Analisis Visualisasi Berdasarkan Pola Kunjungan Terhadap Waktu dan Lokasi Di Kota New York Dan Kota Tokyo

Anggota Kelompok:

1. Eksanty F Sugma Islamiaty - 122450001 (Ketua Kelompok)
2. Muhammad Bayu Syuhada - 122450007 (Anggota Kelompok)
3. Nadia Fitri Yani - 121450101 (Anggota Kelompok)

Deskripsi pembagian pekerjaan:

1. Eksanty F Sugma Islamiaty (Ketua Kelompok)
   Deskripsi: Menentukan judul, bagian 1, analisis visualisasi pertanyaan 1,2, dan 3, dan merapihkan laporan.
2. Muhammad Bayu Syuhada (Anggota Kelompok)
   Deskripsi: Ide dataset, pengolahan data, visualisasi, streamlit dan bagian 2.
3. Nadia Fitri Yani (Anggota Kelompok)
   Deskripsi: Analisis output visualisasi pertanyaan 4 dan 5, membuat PPT, dan infografik.

1. Pre-Implementation

1.1. Dataset dan Deskripsi

Dataset : NYC and Tokyo Check-in

Deskripsi: : Dataset ini menyimpan informasi tentang aktivitas kunjungan pengguna ke berbagai tempat di dua kota besar, yaitu New York dan Tokyo. Setiap baris dalam dataset ini mencatat satu kunjungan ke sebuah lokasi tertentu. Informasi yang tersedia mencakup identitas unik dari pengguna yang melakukan kunjungan, serta identitas tempat yang mereka kunjungi. Setiap tempat juga dikategorikan, misalnya sebagai toko kosmetik, restoran ramen, pusat medis, atau jenis lokasi lainnya. Selain itu, dataset ini mencantumkan koordinat geografis (latitude dan longitude) yang menunjukkan posisi tepat dari setiap tempat di peta. Terdapat juga informasi mengenai perbedaan zona waktu tempat tersebut dari UTC, yang membantu mengetahui perbedaan waktu lokal dengan waktu standar. Akhirnya, setiap kunjungan disertai dengan waktu kunjungan dalam format UTC yang menunjukkan kapan kunjungan itu terjadi. Dengan dataset ini, kita dapat menganalisis pola kunjungan pengguna, seperti jenis-jenis tempat yang paling sering dikunjungi, distribusi lokasi tempat-tempat di kota-kota tersebut, dan pola waktu kunjungan di kedua kota ini.

Pada Tabel 1 di bawah ini merupakan data check-in pada negara Tokyo yang terdiri atas 6 atribut yang tertera pada kolom tabel dan 573703 baris data.

Tabel 1. Data check-in negara Tokyo

Tokyo					
User ID	**Vanue Category**	**Latitude**	**Longtitude**	**Timezone**	**Timestamp**

1541	Cosmetics Shop	35.70510109	139.61959	540	Tue Apr 03 18:17:18 +0000 2012
868	Ramen / Noodle House	35.71558112	139.800317	540	Tue Apr 03 18:22:04 +0000 2012
114	Convenience Store	35.71454217	139.480065	540	Tue Apr 03 19:12:07 +0000 2012
868	Food & Drink Shop	35.72559199	139.776633	540	Tue Apr 03 19:12:13 +0000 2012
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1502	Tea Room	35.70174848	139.771216	540	Sat Feb 16 02:34:55 +0000 2013
408	Fast Food Restaurant	35.67046494	139.768348	540	Sat Feb 16 02:35:17 +0000 2013
1050	Record Shop	35.70406869	139.579496	540	Sat Feb 16 02:35:29 +0000 2013

Pada Tabel 2 di bawah ini merupakan data check-in pada negara New York yang terdiri atas 6 atribut yang tertera pada kolom tabel dan 227428 baris data.

Tabel 2. Data check-in negara New York

New York					
User ID	**Vanue Category**	**Latitude**	**Longtitude**	**Timezone**	**Timestamp**

470	Arts & Crafts Store	40.71981038	-74.002581	-240	Tue Apr 03 18:00:09 +0000 2012
979	Bridge	40.60679958	-74.04417	-240	Tue Apr 03 18:00:25 +0000 2012
69	Home (private)	40.71616168	-73.88307	-240	Tue Apr 03 18:02:24 +0000 2012
395	Medical Center	40.7451638	-73.982519	-240	Tue Apr 03 18:02:41 +0000 2012
87	Food Truck	40.74010383	-73.989658	-240	Tue Apr 03 18:03:00 +0000 2012
.	.	.	.	.	.
.	.	.	.	.	.
945	Home (private)	40.8543645	-73.88307	-300	Sat Feb 16 02:33:16 +0000 2013
671	Bar	40.73598132	-74.029309	-300	Sat Feb 16 02:34:31 +0000 2013
942	Bar	40.726805	-73.957422	-300	Sat Feb 16 02:35:36 +0000 2013

1.2. Pertanyaan Analisis Visualisasi

Berdasarkan ide cerita yang ada, Berikut beberapa pertanyaan terkait data dan visualisasi yang akan dibuat yaitu:

1. Apakah ada pola perilaku dari check-in pengguna?
   Langkah Kerja : Membuat pola dengan line chart (seperti setiap hari dan jam berapa sama kebanyakan user check-in).
   Metode : Visualisasi data temporal.

2. Apa saja tempat yang sering dikunjungi?
   Langkah Kerja : Pembuatan visualisasi dengan bentuk map atau peta.
   Metode : Visualisasi data geospasial.
3. Apakah ada pengaruh antara perbedaan zona waktu dengan waktu kunjungan?
   Langkah Kerja : Pembuatan bar plot berdasarkan waktu dan jumlah kunjungan.
   Metode : Visualisasi kuantitatif dengan heat map.

Dari ketiga pertanyaan ini, akan dibuatkan visualisasi yang membandingkan antara kota Tokyo dan New York untuk dianalisis adakah perbedaan antara kedua kota ini.

1.3. Ide Cerita

Terdapat beberapa ide cerita yang sudah peneliti dapatkan yaitu sebagai berikut:

1. Peta Sebaran Lokasi: menggunakan plot titik-titik latitude dan longitude untuk melihat distribusi tempat yang dikunjungi di kota New York dan Tokyo yang dapat divisualisasikan menggunakan scatter plot pada peta untuk menunjukkan kepadatan lokasi.
2. Distribusi Kategori Tempat: Visualisasi dengan diagram batang untuk menunjukkan distribusi frekuensi dari berbagai kategori tempat, sehingga kita dapat memahami kategori yang paling sering dikunjungi.
3. Frekuensi Kunjungan Berdasarkan Waktu: Histogram dan line chart yang menampilkan jumlah kunjungan berdasarkan waktu (misalnya, jam atau hari). Visualisasi ini dapat membantu dalam mengidentifikasi pola kunjungan berdasarkan waktu.
4. Analisis Zona Waktu: Heat Map untuk melihat bagaimana perbedaan zona waktu mempengaruhi waktu kunjungan, yang mungkin berguna dalam analisis perilaku pengguna lintas zona waktu.

2. Implementation

2.1. Preprocessing Dataset

Sebelum dilakukan visualisasi data, beberapa langkah pengolahan dataset telah dilakukan untuk memastikan kualitas data yang baik. Pertama, dataset dimuat menggunakan fungsi ‘pd.read_csv()’ dengan parameter ‘na_values=['#name?']’, yang bertujuan menggantikan data yang tidak valid atau kosong dengan nilai NaN. Langkah ini penting untuk menangani missing values yang mungkin ada dalam dataset. Selanjutnya, metode ‘head(5)’ digunakan untuk menampilkan lima baris pertama dari dataset.

Kode Program 1

```
import numpy as np
import pandas as pd
df1=pd.read_csv('/content/dataset_TSMC2014_NYC.csv',na_values=['#n
ame?'])
print(df1.head(5))
df2=pd.read_csv('/content/dataset_TSMC2014_TKY.csv',na_values=['#n
ame?'])
```

```
print(df2.head(5))
```

Output 1

```
userId                   venueId           venueCategoryId  \
0     470  49bbd6c0f964a520f4531fe3  4bf58dd8d48988d127951735
1     979  4a43c0aef964a520c6a61fe3  4bf58dd8d48988d1df941735
2      69  4c5cc7b485a1e21e00d35711  4bf58dd8d48988d103941735
3     395  4bc7086715a7ef3bef9878da  4bf58dd8d48988d104941735
4      87  4cf2c5321d18a143951b5cec  4bf58dd8d48988d1cb941735

         venueCategory   latitude  longitude  timezoneOffset  \
0  Arts & Crafts Store  40.719810 -74.002581          -240.0
1               Bridge  40.606800 -74.044170          -240.0
2       Home (private)  40.716162 -73.883070          -240.0
3       Medical Center  40.745164 -73.982519          -240.0
4           Food Truck  40.740104 -73.989658          -240.0

                    utcTimestamp
0  Tue Apr 03 18:00:09 +0000 2012
1  Tue Apr 03 18:00:25 +0000 2012
2  Tue Apr 03 18:02:24 +0000 2012
3  Tue Apr 03 18:02:41 +0000 2012
4  Tue Apr 03 18:03:00 +0000 2012
   userId                   venueId           venueCategoryId  \
0    1541  4f0fd5a8e4b03856eeb6c8cb  4bf58dd8d48988d10c951735
1     868  4b7b884ff964a5207d662fe3  4bf58dd8d48988d1d1941735
2     114  4c16fdda96040f477cc473a5  4d954b0ea243a5684a65b473
3     868  4c178638c2dfc928651ea869  4bf58dd8d48988d118951735
4    1458  4f568309e4b071452e447afe  4f2a210c4b9023bd5841ed28

           venueCategory   latitude   longitude  timezoneOffset  \
0         Cosmetics Shop  35.705101  139.619590             540
1  Ramen /  Noodle House  35.715581  139.800317             540
2      Convenience Store  35.714542  139.480065             540
3      Food & Drink Shop  35.725592  139.776633             540
4    Housing Development  35.656083  139.734046             540

                    utcTimestamp
0  Tue Apr 03 18:17:18 +0000 2012
1  Tue Apr 03 18:22:04 +0000 2012
2  Tue Apr 03 19:12:07 +0000 2012
3  Tue Apr 03 19:12:13 +0000 2012
4  Tue Apr 03 19:18:23 +0000 2012
```

Dari Kode program 1, menampilkan lima baris data awal dari dua dataset Kota Tokyo dan Kota New York dengan delapan atribut yaitu ‘userId’, ‘venueId’, ‘venueCategoryId’, ‘venueCategory’, ‘latitude’, ‘longitude’, ‘timezoneOffset’, dan ‘utcTimestam’.

Kemudian, fungsi ‘info()’ diaplikasikan pada kedua dataset untuk memeriksa jumlah total data, jumlah nilai non-null di setiap kolom struktur data awal, termasuk nama kolom, tipe data, dan contoh nilai di setiap kolom. Dari sini, dapat dianalisis apakah terdapat kolom yang memiliki missing values serta jenis tipe data yang perlu diolah lebih lanjut. Untuk lebih mendetail, metode ‘isnull().sum()’ digunakan untuk menghitung jumlah nilai kosong di setiap kolom, yang dapat menjadi indikasi perlunya imputasi atau pembersihan data. Selain itu, langkah deteksi duplikasi data dilakukan dengan ‘df1[df1.duplicated()]’ dan ‘df2[df2.duplicated()]’. Jika ditemukan duplikasi, fungsi ‘drop_duplicates()’ digunakan untuk menghapus data redundan.

Seluruh proses ini memastikan dataset bersih dari anomali seperti nilai kosong atau duplikat yang dapat mempengaruhi kualitas analisis dan visualisasi. Dengan langkah-langkah ini, dataset menjadi lebih siap untuk eksplorasi lanjutan, seperti pembuatan visualisasi yang akurat dan bermakna.

Kode Program 2

```
df1.info()
df2.info()
df1.isnull().sum()
df2.isnull().sum()
```

Output 2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 227428 entries, 0 to 227427
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   userId           227428 non-null  int64
 1   venueId          227428 non-null  object
 2   venueCategoryId  227428 non-null  object
 3   venueCategory    227428 non-null  object
 4   latitude         227428 non-null  float64
 5   longitude        227428 non-null  float64
 6   timezoneOffset   227428 non-null  int64
 7   utcTimestamp     227428 non-null  object
dtypes: float64(2), int64(2), object(4)
memory usage: 13.9+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 573703 entries, 0 to 573702
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   userId           573703 non-null  int64
 1   venueId          573703 non-null  object
 2   venueCategoryId  573703 non-null  object
 3   venueCategory    573703 non-null  object
```

```
 4   latitude         573703 non-null  float64
 5   longitude        573703 non-null  float64
 6   timezoneOffset   573703 non-null  int64
 7   utcTimestamp     573703 non-null  object
dtypes: float64(2), int64(2), object(4)
memory usage: 35.0+ MB

                  0
userId            0
venueId           0
venueCategoryId   0
venueCategory     0
latitude          0
longitude         0
timezoneOffset    0
utcTimestamp      0
```

Berdasarkan output 2 terdapat dua dataset dengan struktur yang mirip tetapi jumlah data berbeda. Pada dataset 1 didapatkan jumlah baris sebanyak 227428, 8 kolom, memori yang digunakan 13.9 mb. Sedangkan pada dataset 2 didapatkan jumlah baris sebanyak 573703, 8 kolom, memori yang digunakan 35.0 MB. Kedua dataset ini memiliki 3 tipe dataset yaitu tipe data int64 sebanyak 2 kolom yaitu userId, dan timezoneOffset, tipe data float64 sebanyak 2 kolom yaitu latitude dan longitude, dan tipe data object sebanyak 4 kolom yaitu venueId, venueCategoryId, venueCategory dan utcTimestamp. dan semua kolom pada tiap dataset lengkap tanpa missing values.

Pengambilan sampel dilakukan untuk mengelola dataset besar secara lebih efisien dan efektif. Langkah ini bertujuan untuk mengurangi beban komputasi sehingga proses analisis atau visualisasi dapat dilakukan lebih cepat, terutama pada tahap eksplorasi awal. Dengan menggunakan sampel, representasi data tetap dapat menggambarkan pola utama dalam dataset asli jika pengambilan dilakukan secara acak. Selain itu, pengambilan sampel membantu menghasilkan visualisasi yang lebih jelas dan mudah dipahami, tanpa risiko plot menjadi terlalu padat atau sulit dibaca. Penggunaan dataset yang lebih kecil juga menghemat sumber daya seperti memori dan daya pemrosesan, yang penting saat bekerja dengan sistem yang memiliki keterbatasan kapasitas. Secara umum, sampel digunakan untuk menguji metode analisis atau preprocessing sebelum diterapkan pada dataset penuh, sehingga memastikan proses lebih efisien tanpa kehilangan esensi data. Pada hal ini peneliti mengambil data sebanyal 150000 pada tiap-tiap dataset.

Kode Program 3

```
# Reducing data sample size for quick visual preview
nyc_sample = df1.sample(150000, random_state=42)
tky_sample = df2.sample(150000, random_state=42)
```

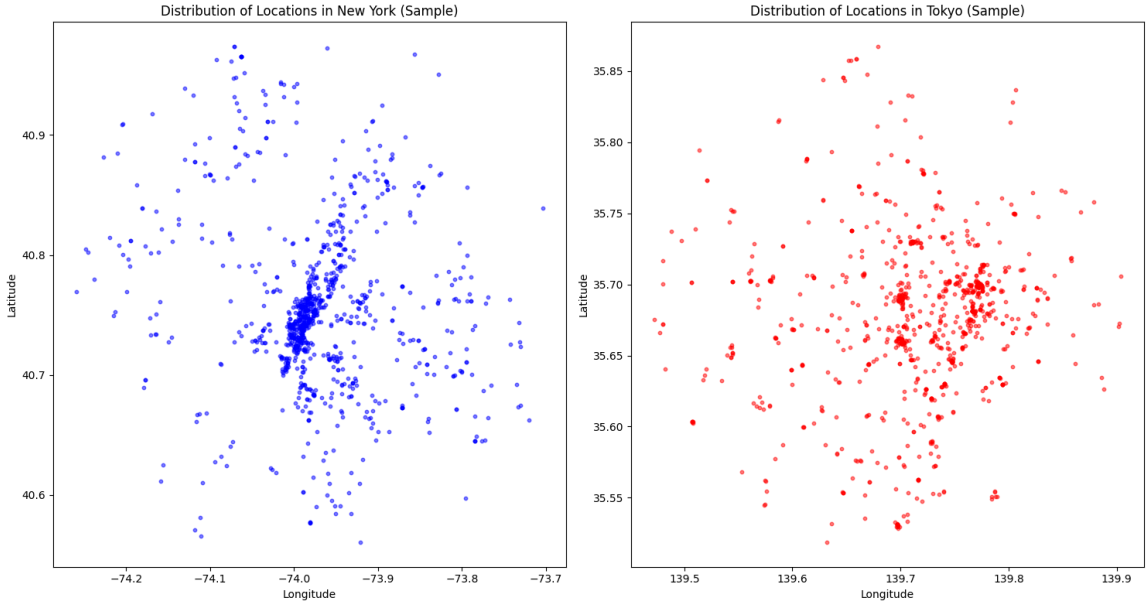
Dari Kode program 3 dataset pertama (df1) yang berisi 227.428 baris data, diambil sampel sebanyak 150.000 baris, yang setara dengan sekitar 65,9% dari total data.

Sementara itu, dari dataset kedua (df2) yang berisi 573.703 baris, diambil sampel dengan jumlah yang sama, yaitu 150.000 baris, sehingga hanya sekitar 26,1% data yang digunakan. Pengambilan sampel dilakukan secara acak menggunakan fungsi sampel dengan parameter random_state=42 untuk memastikan hasil sampel konsisten dan analisis dapat direproduksi. Dengan data yang lebih kecil, analisis menjadi lebih efisien secara komputasi, tanpa kehilangan representasi keseluruhan jika dataset asli tidak memiliki bias atau distribusi yang tidak merata. Namun, untuk dataset Tokyo, penggunaan hanya seperempat data dapat berisiko kehilangan pola kecil yang mungkin signifikan dalam dataset besar. Langkah ini ideal untuk memberikan gambaran awal tentang data sebelum melakukan analisis lebih mendalam dengan dataset penuh.

2.2. Visualization

Berdasarkan 1.2,

2.2.1. Pertanyaan 1: Bagaimana Distribusi Lokasi Pada Setiap Kota?



Gambar 1. Perbandingan visual distribusi lokasi di New York (biru) dan Tokyo (merah)

Berdasarkan titik-titik yang menyebar pada Gambar 1 tersebut, dapat diketahui persebaran lokasi yang paling banyak dikunjungi melalui info wilayah berdasarkan *Longitude* (sumbu X) dan *Latitude* (sumbu Y).

Hasilnya berdasarkan distribusi lokasi di New York (kiri-biru) menunjukkan kepadatan yang sangat tinggi pada satu titik pusat, kemudian menyebar dengan intensitas menurun ke area sekitarnya. Pada scatter plot ini, didapatkan bahwa kepadatan berada pada sekitar longitude -74, dan latitude 40.7 - 40.8 yang mana berdasarkan informasi ini didapatkan bahwa daerah padat ini adalah Midtown Manhattan yang berdekatan dengan tepi Sungai Hudson. Lokasi ini kemungkinan besar berada di sekitar Hell's Kitchen atau Hudson Yards, yang terkenal dengan aktivitas wisata, restoran, dan perkembangan infrastruktur modern. Area ini juga tidak jauh dari destinasi seperti Times Square dan Madison Square Garden.

Sedangkan berdasarkan distribusi lokasi di Tokyo (kanan-merah) menunjukkan pola yang padat pada satu titik pusat dengan penyebaran lebih merata ke arah sekeliling.

Pada scatter plot di bawah ini terlihat bahwa kepadatan berada di longitude 139.75 - 139.80 dan latitude 35.65 - 35.70 yaitu sekitar Chiyoda, Shinjuku, dan Shibuya, kedua kota ini merupakan pusat kota di Tokyo yang terkenal dengan sejarah, perbelanjaan, taman kota, dan kehidupan perkotaan modern seperti Imperial Palace, Hibiya Park, Tokyo Station, Tokyo Metropolitan Government Building, dan Shinjuku Gyoen National Garden.

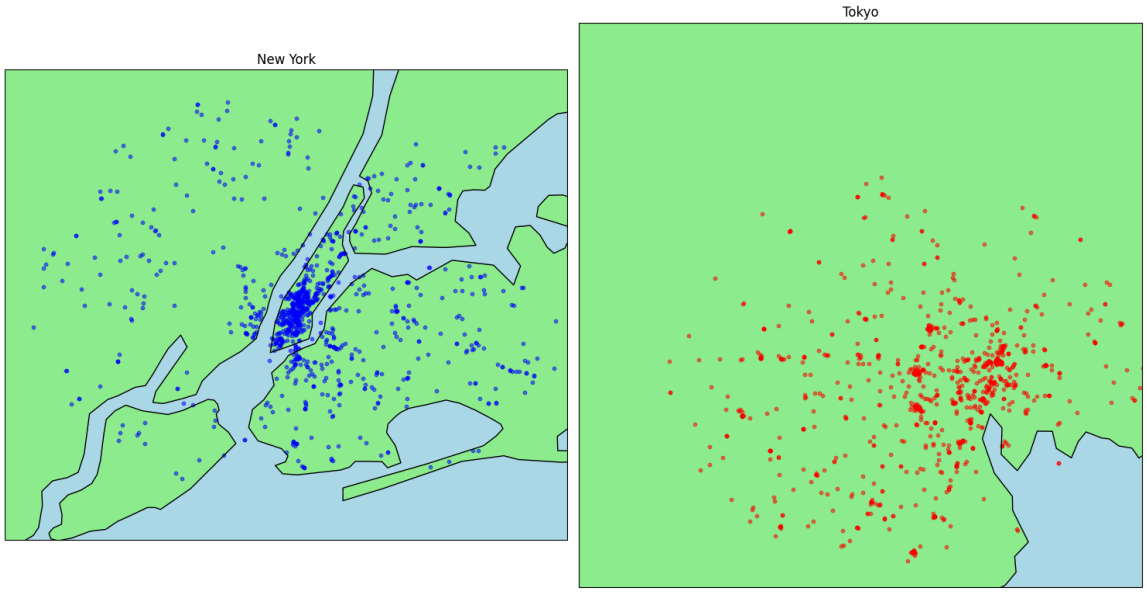
Kode Program 4

```
# Reducing data sample size for quick visual preview
nyc_sample = df1.sample(1000, random_state=42)  # Sample 1000 data
points for NYC
tky_sample = df2.sample(1000, random_state=42)  # Sample 1000 data
points for Tokyo

# 1. Peta Sebaran Lokasi untuk NYC dan Tokyo (dengan sample data)
fig, ax = plt.subplots(1, 2, figsize=(15, 8))

# Plot NYC sample data
ax[0].scatter(nyc_sample['longitude'], nyc_sample['latitude'],
alpha=0.5, s=10, color='blue')
ax[0].set_title("Distribution of Locations in New York (Sample)")
ax[0].set_xlabel("Longitude")
ax[0].set_ylabel("Latitude")

# Plot Tokyo sample data
ax[1].scatter(tky_sample['longitude'], tky_sample['latitude'],
alpha=0.5, s=10, color='red')
ax[1].set_title("Distribution of Locations in Tokyo (Sample)")
ax[1].set_xlabel("Longitude")
ax[1].set_ylabel("Latitude")

plt.tight_layout()
plt.show()
```

Kode ini digunakan untuk memvisualisasikan sebaran lokasi (longitude dan latitude) di dua kota, New York City (NYC) dan Tokyo, dengan menggunakan scatter plot. Terdapat beberapa tahapan yang ada dalam pembuatan scatter plot ini, yaitu membuat plot sebaran lokasi dengan membuat dua plot berdampingan (1,2) dan mengatur ukuran plot

(15,8), dan tampilan scatter plot dengan sumbu-x adalah longitude, sumbu-y adalah latitude, dengan alpha (jarak antar titik) 0.5, warna biru untuk New York dan merah untuk Tokyo.

Dari Kode program 4, peneliti menghasilkan visualisasi lebih jelas lagi dengan visualisasi geospasial atau data mapping yaitu scatter plot map berdasarkan koordinat geografis dari latitude dan longitude.



Gambar 2. Scatter Plot Map distribusi lokasi di New York (biru) dan Tokyo (merah)

Kode Program 5

```
import cartopy.crs as ccrs
import cartopy.feature as cfeature
import matplotlib.pyplot as plt

nyc_samplee = df1.sample(1000, random_state=42)  # Sample 1000 data
points for NYC
tky_samplee = df2.sample(1000, random_state=42)  # Sample 1000 data
points for Tokyo

fig, ax = plt.subplots(1, 2, figsize=(15, 8),
subplot_kw={'projection': ccrs.PlateCarree()})

ax[0].set_extent([-74.3, -73.7, 40.5, 41.0],
crs=ccrs.PlateCarree())
ax[0].add_feature(cfeature.LAND, facecolor='lightgreen')
ax[0].add_feature(cfeature.OCEAN, facecolor='lightblue')
```

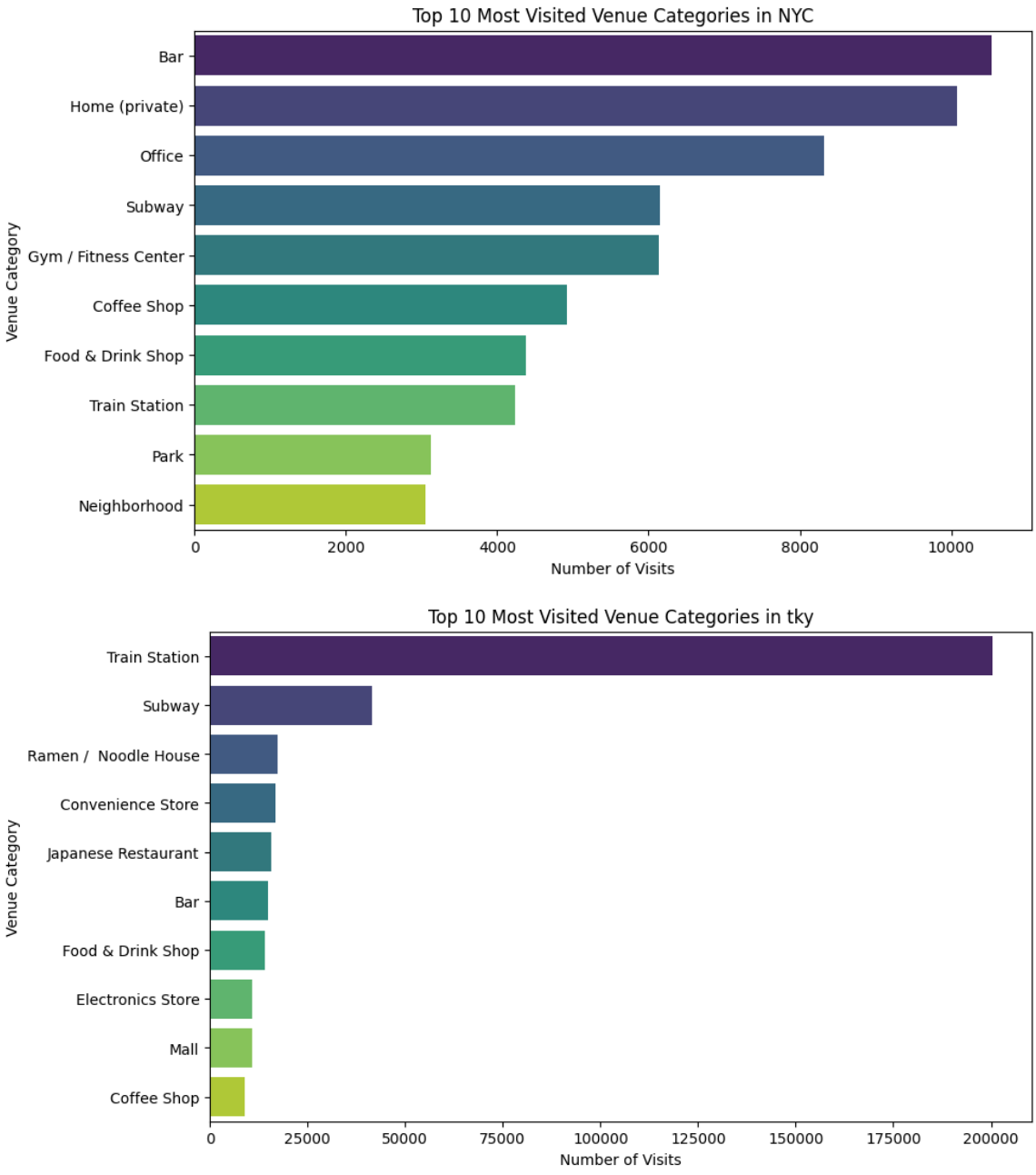
```
ax[0].add_feature(cfeature.COASTLINE)
ax[0].add_feature(cfeature.BORDERS, linestyle=':')
ax[0].scatter(nyc_samplee['longitude'], nyc_samplee['latitude'],
alpha=0.5, s=10, color='blue', label="NYC Locations")
ax[0].set_title("New York")

ax[1].set_extent([139.4, 139.9, 35.5, 36.0],
crs=ccrs.PlateCarree())
ax[1].add_feature(cfeature.LAND, facecolor='lightgreen')
ax[1].add_feature(cfeature.OCEAN, facecolor='lightblue')
ax[1].add_feature(cfeature.COASTLINE)
ax[1].add_feature(cfeature.BORDERS, linestyle=':')
ax[1].scatter(tky_samplee['longitude'], tky_samplee['latitude'],
alpha=0.5, s=10, color='red', label="Tokyo Locations")
ax[1].set_title("Tokyo")

plt.tight_layout()
plt.show()
```

Kode Program 5 di atas digunakan untuk membuat visualisasi geospasial sederhana yang membandingkan distribusi lokasi di New York City (NYC) dan Tokyo. Dalam visualisasi ini, data sampel sebanyak 1.000 titik lokasi diambil dari masing-masing dataset menggunakan fungsi sample dengan 'random_state=42' untuk memastikan kekonsistenan hasil.

Visualisasi dilakukan dengan menggunakan pustaka **Cartopy** untuk menampilkan peta geografis dalam proyeksi **PlateCarree**. Dua subplot dibuat, masing-masing menampilkan peta NYC dan Tokyo dengan pengaturan rentang geografis spesifik: NYC mencakup koordinat [-74.3, -73.7] untuk garis bujur dan [40.5, 41.0] untuk garis lintang, sementara Tokyo mencakup [139.4, 139.9] untuk garis bujur dan [35.5, 36.0] untuk garis lintang.

Peta dihiasi dengan fitur-fitur seperti daratan berwarna hijau, lautan biru muda, garis pantai, dan batas negara untuk memberikan konteks geografis. Lokasi data sampel diplot pada peta masing-masing menggunakan titik-titik biru untuk NYC dan merah untuk Tokyo, dengan transparansi (alpha=0.5) dan ukuran kecil (s=10) untuk memudahkan identifikasi pola distribusi.

2.2.2. Pertanyaan 2: Apa saja tempat yang sering dikunjungi?



Gambar 3. Perbandingan 10 tempat yang sering dikunjungi

Pada kota New York tempat yang memiliki distribusi tertinggi dikunjungi adalah bar, lalu rumah, dan juga tempat kerja. Dari data tersebut, terlihat bahwa aktivitas sehari-hari di Kota New York mencakup keseimbangan antara pekerjaan, transportasi, hiburan malam, kebugaran, dan rekreasi terbuka. Sedangkan pada kota Tokyo distribusi tempat yang paling sering dikunjungi adalah stasiun kereta api mendominasi menjadi kategori paling sering dikunjungi dengan jumlah yang sangat besar dibandingkan kategori lainnya. Lalu kereta bawah tanah, dan toko mie atau ramen. Berdasarkan Gambar 3 kota Tokyo menunjukkan bahwa mobilitas, kuliner, dan belanja adalah aktivitas dominan.

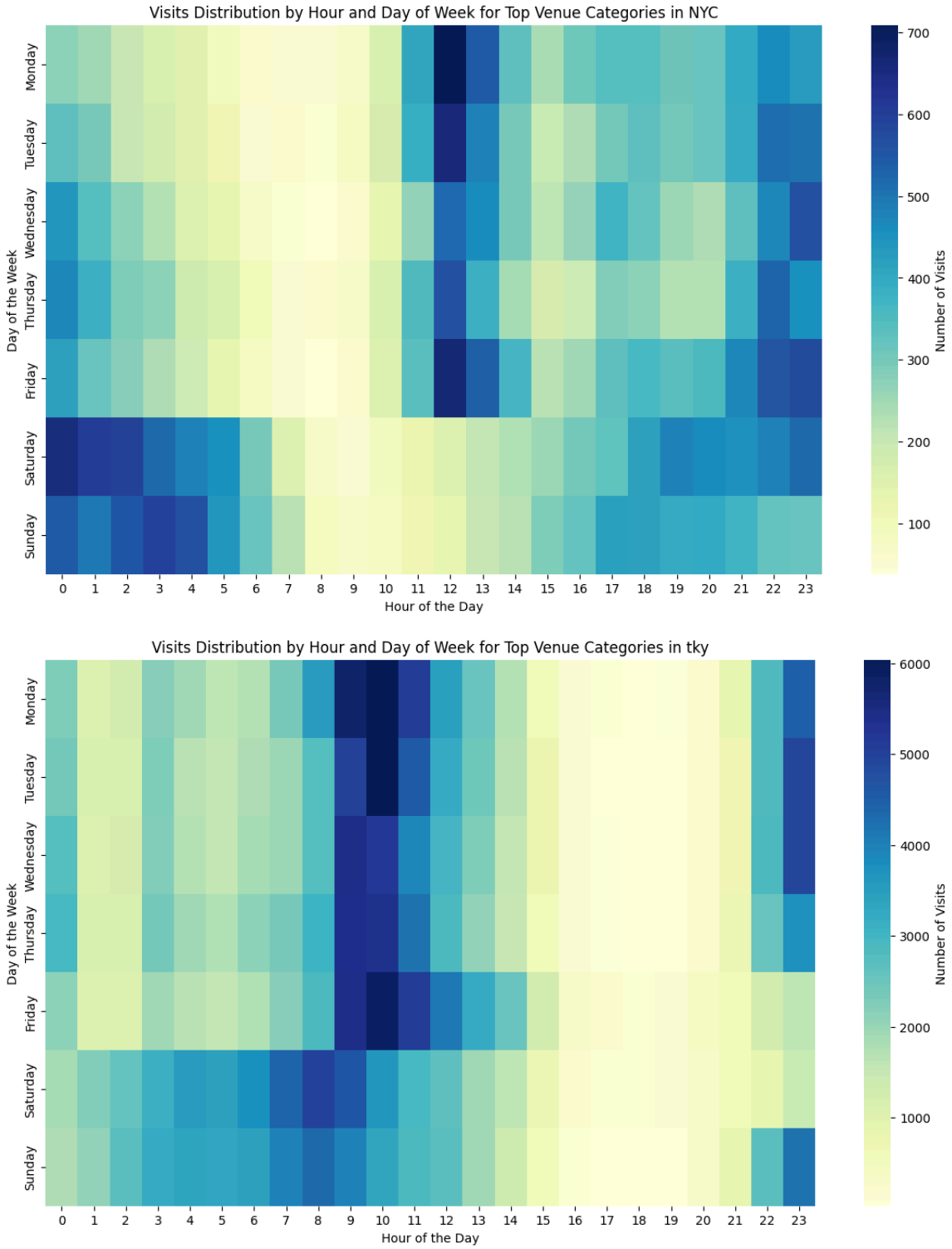
Dari kedua kota ini, terlihat sangat berbeda dari segi tempat yang paling banyak dikunjungi. Artinya kota mempengaruhi tempat yang sering dikunjungi.

Kode Program 6

```
import matplotlib.pyplot as plt  # Import matplotlib.pyplot
import seaborn as sns
import pandas as pd

nyc_data = df1

# Convert 'utcTimestamp' column to datetime objects using
pd.to_datetime
nyc_data['utcTimestamp'] = pd.to_datetime(nyc_data['utcTimestamp'],
errors='coerce')

# Extract day of the week and hour
nyc_data['day_of_week'] = nyc_data['utcTimestamp'].dt.day_name()
nyc_data['hour'] = nyc_data['utcTimestamp'].dt.hour

# Visualize top 10 most visited venue categories
top_venue_categories =
nyc_data['venueCategory'].value_counts().head(10)

plt.figure(figsize=(10, 6))
sns.barplot(y=top_venue_categories.index,
x=top_venue_categories.values, palette="viridis")
plt.title("Top 10 Most Visited Venue Categories in NYC")
plt.xlabel("Number of Visits")
plt.ylabel("Venue Category")
plt.show()
```

Kode program 6 ini mengambil data dalam format timestamp dan mengkonversinya menjadi format datetime (senin-minggu, 00.00-23.59) yang akan menghasilkan waktu berupa hari dan jam untuk analisis lebih lanjut. Selanjutnya menghitung 10 kategori tempat teratas berdasarkan jumlah kunjungan dan data yang dihasilkan akan dibuat grafik berupa grafik batang horizontal untuk memvisualisasikan 10 kategori tempat teratas beserta jumlah kunjungannya dengan sumbu-x adalah jumlah kunjungan dan sumbu-y adalah nama kategori tempat.

2.2.3. Pertanyaan 3: Apakah ada pengaruh antara perbedaan zona waktu dengan waktu kunjungan?



Gambar 4. Kunjungan per Jam dan Hari di New York City dan Tokyo

Grafik yang ditampilkan pada Gambar 4 adalah heatmap yang menunjukkan distribusi jumlah kunjungan berdasarkan jam dalam sehari dan hari dalam seminggu untuk kategori tempat di NYC. Sumbu-y: Hari dalam seminggu (Senin - Minggu). Sumbu-x:

Jam dalam sehari (0.00-23.29). Dan Warna: Jumlah kunjungan, dengan warna yang lebih gelap menunjukkan lebih banyak kunjungan.

Kota New York mendapati jumlah kunjungan cenderung banyak terjadi pada pagi hari (sekitar pukul 8.00-10.00) dan sore/malam hari (sekitar pukul 17.00-23.00). Kunjungan lebih banyak terjadi pada hari libur di malam hari sedangkan pada hari-hari kerja, terdapat penurunan kunjungan yang signifikan. Artinya Zona waktu berpotensi mempengaruhi perilaku aktivitas manusia, terutama aktivitas bisnis dan sosial.

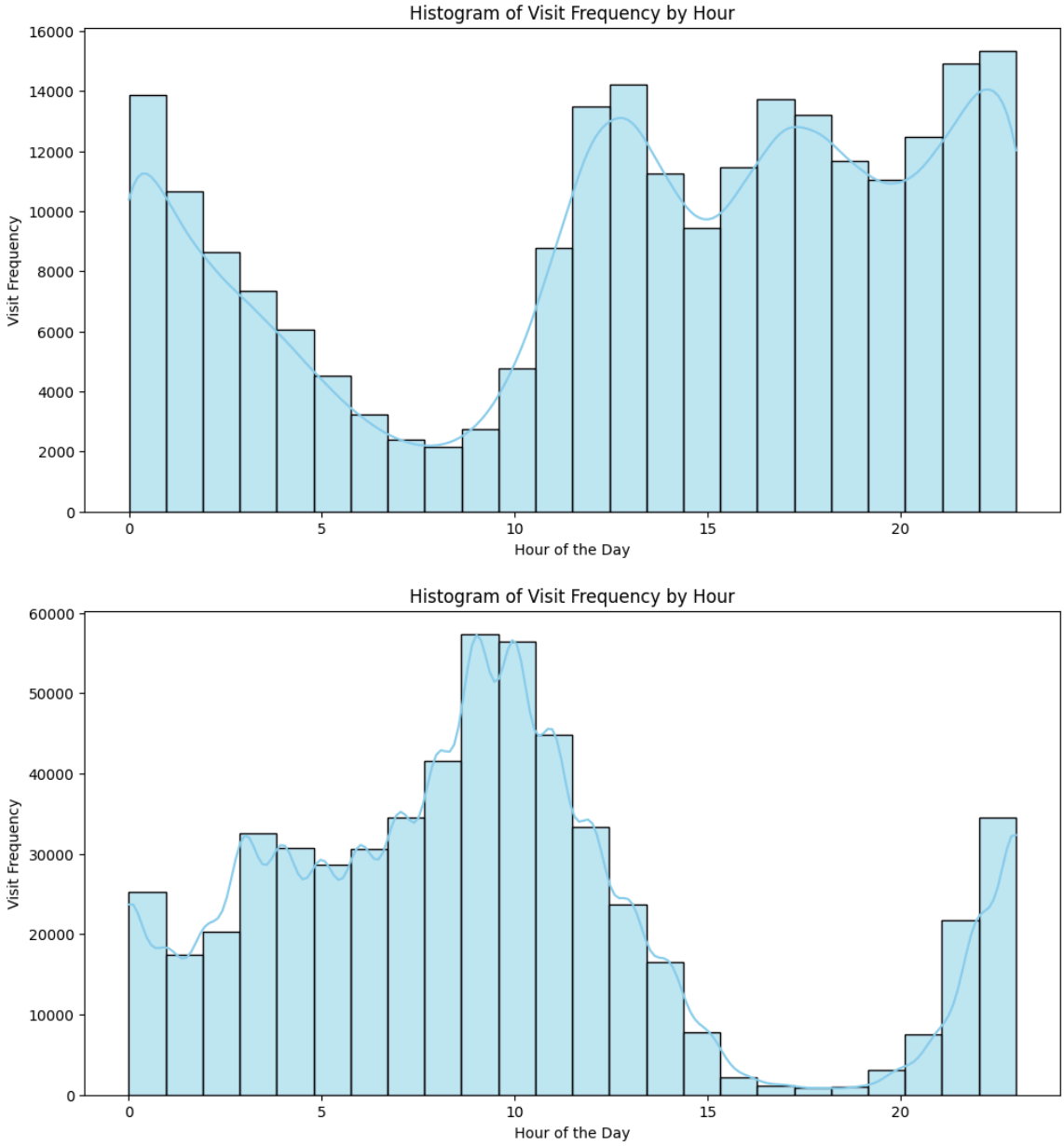
Pada Kota Tokyo, kunjungan memuncak pada pagi hari, sekitar pukul 08.00 - 10.00. Jumlah kunjungan menurun drastis pada siang hari (sekitar 12:00 - 17:00) dan dini hari (0:00 - 5:00). Terdapat peningkatan kunjungan yang signifikan pada malam hari sekitar pukul 21:00 - 23:00, terutama pada hari Senin dan Selasa. Pada semua hari, jam 0:00 hingga 5:00 memiliki jumlah kunjungan paling rendah, yang wajar karena mayoritas aktivitas publik berhenti pada jam ini.

Kode Program 7

```
# Filter data for top venue categories only
top_venue_data =
nyc_data[nyc_data['venueCategory'].isin(top_venue_categories.index)
]

# Create a pivot table to count visits by day of week and hour for
each top venue category
venue_heatmap_data = top_venue_data.pivot_table(
    index='day_of_week', columns='hour', values='venueCategory',
aggfunc='count'
).fillna(0)

# Reorder days of the week for better visual alignment
days_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday',
'Friday', 'Saturday', 'Sunday']
venue_heatmap_data = venue_heatmap_data.reindex(days_order)

# Plot heatmap
plt.figure(figsize=(14, 8))
sns.heatmap(venue_heatmap_data, cmap="YlGnBu", annot=False,
cbar_kws={'label': 'Number of Visits'})
plt.title("Visits Distribution by Hour and Day of Week for Top
Venue Categories in NYC")
plt.xlabel("Hour of the Day")
plt.ylabel("Day of the Week")
plt.show()
```

Kode program 7 bertujuan untuk memfilter dataset New York dan Tokyo hanya untuk kategori venue yang termasuk dalam 10 kategori teratas. membuat pivot tabel dengan menghitung jumlah kunjungan berdasarkan hari dalam seminggu (sumbu-y) dan jam dalam sehari (sumbu-x) adapun aggfunc='count': Menghitung jumlah kunjungan untuk setiap kombinasi hari dan jam. Ketika semua sudah diatur, lalu visualisasi heatmap ditampilkan memperhitungkan jumlah kunjungan maka warna pada heatmap akan lebih gelap menandakan jumlah kunjungan lebih tinggi

2.2.4. Pertanyaan 4: **Apakah waktu mempengaruhi pola kunjungan di kedua kota?**



Gambar 5. Histogram Frekuensi Kunjungan per Jam di Tokyo dan New York

Pola kunjungan di Kota New York dan Tokyo menunjukkan perbedaan yang signifikan dalam distribusi waktu aktivitas harian. Di Kota New York, aktivitas cenderung memuncak pada malam hari, khususnya antara pukul 20:00 hingga 23:00, yang menandakan bahwa masyarakat New York lebih dominan melakukan kegiatan di malam hari. Pola ini dapat dikaitkan dengan budaya urban yang dinamis dan kehidupan malam yang aktif, seperti hiburan, restoran, dan acara sosial yang menjadi daya tarik utama. Sebaliknya, aktivitas mulai menurun secara signifikan pada dini hari sekitar 02:00 hingga 06:00, yang merupakan waktu istirahat atau tidur.

Sementara itu, pola kunjungan di Tokyo menunjukkan distribusi aktivitas yang lebih seimbang dan berfokus pada pagi hari. Puncak aktivitas terjadi antara pukul 09:00 hingga 11:00, yang menunjukkan bahwa masyarakat Tokyo lebih aktif di pagi hari, mungkin terkait dengan rutinitas kerja, sekolah, atau kegiatan produktif lainnya. Aktivitas kemudian menurun pada sore hingga awal malam, antara pukul 15:00 hingga 20:00, yang tampaknya menjadi waktu istirahat atau transisi setelah aktivitas pagi. Namun, frekuensi kunjungan kembali meningkat pada malam hari, sekitar pukul 21:00 hingga 23:00, meskipun tidak setinggi puncak pagi. Pola ini menunjukkan bahwa masyarakat Tokyo memiliki fokus aktivitas di pagi hari, namun tetap aktif di malam hari meski dengan intensitas yang lebih rendah dibandingkan New York

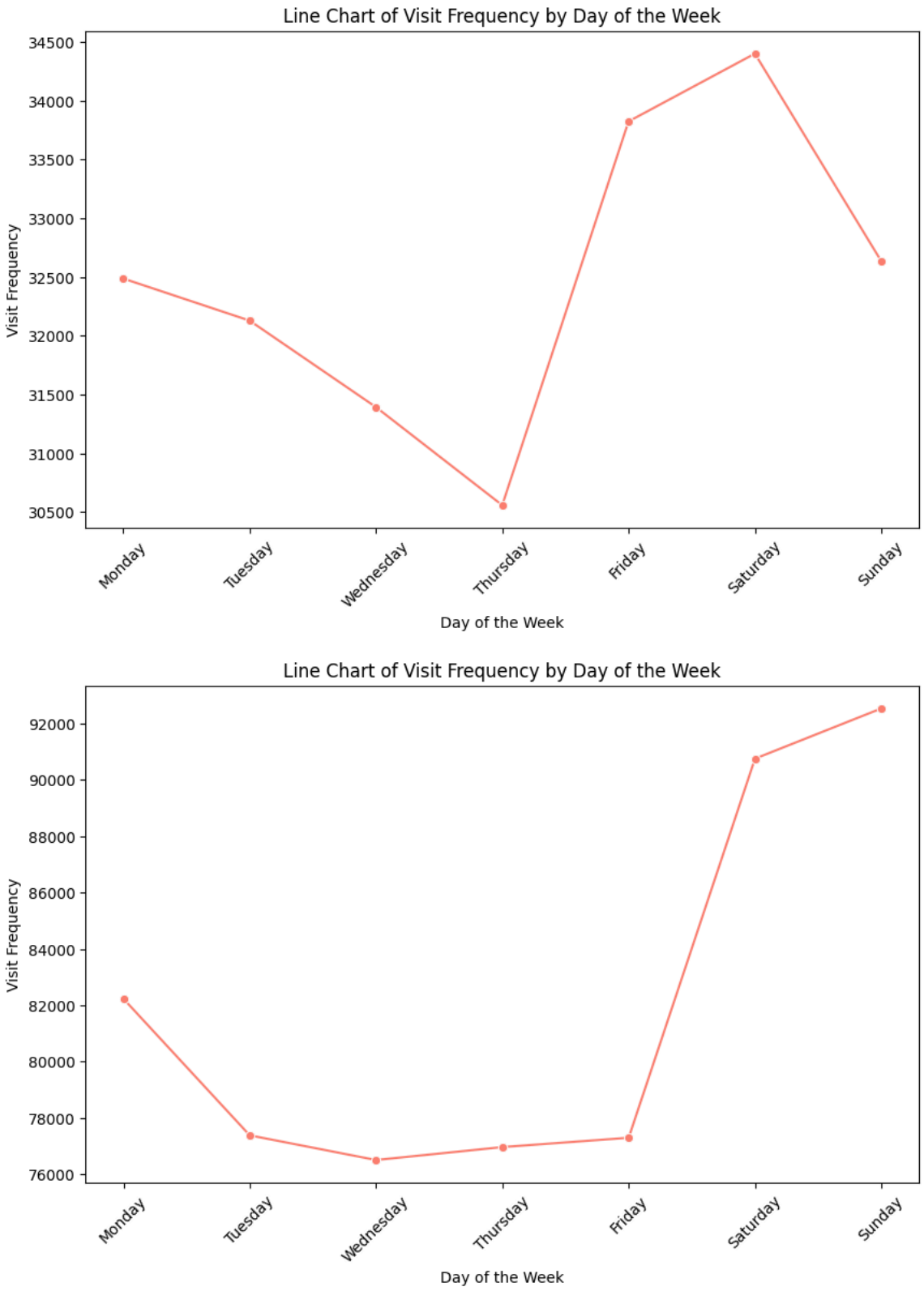
Kode Program 8

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data =df1

# Pastikan kolom 'utcTimestamp' sudah ada dan dapat dikonversi
if 'utcTimestamp' in data.columns:
    # Ubah kolom waktu ke format datetime
    data['utcTimestamp'] = pd.to_datetime(data['utcTimestamp'],
errors='coerce')

    # Ambil hari dan jam dari kolom waktu
    data['day_of_week'] = data['utcTimestamp'].dt.day_name()
    data['hour'] = data['utcTimestamp'].dt.hour

    # Urutkan hari dalam minggu
    days_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday',
'Friday', 'Saturday', 'Sunday']
    data['day_of_week'] = pd.Categorical(data['day_of_week'],
categories=days_order, ordered=True)
```

```
# Histogram berdasarkan jam
plt.figure(figsize=(12, 6))
sns.histplot(data['hour'], bins=24, kde=True, color="skyblue")
plt.title("Histogram of Visit Frequency by Hour")
plt.xlabel("Hour of the Day")
plt.ylabel("Visit Frequency")
plt.show()
```

Kode program 8 di atas bertujuan untuk menganalisis frekuensi kunjungan berdasarkan waktu pada data yang memiliki kolom utcTimestamp. Proses dimulai dengan memastikan bahwa kolom utcTimestamp ada dalam DataFrame dan dapat dikonversi ke format waktu yang sesuai. Setelah itu, dua kolom baru ditambahkan yaitu 'day_of_week' yang menunjukkan nama hari dalam seminggu, dan 'hour' yang menunjukkan jam dari waktu yang tercatat. Hari-hari dalam seminggu diurutkan mulai dari Senin hingga Minggu. Kemudian, Kode ini membuat histogram yang menggambarkan frekuensi kunjungan berdasarkan jam dalam sehari. Histogram ini menggunakan 24 batang, masing-masing untuk setiap jam, dengan tambahan kurva distribusi kernel (KDE) untuk memberikan gambaran yang lebih halus mengenai pola distribusi data. Hasil dari analisis ini memberikan wawasan tentang pola waktu kunjungan pengguna dan kapan kunjungan paling sering terjadi dalam rentang waktu sehari. Dengan visualisasi ini, pengguna dapat lebih mudah mengidentifikasi jam-jam sibuk atau periode dengan frekuensi kunjungan yang lebih rendah.

2.2.5. Pertanyaan 5: Hari apa saja yang sering dikunjungi di kedua kota tersebut?



Gambar 6. Frekuensi kunjungan berdasarkan hari per minggu dengan line plot

Berdasarkan grafik, pola kunjungan di Kota New York memiliki puncak pada hari Jumat dan Sabtu. Di awal minggu, frekuensi kunjungan cenderung menurun secara

bertahap, dimulai dari hari Senin hingga mencapai titik terendah pada Kamis. Hal ini dapat dikaitkan dengan rutinitas pekerjaan atau kegiatan harian yang membuat masyarakat lebih fokus pada aktivitas produktif. Menjelang akhir pekan, khususnya pada hari Jumat, terjadi peningkatan signifikan yang kemudian mencapai puncaknya pada Sabtu. Ini menunjukkan bahwa masyarakat New York cenderung mulai beraktivitas sosial atau melakukan perjalanan wisata sejak Jumat. Namun, pada hari Minggu, frekuensi kunjungan mulai menurun, yang mungkin disebabkan oleh persiapan untuk kembali ke rutinitas awal minggu.

Berbeda dengan Kota New York, pola kunjungan di Tokyo lebih stabil di awal minggu. Aktivitas kunjungan di Tokyo memiliki titik terendah pada hari Rabu, dan mulai meningkat perlahan menjelang akhir pekan. Lonjakan frekuensi yang signifikan terjadi pada Sabtu dan mencapai puncaknya pada hari Minggu, menjadikannya hari dengan kunjungan tertinggi. Pola ini mengindikasikan bahwa masyarakat Tokyo lebih memanfaatkan akhir pekan penuh, terutama pada Sabtu dan Minggu, untuk melakukan aktivitas rekreasi, berbelanja, atau berkumpul bersama keluarga dan sebagainnya. Jika dibandingkan, kedua kota memiliki pola yang serupa di mana akhir pekan menjadi waktu dengan kunjungan tertinggi. Namun, terdapat perbedaan dalam puncak aktivitasnya. Kota New York menunjukkan lonjakan kunjungan lebih awal, yakni pada Jumat, sedangkan Tokyo lebih memusatkan aktivitas pada Sabtu dan Minggu. Perbedaan ini mencerminkan karakteristik kehidupan masyarakat masing-masing kota. Secara keseluruhan, kedua kota menunjukkan bahwa akhir pekan adalah periode yang paling banyak dimanfaatkan oleh masyarakat untuk beraktivitas di luar rutinitas harian.

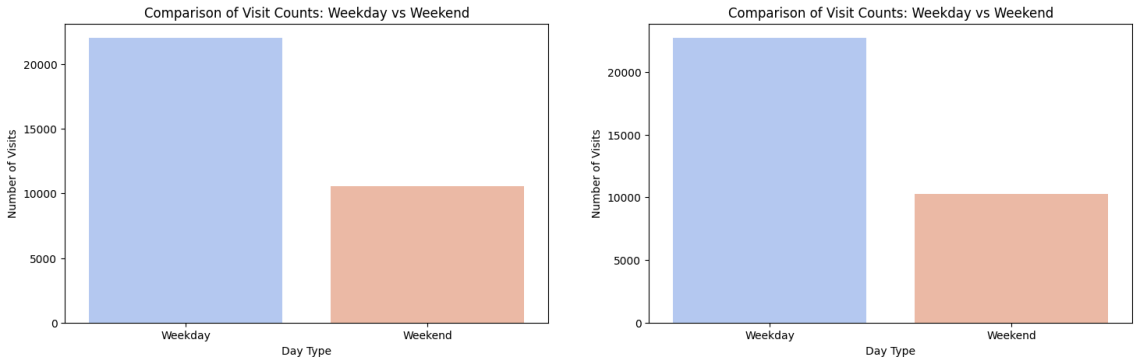
Kode Program 9

```
# Line chart berdasarkan hari
    visit_by_day =
data['day_of_week'].value_counts().reindex(days_order)

    plt.figure(figsize=(10, 6))
    sns.lineplot(x=visit_by_day.index, y=visit_by_day.values,
marker='o', color="salmon")
    plt.title("Line Chart of Visit Frequency by Day of the Week")
    plt.xlabel("Day of the Week")
    plt.ylabel("Visit Frequency")
    plt.xticks(rotation=45)
    plt.show()
else:
    print("Kolom 'utcTimestamp' tidak ditemukan dalam data.")
```

Kode program 9 bertujuan untuk menganalisis frekuensi kunjungan berdasarkan hari dalam seminggu dan memvisualisasikannya menggunakan line chart. Langkah pertama adalah menghitung jumlah kunjungan untuk setiap hari dengan menggunakan metode ‘value_counts()’ pada kolom ‘day_of_week’. Hasilnya kemudian disusun ulang menggunakan ‘reindex()’ agar sesuai dengan urutan hari yang sesuai, yaitu dari Senin

hingga Minggu (days_order). Setelah data frekuensi terurut, Kode membuat line chart dengan ‘sns.lineplot’ untuk menampilkan perubahan jumlah kunjungan sepanjang minggu. Setiap titik pada grafik diberi penanda (marker='o') untuk memperjelas nilai masing-masing hari, dan grafik diberi warna salmon agar lebih menarik secara visual. Label sumbu X (hari) dan Y (frekuensi kunjungan) ditambahkan untuk memudahkan interpretasi, sementara rotasi pada label sumbu X memastikan keterbacaan. Jika kolom utcTimestamp tidak ditemukan dalam data, pesan akan ditampilkan untuk memberi tahu pengguna tentang masalah tersebut. Grafik ini bertujuan untuk memberikan wawasan tentang pola kunjungan berdasarkan hari, seperti hari dengan kunjungan tertinggi atau terendah.

2.2.6. Pertanyaan 6: Apakah faktor waktu mempengaruhi pengguna dalam mengunjungi lokasi tempat?



Gambar 7. Pengaruh hari terhadap kunjungan suatu tempat di New York (Kiri) dan di Tokyo (Kanan)

Faktor waktu tepatnya hari dipilih karena memiliki pengaruh signifikan pada pola aktivitas masyarakat. hari kerja (weekday) dan akhir pekan (weekend) mencerminkan perubahan rutinitas masyarakat, seperti perbedaan antara aktivitas kerja, rekreasi, atau kegiatan lainnya.

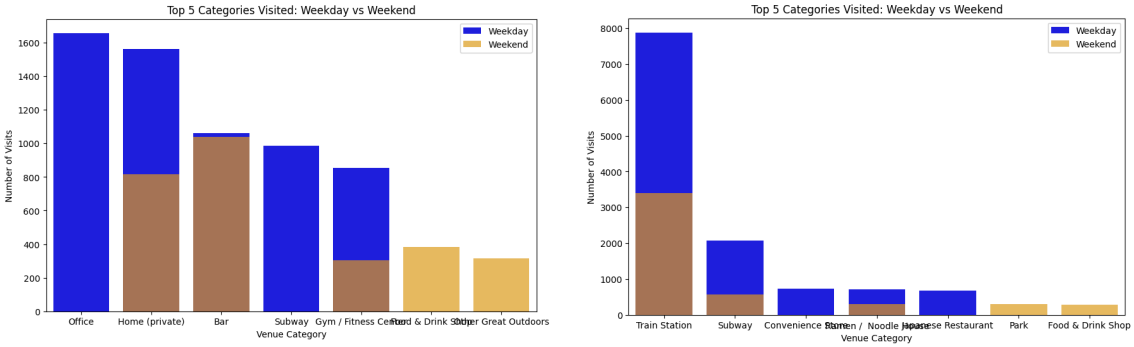
Pada Gambar 7 terlihat bahwa di New York maupun di Tokyo menunjukkan perbandingan jumlah kunjungan antara hari kerja dan akhir pekan. Dari Gambar 7 terlihat bahwa jumlah kunjungan pada hari kerja jauh lebih tinggi dibandingkan akhir pekan, dengan selisih yang cukup signifikan. Hal ini dapat mengindikasikan bahwa mayoritas aktivitas atau kunjungan terjadi selama hari kerja, mungkin karena keterkaitan dengan rutinitas kerja atau keperluan lainnya yang lebih dominan pada periode tersebut. Sebaliknya, pada akhir pekan, jumlah kunjungan menurun, yang mungkin disebabkan oleh pergeseran aktivitas masyarakat ke kegiatan rekreasi atau waktu bersama keluarga.

Kode Program 10

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns

# Pisahkan weekday dan weekend
df1['day_of_week'] =
pd.to_datetime(df1['utcTimestamp']).dt.day_name()
df1['is_weekend'] = df1['day_of_week'].isin(['Saturday', 'Sunday'])

# Jumlah kunjungan weekday vs weekend
visit_counts = df1.groupby('is_weekend')['venueCategory'].count()

# Visualisasi jumlah kunjungan
plt.figure(figsize=(8, 5))
sns.barplot(x=visit_counts.index, y=visit_counts.values,
palette="coolwarm")
plt.xticks([0, 1], ['Weekday', 'Weekend'])
plt.title("Comparison of Visit Counts: Weekday vs Weekend")
plt.ylabel("Number of Visits")
plt.xlabel("Day Type")
plt.show()
```

Kode program 10 tersebut bertujuan untuk menganalisis dan memvisualisasikan jumlah kunjungan berdasarkan hari kerja (Weekday) dan akhir pekan (Weekend) menggunakan data dari check-in user di New York dan Tokyo. Pertama, kode menambahkan dua kolom baru ke dalam DataFrame: day_of_week, yang mengekstraksi nama hari dari kolom ‘utcTimestamp’, dan is_weekend, yang merupakan kolom boolean untuk mengidentifikasi apakah hari tersebut adalah akhir pekan (Saturday atau Sunday). Selanjutnya, jumlah kunjungan dihitung berdasarkan kategori hari (is_weekend) dengan menghitung jumlah nilai dalam kolom ‘venueCategory’ menggunakan metode groupby. Hasil pengelompokan ini divisualisasikan menggunakan grafik batang (bar plot) dengan bantuan pustaka Seaborn. Grafik ini menggunakan palet warna "coolwarm" untuk membedakan antara hari kerja dan akhir pekan, serta diberi label yang jelas pada sumbu X dan Y, dengan judul "Comparison of Visit Counts: Weekday vs Weekend."



Gambar 8. Kunjungan tempat berdasarkan waktu di New York (Kiri) dan di Tokyo (Kanan)

Pada Gambar 8 antara Tokyo dan New York memiliki perbedaan distribusi kunjungan tempat berdasarkan hari. Di New York menunjukkan kategori tempat yang paling sering dikunjungi berdasarkan jumlah kunjungan pada hari kerja (Weekday) dan akhir pekan (Weekend). Kategori tempat yang paling sering dikunjungi adalah kantor (Office) dengan dominasi kunjungan pada hari kerja. Tempat tinggal (Home/private) juga menunjukkan angka kunjungan yang tinggi, dengan distribusi yang lebih merata antara hari kerja dan akhir pekan, kategori bar dan gym/fitness center menunjukkan pola yang menarik. Bar memiliki lebih banyak kunjungan pada akhir pekan, sedangkan gym/fitness center didominasi oleh kunjungan pada hari kerja. Tempat makan dan area outdoor memiliki kunjungan yang lebih rendah dibandingkan dengan kategori lainnya, tetapi proporsi kunjungannya lebih tinggi pada akhir pekan.

Pada Tokyo, kategori dengan jumlah kunjungan tertinggi adalah stasiun kereta (Train Station), dengan dominasi kunjungan pada hari kerja. Hal ini menunjukkan bahwa stasiun kereta merupakan pusat aktivitas transportasi utama, terutama untuk kebutuhan kerja atau perjalanan harian. Subway menempati posisi kedua dengan pola kunjungan yang juga lebih tinggi pada hari kerja, namun dengan selisih yang lebih kecil dibandingkan dengan stasiun kereta. Kategori seperti toko serba ada (Convenience Store) dan restoran (Japanese Restaurant) memiliki volume kunjungan yang jauh lebih rendah, tetapi tetap menunjukkan dominasi kunjungan pada hari kerja. Sementara itu, kategori taman (Park) dan toko makanan dan minuman (Food & Drink Shop) menunjukkan pola kunjungan yang relatif seimbang, dengan kontribusi yang lebih besar dari akhir pekan. Ini mengindikasikan bahwa tempat-tempat ini lebih sering dikunjungi untuk aktivitas rekreasi atau santai pada akhir pekan.

Kode Program 11

```
# Distribusi kategori tempat antara weekday dan weekend
category_weekday =
df2[~df2['is_weekend']]['venueCategory'].value_counts().head(5)
category_weekend =
df2[df2['is_weekend']]['venueCategory'].value_counts().head(5)
```

```
plt.figure(figsize=(10, 6))
sns.barplot(x=category_weekday.index, y=category_weekday.values,
color='blue', label='Weekday')
sns.barplot(x=category_weekend.index, y=category_weekend.values,
color='orange', alpha=0.7, label='Weekend')
plt.title("Top 5 Categories Visited: Weekday vs Weekend")
plt.ylabel("Number of Visits")
plt.xlabel("Venue Category")
plt.legend()
plt.show()
```

Kode program 11 menghasilkan visualisasi berupa grafik batang yang menunjukkan distribusi lima kategori tempat paling sering dikunjungi pada hari kerja (Weekday) dan akhir pekan (Weekend). Distribusi data dihitung berdasarkan kolom 'venueCategory' yang difilter dengan kondisi is_weekend. Untuk hari kerja, jumlah kunjungan ke masing-masing kategori dihitung, begitu pula untuk akhir pekan, kemudian diambil lima kategori teratas dari masing-masing distribusi. Dari grafik yang dihasilkan, warna biru menunjukkan kunjungan pada hari kerja, sementara warna oranye (dengan transparansi) menunjukkan kunjungan pada akhir pekan.

2.3. Kesimpulan

Analisis pola kunjungan di Kota New York dan Tokyo menggunakan dataset check-in menunjukkan perbedaan yang signifikan dalam distribusi aktivitas berdasarkan waktu, lokasi, dan kategori tempat yang sering dikunjungi. Beberapa poin utama yang menjadi sorotan dalam analisis ini adalah:

1. Perbedaan Pola Aktivitas Harian
   - Di New York, aktivitas cenderung memuncak pada malam hari (pukul 20:00–23:00), mencerminkan aktivitas yang aktif di malam hari.
   - Di Tokyo, aktivitas lebih aktif pada pagi hari (pukul 09:00–11:00), menunjukkan fokus pada rutinitas kerja dan aktivitas produktif.
2. Perbedaan Kategori Tempat yang Sering Dikunjungi
   - Di New York, kategori tempat seperti bar, rumah, dan tempat kerja mendominasi kunjungan, menyoroti keseimbangan antara pekerjaan dan kehidupan sosial.
   - Di Tokyo, stasiun kereta api, toko ramen, dan tempat belanja menjadi kategori yang paling sering dikunjungi, mencerminkan budaya mobilitas tinggi dan konsumsi kuliner.
3. Pengaruh Zona Waktu dan Waktu Kunjungan

   Zona waktu mempengaruhi distribusi kunjungan, di mana jam-jam puncak di kedua kota berbeda. Di New York, kunjungan memuncak pada pagi dan malam hari, sementara di Tokyo, puncak kunjungan lebih dominan di pagi hari.
4. Hari dengan Frekuensi Kunjungan Tinggi
   - Di New York, kunjungan memuncak pada Jumat dan Sabtu, dengan penurunan pada Minggu.

- Di Tokyo, kunjungan lebih tinggi pada akhir pekan, dengan puncak pada Minggu.

5. Distribusi Lokasi

   Distribusi lokasi kunjungan di New York menunjukkan kepadatan tinggi di Midtown Manhattan, sedangkan di Tokyo, kepadatan tersebar lebih merata di pusat kota seperti Shinjuku dan Shibuya.

Dari kelima poin utama ini artinya kedua kota memiliki perbedaan budaya, gaya hidup, dan kebiasaan masyarakat. Adapun, Faktor konteks waktu, di New York aktivitas pada hari kerja cenderung terpusat di kantor dan gym, sementara aktivitas akhir pekan lebih terdistribusi ke tempat-tempat hiburan dan area rekreasi. Sedangkan, di Tokyo aktivitas pada hari kerja lebih terfokus pada transportasi dan kebutuhan sehari-hari, sementara akhir pekan lebih cenderung digunakan untuk rekreasi atau kegiatan santai.

Link GitHub:
https://github.com/nadiaftryani/Analisis-Visualisasi-Berdasarkan-Pola-Kunjungan-Terhadap-Waktu-dan-Lokasi-di-Kota-New-York-Dan-Tokyo