Capstone Project

MODULE 2

ELISA | JCDSOL-016 Group 2

Case Study :
New York City TLC Trip Record

Milestone

NYC TAXI & Limousine Commission

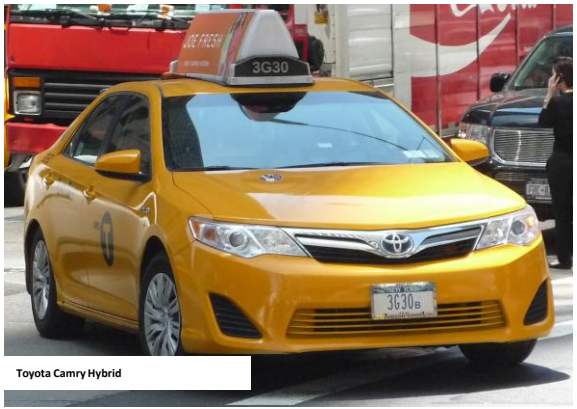
Tahun	Milestone
1897	**Awal Mula Taksi Listrik** - Taksi listrik pertama kali diperkenalkan, diikuti oleh taksi berbahan bakar bensin di awal 1900-an.
1915	**Kemunculan Taksi Kuning** - Harry N. Allen meluncurkan armada taksi berwarna kuning untuk meningkatkan visibilitas dan daya tarik.
1937	**Sistem Medali Taksi** - NYC menerapkan sistem lisensi medali untuk mengatur jumlah taksi dan mencegah kelebihan armada.
1950-1980	**Era Checker Cab** - Checker cabs yang ikonik terkenal dengan kabin luas dan keawetannya, menjadi pilihan populer.
1990an	**Penerapan Teknologi Baru - Pemasangan argo, GPS, dan sistem pembayaran kartu kredit** untuk meningkatkan kenyamanan penumpang.
2010an	**Disrupsi Layanan Ride-Hailing - Kehadiran Uber, Lyft,** dan layanan berbasis aplikasi lainnya mengubah lanskap industri taksi.
2020an	**Inisiatif Hijau & Elektrifikasi** - Peralihan ke kendaraan listrik dan hybrid untuk mengurangi emisi dan memodernisasi armada taksi.

Armada & Kapasitas

Nissan NV200

Ford Crown Victoria

Toyota Prius



Toyota Camry Hybrid

- Tahun 2016, baik CMT maupun VeriFone melayani armada yang terdiri dari berbagai kendaraan dengan kapasitas standar untuk taksi, yaitu 4 hingga 5 penumpang.
- Namun, **untuk kapasitas 6 penumpang**, biasanya diperlukan kendaraan yang lebih besar seperti **van atau SUV**.

NY Taxi SUV

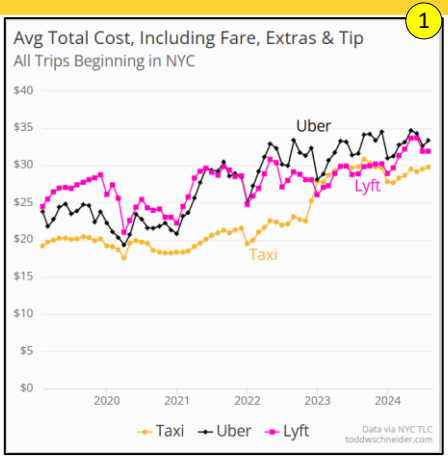
Vendor

Dua perusahaan utama, **Creative Mobile Technologies (CMT)** dan **VeriFone Inc.**, bertanggung jawab atas sistem pembayaran dan teknologi dalam taksi di New York City (**Yellow Cabs** dan **Green Cabs).**

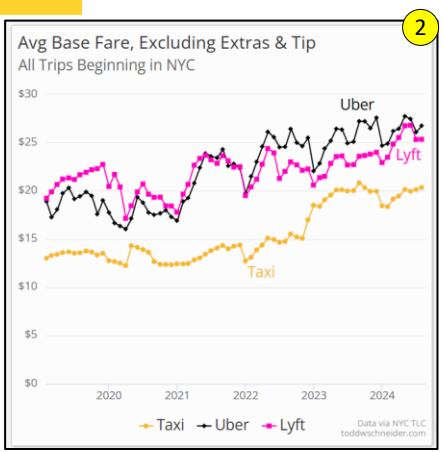
Latar Belakang Masalah

NYC Taxi, Uber & Lyft Price Comparison | To : All

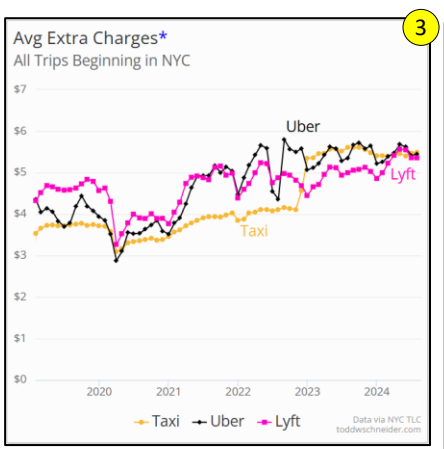
1



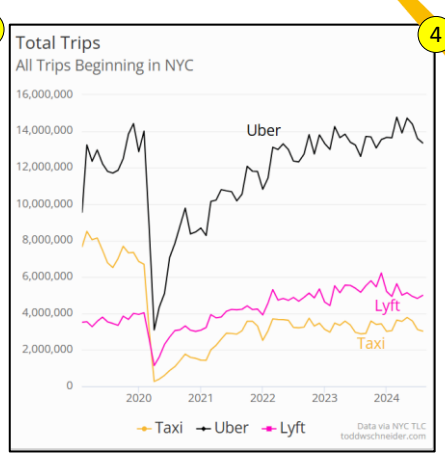
2



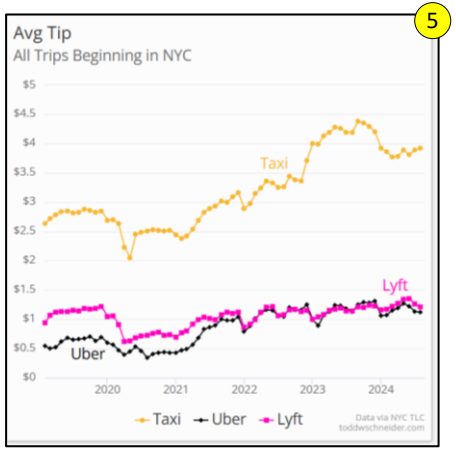
3



4



5



Penjelasan dari Chart 1-5 adalah sbb:

Chart	Metrik	Taksi	Uber	Lyft	Ancaman Utama & Opportunity
1	Rata-rata Biaya Total	↑ (sedikit)	↑ (signifikan)	↑ (signifikan)	**Inflasi dan Kenaikan Biaya Operasional:** Ancaman terbesar bagi ketiga layanan. **Opportunity:** Mencari cara untuk mengoptimalkan rute, mengurangi waktu kosong kendaraan, dan negosiasi ulang kontrak dengan pemasok bahan bakar.
2	Rata-rata Tarif Dasar	↑ (sedikit)	↑ (signifikan)	↑ (signifikan)	**Persaingan Harga:** Tekanan untuk menaikkan tarif dasar dapat membuat pelanggan beralih ke pesaing. **Opportunity:** Menawarkan layanan tambahan bernilai tambah (misal, hiburan dalam perjalanan, pilihan kendaraan) untuk membenarkan kenaikan tarif.
3	Rata-rata Biaya Tambahan	↑ (sedikit)	Fluktuatif	Fluktuatif	**Regulasi Pemerintah:** Perubahan regulasi terkait biaya tambahan (misal, biaya kemacetan) dapat mempengaruhi pendapatan. **Opportunity:** Memantau perkembangan regulasi dan beradaptasi dengan cepat.
4	Jumlah Perjalanan Total	↓ (signifikan)	↑ (signifikan)	↑ (signifikan)	**Pergeseran Preferensi Pelanggan:** Pelanggan semakin memilih layanan *ride-hailing* karena kemudahan dan fleksibilitas. **Opportunity:** Meningkatkan kualitas layanan, menawarkan program loyalitas, dan memperluas jangkauan layanan.
5	Rata-rata Tip	↑ (signifikan)	↑ (signifikan)	↑ (signifikan)	**Persaingan Tip:** Peningkatan persaingan membuat pengemudi harus memberikan layanan yang lebih baik untuk mendapatkan tip yang lebih tinggi. Opportunity: Mengimplementasikan sistem rating yang transparan untuk mendorong perilaku pengemudi yang baik.

Sumber :https://toddwschneider.com/dashboards/nyc-taxi-uber-lyft-fare-and-driver-pay-comparison/?from=all&to=all&since=2019

"**Rising costs** and **falling trips**—are we really surviving, or just hanging on?"

1. Domain / Business Knowledge

- **Industri**: **Transportasi** (layanan Taksi);
- **Area Layanan**: **Kota New York**, dengan kepadatan penumpang tinggi dan tantangan lalu lintas yang sering terjadi.
- **Operasi Bisnis**: Meliputi **analisis permintaan perjalanan dan optimisasi rute**
- **Tantangan**: **Bersaing dengan layanan ride-sharing,** menyeimbangkan *supply-demand* saat jam sibuk, dan kepatuhan regulasi (misal: tarif kemacetan).

2. Konteks Pemahaman Bisnis

- Analisis ini **bertujuan untuk memahami operasional taksi di NYC, khususnya pola perjalanan, puncak permintaan, distribusi tarif, dan faktor yang memengaruhi durasi perjalanan serta kepuasan pelanggan.**
- Ini membantu memberikan wawasan untuk **efisiensi, optimisasi tarif, dan alokasi sumber daya.**

3. Masalah Bisnis yang harus diselesaikan

Optimalisasi Rute : Memaksimalkan efisiensi perjalanan.**Tujuan**:

- Memperpendek durasi perjalanan.
- Mengurangi biaya operasional, terutama bahan bakar.

Optimalisasi Armada : Pengelolaan armada kendaraan secara keseluruhan. **Tujuan**:

- Mengelola dan memanfaatkan kendaraan secara efektif untuk memenuhi permintaan penumpang.
- Menjadwalkan kendaraan untuk memaksimalkan penggunaan dan efisiensi.

4. Stakeholders / Audiens

- BOD (Board of Director) Operational
- Divisi relevan:
  1. Operational: Optimisasi jadwal dan rute pengemudi;
  2. Keuangan: Analisis struktur tarif dan dampak pendapatan;
  3. Pengalaman Pelanggan: Meningkatkan kepuasan berdasarkan pola perjalanan dan permintaan.
- SHE : Memastikan kepatuhan terhadap regulasi lalu lintas dan lingkungan

Note : SHE (Safety, Health & Environment)

5. Business Goals

1. **Meningkatkan Efisiensi Operasional** : Meminimalkan durasi perjalanan, mengurangi waktu tunggu, dan optimisasi rute.
2. **Meningkatkan Pendapatan:** Memaksimalkan pendapatan per perjalanan melalui peningkatan efisiensi dan penyesuaian tarif saat permintaan tinggi.
3. **Meningkatkan Kepuasan Pelanggan** : Mengurangi keterlambatan perjalanan dan mengoptimalkan layanan saat jam sibuk.- Kepatuhan
4. **Regulasi:** Memastikan kepatuhan terhadap regulasi lalu lintas di NYC.

6. Scope of Business

- **Fokus Geografis:** Area operasional taksi di New York City, menggunakan data dari NYC Taxi and Limousine Commission (TLC).
- **Waktu Analisis:** Data dari Desember 2022 hingga Januari 2023 untuk mengidentifikasi tren musiman dan perubahan perilaku pengguna.
- **Karakteristik Perjalanan:**
  - Analisis frekuensi perjalanan (jam dan hari).
  - Rute dan tujuan perjalanan yang umum.
  - Durasi perjalanan dan waktu tunggu.
  - Biaya perjalanan dan tarif.

6. Scope of Business...

- **Perilaku Pengguna:** Analisis demografi dan pola penggunaan untuk memahami preferensi pengguna.
- **Metodologi:** Gunakan analisis statistik deskriptif dan regresi untuk menggambarkan data dan faktor-faktor yang mempengaruhi perilaku pengguna.

6. Business Questions

1. Jam berapa saja yang potensial untuk dilakukan pricing strategy untuk memaksimalkan revenue bagi perusahaan?
2. Upaya apa saja yang dapat dilakukan untuk meningkatkan pendapatan di shit malam untuk menekan disparitas siang v.s malam?
3. Pembayaran tipe apa yang perlu diperbanyak untuk meningkatkan kemudahan pelanggan?
4. Bagaimana trend disparitas speed per jam di semua vendor?

Data Dictionary

No.	Kolom	Deskripsi
1	**VendorID**	**ID penyedia layanan taksi :** **1 = Creative Mobile Technologies LLC.** **2 = VeriFone Inc.**
2	**lpep_pickup_datetime**	Tanggal dan waktu saat penjemputan terjadi.
3	**lpep_dropoff_datetime**	Tanggal dan waktu saat pengantaran terjadi.
4	**store_and_fwd_flag**	Menunjukkan apakah trip disimpan untuk diteruskan; berkaitan dengan akurasi dan integritas data.
5	**RatecodeID**	Kode yang menunjukkan jenis tarif yang diterapkan:
		1: Tarif Dasar
		2: Tarif ke Bandara JFK
		3: Tarif ke Bandara Newark
		4: Tarif ke Nassau atau Westchester
		5: Tarif yang dinegosiasikan
		6: Tarif untuk perjalanan kelompok
6	**PULocationID**	Identifikasi lokasi penjemputan.
7	**DOLocationID**	Identifikasi lokasi pengantaran.
8	**passenger_count**	Jumlah penumpang dalam kendaraan.
9	**trip_distance**	Jarak yang ditempuh selama perjalanan (Miles).
10	**fare_amount**	Total biaya yang dikenakan untuk perjalanan (USD).

No.	Kolom	Deskripsi
11	**extra**	Biaya tambahan yang dikenakan (USD) untuk kondisi tertentu.
12	**mta_tax**	Pajak yang dikenakan oleh MTA (USD).
13	**tip_amount**	Tip yang diberikan kepada pengemudi (USD).
14	**tolls_amount**	Total biaya tol yang dibayarkan (USD).
15	**ehail_fee**	Biaya yang dikenakan saat menggunakan aplikasi untuk memesan taksi (USD).
16	**improvement_surcharge**	Biaya perbaikan yang dikenakan (USD).
17	**total_amount**	Total biaya perjalanan termasuk semua biaya (USD).
18	**payment_type**	Metode pembayaran yang digunakan:
		1. Credit card: Kartu kredit
		2.Cash: Tunai
		3. No charge: Tidak dikenakan biaya
		4. Dispute: Perselisihan
		5. Unknown: Tidak diketahui
		6. Voided trip: Perjalanan dibatalkan
19	**trip_type**	Menunjukkan jenis perjalanan:
		Type 1: Pemanggilan taksi melalui aplikasi
		Type 2: Perjalanan dengan reservasi sebelumnya (Dispatch)
20	**congestion_surcharge**	Biaya yang dikenakan saat terjadi kemacetan (USD).

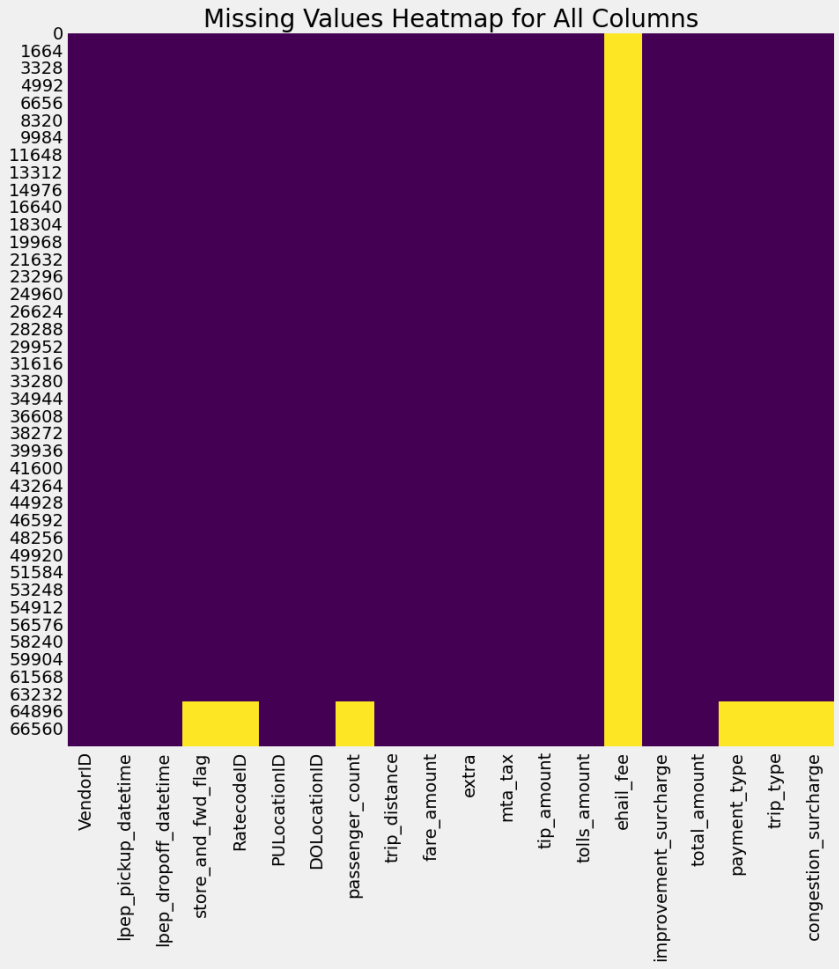
Data Understanding

Dataset Awal : 68211 row data 20 kolom

```
Jumlah baris dan kolom di dataset df adalah (68211, 20)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68211 entries, 0 to 68210
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   VendorID               68211 non-null  int64
 1   lpep_pickup_datetime   68211 non-null  object
 2   lpep_dropoff_datetime  68211 non-null  object
 3   store_and_fwd_flag     63887 non-null  object
 4   RatecodeID             63887 non-null  float64
 5   PULocationID           68211 non-null  int64
 6   DOLocationID           68211 non-null  int64
 7   passenger_count        63887 non-null  float64
 8   trip_distance          68211 non-null  float64
 9   fare_amount            68211 non-null  float64
 10  extra                  68211 non-null  float64
 11  mta_tax                68211 non-null  float64
 12  tip_amount             68211 non-null  float64
 13  tolls_amount           68211 non-null  float64
 14  ehail_fee              0 non-null      float64
 15  improvement_surcharge  68211 non-null  float64
 16  total_amount           68211 non-null  float64
 17  payment_type           63887 non-null  float64
 18  trip_type              63877 non-null  float64
 19  congestion_surcharge   63887 non-null  float64
dtypes: float64(14), int64(3), object(3)
memory usage: 10.4+ MB
```

1. **Data Understanding:** Mengenal dataset untuk mengidentifikasi anomali yang perlu ditangani.
2. **Anomali Identifikasi**: Menentukan jenis anomali yang ada dalam dataset untuk perbaikan.
3. **Justifikasi Penanganan**: Menyertakan alasan penanganan anomali berdasarkan pengetahuan domain dan statistik.
4. Data Kosong: Terdapat beberapa kolom dengan data NaN:
   - store_and_fwd_flag, RatecodeID, passenger_count, payment_type, dan congestion_surcharge **(6.339%).**
   - ehail_fee **(100% kosong).**
   - trip_type **(6.354% kosong).**
5. Format Tanggal: Kolom lpep_pickup_datetime dan lpep_dropoff_datetime perlu dikonversi ke tipe datetime untuk analisis yang lebih mudah.
6. Pemeriksaan Kolom:
   - passenger_count: Cek min dan max, perhatikan kemungkinan nilai 0.
   - ehail_fee: Isi data kosong berdasarkan referensi internet.
   - trip_distance: Konversi dari miles ke kilometer untuk analisis.

Data Understanding & Data Cleaning



Metode Penanganan Missing Value:

1. **Hapus Baris/Kolom**: Tidak disarankan karena tingginya jumlah missing value (contoh: ehail_fee 100% kosong).
2. **Isi Data yang Hilang**: Lebih disarankan untuk mengisi missing value dengan nilai yang mendekati nilai asli, menggunakan kolom lain yang relevan.
3. **Strategi Pengisian**:
   - Utamakan pengisian berdasarkan hubungan domain knowledge atau statistik.
   - Jika tidak memungkinkan, gunakan mean, median, atau modus.
   - **Menghapus data adalah opsi terakhir!!.**

	Column Name	Data Type	Missing Values	Missing Percentage
0	VendorID	int64	0	0.000
1	lpep_pickup_datetime	object	0	0.000
2	lpep_dropoff_datetime	object	0	0.000
3	store_and_fwd_flag	object	4324	6.340
4	RatecodeID	float64	4324	6.340
5	PULocationID	int64	0	0.000
6	DOLocationID	int64	0	0.000
7	passenger_count	float64	4324	6.340
8	trip_distance	float64	0	0.000
9	fare_amount	float64	0	0.000
10	extra	float64	0	0.000
11	mta_tax	float64	0	0.000
12	tip_amount	float64	0	0.000
13	tolls_amount	float64	0	0.000
14	ehail_fee	float64	68211	100.000
15	improvement_surcharge	float64	0	0.000
16	total_amount	float64	0	0.000
17	payment_type	float64	4324	6.340
18	trip_type	float64	4334	6.350
19	congestion_surcharge	float64	4324	6.340

Data Wrangling (Feature Engineering)

```
<class 'pandas.core.frame.DataFrame'>
Index: 66621 entries, 2 to 2
Data columns (total 41 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   lpep_pickup_datetime    66621 non-null  datetime64[ns]
 1   lpep_dropoff_datetime   66621 non-null  datetime64[ns]
 2   store_and_fwd_flag      66621 non-null  object
 3   RatecodeID              66621 non-null  int64
 4   PULocationID            66621 non-null  int64
 5   DOLocationID            66621 non-null  int64
 6   passenger_count         66621 non-null  int64
 7   trip_distance           66621 non-null  float64
 8   fare_amount             66621 non-null  float64
 9   extra                   66621 non-null  float64
 10  mta_tax                 66621 non-null  float64
 11  tip_amount              66621 non-null  float64
 12  tolls_amount            66621 non-null  float64
 13  ehail_fee               66621 non-null  float64
 14  improvement_surcharge   66621 non-null  float64
 15  total_amount            66621 non-null  float64
 16  payment_type            66621 non-null  float64
 17  trip_type               66621 non-null  float64
 18  congestion_surcharge    66621 non-null  float64
 19  pickup_date             66621 non-null  datetime64[ns]
 20  pickup_time             66621 non-null  object
 21  dropoff_date            66621 non-null  datetime64[ns]
 22  dropoff_time            66621 non-null  object
 23  Pickup_Hour             66621 non-null  int64
 24  Pickup_Day              66621 non-null  object
```

Before — **Dataset Awal : 68211 row data 20 kolom**

After — **Dataset clean : 66621 row data 41 kolom**

```
 25  Pickup_Month            66621 non-null  int64
 26  Pickup_Year             66621 non-null  int64
 27  pola_2shift             66621 non-null  object
 28  pola_3shift             66621 non-null  object
 29  category_time           66621 non-null  object
 30  trip_distance(Km)       66621 non-null  float64
 31  trip_duration(Hrs)      66621 non-null  float64
 32  trip_duration(Min)      66621 non-null  float64
 33  speed(Kph)              66621 non-null  float64
 34  vendor_name             66621 non-null  object
 35  pendapatan_per_trip     66621 non-null  float64
 36  fuel_cost_per_trip      66621 non-null  float64
 37  maintenance_cost_per_trip  66621 non-null  float64
 38  fixed_costs_per_trip    66621 non-null  float64
 39  total_costs_per_trip    66621 non-null  float64
 40  laba_bersih_per_trip    66621 non-null  float64
dtypes: datetime64[ns](4), float64(22), int64(7), object(8)
memory usage: 21.3+ MB
```

Referensi Eksternal untuk category_time terkait pola **normal / rush hour** di kota New York

Hari	Jam Sibuk Pagi	Jam Normal	Jam Sibuk Sore	Jam Sibuk Akhir Pekan	Jam Normal Akhir Pekan	Jam Malam
Senin	7:00 AM - 10:00 AM	10:00 AM - 4:00 PM	4:00 PM - 7:00 PM	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Selasa	7:00 AM - 10:00 AM	10:00 AM - 4:00 PM	4:00 PM - 7:00 PM	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Rabu	7:00 AM - 10:00 AM	10:00 AM - 4:00 PM	4:00 PM - 7:00 PM	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Kamis	7:00 AM - 10:00 AM	10:00 AM - 4:00 PM	4:00 PM - 7:00 PM	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Jumat	7:00 AM - 10:00 AM	10:00 AM - 4:00 PM	4:00 PM - 7:00 PM	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Sabtu	-	-	-	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM
Minggu	-	-	-	11:00 AM - 3:00 PM	3:00 PM - 10:00 PM	10:00 PM - 3:00 AM

Referensi Eksternal pola shift yang memungkinkan pada driver NYC Taxi di kota New York sebagai pendeketan analisis.

Pola Shift	Shift Pagi	Shift Sore	Shift Malam
2 Shift	6:00 AM - 6:00 PM	6:00 PM - 6:00 AM	-
3 Shift	5:00 AM - 1:00 PM	1:00 PM - 9:00 PM	9:00 PM - 5:00 AM

Data Wrangling (Feature Engineering)

Pendapatan / Revenue

Pendapatan = fare_amount + extra + mta_tax + tip_amount + tolls_amount + ehail_fee + improvement_surcharge + congestion_surcharge

Penjelasan komponen pendapatan:

- **fare_amount:** Tarif dasar perjalanan.
- extra: Biaya tambahan (misal untuk waktu malam atau cuaca buruk).
- **mta_tax:** Pajak Metropolitan Transportation Authority.
- **tip_amount:** Tip yang diberikan penumpang.
- **tolls_amount:** Biaya tol yang dilalui selama perjalanan.
- **ehail_fee:** Biaya e-hailing, jika ada.
- **improvement_surcharge:** Biaya perbaikan fasilitas transportasi.
- **congestion_surcharge:** Biaya tambahan saat terjadi kemacetan.

Total ini mewakili semua biaya yang dikenakan kepada pelanggan yang kemudian menjadi pendapatan.

Biaya Operasional (Cost)

Cost = Fuel Cost + Maintenance Cost + Fixed Costs

Penjelasan komponen cost:

- Fuel Cost = trip_distance / avg_fuel_consumption / fuel_price_per_liter
- Maintenance Cost = trip_distance /maintenance_cost_per_km
- Fixed Cost

Laba Bersih

Laba bersih = Pendapatan – (Fuel Cost + Maintenance Cost + Fixed Costs)

$$\text{Distance (Km)} = \text{trip\_distance (Miles)} \times 1.60934$$

$$\text{Speed (Kph)} = \frac{\text{Distance (Km)}}{\text{trip\_duration(Hrs)}}$$

Asumsi perhitungan :

- **avg_fuel_consumption = 10 # km per liter**
- **fuel_price_per_liter = 1.00 # USD per liter**
- **maintenance_cost_per_km = 0.15 # USD per km**
- **fixed_costs_per_trip = 0.5 # USD per trip**
- **ehail_fee : 1.25 #USD**

- Dalam sistem taksi New York City (NYC), **biaya e-hail ditetapkan sebesar $1.25 per perjalanan untuk taksi kuning (Yellow Taxi) dan hijau (Green Taxi)** yang dipesan melalui aplikasi e-hail yang disetujui.
- Biaya ini bersifat tetap dan berlaku sama, baik untuk taksi yang menggunakan sistem pembayaran dari Creative Mobile Technologies (CMT) maupun VeriFone Inc., tanpa perbedaan berdasarkan vendor.

Outlier Detection & Treatment

Berikut beberapa anomali-anomali pada dataset yang dilakukan penanganan :

1. Anomali row data **passenger_count >0, trip_duration =0, trip_distance(Kph)=0** ada 774 row data = 1.135%. Sehingga anomali ini di-*drop* dari data frame;
2. Outliers treatment untuk **passenger_count dengan jumlah 0,7, dan 8;**
3. Outlier **trip_distance(Km) >100KM;**
4. Melakukan treatment terhadap **outlier TRIP DURATION dengan durasi lebih dari 24 jam atau 0 jam;**
5. Melakukan treatment terhadap outlier **fare_amount NEGATIF;**
6. Melakukan *treatment* terhadap outlier **Speed dengan nilai Infinite;**
7. Melakukan drop data anomali **pickup_year 2009;**
8. Outlier handling untuk menangani **mta_tax, improvement_surcharge, congestion_surcharge, laba_bersih_per_trip tidak boleh minus;**
9. Menangani Anomali pada **RateCodeID karena tidak sesuai dengan standar sistem nilai 99.**

Step By Step EDA (Exploratory Data Analysis)

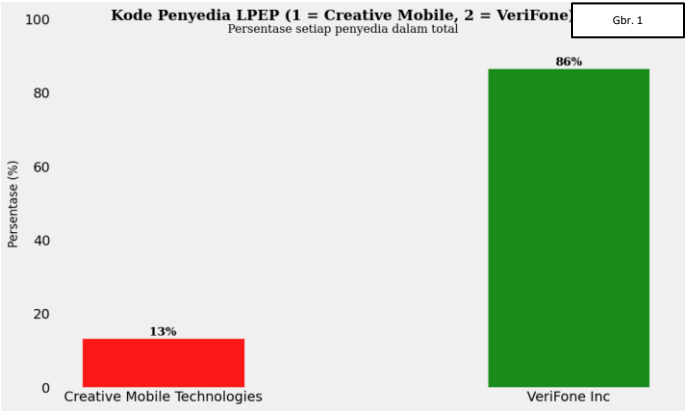
EDA	dataframe	action	null	%null	keterangan:	Row Data
Data Understanding	df	20 kolom			68211 row data	
Data Cleaning	df1	Store_and_fwd_flag	4324	6.339%	Isi data kosong dengan nilai **modus**	68211
Data Cleaning	df2	RatecodeID	4324	6.339%	Isi data kosong dengan nilai **modus**	68211
Data Cleaning	df3	passenger_count	4324	6.339%	Isi data kosong dengan nilai **median** (data tidak berdistribusi normal, disi dengan 1)	68211
Data Cleaning	df4	ehail_fee4	68211	100%	Diisi dengan nilai $1.25/trip (ref. Internet)	68211
Data Cleaning	df5	payment_type	4324	6.339%	Isi data kosong dengan nilai **modus**	68211
Data Cleaning	df6	trip_type	4334	6.354%	Isi data kosong dengan **median**	68211
Data Cleaning	df7	congestion_surcharge	4324	6.339%	Isi data kosong dengan **median**	68211
Data Wrangling	df7	ganti tipe data			**lpep_pickup_datetime** dan **lpep_dropoff_datetime** ke tipe datetime	68211
Data Wrangling	df8	8 kolom baru			Menambahkan kolom baru **pickup_date, pickup_time,dropoff_date, dropoff_time, Pickup_Hour dan Pickup_Day,Pickup Month,Pickup Year**	68211
Data Wrangling	df9	2 kolom baru			Menambahkan **pola_2shift** dan **pola_3shift**	68211
Data Wrangling	df10	1 kolom baru			Menambahkan **category_time**	68211
Data Wrangling	df11	1 kolom baru			Konversi trip_distance dalam Miles --> **trip_distance(KM)**	68211
Data Wrangling	df12	1 kolom baru			Menambahkan kolom **trip_duration(hrs)**	68211
Data Wrangling	df13	2 kolom baru			Menambahkan kolom **speed(Kph)** dan **vendor_name**	68211
Data Wrangling	df14	6 kolom baru			Menambahkan kolom **pendapatan_per_trip, fuel_cost_per_trip,maintenance_cost_per_trip, fixed_costs_per_trip, total_costs_per_trip, laba_bersih_per_trip**	68211
Data Cleaning	df15	cek outliers Passenger Count >0, trip_duration =0, trip_distance(Kph)=0	774	1.135%	**Drop Passenger Count >0, trip_duration =0, trip_distance(Kph)=0** sebanyak 774 row data sehingga menjadi ==> 67437. Serta update perubahan tipe data **lpep_pickup_datetime, lpep_dropoff_datetime , pickup_date** dan **dropoff_date** menjadi tipe data datetime	67437
Data Cleaning	df16	cek outliers passenger count =0,7,8	339	0.503%	**Drop passenger_count=0, 7 dan 8**, row data menjadi 67098	67098
Data Cleaning	df17	trip_distance max: 120098.840	38	0.057%	trip_distance(Km)>100 Km	67060
Data Cleaning	df18	trip_duration(Hrs)	277	0.413%	trip_duration(Hrs) Durasi lebih dari 24 jam atau 0 jam	66783
Data Cleaning	df19	fare_amount	139	0.208%	Melakukan treatment terhadap outlier fare_amount NEGATIF	66644
Data Cleaning	df20	Pickup_Year	1	0.002%	Ada pencilan data 2009	66643
Data Cleaning	df21	mta_tax, improvement_surcharge, congestion_surcharge laba_bersih_per_trip	12	0.018%	mta_tax, improvement_surcharge, congestion_surcharge, laba_bersih_per_trip tidak boleh minus	66631
Data Cleaning	df22	RatecodeID	10	0.015%	anomali RateCodeID 99	66621


ANALYSIS 1
ANALYSIS 2
ANALYSIS 3
20%
8%
6%
9%
23%
90%
75%

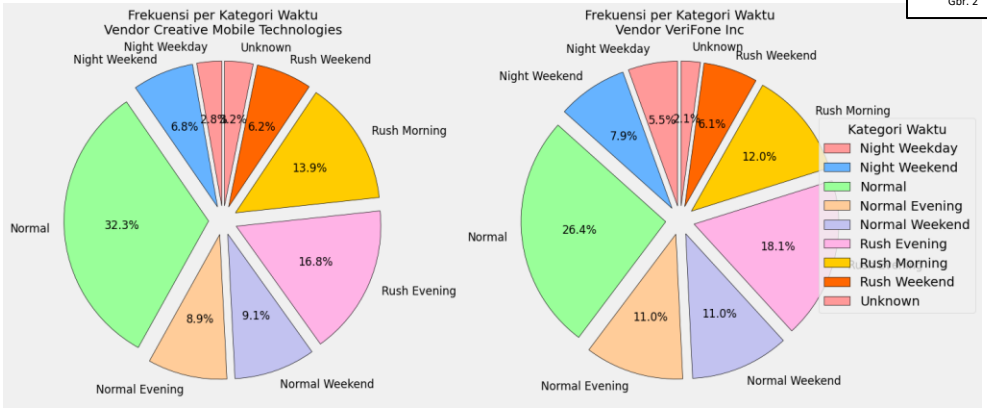
1 Data Analisis (Evaluasi Efektifitas per Vendor)

Sumber Data 9 Des 2022 dan data 1 Jan -1 Feb 2023 : 66621 row data sample

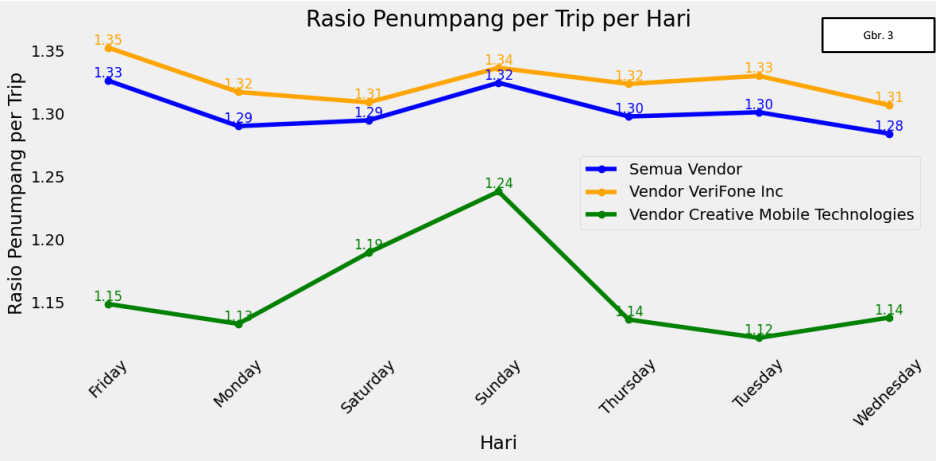
Gbr. 1



Gbr. 2



Gbr. 3

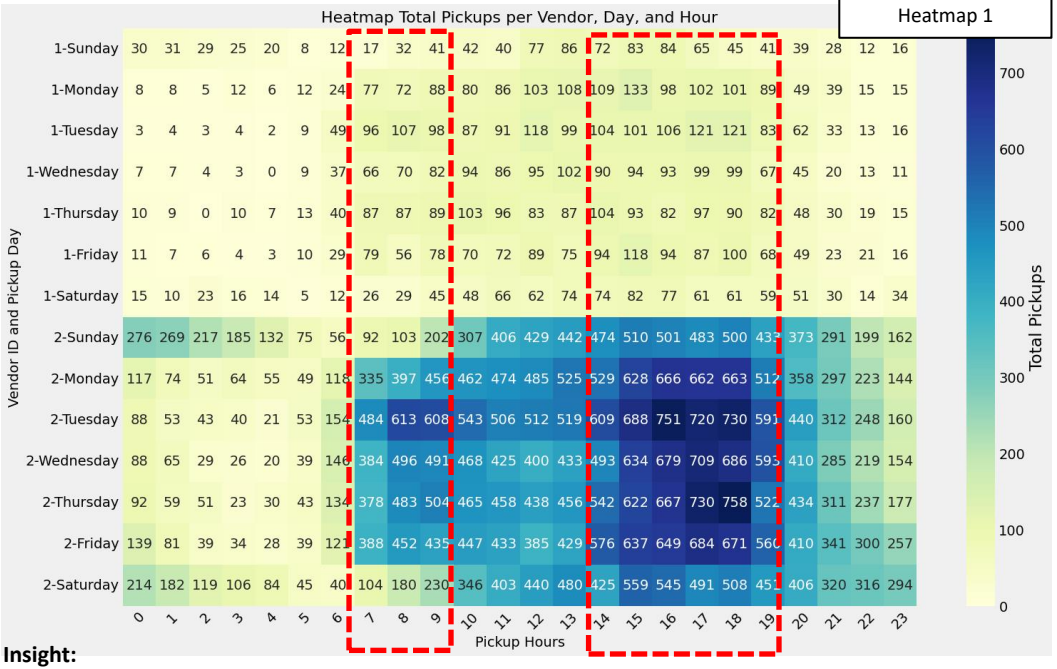


Insight:

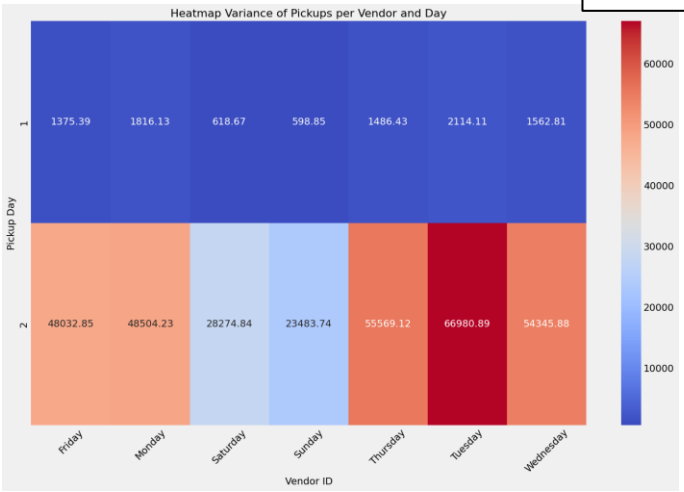
- Analisis terkait kedua vendor (Gbr 1&2) antara : 1 = Creative Mobile Technologies, LLC, 2 = VeriFone Inc. 86 % total sumbangsih pendapatan NYC didapatkan dari **kontribusi positif 86% Verifone Inc. Proporsi *rush hours* baik morning dan evening** setara dengan jam normal**.**
- Jika tanpa melihat rasio penumpang per-trip hanya dari hari dengan trip terbanyak ada di hari Selasa dengan koef. Variansi (KV) yang masih cukup lebar di 10.69%.
- Namun jika ditelisik lagi trend **ratio penumpang per trip** (Gbr 3), maka perlu dicek lebih lanjut kenapa di hari Jumat rationya bisa lebih besar dibanding *weekend*. Kondisi ini membuktikan bahwa fenomena peningkatan ratio penumpang tidak hanya terpusat di *weekend,* untuk vendor Verifone "peak ratio" 1.35 ada di Jumat. Sedangkan pola vendor Creative Mobile Tech. Cenderung *so so* saja. **Jumlah penumpang per trip merupakan indikator efisiensi operasional taksi.**
  a) Rasio yang lebih tinggi menunjukkan bahwa kendaraan mengangkut lebih banyak penumpang dalam satu perjalanan, yang dapat meningkatkan pendapatan dan mengurangi biaya per penumpang.
  b) Selain itu juga **positif bagi aspek Safety terkait mengurangi jumlah kendaraan di jalan, yang berdampak positif terhadap lingkungan dengan mengurangi kemacetan dan emisi karbon (Lebih ke isu ESG/ Energi).**

  Dengan mempertimbangkan rasio ini dapat memberikan keuntungan kompetitif dan mendorong keberlanjutan dalam operasi.

Data Analisis...



Heatmap 1



Heatmap 2

Insight:

- Analisis terkait kedua vendor antara : 1 = Creative Mobile Technologies, LLC, 2 = VeriFone Inc. 86 % total sumbangsih pendapatan NYC didapatkan dari kontribusi positif 86% Verifone Inc.
- **Jika dilihat pada heatmap (1), total pickup berdasarkan jam, vendor, dan hari dapat diidentifikasi jam puncak pagi di 7-9 (jam sibuk pagi), dan 14-19 dengan "jendela" yang lebih lebar**. Logikanya ada pola sore menjelang malam di setiap hari pada vendor 2. Verifone mengalami kenaikan pengguna layanan.
- Heatmap (2) menggambarkan variansi dalam total pickups. Wilayah dengan warna lebih terang menunjukkan variansi yang tinggi, yang berarti ada fluktuasi signifikan dalam jumlah pickups selama jam dan hari tertentu. **Kombinasi vendor, hari, dan jam dengan variansi tinggi bisa menandakan bahwa permintaan tidak konsisten, dan ada faktor-faktor tertentu yang memengaruhi permintaan**.

Rekomendasi :

1. **Promosi dan Diskon:** Tawarkan diskon pada jam sepi (10:00 - 16:00, hari kerja) untuk menarik pelanggan.
2. **Penyesuaian Armada**: Sesuaikan jumlah kendaraan sesuai pola permintaan untuk efisiensi biaya operasional.
3. **Kampanye Pemasaran:** Luncurkan kampanye untuk meningkatkan kesadaran layanan selama jam sepi.Fokus pada acara lokal dan tawarkan layanan khusus di waktu tersebut.
4. **Program Loyalitas:** Kembangkan program loyalitas untuk pelanggan yang menggunakan layanan pada jam sepi.Analisis Data:Gunakan data untuk mengidentifikasi tren permintaan dan mengoptimalkan jadwal operasional.
5. **Kemitraan Strategis**: Jalin kemitraan dengan bisnis lokal untuk paket promosi yang saling menguntungkan

2 Data Analisis (Evaluasi Efektifitas per Vendor)

avg. Cost per Trip	Creative Mobile Technologies	VeriFone Inc	% gap dari vendor 1
1 = Standard rate	1.4	1.6	14.9%
2 =JFK	7.0	6.1	-12.2%
3 =Newark	1.8	6.7	275.2%
4 =Nassau or Westchester	3.9	7.7	96.3%
5 =Negotiated fare	1.9	2.2	17.5%

Avg. Pendapatan per Trip	Creative Mobile Technologies	VeriFone Inc	% gap dari vendor 1
1 = Standard rate	24.1	22.7	-5.7%
2 =JFK	89.3	88.4	-1.0%
3 =Newark	33.9	101.2	198.3%
4 =Nassau or Westchester	49.8	112.7	126.4%
5 =Negotiated fare	27.0	36.9	36.8%

Note : dalam USD

Insight:

Pendapatan per-trip Verifone untuk 98% mayoritas di standard rate cenderung lebih murah dibanding vendor Mobil Tech. Namun cost per tripnya lebih boros dibanding vendor 1

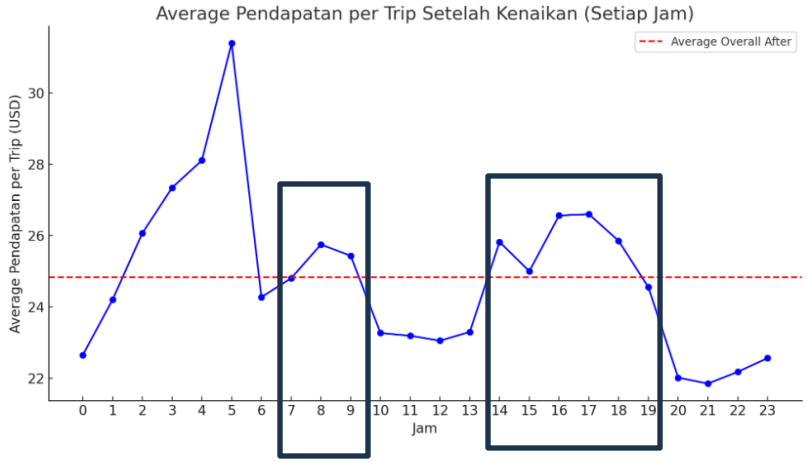
- **Penerapan Pricing Strategy pada rush hours 7,8,9 (morning) dan 14,15,16,17,18,19 (rush evening) khususnya untuk Verifone Inc. Uji coba harga start di 5% sambil melihat elastisitas pasar.**
- **Buat regulasi dan kebijakannya serta perubahan harga harus dipastikan *align* dengan sistem.**

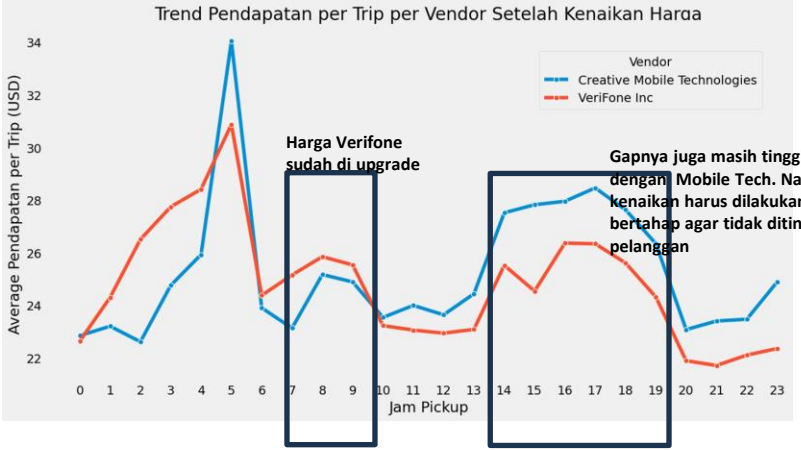
- **Before : Pendapatan per-trip di 23.34 USD/ trip**
- **After: dinaikkan ke 10% menjadi 25.67 USD/trip**

Kenaikan ini juga ada Pros & Cons yang wajib diperhatikan:

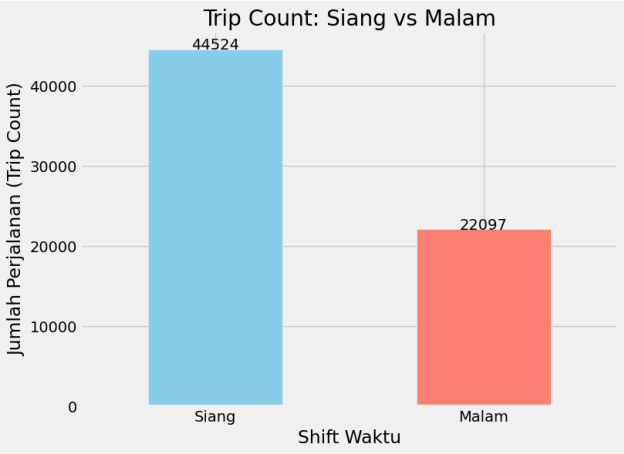
Pros: Increase revenue, manajemen permintaan, optimalisasi armada, & stabilitas lalu lintas;

Cons: kehilangan pelanggan, ketidakpuasaan pelanggan, persepsi negatif (citra buruk karena terkesan "memanfaatkan" situasi yang menguntungkan terutama saat jam-jam sibuk.





2 Data Analisis (Waktu Pickup Paling Ramai)



Gap persentase antara jumlah trip Malam dan Siang adalah **-50.37%**

Hipotesis

1.Hipotesis nol (H0): proporsi waktu malam = 0.5

2.Hipotesis alternatif (Ha): proporsi waktu malam < 0.5

Metode : Langkah Uji Proporsi Satu Sampel

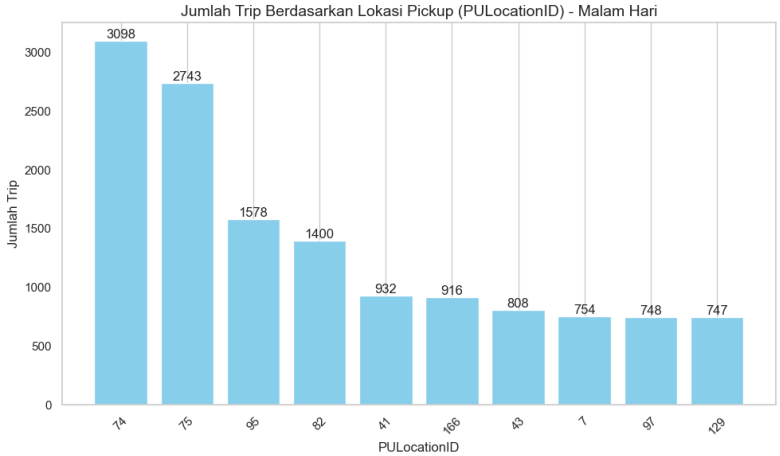
```
z-statistik: -92.27 p-value: 0.0000 Tolak H0:
Ada bukti bahwa proporsi waktu malam lebih
kecil dari 0.5.
```

Interpretasi Hasil

•Jika p-value < 0.05, kita menolak H0, yang berarti kita memiliki cukup bukti untuk menyimpulkan bahwa proporsi waktu malam **lebih kecil dari 0.5**.

•Jika p-value ≥ 0.05, kita gagal menolak H0, yang berarti tidak ada cukup bukti untuk menyimpulkan bahwa proporsi waktu malam lebih kecil dari 0.5, dan H0 diterima.

Setelah divalidasi perjalanan malam hanya 33.17%,



Insight : Fokus identifikasi lokasi-lokasi dengan permintaan tinggi pada malam hari (to ten loacation). **Hal ini dapat membantu dalam mengarahkan pengemudi untuk lebih fokus pada area-area tersebut selama shift malam & termasuk penyesuaian harga untuk memaksimalkan revenue**.

Dengan proporsi perjalanan malam yang hanya mencapai 33.17%, namun memberikan pendapatan lebih tinggi, fokus harus diarahkan pada peningkatan kualitas dan daya tarik shift malam, baik dari sisi pengemudi maupun penumpang.

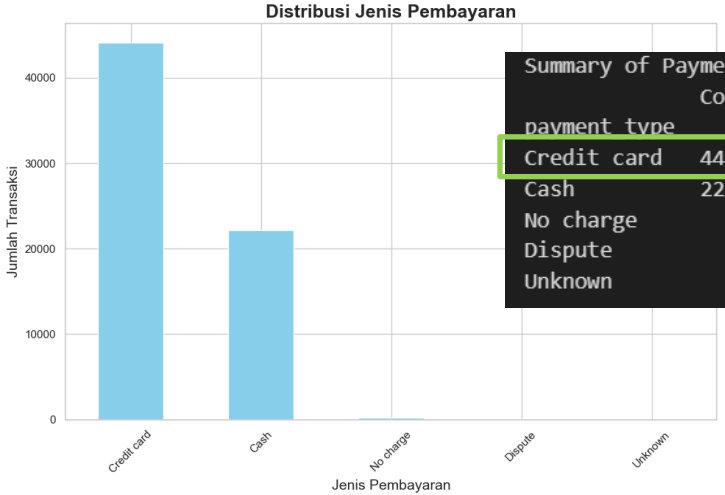
Insentif untuk pengemudi shift malam, dengan dua pendekatan :

- Bonus atau Insentif Finansial
- Fleksibilitas Shift Schedulling

Kampanye untuk menarik penumpang malam hari

- **Promosi untuk penumpang**
- **Target pengalokasian di dekat area yang aktifitas malamnya tinggi. Mis: tempat hiburan malam, restoran, tempat wisata , dsb.**

3 Data Analisis (Payment Type Terfavorit)



Summary of Payment Types:

payment_type	Count	Percentage
Credit card	44133	66.245
Cash	22190	33.308
No charge	247	0.371
Dispute	50	0.075
Unknown	1	0.002

Jenis Pembayaran Terbanyak per Vendor:

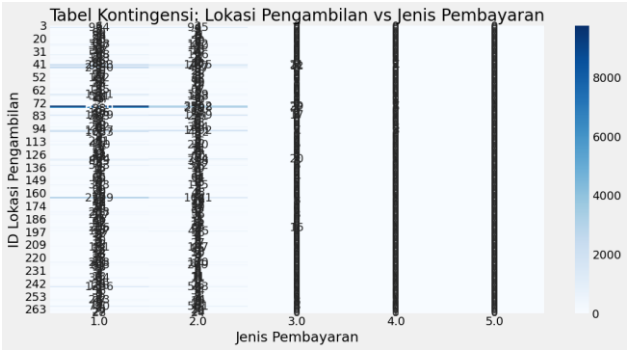
vendor_name	
Creative Mobile Technologies	Credit card
VeriFone Inc	Credit card

Name: payment_type, dtype: object

Pivot Table (RatecodeID vs Payment Type):

RatecodeID \ payment_type	Cash	Credit card	Dispute	No charge	Unknown
JFK	36	98	0	3	0
Nassau or Westchester	35	21	1	0	0
Negotiated fare	313	601	1	12	0
Newark	8	15	0	0	0
Standard rate	21798	43398	48	232	1

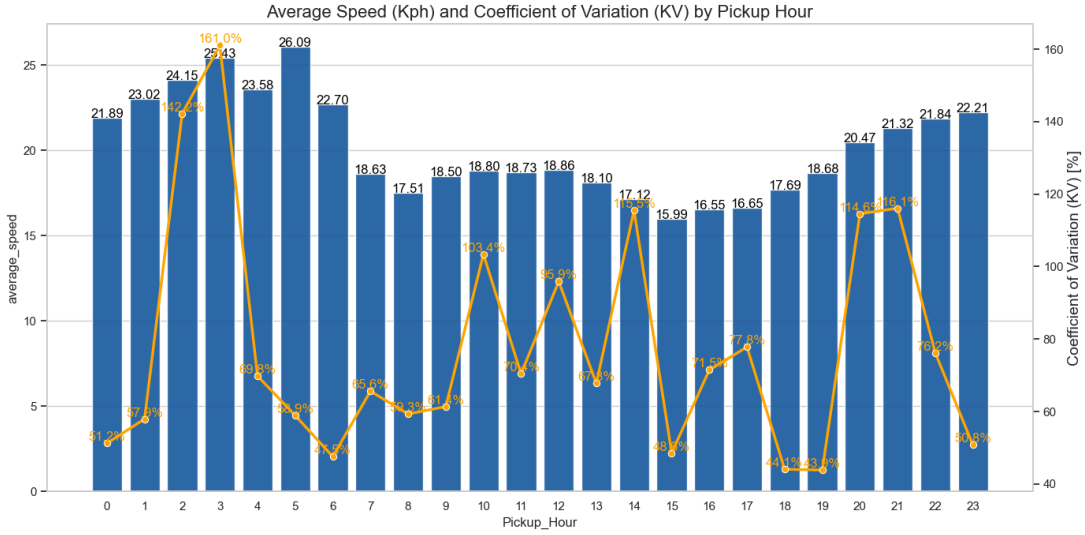
Dilakukan uji Chi-Squared: 13123.723348747879, **P-value: 0.0 Terdapat hubungan signifikan antara lokasi pengambilan dan jenis pembayaran.**

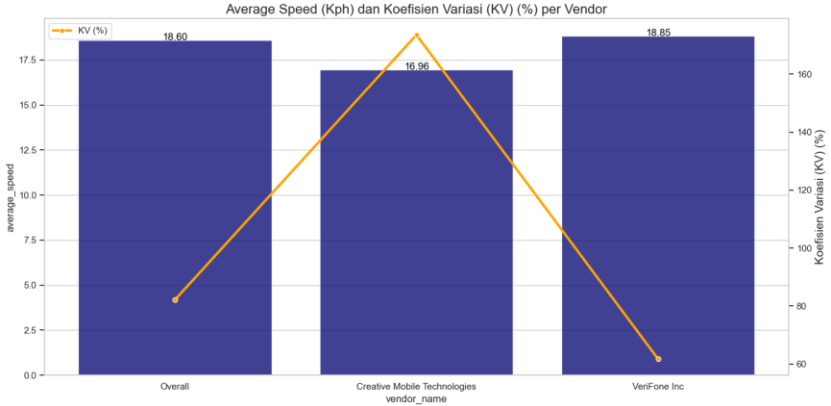


Insight :

- **Mayoritas 98% untuk 5 tipe perjalanan RateCodeID (standard rate, JFK, Nassau or Westchester, Negotiated Fare, Group Ride) menggunakan metode Credit Card dan Cash. Kedua vendor juga mayoritas menggunakan Credit Card.**
- **Implikasi bagi NYCTaxi Trip adalah dengan menerapkan beberapa strategi untuk tetap dapat menjaga kepuasaan pelanggan , yakni:**
  1. **Adopsi Pembayaran Disgital**
     - **Penggunaan Pembayaran Elektronik (Perluas penerimaan pembayaran digital, termasuk opsi dompet digital dan aplikasi pembayaran);**
     - **Keamanan Transaksi**
  2. **Analisis Perilaku Pelanggan**
     - **Preferensi Pelanggan : Pelanggan NYC menunjukkan seacara data 66.2% lebih suka metode non tunai, yang menjadi fokus utama dalam pemasaran;**
     - **Kenyamanan Pelanggan.**
  3. **Dampak pada Penjadwalan dan Manajemen Armada**
     - **Optimasi Penjadwalan ke lokasi tujuan yang permintaannya tinggi untuk pembayaran non-tunai.**
  4. **Inovasi untuk integrasi dengan Aplikasi pemesanan.**

4 Data Analisis (Disparitas Speed -Kph)





Insight :

Koefisien Variansi (KV) adalah ukuran statistik yang digunakan untuk menunjukkan seberapa besar variabilitas relatif suatu data dibandingkan dengan rata-ratanya. Koefisien ini dinyatakan dalam persentase dan dihitung dengan rumus:

$$\text{KV} = \left(\frac{\text{Standar Deviasi}}{\text{Rata-rata}}\right) \times 100\%$$

- **Disparitas dalam kecepatan (speed) selama perjalanan taksi di New York City (NYC) merupakan aspek yang sangat penting dan memiliki beberapa implikasi bagi operasi taksi dan layanan transportasi secara umum. Berikut adalah beberapa alasan mengapa disparitas ini penting:**
  1. **Pengaruh Biaya: Biaya Operasional dan Harga / Tarif;**
  2. **Efisiensi Oeparasional : Pengelolaan waktu dan Perencanaan rute.**
- **Dari data menunjukkan bahwa avg speed all vendor di 18.6Kph. Untuk vendor 1 Mobile Technologis rerata speednya di 16.96 Kph lebih rendah ~10% dibandingkan Verifone dengan rerata speed 18.85 Kph.**
- **Idealnya adalah nilai Avg Speed yang tinggi dengan KV <=10%, kalo melihat grafik di samping ada keanehan pada jam 3 dini hari karena walaupun speednya tinggi di 25.43Kph, namun disparitas/ variancenya tinggi. Hal ini perlu dilakukan observasi lebih lanjut apakah ada error di sistem atau memang perlu dilakukan perbaikan terkait perilaku,**
- **Disparitas Vendor 1 sangat buruk, hal ini dapat menjadi tinjauan lebih lanjut untuk melakukan improvement terkait "perlambatan" agar perusahaan dapat mengukur *treshold cycle 1 trip per km* durasi yang masih dikatakan normal/ tidak sebagai bentuk pengendalian behavior operator (perilaku pengemudi) dan untuk menjaga efisiensi fuel_cost.**

Thank You

If you torture data long enough, it will confess to anything!

ELISA | CAPSTONE PROJECT MODULE 2