Zillow

ZESTIMATE

Zillow’s Home Value Prediction

Presented by Dương Tiến Thành
Trần Sơn Tùng
Mai Ngọc Hiếu

Zillow

OVERVIEW

1. Introduction
2. Data Preprocessing
3. Exploratory Data Analysis
4. Feature Engineering
5. Building Regression Model
6. Result Evaluation

Zillow

INTRODUCTION

- Zillow is an online real estate marketplace.
- Connect and support buyers, sellers, renters, etc.
- Find and share important information related to homes, real estate, mortgages.

Zillow

- Zestimates: automated home valuation tool
- Provides quick property value estimates for user reference
- predict the log-error between Zestimate and the actual sale price

logerror = log(Zestimate) - log(SalePrice)

Zillow

OUR TARGET

1. Improved accuracy of Zestimates
2. Better user experience
3. Optimized business processes
4. More effective risk management

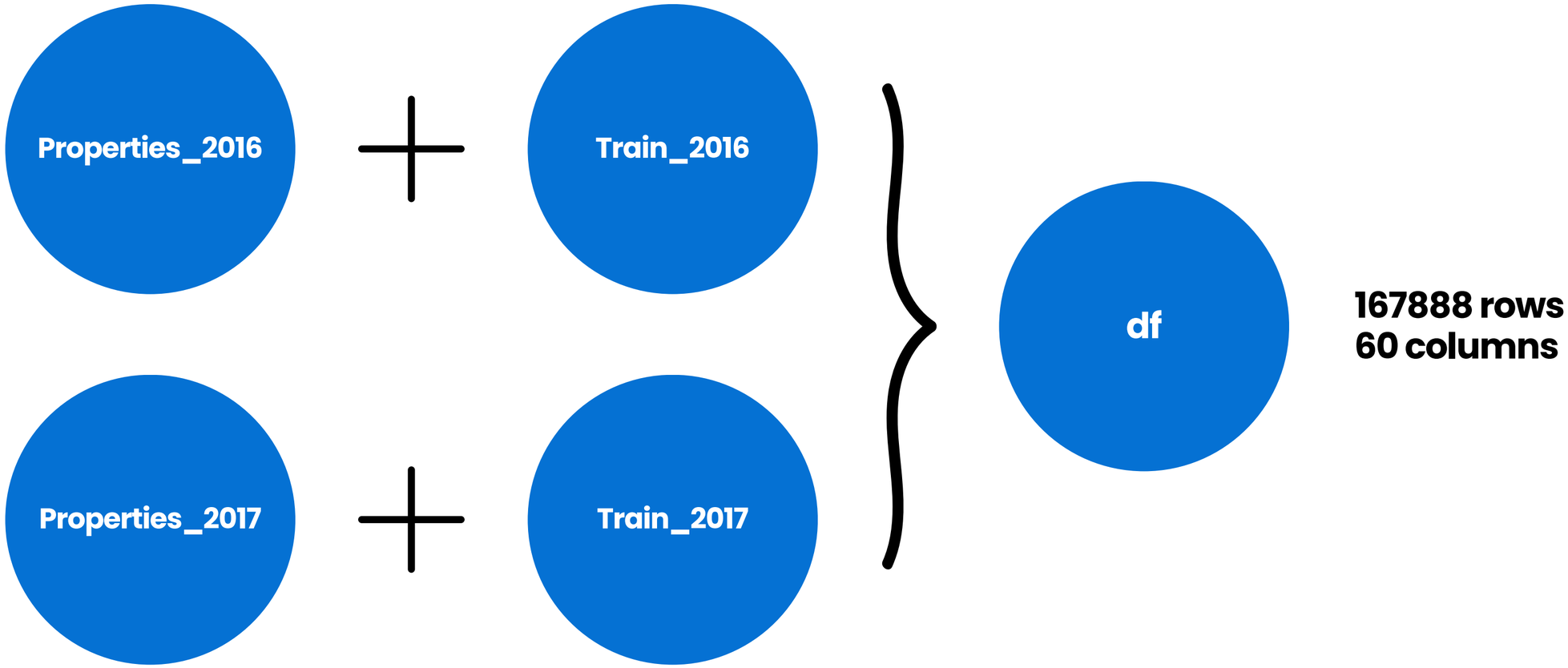
Zillow

DATA

1. properties_2016.csv, properties_2017.csv
2. train_2016_v2.csv, train_2017.csv
3. Zillow_data_dictionary.xlsx

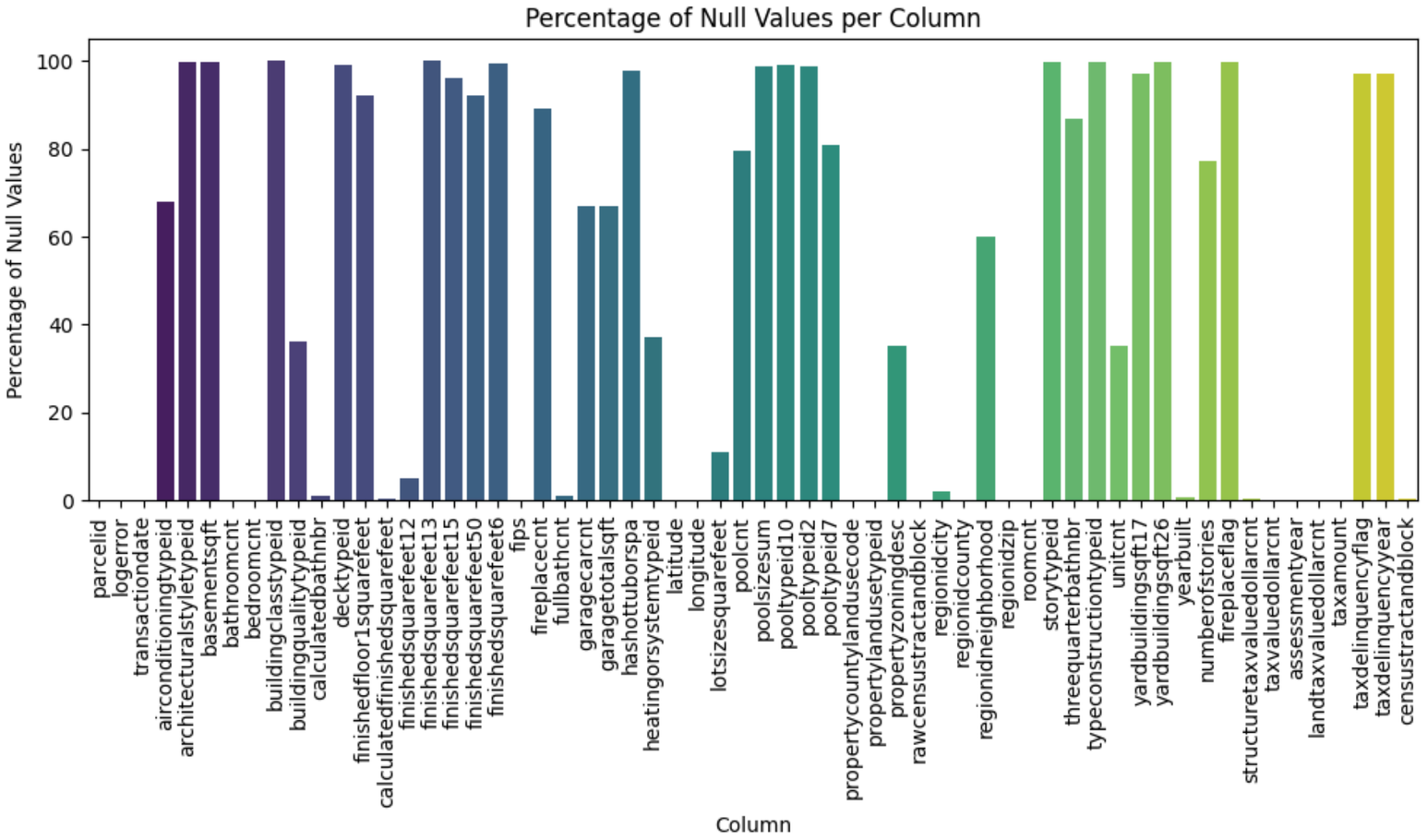
Zillow

Data Preprocessing



Zillow

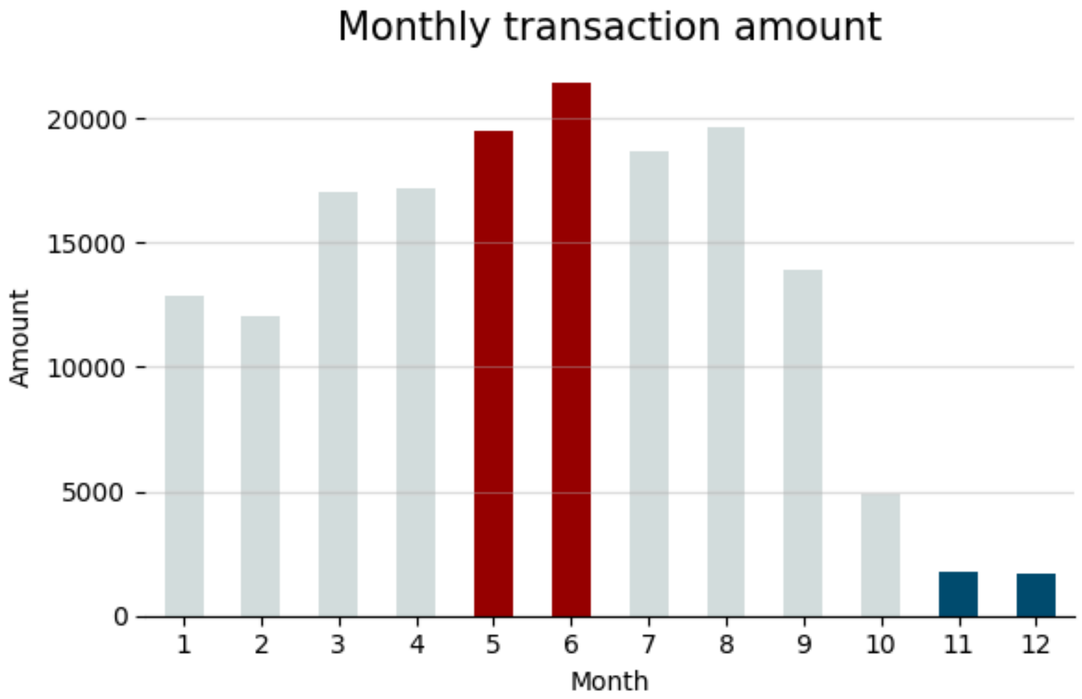
Null and Duplicate Values

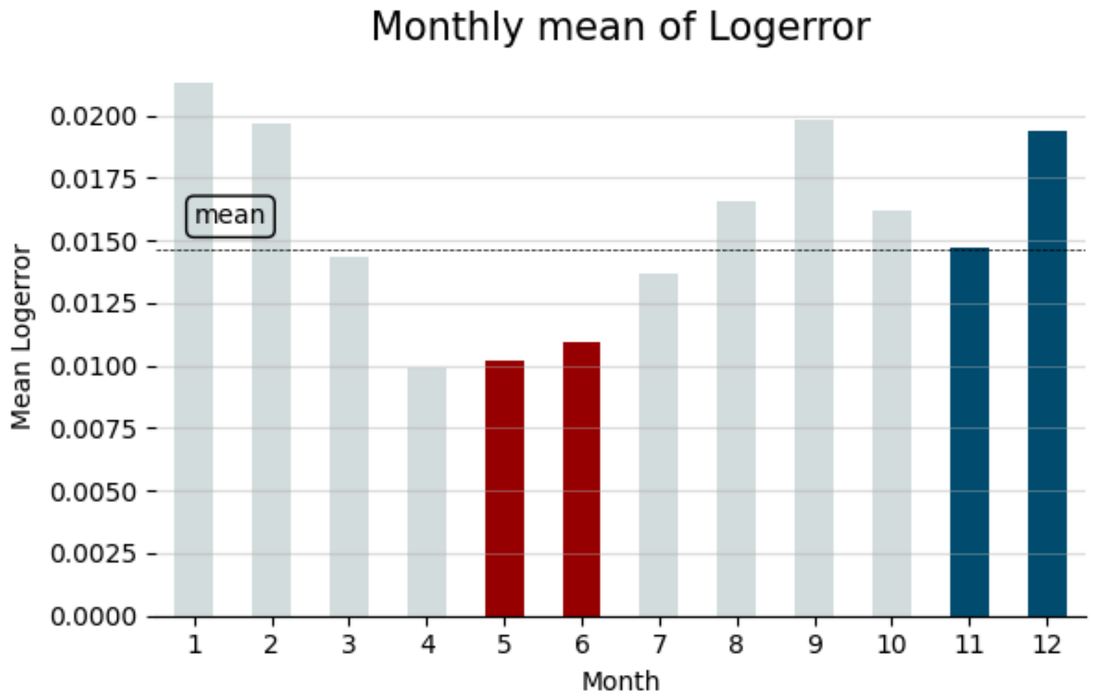


- 2678 duplicate rows
- Remove columns with Null >= 60%
- Remain columns with Null <= 2%
  -> Remove rows
- The others:
  - Replace with mean value
  - Replace with mode value

Zillow

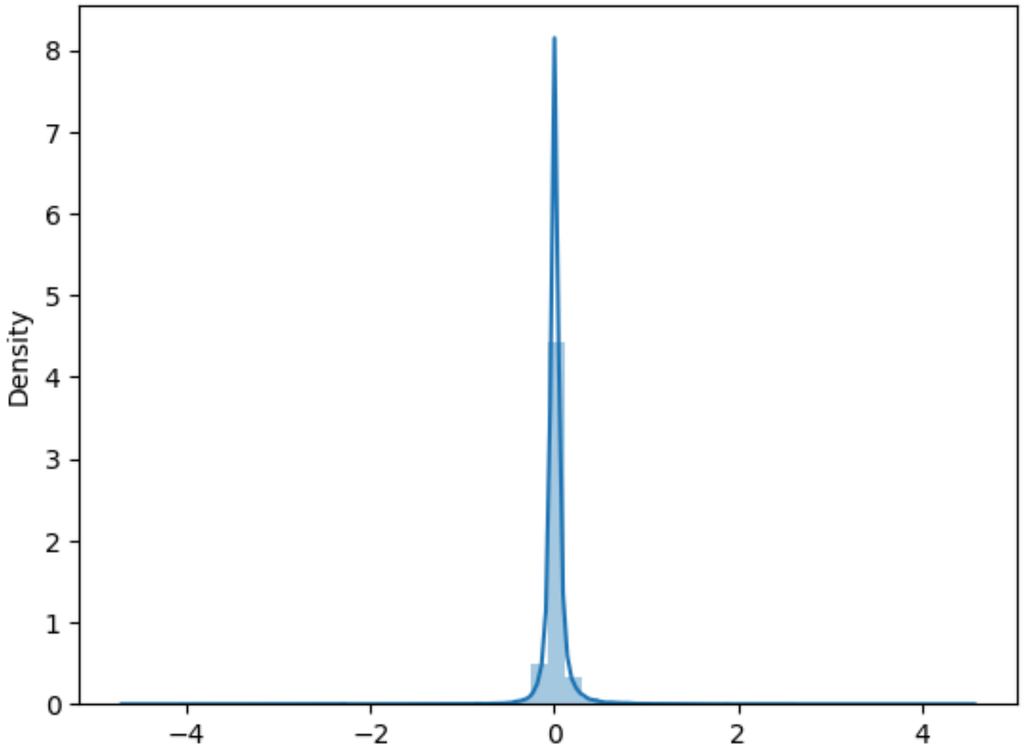
Exploratory Data Analysis

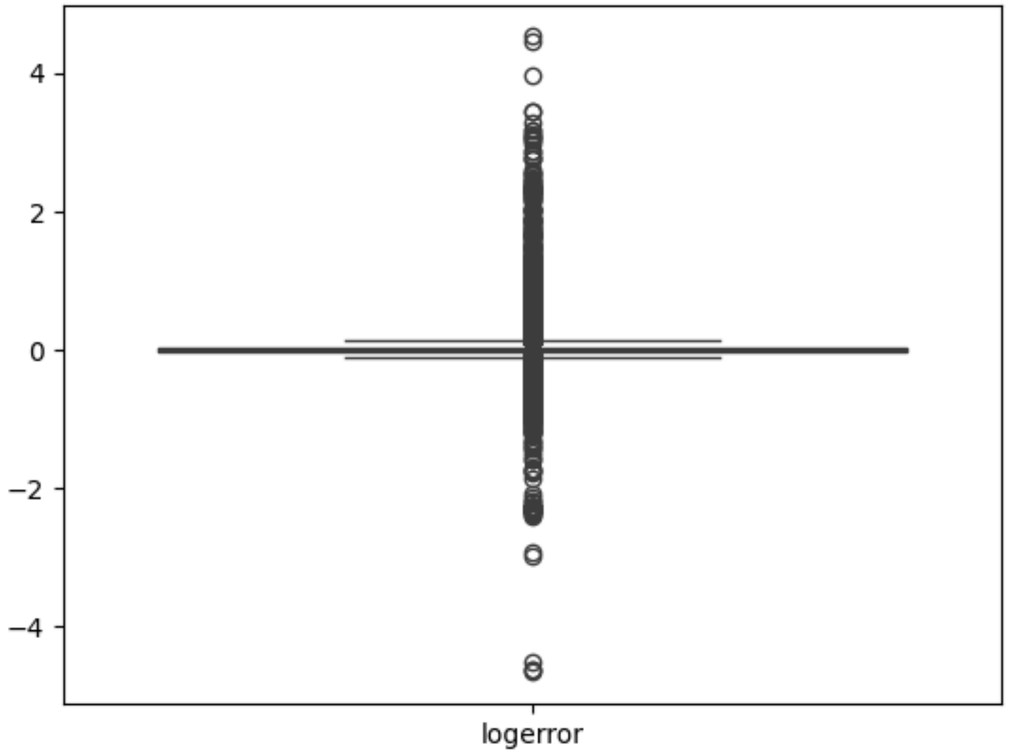




Zillow

Exploratory Data Analysis





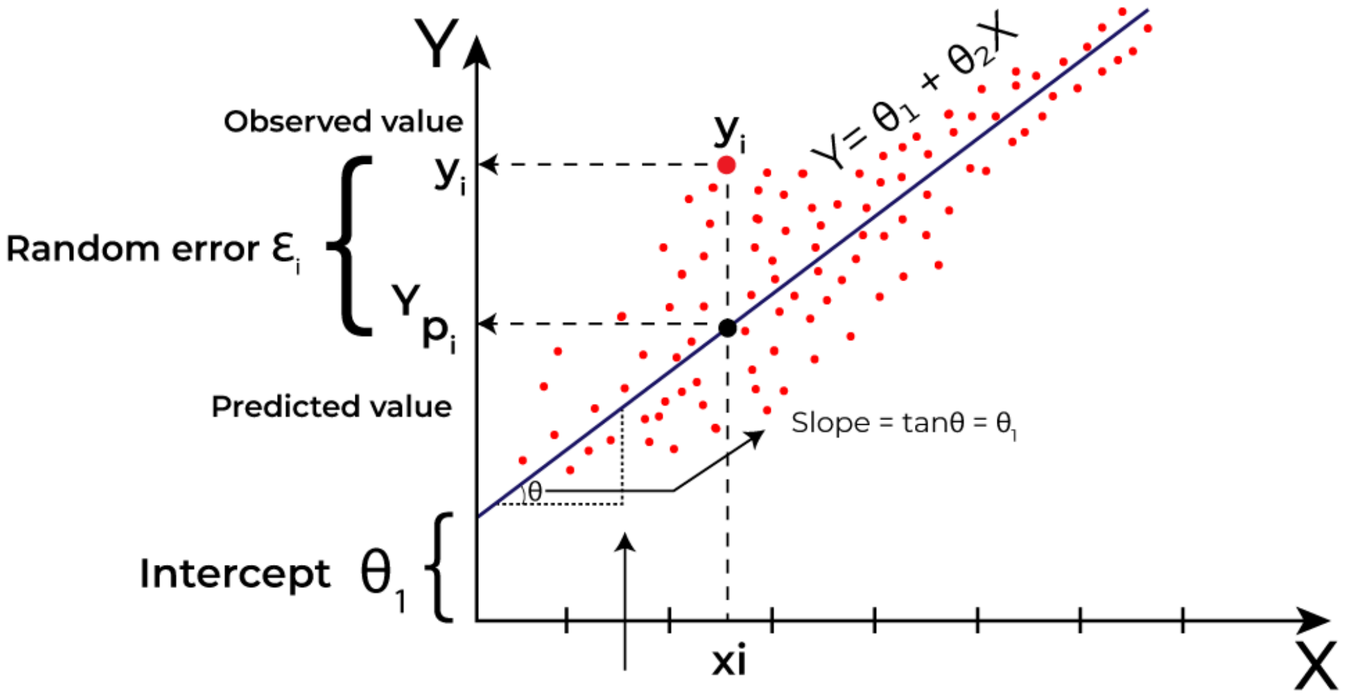
Zillow

Feature Engineering

- ***Creating features from transaction dates*:**
  Columns related to year, month, day, and weekday are derived from the property transaction date.
- ***Calculating the age of the house:***
  The house age is calculated by finding the difference between the year it was built and the transaction year.
- ***Total number of rooms:***
  The total number of rooms is calculated by adding the number of bathrooms and bedrooms.
- ***Garden area:***
  The garden area is computed by subtracting the finished living area from the total lot size.
- ***Living area ratio:***
  Another important feature is the living area ratio, which is the proportion of finished living space relative to the total lot size.

Zillow

Linear Regression



- Simple and Easy to Understand
- Easy to Deploy
- Resource-Efficient

Zillow

Model evaluation method

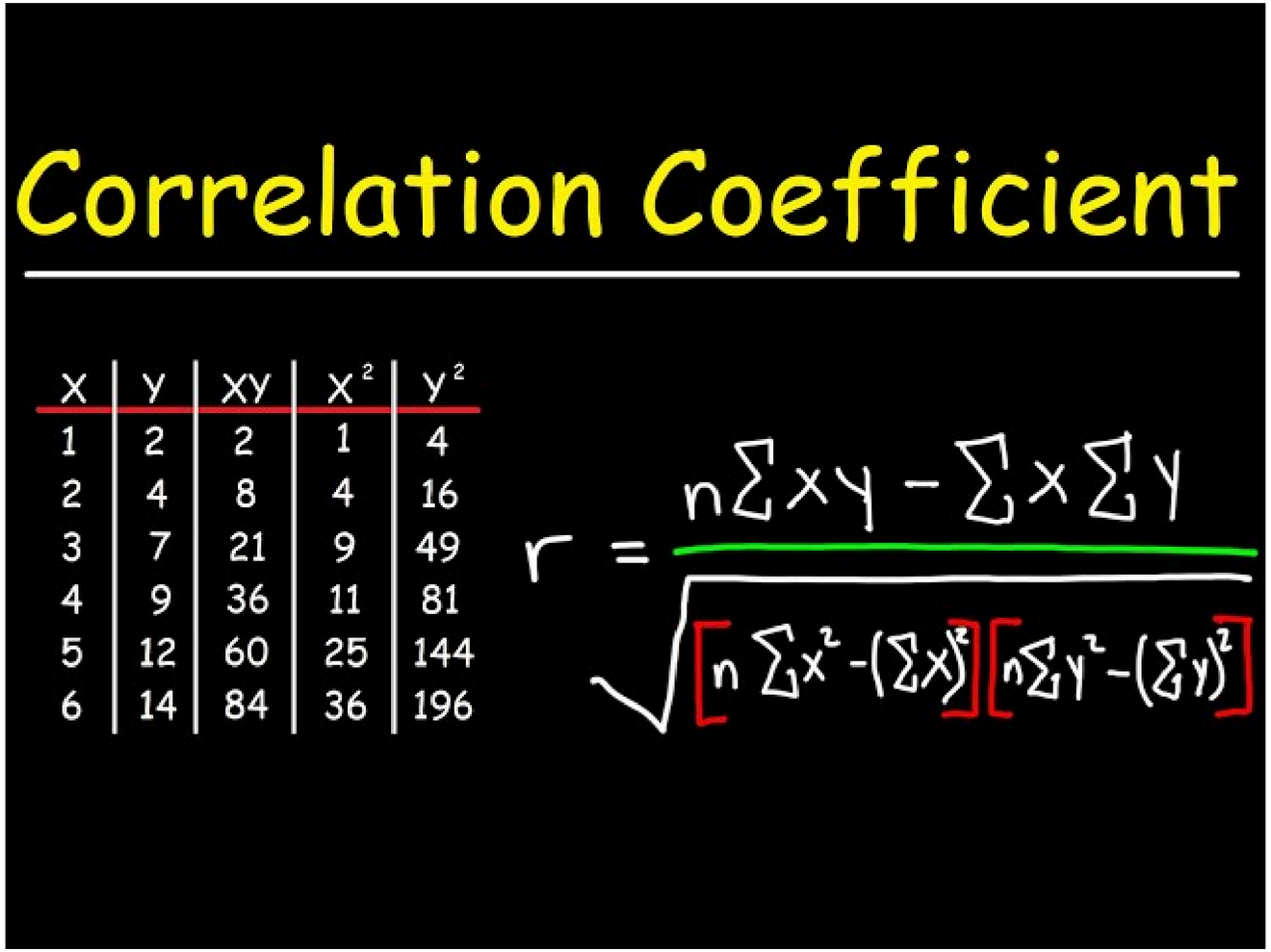
- **Root Mean Squared Error (RMSE):**
  - Simplicity
  - Sensitive to Outliers
- **Mean Absolute Error (MAE):**
  - Easy to Calculate and Implement
  - Handle Outliers well

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m}\left|h(x^{(i)}) - y^{(i)}\right|^2$$

Zillow

Correlation



Zillow

Selected input variable

- *"bedroomcnt',*
- *'buildingqualitytypeid',*
- *'calculatedbathnbr',*
- *'finishedsquarefeet12',*
- *'heatingorsystemtypeid',*
- *'lotsizesquarefeet',*
- *'rawcensustractandblock',*
- *'roomcnt',*
- *'unitcnt'*
- *‘yearbuilt'*
- *'structuretaxvaluedollarcnt'*
- *'assessmentyear',*
- *'landtaxvaluedollarcnt',*
- *'censustractandblock',*
- *'house_age',*
- *'total_room',*
- *'garden',*
- *'living_area_ratio'*
- *‘yearbuilt',*
- *'structuretaxvaluedollarcnt’*

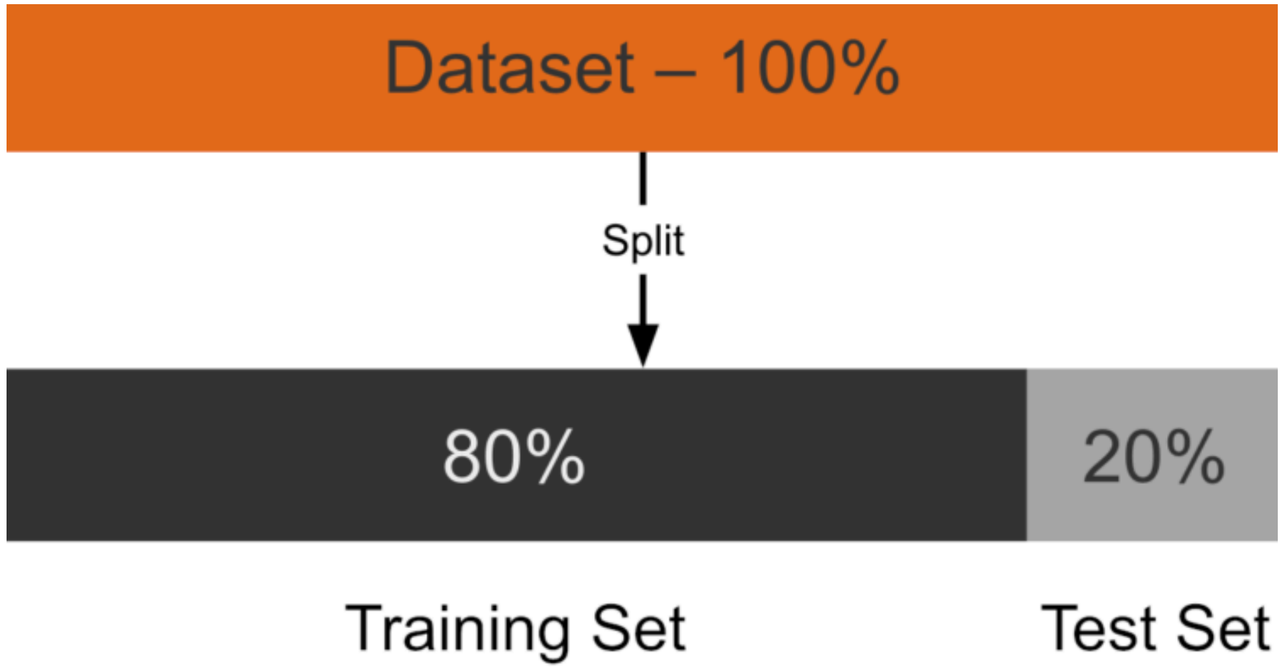
Zillow

Building Regression Model

Target variable

Logerror

Split data set



Zillow

The linear regression model combined with Cross-Validator.

Mean Absolute Error(MAE)	*Root Mean Squared Error(RMSE)*
0.0566296	0.0959415

Zillow

Suggestions for improving the model

- **Collect more Data from other resources**
- **Utilize the Data that were given to gain more insight**
- **Dive more into Feature Engineering**
- **Applying Advanced Machine Learning Techniques**

THANK YOU