CCT College Dublin

Assessment Cover Page

Module Title:	Statistics for Data Analytics Programming for Data Analytics Data Preparation and Visualisation Machine Learning for Data Analytics
Assessment Title:	CA1 50% Integrated Assessment
Lecturer Name:	John O'Sullivan, Sam Weiss, David McQuaid, Muhammad Iqbal
Student Full Name:	Zhongjie Fei
Student Number:	2022173
Assessment Due Date:	11/11/2022
Date of Submission:	11/11/2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

College Dublin
Computing • IT • Business

Author: Zhongjie Fei
Student ID: 2022173

Public Bike Parking in Dublin City

MSc in Data Analytics

e-mail: 2022173@student.cct.ie

Abstract

The cyclist volumes have increased significantly in Dublin over the past decades. This addresses the importance of providing ample parking space for the cyclists. Dublin City Council recently released a dataset that includes information of the public bike park stands across Dublin. This data was studied in order to analyse the characteristics and distribution of the parking stands. It shows that the majority type of the stands is Sheffield Stand and the high density of parking space locates in the central and south side of Dublin. Most locations own 5 to 15 numbers of parking stands. After analysing, two machine learning models were applied to make predications.
Keywords: bike, bike parking, parking stands, data analysis, geographic information system, machine learning

0. Introduction

In modern days, people use various ways to travel, such as planes, trains, buses, etc. Cycling nowadays is becoming more and more popular due to various reasons. Therefore, providing enough parking space is vital for the cities. In Dublin, there are different types of park styles. However, in public, Sheffield Stand earns its most popularity. In this study, we downloaded a dataset that was released by Dublin City Council to analyse the distribution and variety of public bike parking stands across Dublin. At the end, we used two machine learning models to predict the area based on the other independent features in our dataset. And the KNN model seems to suit better than the SVM model.

1. Bike parking

Since the invention of bicycle, often called bike or cycle, in the 19th century, it has transformed into one of the fundamental ways people commute. According to Dublin City Council, ‘cycling numbers in Dublin have more than doubled in the last decade’ (Dublin City Council, 2021). The reasons, to list only a few, why people choose to cycle are because not only does cycling save sufficient money, but cycling also brings modern-day convenience considering the traffic situation nowadays. Cultivated by current culture, people also increasingly cycle to keep themselves fit and protect the environment. These therefore address the great importance of providing ample and secured bike parking facilities for cyclists.

Recently in Dublin, although 'the implementation of Covid-19 mobility-related interventions led to an expansion of higher quality cycling facilities, the majority of the current cycle parking comprises of unsheltered Sheffield Stand cycle racks', demonstrating the lack of diversity (Egan et al., 2022: 1933). This tendency is also exhibited in the data we collected for our report, as well as through our observation around the city.

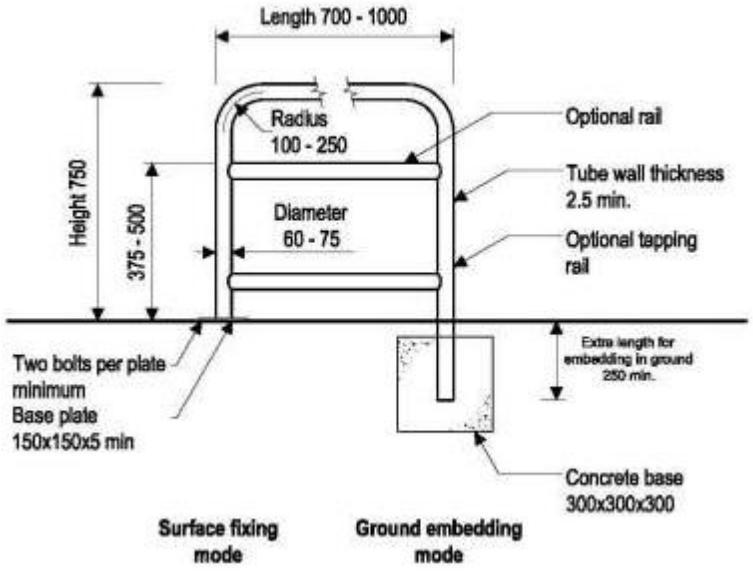

Figure 1: Sheffield Stand details

(Dún Laoghaire-Rathdown County Council Municipal Services Department, 2018: 4)

Figure 1 shows Sheffield Stand, a preferred type of bike parking stand, which is 'usually made of a single metal tube bent to form a stand which will support a bicycle and permit locking of both the frame and front and rear wheels to the stand' (Dún Laoghaire-Rathdown County Council Municipal Services Department, 2018: 4). Unlike stacked cycle parking which is difficult to use or low cycle racks (e.g. "Hoops") which could cause possible damage to the wheels, Sheffield Stand provides more flexibility and protection (ibid.).

2. Data collection

The main purpose of this report was to examine and visualise the distribution of the quantity of different bike parking stands across Dublin City via the data we collected from Ireland's Open Data Portal website (*https://data.gov.ie/*). Furthermore, we applied two machine learning models with the processed data to compare and suggest a relatively more appropriate approach for prediction.

Initially published in 2015 by Dublin City Council, the data was then updated in September 2022. It is worth noting that the listing has not been modified considering some stands having been subsequently removed after installation due to construction or other specific reasons (Dublin City Council, 2022a).

Jupyter notebook version 6.4.8 and Python version 3.9.12 were used for exploratory data analysis (EDA), visualisation and machine learning in our report.

3. Data preparation and EDA

Although the heart of data science lies in visualisation and machine learning, the time that an analyst spend on data preparation takes up to or even more than 80% of the whole project (McKinney, 2018). Data processing is a 'critical piece' in a data science job which involves several steps performing in a 'iterative manner' (Squire, 2015: 3). In other words, the task can be repetitive in order to generate promising result, and continuous revisiting of data preparation is key. Typically, data preparation includes 'loading, cleaning, transforming, rearranging' (McKinney, 2018: 203).

Exploratory data analysis (EDA), on the other hand, is an essential step to gain an in-depth understanding of the datasets and create useful insights through data manipulation and visualisation (Peng *et al.*, 2021). For instance, 'discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures' (Mukhiya and Ahmed, 2020: 8). It often starts after data processing and cleaning.

3.1 Overview of original dataset

Table 1 depicts details of the features in our original data after being read in Jupyter notebook.

Table 1: Details of features in original data

FN	Attribute name	Description
1	type_stands	types of stands in Dublin (7 types)
2	X	longitude
3	Y	latitude
4	Easting	easting
5	Northing	northing
6	location_stand	address where stands locate
7	no_stands	numbers of stands in the location

There are 937 rows and 7 columns in our dataset, which means there are 7 features in total with 937 observations in each one of them. In these 7 features, three are categorical data and four are numerical data. There are no duplicated rows in our dataset, however missing values were found in some features.

3.1.1 Coordinate reference system

Due to the importance of coordinate system in our data, a good knowledge of it is critical. Holdgraf and Wasser (2018) also echo the importance of understanding the coordinate system used in the data, especially if the data is from different resources and stored in different coordinate systems. This can consequently cause difficulty and confusion in analysis.

A coordinate reference system (CRS) is 'a methodology to define a location of a feature in space' (Janssen, 2009: 43). There are two different types of coordinate reference systems, namely geographic coordinate systems (e.g. longitude/latitude) and projected coordinate systems (e.g. easting/northing, UTM, Robinson). In addition, curvilinear coordinates are more useful for locating points on Earth, and grid coordinates are more ideal for measuring distance between points.

Importantly, since coordinate reference systems are 'idealised abstractions', coordinates datums (or reference frames), designated as 'physical materialisation (or realisation)', help to effectively define the origin and orientation of the coordinate reference system at a certain epoch (ibid.). Janssen (2009) illustrates a number of popular datums for geographic coordinate systems (e.g. WGS84, NAD27 & NAD83, ED50 & ETRS89) and projected coordinate systems (e.g. LCC projection and UTM projection).

3.2 Data cleaning and rearranging

Based on checking the unique values of four features that are coordinate inputs, along with research and testing using EPSG.io (*https://epsg.io*) published by Klokan Technologies, Switzerland, three types of coordinate systems were identified in the data. Firstly, a geographic coordinate system realises by **EPSG: 4326** (i.e. WGS 84) datum in the form of longitude and latitude (e.g. -6.22726053, 53.34779261). Secondly, a projected coordinate system realises by **EPSG: 29902** (i.e. TM65/Irish Grid) datum in the form of easting and northing (e.g. 316615.57, 234148.652). Thirdly, a projected coordinate system realises by **EPSG: 2157** (i.e. IRENET95/Irish Transverse Mercator)

datum in the form of itm x and itm y (e.g. 714060.7656, 733161.9352), in accordance with Irish Grid Reference Finder (*https://irish.gridreferencefinder.com/*). It is worth mentioning that there is an "identical datum" applicable for the second coordinate system, however, the one described above was chosen for this study.

Following deeper analysis of the unique values in feature X and Y, four issues were found: i) Table 2 indicates that the data type of feature Y is object whereas the entries are apparently continuous numerical data; ii) the values in both features are a wrong mix of geographic coordinates (i.e. longitude and latitude) and projected coordinates (i.e. itm x and itm y); iii) besides the mixture of two different coordinate systems in the two columns, longitude and latitude values are also inconsistently entered, that is, values in feature X which are expected to be longitudes turns into latitudes in lower half and vice versa; iv) there are 249 values missing in total from 937 rows in both features, which can significantly affect our analysis since coordinate information is exceptionally unique.

Table 2: Data type and missing value sum

FN	Attribute name	Dtype	Isnull
1	type_stands	object	0
2	X	float64	249
3	Y	object	249
4	Easting	float64	0
5	Northing	float64	0
6	location_stand	object	47
7	no_stands	float64	49

Comparatively, feature Easting and Northing combining as the other coordinate system are more reliable and have no missing value. Thus, feature X and Y were dropped to prevent problems. A conversion method using pyproj package was then applied to transform Easting and Northing into their corresponding geographic coordinates (i.e. longitude and latitude) due to their advantage for locating on map and availability for visualising in python. Two new columns were then added to the dataset with the transformed longitude and latitude values.

In regards to locations of the stands, they can be viewed as reiteration of the coordinates, and are expected to be distinct values. Kanani (2020) has demonstrated that features including only distinct values have no use for providing insights and making

predictions. We agree that values of the feature location_stand can be treated simply as identification of the addresses which are of less use. We however argue that the coordinate numbers themselves are though distinctive, there is rich information underlying throughout manipulation and visualisation. For example, combining the coordinates to form spatial data (i.e. points, lines, polygon) can provide us lots of insights. Moreover, there are 47 null values and 79 duplicated values in stand_location, showing that the list was not precisely built by its owner. Since the full list of unique coordinate values can help to identify every location, this column was then dropped.

Interestingly, after dropping the columns after adequate analysis, one duplicated row was found. With thorough observation, two identical entries were found, and the reason why they were not spotted before is because the stand locations of these two rows were entered differently. This therefore addresses the importance of adequate data preparation. This one duplicated row was then dropped.

Thereafter, we renamed the columns and changed their order to form a new dataset for further processing.

3.2.1 Missing value

Missing values are ‘an ubiquitous problem’ in data analysis, however, ‘all standard statistical techniques for analysing the data require complete cases without any missing observation’ (Faisal, 2018: 1, 9). Hence, an inadequate handling approach may draw out biased results and misleading inference.

Little and Rubin (2002, cited in Nakai *et al.*, 2014: 28-29) developed three mechanisms of missing data:

i). Missing at random (MAR) represents that the probability an observation is missing depends on the observed part of Y (Y_{obs}), but not on the missing part of Y (Y_{mis}). The missingness is conditionally independent on current and future responses.

ii). Missing completely at random (MCAR) is the special case of MAR, and the probability that an observation is MCAR depend neither on Y_{obs} or $Y_{mis.}$ It has stronger assumptions than MAR.

iii). In contrast to MAR and MCAR which can be referred to as an ignorable mechanism, not missing at random (NMAR) is referred as a non-ignorable mechanism. The probability that an observation is missing not at random depends on both Y_{obs} and Y_{mis}.

Faisal (2018: 12-13) illustrates three traditional methods for handling missing data. First, deletion methods simply disregard any unknown attribute values in a dataset. When the size of a dataset is large enough, analysis can still be executed "correctly" after deletion as if there is no missingness in the raw data. However, ignoring missing values even in this case can potentially cause loss of information and reduction of statistical power, which may conclude inadequate results. Second, substitution methods replace missing values with a global constant or simple estimates. They can be 0 or null value, as well as mean or median values from the original data. This method can cause biased representation of the data and affect its quality (Li, 2009). Third, imputation is to fill the missing values with plausible values using different algorithms. It can be divided into single imputation and multiple imputation.

With the development of research, a number of more sophisticated approaches have been proposed to deal with missing values in data analysis. For example, Bayesian approaches and surrogate split methods for categorical missing data; regression methods for numeric missing data (ibid.).

Nearest neighbour imputation method, which was chosen to deal with the missing values in our data, observes the neighbourhood of the missing data based on some distance measure from available data to impute missing values (Faisal, 2018). It is simple and easy to implement and fairly outruns other methods. The missing values in our data are the number of stands that locate in certain points, which means by measuring the neighbouring locations of these missing points, reliable numbers can be obtained to fill these missing values. KNNImputer function from scikit-learn was hence imported to handle the missing values in our data.

Additionally, there was one entry with number 0. In reality, zero number of parking stand is impossible to happen. Hence, the value 0 was treated as missing value and replaced by a nan in our case, although sometimes zero can be important in a dataset (e.g. sparse data).

After the imputation, a final dataset was saved and ready for analysis (Figure2).

	stand_type	easting	northing	long	lat	stand_numbers
0	Sheffield Stand	318106.414	234502.185	-6.227256	53.347793	4
1	Sheffield Stand	315620.753	232932.916	-6.265134	53.334248	1
2	Sheffield Stand	316615.570	234148.652	-6.249761	53.344949	2
3	Sheffield Stand	316520.200	234543.281	-6.251047	53.348514	1
4	Sheffield Stand	317119.980	233940.217	-6.242268	53.342965	3

Figure 2: Head of the final dataset after preparation

Also when practicing visualisation, we strongly intended to show the detailed location information (i.e. addresses) of each parking stands on the interactive map. Therefore, a geocoders method from geopy package was applied using Nominatim to transform geographic data into actual addresses. A new column stand_location was then added to the final dataset and saved using a different name for convenience since the conversion code takes long time to run considering there are over 900 rows to be transformed (Figure 3).

	stand_type	easting	northing	long	lat	stand_numbers	geom	stand_location
0	Sheffield Stand	318106.414	234502.185	-6.227256	53.347793	4	53.34779326636702, -6.227255782056404	North Docklands, North Wall, North Dock B ED, ...
1	Sheffield Stand	315620.753	232932.916	-6.265134	53.334248	1	53.334248439303025, -6.265133791780888	39, Camden Street Lower, Saint Kevin's ED, Dub...
2	Sheffield Stand	316615.570	234148.652	-6.249761	53.344949	2	53.344948684334575, -6.2497608135882246	HSE Primary Care Centre, Mark's Lane, Mansion ...
3	Sheffield Stand	316520.200	234543.281	-6.251047	53.348514	1	53.34851376066773, -6.251046948776474	IFSC House, Custom House Quay, North Dock C ED...
4	Sheffield Stand	317119.980	233940.217	-6.242268	53.342965	3	53.342965466086866, -6.242268049500275	Saint Andrew's Resource Centre, 114-116, Pears...
5	Sheffield Stand	315786.687	234487.924	-6.262076	53.348178	2	53.348177511686664, -6.262076180033214	97, Middle Abbey Street, North City ED, Dublin...
6	Sheffield Stand	316463.739	234911.200	-6.251759	53.351830	3	53.35183038824942, -6.251759096280343	Liberty Saints R.F.C, Foley Street, Mountjoy A...
7	Sheffield Stand	315358.411	235444.966	-6.268155	53.356866	3	53.35686606014675, -6.268155451509455	29, Blessington Street, Inns Quay B ED, Dublin...

Figure 3：Head of the final dataset with added location information

3.2.2 Geospatial data

In geographic information system (GIS), there are many different formats to store geospatial data, such as shapefiles and geojson files. There are two primary forms of geospatial data: vector data and raster data. Vector data which matches our study contains features (i.e. points, lines, polygons or multipolygons) to present houses, roads and areas etc. Raster data is pixelated or gridded cells which are identified according to row and column. It is more complicated.

To access these types of data in python, geopandas library is mandatory, which is very similar to the famous pandas library. The most fundamental element that differentiates geodataframes from pandas dataframes is the geometry column.

Dublin City Council divides Dublin into five local administrative areas (i.e. *NORTH WEST, NORTH CENTRAL, CENTRAL, SOUTH CENTRAL and SOUTH EAST*) in order to coordinate the delivery of services into local communities (Dublin City Council, 2022b). Apart from analysing the data in aspect of the whole city, we also initiated splitting the data into the five administrative areas, which concludes in providing more useful insights and a plausible dependent variable for machine learning.

In order to achieve this, the shapefile of the five administrative areas were downloaded (*https://data.smartdublin.ie/dataset/administrative-areas-dcc*; Figure 5). Our finalised data was also transformed into a shapefile using geopandas. We then join these two shapefiles and formed a final geodataframe for analysing. Finally, this geodataframe was saved into a geojson file for future reference, and Figure 5 shows its basic features.

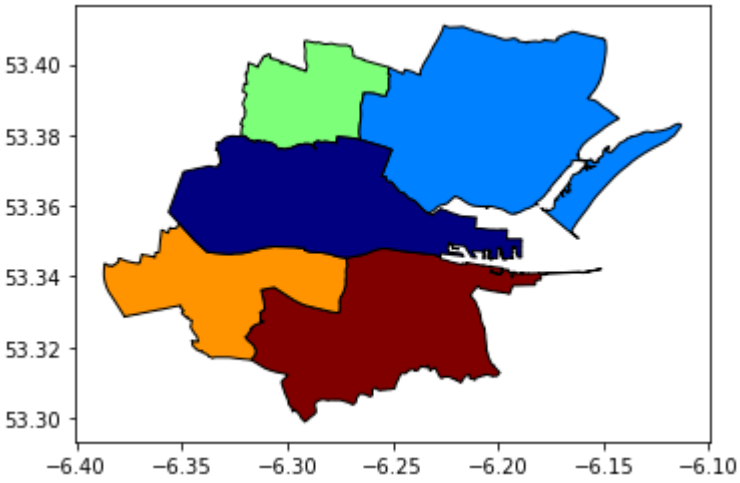

Figure 4: Five committee administrative areas of Dublin

	stand_type	long	lat	stand_numbers	geometry	area
0	Sheffield Stand	-6.227256	53.347793	4	POINT (-6.22726 53.34779)	CENTRAL
3	Sheffield Stand	-6.251047	53.348514	1	POINT (-6.25105 53.34851)	CENTRAL
5	Sheffield Stand	-6.262076	53.348178	2	POINT (-6.26208 53.34818)	CENTRAL
6	Sheffield Stand	-6.251759	53.351830	3	POINT (-6.25176 53.35183)	CENTRAL
7	Sheffield Stand	-6.268155	53.356866	3	POINT (-6.26816 53.35687)	CENTRAL

Figure 5: Head of the finalised geodataframe

Overall, three csv datasets were saved. One for main analysis and the other two for interactive visualisation. One vector data was created for visualisation and helping machine learning.

3.4 EDA

Since there is geographic information in our data, we separate our visualisation into two main parts: general and interactive. Statistical analysis is integrated in between.

3.4.1 General visualisation

In this section, we provided some general visualisation and statistical analysis that have been performed.

Peng *et al.* (2021) created a powerful package named DataPrep.EDA (*https://github.com/sfu-db/dataprep*) that maps common EDA tasks. It also generates some interactive charts and statistical information that provided us useful insights and guides to implement visualisation by matplotlib and seaborn (see attached).

The correlations between the numeric features were checked first. Figure 6 demonstrates that the correlations are extremely poor, and the best score received was 0.046.

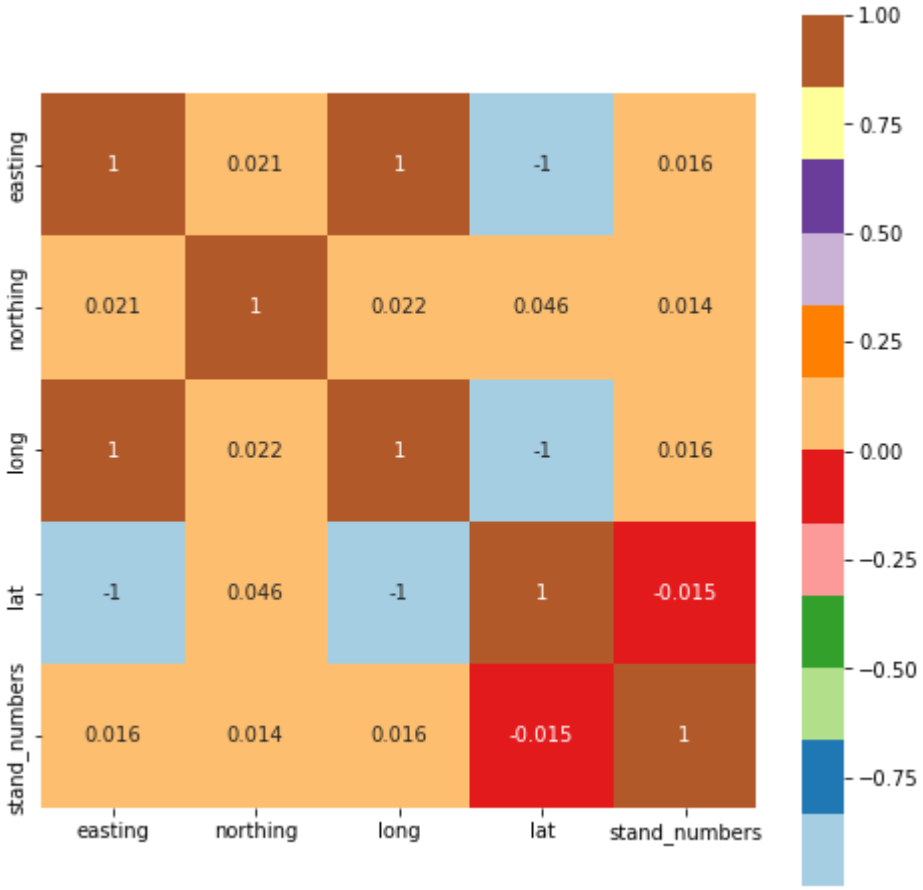

Figure 6: Correlation of the numeric features

Next, Figure 7 shows the counts of each stand type in feature stand_type. It is obvious that the distribution of stand types is extremely imbalanced. A pie chart created by DataPrepEDA reveals that Sheffield Stand consumes more than 90% of all the numbers among the seven types. They both prove the lack of diversity in bike parking stand types across Dublin. This also motivated us to bring in another element (Dublin areas) for learning.

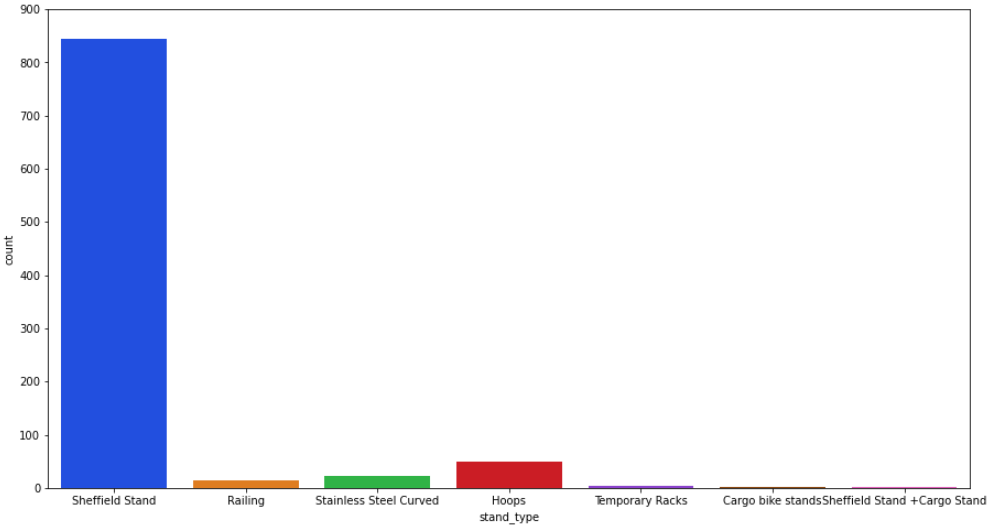

Figure 7: Counts of each stand type

In Figure 8, it shows the appearance of every unique values in stand_numbers. The distribution falls mainly between 1 and 15. Five numbers of parking stands own the highest value being over 250 (see attached).

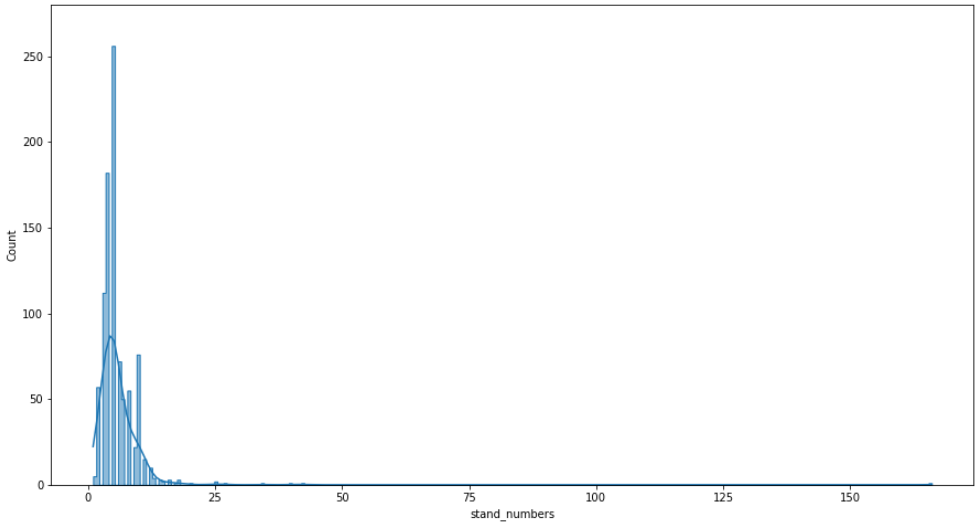

Figure 8: Counts of stand_numbers

The boxplot in Figure 9 demonstrates the Interquartile Range (IQR) of the numbers of each stand type. Notably, there are significant “outliers” in Sheffield Stand, and the reason is because there are some hot spots where the council had assembled more stands. For example, the one that has more than 160 counts is located in Heuston Station which is the largest train station in Ireland for people to travel across the

country, therefore significant amount of parking stands are required (see also in the interactive visualisation section).

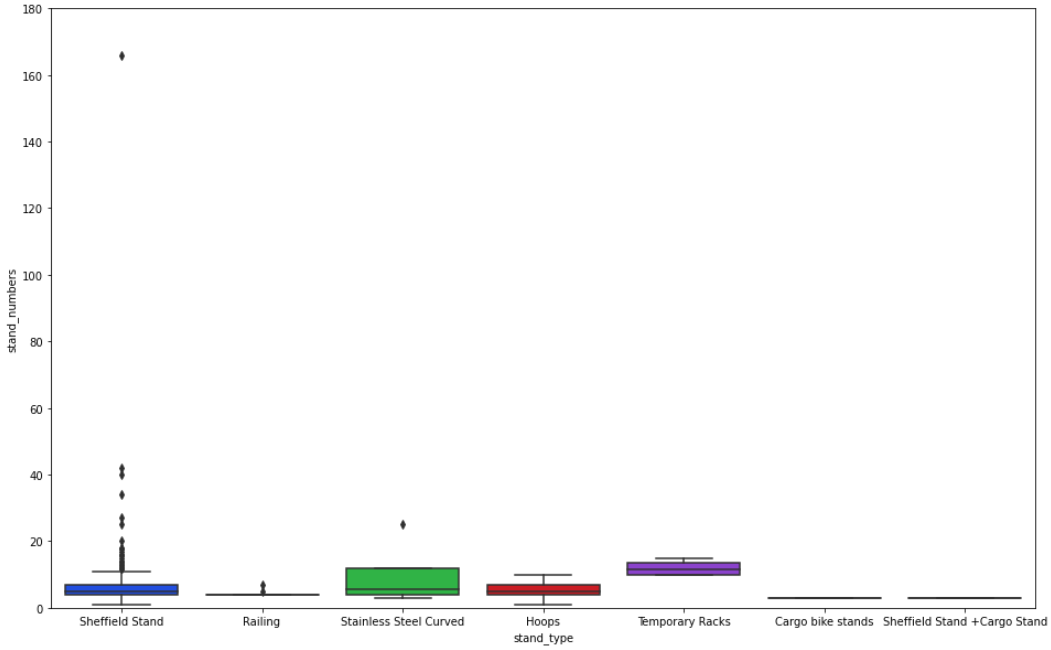

Figure 9: Interquartile range of the stand types

Lastly, a pie chart was created based on the stand counts in the five areas of Dublin (Figure 10). It demonstrates that the southeast and central parts of Dublin own the most numbers of parking stands.

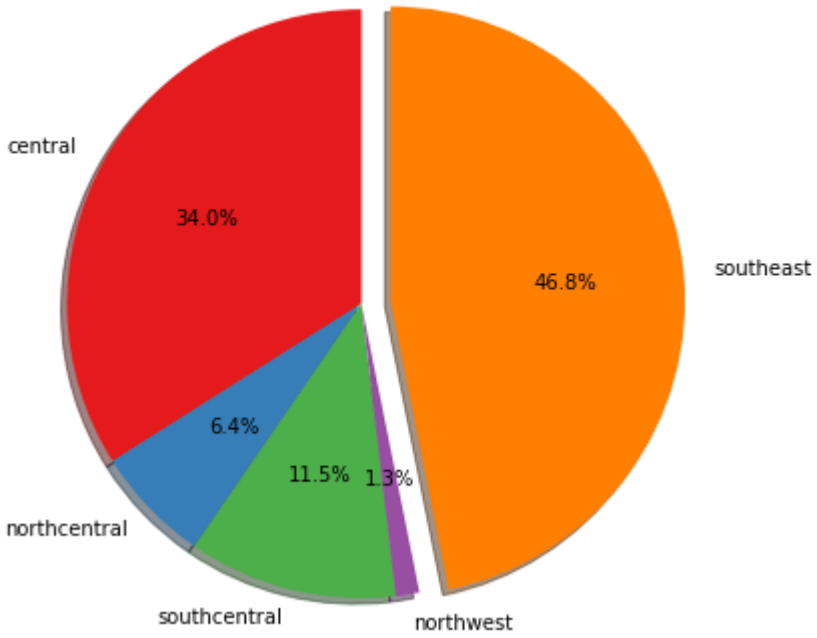

Figure 10: Percentage of the stand numbers in each area

To summarise, from the general visualisation, we understood that Sheffield Stand possesses the biggest proportion of stand types comparing to the others. Stand numbers mostly range from 5 to 15, and the council seems to focus more on developing the parking facilities in central and south Dublin.

3.4.2 Statistical analysis

The importance of statistics in data science and data analytics cannot be underestimated. Statistics provides tools and methods to find structure and to give deeper data insights (Aslanyan, 2021). There are plenty of topics in statistics, from random variables to probability distribution functions, from Bayes Theorem to inferential statics etc. Different statistics may apply to different data according to its characteristic.

The most commonly known are mean, median and standard deviation etc. Python's describe function provides overview of these statistics (Figure 11). Moreover, another popular statistic is the probability distribution function. It describes all the possible values, the sample space, and the corresponding probabilities that a random variable can take within a given range, bounded between the minimum and maximum possible values. This function has two categories: discrete (e.g. Binomial and Poisson Distribution) and continuous (e.g. Normal and Continuous Distribution).

	easting	northing	long	lat	stand_numbers
count	9.360000e+02	936.000000	936.000000	936.000000	936.000000
mean	3.188461e+05	234259.146073	-6.222243	53.338613	5.929487
std	9.291506e+04	1604.292475	1.210941	0.229736	6.294157
min	3.084301e+05	229372.000000	-6.373332	46.331202	1.000000
25%	3.151364e+05	233365.512000	-6.271998	53.338002	4.000000
50%	3.157096e+05	234124.392500	-6.263309	53.344813	5.000000
75%	3.164988e+05	234831.379000	-6.251631	53.351273	7.000000
max	3.157997e+06	240413.000000	30.777691	53.400065	166.000000

Figure 11: Common statistics in our data

Normal Distribution and Continuous Distribution were firstly calculated. Figure 12 depicts the Normal Distribution in our data. The distribution is not standard and has a skewness of about 18. The probability density is squeezing towards the left. It shows

that the chance of having small numbers in our dataset is much higher than large numbers.

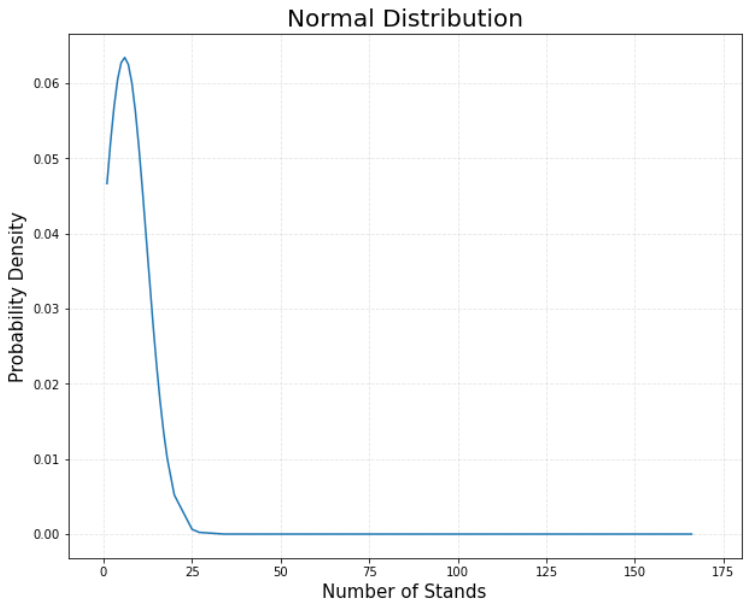

Figure 12: Normal Distribution

In Figure 13, Continuous Normal Distribution in our data explicitly demonstrates similar trend comparing to Normal Distribution. The smaller numbers build up a major part of the probability density, that is, the probability sharply cumulates to close to one before numbers 25.

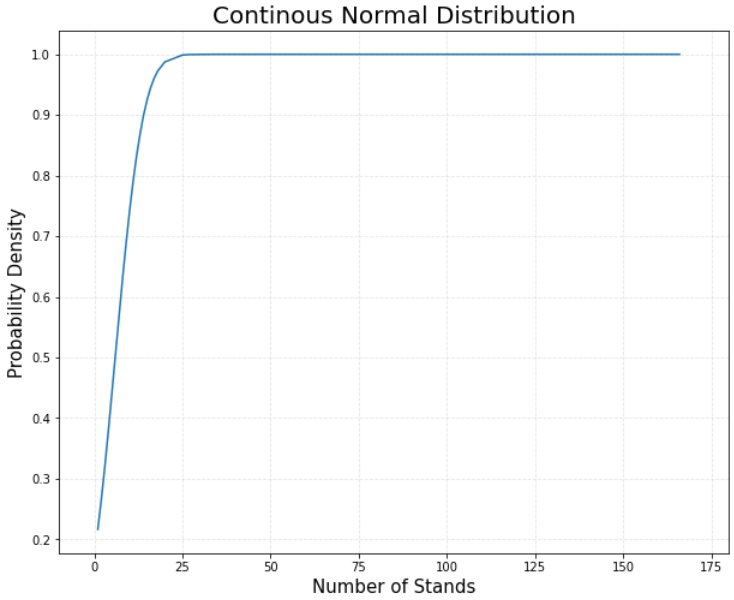

Figure 13: Continuous Normal Distribution

We also checked Poisson Distribution in our data using different functions, however, it does not show this type of distribution (Figure 14) since Poisson Distribution expresses the probability of a given number of events occurring in a fixed interval of time or space.

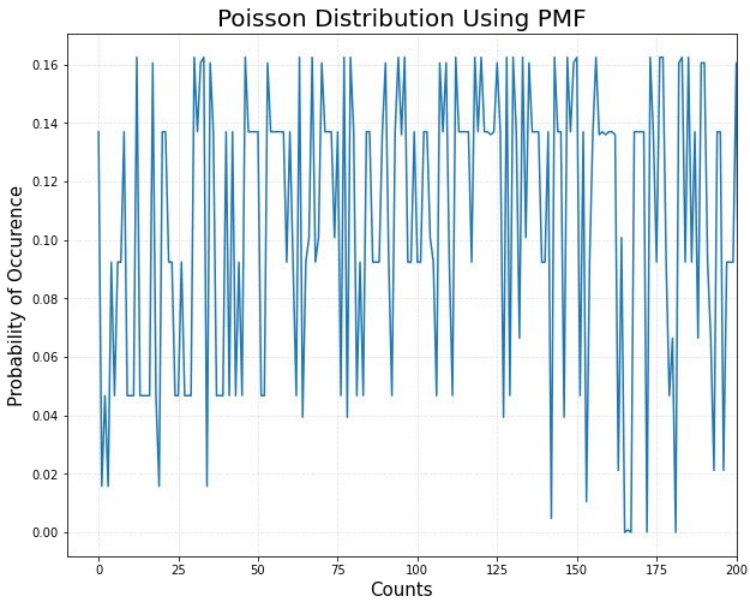

Figure 14: Poisson Distribution

3.4.2 Interactive visualisation

In this section, a series of interactive visualisation had been performed, i.e. mapping. To initiate this, the powerful folium package was installed in python. Folium is a package that was developed by Leaflet that allows users to integrate interactive maps for visualising geospatial data. In our study, the geometry, showing as points, help to mark every parking stand onto the map (Figure 15). Extra arguments, such as popup, were applied to display more detailed information of each stand, including stand type, stand numbers and locations.

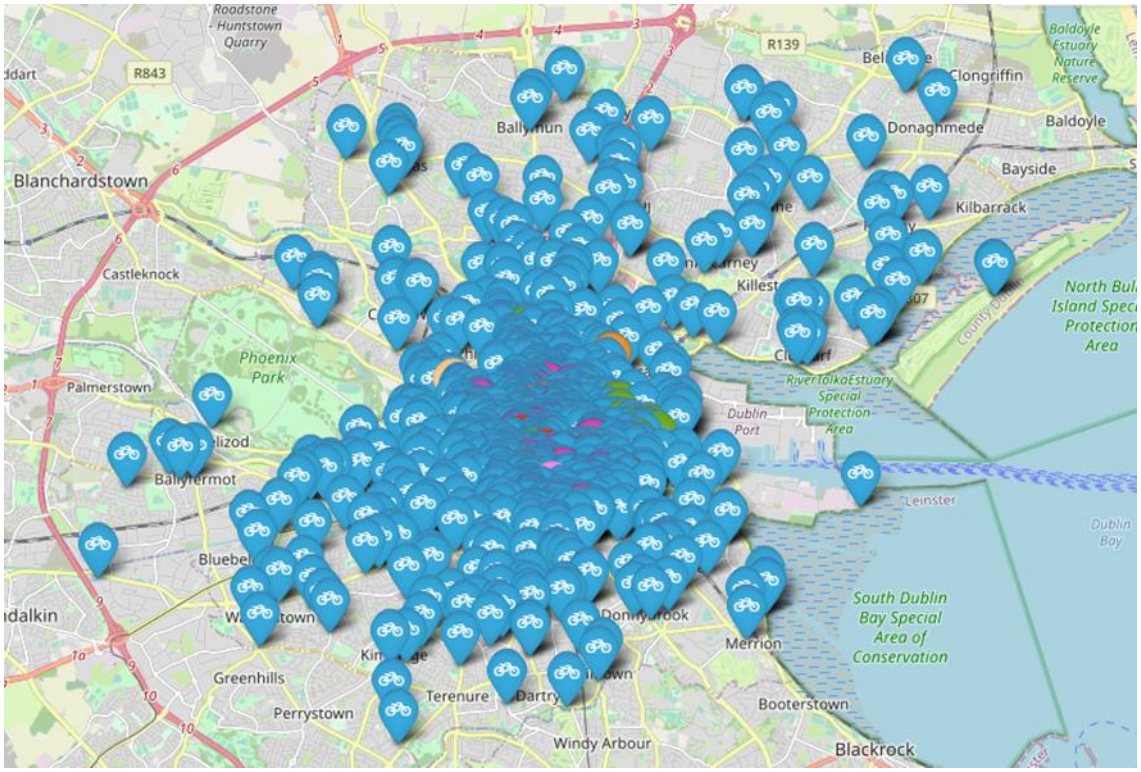

Figure 15: Map with markers

A circle map function was implemented after. The advantage of this visualisation is that it is not only direct but also informative. Human beings are the most sensitive to colours and then followed by sizes. The first perception received from this map is the red circle that has the biggest radius, which is our farthest outlier. Then the relatively smaller and green circles indicate that these spots are popular for bike parking in Dublin (Figure 16). Besides, a heat map was also created in our notebook which shares similar information as the circle map (see attached).

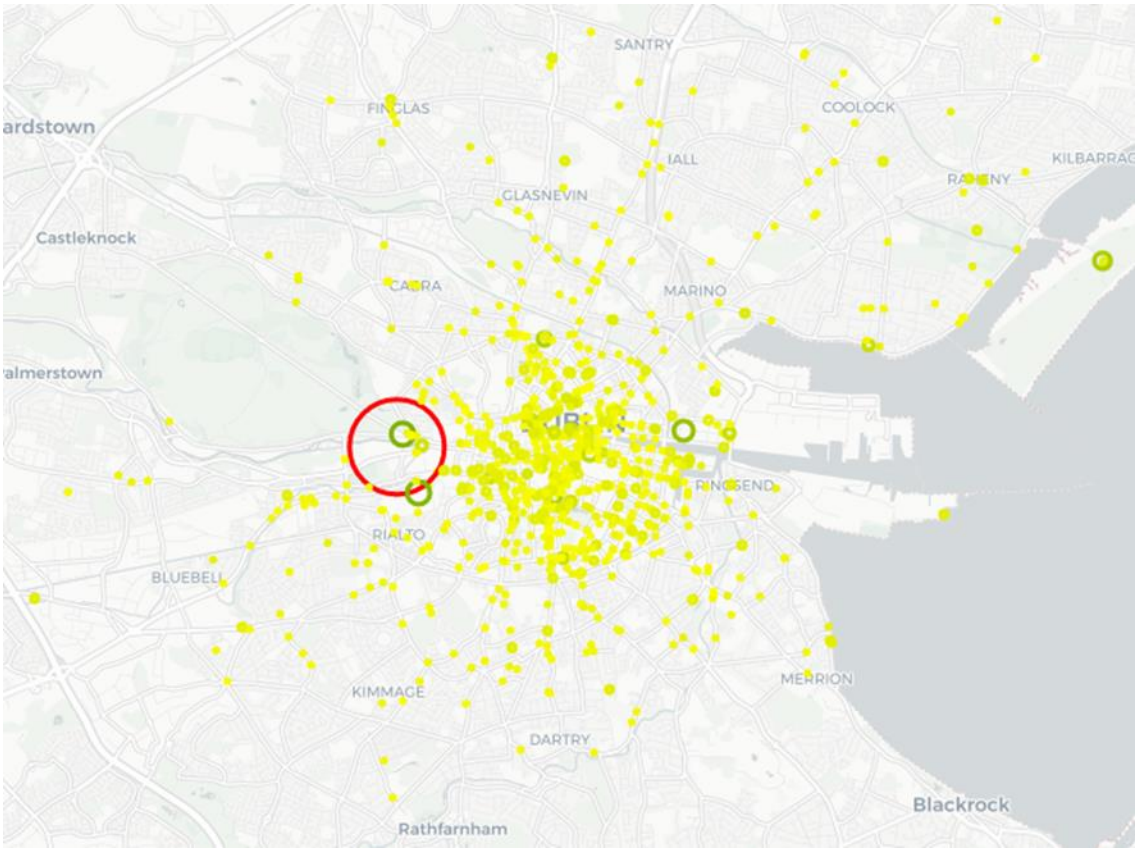

Figure 16: Circle map

Lastly, a choropleth map was created to show the density of bike parking stand in the five areas of Dublin (Figure 17). In order to get the density in each area, we calculated the areas of each part in km^2, and then we checked the total numbers of parking stands in each part. By dividing the numbers with the areas calculated using geopandas, we get the density which ranges from 1 to 16. Visually, this shows us the density of parking stands is higher on the south side of Dublin.

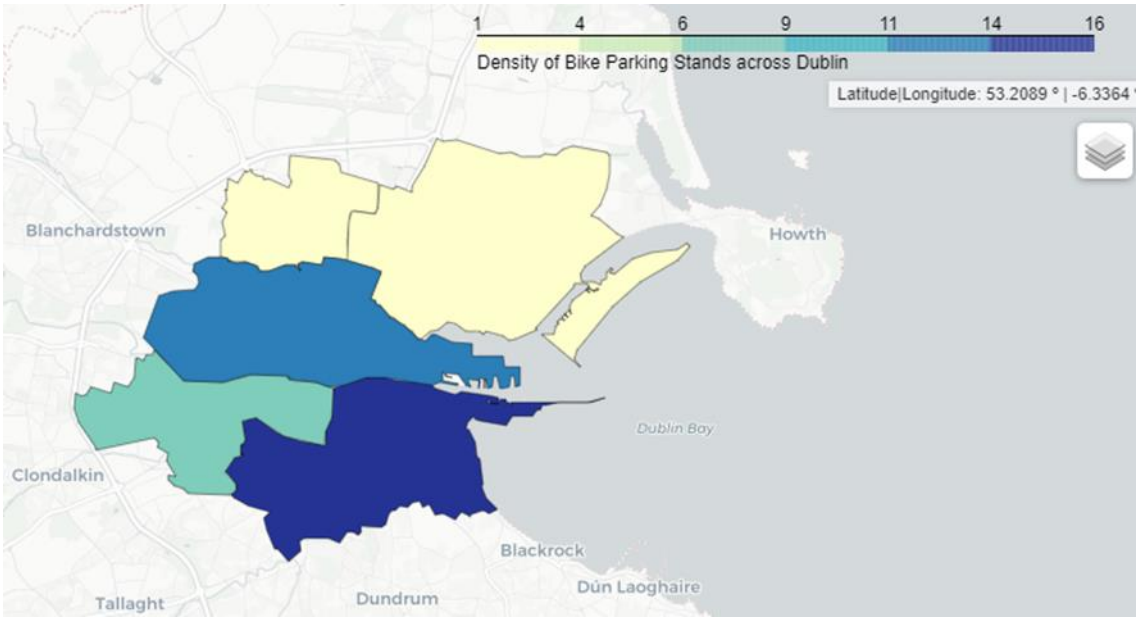

Figure 17: Choropleth map

4. Machine Learning

Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment (El Naqa and Murphy, 2015). It has been applied successfully in various industries, such as finance, computational biology and medical applications which has helped a lot of people in real life.

There are three types of machine learning: supervised learning, unsupervised learning and reinforcement learning (Sebastian Raschka and Vahid Mirjalili, 2019: 2). Supervised learning uses labelled data, and the users receive direct feedback. It can predict outcome and future. Unsupervised learning uses no label, and no feedback is expected. However, it can find hidden structure in data. Reinforcement learning uses reward system to learn a series of actions.

In our study, we applied two supervised learning models to train and test our data and make predictions. It is worth mentioning that since the original categorical data stand_type is practically distinctive, there is no point in learning on this feature. So we brought in the five areas of Dublin as our dependent learning feature. Before starting machine learning, we encoded feature stand_type using installed package.

4.1 KNN model

KNN is one of the oldest and laziest classification algorithms or statistical learning techniques for supervised learning (Ali *et al.*, 2021). It works through observation of the

k-nearest neighbours of a given data and analyse which neighbourhood it is closest to and then assume that the testing data has the same characteristics with that group of neighbours. This suits our data features since in close or same area, the parking numbers are most likely to be similar.

Scaling numeric data is also important for machine learning accuracy. The scaling helps to normalise the numeric data and keep them in a reasonable range, especially when there are outliers and they are also important in the data. In our study, we compared the outcomes of machine learning before and after standardscaling (see attached). Before scaling, the accuracy we got was 0.74. After we applied standardscale method, the accuracy increased to 0.94. The explanation is that in our analysis, we saw numbers of outliers in the dataset. Without scaling, the biases they generated were much higher, concluding in lower accuracy.

4.2 SVM model

Support vector machine (SVM) algorithm finds a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. These decision boundaries help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. It is plausible in our dataset considering the dimensional distribution of different park stands in different area.

In our study, we mainly tested the model based on standardscaled data using Gaussian Radial Basis Function (rdf) kernel. The accuracy before scaling was 0.44, and after it increased to 0.92. Besides, there are two important arguments (i.e. C and gamma) that can potentially affect our model. GridSearchCV was used to test and find out the best hyperparameter. However, after we implement the best scores of these two arguments, the accuracy stayed the same.

5. Conclusion

In order to promote cycling, the assembly of abundant and safe bike parking facilities is extremely important. Through our study, we found out that in Dublin there is not only a lack of diversity in the type of bike parking stands but also imbalanced distribution across the city. The analysis may address the required attention for other areas, northside especially, as well as the openness into other type of stands.

6. Discussion

In this section, the author discusses and reflects on this study.

The author is aware of the possible biases generated by the specific handling method we chose to deal with missing value in our data. Deletion method is plausible since the percentage of the missing values is lower than 5% in the whole dataset.

The target audience for this study is aiming mainly to the Dublin City Council. If extra data is collected considering the cyclists preference. For instance, the promotion of Bike Locker (*https://www.bikelocker.ie/*) and Cyc-Lok(*https://cyc-lok.ie/*), different insights may be received consequently.

Lastly, the author is also aware the difference of handling geospatial data than normal data and the variety of tools that can be used to generate better analysis, such as GDAL and so on.

References

Ali, M.M. *et al.* (2021) 'Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison', *Computers in Biology and Medicine*, 136, pp. 1-10 Available at: https://doi.org/10.1016/j.compbiomed.2021.104672.

Aslanyan, T. (2021) *Fundamentals Of Statistics For Data Scientists and Analysts Towards Data Science*. Available at: https://towardsdatascience.com/fundamentals-of-statistics-for-data-scientists-and-data-analysts-69d93a05aae7 (Accessed: 10 November 2022).

Dublin City Council (2021) *Cycle Parking.* Dublin City Council. Available at: https://www.dublincity.ie/residential/transportation/active-travel/cycling-dublin-city/cycle-parking (Accessed: 14 October 2022).

Dublin City Council (2022a) Public Cycle Parking Stands DCC. DATA.GOV.IE. https://data.gov.ie/dataset/dcc_public_cycle_parking_stands?package_type=dataset.

Dublin City Council (2022b) Administrative Areas DCC. DATA.SMARTDUBLIN.IE. https://data.smartdublin.ie/dataset/administrative-areas-dcc.

Dún Laoghaire-Rathdown County Council Municipal Services Department (2018) *Standards for Cycle Parking and associated Cycling Facilities for New Developments*. Dún Laoghaire-Rathdown County Council Municipal Services Department. https://www.dlrcoco.ie/sites/default/ffles/atoms/ffles/dlr_cycle_parking_standards_0.pdf

El Naqa, I. and Murphy, M.J. (2015) 'What Is Machine Learning?', in I. El Naqa, R. Li, and M.J. Murphy (eds) *Machine Learning in Radiation Oncology: Theory and*

Applications. Cham: Springer International Publishing, pp. 3–11. Available at: https://doi.org/10.1007/978-3-319-18305-3_1.

Egan, R., Dowling, C.M. and Caulfield, B. (2022) 'Planning by Cycle Parking Type: A Cycle Parking Preference Typology for Cyclists', *Case Studies on Transport Policy*, 10, pp. 1930–1944. Available at: https://doi.org/10.1016/j.cstp.2022.08.007.

Faisal, S. (2018) *Nearest Neighbor Methods for the Imputation of Missing Values in Low and High-Dimensional Data*. Göttingen: Cuvillier Verlag.

Holdgraf, C. and Wasser, L. (2018) *Geographic vs projected coordinate reference systems - GIS in Python, Earth Data Science - Earth Lab*. Available at: https://www.earthdatascience.org/courses/use-data-open-source-python/intro-vector-data-python/spatial-data-vector-shapefiles/geographic-vs-projected-coordinate-reference-systems-python/ (Accessed: 26 October 2022).

Irish Grid Reference Finder (no date). Available at: https://irish.gridreferencefinder.com/ (Accessed: 27 October 2022).

Janssen, V. (2009) 'Understanding coordinate reference systems, datums and transformations', *International Journal of Geoinformatics*, 5(4), pp. 41–53.

Kanani, B. (2020) 'Pandas - How to remove DataFrame columns with only one distinct value?', *Machine Learning Tutorials*. Available at: https://studymachinelearning.com/pandas-how-to-remove-dataframe-columns-with-only-one-distinct-value/ (Accessed: 31 October 2022).

Klokan Technologies, G. (2022) *EPSG.io: Coordinate Systems Worldwide*. Available at: https://epsg.io (Accessed: 27 October 2022).

Li, X.-B. (2009) 'A Bayesian Approach for Estimating and Replacing Missing Categorical Data', *Journal of Data and Information Quality*, 1(1), Article 3, pp. 1-11. Available at: https://doi.org/10.1145/1515693.1515695.

McKinney, W. (2018) *Python for data analysis: data wrangling with pandas, NumPy, and IPython Wes McKinney*. Second edition. Beijing: O'Reilly.

Mukhiya, S.K. and Ahmed, U. (2020) *Hands-On Exploratory Data Analysis with Python : Perform EDA Techniques to Understand, Summarize, and Investigate Your Data*. Birmingham, UK: Packt Publishing.

Nakai, M., Chen, D.-G., Nishimura, K. and Miyamoto, Y. (2014) 'Comparative Study of Four Methods in Missing Value Imputations under Missing Completely at Random Mechanism', *Open Journal of Statistics*, 4(1), pp. 27–37. Available at: https://doi.org/10.4236/ojs.2014.41004.

Peng, J., Wu, W., Lockhart, B., Bian, S., Yan, J.N., Xu, L., Chi, Z., Rzeszotarski, J.M. and Wang, J. (2021) 'DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python', in *Proceedings of the 2021 International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery (SIGMOD '21), pp. 2271–2280. Available at: https://doi.org/10.1145/3448016.3457330.

Sebastian Raschka and Vahid Mirjalili (2019) *Python Machine Learning : Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2, 3rd Edition*. [S.l.]: Packt Publishing.

Squire, M. (2015) *Clean Data*. Birmingham, UK: Packt Publishing.