**DAV Final Project in Citibike Bike Sharing**

The endeavor focuses on Exploratory Data Analysis (EDA) with Citi Bike's dataset to discern pivotal insights. Tableau will be the primary tool for analysis and visual representation, complemented by Python for advanced multidimensional visualizations, such as T-sne. The objective is to uncover relationships and patterns among variables across different situations through visualization.

Citi Bike, sponsored by Citigroup and operated by Motivate (formerly Alta Bicycle Share), is a private-public bike-sharing initiative across New York City boroughs and Jersey City. The system, using PBSC Urban Solutions' BIXI technology, was under Jay Walder's leadership until **Lyft** acquired it in 2018.

**Objective –**

* EDA
* Basic Feature Engineering
* Predict Total Customer and Subscriber
* Monthly trip details based on Customer and Subscriber
* Predicting Busiest Start Stations and End Stations

This dataset contains ride details for the year 2021 from January to April. It contains 15 columns each:

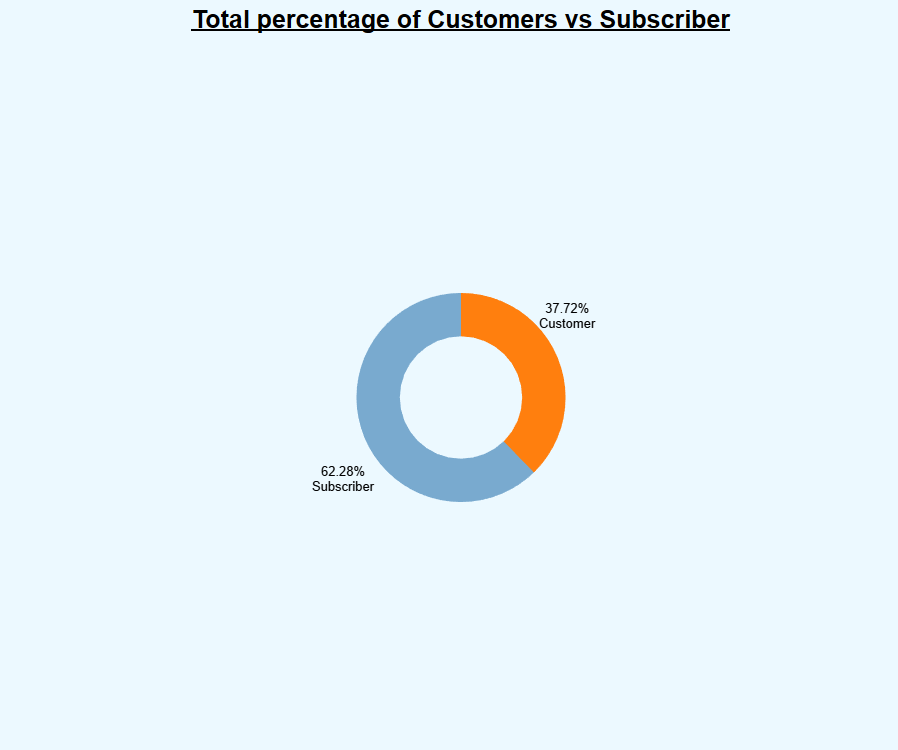
|  |  |
| --- | --- |
| tripduration | Duration in Seconds |
| starttime | Start Time and Date |
| stoptime | Stop Time and Date |
| start station id | ID of Start Station |
| start station name | Name of Start Station |
| start station latitude | Latitude of start station |
| start station longitude | Longitude of start station |
| end station id | ID of End Station |
| end station name | Name of End Station |
| end station latitude | Latitude of End station |
| end station longitude | Longitude of End station |
| Bike ID | Bike ID |
| usertype | Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member |
| birthyear | Year of Birth |
| gender | Zero=unknown;   1=male; 2=female |

**Donut Chart:**

For the very first Visual we have we are representing the total percentage of Customers and Subscribers with the help of Donut Chart.

**The Variable used here are:**

UserType: Customers vs Subscribers



**Tree map:**

The second visualization is a treemap that illustrates the total number of customer trips originating from each start station and ending at each destination station. The treemap's color intensity and the size of each segment represent the volume of customer trips, with the 5 Corner Library to Liberty Light Rail route accounting for 41,551,198 trips, making up 10.31% of the dataset. Larger, darker segments in the treemap correspond to areas with higher customer traffic.

**The Variable used here are:**

End station

Start Station

Aggregated Customer trip

Aggregated Subscriber trip

A screenshot of a green chart

Description automatically generated

**Side-by-Side bar graph:**

In this subsequent visualization, we scrutinize the aggregate monthly trips by customers and subscribers. The data distinctly shows that customers often surpass subscribers in trip volume. June stands out as the month with the highest biking engagement, boasting an average of 63 million customer trips. Conversely, the peak months for subscribers are August and September, each with about 25 million trips. The chart suggests various factors contributing to the surge in June, which could include favorable weather, city events, or tourism peaks.

* **Weather Conditions**: June usually has more favorable weather for outdoor activities like biking, with longer daylight hours and warmer temperatures.
* **Tourism and Events**: June is a popular month for tourists, and there might be more events encouraging outdoor activities.
* **Commuter Patterns**: Changes in commuting behavior with the onset of warmer weather could lead to more people choosing bikes over other modes of transport.
* **School Vacations**: The beginning of school vacations might increase leisure activities, including bike riding.

**The Variable used here are:**

UserType

Month

Aggregated Customer trip

Aggregated Subscriber trip

A graph of trips per month

Description automatically generated

Previous visualizations indicate that Citi Bike's customer base outnumbers its subscribers. To delve deeper, we analyzed the most frequented starting and ending stations among both customer and subscriber groups.

**TOP 5- START STATION:**

The chart illustrates the most popular origin stations for Citi Bike users, distinguishing between customers and subscribers**. Liberty Light R**ail leads for customers with 40.95% of starts, while Newport Pkwy is the top for subscribers at 28.58%. Conversely, **Newport Pkw**y is the second most popular for customers, and Liberty Light Rail for subscribers, both at 20.24%.

**The Variable used here are:**

User Type

Aggregated Total Trips

Start Stations Name

A graph with different colored bars

Description automatically generated with medium confidence

**TOP 5- END STATION:**

The chart outlines the five most utilized end stations by Citi Bike customers and subscribers. **Liberty Light Rail tops** for customers, accounting for 37.73% of their trips, while **Newport Pkwy** follows with 20.94%. On the subscriber side, **Newport Pkwy** ranks highest with 26.23% of trips, and **Liberty Light Rail** comes in at 20.39%.

**The Variable used here are:**

User Type

Aggregated Total Trips

End Stations Name

**A graph with different colored bars

Description automatically generated**

**START MAP:**

On the "Start Map," orange dots represent trips initiated by casual customers, while blue dots signify those begun by subscribers. This color coding visually distinguishes the two user types, allowing for an immediate understanding of their spatial distribution and relative activity levels across the map. It illustrates which areas are frequented more by casual riders versus subscribers, providing valuable insights for targeted marketing and resource allocation.

The visualization indicates that Jersey City has the most significant concentration of bike usage, with customers predominantly opting for bicycle travel.

**The Variable used here are:**

Start Station Longitude

Start Station Latitude

A map of a city with orange dots

Description automatically generated

**END MAP:**

In the "End Map" visualization, the use of orange dots to represent customers and blue dots for subscribers would allow for a clear distinction between where each user type ends their bike rides. This differentiation helps to identify which locations are preferred by casual riders compared to regular subscribers, offering insights into usage patterns that can inform service improvements and marketing strategies.

These dots indicate the volume or frequency of rides concluding at various stations, with different sizes possibly reflecting more or fewer rides. The tooltip suggests that the data includes details such as station latitude and longitude, station names, and user types, which could be customers or subscribers. This map provides insight into which stations are most commonly used as final destinations within the bike-sharing network.

The map highlights Jersey City as the epicenter for bike usage among customers, while in New York, the usage by customers and subscribers is more evenly spread out.

**The Variable used here are:**

End Station Longitude

End Station Latitude

**A map of a city

Description automatically generated**

**Heat Map:**

The map provided appears to be a heatmap visualizing the relationship between "Start Station Name" and "End Station Name" for a bike-sharing system. Each cell's color intensity indicates the frequency of trips from a specific start station to a specific end station. Darker or more vivid colors typically represent a higher number of trips, suggesting popular routes or commonly traveled paths between stations. The heatmap allows for quick identification of patterns in rider behavior, such as preferred routes or stations with high connectivity.

Looking at the map we can see that the color red represents the highest trip frequency Liberty Light Rail with a total Trip of 34,528,695. The color orange represents the second highest trip frequency at Newport Pkwy with total trips 15,406,941.

**The Variable used here are:**

End Station Name

Start Station Name

A screenshot of a computer screen

Description automatically generated

**T-sne:**

This chart visualizes the multidimensional data of bike share usage using t-SNE, a technique for dimensionality reduction. Each point represents a combination of a start and end bike station, colored by the logarithmically transformed duration of the trips. The spatial clustering of points reveals patterns in station usage and trip durations, providing insights into user behavior and network dynamics. The data was sampled from a larger dataset and then projected onto two dimensions to enable this visualization.

A screen shot of a graph

Description automatically generated

**CONCLUSION:**

The compilation of visualizations from the Citibike Bike Sharing project provides a comprehensive analysis of usage patterns. The data reveals a higher number of customer trips compared to subscribers. Deep dives into the most active start and end stations show distinct preferences for certain locations, with Liberty Light Rail and Newport Pkwy being prominent for both user types. Monthly trip data exhibits seasonal trends and peak usage times, particularly in June for customers and August-September for subscribers.   
Despite subscribers constituting a higher percentage in the dataset, the actual number of trips made by customers surpasses that of subscribers. Each individual ride, lasting 30 minutes, incurs a charge of $4.49, while an annual subscription costs $205. Additionally, the appeal of New York as a prominent tourist destination could contribute to this observation.  
These insights can guide service optimization and targeted marketing strategies.