# CCT College Dublin

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation for Data Analytics  Programming for Data Analytics  Machine Learning for Data Analytics  Statistics for Data Analytics |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | David McQuaid, Dr. Muhammad Iqbal, Marina Iantorno |
| **Student Full Name:** | Vitor Tiago Torres Notaro |
| **Student Number:** | sba20229 |
| **Assessment Due Date:** | 11/04/2022 |
| **Date of Submission:** | 11/04/2022 |

## 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.

Table of Contents

[CCT College Dublin 1](#_Toc100512963)

[Declaration 1](#_Toc100512964)

[Abstract 3](#_Toc100512965)

[Methodology 3](#_Toc100512966)

[Introduction 4](#_Toc100512967)

[Programming 4](#_Toc100512968)

[Data Visualization 4](#_Toc100512969)

[Statistics 5](#_Toc100512970)

[Central Tendency Measures 5](#_Toc100512971)

[Variation Measures 5](#_Toc100512972)

[Frequency Distribution 6](#_Toc100512973)

[Exploratory Data Analyses (EDA). 8](#_Toc100512974)

[Dublin Bike Dataset Insights from Dataprep Report 9](#_Toc100512975)

[Data Preparation and Visualization. 10](#_Toc100512976)

[Data Prep and Feature Engineering. 11](#_Toc100512977)

[Cleansing. 12](#_Toc100512978)

[Reshape. 13](#_Toc100512979)

[Hierarchy for data preparation. 13](#_Toc100512980)

[Visualizing the Station patterns. 15](#_Toc100512981)

[PCA for dimensionality reduction. 16](#_Toc100512982)

[Machine Learning Models 17](#_Toc100512983)

[Hierarchy Clustering. 18](#_Toc100512984)

[KMeans Clustering 19](#_Toc100512985)

[Model comparison 20](#_Toc100512986)

[Silhouette Score. 20](#_Toc100512987)

[Model Evaluation and Visualization of Results. 21](#_Toc100512988)

[Conclusion: 22](#_Toc100512989)

[Appendix 1: Reflection Report 23](#_Toc100512990)

[Reference List: 23](#_Toc100512991)

# Abstract

The world's top leaders seem to have understood the consequences of not having a greener world, leading to the implementation of initiatives in various sectors as they look to create more environmentally friendly cities. The Transport sector has been making important contributions to help reach this goal. As an example, bike sharing has been highlighted an interesting alternative. This paper will discuss the benefits of using programming, statistics, Machine Learning, and the significance of data visualisation when there is so much data to be explored. How can we use Data Analytics applied to Dublin Bike Station datasets to support us through the process of being green?

# Methodology

This assessment has been made utilising the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology as framework.

CRISP-DM Overview Representatives from SPSS, Teradata, Daimler, NCR, and OHRA developed the Cross industry standard process for data mining (CRISP-DM) in 1996 to standardize a data mining process across industries. CRISP-DM describes six major iterative phases, each with their own defined tasks and set of deliverables such as documentation and reports (www.ibm.com, n.d.).

1. Business Understanding: determine business objectives; assess situation; determine data mining goals; produce project plan

2. Data Understanding: collect initial data; describe data; explore data; verify data quality

3. Data Preparation (generally, the most time-consuming phase): select data; clean data; construct data; integrate data; format data

4. Modelling: select modelling technique; generate test design; build model; assess model

5. Evaluation: evaluate results; review process; determine next steps

6. Deployment: plan deployment; plan monitoring and maintenance; produce final report; review project

Diagram

Description automatically generated

Figure 1 CRISP-DM process diagram (Wikimedia Commons, 2012).

# Introduction

Dublin Bikes is a bike sharing scheme in operation from bicycle docks and stations in Dublin City. Dublin City Council provides Dublin Bike datasets quarterly which contain information regarding historic static data in different formats; for this project the CSV file format will be used. The static data provides stable information such as bike station position, stands available and number of bikes available etc. With updates every three minutes these datasets are interesting sources for research such as the proposals in this project.

Using a quarterly sample from 01-April 2021 till 30- Jun 2021 we aim to learn more about Dublin Bike Stations by answering questions such as:

* How many stations do Dublin Bike have?
* Do all stations have the same number of Bike Stands?
* What else can be learned about Dublin Bike if applied Statistics, Machine Learning, and Data Visualization?

It is clear that pencils and calculators will not suffice as we address these questions and others that may arise.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2 Dublin Bike Dataset head

# Programming

To perform analyses on such large volumes of data (millions of rows) we need the help of tools that allow us process and manipulate our data more easily. Programming language plays an important role in this task and is one of the main tools in data analysis. Although there are several options, each with its specific features, we will use Python to facilitate our research for this project as it is very easy to learn and understand, fewer lines of code are enough to perform complex tasks, it is free, open source and the python community provide great support with excellent packages and libraries available without cost (Grus, 2021).

# Data Visualization

A fundamental field of study for the development of this project is Data Visualization. With such a huge number of observations in our dataset, we must provide a way of presenting all insights rather than producing infinity tables with millions of rows. Data Visualization is the answer. Data visualization provides a perspective on data by showing its meaning in the larger scheme of things. It demonstrates how particular data references stand with respect to the overall data picture. Data visualization puts the data into the correct context, saves time and tells the data story. (GeeksforGeeks, 2021). According to (C Wilke, 2019 p. 1) “Data visualization is part art and part science”. In the book “*Fundaments of Data Visualization*” the author argues that the challenge is to get the right art without compromising the science and vice versa (C Wilke, 2019). Data visualization allow us to learn from the data using colours, charts, discovering trends in data, and by simplifying the comprehension of complex things. However, to achieve such a high level, a lot of reading, research and practice is required. With python, it is easy to create visualizations, but the challenge is to create good ones.

# Statistics

“Statistics refers to the mathematics and techniques with which we understand data” (Grus, 2021, p. 65). As mentioned earlier, our dataset contains a vast volume of data. To properly understand and interpret all this data, we use statistical concepts. The dataset contains 11 variables. However, it is important to be able to describe more than just the number of columns present on a dataset. For this, descriptive statistics are useful. Descriptive statistics comprise three main categories – [Central Tendency](https://corporatefinanceinstitute.com/resources/knowledge/other/central-tendency/) Measures, Variation Measures and Frequency Distribution as per (Weiss 2017).

## Central Tendency Measures

These measures will allow us to summarize in a single number all the values reflecting the center of the data distribution. In statistics there are three common measures of central tendency; mean, median and mode. Each one contains advantages and disadvantages. For example, the mean is easy calculated and interpreted. It is always unique. However, it is affected by extreme values, known as outliers. These three measures isolated are also not enough once they do not capture the variance of the data.

Example of how to calculate the mean (Agresti and Kateri, 2021 p. 13):

## Variation Measures

These measures help us to get values that determine the level of homogeneity within the observations. In other words, we can see how different or similar the values on each variable are and how far they are from each other or how far they are from the Central Tendency values identified using the central tendency measures. Examples of variation measures are variance and [standard deviation](https://corporatefinanceinstitute.com/resources/knowledge/standard-deviation/).

Example of how to calculate the standard deviation (Agresti and Kateri, 2021 p .14):

As previously mentioned, python provides packages and libraries to process these calculations in large scale in just a few seconds. Below is an output example:

Table

Description automatically generated

Figure 3 Dublin Bike dataset Descriptive Statistics

Interpreting the above results, we see that the average of the variable **AVAILABLE BIKES** is 12 Bikes, with 0 bikes as the minimum value observed and 40 bikes available as the maximum, with a standard deviation of 7 bikes. Note that the maximum value of 40 is 4 standard deviations (7x4=28) from the mean of 12 which suggests this variable might present outliers. The frequency distribution is useful to test if this hypothesis is true.

## Frequency Distribution

Frequency distribution can be defined as an arrangement of statistical data that exhibits the frequency of the occurrence of the values of a variable. It is usually represented with histograms that allow a good comprehension about the shape distribution of the data.

Chart, histogram

Description automatically generated

Figure 4 Histogram – Available Bikes

Another way of statically representing the distribution is using a measure of position with Box and Whisker Plots to represent percentiles, quartiles, median, min, max, outliers.

Diagram

Description automatically generated Chart, bar chart

Description automatically generated

Figure 5 Boxplot definition (left) and Boxplot Available Bikes

Notice the Available Bikes boxplot above confirmed the existence of outlier values for the Available Bikes variable. It is important to decide how to deal with outliers. There is no standard approach as outliers can be caused by different reasons and only through investigation can we determine the best approach, which could be keeping the observations, discarding them, or replacing the values etc. However, remember that, depending on the interventions made, the central tendency measures might be affected. That is also the main reason why outliers cannot be ignored as those are the fundaments to calculate probability which in turn is one of the fundaments for Machine Learning.

Applying descriptive analyses to understand more about the Dublin Bike dataset produces important insights that can also be used to calculate probabilities. Probability distribution is a statistical function that is used to show all the possible values and likelihoods of a random variable in a specific range. In other words, probability distribution helps us know all possible outcomes of a particular experiment.

There are different types of probability distributions such as Binomial, Poisson and Normal. Binomial probability distribution is useful to answer questions such as the following:

It is known by my analyses that 25% of the Dublin Bike Stations have 30 Stands. If we randomly choose 5 Stations, what is the probability to find exactly 3 Stations with 30 Stands?

X = number of Stations with 30 Stands (within 5 Stations)

n = 5 number of my sample (limit) 5 stations

p = 0.25 probability of a random element to have the characteristic or attribute

q = 0.75 (q= 1-P for this case q is equal to 1-0.25)

The Binomial distribution probability was calculated as per (Weiss 2017 p. 264).

Using Python, to answer this question: 9% is the probability.

On the other hand, the Normal distribution calculates a cumulation of probabilities. Therefore, we cannot calculate the probability to get an exact value. It will be always greater or less than, but never equals to. The Normal distribution is always symmetric, which means that the curve will never be skewed to one side, and the expected value (average) is always in the middle of the bell curve. Normal probability distribution is useful to answer questions such as the following:

If the number of available bikes stands across Dublin Bike Stations are normally distributed with an average of 20 stands per station and a standard deviation of 10 points, what is the probability of getting one Station with more than or equal to 15 Stands Available?

The Normal distribution probability was calculated as per (Agresti and Kateri 2021 p. 53).

Using Python to answer this question: 69% is the probability

# Exploratory Data Analyses (EDA).

“You might be tempted to dive in and immediately starts building models and getting answers. But you should resist this urge. Your first step should be to explore your data” (Grus, 2021, p. 133).

Exploratory data analysis is used to analyse and investigate data sets and summarize their main characteristics. Often employing data visualization methods, it helps determine how best to manipulate data sources to get the needed answers, making it easier to discover patterns, test a hypothesis, check assumptions or spot anomalies (IBM Cloud Education, 2020a). It was originally developed by John Tukey, an American mathematician, in the 1970s. A benefit of using python are the available libraries, an example being the dataprep library that help us automate the creation of a report to explore and analyse the data.

## Dublin Bike Dataset Insights from Dataprep Report

Some of the observations identified by the dataprep report are:

* No Missing Data has been identifying in any feature

Chart, bar chart

Description automatically generated

Figure 6 - Missing Values bar plot from Dublin Bike EDA

* **STATUS** feature has constant value "Open" (does not add value to analysis).
* **STATION ID** is a unique identifier of each Station (does not add value to analysis).
* **ADDRESS** has a high cardinality: 109 distinct values (does not add value to analysis once we have Lat and Long).
* **AVAILABLE BIKES** is the only feature that presents outliers (around 1% of the values are considered outliers). However, I could not find any values over 40 as that is the maximum possible value once there are no Bike stations with more than 40 stands.

Chart

Description automatically generated

Figure 7 - Scatter plot from Dublin Bike EDA

* It was not possible to find evidence of errors since in none of the observations, the number of bicycles available at the station was greater than the total number of Stands. For that reason, the outliers will be considered acceptable and part of the nature of the operation of the business.
* **BIKE STANDS** is skewed left, the stations have different numbers of stands. It is important to consider this so as to compare fairly.

Chart

Description automatically generated

Figure 8 – Descriptive statistics analysis and Histogram from Dublin Bike EDA

* **AVAILABLE BIKE STANDS** is skewed but close to Normal Distribution.

Chart, histogram

Description automatically generated

Figure 9 - Descriptive statistics analysis and Histogram from Dublin Bike EDA

# Data Preparation and Visualization.

Before getting started with data preparation, it is also a good check to visualize the Stations plotted in Dublin. This will give us an idea as to how the Stations are spread in Dublin.

Map

Description automatically generated

Figure 10 - Dublin Bike Stations position using Follium

After visualizing the stations in a Map and thinking about all the data exploration steps previously made, some questions arise instinctively such as:

* There are differences or similarities between Dublin Bike Stations?
* Is it possible to cluster the Stations based on mean usage?
* How many groups would exist?

Once targets have been defined, data preparation and feature engineering can commence.

## Data Prep and Feature Engineering.

Dublin Bike dataset has two similar columns that refer to times. Both carry similar information with small differences of around 3 or 4 minutes. Only one is required and I selected LAST\_UPDATED because it carries the appropriate picture of the exact date and time when the observation was collected. However, from previous EDA steps, this column is set up as an object type and must to converted to Date Time type.

Extracting important features from Date Time to create new columns is known as feature engineering. This can be useful as there may be big differences between usages on weekends. Conceptually speaking it is likely searching for days which could potentially be considered outliers as they would be distant from the average usage. To make this analysis, two new columns will be created DAY\_NUMBER (starting with 0 = Monday) and DAY\_TYPE (if Weekday of Sunday/Saturday)

Graphical user interface

Description automatically generated with medium confidence

Figure 11 New features created

As there are differences between the updated times across each observation it’s going to be important to create a precise and standard time to be able to compare the stations. Conceptually speaking it is a way to improve the distribution once the previous EDA exploration confirmed that the column LAST\_UPDATED has a high cardinality with 1311282 distinct values. To do that, a new feature must be created using LAST\_UPDATED as a reference, rounding the time each 10 minutes. With this step completed, another new feature can be created extracting only the rounded time from the previous created.

Table

Description automatically generated

Figure 12 New features created

Another issue to address it is the fact each Station has a different amount of Bike Stands; it is important to create a new variable calculating the percentage occupancy by dividing the number of available bikes by the number of Bike Stands each station has. This enables a comparison of the situation of each Station across my timeline. Conceptually this would be a way to put all stations on the same page to compare fairly.

## Cleansing.

Data cleansing is also an important step. As there are many redundant columns not useful for the exercise its recommended to drop them.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 13 Dataset head after some stages of data preparation

## Reshape.

At this stage it is important to mention that with the current format displayed above, the dataset observations are those specific events that took place in each station in a very specific time frame. As previously mentioned, a new column DAY\_NUMBER was created to analyse the patterns of each weekday. However, to perform some analyses it is sometimes necessary to reshape the data adjusting the observations for what needs to be observed. A pivot table function will be used setting the DAY\_NUMBER as the observations defining NEW\_TIME column as the attributes, populating my attribute values with the mean of PERCENTAGE\_OCCUPANCY.

## Hierarchy for data preparation.

The Hierarchy Cluster model will be used to check which days of the week are similar to each other and see if Clusters among the different days of the week could be generated using Euclidean distance.

A picture containing graphical user interface

Description automatically generated

Figure 14 Euclidean Distance - equation presented in class by Dr. Muhammad Iqbal

Dendrograms are useful to represent the relationship between objects, they also display the distance between each pair of sequentially merged objects in a feature space. Dendrograms are commonly used in studying the hierarchical clusters before deciding the appropriate number of clusters.

Chart, box and whisker chart

Description automatically generated

Figure 15 Demdrogram of Weekday Type showing possible Clusters

The above plot confirmed the weekends (Saturday 5 and Sunday 6) have similar patterns. We can also see that Tuesdays (1) and Thursdays (3) are likely similar. Wednesdays (2), Fridays (4) and Mondays (0) are also similar. However, one common approach is to analyse the dendrogram and look for groups that combine at a higher dendrogram distance. The Grey dashed draw line crosses three groups and suggests 0.01 as distance (from 0.05 to 0.06). The Red dashed draw line crosses two groups and suggests a distance greater than 0.09 (from 0.07 to 0.16). For that reason, two clusters (weekends and workdays) is the most appropriate. An easy way to visualize the trend differences is by plotting a Line chart. Using line charts to represent time series is a generally accepted practice. However, the dots are frequently omitted altogether (C Wilke 2019 p. 133).

Graphical user interface, chart, scatter chart

Description automatically generated

Figure 16 Line Chart – Mean Occupancy per day type

The above chart confirmed the weekends have trends and patterns completely different from the workdays. For that reason, Saturday and Sunday will be interpreted as outliers when compared with the workdays Monday to Friday. Outliers in input data can skew and mislead the training process of machine learning algorithms, resulting in longer training times, less accurate models, and ultimately poorer results. For this reason, observations from those two days will be excluded of the next phase.

## Visualizing the Station patterns.

To visualize the mean trend of each station across the Time Frame, it is necessary to adjust observations one more time using a Pivot Table function setting the NEW\_TIME as the observations, defining STATIONS column as the attributes, populating the attribute values with mean PERCENTAGE\_OCCUPANCY.

It is known this visualization will be very poor and extremely difficult to interpret, although is exactly the point to be proved. As there are 109 different Stations, for our human eyes it is impossible to find out trends and patterns only looking at that Line Chart.

Chart

Description automatically generated

Figure 17 Line Chart Mean Occupancy for 109 Stations

The above plot seems aggressive to the eyes. A better way to cluster those stations is needed and for that reason Machine Learning models will be used to complete this task. To perform this exercise, the dataset must be reshaped using a Pivot Table function setting the NAME as my observations (because the Stations are the objects to cluster), defining NEW\_TIME column as the attributes, populating the attribute values with mean PERCENTAGE\_OCCUPANCY

A picture containing text

Description automatically generated

Figure 18 Pivot Table – Reshape Dublin Bike Dataset

For obvious reasons the new dataset contains 109 observations (109 stations) and 144 columns (24 hours split each 10 minutes of the day). This is a perfect opportunity to test if dimensionality reduction would be applicable for this new dataset without losing its properties. The questions to be addressed are:

* Is it possible to perform dimensionality reduction on this dataset without losing its properties?
* How many components would I need to explain at least 90% of the Variance of my Station Dataset?

To answer these questions, PCA (Principal Components Analysis) will be applied.

## PCA for dimensionality reduction.

Principal component analysis (PCA) is an unsupervised learning method used for dimensionality reduction.



Figure 19 PCA equation presented in class by Dr. Muhammad Iqbal

However, before applying PCA, a good approach is pre-processing the data. Standardization of a dataset is a common requirement for many machine learning estimators (except the ones based on decision tree). They might behave badly if the individual features do not more or less look like a standard normally distributed data. The one used for this experiment is called StandardScaler from Scikit Learn, it standardizes the features onto unit scale (mean = 0 and standard deviation = 1) This is a requirement for the optimal performance of many Machine Learning algorithms including PCA. After PCA is applied, the output obtained was that 2 Principal components are needed to explain 92% of the variance of this dataset, with the first component responsible for explaining 68% and the second component 23%. For that reason, using only the two principal components instead of 144 columns is a good choice.

It is also a possible approach using the two principal components in a plot trying to visualize the existence of possible groups. However, when observing the below plot, human eyes are not able to clearly identify different groups. This reinforces the idea that Machine Learning Models are important tools to solve complex problems.

Chart, histogram

Description automatically generated

Figure 20 Pair plot of Principal Components

# Machine Learning Models

Although two models have been already used, **PCA** for dimensionality reduction and **Hierarchical** when researching for different groups on weekday types (weekends or workdays) during the data preparation phase, it is important to recap some definitions.

What is machine learning? According to IBM, Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on using data and algorithms to imitate how humans learn, gradually improving accuracy. It uses different mathematical algorithms to work through large amounts of data & return easy, visual results of the data (IBM Cloud Education, 2020).

There are three types of machine learning: supervised, unsupervised and reinforcement learning. Supervised models are useful to perform Classifications and Regressions models which implicates a label, or a dependent variable is needed. Unsupervised learning is a type of Machine Learning algorithm that learns patterns from untagged data. The goal of the algorithm is to find relationships within the data and group data points based on the input data creating clusters or groups. On the other hand, reinforcement learning is an approach to machine learning that learns by doing. While the first two machine learning techniques learn by passively taking input data and finding patterns within it, reinforcement learning uses training agents to actively make decisions and learn from their own outcomes.

Considering the defined targets from the previous phase, it is clear Unsupervised Learning is the proper type to be applied to address those previous questions. For this next exercise, two algorithms will be experienced, Hierarchy and KMeans as we would like to cluster using the Euclidean Distance between points.

A picture containing graphical user interface

Description automatically generated

Figure 21 – Euclidean Distance function presented in class by Dr. Muhammad Iqbal

The below table from scikit-learn show some of the methods available.

A picture containing table

Description automatically generated

Figure 22 Overview of Cluster methods available on scikit-learn

## Hierarchy Clustering.

Applying Hierarchy Cluster using the new dataset that contains only the two principal components for each observation, the output can be reproduced using a Dendrogram.

Chart, box and whisker chart

Description automatically generated

Figure 23 Demdrogram of Dublin Bike Stations showing possible Clusters

The above plot confirmed some stations have similar patterns. As previously mentioned, one common approach is to analyse the dendrogram and look for groups that combine at a higher dendrogram distance. The Grey dashed draw line crosses three groups and suggests a distance around 19 (from 56 to 75) when it touches the red line. The Red dashed draw line crosses two groups and suggests a distance greater than at least 30 (from 75 to 110). For this reason, two clusters are the most appropriate.

Chart, scatter chart

Description automatically generated

Figure 24 Pair plot of Hierarchy Cluster

## KMeans Clustering

Different from Hierarchy, KMeans algorithm requires the number of clusters to be specified before being applied. Elbow method can be used to get insights of the possible number of clusters for parameter K (number of clusters).

A picture containing text, watch

Description automatically generated

Figure 25 - Within-Cluster Sum of Squares (WCSS) – Elbow Method function presented in class by Dr. Muhammad Iqbal

Chart, line chart

Description automatically generated

Figure 26 - Elbow Method

From the above plot I can see the line starts to become flat around 4 but the optimum number of neighbours could be also 5 or 6. For that reason, the 3 possibilities were tested and measured with the Silhouette score to compare the results.

# Model comparison

## Silhouette Score.

Silhouette Coefficient, also known as silhouette score, is a metric used to calculate the efficiency of a clustering technique.

A picture containing diagram

Description automatically generated

Figure 27- Formula for the silhouette score function presented in class by Dr. Muhammad Iqbal

Its value ranges from -1 to 1, where 1 means clusters are well apart from each other and clearly distinguished, 0 means clusters are indifferent, or we can say that the distance between clusters is not significant, and -1 means clusters are assigned in the wrong way.

Text

Description automatically generated

Figure 28 - Silhouette score results

The Silhouette Score test indicates KMeans with 5 Cluster as the best option with a higher score than Hierarchy with only 2 clusters.

Text

Description automatically generated

Figure 29- KMeans - Sum of Squared Error (SSE) function presented in class by Dr. Muhammad Iqbal

## Model Evaluation and Visualization of Results.

A pair plot for KMeans will print the results showing the Clusters.

Chart, scatter chart

Description automatically generated

Figure 24 - Pair plot of KMeans Cluster and Cluster size

After visualising the 5 groups produced by KMeans the cluster classes can also be included into the original dataset and used to count the number of Stations on each cluster and re-plot the Station Maps now Clustered.

A picture containing background pattern

Description automatically generated

Figure 30 - Dublin Bike Stations position using Follium and Clustered by KMeans

It is also possible to revisit the line chart Station Plot Usage to visualise the Clusters by mean occupancy.

Chart, line chart

Description automatically generated

Figure 31 – Line Chart

The line chart is now clean and intuitive. With only 5 different groups, even for our human eyes it is easy to see the trends, patterns and possible correlations just by analysing the Line Chart. This also confirms the importance of Data Visualization in providing a perspective on data by showing its meaning in the larger scheme of things. It demonstrates how particular data references stand with respect to the overall data picture. Data Visualization puts the data into the correct context, saves time and tells the data story (GeeksforGeeks, 2021).

# Conclusion:

Upon review, I found that the CRISP\_DM framework is an excellent tool to keep focus on the tasks at hand and I will seek to make use of this in future projects. Prior to signing off on this project, reassessments were made to ensure that all proposals and objectives initially addressed were met and that all questions were answered appropriately with sufficient evidence and rationale for decisions made. It is evident that Data Science applied in any dataset can help us to better understand the world and make better decisions as human beings, with a mindset of recognising the impacts for future generations.

Some points to highlight regarding this assessment is the importance of:

* having a good methodology.
* having the right tools available.
* discipline for reading, researching, and developing skills and the necessary knowledge in Programming, Statistics, Data Preparation, Machine Learning and Data Visualization.

With the project findings presented, it is clear that further research could be made to progress more about the topic. There are clearly some correlations between the different groups of Dublin Bike Stations and maybe opportunities for a better distribution of the Bikes among the stations to improve the service provide. This might become a topic for future discussions.

# Appendix 1: Reflection Report

I faced numerous challenges in accomplishing a successful outcome to this project: the initial impact of working individually, the initial lack of a consistent theoretical basis on the various subjects, and the initial lack of experience writing a report in English. My belief is that I have overcome these hurdles as I delved further into research, both theoretical and practical. The challenge proposed by David, Muhammad and Marina to develop our work based on rational justifications forced me to review with a more technical examination of each step made and brought me great benefits and learning in this project. In the end, I could improve some of my approaches in order to deliver a better project throughout what has been a steep learning curve.

In summary, I have greatly enjoyed the challenges set out in this assessment and forward to continuing my educational journey with more projects and classes.

# Reference List:

Agresti, A. and Kateri, M. (2021). *Foundations of statistics for data scientists : with R and Python*. First Edition ed. Boca Raton: Crc Press.

Bhardwaj, A. (2020). *Silhouette Coefficient : Validating clustering techniques*. [online] Medium. Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

Blackmist (n.d.). *Train and deploy a reinforcement learning model (preview) - Azure Machine Learning*. [online] docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-reinforcement-learning

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

C Wilke (2019). *Fundamentals of data visualization : a primer on making informative and compelling figures*. Sebastopol, Ca: O’reilly Media.

Chun-Houh Chen and Al, E. (2016). *Handbook of data visualization*. Berlin: Springer.

Connors, L. (2021). *Creating a Simple Map with Folium and Python*. [online] Medium. Available at: https://towardsdatascience.com/creating-a-simple-map-with-folium-and-python-4c083abfff94

Das, A. (2020). *Hierarchical Clustering in Python using Dendrogram and Cophenetic Correlation*. [online] Medium. Available at: https://towardsdatascience.com/hierarchical-clustering-in-python-using-dendrogram-and-cophenetic-correlation-8d41a08f7eab

data.smartdublin.ie. (n.d.). *Dublinbikes DCC - data.smartdublin.ie*. [online] Available at: https://data.smartdublin.ie/dataset/analyze/33ec9fe2-4957-4e9a-ab55-c5e917c7a9ab

dataprep.ai. (n.d.). *DataPrep — The easiest way to prepare data in Python*. [online] Available at: https://dataprep.ai/

fontawesome.com. (n.d.). *Font Awesome*. [online] Available at: https://fontawesome.com/icons?d=gallery

Galarnyk, M. (2017). *PCA using Python (scikit-learn)*. [online] Medium. Available at: https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

GeeksforGeeks. (2021). *What is Data Visualization and Why is It Important?* [online] Available at: https://www.geeksforgeeks.org/what-is-data-visualization-and-why-is-it-important/#:~:text=Data%20visualization%20is%20very%20critical%20to%20market%20research

Grus, J. (2021). *DATA SCIENCE FROM SCRATCH : first principles with python.* Second ed. O’Reilly.

IBM Cloud Education (2020a). *What is Exploratory Data Analysis?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/exploratory-data-analysis.

IBM Cloud Education (2020b). *What is Machine Learning?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/machine-learning

James (2017). *Usage patterns of Dublin Bikes stations*. [online] Medium. Available at: https://towardsdatascience.com/usage-patterns-of-dublin-bikes-stations-484bdd9c5b9e

Jeffares, A. (2019). *How I used Machine Learning to improve my Dublin Bikes transit*. [online] Medium. Available at: https://towardsdatascience.com/how-i-used-machine-learning-to-improve-my-dublin-bikes-transit-b6bdc7c2b5cb

Jeffares, A. (2021a). *Data Science Nanodegree*. [online] GitHub. Available at: https://github.com/alanjeffares/data-science-nanodegree/blob/master/dublin-bikes-analysis/get\_nearest\_available\_bike.py

Jeffares, A. (2021b). *Data Science Nanodegree*. [online] GitHub. Available at: https://github.com/alanjeffares/data-science-nanodegree/blob/master/dublin-bikes-analysis/data\_load\_processing\_viz.ipynb

Lawlor, J. (2021). *dublin-bikes-timeseries-analysis*. [online] GitHub. Available at: https://github.com/jameslawlor/dublin-bikes-timeseries-analysis

Martinez, J.C. (2021). *How to plot your data on maps using Python and Folium*. [online] livecodestream.dev. Available at: https://livecodestream.dev/post/how-to-plot-your-data-on-maps-using-python-and-folium/

Mckinney, W. (2018). *Python for data analysis : data wrangling with pandas, NumPy, and IPython*. Second ed. Sebastopol, Ca: O’reilly Media, Inc., October.

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O’reilly.

Nichani, P. (2020). *OutLiers in Machine Learning*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660#:~:text=Outlier%20is%20an%20observation%20that%20appears%20far%20away

plotlygraphs (2019). *Line Charts*. [online] plotly.com. Available at: https://plotly.com/python/line-charts/

rachelbreslin (2022). *dublin\_bikes/Dublin Bikes Analysis.ipynb at main · rachelbreslin/dublin\_bikes*. [online] GitHub. Available at: https://github.com/rachelbreslin/dublin\_bikes/blob/main/Dublin%20Bikes%20Analysis.ipynb

Ranjan, A. (2020). *Hierarchical Clustering (Agglomerative)*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/hierarchical-clustering-agglomerative-f6906d440981

scikit-learn.org (n.d.). *sklearn.metrics.adjusted\_rand\_score — scikit-learn 0.23.1 documentation*. [online] scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted\_rand\_score.html

Scikit-learn.org. (2019). *sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Summerfield, M. (2010). *Programming in Python 3 : a complete introduction to the Python language*. Second ed. Upper Saddle River, New Jersey: Addison-Wesley.

Team, T.A.E. (2021). *Principal Component Analysis (PCA) with Python Examples — Tutorial*. [online] Medium. Available at: https://pub.towardsai.net/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa#5e9a

Weiss, N.A. (2017). *Introductory Statistics*. 10th ed. Pearson Education.

Wilke, C.O. (n.d.). *Fundamentals of Data Visualization*. [online] *clauswilke.com*. Available at: https://clauswilke.com/dataviz/directory-of-visualizations.html.

www.dublinbikes.ie. (n.d.). *DublinBikes*. [online] Available at: https://www.dublinbikes.ie/

www.ibm.com. (n.d.). *CRISP-DM Help Overview*. [online] Available at: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview.

www.youtube.com. (2018). *PyData Dublin: Usage patterns of Dublin Bikes stations - James Lawlor*. [online] Available at: https://www.youtube.com/watch?v=59ck\_Z75cEY

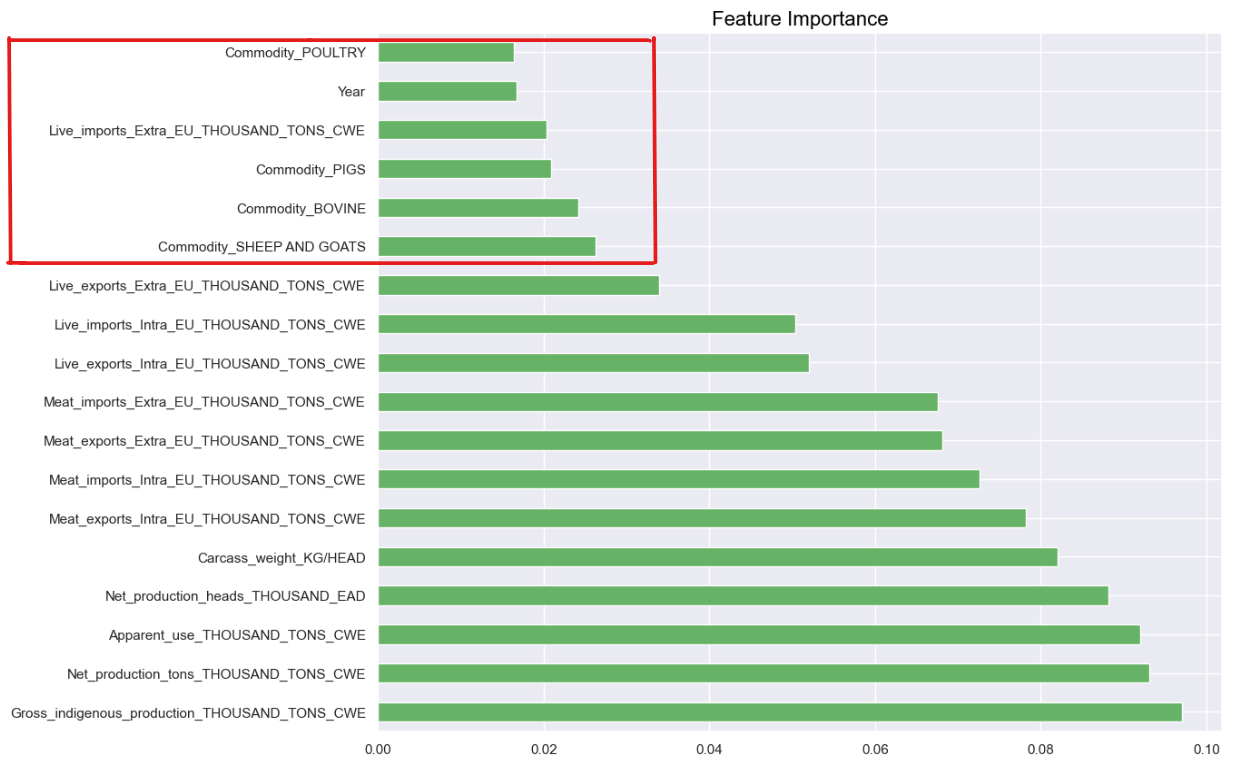


Figure - Variables to exclude (they did not impact models performance)

Graphical user interface, text, application, email

Description automatically generated

Figure Data Leakage detected while running Regression Model.



Figure Results using all variables.

Text

Description automatically generated