

How R powers data journalism

Chris Knox

Data journalist

The New Zealand Herald

WHAT IS DATA JOURNALISM?

Election 2023: Electoral Commission reviewing voting results after Herald investigation led to discovery of three wrong counts



By Chris Knox

Data Editor and Head of Data Journalism - NZ Herald · 7 Nov, 2023 06:23 PM · 5 mins to read



The Electoral Commission is undertaking "a full check" of the country's voting results after a Herald investigation led to three booths being found to have wrongly assigned hundreds of votes to fringe parties, while the National Party received none.

The commission has admitted a data entry error led to hundreds of votes being assigned to the Leighton Baker Party and the New Conservatives.

"A full check of all voting place results is being undertaken immediately to establish if there are any other transcription errors," the commission said.

"Data can be the source of data journalism, or it can be the tool with which the story is told – or it can be both."

- Paul Bradshaw, Birmingham City University

"Data journalism is simply journalism."

"The former is a new and trendy term but ultimately, it is just a way of describing journalism in the modern world."

- Bureau of Investigative Journalism

I like to think of it as stories where data can be thought of as a source, or the only source. And R is one of the best ways of talking to that source.

PROGRAMMING AND ME

- 1999** - Python is better than Fortran (and BASIC).
- 2005** - C++ is necessary for 3D computer graphics.
- 2007** - Classes are overrated - C is fine for everything.
- 2009** - Python plays well with C/C++.
- 2011** - D3.js make javascript awesome for data visualization.
- 2013** - Hello R. Why are you so strange?
- 2014** - Hello SQL. How did I not know about you?
- 2016** - Everything should be functional.
- 2018** - Hello again R. You're awesome - why did I hate you?
- 2019** - Svelte (Javascript) for web, R for everything else.
- 2021** - R with tidyverse, RMarkdown, and targets does everything.
- 2022** - Hello quarto, R for everything (except web and serverless).

WHY IS R SO GOOD FOR DATA JOURNALISM?

Data journalism is all about both:

- Quick exploratory analysis, and
- Regularly recurring analysis.

R excels at both!

Ironically, while writing this slide I got an email with an updated CEO salary for our CEO pay article.

Data flows into the article from a targets-based R pipeline.

The longest step in refreshing all four interactives was updating my AWS credentials.

Ebos reports in Australian dollars, given that the group generates about 80% of its earnings in Australia. The company, which focuses on marketing and distributing medical consumable products, has delivered an 89% shareholder return over five years.

The ten highest paid NZ CEOs

1	 John Cullity \$7,273,119 EBOS	6	 Greg Foran \$4,187,571 AIR NEW ZEALAND
2	 David Bortolussi \$6,586,798 A2 MILK COMPANY	7	 Don Braid \$3,781,227 MAINFREIGHT
3	 Miles Hurrell \$5,924,782 FONTERRA	8	 Lewis Gradon \$3,533,971 FISHER & PAYKEL HEALTHCARE
4	 Vittoria Shortt \$5,154,928 ASB BANK	9	 Dan Huggins \$3,371,070 BNZ
5	 Nick Grayston \$4,220,000 THE WAREHOUSE GROUP	10	 Neal Barclay \$3,246,658 MERIDIAN ENERGY

ELECTION RESULTS "INVESTIGATION"

```
in_parliament <-  
  "ACT|Green|Labour|Māori|National|Zealand First"  
party_votes |>  
  mutate(Total = sum(votes), .by = vp_id) |>  
  mutate(Percent = votes/Total*100) |>  
  filter(!grepl(in_parliament, short_name)) |>  
  arrange(desc(Percent))
```

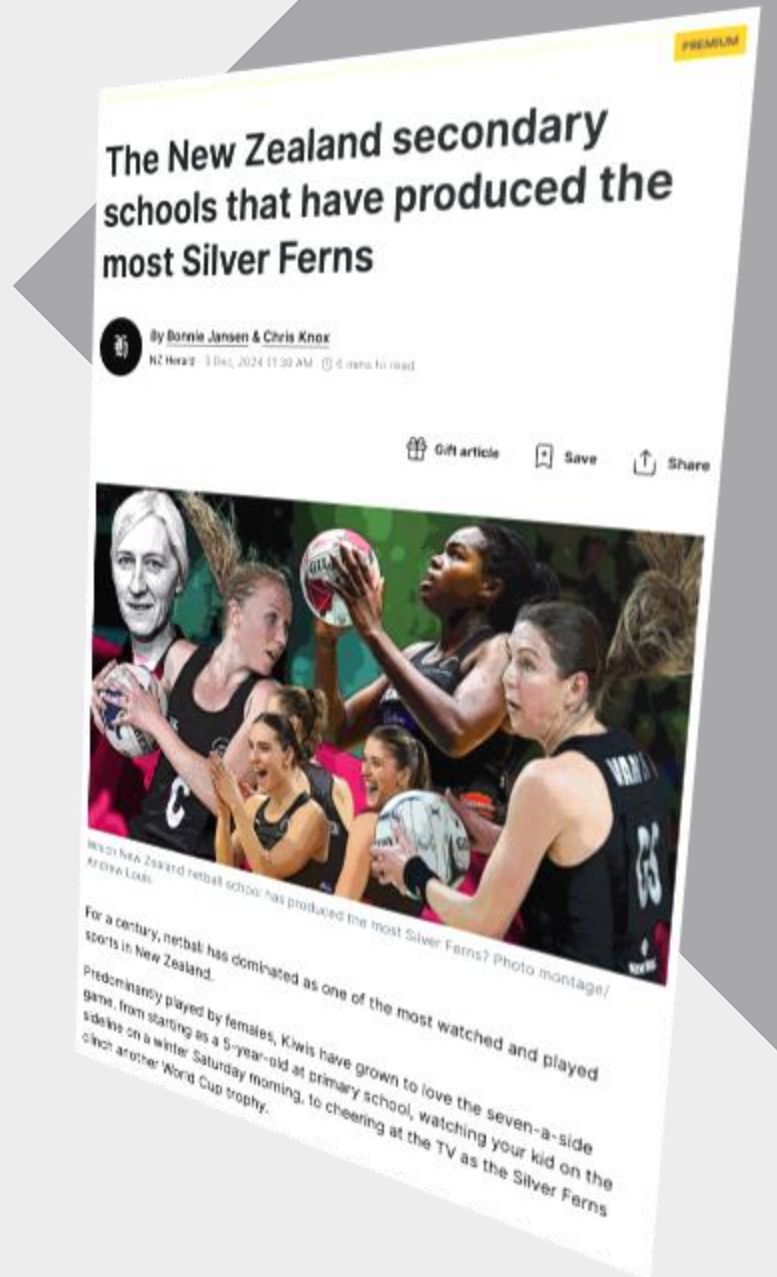
A tibble: 23,403 × 12

	p_no	votes	short_name	abbrev	e_no	location	vp_e_no	electorate	vp_address	vp_id	Total	Percent
	<dbl>	<dbl>	<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	55	505	Leighton Baker	LBP	39	Port Waikato	39	Port Waikato	Pukekohe Intermediate School, 44 Queen Street, Pukekohe	781	1199	42.1
2	54	49	NZ Loyal	NZL	61	West Coast-Tasman	61	West Coast-Tasman	Ngatimoti School, 4 Greenhill Road, Ngātīmoti	1028	214	22.9
3	51	27	Freedoms NZ	FNZ	61	West Coast-Tasman	61	West Coast-Tasman	Pop-up Voting Services, Various Locations - Greymouth, Taken in West Coast-Tasman	3731	164	16.5
4	54	12	NZ Loyal	NZL	61	West Coast-Tasman	61	West Coast-Tasman	Puramāhoi Hall, 602 Takaka-Collingwood Highway (State Highway 60), Puramāhoi	1126	88	13.6
5	54	12	NZ Loyal	NZL	16	Invercargill	16	Invercargill	Waikawa Public Hall, 11 Larne Street, Waikawa	1419	90	13.3

MESSY DATA

Data about schools is often messy

- School names change over time
- Schools have multiple names
- Sometimes people just can't spell



MESSY DATA FIXES

```
sf_raw |>
  mutate(
    HS = str_replace_all(`LAST HS`, " ,", " ") |>
      str_replace_all("\\r\\n", " ") |>
      str_replace_all("Girls ", "Girls' ") |>
      str_replace_all("School ", "School, ") |>
      str_replace_all("Gisbourne", "Gisborne") |>
      str_replace_all("Secndry", "Secondary") |>
      str_replace_all("Scondary", "Secondary") |>
      str_replace_all("Rutherforld", "Rutherford") |> ...
```

THE PRICE OF CHEESE

Some of the most interesting data in Stats NZ's monthly price data release is buried in a 55k row long-format csv file

I love big long-format csv files.

There should be more of them, but only because R makes working with them easy.



FINDING PRICE CHANGES

```
food |>
```

```
  filter(Period >= ym("201501")) |>
```

```
  select(Series_title_1, Period, Data_value) |>
```

```
  pivot_wider(names_from = Period, values_from = Data_value) |>
```

```
  mutate(change = (`2025-02-01` - `2024-02-01`) / `2024-02-01` * 100) |>
```

```
  arrange(desc(change)) |>
```

```
  mutate(`Feb 2024` = `2024-02-01`, `Feb 2025` = `2025-02-01`) |>
```

```
  relocate(Series_title_1, `Feb 2024`, `Feb 2025`, change) |>
```

```
  clipr::write_clip()
```

PIPELINES IN R

A (very) short overview of {targets}

{targets} was developed by Will Landau at Eli Lilly and is now released by the rOpenSci organization.

It takes bit to learn the targets workflow, but the payoffs are well worth it.

<https://books.ropensci.org/targets/>



WHAT DOES {targets} DO?

Each step in your analysis can be encapsulated and then cached.

Targets will automatically detect any downstream steps and will only re-run them when things change.

```
library(targets)
library(tarchetypes)
tar_source()
tar_option_set(packages = c("readr", "dplyr", "ggplot2"))
list(
  tar_file(file, "data.csv"),
  tar_target(data, get_data(file)),
  tar_target(model, fit_model(data)),
  tar_target(plot, plot_model(model, data))
)
```

INTERACTIVE {targets}

I use a file called interactive.R for exploring data

```
library(targets)
library(tidyverse)
tar_load(c(expenditure, expenditure_est))
all_expenditure <- expenditure |>
  count(year, wt=amt, name = 'Actual expenditure') |>
  inner_join(expenditure_est |> count(year, wt=amt, name = 'Estimated expenditure'))

all_expenditure |>
  pivot_longer(-year) |>
  ggplot(aes(x=year, y=value, color=name)) +
  geom_line()
```



NZ
ME.

**NEW ZEALAND
MEDIA AND
ENTERTAINMENT**

chris.knox@nzme.co.nz