Menjelajah Dinamika Tokyo dan New York

Pola Tersembunyi dalam Data Check-In

Anggota Kelompok :

1. Naufal Fakhri 122450089 (Ketua Kelompok)
2. Chevando Daffa Pramanda 122450095 (Anggota Kelompok)
3. Ferdi Kevin 122450107 (Anggota Kelompok)

Deskripsi pembagian pekerjaan:

1. Naufal Fakhri
    - Membuat Tabel Dataset dan statistik deskriptif serta analisis tabel
    - Membuat codingan dan membuat visualisasi
    - Membuat Pertanyaan Analisis Visualisasi
    - Membuat Ide cerita
    - Membuat Infografis
2. Chevando Daffa Pramanda
    - Analisis codingan

3. Ferdi Kevin

    - Analisis Visualisasi

1. Pre-Implementation

1.1 Dataset dan Deskripsi

Dataset ini berisi catatan check-in dari platform Foursquare yang merekam aktivitas pengguna di dua kota besar, Tokyo dan New York City, antara April 2012 hingga Februari 2013

1. **userId**: ID pengguna, semua data pengguna tersedia.
2. **venueId**: ID tempat, tidak ada data tempat yang hilang.
3. **venueCategoryId**: ID kategori tempat, lengkap untuk semua entri.
4. **venueCategory**: Nama kategori tempat, tidak ada nilai yang hilang.
5. **latitude**: Koordinat lintang, data lengkap untuk setiap lokasi check-in.
6. **longitude**: Koordinat bujur, data lengkap tanpa kehilangan nilai.

7. **timezoneOffset**: Perbedaan zona waktu, semua data tersedia.
8. **utcTimestamp**: Stempel waktu dalam UTC, lengkap tanpa null.

Dataset Tokyo (TSMC2014_TKY)

- Terdapat 247 kategori unik, contoh Cosmetic Shop ,Ramen/Noddle House,Convenience Store
- Timezone offset : +540 (JST)

Dataset New York City (TSMC2014_NYC)

- Terdapat 251 Kategori Unik , contoh: Arts & Crafts Store, Bridge, Home (private), Medical Center, Food Truck.
- Timezone offset: -240 (EDT)

New York city

index	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp
0	470	49bbd6c0f964a520f4531fe3	4bf58dd8d48988d127951735	Arts & Crafts Store	4.071.981.038	-7.400.258.103	-240	Tue Apr 03 18:00:09 +0000 2012
1	979	4a43c0aef964a520c6a61fe3	4bf58dd8d48988d1df941735	Bridge	4.060.679.958	-7.404.416.981	-240	Tue Apr 03 18:00:25 +0000 2012
2	69	4c5cc7b485a1e21e00d35711	4bf58dd8d48988d103941735	Home (private)	4.071.616.168	-7.388.307.006	-240	Tue Apr 03 18:02:24 +0000 2012
3	395	4bc7086715a7ef3bef9878da	4bf58dd8d48988d104941735	Medical Center	407.451.638	-7.398.251.878	-240	Tue Apr 03 18:02:41 +0000 2012
4	87	4cf2c5321d18a143951b5cec	4bf58dd8d48988d1cb941735	Food Truck	4.074.010.383	-7.398.965.836	-240	Tue Apr 03 18:03:00 +0000 2012
5	484	4b5b981bf964a520900929e3	4bf58dd8d48988d118951735	Food & Drink Shop	4.069.042.712	-7.395.468.678	-240	Tue Apr 03 18:04:00 +0000 2012
6	642	4ab966c3f964a5203c7f20e3	4bf58dd8d48988d1e0931735	Coffee Shop	4.075.159.143	-739.741.214	-240	Tue Apr 03 18:04:38 +0000 2012
7	292	4d0cc47f903d37041864bf55	4bf58dd8d48988d12b951735	Bus Station	4.077.942.173	-7.395.534.113	-240	Tue Apr 03 18:04:42 +0000 2012
8	428	4ce1863bc4f6a35d8bd2db6c	4bf58dd8d48988d103941735	Home (private)	4.061.915.107	-740.358.876	-240	Tue Apr 03 18:06:18 +0000 2012
9	877	4be319b321d5a59352311811	4bf58dd8d48988d10a951735	Bank	4.061.900.594	-7.399.037.473	-240	Tue Apr 03 18:06:19 +0000 2012

Tokyo city

index	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp
0	1541	4f0fd5a8e4b03856eeb6c8cb	4bf58dd8d48988d10c951735	Cosmetics Shop	3.570.510.109	13.961.959	540	Tue Apr 03 18:17:18 +0000 2012
1	868	4b7b884ff964a5207d662fe3	4bf58dd8d48988d1d1941735	Ramen / Noodle House	3.571.558.112	1.398.003.173	540	Tue Apr 03 18:22:04 +0000 2012
2	114	4c16fdda96040f477cc473a5	4d954b0ea243a5684a65b473	Convenience Store	3.571.454.217	139.480.065	540	Tue Apr 03 19:12:07 +0000 2012
3	868	4c178638c2dfc928651ea869	4bf58dd8d48988d118951735	Food & Drink Shop	3.572.559.199	1.397.766.326	540	Tue Apr 03 19:12:13 +0000 2012
4	1458	4f568309e4b071452e447afe	4f2a210c4b9023bd5841ed28	Housing Development	3.565.608.309	1.397.340.455	540	Tue Apr 03 19:18:23 +0000 2012
5	1541	4b83b207f964a5202c0d31e3	4bf58dd8d48988d1f8941735	Furniture / Home Store	3.570.507.418	1.396.195.023	540	Tue Apr 03 19:20:09 +0000 2012
6	1541	4ea281c302d529c116a57755	4d954b0ea243a5684a65b473	Convenience Store	3.570.627.722	1.396.177.822	540	Tue Apr 03 19:21:00 +0000 2012
7	114	4b3eae5cf964a520b4a025e3	4bf58dd8d48988d129951735	Train Station	3.570.025.263	1.394.802.547	540	Tue Apr 03 19:35:36 +0000 2012
8	1635	4cca7bd67965b60c80f0858a	4bf58dd8d48988d162941735	Other Great Outdoors	3.575.575.922	1.397.335.732	540	Tue Apr 03 19:51:50 +0000 2012
9	2033	4b5c7671f964a5200831 29e3	4bf58dd8d48988d1d1941735	Ramen / Noodle House	3.569.312.098	1.396.994.475	540	Tue Apr 03 19:51:59 +0000 2012

1.1.1 Statistik deskriptif New York CIty

	latitude	Longitude	TimezoneOffset
Count	573.703.000.000	573.703.000.000	573.703.000.000
Mean	35.676.370	139.713.214	539.966.742
Std	0.058958	0.074697	5.451.372
Min	35.510.185	139.470.878	-480.000.000
25%	35.650.417	139.691.390	540.000.000
50%	35.685.867	139.719.274	540.000.000
75%	35.704.023	139.767.101	540.000.000
Max	35.867.150	139.912.593	600.000.000

- **Latitude** dan **Longitude**
  - **Mean** (35.676 untuk latitude dan 139.713 untuk longitude) menunjukkan rata-rata koordinat geografis di dataset ini.
  - **Standar deviasi (Std)** relatif kecil (0.058958 untuk latitude dan 0.074697 untuk longitude), yang berarti variasi lokasi check-in di New York City cenderung terkonsentrasi pada area yang sempit.
  - Rentang nilai (Min dan Max)
    - Latitude: 35.510 hingga 35.867.
    - Longitude: 139.471 hingga 139.913.

- **Kuartil (25%, 50%, 75%)**:

Rentang antar kuartil menunjukkan distribusi lokasi check-in yang tidak terlalu menyebar, mempertegas konsentrasi check-in di area tertentu

- **TimezoneOffset**:
    - Mean sebesar **539.966.742** menunjukkan rata-rata pergeseran zona waktu, kemungkinan ini dalam satuan detik atau offset tertentu.
    - Min dan Max memiliki rentang yang luas (-480.000 hingga 600.000), yang mencerminkan adanya variasi signifikan dalam zona waktu pada dataset.

1.1.2 Statistika deskriptif Tokyo City

	latitude	Longitude	TimezoneOffset
Count	227.428.000.000	227.428.000.000	227.428.000.000
Mean	40.754.045	-73.974.556	-253.392.019
Std	0.058958	0.086209	43.234.750
Min	40.550.852	-74.274.766	-420.000.000
25%	40.718.330	-74.000.633	-240.000.000
50%	40.747.745	-73.983.479	-240.000.000
75%	40.778.374	-73.945.709	-240.000.000
Max	40.988.332	-73.683.825	600.000.000

- **Latitude** dan **Longitude**:
    - Mean latitude (**40.754**) dan longitude (**-73.975**) berbeda jauh dari New York City, karena lokasi geografisnya berada di hemisfer yang berbeda.
    - Standar deviasi latitude (**0.058958**) dan longitude (**0.086209**) menunjukkan pola yang mirip dengan New York City—check-in terkonsentrasi di lokasi tertentu
    - **Rentang nilai**:
        - Latitude: 40.551 hingga 40.988.
        - Longitude: -74.275 hingga -73.684.
    - Distribusi kuartil serupa dengan New York City, menunjukkan penyebaran check-in yang tidak terlalu luas.
- **TimezoneOffset**:
    - Mean timezone offset (**-253.392.019**) menunjukkan rata-rata zona waktu di Tokyo, yang jauh berbeda dari New York City, sesuai dengan perbedaan geografis.

- Standar deviasi yang tinggi (**43.234.750**) mengindikasikan adanya keragaman data check-in dari zona waktu yang berbeda.

Perbandingan

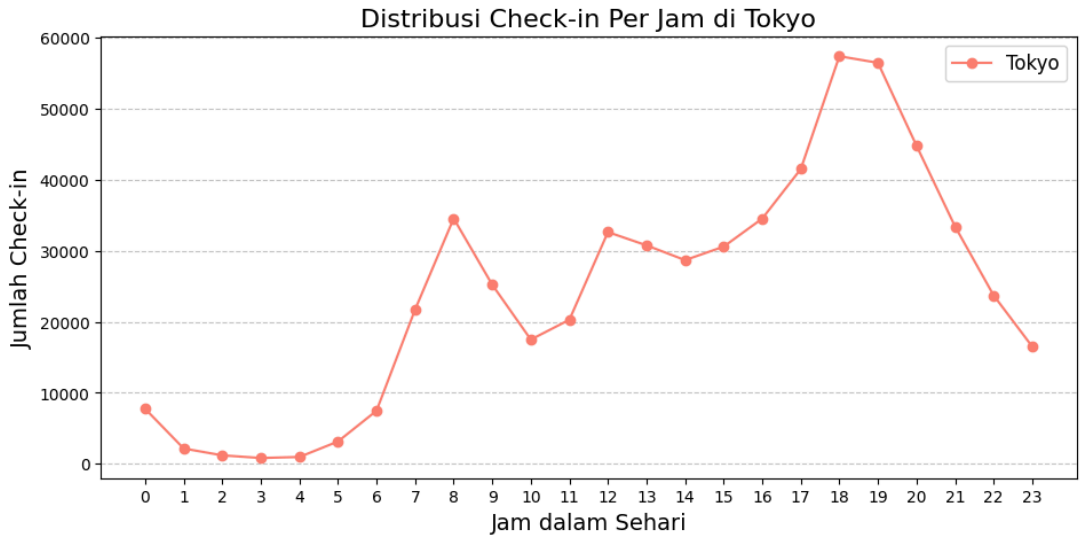
- Koordinat Geografis
  - Distribusi latitude dan longitude di kedua kota menunjukkan konsentrasi check-in pada area tertentu, tetapi lokasi geografis jelas berbeda karena posisi hemisfer.
- Variasi Data
  - Tokyo memiliki standar deviasi longitude yang sedikit lebih besar dibandingkan New York, yang dapat menunjukkan lokasi check-in lebih menyebar dibandingkan dengan New York City
- Timezone Offset
  - New York memiliki mean timezone offset yang jauh lebih tinggi, sedangkan Tokyo memiliki nilai negatif. Hal ini sesuai dengan lokasi zona waktu global (GMT+ di Tokyo vs GMT- di New York)

Analisis statistik deskriptif menunjukkan perbedaan distribusi koordinat geografis (latitude dan longitude) serta perbedaan **timezoneOffset** antara Tokyo dan NYC. Rata-rata koordinat sesuai dengan lokasi geografis masing-masing kota, sementara variabilitas yang rendah menunjukkan distribusi lokasi check-in yang terkonsentrasi pada area tertentu di kedua kota.

Visualisasi

1. Line Chart: Distribusi Check-in Per Jam di Tokyo

Visualisasi pola check-in selama 24 jam untuk Tokyo dan NYC secara terpisah.

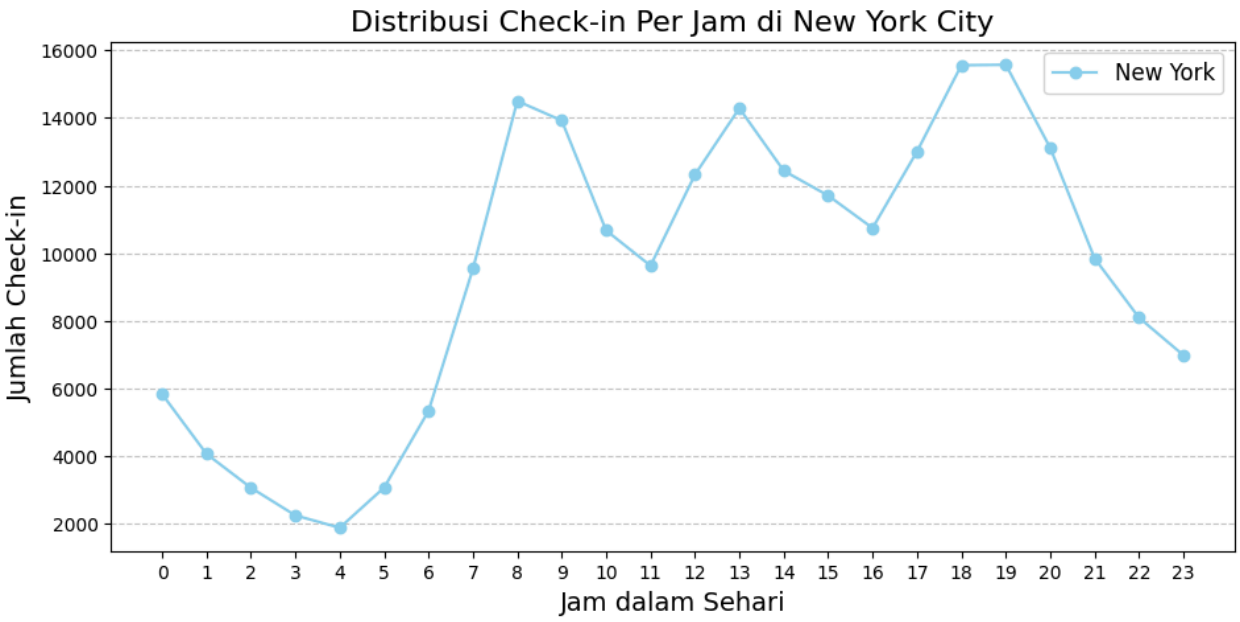


Visualisasi ini menunjukkan Distribusi Check-in Per Jam di Tokyo dalam bentuk Line Chart.

- **Pola:**
- Aktivitas check-in mulai meningkat signifikan setelah pukul 6 pagi.
- Terdapat puncak pertama pada pukul 8 pagi, di mana jumlah check-in mencapai titik tinggi.
- Aktivitas menurun setelah pukul 9 pagi hingga sekitar pukul 13 siang.
- Setelah itu, terjadi peningkatan secara bertahap mulai dari pukul 14 siang.
- Puncak kedua:
- Jumlah check-in kembali mencapai puncak tertinggi antara pukul 18 hingga 20 malam.
- Setelah pukul 20, jumlah check-in mulai menurun drastis.

- **Jam sibuk:**
- Jam sibuk check-in tampaknya berada di pagi hari (sekitar pukul 7-8) dan sore hingga malam (pukul 18-20).
- Aktivitas relatif rendah pada dini hari hingga subuh (pukul 0-5).

  Aktivitas check-in menunjukkan pola yang berhubungan dengan rutinitas harian, seperti jam sibuk kerja di pagi hari dan aktivitas sosial atau rekreasi di malam hari.



Line chart Distribusi Cek in per jam di New York

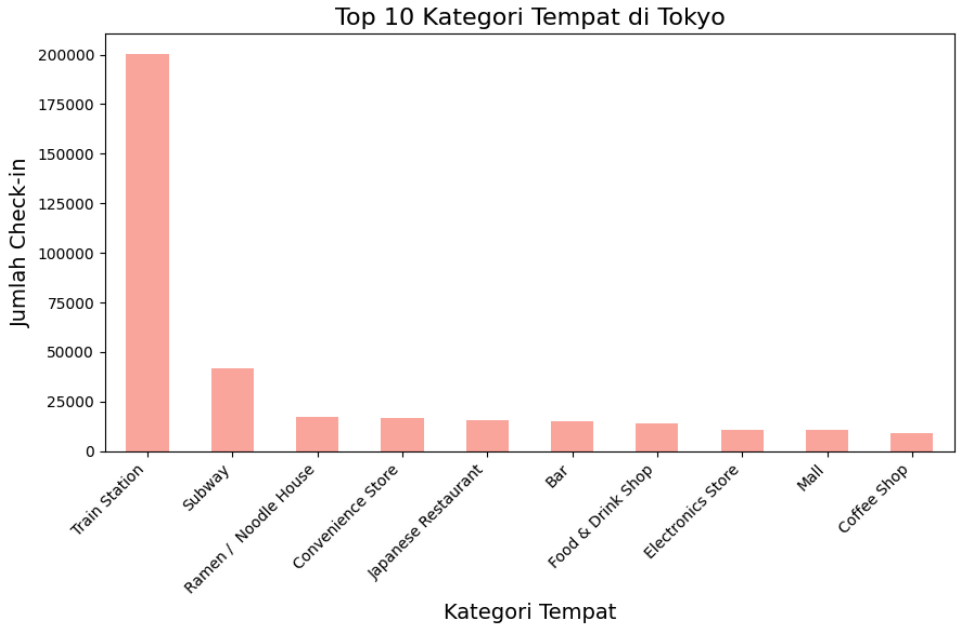
Grafik ini menunjukkan Distribusi Check-in Per Jam di New York City dalam bentuk line chart.

1. Pola:
    - Aktivitas check-in mulai meningkat secara signifikan sekitar pukul 6 pagi.
    - Puncak pertama terjadi pada pukul 8 pagi, dengan jumlah check-in yang tinggi.
    - Aktivitas menurun setelah pukul 9 pagi hingga sekitar pukul 12 siang.
    - Setelah itu, aktivitas naik-turun di siang hari dengan peningkatan yang stabil mulai pukul 17 sore.
2. Puncak kedua:
    - Jumlah check-in mencapai puncak tertinggi kedua pada pukul 18 hingga 20 malam.
    - Setelah pukul 20 malam, jumlah check-in menurun tajam hingga mendekati pukul tengah malam.
3. Jam sibuk:
    - Waktu sibuk check-in di New York terjadi pada pagi hari (sekitar pukul 7-9) dan sore hingga malam hari (pukul 18-20).
4. Perbandingan dengan Tokyo:
    - Dibandingkan dengan Tokyo, jumlah check-in di New York lebih rendah secara keseluruhan.
    - Pola waktu puncaknya cukup mirip, dengan dua puncak utama di pagi dan malam hari.

    Pola ini mencerminkan aktivitas harian di New York yang juga terkait dengan jam kerja dan aktivitas rekreasi di malam hari. Waktu-waktu ini bisa digunakan untuk optimasi bisnis seperti penawaran khusus atau promosi selama jam sibuk tersebut

2. Bar Chart: Top 10 Kategori Tempat

Menampilkan kategori tempat paling populer di Tokyo dan NYC secara terpisah.

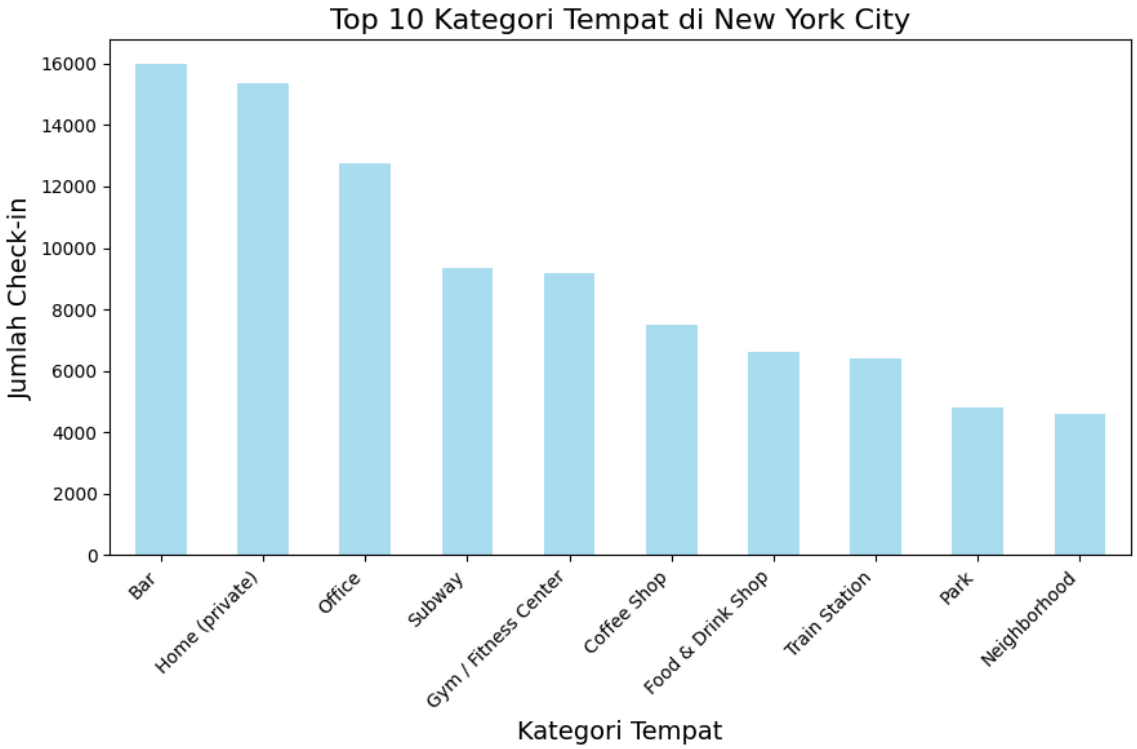


Visualisasi ini menampilkan Top 10 Kategori Tempat di Tokyo berdasarkan jumlah check-in.

1. Kategori Dominan
    - Train Station adalah kategori yang paling dominan, dengan jumlah check-in jauh lebih tinggi dibandingkan kategori lainnya, mencapai hampir 200.000.
    - Hal ini mencerminkan tingginya aktivitas di stasiun kereta di Tokyo, yang mungkin disebabkan oleh tingginya penggunaan transportasi umum di kota tersebut.
2. Kategori Lain
    - Kategori kedua adalah Subway, dengan jumlah check-in yang cukup signifikan tetapi jauh di bawah stasiun kereta.
    - Ramen/Noodle House berada di peringkat ketiga, menunjukkan popularitas tempat makan ramen sebagai destinasi favorit.
    - Kategori seperti Convenience Store, Japanese Restaurant, dan Bar memiliki jumlah check-in yang cukup seimbang, mencerminkan aktivitas sosial dan kebutuhan sehari-hari.
3. Kategori dengan Check-in Lebih Rendah
    - Electronics Store, Mall, dan Coffee Shop berada di posisi terbawah dalam daftar, meskipun masih cukup populer.

4. Kesimpulan
    - Data ini menunjukkan bahwa transportasi publik merupakan aspek utama aktivitas masyarakat di Tokyo, diikuti oleh aktivitas makan dan belanja.

- Bisnis di sekitar stasiun kereta dan subway memiliki peluang besar untuk menarik pengunjung karena tingginya jumlah check-in.



Barchart Kategori tempat di New York City

Visualisasi ini menampilkan Top 10 Kategori tempat di New York City berdasarkan jumlah check-in.

1. Kategori Tempat dengan Jumlah Check-in Tertinggi
   - Bar menempati posisi pertama dengan jumlah check-in tertinggi, menunjukkan bahwa bar adalah salah satu tempat populer di New York City untuk aktivitas sosial.
   - Home (private) berada di posisi kedua, yang mencerminkan tingginya aktivitas check-in dari lokasi pribadi, mungkin dilakukan oleh pengguna aplikasi untuk melacak lokasi mereka sendiri.
2. Kategori Tempat Kerja dan Transportasi
   Office dan Subway memiliki jumlah check-in yang signifikan, menunjukkan frekuensi tinggi aktivitas yang terkait dengan pekerjaan dan transportasi.
3. Gaya Hidup dan Hiburan
   - Tempat-tempat seperti Gym/Fitness Center, Coffee Shop, dan Food & Drink Shop juga memiliki jumlah check-in yang tinggi, menyoroti minat masyarakat terhadap kebugaran, kopi, dan makanan.

4. Transportasi Publik dan Alam
    - Train Station memiliki angka check-in yang cukup tinggi, mencerminkan tingginya penggunaan transportasi publik.
    - Park memiliki angka check-in yang lebih rendah, tetapi tetap relevan sebagai tempat rekreasi.
5. Kesimpulan Umum
    - Pola check-in menunjukkan keseimbangan antara tempat hiburan, lokasi kerja, dan aktivitas pribadi.
    - Bar, tempat kerja, dan transportasi menjadi kategori dengan check-in paling banyak, mencerminkan pola aktivitas masyarakat di kota besar seperti New York City.

1.2 Pertanyaan Analisis Visualisasi

1. Pola Aktivitas dan Waktu Check-In
    - Apakah aktivitas check-in berbeda secara signifikan antara hari kerja dan akhir pekan sepanjang hari?
2. Preferensi Jenis Tempat
    - Bagaimana distribusi kunjungan berdasarkan kategori tempat di kedua kota? (Apakah kategori tertentu lebih dominan di Tokyo atau NYC?)
3. Tren Berdasarkan Hari Kerja dan Akhir Pekan
    - Apakah ada perbedaan kategori tempat yang lebih sering dikunjungi pada hari kerja vs akhir pekan?
4. Pendekatan Keramaian dan Interaksi Sosial
    - Apakah terdapat perbedaan kepadatan check-in di pusat kota dan area pinggiran?
5. Pertanyaan Distribusi Periode Waktu
    - Apakah ada perbedaan aktivitas check-in berdasarkan periode waktu (pagi, siang, sore, malam)?

1.3 Ide Cerita

Melalui data check-in, kita dapat melihat dinamika kehidupan di dua kota besar dunia, Tokyo dan New York. Visualisasi ini tidak hanya menunjukkan angka, tetapi juga menceritakan bagaimana budaya, waktu, dan preferensi masyarakat di kedua kota ini berbeda.

Pola Aktivitas dan Waktu Check-In

- New York: Masyarakat aktif hingga larut malam, terutama di pusat hiburan seperti bar dan restoran, mencerminkan budaya “kota yang tidak pernah tidur”. Aktivitas check-in mencapai puncak di malam hari.

- Tokyo: Aktivitas lebih teratur dengan puncak di sore hingga awal malam, mencerminkan rutinitas yang disiplin. Tokyo menunjukkan keseimbangan antara kehidupan kerja dan waktu pribadi.

Preferensi Jenis Tempat

- New York: Lebih condong ke tempat hiburan modern seperti restoran, bar, galeri seni, dan pusat perbelanjaan, menggambarkan gaya hidup urban yang dinamis.
- Tokyo: Menonjolkan lokasi tradisional seperti kuil, taman, dan stasiun kereta, yang menunjukkan penghormatan terhadap budaya dan keteraturan sehari-hari.

Puncak Aktivitas Berdasarkan Jam dan Kategori

- New York: Bar dan restoran mencapai puncak aktivitas di malam hari, ketika masyarakat menikmati waktu rekreasi.
- Tokyo: Kafe dan taman lebih ramai di pagi hingga sore, menunjukkan preferensi untuk aktivitas santai di siang hari.

Tren Hari Kerja vs. Akhir Pekan

- New York: Tempat hiburan malam seperti bar dan klub cenderung lebih ramai di akhir pekan. Hal ini menggambarkan masyarakat yang memanfaatkan akhir pekan untuk bersosialisasi.
- Tokyo: Akhir pekan diisi dengan aktivitas yang lebih santai seperti mengunjungi kafe dan pusat perbelanjaan, menunjukkan preferensi untuk waktu istirahat berkualitas.

Pendekatan Terhadap Keramaian dan Interaksi Sosial

- New York: Kota ini menunjukkan kenyamanan dengan keramaian, di mana masyarakat berkumpul di pusat hiburan dan lokasi sosial.
- Tokyo: Sebaliknya, Tokyo lebih menghargai ruang pribadi, dengan check-in yang lebih sering di tempat-tempat dengan keteraturan, seperti taman atau transportasi umum.

Aktivitas Berdasarkan Musim

- Tokyo: Aktivitas outdoor seperti mengunjungi taman lebih menonjol di musim semi dan gugur, ketika cuaca mendukung kegiatan di luar ruangan.
- New York: Tempat indoor seperti museum dan pusat perbelanjaan menjadi pilihan utama di musim dingin, menunjukkan adaptasi terhadap musim

Tren Check-in Berdasarkan Kategori dan Waktu

- New York: Bar dan restoran populer di malam hari, sementara tempat transportasi seperti subway lebih sibuk di pagi hari.
- Tokyo: Kafe dan taman lebih sering dikunjungi di pagi hingga sore, menunjukkan pola aktivitas yang lebih terstruktur.

2. Implementation

2.1 Preprocessing Dataset

2.1.1 Input Data

```
data_tky = pd.read_csv('dataset_TSMC2014_TKY.csv')
data_nyc = pd.read_csv('dataset_TSMC2014_NYC.csv')
data_tky.head(5)
data_nyc.head(5)
```

Kode ini digunakan untuk memuat dan meninjau dataset dalam format CSV menggunakan library pandas. Terdapat dua dataset, yaitu 'dataset_TSMC2014_TKY.csv' dan 'dataset_TSMC2014_NYC.csv', masing-masing data tersebut dimuat ke dalam DataFrame bernama 'data_tky' dan 'data_nyc' dengan menggunakan fungsi pd.read_csv(). Dataset berisikan data terkait aktivitas pada suatu lokasi di Tokyo dan New York City.

Fungsi head(5) dipanggil untuk menampilkan lima baris pertama dari masing-masing dataset. Hal ini bertujuan untuk memberikan gambaran awal terkait struktur dan isi data, termasuk kolom yang tersedia, jenis data, dan sampel nilainya.

2.1.2 Mengecek Missing Value

```
print("Null values in Tokyo dataset:")
print(data_tky.isnull().sum())
print("\nNull values in NYC dataset:")
print(data_nyc.isnull().sum())
```

Kode ini digunakan untuk memeriksa keberadaan nilai null dalam dataset data_tky dan data_nyc. Fungsi isnull() pada DataFrame digunakan untuk mengidentifikasi elemen null pada setiap kolom, menghasilkan nilai boolean (True untuk null dan False untuk bukan null(). Fungsi sum() menghitung jumlah nilai null di setiap kolom dengan menjumlahkan semua nilai True

2.1.3 Menghapus Missing Value

```
data_tky = data_tky.dropna()
data_nyc = data_nyc.dropna()

print("Null values in Tokyo dataset:")
print(data_tky.isnull().sum())
print("\nNull values in NYC dataset:")
print(data_nyc.isnull().sum())
```

fungsi dropna() digunakan pada DataFrame "data_tky" dan "data nyc" untuk menghapus baris yang terdapat nilai null. Hasilnya, dataset baru tanpa nilai kosong menggantikan dataset awal. Setelah nilai null dihapus, selanjutnya memeriksa ulang keberadaan nilai null dengan memanfaatkan fungsi isnull() untuk mendeteksi element null, diikuti oleh sum() untuk menghitung jumlah nilai kosong di setiap kolom.

Output:

```
Null values in Tokyo dataset:
userId              0
venueId             0
venueCategoryId     0
venueCategory       0
latitude            0
longitude           0
timezoneOffset      0
utcTimestamp        0
dtype: int64
```

```
Null values in NYC dataset:
userId              0
venueId             0
venueCategoryId     0
venueCategory       0
latitude            0
longitude           0
timezoneOffset      0
utcTimestamp        0
dtype: int64
```

Output menunjukkan bahwa baik dataset Tokyo maupun New York City tidak memiliki nilai null di semua kolomnya, seperti **userId**, **venueId**, **venueCategoryId**, **venueCategory**, **latitude**, **longitude**, **timezoneOffset**, dan **utcTimestamp**. Hal ini menandakan bahwa data bersih dan lengkap, sehingga tidak memerlukan langkah tambahan untuk menangani nilai yang hilang. Dengan struktur data yang seragam dan bebas dari missing values, dataset ini siap untuk analisis lebih lanjut, seperti eksplorasi data, visualisasi, atau penerapan model analitik tanpa risiko bias akibat data yang tidak lengkap.

2.1.4 Menambahkan waktu lokal

```
data_tky['utcTimestamp'] = pd.to_datetime(data_tky['utcTimestamp'])
data_nyc['utcTimestamp'] = pd.to_datetime(data_nyc['utcTimestamp'])
data_tky['localTime'] = data_tky['utcTimestamp'] + pd.to_timedelta(data_tky['timezoneOffset'], unit='m')
data_nyc['localTime'] = data_nyc['utcTimestamp'] + pd.to_timedelta(data_nyc['timezoneOffset'], unit='m')
```

Kode ini digunakan untuk menambahkan kolom baru bernama 'localTime' ke dalam dataset 'data_tky' dan 'data_nyc' dengan menghitung waktu lokal berdasarkan waktu UTC 'utcTimestamp' dan selisih zona waktu 'timezoneOffset'. Kolom 'utcTimestamp' dikonversi ke format datetime menggunakan fungsi 'pd.to_datetime()' untuk memastikan data waktu dapat dimanipulasi dengan mudah. Kemudian, nilai 'timezoneOffset', yang berisi selisih waktu zona dalam satuan menit, diubah menjadi objek timedelta menggunakan fungsi 'pd.to_timedelta()'. Objek ini ditambahkan ke kolom 'utcTimestamp' untuk menghasilkan waktu lokal yang disimpan dalam kolom baru 'localTime'. Hasilnya, setiap baris dalam dataset memiliki informasi waktu lokal yang akurat, memungkinkan analisis berbasis waktu yang lebih relevan untuk lokasi Tokyo dan New York City.

2.1.5 Mengubah data waktu menjadi waktu lokal

```
data_tky['hour'] = data_tky['localTime'].dt.hour
data_tky['day_of_week'] = data_tky['localTime'].dt.dayofweek
data_nyc['hour'] = data_nyc['localTime'].dt.hour
data_nyc['day_of_week'] = data_nyc['localTime'].dt.dayofweek
```

Kode ini memproses data waktu pada dua dataset bernama 'data_tky' dan 'data_nyc'. Kode ini menambahkan dua kolom baru pada masing-masing dataset, yaitu 'hour' dan 'day_of_week', yang dihasilkan dengan memanfaatkan kolom waktu 'localTime'. Untuk setiap baris data, kolom 'hour' diisi dengan informasi jam (dalam format 24-jam) yang diekstrak dari nilai waktu pada kolom 'localTime', sementara kolom 'day_of_week' diisi dengan informasi hari dalam bentuk angka (0 untuk Senin, 6 untuk Minggu). Proses ini dilakukan baik pada dataset 'data_tky' (mungkin mewakili Tokyo) maupun 'data_nyc' (mungkin mewakili New York City).

2.1.6 Menambahkan kolom periode waktu lokal

```
def time_period(hour):
    if 5 <= hour < 12:
        return 'Pagi'
    elif 12 <= hour < 17:
        return 'Siang'
    elif 17 <= hour < 22:
        return 'Sore'
    else:
        return 'Malam'

data_tky['period'] = data_tky['hour'].apply(time_period)
data_nyc['period'] = data_nyc['hour'].apply(time_period)
```

Kode ini mendefinisikan sebuah fungsi bernama ‘time_period’ yang digunakan untuk mengelompokkan waktu berdasarkan jam dalam kategori periode tertentu: pagi, siang, sore, atau malam. Fungsi ini menerima parameter ‘hour’ (jam dalam format 24-jam) dan mengembalikan kategori waktu berdasarkan logika berikut:

- Jika jam berada di antara 5 hingga kurang dari 12, maka dikategorikan sebagai "Pagi".
- Jika jam berada di antara 12 hingga kurang dari 17, maka dikategorikan sebagai "Siang".
- Jika jam berada di antara 17 hingga kurang dari 22, maka dikategorikan sebagai "Sore".
- Untuk semua waktu lainnya (22 hingga sebelum 5), dikategorikan sebagai "Malam".

Setelah mendefinisikan fungsi ini, kode tersebut menggunakannya untuk menambahkan kolom baru bernama ‘period’ ke dalam dataset ‘data_tky’ dan ‘data_nyc’ Kolom ini diisi dengan kategori periode waktu berdasarkan nilai jam yang ada di kolom ‘hour’ dengan menggunakan metode ’apply’ untuk menerapkan fungsi ‘time_period’ pada setiap baris.

2.1.7 Mengkombinasi dataset untuk Komparasi

```
data_tky['city'] = 'Tokyo'
data_nyc['city'] = 'New York'
data_combined = pd.concat([data_tky, data_nyc], ignore_index=True)
```

Kode ini menggabungkan dua dataset, yaitu ‘data_tky’ dan ‘data_nyc’, setelah menambahkan kolom baru bernama ‘city’ pada masing-masing dataset. Kolom ‘city’

pada dataset ‘data_tky’ diisi dengan nilai ‘Tokyo’, menandakan bahwa data tersebut berasal dari Tokyo, sedangkan pada dataset ‘data_nyc’, kolom ini diisi dengan nilai ‘New York’, menunjukkan asal data dari New York. Setelah itu, kedua dataset digabungkan menggunakan fungsi ‘pd.concat()’ dengan parameter ‘ignore_index=True’, yang memastikan bahwa indeks pada dataset gabungan, ‘data_combined’, diatur ulang menjadi numerik kontinu tanpa mempertahankan indeks asli dari dataset masing-masing. Dataset hasil gabungan ini memuat informasi dari kedua kota, lengkap dengan label kota pada kolom ‘city’, sehingga memudahkan analisis lintas lokasi dalam satu struktur data yang terpadu

2.1.8 membagi data latih dan data uji

```
train_tky, test_tky = train_test_split(data_tky, test_size=0.2, random_state=42)
train_nyc, test_nyc = train_test_split(data_nyc, test_size=0.2, random_state=42)

print(f"Tokyo training set size: {len(train_tky)}")
print(f"Tokyo testing set size: {len(test_tky)}")
print(f"NYC training set size: {len(train_nyc)}")
print(f"NYC testing set size: {len(test_nyc)}")
```

Kode Python tersebut membagi dataset ‘data_tky’ dan ‘data_nyc’ menjadi dua bagian: set pelatihan (training) dan set pengujian (testing) menggunakan fungsi ‘train_test_split’ dari pustaka scikit-learn. Untuk dataset Tokyo (‘data_tky’), 80% data dialokasikan sebagai set pelatihan (‘train_tky’), sedangkan 20% sisanya digunakan sebagai set pengujian (‘test_tky’). Pembagian yang sama dilakukan untuk dataset New York (‘data_nyc’), menghasilkan ‘train_nyc’ dan ‘test_nyc’. Parameter ‘random_state=42’ digunakan untuk memastikan bahwa pembagian dilakukan secara deterministik sehingga hasilnya dapat direproduksi. Setelah pembagian, ukuran masing-masing set dicetak untuk memverifikasi bahwa data telah dibagi dengan proporsi yang sesuai, memberikan informasi tentang jumlah data dalam set pelatihan dan pengujian untuk kedua dataset.

Output:

```
Tokyo training set size: 458962
Tokyo testing set size: 114741
NYC training set size: 181942
NYC testing set size: 45486
```

Dataset Tokyo memiliki jumlah data pelatihan (458,962) dan pengujian (114,741) yang jauh lebih besar dibandingkan NYC (181,942 untuk pelatihan dan 45,486 untuk pengujian), dengan rasio pembagian 80:20 untuk keduanya,Perbedaan ukuran ini dapat

memberikan keunggulan pada model Tokyo dalam menangkap pola data yang lebih kompleks.

2.1.9 melakukan save hasil preprocessing data latih dan uji

```
train_tky.to_csv('train_tky.csv', index=False)
test_tky.to_csv('test_tky.csv', index=False)
train_nyc.to_csv('train_nyc.csv', index=False)
test_nyc.to_csv('test_nyc.csv', index=False)
```

Kode ini menggunakan metode .to_csv() dari pustaka pandas untuk mengekspor dataset ‘train_tky’, ‘test_tky’, ‘train_nyc’, dan ‘test_nyc’ ke file CSV. Dataset pelatihan untuk Tokyo disimpan sebagai file ‘train_tky.csv’, sedangkan dataset pengujian untuk Tokyo disimpan sebagai ‘test_tky.csv’. Hal serupa dilakukan untuk dataset New York, di mana dataset pelatihan disimpan dalam file ‘train_nyc.csv’ dan dataset pengujian dalam file ‘test_nyc.csv’. Parameter ‘index=False’ digunakan untuk memastikan bahwa indeks dari DataFrame tidak disertakan dalam file CSV yang dihasilkan. Proses ini memungkinkan penyimpanan hasil pembagian dataset ke dalam format yang dapat diakses atau digunakan di luar lingkungan Python.

Data test Tokyo City

index	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp	localTime	hour	day_of_week	period	venueCategoryGrouped	city
0	1251	4e2183d918a88345f04c0e1e	4bf58dd8d48988d129951735	Train Station	35.56263056	139.716053	540	2012-04-28 06:03:14+00:00	2012-04-28 15:03:14+00:00	15	5	Siang	Other	Tokyo
1	422	4e02bdf0b0fb88a1209b637b	4bf58dd8d48988d172941735	Post Office	35.70960439	139.7971872	540	2012-12-05 07:48:01+00:00	2012-12-05 16:48:01+00:00	16	2	Siang	Other	Tokyo
2	1575	4b0bc75ff964a520923323e3	4bf58dd8d48988d114951735	Bookstore	35.70042688	139.7717518	540	2012-05-12 08:13:14+00:00	2012-05-12 17:13:14+00:00	17	5	Sore	Other	Tokyo
3	1492	4b1ac490f964a5205af123e3	4bf58dd8d48988d1fd931735	Subway	35.72949574	139.7913051	540	2012-06-13 23:19:38+00:00	2012-06-14 08:19:38+00:00	8	3	Pagi	Other	Tokyo
4	2150	4b77a899f964a520e0a52ee3	4bf58dd8d48988d1df941735	Bridge	35.67918693	139.7815901	540	2012-07-02 14:30:11+00:00	2012-07-02 23:30:11+00:00	23	0	Malam	Other	Tokyo

1. Lokasi Check-In
    - Dataset menunjukkan lokasi check-in di Tokyo, berdasarkan koordinat latitude dan longitude.
2. Kategori Tempat
    - Kategori tempat meliputi lokasi umum seperti: Train Station, Post Office, Bookstore, Subway, dan Bridge. Sebagian besar tergolong "Other" pada kolom pengelompokan kategori.
3. Periode Waktu

    - Periode waktu check-in bervariasi, mulai dari pagi (Subway), siang (Train Station, Post Office), sore (Bookstore), hingga malam (Bridge).
4. Distribusi Hari
    - Hari dalam seminggu menunjukkan aktivitas check-in pada berbagai hari (dari 0 hingga 6). Sebagai contoh:
        - Train Station: Hari Sabtu (day_of_week = 5).
        - Post Office: Hari Selasa (day_of_week = 2).
5. Zona Waktu
    - Waktu lokal dan UTC menunjukkan perbedaan zona waktu sebesar 540 menit (+9 jam), yang sesuai dengan Tokyo.

Insight Potensial:

- Kategori Populer: Dengan analisis tambahan, kita dapat mengetahui kategori tempat yang paling sering dikunjungi berdasarkan jumlah check-in.
- Pola Waktu: Jam dan hari check-in dapat membantu mengidentifikasi waktu puncak aktivitas di lokasi tertentu.
- Distribusi Lokasi: Dengan koordinat geografis, peta visual dapat dibuat untuk memvisualisasikan penyebaran check-in.

Data test New York City

userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp	localTime	hour	day_of_week	period	venueCategoryGrouped	city
419	4a6f8c85f964a52079d61fe3	4bf58dd8d48988d1f6941735	Department Store	40.792629	-74.041886	-240	2012-06-18 10:17:29+00:00	2012-06-18 06:17:29+00:00	6	0	Pagi	Other	New York
13	4daddba56a23e6c9347e1092	4bf58dd8d48988d1a3941735	College Academic Building	40.911190	-73.907111	-300	2012-12-15 19:30:04+00:00	2012-12-15 14:30:04+00:00	14	5	Siang	Other	New York
349	4f79a81ae4b09489387bafb6	4bf58dd8d48988d146941735	Deli / Bodega	40.832612	-73.915352	-300	2013-01-26 13:26:41+00:00	2013-01-26 08:26:41+00:00	8	5	Pagi	Other	New York
216	4dbb09d5ffcbcd48566371bf	4c38df4de52ce0d596b336e1	Parking	40.763384	-73.985095	-240	2012-05-15 22:59:41+00:00	2012-05-15 18:59:41+00:00	18	1	Sore	Other	New York
545	4dab7425fa8cc764974da770	4f2a25ac4b909258e854f55f	Neighborhood	40.682690	-73.931737	-240	2012-07-15 13:42:05+00:00	2012-07-15 09:42:05+00:00	9	6	Pagi	Other	New York

Gambar ini adalah contoh dataframe yang menampilkan data check-in di New York City. Berikut adalah penjelasan kolom-kolom yang ada dalam tabel tersebut:

Kolom-Kolom:

1. userId:
    - ID unik pengguna yang melakukan check-in.
2. venueId:
    - ID unik dari tempat (venue) di mana check-in dilakukan.
3. venue CategoryId:
    - ID unik untuk kategori tempat tersebut.
4. venueCategory:
    - Kategori tempat, seperti "Department Store", "College Academic Building", dll.
5. latitude dan longitude:
    - Koordinat geografis lokasi tempat check-in.
6. timezoneOffset:
    - Selisih waktu dalam menit dari UTC (zona waktu tempat check-in).
7. utcTimestamp:
    - Waktu check-in dalam format UTC.
8. localTime:
    - Waktu check-in dalam zona waktu lokal.
9. hour:
    - Jam dalam format 24-jam ketika check-in dilakukan.
10. day_of_week:
    - Hari dalam minggu (0 = Senin, 6 = Minggu).
11. period:
    - Periode waktu berdasarkan jam, seperti:
        - Pagi (6-11)
        - Siang (12-15)
        - Sore (16-18)
        - Malam (19-23).
12. venueCategoryGrouped:
    - Kategori tempat yang dikelompokkan (misalnya, "Other" untuk kategori yang kurang umum).
13. city:
    - Kota di mana check-in dilakukan (semua entri pada contoh ini adalah New York).

Observasi:

- Data Multi-dimensi: Informasi ini memberikan detail lokasi, waktu, dan kategori tempat check-in, memungkinkan analisis seperti:
    - Distribusi check-in berdasarkan waktu (siang/malam).
    - Lokasi dengan check-in terbanyak.
    - Kategori tempat populer.

- Fokus Kota: Data ini berfokus pada New York City, tetapi formatnya dapat digunakan untuk kota lain.

Data Latih Tokyo city

	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp	localTime	hour	day_of_week	period	venueCategoryGrouped	city
1614	4b243a7df964a520356424e3	4bf58dd8d48988d129951735	Train Station	35.729865	139.710956	540	2013-01-30 14:59:22+00:00	2013-01-30 23:59:22+00:00	23	2	Malam	Other	Tokyo
437	4b554874f964a520bde027e3	4bf58dd8d48988d1f8941735	Furniture / Home Store	35.646995	139.517577	540	2012-06-07 03:22:25+00:00	2012-06-07 12:22:25+00:00	12	3	Siang	Other	Tokyo
1648	4d7dfe62cd09224bd4934130	4bf58dd8d48988d1fe931735	Bus Station	35.655666	139.711070	540	2012-06-07 11:59:39+00:00	2012-06-07 20:59:39+00:00	20	3	Sore	Other	Tokyo
2057	4e60e20afa76cd64cd8a6022	4bf58dd8d48988d1fd941735	Mall	35.753287	139.709707	540	2012-11-18 07:36:50+00:00	2012-11-18 16:36:50+00:00	16	6	Siang	Modern Entertainment	Tokyo
1699	4b5bf03cf964a520131e29e3	4bf58dd8d48988d122951735	Electronics Store	35.699299	139.769603	540	2012-04-22 04:01:53+00:00	2012-04-22 13:01:53+00:00	13	6	Siang	Other	Tokyo

Data Latih New York City

	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp	localTime	hour	day_of_week	period	venueCategoryGrouped	city
213	4ba37c86f964a520414138e3	4bf58dd8d48988d139941735	Synagogue	40.773157	-73.955198	-240	2012-05-27 13:57:50+00:00	2012-05-27 09:57:50+00:00	9	6	Pagi	Other	New York
26	451ad4f2f964a5206a3a1fe3	4bf58dd8d48988d116941735	Bar	40.872853	-74.021454	-240	2012-05-12 19:56:03+00:00	2012-05-12 15:56:03+00:00	15	5	Siang	Modern Entertainment	New York
954	4d31e790c6cba35dfad61b7a	4bf58dd8d48988d132941735	Church	40.891131	-74.002018	-240	2012-08-05 16:37:18+00:00	2012-08-05 12:37:18+00:00	12	6	Siang	Other	New York
819	4c74429d66be6dcbc83bbd0f	4bf58dd8d48988d130941735	Building	40.673900	-73.872427	-240	2012-06-18 01:59:14+00:00	2012-06-17 21:59:14+00:00	21	6	Sore	Other	New York
395	4ace6c89f964a52078d020e3	4bf58dd8d48988d1ed931735	Airport	40.773839	-73.871220	-240	2012-05-11 12:59:39+00:00	2012-05-11 08:59:39+00:00	8	4	Pagi	Other	New York

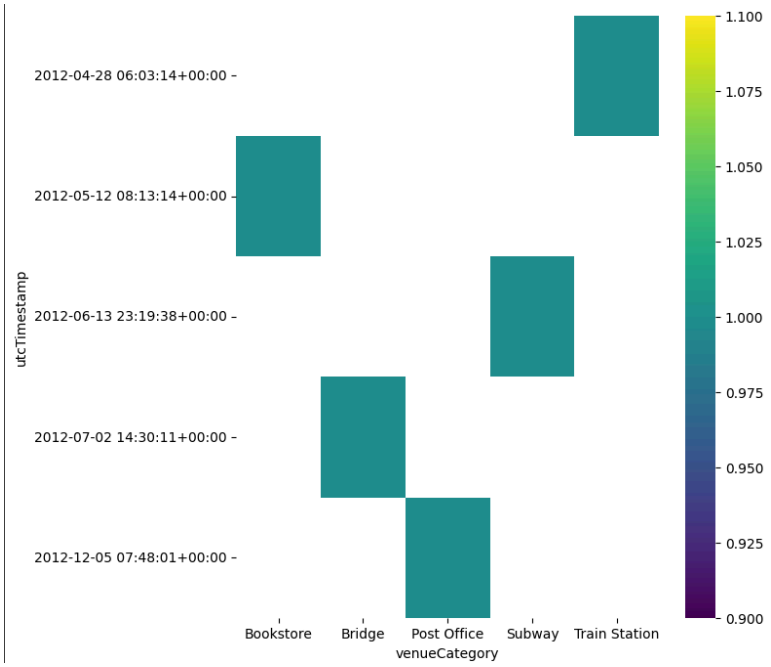
Gambar ini adalah cuplikan tambahan dari dataframe check-in di New York City dengan struktur data serupa seperti sebelumnya. Berikut penjelasan khusus untuk data dalam cuplikan ini:

Observasi Baru dari Dataframe:

1. Kategori Tempat:
    - Data ini menambahkan tempat seperti Synagogue, Church, Bar, Airport, dan Building sebagai kategori.
    - Kategori tempat ini mencerminkan berbagai aktivitas dan tujuan pengguna di New York.
2. Periodisasi Waktu:
    - Periode waktu seperti:
        - Pagi (9 AM): Synagogue dan Airport.
        - Siang (12-3 PM): Church dan Bar.
        - Sore (9 PM): Building.
3. Distribusi Lokasi:

- Koordinat latitude dan longitude tersebar di sekitar wilayah New York City.
- Bisa digunakan untuk membuat peta lokasi check-in untuk kategori tempat tertentu.

4. Keterlibatan Zona Waktu:
    - Zona waktu memiliki offset -240 menit dari UTC, sesuai waktu musim panas di New York (Eastern Daylight Time).



Grafik yang ditampilkan merupakan representasi hubungan antara kategori tempat (venueCategory) dan waktu UTC (utcTimestamp), dengan warna sebagai representasi skala tertentu (mungkin intensitas atau nilai terkait data)

1. Sumbu x (venueCategory)
    - Menampilkan berbagai kategori tempat, seperti Bookstore, Bridge, Post Office, Subway, dan Train Station.
2. Sumbu y (utcTimestamp)
    - Menunjukkan waktu check-in dalam format UTC, diurutkan secara kronologis dari 2012-04-28 hingga 2012-12-05. Hal ini merepresentasikan check-in pada berbagai waktu selama periode tahun tersebut.
3. Warna pada Grafik
    - Warna bar memiliki gradasi dari ungu hingga kuning.

- Interpretasi warna dapat menunjukkan intensitas suatu nilai (misalnya frekuensi check-in, skor tertentu, atau atribut tambahan). Dalam grafik ini, bar dengan nilai warna mendekati kuning menunjukkan nilai lebih tinggi dibandingkan yang ungu.

4. Distribusi Berdasarkan Waktu
   Beberapa tempat memiliki waktu check-in yang terpisah jauh, menunjukkan aktivitas yang sporadis:
   - Train Station: Waktu check-in pada awal grafik (2012-04-28).
   - Bookstore: Check-in pada bulan Mei (2012-05-12).
   - Bridge: Check-in di bulan Juni (2012-06-13).
   - Subway: Check-in di bulan Juli (2012-07-02).
   - Post Office: Check-in di bulan Desember (2012-12-05).

Catatan Lain:

- Visualisasi ini menggambarkan data individual, dan frekuensi kategori tempat terlihat tidak seragam.
- Warna gradasi dapat menunjukkan pentingnya atau intensitas suatu atribut pada check-in di lokasi tertentu.

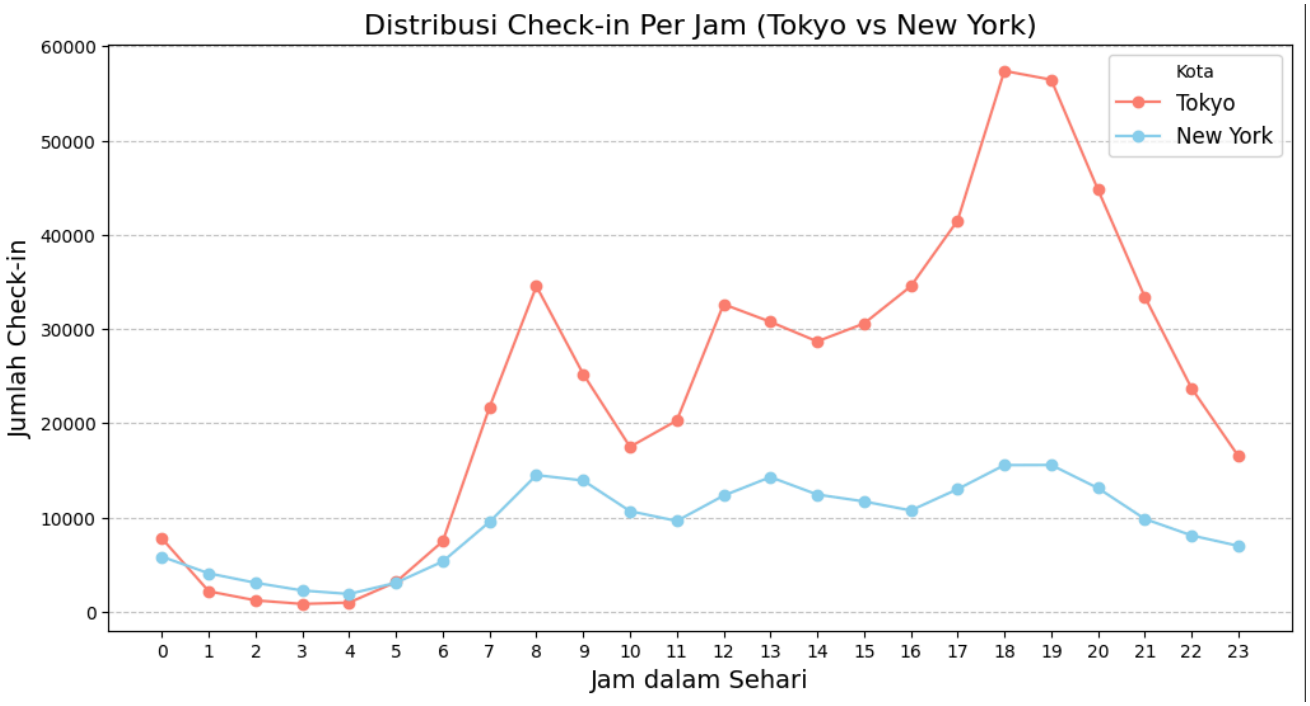
2.2 Visualisasi terkait pertanyaan

2.2.1 Pertanyaan Pola Aktivitas dan Waktu Check-In

- Apakah aktivitas check-in berbeda secara signifikan antara hari kerja dan akhir pekan sepanjang hari?

```
import matplotlib.pyplot as plt

# Hitung jumlah check-in per jam untuk setiap kota
checkin_tky = data_tky.groupby('hour').size()
checkin_nyc = data_nyc.groupby('hour').size()

# Plot line chart
plt.figure(figsize=(12, 6))
plt.plot(checkin_tky.index, checkin_tky.values, label='Tokyo', marker='o', color='salmon')
plt.plot(checkin_nyc.index, checkin_nyc.values, label='New York', marker='o', color='skyblue')

# Tambahkan detail pada plot
plt.title('Distribusi Check-in Per Jam (Tokyo vs New York)', fontsize=16)
plt.xlabel('Jam dalam Sehari', fontsize=14)
plt.ylabel('Jumlah Check-in', fontsize=14)
plt.xticks(range(0, 24))
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.legend(title='Kota', loc='upper right', fontsize=12)
plt.show()
```

Kode ini dimulai dengan menghitung jumlah check-in per jam untuk masing-masing kota menggunakan metode ‘groupby(‘hour’).size()’ pada dataset ‘data_tky’ dan ‘data_nyc’. Hasil pengelompokan tersebut disimpan dalam variabel ‘checkin_tky’ untuk Tokyo dan ‘checkin_nyc’ untuk New York. Kemudian, sebuah grafik garis dibuat menggunakan pustaka matplotlib dengan ukuran 12x6. Dua garis terpisah digambar, satu untuk Tokyo dengan warna salmon, dan satu untuk New York dengan warna biru langit, masing-masing menggunakan simbol penanda (marker= ‘o’) pada titik data. Grafik diberi judul "Distribusi Check-in Per Jam (Tokyo vs New York)", serta sumbu x dan y diberi label yang sesuai. Rentang sumbu x disesuaikan untuk mencakup seluruh jam dalam sehari (0-23), dan garis grid horizontal ditambahkan untuk membantu visualisasi. Terakhir, legenda dengan label kota ditampilkan di sudut kanan atas, dan grafik ditampilkan menggunakan plt.show().



Sumbu dan Data:

1. Sumbu x (Jam dalam Sehari):
    - Merepresentasikan waktu dalam sehari, dari jam 0 (tengah malam) hingga jam 23 (pukul 11 malam).
2. Sumbu y (Jumlah Check-in):
    - Menunjukkan jumlah check-in pada setiap jam dalam sehari.
3. Garis:
    - Tokyo (merah): Garis dengan jumlah check-in lebih tinggi secara keseluruhan.
    - New York (biru): Garis dengan jumlah check-in lebih rendah, lebih stabil dibandingkan Tokyo.
4. Legenda:
    - Warna merah mewakili data check-in Tokyo, dan biru untuk New York.

Observasi:

1. Tokyo:
    - Aktivitas check-in mencapai puncak di dua periode utama:
        - Pukul 8 pagi: Ada lonjakan besar, kemungkinan terkait dengan aktivitas pagi seperti perjalanan kerja atau sarapan.
        - Pukul 6 sore hingga 7 malam: Puncak kedua, mungkin mencerminkan aktivitas setelah jam kerja atau aktivitas malam.
    - Aktivitas cenderung lebih rendah antara pukul 1 pagi hingga 5 pagi.

2. New York:
    - Aktivitas check-in lebih merata sepanjang hari, tanpa lonjakan besar seperti di Tokyo.
    - Aktivitas check-in tertinggi terjadi sekitar pukul 10 pagi hingga 3 sore, menunjukkan fokus pada aktivitas siang.
3. Perbedaan Pola:
    - Tokyo memiliki pola yang lebih tajam dengan dua puncak utama (pagi dan sore), menunjukkan rutinitas harian yang terstruktur.
    - New York memiliki distribusi yang lebih seimbang sepanjang hari, mungkin mencerminkan gaya hidup yang lebih fleksibel atau beragam.

Kesimpulan:

- Tokyo menunjukkan pola aktivitas yang sangat terikat dengan waktu kerja dan perjalanan.
- New York memiliki pola yang lebih merata, menunjukkan aktivitas yang tersebar sepanjang hari.
- Pola ini dapat mencerminkan perbedaan budaya, gaya hidup, dan penggunaan tempat di kedua kota.

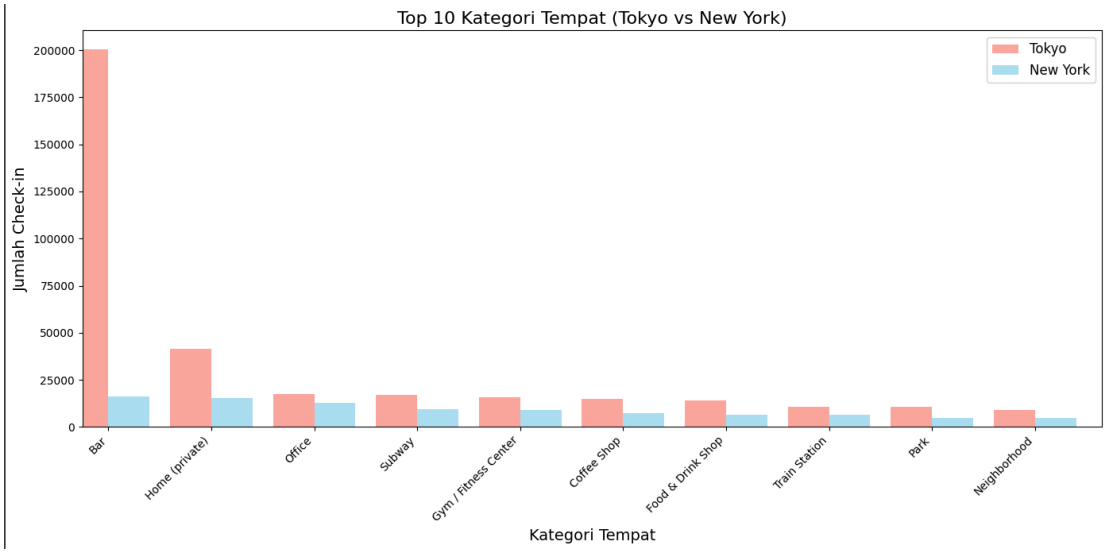
2.2.2 Pertanyaan Preferensi Jenis Tempat yang Dikunjungi

- Bagaimana distribusi kunjungan berdasarkan kategori tempat di kedua kota? (Apakah kategori tertentu lebih dominan di Tokyo atau NYC?)

```
top_categories_tky = data_tky['venueCategory'].value_counts().head(10)
top_categories_nyc = data_nyc['venueCategory'].value_counts().head(10)

plt.figure(figsize=(14, 7))
top_categories_tky.plot(kind='bar', color='salmon', alpha=0.7, label='Tokyo', position=1, width=0.4)
top_categories_nyc.plot(kind='bar', color='skyblue', alpha=0.7, label='New York', position=0,
width=0.4)
plt.title('Top 10 Kategori Tempat (Tokyo vs New York)', fontsize=16)
plt.ylabel('Jumlah Check-in', fontsize=14)
plt.xlabel('Kategori Tempat', fontsize=14)
plt.legend(fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

Kode ini pertama-tama menghitung jumlah check-in untuk setiap kategori tempat pada dataset data_tky dan data_nyc dengan menggunakan metode value_counts(), dan memilih 10 kategori teratas menggunakan head(10). Selanjutnya, grafik batang dibuat dengan ukuran 14x7 menggunakan pustaka matplotlib. Dua set grafik batang digambar, satu untuk Tokyo dengan

warna salmon dan satu untuk New York dengan warna biru langit. Batang untuk Tokyo dan New York ditempatkan berdampingan menggunakan parameter position dan width agar mudah dibandingkan. Grafik ini diberi judul "Top 10 Kategori Tempat (Tokyo vs New York)", serta label pada sumbu x untuk kategori tempat dan sumbu y untuk jumlah check-in. Label sumbu x diputar 45 derajat dengan plt.xticks(rotation=45, ha='right') untuk memastikan keterbacaan kategori tempat. Fungsi plt.tight_layout() digunakan untuk memastikan bahwa semua elemen grafik tertata rapi, dan grafik tersebut akhirnya ditampilkan dengan plt.show().



Grafik ini membandingkan Top 10 Kategori Tempat berdasarkan jumlah check-in antara Tokyo dan New York. Berikut adalah analisisnya:

Informasi Grafik:

1. Sumbu x (Kategori Tempat):
    - Menampilkan 10 kategori tempat teratas berdasarkan check-in, seperti Bar, Home (private), Office, Subway, dan lainnya.
2. Sumbu y (Jumlah Check-in):
    - Menunjukkan jumlah total check-in untuk setiap kategori tempat.
3. Warna:
    - Merah: Check-in di Tokyo.
    - Biru: Check-in di New York.
4. Legenda:
    - Menjelaskan perbedaan warna antara data Tokyo dan New York.

Observasi:

1. Dominasi Bar:
    - Di Tokyo, kategori *Bar* memiliki jumlah check-in yang sangat besar, hampir mendominasi total check-in dibanding kategori lain.
    - New York juga memiliki jumlah check-in yang signifikan di *Bar*, tetapi jauh lebih kecil dibandingkan Tokyo.
2. Home (private):
    - Tempat tinggal pribadi adalah kategori dengan check-in terbesar kedua di kedua kota, tetapi lebih banyak di Tokyo dibandingkan New York.
3. Kategori dengan Pola Serupa:
    - *Office*, *Subway*, *Coffee Shop*, dan *Food & Drink Shop* memiliki jumlah check-in yang relatif lebih seimbang antara Tokyo dan New York, meskipun Tokyo tetap lebih unggul.
4. Kategori Lainnya:
    - Beberapa kategori seperti *Train Station* dan *Park* memiliki kontribusi check-in yang kecil, dengan jumlah yang hampir sama antara kedua kota.

Kesimpulan:

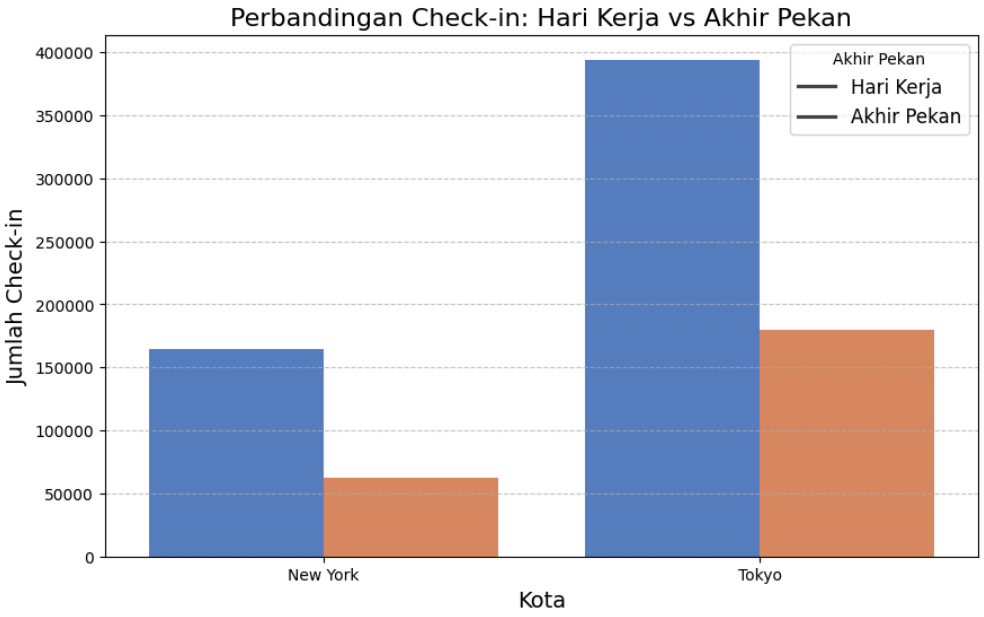
- Tokyo memiliki jumlah check-in yang sangat tinggi di kategori Bar dan Home (private), menunjukkan fokus sosial atau budaya di tempat hiburan malam dan rumah pribadi.
- New York memiliki pola check-in yang lebih merata, meskipun Bar tetap menjadi salah satu tempat favorit.
- Data ini mencerminkan perbedaan budaya dan gaya hidup antara Tokyo dan New York, di mana Tokyo cenderung lebih terkonsentrasi di kategori tertentu.

3. Pertanyaan Tren Berdasarkan Hari Kerja dan Akhir Pekan

- Apakah ada perbedaan kategori tempat yang lebih sering dikunjungi pada hari kerja vs akhir pekan?

```
data_combined['weekend'] = data_combined['day_of_week'] >= 5
weekend_data = data_combined.groupby(['city', 'weekend']).size().reset_index(name='count')

plt.figure(figsize=(10, 6))
sns.barplot(data=weekend_data, x='city', y='count', hue='weekend', palette='muted')
plt.title('Perbandingan Check-in: Hari Kerja vs Akhir Pekan', fontsize=16)
plt.xlabel('Kota', fontsize=14)
plt.ylabel('Jumlah Check-in', fontsize=14)
plt.xticks(rotation=0)
plt.legend(title='Akhir Pekan', labels=['Hari Kerja', 'Akhir Pekan'], fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

Kode ini dimulai dengan menghitung jumlah check-in untuk setiap kategori tempat pada dataset Tokyo dan New York menggunakan metode value_counts() pada kolom venueCategory. Hasil ini diambil 10 kategori teratas dengan fungsi head(10) dan disimpan dalam variabel top_categories_tky untuk Tokyo dan top_categories_nyc untuk New York. Grafik batang dibuat menggunakan pustaka matplotlib dengan ukuran 14x7. Batang untuk kategori Tokyo digambar dengan warna salmon, sementara batang untuk kategori New York menggunakan warna biru langit. Parameter position dan width digunakan untuk mengatur posisi batang sehingga saling bersebelahan untuk memudahkan perbandingan. Grafik diberi judul "Top 10 Kategori Tempat (Tokyo vs New York)" dan sumbu x serta y diberi label yang relevan. Rotasi label pada sumbu x diatur ke 45 derajat agar kategori tempat dapat terbaca dengan jelas. Akhirnya, grafik diperindah dengan pengaturan layout yang rapat menggunakan plt.tight_layout() dan ditampilkan menggunakan plt.show()



Grafik ini menunjukkan perbandingan jumlah check-in antara hari kerja dan akhir pekan untuk kota New York dan Tokyo. Berikut adalah analisisnya:

Informasi Grafik:

1. Sumbu x (Kota):
    - Menampilkan dua kota, yaitu New York dan Tokyo.
2. Sumbu y (Jumlah Check-in):
    - Menunjukkan jumlah check-in yang terjadi di masing-masing kota, baik pada hari kerja maupun akhir pekan.

3. Warna:
    - Biru: Check-in pada hari kerja.
    - Oranye: Check-in pada akhir pekan.
4. Legenda:
    - Menjelaskan kategori hari kerja dan akhir pekan.

Observasi:

1. New York:
    - Jumlah check-in pada hari kerja (biru) lebih tinggi daripada akhir pekan (oranye).
    - Check-in pada hari kerja di New York menunjukkan aktivitas yang lebih dominan, mencerminkan tingginya mobilitas masyarakat di hari kerja.
2. Tokyo:
    - Jumlah check-in di Tokyo jauh lebih tinggi dibandingkan New York, baik pada hari kerja maupun akhir pekan.
    - Check-in pada hari kerja di Tokyo sangat mendominasi, namun aktivitas di akhir pekan juga cukup signifikan, hampir mendekati setengah dari check-in hari kerja.
3. Perbandingan Hari Kerja dan Akhir Pekan:
    - Di kedua kota, check-in lebih tinggi pada hari kerja dibanding akhir pekan, tetapi Tokyo menunjukkan perbedaan yang lebih kecil antara hari kerja dan akhir pekan dibandingkan New York.

Kesimpulan:

- New York: Aktivitas check-in lebih terpusat pada hari kerja, mungkin terkait dengan rutinitas kerja atau aktivitas bisnis.
- Tokyo: Meskipun check-in pada hari kerja lebih tinggi, aktivitas akhir pekan cukup signifikan, mencerminkan gaya hidup yang lebih aktif sepanjang minggu.
- Bagaimana pola check-in di tempat transportasi umum dibanding tempat hiburan pada hari kerja dan akhir pekan?

4. Pendekatan Keramaian dan Interaksi Sosial

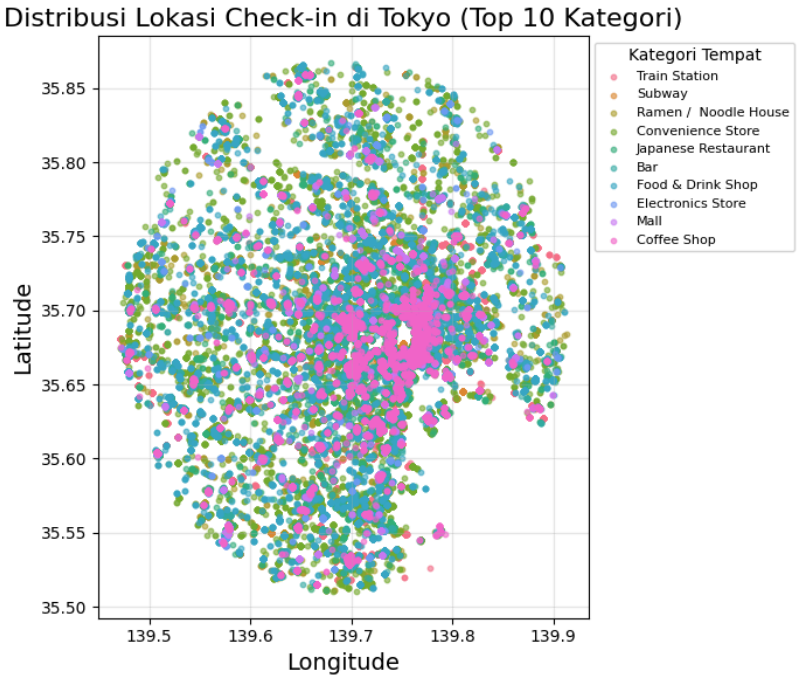
- Apakah terdapat perbedaan kepadatan check-in di pusat kota dan area pinggiran?

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

# Top 10 Kategori Tempat untuk Tokyo dan NYC
top_categories_tky = data_tky['venueCategory'].value_counts().head(10).index
top_categories_nyc = data_nyc['venueCategory'].value_counts().head(10).index

# Buat Subset Data untuk Top 10 Kategori
data_tky_filtered = data_tky[data_tky['venueCategory'].isin(top_categories_tky)]
data_nyc_filtered = data_nyc[data_nyc['venueCategory'].isin(top_categories_nyc)]

# Palet Warna untuk Top 10 Kategori
palette_tky = sns.color_palette('husl', len(top_categories_tky))
palette_nyc = sns.color_palette('husl', len(top_categories_nyc))

# Map Warna ke Kategori
color_map_tky = dict(zip(top_categories_tky, palette_tky))
color_map_nyc = dict(zip(top_categories_nyc, palette_nyc))
```

Kode ini dimulai dengan mengambil 10 kategori tempat teratas dari dataset data_tky dan data_nyc berdasarkan jumlah check-in, menggunakan metode value_counts().head(10).index untuk mendapatkan indeks kategori tersebut. Kemudian, dua subset data dibuat, yaitu data_tky_filtered dan data_nyc_filtered, yang hanya memuat baris dengan kategori tempat yang termasuk dalam 10 besar kategori untuk masing-masing kota. Selanjutnya, palet warna disiapkan menggunakan pustaka seaborn untuk masing-masing kota. Fungsi sns.color_palette('husl', len(top_categories_tky)) dan sns.color_palette('husl', len(top_categories_nyc)) menghasilkan palet warna yang memiliki jumlah warna yang sama dengan jumlah kategori teratas. Kemudian, sebuah pemetaan warna (color_map_tky dan color_map_nyc) dibuat dengan memetakan setiap kategori ke warna tertentu dari palet yang telah dibuat. Ini memungkinkan kategori-kategori tempat di Tokyo dan New York untuk diwarnai dengan warna yang berbeda pada visualisasi berikutnya, mempermudah perbandingan antar kategori.

```python
# Scatter Plot untuk Tokyo (Top 10 Kategori)
plt.figure(figsize=(7, 6))
for category, color in color_map_tky.items():
    subset = data_tky_filtered[data_tky_filtered['venueCategory'] == category]
    plt.scatter(subset['longitude'], subset['latitude'], alpha=0.6, s=10, color=color, label=category)

plt.title('Distribusi Lokasi Check-in di Tokyo (Top 10 Kategori)', fontsize=16)
plt.xlabel('Longitude', fontsize=14)
plt.ylabel('Latitude', fontsize=14)
plt.legend(loc='upper left', bbox_to_anchor=(1, 1), fontsize=8, title="Kategori Tempat")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()
```

Kode ini membuat visualisasi berupa scatter plot yang menunjukkan distribusi geografis check-in untuk kategori tempat teratas di Tokyo. Setiap kategori tempat ditampilkan dengan warna yang berbeda, yang sudah ditentukan sebelumnya menggunakan color_map_tky. Proses dimulai dengan iterasi melalui setiap kategori tempat yang terdapat dalam color_map_tky. Untuk setiap kategori, subset data Tokyo (data_tky_filtered) yang memiliki kategori tersebut diambil, lalu posisi longitude dan latitude untuk check-in yang termasuk dalam kategori tersebut digambarkan dengan titik pada scatter plot. Setiap titik diberi transparansi (alpha=0.6), ukuran kecil (s=10), dan warna sesuai kategori. Grafik ini diberi judul "Distribusi Lokasi Check-in di Tokyo (Top 10 Kategori)" dan label untuk sumbu x (longitude) dan y (latitude). Legenda diatur di luar plot untuk memudahkan identifikasi kategori tempat. Grid pada plot diatur dengan transparansi rendah (alpha=0.3) untuk memperjelas titik data. Fungsi plt.tight_layout() memastikan bahwa layout plot tidak terpotong, dan plot akhirnya ditampilkan menggunakan plt.show().



Visualisasi tersebut adalah scatter plot yang menunjukkan distribusi lokasi check-in di Tokyo berdasarkan koordinat geografis (latitude dan longitude) untuk 10 kategori tempat teratas.

Elemen Visualisasi:

1. Sumbu X (Longitude) dan Sumbu Y (Latitude):
    - Menunjukkan lokasi geografis check-in di wilayah Tokyo.
2. Warna Titik:
    - Setiap warna mewakili kategori tempat tertentu, seperti *Train Station*, *Subway*, *Coffee Shop*, dll. Legenda di sisi kanan memberikan keterangan untuk setiap kategori.
3. Pola Distribusi:
    - Banyaknya titik mengindikasikan lokasi check-in populer di setiap kategori.

Observasi:

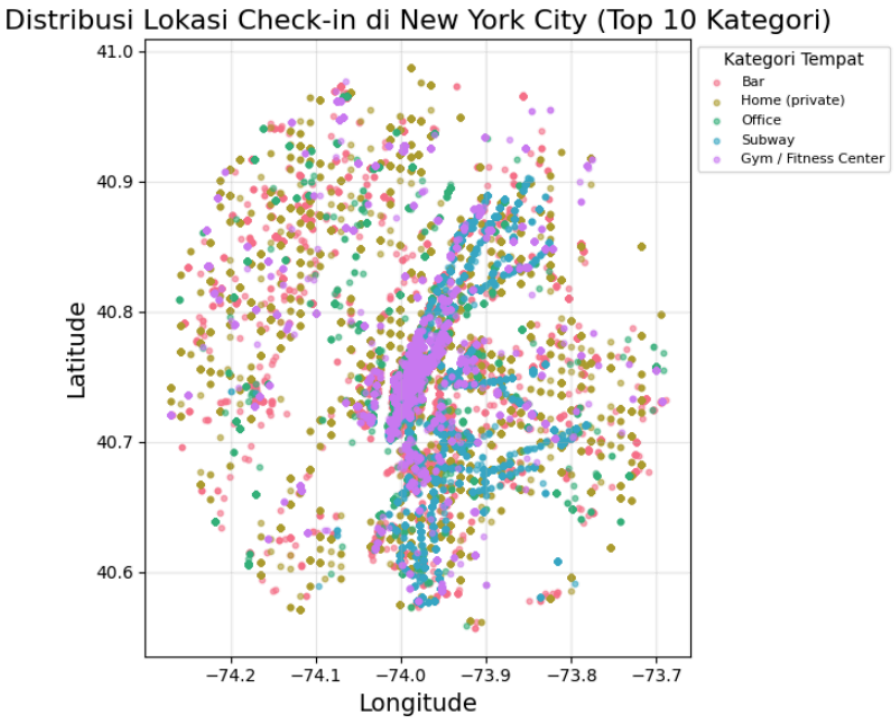
1. Konsentrasi Check-in:
    - Terdapat konsentrasi tinggi di area tertentu, terutama di pusat kota Tokyo (di sekitar koordinat longitude 139.7 dan latitude 35.7). Hal ini mungkin mencerminkan area yang padat seperti distrik bisnis atau pusat hiburan.
2. Kategori yang Dominan:
    - Kategori seperti *Coffee Shop* (warna magenta) memiliki distribusi yang luas, menunjukkan popularitas yang tinggi di berbagai area.
    - *Train Station* dan *Subway* (warna merah muda dan abu-abu) terkonsentrasi di jalur transportasi utama.
3. Sebaran Tempat Lain:
    - Lokasi seperti *Ramen/Noodle House* dan *Convenience Store* juga tersebar luas, yang mencerminkan sifat tempat-tempat ini sebagai bagian dari kehidupan sehari-hari.

Interpretasi:

- Lokasi check-in paling padat terjadi di pusat kota dan sepanjang jalur transportasi, yang menunjukkan aktivitas tinggi di wilayah-wilayah tersebut.
- Kategori seperti *Coffee Shop* dan *Convenience Store* menunjukkan distribusi yang lebih merata, karena tempat-tempat ini biasanya tersedia di berbagai lokasi.

```
# Scatter Plot untuk NYC (Top 10 Kategori)
plt.figure(figsize=(7, 6))
for category, color in color_map_nyc.items():
    subset = data_nyc_filtered[data_nyc_filtered['venueCategory'] == category]
    plt.scatter(subset['longitude'], subset['latitude'], alpha=0.6, s=10, color=color, label=category)

plt.title('Distribusi Lokasi Check-in di New York City (Top 10 Kategori)', fontsize=16)
plt.xlabel('Longitude', fontsize=14)
plt.ylabel('Latitude', fontsize=14)
plt.legend(loc='upper left', bbox_to_anchor=(1, 1), fontsize=8, title="Kategori Tempat")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()
```

Kode ini membuat scatter plot untuk menampilkan sebaran geografis check-in berdasarkan kategori tempat teratas di New York City. Setiap kategori tempat ditampilkan dengan warna yang berbeda, yang sudah dipetakan sebelumnya menggunakan color_map_nyc. Untuk setiap kategori, subset data New York (data_nyc_filtered) yang mencakup check-in pada kategori tersebut diambil, lalu posisi longitude dan latitude untuk setiap check-in tersebut digambarkan sebagai titik pada plot. Titik-titik ini diberi transparansi (alpha=0.6), ukuran kecil (s=10), dan warna sesuai kategori tempat. Plot ini diberi judul "Distribusi Lokasi Check-in di New York City (Top 10 Kategori)", dan label untuk sumbu x (longitude) serta y (latitude) ditambahkan untuk memperjelas koordinat. Legenda disusun di luar plot untuk memudahkan identifikasi kategori, dengan ukuran font yang lebih kecil. Grid pada plot ditambahkan dengan transparansi rendah (alpha=0.3) untuk membantu visualisasi tanpa mengganggu titik data. Fungsi plt.tight_layout() digunakan untuk memastikan tata letak plot tidak terpotong, dan plot tersebut akhirnya ditampilkan menggunakan plt.show().



Grafik ini menunjukkan distribusi lokasi check-in di New York City berdasarkan koordinat latitude dan longitude. Data ini berfokus pada Top 10 kategori tempat, dengan masing-masing kategori ditandai menggunakan warna yang berbeda.

Elemen Grafik:

1. Sumbu X (Longitude) dan Sumbu Y (Latitude):
    - Lokasi geografis tempat check-in di New York City.
2. Legenda Kategori Tempat:
    - Kategori utama:
        - Bar (merah muda)
        - Home (private) (hijau muda)
        - Office (kuning)
        - Subway (biru muda)
        - Gym / Fitness Center (ungu)
    - Setiap titik mewakili satu check-in pada lokasi tertentu, dengan warna menunjukkan kategori tempatnya.
3. Kepadatan Titik:
    - Area dengan konsentrasi titik lebih tinggi menunjukkan wilayah yang lebih sering dikunjungi.

Observasi:

1. Konsentrasi Titik:
    - Titik-titik cenderung terkonsentrasi di area pusat kota, kemungkinan besar di sekitar Manhattan, yang merupakan pusat bisnis dan hiburan New York City.
    - Area di bagian luar Manhattan memiliki penyebaran titik yang lebih jarang.
2. Kategori Dominan:
    - Subway (biru muda) dan Bar (merah muda) tampak tersebar luas di seluruh kota, mencerminkan popularitasnya sebagai tempat yang sering dikunjungi.
    - Home (private) (hijau) memiliki distribusi lebih merata, mencerminkan aktivitas check-in dari area pemukiman.
    - Office (kuning) terkonsentrasi lebih banyak di area bisnis.
3. Kategori yang Spesifik:
    - Gym / Fitness Center (ungu) memiliki distribusi yang lebih jarang, tetapi terlihat terkonsentrasi di lokasi-lokasi tertentu, kemungkinan besar di area dengan fasilitas kebugaran terkenal.

Kesimpulan:

- Lokasi check-in di New York City sangat terpusat di kawasan Manhattan, menunjukkan aktivitas yang tinggi di area ini.
- Subway dan Bar merupakan kategori dengan check-in yang paling banyak terlihat, mendukung peran transportasi umum dan hiburan dalam kehidupan kota.
- Lokasi Home (private) tersebar luas, mencerminkan check-in dari tempat tinggal di berbagai bagian kota.

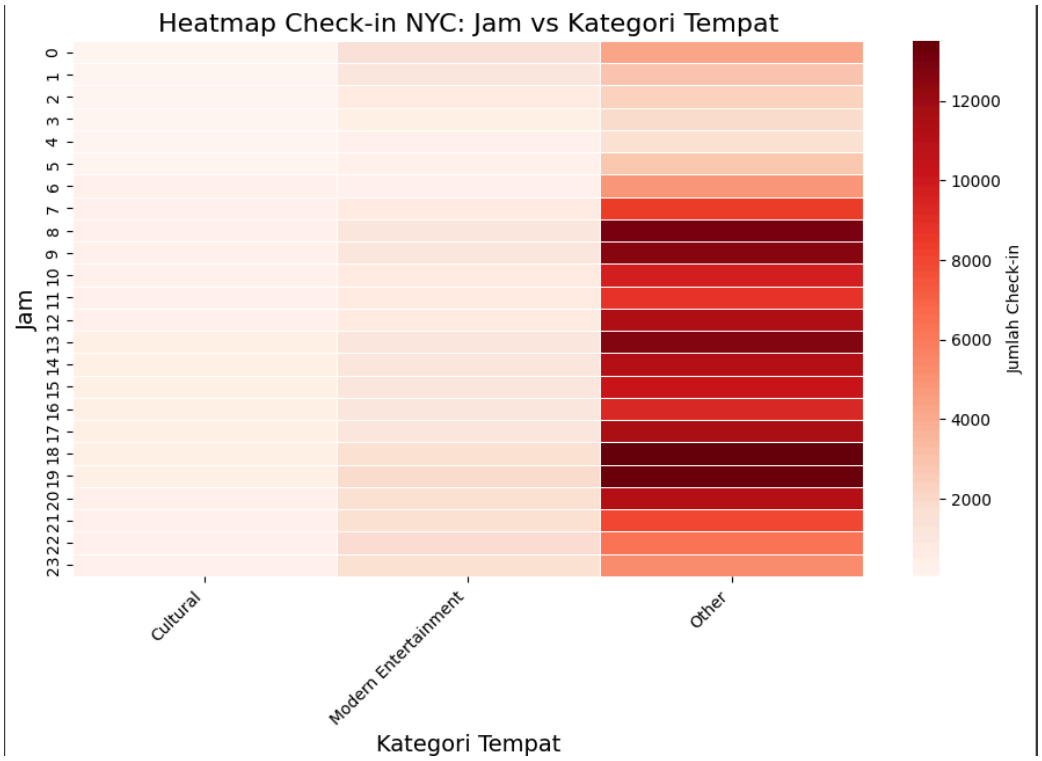
5. Pertanyaan Distribusi Periode Waktu

- Apakah ada perbedaan aktivitas check-in berdasarkan periode waktu (pagi, siang, sore, malam)?

```
plt.figure(figsize=(11, 6))
sns.heatmap(pivot_nyc, cmap='Reds', cbar_kws={'label': 'Jumlah Check-in'}, linewidths=0.5)
plt.title('Heatmap Check-in NYC: Jam vs Kategori Tempat', fontsize=16)
plt.xlabel('Kategori Tempat', fontsize=14)
plt.ylabel('Jam', fontsize=14)
plt.xticks(rotation=45, ha='right')
plt.show()
```

Kode ini membuat sebuah heatmap menggunakan pustaka seaborn untuk memvisualisasikan hubungan antara jam dan kategori tempat di New York dengan jumlah

check-in. Heatmap ini dibuat berdasarkan data yang telah dipivot sebelumnya (misalnya, pivot_nyc) yang memiliki kategori tempat pada sumbu x, jam pada sumbu y, dan nilai jumlah check-in sebagai intensitas warna. Fungsi sns.heatmap() digunakan dengan palet warna 'Reds', yang memberikan nuansa warna merah untuk menunjukkan intensitas yang lebih tinggi, di mana semakin merah semakin banyak jumlah check-in. Parameter cbar_kws={'label': 'Jumlah Check-in'} digunakan untuk menambahkan label pada color bar, yang menunjukkan skala jumlah check-in. Ukuran figure ditetapkan 11x6 untuk memberikan ruang yang cukup bagi visualisasi. Grafik ini diberi judul "Heatmap Check-in NYC: Jam vs Kategori Tempat", dan label sumbu x serta y ditambahkan untuk menjelaskan bahwa sumbu x adalah kategori tempat dan sumbu y adalah jam. Label sumbu x diputar 45 derajat menggunakan plt.xticks(rotation=45, ha='right') agar lebih mudah dibaca.



Grafik ini adalah heatmap check-in di New York City, yang menunjukkan distribusi check-in berdasarkan jam dan kategori tempat. Berikut adalah analisisnya:

Informasi Grafik:

1. Sumbu Y (Jam):

    - Representasi waktu dalam 24 jam, dari pukul 0 (tengah malam) hingga 23 (pukul 11 malam).
2. Sumbu X (Kategori Tempat):
    - Tempat dikategorikan menjadi tiga kelompok utama:
        - Cultural: Tempat budaya.
        - Modern Entertainment: Hiburan modern.
        - Other: Kategori lain.
3. Warna Heatmap:
    - Warna lebih gelap (merah tua) menunjukkan jumlah check-in yang lebih tinggi.
    - Warna lebih terang (putih) menunjukkan jumlah check-in yang lebih rendah.
4. Skala Warna (Jumlah Check-in):
    - Menunjukkan intensitas check-in mulai dari rendah hingga tinggi.

Observasi:

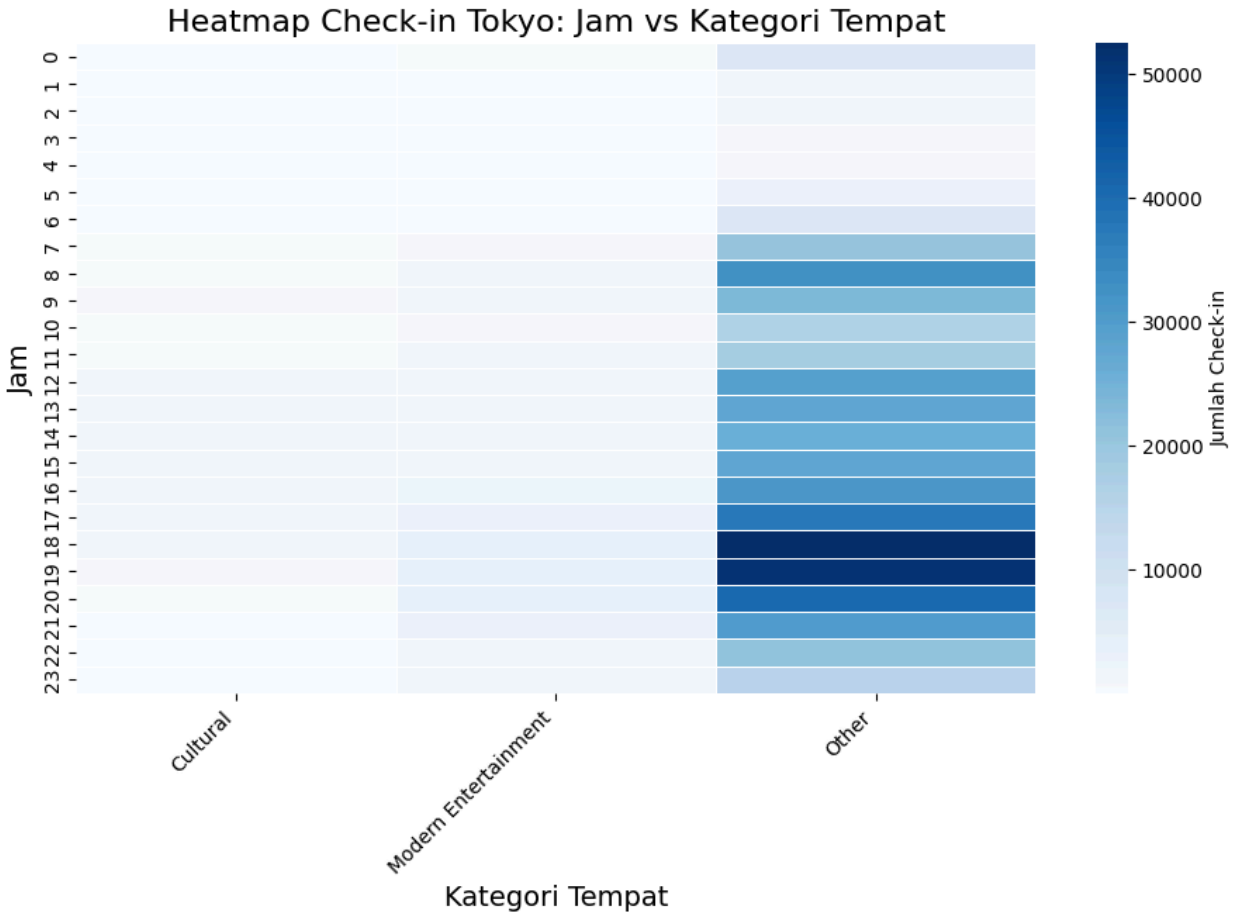
1. Jam dengan Aktivitas Tertinggi:*
    - Aktivitas check-in paling tinggi terjadi antara pukul 12 siang hingga 8 malam, dengan puncak di sore hari.
    - Setelah pukul 8 malam, intensitas check-in mulai menurun.
2. Kategori Tempat:
    - Cultural: Aktivitas check-in lebih menyebar, dengan jumlah yang tidak terlalu signifikan dibanding kategori lainnya.
    - Modern Entertainment: Memiliki intensitas check-in yang lebih tinggi, terutama pada sore hingga malam hari.
    - Other: Menunjukkan puncak aktivitas yang lebih terfokus, serupa dengan kategori Modern Entertainment.
3. Perbandingan Intensitas:
    - Kategori Modern Entertainment mendominasi jumlah check-in dibandingkan kategori lain.

Kesimpulan:

- Tempat hiburan modern adalah kategori yang paling populer di NYC, dengan jam puncak aktivitas pada sore hingga malam hari.
- Tempat budaya memiliki aktivitas check-in yang lebih stabil sepanjang hari, tetapi tidak sebesar tempat hiburan modern.
- Grafik ini menunjukkan bahwa masyarakat cenderung lebih aktif di tempat hiburan pada jam-jam rekreasi, seperti sore hingga malam hari.

```
pivot_tky = pd.crosstab(data_tky['hour'], data_tky['venueCategoryGrouped'])
pivot_nyc = pd.crosstab(data_nyc['hour'], data_nyc['venueCategoryGrouped'])

plt.figure(figsize=(11, 6))
sns.heatmap(pivot_tky, cmap='Blues', cbar_kws={'label': 'Jumlah Check-in'}, linewidths=0.5)
plt.title('Heatmap Check-in Tokyo: Jam vs Kategori Tempat', fontsize=16)
plt.xlabel('Kategori Tempat', fontsize=14)
plt.ylabel('Jam', fontsize=14)
plt.xticks(rotation=45, ha='right')
plt.show()
  return go(f, seed, [])
}
```

Kode ini membuat tabel kontingensi (cross-tabulation) menggunakan fungsi pd.crosstab(), yang menghasilkan matriks jumlah check-in untuk setiap kombinasi antara jam (sumbu x) dan kategori tempat yang dikelompokkan (venueCategoryGrouped). Tabel kontingensi ini disimpan dalam variabel pivot_tky untuk Tokyo dan pivot_nyc untuk New York. Selanjutnya, sebuah heatmap dibuat untuk Tokyo menggunakan pustaka seaborn dengan palet warna 'Blues', yang memberikan gradasi biru untuk menggambarkan intensitas jumlah check-in—semakin gelap birunya, semakin banyak jumlah check-in pada kombinasi jam dan kategori tempat tertentu. Fungsi sns.heatmap() digunakan untuk menggambarkan heatmap dengan label pada color bar (cbar_kws={'label': 'Jumlah Check-in'}) yang menunjukkan skala jumlah check-in. Ukuran plot ditetapkan 11x6, dan grafik ini diberi judul "Heatmap Check-in Tokyo: Jam vs Kategori Tempat". Label untuk sumbu x dan y ditambahkan untuk menunjukkan bahwa sumbu x mewakili kategori tempat dan sumbu y mewakili jam. Label sumbu x diputar 45 derajat dengan plt.xticks(rotation=45, ha='right') untuk memudahkan pembacaan kategori tempat.



Gambar ini adalah heatmap yang mirip dengan yang sebelumnya, menampilkan hubungan antara waktu (jam) dan kategori tempat di Tokyo berdasarkan jumlah check-in. Berikut adalah analisis lebih lanjut berdasarkan gambar tersebut:

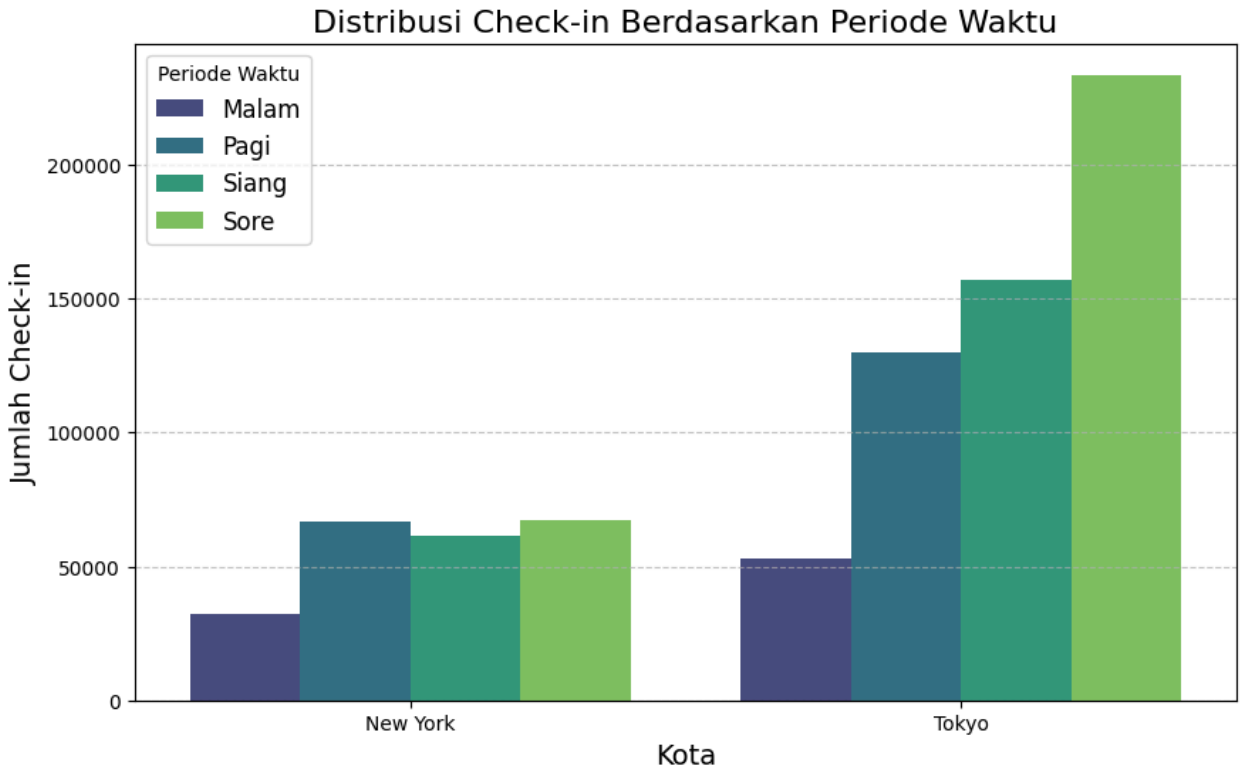
1. Sumbu Vertikal (Jam):
    - Menampilkan rentang waktu dari 0 (pukul 00:00) hingga 23 (pukul 23:00), yang menggambarkan aktivitas check-in berdasarkan jam.
2. Sumbu Horizontal (Kategori Tempat):
    - Meliputi beberapa kategori seperti *Cultural*, *Modern Entertainment*, dan *Other*.
3. Warna Heatmap:
    - Warna biru tua menunjukkan jumlah check-in tertinggi.
    - Warna biru terang hingga putih menunjukkan jumlah check-in yang rendah.
4. Colorbar di Sisi Kanan:
    - Memberikan informasi skala jumlah check-in, mulai dari nilai terendah (putih) hingga tertinggi (biru tua), dengan puncak di angka lebih dari 50.000 check-in.

Observasi

1. Waktu dengan Aktivitas Tertinggi:
    - Aktivitas check-in paling tinggi terjadi di malam hari antara pukul 18:00 hingga 21:00, terutama pada kategori "Modern Entertainment" dan "Other."
2. Kategori "Cultural":
    - Jumlah check-in sangat rendah di kategori ini sepanjang hari.
3. Tren Umum:
    - Aktivitas check-in memuncak pada sore hingga malam hari, mengindikasikan waktu-waktu tersebut populer untuk aktivitas hiburan.

```
period_data = data_combined.groupby(['city', 'period']).size().reset_index(name='count')

plt.figure(figsize=(10, 6))
sns.barplot(data=period_data, x='city', y='count', hue='period', palette='viridis')
plt.title('Distribusi Check-in Berdasarkan Periode Waktu', fontsize=16)
plt.xlabel('Kota', fontsize=14)
plt.ylabel('Jumlah Check-in', fontsize=14)
plt.legend(title='Periode Waktu', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

Kode ini mengelompokkan data gabungan (data_combined) berdasarkan kota (city) dan periode waktu (period), menggunakan fungsi groupby(). Fungsi size() menghitung jumlah check-in untuk setiap kombinasi kota dan periode waktu, dan hasilnya disimpan dalam variabel period_data. Setelah itu, sebuah grafik batang dibuat menggunakan pustaka seaborn (sns.barplot()) dengan sumbu x menunjukkan kota, sumbu y menunjukkan jumlah check-in, dan hue (warna) mewakili periode waktu. Palet warna 'viridis' digunakan untuk memberikan gradasi warna yang menarik. Grafik diberi judul "Distribusi Check-in Berdasarkan Periode Waktu" dan label untuk sumbu x dan y masing-masing "Kota" dan "Jumlah Check-in". Legenda ditambahkan untuk menjelaskan periode waktu dengan ukuran font yang sesuai. Untuk memperjelas pembacaan grafik, grid horizontal diaktifkan dengan transparansi rendah (alpha=0.7).



Grafik ini menunjukkan distribusi check-in berdasarkan periode waktu di dua kota, yaitu New York dan Tokyo. Periode waktu dibagi menjadi Malam, Pagi, Siang, dan Sore.

Analisis Grafik:

1. Sumbu X (Kota):
    - Mencakup dua kota, yaitu New York dan Tokyo.
2. Sumbu Y (Jumlah Check-in):
    - Menunjukkan jumlah check-in untuk setiap periode waktu di masing-masing kota.
3. Warna Bar (Periode Waktu):
    - Malam (ungu tua), Pagi (biru tua), Siang (hijau tua), dan Sore (hijau terang).

Observasi:

1. New York:
    - Sore memiliki jumlah check-in tertinggi dibandingkan periode lainnya.
    - Aktivitas check-in di periode Pagi dan Siang hampir setara, tetapi lebih rendah dari Sore.
    - Periode Malam memiliki jumlah check-in terendah.
2. Tokyo:
    - Sore mendominasi jumlah check-in, dengan angka jauh lebih tinggi dibandingkan periode lain.

    - Aktivitas check-in pada Siang cukup signifikan, tetapi tidak setinggi Sore.
    - Periode Malam memiliki jumlah check-in terendah, mirip dengan New York.
3. Perbandingan Antar Kota:
    - Tokyo memiliki jumlah check-in yang jauh lebih tinggi di semua periode waktu dibandingkan dengan New York.
    - Periode Sore di Tokyo hampir dua kali lipat jumlah check-in dibandingkan di New York.

2.3 Kesimpulan

- Periode sore tetap menjadi waktu puncak aktivitas check-in di Tokyo dan New York City. Hal ini mencerminkan rutinitas yang umum, di mana orang cenderung beraktivitas di luar rumah setelah jam kerja. Namun, intensitas check-in di Tokyo jauh lebih tinggi dibandingkan dengan New York pada semua periode waktu. Hal ini dapat mencerminkan perbedaan tingkat kepadatan penduduk, budaya, atau tingkat adopsi teknologi check-in pada aplikasi tertentu..
- Tokyo menunjukkan distribusi check-in yang lebih terpusat di waktu-waktu tertentu seperti pagi dan sore, yang mencerminkan pola kehidupan yang lebih terstruktur. Sebaliknya, New York memiliki distribusi aktivitas yang lebih merata sepanjang hari, meskipun tetap menunjukkan penurunan yang signifikan pada malam hari.
- Selain itu, perbedaan signifikan pada aktivitas check-in malam hari mengindikasikan bahwa masyarakat Tokyo cenderung memiliki pola hidup yang lebih konservatif terkait jam malam, sementara masyarakat New York, meskipun aktivitasnya lebih rendah pada malam hari, tetap menunjukkan pola yang lebih fleksibel. Ini dapat mencerminkan perbedaan gaya hidup dan dinamika kota yang berbeda antara kedua lokasi.

Link Github : https://github.com/naufalfakhri14/Kelompok-8_Visualisasi-Data_RB