Exploratory Data Analysis (EDA)

Dataset: NYC Taxi Trip Duration
Target Variable: trip_duration (in seconds)

---

1. 🧾 Dataset Overview

- The dataset contains **trip-level data** including pickup/dropoff times, locations, passenger counts, vendor identifiers, and flags.
- **Datetime fields** (pickup_datetime, dropoff_datetime) were converted to proper formats.
- The target variable, trip_duration, was retained in seconds for analysis and later log-transformed to reduce skewness.

---

2. 🔍 Data Quality and Cleaning

- **No major missing values** in critical features.
- **Invalid records** (e.g., trip_duration <= 0, passenger_count = 0) were removed.
- Time features such as **pickup hour**, **weekday**, and **month** were extracted to enable temporal analysis.

---

3. 📊 Univariate and Bivariate Insights

📈 Trip Duration

- Raw distribution of trip_duration is **heavily right-skewed**, with many short trips and a long tail of high durations.
- After applying a **log transformation** (log1p), the distribution became more symmetric and suitable for modelling.
- Median trip duration is around **650 seconds (~11 minutes)**.

👥 Passenger Count

- The majority of trips (over 70%) had **1 passenger**.
- No significant trend between **passenger count** and trip duration beyond single passengers.

⏱️ Time of Day & Week

- **Rush hours (7–9 AM and 5–7 PM)** exhibit longer median trip durations, aligning with NYC traffic patterns.
- **Weekdays** tend to have slightly higher durations than weekends, especially during commuting times.

---

4. 🧭 Geospatial Patterns

🚕 Pickup & Dropoff Clusters

- Dense clusters of trips are observed in **Manhattan**, especially Midtown and Downtown.
- Some activity around **airports (JFK, LaGuardia)** and along major thoroughfares.
- Visualized using **scatter plots and hexbin maps** of coordinates.

---

5. 🔗 Relationships with Trip Duration

Variable	Insight
Passenger Count	Minimal effect beyond single-rider majority.
Pickup Hour	Strong impact; peak hours increase duration.
Vendor ID	Minor variation in median duration between vendors.
Store-and-Forward Flag	Negligible difference in duration distribution.

---

6. 📌 Key Takeaways

- The dataset is **clean and rich**, with good temporal and spatial granularity.
- Trip durations are **log-normally distributed**, with peak travel times influencing length.
- **Time-based features** (hour, weekday) are the strongest correlates of trip duration.
- **Geospatial hotspots** suggest high traffic in business and transport zones.
- The dataset is well-suited for **modelling trip durations** with proper feature engineering.