MSc Data Science & Analytics Thesis

Analysing and Estimating Bike-Share Trips in Dublin Bike Sharing System

Author

Sreevathsa Devanahallibokksam

Student Number - 21250214

Supervisor

Dr. Joseph Timoney

A thesis submitted in fulfilment of the requirements for the degree of MSc in Data Science and Analytics 2021-2022

in the

Department of Computer Science Maynooth University

August 08, 2022

Declaration

I hereby certify that this material, which I now submit for assessment on the program of study as part of MSc. in Data Science and Analytics qualification, is entirely my own work and has not been taken from the work of others - save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Sreevathsa Devanahallibokksam Date: 08-08-2022

`

Acknowledgements

First and foremost, I am extremely grateful to my supervisor, Dr. Joseph Timoney, for his invaluable advice, continuous support, and patience during my project work. Second, I would like to thank all the members of the Mathematics and Statistics Department Faculty. Their immense knowledge and plentiful experience have encouraged me throughout my academic research and daily life. Their kind assistance and support have made my studies and life at Maynooth University a memorable experience. Finally, I would like to express my gratitude to my parents and family members. Without their tremendous understanding and encouragement over the past few years, it would be impossible for me to complete my studies.

`

Abstract

A public bicycle rental program called Dublin Bikes has been operating since 2009. Dublin was the 17th city to launch a similar program, which quickly became one of the most popular dock-based bicycle sharing programs worldwide due to the bike's affordability, environmental friendliness, and convenience. This dissertation examines Dublin Bikes data, which includes seasonal, hourly, weather, public holiday, and geographic information, to better understand bike availability, usage patterns, activity, and journeys across all stations. K-means clustering is implemented on this transformed data to find similar stations based on their usage pattern. This report's primary focus is on implementing Chardon's methods for predicting bike-sharing trips taken in a system without journey-level information.

Additionally, other classification and regression models have been proposed that forecast the possible number of journeys on any given day were also put forth. These models are compared and examined to establish which models predicts with better accuracy.

Although the suggested method may be less precise, it provides a solution for estimating daily trips without trip data or daily trip counts. This estimated trip data can generate discussion and analysis about the efficacy of a BSS and whether it may be related to local legislation, cycling infrastructure, scheme pricing, station density, urban structure, or provisioner management. Additionally, if an organization knows how many trips are going to happen on a particular day in advance, they can rebalance the bikes between stations to meet demand and improve customer satisfaction.

KEYWORDS: bicycle-sharing systems, docking stations, trip estimation, rebalancing

`

Table of Contents

`

`

List of Figures

`

List of Tables

`

Chapter One : Introduction

1.1 Bike Sharing Schemes

Bike-sharing schemes (BSS) are short-term urban bicycle rental programs that allow users to pick up and drop off their bicycles at any self-serve station, making them ideal for the shortest journeys. They are also referred to as Public-use bicycles (PUBs) or city bikes.

Generations of Bike Sharing System

In 1965, Amsterdam saw the debut of the first BSS generation. They were referred to as "white bicycles," and their intended usage was just once before being unlocked and made accessible to the following user. In Amsterdam, the BSS contributed to an increase in cyclists, particularly those without bicycles.

The second generation of the BSS was introduced in Copenhagen, Denmark. The ability to lock the bicycle, the implementation of a "coin refund" system that could enhance the customer experience (promoting the return of bicycles which in turn reduced theft), as well as the addition of an annual membership and fee were the key differences between the first and second generations of BSS. This was the first BSS, offering access to thousands of bicycles [1][2].

The third generation of BSS represents the most important development in the history of bike sharing systems, and it has also been successfully adopted outside of Europe for the first time, with advantages that have been broadly embraced across the world. In Rennes, France, the third generation was unveiled in 1998. In various aspects, the system is notably different from the first and second generations of bike-sharing programs, including:

1. Access through a smart card.
2. Automatic stations and docks.

3. Real-time information such as the number of bicycles and docks available for users.
4. The first 30 minutes are free per day.

Despite the fact that the bike sharing system was developed in Europe, it is currently being implemented in nations all over the world, including China, the United States, India, the United Kingdom, Mexico, Australia, and a few other countries, as shown in (Figure 1.1).

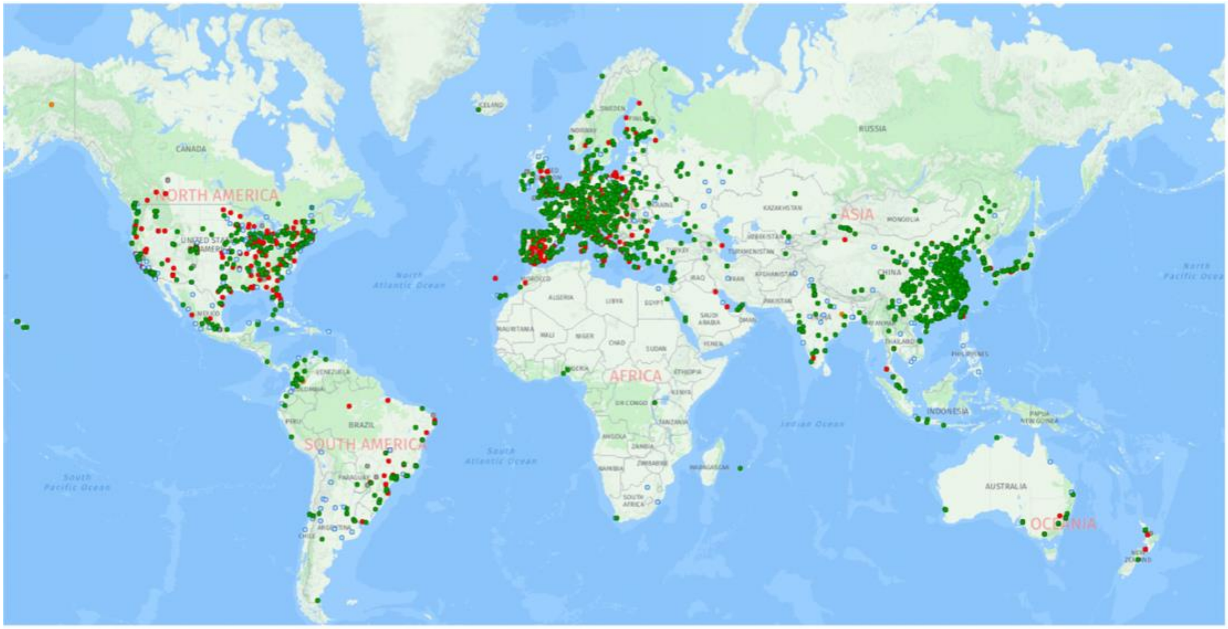


Figure 1.1 Implementation of BSS Worldwide [45]

1.2 Dublin Bikes

Dublin Bikes has been operational in Dublin since 2009 and is sponsored by JCDecaux. At the time of launch, there were 40 stations with 450 French-made unisex bikes. In 2011, four more stations were added, with the addition of 100 more cycles in Dublin city. In 2013, a multi-million euro expansion took place, bringing the total number of bikes to 1,500 by July 2014 and adding 58 new stations [3][4], making it the world's most successful bike-share program. However, a significant loss (€376,000 annually) occurred in 2016, which prevented the expansion of bikes. On July 20, 2017, Just Eat partnered with JCDecaux for the next three years by investing €2.25 million in the

`

scheme. Since then, Dublin Bikes has become one of the world's most successful bike rental schemes, with more than 58,000 subscribers and 2.2 million rentals [6]. The secret behind the success of this scheme is its affordability, with an annual fee of €35 and the first 30 minutes of each journey being free, after rental charges apply, and also no fear of robbery. The service is free for subscribers as 97 % of trips within the city take just under 30 minutes [5][6]. “Now TV” and “JCDecaux” have formed a partnership to provide a bike-sharing service in Dublin. The docking station, which can be seen in Figure 1.2, is where the bikes are parked when they are not in use. Figure 1.3, depicts how the number of Dublin stations has changed throughout time.

Figure 1.2 Picture of Dublin Bike Stands [46]

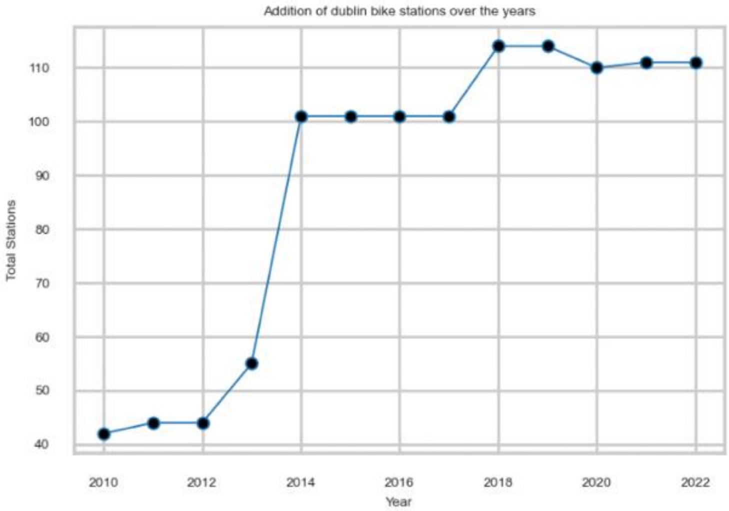


Figure 1.3 Addition of Dublin bikes stations over the years

`

1.3 Motivation

Cities like Paris, London, Barcelona, London, Chicago, Montreal, Moscow, and Dublin use dock-based bike stations for their bike-sharing schemes. These bike-sharing programs provide real-time data on the number of bicycles accessible at each of their docking stations, allowing researchers studying urban mobility to better understand bike journeys at each docking station. Several studies have been conducted to determine the load balance between different stations and make a precise forecast of bike availability at specific intervals of time. However, there has been little study on the quantitative analysis of bike journeys at docking stations to investigate spatiotemporal changes in how specific bike stations are used. The paucity of research is due to a lack of relevant data, which does not include origin and destination data for the cycle in the station-level data. Despite the challenges, several studies have developed a method to analyse or predict bike trips using station-level data in big cities worldwide. Still, research on the Dublin bike-sharing scheme has been rare.

1.4 Problem Statement

The Dublin Bikes dataset lacks origin-destination and cycle-level information, making it difficult to estimate bike journeys that take place daily. As bike journeys were a crucial indicator of the bike sharing scheme's efficacy, a technique was needed that could estimate how many journeys take place each day. This project proposes a technique and implement Chardon’s method to calculate the number of bike journeys and analyse the pattern of bike journeys across the Dublin bike stations using different influential factors.

1.5 Approach

According to the thesis, there are five main stages in which the research was conducted:

- Gathering of information, knowledge, and data related to Dublin Bikes.
- Data cleaning and Data Transformation.

`

- Exploratory Data Analysis and Data Visualisation.
- Predicting Bike Journeys using Chardon's technique, Regression and Classification Models.
- Evaluating and comparing the results from the models.

1.6 Metrics

To evaluate the accuracy of Chardon's four linear regression models, all linear assumptions (linearity, normality, autocorrelation, and homoscedasticity) are verified and the best model is chosen. Regression models were assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Mean Square Logarithmic Error (RMSLE), and Normalised Root Mean Squared Error (NRMSE), whereas classification models used confusion matrix, accuracy score, precision score, recall score, and F1-score. All of the metrics that are being highlighted here will be discussed in Chapter 3.

1.7 Dissertation Outline

There are seven chapters in this dissertation, which are outlined below.

- Chapter One is the "Introduction", where the overview of the project is presented, including an introduction to bike sharing schemes and Dublin Bikes, the motivation, problem statement, approach to the problem, metrics, as well as the dissertation structure.
- Chapter Two is the "Background Literature Review", which discusses the evolution of BSS and its implementation across the globe. The literature review typically begins with analysing the various research implemented on bike-sharing systems worldwide before exploring the Dublin Bike Sharing System. Additionally, this chapter briefly explains the models implemented to estimate bike-share trips using station-level data.

`

- Chapter Three is the "Methodologies Used", and it is meant to give a quick overview of the terms, definitions, calculations, algorithm, predictive models, and statistical tests that have been used throughout this study for better understanding.
- In Chapter Four, "Data Pre-Processing and Feature Engineering", the characteristics of the Dublin Bikes dataset, the methods involved in data cleaning, and data transformation are discussed.
- In Chapter Five, "Exploratory Data Analysis," transformed data are analysed to get insights about bike activity, bike availability, and bike-sharing trips across the stations. This analysis uses month, day of the week, holidays, weather, seasonal, and spatiotemporal information as influential factor and clustering analysis on stations is also performed to identify similar behaviour between stations.
- In Chapter Six, "Model Building and Evaluation," discusses about implementation and evaluation of results obtained from Chardon's model and few traditional machine learning models to estimate bike-sharing trips based on station-level data.
- Finally, Chapter Seven, "Conclusion and Future Work," concludes and summarizes the research contributions, discusses the research's limits as well as the scope for future work.

Chapter Two : Background Literature Review

In this chapter, we look at the existing literature on the bike-sharing system and, more specifically, the depth of the literature and research on the Dublin-bike sharing system. The literature review typically begins with analysing the research on bike-sharing systems worldwide before exploring the Dublin Bike Sharing System.

Public cycling programs, also known as bike-sharing schemes, have become more popular in recent years as more people started to support them due to their ability to improve first-mile and last-mile connections to other modes of transportation and reduce environmental impact. Over the past 45 years, three different generations of bike-sharing systems have been launched [7][8]. The first bike-sharing program was launched in Amsterdam in 1965, and the second generation of bike-sharing schemes was launched in Denmark (Fars, Gren, Denmark) in 1991–1993. These programs were launched on a small scale, with fewer bicycles available for use by the general public. Having seen the success rate of the bike sharing schemes, Copenhagen was able to launch them on a large scale in the year 1995. Although the system was free to use, many bicycles were damaged or stolen due to the user's anonymity. This enabled the development of third-generation bikes, which required magnetic stripe cards from users to operate. Third-generation cycles were adopted in nations including France, Paris, Brazil, and Taiwan due to the benefits of the new technology. Around 60 third-generation bike sharing schemes were launched worldwide by the end of 2007 (DeMaio 2007). However, research and study on these programs were extremely rare because there was a lack of data up until the advent of third-generation bike-sharing schemes. Due to this, the bike-sharing companies started to record docking and undocking information at each station.

There are two types of data in the bike-sharing system:

Trip-level data: A time, date, and duration for each bicycle trip between an origin and destination station is provided for each trip. Currently, this type of information is only shared by a very small number of bike sharing companies.

`

Station-level data: The number of bikes and docks currently available at every station are made available in this dataset. Periodic API requests or consistent scraping are used to collect this data.

The first research to look at such recorded data was conducted on Bicing, the bike-sharing program in Barcelona. On a dataset that covered a period of 13 weeks, geographical and temporal analysis were performed, and clustering methods were utilized to discover patterns among stations. Also, four predicting models were proposed to predict bike availability at each station [9]. Ahmadreza Faghih-Iman et al. [10] used a mixed linear model in order to study and examine how user arrival and departure times are impacted by bike infrastructure as well as socio-demographic and land-use characteristics. They also developed a binary logistic model and a regression model to estimate rebalancing between the time periods. Using Bicing dataset by Spain BSS, the time series models have been utilized in other research on the availability and use of bikes to assess user mobility data utilizing variable like the number of bikes at docking stations. In addition, this study looked into Barcelona's temporal and spatial movement patterns and used those trends to forecast when bikes would be accessible at a docking station [11]. Xiaolu Zhou in the year 2015, analysed the massive volumes of Chicago's BSS dataset. This study revealed users' distinct weekday and weekend travel patterns by using fast greedy algorithms to detect spatial communities of bicycle flows. Furthermore, the hierarchical clustering approach was used to study bike demand patterns [12]. Yuanxuan Yang et al. [13] analysed the dockless bike sharing scheme in Nanchang, China over a period of time when new metro line was introduced. This study revealed that introducing a metro service increased bike demand over time using geographic data. From this analysis, we can derive that people utilise bicycles for last-mile connectivity. Kaltenbrunner et al. [14] utilized an ARIMA time series model to estimate the number of bikes available. Giuseppe et al. [15] studied the "ToBike" bike-sharing system in Italy. Using a sparse optimization approach and closed queueing networks, this study estimated customer arrival rates that are not available in the dataset. Based on the parameters obtained, the researchers predicted the system's behaviour based on an unknown validation dataset.

`

Li et al. [16] predicted the pick-up and drop-off of bikes at each clustered station using a bipartite clustering algorithm employing station locations and their transition pattern information. In similar research, Divya Singhvi et al. [17] proposed a model to predict the number of bikes available at a specific interval by considering taxi use, nearby station usage, and weather. Vogel et al. [18] studied Vienna, Austria's cycling information using a data mining method to investigate user and bike activity trends. To forecast the number of bikes, Feihu Huang et al. [19] developed a unique methodology using a Bimodal Gaussian Inhomogeneous Poisson algorithm. This research uses a Bimodal Gaussian function to explain the intensity of an Inhomogeneous Poisson process that describes the user's arrival to pick up or return bikes at bike stations. In addition to this research, the check-in and check-out orders are used to predict the availability of bikes. Similarly, several studies and research use weather, seasons, and public holiday information to estimate bike availability and develop rebalancing techniques. Even though there have been numerous studies on bike availability and rebalancing strategies, every one of them is unique because the analysis depended on a various of factors, including the city's bike-sharing program, the factors taken into account during the analysis, the algorithms used for predictions, the size and quality of the dataset, etc.

In 2009, JCDecaux launched the Dublin Bikes program, which aims to promote cycling in the city by introducing 450 bikes to be used at forty active stations in the city. The organization has further adopted a station-level data architecture that allows them to collect data at each of the stations, in order to ensure that the collection of data is as efficient as possible. As soon as a sufficient amount of data was gathered for analysis, Peter Mooney et al. [20] 2010 published preliminary results from the Dublin Bike Data, which described the check-out activity of bikes at different stations, as well as the reasons for the busy and quiet stations. As part of this analysis, a GIS tool was also used to visualize the activity levels at each station using the data. Once the bike-sharing system had been increased to 101 stations, Pham Thi et al. [21] conducted several studies using statistical analysis to determine which stations were the busiest and quietest. To find time-dependent trends within the clusters, stations were organized into groups using clustering algorithms based on the availability and activity of bikes. Additionally, they have examined and depicted how stations behave on weekends, casual days, and public holidays. This study showed that, in comparison to regular days, activity and bike availability are drastically reduced during public holidays. The research on Dublin bikes was continued by Joe Timoney et al. [22] where they analysed the behaviour of a few active stations, considering many influential factors. All the Dublin stations are grouped into 4 clusters using K-means clustering to visualize the mean bike availability throughout the day. They found an interesting pattern where two of the clusters behaved quite the opposite of each other. With the results obtained, it is concluded that the stations with these patterns help the user commute from their house to their working place and vice-versa. A comparison of various predictive models for bike numbers using a dataset from a different date range was being conducted in this project to investigate an alternative approach to modelling bike availability numbers.

In the same year, Nidhin George [23] adopted the idea of Joe's research to further extend the prediction analysis on Dublin Bikes. The report uses machine learning techniques to develop regression and classification models. Day of the week, weather, time, holiday, and historical usage, are covariates. This study proposed two types of prediction models for predicting bike availability at stations: short-term predictions using regression algorithms such as Boosted Linear Regression, Random-Forest Regression, Gradient Booting, and LSTM; and long-term predictions using classification algorithms such as Random Forest Classifier, K-Nearest Neighbours, Linear, and Quadratic Discriminant Analysis.

Mishaal Ijaz, 2020 [24] extended Nidhin's study by conducting a similar analysis on the Dublin Bike dataset. This study looked at the activity and the availability of bicycles at bike stations. This study proposed regression and classification prediction models to estimate bike availability and activity at stations: Ridge Regression, Decision Tree, Random Forest, and Gradient Boosting for regression; and KNN, Random Forest, Logistic Regression, and Naive Bayes for classification. Nidhin's and Mishaal's study concluded that considering all these models to predict bike activity or bike availability would yield accurate results. According to both studies, the Random Forest algorithm performed well in regression and classification.

`

Despite extensive research on estimating bike activity and availability, there has been little research on bike-share trips because of a lack of appropriate data on the trip's source, destination, and cycle-level information, which makes estimation of bike-share trips challenging. However, in 2015, Chardon et al. [25] proposed a method to estimate bike-share trips based on station-level data. This research analyses multiple methodologies for estimating the number of daily trips, the crucial indicator of BSS usage, using commonly available data and the number of bicycles available at a station over time. This paper proposes four spatial and temporal aggregate models(Linear Regression Models) called the individual aggregated model (iDAM), the combined day aggregated model (cDAM), the interval aggregation model (IAM), and finally, the station aggregation model (SAM), to estimate the daily bike trips. In this model, the most important features used to create the model are the trip and active station counts at particular intervals. Without public access to trip data, this method can determine whether the BSS is the best investment for non-motorized transport modes. In the absence of trip data or daily trip counts, this study provides a method for estimating daily trips to generate discussion and analysis about the efficacy of a BSS and whether it may be related to local legislation, cycling infrastructure, scheme pricing, station density, urban structure, or provisioner management. James Todd et al. [26] adopted the techniques used by Chardon to accurately predict the number of journeys for a few cities like Paris, London, Chicago, and Moscow for which there is no journey-level data. Based on the results of this study, it can be concluded that the Chardon method is reasonably successful in cases where dock count data is only available.

On the Dublin Bike dataset, Maxim Le Cloerec [27] implemented all four models proposed by Chardon. As the proposed models are linear regression models, this project aimed to verify all of the linear regression assumptions made on residuals, such as linearity, homoscedasticity, normality, and independent residual error terms. However, none of the models satisfied the linear assumptions even after applying transformations such as the square root and Napierian logarithm. Our hypothesis is that the reason for the failure is the residuals being heteroscedastic. In addition to this, we believe that the trip calculation was not carried out correctly, that

`

identifying extremes for rebalancing values was not performed correctly, and, most importantly, that the data used to draw conclusions was significantly smaller than required.

In this project, our major goal is to implement Chardon's linear regression models [25] for predicting the trip generated by users using the Dublin bikes dataset, which was previously not possible. In each of these predictive algorithms, the total journeys resulting from both users and rebalancing activities are taken as input and the journeys resulting only from users are predicted. To compare and evaluate all these models, we verified the linear regression assumptions for all the models. The research was further extended by integrating other machine learning models (Regression and Classification) with appropriate variables, such as month, season, weekday, and public holiday information, in addition to Chardon's linear regression models. During the background study, we observed that the regression models (Linear Regression, Ridge Regression, Decision Tree, Gradient Boosting, and Random Forest Regressor) and classification models (Logistic Regression, KNN, Naive Bayes, Decision Tree, LDA, and Random Forest Classifier) implemented by Nidhin’s [23] and Mishaal’s [24] research, performed well with higher accuracy with similar dependent variables to estimate bike activity and availability at bike stations. As a result, we adopted similar models and metrics for estimating daily station bike trips. Detailed discussions of their implementation and results will be presented in later chapters.

`

Chapter Three : Methodologies Used

This chapter is intended to provide a brief overview of the definitions, terminologies, calculations, predictive models, and statistical tests used throughout this study in order to provide a better understanding.

3.1 Definitions & Calculations

The number of bike journeys is calculated via the following equation [25]:

$$\boldsymbol{trips = \frac{|interactions - rebalance| + interactions\ collisions}{2}}$$

Interactions(X): A bicycle interaction is when a cyclist removes or returns a bicycle from a BSS station for a trip. The value is calculated by subtracting the current bike availability from the previous time stamp's bike availability after ordering the data by timestamp and station ID. A negative number indicates the beginning, while a positive number indicates the end of the journey.

Rebalance(R): The change in the number of bikes can also be due to technical issues, rebalancing operations, or maintenance, referred to as "rebalancing" for simplicity.

- The station-level data makes it extremely difficult to determine whether a change in bike count is the result of user interaction or whether it is the result of rebalancing activity or routine maintenance that caused the change.

In the absence of origin-destination information, the calculations needed for rebalancing were quite challenging. Therefore, it was necessary to find an alternate method of estimating this rebalancing. Taking Maxim's base idea into consideration, we proposed a novel approach to this problem.

Algorithm to calculate Rebalance

Step 1: JCDecaux collects data every 10 minutes. Several samples had non-10-minute timestamps. To ensure quality, the recorded timestamp's minute value is rounded to the nearest minute that is divisible by 10.

Step 2: Filter the rows whose interaction value is greater than 0, which helps to remove the rows that had no bike-activity.

Step 3: Using the K-means clustering algorithm, stations are clustered based on mean bike activity at different times of the day, and the resultant cluster number is added to the original dataset.

Step 4: At intervals of 10 minutes, determine the average number of interactions for each cluster in the dataset.

Step 5: Using Inter Quartile Range, find the extreme values in each cluster's interaction column(lower and upper) limits and store them for use later.

- As the interaction values are heavily skewed and it doesn’t follow a normal distribution which can be clearly seen in density plot Figure 3.1, we decided to extend the probability for Q1 as 0.1 and Q3 as 0.9. So, that extreme values can be identified.

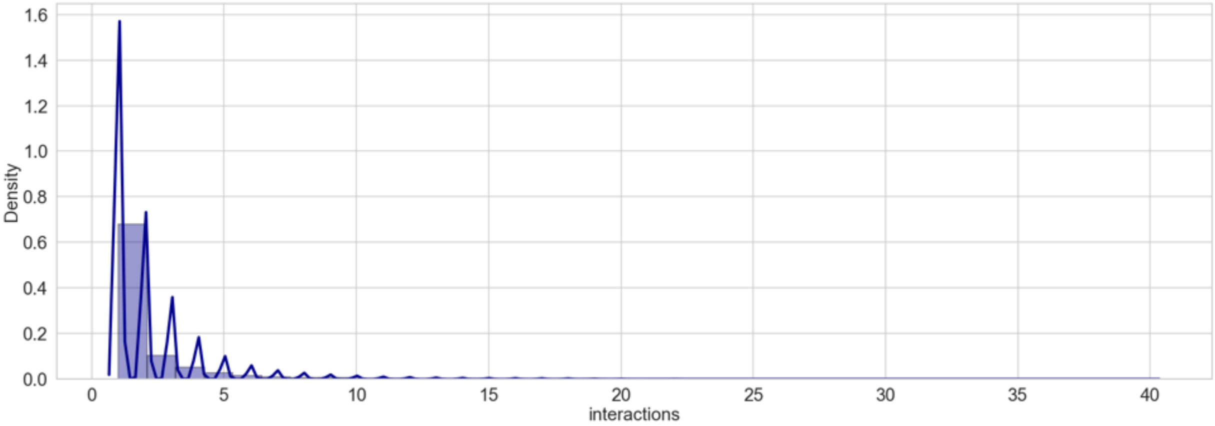


Figure 3.1 Density Plot for Interactions

`

Step 6: Create a new column called "activity_type" that you can use to classify the interactions into rental and rebalancing by using the criteria below.

```
if [interactions < lower limit] or [interactions > Upper limit]
then activity_type = “rebalance”
else activity_type = “rental”
```

Step 7: Create a column called "rebalance" using the below conditions.

```
if activity_type is "rebalancing"
then rebalance =  (interactions - cluster mean interaction)
else rebalance  = 0
```

Example :

Station 1 belongs to Cluster 1, and if the interaction value at 6:00 am for Station 1 is 13, but Cluster1's mean interaction value at 6:00 am is 4, then rebalance is 13-4 = 9.

This can be interpreted as nine interactions caused by rebalancing activity, and four interactions are caused by a user.

Interactions Collisions(C): An interaction collision occurs when there is a missing interaction (the removal and return of bikes) between observations that are not captured by the system.

Example:

Assuming the data from each station is recorded every 10-minute interval.

At 01-01-2022 00:00:00, there are 20 bikes available; at 01-01-2022 00:10:00, there are 20 bikes available. There might be a chance that three bikes were returned and three cycles picked up between the 10-minute intervals. This scenario clearly explains the loss of data.

- In order to calculate the number of collisions per day, we added the difference in bike changes without applying a modulus to the difference. The collision value should generally be zero if each bike has completed the trip and arrived at the station.

`

Using the calculations above, we can now obtain the values of interactions X, rebalancing R, and collision C. Now by using trips formula and values(X, R, C), we can calculate how many bike trips took place on a given day.

3.2 Chardon's aggregated models

We will discuss the four aggregated models that Chardon proposed in his paper to estimate bike journeys. These models are known as the individual aggregated model, the combined day aggregated model, the interval aggregation model, and the station aggregation model [25].

The following model is based on the hypothesis that there is a relationship between the number of trips per day ($\mathbf{T_d}$) and the number of trips which include both the rebalance and collision days ($\mathbf{T_{x\Delta d}}$) .

whereas $\mathbf{T_{x\Delta d}} = \frac{\boldsymbol{interactions}}{\mathbf{2}}$, $\mathbf{T_d} = \frac{|\boldsymbol{interactions - rebalance}| + \boldsymbol{interactions\ collisions}}{\mathbf{2}}$

Model specifications

3.2.1 Individual day aggregated model (iDAM)

In order to compensate for collisions that might occur on high activity days, a quadratic factor ($\mathbf{T^2_{x\Delta d}}$) is applied. Since we will be estimating the count data, it is crucial that we prevent the models from estimating negative values. Therefore, the intercepts are locked to the origin.

The iDAM equation is given by :

$$\mathbf{T_d = 0 + \beta_1 T_{x\Delta d} + \beta_2 T^2_{x\Delta d} + \varepsilon}$$

`

3.2.2 Combined day aggregated model (cDAM)

The second model (cDAM) uses the same hypothesis as the iDAM model, but adds a parameter for active stations in order to normalize the density of activity.

The cDAM equation is given by :

$$\mathbf{T_{d} = 0 + \beta_1 T_{x\Delta d} + \beta_2 (T^2_{x\Delta d} / A_{x\Delta d}) + \varepsilon}$$

$\mathbf{A_{x\Delta d}}$ <- Denotes the number of times that a particular station has been active throughout the day in terms of interactions (interactions greater than zero).

3.2.3 Interval aggregation model (IAM)

In contrast to previous models, the third model (IAM) estimates interactions rather than the number of trips taken on a daily basis. In this model, the interactions between all the stations that are part of the BSS are summed up for each interval duration. In addition to this, the model includes a normalization variable called active stations, as well as its square to compensate for the losses caused by collisions on days of higher activity.

The IAM equation is given by:

$$\mathbf{i_{\Delta dt} = 0 + \beta_1 X_{\Delta dt} + \beta_2 A_{x\Delta dt} + \beta_3 A^2_{x\Delta dt} + \varepsilon}$$

3.2.4 Station aggregation model (SAM)

In this model, all that is considered is the frequency of stations throughout the day instead of the number of active stations in the system itself.
The SAM equation is given by:

$$\mathbf{i_{\Delta sd} = 0 + \beta_1 X_{\Delta sd} + \beta_2 A_{x\Delta sd} + \beta_3 A^2_{x\Delta sd} + \varepsilon}$$

`

3.3 Assumptions for Linear Regression Models

The linear regression technique is one of several kinds of statistical analysis that can be used to find out how two variables, x and y, are related to one another and how their relationship is determined. In order to begin the process of linear regression, the following four assumptions need to be satisfied in order to proceed.

3.3.1 Linearity

It is important to make sure that the relationship between the independent variable (x) and the dependent variable (y) is linear [28].

- Linearity can be tested by visualising the independent and dependent variable using scatterplots.
- Figure 3.2 (Linear) scatterplots shows that if the two variables are linearly related, the points will fall along a straight line whereas in Figure 3.2 (No Linear) scatterplot, the two variables are related but do not have a linear relationship.

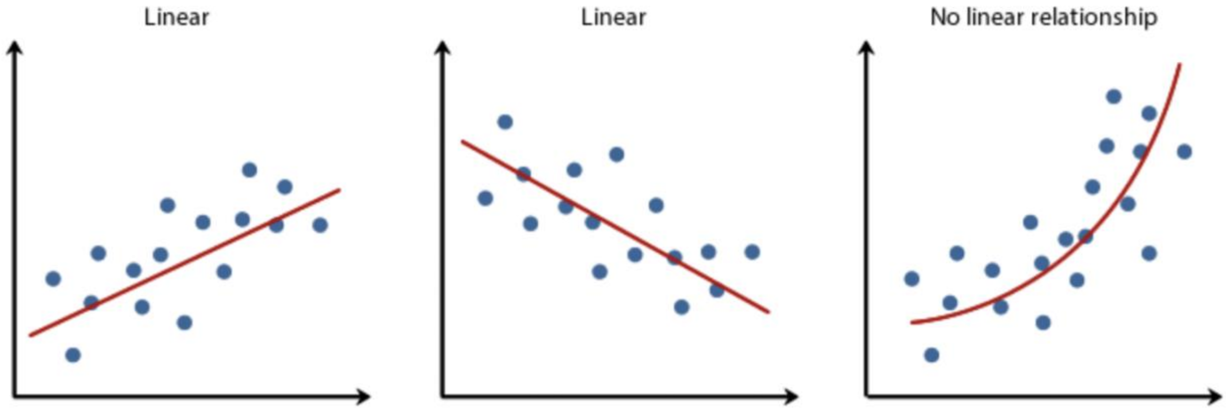


Figure 3.2 Scatterplot plot for Linearity [47]

`

3.3.2 Autocorrelation

It is important that the residuals (error terms) have no autocorrelation between them. A residual value that depends on a previous residual value usually indicates autocorrelation, which is usually seen in time-series data.

- Autocorrelation plot, the Ljungbox test, or the Durbin Watson test are three known ways to investigate whether or not residual values exhibit autocorrelation.

Autocorrelation plot : The presence of any spikes outside the range of confidence interval in blue color, indicates that there is an autocorrelation, otherwise no autocorrelation exists.

- As can be seen in Figure 3.3, there is no evidence of autocorrelation, as the spikes fall within the range of confidence intervals.

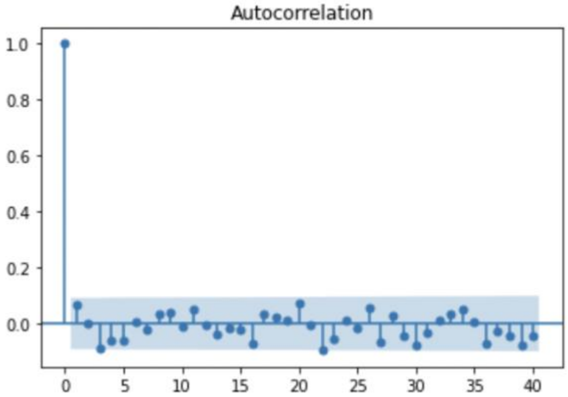


Figure 3.3 Autocorrelation plot

Ljungbox test: The Ljung Box test is a method used to determine whether there is presence or absence of serial autocorrelation up to a certain lag k. [29]

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

Figure 3.4 Ljungbox Test Formula [48]

Null Hypothesis(H_0) : Autocorrelation is absent.

Alternative Hypothesis(H_A): Autocorrelation is present.

- When the p-value is greater than 0.05, there is strong evidence for the null hypothesis.

`

Durbin Watson Test: Using the Durbin Watson statistic, we can determine whether the residuals obtained from regression models have autocorrelation. Its value ranges from 0 to 4 [57].

DW value = 2.0 indicates no autocorrelation
DW value > 2.0 indicates negative autocorrelation
DW value < 2.0 indicates positive autocorrelation

3.3.3 Homoscedasticity

As it stands, homoscedasticity refers to the variances of the residuals being the same or almost the same along the regression line. Plotting the error terms against the predicted terms allows us to verify whether there exists any pattern (Cone-Shaped) in the error terms and also Goldfeld Quandt Test can also be performed to test homoscedasticity.

Goldfeld Quandt Test: It is a test commonly used to check for homoscedasticity when doing regression analysis, and it is known as the Goldfeld Quandt Test.
Null Hypothesis(H_0): Error terms are homoscedastic.
Alternative Hypothesis(H_A): Error terms are heteroscedastic.

- When the p-value is greater than 0.05, there is strong evidence for the null hypothesis.

3.3.4 Normality

Lastly, the fourth assumption of linear regression is that the residuals should be normally distributed. If the residuals are not randomly distributed, the model cannot explain the relationship in the data since their randomness is lost. It is relatively easy to test the normality of the residuals obtained from the model using either a histogram or a QQ plot in order to ensure that the residuals are indeed normal.

`

- In the following histogram, Figure 3.5 (Normal errors), the distribution of the residuals forms a bell-shaped curve, suggesting that the error terms are normally distributed. In contrast, Figure 3.5 (Non-normal errors), the residuals are not in the bell-shaped curve, meaning they are not normally distributed.

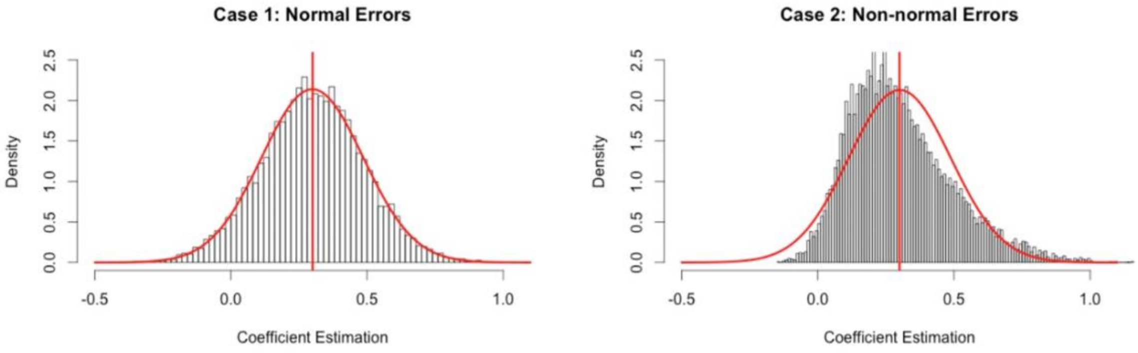


Figure 3.5 Histogram for Normal and Non-Normal Distribution [49]

- In the following QQplot Figure 3.6 (Case 1), the residual points fall along a straight line, supporting the assumption that residual terms are normally distributed. In contrast, in Figure 3.6 (Case 2), residual points are not following the straight line, meaning that residuals do not follow normality.

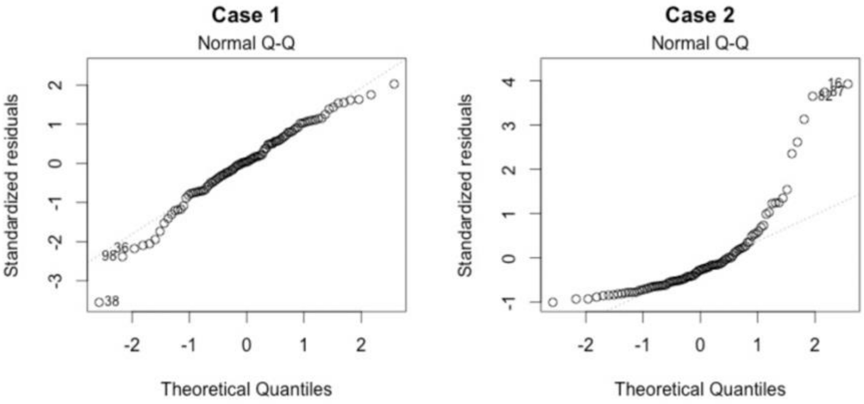


Figure 3.6 QQplot for Normal and Non-Normal Distribution [50]

`

- It is important to note that if any of these assumptions are violated, then the results of our linear regression could be unreliable or even misleading.

3.4 Metrics

A metric is a measure of quantitative assessment that is commonly used for comparing and tracking the performance or production of a process or product. The development, establishment, and definition of program progress using metrics allows for greater accountability, and adds a greater level of value to the program.

3.4.1 Regression Metrics

Mean Absolute Error

The Mean Absolute Error, or MAE, is a popular method of calculating error scores. This method uses units that are identical to those used for the target value that we are trying to predict. In the MAE, error types are not weighted differently, and the scores increase linearly as the errors increase [30].

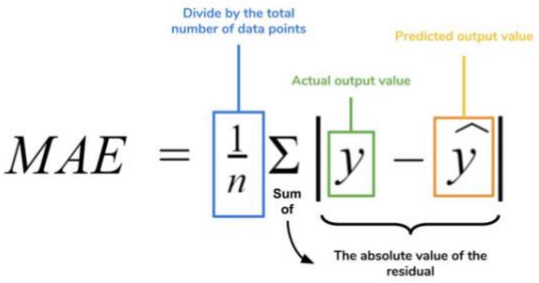

Divide by the total number of data points
Predicted output value
Actual output value
$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$
Sum of
The absolute value of the residual

Figure 3.7 Mean Absolute Error Formula [51]

Root Mean Square Error

A root mean square error or RMSE is one of the most commonly used methods for estimating prediction quality. In order to illustrate how much the predicted value differs from the actual value, Euclidean distance is used and also RMSE is robust to outliers [31].

`

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n} |y_i - x_i|^2}$$

x_i = actual value n = sample size
y_i = predictions

Figure 3.8 Root Mean Square Error Formula [52]

Normalised Root Mean Square Error

An extension of RMSE is NRMSE, which is calculated by normalizing the RMSE. Normalizing the RMSE can be accomplished in two ways: using the mean or the range of the actual values (the difference between the minimum and maximum values [31].

$$NRMSE = \frac{RMSE}{mean(y)} \quad OR \quad NRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

Figure 3.9 Normalised Root Mean Square Error Formula [53]

Root Mean Square Logarithmic Error

The Root Mean Square Logarithmic Error or RMSLE measures the difference between the actual and predicted values. Using logarithms, huge differences between actual and predicted values are not penalized. In comparison studies, RMSLE is usually used as the final metric to measure a regression machine learning models' efficiency. With RMSLE, the small and the large errors are handled equally[33].

$$\text{RMSLE} = \sqrt{(log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

Figure 3.10 Root Mean Square Logarithmic Error Formula [54]

`

R Squared

A regression model's R-Squared (R^2) indicates how much of the variation in the dependent variable is explained by the independent variables. Ideally, R-squared should range between 0 and 1, and if it is closer to 0, the model is not capturing trends well. In general, the closer R-squared is to 1, the better our model performs.

3.4.2 Classification Metrics

Confusion Matrix

One of the most straightforward metrics for determining the correctness and accuracy of the model is the confusion matrix. Classification problems with two or more output types can be solved with it [34]. It is important to note that the diagonal elements in the matrix represents the total correct values predicted per class. The graph shows the following results to help us visualize our model's performances

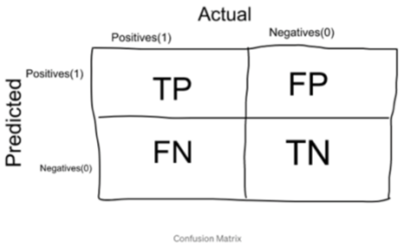


Figure 3.11 Confusion Matrix [34]

Confusion matrix terms

True Positives (TP): A true positive is when the actual and predicted classes of a data point are both 1 (True).

True Negatives (TN): A true negative is when the actual and predicted classes of a data point are both 0 (False).

False Positives (FP): A false positive is when the actual class of the data point is 0 (False) and the predicted class is 1 (True).

`

False Negatives (FN): A false negative is when the actual class of the data point is 1 (True) and the predicted class is 0 (False).

Accuracy

In classification problems, accuracy refers to the percentage of correct predictions made by the model out of all possible predictions. When the target variable classes are nearly balanced, accuracy is a good measure, but it's not recommended when one class dominates the data.

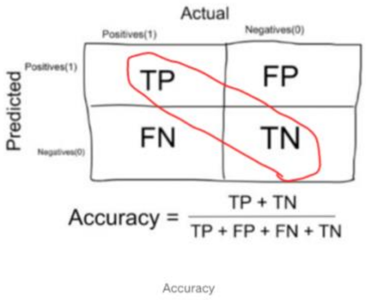


Figure 3.12 Accuracy Formula [34]

Precision

The precision is calculated as the ratio between correctly classified Positive samples and all Positive samples (correctly or incorrectly). Positive classification accuracy is measured by precision [35].

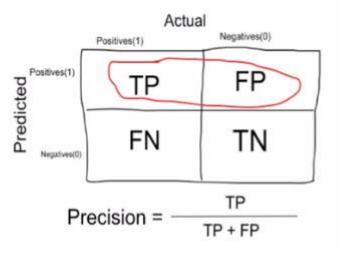


Figure 3.13 Precision Formula [34]

`

Recall or Sensitivity

The recall is the percentage of Positive samples that were correctly classified to the total number of Positive samples. This parameter measures the model's ability to detect positive samples. Higher recall means more positive samples [35].

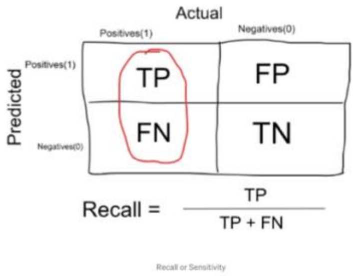


Figure 3.14 Recall or Sensitivity Formula [34]

F1-Score

F1 is a weighted harmonic mean of recall and precision, where 1.0 is the best and 0.0 is the lowest [36].

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Figure 3.15 F1-Score Formula [36]

`

Chapter Four : Data Pre-Processing and Feature Engineering

4.1 Data Collection

4.1.1 Dublin Bikes

A literature review shows the Dublin bike analysis began in 2010 with a paper by Peter Mooney[20]. Since then, more studies and analyses have been conducted to enhance Dublin Bikes BSS. Throughout this research, data was collected regularly, combined as necessary, and stored in a CSV. From 2011 -2019, data was covered in this dataset. According to previous studies, the dataset had the following structure shown in Table 4.1.

Column	Description
tfl_id	bike station id
bikes	Available bikes at that time
spaces	Available docks at that time
total_docks	Total number of stands in station
timestamp	Timestamp of recorded entry

Table 4.1 Historical Dublin Bikes dataset description

For this project, we obtained the latest data from Smart Dublin in a CSV file and combined it with previously gathered data for further analysis and modelling purposes. Now, we have the data from 2011 – 2021. It is important to note that the most recent data was collected every five minutes, whereas the historical data was collected every ten minutes. The following structure in Table 4.2, belongs to the latest dataset which is published by Smart Dublin.

`

Column	Description
STATION ID	Bike Station Id
TIME	Timestamp of recorded entry
LAST UPDATED	Timestamp of last updated information
NAME	Name of the station
BIKE STANDS	Total number of bike stands in station
AVAILABLE BIKE STANDS	Available bike stands at that time
AVAILABLE BIKES	Available bikes at that time
STATUS	Station Status (Open/Close)
ADDRESS	Address of the station
LATITUDE	Latitude of the station
LONGITUDE	Longitude of the station

Table 4.2 Latest Dublin Bikes Dataset Description

4.1.2 Weather information

To analyse the behaviour of bike activities in various weather conditions, Met Eireann's Dublin Airport Hourly Data was collected for Dublin [37]. The only columns that were collected from this dataset were those for air temperature and wind speed. Using windspeed, we were able to classify the wind according to the Beaufort Scale [38].

4.2 Data Preparation and Feature Engineering

4.2.1 Data Cleaning

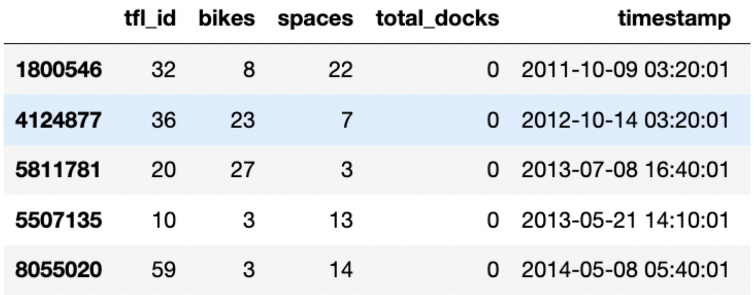
As a part of the data cleaning process, we assessed the quality of the historical datasets collected from previous studies and the quality of the latest data collected from the website during the data cleaning process. In this assessment, we discovered a few issues that needed to be addressed, and the appropriate measures were taken to overcome them.

`

The following is a list of the issues that we encountered:

Noisy data

In the real world, noisy data is meaningless, meaningless information that cannot be understood by machines. This kind of data is caused by faulty data collection, incorrect data entry, etc. In the Dublin bike data, we found a few samples whose total_docks values were set to zero as shown Figure 4.1, which is logically not possible where the bikes and spaces values are not zero. Therefore, a simple imputation method was performed on incorrect data by summing the number of bikes and the number of docks available for each row. The simple formula is given by [total_docks = bikes + space].



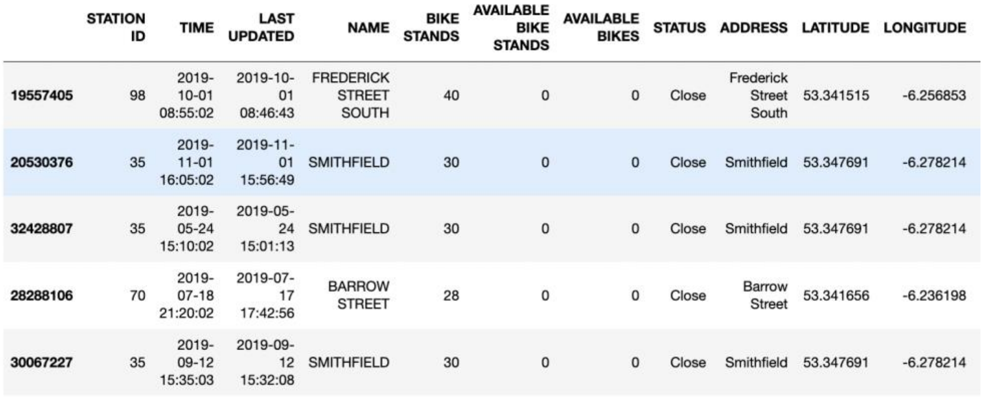
	tfl_id	bikes	spaces	total_docks	timestamp
1800546	32	8	22	0	2011-10-09 03:20:01
4124877	36	23	7	0	2012-10-14 03:20:01
5811781	20	27	3	0	2013-07-08 16:40:01
5507135	10	3	13	0	2013-05-21 14:10:01
8055020	59	3	14	0	2014-05-08 05:40:01

Figure 4.1 Sample dataset for Incorrect total_docks values

Removing bad samples

As shown in Figure 4.2, we also found 10,6080 records whose station status was set to "Close" meaning no activity took place at that station during that period. Our analysis and predictions would be inaccurate if we considered these samples. We eliminated all such rows in order to keep our data clean.

`

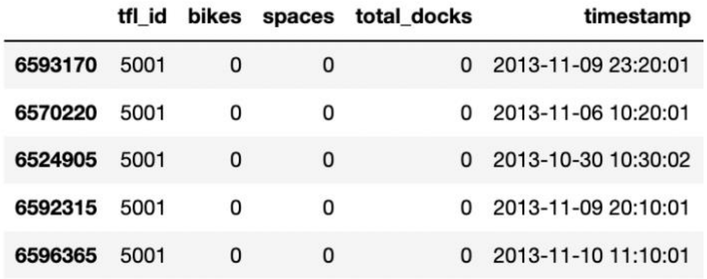


	STATION ID	TIME	LAST UPDATED	NAME	BIKE STANDS	AVAILABLE BIKE STANDS	AVAILABLE BIKES	STATUS	ADDRESS	LATITUDE	LONGITUDE
19557405	98	2019-10-01 08:55:02	2019-10-01 08:46:43	FREDERICK STREET SOUTH	40	0	0	Close	Frederick Street South	53.341515	-6.256853
20530376	35	2019-11-01 16:05:02	2019-11-01 15:56:49	SMITHFIELD	30	0	0	Close	Smithfield	53.347691	-6.278214
32428807	35	2019-05-24 15:10:02	2019-05-24 15:01:13	SMITHFIELD	30	0	0	Close	Smithfield	53.347691	-6.278214
28288106	70	2019-07-18 21:20:02	2019-07-17 17:42:56	BARROW STREET	28	0	0	Close	Barrow Street	53.341656	-6.236198
30067227	35	2019-09-12 15:35:03	2019-09-12 15:32:08	SMITHFIELD	30	0	0	Close	Smithfield	53.347691	-6.278214

Figure 4.2 Sample dataset where stations are closed

Inconsistent Values

The Dublin bikes dataset contains a few uncommon station IDs, such as 507 and 5001, as shown in Figure 4.3. Although the exact cause of such entries is unknown, we infer that it may be related to a technical glitch or new functionality being tested in production server. As a result, we removed those station IDs from our dataset.



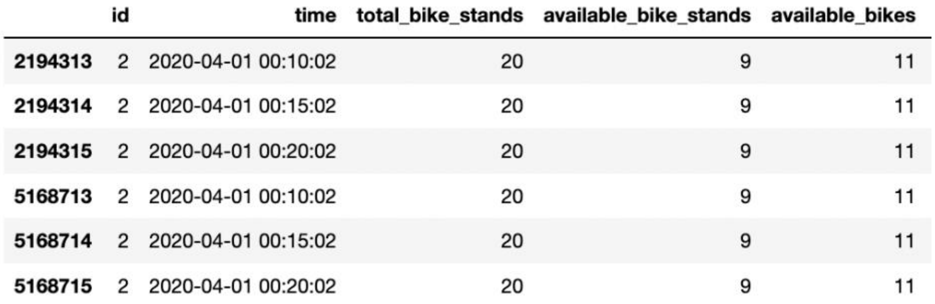
	tfl_id	bikes	spaces	total_docks	timestamp
6593170	5001	0	0	0	2013-11-09 23:20:01
6570220	5001	0	0	0	2013-11-06 10:20:01
6524905	5001	0	0	0	2013-10-30 10:30:02
6592315	5001	0	0	0	2013-11-09 20:10:01
6596365	5001	0	0	0	2013-11-10 11:10:01

Figure 4.3 Sample Dataset with station ID 5001

Duplicate values

We found that there were duplicate records in the dataset, which must be eliminated in order to deal with the issue. As an example, Figure 4.4 clearly shows the fact that each row has been repeated twice. In a similar manner 2,70,082 records were also found to be duplicated.

`



	id	time	total_bike_stands	available_bike_stands	available_bikes
2194313	2	2020-04-01 00:10:02	20	9	11
2194314	2	2020-04-01 00:15:02	20	9	11
2194315	2	2020-04-01 00:20:02	20	9	11
5168713	2	2020-04-01 00:10:02	20	9	11
5168714	2	2020-04-01 00:15:02	20	9	11
5168715	2	2020-04-01 00:20:02	20	9	11

Figure 4.4 Sample Dataset where Duplicate entries are present

4.2.2 Data Transformation

In data cleaning, we have already begun making changes to our data, but data transformation begins the process of making the data suitable for analysis and downstream processing. As part of the pre-processing of the data, the following steps were taken:

1. In order to analyse the Dublin bikes on a broad range, we have merged the old dataset with the latest dataset after it has been cleaned.
2. By using pandas and NumPy libraries, we extracted year, month, day, hour, minute, and weekday for each entry from the timestamp column, and also created a column called season based on the month's column.
3. Besides Dublin Bike datasets, we also gathered hourly weather information, Dublin Bike spatial information, and public holidays. These datasets were all combined into Dublin Bikes. Further, we classified wind speed according to the Beaufort Scale [37][38].
4. Different transformations were used to create crucial columns for our analysis, such as bike availability percentage, bike arrival, bike departure, and interactions. Additionally, we renamed the columns to make them easier to understand and more convenient to use.

`

The following structure in Table 4.3 are descriptions for newly created columns:

Columns	**Description**
weekday	Categorical column that denotes day of the week.
season	Categorical column that provides season('Winter', 'Spring', 'Summer', 'Autumn')
wdsp	Mean Wind Speed (knot).
wdsp_classified	Categorical column to indicate the level of windspeed (Beaufort Scale).
holiday	Categorical column to indicate whether the day is a public holiday or casual day.
availability_percentage	Numerical Column that denotes percentage of bikes available at the station at that interval.
bike_arr_dep	A station's number of bikes has changed since the previous row.
bike_arr	Number of bikes returned to the station compared to previous record.
bike_dep	Number of bikes picked from the station compared to previous record.

Table 4.3 Data Description for Transformed dataset

`

Chapter Five : Exploratory Data Analysis

An exploratory data analysis is performed on a pre-processed dataset to identify trends or patterns in how bikes are used in the city. Also, the bike rental activity in Dublin city was compared with several variables that could have a bearing on the use of bicycles in the city.

5.1 Data Supply Analysis

Our interest was piqued by the statistics about data collection [39]. A time-series graph Figure 5.1, revealed that data supply for Dublin Bikes increased over time. There are several factors that may explain this, including the growth of Dublin bike stations, as shown in Figure 1.3, increased activity, and more efficient data collection. A million data points were collected by JCDECaux in 2019. The data collection helps us uncover more trends across all stations and provide more insights into bike activity.

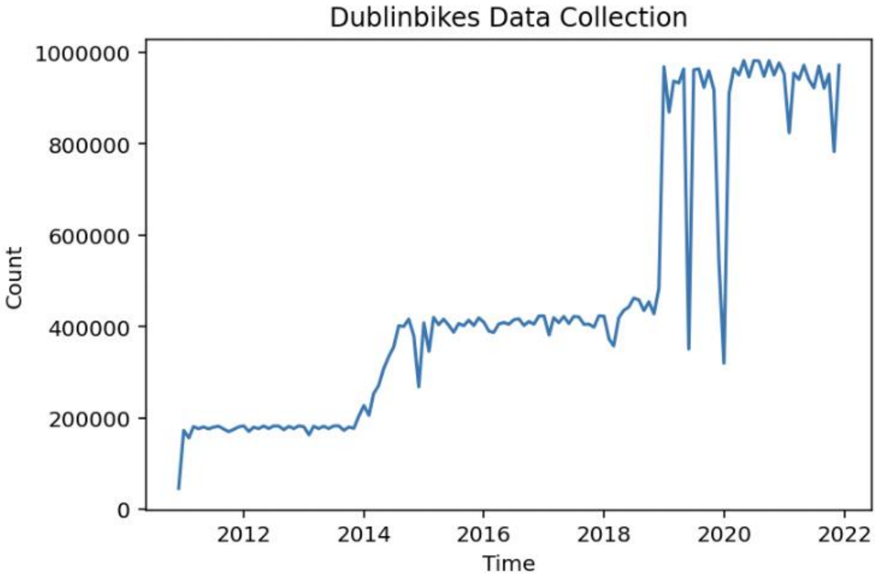
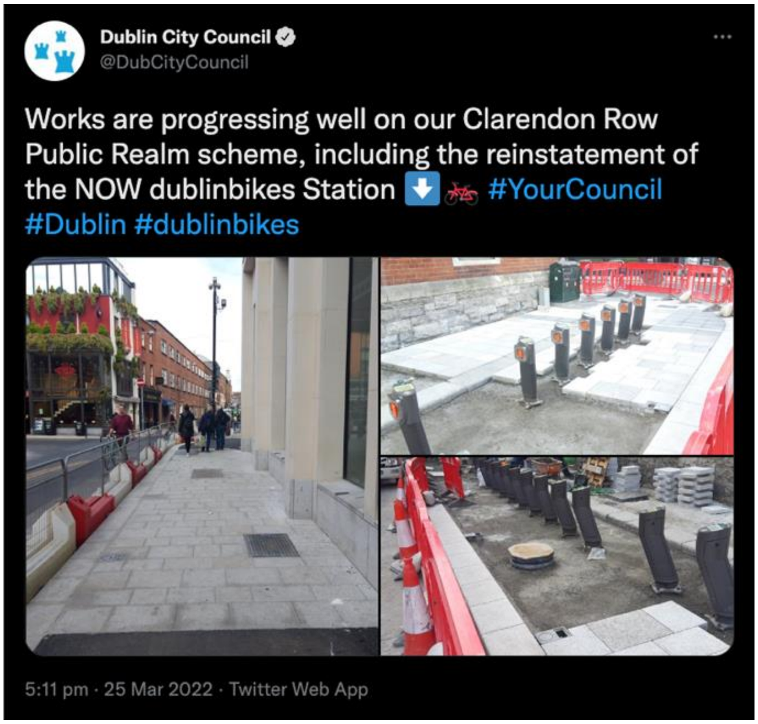


Figure 5.1 Data Collection Trend for Dublin Bikes over the years

According to our research, the "Clarendon Row" station, which has been recording data since September 2009, stopped recording data in February 2019. Dublin City Council (DCC) initiated the Clarendon Row Improvement Scheme that led to this station being demolished and relocated from outside Butler's Chocolates on Chatham Street [40]. As shown in (Figure 5.2), we have also found a recent tweet from DCC regarding the reinstallation of the station.



Dublin City Council
@DubCityCouncil

Works are progressing well on our Clarendon Row Public Realm scheme, including the reinstatement of the NOW dublinbikes Station ⬇️ 🚲 #YourCouncil #Dublin #dublinbikes

5:11 pm · 25 Mar 2022 · Twitter Web App

Figure 5.2 Tweet from Dublin City Council on Clarendon Row [56]

5.2 Bike Activity Analysis

BSS usage is primarily determined by the amount of bicycle activity at a station, which is the most reliable indicator of BSS usage. As a result, it is essential to analyse and study its influencing factors in order to determine the best way of planning and managing Dublin bikes. Moreover, we wanted to explore and visualise bike activity fluctuations in different stations, including some of Dublin's most well-known and central stations, as well as suburban stations, as part of our study.

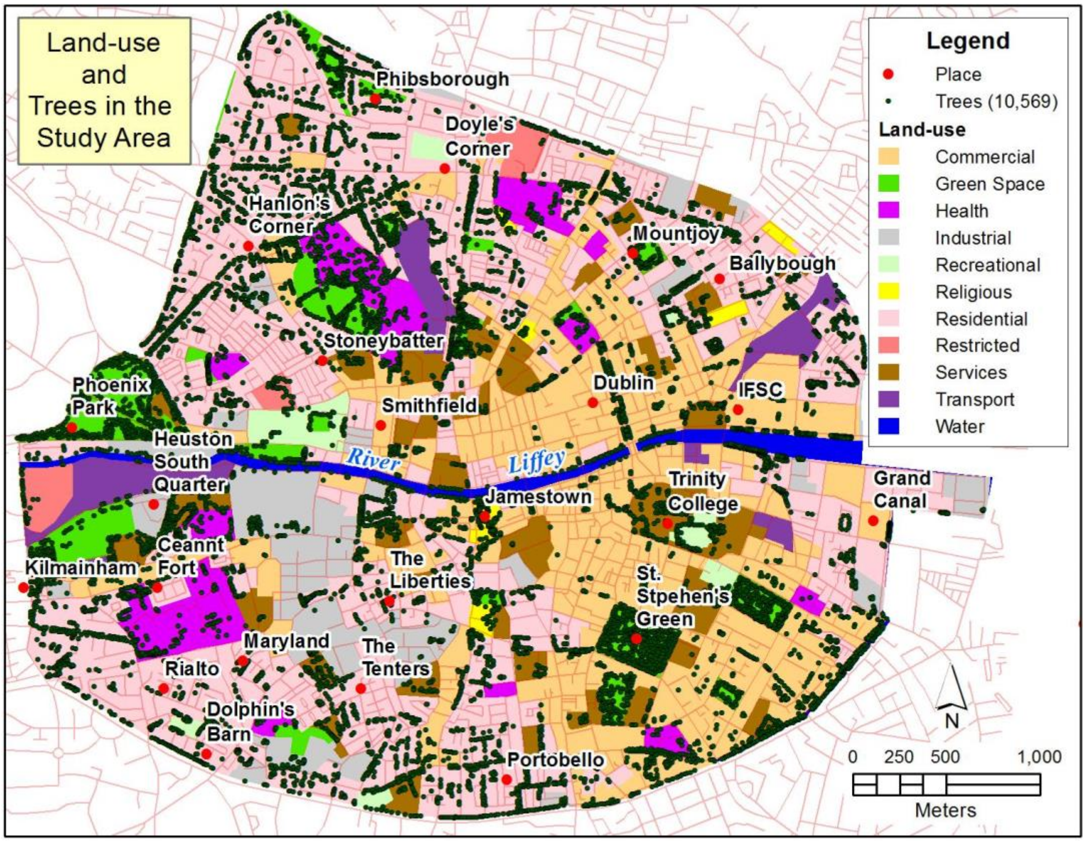
We introduced a new column called "interactions," which is a difference between the bike availability values for each entry in relation to the previous record, whose definition is discussed in Chapter 3. To remove negative values, this value is taken as an absolute value. In this case,

`

interactions represent the amount of activity occurring at a station which is measured by the change or difference in values.

Our analysis of bike activity was limited to the latest six years' data due to a lack of computing resources, although data were available from 2011 to 2021. Based on the data from these six years, we believed that it would be sufficient to analyse and draw conclusions.

In order to explain the reasons for the biking activities occurring at various stations, we utilized the following Figure 5.3, which depicts the neighbourhoods of Dublin city. It includes geographical information about the area whether it is surrounded by residential, commercial, or industrial properties, etc. Using this Figure 5.3, we are able to do a geospatial analysis on the results obtained from each plots.



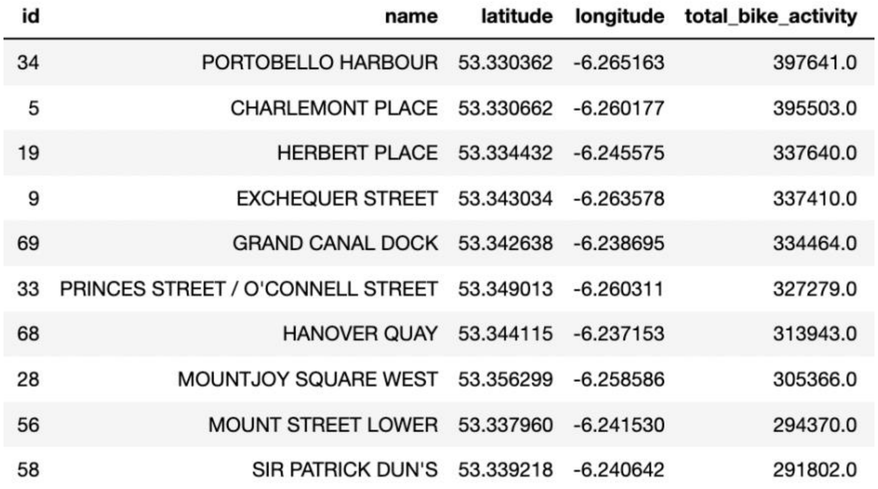
.

Figure 5.3 Neighbourhoods across Dublin City [41]

`

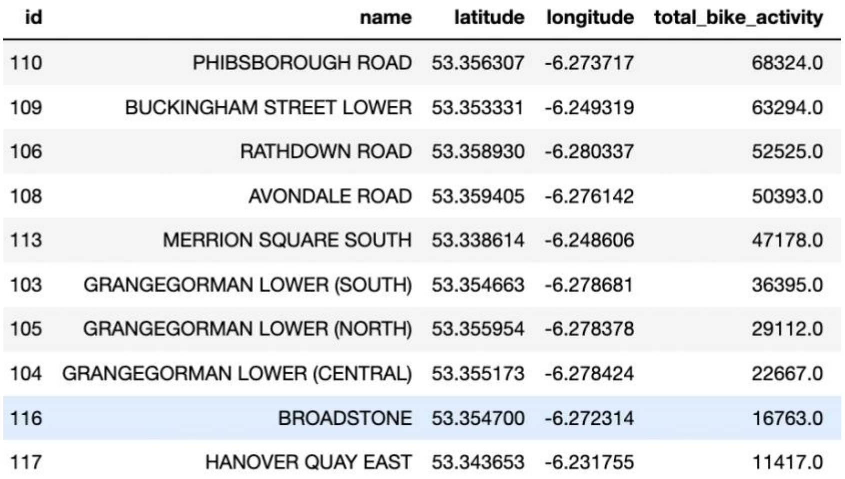
5.2.1 Busiest and Quietest Stations

A calculation was carried out to determine the top 10 bike stations in Dublin with the highest bike rental activity and the bottom 10 stations with the least amount of activity among 117 stations, based on the total activity of the bike stations. Figure 5.4 and Figure 5.5 shows the stations with the greatest and least amount of activity. The term 'total_bike_activity' here refers to the total number of pick-ups and drops.



id	name	latitude	longitude	total_bike_activity
34	PORTOBELLO HARBOUR	53.330362	-6.265163	397641.0
5	CHARLEMONT PLACE	53.330662	-6.260177	395503.0
19	HERBERT PLACE	53.334432	-6.245575	337640.0
9	EXCHEQUER STREET	53.343034	-6.263578	337410.0
69	GRAND CANAL DOCK	53.342638	-6.238695	334464.0
33	PRINCES STREET / O'CONNELL STREET	53.349013	-6.260311	327279.0
68	HANOVER QUAY	53.344115	-6.237153	313943.0
28	MOUNTJOY SQUARE WEST	53.356299	-6.258586	305366.0
56	MOUNT STREET LOWER	53.337960	-6.241530	294370.0
58	SIR PATRICK DUN'S	53.339218	-6.240642	291802.0

Figure 5.4 Top 10 Busiest Stations



id	name	latitude	longitude	total_bike_activity
110	PHIBSBOROUGH ROAD	53.356307	-6.273717	68324.0
109	BUCKINGHAM STREET LOWER	53.353331	-6.249319	63294.0
106	RATHDOWN ROAD	53.358930	-6.280337	52525.0
108	AVONDALE ROAD	53.359405	-6.276142	50393.0
113	MERRION SQUARE SOUTH	53.338614	-6.248606	47178.0
103	GRANGEGORMAN LOWER (SOUTH)	53.354663	-6.278681	36395.0
105	GRANGEGORMAN LOWER (NORTH)	53.355954	-6.278378	29112.0
104	GRANGEGORMAN LOWER (CENTRAL)	53.355173	-6.278424	22667.0
116	BROADSTONE	53.354700	-6.272314	16763.0
117	HANOVER QUAY EAST	53.343653	-6.231755	11417.0

Figure 5.5 Bottom 10 Quietest Stations

`

According to this analysis, Portobello Harbour and Charlemont Street are the busiest stations in southern Dublin. In contrast, Broadstone, Hanover Quay East, and Grangegorman are the quietest stations to be found in the city's north.

An interactive map Figure 5.6, is also created to perform geospatial analysis on these results where “red circles” indicates busiest stations and “blue circles” indicates quietest stations The “red circles” on the below map are busier stations that are spread across the south part of the city, with the exception of Mountjoy Square West and Princes Street / O’Connell Street stations. Also, most of the stations marked in “blue circles” in the map are located in northern areas, except Merrion Square South. The station's activities is closely linked with the use of the surrounding land. There are commercial and residential areas around the most active stations, whereas the lowest active stations are surrounded by parks and hospitals.

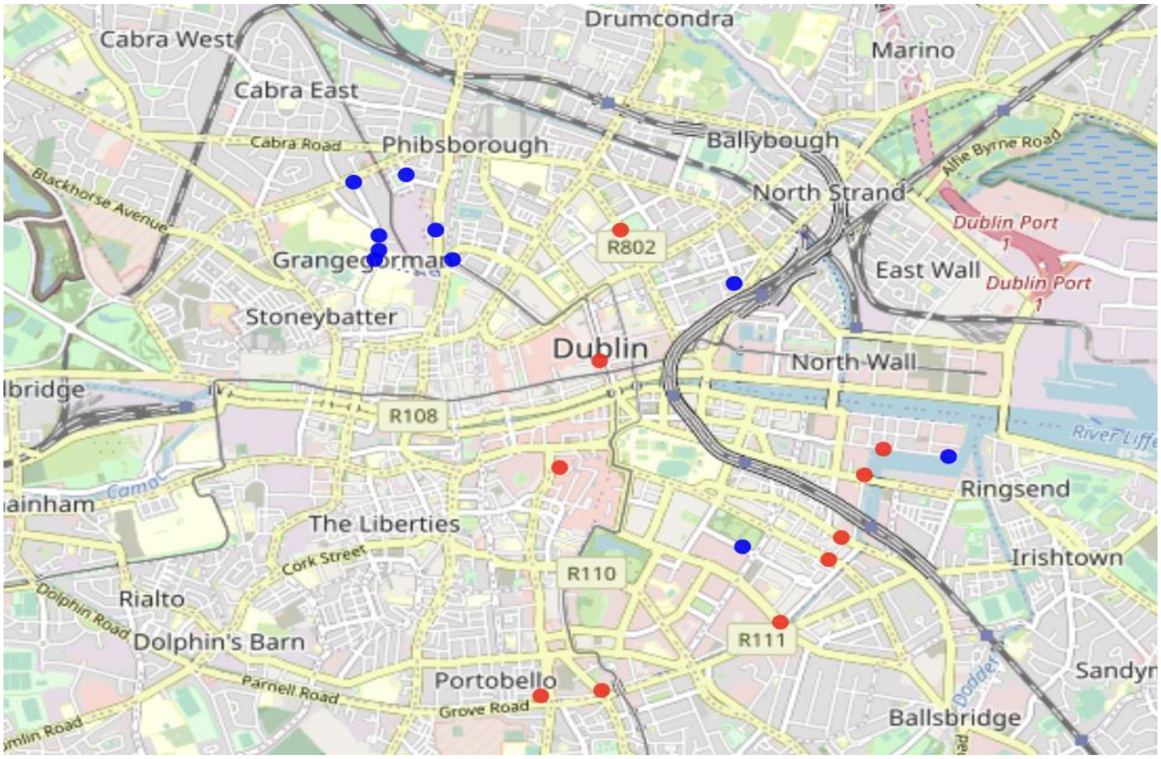


Figure 5.6 Geo-Spatial Map for Quietest and Busiest bike stations

`

5.2.2 Weekday and Hourly trends

As seen in the Figure 5.7, Tuesday, Wednesday, Thursday, and Friday seem to have the highest bike activity, especially between 6-10 am and 4-7 pm. On the other hand, Monday is a quieter day than the rest of the weekdays. According to this, most working professionals use bikes to commute between their homes and their workplaces. Weekends, on the other hand, are quieter, with a steady stream of cyclists at noon who use bikes to travel to shopping, parks, and entertainment on the weekends.

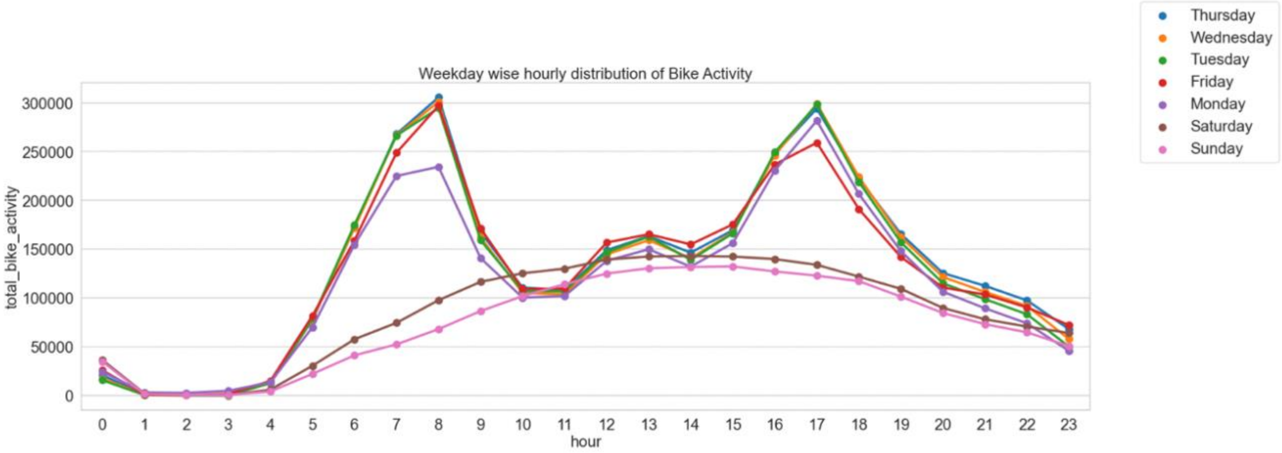


Figure 5.7 Bike Activity trend (Weekdays)

5.2.3 Holidays Trends

The bike activity was plotted with respect to several Irish public holidays in order to explore the patterns of biking on these holidays. Figure 5.8 illustrates the interesting results that were produced by this investigation. On Christmas Day, all shops and bars are shut down, and there are no public transit options. On New Year's Day, pubs will open later and close earlier with minimal public transportation. Therefore, these holidays have less bicycle riders than other holidays. On the other hand, The October and June bank holidays have the most cycling activity. Even if banks are closed, stores and bars in Dublin are often open, leading to higher activity [42].

`

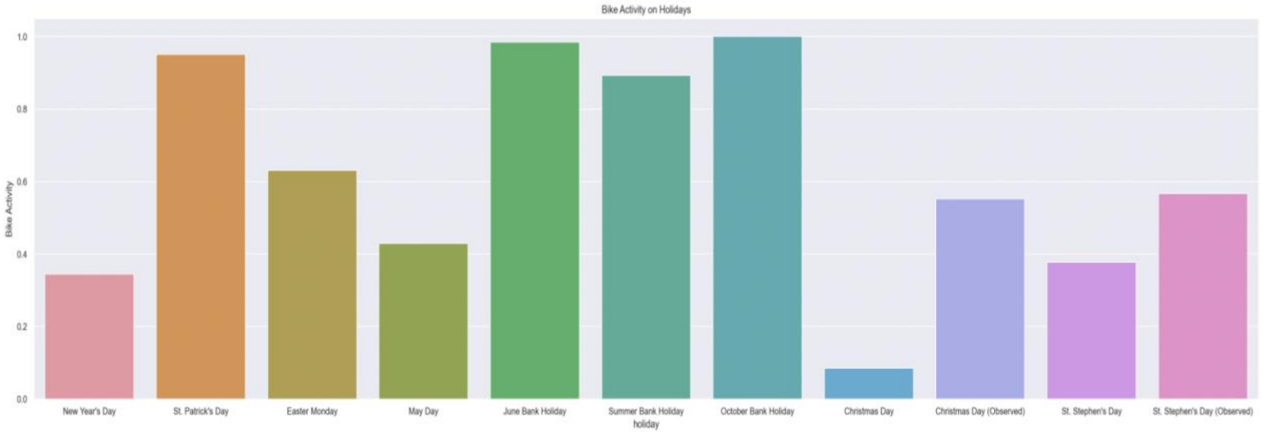


Figure 5.8 Bike Activity trend in Public Holidays

5.2.4 Seasonal Trends

According to Figure 5.9, bicycle activity was constant throughout each season. This demonstrates clearly that Irish bike riders don't consider the weather while they ride.

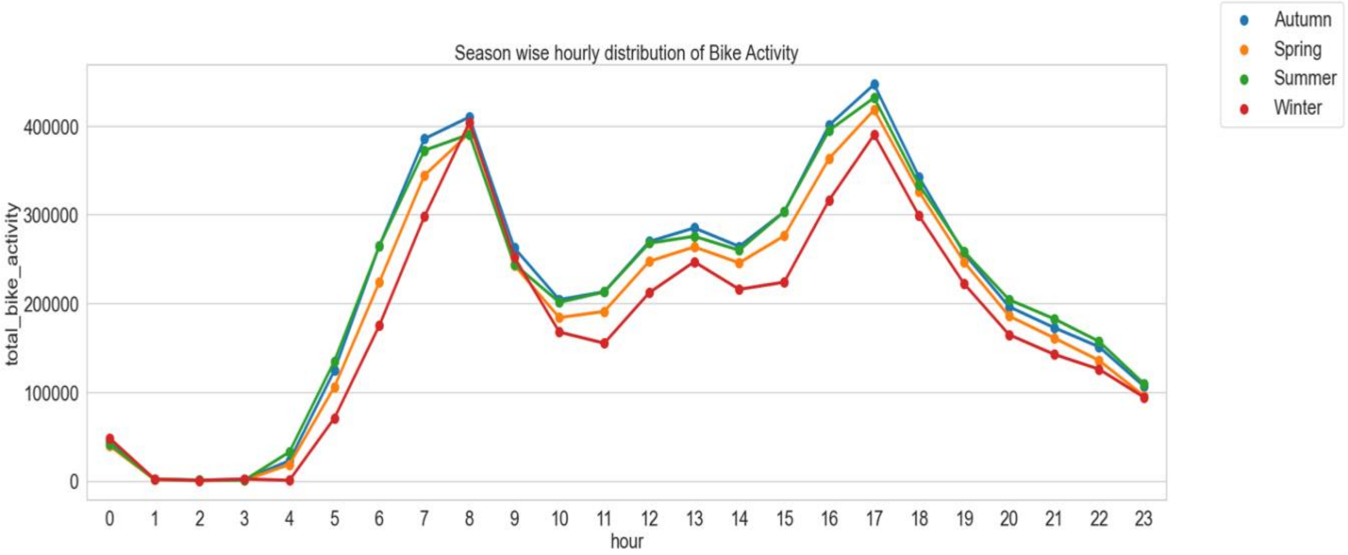


Figure 5.9 Bike Activity trend in different Seasons

`

5.2.5 Effect of Covid-19

We were also curious to know how the Covid pandemic affected bike usage and whether easing restrictions had led to a noticeable recovery in bike usage since then. Therefore, we visualized bike activity by plotting the bike activity for six years monthly, covering pre-covid and post-covid.

Figure 5.10, shows a continuous decline in bike activity following Ireland's shutdown in March 2020. Bikes usage has reduced to one-third from previous years which is evident that COVID-19 has impacted Dublin bikes usage significantly. Several transport-related reforms were introduced soon after the pandemic, altering travel behaviour and public transportation safety perceptions. Also, the transition to remote working is a significant factor in the decline in bike usage. As the restrictions are being relaxed and most businesses have adopted a hybrid working model, bicycle activity may rise in the future.

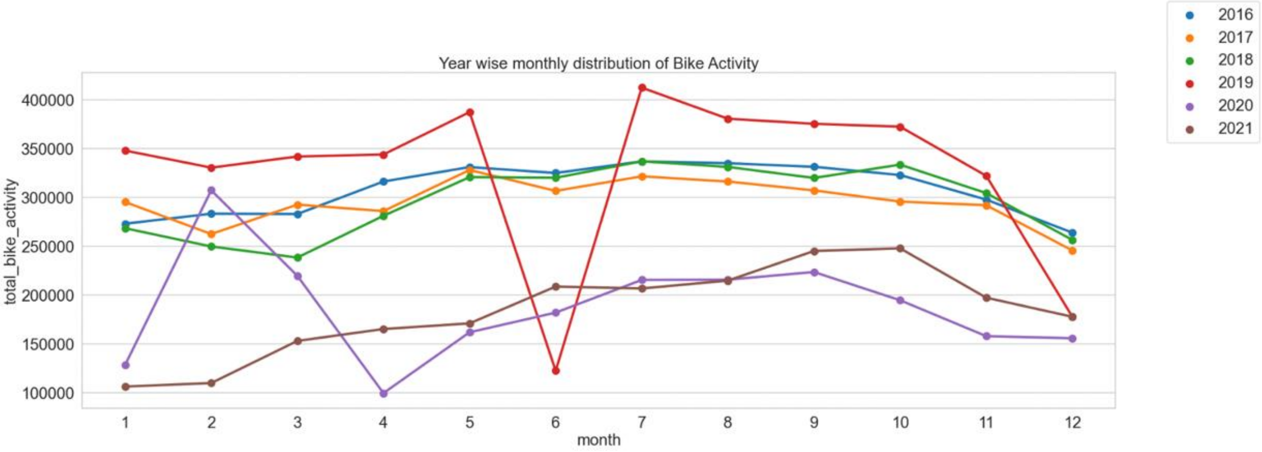


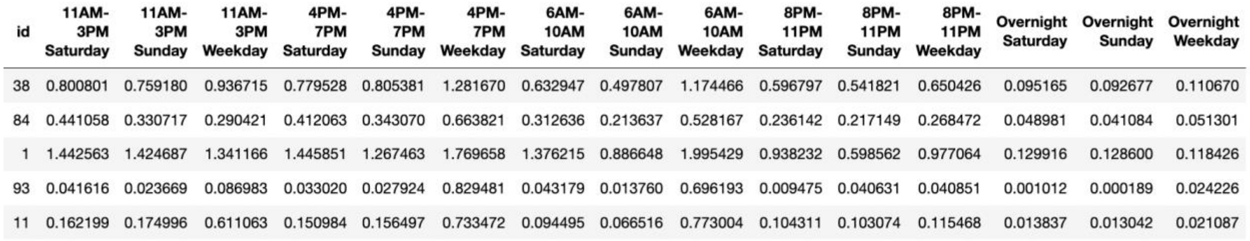
Figure 5.10 Bike activity trend during Covid-19

Furthermore, we found that bike activity dropped suddenly during June 2019, which was not caused by COVID-19. Despite our efforts, we could not determine the cause of the decline. As a result, we assumed that data collection was halted owing to the server that collects or stores the data being under maintenance or malfunctioning.

`

5.2.6 Clustering

We wanted to determine whether any patterns or behaviors could be generalized over all stations and, if so, where they might occur in the city. As bikes are utilized for commuting in the city, we believe there should be a spatial relationship between the stations. To achive this, we used K-means clustering algorithm which is one of the best unsupervised machine learning algorithm. The K-means algorithm works on the assumption that data can be categorized into 'k' different clusters or categories before computation. To acquire better clustering results using the K-Means method, we adopted Rachel's method [58], to transform the data and implement. Therefore, to implement this transformation, it was necessary to group the mean bike activity for each station by "day type" (weekday; Saturday; Sunday) and "time type" (home to work (6 am-10 am); working hours and lunch (11 am-3 pm); arriving home from work (4 pm-7 pm); evening activities (8 pm-11 pm; overnight). The algorithm uses the transformed data that is shown Figure 5.11, as input to identify stations with similar behavior.



id	11AM-3PM Saturday	11AM-3PM Sunday	11AM-3PM Weekday	4PM-7PM Saturday	4PM-7PM Sunday	4PM-7PM Weekday	6AM-10AM Saturday	6AM-10AM Sunday	6AM-10AM Weekday	8PM-11PM Saturday	8PM-11PM Sunday	8PM-11PM Weekday	Overnight Saturday	Overnight Sunday	Overnight Weekday
38	0.800801	0.759180	0.936715	0.779528	0.805381	1.281670	0.632947	0.497807	1.174466	0.596797	0.541821	0.650426	0.095165	0.092677	0.110670
84	0.441058	0.330717	0.290421	0.412063	0.343070	0.663821	0.312636	0.213637	0.528167	0.236142	0.217149	0.268472	0.048981	0.041084	0.051301
1	1.442563	1.424687	1.341166	1.445851	1.267463	1.769658	1.376215	0.886648	1.995429	0.938232	0.598562	0.977064	0.129916	0.128600	0.118426
93	0.041616	0.023669	0.086983	0.033020	0.027924	0.829481	0.043179	0.013760	0.696193	0.009475	0.040631	0.040851	0.001012	0.000189	0.024226
11	0.162199	0.174996	0.611063	0.150984	0.156497	0.733472	0.094495	0.066516	0.773004	0.104311	0.103074	0.115468	0.013837	0.013042	0.021087

Figure 5.11 K-Means Input Data Frame – Bike Activity

We found the optimal K-value to be 3 using the Elbow method [59]. Therefore, we decided to classify the stations into three clusters. For further analysis, the K-Means cluster values are merged with the bike activity dataset. Finally, we created interactive maps as shown Figure 5.12, showing Dublin bike stations as points with colours according to their clusters. This led to some exciting discoveries.

`

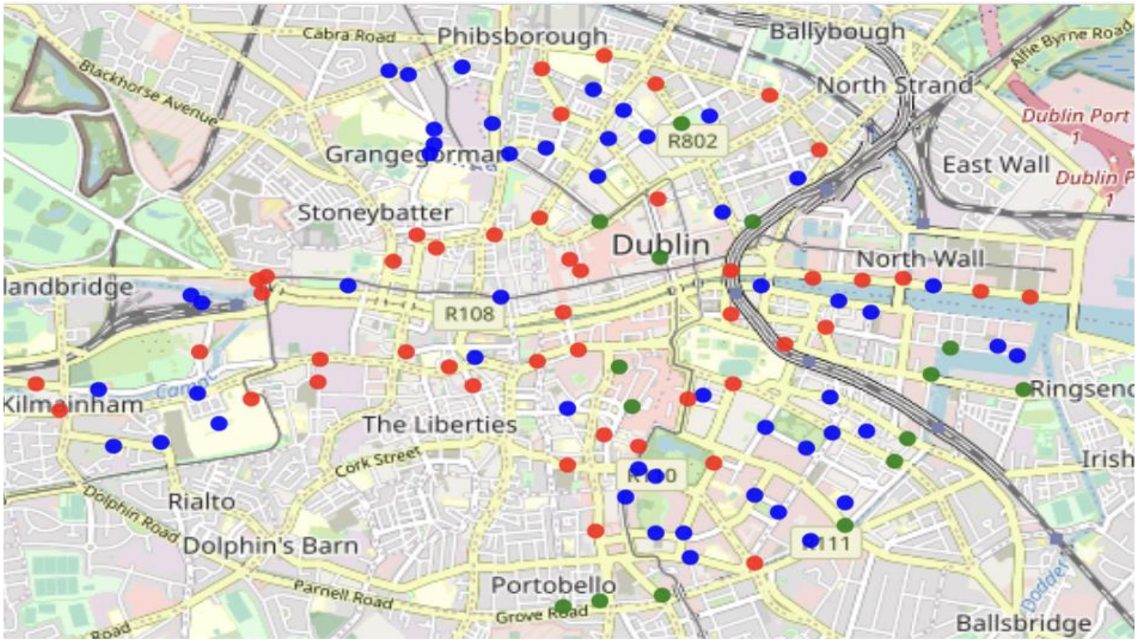


Figure 5.12 Geo-Spatial Map for all clustered stations

Labels - (green – high, red – medium, blue - low)

The "green circles" represents the stations with the most bike traffic. Stations like Portobello Road, Exchequer Street, and Grand Canal Dock are surrounded by commercial properties like pubs, retail complexes, big IT businesses, bus stops, and restaurants, making it one of the busiest stations. These stations provide last-mile connectivity to commuters. Contrarily, the majority of the “blue circles” stations are situated outside of Dublin, making them the least used stations. All these stations are surrounded by residential areas, hospitals, green spaces, and empty lands where human activity is relatively lower compared to other parts of Dublin. The majority of the stations in the “red circles” are situated in the city's heart and have a medium level of activity. The majority of these stations are within a two-kilometre radius of Luas and buses. As a result, users visit these stations less often than the green circled stations.

5.2.5 Station analysis

In this investigation, we selected a limited number of stations at random from each cluster and looked at their activity throughout the day on weekdays, Saturdays, and Sundays.

`

According to the Figure 5.13, weekends and weekdays are busy at Charlemont Place station, where several neighbouring stores, establishments, pubs, and residences are surrounded around this station. Because many individuals ride bikes from home to the workplace, bike traffic peaks between 5 and 10 am on weekdays. Due to the possibility that users won't use bicycles during working hours, bicycle activity is lower around midday. However, activity increases once again between 3 and 7 pm when individuals use bicycles to return home or to bus or train terminals. Weekend bike traffic is higher around noon since more people enjoy the outdoors. Since there is no commercial property in Grangegorman Lower (Central), which is near to TUDublin University and St. Brendan's Hospital, there is very little traffic on weekdays or weekends. Despite the station's proximity to the university, students choose not to ride because of the lack of recreational facilities such as parks, pubs, theatres, hotels, and restaurants. As a result, when it comes to mode of transportation, they choose public transport. On the other hand, because of the similarity of their surrounding areas, Kilmainham Gaol and Upper Sherrard Street stations see similar levels of bicycle activity on a daily basis.

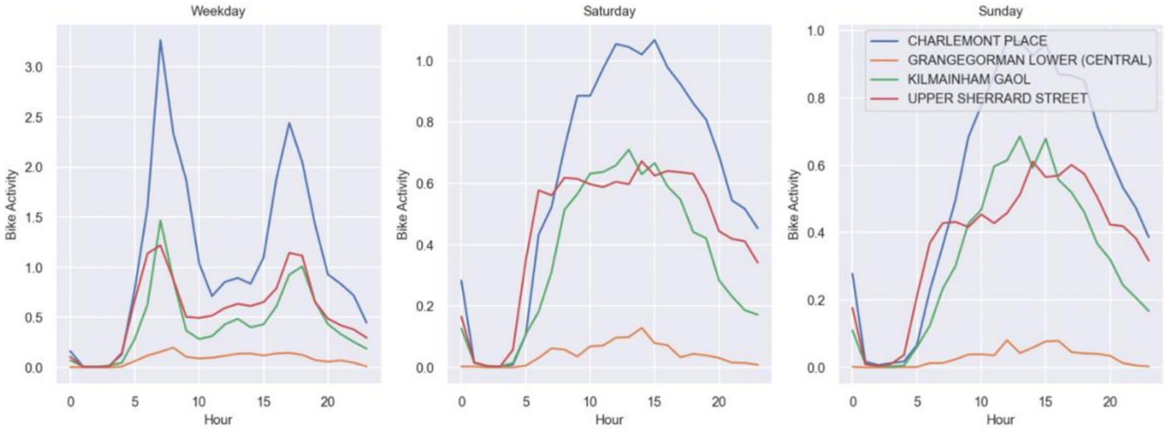


Figure 5.13 Bike Activity levels on Weekdays and Weekends for few Dublin Stations

`

5.3 Bike Availability Analysis

The Dublin bikes program relies heavily on the availability of bicycles. It's a way to figure out which stations get the most usage from cyclists and which ones are less used . This analysis helps us to find which stations require most rebalancing to meet users demand.

5.3.1 Station Analysis

To identify which stations had the highest and lowest bike availability, the percentage of bikes at each station was calculated, and the average for each station was calculated. The results obtained for each station were visualised using bar plots as show in Figure 5.14. It was found that Grangegorman Lower (South, North, and Central) and Hanover Quay East, the two stations closest to TU Dublin, had the fewest bicycles accessible in total. These results may confuse us considering that low availability is due to the high usage of bicycles. In the section 5.2, we conducted a bike activity study and concluded that these stations had lower activity. Less availability is due to fewer bike returns or inadequate rebalancing by the bike organization. On the other hand, Heuston station(South, North) and Smithfield North have the highest bike availability, meaning fewer people use them.

Despite the fact that the stations at Portobello Harbour and Charlemont Place are well recognized for being quite busy, there are more bikes available than usual. This leads us to the conclusion that users are not only picking the bikes but also returns them, and the organization is keeping up with the rebalancing activities to satisfy user demand.

`

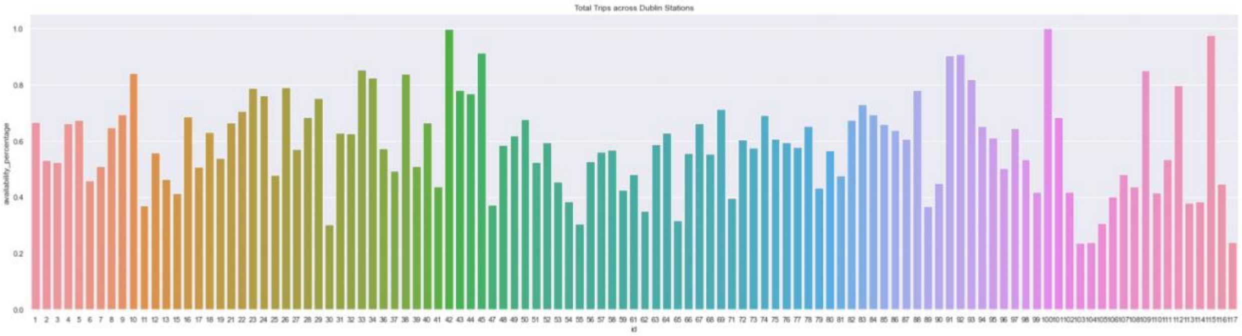


Figure 5.14 Bike Availability Percentage for Dublin Stations

5.3.2 Clustering

In a similar way to the implementation of clustering for bike activity (Section 5.2.6), we wanted to identify patterns or commonalities between the Dublin Bike stations. Therefore, we grouped all the stations using K-Means using four clusters and then plotted the availability of bikes on weekdays and weekends. After doing this study, we discovered several intriguing things about the stations' behaviour.

Weekday

From the Figure 5.15, on weekdays the availability of bikes at green clustered stations decreases while the availability of bikes at red clustered stations increases between 6 a.m. and 10 a.m. People commute from green-cluster stations to red-cluster stations because they move from their homes or bus stops to their workplaces, which makes sense. Green cluster stations are more likely to be in residential areas, whereas red clusters are more likely to be found in business regions (City Center). In the evening, employees return to their homes; therefore, there is high activity.

`

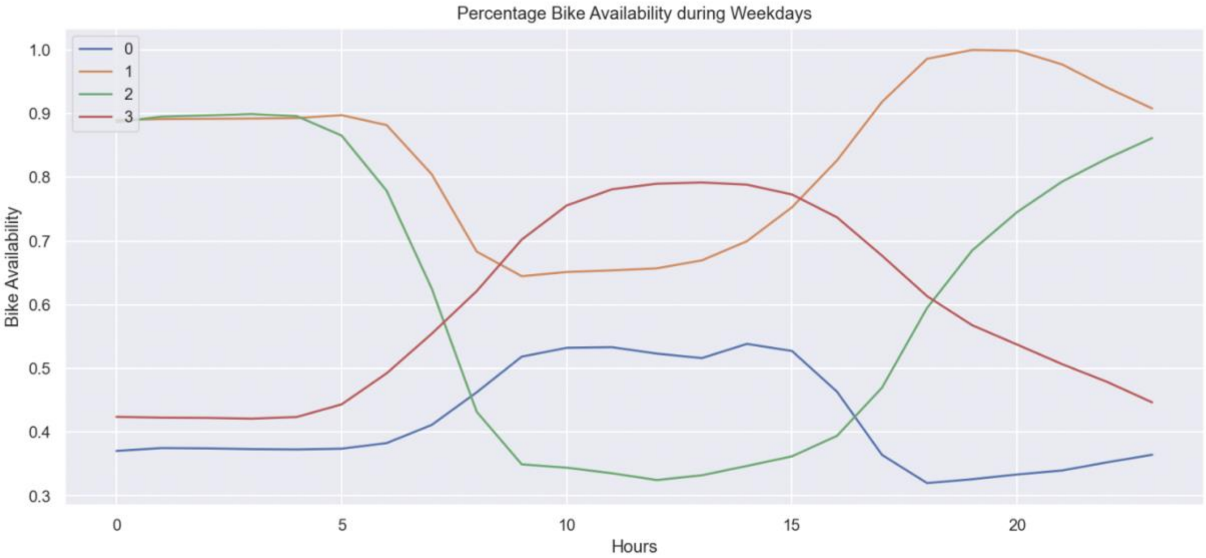


Figure 5.15 Bike Availability Trends for Clustered Station on Weekdays

Weekend

During weekends (see Figure 5.16), there is a contrasting behaviour between the red and green cluster stations, where the bike availability changes during the afternoon and evening, which clearly says that users travel from their residences to commercial places where there are more outdoor activities like parks, theatres, hotels, pubs, bars, and restaurants, etc. On the other hand, there is no noticeable change in bike availability in the orange and blue clustered stations.

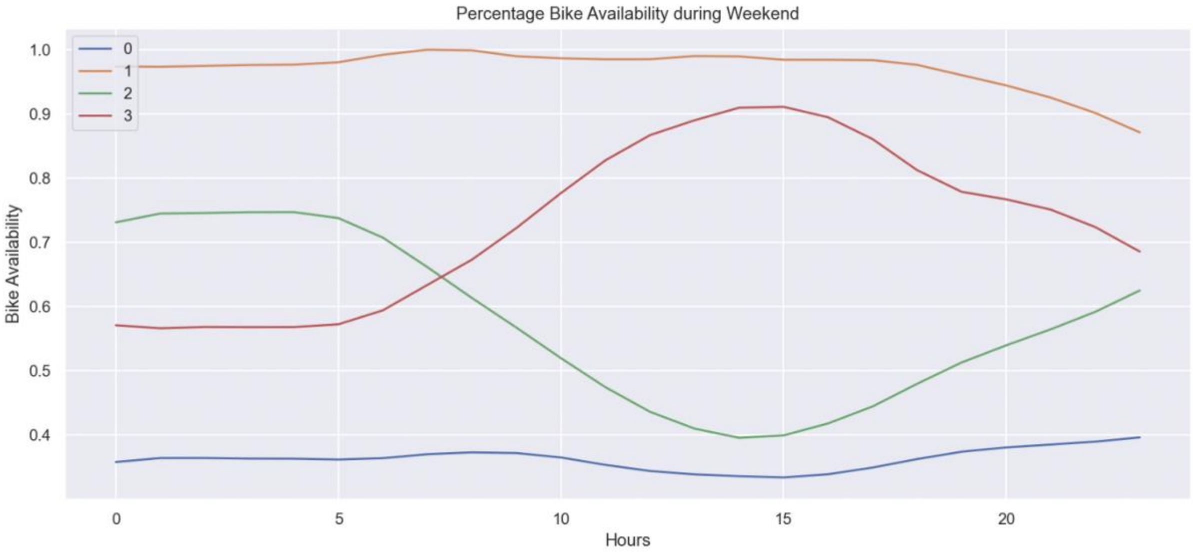


Figure 5.16 Bike Availability Trends for Clustered Station on Weekends

`

5.4 Bike Sharing Journey Analysis

This investigation focuses primarily on the number of bike trips (pickups and drops) at each station. To determine the actual journey made by the user, we used the formula described in Section 3.1.

5.4.1 Stations Analysis

The locations with the greatest user journeys were Charlemont Place, Portobello Harbour, and Herbert Place, as seen in the Figure 5.17, the cause of which is discussed in previous chapters. The lowest bike sharing trips are those made on York Street West (West), Grangegoram Lower (Central, North, South), and Buckingham Street Lower (Central, North, South).

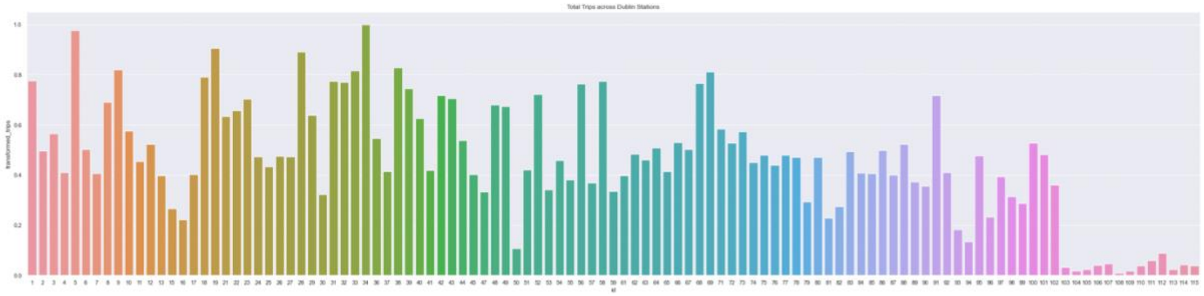


Figure 5.17 Bike Journeys trends across Dublin Stations

5.4.2 Hourly Distribution Analysis

Figure 5.18, shows that the number of bike trips count increases from five in the morning, reaching its peak at eight, then decrease till 10. This large value is understandable given the long distances that both students and employees must travel to go to their places of employment and educational institutions. As most people were already at their places of employment, the number of trips was comparatively low at midday compared to the morning. Once again, after 3 o'clock, people go back home, which necessitates numerous trips. Many businesses, including shops,

`

pubs, and motels, are closed at night, thus there is no movement of people. As a consequence, there are no trips shown at this time.

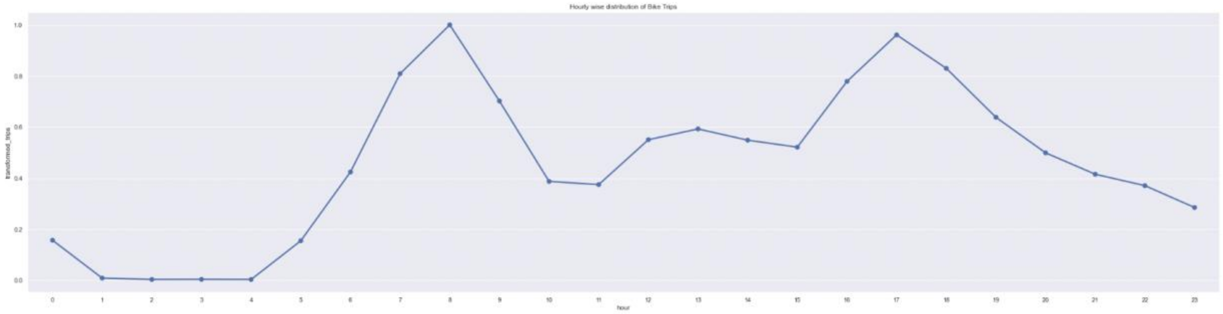


Figure 5.18 Bike Journeys trends along the day

5.4.3 Weekly and Hourly Journeys Trends

We were interested in learning how many trips may be made at the few random stations during the course of each and every workday. We created a heatmap that is appropriate for this sort of study. In the heat map, as the value increases, the rectangular box's colour intensity also increases. In other words, it gives extreme colours to extreme values, making them easily visible to the naked eye.

Princes Street / O'Connell Street

O'Connell Street runs northward from O'Connell Bridge towards Parnell Square on the north side of Dublin city. Based on the geospatial analysis, we found famous shopping centres like Iliac and Jervis and An Post's headquarters are nearby to O'Connell Street. In addition, there are bus stops and Luas stops all along the roadway. The facts above made us think that looking into the Princes Street/O'Connell Street bike station would provide some insightful information.

Thus, a heatmap was used to visualize the number of trips occurring every weekday throughout the day. According to the heatmap Figure 5.19, there is always a noticeable number of journeys at this station, with most rectangular boxes being light blue to dark blue. Unlike other bike station

(Figure 5.20 and Figure 5.21), we can see trips counts after 10 pm. Due to its location in the city, this station experiences the highest number of journeys compared to other stations.

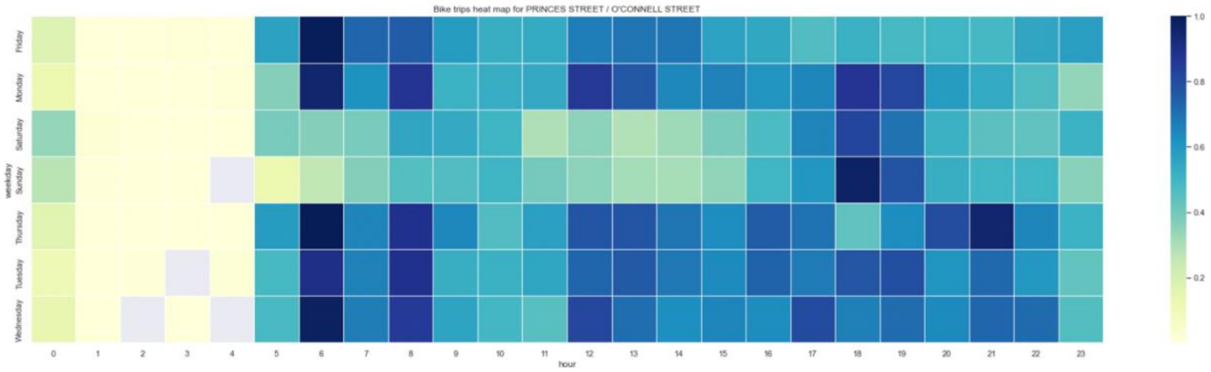


Figure 5.19 Bike Journeys at Princes Street / O'Connell Street

Grand Canal Dock

From Figure 5.20, it is clearly seen this station has more trips during all the weekdays except on Mondays at 7-8 am and 4-5pm who uses bikes for commuting. During afternoon, the number of trips looks average as most of the users are busy in their work. Several hotels, parks, IT companies like Google, Facebook, Twitter, shopping centres, and a canal surround this station. In the evening, many people visit the dock and enjoy the scenery of the river, making this station busier than in the morning.

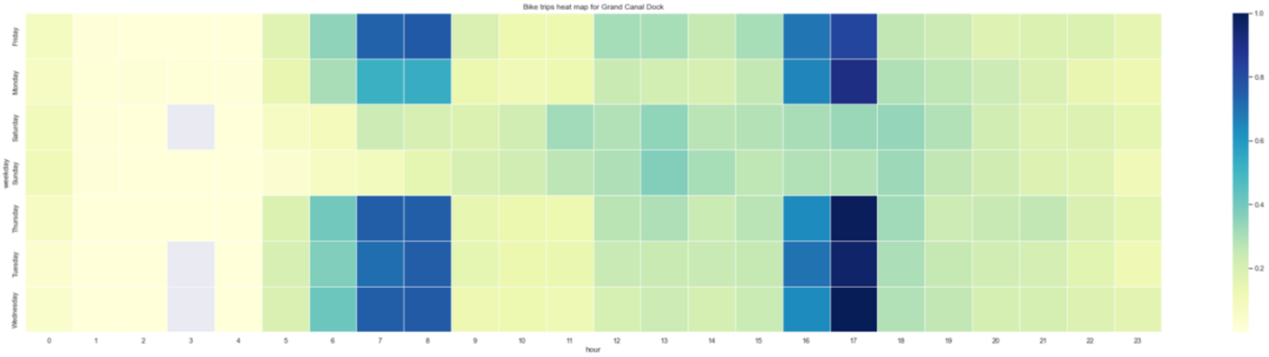


Figure 5.20 Bike Journeys at Grand Canal Dock

`

Heuston Station Dock (Car Park)

Analysis of the bike journey at Heuston Station (Figure 5.21). Unlike Princes Street/O'Connell Street and Grand Canal Dock stations, this station always seems quiet during peak business hours. Although there are a few options for outdoor activities nearby, such as hotels, parks, and restaurants, the station is able to make only fewer trips. We were intrigued to learn what was causing such unusual behaviour. The only possible explanation for this is inadequate rebalancing management by the organization.

To verify this, we looked at the bike availability at the station in Figure 5.22. We discovered that there were no bikes available during the peak biking hours of 9 a.m. to 3 p.m. This happens because, as the day's first train arrives, people start using their bicycles to go to work, leaving no bikes on the racks. However, as soon as people start returning their home from work, the stands start to fill up with bicycles. In addition, many people don't prefer to ride bikes and travel in Luas, especially if they travel to work after commuting in from the suburbs.

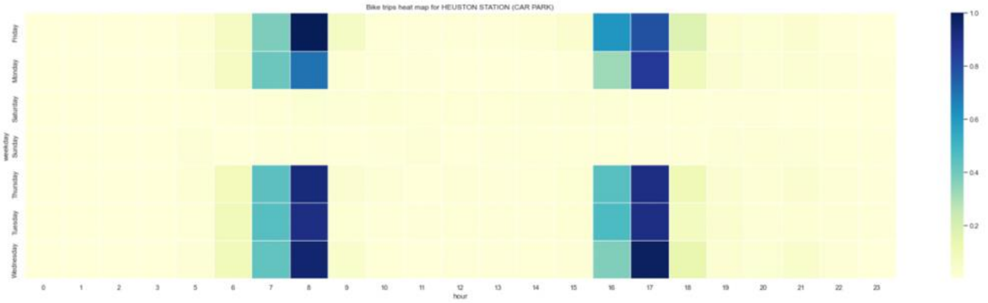


Figure 5.21 Bike Journeys at Heuston Station Dock

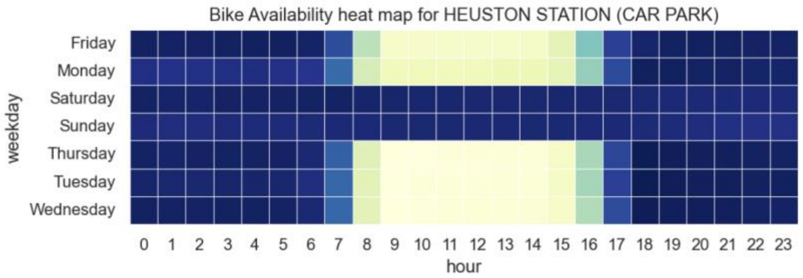


Figure 5.22 Bike Availability at Heuston Station Dock

`

Chapter Six : Model Building and Evaluation

In this chapter, we put Chardon's models into practice and evaluate the linear assumptions that is discussed in Section 3.2 & 3.3 and also implemented few traditional machine learning to estimate bike-sharing trips based on station-level data.

For Chardon's model to work, interaction data must be available in order for it to estimate bike-sharing trips. However, in the real world, making the interaction data available is a challenge, as it involves several forms of transformation, and the number of bike trips depends on a wide range of factors that contribute to the frequency of bike trips. It was therefore necessary to develop machine learning models like regression and classification that take input as influential factors like month, holiday, weekday, etc into account and estimate the number of daily bike journeys.

In the regression model, a bike-share trip value is estimated as the response, whereas in the classification model, a day is classified into low, medium, and high levels of bike trips. Finally, when all of the models have been implemented, the metrics outlined in Section 3.4 were used to compare and contrast them. In the end, the best-performing model is selected to provide the most accurate forecasts possible.

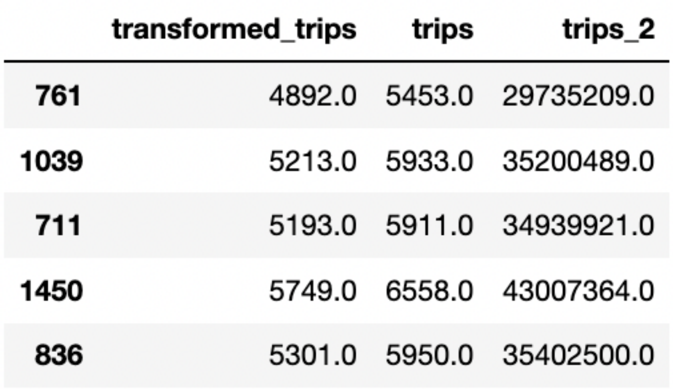
6.1 Chardon’s models

Using the algorithm and formula described in Section 3.1 & 3.2, the data is transformed based on the requirements of each model. As any machine learning model involves two phases of training and testing, the transformed dataset is randomly split at 70:30 to accommodate the training and testing phases of the model. As a final step, all four models are constructed using ordinary least squares (Linear Regression), and linear assumptions are verified and discussed along with the model summary.

`

6.1.1 Individual day aggregated model (iDAM)

The iDAM model, which is discussed in Section 3.2.1, is implemented using transformed data. A transformed sample dataset is displayed in Figure 6.1.

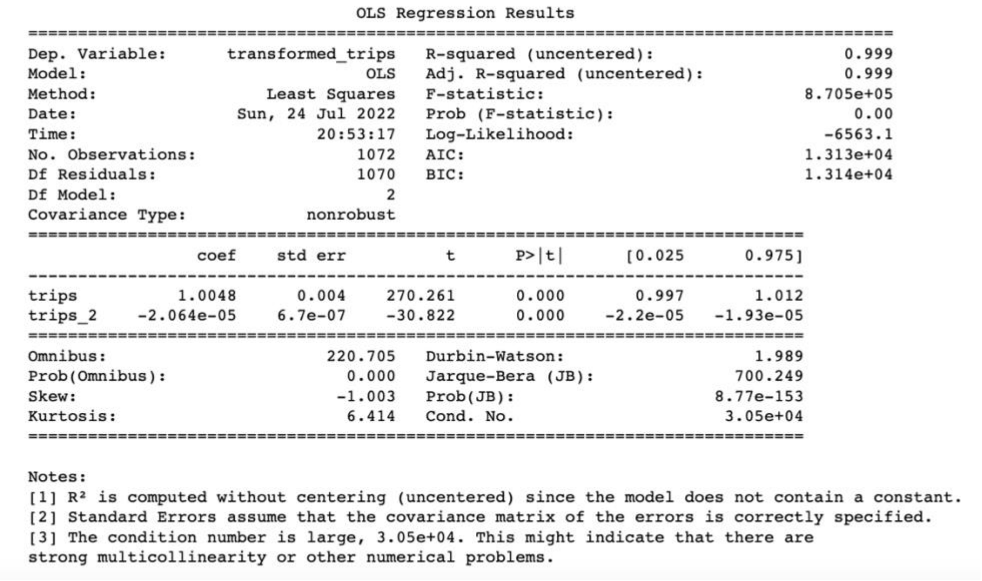


	transformed_trips	trips	trips_2
761	4892.0	5453.0	29735209.0
1039	5213.0	5933.0	35200489.0
711	5193.0	5911.0	34939921.0
1450	5749.0	6558.0	43007364.0
836	5301.0	5950.0	35402500.0

Figure 6.1 Sample Dataset for iDAM Model

Model Summary:

From the Figure 6.2, the p-value for the dependent variables (trips, trips_2) is less than 0.05 which means that they are highly significant to the response variable (transformed trips). Here R^2 and Adjusted-R^2 values are closer to 1 in Figure 6.2. This means that the model is performing well to predict bike sharing trips.



```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:      transformed_trips   R-squared (uncentered):                   0.999
Model:                            OLS   Adj. R-squared (uncentered):              0.999
Method:                 Least Squares   F-statistic:                          8.705e+05
Date:                Sun, 24 Jul 2022   Prob (F-statistic):                        0.00
Time:                        20:53:17   Log-Likelihood:                         -6563.1
No. Observations:                1072   AIC:                                  1.313e+04
Df Residuals:                    1070   BIC:                                  1.314e+04
Df Model:                           2                                                  
Covariance Type:            nonrobust                                                  
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
trips          1.0048      0.004    270.261      0.000       0.997       1.012
trips_2    -2.064e-05    6.7e-07    -30.822      0.000    -2.2e-05   -1.93e-05
==============================================================================
Omnibus:                      220.705   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              700.249
Skew:                          -1.003   Prob(JB):                    8.77e-153
Kurtosis:                       6.414   Cond. No.                     3.05e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 3.05e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 6.2 Model Summary from iDAM model

`

Linear Assumptions

Linearity

Analysis of the relationships between the dependent and independent variables is carried out using a scatterplot. The dependent (trips) and independent (transformed trips) variable points lie along a straight line in Figure 6.3, indicating that the two variables are linearly related. As a result, the linearity assumption is satisfied.

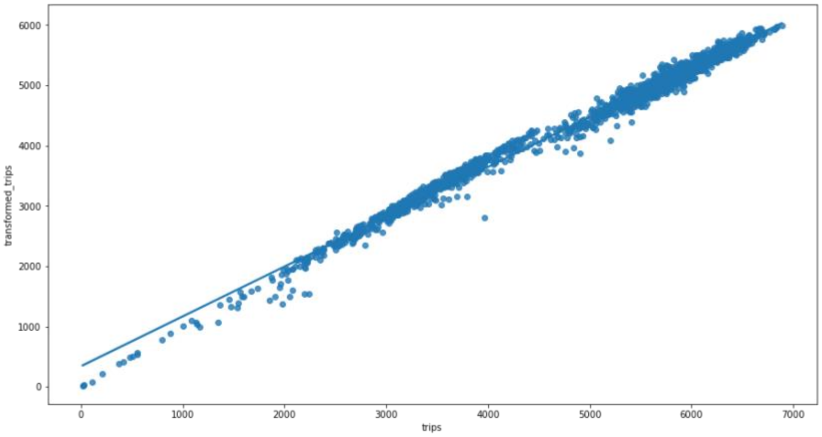


Figure 6.3 Trips vs Transformed Trips Scatterplot (iDAM)

Heteroscedasticity

Although many ways of detecting heteroscedasticity exist, we opted for the residual Vs fitted scatterplot plot and Goldfeld Quandt test. As discussed in Section 3.3.3, we used the residual Vs fitted scatterplot plot and Goldfeld Quandt test to detect heteroscedasticity.

Figure 6.4 shows residuals on the y-axis and predicted values on the x-axis. This figure shows that the residual points lie along the regression line and that there is no pattern in the residual points, indicating that the variance of error terms with respected to overall y prediction is constant. As a result, there is evidence that residuals are homoscedastic. Because there are few outliers in the Figure 6.4, we conducted the Goldfeld Quandt test to confirm the observation from the plot.

`

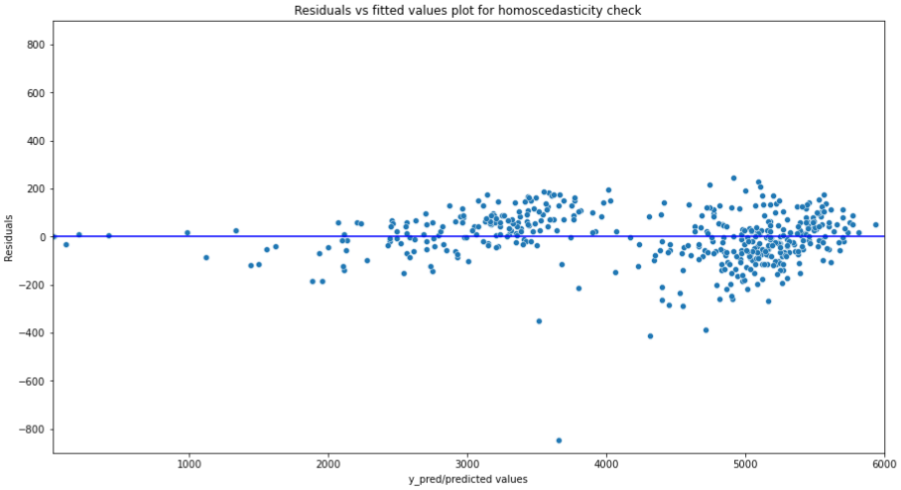
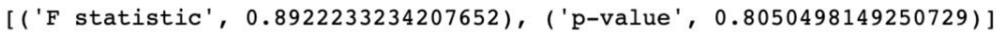


Figure 6.4 Residual Vs Fitted plot (iDAM)

- **Goldfeld Quandt Test :** The p-value obtained by the Goldfeld Quandt was 0.805. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis which means that residuals are homoscedastic. Therefore, the assumption is satisfied.



```
[('F statistic', 0.8922233234207652), ('p-value', 0.8050498149250729)]
```

Figure 6.5 Goldfeld Quandt Test Results (iDAM)

Normality

In order to verify the normality, histogram and qqplot are be generated from residual values. According to histogram (Figure 6.6), the residuals formed a bell-shaped curve which indicates that residuals are normally distributed. In addition from qqplot (Figure 6.7), all the residual points lie on the regression line. From this, we can conclude that normality assumption is satisfied.

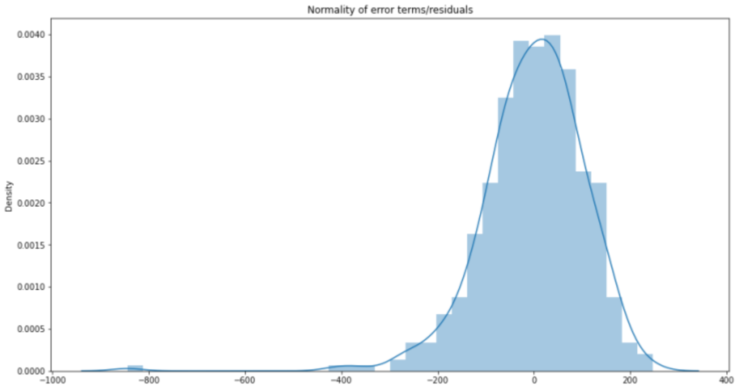


Figure 6.6 Histogram for residuals (iDAM)

`

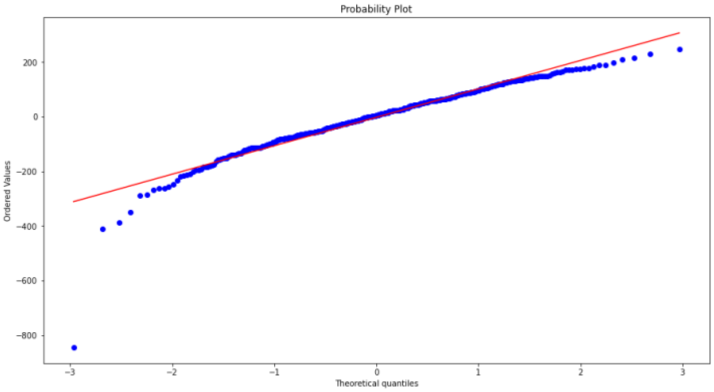


Figure 6.7 QQplot for residuals (iDAM)

Autocorrelation

In order to identify autocorrelations within the residuals, tests like Ljungbox and Durbin-Watson are used along with autocorrelation plots are implemented. As shown in Figure 6.8, there are no spikes outside the blue confidence interval which means that the residuals are not autocorrelated.

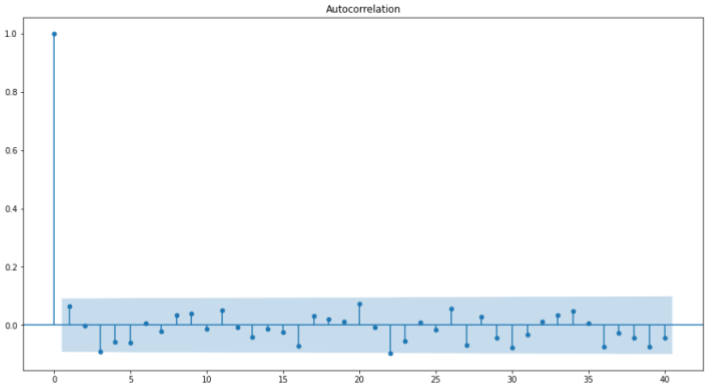


Figure 6.8 Autocorrelation Plot for Residuals (iDAM)

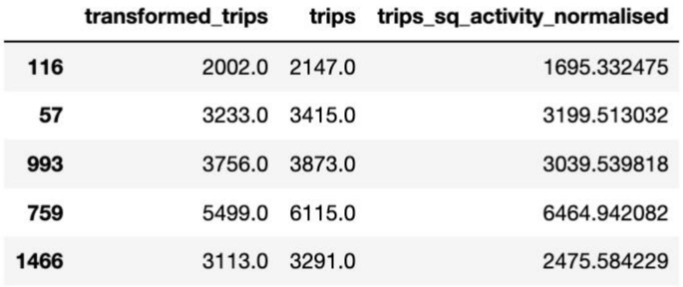
- **Ljungbox test:** The p-value obtained by the Ljungbox test was 0.1068. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis, meaning that residuals are not auto-correlated to each other.
- **Durbin Watson Test:** According to the model summary Figure 6.2, the score for the Durbin Watson test is closer to 2, indicating that there is no autocorrelation between residuals.

Considering overall results from the autocorrelation evaluation, we can say that residuals are not autocorrelated. Therefore, the assumption is satisfied.

`

6.1.2 Combined day aggregated model (cDAM)

The cDAM model, which is discussed in Section 3.2.2, is implemented using the transformed data. A transformed sample dataset is displayed in Figure 6.9.

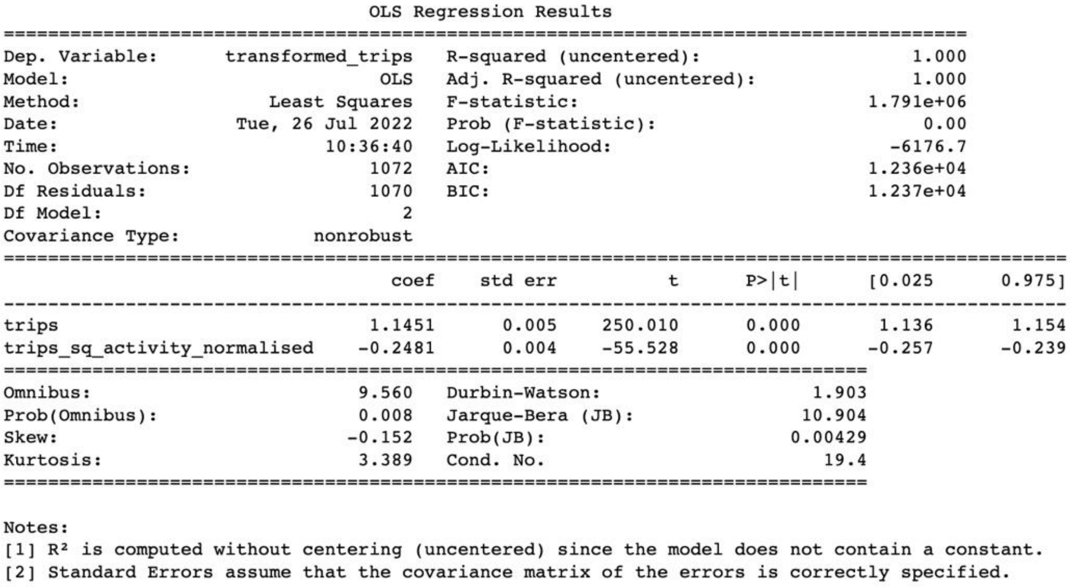


	transformed_trips	trips	trips_sq_activity_normalised
116	2002.0	2147.0	1695.332475
57	3233.0	3415.0	3199.513032
993	3756.0	3873.0	3039.539818
759	5499.0	6115.0	6464.942082
1466	3113.0	3291.0	2475.584229

Figure 6.9 Sample Dataset for cDAM Model

Model Summary:

From the Figure 6.10, the p-value for the dependent variables (trips, trips_sq_activity_normalised) is less than 0.05 which means that they are highly significant to the response variable (transformed trips). Here R^2 and Adjusted-R^2 values are closer to 1 in Figure 6.10. This means that the model is performing well to predict bike sharing trips.



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:      transformed_trips   R-squared (uncentered):                   1.000
Model:                            OLS   Adj. R-squared (uncentered):              1.000
Method:                 Least Squares   F-statistic:                          1.791e+06
Date:                Tue, 26 Jul 2022   Prob (F-statistic):                        0.00
Time:                        10:36:40   Log-Likelihood:                         -6176.7
No. Observations:                1072   AIC:                                  1.236e+04
Df Residuals:                    1070   BIC:                                  1.237e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
trips                          1.1451      0.005    250.010      0.000       1.136       1.154
trips_sq_activity_normalised  -0.2481      0.004    -55.528      0.000      -0.257      -0.239
==============================================================================
Omnibus:                        9.560   Durbin-Watson:                   1.903
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.904
Skew:                          -0.152   Prob(JB):                      0.00429
Kurtosis:                       3.389   Cond. No.                         19.4
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 6.10 Model Summary from cDAM model

`

Linear Assumptions

Linearity

The dependent (trips) and independent (transformed trips) variable points lie along a straight line in Figure 6.11, indicating that the two variables are linearly related. As a result, the linearity assumption is satisfied.

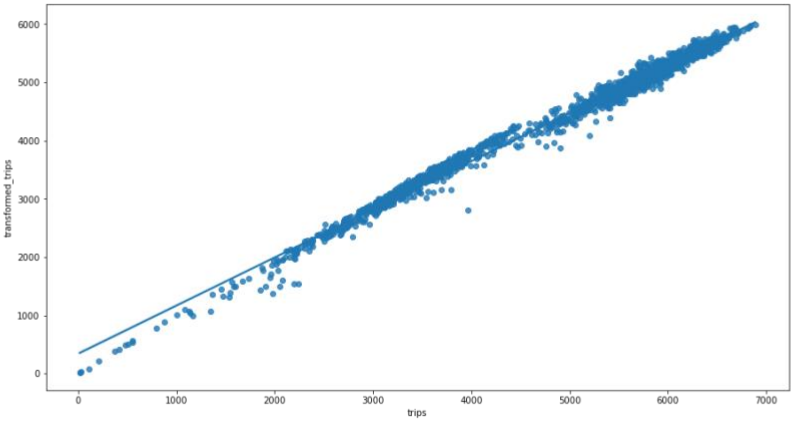


Figure 6.11 Trips vs Transformed Trips Scatterplot (cDAM)

Heteroscedasticity

Figure 6.12, demonstrates that the residual points are scattered over the regression line rather than falling on it. However, the variance does not increase as the overall y prediction increases, indicating no pattern (Cone-Shape) in the residual points. Because there are a few outliers in the figure, we conducted the Goldfeld Quandt test to confirm the observation from the plot.

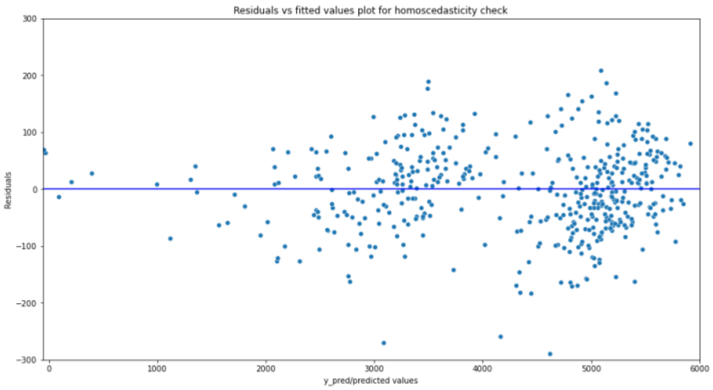
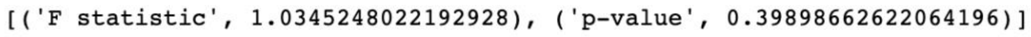


Figure 6.12 Residual Vs Fitted plot (cDAM)

`

- **Goldfeld Quandt Test :** The p-value obtained by the Goldfeld Quandt was 0.398. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis, meaning that residuals are homoscedastic. Therefore, the assumption is satisfied.



```
[('F statistic', 1.0345248022192928), ('p-value', 0.39898662622064196)]
```

Normality

According to histogram (Figure 6.13), the residuals formed a bell-shaped curve which indicates that residuals are normally distributed. In addition, from qqplot (Figure 6.14), all the residual points lie on the regression line. From this, we can conclude that normality assumption is satisfied.

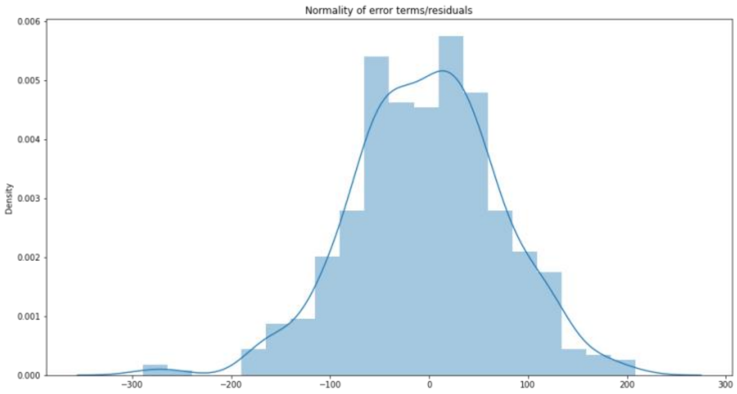


Figure 6.13 Histogram for residuals (cDAM)

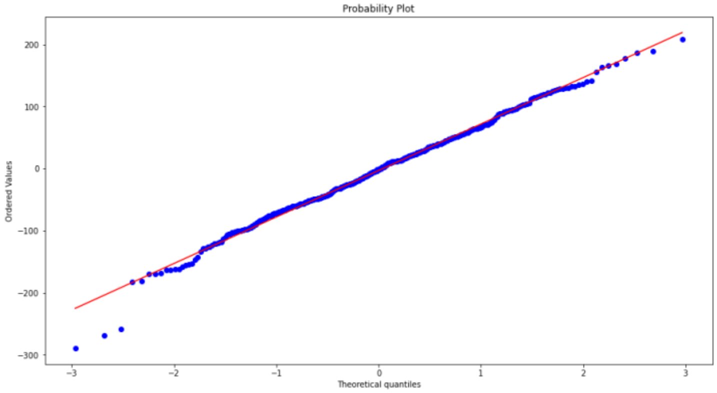


Figure 6.14 QQPlot for residuals (cDAM)

`

Autocorrelation

As shown in Figure 6.15, there are no spikes outside the blue confidence interval which means that the residuals are not autocorrelated.

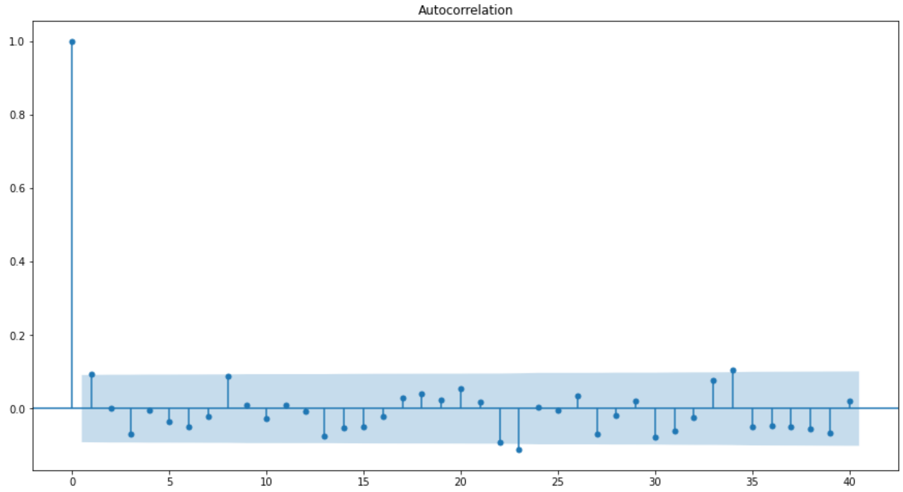


Figure 6.15 Autocorrelation Plot for Residuals (cDAM)

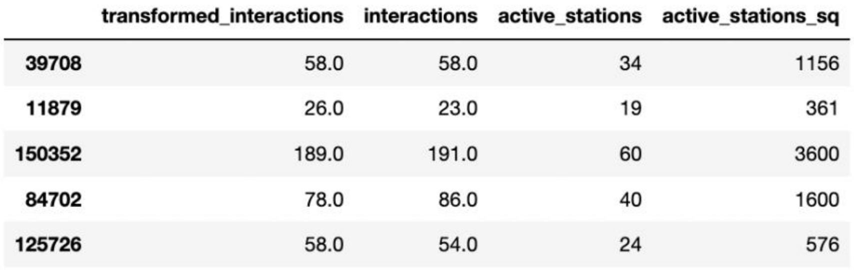
- **Ljungbox test:** The p-value obtained by the Ljungbox test was 0.0416, which is closer to the significance value(0.05). According to this test, we reject the null hypothesis if the p-value is less than 0.05. Therefore, as per the Ljungbox test, residuals are auto-correlated.
- **Durbin Watson Test:** According to the model summary Figure 6.10, the score for the Durbin Watson test is closer to 2, indicating that there is no autocorrelation between residuals.

Durbin-Watson and autocorrelation plot evaluations were both successful, whereas the Ljung-Box test failed with just a small deviation from a significant value which can be ignored. Therefore, we can conclude that residuals are not autocorrelated.

`

6.1.3 Interval aggregation model (IAM)

The IAM model, which is discussed in Section 3.2.3, is implemented in this part using transformed data. A transformed sample dataset is displayed in Figure 6.16.

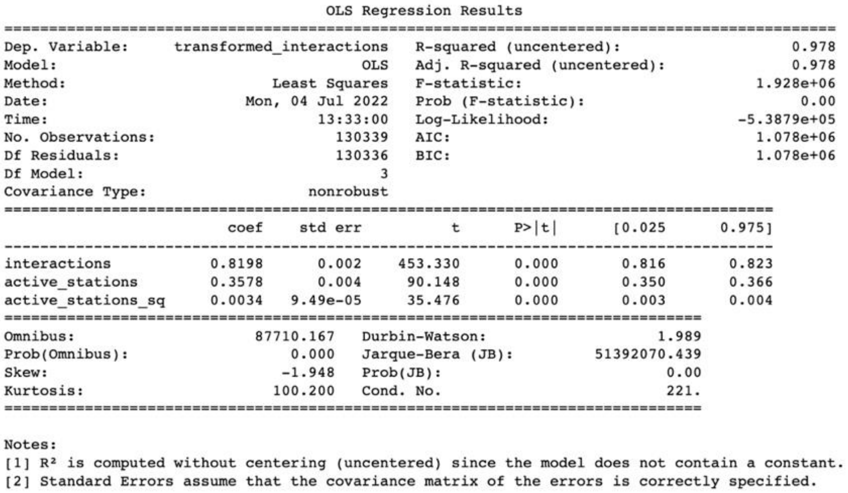


	transformed_interactions	interactions	active_stations	active_stations_sq
39708	58.0	58.0	34	1156
11879	26.0	23.0	19	361
150352	189.0	191.0	60	3600
84702	78.0	86.0	40	1600
125726	58.0	54.0	24	576

Figure 6.16 Sample Dataset for IAM model

Model Summary:

From the Figure 6.17, the p-value for the dependent variables (interactions, active stations, active station sq) is less than 0.05 which means that they are highly significant to the response variables (transformed interactions). Here R^2 and Adjusted-R^2 values are closer to 1 in Figure 6.17. This means the model is performing well to predict interactions.



```
                                 OLS Regression Results
=============================================================================================
Dep. Variable:     transformed_interactions   R-squared (uncentered):                   0.978
Model:                                  OLS   Adj. R-squared (uncentered):              0.978
Method:                       Least Squares   F-statistic:                          1.928e+06
Date:                      Mon, 04 Jul 2022   Prob (F-statistic):                        0.00
Time:                              13:33:00   Log-Likelihood:                     -5.3879e+05
No. Observations:                    130339   AIC:                                  1.078e+06
Df Residuals:                        130336   BIC:                                  1.078e+06
Df Model:                                 3
Covariance Type:                  nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
interactions           0.8198      0.002    453.330      0.000       0.816       0.823
active_stations        0.3578      0.004     90.148      0.000       0.350       0.366
active_stations_sq     0.0034   9.49e-05     35.476      0.000       0.003       0.004
==============================================================================
Omnibus:                    87710.167   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         51392070.439
Skew:                          -1.948   Prob(JB):                         0.00
Kurtosis:                     100.200   Cond. No.                         221.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 6.17 Model Summary from IAM model

`

Linear Assumptions

Linearity

As shown in figure 6.18, majority of dependent (interactions) and independent (transformed interactions) variable points lie along a straight line except few outliers which are spread across the line. Due to this we cannot consider that two variables are linearly related. Therefore, the assumption is failed.

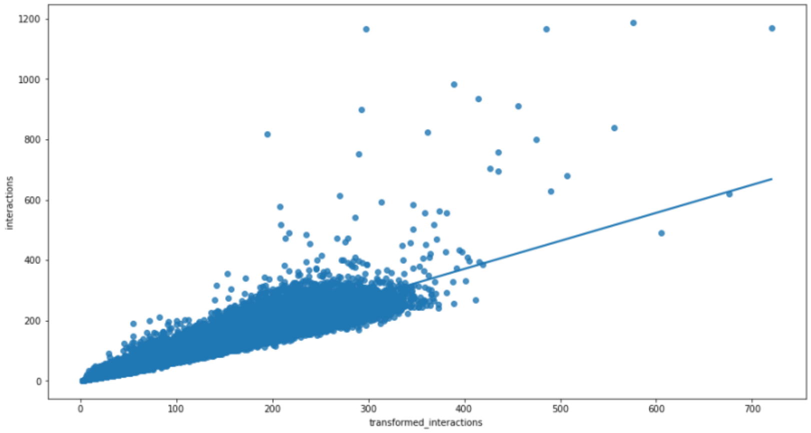


Figure 6.18 Interactions vs Transformed Interactions Scatterplot (IAM)

Heteroscedasticity

Figure 6.19, shows that the error terms varied marginally when the predictor variable changed(increasing trend & Cone shaped), which means that the variance of error terms is not constant. This suggests that the residuals are heteroscedastic. To validate the observations from the plot, we used the Goldfeld Quandt test.

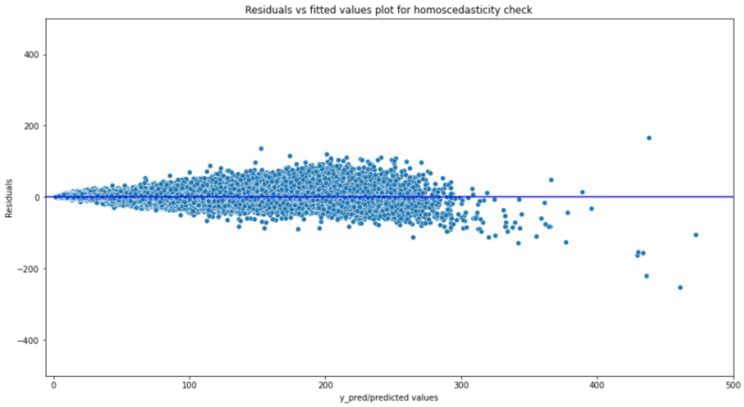
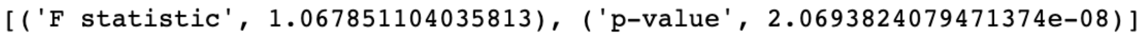


Figure 6.19 Residual Vs Fitted Plot (IAM)

`

- **Goldfeld Quandt Test :** The p-value obtained by the Goldfeld Quandt test was less than zero. According to this test, we reject the null hypothesis if the p-value is less than 0.05. Therefore, as per the Goldfeld Quandt test, the residuals are not homoscedastic.



```
[('F statistic', 1.067851104035813), ('p-value', 2.0693824079471374e-08)]
```

Considering overall results from Homoscedasticity evaluation, we can say that residuals are not homoscedastic. Therefore, the assumption is failed.

Normality

The residuals did not form a bell-shaped curve, but rather a narrow curve, as seen in histogram Figure 6.20. Another interesting finding is that the residual points close to the tails of the regression line are outliers that do not lie on the regression line, as shown in qqplot (Figure 6.21). If outliers are removed, then all the points fall on the line. Therefore, we conclude that residuals are normally distributed which means that the normality assumption is satisfied.

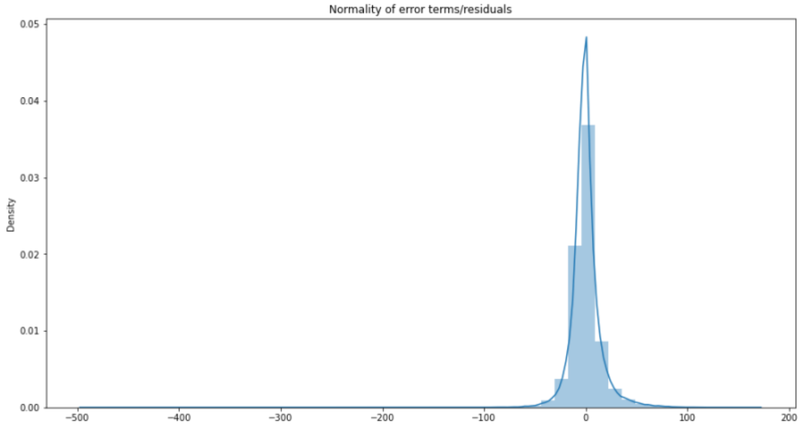


Figure 6.20 Histogram for residuals (IAM)

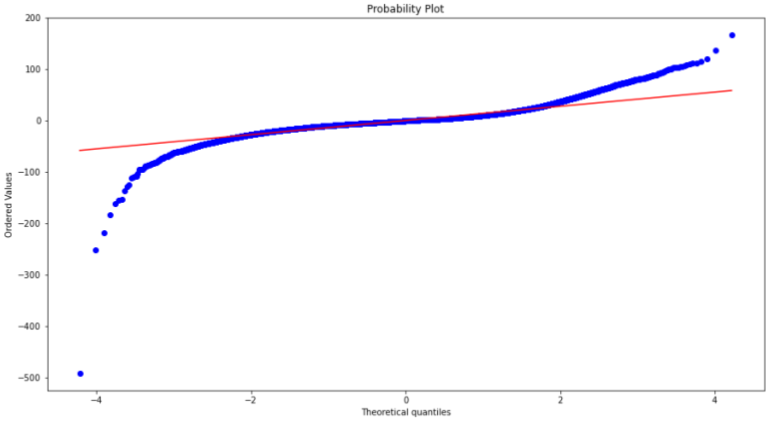


Figure 6.21 QQPlot for residuals (IAM)

`

Autocorrelation

As shown in Figure 6.22, there are no spikes outside the blue confidence interval which means the residuals are not autocorrelated.

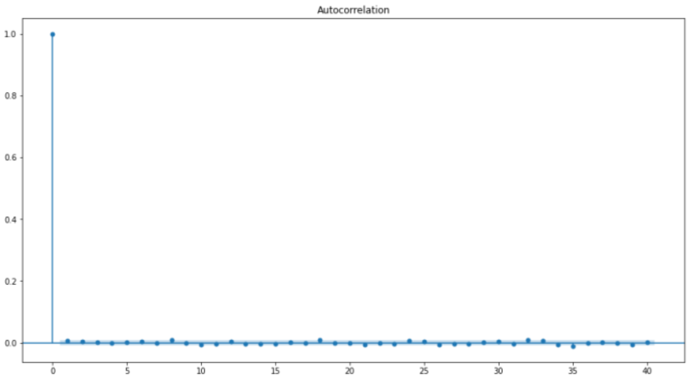


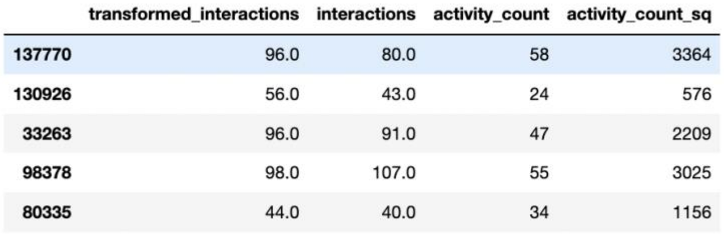
Figure 6.22 Autocorrelation plot for residuals (IAM)

- **Ljungbox test:** The p-value obtained by the Ljungbox test was 0.074. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis, meaning that residuals are not auto-correlated to each other.
- **Durbin Watson Test:** According to the model summary Figure 6.17, the score for the Durbin Watson test is closer to 2, indicating that there is no autocorrelation between residuals.

Considering overall results from the autocorrelation evaluation, we can say that residuals are not autocorrelated which means that normality assumption is satisfied.

6.1.4 Station aggregation model (SAM)

The SAM model, which is discussed in Section 3.2.4, is implemented in this part using transformed data. A transformed sample dataset is displayed in Figure 6.23.



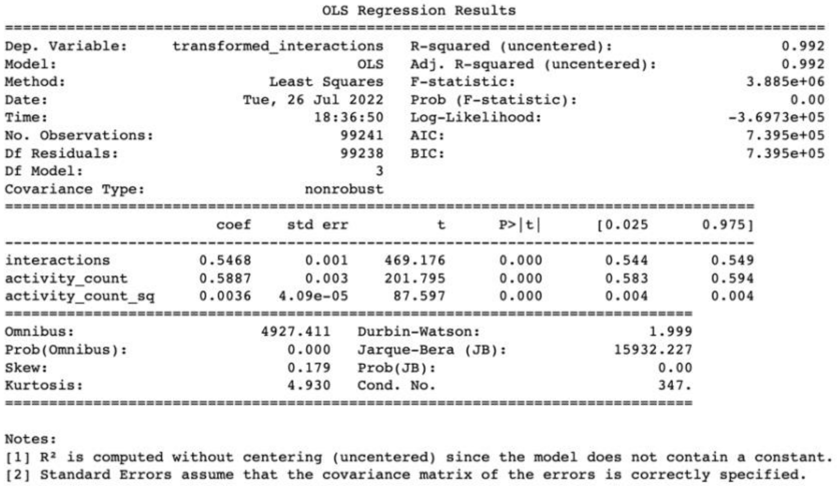
	transformed_interactions	interactions	activity_count	activity_count_sq
137770	96.0	80.0	58	3364
130926	56.0	43.0	24	576
33263	96.0	91.0	47	2209
98378	98.0	107.0	55	3025
80335	44.0	40.0	34	1156

Figure 6.23 Sample Dataset for SAM model

`

Model Summary:

From the Figure 6.24, the p-value for the dependent variables (interactions, activity count, activity count sq) is less than 0.05 which means that they are highly significant to the predictor (transformed interactions). As R^2 and Adjusted-R^2 values are closer to 1 in Figure 6.24. This means the model is performing well to predict interactions.



```
                                 OLS Regression Results
=============================================================================================
Dep. Variable:     transformed_interactions   R-squared (uncentered):                   0.992
Model:                                  OLS   Adj. R-squared (uncentered):              0.992
Method:                       Least Squares   F-statistic:                          3.885e+06
Date:                      Tue, 26 Jul 2022   Prob (F-statistic):                        0.00
Time:                              18:36:50   Log-Likelihood:                     -3.6973e+05
No. Observations:                     99241   AIC:                                  7.395e+05
Df Residuals:                         99238   BIC:                                  7.395e+05
Df Model:                                 3
Covariance Type:                  nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
interactions          0.5468      0.001    469.176      0.000       0.544       0.549
activity_count        0.5887      0.003    201.795      0.000       0.583       0.594
activity_count_sq     0.0036   4.09e-05     87.597      0.000       0.004       0.004
==============================================================================
Omnibus:                     4927.411   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15932.227
Skew:                           0.179   Prob(JB):                         0.00
Kurtosis:                       4.930   Cond. No.                         347.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 6.24 Model Summary from SAM model

Linear Assumptions

Linearity

The dependent (interactions) and independent (transformed interactions) variable points lie along a straight line in Figure 6.25, indicating that the two variables are linearly related. As a result, the linearity assumption is satisfied.

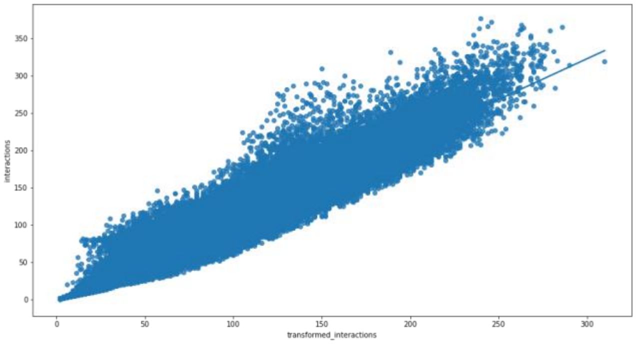


Figure 6.25 Interactions vs Transformed Interactions Scatterplot (SAM)

`

Heteroscedasticity

Figure 6.26, displays that the residual points lie along the regression line and there is no pattern in the residual points, indicating that the variance of error terms with respected to overall y prediction terms is constant. As a result, there is evidence that residuals are homoscedastic.

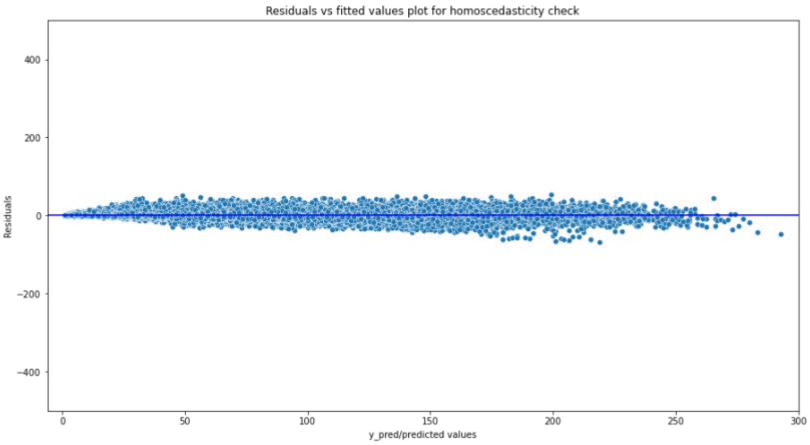
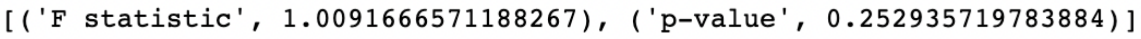


Figure 6.26 Residual Vs Fitted Plot (SAM)

- **Goldfeld Quandt Test :** The p-value obtained by the Goldfeld Quandt was 0.252. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis which means that residuals are homoscedastic which means that the homoscedasticity assumption is satisfied.



```
[('F statistic', 1.0091666571188267), ('p-value', 0.252935719783884)]
```

Normality

According to histogram (Figure 6.27), the residuals formed a bell-shaped curve which indicates that residuals are normally distributed. In addition, from qqplot (Figure 6.28), all the residual points lie on the regression line. From this, we can conclude that normality assumption is satisfied.

`

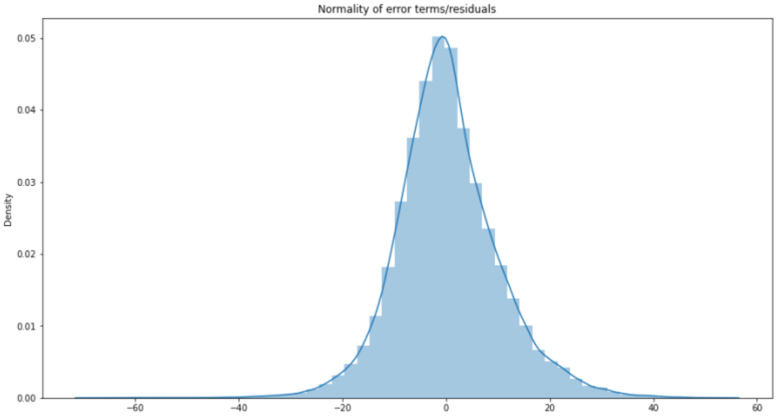


Figure 6.27 Histogram for residuals (SAM)

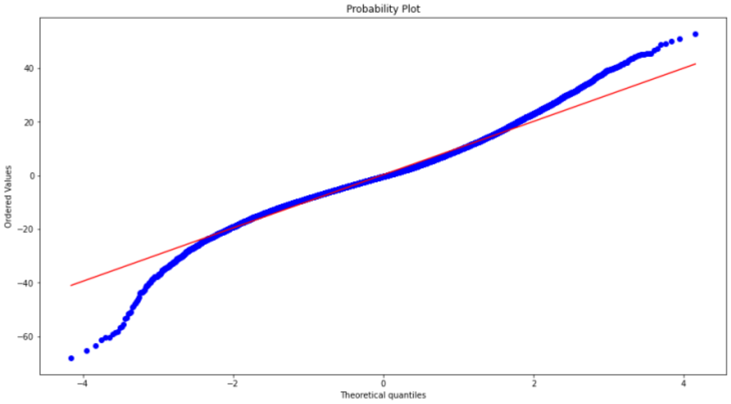


Figure 6.28 QQPlot for residuals (SAM)

Autocorrelation

As shown in Figure 6.29, there are no spikes outside the blue confidence interval which means the residuals are not autocorrelated.

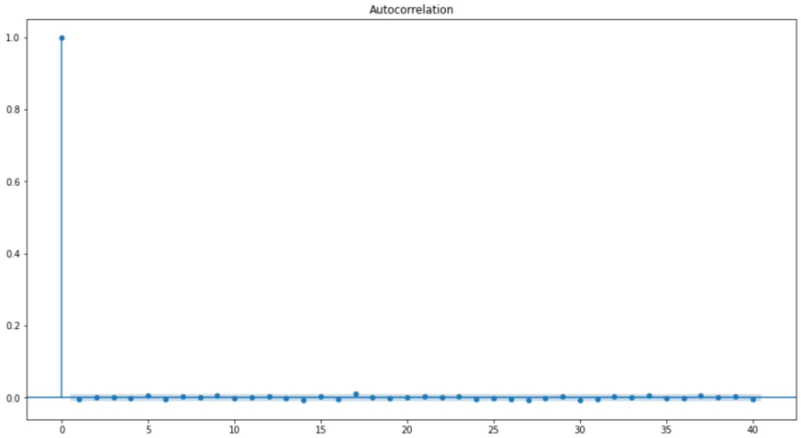


Figure 6.29 Autocorrelation plot for residuals (SAM)

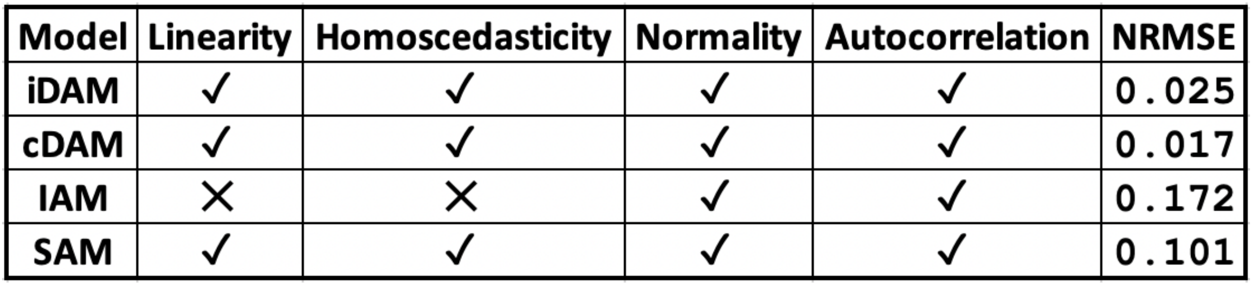
`

- **Ljungbox test:** The p-value obtained by the Ljungbox test was 0.502. According to this test, if the p-value is greater than 0.05, there is strong evidence for the null hypothesis, meaning that residuals are not auto-correlated to each other.
- **Durbin Watson Test:** According to the model summary Figure 6.24, the score for the Durbin Watson test is closer to 2, indicating that there is no autocorrelation between residuals.

Considering overall results from the autocorrelation evaluation, we can say that residuals are not autocorrelated. Therefore, the assumption is satisfied.

6.1.5 Summary

A summary of validation results for all models can be found in Table 6.1. Based on the results obtained, all assumptions of the linear models are satisfied by the iDAM, cDAM, and SAM models. Therefore, their predictions will be more accurate with lower error rates. Moreover, the iDAM model, which has a simpler implementation than other models, is capable of providing the best estimates of daily trips. The IAM model, on the other hand, did not satisfy all linear assumptions, leaving out linearity and homoscedasticity. Heteroscedastic residuals and the presence of outliers are the primary reasons for failure of linearity and homoscedasticity. These failures can be resolved by eliminating the outliers and applying transformations such as the Napierian logarithm or its square root to the dependent variables [60].



Model	Linearity	Homoscedasticity	Normality	Autocorrelation	NRMSE
iDAM	✓	✓	✓	✓	0.025
cDAM	✓	✓	✓	✓	0.017
IAM	✕	✕	✓	✓	0.172
SAM	✓	✓	✓	✓	0.101

Table 6.1 Comparison of Chardon's Models

`

6.2 Regression Models

Based on the background literature research in Chapter 2, regression models such as Linear Regression, Ridge Regression, Decision Tree, Gradient Boosting, and Random Forest Regression were implemented and evaluated as they had been found to be suitable and effective in predicting daily bike journeys.

Transforming the dataset into an appropriate format before training the model is essential for improving accuracy. As a result, the column's datatype was converted to categorical variables, and then the categorical variable was converted into dummy/indicator variables using the get_dummies() function that is available in python. As a final step, the transformed dataset is randomly split into 70:30 to accommodate the training and testing phases.
The following variables are used in regression models.

Response:

- transformed_trips : A continuous variable denoting the number of bike-share trips per day.

Predictors:

- season: A categorical value indicating the season in which the event occurred.
- weekday: A categorical variable that indicates which day of the week it is.
- holiday: A categorical value indicating whether a day is a national holiday or casual.
- month: A categorical variable that indicates which month it is.
- trip_ind: Categorical variable that denotes the level of bikes trips in a given day.

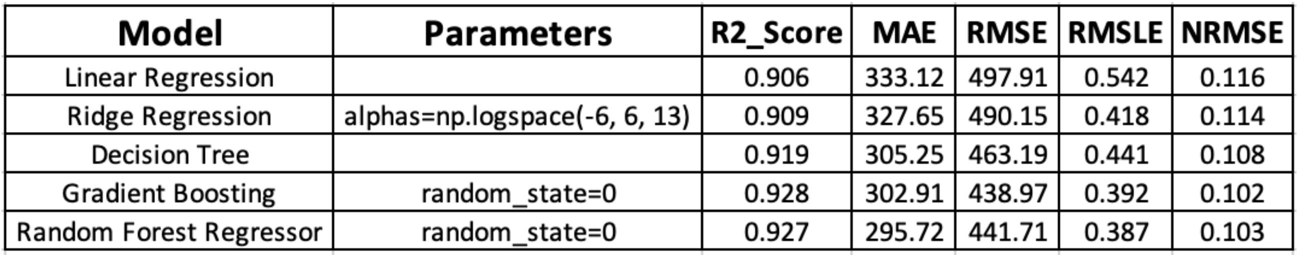
bike trips count > 3000 -> 0 (Low)
bike trips count <= 3000 & bike trips count <= 5000 -> 1 (Medium)
bike trips count > 5000 -> 2 (High)

As the weather in Ireland varies frequently and is unpredictable, we cannot use weather information to predict the number of bike rides for an entire day.

`

Regression models were trained and tested using the pre-processed dataset. The predicted bike-share trip count that are obtained from the testing dataset is evaluated by the regression metrics like R^2, MAE, RMSE, NRMSE, and RMSLE, discussed in Section 3.4.1. Lastly, the results were displayed in tabular format to compare and analyse the each model's performance.

Results :



Model	Parameters	R2_Score	MAE	RMSE	RMSLE	NRMSE
Linear Regression		0.906	333.12	497.91	0.542	0.116
Ridge Regression	alphas=np.logspace(-6, 6, 13)	0.909	327.65	490.15	0.418	0.114
Decision Tree		0.919	305.25	463.19	0.441	0.108
Gradient Boosting	random_state=0	0.928	302.91	438.97	0.392	0.102
Random Forest Regressor	random_state=0	0.927	295.72	441.71	0.387	0.103

Table 6.2 Comparison of Regression Models

In terms of model metrics displayed in Table 6.2, we found that the Random Forest Regressor and Gradient Boosting models were the most accurate. The gradient boosting algorithm was found to have an R^2 of 0.928, a MAE of 302.91, a RMSE of 438.97, a RMSLE of 0.392, and a NRMSE of 0.102, while a Random Forest algorithm had an R^2 of 0.927, a MAE of 295.72, a RMSE of 441.71, a RMSLE of 0.387, and a NRMSE of 0.103. It is acceptable for the metric values to be higher than usual due to the range of the response variable from 1000 to 6000. The least accurate models were linear regressions, which suggests that the data is not linear in nature.

`

6.3 Classification Models

Based on the background literature research in Chapter 2, classification models such as Logistic Regression, KNN, Naive Bayes, Decision Tree, LDA, and Random Forest Classifier were implemented and evaluated as they had been found to be suitable and effective in classifying the levels of bike journeys on a given day.

The following variables are used in classification models:

Response:

- trip_ind: A Categorical variable that denotes the level of bikes trips on a given day.

bike trips > 3000 -> 0 (Low)
bike trips <= 3000 & bike trips <= 5000 -> 1 (Medium)
bike trips > 5000 -> 2 (High)

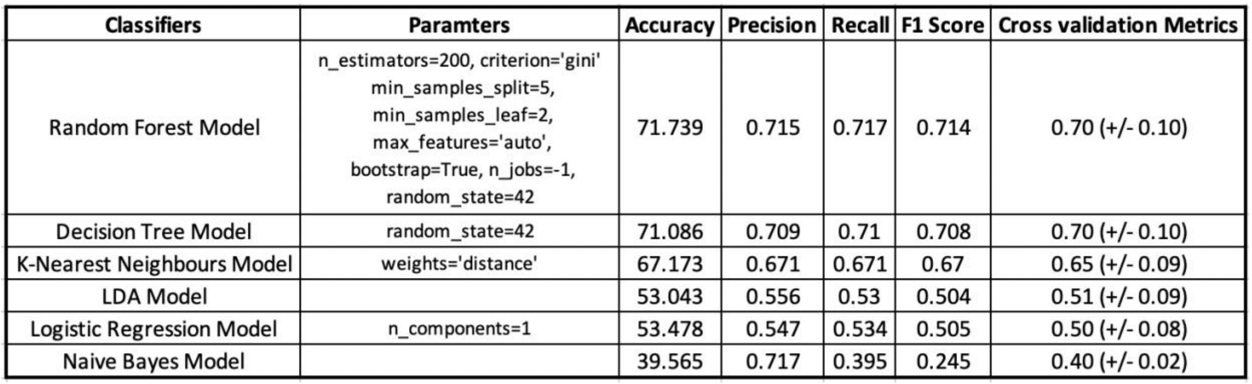
Predictors:

- season: A categorical value indicating the season in which the event occurred.
- weekday: A categorical variable that indicates which day of the week it is.
- holiday: A categorical value indicating whether a day is a national holiday or casual.
- month: A categorical variable that indicates which month it is.

Results

In order to evaluate the results obtained from classification models, the following classification evaluation metrics were applied: accuracy, precision, recall, F1 score, cross validation metrics, and confusion matrix. As shown in the following Table 6.3, the results are as follows:

`



Classifiers	Paramters	Accuracy	Precision	Recall	F1 Score	Cross validation Metrics
Random Forest Model	n_estimators=200, criterion='gini' min_samples_split=5, min_samples_leaf=2, max_features='auto', bootstrap=True, n_jobs=-1, random_state=42	71.739	0.715	0.717	0.714	0.70 (+/- 0.10)
Decision Tree Model	random_state=42	71.086	0.709	0.71	0.708	0.70 (+/- 0.10)
K-Nearest Neighbours Model	weights='distance'	67.173	0.671	0.671	0.67	0.65 (+/- 0.09)
LDA Model		53.043	0.556	0.53	0.504	0.51 (+/- 0.09)
Logistic Regression Model	n_components=1	53.478	0.547	0.534	0.505	0.50 (+/- 0.08)
Naive Bayes Model		39.565	0.717	0.395	0.245	0.40 (+/- 0.02)

Table 6.3 Comparison of Classification Models

Confusion Matrix

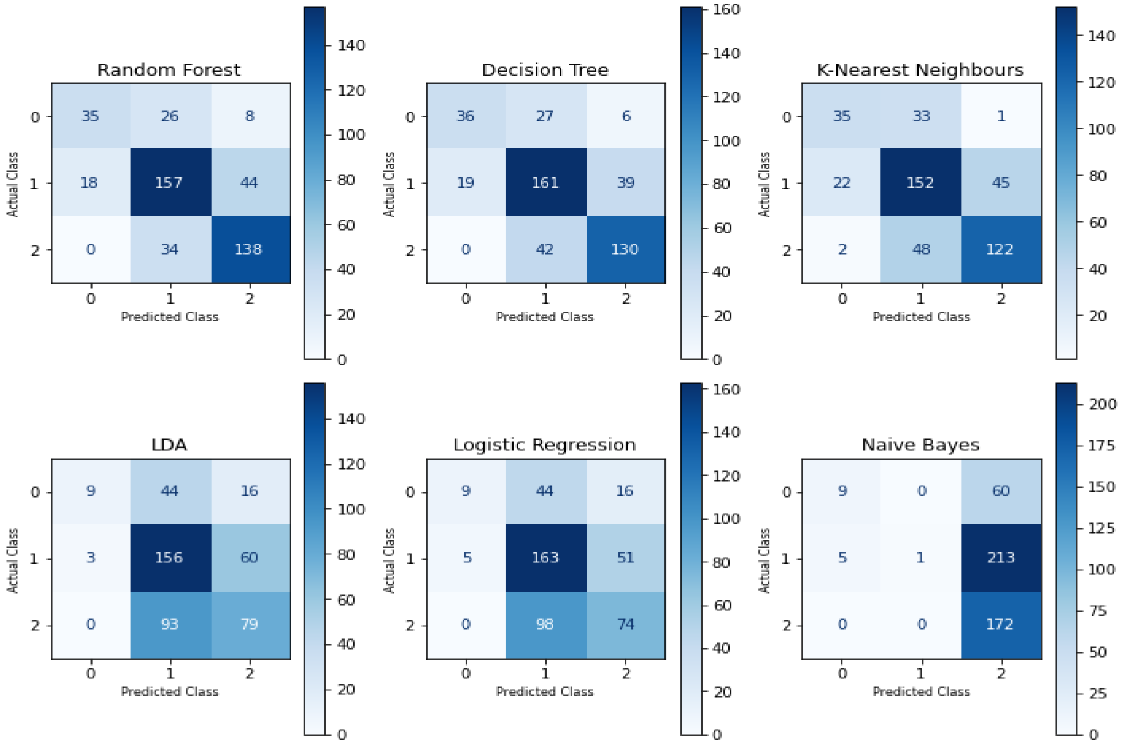


Figure 6.30 Confusion Matrix for Classification Models

From the Table 6.3, Random Forest and Decision Tree algorithms showed the most accurate findings, with Accuracy rate of 71.73% and 71.08% respectively, in categorizing a type of day into

`

low, medium, and high bike-sharing trips when compared to the metrics from each classification model. In contrast, the Naive Bayes method had the lowest level of accuracy when classifying. Based on the confusion matrix shown in Figure 6.30, the Random Forest method was able to categorize classes 1 and 2 properly, with 156 out of 219 instances correctly classified for class 1 and 138 out of 172 instances correctly identified for class 2. In spite of this, the accuracy of the model for class 0 is lower than that of other classes since the training and testing dataset only comprises a small number of class 0 elements. In contrary, the Naive Bayes model had a relatively low level of accuracy when it came to the classifying.

6.4 Discussion

Linear Regression's low accuracy indicates that our Dublin Bike dataset is non-linear, skewed, unbalanced, and includes outliers. However, repeatedly Random Forest performed well on our data both in regression as well as classification. Understanding the Random Forest algorithm will help us examine the findings and justify them based on the data and research we've conducted. Random Forest is one of the most powerful algorithm which works on the bagging technique. This algorithm selects observations and variables at random to create multiple decision trees, and the results from each decision tree are averaged to get a more accurate and robust prediction which makes the Random Forest algorithm robust to outliers, non-linear, skewed and unbalanced data. Also, Random forest algorithm has built-in methods that can automatically handle datasets of this kind which makes it less impacted by noise. These advantages of Random Forest make it robust to outliers, non-linear, skewed, and unbalanced data. Because of this, the Random Forest approach outperformed all other machine learning models on the Dublin Bikes dataset. However, one of the main drawbacks of the Random Forest method is that it requires more time and computational resources to build many trees, slowing down performance.

`

Chapter Seven : Conclusion and Future Work

7.1 Conclusion

The purpose of this study is to develop prediction algorithms for estimating bike-share journeys using Dublin Bikes station data. To accomplish this, Chardon's approach was used, and a novel strategy for calculating rebalancing was developed. Additionally, several machine learning algorithms were used to predict daily bike journeys based on different factors. Also, bike activity, bike availability, and bike-share trips between stations were extensively explored using clustering methods to unravel the hidden patterns.

Based on the results from each of Chardon's models, the individual day aggregated model, combined day aggregated model and, station aggregated model were promising as they satisfied all linear model assumptions, so we can rely on this model for estimating daily bike-share trips. Also, the Random Forest algorithm gave the most accurate results for both regression and classification compared to all other models. However, in real-world applications, classification models have proven to be effective as they enable the stakeholders to understand the results easily and draw conclusions.

Analysing and estimating daily bike trips can generate discussion and analysis about the efficacy of a BSS and whether it may be related to local legislation, cycling infrastructure, scheme pricing, station density, urban structure, or provisioner management. Furthermore, knowing the number of journeys that might occur on a particular day in advance can assist the organization in rebalancing the bikes between stations in order to meet demand and improve customer satisfaction.

`

7.2 Limitations

The dataset JCDecaux made accessible to the general public was restricted to have only availability of bikes/docks at various time zones. As a result, this data lacked essential information such as the origin and destination of the journey, information about the user's profile, and the subscriber information, all of which may have helped to unfold user behaviours and bike flow patterns. Due to this, we were not able to verify our estimated daily bike-share trips by models with actual values. Also, several data samples had to be dropped due to data entry issues in the dataset, and this, in turn, may have affected the research to some extent.

7.3 Future Work

Future work should focus on creating an algorithm with more tuning and statistical understanding to estimate rebalancing, which is crucial in the absence of cycle-level data to estimate the daily bike trips. Additionally, it would be beneficial to gather information about the actual bike journeys taken by a user from the organization and assess how well Chardon's models and the suggested machine learning performed in estimating the daily bike journeys. This research used traditional machine learning models to estimate daily bike trips. However, deep-learning models could be considered in the future to enhance accuracy. In addition, stationless bicycle sharing programs such as Urbo and Bleeper bike [43][44], have been made available in Dublin during the last few of years. These stationless bicycles, in contrast to Dublin Bikes, do not need riders to park them at predetermined locations, therefore they are more adaptable and require less maintenance. As a result, it would be beneficial to compare and contrast both stationed and stationless approaches.

`

References

[1] DeMaio, P., 2009. Bike-sharing: History, Impacts, Models of Provision, and Future. Journal of Public Transportation, 12(4), pp.41-56.

[2] ECF. 2020. *Bike Share Schemes (BSS)*. [online] Available at: <https://ecf.com/what-we-do/urban-mobility/bike-share-schemes-bss> [Accessed 7 August 2022].

[3] En.wikipedia.org. 2010. *Dublinbikes - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Dublinbikes> [Accessed 31 July 2022].

[4] Ryan, N., 2013. *First new Dublin Bikes stations launched, part of €35 million expansion*. [online] TheJournal.ie. Available at: <https://www.thejournal.ie/dublin-bike-scheme-expansion-stations- bicycles-1196211-Nov2013/> [Accessed 31 July 2022].

[5] Dublinbikes.ie. *DublinBikes*. [online] Available at: <https://www.dublinbikes.ie/en/tutorial> [Accessed 31 July 2022].

[6] Daly, M., 2011. *What's the secret of the Dublin bike hire scheme's success? | Maria Daly*. [online] the Guardian. Available at: <https://www.theguardian.com/environment/bike-blog/2011/aug/04/dublin-bike-hire-scheme> [Accessed 31 July 2022].

[7] DeMaio, P., 2003. [online] Metrobike.net. Available at: <https://www.metrobike.net/wp-content/uploads/2013/10/Smart-Bikes.pdf> [Accessed 31 July 2022].

[8] DeMaio, P. and Gifford, J., 2004. Will Smart Bikes Succeed as Public Transportation in the United States?. *Journal of Public Transportation*, 7(2), pp.1-15.

[9] Jon, F., Joachim, N. and Nuria, O., 2009. [online] Aiweb.cs.washington.edu. Available at: < http://aiweb.cs.washington.edu/research/projects/aiweb/media/papers/tmpxUhx0M.pdf> [Accessed 31 July 2022].

[10] Faghih-Imani, A., Hampshire, R., Marla, L. and Eluru, N., 2017. An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. Transportation Research Part A: Policy and Practice, [online] 97, pp.177-191. Available at: <https://www.sciencedirect.com/science/article/pii/S0965856416311648?via%3Dihub>.

[11] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J. and Banchs, R., 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4), pp.455-466.

[12] Zhou, X., 2015. Understanding Spatiotemporal Patterns of Biking Behavior by Analyzing Massive Bike Sharing Data in Chicago. *PLOS ONE*, [online] 10(10), p.e0137922. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4596835/>.

[13] Yang, Y., Heppenstall, A., Turner, A. and Comber, A., 2019. A spatiotemporal and graph- based analysis of dockless bike sharing patterns to understand urban flows over the last mile. *Computers, Environment and Urban Systems*, 77, p.101361.

[14] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J. and Banchs, R., 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4), pp.455-466.

[15] Calafiore, G., Portigliotti, F. and Rizzo, A., 2017. A Network Model for an Urban Bike Sharing System. *IFAC-PapersOnLine*, 50(1), pp.15633-15638.

[16] Li, Yexin & Zheng, Yu & Zhang, Huichu & Chen, Lei. (2015). Traffic prediction in a bike-sharing system. 1-10. 10.1145/2820783.2820837.

[17] Singhvi, Divya et al. “Predicting Bike Usage for New York City's Bike Sharing System.” *AAAI Workshop: Computational Sustainability* (2015).

`

[18] Vogel, P., Greiser, T. and Mattfeld, D., 2011. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, 20, pp.514-523.

[19] Huang, F., Qiao, S., Peng, J. and Guo, B., 2019. A Bimodal Gaussian Inhomogeneous Poisson Algorithm for Bike Number Prediction in a Bike-Sharing System. *IEEE Transactions on Intelligent Transportation Systems*, 20(8), pp.2848-2857.

[20] Mooney, Peter and Corcoran, Paul and Winstanley, Adam C. (2010) *Preliminary Results of a Spatial Analysis of Dublin City's Bike Rental Scheme.* In: GISRUK 2010: GIS Research UK 18th Annual Conference, 14-16 April 2010, London.

[21] Pham Thi, Thanh Thoa & Timoney, Joe & Ravichandran, Shyram & Mooney, Peter & Winstanley, A.. (2017). Bike Renting Data Analysis: The Case of Dublin City.

[22] Timoney, Joseph and Do Amaral, Carlos Siqueira and Thi, Thanh Thoa Pham and Winstanley, Adam C. (2018) *A continuation on the data analysis for the Dublin Bike rental scheme.* In: GISRUK 2018, 17-20 April 2018, University of Leicester.

[23] NIDHIN, G., 2018. Implementation and Comparison of Strategies to Predict the Availability ofDublin Bikes.

[24] Mishaal, I., 2020. Dublin Bikes Analysis.

[25] Médard de Chardon, C. and Caruso, G., 2015. Estimating bike-share trips using station level data. *Transportation Research Part B: Methodological*, 78, pp.260-279.

[26] James, T., Oliver, O. and James, C., 2019. Detecting Journeys in Bicycle Sharing Systems from Docking Station Counts. *GISRUK*,.

[27] Maxim, L., 2019. Estimating bike-share trips : Dublin City.

[28] Zach, 2020. *The Four Assumptions of Linear Regression - Statology*. [online] Statology. Available at: <https://www.statology.org/linear-regression-assumptions/> [Accessed 1 August 2022].

[29] Stephanie, G., 2018. *Ljung Box Test: Definition*. [online] Statistics How To. Available at: <https://www.statisticshowto.com/ljung-box-test/> [Accessed 1 August 2022].

[30] Brownlee, J., 2016. *Regression Metrics for Machine Learning*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/regression-metrics-for-machine-learning/> [Accessed 1 August 2022].

[31] C3 AI. 2022. *Root Mean Square Error (RMSE) - C3 AI*. [online] Available at: <https://c3.ai/glossary/data-science/root-mean-square-error-rmse/> [Accessed 7 August 2022].

[32] Jedox. 2022. *Error Metrics: How to Evaluate Your Forecasting Models*. [online] Available at: <https://www.jedox.com/en/blog/error-metrics-how-to-evaluate-forecasts/> [Accessed 1 August 2022].

[33] Jachner, S., Gerald van den Boogaart, K. ., & Petzoldt, T. (2007). Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV). *Journal of Statistical Software*, *22*(8), 1–30. https://doi.org/10.18637/jss.v022.i08

[34] Sunasra, M., 2017. *Performance Metrics for Classification problems in Machine Learning*. [online] Medium. Available at: <https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b> [Accessed 1 August 2022].

[35] Fawzy Gad, A., 2020. *Accuracy, Precision, and Recall in Deep Learning | Paperspace Blog*. [online] Paperspace Blog. Available at: <https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/> [Accessed 1 August 2022].

[36] Shivam, K., 2019. *Understanding a Classification Report For Your Machine Learning Model*. [online] Medium. Available at: <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397> [Accessed 1 August 2022].

`

[37] Data.gov.ie. 2022. *Dublin Airport Hourly Data - Dublin_Airport Hourly Data - data.gov.ie*. [online] Available at: <https://data.gov.ie/dataset/dublin-airport-hourly-data/resource/a818dbd2-c82a-4514-8ff3-62a63c0b9ba3> [Accessed 1 August 2022].

[38] Met.ie. 2022. *Beaufort Scale - Met Éireann - The Irish Meteorological Service*. [online] Available at: <https://www.met.ie/forecasts/marine-inland-lakes/beaufort-scale> [Accessed 1 August 2022].

[39] IrishCycle.com. 2015. *DublinBikes facts and figures -- IrishCycle.com*. [online] Available at: <https://irishcycle.com/dublinbikes/> [Accessed 1 August 2022].

[40] Dublin City Council. 2022. *Clarendon Row Improvement Scheme*. [online] Available at: <https://www.dublincity.ie/residential/transportation/roads-and-traffic-projects/clarendon-row-improvement-scheme> [Accessed 1 August 2022].

[41] Logmytree.blogspot.com. 2012. *Dublin Urban Trees*. [online] Available at: <https://logmytree.blogspot.com/> [Accessed 5 August 2022].

[42] Ireland.com. 2022. *Public holidays in Ireland | Ireland.com*. [online] Available at: <https://www.ireland.com/en-se/help-and-advice/practical-information/public-holidays/> [Accessed 1 August 2022].

[43] Duffy, R., 2018. *Stationless bike hire scheme launched in Dublin with 200 bikes hitting the streets right away*. [online] TheJournal.ie. Available at: <https://www.thejournal.ie/dcc-stationless-bikes-4044090-May2018/> [Accessed 1 August 2022].

[44] McGowran, L., 2018. *New stationless bike scheme launched in Dublin today*. [online] Green News Ireland. Available at: <https://greennews.ie/stationless-bike-scheme-launched-in-dublin-today/> [Accessed 1 August 2022].

[45] Bikesharingworldmap.com. 2021. *The Meddin Bike-sharing World Map*. [online] Available at: <https://bikesharingworldmap.com/> [Accessed 1 August 2022].

[46] Dublinbikes.ie. 2022. *DublinBikes*. [online] Available at: <https://www.dublinbikes.ie/en/tutorial/groups?tab=RIDE> [Accessed 1 August 2022].

[47] Statistics.laerd.com. 2018. *Linear Regression Analysis in SPSS Statistics - Procedure, assumptions and reporting the output.*. [online] Available at: <https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php> [Accessed 1 August 2022].

[48] En.wikipedia.org. 2006. *Ljung–Box test - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test> [Accessed 1 August 2022].

[49] Kim, B., 2015. *Should I always transform my variables to make them normal? | University of Virginia Library Research Data Services + Sciences*. [online] Data.library.virginia.edu. Available at: <https://data.library.virginia.edu/normality-assumption/> [Accessed 1 August 2022].

[50] Kim, B., 2015. *Understanding Diagnostic Plots for Linear Regression Analysis | University of Virginia Library Research Data Services + Sciences*. [online] Data.library.virginia.edu. Available at: <https://data.library.virginia.edu/diagnostic-plots/> [Accessed 1 August 2022].

[51] Pascual, C., 2018. *Understanding Regression Error Metrics*. [online] SunJackson Blog. Available at: <https://sunjackson.github.io/2018/09/26/2cf12da6359138289cad4abcb69a7612/> [Accessed 1 August 2022].

[52] Müller, F., 2020. *Measuring Regression Errors with Python*. [online] relataly.com. Available at: <https://www.relataly.com/regression-error-metrics-python/923/> [Accessed 1 August 2022].

[53] Jedox. 2022. *Error Metrics: How to Evaluate Your Forecasting Models*. [online] Available at: <https://www.jedox.com/en/blog/error-metrics-how-to-evaluate-forecasts/> [Accessed 1 August 2022].

[54] Padhma, M., 2021. *Evaluation Metric for Regression Models - Analytics Vidhya*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/> [Accessed 1 August 2022].

`

[55] Srivastava, T., 2019. *Evaluation Metrics Machine Learning*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/> [Accessed 1 August 2022].

[56] Twitter. 2022. *Dublin City Council*. [online] Available at: <https://twitter.com/dubcitycouncil/status/1507404747979980801> [Accessed 1 August 2022].

[57] Investopedia. 2021. *Understanding the Durbin Watson Statistic*. [online] Available at: <https://www.investopedia.com/terms/d/durbin-watson-statistic.asp#:~:text=The%20Durbin%20Watson%20statistic%20is,above%202.0%20indicates%20negative%20autocorrelation.> [Accessed 1 August 2022].

[58] Breslin, R., 2020. *What Dublin Bikes data can tell us about the city and its people*. [online] Medium. Available at: <https://towardsdatascience.com/what-dublin-bikes-data-can-tell-us-about-the-city-and-its-people-63fde77ee383#:~:text=The%20%E2%80%9CRed%20Cluster%E2%80%9D%20and%20%E2%80%9C,areas%20with%20Bicycle%20Stations%20present.> [Accessed 4 August 2022].

[59] GeeksforGeeks. 2022. *Elbow Method for optimal value of k in KMeans - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/> [Accessed 5 August 2022].

[60] Zach, 2019. *Understanding Heteroscedasticity in Regression Analysis - Statology*. [online] Statology. Available at: <https://www.statology.org/heteroscedasticity-regression/> [Accessed 6 August 2022].

Appendices

GitHub Source Code Repository Link

https://github.com/sreevathsadb/Dublin-Bikes-Analysis

`