**PRAKTIKUM DATA WAREHOUSING DAN DATA MINING**

**MODUL 10**

**REGRESI LINIER SEDERHANA**

****

**Disusun oleh:**

**Adinda Aulia Hapsari**

**L200220037**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS KOMUNIKASI DAN INFORMATIKA**

**UNIVERSITAS MUHAMMADIYAH SURAKARTA**

**TAHUN 2024**

|  |  |
| --- | --- |
| Setelah kegiatan selesai, lembar kerja ini dicetak (di-print) dan dikumpulkan ke asisten.  NIM : L200220037  Nama : Adinda Aulia Hapsari  Nama Asisten : Diva Halimah  Tanggal Praktikum : 6 Desember 2024 | (Diisi oleh Asisten)  Nilai Praktek :  Tanda Tangan : |

**KEGIATAN PRAKTIKUM**

**Contoh Kasus:**

Perusahaan asuransi kesehatan hanya dapat menghasilkan uang jika pemasukan yang didapatkan lebih banyak dibanding dengan biaya yang dikeluarkan untuk pembiayaan kesehatan pasien penerima manfaat. Namun, memprediksi biaya kesehatan sangatlah sulit karena setiap individu memiliki karakteristik yang berbeda-beda. Tujuan dari eksperimen ini adalah untuk memprediksi biaya asuransi secara akurat berdasarkan data orang, termasuk usia, indeks masa tubuh, merokok atau tidak, dll. Selain itu, kita juga akan menentukan variabel terpenting yang mempengaruhi biaya asuransi. Prediksi ini akan bermanfaat bagi perusahaan asuransi untuk menentukan besaran premi yang harus dibayarkan per bulannya. Dari deskripsi di atas dapat kita simpulkan bahwa permasalahan kasus ini dapat diselesaikan dengan teknik regresi.

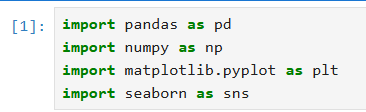
Dataset yang akan kita gunakan pada eksperimen ini tersedia di tautan yang tertera di atas dengan nama file insurance.csv. Dataset ini memiliki 1338 baris dan 7 kolom. Berikut deskripsi dari masing-masing kolom:

1. Usia: usia penerima manfaat asuransi
2. Jenis kelamin: jenis kelamin penerima manfaat asuransi
3. BMI : Indeks Massa Tubuh, memberikan pemahaman tentang berat badan yang relatif tinggi atau rendah relatif terhadap tinggi badan, indeks objektif berat badan (kg/m²) menggunakan rasio tinggi terhadap berat, idealnya 18,5 hingga 24,9
4. Anak: jumlah anak yang ditanggung oleh asuransi Kesehatan
5. Perokok: merokok atau tidak
6. Wilayah: daerah perumahan penerima di Amerika Serikat yang meliputi timur laut, tenggara, barat daya, barat laut.
7. Biaya: biaya pengobatan individu yang ditagih oleh asuransi Kesehatan

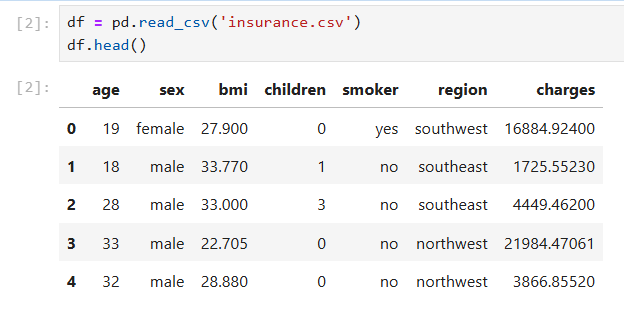
Karena kita akan memprediksi biaya asuransi, maka yang akan menjadi target fitur (fitur yang akan diprediksi) adalah kolom biaya.

**Memprediksi biaya asuransi berdasarkan fitur-fitur yang tersedia.**

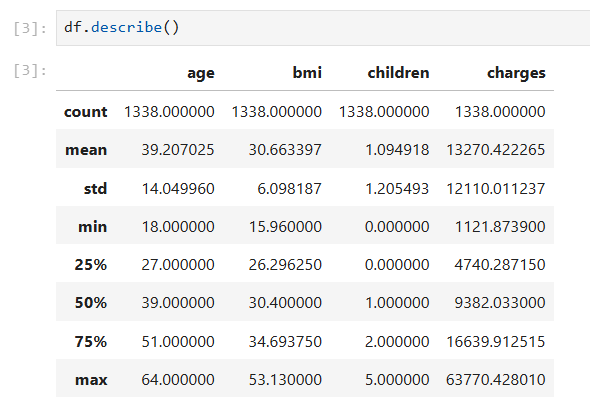
1. Pada Langkah pertama kita perlu memasukan beberapa library dan kelas yang akan kita gunakan pada eksperimen ini. Berikut beberapa library yang akan kita gunakan.



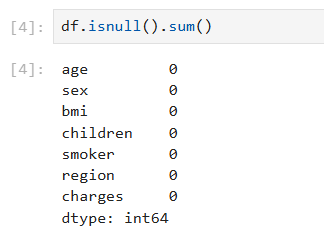
1. Selanjutnya, kita akan membaca dataset yang sudah kita unduh dengan menggunakan library pandas. Untuk membaca dataset dan melihat 5 data teratas



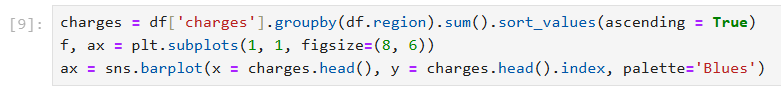
1. Selanjutnya kita juga bisa melihat gambaran dan ringkasan secara umum dari dataset yang kita gunakan.



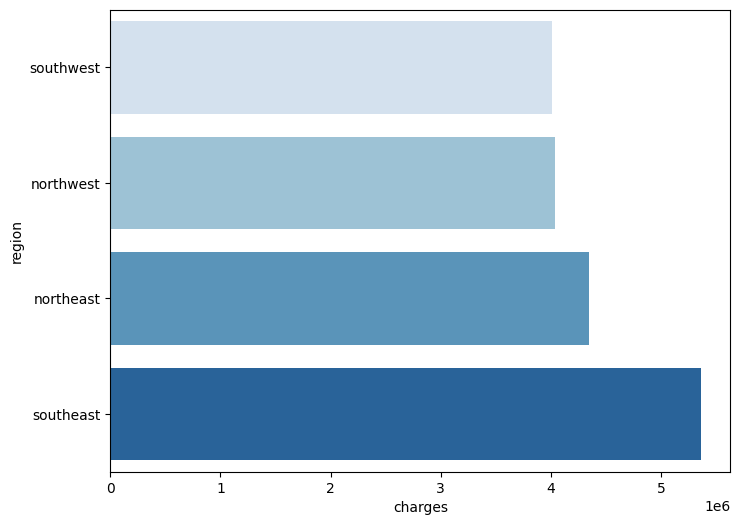
1. Sebelum kita bisa menggunakan dataset, kita perlu memastikan dulu bahwa tidak ada data yang null (kosong). Sehingga dapat dipastikan bahwa data konsisten dan tidak menyebabkan error.



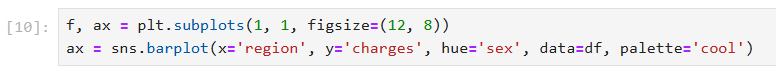
1. Selanjutnya kita bisa melakukan beberapa visualisasi terhadap data yang ada untuk meningkatkan pemahaman kita tentang dataset yang akan digunakan dalam eksperimen. Pertama kita bisa lihat distribusi biaya kesehatan untuk masing-masing daerah dengan kode dibawah ini.

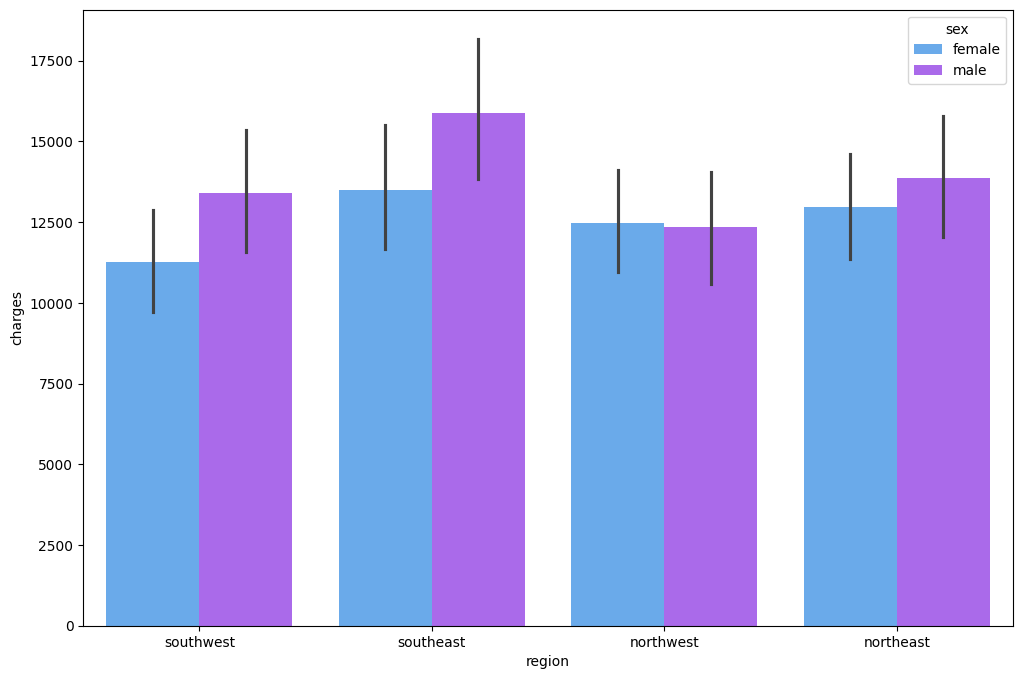


Jadi secara keseluruhan biaya pengobatan tertinggi ada di wilayah southeast dan terendah di southwest.



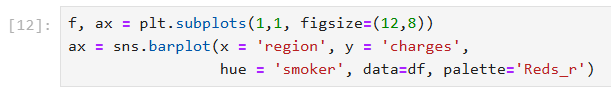
1. Selanjutnya kita juga bisa melakukan analisa terhadap data-data yang lain meliputi jenis kelamin, merokok atau tidak dan jumlah anak (khusus data yang dalam format kategorikal). Dalam hal ini kita akan menganalisa data tersebut berdasarkan wilayah yang berbeda. Sekarang kita mulai dengan jenis kelamin terlebih dahulu.

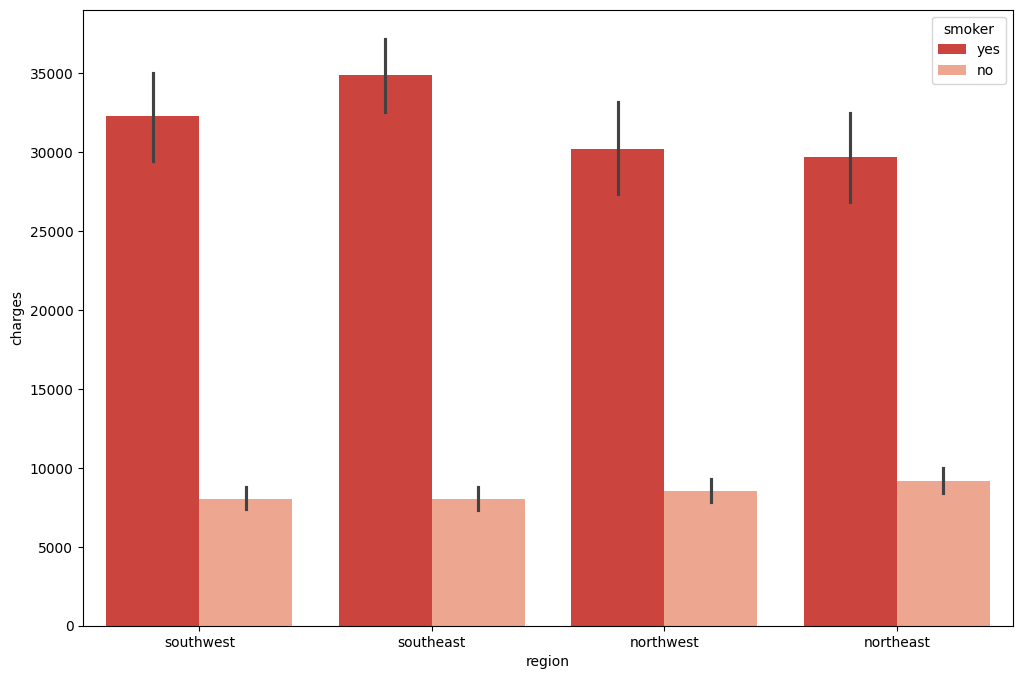




Dari gambar di atas dapat dilihat bahwa pasien laki-laki memiliki biaya lebih banyak di 3 dari 4 wilayah.

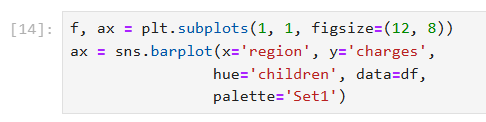
1. Selanjutnya kita akan melihat pengaruh variabel merokok atau tidak terhadap biaya kesehatan. Berikut potongan kode yang digunakan untuk memvisualisasikan pengaruh variabel merokok.

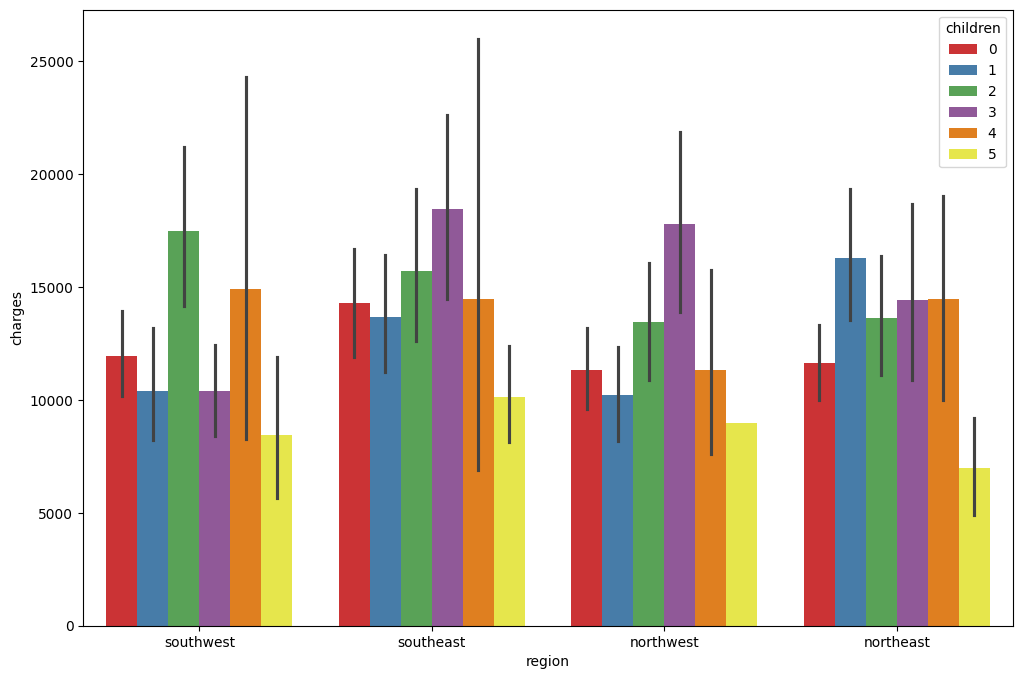




Dari gambar di atas dapat dilihat bahwa pasien merokok jauh menggunakan biaya kesehatan yang lebih tinggi dibandingkan yang tidak merokok di semua wilayah.

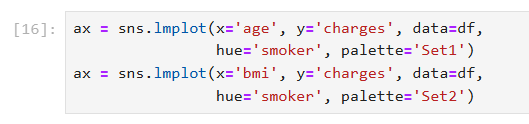
1. Terakhir kita akan melihat pengaruh jumlah anak yang dimiliki oleh pasien terhadap biaya kesehatan yang diklaim. Untuk melihat analisa tersebut bisa dengan menggunakan kode di bawah ini.

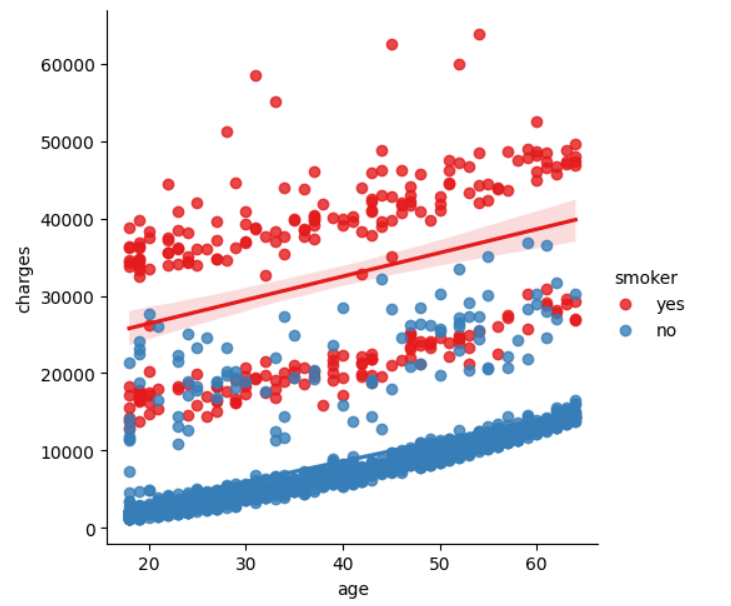


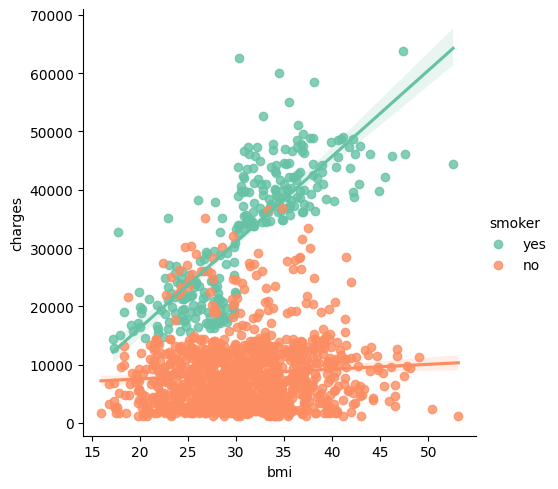


Dari gambar di atas dapat disimpulkan juga bahwa orang dengan anak-anak cenderung memiliki biaya medis yang lebih tinggi secara keseluruhan.

1. Dari sini kita sudah paham bahwa faktor merokok atau tidak merupakan variabel yang sangat penting dan berpengaruh terhadap biaya medis pasien. Untuk menganalisa dua variabel selanjutnya yaitu umur dan BMI, kita akan sekaligus menyertakan variabel merokok atau tidak. Untuk melakukan itu, silahkan gunakan kode di bawah ini.

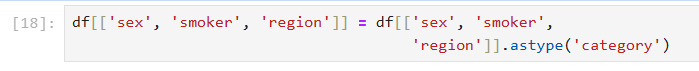






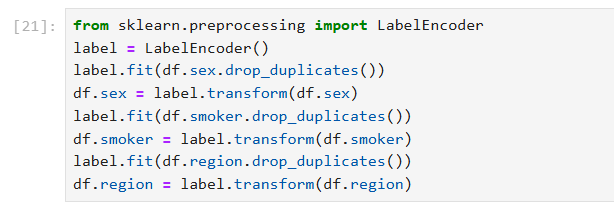
Dari kedua gambar di atas kita bisa simpulkan bahwa merokok memiliki dampak tertinggi pada biaya medis, meskipun biayanya meningkat seiring bertambahnya usia dan BMI.

1. Setelah kita selesai untuk menganalisa data dengan melakukan visualisasi dan interaksi antar data, selanjutnya kita akan melakukan prediksi biaya medis suatu pasien dengan menggunakan algoritma regresi linear. Langkah pertama yang harus kita lakukan adalah mengubah data yang yang sifatnya label atau teks menjadi data kategorikal. Dari data di atas yang memiliki data label ada 3 yaitu jenis kelamin, merokok atau tidak, dan wilayah. Untuk mengubah dari teks menjadi kategorikal dapat dilakukan dengan kode di bawah ini.

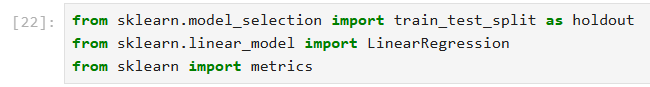


Kode di atas akan mengubah data jenis kelamin, merokok atau tidak, dan wilayah menjadi data kategorikal.

1. Selanjutnya kita perlu mengubah data yang sudah diubah menjadi kategorikal pada langkah sebelumnya ke data numerikal. Untuk melakukan hal tersebut, kita bisa gunakan library LabelEncoder seperti potongan kode di bawah ini.

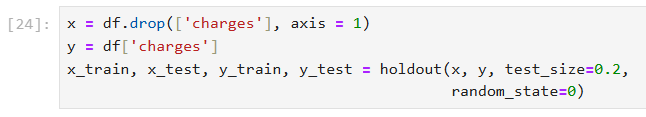


1. Untuk data yang lain tidak perlu dilakukan perubahan karena sudah dalam format numerikal. Setelah semua data telah siap digunakan, kita dapat menggunakannya untuk eksperimen prediksi biaya kesehatan. Seperti yang sudah disampaikan sebelumnya, bahwa kita akan menggunakan algoritma linear regression. Pertama kita perlu melakukan pemanggilan terhadap library yang akan digunakan untuk mengeksekusi algoritma linear regression dengan kode seperti berikut.



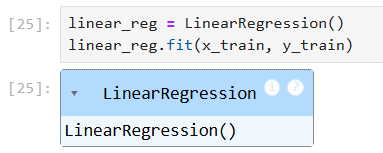
Baris pertama kode di atas digunakan untuk membagi data yang kita miliki menjadi dua porsi yaitu data training dan data testing. Baris kedua digunakan untuk memanggil algoritma linear regression. Baris ketiga digunakan untuk mengevaluasi performa dari algoritma linear regression dalam melakukan prediksi.

1. Langkah selanjutnya, kita perlu mendefinisikan data training dan data testing yang akan kita gunakan dalam eksperimen. Untuk melakukan itu, gunakan kode berikut ini.



Variabel ‘x’ digunakan untuk menampung atribut yang digunakan dalam eksperimen, di mana semua data digunakan kecuali data ‘charges’. Kemudian variabel ‘y’ digunakan untuk menampung atribut target, yaitu atribut yang akan diprediksi, di mana yang digunakan sebagai atribut target adalah data ‘charges’. Kode baris ketiga digunakan untuk membagi data menjadi data training dan data testing, dimana 80% data akan digunakan sebagai data training dan 20% data digunakan sebagai data testing.

1. Selanjutnya kita bisa lakukan proses training (fitting) model linear regression kita dengan menggunakan data training yaitu x\_train dan y\_train. Untuk melakukan training bisa digunakan kode di bawah ini.



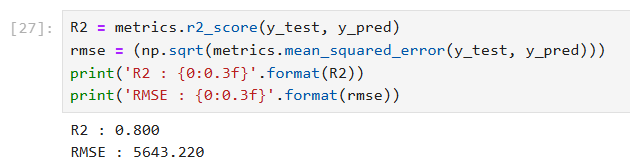
Baris kode pertama digunakan untuk menginisialisasi model linear regression. Selanjutnya baris kedua digunakan untuk melakukan training dengan menggunakan data training, dimana x\_train (berisi f itur yang akan dipelajari berisi data-data karakteristik pasien) akan di fitting dengan y\_train (berisi target fitur berupa biaya kesehatan).

1. Setelah proses training, kita bisa melakukan testing dengan menggunakan data testing yang telah kita persiapkan. Pada proses testing ini, kita akan memprediksi data testing dengan menggunakan model linear regression yang telah kita train. Untuk melakukan testing dapat digunakan kode di bawah ini.

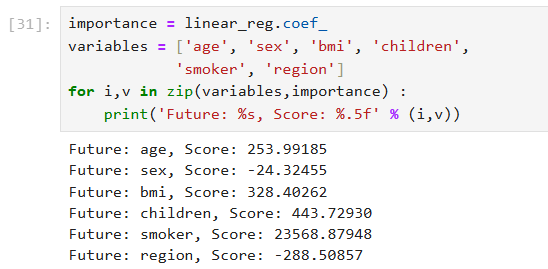


Hasil prediksi akan ditampung di variabel ‘y\_pred’.

1. Selanjutnya sudah tentu kita ingin melakukan evaluasi terhadap performa prediksi model linear regression yang telah kita kembangkan. Untuk melakukan evaluasi regresi, metrik yang paling umum digunakan adalah R2 dan RMSE (root mean squared error). Gunakan kode berikut untuk mengevaluasi hasil prediksi.



1. Selanjutnya kita juga masih bisa mencoba meningkatkan performa model linear regression dalam melakukan prediksi dengan mengeliminasi fitur yang tidak terlalu berpengaruh. Sebelum mengeliminasi fitur yang kurang penting, kita perlu mengetahui peranan masing-masing fitur dalam eksperimen ini. Untuk melihat peranan masing-masing fitur, gunakan kode di bawah ini.



1. Dari hasil analisa peranan masing-masing fitur dapat disimpulkan bahwa jenis kelamin dan wilayah memiliki peranan yang tidak signifikan dalam menentukan biaya medis pasien asuransi. Selanjutnya kita juga bisa menyimpulkan bahwa faktor merokok atau tidaknya pasien merupakan fitur paling penting untuk menentukan biaya medis pasien. Penemuan ini sesuai dengan analisa yang telah kita lakukan pada bagian awal eksperimen ini.

**TUGAS**

**Dikerjakan saat ini, jika tidak selesai bisa dilanjutkan di rumah.**

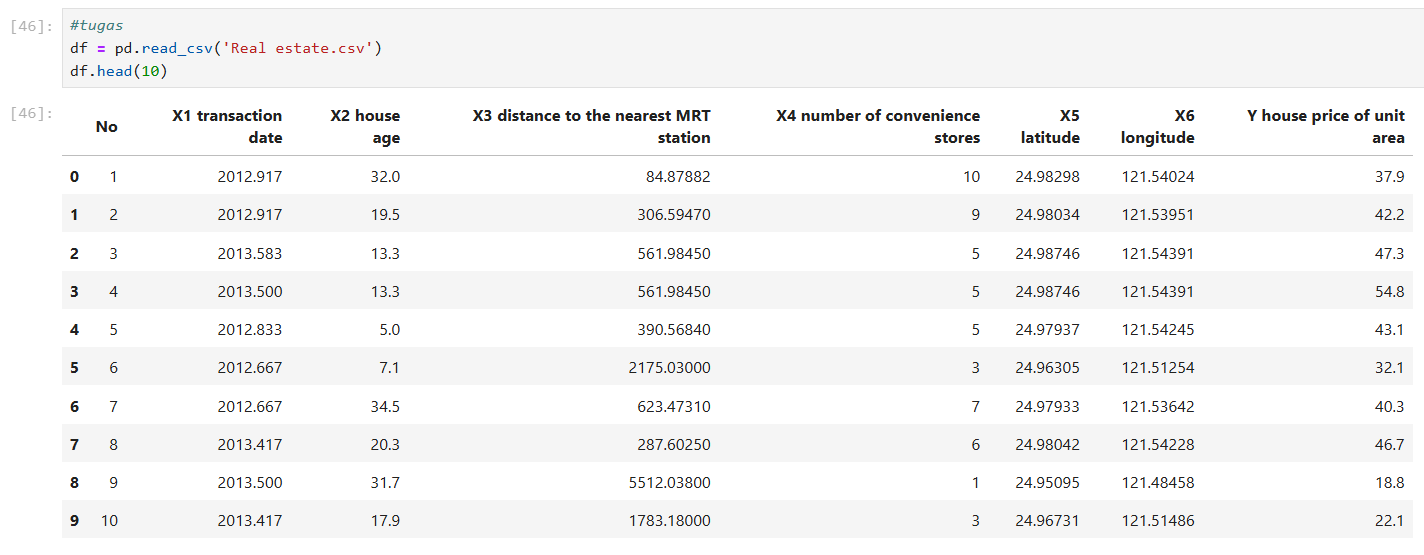
**Kasus:**

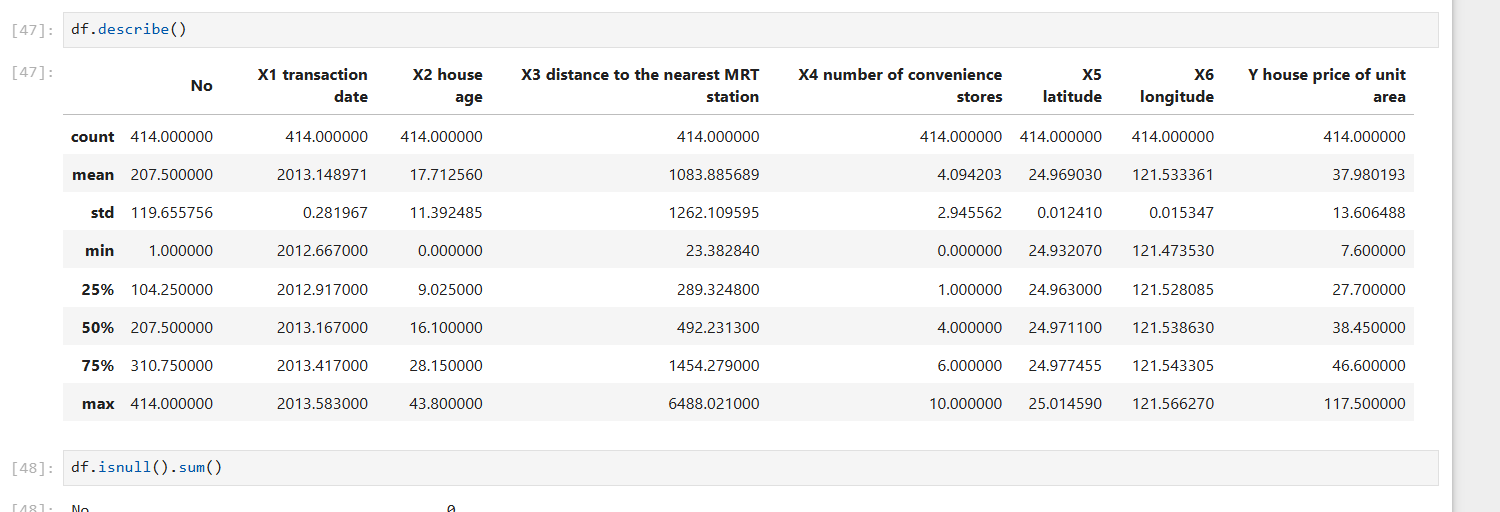
Pada tugas ini, anda akan diberikan sebuah dataset tentang karakteristik beberapa rumah beserta dengan harganya. Tugas anda adalah memprediksi harga rumah dengan menggunakan algoritma regresi linear. Dataset dapat diunduh pada tautan yang tercantum bagian alat dan bahan dengan nama file Real estate.csv. Pada dataset ini terdapat beberapa kolom diantaranya:

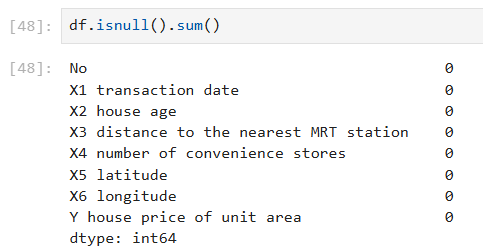
1. transaction date
2. house age
3. distance to the nearest MRT station
4. number of convenience stores
5. latitude
6. longitude
7. house price of unit area

Dari data di atas data nomor 7 yang akan menjadi kelas target. Silahkan ikuti langkah-langkah praktikum di atas untuk mengerjakan tugas ini. Beberapa hal yang perlu kalian perhatikan adalah:

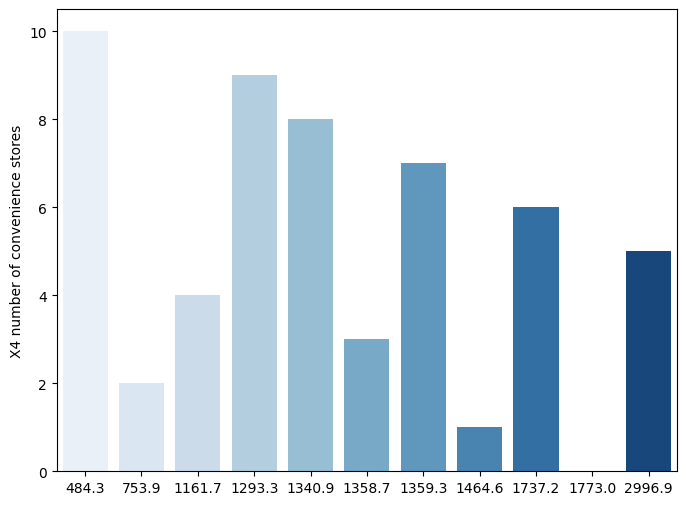
1. Silahkan kerjakan tugas ini dengan mengikuti semua langkah langkah di atas. Untuk visualisasi dan analisa data (langkah 5 sampai langkah 9) bersifat opsional. Namun yang mengerjakan langkah 5 sampai 9 akan mendapatkan nilai yang maksimal. 188 Data Warehousing dan Data Mining (Modul Praktikum)

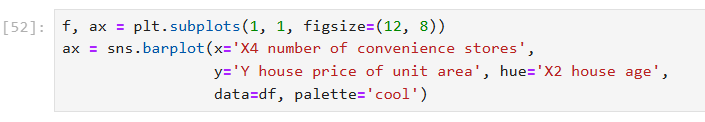


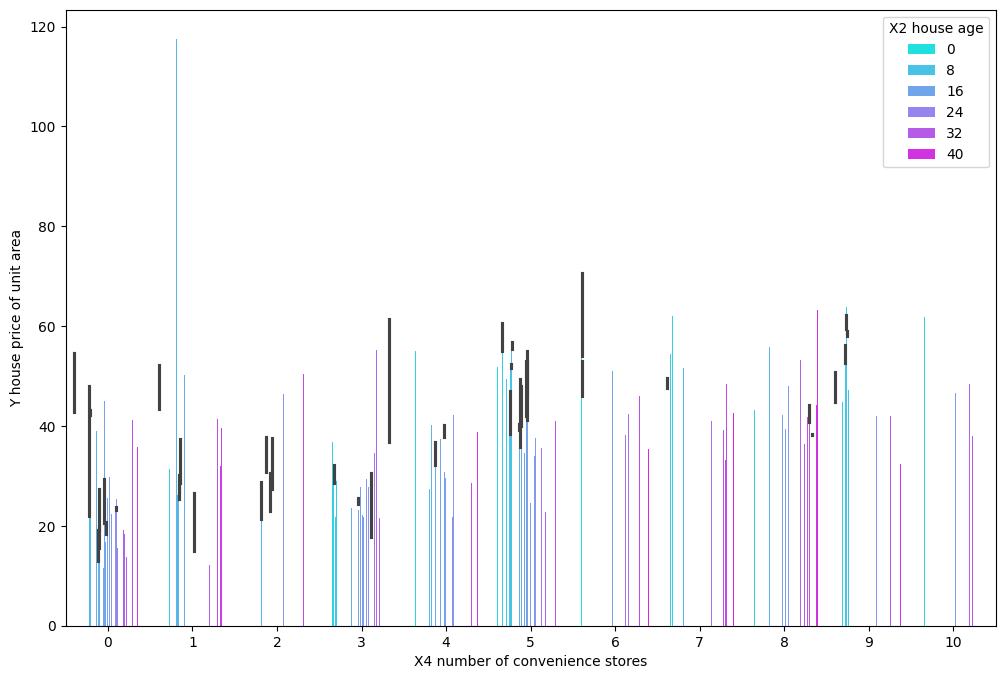


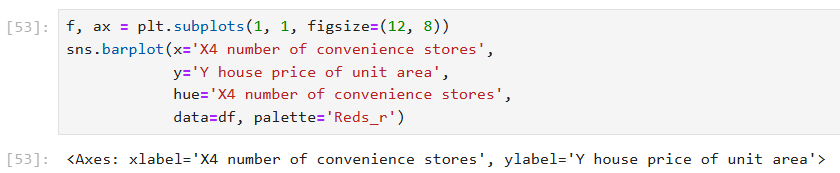


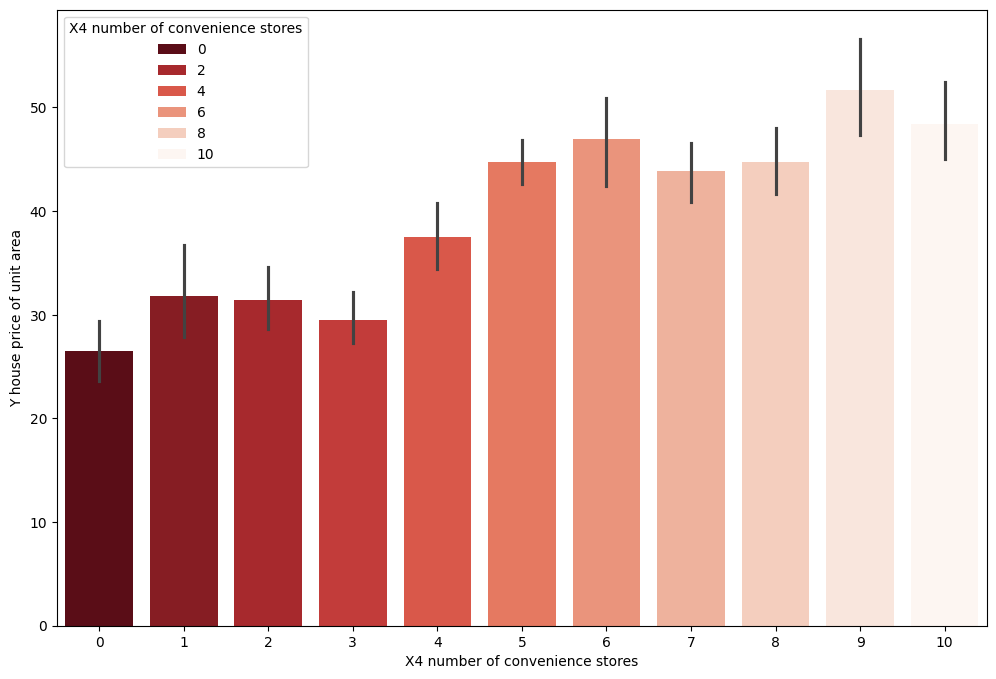




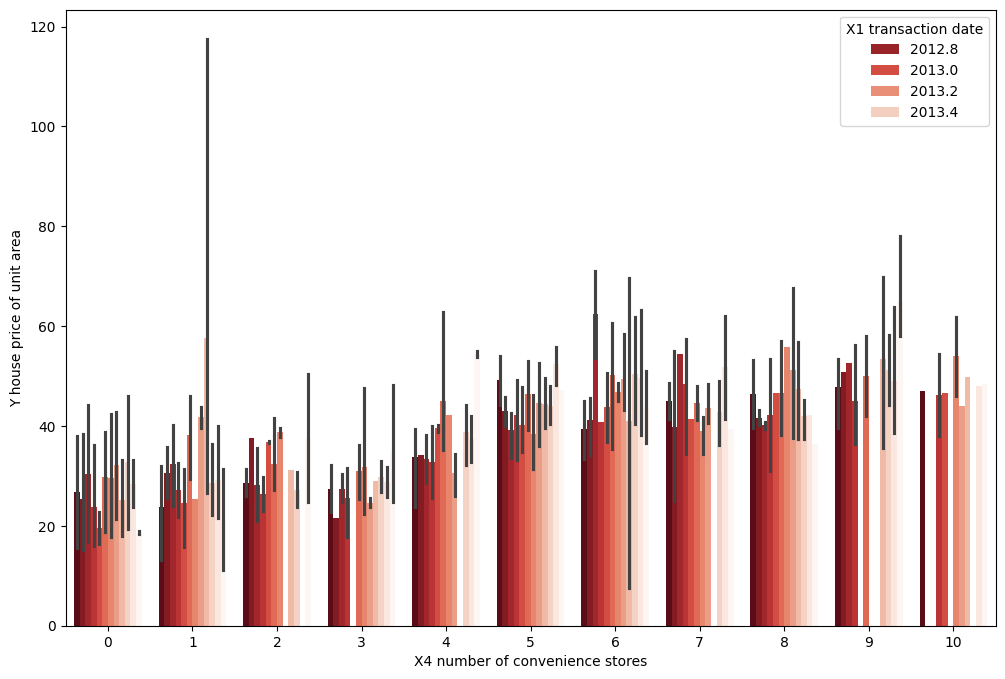


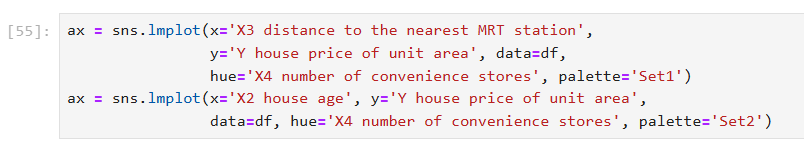


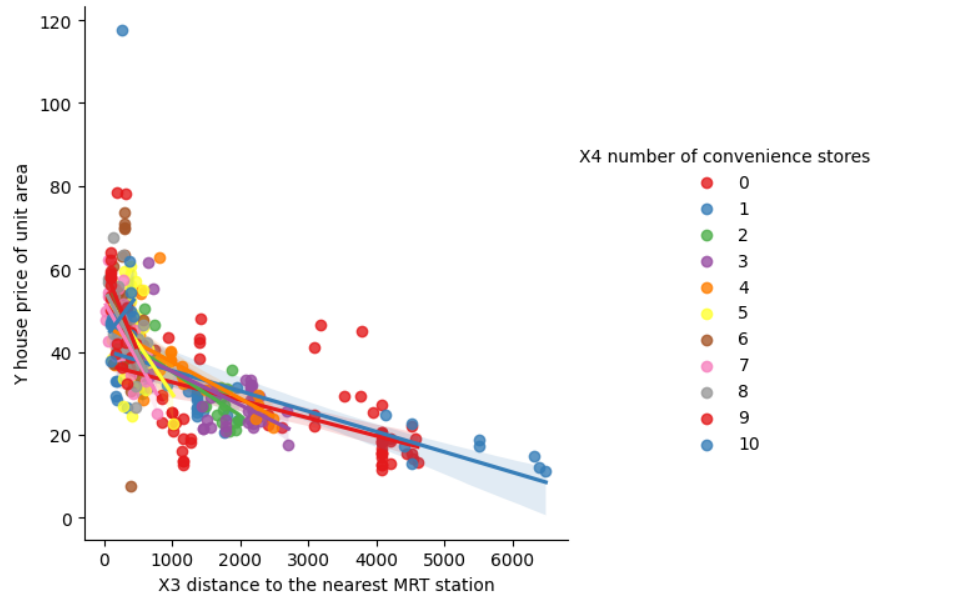


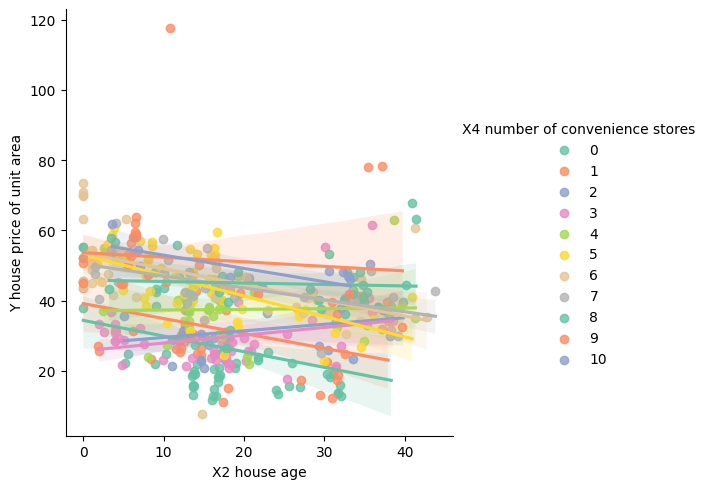


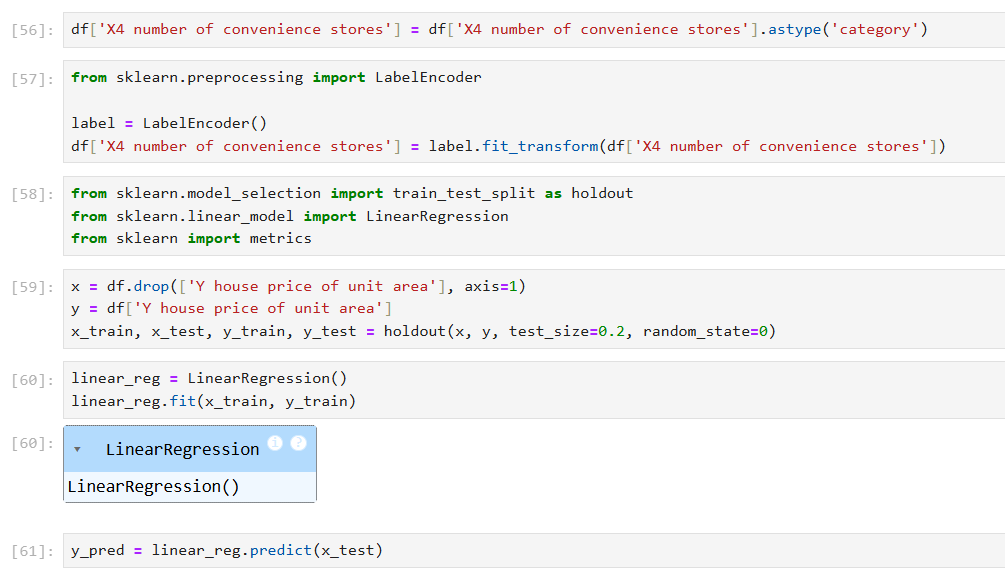


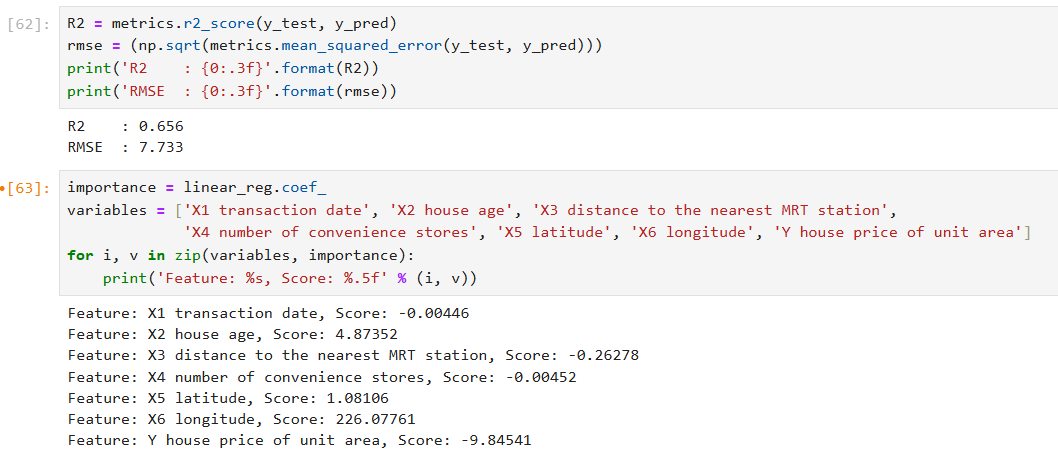












1. Jika menurut anda ada fitur atau data yang tidak berpengaruh terhadap harga rumah, anda bisa mengeliminasi fitur tersebut. Kemudian bandingkan hasilnya dengan ketika menggunakan semua fitur.



