Thank you for providing the data in CSV format, which is typically more efficient to read than Excel. I'll start by loading the CSV file and then provide a descriptive summary of its contents.

Finished working

Show work

Here's a descriptive summary of the dataset:

1. **ride\_id**: Unique identifiers for rides with a total of 274,571 unique entries.
2. **rideable\_type**: Types of bikes used for the rides. There are 3 unique types, with the 'classic\_bike' being the most frequent.
3. **started\_at** & **ended\_at**: Timestamps indicating the start and end times of rides. They have unique values, showing the granularity of the data.
4. **start\_station\_name** & **end\_station\_name**: Names of the start and end stations for rides. There are 801 unique starting stations and 793 unique ending stations. The most frequent station for both starting and ending is 'Kingsbury St & Kinzie St'.
5. **start\_lat**, **start\_lng**, **end\_lat**, & **end\_lng**: Latitude and longitude coordinates for the start and end stations.
6. **member\_casual**: Indicates if the rider is a member or a casual rider. The dataset has more members (206,473 entries) than casual riders.
7. **start\_date** & **end\_date**: The dates when the rides started and ended. The data spans 90 unique dates.
8. **start\_hour**: The hour when the ride started. The average starting hour is approximately 13:43 (1:43 PM).
9. **day\_of\_week**: Day of the week for the rides. There are 7 unique values (as expected) with Wednesday being the most frequent.
10. **start\_month**: The month when the ride started. There are 3 unique months, with March being the most frequent.
11. **trip\_duration**: Duration of the trips in minutes. The average trip duration is approximately 13.23 minutes, with a maximum of 854 minutes.
12. **trip\_distance**: Distance of the trips in miles. The average trip distance is approximately 1.37 miles.
13. **IsOutlier**: A binary column, possibly indicating whether a trip is considered an outlier in some context. All values are 0, indicating no outliers based on this column.

This dataset appears to contain detailed information about bike rides, including when and where they started and ended, the type of bike used, and whether the rider was a member or a casual user.

If you have specific questions or analyses you'd like to perform on this data, please let me know!



Perform detailed analysis on data

ChatGPT

Certainly! A detailed analysis can encompass a range of perspectives and techniques. Given the nature of the dataset, I'll start with the following analyses:

1. **Distribution Analysis**:
   * Rideable type distribution
   * Distribution of rides based on member status (member vs. casual)
   * Distribution of rides by day of the week
   * Distribution of ride start hours
2. **Trend Analysis**:
   * Number of rides over time
   * Average trip duration over time
   * Average trip distance over time
3. **Station Analysis**:
   * Most and least popular start and end stations
4. **Ride Characteristics**:
   * Average trip duration for members vs. casual riders
   * Average trip distance for members vs. casual riders

Let's begin with the **Distribution Analysis**.

Finished working

Show work

A group of different colored bars

Description automatically generated with medium confidence

Here's a breakdown of the distribution analyses:

1. **Distribution of Rideable Types**:
   * The most popular type of bike is the classic\_bike, followed by the docked\_bike. The electric\_bike has the fewest rides among the three types.
2. **Distribution of Rides based on Member Status**:
   * Members take significantly more rides than casual riders. This indicates that members are more active users of the bikeshare service.
3. **Distribution of Rides by Day of the Week**:
   * The number of rides is relatively consistent across weekdays, with a slight decrease on the weekends. This suggests that the bikeshare service might be popular among commuters and daily users, as opposed to just recreational users during weekends.
4. **Distribution of Ride Start Hours**:
   * There are two clear peaks in the distribution of ride start hours: one in the morning around 8 AM and another in the late afternoon around 5 PM. This further supports the idea that many users are commuting to and from work or school.
   * Additionally, there's a relatively consistent number of rides during midday hours, indicating ongoing usage throughout the day.

Next, we'll move on to the **Trend Analysis** to study the patterns over time.

Finished working

Show work

A screenshot of a graph

Description automatically generated

Here's a breakdown of the trend analyses:

1. **Number of Rides over Time**:
   * The number of rides seems to fluctuate slightly over the 3-month period. There are some peaks and troughs, which might be influenced by factors like weekdays versus weekends, weather conditions, or specific events in the area.
   * Toward the end of the period, there seems to be a noticeable increase in the number of rides. This could be due to factors like warmer weather, promotions, or other reasons that would need further investigation.
2. **Average Trip Duration over Time**:
   * The average trip duration remains relatively consistent over the period, with a few minor fluctuations. The average duration typically hovers around 10-15 minutes.
   * There's a noticeable dip in the trip duration at certain points. These dips might correspond to specific days where shorter rides were more common, and understanding the reasons would require more detailed analysis.
3. **Average Trip Distance over Time**:
   * Like trip duration, the average trip distance remains relatively steady over the period, typically around 1 to 1.5 miles.
   * There are also a few dips and peaks in this metric, which could be influenced by various factors.

Next, we'll move on to the **Station Analysis** to identify the most and least popular start and end stations.

Finished working

Show work

A group of colorful bars

Description automatically generated

Here's a summary of the station analyses:

1. **Top 10 Most Popular Start Stations**:
   * The chart visualizes the ten most frequently used start stations. "Kingsbury St & Kinzie St" stands out as the most popular start station.
2. **Top 10 Most Popular End Stations**:
   * Similarly, "Kingsbury St & Kinzie St" is also the most popular end station, followed closely by several other stations. This suggests that certain stations, possibly due to their strategic location or nearby amenities, attract a higher volume of users.
3. **Top 10 Least Popular Start Stations**:
   * The chart shows the ten least frequently used start stations. These stations might be in less accessible or less populated areas, or there might be other factors affecting their low usage.
4. **Top 10 Least Popular End Stations**:
   * The least popular end stations mirror the start stations in terms of low usage. Understanding the reasons behind the low popularity of these stations could offer insights for potential improvements or service adjustments.

Lastly, let's examine the **Ride Characteristics** to understand the differences between member and casual riders in terms of trip duration and distance.

Finished working

Show work

A diagram of a rectangular and a rectangular object

Description automatically generated with medium confidence

Here's a summary of the ride characteristics analyses between member and casual riders:

1. **Average Trip Duration**:
   * Casual riders tend to have slightly longer trip durations compared to members. The median trip duration for casual riders is higher, and the spread (interquartile range) is broader. This might suggest that casual riders are more leisurely in their rides or possibly use the service for more exploratory or recreational purposes.
   * Members, on the other hand, have a tighter distribution of trip durations, indicating more consistent and possibly more purpose-driven rides.
2. **Average Trip Distance**:
   * The trip distances between members and casual riders are relatively similar. Both groups have a median trip distance of around 1 mile.
   * However, like the trip duration, casual riders show a slightly broader spread in trip distances, indicating more variability in the distances they travel.

These insights suggest that while members and casual riders might travel similar distances, the purposes and durations of their rides might differ. Members likely use the service more consistently for regular commutes or short trips, while casual riders might be more exploratory.

If there are any additional analyses or specific areas you'd like to delve into further, please let me know!



Add a loess line the Trend Analysis plots

ChatGPT

Sure! A LOESS (Locally Estimated Scatterplot Smoothing) line will help in visualizing the underlying trend in the data by smoothing out short-term fluctuations.

Let's add a LOESS line to the Trend Analysis plots we generated earlier:

1. Number of Rides over Time
2. Average Trip Duration over Time
3. Average Trip Distance over Time

Finished working

Show work

It seems there was an error when trying to apply the LOESS smoothing line on the timestamp data. The issue arises from trying to fit a regression line directly to timestamp data in the regplot function.

To address this, I'll convert the timestamp data to a numeric format, fit the LOESS line, and then adjust the x-axis labels to display dates. Let's try this approach.

Finished working

Show work

A screenshot of a graph

Description automatically generated

Here are the updated trend analysis plots with the LOESS (Locally Estimated Scatterplot Smoothing) lines:

1. **Number of Rides over Time**:
   * The red LOESS line captures the underlying trend in the data, smoothing out short-term fluctuations. It provides a clearer view of the general direction and pattern of the data over time. We can observe that the number of rides has been showing an upward trend toward the end of the period.
2. **Average Trip Duration over Time**:
   * The LOESS line shows some periodic fluctuations in average trip duration over time, with certain dates having noticeably shorter average durations.
3. **Average Trip Distance over Time**:
   * The smoothed line indicates that the average trip distance remains relatively steady throughout the observed period, with some minor fluctuations.

The LOESS line provides a clearer understanding of the underlying patterns in the data by highlighting the general trend while ignoring short-term noise.

If there are any other specific analyses or questions you have, please let me know!



Plot the total number of rides, total trip duration and total trip distance

ChatGPT

Finished working

Show work

A screenshot of a graph

Description automatically generated

Here are the plots depicting the cumulative metrics over time:

1. **Total Number of Rides over Time**:
   * This plot shows the total number of rides each day. Similar to our earlier observation, there's an upward trend in the number of rides, especially toward the end of the period.
2. **Total Trip Duration over Time**:
   * The total trip duration each day shows some fluctuations but generally follows the pattern we observed in the number of rides. Some days have notably higher total durations, indicating days with either more rides or longer average ride durations.
3. **Total Trip Distance over Time**:
   * The total trip distance each day remains relatively consistent, mirroring the trends observed in the number of rides and total duration. Some spikes indicate days where riders traveled greater distances collectively.

These plots provide an aggregate view of the bikeshare usage over the three-month period, highlighting days of high and low activity.

If there's anything else you'd like to explore or analyze, please let me know!



Plot cumulative totals for rides, duration and distance

ChatGPT

Finished working

Show work

A graph of a graph

Description automatically generated with medium confidence

Here are the plots depicting the cumulative metrics over time:

1. **Cumulative Total Number of Rides over Time**:
   * The plot displays the running total of rides throughout the three-month period. As expected, this line consistently moves upwards, showing the accumulation of rides over time.
2. **Cumulative Total Trip Duration over Time**:
   * This line represents the running total of trip durations. We can observe the steady increase in cumulative duration, reflecting the accumulation of total ride minutes.
3. **Cumulative Total Trip Distance over Time**:
   * The line showcases the running total of the trip distances. The steady incline indicates the accumulation of miles traveled over the period.

These cumulative plots provide a sense of the aggregate growth of the bikeshare's usage metrics over the three-month span.

If there's anything else you'd like to explore or any further questions, please let me know!



Predict cumulative totals for next 30 days, and plot

ChatGPT

Predicting future values based on historical data is typically done using time series forecasting methods. One of the popular methods for this type of analysis is the ARIMA (AutoRegressive Integrated Moving Average) model.

Here's a breakdown of the approach:

1. **Model Selection**:
   * I'll fit an ARIMA model to each of the three time series datasets (cumulative rides, cumulative duration, and cumulative distance).
2. **Forecasting**:
   * Once the models are fit, I'll use them to forecast the next 30 days for each metric.
3. **Visualization**:
   * I'll visualize the actual historical data alongside the forecasted data for the next 30 days.

Let's start by fitting an ARIMA model to the cumulative rides data and then proceed with the other metrics.

Finished working

Show work

It seems there was an error due to a change in the newer version of the statsmodels library where the disp argument is no longer available for the fit method of the ARIMA model.

I'll correct this and rerun the forecasting.

Finished working

Show work

I apologize for the oversight. It seems I forgot to define the extended\_dates variable, which is essential for plotting the forecasts for the next 30 days.

Let me correct this and plot the forecasts again.

Finished working

Show work

A graph of a graph

Description automatically generated with medium confidence

Here are the plots depicting the cumulative metrics along with their forecasts for the next 30 days:

1. **Cumulative Total Number of Rides with Forecast**:
   * The blue line represents the actual cumulative total number of rides over the past three months. The red dashed line extends this to forecast the cumulative total for the next 30 days. As expected, the forecast shows a steady increase in cumulative rides.
2. **Cumulative Total Trip Duration with Forecast**:
   * Similarly, the blue line represents the actual cumulative total trip duration, and the red dashed line forecasts the cumulative total for the next 30 days. The forecast indicates a consistent growth in total duration.
3. **Cumulative Total Trip Distance with Forecast**:
   * The blue line showcases the actual cumulative total trip distance, while the red dashed line forecasts the cumulative total for the next month. The forecast suggests a steady increase in the total distance traveled by riders.

The forecasts provide an estimate based on past trends, but it's essential to note that real-world factors could influence these predictions. Regularly updating the model with fresh data can help improve its accuracy.

If there's anything else you'd like to explore or any further questions, please let me know!



Add a label for the last predicted value to the plots

ChatGPT

Finished working

Show work

I apologize for the oversight. It seems I made an error when attempting to extract the last value from the forecasted series. I'll correct this mistake and plot the cumulative metrics with labels for the last predicted values.

Finished working

Show work

A graph of a graph

Description automatically generated with medium confidence

Here are the plots depicting the cumulative metrics along with their forecasts for the next 30 days, now with labels for the final predicted values:

1. **Cumulative Total Number of Rides with Forecast**:
   * The blue line represents the actual cumulative total number of rides over the past three months. The red dashed line extends this to forecast the cumulative total for the next 30 days. The labeled point at the end of the forecasted line indicates the predicted cumulative total number of rides at the end of the 30-day forecast period.
2. **Cumulative Total Trip Duration with Forecast**:
   * Similarly, the blue line represents the actual cumulative total trip duration, and the red dashed line forecasts the cumulative total for the next 30 days. The labeled point provides the predicted cumulative total trip duration at the end of the forecast.
3. **Cumulative Total Trip Distance with Forecast**:
   * The blue line showcases the actual cumulative total trip distance, while the red dashed line forecasts the cumulative total for the next month. The labeled point provides the predicted cumulative total trip distance at the end of the forecasted period.

Having labels for the final predicted values allows for easier interpretation and understanding of the forecasted end values.

If there's anything else you'd like to explore or any further questions, please let me know!



What is the max value for each plot

ChatGPT

Finished working

Show work

Here are the maximum values for each of the cumulative metrics:

1. **Cumulative Rides**:
   * Actual Maximum: 274,571 rides
   * Forecasted Maximum (at the end of the next 30 days): approximately 364,966 rides
2. **Cumulative Duration (Minutes)**:
   * Actual Maximum: 3,632,507 minutes
   * Forecasted Maximum (at the end of the next 30 days): approximately 4,396,353 minutes
3. **Cumulative Distance (Miles)**:
   * Actual Maximum: 375,514 miles
   * Forecasted Maximum (at the end of the next 30 days): approximately 488,717 miles

The forecasted maximum values provide an estimate of where each metric might stand at the end of the next 30 days based on the current trend.

If you have any other questions or need further insights, let me know!



Add 95 % confidence bands

ChatGPT

To add 95% confidence intervals to our forecast, we'll need to modify our forecasting function. The ARIMA model's get\_forecast method can provide confidence intervals for our predictions.

After obtaining the confidence intervals, we can then visualize them as shaded regions around our forecasted values on the plots.

Let's proceed with updating our forecasting function and then plotting the results with 95% confidence bands.

Finished working

Show work

A graph of a graph

Description automatically generated with medium confidence

Here are the updated plots depicting the cumulative metrics along with their forecasts for the next 30 days, now enhanced with 95% confidence bands:

1. **Cumulative Total Number of Rides with Forecast**:
   * The blue line represents the actual cumulative total number of rides. The red dashed line provides the forecast for the next 30 days. The shaded pink region around the forecasted values represents the 95% confidence interval.
2. **Cumulative Total Trip Duration with Forecast**:
   * The blue line represents the actual cumulative total trip duration, while the red dashed line gives the forecast for the next 30 days. Again, the shaded pink region depicts the 95% confidence interval around the forecast.
3. **Cumulative Total Trip Distance with Forecast**:
   * The blue line showcases the actual cumulative total trip distance, with the red dashed line providing the forecast for the next month. The pink shaded region highlights the 95% confidence interval for the forecasted values.

The 95% confidence bands give an interval where we expect the future values to lie with 95% confidence, based on the current data and model. It provides a range of plausible values for our forecasts, considering the inherent uncertainty in forecasting.

If you have any more questions or further insights you'd like to explore, let me know!