The `Future` of `NA` Data

Nicholas Tierney

Telethon Kids Institute

`rstudio::conf(2022)`

Hi, my name is Nick Tierney, and I'm going to talk about the future missing data.

I hate missing data

I fully hate missing data. It disrupts your data analysis, because you need to stop, and work out how much is missing, and think: why it is missing, why!?

~~Redacted~~

> I fully hate missing data. It disrupts your data analysis, because you need to stop, and work out how much is missing, and think: why it is missing, why!?
>
> I ~~**redacted**~~ hate missing data. It ~~**redacted**~~ your data analysis. Because ~~**redacted**~~, and work out how ~~**redacted**~~ is missing and think: why ~~**redacted redacted redacted**~~why!?

Missing data contorts messages. I mean, imagine what I just said then was bleeped out, you might get a different idea of what I was saying: I redacted hate missing data. It redacted your data analysis. Because redacted, and work out how redacted is missing and think: why redacted redacted redacted why!?

I ❤️Missing Data

But i've grown to love missing data. Understanding it is hard, but it adds an extra challenge, a little bit of spice, to a data analysis.

And I've written two R packages to help you explore your data, and explore missingness: visdat, and naniar.

Plan

1. Explore missing data: Overview –> Relationship
2. Brief tour of missing data visualisations.

And today, I'm going to talk about two things: A principle of going from overview to relationship to explore missing data A brief tour of missing data visualisations. Emphasis on the brief. I won't have time to explore them all in detail.

Overview

```r
vis_miss(oceanbuoys)
```

The first one I'd recommend is vis_miss. We are looking at a heatmap of the missingness of your data - as if you are looking at your data from a birds eye view. The rows and columns of your data are shown as either missing - in black, or present - in gray.

What we learn from this is that the variables with the most missings are "air_temp_c" and "humidity", but there are some missing values that go missing at the same time across these variables - which you can see as these black horizontal lines stretching across multiple variables.

We also get information on the percent of missing data in each of the variables, with air temperature and humidity having 11 and 12 % missing data, respectively.

Now, we go from the overview, to the relationship - in this case, the relationship between these two variables.

~~Missing~~ Relationship

```r
ggplot(oceanbuoys, aes(x = air_temp_c, y = humidity)) +
  geom_point()
```

```
Warning: Removed 171 rows containing missing values (geom_point).
```

We can explore the relationship in a scatterplot using ggplot2. And constructing a ggplot call like so, of humidity and air temperature - we encounter a substantial problem, which is that the missing values are removed. Although I must say that it does this loudly, which is very useful!

Missing Relationship

```r
ggplot(oceanbuoys, aes(x = air_temp_c, y = humidity)) +
  geom_miss_point()
```

We can instead use the function `geom_miss_point()` from naniar, which imputes the values below the range of the data, so that they show up on the plot, but in a different position, and also a different colour. Let's break this down. Let's look at each axis, one at a time.

Missing relationship

```r
ggplot(oceanbuoys, aes(x = air_temp_c, y = humidity)) +
  geom_miss_point()
```

The values in red on the X axis are air temperature values, which are missing for humidity. We see that this distribution of values here matches this first cluster on the left.

Missing relationship

```r
ggplot(oceanbuoys, aes(x = air_temp_c, y = humidity)) +
  geom_miss_point()
```

And on the Y axis are the humidity values that have missing air temperature values - interesting again is that this cluster of values seems to match the second cluster. So, what we learn from this is that the missingness is aligned with these two clusters of the air temperature and humidity values.

Let's explore this data by facetting the data along another column: year.

Missingness relationship + explore

```r
ggplot(oceanbuoys, aes(x = air_temp_c, y = humidity)) +
  geom_miss_point() +
  facet_wrap(~year)
```

This is an important feature of this geom, is that it allows you to construct ggplots as you would normally - we can add some regular code to facet by year. What we learn from this, is that the missingness of each of these appears to be aligned by year, very neat!

Moar missingness vis

Now, some more missing data visualisations

Missingness in Variables

```r
gg_miss_var(oceanbuoys)
```

Gg_miss_var to quickly show the amount of missings in variables

Missingness in Variables %

You can even show the percentage of missing

```r
gg_miss_var(oceanbuoys, show_pct = TRUE)
```

Missingness in Variables + facetted

And facet by another variable to explore - hey look, we found it again, the pattern of missing values for each year.

```r
gg_miss_var(oceanbuoys, facet = year)
```

Combinations of missings

Use gg_miss_upset to explore combinations of missingness in a simple datases

```r
gg_miss_upset(oceanbuoys)
```

(Complex) Combinations of missings

or a more complex one.

```r
gg_miss_upset(riskfactors)
```

`gg_miss_fct()`

And we can use gg_miss_(factor) to explore the percent of missingness of all variables grouped by another factor. Here we see the percent of missing data over all variables for different marital statuses.

```r
gg_miss_fct(x = riskfactors, fct = marital)
```

Future work: moar `geom` s

With some help from the R Consortium, we'll be adding some more tools for visualising missing data. I'll briefly talk about just two new additions that I think will be really cool:

Future work: `geom_miss_histogram()`

Geom_miss_histogram will show the amount of missing values alongside a univariate distribution, by imputing it below the range of the data. In this plot, this gives us a general sense of the amount of missing data in the air t emperature data.

Future work: `geom_imputed_point()`

geom_imputed_point() will take in a dataset and identify any imputed values, and will also show any values that are still missing

A book: "The Missing Book"

Nicholas Tierney & **Allison Horst**

And in some very exciting news, I'd like to introduce the first draft of a book that Allison Horst and I are working on: "The Missing Book".

It contains exercises on missing data, and general workflows on exploring missing values and imputations. The aim is to have a general book to guide you through exploring data, with some case studies of real missing data.

The Future of Missing Data is `presence`

I'll wrap up now by saying that the future of missing data is it's presence in software, in data analysis workflows, in tutorials, and in how we think about data generally.

Funnily enough, missing data is almost always present. So let's make sure we don't forget it.