**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_CA1* |
| **Lecturer Name:** | *Sam Weiss*  *Taufique Ahmed/* *Marina Iantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Ben Wilson |
| **Student Number:** | sbs24004 |
| **Assessment Due Date:** | 7/4/2024 |
| **Date of Submission:** | 7/4/2024 |

**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. |

# ****Statistical and Clustering Analysis of St. Patrick's Day Footfall Counts in Dublin City Centre****

### Prepared by Ben Wilson, sbs24004, CCT

## Introduction

This project aims to uncover patterns and trends in pedestrian footfall within Dublin City Centre, focusing particularly on the dynamics surrounding St. Patrick's Day, and identifying similar behaviour of other events throughout the year. As Ireland's national holiday, St. Patrick's Day attracts thousands of both local and international tourists to the capital's streets. The project will mainly utilise footfall count data from 2023 to highlight trends, peak activity, and heavily used areas. By comparing everyday footfall with that on St. Patrick's Day, to reveal changes in footfall dynamics and tourist visitor volume. Through the use of unsupervised k-means clustering, identify signature footfall patterns associated with large-scale events, and group them with other periods of similar activity. This should serve to provide future insights valuable for large-scale event organisation, tourism boards, festival coordinators, and local businesses.

"Fáilte Ireland is pleased to welcome thousands of visitors to locations across the country this weekend for St. Patrick’s Day celebrations. Festivals and events play a key role in delivering brilliant experiences, providing a unique reason for visitors to choose a destination and increase footfall for local businesses, supporting jobs and revenue generation. Fáilte Ireland estimates that last year 570,000 people attended the Saint Patrick’s Festival Dublin over the bank holiday weekend generating €113million in revenue for Dublin."  
  
Source: [Fáilte Ireland 2024](https://www.failteireland.ie/Utility/News-Library/failte-ireland-welcomes-visitors-st-patricks-day.aspx)

## Objectives:

* To perform statistical analysis of footfall count patterns in Dublin City Centre
* To quantify the effect of St. Patrick's Day Festival on pedestrian footfall and compare counts with previous years
* To create a k-means clustering machine learning model for identifying similar periods of activity and their characteristics

## Data Source

The data used in this project was downloaded from Smart Dublin; a data set entitled [Pedestrian Footfall DCC](https://data.smartdublin.ie/dataset/dublin-city-centre-footfall-counters). According to the "Dublinked Open Data Store", the counts are recorded using "a network of PYRO-Box people counters located throughout central Dublin". The data is supplied by Dublin City Council and the NTA (National Transport Authority).

## Exploratory Data Analysis *(section 1)*

Exploring the data is the first critical step in the data analysis process. This involves understanding the dataset by summarizing its main characteristics, such as statistics, and often employing visual methods to comprehend the structure, patterns, and relationships between variables. In this project, there were several iterations of exploration and data preparation, aiming to refine and obtain a fully processed dataset. This section details some of the initial observations made.

### Counter Locations

The counter locations data contained information on counter coordinates, which would be useful for visualizing the geographic locations of counts. Observations included:

* There were columns irrelevant to the analysis, including “\_id” and “User Type.”
* A column called “Eco-Visio Oupput” contained the location names, which matched those in the footfall count data.
* The locations were in no particular order.

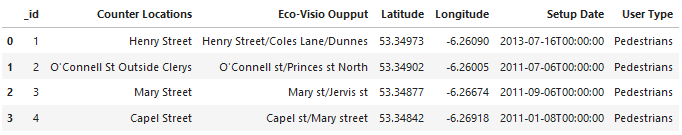


Table 1 A portion of the Counter Location data published by Dublinked

### Footfall Counts 2023

The footfall counts for the year 2023 are perhaps the most important data in this project. Observations included:

* All count values in the dataset were floats, continuous, and numerical.
* There were columns labelled “IN” and “OUT” in addition to each counter column.
* There was a time column with time as an object type.
* Approximately a quarter of the dataset had significant missing values, with spurious NaNs at many locations, and some locations had no data at all.
* There was a wide distribution of values between locations, considering parameters such as mean, standard deviation, and interquartile ranges.

## Data Cleaning and Preparation *(section 2)*

Real-world data is often in a state requiring significant preprocessing for analysis. The observations made from the exploratory analysis of the datasets in this project underscore this point. The following section will detail some of the methods used to prepare the data, and justifications for those methods.

### Counter Locations

* Columns irrelevant to the project's scope were eliminated to retain only pertinent data. This was achieved using the **.drop** method.
* Locations identified in the footfall data for removal, as a list, were also removed to ensure consistency between datasets.
* The column labelled “Eco-Visio Oupput” was renamed to “Counter Location” via the **.rename** method to enhance clarity and readability.
* Counter locations were alphabetically sorted, and the index was reset to facilitate data navigation and ensure symmetry across datasets. This process utilized the **.sort\_values** and **.reset\_index** functions.
* The application of these methods yielded a streamlined dataset, well-ordered and harmonized with the main dataset, enhancing usability for future analysis.

### Footfall Counts 2023

* Columns marked with “IN” or “OUT”, although potentially valuable for broader analyses, were deemed extraneous for the purposes of this project, due to its limited scope. A user function was devised to systematically remove these columns across all datasets, detailed in the file **UserDefinedFunctions** under **remove\_IN\_OUT()**.

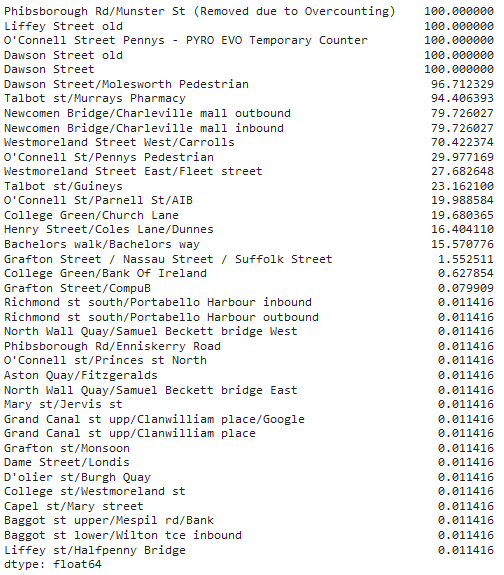


Table 2 Percentage of missing values at each location

* The 'time' column was converted into a datetime object and set as the new index. This change facilitated time series analysis.
* As highlighted in Table 2, the dataset exhibited a considerable proportion of missing data. In prioritizing data integrity—opting for quality over quantity—locations with more than 2% missing values were excluded. This approach aimed to minimize reliance on imputation techniques that could potentially distort underlying signals.

### Missing Data

Despite excluding a substantial amount of location data, there was still missing values that needed to be addressed. These gaps spanned various durations, including single hours, several hours, and whole days. Given the cyclic nature of footfall counts, each of these voids was addressed individually, depending on the nature of the absence.

### Single Period NaNs

All locations were missing data for a single hour at the same time—2 am on March 26, 2023. Using the premise that footfall would change gradually over a short period, linear interpolation was deemed suitable for imputing the missing hour. This method was applied over that period.

### Short Period NaNs

“Grafton Street/CompuB” experienced a short null period lasting 5 hours. Applying the same logic from the previous step, linear interpolation over this period was considered appropriate to fill the NaNs. Additionally, this method was deemed particularly suitable here, given the start and end times—from 4 am to 8 am, as in Figure 1. As this time of day is well outside working hours, it is reasonable to assume that footfall would generally be low and consistent.

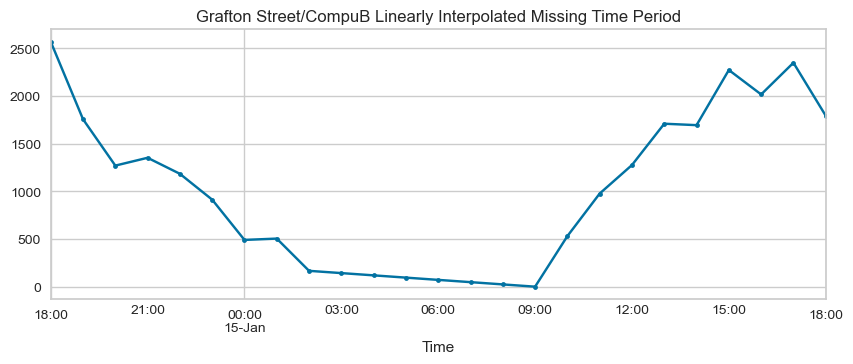
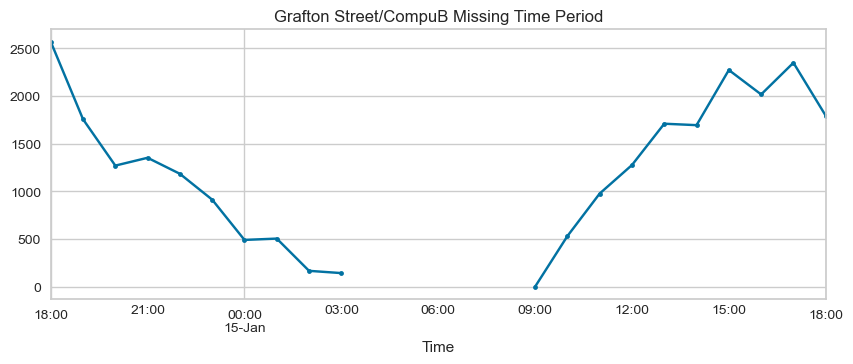


Figure 1 Line graphs showing footfall counts over "Grafton Street/CompuB" missing time period

### Long Period NaNs

There were four missing time periods, each approximately a day long, with two occurring at "College Green/Bank Of Ireland" and two at "Grafton Street / Nassau Street / Suffolk Street." The longer missing periods present a greater challenge for imputation due to the cyclic and dynamic behaviour of footfall. Linear interpolation would not be suitable over such extended periods, as it fails to accurately reflect the characteristics observed in footfall count data. “IterativeImputer” was deemed an appropriate solution, utilizing the scikit-learn library. This method leverages all features to impute missing data. As can be observed in the heat maps Figure 3, 4, 5, all features/locations exhibit a reasonable semblance of correlation. Given the strong correlations among many features in this dataset, this module was considered an adequate solution. The results, shown in Figure 2, show visually the cyclic characteristic has been captured, although a statistical method of evaluation would be preferable.

[Iterative Imputer](https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html) is an experimental module from the scikit-learn library. As an estimator, it uses a multivariate strategy for the imputation of missing data. From [6.4.3. Multivariate feature imputation](https://scikit-learn.org/stable/modules/impute.html#iterative-imputer), the class:

*"...models each feature with missing values as a function of other features, and uses that estimate for imputation. It does so in an iterated round-robin fashion: at each step, a feature column is designated as output y and the other feature columns are treated as inputs X. A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature in an iterative fashion, and then is repeated for max\_iter imputation rounds. The results of the final imputation round are returned."*

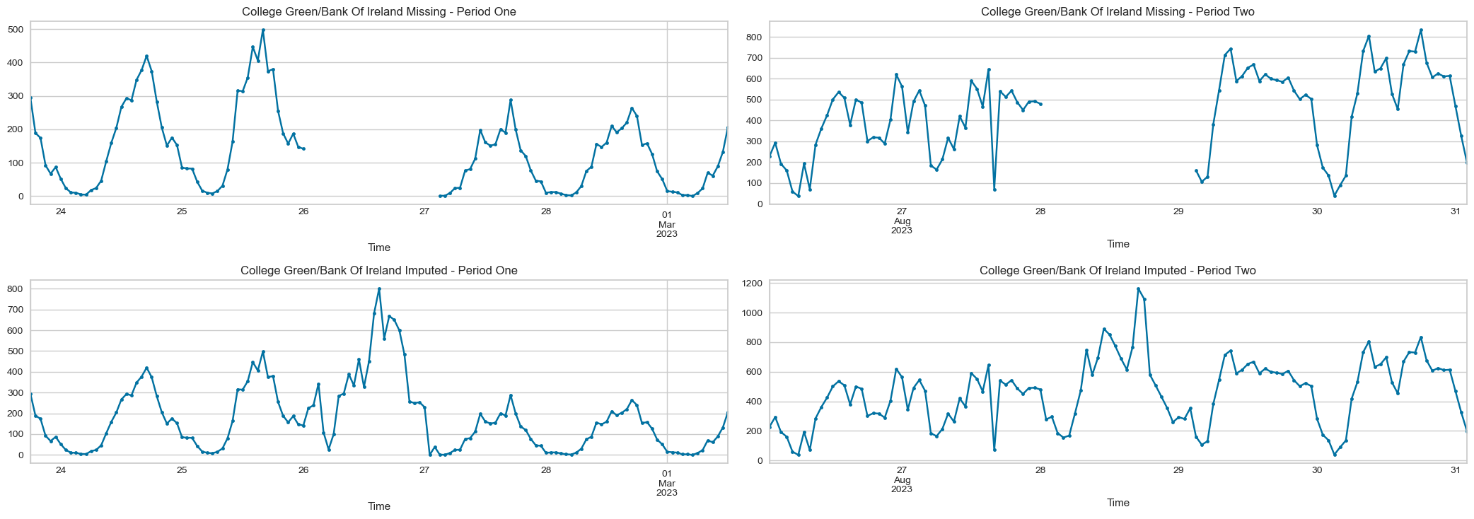
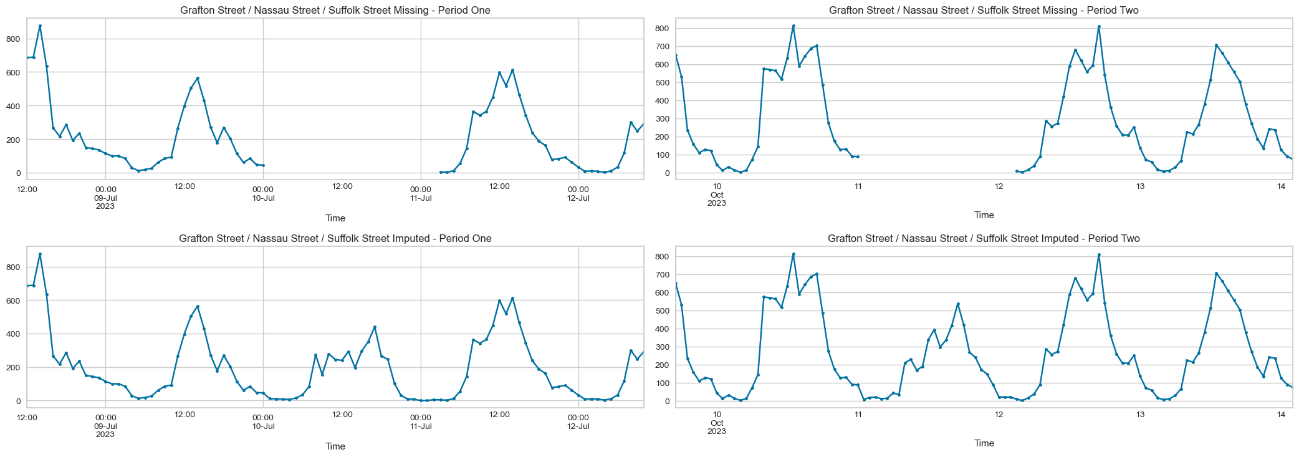


Figure 2 Line graphs displaying the results before and after processing with the Iterative Imputer for extended periods of missing data (NaNs)

### Correlation Heat Maps



Figure 3 Correlation Heat Map showing the correlation between all location counts through the year 2023

Correlation heat maps serve as a powerful tool for exploratory data analysis, providing a visual representation of the relationships between variables. In this project, they are utilized to decipher the relationship between footfall counts at various locations. As observed in Figures 3, 4, and 5, all locations exhibit a degree of positive correlation. This finding is logical and aligns with expectations, considering that footfall dynamics are generally regarded as interdependent. Such insights prove useful for validating our imputation methodology, enriching our overall analysis, and guiding the selection of focal areas for deeper examination.

A notable variation in pattern emerges when comparing the heat map of 2023 in Figure 3 with those of St. Patrick's Day in Figures 4 and 5. Specifically, 'College Green/Bank of Ireland' shows a weaker correlation with other locations throughout the year. However, on St. Patrick's Day, it demonstrates a significantly stronger correlation.

## 

Figure 4 Correlation Heat Map showing the correlation between all location counts through St. Patrick's Day 2023

Figure 5 Correlation Heat Map showing the correlation between all location counts through St. Patrick's Day 2022

## 

### Visualization of Counts with Line Plots

Line plots are among the simplest ways to visualize data, proving particularly useful for time series analysis. In our dataset, they have been instrumental in identifying trends, outliers, cyclical patterns, and the dynamics of footfall counts. Given the extensive range of features, dynamic plots created with the Plotly library have enhanced our ability to visualize these data. This functionality has facilitated the segmentation, isolation, or removal of specific patterns, and has been invaluable for focusing on particular time periods.

In Figure 6, a typical flow of footfall in Dublin city centre throughout a day is depicted. In contrast, Figure 7 reveals more sporadic variations and higher counts. Analysing specific locations by hour has yielded insightful observations, such as the clear spike in footfall on St. Patrick's Day observed in the middle of the day, depicted in Figure 8, with similar spikes visible at other times of the year.

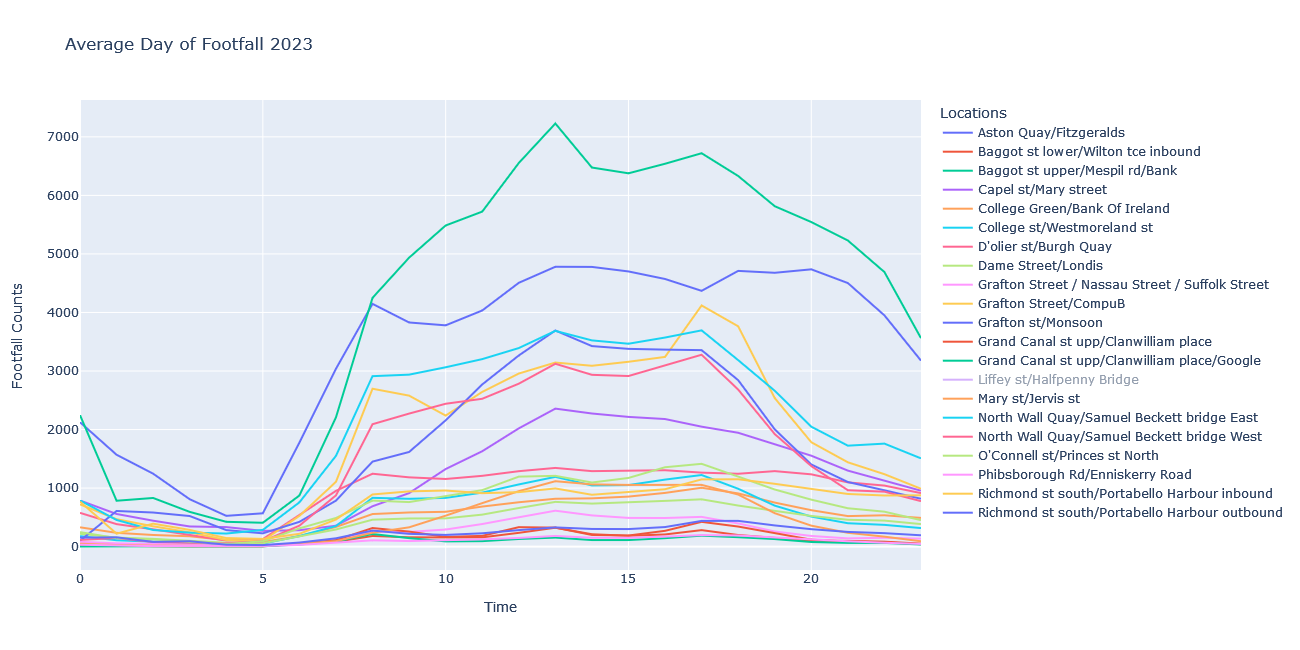
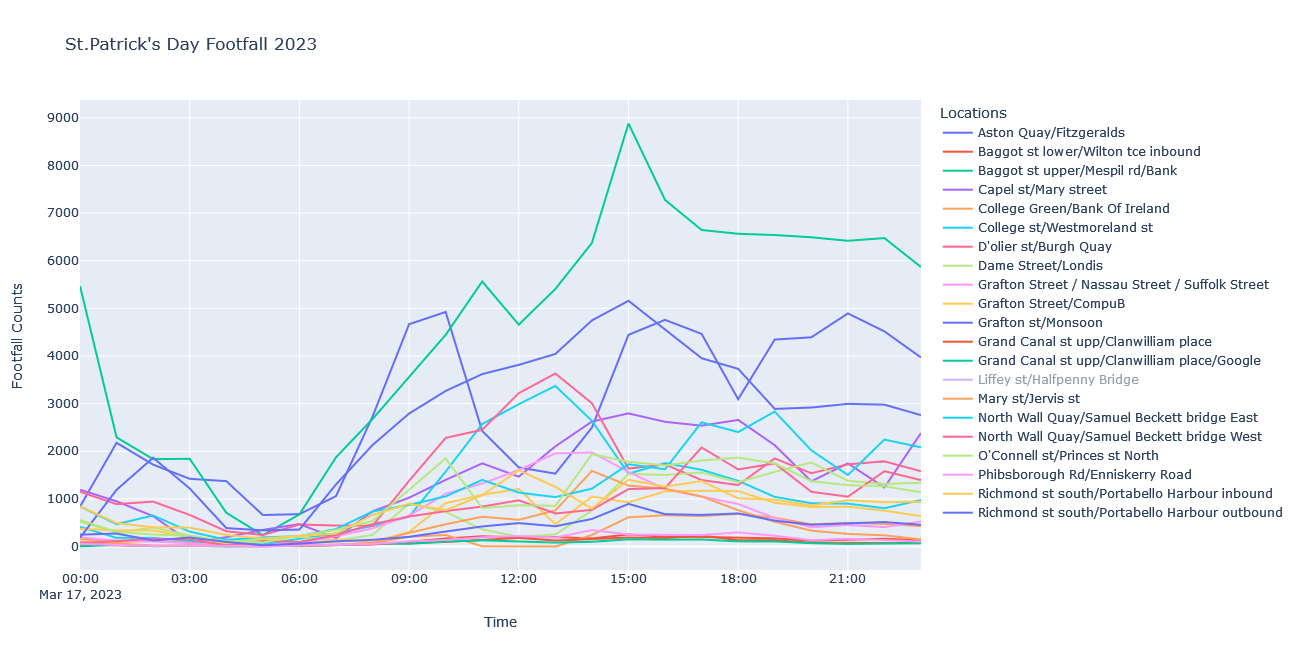


Figure 6 Line plot of average footfall count at each location by hour

Figure 7 Line plot of footfall counts at each location on St. Patrick’s Day 2023



### Sorting Locations by Footfall Volume

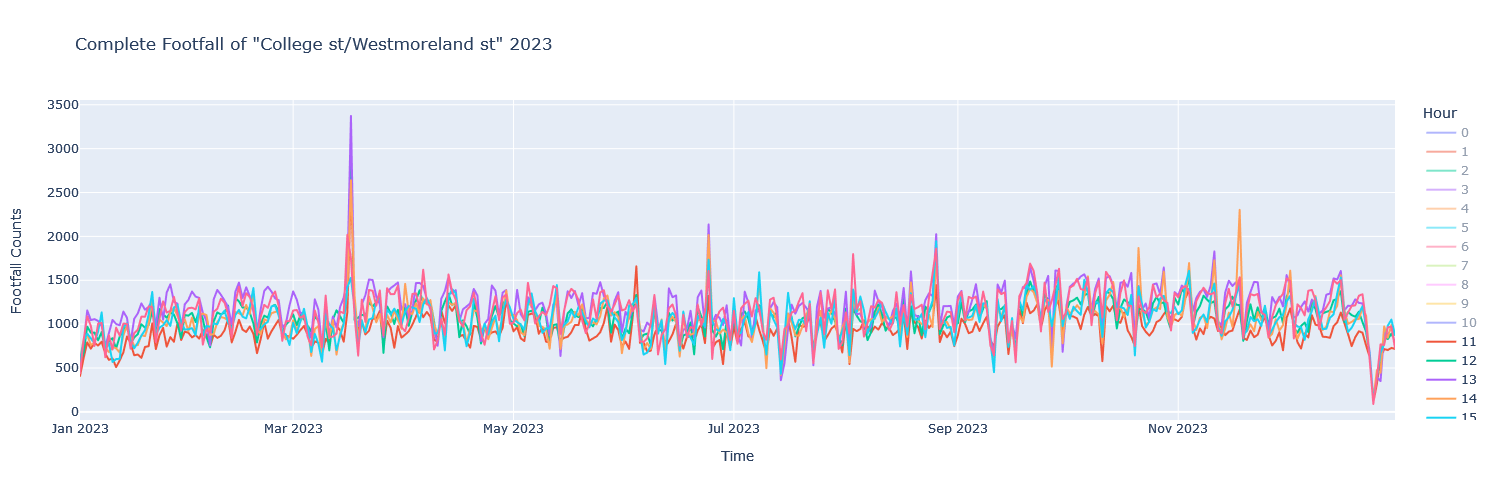


Figure 8 Line plot showing footfall counts over 2023 at "College st/Westmoreland st", between hours 11am and 4pm

Bar charts were utilized to pictorially represent the volume of footfall at various locations, offering a visually engaging means to grasp the distribution of activity from the busiest to the quietest sites, as in Figure 9. This approach not only enhances comprehension of footfall volumes but also effectively highlights the ranking of locations based on their activity levels.

This visual method proves valuable for gauging the scale of footfall, providing a clear, comparative perspective on the dynamics of pedestrian footfall across different areas. For instance, streets such as “D’Olier St/Burgh quay” are observed to experience a proportionate increase in footfall on St. Patrick's Day compared to an average day.

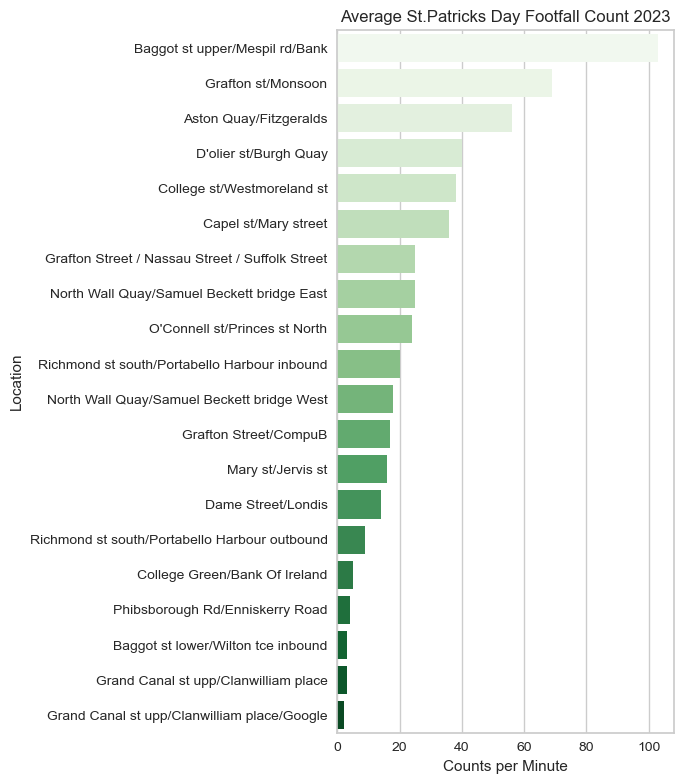
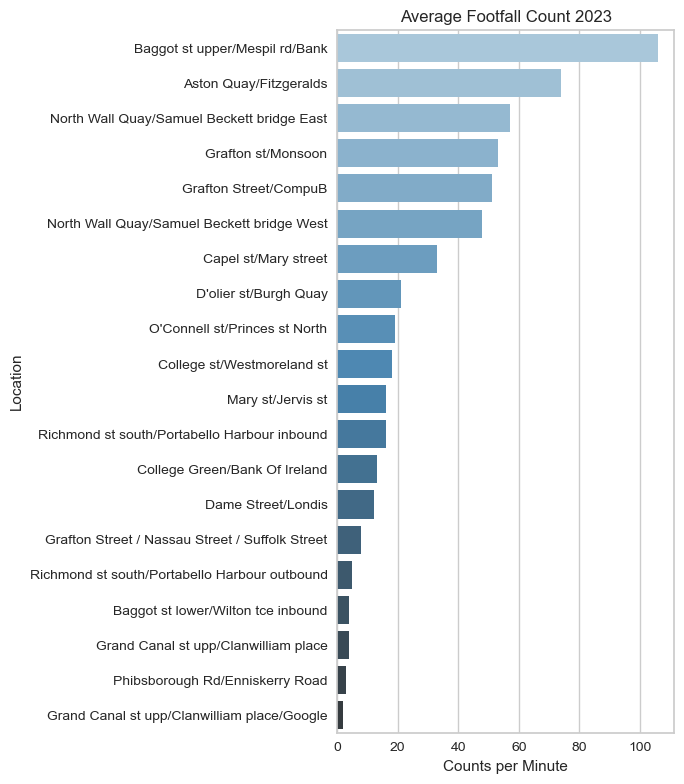


Figure 9 Bar charts showing locations sorted from busiest to quietest, in Counts per Minute

### Visualisation of Counter Locations

Creating a geographic visualization can enhance our understanding of the geospatial significance of each location. For example, 'College Green/Bank Of Ireland' may play a more important role in highlighting the main pathways of inner-city footfall compared to other sites. When this technique is applied to the project datasets, it not only allows us to visualize the coordinate locations but also enables the use of the size of each marker to represent the volume of footfall. Figure 10 illustrates the footfall volume on a typical day, while Figure 11 shows the footfall on St. Patrick’s Day. This method provides insights into various aspects, such as tourist experiences or the effects of events like the St. Patrick’s Day parade, which follows a specific route. Such visualization offers a more intuitive understanding of how footfall is distributed across different areas.

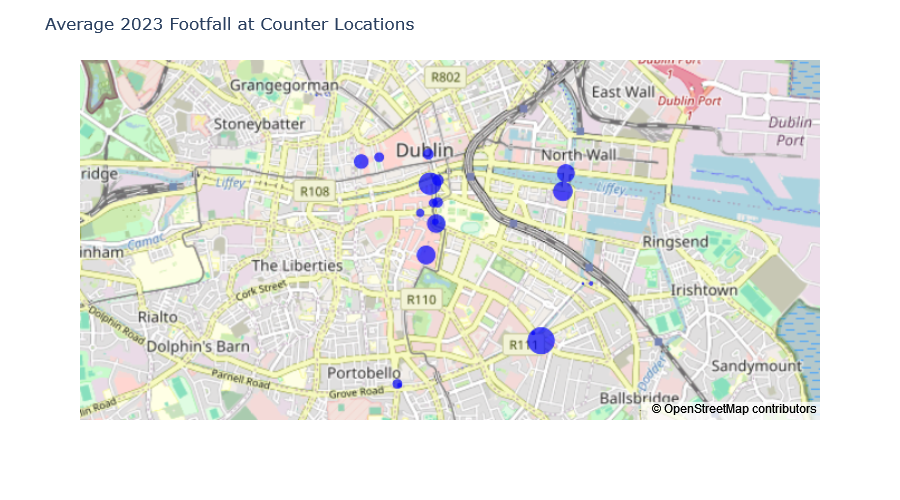


Figure 10 Visualisation of counter locations and footfall counts across the year 2023

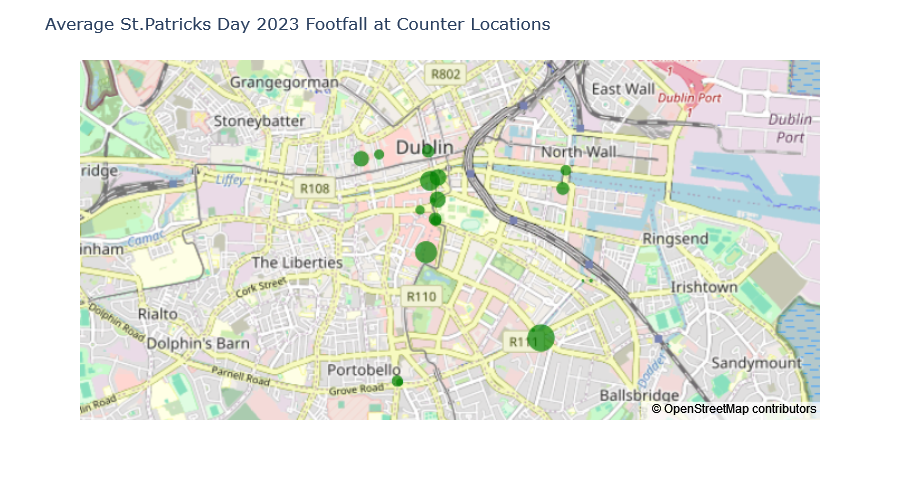
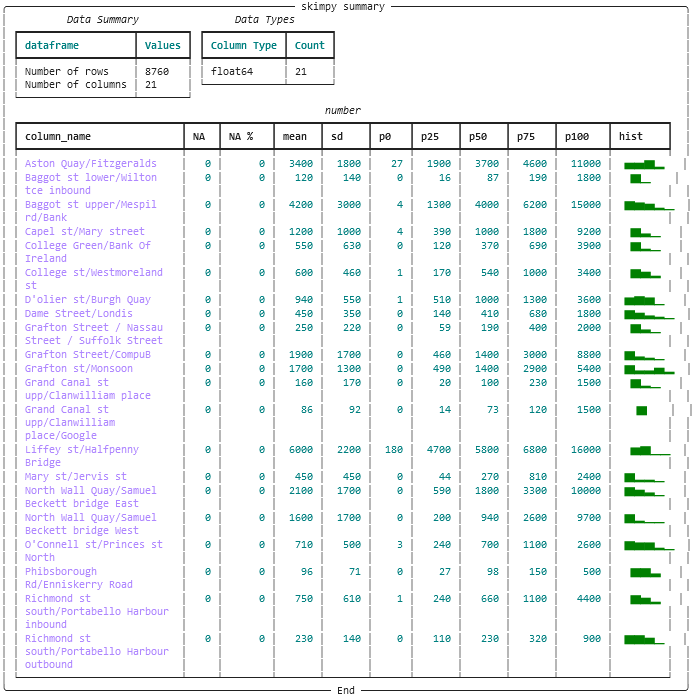


Figure 11 Visualisation of counter locations and footfall counts on St. Patrick's Day 2023



## Statistical Analysis *(Section 3)*

Statistics refers to the mathematics and techniques used to interpret data, and there are a wide range of analytical methods which can be employed. In this project, descriptive statistics—mean and median—were applied to provide a sense of typical footfall counts. Additionally, the standard deviation, minimum, and maximum values were analysed to understand the variations between high and low counts, thus offering a scale of footfall variability. Percentiles were also calculated, aiding in the identification of periods of higher or lower footfall. Moreover, various visualizations, including histogram plots, box plots, and distributions, were utilized to illustrate the findings. This section aims to summarize the results of the statistical analysis conducted on the dataset.

### Summary of Descriptive Statistics

In Table 3, the values for the mean, standard deviation, and percentiles are presented. Notable observations from this data include the relatively large mean of 6000 counts at "Liffey St/Halfpenny Bridge", and a high standard deviation at Grafton Street locations, indicating a wide range of counts.

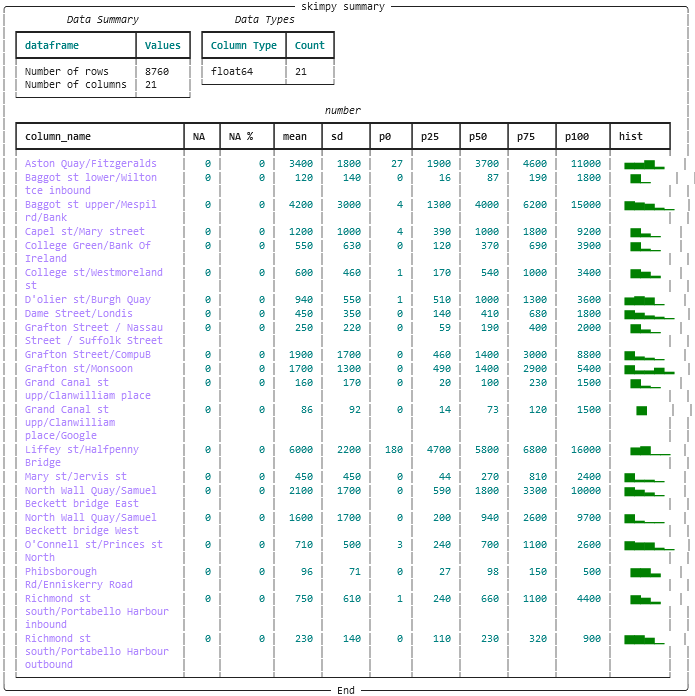


Table 3 Table showing a summary of descriptive statistics from "skimpy" module, of all counts over 2023

### Histograms

Histograms serve as a method to visualize frequency distributions, utilizing equally spaced "bins" on the x-axis to represent different ranges of values, with the proportion or frequency of those values plotted on the y-axis. In this context, "frequency" denotes the number of times an hourly footfall count is observed. The independent variable in these histograms is the Footfall count, while the dependent variable is its frequency A summary of observations for the year 2023 and St. Patrick's Day individually is presented below. The graphs for 2023 are displayed in Figure 12, while the histograms for St. Patrick's Day are shown in Figure 13.

**Histograms of counts over 2023**

* Generally, the histograms exhibit right skewness with a long tail and some centred peaks in footfall. This pattern suggests that, for most of the time, there is a lower proportion of footfall on the streets, with centred peaks indicating periods of high footfall.
* The right skewness may be influenced by the inclusion of nighttime hours in the data, suggesting that separating daylight from nighttime hours could offer better insights into footfall patterns.
* Locations like “Aston Quay/Fitzgeralds” and “D'Olier Street/Burgh Quay” display centralized high peaks in their histograms, indicating a consistently higher footfall.
* "Liffey St/Halfpenny Bridge" demonstrates a more normalized distribution, albeit with a noticeable absence of lower counts.
* “Capel St/Mary Street” and “Grafton Street/Nassau Street/Suffolk Street” both exhibit a similar shape of right skewness. However, Capel Street/Mary Street generally experiences significantly higher footfall.
* “College Green/Bank Of Ireland” demonstrates a wide range of footfall, with a consistent distribution tail between 1000 and 4000.

**Histograms of counts over St. Patrick’s Day 2023**

* A tendency towards right skewness is observed, as seen with “D'Olier St/Burgh Quay” and “Grand Canal St Upp/Clanwilliam Place”.
* Most distributions show slight bimodality, characterized by two peaks. This pattern could be attributed to the distinction between peak and off-peak hours, along with the limited sample size of counts over a 24-hour period.
* "Liffey St/Halfpenny Bridge" experiences a proportionately large number of high counts.
* “Baggot st upper/Mespil rd/Bank” and “Grafton st/Monsoon” have a high proportion of large counts

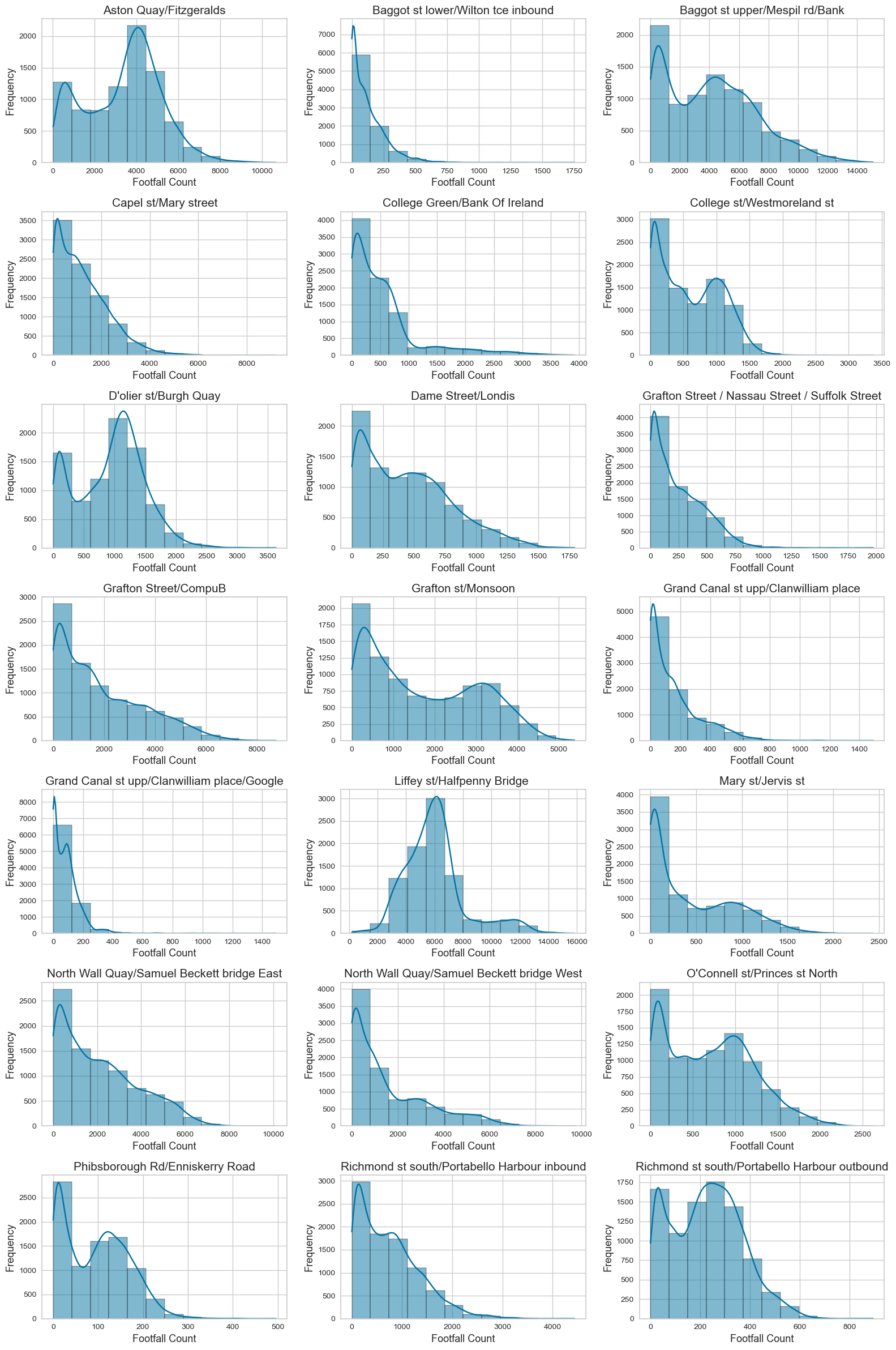


Figure 12 Histograms showing the distribution footfall counts at each location, over the year 2023



Figure 13 Histograms showing the distribution footfall counts at each location, over St. Patrick's Day 2023

### Percentiles and Boxplots[¶](http://localhost:8892/notebooks/ca1-sbs24004/TourismFootfallStPatricksDay.ipynb#3.1.2-Percentiles-and-Boxplots)

Percentiles are instrumental in exploring the spread of data and summarizing the entire distribution. Boxplots are employed to visualize these percentiles, offering an immediate preliminary interpretation. For this analysis, boxplots were created for counter locations for both the entire year, as seen in Figure 14, and specifically for St. Patrick's Day 2023 and 2022, Figures 15 and 16 respectively, providing insight into the variability and distribution of footfall counts, as well allowing for comparative analysis.

**Box Plots of footfall counts over the year 2023**

* Certain streets, such as “Baggot St Upper/Mespil Rd/Bank”, exhibit a very wide interquartile range, suggesting significant variation in footfall counts throughout the year.
* Locations like “Aston Quay/Fitzgeralds”, “Baggot St Upper/Mespil Rd/Bank”, and “Liffey St/Halfpenny Bridge” have considerably higher medians compared to other areas, indicating more consistent footfall.
* Upper bound outliers are present across all locations, potentially signalling periods of increased footfall due to events like the St. Patrick's Day festival. These outliers could be useful in isolating and identifying specific events.
* “Liffey St/Halfpenny Bridge” is notable for a significant number of outliers, both above and below the interquartile range, highlighting both exceptionally high and low footfall counts.

**Box Plots of footfall counts over St. Patrick's Day 2023**

* “Liffey St/Halfpenny Bridge” features very large outliers and did not record any low counts throughout the day, indicating consistently high footfall.
* Locations such as “D'Olier St/Burgh Quay” and “Mary St/Jervis St” have upper outliers, reflecting spikes in footfall.
* “Aston Quay/Fitzgeralds” and “Grafton St/Monsoon” show similar distributions over the day, though the latter has a higher median, suggesting a denser central distribution of footfall.
* “Baggot St Upper/Mespil Rd/Bank” demonstrates a large variation throughout the day, with both high and low footfall counts observed.

**Box Plots of footfall counts over St. Patrick's Day 2022**

* Both “College St/Westmoreland St” and “Grafton Street/Nassau Street/Suffolk Street” experience upper outliers, indicating instances of high footfall.
* “Grand Canal St Upp/Clanwilliam Place/Google”, show a similar distribution of low counts over the day, suggesting areas of lower footfall.
* “Mary St/Jervis St” exhibits large counts exceeding its median, highlighting periods of significantly increased footfall

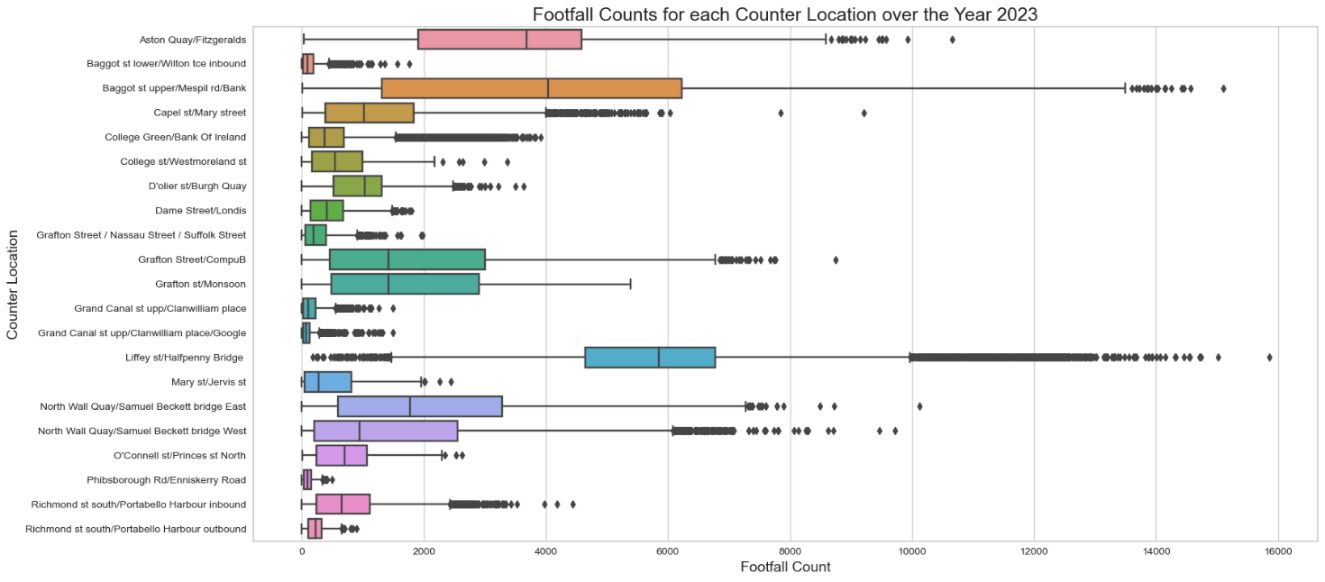


Figure 14 Box Plots of footfall counts over the year 2023

Figure 15 Box Plots of footfall counts over St. Patrick's Day 2023

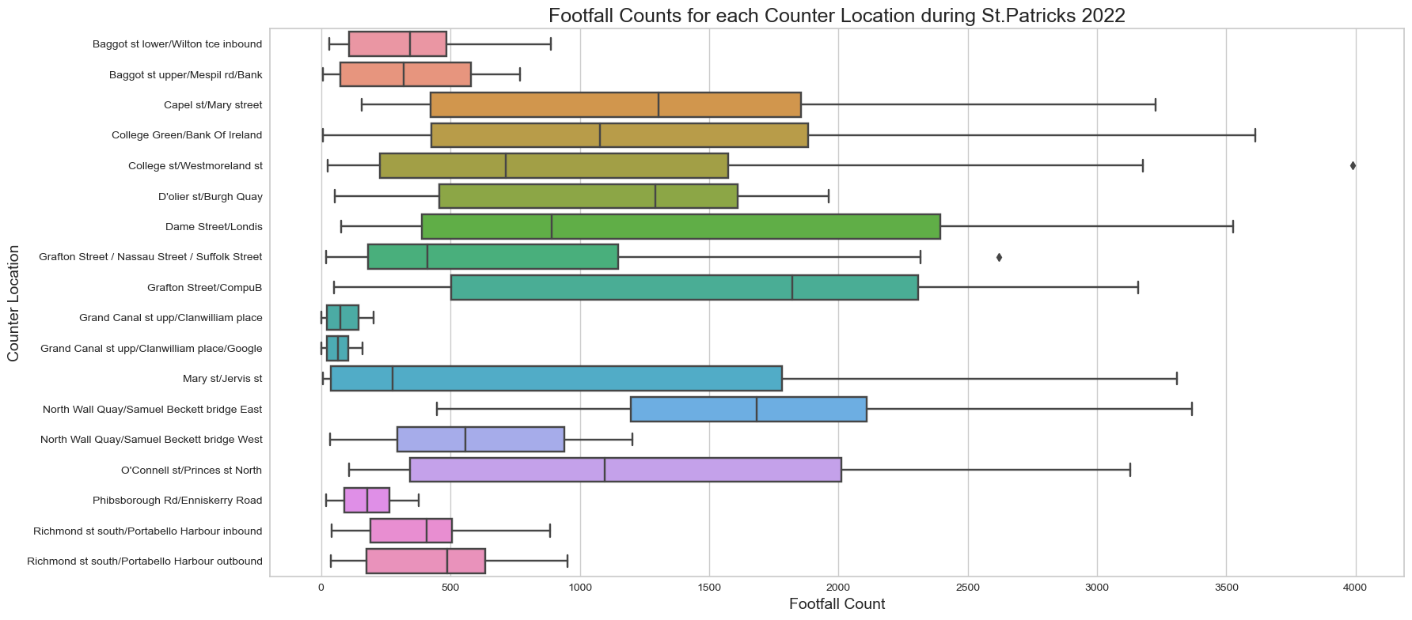
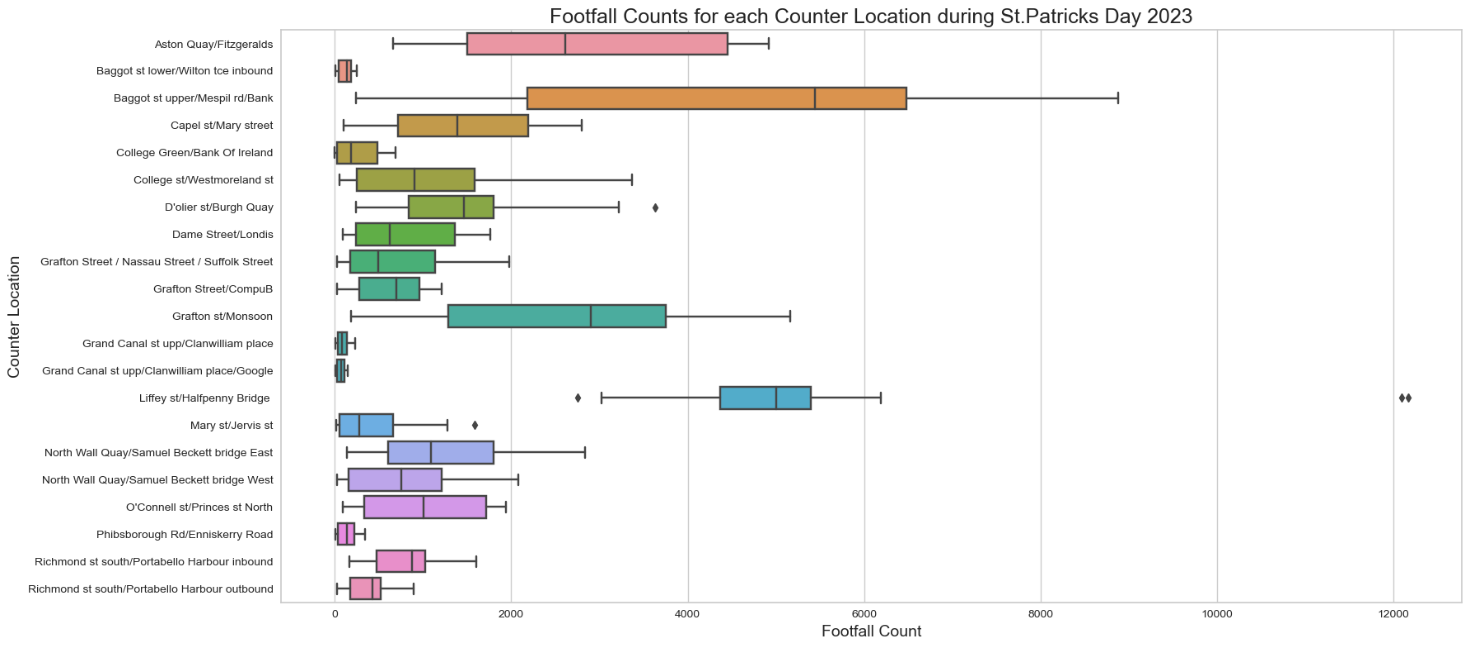


Figure 16 Box Plots of footfall counts over St. Patrick's Day 2022

### “Liffey st/Halfpenny Bridge” Analysis and the Poisson Distribution

Given the significant footfall counts, outliers, and variances observed at “Liffey Street/Halfpenny Bridge”, this analysis delves into the application of the Poisson Distribution to evaluate the likelihood of such levels of footfall. The Poisson distribution is apt for modelling the frequency of events occurring within a fixed interval, either in time or space, making it an ideal framework for this examination. For this purpose, the average footfall counts recorded across all locations during the entire year of 2023 at 3 am against the average footfall counts specifically observed at the same time at “Liffey Street/Halfpenny Bridge” were utilised. This comparative approach aims to elucidate the deviation of footfall counts at “Liffey Street/Halfpenny Bridge” from the normal patterns.

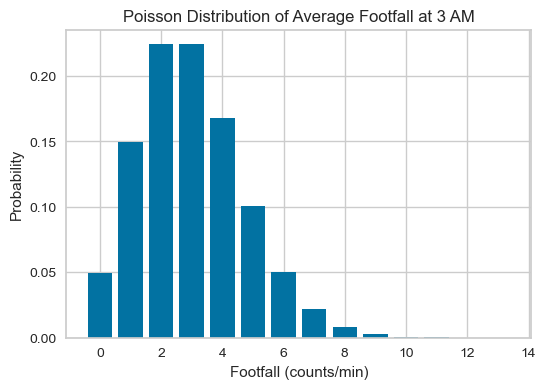


Figure 17 Poisson Distribution of Average Footfall at 3 AM, with counts observed being 110 per minute

**Key Findings:**

* The average footfall counted at “Liffey Street/Halfpenny Bridge” at 3 am is markedly higher than that observed at other locations, falling significantly outside the expected range.
* The likelihood of any location registering an average footfall exceeding 110 per minute at 3 am is calculated to be nearly 0%, suggesting an extreme anomaly in the data.
* Owing to the highly improbable nature of such footfall figures, the counts recorded at Liffey Street/Halfpenny Bridge were deemed unrepresentative of true footfall levels and were excluded from analysis.

### Total Daily Footfall: Normal Distribution and the Central Limit Theorem

The Normal Distribution has served as a cornerstone in the development of modern statistics, providing foundational tools for estimating uncertainty and variability. While raw data may not always adhere to a normal distribution, the Central Limit Theorem has indicated that the distribution of averages or totals from large samples typically approximates the familiar "bell curve", allowing for assumptions about the data. From the histograms of individual footfall counts across various locations, it was observed that they do not exhibit a normal distribution. However, by resampling these counts to calculate the total daily footfall at each location, as in Figure 18, it was found that these totals align more closely with a normal distribution.

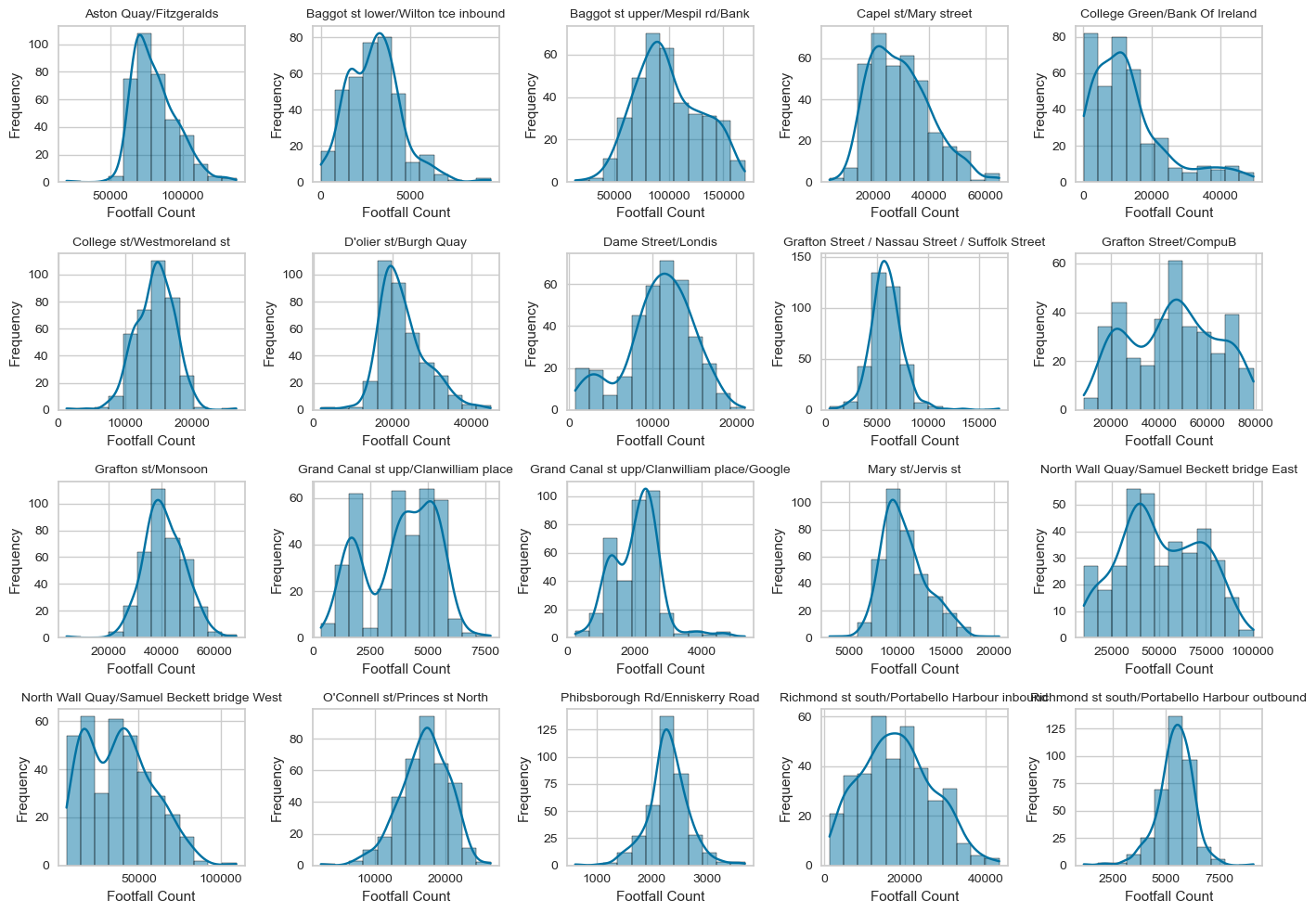


Figure 18 Histograms of total daily footfall counts by location, displaying conformity to a normal distribution

**Key Findings:**

* The aggregated daily counts across all locations tend to exhibit characteristics of a normal distribution, particularly when contrasted with the distribution of the counts themselves.
* Notably, several histograms, including those for “Grafton Street/CompuB” and “Grand Canal St Upp/Clanwilliam Place”, do not wholly exhibit the characteristic bell shape, indicating deviations from normal distribution in these specific instances.

### QQ Plots

A QQ plot, or quantile-quantile plot, serves as a visual representation of how closely a sample set aligns with a given distribution, in this instance, a normal distribution. In this plot, the Z-scores from the sample data are plotted on the y-axis, and the corresponding quantiles of the normal distribution are plotted on the x-axis. Should the dataset adhere to the specified distribution, the plotted points will closely align with the red diagonal line. Applying QQ plots to the total footfall counts of each location enables a check on how well these totals conform to a normal distribution. It is important to note that converting data to Z-scores does not inherently normalize the data but rather scales it for comparison purposes with the normal distribution.

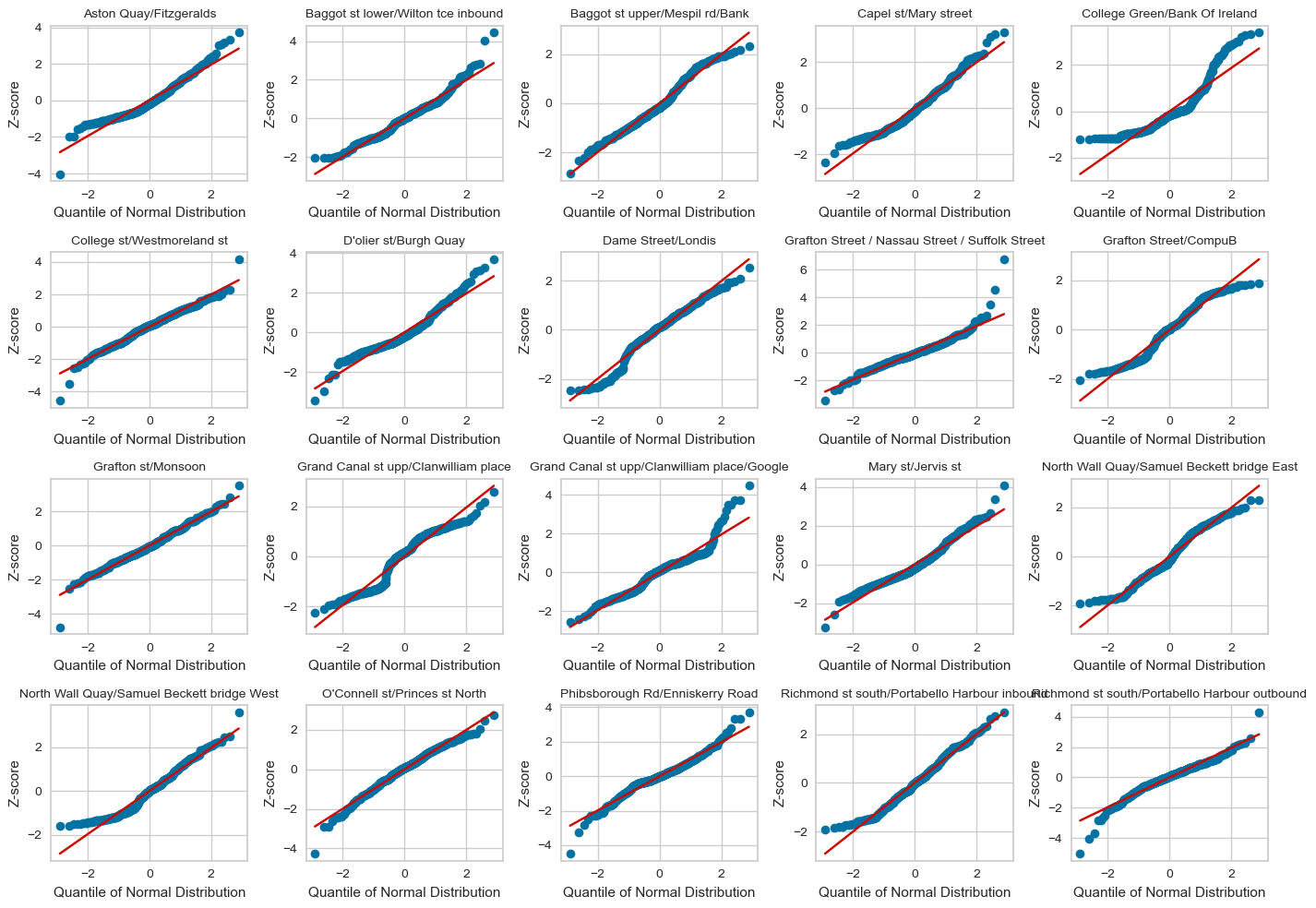


Figure 19 QQ plots of total daily footfall counts by location

**Key Findings:**

* The QQ plots, shown in figure 19 provide a clearer illustration of how well the total daily counts match a normal distribution.
* Generally, all locations exhibit a good match with a normal distribution.
* Locations such as "College St/Westmoreland St" and "Grafton St/Monsoon" closely reflect the normal distribution.
* Conversely, locations like "College Green/Bank Of Ireland" and "Grand Canal St Upp/Clanwilliam Place" show a poorer reflection of the normal distribution compared to other locations, yet they still follow it reasonably well.

## Machine Learning *(Section 4)*

### Project Management Frameworks

1. CRISP-DM, (Cross-Industry Standard Process for Data Mining)

* Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.
* Scenario: A tourism board wants to boost local tourism by identifying and targeting visitors based on their interests
* Justification: Clear business goals and strategy and to enhance tourism through targeted marketing

1. KDD (Knowledge Discovery in Databases))

* selection, preprocessing, transformation, data mining, and interpretation/evaluation.
* Scenario: A tour operator wants to improve their services by analysing customer feedback
* Justification: The action is exploratory, and may uncover insights or preferences for tailoring customer experience

1. SEMMA (Sample, Explore, Modify, Model, Assess)

* Structured, focused on model building, validation, and accuracy
* Scenario: A travel website wants to develop a recommendation system for users based on their history and preferences
* Justification: The goal is to create a prediction model which is both accurate and reliable

### Unsupervised Learning of Footfall Count Data

For this project, the objective was to segregate daily footfall patterns and identify which patterns resembled those of high footfall events, such as observed on St. Patrick's Day. Unsupervised learning models were selected for their ability to identify natural groupings within data. By applying k-means clustering, it was possible to segment days based on similar footfall patterns. The analysis was exploratory in nature due to the complexity of footfall dynamics and time series analysis. It was an uncertainty as to whether similar patterns were present during other periods of Dublin events.

Several approaches were assessed, and for the analysis, k-means clustering of “College St/Westmoreland” was specifically chosen to evaluate distinctions in the number of clusters, as well as to provide a complementary example with agglomerative clustering. The locations “Grafton Street / Nassau Street / Suffolk Street” and “Dame Street/Londis” were also evaluated using k-means clustering to identify the closest grouping to the footfall patterns observed on St. Patrick's Day.

### k-means Clustering of "College st/Westmoreland"

This approach utilized the k-means algorithm from the scikit-learn library, along with visualization tools from Yellowbrick. Data for the specified location was segmented and organized by day and hour. Given that clustering algorithms are sensitive to scale, the data was normalized using the StandardScaler from the scikit-learn library. The k-means clustering algorithm was then applied, with both inertia and silhouette scores visualized to assess the model's performance.

An inertia, or elbow diagram, as depicted in Figure 20, aids in determining the optimal number of clusters by identifying the "elbow point," typically between 3 to 5 clusters. In this case, the optimal number was found to be 4. This diagram helps in assessing the compactness of the clusters, with the elbow point indicating where adding more clusters does not significantly improve the fit of the model.

Silhouette scores per cluster were also visualized, as shown in Figure 21, with an optimal score identified at 2. The silhouette score measures how similar an object is to its own cluster compared to other clusters, providing a clear visualization of how well each object has been classified.

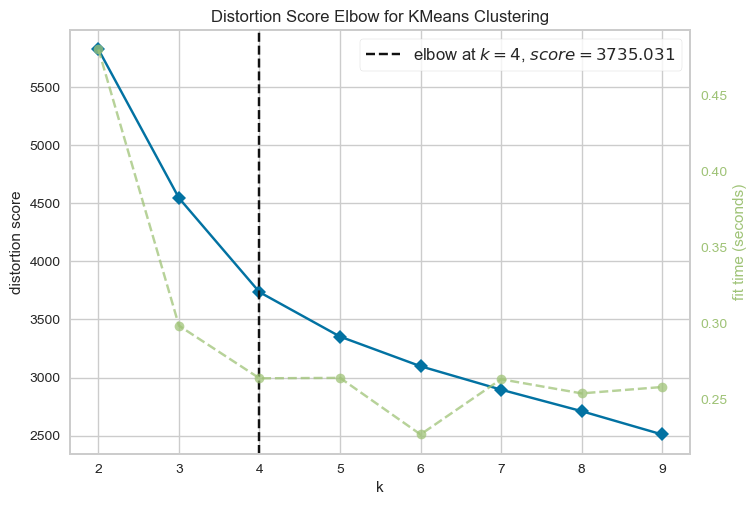


Figure 20 Elbow diagram for k means clustering for "College st/Westmoreland"

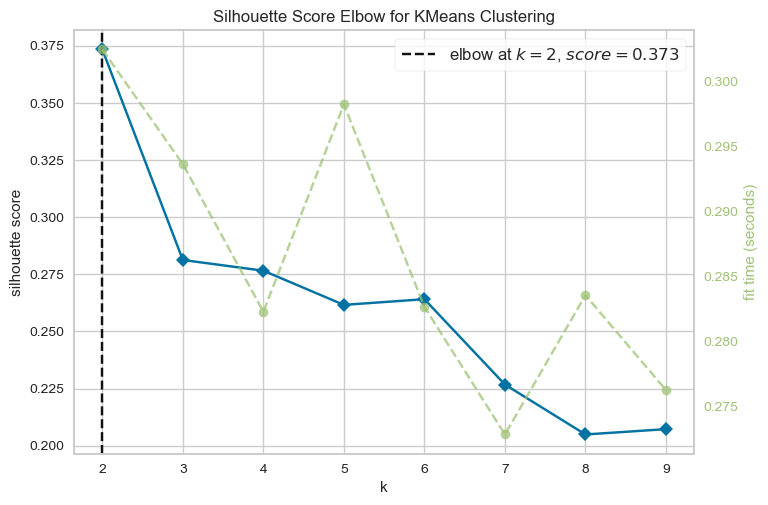


Figure 21 Silhouette Score Plot for k means clustering of "College st/Westmoreland"

### Silhouette Diagrams

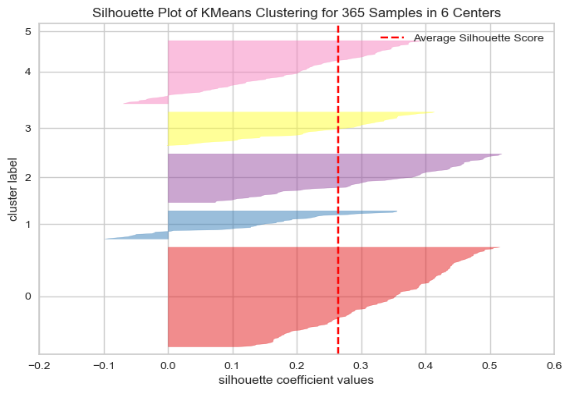
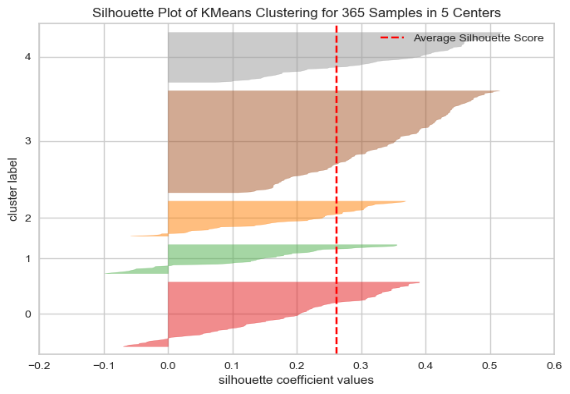
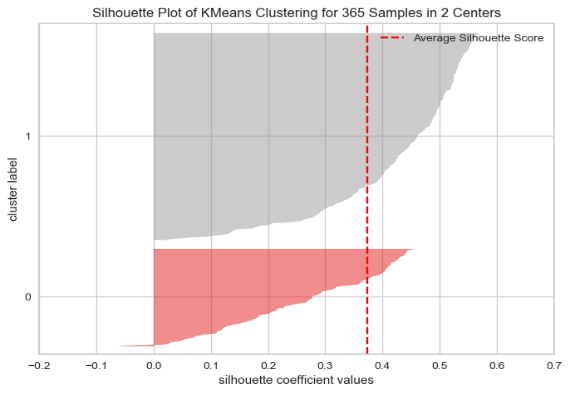
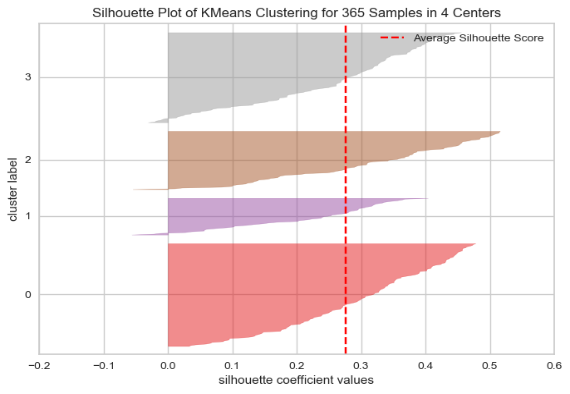
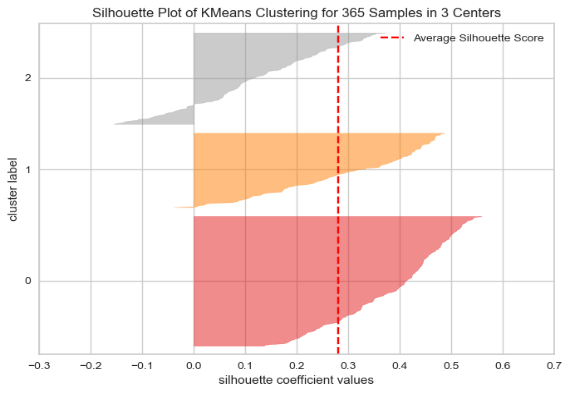


Figure 22 Silhouette Diagrams for k means clustering of "College st/Westmoreland"

Additionally, silhouette diagrams were created to visualize the grouping of each cluster, offering a deeper insight into the distribution of data within and across the clusters. This was part of a broader effort to explore hyperparameters and compare models, as highlighted in Figure 22, aiming to fine-tune the clustering process and achieve more precise groupings that closely match the footfall patterns observed on St. Patrick's Day.

### Heat Maps

To further evaluate the outcomes of the k-means clustering analysis, heat maps representing each clustering value from 2 to 6 were generated, as displayed in the left column in Figure 23. These heat maps provide a visual representation of how the data was grouped, emphasizing similarity between patterns of footfall counts.

### Centroid Visualisation

The centroids from the k-means clustering were plotted to visualize the grouping of footfall patterns, as shown in the right column of Figure 23. Each cluster's centroid represents the average position of all points within that cluster. The centroids can also be compared to their respective heat maps, to discern the central tendencies and variances within each cluster. At k=6, St. Patrick’s Day footfall counts are clearly isolate from the rest. At k=6, the footfall counts for St. Patrick's Day are clearly isolated from the rest. Following, at k=5, the footfall for St. Patrick's Day is grouped closest with patterns observed on other days.

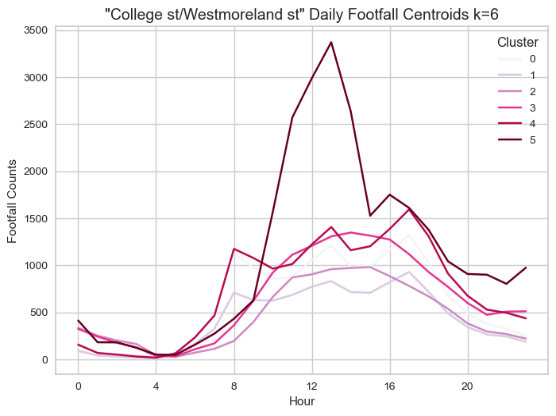
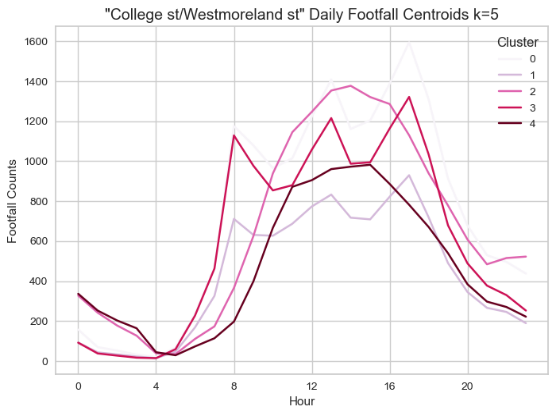
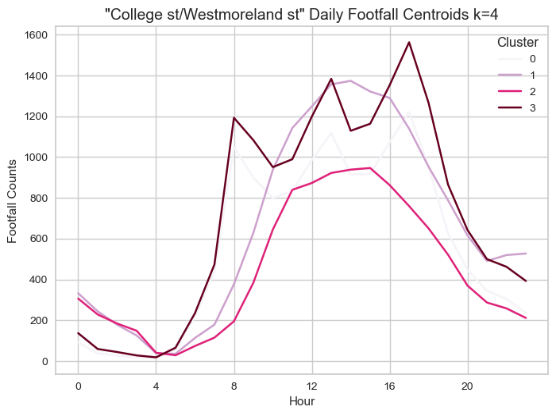
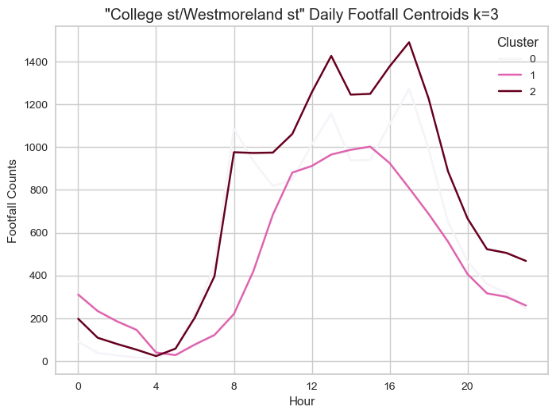
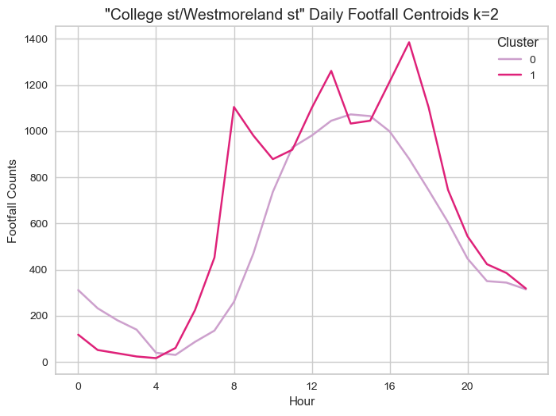
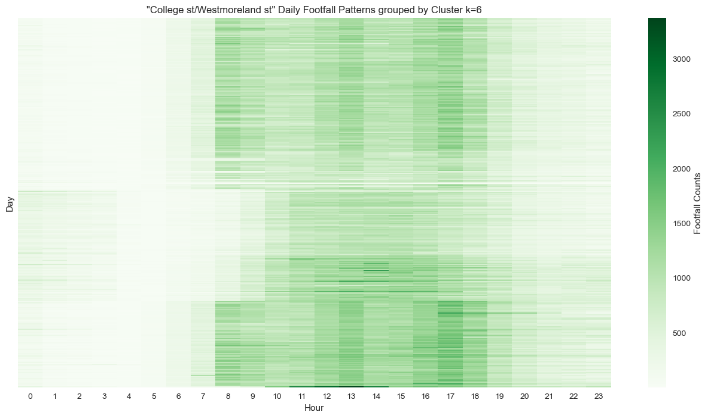
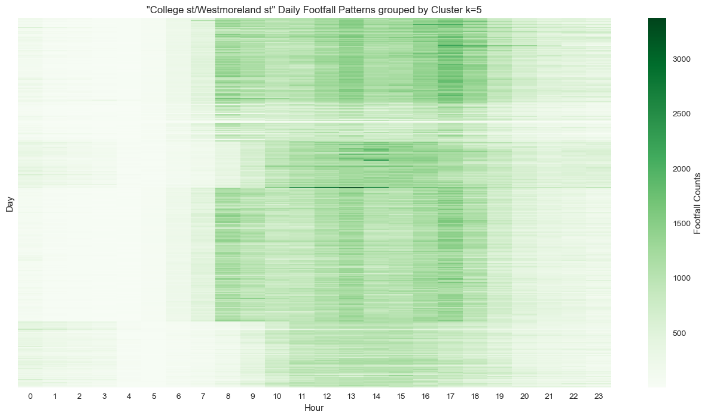
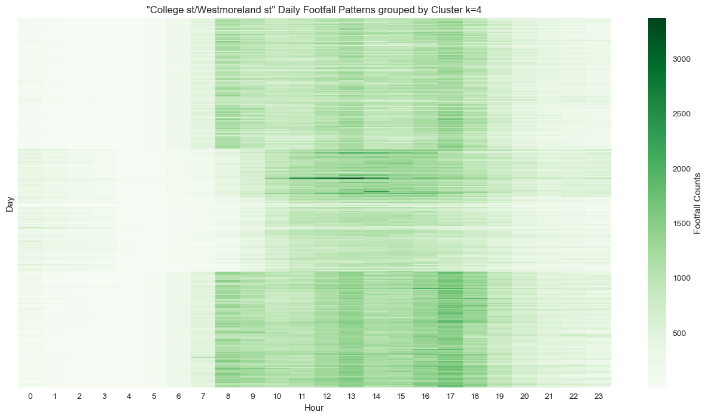
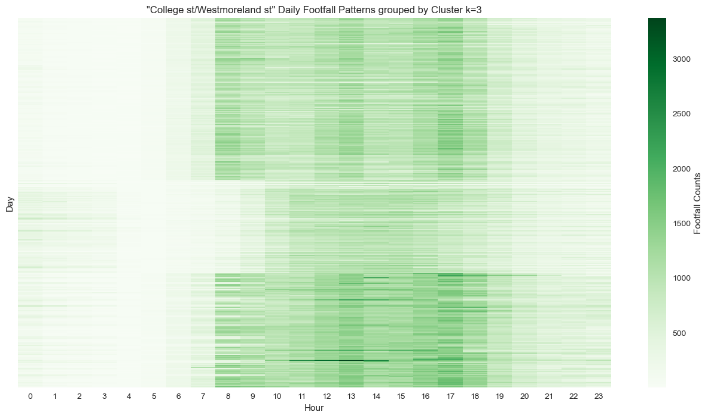
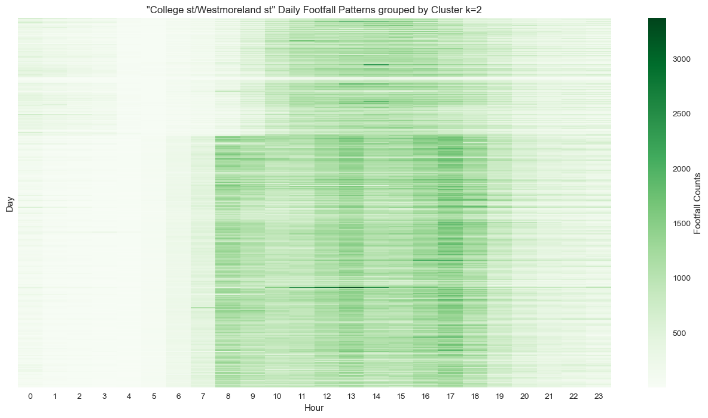


Figure 23 Visualization of results from k=2 to k=6 after k-means clustering is presented with heat maps in the left column, and centroids in the right column, showcasing daily footfall patterns

### k-means Clustering of "Grafton Street / Nassau Street / Suffolk Street" and “Dame Street/Londis”

The same approach was applied to locations "Grafton Street / Nassau Street / Suffolk Street" and   
“Dame Street/Londis”. The results are presented in Figure 24, and k values were chosen to cluster the most similar patterns to those on St. Patrick’s Day.

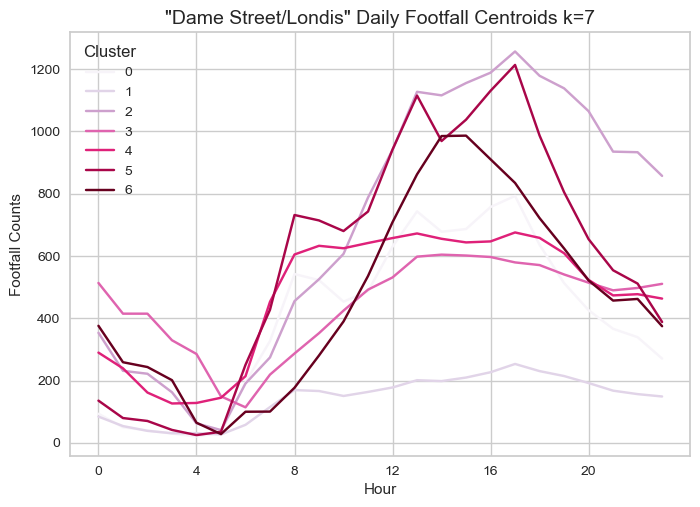
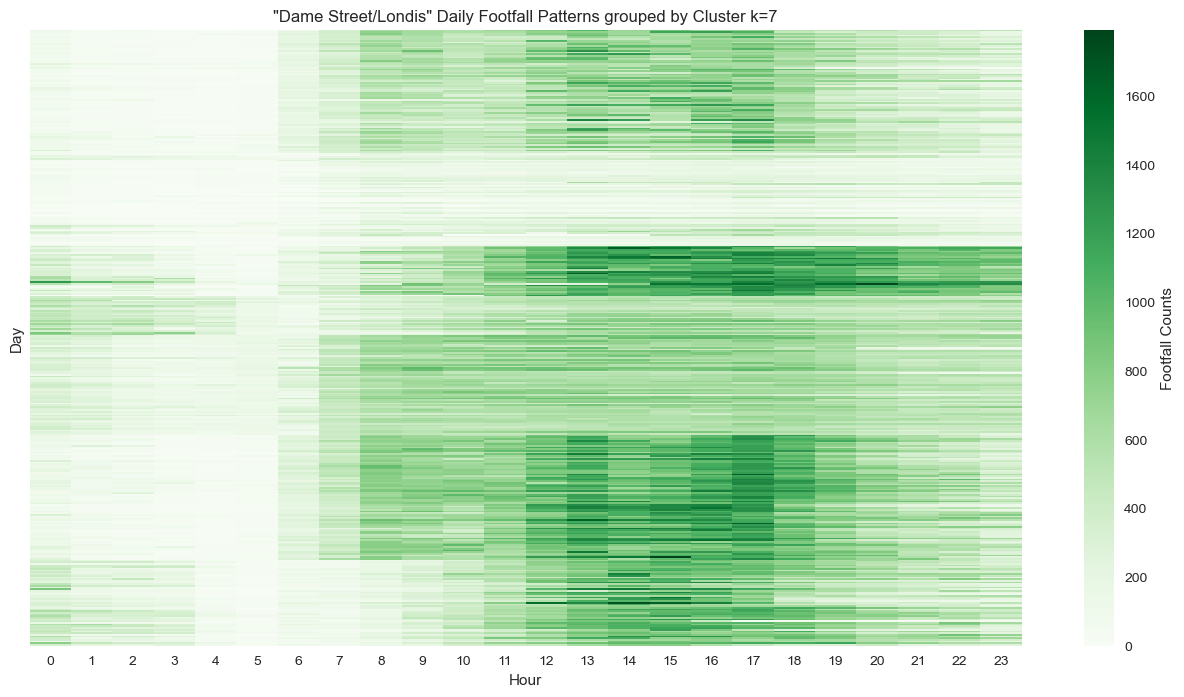
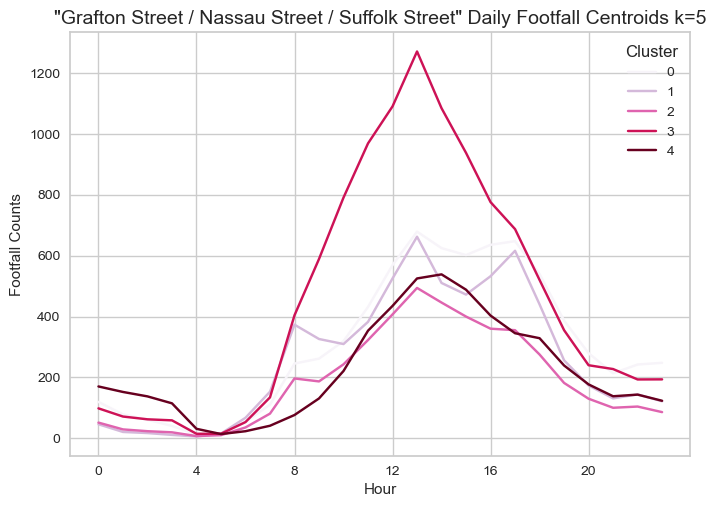
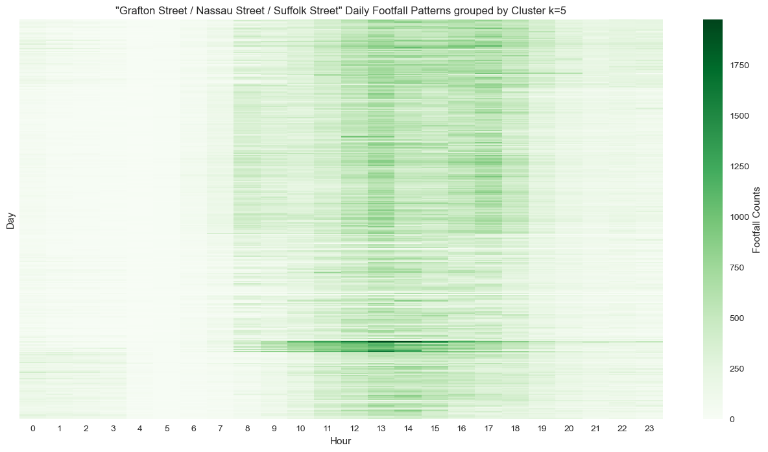


Figure 24 Visualisation of k means clustering results for "Grafton Street / Nassau Street / Suffolk Street" , k=5, and of "Dame Street/Londis", k=7

### Hierarchical Clustering of "College st/Westmoreland"

Hierarchical or Agglomerative Clustering is a form of unsupervised learning that creates clusters from the bottom up. This process is often likened to tiny bubbles on water, gradually merging to form a single, larger group of bubbles. The agglomerative clustering algorithms from scikit-learn were used to analyse "College St/Westmoreland", k=5, allowing for a comparison of results with those from k-means clustering, as depicted in Figure 24. A "Hierarchical Clustering Dendrogram" visualizes how clusters are formed at various levels of similarity. In Figure 25, the dendrogram illustrates the process of clustering for "College St/Westmoreland".

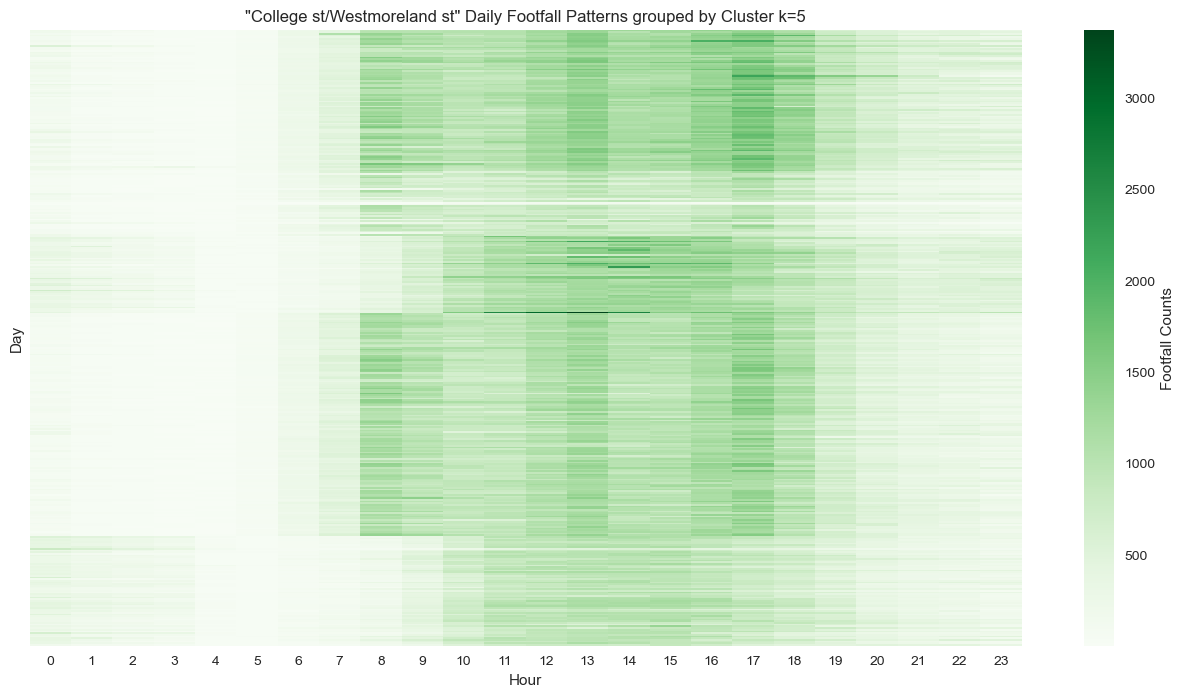
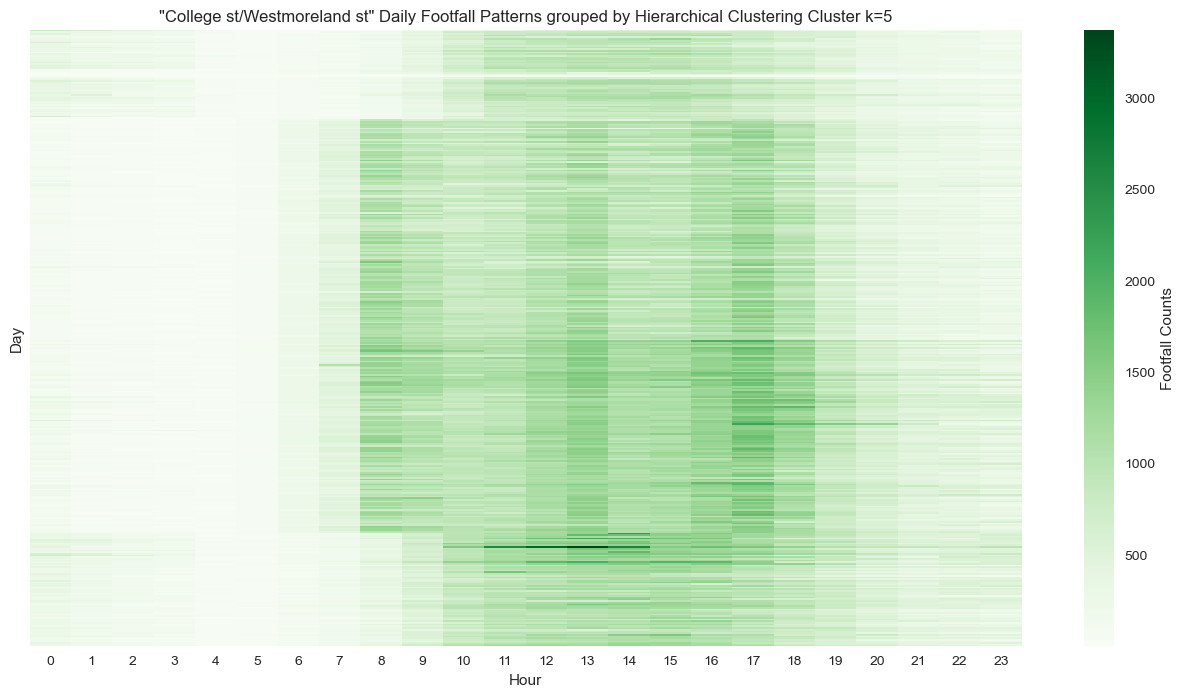


Figure 25 Heat maps comparing results from k means (Left) clustering to agglomerative clustering (Right) at k=5

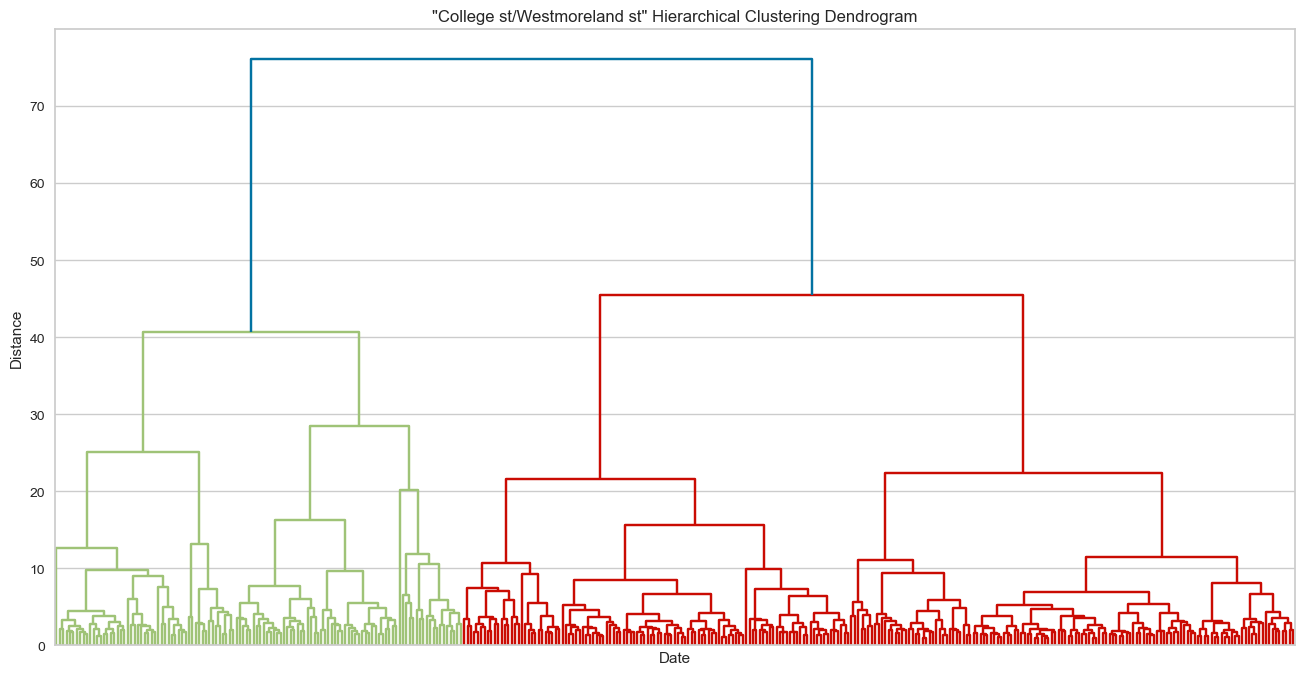


Figure 26 Hierarchical Clustering Dendrogram illustrating the process of clustering for "College St/Westmoreland"

## Results

The table below presents the dates clustered by the algorithms, aligning similar dates adjacent to each other for a clear comparison.

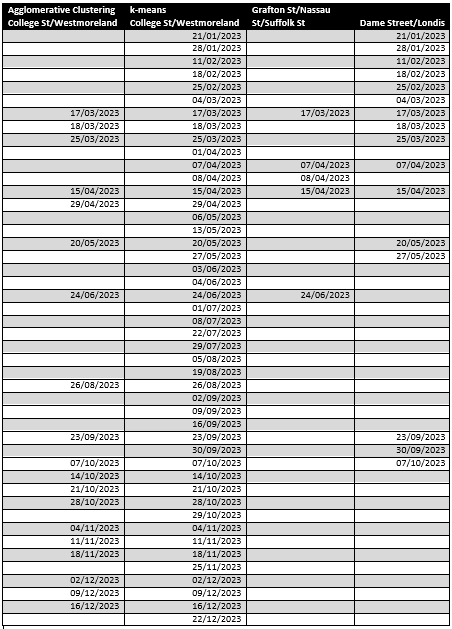


Table 4 Dates grouped by clustering algorithm

## Tufte’s Principles

Tufte's principles, embodying the ethos of clarity, efficiency, and clear communication of data through visual representation, were diligently adhered to throughout this project. This was achieved by focusing on the crucial aspects of the data, ensuring colour matching was aesthetically pleasing, and combining data in a way that complements each other.

## Programming Discussion

Python is one of the most popular programming languages and has fostered a large and active community in both scientific computing and data analysis. This project was constructed using version Python 3.11.5 in a Jupyter Notebook environment, with all code thoroughly annotated. This section discusses the libraries and specific programming techniques used, along with the reasons for their selection. Several specialized tools and libraries were utilized:

* **NumPy (Numerical Python):** It provides the data structures and algorithms necessary for most scientific applications involving numerical data, including multi-dimensional arrays and mathematical operations between arrays. Most other libraries are built upon this, and it is foundational for general operation. Instances of direct use in this project were in the application of ‘np.arange’, which creates a range of values, for plotting Poisson distributions over a certain discrete interval. See sections 3.6 and 3.7.2 in the notebook.
* **Pandas:** This library offers high-level data structures, making working with structured or tabular data intuitive and flexible. The most notable structures are the DataFrame, a data structure with rows and columns, and Series, a one-dimensional labelled array. It was used extensively in the project for all manipulation and data wrangling. Some Pandas functions employed include:
  + - **pd.read\_csv,** in section 1, used to import CSV files into the notebook as a DataFrame, which is important for working with the published footfall count data.
    - **pd.to\_datetime,** in section 2.2, used to convert time in an object format into a datetime format, facilitating time series analysis later on.
    - **pd.merge,** in section 3.10, where it was used to merge a counter coordinates DataFrame with a DataFrame containing footfall counts at that location, enabling dynamic and volume visualization with the Plotly library.
    - **pd.DataFrame,** in section 4.1.3, to create a DataFrame after performing a melt operation on KMeans centroid data. The centroid data, stored as an array, and after converting to a DataFrame, allowed visualization of clusters with sns.
* **Matplotlib and Seaborn:** Both are popular libraries for producing plots and data visualizations. Matplotlib, the more fundamental visualization library, is designed for creating publication-suitable plots. Seaborn, built on top of Matplotlib, facilitates the creation of attractive and informative statistical graphics, such as histograms, KDE plots, bar plots, and line plots. These libraries were used for most visualizations in this project, including, but not limited to:
  + - **sns.histplot,** in section 3.1.1, used in conjunction with a for loop to iterate over data and produce histograms with kernel density estimates for footfall at each location.
    - **sns.boxplot,** in section 3.1.2, used for visualizing percentiles and creating boxplots of footfall counts, useful for understanding the spread of data.
    - **sns.barplot,** in section 3.11, used for creating bar plots in order to visualize the volume of footfall counts at each location and ranking locations from busiest to quietest.
* **Plotly:** This library specializes in creating interactive visualizations, allowing users to dynamically explore data. In this project, it was used for creating interactive line plots of footfall counts, featuring zoom and pan capabilities, and for dynamically mapping the locations of counters to visualize the geographic spread of data and the volume of counts at each location.
  + - **px.line,** in section 3.5.1, was used for visualizing footfall counts at each location around the period of interest.
    - **px.scatter\_mapbox,** in section 3.10, was used to create a dynamic map of counter locations over Dublin city centre, with the size parameter being proportional to footfall volume.
* **SciPy Stats:** A module from the SciPy library, it offers tools related to statistical tests and descriptive statistics. SciPy itself integrates well with Scikit-learn, and in this project, it was used to visualize a dendrogram of agglomerative clustering results.
  + - **poisson.pmf,** in section 3.6, was used to calculate the probability mass function, as well as Poisson scores below or above a certain metric.
    - **stats.zscore,** in section 3.8.2, was used to calculate the z-score of footfall counts, in the creation of QQ plots, to be scaled with the normal distribution.
    - **stats.probplot,** in section 3.8.2, was used to create the QQ plots, for determining how well the counts followed a normal distribution.
* **Scikit-learn:** Known as a comprehensive machine learning toolkit, it includes numerous submodules for classification, regression, clustering, dimensionality reduction, and preprocessing.
  + - **IterativeImputer,** in section 2.4.4, was used to impute data over large time periods, where linear regression was not reflective of the data.
    - **StandardScaler,** in section 4, was used to scale data, important for processing the count data which was over large ranges, for ML algorithms which are sensitive to scale.
    - **KMeans**, in section 4.1, was applied for k-means clustering, and producing cluster labels at selected k values.
    - **AgglomerativeClustering,** in section 4.4, was employed to apply another unsupervised learning algorithm for clustering, comparing hierarchical clustering results against K-means.
* **Yellowbrick:** Designed for visualizations of machine learning model evaluation, it is built to integrate with Scikit-learn, enabling hyperparameter comparison and cluster visualization.
  + - **KElbowVisualizer,** in section 4.1, was utilized for generating Inertia or “Elbow” plots for evaluating K-means hyperparameters.
    - **SilhouetteVisualizer**, in section 4.1.1, for clustering silhouette score plots and silhouette diagrams, another method for evaluating clustering.

### Code Quality Standards

### An essential aspect of enabling the sharing of project code and facilitating collaboration among data scientists is its legibility and understandability. A key practice in this regard is the use of meaningful variable names that clearly reflect their purpose. In this project, variables and data frames predominantly followed the “snake\_case” convention, such as “footfall23\_df”, which denotes the footfall counts for the year 2023 DataFrame. Code indentation and spacing were applied to enhance visual organization and comprehensibility. All relevant code sections were annotated to explain the purpose of the operations being performed. Several common libraries, as listed above, were utilized to support the project's objectives. The code underwent numerous rounds of checking, testing, and refinement to ensure quality and reliability. The version control system GitHub was employed to document progress and changes at significant intervals, further enhancing collaboration and code management.

Table 5 A list of significant DataFrames and their definitions that were used in this project

### User Defined Functions

In alignment with code quality standards, a separate file named **UserDefinedFunctions.py** was created to store user-defined functions. Storing these unique functions in a separate file helps reduce clutter in the notebook and enhances functionality by allowing the code to be called from any point within the notebook. This approach is part of the modular programming paradigm, which advocates for breaking down a program into smaller, independent modules of code. In the context of this project, two user-defined functions were developed, which are outlined here:

* **remove\_IN\_OUT(df);** This function accepts a DataFrame as an argument and iterates through the column names of that DataFrame to create a list. If the column names contain the string "IN" or "OUT", those columns are added to a list, which is then used with .drop to remove those columns. Essentially, a DataFrame containing "IN" or "OUT" in its column names will have those columns removed. The purpose of creating this function in this project is that there were several CSV files of footfall counts, all with these erroneous columns. Instead of executing specific code each time, this function facilitates a repeatable cleaning step at any point in the notebook.
* **outlier\_dates(df\_column);** This function takes a DataFrame column as an argument, in the form “df[col]”. It calculates the Interquartile Range (IQR) for that column and returns the rows where the column values exceed the upper quantile. The purpose of this function is to identify upper outliers in a column/location, under the assumption that these dates represent periods of unusually high activity.