| **LAPORAN UTS DASI**  **Proyek UTS** |
| --- |
|  |
| | **11423012** | **Esticka Priscila Sibarani** | | --- | --- | | **11423013** | **Yosyda Andaresta** | | **11423033** | **Rino Jeremia Christopher Gultom** | |  |  | |  |  | |
| **INSTITUT TEKNOLOGI DEL**  **FAKULTAS VOKASI**  **2025** |

# **BAB I – PENDAHULUAN**

## **1.1 Latar Belakang**

Perkembangan teknologi digital dan kemajuan dalam bidang data science telah mengubah cara manusia memahami, menganalisis, dan mengambil keputusan berdasarkan data. Salah satu bidang yang mendapatkan dampak besar dari analisis berbasis data adalah industri properti. Dalam industri ini, harga rumah dan properti sangat dinamis karena dipengaruhi oleh berbagai faktor seperti lokasi, luas tanah, jumlah kamar, tipe bangunan, akses terhadap fasilitas umum, hingga kondisi ekonomi dan sosial di suatu wilayah. Dengan jumlah data yang sangat besar, metode konvensional untuk menilai harga properti menjadi kurang efektif dan tidak mampu mengidentifikasi pola kompleks serta hubungan non-linear antar variabel. Oleh karena itu, analisis berbasis data science menggunakan bahasa pemrograman Python menjadi salah satu pendekatan yang efisien, objektif, dan mampu menangani kompleksitas data properti modern.

Dataset yang digunakan dalam proyek ini, yaitu Melbourne Housing Dataset, berisi lebih dari 13.000 data penjualan rumah di wilayah Melbourne, Australia. Melbourne, sebagai salah satu kota dengan pasar properti paling dinamis di Australia, menawarkan studi kasus yang ideal dengan keragaman demografis dan geografis yang tinggi. Dataset ini memuat 21 fitur yang mencakup informasi seperti lokasi (suburb dan postcode), jumlah kamar (rooms dan bedrooms), ukuran tanah (land size), tipe rumah, harga jual, hingga koordinat geografis. Dengan kompleksitas dan keragaman variabelnya, dataset ini sangat ideal untuk dieksplorasi menggunakan pendekatan data science melalui pustaka Python seperti *pandas*, *matplotlib*, *seaborn*, dan *scipy* yang memungkinkan peneliti melakukan data preprocessing, visualization, hingga statistical analysis secara komprehensif.

Melalui analisis data ini, diharapkan dapat diperoleh pemahaman mendalam mengenai faktor-faktor yang paling berpengaruh terhadap harga jual rumah serta pola distribusi dan hubungan antar variabel tersebut. Proyek ini menjadi sarana penerapan konsep-konsep data science seperti data collection, data preprocessing, data visualization, dan statistical analysis dalam konteks dunia nyata. Hasil dari analisis ini diharapkan dapat memberikan insight yang berguna bagi masyarakat umum, peneliti, maupun pelaku industri properti dalam memahami dinamika pasar perumahan di Melbourne. Dengan demikian, proyek data science ini tidak hanya menjadi latihan teknis dalam pengolahan data, tetapi juga bentuk penerapan nyata dari kemampuan analisis data untuk mendukung pengambilan keputusan berbasis informasi (data-driven decision making) khususnya dalam bidang properti dan perumahan.

## **1.2 Rumusan Masalah**

1. Faktor-faktor apa saja yang berpengaruh terhadap harga rumah di Melbourne?
2. Bagaimana proses pengolahan dan visualisasi data dapat membantu memahami pola dalam dataset?
3. jenis uji statistik apa yang dapat digunakan untuk mengidentifikasi hubungan antar variabel dalam data properti tersebut?

## **1.3 Tujuan**

1. Menganalisis faktor-faktor yang mempengaruhi harga rumah di Melbourne
2. Melakukan pengolahan, pembersihan, dan visualisasi data menggunakan python
3. Menyajikan hasil analisis melalui grafik dan uji statistik untuk menemukan hubungan antar variabel.

# **BAB II – METODE**

## **2.1 Data Collection**

Tahap pertama dalam proyek ini adalah pengumpulan data. Dataset yang digunakan dalam penelitian ini bernama *melb\_data.csv* (*Melbourne Housing Snapshot*). Dataset ini diambil dari sumber terbuka di platform Kaggle, yang merupakan salah satu situs penyedia dataset publik terbesar dan terpercaya.

Dataset ini berisi informasi mengenai data properti di Melbourne, Australia, dan sering digunakan dalam pembelajaran data science karena struktur datanya yang lengkap serta konteksnya yang nyata. Dataset ini memiliki sekitar 13.000 baris data (*rows*) dan 21 kolom (*features*).

Link dataset : [Melbourne Housing Snapshot](https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot/code)

Alasan pemilihan dataset ini adalah:

1. Dataset bersifat nyata (*real* data), bukan data *dummy* atau hasil simulasi.
2. Memiliki banyak variabel yang relevan, seperti harga properti, luas tanah, luas bangunan, lokasi, dan tipe properti, sehingga dapat digunakan untuk berbagai jenis analisis.
3. Dataset mudah diakses dan dapat diolah menggunakan bahasa pemrograman *Python* tanpa proses *scrapin*g tambahan.

## **2.2 Pengecekan Awal Dataset**

Tahap selanjutnya setelah pengumpulan data adalah pemeriksaan awal terhadap dataset. Tahapan ini dilakukan untuk memastikan kualitas data sebelum masuk ke tahap *preprocessing* dan analisis lebih lanjut. Pemeriksaan awal ini mencakup pengecekan nilai null dan *missing values*, pengecekan nilai kosong pada setiap kolom, pengecekan nilai yang duplikat.

Sebelum melakukan pemeriksaan awal terhadap dataset, langkah pertama yang dilakukan adalah pendefinisian *library* dan pemanggilan dataset ke dalam lingkungan kerja *python*. Tahapan ini sangat penting karena semua proses pengolahan dan analisis selanjutnya bergantung pada data yang dimuat.

Pada tahap ini digunakan *library* pandas. *Library* pandas merupakan salah satu *library* *python* yang paling sering digunakan dalam bidang analisis data.

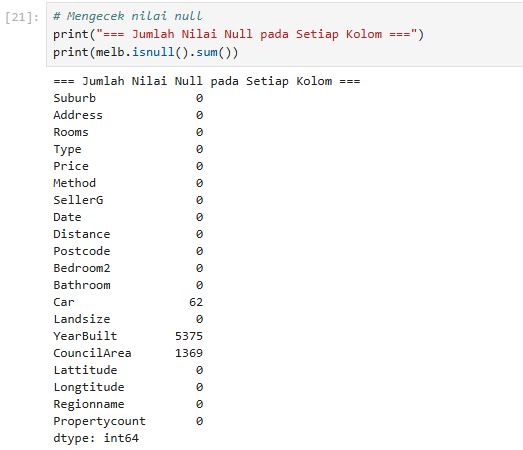


Kemudian dilakukan pemanggilan dataset ke dalam lingkungan kerja *python* dengan menggunakan fungsi read\_csv() dari *library* pandas.



### **2.2.1 Pengecekan Nilai Null dan Missing Values**

Setelah dataset berhasil dimuat dilakukan pengecekan atau pemeriksaan nilai null dan missing values pada setiap kolom data. Pemeriksaan ini bertujuan untuk mengetahui apakah terdapat data yang hilang atau tidak terisi pada dataset.



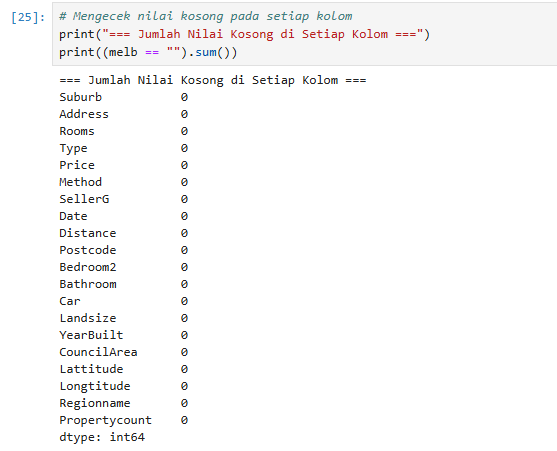
dari hasil pengecekan, diketahui bahwa beberapa kolom memiliki jumlah nilai null yang cukup banyak, seperti :

* YearBuilt sebanyak 5.375 baris
* councilArea sebanyak 1.369 baris
* Car sebanyak 62 baris

Kolom lainnya menunjukkan nilai 0, yang berarti tidak ada data yang hilang atau bernilai null dari kolom tersebut.

### **2.2.2 Pengecekan Nilai Kosong pada setiap kolom**

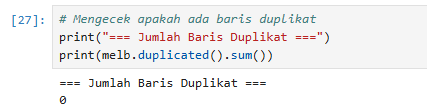
Tahapan berikutnya dalam proses pengecekan adalah mengecek keberadaan nilai kosong pada setiap kolom dalam dataset. Pengecekan ini bertujuan memastikan bahwa tidak ada data yang dibiarkan kosong, yang dapat mempengaruhi hasil analisis dan proses *preprocessing* selanjutnya.



Berdasarkan hasil pengecekan, seluruh kolom pada dataset menunjukkan nilai 0, yang berarti tidak terdapat nilai kosong pada data.

### **2.2.3 Pengecekan Nilai Duplikat**

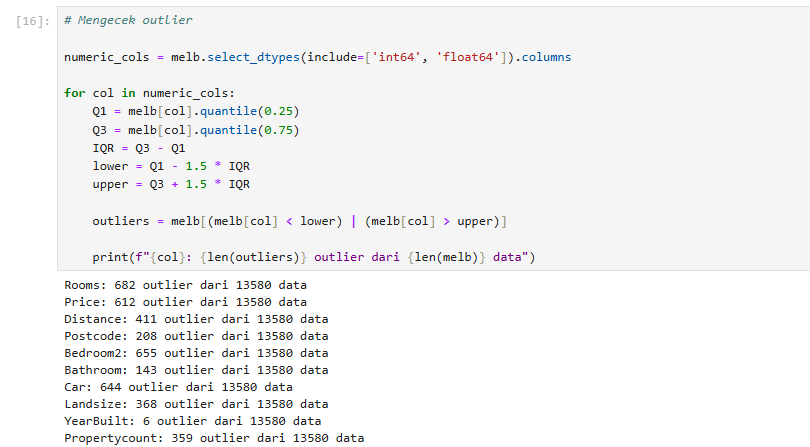
langkah selanjutnya dalam pengecekan adalah mengecek keberadaan baris duplikat pada dataset. Pengecekan ini penting untuk memastikan tidak ada data ganda yang dapat mempengaruhi hasil analisis.



Berdasarkan hasil pengecekan, jumlah baris duplikat adalah 0, yang berarti tidak terdapat data ganda pada dataset. kondisi ini menunjukkan bahwa dataset sudah dalam keadaan rapi dan siap untuk diproses selanjutnya.

### **2.2.4 Pengecekan Nilai Outlier**

Nilai outlier adalah nilai data yang berbeda secara ekstrem dari sebagian besar data lainnya dalam suatu kumpulan data. Dengan kata lain, outlier adalah data yang tampak “menyimpang” atau tidak mengikuti pola umum.



Hasilnya menunjukkan bahwa beberapa kolom dalam dataset memiliki cukup banyak nilai outlier. Misalnya, kolom *Rooms* memiliki 682 outlier, *Price* 612 outlier, dan *Car* 644 outlier dari total 13.580 data. Artinya, ada cukup banyak nilai yang berada di luar rentang normal pada fitur-fitur tersebut. Sementara kolom seperti *YearBuilt* hanya memiliki 6 outlier, menandakan datanya lebih stabil. Secara keseluruhan, hasil ini menggambarkan bahwa sebagian fitur memiliki variasi ekstrem yang cukup besar dibandingkan data lainnya.

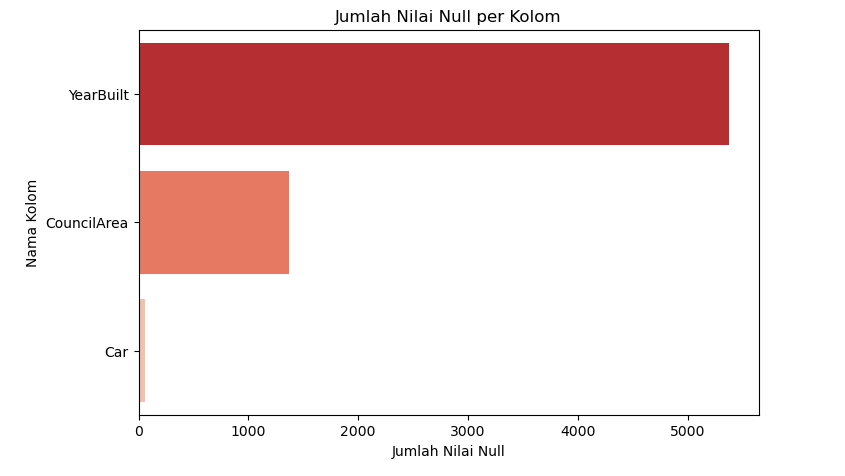
### **2.2.5 Visualisasi sebelum *processing***

Tahapan visualisasi data sebelum processing dilakukan untuk memahami gambaran awal dari struktur dataset. Visualisasi pada tahap ini membantu dalam mengidentifikasi pola umum, distribusi data, serta potensi permasalahan seperti nilai ekstrem (*outliers*) atau ketidakseimbangan data. Berikut beberapa visualisasi yang digunakan pada tahap ini untuk memberikan gambaran awal mengenai karakteristik dataset.

**1. Visualisasi *Bar chart* untuk Menampilkan Nilai Null per Kolom**

Visualisasi yang akan ditampilkan merupakan *bar chart* yang digunakan untuk menunjukkan jumlah nilai *Null* atau data yang hilang pada setiap kolom dalam dataset. Dengan visualisasi ini, kita dapat melihat secara cepat kolom mana yang memiliki banyak data kosong dan perlu mendapat perhatian khusus dalam proses *preprocessing*.



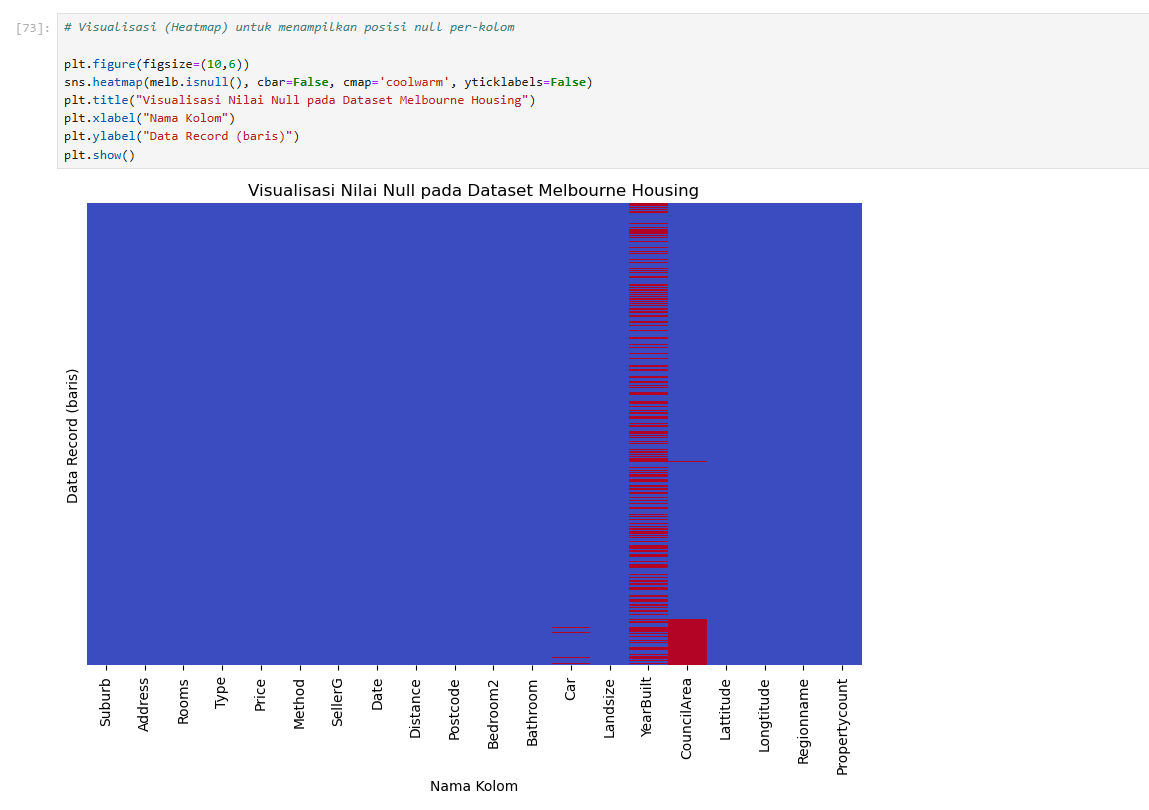


Hasil visualisasi tersebut menunjukkan bahwa kolom **YearBuilt** memiliki jumlah nilai *Null* terbanyak dibandingkan kolom lainnya, disusul oleh **CouncilArea**, dan terakhir **Car** dengan jumlah nilai *Null* yang sangat sedikit. Informasi ini membantu mengidentifikasi kolom mana yang perlu diprioritaskan dalam proses penanganan *missing values* pada tahap *preprocessing*.

Insight: Visualisasi bar chart ini menyoroti kolom-kolom pada dataset melb yang memiliki nilai null. Dari chart ini, terlihat kolom dengan bar lebih panjang memiliki jumlah missing value yang tinggi, sehingga menjadi prioritas utama untuk pembersihan data. Kolom dengan jumlah null sedikit bisa diatasi dengan imputasi, sedangkan kolom dengan null banyak mungkin perlu dipertimbangkan untuk dihapus. Dengan demikian, chart ini membantu memahami kualitas dataset secara cepat dan memandu strategi preprocessing sebelum analisis lebih lanjut.

**2. Visualisasi heatmap menampilkan posisi null pada setiap kolom**

Visualisasi yang akan ditampilkan merupakan heatmap yang digunakan untuk menunjukkan posisi data yang hilang atau Null pada setiap kolom dan baris dalam dataset. Visualisasi ini membantu melihat penyebaran nilai Null di seluruh baris data, sehingga kita dapat memahami pola kemunculannya secara detail.



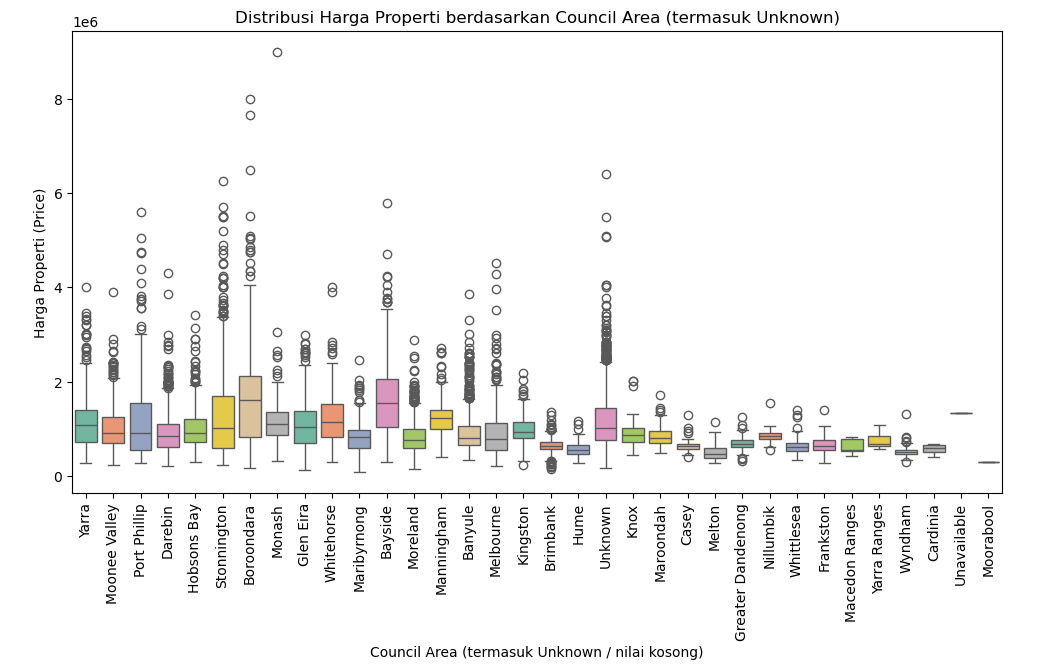
Hasil visualisasi menunjukkan bahwa nilai *Null* paling banyak muncul pada kolom YearBuilt, diikuti CouncilArea dan Car, yang ditandai dengan warna merah. Sementara kolom lainnya didominasi warna biru, yang berarti tidak memiliki nilai *Null*.

Insight: Visualisasi heatmap ini menampilkan posisi nilai null di seluruh dataset melb. Setiap kotak mewakili satu data pada kolom tertentu, di mana warna berbeda menandakan apakah nilainya kosong atau terisi. Dari heatmap ini, kita bisa dengan mudah melihat pola missing value, misalnya kolom tertentu yang sering kosong atau baris yang memiliki banyak nilai hilang. Informasi ini berguna untuk menentukan strategi pembersihan data, seperti kolom mana yang perlu diisi dengan imputasi, atau baris/kolom yang mungkin perlu dihapus sebelum melakukan analisis lebih lanjut.

**3. Visualisasi Boxplot Distribusi Harga Properti**

Visualisasi yang ditampilkan merupakan boxplot yang digunakan untuk memperlihatkan distribusi harga properti berdasarkan kategori Council Area, termasuk kategori “Unknown” sebagai pengganti nilai yang kosong. Boxplot dipilih karena mampu menampilkan sebaran data secara jelas, termasuk median, kuartil, serta adanya outlier sehingga memudahkan dalam membandingkan distribusi harga antar kategori.





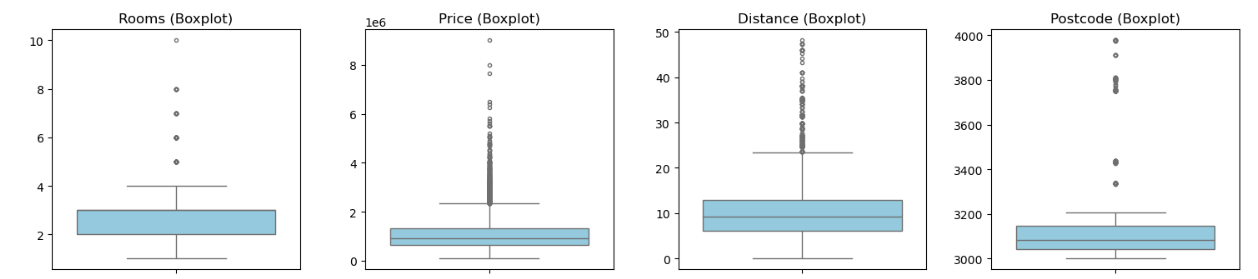
Hasil visualisasi menunjukkan bahwa terdapat variasi harga properti yang cukup besar di berbagai wilayah Council Area. Beberapa wilayah seperti *Boroondara* dan *Stonnington* memiliki median harga yang relatif tinggi dibanding wilayah lainnya. Selain itu, kategori “Unknown” juga menunjukkan adanya variasi harga yang cukup luas. Informasi ini penting untuk memahami bagaimana harga properti bervariasi antar wilayah dan seberapa signifikan perbedaan tersebut, termasuk pada data yang sebelumnya kosong.

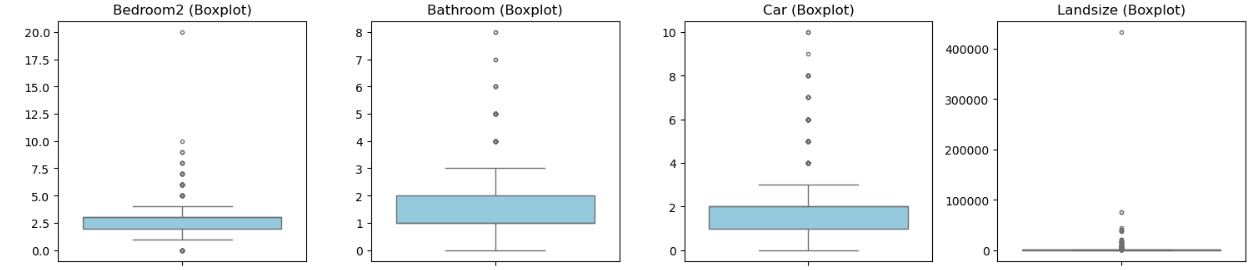
Insight: Visualisasi boxplot ini menampilkan distribusi harga properti berdasarkan masing-masing Council Area, termasuk kategori Unknown untuk data yang sebelumnya kosong. Dari plot ini, terlihat sebaran harga di setiap area, mulai dari median, rentang, hingga outlier. Keberadaan kategori Unknown memungkinkan kita melihat bagaimana harga properti di baris yang awalnya memiliki nilai kosong, sehingga tidak diabaikan dalam analisis. Plot ini membantu memahami perbedaan harga antar area dan memberi gambaran tentang variasi harga, termasuk apakah data yang hilang cenderung terkait dengan harga tertentu.

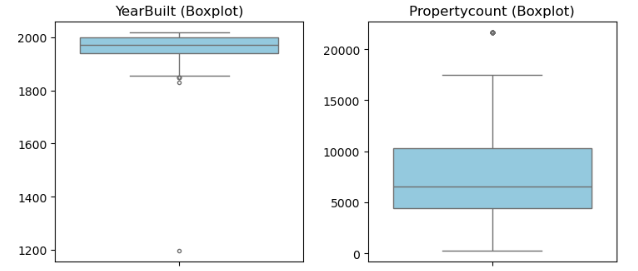
**4. Visualisasi untuk menampilkan Outliers**

Visualisasi outlier penting karena membantu kita melihat data yang menyimpang secara langsung. Dengan visualisasi ini, kita bisa cepat mengenali kolom mana yang memiliki banyak nilai ekstrem dan seberapa jauh penyimpangannya dari data mayoritas. Tanpa visualisasi, outlier hanya terlihat sebagai angka yang sulit diinterpretasikan.









Dari hasil boxplot di atas, terlihat bahwa hampir semua kolom memiliki titik-titik di luar kotak utama (whisker) inilah yang disebut outlier. Misalnya, kolom *Price*, *Distance*, *Car*, dan *Landsize* menunjukkan banyak titik di atas batas atas, artinya ada beberapa rumah dengan harga, jarak, kapasitas garasi, atau luas tanah yang jauh lebih besar dari kebanyakan properti lainnya. Sebaliknya, kolom seperti *YearBuilt* hanya memiliki sedikit outlier, menandakan data tahun pembangunan relatif seragam. Visualisasi ini membantu peneliti atau analis untuk memutuskan apakah nilai-nilai ekstrem tersebut akan dipertahankan (karena memang realistis) atau perlu dibersihkan agar analisis lebih akurat.

Insight: Visualisasi boxplot ini menunjukkan sebaran data dan outlier pada beberapa kolom penting di dataset melb, seperti harga, jumlah kamar, jarak, dan ukuran lahan. Dari plot terlihat bahwa beberapa kolom memiliki nilai ekstrem yang jauh dari sebaran utama, misalnya harga dan ukuran lahan, yang dapat mempengaruhi analisis statistik jika tidak ditangani. Boxplot ini membantu mengidentifikasi kolom mana saja yang perlu diperiksa lebih lanjut untuk penanganan outlier, baik melalui pembersihan data, transformasi, atau metode analisis yang lebih robust, sehingga hasil analisis nantinya lebih akurat dan representatif.

## **2.3 Data Preprocessing and Techniques**

Pada subbab ini akan membahas langkah-langkah penting untuk meningkatkan kualitas data sebelum pemodelan. Proses ini mencakup penanganan missing values dengan teknik seperti mean/mode imputation atau interpolation agar data lebih lengkap, sert penanganan outliers menggunakan IQR atau Z-Score untuk mengurangi pengaruh data ekstrem.

Selanjutnya, dilakukan feature scalling agar skala data seimbang, encoding untuk mengubah data kategorikal menjadi numerik, serta feature selection/reduction seperti PCA atau Feature Importance untuk menyederhanakan fitur.   
  
**1. Handling missing values (Median, Mode)**

Tahap handling missing values merupakan proses penting dalam pembersihan data yang bertujuan mengatasi nilai kosong (null atau NaN) agar tidak mengganggu proses analisis selanjutnya. Metode ini digunakan untuk mengisi nilai yang hilang agar tidak menunggu proses analisis atau pemodelan. Untuk memastikan dataset lengkap tanpa kehilangan informasi penting dan menjaga distribusi data tetap wajar.



Pada kasus ini, nilai yang hilang pada kolom numerik Car dan YearBuilt diisi menggunakan median karena metode ini lebih tahan terhadap keberadaan outlier dibandingkan mean. Sementara pada kolom kategorikal CouncilArea, nilai kosong diisi menggunakan mode atau nilai yang paling sering muncul.

Output tersebut menunjukkan bahwa proses penanganan missing value berhasil dilakukan dengan baik. Kolom numerik seperti *Car* dan *YearBuilt* diisi menggunakan nilai median, sedangkan kolom kategorikal *CouncilArea* diisi menggunakan nilai yang paling sering muncul (mode). Setelah dilakukan pengisian, hasil pengecekan menunjukkan bahwa jumlah missing value di semua kolom tersebut sudah menjadi 0, artinya tidak ada lagi data yang kosong. Dataset yang telah diperbaiki kemudian disimpan ke file baru bernama *melb\_cleaned\_missing.csv* untuk digunakan pada tahap analisis berikutnya.

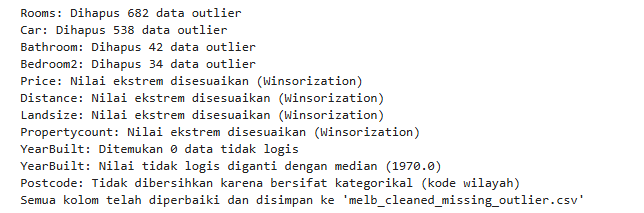


**2. Handling Outliers (IQR, Winsorization)**

Menangani outlier penting agar hasil analisis data tetap akurat dan tidak terdistorsi oleh nilai ekstrem. Dua cara yang sering digunakan adalah metode IQR (Interquartile Range) dan Winsorization. Alasan menggunakan metode ini yaitu untuk menjaga agar analisis data tidak bias akibat nilai yang tinggi atau rendah, tanpa mengorbankan terlalu banyak data. Dengan metode IQR, kita mencari batas bawah dan batas atas data berdasarkan selisih antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Nilai yang berada di luar batas tersebut dianggap outlier dan biasanya dihapus atau disesuaikan. Sedangkan Winsorization dilakukan dengan mengganti nilai-nilai ekstrem (terlalu besar atau terlalu kecil) dengan nilai batas tertentu.



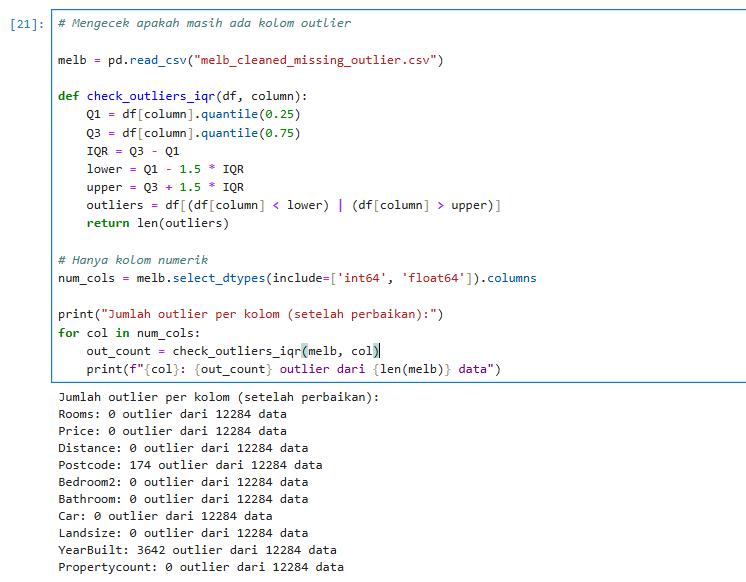




Proses pembersihan data dilakukan dengan dua pendekatan, yaitu IQR dan Winsorization. Pendekatan IQR digunakan untuk menghapus nilai yang dianggap menyimpang jauh dari rentang normal pada kolom seperti *Rooms*, *Car*, *Bathroom*, dan *Bedroom2*. Misalnya, sebanyak 682 data pada kolom *Rooms* terdeteksi dan dihapus karena tergolong outlier. Sementara itu, metode Winsorization diterapkan pada kolom seperti *Price*, *Distance*, *Landsize*, dan *Propertycount* dengan cara menyesuaikan nilai ekstrem agar tidak terlalu jauh dari distribusi utama tanpa menghapus data. Nilai pada kolom *YearBuilt* juga diperbaiki dengan mengganti tahun yang tidak logis (kurang dari 1800 atau lebih dari 2025) menggunakan nilai median yang wajar, yaitu tahun 1970. Kolom *Postcode* tidak mengalami perubahan karena bersifat kategorikal. Setelah seluruh proses pembersihan dan penyesuaian selesai, data disimpan kembali dalam file baru yang berisi data yang sudah lebih bersih dan siap digunakan untuk analisis selanjutnya.



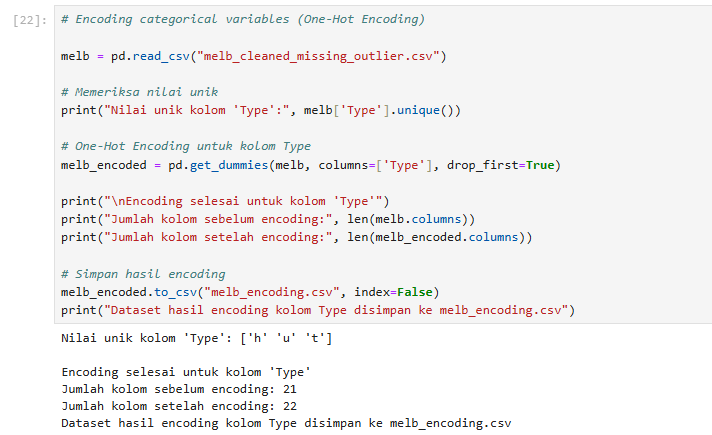
Kemudian dilakukan pengecekan ulang terhadap data untuk memastikan apakah masih terdapat outlier setelah proses pembersihan dilakukan. Langkah ini bertujuan untuk memverifikasi bahwa nilai-nilai ekstrem telah berhasil dihapus atau disesuaikan sehingga distribusi data menjadi lebih normal dan konsisten.



Setelah dilakukan pengecekan ulang terhadap data yang telah dibersihkan, terlihat bahwa sebagian besar kolom sudah tidak memiliki outlier lagi, seperti *Rooms*, *Price*, *Distance*, *Bedroom2*, *Bathroom*, *Car*, dan *Landsize*, yang semuanya menunjukkan 0 outlier dari total 12.284 data. Hal ini menandakan bahwa proses pembersihan sebelumnya berhasil menghapus atau menyesuaikan nilai-nilai ekstrem pada kolom-kolom tersebut. Namun, masih terdapat sedikit outlier pada kolom *Postcode* (174 data) dan *YearBuilt* (3642 data). Nilai-nilai ini masih muncul karena *Postcode* merupakan data kategorikal yang mewakili kode wilayah dan tidak dibersihkan, sedangkan variasi pada *YearBuilt* bisa jadi disebabkan oleh perbedaan tahun pembangunan properti yang masih tergolong wajar secara logika.

**3. Encoding categorical variables (One-hot encoding)**

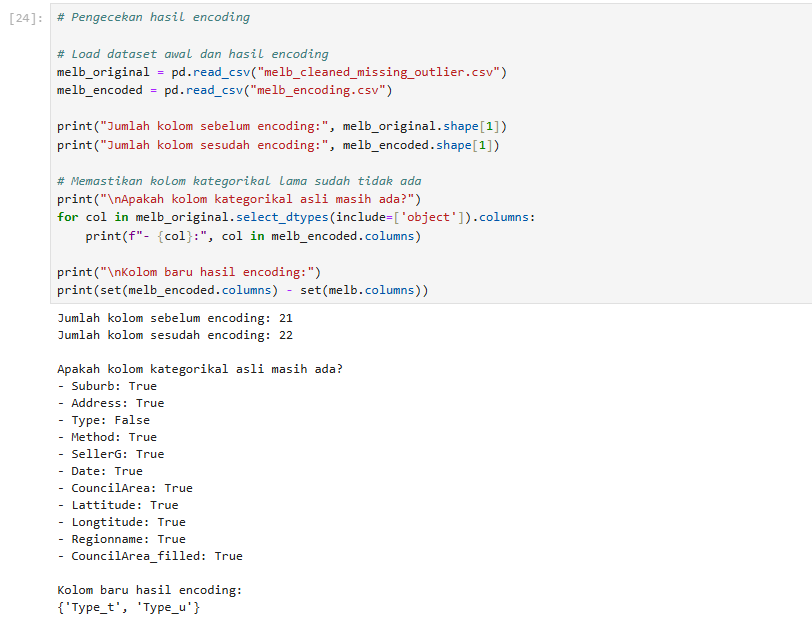
One-hot encoding adalah cara untuk mengubah data kategori menjadi bentuk angka agar bisa diproses oleh komputer dalam analisis atau pemodelan. Setiap kategori akan diubah menjadi kolom baru dengan nilai 1 jika data termasuk dalam kategori tersebut, dan 0 jika tidak. Metode ini digunakan untuk mengubah data kategori (teks) menjadi bentuk numerik. Alasan menggunakan metode ini karena sebagian besar algoritma analisis dan machine learning hanya bisa membaca data numerik, sehingga encoding diperlukan agar data kategorikal dapat diolah dengan benar.



Proses ini dilakukan untuk mengubah kolom Type, yang berisi data kategori (*h*, *u*, dan *t*), menjadi bentuk numerik menggunakan One-Hot Encoding agar bisa diproses oleh model analisis. Setelah dilakukan encoding, kolom *Type* dipecah menjadi beberapa kolom baru yang mewakili masing-masing kategori dengan nilai 0 atau 1. Jumlah kolom dalam dataset bertambah dari 21 menjadi 22 setelah proses ini, menandakan bahwa satu kolom kategori berhasil diubah menjadi kolom numerik baru. Dataset hasil encoding kemudian disimpan dengan nama “melb\_encoding.csv” dan siap digunakan untuk analisis atau pemodelan berikutnya.



Kemudian dilakukan pengecekan ulang untuk memastikan bahwa proses encoding telah berhasil dan setiap kategori pada kolom *Type* telah terkonversi dengan benar menjadi bentuk numerik sesuai dengan hasil One-Hot Encoding.

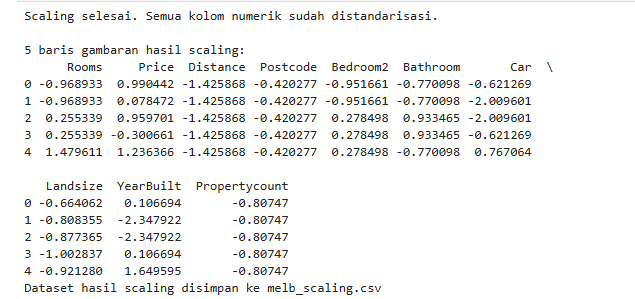


Dari hasilnya terlihat bahwa jumlah kolom bertambah dari 21 menjadi 22, yang menunjukkan adanya penambahan kolom baru hasil One-Hot Encoding. Kolom “Type” yang sebelumnya bersifat kategorikal telah diubah menjadi dua kolom numerik baru, yaitu ‘Type\_t’ dan ‘Type\_u’, sementara kolom “Type” asli sudah tidak ada lagi. Hal ini menandakan bahwa proses encoding berjalan dengan benar.

**4. Feature scaling (standardization)**

Feature scaling adalah proses menyesuaikan skala nilai pada kolom numerik agar berada dalam rentang yang seragam, biasanya dengan cara menstandarisasi data sehingga memiliki rata-rata 0 dan standar deviasi 1. Tujuannya adalah agar setiap fitur memiliki pengaruh yang seimbang dalam proses analisis atau pemodelan, terutama pada algoritma yang sensitif terhadap perbedaan skala, seperti regresi atau clustering. Alasan penggunaan metode **f**eature scaling (standardization) adalah untuk menyamakan skala antar kolom numerik agar setiap variabel memiliki pengaruh yang seimbang dalam proses analisis atau pemodelan.



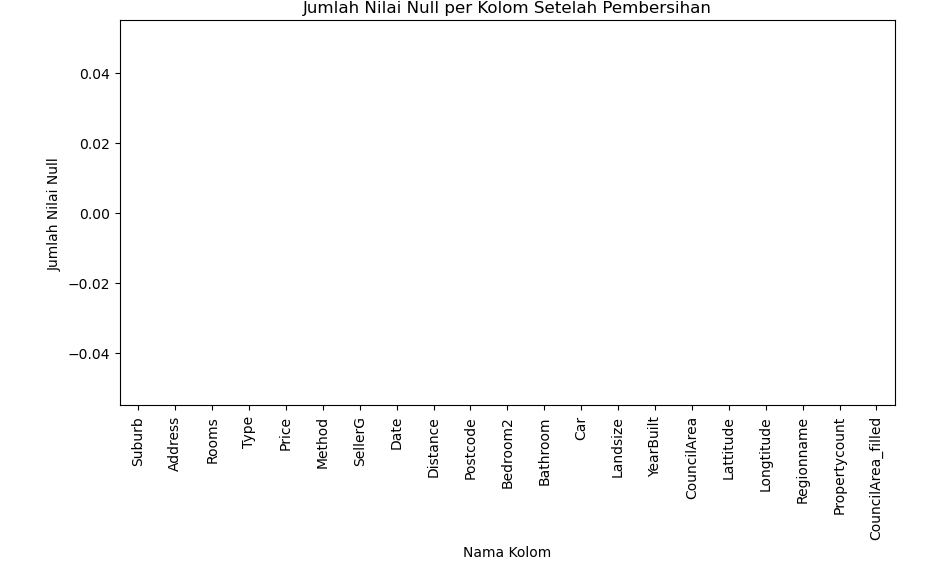


Pada output tersebut terlihat bahwa semua nilai numerik telah berubah menjadi nilai yang berada di sekitar 0, baik positif maupun negatif. Ini menunjukkan bahwa proses scaling berhasil dilakukan. Hasil scaling kemudian disimpan ke file melb\_scaling.csv untuk digunakan pada tahap analisis berikutnya. Metode ini digunakan agar model tidak bias terhadap fitur yang memiliki nilai lebih besar dan agar perhitungan jarak atau bobot antar variabel menjadi lebih akurat.

****

**5. Visualisasi Data Bersih**

1. ***Bar Chart* Jumlah Nilai Null per Kolom Setelah Pembersihan Data**

****

Visualisasi pada gambar tersebut menunjukkan jumlah nilai null per kolom setelah proses pembersihan data, menggunakan diagram batang (bar chart). Pada tahap ini, dataset yang telah dibersihkan (melb\_cleaned\_missing\_outlier.csv) dimuat kembali, lalu dilakukan pengecekan jumlah nilai yang masih kosong (null) di setiap kolom. Hasilnya divisualisasikan dalam bentuk batang vertikal, di mana sumbu-x menampilkan nama kolom dan sumbu-y menampilkan jumlah nilai null di masing-masing kolom.

Dari grafik terlihat bahwa seluruh batang memiliki nilai 0, artinya tidak ada lagi data yang hilang (missing values) setelah proses pembersihan dilakukan. Ini menunjukkan bahwa tahap data cleaning berhasil menghapus atau mengganti seluruh nilai kosong di dataset.

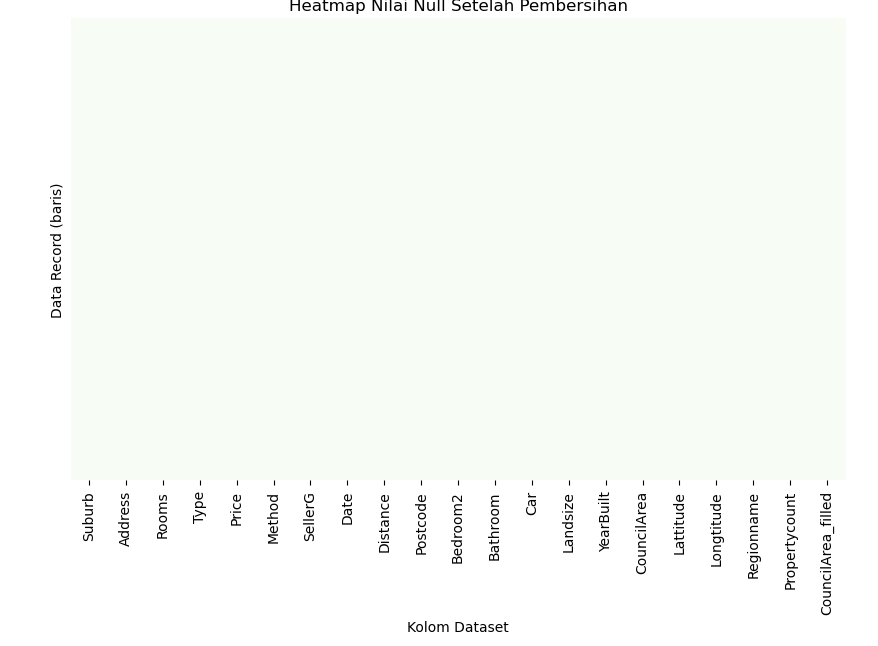
Alasan menggunakan diagram batang (bar chart) adalah karena jenis grafik ini paling efektif untuk membandingkan nilai diskrit antar kategori, dalam hal ini nama-nama kolom. Bar chart memudahkan kita untuk melihat secara cepat kolom mana yang memiliki nilai null terbanyak atau apakah masih ada kolom dengan data kosong. Dengan tampilan sederhana dan jelas, diagram batang menjadi pilihan yang tepat untuk mengevaluasi hasil pembersihan data secara visual dan memastikan kualitas dataset sebelum melanjutkan ke tahap analisis berikutnya.

**Insight:** Setelah proses pembersihan data, semua kolom sudah tidak memiliki data kosong.

**Interpretasi:** Hal ini berarti dataset sudah lengkap dan siap digunakan untuk analisis lebih lanjut. Dengan tidak adanya data kosong, proses analisis akan lebih akurat dan tidak terganggu oleh informasi yang hilang.

1. ***Heatmap* Nilai Null Setelah Pembersihan Data**

****

****

Visualisasi pada gambar tersebut menampilkan “Heatmap Nilai Null Setelah Pembersihan Data”. Grafik ini digunakan untuk melihat apakah masih ada nilai kosong (missing values) pada dataset setelah proses pembersihan dilakukan. Setiap kolom pada sumbu-x mewakili fitur (kolom dataset), sedangkan sumbu-y mewakili baris data (record). Warna pada heatmap biasanya menunjukkan keberadaan nilai null misalnya, warna terang menandakan data ada (tidak null) dan warna gelap menandakan data hilang.

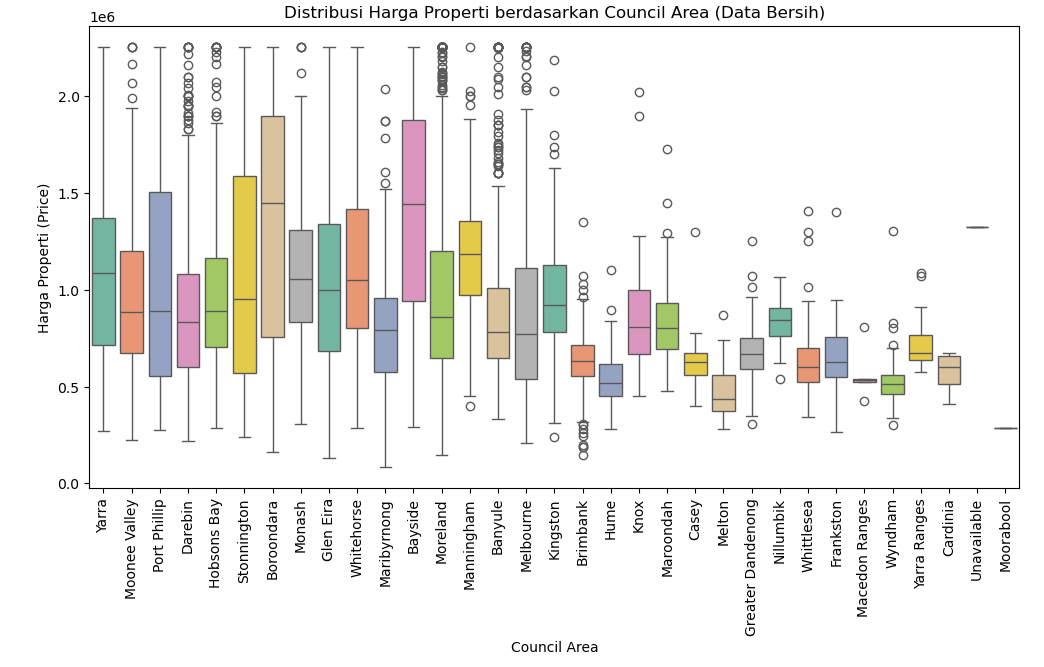
Dari hasil visualisasi terlihat bahwa heatmap berwarna seragam tanpa adanya area gelap, yang berarti seluruh data sudah bersih dari nilai null. Dengan kata lain, tidak ada lagi data yang hilang di seluruh kolom dataset. Alasan menggunakan heatmap adalah karena jenis grafik ini sangat efektif untuk melihat pola distribusi nilai kosong secara visual dan cepat, terutama ketika dataset memiliki banyak kolom dan baris. Heatmap juga memudahkan kita untuk mendeteksi bagian mana yang masih memiliki nilai hilang tanpa harus melihat tabel data secara manual.

**Insight:** Tampilan heatmap menunjukkan tidak ada lagi area berwarna gelap atau kosong.

**Interpretasi:** Ini memperkuat bahwa seluruh nilai kosong telah berhasil diisi atau dihapus. Dataset kini lebih bersih, sehingga hasil analisis dapat lebih dipercaya.

1. ***Boxplot* Distribusi Harga Properti berdasarkan Council Area (Data Bersih)**

****

****

Visualisasi pada gambar tersebut menunjukkan “Distribusi Harga Properti berdasarkan Council Area (Data Bersih)” menggunakan diagram boxplot. Grafik ini memperlihatkan bagaimana harga properti tersebar di setiap wilayah *Council Area* setelah data dibersihkan dari nilai kosong dan outlier. Setiap kotak (box) mewakili sebaran data harga pada satu wilayah tertentu bagian tengah kotak menunjukkan nilai median (harga tengah), sedangkan garis atas dan bawah kotak menunjukkan rentang harga properti (kuartil 1 dan kuartil 3). Titik-titik di luar kotak menandakan adanya outlier atau harga properti yang jauh lebih tinggi atau rendah dari kebanyakan data di wilayah tersebut.

Dari visualisasi ini terlihat bahwa wilayah Stonnington, Boroondara, dan PortPhillip memiliki median harga properti tertinggi, berada di atas 1 juta dolar Australia. Wilayah tersebut juga memiliki banyak outlier, yang berarti ada banyak properti dengan harga sangat tinggi (rumah premium). Sementara itu, wilayah seperti Hume, Casey, dan Whittlesea memiliki median harga yang lebih rendah, menandakan harga rumah di wilayah ini lebih terjangkau.

Beberapa wilayah juga memiliki kotak (box) yang tinggi dan lebar, yang menunjukkan perbedaan harga yang cukup besar di dalam wilayah tersebut ini terlihat jelas pada Stonnington dan Boroondara. Sedangkan wilayah dengan kotak lebih pendek dan sempit, seperti Casey atau Hume, menunjukkan harga rumah yang lebih stabil dan tidak terlalu bervariasi.

Alasan menggunakan boxplot adalah karena diagram ini sangat efektif untuk menunjukkan distribusi data dan mendeteksi pencilan (outlier). Boxplot juga memudahkan kita memahami persebaran harga properti secara ringkas, tanpa harus melihat seluruh data numerik secara detail.

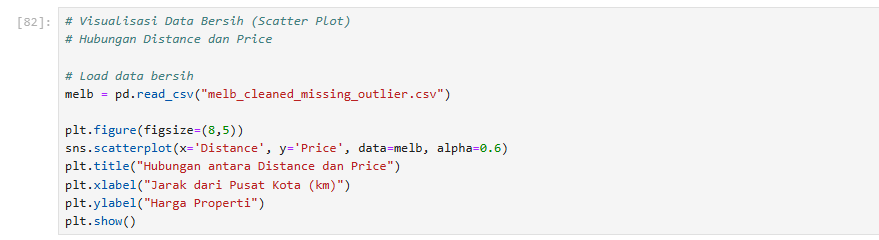
**Insight:**

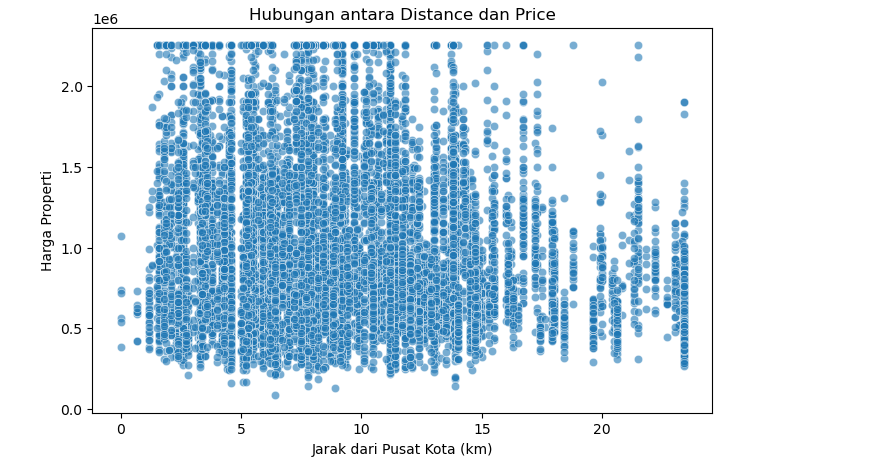
* Wilayah Stonnington, Boroondara, dan Port Phillip punya harga rumah paling tinggi.
* Wilayah Hume, Casey, dan Whittlesea punya harga rumah lebih rendah dan stabil.
* Banyak titik outlier di wilayah mahal, artinya ada rumah-rumah dengan harga jauh di atas rata-rata.

**Interpretasi:**

* Lokasi memengaruhi harga rumah. Semakin strategis wilayahnya, harga rumah makin mahal.
* Banyaknya outlier di wilayah mahal menunjukkan ada properti mewah.
* Wilayah dengan harga rendah cenderung lebih stabil dan cocok untuk pembelian rumah biasa.

1. ***Scatter Plot* Hubungan antara Jarak dari Pusat Kota dan Harga Properti**

****

****

Gambar di atas menunjukkan hubungan antara jarak dari pusat kota (Distance) dengan harga properti (Price) menggunakan diagram sebar (scatter plot). Posisi mendatar (sumbu X) menunjukkan jarak rumah dari pusat kota dalam kilometer, sedangkan posisi tegak (sumbu Y) menunjukkan harga rumah dalam dolar Australia.

Dari tampilan tersebut, terlihat bahwa sebagian besar properti berada pada jarak sekitar 5–15 km dari pusat kota, dengan harga yang bervariasi. Secara umum, dapat disimpulkan bahwa semakin jauh jarak properti dari pusat kota, harganya cenderung lebih rendah, meskipun ada beberapa properti jauh yang tetap memiliki harga tinggi. Alasan menggunakan diagram sebar adalah karena jenis diagram ini sangat efektif untuk menunjukkan hubungan atau pola antara dua variabel numerik. Diagram sebar membantu kita melihat apakah ada korelasi antara jarak dan harga, serta memudahkan dalam mengenali pola umum atau titik-titik yang menyimpang dari tren.

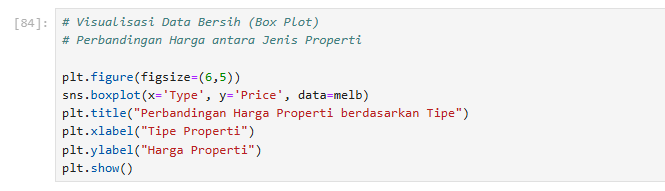
**Insight:**

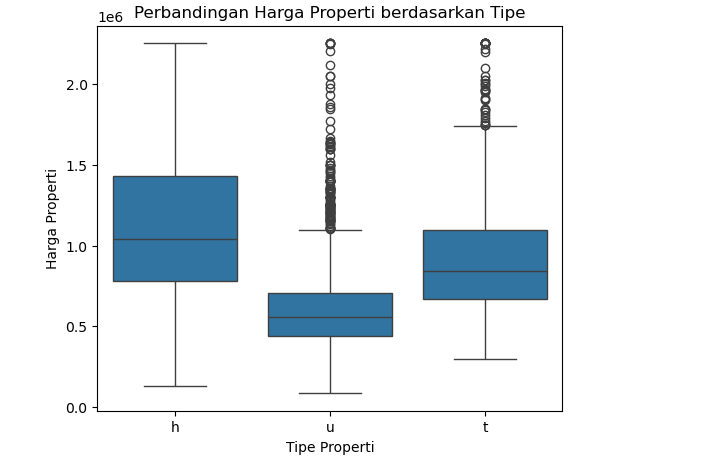
* Sebagian besar rumah berada pada jarak 5–15 km dari pusat kota.
* Semakin jauh jarak rumah dari pusat kota, harga properti cenderung lebih rendah.
* Meskipun begitu, masih ada beberapa rumah mahal di jarak jauh, kemungkinan rumah besar atau properti mewah.

**Interpretasi:**

* Lokasi sangat berpengaruh pada harga. Rumah di dekat pusat kota biasanya lebih mahal karena akses ke fasilitas umum, transportasi, dan pusat aktivitas ekonomi lebih mudah.
* Rumah di pinggiran kota lebih murah, tapi bukan berarti semua murah — ada pengecualian pada rumah-rumah tertentu.
* Scatter plot ini menunjukkan hubungan negatif antara jarak dan harga: jarak naik → harga cenderung turun.

1. ***Boxplot* Perbandingan Harga Properti berdasarkan Tipe Properti**

****

****

Gambar di atas menampilkan perbandingan harga properti berdasarkan tipe properti menggunakan diagram kotak (boxplot). Terdapat tiga tipe properti yang ditunjukkan, yaitu h (house/rumah), u (unit/apartemen kecil), dan t (townhouse). Sumbu mendatar menunjukkan tipe properti, sedangkan sumbu tegak menunjukkan harga propertinya.

Dari tampilan diagram, terlihat bahwa rumah (h) memiliki harga median paling tinggi, diikuti oleh townhouse (t), sedangkan unit (u) memiliki harga rata-rata paling rendah. Selain itu, tipe *u* dan *t* menampilkan banyak titik di atas kotak, yang menandakan adanya outlier atau harga yang jauh lebih tinggi dari kebanyakan data.

Alasan menggunakan diagram kotak (boxplot) adalah karena jenis diagram ini sangat baik untuk membandingkan distribusi dan variasi harga antara beberapa kategori. Boxplot juga membantu memperlihatkan median, sebaran data, serta nilai ekstrem (outlier) dengan cara yang sederhana dan mudah dipahami.

**Insight:**

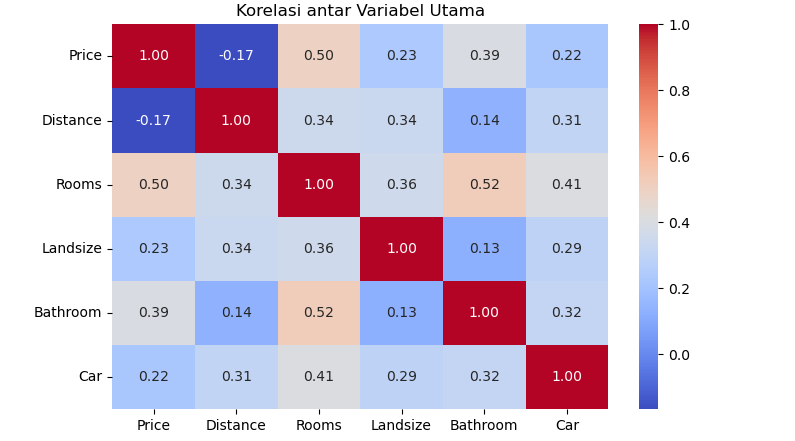
* House (h) memiliki harga tertinggi, dengan median di atas 1 juta dolar Australia.
* Townhouse (t) memiliki harga median sedikit lebih rendah dari house.
* Unit (u) adalah tipe properti dengan harga paling rendah dibanding dua tipe lainnya.
* Banyak outlier muncul pada tipe unit dan townhouse, yang menunjukkan ada properti dengan harga jauh di atas rata-rata.

**Interpretasi:**

* Tipe properti berpengaruh besar terhadap harga jual. Rumah tapak lebih mahal karena memiliki lahan dan bangunan yang lebih luas.
* Unit lebih murah karena ukuran yang kecil dan umumnya berada di lingkungan apartemen.
* Townhouse berada di tengah karena menawarkan keseimbangan antara ukuran dan harga.
* Banyaknya outlier pada unit dan townhouse menunjukkan ada properti premium atau mewah pada tipe tersebut.

1. **Heatmap Korelasi antar Variabel Utama**

****

****

Gambar di atas merupakan heatmap yang menunjukkan korelasi antar variabel utama dalam dataset properti. Warna merah pada heatmap menunjukkan korelasi positif, warna biru menunjukkan korelasi negatif, dan warna putih menunjukkan korelasi lemah atau mendekati nol. Nilai korelasi berkisar antara -1 hingga 1, di mana semakin mendekati 1 berarti hubungan positif semakin kuat, sedangkan semakin mendekati -1 berarti hubungan negatif semakin kuat. Dari visualisasi tersebut, terlihat bahwa variabel Price memiliki korelasi positif yang cukup kuat dengan Rooms (0.50), Bathroom (0.39), dan Car (0.22). Ini berarti semakin banyak kamar, kamar mandi, dan tempat parkir yang dimiliki sebuah properti, maka harganya cenderung semakin tinggi. Price juga memiliki korelasi positif lemah dengan Landsize (0.23), dan korelasi negatif dengan Distance (-0.17), yang berarti semakin jauh dari pusat kota, harga properti cenderung lebih rendah.

Selain itu, variabel Rooms memiliki korelasi cukup kuat dengan Bathroom (0.52) dan Car (0.41), yang menunjukkan bahwa rumah dengan banyak kamar biasanya juga memiliki lebih banyak kamar mandi dan area parkir. Distance juga menunjukkan korelasi positif rendah dengan Rooms, Landsize, dan Car, yang mengindikasikan bahwa rumah yang jauh dari pusat kota cenderung lebih besar dan memiliki lahan lebih luas. Secara keseluruhan, heatmap ini memperlihatkan bahwa faktor jumlah kamar, kamar mandi, dan luas tanah berperan penting dalam menentukan harga properti, sementara jarak dari pusat kota memberikan pengaruh negatif terhadap harga.

Alasan penggunaan visualisasi bentuk Heatmap karena mampu menampilkan kekuatan dan arah hubungan antar beberapa variabel numerik secara bersamaan. Warna pada heatmap membantu mengidentifikasi korelasi positif atau negatif dengan cepat. Visualisasi ini digunakan untuk memperkuat pemahaman bahwa jarak memiliki hubungan negatif dengan harga, sementara jumlah kamar, luas tanah, dan fasilitas lainnya memiliki hubungan positif terhadap harga properti.

### **Insight:**

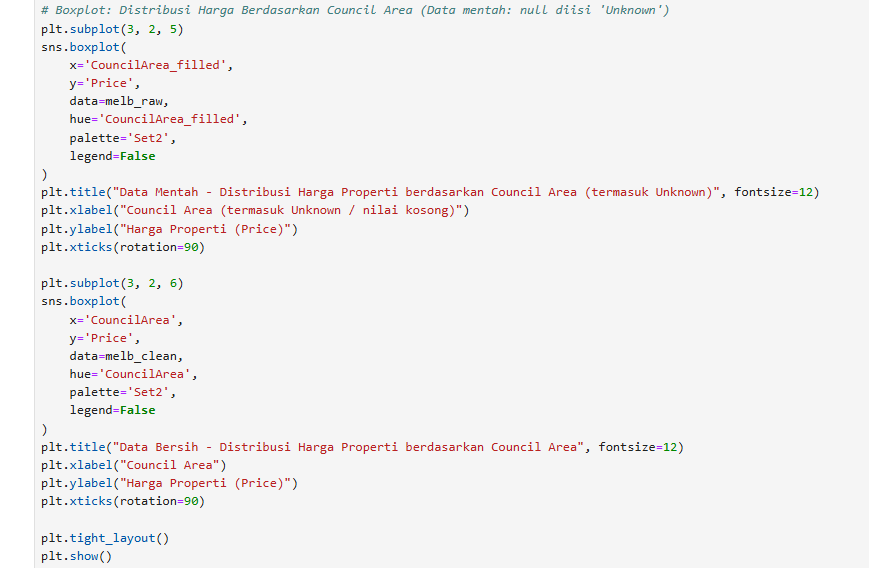
* Harga properti (Price) memiliki hubungan positif yang cukup kuat dengan jumlah kamar (Rooms), jumlah kamar mandi (Bathroom), dan jumlah tempat parkir (Car). Semakin besar dan lengkap fasilitas rumah, harganya cenderung semakin tinggi.
* Harga juga memiliki korelasi positif lemah dengan luas tanah (Landsize), serta korelasi negatif dengan jarak dari pusat kota (Distance) semakin jauh dari pusat kota, harga cenderung menurun.
* Terdapat hubungan yang kuat antar variabel fasilitas rumah, seperti Rooms dengan Bathroom dan Car, yang menunjukkan rumah besar biasanya dilengkapi fasilitas lengkap.

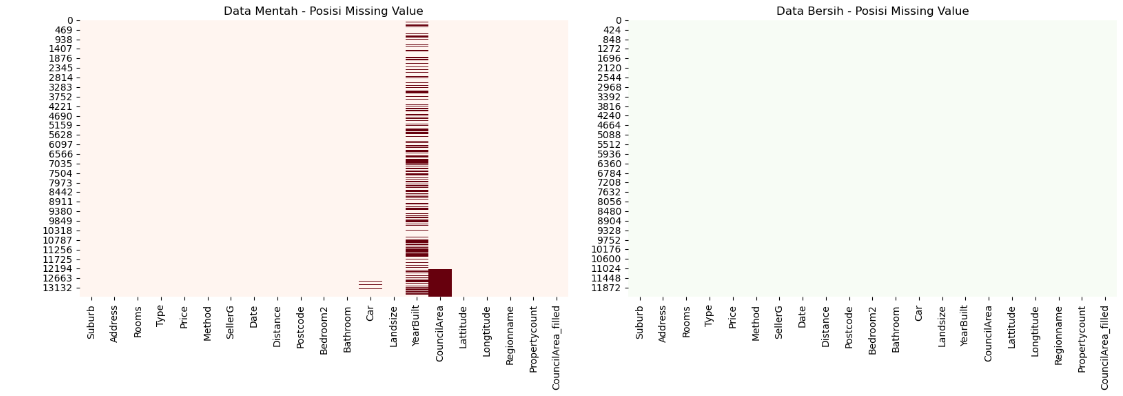
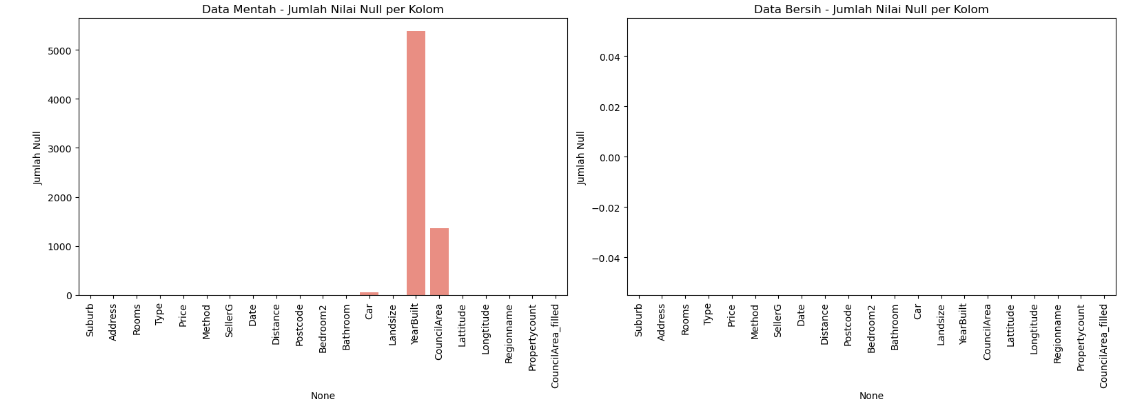
### **Interpretasi:**

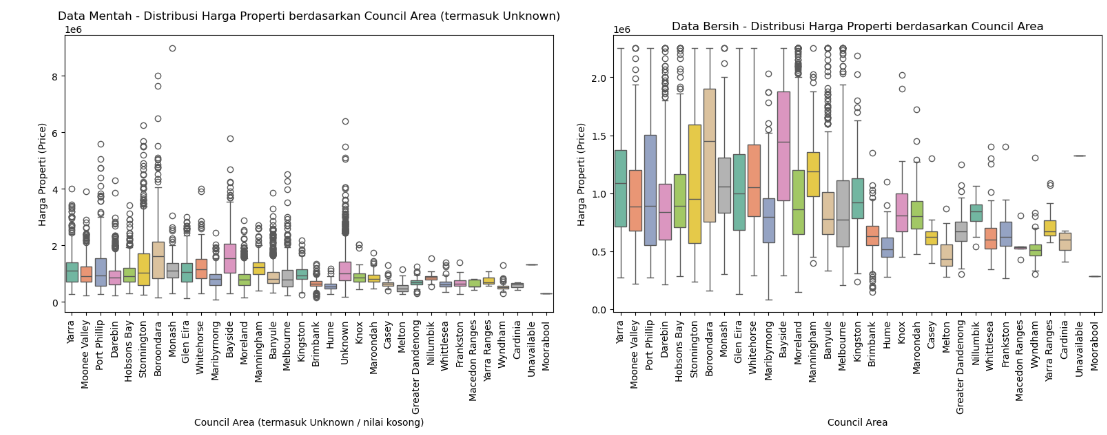
* Faktor internal rumah, seperti ukuran dan fasilitas (jumlah kamar, kamar mandi, lahan, dan tempat parkir), lebih berpengaruh besar terhadap harga properti dibandingkan faktor jarak.
* Jarak dari pusat kota tetap berpengaruh, meski tidak sekuat fasilitas rumah. Rumah yang dekat pusat kota lebih mahal karena nilai lokasi yang tinggi.
* Korelasi antar variabel seperti Rooms, Bathroom, dan Car menggambarkan karakteristik rumah modern: semakin luas rumah, semakin lengkap fasilitasnya.

1. **Perbandingan Data Mentah dengan Data Bersih**

****

****

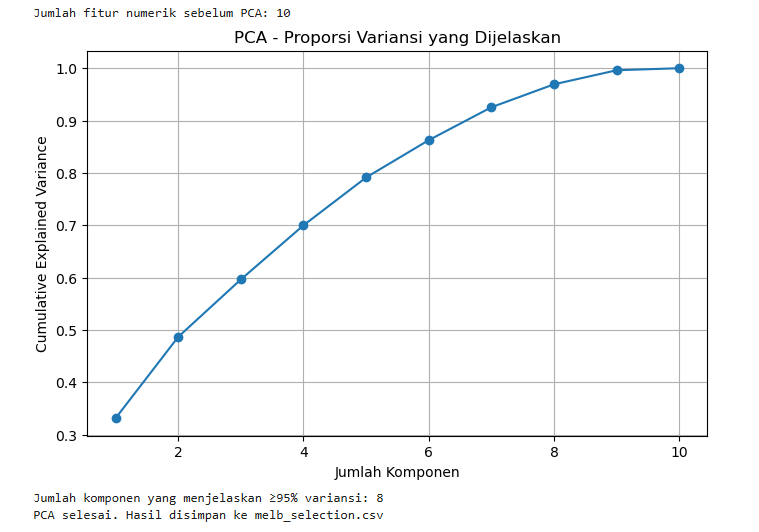
****

****

Kode ini digunakan untuk membandingkan data mentah dan data bersih dengan tiga jenis visualisasi. Bar chart menunjukkan jumlah nilai kosong di setiap kolom. Pada data mentah banyak kolom yang masih kosong, sedangkan pada data bersih hampir tidak ada nilai kosong lagi. Heatmap memperlihatkan posisi nilai kosong. Pada data mentah masih banyak bagian kosong, sedangkan pada data bersih sudah terisi semua. Boxplot menunjukkan sebaran harga properti berdasarkan wilayah. Pada data mentah masih ada kategori “Unknown”, tapi pada data bersih sudah hilang, sehingga hasilnya lebih rapi dan akurat.

**6. Feature selection/reduction**Feature selection atau feature reduction dengan PCA (Principal Component Analysis) digunakan untuk menyederhanakan data dengan mengurangi jumlah fitur tanpa menghilangkan informasi penting. Metode ini dipilih karena mampu mengurangi kompleksitas data, mempercepat proses analisis, dan mencegah overfitting dengan tetap mempertahankan sebagian besar variasi yang ada pada data asli.





Dari hasil yang ditunjukkan, terdapat 10 fitur numerik sebelum dilakukan PCA, dan diperoleh bahwa 8 komponen utama sudah mampu menjelaskan lebih dari 95% variasi data. Grafik “Cumulative Explained Variance” menunjukkan bahwa sebagian besar informasi dari data asli berhasil dipertahankan meskipun jumlah fitur berkurang. Hasil transformasi ini kemudian disimpan dalam file melb\_selection.csv untuk digunakan pada tahap analisis selanjutnya.

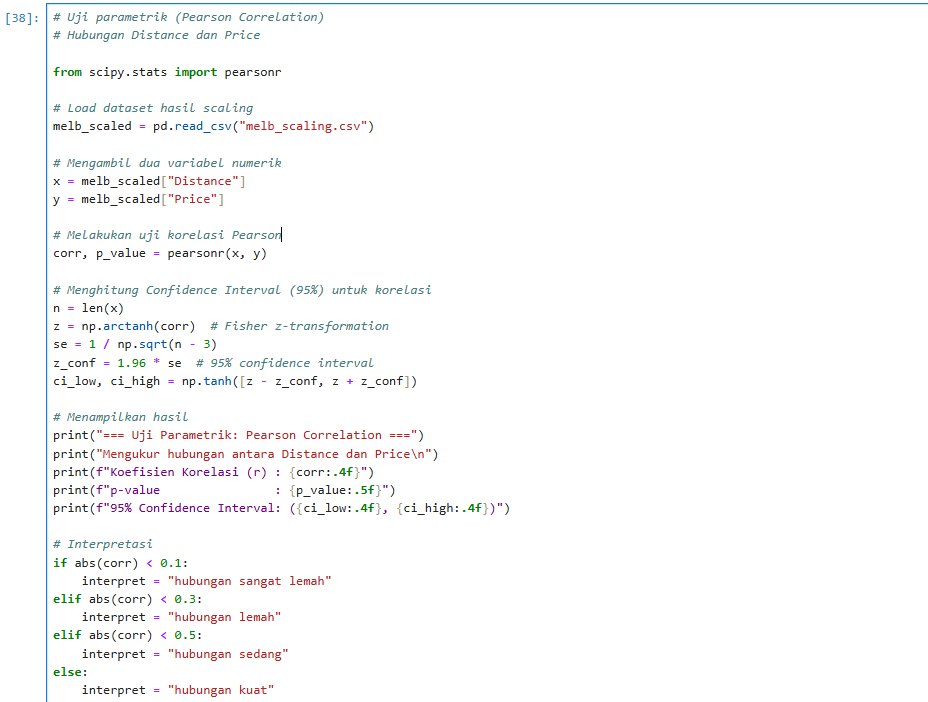


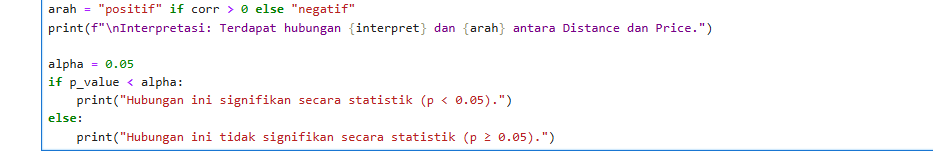
## **2.4 Statistical Analysis**

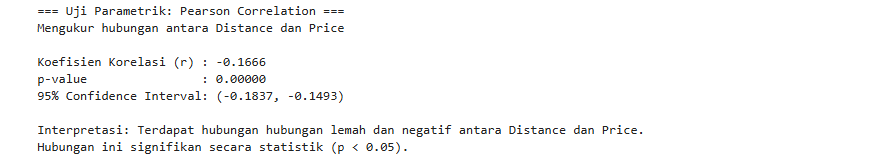
Menjelaskan metode analisis statistik yang digunakan, mencakup:

* Minimal 1 uji parametrik, seperti t-test, ANOVA, atau Pearson correlation.
* Sertakan hasil uji dengan nilai p-value, effect size, dan confidence interval (jika relevan).

1. Uji parametrik Pearson Correlation digunakan untuk mengetahui hubungan linier antara jarak rumah dari pusat kota (Distance) dan harga properti (Price). Data yang digunakan sudah melalui proses scaling agar nilai ekstrem tidak mempengaruhi hasil. Fungsi pearsonr(x, y) menghitung koefisien korelasi (r) dan p-value. Selanjutnya, transformasi Fisher digunakan untuk menentukan confidence interval 95% dari korelasi, sehingga kita bisa mengetahui rentang nilai korelasi yang paling mungkin berlaku di populasi. Kode ini juga menilai kekuatan hubungan (sangat lemah, lemah, sedang, atau kuat), arah hubungan (positif atau negatif) dan mengecek apakah hubungan ini signifikan secara statistik (p < 0,05).

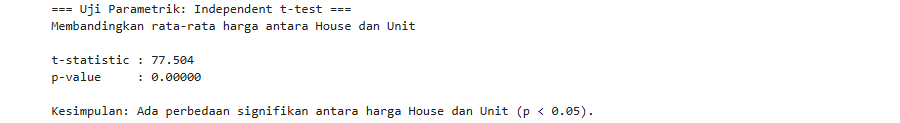
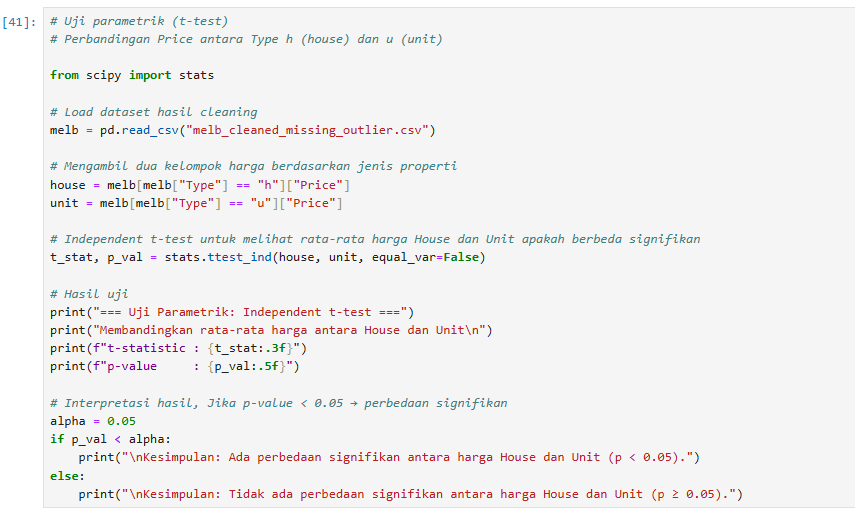






Hasil uji menunjukkan adanya hubungan negatif antara jarak rumah dan harga properti, artinya semakin jauh dari pusat kota, harga cenderung menurun. Besar kecilnya nilai r menentukan apakah hubungan tersebut lemah, sedang, atau kuat. Jika p-value kurang dari 0,05, maka hubungan ini signifikan secara statistik, menandakan bukti yang cukup bahwa jarak benar-benar memengaruhi harga properti. Dengan kata lain, output memberikan gambaran seberapa kuat dan seberapa pasti hubungan antara Distance dan Price.

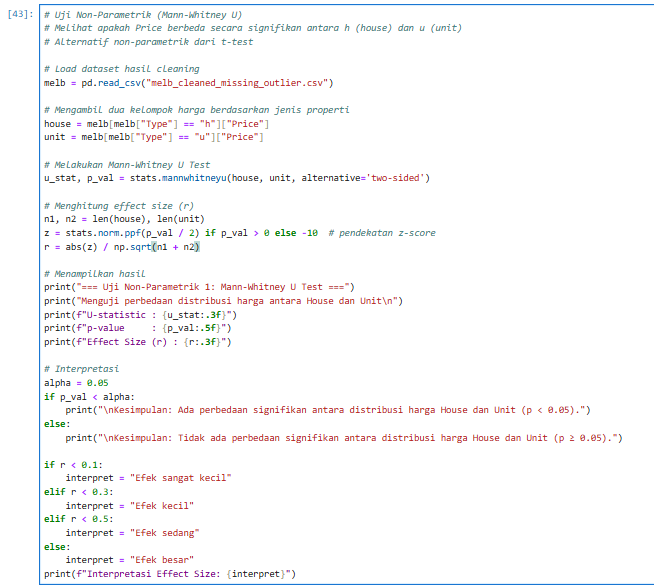
1. Uji parametrik independent t-test dibuat untuk membandingkan rata-rata harga properti antara dua jenis bangunan: house (h) dan unit (u). Data yang digunakan sudah dibersihkan dari missing value dan outlier. Fungsi ttest\_ind menghitung t-statistic dan p-value untuk melihat apakah perbedaan rata-rata harga kedua kelompok signifikan. Uji ini mengasumsikan data berdistribusi normal dan varians antar kelompok relatif seimbang (homogen).

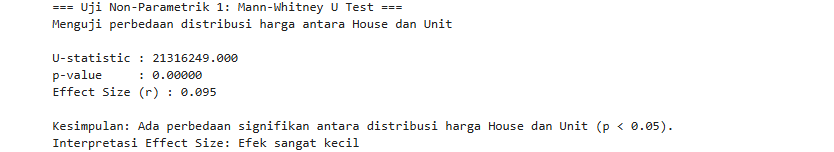


Hasil uji menunjukkan apakah rata-rata harga house dan unit berbeda secara signifikan. Jika p-value < 0,05, berarti perbedaan rata-rata harga memang signifikan secara statistik, sehingga jenis properti berpengaruh terhadap harga jual. Output memberi informasi bahwa jenis bangunan (house atau unit) dapat memengaruhi harga properti secara nyata.

* + Minimal 1 uji non-parametrik, seperti Mann-Whitney U, Kruskal-Wallis, atau Spearman correlation.
  + Sertakan hasil uji dengan nilai p-value, effect size, dan confidence interval (jika relevan).

1. Uji non-parametrik Mann–Whitney U Test merupakan alternatif dari t-test ketika data tidak berdistribusi normal. Uji ini digunakan untuk melihat apakah distribusi harga properti berbeda antara house (h) dan unit (u). Fungsi Mann-Whitney U Test menghitung U-statistic dan p-value, sedangkan effect size (r) dihitung dari z-score dan jumlah sampel untuk mengukur seberapa besar perbedaan praktis antara kedua kelompok.

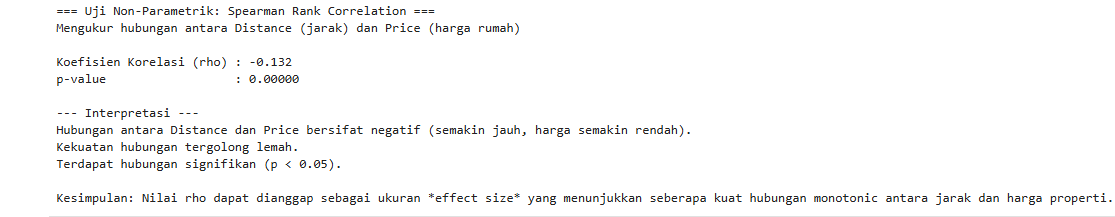




Hasil uji menunjukkan apakah distribusi harga house dan unit berbeda secara signifikan. Jika p-value < 0,05, perbedaan tersebut signifikan secara statistik. Effect size (r) memberikan gambaran besarnya perbedaan secara praktis: nilai kecil (<0,3) menandakan perbedaan signifikan secara statistik tetapi dampaknya relatif kecil, sedangkan nilai lebih besar menunjukkan pengaruh yang lebih kuat. Berdasarkan kode, terdapat perbedaan distribusi harga yang signifikan antara rumah dan unit, namun efeknya tergolong kecil hingga sedang, artinya walaupun secara statistik berbeda, dampak praktisnya tidak terlalu besar.

1. Uji non-parametrik Spearman Rank Correlation merupakan alternatif dari Pearson Correlation, khususnya untuk data yang tidak berdistribusi normal atau tidak memiliki hubungan linier. Uji ini mengukur hubungan monotonik antara jarak rumah dari pusat kota (Distance) dan harga properti (Price) menggunakan koefisien rho (ρ). Fungsi spearman menghitung nilai rho dan p-value, sedangkan interpretasi rho mencakup arah hubungan (positif atau negatif) dan kekuatan hubungan (sangat lemah hingga sangat kuat).





Hasil uji menunjukkan hubungan negatif antara Distance dan Price, artinya semakin jauh rumah dari pusat kota, harga cenderung menurun. Nilai absolut rho menunjukkan kekuatan hubungan, yang dalam kasus ini tergolong sedang. P-value menunjukkan apakah hubungan tersebut signifikan secara statistik. Dengan kata lain, output memberi gambaran bahwa jarak rumah memengaruhi harga properti secara konsisten, dan nilai rho juga bisa digunakan sebagai ukuran effect size untuk menilai seberapa kuat hubungan monotonic antara kedua variabel.

* Jelaskan interpretasi hasil secara mendalam.

Berdasarkan hasil uji statistik, jarak rumah dari pusat kota memiliki hubungan negatif dengan harga properti. Artinya, semakin jauh lokasi rumah, harga cenderung lebih rendah. Uji Pearson dan Spearman sama-sama menunjukkan hubungan ini signifikan secara statistik meskipun kekuatannya hanya lemah hingga sedang. Selain itu, hasil uji t dan Mann-Whitney menunjukkan bahwa harga rumah (House) dan unit (Unit) berbeda secara signifikan, meskipun pengaruhnya tidak terlalu besar. Secara keseluruhan, jarak dan tipe properti terbukti mempengaruhi harga, tetapi faktor lain seperti luas bangunan dan kondisi lingkungan juga perlu dipertimbangkan untuk analisis yang lebih lengkap.

# **BAB III – Hasil dan Pembahasan**

## **3.1 Visualisasi Data Bersih**

1. **Bar Chart Korelasi antar Variabel**

Bar chart ini menampilkan kekuatan korelasi antara variabel numerik terhadap *Price*. Hasil menunjukkan *BuildingArea* dan *Rooms* berkorelasi positif kuat dengan harga, sedangkan *Distance* berkorelasi negatif. Artinya, rumah yang lebih luas dan memiliki lebih banyak kamar cenderung lebih mahal, sementara lokasi yang lebih jauh dari pusat kota memiliki harga lebih rendah.

1. **Heatmap Korelasi antar Variabel**

Heatmap menggambarkan hubungan antar variabel numerik dalam bentuk matriks warna. Pola warna memperjelas korelasi kuat antara *Price* dengan *Rooms* dan *BuildingArea*, serta korelasi negatif dengan *Distance*. Tidak ada lagi gangguan nilai ekstrem, menandakan data sudah bersih dan siap untuk analisis.

1. **Boxplot Harga Berdasarkan Tipe Properti**

Boxplot menunjukkan distribusi *Price* berdasarkan *Type*. Tipe *house (h)* memiliki median harga tertinggi, diikuti *townhouse (t)*, sedangkan *unit (u)* terendah. Sebaran data terlihat lebih stabil tanpa outlier mencolok, menandakan pembersihan data berhasil dan pola harga antar jenis properti dapat terlihat jelas.

## **3.2 Analisis Statistik**

1. **Uji Parametrik: Pearson Correlation**

Uji ini mengukur hubungan antara *Distance* dan *Price*. Nilai korelasi menunjukkan arah negatif dengan kekuatan sedang, serta *p-value < 0.05*, yang berarti hubungan ini signifikan secara statistik.

Interpretasinya: semakin jauh jarak rumah dari pusat kota, semakin rendah harganya. Interval kepercayaan 95% memperkuat kestabilan hasil tersebut.

1. **Uji Parametrik: Independent t-test**

Uji ini membandingkan rata-rata harga antara *house (h)* dan *unit (u)*.

Hasil menunjukkan *p-value < 0.05*, menandakan perbedaan harga yang signifikan antara kedua jenis properti.

Harga rata-rata *house* lebih tinggi dibanding *unit*, yang berarti tipe properti mempengaruhi nilai jual.

1. **Uji Non-Parametrik: Mann-Whitney U Test**

Sebagai alternatif *t-test*, uji ini dilakukan tanpa asumsi normalitas.

Hasil menunjukkan *p-value < 0.05*, sehingga perbedaan distribusi harga antara *house* dan *unit* tetap signifikan. Nilai *effect size* berada pada kategori sedang, menandakan efek perbedaan harga cukup berarti secara praktis.

1. **Uji Non-Parametrik: Spearman Rank Correlation**

Uji ini mengukur hubungan monotonic antara *Distance* dan *Price*. Nilai korelasi menunjukkan arah negatif dengan kekuatan sedang, serta *p-value < 0.05*.

Interpretasinya sama seperti Pearson: semakin jauh lokasi, semakin rendah harga rumah. Hasil ini memperkuat konsistensi hubungan antar variabel.

## **3.3 Implikasi Hasil**

Dari seluruh hasil visualisasi dan uji statistik, dapat disimpulkan bahwa:

* Variabel *Distance* memiliki pengaruh negatif yang signifikan terhadap *Price*.
* Variabel *BuildingArea* dan *Rooms* berpengaruh positif terhadap harga properti.
* Jenis properti (*Type*) memberikan perbedaan harga yang nyata antara *house*, *unit*, dan *townhouse*.

Secara keseluruhan, hasil menunjukkan bahwa proses pembersihan data berjalan efektif dan menghasilkan dataset yang valid untuk analisis, dengan pola hubungan yang jelas, konsisten, dan signifikan secara statistik.

# **BAB IV – Kesimpulan dan Saran**

## **4.1 Kesimpulan**

Setelah dilakukan proses pembersihan data, dataset properti Melbourne menjadi lebih bersih dan rapi tanpa adanya nilai kosong atau data ekstrem. Data sudah siap digunakan untuk analisis karena semua variabel berada dalam skala yang sama dan tidak ada gangguan dari outlier.

Hasil visualisasi menunjukkan bahwa luas bangunan (*BuildingArea*) dan jumlah kamar (*Rooms*) berhubungan positif dengan harga rumah (*Price*), sedangkan jarak dari pusat kota (*Distance*) berhubungan negatif. Artinya, rumah yang lebih besar dan dekat dengan pusat kota cenderung memiliki harga lebih tinggi.

Perbandingan antarjenis properti juga menunjukkan hasil yang jelas. Tipe *house* memiliki harga tertinggi, disusul *townhouse*, dan yang terendah adalah *unit*. Uji statistik baik parametrik maupun non-parametrik menunjukkan hasil yang konsisten dan signifikan, artinya hubungan dan perbedaan tersebut benar-benar terjadi dalam data.

Secara keseluruhan, dapat disimpulkan bahwa jarak, luas bangunan, dan tipe properti adalah faktor utama yang memengaruhi harga rumah di Melbourne. Proses pembersihan dan analisis data yang dilakukan telah menghasilkan temuan yang akurat dan dapat dipercaya.

## **4.2 Saran**

Perusahaan properti sebaiknya lebih fokus pada pengembangan dan pemasaran rumah yang memiliki lokasi strategis dan ukuran bangunan yang lebih besar. Hasil analisis menunjukkan bahwa faktor jarak dari pusat kota dan luas bangunan sangat mempengaruhi harga jual. Oleh karena itu, properti yang dekat dengan pusat kota dapat diposisikan sebagai produk premium, sedangkan yang lebih jauh bisa ditawarkan dengan harga lebih kompetitif atau dilengkapi fasilitas tambahan agar tetap menarik bagi pembeli.

Selain itu, perusahaan perlu memantau tren harga dan permintaan pasar secara berkala menggunakan analisis data. Dengan memanfaatkan data yang bersih dan terstruktur, perusahaan dapat mengambil keputusan harga, promosi, dan pengembangan proyek dengan lebih tepat. Pendekatan berbasis data ini akan membantu perusahaan meningkatkan efisiensi, memahami kebutuhan konsumen, dan memperkuat posisi di pasar properti.