Cahya Wimar W

Agenda

- Background, Business Problem, Key Questions & Goal Settings
- Executive Summary & Recommendation
- Project Framework
  - Data Pre Processing
  - Data Processing
- Output Data Pre Processing: Dataset knowledge & data clean plan
- Output Data Processing:
  - Identifikasi listings belum optimal
  - Insights karakteristik skema penawaran pada listing yang belum optimal
  - Insights karakteristik visitor pada listing yang belum optimal
- Tableau Data Visualization

airbnb

Background

Background

Tentang:
Airbnb merupakan platform markeplace yang menghubungkan pemilik properti (host) dengan visitor. Platofrom tersebut memampukan visitor untuk mencari dan memilih properti dengan keunikanya masing-masing.

Model Bisnis:
Airbnb mendapatkan revenue dari fee hosting dan fee booking. Fee hosting didapatkan dari host yang propertinya tersewa oleh visitor dengan besaran 3% dari transaksi booking. Divisi Supply merupakan pihak yang bertanggung jawab pada proses pengelolaan hosting tersebut.

Airbnb 2024 Strategy:
Salah satu pilar strategi Airbnb 2024 yaitu “Making Hosting Mainstream” yang fokus pada penciptaan persepsi di kalangan pemilik properti (host) bahwa hosting di Airbnb adalah hal yang menyenangkan dan mudah. Objective dari pilar strategi ini adalah Promote Hosting, Highlights the Benefits dan Support for Host.

Strategic Initiative of Property Supply Dept:
Upaya Property Supply Dept untuk merealisasikan strategi tersebut adalah dengan melakukan revenue optimization melalui ‘program targeted ads’ pada listing yang revenue nya belum optimal.

Business Problem, Key Questions & Goal Settings

Business Problem

Property Supply Dept perlu melakukan analisis untuk mengidentifikasi karakteristik listing yang revenue nya belum optimal. Sehingga strategi ‘program targeted ads’ dapat terlaksana dengan lebih efektif dan terarah.

NB:
`remaining_potential_revenue` adalah metriks yang digunakan untuk mengukur revenue listing yang belum optimal dengan cara mengkalikan `price` dengan `availability_365`

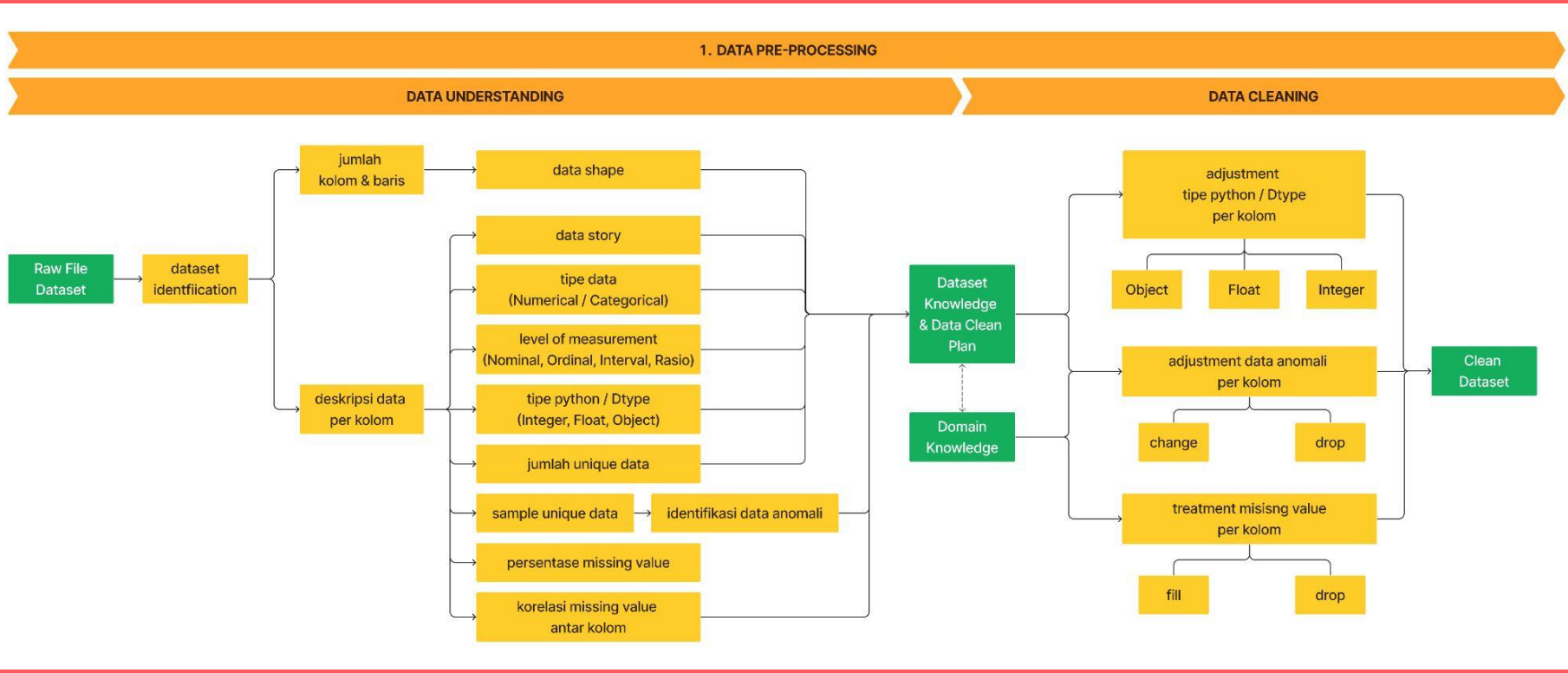
Key Questions & Goal Settings

- Berapa nilai minimal `remaining_potential_revenue` yang digunakan untuk identifikasi listing belum optimal dan optimal? -> Menggunakan parameter pada statistik deskriptif untuk menentukan nilai minimal `remaining_potential_revenue`
- Berapa dan apa saja listings yang belum optimal? -> Identifikasi jumlah listing belum optimal
- Bagaimana karakteristik listings yang belum optimal berdasarkan skema penawaran? -> EDA untuk generate insights karakteristik skema penawaran pada listing yang belum optimal berdasarkan `room_type` & `minimum_nights`
- Bagaimana karakteristik listings yang belum optimal berdasarkan visitor engagement? -> EDA untuk generate insi karakteristik visitor engagement pada listing yang belum optimal berdasarkan `number_of_review` & `review_per_month`

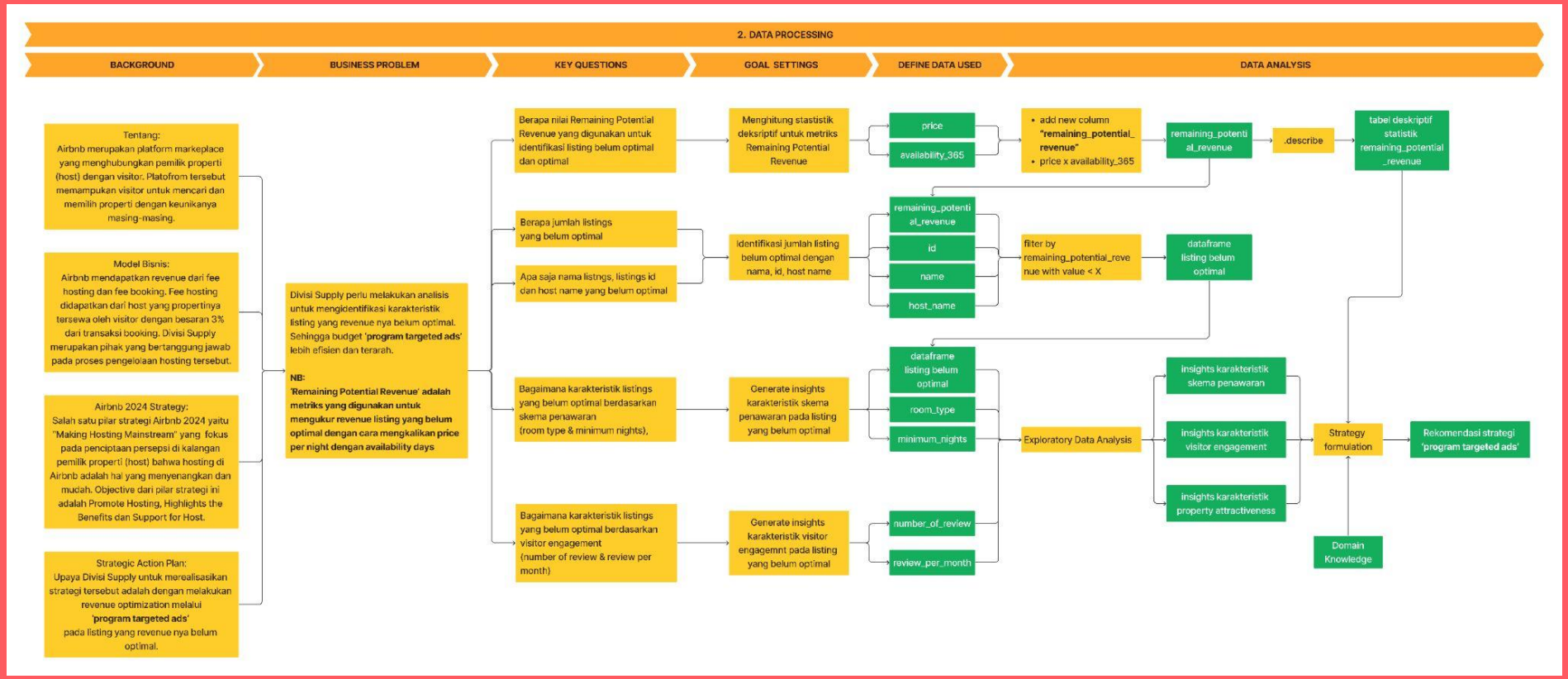
Executive Summary & Recommendation

- Project analisis dataset listings Airbnb di Bangkok, Thailand untuk memberikan insights kepada Divisi Supply dalam merencanakan strategi ‘Program Targeted Ads’
- Raw dataset terdiri dari 16 kolom dan 15.937 baris, setelah pre-processing data menjadi 14.883 baris.
- Data pre processing menghasilkan 5 ‘dataset knowledge & data clean plan’
- Pada kolom ‘neighbourhood’ data anomali 99.52% dan tidak adanya cara lain untuk mengisi data tersebut, sehingga tidak bisa digunakan. Saran untuk tim product untuk melakukan improvement pada fitur listing data entry
- Terdapat 7.434 listing masuk kategori belum optimal berdasarkan cut-off remaining potential revenue sebesar 1.151.434.647
- Tipe *entire home/apt* dan *private room* adalah tipe yang paling banyak pada listings yang belum optimal dengan jumlah 6.796 listings atau 91% dimana nilai remaining potential revenue sebesar 1.069.647.277. Sehingga sebaiknya ‘program targeted ads’ dapat fokus untuk menawarkan tipe *entire home/apt* dan *private room*.
- Untuk ads *entire home/apt* sebaiknya menargetkan visitor dengan budget penginapan pada range 333 s.d 00.000 per malam. Sedangkan untuk *private room* range nya sekitar 278 s.d 161.516 per malam.
- Pada tipe room *entire home/apt* dan *private room* sama-sama memiliki jumlah minimum sewa terbesar pada kategori 1 malam dan 2-7 malam. Sehingga sebaiknya target visitor adalah mereka yang cenderung menghabiskan waktu liburan <1 minggu.
- Berdasarkan listings yang memiliki nilai anomali pada kolom `reviews_per_month` dan `number_of_reviews` terlihat bahwa mayoritas visitor lebih tertarik pada listings dengan jumlah minimum sewa pada kategori 1 sd 7 malam.

Project Framework: Data Pre Processing



Project Framework: Data Processing



Output Data Pre Processing

Dataset knowledge & data clean plan 1:

Dataset merupakan data listing property yang disewakan pada platform Airbnb di Bangkok, Thailand. Data tersebut merupakan hasil dari inputan keterangan tentang property yang disewakan oleh host dengan rincian sebagai berikut:

- `no` -> nomor baris -> **numerical** -> **integer**
- `id` -> unique id untuk listing Airbnb -> **numerical** -> **integer**
- `name` -> nama listing -> **categorical** -> **object**
- `host_id` -> unique id untuk host Airbnb -> **numerical** -> **integer**
- `host_name` -> nama host (property owner) -> **categorical** -> **object**
- `neighbourhood` -> lingkungan di-geocode menggunakan latitude dan longitude berdasarkan data shapefile publik atau terbuka -> **categorical** -> **object**
- `latitude` -> proyeksi World Geodetic System (WGS84) untuk latitude dan longitude -> **numerical** -> **float**
- `longtitude` -> proyeksi World Geodetic System (WGS84) untuk latitude dan longitude -> **numerical** -> **float**
- `room_type` -> tipe kamar -> **categorical** -> **object**
- `price` -> harga sewa per malam -> **numerical** -> **float**
- `minimum_nights` -> jumlah malam minimum untuk menginap -> **numerical** -> **integer**
- `number_of_reviews` -> jumlah review oleh visitor -> **numerical** -> **integer**
- `last_review` -> tanggal review terbaru -> **date/time** -> **datetime64[ns]**
- `reviews_per_month` -> rata-rata review per bulan -> **numerical** -> **float**
- `calculated_host_listings_count` -> jumlah listing yang dimiliki host di kota/region saat ini -> **numerical** -> **integer**
- `availability_365` -> ketersediaan listing dalam x hari ke depan max 365 hari. (karena telah dipesan atau diblokir oleh host) -> **numerical** -> **integer**
- `number_of_reviews_ltm` -> jumlah review oleh visitor dalam 12 bulan terakhir -> **numerical** -> **integer**

Output Data Pre Processing

Dataset knowledge & data clean plan 2:

- Jika dilihat sekilas, terdapat baris data duplikat pada kolom `name` `host_id` `host_name` `neighbourhood`, kemungkinan host mengupload property nya lebih dari 1x pada platform Airbnb. Sehingga penanganan untuk data cleaning adalah **menghapus baris data duplikat berdasarkan duplikasi data pada kolom** `name` `host_id` `host_name` `neighbourhood`.

Dataset knowledge & data clean plan 3:

- Pada kolom `latitude` dan `longtitude` memiliki jumlah digit yang berbeda-beda. Sedangkan format digit lat-long sebaiknya adalah Decimal Degree. Sehingga penanganan untuk data cleaning adalah **mengubah format** `latitude` **dan** `longtitude` **menjadi decimal degree**.

Dataset knowledge & data clean plan 4:

- Berdasarkan hasil identifikasi missing value, kolom `last_review` dan `reviews_per_month` dengan persentase missing value terbesar yaitu 36% atau 5840-an baris data. Sedangkan sisanya memiliki persentase missing value yang sangat kecil yaitu berkisar <1% sd 1,1% atau 16-182 baris data.
- Berdasarkan heatmapping missing value terlihat bahwa missing value pada kolom `last_review` dan `review_per_month` berada pada baris yang sama.
- Jumlah missing value pada kolom `last_review`: 5841 dan `reviews_per_month` : 5842
- Kemudian jumlah data pada kolom `number_of_reviews` yang bernilai 0.0 adalah 5732
- Dapat disimpulkan bahwa missing value pada kolom `last_review` dan `reviews_per_month` diakibatkan karena `number_of_reviews` bernilai 0 atau memang tidak ada review pada listings tersebut.
- **Sehingga penanganan missing value pada kolom** `last_review` **dan** `reviews_per_month` **akan di isi dengan (no review), namun hanya sebatas pada** `number_of_reviews` **yang bernilai 0 saja**

Output Data Pre Processing

Dataset knowledge & data clean plan 5:

- `no` -> terdapat 15939 unique data dari total 15939 data -> tidak menunjukan adanya data anomali
- `id` -> terdapat 12607 unique data dari total 15939 data -> terlihat ada data dengan jumlah digit 5, 6 dan 7, walaupun cukup aneh karena jumlah digit listing id pada umumnya seragam namun pada analisis kali ini tidak dilakukan intervensi data
- `name` -> terdapat 14859 unique data dari total 15939 data -> terlihat ada beberapa data dengan format penulisan nama yang tidak rapi dan terkesan penjelasan/deskripsi fasilitas. pada analisis ini tidak akan dilakukan intervensi data, namun sebagai saran untuk tim product agar dapat menambahkan field `property_description` dan setting policy format input `name` dengan font yang lebih rapi
- `host_id` -> terdapat 6632 unique data dari total 15939 data -> data terlihat normal dengan 6 digit dan 7 digit sehingga tidak dilakukan intervensi data
- `host_name` -> terdapat 5347 unique data dari total 15939 data -> data terlihat normal sehingga tidak dilakukan intervensi data
- `neighbourhood` -> terdapat 126 unique data dari total 15939 data -> terlihat ada sebagian data anomali berupa data longtitude, **sehingga perlu di teliti lebih lanjut pada data cleaning untuk jumlah data anomali nya, jika >10% maka kolom ini tidak akan digunakan untuk analisis karena tidak ada opsi untuk mengisi data tersebut**
- `latitude` -> terdapat 9555 unique data dari total 15939 data -> data masih memiliki format bilangan desimal, namun secara keseluruhan berada pada garis lintang 13derajat. **sehingga penanganan data cleaning berupa select 7 digit dari depan dan konversi ke decimal degree**
- `longtitude` -> terdapat 10222 unique data dari total 15939 data -> data masih memiliki format bilangan desimal, namun secara keseluruhan berada pada garis bujur 100derajat. **sehingga penanganan data cleaning berupa select 8 digit dari depan dan konversi ke decimal degree**
- `room_type` -> terdapat 22 unique data dari total 15939 data -> seharus nya hanya 4 uniqe data, **sehingga perlu di teliti lebih lanjut pada data cleaning untuk jumlah data anomali nya dan korelasi dengan kolom lainya untuk intervensi data**
- `price` -> terdapat 3046 unique data dari total 15939 data -> tidak menunjukan adanya data anomali
- `minimum_nights` -> terdapat 120 unique data dari total 15939 data -> terlihat ada 2 jenis data anomali. Pertama jumlah minimum hari >30 hari (1bulan) yang terlihat kurang masuk akal karena satuan sewa harga per malam. Kedua terdapat data dengan format tanggal (dd/mm/yyyy). **sehingga penanganan data cleaning akan menghitung jumlah data anomali serta opsi untuk intervensi data melalui mengganti data dengan nilai median**
- `number_of_reviews` -> terdapat 318 unique data dari total 15939 data -> tidak menunjukan adanya data anomali
- `last_review` -> terdapat 1676 unique data dari total 15939 data -> tidak menunjukan adanya data anomali
- `reviews_per_month` -> terdapat 554 unique data dari total 15939 data -> hanya terlihat missing value dan tidak menunjukan adanya data anomali
- `calculated_host_listings_count` -> terdapat 52 unique data dari total 15939 data -> terlihat beberapa data yang aneh berupa data >10 listing per host. jika dilihat dari total 15939 listing, `host_name` nya 5347 artinya asumsi rasio 1 host 3 listing **sehingga penangangan data cleaning perlu melihat pengelompokan `id` berdasarkan `host_name` dan menghitung jumlah data anomali serta opsi untuk intervensi data melalui mengganti data dengan nilai median**
- `availability_365` -> terdapat 366 unique data dari total 15939 data -> tidak menunjukan adanya data anomali
- `number_of_reviews_ltm` -> terdapat 85 unique data dari total 15939 data -> tidak menunjukan adanya data anomali

Identifikasi Listings Belum Optimal

2.4.2 Identifikasi jumlah listing belum optimal

```python
# Membuat dataframe baru dengan nilai 'remaining_potential_revenue' di bawah median
no_optimal_df = clean_df[clean_df['remaining_potential_revenue'] < median_value]

# Menambahkan kolom 'status' dengan nilai "no optimal" jika 'remaining_potential_revenue' di bawah median
no_optimal_df['status'] = np.where(no_optimal_df['remaining_potential_revenue'] < median_value, 'no optimal', 'optimal')

# Menampilkan hasil
display(no_optimal_df.head())
```

Python

	no	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	room_type_code	...	number_of_reviews
2	2	28745	modern-style apartment in Bangkok	123784	Familyroom	Bang Kapi	1.375232	1.006240	Private room	2	...	0
9	9	952677	Standard Room Decor do Hostel	5171292	Somsak	Khlong San	0.137204	1.005076	Private room	2	...	4
23	23	1808600	Contemporary Modern Duplex-Thong Lo	9478184	Shine	Khlong Toei	1.372097	1.005782	Entire home/apt	1	...	83
30	30	156583	Studio near Chula University/Silom walk to MRT/BTS	58920	Gael	Bang Rak	0.137285	1.005231	Entire home/apt	1	...	63

Insights karakteristik skema penawaran pada listing yang belum optimal

Menggunakan asumsi bahwa salah satu faktor yang menentukan preferensi visitor untuk memilih listings property adalah jenis ruangan yang ditawarkan serta batasan minimum malam diwajibkan oleh host. Hal tersebut masuk akal karena visitor perlu mencari property yang sesuai dengan kebutuhan ruangan serta jangka waktu liburan/menginap nya.

```python
# Mengelompokkan dan menghitung agregat remaining_potential_revenue
agg_functions = {
    'remaining_potential_revenue': ['count','sum', 'min', 'median', 'max'],
    'price': ['count','sum', 'min', 'median', 'max']
}

# Membuat dataframe baru dengan agregat
room_type_summary = no_optimal_df.groupby('room_type').agg(agg_functions).reset_index()

# Mengurutkan berdasarkan sum remaining_potential_revenue secara descending
room_type_summary_sorted = room_type_summary.sort_values(by=('remaining_potential_revenue', 'sum'), ascending=False)

# Menampilkan hasil
display(room_type_summary_sorted.head())
```

Python

	room_type	remaining_potential_revenue					price				
		count	sum	min	median	max	count	sum	min	median	max
0	Entire home/apt	4111	633944848.0	0.0	160500.0	313000.0	4111	5837506.0	332.0	1085.0	20000.0
2	Private room	2685	435702429.0	0.0	170325.0	313040.0	2685	3077063.0	278.0	800.0	161516.0
3	Shared room	383	49822290.0	0.0	128115.0	298749.0	383	206318.0	280.0	450.0	4800.0
1	Hotel room	255	31965080.0	0.0	126448.0	310250.0	255	506505.0	0.0	955.0	23629.0

Insights karakteristik skema penawaran pada listing yang belum optimal

Menggunakan asumsi bahwa salah satu faktor yang menentukan preferensi visitor untuk memilih listings property adalah jenis ruangan yang ditawarkan serta batasan minimum malam diwajibkan oleh host. Hal tersebut masuk akal karena visitor perlu mencari property yang sesuai dengan kebutuhan ruangan serta jangka waktu liburan/menginap nya.

```python
# Mengelompokkan dan menghitung agregat remaining_potential_revenue
agg_functions = {
    'remaining_potential_revenue': ['count','sum', 'min', 'median', 'max'],
    'price': ['count','sum', 'min', 'median', 'max']
}

# Membuat dataframe baru dengan agregat
room_type_summary = no_optimal_df.groupby('room_type').agg(agg_functions).reset_index()

# Mengurutkan berdasarkan sum remaining_potential_revenue secara descending
room_type_summary_sorted = room_type_summary.sort_values(by=('remaining_potential_revenue', 'sum'), ascending=False)

# Menampilkan hasil
display(room_type_summary_sorted.head())
```

	room_type	remaining_potential_revenue					price				
		count	sum	min	median	max	count	sum	min	median	max
0	Entire home/apt	4111	633944848.0	0.0	160500.0	313000.0	4111	5837506.0	332.0	1085.0	20000.0
2	Private room	2685	435702429.0	0.0	170325.0	313040.0	2685	3077063.0	278.0	800.0	161516.0
3	Shared room	383	49822290.0	0.0	128115.0	298749.0	383	206318.0	280.0	450.0	4800.0
1	Hotel room	255	31965080.0	0.0	126448.0	310250.0	255	506505.0	0.0	955.0	23629.0

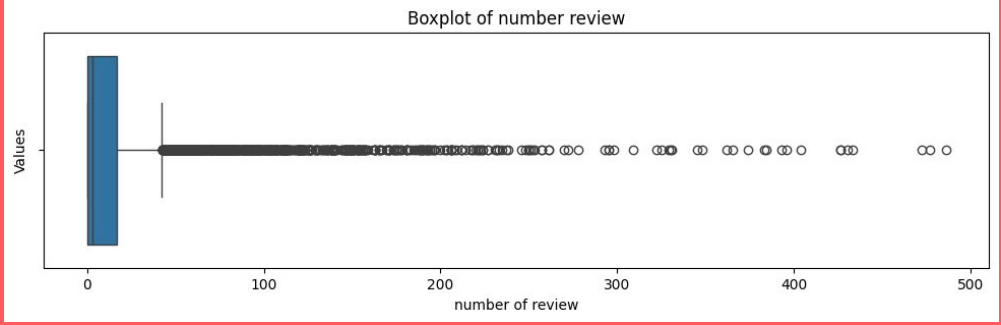
```python
# Membuat crosstab antara room_type dan minimum_nights_code
room_type_cross_min_nights = pd.crosstab(no_optimal_df['room_type'], no_optimal_df['minimum_nights_code'], margins=True, margins_name=

# Mengurutkan berdasarkan total secara descending
room_type_cross_min_nights_sorted = room_type_cross_min_nights.sort_values(by=('Total'), ascending=False)

# Mengatur urutan kolom sesuai urutan yang diinginkan
column_order = ['1', '2-7', '8-14', '15-30', '>30', 'Total']
room_type_cross_min_nights_sorted = room_type_cross_min_nights_sorted[column_order]

# Menampilkan hasil
display(room_type_cross_min_nights_sorted.head())
```

Python

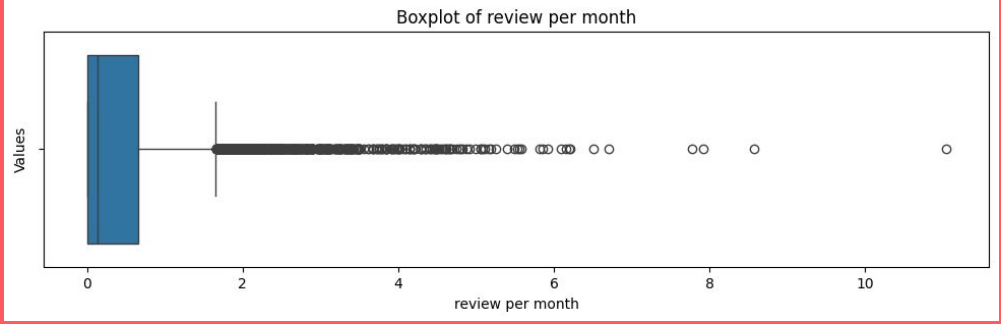
minimum_nights_code	1	2-7	8-14	15-30	>30	Total
room_type						
Total	3880	1923	218	779	634	7434
Entire home/apt	1476	1309	178	601	547	4111
Private room	1816	575	37	176	81	2685
Shared room	353	26	3	1	0	383
Hotel room	235	13	0	1	6	255

Insights karakteristik visitor pada listing yang belum optimal

Menggunakan asumsi bahwa listings dengan nilai anomali pada `number_of_reviews` dan `reviews_per_month` akan memberikan kita insights terkait dengan listings seperti apa yang di minati oleh visitor.




```
nilai minimal number_of_reviews adalah 0.0
nilai median number_of_reviews adalah 3.0
nilai maximal number_of_reviews adalah 17.0
nilai lower fence number_of_reviews adalah -25.5
nilai upper fence number_of_reviews adalah 42.5
```




```
nilai minimal reviews_per_month adalah 0.0
nilai median reviews_per_month adalah 0.13
nilai maximal reviews_per_month adalah 0.66
nilai lower fence reviews_per_month adalah -0.99
nilai upper fence reviews_per_month adalah 1.65
```

```python
# Filter berdasarkan kolom 'reviews_per_month' >1.65 dan 'number_of_reviews' >42.5
review_anomalous_df = no_optimal_df[(no_optimal_df['reviews_per_month'] > 1.65) & (no_optimal_df['number_of_reviews'] > 42.5)]

# Menampilkan hasil
display(review_anomalous_df.head())
```

Python

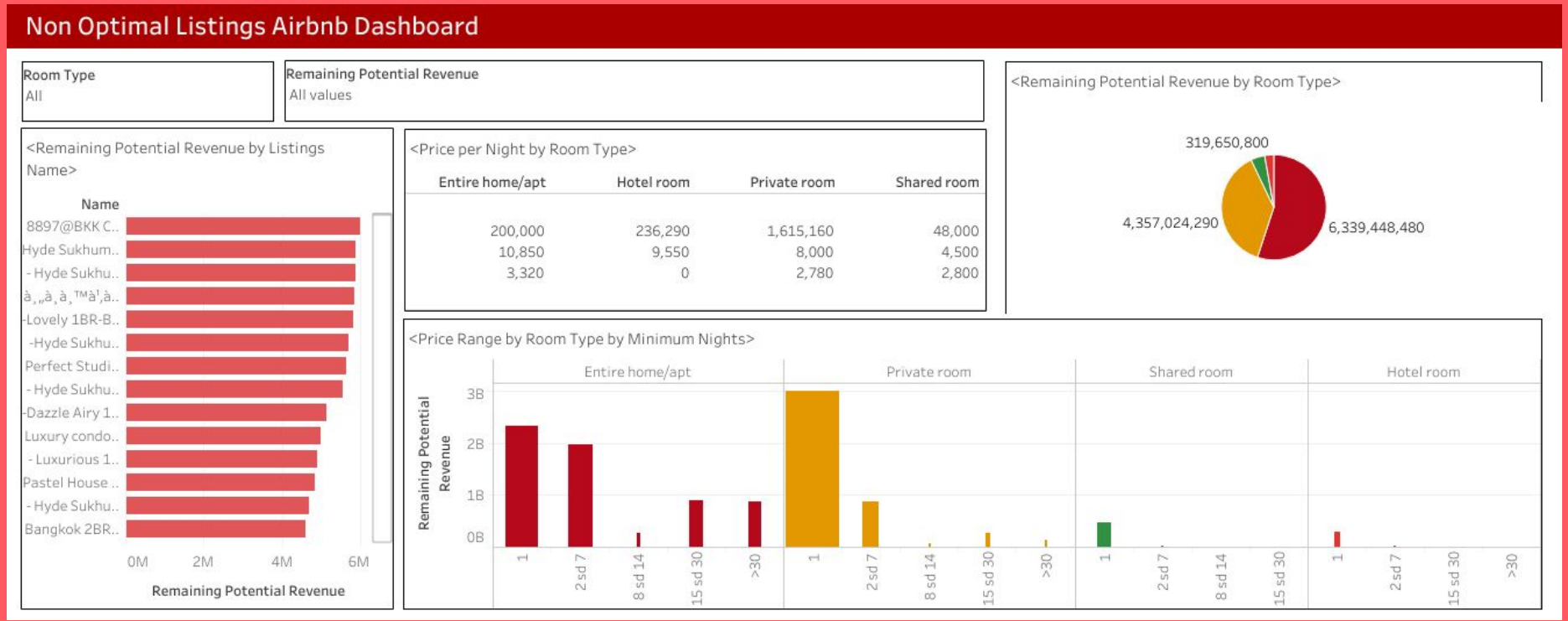
	no	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	room_type_code	...	number_of_revi
53	53	1026451	♡Chic Studio, Easy Walk to Pier & BTS Taksin♡	3346331	Bee	Sathon	1.371192	1.005154	Entire home/apt	1	...	
61	61	1041976	Long-stay special rate spacious entire floor Siam	5735895	Pariya	Parthum Wan	1.374814	1.005202	Entire home/apt	1	...	
72	72	1943048	Best nr Chatujak, MRT, BTS free wifi&fNetflix	9906827	Nokiko	Chatu Chak	1.381694	1.005645	Entire home/apt	1	...	
73	73	385130	✪Citycenter✓Subway station✓Private Bathroom4Aircon	1927968	Evan	Sathon	1.372062	1.005471	Entire home/apt	1	...	
80	80	393066	❀❀99 feet in the sky❀❀	1927968	Evan	Sathon	1.372062	1.005471	Entire home/apt	1	...	

```python
# Membuat crosstab antara room_type dan minimum_nights_code pada listings yang review nya anomali
room_type_cross_min_nights_anomali = pd.crosstab(review_anomalous_df['room_type'], review_anomalous_df['minimum_nights_code'], margins

# Mengurutkan berdasarkan total secara descending
room_type_cross_min_nights_anomali_sorted = room_type_cross_min_nights_anomali.sort_values(by=('Total'), ascending=False)

# Mengatur urutan kolom sesuai urutan yang diinginkan
column_order = ['1', '2-7', '8-14', '15-30', '>30', 'Total']
room_type_cross_min_nights_anomali_sorted = room_type_cross_min_nights_anomali_sorted[column_order]

# Menampilkan hasil
display(room_type_cross_min_nights_anomali_sorted.head())
```

Python

minimum_nights_code	1	2-7	8-14	15-30	>30	Total
room_type						
Total	209	89	4	23	19	344
Entire home/apt	135	70	4	23	17	249
Private room	57	17	0	0	1	75
Hotel room	16	1	0	0	1	18
Shared room	1	1	0	0	0	2

Tableau Data Visualization



https://public.tableau.com/views/Capstone_Project2_Airbnb/Dashboard1?:language=en-GB&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

airbnb

Thank You