The dataset contains missing values in the following columns:

1. `start\_station\_name` - 122,943 missing values

2. `start\_station\_id` - 122,943 missing values

3. `end\_station\_name` - 130,304 missing values

4. `end\_station\_id` - 130,304 missing values

5. `end\_lat` - 1,254 missing values

6. `end\_lng` - 1,254 missing values

### Data Cleaning Steps

1. \*\*Handling Missing Values:\*\*

- For `start\_station\_name` and `start\_station\_id`, and `end\_station\_name` and `end\_station\_id`, we could either fill them with a placeholder indicating unknown locations or drop the rows if they are not crucial.

- For `end\_lat` and `end\_lng`, we could similarly fill them with a placeholder or drop the rows.

2. \*\*Incorrect Data Types:\*\*

- Convert `started\_at` and `ended\_at` columns to datetime objects for proper time-series analysis.

Let's proceed with these steps.

#### Step 1: Handle Missing Values

We will fill missing values for station names and IDs with "Unknown", and for latitudes and longitudes with `0` indicating an unknown location.

#### Step 2: Convert Date Columns to Datetime

We will convert the `started\_at` and `ended\_at` columns to `datetime`.

Let's apply these cleaning steps.

The data cleaning process is complete. Here are the main changes applied:

1. \*\*Missing Values:\*\*

- Filled `start\_station\_name` and `start\_station\_id` with "Unknown".

- Filled `end\_station\_name` and `end\_station\_id` with "Unknown".

- Filled `end\_lat` and `end\_lng` with `0`.

2. \*\*Data Type Conversion:\*\*

- Converted `started\_at` and `ended\_at` columns to datetime format.

The updated data types and a preview of the first few rows are shown below:

```plaintext

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 767650 entries, 0 to 767649

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ride\_id 767650 non-null object

1 rideable\_type 767650 non-null object

2 started\_at 767650 non-null datetime64[ns]

3 ended\_at 767650 non-null datetime64[ns]

4 start\_station\_name 767650 non-null object

5 start\_station\_id 767650 non-null object

6 end\_station\_name 767650 non-null object

7 end\_station\_id 767650 non-null object

8 start\_lat 767650 non-null float64

9 start\_lng 767650 non-null float64

10 end\_lat 767650 non-null float64

11 end\_lng 767650 non-null float64

12 member\_casual 767650 non-null object

dtypes: datetime64 , float64(4), object(7)

memory usage: 76.1+ MB

```

| 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 |

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

| 9340B064F0AEE130 | electric\_bike | 2023-07-23 20:06:14 | 2023-07-23 20:22:44 | Kedzie Ave & 110th St | 20204 | Public Rack - Racine Ave & 109th Pl | 877 | 41.692406 | -87.700905 | 41.694835 | -87.653041 | member |

| D1460EE3CE0D8AF8 | classic\_bike | 2023-07-23 17:05:07 | 2023-07-23 17:18:37 | Western Ave & Walton St | KA1504000103 | Milwaukee Ave & Grand Ave | 13033 | 41.898418 | -87.686596 | 41.891578 | -87.648384 | member |

| DF41BE31B895A25E | classic\_bike | 2023-07-23 10:14:53 | 2023-07-23 10:24:29 | Western Ave & Walton St | KA1504000103 | Damen Ave & Pierce Ave | TA1305000041 | 41.898418 | -87.686596 | 41.909396 | -87.677692 | member |

| 9624A293749EF703 | electric\_bike | 2023-07-21 08:27:44 | 2023-07-21 08:32:40 | Racine Ave & Randolph St | 13155 | Clinton St & Madison St | TA1305000032 | 41.884112 | -87.656943 | 41.882752 | -87.641190 | member |

| 2F68A6A4CDB4C99A | classic\_bike | 2023-07-08 15:46:42 | 2023-07-08 15:58:08 | Clark St & Leland Ave | TA1309000014 | Montrose Harbor | TA1308000012 | 41.967088 | -87.667291 | 41.963982 | -87.638181 | member |

Would you like to proceed with any specific analysis or further operations on this cleaned data?

Let's start with a few basic analyses on the cleaned data, including summary statistics and trends over time.

### Summary Statistics

We can begin with summary statistics for numerical columns and a count of unique values for categorical columns.

### Trends Over Time

We can analyze trends over time such as:

- Number of rides per day.

- Average duration of rides per day.

- Comparison of usage between members and casual riders.

Let's proceed with these analyses.

#### Summary Statistics

1. Basic descriptive statistics for numerical columns.

2. Count of unique values for categorical columns.

#### Trends Over Time

1. Number of rides per day.

2. Average ride duration per day.

3. Comparison of usage between members and casual riders.

Let's start by calculating and displaying these metrics.

### Summary Statistics

\*\*Numerical Columns:\*\*

| Statistic | start\_lat | start\_lng | end\_lat | end\_lng |

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

| count | 767650 | 767650 | 767650 | 767650 |

| mean | 41.905647 | -87.646413| 41.837598| -87.503467|

| std | 0.042675 | 0.026904 | 1.692887 | 3.539650 |

| min | 41.648501 | -87.920000| 0.000000 | -87.920000|

| 25% | 41.883073 | -87.660000| 41.883165| -87.660000|

| 50% | 41.901087 | -87.642962| 41.902924| -87.643118|

| 75% | 41.931931 | -87.629544| 41.932418| -87.629544|

| max | 42.070000 | -87.528232| 42.180000| 0.000000 |

\*\*Categorical Columns:\*\*

| Column | Unique Values |

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

| ride\_id | 767650 |

| rideable\_type | 3 |

| start\_station\_name | 1254 |

| start\_station\_id | 1224 |

| end\_station\_name | 1251 |

| end\_station\_id | 1222 |

| member\_casual | 2 |

### Trends Over Time

\*\*Number of Rides Per Day:\*\*

| Date | Number of Rides |

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

| 2023-07-01 | 22802 |

| 2023-07-02 | 6162 |

| 2023-07-03 | 26860 |

| 2023-07-04 | 28234 |

| 2023-07-05 | 17029 |

\*\*Average Ride Duration Per Day (in minutes):\*\*

| Date | Average Duration |

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

| 2023-07-01 | 29.75 |

| 2023-07-02 | 23.62 |

| 2023-07-03 | 27.27 |

| 2023-07-04 | 30.62 |

| 2023-07-05 | 24.50 |

\*\*Usage Comparison Between Members and Casual Riders:\*\*

| Rider Type | Count |

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

| Member | 436292 |

| Casual | 331358 |

### Next Steps

- \*\*Visualize trends\*\*: We can create plots to visualize the trends in the number of rides per day and average ride duration.

- \*\*Further analysis\*\*: Perform deeper analysis on ride patterns, geographical trends, and differences between member and casual riders.

Would you like to proceed with visualizations or any specific analysis?

Here are the visualizations for the data:

1. \*\*Number of Rides Per Day:\*\*

- This line plot shows the daily variation in the number of rides. Peaks and troughs indicate days with higher and lower ride counts.

2. \*\*Average Ride Duration Per Day (in minutes):\*\*

- This line plot illustrates the daily average ride duration. It helps identify days when rides were generally longer or shorter.

3. \*\*Usage Comparison Between Members and Casual Riders:\*\*

- This bar chart compares the total number of rides taken by members and casual riders. It highlights the difference in usage patterns between the two groups.

If you need further analysis or additional visualizations, please let me know!