Chapter 2: Data Visualisation Matters

Steph de Silva

03/11/2017

Introduction

One of the key ways of understanding data and extrapolating ideas from it is data visualisation. It's been around for as long as people have recognised that they can use data to understand the world around them.

For some beautiful examples of historic, hand drawn data visualisation, check out Florence Nightingale's work here:

https://www.theguardian.com/news/datablog/2010/aug/13/florence-nightingale-graphics (https://www.theguardian.com/news/datablog/2010/aug/13/florence-nightingale-graphics)

(Yes, THAT Florenece Nightingale was also a statistician of considerable insight!)

Back in the bad old days it was called "charts" and in the early days of computers in statistics, it was just a bunch of ASCII characters on a page. Ever tried to draw a picture with just +, -, 1 and 0? Things have come on since then!

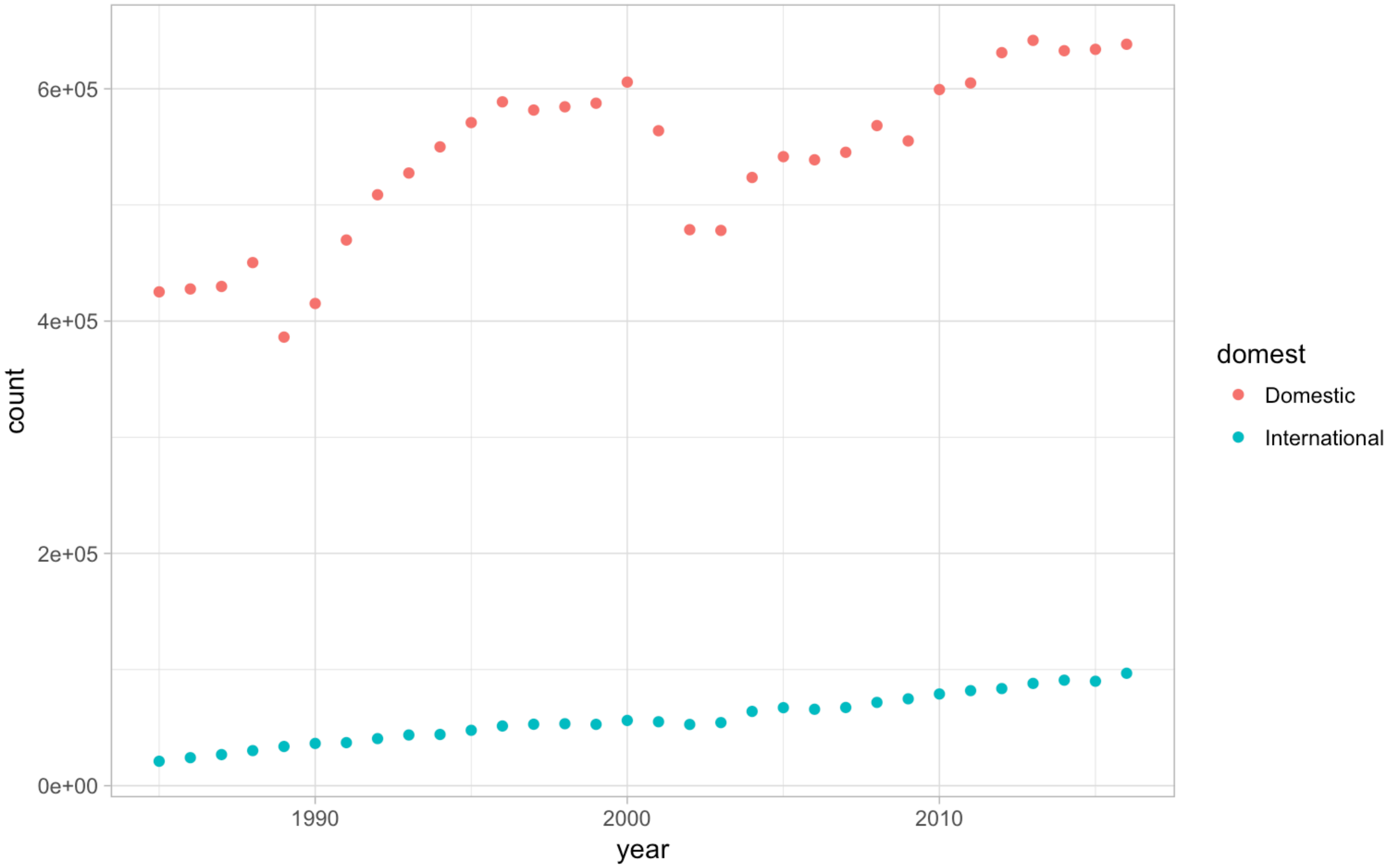
Data visualisation tells a story, and there are some wonderful things being done with it these days. For example, this stark piece by data journalists at 538 is well worth a look: https://fivethirtyeight.com/features/gun-deaths/ (https://fivethirtyeight.com/features/gun-deaths/)

Nathan Yau does some great work visualising data. What do you think happened to marital status between 1900 and 2015? Nathan can tell you in detail without saying very much at all: http://flowingdata.com/2017/07/17/marrying-age-over-the-past-century/ (http://flowingdata.com/2017/07/17/marrying-age-over-the-past-century/)

I expect I've convinced you that data visualisation is an important part of data literacy at this point, but we're going to start out with the basics!

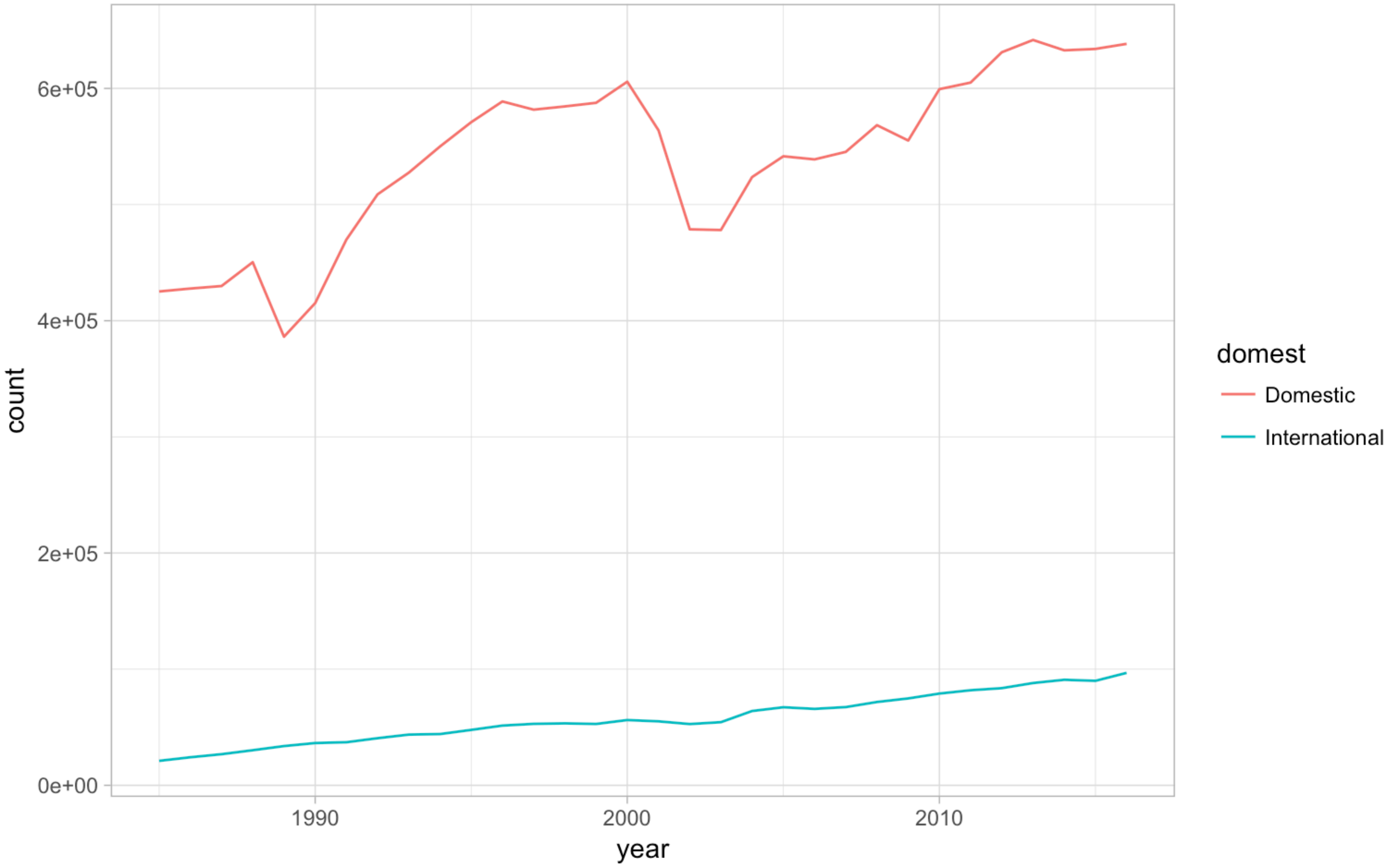
Continuous data, try a scatter plot first

Let's take a look at the total airline movements in Australia over the last few decades as an example. Here I've put the the domestic movements in red and the international ones in blue Our scatter plot shows us that over time both are increasing, but they have different levels!



What does this simple chart show us? It shows us the domestic pilot's dispute in about 1986 (see the dip in the red dots?). That's pretty great for a pretty simple chart.

We could also describe this continuous data as a line chart, let's see how:



Review questions

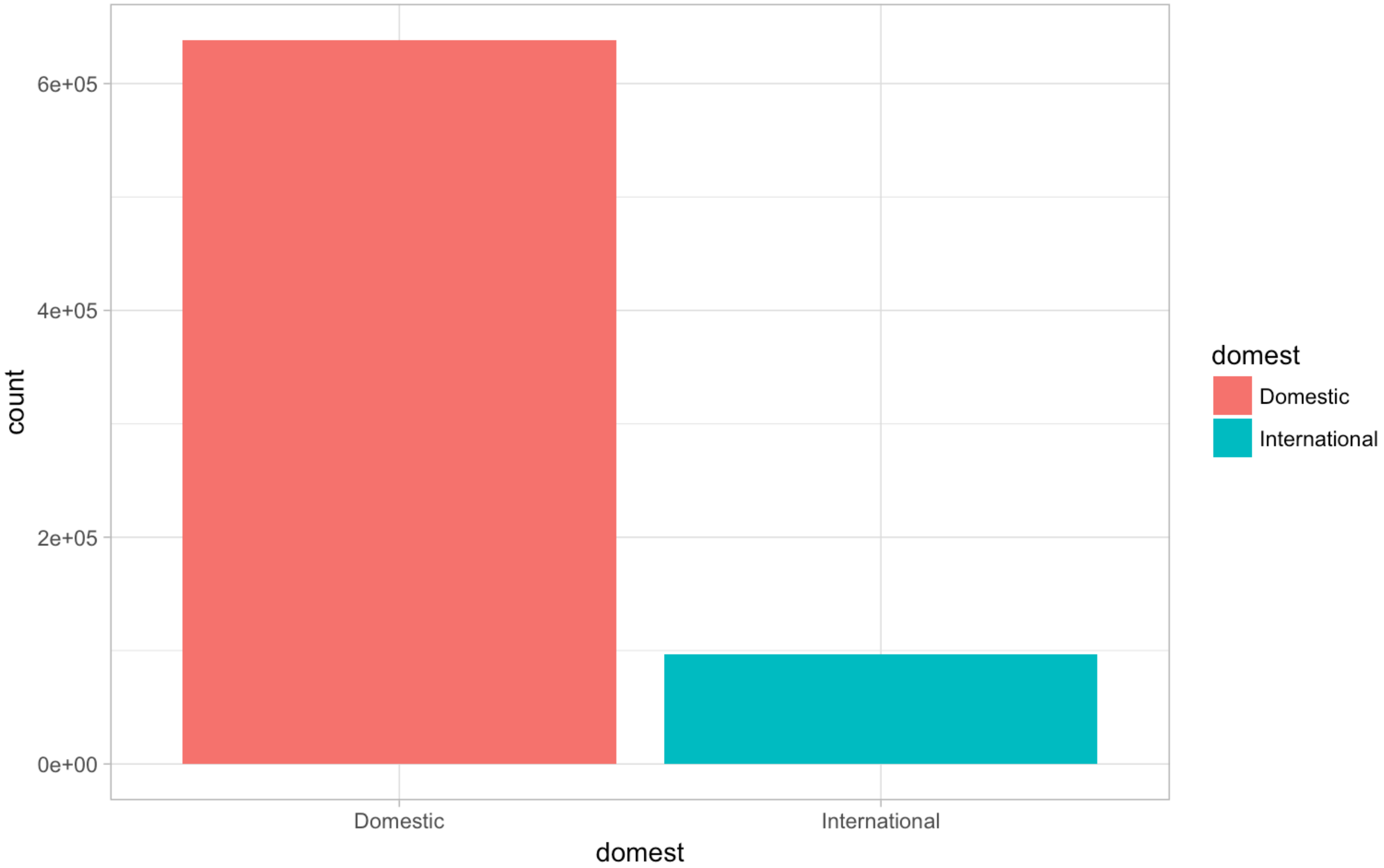
1. What are the differences between domestic movements and international movements?
2. The Sydney Olympics was in 2000. Were there any changes in air movement at about that time?

Review answers

1. There are far fewer international movements, but there tends to be less variation - a more stable upwards trend over time. The number of domestic flights tend to be much more variable.
2. Not a big increase in 2000 at all. But if you look at the period 2001-2003 there's a substantial drop. Do you have any thoughts why that may have been?

Categorical data needs different handling.

Categorical data is often better expressed as a bar or column chart. For instance, let's look at the total number of domestic and international flights in 2016.



There's alot more domestic movements than international! **Notice how the bars don't touch here? This is important- it means we are looking at a bar or column chart, not a histogram.** More on those later!

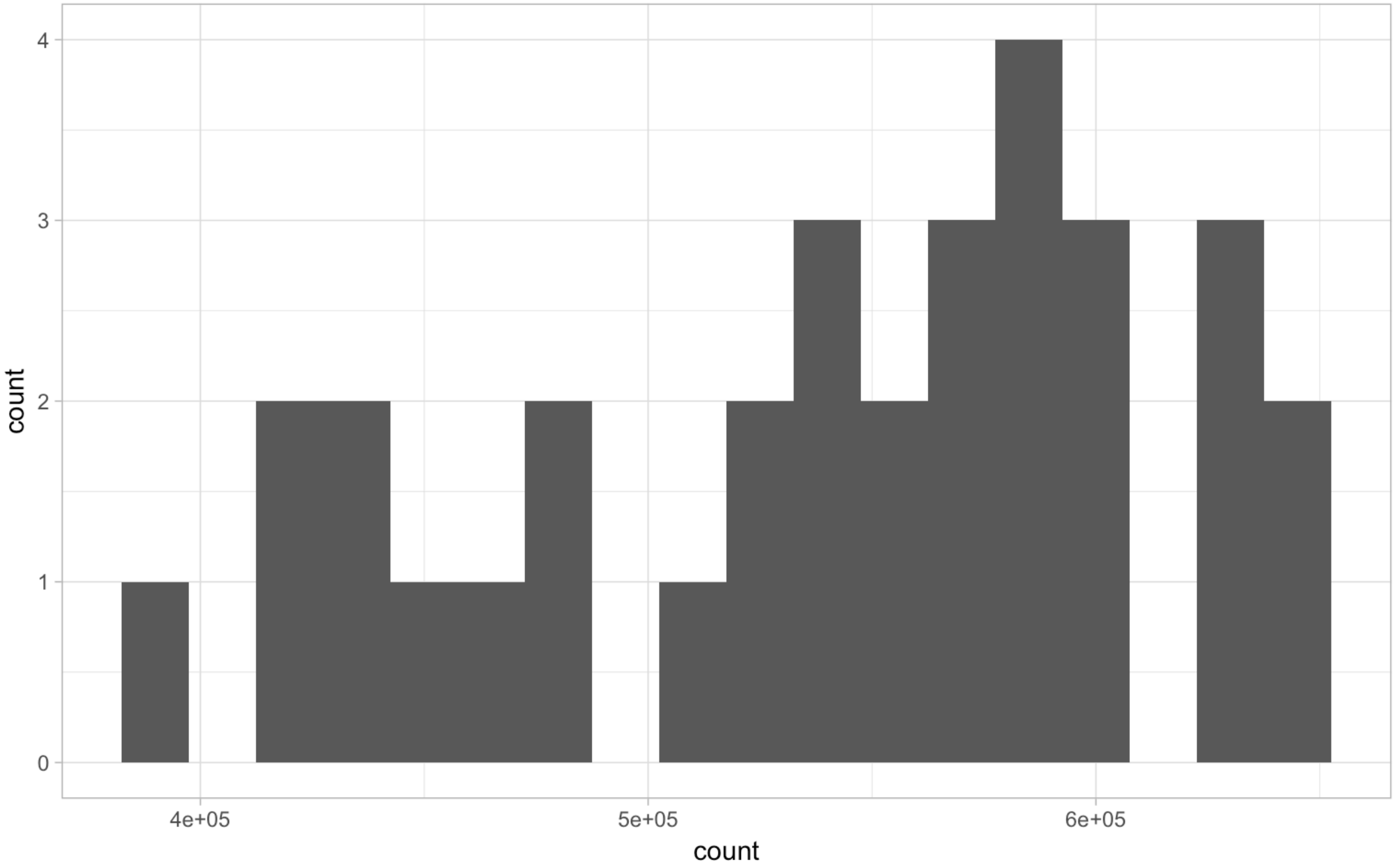
Pro tip: pie charts are considered pretty lame in practice, almost nobody uses them any more. The technical reasons for this is they don't encode data very well, but for our purposes "lame" covers it. 3D should be avoided at all costs.

Frequency distributions are hella useful

Note for those of us over thirty-five: apparently in young people's parlance *hella* is some kind of moderator indicating *very* or *most*.

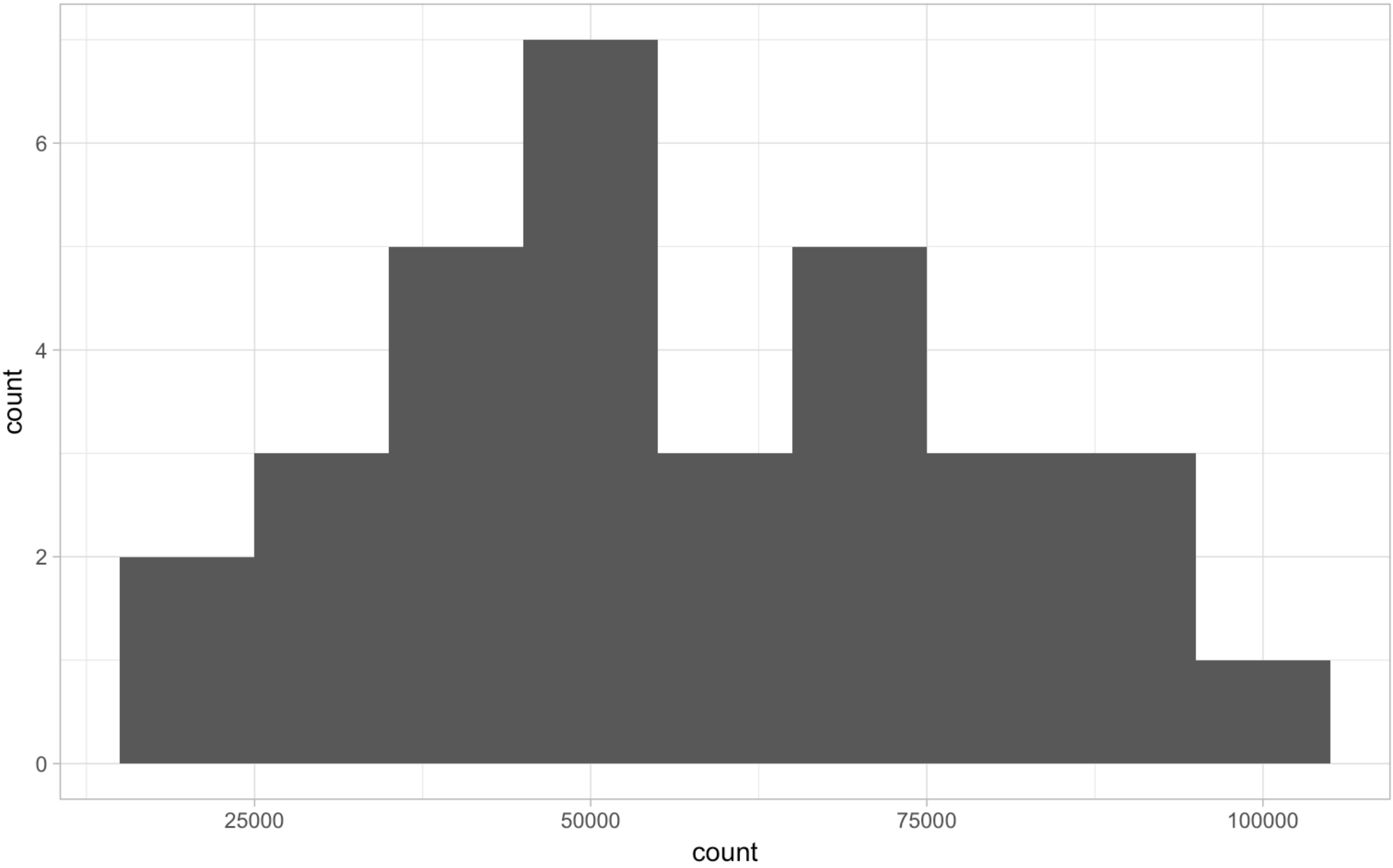
Frequency distributions are one way of describing how data is shaped. This in turn can tell us quite a bit about what's going on in any given data set. We can create frequency distributions for continuous data using *histograms*.

Let's look at a frequency distribution for the number of movements over time for domestic flights.



This data is clearly not symmetrical, most of it is heaped over to the right hand side. We call this right-skewed. Most of the data is heaped to the higher end of the range.

Let's look at international movements now.



This is much more symmetric, but if anything it's skewing left. Most of the data is heaped towards the middle and the bottom. Notice how all the bars touch? **This is a feature you can use to distinguish between a histogram and a bar chart.**

The frequency distribution allows us to see already that the two sets of data are different - most years have fewer international flights, while domestic tends to have more.

And we haven't even done any stats yet!

Visualisation helps us know what to look for and how to explain it.

Want to try some yourself?

This video https://www.youtube.com/edit?o=U&video_id=glnfkgHWhf8 (https://www.youtube.com/edit?o=U&video_id=glnfkgHWhf8) will help you get started in Excel and to make some of your own charts.

This one shows you how to build a histogram in Excel: https://www.youtube.com/edit?o=U&video_id=LQt_uS1cDLw (https://www.youtube.com/edit?o=U&video_id=LQt_uS1cDLw)

Where is this data from?

This data is from BITRE, it uses the annual airport traffic data from the Australian Government Department of Infrastructure and Regional Development located at: https://bitre.gov.au/publications/ongoing/airport_traffic_data.aspx (https://bitre.gov.au/publications/ongoing/airport_traffic_data.aspx)

This chapter uses the ozflights package from Github, written in R.