Analisis Data airbnb New York City

by

Kelompok 12

Rafi Iqbal Rendy S. (3.34.23.3.19) & Tsabitah Hilyatul Aulia (3.34.23.3.23)

Introduction

Analisis Dataset Airbnb New York City

Notebook ini disusun sebagai bagian dari **Ujian Tengah Semester** mata kuliah **Kecerdasan Buatan**.

Analisis ini memanfaatkan dataset **Airbnb New York City**, yang berisi informasi seputar properti sewaan jangka pendek, mencakup:

- Lokasi geografis properti
- Harga sewa
- Jenis kamar
- Wilayah administratif

Tujuan Analisis

Analisis ini bertujuan untuk:

- 📍 **Menelusuri persebaran listing Airbnb** berdasarkan lokasi geografis
- 💲 **Menganalisis pola harga sewa** di berbagai wilayah kota
- 🏠 **Menggambarkan karakteristik umum** dari properti yang tersedia di Airbnb New York City

Import library

```
# Import library numpy sebagai np, digunakan untuk operasi numerik seperti array, aljabar linear, dan statistik
import numpy as np

# Import library pandas sebagai pd, digunakan untuk manipulasi dan analisis data dalam bentuk tabel (DataFrame)
import pandas as pd

# Import library matplotlib (inti) dan modul pyplot sebagai plt, digunakan untuk membuat grafik statis seperti line plot, bar chart, dll
import matplotlib
import matplotlib.pyplot as plt

# Import library seaborn sebagai sns, digunakan untuk visualisasi statistik yang lebih menarik dan informatif, berbasis matplotlib
import seaborn as sns

# Import modul plotly.express sebagai px, digunakan untuk membuat visualisasi data interaktif dengan sintaks yang sederhana
import plotly.express as px


df_airbnb = pd.read_csv('./airbnb.csv')
df_airbnb.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nigh
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	

```
# menampilkan informasi dasar dataset
df_airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

Load and Explore the Data

```
df_airbnb = pd.read_csv('./airbnb.csv')
df_airbnb.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nigh
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	

```
# menampilkan informasi dasar dataset
df_airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

```
df_airbnb.shape
```

```
(48895, 16)
```

```
# menampilkan data statistik dataset
df_airbnb.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	cal
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	

```
# cek missing values
df_airbnb.isnull().sum()
```

```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
price_scaled                      0
price_capped                      0
dtype: int64
```

```
# Cek baris yang duplikat
df_airbnb.duplicated()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
48890    False
48891    False
48892    False
48893    False
```

```
48894    False
Length: 48895, dtype: bool
```

Data Cleaning and Preprocessing

Handle Missing Values

```
# Mengisi kolom 'name' dan 'host_name' yang hilang dengan 'Unknown'
df_airbnb['name'] = df_airbnb['name'].fillna('Unknown')
df_airbnb['host_name'] = df_airbnb['host_name'].fillna('Unknown')

# Mengisi 'reviews_per_month' dengan 0, karena tidak ada review
df_airbnb['reviews_per_month'] = df_airbnb['reviews_per_month'].fillna(0)

# Menghapus baris yang hilang pada kolom 'last_review'
df_airbnb = df_airbnb.dropna(subset=['last_review'])
```

```
# cek missing values
df_airbnb.isnull().sum()
```

```
id                              0
name                            0
host_id                         0
host_name                       0
neighbourhood_group             0
neighbourhood                   0
latitude                        0
longitude                       0
room_type                       0
price                           0
minimum_nights                  0
number_of_reviews               0
last_review                     0
reviews_per_month               0
calculated_host_listings_count  0
availability_365                0
dtype: int64
```

Analysis

1.Harga Termurah Untuk Listing Di Manhattan

```
# Filter hanya Manhattan
manhattan_listings = df_airbnb[df_airbnb['neighbourhood_group'] == 'Manhattan']

# Ambil 5 listing termurah (harga > 0 untuk menghindari error listing gratis)
top5_cheapest = manhattan_listings[manhattan_listings['price'] > 0].sort_values(by='price').head(5)

top5_cheapest[['name', 'neighbourhood', 'price', 'latitude', 'longitude']]
```

	name	neighbourhood	price	latitude	longitude
31407	Cozy feel at home studio	Kips Bay	10	40.74408	-73.97803
24100	Girls only, cozy room one block from Times Square	Hell's Kitchen	10	40.75812	-73.98935
22287	Jen Apt	SoHo	10	40.72237	-73.99817
31066	Very Spacious bedroom, steps from CENTRAL PARK.	Upper West Side	10	40.76844	-73.98333
23256	Quiet, Cozy UES Studio Near the Subway	Upper East Side	10	40.76844	-73.95341

2.Plot semua listing di Manhattan + tandai Top 5 listing termurah

```
fig = px.scatter_map(
    combined_df,
    lat="latitude",
    lon="longitude",
    color="kategori",
    hover_name="name",
    hover_data={"price": True, "neighbourhood": True},
    zoom=11,
    height=600,
```

```
    size=combined_df['price'].apply(lambda x: 15 if x in top5_cheapest['price'].values else 5),
    size_max=15,
    color_discrete_map={'Semua Listing': 'lightgray', 'Top 5 Termurah': 'red'}
)

fig.update_layout(
    mapbox_style="carto-positron",  # gaya peta minimalis dan clean
    title="Peta Lokasi 5 Listing Termurah di Manhattan",
    title_x=0.5,
    legend_title="Kategori",
    margin={"r": 0, "t": 40, "l": 0, "b": 0}
)

fig.show()
```

Performa Host & Kepemilikan Properti

3.Top 10 Host dengan Jumlah Listing Terbanyak

```
# Ambil top 10 host berdasarkan jumlah listing
top_hosts = df_airbnb['host_id'].value_counts().head(10).reset_index()
top_hosts.columns = ['host_id', 'jumlah_listing']

# Gabungkan dengan df_airbnb untuk mengambil nama host
host_names = df_airbnb[['host_id', 'host_name']].drop_duplicates()

# Merge untuk dapatkan nama host
top_hosts = top_hosts.merge(host_names, on='host_id', how='left')

# Tampilkan hasil
print(top_hosts)
```

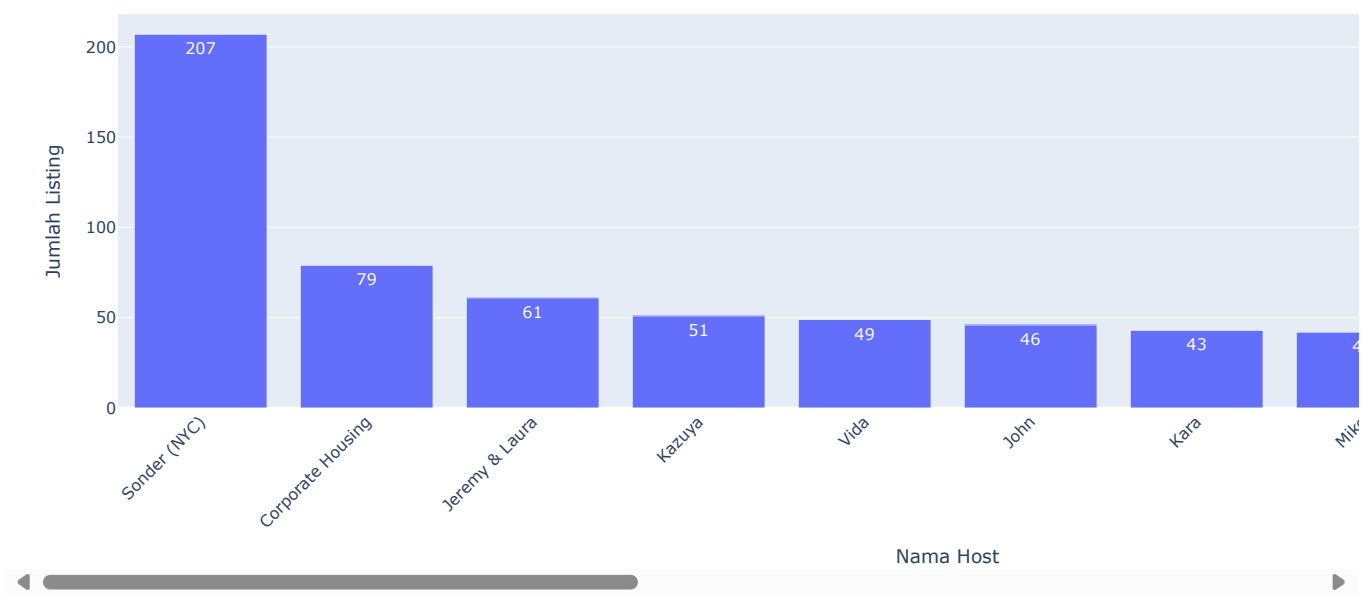
```
      host_id  jumlah_listing         host_name
0  219517861             207      Sonder (NYC)
1   61391963              79  Corporate Housing
2   16098958              61    Jeremy & Laura
3  137358866              51            Kazuya
4    7503643              49              Vida
5  190921808              46              John
6   30283594              43              Kara
7    1475015              42              Mike
8  120762452              40           Stanley
9    2119276              39              Host
```

4.Top 10 Host dengan Jumlah Listing Terbanyak - VISUALISASI

```
fig = px.bar(
    top_hosts,
    x='host_name',
    y='jumlah_listing',
    title='Top 10 Host dengan Listing Terbanyak di NYC',
    text_auto=True,
    labels={'host_name': 'Nama Host', 'jumlah_listing': 'Jumlah Listing'},
    hover_data=['host_id']
)

fig.update_layout(
    xaxis_tickangle=-45,
    title_x=0.5
)

fig.show()
```



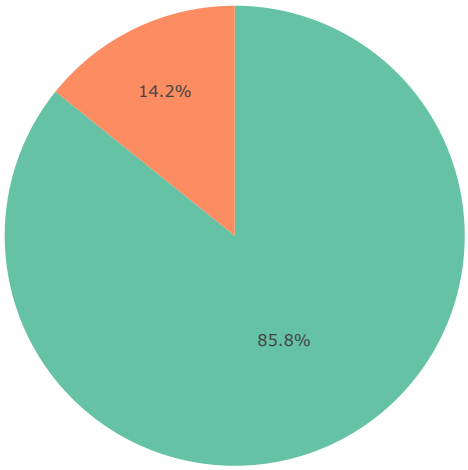
5.Single Property vs Multiple Property

```
host_counts = df_airbnb.groupby("host_id").size().reset_index(name="jumlah")
host_counts["host_type"] = host_counts["jumlah"].apply(lambda x: "Multiple" if x > 1 else "Single")
summary = host_counts["host_type"].value_counts().reset_index()
summary.columns = ["host_type", "count"]
summary
```

	host_type	count
0	Single	25949
1	Multiple	4302

6. Visualisasi host yang memiliki Single Property vs Multiple Property

```
fig = px.pie(
    summary,
    values='count',
    names='host_type',
    title='Proporsi Host: Single vs Multiple Property',
    color_discrete_sequence=px.colors.qualitative.Set2
)
fig.show()
```



Kepadatan Listing per Kecamatan (Neighbourhood)

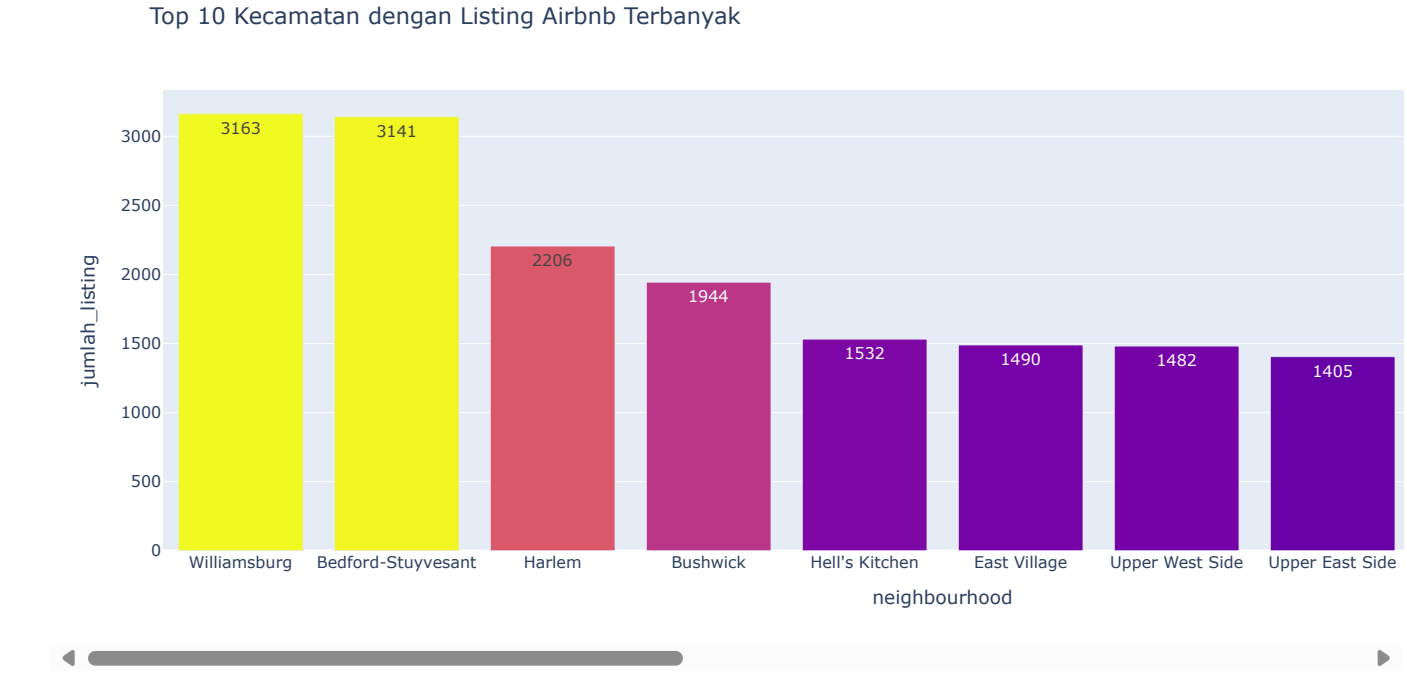
7.Neighbourhood dengan Listing Terbanyak

```
top_neigh = df_airbnb['neighbourhood'].value_counts().head(10).reset_index()
top_neigh.columns = ['neighbourhood', 'jumlah_listing']
top_neigh
```

	neighbourhood	jumlah_listing
0	Williamsburg	3163
1	Bedford-Stuyvesant	3141
2	Harlem	2206
3	Bushwick	1944
4	Hell's Kitchen	1532
5	East Village	1490
6	Upper West Side	1482
7	Upper East Side	1405
8	Crown Heights	1265
9	Midtown	986

8.Neighbourhood dengan Listing Terbanyak- VISUALISASI

```
fig = px.bar(
    top_neigh,
    x='neighbourhood',
    y='jumlah_listing',
    title='Top 10 Kecamatan dengan Listing Airbnb Terbanyak',
    text_auto=True,
    color='jumlah_listing'
)
fig.show()
```



9.Persebaran Lokasi Properti Airbnb di New York City

```
# Membatasi harga maksimal ke 500 agar warna lebih terlihat variasinya
df_airbnb["price_capped"] = df_airbnb["price"].clip(upper=500)

fig = px.scatter_map(
    df_airbnb,
    lat="latitude",
    lon="longitude",
    hover_name="name",
    hover_data=["price", "neighbourhood_group"],
    color="price_capped",  # Gunakan price yang sudah dibatasi
    color_continuous_scale="Viridis",
    size_max=10,
    zoom=10
)

fig.update_layout(
    title="Peta Persebaran Properti Airbnb di New York City",
    title_x=0.5,
    mapbox_style="carto-positron"
)

fig.show()
```

Peta Persebaran Properti Airbnb di New York City

Dominasi Tipe Kamar di Setiap Borough

10. Tipe Kamar Paling Umum per Borough

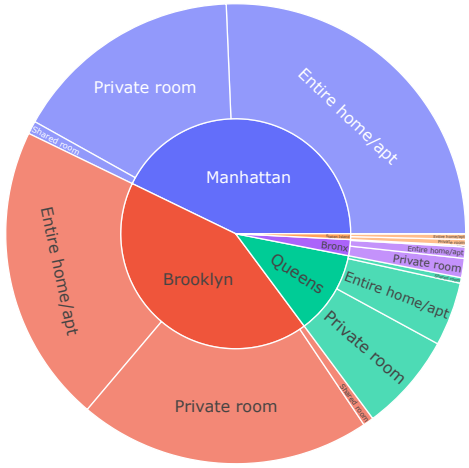
```
dominant_room = df_airbnb.groupby(['neighbourhood_group', 'room_type']).size().reset_index(name='count')
dominant_room_sorted = dominant_room.sort_values(['neighbourhood_group','count'], ascending=[True, False])
dominant_room_sorted.groupby('neighbourhood_group').first().reset_index()
```

	neighbourhood_group	room_type	count
0	Bronx	Private room	524
1	Brooklyn	Entire home/apt	8164
2	Manhattan	Entire home/apt	9967
3	Queens	Private room	2680
4	Staten Island	Private room	159

11.Tipe Kamar Paling Umum per Borough -VISUALISASI

```
fig = px.sunburst(
    dominant_room,
    path=['neighbourhood_group', 'room_type'],
    values='count',
    title='Dominasi Tipe Kamar per Borough'
)
fig.show()
```

Dominasi Tipe Kamar per Borough



Karakteristik Minimum Malam Menginap di Tiap Wilayah NYC

12.Tabel Rata-rata Minimum Nights per Wilayah

```
# Hapus outlier (anggap 30 hari ke atas adalah untuk keperluan jangka panjang)
df_min_nights = df_airbnb[df_airbnb['minimum_nights'] <= 30]

# Hitung rata-rata minimum nights per wilayah
avg_min_nights = df_min_nights.groupby('neighbourhood_group')['minimum_nights'].mean().reset_index()
avg_min_nights.columns = ['Wilayah', 'Rata-rata Minimum Nights']

# Tampilkan tabel
avg_min_nights.style.set_caption("Rata-rata Minimum Malam Menginap per Wilayah NYC")
```

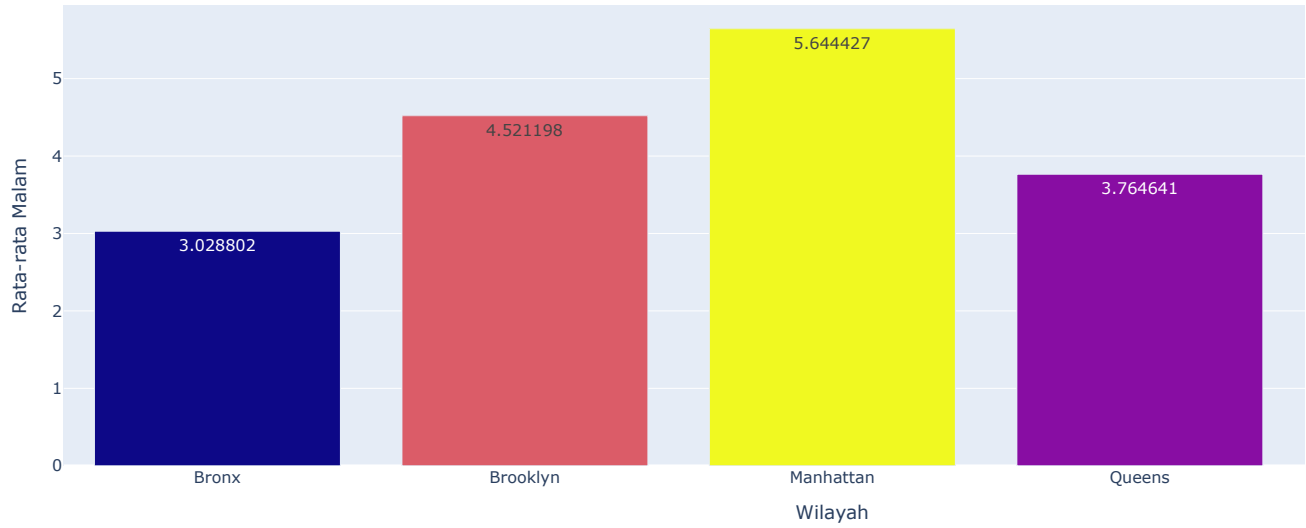
Rata-rata Minimum Malam Menginap per Wilayah NYC

	Wilayah	Rata-rata Minimum Nights
0	Bronx	3.028802
1	Brooklyn	4.521198
2	Manhattan	5.644427
3	Queens	3.764641
4	Staten Island	2.225870

13.Visualisasi Rata-rata Minimum Nights per Wilayah

```
fig = px.bar(
    avg_min_nights,
    x='Wilayah',
    y='Rata-rata Minimum Nights',
    title='Rata-rata Minimum Malam Menginap di Tiap Wilayah NYC',
    text_auto=True,
    labels={'Wilayah': 'Wilayah', 'Rata-rata Minimum Nights': 'Rata-rata Malam'},
    color='Rata-rata Minimum Nights',
    color_continuous_scale='Plasma'
)

fig.update_layout(title_x=0.5)
fig.show()
```



Rata-rata Review per Bulan

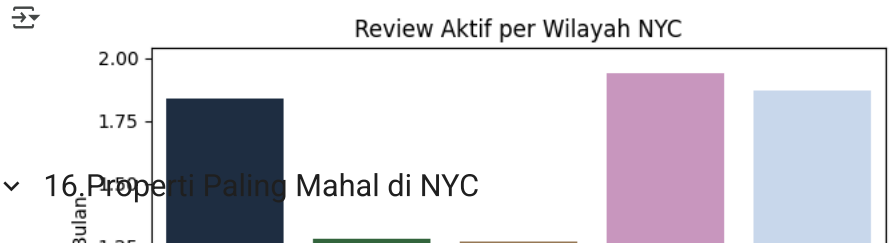
14. tabel rata-rata Review per Bulan

```
active_reviews = df_airbnb.groupby('neighbourhood_group')['reviews_per_month'].mean().reset_index()
active_reviews.columns = ['Wilayah', 'Rata-rata Review/Bulan']
active_reviews
```

	Wilayah	Rata-rata Review/Bulan
0	Bronx	1.837831
1	Brooklyn	1.283212
2	Manhattan	1.272131
3	Queens	1.941200
4	Staten Island	1.872580

15. rata rata review per Bulan- VISUALISASI

```
sns.barplot(data=active_reviews, x='Wilayah', y='Rata-rata Review/Bulan', hue='Wilayah', palette='cubehelix', legend=False)
plt.title('Review Aktif per Wilayah NYC')
plt.tight_layout()
plt.show()
```



16. Properti Paling Mahal di NYC

```
fig = px.bar(
    top_price,
    x='name',
    y='price',
    color='neighbourhood_group',
    title='10 Listing Termahal di NYC',
    text='price'
)
fig.update_layout(xaxis_tickangle=45, title_x=0.5)
fig.show()
```

17. Tempat dengan Review Terbaik

```
# Filter listing dengan jumlah review tinggi (misal >100 review)
top_reviews = df_airbnb[df_airbnb['number_of_reviews'] > 100]

# Ambil 10 listing dengan jumlah review terbanyak
top_reviews_sorted = top_reviews.sort_values('number_of_reviews', ascending=False).head(10)

# Tampilkan kolom penting
top_reviews_sorted[['name', 'host_name', 'neighbourhood_group', 'room_type', 'number_of_reviews', 'price']]
```

	name	host_name	neighbourhood_group	room_type	number_of_reviews	price
11759	Room near JFK Queen Bed	Dona	Queens	Private room	629	47
2031	Great Bedroom in Manhattan	Jj	Manhattan	Private room	607	49
2030	Beautiful Bedroom in Manhattan	Jj	Manhattan	Private room	597	49
2015	Private Bedroom in Manhattan	Jj	Manhattan	Private room	594	49