

# House Pricing Prediction

*Alan Turing*



# Agenda

1

**Project Background**

2

**Data Understanding**

3

**Data Cleansing**

4

**Data Modeling**

5

**Model Evaluation**

6

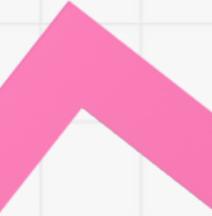
**Model Deployment**

# Project Background

# Background



Real Estate merujuk pada properti berwujud seperti tanah, bangunan, atau struktur lainnya, yang dapat menjadi objek investasi atau sumber pendapatan dalam bisnis melalui sewa atau penjualan properti.



Machine Learning berperan penting di industri real estate dengan memprediksi harga properti berdasarkan faktor-faktor seperti lokasi, ukuran, fasilitas, dan kondisi properti, membantu agen, investor, dan penjual menentukan harga yang tepat.



# Goals

Membangun model machine learning yang dapat memprediksi harga jual rumah di kawasan Ames, Iowa, US.



# Problem Importance

Keputusan Investasi  
yang Tepat

Kebutuhan Pasar yang  
Bervariasi

Peningkatan  
Efisiensi

Peningkatan  
Kepercayaan Customer

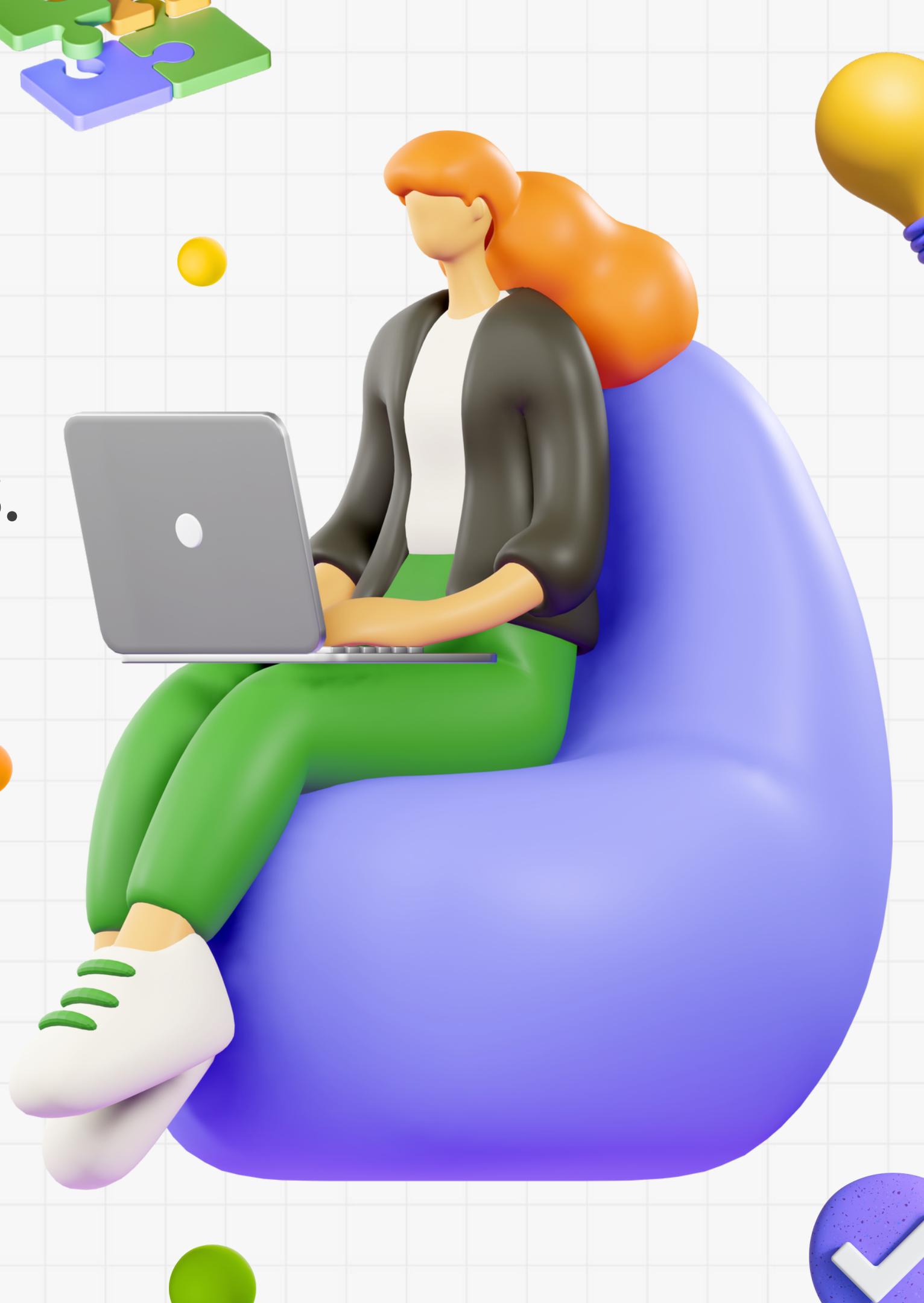
# Data Understanding

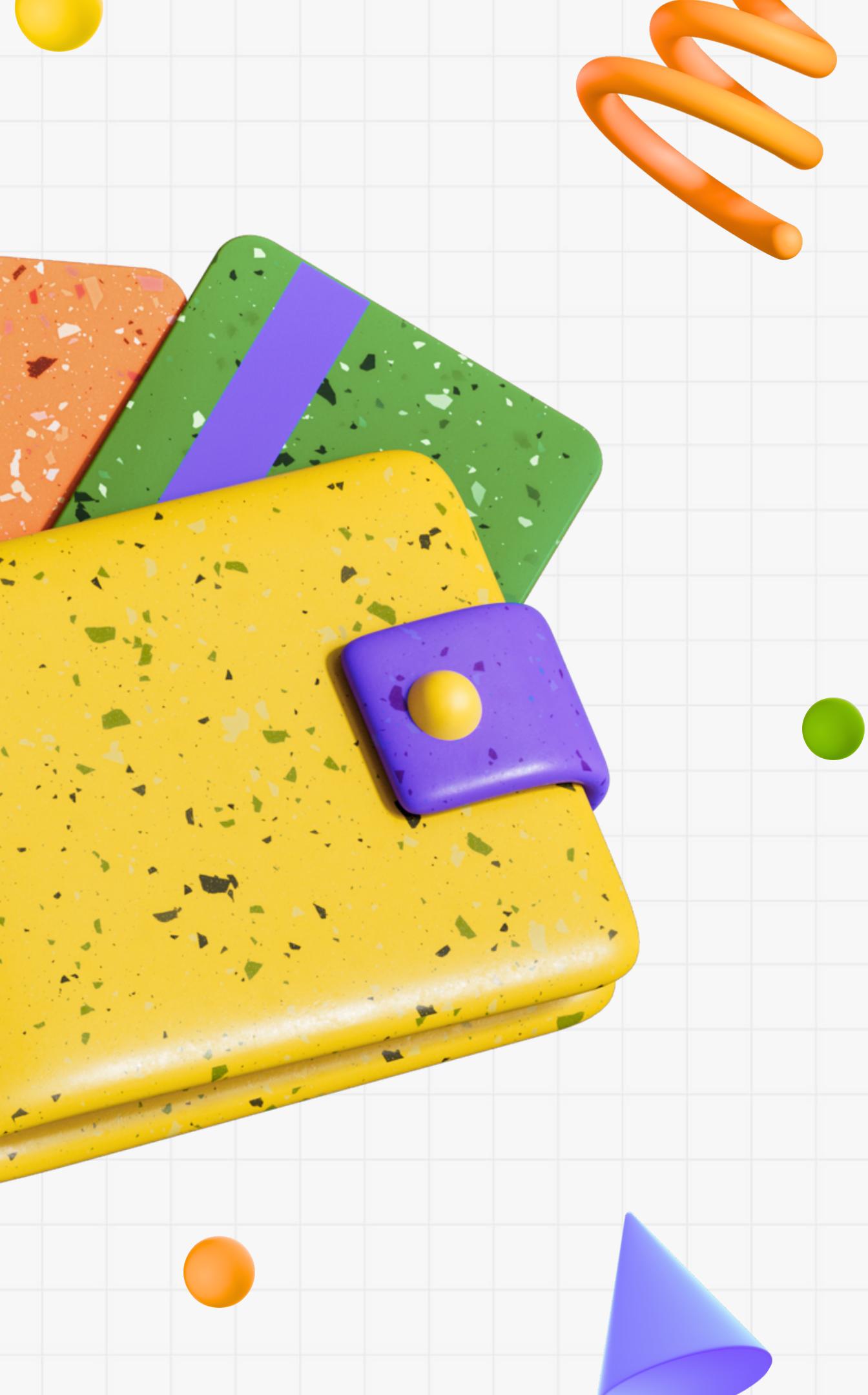
# Dataset

House Pricing kawasan Ames, Lowa, US.

Indonesia AI

- **Train Set:** Terdiri dari 1.460 data dengan 80 fitur + 1 Target
- **Test Set:** Terdiri dari 1.459 data dengan 80 fitur.

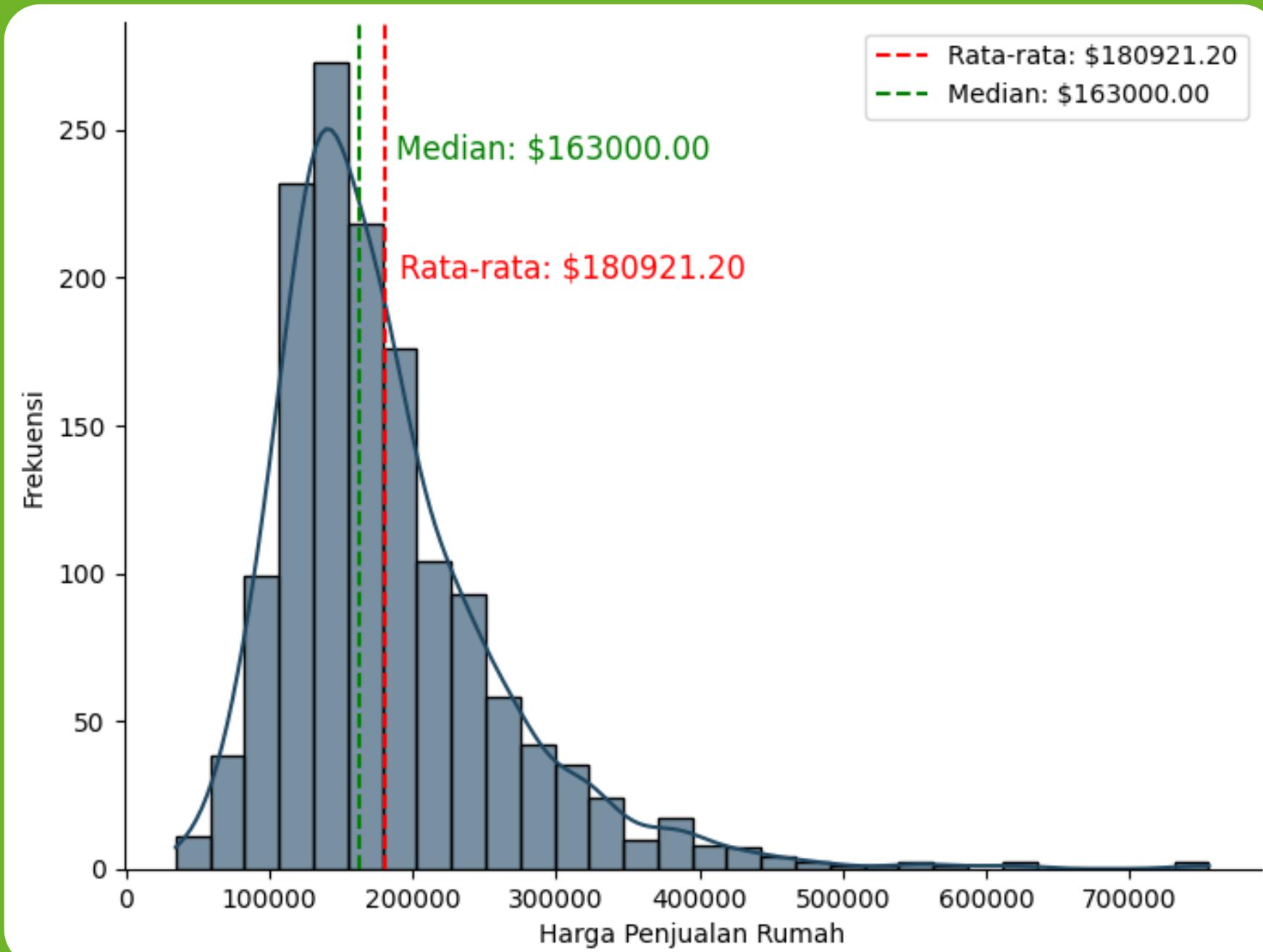




# Business Question?

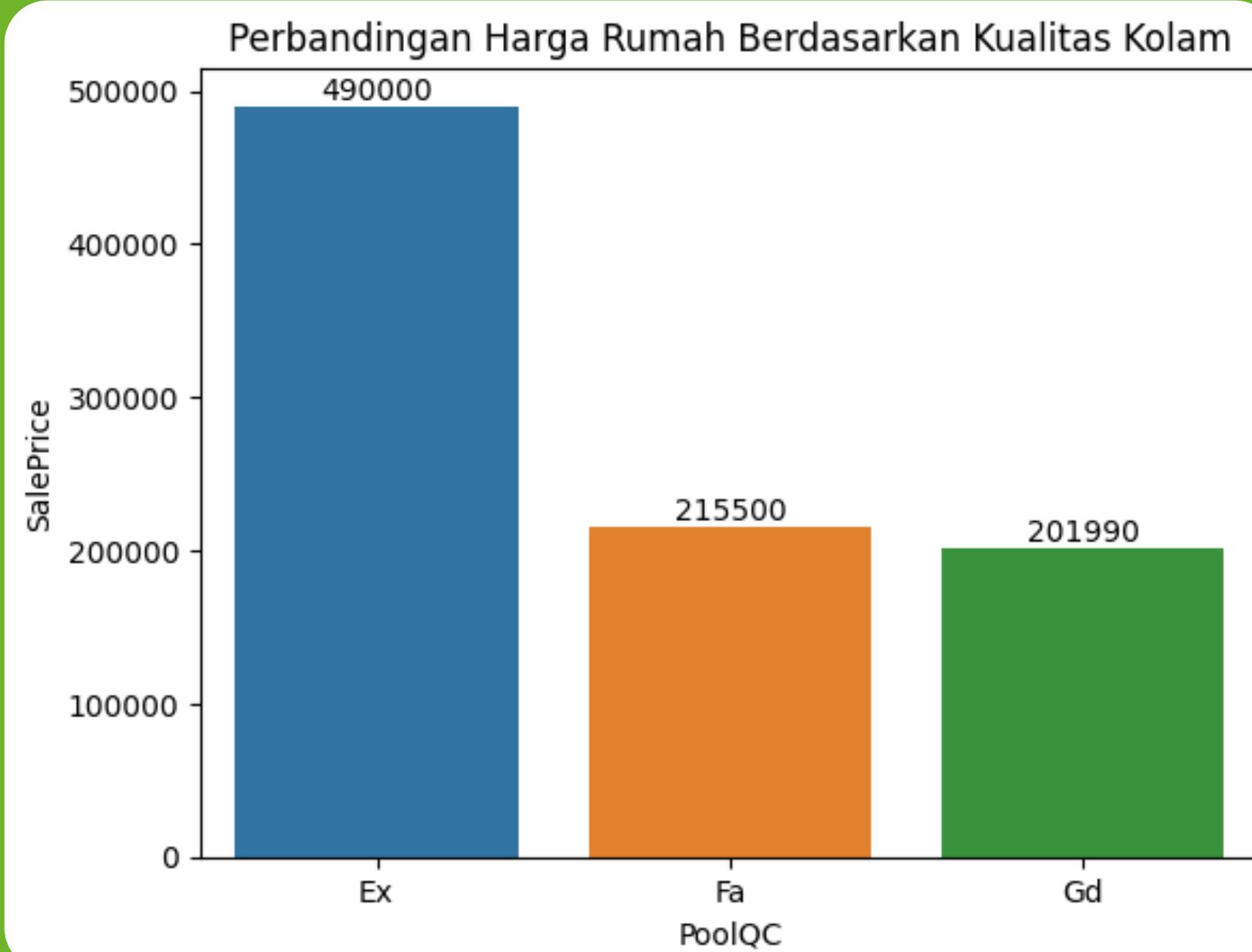
- 1 Bagaimana distribusi harga rumah pada data yang tersedia?
- 2 Bagaimana profiling harga rumah pada data yang tersedia?
- 3 Apa faktor-faktor utama yang mempengaruhi harga penjualan rumah di kawasan tersebut?

# Distribusi Harga Rumah



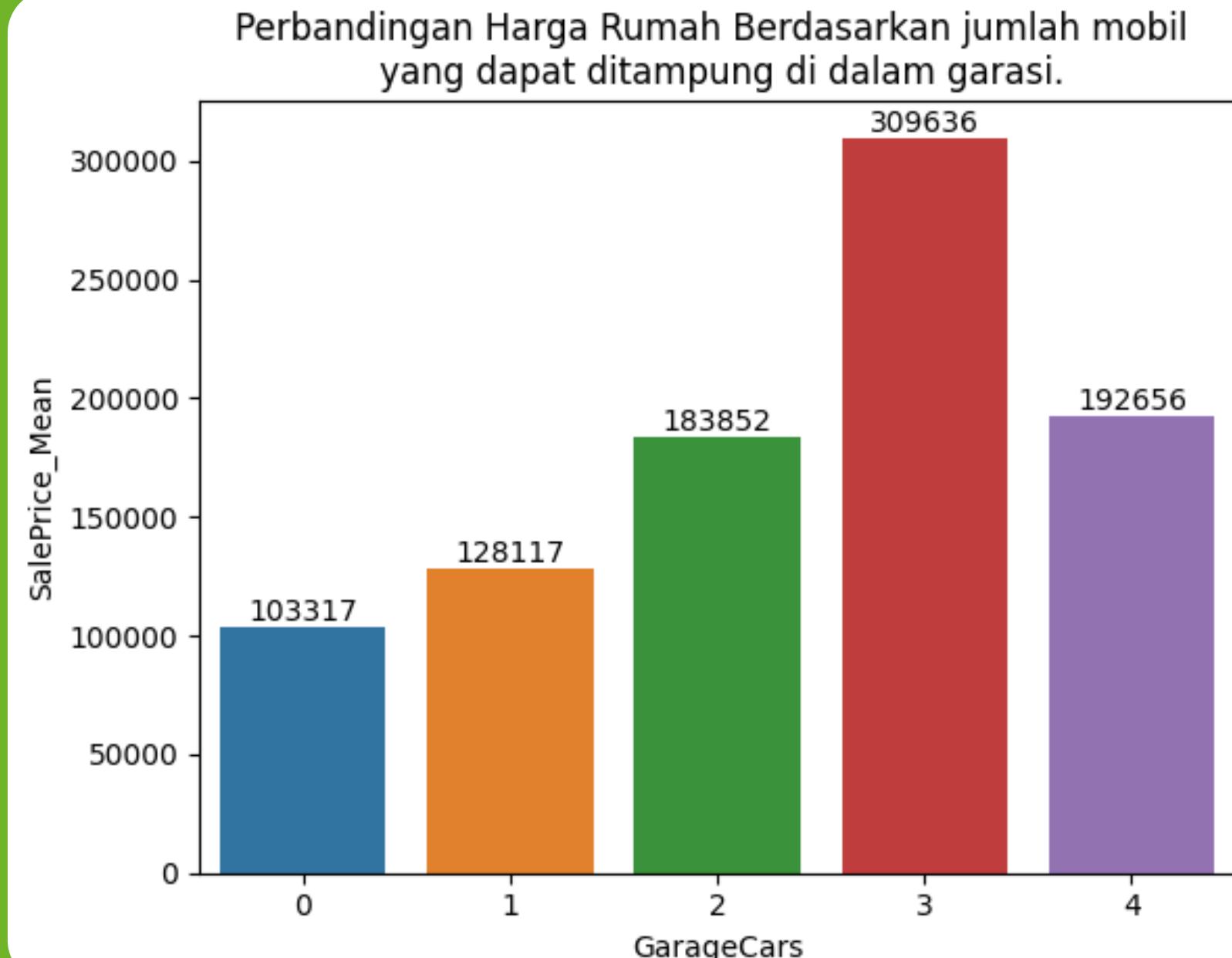
Distribusi harga penjualan rumah pada dataset memiliki rentang yang cukup luas, dari 34.900 hingga 755.000 dolar, dengan rata-rata harga sekitar 180.921 dolar. Dengan nilai median (50th percentile) sebesar 163.000 dolar, distribusi cenderung agak condong ke kiri.

# Harga Rumah dengan Kolam Renang



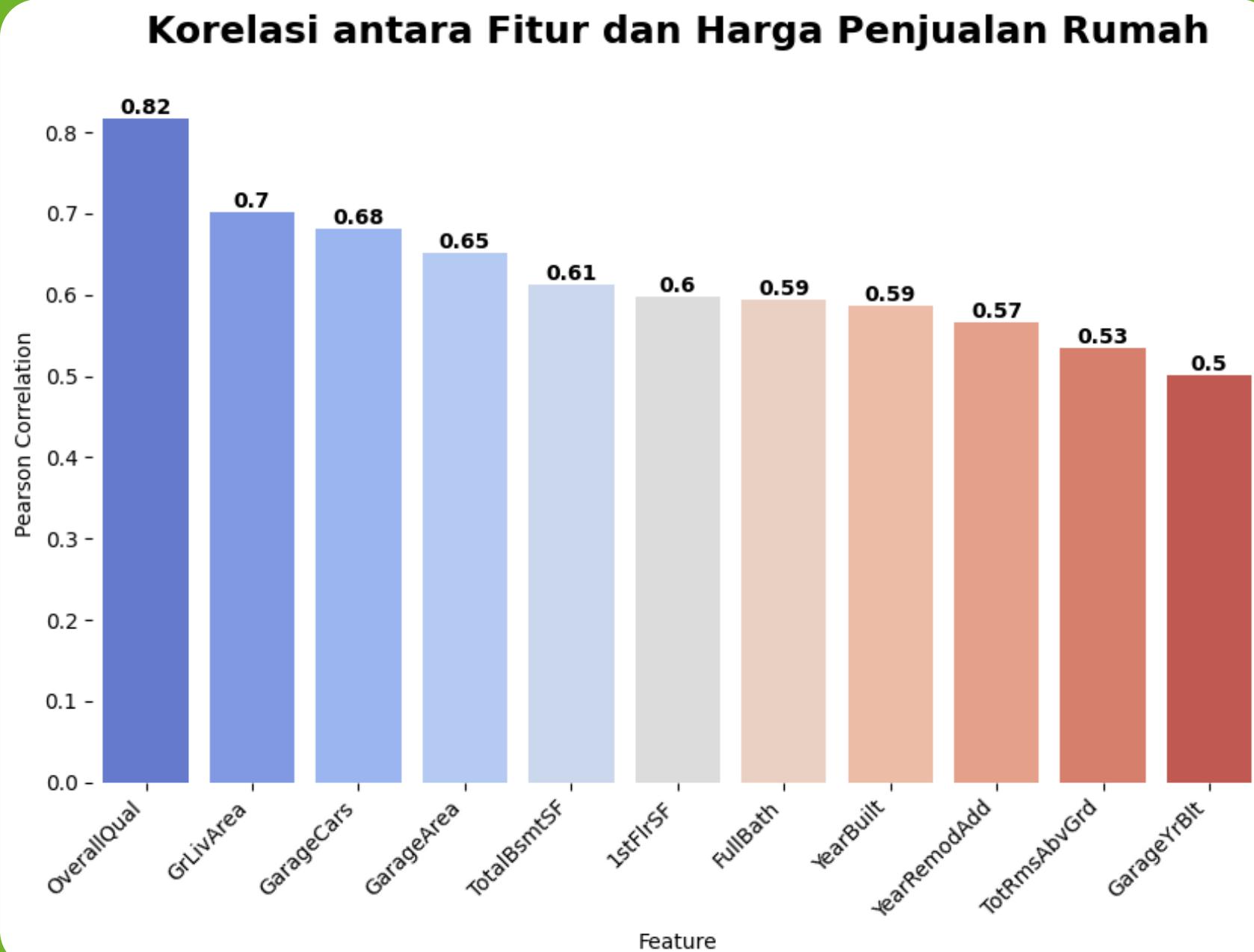
Rumah dengan kolam renang kualitas "Excellent" memiliki harga lebih tinggi karena memberikan nilai tambah yang signifikan.

# Harga Rumah dengan Garasi



Meskipun umumnya harga properti meningkat seiring dengan jumlah tempat parkir garasi yang lebih banyak, terdapat pengecualian pada properti dengan empat tempat parkir garasi (GarageCars = 4).

# Fitur Pengaruh Harga Rumah



Visualisasi di samping merupakan grafik yang menampilkan korelasi pearson fitur terhadap harga jual rumah (Sale Price). Terdapat 11 fitur yang memiliki korelasi di atas 50%.

# Fitur Pengaruh Harga Rumah

## OverallQual

Menggambarkan kualitas keseluruhan rumah secara umum berdasarkan skala penilaian.

## GrLivArea

Menunjukkan luas area tinggal di atas tanah dalam satuan kaki persegi.

## GarageCars

Merepresentasikan jumlah mobil yang dapat ditampung di dalam garasi.

## GarageArea

Menyatakan luas area garasi dalam satuan kaki persegi.

## TotalBsmtSF

Mengindikasikan total luas area ruang bawah tanah dalam satuan kaki persegi.

## 1stFlrSF

Merupakan luas area lantai pertama dalam satuan kaki persegi.

# Fitur Pengaruh Harga Rumah

## FullBath

Menyatakan jumlah kamar mandi penuh di dalam rumah.

## YearBuilt

Menyatakan tahun pembangunan rumah.

## YearRemodAdd

Merupakan tahun terakhir rumah direnovasi.

## TotRmsAbvGrd

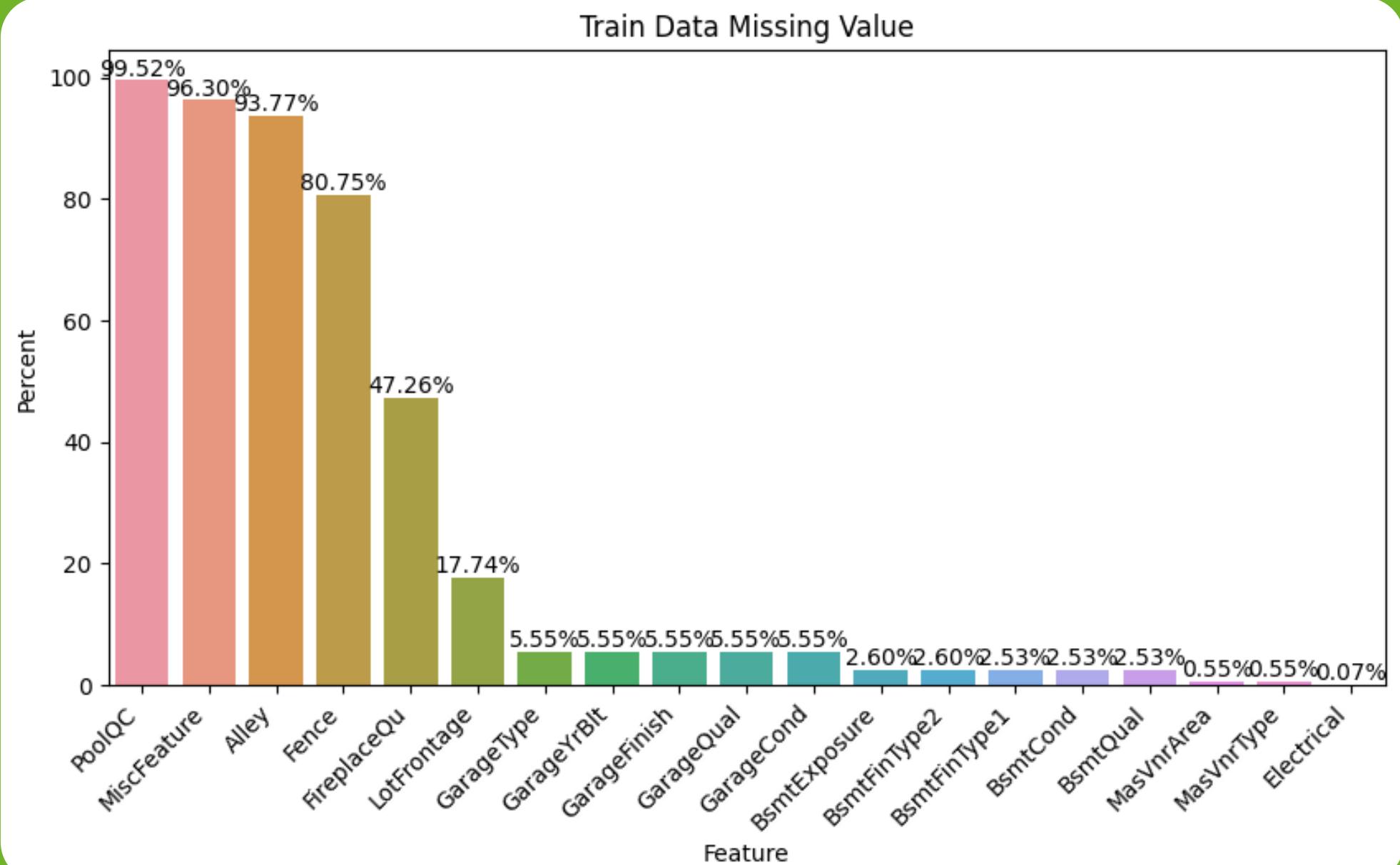
Menggambarkan total jumlah kamar di atas tanah, tanpa termasuk kamar mandi.

## GarageYrBlt

Menunjukkan tahun pembangunan garasi.

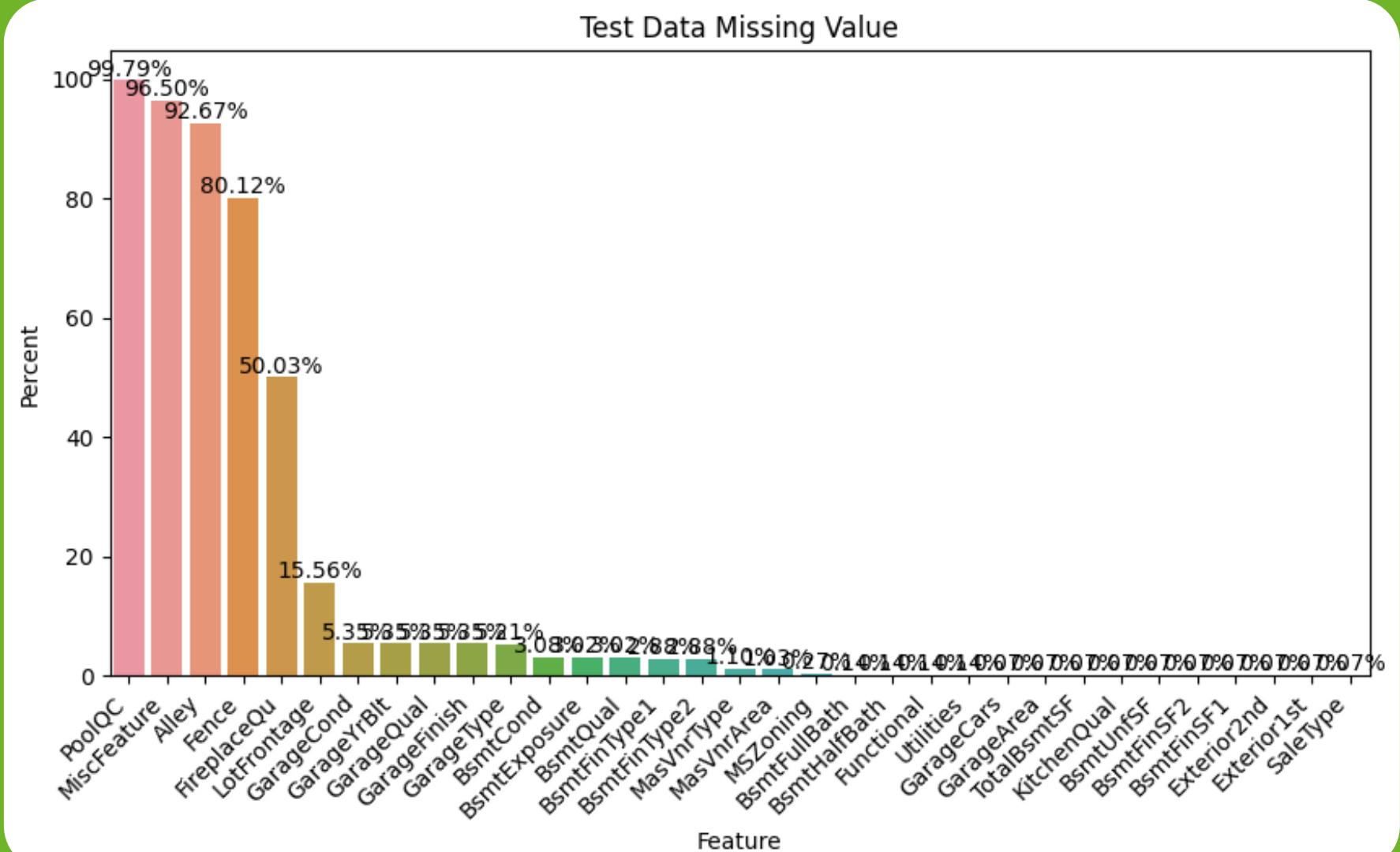
# Data Cleansing

# Missing Value



Terdapat 19 fitur yang memiliki missing value. Terdapat 4 fitur diantaranya yang memiliki missing value di atas 80% yaitu PoolQC, MiscFeature, Alley, dan Fence. Fitur ini akan di drop karena memiliki missing value yang terlalu banyak.

# Missing Value



Dari total 33 fitur dengan nilai yang hilang, 4 fitur dalam dataset train memiliki lebih dari 80% nilai yang hilang, termasuk PoolQC, MiscFeature, Alley, dan Fence. Sementara dalam dataset test, FireplaceQU memiliki lebih dari 50% nilai yang hilang. Oleh karena itu, fitur-fitur ini akan dihapus dari analisis.

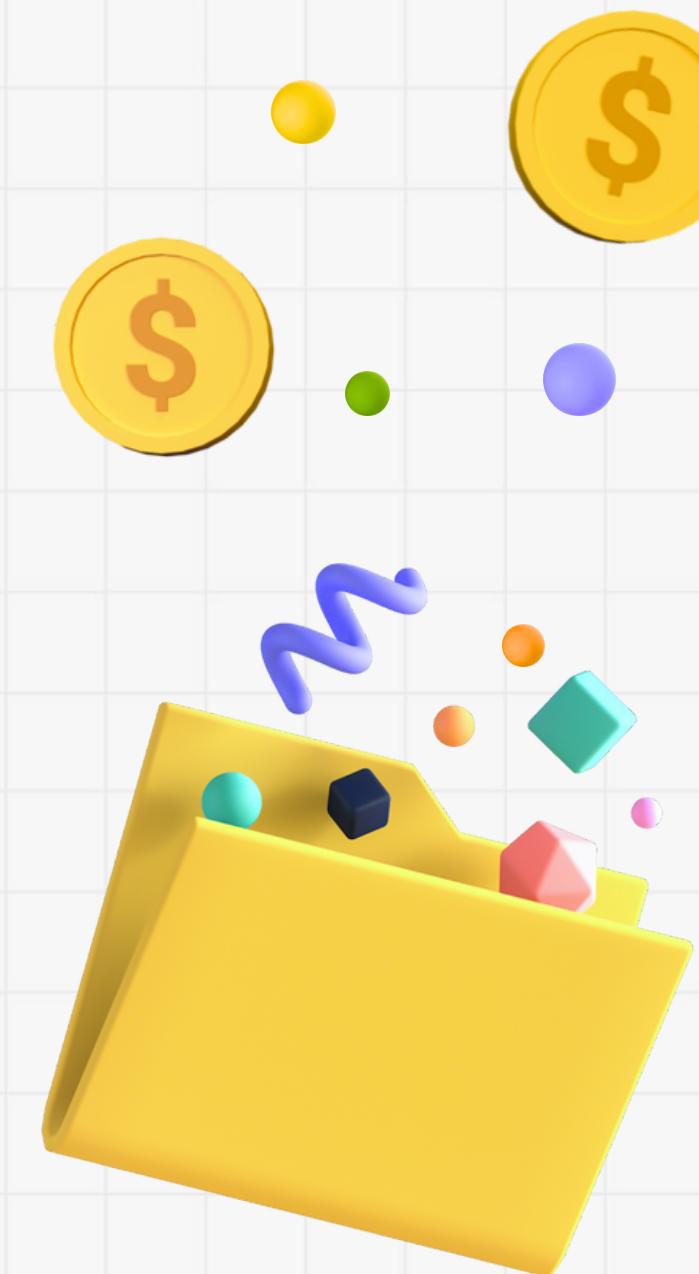
# Handling Missing Value

## Kategorikal

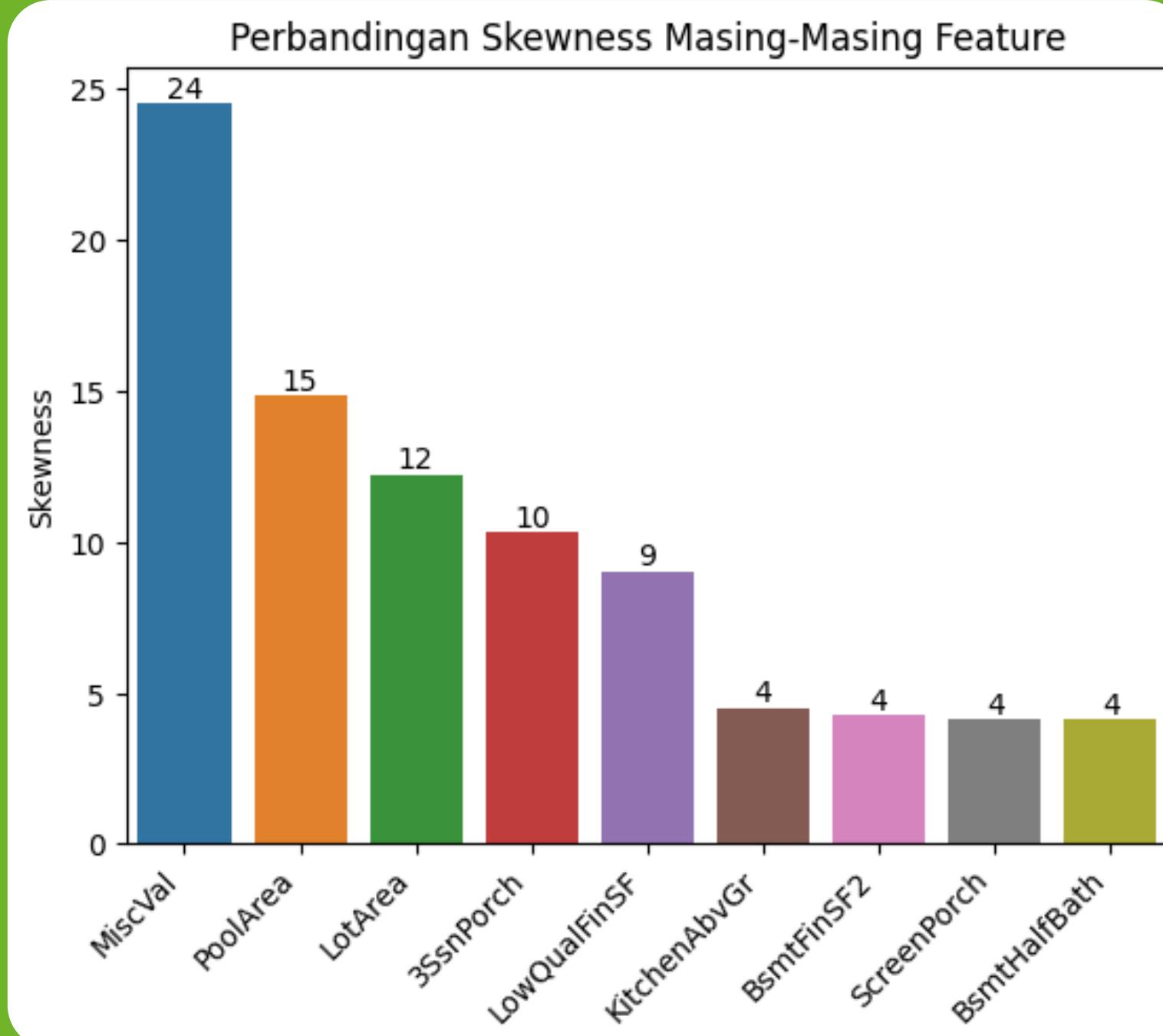
Missing value yang null akan diisi menggunakan nilai modus suatu kategori pada data train

## Numerik

Suatu fitur null akan di check terlebih dahulu, apabila skew akan diisi menggunakan median dan apabila tidak skew akan diisi menggunakan mean

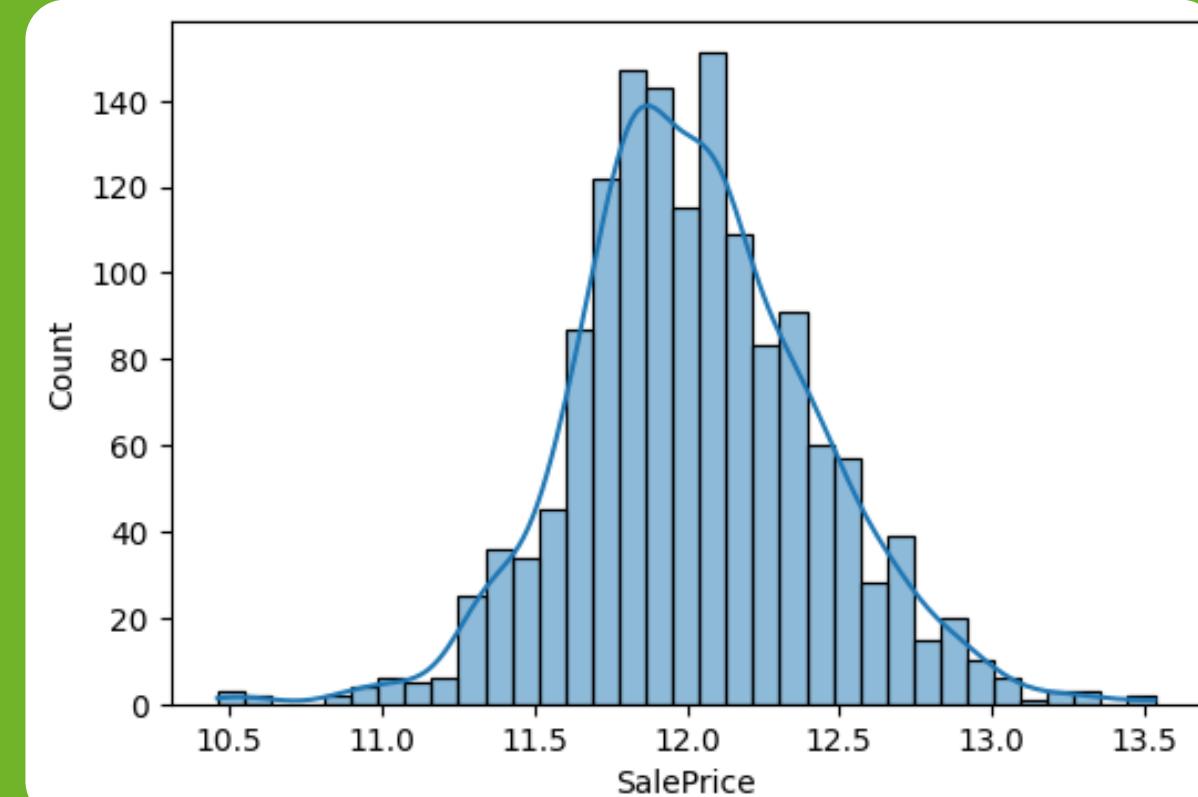
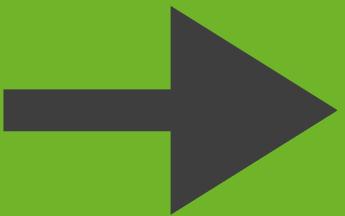
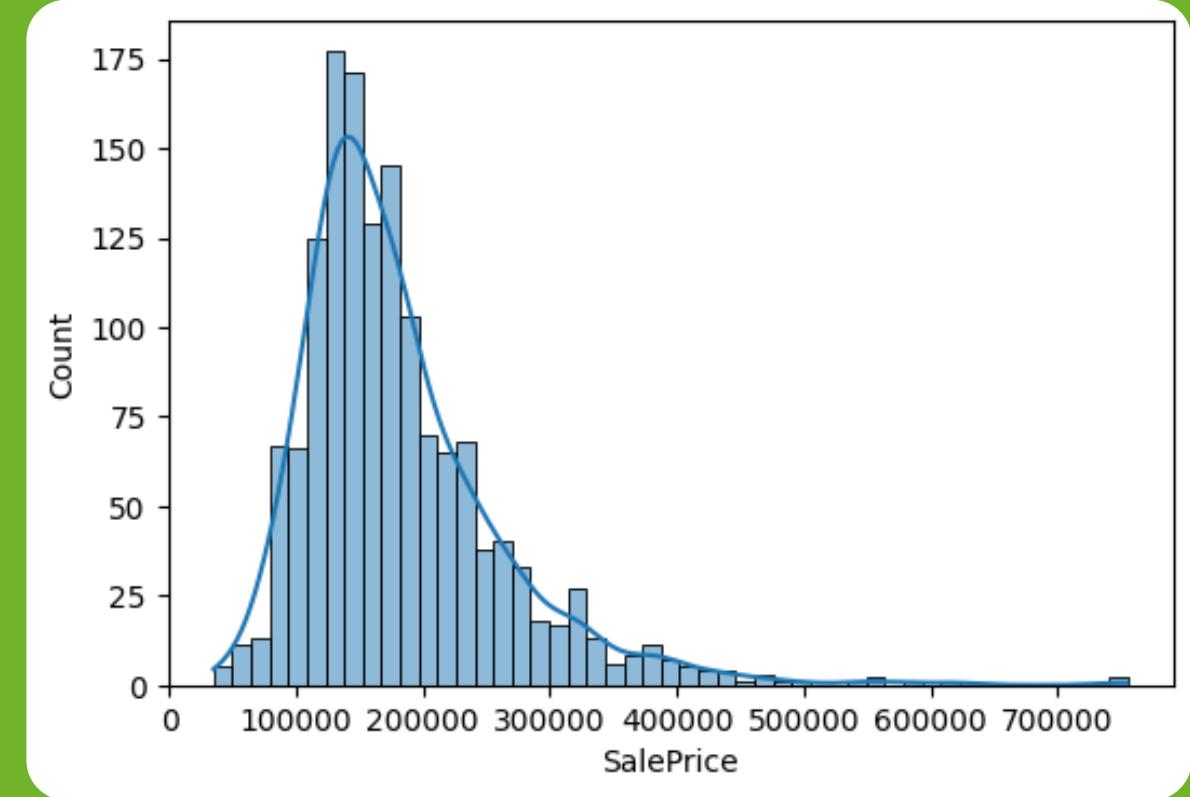


# Skewness



Beberapa fitur seperti MiscVal, PoolArea, dan LotArea menunjukkan kemiringan yang signifikan, menandakan distribusi nilai yang tidak simetris. Untuk mengatasinya, akan dilakukan transformasi logaritma pada fitur-fitur tersebut guna memperoleh distribusi yang lebih simetris.

# Log Transform



Selain itu, log transform juga dilakukan pada target Sale Price. Hasil log transform membuat distribusi Sale Price menjadi lebih normal. Dengan begitu, model regresi dapat melakukan prediksi dengan lebih baik.

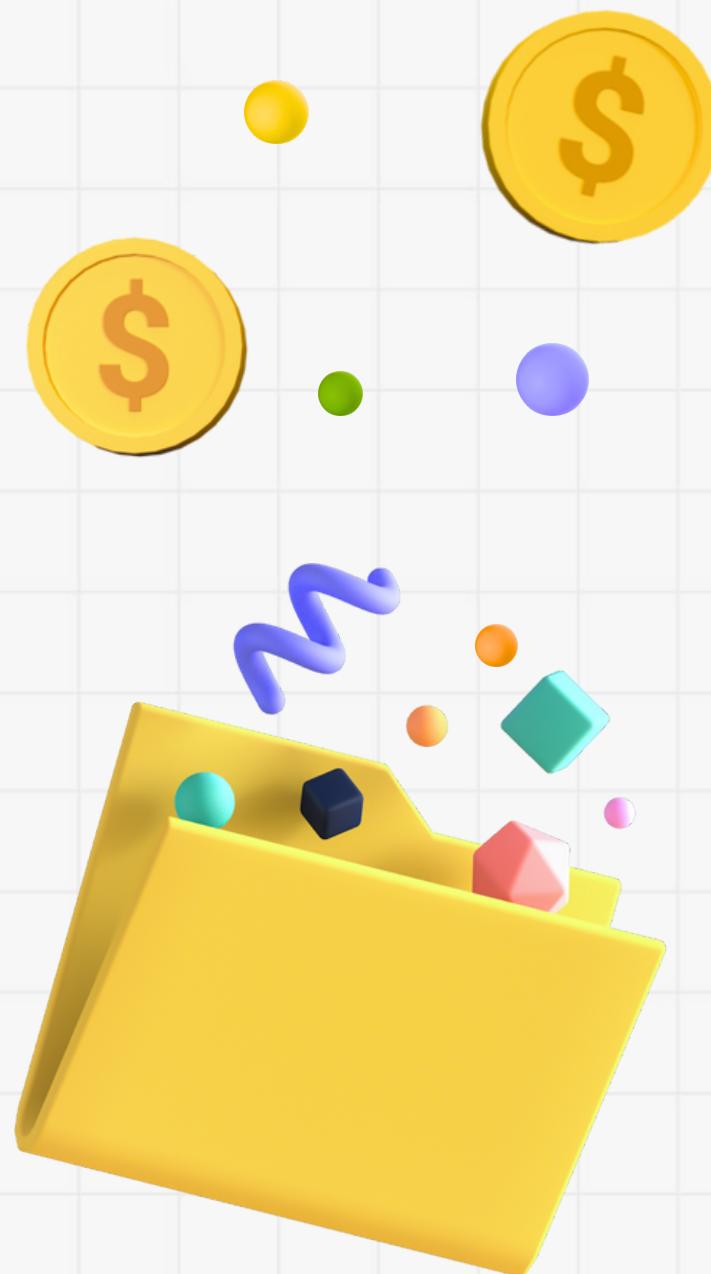
# Feature Selection

## Kategorikal

Jika ada fitur yang memiliki dampak pada Sale Price setelah dianalisis menggunakan Boxplot, maka fitur tersebut akan tetap dipertahankan. Namun, jika tidak ada dampak yang terlihat, fitur tersebut akan dihapus.

## Numerik

Setiap fitur numerik akan diperiksa korelasinya menggunakan korelasi Pearson dan statistik p value. Fitur dengan korelasi Pearson rendah dan p value di bawah 0,05 menandakan tidak signifikan, dan akan dihapus.



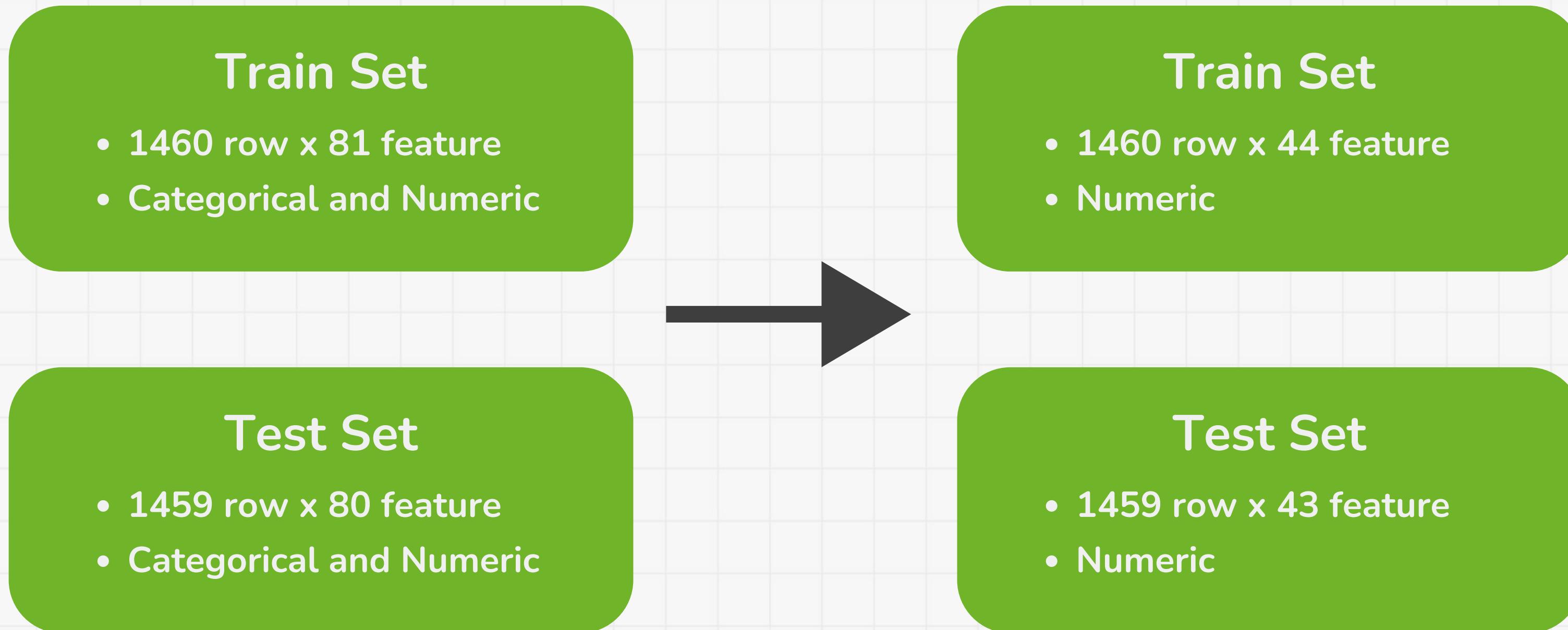
# Feature Engineering

- 1 **ExterQual (Ordinal)**
- 2 **BsmtQual (Ordinal)**
- 3 **KitchenQual (Ordinal)**
- 4 **Neighborhood (Nominal)**
- 5 **MasVnrType (Nominal)**

Setelah melalui feature selection tersisa 5 fitur kategorikal yang harus diubah menjadi numerik.

- Label Encoding akan dilakukan pada fitur yang memiliki data ordinal
- One Hot Encoding akan dilakukan pada fitur yang memiliki data nominal

# Preprocessing Result



# Data Training

# Method

- Scale data menggunakan Standard Scaler
- HPO Process
  - Penggunaan BayesSearchCV
  - Menambahkan Early Stopping untuk mempercepat proses
  - Scoring menggunakan negative MSE
- Menguji Data Test dengan parameter terbaik hasil HPO
- Tampilkan Metrik Evaluasi

# Model Evaluation

# Model Comparison

*Before HPO*

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
SGDRegressor	1.05	0.93	0.11	0.02
GradientBoostingRegressor	1.05	0.93	0.11	0.81
LinearRegression	1.05	0.93	0.11	0.03
Ridge	1.05	0.93	0.11	0.02

# HPO

## *Optimal Parameter*

### GradientBoostingRegressor:

- Learning Rate: 0.054
- Max Depth: 4
- Min Samples Split: 59
- Number of Estimators: 432

### SGDRegressor:

- Alpha: 0.002
- L1 Ratio: 0.172
- Learning Rate: 'invscaling'
- Max Iterations: 822

### Ridge Regression:

- Alpha: 3.252

# Evaluation Metrics

## Glosarium

### RMSE

- RMSE mengukur besarnya rata-rata kesalahan antara nilai prediksi dan nilai aktual.
- Nilai RMSE yang lebih rendah menunjukkan kinerja model yang lebih baik.

### R-kuadrat (R<sup>2</sup>)

- R-squared mengukur proporsi varians dalam variabel target yang dijelaskan oleh model.
- Nilai R-squared yang lebih tinggi (mendekati 1) mengindikasikan kecocokan yang lebih baik antara model dengan data.

### Mean Absolute Error(MAE):

- MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual.
- Nilai MAE yang lebih rendah mengindikasikan akurasi model yang lebih baik.

### Median Absolute Error (MedAE)

- MedAE mirip dengan MAE namun menggunakan median dan bukan mean.
- MedAE kurang sensitif terhadap outlier dalam data.

### Explained Variance Score

- Explained Variance Score mengukur seberapa baik model menangkap varians dalam variabel target.
- Nilai yang lebih tinggi (mendekati 1) menunjukkan kinerja yang lebih baik.

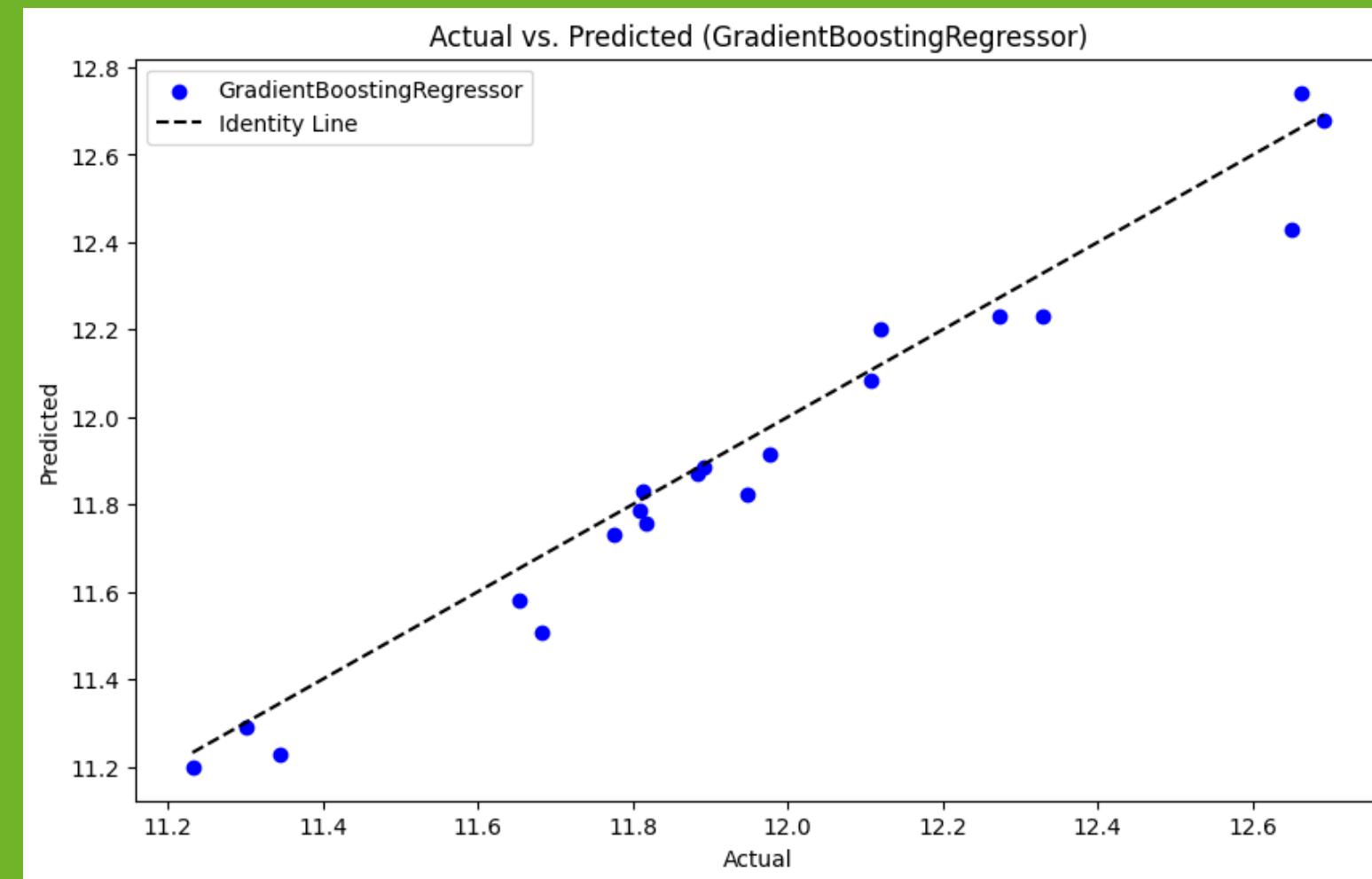
# Model Comparison

After HPO

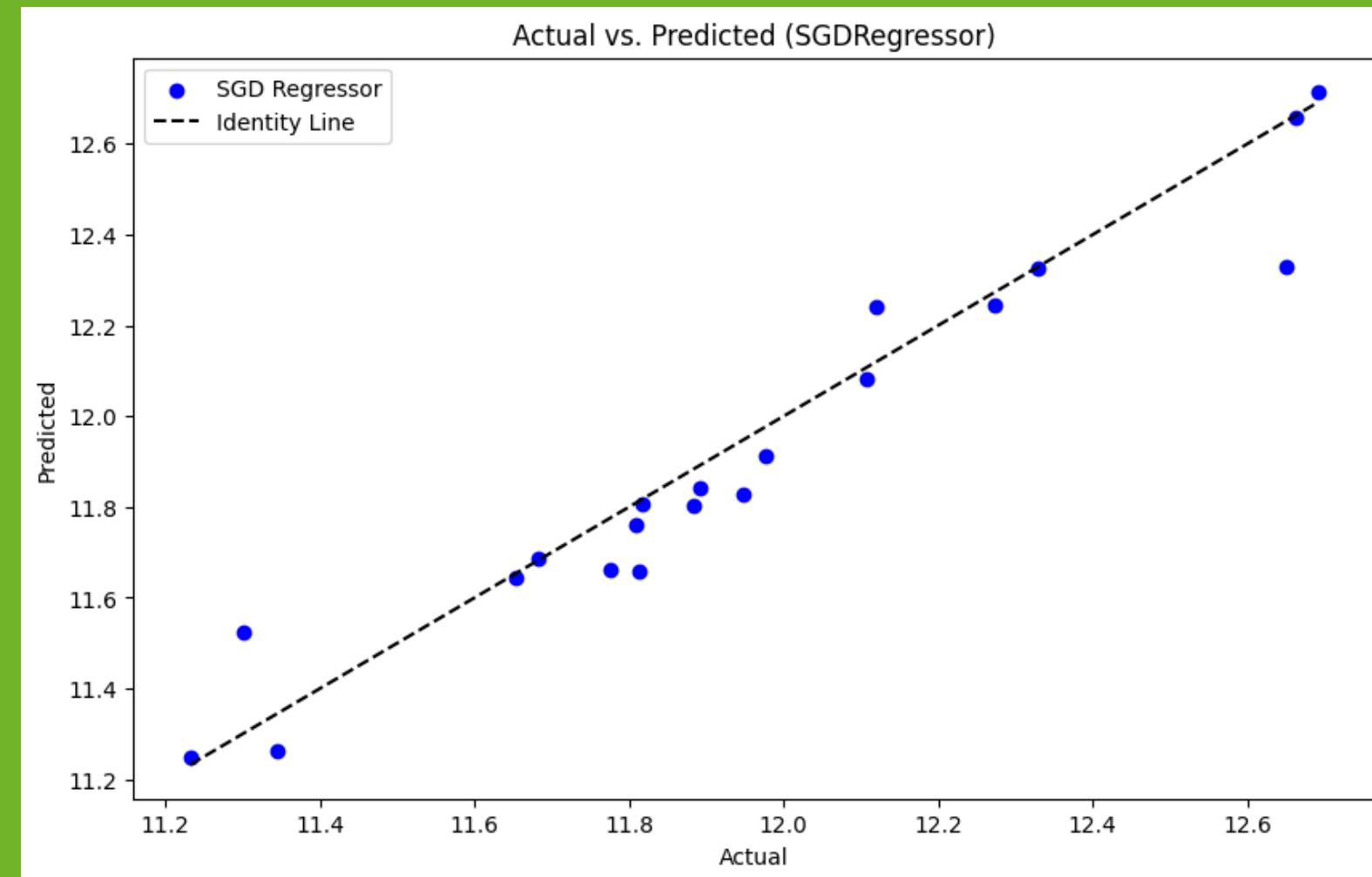
Model	RMSE	R2	MAE	Explained Variance Score	MedAE
SGDRegressor	0.109	0.929	0.074	0.936	0.050
GradientBoostingRegressor	0.086	0.955	0.066	0.969	0.052
LinearRegression	0.106	0.933	0.072	0.939	0.053
Ridge	0.107	0.932	0.072	0.938	0.052

# Actual vs Predicted Plot

GradienBoost

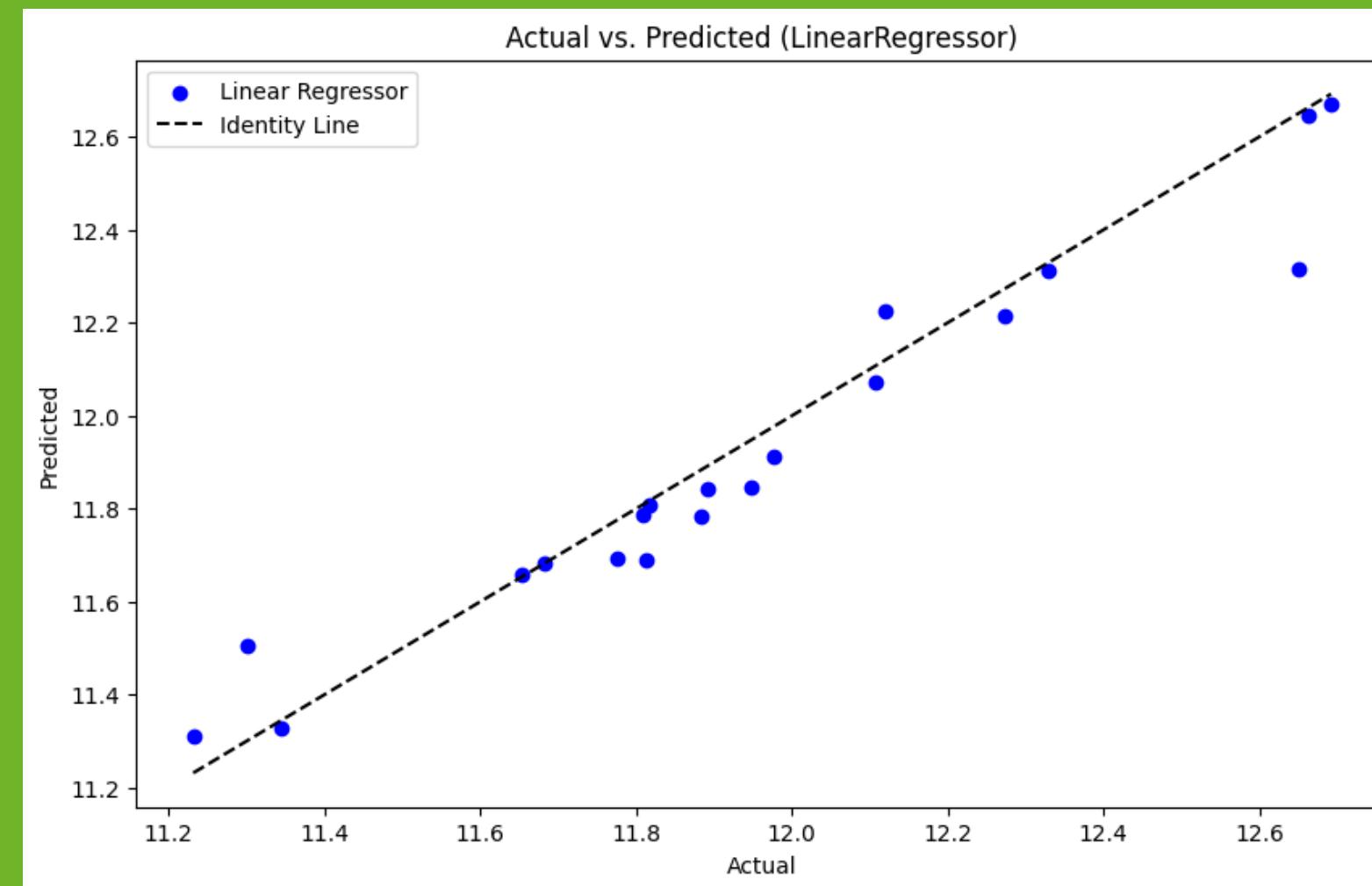


SGD

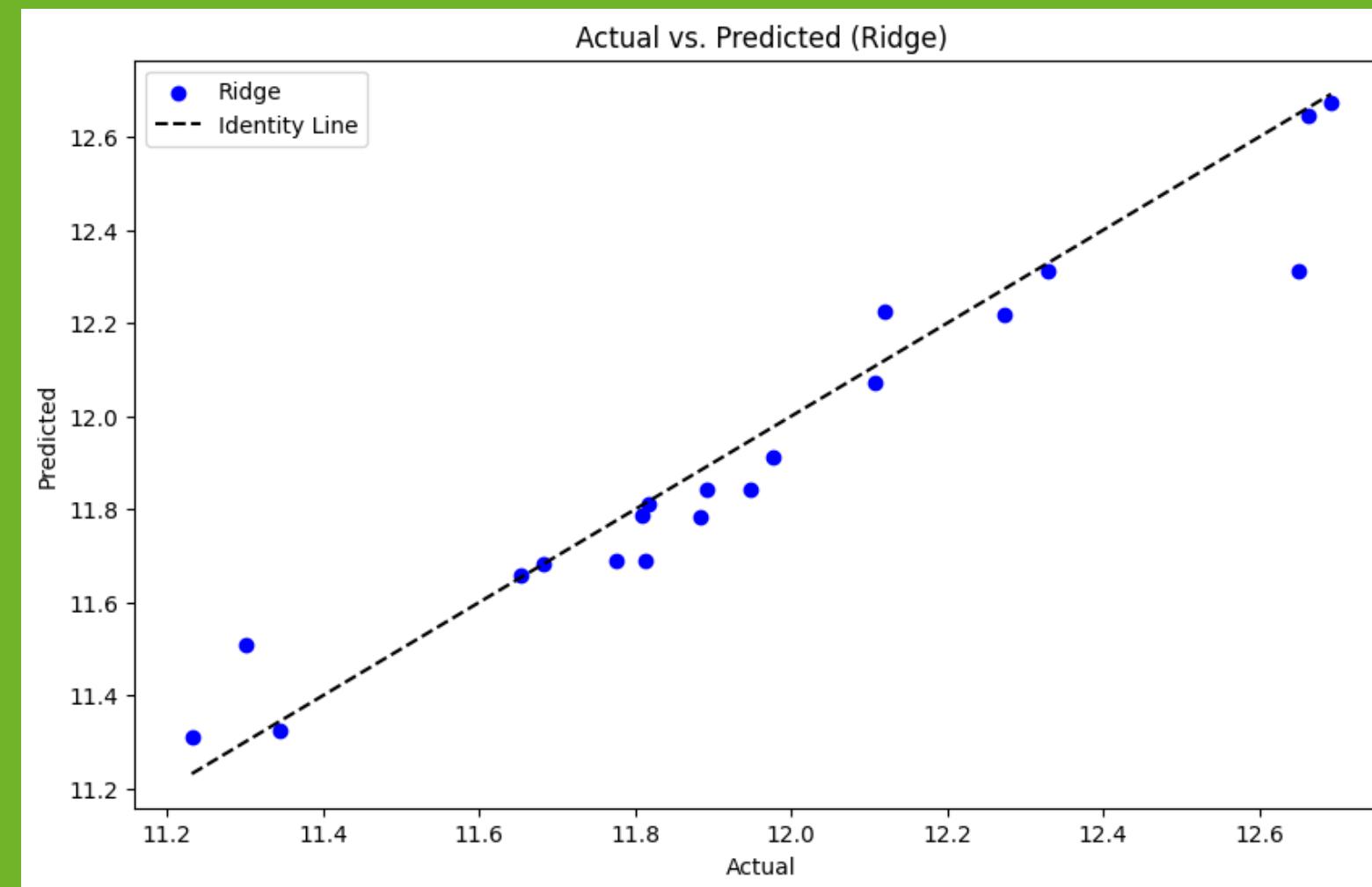


# Actual vs Predicted Plot

LinearRegressor



Ridge



# Model Deployment

# Model Deployment



Pada tahap deployment, kami mengekspor encoder, scaler, dan model machine learning. Langkah ini penting untuk menjalankan model dengan efisien dan terintegrasi dalam sistem.

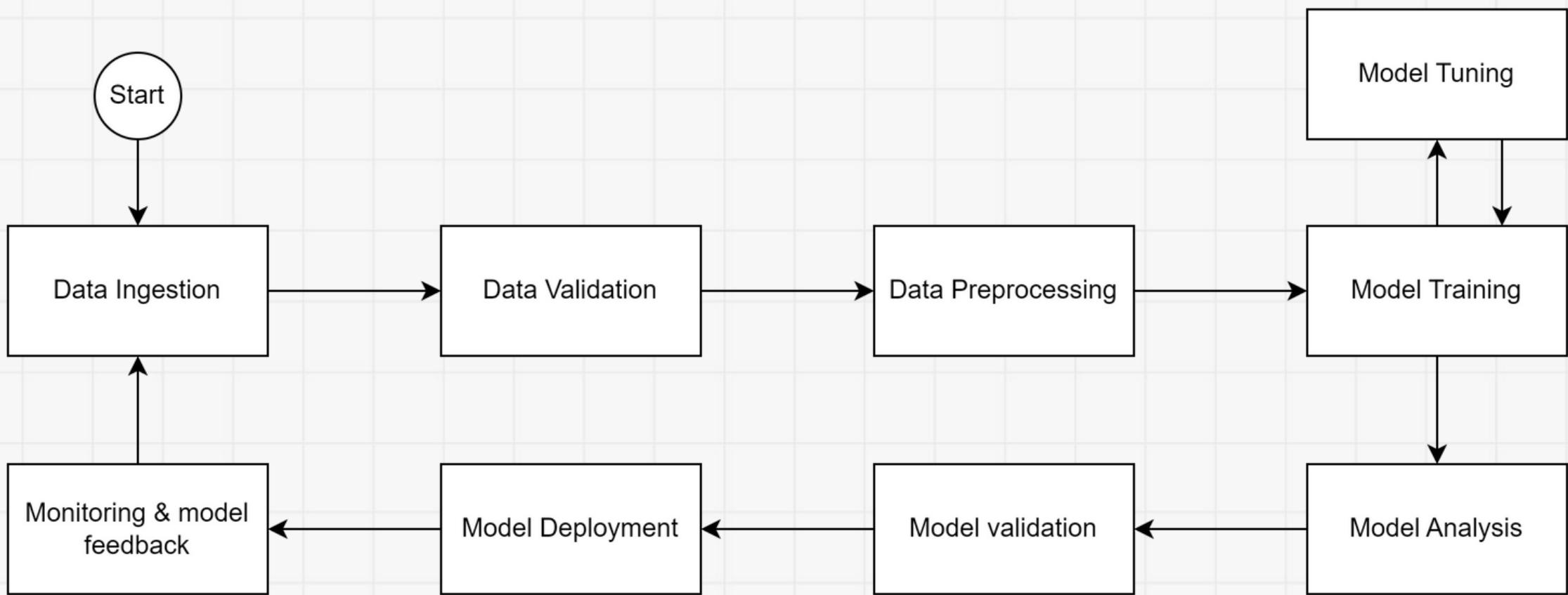


Setelah berhasil diekspor, model tersebut kami jalankan di platform Streamlit. Streamlit adalah sebuah platform open-source yang memungkinkan kami untuk membuat aplikasi web interaktif dengan cepat dan mudah.

## Link Deployment



# Deployment & Maintenance Plan



Source: Dicoding (MLOps) - Pipeline