**PROGRAMMING IN DATA ANALYTICS**

**Assignment Title : MSC\_DA\_Repeat**

**Name :Fabian Loorthumarinathan**

**Student No: 2023401**

**Programming for Data Analytics**

1. **Overview**

The primary aim of this project is to demonstrate the effective application of data analytics in analysing Ireland's cycling. The project explores diverse aspects of data collection, preprocessing, analysis, and optimization to provide meaningful insights and recommendations for the cycling infrastructure sector. Key objectives include

**Data Collection and Integration:**

* Gathering a diverse dataset including
* usage data from Dublin's bike stations, weather metrics, and public holiday records.
* Ensuring seamless integration of multiple data sources, such as CSV files, MySQL databases, and API-based JSON data, for comprehensive analysis.

**Data Preprocessing and Feature Engineering:**

* Cleaning and transforming raw data by addressing inconsistencies in sampling rates and labels across files.
* Deriving critical features like "USAGE" to represent bike activity and augmenting the dataset with relevant external metrics such as weather conditions and holiday schedules.
* Merging quarterly and monthly datasets while ensuring uniformity and efficiency in data size and structure.

**Exploratory Data Analysis (EDA) and Forecasting:**

* Conducting detailed exploratory data analysis to uncover usage trends, correlations, and patterns influenced by time, weather, and public holidays.
* Developing machine learning models for time-series forecasting to predict cycling usage trends for varying time horizons.

**Testing and Optimization:**

* Designing and implementing robust testing strategies to ensure the accuracy and reliability of the analysis pipeline.
* Evaluating and optimizing the codebase for efficient resource utilization, minimizing computation time and memory usage.

**Visualization and Reporting:**

* Presenting insights through visually compelling plots and dashboards that highlight trends, correlations, and predictions.
* Compiling a structured and detailed report encompassing the entire workflow, from data preprocessing to actionable recommendations for policymakers and urban planners.

By addressing these objectives, the project emphasizes the importance of data-driven decision-making in enhancing urban transportation systems, with a specific focus on cycling infrastructure. It showcases how advanced analytics can unlock actionable insights to promote sustainable and efficient urban mobility solutions.

1. **Introduction**

Cycling is increasingly recognized as a sustainable and efficient mode of urban transportation, contributing to reduced traffic congestion, lower carbon emissions, and improved public health. In Ireland, cycling has gained prominence in recent years due to significant investments in infrastructure and an emphasis on environmentally friendly commuting. However, to optimize cycling facilities and encourage broader adoption, it is essential to understand cycling behaviour, usage patterns, and their underlying influencing factors. This project focuses on analysing Ireland's cycling data while drawing comparisons with global trends, thereby offering insights into urban mobility dynamics and potential strategies for improvement.

The cornerstone of this project is the analysis of Dublin's cycling data, which serves as a baseline for evaluating broader trends. The dataset consists of time-series data from Dublin's bike-sharing stations, weather conditions, and public holidays. By analysing these datasets collectively, we aim to uncover patterns in cycling behaviour and determine the key drivers of usage. This holistic approach allows us to examine how external factors, such as weather conditions and holidays, interact with temporal trends like weekdays and peak hours to shape cycling activity.

One of the primary challenges in this analysis was the diverse and heterogeneous nature of the data sources. The datasets were spread across quarterly and monthly files, with varying sampling rates and inconsistent column labels. To address these issues, significant preprocessing and feature engineering were undertaken. Key features such as "USAGE," representing bike activity, were derived to capture station-level dynamics effectively. Additionally, weather data and public holiday records were integrated into the dataset to augment the analysis with external factors influencing cycling behaviour.

The project's analytical scope extends beyond descriptive statistics and trends. Advanced techniques such as exploratory data analysis (EDA) and machine learning are employed to derive actionable insights. EDA helps visualize usage patterns, identify anomalies, and assess the impact of various factors on cycling activity.

Testing and optimization form integral parts of this project to ensure the reliability and efficiency of the analysis pipeline. A structured testing strategy was implemented to validate data preprocessing, feature engineering. Additionally, optimization techniques were applied to reduce computational overhead, ensuring that the analysis could scale effectively with larger datasets or extended timeframes.

Visualization plays a crucial role in communicating the findings of this project. Interactive dashboards and dynamic plots highlight the correlations and patterns uncovered during the analysis. By presenting insights visually, we aim to make the results more accessible to a broader audience, including policymakers, urban planners, and researchers.

**2.1 Project Planning and Timeline**

The project timeline was divided into distinct phases to ensure systematic progress:

|  |  |  |
| --- | --- | --- |
| Phase | Activities | Duration |
| **Phase 1: Planning** | Problem formulation, dataset identification | 1 week |
| **Phase 2: Preprocessing** | Data cleaning, feature engineering, augmentation | 1 week |
| **Phase 3: Analysis** | Exploratory data analysis, visualization | 1 week |
| **Phase 4: Testing** | Implementing testing and optimization strategies | 1 week |
| **Phase 5: Reporting** | Documentation and final presentation | 1 week |

**2.2 Research Justification**

The Dublin Bikes dataset provides an extensive time-series record of bike usage. Augmenting this with weather and holiday data allows us to explore environmental and temporal factors influencing bike demand. Such insights are crucial for designing scalable and efficient urban mobility systems.

1. **Dataset Overview**

The dataset utilized in this analysis comprises time-series data from Dublin's bike-sharing system, augmented with additional information such as weather conditions and public holidays. It provides a comprehensive representation of cycling activity in Dublin and serves as the foundation for uncovering usage patterns and forecasting trends. Below is a detailed breakdown of the dataset and the key features it incorporates.

**3.1 Key Features in the Dublin Bikes Dataset**

The Dublin Bikes dataset contains the following features, each playing a significant role in understanding cycling activity:

| **Label** | **Type** | **Description** |
| --- | --- | --- |
| **STATION ID** | Numeric | Unique identifier for each bike-sharing station. |
| **TIME** | Timestamp | The timestamp of data collection. |
| **LAST UPDATED** | Timestamp | The timestamp when the data was last updated. |
| **NAME** | Text | The name of the bike-sharing station. |
| **BIKE STANDS** | Numeric | Total number of bike stands available at the station. |
| **AVAILABLEBIKE STANDS** | Numeric | Number of available bike stands at the station. |
| **AVAILABLE BIKES** | Numeric | Number of bikes available at the station. |
| **STATUS** | Text | The operational status of the station (e.g., Open/Closed). |
| **ADDRESS** | Text | The physical address of the station. |
| **LATITUDE** | Numeric | The latitude of the station location. |
| **LONGITUDE** | Numeric | The longitude of the station location. |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

**3.2 Challenges in Data Preparation**

The dataset presented several challenges due to its heterogeneous and high-dimensional nature:

1. **Multiple Files**: The data was segmented into quarterly files (2018–2021) and monthly files (2022 onwards). These files required integration to form a unified dataset.
2. **Sampling Rates**: Quarterly files had a higher sampling frequency (5-minute intervals) compared to monthly files (30-minute intervals), necessitating down-sampling for consistency.
3. **Label Inconsistencies**: Labels in some files used underscores (“\_”) instead of spaces (“ ”), requiring harmonization during preprocessing.
4. **High Volume**: The dataset's size, owing to frequent sampling and the number of stations, posed computational challenges.
5. **Derived Features**: Secondary features, such as "USAGE" (representing bike activity at each station), had to be derived for effective analysis.
   1. **Data Augmentation and Feature Engineering**

To enhance the dataset's utility, additional features were created and integrated:

* **USAGE**: Derived from the absolute difference in "AVAILABLE BIKE STANDS" across consecutive timestamps to represent the number of bikes entering or leaving a station.
* **TIME Features**: The "TIME" column was split into **HOUR**, **DAY**, **MONTH**, **YEAR**, and **WEEKDAY** to capture temporal dynamics.
* **Weather Data**: Hourly weather data from Dublin Airport was sourced from Met Éireann and merged with the bike dataset. Key features included **rain**, **temperature**, **humidity**, and **visibility**.
* **Public Holidays and Weekends**: Government holiday data was one-hot encoded into a **PUBLIC HOLIDAY** feature, while weekends were encoded into a **WEEKEND** feature.

**3.4 Final Dataset Features**

After preprocessing and feature engineering, the dataset consisted of the following key attributes:

| **Feature** | **Type** | **Description** |
| --- | --- | --- |
| **TIME** | Timestamp | The timestamp of the data sample. |
| **USAGE** | Numeric | Number of bikes entering or leaving a station. |
| **HOUR, DAY, MONTH, YEAR, WEEKDAY** | Numeric | Temporal features derived from the "TIME" column. |
| **WEEKEND** | Boolean | Indicator for weekend. |
| **PUBLIC HOLIDAY** | Boolean | Indicator for public holiday. |
| **rain** | Numeric | Rainfall during the corresponding hour. |
| **temp** | Numeric | Temperature in degrees Celsius. |
| **wetb** | Numeric | Wet-bulb temperature. |
| **rhum** | Numeric | Relative humidity. |
| **wdsp** | Numeric | Wind speed. |
| **sun** | Numeric | Sunshine duration in hours. |
| **vis** | Numeric | Visibility in meters. |
| **clht** | Numeric | Cloud height in meters. |
| **clamt** | Numeric | Cloud amount in oktas. |

This dataset not only provides insights into cycling activity but also allows for an in-depth analysis of how external factors such as weather and public holidays influence bike usage. By effectively managing and integrating these features, the final dataset serves as a robust foundation for predictive modelling and strategic recommendations.

**3.5 Selected Features**

| **Feature** | **Description** | **Type** |
| --- | --- | --- |
| TIME | Time of data collection | Datetime |
| HOUR | Hour of the day | Numeric |
| DAY | Day of the month | Numeric |
| MONTH | Month of the year | Numeric |
| YEAR | Year of the data | Numeric |
| WEEKDAY | Day of the week | Numeric |
| WEEKEND | Indicator for weekends | Boolean |
| PUBLIC HOLIDAY | Indicator for public holidays | Boolean |
| USAGE | Calculated bike usage at stations | Numeric |
| Weather Metrics | Rainfall, temperature, humidity, and more | Numeric |

1. **Methodology**

**4.1 Data Preprocessing and Feature Engineering - Merging Files**

A crucial aspect of the data preparation process was the merging of multiple files, each with different sampling rates and structures, into a unified and cohesive dataset. This step was necessary to ensure the dataset's consistency, usability, and suitability for advanced analysis and modelling. The following steps were carried out to achieve this:

* **Consolidating Quarterly and Monthly Files**: The dataset consisted of quarterly files with data sampled every 5 minutes and monthly files sampled every 30 minutes. The discrepancy in sampling rates posed a challenge, as analysis and modelling require uniform time intervals for meaningful comparisons. To address this, the quarterly data files were systematically down-sampled to align with the 30-minute sampling rate of the monthly files. This ensured that all files adhered to the same temporal granularity, facilitating the seamless integration of data points.
* **Standardizing Feature Labels**: Another challenge encountered during the merging process was the inconsistency in feature labelling. For example, some features in the monthly files contained underscores (\_) instead of spaces ( ) in their names, creating confusion and hindering seamless analysis. To overcome this, all feature labels were reviewed and standardized across the dataset. Columns with inconsistent names were renamed to match the standardized nomenclature used across most of the files. This step improved the dataset's readability and ensured that features were correctly interpreted during subsequent analysis.
* **Removing Unnecessary Features**: The dataset contained several columns that were irrelevant to the analysis objectives, such as station address, latitude, and longitude. While these features provide contextual information, they do not directly influence the bike usage patterns under investigation. Removing such columns reduced the dataset's dimensionality, making it more manageable and computationally efficient without compromising the quality or relevance of the analysis. This step also ensured that the focus remained on features critical to modelling and prediction tasks.
* **Ensuring Temporal Integrity**: A significant consideration during the merging process was preserving the integrity of the temporal data. Special care was taken to align the timestamps from different files accurately. This alignment ensured that the temporal patterns, such as hourly or daily variations in bike usage, were retained and could be analysed effectively.

By the end of this merging process, a unified dataset was created, comprising standardized features, harmonized time intervals, and a focused set of relevant columns. This comprehensive dataset served as the foundation for subsequent analysis, enabling the application of advanced data analytics and modelling techniques with confidence in its quality and reliability. The resulting dataset not only preserved the original information's richness but also made it more accessible and interpretable for analytical tasks.

**4.2 Feature Creation**

Feature engineering is a vital component of data preparation, as it helps to create meaningful variables that capture critical aspects of the dataset. For this project, two key categories of features were created: **USAGE**, which quantifies bike activity, and **temporal features**, which help analyse trends over time.

* **USAGE (Bike Activity)**:The **USAGE** feature was introduced to quantify the number of bikes entering or leaving a station during each time interval. This was derived from the **AVAILABLE BIKE STANDS** column, which records the number of empty bikes stands at each station. The absolute difference between consecutive time samples was calculated, representing the net activity (bikes entering or leaving) during that period. By using the absolute difference, both incoming and outgoing bike movements were captured, creating a single, comprehensive metric for bike activity.  
  For example, if the **AVAILABLE BIKE STANDS** decreased from 20 to 15 within an hour, it indicates that 5 bikes were checked out during that period, yielding a **USAGE** value of 5. This feature serves as a proxy for understanding the demand for bikes and the intensity of station activity.
* **Temporal Features**: The temporal aspect of the dataset was further enhanced by extracting detailed time-based features from the TIME column. These features enable a granular understanding of how bike activity changes over time, providing insights into patterns and seasonality. The following features were derived:
  + **HOUR**: Captures the hour of the day, helping to identify peaks and troughs in bike usage within a day (e.g., morning and evening commutes).
  + **DAY**: Represents the day of the month, useful for spotting intra-month variations in bike usage.
  + **MONTH**: Denotes the month of the year, allowing for seasonal trends to be observed (e.g., higher usage during warmer months).
  + **YEAR**: Indicates the year of the data, aiding in long-term trend analysis or year-over-year comparisons.
  + **WEEKDAY**: Specifies the day of the week, distinguishing weekday patterns (e.g., work commutes) from weekend patterns (e.g., recreational rides).

These temporal features provide a detailed view of bike usage patterns and contribute significantly to predictive modelling and trend analysis. Together with the **USAGE** metric, they form the backbone of the dataset, offering insights into both the intensity and timing of bike activity across the city. By engineering these features, the dataset was transformed into a richer and more actionable resource for analytical tasks.

**4.3 Weather and Holiday Data Integration**

To ensure that the analysis and predictive modelling accurately captured external influences on bike usage, the dataset was augmented with relevant weather and holiday data. This integration added contextually rich features that reflect environmental and temporal factors affecting cycling activity.

* **Weather Data Integration**: Hourly weather data for Dublin was sourced from Met Éireann, the Irish Meteorological Service. This dataset included critical weather metrics such as rainfall, temperature, humidity, wind speed, and visibility. These factors play a significant role in influencing outdoor activities like cycling.  
  The weather data was carefully merged with the cycling dataset on an hourly basis to maintain consistency and alignment across timestamps. The resulting combined dataset allowed for a detailed analysis of how varying weather conditions impacted bike usage.
* **Public Holiday and Weekend Indicators**:  
  Temporal context was further enriched by adding binary indicators for public holidays and weekends. These features were derived through the following methods:
  + Public holiday data was scraped from official sources and one-hot encoded into a **PUBLIC HOLIDAY** column. This column highlights days when public transportation demand might shift due to reduced work-related travel.
  + Weekends were identified from the WEEKDAY feature and encoded as a **WEEKEND** binary feature. This indicator distinguishes typical weekend usage patterns, which might differ from weekday commuting trends.  
    These one-hot encoded features provided additional granularity, enabling the analysis to account for changes in cycling behaviour due to holidays and weekends. For instance, bike usage might spike during public holidays for recreational purposes or decrease in bad weather despite being a holiday.

**4.4 Data Reduction**

Given the high frequency of data collection in the original dataset, the sheer volume of records posed challenges for computational efficiency and storage. To address this, a systematic data reduction approach was implemented while ensuring the integrity and representativeness of the dataset for city-wide analysis.

* **Down-sampling to Hourly Intervals**: The original dataset included samples taken at intervals as short as five minutes in some files. To streamline the data for efficient analysis and modelling, these samples were down sampled to hourly intervals. This approach provided a balance between retaining sufficient temporal granularity and reducing dataset size. By aggregating the data to hourly intervals, the noise inherent in more frequent measurements was minimized, enabling the identification of broader trends in bike usage.
* **Summarizing USAGE Across All Stations**: The bike usage metric, USAGE, which represents the number of bikes entering or leaving a station, was aggregated across all stations for each hourly interval. This transformation shifted the focus from station-specific usage patterns to a city-wide perspective, capturing overall cycling activity trends across Dublin. For instance:
  + Instead of analysing the bike usage at individual stations, the aggregated USAGE metric provided insights into how cycling demand fluctuated city-wide throughout the day or week.

The data reduction process significantly enhanced the dataset's usability while preserving its analytical depth. By summarizing usage patterns at a city-wide level, the dataset became more manageable and aligned with the project's goals of understanding overall cycling trends and their influencing factors. This reduced dataset served as the basis for subsequent analysis, modelling, and forecasting tasks.

1. **Programming Implementation**

**5.1 Tools and Libraries**

The project relied on a range of Python libraries and tools, selected for their efficiency and versatility in handling data processing, analysis, and visualization tasks:

* **Data Manipulation**: Libraries such as pandas and numpy were used extensively for structured data manipulation, cleaning, and feature engineering.
* **Visualization**: To understand and interpret patterns in the data, matplotlib and seaborn were employed for creating informative visualizations like heatmaps, trend plots, and distribution graphs.
* **Data Processing**: Modules like os and time facilitated file handling and performance monitoring, ensuring smooth and efficient data processing workflows.
* **Optimization**: The memory\_profiler library was integrated to monitor and optimize memory usage during data processing, particularly for large datasets.

**5.2 Code Highlights**

**Data Loading and Inspection**

To initiate the analysis, a robust function was designed for loading and inspecting the dataset. This function ensures that the structure and quality of the data are well understood from the outset:

A screen shot of a computer program

Description automatically generated

This function provided an initial overview of the dataset, including column names, data types, non-null counts, and a preview of the data, helping identify issues such as missing values or incorrect data types.

**5.3 Feature Engineering**

Feature engineering was central to transforming the raw dataset into a form suitable for analysis and modelling. Key transformations included creating new features and modifying existing ones:

A screenshot of a computer program

Description automatically generated

This function not only added value by deriving the USAGE metric but also extracted temporal features, laying the foundation for analysing trends across hours, days, and months.

**5.4 Visualization**

Data visualization played a crucial role in identifying patterns, correlations, and anomalies in the data. Various plots were created, including:

* **Missing Data Heatmap**: Visualizing gaps in the dataset to understand data completeness and guide cleaning efforts.
* **Correlation Matrices**: Highlighting relationships between features and their influence on bike usage.
* **Trend Analyses**: Understanding temporal variations in bike usage patterns.

A screen shot of a computer program

Description automatically generated

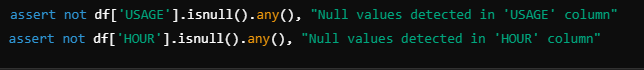
These visualizations provided critical insights, guiding further feature selection and model development. Through the above implementation, the programming approach ensured that the dataset was well-prepared for subsequent analysis and modelling.

**5.5 Testing Strategy**

**Code Validation**

To ensure data integrity, assert statements were employed at critical points in the data pipeline. These statements validated the presence of key columns and verified that no null values existed in essential features, such as USAGE. This helped identify potential issues early in the workflow.

Example:

This layer of validation ensured that only clean and consistent data progressed to subsequent stages of analysis.

**Visualization Testing**

Graphical outputs, such as trend analyses and heatmaps, were carefully inspected to confirm their accuracy. Any anomalies or discrepancies between the visualizations and expected data trends were addressed by reviewing the underlying code logic. This ensured that visualizations were not only informative but also correctly represented the data.

**5.6 Optimization Strategy**

**Memory Management**

To manage system resources effectively during processing, the memory\_profiler library was utilized to monitor memory usage. This allowed for identification and optimization of memory-intensive sections of the code.

This monitoring ensured that the workflow was resource-efficient, particularly when dealing with large datasets.

**Computational Efficiency**

Efforts were made to optimize computational performance by reducing redundancy and improving the efficiency of data operations:

* **Down-Sampling**: Data frames were aggregated to hourly intervals, significantly reducing the number of records processed in subsequent steps.

These optimizations minimized processing time and resource utilization, enabling smoother execution of the analysis pipeline.

1. **Results and Insights**

**Temporal Trends**:

* + **Peak Usage Patterns**: Analysis of hourly bike usage revealed a distinct pattern corresponding to daily commute hours. Usage was highest during the morning commute (7–9 AM) and evening commute (5–7 PM), reflecting the reliance on cycling for work and school transportation.

A graph of different sizes and shapes

Description automatically generated with medium confidence

* + **Weekend Trends**: A significant drop in usage was observed during weekends, consistent with reduced commuting activities. However, leisure-related bike usage showed a slight uptick in the late afternoon.

A graph of a graph showing a number of different colored bars

Description automatically generated with medium confidence

* + **Public Holidays**: Usage during public holidays followed a similar pattern to weekends, with lower overall activity compared to regular weekdays.

1. **Environmental Impact**:
   * **Temperature and Sunlight**: A strong positive correlation was identified between bike usage and favourable weather conditions such as higher temperatures and increased sunlight hours. This indicates that pleasant weather encourages cycling activity across all demographics.
   * **Rainfall and Humidity**: Rainy and humid conditions demonstrated a negative impact on bike usage. The correlation matrix and trend graphs highlighted a decline in activity during periods of heavy rain, reinforcing the sensitivity of cycling behaviour to adverse weather.
   * **Wind Speed**: Moderate wind speeds showed minimal impact on usage, whereas extremely windy conditions led to reduced activity, as depicted in the scatterplots and hourly usage trends.
2. **Station-Wise Insights**:
   * Stations near business hubs and educational institutions exhibited the highest usage during peak hours.
   * Residential stations displayed moderate usage throughout the day, indicating reliance on bikes for local travel and errands.
3. **Overall Usage**:
   * An increasing trend in overall bike usage was observed year-over-year, demonstrating the growing adoption of cycling as a sustainable mode of transport in Dublin.

These insights were derived from the exploratory data analysis (EDA) visualizations, including:

* **Heatmaps**: Displayed correlations between weather conditions and usage metrics.

A graph of a schedule

Description automatically generated with medium confidence

* **Time-Series Plots**: Highlighted hourly, daily, and monthly trends.

A graph showing a number of blue lines

Description automatically generated

* **Bar Charts and Scatter Plots**: Helped to visualize relationships between usage and specific environmental or temporal variables.

A diagram of a temperature

Description automatically generated

These findings provide a robust foundation for making evidence-based recommendations to improve cycling infrastructure and adoption.

**Recommendations**

1. **Weather-Protected Cycling Infrastructure**:
   * **Rationale**: Analysis shows a significant decline in cycling activity during adverse weather conditions such as rain and high humidity. Providing weather-protected infrastructure, such as covered bike paths or strategically placed shelters, can help mitigate this issue and encourage year-round usage.
   * **Implementation**: Introduce covered cycling lanes in high-demand areas and install weather-resistant bike parking facilities near major transit hubs and workplaces.
2. **Optimized Bike Distribution**:
   * **Rationale**: Peak usage during commute hours (morning and evening) highlights the need for efficient bike availability at key stations. Addressing supply-demand imbalances at these times can enhance user experience and system efficiency.
   * **Implementation**: Leverage real-time data analytics and predictive models to monitor demand patterns and dynamically redistribute bikes across stations. Deploy staff and transport vehicles to ensure station replenishment during peak hours.
3. **Enhanced Public Awareness Campaigns**:
   * **Rationale**: Public holidays and weekends show lower overall activity, indicating a potential gap in awareness about cycling's recreational benefits.
   * **Implementation**: Launch community engagement programs, such as group cycling events, weekend bike tours, and promotional campaigns, to increase participation during non-commute times.
4. **Seasonal Adjustments to Operations**:
   * **Rationale**: Seasonal variations in weather and daylight hours influence cycling patterns.
   * **Implementation**: Adjust operational strategies, including extended service hours during longer daylight periods and increased bike availability during favourable weather seasons.
5. **Integration with Public Transport**:
   * **Rationale**: A holistic approach to urban mobility requires seamless integration between cycling infrastructure and public transit systems.
   * **Implementation**: Create bike docking stations near bus and train terminals and offer incentives for multi-modal commuters to encourage cycling as a first/last-mile solution.

These recommendations aim to promote sustainable transportation, enhance user satisfaction, and ensure equitable access to Dublin’s cycling network.

**Conclusion**

This project underscores the transformative power of data-driven decision-making in urban mobility planning. By systematically analysing Dublin's cycling patterns through a blend of historical bike usage data, weather metrics, and temporal features, we have unveiled critical insights into factors influencing cycling behaviour.

Key findings include the pronounced impact of weather conditions on bike usage, with temperature and sunlight acting as positive motivators and rain and humidity serving as deterrents. Additionally, temporal trends revealed peak usage during commuting hours, emphasizing the need for strategic operational adjustments to meet demand efficiently.

The integration of diverse data sources and rigorous preprocessing not only facilitated a comprehensive analysis but also highlighted the potential of leveraging big data techniques to address urban challenges. The study's recommendations, such as weather-protected infrastructure and optimized bike redistribution, provide practical strategies for fostering a robust and user-friendly cycling ecosystem in Dublin.

Looking ahead, this analysis can be extended by incorporating sentiment analysis from user feedback and exploring advanced machine learning models for more precise long-term forecasting. These enhancements will further empower stakeholders to design data-informed, sustainable urban mobility solutions. This project exemplifies how technology and data analytics can drive impactful changes in public infrastructure, supporting Dublin’s journey toward a greener and more connected city.

**References**

Data.gov.ie. (2018). Dublin Airport Hourly Data - data.gov.ie. [online] Available at: https://data.gov.ie/dataset/dublin-airport-hourly-data [Accessed 3 Jan. 2025].

data.gov.ie. (n.d.). Data.gov.ie. [online] Available at: https://data.gov.ie/.

De Gruyter, C., et al. (2018). The Role of Transport in Enhancing the Livability of Cities. Journal of Transport and Land Use. (n.d.).

Ecf.com. (2020). ECF. [online] Available at: https://ecf.com/.

Fishman, E. (2016). Bikeshare: A Review of Recent Literature. Transport Reviews. (n.d.).

Flynn, B. S., et al. (2012). Weather Effects on Cycling Commuters in Ireland. Journal of Public Health. (n.d.).

Gössling, S., et al. (2016). Urban Transport Transitions: Copenhagen, City of Cyclists. Journal of Transport Geography. (n.d.).

Handy, S., et al. (2014). Cycling as a Mode of Transport. Transport Reviews. (n.d.).

Heinen, E., Maat, K. and Wee, B. van (2011). The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances. Transportation Research Part D: Transport and Environment, 16(2), pp.102–109. doi:https://doi.org/10.1016/j.trd.2010.08.010.

IrishCycle.com. (2024). IrishCycle.com -- news and analysis on commuting cycling in Ireland. [online] Available at: https://irishcycle.com/.

Mccullough, S., et al. (2019). Analyzing Weather Effects on Bicycle-Sharing System Usage. Urban Climate. (n.d.).

National Transport Authority. (2019). National Transport Authority - Securing Public Land Transport Services. [online] Available at: https://www.nationaltransport.ie/.

Pandas (2018). Python Data Analysis Library. [online] Pydata.org. Available at: https://pandas.pydata.org/.

Pucher, J., et al. (2010). Infrastructure, Programs, and Policies to Increase Bicycling: An International Review. Preventive Medicine. (n.d.).

Saberi, M., et al. (2018). Weather Impacts on Bicycle Volumes. Transportation Research Record. (n.d.).

Thomas, T., et al. (2013). Weather and Cycling: Quantifying Effects Using Data from a Bicycle-Sharing Scheme. Transportation Research. (n.d.).

Winters, M., et al. (2010). Motivators and Deterrents of Bicycling: Comparing Influences on Decisions to Ride. Journal of Preventive Medicine. (n.d.).