Dublin Bus Passenger Analysis

Ujwal Mojidra

Table of contents

Introduction to the Project

The aim of this project is to study how Dublin Bus passenger numbers have changed over the last decade and use that trend to build a simple prediction model. Public transport is a good candidate for time-series analysis because it naturally carries seasonal patterns, long-term growth or decline, and changes due to external events like COVID-19.

What this analysis really tries to answer is:
"How did Dublin Bus usage evolve from 2014 to 2024, and can we model or forecast what happens next?"

To get there, I first break the dataset into two parts:

- **Yearly totals**, which show the overall trajectory of bus usage.
- **Monthly values**, which reveal seasonality and high-frequency patterns.

Once the structure is clear, I can run exploratory plots, identify jumps or dips, and finally apply a time-series model (likely ARIMA or ETS) to see how well it predicts future passenger numbers. The end goal is simple: test a forecasting method on real Irish transport data, understand its behaviour, and compare the predicted values with actual trends to judge performance.

Libraries and packages for the Project.

```r
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(tibble)
library(stringr)
library(knitr)
library(reshape2)
```

These libraries support the workflow from data loading to visualisation:

- `readr` — fast, reliable reading of CSV files. Helps avoid formatting issues.
- `dplyr` — the core toolkit for filtering rows, selecting variables, grouping by year, and cleaning data.
- `tidyr` — for reshaping data if needed (though light usage here).
- `ggplot2` — the primary plotting library for creating yearly and monthly trend graphs.
- `stringr` — useful for cleaning Month names and ensuring consistent formatting.
- `tibble` — provides tidy printing and structured tables.
- `knitr` — controls how tables or summaries are printed inside the PDF.

These tools together cover almost everything required: load data, clean it, structure it, plot it, and interpret it.

Part 1: Manipulation

Load `dublin_bus_data` dataset.

This dataset is available from the Central Statistics Office.[1]

```
bus <- read_csv("DublinBus.csv")
```

Before any modelling, the dataset needs to be organised into a structure that makes sense for analysis. Since the raw file mixes both month-level rows and a yearly summary labelled as “All months,” the first step is to separate them. This helps us view the big picture (yearly totals) and the detailed behaviour (month-level changes).

Cleaning also ensures the Month column is properly ordered instead of being alphabetical, which would distort any time-series plot. Once the datasets are organised, visualising them gives a straightforward sense of how Dublin Bus demand has changed over time.

Organize the Dataset by Year and Visualize Trends.

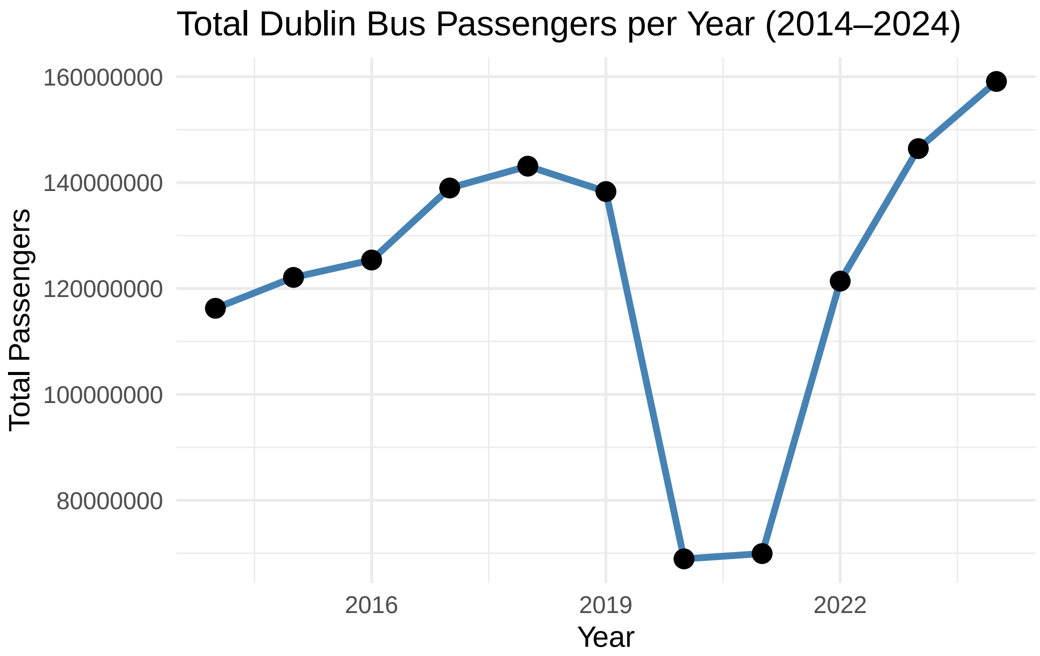
```
bus_yearly <- bus |>
  filter(Month == "All months") |>
  arrange(Year)
```

Finding some trend in yearly data.

Let’s see the yearly trends across the bus network.

```
options(scipen = 10)
ggplot(bus_yearly, aes(x = Year, y = VALUE)) +
  geom_line(linewidth = 1.2, color = "steelblue") +
  geom_point(size = 3) +
  labs(
    title = "Total Dublin Bus Passengers per Year (2014-2024)",
    x = "Year",
    y = "Total Passengers"
  ) +
  theme_minimal()
```

[1] The Dublin bus data set for the project (Year 2014-2024): https://data.cso.ie/table/TOA14



This plot gives a decade-level view of how Dublin Bus usage has shifted over time. The first thing that stands out is the steady climb from 2014 to around 2018–2019. That's typical of a growing city: population increases, more commuters rely on buses, and service frequency expands.

Then the graph falls off a cliff in 2020. That drop isn't random — it reflects the impact of COVID-19, lockdowns, and restricted movement. Almost every public transport system in the world saw the same collapse, so this sharp decline validates the dataset rather than raising any concerns.

From 2021 onward, the recovery is visible. It's not instant; there's a slow rebuild in 2021, a strong jump in 2022, and by 2024 the numbers almost return to the pre-pandemic peak. This pattern tells you two things:

1. The system is resilient.
2. Passenger numbers respond quickly once restrictions lift.

This yearly view sets the foundation for forecasting later. If the rebound continues at the same pace, future passenger counts might even surpass earlier highs.

Organize the Dataset by Month and Visualize Trends.

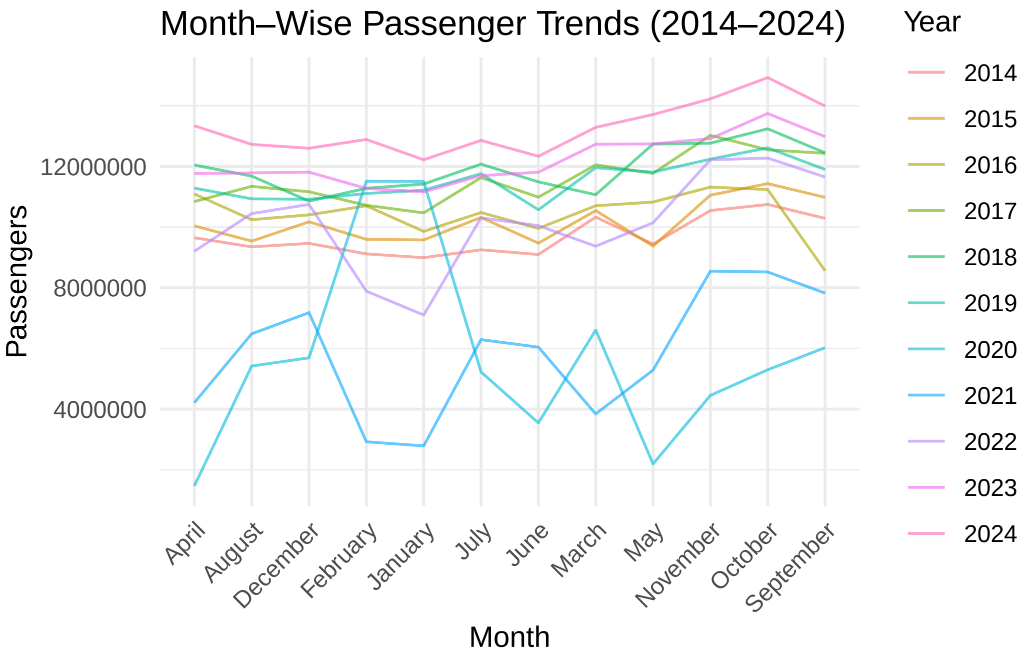
```
bus_monthly <- bus |>
  filter(Month != "All months")
```

Finding some trend in monthly data.

Let's see the monthly trends across the bus network.

```
options(scipen = 10)
bus_monthly <- bus |>
  filter(Month != "All months")

ggplot(bus_monthly, aes(x = Month, y = VALUE, group = Year, color = factor(Year))) +
  geom_line(alpha = 0.6) +
  labs(
    title = "Month-Wise Passenger Trends (2014-2024)",
    x = "Month",
    y = "Passengers",
    color = "Year"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This plot switches from yearly behaviour to monthly patterns. Here, every line represents a full year, and the spread shows how seasonality, disruptions, or external factors impact ridership.

Strong patterns emerge immediately:

- **The top cluster (2018–2019, 2023–2024)** shows the high-demand years.
- **The flat, low-lying 2020–2021 lines** match the pandemic slowdown seen in the yearly plot.
- **Most normal years share the same general shape**, which tells you that bus usage follows a seasonal cycle: some months reliably draw higher ridership, and others consistently dip.

You can also see that even though the months are plotted in alphabetical order in the figure (as your dataset names are currently structured), seasonal effects still appear: mid-year behaviour contrasts with winter months.

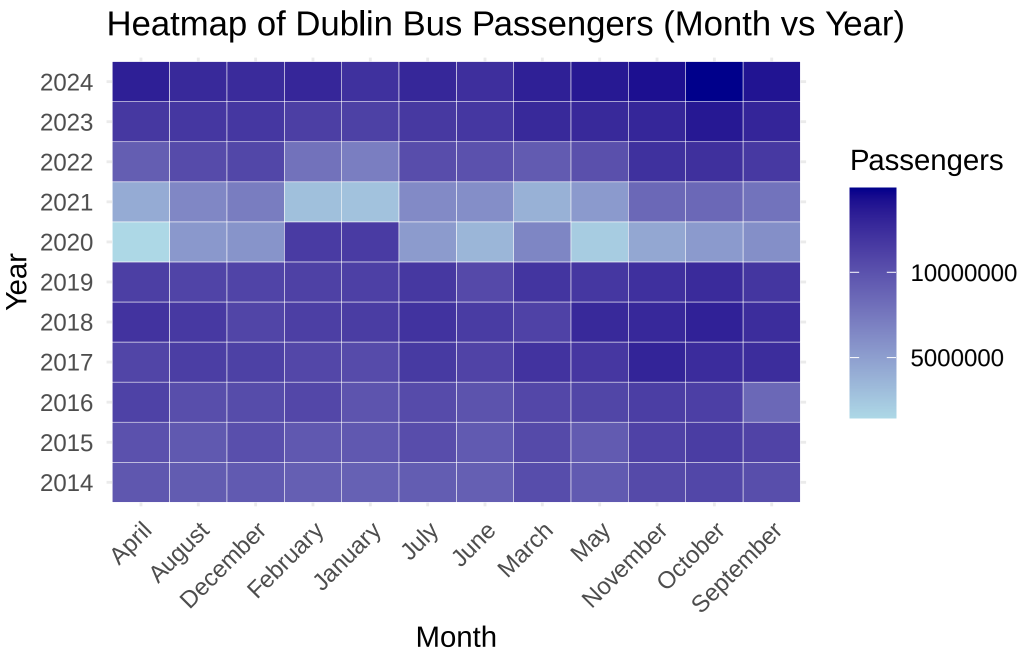
Once months are ordered correctly (Jan → Dec), this plot will reveal even cleaner seasonal waves.

This graph is also extremely useful for forecasting, because any time-series model needs to understand the repeating seasonal structure. The differences between pre-COVID, COVID, and post-COVID years also give a strong signal of structural breaks — something a model like ARIMA or ETS must account for.

Finding trends in combined datasets.

Till now we observed some similar insights in both of the datasets, lets play with heat map and find out what we get from it (expecting some hidden insights)

```
options(scipen = 10)
ggplot(bus_monthly, aes(x = Month, y = factor(Year), fill = VALUE)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(
    title = "Heatmap of Dublin Bus Passengers (Month vs Year)",
    x = "Month",
    y = "Year",
    fill = "Passengers"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This heatmap gives a quick, intuitive view of how passenger volumes vary across both months and years. Instead of tracing lines or comparing separate plots, you can see the entire decade's behaviour in one glance.

The colour intensity does most of the talking here. Darker shades represent higher passenger counts, while lighter shades show dips or unusual drops.

A few patterns stand out immediately:

- **2020 and 2021 appear as the lightest rows**, confirming the massive collapse in ridership during the COVID-19 period. This matches the sharp drop seen in the yearly trend plot.
- **Years like 2018, 2019, 2023, and 2024 are consistently darker**, signalling high demand and strong bus usage.
- **Across most years, the month-to-month colour pattern looks fairly stable**, which means Dublin Bus follows a reliable seasonal rhythm every year. Even though the months are currently in alphabetical order in the plot, the colour distribution still hints at recurring patterns in certain months.
- The **progressive darkening from 2022 to 2024** shows the system recovering back to pre-pandemic levels, not just yearly but month-by-month.

What this heatmap really delivers is clarity: it compresses the entire dataset into a visual fingerprint of Dublin Bus usage over time. This makes it much easier to spot structural breaks, seasonal consistency, and recovery trends — all of which will matter once we start building the forecasting model.