# Executive Summary Report - House Price in New York

| 490585676 | 510006222 | 500002526 | 510239602 | 510323026

This version was compiled on November 6, 2022

**The aim of this report is to predict which factors influence the house price the most. The dataset we used in this report is House price in New York. We use some models to analyze the result depending on many variables such as living area, land value and ages of the house. First, we read in the dataset and cleaned the missing value and Na value. Then we visualized the distribution of dependent variables to check the normality and linearity. We create 5 models to check which one is the most suitable for our topic. Afterwards we are using forward and back forward selections. Finally, we keep the 6 most relevant variables. We checked each model's RMSE, MAE, AIC. The log-log model is the most appropriate model. Finally, we discussed the result and raised some limitations of our report and the points we can improve.**

House Price | Multiple Linear Regression

## Introduction

The main idea of our report is to predict the house price will be influenced by what variables. The dataset was taken from full Saratoga Housing Data. We will use some models to predict the result and analyze the best model for our topic. The property development will be our target audience. Through the report we will discuss the relationship between each variable and house price in New York. The main goal of property development is making money. Through this report they can know which variables affect house price the most. Some obvious variables can influence the house price for example living area.

## Data description

The dataset was a random sample of 1734 houses. It is taken from full saratoga housing data. There are 1734 rows and 17 columns in this dataset.The house price is the main variable. There are 6 key factors are lot size, water frontage, land value, living area, bathrooms, and new construction. The land value and living area shows the value of the land and the size of the living area. Water frontage represents the property including the waterfront or not. Bathrooms show the number of bathrooms and new construction shows if it is new. Limitations: This dataset is a random sample, so the standard is not the same. The dataset shows some variables that may affect the house price. But it does not cover many fields, for example transportation or a good shopping mall. These two things can convenient the life convenient and affect the price. This dataset is more focused on the house itself, not from the owner's point of view. Some description of the data is not clear and there are some missing values in this dataset.

## Analysis

**Model selection:.** To make the most appropriate model, we create five models. Through the linear-linear full model and the residual plot, the RSE is high, and the residuals are scattered and high.

And in our IDA , the box plot also shows there are a number of outliers. In this regard, we preliminarily judge that we need to perform log transformation on the dependent variable (price). In order to get a more suitable model, we need to get the correlation of each variable with the dependent variable, and then leave the variables with high correlation. We decided to filter the variables by AIC which is a very scientific and effective model selection method. After using forward and backward selection, variables that don't have much effect on the price have all been removed, we keep the 6 most relevant variables. We have previously observed that the land value was the least normal and interested in it. So, we do log transformation on the land value and obtain a linear-log model, however, the RSE is still high.

As mentioned, we do log transformation on the price and obtain a log-linear model (Fig.2). The residual standard error is significantly lower. Finally, log two variables simultaneously and obtain the log-log model (Fig.3), the result is very close to the previous model.

To get the most appropriate model, we use 10-fold cross validation. By comparing the results (Fig.4 and Fig.5), we find that the log-log model has the highest R-squared and lowest RMSE and MAE. Therefore, the log-log model is our most appropriate model.

**Assumption check:.** In this part, we will focus on the figure 6 provided in the appendix. First, we will be looking for **linearity**. Based on the residuals versus fitted values plot, we can observe that there are some outliers that occur in the plot. The residuals roughly form a horizontal band around the 0 line. This suggests that the variances of the error terms are equal. The equally spread residuals around a horizontal line without distinct patterns are a good indication of having the linear relationships. Second, we will be looking for **independence**. From the residuals versus fitted values plot, we can again discover that the relation can be also considered as an independence observation. Third, we will be looking for **homoscedasticity**. Since the residuals are spread along the range of predictors, and there was not a funnel shape in the plot, we can say that the data is homoscedasticity. Finally, we will be looking for **normality**. The normality assumption is satisfied because the points in the QQ plot follow the line closely and therefore the residuals can reasonably be assumed to follow a normal distribution.

## Result

$$\begin{aligned} log(price) = 10.10335 + 0.03328\ lot\ size + 0.53084\ waterfront \\ +0.12108\ log(land\ value) - 0.08690\ new\ construct \\ +0.00032\ living\ area + 0.14073\ bathroom \end{aligned}$$

The independent variable is New York house prices, and the dependent variable is six influencing factors. For house prices, the change in price changes as the dependent variable changes. The best fit among the five models is the log-log model. We can know that on average, a one percent increase in land value will result in a 0.12108% change in house price in New York.

## Conclusion and discussion

In the above study, we found that there are several key factors that influence prices in New York throughout the data. These factors are lot size, water frontage, land value, living area, bathrooms, and new construction. In addition to that, we got log -log is the most suitable model for our data. After our discussion, we agreed that this article would be most beneficial for real estate developers in New York. This is because they can decide how to maximize their profits when selling their homes based on the factors analyzed in this article. Of course, there are some limitations to the timing of the data release. For example, because the data was released so early, it has no reference value for current New York home prices.

## References


Eddelbuettel, D. (2021, November 4). eddelbuettel/pinp. GitHub. https://github.com/eddelbuettel/pinp

Valiente, C., Swanson, J., & Eisenberg, N. (2011). Linking Students' Emotions and Academic Achievement: When and Why Emotions Matter. Child Development Perspectives, 6(2), 129–135. https://doi.org/10.1111/j.1750-8606.2011.00192.x

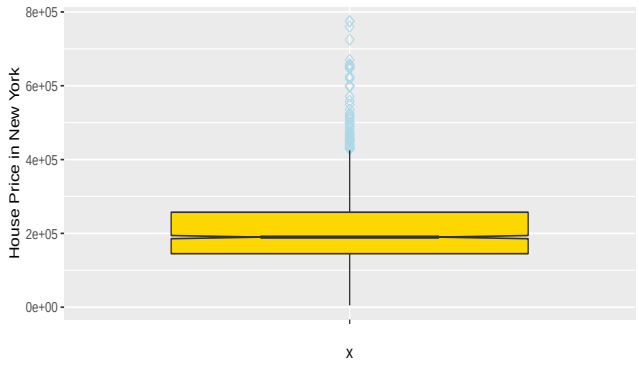
Gujarati, Damodar N.; Porter, Dawn C. (2009). "How to Measure Elasticity: The Log-Linear Model". Basic Econometrics. New York: McGraw-Hill/Irwin. pp. 159–162. ISBN 978-0-07-337577-9.

Making PowerPoint Slides with R. (n.d.). Rstudio-Pubs-Static.s3.Amazonaws.com. Retrieved October 23, 2022, from https://rstudio-pubs-static.s3.amazonaws.com/271122_ab8134500037448f829d1768e5364c14.html

Xie, Y. (2022, October 22). xaringan. GitHub. https://github.com/yihui/xaringan

## Appendix

Fig. 1.Box plot for House Price in New York




```
#
#  Call:
#  lm(formula = loPrice ~ lot_size + land_value + wate
#      living_area + bathrooms, data = data)
#
#  Residuals:
#       Min       1Q   Median       3Q      Max
#  -3.7169 -0.1583  0.0186  0.1757  1.3656
#
#  Coefficients:
#                  Estimate Std. Error t value Pr(>|t|)
#  (Intercept)    1.119e+01  2.400e-02 466.408  < 2e-1
#  lot_size       3.537e-02  1.053e-02   3.360 0.00079
#  land_value     3.485e-06  2.331e-07  14.949  < 2e-1
#  waterfront     4.312e-01  7.889e-02   5.465 5.29e-0
#  new_construct -1.391e-01  3.609e-02  -3.853 0.00012
#  living_area    3.169e-04  1.796e-05  17.643  < 2e-1
#  bathrooms      1.464e-01  1.585e-02   9.237  < 2e-1
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
#
#  Residual standard error: 0.301 on 1726 degrees of f
#  Multiple R-squared:  0.5659, Adjusted R-squared:  0
#  F-statistic: 375.1 on 6 and 1726 DF,  p-value: < 2.2e-16
```

**Fig. 2. Log-linear model**

```
#
#  Call:
#  lm(formula = lorrice ~ lot_size + waterfront + lorla
#      living_area + bathrooms, data = data)
#
#  Residuals:
#       Min       1Q   Median       3Q      Max
#  -3.7627 -0.1612  0.0139  0.1761  1.4238
#
#  Coefficients:
#                   Estimate Std. Error t value Pr(>|t|)
#  (Intercept)    10.1033500  0.0721861 139.963  < 2e-16
#  lot_size        0.0332755  0.0104579   3.182  0.00149
#  waterfront      0.5308395  0.0776950   6.832 1.15e-11 ***
#  lorland         0.1210839  0.0076705  15.786  < 2e-16 ***
#  new_construct -0.0868981  0.0354929  -2.448  0.01445 *
#  living_area     0.0003208  0.0000177  18.124  < 2e-16 ***
#  bathrooms       0.1407319  0.0157474   8.937  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
#
#  Residual standard error: 0.2991 on 1726 degrees of freedom
#  Multiple R-squared:  0.5716, Adjusted R-squared:  0.5701
#  F-statistic: 383.8 on 6 and 1726 DF,  p-value: < 2.2e-16
```

**Fig. 3. Log-log model**

```
#  Linear Regression
#
#  1733 samples
#     6 predictor
#
#  No pre-processing
#  Resampling: Cross-Validated (10 fold)
#  Summary of sample sizes: 1561, 1559, 1560, 1558, 1560, 1560,
#  Resampling results:
#
#    RMSE       Rsquared   MAE
#    0.2983242  0.5710149  0.2112457
#
#  Tuning parameter 'intercept' was held constant at a value of
```

**Fig. 4. Log-linear model after 10-fold cross validation**

```
#  Linear Regression
#
#  1733 samples
#     6 predictor
#
#  No pre-processing
#  Resampling: Cross-Validated (10 fold)
#  Summary of sample sizes: 1560, 1561, 1560, 1558, 1559, 1561,
#  Resampling results:
#
#    RMSE       Rsquared   MAE
#    0.2978949  0.5774116  0.2111992
#
#  Tuning parameter 'intercept' was held constant at a value of
```

**Fig. 5. Log-log model after 10-fold cross validation**

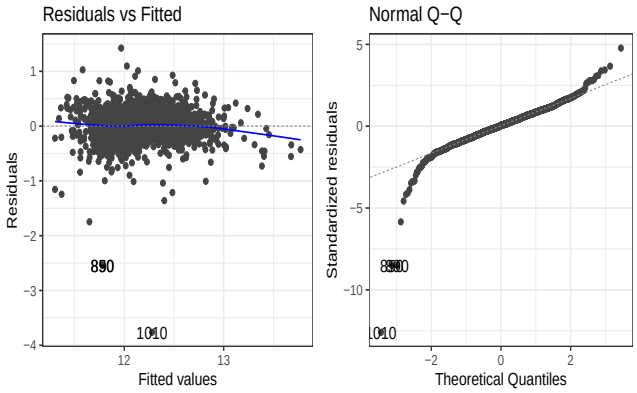



**Fig. 6. Assumption check**