**Seattle AirBnB Listing Dataset (Rakamin Academy Final Project)**

Dalam final project ini, kelompok kami ingin membuat machine learning yang bertujuan untuk memberikan wawasan yang diperlukan dan dapat membantu AirBnB dalam meningkatkan jumlah customer di Kota Seattle sehingga posisi mereka di pasar Kota Seattle semakin kuat.

**Kelompok 2 ResNet**

1. Project Manager : Mohammad Fauzan
2. Data Analyst : Indah Mutiah Utami. MZ
3. Data Engineer : Yusuf Nafi Farhan
4. Data Scientist : Julian

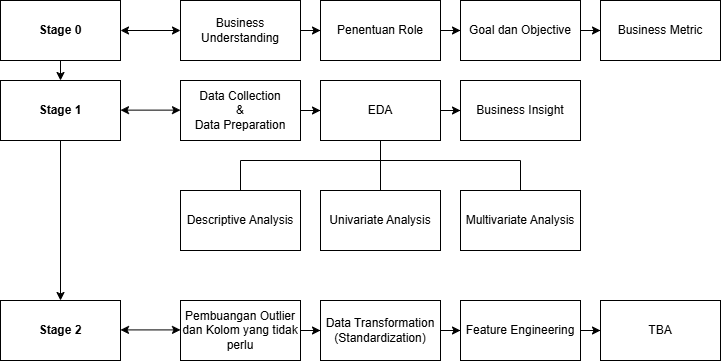
**Daftar Isi**

* Prerequisites
* Penjelasan Sebelum EDA
* EDA
* Business Insight

**Prerequisites**

1. Download data here
2. Clone repositori ini :

git clone <https://github.com/Podjan/ResNet2.git>



**Penjelasan Sebelum EDA**

**Dataset**

Ada tiga dataset pada pekerjaan kali ini.

1. Calendar yang berisi tentang data penghasilan dari AirBnB selama setahun.
2. Listings yang berisi tentang data detail lengkap mengenai setiap listings.
3. Reviews yang berisi tentang ulasan dari setiap listing.

Beberapa hal yang perlu diperhatikan

1. Penentuan kolom dan dataset yang diambil dilakukan diawal untuk meminimalisir resiko dan membuang yang tidak perlu.
2. Poin pertama bukan berarti selanjutnya tidak ada pengambilan keputusan untuk membuang kolom. Jika ada kolom yang korelasinya mendekati 1 kemungkinan akan dibuang juga.
3. Penentuan kolom dan dataset diawal juga didasari oleh goal dan objective yang ingin dicapai.

**Goal dan Objectives**

**Goal**

Goals yang ingin dicapai dalam studi case ini adalah meningkatkan jumlah customer AirBnB di tahun berikutnya dengan memberikan wawasan yang akurat kepada pemilik AirBNB mengenai jenis properti yang paling diminati oleh tamu dan tingkat kepuasan para tamu, sehingga mereka dapat mengoptimalkan strategi pemasaran, penetapan harga, dan pengelolaan inventaris di tahun berikutnya.

**Objectives**

1. Memprediksi jumlah pengunjung di tahun berikutnya berdasarkan kualitas dan harga.
2. Meningkatkan tingkat kepuasan customer terhadap kualitas hotel yang pengunjungnya sepi.
3. Memprediksi harga hotel di tahun depan untuk memberikan rekomendasi harga yang tepat di setiap hotel.

**Langkah awal**

1. Memisahkan kolom date menjadi kolom bulan dan tahun pada dataset calendar
2. Melakukan dropna karena pada dataset calendar, na itu avaiability dari listingnya false dan mengakibatkan tidak ada pemasukan untuk listing tersebut.
3. Untuk memudahkan melakukan analisa yang lain, kami melakukan groupby di dataset calendar yang bertujuan mendapatkan median price dan jumlah customer (jumlah id) dari masing listing id di setiap bulannya.

df1g = df1.groupby(['listing\_id', 'tahun', 'bulan\_num']).agg(

pendapatan\_per\_bulan=('pendapatan', 'sum'),

median\_price=('pendapatan', 'median'),

jumlah\_customer=('listing\_id', 'size')

).reset\_index()

1. Mengubah beberapa kolom di dataset listing menjadi tipe yang diinginkan

# Mengubah data yang bertipe string ke float

dfl2['price'] = dfl2['price'].str.replace('$', '').str.replace(',', '').astype(float)

dfl2['weekly\_price'] = dfl2['weekly\_price'].str.replace('$', '').str.replace(',', '').astype(float)

dfl2['monthly\_price'] = dfl2['monthly\_price'].str.replace('$', '').str.replace(',', '').astype(float)

dfl2['host\_response\_rate'] = dfl2['host\_response\_rate'].str.replace('%', '').astype(float)

dfl2['host\_acceptance\_rate'] = dfl2['host\_acceptance\_rate'].str.replace('%', '').astype(float)

# Mengubah data yang beripe integer ke string

dfl2['host\_id'] = dfl2['host\_id'].astype(str)

dfl2['latitude'] = dfl2['latitude'].astype(str)

dfl2['longitude'] = dfl2['longitude'].astype(str)

1. Memasukkan kolom jumlah id dari dataset calendar ke dataset listing.

df1g2 = df1g [['listing\_id', 'pendapatan\_per\_bulan','jumlah\_customer']]

df1g3 = df1g2.groupby(['listing\_id']).agg(

total\_pendapatan=('pendapatan\_per\_bulan', 'sum'),

total\_customer=('jumlah\_customer', 'sum')).reset\_index()

# Menggabungkan df1g3 dan dfl2 berdasarkan 'listing\_id dengan tujuan memembuat kolom jumlah\_id di dataset listing

dfl3 = pd.merge(df1g3, dfl2, on='listing\_id', how='left')

1. Sekarang kita punya dua dataframe/dataset yang akan diolah

df1

df2 = dfl3

**Fitur yang akan digunakan**

df1 (dataset calendar final)

* listing\_id
* date
* bulan\_num (tambahan fitur di dataset calendar)
* tahun (tambahan fitur di dataset calendar)
* pendapatan

df2 (dataset listing final)

* listing\_id
* total\_pendapatan (tambahan fitur di dataset listing)
* total\_customer (tambahan fitur di dataset listing)
* name
* host\_id
* host\_response\_time
* host\_response\_rate
* host\_acceptance\_rate
* host\_is\_superhost
* host\_identity\_verified
* zipcode
* latitude
* longitude
* is\_location\_exact
* property\_type
* room\_type
* accommodates
* bathrooms
* bedrooms
* beds
* bed\_type
* price
* weekly\_price
* monthly\_price
* guests\_included
* minimum\_nights
* maximum\_nights
* review\_scores\_rating
* review\_scores\_accuracy
* review\_scores\_cleanliness
* review\_scores\_checkin
* review\_scores\_communication
* review\_scores\_location
* review\_scores\_value
* instant\_bookable
* cancellation\_policy
* require\_guest\_profile\_picture
* require\_guest\_phone\_verification

Note: Untuk dataset review, dataset tersebut tidak kami masukkan sebagai model utama machine learning. Tetapi data itu akan kami gunakan sebagai model untuk melihat sentimen customer dari masing-masing listing. Kami menggunakan geminiAI untuk mengerjakan model tersebut. Untuk saat ini, codingnya ada di repository ini dengan nama API.ipynb.

**EDA**

**Descriptive Statistic**

Pada proses descriptive statistic ini, berguna untuk melihat nilai statistika dari setiap data yang sudah diolah sebelumnya. Pada bagian ini descriptive statistic menampilkan 2 data frame yang berguna untuk mempertimbangkan analisis selanjutnya.

**df1**

df1 merupakan data frame yang data sourcenya diambil dari dataset calendar. Sebelum dilakukan data cleansing terdapat sekitar 1.393.570 rows yang kemudian diolah dengan membuang missing valuenya sehingga feature payment yang awalnya 459028 didrop semua menjadi 0.

|  |  |  |
| --- | --- | --- |
| Features | Before | After |
| Pendapatan | 459028 | 0 |

Salah satu pertimbangan melakukan drop terhadap missing value adalah karena listing yang available saat itu berstatus false artinya pada saat itu penginapannya tidak available dan tidak ada proses transaksi pada listing tersebut. Hal ini sangat mempengaruhi proses selanjutnya mengingat goal yang ingin dicapai yaitu memprediksi jumlah pengunjung di tahun berikutnya dan memprediksi harga hotel di tahun depan, jadi membersihkan data null sangatlah penting dalam proses ini.

Saat ini df1 mempunyai jumlah row sebanyak 934.542. Data tersebut adalah data transaksi dari semua listing.

Adapun hasil descriptive statistic nya adalah sbb:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | listing\_id | date | Bulan\_num | tahun | pendapatan |
| count | 934542 | 934542 | 934542 | 934542 | 934542 |
| mean | 5.30552e+06 | 2016-07-08 08:12:42.911458560 | 6.66112 | 2016.01 | 137.945 |
| min | 3335 | 2016-01-04 00:00:00 | 1 | 2016 | 10 |
| 25% | 2.87598e+06 | 2016-04-07 00:00:00 | 4 | 2016 | 75 |
| 50% | 5.61562e+06 | 2016-07-08 00:00:00 | 7 | 2016 | 109 |
| 75% | 7.87334e+06 | 2016-10-09 00:00:00 | 10 | 2016 | 160 |
| Max | 1.03402e+07 | 2017-01-02 00:00:00 | 12 | 2017 | 1650 |
| std | 2.97482e+06 | nan | 3.4464 | 0.0778456 | 105.063 |

Dari hasil descriptive statistic tidak menunjukkan adanya data yang invalid dan sebagian besar datanya berasal dari tahun 2016 dengan rata-rata distribusi datanya berada dipertengahan bulan sekitar Juni-Juli. median dan mean pendapatan yang lebih tinggi dari standar devian menunjukkan data terkonsentrasi di nilai tengah. Karena mean pendapatan lebih tinggi dari median, maka data ini berbentuk right skewed. Standar devian yang mendekati nilai median pendapatan juga menunjukkan sebaran data yang lebar. Bentuk dari date bisa diperbaiki saat data transformation.

**df2**

df2 adalah data frame yang data source nya diambil dari dataset listing\_id. data ini sangat berguna dalam analisis tingkat kepuasan customer.

Pada df2 total data sebelum diolah ada sekitar 3818 rows listing penginapan. Selanjutnya, data terebut diolah dengan melakukan groupby berdasarkan listing\_id untuk mendapatkan total customer dari masing-masing listing.

df1g3 = df1g2.groupby(['listing\_id']).agg(

total\_pendapatan=('pendapatan\_per\_bulan', 'sum'),

total\_customer=('jumlah\_customer', 'sum')).reset\_index()

Sehingga dari hasil tersebut didapatkan total keseluruhan row adalah 3723 dengan 38 kolom. Adapun kolom-kolom yang masih memiliki missing value adalah sebagai berikut:

|  |  |
| --- | --- |
| host\_response\_time | 480 |
| host\_response\_rate | 480 |
| host\_acceptance\_rate | 723 |
| host\_is\_superhost | 2 |
| host\_identity\_verified | 2 |
| zipcode | 7 |
| property\_type | 1 |
| bathrooms | 16 |
| bedrooms | 6 |
| beds | 1 |
| weekly\_price | 1750 |
| monthly\_price | 2231 |
| review\_scores\_rating | 622 |
| review\_scores\_accuracy | 633 |
| review\_scores\_cleanliness | 628 |
| review\_scores\_checkin | 633 |
| review\_scores\_communication | 626 |
| review\_scores\_location | 630 |
| review\_scores\_value | 631 |

Adapun missing value tersebut akan diolah lebih lanjut ditahap berikutnya.

Berikut hasil descriptive statistic dari df2:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| listing\_id | 3723 | 5.54805e+06 | 2.96979e+06 | 3335 | 3.24243e+06 | 6.11982e+06 | 8.0368e+06 | 1.03402e+07 |
| total\_pendapatan | 3723 | 34626.7 | 33041.2 | 49 | 13737.5 | 26935 | 44201.5 | 362941 |
| total\_customer | 3723 | 251.019 | 122.12 | 1 | 139 | 313 | 360 | 365 |
| host\_response\_rate | 3243 | 94.8545 | 11.9244 | 17 | 98 | 100 | 100 | 100 |
| host\_acceptance\_rate | 3000 | 99.9667 | 1.82574 | 0 | 100 | 100 | 100 | 100 |
| accommodates | 3723 | 3.33306 | 1.96551 | 1 | 2 | 3 | 4 | 16 |
| bathrooms | 3707 | 1.25681 | 0.587699 | 0 | 1 | 1 | 1 | 8 |
| bedrooms | 3717 | 1.29782 | 0.873499 | 0 | 1 | 1 | 2 | 7 |
| beds | 3722 | 1.72864 | 1.13387 | 1 | 1 | 1 | 2 | 15 |
| price | 3723 | 126.993 | 88.5101 | 20 | 75 | 100 | 150 | 1000 |
| weekly\_price | 1973 | 787.725 | 534.051 | 100 | 455 | 650 | 950 | 6300 |
| monthly\_price | 1492 | 2604.68 | 1724.23 | 500 | 1507.5 | 2200 | 3100 | 19500 |
| guests\_included | 3723 | 1.65834 | 1.29537 | 0 | 1 | 1 | 2 | 15 |
| minimum\_nights | 3723 | 2.36557 | 16.5044 | 1 | 1 | 2 | 2 | 1000 |
| maximum\_nights | 3723 | 785.906 | 1702.45 | 1 | 60 | 1125 | 1125 | 100000 |
| review\_scores\_rating | 3101 | 94.5485 | 6.62001 | 20 | 93 | 96 | 99 | 100 |
| review\_scores\_accuracy | 3090 | 9.63625 | 0.699891 | 2 | 9 | 10 | 10 | 10 |
| review\_scores\_cleanliness | 3095 | 9.56058 | 0.794371 | 3 | 9 | 10 | 10 | 10 |
| review\_scores\_checkin | 3090 | 9.78835 | 0.594455 | 2 | 10 | 10 | 10 | 10 |
| review\_scores\_communication | 3097 | 9.80949 | 0.568215 | 2 | 10 | 10 | 10 | 10 |
| review\_scores\_location | 3093 | 9.60782 | 0.631045 | 4 | 9 | 10 | 10 | 10 |
| review\_scores\_value | 3092 | 9.45343 | 0.749484 | 2 | 9 | 10 | 10 | 10 |

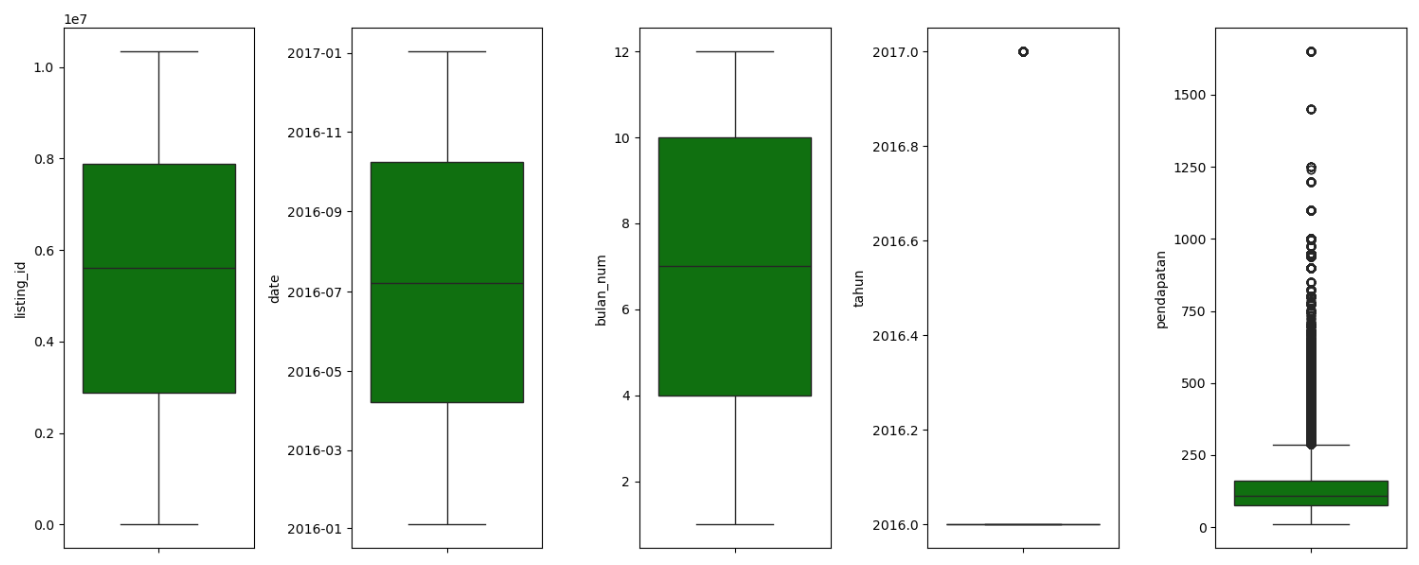
Dari data tersebut tidak menunjukkan adanya data yang invalid. Selain itu, beberapa point yang bisa kita dapat dari data tesebut adalah :

1. Nilai rata-rata jumlah\_id berkisar 251 dengan variasi sebesar 122 dan median 310, artinya beberapa listing memiliki sebagian besar total\_customer nya 310 yang mana itu mengindikasikan bahwa total\_customer memilik median > mean sehingga distribusinya cenderung negative skewed.
2. Nilai rata-rata total pendapatan setiap listing sekitar 34-35k dollar dan mediannya adalah 26,9k dollar. Hal ini berarti datanya bersifat right skewed, artinya ada listing-listing yang mempunyai pendapatan yang sangat tinggi dibandingkan kebanyakan listing-listing lain.
3. Untuk nilai kepuasan didapatkan dari kolom host\_response\_rate, host\_acceptance\_rate, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, menunjukkan skor rata-rata untuk semua aspek mendekati nilai maksimum (9,5-10) yang artinya kebanyakan listing memiliki ulasan yang sangat positif.
4. Pada kolom price menunjukkan adanya perbedaan harga minimum dan maximumnya yang cukup jauh, artinya ada listing dengan harga yang sangat mahal dan sangat murah. Ini menandakan adanya outlier pada data ini. Selain itu, sebagian besar listing price per malamnya berkisar $105.
5. Accomodates menunjukkan jumlah tamu yang dapat ditampung oleh suatu listing. Rata-rata kapasitas yang bisa ditampung adalah sekitar 3 tamu dengan sebagian besar listing mampu menampung 3-4 tamu.
6. Sebagian besar listing hanya memiliki satu kamar mandi, satu kamar tidur, dan satu hingga dua tempat tidur. Namun beberapa listing memiliki jumlah fasilitas yang tinggi, dengan nilai maksimum 8 kamar mandi, 7 kamar tidur, dan 20 tempat tidur.

**Univariate Analysis**

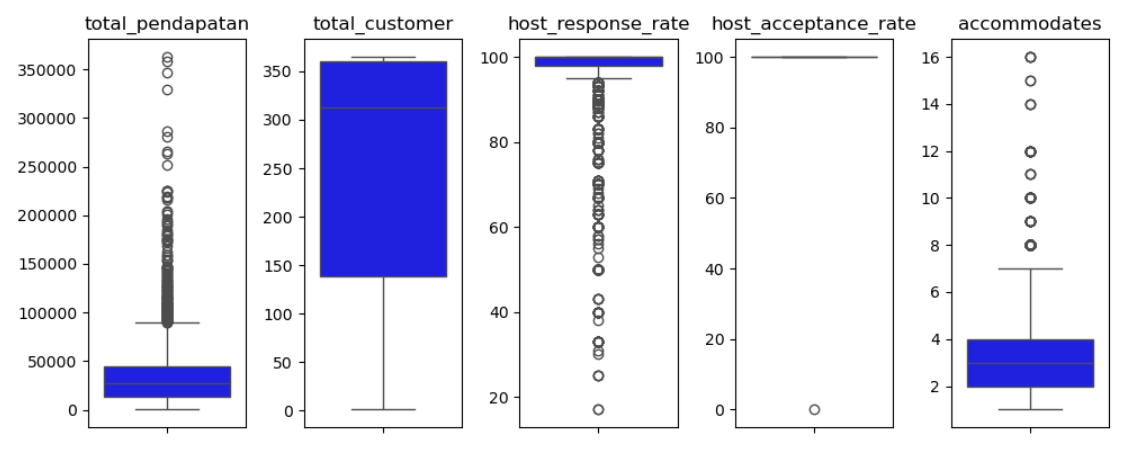
Pada analisis univariate, untuk df2, kami mengelompokkan kolom yang bertipe float dan integer ke kolom bernama nums (numerical). Karena ada 21, kami membagi nums tersebut menjadi 5, yaitu nums21,nums22, nums23, nums 24, dan nums25. Tujuannya, agar tidak memberatkan pc saat melakukan run codingnya.

**df1**

****

1. Listing ID: Rentang nilai Listing ID sangat lebar. Tidak ada outlier yang signifikan.
2. Date: Distribusi data sangat merata di sekitar bulan 4-10.
3. Bulan\_num: Distribusi data sangat merata dengan rentang bulan 4-10 (sama seperti date).
4. Pendapatan: Terdapat outlier yang sangat banyak mulai diatas 250 dollar.
5. Tahun: Data sebagian besar ada di 2016 akan tetapi ada outlier pada 2017 yang harus di analisis. Untuk tahun 2017, data terakhir tahun 2017 adalah tanggal 2 bulan Januari 2017. Untuk date pertama di tanggal 4 bulan Januari 2016

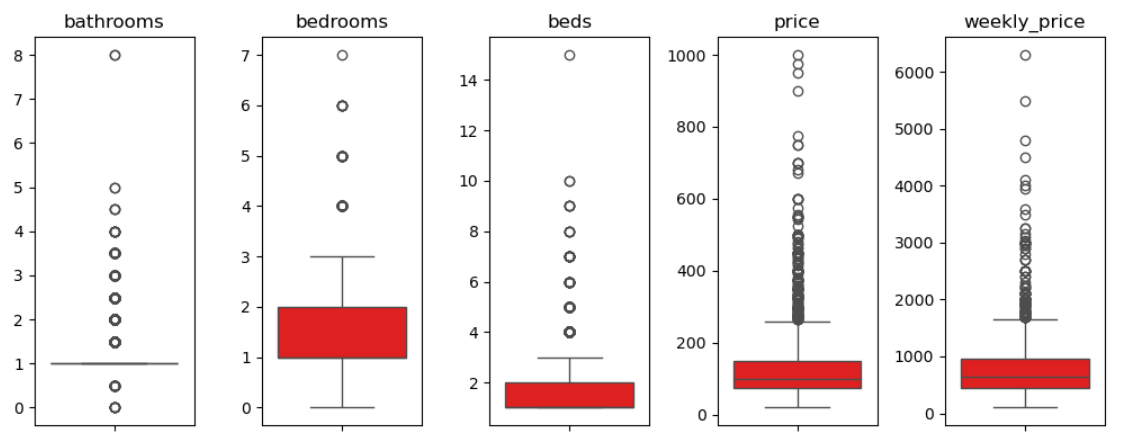
**df2**

****

**nums21**

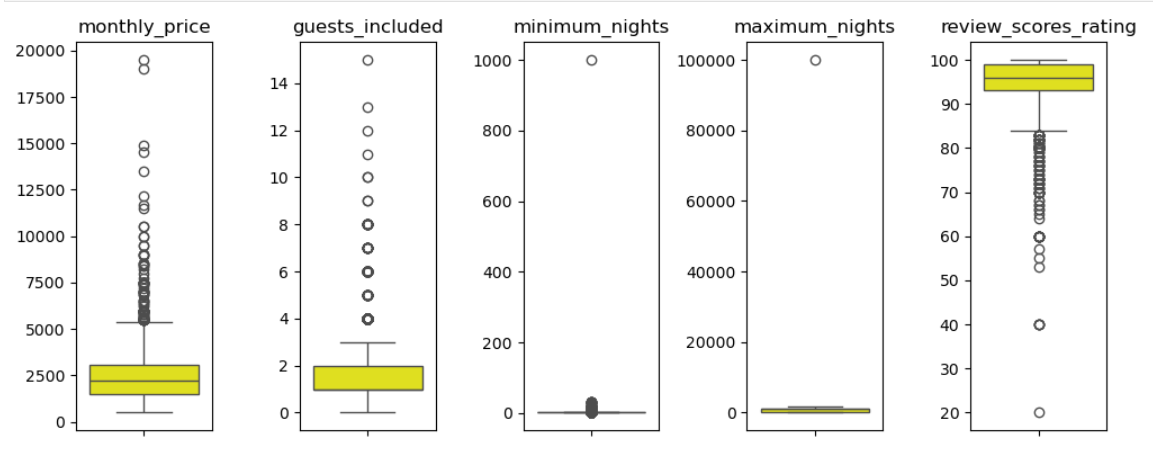
1. total\_pendapatan: Distribusi pendapatan semua listing terpusat di sekitar 250k, dengan outlier mulai di atas 100k
2. total\_customer: Distribusi terpusat di sekitar angka 300, menunjukkan total customer yang konsisten antar data. Kemungkinan mewakili jumlah ulasan atau interaksi, dengan nilai tinggi mengindikasikan listing yang populer.
3. host\_response\_rate: Banyak host dengan tingkat respons tinggi, tetapi ada juga yang sangat rendah. Hal ini menunjukkan bahwa responsivitas host cukup beragam, dengan yang cepat lebih disukai tamu.
4. host\_acceptance\_rate: Mayoritas host memiliki tingkat penerimaan mendekati 100%, tetapi ada juga yang lebih selektif. Host dengan penerimaan tinggi cenderung populer atau memiliki kebijakan khusus.
5. accommodates: Sebagian besar listing menampung hingga 4 orang, dengan beberapa outlier yang menampung lebih banyak. Kapasitas akomodasi beragam, dari 1-2 hingga lebih dari 8 orang.

**nums22**

****

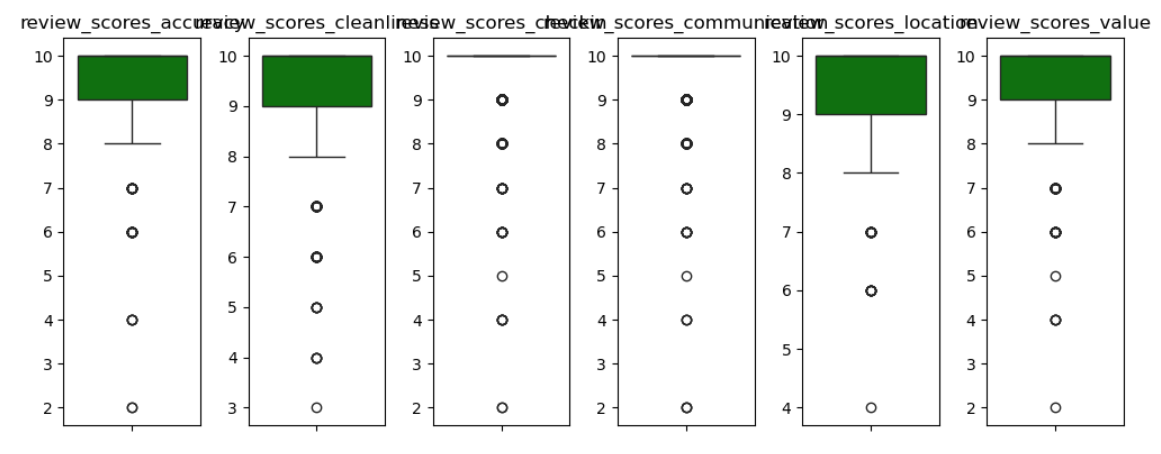
1. bathrooms: Sebagian besar listing memiliki 1-2 kamar mandi, dengan beberapa outlier yang memiliki lebih banyak. Fasilitas kamar mandi bervariasi, meski mayoritas terbatas.
2. Bedrooms: Sebagian besar listing memiliki 1-2 kamar tidur, cocok untuk individu atau keluarga kecil. Listing dengan lebih banyak kamar tidur biasanya untuk kelompok besar atau masa inap lebih lama.
3. Beds: Umumnya ada 2-4 tempat tidur per listing, dengan beberapa memiliki lebih banyak. Beberapa listing menawarkan tempat tidur tambahan, seperti sofa bed.
4. Price: Harga per malam bervariasi dengan median rendah, namun ada outlier dengan harga tinggi, dipengaruhi oleh lokasi, ukuran, dan fasilitas.
5. Weekly\_price: Harga mingguan mirip dengan harga harian, biasanya kelipatan dengan diskon atau biaya tambahan.

**nums23**

****

1. Monthly\_price: Harga bulanan cenderung lebih murah per hari dibandingkan harga harian, sering kali disertai diskon.
2. guests\_included: Sebagian besar listing dapat menampung hingga 4 tamu, dengan beberapa yang bisa menampung lebih dari 8 orang. Listing dengan kapasitas besar cocok untuk keluarga atau kelompok besar.
3. minimum\_nights: Minimum menginap umumnya rendah (1-2 malam), dengan beberapa listing yang mensyaratkan durasi lebih lama untuk masa inap jangka panjang.
4. maximum\_nights: Mayoritas listing memiliki batas maksimum menginap yang fleksibel atau tanpa batas, namun ada beberapa yang membatasi masa inap lebih pendek.
5. review\_scores\_rating: Sebagian besar listing memiliki rating tinggi (di atas 80), menunjukkan kualitas yang baik dan tamu yang puas.

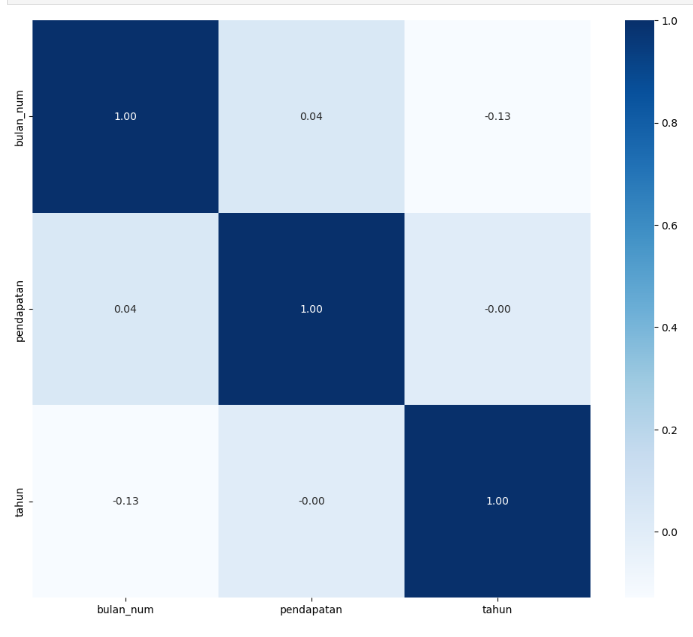
**nums24**

****

1. review\_scores\_accuracy: Skor akurasi deskripsi umumnya tinggi (di atas 8), menunjukkan deskripsi listing yang sesuai dengan ekspektasi tamu.
2. review\_scores\_cleanliness: Skor kebersihan mayoritas tinggi (di atas 8), menunjukkan host menjaga standar kebersihan yang baik.
3. review\_scores\_checkin: Skor check-in umumnya tinggi, meski ada beberapa kesulitan di beberapa listing.
4. review\_scores\_communication: Skor komunikasi juga tinggi, menandakan host responsif dan mudah diajak berkomunikasi.
5. review\_scores\_location: Skor lokasi tinggi, menunjukkan lokasi listing umumnya strategis dan mudah diakses.
6. review\_scores\_value: Skor nilai tinggi, menunjukkan tamu merasa puas dengan harga yang dibayar sesuai fasilitas yang diterima.

**Multivariate Analysis**

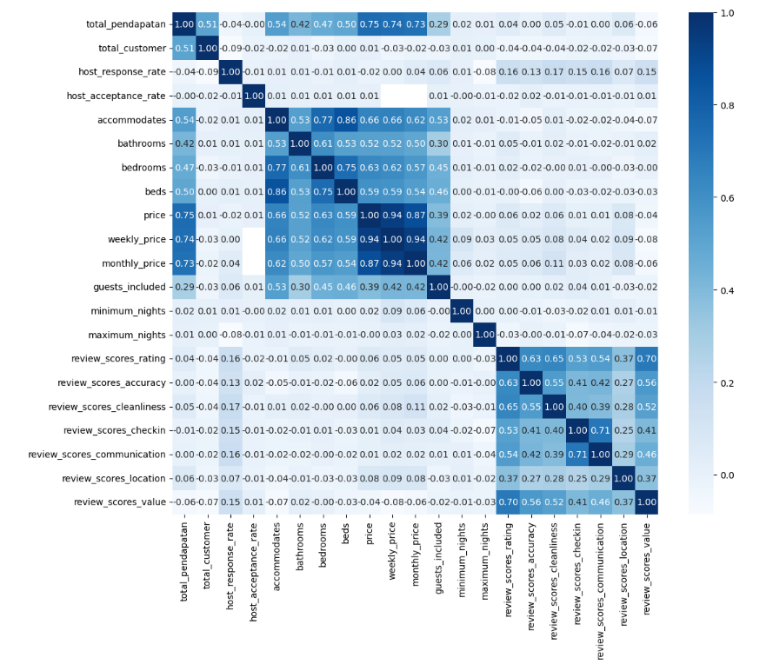
**df1**



matriks korelasi antar variabel dalam dataset, yaitu bulan\_num, pendapatan, dan tahun. Berikut penjelasan dari korelasi antar variabel:

1. tahun dengan bulan\_num (-0.13): Korelasi negatif lemah, yang menunjukkan bahwa ketika tahun meningkat, bulan\_num cenderung sedikit menurun. Namun, hubungan ini tidak terlalu kuat, sehingga tidak ada keterkaitan yang jelas antara tahun dan bulan\_num.
2. tahun dengan pendapatan (-0.00): Korelasinya negatif (hampir nol (karena angka dibelakang komanya hanya 2 jadi kelihatannya 0 padahal tidak)), yang menunjukkan bahwa ketika tahun naik maka pendapatan turun. Tetapi karena korelasinya sangat mendekati 0, maka tidak ada hubungan linear antara tahun dan pendapatan. Hal ini juga dikarenakan data tahun 2017 sangat sedikit, berakhir di tanggal 6 Januari 2017.
3. bulan\_num dengan pendapatan (0.04): Korelasi positif, ketidka bulan semakin mendekati akhir tahun maka pendapatan juga meningkat. Tetapi korelasinya hampir nol, menunjukkan tidak ada hubungan linear yang signifikan antara bulan\_num dan median\_price. Ini berarti bahwa perubahan dalam bulan\_num tidak mempengaruhi pendapatan secara konsisten.

**df2**



1. Korelasi Antar Variabel Fasilitas Properti

* (accommodates, bathrooms, bedrooms, dan beds): Terdapat korelasi yang cukup tinggi di antara variabel ini. Contohnya:
* accommodates dan bedrooms memiliki korelasi sebesar 0.77.
* bedrooms dan beds memiliki korelasi sebesar 0.75.
* Korelasi yang tinggi ini menunjukkan bahwa jumlah kamar tidur (bedrooms), kamar mandi (bathrooms), dan tempat tidur (beds) cenderung bertambah seiring bertambahnya kapasitas (accommodates) properti. Hal ini logis karena properti yang lebih besar biasanya memiliki lebih banyak fasilitas.
* (price, weekly\_price, dan monthly\_price): Korelasi yang sangat tinggi antara price, weekly\_price, dan monthly\_price (0.94 hingga 1.00) menunjukkan bahwa harga harian, mingguan, dan bulanan sangat berhubungan dan sejalan. Ketika harga harian meningkat, harga mingguan dan bulanan juga cenderung meningkat.

1. Korelasi Antar Variabel Skor Ulasan Variabel skor ulasan seperti review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin review\_scores\_communication, review\_scores\_location, dan review\_scores\_value memiliki korelasi positif yang cukup tinggi satu sama lain. Sebagai contoh:

* review\_scores\_rating memiliki korelasi sebesar 0.70 dengan review\_scores\_cleanliness, 0.65 dengan review\_scores\_checkin, dan 0.70 dengan review\_scores\_value. Korelasi positif yang tinggi ini menunjukkan bahwa ketika suatu properti memiliki skor tinggi dalam satu aspek (misalnya cleanliness atau checkin), biasanya aspek lainnya juga mendapatkan skor tinggi. Ini dapat menunjukkan bahwa properti yang bagus dalam satu faktor cenderung berkinerja baik dalam faktor ulasan lainnya.

1. Hubungan Fasilitas Properti dengan Harga

* (accommodates, bathrooms, bedrooms, dan beds dengan price): Semua variabel ini memiliki korelasi positif moderat dengan price (0.53 hingga 0.63), menunjukkan bahwa semakin banyak kapasitas atau fasilitas yang dimiliki properti (lebih banyak kamar tidur, kamar mandi, atau tempat tidur), maka harga cenderung lebih tinggi.
* (guests\_included dengan price): Korelasi positif antara guests\_included dan price (0.46) menunjukkan bahwa harga cenderung lebih tinggi untuk properti yang mencakup lebih banyak tamu dalam tarif dasar.

1. Hubungan Skor Ulasan dengan Variabel Lain

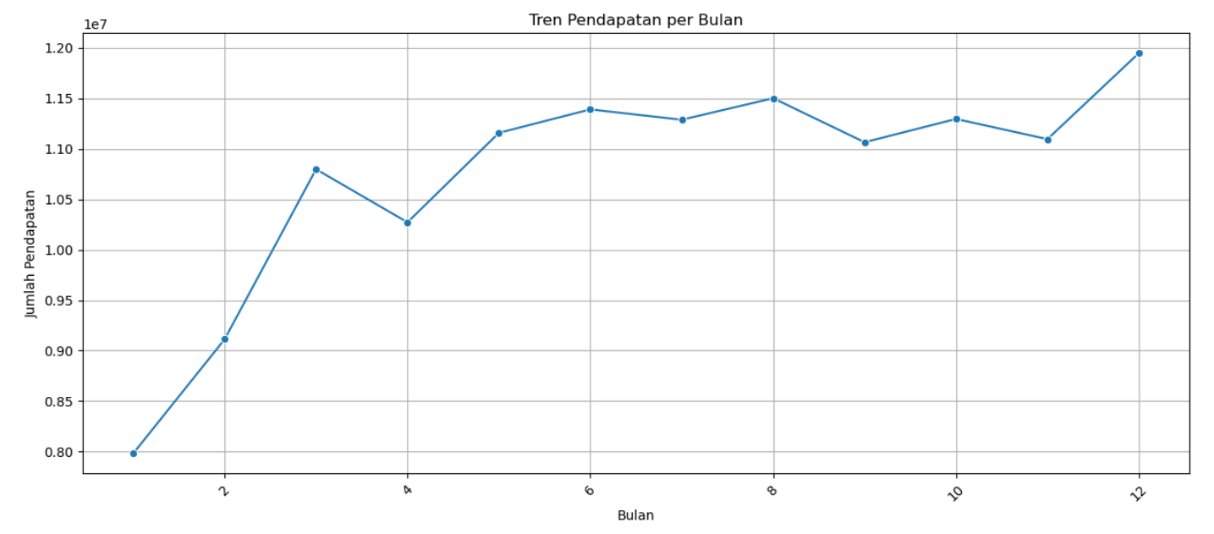
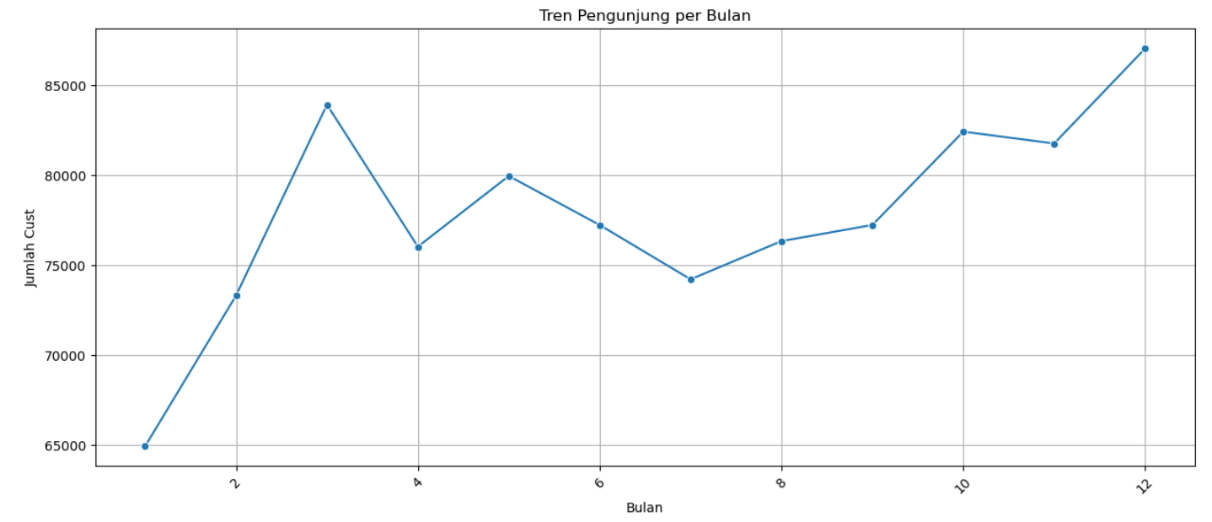
* (review\_scores\_location dengan review\_scores\_value (0.37)): Korelasi positif moderat menunjukkan bahwa properti yang memiliki nilai baik dalam location cenderung memiliki nilai tinggi juga dalam value.
* (review\_scores\_value dengan accommodates dan price (-0.07)): Korelasi negatif kecil menunjukkan bahwa nilai ulasan value cenderung sedikit menurun jika kapasitas atau harga meningkat, yang bisa diartikan bahwa tamu merasa harga yang lebih tinggi atau kapasitas lebih besar tidak selalu sepadan.

1. Korelasi Total Pendapatan dengan Harga Total Pendapatan dengan kelompok harga (price, weekly\_price, monthly\_price) memiliki korelasi yang tinggi (0.75,0.74,0.73). Hal ini menunjukkan kelompok harga mempengaruhi pendapatan, ketika harga dinaikkan maka pendapatan akan naik.

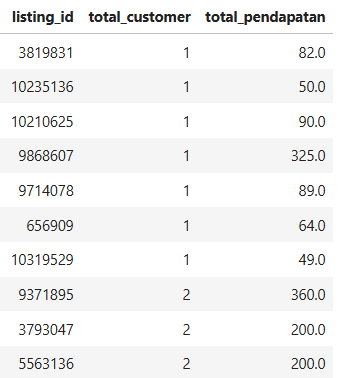
**Business Insight**

1. Jika kita melihat dua grafik dibawah, walaupun tren jumlah customer dan total pendapatan naik, tetapi di bulan Juni, terjadi keanehan antara dua tren tersebut, dimana jumlah customernya turun tetapi total pendapatannya naik. Kemungkinannya:

* Ada beberapa listing yang mempunyai nilai price yang lebih tinggi dan listing-listing tersebut mengalami kenaikan customer, sedangkan listing-listing yang punya nilai price yang rendah customernya mengalami penurunan melebihi listing yang dijelaskan sebelumnya.
* Ada beberapa customer yang menginap mingguan, sehingga mereka memilih harga weekly\_price yang lebih tinggi daripada price.
* Ada beberapa customer yang membayar biaya lain selain kelompok harga (price, weekly\_price, day\_price) dan hal itu melebih akumulasi nilai penurunan customer yang membayar nilai yang rendah (harga price saja di listing yang murah).



1. Listing yang mempunyai customer yang rendah, belum tentu mempunyai pendapatan yang rendah juga. Begitupun sebaliknya. Hal ini karena didasari oleh banyak faktor, seperti kelompok harga (price, weekly\_price, day\_price) yang berbeda di setiap listing, akomodasi dari listing tersebut yang mungkin bisa mempengaruhi kecenderungan orang memilih listing tersebut, biaya tambahan lainnya, dan banyak hal lain. Catatan: Dua tabel dibawah masih harus dibuang outliernya. Karena dari boxplot outlier dari total pendapatan tidak ada yang dibawah q1, tetapi listing yang hanya mendapatkan satu customer ada dibawah q1 (q1 nya adalah 139) jadi untuk membuangnya akan memakai



1. Mempunyai rating score yang tinggi belum tentu mempunyai pendapatan yang tinggi juga. Seperti padagrafik dibawah adalah top 10 listing dengan rating tertinggi dan yang kedua adalah top 10 listing dengan rating terendah. Pada tabel pertama, hanya dua listing dengan rating 100 yang pendapatannya melebihi q3 total pendapatan (44,2k dollar). Sedangkan pada tabel kedua, tidak ada yang menyentuh q3 total pendapatan. Perbdedaan lain dari dua tabel tersebut adalah rating yang tinggi memeiliki kecenderungan mempunyai customer yang lebih banyak dibandingkan rating yang randah. Hal ini bisa kita simpulkan sekilas bahwa rating rendah akan mempunyai kecenderungan pendapatan dan jumlah customer yang lebih sedikit. Oleh karena itu, listing yang mempunyai rating yang rendah harus menaikkan kualitas listing mereka.

