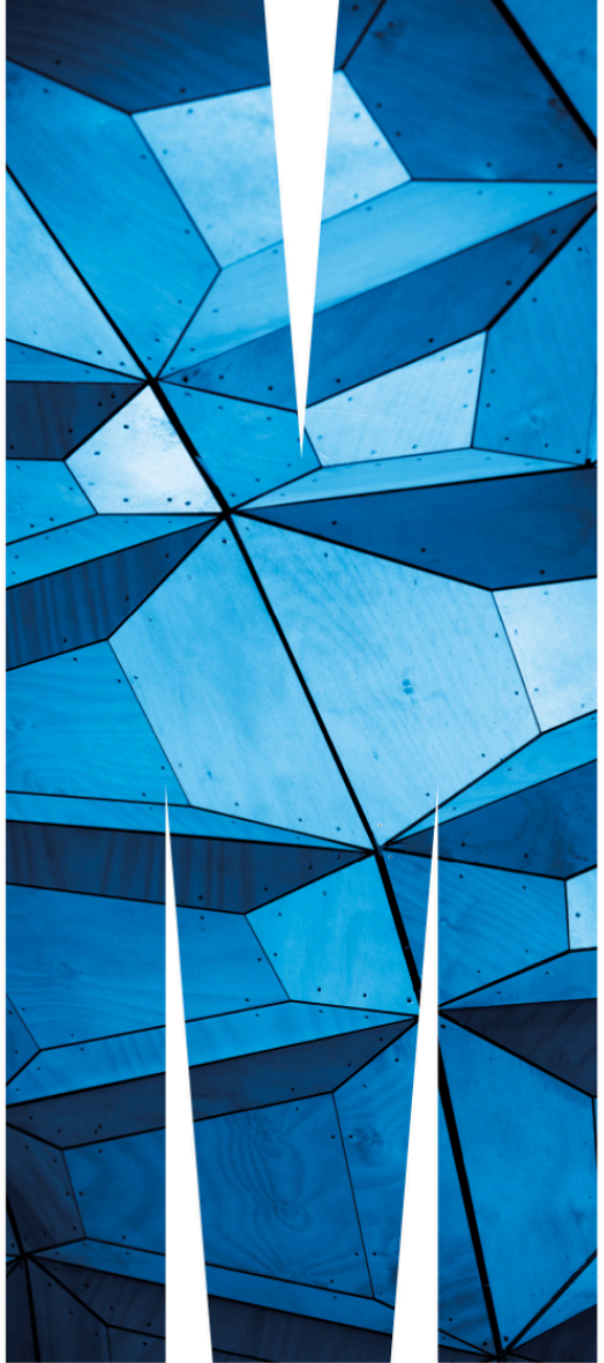
MONASH University

ETC5512: Wild Caught Data

Case Study: COVID-19 Decision Making



Lecturer: *Michael Lydeamore*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 10

Today you will:

- Learn about exploratory data techniques and analysis
- Consider ethical standpoints when making recommendations
- Combine multiple sources of data to increase information

Bring your minds back

It's August 2020. You're in Victoria, working as a public health advisor to the state government.

Movement restrictions are the tighest they've ever been, and for a long time too.

The number of COVID-19 cases is staying stubbornly high. You've been tasked with working out what should be done, and when.

What should you do?

Let's get some data

COVID-19 was an unusual succses of open data, **but**

- No standard format
- No standard reporting window
- No standard distribution method

These days, a lot of the COVID data sources are no longer being (regularly) updated.

The Victorian government has a dataset of cases available on their website.

```r
library(tidyverse)
cases <- read_csv("./data/ncov_cases_by_postcode_lga.csv")

glimpse(cases)
```

```
## Rows: 295,148
## Columns: 7
## $ diagnosis_date   <date> 2020-01-25, 2020-01-28, 2020-01-30, 2020-01-31, 202
## $ postcode         <dbl> 3149, 3150, 3006, 3008, 3058, 3931, 3109, 3104, 3802
## $ lga_name         <chr> "Monash (C)", "Monash (C)", "Melbourne (C)", "Melbou
## $ lga_code         <dbl> 24970, 24970, 24600, 24600, 25250, 25340, 24210, 211
## $ RAT_case_count   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ PCR_case_count   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ Total_case_count <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
```

What is the structure of this dataset?

What do the columns mean?

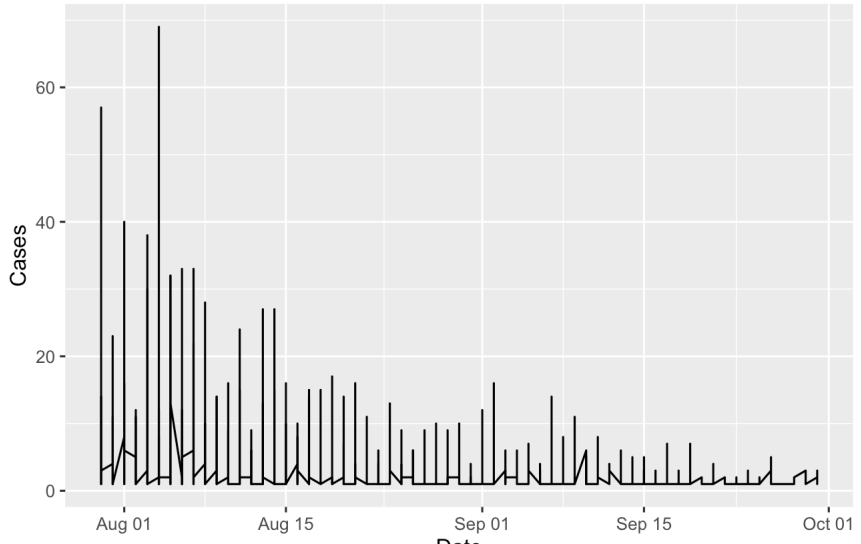
Some definitions

- Local Government Area (LGA): A geographic area mostly represented by a single council.
- PCR: Polymerase-chain reaction. A method for replicating virus to detect it's presence/absence.
- RAT: Rapid antigen test. A less specific and sensitive test that provides presence/absence of an organism.

Data exploration

Today is August 1, 2020. Let's have a look back at the last 2 months:

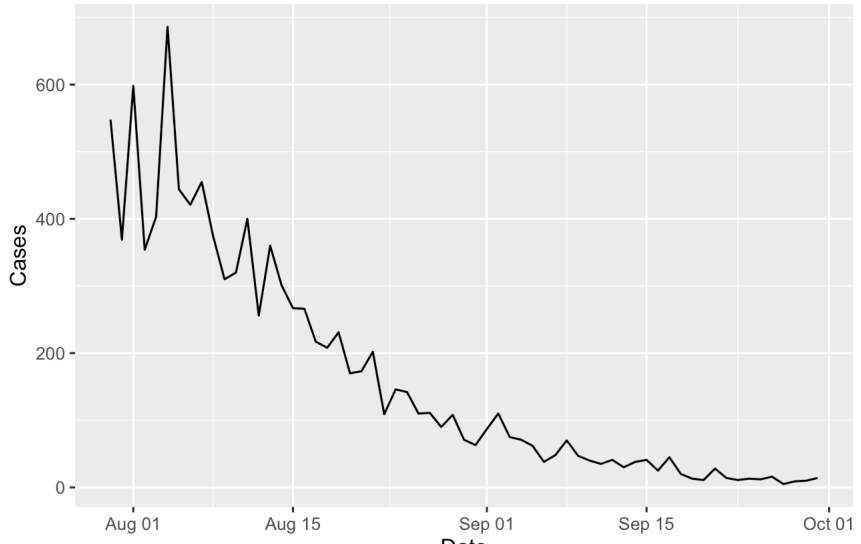
```
date_interval <- interval(ymd("2020-09-30") - months(2), ymd("2020-09-30"))
cases %>%
  filter(diagnosis_date %within% date_interval) %>%
  group_by(diagnosis_date) %>%
  ggplot(aes(x = diagnosis_date, y = Total_case_count)) +
  geom_line() +
  labs(x="Date", y="Cases")
```



Data exploration

Cases are listed separately by geography, so we need to group them up:

```
cases %>%
  filter(diagnosis_date %within% date_interval) %>%
  group_by(diagnosis_date) %>%
  summarise(n = sum(Total_case_count)) %>%
  ggplot(aes(x = diagnosis_date, y = n)) +
  geom_line() +
  labs(x = "Date", y = "Cases")
```



We've generated evidence of the problem.

What are some ideas for solutions?

Spatial variation

There is much talk about differences in geographical response to COVID. It started as differences between the states, now people are talking about within the state itself.

Senior people are talking about restrictions "not working": what does that mean?

Let's have a look at spatially disaggregated data.

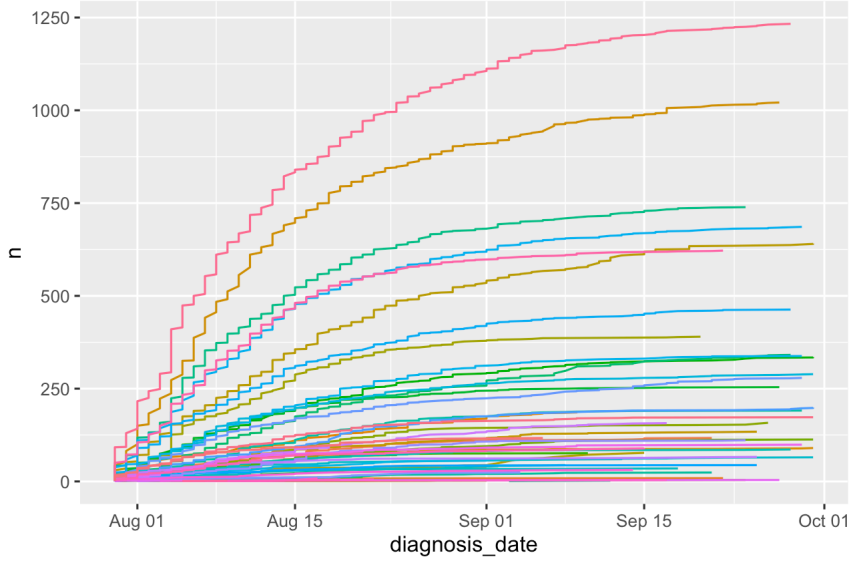
Spatial variation

```
cases %>%
  filter(diagnosis_date %within% date_interval) %>%
  group_by(diagnosis_date, lga_name) %>%
  summarise(n = sum(Total_case_count)) %>%
  ggplot(aes(x = diagnosis_date, y = n, colour = lga_name)) +
  geom_line()
```

A little hard to see...

Spatial variation

```
cases %>%
  filter(diagnosis_date %within% date_interval) %>%
  group_by(lga_name) %>%
  arrange(diagnosis_date) %>%
  mutate(n = cumsum(Total_case_count)) %>%
  ggplot(aes(x = diagnosis_date, y = n, colour = lga_name)) +
  geom_line() + guides(colour="none")
```



What do you see?

The order of the lines doesn't change much

Perhaps restrictions aren't as effective in certain places

Mobility data

Google released a series of "mobility reports" that attempted to quantify changes in mobility patterns using mobile phone data.

The reports are no longer updated but are available online:

https://www.google.com/covid19/mobility/

Let's take a peak

```
mobility <- read_csv("data/2020_AU_Region_Mobility_Report.csv")
head(mobility)
```

```
## # A tibble: 6 × 15
##   country_region_code country_region sub_region_1 sub_region_2 metro_area iso_3166_2_code cen
##   <chr>               <chr>          <chr>        <chr>        <lgl>      <chr>           <lg
## 1 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## 2 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## 3 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## 4 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## 5 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## 6 AU                  Australia      <NA>         <NA>         NA         <NA>            NA
## # ℹ abbreviated names: ¹retail_and_recreation_percent_change_from_baseline, ²grocery_and_phar
## # ℹ 3 more variables: transit_stations_percent_change_from_baseline <dbl>, workplaces_percent
```

Quite big: 85330 rows and 15 columns.

Let's try to find Victoria...

Let's take a peak

```
vic_mobility <- mobility %>%
  filter(sub_region_1 == "Victoria", !is.na(sub_region_2))

head(vic_mobility)
```

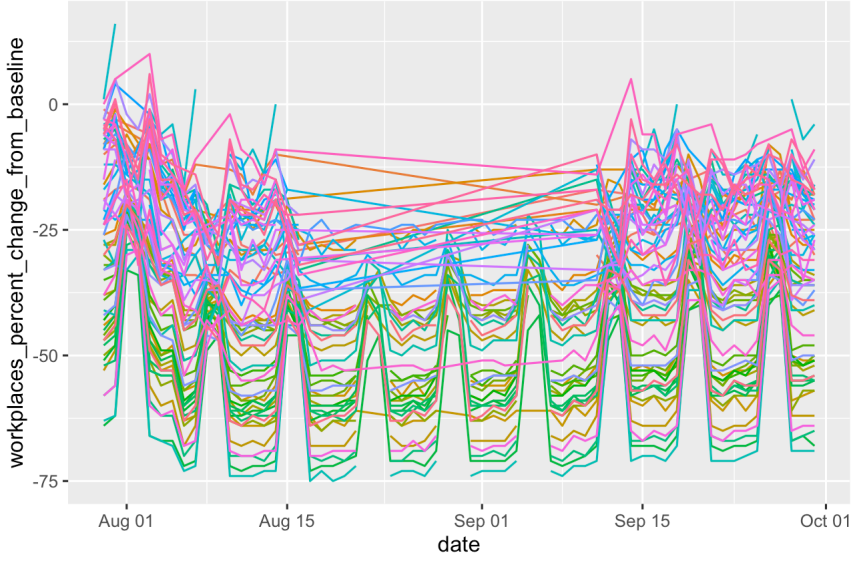
```
## # A tibble: 6 × 15
##   country_region_code country_region sub_region_1 sub_region_2 metro_area iso_3166_2_code cen
##   <chr>               <chr>          <chr>        <chr>        <lgl>      <chr>           <lg
## 1 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## 2 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## 3 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## 4 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## 5 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## 6 AU                  Australia      Victoria     Alpine Shire NA         <NA>            NA
## # ℹ abbreviated names: ¹retail_and_recreation_percent_change_from_baseline, ²grocery_and_phar
## # ℹ 3 more variables: transit_stations_percent_change_from_baseline <dbl>, workplaces_percent
```

Closer...

Incidentally, this is the classic quest of the analyst. A lot of time manipulating data to answer a seemingly straightforward question.

Let's take a peak

```
vic_mobility %>%
  filter(date %within% date_interval) %>%
  ggplot(aes(x = date, y = workplaces_percent_change_from_baseline, colour = sub
  geom_line() + guides(colour = "none")
```



Unreadable.

Suggestions for fixes?

So back to the question

After playing around with our data, I find it useful to return to the question. For us that was:

What should be done to try to reduce COVID cases, and why?

You've heard talks of movement restrictions being further tightened and are expecting to be asked about that. Let's see if we can get a handle on how explanatory mobility is on cases...

Mobility and cases

In theory, mobility should be tied closely to cases. For an infectious disease to spread between two people:

- There has to be contact between two people
- One has to be *susceptible*
- One has to be infectious
- Virus has to somehow move from the susceptible person to the infectious person

It is hard to control points 2, 3 and 4, so a lot of effort goes to point 1.

Mobility restrictions *decrease* the amount of time an individual spends in community, so while they may still infect those close to them, they should be less likely to see transmission outside their household.

Mobility and cases

Let's join together our two datasets:

```
cases %>%
  left_join(mobility, by = join_by(diagnosis_date == date, lga_name == sub_region_1))
```

```
## # A tibble: 295,148 × 20
##    diagnosis_date postcode lga_name    lga_code RAT_case_count PCR_case_count Total_case_coun
##    <date>            <dbl> <chr>          <dbl>          <dbl>          <dbl>            <dbl
##  1 2020-01-25         3149 Monash (C)     24970              0              1
##  2 2020-01-28         3150 Monash (C)     24970              0              1
##  3 2020-01-30         3006 Melbourne …    24600              0              1
##  4 2020-01-31         3008 Melbourne …    24600              0              1
##  5 2020-02-22         3058 Merri-bek …    25250              0              1
##  6 2020-02-22         3931 Mornington…    25340              0              1
##  7 2020-02-25         3109 Manningham…    24210              0              1
##  8 2020-03-01         3104 Boroondara…    21110              0              1
##  9 2020-03-01         3802 Casey (C)      21610              0              1
## 10 2020-03-04         3006 Melbourne …    24600              0              1
## # ℹ 295,138 more rows
## # ℹ 6 more variables: retail_and_recreation_percent_change_from_baseline <dbl>, grocery_and_p
## #   transit_stations_percent_change_from_baseline <dbl>, workplaces_percent_change_from_baseli
```

Mobility and cases

We know the focus is on metropolitan Melbourne, so let's focus down on just those at least:

```
metro_lgas <- c(
  "Melbourne", "Port Phillip", "Yarra", "Stonnington", "Bayside", "Boroondara",
  "Melton", "Brimbank", "Hobsons Bay", "Wyndham", "Moonee Valley", "Maribyrnong"
  "Banyule", "Whittlesea", "Nillumbik", "Hume", "Moreland", "Darebin",
  "Manningham", "Whitehorse", "Knox", "Yarra Ranges", "Maroondah", "Monash",
  "Kingston", "Frankston", "Cardinia", "Casey", "Greater Dandenong", "Mornington
)
```

What's coming next is the classic plight of the data analyst...

Inconsistent naming conventions.

Challenge: Making the LGA names match

Download the mobility data and the case data from these two websites:

- Case data: https://discover.data.vic.gov.au/dataset/victorian-coronavirus-data
- Mobility data: https://www.google.com/covid19/mobility/

Have a go at getting `sub_region_2` to match `lga_name`.

When you make some progress, shout it out - let's try to solve it together.

Challenge: Making the LGA names match

Let's look at the total cases over the previous month to now, and compare that to mobility.

```
cases_vs_mobility <- cases %>%
  group_by(diagnosis_date, lga_name) %>%
  summarise(total_cases = sum(Total_case_count)) %>%
  left_join(vic_mobility, by = join_by(lga_name == new_lga_name, diagnosis_date
  group_by(lga_name) %>%
  summarise(
    n = sum(total_cases),
    mean_mobility_change = mean(workplaces_percent_change_from_baseline, na.rm =
  )
```

Challenge: Making the LGA names match

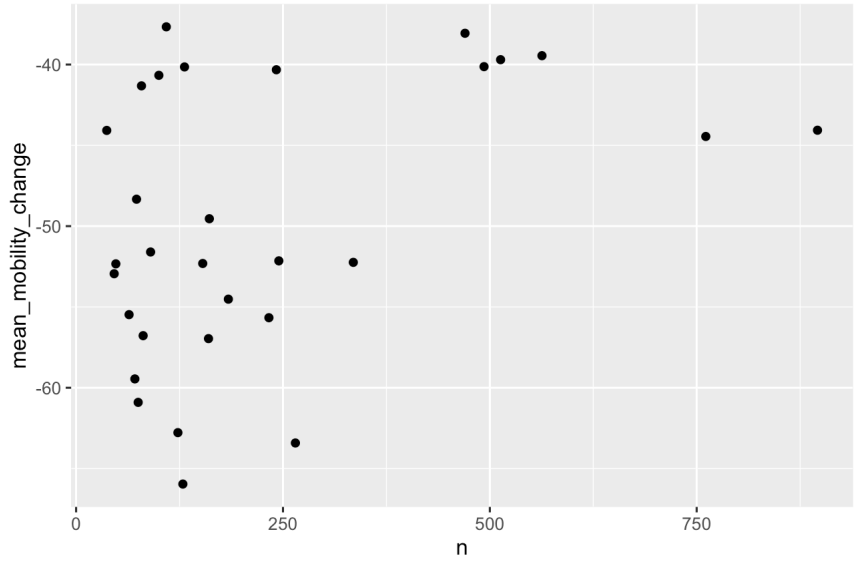
```
cases_vs_mobility
```

```
## # A tibble: 30 × 3
##    lga_name              n mean_mobility_change
##    <chr>             <dbl>                <dbl>
##  1 Banyule              90                -51.6
##  2 Bayside             160                -57.0
##  3 Boroondara           71                -59.5
##  4 Brimbank            761                -44.5
##  5 Cardinia            109                -37.7
##  6 Casey               470                -38.1
##  7 Darebin             335                -52.2
##  8 Frankston           131                -40.1
##  9 Glen Eira            81                -56.8
## 10 Greater Dandenong   242                -40.3
## # ℹ 20 more rows
```

Let's try and plot

Plotting our exploratory analysis

```
cases_vs_mobility %>%
  ggplot(aes(x = n, y = mean_mobility_change)) +
  geom_point()
```

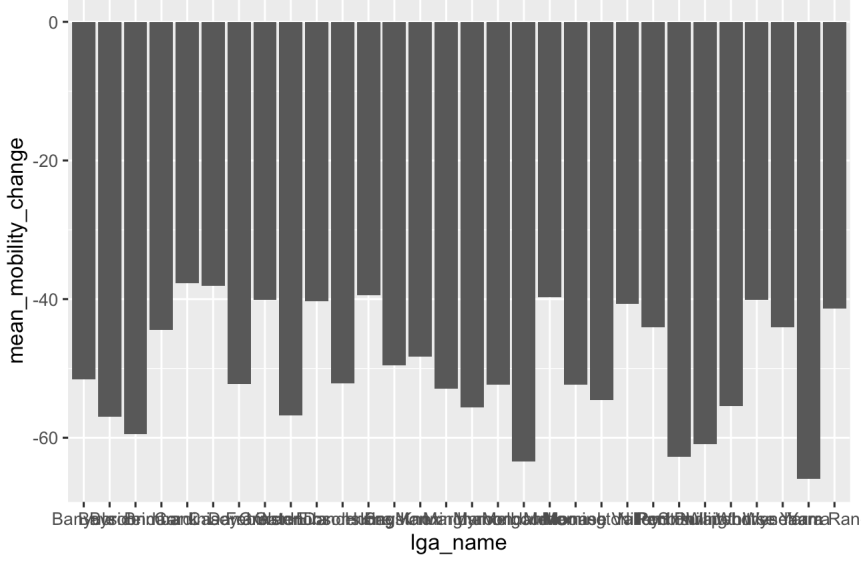


Thoughts?

Plotting our exploratory analysis

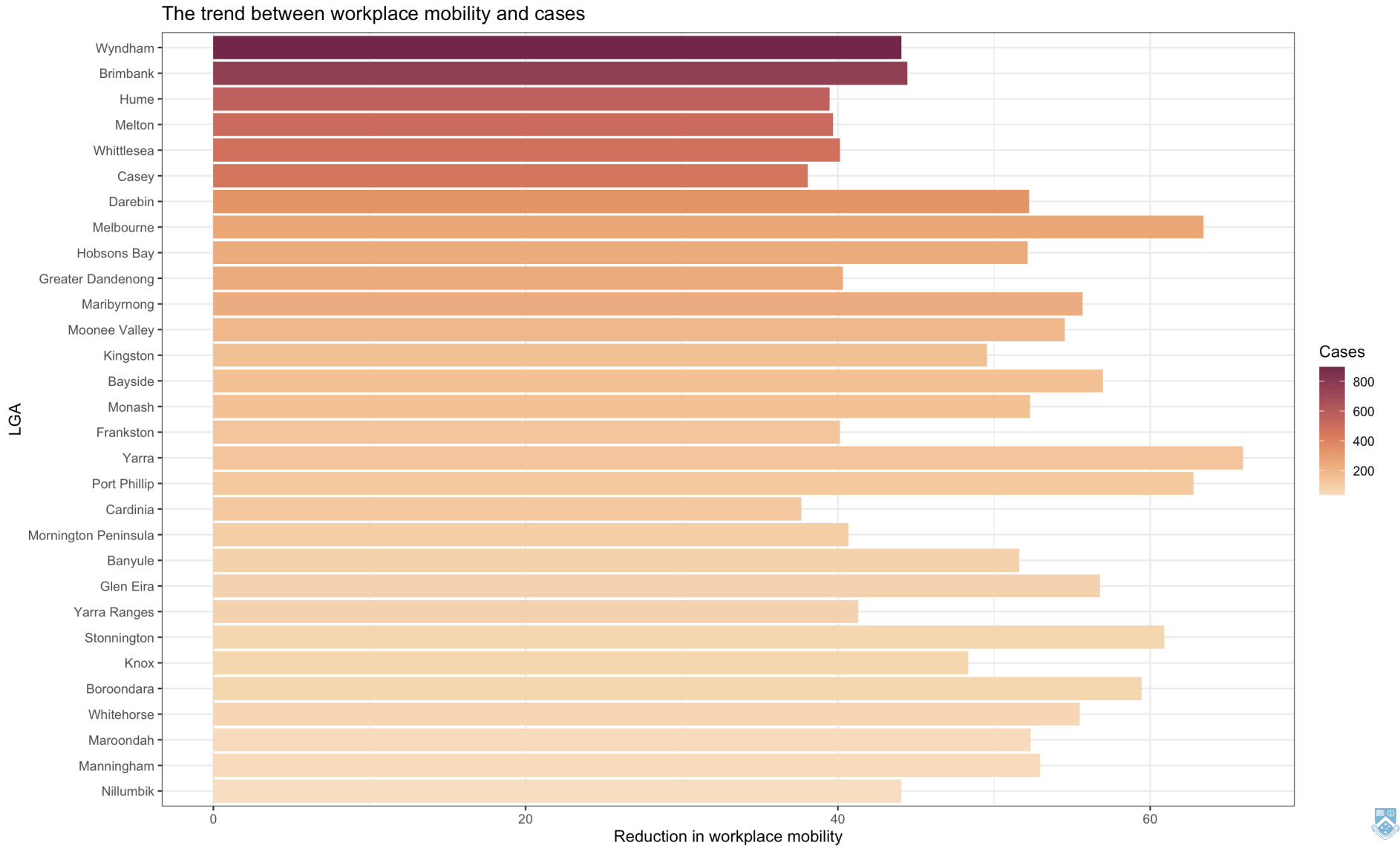
An aside: Let's think about how to format this for a report. Non-technical readers will struggle with the previous plot, and we can make the point much more clearly.

```
cases_vs_mobility %>%
  ggplot(aes(x = lga_name, y = mean_mobility_change)) +
  geom_col()
```



Column graphs like this are very easy to read and understand. This one could do with some work.

Challenge: Create this plot



Interpreting our results

Before we go to writing our report, it is good practice to think about the ethical dilemma of this situation.

✅ We have a deidentified, aggregated dataset, which we have further aggregated

✅ We have filtered the dataset to only "high" numbers of cases

But our conclusions here have real-world impacts: People in a position of power are relying on this evidence to decide whether to increase lockdowns.

One doesn't have to look far to see when this can go wrong

This is a different type of ethics to what we've discussed previously, but is an important part of generating new knowledge.

An ethics committee will only approve research if **the potential benefits of the project justify any risks**.

Do you think that is the case here?

Interpreting our results

What is the conclusion here?

We have *some* (weak) evidence that mobility is associated with lower cases.

But there is little to no evidence that further decreasing mobility will decrease cases more.

Why?

Conclusion

Challenge for 5 minutes: Write a two sentence executive summary outlining what you think the findings are.

Then, swap with someone else in the room and see if you agree.

Conclusion and communication

The reality is, most stakeholders won't read past your executive summary and a key picture.

So, **spend time on the things that matter**!

Here, we set out to think about what COVID measures may or may not further impact transmission. We have *answered what probably won't work* but not provided any solutions.

- In most cases this will get sent back with a question: "So what should we do?"

How do you deal with that?

Keep looking!

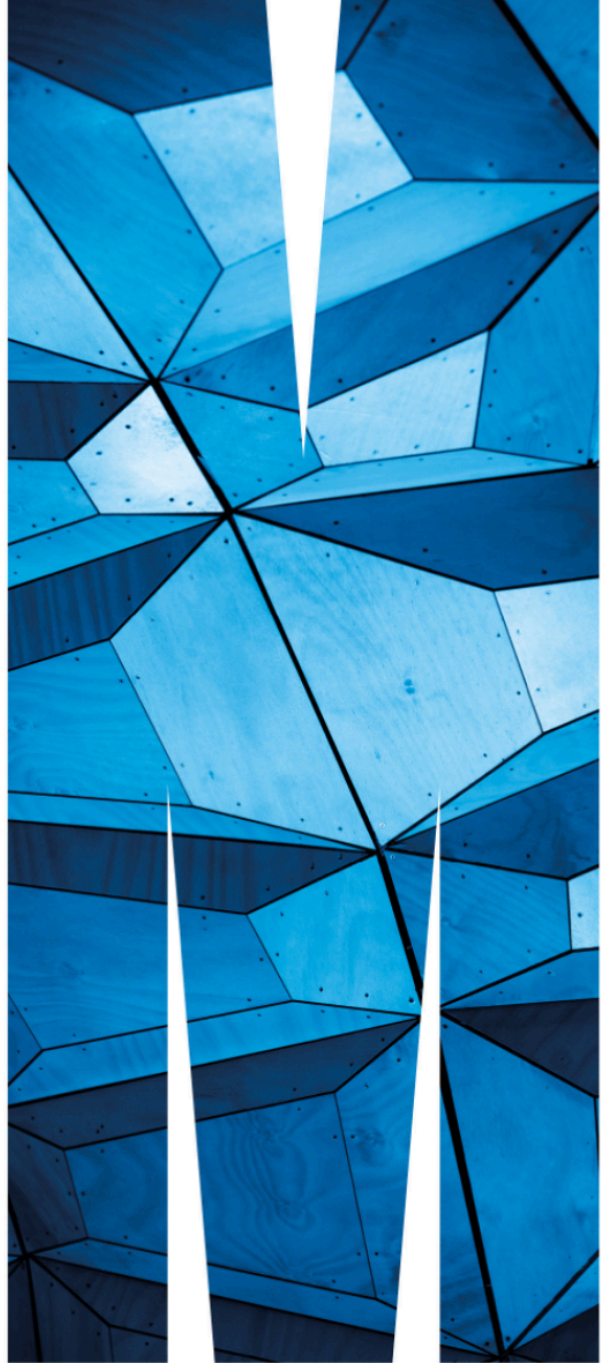
More exploration will hopefully lead to more possibilities.

Summary

- We looked at using open data to answer a specific question
- We generated evidence *against* a specific policy
- We thought about how to communicate this, both written and visually
- We learnt about the plight of no standards being adhered to

Slides created by Dr Michael Lydeamore

MONASH University



Lecturer: *Michael Lydeamore*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 10