MSc in Data Analytics – CA2

An Analysis of Dublin Bike Rental trends and comparisons to London

Author: Stephen Burke
e-mail: sba23027@student.cct.ie
Student ID: 23027
Github: https://github.com/step-hen-burke/CA2

Accompanying notebooks:

- https://github.com/step-hen-burke/CA2/blob/master/1_data_processing.ipynb
- https://github.com/step-hen-burke/CA2/blob/master/2_dublin_london_comparison.ipynb
- https://github.com/step-hen-burke/CA2/blob/master/3_rental_prediction.ipynb
- https://github.com/step-hen-burke/CA2/blob/master/4_bike_rental_dashboard.ipynb

Wordcount:
Total: 3843
Figures: 80
Appendix & references: 489
Final count: 3274

Introduction

Public bike rental schemes have been launched in many cities around Europe, including successful networks in Irish towns and cities such as Dublin, Cork, and Galway. These provide local active transport links for commuting and leisure. This project explores the publically available data pertaining to the Dublin bike network, anlayses the sentiment towards related terms using reddit comment data, makes comparisons to the London bike network, and makes predictions that could be used to manage and improve the network.

Methodology

Data Retrieval

Data was downloaded with regards to the dublinbikes service run by JCDecaux, and the equivalent service in London, which is operated by Santander. Data is openly available from public APIs, and public datasets are hosted by the Irish and UK governments.

The historical data for Irish bike rentals was retrieved as a set of csv files totalling 4.5Gb. A comparison was carried out between the csv and parquet file formats to determine which format would be used for the rest of the analysis. It was found that the parquet format was several times faster at both read and write operations, as well as being nearly 20 times more efficient storage-wise due to its in-built compression. Results of the comparison can be found in the table below, and full details can be found in section 1.2.1 of the accompanying notebooks.

Format	Read Time (s)	Write Time (s)	Disk Space (Mb)
CSV	48.6	156.6	263.0
Parquet	16.7	19.9	4489.1

Historical data for London was retrieved as an xlsx file. This contained daily, monthly, and yearly rental figures in addition to YoY changes etc. This required additional processing as the data was not stored in a "tidy" format (Wickham 2011), details of which can be found in section 1.1.2 and 1.2.3 of the notebooks.

A catalogue of each dublin bike station was retrieved from the JCDecaux API in json format. This provdied each station's id and address, as well as their location in latitude and longitude (see notebok section 1.1.1).

Supplementary data was retrieved from met eireann in order to get a sense of the weather conditions throughout the history of bike rentals retrieved, and from the european population grid in order to augment the retrieved station data with the local population levels. The weather data was retrieved as a zip file contining a csv, and the grid data was retrieved as a parquet file.

In each of these cases the licenses were permissive provided the data was not used commercially and was properly attributed. Links to the relevant datasets can be found in the references section and further discussion and links can be found in notebook section 1.1.

Reddit data was accessed using their open API. Comment data was retrieved relating to a number of relevant search terms - "bikes", "cycling", "cyclists", "transport", and "bike share". These comments were retrieved from both the r/london and r/dublin subreddits. The access conditions for the API were that an app was registered to a reddit account in order to generate a token, and this token was used with the oauth.reddit.com site for all requests. This was adhered to for all retrievals.

Data Processing

The Dublin bike data was retrieved at a mix of 5 minute and half-hourly cadences, whereas the lowest granularity from the London data is daily. In order to compare the two datasets in subsequent sections the rentals were estimated for Dublin by looking at the differences in numbers of available bikes at each station at subsequent times and summing these up to daily figures. This had the additional benefit that the data volume for daily aggregated data was much less than the half-hourly data across every station, and as such was much less computationally intensive work with. Details of this can be found in notebook section 1.3.2.

The population data retrieved from the European grid was used to augment the station data with the local population values from the nearest square kilometer as defined by the grid. "Near" in this case was calculated by converting the Latitude and Longitude retrieved from JCDecaux's station index data to Northings and Eastings used by the European grid. As these are expressed in meters, closeness could then be calculated using a metric such as the Euclidean or Manhattan distance metric, without having to perform calculations using spherical coordinates. The transformations between both coordinate systems were unit tested with a known pair of points to ensure that the correct conversion was occurring.

As only some years had population data available, an assumption wass made that population would increase linearly in each grid square and as such, a linear model was fit to each grid subset of population values to impute the missing years. Full details of this can be found in notebook section 1.2.2.

Some additional processing was also necessary on the met eireann data retrieved: the date column was encoded as a string, and null values were encoded as empty strings. Additionally, missing values were found, but could be safely ignored as they occurred before the beginning of the retrieved bike data. Details of the rectification of these issues can be found in notebook section 1.2.4.

The reddit data was retrieved as a series of json files, containing up to 100 comments in each. A helper function was defined to extract the relevant information from each comment blob, and unit tests were defined using a sample json object in order to ensure that they were consistently processed (see notebook section 1.2.5).

Sentiment Anlaysis

The reddit comment data was analysed in order to get a sense of the sentiment of the Dublin and London subreddits towards bike-related terms. This was carried out using the nltk library and its in-built sentiment analysis class, which is powered by VADER.

The text processing flow was developed using a test driven development workflow. As the output format of the text was known ahead of time, a set of tests was defined using the unittest framework which set out the expectations for the output, and the text processing function was iteratively changed until these tests were satisfied.

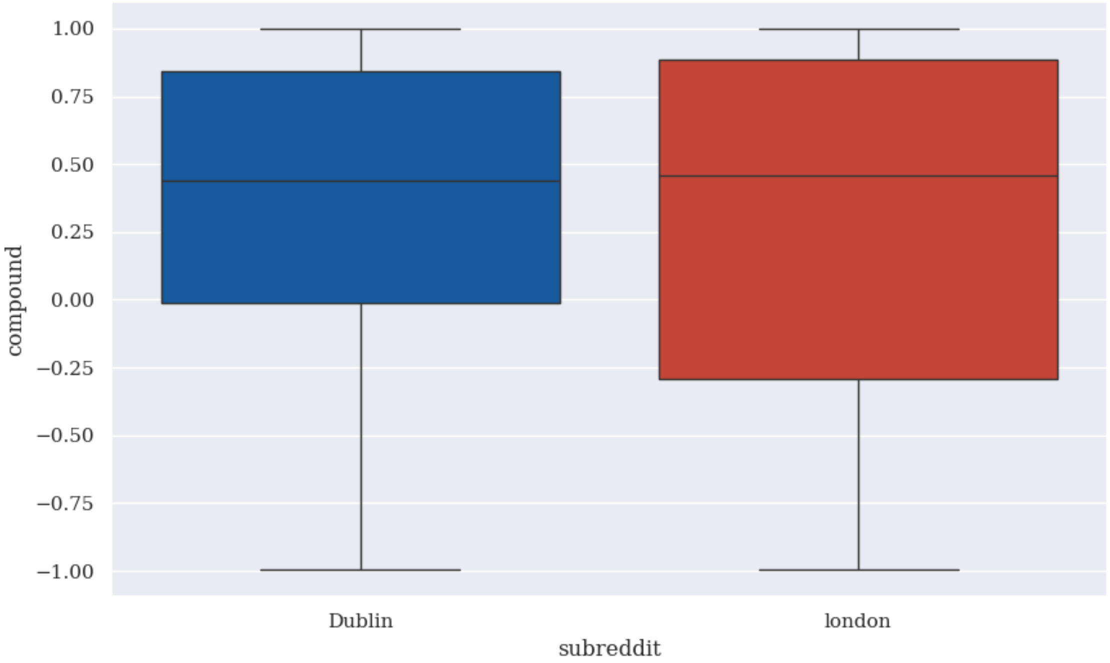
The processing steps were as follows:

1. The text was normalised - converted to lowercase, stripped of excess whitespace and punctuation

2. Stop words were removed as these do not contribute to the overall sentiment of a comment and would dilute the results if left in

3. Words were stemmed to their base form so that equivalent words are not treated differently due to their conjugation.

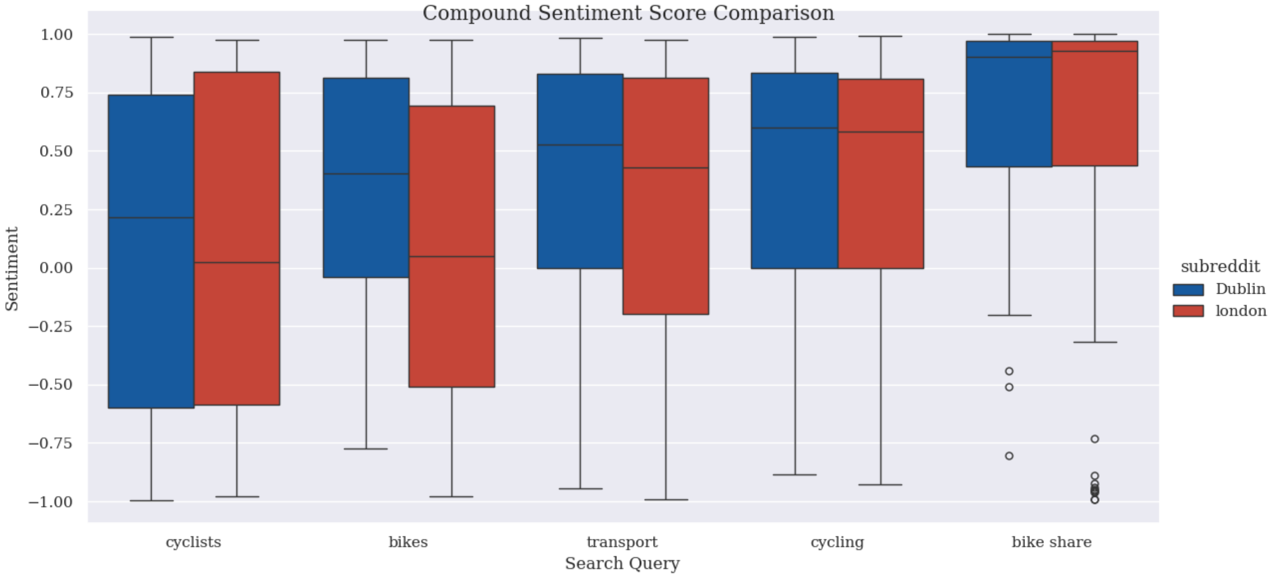
Visualising the sentiments of the subreddits in question as a boxplot (figure 1), it can be seen that the average sentiment is roughly the same, however the sentiment of r/london is skewed more negatively that r/dublin.



(figure 1: Sentiment comparison between Dublin and London)

In this visualisation (and subsequent visualisations where apropriate). A variety of visualisation recommendations made by Tufte (2001) are adhered to. The seaborn default style is set to "darkgrid", which has a muted grey background so as to not be distracting, and has gridlines which use minimal ink. The font family uses serifs in order to create more "friendly" graphics, and the default linewidth is thinner than seaborn's default in order to use less data ink, and thus result in more aesthetically pleasing graphics.

It was found that for nearly all terms r/dublin was more positive than r/london. Particularly for the terms "cyclists" and "bikes", however r/london was slightly more positive towards the phrase "bike share". This can be seen in figure 2.



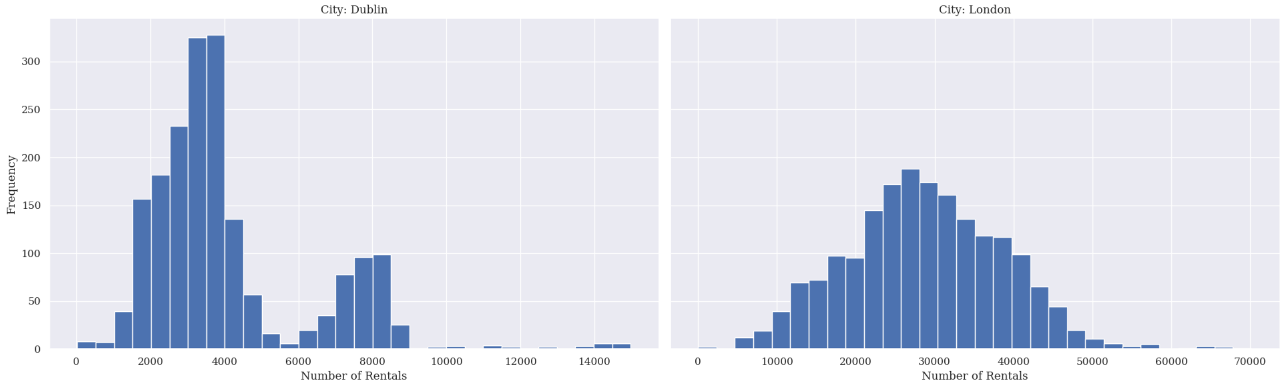
(figure 2: Sentiment comparison between Dublin and London by search Term)

This figure demonstrates another technique for the display of visual information that Tufte advocates - that the use of multiple plots side by side allows information to be easily compared.

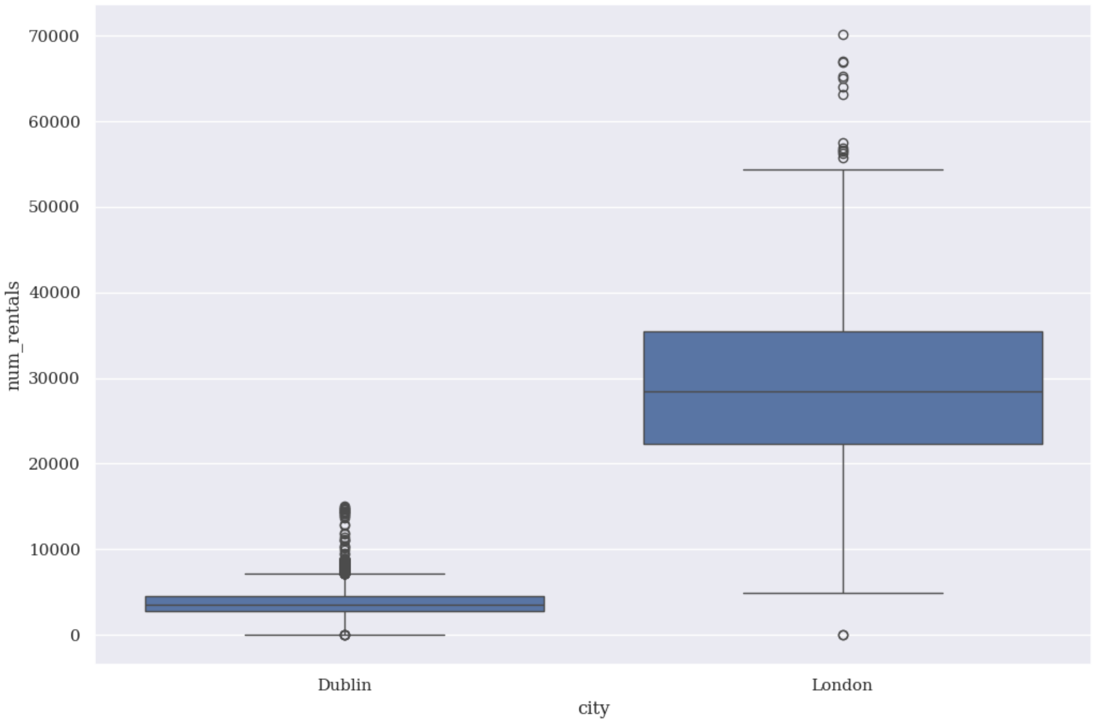
Full details of the sentiment analysis can be found in section 1.4 of the accompanying notebooks.

Dublin and London Bike rental Comparison

The mean daily rentals was compared between Dublin and London using a t-test (Weiss, 2017). In this case H0 was that the mean levels were the same, and an alpha level of 0.05 was chosen by convention. The sample size was sufficiently large that the normality assumption of the test could be assumed to be satisfied (Lumley et al., 2002). It was found that the difference in means is significant, as the p-value obtained was very close to 0, and so it can be concluded that the average daily rentals is different between the two cities. This is most likely due to the difference in population between the cities, and this difference in scale was important to note in subsequent analyses. This scale difference in population was estimated to be a factor of 8.11. This finding was not dissimilar to another estimate of the population scale factor with some caveats pertaining to the relative coverage of the rental networks and the population estimates taiking the metropolitan area population into account. More details and discussion can be found in notebook section 2.1.1.



(figure 3: Daily rentals in Dublin and London (histogram))

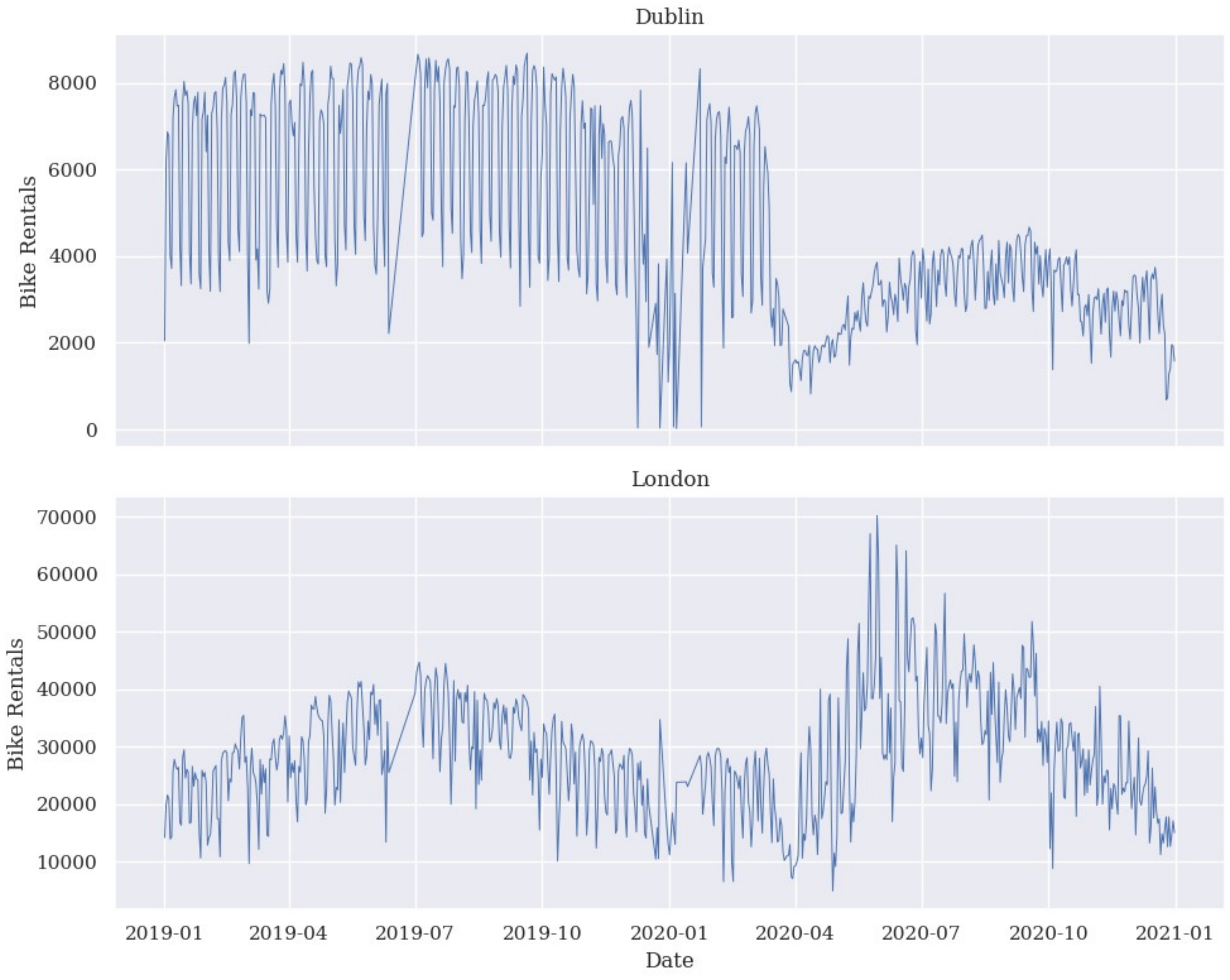


(figure 4: Daily rentals in Dublin and London (boxplot))

Comparing rentals between 2019 and 2020 in both London and Dublin, it was found that the number of rentals was significantly different between the two years in Dublin, but not in London. This can be seen on the visualization below (figure 5).

A Wilcoxon test (Weiss, 2017) was used for this comparison, as the data can be thought of as paired samples before and after an "intervention", which in this case is the pandemic. The null hypothesis for both cities was that there was no difference in rentals between 2019 and 2020, and once again an alpha value of 0.05 was used. For Dublin, the p-value obtained was very close to 0, whereas for London, a p-value of 0.62 was obtained.

It can be seen that while bike rentals decreased dramatically in Dublin at the beginning of the Covid19 pandemic, rentals appear to increase in London. This may be indicative of differences in approach to lockdown in both countries (notebook section 2.1.2). In notebook section 3.1, this result was noted and an indicator variable for before and after the beginning of the pandemic was defined for machine learning purposes.



(figure 5: Rentals at the beginning of the Covid19 pandemic)

As the bike rental data is a time series, potential seasonal effects were taken into account and investigated as these could influence variable selection in further machine learning applications.

Firstly, a mann-whitney U test was carried out to investigate whether or not there is a significant difference between the shapes of the distributions of monthly rentals across both cities. In this case, the null hypothesis was that there is no difference between the distributions of the two.

Additionally, a chi-square test was carried out to investigaye the relationship between the categorical variables of month and city in order to determine whether there are differing seasonal effects between the two cities. The null hypotheis for this test was that there is no association between distributions of bike rentals across the months between cities.

Finally, an ANOVA was carried out to determine whether there are differences in the mean rentals per month, or rather, whether seasonal effects exist at all.

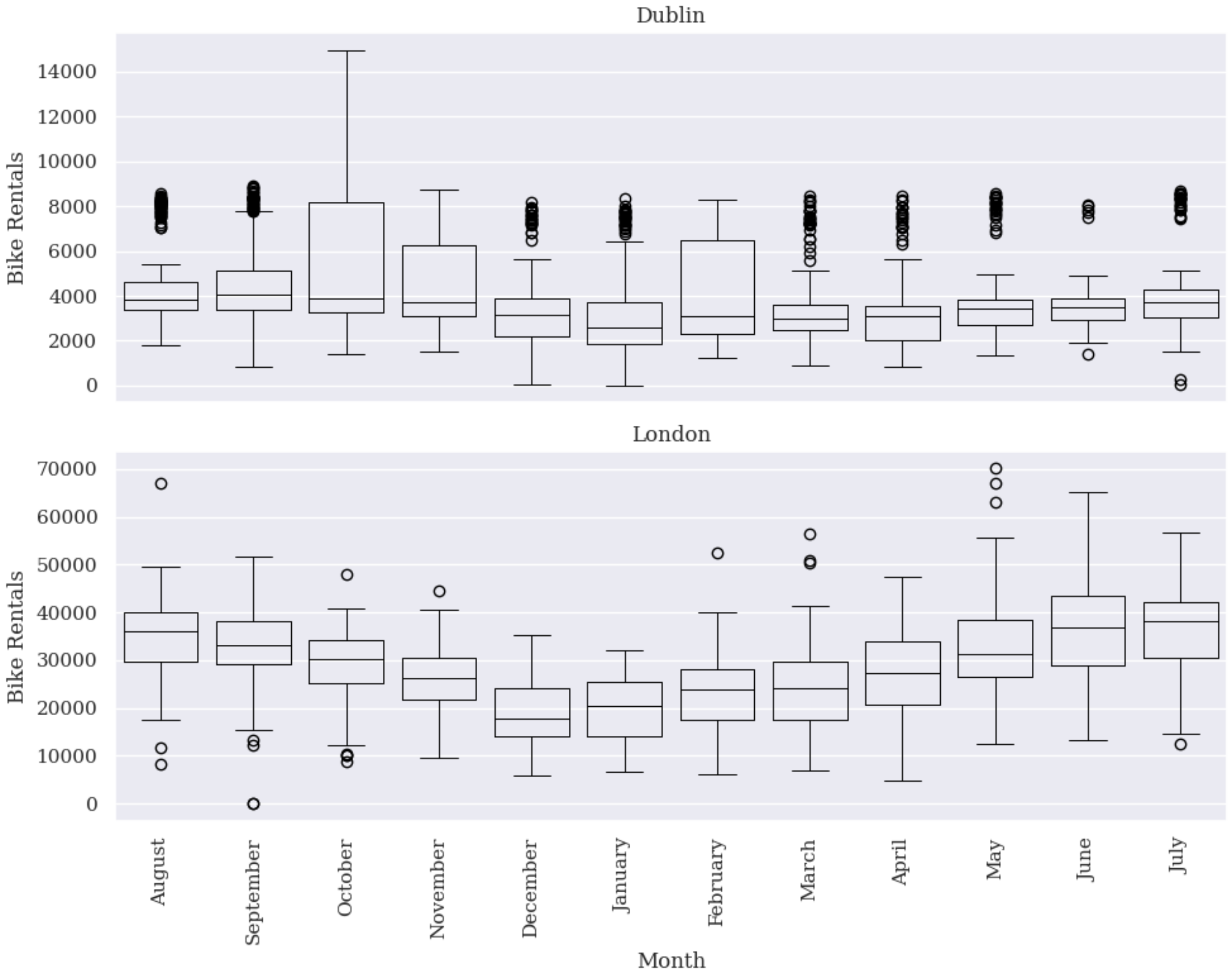
For each of these tests, the proportion of total rentals in the city for the year was analysed in order to control for differences in scale between the two cities. This is known to be in important consideration from the results obtained in notebook section 2.1.1. Again, by convention, a threshold value of 0.05 was chosen in all cases.

The p-value obtained for the U test was 0.46, meaning that the null hypothesis was not rejected. This can be interpreted as the monthly rentals (when expressed as a proportion of the rentals for the year) are distributed similarly in both Dublin and London.

The p-value obtained for the chi-square test was 0.99, meaning that again, the null hypothesis was not rejected. This indicates that the association between the city and month variables is not significantly different. In other words, there is insufficient evidence to claim that the impact of the seasons on bike rentals is different in London than in Dublin.

Two ANOVAs were carried out - one for each city. For Dubin a p-value of 0.18 was obtained, indicating that there is no significant seasona effect. In London however, a p-value of 0.02 was obtained, indicating that seasonality has a significant impact on bike rentals.

Visually, it can be seen that the London seasonal effect is distinct, with fewer rentals in the winter than the summer, whereas Dublin's rentals are not as easily visually separated, and have much wider spreads from month to month. This is illustrated in figure 6.

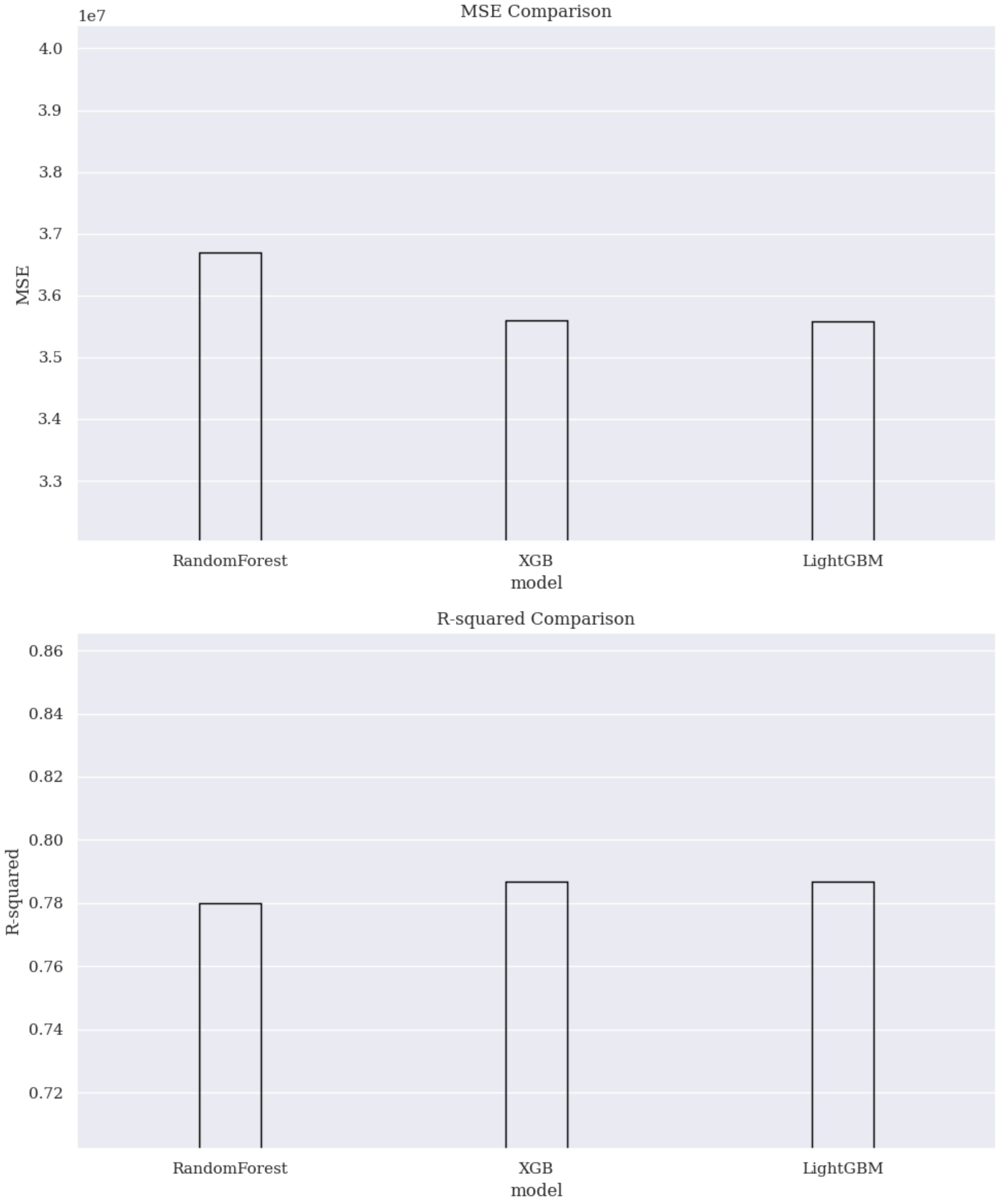


(figure 6: Seasonal comparison between Dublin and London)

Bike rental prediction

Predicting the number of daily rentals in both Dublin and London can be framed as a supervised regression problem. Further to the analysis carried out in section 3 of the notebooks, various date based features were defined such as the day of the week, the month of the year, whether or not the date was a public holiday, and whether the date was before or after the beginning of the Covid-19 pandemic.

Several tree-based regressors were tuned and compared. It was found that LightGBM (Ke et al., 2017) outperformed Random Forest and XGBoost (Chen & Geustrin, 2016), having both a lower mean squared error and higher R2 score (figure 7). Full details of this can be found in section 3.1 of the accompanying notebooks.

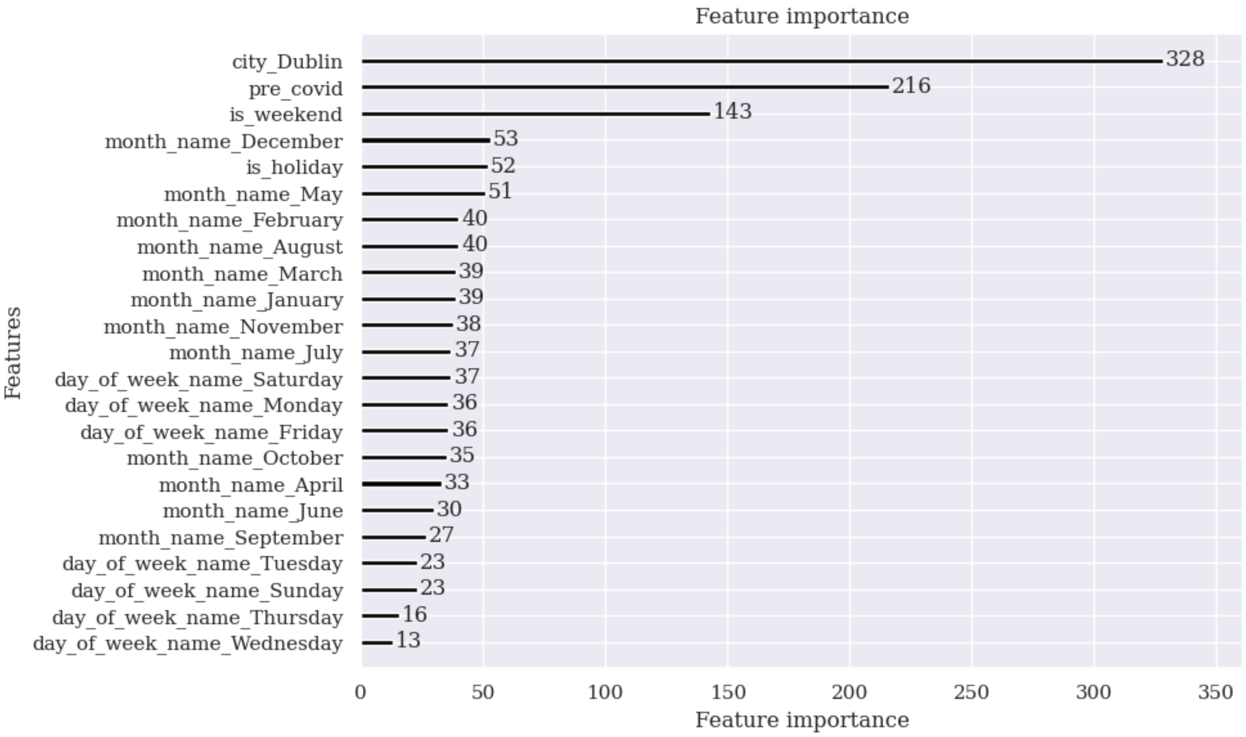


(figure 7: Regression model comparison)

Looking at the feature importance chart (figure 8) for the best performing model (LightGBM), it can be seen that the city=Dublin boolean is the most important feature, which is to be expected given the results from notebook section 2.1.1, followed by the pre_covid boolean. This also makes sense as a shift in behaviour post-covid was noted in section 2.1.2.

Whether the day is a weekend is the next most important feature, which again makes sense as one would imagine that one of the primary uses of the bike network in both cities is to commute to work. The is_holiday variable is also seen to rank highly, which could be due to a similar phenomena.

It can also be noted that the December dummy variable is the most important of the seasonal dummies. This is possibly due to cycling being a less attractive transport option in Winter due to poor weather conditions, a trend that can be observed in some other parts of Europe (Hudde, 2022).



(figure 8: LightGBM Feature Importance)

In these figures (7 & 8), the coloured fill has been removed from each bar in accordance with Tufte's recommendation to minimise data ink use. The outlines of the bars are sufficient to convey the necessary information.

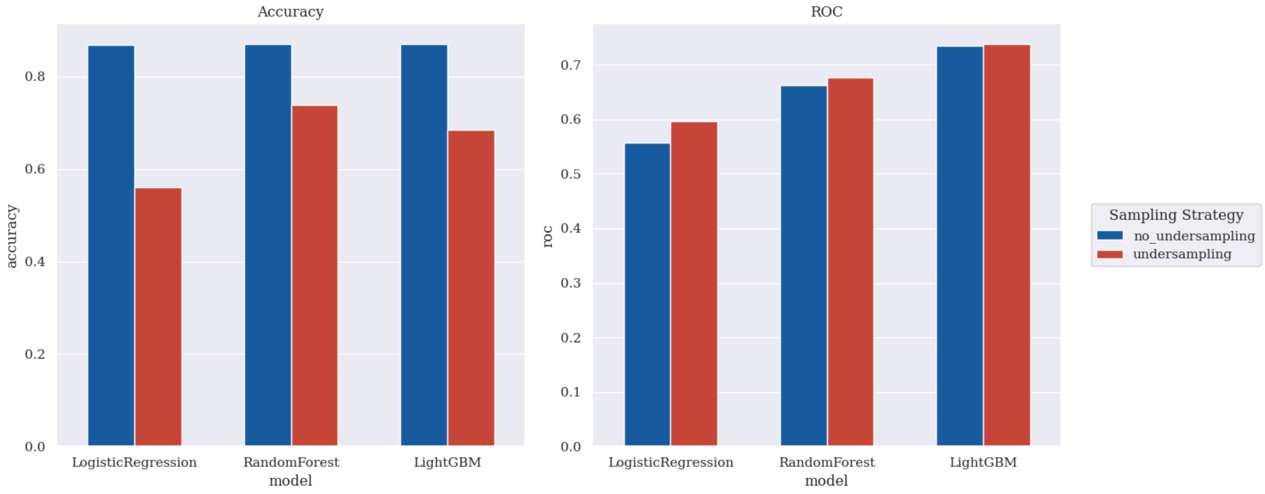
In order to successfully manage a bike network, it would be useful to be able to predict peak demand both for bikes and for stations. Action should be taken either to replenish the bikes at empty stations, or to redistribute bikes from full stations. Some examples of actions that could be taken by the network operators have been identified in notebook including: scheduling bike collection to line up with demand spikes around Heuston Station, more aggressive bike replenishment around the TUD campus, and increasing capacity at the periphery of the network.

Identifying the need for intervention can be framed as a supervised classification problem, with the target being a binary variable denoting whether a station needs attention (notebook section 3.2).

Three models were compared – LightGBM, RandomForest, and LogisticRegression. It should be noted that the data used for this section of the analysis was the hourly data for every dublinbikes station in dublin for the first half of 2023. Several issues arose due to the scale of this, for example, SVM was initially considered for comparion, but it was deemed to be computationally infeasible to train. Also, dimensional reduction was considered at this point, however the applicability these techniques was limited due to the scale of the data – none of the methods tried concluded in a reasonable amount of time.

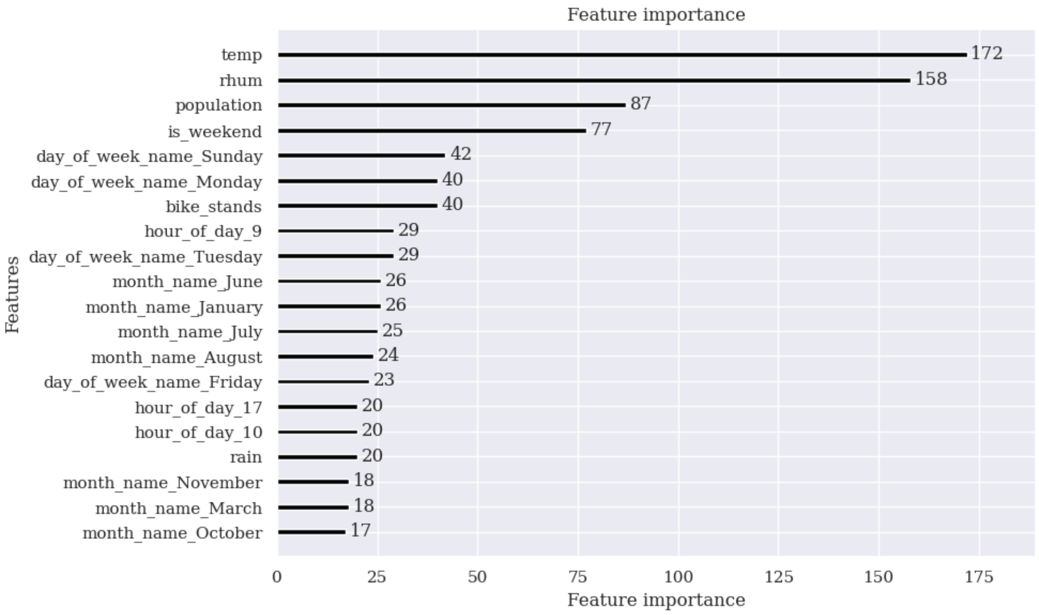
It was found that the models were preferentially predicting the negative class when provided the data as-is, so resampling techniques were employed. Due to the scale of the data, undersampling was chosen rather than oversampling, again to save computational resources.

It was found that although the accuracy of each model decreased when undersampling was applied, the AUC of each model improved. A comparison of these metrics before and after resampling can be found in figure 9. LightGBM was again selected as the best performing model due to its high AUC score.



(figure 9: A comparison of classification metrics)

Looking at the top 20 features for this LightGBM model (figure 10) we can see that the most important features are those related to weather (temperature and humidity), the number of people living in the local area, seasonal indicators in Summer and Winter (January, June, July, etc.) and those related to commuting - time of day is 9am or 5pm, whether the day is a weekend etc. This reinforces some of the intuitions developed in section 2.



(figure 10: LightGBM classification feature importance (top 20))

Dashboard

A dashboard was designed to aid in the management of the dublinbikes network. As mentioned in notebook section 3.2, it is important to be aware of when stations need attention. In aid of this, multiple interactive charts were defined:

A scatterplot of each station overlayed on a map of Dublin is provided, where the size of each point denotes the number of available bikes at each station at the selected time and date. The stations that "need attention" were coloured in red, and in the case of having 0 bikes available their sizes were artificually inflated. Both of these choices draw the user's eye towards the most relevant information on the map.

A line plot of the hourly rentals for the week containing the selected date is provided so that the user can see the weekly trends in rental – this allows them to pick out points in the week where the stations need attention, as well as see the rentals on weekdays vs weekends and the hourly commute effect, both of which were seen to be important factors in section 3.2.

A summary of the weekly weather events is given which lines up with the weekly rental plot. Temperature was identified as the most important predictive variable in section 3.2, so having both of these charts side by side allows the user to immediately geta sense of weather’s influence on rentals. Tufte advocates for having multiple plots placed such that they can be quickly compared in this manner.

Both this and the weekly rental chart have an indicator highlighting which point in the week the map figure is displaying. There is also a summary of the stations for the selected time, displaying the prediction from section 3.2’s LightGBM classifier for each station. Tufte argues that tables are often the best way to show exact values when understanding the data requires many localized comparisons, as is the case here.

Additionally, the bootstrap css framework was used to control the layout of the dashboard such that all relevant information is immediately visible, and to provide a minimal theme that does not distract the user from the information provided. Tufte argues for reducing the data ink used, so this minimal layout with few distracting elements was deemed to be appropriate.

Screenshots of the dashboard can be found in the appendix, and the dashboard code and additional details can be found in section 4 of the notebooks.

Conclusion

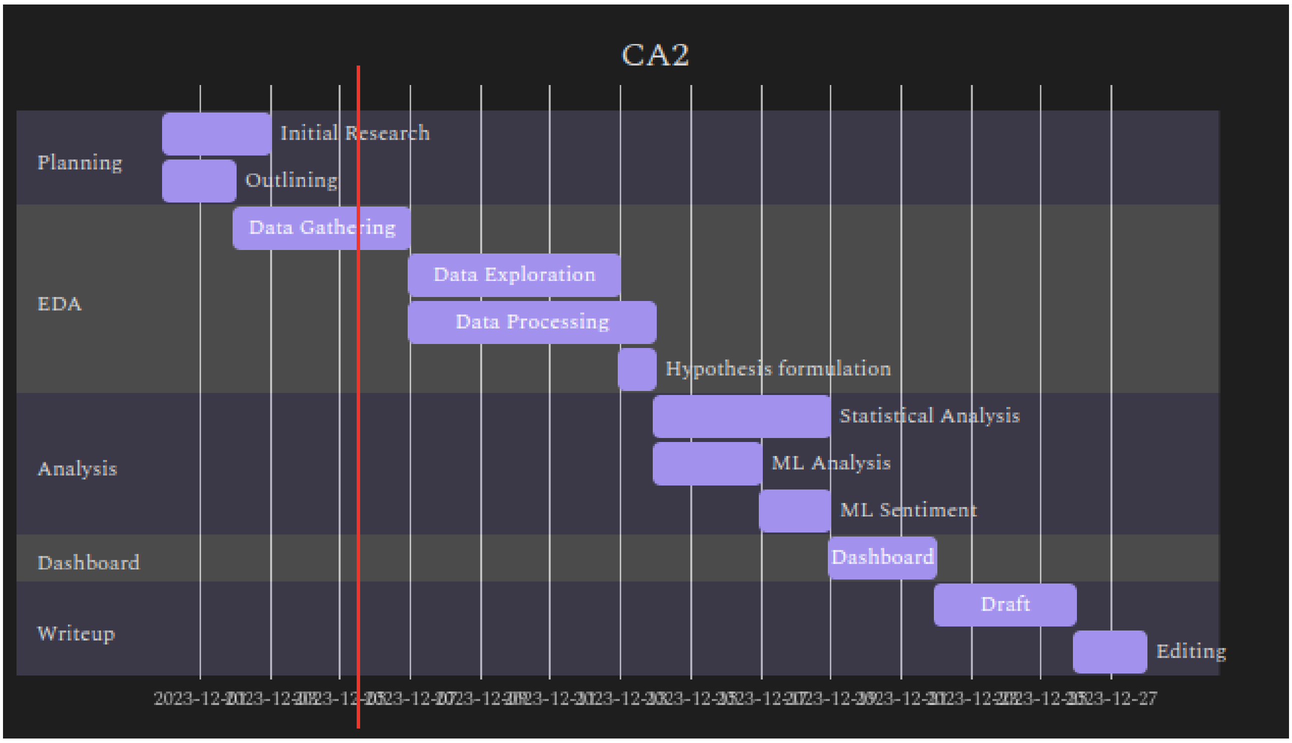
This analysis of Dublin's public bike network and comparison to London has revealed differences in bike rental and usage patterns both seasonally and in each city's response to the Covid19 pandemic.

It has been shown that each city's subreddit have distinct attitudes towards a range of terms related to cycling and bike rental, and it has been shown that supervised learning models can be used to predict both the daily rental levels in a regression context and when a station needs intervention from a network manager in a classification context.

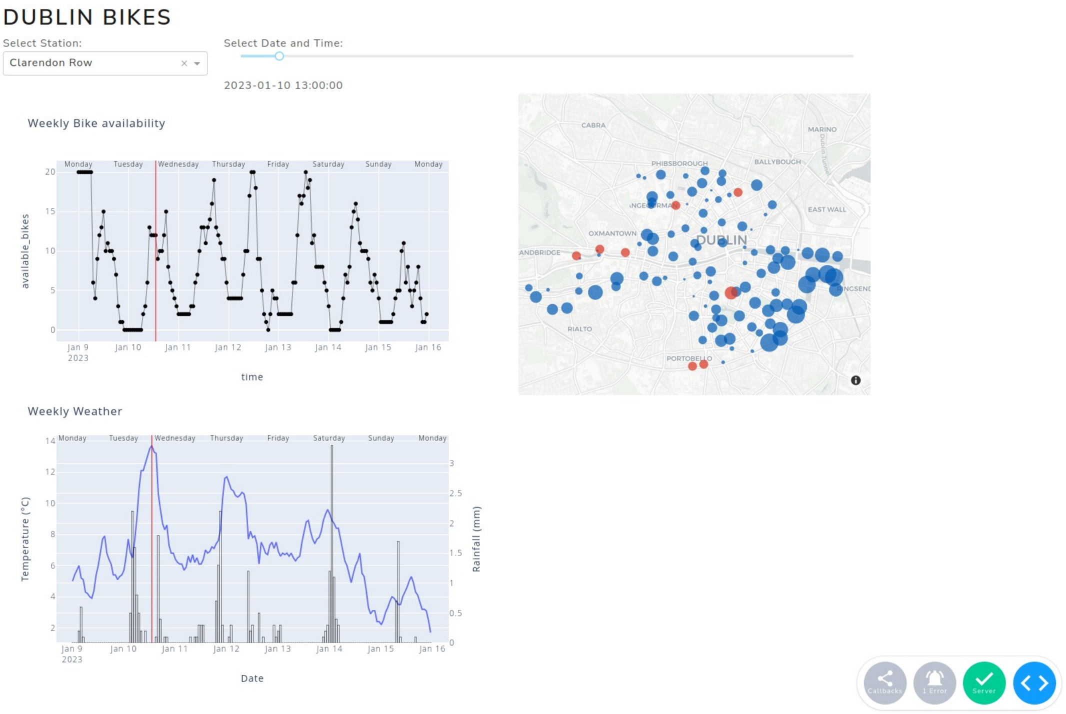
Appendix

This analysis was initially planned by identifying each sub-task that would feed into the final output, estimating the effort (in days of work) for each of these, and arranging them in a logical sequence (some tasks could proceed in parallel).

The gantt chart below is indicative of the workflow structure, some tasks took more/less time than anticipated, and the plan does not take into account elements such as Christmas. This chart was produced using mermaidJS in the note taking tool Obsidian, where much of the initial outlining and draft-work took place.



Screenshots of the Dublin Bike dashboard built in dash with plotly express and boostrap:



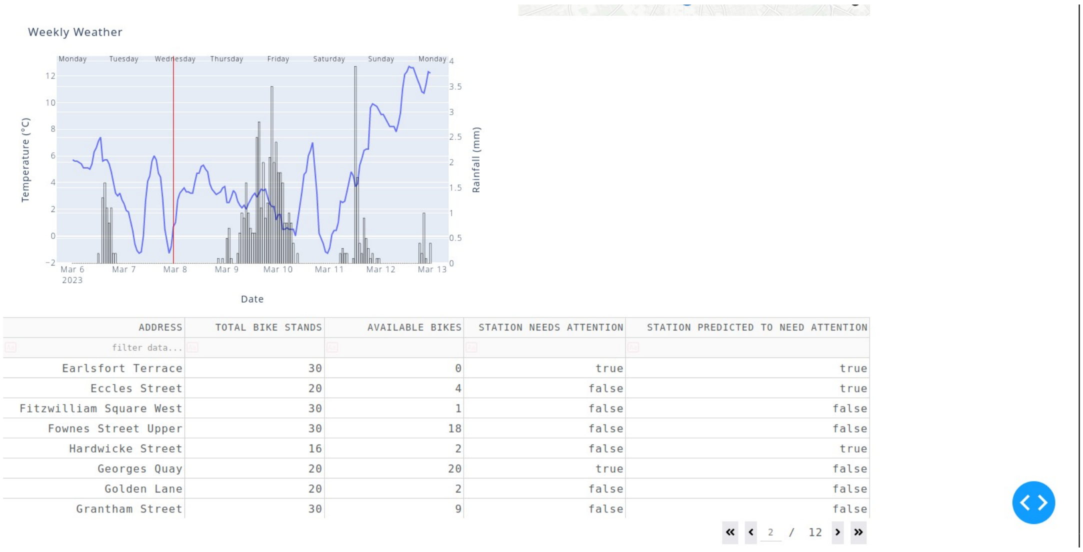
DUBLIN BIKES

Select Station: Clarendon Row

Select Date and Time: 2023-01-10 13:00:00

Weekly Bike availability

Weekly Weather



Weekly Weather

ADDRESS	TOTAL BIKE STANDS	AVAILABLE BIKES	STATION NEEDS ATTENTION	STATION PREDICTED TO NEED ATTENTION
filter data...				
Earlsfort Terrace	30	0	true	true
Eccles Street	20	4	false	true
Fitzwilliam Square West	30	1	false	false
Fownes Street Upper	30	18	false	false
Hardwicke Street	16	2	false	true
Georges Quay	20	20	true	false
Golden Lane	20	2	false	false
Grantham Street	30	9	false	false

« ‹ 2 / 12 › »

References

- Burkov, A., 2019. The hundred-page machine learning book. Andriy Burkov, Polen.
- Central Statistics Office, 2023. Census 2022 Population Grid (https://ie-cso.maps.arcgis.com/apps/webappviewer/index.html?id=0fe164e96d254776866425e2fd3e73af).
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. https://doi.org/10.1145/2939672.2939785
- eurostat, 2023. GISCO grids (https://gisco-services.ec.europa.eu/grid/GISCO_grid_metadata.pdf).
- Hastie, T., Tibshirani, R., Friedman, J.H., 2017. The elements of statistical learning: data mining, inference, and prediction, Second edition, corrected at 12th printing 2017. ed, Springer series in statistics. Springer, New York, NY. https://doi.org/10.1007/b94608
- Hudde, A., 2022. It's the mobility culture, stupid! Winter conditions strongly reduce bicycle usage in German cities, but not in Dutch ones (preprint). SocArXiv. https://doi.org/10.31235/osf.io/yejxf
- Hutto, C., 2023. VaderSentiment (https://vadersentiment.readthedocs.io/en/latest/).
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7
- JCDecaux, 2023. Self-service bicycles Open Data (https://developer.jcdecaux.com/#/opendata/vls?page=static&contract=dublin).
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 3149–3157.
- London Datastore, 2023a. Cycle Hire Availability (https://data.london.gov.uk/dataset/cycle-hire-availability).
- London Datastore, 2023b. Number of Bicycle Hires (https://data.london.gov.uk/dataset/number-bicycle-hires).
- Lumley, T., Diehr, P., Emerson, S., Chen, L., 2002. The Importance of the Normality Assumption in Large Public Health Data Sets. Annu. Rev. Public Health 23, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546
- Martin, R.C., 2012. Clean code: a handbook of agile software craftsmanship, Repr. ed, Robert C. Martin series. Prentice Hall, Upper Saddle River, NJ Munich.
- Met Eireann, 2023. Historical Weather Data (https://www.met.ie/climate/available-data/historical-data).
- Open Data Unit, 2023. Dublinbikes API (https://data.gov.ie/dataset/dublinbikes-api).
- Reddit, 2023. Reddit API (https://www.reddit.com/dev/api/).
- Tufte, E., 2001. The visual display of quantitative information / E.R. tufte. American Journal of Physics 31. https://doi.org/10.1109/MPER.1988.587534
- Vohra, D., 2016. Apache Parquet, in: Practical Hadoop Ecosystem. Apress, Berkeley, CA, pp. 325–335. https://doi.org/10.1007/978-1-4842-2199-0_8
- Weiss, N.A., 2017. Introductory statistics, 10th edition, global edition. ed. Pearson, Boston; Columbus; Indianapolis New York.
- Wickham, H., 2014. Tidy Data. J. Stat. Soft. 59. https://doi.org/10.18637/jss.v059.i10
- Wirth, R., Hipp, J., 2000. Crisp-dm: towards a standard process model for data mining.