**Ford GoBike Data Analysis**

**Overview**

This project analyzes Ford GoBike trip data to uncover key insights into bike-sharing usage patterns, user demographics, and trip durations. By examining ride details such as start and end times, station locations, user information, and membership types, the study provides a deeper understanding of user behavior. Through data cleaning, visualization, and statistical analysis, the project highlights trends that can help improve bike-sharing services and enhance user experience.

**Dataset**

The dataset used in this analysis is the April 2019 Ford GoBike trip data, downloaded from:  
[Ford GoBike Data](https://www.fordgobike.com/)

**Objectives**

* Clean and preprocess the dataset (handle missing values, outliers, and data type conversions).
* Perform exploratory data analysis (EDA) to identify key trends and patterns.
* Visualize trip behaviors based on station popularity, user demographics, and trip durations.

**Steps Taken**

**1. Data Cleaning**

* Dropped irrelevant columns (latitude, longitude).
* Converted data types for better analysis (e.g., start time to datetime, birth year to integer).
* Removed missing values in critical columns like station names, user gender, and birth year.

**2. Feature Engineering**

* Created a duration\_in\_minutes column from duration\_sec.
* Derived user age from birth year and categorized users into age groups (Teenagers, Young Adults, Middle-Aged, and Older Adults).

**3. Exploratory Data Analysis**

* Identified the most popular start and end stations.
* Determined the most frequently used bike.
* Analyzed user demographics (age distribution, gender breakdown).
* Measured the number of round trips at the busiest stations.
* Examined trip patterns across different days of the week.
* Compared trip behaviors between bike-share-for-all users and other riders.

**4. Outlier Removal**

* Used the Interquartile Range (IQR) method to detect and remove outliers in ride duration and age.
* Applied a stricter IQR factor to refine the dataset further.

**5. Data Visualization**

* Bar charts for station popularity, age distribution, and membership type.
* Pie chart for gender distribution.
* Line plot showing trips per day of the week.
* Boxplots to visualize outliers before and after cleaning.

**Key Findings**

* The majority of users are subscribers, with fewer casual users.
* Most trips occur on weekdays, suggesting that the service is widely used for commuting.
* The most common age group among riders is middle-aged adults (31-40 years old).
* Trip durations differ significantly between bike-share-for-all users and other members.
* Customers take trips that are, on average, twice as long as those taken by subscribers.

**Tools Used**

* **Python** (Pandas, NumPy) for data manipulation.
* **Matplotlib, Seaborn** for data visualization.

**How to Use**

1. Run the script in a Jupyter Notebook or Python environment.
2. Ensure required libraries are installed:
3. pip install pandas numpy matplotlib seaborn
4. Execute the script step by step to explore the cleaned dataset and visualizations.

**Future Improvements**

* Incorporate additional datasets to analyze trends over multiple months.
* Perform predictive modeling to forecast bike demand.