# Introduction

Aviation is a growing necessity in keeping our modern world connected. At any moment, hundreds of planes are in the air, and millions of passengers are transiting every month.

Australia has a strong reliance on the aviation industry due to our isolated geography – with no land borders, air travel is key for international passengers to travel in and out of the country. This report will analyse and visualise aviation passenger data within Australia for the past few years, covering both international and domestic flights.

By analysing these trends, I hope to see how the industry has changed and where it is headed in the future; especially after these uncertain times. I also hope to examine how major international hubs connected to Australia have changed.

# Data Source

## Origins

The data sources used in this project are publicly provided by the Bureau of Infrastructure, Transport, and Regional Economics (BITRE); a division of the Department of Infrastructure, Transport, Regional Development and Communications.

BITRE publishes various transportation statistics, including a wide variety of Aviation statistics. This report focuses on Domestic aviation activity and International airline activity. Time series data is provided, with monthly information for routes provided as far back as 1984.

One caveat with this dataset is that reporting of data by Qantas Airways changed in 2003 – for example, a flight reported as Adelaide to London in January 2002 (no direct services between these two cities), would be reported in January 2003 as either Adelaide to Singapore or Melbourne/Sydney to London. This makes it difficult to directly compare data before 2003 to data after 2003. To deal with this in my visualizations, most of my visualizations will only use data from 2003 to 2019 – if required, data before 2003 will not be compared to data after 2003.

## Processing

There was no element of data collection or web scraping in this project.

BITRE provided the data in multiple Excel spreadsheets – for example, the international data was split into 1985-1998, 1989-1993, 1994-1998, 1999-2003, 2004-2008, and 2009-2020.

To combine these spreadsheets together, the Python package **pandas** was used. Multiple spreadsheets are loaded at once using the **read\_excel** function, and then the data frames are appended together to create one large data frame.

Since the data is inherently multi-dimensional (Origin, Destination, Time); I also used the package **xarray** to easily work with the dimensions. Xarray provides a powerful n-dimensional data structure with dimensional labelling. These structures are much easier to work with for analysis and visualization purposes.

# Results

## Preliminary Analysis

Since this is a large dataset with a wide variety of origins and destinations, it’s important to get a look at the bigger picture before jumping in too deep.

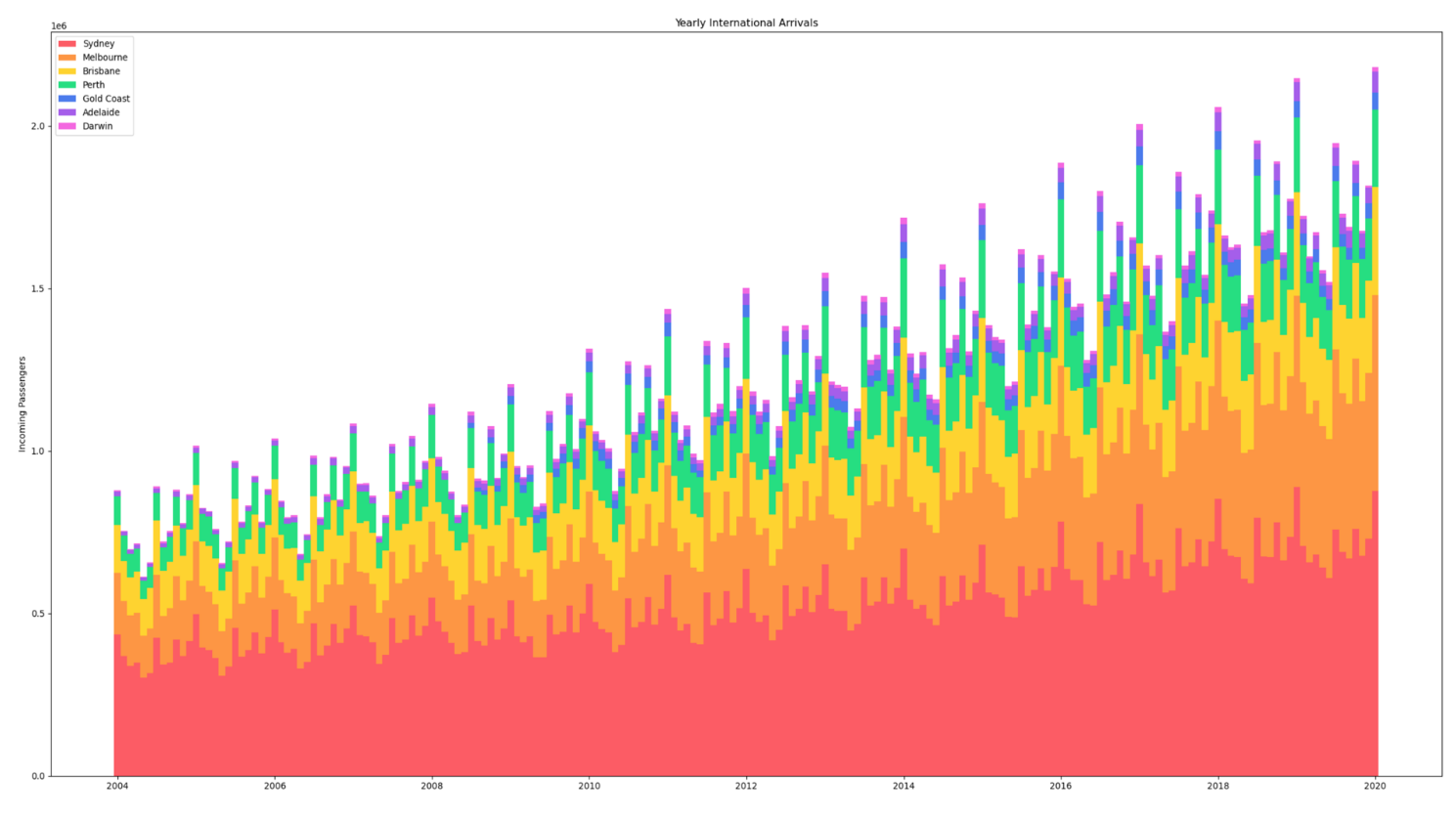


Figure 1: Stacked bar chart, showing monthly international arrivals in stacked barchart form.

This stacked bar chart shows two key properties of this dataset. Firstly, activity trends upwards over time. Secondly, aviation activity follows a season pattern, reaching a low in the winter months before reaching a peak around December each year.

This visualization is a little ugly, but the bright contrast of hues should help identify each section clearly, even if colour-blind. The order of each bar is also consistent.

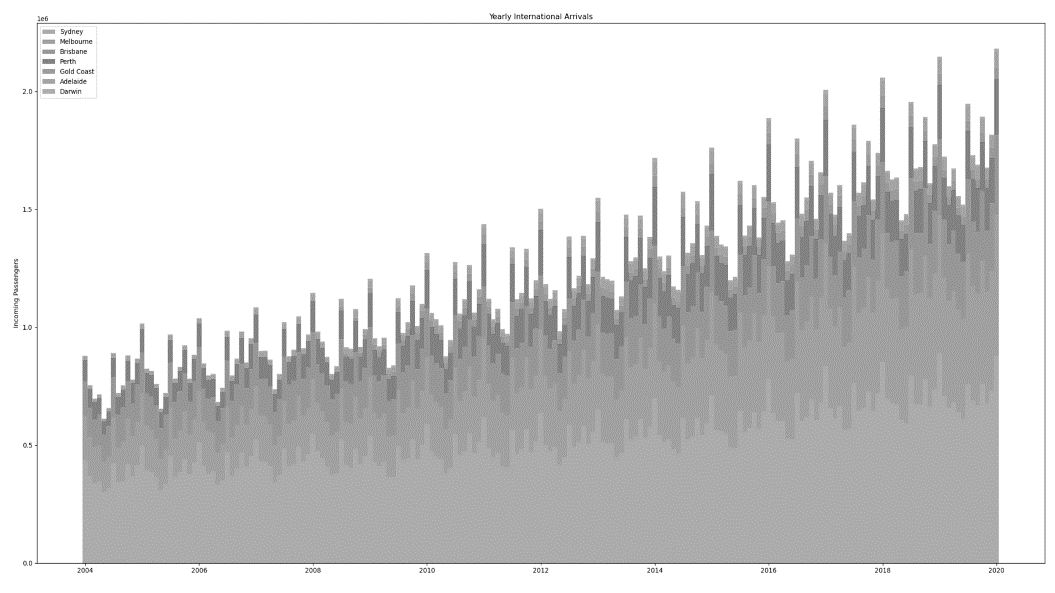


Figure 2: Grayscale version of the above chart, showing that the colour choices are still legible even in the case of vision impairment

# Reflection

Processing the data was a difficult step in the process, but I’m incredibly happy with my decision to use xarray for this task. The multi-dimensional structure provided by xarray then made it incredibly easy to aggregate and simplify the dimensions of this dataset. The main difficult with the dataset processing was I designed the majority of the original code around the international spreadsheets, only to find that the domestic spreadsheets were very different.

I’m reasonably happy with the quality of the visualisations that were created. There was a degree of difficulty determining the most accurate charts to use, as the data was inherently multi-dimensional. Having two categorical dimensions and one continuous dimension meant that the data didn’t fit easily into some common visualisations, and I did not have the time to make visualisations completely from scratch.

My narrow scope of the project ultimately brought some of the insights down – if I repeated this project, I would spend more time focusing on data such as seat utilisation factors and number of flights, or possibly other forms of aviation such as freight.

Aviation data is only released on a biannual basis, which unfortunately means that data for February to April 2020 is unavailable. The trends that occurred in these months would be incredibly interesting to look at, and I’d love the opportunity to come back in a year or two and examine the drastic changes that occurred in 2020.

# Appendix

All code, processed data, and resulting figures can be found in my GitHub repository, <https://github.com/rafraser/COSC3000/tree/master/Visualization>

This repository was kept private until after the due date of the project.

