Tugas 4 PDIB: Data Preparation

Oleh: Muhammad Iqbal Asrif

1. Pendahuluan

   Terdapat *raw data* harga rumah di India yang akan dilakukan *data preparation* terhadapnya. *Data preparation* mencakup mengumpulkan, menggabungkan, menyusun, dan mengatur data. *Preparation* ini sangat diperlukan karena *raw data* yang ada masih memiliki banyak *missing* atau *inconsistent values*. *Raw data* juga memiliki *outliers* yang perlu di-*handle* sesuai dengan kebutuhan. Adapula data yang tidak sesuai dengan data dunia nyata, sehingga perlu dilakukan riset terlebih dahulu untuk meng-handle data-data tersebut. Setelah *data preparation* selesai dilakukan, akan dilanjutkan dengan analisa data lebih lanjut untuk mendapatkan *insight* dan informasi yang lebih mendalam.

2. Deskripsi Data

   a. diposting_oleh: Berisi mengenai value pemilik properti rumah, memiliki nilai Dealer, Owner, Builder
   b. sedang_pembangunan: Status masa pembangunan, 1 berarti dalam masa pembangunan, 0 berarti tidak dalam masa pembangunan
   c. disetujui_pemerintah: Status persetujuan pemerintah, 1 berarti properti sudah disetujui, 0 berarti belum disetujui
   d. total_ruangan: Total banyaknya ruangan pada suatu properti
   e. tipe_a_atau_b: Tipe dari suatu properti rumah
      - 1 Tipe a = Terdiri atas 1 kamar tidur, 1 ruang tengah / aula, dan 1 dapur
      - 1 Tipe b = Terdiri atas 1 ruangan dan 1 dapur

      Perbedaan mendasar antara keduanya adalah tidak adanya aula di Tipe b. 1 Tipe a sangat cocok untuk keluarga kecil dan menengah ke bawah karena ada kamar, ruang tengah / aula, dan dapur terpisah sehingga keluarga pada akhirnya akan mengubah aula menjadi ruangan lain saat anak-anak keluarga tumbuh dewasa. Tapi di Tipe b hanya ada kamar dan dapur terpisah tanpa aula yang bisa diubah menjadi kamar lain untuk anak-anak Anda yang sudah dewasa, jadi mungkin hanya bagus untuk para bujangan. Tipe a menawarkan lebih banyak ruang sementara Tipe b terbatas dan biaya Tipe a biasanya lebih tinggi dari Tipe b. Sehingga biasanya disarankan untuk memiliki rumah Tipe a daripada Tipe b.

Apabila properti memiliki jumlah ruang selain 2 dan 3, maka akan dikategorikan sebagai selain Tipe A dan Tipe B

f. kaki_persegi: Luas total rumah dalam satuan square feet atau kaki persegi
g. siap_pindah: Status siap huni properti, 1 berarti siap huni, 0 berarti tidak siap huni
h. dijual_kembali: Status penjualan rumah, 1 berarti rumah yang dijual kembali, 0 berarti rumah baru belum pernah terjual
i. alamat: Alamat dari suatu properti
j. longitude: Longitude dari suatu properti
k. latitude: Latitude dari suatu properti
l. harga: Harga rumah dalam satuan rupee lakh, 1 lakh = 100000
m. harga_per_kaki_persegi: Harga rumah per kaki persegi, untuk membantu preparasi dan visualisasi data
n. harga_per_ruangan: Harga rumah per ruangan, untuk membantu preparasi dan visualisasi data

3. Data Preparation
   a. Cek missing values

```
def cek_null(df):
    col_na = df.isnull().sum().sort_values(ascending=True)
    percent = col_na / len(df)

    missing_data = pd.concat([col_na, percent], axis=1, keys=['Total', 'Percent'])

    if (missing_data[missing_data['Total'] > 0].shape[0] == 0):
        print("Tidak ditemukan missing value pada dataset")

    else:
        print(missing_data[missing_data['Total'] > 0])
```

```
cek_null(data)
```

   b. Handle inappropriate values

```
for i in data:
    display(f"{i}: {data[i].unique()}")
```

```
data['kaki_persegi'].mask(data['kaki_persegi'].isin(['1-1-1-',
'a', 'asik', 'awa', 'benar', 'salah', 'testes']), None,
inplace=True)
```

```
data['siap_pindah'].mask(data['siap_pindah'].isin(['testes', 'a',
'awa', '1-1-1-', 'benar', 'salah', 'asik']), None, inplace=True)
```

```
data['dijual_kembali'].mask(data['dijual_kembali'].isin(['Aku
paling benar', 'SALAH', 'Semangat Dek']), None, inplace=True)
```

c. Handle missing values

```
for i in ['total_ruangan', 'kaki_persegi', 'latitude']:
    data[i].fillna(data[i].median(), inplace=True)
```

```
for i in ['siap_pindah', 'dijual_kembali', 'sedang_pembangunan']:
    data[i].fillna(data[i].mode()[0], inplace=True)
```

d. Handle column data types

```
for i in ['sedang_pembangunan', 'total_ruangan', 'siap_pindah',
'dijual_kembali']:
    data[i] = data[i].astype(np.int64)

data['kaki_persegi'] = data['kaki_persegi'].astype(np.float64)
```

e. Handle inconsistent values

```
data['tipe_a_atau_b'].mask(data['total_ruangan'] == 3, 'tipe_a',
inplace=True)
```

```
data['tipe_a_atau_b'].mask(data['total_ruangan'] == 2, 'tipe_b',
inplace=True)
```

```
data['tipe_a_atau_b'].mask(~data['total_ruangan'].isin([2, 3]),
'selain tipe_a dan tipe_b', inplace=True)
```

f. Handle duplicates

```
data.drop_duplicates(inplace=True)
```

g. Create new features to help with outlier and data visualization

```
data['harga_per_kaki_persegi'] = data.apply(lambda row:
row['harga'] / row['kaki_persegi'], axis=1)
```

```
data['harga_per_ruang'] = data.apply(lambda row: row['harga'] /
row['total_ruangan'], axis=1)
```

h. Handle outlier

```
Q1 = data['harga_per_kaki_persegi'].quantile(0.25)
Q3 = data['harga_per_kaki_persegi'].quantile(0.75)
IQR = Q3 - Q1
```

```
data = data[~((data['harga_per_kaki_persegi'] < (Q1 - 1.5 * IQR))
| (data['harga_per_kaki_persegi'] > (Q3 + 1.5 * IQR)))]
data.shape
```

```
Q1 = data['harga_per_ruang'].quantile(0.25)
Q3 = data['harga_per_ruang'].quantile(0.75)
IQR = Q3 - Q1
```

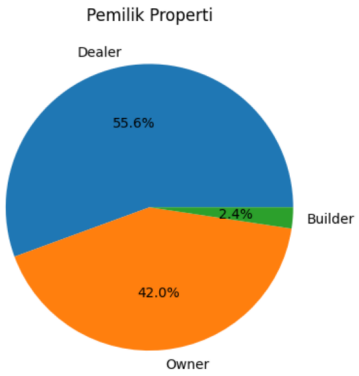
```
data = data[~((data['harga_per_ruang'] < (Q1 - 1.5 * IQR)) |
(data['harga_per_ruang'] > (Q3 + 1.5 * IQR)))]
data.shape
```

```
Q1 = data['kaki_persegi'].quantile(0.25)
Q3 = data['kaki_persegi'].quantile(0.75)
IQR = Q3 - Q1
```

```
data = data[~((data['kaki_persegi'] < (Q1 - 1.5 * IQR)) |
(data['kaki_persegi'] > (Q3 + 1.5 * IQR)))]
data.shape
```

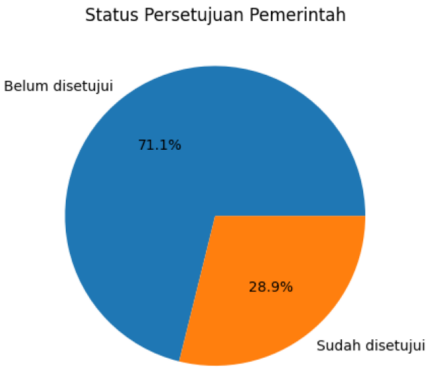
4. Hasil dan Temuan

a. Distribusi Pemilik Properti



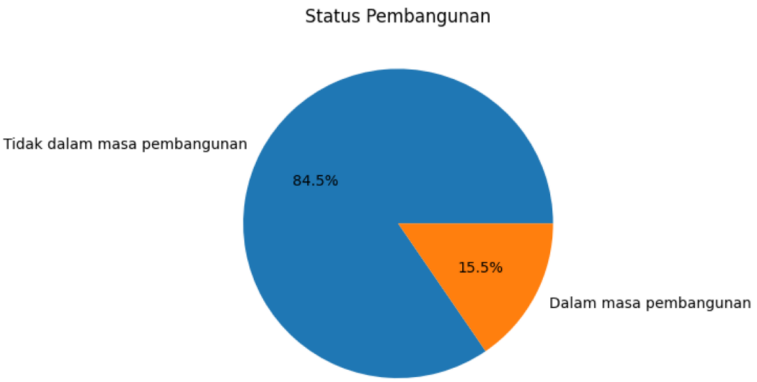
Terlihat bahwa mayoritas pemilik yang memposting propertinya adalah dealer, dengan jumlah sebanyak 55.6%, diikuti dengan owner sebanyak 42%. Hanya sedikit builder yang memposting propertinya yaitu sebanyak 2.4%

b. Status Persetujuan Pemerintah



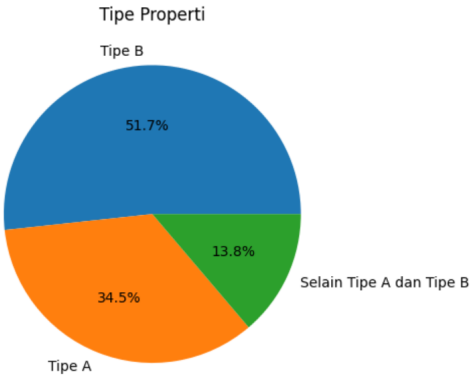
Lebih dari 2/3 izin properti belum disetujui pemerintah, dengan properti yang sudah memiliki perizinan hanya sebanyak 28.9%

c. Status Pembangunan



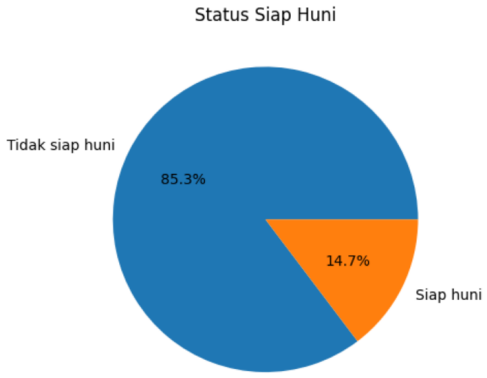
Lebih dari 4/5 properti sudah selesai dibangun, dengan 15.5% properti dalam masa pembangunan

d. Tipe Properti



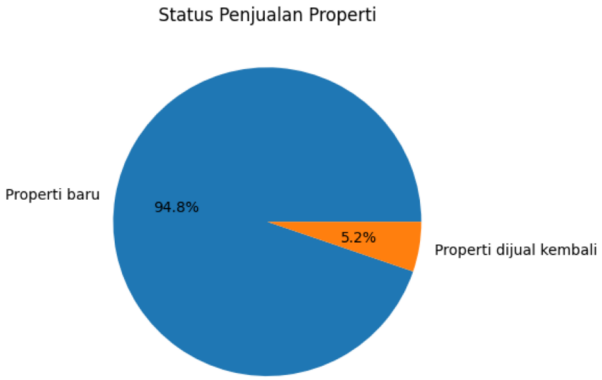
Properti Tipe B, dengan harga yang lebih murah daripada Tipe A, mendominasi tipe properti yang ada, dengan Tipe B sebanyak 51.7% dan Tipe A sebanyak 34.5%. Adapun properti yang tidak dalam kategori Tipe A dan Tipe B ada sebanyak 13.8%

e. Status Siap Pindah



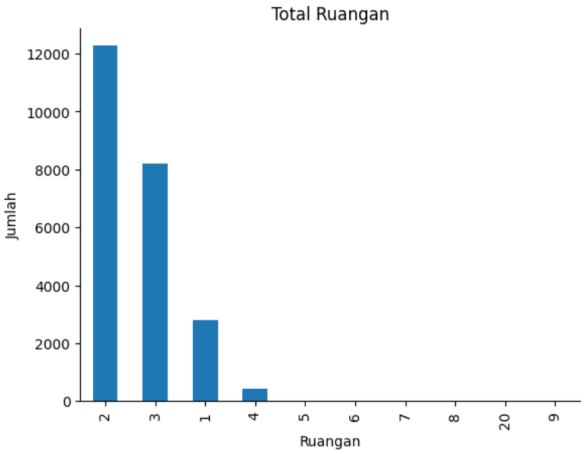
Lebih dari 4/5 properti belum siap huni, dengan jumlah properti siap huni hanya sebanyak 14.7%

f. Status Penjualan Rumah



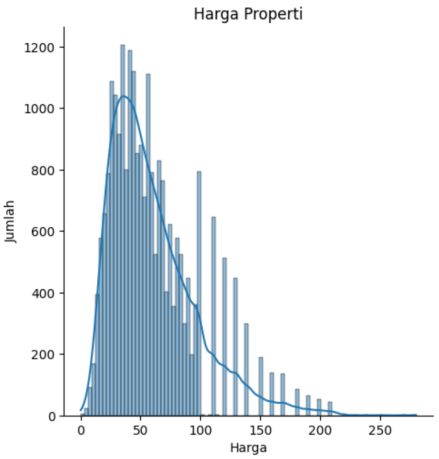
Hampir seluruh properti yang dijual merupakan properti baru, dengan jumlah properti yang dijual kembali hanya sebanyak 5.2%

g. Total Ruangan



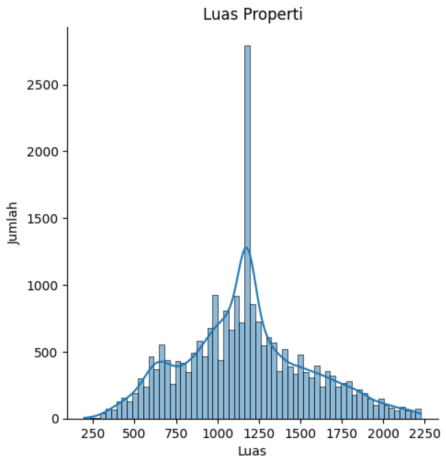
Mayoritas properti memiliki 2 dan 3 ruangan, yaitu Tipe B dan Tipe A respectively. Adapun untuk properti selain Tipe A dan Tipe B, didominasi oleh properti dengan 1 dan 4 ruangan

h. Harga Properti



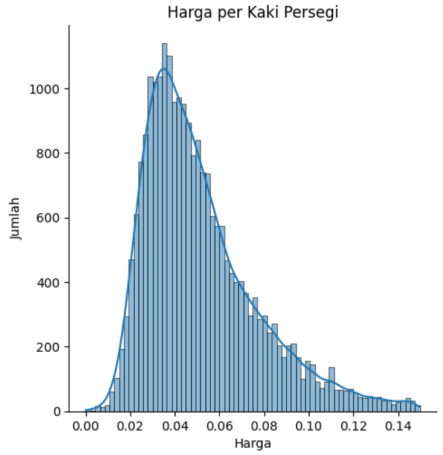
Mayoritas properti memiliki rentang harga antara 0 - 100, dengan properti murah yang memiliki harga mendekati 0 sangat sedikit. Adapun properti mewah dengan harga diatas 100 cukup banyak terutama pada rentang 100 - 200

i. Luas Rumah



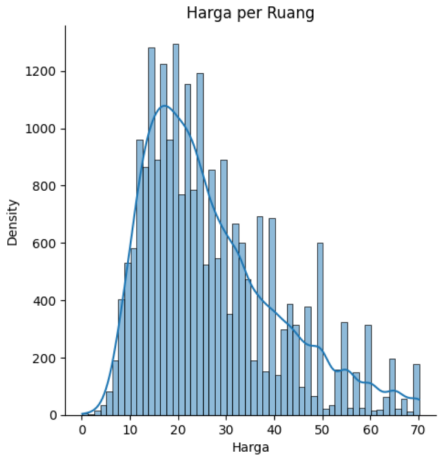
Terdapat sebuah outlier pada point luas antara 1000 - 1250 dimana jumlah properti mencapai lebih dari 2500. Mayoritas properti memiliki luas antara 750 - 1500. Properti dengan luas dibawah 750 hanya ada sedikit. Sedangkan properti dengan luas diatas 1500 cukup banyak, sampai dengan luas 2250

j. Harga Per Kaki Persegi



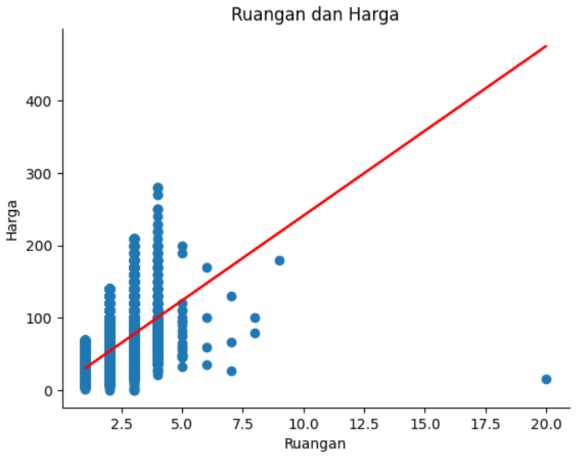
Mayoritas properti memiliki harga 0.02 - 0.10 Lakh Rupee. Adapun jumlah properti murah sangat sedikit jika dibandingkan dengan jumlah properti mewah yang memiliki harga jauh diatas median

k. Harga Per Ruang



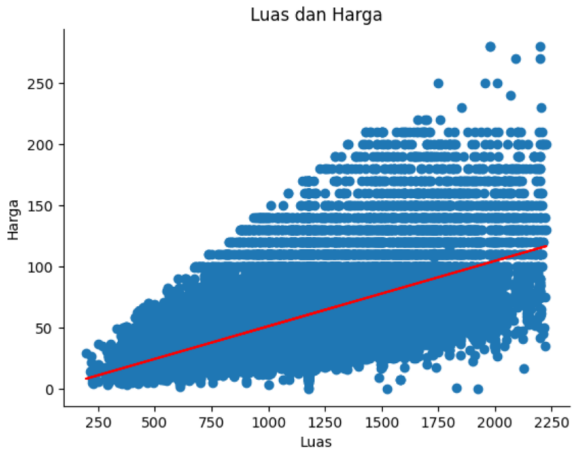
Mayoritas harga per ruang berada dalam rentang 10 - 50. Adapun jumlah properti yang memiliki harga per ruang murah sangat sedikit, jika dibandingkan jumlah properti yang memiliki harga per ruang yang mahal

l. Korelasi Ruang Dan Harga



Terlihat bahwa ruangan memiliki korelasi positif dengan nilai yang cukup tinggi terhadap harga, dimana semakin banyak ruangan maka harga akan meningkat dengan cepat. Adapun terdapat outlier properti dengan 20 ruangan namun dengan harga yang sangat rendah

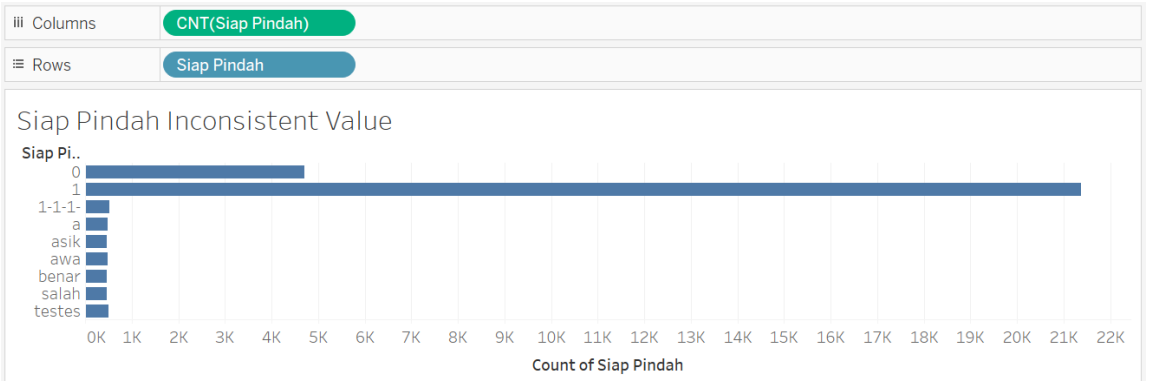
m. Korelasi Luas Dan Harga



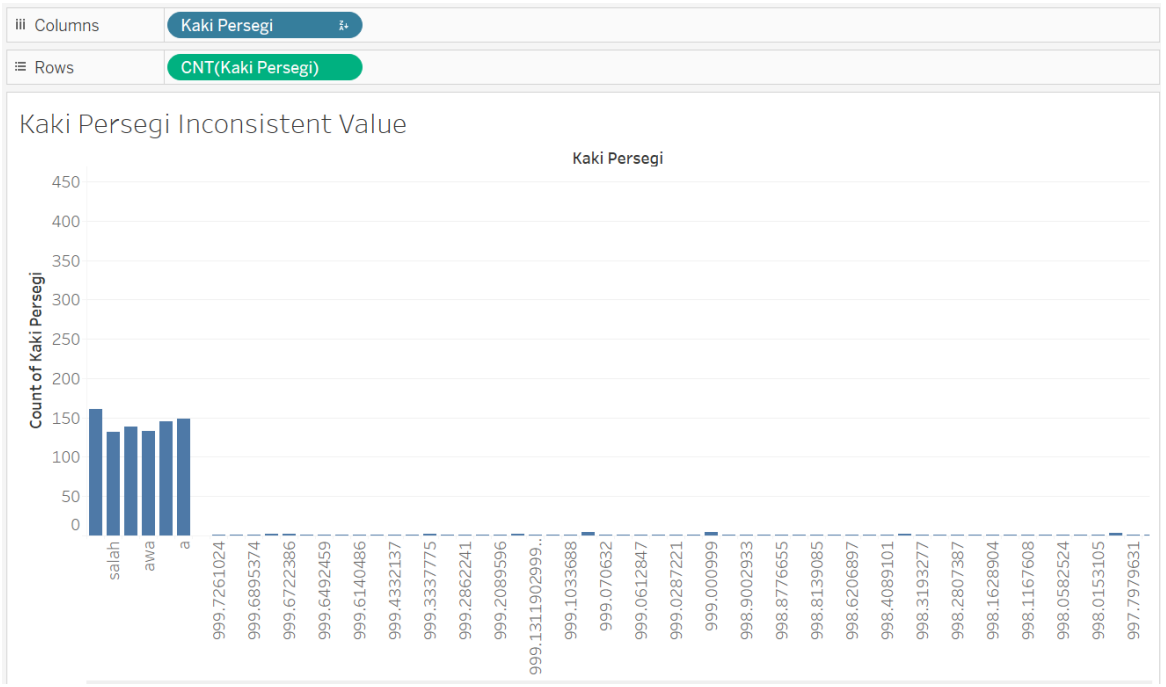
Terlihat bahwa luas memiliki korelasi positif terhadap harga, dimana semakin banyak ruangan maka harga akan meningkat

5. Tableau Data Visualization
    a. Data Noise / Outlier

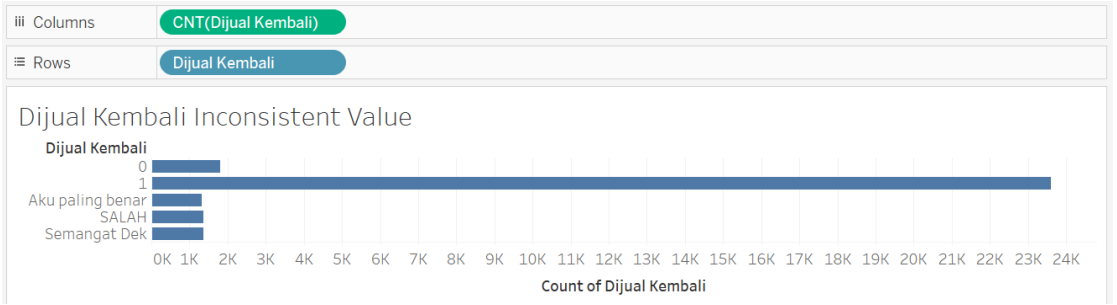
    siap_pindah noise



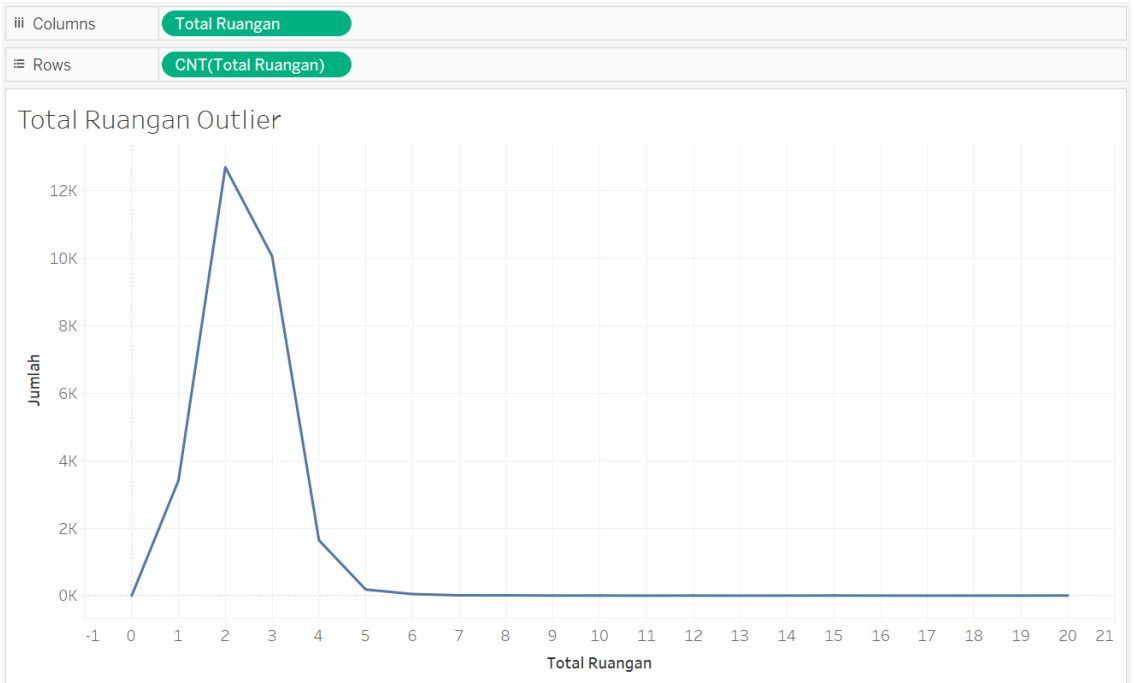
    kaki_persegi noise



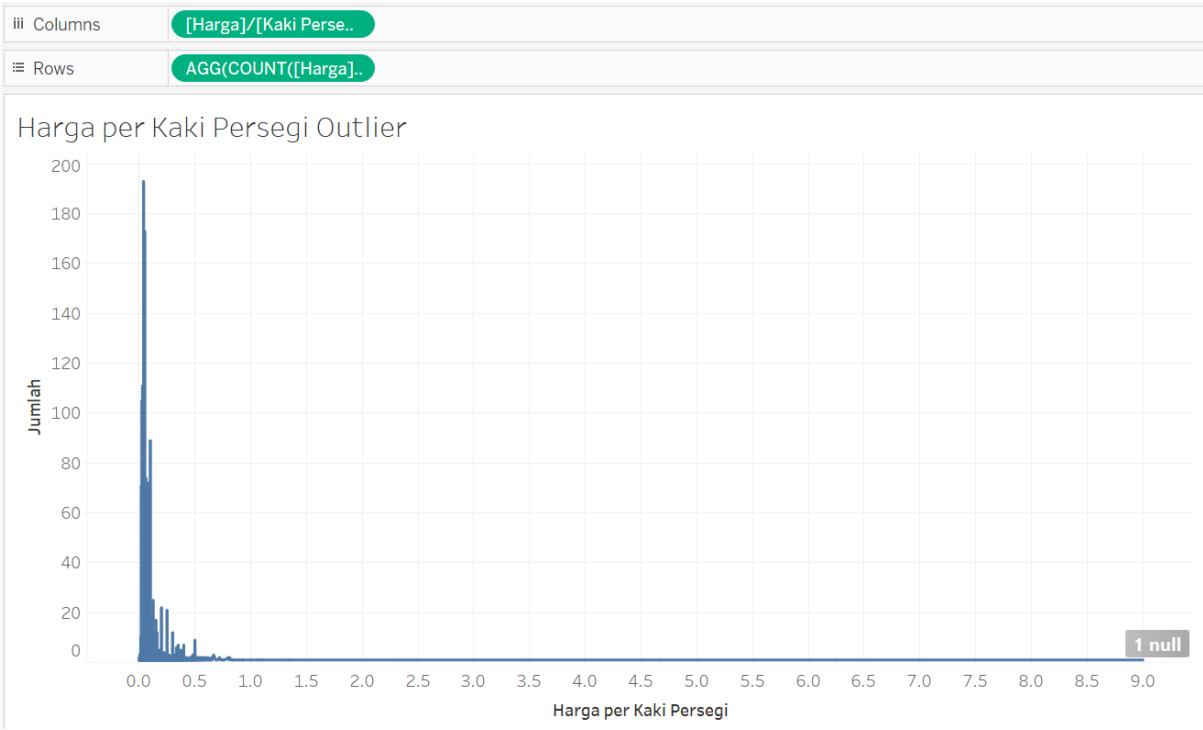
dijual_kembali noise



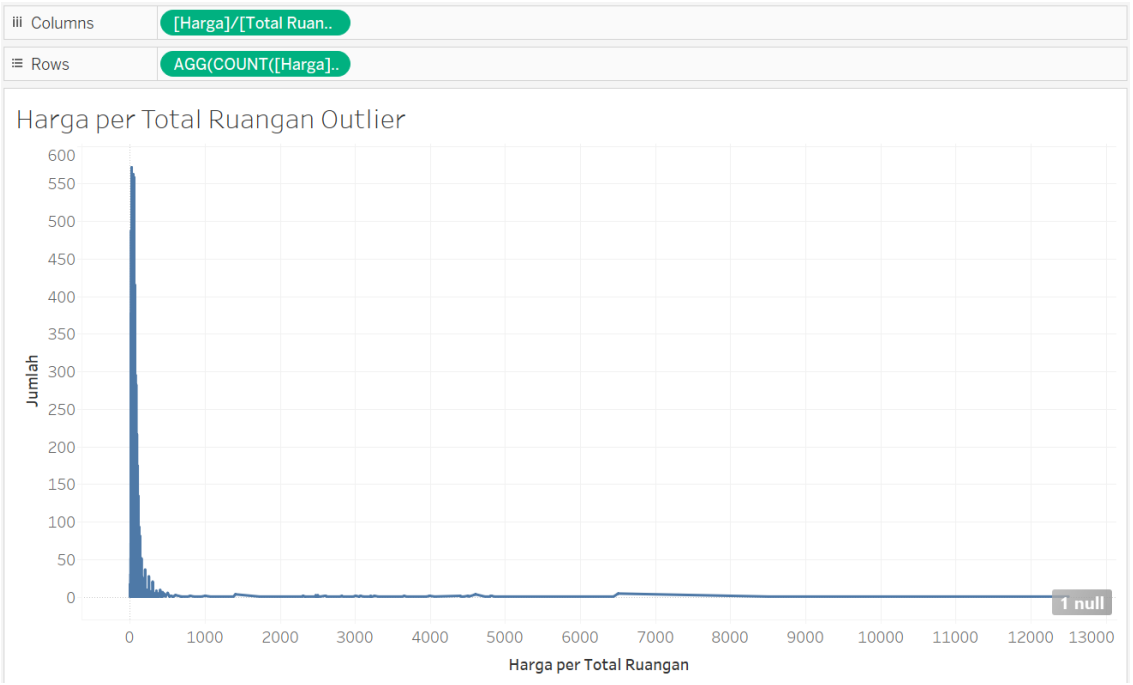
total_ruangan outlier



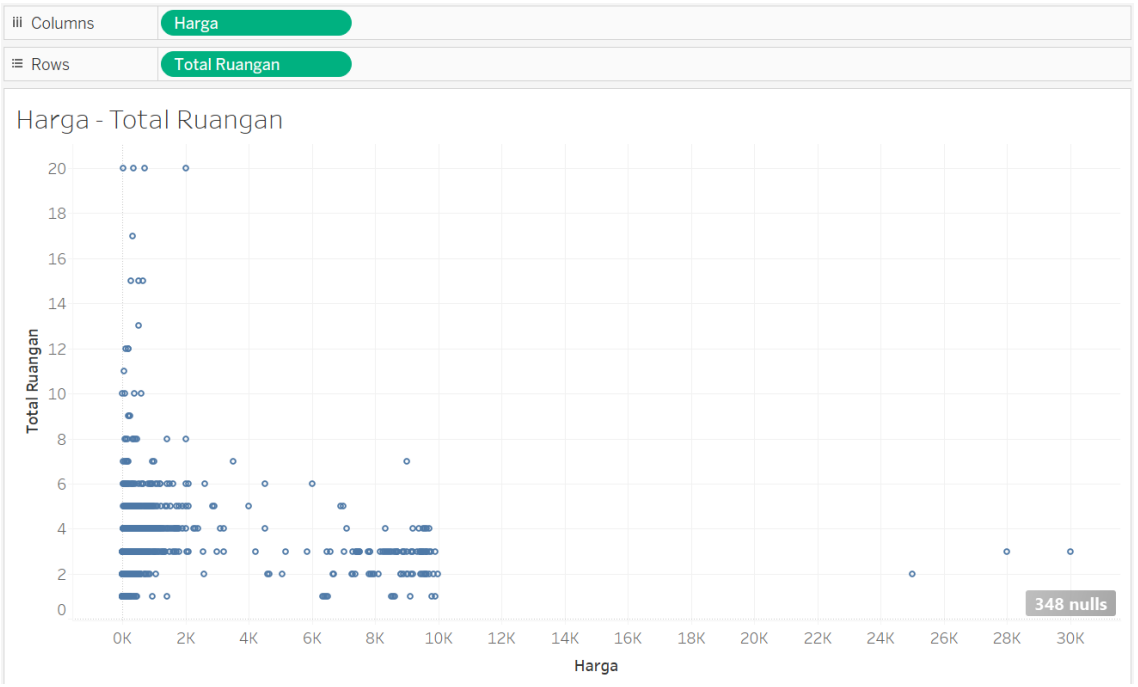
harga_per_kaki_persegi outlier and noise



harga_per_total_ruangan outlier and noise

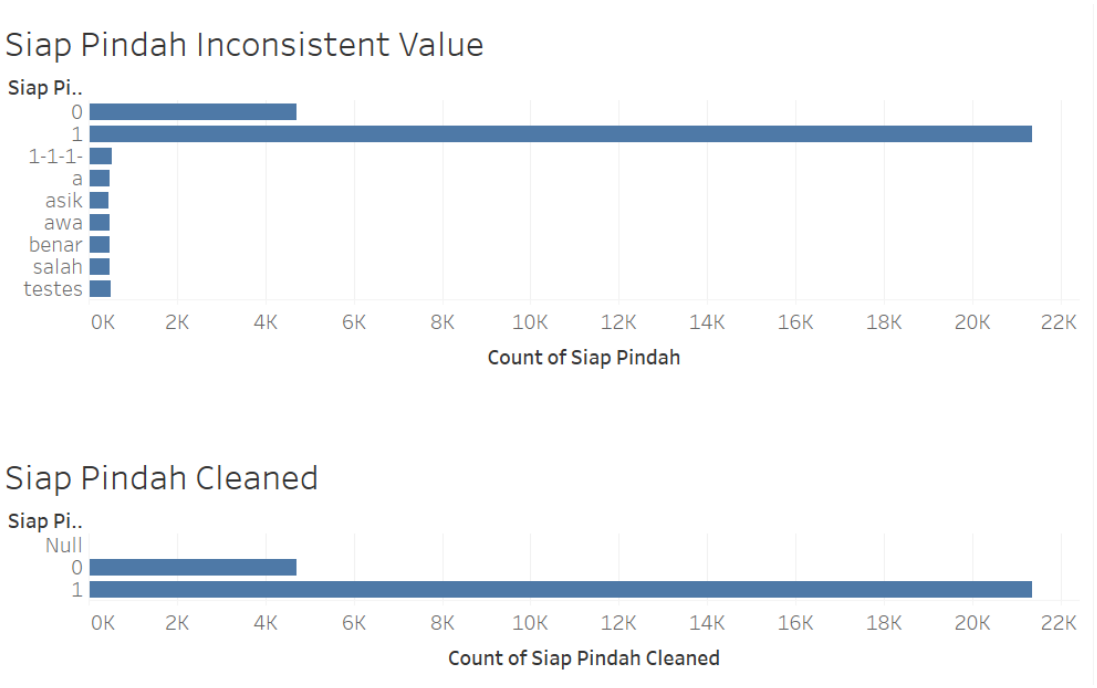


Harga - Total Ruangan outlier and noise

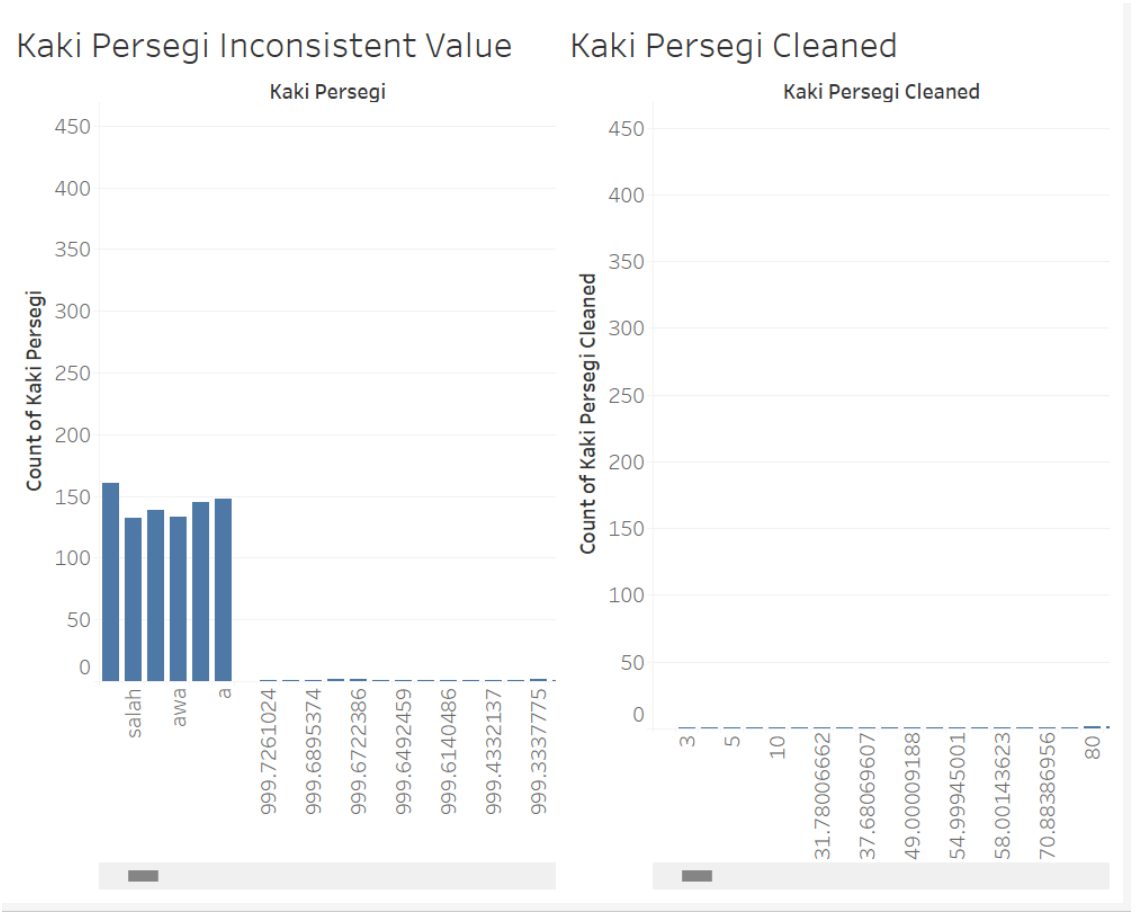


b. Before - After Data Preparation

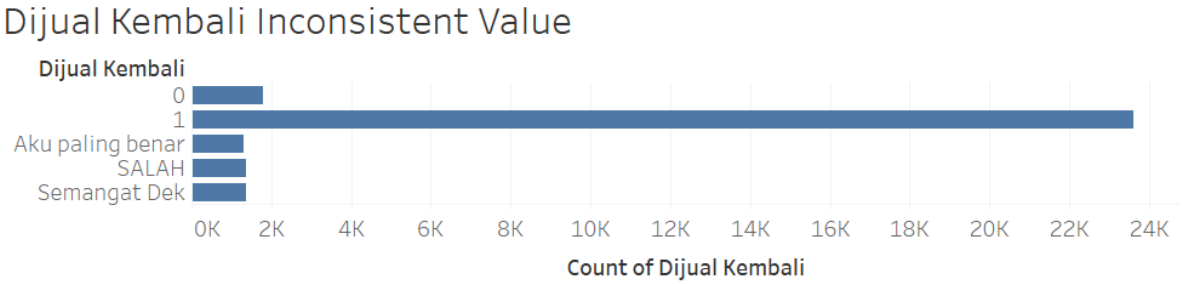
siap_pindah

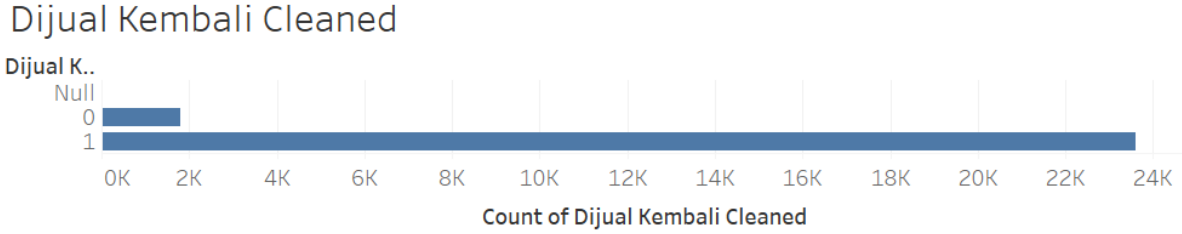


kaki_persegi

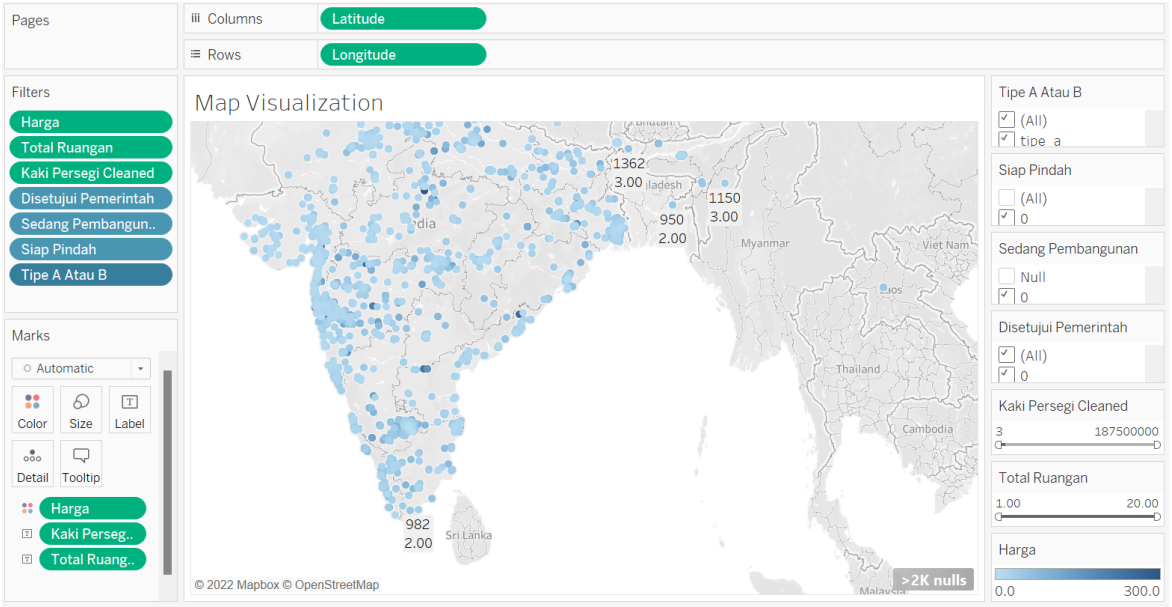


dijual_kembali





c. Map Data Visualization



Visualisasi Peta dilakukan berdasarkan fitur latitude dan longitude. Adapun telah diketahui bahwa latitude dan longitude tertukar, sehingga perlu ditukar *geographic role*. Penampilan titik rumah pada peta diberikan tiga label utama, yaitu harga, kaki persegi, serta total ruang. Adapun dibuat filter berdasarkan fitur yang ada, mencakup harga, total ruangan, kaki persegi, disetujui pemerintah, sedang pembangunan, siap pindah, dan tipe A atau B. Filters ini dibuat untuk mempermudah user mencari rumah secara lebih spesifik.

6. Kesimpulan

Tersedia data atas properti di India yang mayoritas berupa properti yang telah selesai dibangun namun belum mendapatkan perizinan resmi dan belum bisa dihuni, dengan hampir seluruh properti merupakan bangunan baru. Lebih dari setengah properti dimiliki oleh *dealer*, lalu diikuti oleh *owner*, dimana kepemilikan oleh *dealer* dan *owner* mencakup hampir seluruh properti. Properti yang disediakan banyak yang memenuhi persyaratan Tipe A dan Tipe B, dengan masing-masing berupa 1 kamar tidur, 1 ruang tengah / aula, dan 1 dapur, serta 1 ruangan dan 1 dapur. Adapula 10%+ bangunan yang tidak masuk persyaratan yang ada, antara memiliki ruangan kurang atau lebih dari persyaratan.

Berdasar data yang ada, dapat disimpulkan bahwa properti berfokus kepada kaum menengah dengan banyak tersedianya properti Tipe A dan Tipe B, dengan distribusi harga, luas, serta total ruangan yang berpusat di tengah agak ke kiri. Adapun properti untuk menengah bawah sangat-sangat terbatas. Properti bagi kalangan atas tersedia

cukup banyak dengan fasilitas dan harga yang cukup beragam. Sehingga dapat disimpulkan bahwa market utama dari data ini adalah kalangan menengah dengan market sekundernya adalah kalangan atas.