Investigating the Impact of Weather on Demand Prediction for Bike Sharing System

MSc
Data Analytics

Dawn Walsh
Student ID: x19190352

School of Computing
National College of Ireland

Supervisor: Pramod Pathak & Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Dawn Walsh
Student ID:	x19190352
Programme:	MScData Analytics
Year:	2021
Module:	Research Project
Supervisor:	Pramod Pathak & Paul Stynes
Submission Due Date:	16/08/2021
Project Title:	Investigating the Impact of Weather on Demand Prediction for Bike Sharing System
Word Count:	
Page Count:	15

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Dawn Walsh
Date:	17th September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Investigating the Impact of Weather on Demand Prediction for Bike Sharing System

Dawn Walsh
x19190352

Abstract

Rebalancing a bike sharing system involves removing bikes from oversubscribed stations and putting them into undersubscribed stations in order to satisfy demand. Predicting the number of spaces required depending on day, time and weather is a challenge. This research proposes to investigate the two most prevalent prediction methods in conjunction with weather data to find whether clustering or tree methods provide a better model for improving demand prediction to aid system rebalancing. The previous research on DC Bikes was replicated, in Dublin however weather has much less of an impact on Dublin Bikes usage. Random Forest Classification gave better demand prediction for rebalancing the bike system at a station level and a model that combines the two methods may well be better overall than either individually.

1 Introduction

Bike sharing systems (BSS) are rising in popularity schemes have expanded world-wide, with most countries in Europe, the Americas and Asia having at least one such scheme in a major city. Further roll-out and expansion of most schemes have slowed due to the difficulties faced in trying to balance a system that is inherently unbalanced. These systems are often touted as the "last mile" solution as a link between public transport and a users workplace Shaheen et al. (2014); Zamir et al. (2017). Early morning usage going from the outside in, leaving outer stations empty and inner stations full up, and vice versa in the evening time. This is not just a problem faced by docked biking systems, it has also been observed in the dockless bike systems Li et al. (2019). While there is significant interest in this area of research the vast majority of the current research is carried out in the Far East and in the United States. There has been little current research carried out in smaller cities or even in Europe in general.

The aim of this research is to investigate to which machine learning frameworks can best assist owners to rebalance their bike sharing system. It is clear that weather does have a major impact on demand in cities such as Washington DC Quach and Malekian (2020) or Seoul Sathishkumar et al. (2020), however in a milder climate like Ireland's where weather extremes are unusual does it have the same kind of impact? There doesn't seem to be a consensus on which ML methods are more effective, however most recent research seems to fall into two camps, Clustering Methods or Tree Methods. It is proposed to investigate the effectiveness of both K-Means Clustering and a Random Forest Classifier for demand prediction on the dublinbikes dataset in conjunction with weather data.

Since most research in this area points to the need for further analysis to be carried out in other cities to expand the knowledge baseLi and Zheng (2020) the major contribution of this paper will be a direct comparison of these two methods on a BSS that is currently unresearched.

Bike systems can tell us so much about how people use the city that they live in. This means that predicting demand for BSS is extremely valuable when it comes to planning and policy-making. The information would feed into infrastructure improvement and facilitate long-term traffic management strategies Xu et al. (2019). The major problems that face bike-sharing schemes are replenishment and re-balancing of stations. If there are no bikes at a station the user can't hire a bike which is a lost customer, too many of those and the BSS won't survive, equally if the station is full up returning a bike can be a frustrating enterprise and may put the user off from future use. Attempting to optimise the system depends on predicting demand at both a station and a system-wide level so that re-balancing can be planned and carried out in a timely fashion.

This paper discusses related work in Section 2 with a focus on demand prediction and traffic analysis. Section 3 contains a discussion of the research methodology employed in this investigation. Section 6 contains details of the experiments carried out and the results achieved. Section 7 contains a discussion of the conclusions that have been drawn from the research alongside future work that could be carried out.

2 Related Work

2.1 Rebalancing

One of the major challenges within BSS is the issue of re-balancing a system that is inherently unbalanced. A novel approach is taken by looking at the Origin and Destination of a bike as a pair by F. et al. (2021). The authors carried out the research with a specific view to tracing mobility flows through a city to aid city planning. The authors only carried out a very rudimentary machine learning investigation for predicting usage as it was not the focus of the paper but have mentioned that initial investigations indicated that using a Random Forest model taking points of interest, weather and demographic data gave good initial prediction results.

Another area in which deep learning methods are utilised is in traffic forecasting. This helps to create a better understanding of how a system can be efficiently rebalanced. Currently Graph Neural Networks are a major area of focus for traffic forecasting Jiang and Luo (2021). These traffic networks look at bicycle flow along with other types of traffic and could be used in conjunction with demand prediction to better aid re-balancing. The lack of overall quality data has hindered models performance in all cities.

If a station has several damaged bikes it will appear to have bikes available but as far as users are concerned it is empty. So the problem of faulty bikes also has an impact on rebalancing and it stands to reason then that we also need to consider repairing/replacing faulty bikes Usama et al. (2020). In this study the authors only considered dockless biking systems, however the findings could be applied to docked systems. The authors found that while their solution were useful for small and medium sized systems it had some serious scaling issues with larger systems. It can be seen how re-balancing BSS in a large urban area is technically challenging and computationally expensive.

2.2 Use Prediction

The decision of the location of docking stations is another challenge to BSS, the stations ought to be placed to ensure visibility, ease of access and balanced usage. In order to do this it is necessary to be able to predict when and where the bikes will be in most demand. Of course the obvious answer is weekdays during rush hours. The current research uses Clustering or various Tree methods to try to predict both short-term and longer term demand within the system Ashqar et al. (2020); Sathishkumar et al. (2020).

Despite there having been numerous researches carried out focusing on global prediction (i.e. predicting the total number of bikes that will be required on a given day), not much has been carried out on predicting at a station level. This is likely to be down to with how unpredictable overall this can be Li and Zheng (2020). This research suggests that only approximately 18% of trips taken on BSS in New York and Washington DC on a daily basis are repeat trips, i.e. most trips are random. In order to more closely predict demand, stations are formed into clusters using their physical locations. This lead to using a hierarchical clustering model in combination with time series data to predict usage at a cluster of geographically close stations. The authors also took into account the impact of weather and weekdays vs weekends.

Random Forest or tree-based methods to predict user-demand are an option that avoids the ever-present risk of over-fitting. Random Forest can deal with both categorical and numerical variables easily and rank the most important ones based on their contribution to a model Ashqar et al. (2020), it is also extremely robust when dealing with colinearity. This research also used Partial Least Squares Regression (PLSR) to reduce the number of models required, PLSR did not work as well as Random Forest at a station level, however when applying it to the whole network it worked well enough given the reduction in complexity. The authors also accounted for days of the week, time of year and weather in their models, but have not gone into much detail as to how this is incorporated.

2.3 Weather impact

Ireland rarely has extreme weather and our annual rainfall is actually quite low, it rains little and often here and this has an impact on road conditions and obviously the bikes themselves (who wants to use a wet saddle). This means that no calculation of demand prediction on BSS can really be considered complete without taking the weather into account. It is known that BSS usage varies by weekday and time of day but it also varies between seasons, indicating that weather, specifically temperature and precipitation have a significant impact on demand Quach and Malekian (2020). In this instance the authors using Washington DC for their data, used k-means Clustering, and unusually instead of forming clusters using spatio-temporal information the clusters were formed based on their usage statistics. With one cluster having a mean number of trips per day that was almost twice as large as the other two clusters. Temperature and precipitation had a significant impact on usage in a manner that was also pretty intuitive, with rain have a negative impact on usage and warm temperatures having a positive impact.

Further investigations into the weather impact were carried out by Sathishkumar et al. (2020) on the Seoul BSS. The authors also came to the conclusion that weather very significantly impacted the demand within BSS. Several Machine Learning techniques including Support Vector Machines (SVM) and XGBoostTree were investigated and perhaps surprisingly showed the Tree models significantly outperformed a basic Linear model but also the more complex SVM as well. A second paper also investigated weather impact on BSS in Korea Kim (2018), however the authors used a hierarchical clustering method to group stations with similar characteristics together to make demand prediction easier. The author found that while temperature seemed to have a significant impact on two of the clusters, particularly high temperatures, it did not have the same impact on the other, leading the author to speculate that this may be to do with the fact that this cluster sees its highest demand in the morning when temperatures would be at their lowest.

With the climate of Ireland being significantly more temperate than either Seoul or Washington DC it will be interesting to see if these findings are borne out on the Dublin Bikes set.

3 Methodology

The research methodology of this research discusses the step-by-step process as shown in Fig 1.

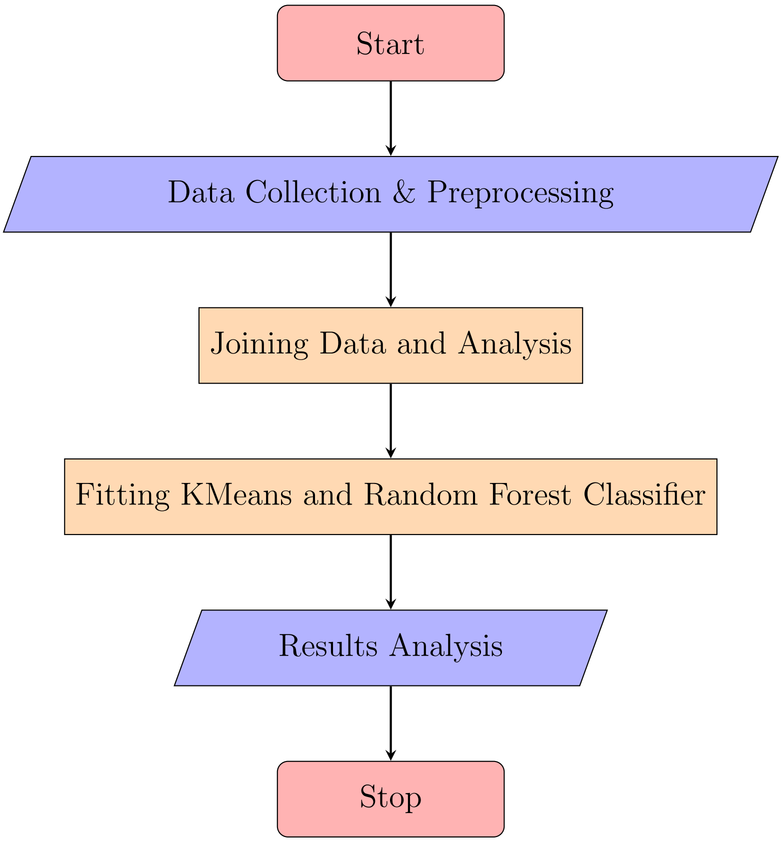

Start
Data Collection & Preprocessing
Joining Data and Analysis
Fitting KMeans and Random Forest Classifier
Results Analysis
Stop

Figure 1: Research Methodology Flow Chart

3.1 Data Collection

For this study Dublin Bikes station data and Met Eireann weather station data is analysed.

3.1.1 Dublin Bikes

The Dublin bikes data is available as an API from the Smart Dublin website [1]. In order to follow on from other works, approximately 2 years of data will be retrieved. Since each station updates its status every five minutes this will be approximately 144 million rows of data. This will be reduced to approximately 110 million or less by disregarding the hours of midnight to 6am when the stations are closed for removal of bikes, it will then be further reduced by creating various prediction windows and averaging the number of available bikes during that period. Most research suggest that 15 minutes Ashqar et al. (2020); Kim (2018) is the optimal window.

An initial investigation of the Dublin Bikes set shows that there is definitely a relationship between time of day and demand, however as in other research Quach and Malekian (2020) we can see that it follows a double peak during weekdays Fig 2, coinciding with traditional rush hours with a little interim spike that falls around lunchtime and during the weekend follows a more normal distribution.

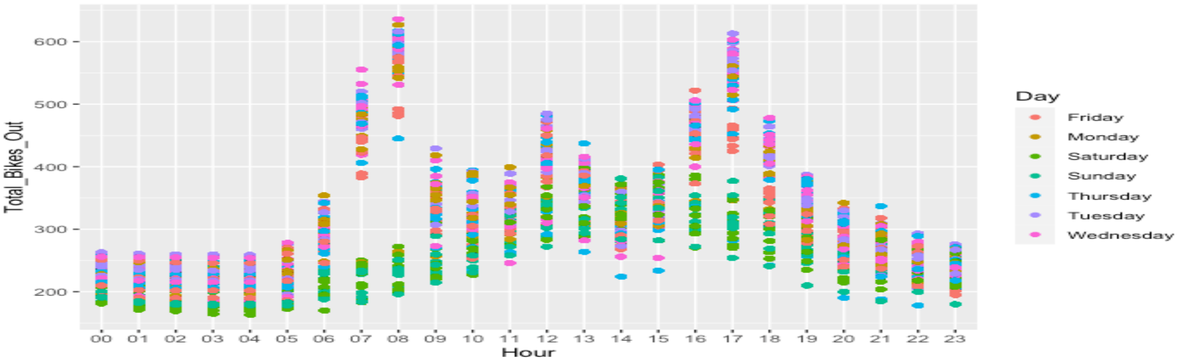


Figure 2: Dublin Bikes Daily Usage

3.1.2 Weather Data

The weather data is available from Met Eireanns website [2]. The information is given by station and can be downloaded as .csv files for specified time period or all the data for a given station can be downloaded as an hourly, daily or monthly series depending on requirement. This data given temperature, precipitation rates and humidity amongst other information, with some station's such as the one at Dublin Airport having much more detailed information than others.

3.2 Data Analysis

This research applied both K-Means clustering and Random Forest Classification (RFC) to analyse the usage of the Dublin Bikes stations according to both their spatial and temporal similarities to one another. The stations are scattered around Dublin city centre, mostly situated around the main business and shopping areas with a western stretch going to the main transport hub of Heuston station along the red Luas line Fig 3.

[1] `https://data.smartdublin.ie/dataset/dublinbikes-api`

[2] `https://www.met.ie/climate/available-data/historical-data`

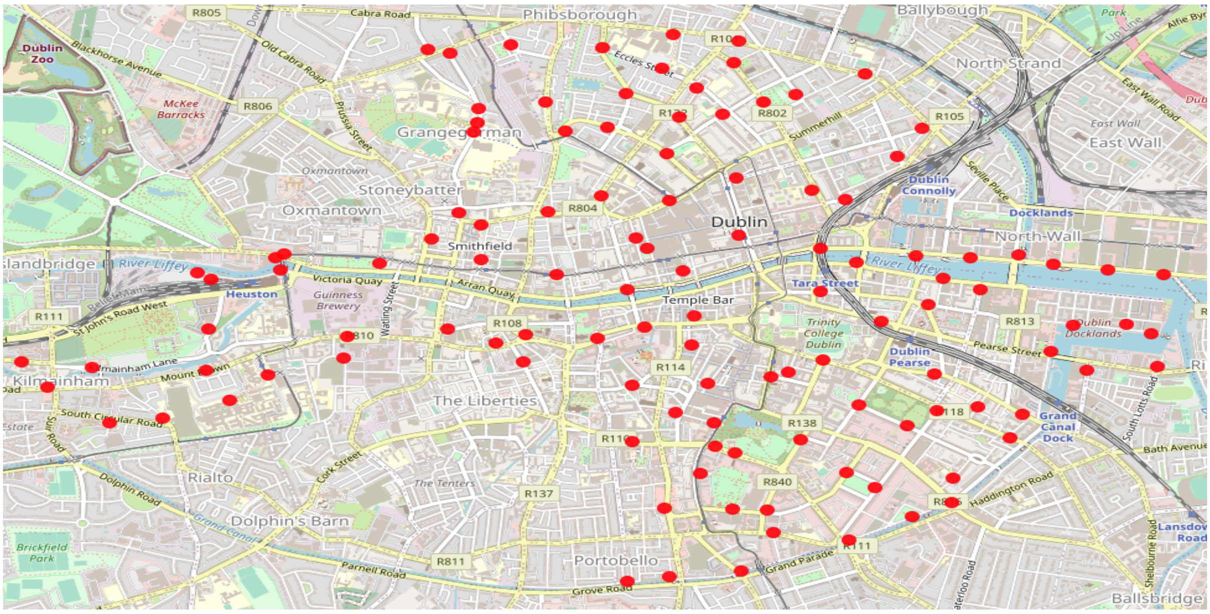
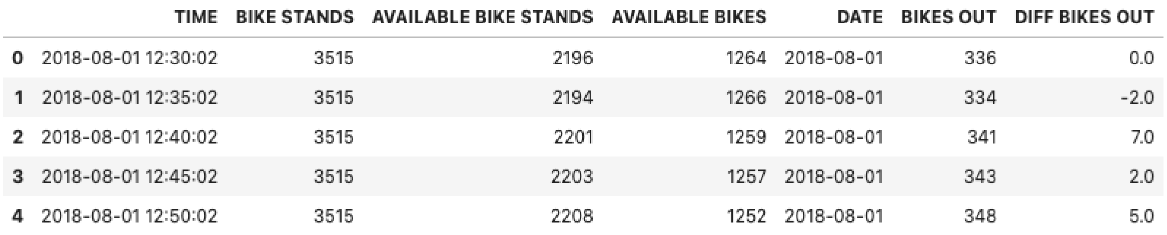


Figure 3: Dublin Bikes Map

3.2.1 Feature Extraction

The Dublin Bikes data does not give trip information so this detail has to be extrapolated by grouping the information by time to figure out how many bikes are out and then differencing from one time stamp to the next, since we are only interested in bikes being taken out (i.e. the start of a trip), negative totals (which indicate more bikes returned than removed and mean the end of trip) are replaced with zeros as shown in Fig 4.



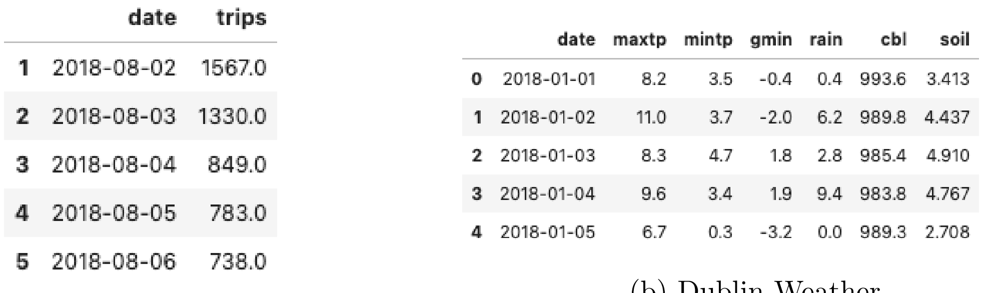
	TIME	BIKE STANDS	AVAILABLE BIKE STANDS	AVAILABLE BIKES	DATE	BIKES OUT	DIFF BIKES OUT
0	2018-08-01 12:30:02	3515	2196	1264	2018-08-01	336	0.0
1	2018-08-01 12:35:02	3515	2194	1266	2018-08-01	334	-2.0
2	2018-08-01 12:40:02	3515	2201	1259	2018-08-01	341	7.0
3	2018-08-01 12:45:02	3515	2203	1257	2018-08-01	343	2.0
4	2018-08-01 12:50:02	3515	2208	1252	2018-08-01	348	5.0

Figure 4: Dublin Bikes Data

The data is finally gathered together to give total trips started on a given day so the Dublin Bikes data set is left boiled down to the date and the number of trips taken Fig 5a which is joined with the weather data Fig 5b

4 Design Specification

The methodology as described in the previous Section 3 is shown in Fig 6 below. The two machine learning methods that were investigated are discussed in more detail in this section.



	date	trips
1	2018-08-02	1567.0
2	2018-08-03	1330.0
3	2018-08-04	849.0
4	2018-08-05	783.0
5	2018-08-06	738.0

(a) Dublin Bikes Trips

	date	maxtp	mintp	gmin	rain	cbl	soil
0	2018-01-01	8.2	3.5	-0.4	0.4	993.6	3.413
1	2018-01-02	11.0	3.7	-2.0	6.2	989.8	4.437
2	2018-01-03	8.3	4.7	1.8	2.8	985.4	4.910
3	2018-01-04	9.6	3.4	1.9	9.4	983.8	4.767
4	2018-01-05	6.7	0.3	-3.2	0.0	989.3	2.708

(b) Dublin Weather

Figure 5: Dublin Bikes & Dublin Weather

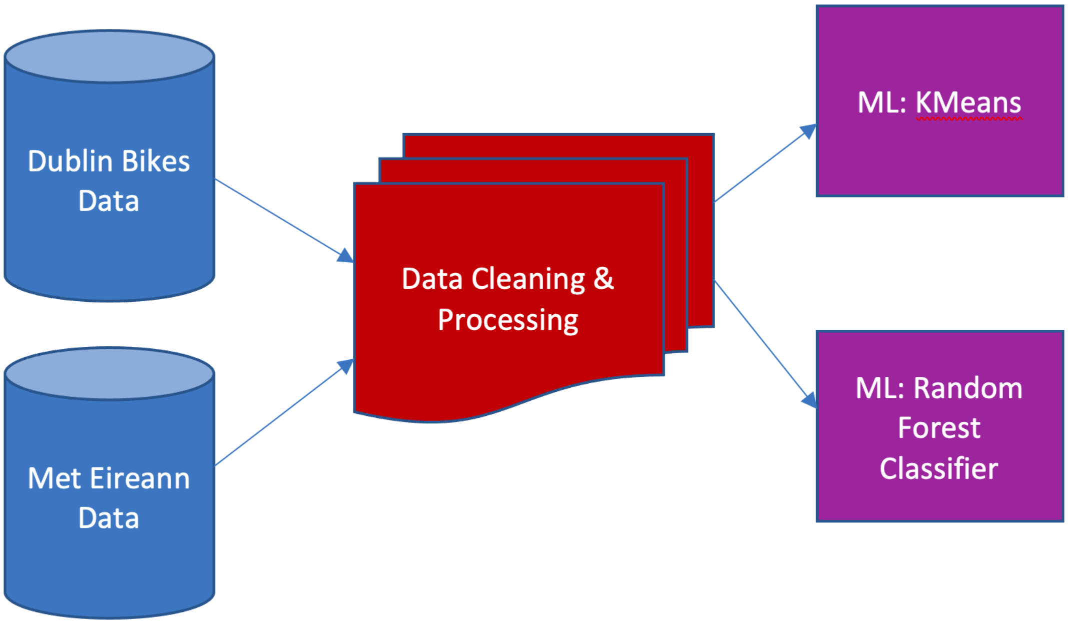


Figure 6: Design Architecture Diagram

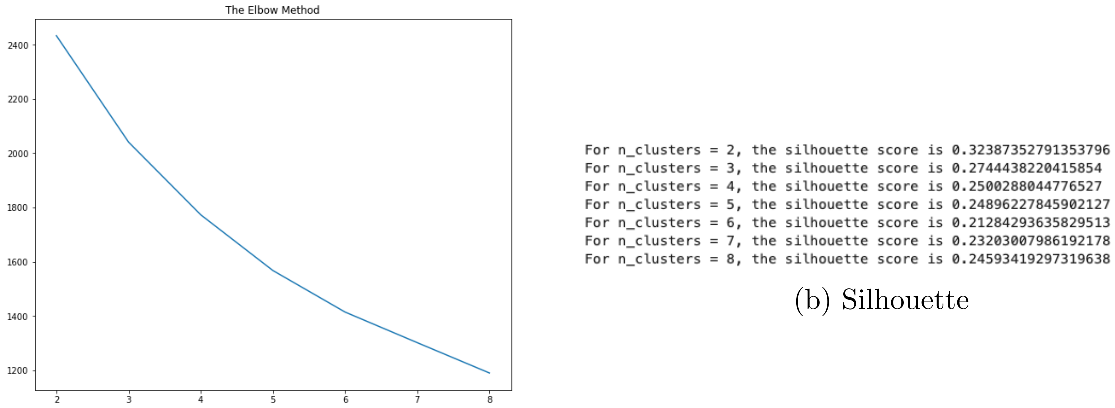
4.1 Cluster Analysis

After the feature extraction was completed, K-Means clustering was applied to the dataset. K-means clustering aims to partition each observation into k distinct clusters based on how ”close” that observation is to the centre of that cluster. In the end the centre of each cluster is the mean of all the data points within that cluster.

Both the Elbow and Silhouette methods were used to help determine the best value for k, both are shown below in Fig 7. The Elbow curve indicates that the best k could be 3 or 5 and the silhouette method confirms that 3 is best as the closer the value is to 1 the more distinct the clusters are.

4.2 Random Forest Classifier

The RFC is an ensemble method which fits multiple decision trees to the data using random splits of the dataset. It works on the divide-and-conquer approach, each tree



```
For n_clusters = 2, the silhouette score is 0.32387352791353796
For n_clusters = 3, the silhouette score is 0.2744438220415854
For n_clusters = 4, the silhouette score is 0.2500288044776527
For n_clusters = 5, the silhouette score is 0.24896227845902127
For n_clusters = 6, the silhouette score is 0.21284293635829513
For n_clusters = 7, the silhouette score is 0.23203007986192178
For n_clusters = 8, the silhouette score is 0.24593419297319638
```

(b) Silhouette

(a) Elbow

Figure 7: Finding Optimal K

gets a vote on the outcome and it is a case of winner takes all at the end. Variable Importance factors are extracted at the end to tell which variables in the data had the largest influence on the outcome of the classifier. This can be used to remove some variables to try and improve the outcomes.

5 Implementation

The time span from August 2018 to mid-March 2020 is explored as August 2018 is the earliest data available for Dublin Bikes and post March 2020 is when the lock-downs due to the COVID-19 pandemic began and people began to work from home rather than commuting to their normal place of work.

The models are implemented using Python using the scikit-learn library Pedregosa et al. (2011). The scikit-learn library supports most current machine learning methods along with evaluation metrics and visualisation tools. Seaborn and matplotlib have been choosen for creating both the preliminary visualisations of the data as well as post fitting of the various machine learning methods to visualise the results.

Following on from the initial cleaning and pre-processing several different clustering and tree methods were fitted, including a Random Forest Classifier to explore whether the supervised or unsupervised methods give better prediction results.

The data from both Met Eireann and Dublin Bikes is very clean, there is no missing data, however the weather data has several columns that are zero-filled which were removed. Other than that all the columns are retained

6 Experiments & Discussion

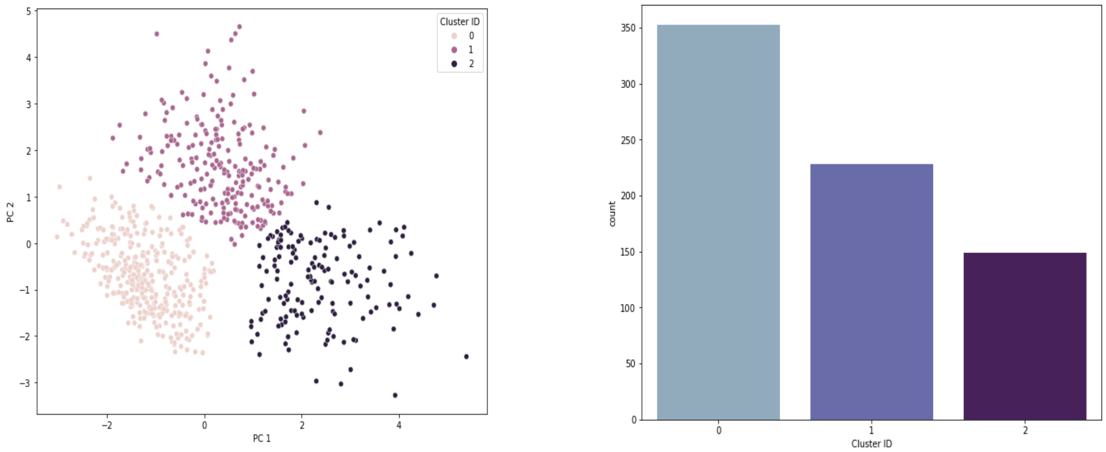
Several experiments were carried out during the investigation into the two methods that are to be compared, some of which have been set out below.

6.1 Experiment 1: DC Bikes Set Clustering Analysis

The aim of this experiment is to replicate the work carried out in “Weather impact on bike sharing using Clustering Analysis” Quach and Malekian (2020) which uses Clustering

Methods with the Washington DC bike set from [3] and weather data from [4].

A K-Means Clustering Analysis was carried out based on their work as set out in the paper above. The data was retrieved from the same sources for the same time period and cleaned in the same fashion. The results achieved replicated those found by the authors. Specifically that 3 distinct clusters were found in the data (there could conceivably be 4 but 3 won out as the marginally better option after the Elbow and Silhouette methods were employed.



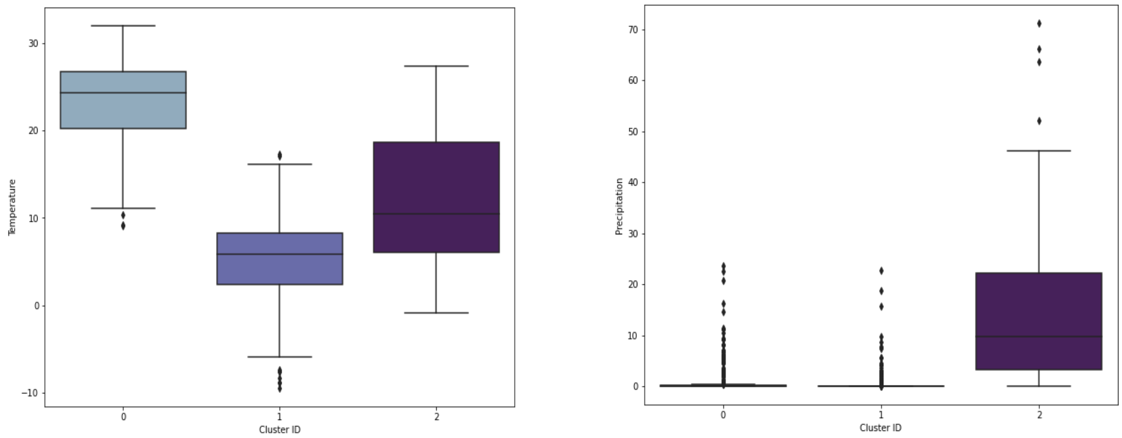
(a) Clusters DC

(b) Clusters Size

Figure 8: Clusters and Cluster Size

The clusters also displayed the same behaviour as discussed in the work. Namely that in terms of size one cluster accounts for just shy of half of the data points. The other two clusters accounting for 30% and 20% respectively.

The clusters displayed the same behaviour as shown in the original work including the usage statistics in terms of the clusters based on both temperature and precipitation



(a) Clusters and Temperature

(b) Clusters and Rainfall

Figure 9: Weather Impact on DC Clusters

[3] https://s3.amazonaws.com/capitalbikesharedata/index.html

[4] https://www.visualcrossing.com/weather-data

It would be useful to determine if the same behaviour is displayed within other Bike sharing systems with different weather data.

6.2 Experiment 2: Clustering on the Dublin Bikes data with Weather

The purpose of this experiment was to carry out a Cluster Analysis on the Dublin Bikes and Dublin Weather data. The Dublin Bikes data is presented in a slightly different manner to the DC Bikes data, in that it does not give actual trip data, only data from the stations that update every 5 minutes. The data is available from August 2018 to January 2021, so the assessment is carried out from August 2018 to mid-March 2020 to get as much time pre-Pandemic as is available, if the data covering the pandemic period is left in it skews the usage statistics and causes the clusters to be far closer together.

The weather and trip data is then joined on date and a correlation matrix calculated. It is very evident that correlation between weather and trips is quite low. The highest correlation is a negative one between precipitation and trips.

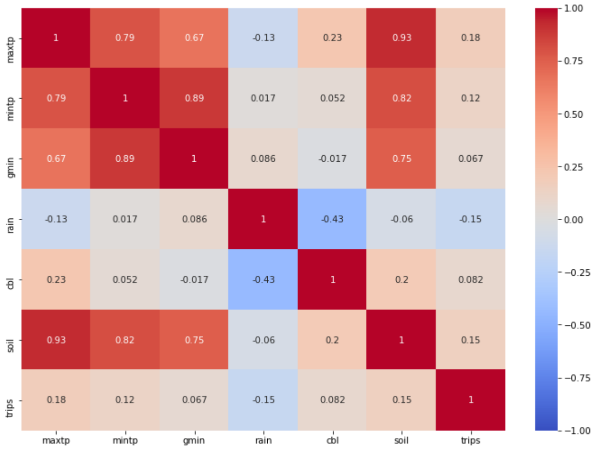


Figure 10: Bikes & Weather Correlation

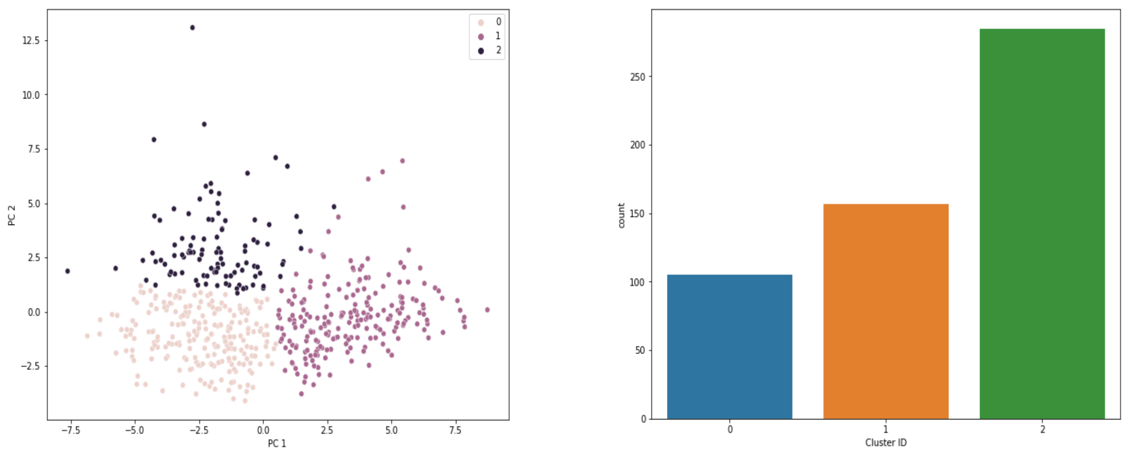
In order to create clusters first the data is Standardised. Using the the Elbow and Silhouette methods it is clear that the optimum number of clusters for the Dublin Bikes data is also 3. Once the clusters are created they can be investigated visually for the differences between them. The clusters are visualised and the number of trips per cluster are investigated to see if they behave similarly to the DC clusters.

Similarly to the DC Clusters 1 of the Clusters seem to contain over half of the overall data points with the other two splitting the remaining points 60/40. It is also clear that weather seems to have very little impact on the overall demand for bikes on a given day Fig 12.

Further investigation of the data on an hourly basis may give better insight.

6.3 Experiment 3: Clustering on the Dublin Bikes data with Weather on an Hourly Basis

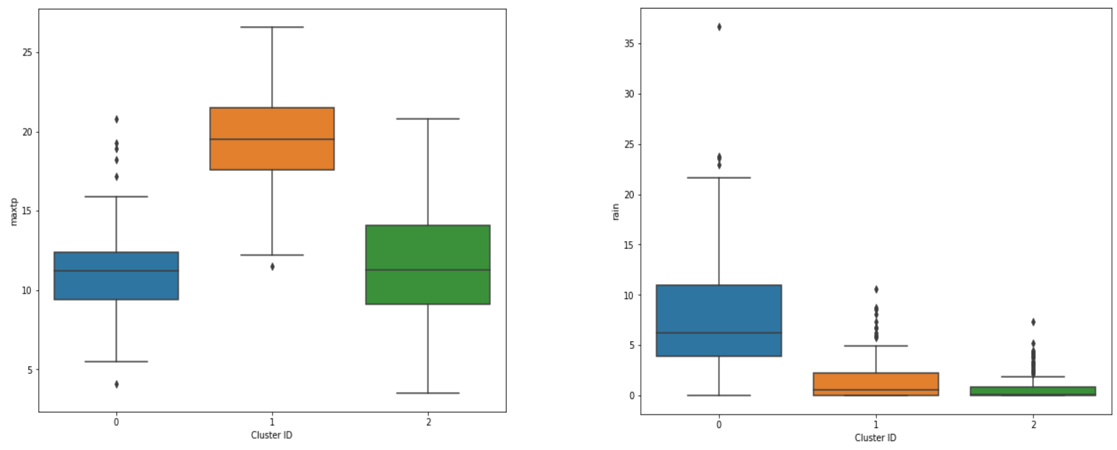
Similar to the previous experiment however the data was looked at on an hourly rather than daily basis. The features had to be engineered slightly differently to allow for the number of trips taken on an hourly basis within the system. The number of clusters



(a) Cluster Visualisation Dublin Bikes

(b) Cluster Size Dublin Bikes

Figure 11: Dublin Bikes Clusters



(a) Clusters and Temperature

(b) Clusters and Rainfall

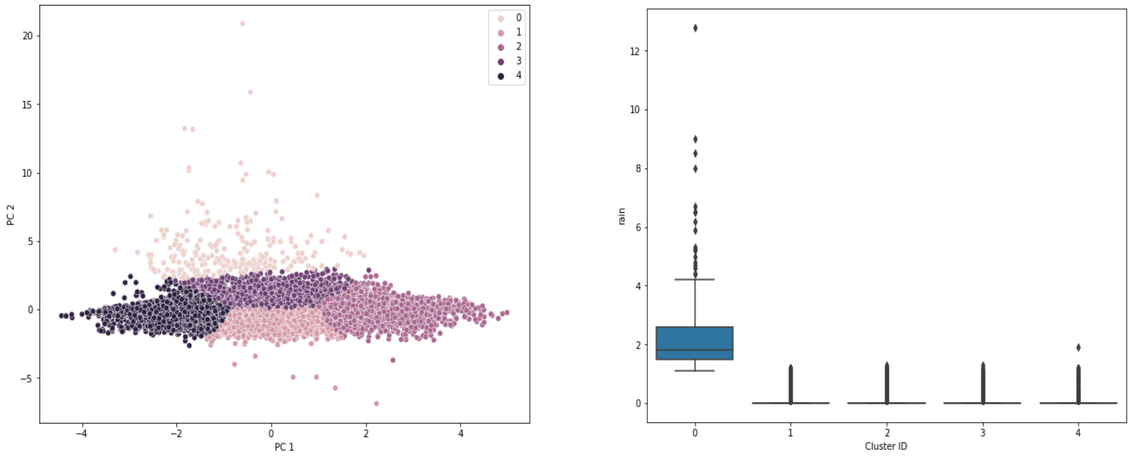
Figure 12: Weather Impact on DB Clusters

chosen was 5 as given by the silhouette method. It can be seen that on an hourly basis weather does have a slightly greater impact on usage, however there is no real clear definition between the clusters when they are visualised Fig 13. It can be seen that precipitation has a large impact on the formation of one of the clusters.

The data points are very close together, probably due to the fact that when we look at the data over the course of 18 months and break it down to hourly trips we end up with nearly 13000 data points. There were two further experiments carried out that are set out briefly next followed by the Random Forest Classifier experiment.

6.4 Experiments 4 & 5: Station IDs and Reducing time to Quarter Hour

Similarly to the previous two experiments the same bikes and weather data were used for both of these experiments. In Experiment 4 the Station ID was retained to see if it had any influence on the daily clusters. This added a lot more data points to the analysis but the clusters behaved in a very similar fashion to the ones in Experiment 2 they just



(a) DB Clusters Hourly

(b) Clusters and Rainfall

Figure 13: Weather Impact on Dublin Bikes hourly Clusters

Occupancy Rate	Class Label	Label Meaning
<10%	0	Empty
10-80%	1	Balanced
>80%	2	Full

Table 1: Dublin-Bikes Class Labels

weren't as clearly defined.

Experiment 5 took the hourly bikes data experiment and tried breaking it into 15 minute sections as had been suggested in Ashqar et al. (2020) however in this instance again there is very little difference in the behaviour of the clusters. Perhaps if the data were configured more like it is in Experiment 6 the clustering on 15 minute increments might give a little more insight.

6.5 Experiment 6: Random Forest Classifier

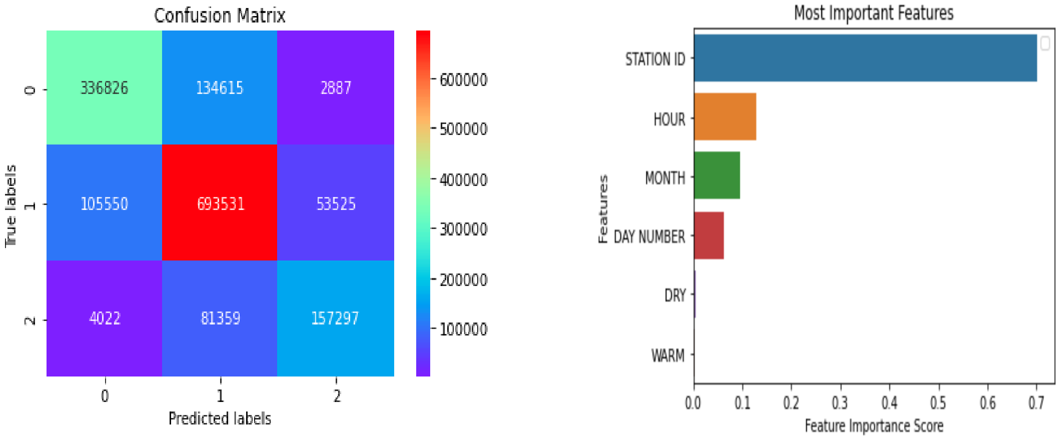
In order to fit a Random Forest Classifier some research suggested that the occupancy rates for the stations should perhaps be split into 3 classes, shown in Table 1. From previous work Quach and Malekian (2020); Sathishkumar et al. (2020) and from the earlier experiments it seems it's only really whether it is wet or not or warm or not that has an affect on usage in terms of weather. So the weather data was boiled down to this, so if there was more than 3mm of rain on a given day it was classed as "Wet" otherwise "Dry" and if the temperature is over 18C then "Warm" otherwise not.

The data was split using an 80:20 holdout and a Random Forest Classifier was fit to the training set and predictions were obtained to see if the status of the station could be accurately predicted. The overall accuracy of predictions was 75.6%. The classification report is given in Table 2.

In order to see how the predictions fell into the various categories a confusion matrix was also generated for the true vs. predicted classes and is shown in Fig 14a. The importance of the variables is also visualised to give an idea of which feature is the most powerful indicator of the current state of the station Fig 14b

class label	precision	recall	f1-score	support
0	0.75	0.71	0.73	474328
1	0.76	0.81	0.79	852606
2	0.74	0.65	0.69	242678
accuracy			0.76	1569612
macro avg.	0.75	0.72	0.74	1569612
weighted avg.	0.76	0.76	0.76	1569612

Table 2: Random Forest Classifier - Classification Report



(a) RFC: Confusion Matrix

(b) RFC: Feature Importance

Figure 14: Dublin Bikes Random Forest Classifier

6.6 Discussion

Unlike previous works in other locations it is clear from both the Clustering experiments and the RFC experiment that weather has little impact on usage in the Dublin bikes system. This may be to do with either the Irish just not being all that bothered by wet weather or more likely to do with the fact that it doesn't really rain that much in Ireland (despite its reputation). The things that have a much weightier impact are location of the station time of day and whether it is a weekday or the weekend. Weather however cannot be discounted and it should be further investigated in other cities with different climates.

Moreover it was clear from the initial investigations of the datasets that the COVID-19 pandemic had a profound impact on the usage of the system overall. There is a potential for future research comparing pre-Pandemic usage and Pandemic usage and seeing if any changes carry over when lockdowns are completely lifted as many of the regular users of the system may never return to city centre offices/workplaces on a full-time basis.

7 Conclusion

In contributing to the continuing efforts to improve bike share systems rebalancing this study looked at both clustering methods and random forest classification to determine which of these two most popular methods is the more useful for predicting demand and usage when combined with weather data.

It is evident that while Clustering gives a good indication of usage patterns of different clusters it is only really on a system-wide basis. RFC gives a much better indication on a station level of usage. Future work that should be investigated would be applying the RFC to the individual clusters to improve the predictions of when and where rebalancing needs to take place.

References

Ashqar, H., Elhenawy, M., Rakha, H., Almannaa, M. H. and House, L. (2020). Network and station-level bike-sharing system prediction: A san francisco bay area case study, *ArXiv* **abs/2009.09367**.

F., K., É.C., F., de Souza H.A., F., D., P., S. and C., R. (2021). Abstracting mobility flows from bike-sharing systems., *Public Transport* p. 1–37.

Jiang, W. and Luo, J. (2021). Graph neural network for traffic forecasting: A survey.
URL: *https://arxiv.org/pdf/2101.11174.pdf*

Kim, K. (2018). Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations, *Journal of Transport Geography* **66**: 309–320.

Li, Y. and Zheng, Y. (2020). Citywide bike usage prediction in a bike-sharing system, *IEEE Transactions on Knowledge and Data Engineering* **32**(6): 1079–1091.

Li, Y., Zhu, Z. and Guo, X. (2019). Operating characteristics of dockless bike-sharing systems near metro stations: Case study in nanjing city, china, *Sustainability* **11**(8).
URL: *https://www.mdpi.com/2071-1050/11/8/2256*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**: 2825–2830.

Quach, J. and Malekian, R. (2020). Exploring the weather impact on bike sharing usage through a clustering analysis, *ArXiv* **abs/2008.07249**.

Sathishkumar, V., Park, J. and Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city, *Computer Communications* **153**: 353 – 366.
URL: *http://www.sciencedirect.com/science/article/pii/S0140366419318997*

Shaheen, S. A., Martin, E. W., Chan, N. D., Cohen, A. P. and Pogodzinski, M. (2014). Public bikesharing in north america during a period of rapid expansion: Understanding business models, industry trends and user impacts.

Usama, M., Zahoor, O., Shen, Y. and Bao, Q. (2020). Dockless bike-sharing system: Solving the problem of faulty bikes with simultaneous rebalancing operation., *Journal of Transport and Land Use* **13(1)**: 491–515.

Xu, F., Chen, F. and Liu, Y. (2019). Bike sharing data analytics for smart traffic management, pp. 69–73.
URL: *https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8904989*

Zamir, K. R., Shafahi, A. and Haghani, A. (2017). Understanding and visualizing the district of columbia capital bikeshare system using data analysis for balancing purposes.