A picture containing icon

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for Data Analytics*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *A Comparative Analysis of Bike Sharing in Dublin and Washington D.C* |
| **Lecturer Name:** | *Marina Iantorno/Taufique Ahmed*  *Sam Weiss/David Gonzalez*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Jose Mario da Cruz Costa |
| **Student Number:** | 2023422 |
| **Assessment Due Date:** | *5th Jan 2024* |
| **Date of Submission:** | *3rd Jan 2024* |

**Declaration**

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

***A Comparative Analysis of Bike Sharing in Dublin and Washington D.C***

# Table of Contents

Table of Contents 2

List of figures 3

List of Tables 3

Abstract 4

Introduction 5

Project Description 5

1 - Data Preparation and Visualization 7

1.1 Process of acquiring raw data: 7

1.2 Exploratory Data Analysis (EDA) Method and Insights 8

2.0 Statistics for Data Analytics 15

2.1 Summarizing the datasets using descriptive statistics and appropriate visualisations 15

2.2 Analysing the variables using appropriate inferential statistics 18

2.3 Undertaking research and finding similarities between Dublin and Washington D.C. 18

2.4 Challenges faced in the process: 20

3.0 Machine Learning for Data Analysis 21

3.1 Time Series Analysis Over Dublin bike share system 21

3.2 Sentimental Analysis 24

4.0 Programming for Data Analysis 26

4.1 Programming: 26

4.2 Data Structures: 26

4.3 Documentation: 26

4.4 Testing & Optimisation: 27

4.5 Data Manipulation: 28

5.0 Bibliography 29

# List of figures

[Figure 1 - CRISP-DM Workflow 5](#_Toc155196883)

[Figure 2 - Project Management - Gantt Chart 6](#_Toc155196884)

[Figure 3 - Map of Bike Stations in Dublin 8](#_Toc155196885)

[Figure 4 - Map of Bike Stations in Washington D.C 9](#_Toc155196886)

[Figure 5 - Comparing Top 10 trips - Dublin and Washington DC 10](#_Toc155196887)

[Figure 6 - Comparing trips by weekday in both cities. 10](#_Toc155196888)

[Figure 7 - Comparing trips by month in both cities 11](#_Toc155196889)

[Figure 8 - Distribution of reviews lengths 11](#_Toc155196890)

[Figure 9 - Polarity scatter plot 12](#_Toc155196891)

[Figure 10 - TextBlob Sentiment Result 12](#_Toc155196892)

[Figure 11 - Plotly Dashboard 13](#_Toc155196893)

[Figure 12 - Distribution of the dataset: “bike\_usage” 14](#_Toc155196894)

[Figure 13 - Distribution and Density of Trips grouped by station on Both Datasets 15](#_Toc155196895)

[Figure 14 - Correlation between features in Dublin dataset 16](#_Toc155196896)

[Figure 15 - Frequency distribution of WeekDay 16](#_Toc155196897)

[Figure 16 - Proportions and confidence Intervals 17](#_Toc155196898)

[Figure 17 - Sampling and ECDFs – Empirical Cumulative Distribution Function 18](#_Toc155196899)

[Figure 18 - RandomForestRegressor over the entire timestamp 20](#_Toc155196900)

[Figure 19- Aplying RandomForesRegressor over the Hour of the day 21](#_Toc155196901)

[Figure 20 - RandomForestRegressor over the day of the week and the hour of the day 21](#_Toc155196902)

[Figure 21 - RandomForestRegressor after Hyperparameter tuning: 22](#_Toc155196903)

[Figure 22 - Linear Regression over the day of the week and the hour of the day 22](#_Toc155196904)

[Figure 23 - Ridge regression with polynomial features 23](#_Toc155196905)

[Figure 24 - distribution of sentment Dublin and Capital Bikeshare 24](#_Toc155196906)

[Figure 25 - Accuracy of TF-IDF and Bag of Words Vectorization 25](#_Toc155196907)

[Figure 26 - Code with documentation TripAdvisor 26](#_Toc155196908)

[Figure 27 - code with explanation Reddit 26](#_Toc155196909)

[Figure 28 - Testing Functions to verify it file exists. 27](#_Toc155196910)

# List of Tables

[Table 1 - Comparing Both Bikeshare Systems 9](#_Toc155196911)

[Table 2 - RandomForesRegressor Results 21](#_Toc155196912)

[Table 3 - Comparation of models and Techiniques 23](#_Toc155196913)

[Table 4 - Comparing Libraries and techniques 28](#_Toc155196914)

# Abstract

***A Comparative Analysis of Bike Sharing in Dublin and Washington D.C***

*This project aims to analyse and predict bike service trends in Dublin and Washington D.C., it employs CRISP-DM framework and the project will dive into four primary datasets.*

*The data was acquired from trusted sources like data.gov.ie and Capital Bikeshare's official website, and reviews were sourced from TripAdvisor and Reddit through APIs.*

*The datasets were cleaned and transformed to produce notable findings, which includes distribution of bike stations, statistical analyses and machine learning models.*

*This project is available on GitHub through the link:*

*https://github.com/jmdtanalyst/MSC\_DA\_CA2\_Transport\_Ireland*

***Keywords:*** *population, transport, public, bike-sharing*

***Word count***: 2.843 words (excluding code, code comments, titles, references and citations)

# Introduction

In recent years, bike sharing systems have emerged as a popular and sustainable mode of transportation in urban areas.

One of the largest bike-sharing systems in Dublin is DublinBikes, It is the oldest and the biggest system. It is operated by a subsidiary of the National Transport Authority (NTA).

With a network of more than 100 docking stations located across key areas of Dublin, DublinBikes has a fleet of 3.647 bikes.

In other hand, the largest bike-sharing service in Washington D.C. is Capital Bikeshare, with a fleet of over 7,000 and 759 bike stations.

This project will compare the usage of Dublin bikeshare systems against Capital bikeshare services.

## Project Description

For this project, CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was applied.

CRISP-DM framework was chosen because it provides a structured approach to ensure that all of the important aspects of the project are considered.

### CRISP-DM framework steps:

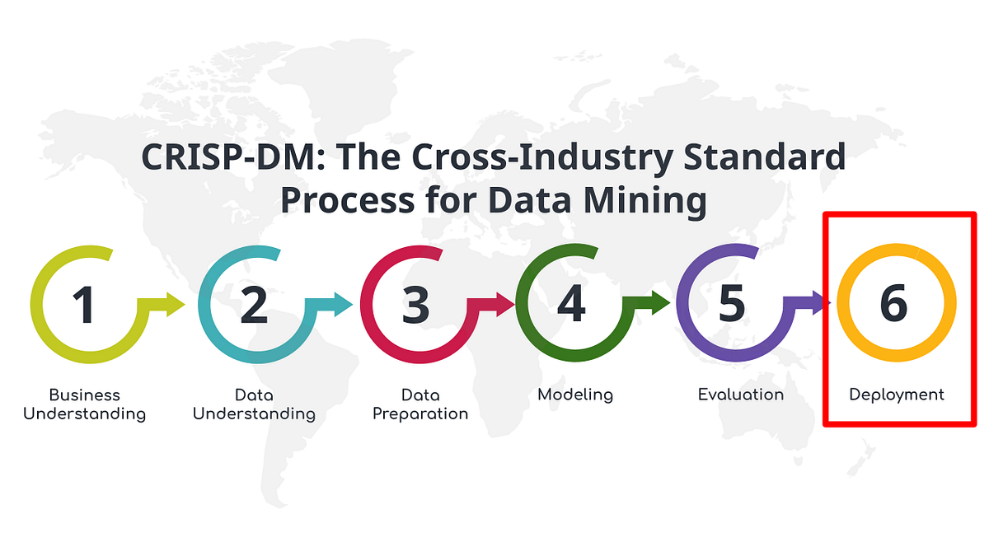


Figure 1 - CRISP-DM Workflow

**1 - Business Understanding**

The business objective Is analysing and understanding people's experiences with bike-sharing services in Dublin and Washington DC, it will dive into patterns, issue and also into sentiment analysis.

**2 - Data Understanding**

The analysis will focus on four datasets: two datasets with historical bike trips, for calculations and predictions of volume of trips, and two datasets with reviews collected from TripAdvisor and Reddit, for sentiment analysis.

**3 - Data Preparation**

The steps of Data Understanding and Data Preparation were performed in Data Preparation and Visualization and Statistics phases**.**

**4 - Modeling**

For predicting the number of bikes trips per day, a time-series analysis was performed applying three Machine Learning Algorithms: RandomForestRegressor, Linear Regression and Ridge Regression, also GridSearchCV was applied to select the best hyperparameters and Sentiment analysis was performed.

**5 – Evaluation**

The model's efficacy was evaluated using key metrics as the Mean Squared Error (MSE) and R^2 score.

**6 – Deployment**

After a satisfactory model performance, the model deployed and saved under the name 'random\_forest\_regressor.pkl'. This allows for easy loading and reuse.

For project management, a Gantt Chart was created on Jupyter Notebook utilizing Plotly Express library to track the project overflow:

A graph of a chart

Description automatically generated with medium confidence

Figure 2 - Project Management - Gantt Chart

# 1 - Data Preparation and Visualization

## 1.1 Process of acquiring raw data:

Data Preparation consists of collecting, cleaning, transforming, and manipulating raw data to make it usable for analysis.

### Raw Data Acquisition:

* **DublinBikes :**

Datasets were collected on data.gov.ie, It was downloaded in CSV file, under the “Creative Commons Attribution 4.0 (CC BY 4.0)” Licence.

**Advantage**:

The data is easy to be collected and this licence allows to copy and use the data for free.

**Disadvantage**:

The data isn’t automatically updated, requiring manual download and analysis.

* **Capital Bikeshare:**

Data sourced from the company's official website, https://ride.capitalbikeshare.com/system-data, under its own Licence. Similar to DublinBikes, the data does not auto-update, requiring manual interventions for updates.

* **Reviews:**

Data were sourced from TripAdvisor and Reddit through APIs.

**Advantages:**

Regular data updates.

**Disadvantages:**

Security of the API\_KEY become crucial aspects to consider.

TripAdvisor and Reddit Terms of Service permit API usage for non-commercial purposes.

**Cleaning and Engineering Process**

Taking into consideration the machine learning, the datasets have been cleaned, duplicated rows and null values have been addressed, engineering was applied on the “TIME” and “Trips” columns.

Given that the dataset is large, a statistical technique called sampling was applied. It is a powerful tool for data analysis that offers several advantages:

* Reduces Data Analysis Time and Cost and Data Storage
* Improves Data Analysis Efficiency:

## 1.2 Exploratory Data Analysis (EDA) Method and Insights

Considering Tufte's principles of data visualization, a scatter Map was plotted, making easy to understand the pattern of bike station around the city.

For Dublin map, the size of the bubble is proportional to the number of bikes available, it will provide a clear bike availability at each station, also, the colour of the bubble was chosen to represent the open and close station.

**Distribution of Bike Station in over the city and comparing both maps.**

A map with green dots and red dots

Description automatically generated

Figure 3 - Map of Bike Stations in Dublin

From the map, it is observed that Dublin bike stations are spread out across the city centre, whereas in the suburbs there are less stations.

Also, the map shows that there are 2 Bikes stations off-line (the red ones)

A map with blue dots

Description automatically generated

Figure 4 - Map of Bike Stations in Washington D.C

Capital Bikeshare has a huge Stations Network across the city, less bike stations in the suburbs, but there are still a number of stations located in key areas.

Table 1 - Comparing Both Bikeshare Systems

|  |  |  |
| --- | --- | --- |
| Feature | Dublin Bikes | Capital Bikeshare |
| Station density | Moderate | High |
| Distribution | spread across the city center | Concentrated in the city center and surrounding |
| Proximity to tourist areas | Good | Excellent |
| Coverage in suburbs | Limited | Moderate |

### Interactive Visualization

It is well-known that modern transport planning relies heavily on visualizations. To address this need while adhering to the principles of Edward Tufte, renowned for his emphasis on clarity and precision in data visualization, we employed Plotly Express and Dash to create an interactive dashboard.

Plotly Express was chosen because it is a powerful library able to generate insightful visualizations, aligned to Dash, a web application framework, its possible to create dynamic dashboard where users can interactively explore the data.

**Top 10 Bike Usage by Station**

A close-up of a graph

Description automatically generated

Figure 5 - Comparing Top 10 trips - Dublin and Washington DC

**Insights:**

The chart shows that Dublinbikes is more popular in the city centre. The busiest stations located in the city centre. Also, some stations with less bike stands have more demand than the number bike stands available.

For Capital Bikeshare, all of the top 10 stations are located in downtown, residential and commercial areas.

**Trips by WeekDay**

A screenshot of a graph

Description automatically generated

Figure 6 - Comparing trips by weekday in both cities.

The chart for Dublin shows that bike usage is highest on Tuesday, Wednesday and Thursday with a considerable decrease at weekend, whereas in Washinbgton D.C the trips are high all days, with a slow decrease on monday.

**Trips by Month**

A graph of different colored bars

Description automatically generated with medium confidence

Figure 7 - Comparing trips by month in both cities

For Dublinbikes, the usage is highest in May and June but for all month, the usage is consistent, between 80k to 100k, in Washington, D.C. there is a big difference between the lowest and the highest month (January and October)

### Visualization for Sentiment Analysis

**Distribution of reviews lengths:**

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 8 - Distribution of reviews lengths

There are more longer reviews for Dublin bikeshare service than for the Washington, D.C. bikeshare service.

**Sentiment analysis with TextBlob:**

A screenshot of a graph

Description automatically generated

Figure 9 - Polarity scatter plot

This Polarity Scatter shows that Dublin has a higher overall polarity than Washington DC. i.e comments from Dublin Bikes are more positive or negative on average.

Also, there are more positive sentiment for Dublin whereas, for Washington DC there are more neutral sentiment.

### Comparing the Sentiment in both bike Systems.

A comparison of a bar graph

Description automatically generated

Figure 10 - TextBlob Sentiment Result

Dublin and Washington D.C have positive sentiments. However, for Capital Bikeshare, there are more neutral than positive sentiment.

These results were made using the TextBlob library, which is a simple sentiment analysis library. There are more accurate methods in machine learning, which will be covered in Machine Learning Session.

### Dashboard

For dynamic visualization, a dashboard was created. It ca be accessed through the address: <http://127.0.0.1:8050> , and its possible view the chart based on topic selected.

A screenshot of a computer screen

Description automatically generated

Figure 11 - Plotly Dashboard

# 2.0 Statistics for Data Analytics

## 2.1 Summarizing the datasets using descriptive statistics and appropriate visualisations

**Descriptive statistics**:

Through Descriptive statistics, it’s possible to identify patterns, trends, and relationships between the data. It is divided **measures of central tendency** and **measures of variability.**

Measures of central tendency describe the centre of the data set whereas measure of dispersion describes how the data is spread out.

In the figure bellow, its plotted the Distribution of the dataset: “bike\_usage”, it represents the bike usage in Dublin Bike System.

### Distribution of the dataset: “bike\_usage”

A group of graphs showing different types of bikes

Description automatically generated

Figure 12 - Distribution of the dataset: “bike\_usage”

**Insights from this figure:**

* **Distribution of Bike Stands by Density**

The chart shows that the majority of bike stands have a density of between 10 and 25 per unit area.

However, there is a small number of stations with a significantly higher or lower number of bike stands than the average.

This suggests that there may be a need to redistribute bike stands to ensure that all stations have a similar number of bike stands available.

* **Distribution of Bikes in Use**

From the Distribution of Bikes in Use, with a mean of 20.65 and a standard deviation of 10.33, its possible assume that the majority of bikes in use have a usage of between 10.33 and 30.97.

This suggests that bikes are typically used for a variety of purposes, from short commutes to longer recreational rides.

* **Distribution of Available Bikes**

The majority of available bikes have a density of between 10 and 20. Also, there are a smaller number of available bikes with densities below 10 or above 20.

That means, the available bikes are typically concentrated in certain areas, rather than being evenly distributed throughout the area.

### Distribution and Density of Trips grouped by station on Both Datasets

A graph of a number of points

Description automatically generated with medium confidence

Figure 13 - Distribution and Density of Trips grouped by station on Both Datasets

**Insights:**

Its possible to see that the distribution is positively skewed, showing that there are a few stations with a very high number of trips and others stations with a lower number of trips.

Also, the mean number of trips per station is higher for Capital DF (5240) than for Dublin Bikes (9197), this indicates that the average station in Capital DF has more trips than Dublin Station.

### Correlation between features in Dublin bike\_share dataframe:

A graph of a bike stand

Description automatically generated with medium confidence

Figure 14 - Correlation between features in Dublin dataset

**Insights**

There is a strong negative correlation between BIKE\_STANDS and BIKES\_IN\_USE. This means that as the number of bike stands at a station increases, the number of bikes in use at that station decreases.

There is a moderate negative correlation between AVAILABLE\_BIKES and TRIPS and a very weak correlation between MONTH and TRIPS. This suggests that the number of bike trips does not vary significantly throughout the year.

### Frequency distribution of WeekDay - Comparing Dublin to Washington DC

A comparison of blue and white bars

Description automatically generated

Figure 15 - Frequency distribution of WeekDay

**Insights:**

The bike usage in Dublin peaks on Thursdays, with a dip on Sundays, whereas n Capital DF, Thursday is also the most popular weekday, but Monday is the least popular, highlighting that there is a different behavior in each city.

## 2.2 Analysing the variables using appropriate inferential statistics

By applying inferential statistics, it is possible to find a confidence interval for the population proportion of trips per month. The proportion\_confint() function is employed to calculate these confidence intervals, taking the number of bike trips in each month, the total number of bike trips across all months, and the significance level (0.05 for a 95% confidence interval)

A graph of a line and a line

Description automatically generated

Figure 16 - Proportions and confidence Intervals

**Insights**:

The proportion of bike trips in Dublin varies throughout the year, with the highest proportion of trips occurring in June, July, August.The proportion of bike trips in Dublin has been increasing over time, while the proportion of bike trips in Capidal\_DF has remained stable.

## 2.3 Undertaking research and finding similarities between Dublin and Washington D.C.

**Applying apply parametric and non-parametric inferential statistical techniques :**

**Sampling**

Sampling data is a fundamental practice in statistics and data analysis. It offers several advantages in data analysis and data science research. By executing properly sampling techniques, it can offering a balance between accuracy, efficiency, and practicality.

For this project, a sample of 5% was applied, which didn’t get a good result on Evaluation (applying ECDFs – Empirical Cumulative Distribution Function). Furthemore, a 20% sampling was applied, wich ended in a good result.

A screenshot of a graph

Description automatically generated

Figure 17 - Sampling and ECDFs – Empirical Cumulative Distribution Function

**Parametric Tests, non-Parametric Test (Wilcoxon) and Chi-Squared Test:**

**Hypothesis:**

The hypothesis was assumed as follows: "Dublin exhibits a higher proportion of bike trips during daytime hours compared to Washington D.C." This assumption was tested using statistical methodologies.

* Parametric tests: assumes that the data has a normal distribution, it provided initial insights into the difference between the two datasets. These tests resulted in a significant t-statistic (-6.5855) and a minuscule p-value (2.62e-05), which indicates the rejection of the null hypothesis.
* Non-Parametric Tests (Wilcoxon): Also, a non-parametric test was applied. The Wilcoxon test, resulted similar conclusion.
* Chi-Squared Test: The chi-squared test was and also confirmed the presence of a significant association between daytime usage and the two cities. Whti a chi-squared value of 2,063.23 and a p-value of 2.2e-161.

**Conclusion:**

After applying the test, With the values:

Total trips in Dublin during daytime: 154411.0

Total trips in Washington D.C. during daytime: 542630

T-statistic: -6.585538607430632

P-value: 2.6198294268193643e-05

The hypothesis was rejected.

“Reject the null hypothesis: Dublin has more bike usage during daytime than Washington D.C.”

**Application of Type I and Type II Errors on this Hypothesis**

For statistical hypothesis testing, understanding potential errors is important. Specifically when comparing the means of two samples, two types of errors can occur: Type I (False Positive) and Type II (False Negative) errors.

**Type I Error:**

A Type I error occurs when the null hypothesis is rejected incorrectly.

**Type II Error:**

A Type II error happens when the null hypothesis is not rejected, when it should be accepted.

To mitigate Type I and Type II errors, t-test was applied, to support the means of daytime bike trips between the two cities based on the p-value result.

## 2.4 Challenges faced in the process:

For analysing and comparing bike usage data from Dublin and Washington D.C. the following challenges were faced:

* Data Cleaning and Preprocessing: Dealing with missing data and bad quality.
* Handling large volume and complexity of the dataset
* Statistical Analysis and Inferential Modelling: Sampling:
* Finding the right size of the sample and applying the appropriate test.
* Hypothesis Testing: Formulating the appropriate hypotheses.
* Understanding the significance level and p-values was important to avoid making false conclusions.

# 3.0 Machine Learning for Data Analysis

Considering that predicting the number of trips for bikeshare services is a classic time series forecasting problem. Time series analysis was applied, it is an approach that allows to identify patterns, trends and utilize historical data to forecast future values.

For this purpose, the dataset was filtered and applied in a month (November), also series object with time(datetime) as index and sum of trips was created.

## 3.1 Time Series Analysis Over Dublin bike share system

### RandomForestRegressor

Initially, the RandomForestRegressor algorithm was employed across various timestamps to evaluate its predictive performance.

A graph with green and orange lines

Description automatically generated

Figure 18 - RandomForestRegressor over the entire timestamp

After Applying this model over the entire timestamp, the result was: R-squared value of -0.15 shows that its is not a good model, the model performed poorly and it did not explain the variance in the target variable. A MSE of 2395.29 ondicates that,the difference between the predicted number of trips and the actual number of trips is high, it shows a larger discrepancy between predicted and actual values.

A graph of a graph

Description automatically generated with medium confidence

Figure 19- Aplying RandomForesRegressor over the Hour of the day

By using the hour of the day as a feature, the RandomForestRegressor model achieved a moderate fit (R^2 of 0.52) and an average prediction error (MSE of 1010.20). which yet is a poor result.

A graph with different colored lines

Description automatically generated

Figure 20 - RandomForestRegressor over the day of the week and the hour of the day

Table 2 - RandomForesRegressor Results

|  |  |
| --- | --- |
| Key Features & Techniques | Result |
| over the entire timestamp | Test-set R^2: -0.20  Test-set MSE: 2683.33 |
| over the Hour of the day | Test-set R^2: 0.53  Test-set MSE: 1060.65 |
| over the day of the week and the hour of the day | Test-set R^2: 0.81  Test-set MSE: 433.23 |

As a result, the best performance was on test: “RandomForestRegressor over the day of the week and the hour of the day” which had a result of:

Test-set R^2: 0.81

Test-set MSE: 433.23

The model applied on the entire timestamp had a negative R^2, that means, it performed worse than a horizontal line. in other hand, focusing on the hour of the day improved the performance, but not enough to consider a great result. whare as, incorporating both the day of the week and hour of the day, got the best result. This means that considering both temporal factors, a more accurate result was reached.

### Hyperparameter Tuning

After applying hyperparameter tuning, the model yielded the following optimal parameters:

*Maximum Depth (max\_depth): 10*

*Minimum Samples per Leaf (min\_samples\_leaf): 4*

*Minimum Samples for Splitting (min\_samples\_split): 2*

*Number of Estimators (n\_estimators): 150*

After getting the best result from RandomForestRegressor, the model was compared to others models:

RandomForestRegressor after Hyperparameter tuning:

Figure 21 - RandomForestRegressor after Hyperparameter tuning:

A graph with numbers and lines

Description automatically generated

### Linear Regression

Figure 22 - Linear Regression over the day of the week and the hour of the day

A graph with blue and orange lines

Description automatically generated

### Ridge Regression with polynomial features

Figure 23 - Ridge regression with polynomial features

A graph of a graph

Description automatically generated with medium confidence

### Evaluating the results

**Below, a table comparing the results:**

|  |  |  |
| --- | --- | --- |
| Model Type | Key Features & Techniques | Result |
| RandomForestRegressor | Hyperparameter tuning **-** GridSearchCV | Test-set R^2: 0.81  Test-set MSE: 425.70 |
| Linear Regression | day of the week andhour of the day | Test-set R^2: 0.08  Test-set MSE: 2071.39 |
| Ridge Regression | One-hot encoding and polynomial features | Test-set R^2: 0.79  Test-set MSE: 463.64 |

Table 3 - Comparation of models and Techiniques

**Conclusion:**

From these 3 models, its possible to observe that RandomForestRegressor model performed better. After applying hyperparameter tuning using GridSearchCV, the result was even better, The Ridge regression with one-hot encoding and polynomial features, performed similar, but slightly behind the RandomForestRegressor, whereas the LinearRegression model performed worse.

## 3.2 Sentimental Analysis

For Sentiment Analysis, the first technique applied was VADER (Valence Aware Dictionary and Sentiment Reasoner), its was chosen because it is specifically designed for analysing sentiments in social media text and has been trained on such data. Also, two machine learning models were used to improve the accuracy of the sentiment classification: Bag of Words (BoW) and TF-IDF Vectorization

### VADER (Valence Aware Dictionary and Sentiment Reasoner) - shows that the sentiment for Dublin Bikeshare is positive and Capital Bikeshare has more neutral reviews.

A comparison of a bar graph

Description automatically generated

Figure 24 - distribution of sentment Dublin and Capital Bikeshare

**Conclusion**

This analysis shows that Dublin Bikeshare is generally viewed in more positive way compared to Capital Bikeshare.

### Machine Learning Models: Bag of Words (BoW) and TF-IDF Vectorization:

Bag of Words (BoW): The accurancy of the model comparing to VADER was 0.6986 for Dublin Bikeshare reviews and 0.5082 for Capital Bikeshare reviews.

TF-IDF Vectorization: The accurancy of the model was 0.6712 for Dublin Bikeshare reviews and 0.5082 for Capital Bikeshare reviews.

A graph with different colored squares

Description automatically generated

Figure 25 - Accuracy of TF-IDF and Bag of Words Vectorization

The Bag of Words models performed better for Dublin Bikeshare reviews. However, the performance of both models was similar for Capital Bikeshare reviews.

# 4.0 Programming for Data Analysis

## **4.1 Programming**:

Through the development of this project, Imperative programming, functional programming and object oriented programming has been applied.

Imperative programming was applied in tasks such: reading and writing data to and from CSV files, Performing data cleaning and preprocessing operations, whereas Functional programming was applied to tasks such: filtering and manipulating dataframes using pandas and creating and applying custom functions, for example, test\_file\_dublin() and test\_file\_capital.

The code was implemented in Jupyter Notebook, the following libraries and tools were implemented: Pandas, NumPy, Matplotlib, Seaborn, SciPy, Scikit-learn, NLTK, Sklearn's TfidfVectorizer & CountVectorizer, GridSearchCV and Joblib:

## **4.2 Data Structures:**

The data collection process was made in two primary formats: CSV files, sourced directly from reputable and trusted platforms: data.gov.ie and Capital Bikeshare's official website, and reviews dataframes were sourced from TripAdvisor and Reddit through APIs, in JSON format.

All data were sourced under its respective Licence.

## **4.3 Documentation:**

All documentation was made in the code, and also explaining why use the specific library, as shown in bellow.

A screenshot of a computer

Description automatically generated

Figure 26 - Code with documentation TripAdvisor

A screenshot of a computer

Description automatically generated

Figure 27 - code with explanation Reddit

## **4.4 Testing & Optimisation:**

To ensure the reliability of the several testing mechanisms were imployed. Firstly, test functions were created in order to avoid errors for missing files and incorrect format and, try catch also was implemented in order to note break the operation.

Furthermore, the integration of the 'try-catch' paradigm was also employed to avoid disruptions.

Optimization was addressed by using vectorized operations and concatenating multiple smaller dataframes.

Cartopy and Plotly Express are two popular options to create Maps, However Plotly Express is considered faster due to its vectorized operations and high-performance backend. Therefore, Plotly Express was chosen to create the maps.

**Testing from the code:**

A screenshot of a computer program

Description automatically generated

Figure 28 - Testing Functions to verify it file exists.

## **4.5 Data Manipulation**:

Addressing this question, a table was create comparing the libraries.

|  |  |  |
| --- | --- | --- |
| **Dublin Bikes and Capital Bikeshare datasets:** | | |
| Libraries for Processing | Techniques for Aggregating | Justification |
| Pandas  Numpy | sum()  groupby() | Chosen: **Both.**  Pandas stands as a cornerstone in data manipulation, renowned for its extensive functionalities. With features like to\_datetime() for seamless date-time conversions and mean() for precise column averaging, it offers a holistic suite for diverse data operations. Concurrently, NumPy's mathematical prowess complements Pandas, providing essential functions that underpin the data analysis tasks in this project. |
| **TripAdvisor API** | | |
| Requests  Beautiful Soup | JSON  Pandas | Chosen: **Requests.**  Requests, a versatile HTTP library, emerges as the preferred choice for interfacing with the TripAdvisor API. Its simplicity and efficiency make it adept at handling varied tasks, ensuring streamlined data retrieval processes. |
| **Reddit API** | | |
| Praw  Beautiful Soup | JSON  Pandas | Chosen: **Praw.**  Praw, being the official API wrapper for Reddit, aligns seamlessly with our objective of extracting and analyzing Reddit comments pertinent to Dublin bikes and Capital Bikeshare. Its integration with Pandas amplifies data manipulation capabilities, facilitating in-depth insights from the collected data. |

Table 4 - Comparing Libraries and techniques

# 5.0 Bibliography

Creswell, J.W. and Plano Clark, V.L., 2018. Introduction to research methods: A global perspective. 5th ed. Sage Publications.

Flicker, S., 2007. The human research subject: The ethics of participant observation. SAGE Publications.

Sieber, J.E., 1992. Planning ethically responsible research: A guide for students and faculty. Jossey-Bass.

Tufte, E.R., 2001. The Visual Display of Quantitative Information. 2nd ed. Graphics Press, Cheshire, CT.

Uysal, M.P., 2021. Machine Learning and Data Science Project Management from an Agile Perspective: Methods and Challenges. In: Contemporary Challenges for Agile Project Management. pp. 43-55. IGI Global.

Dang, T., McKelvie, W., Valiant, P. and Wang, H., 2023. Improving Pearson's chi-squared test: hypothesis testing of distributions.

Huo, K., Kelly, K. and Webb, A., 2022. The Beneficial Learning Effects of Combining a Hypothesis-Testing Mindset with a Causal Model. Accounting Review, 97(5), pp. 325-348.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2014. Introduction to machine learning. 3rd ed. The MIT Press.

Grus, J., 2015. Data Science from Scratch: First Principles with Python. O'Reilly Media.