|  |  |
| --- | --- |
| **Module Title:** | **Strategic Thinking** |
| **Assessment Title:** | **CA2 Capstone Project Proposal** |
| **Lecturer Name:** | **James Garza** |
| **Student Full Name:** | **Ciara Barron** |
| **Student Number:** | **sba23419** |
| **Assessment Due Date:** | **Sunday, 17 December 2023, 11:59 PM**  **Extension received: 23/12/23** |
| **Date of Submission:** | **23/12/23** |

**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.** |

Contents

[Title 3](#_Toc154223680)

[Introduction 3](#_Toc154223681)

[Objectives 3](#_Toc154223682)

[Problem Definition 4](#_Toc154223683)

[Scope 4](#_Toc154223684)

[Data Sources 5](#_Toc154223685)

[Ethical Consideration 5](#_Toc154223686)

[Project Management Methodology 6](#_Toc154223687)

[Data Understanding 6](#_Toc154223688)

[Machine Learning Implementation 11](#_Toc154223689)

[Summary of Findings and Conclusions 15](#_Toc154223690)

[Future Recommendations 15](#_Toc154223691)

[Reference list 16](#_Toc154223692)

Title**:** Bike Sharing- An In-Depth Analysis of the Potential Benefits of Introducing Bike Sharing in Wexford

# Introduction

Currently there are over 500 bike-sharing programs around the world. They have a huge role to play in traffic, environmental and health issues. Despite being 50 years old, bike sharing is only recently seeing international growth (smartcitiesdive.com, n.d.). This proposal will briefly define potential problems in relation to introducing bike sharing and elaborate on the significance of an exploration of the issue. It will outline the objectives and scope of the proposed report on the issue of introducing bike sharing in Wexford. Finally, this essay will identify the appropriate data sources and ethical considerations for the proposed report.

# Objectives

Using several data sources where studies have been conducted on the real-world application of bike sharing systems, the main objective of this report will be to examine economic viability of implementing a bike sharing system, to consider potential cost savings, revenue generation and job creation. Other objectives include evaluating environmental impact, promoting healthier lifestyles, and enhancing urban mobility.

# Problem Definition

In Ireland and Wexford specifically, there are several pressing environmental and transport issues such as traffic congestion, air pollution and a scarcity of sustainable transport alternatives. There is a reliance on traditional modes of transport which contribute negatively on the environment and on citizen’s quality of life. To address these issues, we need to look at innovative solutions such as introducing bike sharing options.

# Scope

This report focuses on assessing the feasibility and potential benefits of introducing a bike sharing system in Wexford. It will look at factors such as infrastructure requirements, community engagement, economic implications, and environmental impact. The scope encompasses a comprehensive analysis of the main dataset (UCI Machine Learning Repository, n.d.) aiming to extract valuable insights that inform decision-making regarding the implementation of a bike sharing program.

This project takes place over two semesters. By the end of semester two, it is expected that a comprehensive report will be delivered detailing the findings on the above objectives.

The aim is to provide valuable insights that can help inform future decisions and policies. The project will not be making specific policy recommendations.

To give a high-level overview of the planned methods, techniques, and approaches that it is hoped will be carried out, an initial review will be undertaken of the information available, beginning with the sources referenced below. The relevant data will then be collected, and statistical analysis techniques will be used to clean the data and identify trends and patterns. Finally, the findings will be documented in a detailed report.

The timeline will consist of four stages divided over three months where the work will be divided equally, roughly three weeks will be allocated to each stage. These stages can be broken down into, Literature Review, Data Collection, Data Analysis and Report Writing. Project management techniques to keep track of the work.

# Data Sources

The main dataset this report will be referencing will be ‘Bike Sharing Dataset’ from UCI Machine Learning Repository. This is a comprehensive repository that includes information on bike sharing usage patterns and related metrics. Supplementary data from data. World will also be used, there are 65 bike sharing datasets available here to the report which may also be referenced. The report will conduct exploratory data analysis and statistical methods to achieve relevant results regarding the potential advantages and challenges associated with implementing a bike sharing scheme in Wexford. The report will also reference reports and journals discussing the relevant issues along with Government strategies and action plans to tackle these issues.

# Ethical Consideration

This report will reference several datasets, mainly open data held by public bodies and open-source data. Open-source data gives “everyone access to non-personal government data which can deliver enhanced economic, social, environmental and democratic benefits to all” (Government of Ireland, n.d.). These datasets are not only easily accessible but also free to use. The authors will be aware of ensuring data accuracy when using different datasets and will highlight the potential issues related to this throughout the report. We will ensure transparency in our methods and findings, using open-source data that aligns with these considerations.

# Project Management Methodology

To implement this project successfully, it is important to adhere to an efficient project plan. A detailed plan will be developed, the business concept will be explored, the dataset will be researched and chosen, a machine learning model will be applied, and an artefact of code will be supplied. There will be a comprehensive report, which will cover each phase of the project with clarity and conciseness.

It is important to note that there are many project management methodologies in use around the world. An agile approach will be chosen here, Hobbs & Petit (2017) have found that these methods allow a flexible approach to breakdown the projects into manageable tasks, analyse and prioritise deliverables and successfully complete the task on time.

Initially a Gannt chart is created, this will serve as a form of documentation for the schedule. It visually represents various aspects of the project. It visualises the entire timeline of the project and shows dependencies between the tasks. It also helps to track milestones; this will allow the project manager and stakeholders to track progress and ensure the project is on schedule.

Figure 1. Gannt Chart

A screenshot of a computer

Description automatically generated

# Data Understanding

The next step is to perform Exploratory Data Analysis (EDA) to explore the data. Insights into the data will become clear here, relevant patterns will be visualised.

Initially the bike. Head function is applied to begin the initial data exploration. This displays the first number of rows of the chosen dataset which consists of 17,379 observations across 17 features of data. This allows us to check the actual values in the data frame, providing an overview of the data structure such as column names and the data types. There are no duplicates and no missing values. Columns were renamed for ease of understanding. Several irrelevant columns were removed; ‘instant’ as it was a Unique Identifier, dteday as it was not relevant to the research, and ‘casual’ and registered’ as they are closely correlated with ‘count’ so all three were not required.

Figure 2. EDA

A screenshot of a computer

Description automatically generated

Visualisations (Figure 3) are created to see if the data is skewed or balanced. Boxplots (Figure 4) are created to see if there are any outliers within the dataset. It is useful to find out the relationships between the features and the target variable.

Figure 3. Visualisations

A screenshot of a graph

Description automatically generated

These visualisations provide a clear overview of the breakdown of each numeric feature, along with indicators for both the mean and the median, making it very easy to visualise whether the data is skewed or has a normal distribution. ‘Hour’ and ‘month’ are the closest to normal as we can see.

Figure 4. Boxplots

A group of rectangular objects with text

Description automatically generated with medium confidence

Boxplots or ‘box-and-whisper’ plots are important to visualise key statistical measures. It also shows presence of outliers in the data. In this dataset there are a small number of features with outliers, most notably ‘windspeed’, this could be for several reasons; geographical factors, seasonal variations or measurement errors could be attributed to the outliers.

Figure 5. Correlation Matrix

A diagram of the weather

Description automatically generated

Applying multivariate analysis, a correlation matrix (Figure 5) was developed to visualise the connection between features.

Several things can be interpreted from this matrix; for example, ‘humidity’ is inversely related to ‘total\_count’ as people would not like to cycle when it’s extremely humid. Similarly, while ‘temp’ and ‘atemp’ are highly related, they both have quite an effect on ‘total\_count’. ‘Weather’ and ‘total\_count’ are inversely related, people will not want to rent bicycles in bad weather. ‘temp’ and ‘atemp’ are closely related so the decision was made to drop one of them.

Once these steps are complete, running parallel with the Gannt chart, the next step is to apply scaling to the features, this is an unsupervised learning algorithm, adjusting the features so that the data representation is more suitable to apply a machine learning algorithm (Muller & Guido, 2017).

For this exercise MInMax scaling was applied, this is a linear transformation that doesn’t alter the distribution of the data. There are many scaling techniques and in future iterations of this agile process, different ones will be applied to test for better applications of the machine learning model.

Figure 6. Scaling

A screenshot of a computer program

Description automatically generated

# Machine Learning Implementation

Figure 7. Gannt Chart 0.2

A screenshot of a computer

Description automatically generated

The final stages of the machine learning implementation section of the Gannt chart (Figure 7) outline the process of training a linear regression model, assessing its performance using cross-validation, obtaining model co-efficients, making predictions and evaluating prediction accuracy.

As per 4.1 in figure 7, ‘ML Algorithm Selection’, for the first sprint in this agile methodology, it was decided to apply linear regression to the dataset. This was chosen for its predictive performance abilities; it can serve as a baseline model for comparison with more complex algorithms. It is a good starting point in the agile methodology. It has no parameters, which is a benefit, but it also has no way to control model complexity.

10-fold cross-validation (scikit-learn, n.d.) was used to calculate the R-squared scores on the training data. It is recommended to use at least five-fold cross-validation. Ten-fold is a common choice and strikes a good balance between computational efficiency and providing a reliable estimate of model performance. In later sprints of the model, different folds can be assessed at this point. When choosing an evaluation metric, it is important to keep the business goals in mind, this is where the agile approach will be beneficial as different techniques can be evaluated and tested to ensure that the business impact is assessed (Muller & Guido, 2017).

Figure 8. Results

A screenshot of a computer

Description automatically generated

Training the model to a linear regression method results in a CV Mean of 0.387, this would indicate that on average, the model explains about 38.7% of the variance in the target variable across the different folds. The STD, standard deviation, result is 0.019 which would suggest that the model’s performance is consistent across the different subsets of data. This is a good start for this sprint, it is important to explore the problem in more detail with more trials to make effective decisions on the problem.

Figure 9. Intercept and Co-efficient

A screenshot of a computer

Description automatically generated

To interpret this intercept and coefficients correctly, it would be important to have another model to use for comparison. In future sprints this would become clearer.

Figure 10. Results 0.2

A grey rectangular sign with black numbers

Description automatically generated

Figure 10 shows the commonly used metrics to assess the performance of the linear regression model.

Figure 11. Final Visualisation

A screenshot of a graph

Description automatically generated

These figures indicate that the model is overfitted, using this visualisation along with the results shown in Figure 10, more work needs to be done on this model to ensure that it is fitted correctly and produce better results for the business problem.

# Summary of Findings and Conclusions

Perhaps too simple a model was chosen for the initial sprint, Muller & Guido (2017) state that the more complex we allow our model to be, the better we will be able to predict on the training data. In future sprints, it is hoped to prove this point, but it is important to be cautious as the opposite is also true, if a model is too complex, ‘we start focusing too much on each individual data point in our training set, and the model will not generate well to new data’. This is why the agile approach chosen is the best methodology to approach this problem. It allows for flexibility in applying different models, while also realising that this is an iterative process, based on feedback, evaluation and new insights, steps can be taken to optimise the desired outcomes.

The Gannt chart was an extremely useful tool to provide a visual representation of the project schedule. It allowed for progress tracking and showed clearly when milestones were achieved and is useful to put the project management methodology into practice.

The execution of the capstone project encountered challenges arising from a transition in project dynamics. Initially structured as a collaborative effort within a group setting, subsequent instructions clarified that the second assessment would be conducted individually. Faced with uncertainty about the specific requirements of the second assessment, an adjustment was made to the initially chosen dataset. This adaptation aimed to align more closely with the updated project guidelines and facilitate a more effective individual assessment process.

# Future Recommendations

In the future, a continuous feedback mechanism will be introduced, this will fit in well with the agile, iterative process that was applied and will help to refine the approached and address challenges to reach the desired outcome.

# Reference list

GeeksforGeeks. (2021). *Data Pre-Processing wit Sklearn using Standard and Minmax scaler*. [online] Available at: <https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/> *accessed 20/12/23*

Hadi Fanaee-T (n.d.). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: <https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset>. *accessed 01/12/23.*

Hobbs, B. and Petit, Y. (2017). Agile Methods on Large Projects in Large Organizations. *Project Management Journal*, 48(3), pp.3–19. doi:https://doi.org/10.1177/875697281704800301.

Mahadevan, M. (2022). *Step-by-Step Exploratory Data Analysis (EDA) using Python*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/#h-step-7-eda-exploratory-data-analysis> *accessed 17/12/23*.

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

Roula Michaelides and Elena Antonacopoulou (2021). Project Management as a Dynamic Collaborative Social Practice—Collaborative Innovation Revisited.

SciKit-Learn (2009). *3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/cross_validation.html>. *accessed 17/12/23*.

Smartcitiesdive.com. (2017). *From Amsterdam to Beijing: The Global Evolution of Bike Share | Smart Cities Dive*. [online] Available at: <https://www.smartcitiesdive.com/ex/sustainablecitiescollective/amsterdam-beijing-global-evolution-bike-share/1100421/>. *accessed 18/12/23*.

Github Repository: https://github.com/CiaraB123/Strategic\_Thinking\_CA2