Final Project Data Science

Prediksi Harga Perumahan di Washington

By : Ragil Setyo Utomo

Deskripsi singkat

Nama : Ragil Setyo Utomo

Minat : Pengolahan data, visualisasi , serta prediksi dengan machine learning untuk mendukung keputusan bisnis.

Latar belakang : Data science, terbiasa dengan data dan mengubahnya menjadi insight yang dipahami.

Tujuan : Membangun karier yang berdampak positif di perusahaan

Education dan Pengalaman

Saya lulusan S-1 Brawijaya teknik Informatika dan sudah mengikuti bootcamp dibimbing untuk data science.

Saya belum mempunyai pengalaman kerja sebagai data science, tetapi pada saat bootcamp, saya sudah mengerjakan proyek untuk memberikan insight dan melakukan prediksi dengan machine learning.

Dibimbing

Overview Project Sebelumnya

Tujuan proyek untuk Mengetahui Churn rate dari dataset. Dengan memakai Random Forest, didapatkan hasil :

- Recall = 1.0
- Precision = 0.67
- AUC = 0.9

Overview Project Sebelumnya

Tujuan proyek untuk mengetahui prediksi harga perumahan di boston dengan model linear regression dengan memakai ridge

```
RMSE for testing data is 5.081102260579794
MAE for testing data is 3.2789357536706483
MAPE for testing data is 0.17613319441071806
R² Score : 0.6479
```

Project Background

Pasar properti di negara bagian Washington seperti seattle, Bellevue mengalami pertumbuhan yang dinamis dalam beberapa tahun akhir.

Agen Properti ingin menentukan harga jual rumah yang tepat dengan menggunakan berbagai faktor seperti lokasi, luas bangunan, serta kondisi rumah. Penentuan harga yang tidak akurat dapat menyebabkan kerugian dan menurunkan minat pembeli.

Melalui projek ini, agen properti ingin membangun model prediksi harga rumah berbasis machine learning dengan memanfaatkan data histori penjualan rumah.

Dengan model prediksi ini, diharapkan agensi dapat memberikan rekomendasi harga properti yang lebih akurat, memahami faktor utama yang memengaruhi nilai rumah, serta meningkatkan kepercayaan pelanggan dan efisiensi bisnis.

Problem Statement

1

Adanya demand dari banyak pelanggan untuk membeli rumah di USA tepatnya di Washington

2

Bagaimana cara memprediksi harga rumah secara akurat dengan data penjualan historis?

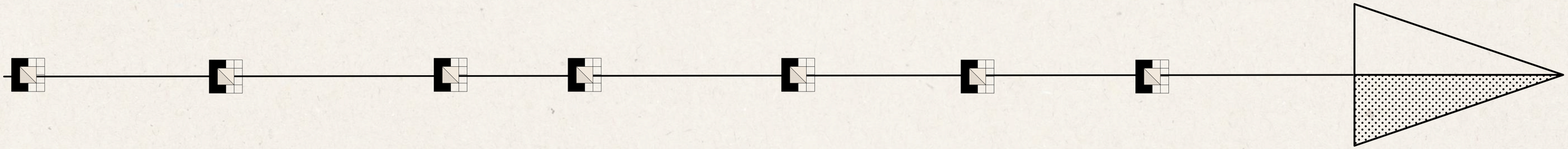
3

Bagaimana model yang dibuat dapat membantu agensi dalam meningkatkan keakuratan penilaian harga rumah? dibanding dengan cara tradisional atau intuisi pasar

Objective & Bussiness Question

Objective	1. Membangun prediksi harga rumah di washington menggunakan algoritma machine learning dengan data historis. 2. Memprediksi fitur yang paling mempengaruhi terhadap harga rumah
Bussiness Question	1. Faktor utama apa yang paling mempengaruhi harga rumah? 2. Bagaimana pengaruh lokasi (city, statezip) terhadap harga rumah? Apakah rumah di area tertentu secara konsisten memiliki harga lebih tinggi? 3. Dapatkah model machine learning memperkirakan harga rumah baru dengan tingkat kesalahan (MAE/RMSE) yang rendah?

Alur Pemodelan



1
Identifikasi Variabel

2
Data Cleaning

3
EDA

4
Data Preprocessing

5
Modeling

6
Evaluasi

7
Kesimpulan & Saran

Identifikasi Variabel

Kolom Deskripsi Tipe:

- date = Tanggal transaksi rumah dijual
- Price = Harga jual rumah
- bedrooms = Jumlah kamar tidur
- bathrooms = Jumlah kamar mandi
- sqft_living = Luas area tempat tinggal (dalam square feet)
- sqft_lot = Luas keseluruhan tanah
- floors = Jumlah lantai rumah
- waterfront = Apakah rumah di tepi air (lake/sea view)
- view = Kualitas pemandangan (semakin tinggi, semakin bagus)
- condition = Kondisi fisik rumah (1 = buruk, 5 = sangat baik)
- sqft_above = Luas bagian atas tanah (tidak termasuk basement)
- sqft_basement = Luas basement (jika ada)
- yr_built = Tahun rumah dibangun
- yr_renovated = Tahun terakhir direnovasi (0 = belum pernah)
- street = Alamat jalan lengkap
- city = Nama kota / wilayah
- statezip = Kode wilayah (misal: WA 98133)
- country = Negara

Jumlah Baris : 4600

Jumlah kolom : 18

Data Cleaning

Pengecekan Missing Value : Tidak ada Missing Value

Pengecekan Duplikat data : Tidak ada duplikat pada dataset

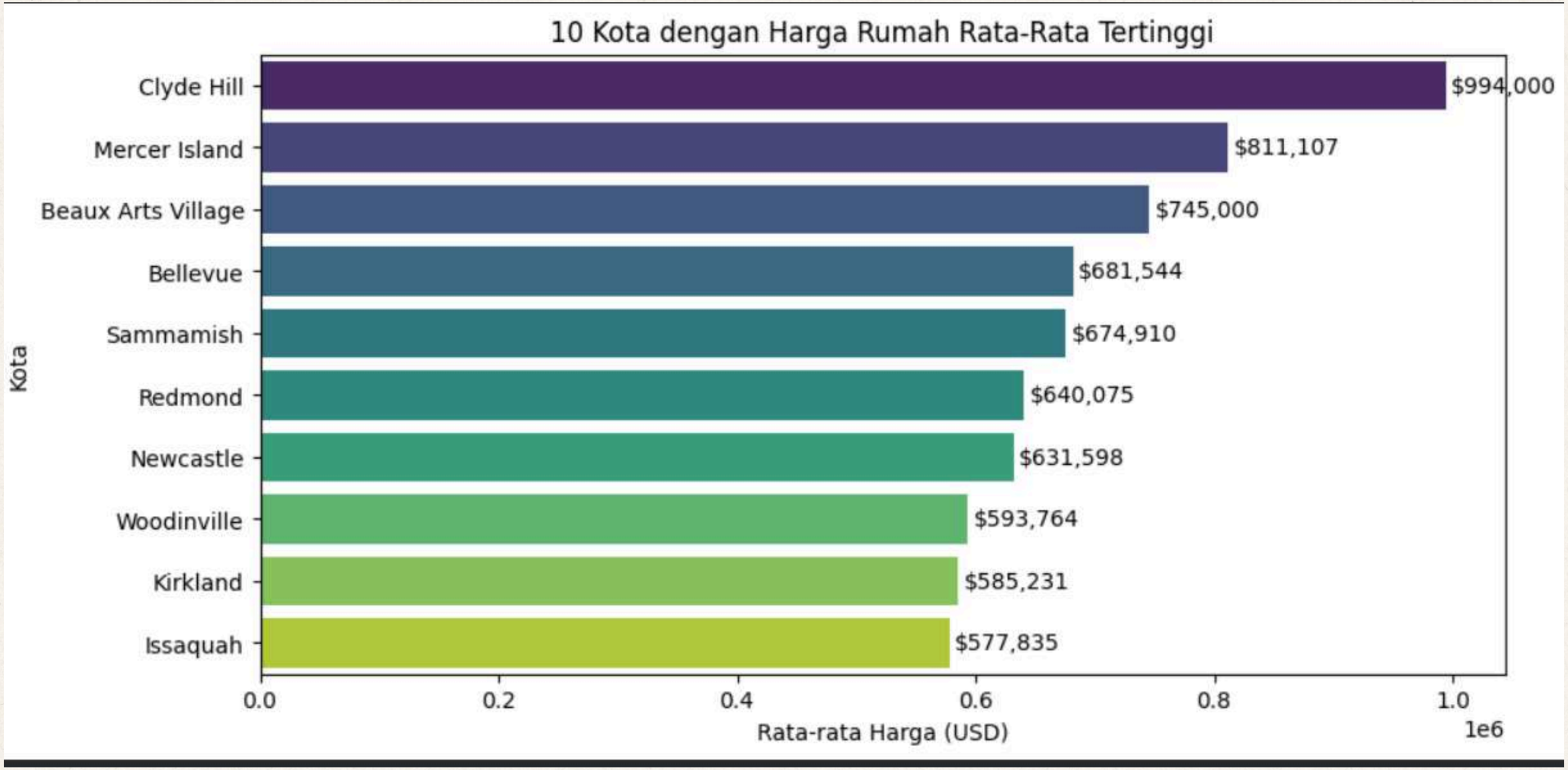
Pengecekan Outlier : Terdapat outlier dan nilai 0

Outlier Handling

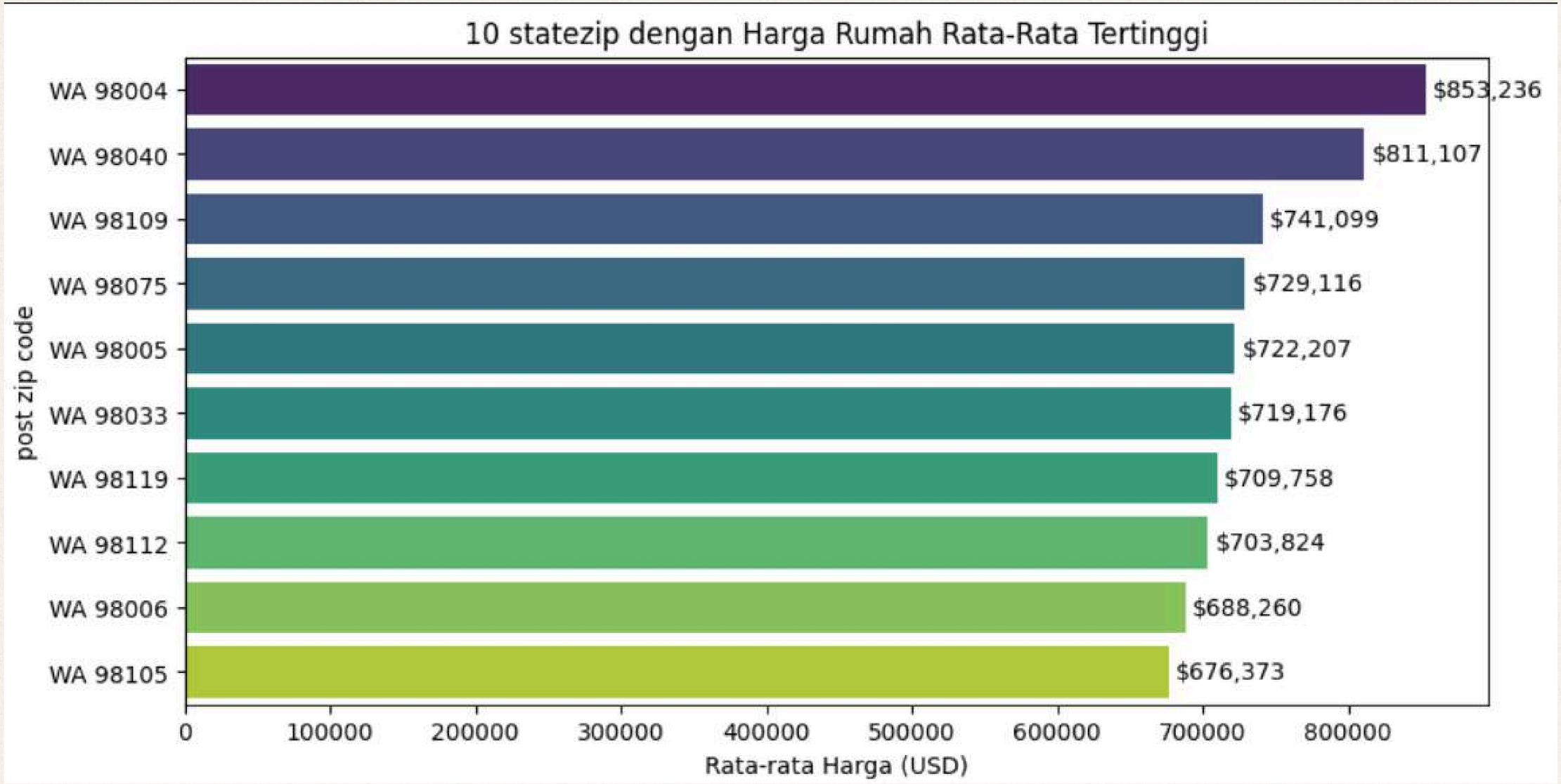
- Jumlah outlier = 240
- drop outlier
- Total baris setelah outlier handling = 4311

Exploratory Data Analysis (EDA)

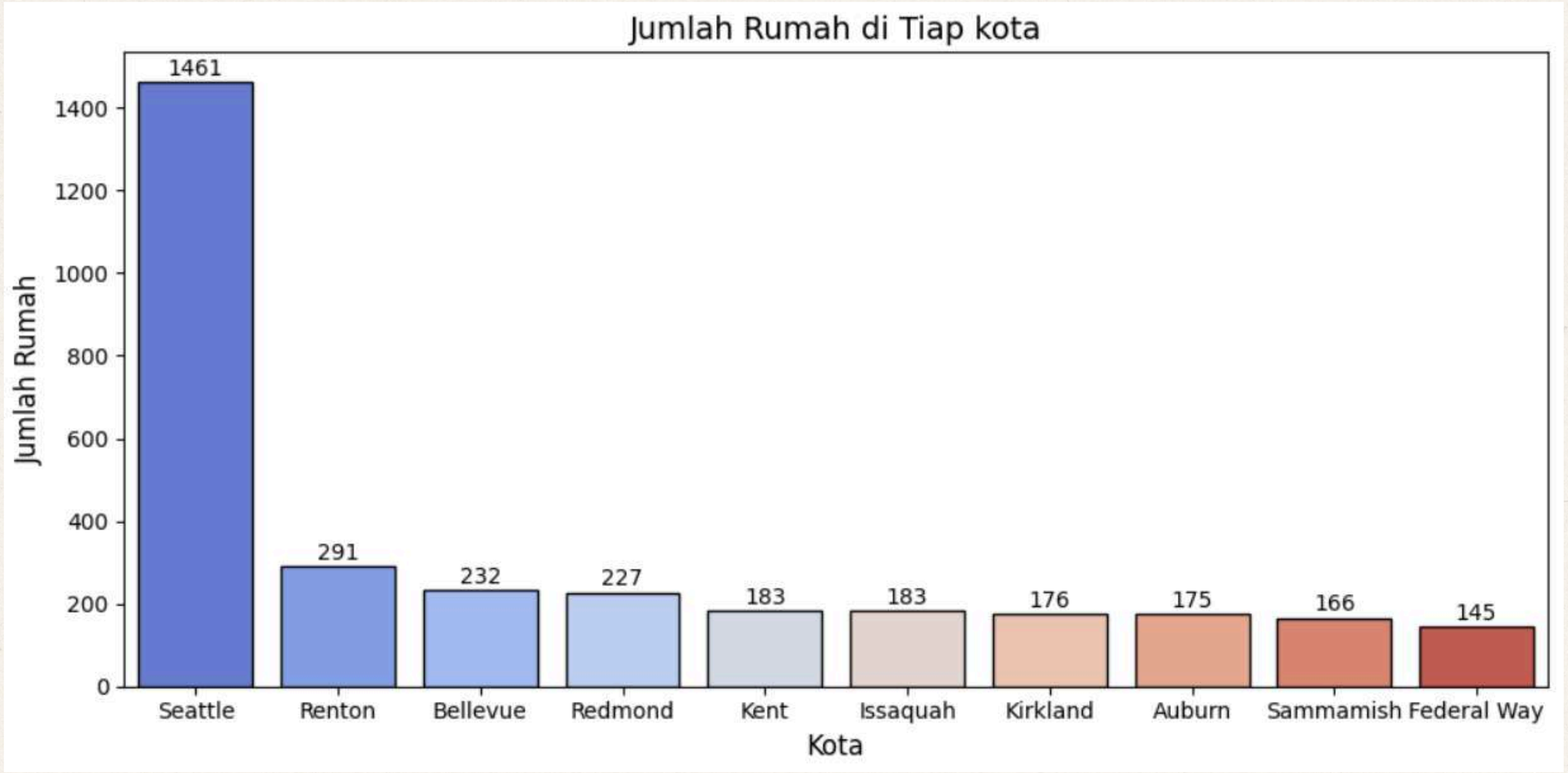
Top 10 Harga Rumah Rata-rata Tertinggi Berdasarkan Kota



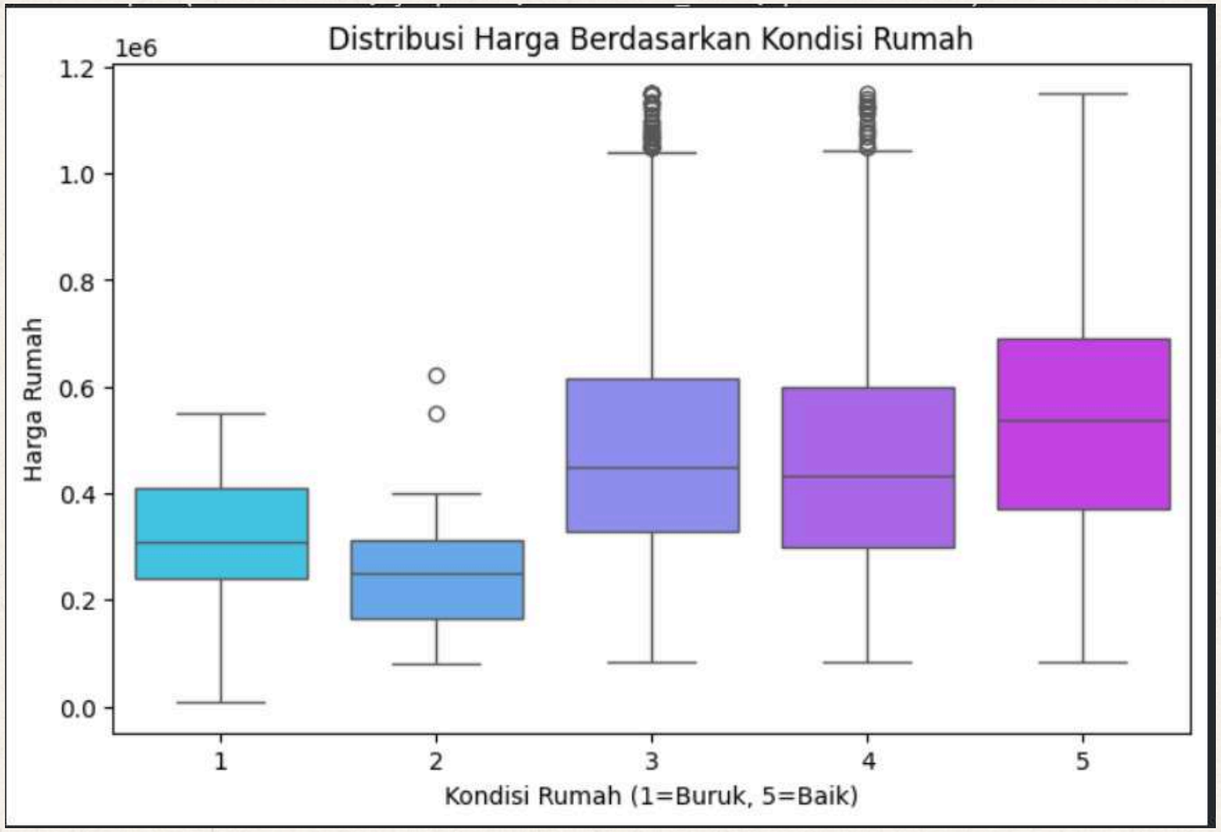
Top 10 Harga Rumah Rata-rata Tertinggi Berdasarkan Statezip

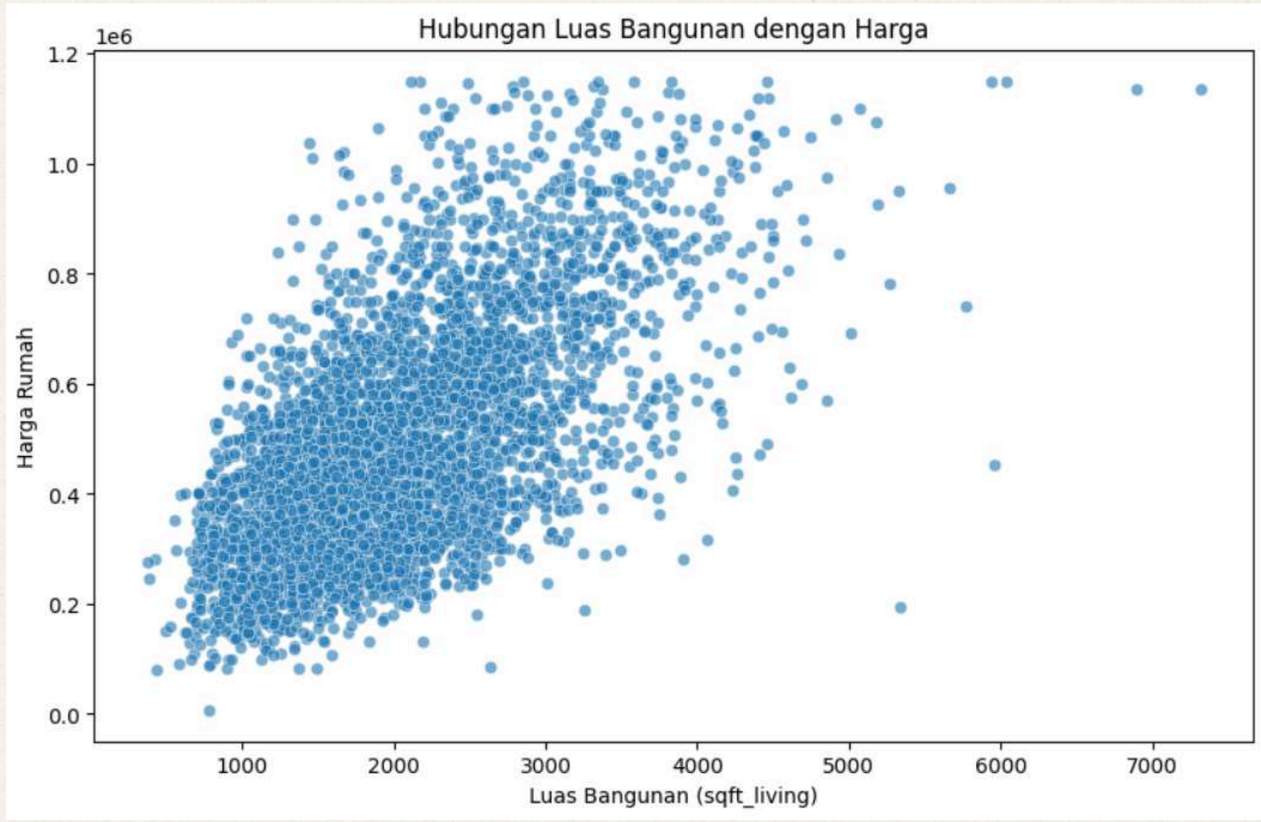


Top 10 Jumlah Rumah di Tiap Kota



Distribusi Data Berdsarkan Harga





Data Preprocessing

Split Dataset

- Data Train = 80%, Test 20%
- Target prediksi = price

Cek Korelasi Data

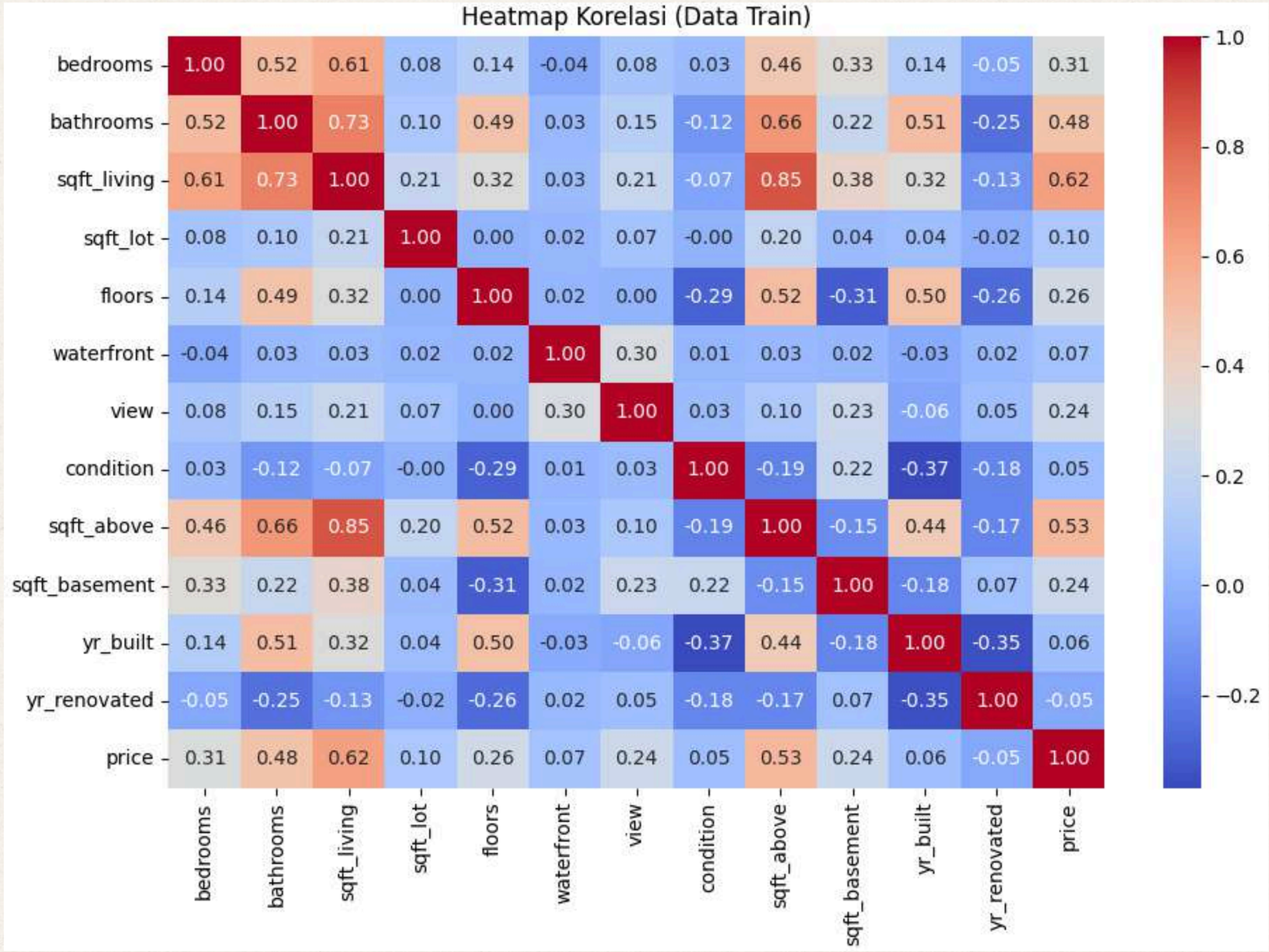
Terdapat kolom/fitur dengan VIF tinggi :

- sqft_living= inf
- sqft_above= inf
- sqft_basement = inf

1	bedrooms	1.748363
2	bathrooms	3.098483
3	sqft_living	inf
4	sqft_lot	1.066035
5	floors	1.937948
6	waterfront	1.104379
7	view	1.210886
8	condition	1.412632
9	sqft_above	inf
10	sqft_basement	inf
11	yr_built	2.059329
12	yr_renovated	1.350211
13	price	1.864279

Pemilihan Fitur

Pada heatmap, diketahui sqft_living pada price mempunyai kolerasi paling tinggi. Sehingga perlu di drop.



Cek nilai Korelasi kembali

Nilai vif score sudah tidak ada yang tinggi lagi.

	feature	vif_score
1	bedrooms	1.748363
2	bathrooms	3.098483
3	sqft_lot	1.066035
4	floors	1.937948
5	waterfront	1.104379
6	view	1.210886
7	condition	1.412632
8	sqft_above	3.222778
9	sqft_basement	2.017274
10	yr_built	2.059329
11	yr_renovated	1.350211
12	price	1.864279

Data Preprocessing

1. Encoding = One-hot Encoding
2. MinMax Scaler = agar distribusi data tidak berjauhan

Modeling

Model 1

Linear Regression

Model 2

Random Forest Regressor

Model 3

XgBoost Regressor

Penjelasan Test model yang dilakukan :

- Metrik yang digunakan adalah R2 Score, MAE, dan RSME
- Pemodelan akan dibandingkan dengan data train

Evaluasi Model

Metrik/Model	Linear Regression Test/Train	Random Forest Regressor Test/Train	XGBoost Regressor Test/Train
R2 Score (Root Mean Squared Error)	**0.423/1.00**	**0.686/0.975**	**0.8/0.975**
MAE (Mean Absolute Error)	**121,560/0.00**	**90,286/26,640**	**66,794/26,640**
RMSE (Root Mean Squared Error)	**162,442/0.00**	**119,891/34.062**	**95,761/34,062**

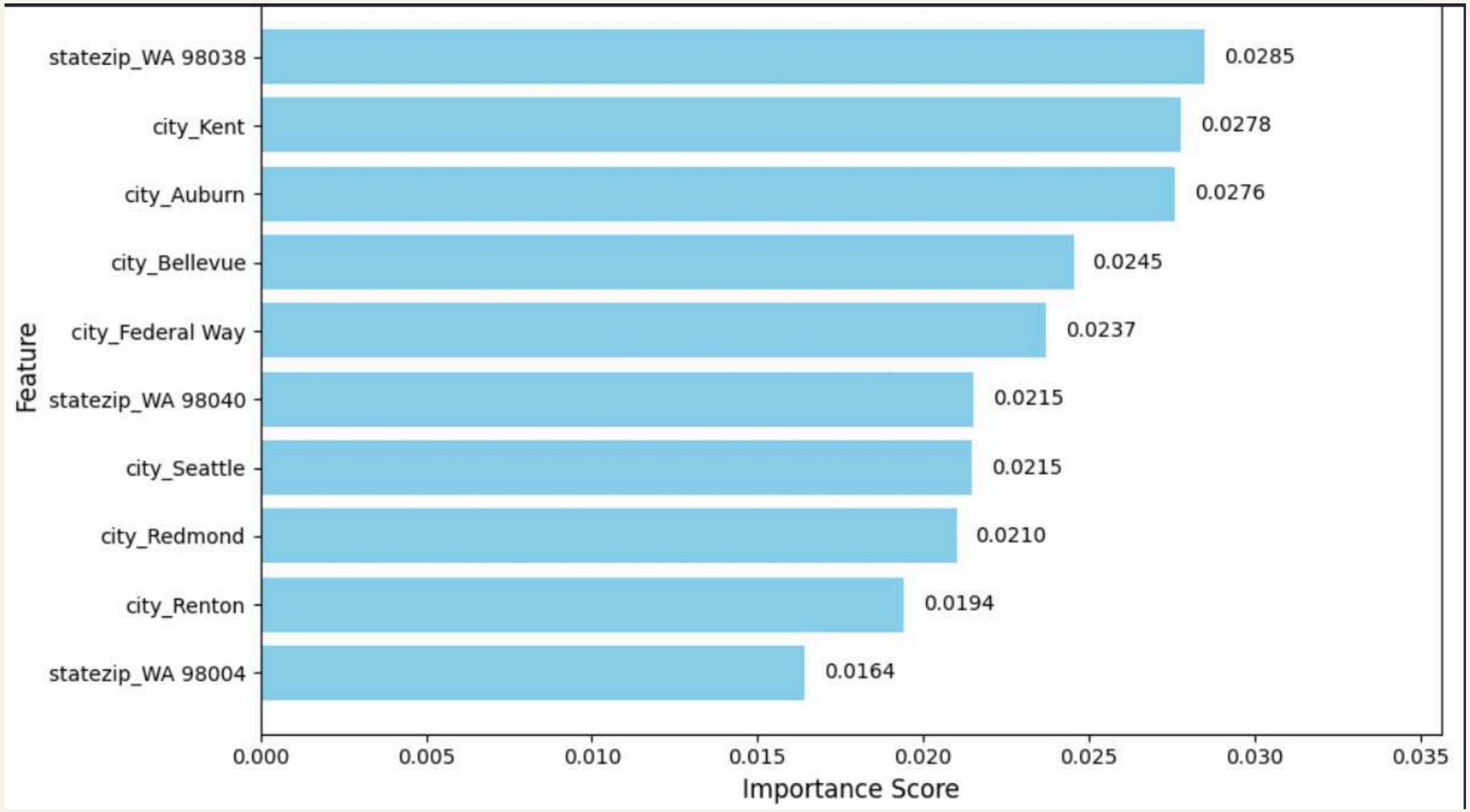
Hasil Pemodelan

XGBoost Regressor mempunyai nilai R2 Score lebih baik dari model lain, yaitu :

- R2 Score = 0.8 atau 80%
- MAE = 66,794 atau rata-rata kesalahan prediksi model = $ 66,794
- RSME = 95,761 atau rata-rata deviasi prediksi model = $ 95,761

Feature Importance

Fitur Importance diambil dari model terbaik



Kesimpulan

Faktor utama yang mempengaruhi nilai pada model XGboost Regressor adalah statezip wa serta city.

Pengaruh lokasi seperti city dan statezip memiliki pengaruh besar seperti yang sudah ditunjukkan pada feature importance.
Penjelasan nilai Metrik model terbaik (XGboost Regressor) :

- Rata-rata prediksi meleset (MAE) = $ 66,794 per rumah
- Rata-rata akar kuadrat prediksi meleset (RMSE) = $ 95,761 per rumah
- R2 Score= 80.00%
- Rata-rata jumlah harga rumah = $ 487,456.90

Kesimpulan

- Akurasi rumah baru dengan rsme dan mae :

a. Berdasarkan MAE

$$\text{Error Relatif} = \frac{66,794}{487,456.90} = 0.137$$

$$\text{Akurasi} = (1 - 0.137) \times 100\% = 86.3\%$$

b. Berdasarkan RMSE

$$\text{Error Relatif} = \frac{95,761}{487,456.90} = 0.196$$

$$\text{Akurasi} = (1 - 0.196) \times 100\% = 80.4\%$$

Saran

- Berdasarkan hasil analisis rata-rata harga rumah, **kota Issaquah dan Kirkland** memiliki harga rumah relatif lebih rendah dibandingkan kota lain di Washington. Lokasi ini cocok untuk pembelian rumah pertama atau hunian pribadi dengan harga lebih ekonomis.

- Berdasarkan analisis feature importance, fitur yang paling berpengaruh terhadap harga rumah adalah **Statezip_WA 98038** dan City (terutama **Maple Valley, Kent, dan Auburn**).
- Lokasi ini berpotensi untuk investasi jangka panjang, karena memiliki nilai pengaruh tinggi terhadap model harga.

- Bisa dilakukan penelitian lebih lanjut dengan menambahkan faktor external seperti fasilitas sekitar, tingkat kriminal, akses pusat kota.

TERIMA KASIH

🔗 GitHub Repository: Link Github Final Project DS
🔗 GitHub Repository: Link Github data Streamlit
🌐 Streamlit App: Deployment Streamlit