OPTIMISING THE DUBLIN BIKES GRID

AUTHORS:
CRIOMTHANN MORRISON **TIANMIN ZHANG**
MASON RICCI **ZHIPENG LIN**

INTRODUCTION

Dublin Bikes users face a common problem: arriving at a station to park their bike, only to find that the station is full. Further, 'full' stations often cluster in the same area, exacerbating the impacts of poor service provision. Such scenarios can lead to negative outcomes for the user, including frustration, fatigue and discomfort, and delayed appointments.

Using Machine Learning methods, this report identifies a section of the city with stations often full in the afternoon, and demonstrates how expanding five bike stations is predicted to improve service provision. Five bike stations may also be reduced with no predicted disruption to user experience.

By reducing the frequency that key bike stations are persistently full, our optimisation proposals will improve the overall user experience and imbue greater public confidence in the Dublin Bikes service.

DATA PREPARATION

TIME_PERIODS

The original dataset reports bike station usage at 5-minute intervals. These are divided into four daily time periods to account for differences in use over the course of an given day, with special consideration for peak commuting times:

MORNING	5 AM - 10:59 AM
AFTERNOON	11 AM - 2:59 PM
EVENING	3 PM - 7:59 PM
NIGHT	8 PM - 4:59 AM

HIGHLIGHTS

Through clustering techniques, we identify five bike stations for increased capacity at the following locations:

- Dame Street
- Fowens Street Upper
- Molesworth Street
- Merrion Square West
- Prince/O'Connell Street

Further, we identify five bike stations for decreased capacity at the following locations:

- York Street East
- Fenian Street
- Newman House
- St. Stephens Green East

Using regression methods, we propose changes that reduce the average frequency of 'full' stations by an estimated 14% over the course of a given afternoon.

FULLNESS

A new feature fullness is generated to represent the number of bikes present at a station at a given time as a proportion of the total number of bike stands installed. This fullness variable exists on a 0-1 scale; where 0 indicates there are no available bikes for users, and 1 indicates a station is full of available bikes. In other words, 0 indicates that all bike stands are empty, while 1 indicates that there are no empty bike stands.

$$Fullness_{BikeStation} = \frac{Available_{BikeStation}}{Total_{BikeStation}}$$

To account for obscure cases of faulty bikes or bike stands, which may obfuscate the true capacity of individual bike stations, a threshold of 0.9 is used to indicate the true "fullness" of available bikes; or contains no empty bike stands.

Note that this relationship does not exist in all clusters, and so future analyses should explore the dynamics of this relationship further.

FULLNESS BY TIME_PERIOD

Additionally, the two variables time_periods and fullness, are combined to generate four new variables:

- fullness_morning
- fullness_afternoon
- fullness_evening
- fullness_night

These variables represent how often a bike stand reaches 90% fullness or higher in the specified time period; where a value of 0.5 for fullness_afternoon indicates the given station is above 90% bike capacity for half of the afternoon.

SELECTING STATIONS

The proximity of 'full' bike stations with others is an important dimension to addressing the issue of persistent fullness of a station (or cluster). Therefore, cluster analysis is used to identify groupings of persistently full bike stations which will most benefit from increased capacity among neighbouring stations. It also highlights groups of stations which have low risk of reaching full capacity when reduced.

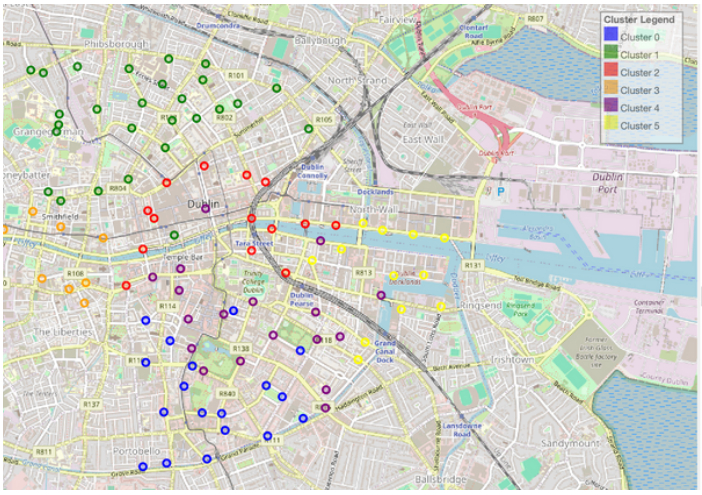

Figure 1: Six Clusters of Stations

K-means clustering by features LONGITUDE and LATITUDE (geographic features) and fullness_afternoon sets a baseline for further analysis of other time periods of the pre-defined groupings of stations.

While the Calinski and GAP criterions are used to estimate the appropriate number of clusters to specify, both algorithms output values greater than 15, which proves too large for practical use. Furthermore, increasing weightings of *fullness* to achieve greater heterogeneity within clusters do not improve results; standardisation of features is more effective. Ultimately, a brute-force approach finds six clusters on standardised features produce the most useful results. Figure 1 displays the result of the clustering on a map of Dublin City.

Cluster 4, encompassing the central and south portions of the city center, experiences abnormally high levels of fullness for many stations during the afternoon. Therefore, our analysis focuses on optimising this section of the Dublin Bikes grid.

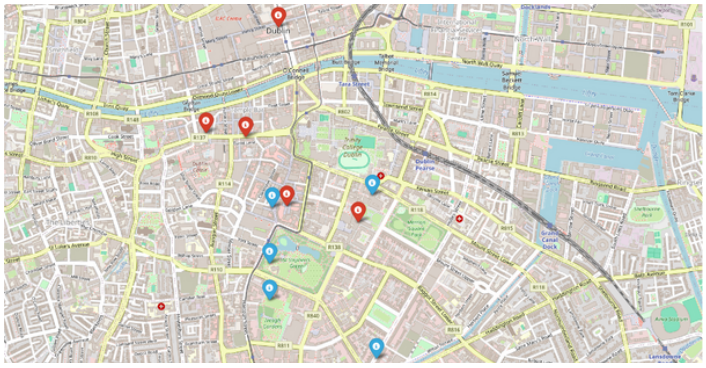

Figure 2: Five Stations to Reduce (Red) and Five Stations to Expand (Blue)

PREDICTING TIME SERIES

Simple linear regressions for the stations within Cluster 4 indicate negative correlation between the total number of bike stands and the average fullness of that station; the trend exists in all time periods. This analysis is extended to assess the impacts recommended for station modifications below. Figure 3 illustrates this trend across each time period.

Note that this relationship does not exist in all clusters, and so future analyses should explore the dynamics of this relationship further.

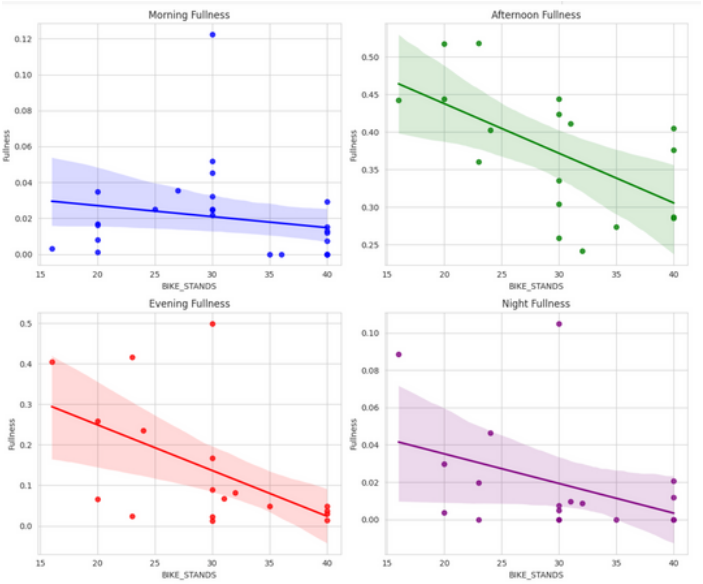

Figure 3: Negative Relationship Between Bike Stand Capacity and Regularity of Fullness Across Morning, Afternoon, Evening, and Night

Additional regression models estimate the average fullness of each modified bike stand, and Figures 4 and 5 illustrate the differences between the original and modified bike stations in average fullness over a given day.

Note that this analysis is based on adding or removing 5 bike stands per station - this value is arbitrarily chosen for reporting purposes and may be altered and re-assessed based on space limitations at selected locations and other considerations.

As the data available and team resources are limited, simple linear regression offers preliminary insights. However, with more resources, polynomial and logistical regression may be used for classifying outcomes and estimating the impacts of the recommendations more effectively.

Note cross-validation is used to ensure the validity of the models. This is a common approach for cases like this study where data is limited yet can be effectively shuffled and re-sampled as training and validation data.

RESULTS

Our analysis shows that the selected stations for additional bike stands are expected to reach full capacity less often throughout the day, most notably in the afternoon period. Further, Figure 5 shows that reduced stations are predicted to experience no more than 0.5 instances of full capacity after changes are implemented.

The results of the regression models produce a Mean Squared Error for our test set was 1060 and 1160 for each model, which are somewhat lower than those for the training sets. This indicates our model is underfitting the data. This should be investigated further in later research.

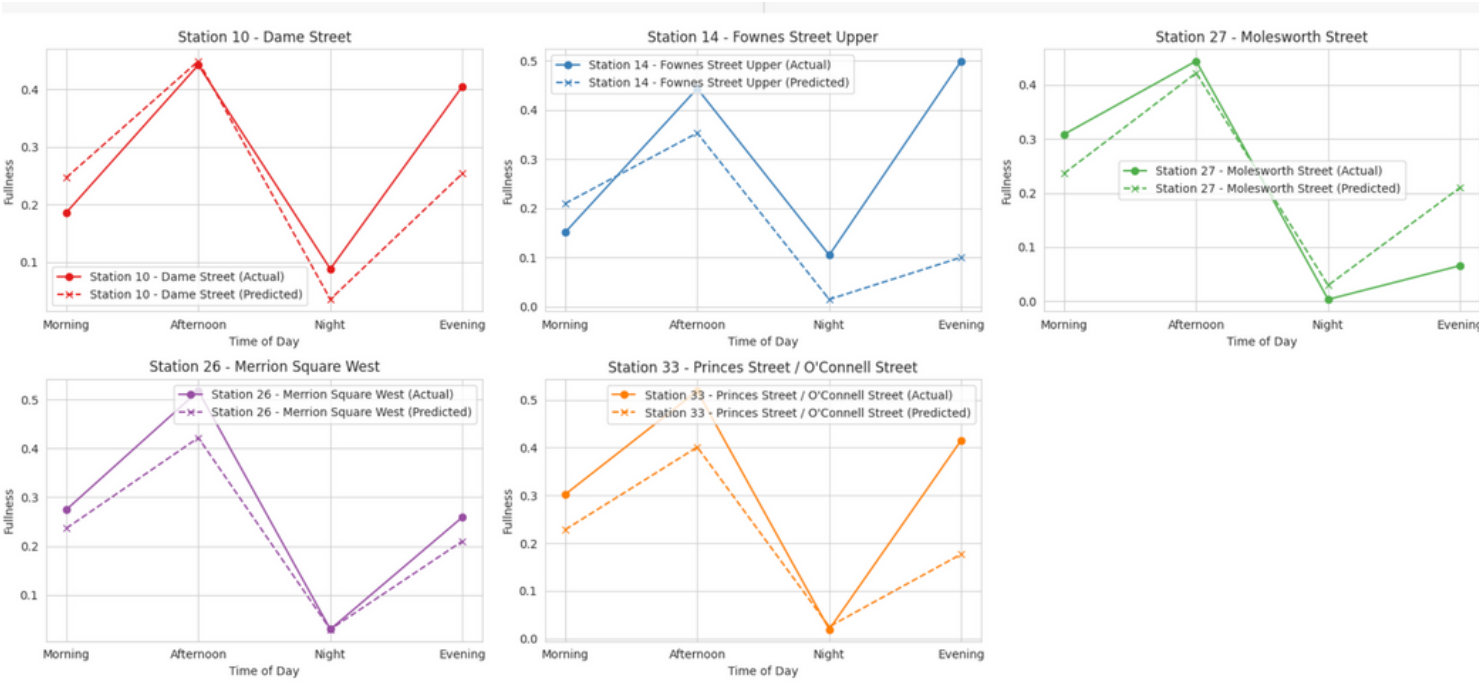

Figure 4 : Change in (Estimated) Fullness for Stations With Added Bike Stands

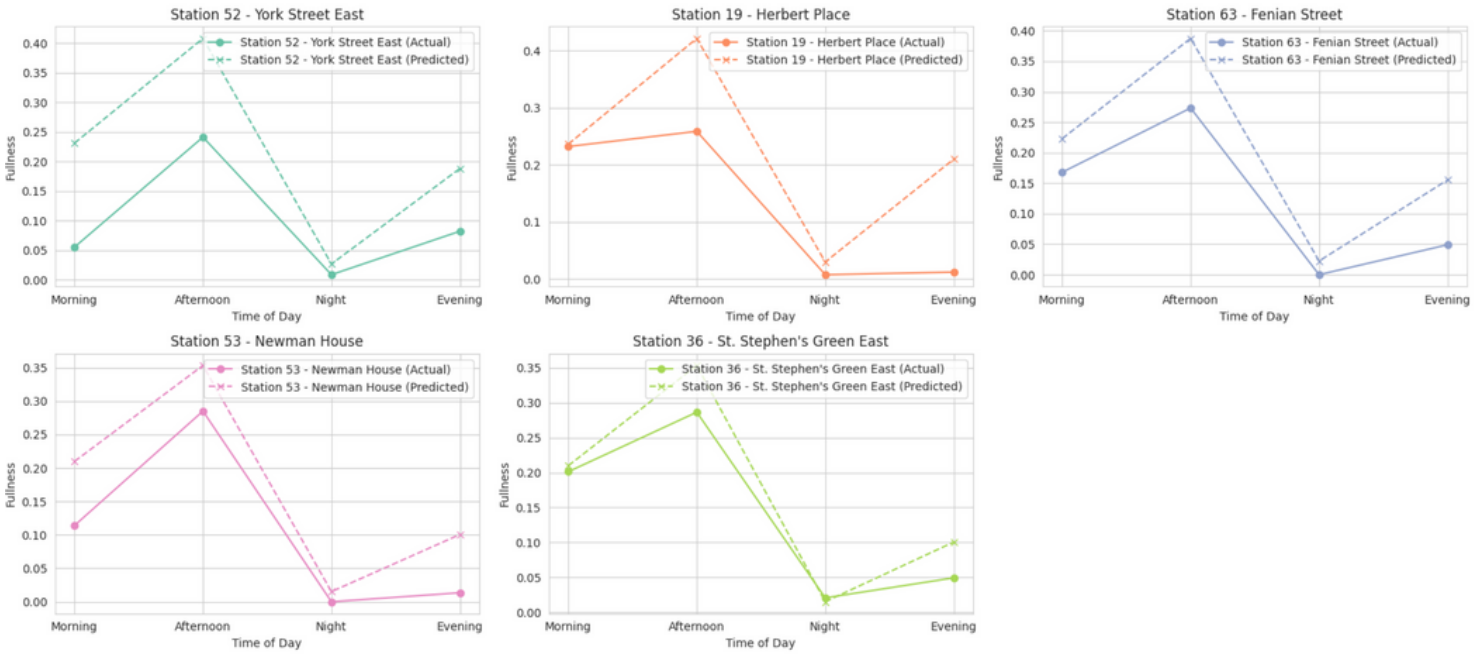

Figure 5: Change in (Estimated) Fullness for Stations With Removed Bike Stands

FURTHER CONSIDERATIONS

Given resource constraints, several factors are not considered in our analysis. For example, Dublin City Council uses several trucks to transport bikes from one stand to another to help alleviate over/under capacity of bike stations over the course of the day. Regrettably, data for this activity is not available.

Furthermore, the data used in this study is limited as it only accounts for Quarter 3 during the 2018 period. Further analyses may implement similar or more comprehensive techniques with larger and more recent datasets, to further expand on our recommendations, for instance, accounting for other variables which affect service usage such as public holidays and seasonality.