

Final Project Stage 1_Homework EDA - Group 5 (DS Batch 50)

- Veraldo Efraim
- Novisna Lintang Negari
- Alexander Panggabean
- Kevin William Markus Simbolon
- Adila

1. Descriptive Statistics (15 poin)

Gunakan function info dan describe pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

C. Apakah ada kolom yang memiliki nilai summary agak aneh?

(min/mean/median/max/unique/top/freq)

- Untuk masing-masing jenis observasi, tuliskan juga jika tidak ada masalah misal untuk A: "Semua tipe data sudah sesuai"

In []:

```
# Soal 1a
import pandas as pd

file_path = 'combined_all_data.csv'
data = pd.read_csv(file_path)

data.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51707 entries, 0 to 51706
Data columns (total 22 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   Unnamed: 0        51707 non-null  int64   
 1   realSum          51707 non-null  float64 
 2   room_type         51707 non-null  object  
 3   room_shared       51707 non-null  bool    
 4   room_private      51707 non-null  bool    
 5   person_capacity   51707 non-null  float64 
 6   host_is_superhost 51707 non-null  bool    
 7   multi             51707 non-null  int64   
 8   biz               51707 non-null  int64   
 9   cleanliness_rating 51707 non-null  float64 
 10  guest_satisfaction_overall 51707 non-null  float64 
 11  bedrooms          51707 non-null  int64   
 12  dist               51707 non-null  float64 
 13  metro_dist         51707 non-null  float64 
 14  attr_index         51707 non-null  float64 
 15  attr_index_norm    51707 non-null  float64 
 16  rest_index         51707 non-null  float64 
 17  rest_index_norm    51707 non-null  float64 
 18  lng                51707 non-null  float64 
 19  lat                51707 non-null  float64 
 20  City               51707 non-null  object  
 21  Day_Type           51707 non-null  object  
dtypes: bool(3), float64(12), int64(4), object(3)
memory usage: 7.6+ MB

```

Dari hasil data.info(), dataset memiliki 22 kolom dengan tipe data sebagai berikut:

1. Kolom dengan tipe data numerik (int64):

- Unnamed: 0
- multi
- biz
- bedrooms

2. Kolom dengan tipe data numerik (float64):

- realSum
- person_capacity
- cleanliness_rating
- guest_satisfaction_overall
- dist
- metro_dist
- attr_index
- attr_index_norm
- rest_index
- rest_index_norm

- lng
- lat

3. Kolom dengan tipe data boolean (bool):

- room_shared
- room_private
- host_is_superhost

4. Kolom dengan tipe data kategorikal (object):

- room_type
- City
- Day_Type

Kesimpulan:

Semua tipe data sudah sesuai dengan konteks masing-masing kolom. **Tidak ada masalah dengan tipe data.**

```
In [ ]: # Soal 1b
missing_values = data.isnull().sum()
print("Jumlah nilai kosong per kolom:")
print(missing_values)
```

```
Jumlah nilai kosong per kolom:
Unnamed: 0          0
realSum            0
room_type          0
room_shared        0
room_private       0
person_capacity   0
host_is_superhost 0
multi              0
biz                0
cleanliness_rating 0
guest_satisfaction_overall 0
bedrooms           0
dist               0
metro_dist         0
attr_index         0
attr_index_norm   0
rest_index         0
rest_index_norm   0
lng                0
lat                0
City               0
Day_Type           0
dtype: int64
```

Berdasarkan hasil dari data.info() dan pemeriksaan menggunakan isnull().sum():

- Semua kolom memiliki jumlah baris 51707, yang sama dengan total baris dataset.
- Tidak ada nilai kosong di kolom manapun.

Kesimpulan:

■ Semua kolom lengkap dan **tidak ada** nilai kosong dalam dataset.

In []:

```
# Soal 1c
data_description = data.describe(include='all').transpose()
print("Statistik deskriptif:")
print(data_description)
```

Statistik deskriptif:

	count	unique	top	freq	\
Unnamed: 0	51707.0	NaN	NaN	NaN	
realSum	51707.0	NaN	NaN	NaN	
room_type	51707	3	Entire home/apt	32648	
room_shared	51707	2	False	51341	
room_private	51707	2	False	33014	
person_capacity	51707.0	NaN	NaN	NaN	
host_is_superhost	51707	2	False	38475	
multi	51707.0	NaN	NaN	NaN	
biz	51707.0	NaN	NaN	NaN	
cleanliness_rating	51707.0	NaN	NaN	NaN	
guest_satisfaction_overall	51707.0	NaN	NaN	NaN	
bedrooms	51707.0	NaN	NaN	NaN	
dist	51707.0	NaN	NaN	NaN	
metro_dist	51707.0	NaN	NaN	NaN	
attr_index	51707.0	NaN	NaN	NaN	
attr_index_norm	51707.0	NaN	NaN	NaN	
rest_index	51707.0	NaN	NaN	NaN	
rest_index_norm	51707.0	NaN	NaN	NaN	
lng	51707.0	NaN	NaN	NaN	
lat	51707.0	NaN	NaN	NaN	
City	51707	10	London	9993	
Day_Type	51707	2	Weekend	26207	
	mean	std	min	25%	\
Unnamed: 0	1620.502388	1217.380366	0.0	646.0	
realSum	279.879591	327.948386	34.779339	148.752174	
room_type	NaN	NaN	NaN	NaN	
room_shared	NaN	NaN	NaN	NaN	
room_private	NaN	NaN	NaN	NaN	
person_capacity	3.161661	1.298545	2.0	2.0	
host_is_superhost	NaN	NaN	NaN	NaN	
multi	0.291353	0.45439	0.0	0.0	
biz	0.350204	0.477038	0.0	0.0	
cleanliness_rating	9.390624	0.954868	2.0	9.0	
guest_satisfaction_overall	92.628232	8.945531	20.0	90.0	
bedrooms	1.15876	0.62741	0.0	1.0	
dist	3.191285	2.393803	0.015045	1.453142	
metro_dist	0.68154	0.858023	0.002301	0.24848	
attr_index	294.204105	224.754123	15.152201	136.797385	
attr_index_norm	13.423792	9.807985	0.926301	6.380926	
rest_index	626.856696	497.920226	19.576924	250.854114	
rest_index_norm	22.786177	17.804096	0.592757	8.75148	
lng	7.426068	9.799725	-9.22634	-0.0725	
lat	45.671128	5.249263	37.953	41.39951	
City	NaN	NaN	NaN	NaN	
Day_Type	NaN	NaN	NaN	NaN	
	50%	75%	max		
Unnamed: 0	1334.0	2382.0	5378.0		
realSum	211.343089	319.694287	18545.450285		
room_type	NaN	NaN	NaN		
room_shared	NaN	NaN	NaN		
room_private	NaN	NaN	NaN		
person_capacity	3.0	4.0	6.0		

host_is_superhost	NaN	NaN	NaN
multi	0.0	1.0	1.0
biz	0.0	1.0	1.0
cleanliness_rating	10.0	10.0	10.0
guest_satisfaction_overall	95.0	99.0	100.0
bedrooms	1.0	1.0	10.0
dist	2.613538	4.263077	25.284557
metro_dist	0.413269	0.73784	14.273577
attr_index	234.331748	385.756381	4513.563486
attr_index_norm	11.468305	17.415082	100.0
rest_index	522.052783	832.628988	6696.156772
rest_index_norm	17.542238	32.964603	100.0
lng	4.873	13.518825	23.78602
lat	47.50669	51.471885	52.64141
City	NaN	NaN	NaN
Day_Type	NaN	NaN	NaN

Hasil dari fungsi describe() menunjukkan beberapa kolom memiliki nilai yang mencurigakan atau outlier:

1. Kolom realSum (Harga Sewa):

- **Mean:** 279.88
- **Std Dev:** 327.95
- **Min:** 34.78
- **Q3 (75%):** 319.69
- **Max:** 18,545.45
- **Analisis:** Nilai maksimum jauh di atas kuartil ketiga ($Q3 = 319.69$) dan rata-rata (279.88). Ini menunjukkan adanya outlier, yaitu harga sewa yang sangat tinggi.

2. bedrooms (Jumlah Kamar):

- **Mean:** 1.16
- **Std Dev:** 0.63
- **Min:** 0
- **Max:** 10
- **Analisis:** Nilai 0 kamar tidak wajar untuk sebuah penginapan. Nilai maksimum 10 kamar juga terlihat sangat tinggi dibandingkan rata-rata (1.16), yang dapat dianggap outlier.

3. dist (Jarak ke pusat):

- **Mean:** 3.19
- **Std Dev:** 2.39
- **Min:** 0.015
- **Max:** 25.28
- **Analisis:** Nilai minimum sangat kecil (0.015), yang perlu diverifikasi apakah relevan. Nilai maksimum (25.28) jauh di atas Q3 (4.26), menunjukkan kemungkinan outlier.

4. attr_index (Indeks Daya Tarik):

- **Mean:** 294.20
- **Std Dev:** 224.75
- **Min:** 15.15
- **Max:** 4,513.56
- **Analisis:** Nilai maksimum jauh di atas Q3 (385.76), menunjukkan adanya outlier.

5. rest_index (Indeks Restoran):

- **Mean:** 626.86
- **Std Dev:** 497.92
- **Min:** 19.58
- **Max:** 6,696.16
- **Analisis:** Nilai maksimum jauh di atas Q3 (832.63), menunjukkan adanya outlier.

Kesimpulan

Kolom yang memiliki nilai mencurigakan atau outlier berdasarkan analisis:

- **realSum:** Harga sewa sangat tinggi (maksimum 18,545.45).
- **bedrooms:** Nilai 0 kamar tidak wajar, dan nilai 10 kamar terlihat ekstrem.
- **dist:** Jarak yang sangat kecil (0.015) atau sangat besar (25.28) perlu diverifikasi.
- **attr_index:** Indeks daya tarik dengan nilai maksimum 4,513.56 terlihat sangat besar.
- **rest_index:** Indeks restoran dengan nilai maksimum 6,696.16 jauh di atas rata-rata.

2. Univariate Analysis (25 poin)

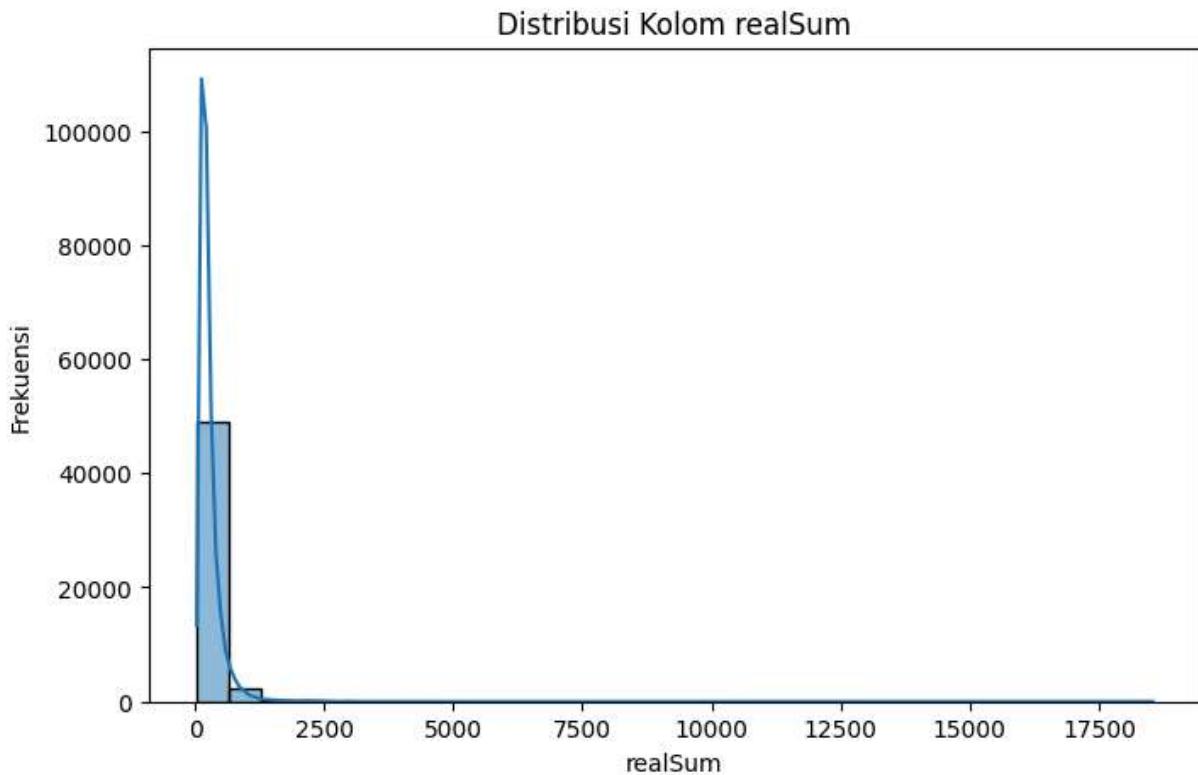
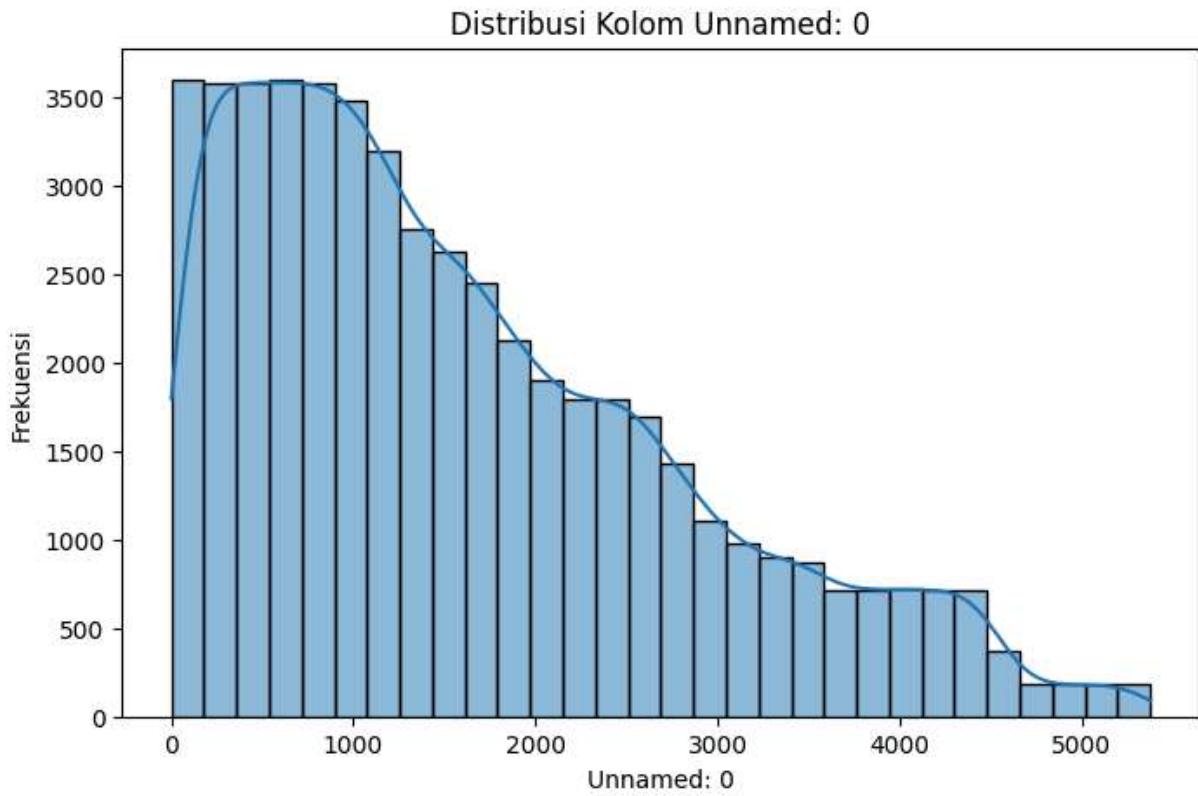
Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

```
In [ ]: import matplotlib.pyplot as plt
import seaborn as sns

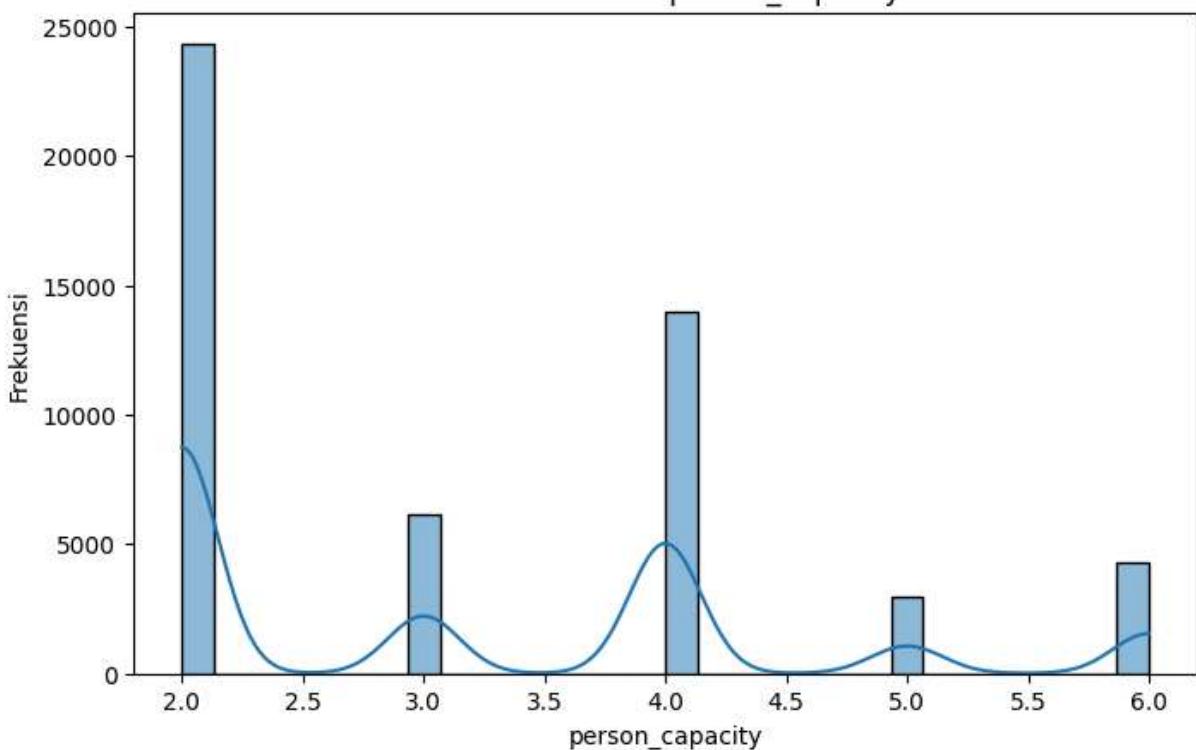
numerical_cols = data.select_dtypes(include=['float64', 'int64']).columns
for col in numerical_cols:
    plt.figure(figsize=(8, 5))
    sns.histplot(data[col], kde=True, bins=30)
    plt.title(f'Distribusi Kolom {col}')
    plt.xlabel(col)
    plt.ylabel('Frekuensi')
    plt.show()

categorical_cols = data.select_dtypes(include=['object', 'bool']).columns
for col in categorical_cols:
    plt.figure(figsize=(8, 5))
    sns.countplot(data=data, x=col, order=data[col].value_counts().index)
```

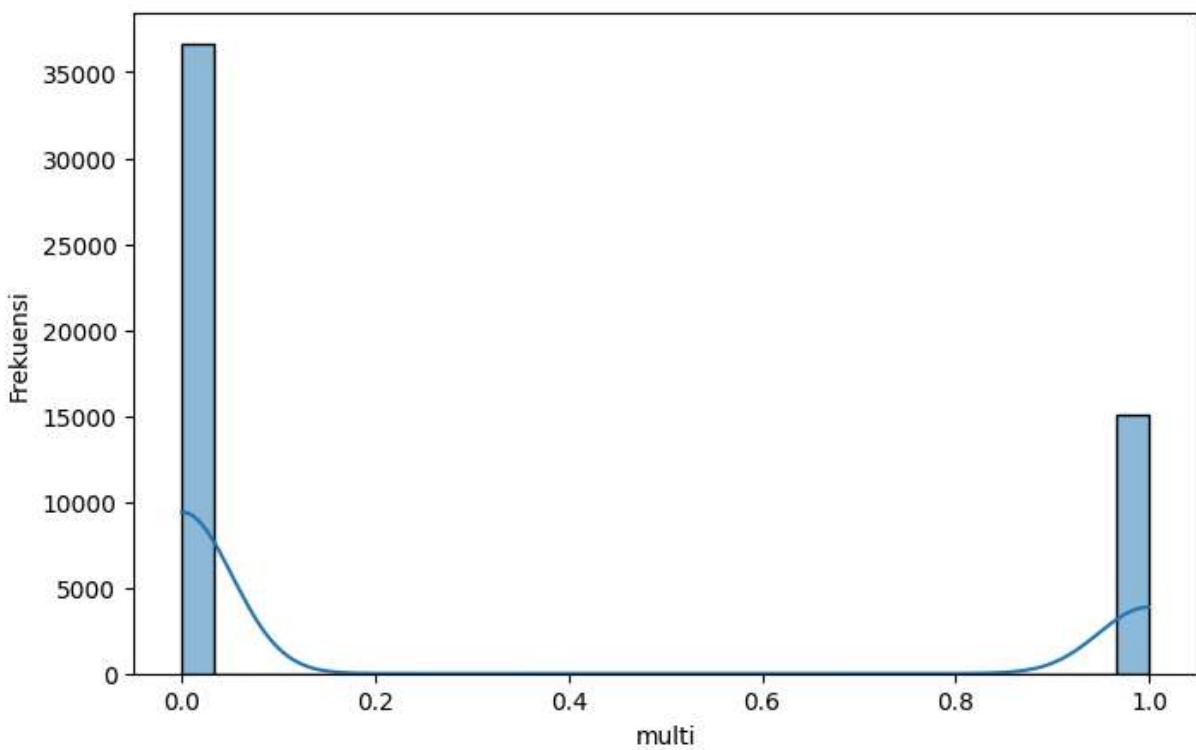
```
plt.title(f'Distribusi Kolom {col}')
plt.xlabel(col)
plt.ylabel('Frekuensi')
plt.xticks(rotation=45)
plt.show()
```



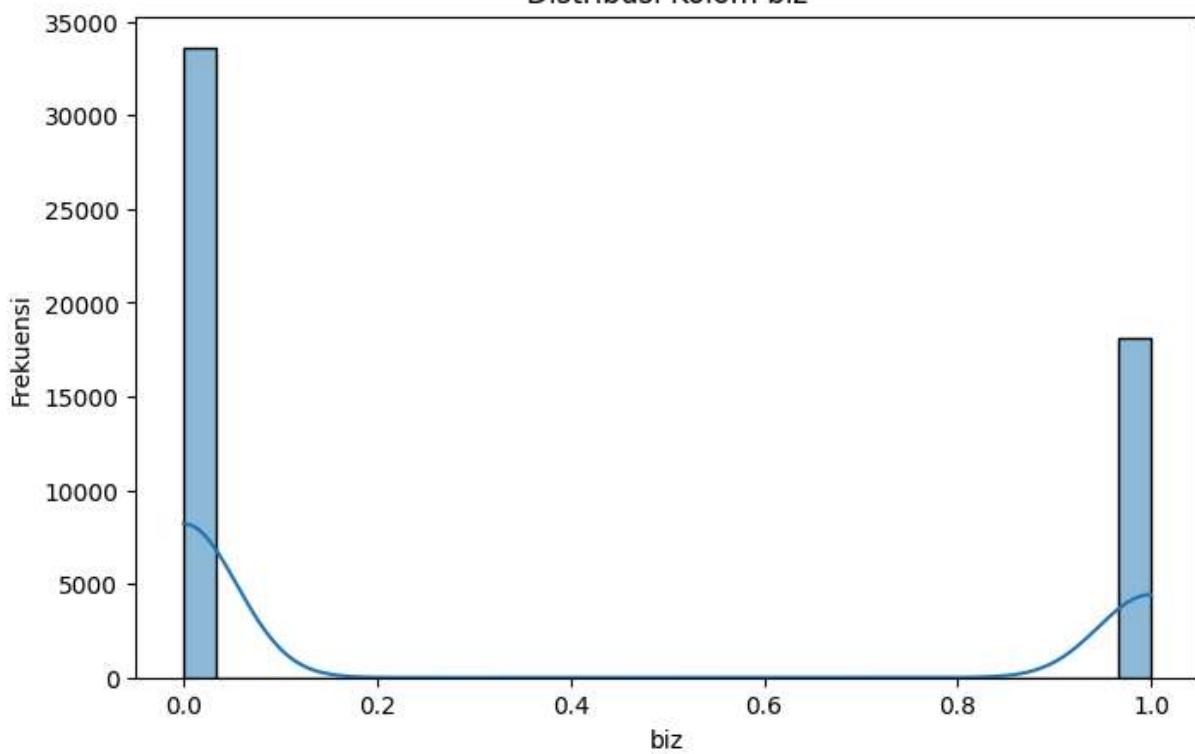
Distribusi Kolom person_capacity



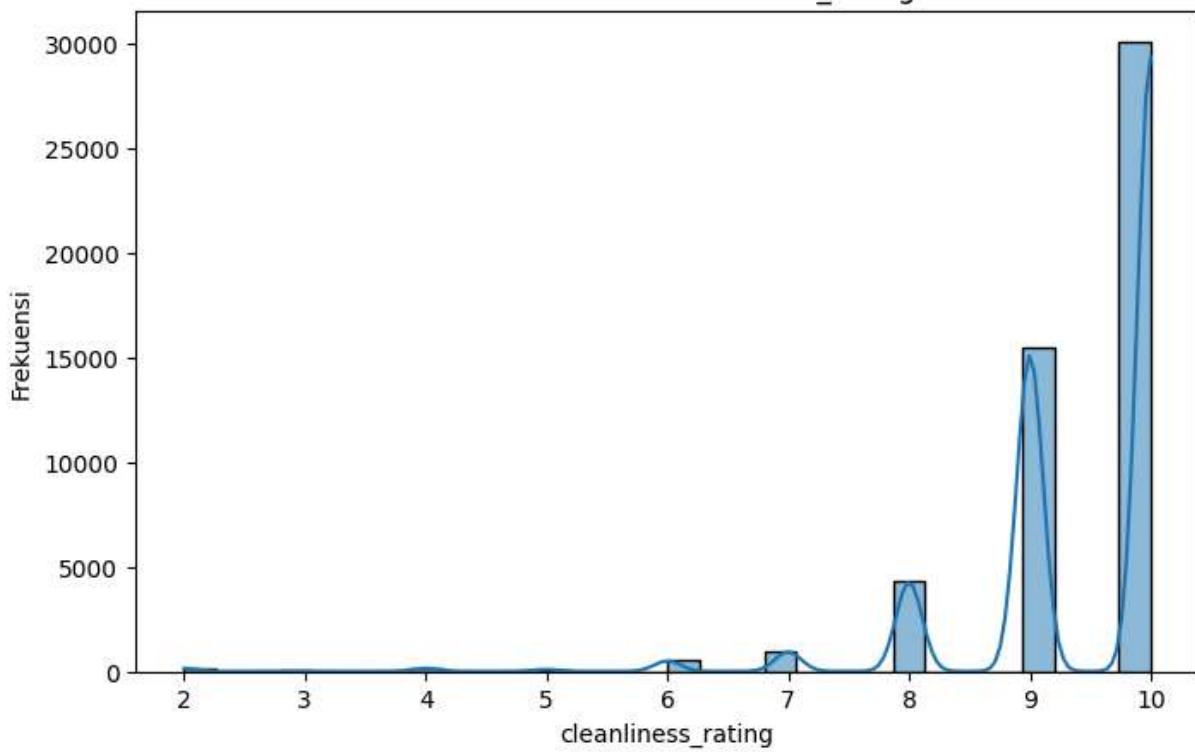
Distribusi Kolom multi



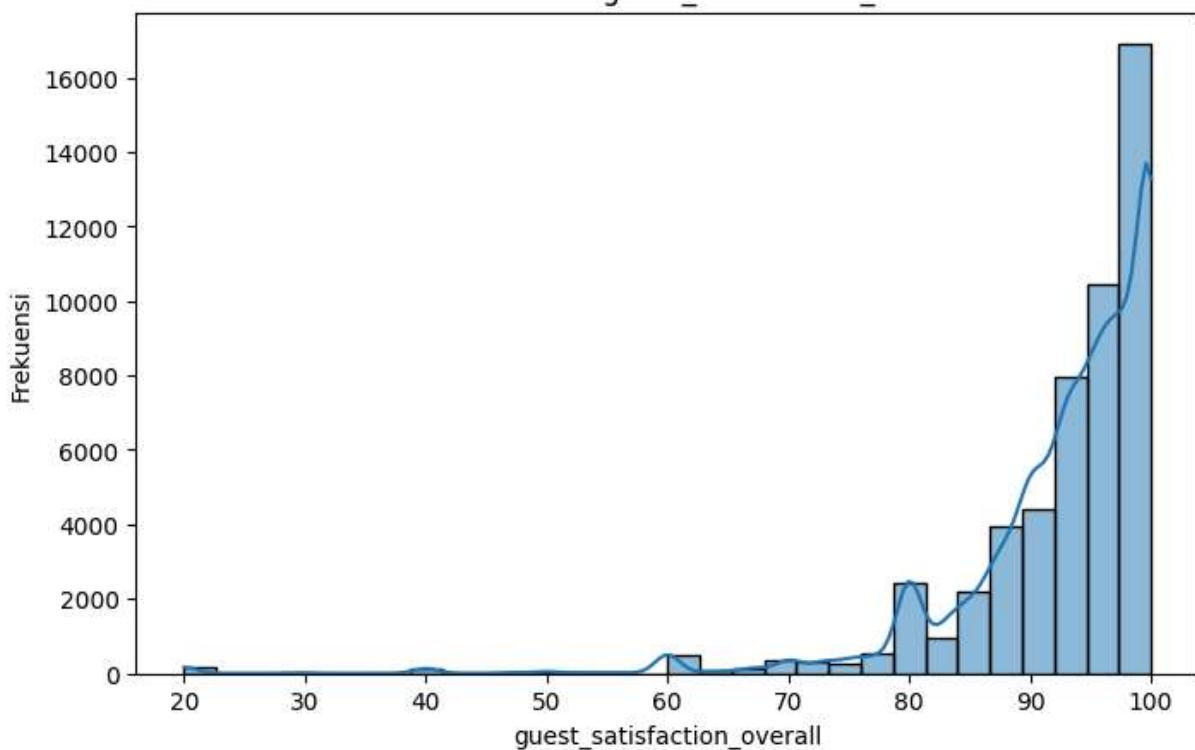
Distribusi Kolom biz



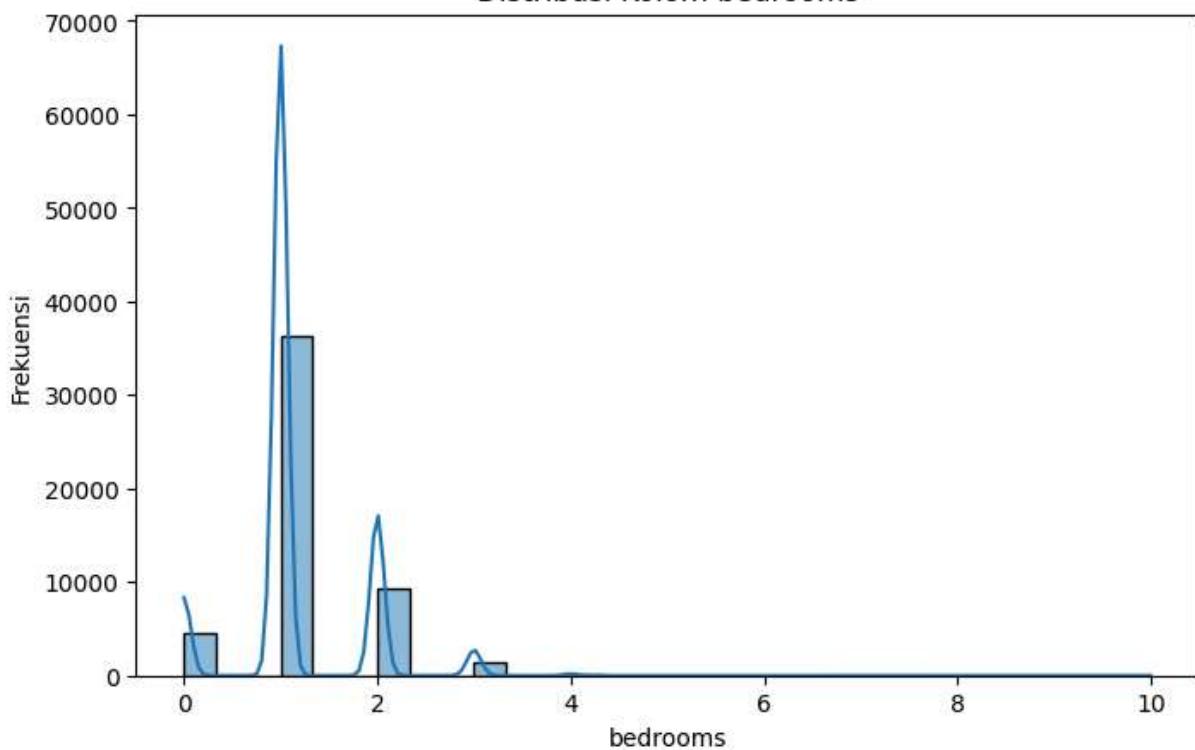
Distribusi Kolom cleanliness_rating



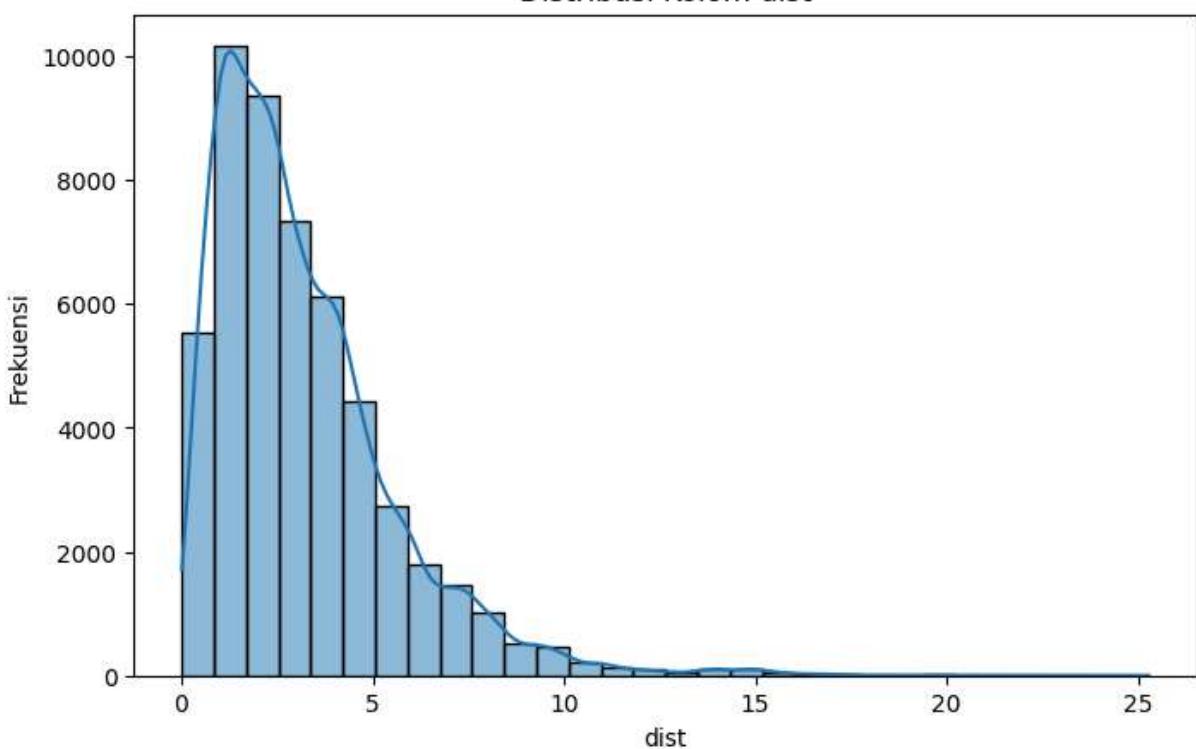
Distribusi Kolom guest_satisfaction_overall



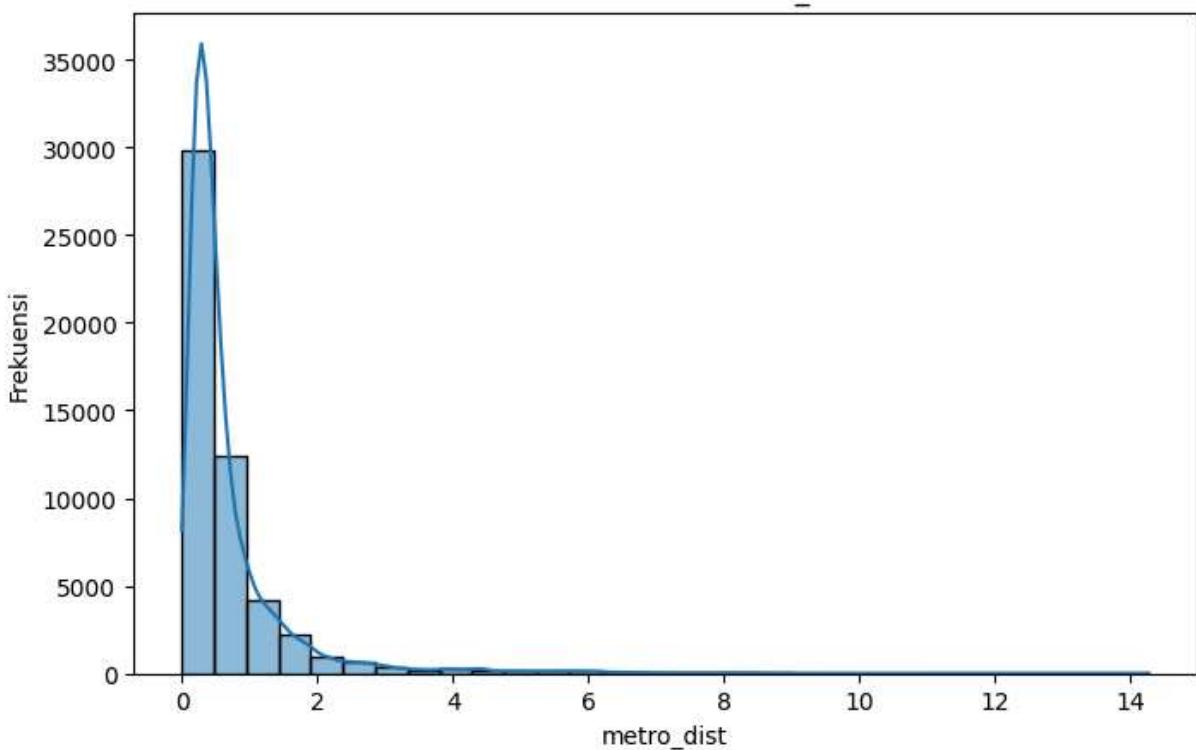
Distribusi Kolom bedrooms



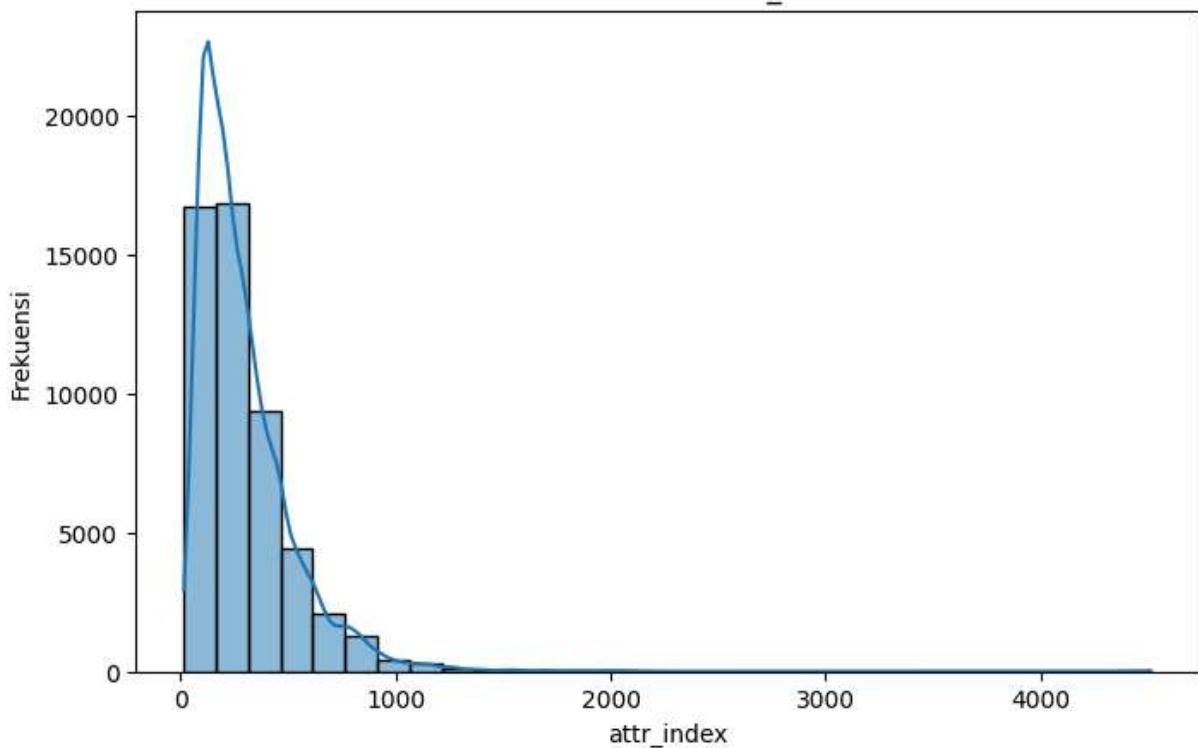
Distribusi Kolom dist



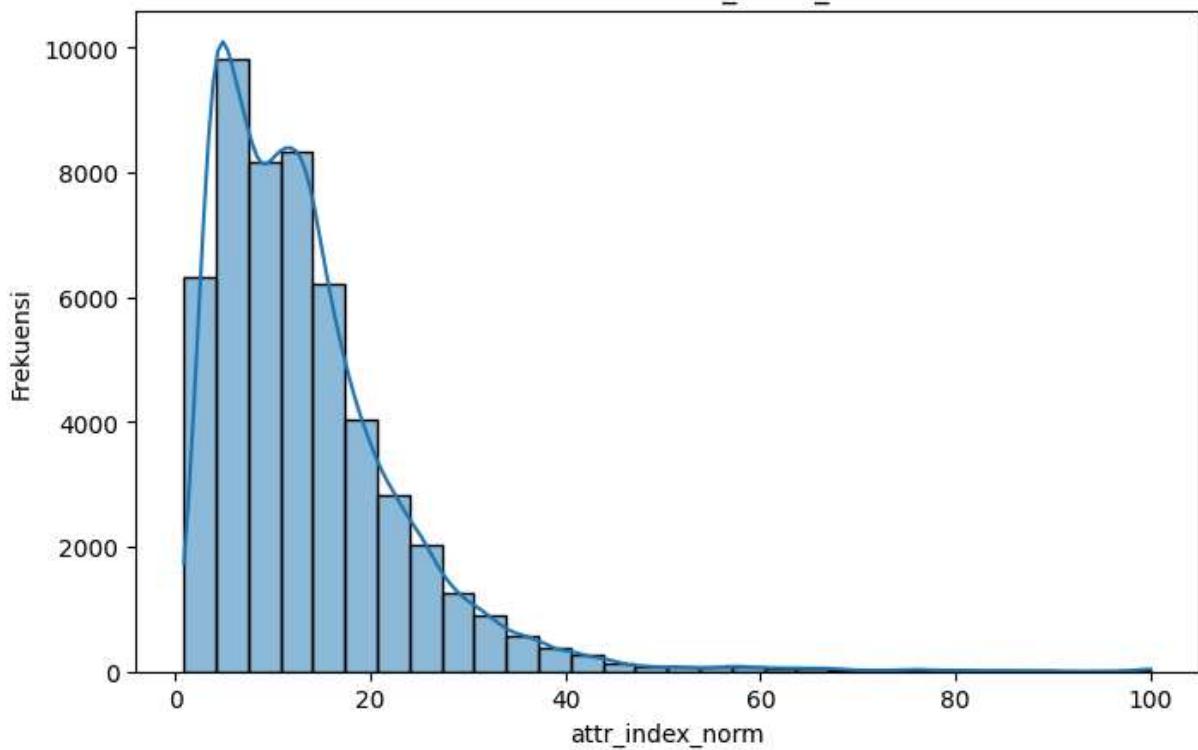
Distribusi Kolom metro_dist



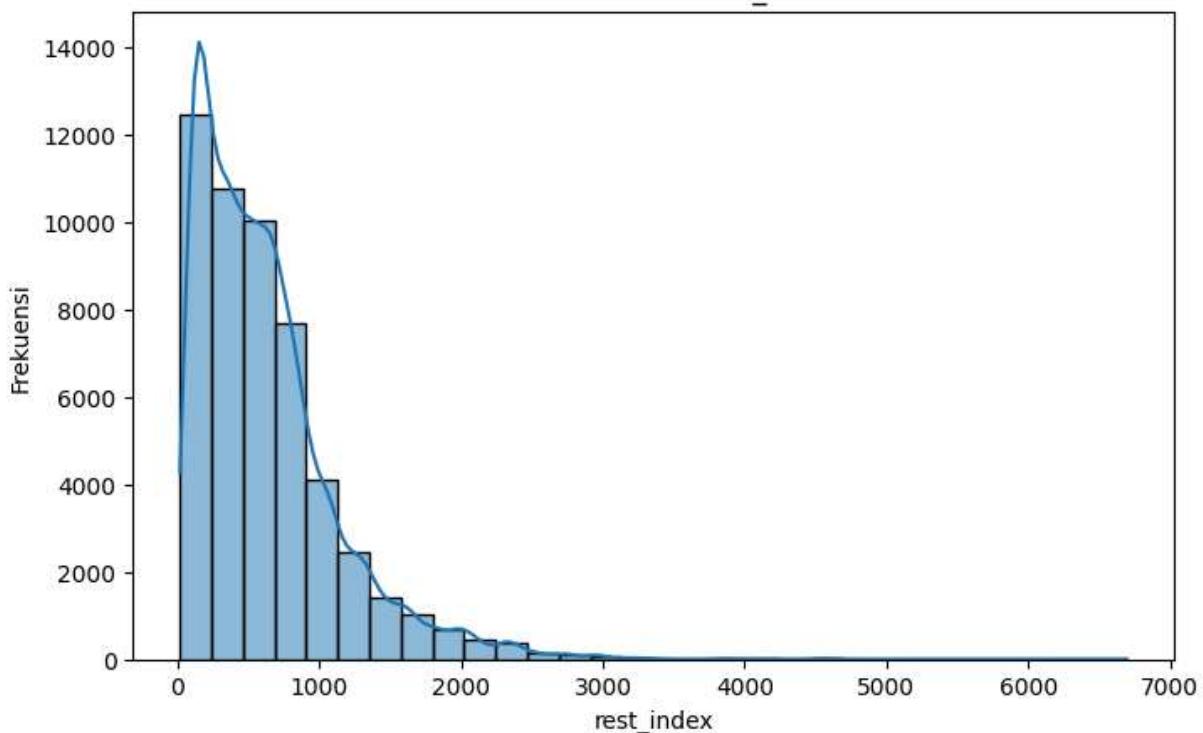
Distribusi Kolom attr_index



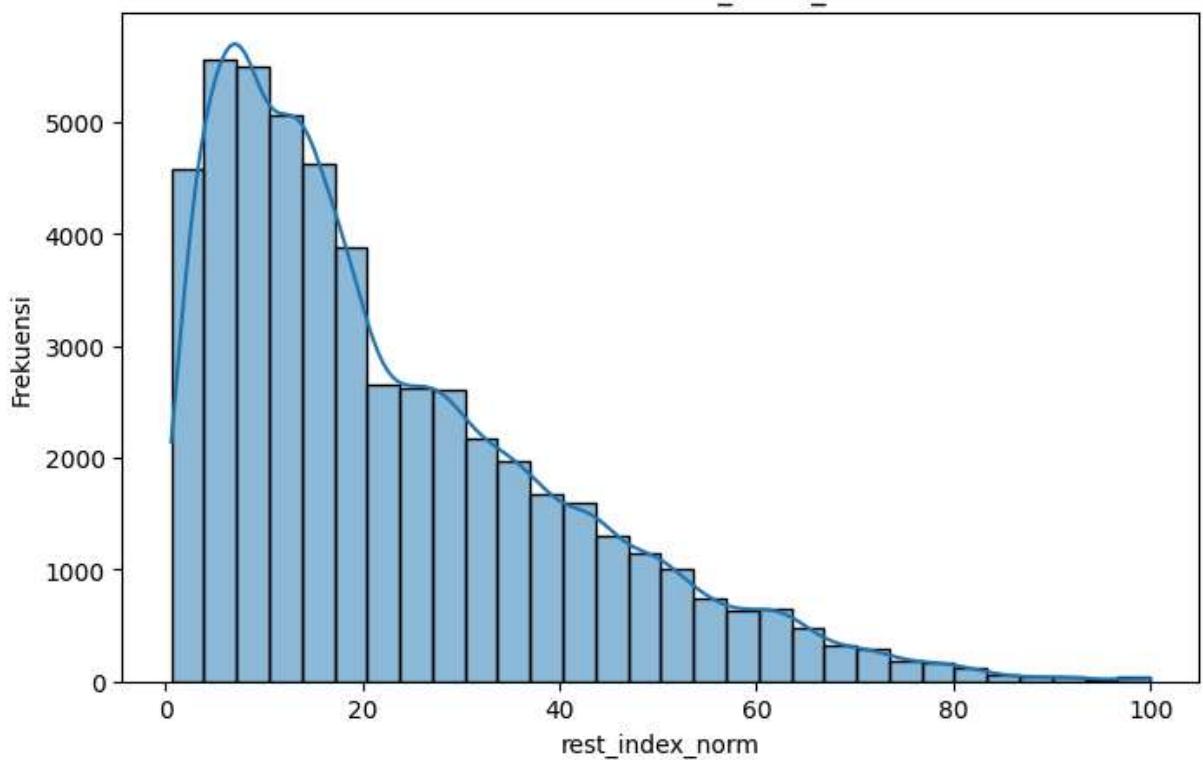
Distribusi Kolom attr_index_norm



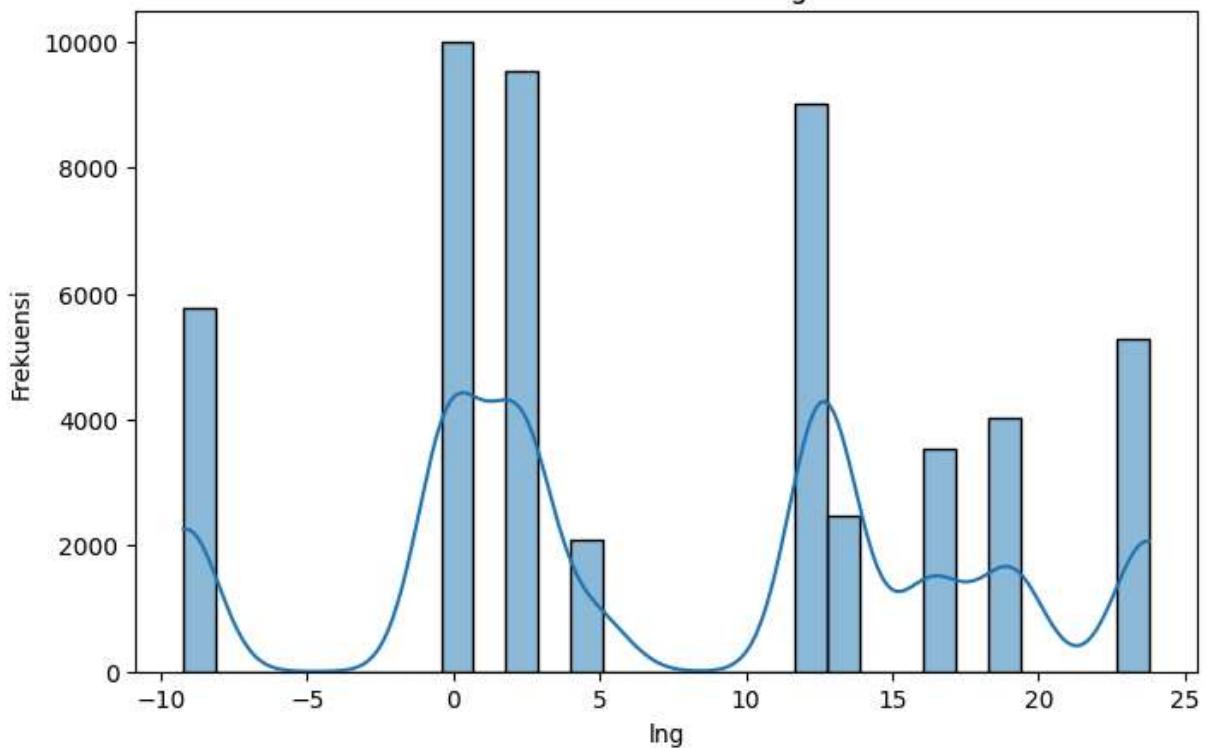
Distribusi Kolom rest_index



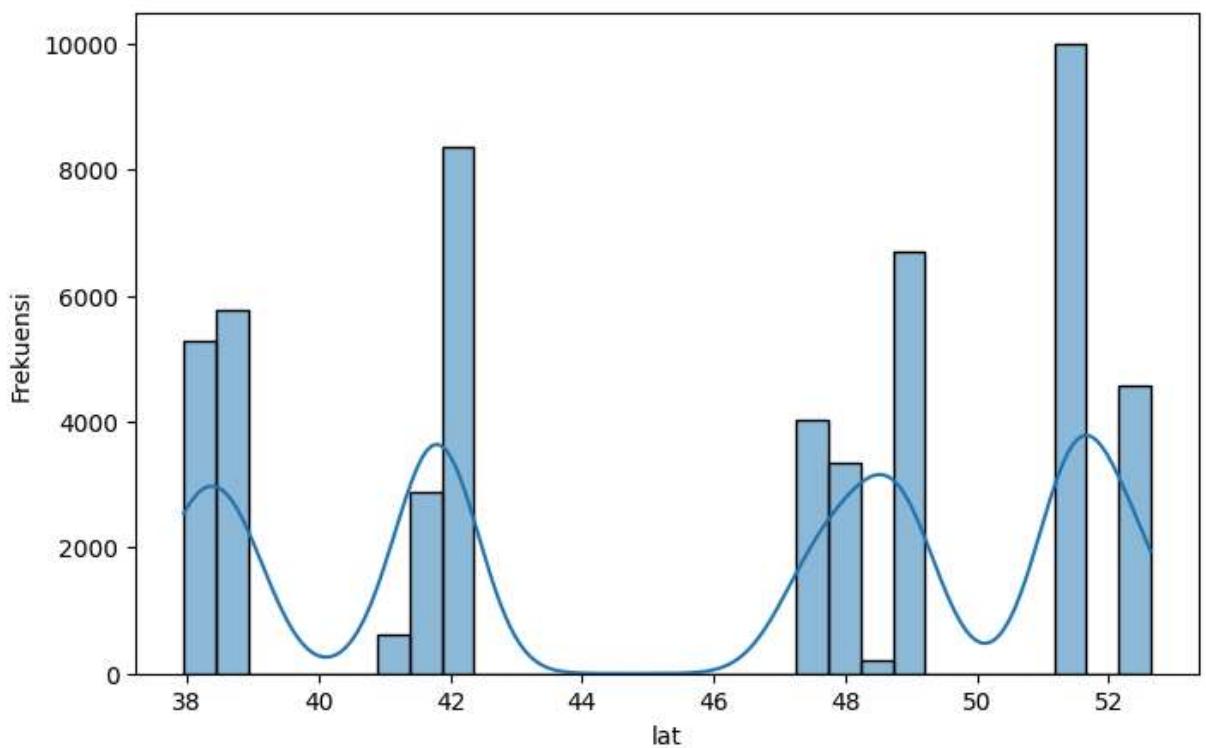
Distribusi Kolom rest_index_norm



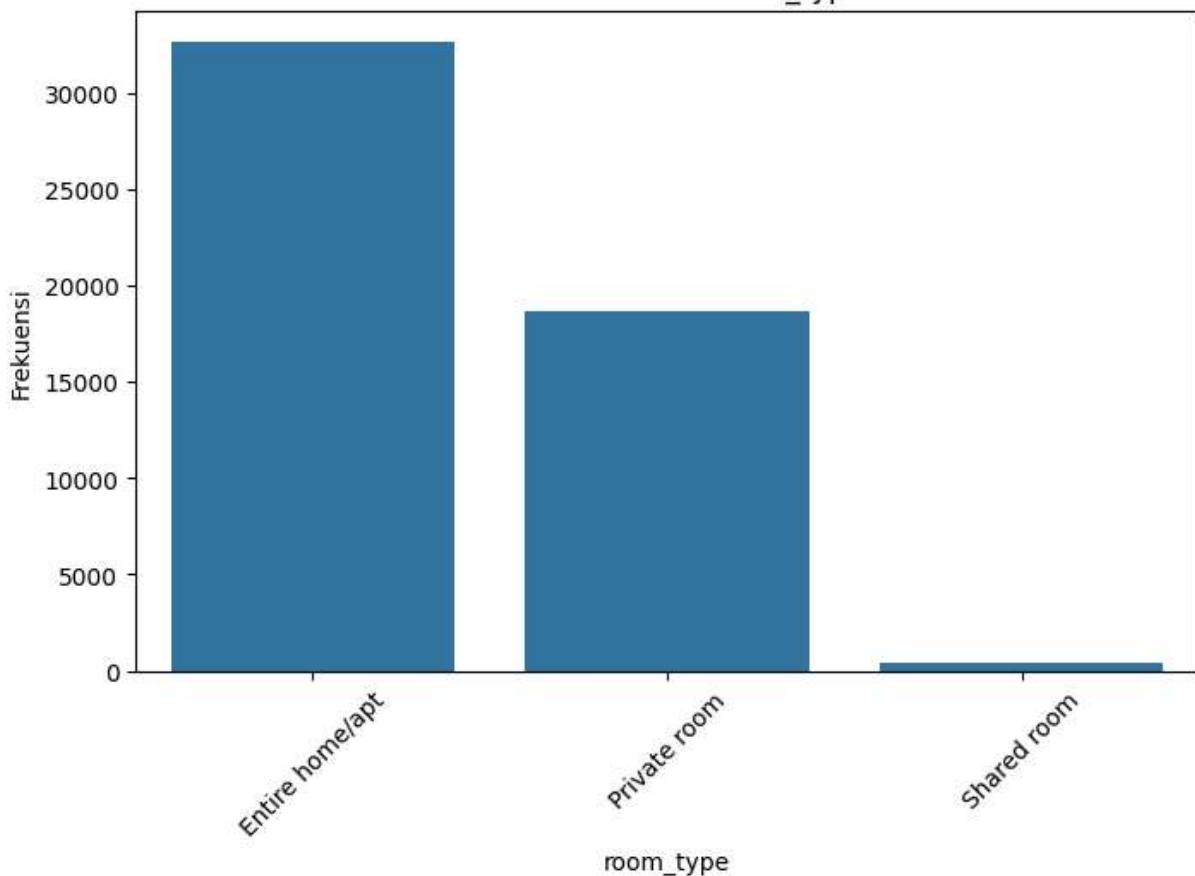
Distribusi Kolom Ing



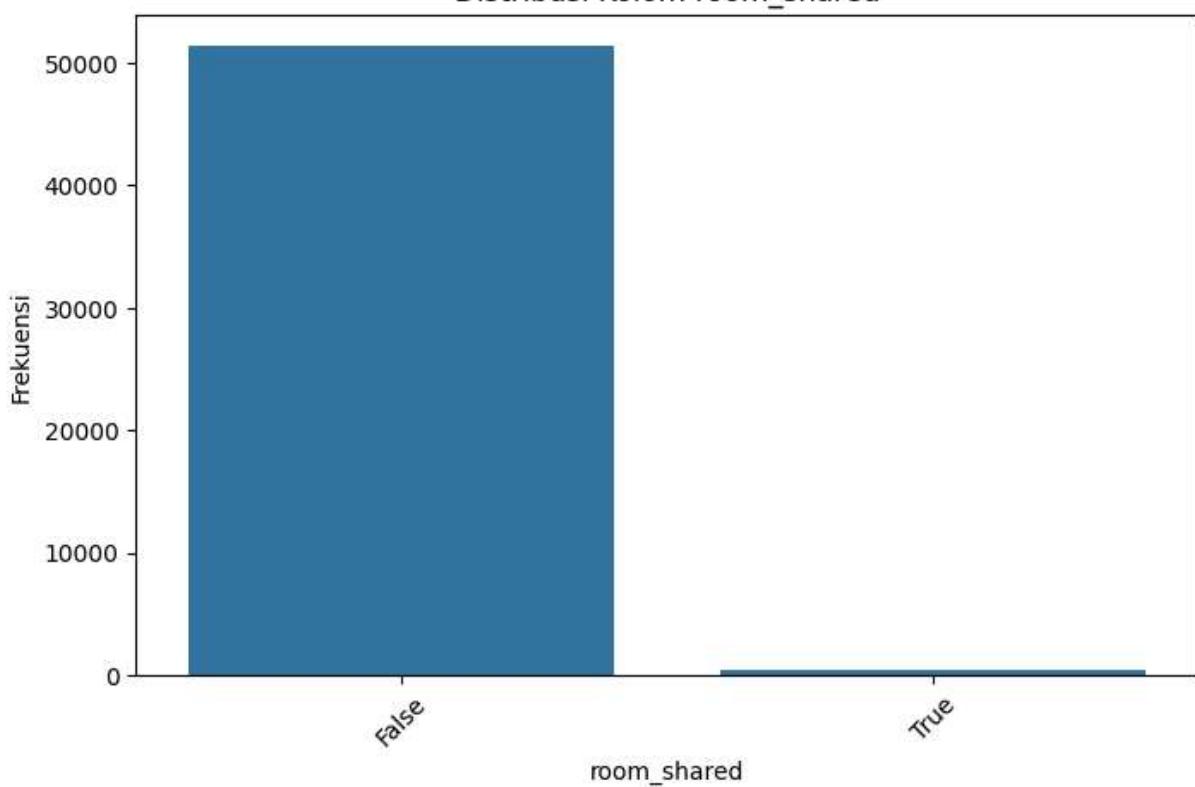
Distribusi Kolom lat



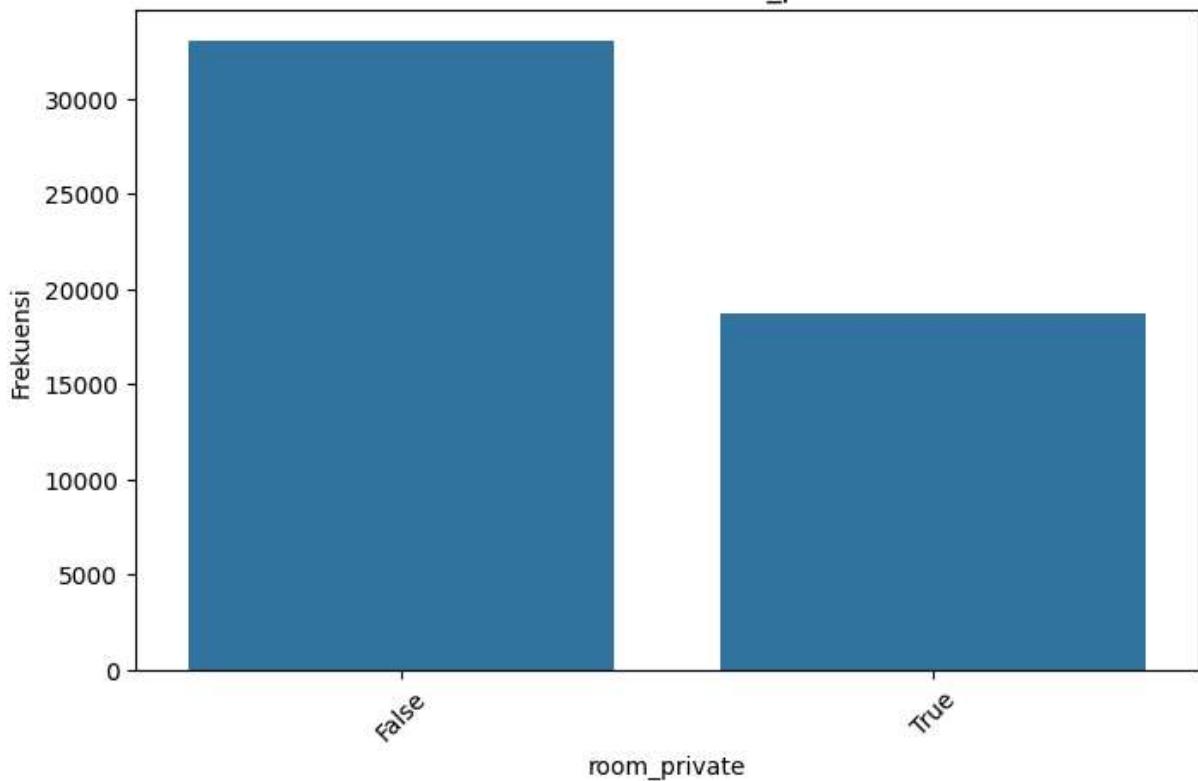
Distribusi Kolom room_type



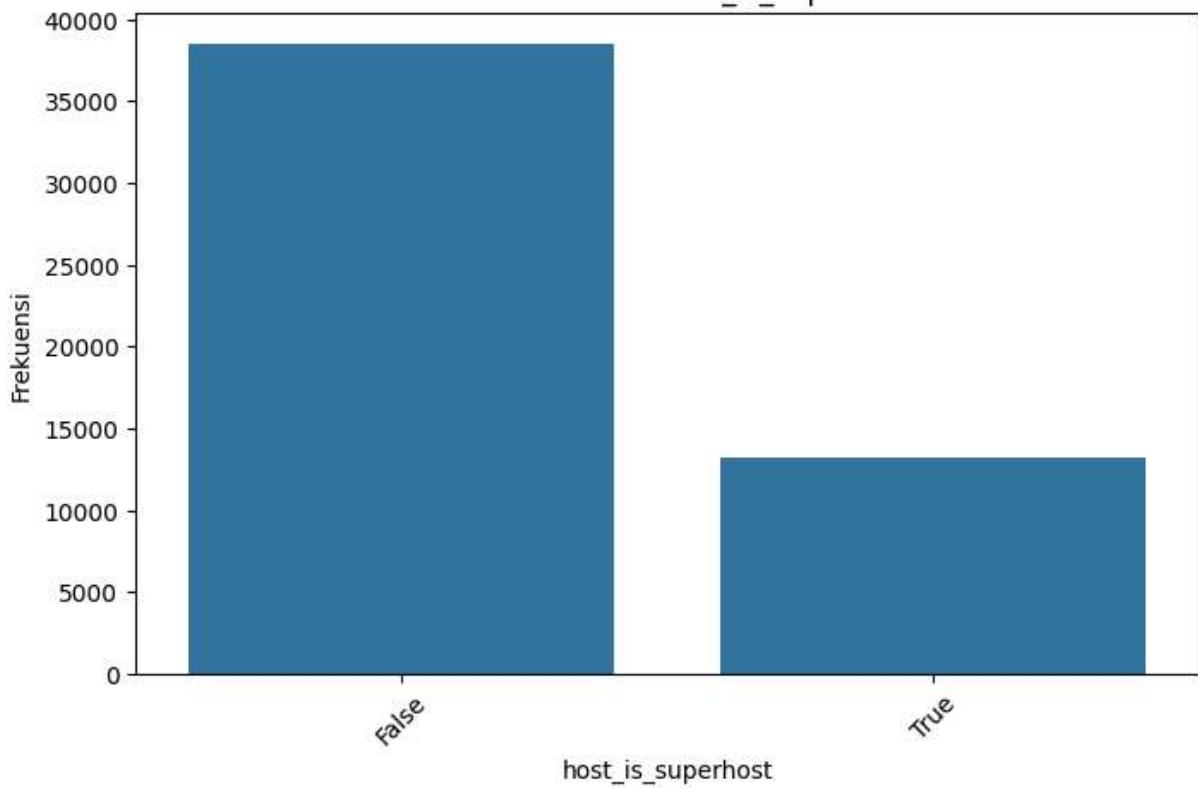
Distribusi Kolom room_shared

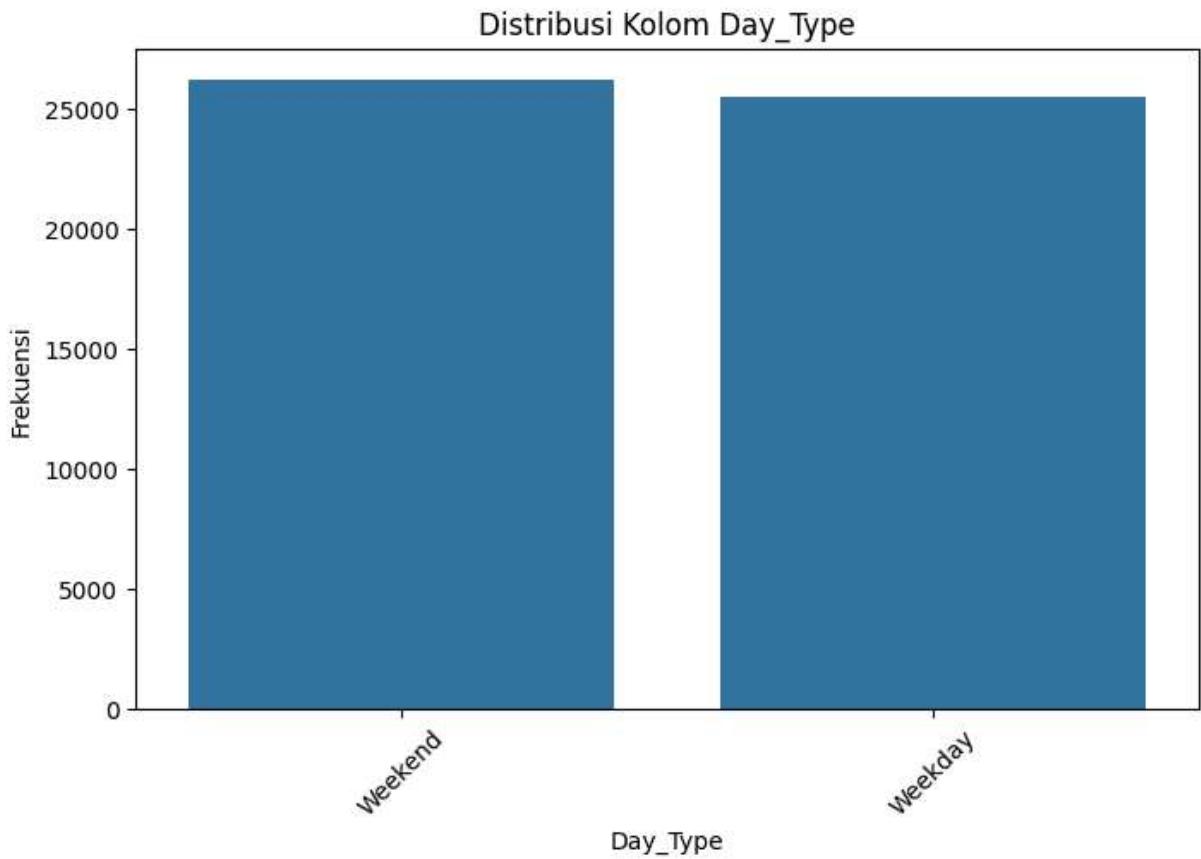
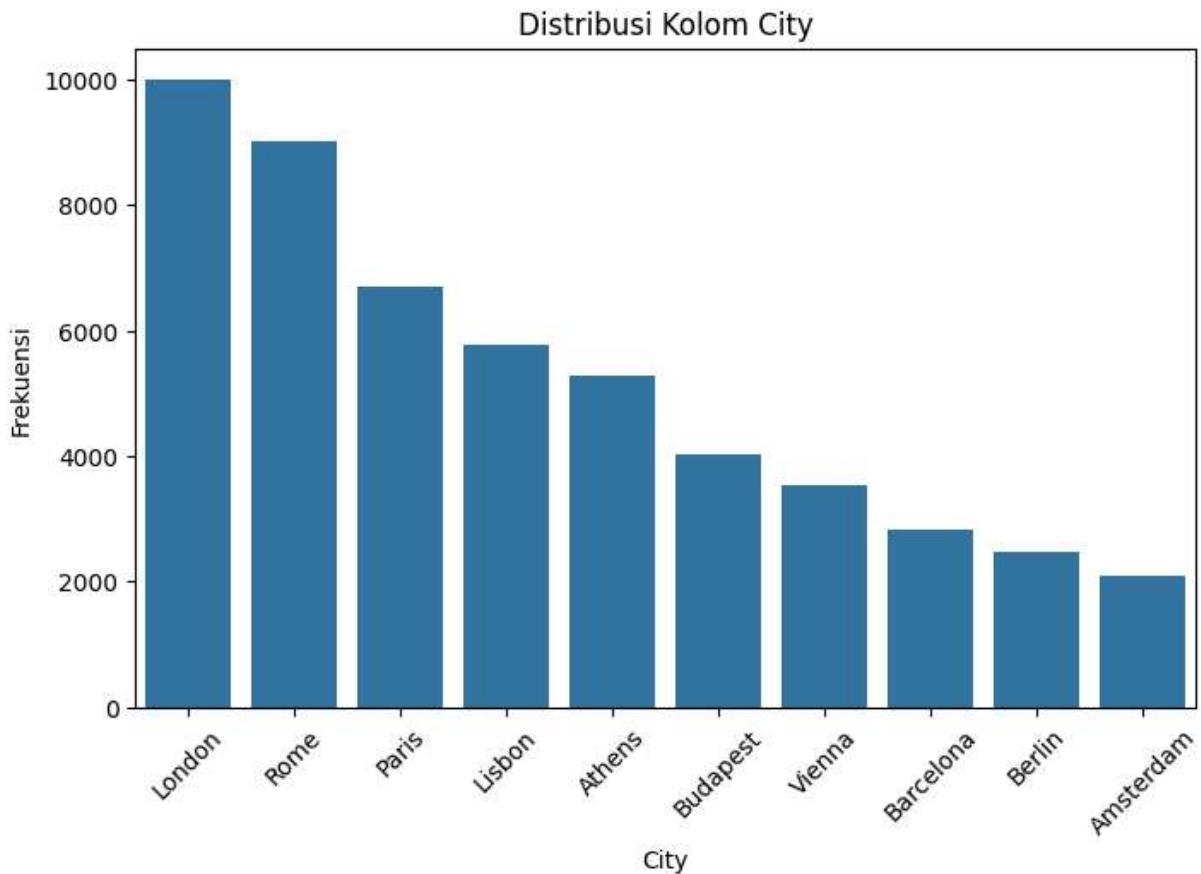


Distribusi Kolom room_private



Distribusi Kolom host_is_superhost





A. Analisis Kolom Numerik

1. realSum (Harga Sewa)

- **Distribusi:** Kolom ini menggambarkan harga sewa properti dalam dataset, di mana sebagian besar nilai terkonsentrasi pada rentang harga yang rendah, sekitar 200-500, dan distribusinya sangat miring ke kanan (right-skewed). Terdapat nilai outlier yang sangat ekstrem, dengan harga tertinggi mencapai lebih dari 18,500.
- **Observasi Menarik:** Distribusi harga yang miring ini menunjukkan bahwa sebagian besar properti memiliki harga yang terjangkau, tetapi ada beberapa properti premium dengan harga sangat tinggi. Nilai ekstrem ini bisa jadi disebabkan oleh properti dengan fasilitas mewah atau kesalahan input data.
- **Follow-up:** Sebagai langkah pre-processing, disarankan untuk melakukan transformasi log agar distribusi lebih normal sehingga analisis statistik atau model prediktif menjadi lebih akurat. Selain itu, outlier harus divalidasi untuk memastikan bahwa nilai-nilai ini benar-benar mencerminkan data aktual dan bukan kesalahan input.

2. dist (Jarak ke Pusat)

- **Distribusi:** Kolom ini menunjukkan jarak properti ke pusat kota. Sebagian besar properti memiliki jarak kurang dari 5 km, tetapi terdapat beberapa nilai yang sangat jauh, hingga mencapai lebih dari 25 km, sehingga distribusinya juga miring ke kanan.
- **Observasi Menarik:** Properti yang terletak sangat jauh dari pusat kota kemungkinan mencerminkan lokasi di area suburban atau rural, yang mungkin memiliki target pasar berbeda. Namun, nilai-nilai ekstrem perlu divalidasi untuk memastikan tidak ada kesalahan dalam data.
- **Follow-up:** Transformasi log juga dapat diterapkan pada kolom ini untuk mendistribusikan data secara lebih merata dan mengurangi pengaruh outlier terhadap analisis. Selain itu, properti dengan jarak ekstrem harus diperiksa apakah benar-benar relevan dengan lokasi geografis atau merupakan kesalahan input.

3. metro_dist (Jarak ke Stasiun Metro)

- **Distribusi:** Distribusi kolom ini mirip dengan kolom dist, di mana sebagian besar nilai berada dalam jarak yang kecil, kurang dari 2 km, tetapi ada beberapa nilai yang jauh lebih tinggi, hingga lebih dari 14 km.
- **Observasi Menarik:** Jarak ke stasiun metro yang sangat jauh kemungkinan mencerminkan properti yang berada di daerah terpencil atau tidak terjangkau oleh transportasi umum. Properti semacam ini mungkin memiliki karakteristik yang berbeda dari properti lain dalam dataset.
- **Follow-up:** Transformasi log sangat dianjurkan untuk mendistribusikan data lebih merata, dan properti dengan jarak ekstrem perlu divalidasi untuk memastikan tidak ada kesalahan dalam data.

4. attr_index dan rest_index

- **Distribusi:** Kedua kolom ini mencerminkan skor akumulasi untuk atribut tertentu, seperti daya tarik dan fasilitas restoran di sekitar properti. Distribusinya sangat miring ke kanan, dengan sebagian besar nilai berada pada rentang yang rendah (<1000), tetapi terdapat beberapa properti dengan skor yang sangat tinggi, melebihi 4000.
- **Observasi Menarik:** Properti dengan skor yang sangat tinggi kemungkinan berada di lokasi premium dengan banyak fasilitas atau daya tarik wisata. Namun, nilai-nilai ekstrem ini perlu diverifikasi untuk memastikan konsistensinya.
- **Follow-up:** Transformasi log direkomendasikan untuk membuat data lebih merata, sehingga analisis statistik dan prediktif menjadi lebih akurat. Validasi terhadap nilai ekstrem juga diperlukan untuk memastikan relevansi data.

5. attr_index_norm dan rest_index_norm

- **Distribusi:** Kedua kolom ini merupakan versi normalisasi dari attr_index dan rest_index. Nilainya sudah distandardisasi dalam rentang 0-100, sehingga distribusi data lebih merata dibandingkan versi aslinya.
- **Observasi Menarik:** Distribusi yang lebih merata ini menunjukkan bahwa normalisasi telah dilakukan dengan baik. Tidak ada outlier yang terlihat pada kolom ini.
- **Follow-up:** Tidak ada tindakan tambahan yang diperlukan karena data sudah siap untuk analisis lebih lanjut.

6. person_capacity

- **Distribusi:** Kolom ini menunjukkan kapasitas maksimum properti, dan distribusinya memiliki pola bimodal dengan dua puncak utama pada kapasitas 2 orang dan 4 orang. Sebagian besar properti dalam dataset dirancang untuk kapasitas kecil hingga sedang.
- **Observasi Menarik:** Pola ini mencerminkan segmen pasar properti yang didominasi oleh pasangan atau keluarga kecil.
- **Follow-up:** Tidak ada tindakan khusus yang diperlukan, tetapi kolom ini dapat digunakan untuk segmentasi berdasarkan ukuran properti.

7. bedrooms (Jumlah Kamar)

- **Distribusi:** Sebagian besar properti memiliki 1-2 kamar tidur, tetapi terdapat beberapa nilai ekstrem, seperti 0 kamar yang dapat mengindikasikan studio, dan 10 kamar yang mungkin mencerminkan properti dengan kapasitas besar atau villa.
- **Observasi Menarik:** Nilai ekstrem seperti 0 kamar dan 10 kamar harus diperiksa lebih lanjut untuk memastikan relevansinya.
- **Follow-up:** Validasi nilai ekstrem diperlukan untuk menentukan apakah data ini mencerminkan kondisi aktual atau kesalahan input. Nilai ekstrem juga dapat diatasi dengan teknik seperti winsorizing jika ditemukan tidak relevan.

8. Ing dan lat (Longitude dan Latitude)

- Distribusi: Observasi Menarik: Nilai ekstrem seperti 0 kamar dan 10 kamar harus diperiksa lebih lanjut untuk memastikan relevansinya.
- Observasi Menarik: Pola bimodal mencerminkan adanya segmentasi geografis berdasarkan kota. Namun, beberapa nilai ekstrem yang berada di luar cakupan wilayah perlu diverifikasi untuk memastikan tidak ada kesalahan input data.

*Follow-up: Validasi rentang longitude dan latitude diperlukan untuk menghindari data yang tidak relevan atau berada di luar cakupan wilayah geografis yang seharusnya. Jika ditemukan data yang tidak sesuai, tindakan seperti penghapusan atau pembetulan data perlu dilakukan.

9. cleanliness_rating

- **Distribusi:** Kolom ini memiliki distribusi dengan rata-rata 9.0 dan sebagian besar nilai berkisar antara 8 hingga 10, mencerminkan properti dengan tingkat kebersihan yang sangat baik. Tidak ditemukan nilai kosong pada kolom ini.
- **Observasi Menarik:** Sebagian besar properti memiliki nilai kebersihan yang sangat baik, yang mencerminkan fokus pada kualitas layanan. Namun, variabilitas yang rendah dapat membatasi penggunaannya untuk analisis prediktif.
- **Follow-up:** Tidak diperlukan tindakan tambahan, tetapi kolom ini relevan untuk analisis lebih lanjut terkait ulasan dan harga.

10. guest_satisfaction_overall

- **Distribusi:** Nilai dalam kolom ini bervariasi dari 50 hingga 100, dengan rata-rata sekitar 90. Sebagian besar properti mendapatkan nilai kepuasan di atas 85, mencerminkan ulasan yang sangat positif dari tamu.
- **Observasi Menarik:** Kolom ini mencerminkan kepuasan tamu yang sangat tinggi di sebagian besar properti. Data ini relevan untuk analisis kualitas properti.
- **Follow-up:** Tidak ada tindakan pre-processing tambahan yang diperlukan karena distribusi mencerminkan data yang konsisten.

B. Analisis Kolom Kategorikal

1. multi (Properti Mendukung Grup Besar)

- **Distribusi:** Sebagian besar properti dalam dataset memiliki nilai 0, menunjukkan bahwa mayoritas tidak mendukung grup besar. Properti dengan nilai 1, yang mendukung grup besar, hanya merupakan sebagian kecil dari total dataset.
- **Observasi Menarik:** Properti dengan nilai 1 mencerminkan segmen pasar yang lebih spesifik untuk kelompok besar, seperti acara keluarga atau perjalanan kelompok besar.
- **Follow-up:** Tidak ada tindakan pre-processing tambahan yang diperlukan untuk kolom ini. Namun, kolom ini relevan untuk segmentasi pasar lebih lanjut berdasarkan jenis penyewaan properti.

2. biz (Properti Cocok untuk Bisnis)

- **Distribusi:** Kolom ini memiliki distribusi yang serupa dengan multi, di mana sebagian besar properti memiliki nilai 0, yang menunjukkan bahwa mereka tidak dirancang untuk keperluan bisnis. Properti dengan nilai 1, yang dirancang untuk kebutuhan bisnis, hanya mencakup sebagian kecil dataset.
- **Observasi Menarik:** Properti yang cocok untuk bisnis dapat menjadi segmen menarik, terutama dalam analisis terkait harga sewa, lokasi strategis, atau fasilitas tambahan yang relevan bagi penyewa bisnis.
- **Follow-up:** Tidak ada tindakan tambahan yang diperlukan, tetapi kolom ini dapat digunakan untuk segmentasi lebih lanjut atau model prediktif.

3. room_shared dan room_private

- **Distribusi:** Pada kolom room_shared, sebagian besar nilai adalah False, menunjukkan bahwa ruang bersama jarang ditawarkan oleh properti dalam dataset. Sebaliknya, kolom room_private memiliki distribusi yang lebih merata antara True dan False, dengan nilai True menunjukkan bahwa banyak properti menawarkan ruang privat.
- **Observasi Menarik:** Properti dengan ruang privat lebih umum dibandingkan ruang bersama, yang mencerminkan preferensi pasar terhadap privasi yang lebih besar bagi penyewa.
- **Follow-up:** Tidak diperlukan tindakan tambahan pada tahap pre-processing untuk kedua kolom ini.

4. host_is_superhost

- **Distribusi:** Sebagian besar host dalam dataset tidak memiliki status superhost, dengan nilai False mendominasi distribusi. Namun, ada sejumlah host yang memiliki status superhost, yang dicerminkan oleh nilai True.
- **Observasi Menarik:** Status superhost dapat digunakan untuk menilai kualitas layanan properti, karena biasanya berhubungan dengan ulasan yang baik dan pengalaman penyewaan yang lebih positif.
- **Follow-up:** Tidak ada tindakan tambahan yang diperlukan, tetapi kolom ini dapat digunakan untuk analisis lebih lanjut mengenai hubungan antara status superhost dan performa properti.

5. City (Kota)

- **Distribusi:** Kota dengan jumlah properti terbanyak dalam dataset adalah London, sementara jumlah properti terendah ditemukan di Amsterdam. Distribusi ini mencerminkan ketidakseimbangan yang signifikan antara jumlah properti di masing-masing kota.
- **Observasi Menarik:** Ketidakseimbangan ini menunjukkan dominasi beberapa kota tertentu dalam dataset, yang dapat memengaruhi analisis perbandingan antar kota.
- **Follow-up:** Jika analisis berbasis kota dilakukan, disarankan untuk menggunakan teknik oversampling atau undersampling untuk menyeimbangkan distribusi data per kota.

6. Day_Type

- **Distribusi:** Data dalam kolom ini hampir seimbang antara dua kategori, yaitu **Weekend** dan **Weekday**, dengan sedikit dominasi kategori. **Weekend**. Distribusi ini mencerminkan data yang representatif untuk kedua jenis hari.
- **Observasi Menarik:** Keseimbangan distribusi ini menunjukkan bahwa dataset tidak bias terhadap salah satu jenis hari tertentu, sehingga cocok untuk analisis pola waktu.
- **Follow-up:** Tidak ada tindakan tambahan yang diperlukan karena distribusi sudah ideal.

7. room_type

- **Distribusi:** Properti dalam dataset didominasi oleh tipe Entire home/apt, diikuti oleh Private room, sementara tipe Shared room hanya mencakup sebagian kecil data. Dominasi ini mencerminkan preferensi pasar untuk privasi penuh.
- **Observasi Menarik:** Tipe Entire home/apt kemungkinan besar lebih mahal dan memiliki target pasar yang berbeda dibandingkan tipe lainnya.
- **Follow-up:** Tidak ada tindakan pre-processing tambahan yang diperlukan. Namun, kolom ini dapat digunakan untuk analisis segmentasi berdasarkan tipe properti.

C. Analisis Kolom Indeks/Tidak relevan untuk analisis

1. Unnamed: 0

- **Distribusi:** Kolom ini adalah indeks unik untuk setiap baris, dan nilainya meningkat secara berurutan dari 0 hingga 51706. Tidak ada variasi yang bermakna karena kolom ini hanya berfungsi sebagai penanda baris.
- **Observasi Menarik:** Kolom ini tidak relevan untuk analisis karena hanya berfungsi sebagai indeks.
- **Follow-up:** Kolom ini dapat dihapus dari dataset untuk menyederhanakan analisis, karena tidak memberikan informasi tambahan.

3. Multivariate Analysis (15 poin)

Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

- Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

In []: `data.sample(5)`

Out[]:

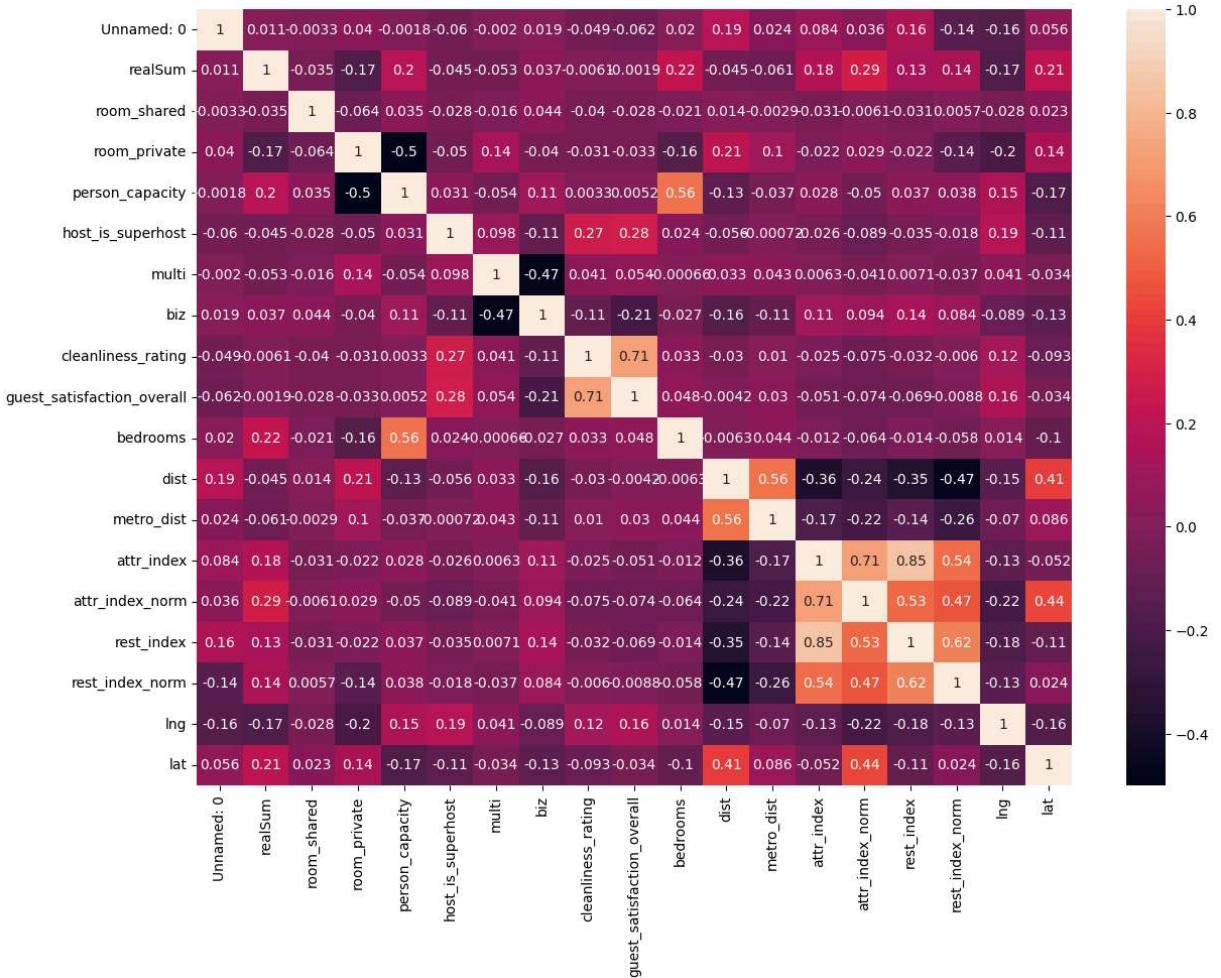
	Unnamed: 0	realSum	room_type	room_shared	room_private	person_capacity	hc
32511	2207	300.336320	Entire home/apt	False	False	2.0	
23207	564	339.118199	Entire home/apt	False	False	2.0	
46025	450	205.972025	Private room	False	True	2.0	
31253	949	631.482396	Entire home/apt	False	False	6.0	
41518	2277	126.949244	Private room	False	True	2.0	

5 rows × 22 columns

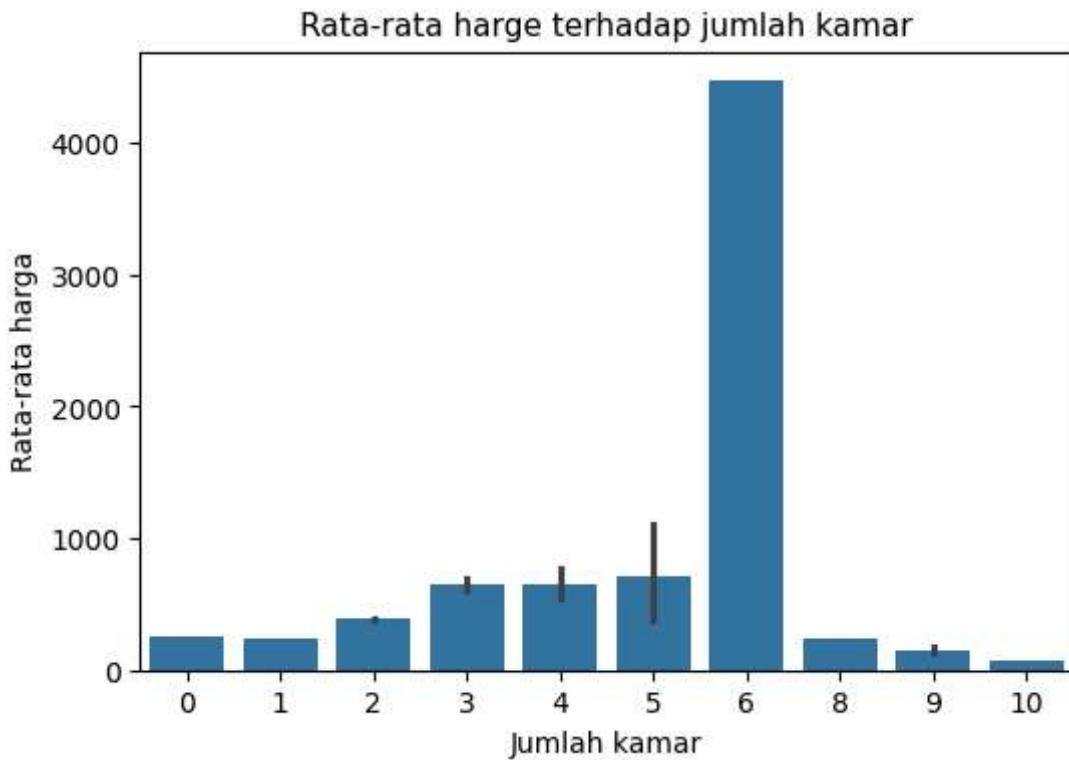
In []:

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(14,10))
sns.heatmap(data.corr(numeric_only=True), annot=True)
plt.show()
```



```
In [ ]: plt.figure(figsize=(6, 4))
sns.barplot(data, x="bedrooms", y="realSum")
plt.title("Rata-rata harga terhadap jumlah kamar", fontsize=11)
plt.ylabel("Rata-rata harga")
plt.xlabel("Jumlah kamar")
plt.show()
```



A. Korelasi antara masing-masing feature dan label

1. **Korelasi positif** : berdasarkan heatmap di atas, features attr_index_norm (0.29), bedrooms (0.22), lat (0.21), person_capacity (0.20) memiliki korelasi positif terhadap feature realSum.

feature attr_index_norm menunjukkan lokasi dengan atraksi lebih lebih tinggi cenderung memiliki harga yang lebih tinggi juga

feature bedrooms menunjukkan jumlah kamar tidur yang lebih banyak akan memiliki harga yang lebih mahal

feature lat menunjukkan hubungan geografis dengan harga

feature person_capacity menunjukkan properti dengan kapasitas tamu dalam jumlah besar akan memiliki harga yang lebih tinggi

2. **Korelasi negatif**: berdasarkan heatmap di atas, features lng (-0.17) dan room_private (-0.17) memiliki korelasi negatif terhadap feature realSum.
3. **Korelasi lemah**: berdasarkan heatmap di atas, features cleanliness_rating (-0.006) dan guest_satisfaction_overall (-0.002)tidak memiliki pengaruh terhadap harga, sehingga **dapat dipertimbangkan untuk dihilangkan jika tidak relevan**

B. Korelasi antar-feature

1. Korelasi antara **attr_index** dan **attr_index_norm** memiliki nilai di atas 0.7 yang menunjukkan kedua feature tersebut **redundant**. Feature redundant juga tercermin

pada feature **rest_index** dan **rest_index_norm**. Untuk selanjutnya, feature attr_index dan rest_index dapat dihilangkan.

2. Feature metro_dist dan dist memiliki kemungkinan korelasi positif, menunjukkan adanya hubungan geografis
3. Feature guest_satisfaction_overall, cleanliness_rating, dan host_is_superhost tidak memiliki korelasi yang signifikan dengan fitur lain. Dapat dipertimbangkan kembali untuk menggunakan feature tersebut karena mungkin feature tersebut kurang relevan.
4. Kombinasi antara fitur geografis seperti lat, lon, dist, dan metro_dist dapat memberikan wawasan tambahan tentang lokasi strategis yang memengaruhi harga.

4. Business Insight (30 poin)

Selain EDA, lakukan juga beberapa analisis dan visualisasi untuk menemukan suatu business insight. Tuliskan minimal 3 insight, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

A. Insight 1: Distribusi harga berdasarkan tipe kamar

Kamar dengan tipe Entire home/apt memiliki rentang harga yang jauh lebih tinggi dibandingkan jenis lainnya, dengan beberapa properti berada di kisaran harga yang sangat tinggi. Sementara itu, Private room dan Shared room memiliki rentang harga lebih terbatas dengan rata-rata yang lebih rendah.

Rekomendasi:

1. Targetkan Segmen Premium untuk Tipe Entire Home/Apt

Pemasaran tipe kamar Entire home/apt dapat ditargetkan kepada segmen pelanggan premium atau kelompok pelanggannya yang mengutamakan kenyamanan.

2. Paket dan Promosi untuk kamar tipe Private Room dan Shared Room

Pemasaran tipe kamar Private room dapat menyoroti nilai ekonomis bagi solo traveler atau pelanggan yang mencari opsi penginapan yang lebih terjangkau.

3. Peningkatan Fitur (Amenities) dan Harga Berdasarkan Tipe Kamar

Tambahan fitur/amenities dalam penginapan, dapat dipersonalisasikan untuk membedakan pengalaman menginap pada setiap jenis tipe kamar. Contohnya adalah menyediakan layanan jemput untuk tipe kamar Entire home/apt (seperti antar-jemput bandara) dan meningkatkan fasilitas lebih untuk Private room dan Shared room. Hal ini dapat meningkatkan daya tarik masing-masing segmen pasar dan dapat menjadi nilai tambahan saat pelanggan sedang mencari penginapan online.

B. Insight 2: Kepuasan tamu dan penilaian kebersihan berdasarkan jenis kamar

Berdasarkan visualisasi diatas, secara umum, kepuasan tamu meningkat dengan penilaian kebersihan yang lebih tinggi, terutama pada Entire home/apt. Shared room menunjukkan lebih banyak variasi dalam kepuasan dibanding tipe lainnya.

Rekomendasi:

Jadikan kebersihan prioritas utama untuk meningkatkan kepuasan tamu, terutama pada tipe Shared room.

Berikan panduan atau pelatihan tambahan kepada host untuk meningkatkan standar kebersihan.

1. Targetkan Segmen Premium untuk Tipe Entire Home/Apt

Karena kepuasan tamu meningkat dengan penilaian kebersihan yang lebih tinggi, penting untuk memastikan kebersihan terjaga di seluruh tipe kamar. Airbnb dapat fokus mensosialisasikan atau investasi untuk meningkatkan standar kebersihan dan membuat sistem audit kebersihan yang lebih ketat untuk memastikan pengalaman pelanggan yang lebih baik.

2. Tingkatkan Pengalaman di Shared Room dengan Meningkatkan Kebersihan dan Fasilitas

Karena shared room memiliki variasi lebih banyak dalam kepuasan dibandingkan dengan tipe lainnya, airbnb dapat fokus dalam sosialisasi dan fokuskan pada peningkatan kebersihan dan fasilitas tambahan di tipe kamar ini. Fasilitas tambahan seperti penyediaan tempat penyimpanan pribadi, perlengkapan tidur yang lebih baik, dan suasana yang lebih nyaman. 3 hal ini dapat dapat membantu mengurangi variasi yang banyak dalam kepuasan tamu dan meningkatkan rating secara keseluruhan.

C. Insight 3: Harga rata-rata berdasarkan city dan day type

Dari perbandingan harga rata-rata weekday dan weekend dalam 10 kota, ditemukan bahwa beberapa kota, seperti Amsterdam dan London, memiliki harga yang lebih tinggi, terutama pada weekend. Pola harga ini menunjukkan suatu peluang untuk memaksimalkan keuntungan pada hari tertentu.

Rekomendasi:

1. Penetapan Harga Dinamis Berdasarkan Hari

Optimalkan strategi harga dinamis untuk meningkatkan harga saat akhir pekan di kota-kota dengan permintaan tinggi. Ini dapat memaksimalkan pendapatan selama peak days, sementara tetap menarik pelanggan pada weekdays dengan harga yang lebih kompetitif.

2. Penawaran Khusus untuk Weekdays

Agar lebih menarik pelanggan, penawaran promosi atau diskon eksklusif dapat dilakukan khusus untuk order pada hari Senin sampai Jumat (Weekdays). Dengan

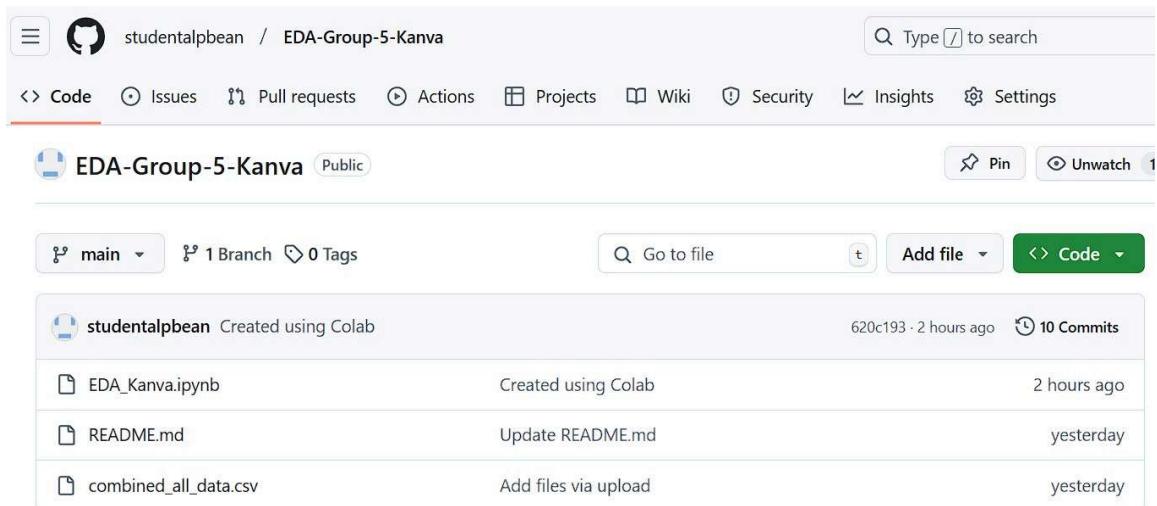
demikian, dapat menarik pelanggan yang mungkin cenderung memilih liburan selama weekend dengan harga yang lebih terjangkau, sekaligus menjaga tingkat pemanfaatan di luar weekend yang lebih padat.

5. Git (15 poin)

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

A. Buat Repository Git: Berikut adalah Link Repository Git:
(<https://github.com/studentalpbean/EDA-Group-5-Kanva>)

B Upload file notebook atau file penggerjaan lainnya pada repository tersebut: File penggerjaan telah diupload pada Repository Git



The screenshot shows a GitHub repository page for 'EDA-Group-5-Kanva'. At the top, there's a navigation bar with links for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. Below the navigation bar, the repository name 'EDA-Group-5-Kanva' is displayed, along with a 'Public' badge and 'Pin' and 'Unwatch' buttons. A commit history section shows a single commit from 'studentalpbean' created using Colab, made 2 hours ago. The commit details show the creation of 'EDA_Kanva.ipynb', an update to 'README.md', and an addition of 'combined_all_data.csv'. The commit was made yesterday.

File	Description	Time
EDA_Kanva.ipynb	Created using Colab	2 hours ago
README.md	Update README.md	yesterday
combined_all_data.csv	Add files via upload	yesterday

C Untuk file README, dapat merupakan summary insight yang telah didapatkan dari EDA: File README di Link Repository Git sudah diupdate dengan Insight Utama