🚕 Project Title: Uber Trip Analysis

📌 Objective

This project aims to identify **spatial clusters of Uber pickups** using unsupervised machine learning (KMeans) and analyze **demand patterns by time and zone**. The final goal is to provide recommendations for **optimal vehicle allocation** per cluster.

📁 Dataset Used

- **Source**: uber-raw-data-apr14.csv
- **Records**: Over 200,000 Uber pickup records for April 2014
- **Features**:
    - Date/Time: Timestamp of pickup
    - Lat: Latitude of pickup
    - Lon: Longitude of pickup
    - Base: Uber base code (not used in this analysis)

---

1. 🖌️ Data Preprocessing

- **Datetime Conversion**: The Date/Time column was converted to datetime format to extract time-based features.
- **Extracted Features**:
    - hour — hour of day (0–23)
    - day — day of the month (1–31)
    - weekday — day of the week (0=Monday)
    - month — constant value (April = 4)
- **Filtering**: Latitude and longitude values were filtered to fall within realistic NYC bounds:
    - Latitude: 40 to 41
    - Longitude: –75 to –72

---

2. 📊 Clustering Analysis (KMeans)

- A **random sample of 10,000 pickup coordinates** was taken from the dataset.
- Standardized the data using StandardScaler.
- Applied **KMeans clustering** with **6 clusters**.
- Visualized the clusters using seaborn.scatterplot, colored by cluster ID.

📈 **Visualization:**

A clear spatial separation of clusters shows distinct high-demand pickup zones across NYC.

---

3. ⏰ Hourly Demand Analysis

- Each pickup was assigned to one of the 6 clusters using the trained KMeans model.
- Grouped pickup counts by cluster and hour to assess hourly demand trends.

📉 **Insights:**

- Clusters have distinct peak hours.
- Some zones are busiest in morning hours, while others peak in the evening or late night.

---

4. 🚚 Fleet Allocation Strategy

- Calculated the **average hourly demand** per cluster.
- Interpreted these as a proxy for **minimum recommended vehicle allocation** per cluster to meet average demand.

🧾 **Recommended Fleet Allocation:**

Cluster	Suggested Vehicles
0	182
1	150
2	171
3	165

Cluster	Suggested Vehicles
4	159
5	138

(Note: These values are hypothetical; actual numbers will vary based on full dataset.)

✅ Conclusion

- **KMeans clustering** successfully identified meaningful zones of Uber activity.
- **Hourly analysis** revealed temporal patterns of ride demand across zones.
- **Fleet allocation strategy** based on average demand can help Uber optimize driver deployment and minimize wait times.

🔮 Recommendations & Future Work

- Include **multiple months** of data to capture seasonality.
- Incorporate **external data** (weather, holidays, events) for deeper insights.
- Use **geo-fencing techniques** to refine cluster boundaries.
- Consider dynamic clustering per day or per hour for real-time optimization.