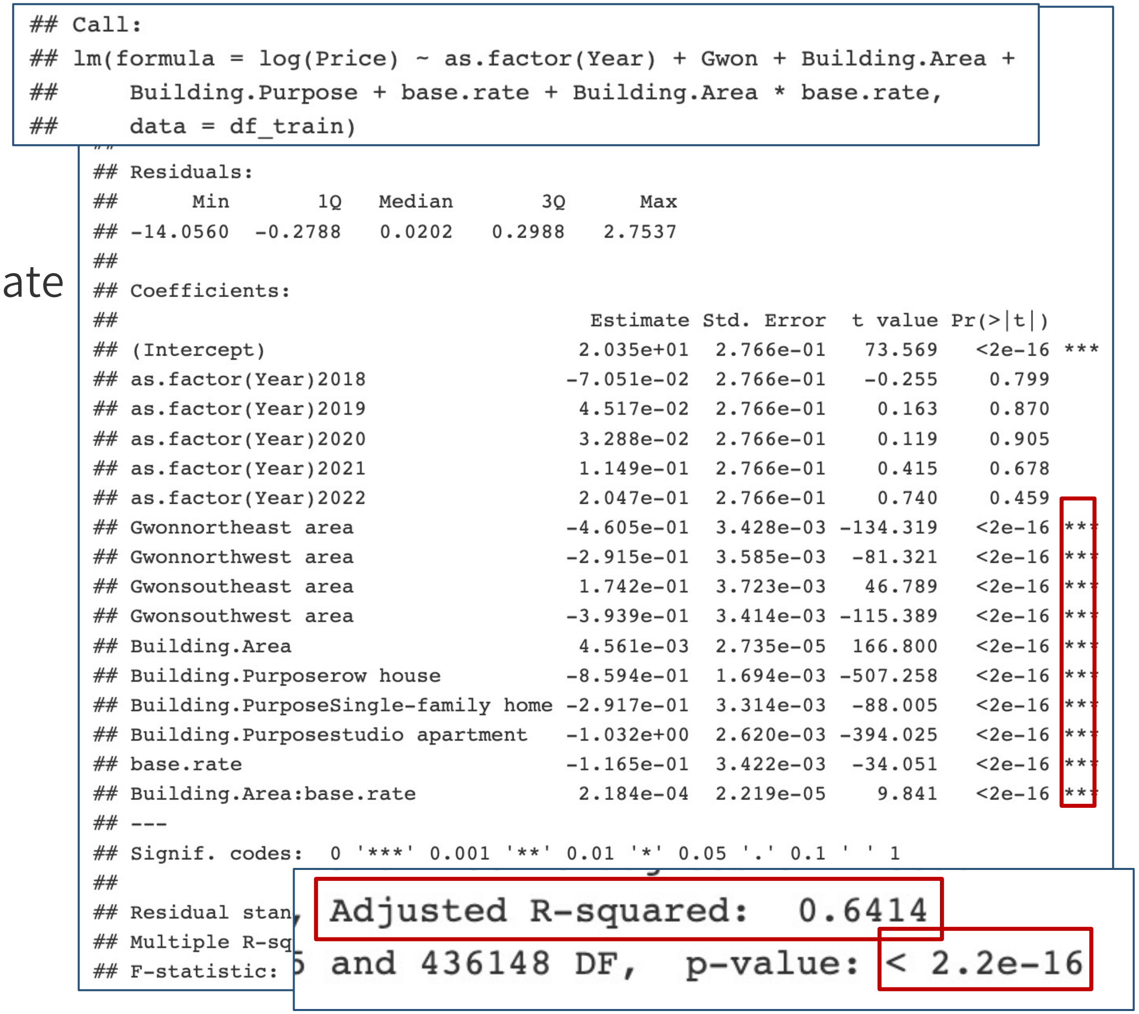
✓ All variables were selected

Third Linear Regression Model (model3)

💡 **Interaction Term**
: The two variables interact to have an effect that is more than the sum of their parts.
→ added Building Area*Base Rate as a new independent variable

Dependent Variable (y)
: log(Price)

Independent Variables (x1, x2, ⋯)
: as.factor(Year), Gwon, Building Area, Building Purpose, Base Rate, Building Area*Base Rate

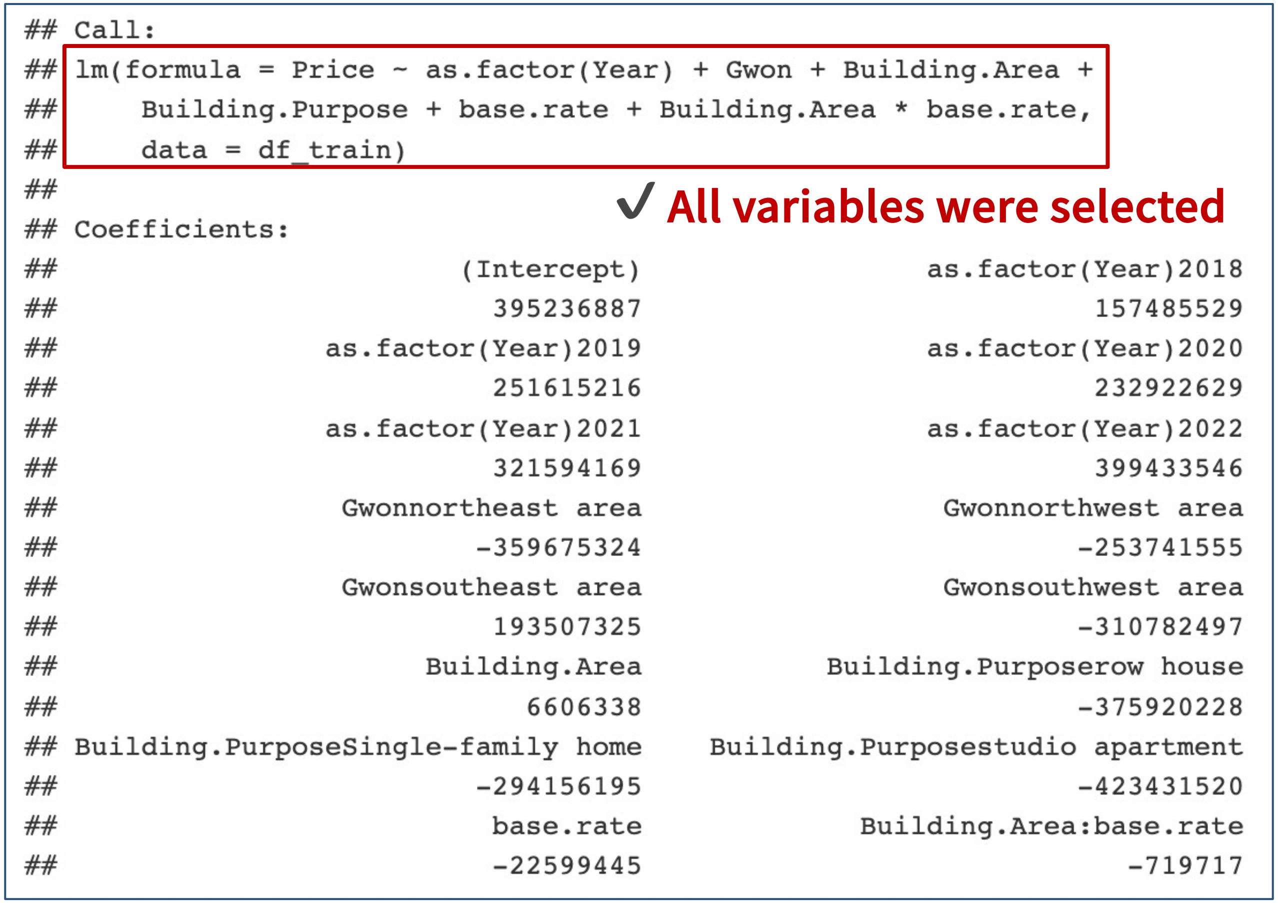


```
## Call:
## lm(formula = log(Price) ~ as.factor(Year) + Gwon + Building.Area +
##     Building.Purpose + base.rate + Building.Area * base.rate,
##     data = df_train)
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0560  -0.2788   0.0202   0.2988   2.7537
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       2.035e+01  2.766e-01   73.569   <2e-16 ***
## as.factor(Year)2018              -7.051e-02  2.766e-01   -0.255    0.799
## as.factor(Year)2019               4.517e-02  2.766e-01    0.163    0.870
## as.factor(Year)2020               3.288e-02  2.766e-01    0.119    0.905
## as.factor(Year)2021               1.149e-01  2.766e-01    0.415    0.678
## as.factor(Year)2022               2.047e-01  2.766e-01    0.740    0.459
## Gwonnortheast area               -4.605e-01  3.428e-03 -134.319   <2e-16 ***
## Gwonnorthwest area               -2.915e-01  3.585e-03  -81.321   <2e-16 ***
## Gwonsoutheast area                1.742e-01  3.723e-03   46.789   <2e-16 ***
## Gwonsouthwest area               -3.939e-01  3.414e-03 -115.389   <2e-16 ***
## Building.Area                     4.561e-03  2.735e-05  166.800   <2e-16 ***
## Building.Purposerow house        -8.594e-01  1.694e-03 -507.258   <2e-16 ***
## Building.PurposeSingle-family home -2.917e-01  3.314e-03  -88.005   <2e-16 ***
## Building.Purposestudio apartment -1.032e+00  2.620e-03 -394.025   <2e-16 ***
## base.rate                        -1.165e-01  3.422e-03  -34.051   <2e-16 ***
## Building.Area:base.rate           2.184e-04  2.219e-05    9.841   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual stan
## Multiple R-sq
## F-statistic:
```

```
Adjusted R-squared:  0.6414
5 and 436148 DF,  p-value: < 2.2e-16
```

Variable Selection (Stepwise Method)

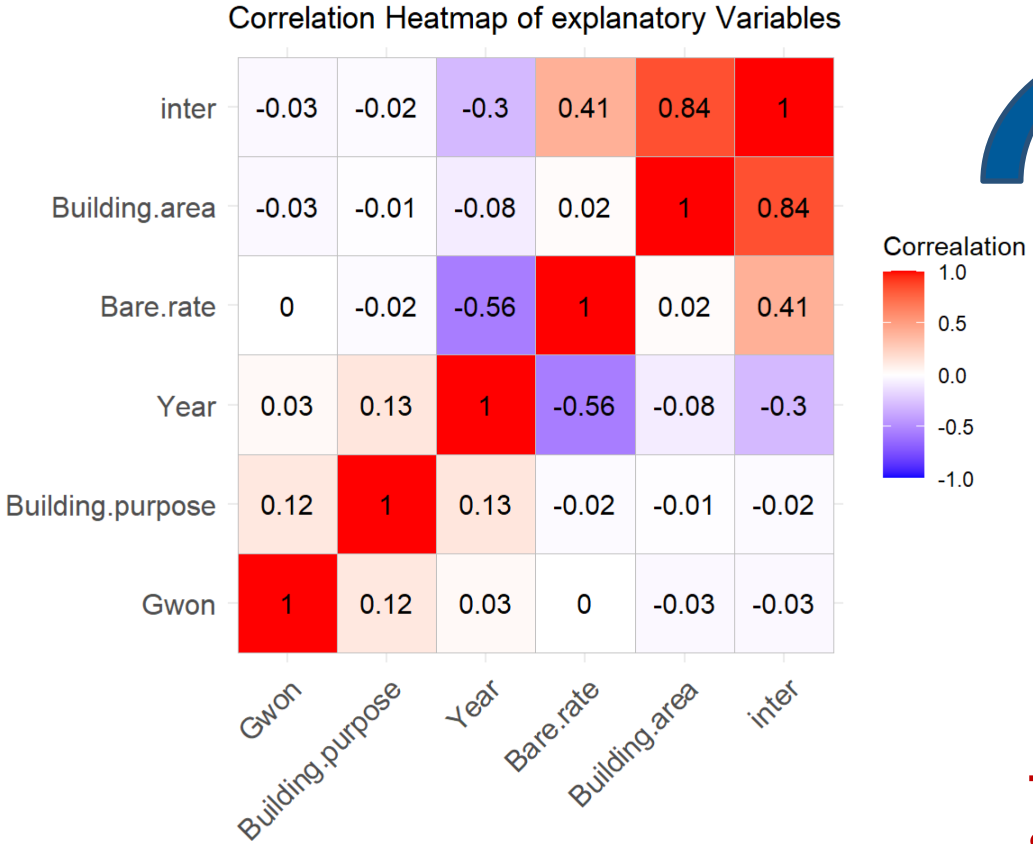


```
## Call:
## lm(formula = Price ~ as.factor(Year) + Gwon + Building.Area + 
##     Building.Purpose + base.rate + Building.Area * base.rate, 
##     data = df_train)
## 
## Coefficients:
##                        (Intercept)                 as.factor(Year)2018  
##                          395236887                           157485529  
##                as.factor(Year)2019                 as.factor(Year)2020  
##                          251615216                           232922629  
##                as.factor(Year)2021                 as.factor(Year)2022  
##                          321594169                           399433546  
##                 Gwonnortheast area                  Gwonnorthwest area  
##                         -359675324                          -253741555  
##                 Gwonsoutheast area                  Gwonsouthwest area  
##                          193507325                          -310782497  
##                      Building.Area           Building.Purposerow house  
##                            6606338                          -375920228  
## Building.PurposeSingle-family home    Building.Purposestudio apartment  
##                         -294156195                          -423431520  
##                          base.rate             Building.Area:base.rate  
##                          -22599445                             -719717  
```

✓ All variables were selected

3 Modeling

Multicolinearity Check



Correlation Heatmap of explanatory Variables

	Gwon	Building.purpose	Year	Bare.rate	Building.area	inter
inter	-0.03	-0.02	-0.3	0.41	0.84	1
Building.area	-0.03	-0.01	-0.08	0.02	1	0.84
Bare.rate	0	-0.02	-0.56	1	0.02	0.41
Year	0.03	0.13	1	-0.56	-0.08	-0.3
Building.purpose	0.12	1	0.13	-0.02	-0.01	-0.02
Gwon	1	0.12	0.03	0	-0.03	-0.03

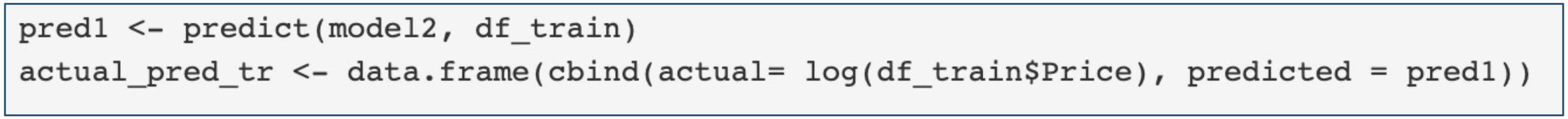
* inter = Building Area x Base Rate

Multicollinearity

→ The second model was selected as the best one

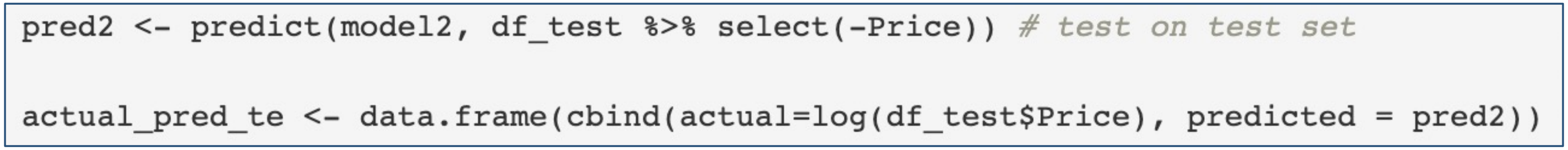
Calculation of Accuracy

# predicting values with training set



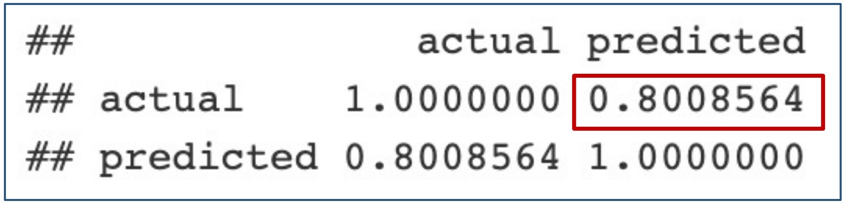
```
pred1 <- predict(model2, df_train)
actual_pred_tr <- data.frame(cbind(actual= log(df_train$Price), predicted = pred1))
```

# test with test set



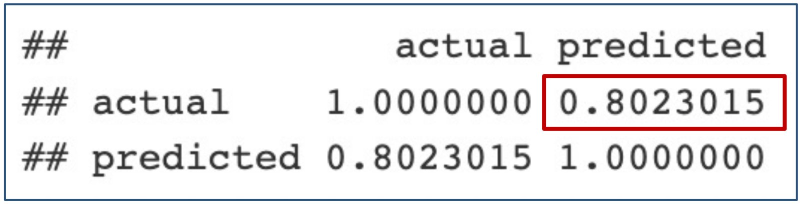
```
pred2 <- predict(model2, df_test %>% select(-Price)) # test on test set

actual_pred_te <- data.frame(cbind(actual=log(df_test$Price), predicted = pred2))
```

💡 Correlation Analysis between predicted values and actual ones



```
##              actual predicted
## actual    1.0000000 0.8008564
## predicted 0.8008564 1.0000000
```

Training Set



```
##              actual predicted
## actual    1.0000000 0.8023015
## predicted 0.8023015 1.0000000
```

Test Set

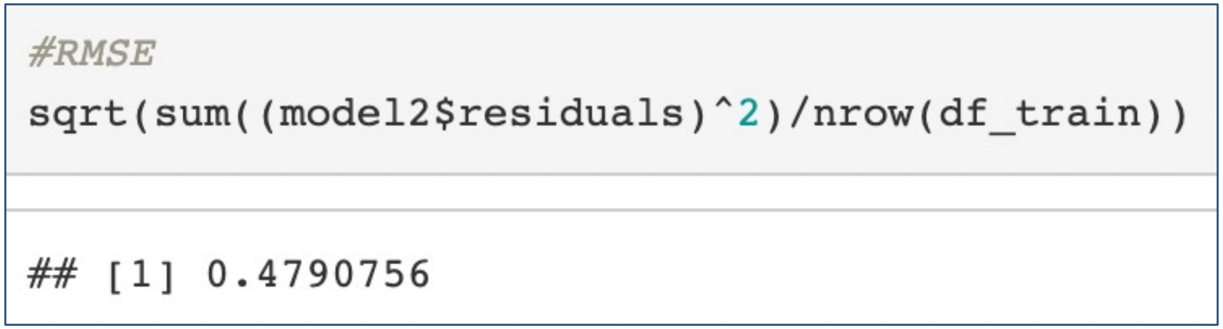
✓ Positively Linearly Related ✓

correlation of about 0.8 with the actual value of the data

Calculation of Accuracy and Error

Root-Mean-Square Error (RMSE)
: A measure of the differences between values predicted by a model and the actual values.

$$\sqrt{\mathrm{E}((\hat{\theta}-\theta)^2)}.$$



```
#RMSE
sqrt(sum((model2$residuals)^2)/nrow(df_train))
```

```
## [1] 0.4790756
```

4

Conclusions

4 Conclusions

Significances

- Transaction year, location of house(Gwon), building purpose, and building area influences on the price of real estates in Seoul. Especially, base rate was found out to have significant influence on the price. (negative correlation)
- By considering various factors of independent variables, it was possible to evaluate three models and select the best one .
- The selected linear regression model could predict the price of real estates with high similarity to the actual values.

Limitations

- The model at some point predicted the price as negative quantity because there was not enough consideration about exogenous variables except for base rate.
- The supply of apartments, LTV(Loan to Value Ratio), DTI(Debt-to-Income Ratio) would be proper additional exogenous variables.
- More precise prediction would have been possible if we could proceed with more specified location data, such as ‘Dong.’

🏠 THANK YOU 🏠