

# CS7CS4 Machine Learning

## Final Project Dublin Bike Analysis

### Minjuan Luo

### 20313326

1. Requirement

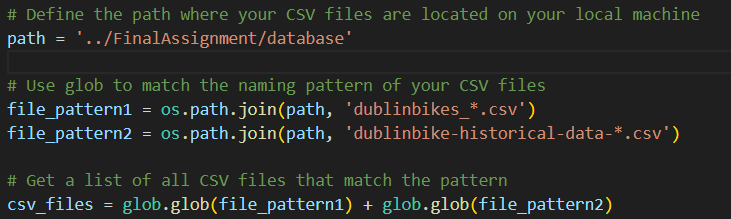
Scenario: You are working for FUTURE-DATA a local company specialized in data science. Dublin City Council hired your company to study the impact of COVID-19 on the city-bikes usage as they are planning to optimize the city-bike system. Dublin City Council had originally structured the city-bike network based on the forecasts of bike usage up to 2030. However, they think that the usage may not match the initial prediction because of the impact of the pandemic on our mobility. FUTURE-DATA decided that their first step should be to investigate the impact of the pandemic on the usage of the city bike network.

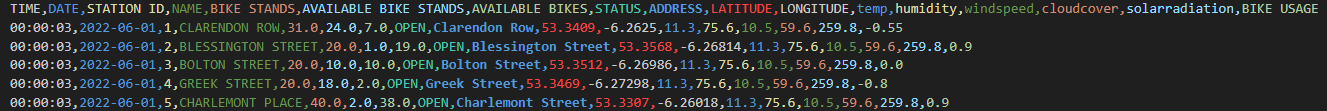
Task: The company agreed with our manager on two goals:

1. To assess the impact of the pandemic on the city-bike usage for the pandemic period.

(2) To assess the impact of the pandemic on the city-bike usage for the post-pandemic period.

1. Cleaning, Preprocessing and Concatenating
2. The current database provided is a Dublin bikes dataset at <https://data.gov.ie/dataset/dublinbikes-api>. The dataset is organised in party quarterly (i.e., four data-files per year) and in part monthly, meaning that the project might involve concatenating different portions of the dataset.
3. So, we first of all concatenate all Dublin bike dataset together from 2018-8-1 to 2023-11-24. Note that monthly dataset’s column name is partially inconsistent with the quarterly dataset. Concatenating two different datasets will in need of further column name changing



1. Relevant weather data from <https://www.visualcrossing.com/weather/weather-data-services> are also imported consistent with the date time including temperature, humidity, wind speed, cloud cover and solar radiation. Split Time column into Time and Date column in order to help merge with daily weather data. 
2. Calculate the bike usage:
   1. Bike Usage Ratio (BUR): This feature can be calculated using the formula:

BUR = AVAILABLE BIKES / BIKE STANDS

This ratio indicates the proportion of bikes currently in use. A higher ratio suggests higher usage of bikes.

* 1. Bike Availability Index (BAI): This can be calculated as:

BAI = AVAILABLE BIKE STANDS / BIKE STANDS

This index indicates the availability of bike stands for returning bikes. A lower index suggests higher bike usage and fewer available stands.

* 1. Combined Usage and Availability Score (CUAS): This feature combines both usage and availability:

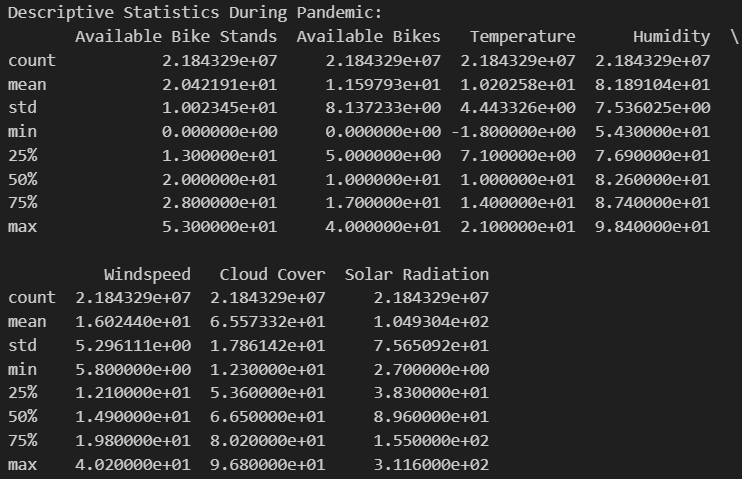
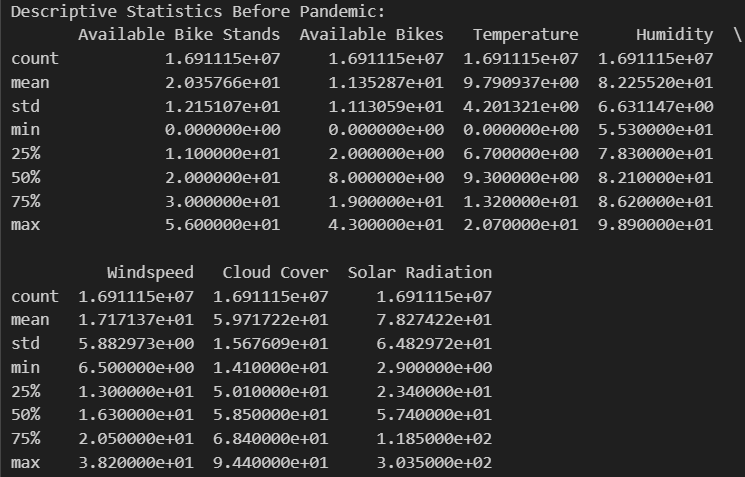
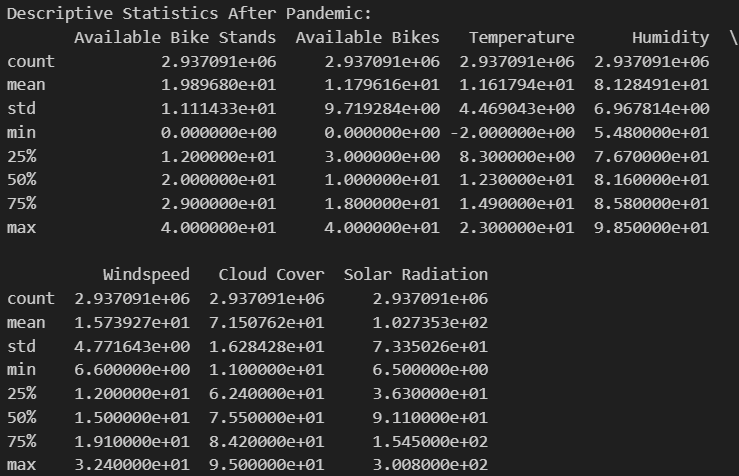
CUAS = BUR − BAI

This score provides a balanced view of both bike usage and stand availability. A positive high score indicates high usage and low availability of stands, which could suggest a need for more bikes or stands in that area.

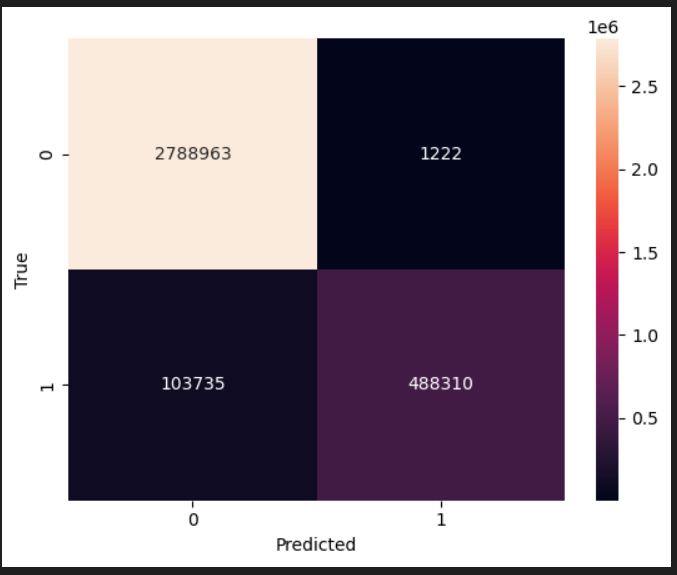
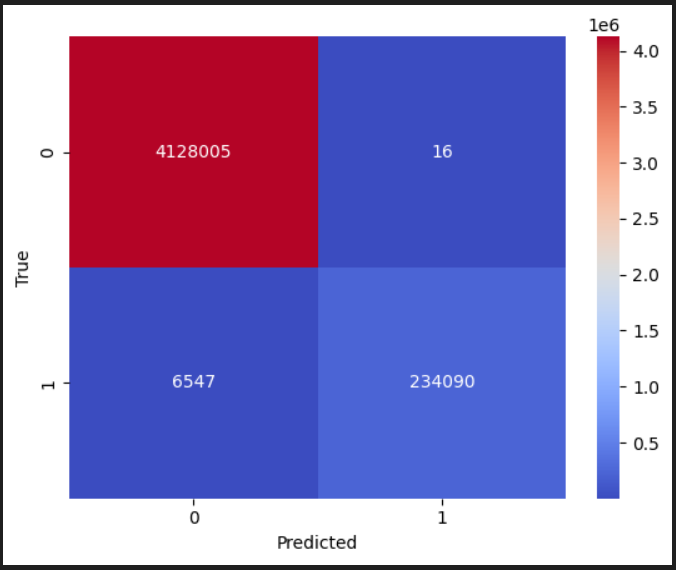
And this is the final BIKE USAGE column

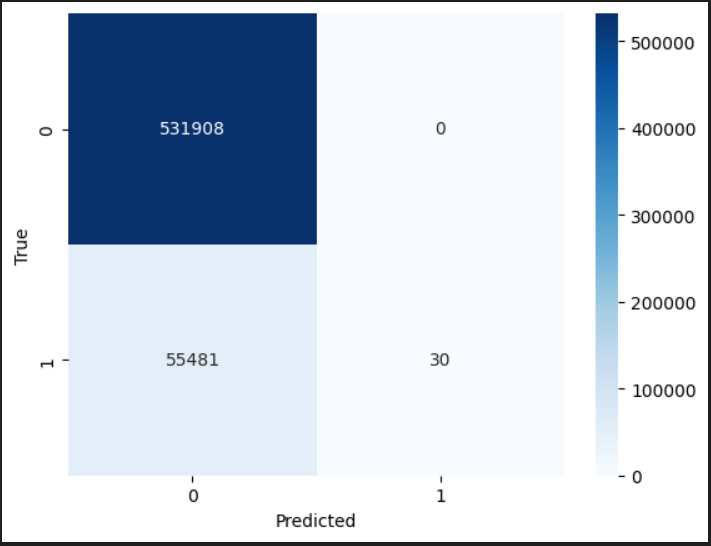
* 1. Finally, split the whole database into 3 parts: before pandemic (2018-8-1 to 2020-3-1), during pandemic (2020-3-1 to 2022-6-1) and after pandemic (2022-6-1 to 2023-11-24).

1. Static Analysis
   1. Descriptive Statistical Analysis
      1. Before pandemic:
         1. available bike max: 4.300000e+01 median: 8.000000e+00 mean: 1.135287e+01
         2. available bike stands max: 5.600000e+01 median: 2.000000e+01 mean: 2.035766e+01
      2. During pandemic:
         1. available bike max: 4.000000e+01 median: 1.000000e+01 mean: 1.159793e+01
         2. available bike stands max: 5.300000e+01 median: 2.000000e+01 mean: 2.042191e+01
      3. After pandemic:
         1. available bike max: 4.000000e+01 median: 1.000000e+01 mean: 1.179616e+01
         2. available bike stands max: 4.000000e+01 median: 2.000000e+01 mean: 1.989680e+01
      4. Base on the max, median and mean number from three different kinds of period, we can easily see that the pandemic had a noticeable effect on the availability of bikes and bike stands, with a decrease in the median number of available bikes. Post-pandemic, these numbers did not return to pre-pandemic levels, suggesting a lasting impact on the patterns of bike usage in Dublin. It's important to note that the data indicates more extreme values for both bikes and stands before the pandemic, which could be due to less consistent usage patterns. During and after the pandemic, the usage patterns appear to have become more stable, but with generally lower availability, indicating higher or more consistent usage.

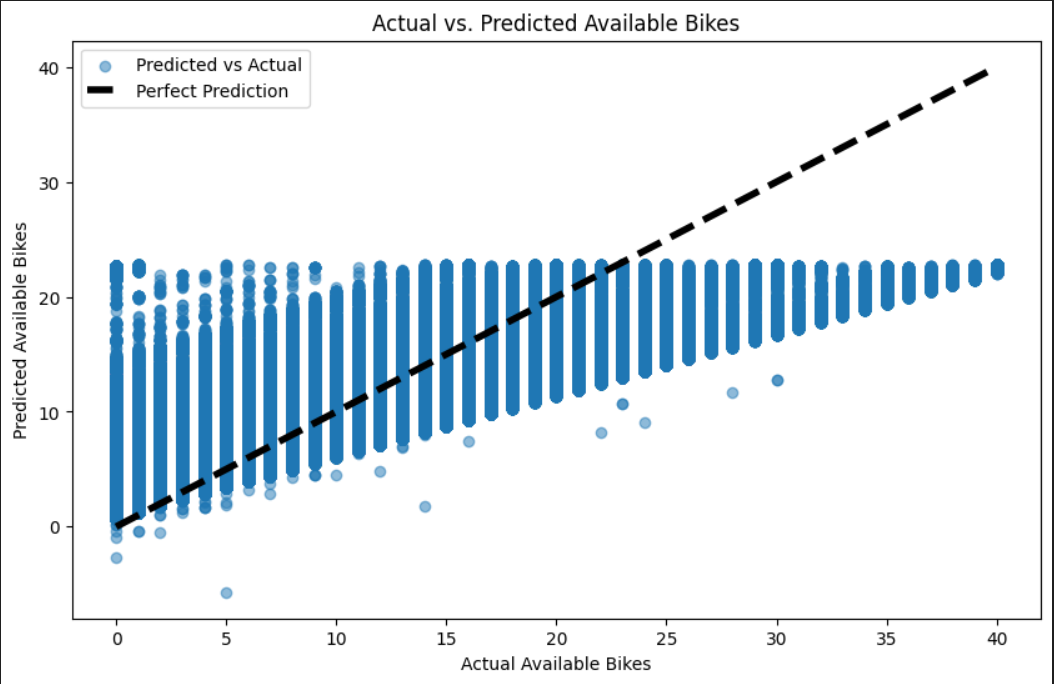
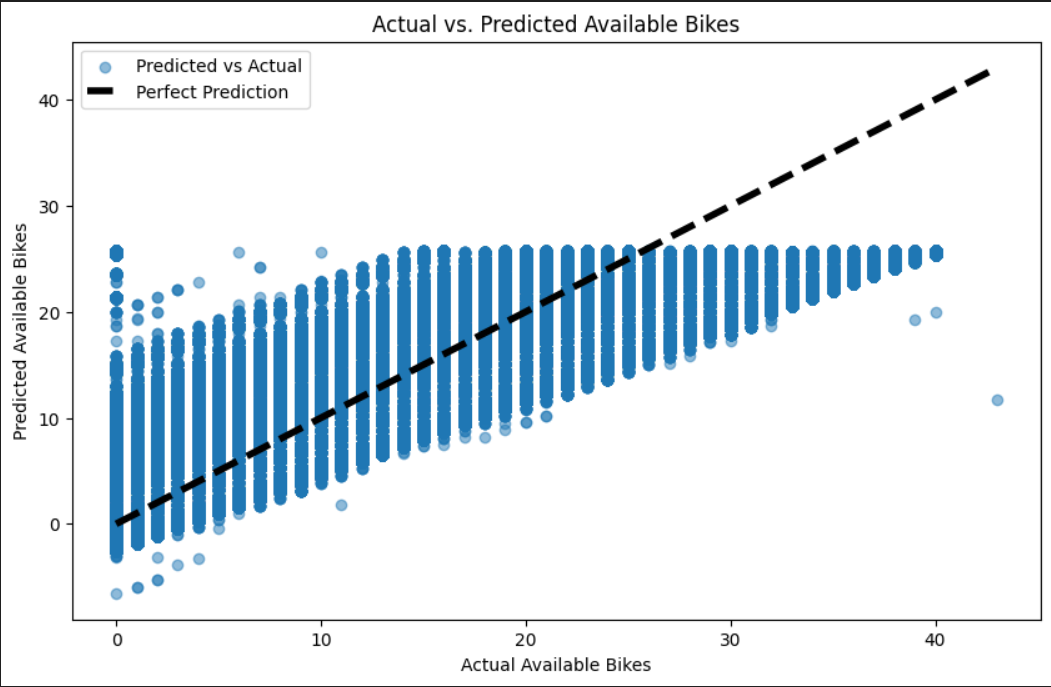
 

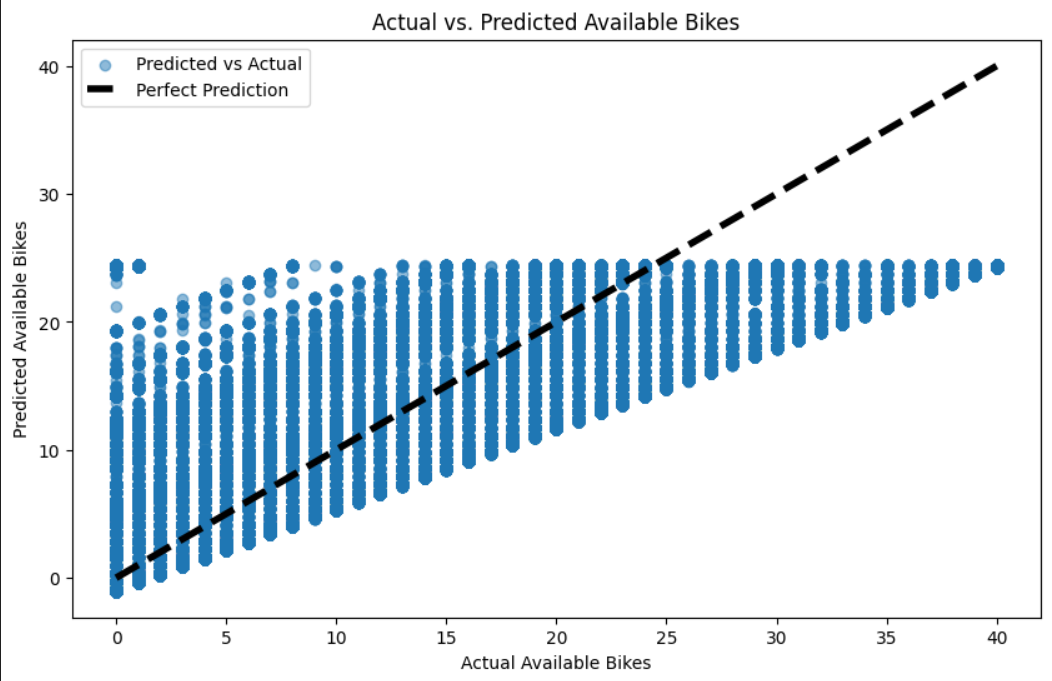
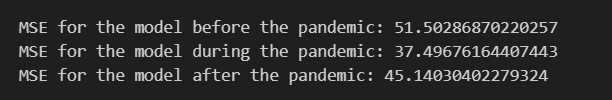
* 1. Confusion Matrix
     1. If we were to build a classification model, we would need a categorical target variable. For example, we could predict whether a bike station is likely to be full or empty (a binary classification)
     2. For the sake of this example, let's assume we're predicting whether a station is "full" (no available bikes) or "not full" (at least one available bike).
        1. True Positives (TP): The number of times the model correctly predicts that a station is full.
        2. True Negatives (TN): The number of times the model correctly predicts that a station is not full (there's at least one bike).
        3. False Positives (FP): The number of times the model incorrectly predicts that a station is full (but there are bikes).
        4. False Negatives (FN): The number of times the model incorrectly predicts that a station is not full when it is actually full.
     3. Conclusion
        1. Sensitivity/Recall (True Positive Rate): The model's ability to correctly predict 'full' stations decreased significantly from before to after the pandemic. This could be due to changes in patterns of bike station usage during these periods.
        2. Specificity (True Negative Rate): The model consistently predicted 'empty' stations well, especially during and after the pandemic.
        3. Precision: Before and during the pandemic, the model had a high precision for 'full' predictions, but this metric likely dropped after the pandemic due to the low number of correct 'full' predictions.
        4. Accuracy: The accuracy of the model (both true positives and true negatives) was highest before the pandemic and seemed to have decreased, particularly after the pandemic.
        5. Adaptability: It appears that the model may not have adapted well to changes in bike station usage patterns, particularly in the post-pandemic period. This could be due to a shift in human behavior that the model has not learned from.

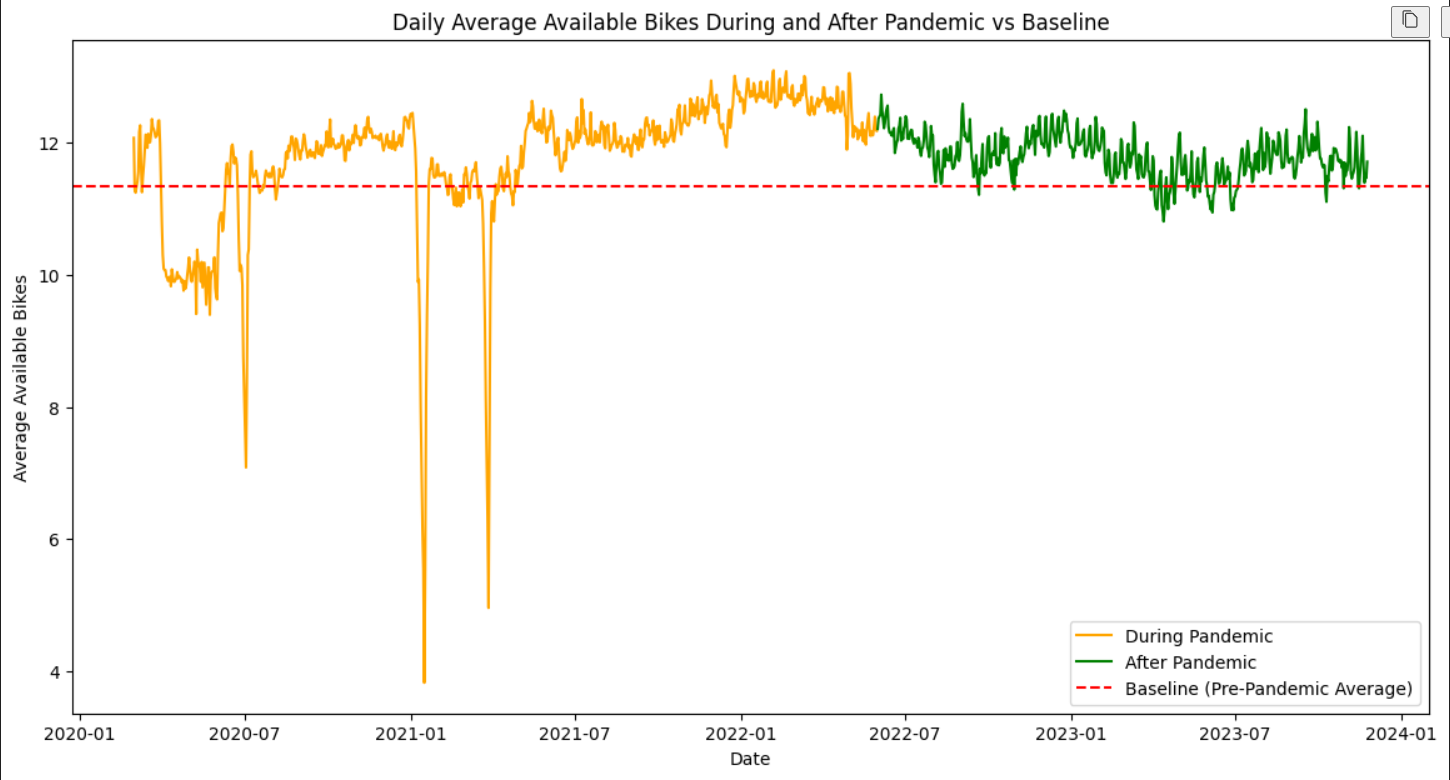


* 1. Linear Regression Model
     1. The scatter plot generated from the code represents the relationship between the actual and predicted values of available bikes from the 'after pandemic' dataset using a linear regression model. Here are the key elements of the plot and what they represent:
     2. Data Points (Scatter Plot): Each point on the scatter plot represents an individual observation from the test dataset. The x-coordinate of a point corresponds to the actual number of available bikes (the true value), and the y-coordinate corresponds to the predicted number of available bikes (the value predicted by the model).
     3. Diagonal Line (Perfect Prediction Line): The diagonal dashed line represents the line of perfect prediction. If a prediction is perfect, the actual and predicted values would be equal, placing the point along this line. The closer the points are to this line, the more accurate the predictions are.
     4. Distribution of Points: The spread of the points around the diagonal line indicates the variance in the model's accuracy. Points that are far from the line represent predictions that are less accurate, whereas points that lie close to the line represent more accurate predictions.
     5. Model Fit: The model had the best fit before the pandemic, with a slight reduction during, and a notable reduction after the pandemic.
     6. Predictive Accuracy: Predictive accuracy seems to have decreased after the pandemic, with greater discrepancies between predicted and actual values.
     7. Outliers: There are outliers in all three periods, but their impact appears to be the least before the pandemic and more pronounced after the pandemic.
     8. Variance: The variance in the prediction errors appears to increase in the later periods, especially after the pandemic, suggesting changes in usage patterns that the model did not account for.
     9. Model Optimization: The model was more accurate during the pandemic, which may be due to less variation in the number of bikes used as people's movements were more restricted.



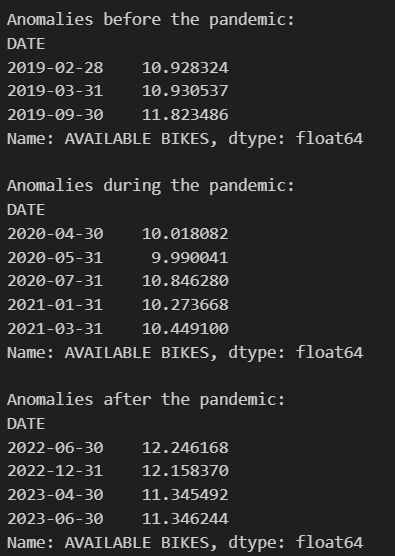
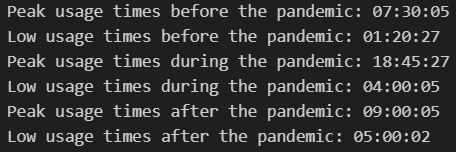
 

* 1. Base line comparison
     1. It calculates the baseline number of available bikes from the pre-pandemic data.
     2. It groups the during-pandemic data by date and calculates the daily average of available bikes.
     3. It finds the difference between the daily averages during the pandemic and the baseline to identify trends.

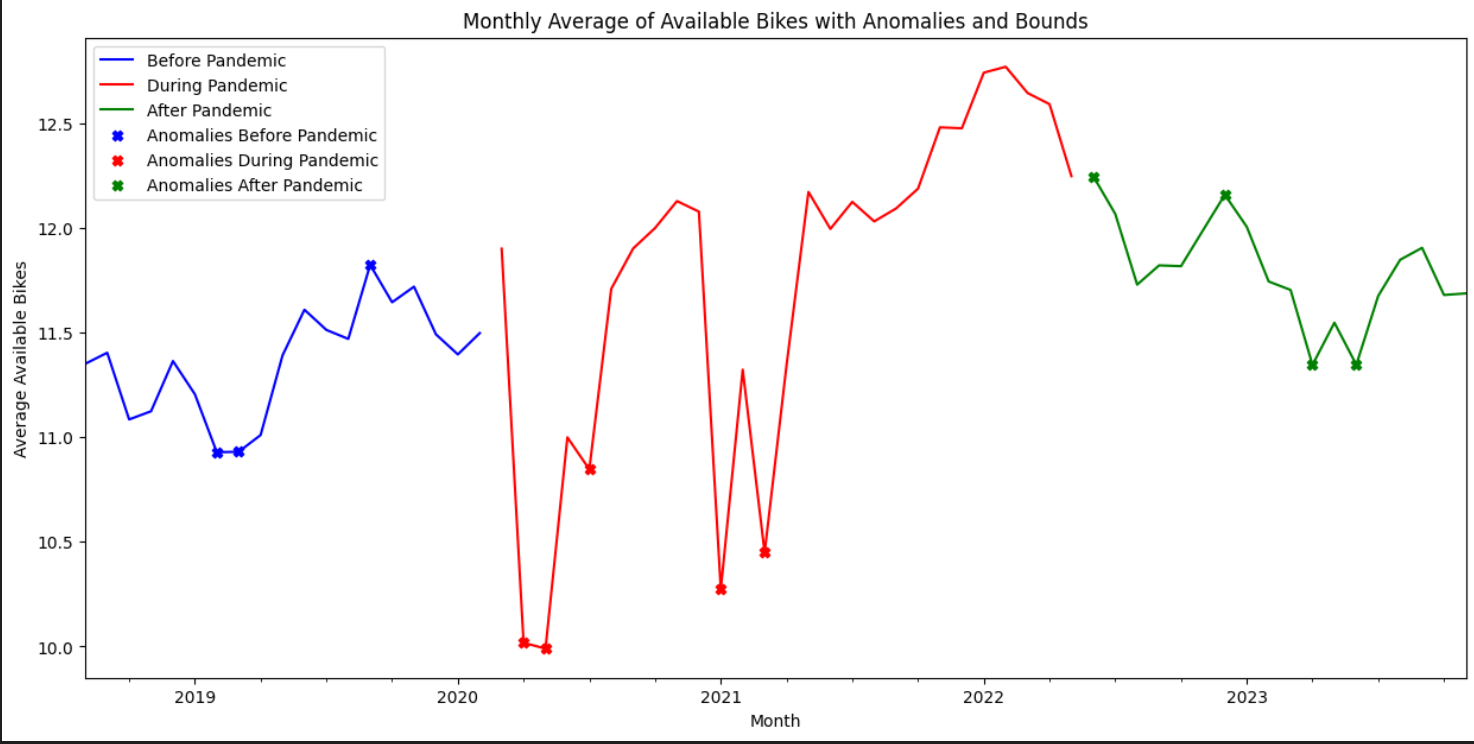


* + 1. Impact of the Pandemic: There was a noticeable impact of the pandemic on the availability of bikes, with lower averages during the pandemic compared to the baseline. This could be due to increased usage or logistical challenges in bike supply and maintenance during this period.
    2. Post-Pandemic Recovery: After the pandemic, there seems to be a recovery or even an increase in the average number of available bikes compared to the baseline, which could be due to a decrease in usage as people return to other modes of transportation or due to an increase in the number of bikes provided by the service.
    3. Variability: The variability in the number of bikes available during the pandemic was high, with sharp drops potentially corresponding to lockdowns or other restrictive measures. After the pandemic, the variability is reduced, indicating a more stable situatio

1. Temperal Dynamic Analysis



* 1. Peak Usage Time
     1. Shift in Commuting Patterns: There is a clear shift in peak usage times from the early morning before the pandemic to the evening during the pandemic, and then a slight delay in the morning after the pandemic. This suggests that the pandemic may have led to lasting changes in commuting habits.
     2. Stable Low Usage Times: The low usage times remain in the early morning hours across all periods, which is consistent with a typical daily pattern where bike usage would be minimal.
     3. Potential for Flexible Work Policies: The shift in peak usage times may reflect wider adoption of flexible work policies, which could have implications for transportation planning and bike-sharing services.
  2. Time Series Analysis
     1. Include monthly average data from all three period of data first, calculate the median and MAD for the provided data and then identifies any data points that fall outside the median ± MAD range as anomalies. These anomalies are then plotted as distinct points on the time series plot, and their values are printed out.
     2. Increased Volatility During Pandemic: The pandemic brought significant volatility to bike availability, with sharp increases and decreases likely reflecting the changing public health guidelines and people's responses to them.
     3. Higher Average Post-Pandemic: After the pandemic, the average availability of bikes is higher than the pre-pandemic levels, which might suggest an increase in the bike fleet or a change in maintenance and supply strategies by bike-sharing providers.
     4. Declining Trend Post-Pandemic: There's a noticeable downward trend in bike availability after the pandemic, which could indicate a normalization of usage patterns or perhaps changes in the management of the bike-sharing system.
     5. Weather and Seasonal Impact: The availability of bikes before the pandemic shows a clear seasonal pattern, likely influenced by weather conditions and related seasonal activities.
     6. Pandemic Disruption: The pandemic significantly disrupted these patterns, with usage possibly driven more by lockdowns and social distancing measures than by seasonality.
     7. Returning Patterns: After the pandemic, there is evidence of a return to seasonal patterns, although the trend line suggests that the average availability has not yet stabilized to pre-pandemic levels.



* 1. Geospacial Analysis
     1. To perform geospatial analysis and map the bike stations using longitude and latitude data, we can use libraries such as matplotlib for plotting and mpl\_toolkits.mplot3d for 3D plotting.
     2. AVAILABLE BIKES
        1. Before the Pandemic:

The distribution of available bikes before the pandemic appears to be relatively uniform across the different locations. Higher bars in certain areas could indicate stations with consistently higher numbers of bikes available or possibly a higher number of stations in a concentrated area.

* + - 1. During the Pandemic:

The availability during the pandemic shows some reduction in the height of the bars in certain areas, suggesting a decrease in the number of available bikes. This could be due to reduced restocking of bikes, stations being temporarily closed, or increased usage at specific locations.

* + - 1. After the Pandemic:

Post-pandemic, the distribution of available bikes shows some areas with increased availability, as seen by taller bars, while others remain at levels similar to during the pandemic. This might indicate a return to normal operations with an additional increase in the bike fleet or a change in the distribution strategy to meet altered demand patterns.

* + 1. BIKE USAGE
       1. Before the Pandemic:

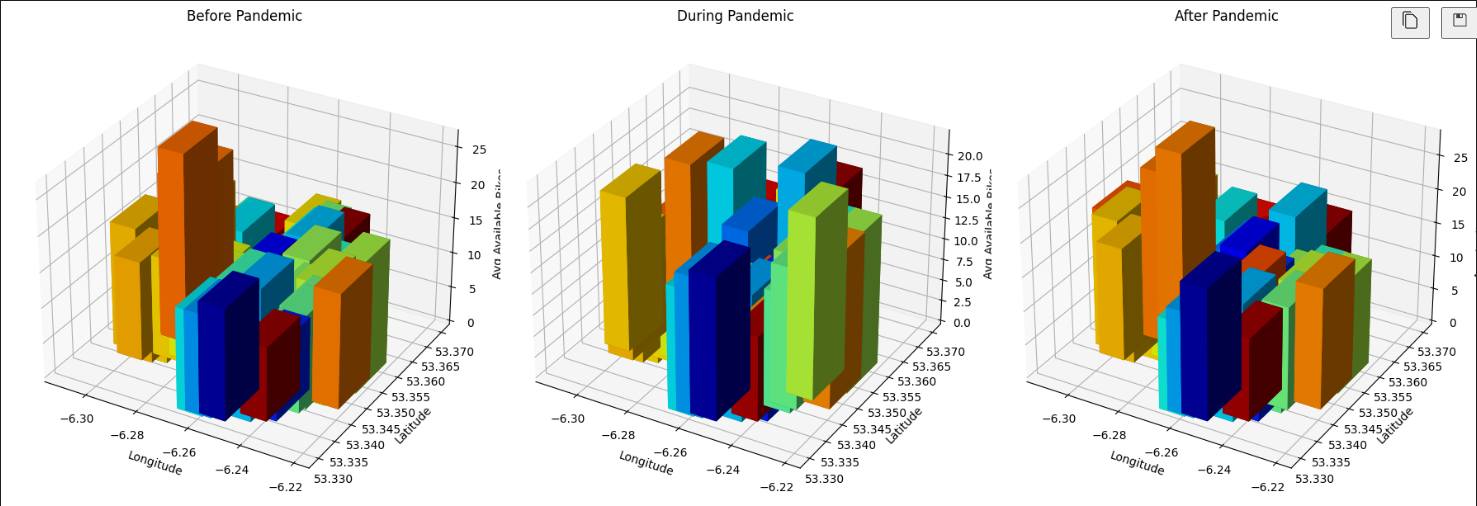
The usage patterns before the pandemic show moderate variation across locations, with some areas having higher usage than others, which may correlate with commercial or residential density, or the presence of key transit hubs.

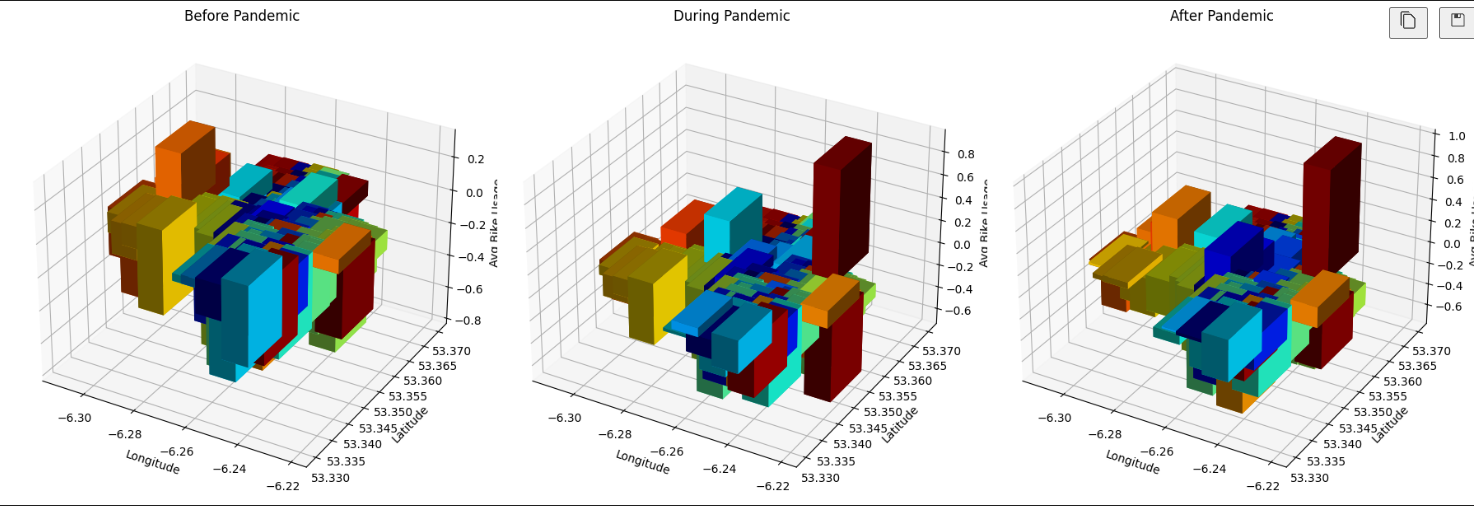
* + - 1. During the Pandemic:

During the pandemic, there's a noticeable shift in usage patterns, with some areas experiencing a significant drop in usage, shown by lower bars, and others a slight increase. This might reflect changes in mobility patterns during lockdowns and restrictions.

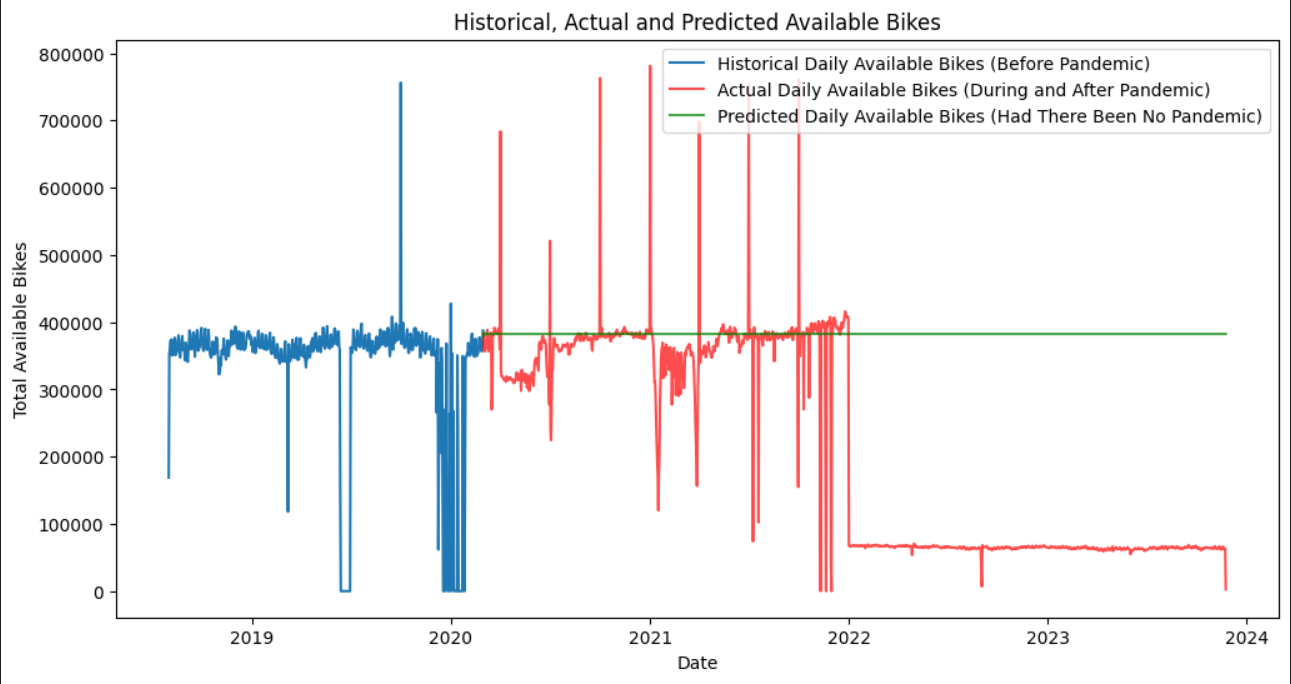
* + - 1. After the Pandemic:

The bike usage after the pandemic appears to have not returned entirely to pre-pandemic patterns, with some areas showing increased usage, while others have not recovered fully, suggesting a lasting impact of the pandemic on mobility behaviors.

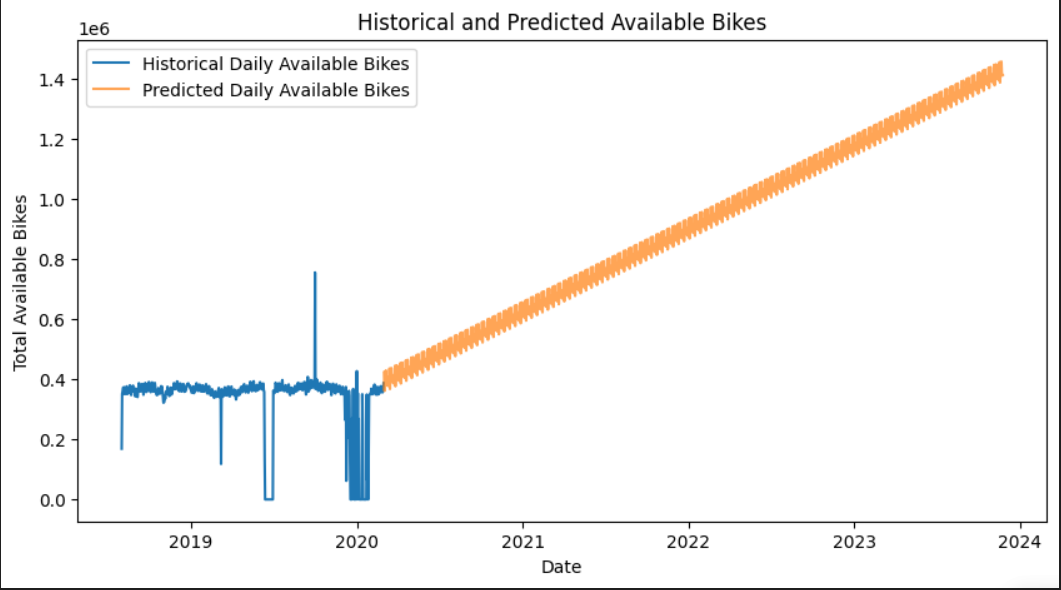
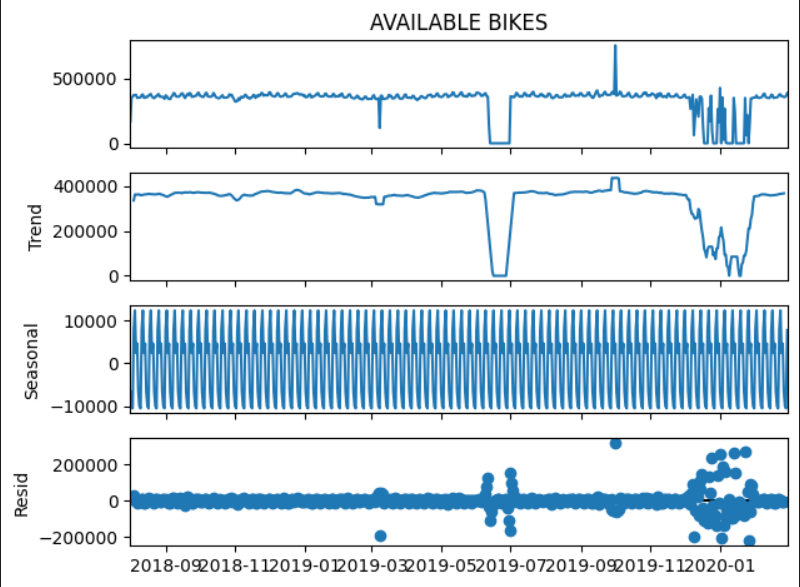




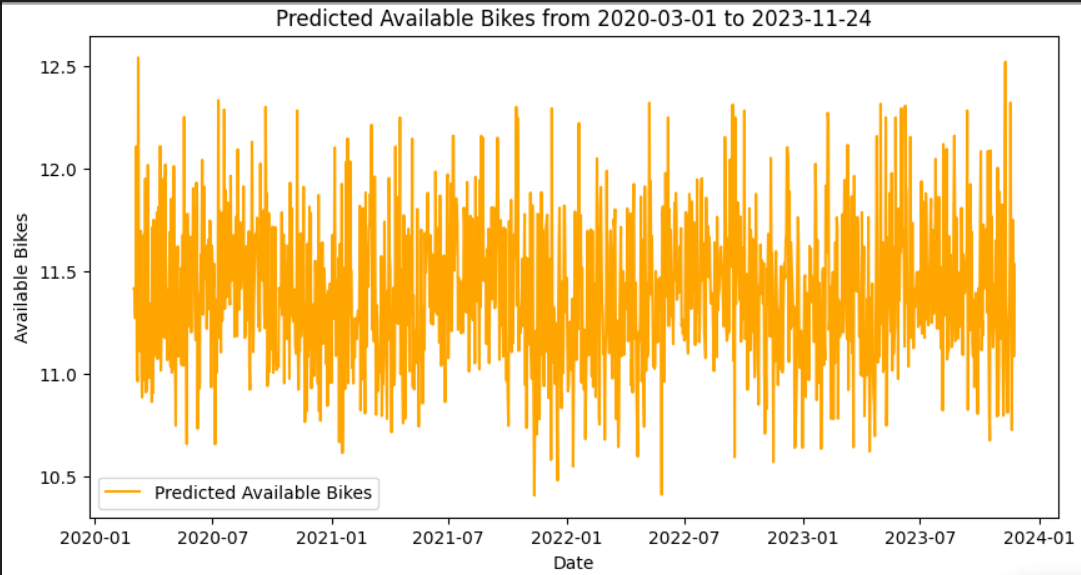
1. Machine Learning Model Analysis
   1. ARIMA (AutoRegressive Integrated Moving Average)
      1. To perform predictive modeling using machine learning to estimate what the available bike counts might have been had the pandemic not occurred, we can use a time series forecasting method. One common approach is to use ARIMA (AutoRegressive Integrated Moving Average), which is suitable for time series data without a clear trend or seasonal pattern.
      2. If there is seasonality in the data, SARIMA (Seasonal ARIMA) might be more appropriate. For this example, we'll use SARIMA as it allows us to model both the non-seasonal and seasonal elements.
      3. Impact of Pandemic: The actual available bikes during and after the pandemic deviate significantly from the predicted (no-pandemic) trend, highlighting the substantial impact of the pandemic on bike-sharing services.
      4. Disruption and Recovery: There is an evident disruption during the pandemic, with a partial recovery that follows. However, the recovery does not reach the expected levels as per the ARIMA model's prediction, indicating a potential lasting change in the usage or operation of the bike-sharing services.
      5. Forecasting Limitations: The ARIMA model could not predict the pandemic's impact, which underscores the limitations of such models in the face of unprecedented events that significantly alter historical patterns.



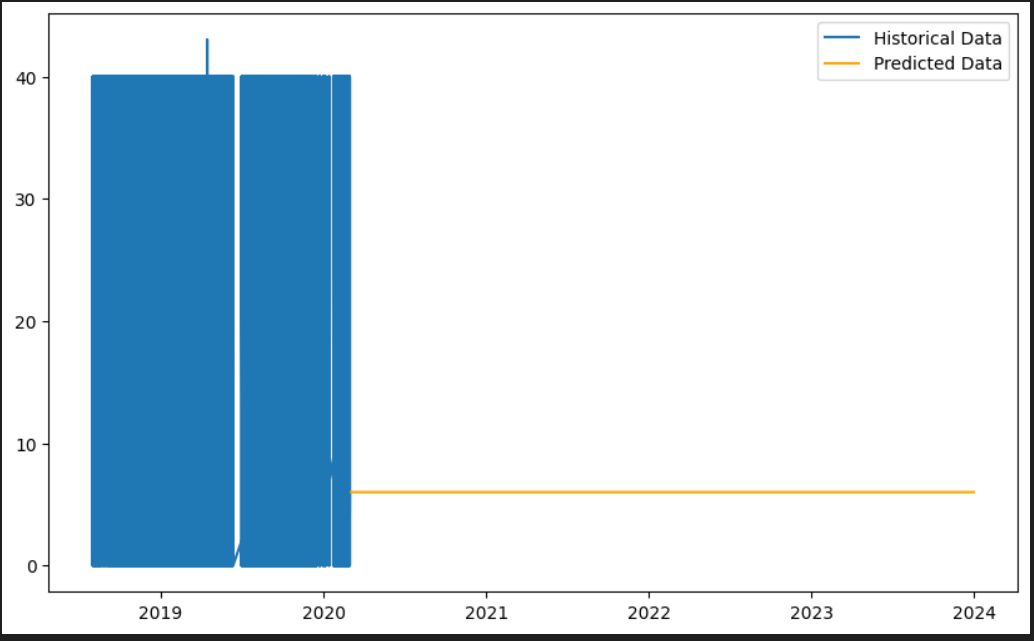
* 1. SARIMA (Seasonal ARIMA)



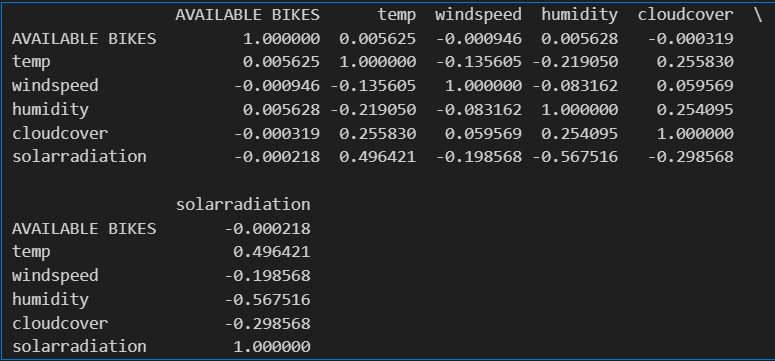
* + 1. Impact of Pandemic: Both the decomposition and the forecast illustrate the significant impact the pandemic had on the availability of bikes, with deviations from historical patterns and the expected trend.
    2. Recovery and Growth: The model predicts a recovery and subsequent growth in bike availability post-pandemic, suggesting that the system may bounce back stronger than before, possibly due to an increase in the bike fleet or improved operational efficiency.
    3. Residual Analysis: The residuals indicate that there may have been factors affecting bike availability that the model did not fully account for, particularly around the pandemic period. This could be due to the model's inability to predict unprecedented events.
  1. Random Forest Regressor
     1. To predict the number of available bikes using a more general machine learning model and including additional features like temperature and windspeed, we can use a model like Random Forest Regressor. These models can handle non-linear relationships and are robust to overfitting, making them suitable for this task.
     2. Model Insights: The Random Forest Regressor seems to capture a complex set of patterns that influence bike availability. It is sensitive to fluctuations, possibly reacting to multiple input features such as weather, time of day, day of the week, and other external factors.
     3. Uncertainty and Variability: The variability and the spikes in the predictions suggest that there might be days with significant deviations from the average expected availability. Operators should be prepared for such uncertainties and have contingency plans in place.

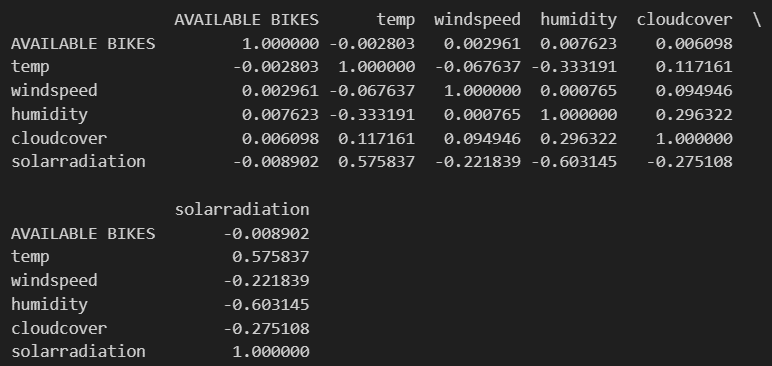


* 1. LSTM model

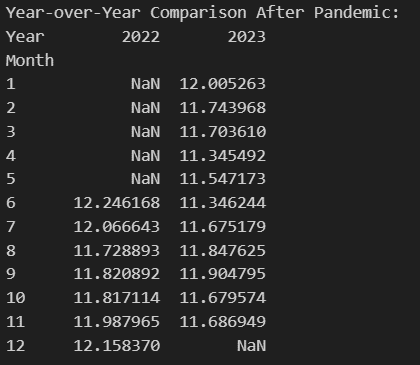
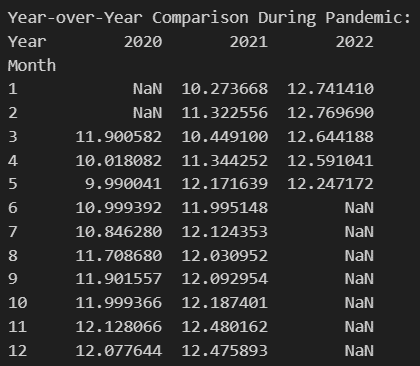
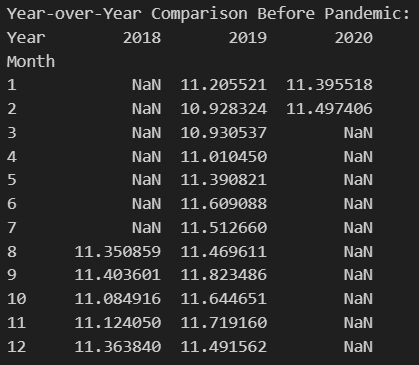


1. Integration of Weather Data
   1. Correlation Analysis
      * 1. In the correlation analysis, a positive correlation value indicates a direct relationship, while a negative value indicates an inverse relationship.
        2. Weather Impact: The correlation between weather conditions and bike availability varies across the three periods, suggesting that the pandemic may have influenced how weather conditions affect bike-sharing usage.
        3. Behavioral Changes: The shifts in correlation patterns during and after the pandemic might reflect changes in user behavior or operational changes in the bike-sharing system in response to the pandemic.

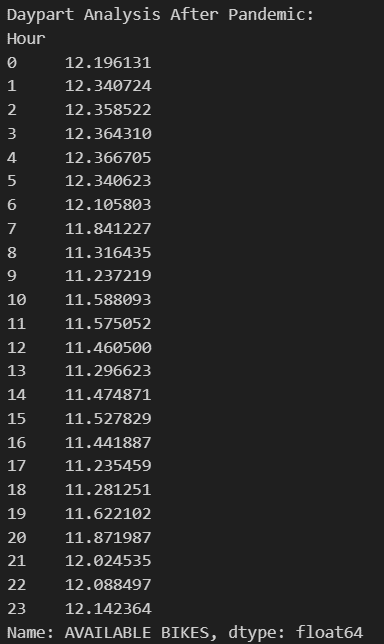
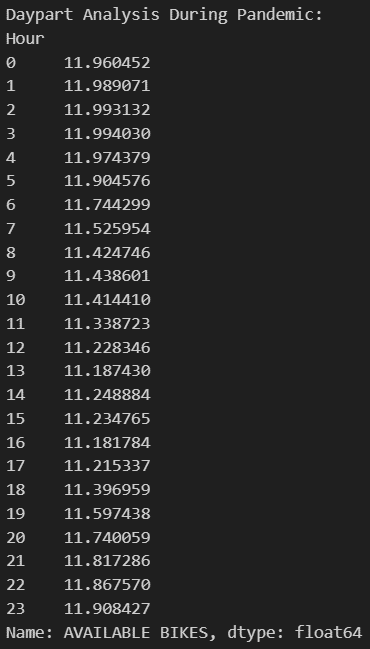
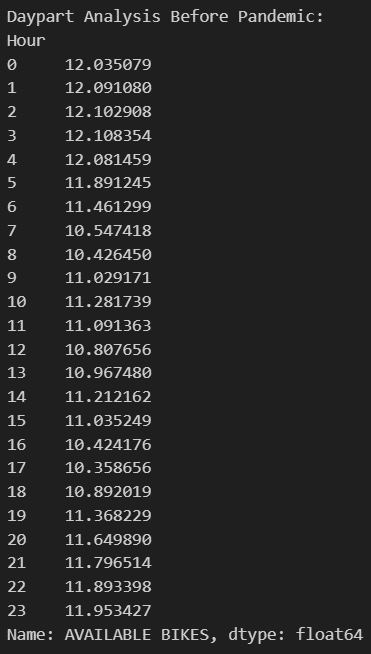




1. Comparative Analysis
   1. Year-over-Year Comparison
      1. From 2019 to 2020, there is a slight increase in bike availability for January and February, with no other direct comparisons available for the other months due to missing data.
      2. The increase in the early months of 2020 might reflect growth in the bike-sharing system or increased seasonal availability.
      3. The start of 2020 shows a significant increase in availability in March compared to the previous year, which could be due to pandemic-related changes in usage patterns.
      4. For 2021, the bike availability continues to rise compared to 2020, suggesting an adjustment to the pandemic situation, potentially with more bikes being made available as the service adapted to new conditions.
      5. There's a consistent year-over-year increase in availability from 2020 to 2021, which may indicate a recovery in bike sharing usage or a deliberate increase in the number of bikes provided.
      6. In 2022, there is a noticeable increase in bike availability compared to 2023, where data starts from May. This might suggest either a reduction in the number of bikes or changes in usage patterns post-pandemic.
      7. The available data shows a drop in availability in May and June of 2023 compared to 2022, indicating a possible normalization of bike usage as the situation stabilizes post-pandemic.



* 1. Daypart Analysis
     1. Before pandemic
        1. Bike availability starts higher at midnight and gradually decreases until the morning hours, reaching the lowest around 7-8 AM. This pattern suggests a high demand for bikes during the morning commute hours.
        2. Availability increases again throughout the day after the morning low, indicating that bikes are returned or redistributed.
        3. There is another lower point in the evening around 5-6 PM, which might reflect the evening commute where bikes are being used to return home.
     2. During pandemic
        1. The range of availability is tighter throughout the day compared to the pre-pandemic period, with less variation between the highest and lowest points. This could suggest a more even distribution of bike usage throughout the day or a general decrease in the usage due to lockdowns and other restrictions.
        2. The lowest points of availability are less pronounced, which may reflect the disruption of typical commute patterns as more people work from home.
     3. After pandemic
        1. The pattern of bike availability appears to be returning to a pre-pandemic-like pattern, with higher availability during the late night and early morning hours, and lower availability during typical commute times.



1. User Behaviour Analysis
   1. Cluster Analysis/Usage Patterns
      1. From the clustering analysis images, it seems that bike stations have been grouped into clusters based on their geographical location and usage patterns. The colors likely represent different clusters. The consistency in cluster formation across the three periods suggests that geographical location remains a strong determinant of usage patterns, with certain areas consistently having higher or lower bike usage.
      2. The numerical results indicate the average bike usage values for the respective periods:
         1. Before the pandemic: -0.278
         2. During the pandemic: -0.257
         3. After the pandemic: -0.239
      3. If we assume that a higher negative value indicates lesser usage or higher availability (which is a reasonable assumption if negative values indicate an excess of available bikes over rentals), this could mean the following:
         1. Before the Pandemic: Bike usage was at its lowest among the three periods, or bikes were more frequently available and not in use.
         2. During the Pandemic: There was a slight increase in bike usage, which could be due to a variety of factors such as people avoiding public transport due to health concerns, or changes in work/commute patterns.
         3. After the Pandemic: There was a reduction in bike usage as compared to during the pandemic, but not to the levels before the pandemic. This could indicate a partial return to pre-pandemic behavior or it could reflect new patterns of work and travel that have persisted after the pandemic.

