Uber Ride Data Analysis Project

Project Overview

This project focuses on analyzing Uber ride data using fundamental data science techniques. The main goal is to understand ride patterns, peak hours, trip behavior, and usage trends based on time, location, purpose, and category of rides. The project is designed and implemented from a **beginner / fresher perspective**, following a clear and step-by-step analytical approach.

The analysis includes data cleaning, preprocessing, exploratory data analysis (EDA), and multiple visualizations to extract meaningful insights from the dataset.

Dataset Description

The dataset used in this project is an Uber ride log dataset containing real ride information. The key features in the dataset include:

- **START_DATE** – Date and time when the ride started
- **END_DATE** – Date and time when the ride ended
- **START** – Starting location of the ride
- **STOP** – Destination location of the ride
- **MILES** – Distance covered during the ride
- **PURPOSE** – Purpose of the ride (Meeting, Errand, Customer Visit, etc.)
- **CATEGORY** – Ride category (Business or Personal)

The dataset was loaded from an Excel file (`Uber_Ride_Analysis_Dataset.xlsx`).

Libraries Used

The following Python libraries were used in this project:

- **Pandas** – Data manipulation and analysis
- **NumPy** – Numerical operations
- **Matplotlib** – Basic data visualization
- **Seaborn** – Advanced statistical visualizations
- **Warnings** – To suppress unnecessary warning messages

Data Preprocessing Steps

Several preprocessing steps were applied to ensure data quality and consistency:

1. **Column Cleaning**

2. Removed special characters like `*` and `-`

3. Converted column names to uppercase for uniformity

4. **Duplicate Removal**

5. Duplicate rows were identified and removed

6. **Missing Value Handling**

7. Numerical columns were filled using median values

8. Categorical columns were filled using mode values

9. **Datetime Conversion**

10. Converted start and end date columns into proper datetime format

11. Extracted date, time, hour, weekday, and month from datetime values

12. **Feature Engineering**

13. Created new features such as:

    - **RIDE_DURATION** (in minutes)
    - **DAY_NIGHT** (Morning, Afternoon, Evening, Night)
    - **ROUND_TRIP** (Yes/No based on same start and stop location)
    - **FARE (11 units per Mile)**

14. **Column Reordering**

15. Rearranged columns for better readability and analysis

Exploratory Data Analysis (EDA)

Exploratory analysis was performed to identify trends and patterns in Uber rides:

- • Ride frequency across different hours of the day

• Ride distribution across weekdays
• Monthly ride trends
• Ride behavior based on category and purpose
• Distance distribution across time periods

Summary statistics were generated to understand numerical and categorical data distributions.

Data Visualizations

The following visualizations were created to support insights:

• **Count Plots** – Rides by hour, weekday, month, category, and purpose
• **Heatmaps** – Relationship between purpose, category, and day/night
• **Box Plots** – Miles traveled across day/night and categories
• **Violin Plots** – Distribution of miles by purpose and time of day
• **Scatter Plots** – Miles vs hour and ride duration
• **Line Plots** – Relationship between ride duration and distance
• **Bar Plots** – Total miles by purpose (day vs night)
• **Swarm Plots** – Individual ride distance distribution

These visualizations help in identifying peak hours, travel behavior, and ride trends.

Key Insights

Some important insights derived from the analysis include:

• Most rides occur during **morning and evening hours**, indicating peak commuting times
• **Business rides** dominate the dataset compared to personal rides
• Certain weekdays show higher ride frequency
• Longer ride distances are generally associated with business purposes
• Round trips are less common compared to one-way trips

Challenges Faced

• Handling missing values in categorical columns
• Datetime format inconsistencies
• Managing multiple derived features without data leakage

These challenges were addressed using appropriate preprocessing techniques.

Conclusion

This project demonstrates a complete data science workflow starting from raw data to meaningful insights. The analysis helps understand Uber ride behavior based on time, distance, purpose, and category. The project follows a beginner-friendly and structured approach, making it easy to understand and extend further.

Future Improvements

• Incorporating fare data for pricing analysis
• Applying machine learning models for ride demand prediction
• Using geospatial analysis for location-based insights
• Creating interactive dashboards for better visualization

Author

Project developed by 'YOGANANDHAM P' Data Science Beginner

This project was created for academic learning and evaluation purposes.