Nama	: Muhammad Hisyam Kamil
NIM	: 202210370311060
Mata Kuliah	: Pemodelan dan Simulasi Data
Kelas	: 6B
Source Code	: https://github.com/hisyam99/dip-task2-kdd
Google Colab (IPYNB)	: https://mil.kamil.my.id/dip-task2-hisyam99
Kaggle Dataset	: https://www.kaggle.com/datasets/nypd/vehicle-collisions

Laporan Tugas 2: Analisis Data Kecelakaan Kendaraan di New York City (NYC)

Pendahuluan

Tujuan utama dari tugas ini adalah menerapkan teknik *Knowledge Discovery in Databases* (KDD) untuk mengekstrak pengetahuan dari dataset dunia nyata, sekaligus memastikan kualitas data melalui tahapan *preprocessing* dan validasi. Proses ini mencakup pemilihan dataset, *preprocessing* data, analisis data, visualisasi, hingga pengambilan *insight* yang dapat memberikan manfaat bagi pihak terkait, dalam hal ini Pemerintah Kota New York City (NYC).

Dataset yang dipilih untuk analisis ini adalah dataset kecelakaan kendaraan di NYC yang diperoleh dari Kaggle, dengan judul **"NYPD Vehicle Collisions"**. Dataset ini mencakup data kecelakaan kendaraan dari tahun 2015 hingga 2017, yang mencakup informasi seperti lokasi kecelakaan, jenis kendaraan, faktor penyebab, jumlah korban, dan waktu kejadian. Dataset ini dipilih karena relevansinya dalam memberikan *insight* untuk meningkatkan keselamatan lalu lintas, yang merupakan isu penting di kota besar seperti NYC.

Laporan ini akan membahas secara rinci setiap langkah yang dilakukan, mulai dari pemilihan dataset, *preprocessing* data (penanganan *missing values*, penghapusan duplikat, normalisasi, pengecekan konsistensi, dan deteksi *outlier*), analisis data, visualisasi, hingga pengambilan *insight* bisnis. Semua proses dilakukan menggunakan bahasa pemrograman Python dengan pustaka seperti Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, dan Plotly.

Langkah 1: Pemilihan Dataset

1.1 Deskripsi Dataset

Dataset yang digunakan dalam analisis ini adalah **"NYPD Vehicle Collisions"**, yang tersedia secara publik di Kaggle melalui tautan berikut: https://www.kaggle.com/datasets/nypd/vehicle-collisions. Dataset ini berisi catatan kecelakaan kendaraan di NYC yang dilaporkan oleh New York Police Department (NYPD) dari tahun 2015 hingga 2017. Dataset ini dipilih karena relevansinya dengan isu keselamatan lalu lintas, yang merupakan perhatian utama di kota metropolitan seperti NYC. Selain itu, dataset ini cukup besar dan memiliki variasi data yang memungkinkan analisis multidimensi, seperti analisis temporal, spasial, dan kategorikal.

Dataset ini memiliki **477.732 baris** dan **29 kolom** pada awalnya, dengan informasi yang mencakup:

- **UNIQUE KEY**: ID unik untuk setiap kecelakaan.
- **DATE** dan **TIME**: Tanggal dan waktu kejadian kecelakaan.
- **BOROUGH**: Wilayah di NYC tempat kecelakaan terjadi (misalnya, Brooklyn, Manhattan).
- **LATITUDE** dan **LONGITUDE**: Koordinat geografis lokasi kecelakaan.
- **VEHICLE 1 TYPE**, **VEHICLE 2 TYPE**, dst.: Jenis kendaraan yang terlibat dalam kecelakaan.
- **VEHICLE 1 FACTOR**, **VEHICLE 2 FACTOR**, dst.: Faktor penyebab kecelakaan untuk setiap kendaraan.
- **PERSONS INJURED**, **PERSONS KILLED**: Jumlah orang yang terluka atau meninggal.
- **PEDESTRIANS INJURED**, **CYCLISTS INJURED**, **MOTORISTS INJURED**, dst.: Jumlah korban spesifik berdasarkan kategori (pejalan kaki, pengendara sepeda, pengemudi).

1.2 Alasan Pemilihan Dataset

Dataset ini dipilih karena:

1. **Relevansi dengan Masalah Nyata**: Kecelakaan kendaraan adalah masalah besar di kota besar seperti NYC, dan analisis data ini dapat memberikan *insight* untuk meningkatkan keselamatan lalu lintas.
2. **Ketersediaan Data Multidimensi**: Dataset ini mencakup dimensi temporal (waktu), spasial (lokasi), kategorikal (jenis kendaraan, faktor penyebab), dan numerik (jumlah korban), sehingga memungkinkan analisis yang mendalam.
3. **Ukuran Dataset yang Cukup Besar**: Dengan 477.732 baris, dataset ini cukup besar untuk analisis statistik, tetapi masih dapat dikelola dengan perangkat komputasi standar.
4. **Tantangan dalam *Preprocessing***: Dataset ini memiliki banyak *missing values*, duplikat, dan potensi *outlier*, sehingga memberikan peluang untuk menerapkan teknik *preprocessing* secara menyeluruh.

1.3 Informasi Awal Dataset

Berikut adalah informasi awal dataset yang diperoleh saat pertama kali dimuat:

- **Jumlah Baris**: 477.732
- **Jumlah Kolom**: 29
- **Tipe Data**:
    - Numerik: 9 kolom (misalnya, PERSONS INJURED, LATITUDE, LONGITUDE).
    - Kategorikal: 16 kolom (misalnya, BOROUGH, VEHICLE 1 TYPE, VEHICLE 1 FACTOR).
    - Datetime: 1 kolom (DATE).
- **Memori yang Digunakan**: 105.7 MB.

Contoh 5 baris pertama dari kolom DATE:

0 2015-01-01
1 2015-01-01
2 2015-01-01
3 2015-01-01
4 2015-01-01

Statistik deskriptif awal menunjukkan adanya nilai ekstrem, seperti:

- PERSONS INJURED: Maksimum 32 orang, yang menunjukkan potensi *outlier*.
- LATITUDE dan LONGITUDE: Terdapat nilai 0, yang tidak valid untuk lokasi di NYC, menunjukkan adanya data yang salah atau hilang.
- BOROUGH: 139.342 nilai hilang (29,17% dari total data).

Langkah 2: *Preprocessing* Data

Langkah *preprocessing* dilakukan untuk memastikan kualitas data sebelum analisis lebih lanjut. Langkah-langkah yang dilakukan meliputi: penanganan *missing values*, penghapusan duplikat, normalisasi dan standarisasi data, pengecekan konsistensi teks, serta deteksi dan penanganan *outlier*. Setiap langkah akan dijelaskan secara rinci di bawah ini, termasuk perbandingan sebelum dan sesudah *preprocessing*.

2.1 Penanganan *Missing Values*

2.1.1 Identifikasi Missing Values

Sebelum *preprocessing*, jumlah *missing values* pada setiap kolom dihitung untuk mengetahui tingkat kelengkapan data. Berikut adalah jumlah dan persentase *missing values* untuk beberapa kolom dengan nilai hilang terbanyak:

- **VEHICLE 5 TYPE**: 476.049 (99,65%).
- **VEHICLE 5 FACTOR**: 475.970 (99,63%).
- **VEHICLE 4 TYPE**: 470.901 (98,57%).
- **VEHICLE 4 FACTOR**: 470.500 (98,49%).

- **VEHICLE 3 TYPE**: 447.468 (93,67%).
- **OFF STREET NAME**: 419.221 (87,75%).
- **CROSS STREET NAME**: 142.158 (29,76%).
- **BOROUGH**: 139.342 (29,17%).
- **LATITUDE** dan **LONGITUDE**: 121.132 (25,36%).

Total *missing values* sebelum *preprocessing* adalah **4.254.154** nilai, yang menunjukkan bahwa dataset ini memiliki banyak data yang tidak lengkap.

2.1.2 Strategi Penanganan Missing Values

Berikut adalah strategi yang digunakan untuk menangani *missing values*:

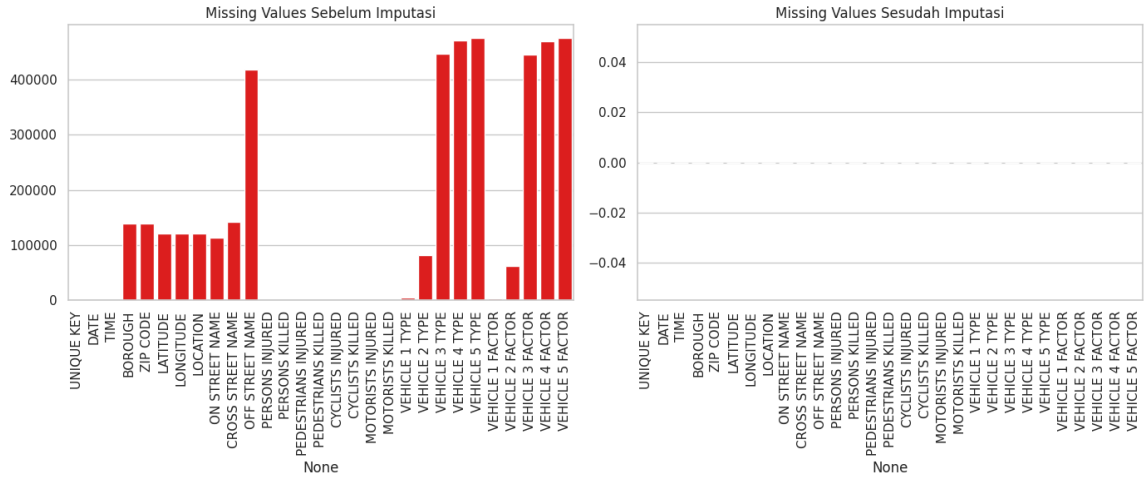
1. **Kolom Numerik**:
    - Kolom seperti LATITUDE dan LONGITUDE diisi dengan rata-rata (mean) dari kolom tersebut. Hal ini dilakukan karena koordinat geografis di NYC memiliki rentang yang relatif kecil, sehingga imputasi dengan rata-rata cukup representatif.
    - Kolom seperti PERSONS INJURED, PERSONS KILLED, dll., tidak memiliki *missing values*, sehingga tidak memerlukan imputasi.
2. **Kolom Kategorikal**:
    - Kolom seperti BOROUGH, VEHICLE 1 TYPE, dan VEHICLE 1 FACTOR diisi dengan nilai "UNKNOWN" untuk menandakan bahwa data tidak tersedia.
    - Kolom seperti VEHICLE 3 TYPE, VEHICLE 4 TYPE, dan VEHICLE 5 TYPE memiliki *missing values* yang sangat tinggi (lebih dari 90%) karena tidak semua kecelakaan melibatkan lebih dari 2 kendaraan. Kolom ini diisi dengan "NONE" untuk menunjukkan bahwa tidak ada kendaraan ketiga, keempat, atau kelima yang terlibat.
    - Kolom seperti ON STREET NAME, CROSS STREET NAME, dan OFF STREET NAME diisi dengan "UNKNOWN".
3. **Kolom Lokasi**:
    - Kolom LOCATION, yang merupakan kombinasi dari LATITUDE dan LONGITUDE, diisi berdasarkan nilai imputasi LATITUDE dan LONGITUDE.

2.1.3 Hasil Setelah Imputasi

Setelah imputasi, semua *missing values* berhasil ditangani. Berikut adalah jumlah dan persentase *missing values* setelah imputasi:

- **Semua kolom**: 0 (0%).

Perbandingan sebelum dan sesudah imputasi divisualisasikan dalam bentuk *bar plot*, yang menunjukkan bahwa semua *missing values* telah berhasil ditangani.



2.2 Penghapusan Duplikat

2.2.1 Identifikasi Duplikat

Duplikat diidentifikasi menggunakan kolom UNIQUE KEY, yang seharusnya bersifat unik untuk setiap kecelakaan. Namun, ditemukan bahwa terdapat **63.644 duplikat** berdasarkan semua kolom, yang setara dengan **13,32%** dari total data.

2.2.2 Penghapusan Duplikat

Duplikat dihapus menggunakan fungsi drop_duplicates() dari Pandas. Setelah penghapusan:

- **Jumlah duplikat**: 0.
- **Persentase duplikat**: 0%.
- **Jumlah baris setelah penghapusan duplikat**: 414.088 (dari 477.732 baris awal).

Penghapusan duplikat ini memastikan bahwa setiap baris mewakili kejadian kecelakaan yang unik, sehingga tidak ada data yang berulang yang dapat memengaruhi analisis.

2.3 Normalisasi dan Standarisasi Data

2.3.1 Kolom yang Dinormalisasi

Kolom Total Victims (dihitung sebagai PERSONS INJURED + PERSONS KILLED) dipilih untuk dinormalisasi dan distandarisasi karena merupakan variabel numerik utama yang akan dianalisis lebih lanjut. Kolom ini mencerminkan tingkat keparahan kecelakaan.

2.3.2 Metode Normalisasi dan Standarisasi

1. **Min-Max Scaling**:
    - Menggunakan MinMaxScaler dari Scikit-learn.
    - Rumus:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

    - Hasil: Nilai Total Victims diskalakan ke rentang [0, 1].

- Statistik sebelum normalisasi:

count	414088.000000
mean	0.243963
std	0.640842
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	32.000000

- Statistik setelah Min-Max Scaling:

count	414088.000000
mean	0.007624
std	0.020026
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

2. **Z-Score Standardization**:
   - Menggunakan StandardScaler dari Scikit-learn.
   - Rumus:

$$X_{std} = \frac{X - \mu}{\sigma}$$

   - Hasil: Nilai Total Victims memiliki rata-rata 0 dan standar deviasi 1.
   - Statistik setelah Z-Score Standardization:

count	414088.000000
mean	0.000000
std	1.000001
min	-0.380692
25%	-0.380692
50%	-0.380692
75%	-0.380692
max	49.553698

Normalisasi dan standarisasi ini mempermudah analisis lebih lanjut, terutama untuk deteksi *outlier* dan visualisasi data.

2.4 Pengecekan Konsistensi Teks

2.4.1 Identifikasi Inkonsistensi

Kolom-kolom kategorikal seperti BOROUGH, VEHICLE 1 TYPE, dan VEHICLE 1 FACTOR diperiksa untuk memastikan konsistensi format teks. Beberapa inkonsistensi yang ditemukan:

- Teks dengan huruf kapital dan kecil bercampur (misalnya, "PASSENGER VEHICLE" vs "passenger vehicle").
- Spasi berlebih atau karakter tidak diinginkan (misalnya, " TAXI " vs "TAXI").

2.4.2 Penyeragaman Teks

Langkah-langkah yang dilakukan:

1. Mengubah semua teks menjadi huruf kapital menggunakan metode str.upper().
2. Menghapus spasi berlebih menggunakan metode str.strip().
3. Mengganti nilai kosong atau tidak valid dengan "UNKNOWN".

2.4.3 Hasil Setelah Penyeragaman

Contoh 5 baris pertama setelah penyeragaman teks:

	BOROUGH	ON STREET NAME	VEHICLE 1 TYPE	VEHICLE 1 FACTOR
0	QUEENS	47 AVENUE	SPORT UTILITY/STATION WAGON	TRAFFIC CONTROL DISREGARDED
1	UNKNOWN	UNKNOWN	PASSENGER VEHICLE	ANIMALS ACTION
2	BROOKLYN	BEDFORD AVENUE	PASSENGER VEHICLE	FATIGUED/DROWSY
3	BROOKLYN	BUFFALO AVENUE	BUS	LOST CONSCIOUSNESS
4	UNKNOWN	RICHMOND TERRACE	UNKNOWN	UNSPECIFIED

Penyeragaman ini memastikan bahwa data kategorikal konsisten dan siap untuk analisis lebih lanjut, seperti pengelompokan atau visualisasi.

2.5 Deteksi dan Penanganan *Outlier*

2.5.1 Deteksi Outlier pada Total Victims

Metode Z-score digunakan untuk mendeteksi *outlier* pada kolom Total Victims. Langkah-langkahnya:

1. Menghitung Z-score menggunakan StandardScaler.
2. Menetapkan *threshold* Z-score sebesar 3 (sesuai standar untuk data yang mungkin memiliki distribusi *skewed*).
3. Mengidentifikasi *outlier* sebagai data dengan Z-score > 3.

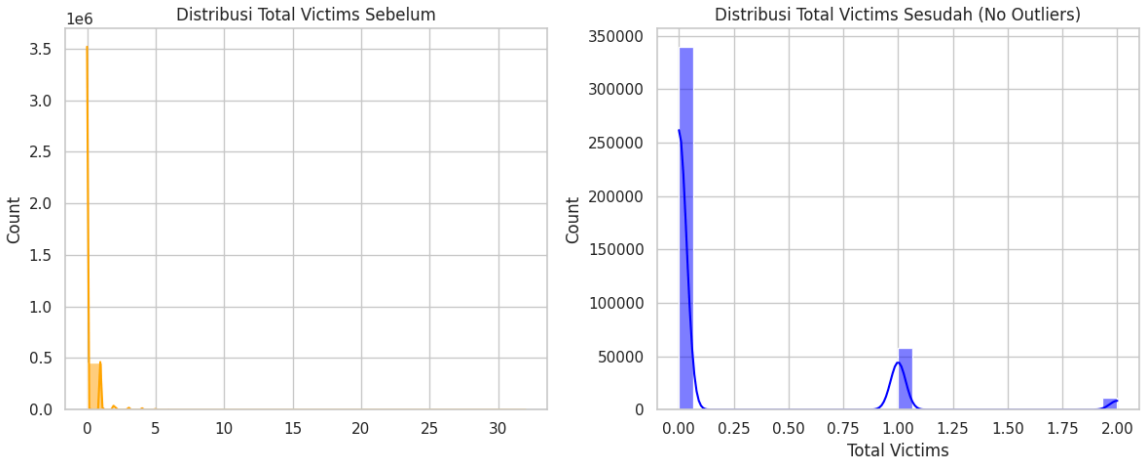
Hasil:

- **Jumlah *outlier***: 5.805 baris.
- **Persentase *outlier***: 1,40%.
- Statistik Total Victims setelah penghapusan *outlier*:

count	408283.000000
mean	0.193946
std	0.457472
min	0.000000
25%	0.000000
50%	0.000000

75%	0.000000
max	2.000000

Distribusi Total Victims sebelum dan sesudah penghapusan *outlier* divisualisasikan menggunakan histogram.



Histogram menunjukkan bahwa setelah penghapusan *outlier*, distribusi menjadi lebih terfokus pada nilai rendah (0 hingga 2 korban), yang lebih mencerminkan kecelakaan tipikal di NYC.

2.5.2 Deteksi Outlier pada LATITUDE dan LONGITUDE

Outlier pada LATITUDE dan LONGITUDE juga dideteksi menggunakan metode Z-score dengan *threshold* 3. Langkah ini dilakukan untuk memastikan bahwa lokasi kecelakaan berada dalam rentang yang valid untuk NYC. Hasil:

- **Jumlah *outlier* pada LATITUDE**: 14 baris.
- **Jumlah *outlier* pada LONGITUDE**: 22 baris.
- Setelah penghapusan *outlier* pada kedua kolom ini, jumlah baris berkurang menjadi **408.261**.

Penghapusan *outlier* ini memastikan bahwa analisis spasial (misalnya, visualisasi peta) hanya mencakup lokasi yang valid di NYC.

2.6 Ringkasan *Preprocessing*

Setelah semua langkah *preprocessing* selesai, berikut adalah ringkasan hasilnya:

- **Baris awal**: 477.732.
- **Baris setelah *preprocessing***: 408.261.
- *Missing values ditangani**: 4.254.154 nilai.
- **Duplikat dihapus**: 63.644 baris.
- *Outlier dihapus**: 5.827 baris (5.805 dari Total Victims + 22 dari LATITUDE dan LONGITUDE).

Langkah 3: Analisis dan Visualisasi Data

Setelah *preprocessing*, dataset siap untuk dianalisis dan divisualisasikan. Analisis dilakukan untuk menjawab pertanyaan-pertanyaan berikut:

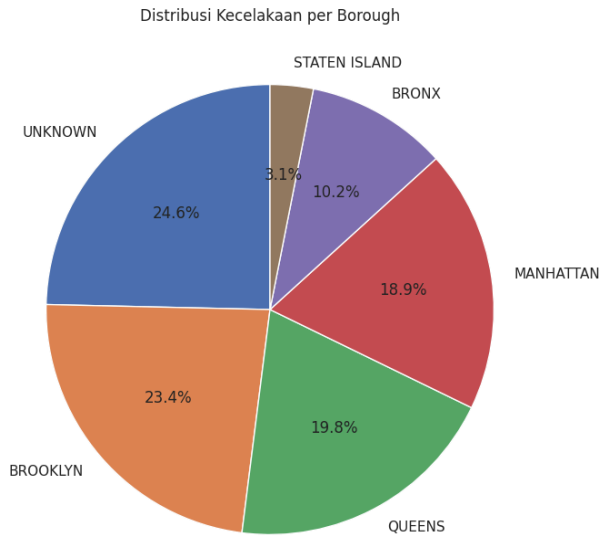
1. Bagaimana distribusi kecelakaan di berbagai wilayah (*borough*) di NYC?
2. Apa jenis kendaraan yang paling sering terlibat dalam kecelakaan?
3. Apa faktor penyebab utama kecelakaan?
4. Bagaimana tren kecelakaan dari waktu ke waktu (per bulan, per jam, per hari)?
5. Bagaimana distribusi korban berdasarkan kategori (pejalan kaki, pengendara sepeda, pengemudi)?
6. Bagaimana distribusi lokasi kecelakaan berdasarkan koordinat geografis?

3.1 Distribusi Kecelakaan per *Borough*

Distribusi kecelakaan dihitung berdasarkan kolom BOROUGH. Hasilnya:

- **UNKNOWN**: 100.615 kecelakaan (24,64%).
- **BROOKLYN**: 95.340 kecelakaan (23,35%).
- **QUEENS**: 80.923 kecelakaan (19,82%).
- **MANHATTAN**: 77.088 kecelakaan (18,88%).
- **BRONX**: 41.593 kecelakaan (10,19%).
- **STATEN ISLAND**: 12.702 kecelakaan (3,11%).

Visualisasi dilakukan menggunakan *pie chart*, dapat dilihat sebagai berikut ini:



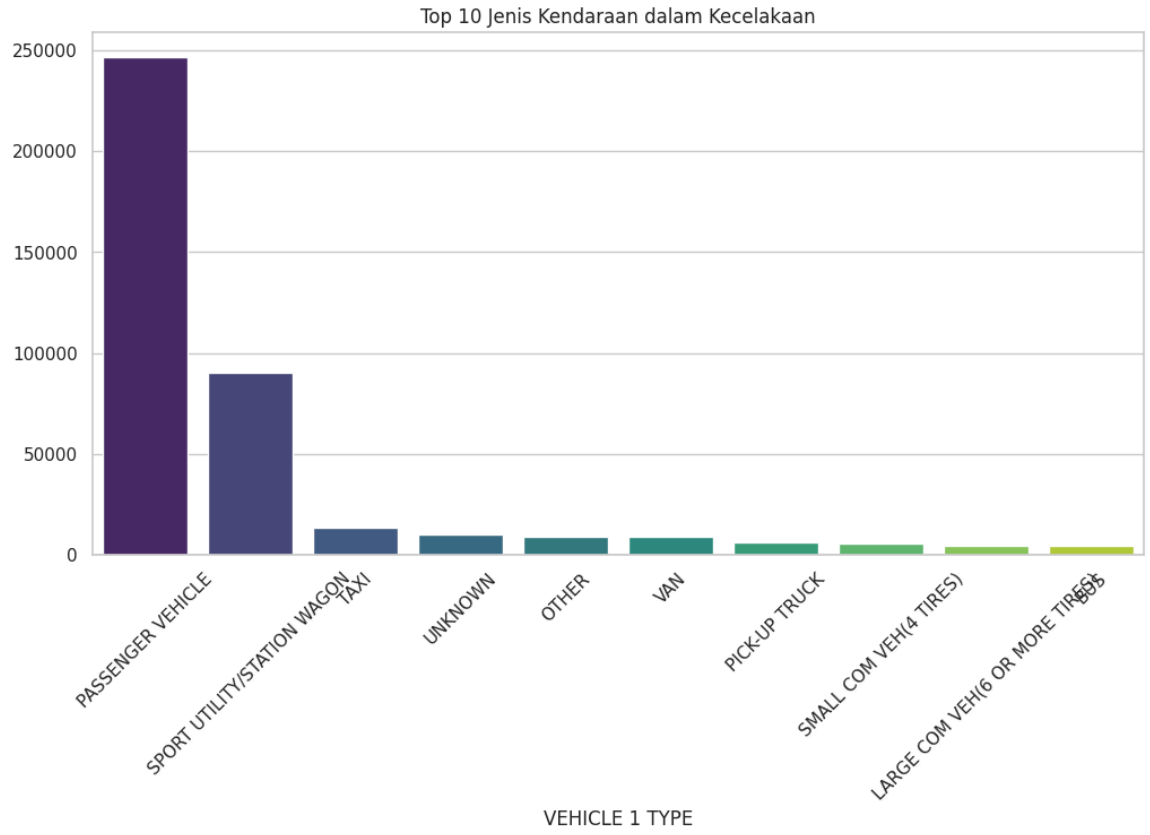
Pie chart ini menunjukkan bahwa Brooklyn dan Manhattan adalah wilayah dengan jumlah kecelakaan tertinggi, yang kemungkinan disebabkan oleh kepadatan lalu lintas dan populasi di kedua wilayah tersebut.

3.2 Jenis Kendaraan yang Paling Sering Terlibat

Jenis kendaraan dianalisis berdasarkan kolom VEHICLE 1 TYPE. Berikut adalah 10 jenis kendaraan teratas:

- **PASSENGER VEHICLE**: 246.436 kecelakaan (60,36%).
- **SPORT UTILITY/STATION WAGON**: 90.330 kecelakaan (22,13%).
- **TAXI**: 13.812 kecelakaan (3,38%).
- **UNKNOWN**: 10.270 kecelakaan (2,52%).
- **OTHER**: 9.253 kecelakaan (2,27%).
- **VAN**: 9.025 kecelakaan (2,21%).
- **PICK-UP TRUCK**: 6.543 kecelakaan (1,60%).
- **SMALL COM VEH(4 TIRES)**: 5.529 kecelakaan (1,35%).
- **LARGE COM VEH(6 OR MORE TIRES)**: 4.594 kecelakaan (1,13%).
- **BUS**: 4.527 kecelakaan (1,11%).

Visualisasi dilakukan menggunakan *bar plot*, dapat dilihat sebagai berikut:



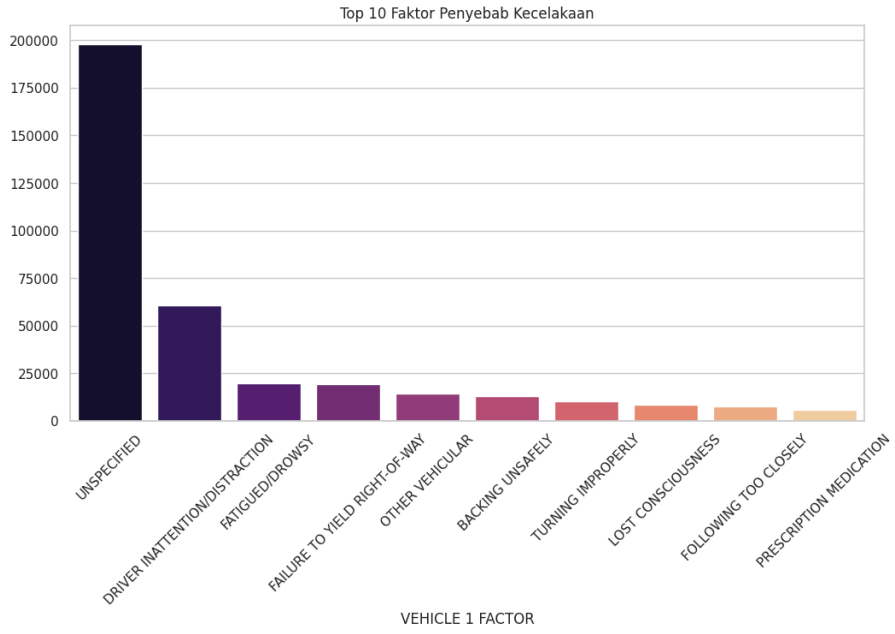
Bar plot ini menunjukkan bahwa kendaraan penumpang dan SUV mendominasi kecelakaan, yang mencerminkan prevalensi jenis kendaraan ini di NYC.

3.3 Faktor Penyebab Kecelakaan

Faktor penyebab dianalisis berdasarkan kolom VEHICLE 1 FACTOR. Berikut adalah 10 faktor teratas:

- **UNSPECIFIED**: 197.976 kecelakaan (48,49%).
- **DRIVER INATTENTION/DISTRACTION**: 60.702 kecelakaan (14,87%).
- **FATIGUED/DROWSY**: 19.561 kecelakaan (4,79%).
- **FAILURE TO YIELD RIGHT-OF-WAY**: 19.188 kecelakaan (4,70%).
- **OTHER VEHICULAR**: 14.471 kecelakaan (3,54%).
- **BACKING UNSAFELY**: 12.943 kecelakaan (3,17%).
- **TURNING IMPROPERLY**: 10.191 kecelakaan (2,50%).
- **LOST CONSCIOUSNESS**: 8.430 kecelakaan (2,06%).
- **FOLLOWING TOO CLOSELY**: 7.540 kecelakaan (1,85%).
- **PRESCRIPTION MEDICATION**: 5.951 kecelakaan (1,46%).

Visualisasi dilakukan menggunakan *bar plot*, yaitu sebagai berikut:

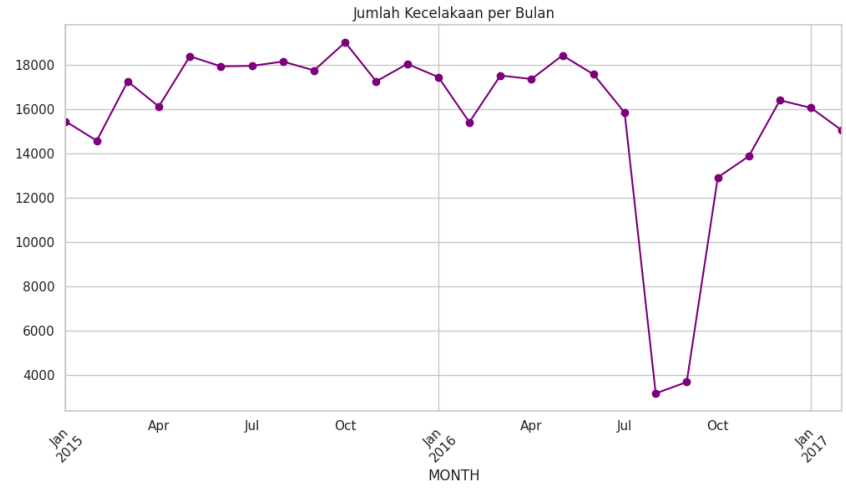


Bar plot ini menunjukkan bahwa faktor "UNSPECIFIED" sangat dominan, yang menunjukkan perlunya pelaporan yang lebih rinci. Faktor utama yang dapat ditindaklanjuti adalah "DRIVER INATTENTION/DISTRACTION", yang menunjukkan bahwa pengemudi sering kehilangan fokus saat mengemudi.

3.4 Tren Kecelakaan Berdasarkan Waktu

3.4.1 Tren per Bulan

Tren kecelakaan per bulan dihitung dengan mengelompokkan data berdasarkan DATE. Hasilnya divisualisasikan menggunakan *line plot*, yaitu sebagai berikut:



Beberapa temuan:

- Puncak kecelakaan terjadi pada **Oktober 2015** (18.999 kecelakaan).
- Penurunan tajam terjadi pada **Agustus 2016** (3.174 kecelakaan), yang mungkin disebabkan oleh kelengkapan data yang rendah pada bulan tersebut.
- Secara keseluruhan, kecelakaan cenderung lebih tinggi pada bulan-bulan musim panas (Mei hingga Oktober), yang mungkin terkait dengan peningkatan aktivitas di jalan raya.

3.4.2 Kecelakaan per Jam

Distribusi kecelakaan per jam dihitung berdasarkan kolom TIME. Hasilnya:

- **Puncak kecelakaan**: Jam 16:00 (30.029 kecelakaan) dan 17:00 (29.018 kecelakaan), yang merupakan jam sibuk sore hari.
- **Kecelakaan terendah**: Jam 03:00 (4.281 kecelakaan), yang merupakan waktu dengan aktivitas lalu lintas rendah.

Visualisasi dilakukan menggunakan histogram, yang menjadi bagian dari dashboard temporal (lihat bagian 3.7).

3.4.3 Kecelakaan per Hari

Distribusi kecelakaan per hari dihitung berdasarkan hari dalam seminggu. Hasilnya:

- **Jumat**: 64.616 kecelakaan (hari dengan kecelakaan tertinggi).
- **Minggu**: 48.213 kecelakaan (hari dengan kecelakaan terendah).

Visualisasi ini juga menjadi bagian dari dashboard temporal.

3.4.4 Kecelakaan per Bulan (Nama Bulan)

Distribusi kecelakaan berdasarkan nama bulan (tanpa mempedulikan tahun) menunjukkan:

- **Januari**: 48.917 kecelakaan (tertinggi).
- **September**: 21.418 kecelakaan (terendah).

Visualisasi ini juga menjadi bagian dari dashboard temporal.

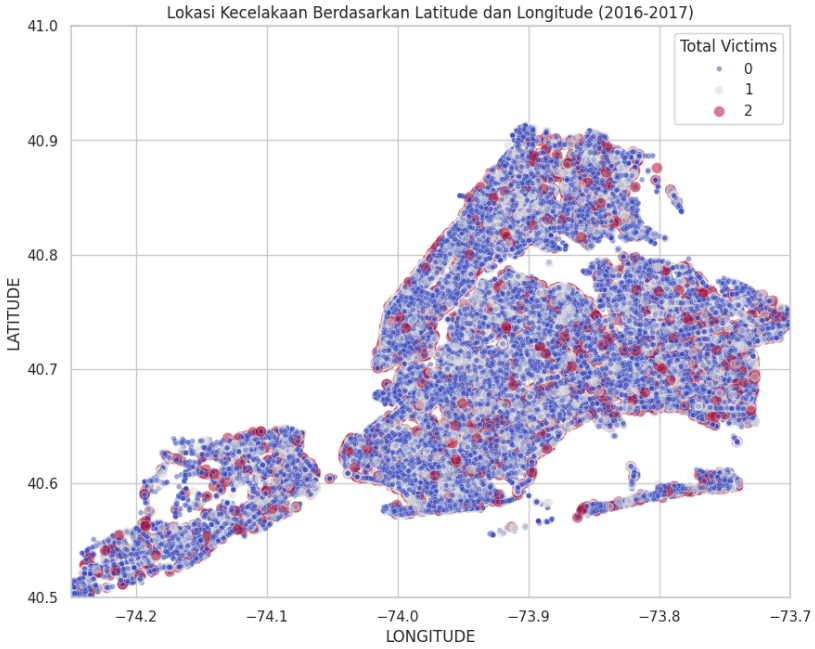
3.5 Distribusi Korban

Distribusi korban dihitung berdasarkan kategori:

- **Pejalan Kaki**: 20.575 korban (20.357 terluka, 218 meninggal).
- **Pengendara Sepeda**: 7.824 korban (7.796 terluka, 28 meninggal).
- **Pengemudi**: 54.722 korban (54.598 terluka, 124 meninggal).

3.6 Analisis Lokasi Kecelakaan

Lokasi kecelakaan divisualisasikan berdasarkan LATITUDE dan LONGITUDE menggunakan *scatter plot*, yaitu sebagai berikut :



Beberapa temuan:

- Kecelakaan terkonsentrasi di wilayah Manhattan, Brooklyn, dan Queens, yang sesuai dengan distribusi per *borough*.
- Titik-titik dengan Total Victims lebih tinggi (ditandai dengan warna merah) tersebar merata di seluruh NYC, menunjukkan bahwa kecelakaan dengan korban besar tidak terbatas pada wilayah tertentu.

Statistik koordinat setelah penghapusan *outlier*:

- **LATITUDE**:

  count 408261.000000
  mean 40.723621
  std 0.071949
  min 40.499135

25%	40.678455
50%	40.722105
75%	40.760776
max	40.912884

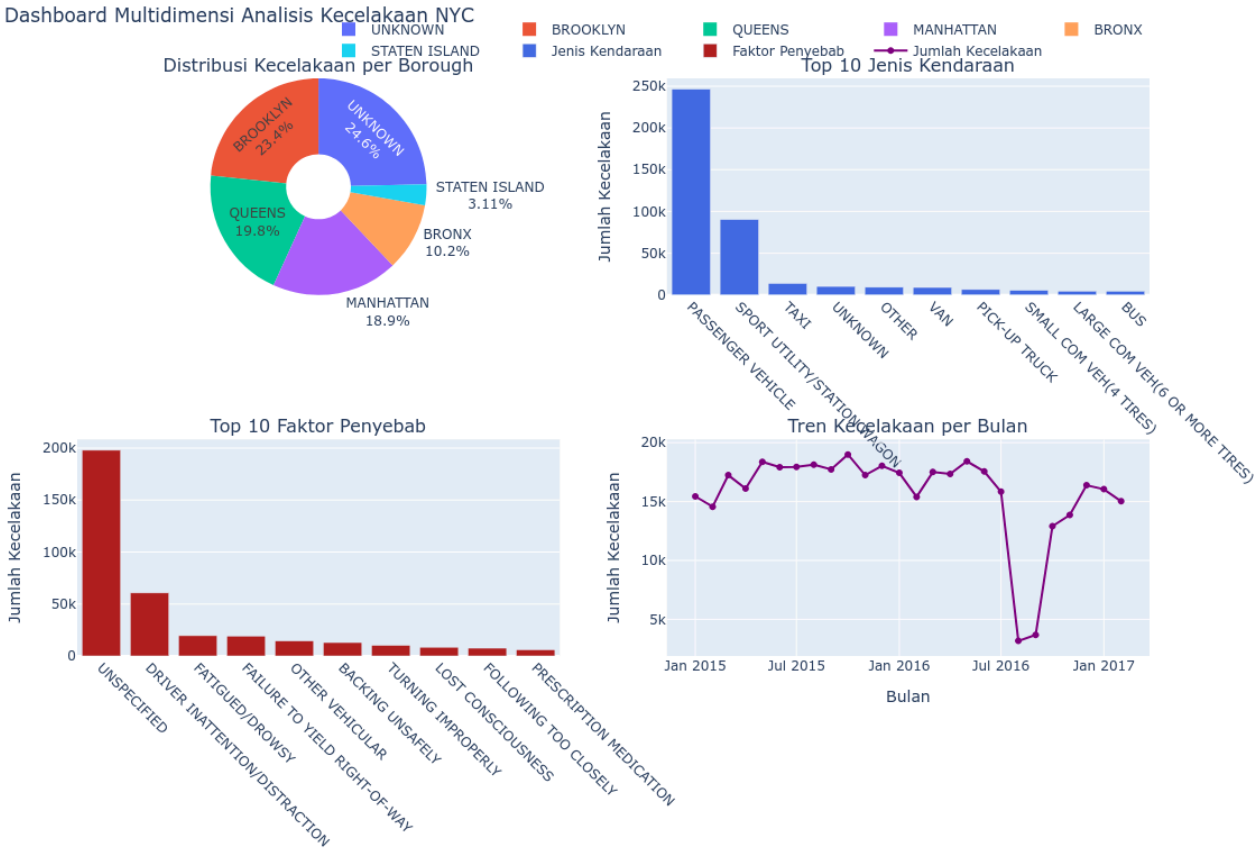
- **LONGITUDE**:

count	408261.000000
mean	-73.920531
std	0.079089
min	-74.253031
25%	-73.969360
50%	-73.919663
75%	-73.880908
max	-73.700597

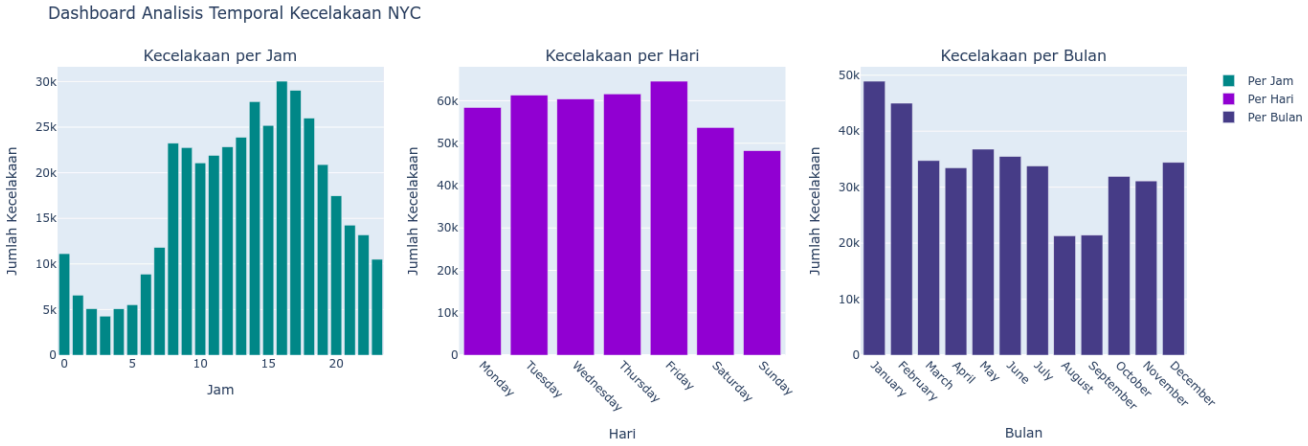
3.7 Dashboard Interaktif

Tiga dashboard interaktif dibuat menggunakan Plotly untuk memberikan visualisasi yang lebih mendalam:

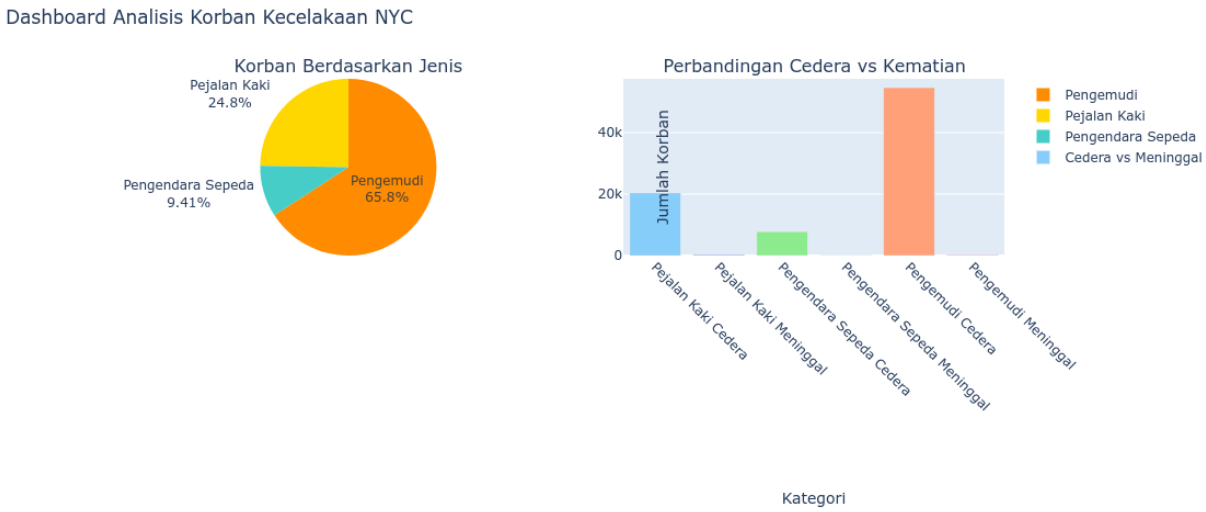
1. **Dashboard Multidimensi**: Mencakup distribusi per *borough*, top 10 jenis kendaraan, top 10 faktor penyebab, dan tren per bulan :



2. **Dashboard Temporal**: Mencakup distribusi kecelakaan per jam, per hari, dan per bulan :



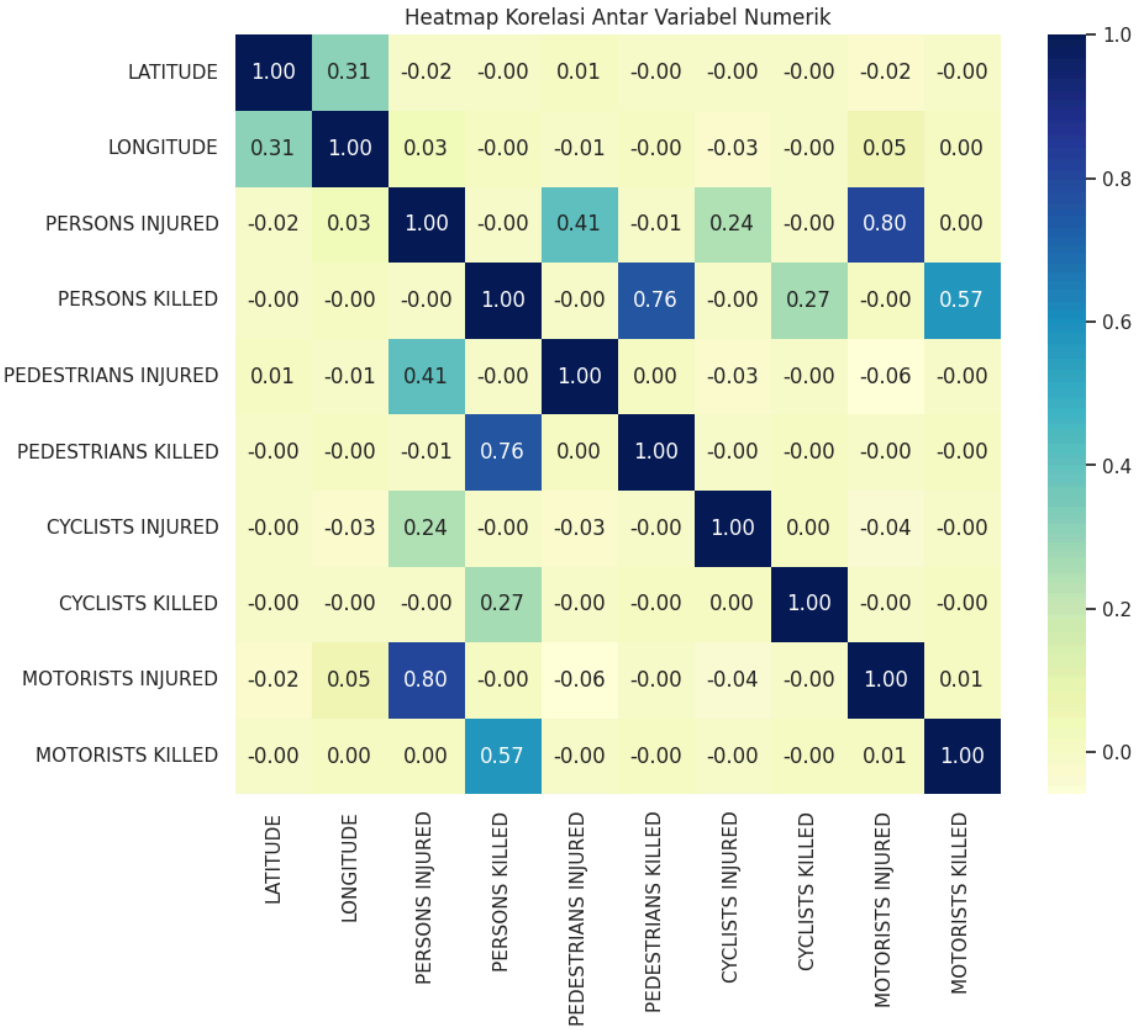
3. **Dashboard Korban**: Mencakup perbandingan korban berdasarkan kategori (pejalan kaki, pengendara sepeda, pengemudi) dan status (terluka atau meninggal) :



3.8 Korelasi Antar Variabel Numerik

Korelasi antar variabel numerik dihitung menggunakan metode Pearson dan divisualisasikan dalam bentuk *heatmap*. Hasilnya:

- **Korelasi Tinggi**:
  - PERSONS INJURED dan MOTORISTS INJURED: 0,80 (karena pengemudi merupakan komponen utama dari total korban).
  - PERSONS KILLED dan PEDESTRIANS KILLED: 0,76 (mayoritas korban meninggal adalah pejalan kaki).
- **Korelasi Rendah**:
  - LATITUDE dan LONGITUDE dengan jumlah korban: <0,05 (lokasi tidak berkorelasi langsung dengan jumlah korban).



Langkah 4: *Insight* Bisnis untuk Pemerintah NYC

Berdasarkan analisis dan visualisasi yang telah dilakukan, berikut adalah *insight* bisnis yang dapat digunakan oleh Pemerintah NYC untuk meningkatkan keselamatan lalu lintas:

1. **Fokus pada Faktor Penyebab Utama**:
    - Faktor utama kecelakaan adalah "DRIVER INATTENTION/DISTRACTION" (14,87%). Pemerintah dapat meluncurkan kampanye kesadaran untuk mengurangi penggunaan ponsel saat mengemudi, seperti melalui iklan di media sosial atau papan reklame.
    - Faktor "UNSPECIFIED" yang sangat tinggi (48,49%) menunjukkan perlunya pelaporan yang lebih rinci oleh petugas kepolisian. Pemerintah dapat melatih petugas untuk mencatat faktor penyebab dengan lebih spesifik.
2. **Waktu Rawan Kecelakaan**:
    - Kecelakaan paling sering terjadi pada jam 14:00-17:00, yang merupakan jam sibuk sore hari. Pemerintah dapat meningkatkan patroli lalu lintas dan menambah rambu peringatan pada jam-jam ini, terutama di wilayah dengan lalu lintas padat seperti Brooklyn dan Manhattan.
    - Jumat adalah hari dengan kecelakaan tertinggi (64.616 kecelakaan), mungkin karena aktivitas akhir pekan. Pemerintah dapat mengadakan operasi khusus pada hari Jumat untuk mengurangi pelanggaran lalu lintas.
3. **Lokasi Rawan Kecelakaan**:
    - Brooklyn (23,35%) dan Manhattan (18,88%) memiliki jumlah kecelakaan tertinggi. Pemerintah perlu mengevaluasi infrastruktur lalu lintas di wilayah ini, seperti menambah lampu lalu lintas, zebra cross, atau jalur sepeda yang lebih aman.
    - Peta lokasi kecelakaan menunjukkan konsentrasi tinggi di pusat kota. Pemerintah dapat memasang kamera pengawas di titik-titik rawan untuk mencegah pelanggaran.
4. **Jenis Kendaraan yang Sering Terlibat**:
    - Kendaraan penumpang (60,36%) dan SUV (22,13%) paling sering terlibat dalam kecelakaan. Pemerintah dapat mempertimbangkan regulasi khusus untuk pengemudi jenis kendaraan ini, seperti pelatihan keselamatan tambahan atau pemeriksaan kendaraan secara berkala.
5. **Perlindungan bagi Pejalan Kaki dan Pengendara Sepeda**:
    - Pejalan kaki (20.575 korban) dan pengendara sepeda (7.824 korban) merupakan kelompok yang rentan. Pemerintah dapat meningkatkan fasilitas keselamatan, seperti jalur sepeda yang terpisah dan trotoar yang lebih lebar, terutama di wilayah dengan kepadatan tinggi seperti Manhattan.
6. **Tren Musiman**:

- Kecelakaan cenderung lebih tinggi pada bulan-bulan musim panas (Mei hingga Oktober). Pemerintah dapat meningkatkan kampanye keselamatan selama periode ini, seperti mengingatkan pengemudi untuk lebih berhati-hati saat libur musim panas.

Kesimpulan

Proses *preprocessing* dilakukan secara menyeluruh untuk memastikan kualitas data, meliputi:

- Penanganan *missing values* (4.254.154 nilai ditangani).
- Penghapusan duplikat (63.644 baris dihapus).
- Normalisasi dan standarisasi data pada kolom Total Victims.
- Penyeragaman teks pada kolom kategorikal.
- Deteksi dan penghapusan *outlier* (5.827 baris dihapus).

Hasil analisis menunjukkan bahwa kecelakaan di NYC dipengaruhi oleh berbagai faktor, seperti kurangnya perhatian pengemudi, waktu sibuk, dan lokasi dengan kepadatan tinggi. Visualisasi yang dibuat memberikan gambaran yang jelas tentang distribusi kecelakaan, jenis kendaraan, faktor penyebab, tren temporal, dan distribusi korban. *Insight* yang dihasilkan dapat menjadi dasar bagi Pemerintah NYC untuk merancang kebijakan yang lebih efektif dalam meningkatkan keselamatan lalu lintas.