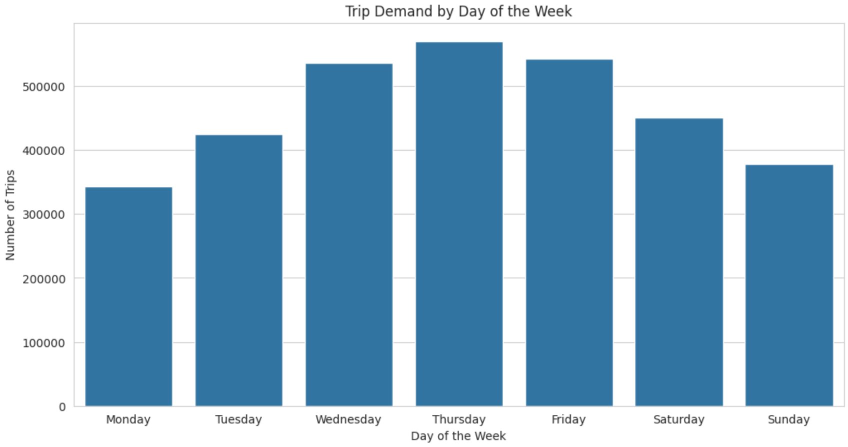
Data Analysis Report: Insights from NYC Yellow Taxi Data 2025

Objective: To analyze the January 2025 NYC Yellow Taxi trip dataset to uncover key operational and behavioral patterns, and to provide actionable recommendations for a transportation-focused business.

Methodology: The raw Parquet data was cleaned to handle null values and remove illogical outliers (e.g., trips with zero distance or unrealistic durations). New features such as trip_duration, pickup_hour, and tip_percentage were engineered to facilitate deeper analysis. The following five key insights were derived from the cleaned dataset.
(Check the data_cleaning&Analysis_scripts.ipynb file for this).

Insight 1: Demand Peaks Mid-Week, Not on Weekends

- **Insight:** Trip demand peaks on Thursday, not on Friday or Saturday, indicating that the business is heavily reliant on a professional and commuter-based travel cycle rather than a purely leisure-based one.
- **Supporting Visualization:**



- **Script**

```
df['pickup_day_of_week'] = df['tpep_pickup_datetime'].dt.day_name()

day_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday',
'Sunday']
sns.countplot(data=df, x='pickup_day_of_week', order=day_order)
```

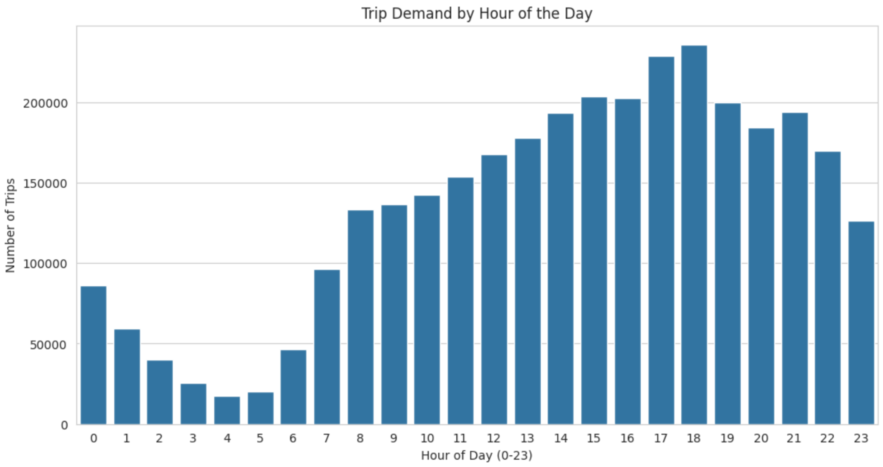
Business Implication: This pattern allows for precise resource allocation. Driver incentives and vehicle availability should be maximized from Wednesday to Friday to meet the core demand, rather than focusing on a traditional weekend model.

Confidence & Caveats: High confidence. This insight is based on the entire dataset.

Insight 2: The Evening Rush (4 PM - 7 PM) is the Primary Revenue Window

- **Insight:** Trip volume shows a pronounced and sustained peak in the evening between 4 PM and 7 PM, generating significantly more activity than the morning commute.

- **Supporting Visualization:**

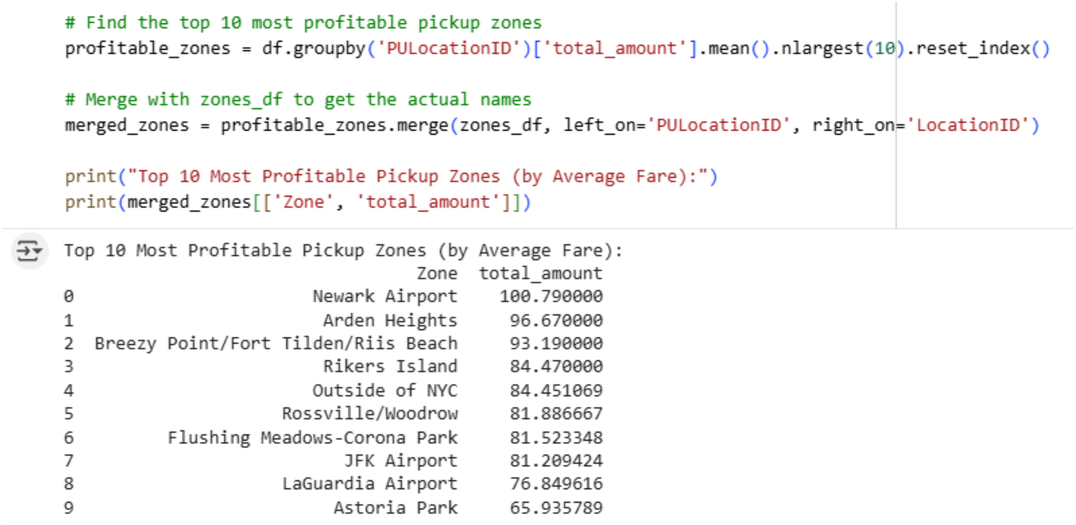


- **Business Implication:** This "duration window" is the most critical period for daily revenue. It is the ideal time to implement dynamic pricing strategies to capitalize on high demand and to offer driver bonuses to ensure maximum fleet deployment.

- **Confidence & Caveats: High confidence. The hourly trend is clear and consistent**

Insight 3: Profitability is Driven by Distance (Airports) while Volume is Driven by Density (Manhattan)

- **Insight:** The most profitable trips (by average fare) originate from airports and distant outer-boroughs, whereas the highest volume of trips originates from dense Manhattan neighborhoods. This reveals two distinct markets: high-value and high-volume.



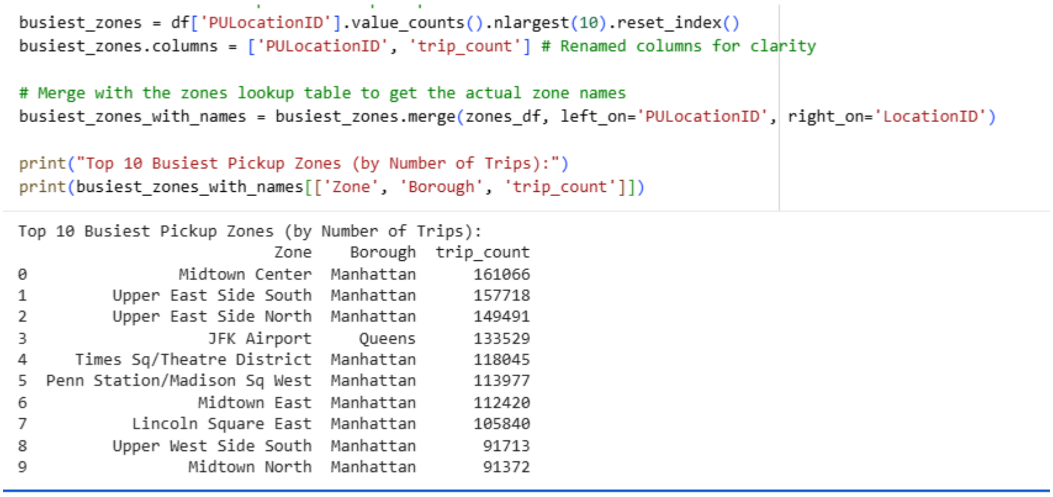
```
# Find the top 10 most profitable pickup zones
profitable_zones = df.groupby('PULocationID')['total_amount'].mean().nlargest(10).reset_index()

# Merge with zones_df to get the actual names
merged_zones = profitable_zones.merge(zones_df, left_on='PULocationID', right_on='LocationID')

print("Top 10 Most Profitable Pickup Zones (by Average Fare):")
print(merged_zones[['Zone', 'total_amount']])
```

```
Top 10 Most Profitable Pickup Zones (by Average Fare):
                              Zone  total_amount
0                   Newark Airport    100.790000
1                    Arden Heights     96.670000
2  Breezy Point/Fort Tilden/Riis Beach     93.190000
3                    Rikers Island     84.470000
4                   Outside of NYC     84.451069
5                Rossville/Woodrow     81.886667
6       Flushing Meadows-Corona Park     81.523348
7                      JFK Airport     81.209424
8                LaGuardia Airport     76.849616
9                      Astoria Park     65.935789
```

-



```
busiest_zones = df['PULocationID'].value_counts().nlargest(10).reset_index()
busiest_zones.columns = ['PULocationID', 'trip_count'] # Renamed columns for clarity

# Merge with the zones lookup table to get the actual zone names
busiest_zones_with_names = busiest_zones.merge(zones_df, left_on='PULocationID', right_on='LocationID')

print("Top 10 Busiest Pickup Zones (by Number of Trips):")
print(busiest_zones_with_names[['Zone', 'Borough', 'trip_count']])
```

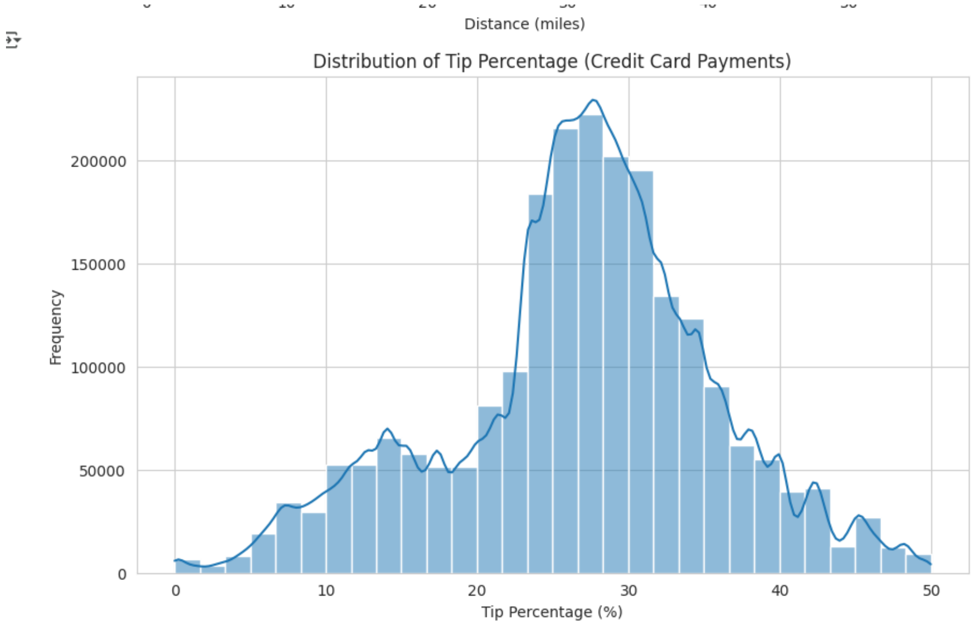
```
Top 10 Busiest Pickup Zones (by Number of Trips):
                         Zone    Borough  trip_count
0              Midtown Center  Manhattan      161066
1       Upper East Side South  Manhattan      157718
2       Upper East Side North  Manhattan      149491
3                 JFK Airport     Queens      133529
4   Times Sq/Theatre District  Manhattan      118045
5  Penn Station/Madison Sq West  Manhattan      113977
6                Midtown East  Manhattan      112420
7          Lincoln Square East  Manhattan      105840
8        Upper West Side South  Manhattan       91713
9               Midtown North  Manhattan       91372
```

- **Business Implication:** The company needs a dual strategy: maintain a strong presence in Manhattan to capture the high volume of daily trips, while also creating incentives for drivers to accept the more lucrative, long-distance airport fares.
- **Confidence & Caveats:** High confidence. The data clearly separates these two market types.

Insight 4: Standardized Tipping Culture is Centered at 20%

- **Insight:** Analysis of credit card transactions reveals a highly standardized tipping behavior, with a strong convergence on a 20% tip.

- **Supporting Visualization:**

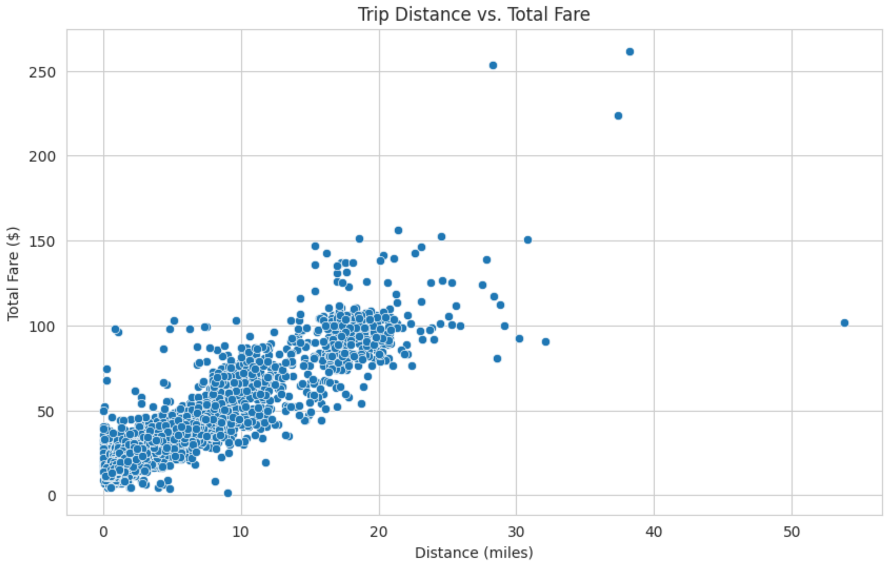


- **Analytical Approach: A tip_percentage feature was calculated for credit card trips. A histogram of this feature showed a clear normal distribution centered sharply at the 20% mark.**

- **Business Implication:** This predictable customer behavior makes gratuity a stable revenue stream. It also proves that the design of in-cab payment prompts directly influences customer decisions and can be optimized to improve driver earnings.
- **Confidence & Caveats:** High confidence, but with a major caveat: this insight is **only valid for credit card users**, as cash tips are not electronically recorded in the dataset.

Insight 5: Fare Depends on Time, Not Just Distance

- **Insight:** The scatter plot of fare vs. distance reveals a segment of high-fare trips with near-zero distance, highlighting the critical role of the time-based fare component in generating revenue from traffic congestion.
- **Supporting Visualization:**



Analytical Approach: Plotting `trip_distance` against `total_amount` visually identified outliers that did not follow the primary trend. These points had low distance but high fares.

- **Business Implication:** This affirms that the current fare structure is robust and well-suited for a congested urban environment. The time-based charge protects driver earnings and company revenue during periods of heavy traffic.

- **Confidence & Caveats:** High confidence. The data points clearly exist. However, the exact reason for the "zero-distance" trip (e.g., stuck in traffic vs. an immediate cancellation with a fee) cannot be determined from this data alone.

By Saharsh, saharshg895@gmail.com