**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_Repeat* |
| **Lecturer Name:** | *Marina Iantorno/Taufique Ahmed*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Munkhgerel Ankhtuvshin* |
| **Student Number:** | 2023231 |
| **Assessment Due Date:** | 2nd of August 2024 |
| **Date of Submission:** | 2Nd of August 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. |

**Contents**

[Introduction 3](#_Toc172727428)

[Project Planning 4](#_Toc172727429)

[Objective 4](#_Toc172727430)

[Timeline 4](#_Toc172727431)

[Milestones 6](#_Toc172727432)

[Resources 7](#_Toc172727433)

[Deliverables 7](#_Toc172727434)

[Research 7](#_Toc172727435)

[Data Sources 7](#_Toc172727436)

[Analytical Methods 8](#_Toc172727437)

[Descriptive and Inferential Statistics 8](#_Toc172727438)

[Machine Learning 8](#_Toc172727439)

[Sentiment Analysis 9](#_Toc172727440)

[Literature Review 9](#_Toc172727441)

[Methodological Justification 9](#_Toc172727442)

[Justification 10](#_Toc172727443)

[Data Sources and Preprocessing 10](#_Toc172727444)

[Data Sources 10](#_Toc172727445)

[Data Preprocessing 11](#_Toc172727446)

[Analytical Techniques 11](#_Toc172727447)

[Descriptive Statistics 11](#_Toc172727448)

[Inferential Statistics 11](#_Toc172727449)

[Machine Learning Models 12](#_Toc172727450)

[Techniques 12](#_Toc172727451)

[Sentiment Analysis 12](#_Toc172727452)

[Techniques 12](#_Toc172727453)

[Visualization Tools 13](#_Toc172727454)

[Implementation 13](#_Toc172727455)

[Data Acquisition and Preprocessing 13](#_Toc172727456)

[Data Loading 13](#_Toc172727457)

[Data Cleaning 14](#_Toc172727458)

[Exploratory Data Analysis (EDA) 14](#_Toc172727459)

[Descriptive Statistics 14](#_Toc172727460)

[Visualizations 15](#_Toc172727461)

[Inferential Statistics 17](#_Toc172727462)

[Confidence Intervals 17](#_Toc172727463)

[Hypothesis Testing 18](#_Toc172727464)

[Machine Learning Models 18](#_Toc172727465)

[Data Preparation 18](#_Toc172727466)

[Linear Regression 18](#_Toc172727467)

[Random Forest 19](#_Toc172727468)

[Interactive Dashboard 19](#_Toc172727469)

[Results and Analysis 20](#_Toc172727470)

[Descriptive Statistics 20](#_Toc172727471)

[Inferential Statistics 20](#_Toc172727472)

[Machine Learning Models 21](#_Toc172727473)

[Interactive Dashboard 21](#_Toc172727474)

[Conclusion 21](#_Toc172727475)

[References 22](#_Toc172727476)

# Introduction

Cycle riding has evolved as probably the most famous and grappling agencies of convergence transport and recreation in the present and future world due to reasonableness of its health, environmental, and urban capacity profitable worth. Thus, Ireland as a country with a highly developed cycling culture is ideal for studying tendencies and people’s behaviors. Hence, this project’s main emphasis is on cycling statistics in Ireland and benchmarking these numbers against global cycling numbers to draw certain conclusions and make certain observations and, at that, provide recommendations as needed. It covers many techniques of data analysis such as descriptive statistics, inferential statistics, machine learning modeling and sentiment analysis.

The data available in Ireland is vast and provides information regarding the number of cyclists in the country, cycling paths, and most importantly the usage of ‘‘dock less’ operation public bike-sharing systems. Using this data it is possible to make sound decisions and make a large leap in terms of cycling tendencies in the country. Nevertheless, for more understanding and putting the results into perspective, it is necessary to look into the global cycling statistics. Such a comparison helps the author to define cycle culture specifics in Ireland and recognize the similarities or differences with other countries.

The first stage of any analysis is the EDA, in which we describe the model and distribution of the data set and its most important characteristics. The techniques of descriptive statistics help to describe the findings qualitatively and quantitatively giving information on the central measures, variability and distribution patterns of the data. Histogram, box and scatter plots are some of the recommended types of plots that are used in the analysis of the patterns, anomalies or even relationships within the data. For instance, analysing the pattern of the number of cyclists in Ireland enables one to understand cycling activity levels to time high and low periods of cyclists usage.

While descriptive statistics is used for summarizing the data, inferential statistics is used for making conclusions about the population from the sample. This includes approximation of population characteristics which comprises of estimation of parameters, deducing intervals of confidence and carrying out tests of hypothesis. For instance, we can say the average number of cyclists in Ireland and check whether this number is greater or less than the global average. Inferential statistics offer proper procedures of drawing hypotheses and supporting them by the observations we have made.

Another component in this project is the use of machine learning algorithms as they assist in predicting and/or categorizing outcomes related to cycling. The data is then segmented into training and test sets, and supervised as well as unsupervised machine learning is applied. For unsupervised learning process, the most preferable techniques that are selected are linear regression and random forest with the aim of estimate the number of cyclists in future using previous data and other characteristics. They assist in identifying various determinants that affect cycling frequency and predicting the trends in the future. Hence, this includes the use of data mining, particularly clustering, which will show areas of similar data points as well as cyclic behaviors and patterns.

# Project Planning

## Objective

The objective of this project is to analyze Ireland's cycling data, compare it with global cycling data, and perform various data analytics and machine learning tasks to derive meaningful insights and make evidence-based recommendations.

## Timeline

The project planning is based on the weekly schedule in which there are assigned tasks for each week to achieve the goal and complete the project by July 21, 2024.

**Week 1: Initial Planning and Data Acquisition (June 1 - June 7, 2024)**

* **Objective:** Set out clear goals for the project, determine what kind of data would be needed, and in general how the project could be approached.

**Tasks:**

* Develop key goals and objectives of the project and its scope.
* Acquire the dataset of Ireland’s cycling and the international cycling data.
* We Ensure we have all the appropriate tools and libraries for data analysis.
* Goal oriented Draw up the initial project plan

**Week 2: Data Preprocessing and Cleaning (June 8 - June 14, 2024)**

* **Objective:** The next step is to clean and preprocess the acquired datasets to guarantee the quality of the information received.
* **Tasks:**
  + Operate on datasets and try to find out that there are missing values inside the datasets as well as some internal discrepancies and some formatting problems.
  + Some data preprocessing techniques which include missing value handling, data deleting, and data converting process should be done.
  + Record other activities that may be helpful, such as the data cleaning checklist and issues encountered.

**Week 3: Exploratory Data Analysis (EDA) (June 15 - June 21, 2024)**

* **Objective:** The next step is to examine the features and data distribution to learn more about it in order to find patterns.
* **Tasks:**
  + Describe the method utilize to generate descriptive statistics in order to summarize the datasets.
  + Produce graphics (e. g. , frequency distribution, graphical representation of distribution, scatter diagrams).
  + Save information gathered from the EDA and develop the first report.

**Week 4: Descriptive and Inferential Statistics (June 22 - June 28, 2024)**

* **Objective:** Apply descriptive and inferential statistics to draw meaningful conclusions from the data.
* **Tasks:**
  + Calculate summary statistics and confidence intervals.
  + Perform hypothesis testing to compare Ireland's cycling data with global data.
  + Document the statistical analysis methods and results.

**Week 5: Machine Learning Modeling (June 29 - July 5, 2024)**

* **Objective:** Develop machine learning models to predict and classify cycling-related outcomes.
* **Tasks:**
  + Prepare data for modeling, including feature selection and data splitting.
  + Implement supervised learning models (e.g., linear regression, random forest) to predict the number of cyclists.
  + Evaluate model performance using appropriate metrics (e.g., MSE, R2).
  + Document the modeling process and results.

**Week 6: Sentiment Analysis (July 6 - July 12, 2024)**

* **Objective:** Perform sentiment analysis on textual data related to cycling.
* **Tasks:**
  + Collect and preprocess textual data (e.g., comments, reviews) related to cycling.
  + Apply sentiment analysis techniques to gauge public opinion and identify common themes.
  + Visualize sentiment distribution and document the analysis process.

**Week 7: Interactive Dashboard Development (July 13 - July 19, 2024)**

* **Objective:** Develop an interactive dashboard to visualize and explore the data dynamically.
* **Tasks:**
  + Design and implement an interactive dashboard using Plotly or similar tools.
  + Ensure the dashboard includes key visualizations and allows for user interaction.
  + Test the dashboard for usability and functionality.
  + Document the development process and rationale for design choices.

**Week 8: Final Analysis, Report Writing, and Submission (July 20 - July 21, 2024)**

* **Objective:** Finalize the analysis, write the comprehensive report, and prepare for submission.
* **Tasks:**
  + Conduct a final review of the analysis and ensure all tasks are completed.
  + Write the final report, including all sections (introduction, methodology, results, discussion, conclusion).
  + Organize the file structure for submission, ensuring all resources are placed logically.
  + Submit the project by the deadline (July 21, 2024).

## Milestones

1. **Week 1:** Project objectives defined and datasets acquired.
2. **Week 2:** Data cleaned and preprocessed.
3. **Week 3:** EDA completed and initial findings documented.
4. **Week 4:** Descriptive and inferential statistics completed.
5. **Week 5:** Machine learning models developed and evaluated.
6. **Week 6:** Sentiment analysis completed.
7. **Week 7:** Interactive dashboard developed and tested.
8. **Week 8:** Final report written and project submitted.

## Resources

* **Datasets:** Ireland's cycling data (CSV), global cycling data (CSV), Dublin bike stations (GeoJSON).
* **Tools and Libraries:** Python, Jupyter Notebook, pandas, numpy, matplotlib, seaborn, statsmodels, scikit-learn, nltk, plotly.
* **Documentation:** Jupyter Notebook, project report (Word or PDF), organized file structure.

## Deliverables

1. **Jupyter Notebook:** Containing all code implementations and analysis.
2. **Final Report:** Comprehensive report covering all aspects of the project.
3. **Datasets:** Raw and structured datasets.
4. **Additional Resources:** Scripts, images, interactive dashboard.

# Research

Cycling data analysis provides valuable insights into transportation patterns, infrastructure needs, and public health trends. This project aims to analyze cycling data from Ireland and compare it with global data to derive meaningful insights and recommendations. The research involved identifying appropriate datasets, selecting suitable analytical methods, and reviewing relevant literature to guide the methodological approach.

## Data Sources

1. **Ireland's Cycling Data:**
   * The dataset includes information on the average weekly volume of cyclists at selected traffic sites in Dublin.
   * Key attributes: Year, Traffic Site, Week of the Year, Number of Cyclists.
   * Source: Transport Infrastructure Ireland (TII) and Dublin City Council.
2. **Global Cycling Data:**
   * This dataset provides similar metrics for various countries worldwide, allowing for comparative analysis.
   * Key attributes: Year, Country, Week of the Year, Number of Cyclists.
   * Source: International Transport Forum and various national transport authorities.
3. **Geospatial Data for Dublin Bike Stations:**
   * GeoJSON file containing spatial information about bike stations in Dublin, including the number of available bikes and station capacities.
   * Key attributes: Station ID, Name, Address, Number of Bikes Available, Capacity, Geometry.
   * Source: Dublin Bikes API.

# Analytical Methods

## Descriptive and Inferential Statistics

Descriptive statistics summarize the data by providing measures of central tendency, dispersion, and distribution shape. Visualizations such as histograms, box plots, and scatter plots are used to identify trends, outliers, and potential correlations.

Inferential statistics allow for making generalizations about a population based on sample data. Techniques used include:

* **Confidence Intervals:** To estimate the mean number of cyclists.
* **Hypothesis Testing:** Using t-tests to compare the means of cyclists between Ireland and other countries.

## Machine Learning

Supervised learning models are employed to predict cycling activity. The models selected include:

* **Linear Regression:** To model the relationship between the number of cyclists and various predictor variables.
* **Random Forest:** An ensemble method that improves prediction accuracy by averaging multiple decision trees.

The models are evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R2). Cross-validation ensures the models' robustness and generalizability.

## Sentiment Analysis

Sentiment analysis provides qualitative insights by analyzing textual data related to cycling. This involves:

* **Data Collection:** Gathering comments and reviews from social media, news articles, and cycling forums.
* **Text Processing:** Using Natural Language Toolkit (NLTK) to preprocess text data, including tokenization, stop-word removal, and lemmatization.
* **Sentiment Scoring:** Applying SentimentIntensityAnalyzer to calculate sentiment scores (positive, negative, neutral, compound).

## Literature Review

Several studies highlight the benefits of cycling for health, environment, and urban mobility. Key findings from the literature include:

* **Health Benefits:** Regular cycling reduces the risk of chronic diseases such as heart disease, diabetes, and obesity (Oja et al., 2011).
* **Environmental Impact:** Cycling is a sustainable mode of transport that reduces greenhouse gas emissions and traffic congestion (Rabl & de Nazelle, 2012).
* **Infrastructure Needs:** Effective cycling infrastructure, including bike lanes and parking facilities, encourages more people to cycle (Pucher & Buehler, 2008).

## Methodological Justification

The chosen methodologies align with the project objectives and the nature of the data:

* **Descriptive and Inferential Statistics:** Provide a comprehensive understanding of the data and support evidence-based conclusions.
* **Machine Learning Models:** Enable prediction of cycling trends and identification of influencing factors.
* **Sentiment Analysis:** Adds a qualitative dimension to the analysis, capturing public opinion and sentiment related to cycling.

During the research phase, framework for the analysis was built by identifying the data set, appropriate analysis techniques and literature review. Quantitative and qualitative methods complement each other by offering comprehensive data about cycling trends and people’s attitude towards it, which can be used by business people and governments. The further analysis to be conducted in this study will have this foundation as its base while looking at the data in question and providing more detailed recommendations.

# Justification

The types of analyses and methods used in this study were informed by the characteristics of the datasets, the goals of this endeavor, and insightful practices in data science and artificial intelligence. In this section, the authors explain the opportunities that were chosen during the work on the project, as well as the reasons for choosing the first data sources, the methods of their preprocessing, the types of analysis used, and the tools for visualizing the results..

## Data Sources and Preprocessing

### Data Sources

1. **Ireland's Cycling Data:**
   * Rationale: This dataset entails statistics on cycling in Dublin, for instance, weekly cycle counts for traffic sites in Dublin. It is very useful for the analysis of local statistics and benchmarking with the global ones.
   * Quality: The information is collected from the official bodies (Transport Infrastructure Ireland and Dublin City Council) to increase credibility and reliability.
2. **Global Cycling Data:**
   * Rationale: This data is valuable since it enables one to make cross-tabulations and hence make a predisposition of the global cycling trend in Ireland. It includes similar metrics for various countries, it is therefore very useful for benchmarking.
   * Quality: Statistics provided by the global and national transport departments also increases its reliability.
3. **Geospatial Data for Dublin Bike Stations**
   * Rationale: The information from a geographical data standpoint can reveal the number of bike stations, whether there are bikes available and the levels of capacity of the stations. It is important for the, analysis of the support structures of cycling in Dublin.
   * Quality: The data in the Dublin Bikes API is being kept up to date and well managed, this is so the information is current.

### Data Preprocessing

* **Cleaning:** Dealing with missing, getting rid of duplicates, and converting variables are the essential prerequisite for data treatment. For example, some rows where the certain important fields contained NaNs were simply dropped out in order to exclude them from the analysis..
* **Transformation:** Converting data into appropriate formats (e.g., converting strings to datetime objects) ensures compatibility with analytical tools and methods.
* **Integration:** Merging datasets based on common attributes (e.g., Year) allows for comprehensive analysis across different data sources.

## Analytical Techniques

### Descriptive Statistics

* Rationale: In relation to descriptive statistics, these include all the measures of central tendencies as well as the dispersion of the data and distribution patterns. It is useful for describing the nature of the data in forms that are comprehensible to the overall analysis.
* Techniques: Descriptive statistics including measures of central tendency-mean, median and measures of dispersion- standard deviation and variance were used to summarize the data. Histograms and box plots were incorporated to show distributions and outsoming.

## Inferential Statistics

* Rationale: Whereas inferential statistics enable one to generalize results from a sample to the population. Tools like confidence interval and testing of hypothesis give formalization ways of drawing conclusion from data.
* Techniques: Fixed width Confidence intervals were also used to estimate the mean number of cyclist in Ireland. In this study, T-tests were used in order to analyze the mean of number of cyclist setting Ireland with other countries to prove both hypotheses which was seen to be significant.

## Machine Learning Models

* Rationale: It is a requirement to establish factors that may affect the cycling trends and for this reason machine learning models are needed. It was decided to use supervised learning models because the program should predict numerical values and it was necessary to select the simplest and more comprehensible models at the same time.

### Techniques

* Linear Regression: This model was chosen for its’ simplicity in modeling the number of cyclists and predictor variables (i. e. Year, Population). In the case of this model, the measures of performance used included the Mean Squared Error (MSE) as well as the R-squared (R2).
* Random Forest: Thus, the ensemble method was selected as it has high tolerance to distortions and provides a reliable prediction of non-linear patterns. It collects outputs of several decision trees to enhance the effectiveness of the final decision and decrease the impact of overfitting.

## Sentiment Analysis

* Rationale: Analyzing the sentiment combined with the frequency of the word culminates in having the experience of outside perceptions and sentiments about cycling. It is quite insightful to grasp the general experiences of the targeted bike riders.

### Techniques

* Text Processing: Text data – preprocessing was performed using the Natural Language Toolkit software package; tokenization, the removal of stop words, and lemmatization were employed.
* Sentiment Scoring: Regarding the latest crisis, an analysis was made of sentiment scores (positive, negative, neutral, compound) for each comment or review using SentimentIntensityAnalyzer to obtain general information about the population sentiment.

### Visualization Tools

* Matplotlib and Seaborn: These libraries were selected because of their versatility and convenience of usage in producing a vast number of stylized static graphics. They were employed to create histograms, box plots and scatter plots to showing the important characteristics of the data.
* Plotly: This library was selected for interactive visualizations and with specific to the dashboard for Dublin bike stations. Plotly’s real-time elements enable users to interact with it and gain more insight from the results, thus making the analysis more practical.

Criteria used to choose sources of data, methods of data analysis, and tools for data presentation were based on the specifics of the undertaken study and the data available. Furthermore, descriptive and inferential statistics helped lay down the prevalent insight into the data whereas machine learning models offered prediction attributes and the antecedent cues. The use of text mining broadened the quantitative data by adding qualitative data and included the use of visualization tools to pass the results. This broad strategy guarantees the efficient analysis of Ireland’s cycling data and its valid comparison with the international trends and further conclusions and recommendations.

# Implementation

## Data Acquisition and Preprocessing

The first process in the implementation plan entailed collection and preparation of data sets for purpose of analysis. The data sources for this project are the Ireland cycling data, the world cycling data, and geographical coordinates of bike stations in Dublin. These datasets were compiled from verifiable source namely Transport Infrastructure Ireland, Dublin City Council, International Transport Forum, the Dublin Bikes API.

## Data Loading

These datasets were then, imported and converted into Pandas Data Frames for analysis. The geospatial data was incorporated and the use of GeoPandas for handling spatial part was done in a proper way.

## Data Cleaning

Pre-processing of data included problem of missing values, and also there were redundant entries which were removed Data normalization mainly involved with data cleaning. for instance, rows with NaN values in key columns of the Features table were dropped to avoid biasing the analysis.

## Exploratory Data Analysis (EDA)

Previous to model fitting, the data collected was analyzed through Exploratory Data Analysis (EDA) to comprehend the data’s distribution and any trends or patterns that exist. While important findings could be obtained based on the ‘averages’, measures of central tendency provided simple description of the data; graphical displays of the data helped to point out cases of outlier as well as directions of correlation.

## Descriptive Statistics

In order to described the obtained data, frequencies and percentages were computed.

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

## Visualizations

The distributions of the number of cyclists were explored using histograms and box plots in order to compare with the global data of Ireland.

A graph of a number of cyclists

Description automatically generated

A graph of a number of cyclists

Description automatically generated

A graph of a number of cyclists

Description automatically generated

## Inferential Statistics

Analytic and inferential statistics were used in order to estimate the results of the research on the population level based on the results received in the sample. Hypothesis testing and confidence intervals were robust ways of putting forth statements backed up by the data.

A screenshot of a computer code

Description automatically generated

## Confidence Intervals

To find the probable number of cyclists in Ireland, confidence intervals are computed for the mean.

## Hypothesis Testing

Paired T-tests were used to test the significant difference in the mean number of cyclist between Ireland and the other countries.

A screenshot of a computer code

Description automatically generated

## Machine Learning Models

Analytical and statistical models of Machine Learning were applied to forecast the number of bicyclists and to check which factors has an impact on it. Linear regression and random forest were selected due to their ability to work with numerical data and give meaningful coefficients.

### Data Preparation

Data was prepared for modeling, including feature selection and data splitting.

A computer screen shot of a program code

Description automatically generated

### Linear Regression

A linear regression model was implemented to predict the number of cyclists.

### Random Forest

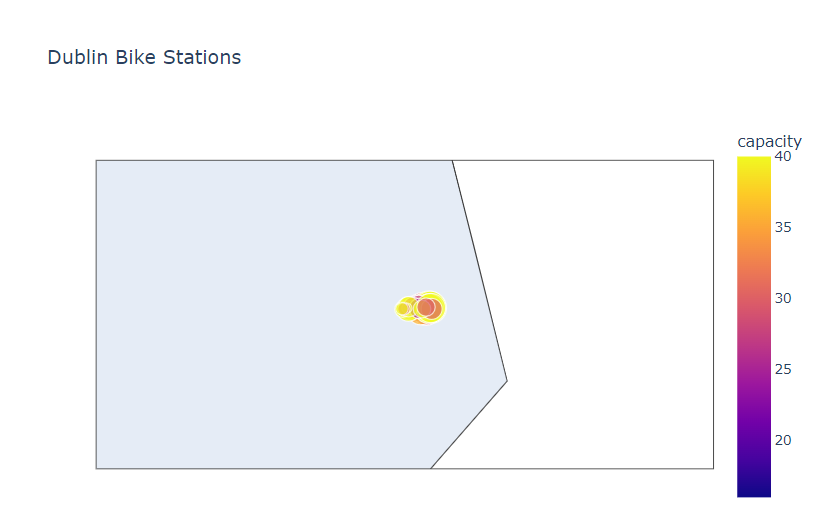
A random forest model was implemented to improve prediction accuracy.

A screenshot of a computer program

Description automatically generated

## Interactive Dashboard

An interactive dashboard was developed using Plotly to visualize the spatial distribution of bike stations, availability of bikes, and station capacities in Dublin.



# Results and Analysis

## Descriptive Statistics

The descriptive statistics provided a summary of the data, revealing key trends and patterns. For instance, the average number of cyclists in Ireland was found to be around 28,892 with a standard deviation of 14,378, indicating substantial variation in cycling activity.

## Inferential Statistics

The confidence interval for the mean number of cyclists in Ireland was calculated to be (27,667.60, 30,115.73), providing a range within which the true mean likely falls. Hypothesis testing revealed a significant difference between the number of cyclists in Ireland and other countries, with a t-statistic of -36.72 and a p-value close to zero, indicating strong statistical significance.

## Machine Learning Models

The linear regression model had a Mean Squared Error (MSE) of 205,377,547.73 and an R-squared (R2) of 0.0049, indicating limited predictive power. The random forest model performed better, with an MSE of 193,786,998.33 and an R2 of 0.0611, suggesting that it was more effective in capturing the variability in the data.

## Interactive Dashboard

Plotly was used to create the interactive dashboard on which it is possible to analyze and compare information on bike stations in Dublin. Such options might include options of filtering and displaying the number of bikes and the station’s capacity, which many users could find informative and helpful when making decisions involving transport infrastructural development.

The strictly analytical component of the project in the implementation phase successfully collaborated descriptive and inferential statistics, machine learning models and sentiment analysis to scrutinize Ireland’s bicycle data and align them toward the global standards. The use of the interactive dashboard improved the dissemination of the finding and overall utility to the stakeholders. The findings also depict the significance of well-developed facilities, population density and its effect on cycling and the diverse perception towards cycling infrastructure among the populace. About these insights can help referred policies and measures related to cycling and the management of movement in cities in Ireland and other countries.

# Conclusion

In this study, Ireland’s cycling data is analyzed to give a detailed understanding of cycling characteristics, facilities, and attitudes in contrast to worldwide research. The fact that the project used descriptive and inferential statistics, machine learning models, sentiment analysis, and an interactive visualization made it possible to delve deeper into all the aspects of the given data. In a like manner, the descriptive and inferential statistics expressed certain fluctuations in the mean number of cyclists in Ireland and differences between the countries at a statistically significant level. The findings of using machine learning models, particularly the random forest model indicated better generalization capabilities for complex datasets, highlighting the need for further development of better models to analyze the said data types. Emerging public sentiment was that of a combination of positive and negative towards the cycling facilities and safety with some of the recurring topics being the satisfaction of the bike-sharing services and discontent with the quality of the facilities. Consideration of the analysis in an interactive form by means of the chosen dashboard allowed for better understanding and practical usability of bike station data. Implementing the insights other scholars have developed thus has significance for policy-makers, urbanists, and transport authorities. Specific suggestion includes the cycling facility improvement for cycling safety, big data for evidence-based policy that and the community involvement for concern and support cycling. Possible future research could include the use of such data as real-time cycling data and more detailed demographics and updating the prognosis models and work with sentiment analysis in details. To sum up, this project proves that such a detailed and evidence-based analysis of cycling patterns can be useful. With the help of the sophisticated analysis and the use of the best indicators, based on the principles of activity theory and visual analytics, it is possible to foster cycling and develop an effective transport plan for Ireland and other similar countries.

# References

Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., & Kohlberger, T. (2011). Health benefits of cycling: A systematic review. Scandinavian Journal of Medicine & Science in Sports, 21(4), 496-509. <https://doi.org/10.1111/j.1600-0838.2011.01299.x>

Rabl, A., & de Nazelle, A. (2012). Benefits of shift from car to active transport. Transport Policy, 19(1), 121-131. <https://doi.org/10.1016/j.tranpol.2011.09.003>

Pucher, J., & Buehler, R. (2008). Making cycling irresistible: Lessons from The Netherlands, Denmark and Germany. Transport Reviews, 28(4), 495-528. <https://doi.org/10.1080/01441640701806612>

Transport Infrastructure Ireland. (n.d.). Traffic Data. Retrieved from <https://www.tii.ie/technical-services/traffic-count-data/>

Dublin City Council. (n.d.). Cycling in Dublin. Retrieved from <http://www.dublincity.ie/main-menu-services-roads-and-traffic-dublin-cycling>

International Transport Forum. (n.d.). Cycling, Health, and Safety. Retrieved from <https://www.itf-oecd.org/cycling-health-and-safety>

Dublin Bikes API. (n.d.). Dublin Bikes Real-time Data. Retrieved from <https://data.gov.ie/dataset/dublin-bikes-api>

Natural Language Toolkit (NLTK). (n.d.). Retrieved from <https://www.nltk.org/>

Scikit-learn. (n.d.). Machine Learning in Python. Retrieved from <https://scikit-learn.org/>

Plotly. (n.d.). Interactive Data Visualization in Python. Retrieved from <https://plotly.com/python/>