Let's start with some basic data analysis tasks:

1. \*\*Summary Statistics\*\*: Compute summary statistics for numerical columns.

2. \*\*Trends Over Time\*\*: Analyze the number of rides over time (e.g., daily, weekly).

3. \*\*Ride Duration\*\*: Calculate ride duration and analyze its distribution.

4. \*\*Popular Stations\*\*: Identify the most popular start and end stations.

5. \*\*Member vs. Casual Riders\*\*: Compare ride patterns between members and casual riders.

### Summary Statistics

Let's first compute the summary statistics for the numerical columns.

### Summary Statistics

- \*\*Latitude and Longitude\*\*:

- The average starting latitude is 41.91, and the average starting longitude is -87.65.

- The average ending latitude is 41.91, and the average ending longitude is -87.65.

- There are some zero values in the `end\_lat` and `end\_lng`, which might indicate rides that ended outside the usual area or data errors.

### Trends Over Time

Let's analyze the number of rides over time. We'll start by looking at daily trends.

### Daily Number of Rides

The plot above shows the daily number of rides for August 2023. The number of rides fluctuates, with some days having significantly more rides than others.

### Ride Duration

Next, let's calculate the ride duration and analyze its distribution.

#### Calculate Ride Duration

We'll compute the duration of each ride in minutes.

### Ride Duration

- \*\*Summary Statistics\*\*:

- Mean ride duration: 16.72 minutes

- Median ride duration: 10.65 minutes

- The ride durations range from -49.53 minutes to 6891.22 minutes, indicating some possible data errors.

- \*\*Distribution\*\*:

- The histogram shows that most rides are under 30 minutes, with a significant number clustered around 5 to 20 minutes.

### Popular Stations

Let's identify the most popular start and end stations.

#### Most Popular Start Stations

### Most Popular Stations

- \*\*Start Stations\*\*:

- The most popular start station is "Unknown" with 118,919 rides, indicating missing data.

- The next most popular start stations are:

- Streeter Dr & Grand Ave: 10,721 rides

- DuSable Lake Shore Dr & North Blvd: 6,697 rides

- DuSable Lake Shore Dr & Monroe St: 6,429 rides

- Michigan Ave & Oak St: 6,134 rides

- \*\*End Stations\*\*:

- The most popular end station is also "Unknown" with 124,352 rides.

- The next most popular end stations are:

- Streeter Dr & Grand Ave: 10,772 rides

- DuSable Lake Shore Dr & North Blvd: 7,777 rides

- Michigan Ave & Oak St: 6,241 rides

- DuSable Lake Shore Dr & Monroe St: 6,044 rides

### Member vs. Casual Riders

Let's compare ride patterns between members and casual riders.

#### Ride Patterns: Members vs. Casual Riders

### Ride Patterns: Members vs. Casual Riders

- \*\*Summary Statistics for Ride Duration\*\*:

- \*\*Casual Riders\*\*:

- Mean ride duration: 21.96 minutes

- Median ride duration: 12.85 minutes

- 75th percentile: 23.28 minutes

- Standard deviation: 49.46 minutes

- Maximum ride duration: 6891.22 minutes

- \*\*Members\*\*:

- Mean ride duration: 13.19 minutes

- Median ride duration: 9.45 minutes

- 75th percentile: 16.03 minutes

- Standard deviation: 23.91 minutes

- Maximum ride duration: 1499.90 minutes

- \*\*Number of Rides\*\*:

- Members: 460,381 rides

- Casual Riders: 310,055 rides

Members tend to have shorter ride durations compared to casual riders. This might indicate that members use the bikes for shorter, more routine trips, while casual riders might use them for longer, more leisurely rides.

Would you like to explore any other specific analyses or visualizations?

The first few rows of the `202308-divvy-tripdata.csv` file are as follows:

| ride\_id | rideable\_type | started\_at | ended\_at | start\_station\_name | start\_station\_id | end\_station\_name | end\_station\_id | start\_lat | start\_lng | end\_lat | end\_lng | member\_casual |

|-----------------|---------------|---------------------|---------------------|----------------------|------------------|--------------------|------------------|-----------|------------|------------|------------|---------------|

| 903C30C2D810A53B| electric\_bike | 2023-08-19 15:41:53 | 2023-08-19 15:53:36 | LaSalle St & Illinois St | 13430 | Clark St & Elm St | TA1307000039 | 41.890721 | -87.631477 | 41.902973 | -87.63128 | member |

| F2FB18A98E110A2B| electric\_bike | 2023-08-18 15:30:18 | 2023-08-18 15:45:25 | Clark St & Randolph St | TA1305000030 | NaN | NaN | 41.884511 | -87.631550 | 41.930000 | -87.64000 | member |

| D0DEC7C94E4663DA| electric\_bike | 2023-08-30 16:15:08 | 2023-08-30 16:27:37 | Clark St & Randolph St | TA1305000030 | NaN | NaN | 41.884981 | -87.630793 | 41.910000 | -87.63000 | member |

| E0DDDC5F84747ED9| electric\_bike | 2023-08-30 16:24:07 | 2023-08-30 16:33:34 | Wells St & Elm St | KA1504000135 | NaN | NaN | 41.903105 | -87.634667 | 41.900000 | -87.62000 | member |

| 7797A4874BA260CA| electric\_bike | 2023-08-22 15:59:44 | 2023-08-22 16:20:38 | Clark St & Randolph St | TA1305000030 | NaN | NaN | 41.885548 | -87.632019 | 41.890000 | -87.68000 | member |

The data includes the following columns:

- `ride\_id`: Unique identifier for the ride

- `rideable\_type`: Type of bike (e.g., electric\_bike)

- `started\_at`: Start time of the ride

- `ended\_at`: End time of the ride

- `start\_station\_name`: Name of the start station

- `start\_station\_id`: Identifier for the start station

- `end\_station\_name`: Name of the end station

- `end\_station\_id`: Identifier for the end station

- `start\_lat`: Latitude of the start location

- `start\_lng`: Longitude of the start location

- `end\_lat`: Latitude of the end location

- `end\_lng`: Longitude of the end location

- `member\_casual`: Type of rider (member or casual)

If you have any specific analysis or tasks you'd like to perform on this data, please let me know!

### Missing Values

- `start\_station\_name`: 118,919 missing values

- `start\_station\_id`: 118,919 missing values

- `end\_station\_name`: 125,568 missing values

- `end\_station\_id`: 125,568 missing values

- `end\_lat`: 1,257 missing values

- `end\_lng`: 1,257 missing values

### Data Types

- `ride\_id`: object

- `rideable\_type`: object

- `started\_at`: object

- `ended\_at`: object

- `start\_station\_name`: object

- `start\_station\_id`: object

- `end\_station\_name`: object

- `end\_station\_id`: object

- `start\_lat`: float64

- `start\_lng`: float64

- `end\_lat`: float64

- `end\_lng`: float64

- `member\_casual`: object

### Steps for Data Cleaning

1. \*\*Convert `started\_at` and `ended\_at` to datetime\*\*.

2. \*\*Handle missing values\*\*:

- For latitude and longitude, drop rows with missing values.

- For station names and IDs, consider the context (e.g., possibly replacing with "Unknown" or dropping if necessary).

Let's start by converting the `started\_at` and `ended\_at` columns to datetime and then handle the missing values.

#### Convert `started\_at` and `ended\_at` to datetime

The `started\_at` and `ended\_at` columns have been successfully converted to datetime.

#### Handling Missing Values

1. \*\*Latitude and Longitude\*\*:

- Drop rows with missing values for `end\_lat` and `end\_lng`.

2. \*\*Station Names and IDs\*\*:

- For missing `start\_station\_name`, `start\_station\_id`, `end\_station\_name`, and `end\_station\_id`, fill with "Unknown".

Let's proceed with these steps.

The data has been successfully cleaned. There are no missing values in the dataset now.

Would you like to proceed with any specific analysis or further exploration of the data?

Here are some visualizations we can create:

1. \*\*Daily Number of Rides\*\*: Line chart to show trends over time.

2. \*\*Ride Duration Distribution\*\*: Histogram to show the distribution of ride durations.

3. \*\*Popular Stations\*\*: Bar charts to show the most popular start and end stations.

4. \*\*Member vs. Casual Riders\*\*: Side-by-side box plots to compare ride durations.

Let's start with these visualizations.

### Daily Number of Rides

We'll create a line chart to show the number of rides per day.

### Daily Number of Rides

The line chart above shows the daily number of rides for August 2023, highlighting the trends and fluctuations over the month.

### Ride Duration Distribution

Next, let's create a histogram to visualize the distribution of ride durations.

### Distribution of Ride Durations

The histogram shows that most rides are under 30 minutes, with a significant number clustered around 5 to 20 minutes. There are a few outliers with very long durations.

### Popular Stations

Let's create bar charts to visualize the most popular start and end stations.

#### Most Popular Start Stations

### Top 10 Most Popular Start Stations

The bar chart above displays the top 10 most popular start stations. "Streeter Dr & Grand Ave" is the most popular start station after accounting for missing data labeled as "Unknown."

#### Most Popular End Stations

Next, let's create a bar chart for the most popular end stations.

### Top 10 Most Popular End Stations

The bar chart above shows the top 10 most popular end stations. "Streeter Dr & Grand Ave" is the most popular end station after excluding the "Unknown" category.

### Member vs. Casual Riders

Finally, let's create side-by-side box plots to compare ride durations between members and casual riders.

### Ride Duration: Members vs. Casual Riders

The box plot above shows the ride durations for members and casual riders:

- \*\*Members\*\* tend to have shorter ride durations with less variability.

- \*\*Casual Riders\*\* have longer ride durations with more variability and some significant outliers.

These visualizations provide insights into the usage patterns and behaviors of Divvy bike users.

If you have any specific analyses or additional visualizations you'd like to see, please let me know!

Here are some visualizations we can create:

1. \*\*Daily Number of Rides\*\*: Line chart to show trends over time.

2. \*\*Ride Duration Distribution\*\*: Histogram to show the distribution of ride durations.

3. \*\*Popular Stations\*\*: Bar charts to show the most popular start and end stations.

4. \*\*Member vs. Casual Riders\*\*: Side-by-side box plots to compare ride durations.

Let's start with these visualizations.

### Daily Number of Rides

We'll create a line chart to show the number of rides per day.

### Daily Number of Rides

The line chart above shows the daily number of rides for August 2023, highlighting the trends and fluctuations over the month.

### Ride Duration Distribution

Next, let's create a histogram to visualize the distribution of ride durations.

### Distribution of Ride Durations

The histogram shows that most rides are under 30 minutes, with a significant number clustered around 5 to 20 minutes. There are a few outliers with very long durations.

### Popular Stations

Let's create bar charts to visualize the most popular start and end stations.

#### Most Popular Start Stations

### Top 10 Most Popular Start Stations

The bar chart above displays the top 10 most popular start stations. "Streeter Dr & Grand Ave" is the most popular start station after accounting for missing data labeled as "Unknown."

#### Most Popular End Stations

Next, let's create a bar chart for the most popular end stations.

### Top 10 Most Popular End Stations

The bar chart above shows the top 10 most popular end stations. "Streeter Dr & Grand Ave" is the most popular end station after excluding the "Unknown" category.

### Member vs. Casual Riders

Finally, let's create side-by-side box plots to compare ride durations between members and casual riders.

### Ride Duration: Members vs. Casual Riders

The box plot above shows the ride durations for members and casual riders:

- \*\*Members\*\* tend to have shorter ride durations with less variability.

- \*\*Casual Riders\*\* have longer ride durations with more variability and some significant outliers.

These visualizations provide insights into the usage patterns and behaviors of Divvy bike users.

If you have any specific analyses or additional visualizations you'd like to see, please let me know!