**NY Citi Bike Data Analysis: Enhancing Service and Customer Experience**

**Introduction**

Citi Bike is the largest bike-share program in the U.S., offering a convenient and eco-friendly transportation option with over 20,000 bikes and 1,300 stations across New York City and Jersey City. With its easy-to-use app, Citi Bike gathers large amounts of data, providing a rich source of insights into rider behavior, demographics, and usage patterns.

To help Citi Bike's stakeholders make informed, data-driven decisions, this analysis aims to answer key questions about service utilization and customer behavior. I will mainly focus on:

1. Identifying the Most Popular Pick-Up Locations & Routes

2. Understanding Trip Duration Variation by Age Group

3. Determining Which Age Group Rents the Most Bikes

4. Exploring Rental Patterns by User Type (One-Time vs. Long-Term Users)

5. Assessing the Impact of Age on Trip Duration

6. Identify the peak hours and days

**Step-by-Step Approach**

1. Data Cleaning and Preprocessing with Data Wrangler in VSCode

Before delving into insights, it's crucial to clean and preprocess the raw data. This includes:

- Handling missing values, outliers, and inconsistencies.

- Formatting variables like 'Start Time', 'Stop Time', and 'Trip Duration' for easier analysis.

- Creating relevant features such as 'Trip Duration in Minutes', 'Age Groups', and weather-related attributes like 'Temperature' and 'Season'.

2. Descriptive Statistics and Exploratory Data Analysis (EDA)

Using Python’s pandas, I will conduct EDA to uncover initial patterns. Key steps include:

- Calculating basic statistics for trip duration and rider demographics.

- Grouping data by age, user type, and day of the week to analyze behavioral trends.

- Identifying the busiest stations and timeframes.

3. Popular Pick-Up Locations

Objective: Optimize station placement and resource distribution.

- Method: Use spatial analysis tools like Power BI or Tableau to create a heatmap of the most frequently used stations. I will also do this using Folium in Python. Analyze 'Start Station ID' and 'Start Station Name' to identify stations with high demand.

- Recommendation: Invest in additional bikes or stations in these areas to meet growing demand.

4. Trip Duration by Age Group

Objective: Understand how trip duration varies with age.

- Method: Group data by 'Age Group' and calculate the average trip duration using Python’s seaborn for visualizing trends.

- Recommendation: Age-specific pricing models or marketing campaigns could be developed to cater to different rider needs.

5. Bikes Rented by Age Group

Objective: Identify which age group rents the most bikes.

- Method: Using SQL or Python, count the number of rentals for each age group and represent the data visually through bar charts in Power BI.

- Recommendation: Create tailored promotions or loyalty programs targeting the most active age groups.

6. Rental Patterns by User Type (One-Time vs. Long-Term)

Objective: Compare the behavior of short-term and long-term users.

- Method: Analyze the 'User Type' variable in conjunction with 'Trip Duration' and 'Weekday' data to observe trends. Use SQL for data filtering and Power BI for weekday-specific comparisons.

- Recommendation: Offer incentives to convert one-time users into subscribers based on their rental behavior.

7. Impact of Age on Trip Duration

Objective: Investigate the relationship between age and trip duration.

- Method: Apply regression analysis using scikit-learn to model the influence of age on trip duration. If the correlation is significant, this insight could inform pricing or station layout strategies.

- Recommendation: Citi Bike could offer age-based features or promotions to optimize rider experiences.

8. Peak Hours and Days

Objective: Identify peak usage times for effective resource allocation and maintenance scheduling.

* Method: Analyze the distribution of rentals by hour of the day and day of the week. Use SQL or Python to group rentals by 'Start Time' and extract the hour and day information. Visualize the data with heatmaps or line graphs to highlight the busiest periods.
* Recommendation: Increase bike availability and maintenance staff during peak hours and days to improve customer experience and reduce downtime.

Advanced Techniques for Portfolio Building

1. Churn Prediction

Goal: Identify users at risk of leaving the service.

Approach: Develop a classification model using scikit-learn (e.g., logistic regression or random forests) to predict churn based on user behavior data.

Recommendation: Implement retention strategies based on key predictors of churn.

2. Time Series Analysis

Goal: Forecast bike demand for better inventory planning.

Approach: Use time series models (ARIMA, Prophet) to predict future demand based on historical patterns.

Recommendation: Adjust bike availability in anticipation of seasonal or event-driven demand spikes.

3. Customer Segmentation

Goal: Segment users for targeted marketing efforts.

Approach: Apply clustering algorithms (e.g., K-means) to group users based on riding behavior and demographics.

Recommendation: Tailor promotions to distinct user segments, such as tourists or commuters.

4. Geospatial Analysis

Goal: Optimize station locations based on user demand.

Approach: Perform spatial analysis using QGIS or Folium to map demand density and suggest new station placements.

Recommendation: Add or relocate stations to minimize distance for users and ensure adequate bike availability.

Conclusion

By analyzing NY Citi Bike’s data, I can provide actionable insights to improve operations, customer experience, and marketing strategies. Leveraging techniques such as machine learning, geospatial analysis, and data visualization not only enhances Citi Bike’s decision-making process but also helps build a strong data analyst portfolio, showcasing a wide range of skills.