REPORT
AFFECT OF DATA PREPARATION IN MACHINE LEARNING

RESEARCH

Systematic comparison of multiple preprocessing techniques across multiple datasets and models. Straightforward conclusion: the accuracy and efficiency of ML algorithms strongly depend on the quality & structure of the input data, data preprocessing is a key step in the pipeline.

Yasodha (2025) – Data Preprocessing Methods for Machine Learning: An Empirical Comparison

Research in the context of industrial production. The author points out that many ML projects fail because of poor data quality, and estimates that data preprocessing accounts for about 80% of the time & resources of an ML project

Frye et al. (2021) – Benchmarking of Data Preprocessing Methods for Machine Learning – Applications in Production

Initial Number of Columns

32

Initial Number of Records

79.884K

Building Classes like: **CONDOMINIUM**, **CO-OP** and Records have **SALE PRICE = 0** will be not analyzed
There are **very little** values of **EASEMENT** and **APARTMENT NUMBER** and **massive values** of **ADDRESS**, so it must be **dropped**

After Number of Columns

29

After Number of Records

24K

LIST OF COLUMNS

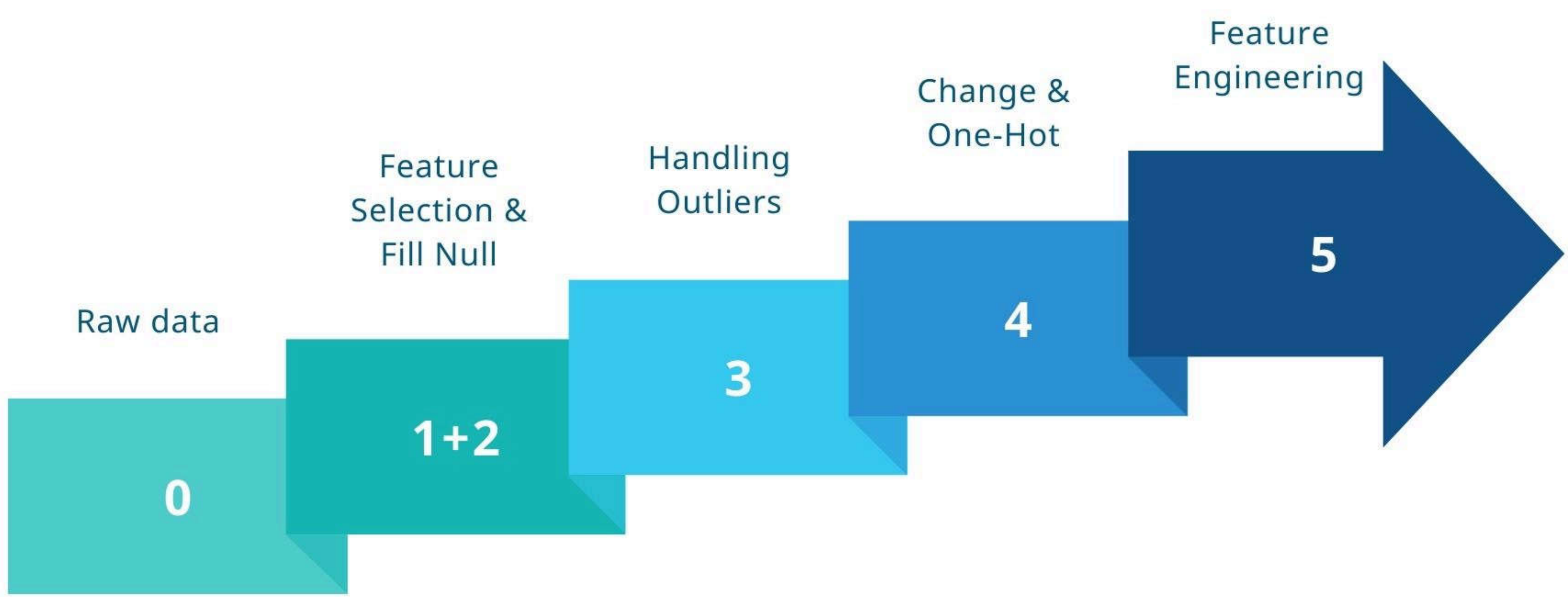
CATEGORICAL

- Neighborhood
- Building class category
- Tax class at present
- Building class at present
- Building class at time of sale

NUMERICAL

- Borough
- Block
- Lot
- Zip code
- Residential unit
- Commercial unit
- Land square feet
- Gross square feet
- Total unit
- Year built
- Tax class at time of sale
- Sale date

4-Step Data Preparation



FEATURE SELECTION

To construct a **machine learning model**, it is first necessary to identify the variables that **influence** the **price of a real estate transaction**.

Variables **BUILDING CLASS CATEGORY** and **TAX CLASS AT TIME OF SALE** represent categorical attributes describing the type of property.
ZIP CODE may serve as a proxy for the broader group of geographical location factors.

Number of ZIP CODE

189

Each **ZIP code** represents a **different postal area**, so this variable can be used as a proxy for geographic location. Since real estate values are **strongly driven by location**, ZIP code is a potentially important feature in the model.

Variables describing **area, age, number of units** captures key characteristic features that are highly considered when assessing housing prices.

FEATURE SELECTION

	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQ. FEET	GROSS SQ. FEET	YEAR BUILT
COMMERCIAL UNITS	0.03					
TOTAL UNITS	0.99	0.20				
LAND SQ. FEET	0.06	0.03	0.07			
GROSS SQ. FEET	0.66	0.38	0.71	0.43		
YEAR BUILT	0.06	-0.01	0.06	0.02	0.06	
SALE PRICE	0.11	0.30	0.16	0.03	0.35	0.01

There is no clear linear relationship between **LAND SQUARE FEET** and **SALE PRICE**. However, **LAND SQUARE FEET** is still included in the model, since it may contribute through **more complex (non-linear or interaction) effects**. The actual importance of this feature will be **examined** in more detail once the model results are available.

Feature Selection

- NEIGHBORHOOD
- BUILDING CLASS CATEGORY
- TAX CLASS AT PRESENT
- BUILDING CLASS AT PRESENT
- BUILDING CLASS AT TIME OF SALE

Categorical

- BOROUGH
- BLOCK
- LOT
- ZIP CODE
- RESIDENTIAL UNIT
- COMMERCIAL UNIT
- LAND SQUARE FEET
- GROSS SQUARE FEET
- TOTAL UNIT
- YEAR BUILT
- TAX CLASS AT TIME OF SALE
- SALE DATE

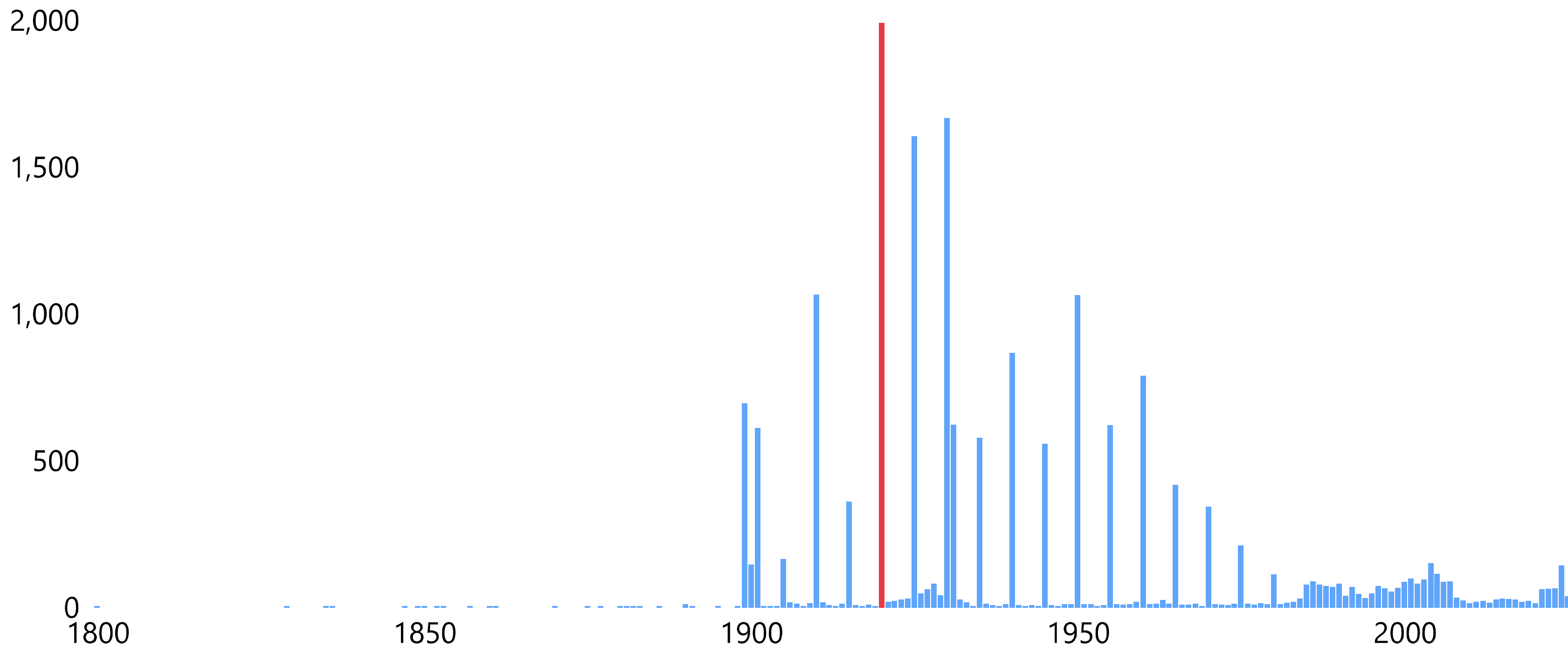
Numerical

FILL NULL VALUE

There are 5 records having **null value** of **ZIP CODE** that can be easily filled by **searching information** based on address

Many **null values** of **YEAR BUILT** will be filled by 1920 in which **the greatest number** of buildings were constructed.

The year 1920 recorded the highest number of buildings constructed.

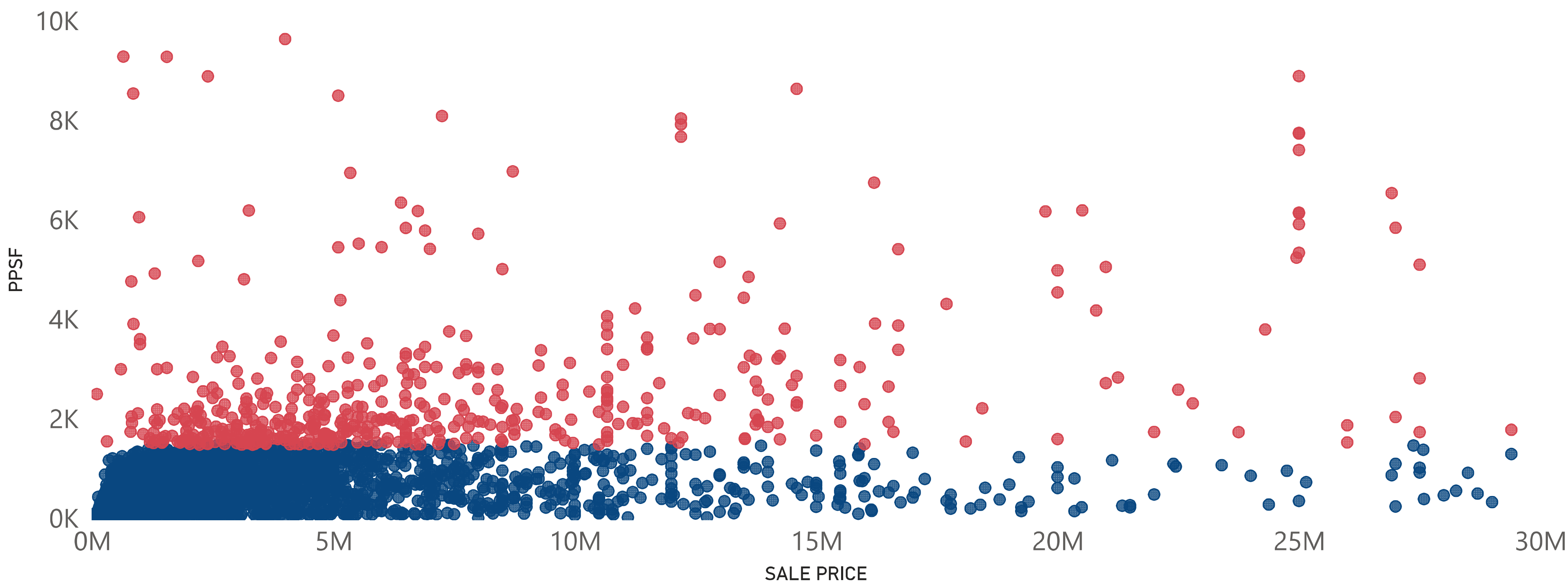


HANDLING OUTLIERS

There are numerous records in which **the property area is very small**, yet **the prices are unusually high.** At the same time, many transactions with a **price of lower 1000**—typically corresponding to transferred or inherited units—**are excluded** from the dataset.

Too Small building with very high price will be dropped

This chart doens't show extremely PPSE and SALE PRICE

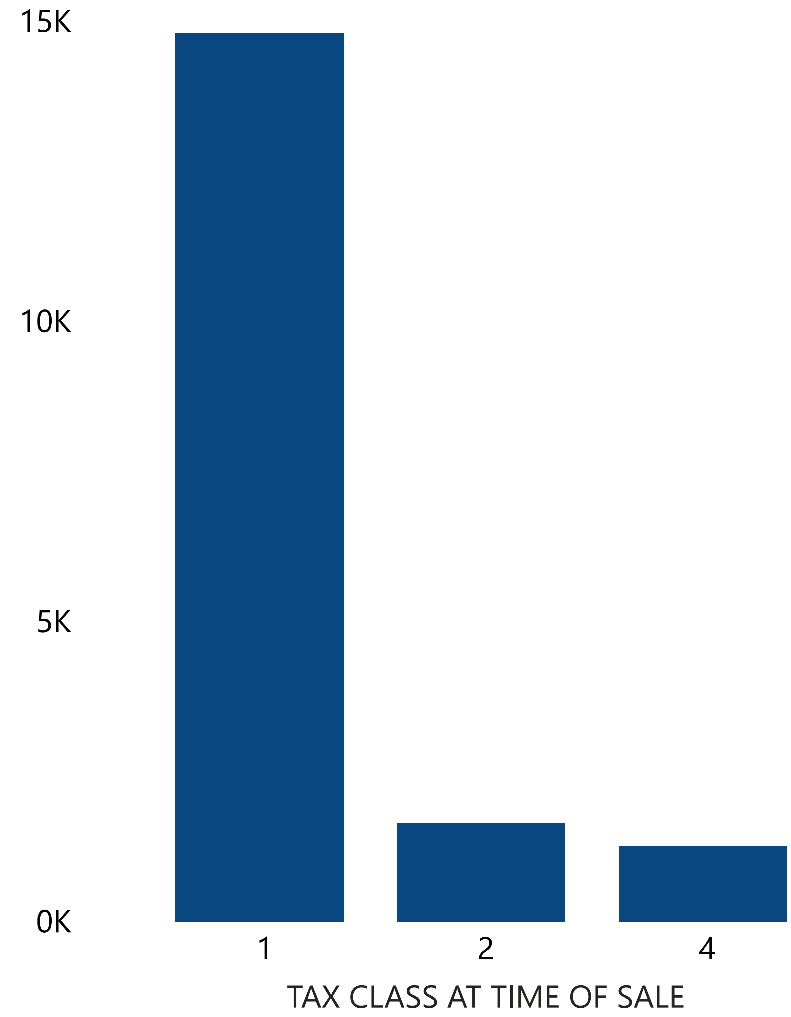


CONVERT FEATURE TYPES

TAX CLASS and **ZIP CODE** are currently treated as numerical variables; however, they **do not carry any mathematical meaning**. They are merely encoded categorical values.

Hence, it is necessary to **transform** them into **character-based** representations.

Distribution of TAX CLASS

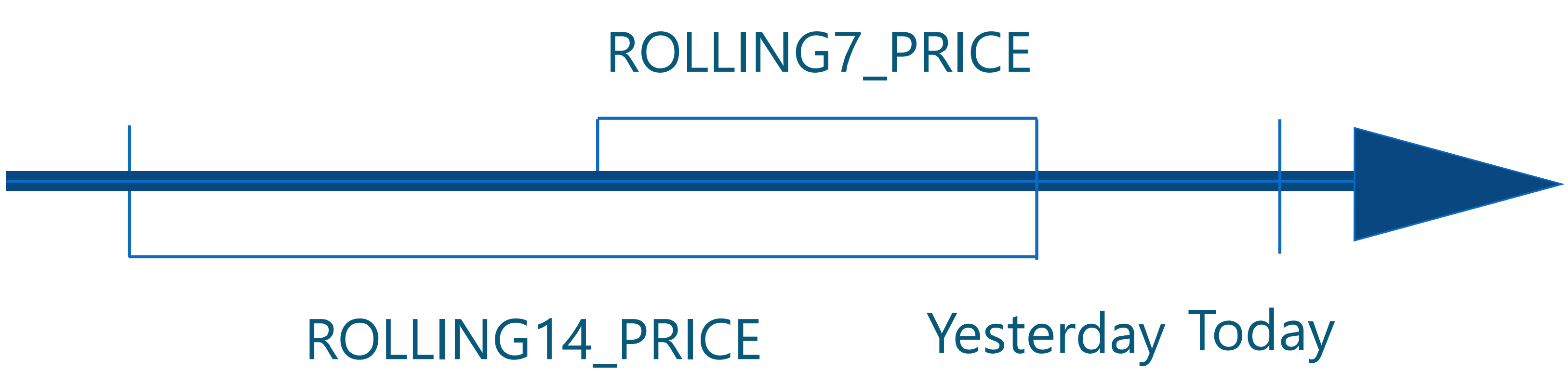


FEATURE ENGINEERING

BUILDING AGE = SALE YEAR - YEAR BUILT

The **year of construction** is converted into the **building's age** to **reduce the feature's range.**

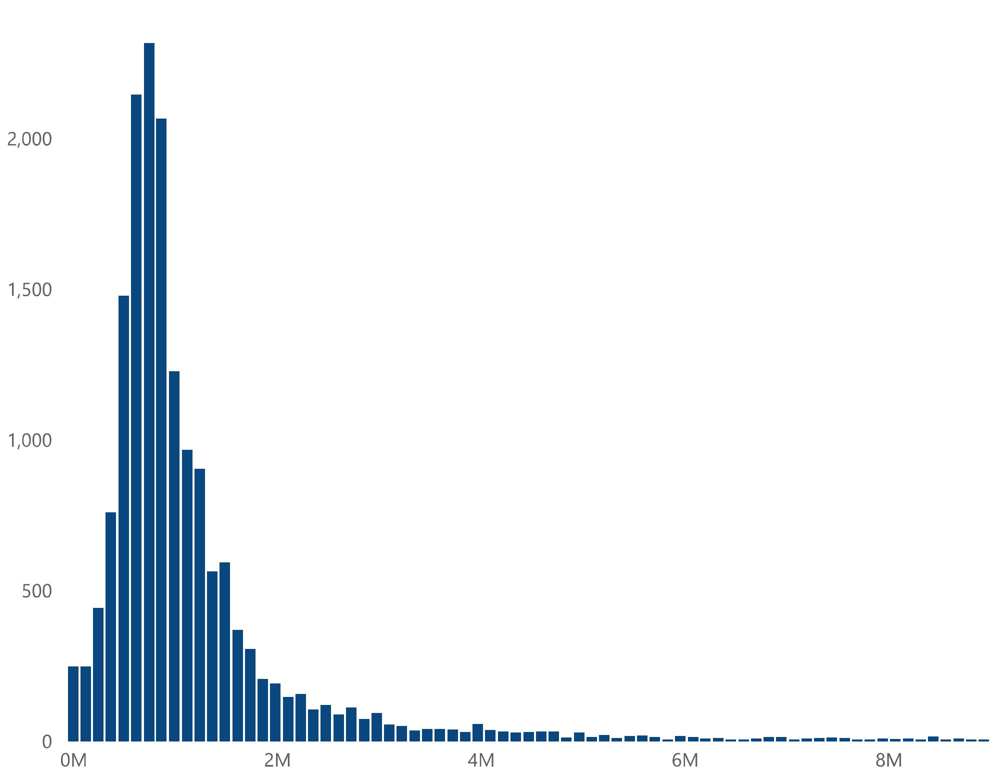
In addition, based on information from **SALE DATE**, the variables **ROLLING7_PRICE** and **ROLLING14_PRICE** are generated by computing the **average** real estate market price within the same area over the preceding 7 and 14 days. These features provide insight into **market conditions** as well as **short-term pricing trends**.



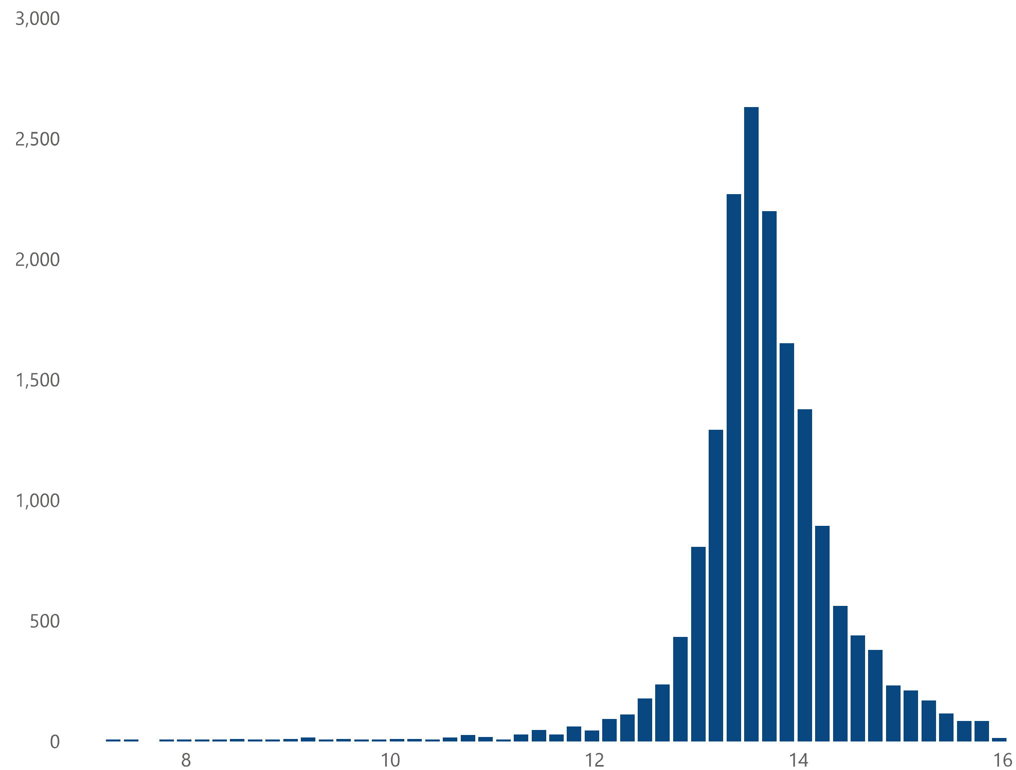
DATA TRANSFORMATION

In order to **normalize** the value distribution, **log-transformations** are applied to ***SALE PRICE***, ***LAND SQUARE FEET***, and ***GROSS SQUARE FEET*.**

SALE PRICE distribution



SALE PRICE distribution after log-transformation

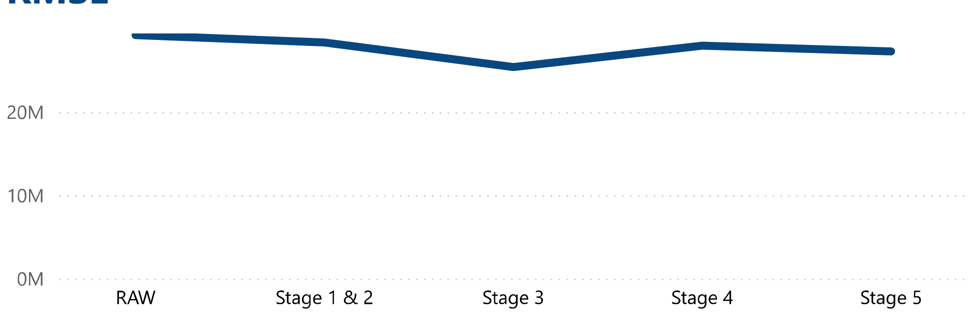


MODEL RESULT

R^2 indicates the proportion of the variability in prices that is explained by the model.
MAE (Mean Absolute Error) measures the average magnitude of the absolute errors.
RMSE (Root Mean Squared Error) measures the square root of the average of the squared errors.

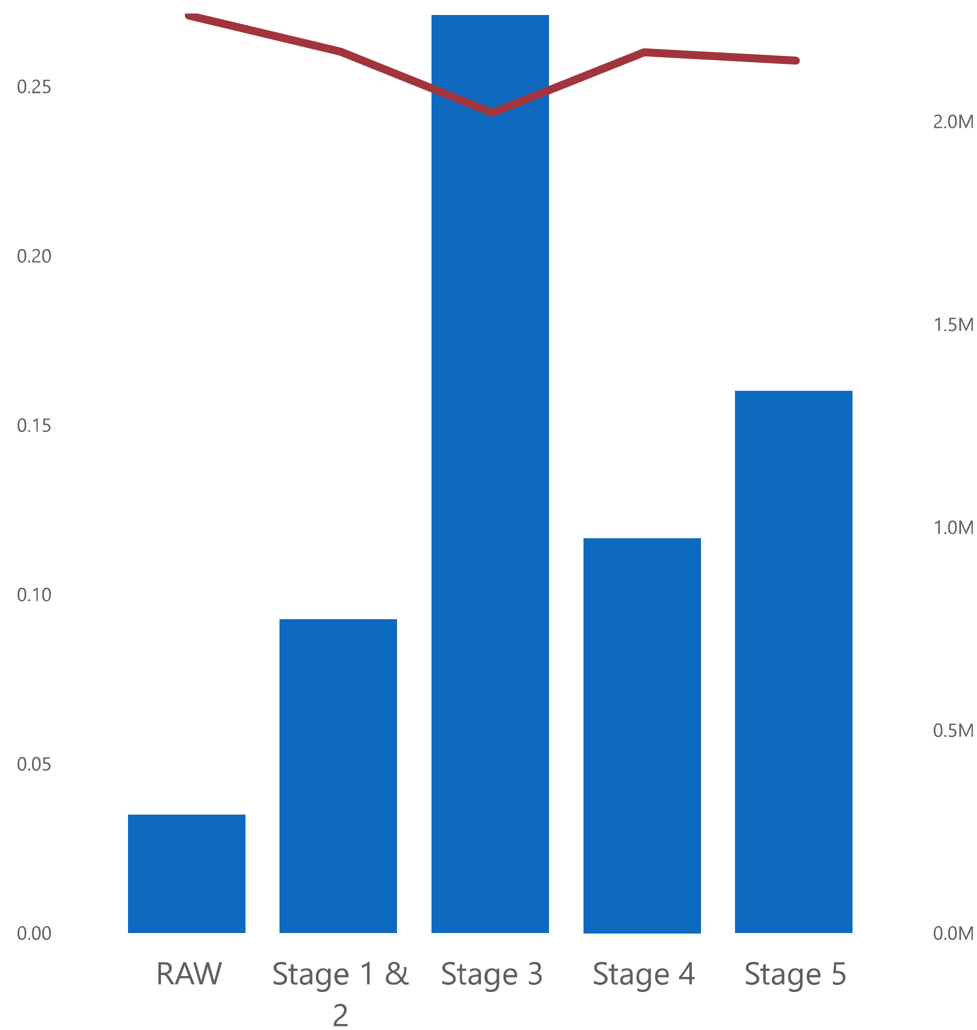
The evaluation metric results **improve** at each step, with a particularly **notable gain at Step 3**, indicating that **noisy values** have a **strongly negative impact** on the model.

RMSE



R^2 and MAE

R^2: line and MAE: column



The real enemy: **Label Noise** or **Feature bias**

This is what **kills your model**. If the Target column is **wrong by more than 20%**, the performance of most algorithms will **plummet**, even **worse** when compared to the baseline

The real enemy: Label Noise or Feature bias

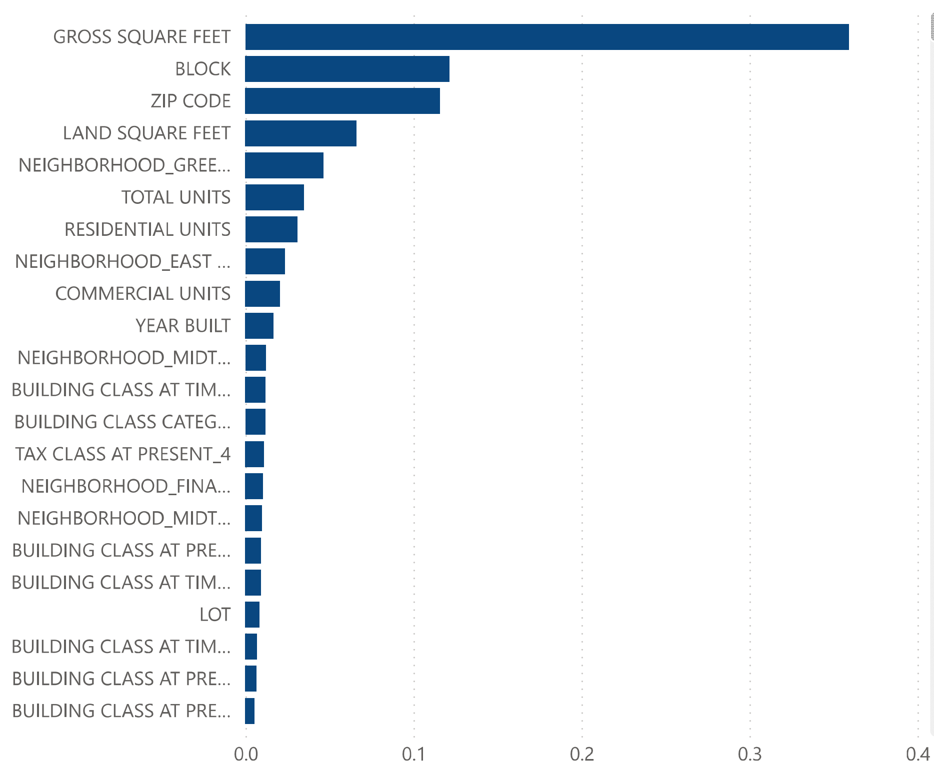
Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., ... & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. Information Systems, 132, 102549.

Data cleaning is **essential**, even if it causes the evaluation metrics to **deteriorate**, because it ensures that the model is trained on **realistic data** rather than on “copied” or **artificially fabricated records** that produce impressive but ultimately **misleading results**.

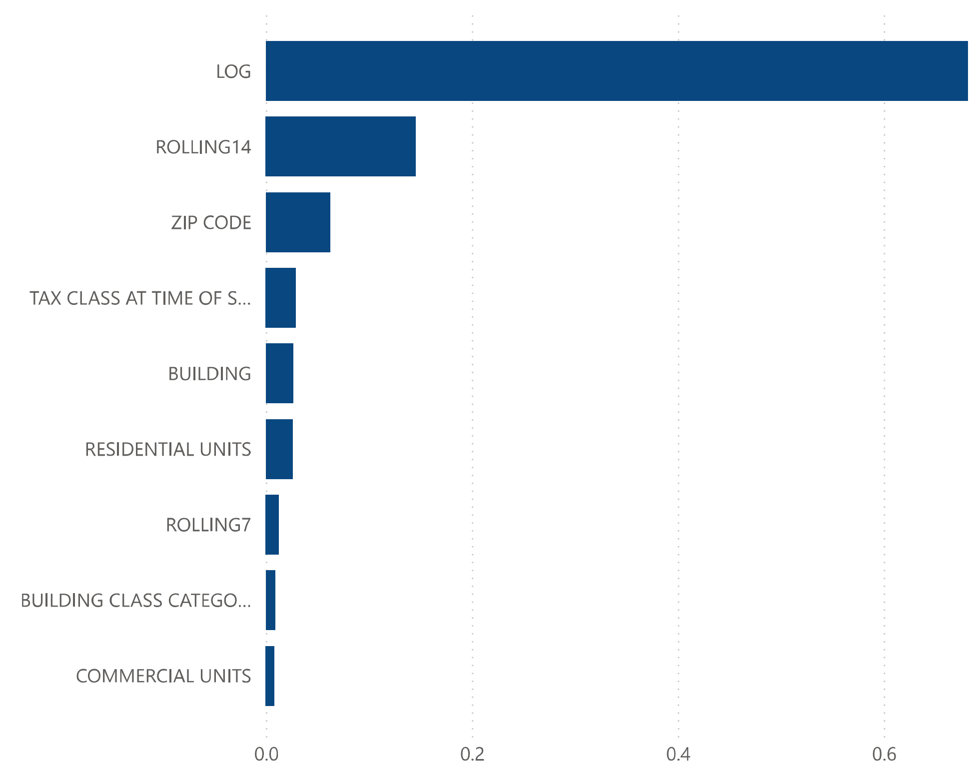
FEATURE IMPORTANCE

Although **LAND SQUARE FEET** does not exhibit a strong linear relationship with price, it still makes a **substantial contribution** to the house price predictions.
GROSS SQUARE FEET provides **the greatest contribution** to the model, which is largely in line with prior expectations.

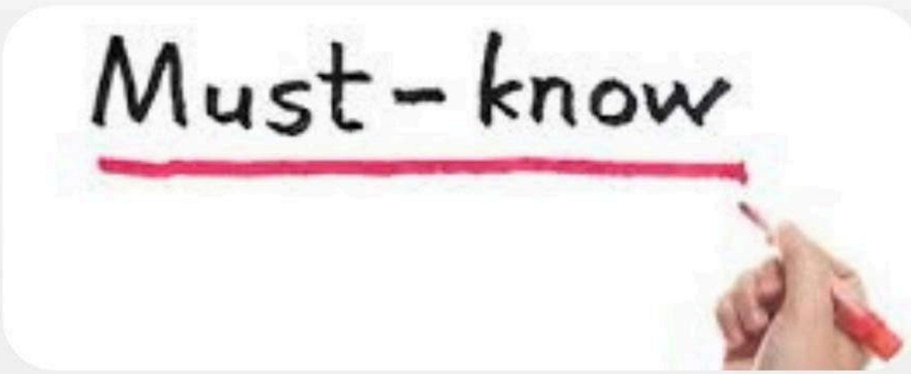
RAW data



Cleaned data



Actions



- **Knowing the difference between clean and raw data**
  Clean data delivers reliable insights; raw data leads to misleading conclusions.
- **Knowing the important of data cleaning**
  Cleaning data ensures accurate predictions and reduces investment risk.
- **Approving clean data priority proposal**
  Prioritizing data cleaning strengthens our investment decisions.