Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

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.

2. Cleaning, Preprocessing and Concatenating

(1) 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.

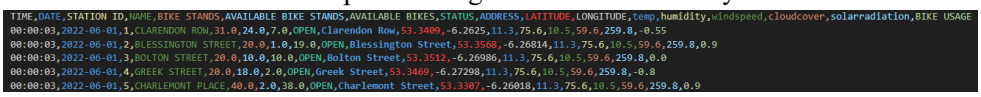
(2) So, we first of all concatenate all Dublin bike dataset together from 2018-8-1 to 2023-11-24. Note that monthly dataset's column name is partially inconsistent with the quarterly dataset. Concatenating two different datasets will in need of further column name changing

```
# Define the path where your CSV files are located on your local machine
path = '../FinalAssignment/database'

# Use glob to match the naming pattern of your CSV files
file_pattern1 = os.path.join(path, 'dublinbikes_*.csv')
file_pattern2 = os.path.join(path, 'dublinbike-historical-data-*.csv')

# Get a list of all CSV files that match the pattern
csv_files = glob.glob(file_pattern1) + glob.glob(file_pattern2)
```

(3) 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.



(4) Calculate the bike usage:

a) 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.

b) 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.

c) 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

d) 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).

3. Static Analysis

a) Descriptive Statistical Analysis

i. 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

ii. 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

iii. 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

iv. 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.

Descriptive Statistics Before Pandemic:

	Available Bike Stands	Available Bikes	Temperature	Humidity
count	1.691115e+07	1.691115e+07	1.691115e+07	1.691115e+07
mean	2.035766e+01	1.135287e+01	9.790937e+00	8.225520e+01
std	1.215107e+01	1.113059e+01	4.201321e+00	6.631147e+00
min	0.000000e+00	0.000000e+00	0.000000e+00	5.530000e+01
25%	1.100000e+01	2.000000e+00	6.700000e+00	7.830000e+01
50%	2.000000e+01	8.000000e+00	9.300000e+00	8.210000e+01
75%	3.000000e+01	1.900000e+01	1.320000e+01	8.620000e+01
max	5.600000e+01	4.300000e+01	2.070000e+01	9.890000e+01

	Windspeed	Cloud Cover	Solar Radiation
count	1.691115e+07	1.691115e+07	1.691115e+07
mean	1.717137e+01	5.971722e+01	7.827422e+01
std	5.882973e+00	1.567609e+01	6.482972e+01
min	6.500000e+00	1.410000e+01	2.900000e+00
25%	1.300000e+01	5.010000e+01	2.340000e+01
50%	1.630000e+01	5.850000e+01	5.740000e+01
75%	2.050000e+01	6.840000e+01	1.185000e+02
max	3.820000e+01	9.440000e+01	3.035000e+02

Descriptive Statistics During Pandemic:

	Available Bike Stands	Available Bikes	Temperature	Humidity
count	2.184329e+07	2.184329e+07	2.184329e+07	2.184329e+07
mean	2.042191e+01	1.159793e+01	1.020258e+01	8.189104e+01
std	1.002345e+01	8.137233e+00	4.443326e+00	7.536025e+00
min	0.000000e+00	0.000000e+00	-1.800000e+00	5.430000e+01
25%	1.300000e+01	5.000000e+00	7.100000e+00	7.690000e+01
50%	2.000000e+01	1.000000e+01	1.000000e+01	8.260000e+01
75%	2.800000e+01	1.700000e+01	1.400000e+01	8.740000e+01
max	5.300000e+01	4.000000e+01	2.100000e+01	9.840000e+01

	Windspeed	Cloud Cover	Solar Radiation
count	2.184329e+07	2.184329e+07	2.184329e+07
mean	1.602440e+01	6.557332e+01	1.049304e+02
std	5.296111e+00	1.786142e+01	7.565092e+01
min	5.800000e+00	1.230000e+01	2.700000e+00
25%	1.210000e+01	5.360000e+01	3.830000e+01
50%	1.490000e+01	6.650000e+01	8.960000e+01
75%	1.980000e+01	8.020000e+01	1.550000e+02
max	4.020000e+01	9.680000e+01	3.116000e+02

```
Descriptive Statistics After Pandemic:
       Available Bike Stands  Available Bikes   Temperature      Humidity  \
count           2.937091e+06     2.937091e+06  2.937091e+06  2.937091e+06
mean            1.989680e+01     1.179616e+01  1.161794e+01  8.128491e+01
std             1.111433e+01     9.719284e+00  4.469043e+00  6.967814e+00
min             0.000000e+00     0.000000e+00 -2.000000e+00  5.480000e+01
25%             1.200000e+01     3.000000e+00  8.300000e+00  7.670000e+01
50%             2.000000e+01     1.000000e+01  1.230000e+01  8.160000e+01
75%             2.900000e+01     1.800000e+01  1.490000e+01  8.580000e+01
max             4.000000e+01     4.000000e+01  2.300000e+01  9.850000e+01

          Windspeed   Cloud Cover  Solar Radiation
count  2.937091e+06  2.937091e+06     2.937091e+06
mean   1.573927e+01  7.150762e+01     1.027353e+02
std    4.771643e+00  1.628428e+01     7.335026e+01
min    6.600000e+00  1.100000e+01     6.500000e+00
25%    1.200000e+01  6.240000e+01     3.630000e+01
50%    1.500000e+01  7.550000e+01     9.110000e+01
75%    1.910000e+01  8.420000e+01     1.545000e+02
max    3.240000e+01  9.500000e+01     3.008000e+02
```

b) Confusion Matrix

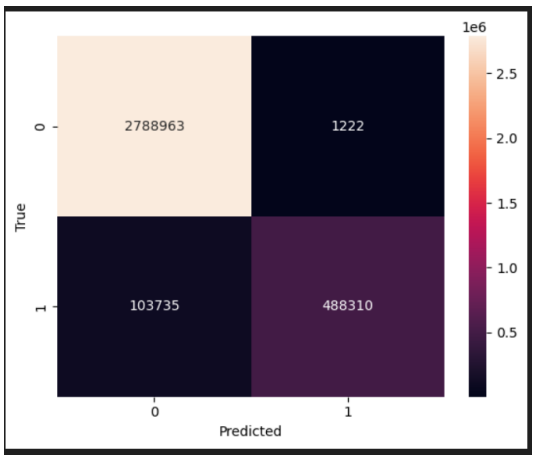
i. If we were to build a classification model, we would need a categorical target variable. For example, we could predict whether a bike station is likely to be full or empty (a binary classification)

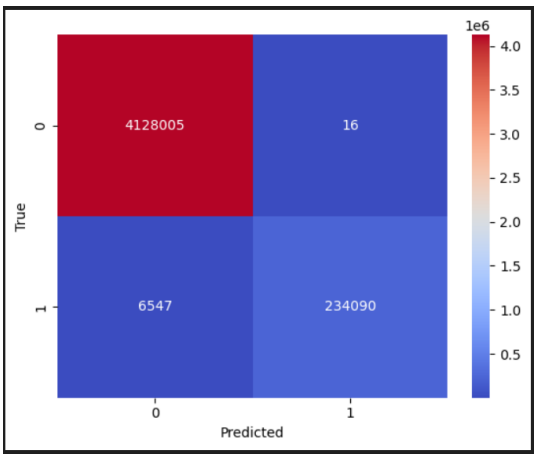
ii. For the sake of this example, let's assume we're predicting whether a station is "full" (no available bikes) or "not full" (at least one available bike).

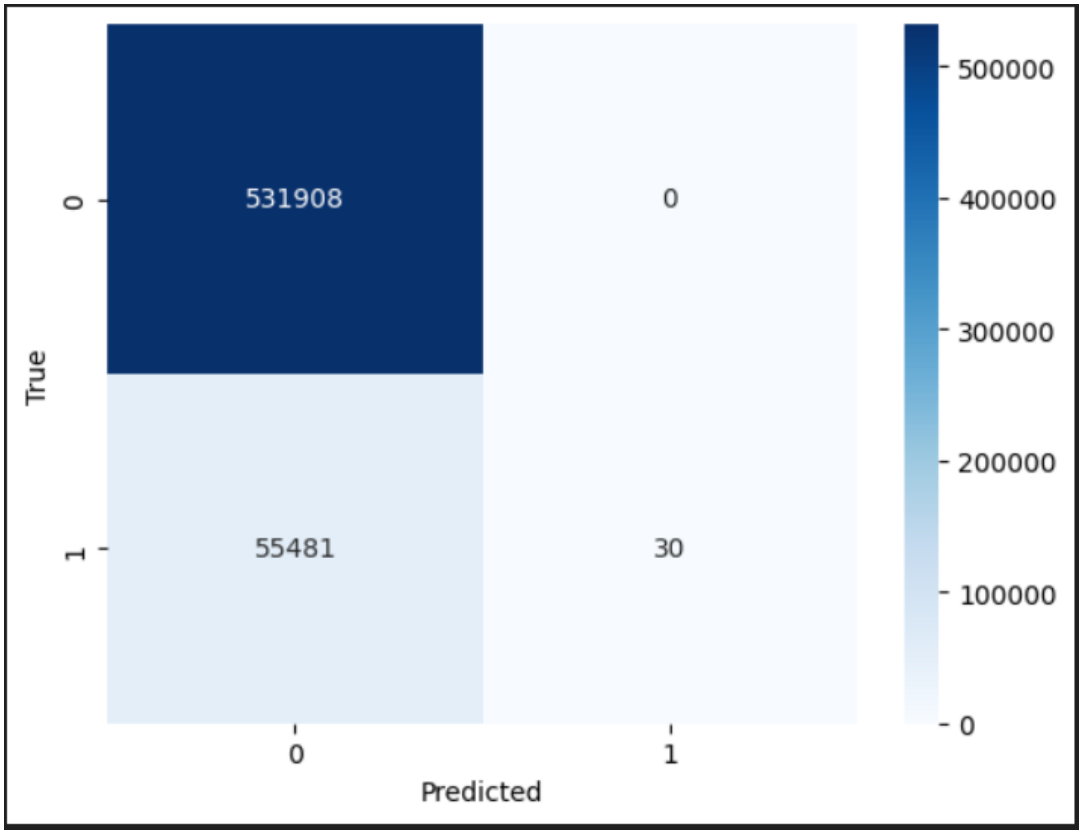
1. True Positives (TP): The number of times the model correctly predicts that a station is full.
2. True Negatives (TN): The number of times the model correctly predicts that a station is not full (there's at least one bike).
3. False Positives (FP): The number of times the model incorrectly predicts 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.

iii. 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.







c) Linear Regression Model

i. 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:

ii. 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).

iii. 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.

iv. 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.

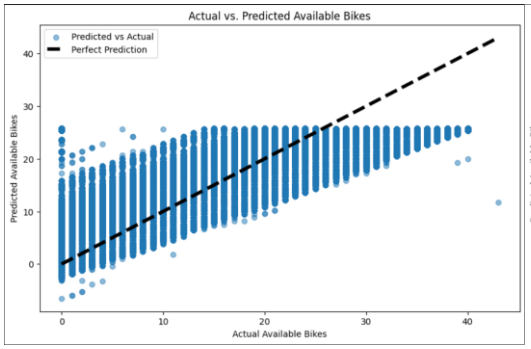
v. Model Fit: The model had the best fit before the pandemic, with a slight reduction during, and a notable reduction after the pandemic.

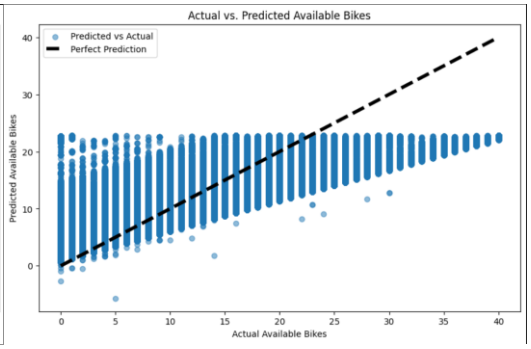
vi. Predictive Accuracy: Predictive accuracy seems to have decreased after the pandemic, with greater discrepancies between predicted and actual values.

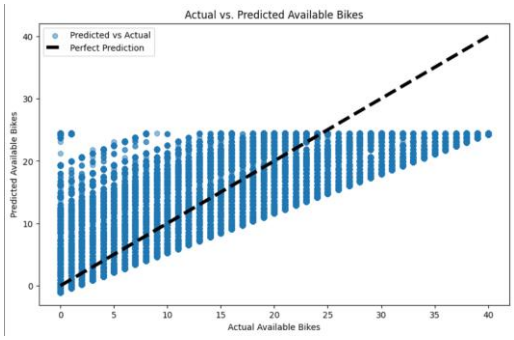
vii. Outliers: There are outliers in all three periods, but their impact appears to be the least before the pandemic and more pronounced after the pandemic.

viii. 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.

ix. 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.







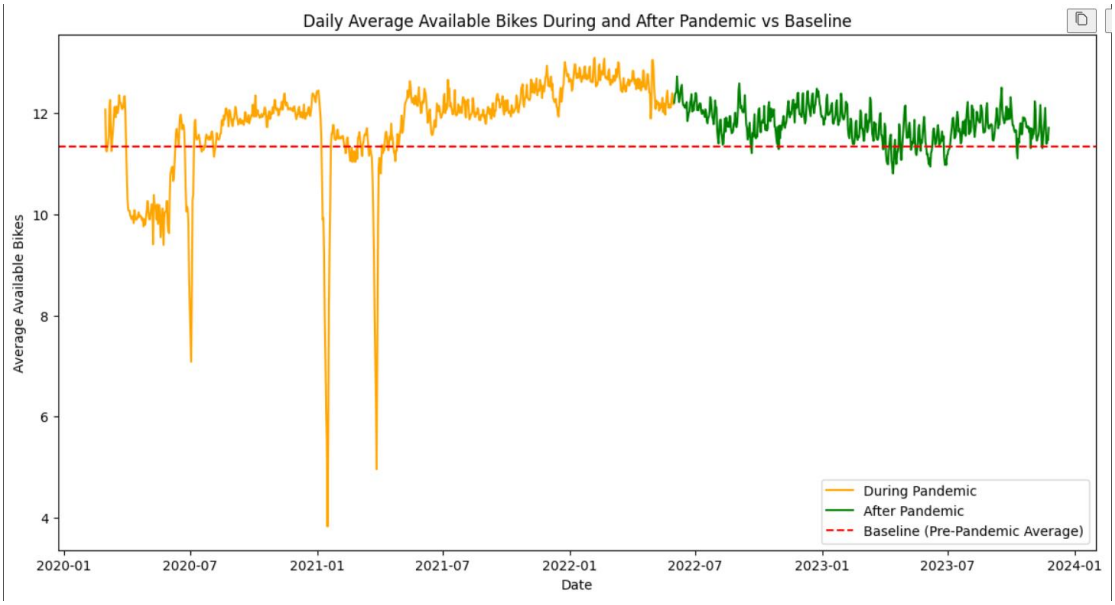

```
MSE for the model before the pandemic: 51.50286870220257
MSE for the model during the pandemic: 37.49676164407443
MSE for the model after the pandemic: 45.14030402279324
```

d) Base line comparison

i. It calculates the baseline number of available bikes from the pre-pandemic data.

ii. It groups the during-pandemic data by date and calculates the daily average of available bikes.

iii. It finds the difference between the daily averages during the pandemic and the baseline to identify trends.

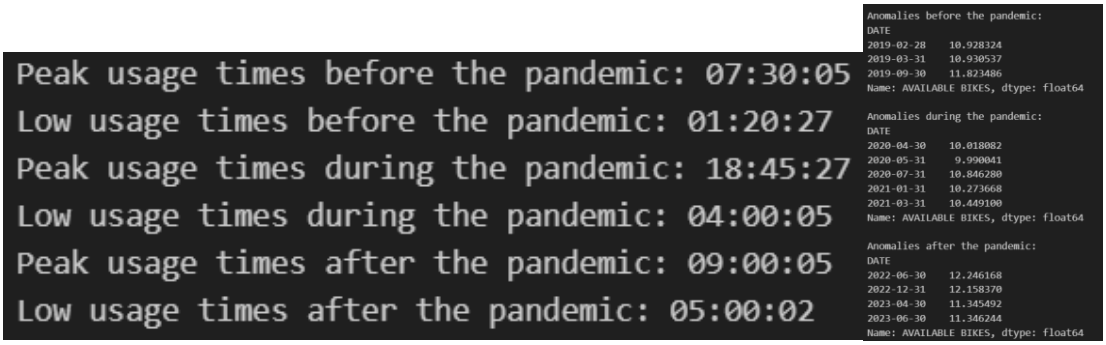


iv. Impact of the Pandemic: There was a noticeable impact of the pandemic on the availability of bikes, with lower averages during the pandemic compared to the baseline. This could be due to increased usage or logistical challenges in bike supply and maintenance during this period.

v. 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.

vi. 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

4. Temperal Dynamic Analysis



a) Peak Usage Time

i. 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.

ii. 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.

iii. 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.

b) Time Series Analysis

i. 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.

ii. 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.

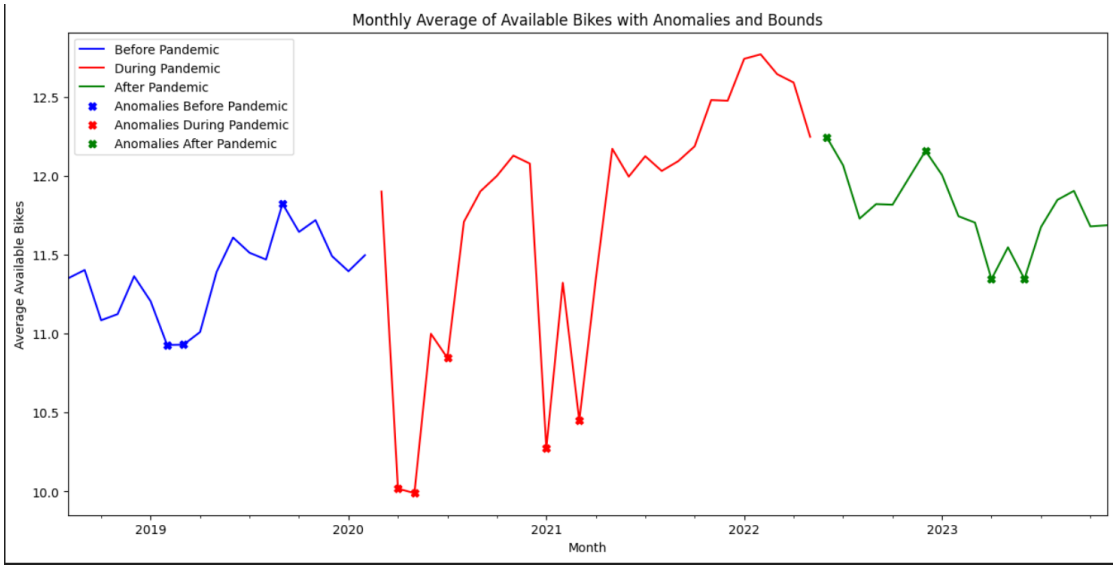
iii. Higher Average Post-Pandemic: After the pandemic, the average availability of bikes is higher than the pre-pandemic levels, which might suggest an increase in the bike fleet or a change in maintenance and supply strategies by bike-sharing providers.

iv. 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.

v. 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.

vi. Pandemic Disruption: The pandemic significantly disrupted these patterns, with usage possibly driven more by lockdowns and social distancing measures than by seasonality.

vii. 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.



c) Geospacial Analysis

i. 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

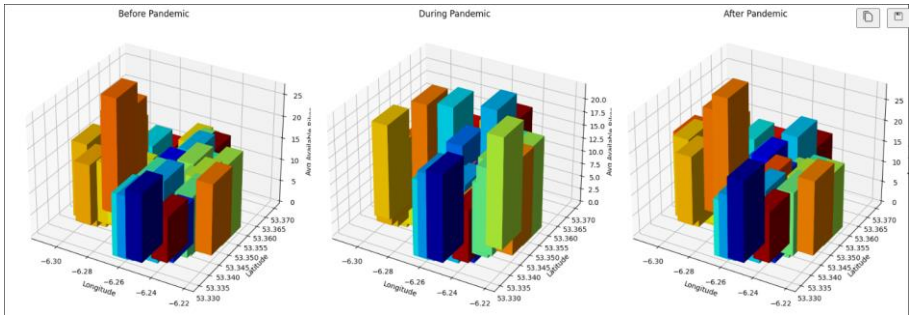
for 3D plotting.

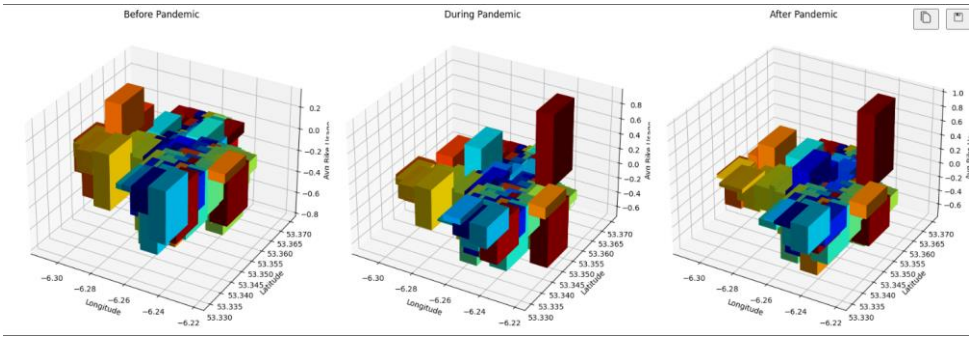
ii. AVAILABLE BIKES

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.
2. 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.
3. 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.

iii. 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.
2. 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.
3. 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.





5. Machine Learning Model Analysis

a) ARIMA (AutoRegressive Integrated Moving Average)

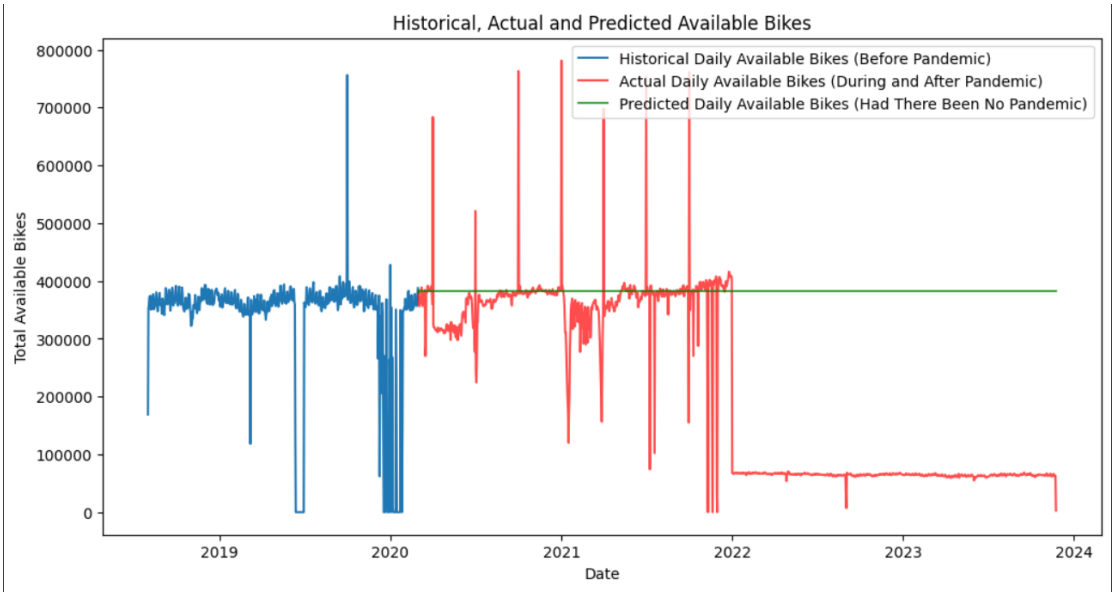
i. To perform predictive modeling using machine learning to estimate what the available bike counts might have been had the pandemic not occurred, we can use a time series forecasting method. One common approach is to use ARIMA (AutoRegressive Integrated Moving Average), which is suitable for time series data without a clear trend or seasonal pattern.

ii. 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.

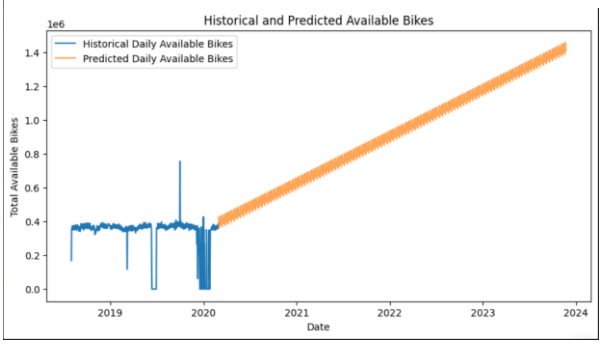
iii. Impact of Pandemic: The actual available bikes during and after the pandemic deviate significantly from the predicted (no-pandemic) trend, highlighting the substantial impact of the pandemic on bike-sharing services.

iv. 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.

v. 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.



b) SARIMA (Seasonal ARIMA)



i. 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.

ii. 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.

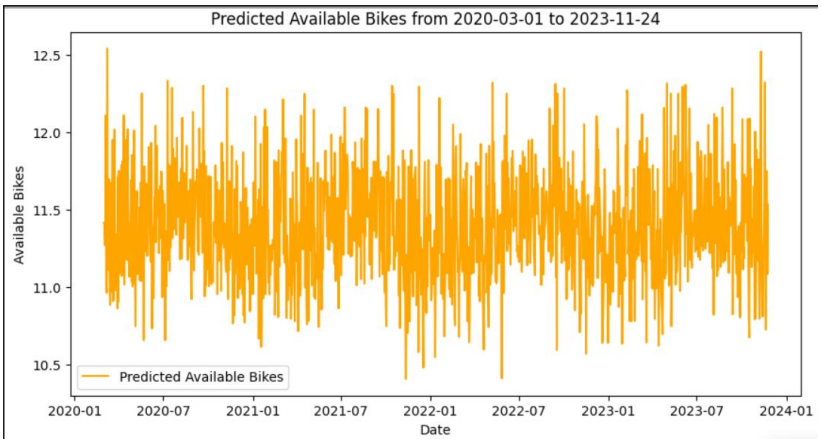
iii. Residual Analysis: The residuals indicate that there may have been factors affecting bike availability that the model did not fully account for, particularly around the pandemic period. This could be due to the model's inability to predict unprecedented events.

c) Random Forest Regressor

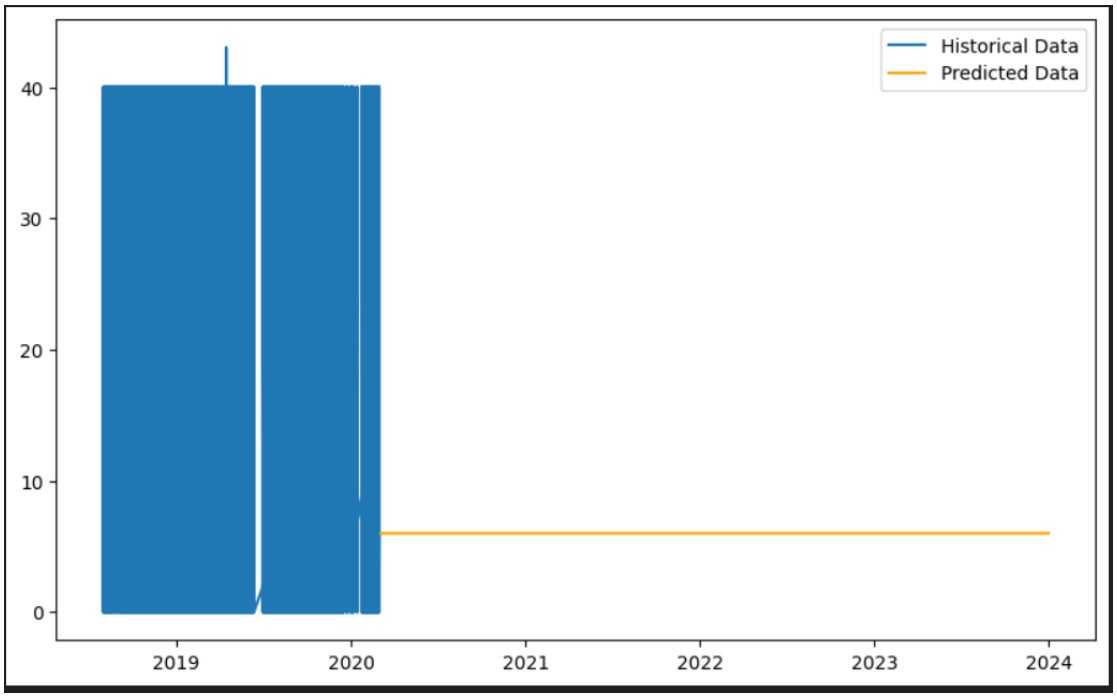
i. To predict the number of available bikes using a more general machine learning model and including additional features like temperature and windspeed, we can use a model like Random Forest Regressor. These models can handle non-linear relationships and are robust to overfitting, making them suitable for this task.

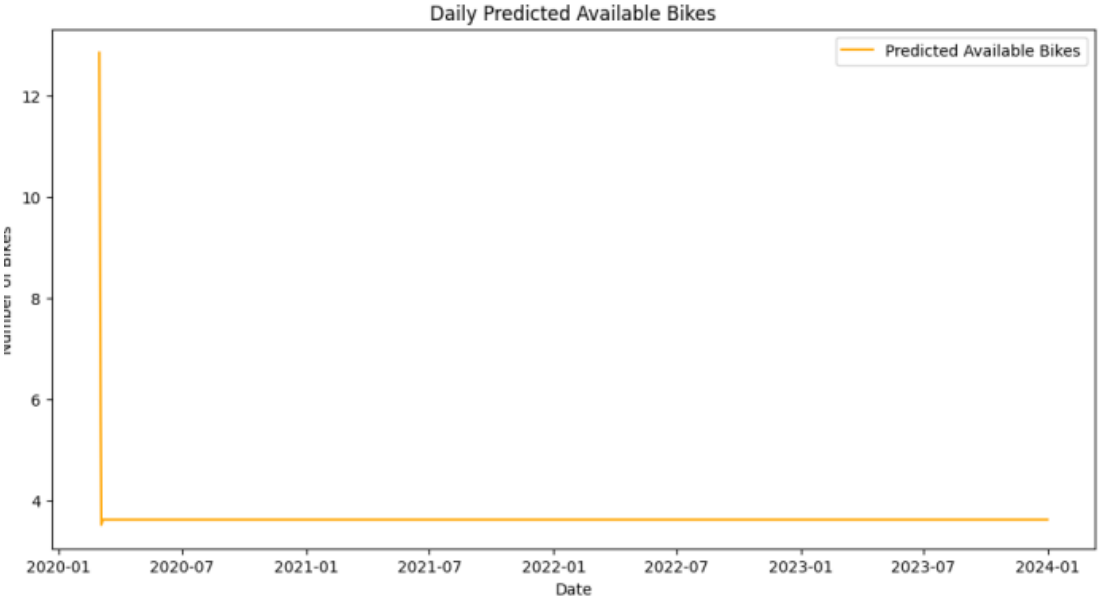
ii. Model Insights: The Random Forest Regressor seems to capture a complex set of patterns that influence bike availability. It is sensitive to fluctuations, possibly reacting to multiple input features such as weather, time of day, day of the week, and other external factors.

iii. 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.



d) LSTM model

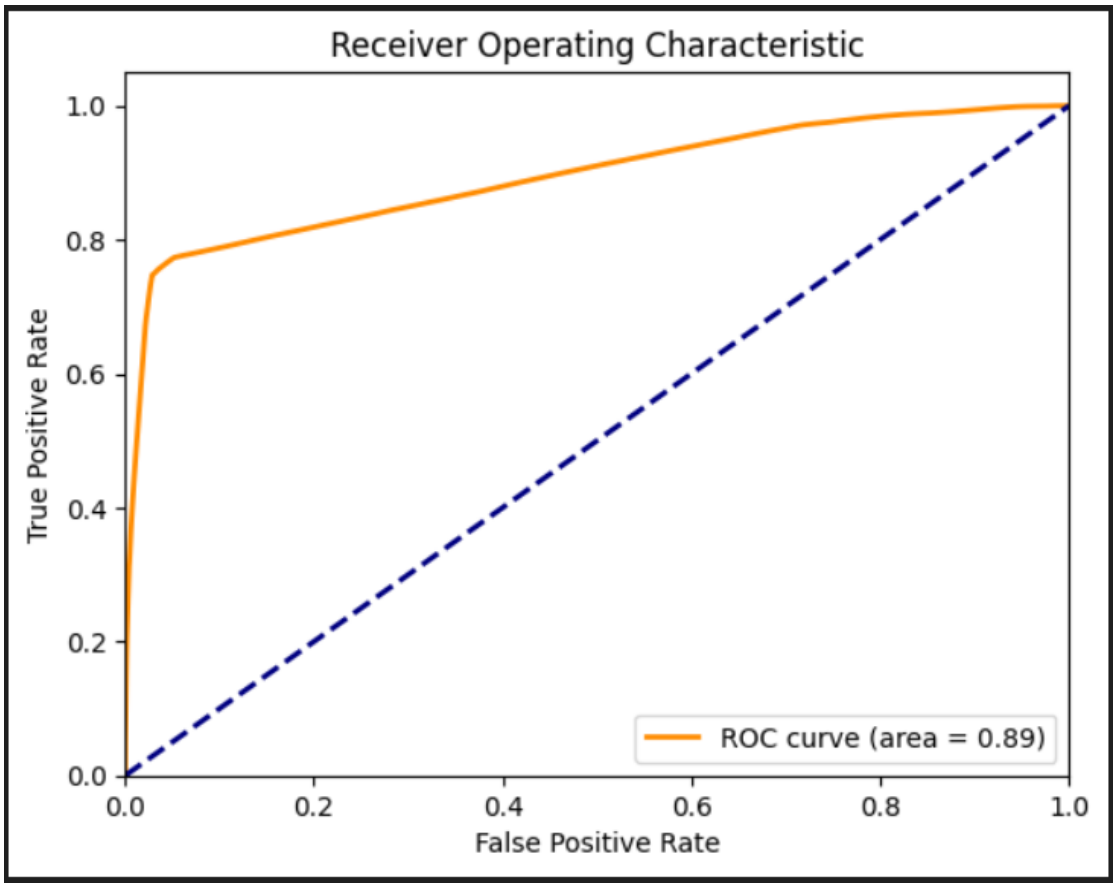




e) ROC curve

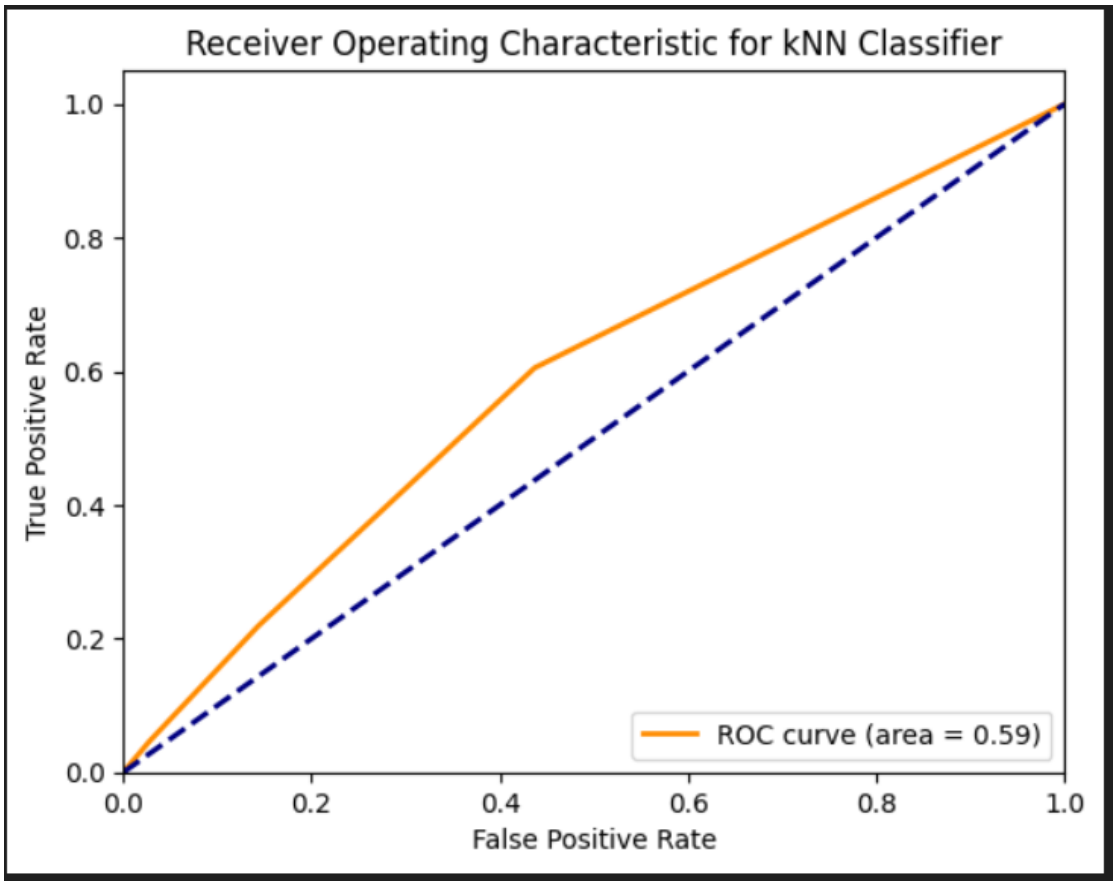
The ROC curve rises sharply towards the top-left corner of the plot, which indicates a high true positive rate (TPR) and a low false positive rate (FPR) at various threshold settings. This is characteristic of a good model performance.

The AUC score is 0.89, which is quite close to 1. An AUC score between 0.5 (no discrimination) and 1 (perfect discrimination) indicates how well the model is distinguishing between the two classes. A score of 0.89 suggests that the model has a very good discriminative ability.



f) k-Nearest Neighbour (kNN) Classifier

Since the curve is closer to the diagonal line of no-discrimination, the kNN classifier does not appear to be much better than random guessing in this case.



6. Integration of Weather Data
   a) 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.

	AVAILABLE BIKES	temp	windspeed	humidity	cloudcover
AVAILABLE BIKES	1.000000	0.005625	-0.000946	0.005628	-0.000319
temp	0.005625	1.000000	-0.135605	-0.219050	0.255830
windspeed	-0.000946	-0.135605	1.000000	-0.083162	0.059569
humidity	0.005628	-0.219050	-0.083162	1.000000	0.254095
cloudcover	-0.000319	0.255830	0.059569	0.254095	1.000000
solarradiation	-0.000218	0.496421	-0.198568	-0.567516	-0.298568

	solarradiation
AVAILABLE BIKES	-0.000218
temp	0.496421
windspeed	-0.198568
humidity	-0.567516
cloudcover	-0.298568
solarradiation	1.000000

	AVAILABLE BIKES	temp	windspeed	humidity	cloudcover
AVAILABLE BIKES	1.000000	0.001285	-0.002815	0.028300	0.018835
temp	0.001285	1.000000	-0.159513	-0.070797	0.087905
windspeed	-0.002815	-0.159513	1.000000	-0.082687	0.116917
humidity	0.028300	-0.070797	-0.082687	1.000000	0.407256
cloudcover	0.018835	0.087905	0.116917	0.407256	1.000000
solarradiation	-0.033140	0.471360	-0.214457	-0.622486	-0.469874

	solarradiation
AVAILABLE BIKES	-0.033140
temp	0.471360
windspeed	-0.214457
humidity	-0.622486
cloudcover	-0.469874
solarradiation	1.000000

	AVAILABLE BIKES	temp	windspeed	humidity	cloudcover
AVAILABLE BIKES	1.000000	-0.002803	0.002961	0.007623	0.006098
temp	-0.002803	1.000000	-0.067637	-0.333191	0.117161
windspeed	0.002961	-0.067637	1.000000	0.000765	0.094946
humidity	0.007623	-0.333191	0.000765	1.000000	0.296322
cloudcover	0.006098	0.117161	0.094946	0.296322	1.000000
solarradiation	-0.008902	0.575837	-0.221839	-0.603145	-0.275108

	solarradiation
AVAILABLE BIKES	-0.008902
temp	0.575837
windspeed	-0.221839
humidity	-0.603145
cloudcover	-0.275108
solarradiation	1.000000

7. Comparative Analysis
    a) Year-over-Year Comparison
        i. 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.
        ii. The increase in the early months of 2020 might reflect growth in the bike-sharing system or increased seasonal availability.
        iii. 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.
        iv. 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.
        v. 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.
        vi. 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.
        vii. 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.

Year-over-Year Comparison Before Pandemic:

Year	2018	2019	2020
Month			
1	NaN	11.205521	11.395518
2	NaN	10.928324	11.497406
3	NaN	10.930537	NaN
4	NaN	11.010450	NaN
5	NaN	11.390821	NaN
6	NaN	11.609088	NaN
7	NaN	11.512660	NaN
8	11.350859	11.469611	NaN
9	11.403601	11.823486	NaN
10	11.084916	11.644651	NaN
11	11.124050	11.719160	NaN
12	11.363840	11.491562	NaN

Year-over-Year Comparison During Pandemic:

Year	2020	2021	2022
Month			
1	NaN	10.273668	12.741410
2	NaN	11.322556	12.769690
3	11.900582	10.449100	12.644188
4	10.018082	11.344252	12.591041
5	9.990041	12.171639	12.247172
6	10.999392	11.995148	NaN
7	10.846280	12.124353	NaN
8	11.708680	12.030952	NaN
9	11.901557	12.092954	NaN
10	11.999366	12.187401	NaN
11	12.128066	12.480162	NaN
12	12.077644	12.475893	NaN

Year-over-Year Comparison After Pandemic:

Year	2022	2023
Month		
1	NaN	12.005263
2	NaN	11.743968
3	NaN	11.703610
4	NaN	11.345492
5	NaN	11.547173
6	12.246168	11.346244
7	12.066643	11.675179
8	11.728893	11.847625
9	11.820892	11.904795
10	11.817114	11.679574
11	11.987965	11.686949
12	12.158370	NaN

    b) Daypart Analysis
        i. 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.
        ii. 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.

iii. 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.

```
Daypart Analysis Before Pandemic:
Hour
0     12.035079
1     12.091080
2     12.102908
3     12.108354
4     12.081459
5     11.891245
6     11.461299
7     10.547418
8     10.426450
9     11.029171
10    11.281739
11    11.091363
12    10.807656
13    10.967480
14    11.212162
15    11.035249
16    10.424176
17    10.358656
18    10.892019
19    11.368229
20    11.649890
21    11.796514
22    11.893398
23    11.953427
Name: AVAILABLE BIKES, dtype: float64
```

```
Daypart Analysis During Pandemic:
Hour
0     11.960452
1     11.989071
2     11.993132
3     11.994030
4     11.974379
5     11.904576
6     11.744299
7     11.525954
8     11.424746
9     11.438601
10    11.414410
11    11.338723
12    11.228346
13    11.187430
14    11.248884
15    11.234765
16    11.181784
17    11.215337
18    11.396959
19    11.597438
20    11.740059
21    11.817286
22    11.867570
23    11.908427
Name: AVAILABLE BIKES, dtype: float64
```

```
Daypart Analysis After Pandemic:
Hour
0     12.196131
1     12.340724
2     12.358522
3     12.364310
4     12.366705
5     12.340623
6     12.105803
7     11.841227
8     11.316435
9     11.237219
10    11.588093
11    11.575052
12    11.460500
13    11.296623
14    11.474871
15    11.527829
16    11.441887
17    11.235459
18    11.281251
19    11.622102
20    11.871987
21    12.024535
22    12.088497
23    12.142364
Name: AVAILABLE BIKES, dtype: float64
```

8. User Behaviour Analysis

a) Cluster Analysis/Usage Patterns

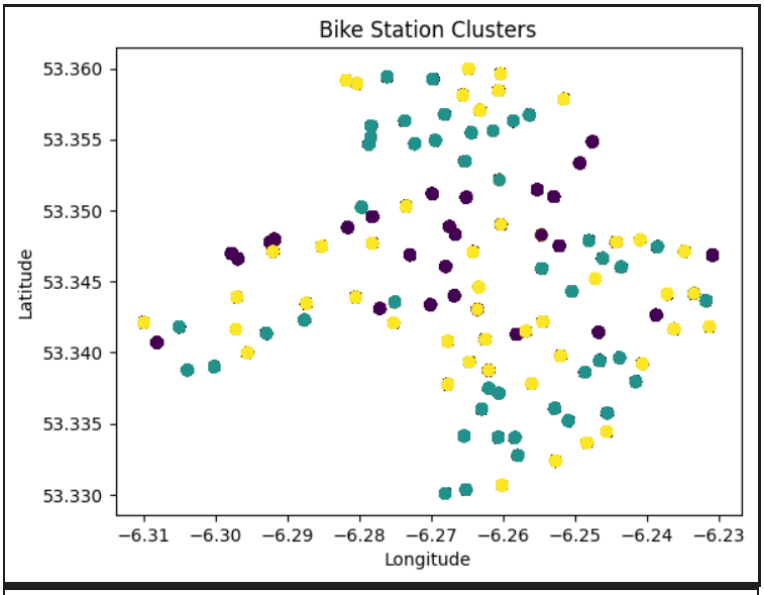
i. From the clustering analysis images, it seems that bike stations have been grouped into clusters based on their geographical location and usage patterns. The colors likely represent different clusters. The consistency in cluster formation across the three periods suggests that geographical location remains a strong determinant of usage patterns, with certain areas consistently having higher or lower bike usage.

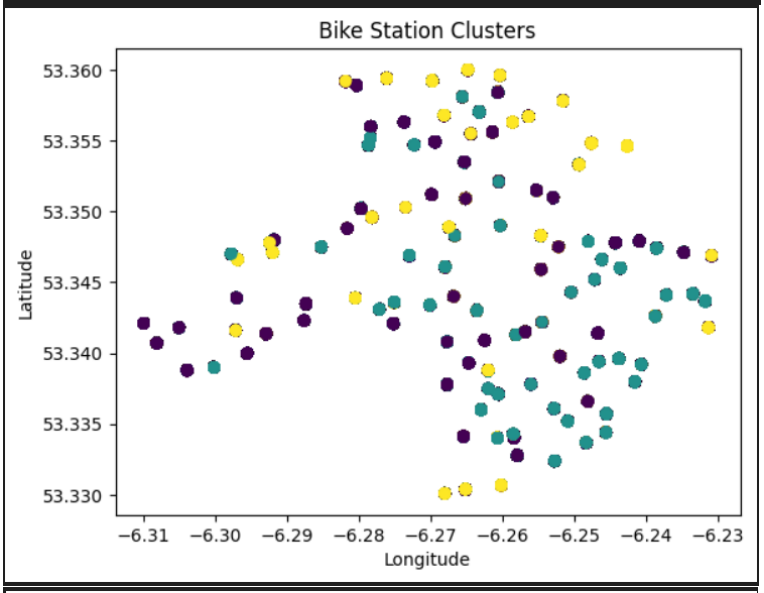
ii. The numerical results indicate the average bike usage values for the respective periods:

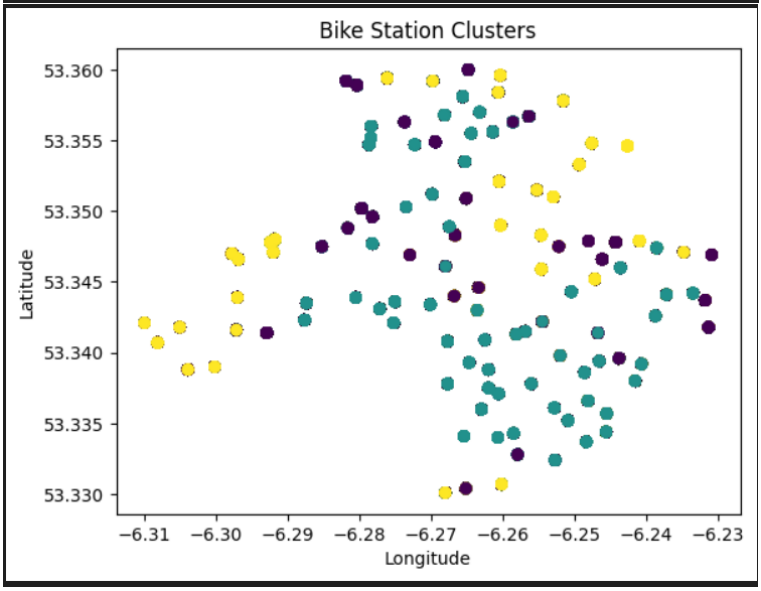
1. Before the pandemic: -0.278
2. During the pandemic: -0.257
3. After the pandemic: -0.239

iii. 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.


Bike Station Clusters
Latitude
Longitude
53.360
53.355
53.350
53.345
53.340
53.335
53.330
−6.31
−6.30
−6.29
−6.28
−6.27
−6.26
−6.25
−6.24
−6.23


Bike Station Clusters
Latitude
Longitude
53.360
53.355
53.350
53.345
53.340
53.335
53.330
−6.31
−6.30
−6.29
−6.28
−6.27
−6.26
−6.25
−6.24
−6.23


Bike Station Clusters
Latitude
Longitude
53.360
53.355
53.350
53.345
53.340
53.335
53.330
−6.31
−6.30
−6.29
−6.28
−6.27
−6.26
−6.25
−6.24
−6.23

(i) What is a ROC curve? How can it be used to evaluate the performance of a classifier compared with a baseline classifier? Why would you use an ROC curve instead of a classification accuracy metric?

1. What is a ROC Curve?

<1> Definition: The ROC curve is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the True Positive Rate (TPR, also known as sensitivity) against the False Positive Rate (FPR, 1 - specificity) at various threshold settings.

<2> TPR (True Positive Rate): It is the ratio of correctly predicted positive observations to the total actual positives.

<3> FPR (False Positive Rate): It is the ratio of incorrectly predicted positive observations to the total actual negatives.

2. Using ROC Curve to Evaluate Performance Compared to a Baseline Classifier

<1> Comparative Analysis: The ROC curve of a classifier can be compared to the ROC curve of a baseline classifier, like a random or a naïve classifier, to understand the relative performance. The area under the ROC curve (AUC) quantifies this comparison. An AUC of 0.5 represents a random or no-skill classifier, so any classifier with an AUC higher than 0.5 is considered better than random. ROC curves allow the examination of the performance of a classifier at various thresholds, which is crucial for understanding how changes in the threshold affect the sensitivity and specificity of the classifier.

3. Why Use an ROC Curve Instead of Classification Accuracy?

<1> Dealing with Imbalanced Classes: Classification accuracy can be misleading when dealing with imbalanced datasets when the number of instances in one class significantly outnumbers the other. ROC curves provide a better analysis in such scenarios because they evaluate the classifier's ability to distinguish between classes.

<2> Threshold Independent: ROC curves provide insights into the performance of a classifier independent of the threshold set for classification. This is particularly useful in scenarios where the threshold might be adjusted based on different contexts or requirements.

<3>More Comprehensive View: While accuracy gives a single metric, ROC curves offer a more comprehensive view of the classifier's performance across different levels of sensitivity and specificity.

(ii) Give two examples of situations where a linear regression would give inaccurate predictions. Explain your reasoning and what possible solutions you would adopt in each situation.

1. Non-Linear Relationships

<1> Situation: Suppose you are trying to predict the growth rate of a certain species of bacteria based on temperature. The relationship between temperature and growth rate might be non-linear, with growth increasing up to an optimal temperature and then decreasing.

<2> Reason for Inaccuracy: Linear regression assumes a constant rate of change between the dependent and independent variables. In this case, the relationship is more complex and doesn't fit a straight line, leading to poor predictions

<3> Possible Solution: Use a non-linear regression model like polynomial regression, which can model the curved relationship. Alternatively, regression models like decision trees or neural networks that can capture non-linearities could be employed.

2. Outliers

<1> Situation: Imagine a dataset for predicting employee salaries based on years of experience. If the dataset includes a few high-level executives with exceptionally high salaries compared to the rest of the employees, these points would be outliers.

<2> Reason for Inaccuracy: Linear regression is sensitive to outliers. The presence of outliers can significantly skew the results of a linear regression model, as it tries to fit these anomalous points, leading to a poor fit for the majority of the data.

<3> Possible Solution: Before building the model, perform outlier detection and either remove or adjust these outliers. Techniques like robust regression, which are less sensitive to outliers, can also be used.

(iii) The term 'kernel' has different meanings in SVM and CNN models. Explain the two different meanings. Discuss why and when the use of SVM kernels and CNN kernels is useful, as well as mentioning different types of kernels.

1. SVM Kernels

<1> Meaning in SVM:

(1) In SVMs, a kernel is a function used to map the input data into a higher-dimensional space. This is done to make it easier to find a linear hyperplane that can separate the data if it's not linearly separable in the original space. Essentially, kernels enable SVMs to solve non-linear classification and regression problems.

<2> Types of SVM Kernels:

(1) Linear Kernel: For linearly separable data.
(2) Polynomial Kernel: Maps data to a polynomial feature space.
(3) Radial Basis Function (RBF) or Gaussian Kernel: Suitable for complex datasets where the decision boundary is irregular.
(4) Sigmoid Kernel: Similar to a two-layer, perceptron neural network.

<3> Usefulness:

(1) Effective in High Dimensional Spaces: SVMs are particularly effective in cases where the number of dimensions is greater than the number of samples, which is a common scenario in fields like genomics and text classification.
(2) Flexibility: Different kernels can be chosen based on the data distribution and the problem at hand.
(3) Robustness: SVMs are known for their robustness, especially to overfitting when the number of features is much larger than the number of samples.
(4) Memory Efficiency

2. CNN Kernels

<1> Meaning in CNN:

(1) In CNNs, a kernel (or filter) is a small matrix used to apply convolution operations to the input data, like an image. It involves sliding the kernel over the input and computing the dot product at each position. This process extracts features like edges,

textures, and other visual elements from the image.

<2> Types of CNN Kernels:

(1) Small-sized Kernels: Commonly used to capture small/local features.

(2) Larger Kernels: For capturing larger patterns, but with increased computational cost.

(3) Specialized Kernels: Designed for specific tasks (e.g., edge detection kernels like Sobel filters).

<3> Usefulness:

(1) Feature Extraction: CNN kernels are adept at automatically learning and extracting relevant features from images, which is critical for tasks like image classification, object detection, etc.

(2) Automatic Feature Learning

(3) Spatial Hierarchy of Features: The architecture of CNNs allows them to learn features with a hierarchy, from simple edges in the early layers to complex patterns in deeper layers. This is beneficial in applications like object detection and segmentation in autonomous vehicles.

(4) Real-time Processing: Once trained, CNNs can process new data (like video frames) in real-time, making them suitable for applications like video analysis, real-time language translation in augmented reality, and more.

(iv) In k-fold cross-validation, a dataset is resampled multiple times. What is the idea behind this resampling i.e. why does resampling allow us to evaluate the generalization performance of a machine learning model. Give a small example to illustrate. Discuss when it is and it is not appropriate to use k-fold cross-validation

1. Idea Behind Resampling in k-Fold Cross Validation

<1> Dividing the Dataset: The dataset is split into 'k' equal-sized subsets or folds.

<2> Model Training and Validation: For each fold, the model is trained on 'k-1' subsets and validated on the remaining subset.

<3> Repeating the Process: This process is repeated 'k' times, with each of the 'k' subsets used exactly once as the validation data.

<4> Aggregating Results: The results from each fold are averaged to provide a single estimation.

<5> The idea behind: To make the most out of the limited data available for both training and validation purposes, which helps in evaluating the model's generalization performance.

Why Resampling Enhances Generalization Evaluation

(1) Utilization of All Data: Each data point gets to be in a validation set exactly once and in a training set 'k-1' times. This ensures that every data point is used for both training and validation, which maximizes the use of available data.

(2) Reducing Bias: Since each data point is used for validation once, the method reduces the bias associated with the random splitting of data into training and test sets.

(3) Variance Reduction: Averaging the results from 'k' different sets reduces the variance associated with a single trial of train/test split.

2. Example

Consider a dataset with 10 data points: [A, B, C, D, E, F, G, H, I, J]. In a 5-fold cross-validation,

this dataset would be divided into 5 subsets:

Fold 1: [A, B], Fold 2: [C, D], Fold 3: [E, F], Fold 4: [G, H], Fold 5: [I, J]

For each fold, the model is trained on 4 subsets and tested on the remaining one. For instance, in the first iteration, the model is trained on [C, D, E, F, G, H, I, J] and tested on [A, B]. This process is repeated for all folds.

3. Appropriate Usage

<1> When to use

(1) Balanced Distribution Across Folds: k-Fold cross-validation is beneficial when you need to ensure that each fold is a good representative of the whole. It's especially useful in cases where the data might have imbalances or varied distributions. Stratified k-fold cross-validation, a variant, ensures that each fold reflects the proportion of each class in the dataset.

(2) Model Selection and Tuning: It's ideal for model selection and hyperparameter tuning. By evaluating a model on different subsets of the data, you can better assess how different models or parameter settings perform across varied data samples.

(3) Small to Medium-Sized Datasets: For datasets that are not extremely large, k-fold cross-validation is a good choice. It maximizes both the amount of data used for training and for validation, which is crucial when data is limited.

(4) Statistically Significant Results: When the goal is to obtain statistically significant model evaluation metrics, k-fold cross-validation provides more reliable and less biased results compared to a single train/test split.

<2> When not to use

(1) Time-Series Data: For time-dependent data, using standard k-fold cross-validation can lead to serious temporal biases because it ignores the time aspect. Techniques like time series cross-validation, where the test sets are always ahead in time compared to the training sets, are more appropriate.

(2) Large Datasets: When datasets are very large, the computational cost of running k-fold cross-validation can be prohibitive. In such cases, a single train/test split, or a limited number of splits, might be sufficient and more practical.

(3) Computational Resources and Time Constraints: If computational resources are limited or there are strict time constraints, running multiple training cycles (as required in k-fold cross-validation) might not be feasible. In such cases, simpler validation techniques might be chosen.

(4) Data Leakage Concerns: In scenarios where there's a risk of data leakage (e.g., when data preprocessing steps like normalization are applied before cross-validation), careful consideration is needed. Data leakage can lead to overly optimistic estimates of model performance.