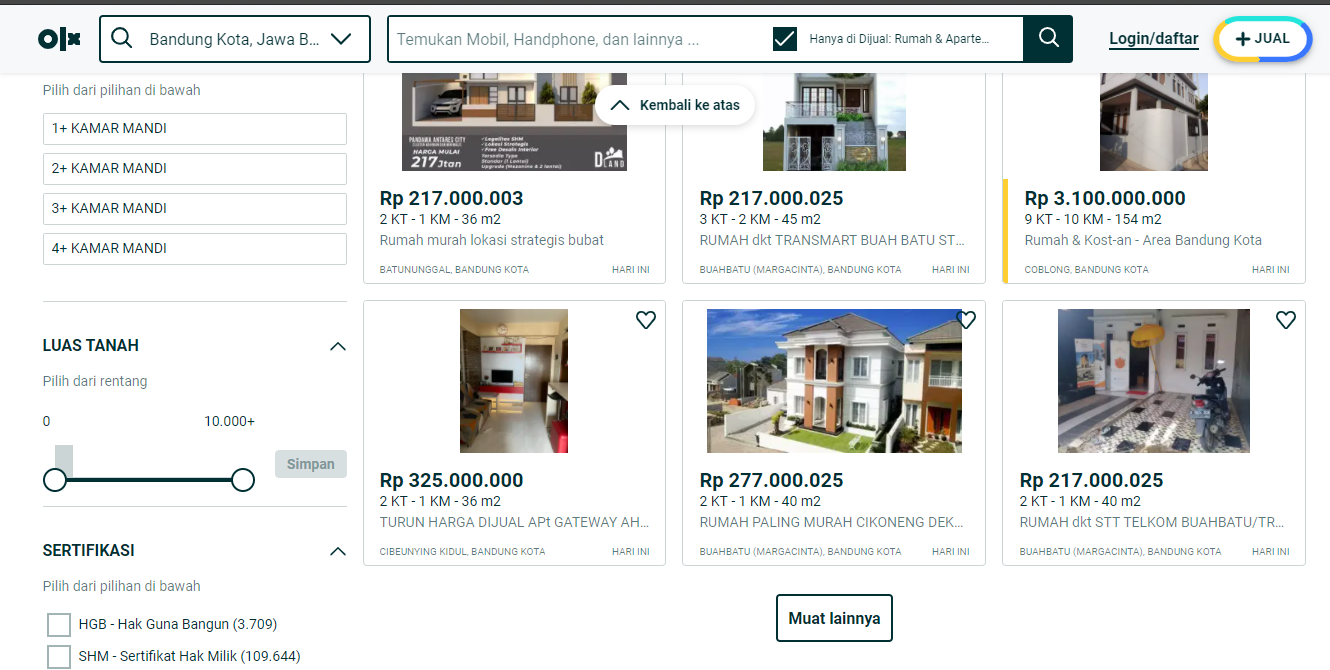
**Analisis Harga Rumah di Wilayah Kota Bandung dan Sekitarnya**

1. **Pengumpulan Data**

Data yang digunakan pada analisis ini adalah data yang diambil dari aplikasi web OLX dengan filter Rumah dijual di Kota Bandung. Data diambil menggunakan *library* bahasa pemrograman python yaitu *beautifulsoup* dan *selenium*. *Selenium* digunakan untuk melakukan interaksi dengan ajax yang ada pada web tersebut, dimana pada OLX terdapat tombol ‘muat lainnya’ untuk memperoleh item yang lebih banyak sehingga dibutuhkan interaksi terhadap tombol tersebut.



Kemudian ditemukan bahwa pada OLX hanya dapat menampilkan 1000 item dalam satu halaman, sehingga data dibagi lagi menjadi beberapa filter berdasarkan harga, yaitu:

* price\_max\_250000000',
* price\_between\_250000000\_to\_400000000',
* price\_between\_400000000\_to\_500000000',
* price\_between\_500000000\_to\_650000000',
* price\_between\_650000000\_to\_750000000',
* price\_between\_750000000\_to\_950000000',
* price\_between\_950000000\_to\_1500000000',
* price\_between\_1500000000\_to\_3000000000',
* price\_between\_3000000000\_to\_5000000000',
* price\_min\_5000000000'

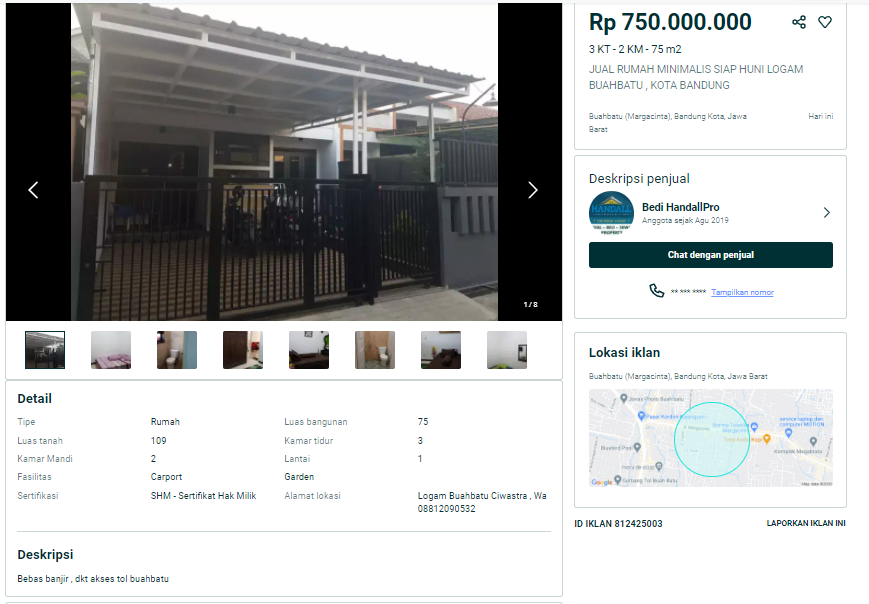
sehingga diperoleh 8597 item. Kemudian item-item tersebut diambil masing-masing url yang menuju halaman produknya menggunakan *beautifulsoup*, hal ini bertujuan untuk memperoleh data yang lebih lengkap karena apabila hanya mengambil data dari itembox saja hanya akan memperoleh data seperti yang ditampilkan dibawah.



Link-link tersebut kemudian dibuka menggunakan urlopen, dan beautifulsoup untuk memperoleh data-data sebagai berikut:

* 'ID':ID Iklan,
* 'title': Judul Iklan,
* 'bedroom': Jumlah Kamar,
* 'bathroom': Jumlah Kamar Mandi,
* 'landsqr': Luas Tanah,
* 'buildingsqr': Luas Bangunan,
* 'cert': Sertifikat,
* 'loc': Lokasi1,
* 'location': Lokasi2,
* ‘floor': Lantai,
* ‘postTime': Waktu iklan dipasang,
* ‘numOfFac': jumlah fasilitas,
* ‘fac': fasilitas,
* ‘desc': deskripsi iklan,
* ‘price': harga

Data yang diperoleh berjumlah 8304, jumlah tidak sama dengan url yang ada dikarenakan beberapa iklan sudah dihapus oleh pemiliki iklan.



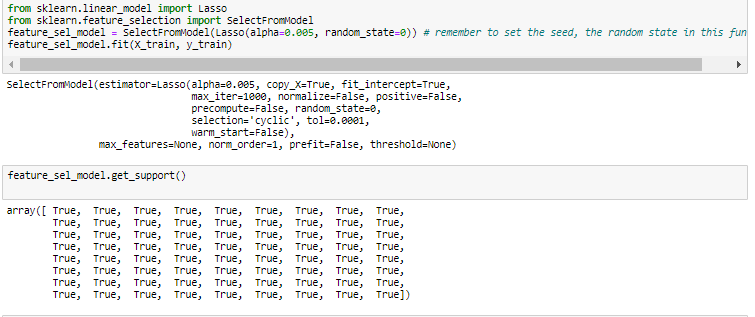
1. **Praproses dan Pembersihan Data**

Terdapat cukup banyak nilai null pada data tersebut hal ini disebabkan karena pemilik iklan yang tiadk mencantumkan data secara lengkap, sehingga beberapa data tersebut dihapuskan, selain itu terdapat beberapa iklan dengan data yang sama (data ganda), data anomali seperti luas tanah 900000m2, jumlah lantai 999 lantai, hal ini kemungkinan diakibatkan pemilik iklan yang mengisi iklan secara asal, selain itu juga terdapat data yang merupakan apartemen sehingga memiliki spesifikasi yang berbeda. Untuk menangani permasalah tersebut data-data tersebut dihapus, sehingga jumlah data yang terakhir ada 7107 baris, proses ini dilakukan menggunakan ms excel dan jupyter notebook.

Untuk melakukan *training* dan membuat model maka atribut yang digunakan adalah bedroom, bathroom, landsqr, buildingsqr, cert, location, floor, numOfFac, dan price, dimana price adalah variabel dependennya. Kemudian data-data tersebut diambil untuk proses *feature* *engineering* untuk selanjutnya dibuat model berdasarkan data-data tersebut. Selain itu juga dilakukan analisis korelasi serta analisis harga rumah terhadap lokasi rumah tersebut.

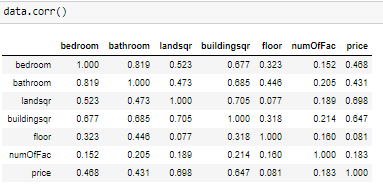
1. ***Feature Engineering***

Dalam proses ini dilakukan dua pekerjaan, yang pertama adalah one hot encoding terhadap data kategori, yaitu location dan cert, dan yang kedua adalah feature selection dimana kedua pekerjaan ini dilakukan untuk memperoleh model yang lebih baik. One hot encoding menambahkan jumlah atribut menjadi 73 atribut total. Proses *feature* *selection* dilakukan menggunakan Lasso regression dengan parameter alpha = 0.005, dan random\_state = 0, dimana hasilnya adalah semua atribut dapat digunakan. Sehingga tidak ada atribut yang digunakan pada proses training model regresi.



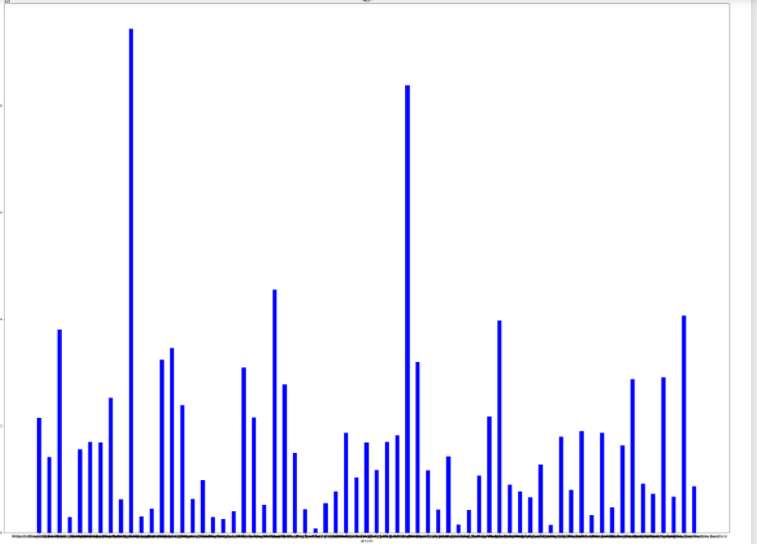
1. **Training dan Analisis**
2. Analisis Korelasi

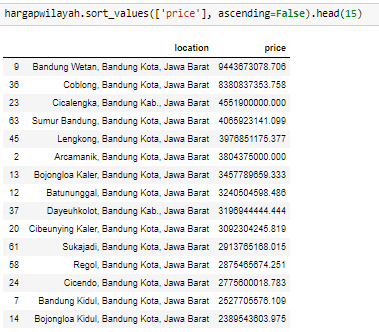
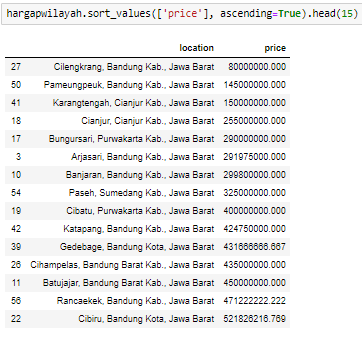
Pada analisis korelasi digunakan fungsi corr, untuk medapatakan nilai korelasi antar variabel, dapat dilihat bahwa korelasi harga terhadap landsqr dan building sqr cenderung signifikan dibandingkan dengan variabel yang lainnya.



1. Analisis harga terhadap lokasi

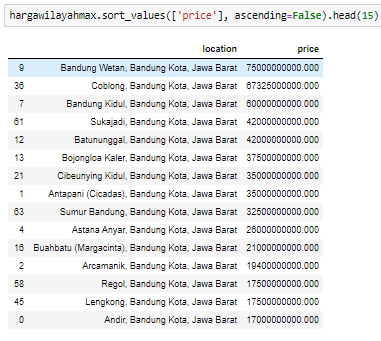
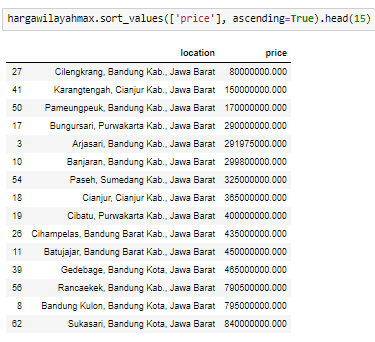
Pada tahapan ini dibagi menjadi 4 kategori yaitu lima belas teratas untuk lokasi dengan rata-rata harga rumah, lima belas teratas untuk lokasi dengan nilai maksimal harga rumah, lima belas terbawah untuk lokasi dengan rata-rata harga rumah, dan lima belas terbawah untuk lokasi dengan nilai maksimal harga rumah. Gambar di bawah ini merupakan plot dari rata-rata harga rumah berdasarkan lokasi, dapat dilihat terdapat beberapa lokasi yang memiliki nilai rata-rata cenderung jauh lebih tinggi dari pada lokasi lainnya.



Lima belas teratas dan lima belas terbawah lokasi dengan rata-rata harga rumah  

dapat dilihat bahwa top 15 dan bottom 15 data diatas top 15 cenderung merupakan pusat kota, dimana memiliki fasilitas publik yang lebih unggul, sehingga memiliki rata-rata harga yang cukup tinggi.

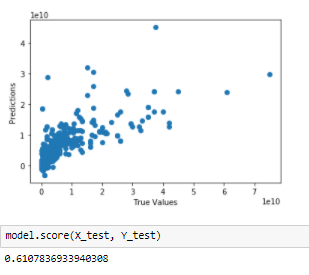
Lima belas teratas dan lima belas terbawah lokasi dengan nilai maximal harga rumah

1. Pembentukan Model

Dalam membentuk model digunakan 2 jenis algoritma, yang pertama adalah linear regression dan yang kedua adalah xgboost regressor. Sebelum melakukan pemodelan data sebelumnya dibagi terlebih dahulu menjadi data training dan data testing menggunakan train test split dari sklearn, dimana jumlah data tesnya yaitu 20 persen dari data yang ada sedangkan data trainingnya dalah 80 persen dari data yang ada.

* Linear Regression memperoleh score 0.61



* XGBoost dengan parameter objective ='reg:linear', colsample\_bytree = 0.5, learning\_rate = 0.5, max\_depth = 50, alpha = 10, n\_estimators = 10, memperoleh score 0.91

