**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | MSC\_DA\_Repeat |
| **Lecturer Name:** | Marina Iantorno/Taufique Ahmed  Sam Weiss,  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Emre Arslan |
| **Student Number:** | 2023184 |
| **Assessment Due Date:** |  |
| **Date of Submission:** | 05.01.2025 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

ABSTRACT

**COMPARISON OF BIKE USAGE IN DUBLIN AND SAN FRANCISCO**

**( OCTOBER-DECEMBER 2020 ) AND VISUALIZATION**

**CCT COLLEGE DUBLIN**

**DATA ANALYTICS**

**LECTURER**

**David McQuaid**

**Muhammad Iqbal**

**Sam Weiss**

**Marina Iantorno/Taufique Ahmed**

**Abstract**

**Problem**

This project aims to compare bike usage in Dublin and San Francisco from October to December 2020. The dataset includes monthly bicycle usage statistics for both cities, as well as information on bicycle preferences (Classic or E-Bike) and community satisfaction. The primary objective is to highlight the similarities and differences in bike usage between the two cities, providing valuable insights for other cities looking to develop sustainable transportation policies.

To achieve this, we will employ a range of machine learning algorithms, including supervised, unsupervised, and semi-supervised learning models, to identify the most accurate approach for predicting bike usage in Dublin and San Francisco. Before implementing these models, we will conduct a comprehensive analysis of the bike usage data using various analytical techniques.

The analysis will reveal key differences and similarities in bicycle usage patterns between the two cities. These findings will be used to assess their potential impact on urban transportation policies, offering a practical framework for other cities aiming to enhance their sustainable transport initiatives.

**Keywords:** Transportation System , Bicycle Usage , San Francisco , Dublin , Machine Learning , Supervised Learning , Unsupervised Learning , Sustainable Transportation , Bicycle Preferences (Classic, E-Bike) , Data Visualization , Sentiment Analysis

CONTENTS

**Page**

[ABSTRACT 2](#_Toc185195490)

[**1.** **INTRODUCTION** 4](#_Toc185195491)

[**2. CONCEPTS** 5](#_Toc185195492)

[2.1. Classic Bike & E-Bike 5](#_Toc185195493)

[2.2 Machine Learning 5](#_Toc185195494)

[2.3 Data Analysis 8](#_Toc185195495)

[2.4 JSON and CSV File 8](#_Toc185195496)

[**3.** **MATERIALS and METHODS** 8](#_Toc185195497)

[3.1 Choosing a Subject and Finding Data 8](#_Toc185195498)

[**4. RESULTS** 18](#_Toc185195499)

[REFERENCES 18](#_Toc185195500)

# **INTRODUCTION**

Bicycles are increasingly viewed as a sustainable solution for urban transportation, particularly in densely populated cities struggling with traffic congestion. Encouraging bicycle use not only alleviates traffic but also promotes healthier lifestyles and supports environmental protection. Micromobility options, such as bicycles and e-bikes, offer low-cost, eco-friendly, and health-beneficial transportation alternatives. Bicycle path and sharing projects further enhance bike usage by creating safe, accessible road networks, reducing dependence on automobiles, and encouraging a broader user base to adopt cycling.

In response to the growing emphasis on sustainable urban mobility, many modern cities are incorporating Internet of Things (IoT) technology into cycling infrastructure. This has prompted urban planners and local governments to prioritize cycling as a key element of sustainable transport. In this context, this study aims to analyze and compare bike usage patterns in Dublin, Ireland, and San Francisco, USA, two cities with distinct urban and cultural characteristics. The study focuses on bicycle usage data from October to December 2020, a period chosen to avoid anomalies caused by the COVID-19 pandemic, thereby providing a reliable basis for comparison.

The analysis examines monthly shared bike usage, bicycle preferences (Classic or E-Bike), and community satisfaction in both cities. Data for Dublin is sourced from the Dublin City Council (specifically the Moby Bike dataset), while data for San Francisco is obtained from DataSF. Identical timeframes were used to facilitate a fair comparison. The study also incorporates bike usage data from different years (May, June, and July) to provide a broader perspective on seasonal trends.

The project employs a range of data visualization tools (e.g., pie charts, bar plots, line plots) and statistical measures (e.g., Bernoulli Distribution, Binomial Distribution, mode, median, variance). Advanced machine learning techniques are also utilized to predict bike usage and identify patterns. Models used include:

* Classification Models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Naive Bayes, Decision Tree, and Random Forest.
* Clustering Models: K-means Clustering.
* Dimensionality Reduction: Principal Component Analysis (PCA).
* Evaluation Methods: Confusion Matrix and K-Fold Cross-Validation.

The comparative analysis highlights the similarities and differences in bike usage between the two cities, offering valuable insights for policymakers and urban planners aiming to develop more sustainable transportation systems. By identifying effective models for predicting bike usage, this study provides a framework for other cities seeking to encourage cycling as a viable alternative to automobile use, ultimately contributing to healthier lifestyles, reduced congestion, and environmental sustainability.

# **CONCEPTS**

1. **Classic Bike & E-Bike**

Bicycles play a vital role in urban transportation in over 200 cities worldwide, with Europe taking significant strides in promoting cycling as a key mode of transport. European countries such as the Netherlands, Denmark, Germany, and Belgium have some of the highest rates of bicycle usage.

In Japan, approximately 15% of the population commutes to work by bicycle, with an impressive 10 million bicycles sold annually. Similarly, in Belgium, a country with a population of 10.83 million, there are approximately 5.2 million bicycles in use.

The Netherlands is often regarded as a global leader in bicycle usage. An estimated 84% of Dutch citizens own at least one bicycle, and with a population of 17 million, the Netherlands has a staggering 23 million bicycles. The ratio of bicycles to cars is approximately 3:1, and bicycles account for 48% of vehicle traffic in urban areas. The average Dutch citizen cycles 3 km daily, the highest figure among European countries. This success is attributed to the country’s safe and efficient cycling infrastructure and the widespread adoption of bicycles as a primary mode of transport across all demographics.

In England, cycling policies aim to increase bike usage through several initiatives, such as:

* Bicycle parking at bus stations and car parks,
* Converting unused railways and canals into bicycle-friendly routes, and
* Developing "green roads", which are strategic cycling corridors.

New York City is another notable example, with over 100,000 people cycling daily, and this number continues to rise. The city has declared May as "Bicycle Month", during which various events are organized to promote cycling. Bicycles in New York are not only seen as a mode of transport but also as a recreational and fitness tool. The city has more daily bicycle commuters than any other U.S. city. To support this growing demand, over 50 km of dedicated bicycle paths have been established in the Manhattan area alone, encouraging more people to adopt cycling as a sustainable transportation option.

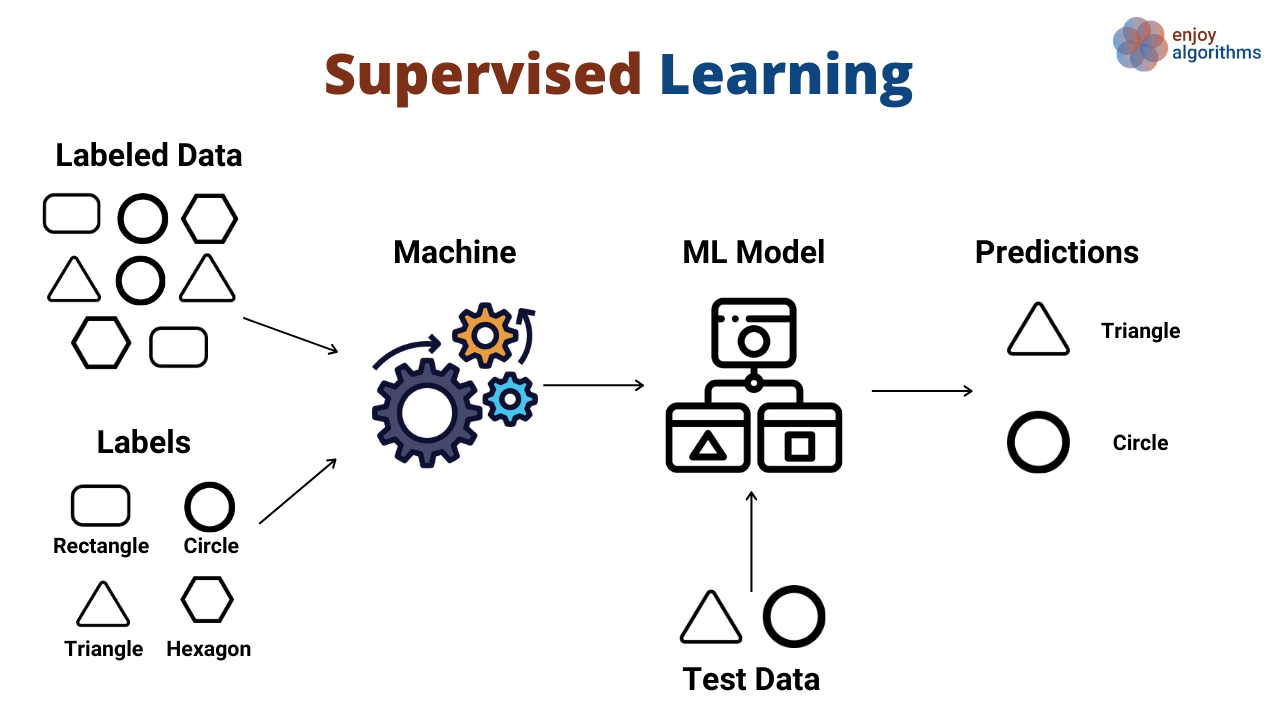
1. **Machine Learning**

Intelligence is at the heart of what makes us human, but the methods we use for identifying, talking about and valuing human intelligence are impoverished. We invest artificial intelligence (AI) with qualities it does not have and, in so doing, risk losing the capacity for education to pass on the emotional, collaborative, sensory and self-effective aspects of human intelligence that define us. (Luckin, 2022) . Machine learning can be applied in diverse ways. For instance, face recognition technology is used in security applications for access control. The system must identify or verify individuals by comparing their faces to predefined categories or assessing whether someone is who they claim to be. Effective face recognition systems must handle various factors such as lighting, facial expressions, and accessories, making it crucial for these systems to learn and adapt to different conditions.

Overall, the field of machine learning has evolved from theoretical exploration to practical application, addressing a wide range of challenges and continuously improving as new techniques and data become available.

**2.2.1 Supervised Learning**

Supervised learning is one of the most widely used ML algorithms. In supervised learning, the training data you use are already labeled. These training data are used to infer a learning algorithm or mapping function from the input variable (X) to the output variable (Y). The correct answers or desired outputs (labels), here, are already known, given a labeled set of input–output pairs, N is simply the number of training examples. The training input is a d-dimensional vector or numbers also known as features, or attributes. The input can be an image, an email message, a time series, a molecular shape, or a graph. The output , also known as a response variable, is a categorical or nominal variable for a classification problem or real value for a regression problem. Classification algorithms and regression techniques are two types of supervised learning widely used to develop predictive models. (Knowledge Discovery in Big Data from Astronomy and Earth Observation, 2020)



Pic 1.0

* **Decision Tree**

Decision Trees are a supervised learning method used for classification and regression. They work by learning simple decision rules from data features to predict the target variable. Essentially, a decision tree splits data into branches to make predictions, creating a piecewise constant approximation.

* **Support Vector Machine**

Nonlinearly transforming the input space into a high-dimensional feature space, the SVM algorithm then searches the high-dimensional feature space for the best linear boundary hyperplane. Linearly separable datasets are ideal for the SVM classification technique since it relies on minimizing empirical risk to get optimal results.

To this end, it is necessary to choose a boundary plane that not only effectively partitions the data into the two classes, but also maximizes the classification cycle or gap between the classes.

Improving the support vector machine's ability to generalize requires increasing the size of its classification interval . (Pisner and Schnyer, 2020)

* **K-Means Clustering**

K-Means is a fundamental unsupervised learning algorithm used to solve clustering problems. It works by classifying a given dataset into a predefined number of clusters, denoted as k. The process begins by initializing k cluster centers, which are strategically placed to maximize the distance between them. Proper placement of these centers is crucial, as it significantly impacts the final clustering outcome.

Once the centers are established, each data point in the dataset is assigned to the nearest cluster center. After all points have been assigned, the first iteration is complete, and an initial grouping is formed. The next step involves recalculating the cluster centers, also known as centroids, which are computed as the mean position (barycenter) of all points within each cluster. This process repeats iteratively until the positions of the centroids stabilize, resulting in well-defined clusters.

* **K Nearest Neighbor**

The K-Nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows (Mahesh, 2019).

* **Naive Bayes**

Naive Bayes is a classification algorithm based on Bayes' Theorem, which assumes that the predictors (features) are independent of one another. In simple terms, it means that the presence or absence of a specific feature in a class does not influence the presence of any other feature.

This method is widely used in text classification, where it excels in tasks such as spam detection, sentiment analysis, and document categorization. Naive Bayes operates by calculating the conditional probability of each class given the input features and assigning the input to the class with the highest probability. It is especially effective when working with large datasets and is known for its simplicity, efficiency, and speed.

**2.2.2 Semi-Supervised Learning**

In contrast to other machine learning algorithms, semi-supervised learning is more effective,

making it more widely applicable in real-world scenarios. To understand the link between the

different features and to produce the classifier model for predictive modeling, its core premise is

to combine labeled or otherwise definitive data with unlabeled data. Both regression and

classification issues may be solved using the approach . (Van Engelen and Hoos, 2020)

**2.2.3 Unsupervised Learning**

Unsupervised learning algorithms identify patterns in data without using labeled outcomes. Unlike

supervised learning, which relies on known output values for training, unsupervised learning explores

the data to uncover its underlying structure and relationships, as there are no predefined labels or

values to guide the process.

1. **Data Analysis**

Data analysis is the process of extracting valuable insights and conclusions from raw data through various steps, including collection, cleaning, processing, and examination. This involves using statistical methods, data mining techniques, and visualization tools to interpret and analyze the data. Data analysis serves a range of purposes, such as assessing business performance, shaping marketing strategies, enhancing operational efficiency, and supporting scientific research. The process starts with gathering data, followed by cleaning it to remove errors and converting it into a suitable format. The data is then thoroughly analyzed, and the findings are presented through graphs, tables, and reports. These results provide key insights for decision-makers and stakeholders, enabling more informed and effective decisions. Data analysis is essential across numerous industries and fields today.

1. **JSON and CSV File**

JSON (JavaScript Object Notation) and CSV (Comma-Separated Values) are both popular formats used for data storage and exchange.

* **JavaScript Object Notation (JSON):**  
  JSON is a lightweight data format designed for easy reading and writing by humans, and simple parsing and generation by machines. It is derived from a subset of the JavaScript programming language and is often used for data exchange between systems.
* **Comma-Separated Values (CSV):**  
  CSV is a simple format commonly used to store tabular data in a spreadsheet or database. Each row in the CSV file corresponds to a row in the table, and the individual fields within that row are separated by commas.

# **MATERIALS and METHODS**

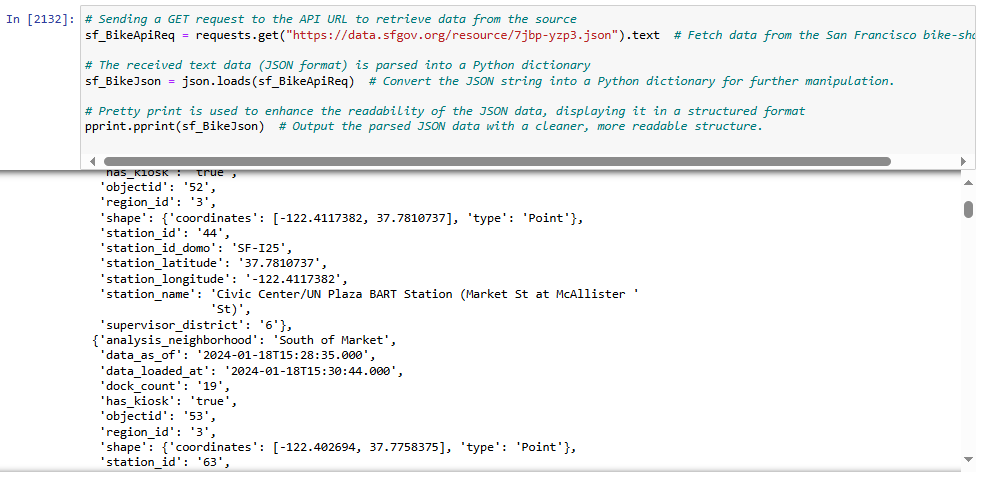
**3.1 Choosing a Subject and Finding Data**

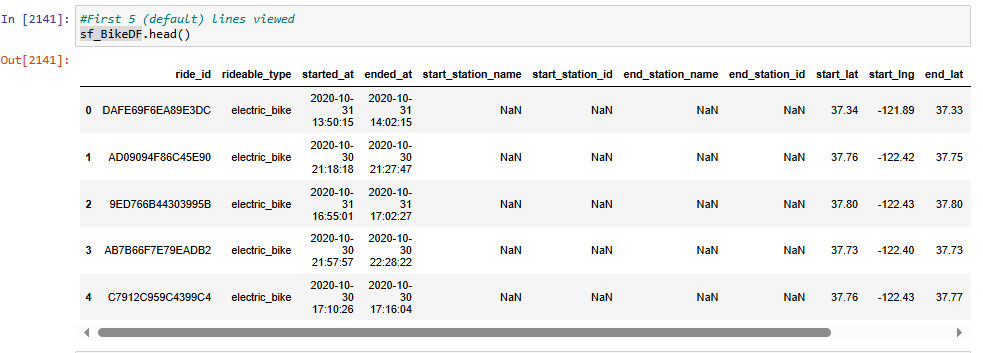
The first step in starting my project was to find relevant data. During my research, I discovered that many countries are actively promoting cycling, but they were not sharing their data publicly. Those that did share data often sold it. After an extensive search, I found that San Francisco had publicly available data, so I decided to focus on this city (https://s3.amazonaws.com/baywheels-data/index.html).

Finding suitable data for Dublin was a bit easier. I looked for a dataset that could be compared to the San Francisco dataset on “https://data.gov.ie/organization/dublin-city-council?q=bike&sort=score+desc%2C+metadata\_modified+desc” the official government database portal. To make a valid comparison between the two cities, I needed to examine similar features. I found that the Moby dataset for Dublin was a good match, so I chose it.

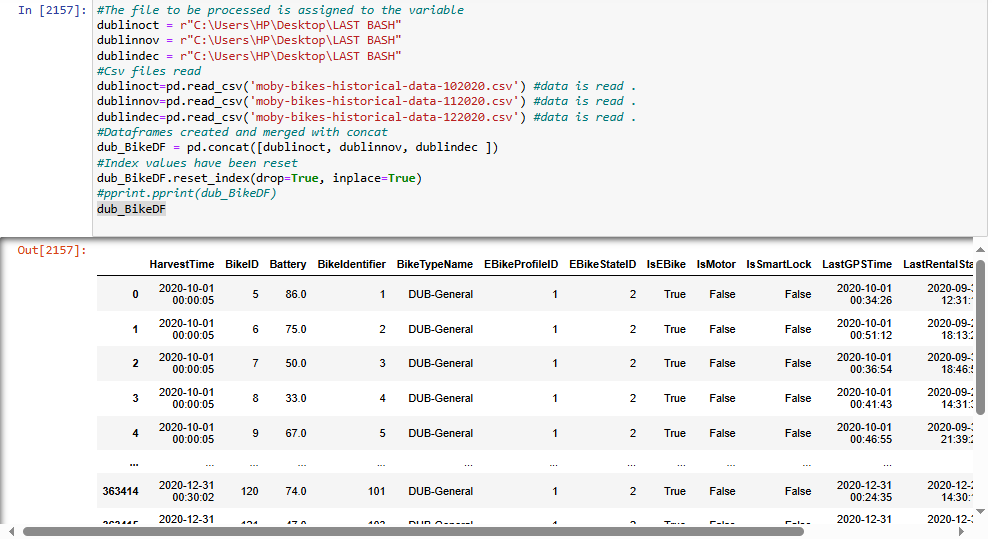
However, I noticed a key difference between the two datasets: they covered different time periods., I aimed to align the data periods. I soon realized that the Dublin data was not available during the COVID-19 period, which further complicated the comparison. Ultimately, I decided to focus on data from Oct to January 2021 .

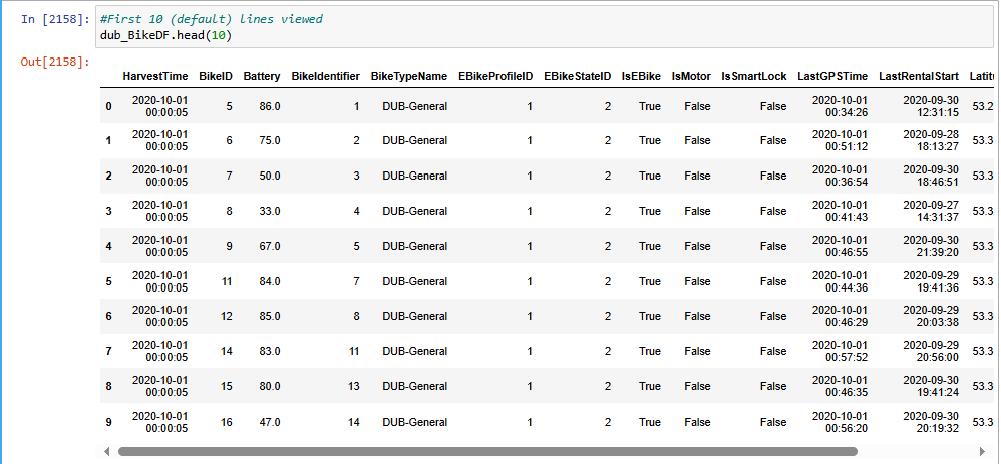
**San Francisco Dataset**





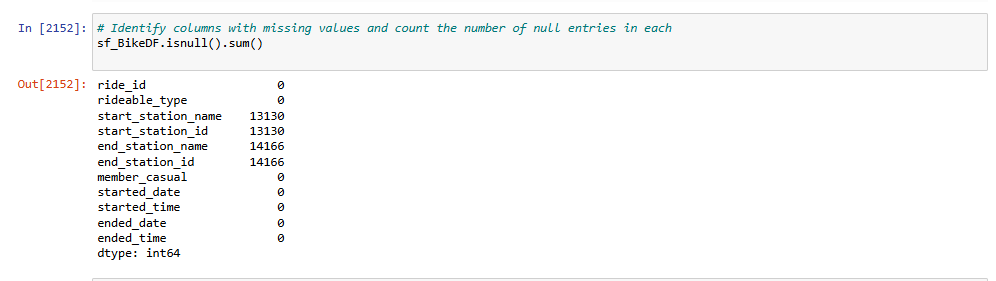
**Dublin Dataset**

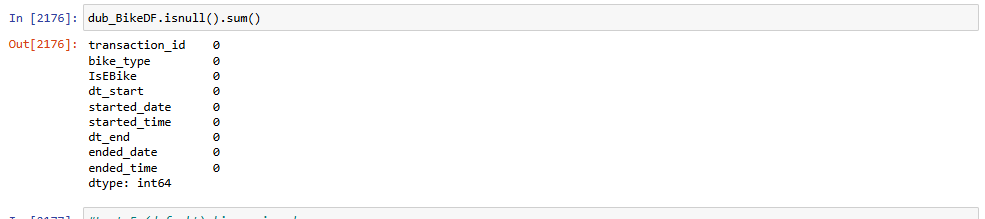




**3.2 Starting to Write Code in Jupyter Notebook**

Both datasets had missing data. However, after removing some unnecessary columns from the Dublin dataset, it no longer had any missing values. In contrast, the San Francisco dataset still had missing data in columns such as 'start\_station\_name' and 'end\_station\_id'. These columns will be dropped in the following steps.





Rows with NaN or NULL values were wiped out with below codes:

“dub\_BikeDF.dropna(inplace=True)”

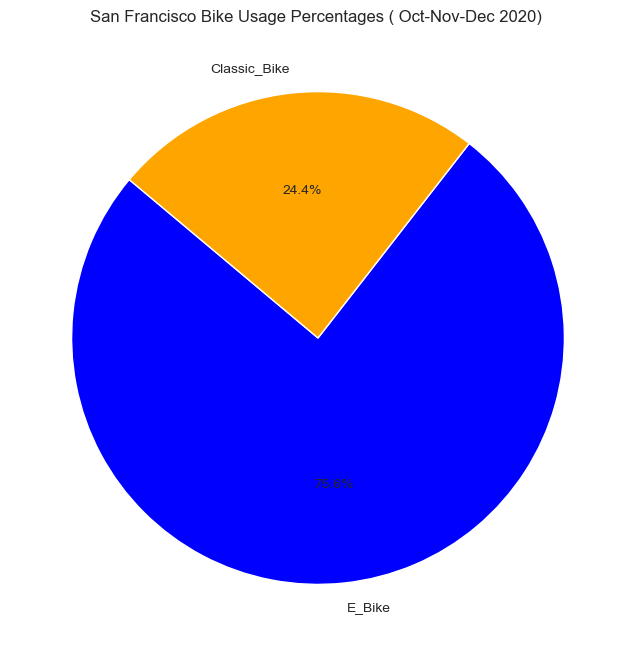
“sf\_BikeDF.dropna(inplace=True)”

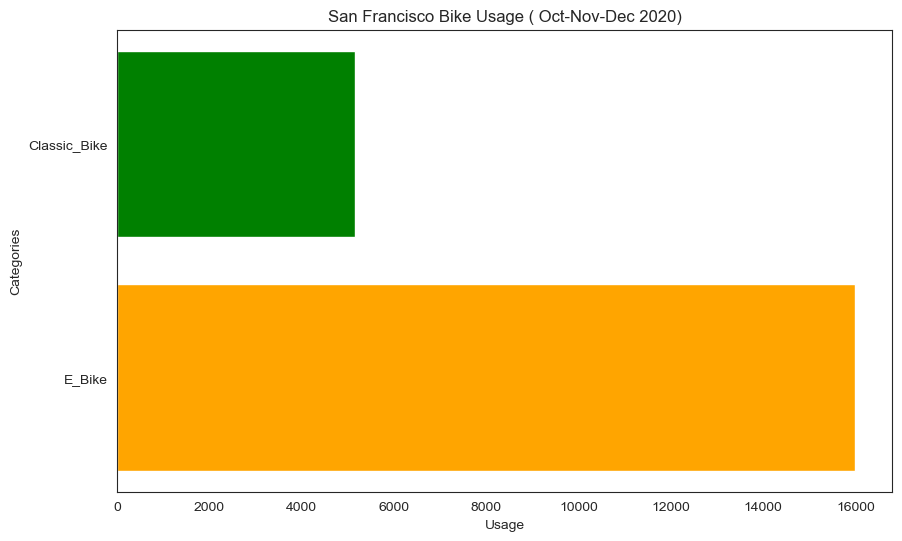
To ignore the missing values, we used the code “dub\_BikeDF[dub\_BikeDF["station\_id"].notna()]”. After running the code, we verified that the null values had been removed.

Subsequently, we explored the data further using functions like “describe” and “unique” to gain more insights.

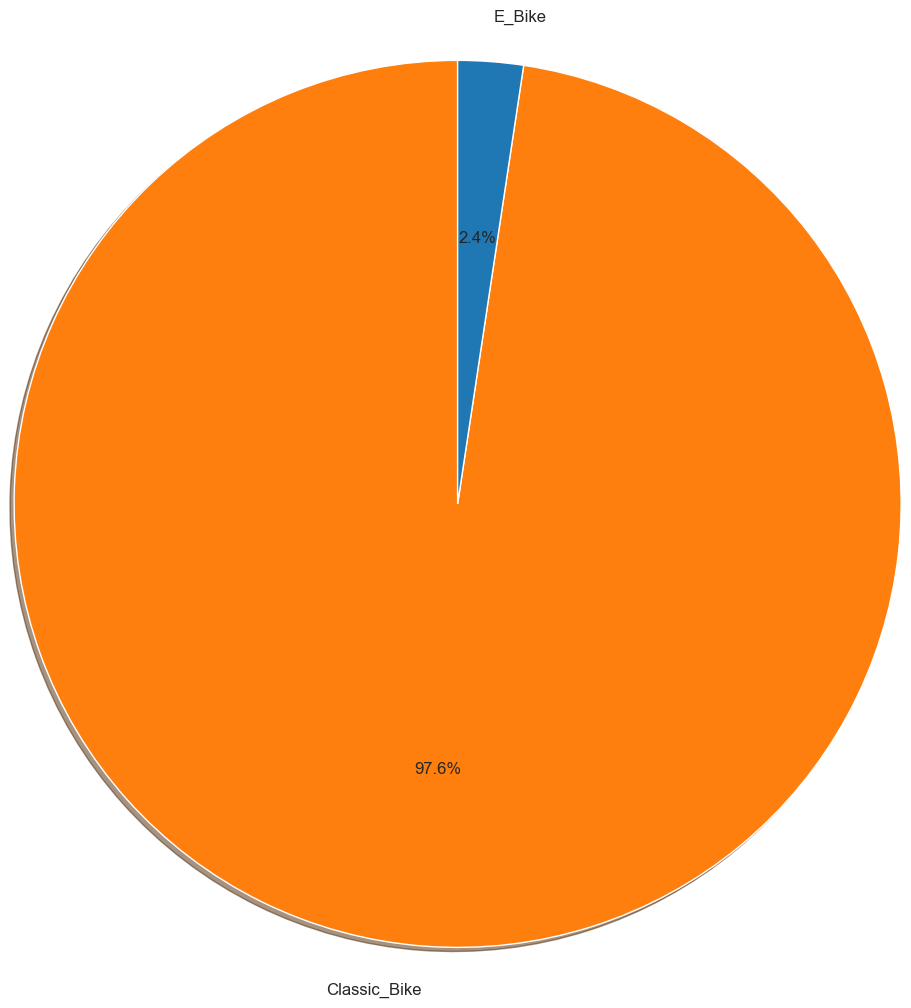
In the visualization step, we first used pie charts and bar charts to analyze the data. We examined the rates of classic bikes and e-bikes for each city. It was observed that e-bikes are quite popular in San Francisco, while their usage is much lower in Dublin. On the other hand, classic bikes are more commonly used in Dublin. The details of these findings are shown below.

San Francisco,

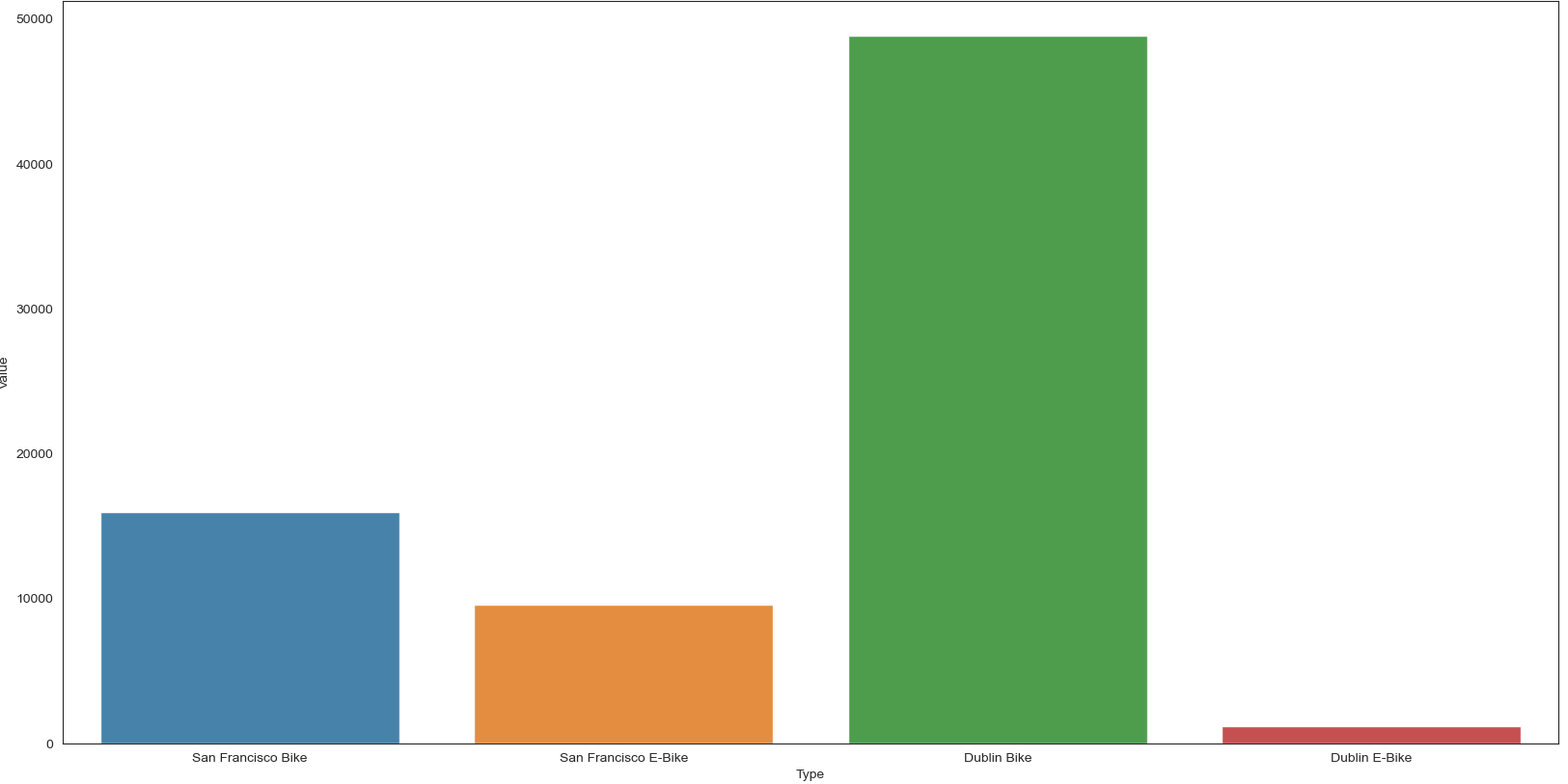




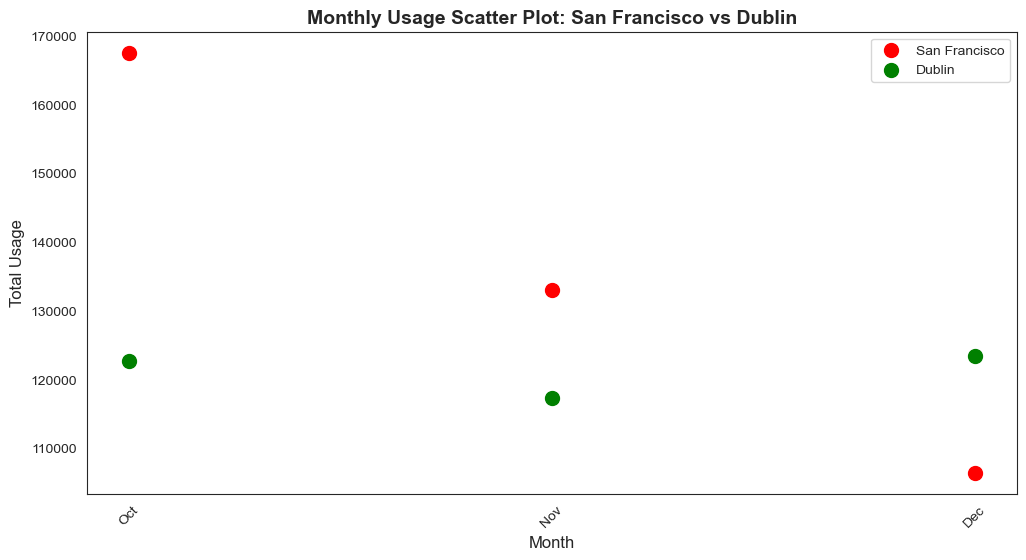
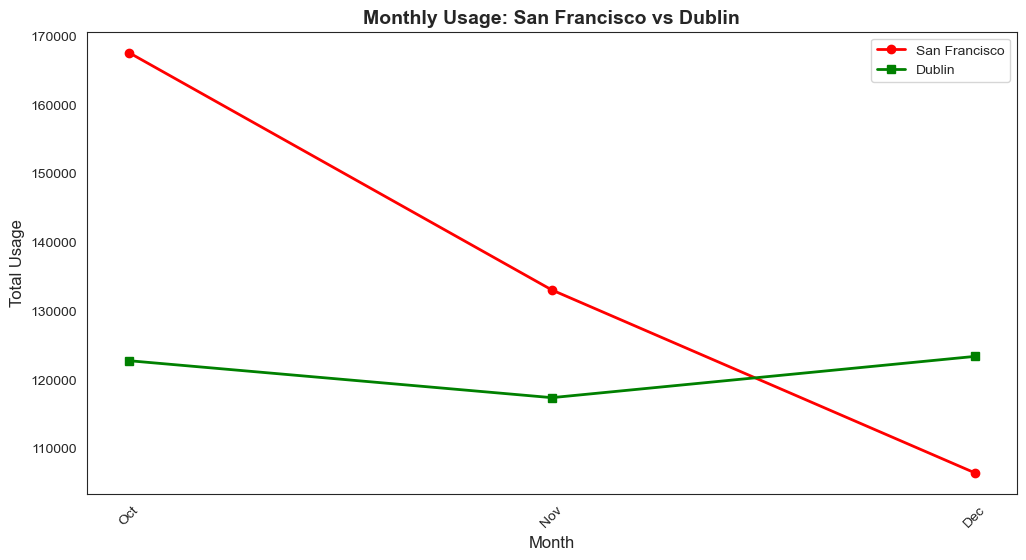
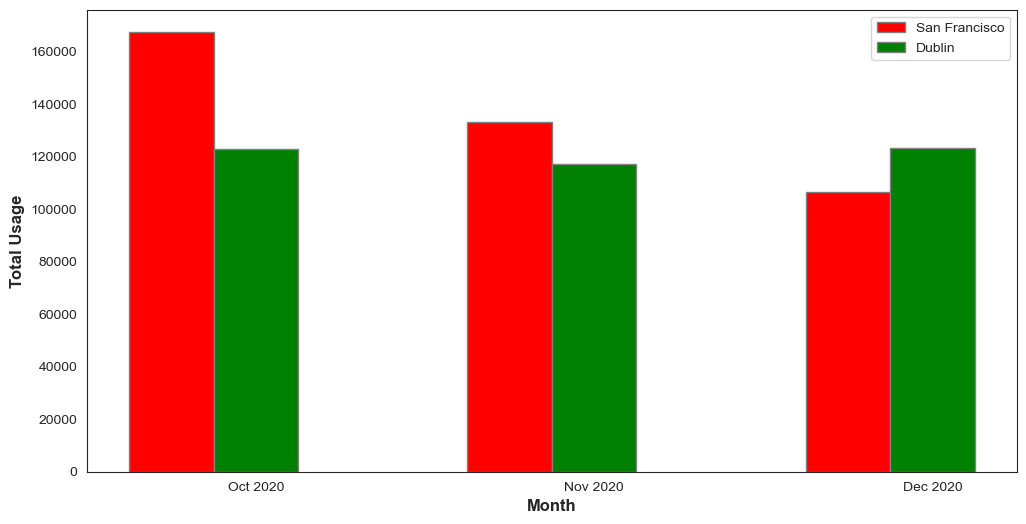
For Dublin,



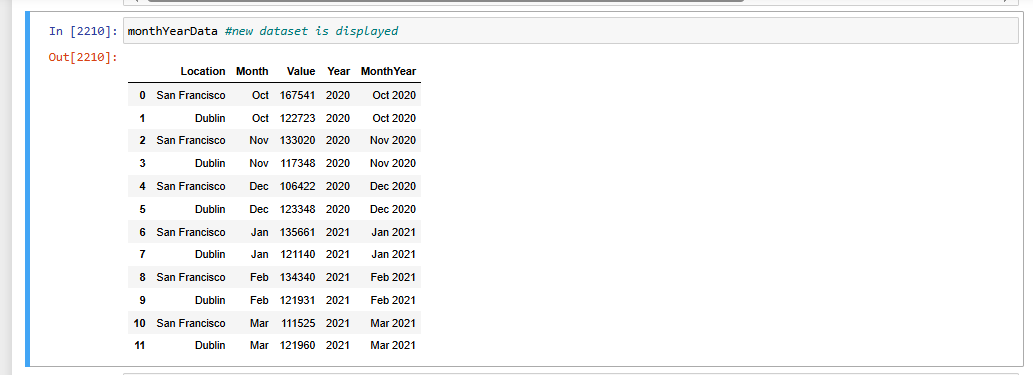
San Francisco and Dublin Bikes Total ( Oct-Dec 2020 ) results visualized with bar plot,

****

San Francisco and Dublin Bikes Monthly ( Oct-Dec 2020 ) results visualized



Enhance the analysis using machine learning techniques and generate more meaningful values, the upcoming months (January, February, March) will be forecasted using statistical methods. Specifically, we will employ methods such as mean, mode, and linear trend analysis to estimate the data for these months. By applying these statistical techniques, we aim to provide a broader set of values, allowing for more accurate predictions and insights. The ultimate goal is to improve the detection and classification of X and Y test values, which will help in refining the predictive model and enhancing its performance. These methods will facilitate a deeper understanding of the data, ensuring that future values are generated more precisely, which will contribute to better decision-making and model testing.



Location Month Value Year MonthYear Above Average Values

0 San Francisco Oct 167541 2020 Oct 2020 1

1 Dublin Oct 122723 2020 Oct 2020 0

2 San Francisco Nov 133020 2020 Nov 2020 1

3 Dublin Nov 117348 2020 Nov 2020 0

4 San Francisco Dec 106422 2020 Dec 2020 0

5 Dublin Dec 123348 2020 Dec 2020 0

6 San Francisco Jan 135661 2021 Jan 2021 0

7 Dublin Jan 121140 2021 Jan 2021 0

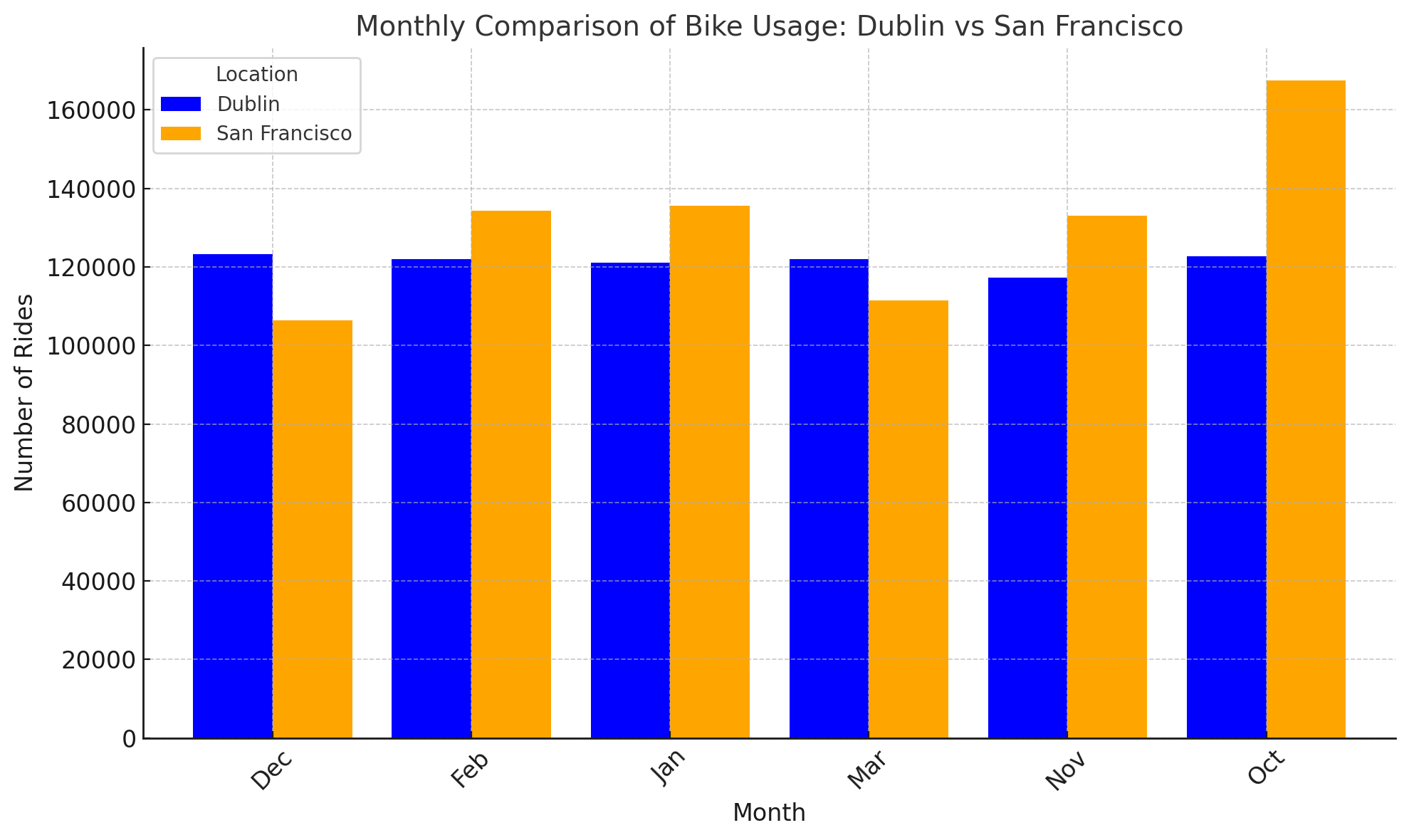
8 San Francisco Feb 134340 2021 Feb 2021 0

9 Dublin Feb 121931 2021 Feb 2021 0

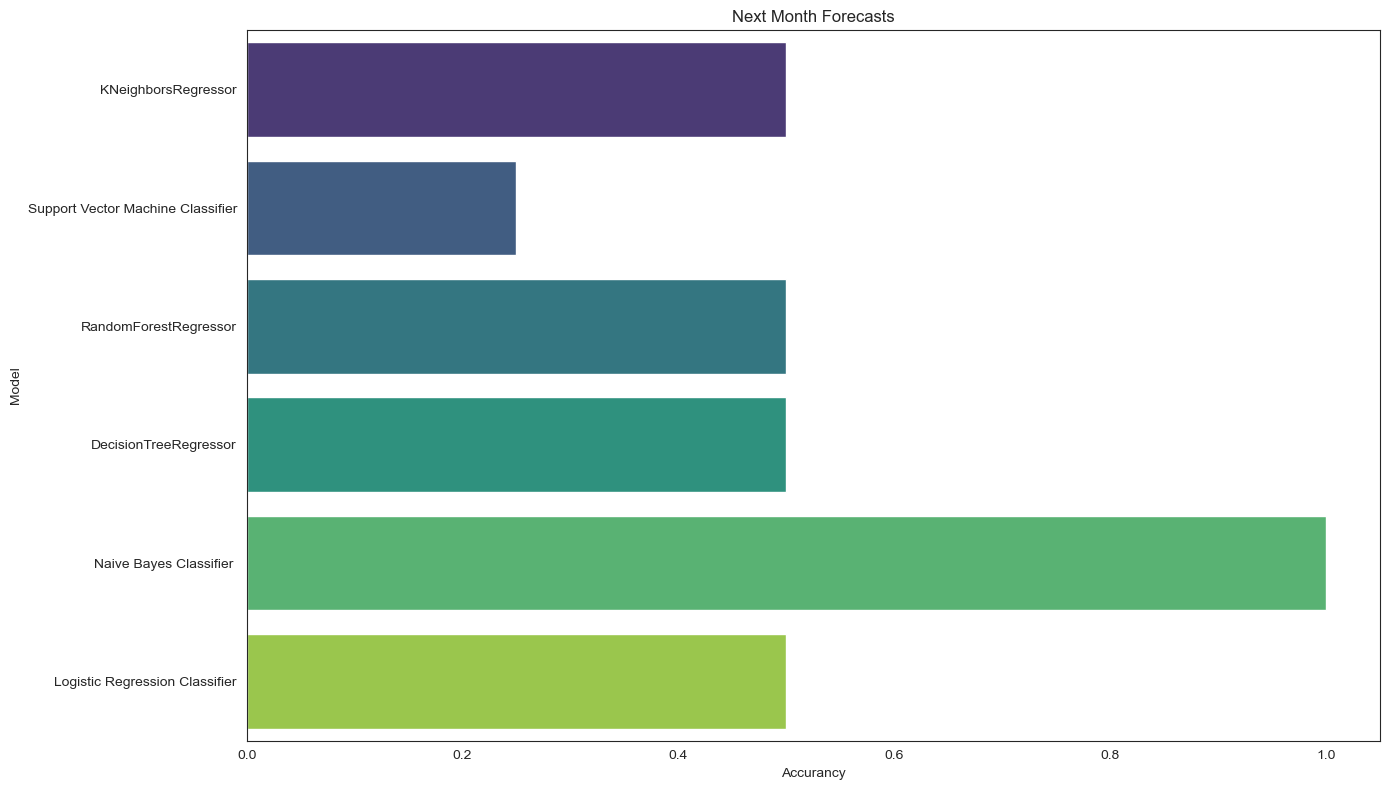
10 San Francisco Mar 111525 2021 Mar 2021 0

11 Dublin Mar 121960 2021 Mar 2021 0

Monthly Bike Usage table is as follows.



Comparison of Machine Learning Methods According to Accurancy



The accuracy rates are mainly 0.25, 0.5, and 1.0, which raises some interesting points.

Naive Bayes Classifier has achieved 100% accuracy, which is unusual and may indicate potential issues like data leakage or overfitting. Dataset was scaled by StandardScaler before .

Support Vector Machine Classifier (SVM) has the lowest accuracy of 25%, suggesting the model might require better data preprocessing, scaling, or hyperparameter tuning.

# **4. RESULTS**

Various analyses were conducted on the statistics of shared bike usage in Dublin and San Francisco, yielding several insights. Bicycle use in San Francisco has increased steadily over the years, demonstrating consistent growth. However, Dublin shows a different trend, which can likely be attributed to the government's policies on shared bikes and its support for cycling.

Another point of comparison is the usage of e-bikes versus classic bikes. In San Francisco, e-bike usage accounts for 75.6%, whereas in Dublin, this figure is just 2.3%. Conversely, classic bikes are far more popular in Dublin, with 97.7% of users opting for them, compared to only 24.4% in San Francisco. However, when looking at total bike usage, Dublin surpasses San Francisco only in December and March. For the remaining months, San Francisco shows a significant lead. San Francisco's mild climate and generally bike-friendly weather encourage cycling, while Dublin's wetter and windier conditions may make e-bike use more challenging.

In terms of algorithm performance, the Support Vector Machine model showed an accuracy of 25%, which suggests it should not be used for this analysis. On the other hand, the K-Nearest Neighbors Classifier, Random Forest Classifier, Decision Tree Classifier, and Logistic Regression Classifier all achieved 50% accuracy. Since these models have a high risk of errors, they are not suitable for this problem. Interestingly, the Naive Bayes Classifier achieved 100% accuracy, indicating that the model perfectly classified the instances. However, despite this perfect accuracy, Naive Bayes is generally more suited for text classification problems, and its performance in this analysis may be misleading.

REFERENCES

A.S.Ahmad, M. Y. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. https://www.researchgate.net/publication/260315410\_A\_review\_on\_applications\_of\_ANN\_and\_SVM\_for\_building\_electrical\_energy\_consumption\_forecasting.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. https://link.springer.com/book/10.1007/978-0-387-45528-0.

*Knowledge Discovery in Big Data from Astronomy and Earth Observation.* (2020). Khadija El Bouchefry PhD, Rafael S. de Souza PhD.

Luckin, R. (2022). Machine Learning and Human Intelligence: The Future of Education for the 21st Century. *Machine Learning and Human Intelligence*, https://eric.ed.gov/?id=ED584840.

Mahesh, B. (2019). *Machine Learning Algorithms.* doi:10.21275/ART20203995

*Pisner and Schnyer.* (2020). Support vector machine. In Machine learning .Academic Press.

*Pisner and Schnyer.* (2020). Support vector machine. In Machine learning .Academic Press.

*Schonlau and Zou.* (2020). The random forest algorithm for statistical learning. The Stata Journal.

*Van Engelen and Hoos.* (2020). A survey on semi-supervised learning. Machine Learning.

*Van Engelen and Hoos.* (2020). A survey on semi-supervised learning. Machine Learning.

on semi-supervised learning. Machine Learning.