Data Science Technical Report

House Price Prediction Nusantara Group

Diserahkan pada tanggal: 14 November 2024

Divisi Data

Nusantara Group

General Information

Penyusun	Taufik Hidayah
Waktu pengerjaan	28 Oktober 2024 – 14 November 2024

Executive Summary

- Brief overview of the project and its objectives

  Nusantara Group adalah perusahaan properti yang fokus pada penjualan rumah. Untuk meningkatkan daya saing di pasar properti, perusahaan berencana menganalisis faktor-faktor yang mempengaruhi harga rumah, seperti jumlah kamar tidur, luas bangunan, kondisi rumah, dan lokasi. Dengan memahami faktor-faktor ini, Nusantara Group berharap bisa menentukan harga yang sesuai nilai pasar dan menyusun strategi pemasaran yang lebih efektif untuk menarik lebih banyak calon pembeli. Melalui upaya ini, perusahaan menargetkan peningkatan pendapatan sebesar 15% pada akhir 2024, peningkatan efektivitas pemasaran pada kuartal pertama tahun 2025, dan kenaikan kepuasan pelanggan menjadi 97% pada kuartal keempat tahun 2024. Target ini diharapkan bisa meningkatkan jumlah pendapatan Perusahaan, meningkatkan efektifitas pemasaran dan meningkatkan kepuasan pelanggan dengan lebih baik.
- Key findings and recommendations
  a. Distribusi Harga Rumah pada Variabel price menunjukkan distribusi yang skewed ke kanan (positively skewed), artinya sebagian besar harga rumah berada pada level rendah hingga sedang, dengan sedikit data pada harga yang sangat tinggi.
  b. Untuk mengatasi nilai yang hilang bisa dilakukan dengan beberapa strategi, imputasi dengan median melihat distribusi dari datanya, supervised imputer atau menggunakan model machine learning untuk imputasi data, dan menghapus data. Penanganan-penanganan tersebut akan kami coba lakukan untuk kebutuhan pemodelan yang optimal.
  c. Sebanyak 50,16% rumah dalam dataset tergolong kategori harga moderate, yaitu rumah dengan rentang harga antara 300 ribu hingga kurang dari 600 ribu. Hal ini menunjukkan bahwa kategori harga moderate mendominasi data.

d. Terdapat Rumah Tanpa Basement sekitar 2.203, menunjukkan bahwa basement bukan fitur umum pada sebagian besar rumah dalam dataset.
e. Terdapat 0,76% atau sekitar 28 rumah yang memiliki nilai 1 pada variabel waterfront, hal ini menandakan bahwa variabel ini sangat jarang ada.
f. Terdapat nilai 0 pada variabel view, yang mencakup sekitar 90% dari seluruh data, padahal rentang nilai untuk variabel ini seharusnya 1 hingga 5. Ada dua kemungkinan pertama, penilaian view menggunakan skala 0–4, atau kedua, rentang view tetap 1–5, namun kebanyakan rumah tidak dinilai dari segi penampilannya, dan tidak ada rumah yang mendapat nilai view 5.
g. Skor Rendah pada Variabel Lain, dimana Kurang dari 3% rumah mendapat skor di bawah 3 dari skala 1 hingga 5, menandakan bahwa sebagian besar rumah memiliki skor menengah hingga tinggi.
h. Rumah yang Telah dan Belum Direnovasi, dari keseluruhan data, sebanyak 2.175 rumah belum direnovasi, sementara 1.505 rumah telah direnovasi, memberikan gambaran umum tentang kondisi rumah.
i. Mayoritas rumah dalam dataset berada di Seattle, dengan jumlah mencapai 1.255 rumah, menunjukkan bahwa Seattle adalah kota dengan rumah terbanyak di dataset ini.
j. Variabel sqft_living (luas bangunan) memiliki korelasi positif yang kuat dengan price (harga rumah), menunjukkan bahwa semakin besar luas bangunan, semakin tinggi pula harga rumah.
k. Terdapat korelasi kuat antara sqft_living dan sqft_above karena banyak rumah memiliki luas bangunan yang sama dengan luas lantai atas.
l. Ada perbedaan median harga antara rumah yang memiliki luas bangunan sama dengan lantai atas dan yang berbeda, mengindikasikan perbedaan nilai pada tipe bangunan ini.
m. Variabel jumlah kamar tidur dan jumlah lantai rumah memiliki hubungan positif dengan harga, walaupun lemah secara individual. Sementara itu, jumlah kamar mandi memiliki korelasi positif yang lebih signifikan dengan harga.
n. Jumlah kamar mandi dan kamar tidur menunjukkan korelasi positif kuat dengan luas bangunan, menandakan bahwa rumah dengan lebih banyak kamar biasanya memiliki bangunan yang lebih luas.
o. Ada satu rumah tanpa kamar tidur dan kamar mandi namun memiliki harga sangat mahal (1,095 juta). Ini bisa menarik bagi pelaku usaha kreatif karena fleksibilitas ruang. Dalam pemodelan, data ini dianggap sebagai noise.
p. Terdapat Lima harga rata-rata terendah didominasi oleh rumah dengan arsitektur pasca-perang, menunjukkan gaya arsitektur ini cenderung memiliki harga lebih rendah.

q. Terdapat lima harga rata-rata tertinggi berasal dari rumah dengan arsitektur era sebelum perang, menunjukkan bahwa rumah klasik memiliki nilai lebih tinggi.
r. Banyak rumah dengan arsitektur klasik (pre-World War II) memiliki harga rata-rata lebih tinggi disebabkan desain unik, bahan berkualitas, dan nilai sejarahnya, serta sering terletak di lokasi premium.
s. Banyak rumah dengan arsitektur masal (post-World War II) memiliki harga rata-rata lebih rendah karena desain yang lebih sederhana dan kualitas bahan yang lebih standar.
t. Harga rata-rata rumah era pre-World War II yang tidak direnovasi dan/atau memiliki basement cenderung lebih tinggi dibandingkan dengan rumah yang sudah direnovasi dan tidak memiliki basement,hal ini dikarenakan rumah dengan karakteristik tersebut sering dianggap lebih bernilai historis atau memiliki potensi renovasi yang lebih besar.
u. Medina, Yarrow Point, dan Clyde Hill memiliki harga rata-rata rumah tertinggi, sementara Seattle memiliki harga tanah per kavling tertinggi meskipun harga dan luas bangunannya tidak tertinggi. Hal ini menunjukkan bahwa meskipun rumah di Seattle mungkin lebih terjangkau, nilai tanah di sana sangat tinggi.
v. Rumah dengan fitur waterfront cenderung memiliki harga rata-rata yang lebih tinggi, dengan sebagian besar masuk dalam kategori harga moderate ke atas. Karena 99,24% rumah tidak memiliki nilai waterfront, sehingga variabel ini akan dipertimbangkan untuk seleksi variabel.
w. Ada satu rumah dengan luas lahan jauh melebihi lainnya namun harga sedikit di atas rata-rata. Rumah ini terletak di area rawan bencana, seperti banjir dan gempa, yang mungkin mempengaruhi harganya.

Business Understanding

1.1. Problem statement and business objectives

Nusantara Group menghadapi tantangan dalam menentukan harga jual rumah yang sesuai dengan nilai pasar. Untuk tetap kompetitif, perusahaan perlu memahami berbagai faktor yang mempengaruhi harga rumah, seperti jumlah kamar tidur, luas bangunan, kondisi rumah, dan lokasi. Dengan menguasai informasi ini, Nusantara Group dapat menetapkan harga yang lebih realistis dan menyusun strategi pemasaran yang lebih kuat untuk menarik lebih banyak calon pembeli. Target bisnis yang ingin dicapai adalah meningkatkan pendapatan, efektivitas pemasaran, dan kepuasan pelanggan. Adapun kriteria suksesnya adalah peningkatan pendapatan sebesar 15% pada akhir 2024, peningkatan efektivitas pemasaran pada kuartal pertama tahun

2025, dan kenaikan kepuasan pelanggan menjadi 97% pada kuartal keempat tahun 2024.

1.2. Stakeholder identification and involvement

Terdapat beberapa pemangku kepentingan utama di setiap tahap pelaksanaannya. Berikut adalah pemangku kepentingan utama beserta tim-tim yang berperan dalam proyek Perusahaan Nusantara Group yaitu:

- Tim Manajemen Proyek, terdiri dari Project Manager (PM), Data Analyst, dan Developer. Mereka bertanggung jawab pada tahap Persiapan dan Perencanaan proyek, memastikan setiap langkah terorganisir dengan baik sesuai timeline. Mereka juga akan mengelola alat manajemen proyek seperti Trello atau Asana untuk pelacakan tugas dan kemajuan proyek.
- Tim Pengumpulan Data, tim ini terlibat dalam Pengumpulan Data dan berperan penting dalam mengakses sumber data seperti database atau survei, dengan dukungan anggaran untuk kebutuhan pengambilan data. Mereka memastikan data yang dikumpulkan sesuai dengan kebutuhan analisis.
- Tim Pengembangan Model, terdiri dari Data Scientist yang akan bekerja pada tahap Pengembangan Model dengan menggunakan software analisis misalnya Python dan vscode. Mereka akan mengembangkan model untuk menganalisis faktor-faktor yang mempengaruhi harga rumah. Tim ini juga memerlukan server untuk pemrosesan data dalam jumlah besar.
- Tim Quality Assurance (QA), tim QA berperan di tahap Pengujian dan Validasi untuk memastikan bahwa model yang dikembangkan memenuhi standar kualitas dan dapat menghasilkan prediksi yang akurat. Mereka akan menggunakan data uji serta software analisis untuk memastikan hasil yang valid.
- Tim IT, Tim ini akan berperan pada tahap Implementasi dan Peluncuran untuk menerapkan model ke dalam sistem operasional. Mereka juga akan memberikan pelatihan kepada pengguna akhir agar dapat menggunakan dashboard atau alat visualisasi dengan mudah.
- Tim Monitoring dan Evaluasi, tim ini bertanggung jawab dalam Evaluasi dan Pemeliharaan yang berkelanjutan untuk memastikan model tetap relevan dan akurat. Mereka akan memonitor kinerja model dan melakukan pembaruan bila diperlukan agar model tetap sesuai dengan perubahan di pasar properti.

Data Understanding

2.1. Data sources and collection methods

Dataset yang digunakan diambil dari SQL Database, tepatnya dari tabel bernama **detail_properti_rumah**. Tabel ini memiliki 16 kolom dengan beberapa jenis data yaitu 3 kolom bertipe float64, 10 kolom bertipe int64, dan 3 kolom bertipe object.

Secara keseluruhan, ada 3680 baris data dalam tabel ini, yang mencatat informasi detail mengenai properti rumah, seperti luas bangunan, jumlah kamar, dan lainnya.

2.2. Data Deskripsi Report

Berikut ini adalah statistik deskriptif untuk kolom bertipe objek dalam data, yang memberikan gambaran tentang distribusi data kategori dan analisis statistik deskriptif pada data numerik seperti berikut:

	count	mean	std	min	25%	50%	75%	max
bedrooms	3,680.00	3.40	0.90	0.00	3.00	3.00	4.00	9.00
bathrooms	3,680.00	2.16	0.78	0.00	1.75	2.25	2.50	6.75
sqft_living	3,680.00	2,137.85	952.49	370.00	1,470.00	1,980.00	2,612.50	10,040.00
sqft_lot	3,680.00	14,814.93	36,122.45	638.00	5,012.50	7,681.50	11,111.00	1,074,218.00
floors	3,680.00	1.51	0.54	1.00	1.00	1.50	2.00	3.50
waterfront	3,680.00	0.01	0.09	0.00	0.00	0.00	0.00	1.00
view	3,680.00	0.24	0.77	0.00	0.00	0.00	0.00	4.00
condition	3,680.00	3.46	0.68	1.00	3.00	3.00	4.00	5.00
sqft_above	3,680.00	1,825.78	857.24	370.00	1,190.00	1,590.00	2,300.00	8,020.00
sqft_basement	3,680.00	312.07	463.45	0.00	0.00	0.00	600.00	4,820.00
yr_built	3,680.00	1,970.56	29.69	1,900.00	1,951.00	1,975.00	1,997.00	2,014.00
yr_renovated	3,680.00	815.56	980.66	0.00	0.00	0.00	1,999.00	2,014.00
price	3,680.00	541,631.26	371,167.47	0.00	323,958.33	460,000.00	650,000.00	7,062,500.00

Analisis statistik deskriptif pada tabel di atas dimana data numerik menunjukkan bahwa terdapat rumah dengan nilai harga price sebesar 0, yang terlihat dari nilai minimum pada variabel tersebut. Hal ini mungkin menunjukkan data yang tidak valid atau kesalahan pengisian data, karena umumnya harga properti tidak

bernilai nol. Kondisi ini perlu dievaluasi lebih lanjut, seperti melakukan pengecekan terhadap entri yang memiliki nilai harga nol untuk memastikan apakah data tersebut valid atau perlu diperbaiki/diimputasi.

	count	unique	top	freq
street	3680	3634	2520 Mulberry Walk NE	3
city	3680	42	Seattle	1255
statezip	3680	77	WA 98103	119

Pada tabel di atas terdapat 3.680 data rumah (street), dengan 3.634 alamat jalan yang berbeda. Alamat jalan yang paling sering muncul adalah Mulberry Walk NE, yang tercatat sebanyak 3 kali. Ada 3.680 data rumah (city), dengan 42 kota yang berbeda. Kota yang paling banyak muncul adalah Seattle, yang tercatat sebanyak 1.255 kali. Terdapat 3.680 data rumah (statezip), dengan 77 kode pos yang berbeda. Kode pos yang paling sering muncul adalah WA 98103, yang tercatat sebanyak 119 kali. Ini menunjukkan bahwa sebagian besar rumah terpusat di Seattle dan ada beberapa alamat yang muncul lebih sering.

2.2.1 Data Description Report

Terdapat 16 kolom dalam dataset, masing-masing berisi 3680 baris, dengan rincian 3 kolom bertipe float, 10 kolom bertipe int64, dan 3 kolom bertipe object. Berikut adalah penjelasan untuk setiap kolom tersebut:

Column Name	Description
bedrooms	Kolom ini menunjukkan jumlah kamar tidur di dalam rumah. Data berisi nilai 0 hingga 9, di mana 0 menunjukkan tidak ada kamar tidur.
bathrooms	Menunjukkan jumlah kamar mandi, dengan rentang nilai dari 0 hingga 6,75. Kenaikan setiap 0,25 menandakan tambahan fasilitas. Kamar mandi biasanya terdiri dari empat fasilitas utama yaitu toilet, wastafel, bak mandi, dan shower. Setengah kamar mandi (0,5) biasanya memiliki toilet dan wastafel saja, biasanya untuk kamar mandi tamu. Tiga perempat kamar mandi (0,75) memiliki toilet, wastafel, dan bak mandi. Seperempat kamar mandi (0,25) memiliki dua versi yaitu hanya toilet atau hanya shower, Kamar mandi tipe ini umumnya terdapat di luar ruangan, misalnya di dekat kolam renang.
sqft_living	Menunjukkan luas bangunan tempat tinggal dalam satuan kaki persegi (square feet atau sqft)
sqft_lot	Menunjukkan luas total tanah tempat rumah berdiri dalam satuan kaki persegi (sqft).

floors	Kolom floors menunjukkan jumlah lantai rumah dalam rentang 1 hingga 3,5. Rumah 1 lantai memiliki seluruh ruang di satu tingkat, yang biasa disebut single-story. Rumah dengan 1,5 lantai biasanya memiliki tambahan ruang di loteng atau mezzanine. Rumah 2 lantai memiliki dua tingkat penuh, dengan ruang tamu dan dapur di lantai bawah, serta kamar tidur di lantai atas. Sementara itu, rumah 2,5 lantai mencakup dua lantai penuh dan tambahan ruang di loteng. Rumah 3 lantai memiliki tiga tingkat penuh, sering ditemukan pada rumah besar atau townhouse, dan untuk 3,5 lantai, biasanya terdiri dari tiga lantai penuh dengan tambahan ruang kecil di atas, seperti di atap
waterfront	Menunjukkan apakah properti memiliki pemandangan atau akses ke perairan seperti danau atau laut. Nilai 0 berarti tidak memiliki pemandangan air, dan 1 berarti memiliki pemandangan air.
view	Menunjukkan kualitas pemandangan properti, dengan skala nilai dari 1 (terendah) hingga 5 (terbaik)
condition	Menunjukkan kondisi fisik keseluruhan properti, berdasarkan skala dari 1 (terburuk) hingga 5 (terbaik).
sqft_above	Luas bangunan di atas permukaan tanah (tidak termasuk basement), dalam satuan kaki persegi (sqft).
sqft_basement	Luas bangunan di bawah tanah atau basement, dalam satuan kaki persegi (sqft).
yr_built	Menunjukkan tahun rumah dibangun. Rentang tahun dalam data ini adalah dari 1900 hingga 2014.
yr_renovated	Menunjukkan tahun terakhir kali properti direnovasi, yaitu tahun 2014.
street	Nama jalan tempat properti berada.
city	Nama kota tempat properti berada.
statezip	Kolom ini berisi kode pos (*ZIP code*) dan kode negara bagian di mana properti berlokasi, yang mengidentifikasi area geografis properti tersebut.
price	Harga properti dalam mata uang dolar, yang menunjukkan nilai jual.

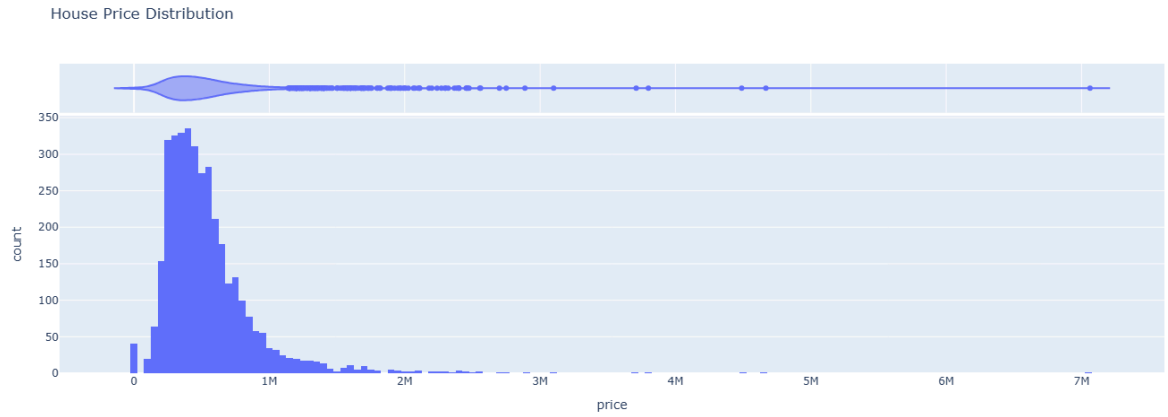
2.3. Data exploration and analysis

2.3.1. Variabel Distribution

a. House Price Distribution

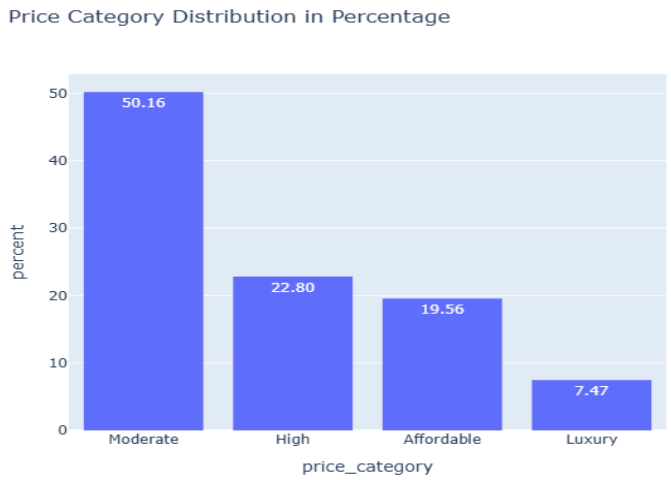
Distribusi data price terlihat positively skewed pada visualisasi di bawah ini dengan banyak nilai ekstrim (outliers). Untuk data yang skewed, ukuran central

tendency yang lebih representatif adalah median, namun juga menggunakan nilai rata-rata sebagai pembanding.



Selanjutnya, membuat variabel baru bernama price_category yang dapat dilihat pada visualisasi di bawah ini, untuk mengkategorikan rumah berdasarkan kisaran harga tertentu. Kategori ini akan membantu dalam analisis yang lebih terfokus terhadap berbagai segmen pasar, yaitu:

- Affordable, Rumah dengan harga di bawah 300,000, yang umumnya ditemukan di daerah pedesaan.
- Moderate, Rumah dengan harga antara 300,000 hingga 600,000, biasanya berada di area pinggiran kota atau wilayah di luar pusat kota.
- High, Rumah dengan harga antara 600,000 hingga 1,000,000, mencerminkan rumah kelas menengah di area perkotaan.
- Luxury, Rumah dengan harga di atas 1,000,000, yang sering dijumpai di pusat kota urban seperti Seattle, Bellevue, dan Kirkland.



Pada dataset yang digunakan, sebagian besar rumah berada dalam kategori harga moderate, yaitu dengan kisaran antara 300 ribu hingga kurang dari 600 ribu. Rumah dengan kategori harga ini mendominasi hingga sekitar 50,16% dari seluruh data, menunjukkan preferensi harga yang lebih umum dan terjangkau bagi sebagian besar pembeli potensial dalam dataset tersebut. Dominasi rumah dengan harga moderate ini mencerminkan segmentasi pasar utama yang mungkin berfokus pada pembeli kelas menengah yang menginginkan hunian dengan nilai yang sesuai namun tetap mempertimbangkan keterjangkauan.

b. Area Information Distribution

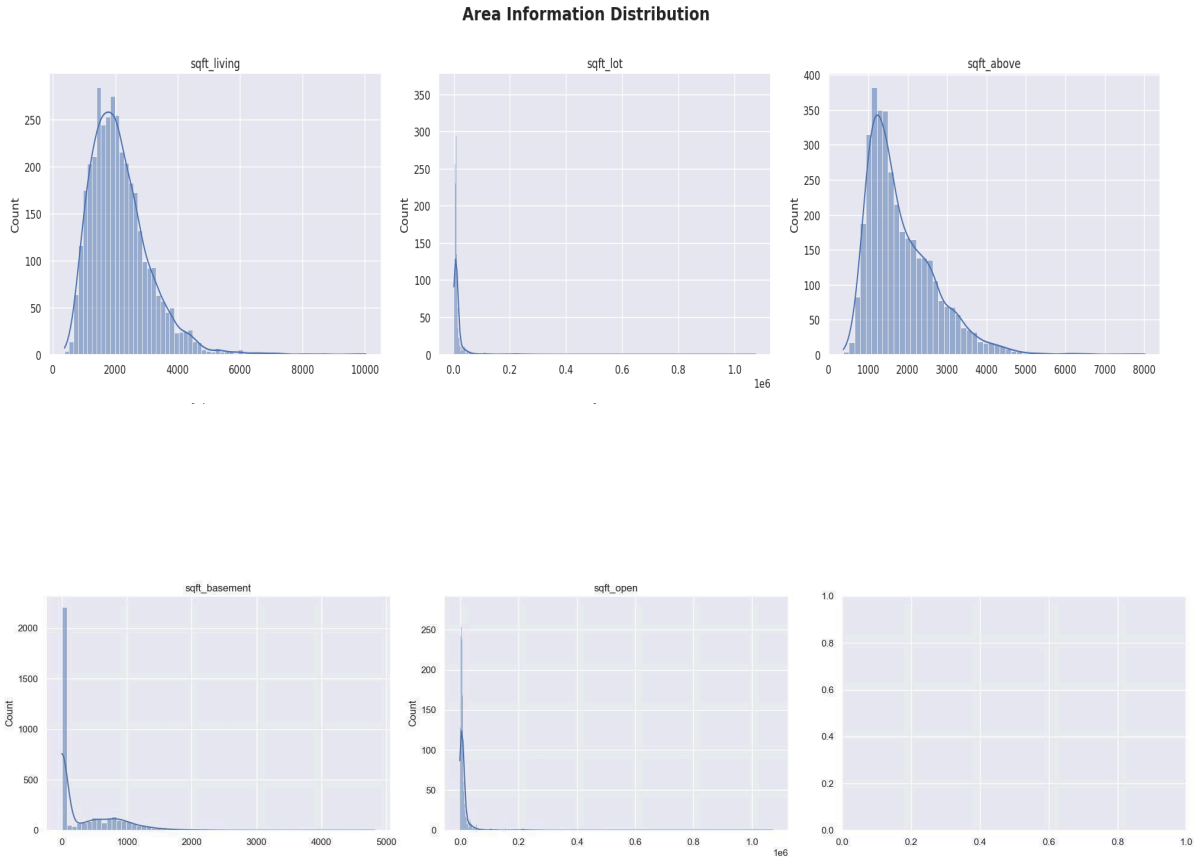
Membuat variabel baru **sqft_open** untuk luas lahan kosong dengan rumus:

$$total\ luas\ lahan\ (sqft\_lot)\ -\ luas\ bangunan\ (sqft\_living)$$

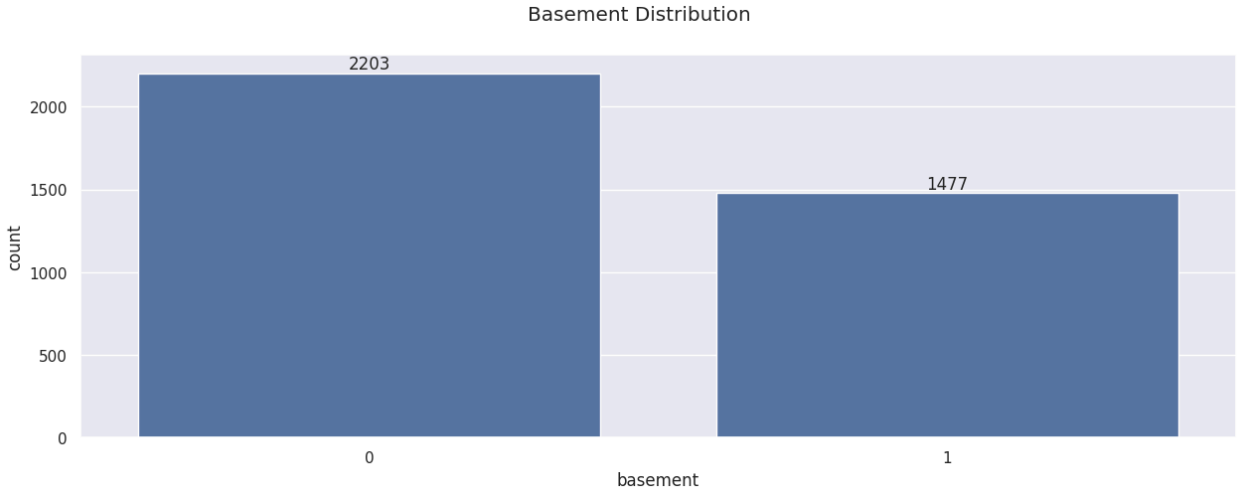
Untuk mengetahui harga per square feet lahan, gunakan variabel **price_per_sqft_lot** dengan rumus:

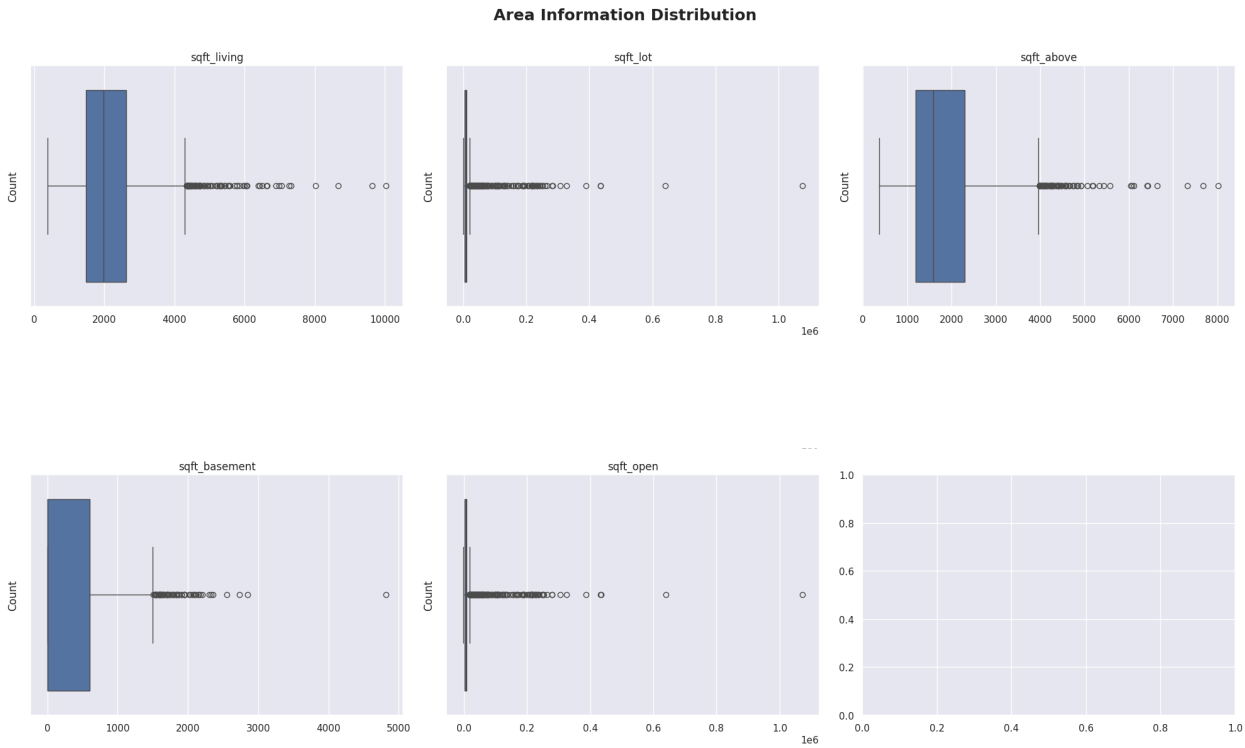
$$\frac{harga\ rumah\ (price)}{luas\ lahan\ (sqft\_lot)}$$

	sqft_open	price_per_sqft_lot
0	14348	25.83
1	6620	89.86
2	7634	31.87
3	900	135.19
...	...	...

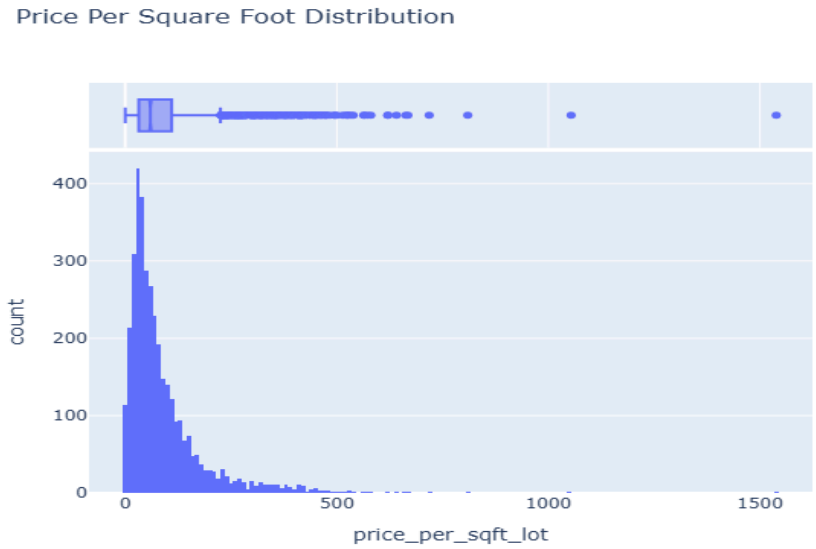


Dalam data, terdapat 2.203 rumah dengan nilai 0 pada luas basement, yang berarti rumah tersebut tidak memiliki basement. Jumlah ini melebihi 50% dari total data, sehingga akan dibuat variabel baru bernama has_basement, di mana rumah dengan basement diberi nilai 1, dan rumah tanpa basement diberi nilai 0. Dapat dilihat pada visualisasi berikut:





Pada variabel area information pada visualisasi di atas, terdapat banyak nilai ekstrim. Salah satu yang menarik adalah pada variabel sqft_lot, di mana terdapat rumah dengan luas keseluruhan lahan lebih dari 1 juta sqft. Nilai ini jauh di atas nilai ekstrim sebelumnya, yaitu 641 ribu sqft, dengan selisih sekitar 350 ribu sqft. Rentang yang sangat besar ini menarik untuk diteliti lebih lanjut.

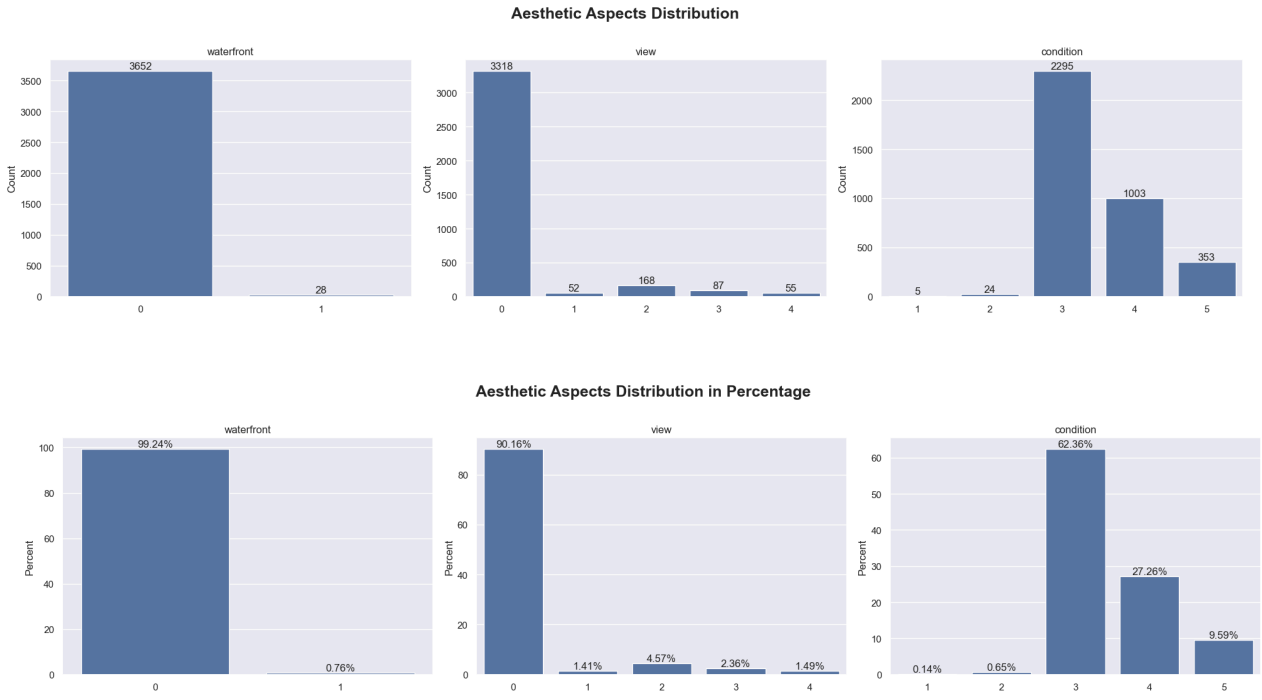


Pada variabel price_per_sqft_lot pada visualisasi di atas, terdapat nilai ekstrim yang cukup jauh dari nilai ekstrim lainnya, menunjukkan adanya perbedaan harga per sqft yang signifikan. Hal ini menarik untuk ditinjau lebih lanjut, karena dapat menunjukkan karakteristik atau kondisi unik pada properti tertentu.

c. Rooms and Floors Distribution

Pada data, ditemukan rumah yang tidak memiliki kamar tidur dan beberapa rumah tanpa toilet yang terlihat pada visualisasi di bawah ini. Selain itu, hanya terdapat satu rumah dengan 9 kamar tidur, serta satu rumah dengan jumlah lantai unik, yaitu 3,5 lantai. Temuan ini menarik untuk dievaluasi lebih lanjut karena mencerminkan fitur properti yang tidak umum.

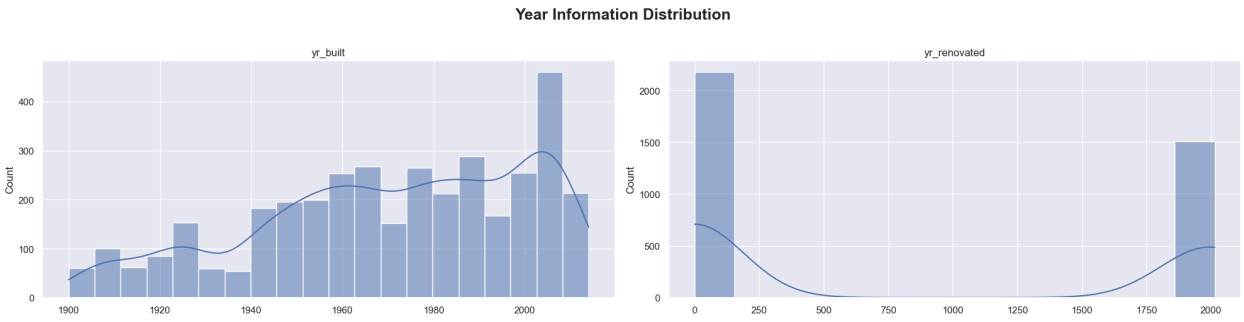
d. Aesthetic Aspects Distribution



Berdasarkan analisis pada visualisasi di atas, aspek estetika mencakup variable waterfront, view, dan condition dari rumah, berikut beberapa *insight* yang didapat:

- Rumah dengan pemandangan ke perairan, Hanya sekitar 0.76% atau sebanyak 28 rumah yang memiliki pemandangan ke perairan. Ini menunjukkan bahwa rumah dengan keunggulan pemandangan ini sangat langka dalam dataset dan mungkin dianggap sebagai properti bernilai lebih tinggi atau premium.
- Nilai *view* sebagian besar rumah sebesar 0, meskipun rentang penilaian untuk atribut ini adalah 1-5. Hal ini dapat diartikan bahwa sekitar 90% rumah dalam dataset tidak dinilai aspek penampilannya, atau kemungkinan besar rumah-rumah ini memang tidak memiliki keistimewaan dalam hal pemandangan.
- Kondisi rumah, dimana Kurang dari 1% rumah memiliki kondisi di bawah 3 dalam skala 1-5. Ini menunjukkan bahwa sebagian besar rumah dalam dataset perusahaan berada pada standar kondisi yang baik hingga sangat baik, yang bisa menjadi daya tarik bagi calon pembeli dan menunjukkan kualitas properti yang tinggi.

e. Year Information Distribution



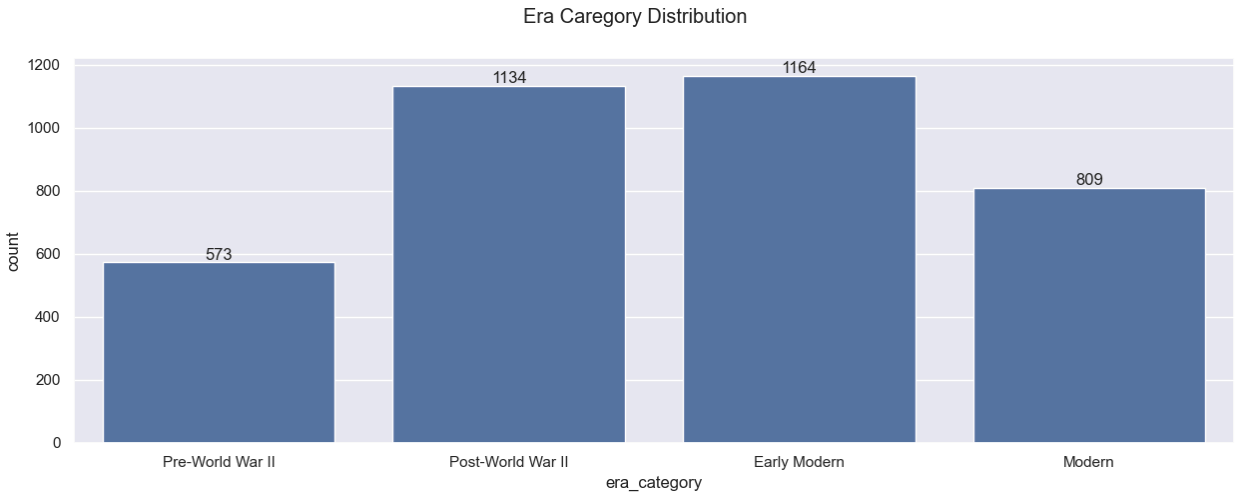
1. Year Built

Untuk memahami lebih baik perbedaan dalam gaya arsitektur dan usia rumah dalam dataset dapat di lihat pada visualisasi di bawah ini, variabel tahun rumah dibangun yang memiliki rentang cukup luas (115 tahun) akan di kategorisasi berdasarkan era dan gaya arsitektur yang khas untuk setiap periode. Kategorisasi ini akan membantu mengidentifikasi tren pasar serta memberikan konteks tambahan terkait usia dan kualitas konstruksi rumah:

- 1900–1939: Era Pra-Perang Dunia II – Rumah dari era ini biasanya memiliki arsitektur klasik dan fitur tradisional yang khas.
- 1940–1969: Era Pasca-Perang Dunia II – Rumah yang dibangun pada masa ini sering kali menunjukkan gaya suburban dan konstruksi massal, seiring dengan pertumbuhan populasi dan kebutuhan perumahan yang meningkat setelah perang.
- 1970–1999: Era Modern Awal – Rumah dari era ini mulai menunjukkan perkembangan dalam teknologi konstruksi dan material bangunan, meskipun belum banyak menerapkan teknologi hemat energi yang canggih.

- 2000–2014: Era Modern – Pada periode ini, rumah sering kali dibangun dengan teknologi terbaru, desain yang lebih efisien energi, dan material yang ramah lingkungan.

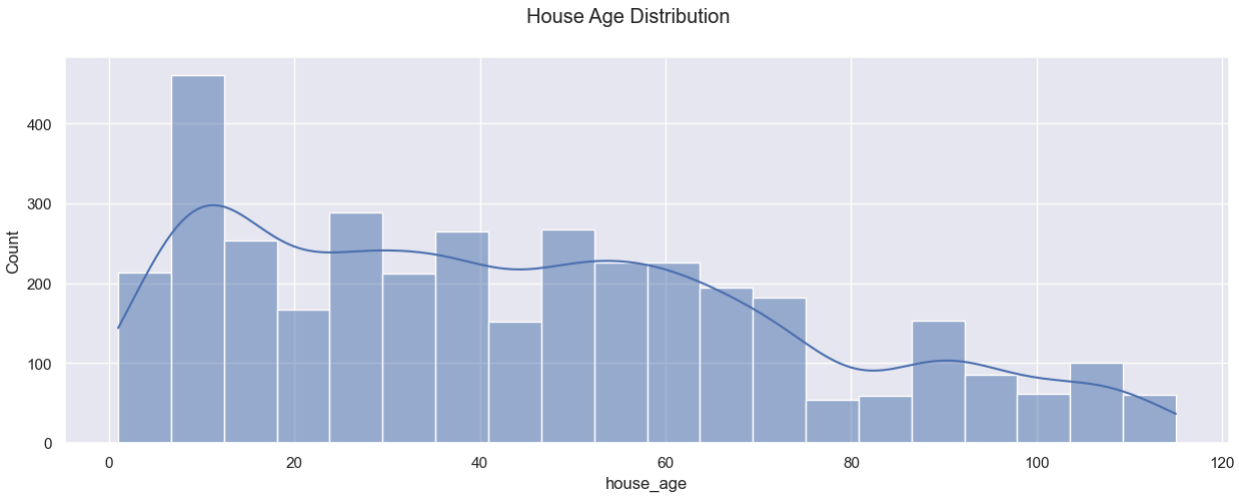
Referensi: Pendekatan ini digunakan dalam analisis properti historis dan penelitian gaya arsitektur, seperti yang tercantum dalam American Housing Survey dan artikel arsitektur historis pada National Register Bulletin.



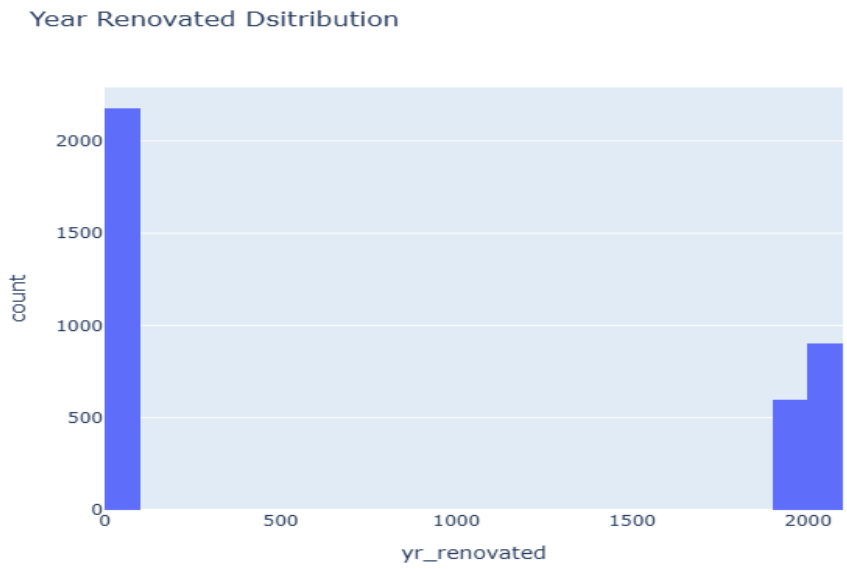
Generate variabel baru house_age untuk menghitung usia rumah, dengan rumus:

$$house\_age = 2015 - year\_built$$

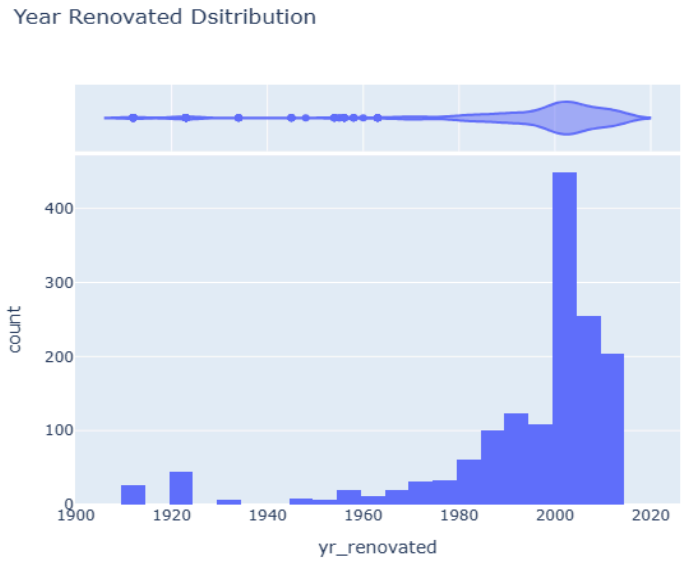
di mana year_built adalah tahun rumah dibangun. Tahun acuan 2015 digunakan agar tidak ada nilai usia 0 tahun, mengingat rumah paling baru dibangun pada tahun 2014. Dapat di lihat pada visualisasi di bawah ini:

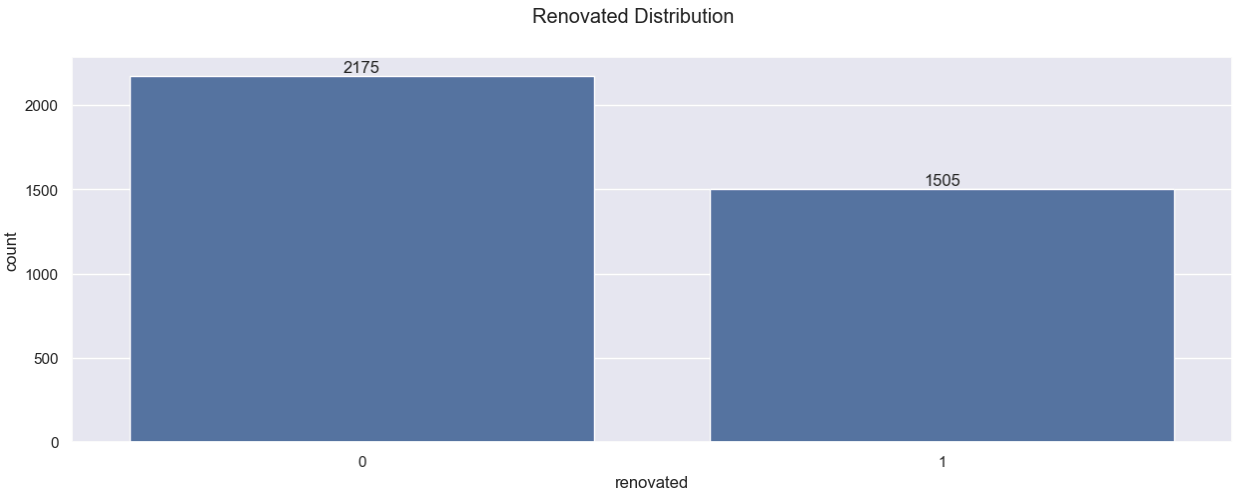


f. Year Renovated



Data menunjukkan perbedaan jumlah yang signifikan antara rumah yang telah direnovasi dan yang belum dilihat pada visualisasi di atas. Oleh karena itu, akan dilakukan kategorisasi biner dengan 1 untuk rumah yang sudah direnovasi dan 0 untuk rumah yang belum direnovasi.





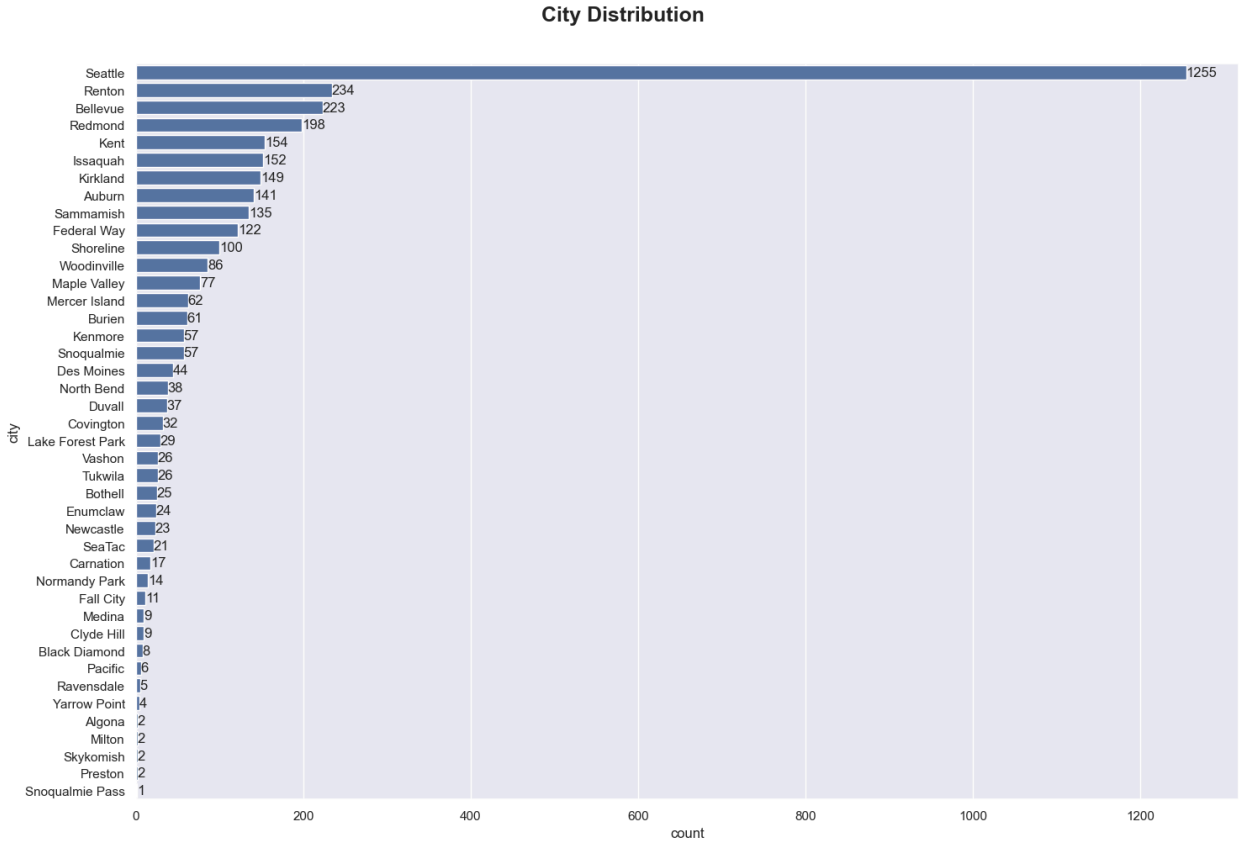
Pada visualisasi di atas masih banyak rumah yang belum direnovasi yakni ada 2175 rumah yang belum direnovasi, dibandingkan dengan 1505 rumah yang sudah direnovasi. Ini menunjukkan bahwa sebagian besar rumah dalam dataset belum menjalani renovasi, yang mungkin mempengaruhi harga dan permintaan pasar.

g. Location Distribution

Berdasarkan analisis, aspek location mencakup tiga variabel utama, yaitu city, street, dan statezip. Ketiga variabel ini memberikan informasi terkait lokasi geografis rumah, mulai dari nama kota, alamat jalan, hingga kombinasi kode pos dan negara bagian. Karena semua kota berada di Washington, terdapat 1 kolom state akan dihapus untuk mengurangi redudansi informasi. Dapat dilihat pada tabel berikut:

	0
state	1
zip_code	77

Sebagian besar rumah dalam dataset terletak di pusat kota Seattle, Washington, dengan jumlah rumah yang signifikan juga ditemukan di daerah Renton dan Bellevue dilihat pada visualisasi di bawah ini. Hal ini menunjukkan bahwa sebagian besar properti berada di area perkotaan yang berkembang pesat di Washington State. Untuk memperkaya informasi lokasi, akan ditambahkan data latitude dan longitude yang terkait dengan masing-masing kota yang dapat dilihat pada tabel di bawah ini:

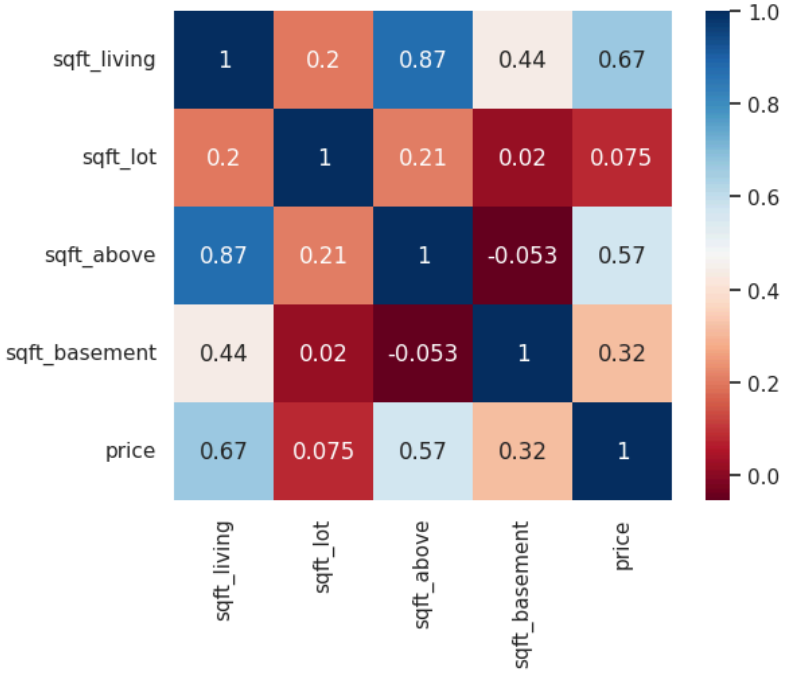


	city	lat	lng
0	Sammamish	47.60	-122.04
1	Kirkland	47.70	-122.21
2	Shoreline	47.76	-122.34
3	Seattle	47.62	-122.32
...	...	...	...

2.3.2. Correlation analysis

a. Area Information vs House Price

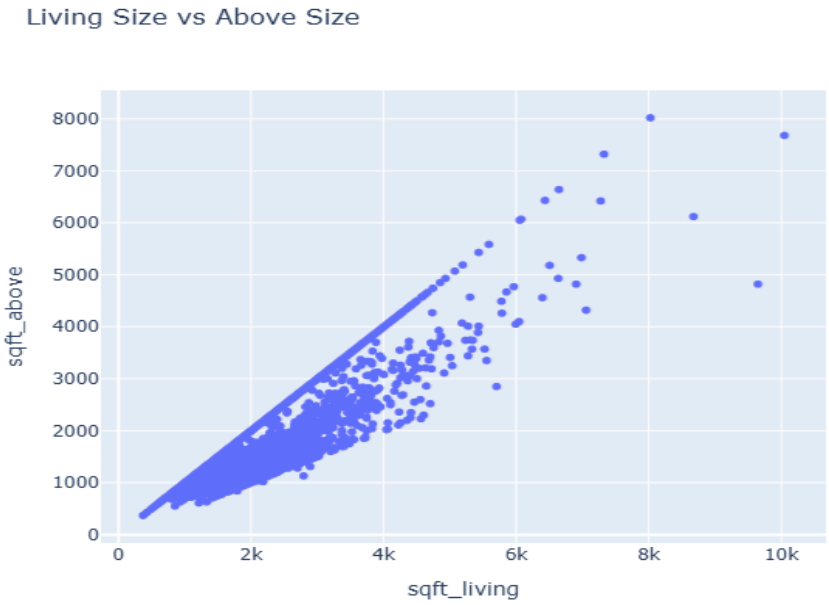
Terdapat banyak anomali atau data unik yang terlihat pada visualisasi di bawah ini. Hal ini akan kita telusuri lebih lanjut pada bagian anomaly analysis untuk memahami lebih detail dan menentukan apakah data tersebut perlu dibersihkan atau diperlakukan secara khusus.



Terdapat beberapa temuan penting dalam analisis data dari hasil korelasi di atas yaitu:

- Korelasi kuat antara variabel sqft_living dan sqft_above, yang menunjukkan bahwa luas bangunan di atas tanah memiliki hubungan yang erat dengan total luas bangunan.
- Luas lahan keseluruhan (sqft_lot) memiliki hubungan yang sangat lemah dengan harga rumah, mengindikasikan bahwa harga rumah lebih dipengaruhi oleh faktor lain selain ukuran lahan.
- Hubungan positif yang kuat ditemukan antara sqft_living dan harga rumah (price), yang menunjukkan bahwa semakin besar luas bangunan, semakin tinggi harga rumah.
- Hubungan positif sedang juga teridentifikasi antara sqft_above dan harga rumah, meskipun tidak sekuat hubungan dengan sqft_living.

Pada kolom sqft_living vs. sqft_above, terdapat hubungan yang sangat linear antara dua variabel, yang ditandai dengan terbentuknya garis lurus pada plot, di mana setiap titik data berada pada garis x = y. Hal ini menunjukkan bahwa kedua variabel tersebut memiliki hubungan yang sangat erat dan konsisten satu sama lain, dengan perubahan pada satu variabel diikuti oleh perubahan yang proporsional pada variabel lainnya dapat dilihat pada visualisasi berikut:



	bedrooms	bathrooms	sqft_living	sqft_lot	floors	...
1	4	2.50	2700	9320	2.00	...
2	2	1.00	790	8424	1.00	...
3	3	2.50	1800	2700	2.00	...

...	...	...	...	...	...	...

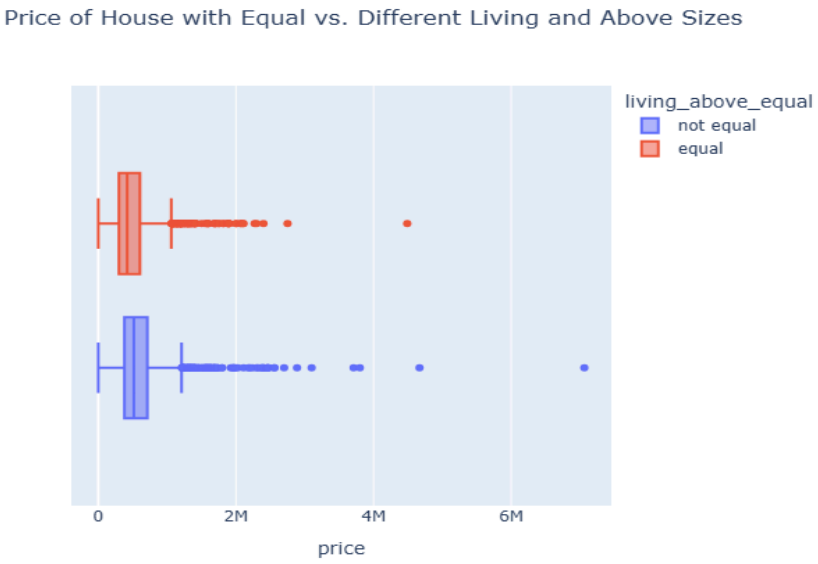
2203 rows x 26 columns

Hal ini menunjukkan bahwa sekitar 59,86% rumah memiliki luas lahan dan luas lantai atas yang sama persis, yaitu sebanyak 2.203 rumah.

Q: Apakah terdapat perbedaan median harga dari rumah yang memiliki luas bangunan dan luas lantai atas yang sama persis dengan yang berbeda?

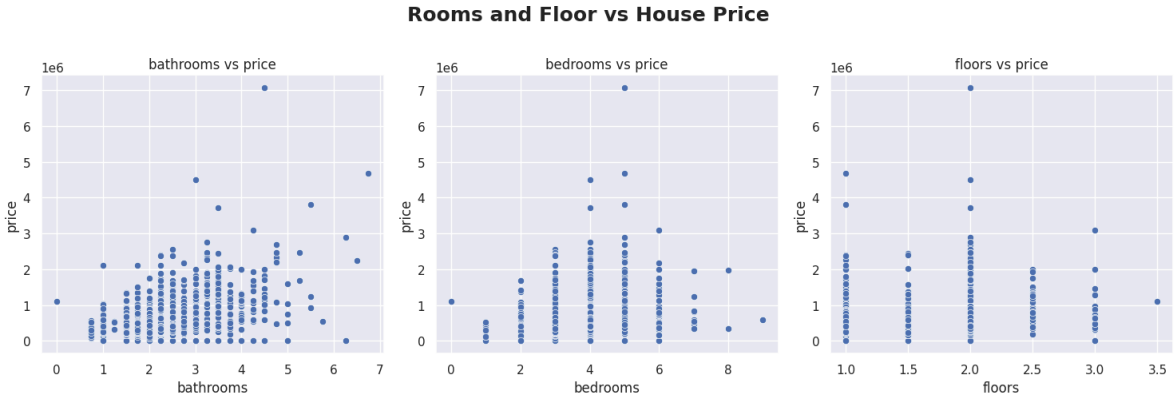
Untuk menjawab pertanyaan apakah terdapat perbedaan median harga rumah antara rumah yang memiliki luas bangunan dan luas lantai atas yang sama persis dengan yang berbeda, kita perlu melakukan analisis perbandingan median harga antara kedua kelompok tersebut:

sqft_living == sqft_above		**sqft_living != sqft_above**	
	price		**price**
count	2,203.00	**count**	1,477.00
mean	491,406.32	**mean**	616,543.62
std	304,265.35	**std**	442,597.33
min	0.00	**min**	0.00
25%	300,000.00	**25%**	380,000.00
50%	420,000.00	**50%**	516,200.00
75%	605,000.00	**75%**	712,000.00
max	4,489,000.00	**max**	7,062,500.00



Terdapat perbedaan median harga yang signifikan antara rumah dengan luas bangunan dan luas lantai atas yang sama persis dan yang berbeda, tanpa mempertimbangkan variabel lain. Hal ini menunjukkan bahwa kesamaan atau perbedaan antara kedua ukuran tersebut dapat memengaruhi harga rumah secara langsung.

b. Rooms and Floor vs House Price

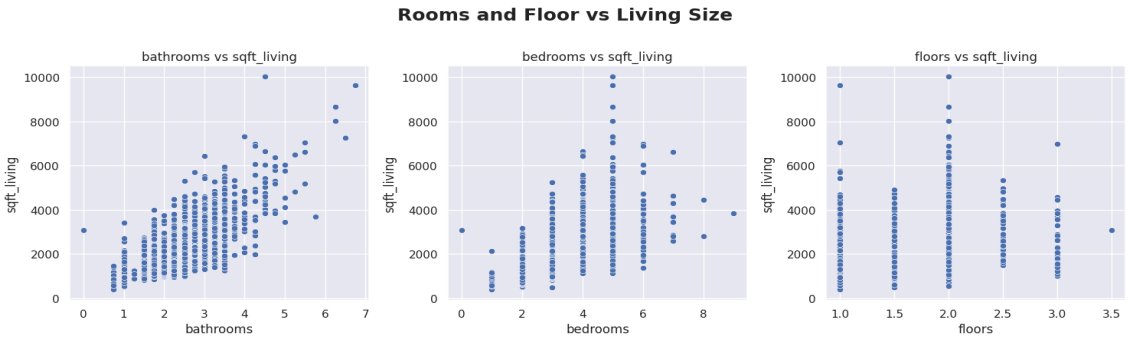


Terdapat beberapa temuan penting terkait hubungan antara fitur rumah dan harga:

- Jumlah kamar tidur dan jumlah lantai memiliki hubungan positif dengan harga rumah, namun dengan hubungan yang lemah secara individual.
- Jumlah kamar mandi memiliki hubungan positif yang signifikan dengan harga rumah, dengan tingkat hubungan yang sedang.
- Visualisasi scatter plot antara jumlah kamar tidur dan jumlah lantai terhadap harga rumah menunjukkan adanya banyak anomali, yang perlu ditelusuri lebih lanjut untuk memahami penyebabnya.

* Tanpa mempertimbangkan variabel lain

c. Rooms and Floor vs Living Size



Non-linear, Terdapat hubungan positif yang cukup signifikan dengan kekuatan korelasi sedang antara jumlah lantai (floors) dan luas bangunan (sqft_living). Ini menunjukkan bahwa semakin banyak jumlah lantai, semakin besar pula luas bangunan rumah, meskipun hubungan ini bersifat non-linear. Hasil uji signifikansi menunjukkan nilai statistik 0.40 dengan p-value sangat kecil (2.34e-144), yang menunjukkan hubungan yang sangat signifikan secara statistik. Dapat dilihat pada visualisasi di atas:
*Tanpa mempertimbangkan variabel lain

	bathrooms	**bedrooms**	**sqft_living**
bathrooms	1.00	0.54	0.76
bedrooms	0.54	1.00	0.60
sqft_living	0.76	0.60	1.00

Linear, Terdapat hubungan linear positif yang sangat signifikan antara jumlah kamar mandi (bathrooms) dan jumlah kamar tidur (bedrooms) dengan luas bangunan rumah (sqft_living), dengan hubungan yang kuat yang dapat dilihat pada tabel di atas. Ini berarti semakin banyak jumlah kamar mandi dan kamar tidur, semakin besar pula luas bangunan rumah tersebut. Tanpa mempertimbangkan variabel lain

d. Aesthetic Aspects vs House Price

- Binary vs Continuous Correlation

Terdapat hubungan positif yang lemah antara variabel waterfront (yang menunjukkan apakah rumah memiliki pemandangan ke area perairan) dan harga rumah. Uji signifikansi menghasilkan nilai statistik 0.205 dengan p-value sangat kecil (2.82e-36), yang menunjukkan hubungan ini cukup signifikan meskipun lemah. Hal ini perlu ditelusuri lebih mendalam, karena spesifikasi waterfront biasanya

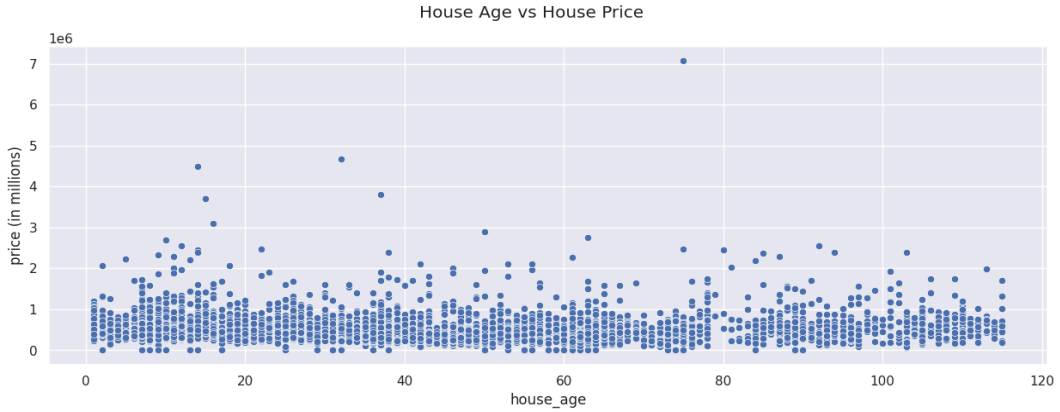
terkait dengan rumah yang memiliki harga lebih tinggi karena pemandangan ke perairan, yang menjadi salah satu faktor penentu harga mahal. Masalah ini akan dibahas lebih lanjut pada bagian deep dive.

- Non-linear Insights

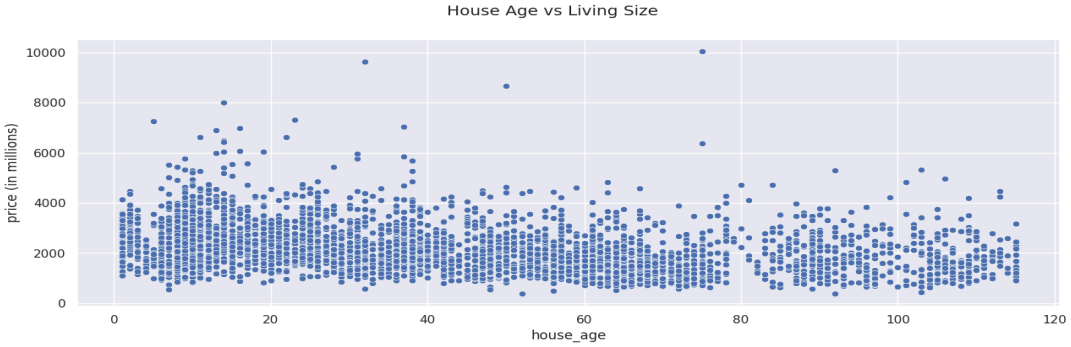
Terdapat hubungan positif yang signifikan antara view dan harga rumah, dengan statistik 0.259 dan p-value sangat kecil (2.77e-57), yang menunjukkan adanya hubungan yang kuat meskipun tidak linear. Untuk variabel condition (kondisi rumah), uji signifikansi menghasilkan statistik 0.018 dan p-value 0.28 , yang menunjukkan bahwa tidak ada hubungan signifikan dengan harga rumah, atau hubungan yang ada sangat lemah.

e. House Age

- *vs House Price*



- *vs Sqft_living*



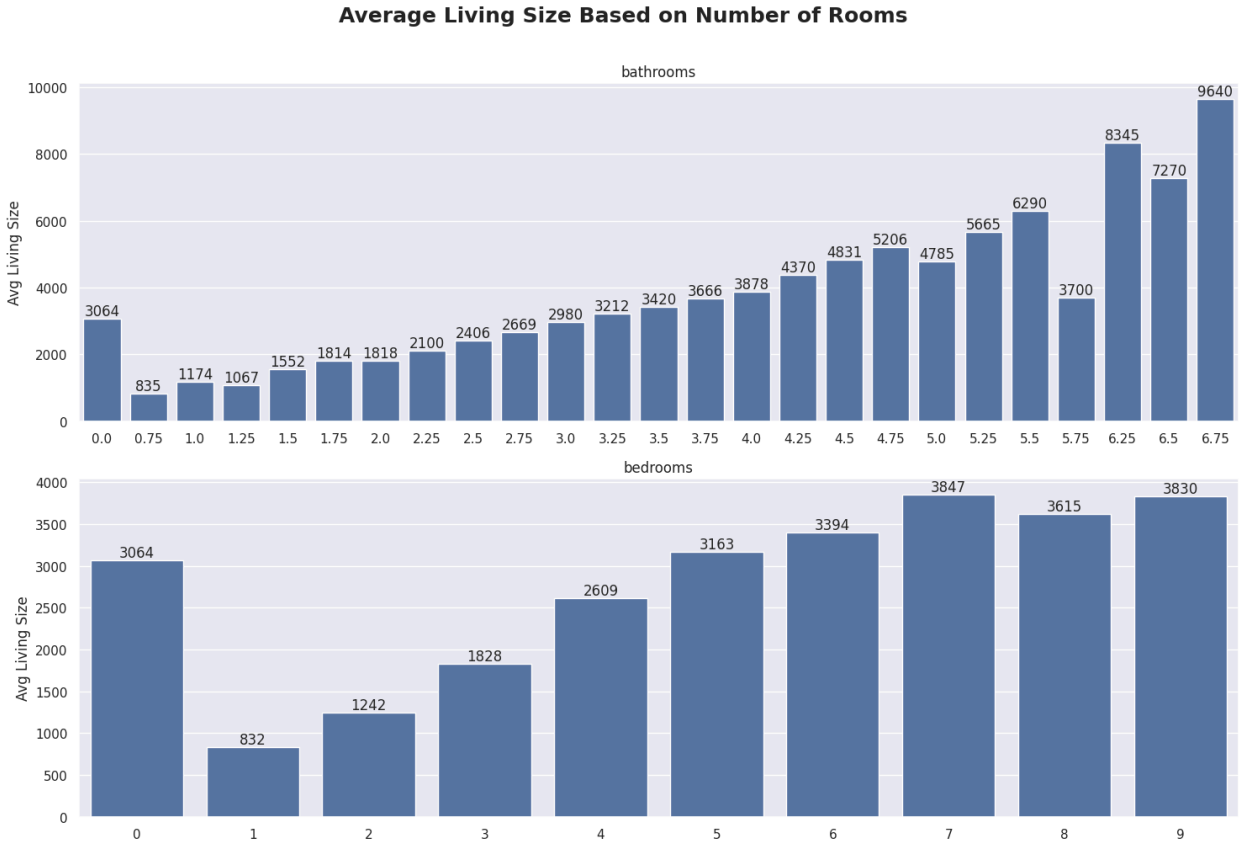
Secara individual, variabel house_age belum menunjukkan hubungan yang signifikan dengan harga rumah (price) maupun luas bangunan (sqft_living). Hal ini mengindikasikan bahwa usia rumah tidak menjadi faktor utama yang mempengaruhi harga atau ukuran bangunan rumah secara langsung.

2.3.3. Deep Dive

a. Strong Correlation Between Rooms and Living Size

Mengacu pada insight dari sub bab 2.3.2. Correlation Analysis point b, terdapat hubungan yang kuat antara jumlah kamar mandi (bathrooms) dan jumlah kamar tidur (bedrooms) secara individual dengan luas bangunan rumah (sqft_living). Hal ini menunjukkan bahwa semakin banyak jumlah kamar mandi dan kamar tidur, semakin besar pula luas bangunan rumah tersebut.

Mari kita telusuri lebih dalam untuk memahami lebih lanjut tentang kekuatan hubungan ini dan apakah ada faktor lain yang mempengaruhi korelasi ini. Kita dapat menggali lebih dalam dengan melihat distribusi data, pola yang lebih kompleks, atau pengaruh variabel lain yang mungkin turut berperan.



- Rumah tanpa kamar mandi memiliki rata-rata luas bangunan yang lebih besar dibandingkan dengan rumah yang memiliki 3 kamar mandi dengan fasilitas lengkap. Hal ini cukup menarik karena biasanya rumah dengan fasilitas lebih banyak (seperti kamar mandi) cenderung lebih kecil dalam ukuran ruang per unit fasilitas. Mungkin saja rumah tanpa kamar mandi ini memiliki ruang luas namun

tidak terfokus pada fasilitas kamar mandi, atau bisa jadi tipe rumah tertentu yang memiliki ukuran lebih besar namun lebih sederhana dalam hal fasilitas.

- Rumah tanpa kamar tidur memiliki rata-rata luas bangunan yang lebih besar dibandingkan dengan rumah yang memiliki 4 kamar tidur. Fenomena ini juga bisa disebabkan oleh desain rumah tertentu, di mana rumah yang tidak memiliki kamar tidur mungkin didesain dengan ruang terbuka yang lebih luas, atau bisa juga rumah tersebut memiliki fungsi yang berbeda (seperti ruang komersial atau ruang besar untuk keperluan lain), meskipun tidak dilengkapi dengan kamar tidur.
- Rata-rata luas bangunan rumah dengan 5.75 kamar mandi (yang terdiri dari 5 kamar mandi dengan fasilitas lengkap dan 1 kamar mandi dengan 3 fasilitas) hampir setara dengan rumah yang memiliki 3.75 kamar mandi. Ini menunjukkan adanya ketidakseimbangan antara jumlah kamar mandi dan ukuran rumah, di mana lebih banyak kamar mandi tidak selalu berbanding lurus dengan ukuran rumah yang lebih besar.

*tanpa mempertimbangkan variabel lain

Next step: Untuk memahami lebih lanjut dan mencari penjelasan

1. Investigae House Without Bathrooms or Bedrooms

Mengacu pada sub bab 2.3.1. Variable Distribution point c, hanya terdapat masing-masing satu rumah yang tidak memiliki kamar mandi atau kamar tidur. Hal ini menunjukkan adanya nilai ekstrim yang jarang ditemukan pada dataset, yang mungkin merupakan kasus unik atau anomali dalam data. Data ini perlu ditelusuri lebih lanjut untuk memahami apakah ada kesalahan pencatatan atau apakah ada alasan tertentu mengapa rumah-rumah tersebut tidak memiliki kamar mandi atau kamar tidur.

bedrooms == 0				
	bedrooms	bathrooms	sqft_living	...
1991	0	0.00	3064	...
bathrooms == 0				
	bedrooms	bathrooms	sqft_living	...
1991	0	0.00	3064	...

- Ternyata rumah yang tidak memiliki kamar mandi dan kamar tidur adalah rumah yang sama, yang terletak di pusat Seattle, Washington, dan memiliki 3.5 lantai dengan luas bangunan dan luas lantai atas yang sama. Rumah ini termasuk dalam kategori luxury house dengan harga yang sangat mahal. Kemungkinan besar, rumah ini didesain untuk peluang industri atau usaha, di mana tidak adanya kamar mandi

dan kamar tidur memberikan fleksibilitas bagi calon pembeli untuk menyusun skema ruangan atau mendesain ulang rumah sesuai kebutuhan (customizable house).

- Karena rumah ini memiliki spesifikasi yang sangat berbeda, terutama dalam hal desain dan fungsionalitas, luas bangunan rumah ini sangat tinggi dibandingkan dengan rata-rata rumah lainnya yang memiliki jumlah kamar mandi dan kamar tidur tertentu. Hal ini dapat menyebabkan nilai rata-rata luas bangunan rumah tersebut jauh lebih tinggi.
- Untuk keperluan pemodelan, rumah ini dianggap sebagai bias karena spesifikasi uniknya sangat berbeda dengan rumah-rumah lain, sehingga pengaruhnya terhadap variabel target price tidak sebanding dengan rumah-rumah lain. Rumah ini perlu diperlakukan secara terpisah untuk menghindari distorsi dalam model prediksi harga rumah.

2. Investiigate house with 5.75 bathrooms less space than house with other bathrooms size

	bedrooms	bathrooms	sqft_living	...
3358	7	5.75	3700	...

- Mengacu pada sub bab 2.3.1. Variable Distribution point c, hanya terdapat satu rumah dengan jumlah kamar mandi 5.75. Ini berarti bahwa nilai rata-rata luas bangunan rumah tersebut adalah nilai aktual luas bangunannya sendiri. Rumah ini termasuk dalam kategori harga moderate (antara 300 ribu hingga 600 ribu). Berdasarkan 2.3.2. Correlation Analysis point a, di mana harga rumah memiliki korelasi kuat dengan sqft_living, rumah ini diharapkan memiliki luas bangunan yang sesuai dengan kategori harga moderate, yaitu ukuran luas bangunan yang sedang.
- Jika dibandingkan dengan rumah dengan jumlah kamar mandi yang berbeda, rumah ini cenderung memiliki luas bangunan yang lebih rendah. Hal ini disebabkan oleh pengaruh harga, yang dalam kategori moderate membatasi ukuran luas bangunan dibandingkan dengan rumah dalam kategori harga yang lebih tinggi.

b. Let's go back to the past

Pada bagian ini kami akan berusaha mengeksplorasi variable year information vs. semua variable:

1. Year vs. area info and price

	yr_built	avg_price

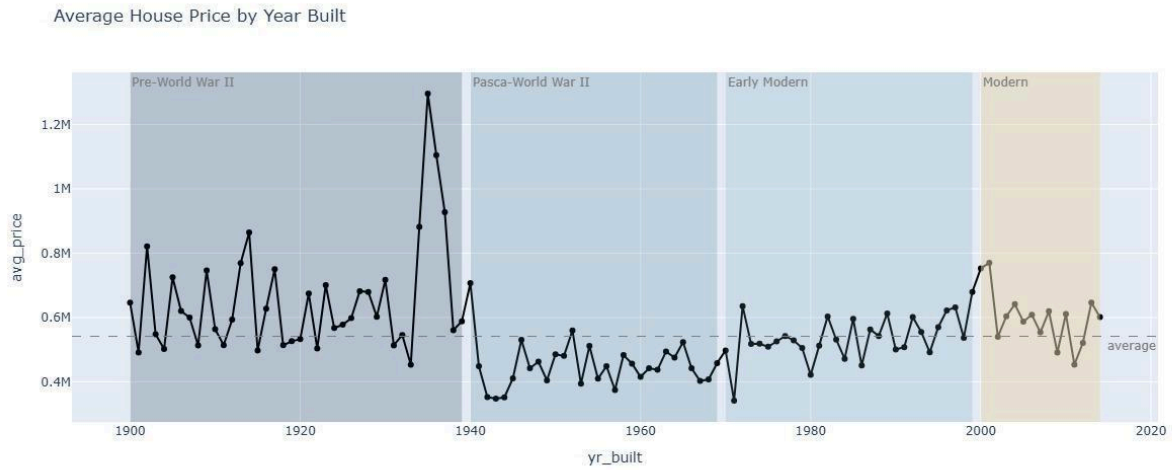
0	1971	341,463.52
1	1943	347,948.64
2	1944	351,663.43
3	1942	352,232.50
4	1957	374,530.25

- Lima rumah dengan harga rata-rata terendah didominasi oleh rumah dengan gaya arsitektur setelah perang (post-war architecture). Hal ini menunjukkan bahwa rumah-rumah bergaya arsitektur ini cenderung berada di segmen harga yang lebih rendah, mungkin karena faktor usia, desain sederhana, atau bahan yang digunakan, yang dapat mempengaruhi nilai pasar rumah tersebut. Dapat dilihat pada tabel di atas ini:

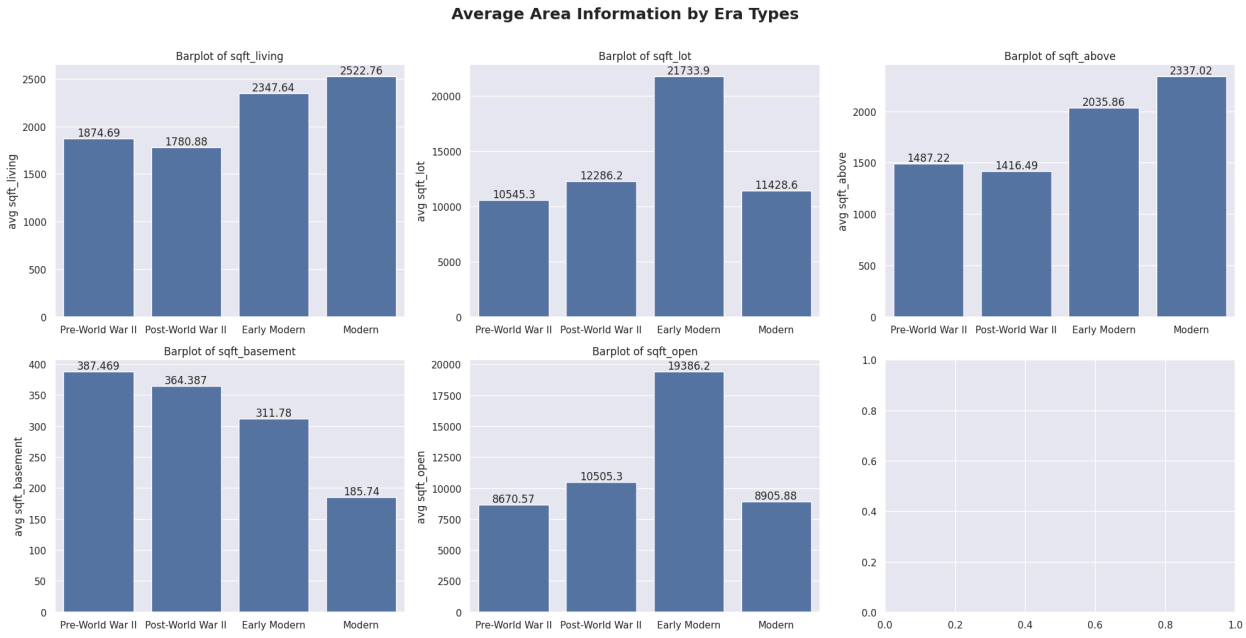
	yr_built	avg_price
0	1935	1,296,166.67
1	1936	1,105,000.00
2	1937	927,218.75
3	1934	881,750.00
4	1914	864,360.00

- Lima rumah dengan harga rata-rata tertinggi semuanya memiliki gaya arsitektur dari era sebelum perang (pre-war architecture). Rumah-rumah ini cenderung memiliki nilai pasar yang lebih tinggi, yang mungkin disebabkan oleh karakteristik unik gaya arsitektur klasik, nilai historis, atau kualitas bahan dan detail desain yang lebih tinggi yang sering ditemukan pada bangunan dari periode tersebut. Dapat dilihat pada tabel di atas.

Rumah dengan gaya arsitektur klasik (pra-World War II) memiliki rata-rata harga per tahun yang seringkali berada di atas harga rata-rata keseluruhan. Sebaliknya, rumah dengan gaya arsitektur masal (pasca-World War II) cenderung memiliki harga rata-rata per tahun yang berada di bawah rata-rata keseluruhan. Selain itu, terdapat anomali pada rata-rata harga rumah untuk tahun 1935 dan 1936. Anomali ini akan ditelusuri lebih lanjut pada bagian Anomaly Analysis untuk memahami penyebabnya. Dapat dilihat pada visualisasi di bawah ini:

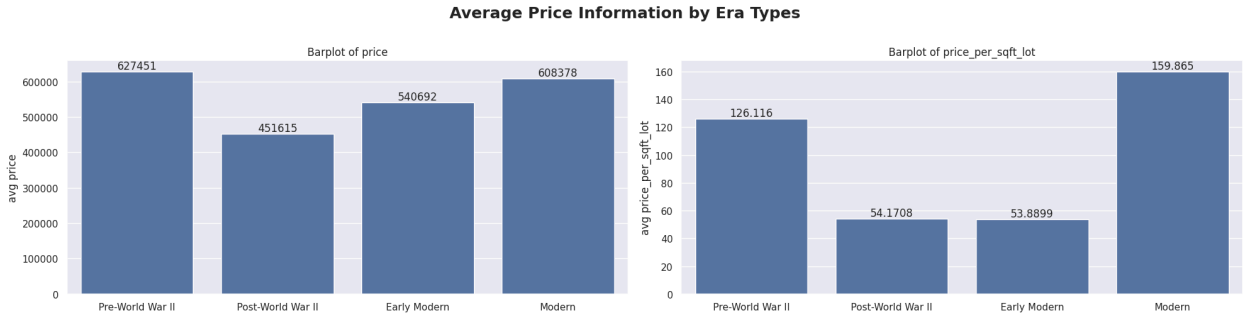


2. era category vs. area info and price



- Terdapat tren kenaikan pada rata-rata luas bangunan dari era Perang Dunia II hingga era modern, yang menunjukkan peningkatan preferensi atau kebutuhan akan ruang lebih besar dalam desain rumah dari waktu ke waktu.
- Di sisi lain, terlihat tren penurunan pada rata-rata luas basement dari era sebelum Perang Dunia II hingga era modern. Kemungkinan besar, di era perang, basement memiliki fungsi tambahan sebagai shelter atau tempat persembunyian, sehingga luas basement cenderung lebih besar pada periode tersebut.

*tanpa mempertimbangkan variabel lain



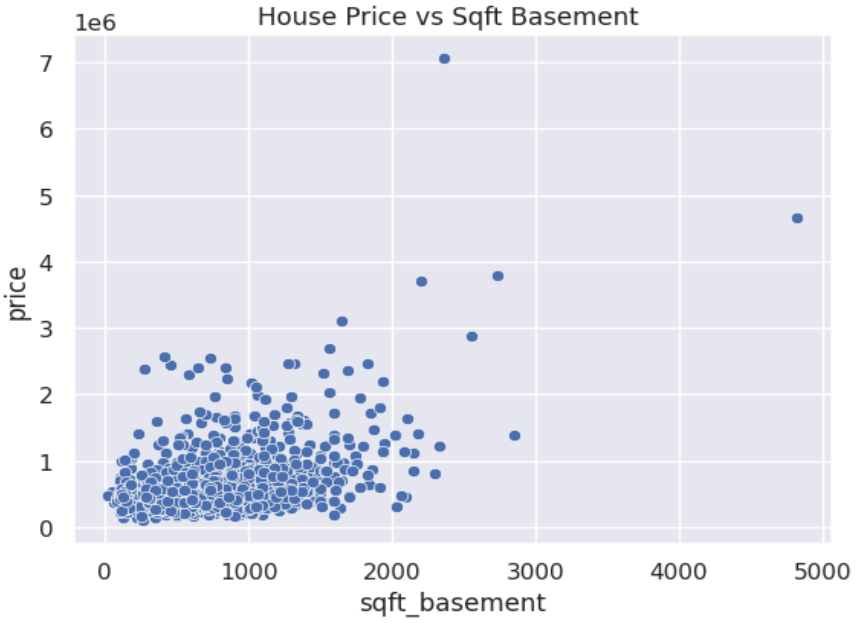
- Terdapat tren kenaikan pada rata-rata harga rumah dari era setelah Perang Dunia II hingga era modern. Namun, rata-rata harga rumah dari era sebelum perang lebih tinggi dibandingkan dengan era lainnya. Hal ini kemungkinan disebabkan oleh nilai historis dan gaya arsitektur unik yang melekat pada rumah-rumah pra-perang, yang sering kali dianggap lebih berharga.
- Sebaliknya, rata-rata harga rumah dari era setelah perang adalah yang paling rendah di antara semua era. Hal ini dapat disebabkan oleh penggunaan konstruksi massal pada periode tersebut, yang menghasilkan gaya arsitektur yang seragam dan sederhana. Selain itu, efek pasca perang mungkin berdampak pada kualitas bahan bangunan, yang bisa lebih terbatas atau efisien dalam biaya.

*tanpa mempertimbangkan faktor lain

Q: Apakah rumah dengan basement memiliki rata-rata harga lebih tinggi dibandingkan dengan rumah tanpa basement?

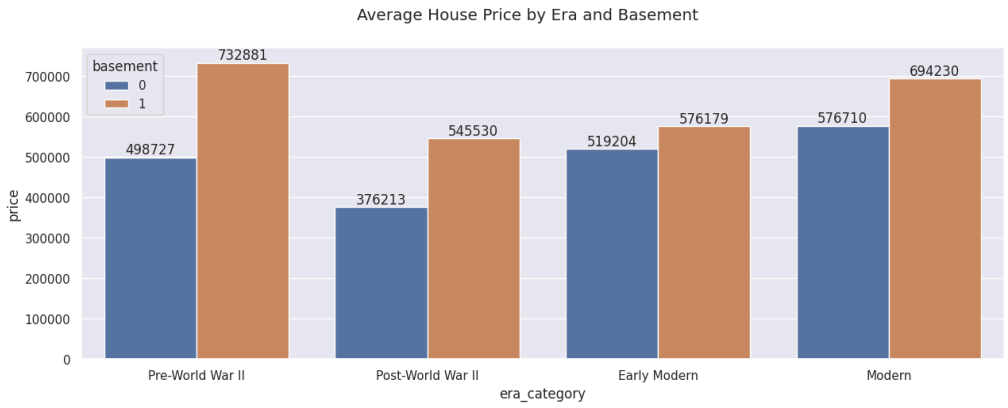
Mari kita telusuri perbedaan harga rata-rata antara rumah dengan basement dan rumah tanpa basement untuk menentukan apakah basement berdampak pada peningkatan nilai rumah.

- Rumah yang memiliki basement memiliki rata-rata harga yang lebih tinggi dibandingkan dengan rumah yang tidak memiliki basement. Hal ini menunjukkan bahwa keberadaan basement dapat memberikan nilai tambah pada properti, kemungkinan karena basement menyediakan ruang tambahan yang bisa digunakan untuk berbagai keperluan, seperti penyimpanan, ruang kerja, atau area rekreasi, yang meningkatkan daya tarik dan nilai rumah di pasar.

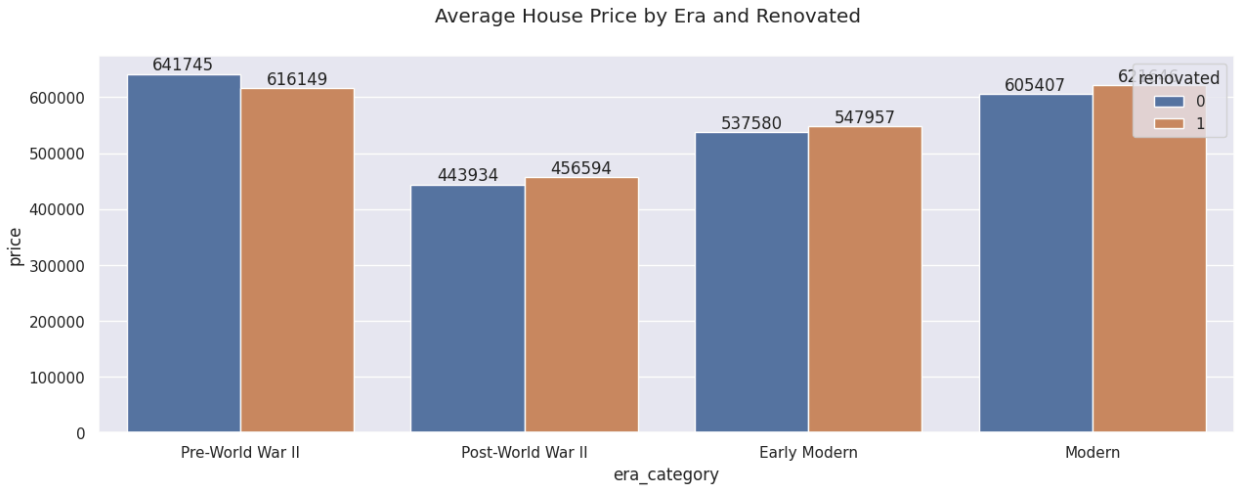


	sqft_basement	price
sqft_basement	1.00	0.44
price	0.44	1.00

- Terdapat korelasi positif sedang antara luas basement dan harga rumah, yang menunjukkan bahwa setiap kenaikan luas basement cenderung diikuti dengan peningkatan harga rumah dalam jumlah sedang. Dari sini, dapat disimpulkan bahwa rumah pada era sebelum perang memiliki harga yang cukup tinggi sebagian karena luas basementnya yang cenderung lebih besar dibandingkan dengan rumah dari era lain. Luas basement yang besar ini kemungkinan menambah nilai, baik sebagai ruang tambahan yang fungsional maupun sebagai fitur unik dari arsitektur klasik, yang turut mendongkrak harga rumah tersebut di pasar.



- Rumah yang tidak direnovasi memiliki rata-rata harga yang sedikit lebih tinggi dibandingkan dengan rumah yang telah direnovasi. Hal ini mungkin disebabkan oleh nilai asli atau karakteristik unik dari rumah yang belum mengalami perubahan, yang dapat dianggap lebih bernilai bagi pembeli tertentu. Selain itu, rumah yang belum direnovasi bisa saja terletak di lokasi yang lebih diinginkan atau memiliki fitur-fitur asli yang diinginkan di pasar properti, sehingga mendongkrak harga rata-ratanya meski tanpa renovasi.

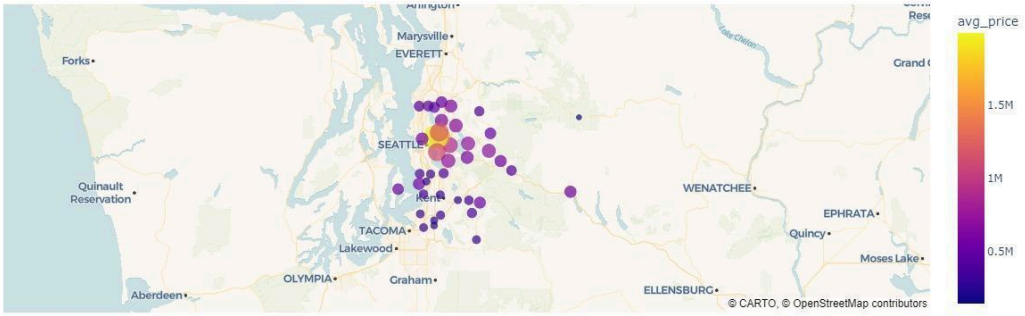


- Dengan mempertimbangkan variabel renovated dan basement, dapat disimpulkan bahwa harga rata-rata rumah era pre-World War II yang tidak direnovasi dan/atau memiliki basement cenderung lebih tinggi dibandingkan dengan rumah yang sudah direnovasi dan tidak memiliki basement. Hal ini kemungkinan disebabkan oleh gaya arsitektur klasik yang masih dipertahankan pada rumah-rumah tersebut, yang memiliki nilai historis dan estetika yang diinginkan di pasar properti. Selain itu, keberadaan basement sebagai ruang tambahan memberikan nilai tambah pada rumah, meningkatkan fungsionalitas dan daya tariknya, yang pada gilirannya meningkatkan harga rumah tersebut.

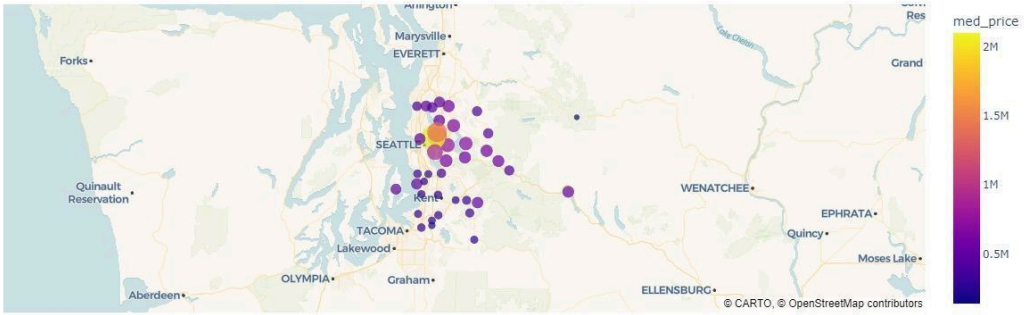
c. Let's take a walk around city

Pada bagian ini kami akan menelusuri lebih dalam sebaran harga dan informasi area terhadap city menggunakan interactive geospatial visualization sebagai berikut:

Average House Price by City

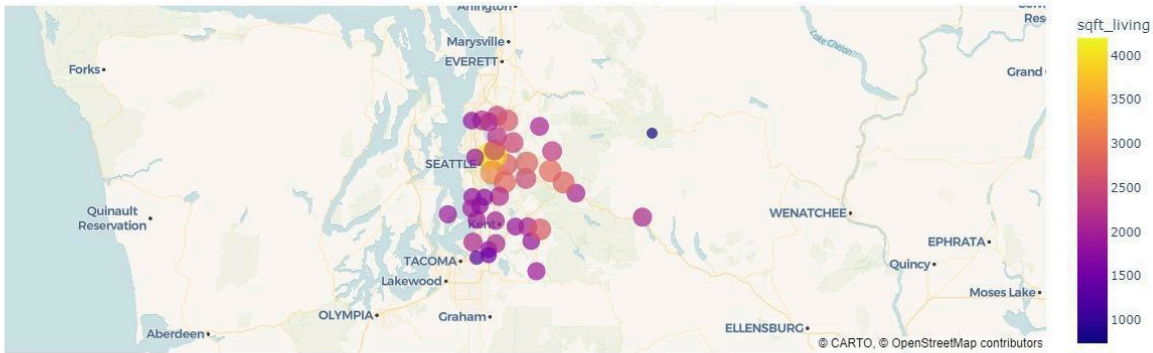


Median House Price by City

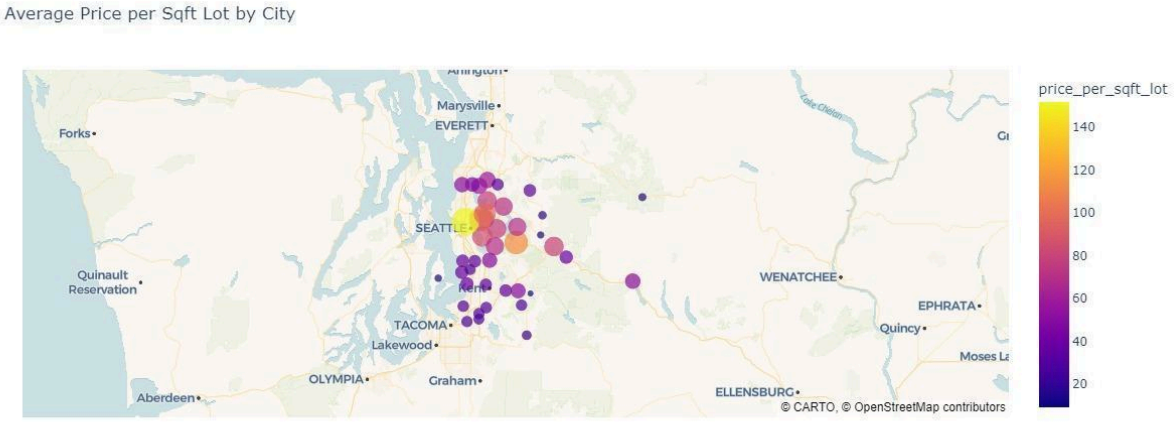


- Berdasarkan analisis rata-rata harga rumah, tiga lokasi dengan harga tertinggi berada di Medina, Yarrow Point, dan Clyde Hill. Hal ini menunjukkan bahwa area tersebut merupakan kawasan premium dengan nilai properti yang tinggi dibandingkan dengan area lainnya. Dapat dilihat pada visualisasi di atas.

Average Living Size by City



- Rata-rata luas bangunan rumah dalam dataset menunjukkan variasi yang cukup signifikan, namun Medina dan Clyde Hill tetap menempati posisi teratas untuk ukuran bangunan terbesar. Hal ini sejalan dengan korelasi positif yang kuat antara luas bangunan dan harga rumah semakin besar luas bangunan, semakin tinggi pula harga rumah di wilayah tersebut. Dapat dilihat pada visualisasi di atas.



- Meskipun pusat kota Seattle tidak memiliki rata-rata harga rumah maupun luas bangunan yang tinggi, kota ini mencatat rata-rata harga tanah per kavling tertinggi dibandingkan kota-kota lainnya. Hal ini menunjukkan bahwa nilai tanah di Seattle sangat tinggi, kemungkinan besar karena daya tarik lokasi strategis dan tingginya permintaan akan lahan di pusat kota. Dapat dilihat pada visualisasi di atas.

d. Are aesthetic aspects important?

Pada bagian deep dive poin d, kami akan mendalami pentingnya variabel-variabel yang terkait dengan aspek estetika, seperti waterfront, view, dan condition rumah. Analisis ini bertujuan untuk memahami apakah faktor-faktor estetika tersebut berpengaruh signifikan terhadap nilai atau daya tarik properti di mata pembeli

1. Waterfront

Dilihat dari dua variabel, yaitu harga rumah dan informasi area rumah, rumah yang dekat dengan area perairan cenderung memiliki nilai yang lebih tinggi, baik dari segi harga rumah maupun luas tanah, dibandingkan dengan rumah yang tidak dekat dengan area perairan. Rumah yang berada di dekat area perairan sering kali dianggap lebih bernilai karena pemandangan yang indah dan akses ke sumber daya alam, yang

membuatnya lebih menarik bagi pembeli. Kategori harga rumah yang dekat dengan area perairan biasanya berada pada kategori moderate ke atas, menunjukkan bahwa lokasi dekat dengan air dapat meningkatkan nilai pasar rumah. Dapat dilihat pada visualisasi di bawah ini:

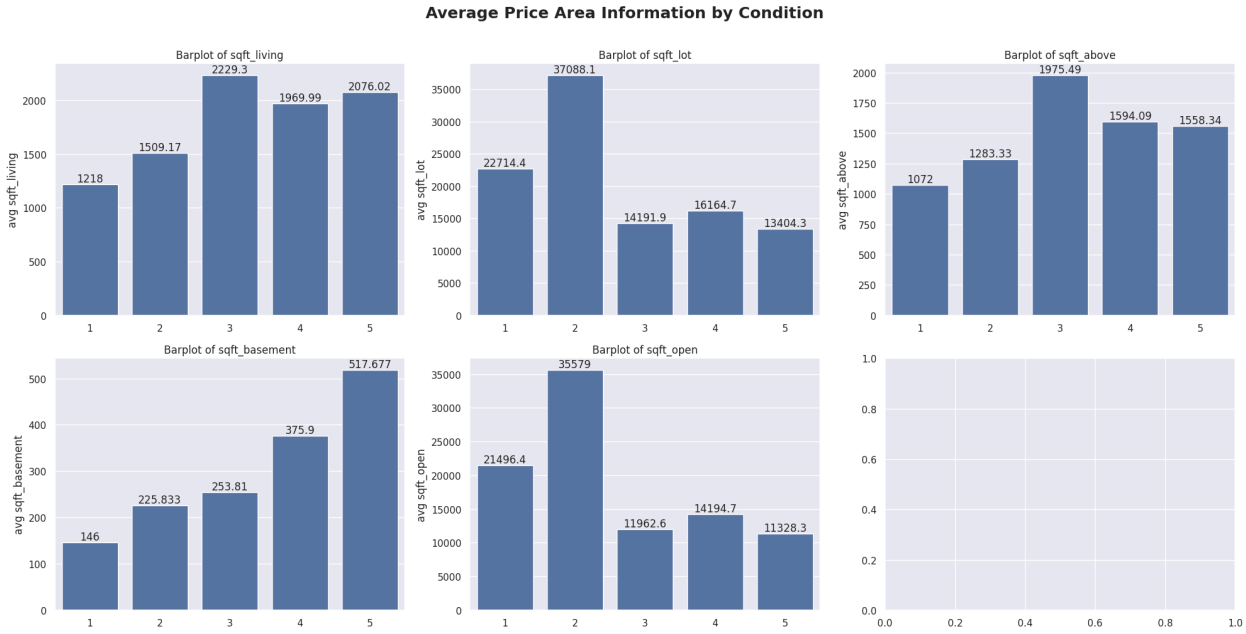
2. View

Terdapat tren kenaikan rata-rata harga rumah pada setiap tingkatan aspek penampilan rumah. Hal ini menunjukkan bahwa rumah dengan penampilan yang lebih baik atau lebih menarik secara visual cenderung memiliki harga yang lebih tinggi. Selain itu, terdapat tren kenaikan pada beberapa area informasi, seperti luas bangunan, luas lantai atas , dan luas basement. Ini menunjukkan bahwa rumah dengan lebih banyak ruang, baik di tingkat bangunan utama, lantai atas, maupun basement, cenderung memiliki harga yang lebih tinggi, yang mencerminkan preferensi pasar terhadap rumah dengan lebih banyak ruang hidup. Dapat dilihat pada visualisasi dibawah ini:

3. Condition

Terdapat tren kenaikan rata-rata harga pada setiap kenaikan kondisi rumah. Artinya, semakin baik kondisi rumah, semakin tinggi pula harga rata-rata rumah tersebut. Hal ini menunjukkan bahwa pembeli cenderung bersedia membayar lebih untuk rumah dengan kondisi yang lebih baik atau lebih terawat. Selain itu, terdapat tren kenaikan pada luas basement. Semakin tinggi kualitas rumah, semakin luas pula basement-nya. Hal ini bisa menunjukkan bahwa rumah dengan kualitas lebih baik seringkali dilengkapi dengan basement yang lebih besar, yang menambah nilai rumah tersebut dan memberikan ruang tambahan untuk berbagai keperluan. Dapat dilihat pada visualisasi dibawah ini:

*tanpa mempertimbangkan variabel lain



e. Anomaly Everywhere!

Pada bagian ini, kami akan mengidentifikasi dan mengungkap anomali-anomali dalam data. Anomali ini dapat mencakup data yang tidak konsisten, nilai ekstrem, atau data yang tampak tidak sesuai dengan pola umum. Mengidentifikasi anomali sangat penting untuk memastikan kualitas dan akurasi analisis berikutnya.

1. House Price Zero Anomaly

Mengacu pada sub-bab 2.2 tentang description report, terdapat rumah dengan harga 0, yang merupakan anomali dan perlu ditelusuri lebih dalam. Berikut adalah statistik deskriptif untuk rumah-rumah dengan harga 0, yang bertujuan untuk memahami karakteristik rumah tersebut dan mencari kemungkinan penyebab dari nilai harga yang tidak wajar ini.

	count	mean	std	min	25%	50%	75%	max
bedrooms	40.00	04.03	1.14	1.00	3.75	4.00	5.00	6.00
bathrooms	40.00	2.82	1.21	1.00	2.19	2.75	3.56	6.25
sqft_living	40.00	2,890.25	1,423.24	720.00	1,982.50	2,810.00	3,682.50	8,020.00
sqft_lot	40.00	16,929.40	29,301.99	4,545.00	6,887.00	9,662.50	17,926.00	188,200.0
floors	40.00	1.54	0.54	1.00	1.00	1.50	2.00	3.00
waterfront	40.00	0.07	0.27	0.00	0.00	0.00	0.00	1.00
view	40.00	0.88	1.52	0.00	0.00	0.00	2.00	4.00
condition	40.00	3.60	0.74	3.00	3.00	3.00	4.00	5.00

sqft_above	40.00	2,424.25	1,364.11	720.00	1,457.50	2,120.00	3,217.50	8,020.00
sqft_basement	40.00	466.00	631.97	0.00	0.00	0.00	920.00	1,950.00
yr_built	40.00	1,973.75	25.81	1,925.00	1,952.75	1,978.00	1,998.75	2,013.00
yr_renovated	40.00	847.40	998.31	0.00	0.00	0.00	1,999.00	2,009.00
price	40.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sqft_open	40.00	14,039.15	29,303.77	1,245.00	5,026.25	7,015.00	12,876.50	186,000.0
price_per_sqft_lot	40.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
basement	40.00	0.40	0.50	0.00	0.00	0.00	1.00	1.00
house_age	40.00	41.25	25.81	2.00	16.25	37.00	62.25	90.00
renovated	40.00	0.42	0.50	0.00	0.00	0.00	1.00	1.00
zip_code	40.00	98,070.95	55.71	98,001.00	98,026.75	98,060.50	98,109.75	98,188.00
lat	40.00	47.54	0.14	47.20	47.47	47.58	47.62	47.76
lng	40.00	-122.20	0.13	-122.34	-122.32	-122.22	-122.14	-121.84

- Spesifikasi rumah dengan harga 0 menunjukkan bahwa data tersebut merupakan anomali, kemungkinan besar disebabkan oleh kesalahan penginputan data. Harga 0 tidak sesuai dengan harga pasar rumah yang seharusnya, sehingga data ini perlu ditinjau lebih lanjut atau diperbaiki sebelum digunakan dalam analisis lebih lanjut. Anomali seperti ini bisa terjadi akibat kesalahan dalam sistem pencatatan atau pengolahan data, dan perlu dibersihkan agar hasil analisis lebih akurat.

2. House Price Anomaly

- Rumah dengan harga melebihi upper fence (yaitu 1.139 juta) menunjukkan bahwa rumah tersebut memiliki nilai median spesifikasi informasi area seperti sqft_living, sqft_lot, sqft_above, dan sqft_basement yang jauh lebih tinggi dibandingkan dengan rumah yang harganya di bawah upper fence. Hal ini menunjukkan bahwa rumah dengan harga lebih tinggi cenderung memiliki ruang yang lebih luas secara keseluruhan, termasuk ruang lantai, tanah, dan basement.
- Selain itu, 50% rumah dengan harga lebih dari 1.139 juta memiliki jumlah kamar tidur lebih dari 4 kamar , dan 25% di antaranya memiliki 5 kamar tidur . Sementara untuk rumah dengan harga di bawah 1.139 juta, 50% diantaranya memiliki lebih dari 3 kamar tidur , dan 25% memiliki lebih dari 4 kamar tidur .
- Namun, yang menarik adalah nilai sqft_lot pada rumah dengan harga kurang dari 1.139 juta memiliki nilai maksimal jauh lebih tinggi dibandingkan dengan

rumah di atas harga upper fence, yaitu 1,074,218 sqft dibandingkan dengan 230,652 sqft . Hal ini menunjukkan bahwa beberapa rumah dengan harga lebih rendah dapat memiliki luas tanah yang jauh lebih besar, meskipun harga totalnya lebih rendah, mungkin karena lokasi atau faktor lainnya.

3. Sqft_lot Anomaly

Insight mengenai anomali yang ditemukan menunjukkan bahwa terdapat rumah dengan luas lahan keseluruhan sebesar 1,074,218 yang memiliki harga di bawah batas upper fence (batas atas distribusi normal). Luas lahan yang sangat besar ini merupakan anomali, karena properti dengan luas lahan ekstrim biasanya memiliki harga yang jauh lebih tinggi. Hal ini mungkin disebabkan oleh kesalahan penginputan data atau properti tersebut memiliki karakteristik khusus yang mempengaruhi harganya. Anomali ini penting untuk ditelusuri lebih lanjut agar tidak mengganggu analisis keseluruhan, didapatkan insight sebagai berikut:

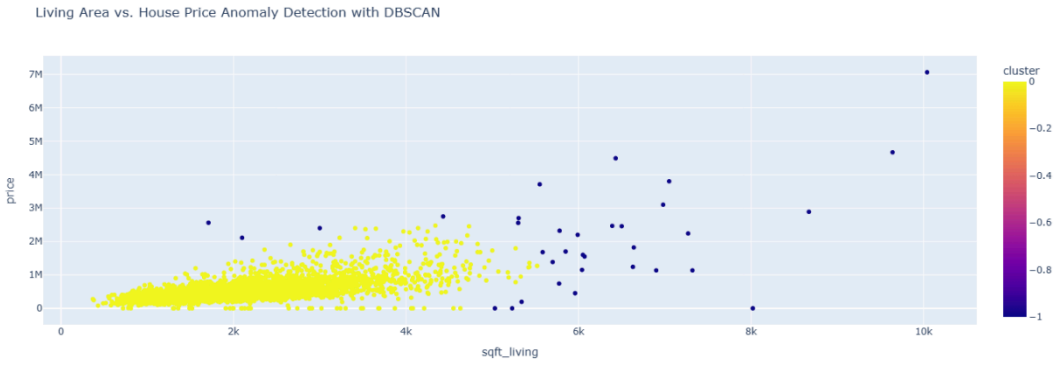
- Rumah ini memiliki spesifikasi sangat baik dengan harga yang relatif murah. Luas lahannya adalah yang terbesar dalam dataset, ditambah luas bangunan, basement, dan lantai atas yang di atas rata-rata. Lokasinya hanya 23 menit dari pusat Seattle, memiliki kondisi maksimal (5 dari skala 1-5), dibangun pada 1931 namun belum direnovasi, serta jumlah kamar tidur dan toilet yang di atas rata-rata. Meskipun spesifikasi rumah hampir semuanya unggul, harga 542.5 ribu sedikit di atas rata-rata (541 ribu), yang menimbulkan pertanyaan mengenai penyebab harganya tetap rendah. Untuk memahaminya lebih dalam, diperlukan kunjungan langsung ke rumah dengan memperhatikan faktor seperti sejarah rumah, kualitas lingkungan, dan kondisi sekitar.
- Rumah tersebut berlokasi di daerah Mirrormont yang termasuk dalam wilayah Issaquah, yang berada di dalam flood plain atau dataran banjir. Berdasarkan artikel di situs City of Issaquah [1], daerah ini memiliki riwayat banjir yang cukup sering, dengan peristiwa besar pada tahun 1986, 1990, 1996, 2009, dan 2020. Banjir tahun 1996 dan 2009 sangat merugikan, menyebabkan kerugian hingga jutaan dolar. Selain itu, Mirrormont juga dikelilingi beberapa pegunungan, yang membuat daerah ini berisiko tinggi terhadap bencana gempa bumi. Menurut situs Augurisk.com [2], Mirrormont memiliki risiko gempa yang tinggi, dengan kategori risiko mencapai 75%. Faktor risiko banjir dan gempa bumi ini bisa menjadi alasan mengapa rumah tersebut dijual dengan harga relatif murah meskipun memiliki spesifikasi yang baik.

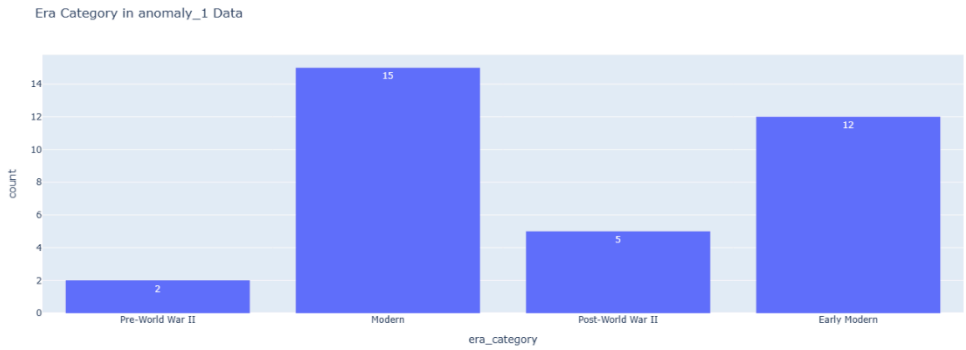
Referensi:

[1] https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.issaquahwa.gov%2FDocumentCenter%2FView%2F24%2FFlood-Brochure%3FbidId%3D&psig=AOvVaw3LaO-9JTe74q6xOGyl3UOv&ust=1730816234025000&source=images&cd=vfe&opi=89978449&ved=0CAYQrpoMahcKEwiQ1ueY78KJAxUAAAAAHQAAAAAQBA

[2] https://www.augurisk.com/neighborhood/washington/issaquah/mirrormont/47.463466013491804/-122.00884475722262

4. living area vs. house price anomaly





- Dari 34 rumah yang terdeteksi sebagai anomali, sekitar 25% memiliki harga di bawah 1.175 juta. Terdapat hanya 3 rumah yang memiliki harga di bawah 1 juta (harga 0 tidak dihitung), sedangkan sisanya lebih dari 1 juta. Dari 31 rumah yang dianalisis (mengabaikan harga 0), sebanyak 28 rumah termasuk dalam kategori harga Luxury.
- Rumah-rumah yang terdeteksi sebagai anomali memiliki rata-rata luas bangunan hampir 3 kali lebih besar dibandingkan dengan rumah pada umumnya. Rata-rata luas bangunan rumah anomali adalah 6,066.85 sqft, sedangkan rata-rata rumah secara keseluruhan adalah 2,101.21 sqft.
- Sebagian besar rumah yang terdeteksi sebagai anomali memiliki gaya arsitektur early modern hingga modern, yang mungkin mempengaruhi harga dan ukuran rumah yang tidak sesuai dengan pola rumah lainnya.

5. Average House Price in 1935 dan 1936 Anomaly

Insight ini menunjukkan bahwa ada beberapa faktor yang menyebabkan harga rumah pada tahun 1935 dan 1936 memiliki perbedaan signifikan dibandingkan dengan

tahun-tahun lainnya, khususnya terkait dengan rumah-rumah di pusat kota Seattle yang terdeteksi sebagai anomaly. Berikut adalah beberapa poin penting:

- Semua rumah anomaly terdeteksi berlokasi di pusat kota Seattle, yang kemungkinan besar berkontribusi pada harga yang lebih tinggi dibandingkan dengan lokasi lainnya.
- Terdapat 5 rumah yang dibangun pada tahun 1935 (3 rumah) dan 1936 (2 rumah).
- Rumah yang dibangun pada tahun 1935 memiliki harga antara 535 ribu hingga 2.4 juta, dengan dua rumah yang memiliki harga sangat tinggi. Rumah yang dibangun pada tahun 1936 memiliki harga antara 840 ribu hingga 1.37 juta , yang juga menyebabkan rata-rata harga menjadi lebih tinggi.
- Dua rumah dengan harga tertinggi pada tahun 1935 dan 1936 memiliki luas lahan keseluruhan yang sangat besar, yaitu 13 ribu sqft dan 16 ribu sqft dengan harga 2.4 juta dan 1.37 juta, masing-masing.
- Rumah yang memiliki harga tertinggi tersebut juga memiliki luas bangunan yang cukup besar , sekitar 4.7 ribu sqft, yang memberikan kontribusi pada harga tinggi.

2.4. Data quality assessment

Berikut adalah insight yang didapat dari penilaian kualitas data dalam proses data quality assessment. Kami telah mengidentifikasi beberapa masalah yang perlu diperbaiki untuk memastikan data yang digunakan dalam pemodelan lebih akurat dan konsisten. Berikut adalah penjelasan untuk masing-masing poin terkait masalah kualitas data yang ditemukan:

- Nilai kosong pada variabel `price_category`dimana variabel ini memiliki nilai kosong karena ada rumah dengan harga 0. Karena variabel ini bergantung pada harga rumah untuk kategori harga, kami akan menghapus data rumah dengan harga 0 dan menghapus nilai kosong pada `price_category` untuk keperluan pemodelan.
- Memastikan tidak ada inkonsistensi pada variabel bertipe objek dengan memeriksa perbedaan penulisan besar-kecil huruf. Hasilnya menunjukkan bahwa semua data dalam variabel tersebut sudah konsisten, tanpa adanya perbedaan penulisan yang tidak sesuai.
- Memastikan bahwa tidak ada nilai duplikat dalam dataset. Setelah dilakukan pemeriksaan, ternyata tidak ditemukan data duplikat, sehingga semua nilai sudah unik dan siap untuk digunakan dalam analisis selanjutnya.
- Variabel `lng` (longitude), variabel ini menunjukkan koordinat geografis (longitude) dari lokasi rumah. Nilai negatif pada variabel ini tidak masalah karena menunjukkan posisi rumah di sisi barat Bumi.

- Variabel `cluster_living_price’, variabel ini menunjukkan apakah rumah tergolong anomali atau tidak, dengan nilai 0 berarti bukan anomali dan -1 berarti anomali. Karena variabel ini melibatkan informasi harga rumah (target), kami tidak akan menggunakannya dalam pemodelan.
- Variabel `sqft_open` (luas lahan kosong), variabel ini dihitung dengan mengurangi luas bangunan dari luas lahan keseluruhan. Jika hasilnya negatif, berarti ada rumah dengan luas bangunan lebih besar daripada luas tanahnya. Kami akan meneliti lebih lanjut dan mengasumsikan bahwa ini terjadi karena kesalahan input data, sehingga data ini akan kami hapus.

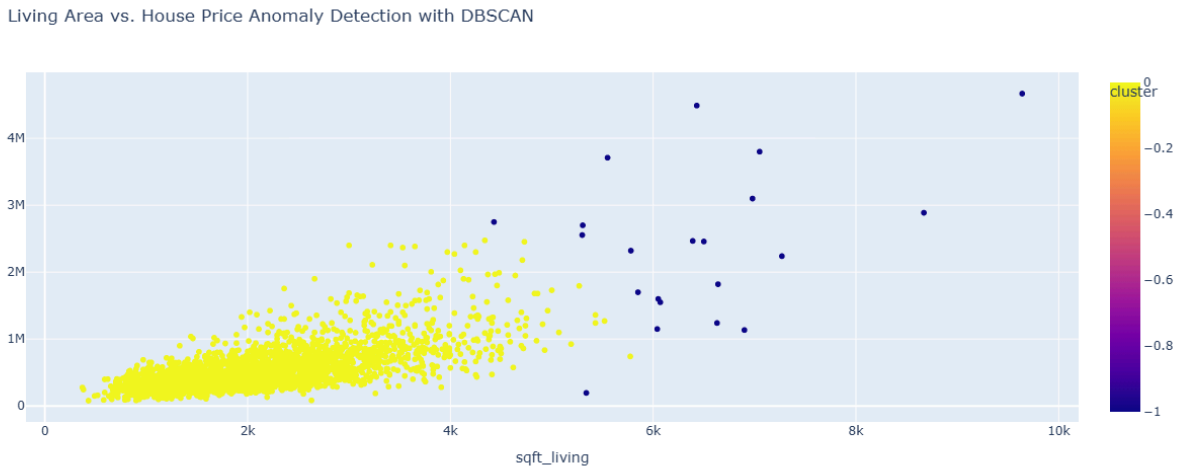
Data Preparation

3.1. Select Data

Seleksi data yang kami lakukan adalah seleksi pada row-level yaitu menghapus baris dengan nilai sqft_living > sqft_lot karena kami asumsikan rumah yang luas bangunannya lebih luas dibandingkan luas keseluruhan lahan merupakan kesalahan dalam penginputan data. Seleksi baris berikutnya adalah rumah dengan harga 0 dengan alasan kami asumsikan rumah tersebut tidak relevan untuk pemodelan. Rumah dengan harga maksimal yaitu sekitar 7 juta juga kami hapus dengan asumsi distribusi data dari rumah tersebut hanya terdapat pada satu rumah saja dan sangat jauh dari data sebelumnya.

3.2. Clean Data

Pembersihan data kami lakukan pada data outlier yang terdapat pada feature sqft_living dan target price. Kedua variabel tersebut mempunyai korelasi positif yang kuat sehingga anomaly pada kedua data tersebut cukup mewakili keseluruhan data. Metode anomaly detection yang digunakan adalah DBSCAN density-based clustering yang terbukti mampu mendeteksi anomaly dengan tapat.



3.3. Construct Data

Terdapat tiga feature engineering yang kami lakukan pada tahap data preparation. Pertama, pembuatan feature baru basement yang didapatkan dari feature sqft_basement dengan nilai 0 untuk yang tidak memiliki basement dan 1 untuk rumah yang mempunyai basement hal ini kita lakukan dengan melihat distribusi rumah yang tidak memiliki basement lebih dari setengah dari keseluruhan data. Feature kedua yang kami buat adalah feature renovated, seperti halnya dengan feature basement. Feature renovated dibuat dengan tujuan untuk membedakan rumah yang telah direnovasi dan belum pernah direnovasi. Terakhir, feature era_category. Feature era_category merupakan feature yang tercipta dari kategorisasi dari feature yr_built dengan pembagian kategori berdasarkan era gaya arsitektur. Dibagi menjadi 4 bagian yaitu Pre-World War II yaitu era sebelum perang dengan gaya arsitektur klasik. Era tersebut dari tahun 1900-1939. Kategori berikutnya adalah era setelah perang dunia (Post-World War II) dengan pembagian tahun 1940-1969 dengan gaya arsitektur masal. Era berikutnya adalah era modern awal atau Early Modern dengan pembagian tahun rumah dibangun pada 1970-1999. Terakhir yaitu rumah dengan gaya arsitektur modern dibangun pada tahun 2000-2014.

Feature street merupakan feature yang menunjukan alamat jalan dari setiap rumah. Feature ini memiliki variasi yang sangat banyak sehingga dirasa kurang informatif. Namun setelah melakukan riset, feature tersebut mampu memberikan informasi yang cukup bagus terhadap harga rumah. Dengan memanfaatkan metode feature extraction TF-IDF pada text processing, didapatkan sejumlah 13 feature penting dari feature street. Dari hasil ekstraksi tersebut akan coba kami kombinasikan dengan feature lain.

3.4. Integrate Data

Tahap berikutnya adalah mengintegrasikan data utama dengan data lain untuk memperkaya informasi dari data. Data baru yang kami tambahkan adalah informasi longitude dan latitude dari feature city. Penggabungan data ini juga dimaksudkan sebagai feature encoding dari feature city yang memiliki kardinalitas yang cukup tinggi untuk mencegah cursed of dimensionality yang dihasilkan oleh metode encoding tradisional yaitu one-hot-encoding untuk categorical data types.

3.5. Reformatted Data

Tahap reformatted data bertujuan untuk mengubah tipe data dari features yang bertipe data kurang sesuai. Feature yang menjadi fokus kami adalah statezip yang berisikan informasi state atau negara bagian dan zip atau kode pos. Dalam dataset tersebut semua record berada pada negara bagian Washington state. Sehingga informasi state yaitu "WA" akan kami hilangkan dan hanya mengambil zipcode nya saja. Karena tipe data awal dari feature statezip adalah object sehingga perlu dilakukan casting tipe data menjadi integer.

3.6. Data Transformation

Bagian data transformation bertujuan untuk mentransformasi data menjadi bentuk yang sesuai dengan keperluan model. Transformasi yang pertama kita lakukan adalah mean encoding pada feature era_category. Mean encoding merupakan jenis encoding yang mengambil nilai rata-rata suatu kategori berdasarkan variable target. Kami menerapkan fitting pada data training untuk menghasilkan mapping kategori dan rata-rata harga setiap kategorinya. Setelahnya kami transform ke feature era_category pada data training dan data testing. Transformasi kedua yang kami lakukan adalah log transform pada beberapa feature yang berdistribusi skewed yaitu features pada area information seperti sqft_living, sqft_lot, dan sqft_above.

3.7. Feature Selection

Pada bagian select data kami telah melakukan penghapusan beberapa baris dengan melihat tingkat relevansi record. Pada tahan feature selection kami menerapkan metode wrapper dengan algoritma RFECV. Wrapper method merupakan metode dalam feature selection dengan mencoba beberapa kombinasi dari feature dan langsung mengevaluasinya ke target dengan mengukur metrik evaluasi yang sesuai. Metode ini sangat memberikan pengaruh dalam performa model, namun limitasi dari metode ini adalah konsumsi waktu yang cukup lama karena ada beberapa algoritma wrapper yang mencoba banyak kombinasi sekaligus. RFECV merupakan salah satu algoritma pada wrapper method. RFECV (Recursive Feature Elemination with Cross Validation). Algoritma tersebut mencari kombinasi feature secara rekursif dengan mengeliminasi feature paling tidak penting pada setiap iterasinya sampai

didapatkan feature yang paling penting buat model. Feature selection kami lakukan dengan tiga kali percobaan. Percobaan pertama menghasilkan 20 feature didapat dari penggabungan feature utama dengan feature extraction dari street. Percobaan kedua menghasilkan 8 feature didapatkan dari feature awal. Terakhir menggabungkan 8 feature pada percobaan kedua dengan feature extraction dari street, didapatkan 21 feature

Modeling

4.1. Model selection and experimentation

Model yang digunakan pada project ini adalah tree-based, boosting, dan ensemble model. Dengan rincian model sebagai berikut, Decision Tree, Gradient Boosting, AdaBoost, XGBoost, lightGBM, Random Forest dan Voting. Percobaan dilakukan dengan beberapa kali perlakuan. Pertama percobaan pada data preparation yaitu data dengan semua feature dan data dengan selected features. Terdapat variasi feature dari hasil feature selection seperti yang dijelaskan pada bab data preparation sub bab feature selection.

Perlakuan pada data preparation berikutnya adalah dengan dan tanpa melalui proses penghapusan outlier, imputasi dan penghapusan missing value. Perlakuan berikutnya kami lakukan pada proses modelling menggunakan log transform pada target. Kami membuat dua prosedur yaitu model_training dan model_training_log_transform. Prosedur pertama digunakan untuk melatih model dan mengevaluasinya tanpa transformasi. Prosedur yang kedua digunakan untuk melatih model dan melihat evaluasinya dengan target yang ditransformasi log.

Berikutnya kami melakukan percobaan hyperparameter tuning menggunakan library optuna karena memiliki fitur optimasi yang baik dibandingkan algoritma hyperparameter tuning lainnya. Pada proses tuning kami gunakan variasi data pada perlakuan data preparation diatas, dihasilkan beberapa tuned_models. Tuned_models merupakan variabel yang berisi komparasi model sebelum dan sesudah melalui hyperparameter tuning beberapa kali.

4.2. Model performance evaluation

Evaluasi model yang kami gunakan pada kasus house price prediction adalah Mean Absolute Error (MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE). Dengan nilai RMSE sebagai evaluasi utama karena mempertimbangkan penalti pada error yang lebih besar.

4.3. Model comparison and selection

Setelah berbagai percobaan dilakukan model XGBoost menghasilkan performa yang lebih baik dibandingkan dengan model-model lainnya. Komparasi performa pada setiap beberapa percobaan dapat diakses melalui link google sheet berikut:

All features

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	21.392,62	92.058,81	853.172.115,33	25.336.249.976,92	29.209,11	159.173,65	0,99	0,73	0,87
Decision Tree	20,12	132.888,48	563.420,77	55.547.092.959,59	750,61	235.684,31	1	0,4	0,06
Random Forest	35.868,06	97.197,16	3.764.366.389,85	28.950.315.659,57	61.354,43	170.147,92	0,96	0,69	4,95
LightGBM	52.046,31	93.221,16	5.898.400.472,40	27.710.049.845,12	76.801,04	166.463,36	0,94	0,7	0,2
Gradient Boosting	80.190,73	95.156,71	14.867.224.713,74	26.870.857.373,05	121.931,23	163.923,33	0,85	0,71	1,22
Ada Boost	187.722,01	198.705,09	50.140.670.673,62	59.270.816.575,32	223.921,13	243.455,98	0,49	0,36	0,69
Voting	35.431,86	89.832,66	2.471.961.903,62	25.530.024.023,49	49.718,83	159.781,18	0,98	0,73	1,79

All features log transform

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	22.485,58	94.224,95	1.204.734.441,08	31.423.777.170,71	34.709,28	177.267,53	0,99	0,66	2,14
Decision Tree	20,12	123.399,60	564.396,50	41.333.988.744,99	751,26	203.307,62	1	0,56	0,13
Random Forest	36.160,83	96.793,81	4.481.986.142,99	30.520.831.153,26	66.947,64	174.702,12	0,95	0,67	4,87
LightGBM	52.226,12	89.930,05	7.254.399.214,35	26.930.535.329,05	85.172,76	164.105,26	0,93	0,71	0,34
Gradient Boostin	81.728,47	97.010,53	18.285.431.864,67	30.620.412.880,36	135.223,64	174.986,89	0,82	0,67	1,64
Ada Boost	119.368,97	118.602,12	36.948.111.340,81	41.583.481.401,68	192.218,92	203.920,28	0,63	0,55	0,85
Voting	36.305,27	90.353,86	3.279.459.541,00	28.675.408.633,85	57.266,57	169.338,15	0,97	0,69	5,65

Selected features 2

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	30.465,53	93.466,16	1.962.663.637,20	28.242.273.956,35	44.301,96	168.054,38	0,98	0,7	1,4
Decision Tree	20,12	116.760,16	563.420,77	35.903.415.666,43	750,61	189.481,97	1	0,62	0,02
Random Forest	36.338,56	94.155,99	3.878.393.574,12	26.920.538.926,10	62.276,75	164.074,80	0,96	0,71	1,29
LightGBM	62.417,75	89.115,79	9.096.989.333,10	26.130.158.198,43	95.378,14	161.648,25	0,91	0,72	0,07
Gradient Boosting	83.067,98	94.300,47	16.751.764.122,54	26.430.904.377,53	129.428,61	162.575,84	0,83	0,72	0,4
Ada Boost	151.494,02	157.396,41	38.593.711.586,65	46.299.769.356,51	196.452,82	215.173,81	0,61	0,5	0,11
Voting	45.324,16	88.366,25	4.368.461.927,13	26.039.234.376,62	66.094,34	161.366,77	0,96	0,72	0,22

Selected features 2 log transform

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	32.157,54	92.142,50	2.671.619.248,30	28.048.096.342,23	51.687,71	167.475,66	0,97	0,7	0,15
Decision Tree	20,12	123.005,74	564.396,50	44.271.031.939,58	751,26	210.406,82	1	0,53	0,02
Random Forest	37.127,18	94.845,19	4.635.068.729,46	28.199.095.424,22	68.081,34	167.925,86	0,95	0,7	1,59
LightGBM	63.446,07	90.069,27	10.856.754.681,42	27.736.442.497,02	104.195,75	166.542,61	0,89	0,7	0,24
Gradient Boostin	85.907,25	97.023,65	20.269.694.125,60	32.481.845.450,08	142.371,68	180.227,21	0,8	0,65	0,82
Ada Boost	121.284,89	121.991,90	38.663.207.442,52	42.592.167.944,43	196.629,62	206.378,70	0,61	0,54	0,46
Voting	46.604,28	88.577,48	5.593.794.153,26	26.742.068.408,87	74.791,67	163.530,02	0,94	0,71	1,46

Selected features 3

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	29.825,35	93.568,17	1.973.585.946,65	26.696.570.868,92	44.425,06	163.390,85	0,98	0,71	2,9
Decision Tree	20,12	123.257,01	563.420,77	45.684.695.791,97	750,61	213.739,79	1	0,51	0,1
Random Forest	35.958,84	94.407,42	3.752.287.828,15	27.512.904.296,35	61.255,92	165.870,14	0,96	0,71	5,27
LightGBM	60.865,22	88.379,39	8.434.169.110,39	25.560.357.702,17	91.837,73	159.876,07	0,91	0,73	0,51
Gradient Boosting	83.059,77	93.211,40	16.667.344.249,79	26.406.894.529,50	129.102,07	162.501,98	0,83	0,72	3,97
Ada Boost	175.559,31	185.984,06	46.727.755.025,12	57.408.053.074,94	216.166,04	239.599,78	0,53	0,38	1,41
Voting	43.903,75	88.823,76	4.129.062.127,05	25.126.734.140,75	64.257,78	158.514,14	0,96	0,73	2,4

Selected features 3 log transform

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	29.719,89	90.483,23	2.290.073.677,74	26.023.479.534,39	47.854,71	161.317,95	0,98	0,72	2,39
Decision Tree	20,12	124.295,23	564.396,50	46.767.023.820,16	751,26	216.256,85	1	0,5	0,06
Random Forest	36.609,56	95.225,42	4.572.990.475,64	29.353.258.804,36	67.623,89	171.327,93	0,95	0,69	3,32
LightGBM	60.742,22	89.905,30	9.852.872.921,16	27.276.992.933,95	99.261,64	165.157,48	0,9	0,71	0,17
Gradient Boostin	85.201,59	94.589,46	20.172.033.923,93	29.735.884.574,64	142.028,29	172.440,96	0,8	0,68	0,87
Ada Boost	120.837,94	120.012,14	38.210.872.808,43	41.172.634.118,57	195.476,02	202.910,41	0,61	0,56	0,7
Voting	44.141,68	87.819,38	4.963.543.672,28	25.373.627.400,54	70.452,42	159.291,01	0,95	0,73	0,92

Tuned models using selected features 3, log transform in feature, without log transform in target

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Training Time (seconds)
XGBoost	29.825,35	93.568,17	1.973.585.946,65	26.696.570.868,92	44.425,06	163.390,85	0,98	0,71	0,18
XGBoost (Tuned)	58.153,71	86.853,60	7.602.320.101,63	22.987.747.891,22	87.191,28	151.617,11	0,92	0,75	0,41
XGBoost (Tuned 2)	54.422,57	88.205,51	6.705.239.961,11	23.568.167.538,13	81.885,53	153.519,27	0,93	0,75	0,12
XGBoost (Tuned 3)	67.737,99	88.917,36	10.993.719.307,73	23.844.840.324,20	104.850,94	154.417,75	0,89	0,74	0,79
XGBoost (Tuned 4)	63.413,05	87.225,00	9.454.514.852,75	23.089.533.780,06	97.234,33	151.952,41	0,9	0,75	0,99
Decision Tree	20,12	123.257,01	563.420,77	45.684.695.791,97	750,61	213.739,79	1	0,51	0,05
Decision Tree (Tuned)	84.090,64	109.795,45	22.699.014.160,37	37.940.315.874,35	150.661,92	194.782,74	0,77	0,59	0,33
Decision Tree (Tuned 2)	102.282,12	118.109,18	26.693.851.490,03	38.970.233.597,57	163.382,53	197.408,80	0,73	0,58	0,01
Random Forest	35.958,84	94.407,42	3.752.287.828,15	27.512.904.296,35	61.255,92	165.870,14	0,96	0,71	1,79
Random Forest (Tuned)	67.047,99	97.315,54	11.441.435.937,71	27.950.092.885,55	106.964,65	167.182,81	0,88	0,7	2,84
LightGBM	60.865,22	88.379,39	8.434.169.110,39	25.560.357.702,17	91.837,73	159.876,07	0,91	0,73	0,09
LightGBM (Tuned)	45.505,88	89.220,46	4.914.238.593,58	25.140.922.270,64	70.101,63	158.558,89	0,95	0,73	1,46
LightGBM (Tuned 2)	74.265,22	92.342,07	13.650.153.523,51	27.514.093.332,75	116.833,87	165.873,73	0,86	0,71	0,25
Gradient Boosting	83.059,77	93.211,40	16.667.344.249,79	26.406.894.529,50	129.102,07	162.501,98	0,83	0,72	0,7
Voting	60.484,39	85.625,98	8.477.387.895,95	23.165.139.911,52	92.072,73	152.200,99	0,91	0,75	1,76

Setelah melakukan beberapa percobaan didapatkan model terbaik dengan treatmenet pada data preparation yaitu penghapusan nilai 0 pada harga rumah, penghapusan outlier dengan DBSCAN, log transform pada feature sqft_lot, sqft_living, dan sqft_above. Feature yang digunakan adalah selected_features_3 tanpa transformasi log pada target, dan dengan model XGBoost Tuned Pertama

Evaluation

5.1. Model evaluation on the test set

Evaluasi model pada test set menunjukan hasil cukup baik dengan nilai MAE 86,853 dan RMSE 151,617.11. Namun hasil dari MAE tersebut masih berada di kisaran 16% dari rata-rata harga sehingga model perlu ditingkatkan lagi performanya. Beberapa Langkah yang bisa dilakukan untuk meningkatkan performa model adalah sebagai berikut:

a) Model masih gagal memprediksi rumah dengan harga tinggi yang berada di kota Issaquah dan Tukwila karena rata-rata harga rumah di sana cenderung rendah. Di kota Issqruah dipengaruhi oleh faktor bencana alam. Untuk kota Tukwila memiliki harga rata-rata yang cukup rendah yaitu sekitar 200 juta. Sehingga perlu adanya penyesuain data dengan menambahkan data rumah yang berlokasi di kedua kota tersebut dengan kelas harga Luxury

b) Error extreme terjadi pada data harga rumah sangat rendah dan sangat tinggi. Sehingga perlu tambahan data untuk harga rumah dengan dua kategori tersebut.