

**Athens University of Economy and Business**

**School of Information Sciences & Technology: Department of Informatics**

**Master of Science in “Data Science”**

**Course:**  Data Visualization and Communication

**“FINAL PROJECT – TEAM 2”**

**Students:** 1) Maria Grammatikopoulou (f3352310)

2) Vasileios Dimopoulos (f3352311)

3) Rafail Mpalis (f3352308)

4) Dimitrios Tsirmpas (f3352315)

**Professors:** 1) Prof. Dimitris Karlis

2) Ass. Prof. Nikos Platis

3) Ass. Prof. Konstantinos Bourazas

Athens, June 2024

**Introduction**

This project analyzes a dataset of bike-sharing rides to uncover patterns and insights about usage trends. The dataset includes detailed information about individual rides, such as start and end times, start and end locations, and whether the user was a member or a casual rider. By exploring these attributes, we aim to understand user behavior and optimize bike-sharing services.

**Purpose**

The primary purpose of this project is to visualize and communicate key aspects of bike-sharing data to provide actionable insights. We focus on identifying the most frequently used start and end stations, analyzing the duration of rides, and examining the income generated by the bike-sharing service. These insights can help improve service planning, enhance user experience, and optimize resource allocation.

**Implementation**

To achieve our objectives, we implemented several visualizations:

1. **Most Frequent Start and End Stations:**

* We created bar plots to highlight the stations with the highest number of ride commencements and terminations. This helps identify popular locations and potential hotspots for bike demand.

1. **Duration of Rides:**

* We plotted the distribution of ride durations to understand how long users typically use the service. This analysis can inform pricing strategies and operational efficiencies.

1. **Income Analysis:**

* We analyzed the revenue generated by distinguishing between casual and member rides. This helps in understanding the financial performance and user preferences, which is crucial for developing targeted marketing and membership strategies.

1. **Interactive Visualization**:

* We created interactive plots to explore the duration of rides in more detail, allowing users to filter and manipulate the data for deeper insights.
* Additionally, we developed an interactive map displaying all bike-sharing stations, providing a geographical perspective on station locations and their usage patterns.

**Dataset Description**

The dataset comprises the following fields:

* **ride\_id:** Unique identifier for each ride
* **rideable\_type:** Type of bike used (e.g., docked bike)
* **started\_at:** Start time and date of the ride
* **ended\_at:** End time and date of the ride
* **start\_station\_name:** Name of the start station
* **start\_station\_id:** Unique identifier for the start station
* **end\_station\_name:** Name of the end station
* **end\_station\_id:** Unique identifier for the end station
* **start\_lat:** Latitude of the start station
* **start\_lng:** Longitude of the start station
* **end\_lat:** Latitude of the end station
* **end\_lng:** Longitude of the end station
* **member\_casual:** User type (member or casual)

By examining these variables, we aim to deliver comprehensive visual insights that support strategic decision-making for bike-sharing services.

**Dataset Preprocessing**

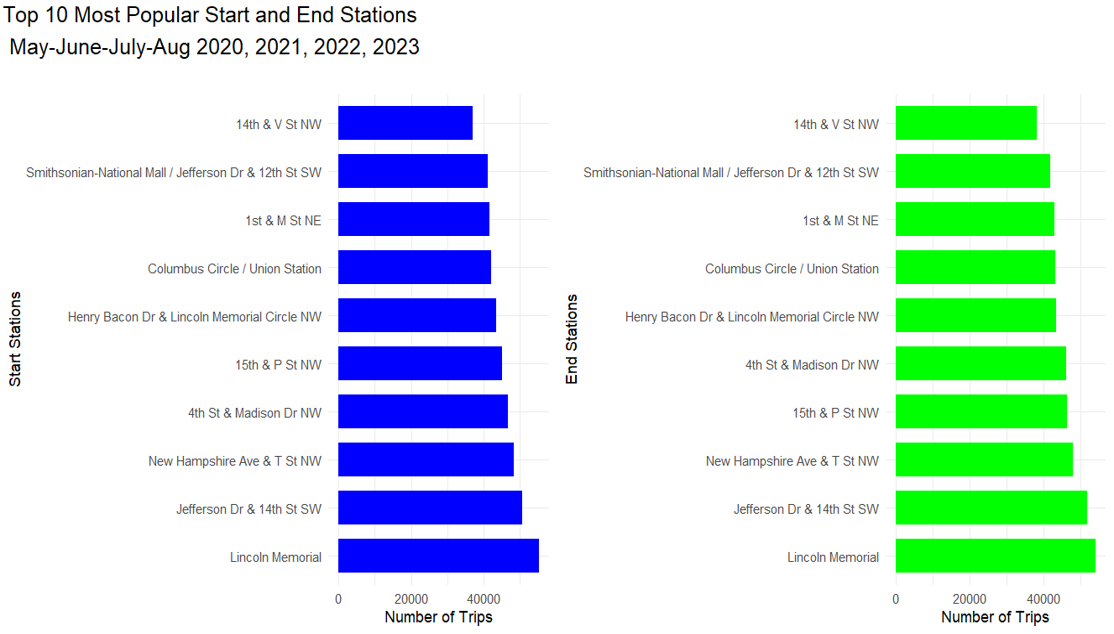
**Manipulation**

1. Before preprocessing the data, a script was required to merge multiple .zip files into a final .csv file to consolidate all the data into a single file.
2. Data preprocessing included the creation of new columns to calculate the duration of each ride.
3. Trips with negative durations (where the start of trip occurring after the end of the trip) were removed from the dataset, as their values are outside of the domain’s valid range.
4. A function was defined to transform ride times into pricing structures.

**Issues**

1. The scatter plot required data aggregation due to the large volume of data.
2. The use of the label "docked" bikes for only one month in the entire dataset was assumed to be incorrect, leading to their reclassification as "classic" bikes.
3. Both Tableau and R encountered issues with the dataset size, causing join operations with other datasets (such as map coordinates) to fail.
4. The dataset contained huge outliers, such as ride durations extending beyond a month, which were deemed improbable and required handling.

**1. Most Frequent Start and End Stations**



Plot 1: Top 10 Most Popular Start and End Stations

**Plot 1:** **Top 10 Most Popular Start and End Stations**

**Purpose:** This plot aims to identify the most frequently used bike-sharing stations for both the starting and ending of rides over a specified period (May-August, 2020-2023). Understanding the popularity of these stations can help in resource allocation, such as station maintenance, bike distribution, and infrastructure planning.

**Display:** The bar charts illustrate the top 10 stations where rides began (left) and ended (right). Each bar represents the number of trips starting or ending at the respective station, providing a clear comparison of station usage.



Plot 2: Most Popular End Station Each Month by Year

**Plot 2: Most Popular Start Station Each Month by Year**

**Purpose:** Similar to the second plot, this visualization focuses on the most popular start stations for each month over multiple years. This helps in understanding which stations are consistently popular for beginning rides and how this changes seasonally and annually.

**Display:** The bar chart displays the number of trips starting at the most popular station each month, with different colors representing different stations. This visual helps identify stable trends and significant changes in the preference of start stations over time.



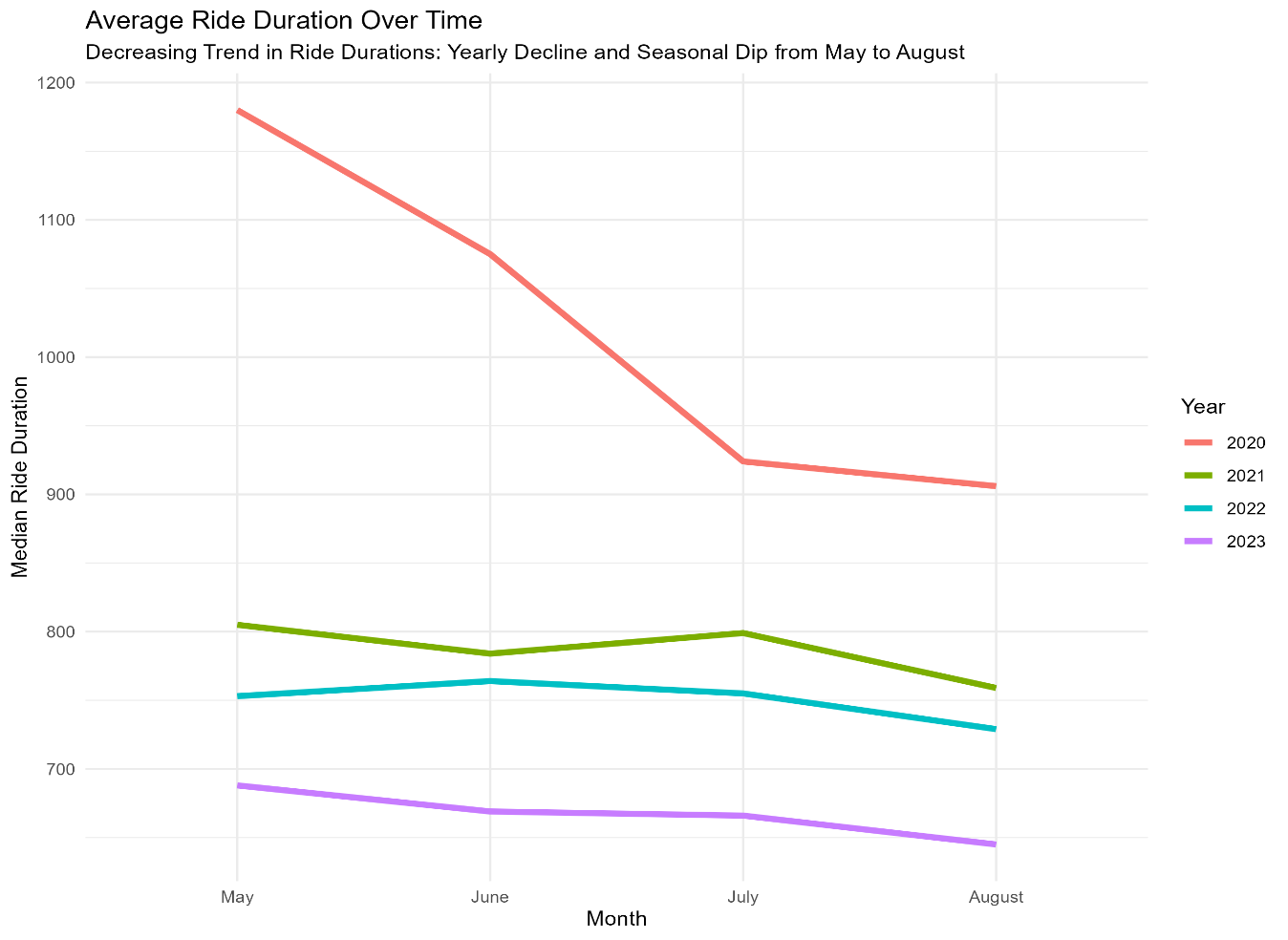
Plot 3: Most Popular End Station Each Month by Year

**Plot 3: Most Popular End Station Each Month by Year**

**Purpose:** This plot tracks the popularity of end stations on a monthly basis over several years. This temporal analysis helps to identify trends and seasonal patterns in bike-sharing usage, allowing for better anticipation of demand changes throughout the year.

**Display:** The bar chart shows the number of trips ending at the most popular station for each month, differentiated by color for each year. Each color represents a specific station, enabling easy identification of trends and shifts in popularity over time.

**2. Duration of Rides**

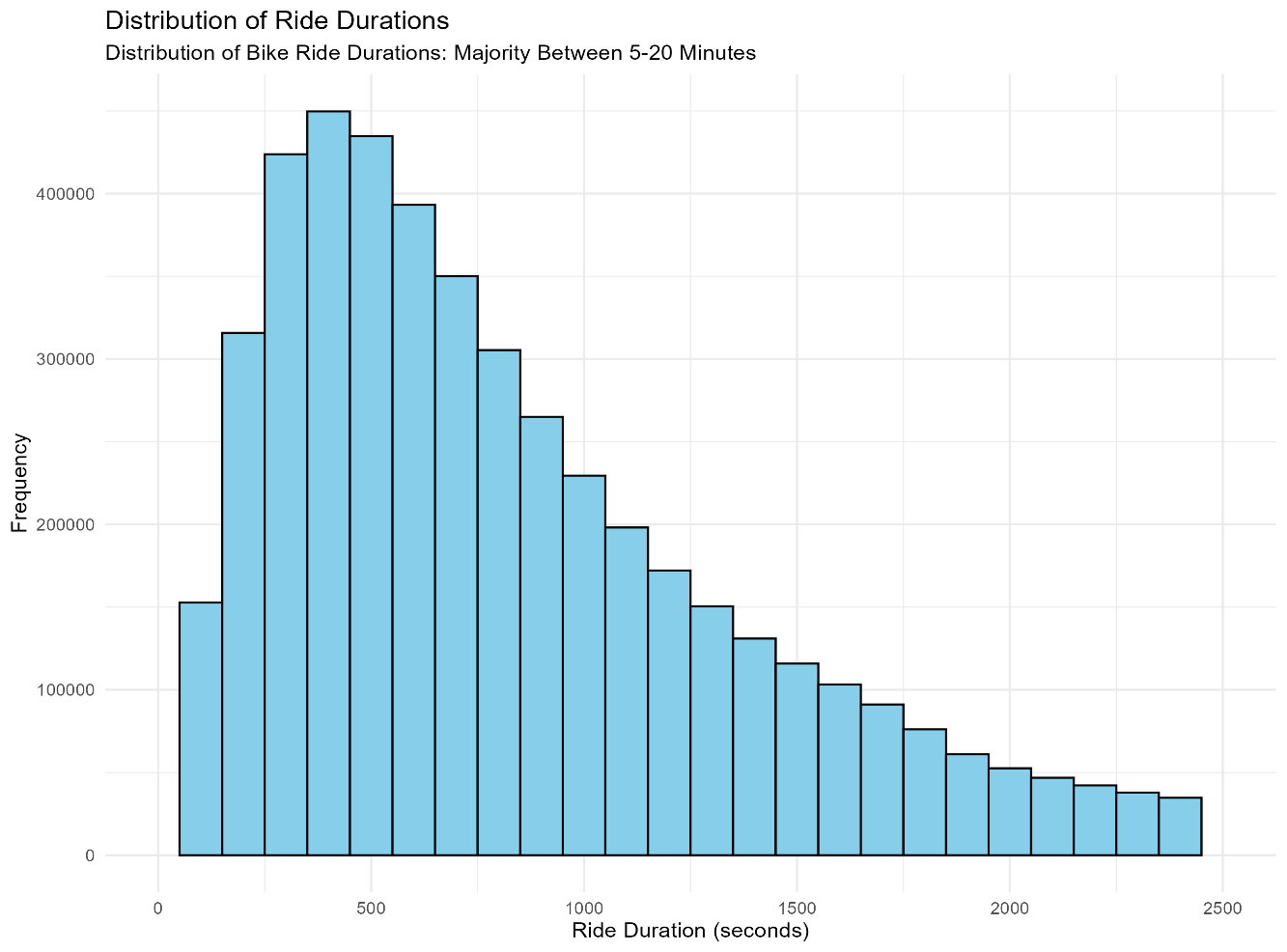


Plot 4: Average Ride Duration Over Time

**Plot 4: Average Ride Duration Over Time**

**Purpose:** This plot tracks the average ride duration over time, revealing trends and seasonal variations. It is useful for assessing the impact of external factors (e.g., weather, events) on ride durations.

**Display:** The line chart illustrates the average ride duration per month over several years. It shows a decreasing trend in ride durations, with a yearly decline and a seasonal dip from May to August, suggesting changes in user behavior and external influences on ride durations.

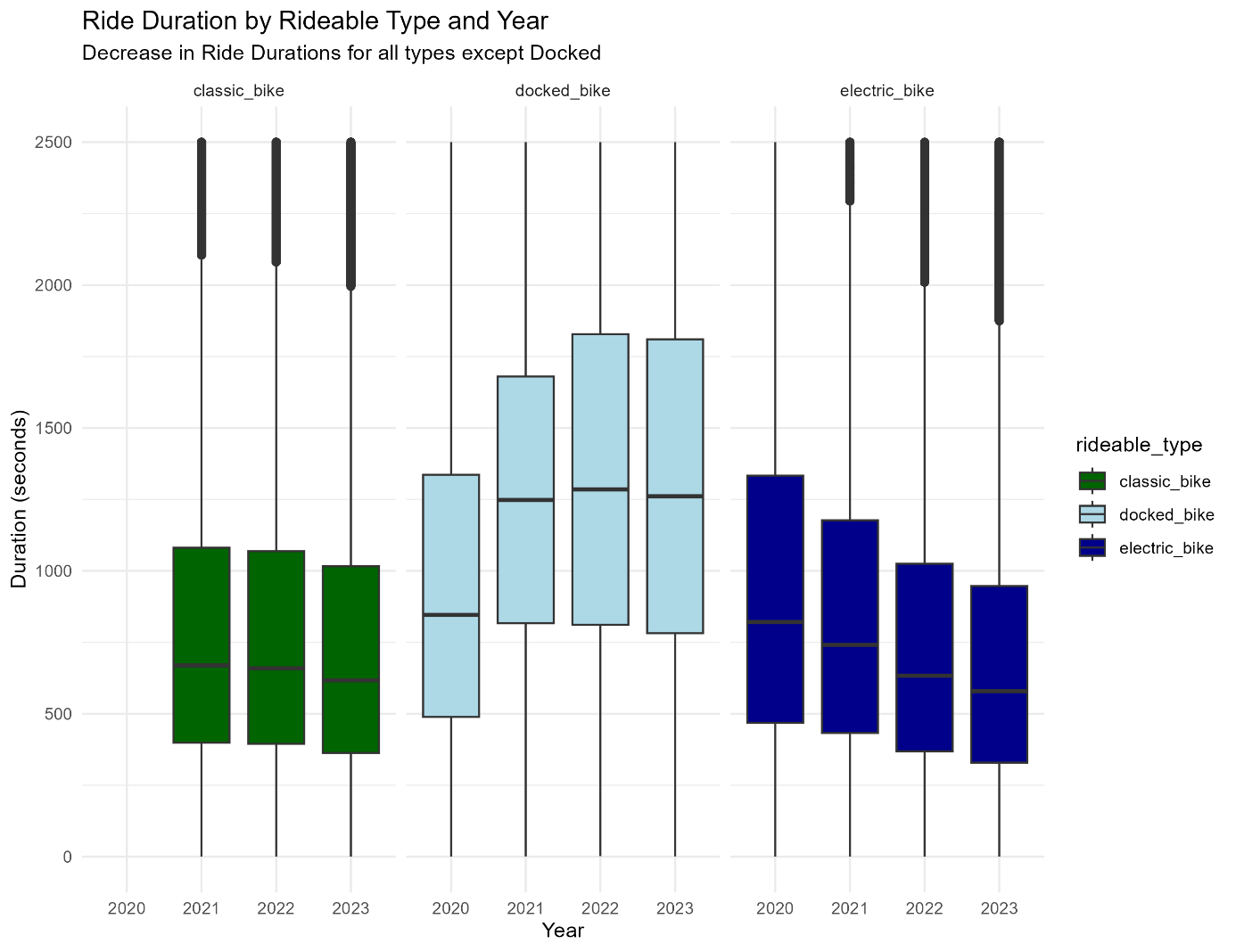


Plot 5: Distribution of Ride Durations

**Plot 5: Distribution of Ride Durations**

**Purpose:** This plot aims to provide an overview of the distribution of ride durations, helping to understand how long users typically spend on each ride. This information is useful for optimizing bike availability and pricing models.

**Display:** The histogram shows the frequency of rides for different duration intervals, with the majority of rides falling between 5-20 minutes. This indicates that most users prefer short trips, which can guide service improvements and marketing strategies.



Plot 6: Ride Duration by Rideable Type and Year

**Plot 6: Ride Duration by Rideable Type and Year**

**Purpose:** This plot compares the duration of rides across different types of bikes (classic, docked, electric) over the years. It helps to identify trends and changes in user behavior concerning different bike types.

**Display:** The box plot presents the median and distribution of ride durations for each bike type across different years. It shows that while docked bike durations remain relatively stable, there is a noticeable decrease in ride durations for classic and electric bikes over time.

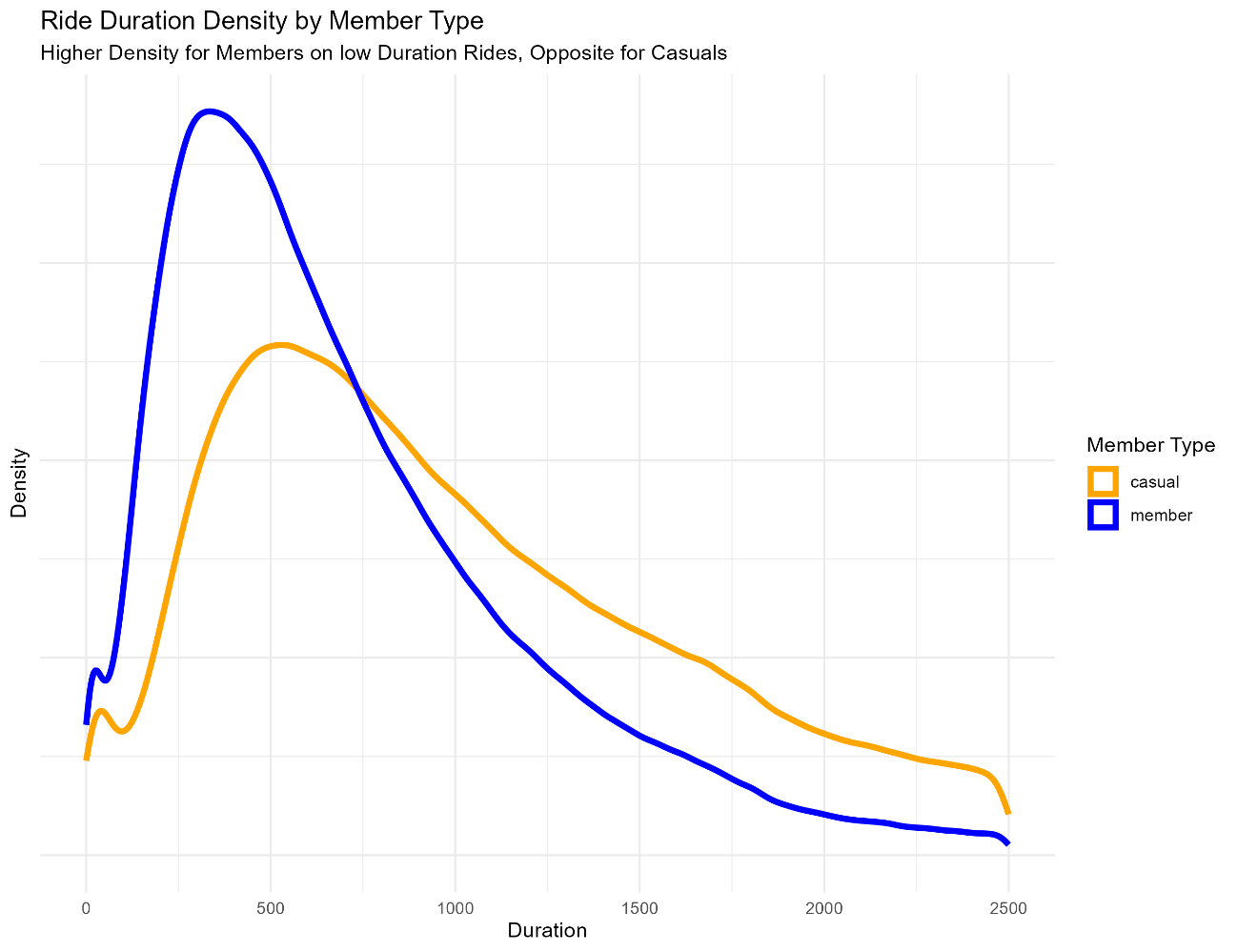


Plot 7: Median Ride Duration by Hour of the Day

**Plot 7: Median Ride Duration by Hour of the Day**

**Purpose:** This plot examines the variation in median ride duration throughout the day, providing insights into peak usage times and rider behavior patterns at different hours.

**Display:** The line chart depicts the median ride duration for each hour of the day across different months. It reveals a decrease in ride durations from 4-8 AM and an increase from 11 AM to 5 PM, highlighting the periods of highest activity and longest rides.



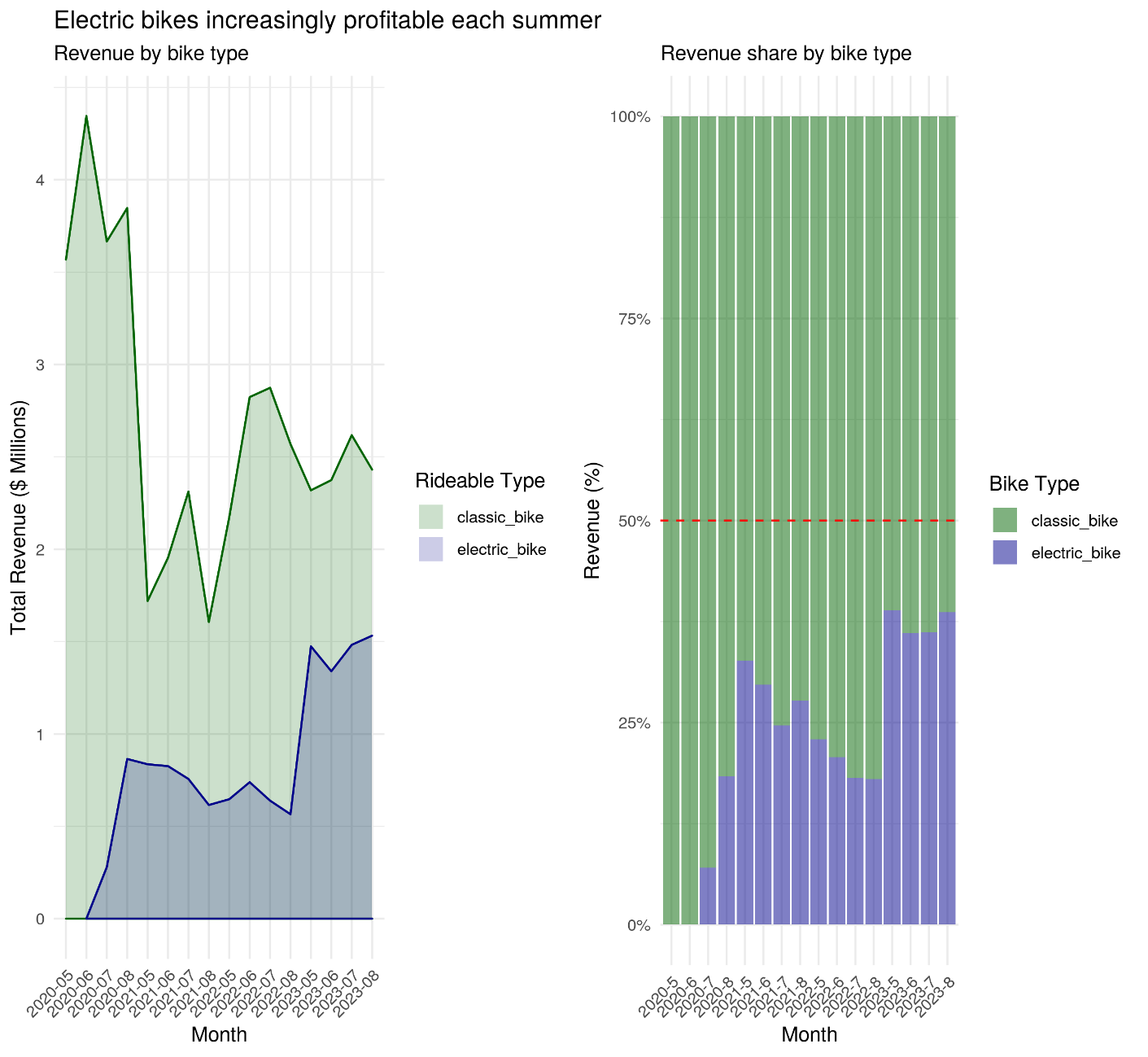
Plot 8: Ride Duration Density by Member Type

**Plot 8: Ride Duration Density by Member Type**

**Purpose:** This plot compares the distribution of ride durations between casual users and members. Understanding these differences helps tailor services and marketing efforts to each group.

**Display:** The density plot shows the ride duration distributions for casual and member riders. Members tend to have higher densities of shorter duration rides, while casual riders have a more spreads out distribution, often taking longer rides.

**3. INCOME ANALYSIS**

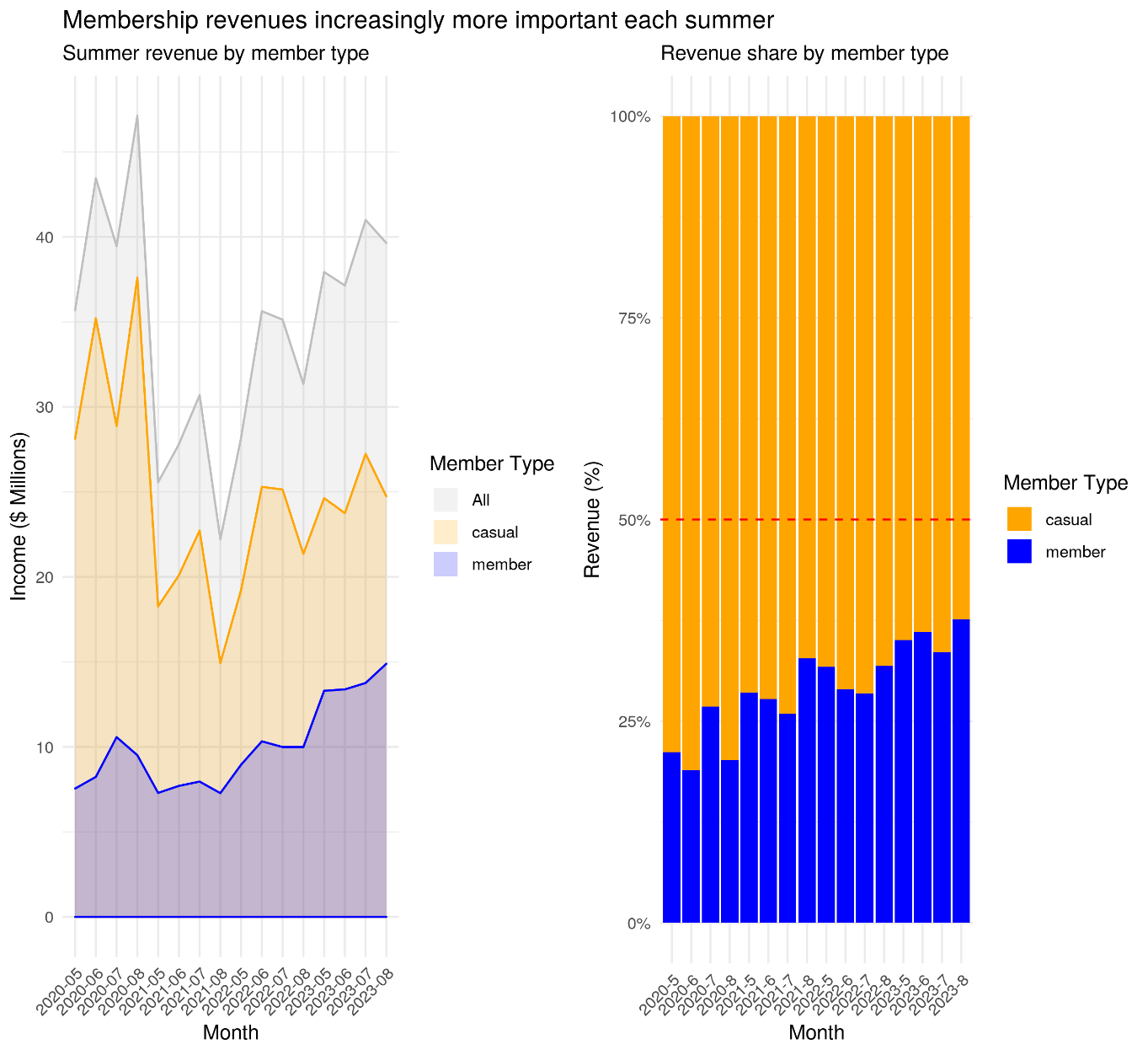


Plot 9: Revenue by Bike Type

**Plot 9: Revenue by Bike Type**

**Purpose:** This plot aims to illustrate the revenue generated by different types of bikes (classic and electric) over time. This helps in understanding the profitability and growing importance of electric bikes in the bike-sharing system.

**Display:** The left panel shows the total revenue in millions of dollars generated by classic and electric bikes over the months, highlighting the increasing contribution of electric bikes each summer. The right panel displays the revenue share by bike type, with electric bikes gradually capturing a larger percentage of the total revenue.

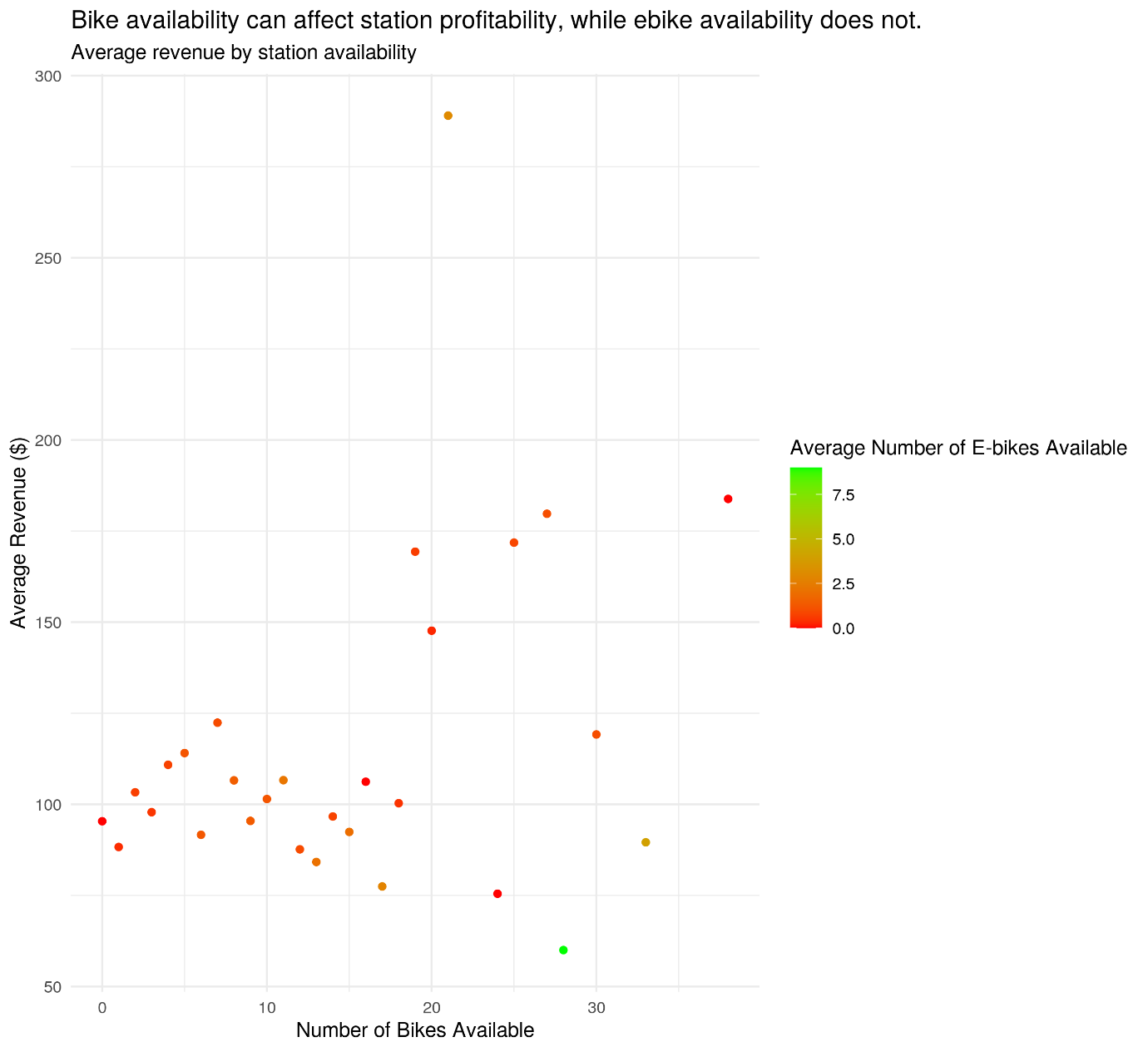


Plot 10: Revenue by Member Type

**Plot 10: Revenue by Member Type**

**Purpose:** This plot examines the revenue generated by different user types (casual and member) over time, with a focus on summer months. It highlights the importance of membership revenue in the overall income of the bike-sharing service.

**Display:** The left panel shows the total income in millions of dollars from casual users and members over the months. The right panel illustrates the revenue share by member type, showing an increasing contribution from members each summer.



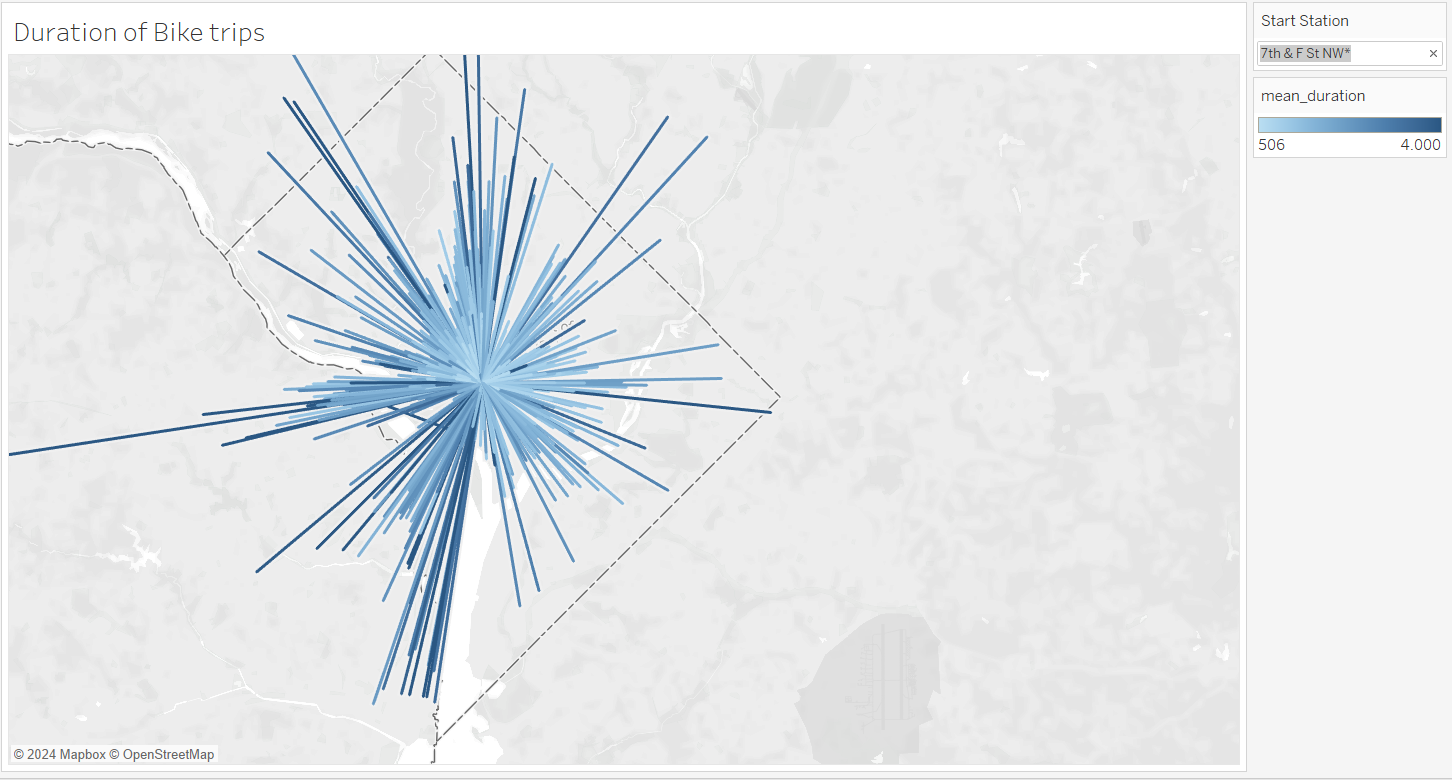
Plot 11: Average Revenue by Station Availability

**Plot 11: Average Revenue by Station Availability**

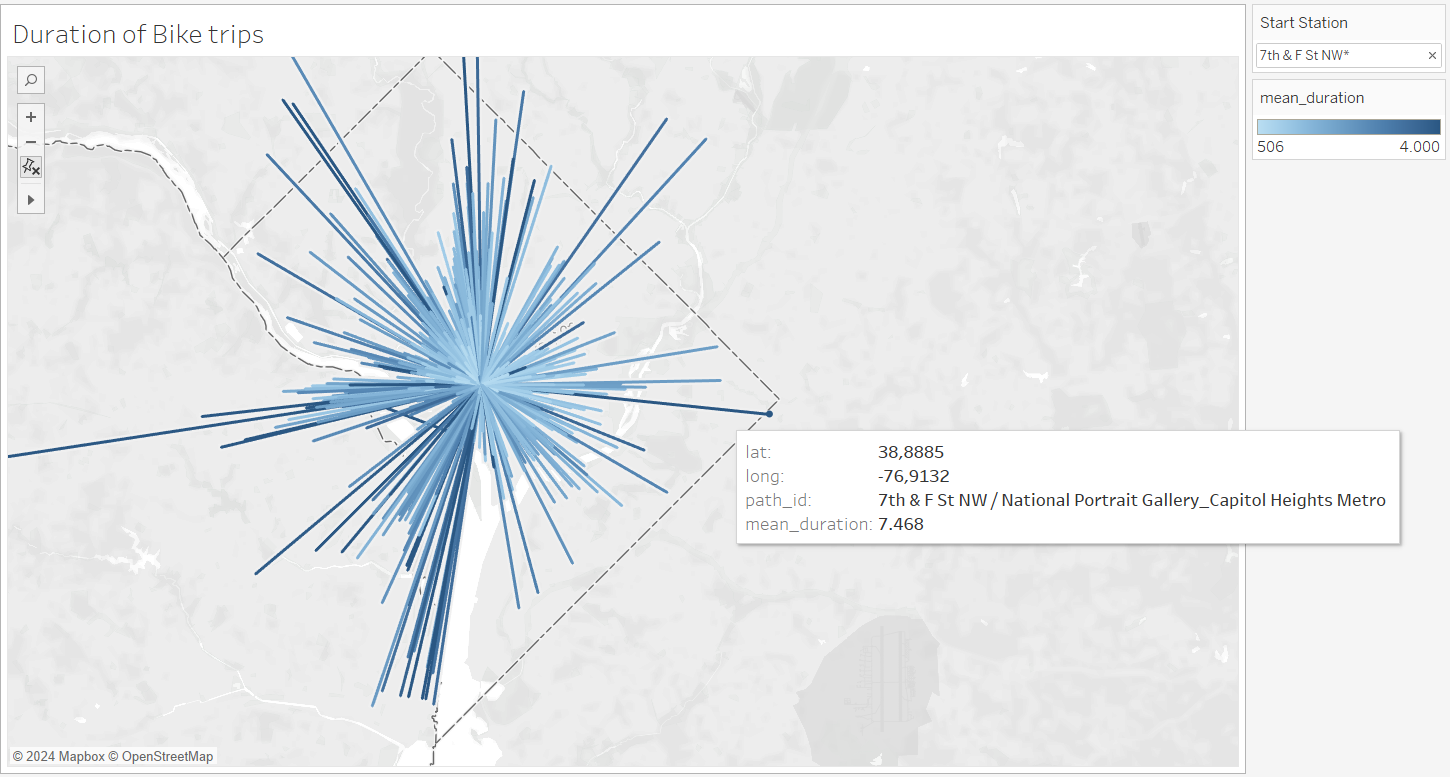
**Purpose:** This plot explores the relationship between bike availability at stations and the average revenue generated. It helps in understanding how the availability of bikes (and e-bikes) at stations affects station profitability.

**Display:** The scatter plot shows the average revenue per station against the number of bikes available, with color indicating the average number of e-bikes available. The plot suggests that while overall bike availability correlates with higher revenue, e-bike availability does not have a significant impact.

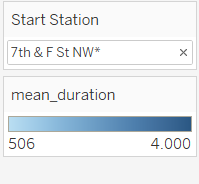
**4. Interactive Visualization**



Interactive Plot 1-a: Duration of Bike Trips by Start Station



Interactive Plot 1-b: Duration of Bike Trips by Start Station with a chosen station

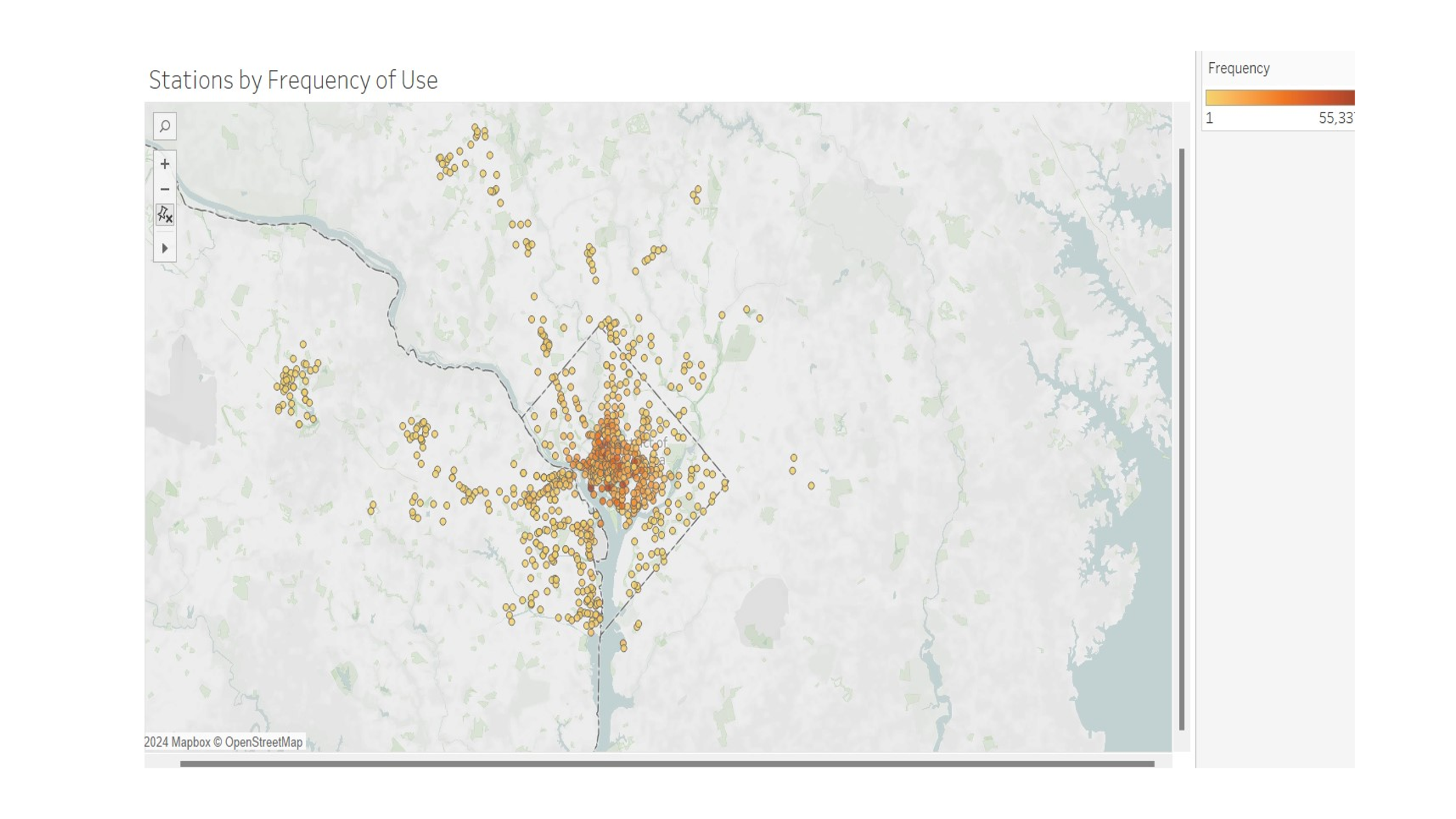


Interactive Plot 1-c: Table in the plot showing the duration of Bike Trips per Bike Station

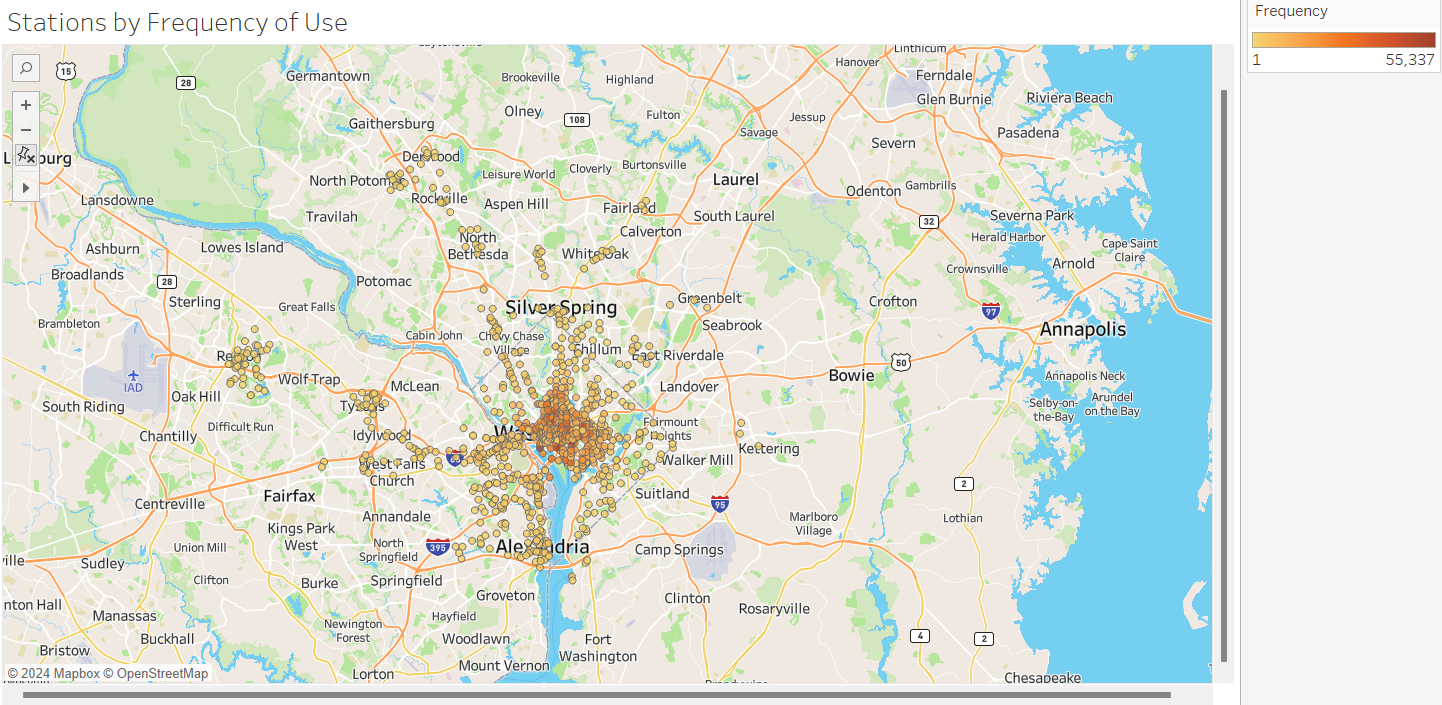
**Interactive Plot 1: Duration of Bike Trips by Start Station**

**Purpose:** This interactive plot aims to visualize the duration of bike trips originating from a selected start station. It helps users understand the distribution and length of trips starting from a specific location.

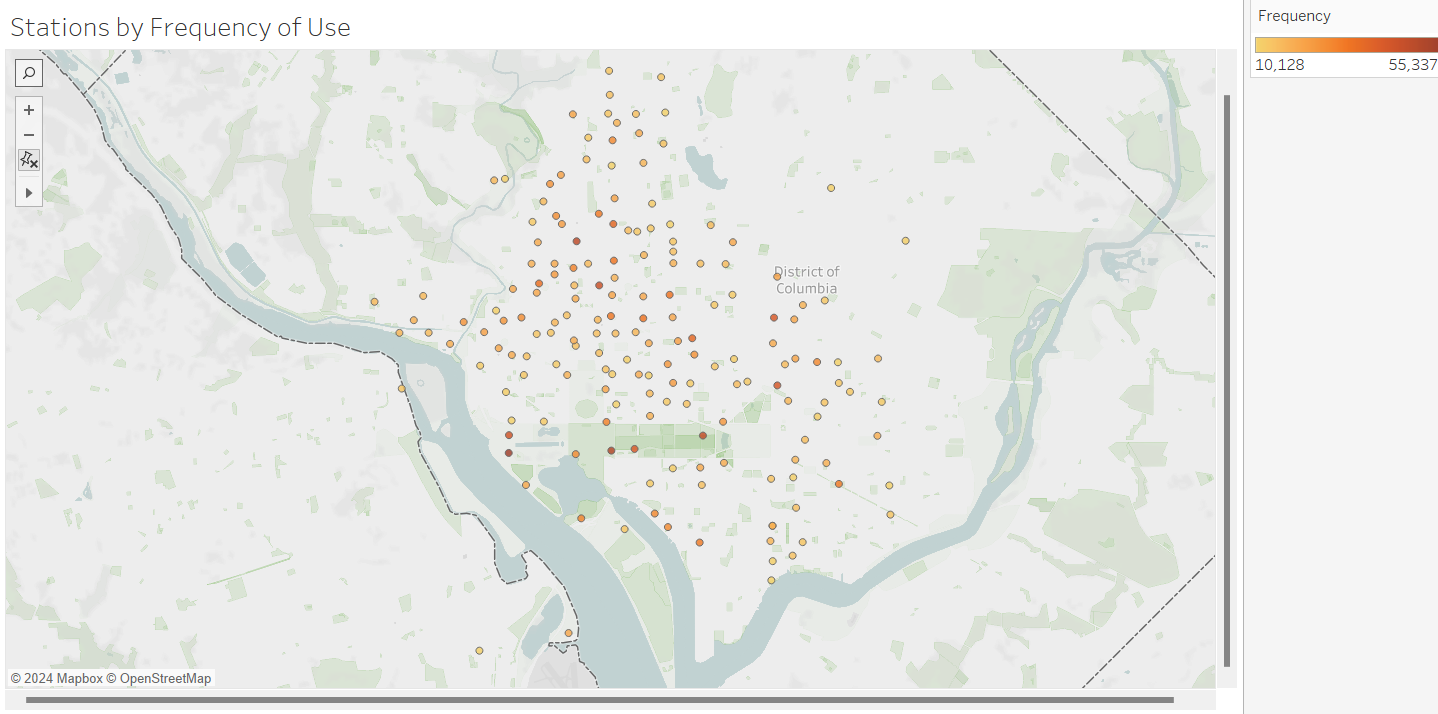
**Display:** The map shows bike trip durations radiating from the selected start station (e.g. "7th & F St NW"). Each line represents a trip, with its length indicating the duration. The color bar on the right provides a gradient scale of the mean duration of trips. Users can interact with the plot by selecting different start stations to see how trip durations vary by location.



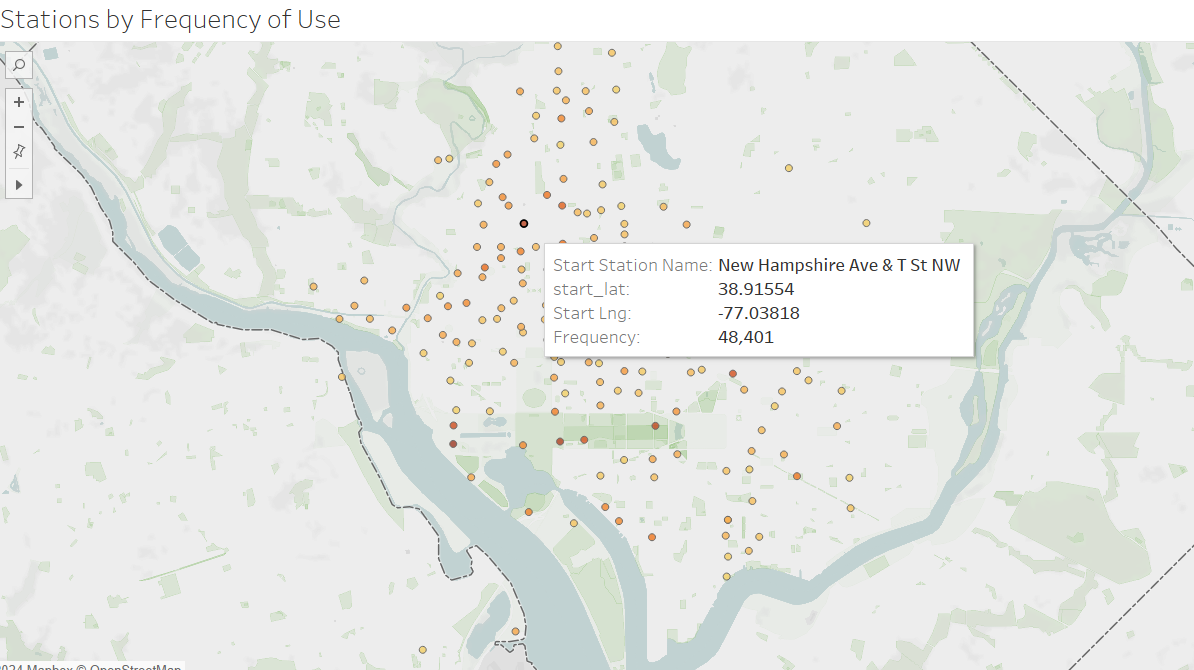
Interactive Plot 2-a: Map with Stations by Frequency of Use



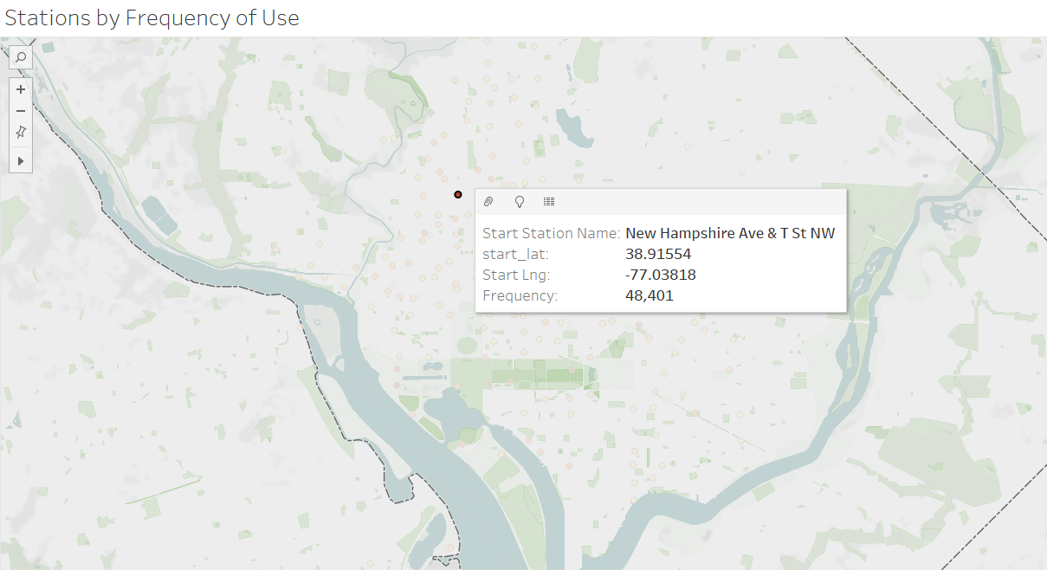
Interactive Plot 2-b: Map with Stations by Frequency of Use (with streets and areas)



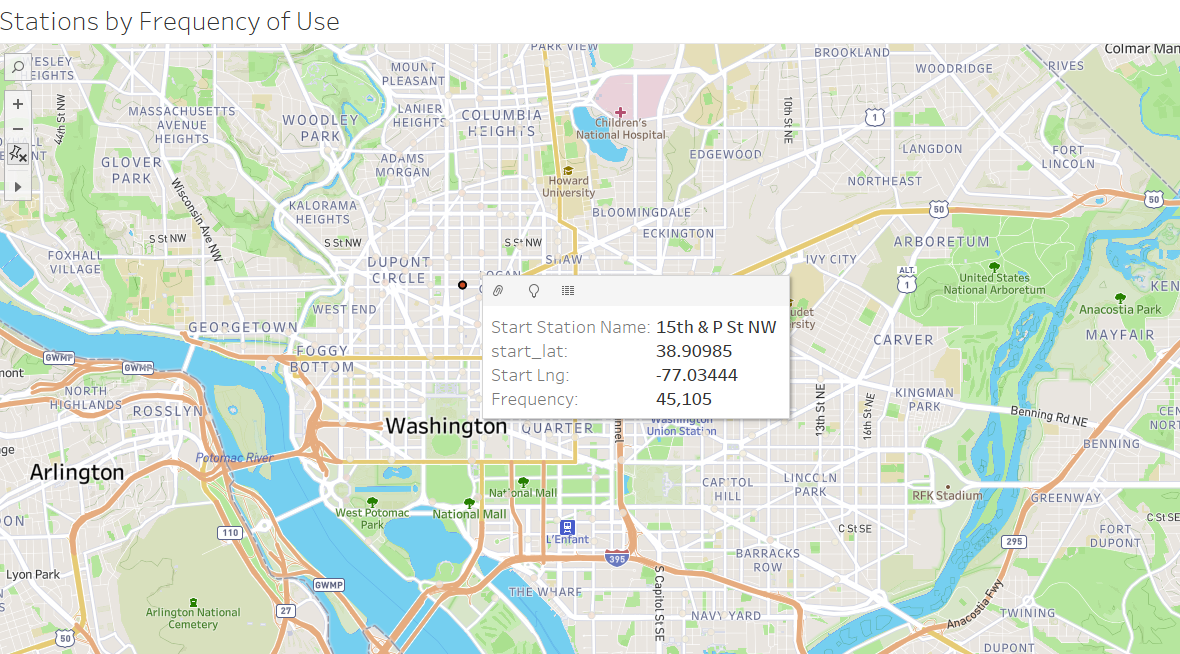
Interactive Plot 2-c: Map with Stations by Frequency of Use frequency of stations bigger than 10k)



Interactive Plot 2-d: Map with Stations by Frequency of Use (with the mouse on a chosen station)



Interactive Plot 2-e: Map with Stations by Frequency of Use (with a chosen station)



Interactive Plot 2-f: Map with Stations by Frequency of Use (with a chosen station, frequency biggen that 10k and the names of streets active)

**Interactive Plot 2: Map with Stations by Frequency of Use**

**Purpose:** This interactive plot aims to visualize the frequency of use for bike-sharing stations. It helps identify which stations are most frequently used and provides insights into spatial usage patterns across the service area.

**Display:** The map displays bike-sharing stations as points, with the color intensity representing the frequency of use. Users can interact with the plot by zooming in and out and clicking on individual stations to see detailed information such as station name, latitude, longitude, and usage frequency. This interaction allows users to explore the geographical distribution and usage intensity of different stations.

**CONCLUSION**

These visualizations collectively provide valuable insights into the usage patterns, ride duration trends, and revenue dynamics of bike-sharing services. In more detail:

1. **Consistent Popular Stations**: Certain stations, such as "14th & V St NW" and "Smithsonian-National Mall / Jefferson Dr & 12th St SW," consistently rank among the top for both starting and ending rides. This indicates that these are key locations for users.
2. **Monthly Shifts**: Over the years, some stations have seen increased or decreased usage, reflecting changes in user preferences, urban development, or station accessibility.
3. **Short Trip Preference**: The majority of rides are short, typically between 5-20 minutes, indicating a preference for quick, convenient trips.
4. **Stable Docked Bike Usage**: While the ride durations for docked bikes remain stable, there is a noticeable decrease in durations for classic and electric bikes over the years.
5. **Peak Usage Hours**: Ride durations peak between 11 AM and 5 PM, suggesting higher demand and longer usage during these hours.
6. **Member vs. Casual Rides**: Members tend to take shorter rides compared to casual users, who often have more varied and longer ride durations.
7. **Decreasing Trend**: There is a general decrease in average ride durations over time, with significant drops observed year-on-year and seasonally from May to August.
8. **Growing Importance of Electric Bikes**: The revenue generated by electric bikes has been increasing steadily each summer, indicating their growing popularity and profitability.
9. **Membership Revenue Significance**: Revenue from members has become increasingly important, especially during summer months, highlighting the value of promoting membership plans.
10. **Impact of Bike Availability on Revenue**: There is a positive correlation between the number of bikes available at stations and the average revenue generated. However, the availability of e-bikes does not show a significant impact on station profitability.

These insights are crucial for optimizing bike distribution, improving user experience, and planning future infrastructure to support the growing demand for bike-sharing services. They also help in strategic decision-making regarding bike fleet management, membership promotions, and resource allocation to maximize revenue and improve service efficiency.

**Participation in the Final Project**

For this assignment each member of the team contributed equally. In more detail:

* Maria Grammatikopoulou: Station Analysis

* Vasileios Dimopoulos: Trip Duration Analysis
* Rafail Mpalis: Report & Presentation
* Dimitrios Tsirmpas: Income Analysis