Laporan Praktikum
Data Science

Hubungan Pola Penggunaan dan Jenis Mode Transportasi terhadap Persentase Pengguna di Masyarakat

Nama Anggota :
41425064 Yosevyn Sipahutar
41425081 Lukas Simatupang
41425083 Inez Cecilia Tiurma

INSTITUT TEKNOLOGI DEL
FAKULTAS VOKASI
Sarjana Terapan Teknologi Rekayasa Perangkat Lunak

BAB I
PENDAHULUAN

1. Latar Belakang

Transportasi memiliki peran vital dalam mendukung aktivitas sosial, ekonomi, dan mobilitas masyarakat. Seiring meningkatnya jumlah penduduk dan perkembangan wilayah perkotaan, kebutuhan akan sistem transportasi yang efisien, aman, dan berkelanjutan semakin mendesak. Pemahaman terhadap pola penggunaan mode transportasi menjadi hal penting bagi pemerintah dan lembaga perencana untuk merumuskan kebijakan transportasi yang tepat sasaran.

Melalui analisis data, pola penggunaan moda transportasi dapat diidentifikasi secara objektif berdasarkan variabel seperti wilayah, etnis, dan tahun penggunaan. Analisis ini tidak hanya membantu mengetahui mode transportasi mana yang paling dominan digunakan oleh masyarakat, tetapi juga menilai sejauh mana perbedaan preferensi transportasi terjadi antar kelompok pengguna. Dengan adanya perbedaan signifikan antar mode transportasi, kebijakan publik dapat diarahkan untuk memperbaiki aksesibilitas, mengurangi kemacetan, serta mendorong penggunaan moda yang lebih ramah lingkungan.

Pemanfaatan metode statistik seperti *Kruskal–Wallis Test* memungkinkan peneliti untuk mengukur perbedaan yang signifikan antar kelompok mode transportasi, sementara visualisasi data melalui *boxplot*, *violin plot*, dan *bar plot* memberikan gambaran yang lebih intuitif terhadap sebaran dan proporsi penggunaannya. Dengan demikian, penelitian ini tidak hanya berfokus pada analisis deskriptif, tetapi juga mendukung pengambilan keputusan berbasis data (*data-driven decision making*) dalam sektor transportasi.

2. Rumusah Masalah

Terdapat rumusan masalah dari pengerjaan ini adalah sebagai berikut.

a. Mode transportasi apa yang memiliki rata-rata percent terbesar?

b. Apakah terdapat perbedaan signifikan percent antar mode transportasi?

3. Tujuan

Tujuan dari pengerjaan ini yaitu, untuk memberikan pemahaman yang lebih mendalam mengenai pola dan perbedaan penggunaan moda transportasi di masyarakat.

Pada tujuan pertama, yaitu mengetahui distribusi penggunaan mode transportasi, dilakukan analisis deskriptif terhadap data untuk melihat bagaimana setiap jenis moda digunakan oleh masyarakat dalam berbagai kondisi dan kelompok wilayah. Analisis ini membantu mengidentifikasi moda transportasi yang paling dominan serta moda yang jarang digunakan, sehingga dapat memberikan gambaran umum tentang preferensi masyarakat dalam memilih alat transportasi.

Sementara itu, tujuan kedua, yaitu menilai apakah ada perbedaan signifikan persentase penggunaan mode antar kelompok mode (mode_name), dilakukan melalui pendekatan statistik menggunakan uji Kruskal–Wallis Test. Uji ini digunakan untuk mengetahui apakah perbedaan rata-rata atau median antar kelompok mode transportasi benar-benar signifikan secara statistik. Dengan demikian, hasil analisis ini dapat memberikan dasar yang kuat dalam memahami perilaku pengguna transportasi serta mendukung pengambilan keputusan yang lebih tepat dalam perencanaan sistem transportasi yang efisien dan berkelanjutan.

BAB II
METODE PENELITIAN

1. Data Collection

Analisis ini menggunakan dataset yang diperoleh dari platform terbuka Kaggle**,** yang berisi informasi terkait pola penggunaan berbagai mode transportasi oleh masyarakat berdasarkan wilayah, etnis, dan tahun. Dataset ini berjudul *Mode of Transportation Dataset* dan berfungsi untuk menganalisis distribusi serta perbedaan tingkat penggunaan antar kelompok moda transportasi.

Dataset ini terdiri dari 24 kolom dan 202.203 baris data yang mencakup atribut seperti *mode_name* (jenis moda transportasi), *percent* (persentase pengguna), *pop_total* (jumlah populasi), *reportyear* (tahun laporan), serta beberapa variabel demografis lainnya. Data ini merepresentasikan hasil survei penggunaan transportasi di berbagai wilayah dan kelompok masyarakat.

2. Data Visualization

Tahap data visualization dilakukan untuk menampilkan hasil analisis secara visual agar pola dan perbedaan penggunaan mode transportasi dapat terlihat dengan jelas. Dalam penelitian ini digunakan beberapa bentuk visualisasi seperti boxplot**,** violin **plot**, dan bar plot. Boxplot digunakan untuk melihat sebaran data dan median dari setiap mode transportasi, di mana hasilnya menunjukkan bahwa mode *Car, Truck, or Van: Drove Alone* memiliki persentase penggunaan tertinggi. Violin plot memperlihatkan bentuk distribusi data pada setiap mode, dan hasilnya memperkuat bahwa moda tersebut paling dominan digunakan. Sementara itu, bar plot digunakan untuk menampilkan rata-rata persentase penggunaan tiap mode, yang juga menunjukkan bahwa moda kendaraan pribadi memiliki rata-rata tertinggi dibandingkan moda lain seperti sepeda atau berjalan kaki. Secara keseluruhan, visualisasi ini membantu memperjelas hasil analisis bahwa terdapat perbedaan nyata dalam tingkat penggunaan antar mode transportasi di masyarakat.

3. Data Processing and Techniques (Advance Preprocessing)

Tahap data processing and techniques (advance preprocessing) dilakukan untuk membersihkan dan menyiapkan data sebelum dianalisis. Proses ini mencakup penghapusan tanda titik pada kolom numerik, pengubahan tipe data menjadi numerik, serta pengisian nilai kosong dengan median atau modus agar data lebih lengkap dan konsisten. Selain itu, dilakukan penanganan *outlier* menggunakan metode IQR dan Winsorization untuk menjaga kestabilan data. Langkah-langkah ini memastikan dataset siap digunakan dan hasil analisis menjadi lebih akurat serta terpercaya.

3.1. Menghapus tanda titik (.) dari kolom numerik

Terdapat 24 kolom dan 202.203 baris yang dihapus tanda (.) yang digunakan untuk membuat nilai dapat diubah menjadi format numerik tanpa error dan memastikan data konversi ke string sebelum diproses.

```
df['pop_total'] = df['pop_total'].astype(str).str.replace('.', '', regex=False)
df['percent'] = df['percent'].astype(str).str.replace('.', '', regex=False)
```

3.2. Mengubah tipe data menjadi numerik

Setelah tanda titik dihapus, tahap selanjutnya adalah mengubah tipe data menjadi numerik yang digunakan agar data dapat diolah dalam analisis statistik dan visualisasi numerik.

```
df['pop_total'] = pd.to_numeric(df['pop_total'], errors='coerce')
df['percent'] = pd.to_numeric(df['percent'], errors='coerce')
```

3.3. Mengisi Nilai Kosong (Missing Values)

Selanjutnya setelah tipe data diubah menjadi numerik, tahapan selanjutnya adalah mengubah nilai yang hilang diisi (imputasi) agar dataset tidak memiliki missing values.

```
df['percent'] = df['percent'].fillna(df['percent'].median())
df['pop_total'] = df['pop_total'].fillna(df['pop_total'].median())
df['mode_name']=df['mode_name'].fillna(df['mode_name'].mode()[0])
df['reportyear'] = df['reportyear'].fillna(df['reportyear'].mode()[0])
```

3.4. Mengecek Nilai yang Hilang

Pengecekan ini dilakukan apakah masih ada nilai NaN setelah dilakukan pengisian. Sehingga kolom dipastikan sudah bersih dan data kosong harus bernilai 0.

```
print("Missing values sesudah imputation:")
print(df.isnull().sum())
```

3.5. Mendeteksi Dan Menangani Outliner Dengan Metode IQR

Pengerjaan ini menggunakan metode IQR (Interquartile Range) untuk mendeteksi outlier. Q1 = kuartil bawah (25%), Q3 = kuartil atas (75%), IQR = selisih antar kuartil. Dimana, nilai yang dianggap outlier adalah yang lebih kecil dari lower_bound atau lebih besar dari upper_bound.

```
def remove_outliers_iqr(df, column):

 Q1 = df[column].quantile(0.25)

Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1 lower_bound = Q1 - 1.5 * IQR

upper_bound = Q3 + 1.5 * IQR
```

3.6. Menangani Outlier dengan Winsorization

Metode ini disebut Winsorization, yaitu cara memperbaiki data ekstrem tanpa menghapusnya. Nilai yang lebih kecil dari lower_bound diganti dengan batas bawah, Nilai yang lebih besar dari upper_bound diganti dengan batas atas. Tujuan nya adalah menjaga agar data tetap utuh tapi tidak terdistorsi oleh nilai ekstrem

```
df[column] = np.where(df[column] < lower_bound, lower_bound, np.where(df[column] > upper_bound, upper_bound, df[column]))
```

BAB III
HASIL DAN PEMBAHASAN

1. Modeling

Modeling adalah tahap dalam analisis data di mana model atau algoritma pembelajaran mesin digunakan untuk mengambil wawasan atau membuat prediksi berdasarkan data yang telah dipersiapkan sebelumnya. Tahap modeling adalah proses yang iteratif, dan beberapa percobaan mungkin diperlukan untuk mencapai model yang optimal. Selama iterasi, analisis hasil dan evaluasi model membimbing penyesuaian dan perbaikan untuk mencapai performa model yang diinginkan.

```
from scipy.stats import kruskal

# --- Kruskal-Wallis Test ---
kruskal_stat, kruskal_p = kruskal(*groups)

print("\n=== Kruskal-Wallis Test ===")
print(f"H-statistic : {kruskal_stat:.4f}")
print(f"p-value     : {kruskal_p:.6f}")

# Interpretasi hasil
if kruskal_p < 0.05:
    print("Kesimpulan : Ada perbedaan signifikan antar mode_name (p < 0.05)")
else:
    print("Kesimpulan : Tidak ada perbedaan signifikan antar mode_name (p >= 0.05)")
```

```
=== Kruskal-Wallis Test ===
H-statistic : 43594.2026
p-value     : 0.000000
Kesimpulan : Ada perbedaan signifikan antar mode_name (p < 0.05)
```

Kode tersebut menjalankan **uji Kruskal–Wallis** menggunakan fungsi kruskal() dari library scipy.stats untuk mengetahui apakah terdapat **perbedaan signifikan** antara beberapa kelompok data independen (groups) berdasarkan nilai median mereka. Hasil uji menghasilkan dua output utama, yaitu nilai **H-statistic** dan **p-value**. Nilai H-statistic menunjukkan seberapa besar perbedaan antar kelompok, sedangkan p-value digunakan untuk menentukan signifikansinya. Jika p-value < 0,05, maka disimpulkan bahwa terdapat perbedaan signifikan antar kelompok (seperti pada hasil yang ditampilkan, p = 0.000000 < 0.05), artinya setidaknya satu kelompok memiliki perbedaan signifikan dibanding kelompok lainnya**.**

Statistical analysis

```
# Filter data untuk mode transportasi yang relevan dengan kemacetan
congestion_modes = ['Car, Truck, or Van: Drove Alone', 'Car, Truck, or Van: Carpooled', 'Public Transportation']
df_congestion = df[df['mode_name'].isin(congestion_modes)]

# Tampilkan statistik deskriptif untuk 'percent' pada mode-mode ini
print("=== Statistik Deskriptif untuk 'percent' pada Mode Terkait Kemacetan ===")
display(df_congestion.groupby('mode_name')['percent'].describe())

# Hitung persentase total pengguna kendaraan pribadi vs transportasi publik (jika data memungkinkan)
# Ini mungkin memerlukan agregasi lebih lanjut atau mempertimbangkan kolom pop_total
# untuk saat ini, kita fokus pada persentase per mode seperti di kolom 'percent'
```

```
=== Statistik Deskriptif untuk 'percent' pada Mode Terkait Kemacetan ===
```

mode_name	count	mean	std	min	25%	50%	75%	max
Car, Truck, or Van: Carpooled	28886.0	1.156383e+14	1.967419e+14	0.0	124.0	6.339144e+13	1.390589e+14	8.311068e+14
Car, Truck, or Van: Drove Alone	28886.0	2.534409e+14	3.326137e+14	0.0	773.0	6.339144e+13	6.619147e+14	8.311068e+14
Public Transportation	28886.0	1.247077e+14	2.097841e+14	0.0	18.0	4.308738e+13	1.375020e+14	8.311068e+14

Terdapat kategori 3 mode transportasi yang di gabung dengan kolom persen untuk menganalisis kemacetan yang disebabkan per kategori.

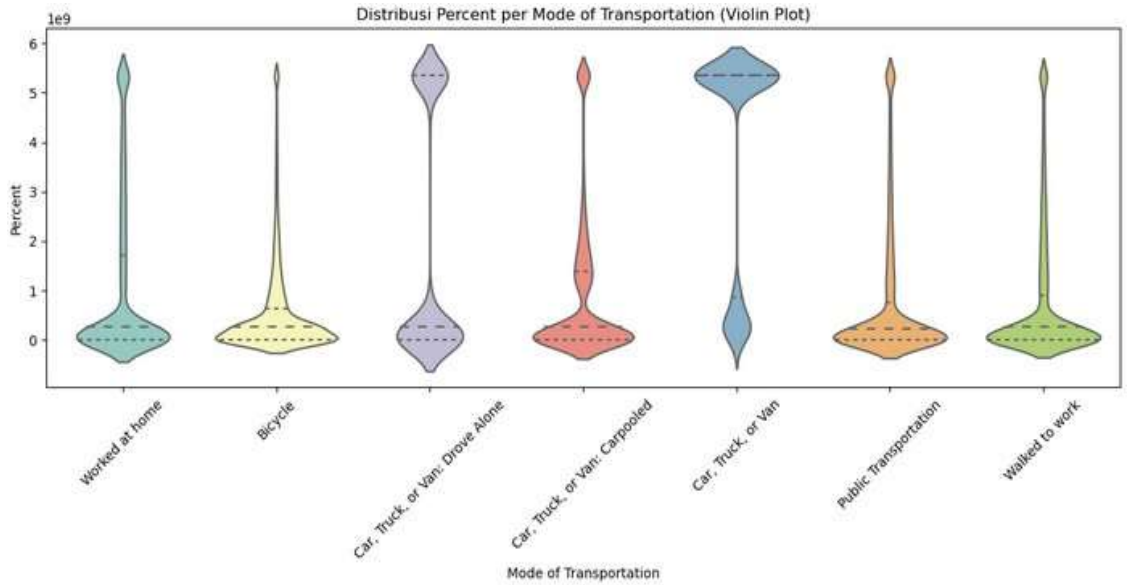
```
import matplotlib.pyplot as plt
import seaborn as sns

# Atur ukuran figure
plt.figure(figsize=(12, 6))

# Buat boxplot untuk melihat distribusi 'percent' per mode_name
sns.boxplot(data=df, x='mode_name', y='percent', palette='Set2')

plt.title('Distribusi Percent per Mode of Transportation')
plt.xlabel('Mode of Transportation')
plt.ylabel('Percent')
plt.xticks(rotation=45)
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

Kode diatas digunakan untuk membuat visualisasi dari data presentase dari mode

transaportasi dengan boxplot. Dengan menggunakan `matplotlib` dan `seaborn`. grafik yang dihasilkan menunjukkan bagaimana nilai `percent` tersebar untuk setiap kategori `mode_name`. Setiap kotak dalam boxplot menggambarkan sebaran data: median, kuartil bawah (Q1), kuartil atas (Q3).

Hasil visualisasinya menunjukkan bahwa mode transportasi "Car, Truck, or Van: Drove Alone" memiliki median tertinggi dan sebaran paling lebar, artinya mode ini paling dominan digunakan oleh responden. Sementara mode seperti "Worked at Home" dan "Bicycle" memiliki median yang lebih rendah dan sebaran yang sempit, menandakan bahwa mode tersebut kurang umum digunakan.

```
plt.figure(figsize=(12, 6))
sns.violinplot(data=df, x='mode_name', y='percent', palette='Set3', inner='quart')
plt.title('Distribusi Percent per Mode of Transportation (Violin Plot)')
plt.xlabel('Mode of Transportation')
plt.ylabel('Percent')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

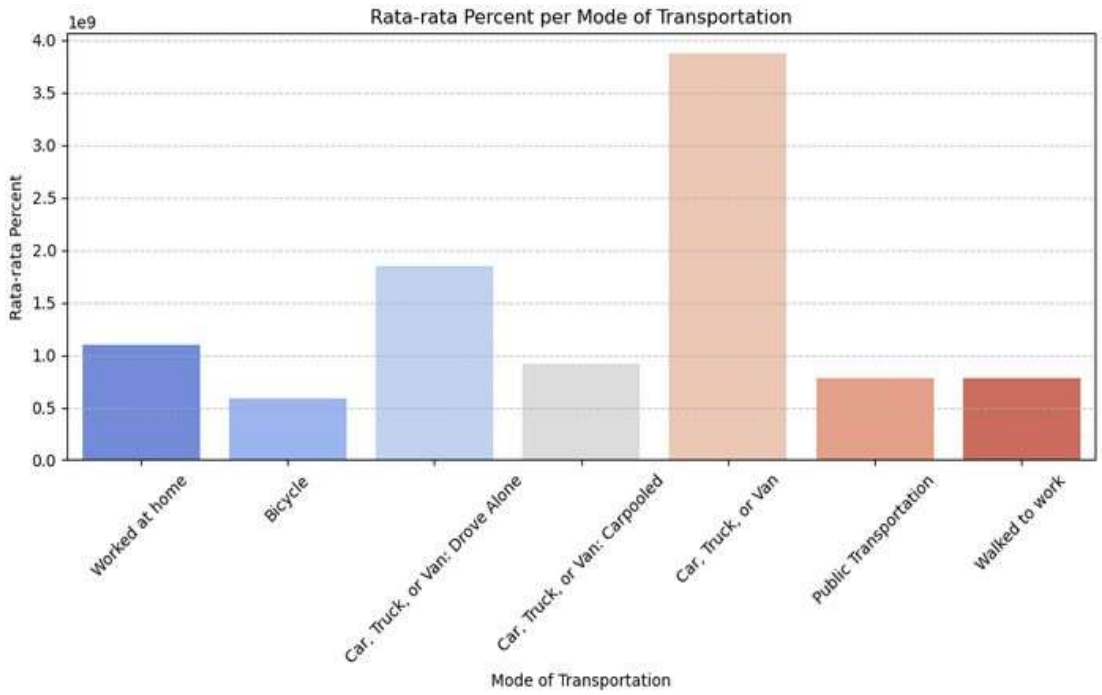


Kode ini digunakan untuk membuat violin plot yang menampilkan distribusi nilai `percent` untuk setiap kategori `mode_name` dalam DataFrame. Violin plot ini menunjukkan bentuk sebaran data secara visual, termasuk kepadatan, median, dan kuartil. Dengan parameter `inner='quart'`, grafik juga menampilkan garis kuartil di dalam bentuk violin.

Dapat dilihat untuk hasil dari visualisasi menunjukkan bentuk violin yang berbeda-beda untuk setiap mode transportasi. Perbedaan bentuk ini terjadi karena distribusi data pada tiap kategori tidak sama. Jika bentuk violin simetris dan ramping, itu menandakan data tersebar merata dan tidak terlalu banyak variasi. Sebaliknya, jika bentuknya melebar di bagian tertentu, itu menunjukkan konsentrasi data yang tinggi di rentang nilai tersebut. Misalnya,

mode "Car, Truck, or Van: Drove Alone" memiliki bentuk violin yang lebar di bagian tengah, menandakan banyak data berkumpul di sekitar median. Mode seperti "Worked at Home" atau "Bicycle" memiliki bentuk yang lebih sempit dan kadang tidak simetris, menunjukkan distribusi data yang lebih kecil dan mungkin adanya outlier.

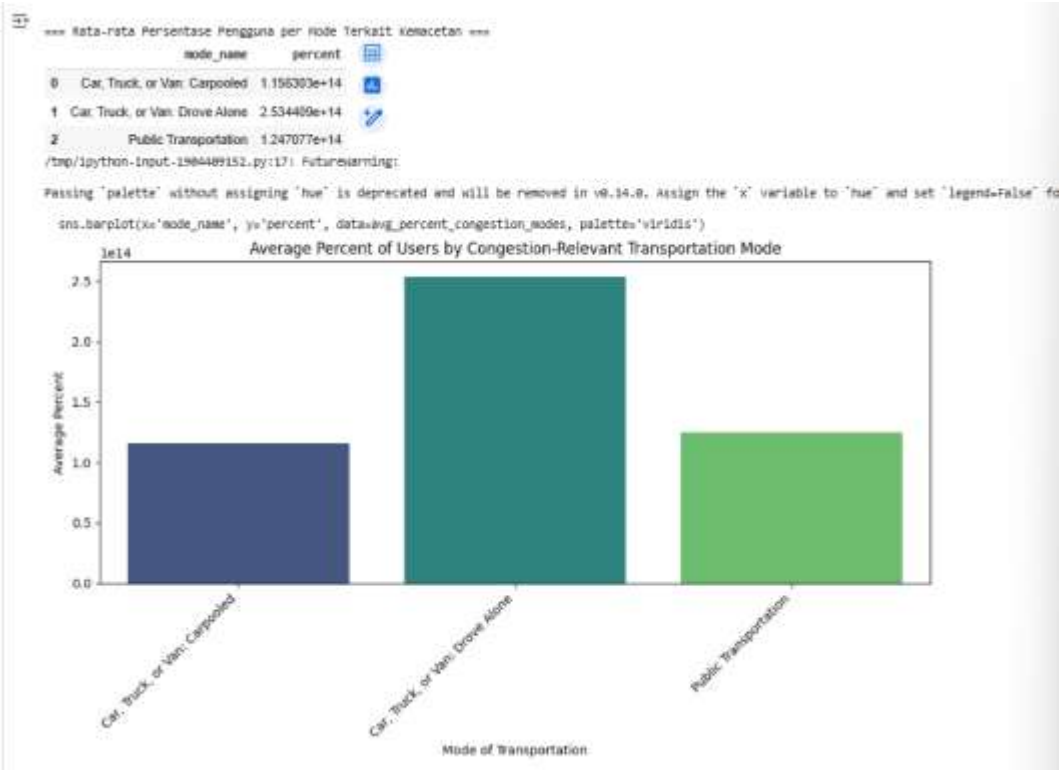
```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10, 6))
sns.barplot(data=df, x='mode_name', y='percent', estimator='mean', ci=None, palette='coolwarm')
plt.title('Rata-rata Percent per Mode of Transportation')
plt.xlabel('Mode of Transportation')
plt.ylabel('Rata-rata Percent')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



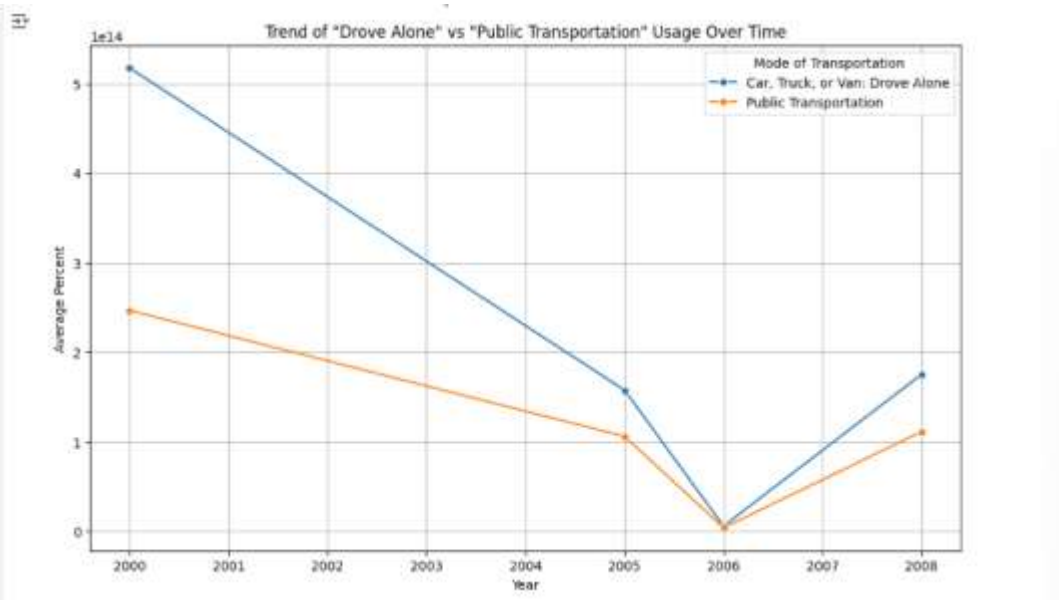
Kode tersebut digunakan untuk membuat grafik batang (bar plot) yang menampilkan rata- rata nilai `percent` untuk setiap kategori `mode_name` dalam DataFrame. Dengan menggunakan pustaka `matplotlib` dan `seaborn`, grafik ini menyajikan perbandingan visual antar mode transportasi berdasarkan seberapa besar rata-rata persentase penggunaannya. Parameter `estimator='mean'` memastikan bahwa nilai yang ditampilkan adalah rata-rata, bukan median atau total, dan `ci=None` menonaktifkan tampilan interval kepercayaan agar grafik lebih bersih.

Hasil dari visualisasi menunjukkan bahwa mode transportasi "Car, Truck, or Van" memiliki batang tertinggi, menandakan bahwa mode ini paling banyak digunakan secara rata-rata oleh

responden. Mode seperti "Walked to work" dan "Public Transportation" memiliki batang yang jauh lebih pendek, menunjukkan tingkat penggunaan yang lebih rendah.



Terdapat 3 kategori yang berpotensi menyebabkan kemacetan volume kategori transportasi drove alone sangat tinggi penggunaannya dibandingkan carpooled dan public transportation



Perbandingan moda transportasi drove alone vs public transportation dari data ini dapat di lihat pertahun 2000 sampai 2008 ada perubahan signifikan dari kategori drove alone yang menurun setiap tahun bahkan penggunaan sampai setara dengan public transportation pada tahun 2006. Kami berasumsi pemerintah sedang berupaya menangani masalah masalah yang disebabkan oleh transportasi contohnya kemacetan, polusi, dll

```
=== Potensi Kemacetan Berdasarkan Rata-rata Persentase Pengguna 'Drove Alone' per County ===
```

	county_name	percent
29	Orange	72.675942
42	Santa Clara	67.970219
18	Los Angeles	65.787669
47	Solano	65.551593
0	Alameda	64.691248
55	Ventura	64.647046
36	San Diego	62.969571
35	San Bernardino	62.472849
40	San Mateo	62.228016
33	Sacramento	61.705696

	county_name	percent
22	Mendocino	34.183215
13	Inyo	33.111150
17	Lassen	32.867144
52	Trinity	30.875336
1	Alpine	23.339323
24	Modoc	22.013827
31	Plumas	18.866688
21	Mariposa	18.619423
45	Sierra	16.884769
25	Mono	15.128770

Prediksi terhadap country yang berpotensi mengalami masalah masalah transportasi seperti kemacetan sehinggan hanya menggunakan persentase pengguanaan transportasi yang berkategori drove alone

BAB IV
KESIMPULAN

1. Kesimpulan

Berdasarkan tujuan pertama, yaitu mengetahui distribusi penggunaan mode transportasi, hasil analisis menunjukkan bahwa setiap mode transportasi memiliki pola penggunaan yang berbeda di masyarakat. Dari hasil visualisasi menggunakan boxplot, violin plot, dan diagram batang, terlihat bahwa mode seperti Car, Truck, or Van memiliki rata-rata persentase penggunaan tertinggi, menandakan bahwa mode ini paling banyak digunakan oleh masyarakat. Sementara itu, mode seperti sepeda, berjalan kaki, dan transportasi umum memiliki persentase penggunaan yang lebih rendah, meskipun tetap digunakan oleh kelompok tertentu. Hal ini menggambarkan bahwa preferensi masyarakat terhadap mode transportasi sangat bervariasi, tergantung pada kebutuhan, kenyamanan, dan kondisi lingkungan masing-masing.

Dengan hasil nya kita dapat menilai dan melihat apakah ada perbedaan signifikan persentase penggunaan antar kelompok mode (mode_name), dilakukan pengujian statistik menggunakan Kruskal–Wallis Test. Hasil pengujian menunjukkan bahwa nilai p-value < 0.05, sehingga dapat disimpulkan bahwa terdapat perbedaan signifikan antar kelompok mode transportasi. Artinya, penggunaan setiap mode tidak bersifat merata dan terdapat perbedaan yang nyata dalam tingkat penggunaannya. Dengan demikian, hasil analisis ini telah menjawab kedua tujuan utama penelitian, yaitu berhasil mengidentifikasi distribusi penggunaan mode transportasi dan membuktikan adanya perbedaan signifikan antar kelompok mode. Kesimpulan ini memberikan gambaran bahwa pilihan mode transportasi masyarakat dipengaruhi oleh berbagai faktor seperti aksesibilitas, jarak tempuh, dan preferensi pribadi, sehingga dapat menjadi bahan pertimbangan dalam penyusunan kebijakan transportasi yang lebih tepat sasaran dan untuk mengatasi masalah kemacetan, polusi, dll pemerintah california harus fokus terhadap penggunaan moda drove alone yang sangat tinggi seperti di daerah Orange dengan persentase 72%, santa clara dengan persentase 67%, los angles 65% dan lainnya.