Declaration on Plagiarism

This form must be filled in and completed by the student submitting an assignment

Name/s:	Ishrat Fatima Syed and Pooja Balloli
Student Number/s:	Ishrat- 22264608 and Pooja- 22261460
Programme:	Msc in Computing (AI) and Msc in Computing (DA)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	06-12-2022
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml,
https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name:	ISHRAT FATIMA SYED	Date: 06-12-2022
Name:	POOJA BALLOLI	Date: 06-12-2022

Data Visualisation

Analysis of Footfall across the busy streets of Dublin over last 7 years

Abstract:

Footfall data, also known as mobility data is a source of information to analyse how people engage with Point of Interests in the city. Footfall Data is vital for city planners as it helps to measure human activity at different locations as well as the changing demographics within the city. It also holds particular significance for the retail stores as they would like to know the busiest hours on a street so that they can ensure more employees are working on the sales floor at that time. In this report, we investigate and analyse the footfall data from more than 100 locations across Dublin collected across the last 7 years to identify hotspots in Dublin and how the footfall has changed over the years and in different weather conditions. The visualisation highlighted that Henry Street, O'connell Street and Grafton Street have consistently attracted the highest number of Dubliners over the last 7 years. Furthermore, we also investigated the impact Covid had on the footfall in Dublin and it came in as no surprise that the Covid restrictions in 2020 and 2021 caused a drop in footfall across all of the streets in dublin. Finally, we also investigated how resilient Dubliners are to the changes in weather conditions and found that footfall increases with rise in temperature which shows that Dubliners enjoy the rare warm sunny weather. It was interesting to see that rain does not stop dubliners from heading out.

Data Collection:

Two datasets have been used in the Visualisation. One dataset consists of footfall data from the year 2016 to 2022 and the second dataset consists of weather data. The footfall dataset was provided by https://data.smartdublin.ie/dataset/dublin-city-centre-footfall-counters in csv files. Separate csv data file was downloaded for all the years. Weather dataset was downloaded from https://www.met.ie/climate/available-data/historical-data .

The data exhibits collected exhibits the characteristics of big data in the following ways.

It has velocity in the sense that the data is at hourly granularity, meaning the data is updated every hour.

The footfall dataset has more than 130 columns initially and over 7000 rows in each dataset file from the year 2016 to 2022. The weather data has around 168408 rows and 11 columns.

Data Exploration, Processing, Cleaning and/or Integration

Data Exploration

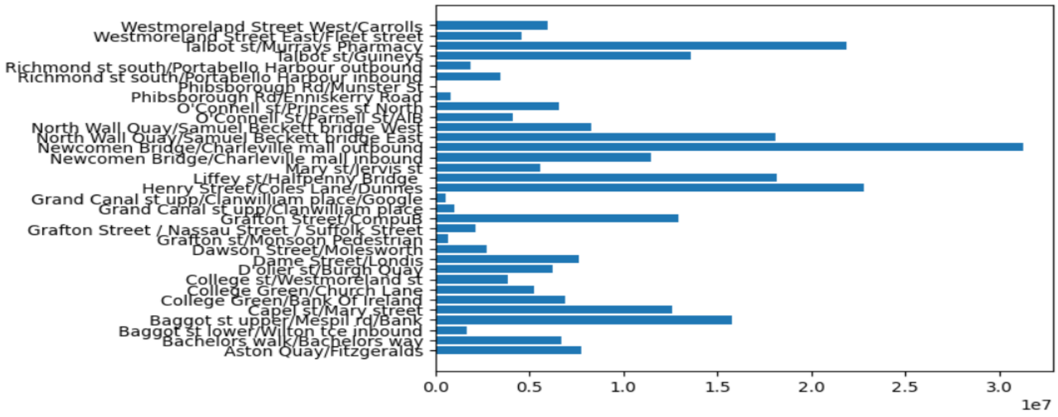
We started out by working out on one of the files, the recent one - footfall files for the year 2022. To explore the data see how spread the data was and what sort of cleaning would be required. The first file we explored had three of the columns with no values at all - These were dropped.

Each location was having location, location IN and location out, where in location = locationIN + location OUT - The IN and OUT columns were dropped as we had the consolidated footfall in the respective location column.

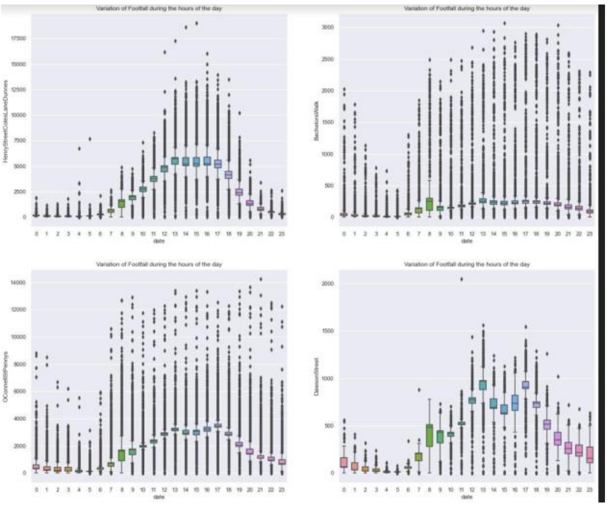
And the default index column added was dropped and instead the first column the “Date” column was set as index.

The initial file explored did have a lot of missing data, resulting in NaN values, which were kept to see if we can map some pattern with the missing values.

What we mapped out initially looked something like this, which gave us a mapping on which place is most crowded, which is moderately crowded and which is least crowded.

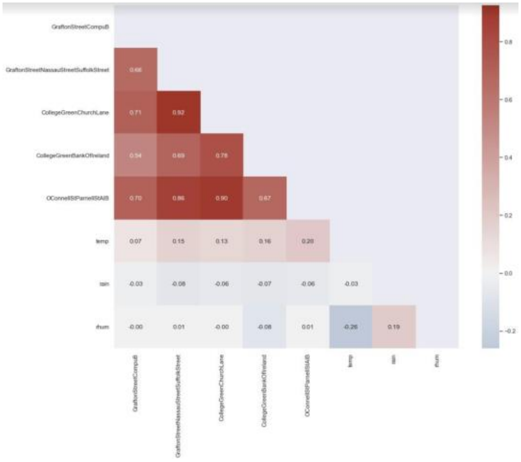


This gave us an insight into the data, and we planned to merge the files for 7 years from 2016 - 2022 and see how footfall is changing over the period of time at these places. and proceed with visualisation.



We also used box plot for visualizing how the footfall changed over the course of the day.

We Observed Bachelor's walk is the most crowded at 8 am. O'Collonel street is the busiest at 4pm and 5 pm.



We also tried to explore the data by creating a correlation matrix to find out any possible correlations of footfall with weather data. We found that footfall traffic increases when the temperature rises, indicating that Dubliners like the infrequently warm and sunny weather. It was amazing to observe that Dubliners go outside even in the rain.

Data Cleaning and Processing

What did you need to do to prepare the dataset(s) to create your graph/chart?

After collecting the data, Data cleaning and preprocessing had to be done so that the data could be used for our visualisations.

Once the files were merged - For data cleaning, firstly we had to rename the date and time column for each of the footfall datasets from the year 2016 to 2022 to one single name 'date' as all of them had different names.

The data types of the data time columns were from float to datetime format to make use of datetime object from the Pandas library.

There were a lot of missing values (NAN Values) in the dataset. We have filled in the missing values using data imputation. We have imputed missing values with the mean values of the same day and same hour of the respective column. We also checked and removed all those columns that did not have at least 30% of the data since it made no sense to impute 70% data in them.

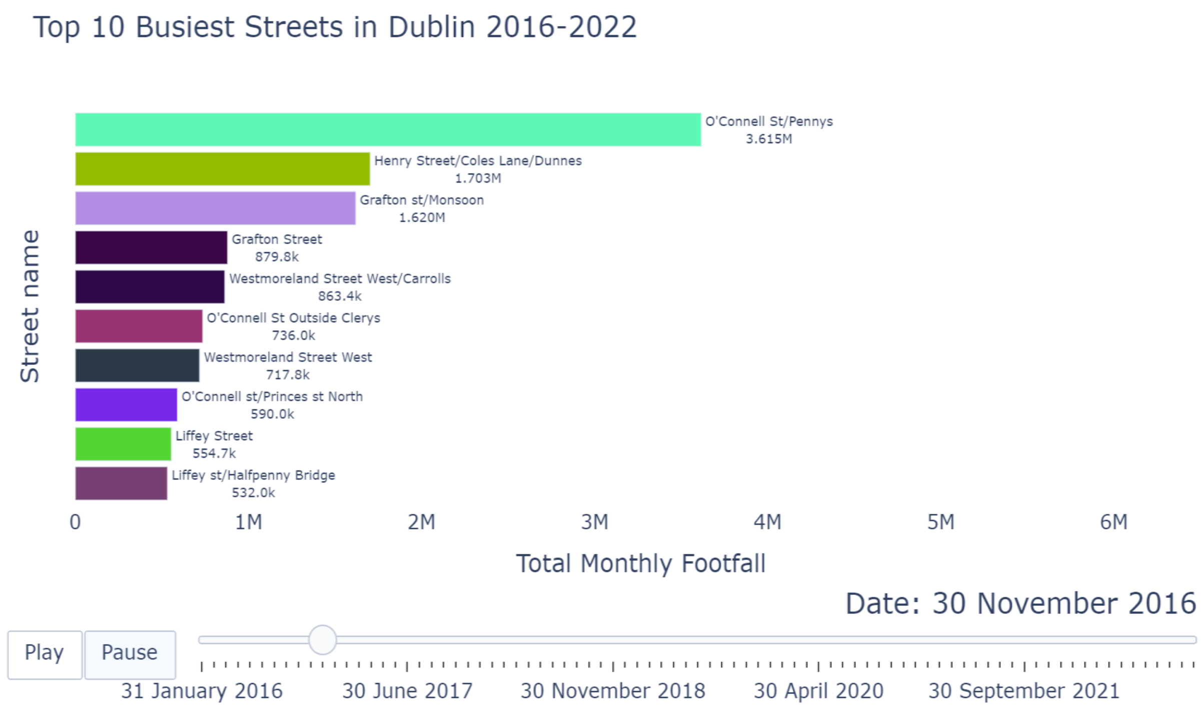
After all the pre-processing steps one single file was generated with 86252 rows and 33 columns.

Data Integration

We have integrated the preprocessed footfall dataset with the weather dataset to create a visualisation that would show correlation between weather and footfall data.

Visualisation

In the **first visualisation**, an animated chart is used to visualise the Top 10 Busiest Streets in Dublin from 2016-2022.



Choice of chart types:

We wanted to visualise the top 10 busiest streets in Dublin from the year 2016 to 2022. The street names are nominal variables, footfall is numerical and the date is a discrete variable. The horizontal bars allow us to see the comparison of footfall between the busy streets. The chart shows the top 10 rankings in order from top to bottom.

Design choices:

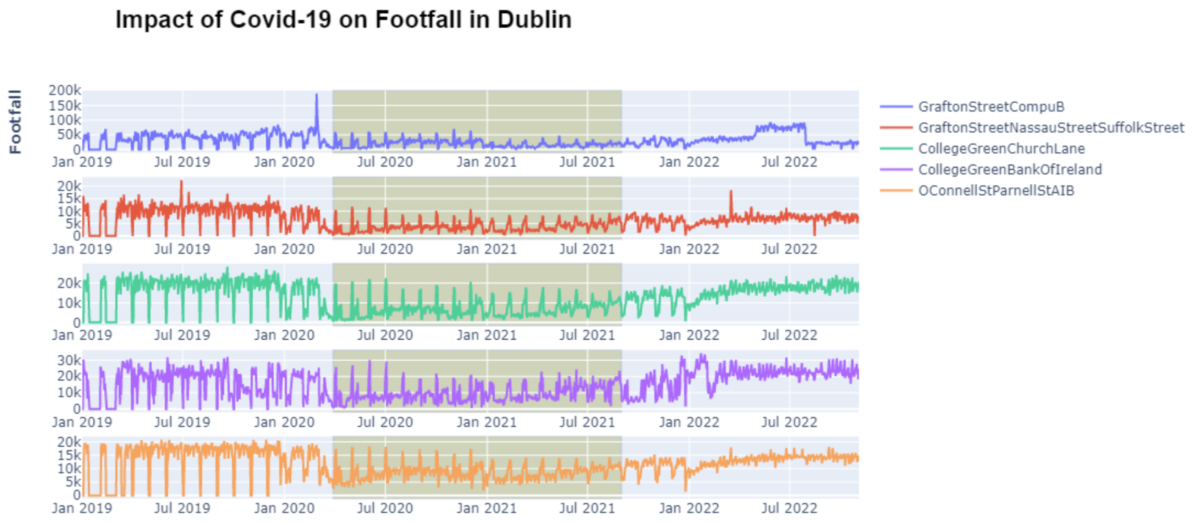
We referred to https://pypi.org/project/raceplotly/ to build our animated visualisation.

- Bar chart animation is used to visualise the change in footfall over the years as it is simple and easy to understand and represents the required information correctly.
- The street names and footfall value is shown outside the bars which improves the readability of the graph.
- The date time format was changed from yyyy-mm-dd to a more easily readable word format
- frame_duration was adjusted so that the animation improves the visualisation of variation footfall data over the years.
- The Orientation of the graph was kept horizontal as the visualisation is moving forward in time and showing the footfall variation over the years
- Default colours have been used in the visualisation and they do not convey any extra information in our dataset.

Animation Choice:

The dataset we had was from the year 2016 to 2022. To show the variation of footfall over the years an animated bar chart is used. By using a moving time frame we are able to show the reader which street had the highest footfall in every single month starting January 2016 and finishing in November 2022.

In the **Second Visualisation** we tried visualising the Impact of COVID-19 on the Footfall in Dublin. It depicts how the daily footfall across Dublin changed during the COVID restrictions and national Lockdown. We divided the data in three periods - Pre-covid, amidst covid restrictions and post-covid and plotted them as timeseries to visually observe the changes in the footfall.



Choice of Chart types:

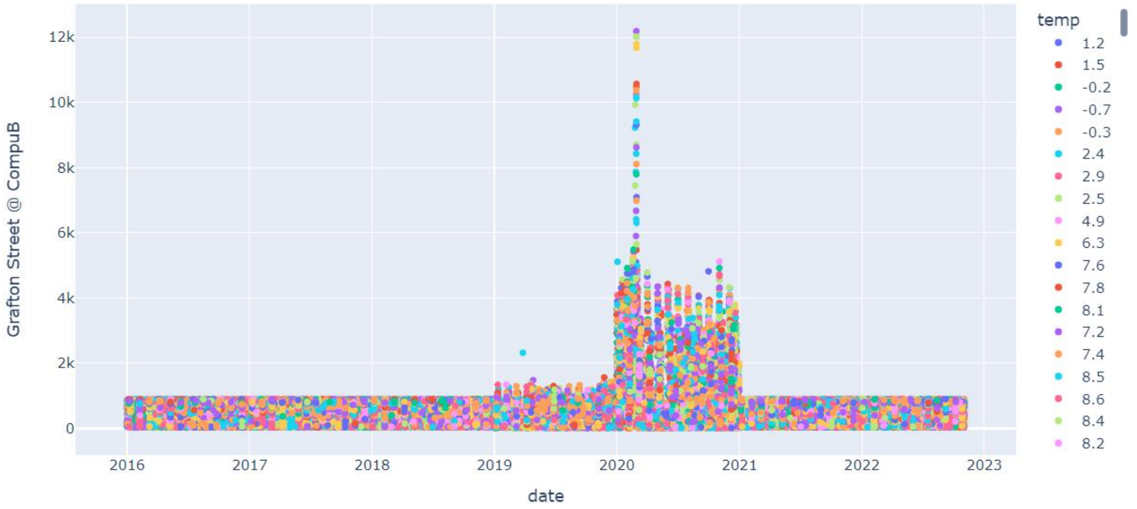
Footfall count is continuous numerical variable and the dates are discrete therefore, Line chart was an obvious choice as we were looking at time series data at daily granularity. We used subplot to show the footfall across multiple streets and we plot them as timeseries in 5 subplot. We used 5 rows since only 5 streets had significant data for all the years that we were analysing. We have different colours for every street and have highlighted the period with Covid restrictions in Olive colour for easier comparison.

Design choices:

- We have used line charts as it is a simple way of visualising timeseries.
- Subplots were used to show that the drop in trend was not restricted to a particular street during the national lockdown
- We have used the colors to distinguish the footfall of one street from the other. The actual color assigned to the street does not hold any significance
- The Period when Ireland was under Lockdown has been highlighted in Olive color to attract users attention
- The chart is interactive and the user can use mouse to zoom in on a specific day or date

In the **third visualisation** we planned on mapping footfall of a place that had significant data i,e - Grafton Street @ CompuB and see how it is varying with temperature

Footfall of a place mapped with temperature



Choice of chart type:

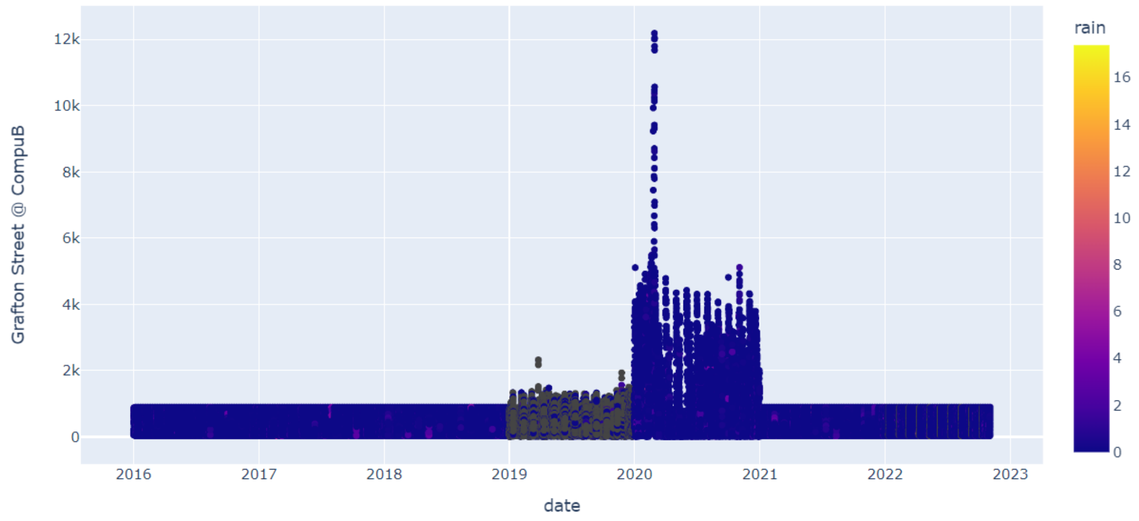
We were seeing the footfall data over the span of 7 years for a specific place, with temperature in play, as we were pairing two specific parameters: Footfall and Date scatter plot seemed like a better choice. We also wanted to relate temperature with this graph so colour was used with the scatter plot, with each colour identifying a specific temperature.

Design Choices:

- Scatter plot is helpful when it comes to visualising relation between two parameters
- As we wanted to see the variation of another parameter "temp" in relation to the ones mapped on x and y axis colours were used.
- Colours used for temperatures are random and just used to uniquely identify a specific temp.
- Indication of which colour represents what temperature is mapped along with the graph.
- The street was particularly very busy on Feb 28, 2020, 15:00 , when the temperature was around 12.9 degrees.
- The chart is interactive for the user to zoom in on a specific area and see more details.

Since we saw the temp variations. In the **fourth visualisation,** we planned on mapping footfall of a place that had significant data i,e - Grafton Street @ CompuB and see how it is varying with rain in picture.

Footfall of a place mapped with rain



Choice of chart type:

Similar to the third visualisation, we were seeing the footfall data over the span of 7 years for a specific place, with rain in play, as we were pairing two specific parameters: Footfall and Date scatter plot seemed like a better choice. We also wanted to relate rain with this graph so colour was used with a scatter plot, with each colour identifying a specific temperature.

Design Choices:

- Scatter plot is helpful when it comes to visualising relation between two parameters

- As we wanted to see the variation of another parameter "rain" in relation to the ones mapped on x and y axis colours were used.
- Colours used to plot rain are random and just used to uniquely identify specific range values pertaining to rain.
- The last scatter plot used colours, temperature values were mentioned by the side of the graph one by one, for this graph we thought, instead of mentioning colours individually, we will show it as range.
- Indication of which colour represents what value for rain is mapped along with the graph.
- The street was particularly very busy on Feb 28, 2020, 15:00 , the rain value for the day shows as 0.1 .
- The chart is interactive for the user to zoom in on a specific area and see more details.

Tools and libraries used

- Jupyter notebook was used to code in Python and create visualisations.
- Python was used to do all the cleaning and processing.
- Pandas and numpy libraries were used in the data cleaning, preprocessing and transformation.
- Python, plotly and raceplotly were used to make the animation and the bar chart.

Conclusion:

We started the analysis with an open mind. We wanted to see how the Footfall has changed over the years across Dublin and how other external events (Pandemic, Temperature, Rain etc) affect the footfall. In order to answer these questions we collected the data for over 103 counters that report data at hourly granularity. We also augmented the data with the weather data for Dublin.

For the Top 10 Busiest Streets in Dublin, we wanted to visualise what places in Dublin recorded maximum monthly footfall starting January 2016 so we decided to make a Bar racechart to highlight this. The visualisation showed that Henry Street, O’connell Street and Grafton Street have consistently attracted the highest number of Dubliners over the last 7 years. We wanted to color code every street to a specific color so that reader could visually track its movement but the time limited us in doing so. we also wanted to add interactive filters to the animation so that reader could select top N streets instead of 10 or select the busiest streets in North Dublin.

For the Second visualisation, we wanted to see how Covid affected the footfall across Dublin. The Interactive visualisation clearly indicates that the footfall has reduced during the phase when Ireland was in lockdown. We visually depicted this by shading the region between 29th March 2020 and 31st August 2021 as Ireland was in Lockdown during this phase. Moreover, We also observe an anomaly on 30th July 2022 when the footfall for GraftonStreetCompuB counter sharply dropped. This warrants further investigation. We wanted to add annotations to the shaded region in Olive but the plotly version kept throwing errors when used with subplot. The annotations would make the user aware of what the shaded region highlights. Furthermore, we could add filtering capability so that the user could pick a street of his or her choice and observe Covid's impact on that street's footfall.

For the third Visualisation we wanted to see temperature in play with footfall for one of the places which had significant data. From the graph we could see that the footfall was around the range 800-1000 from 2016-2019. And started seeing slight increases from 2019 with the highest spike recorded during 2020.Then we see a drop and we see footfall falling in the same range i.e. around 800-1000 from 2021 to 2023. Particularly during the month of feb, as the first five highest footfall recordings fell in the same month

First - Feb 28, 2020 15:00 footfall - 12188 temp 12.9, Second - Feb 29, 2022 09:00 footfall - 12043 temp 3,Third - feb 29, 2020 06:00 footfall - 12016 temp 2.4,Fourth - Feb 28 , 2020 12:00 footfall 12014 temp 12.5,Fifth - Feb 29, 2020, 08:00 footfall 11794 temp 2.1. On the busiest days. Feb 28 - temp was 4.3 to 13.7 and on Feb 29 - temp was 2.1 to 10.9 degrees. As the graph is interactive, it can be zoomed in to see what the temperature range was for a particular day to infer the footfall recordings and temp. Highest temperature recorded was 8.5 which was a no of days, selecting this temp from graph and noticing footfall should give us an idea on footfall and temp relation.

For the fourth visualisation, we mapped the footfall with rain

While the range of footfall from 2016-2019, the spike and dip in numbers, and the busiest days remained the same, what rain values looked like for the busiest day were as follow

Feb 28 was between 0 - 0.9 and Feb 29 was between 0- 1.5. The interactive graph could be zoomed in to see the relation between rain and footfall for Grafton Street for any given day.

Contributions and Working as a pair:

The biggest challenge we came across was cleaning and pre-processing the data as the datasets were downloaded from the government's website in a CSV form. There were a large number of missing values, the data for all years were in separate CSV files, the column names had special characters which made it tricky to work with, the data type of columns were all objects even for numbers. Pooja worked on exploring the data with regards to the recent footfall file i.e FootFall file for the year 2022. Where necessary cleaning required to visualise the data like dropping columns with no values, dropping redundant columns, dropping default index column and setting it to date column were handled. In the file used for exploration the NaN values were kept with the intention to see if it can be used for some analysis like checking for emerging patterns for missing values. This analysis could not be taken up due to time crunch. Merged file was handled by Ishrat. Collection, cleaning, preprocessing and integration of the data were all done by Ishrat Syed.

The workload for data exploration was shared, Ishrat worked on the Correlation matrix for finding the correlation between weather and footfall, and the box plot for visualizing how

the footfall changed over the course of the day. While Pooja worked on building a visualisation of the top 10 busy streets in 2022 .
For the visualisations, we decided that we will split a total of 4 visualisations wherein Ishrat generated the first two visualisation which are, The animated raceplot for Top 10 busiest streets and Interactive plotly chart for analysing the impact of Covid 19. While Pooja worked on the third and the fourth interactive graphs mapping one of the busiest streets and seeing the footfall variation in relation with temperature and rain.

Finally, we both worked on the report and explained the findings from the visualisation we had generated.

References:

1. https://data.smartdublin.ie/dataset/dublin-city-centre-footfall-counters
2. https://www.met.ie/climate/available-data/historical-data
3. https://pypi.org/project/raceplotly/
4. https://visme.co/blog/scatter-plot/#:~:text=Use%20a%20scatter%20plot%20when%20you%20have%20two%20variables%20that,a%20positive%20or%20negative%20correlation.
5. https://chartio.com/learn/charts/what-is-a-scatter-plot/
6. https://plotly.com/python/line-and-scatter/
7. https://plotly.com/python-api-reference/generated/plotly.express.scatter.html