# An Analsis of 2019-2020 Dublin City Centre Pedestrian Patterns

Benjamin Guilfoyle, Declan Hill, Kevin Horan, Madhusudan Panwar

## 1 Introduction

This project uses footfall data for 23 Dublin city streets. Two years worth of data was used, 2019 and 2020. The goal here is to analyze patterns in pedestrian traffic with respect to time. With that in mind three distinct skews were taken when analyzing the data.

- Microscopic View: Looking at individual streets and the pedestrian frequency of each and comparing the two years.
- Macroscopic View: Analyzing Dublin as an entire city rather than individual streets with a focus on 2020.
- Predictive View: Trying to use 2020 as the basis for a predictive model and compare it against 2019.

These three approaches lead to the following conclusions. In the microscopic view one can see a clear change in the number of pedestrians on the streets of Dublin between 2019 and 2020. Looking on the macroscopic scale of 2020, this analysis shows Henry Street had the highest mean pedestrian footfall. Furthermore, one can look at the streets North, and South of the River Liffey that divides the city. There is a higher median number of pedestrians on the North side of the river. The busiest times for Dublin appear to be around 13:30, while the busiest month is February. This is likely due to the Ireland launching a nationwide COVID-19 lockdown starting in March. Finally, the predictive model led to the conclusion that 2020 was indeed abnormal year for pedestrian traffic. We attempted to use 2020 data to predict which streets in 2019 would be the busiest, and this proved unsuccessful, highlighting 2020 as a highly exceptional year for footfall traffic.
The project tasks were divided as follows

- Benjamin Guilfoyle: Predictive Modeling
- Declan Hill: Microscopic
- Kevin Horan: Macroscopic
- Madhusudan Panwar: Geographic Analysis & Feature Engineering

## 2 Libraries & Datasets

Several libraries were used throughout this project, highlighted below.

- tidyverse: Gives access to tools such as dplyr, ggplot2, and forecast to aid in data analysis.
- lubridate: Enables easier manipulation of dates.
- randomForest: Allows for the creation of randomForest objects for predictive modeling.
- gridExtra: Used to align ggplot objects like par(mfrow)

## 3 Data Cleaning

Each aspect of the project requires data cleaning in its own unique way. One thing that had to be dealt with across all skews was the treatment of NA values. One column was 100% NA values, that being Dawson Street 2020. This was with mostly through the use of na.omit, and na.rm. In some cases, such as in the predictive they had to converted to zeros as some of the methods used do not accept NA values. These values converted to zero were only done so after any arithmetic was applied.

## 4 Analysis Skews

### 4.1 Microscopic View

### 4.1.1 Data Analysis

2019 data is measured in 15-minute intervals, as opposed to 1-hour intervals in 2020 data. To be comparable, 2019 data must be reduced to same interval length.

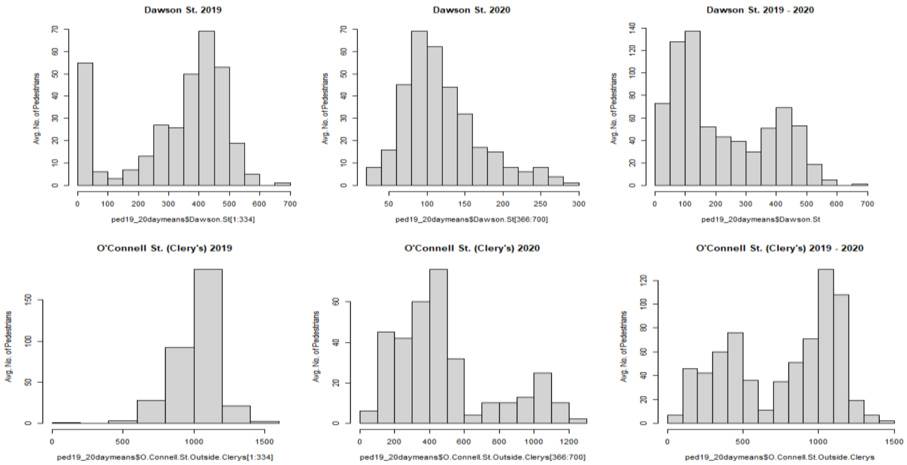
2019 data now reduced to hourly intervals, with the appropriate rows of the date/time column reattached. Next, the two datasets were merged into one data frame. This required changing some column names. It was also noted that two entries appeared to be missing (31st March 2019, at 1:00 am and 29th March 2020 at 1:00 am). This was established to be due to the switch over to Summertime when the clocks were shifted forward by one hour. Empty rows of data were created and appended to the data frame to account for this. The timestamp column was updated and subsequently altered to a POSIXlt format, with the time zone set to UTC, which ignores Daylight Savings.

Dawson.Street and Dawson.Street.Replacement were combined into one column.

The dataset was merged and scaled up to show counts for full days. The dataframe produced contains counts for each day, with only dates in the first column. A function was created to calculate the mean count for each day.

#### 4.1.2 Results

These graphs are simple comparisons of the daily average footfall on certain streets between January and November in 2019 and 2020 (December 2019 ommitted for fair comparison with 2020 data), as well as a graph of both years combined. Each street shows a clear change in the numbers of people, and in the patterns of footfall. There is some degree of normality to the data, and the two-year graphs clearly show distinct bimodality in the data for many streets. Highlighting a change in how pedestrians move around the city.



### 4.2 Macroscopic View

#### *4.2.1.1 Introduction.*

These locations are what might be referred to as the extreme centre of Dublin. They feature the principal thoroughfares of the city (Grafton St, O'Connell St, Henry St) and the nexus of connecting and feeding streets in their vicinity.

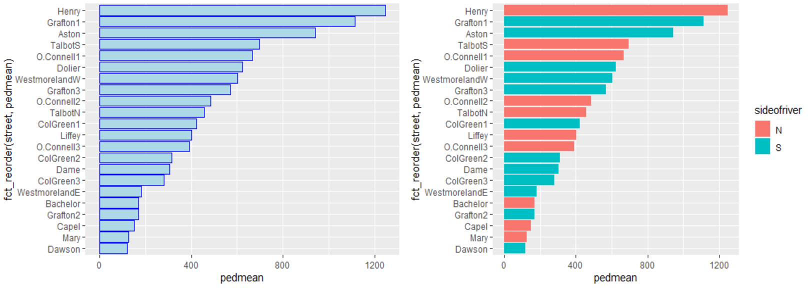
We will begin by examining each of these streets on an individual level to gain insight into their relative level of footfall. We will then change approach and focus instead on the broad totals of population information and discuss how this changed based on the time of day, or month of the year. As there was such a dramatic occurrence in March of this year, we will then turn to data from 2019 to derive further insight.

#### 4.2.2 Data Analysis

#### *4.2.2.1 Streets*

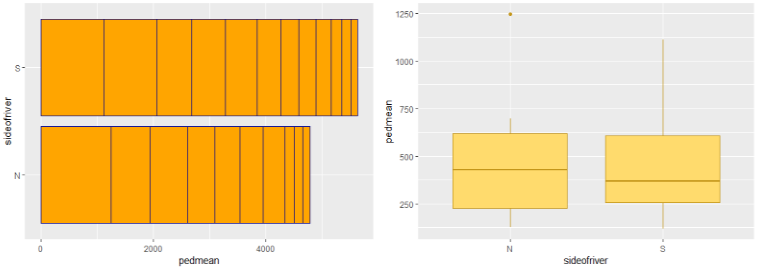
We start by examining each street on an individual basis.

The following bar-chart shows the average hourly observations on each street throughout the entire time period. We can see that Henry Street is the busiest, followed by one of the Grafton Street observations, all the way to Dawson Street which has the lowest observed pedestrian traffic.



We can seek a further insight by dividing these streets into those lying north of the River Liffey and those lying to the south. The previous plot shows that streets on either side of the river are quite evenly distributed with 5 from each side featuring in the busiest 10 streets.

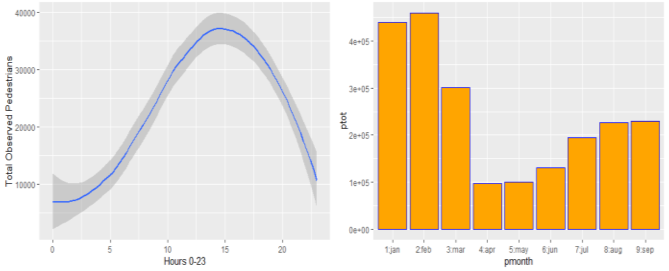
The following box plot shows us that the median observation from a location on the north side is larger than that found on the other side of the river, but not by much.



#### 4.2.2.2 *Hour & Month*

Let us start by examining the hours of the day on Dublin's streets. The following plot shows recorded footfall per hour from all the observation points combined, totalled across the year. We can see that Dublin pedestrian traffic on average throughout the year peaks at 13:30.

We will now perform an analysis on a month-by-month basis. The results on the following plot are quite striking. We see one level of population occurring at the beginning of the year which suddenly drops in March. By April, it is at its minimum and then recovers later in the year. This, of course, is explained by the government-mandated "lockdown' which manifested itself in the form of a series of stringent restrictions on people's movements. This had a very clear effect on the observed footfall readings. We then see these figures recover somewhat as restrictions were eased.



## 4.3 Predictive Powers

#### 4.3.1.1 *Introduction to Predictive Modeling*

The scale of the data is quite vast, with over 7000 rows of data across over 20 streets there is a wide range of data to be experimented with. Given this it was decided a predictive model would be appropriate to gleam some insight from the data. Our model aims to predict the busiest street in Dublin at a given date/time.

The model decided upon is a "Random Forest" algorithm. This is ideal in classification problems here we are classifying which street will be the busiest. This algorithm works on the concept of decision trees. A decision tree is a branching path, like an if else statement. For example, our decision tree may take the form, if it is later than 10pm, check the day of of the week, if the day of the week is Monday check the month, if it is November or December the busiest street will be "Dawson Street". One decision tree is not enough to be a truly robust classification tool; hence we build a forest. We generate many trees, in our case 1000. Each tree is unique and will assign different levels of importance to the various parameters fed into it. Once all 1000 trees have come up with an answer, whichever answer is the most frequent will be chosen as the answer. Therefore if 600 trees say, "Dawson Street" will be busiest, and 400 say "Capel Street" is busiest we will go with "Dawson Street".

In this case 2020 data will be used to generate the model. This will then be applied to 2019 data which one could count as a "typical year". If the 2020 model is capable of being applied to 2019 with a high degree of accuracy, we can conclude pedestrians in Dublin behave similarly in 2020 and 2019 despite the pandemic. While there are less pedestrians on the streets in 2020, we hope to show their overall patterns are similar.

#### *4.3.1.2 Feature Engineering*

The data supplied features very few features besides that of raw street data. 4 test features were designed, and one target feature.

- **hour:** Integer representing the hour of the day, 0:23.
- **month:** Integer representation of the current month, 1:12
- **colMin:** Index of the street that had the least people at the corresponding time.
- **dayOfWeek:** String, the day of the week Monday, Tuesday etc.
- **colMax:** Index of the street with the most people at the corresponding time. This is the target.

These parameters while somewhat limited are enough to build a model. In the future one may take additional parameters into consideration from external datasets to improve the accuracy of the model, for example using a weather dataset as a potential pull/push factor for pedestrians.

#### *4.3.1.3 Result*

The random forest model was applied to each element of 2019, and busiest street was predicted. The output can be seen in the 2019predictions.csv. The 2020 model had an estimated error of 36.53%. For the 2020 model to be considered applicable in a general use case we hope that there would be a similar error rate when the predictions were made. Upon comparing the 2 years there was a rather pitiful 6% match, or over 94% error. This clearly exemplifies the abnormal nature of 2020. There are countless factors that contribute to this. COVID-19 led to work from home restrictions thereby making business districts less busy, furthermore many retail outlets were closed meaning pedestrian traffic in shopping districts would fall.

#### *4.3.1.4 Conclusion*

All this leads to the conclusion that 2020 is indeed an outlier year when modeling pedestrian traffic. The model built was simply unable to predict the footfall for more normal year. Despite that this model may not be entirely useless. This style of analysis may prove useful if Ireland is to go into lockdown again in 2021 if COVID-19 persists due to slow roll-out of vaccines. In the future we would like to test this model on 2021/2022 data and see if there are any long-term effects on pedestrian traffic thanks to companies post COVID-19 offering long term work from home schemes, and online grocery shopping becoming more widely used.

## 5 UNFINISHED BUSINESS

There were several ideas that started development but were unfortunately remove due to time constraints. Firstly was an animated heatmap of Dublin. This was planned to be an eyecatching piece that used the 'leaflet' library and 'shiny' to create a map that changed colour with respect to time, and the number of people on a given street at that time.

Another idea that had to be removed was to make inferences on the bulk movement of people using the data. Unfortunately this proved beyond the scope of the project given the data.

Finally for the predictive model we hoped to use weather, and bus schedule data to predict the busiest street. However this felt beyond the scope of the project due to the time allotted.

Rough/Unfinished code for these attempts can be supplied as required.