Rochester Institute of Technology
RIT Digital Institutional Repository

Theses

5-13-2022

Hourly Demand Prediction of Shared Mobility Ridership

Eman Asa'ad
ea7082@rit.edu

Follow this and additional works at: https://repository.rit.edu/theses

Recommended Citation
Asa'ad, Eman, "Hourly Demand Prediction of Shared Mobility Ridership" (2022). Thesis. Rochester Institute of Technology. Accessed from

This Master's Project is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

HOURLY DEMAND PREDICTION OF SHARED MOBILITY RIDERSHIP

by

Eman Asa'ad

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

13/5/2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Eman Asa'ad**

Graduate Capstone Title**: Hourly Demand Prediction of Shared Mobility Ridership**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ioannis Karamitsos

Date:

Member of committee

Acknowledgments

I would like to thank my esteemed supervisor – Dr. Ioannis Karamitsos for his invaluable supervision, support, and tutelage during the Graduate Capstone. My gratitude extends to the Faculty of the Data Analytics Program in Rochester Institute of Technology -Dubai led Dr. Sanjay Modak – Program Chair. Additionally, I would like to express gratitude to Dr. Mick McQuaid and Dr. Erik Golen treasured support which was influential in shaping my data analytics methods and knowledge. I would like to thank my friends, lab mates, colleagues, for a cherished time spent together in the program. My appreciation also goes out to my family and friends for their encouragement and support all through my studies.

Abstract

This research focuses on predicting the hourly number of bikes needed using Citi bike data. Micro mobility is the new trend that serves the transportation sector in any city. With the development of technology and introduction of new modes, comes new challenges. Bike sharing is the most developed and standard micro mobility device with extensive data sources. In this research we introduce the rebalancing bike sharing problem, which is very recent and interesting problem. Bikes are being ridden from a station and returned to another, not necessarily the same one of departure, this procedure can cause some stations to be empty while others to be full, as a result, there is a need for a method by which distribution of bikes among stations are done. Using year-round historical trip data obtained from one of the famous bike operators in New York that is Citi bike. The study aims to find the factors affecting bike ridership and then by utilizing some predictive algorithms such as, regression models, k-means, decision trees and random forest a model will be created to estimate the number of bikes needed in an hourly basis regardless of any specific stations initially. Where accuracy will be eventually calculated.

The testing will be initially evaluating the data of Citi bike in New York, however, the same can be utilized to evaluate data from other cities worldwide and operators, as well as other micro mobility modes such as e-scooters, mopeds, and others. Initially the Prediction problem will be evaluated against the current data available in the open-source Citi-Bike data, however, weather factors, bike infrastructure, and some other open-source data can be integrated for better results.

Keywords:

Shared mobility, Micro mobility, Bike sharing, Hourly prediction, Citi-Bike Data

CONTENTS

List of Figures

List of Tables

CHAPTER 1

1.1 Introduction

Transportation methods have dramatically changed over time with cars ruling the roads over the past century, which created problems from noise to air pollution to congestion. Over the past years there was a clear need for a green and more sustainable means of transportation with a focus on public transport as an alternative.

However, as public transport is limited to specific stations and stops, daily commuters shifting from using cars find it challenging to easily reach to such stations from their place of residence or offices. This term in transportation is called the ***first- and- last mile***. The key challenge is how to encourage commuters to leave their initial origin walk to reach to such a stop before taking a bus or a metro to the next stop or station, then walk back to their destination. This required to introduce modes to make the trip faster, easier and would not cost much. This change came in a form of new shared micro-mobility options, such as bike-sharing, e-bikes, e-scooters, and other devices which had a significant presence on cities all around the world.

Shared micro-mobility came with clear benefits and effects in terms of reducing both pollution and congestion in addition to notable health benefits. The first bike sharing initiative started on Europe on the late 20th century around 1960’s and 70’s, but this wasn’t made hugely beneficial till mid-2000’s with the rise of the fourth industrial revolution and the significant adaption of digital platforms, social media, smart phones applications, in addition to installing sensors so live data can be obtained and used to plan and manage the bike-sharing services.

Bike sharing described as “a pool of publicly available bicycles placed around the city and ready to be used for a low payment.” [1] Bike sharing main aim is to complement the current transportation system. This came with added advantages, such as convenience, healthier way to travel, it is environmentally friendly, inexpensive, costs saving and helps reduce traffic jams, in addition to increasing the popularity of areas for both tourists and investors. When compared to cars traveling the same distance, bike sharing systems have the potential to prevent 37,000 kg of carbon dioxide every day [2]. However, it came also with its own downsides, there were cases where new bike sharing services came out to be unsuccessful and operations had to shut down shortly after its initiation which caused much loss to operators due to issues such as imbalance in distribution. Since operators seek profit, it is vital for them that the devices are in continuous move with the least possible costs of maintenance and operation. Hence, it is important that the operator does not have to have a team to return the devices from destinations of low usage to origins of high demand. Furthermore, the imbalance of bikes distribution has another major issue, its particularly vital that users can find bikes when needed. Thus, bikes relocation in different stations required in a way that bikes from overcrowded stations are transferred to those with less bikes, but there is a cost attached to this process and there is a need to predict the number of bikes needed in different stations, so the rebalancing procedure is done properly. This generated the need to research in rebalancing bikes or any other mode around different stations.

1.2 Statement of the problem

This research aims on understanding the success factors for implementing the micro mobility services and aims to answer the following Question:

- ***How to predict the number of bikes/ devices needed on an hourly basis in various locations so that the bike sharing system can relocate them as needed with the maximum profitable way?***

The aim of this study is to forecast bike sharing demand to address the rebalancing problem of the bike sharing system by assessing different machine learning models as random forest, Gradient Boosting Regression Tree.

1.3 Research Methodology

Utilizing a data mining framework/methodology helps to organize the work of a data analytics project and further achieve its goals and objectives. It provides an overview regarding the sequential steps of managing the data mining task at hand. The Cross-industry Standard Process for Data Mining **(CRISP-DM)** methodology was implemented for this project.

- **(CRISP-DM)** methodology dates to early 2000s. it has existed for 20 years and still used in different projects and research up to date.

The need for **(CRISP-DM)** methodology came from the need of Data mining methods to a standard approach which will help translate business problems into data mining tasks, propose suitable data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and lastly documenting the experience. [3]

CRISP-DM methodology aims to achieve all the above in addition to its contribution to increase the use of machine learning and data mining over a variety of business applications. It is a very highly flexible and cyclical model. It consists of six steps as described below:

1. **Business Understanding**

   This phase is the most important part of the project, in this study the aim was to understand the challenges facing the bike sharing demand systems and identifying a clear intention of the whole project with clear goals, objectives and requirements.

2. **Data Understanding**

   This phase depends heavily on business understanding. Data was collected from the Cit-bike open-source data. Clear understanding of the objective of the study which is predicating the hourly need of bikes was determined to predict what data we need to focus at and from what sources and by which methods. In our case the data from the year of 2021 was collected and studied.

   “The initial collection of the data is followed by getting familiar with the data to discovering insights and detecting any interesting hypotheses.” [3]

3. **Data preparation**

   This phase come after the collection and understanding of the data, in which several steps were conducted from features selections and transformations, new features formulation, to data cleaning. all this aimed to transform the data into a final form that can be easily used in further steps and fed into the modeling tools

4. **Modeling**

   This is the phase where suitable models were created to give useful insights and to create useful knowledge out of the data. Patterns were revealed in this stage and the feature of interest were determined. In this stage different modeling techniques were chosen and used.

5. Evaluation

At this stage, all models were evaluated to determine the effectives of the different models, and which one achieve the main objective of the study. Three models in our case were evaluated, random forest, gradient boosting and multivariant regression model.

6. Deployment

This is the stage where the model will be used on a new data set

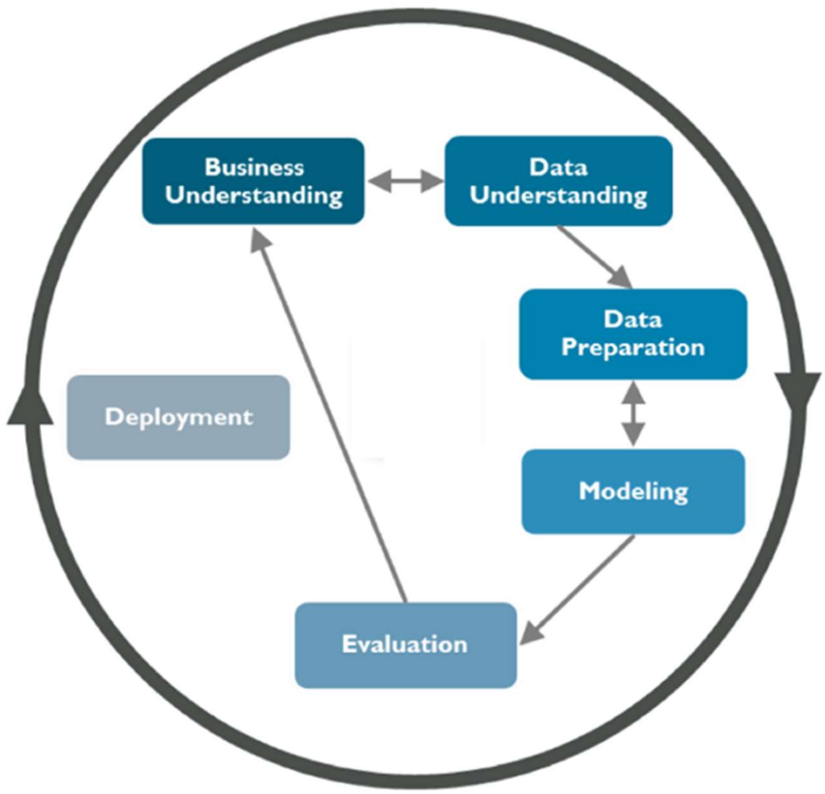

Figure 1 Deployment Method- CRIP Methodology

1.4 Project Goals, Aims and Objectives

- The main Goal of this project is to: "***predict the hourly number of bikes needed within a system and/or area***." Using historical data sets along with independent variables, such as day, time and others, the model will predict the number of bikes that will be rented in that certain hour where they are located and what are the numbers needed.

- The aim of the project is to "***reduce the cost for moving bikes from one station to another in addition to increase customers satisfaction and provide them with a pleasant experience through availability of shared mobility devises as needed"***
- The key objective is ***"to reach Balanced status."*** The term "balanced status" refers to the scenario in which bike supply and demand are equal. Demand fluctuates spatially and temporally in both station-free and station-based bike sharing systems, resulting in an imbalance problem. The imbalance problem may significantly affect system performance unless timely rebalancing measures are made.

1.5 Limitations of the Study

- There are limitations that affected the study initially:
    - **Main Data Sources:** There was a shortage of the availability of open-source data, which made studying the bike sharing system locally difficult. In addition, available data sources are also not complete and would require further data to be collected and shared. However, the model is going to be generated in a way that once local data is available, it can be transformed into the format accepted by the models and hourly prediction can be conducted.
    - **Data features:** Another limitation of the study was the limited number of features in the open data set, further features would enhance the study. Data must be collected from various sources and from various agencies, while the data can be collected from a single source which will provide a more set of accurate results.
    - **Number of Stations/ docks per station:** The models were conducted on all stations. However, the unavailability of key information on the capacity per station limits the ability to assess the percentage utilization per station. However, this can be further utilized on specific or top stations whenever the need arises.

- Neural network continues to prove creating particularly satisfactory results in similar problems, the utilization of neural networks would enhance the study. In addition, the models can be further enhanced with the city data supporting the data of the service provider such weather information, population and census data, infrastructure information, and any other available data that would be vital to the study.

CHAPTER 2

2.1 Literature Review

Understanding the problem and solving it requires a deep understanding of the issues that arise. Hence, it is vital to have a literature review to understand the challenges that occurred and the different solutions that were evaluated along with the motivations and history behind the increase interest in shared mobility.

“Shared mobility - the shared use of a vehicle, bicycle, or other mode - is an innovative transportation strategy that enables users to gain short-term access to transportation modes on an “as-needed” basis. The term shared mobility includes various forms of carsharing, bike sharing, ridesharing (carpooling and vanpooling), and on-demand ride services. [4] Looking at micro and shared mobility in general it can be noted that among the different mode shares, bike sharing is the most mature as it is the oldest between them, even most modes have emerged from bikes. studies were conducted to better understand its benefits and enhance its operations, and this would be our concentration int this paper.

Add to that, and to merge both the online and offline worlds and because of the superiority of resource scarcity, a rapid increase in shared economy has been witnessed in the past two decades. [5]

Shared economy is described as an approach which relies on the interaction between the social and economic factors, and which make it possible for services and good to be exchanged between organizations or individuals to better utilize resources and improve its efficiency. Where transportation and shared mobility is one of the major sections where this concept has been applied. [6]

“Bike sharing growth has undergone three evolutionary stages including: first-generation white bikes (or free bike systems), second-generation coin-deposit systems, and third-generation IT-based systems" [7]. However, our main concern in this paper shall be the docked bike sharing system with proper tracking techniques.

Bike sharing has significant environmental advantages, in a study which was conducted in Shanghai (China's economic center) in 2018, bike sharing in Shanghai saved 8358 tons of petrol and decreased CO_2 and NOX emissions by 25,240 and 64 tons. The authors concluded in their research that energy intake and emissions can be decreased dramatically using bike sharing systems. [8]

Another study was conducted in Europe to study the health impact of bike sharing systems, where a quantitative model was created, and data was collected from transportation, health, and environmental surveys. In all scenarios and cities, the health benefits of physical activity outweighed the health risk of traffic fatalities and air pollution in which they concluded in their research that bike use can significantly increase health benefits. [9]

Other studies have shown that the most common factors which can have an influence on bike sharing systems were established and studied to help in better understanding the factors behind the bike sharing demand, and can be used as a guideline for planners, policy makers and researchers. In their research, Ezgi Eren and Volkan Emre Uz, found out that Weather as expected came as one of the major factors, in particular rain which negatively impacted bike trips for both weekdays and weekends. Another factor which was studied was built environment and land use factors which could have an impact on safety of users. As for the effect of age, it was not possible to make definitive finding in their study, however, young-adult individuals are more likely to use bike sharing than other age groups. Station buffer distance came also to be as one of the determinants in the decision process. [10]

As more bike sharing systems are being implemented in more cities around the world, more challenges are starting to appear, one of which is the rebalancing problem. That is, "the efforts of restoring the number of bikes in each station to its target value by routing vehicles through pick-up and drop-off operations." [11] this led to a continuous demand to predict the number of available bikes in different stations to increase the efficiency of the bike's stations.

Some studies were done to gain more knowledge to enhance the rebalance operations, and some others to find the correlations between demand and the various other factors. Which would help micro mobility start-ups, policy makers and others to better tackle the problem. Studies varied from univariant to multi variate. While their approaches relied on time series forecasting methods in addition to Graph based methods.

In a paper titled "Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data," the authors used both K-Nearest Neighbor and Liner Regression to find the relation between users' behavior and weather data. where temperature proved to be a very major factor. [12]

In another research, Random Forest and Least-Square Boosting algorithms were used to build univariate prediction model, and that is to predict the number of available bikes at each station, Partial Least-Squares Regression was also used as multivariate regression algorithm, in which the conclusion came to be that station neighbors, prediction horizon time, and weather features are among the very significant factors in modelling the number of available bikes. [13]

Another study from China, tackled the problem of rebalancing bikes among stations, stations can be empty or saturated in various times, which rise the need to some means of bikes distribution, and that was by driving trucks to redistribute bikes which lead to unnecessary human resources in addition to being ineffective operators and inconvenient to users. Wanga & Kim, used in their research different machine learning techniques such as Random Forest, long short-term memory neural network and Gated Recurrent Units techniques. Random Forest showed better performance than others, however, both LTSM and GRU are similar on predicting behaviors, but GRU has more accurate results and faster training time than LTSM. [14]

Another case of a technique being proposed is a multi-graph convolutional neural network to predict the number of bikes at station level. In their paper, Chai, Wang, & Yang, described multiple interstation graphs for a bike sharing system as follows:

- Stations are presented as nodes, where the edges between the nodes or stations represent sort of relation (distance, ridership). In addition, graphs were created to indicate the different relations available.
- The graphs are being fused and convolutional layers on the fused graph were applied to predict station-level future bike flow. Their prediction model could surpass a number of advanced station-level prediction models. [15]

Another recent study was conducted in Seoul, using both Seoul Bike and Capital Bikeshare data, along with weather data. five models were created using a repeated cross validation approach followed by different testing methods. The models created were CUBIST, Regularized Random Forest, Classification and Regression Trees, K Nearest Neighbor and Conditional Inference Tree. “The most significant variables from all models were obtained, the most variables came out to be Hour of the day and Temperature” as per their findings. [16]

In another study, a dynamic rebalancing strategy was proposed where historical data was used for modelling purposes. Birth-Death processes were used to determine how bikes were being distributed. In addition to using graph theory so paths and stations can be chosen for the rebalancing problem. The framework was validated on New York City’s bike sharing system. Their findings came to proof that the dynamic methods were better able to adapt to the fluctuating nature of the network and it outperformed the rebalancing techniques used on static methods. [17]

Another method was proposed using a regressor and predictor to predict the pickup and drop off demand. Based on vast number of historical records a meteorology Similarity Weighted K-Nearest-Neighbor regressor was developed aiming to predict the pickup demand of stations in an hourly basis. In addition, they calculated station drop off demand by developing Inter Station Bike Transition Predictor. These were used to predict the station inventory targets. Add to that, they proposed a hierarchical optimization model for finding the optimum solution for the rebalancing problem. [18]

2.2 Key Findings

Studies were conducted defining shared mobility, better understanding its benefits, and enhancing its operations

- Studies varied from univariant to multi variate
- Most famous methods used are Random Forest, Least-Square Boosting algorithms, long short-term memory neural network and Gated Recurrent Units to predict the number of available bikes at each station
- Recent studies used models created like CUBIST, Regularized Random Forest, Classification and Regression Trees, KNN and Conditional Inference Tree
- Birth-Death processes also was used to determine how bikes are being distributed. In addition to graph theory

2.3 Assessing the Gaps

In our research, different methods have been assessed against finding the best model for hourly based prediction, which is not popular in the previously available research. Much research was focusing the previously mentioned methods which are cluster based, area based, and station based. Where insufficient research has been done on hourly based station prediction despite of its importance in resolving the rebalancing problem. In addition to that, our models are general and can be applicable to any other modes of smart mobility devises such as e-scooter, e-bikes and much more.

The key GAP is in the data that is utilized, despite the fact that the Data utilized in this study is acquired from an open source of the service provider, this information is not complete in assessing the capacity of each station, compared to other studies conducted in China in which the assessment of stations capacity was conducted due to the availability of data.

CHAPTER 3

3.1 PROJECT DESCRIPTION

Micro-mobility – or “Micro-vehicles” are a group of small, light vehicles that are driven by individuals at low speed of a maximum 25kph, and particularly located in city congested streets for a short distance that would average around less than 5 miles. Those devices are considered as a solution to different problems including the first- and- last mile or even as a very affordable and convenient option for short trips in a more sustainable way. Those devices vary with the most common current types that are shared bikes and scooters.

Bike-sharing systems are becoming increasingly popular in cities around the world because they are cheap, efficient, healthy, and green. As the population continues to grow rapidly cities are expanding and this is putting pressure on developing major roads infrastructure, while balancing that with the mode shift to public transport. Therefore, residents’ movement is getting more challenging and complicated. Hence, public transport means remains the most efficient in moving considerable number of people for lengthy distances, getting people to and from traditional means (from cars to buses to trains) the first-and- last mile challenge remains difficult. If it is difficult for people to use public transport, they will end using their own vehicles which would increase air and noise pollution, that would cause more traffic jams and accidents. In addition, lack of affordable short trip solutions might discourage people to travel in the first place, neglecting job opportunities, medical care, or even healthy food.

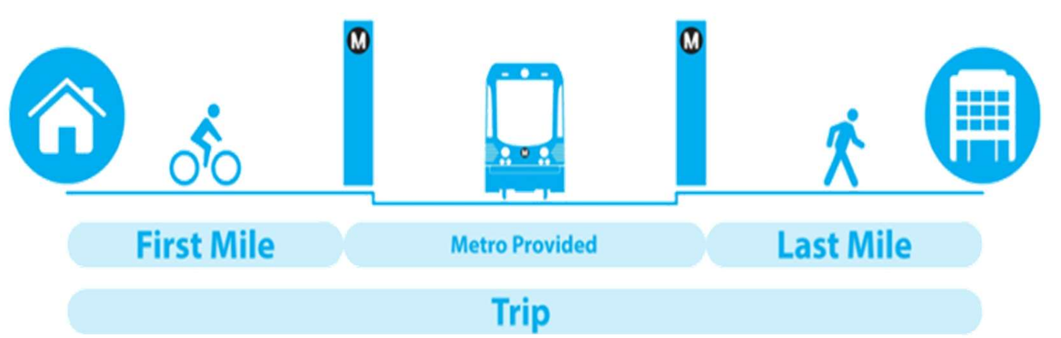

Figure 2 First Last Mile

Bike sharing is emerging as an alternative to nowadays challenges, filling the gap in public transport routes plus its environmental and sustainable benefits. Bicycles can be picked up and dropped off at any time and from any location where a station is in the bike sharing system thus, it comes with its own challenges and problems. As commuters cannot take their personal devices with them on busses and/or trains to support their overall trip, they must rely mainly on service providers leasing micro mobility devices and charging them based on either distance or time. The suppliers of such service are either big established international organization or even start-ups, which are licensed by cities to encourage entrepreneurships. No matter the size of the organization it is important that they can make profit to sustain their operation. Failing to make any profit would result in firms running out of business.

Luckily, with the technological enhancements, which are data driven, and support of IOT the issue can be assessed, and a model can be developed and trained to evaluate the best business model that would lead to increase in profit by maximizing the usage of devices, and reduction of operation cost. Each bike or even micro mobility mode is being monitored within a georeferenced zone, key information is collected by the bike and stored.

3.2 History of Bike Sharing

Five main generations of bike sharing have been developed over the past 45 years.

- **The first generation** started in Amsterdam in the year of 1965 with what was known as White bikes. White bikes were ordinary *free access* bikes painted in white and used as a public means of transportation. The big challenge that faced the first generation was the misusage of bikes, bikes were thrown in the canals or taken for private use, the program *stayed for few days* only.

- **The second generation** was born in Denmark in the year of 1991 and 1993 respectively, the services were small at that time, in 1995 the first large-scale second-generation bike sharing system was developed with a lot of advancement and enhancement over the first generation. Where *bikes had to be returned and picked up* from certain locations with a coin deposit. Still users were anonymous and there was no way to track customers at that point. It was free usage.
- **The third-generation** bike sharing system was released in the year of 1996 in England, where technological improvements were added to the previous generations, at that point a magnetic strip card was used to rent a bike, so users are being identified at that stage, it had locking racks, telecommunication systems and smart phone access. However, it was not until 2005 in Lyon that the biggest third generation system was released with noticeable impact with around 15,000 members and bikes. In 2007 Paris launched its own bike-sharing network named Velib which was expanded to 23,600 bikes in the city which had significant impact and created noticeable interest.
- **The fourth-generation** bike systems, it was equipped with intelligent transportation technology and real time information provision that allowed the bikes to be accessed through mobile app in an integrated traffic management system. [19]
- **The fifth-generation** bike systems introduced dock-less bikes and big data management possibilities. [20]. Those are the systems we see in some of the cities around the world these days.

3.3 Impacts Affecting Bike Sharing

Despite its popularity, bike sharing tends to be affected by a lot of factors, the main factors as they were summarized in the paper “A review on bike-sharing: The factors affecting bike-sharing demand” [21] and can be summarized as follows:

- **Weather condition**, which is one of the main factors, it varies slightly based on attributes, but rain, snow and extreme temperatures have negative correlation with bike sharing demand.
- **Wind speed and relative humidity** in which many research papers agree on the negative correlation between wind speed and relative humidity and bike sharing demand.
- **Built environment** is another factor, the better infrastructure for cycling usage the more the demand, such as number of cycling lanes, and the existence of isolated lanes, streetlight availability, safe parking areas, locations of stations are effective factors. [22]
- **Land use**, bike sharing often decreases when the slope increases, for instance up-hill slopes decreased the demand significantly, in addition the existence of shopping centers, tourisms areas, hotels and restaurants can all affect bike sharing usage. lastly the
- **Public transportation impact** factors, the existence of other mood of transportations next to the bike sharing systems, tend to encourage user to better utilize the systems to complete their trips.

3.4 Project Approach

As stated, bikes ridership has increased tremendously as it is a healthy sustainable transportation mode, for instance, in the city of New York Citi bike ridership has increased from hundreds of thousands rides per month back in the year of 2013 to million trips in 2019. With the increase in demand this number is still on the rise. As the number of ridership increases, it becomes more difficult to find bikes where and when people need them. The reason is that most people tend to ride in a specific pattern. They tend to start and end in public transit locations such a metro or bus station. They also tend to travel at certain peak hours early morning, during the lunch break and at the end of the working day. Research has also shown that people like to ride downhill rather than uphill, during daytime rather than nighttime.

As a result, having the bikes back for other people to find and use them is one of the most challenging problems. As it takes so much time for balancing bikes after imbalance happens.

Currently Rebalancing methods used by Citi-bike are:

- **Valets:** dedicated employees that move bikes to their popular stations
- **Vans:** transfer bikes between full and empty stations
- **Bike angels:** program which rewards riders with points upon riding bikes from full stations to empty ones.

However, Empty docks remain a problem in certain parts of the city especially that Citi bike ridership has increased tremendously, from 600,000 monthly trips in 2013 to 2.1 million trips in October 2019.

For the problem to be solved, one part of it is to predict the number of bikes needed in certain stations during different timing to better prepare for future needs. That is by creating a model that would predict the hourly demand for the top stations in New York city using the open-source Citi-bike data.

By implementing a data driven approach and to perform our analysis, different data sets from open-source resources will be used to better understand the factors behind the hourly need. The aim of this study is to forecast bike sharing demand to address the rebalancing problem of the bike sharing system by assessing different machine learning models as random forest, Gradient Boosting Regression Tree

Creating the model began by studying prediction models that had similar targets, followed by obtaining the data from a service provider with extensive available data. Then the data from all months of the year 2021 were merged, cleaned, and formatted properly, where other features were added, afterwards data was analyzed to better understand the independent variables that will increase or decrease the demand, and different predictive models were designed and then implemented. Their performances were accessed and then they were deployed on new data.

CHAPTER 4

4.1 Data Understanding – Data Source

The open-source trip data used in this study came from the Citi Bike bike-sharing system in New York and is made available at [Citi Bike System Data | Citi Bike NYC](). It is a privately operated bike-sharing system that began in May 2013, and it has gained significant importance since then. Lyft currently runs Citi Bike. It has more than 750 stations where docked bikes are being placed. Recently e-bikes are being added to the operational fleet of New York City.

- Trip data within New York City has been used, the trip data set includes information about, *ride id*, *rideable type*, *start time* and *stop time, start station name*, *end station name*, *date*, *station id*, *station latitude* and *longitude*, *type of rider* (member or casual) the data has been taken for the year of **2021**.
- As for January month in particular, the trip data set includes information about trip duration, *start time, stop time, start station id, start station name, start station latitude, start station longitude, end station id, end station name, end station latitude, end station longitude, bike id, user type, birth year,* and *gender*
- After performing premilitary steps on the January file, the data was merged into one data frame that contains **644,443** records and **sixteen** attributes.

Below shows the first five rows and summary statistics of the attributes in the dataset.

end_station_name	end_lat	end_lng	ride_id	member_casual	birth year	gender	rideable_type
Brunswick St	40.724176	-74.050656	42494	Subscriber	1988.0	1.0	NaN
Van Vorst Park	40.718489	-74.047727	45343	Customer	1996.0	2.0	NaN
Van Vorst Park	40.718489	-74.047727	31794	Customer	1995.0	1.0	NaN
Newport Pkwy	40.728745	-74.032108	42316	Customer	1969.0	0.0	NaN

tripduration	started_at	ended_at	start_station_id	start_station_name	start_lat	start_lng	end_station_id
266.0	2021-01-01 00:03:35.5100	2021-01-01 00:08:01.7770	3273	Manila & 1st	40.721651	-74.042884	3209
1543.0	2021-01-01 00:23:32.9250	2021-01-01 00:49:16.0830	3681	Grand St	40.715178	-74.037683	3213
1461.0	2021-01-01 00:23:50.7940	2021-01-01 00:48:12.5660	3681	Grand St	40.715178	-74.037683	3213
793.0	2021-01-01 00:31:09.0770	2021-01-01 00:44:22.9430	3185	City Hall	40.717733	-74.043845	3199

Table 1 Summary Statistics and Attributes of the Dataset

4.2 Data Preparation / Pre-processing

Datasets in the real world are often messy; however, the bike sharing dataset is almost clean and simple. The preprocessing steps performed on the dataset are explained as follows.

- Data preparation includes different steps such as exploring and cleaning the data, to combining the data with other resources for further analysis.
- The first step in our process was data merging or combining, the data from Citi-bike is being stored in a monthly basis, we aimed to study it on yearly basis and that is in the year of 2021, thus the twelve files were combined into one data frame for analysis and models creations. For this to be made possible the following steps were performed:
    - The names of the attributes in the January file were changed to match the rest of the months.
    - few columns were dropped from the January file as they did not exist in the 11 other months which are *date of birth* and *age*
    - values of *subscriber* and *customer* in January were converted to *member* and *casual* so that it is consistent with the rest of the months.
- The data merging step was followed by data cleaning and formatting, which include the following main steps:
    - All date-time data were converted to their corresponding type. for example, *start time* has been converted to date-time, some cases had padded values at the end of time values which could lead to inefficient conversion to data time object thus the extra values at the end were trimmed. a good example is in case where start time is *'2021-01-01 00:03:35.5100'*. the extra values at the end were trimmed before conversion.
    - Missing values were computed and managed, in our case since missing values were little, they were kept as is and they were not dropped from the dataset.
    - New features were calculated from the current attributes for better data expletory, for instance start time and end time were used to calculate trip

duration. Day of week feature has been added by extracting the day out of the started_at attribute, month has been computed by extracted the name of the month from the started at feature, started_at_timezone was computed to better understand the part of the day in which the most demand used to occur, the started_at_timezone takes the values of morning, midday, evening, and late night. finally, trip duration was calculated from start time and end time attributes.

4.3 Exploratory analysis of Bike-sharing Trips

1. Locations and Users

As can be seen from the figure below the majority of both start and end stations are in Manhattan. Manhattan has the maximum number of populations compared to other boroughs in New York City, in addition it is geographically the smallest. Manhattan is regarded as one of the most important commercial, financial, and cultural centers in the world. It is well-known for its attractions. It has one of the world's most famous streets which is the Broadway. add to that it has various skyscrapers such as the Empire State Building; Greenwich Village, Harlem, and Central Park; the United Nations headquarters; and various cultural and educational institutions. Which explains the high density of trips in that borough specifically.

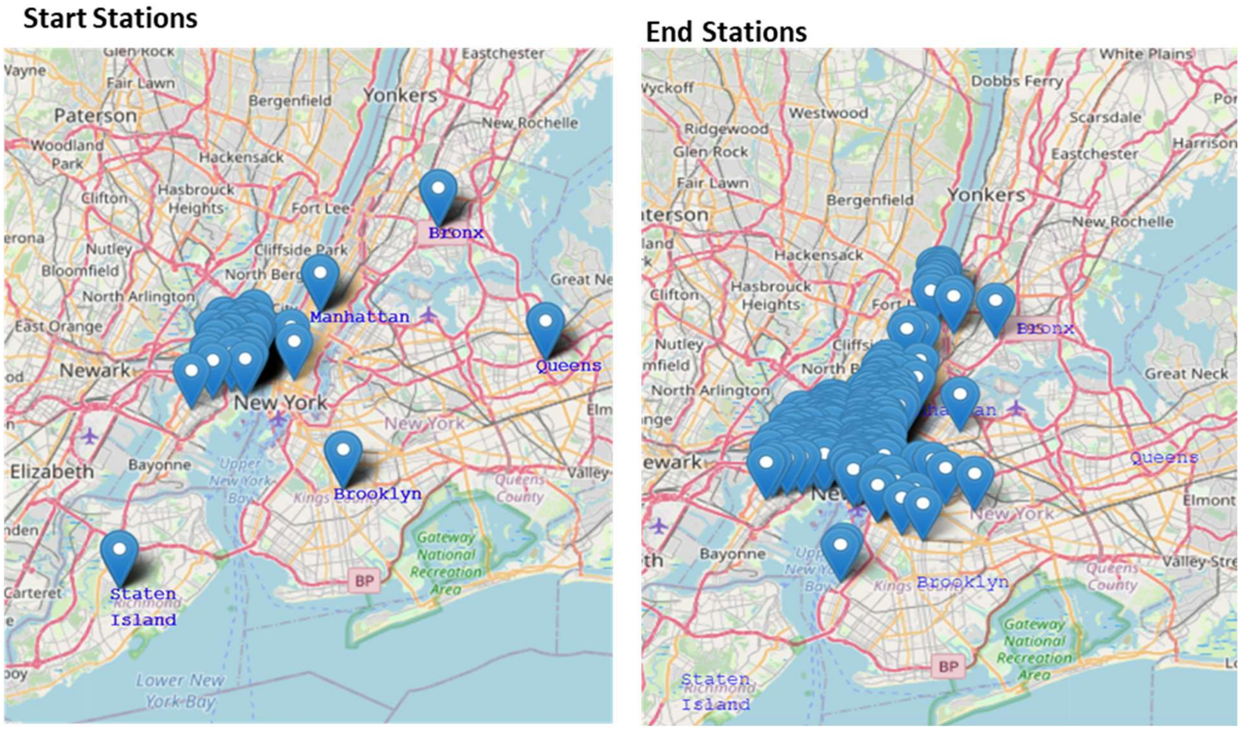

Figure 3 Start and End Stations

The figure below shows that approximately 60% of the users were members. While 40% were casual users

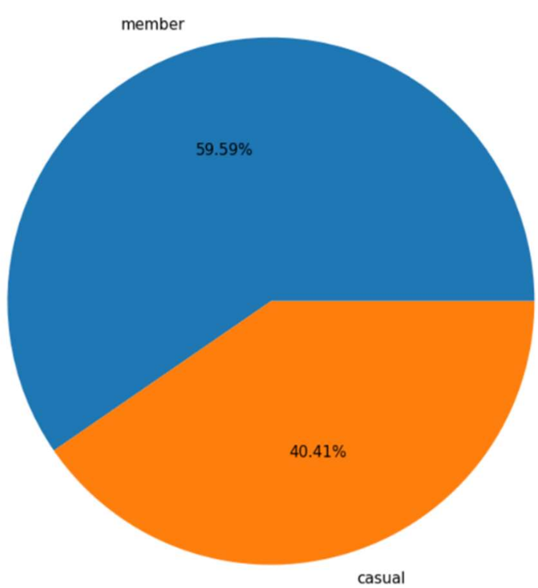

Figure 4 Customer Type Distribution

As can be seen from the figure below, casual users' trips are higher only during the weekend, while through the week, member users have taken most trips. implying that member users are usually either employees or students who need to use the service in a regular basis.

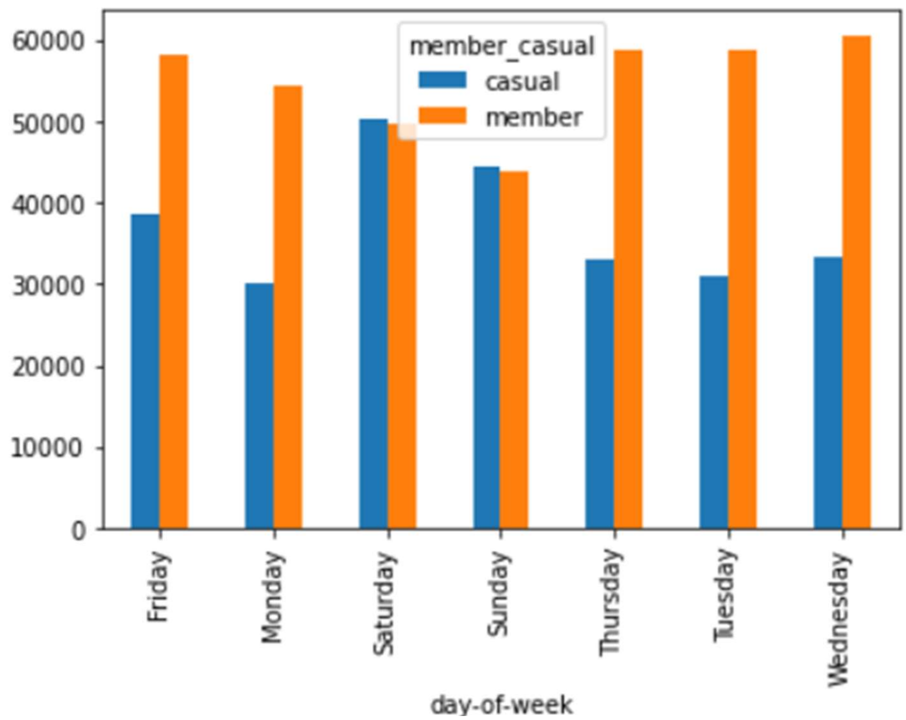

Figure 5 memeber_casual through weekdays

Also, the below figure Shows that casual users' trips where higher than member trips specifically in late night hours, which also explains the nature of trip done by causal member which are for entertainment.

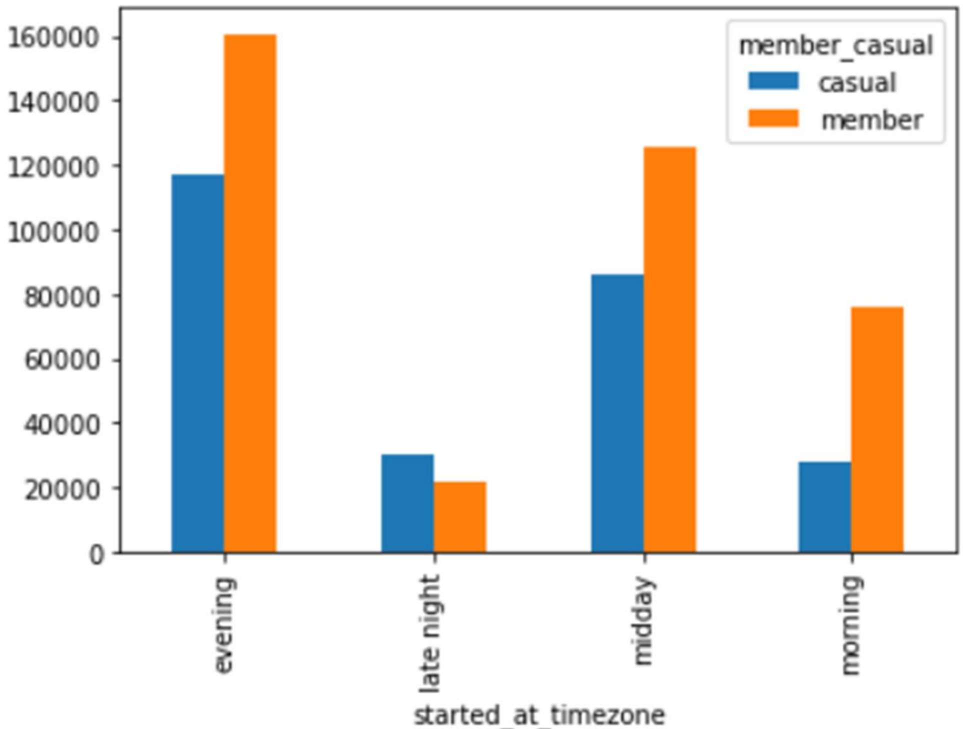

Figure 6 member/casual started at time distribution

From the below figure it can be clearly seen that the start hour for casual trips used to be higher than member trips which agrees to the above observations about the nature of the trips done by member users.

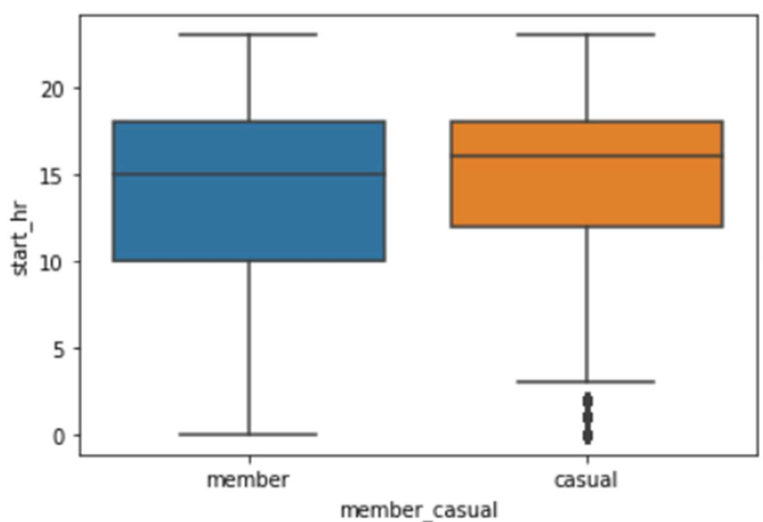

Figure 7 Start hour for member/casual trips

2. Weekly, monthly, and yearly distribution

As for both member and causal riders, August and September were the months with the greatest number of rides.

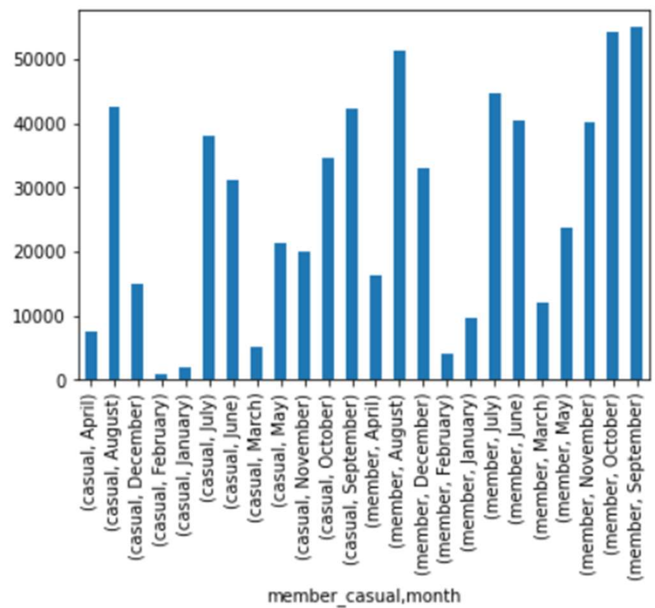

Figure 8 Months distribution for member/casual trips

Regardless of the type of users, the figure below shows which months had the largest number of trips, clearly summer season (July, August, September) had the greatest number of tips followed by Autumn and Spring, which is attributed to weather conditions, as cycling in rainy and wintry conditions is more tedious and difficult.

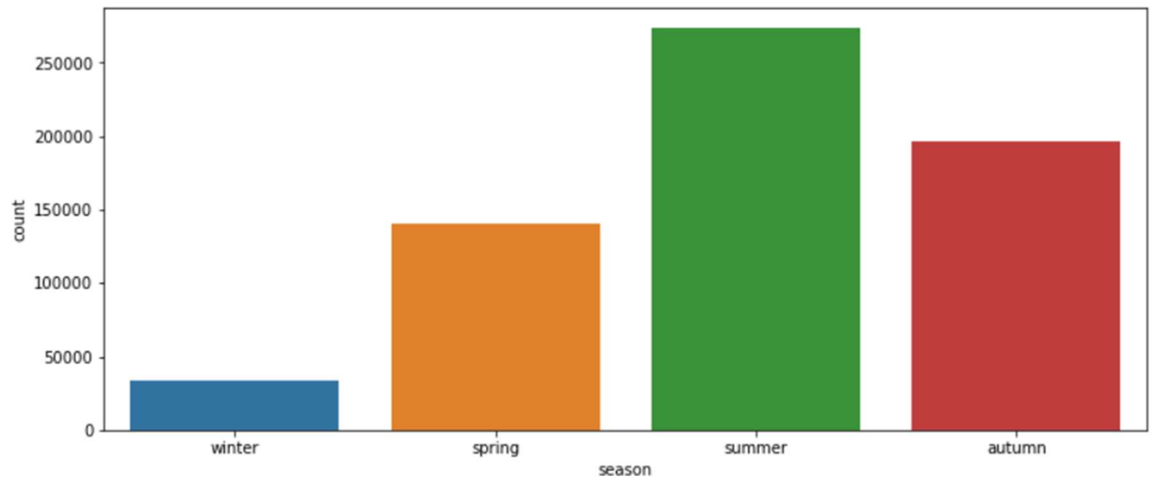

Figure 9 ridership through the seasons

As for the weekdays, the figure below shows that most rides were taken place on Friday and Saturday, due to the increase of the casual trips

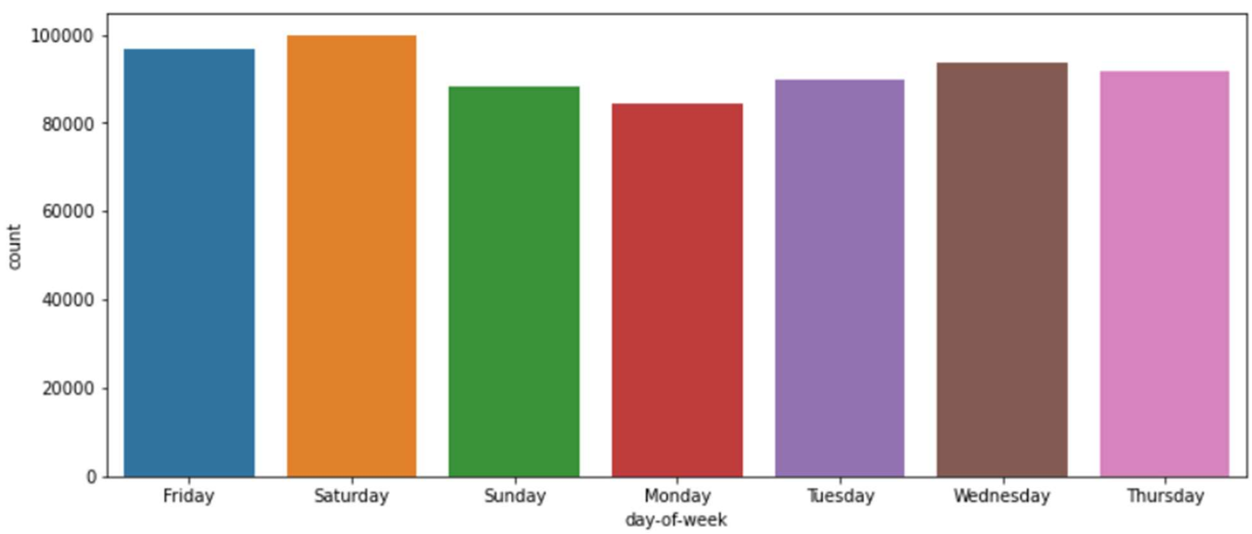

Figure 10 ridership through weekdays

The below two figures show that during the weekdays, riders tend to be the most during morning time and that is between 7-8 am and again in the afternoon at around 5-6 pm which is as previously found out is due to the nature of the users during weekdays which are more likely to be students and employees. Compared to the peak hours during the weekend which is around 12 pm.

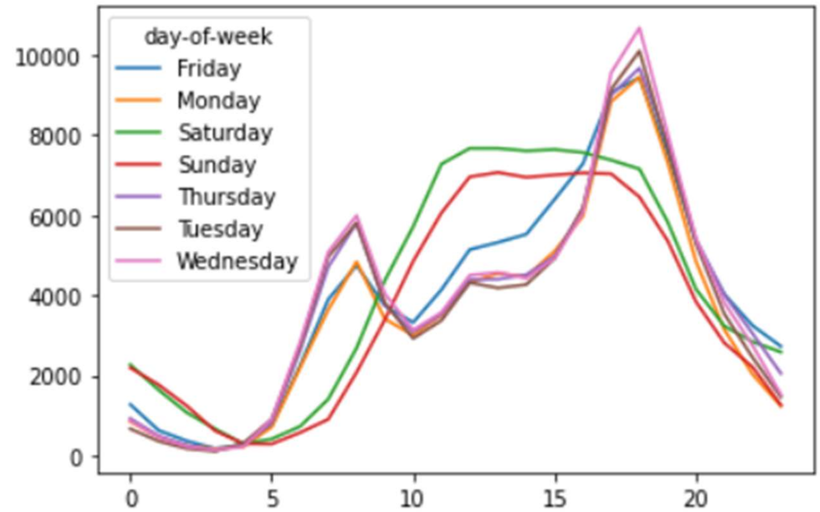

Figure 11 start hour throughout weekdays

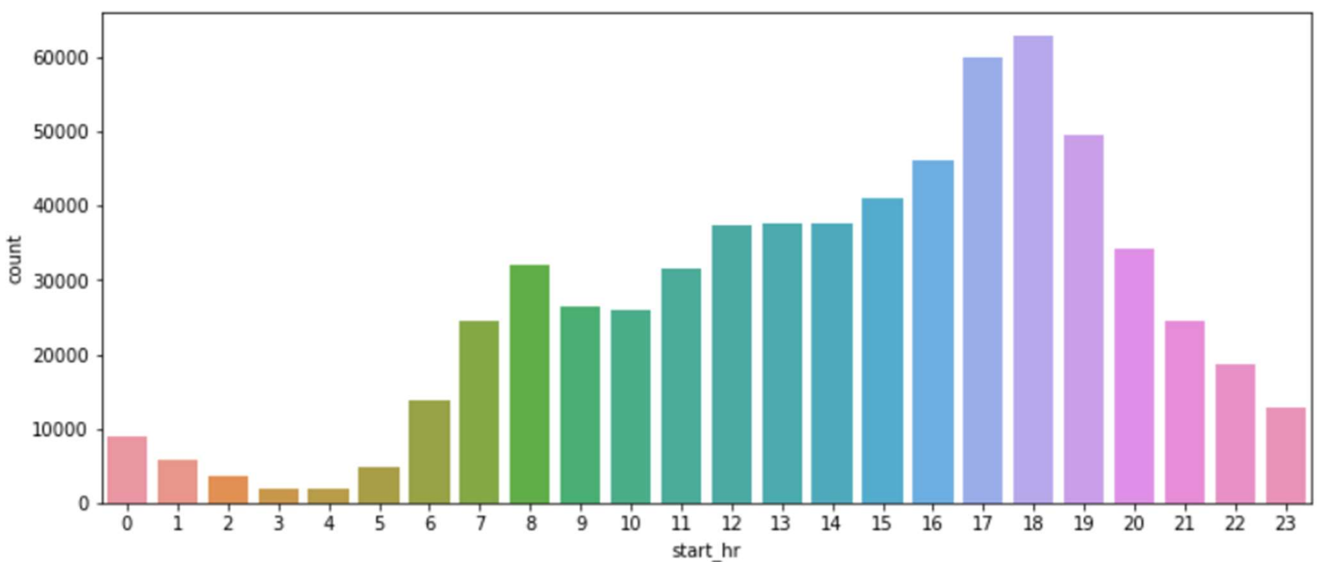

Figure 12 ridership count throughout the day

3. Trip Duration

As can be seen from the figure below, the distribution of trip duration is left skewed where a number of trips lasted less than 5 minutes. Trips were very minimal after around 65 minutes

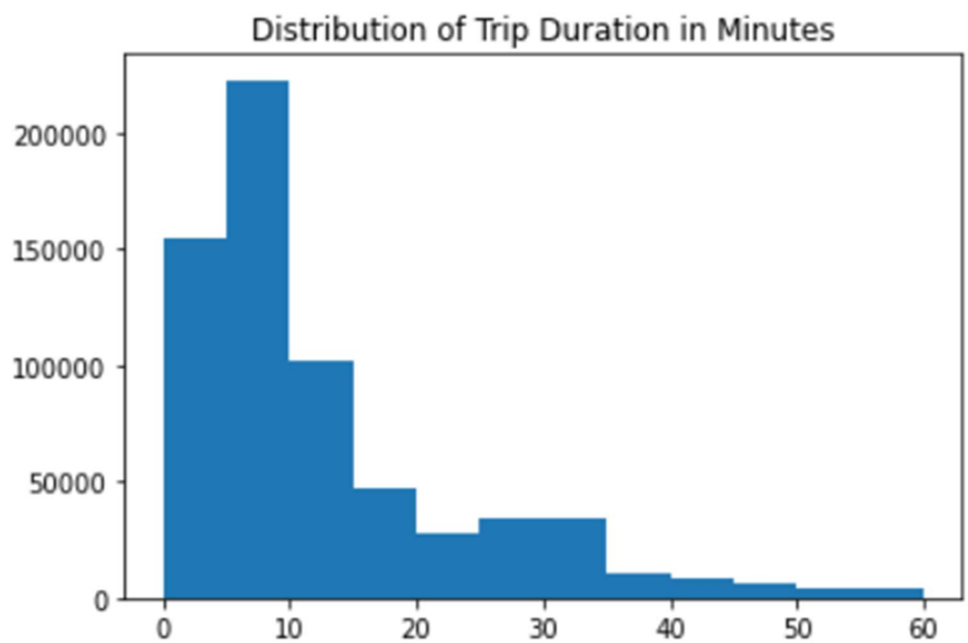

Figure 13 Trip duration distribution

As can be seen from the figure below, trip durations tend to be short for members while its way lengthier for casual members. This is due to the nature of trips being made, members usually drive to their work or university while for casual user its mostly for entertainment.

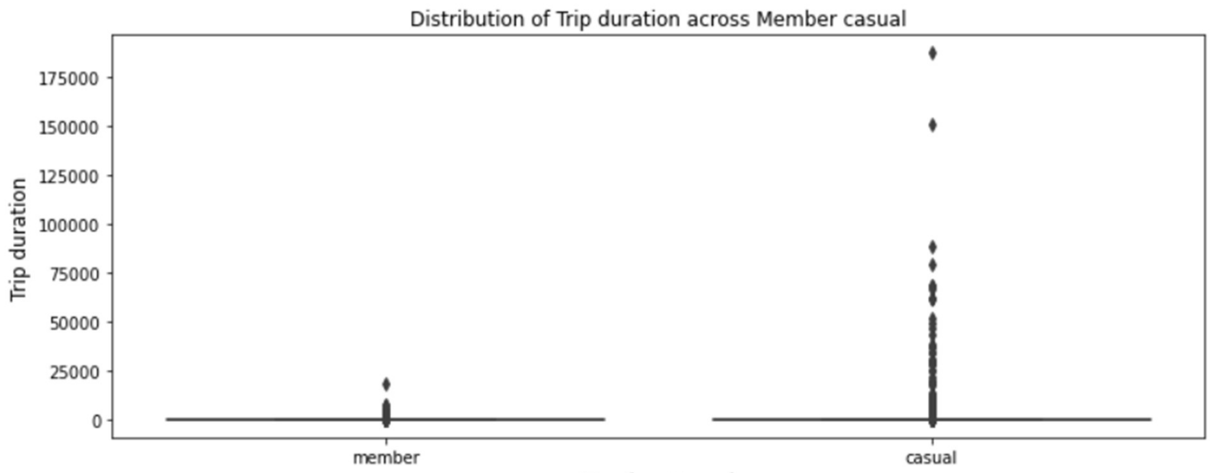

Figure 14 trip duration for member/casual users

As for the months distribution for trip durations, June has the highest distribution rate due to the excellent weather conditions

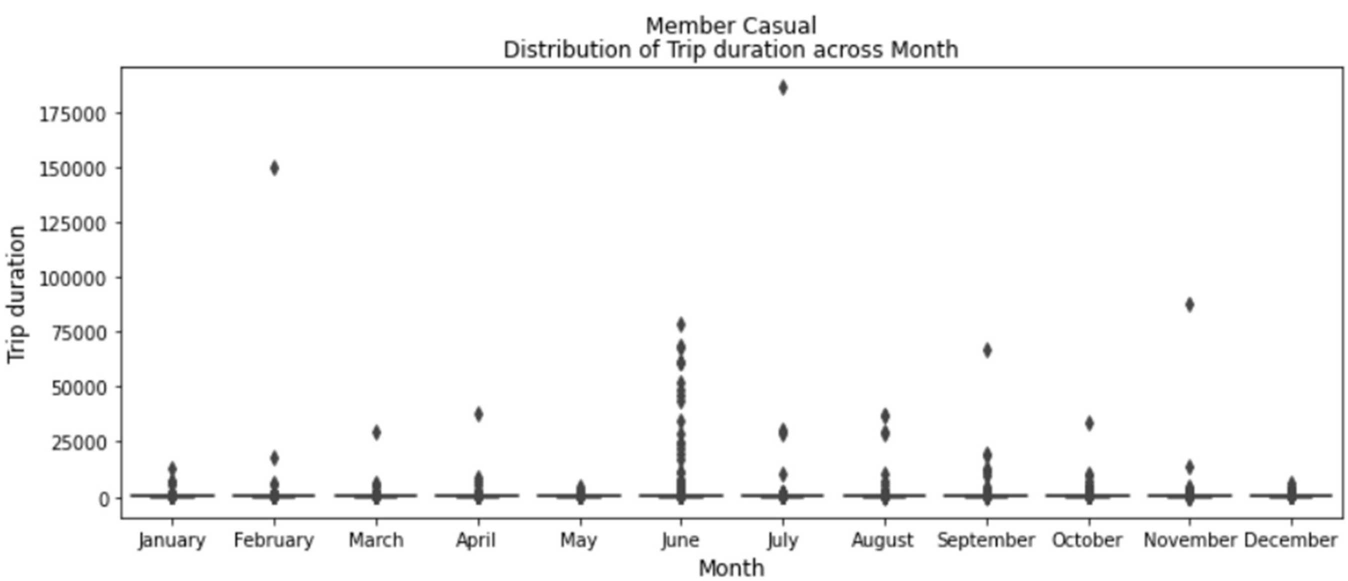

Figure 15 trip duration for different months

4. Types of bikes – classic, docked, electric

The figure below shows that most users both casual and member used classic bikes the most. But as for casual users their second preferred were electric bikes compared to docked bikes for member users.

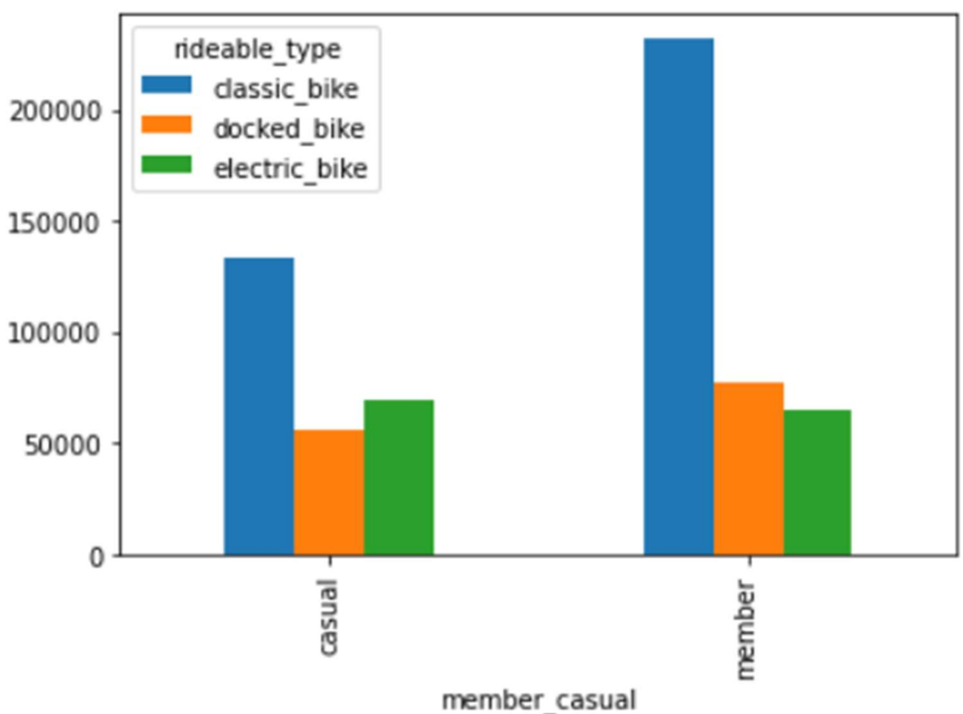

Figure 16 rideable type for member/casual users

5. Top stations

Finally, the blow figure shows the top start and end stations, Groove ST PATH followed by South waterfront walkway followed by Hoboken terminal were the ones with the highest demand for both start and end stations.

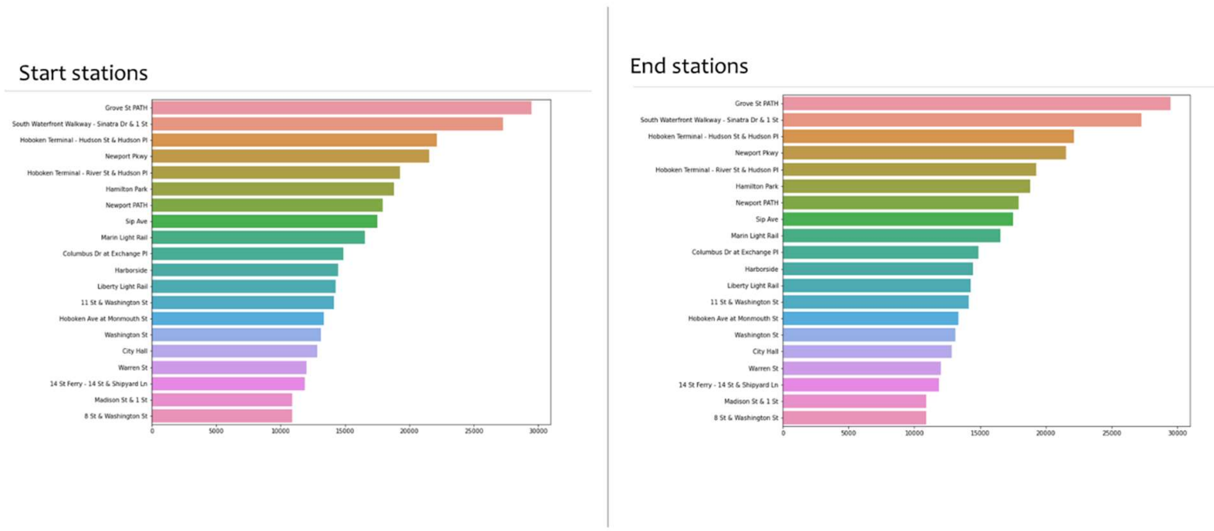

Figure 17 Top Station Distributions

From the observations above the following features were selected for model creation which are member_casual, season, month, day-of-week

4.4 Modeling

The data was aggregated by counting hourly number of rides for each combination of *'member_casual','season','month','day-of-week'*, these grouping variables will be used as the predictors during modelling.

Five different models were chosen as described below:

1. **Poisson Regression**

   It is a generalized linear model used primarily to predict count data. it assumes that the target variable has a Poisson distribution. Since the count data in our case is left skewed and not normally distributed, Poisson regression was chosen.

Poisson regression is a log-liner model, it is a form of regression analysis used to model count data (number of bikes used in an hourly basis) especially when we have several categorical variables as independent variables.
A Poisson regression was used to model the number of rides as dependent on Season, Member/casual, day of the week and month.
below is the model equation:

$$log(\lambda_i)=\beta_0+\beta_1 x_1+\ldots+\beta_n x_n$$

where $Y_i\ Poisson(\lambda)$

λ_i **- is the mean number of counts in hour i**

The model shows that all the predictors significantly influence the average number of cases

	Metric	Poisson regression
0	MAE	2.414624e+01
1	SSE	3.883203e+06
2	MSE	1.225759e+03
3	RMSE	3.501084e+01
4	R_square	3.669083e-01

Table 2 Poisson Regression

2. Decision Tree

The decision tree is a supervised machine learning algorithm used for both classification and regression problems. It is a method for decision making over time with uncertainty. Decision trees classify data by sorting them down from the root to one of the leaf nodes, where the leaf represents the classification to the data. The process is recursive and is repeated for every subtree. Hunt's algorithm is usually used to build decision trees. It starts by assuming that all data instances belong to the majority class. The best attribute is chosen to be the root of the tree based on how well the attribute splits the data into diverse groups using a node purity metric. Leaf nodes are created when a decision from a node split leads to data instances of only one class. Internal nodes are made using the best remaining attribute when a decision leads to data instances of two or more classes. The decision tree classifier/regressor was built/fit using the training set and used to predict new values from the test set. Inside the classifier, the node purity measure was specified as "entropy" for the information gain

The general concept of decision tree regressor is the same as the decision tree classifier, we recursively split the data using a binary tree until we are left with leaf nodes. Regression trees are type of decisions trees where each leaf node represents numerical value in contrast to classification trees where leaf nodes have binary values (True or False) or other discrete category.

	Metric	Regression Tree
0	MAE	1.307769e+01
1	SSE	1.431271e+06
2	MSE	4.517902e+02
3	RMSE	2.125536e+01
4	R_square	7.666550e-01

Table 3 Regression Tree Performance

3. Support Vector Machine

Support Vector Machine (SVM) is a popular supervised machine learning algorithm that analyzes the data where each item is plotted as a point in n-dimensional space for both classification and regression problems.

The target of SVMs for regression problems is to create the largest possible hyperplane where instances are fit within that hyperplane in addition to limiting margin violations.

	Metric	Support vector machine
0	MAE	2.466917e+01
1	SSE	4.000457e+06
2	MSE	1.262771e+03
3	RMSE	3.553548e+01
4	R_square	3.477920e-01

Table 4 Support Vector Machine

4. Random Forest

Random Forest was introduced in 2001. It is a widely used machine learning algorithm. In random forest we have different decision trees that are being created differently. The sum of these decision trees creates a random forest.

Random forest can manage both classification and regression problems. In our case the problem is a regression one. By building more trees in random forest we get better chance to reach the correct prediction and reduce the chance of overfitting of a single tree.

Random forest was chosen for its ease of use and flexibility in addition to these three reasons

- it can manage both categorical and numerical values without normalization
- it can predict the relative importance for the different variables used
- it can manage big data with fast computational speed

The random forest method is like the decision tree method but addresses its weaknesses.

	Metric	Random Forest
0	MAE	1.223598e+01
1	SSE	1.369371e+06
2	MSE	4.322509e+02
3	RMSE	2.079064e+01
4	R_square	7.767469e-01

Table 5 Random Forest

5. Boosted Regression

Boosting is defined as a method which aims to combine several simple models into one composite model that outperform the performance of the simple models from which it was made.

Boosted regression is a recent machine learning technique that has shown considerable success in predictive accuracy. [23]
In boosted regression, weak models are being ensembled to give a better prediction model.

Gradient boosting regression is used to predict numerical outputs, so the dependent variable as in our case count must be numeric.

Gradient Boosting is an efficient machine learning algorithm. It is a sort of boosting algorithm or It assumes that when the best potential next model is coupled with prior models, the overall prediction error is minimized. Friedman developed Gradient Boosting Regression Tree. It can be used for both classifications and regression problems. gradient descent is used to minimize the loss in the model and thus the term “gradient” in “gradient boosting.”

The results obtained from running the model is as seen in the table below:

	Metric	Gradient boosted
0	MAE	1.521024e+01
1	SSE	1.649527e+06
2	MSE	5.206839e+02
3	RMSE	2.281850e+01
4	R_square	7.310721e-01

Table 6 Gradient Boosting

6. K-Nearest Neighbors

K-Nearest-Neighbor (KNN) is a non-parametric, lazy learning algorithm and is considered among the simplest supervised machine learning algorithms. KNN has no explicit training phase and does not build the model explicitly. KNN is based on feature similarity and relies on the assumption that similar things exist in proximity. KNN computes distances to other training records (using a distance metric like Euclidean), identifies k-nearest neighbors and determines the class label of the unknown record by taking the majority vote (mode) of the class labels of k-nearest neighbors. KNN has been chosen for its simplicity, ease of implementation and high accuracy.

In regression problems as in our case, the KNN algorithm will predict a new data point's continuous value by returning the average of the k neighbors' values.
The results achieved is as follows:

	Metric	Gradient boosted
0	MAE	1.306292e+01
1	SSE	1.532666e+06
2	MSE	4.837961e+02
3	RMSE	2.199537e+01
4	R_square	7.501243e-01

Table 7 KNN

4.5 Tuning the Models

Hyperparameters are values that cannot be determined using the training data set, but at the same time their values determine the accuracy of the model created. Different algorithms can be used to determine the values of the hyperparameters. In our case the *grid search* and *cross validation* algorithms was chosen then the accuracy of the different models created were compared.

1. Tuning Decision Trees

The following are the main parameters that were set for better performance:

- **Splitter:** best strategy to split each node
- **max_depth:** which indicate the maximum depth of the tree
- **min_samples_leaf**: minimum number of samples needed to be at a leaf node

- **min_weight_fraction_leaf minimum** fraction of weights in each leaf
- **max_features**: number of features to look at for the best split
- **max_leaf_nodes:** maximum number of leaf nodes

2. Tuning Random Forest

The following are the main parameters that were set for better performance:

- **N_estimators:** then number of trees to be built
- **Max depth:** maximum depth of each tree
- **Min_samples_leaf:** minimum number of samples in a leaf node
- **Max_features:** max numbers of features to look at when splitting a node
- **Max_leaf_nodes:** maximum number of leaf nodes in the random forest model

3. Tunning Support Vector Machine

The following are the main parameters that were set for better performance:

- **Gamma:** It is the kernel coefficient for 'rbf,' 'poly,' and 'sigmoid', it defines how far the influence of a single training example reaches
- **C:** regularization parameter

4. Tuning gradient boosted

The following are the parameters that were set for better performance:

- **N_estimators:** this value indicates the number of trees in the forest.
- **Max_depth: the depth of the built tree**
- **Min_samples_split:** minimum number of samples needed to split an inner node
- **Learning_rate:** the speed or rate at which gradient boosted algorithm updates the parameter estimates or learns the values of the parameters.

5. Tuning KNN

The following are the parameters that were set for better performance:

- value of K or number of neighbors
- distance of new data with training data.

- nearest K-neighbors from the new data
- New Data Class Calculation

4.6 Model comparisons

The following are the key outcomes of the model comparison:

- Boosted regression model had the highest performance in terms of R-squared 81.18% followed by Regression tree with (77.76%). the worst model was Poisson regression with 36.69%
- The model shows that all the predictors significantly influence the average number of rides
- For customer group, casual member was used as the reference. while holding other factors constant, the expected number of rides on the group for members is $1\text{-}e^{(0.3851)}$ = 47% times higher compared to casual.
- For seasons, Autumn was used as the reference category, there is a significant difference on mean number of rides between the rest of seasons and Autumn. While holding other factors constant the expected number of rides in winter is $1\text{-}e^{(-0.3831)}$ = 18% times higher, compared to autumn, while in summer the expected number of rides is 196%. The expected number of rides on spring is 13% higher than in Autumn.

CHAPTER 5

5.1 SUMMARY & CONCLUSION

The challenge that majors cities in the world are facing is having to create a modal shift from the use of private transport systems onto the public transport system. The aim is not just to reduce congestion and/or enhance the safety as stated earlier but also with the aim that they want to develop a sustainable, smart, and environmentally friendly system with a goal to provide happiness to the public.

One of the major problems that is facing the shift in the e-bikes sharing systems is the rebalancing problem, in this research we aimed to create models that can predict the hourly bikes needed and further enhance it to make it relevant to the main stations where most of the rebalancing problem occur. Different machine learning algorithms have been tested, namely Poisson regression, decision tree, support vector machine, KNN, random forest and boosted regression. After tunning the models, the results came as in the table below:

	Metric	Poisson regression	Decision tree	Random Forest	Support Vector	Gradient boosted	KNN
0	MAE	2.414624e+01	1.260543e+01	1.552338e+01	1.806002e+01	1.205722e+01	1.376210e+01
1	SSE	3.883203e+06	1.369706e+06	1.748227e+06	2.726822e+06	1.154442e+06	1.408034e+06
2	MSE	1.225759e+03	4.323566e+02	5.518393e+02	8.607391e+02	3.644071e+02	4.444551e+02
3	RMSE	3.501084e+01	2.079319e+01	2.349126e+01	2.933836e+01	1.908945e+01	2.108210e+01
4	R_square	3.669083e-01	7.766923e-01	7.149807e-01	5.554371e-01	8.117875e-01	7.704435e-01

Table 8 Tunning Outcomes

Boosted regression model had the highest performance in terms of R^2 of 81.18% followed by Decision tree with (77.76%). the worst model was support vector machine with 55.54%

The models can predict which stations have the highest demand and what factors that affects it and its operation, as an example the highest stations are in Manhattan and the highest demand occurs between 7-8 am and again in the afternoon at around 5-6 pm. During the months of summer (August and September).

The developed models can be utilized for **any** shared mobility device current and/or future to assess areas of high demand and peak times throughout the year.

As per the findings, Manhattan has the highest density of rides despite it being geographically the smallest (which results in less ability to use personal cars and taxis in daily commutes) and that is due to its high population and the concentration of different attractions and famous streets that would attract tourists, politicians, artists, in addition to its residents. In addition, Manhattan has the highest annual income compared to other New York city boroughs however, it is not convenient for residents to use their cars and depend more on public transportation for their daily commutes.

5.2 RECOMMENDATION

1. Based on the available data, vital information is still missing which includes the number of docks per station. This information is of importance to estimate and assess among those high demand stations what is the percentage of occupation it reaches and how / when to be able to redistribute the bikes ahead of the peak hours to cater to the high demand.
2. In docking Stations, real time sensors need to be installed to collect further information about the weather (temperature, wind speed, humidity…etc.)
3. Service providers need to assess the use of modern technologies such as self-locking shared mobility devices rather than docked shared mobility devices which will have a positive impact in tracking and increase in spaces allocated for the return of devices. In addition, change in devises can be part of the solution, which is the current trend in various cities. These days, electric scooters are taking over the place of bikes as they can be parked around the city easily, easily taken and returned (no need for any dock) smaller in size, faster with less effort required to operate.
4. Service providers around the world are encouraged to provide their data in a form of open-source data that would allow researchers to study and assess

enhancement to the system that would encourage the transformation towards micro-mobility which in turn will positively impact the increase of percentage of public transportation within congested cities.

Works Cited

[1] C. M. Vallez, M. Castro and D. Contreras, "Challenges and Opportunities in Dock-Based Bike-Sharing," MDPI, Spain, 2021.

[2] Shaheen, Guzman and H. Zhang, "BIKESHARING IN EUROPE, THE AMERICAS, AND ASIA:," 2010.

[3] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data," 2000.

[4] S. Shaheen, N. Chan, A. Bansal and A. Cohen, "Shared Mobility, Definitions, Industry Developments, and Early Understanding," 2015.

[5] M. Cheng, "Sharing economy: A review and agenda for future research," *International Journal of Hospitality Management,* pp. 60-70, 2016.

[6] M. Hossain, "Sharing economy: A comprehensive literature review," *International Journal of Hospitality Management,* vol. 87, May 2020.

[7] S. A. Shaheen, S. Guzman and H. Zhang, "BIKESHARING IN EUROPE, THE AMERICAS, AND ASIA:," 2010.

[8] L.-Y. Qiu, He and Ling-Yun, "Bike Sharing and the Economy, the Environment,," MDPI, China, 2018.

[9] M. Oteroa, Nieuwenhuijsen and D. Rojas-Rueda, "Health impacts of bike sharing systems in Europe," *Environment International,* pp. 387-394, 2018.

[10] E. Eren and V. E. Uz, "A review on bike-sharing: The factors affecting bike-sharing demand," *Sustainable Cities and Society,* 2020.

[11] L. Sun, W. Chen, J. Liu and H. Xiong, "Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization," in *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*, San Fransisco, USA, 2016.

[12] G. Cantelmo, R. Kucharski and C. Antoniou, "Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data," p. 132–144, 2020.

[13] H. Ashqar, M. Elhenawy, M. Almannaa, A. Ghanem, H. Rakha and L. House, "Modelling Bike Availability in a Bike-Sharing System Using Machine Learning," Naples, Italy, 2017.

[14] B. Wanga and I. Kim, "Short-term prediction for bike-sharing service using machine," *Transportation Research Procedia,* pp. 171-178, 2018.

[15] D. Chai, L. Wang and Q. Yang, "Bike Flow Prediction with Multi-Graph Convolutional Networks," Washington, DC, USA, 2018.

[16] S. V. E and Y. Cho, "A rule-based model for Seoul Bike sharing demand," *European Journal of Remote Sensing,* 2020.

[17] F. Chiariotti, C. Pielli, A. Zanella and M. Zorzi, "A Dynamic Approach to Rebalancing Bike-Sharing Systems," *The Architectures, Systems, and Applications of Internet of Things for Smart Cities,* 2018.

[18] J. Liu, L. Sun, W. Chen and H. Xiong, "Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization," in *KDD: Knowledge Discovery and Data Mining*, 2016.

[19] DeMaio, "Bike-sharing: History, Impacts,Models of Provision, and Future," *Journal of Public Transportation ,* vol. 12, pp. 41-56, 2009.

[20] Chen, Turon, Klos, Czech, Pamula and Sierpinski, "Fifth Generation of Bike-Sharing Systems," *Scientific Journal of Silesian University of Technology. Series Transport,* vol. 99, 2018.

[21] E. Eren and V. E. Uz, "A review on bike-sharing: The factors affecting bike-sharing demand," *Sustainable Cities and Society,* vol. 54, 2020.

[22] E. Eren and V. E. Uz, "A review on bike-sharing: The factors affecting bike-sharing demand," *Sustainable Cities and Society,* vol. 54, March 2020.

[23] M. Schonlau, "Boosted Regression (Boosting): An Introductory Tutorial and a Stata Plugin," 2005.

[24] C. M. Vallez, M. Castro and D. Contreras.

[25] X. Jin and D. Tong, "Station-Free Bike Rebalancing Analysis: Scale,Modeling, and Computational Challenges," *International Journal of Geo-Information ,* 2020.

[26] D. Chai, L. Wang and Q. Yang, "Bike Flow Prediction with Multi-Graph Convolutional Networks," 2018.