

**Department of Industrial and Systems Engineering**

**ISYE 570 Final Project Report**

**New York City Bike Data Analysis**

Group Members:

**Fouzan Abdullah Z1978719**

**Mohammed Ali Shajee Z2008300**

**Md Abu Naeem Khan Z1985093**

ISYE 570: Intro Data Analytics Engineers

Dr. Christine Nguyen

December 12, 2024

[2.0 Data Preparation & Cleaning 3](#_Toc1360109428)

[2.1 Getting a Hang of the Data: 3](#_Toc2093193843)

[2.2 Merging of Data: 4](#_Toc1053487879)

[2.3 Dropping Unwanted Columns: 4](#_Toc434436747)

[2.4 Handling Data Types and Conversions: 4](#_Toc1082307560)

[3.0 Analysis & Results 5](#_Toc259700359)

[3.1 Tableau 5](#_Toc989391645)

[3.1.1 Busiest Route Analysis 5](#_Toc1461886562)

[3.1.2 Monthly Trip Volume Analysis 6](#_Toc1583657798)

[3.1.3 Weekly Trip Volume Analysis 8](#_Toc1663214216)

**List of Tables**

Table 1 Raw Data - Fields & Data Type …………………………………………3

Table 2 Data Preparation & Cleaning ....…………………………………………4

**List of Figures**

Figure 1 Busiest Route…...........................………………………………………7

Figure 2 Count of Trips over the Months in 2022 ....…………………………….……8

Figure 3 Count of Trips over the Week in 2022 ………………………………..…9

Figure 4 Trips Hour v/s Month...........................………………………………..…10

Figure 5 Trips Hour v/s Week..............................................................................................11

Figure 6 Top stations with over 20,000 trips ………………………………………….12

Figure 7 Peak Hours of Trips Taken ………………………………………………....13

Figure 8 Rideable Type used over the Months ………………………………………..14

Figure 9 Rideable type used by Different Memberships ………………………………15

Figure 10 Trips Week v/s Months …………………………………………………..16

Figure 11 Test Output ………………………………………………………………17

Figure 12 Heat Map – Correlation …………………………………………………..18

Figure 13 Correlation Ride Duration vs Time of Day Output........................................…19

Figure 14 Correlation Ride Duration vs Hour of Day Output ………………………………19

## **1.0 Introduction**

In the vibrant landscapes of cities like New York, “The City that Never Sleeps”, bike-sharing programs such as Citi Bike have become a preferred mode of transportation. This is particularly evident in a crowded metropolis like New York City, where traditional transport options often come with congestion and jams. Our project's objective is to evaluate the operational effectiveness of the Citi Bike program based on usage patterns throughout 2022.

Our dataset encompasses detailed records from trips made via Citi Bikes, capturing each journey's specifics such as start and end times, and the locations of departure and arrival. The data spans all twelve months of 2022, offering a comprehensive view from January through December. Each record in the dataset is tied to unique trip identifiers and includes additional information such as the type of bike used, whether the trip was one way or round trip, and the membership status of the user. This allows us to analyze the data down to specific details like trip durations and the geographic coordinates of stations, presented in a time-series format.

Moreover, the dataset includes station identifiers along with names, which we will map to their respective locations. This analysis will not only help us in understanding the temporal and spatial usage patterns of Citi Bike but also assist in identifying potential areas for system improvement and expansion (if any).

Table 1 Raw Data - Fields & Data Type

|  |  |  |
| --- | --- | --- |
| S. No. | Field | Data Type |
| 1 | ride\_id | object |
| 2 | rideable\_type | object |
| 3 | started\_at | object |
| 4 | ended\_at | object |
| 5 | start\_station\_name | object |
| 6 | start\_station\_id | object |
| 7 | end\_station\_name | object |
| 8 | end\_station\_id | object |
| 9 | start\_lat | float64 |
| 10 | start\_lng | float64 |
| 11 | end\_lat | float64 |
| 12 | end\_lng | float64 |
| 13 | member\_casual | object |

# **2.0 Data Preparation & Cleaning**

Data cleaning and preparation is the most pivotal and primary step in any data analysis. Data cleaning helps to remove inaccurate information, which may lead to incorrect analysis. It aims to identify missing values, eliminate duplicates, eliminate outliers, and delete incorrectly formatted data. Data cleaning helps to ensure that the data is valid, correct, and consistent by deleting faulty records to perform further analysis. The following sections describe the different phases involved in the data cleaning and data preparation phase:

## **2.1 Getting a Hang of the Data:**

The first step in our data analysis involved getting to know our data. For this first we loaded all the files for the twelve months in a list using pandas read csv function. Thus we got 37 total dataframes for the twelve months.

## **2.2 Merging of Data:**

Since the data was available for each month before performing any analysis, it was necessary to merge the records in all twelve months. The individual data was merged with the help of concatenate function in python and union in Tableau. After merging all monthly data, there were 29,838,806 records with 13 different attributes.

## **2.3 Dropping Unwanted Columns:**

The dataset consisted of one column that was not required to perform analysis. The 'ride\_id’ column was dropped to make it easier to concentrate on the remaining columns as a new id was generated on each rental. During further analysis, rows containing null values were dropped again.

## **2.4 Handling Data Types and Conversions:**

To facilitate precise analysis, it was essential to convert various data types to formats more suited to analytical procedures. Timestamps originally stored as strings were converted into DateTime format, enabling more effective time series analysis. Similarly, fields that numerically represented but stored as strings were transformed to appropriate numeric formats, enhancing the computational efficiency.

Table 2 Data Preparation & Cleaning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Field | Description | Expected Cardinality | Actual Cardinality | Data Type | | Encoded Data Type | | Missing Values |
| Before Data Cleaning | After Data Cleaning |
| ride\_id | Unique identification for trip | Many | 29,768,714 | Quantitative | Discrete | object | object | 0 |
| rideable\_type | Classic ride/electric ride | 2 | 2 | Qualitative | Nominal | object | object | 0 |
| started\_at | The date and time the trip began | Many | 29,763,447 | Qualitative | Interval | object | datetime64 | 0 |
| ended\_at | The date and time the trip ende | Many | 29,733,761 | Qualitative | Interval | object | datetime64 | 0 |
| start\_station\_name | Start station name | Many | 1,761 | Qualitative | Nominal | object | object | 49 |
| start\_station\_id | The station ID from where the trip started | Many | 3,575 | Quantitative | Discrete | object | object | 49 |
| end\_station\_name | End station name | Many | 1,841 | Qualitative | Nominal | object | object | 70,092 |
| end\_station\_id | The station ID from where the trip ended | Many | 3,573 | Quantitative | Discrete | object | object | 70,092 |
| start\_lat | Latitude of start station | Many | 1,204,063 | Quantitative | Interval | float64 | float64 | 0 |
| start\_lng | Longitude of start station | Many | 892,870 | Quantitative | Interval | float64 | float64 | 0 |
| end\_lat | Latitude of end station | Many | 2,636 | Quantitative | Interval | float64 | float64 | 37,392 |
| end\_lng | Longitude of end station | Many | 2,627 | Quantitative | Interval | float64 | float64 | 37,392 |
| member\_casual | Member or Casual Rider | 2 | 2 | Qualitative | Nominal | object | object | 0 |

## **3.0 Analysis & Results**

In order to understand any trend in the data, an exploratory analysis of the data was performed. The following sections of the report discuss the results, and the analysis approach adopted for each variable

## **3.1 Tableau**

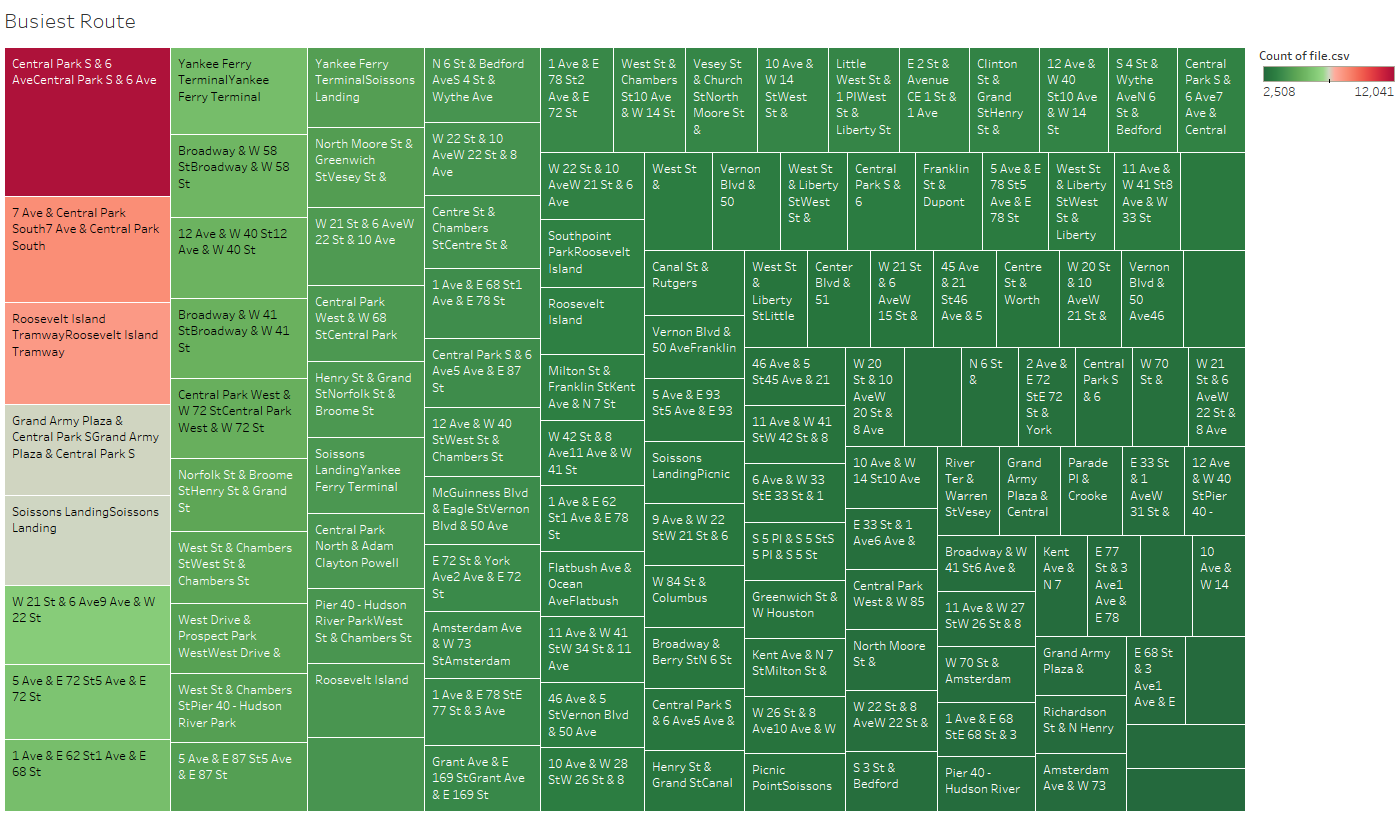
## **3.1.1 Busiest Route Analysis**

In this visualization, we focused on identifying the busiest routes utilized within the Citi Bike program. The graphic is organized in a grid format, highlighting the most frequently traveled routes between pairs of stations across New York City. Each cell represents a specific start and end station, less greener cells indicating higher traffic volumes and darker cells indicating lower traffic.

The routes from Central Park South to Grand Army Plaza and from Broadway & W 58 St to 6 Ave are among the most traveled, suggesting that areas with high recreational or commercial attractions generate significant bike traffic. This heatmap helps in visually identifying patterns of movement and popular cycling corridors in the city.

We can see those routes involving major parks, tourist attractions, and transit hubs tend to be busier. For example, the route from Roosevelt Island Tramway to Roosevelt Island shows considerable activity, possibly due to the unique appeal of the tramway and the limited transportation options on the island.

The counts on the right side of the visualization quantify the total trips recorded for the dataset, providing a numeric reference to gauge the usage scale. We plotted this heat map based on stations which covers at least 2508 trips over the year of 2022. This analysis is crucial for understanding which areas require more bikes and better infrastructure to accommodate users effectively.

Figure 1 Busiest Route

## **3.1.2 Monthly Trip Volume Analysis**

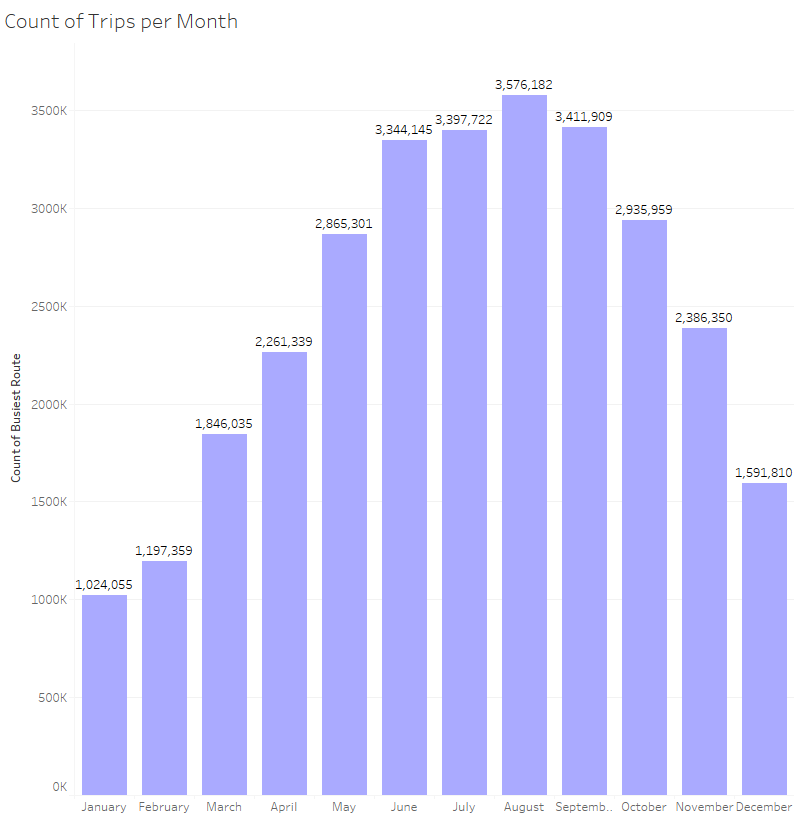


Figure 2 Count of Trips over the Months in 2022

Figure 2 bar chart titled "Count of Trips per Month" provides a clear overview of the monthly usage trends of the Citi Bike program throughout 2022. Each bar represents the total number of trips taken in a specific month, and the data is labeled directly on the bars for easy reading.

From the chart, it is evident that the bike usage peaks during the warmer months, with August recording the highest usage at approximately 3.6 million trips. The chart shows a gradual increase in trip volume starting from March, with a steady climb through August, followed by a decline as temperatures begin to drop in the fall and winter months.

The lowest number of trips occurred in January, with jt over 1 million trips, which can be attributed to the colder weather, which is less conducive to biking. Overall, this visualization helps us understand the cyclical nature of bike-sharing usage, which appears to be heavily influenced by seasonal weather changes.

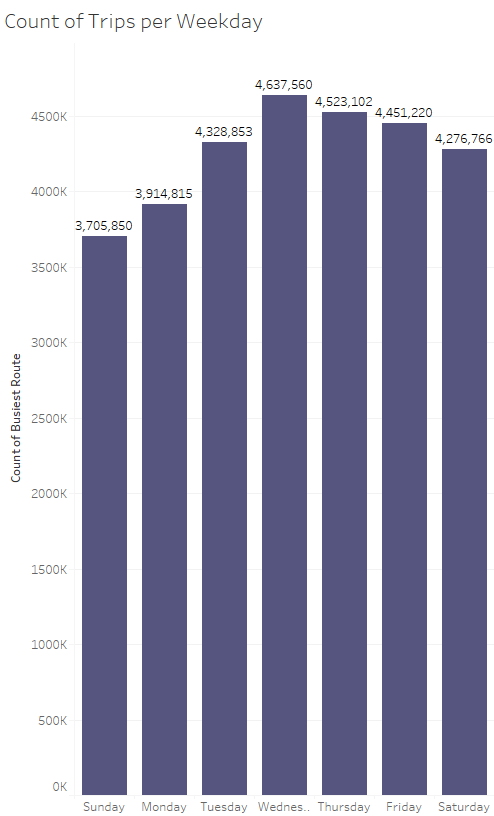


Figure 3 Count of Trips over the Week in 2022

### **3.1.3 Weekly Trip Volume Analysis**

This bar chart, showcases the number of trips taken on each day of the week throughout 2022, starting with Monday. The visualization highlights how bike usage varies from Monday to Sunday, giving us insights into weekly usage patterns within the Citi Bike program.

From the chart:

* Monday starts the week with about 3.9 million trips, showing substantial usage as many people begin their work week.
* Tuesday sees a slight increase in usage, with trips climbing to just over 4.3 million.
* Wednesday marks the peak of the week with the highest number of trips at nearly 4.6 million, suggesting it's the busiest day for Citi Bike.
* The trend continues with a high volume on Thursday, though there's a slight drop to approximately 4.5 million trips.
* Friday also shows robust activity but sees a further slight dip to about 4.4 million trips as the workweek winds down.
* Saturday shows reduced activity compared to weekdays but remains relatively busy with around 4.3 million trips, indicating that many still use the service for leisure activities during the weekend.
* Sunday shows the least activity with approximately 3.7 million trips, which could be due to a quieter city pace and possibly fewer commuting needs.

The data indicates that Citi Bike is primarily used during the weekdays, especially for commuting purposes, with a noticeable decrease during the weekend.

**3.1.4 Hourly and Monthly Usage Analysis**

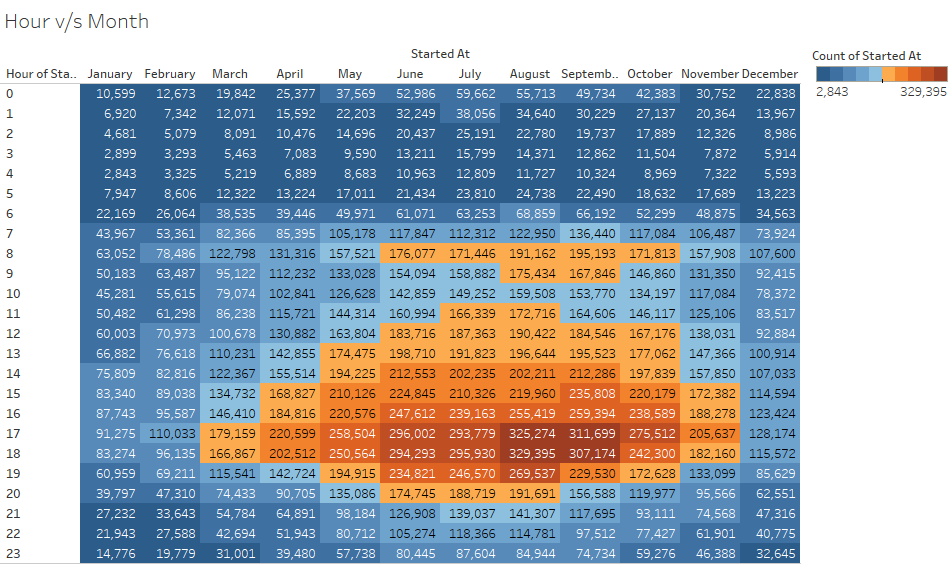


Figure 4 Trips Hour v/s Month

This heat map represents the number of Citi Bike trips started at each hour throughout each month of the year. The table layout helps to easily compare the variations in trip counts across different times and months.

* Early Morning Hours (0 to 5 AM): There's generally lower activity during these hours across all months. However, there is a slight increase during the warmer months, with June and July showing the highest counts, especially at 5 AM, indicating some early risers or late-night travelers.
* Morning to Afternoon (6 AM to 4 PM): Starting at 6 AM, there's a significant increase in trips, peaking between 8 AM and 9 AM, which suggests a heavy use of bikes for morning commutes. This pattern remains consistently high during the warmer months from May through October, indicating regular usage for commuting or midday outings.
* Evening Peak (5 PM to 7 PM): This time frame shows the highest usage across all months, with the peak at 6 PM. The numbers rise remarkably during the warmer months, with the highest counts in June, July, and August, reflecting heavy commuting in the evenings or leisure activities after work hours.
* Late Evening (8 PM onwards): Post 8 PM, there's a gradual decline in trips, but usage remains higher than the early morning hours. During summer, the counts remain relatively higher even late into the evening, suggesting that people continue to engage in recreational activities or utilize the bikes for late commutes.

**3.1.5 Daily and Hourly Usage Trends**

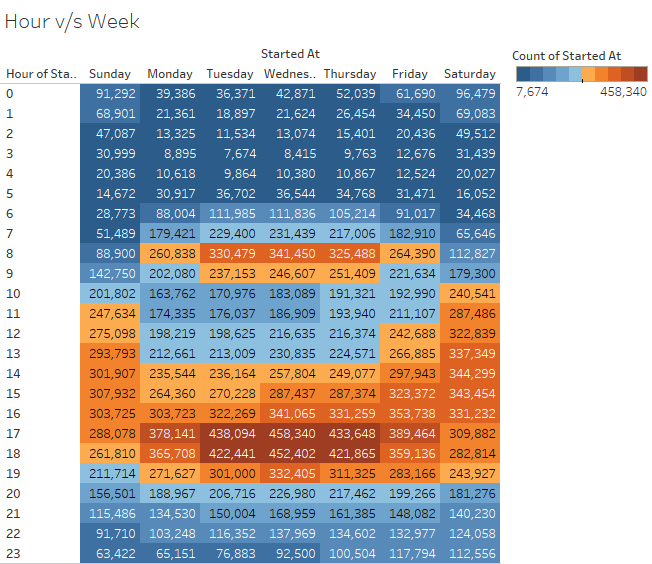


Figure 5 Trips Hour v/s Week

Figure 5 shows the number of Citi Bike trips that started each hour on each day of the week throughout the year. Each cell in the heat map has a number that tells us how many trips started at a certain time on a certain day.

* Early Morning (Midnight to 5 AM): Not many people use bikes during these hours. The numbers are lowest early in the morning but start to increase as the morning goes on.
* Morning (6 AM to 9 AM): There is a big increase in trips starting at 6 AM, especially on weekdays. It looks like a lot of people use Citi Bike to get to work or school, especially around 8 AM.
* Midday (10 AM to 3 PM): The number of trips stays pretty high during these hours. This might be when people are going out for lunch or running errands, especially on Saturdays when we see a lot of trips all day.
* Afternoon and Evening (4 PM to 7 PM): This is another busy time for Citi Bike, with the most trips starting around 5 PM to 6 PM. It’s likely that people are using the bikes to head home from work or go out in the evening.
* Night (8 PM onwards): The number of trips drops off after 8 PM, but there are still quite a few people using the bikes, possibly for getting home after a night out or for late-night events.

This helps us see when Citi Bikes are most often used during the day and week, showing us that weekdays, especially morning and late afternoon, are the busiest times.

**3.1.6 Analysis of Stations with more than 20,000 Trips**

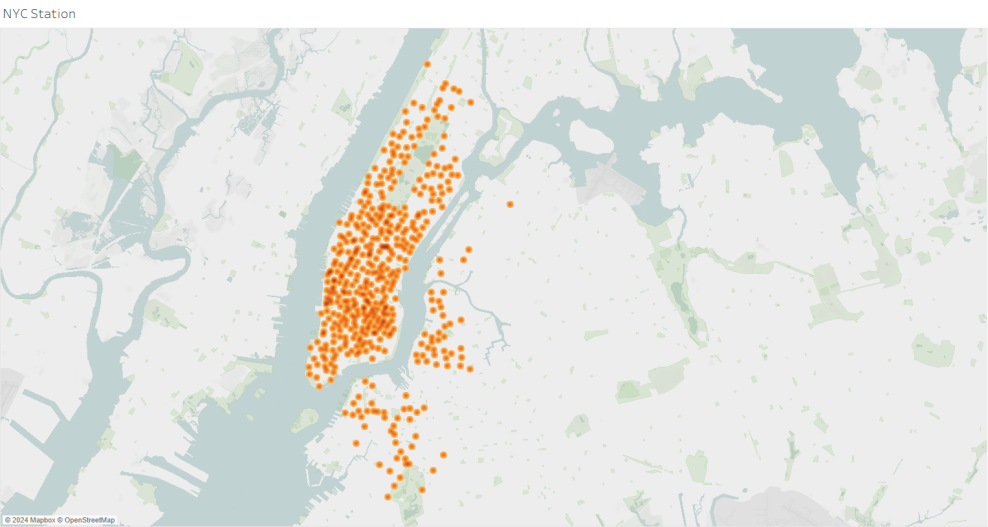


Figure 6 Top stations with over 20,000 trips

Here the map shows different bike stations found in New York City for bike stations that have had more than 20000 trips in the last one year. It underlines the zones that are heavily used by bikes. This information can be useful in the future for redevelopment and decide with more certainty whether it is necessary to increase resources in these sectors. Furthermore, from the graph we know the city's areas where the people of New York use bikes most frequently are revealed frequently.

**3.1.7 Analysis of Peak Hours for Trips**

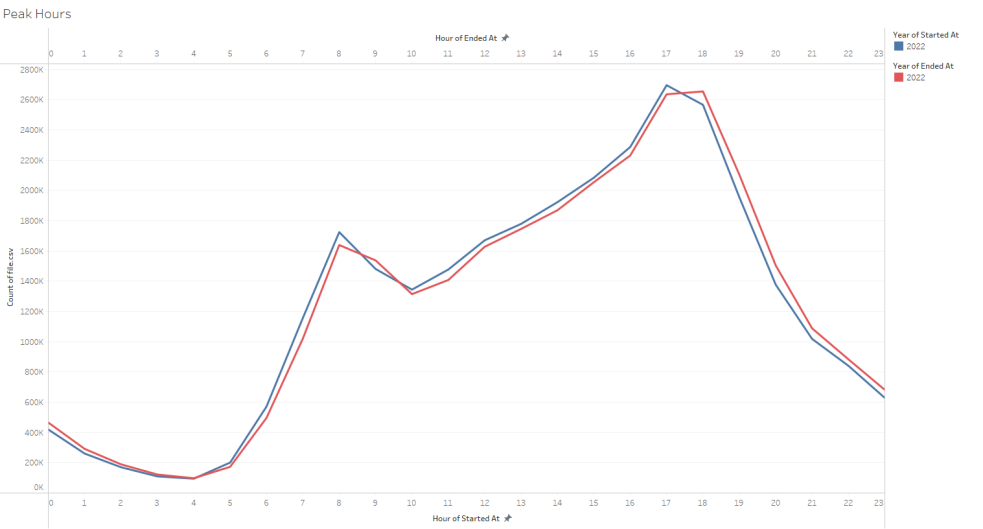


Figure 7 Peak Hours of Trips Taken

The dual line graph illustrates the number of trips taken across different hours of the day throughout 2022. The graph shows two lines, one for when trips start (blue line) and one for when trips end (red line), providing a clear visual representation of biking activity during the day.

* Early Morning (Midnight to 5 AM): Both lines are at their lowest, indicating very few trips start or end during these hours.
* Morning Commute (6 AM to 9 AM): There is a sharp rise in trips starting around 6 AM, peaking between 8 AM and 9 AM. This corresponds to people using Citi Bike to commute to work or school.
* Midday (10 AM to 3 PM): The numbers dip slightly after the morning peak but remain steady, suggesting that bikes are still frequently used for midday errands or short trips.
* Evening Commute (4 PM to 7 PM): There is another significant peak from 4 PM to 6 PM, with the highest number of trips around 5 PM and 6 PM, indicating that many are using bikes to return home or go to evening activities.
* Night (8 PM to 11 PM): The numbers decline rapidly after 7 PM, with a quieter period until the cycle starts again.

This graph helps us understand that the busiest times for Citi Bike are during typical commuting hours in the morning and late afternoon. The similarity in the trends of both starting and ending trips confirms that the bikes are used mostly for commuting purposes during these peak hours.

**3.1.8 Analysis of Rideable Type Usage Over the Months**

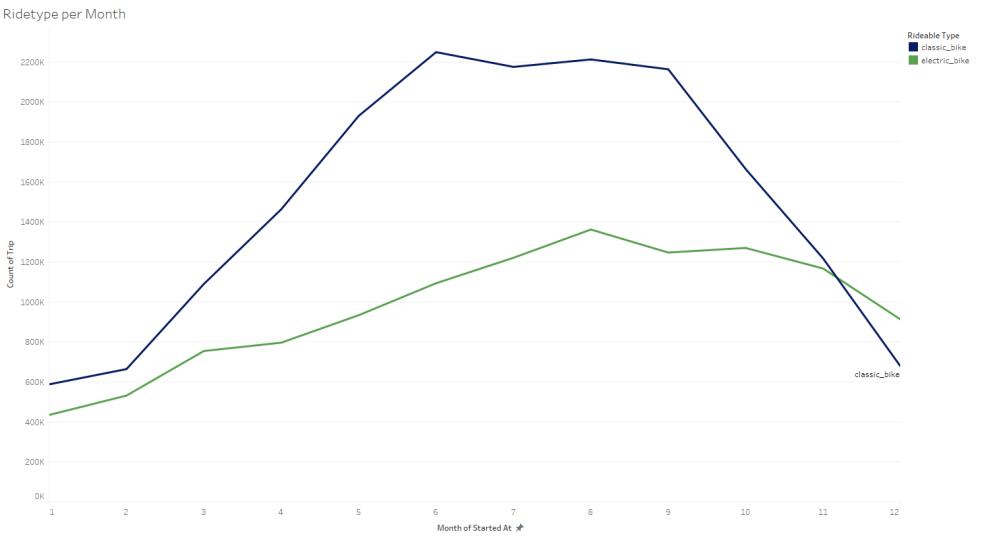


Figure 8 Rideable Type used over the Months

Figure 8 dual line graph shows the monthly usage patterns of two types of Citi Bike rideables; classic bikes and electric bikes throughout the year 2022.

* Early Year Trends (January to March): Both lines start relatively low in January. The usage of classic bikes (Blue Line) and electric bikes (Green Line) begins to increase as the months progress into spring, with classic bikes seeing a more pronounced rise.
* Spring to Summer Peak (April to July): The usage of classic bikes surges significantly during these months, peaking in July. This indicates that classic bikes are particularly favored during the warmer months. Electric bikes also see an increase but at a steadier, more gradual pace, peaking in August.
* Late Summer to Early Fall (August to September): While classic bikes begin a sharp decline after July, electric bikes maintain a plateau through August before starting to decline in September, suggesting that they might be preferred for their ease of use during the hotter days of late summer.
* Fall to Winter Drop-off (October to December): The usage of both bike types sharply decreases as the weather cools down, with classic bikes showing a steeper decline compared to electric bikes, which decrease gradually.

This graph provides insights into seasonal preferences for bike types. Classic bikes dominate during the peak of summer, due to leisure rides and the sheer enjoyment of biking in pleasant weather, while electric bikes show sustained use slightly longer into the year, due to their appeal among people who prefer assistance with pedaling in varying temperatures.

**3.1.9 Analysis of Rideable type by Different Memberships.**

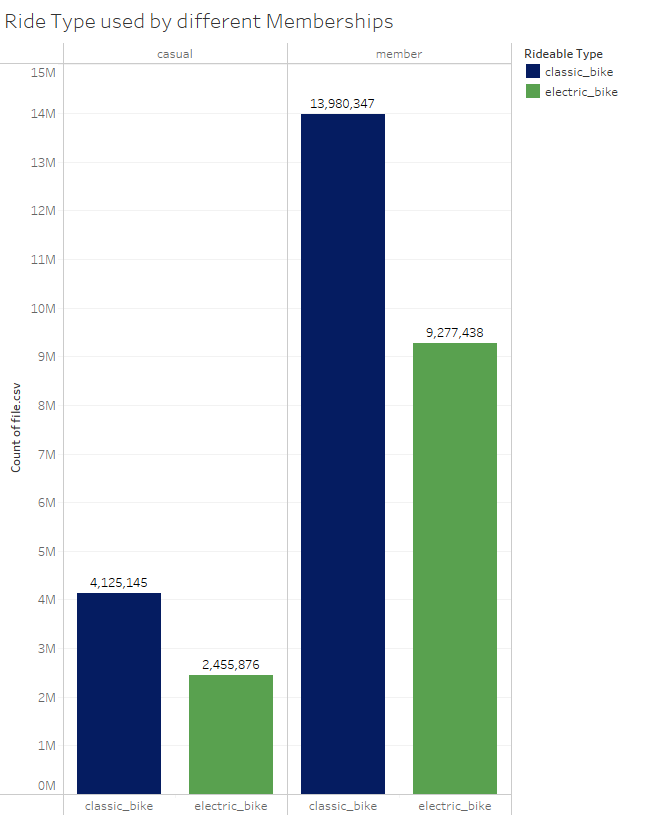


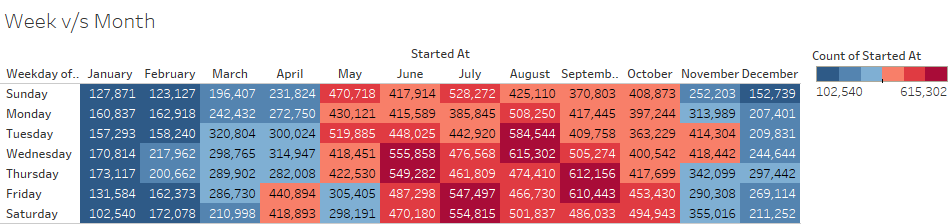
Figure 9 Rideable type used by Different Memberships

The bar chart clearly illustrates the preferences for classic and electric bikes among different Citi Bike membership types; casual and member.

* Classic Bikes:
  + Casual Users: Over 4 million casual users chose classic bikes, showing a significant preference for this bike type among non-members.
  + Members: Nearly 14 million members used classic bikes, which indicates a strong preference for classic bikes among regular users of the service.
* Electric Bikes:
  + Casual Users: Approximately 2.5 million casual users opted for electric bikes, which is less than the number who chose classic bikes but still notable.
  + Members: More than 9 million members used electric bikes, showing that a substantial number of members appreciate the benefits of electric bikes.

Classic Bikes are more popular than electric bikes across both user groups, with members showing a particularly strong preference for them. Whereas Electric Bikes are less popular but still chosen by millions of riders, suggesting they are valued for specific uses or by those who prefer a less strenuous ride. Both groups show a stronger inclination towards classic bikes.

**3.1.10 Analysis of Trip Frequency by Weekday and Month**

Figure 10 Trips Week v/s Months

The heatmap provides a comprehensive overview of the number of Citi Bike trips started on each day of the week for every month throughout the year.

* Weekday Trends:
  + Weekdays (Monday to Friday): There is a consistent increase in trip counts from January to August, with the numbers peaking in August for most weekdays. This suggests high usage during typical workdays throughout the warmer months, possibly due to commuting or midday trips in city areas.
  + Weekends (Saturday and Sunday): Saturday trips also follow a similar trend, albeit with slightly lower counts than weekdays. Sundays typically have the lowest trip counts, which might indicate a more relaxed pace or less demand for commuting purposes.
* Monthly Patterns:
  + January and February: These months show the lowest activity, which gradually increases as the weather improves.
  + March to May: A significant rise in trips indicates the onset of warmer weather, making bike rides more appealing.
  + June to August: These are the peak months with the highest trip frequencies, reflecting the most favorable weather conditions for biking.
  + September to December: A gradual decrease in trips is observed as the weather cools, with December showing a sharp decline, likely due to cold weather impacting the desire or ability to bike.

The heatmap clearly shows that Citi Bike usage is seasonally affected, with peak usage in the summer months and lower engagement during the colder months. Weekday trips outnumber weekend trips, which could be attributed to commuting habits during workdays.

**3.2 Python**

**3.2.1 T-Test Analysis: Ride Duration vs. User Type**



Figure 11 Test Output

The T-test comparing ride durations between members and casual users yielded a T-statistic of -79.49 and a P-value of 0.0.

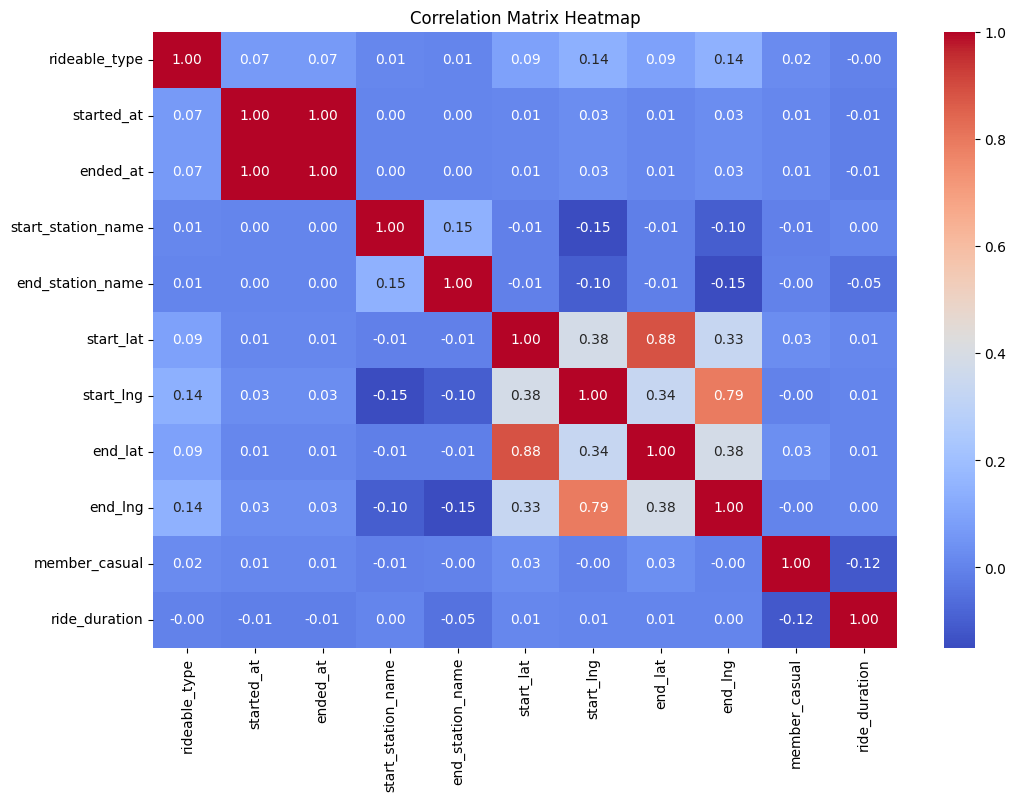
Key Findings:

* Significant Difference: The extremely low P-value indicates a statistically significant difference in ride durations between the two user groups.
* Negative T-statistic: The negative value of the T-statistic suggests that members generally have shorter ride durations compared to casual users.

This analysis clearly shows that members tend to use bikes for quicker trips, while casual users tend to ride for longer periods.

**3.2.2 Correlation Analysis:**

The strength of the relationship between variables and the associated direction is measured by correlation. Connection values typically range from -1 to +1, with -1 indicating significant negative bonding between two variables, +1 indicating strong positive bonding between two variables, and 0 indicating no correlation. The correlation's significance is determined by the p-value. When a correlation's p-value is less than the level of significance (alpha), the correlation is significant.

Figure 12 Heat Map – Correlation

Here the map shows different bike stations found in New York City for bike station that has more than 20000 trips in the last one year. It underlines the zones that are heavily used with the bikes. This information can be useful in the future for redevelopment and decide with more certainty whether it is necessary to increase resources in these sectors. Further, before proceeding with the construction of new bike lanes, it reveals the areas of the city that people of New York use bikes most frequently.

**3.2.2.1 Ride Duration vs. Time of Day**

The analysis of the relationship between ride duration and the time of day yielded a correlation coefficient of 0.0011 and a P-value of 4.96e-10.



Figure 13 Correlation Ride Duration vs Time of Day Output

Key Findings:

* Very Weak Positive Correlation: The correlation coefficient is very close to zero, indicating that there is almost no relationship between ride duration and the time of day. However, the positive sign suggests that ride durations slightly increase as the day progresses.
* Statistical Significance: Despite the correlation being very weak, the extremely low P-value indicates that the relationship is statistically significant, though the actual effect on ride duration is minimal.

This result implies that while there may be a slight increase in ride durations later in the day, the effect is so small that it likely has little practical impact on how bike rides are distributed through the day.

**3.2.2.2 Ride Duration vs. Hour**

The correlation matrix reveals the relationship between ride duration and the hour of the day, with a correlation coefficient of 0.0011.

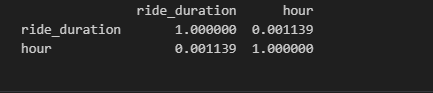


Figure 14 Correlation Ride Duration vs Hour of Day Output

We see that the correlation coefficient is very small, indicating a weak relationship between ride duration and the hour of the day. The positive value suggests that ride durations increase slightly as the day progresses, but the effect is minimal and likely of little practical significance.

This analysis indicates that the time of day has almost no impact on how long rides last, with only a negligible increase in duration later in the day.