Data Analytics

Case Study NYC AirBNB

Siti Apryanti K

linkedin.com/in/sitiapryantii

s.apryanti17@gmail.com

081220554609

Pemahaman Masalah

Dataset

Dataset yang digunakan adalah berformat CSV
Dataset ini berisi mengenai sekumpulan data AirBNB

Identifikasi Masalah

- Bagaimana Perbedaan Bentuk Pemerintahan?
- Dimana sebagian Besar Properti berada?

Tools yang Dipakai

Bahasa Pemrograman Python

Ms. Excel

Jupyter Notebook

```python
import numpy as np
import pandas as pd
import io
import matplotlib.pyplot as plt
import seaborn as sns

loaddataset = pd.read_csv('dataset_airbnb.csv')
loaddataset
```

#import library yang akan digunakan

 # Menampilkan dataset AirBNB

#Isi Dataset AirBNB

	id	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	la
0	2539	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	1
1	2595	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	
2	3647	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	
3	3831	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	
4	5022	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	1
...	...	...	...	...	...	...	...	...	...	...	...	
48890	36484665	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	
48891	36485057	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	
48892	36485431	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	
48893	36485609	30985759	Taz	Manhattan	Hell's Kitchen	40.75751	-73.99112	Shared room	55	1	0	
48894	36487245	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404	-73.98933	Private room	90	7	0	

48895 rows × 15 columns

 # Memeriksa dasar informasi tentang kumpulan data yang digunakan

```python
loaddataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 15 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   host_id                         48895 non-null  int64
 2   host_name                       48874 non-null  object
 3   neighbourhood_group             48895 non-null  object
 4   neighbourhood                   48895 non-null  object
 5   latitude                        48895 non-null  float64
 6   longitude                       48895 non-null  float64
 7   room_type                       48895 non-null  object
 8   price                           48895 non-null  int64
 9   minimum_nights                  48895 non-null  int64
 10  number_of_reviews               48895 non-null  int64
 11  last_review                     38843 non-null  object
 12  reviews_per_month               38843 non-null  float64
 13  calculated_host_listings_count  48895 non-null  int64
 14  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(5)
memory usage: 5.6+ MB
```

 # Cek Heading pada sebuah dataset AirBNB

```
1 loaddataset.head()
```

	id	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
0	2539	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	10/19/2018
1	2595	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	5/21/2019
2	3647	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN
3	3831	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	7/5/2019
4	5022	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	11/19/2018

 # Cek hitungan, mean, std, min, maks, 25%, 50%, dan 75% persentil dari setiap atribut pada dataset AirBNB

```
1 loaddataset.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327

#cek jumlah baris dan kolom pada dataset AirBNB terdapat 48895 baris & 15 kolom

```
num_rows = loaddataset.shape[0]
num_cols = loaddataset.shape[1]
print(num_rows)
print(num_cols)
```

```
48895
15
```

#Menampilkan kolom yang memiliki nilai kosong

```
nilai_nuls = set(loaddataset.columns[loaddataset.isnull().mean()==0])
nilai_nuls
```

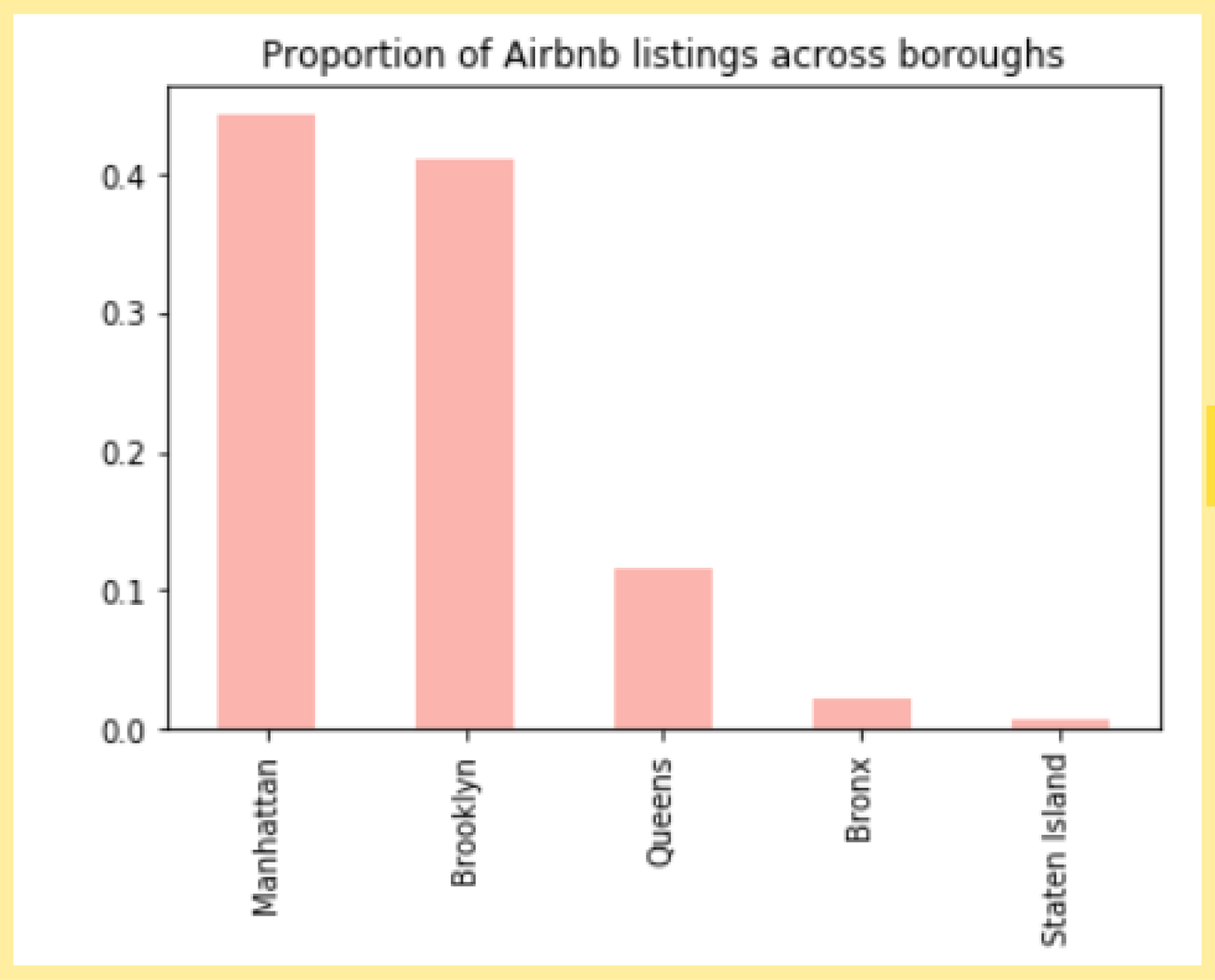
```
{'availability_365',
 'calculated_host_listings_count',
 'host_id',
 'id',
 'latitude',
 'longitude',
 'minimum_nights',
 'neighbourhood',
 'neighbourhood_group',
 'number_of_reviews',
 'price',
 'room_type'}
```

#cek nilai yang hilang/kosong dalam dataset AirBNB

```
loaddataset.isnull().sum()
```

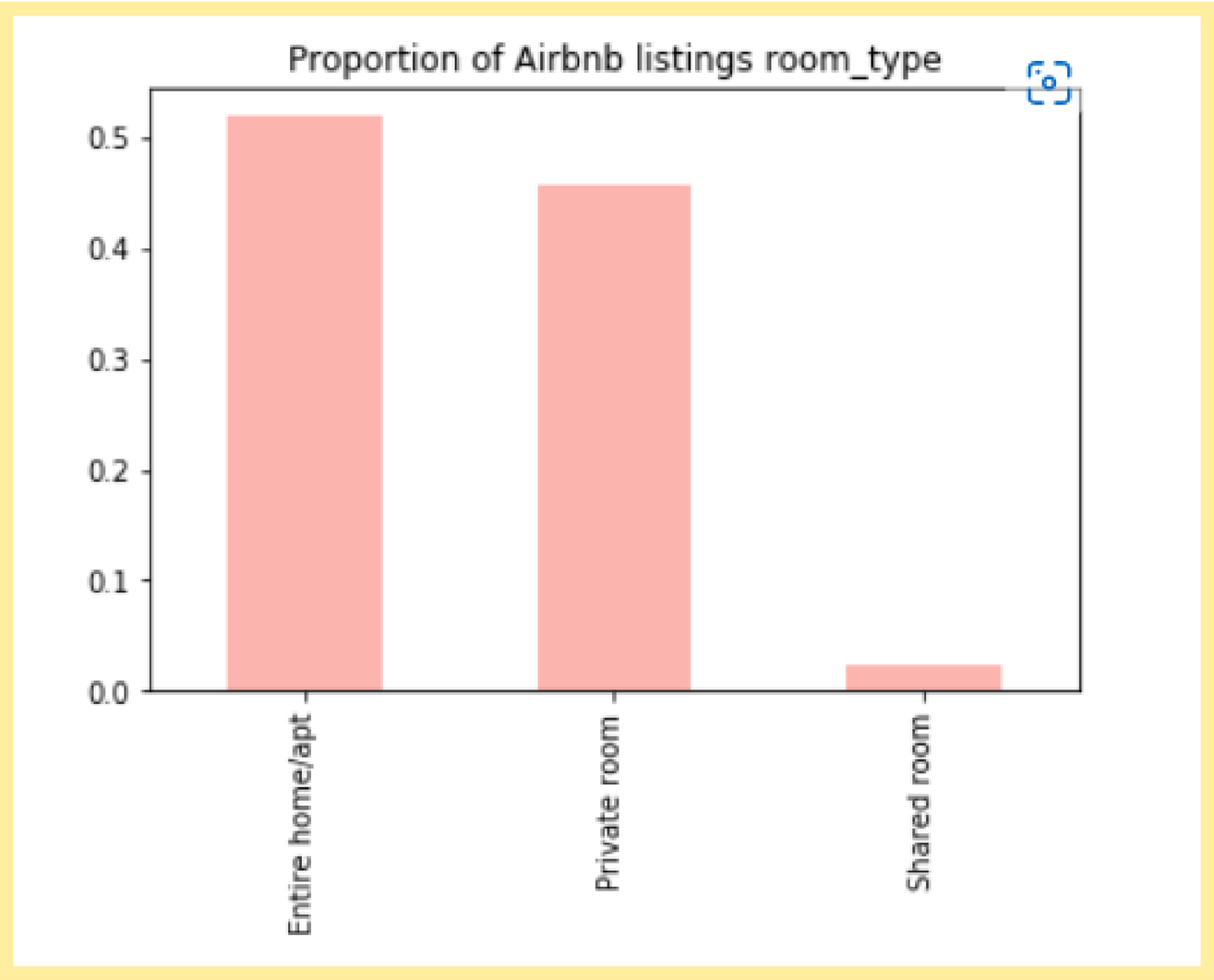
```
id                                  0
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

```
(loaddataset.neighbourhood_group.value_counts()/loaddataset.shape[0]).plot.bar(cmap="Pastel1", title="Proportion of Airbnb l
```



#Menampilkan daftar proporsi dikelima borought pada dataset AirBNB menggunakan diagram batang.
Diketahui dari diagram tersebut bahwa manhattan memiliki nilai tertinggi sedangkan staten island mendapatkan nilai paling rendah.

```
(loaddataset.room_type.value_counts()/loaddataset.shape[0]).plot.bar(cmap="Pastel1", title="Proportion of Airbnb listings ro
```



#Menampilkan daftar proporsi pada kategori room_type menggunakan diagram batang pada dataset AirBNB
diketahui bahwa rumah/apartemen memiliki proporsi paling tinggi, sedangkan proporsi paling rendah didapatkan oleh type kamar bersama (shared room)

Persiapan Data

```
loaddataset['last_review'] = pd.to_datetime(loaddataset['last_review'])
loaddataset['review_year'] = loaddataset['last_review'].apply(lambda last_review:last_review.year)
loaddataset['review_year'] = loaddataset['review_year'].fillna(0)
loaddataset['review_year'] = loaddataset.review_year.astype(int)
loaddataset = pd.concat([loaddataset[(loaddataset['availability_365']==0) & (loaddataset['review_year']==2019)], loaddataset
```

 # availability_365 = 0 menyarankan dua hal: a. ketika listing telah dihapus, b. Kapan kamar sudah dipesan
 # untuk menghapus kondisi a dari data, Saya memilih 0 posting ketersediaan itu ketika mereka memiliki setidaknya satu ulasan

```
loaddataset['review_per_month'] = loaddataset['reviews_per_month'].fillna(0)
```

 # mengisi data NA dengan 0
 # isi data NA dengan 0 di kolom reviews_per_month

```
loaddataset.info()
```

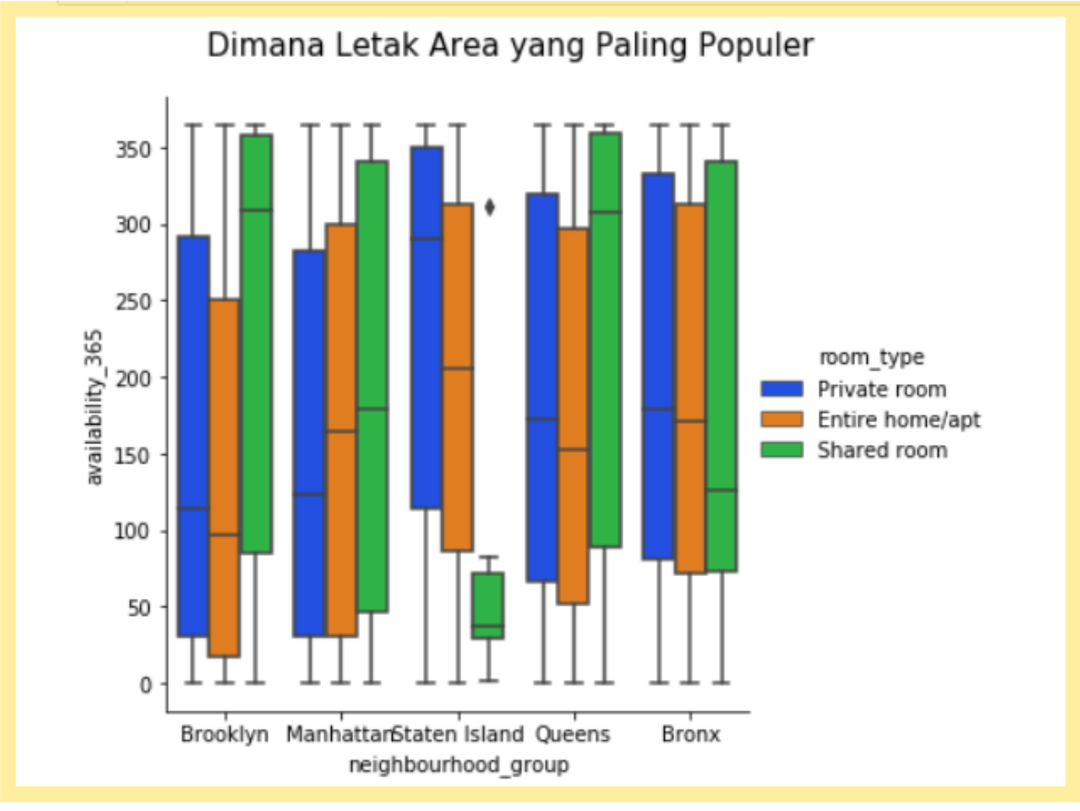
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34450 entries, 132 to 48894
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              34450 non-null  int64
 1   host_id                         34450 non-null  int64
 2   host_name                       34440 non-null  object
 3   neighbourhood_group             34450 non-null  object
 4   neighbourhood                   34450 non-null  object
 5   latitude                        34450 non-null  float64
 6   longitude                       34450 non-null  float64
 7   room_type                       34450 non-null  object
 8   price                           34450 non-null  int64
 9   minimum_nights                  34450 non-null  int64
 10  number_of_reviews               34450 non-null  int64
 11  last_review                     29243 non-null  datetime64[ns]
 12  reviews_per_month               29243 non-null  float64
 13  calculated_host_listings_count  34450 non-null  int64
 14  availability_365                34450 non-null  int64
 15  review_year                     34450 non-null  int32
 16  review_per_month                34450 non-null  float64
dtypes: datetime64[ns](1), float64(4), int32(1), int64(7), object(4)
memory usage: 4.6+ MB
```

#menampilkan kembali informasi dataset setelah dilakukan data cleaning

Visualisasi Data

Bagaimana Perbedaan Bentuk Pemerintahan?

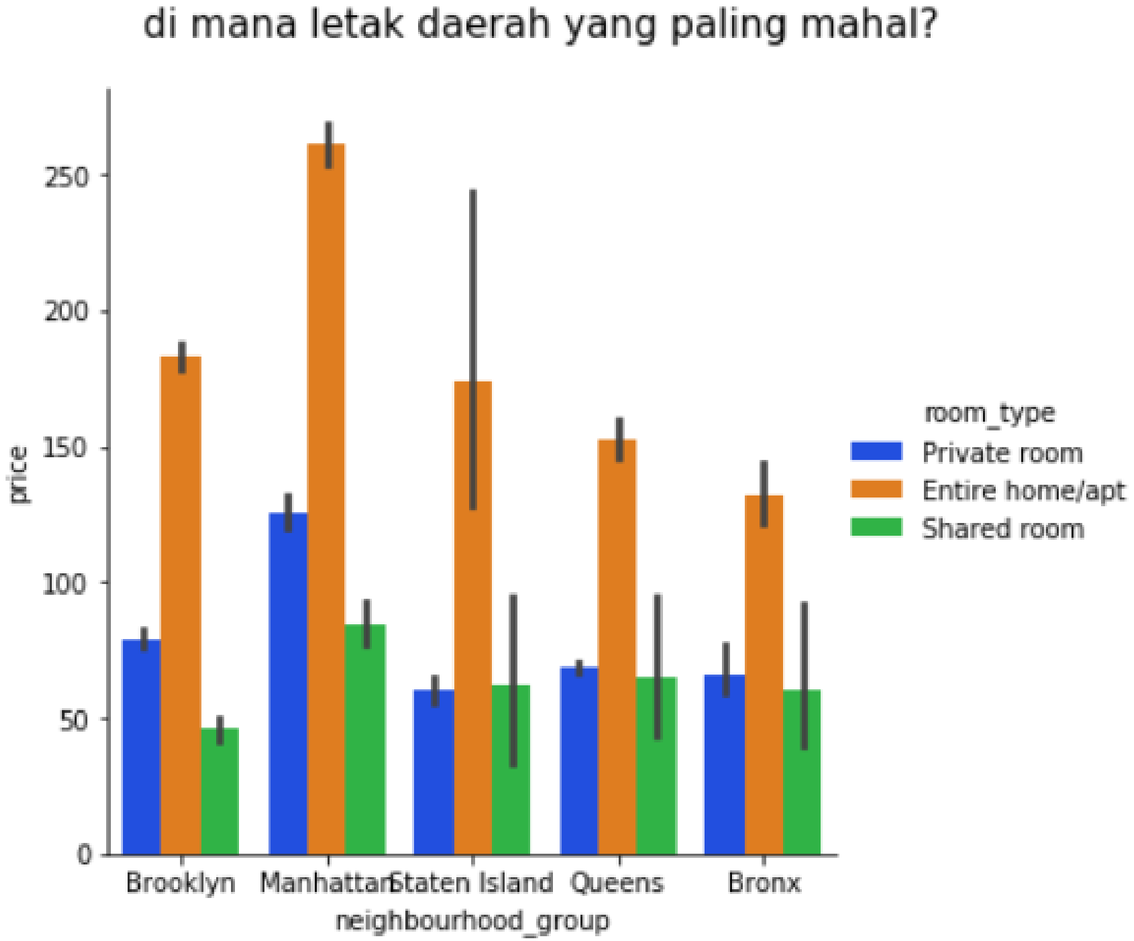
```
fig = sns.catplot(x='neighbourhood_group', y='availability_365', kind='box', hue='room_type',data=loaddataset, palette='brig
fig.fig.suptitle('Dimana Letak Area yang Paling Populer', fontsize=15, y=1.05)
fig.savefig("popular area.png", bbox_inches = 'tight')
```



- **Brooklyn pada seluruh rumah/apartemen memiliki ketersediaan terendah/permintaan tertinggi, diikuti oleh seluruh rumah atau apartemen di manhattan**

- **Ruang permintaan paling populer/tinggi adalah tipe kamar pribadi dan keseluruhan rumah/apartemen di sebagian besar berlokasi di Staten Island dan Bronx**

```
fig = sns.catplot(x='neighbourhood_group', y='price', data=loaddataset,kind='bar', hue='room_type', palette='bright')
fig.fig.suptitle('di mana letak daerah yang paling mahal?', fontsize=15, y=1.05)
fig.savefig("price area.png", bbox_inches = 'tight')
```



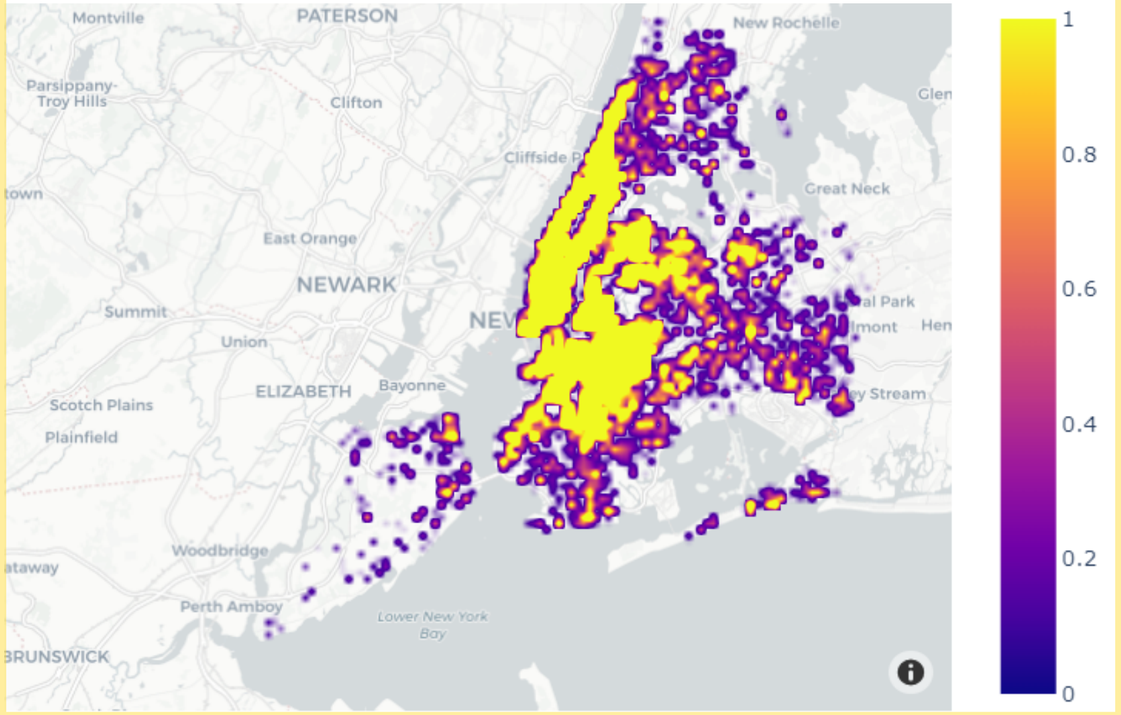
- **Seluruh rumah/apartemen yang berlokasi di daerah Manhattan memiliki harga yang paling mahal pertama.**
- **Lokasi kedua rumah/apartemen termahal terdapat pada lokasi Broklyn**
- **Sedangkan untuk tipe kamar yang memiliki harga tertinggi adalah tipe kamar pribadi (private room), dan apartemen.**
- **sebagian besar daerah yang paling mahal terletak di pulau staten (staten island) dan bronx**

Di mana sebagian besar properti berada?

```python
import plotly.express as px

lat = np.mean(loaddataset['latitude'])
lon = np.mean(loaddataset['longitude'])

fig = px.density_mapbox(loaddataset,lat ='latitude', lon="longitude",
                        radius=2, center=dict(lat=lat, lon=lon), zoom=10,
                         mapbox_style="carto-positron")
fig.show()
```



- **Berdasarkan plot maps disamping kita dapat mengetahui bahwa sebagian besar properti berada di Manhattan, di sisi selatan Central Park, dan juga disisi utara Brooklyn di sekitar Willamsburg**
- **Lokasi ini menawarkan harga trasansportasi yang paling nyaman dan banyak diminati wisatawan.**

Hasil Analisis

Hasil dari analisis ini, didapatkan beberapa faktor utama yang sangat berpengaruh. Diantaranya: Wisatawan/pelanggan lebih memilih lokasi yang dekat ke pusat kota, memiliki harga lebih murah dan semua kamar yang ditawarkan tipe kamar pribadi (private room).
hal ini dapat dipertimbangkan pihak AirBNB untuk memasarkan properti merek secara online. Hal ini dapat dipertimbangkan untuk tuan rumah airbnb saat memposting properti mereka secara online.

WELL DONE!

RevoU Mini Course

Issued 24 June 2022

SITI APRYANTI

has been awarded a certificate of completion for the

Intro to Data Analytics

a 2-weeks certified online course offered by RevoU

Matteo Sutto
CEO and Co-Founder
PT Revolusi Cita Edukasi

Verify at https://certificates.revou.co/siti-apryanti-certificate-completion-damc22.pdf
Revou has confirmed the identity of this individual and their participation in the course

CERTIFICATE OF COMPLETION

Linkedin.com/in/sitiapryantii/