**1.1 Problem Statement**

Nearly thirty percent of current greenhouse gas emissions (GHG) are associated with transport. With increased urbanization and densification, promoting active transport is a needed step towards emission reduction, but also to improve the health and wellbeing of city residents. In order to help motivate active transport and in particular cycling, an understanding of the reasons behind the choice to ride a bike is essential to guide urban planners, policy makers and stakeholders in the transport field to make the right design decisions.

In this context, the proposed study aims to investigate the relationship between outdoor thermal comfort and the decision to ride a bike. The study will focus on Cambridge, MA, and, in particular, the three most popular and distinct bike routes for Bluebikes, Boston’s bike sharing service. Data sets include a combination of publicly accessible data and data extracted from an annual hourly simulation of outdoor thermal comfort along those routes. The findings of the study would ideally help identify whether outdoor comfort affects the choice to ride a bike, by how much it does, and build a predictive model augmented with other logical parameters such as bike availability.

**1.2 Project Goals**

* Perform an exploratory data analysis on the dataset combining cycling data, weather data and outdoor thermal comfort simulation data to identify patterns and relationships.
* Build a predictive model for number of bike trips per hour and bike route choice based on logistical parameters (dock-ability, rideability of stations), campus use (when MIT is in session), weather parameters (precipitation) and outdoor thermal comfort along individual routes.

**1.2 Data Overview**

In this project, we hope to understand the relationship between thermal comfort, measured by the Universal Thermal Climate Index (UTCI), and bike trips in Cambridge, MA. Bluebikes, the bike sharing provider in the Boston area, provides yearly/monthly bike ridership data to the public on their website where each data point was a single bike trip with several trip attributes. We sorted all trips according to their start-station and end-station and identified three station pairs with some of the most bike trips taken by subscribers. Table 1 presents the three station pairs and the number of data points available between 2017 and 2019.

*Table 1: bike-station pairs*

|  |  |  |  |
| --- | --- | --- | --- |
| Route | Start Station | End Station | # of Trips |
| 1 | MIT at Mass Ave / Amherst St | Beacon St/Mass Ave | 13,291 |
| 2 | Davis Square | Linear Park - Mass. Ave. at Cameron Ave. | 8,161 |
| 3 | MIT Pacific St at Purrington St | MIT at Mass Ave / Amherst St | 9,077 |

We created datasets of hourly UTCI values for three years based on local weather data and environmental simulations. Each data point was a single hour in a year with a UTCI value and other weather attributes, such as ambient temperature, relative humidity, wind speed and precipitation. For each data point, three variants of UTCI were included varying in degrees of simulation accuracy (1 being the least, 3 being the most accurate). To relate the bike trips data to the weather data, we created a python script that would count the number of trips between each station pair in a given hour and append this as a new column to the weather data.

*Figure 1: Data processing flow chart*

Timeline

Description automatically generated

**Exploratory Data Analysis (EDA)**

We explored the distributions of our data variables, the relationships between them, with a major focus on their relationship with our target predictor ‘bike count’.

Based on the extensive EDA, the key findings were:

* Weather variables (DBT, RelHum, WSp, Sun Elevation) are highly correlated with UTCI because they inherently define its value. This indicates the potential of exploring separating them in different predictor sets to avoid collinearity in our models.
* Some positive correlations between bike count and DBT, UTCI and Hour are observed and negative correlations with precipitation and wind speed.
* The different UTCI variants do not reflect any noticeable difference in the correlations and one of them may be sufficient to include as a predictor.
* The distribution of bike counts is skewed towards zero count. This may suggest that it might be useful to explore predictor sets that address this limitation. Those can include balancing the dataset, developing a model that focuses on predicting non-zero bike counts or one which excludes data hours where very limited bike use is expected (midnight).
* The spread of data points for respective variables against bike count suggests that those predictors affect the maximum possible bike count more than they clearly define the count.

Initially, we wanted to analyze the three routes separately and compare our model’s performance between routes. However, after performing the EDA, we realized that there were large benefits to combining the datasets into one larger set, especially in training, validating and testing of models.

*Figure 2: Data distribution for each data variable*

Diagram

Description automatically generated

*Figure 3: Relationship between main data variables and bike count*

Diagram

Description automatically generated