Clustering and Forecasting Dublin Bikes Analysis

December 7, 2022

Contents

Acronyms **4**

Glossary **5**

1 Introduction **6**
- 1.1 Introduction . . . 6
- 1.2 Motivation . . . 6
- 1.3 Impact of the COVID-19 on bike sharing . . . 6
- 1.4 History of bike sharing . . . 6
- 1.5 Dublin Bikes . . . 7
- 1.6 Data Analysis . . . 7

2 Data Loading, Storage, and File Formatting **9**
- 2.1 Loading the files to disk . . . 9
- 2.2 Dublin Bikes station occupancy data dictionary . . . 9
- 2.3 Historical weather data from Phoenix park . . . 10

3 Data Analysis and Preparation **12**
- 3.1 Introduction . . . 12
- 3.2 The Dublin Bikes station occupancy data . . . 12
- 3.3 Split our "Last_updated" field into date and time . . . 14
- 3.4 Feature Engineering . . . 14
- 3.5 Bike Feature Engineering . . . 14

4 Preparng the weather Data **15**
- 4.1 Data Dictionary . . . 15

5 Results of Cluster analysis **16**
- 5.1 Folium map of clusters . . . 18
- 5.2 Daily Occupancy Patterns . . . 18
- 5.3 Informing the future network re-balancing schedule . . . 18
- 5.4 Influence of stations as a predictor variable . . . 19
- 5.5 Usage patterns based on day type . . . 21
- 5.6 Machine learning quantification of accuracy of full and empty calculations . . . 21

6 Conclusion **23**

List of Tables

2.1 Data dictionary for data from Dublin Bikes. 9

3.1 df.info() 12
3.2 unique entries for df 13
3.3 This df.isnull().sum() table tellsus that the Bike data has no missing data. 13
3.4 Bike data sample after removing duplicates, dropping STATUS and TIME fields and improving readability 14

4.1 Data Dictionary for the Weather Data 15

5.1 Clustering group output 17
5.2 Impact of Stations and weather conditions 20
5.3 This model can predict whether a station would be too full or empty to a certainty of 79%. 21

List of Figures

1.1 Subscription options range from daily to yearly with an option to rent a compatible battery that you maintain yourself *DublinBikes* (2022). 8

2.1 Dublinbikes: Real Time API, Near Real Time API, and Historical Data 10
2.2 Historical climate data is available hourly, daily and monthly for a range of parameters incuding precipitation, temperature and windspeed 11

5.1 Elbow K-means clustrering showing optimal k 16
5.2 The scikit-learn K-Means library in Python for the clustering and plotted the output on a map using the Folium library based on type of day field and time slot in day 18
5.3 Average Station Occupancy per cluster per hour, showing different weekday vs weekend trends 19
5.4 This plot shows rentals based on dry (blue) and wet (orange days). Clearly the users of the service are not much influenzed by the weather. No harm considering the irish climate! 19
5.5 This plot shows rentals based on dry (blue) and wet (orange days). Clearly the users of the service are not much influenzed by the weather. No harm considering the irish climate! 21
5.6 Plot of most influencial model features 22

Acronyms

AWS automatic weather station. 10

BSS Bike Sharing System. 7, 8, 14

DCC Dublin City Council. 6, 23

NPF National Planning Framework. 6

Glossary

MET Éireann The Irish Meteorological Service. 10

Chapter 1

Introduction

1.1 Introduction

Dublin City Council (DCC) are responsible for planning Transport and Infrastructure developments in the Greater Dublin Area. In this regard, the DCC are expected to meet the targets committed to by the Irish Government in February 2018 when the National Planning Framework (NPF) was agreed upon (Carroll and O'Sullivan 2020). The NPF is a national development and capital investment plan with targets for 2040. Ongoing and projected rapid economic and demographic change in the Greater Dublin Area leave policy makers in DCC facing two related challenges. They must:

- meet the increased transport demands that the projected one million population increase entails
- do so in a balanced and sustainable manner (Commins and Nolan 2008).

The Ireland Data Portal aims to promote innovation and transparency through the open publishing of Irish Public Sector data. Filtering through this data at DATA.GOV.IE for transport and infrastructure datasets relating to DCC gives a set of 26 datasets ranging in topic from Clamping Appeals to Multistorey Car Parking Space Availability to Telecoms Underground Infrastructure etc.

Specifically a relatively small data-set entitled **Modes of Travel in DCC** let me to a much larger historical datasets on Dublin Bikes Station occupancy's and to weather data that can be merged based on date and time and lead to several demand forecasting models that could inform DCC's planning. My modelling question asks what factors affect bike and docking demand across time and stations.

1.2 Motivation

As a global citizen I want to understand and promote more sustainable forms of travel such as public bike sharing and I have personally been a user for many years (Carroll and O'Sullivan 2020).

1.3 Impact of the COVID-19 on bike sharing

Ireland went into a pandemic induced lockdown on the 12 March as a response to COVID-19 (Colfer 2020). The 'Working from Home' (home office) strategy that this entailed had the desired effect of curtailing the spread of the coronavirus to ensure public safety (Chen et al. 2022). While of negligible concern in comparison to the impact of the disease on the population this strategy had detrimental impact on the usage rates of Dublin's urban transportation systems. It did nevertheless lead to a pronounced re-balancing of the modes of public transit being used in Dublin city (Buehler and Pucher 2022). In particular there was a marked increase in the proportional share of shared bicycle riders and the Dublin Bikes ridership proportion significantly increased during lockdown and continues beyond lockdown (Chen et al. 2022).

1.4 History of bike sharing

Bike-sharing started being promoted by governments in the 1960s as a sustainable transportation alternative mitigating against motorisation and climate change. Bike sharing has evolved from this time from through various models of provision called generations from 1st generation programs in the late 1960s to the to present day 3rd generation programs which is the main focus of this study (DeMaio 2009; Shaheen, Guzman, and Zhang 2010).

1.4.1 1st generation unlocked free bikes

Public bike sharing began in 1968 with a 1st generation program called "white bikes" in the so called city of bikes, Amsterdam. Fifty ordinary bikes were painted white and left unlocked distributed throughout the city for unmonitored communal use. Unsurprisingly, the program was abused with theft and vandalism and was abandoned in less than a week.

1.4.2 2nd generation coin-deposit bike sharing systems

Coins were stored securely in a box on the bike but the system had no monitoring or control of fleet availability and distribution.

1.4.3 3rd generation Information technology-based systems

These use real-time information accessible by API.

(d) insurance and liability concerns, and (e) prelaunch considerations. Although limited in number, several studies have documented bikesharing's social and environmental benefits, which include reduced auto use, increased bicycle use, and a growing awareness of bikesharing as a daily mobility option. Despite bike-sharing's ongoing growth, obstacles and uncertainty remain: these include future demand, safety, sustainability of business models, limited cycling infrastructure, challenges to integrate with public transportation systems, technology costs, and user convenience (e.g., limited height adjustment on bicycles, lack of cargo space, and exposure to weather). More research is needed for a better understanding of bikesharing's effects, operations, and business models in light of its reported growth and benefits.

1.5 Dublin Bikes

Dublin Bikes is a bike sharing scheme in operation from bicycle docks and stations in Dublin City.

Dublin Bikes is one of the earliest 3rd generation Bike Sharing System (BSS) providing on-demand bike rentals for customers in Dublin in conjunction with their own data-driven real time app showing customers the numbers of free bikes and docks available in real time across the network. Users can unlock bikes from any one of 111 stations throughout the city, and return them to the same or any other station. They pay for the service either through a yearly subscription or by purchasing 3-day or 24-hour passes. Users can make an unlimited number of trips, with trips under thirty minutes in length having no additional charge; longer trips incur overtime fees." They even rent out batteries on a yearly basis for €60 which modify the push bike to an electrically propelled one. They provide a bank of quarterly historic bike data in static data files that range from the final two quarters of 2018 to the first few days in 2022. These datasets provide information like the number of bike stands, number of bikes and number of free docks resolved into ten-minute periods for all 105 Bike Stations.

As the research question here is to forecast demand and optimum resource re-balancing schedules and as the lock-downs due to the COVID-19 pandemic have led to a p Clustering and Forecasting-driven analysis of Dublin Bikes public data to examine movement patterns and behaviours

1.6 Data Analysis

Data analysis tends to follow a step-by-step process and this project is no exception. Each stage requires different skills and know-how. To get meaningful insights, though, it's important to understand the process as a whole. An underlying framework is invaluable for producing results that stand up to scrutiny.

In this post, we'll explore the main steps in the data analysis process. This will cover how to define your goal, collect data, and carry out an analysis. Where applicable, we'll also use examples and highlight a few tools to make the journey easier. When you're done, you'll have a much better understanding of the basics. This will help you tweak the process to fit your own needs.

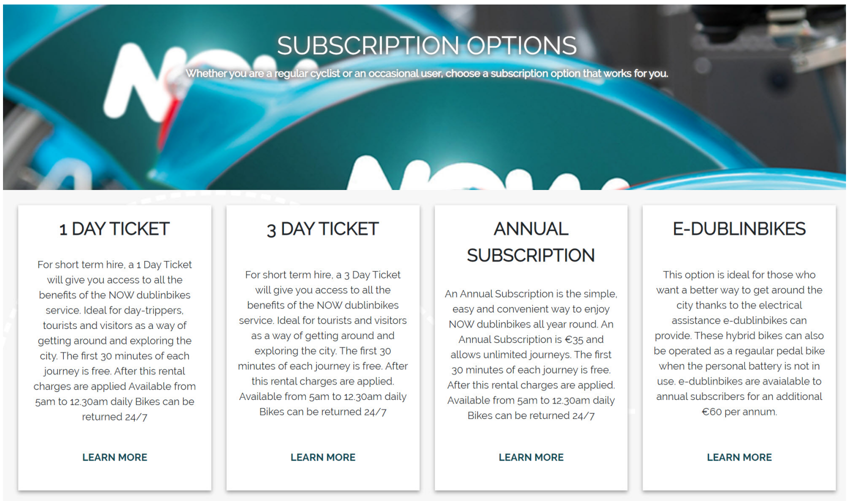

Figure 1.1: Subscription options range from daily to yearly with an option to rent a compatible battery that you maintain yourself *DublinBikes* (2022).

In more and more cities a Bike Sharing System (BSS) are playing a critical roles in transportation services.

Their on-demand scheduling and flexible routing are examples of the many factors that influence service demand at the individual level (Zhou, Wang, and D. Li 2019). Bike Sharing System (BSS) have added new dimensions to urban mobility. A primary challenge of providing public bike sharing services is to effectively plan resources so that bike distribution within the system is re-balanced. Classification studies on the available historical data show that stations show a daily temporal pattern that in the extreme can lead to stations without bikes and stations without free stands. Both of these cases have a negative impact on individual users and ultimately to the profits and viability of the service provider Thoa Pham Thi et al. (2017).

Chapter 2

Data Loading, Storage, and File Formatting

This stage of data analysis takes place before the data preparation and data visualisation stages and focuses on retrieving the required data from secondary sources in a uniform and cross-platform manner.

2.1 Loading the files to disk

The Data Loading Jupyter Notebook downloads data directly from the web and defines data files names and location in the users computer. In particular, it accesses the primary data directly from its source using a URL but avoids the inefficiency of repeatedly downloading large files by always checking for the data at a local level prior to retrieval from the primary source. In other words the programming is setup to ensure that large data files are only retrieved once from the internet to any one computer.

Python could have simply been programmed to read the individual files from their respective urls but this would be repetitive and inefficient programming. Instead it was chosen to use an *urllib.request* module to iteratively download the quarterly CSV files and in turn concatenate them to the main dataframe **df** .

2.2 Dublin Bikes station occupancy data dictionary

Table 2.1 gives what we call a data dictionary for the Dublin-Bikes data. The uppercase column headers are not reader friendly and several columns will be discarded or filtered down to leaner data.

Table 2.1: Data dictionary for data from Dublin Bikes. Analysis will reveal that TIME and STATUS can both be dropped from our dataframe and that several types aer misrepresented in pandasdue to mixed types in the field.

Data Dictionary for raw Dublin Bikes- creative commons attribution			
Column	**Type**	**Label**	**Description**
STATION ID	numeric	STATION ID	Globally unique identifier of station.
TIME	timestamp	TIME	Time of fetching the data.
LAST UPDATED	timestamp	LAST UPDATE	Time of last updated information.
NAME	text	NAME	Station name.
BIKE STANDS	numeric	BIKE STANDS	Station total number of bike stands.
AVAILABLE BIKE STANDS	numeric	AVAILABLE BIKE STANDS	Station available bike stands.
AVAILABLE BIKES	numeric	AVAILABLE BIKES	Station available bikes.
STATUS	text	STATUS	Station status (Open/Close).
ADDRESS	text	ADDRESS	Station address.
LATITUDE	numeric	LATITUDE	Station latitude.
LONGITUDE	numeric	LONGITUDE	Station longitude.

Figure 2.1 shows the range of data being made available by Dublin Bikes. The Real Time API data feeds the dynamic mobile apps that users use to plan their transport and the historical data can be used

for more long term planning and implementation.

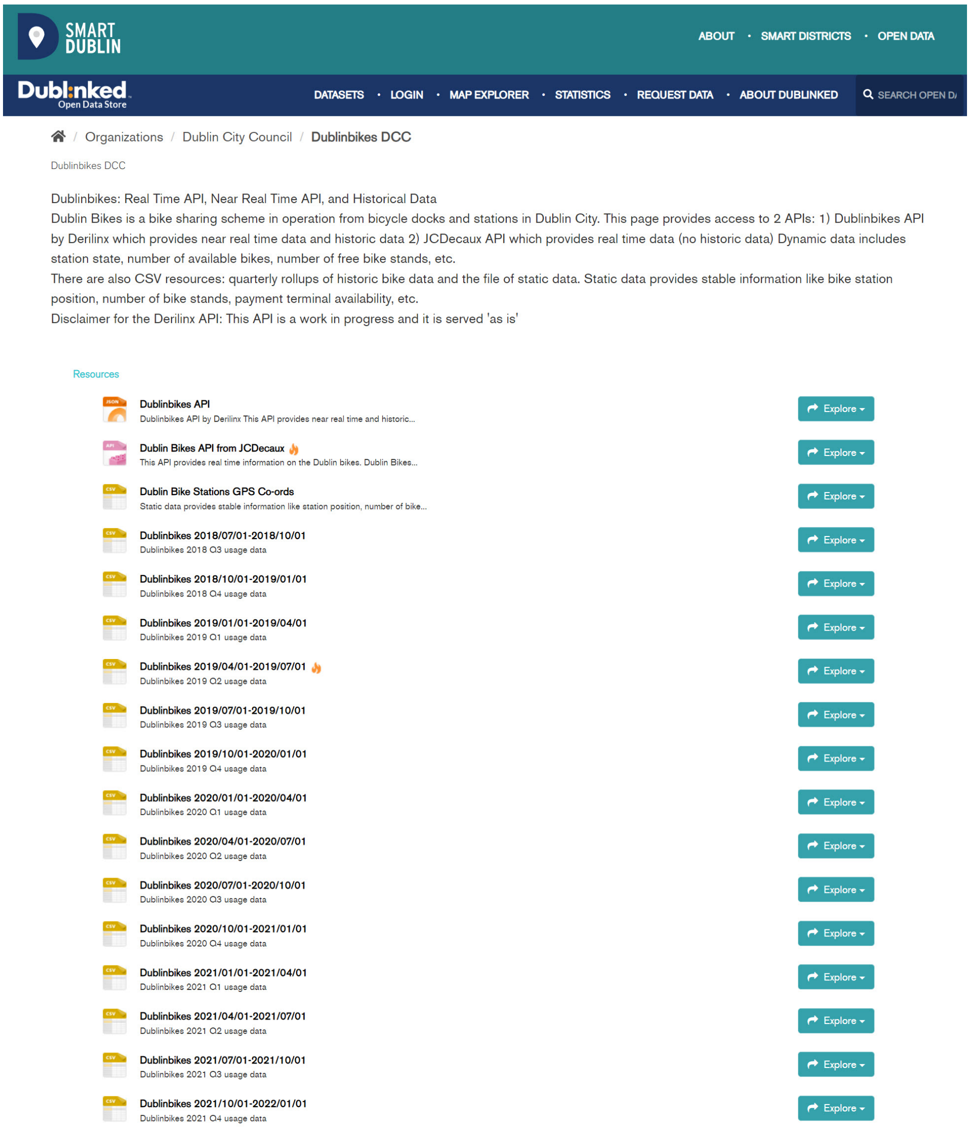

Figure 2.1: Dublinbikes: Real Time API, Near Real Time API, and Historical Data

2.3 Historical weather data from Phoenix park

There are three MET Éireann weather stations in the Greater Dublin Area, casement, Dublin Airport and Phoenix Park.

The Phoenix Park weather station is an automatic weather station (AWS) situated within the grounds of the Ordnance Survey of Ireland in the Phoenix Park, Dublin. It has been in place since 2003 when it replaced a manual climate station which was established back in 1829.

A Historical Database Query was made where the data resolution could be chosen as one of the following:(1) Hourly (2)Daily and (3) Monthly.

The naturally resolution was **hourly** because weather conditions in Ireland change several times a day so if meaningful predictor parameters are to be leveraged in our machine learning model they need to be updated on an hourly timeframe. Spatially there was a choice for County and individual weather Station using drop down-arrows and for simplicity Phoenix Park was chosen and will be assumed to be an accurate proxy for weather conditions across the entire Dublin-Bike terrain.

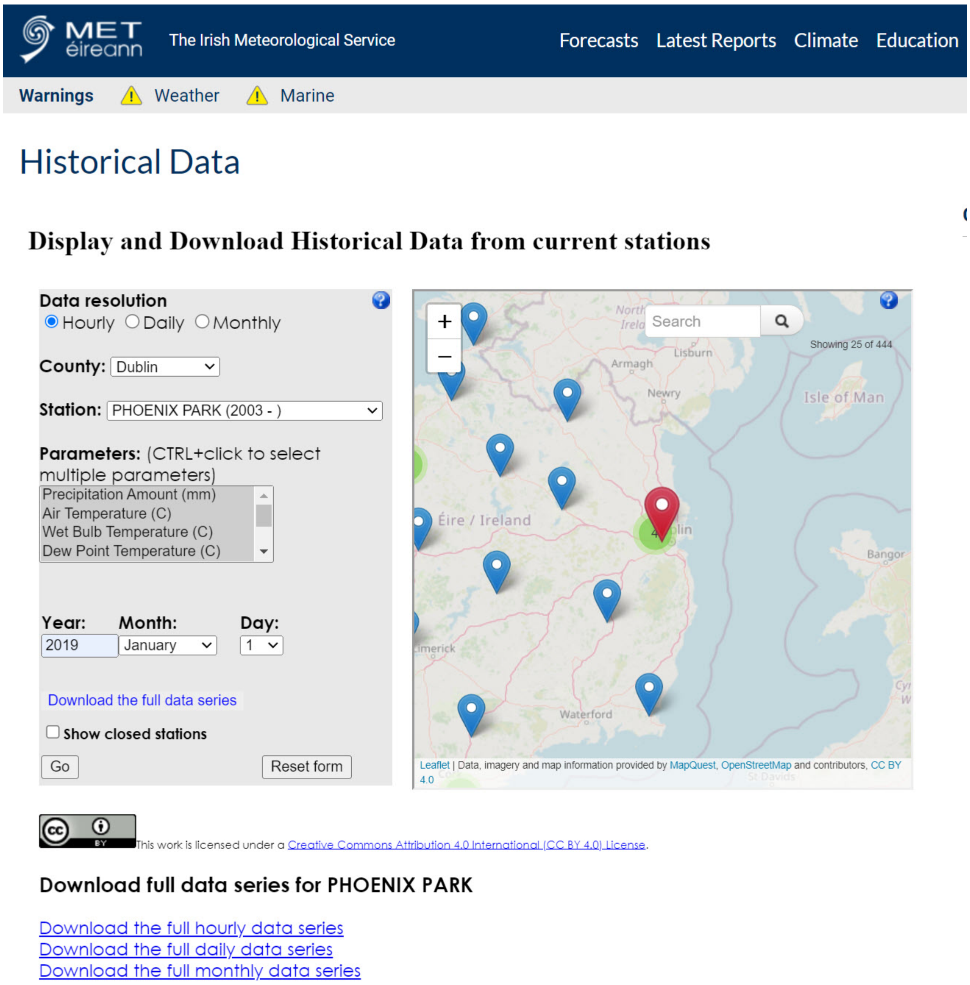

Figure 2.2: Historical climate data is available hourly, daily and monthly for a range of parameters incuding precipitation, temperature and windspeed

Chapter 3

Data Analysis and Preparation

3.1 Introduction

After running the initial ”01 Load” Jupyter Notebook a few key data files are stored in the ”data” folder. In particular 01-loaded-bikes.csv and 01-loaded-weather.csv will form the bike and weather dataframes.

3.2 The Dublin Bikes station occupancy data

The Dublin Bikes station occupancy data publicly available on the on the Smart Dublin Open Data Store, is a set of quarterly files available from the third quarter of 2018 through to the end of 2021. This data forms the backbone for this study. Taking some fields directly from the raw data, feature engineering some new ones and merging with historical weather data based on time and date stamps various classification and regression models will be developed and assessed for performance.

3.2.1 Information for the Dublin Bikes data

Table 3.1 shows data types for the Dublin Bikes data. Pandas uses ”object” for genuine categorical data and for mixed data types.

3.2.2 Unique entry count for the Dublin Bikes data

In table 3.2 we learn from the data that there are 111 Bike Stations in the network. The headers are cumbersome and not very readable due to all uppercase.

3.2.3 Missing bike data

Table 3.3 indicates that there is no missing data in the entire dataframe.

Table 3.1: <class ’pandas.core.frame.DataFrame’>RangeIndex: 11283524 entries, 0 to 11283523 Data columns (total 11 columns): dtypes: float64(2), int64(4), object(5) memory usage: 947.0+ MB

#	Column	Dtype
0	STATION ID	int64
1	TIME	object
2	LAST UPDATED	object
3	NAME	object
4	BIKE STANDS	int64
5	AVAILABLE BIKE STANDS	int64
6	AVAILABLE BIKES	int64
7	STATUS	object
8	ADDRESS	object
9	LATITUDE	float64
10	LONGITUDE	float64

Table 3.2: Number of unique entries for df]df.nunique (axis=0, dropna=True)

STATION ID	111
TIME	103038
LAST UPDATED	5039168
NAME	111
BIKE STANDS	18
AVAILABLE BIKE STANDS	41
AVAILABLE BIKES	41
STATUS	2
ADDRESS	111
LATITUDE	111
LONGITUDE	111

Table 3.3: This df.isnull().sum() table tellsus that the Bike data has no missing data.

df.isnull().sum()

STATION ID	0
TIME	0
LAST UPDATED	0
NAME	0
BIKE STANDS	0
AVAILABLE BIKE STANDS	0
AVAILABLE BIKES	0
STATUS	0
ADDRESS	0
LATITUDE	0
LONGITUDE	0

dtype: int64

Table 3.4: Bike data sample after removing duplicates, dropping STATUS and TIME fields and improving readability

	Id	Last_updated	Name	Total	Docks	Bikes	Address	Latitude	Longitude
9417969	49	2021-10-30 21:58:09	GUILD STREET	40	23	16	Guild Street	53.347931	-6.240928
4497842	31	2021-05-26 13:25:47	PARNELL STREET	20	13	7	Parnell Street	53.350929	-6.265125
4627456	50	2021-05-30 03:22:28	GEORGES LANE	40	16	24	George's Lane	53.350231	-6.279696
9441416	18	2021-10-31 07:51:06	GRANTHAM STREET	30	16	14	Grantham Street	53.334122	-6.265436
5518646	100	2021-06-27 11:18:53	HEUSTON BRIDGE (SOUTH)	25	2	22	Heuston Bridge (South)	53.347107	-6.292041
3817417	68	2021-05-04 17:15:29	HANOVER QUAY	40	24	16	Hanover Quay	53.344116	-6.237153
6690674	96	2021-08-04 04:50:30	KILMAINHAM LANE	30	27	3	Kilmainham Lane	53.341805	-6.305085
6795037	16	2021-08-08 15:37:45	GEORGES QUAY	20	11	9	Georges Quay	53.347507	-6.252192
5086776	6	2021-06-14 04:20:37	CHRISTCHURCH PLACE	20	20	0	Christchurch Place	53.343369	-6.270120
7048828	26	2021-08-16 20:52:14	MERRION SQUARE WEST	20	8	12	Merrion Square West	53.339764	-6.251988
3620257	36	2021-04-28 03:16:26	ST. STEPHEN'S GREEN EAST	40	39	1	St. Stephen's Green East	53.337826	-6.256035
10499070	57	2021-12-08 04:45:30	GRATTAN STREET	23	5	18	Grattan Street	53.339630	-6.243778
8603480	17	2021-10-04 21:37:27	GOLDEN LANE	20	17	3	Golden Lane	53.340801	-6.267732
3407187	61	2021-04-21 07:28:40	HARDWICKE PLACE	25	15	10	Hardwicke Place	53.357044	-6.263232
5805337	114	2021-07-05 22:12:37	WILTON TERRACE (PARK)	40	28	11	Wilton Terrace (Park)	53.333652	-6.248345
4986779	101	2021-06-10 18:57:43	KING STREET NORTH	30	4	26	King Street North	53.350292	-6.273507
7146870	50	2021-08-19 18:09:30	GEORGES LANE	40	31	9	George's Lane	53.350231	-6.279696
9138583	69	2021-10-21 15:55:22	GRAND CANAL DOCK	40	21	19	Grand Canal Dock	53.342636	-6.238695
6449208	13	2021-07-28 02:21:59	FITZWILLIAM SQUARE WEST	30	23	7	Fitzwilliam Square West	53.336075	-6.252825
3521372	17	2021-04-25 18:53:27	GOLDEN LANE	20	8	12	Golden Lane	53.340801	-6.267732
9087868	87	2021-10-19 15:11:37	COLLINS BARRACKS MUSEUM	38	36	2	Collins Barracks Museum	53.347477	-6.285250
6189156	8	2021-07-19 04:00:02	CUSTOM HOUSE QUAY	30	8	22	Custom House Quay	53.347885	-6.248048

3.2.4 Cleaning bike data

A major part of the data cleaning process is to remove data that will not contribute meaningfully to the analysis. A first step in finding this data is to take a look at how many unique classifications there are across the features recorded. Numbers of unique outputs within each field of the main bike-stand dataframe. The status has two categories which are Open or Closed. A simple cleaning task is to filter to open stations and the drop the status field altogether. Table **??** shows this unique data.

We need to figure out whether we have "LAST UPPDATED" duplicate data and if so to drop all but one.

We could use duplicated() and drop$_duplicates(), forfindingandremovingduplicaterows.FindingduplicaterowsCount$
$duplicaterowsExtractingduplicaterowswithlocDeterminingwhichduplicatestomarkwithkeepDroppingduplicaterows$

94,354 duplicate rows are removed.

359 rows representing a STATUS of "closed" were removed.

3.3 Split our "Last_updated" field into date and time

3.4 Feature Engineering

We now split our "Last_updated" field into date and time using the following code.

```
#Splitting "Last_updated"
df['DATETIME'] = [dt.datetime.strptime(d, "%Y-%m-%d %H:%M:%S") for d in df["Last_updated"]]
df['Last_updated'] = [dt.datetime.time(d) for d in df['DATETIME']]
df['DATE'] = [dt.datetime.date(d) for d in df['DATETIME']]
df['date_for_merge'] = df['DATETIME'].dt.round('H')
#Bike Feature Engineering
df['Occupy_Pct'] = df['Bikes'] / df['Total']
df['Saturated'] = np.where(df['Occupy_Pct'] == 0, 1,0 )
df['Empty'] = np.where(df['Occupy_Pct'] == 1, 1,0 )
```

3.5 Bike Feature Engineering

To understand and anticipate the future demands on the Dublin Bikes Bike Sharing System (BSS) it is necessary to carry out feature engineering techniques on the given raw data in order to express the spatial-temporal dependencies in this historical data that can be fed into machine learning algorithms and make meaningful, purposeful predictions and forecast (X. Li et al. 2021).

Chapter 4

Preparng the weather Data

4.1 Data Dictionary

Table 4.1 gives a Data Dictionary for the weather data.

Table 4.1: Data Dictionary for the Weather Data

date	ind	rain	ind.1	temp	ind.2	wetb	dew
Date		Precipitation Amount (mm)		Air Temperature (C)		Wet Bulb Temperature (C)	Dew

Chapter 5

Results of Cluster analysis

```
    #merge clusters back into main dataset

merged_with_clusters = merged_data
cluster_output = locations[['STATION ID', 'Cluster']]
cluster_output.drop_duplicates(keep = 'first', inplace = True)
del merged_data
merged_with_clusters = pd.merge (merged_with_clusters, cluster_output, on = 'STATION ID', how = 'lef
merged_with_clusters['BIKE_ARR_DEP_ABS'] = abs(merged_with_clusters['BIKE_ARR_DEP'])
merged_with_clusters.sample(5)
```

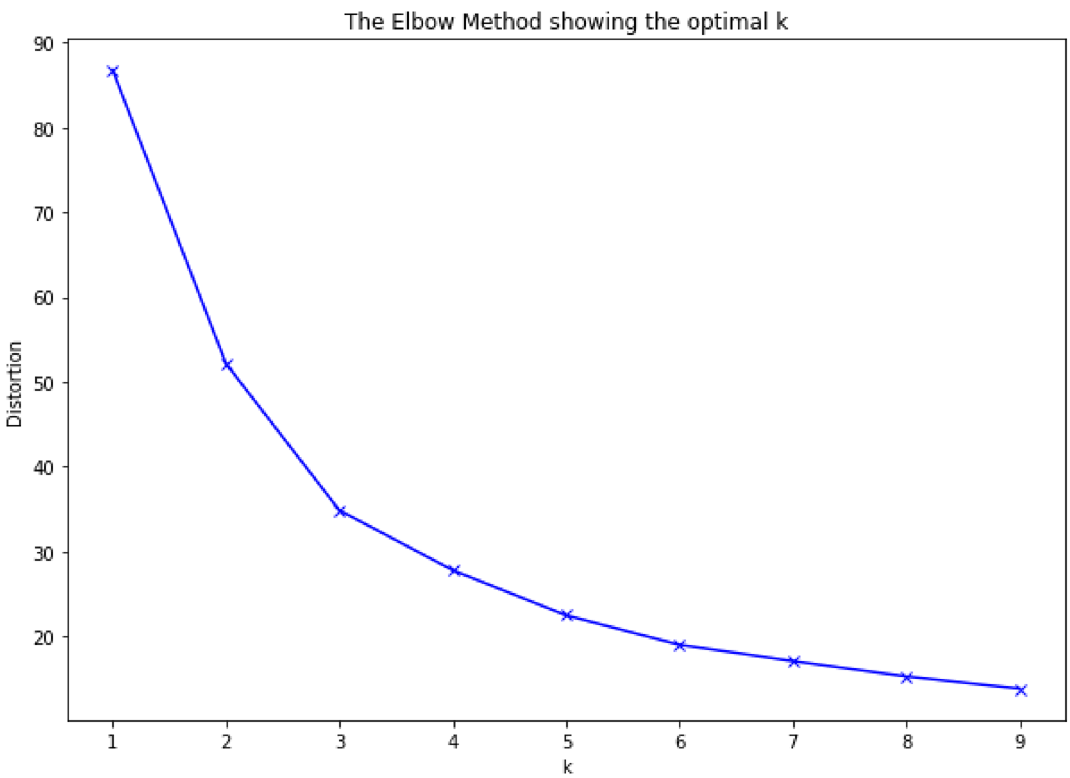

Figure 5.1: Elbow K-means clustrering showing optimal k

Table 5.1 shows the outcome of the clustering algorithmn.

Table 5.1: Clustering group output

CLUSTER_GROUP	NAME	STATION ID	LATITUDE	LONGITUDE	11AM-3PM Saturday	11AM-3PM Sunday	11AM-3PM Weekday	4PM-7PM Saturday	4PM-7PM Sunday	4PM-7PM W
59	KEVIN STREET	71	53.337757	-6.267699	0.304642	0.410472	0.125568	0.341096	0.552481	0.119685
83	PARNELL SQUARE NORTH	30	53.353462	-6.265305	0.173251	0.182215	0.140112	0.226194	0.189375	0.085246
94	SIR PATRICK DUN'S	58	53.339218	-6.240642	0.233121	0.360565	0.780507	0.171545	0.353433	0.183742
76	NEW CENTRAL BANK	66	53.347122	-6.234749	0.121657	0.256576	0.495556	0.174265	0.297717	0.092155
49	HERBERT PLACE	19	53.334431	-6.245575	0.489407	0.514594	0.710475	0.369441	0.475592	0.117246
101	TALBOT STREET	38	53.350975	-6.252940	0.758586	0.793227	0.191219	0.774761	0.608752	0.530055
40	GRANTHAM STREET	18	53.334122	-6.265436	0.345837	0.437829	0.170726	0.525248	0.580548	0.220809
42	GREEK STREET	4	53.346874	-6.272976	0.456860	0.634967	0.161102	0.613695	0.678916	0.183057
37	GRANGEGORMAN LOWER (CENTRAL)	104	53.355171	-6.278424	0.140772	0.081789	0.088553	0.105932	0.067460	0.073878
91	ROTHE ABBEY	85	53.338776	-6.303950	0.222477	0.091853	0.082668	0.235726	0.212328	0.453113
36	GRAND CANAL DOCK	69	53.342636	-6.238695	0.301610	0.382088	0.789709	0.369934	0.489441	0.190457
102	THE POINT	67	53.346867	-6.230852	0.249197	0.233178	0.577421	0.369254	0.384362	0.385355
22	ECCLES STREET	12	53.359245	-6.269779	0.093542	0.090550	0.086365	0.121464	0.129592	0.261289
60	KILLARNEY STREET	115	53.354843	-6.247579	0.273933	0.226020	0.193540	0.493421	0.317822	0.722040
25	EXCHEQUER STREET	9	53.343033	-6.263578	0.838669	0.900400	0.669069	0.748462	0.645783	0.466817
41	GRATTAN STREET	57	53.339630	-6.243778	0.163207	0.207926	0.853889	0.191706	0.288481	0.270259
29	FITZWILLIAM SQUARE WEST	13	53.336075	-6.252825	0.140430	0.221265	0.787322	0.168051	0.232829	0.191592
0	AVONDALE ROAD	108	53.359406	-6.276142	0.154090	0.044511	0.076187	0.164268	0.084199	0.235650
34	GEORGES QUAY	16	53.347507	-6.252192	0.800128	0.866134	0.370548	0.779801	0.703460	0.645451
89	PRINCES STREET / O'CONNELL STREET	33	53.349014	-6.260311	0.899466	0.907180	0.587238	0.759993	0.714499	0.811844
77	NEWMAN HOUSE	53	53.337132	-6.260590	0.189749	0.397296	0.753578	0.217007	0.452431	0.216181
100	ST. STEPHEN'S GREEN SOUTH	37	53.337494	-6.261990	0.328673	0.644534	0.521212	0.436374	0.824583	0.169878
19	DENMARK STREET GREAT	59	53.355610	-6.261397	0.132893	0.054448	0.102178	0.142795	0.099194	0.104090
26	EXCISE WALK	48	53.347778	-6.244239	0.405927	0.430998	0.593624	0.362329	0.507299	0.587542
44	HANOVER QUAY	68	53.344116	-6.237153	0.179985	0.214744	0.772331	0.199796	0.298597	0.131710

5.1 Folium map of clusters

```
    colordict = {0: 'blue', 1: 'red', 2: 'orange', 3: 'green', 4: 'purple'}
dublin_map = folium.Map([53.345, -6.2650], zoom_start=13.5)
for LATITUDE, LONGITUDE, Cluster in zip(locations['LATITUDE'],locations['LONGITUDE'], locations['Clu
    folium.CircleMarker(
        [LATITUDE, LONGITUDE],
        color = 'b',
        radius = 8,
        fill_color=colordict[Cluster],
        fill=True,
        fill_opacity=0.9
        ).add_to(dublin_map)
dublin_map
```

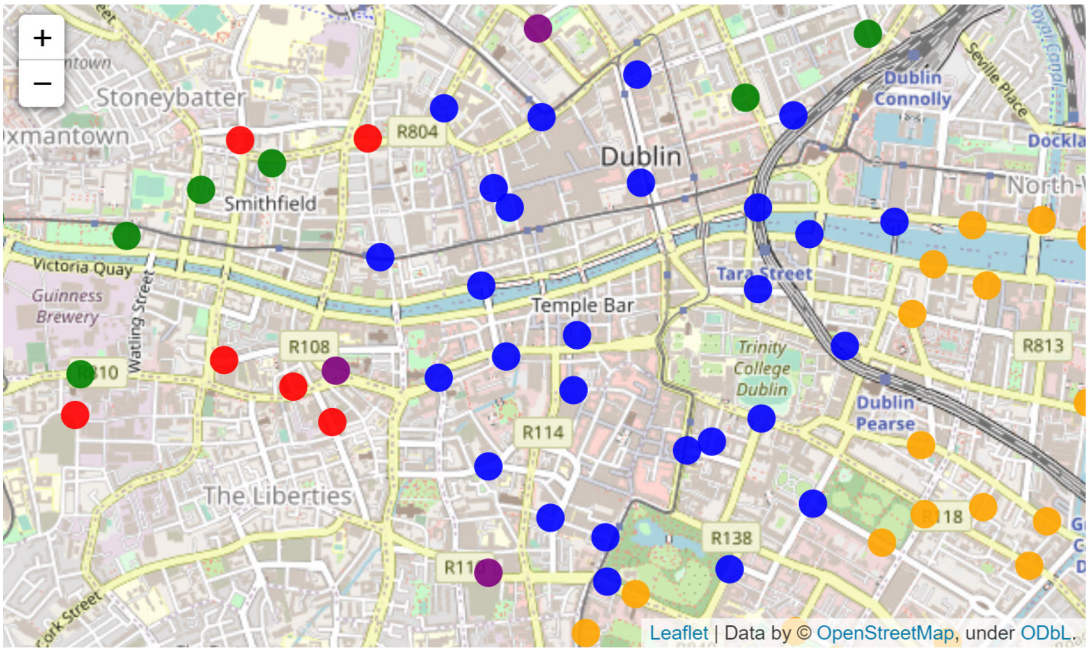

Figure 5.2: The scikit-learn K-Means library in Python for the clustering and plotted the output on a map using the Folium library based on type of day field and time slot in day

5.2 Daily Occupancy Patterns

The clustering map shown in figure 5.2 shows a clear geographical patterns in usage. Clusters of Stations are geographically linked. For instance the "Blue Cluster" is the center of Dublin City and the commercial and recreational hub. The "Purple Cluster" can be associated with transport routes etc.

5.3 Informing the future network re-balancing schedule

Looking at figure 5.6 valuable insights are apparent as to when Dublin bikes should be focusing their rebalancing efforts to meet network trends across the day, week and cluster map. On weekdays the orange office zone should be partially emptied during the working day to avoid FULL stations.

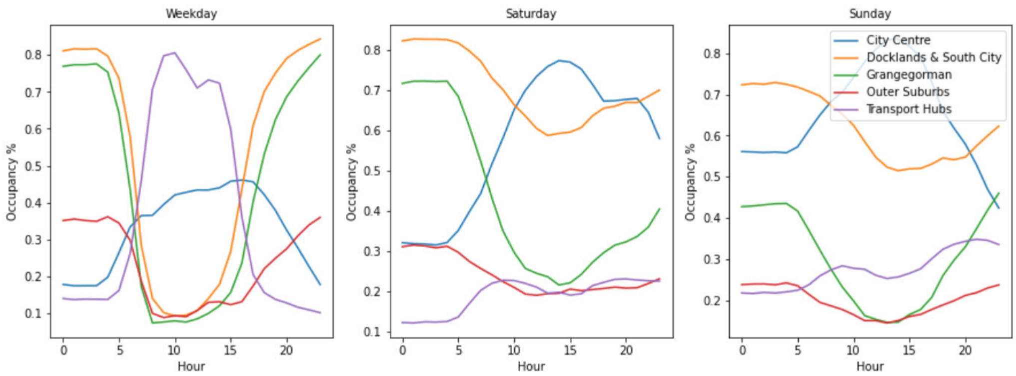

Figure 5.3: Average Station Occupancy per cluster per hour, showing different weekday vs weekend trends

5.4 Influence of stations as a predictor variable

```
    # Impact of Stations
join_table= merged_with_clusters.groupby(['STATION ID','NAME', 'DATE']).agg(rain=('rain', 'sum'), TO
join_table =join_table.reset_index()
join_table['WET/DRY DAY'] = np.where(join_table['rain'] > 3, "Wet", "Dry")
join_table = join_table.drop(['rain'], axis = 1)
join_table =join_table.reset_index()
merged_with_clusters_wetdry = pd.merge(merged_with_clusters, join_table, on = ['STATION ID', 'NAME',

wetday_df= merged_with_clusters_wetdry.groupby(['STATION ID', 'NAME', 'WET/DRY DAY']).agg(AVG_CHANGE
wetday_df =wetday_df.reset_index()
difference_df = wetday_df.pivot(index=['NAME'], columns='WET/DRY DAY', values='AVG_CHANGES').reset_i
difference_df['Change'] = difference_df['Dry'] - difference_df['Wet']
difference_df.sort_values(by = 'Change', ascending=False).head(30)
```

Table 5.2 shows the stations that are most impacted by weather conditions in descending order. This can inform rebalancing in the future.

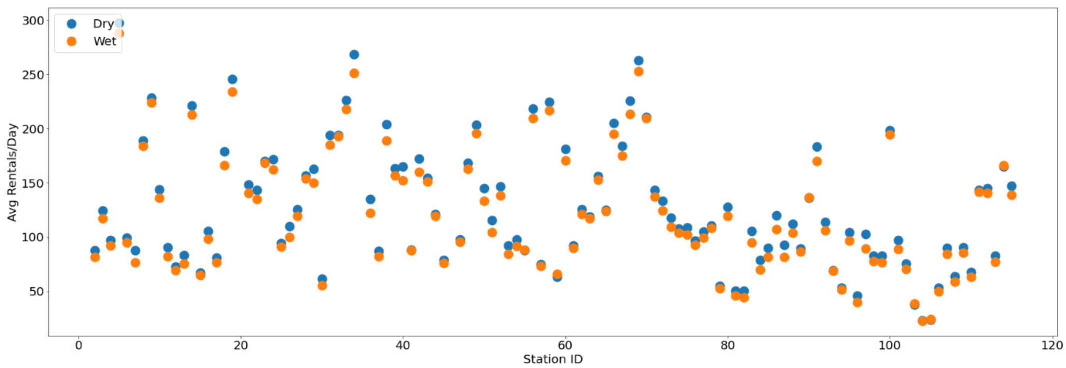

Figure 5.4: This plot shows rentals based on dry (blue) and wet (orange days). Clearly the users of the service are not much influenzed by the weather. No harm considering the irish climate!

Table 5.2: Impact of Stations and weather conditions

WET/DRY DAY	NAME	Dry	Wet	Change
87	PORTOBELLO HARBOUR	268.667020	251.299090	17.367930
101	TALBOT STREET	204.112653	189.169696	14.942957
61	KILMAINHAM GAOL	102.674665	89.138224	13.536442
96	SOUTH DOCK ROAD	183.418196	170.069310	13.348886
99	ST. STEPHEN'S GREEN EAST	135.003279	122.134989	12.868291
40	GRANTHAM STREET	178.837201	165.983982	12.853219
57	JERVIS STREET	165.056634	152.249448	12.807186
81	ORMOND QUAY UPPER	162.692758	150.021584	12.671175
82	PARKGATE STREET	119.818111	107.381632	12.436479
95	SMITHFIELD NORTH	172.240466	159.806675	12.433792
44	HANOVER QUAY	225.873719	213.616577	12.257142
33	GEORGES LANE	144.933547	133.174342	11.759205
49	HERBERT PLACE	245.476967	233.916103	11.560864
110	YORK STREET WEST	115.656987	104.333149	11.323838
14	COLLINS BARRACKS MUSEUM	92.865893	81.577540	11.288353
55	HIGH STREET	87.783011	76.553509	11.229502
24	EMMET ROAD	105.405440	94.909504	10.495936
78	NORTH CIRCULAR ROAD	181.258036	170.873490	10.384546
70	MERRION SQUARE WEST	109.853431	99.939640	9.913791
8	CATHAL BRUGHA STREET	171.860846	162.028754	9.832092
36	GRAND CANAL DOCK	262.869244	253.110929	9.758314
76	NEW CENTRAL BANK	204.893706	195.210987	9.682718
9	CHARLEMONT PLACE	297.536124	288.175189	9.360936
84	PARNELL STREET	194.238089	184.920834	9.317255
73	MOUNT STREET LOWER	218.673234	209.530249	9.142984
102	THE POINT	184.105100	174.993564	9.111537
6	BROOKFIELD ROAD	78.990582	70.014854	8.975728
58	JOHN STREET WEST	133.448011	124.555044	8.892967
21	EARLSFORT TERRACE	90.764412	81.884657	8.879755
3	BLACKHALL PLACE	112.342396	103.589903	8.752493

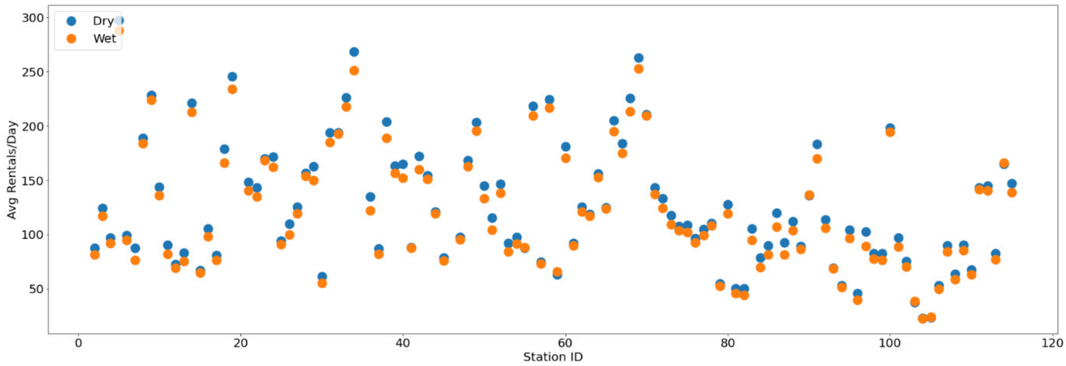

Figure 5.5: This plot shows rentals based on dry (blue) and wet (orange days). Clearly the users of the service are not much influenzed by the weather. No harm considering the irish climate!

Table 5.3: This model can predict whether a station would be too full or empty to a certainty of 79%.

Accuracy: 0.7879801828927795					
precision	recall	f1-score	support		
0	0.79	0.75	0.77	133371	
1	0.79	0.82	0.81	234832	
2	0.78	0.75	0.77	79491	
accuracy	0.79	447694			
macro avg	0.79	0.77	0.78	447694	
weighted avg	0.79	0.79	0.79	447694	

5.5 Usage patterns based on day type

5.6 Machine learning quantification of accuracy of full and empty calculations

Some independent variables are more influential than others and the final algorithm quantifies and maps this.

Plot of most influencial model features

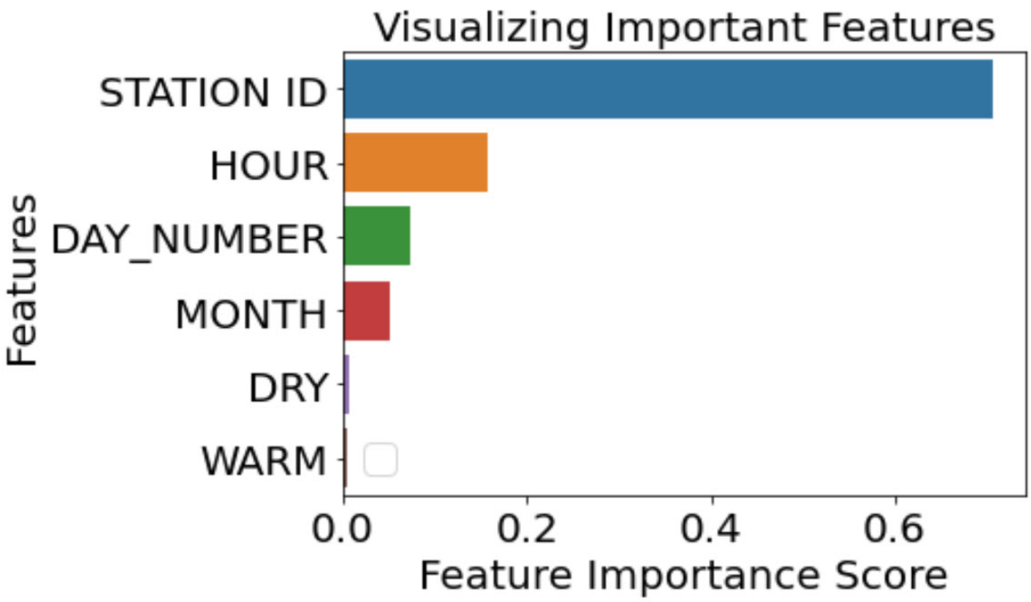

Figure 5.6: Plot of most influencial model features

Chapter 6

Conclusion

This simple Dublin-Bikes dataset when merged with related historical weather data has uncovered otherwise hidden trends in usage on the network. Dublin's population will grow rapidly in the coming years and Dublin City Council (DCC) need to use data driven technologies such as machine learning to plan and develop capacity for societies needs including shared-bike transport. Rebalancing to avoid empty and full stations is a cheap and effective way of improving the service and several daily, weeklu and spatial patterns have been developed here that can inform such future efforts.

Bibliography

Buehler, Ralph and John Pucher (2022). “Cycling through the covid-19 pandemic to a more sustainable transport future: evidence from case studies of 14 large bicycle-friendly cities in Europe and North America”. In: *Sustainability* 14.12, p. 7293.

Carroll, Páraic and Peter O’Sullivan (2020). “Forecasting the impact of the Planning, Land Use and Transport Outlook (PLUTO) Project–a 2040 Ireland case study”. In: *Transportation Research Procedia* 49, pp. 70–82.

Chen, Yan et al. (2022). “The impact of the COVID-19 pandemic on the behaviour of bike sharing users”. In: *Sustainable Cities and Society* 84, p. 104003.

Colfer, Barry (2020). “Herd-immunity across intangible borders: Public policy responses to COVID-19 in Ireland and the UK”. In: *European Policy Analysis* 6.2, pp. 203–225.

Commins, Nicola and Anne Nolan (2008). “The Determinants of Mode of Transport to Work in the Greater Dublin Area. ESRI WP268. December 2008”. In.

DeMaio, Paul (2009). “Bike-sharing: History, impacts, models of provision, and future”. In: *Journal of public transportation* 12.4, p. 3.

DublinBikes (2022). Dublinbikes.ie. URL: `https://www.dublinbikes.ie/en/offers/groups` (visited on 12/03/2022).

Li, Xinyu et al. (2021). “Short-term forecast of bicycle usage in bike sharing systems: a spatial-temporal memory network”. In: *IEEE Transactions on Intelligent Transportation Systems.*

Shaheen, Susan A, Stacey Guzman, and Hua Zhang (2010). “Bikesharing in Europe, the Americas, and Asia: past, present, and future”. In: *Transportation research record* 2143.1, pp. 159–167.

Thoa Pham Thi, Thanh et al. (2017). “Bike Renting Data Analysis: The Case of Dublin City”. In: *arXiv e-prints*, arXiv–1704.

Zhou, Xiaolu, Mingshu Wang, and Dongying Li (2019). “Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning”. In: *Journal of transport geography* 79, p. 102479.