Report Progress

Stage 1

Unison

Muhammad Ath Thaariq Amir Chandra
Sutan Faisal
Farah Nelia Satriani
Hafizh Fakhri Alfarisy

Salsabila Nur Yasmin

Rakamin Academy

Unison

M Ath Thaariq A C

Data Analyst

Rakamin Academy

Sutan Faisal

Data Scientist

Rakamin Academy

Farah Melia S

Project Manager

Rakamin Academy

Hafizh Fakhri A

Data Engineer

Rakamin Academy

Stage 1

Dataset : Travel and Tourism Data

1. Data Quality Assesment

Pada dataset terdapat 3 macam CSV calender.csv, listing.csv, dan reviews.csv. Pada 3 macam CSV ini kami melakukan pencarian missing values, ouliers, dsb. Untuk mengetahui seberapa bersih dari isi data dari 3 macam csv tersebut . Didapatkan :

1. Calender.csv berisi 4 kolom mengenai lisitng, tanggal, avaibility unit dan harga :

```
Missing values in Calendar Dataset:
listing_id         0
date               0
available          0
price         459028
```

```
Inconsistent Dates in Calendar Dataset:
Empty DataFrame
Columns: [listing_id, date, available, price]
Index: []
```

a. Jumlah Data: 1.393.570 baris, 4 kolom
b. Missing Values: Kolom price memiliki 459.028 nilai yang hilang.
c. Potensi Inconsistent Entries: Kolom available berbentuk string, perlu dipastikan hanya memiliki nilai valid seperti 't' atau 'f'.
d. Potensi Outliers: Kolom price (jika berisi nilai numerik setelah pembersihan) bisa diperiksa lebih lanjut.

1. Data Quality Assesment

2. Listing.csv

```
Inconsistent Prices in Listings Dataset:
Empty DataFrame
Columns: [listing_id, listing_url, scrape_id, last_scr
Index: []
```

```
Missing values in Listings Dataset:
listing_id                          0
listing_url                         0
scrape_id                           0
last_scraped                        0
name                                0
                                   ...
cancellation_policy                 0
require_guest_profile_picture       0
require_guest_phone_verification    0
calculated_host_listings_count      0
reviews_per_month                 627
Length: 92, dtype: int64
```

a. Jumlah Data: 3.818 baris, 92 kolom
b. Missing Values: Beberapa kolom yang memiliki banyak nilai hilang:
c. square_feet (hanya 97 data yang tersedia, kemungkinan besar sebagian besar kosong).
d. license (semua data kosong).
e. Kolom harga (weekly_price, monthly_price, security_deposit, cleaning_fee) memiliki banyak nilai kosong.
f. Beberapa kolom review (review_scores_rating, review_scores_accuracy, dsb.) juga memiliki nilai yang hilang.
g. Potensi Inconsistent Entries: *Format zipcode (kode pos) bisa bervariasi dalam format.
h. Kolom host_response_rate dan host_acceptance_rate berbentuk string, perlu dikonversi ke numerik jika diperlukan.
i. Potensi Outliers:

*Kolom harga (price, weekly_price, monthly_price) perlu dicek distribusinya.
*Kolom minimum_nights dan maximum_nights bisa memiliki outliers ekstrem.

1. Data Quality Assesment

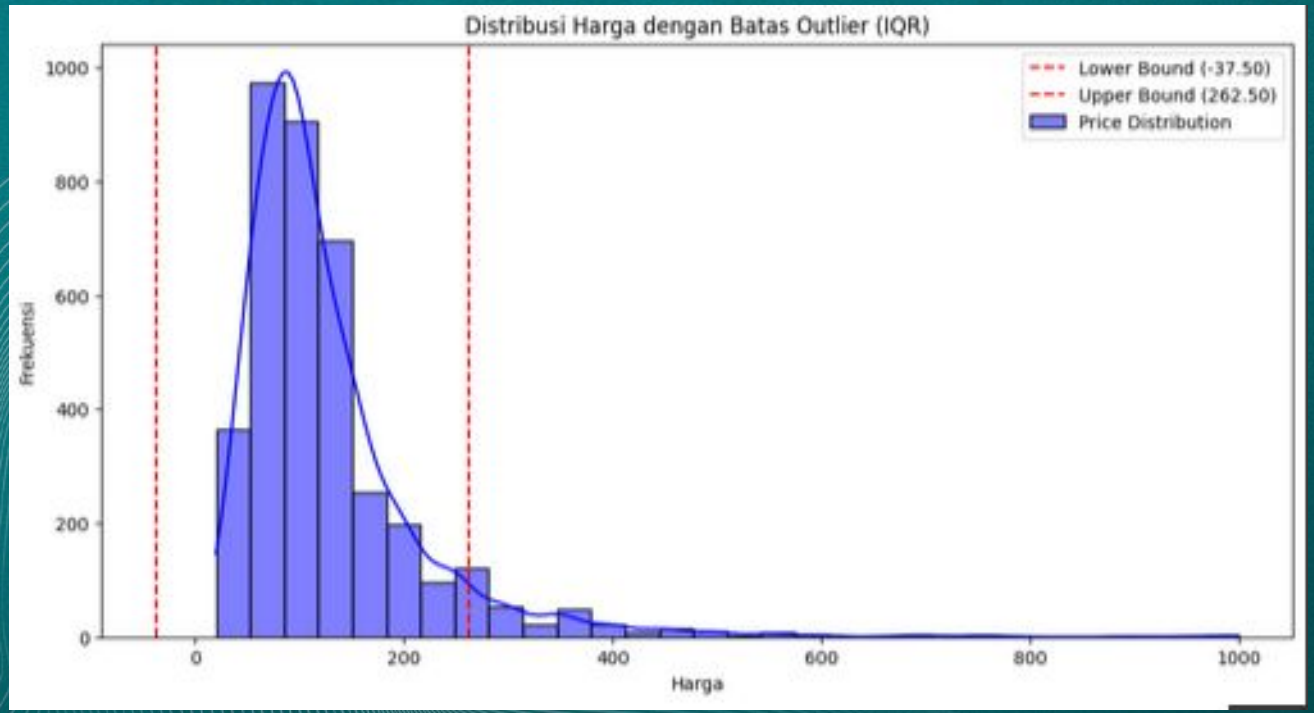
3. Reviews.csv

```
Missing values in Reviews Dataset:
listing_id        0
id                0
date              0
reviewer_id       0
reviewer_name     0
comments         18
dtype: int64
```

Dataset reviews.csv
a. Jumlah Data: 84.849 baris, 6 kolom
b. Missing Values: Kolom comments memiliki 18 nilai kosong.
c. Potensi Inconsistent Entries: Kolom date perlu dicek formatnya agar seragam.
d. Potensi Outliers: Tidak ada kolom numerik yang signifikan untuk outlier di dataset ini.

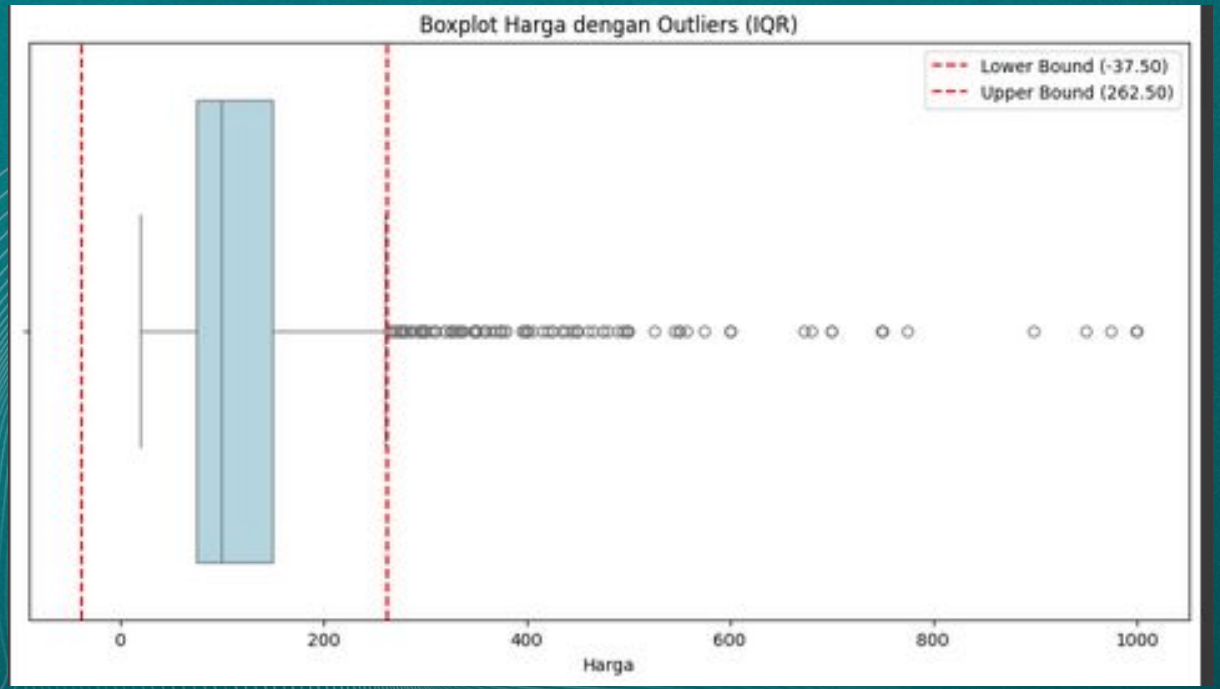
1. Data Quality Assesment

Visualisasi Outlier



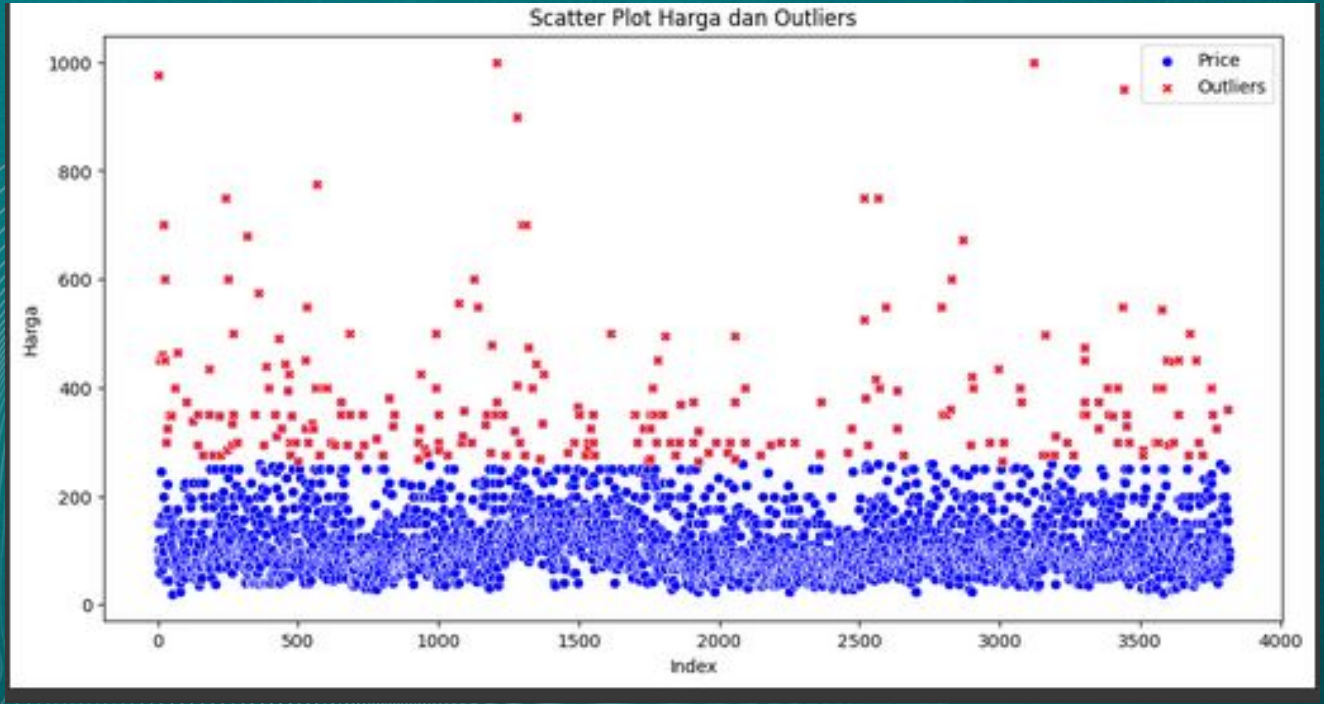
1. Data Quality Assesment

Visualisasi Outlier



1. Data Quality Assesment

Visualisasi Outlier



2. Data Cleaning

A. Handling missing values:

a. Menghapus simbol mata uang dan mengonversi kolom 'price' menjadi tipe data numerik

b. Mengisi missing values dengan median untuk kolom numerik (contoh kolom harga)

c. Mengisi missing values pada kolom kategorikal (misalnya, 'room_type')

d. Mengisi missing values pada kolom tanggal (misalnya, menggunakan '2022-01-01' sebagai placeholder)

B. Correcting inconsistent entries:

a. Mengonversi kolom 'date' ke format datetime yang konsisten

b. Menyamaratakan nilai kategorikal dalam 'room_type

Removing duplicates:

C. Menghapus duplikasi data di dalam dataset.

2. Data Cleaning

```
  calendar_df['date'].fillna('2022-01-01', inplace=True
Duplikasi pada calendar_df: 0
Duplikasi pada listings_df: 0
Duplikasi pada reviews_df: 0
Statistik setelah pembersihan calendar_df:
         listing_id                           date
count  1.393570e+06                        1393570
mean   5.550111e+06  2016-07-04 00:00:00.000000256
min    3.335000e+03             2016-01-04 00:00:00
25%    3.258213e+06             2016-04-04 00:00:00
50%    6.118244e+06             2016-07-04 00:00:00
75%    8.035212e+06             2016-10-03 00:00:00
max    1.034016e+07             2017-01-02 00:00:00
std    2.962274e+06                             NaN
```

2. Data Cleaning

```
Statistik setelah pembersihan listings_df:
         listing_id     scrape_id       host_id  host_listings_count  \
count  3.818000e+03  3.818000e+03  3.818000e+03          3816.000000
mean   5.550111e+06  2.016010e+13  1.578556e+07             7.157757
std    2.962660e+06  0.000000e+00  1.458382e+07            28.628149
min    3.335000e+03  2.016010e+13  4.193000e+03             1.000000
25%    3.258256e+06  2.016010e+13  3.275204e+06             1.000000
50%    6.118244e+06  2.016010e+13  1.055814e+07             1.000000
75%    8.035127e+06  2.016010e+13  2.590309e+07             3.000000
max    1.034016e+07  2.016010e+13  5.320861e+07           502.000000

       host_total_listings_count     latitude    longitude  accommodates
count                3816.000000  3818.000000  3818.000000   3818.000000
mean                    7.157757    47.628961  -122.333103      3.349398
std                    28.628149     0.043052     0.031745      1.977599
min                     1.000000    47.505088  -122.417219      1.000000
25%                     1.000000    47.609418  -122.354321      2.000000
50%                     1.000000    47.623601  -122.328874      3.000000
75%                     3.000000    47.662694  -122.310800      4.000000
max                   502.000000    47.733358  -122.240607     16.000000
```

2. Data Cleaning

```
Statistik setelah pembersihan reviews_df:
        listing_id            id                                date   reviewer_id
count  7.573000e+04  7.573000e+04                               75730  7.573000e+04
mean   3.022982e+06  3.075028e+07  2014-12-30 04:34:52.782252800  1.753278e+07
min    4.291000e+03  3.721000e+03            2009-06-07 00:00:00  1.500000e+01
25%    8.150170e+05  1.743744e+07            2014-08-12 00:00:00  5.395312e+06
50%    2.520890e+06  3.272610e+07            2015-05-20 00:00:00  1.493526e+07
75%    4.718921e+06  4.465204e+07            2015-08-28 00:00:00  2.838107e+07
max    1.024814e+07  5.873651e+07            2016-01-03 00:00:00  5.281274e+07
std    2.473498e+06  1.631777e+07                            NaN  1.362895e+07
```

2. Data Cleaning

```
Missing values setelah pembersihan calendar_df:
listing_id         0
date               0
available          0
price         459028
dtype: int64
Missing values setelah pembersihan listings_df:
listing_id                                0
listing_url                               0
scrape_id                                 0
last_scraped                              0
name                                      0
                                        ...
cancellation_policy                       0
require_guest_profile_picture             0
require_guest_phone_verification          0
calculated_host_listings_count            0
reviews_per_month                       627
Length: 92, dtype: int64
```

```
Missing values setelah pembersihan reviews_df:
listing_id        0
id                0
date              0
reviewer_id       0
reviewer_name     0
comments         17
dtype: int64
Duplikasi setelah pembersihan calendar_df: 0
Duplikasi setelah pembersihan listings_df: 0
Duplikasi setelah pembersihan reviews_df: 0
```

3. Data Integration

Berikut adalah langkah-langkah yang dapat Anda ikuti untuk melakukan Data Integration:
1. Identifikasi Kolom Key

* calendar.csv: Dataset ini kemungkinan memiliki kolom seperti listing_id yang menghubungkan data dengan listings.csv, serta date untuk menghubungkannya dengan data berdasarkan tanggal.

* listings.csv: Kolom listing_id di sini menjadi kunci utama yang menghubungkan dataset ini dengan dataset lainnya.

* reviews.csv: Dataset ini kemungkinan memiliki kolom listing_id yang menghubungkannya dengan listings.csv, serta date yang dapat digunakan untuk menghubungkan dengan calendar.csv.

3. Data Integration

2. Metode Penggabungan

Anda bisa menggunakan merge() dari Pandas untuk menggabungkan dataset berdasarkan kolom yang relevan. Di sini kita akan menggabungkan:

* listings_df dan calendar_df berdasarkan listing_id.

* listings_df dan reviews_df berdasarkan listing_id.

Penggabungan ini bisa dilakukan dengan menggunakan join kiri (left join) untuk memastikan bahwa semua data dari dataset utama (misalnya listings_df) tetap ada meskipun tidak ada kecocokan di dataset lain.

3. Data Integration

3. Periksa Konsistensi Format Data
Pastikan bahwa kolom yang digunakan untuk menggabungkan memiliki format yang konsisten di setiap dataset (misalnya, pastikan bahwa listing_id adalah angka atau string yang seragam).

4. Proses Penggabungan

Berikut adalah contoh kode untuk menggabungkan ketiga dataset tersebut:

4. Feature Engineer

Untuk Feature Engineering dengan menggunakan dataset calendar.csv, listings.csv, dan reviews.csv, kita dapat membuat beberapa fitur baru yang bisa meningkatkan kinerja model prediktif. Berikut adalah beberapa langkah yang bisa dilakukan untuk mengolah fitur baru dari data yang sudah ada.

Langkah-langkah:

1. Identifikasi Kolom yang Bisa Dibuat Fitur Baru Dalam ketiga dataset tersebut, ada beberapa kolom yang dapat digunakan untuk membuat fitur baru, seperti:
    - Price di listings.csv: Dapat digunakan untuk membuat fitur baru seperti log price (untuk mengurangi distribusi harga yang skewed).
    - Date di calendar.csv: Bisa digunakan untuk membuat fitur baru seperti month atau weekday, yang mungkin berguna untuk menganalisis pola musiman atau minggu.
    - Review_scores_rating di reviews.csv: Bisa digunakan untuk membuat fitur seperti average review score per listing.

4. Feature Engineer

2. Feature Engineering yang Bisa Dilakukan Berikut adalah beberapa contoh fitur baru yang bisa dibuat:
a. Logarithmic Transformation pada price Untuk mengurangi skewness pada data harga, kita dapat menggunakan transformasi logaritmik pada harga.
b. Membuat Fitur month dan weekday dari date Dari kolom date di calendar.csv, kita bisa membuat dua fitur baru: bulan (month) dan hari d# Mengambil bulan dan hari dalam minggu dari kolom 'date' di calendar_df
c. Menambahkan Fitur price_per_night Jika ada informasi tentang jumlah malam yang tersedia (misalnya, di listings_df), kita bisa menambahkan fitur baru yang mengukur harga per malam. Jika tidak ada, kita bisa langsung menggunakan price sebagai pengganti.
d. Pengelompokan (Binning) Berdasarkan Harga Kita dapat mengelompokkan harga menjadi beberapa kategori (misalnya, murah, sedang, mahal) untuk analisis lebih lanjut. e. Feature dari Review (Jumlah Ulasan) Di dataset reviews.csv, kita bisa menambahkan fitur baru yang menunjukkan jumlah ulasan untuk setiap listing.
f. Rata-rata Skor Ulasan Kita bisa menghitung rata-rata skor review dari setiap listing untuk fitur baru.
g. Fitur Kategorikal (One-Hot Encoding) Beberapa kolom kategorikal seperti room_type dapat diubah menjadi fitur biner (one-hot encoding) untuk digunakan dalam model prediktif.
h. Fitur Hari Libur Kita bisa membuat fitur baru yang menunjukkan apakah suatu tanggal adalah hari libur atau bukan. Misalnya, kita bisa membuat fitur baru is_holiday di calendar.csv berdasarkan data hari libur.

4. Feature Engineer

```
Listing Data (Head):
   listing_id                          listing_url     scrape_id
0      241032   https://www.airbnb.com/rooms/241032  2.016010e+13
1      953595   https://www.airbnb.com/rooms/953595  2.016010e+13
2     3308979  https://www.airbnb.com/rooms/3308979  2.016010e+13
3     7421966  https://www.airbnb.com/rooms/7421966  2.016010e+13
4      278830   https://www.airbnb.com/rooms/278830  2.016010e+13

  last_scraped                                  name  \
0     1/4/2016          Stylish Queen Anne Apartment
1     1/4/2016   Bright & Airy Queen Anne Apartment
2     1/4/2016  New Modern House-Amazing water view
3     1/4/2016                   Queen Anne Chateau
4     1/4/2016       Charming craftsman 3 bdm house

                                             summary  \
0                                                NaN
1  Chemically sensitive? We've removed the irrita...
2  New modern house built in 2013.  Spectacular s...
3  A charming apartment that sits atop Queen Anne...
4  Cozy family craftman house in beautiful neighb...
```

4. Feature Engineer

```
                                                space  \
0  Make your self at home in this charming one-be...
1  Beautiful, hypoallergenic apartment in an extr...
2  Our house is modern, light and fresh with a wa...
3                                                NaN
4  Cozy family craftman house in beautiful neighb...

                                          description experiences_offered  \
0  Make your self at home in this charming one-be...                 none
1  Chemically sensitive? We've removed the irrita...                 none
2  New modern house built in 2013.  Spectacular s...                 none
3  A charming apartment that sits atop Queen Anne...                 none
4  Cozy family craftman house in beautiful neighb...                 none

                             neighborhood_overview  ... price_bin  \
0                                              NaN  ...     Murah
1  Queen Anne is a wonderful, truly functional vi...  ...    Sedang
2  Upper Queen Anne is a charming neighborhood fu...  ...     Mahal
3                                              NaN  ...    Sedang
4  We are in the beautiful neighborhood of Queen ...  ...    Sedang
```

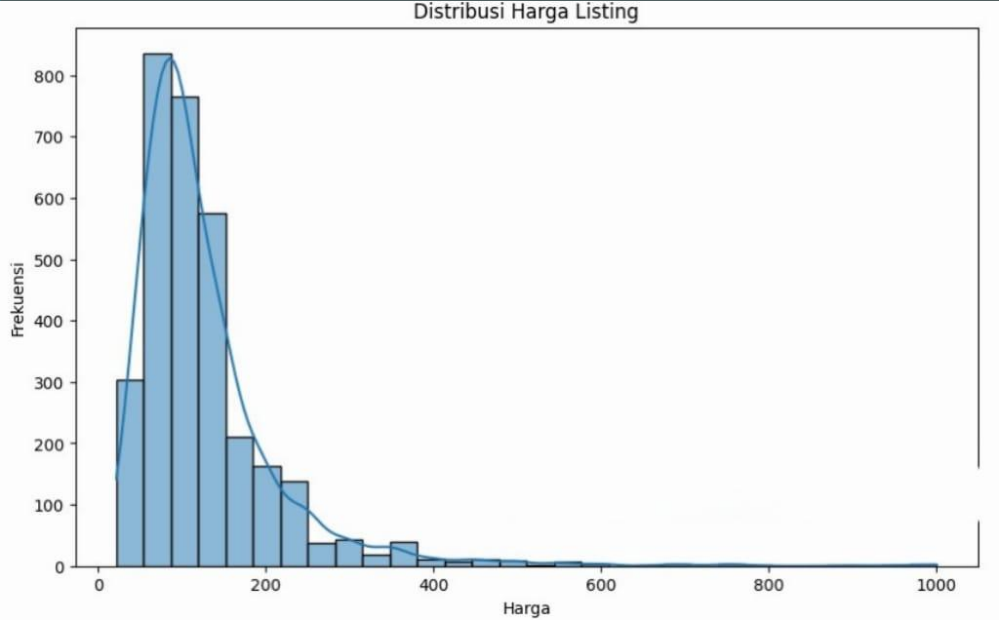
4. Feature Engineer

```
   review_count room_type_Private room room_type_Shared room  \
0         207.0                  False                 False
1          43.0                  False                 False
2          20.0                  False                 False
3           NaN                  False                 False
4          38.0                  False                 False

   review_count_reviews_count room_type_Private room room_type_Shared room  \
0                       207.0                  False                 False
1                        43.0                  False                 False
2                        20.0                  False                 False
3                         NaN                  False                 False
4                        38.0                  False                 False

   review_count_reviews_count room_type_Private room room_type_Shared room
0                       207.0                  False                 False
1                        43.0                  False                 False
2                        20.0                  False                 False
3                         NaN                  False                 False
4                        38.0                  False                 False

[5 rows x 104 columns]
```

4. Feature Engineer

```
Calendar Data (Head):
   listing_id        date available   price  month  weekday  is_holiday
0      241032  2016-01-04         t  $85.00      1        0           0
1      241032  2016-01-05         t  $85.00      1        1           0
2      241032  2016-01-06         f     NaN      1        2           0
3      241032  2016-01-07         f     NaN      1        3           0
4      241032  2016-01-08         f     NaN      1        4           0
   listing_id  review_count
0      241032         207.0
1      953595          43.0
2     3308979          20.0
3     7421966           NaN
4      278830          38.0
```

5. Data Visualisation

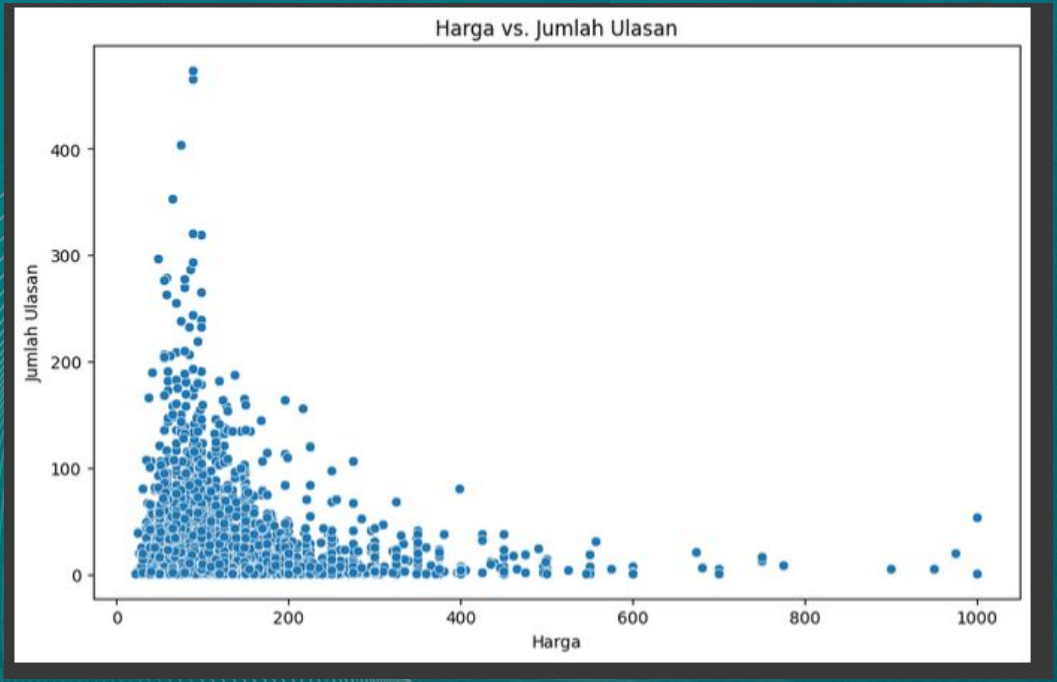


Data Insight :
Dari data yang ada listing yang tersedia memiliki variasi harga berkisar 20 hingga 1000. Menandakan adanya ketersediaan hunian berdasarkan budget yang dimiliki

Harga yang tersedia di platform AirBNB memiliki range harga terbanyak di kelas harga Ekonomis ke Menengah

Dari sini kita bisa memberikan pertimbangan harga apabila rerata harga yang ditawarkan dirange tersebut adalah normal.

5. Data Visualisation



Sebaran Data:

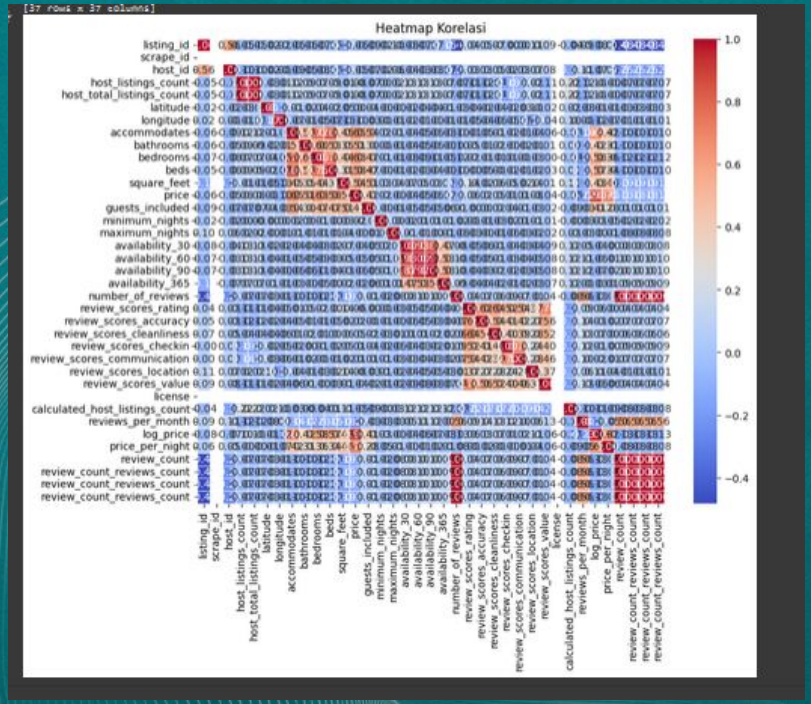
a. Properti dengan harga rendah tampaknya lebih banyak ulasan
b. Properti yang lebih mahal lebih sedikit mendapatkan ulasan tapi ada beberapa pengecualian

Korelasi :

a. Jika pola menunjukan tren naik artinya properti dengan harga lebih tinggi lebih popo[uler
b. Jika pola menyebar secara acak maka harga tidak terpengaruh langsung terhadap jumlah ulasan

Dari grafik ini kita bisa menyimpulkan bahwa harga properti tidak menjadi faktor utama dalam menentukan jumlah ulasan dan ada faktor lain seperti lokasi atau fasilitas yang lebih berpengaruh..

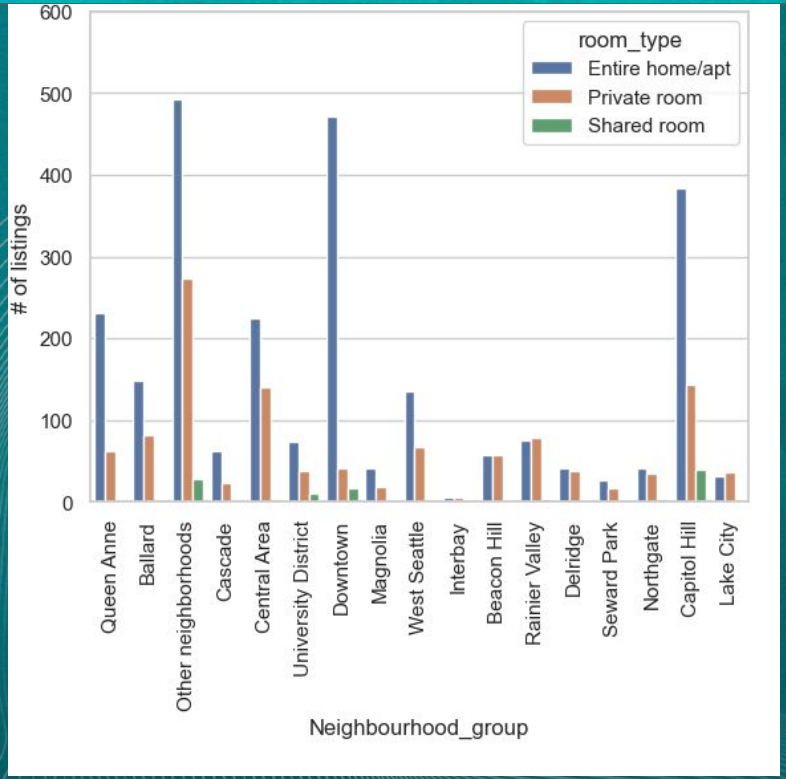
5. Data Visualisation



Insight :
Hubungan antara harga dan jumlah ulasan bisa dilihat apakah memiliki korelasi signifikan atau tidak
Jika harga memilii korelasi rendah terhadap jumlah ulasan ini menunjukan bahwa faktor lain(lokasi, fasilitas,kebijakan pemilik) lebih menentukan banyak ulasan

Identifikasi Korelasi yang signifikan :
Variabel dengan korelasi tinggi (>0.5 atau <-0.5) mungkin memiliki hubungan yang erat dan bisa dieksplorasi lebih lanjut.
Variabel yangkorelasi mendekati nol berarti tidak ada hubungan yang kuat.

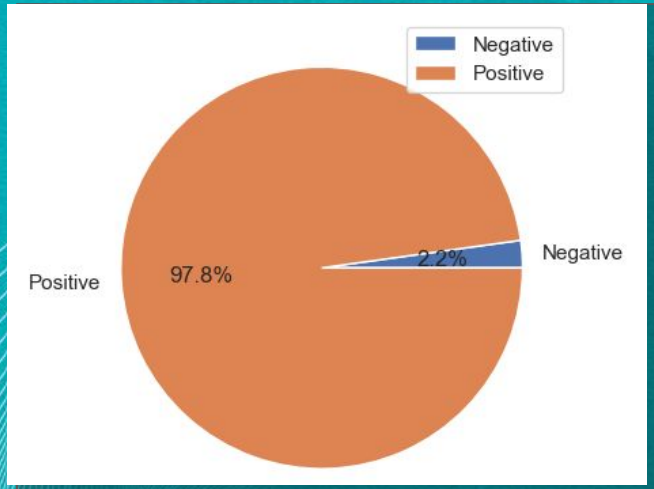
5. Data Visualisation

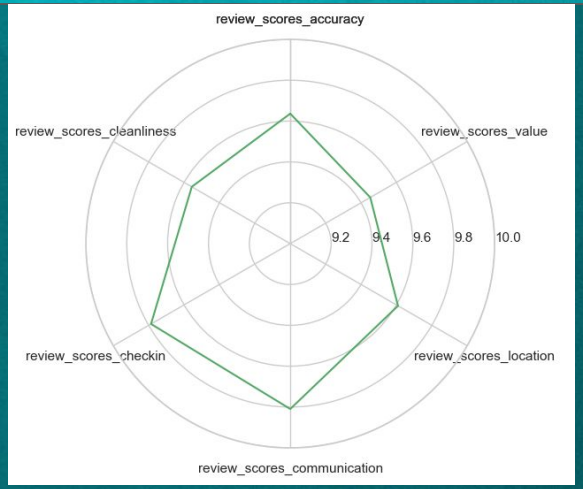


Sebagian besar pelanggan memilih tipe kamar **Entire Home/Apt**, yang menunjukkan preferensi mereka terhadap hunian yang lebih nyaman dan privat saat menginap.

Lokasi juga menjadi faktor penting dalam pemilihan hunian. Beberapa area yang paling diminati oleh pelanggan adalah **Downtown, Capitol Hill, Queen Anne, dan Central Area**, yang menawarkan akses mudah ke berbagai fasilitas dan destinasi populer.

5. Data Visualisation

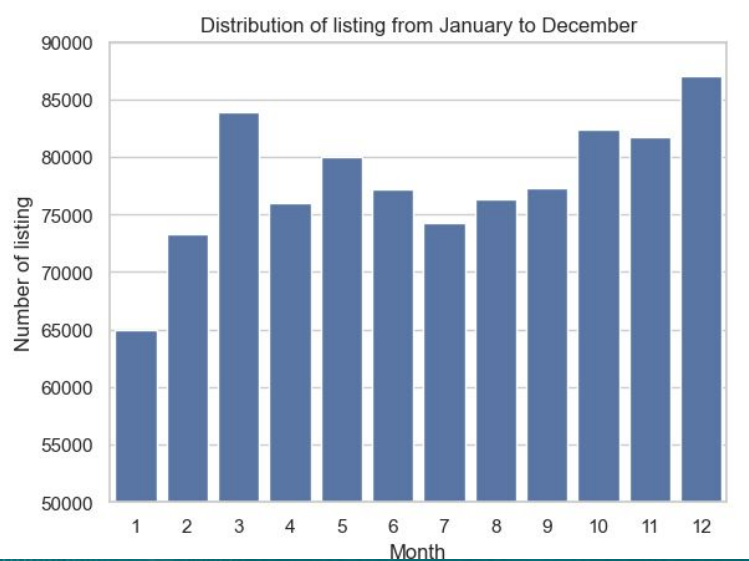




Berdasarkan Score review 97,8 % pelanggan memberikan review terhadap hunian yang disewa sangat positif, Hanya terdapat 2,2 % yang memiliki review negatif

Berdasarkan radar chart pelanggan memberikan review bagus mengenai aspek Komunikasi, Checkin, dan Value. Dimana menggambarkan respon dan proses host menerima tamu sangat baik dan harga sewa yang dianggap sepadan dengan fasilitas yang diberikan

5. Data Visualisation



Januari memiliki jumlah listing terendah. Berbeda dengan musim panas, hal ini bisa disebabkan oleh faktor musiman seperti **cuaca dingin, rendahnya permintaan perjalanan, serta berkurangnya jumlah listing baru**. Selain itu, banyak pemilik properti yang mungkin menonaktifkan listing mereka sementara waktu setelah musim liburan berakhir.

Periode liburan musim panas (Juni, Juli, Agustus) memiliki lebih sedikit listing yang tersedia dibandingkan bulan lainnya. Hal ini kemungkinan besar disebabkan oleh banyaknya pelanggan yang telah melakukan reservasi jauh-jauh hari, sehingga tingkat hunian menjadi tinggi.

Pada periode **Oktober hingga Desember,** jumlah listing mengalami peningkatan signifikan, dengan puncaknya terjadi di bulan Desember. Peningkatan ini kemungkinan besar dipengaruhi oleh musim liburan akhir tahun, seperti Thanksgiving, Natal, dan Tahun Baru, yang mendorong lebih banyak pemilik properti untuk membuka listing mereka guna memenuhi tingginya permintaan

Link Colab :
https://colab.research.google.com/drive/1V2T7I2Fk-NHf4bIke114CddKP5mJ3DSW?usp=sharing

Link Dataset :
https://drive.google.com/drive/folders/1kykWeF39WxU8stu5u1uG3TdOjcmDMvpx?usp=sharing


Thank you!
Raih Mimpi #TanpaBatas