ETC5521: Exploratory Data Analysis

Exploring data having a space and time context

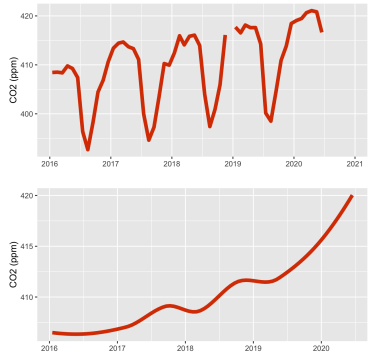
Lecturer: *Di Cook*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 9 - Session 1

What is temporal data?

🕘 Remember the pedestrian sensor data

Sensor	Date_Time	Date	Time	Count
Birrarung Marr	*2015-02-14 22:00:00*	2015-02-14	22	7081
Birrarung Marr	*2015-02-21 21:00:00*	2015-02-21	21	8363
Birrarung Marr	*2015-02-21 22:00:00*	2015-02-21	22	9658
Birrarung Marr	*2015-02-21 23:00:00*	2015-02-21	23	10121
Birrarung Marr	*2015-02-22 00:00:00*	2015-02-22	0	8441
Birrarung Marr	*2015-03-07 20:00:00*	2015-03-07	20	7144
Birrarung Marr	*2015-03-07 21:00:00*	2015-03-07	21	7238
Birrarung Marr	*2015-03-08 13:00:00*	2015-03-08	13	7092
Birrarung Marr	*2015-03-08 14:00:00*	2015-03-08	14	7031
Birrarung Marr	*2015-03-08 15:00:00*	2015-03-08	15	6951
Birrarung Marr	*2015-03-08 16:00:00*	2015-03-08	16	7167
Birrarung Marr	*2015-03-08 17:00:00*	2015-03-08	17	7246

What is temporal data?

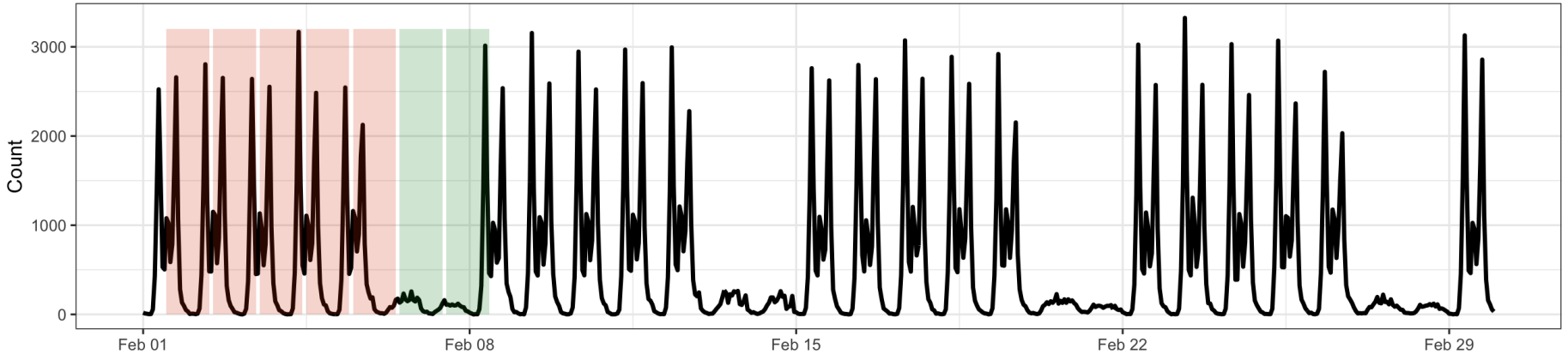


- Temporal data has date/time/ordering index variable, call it time.
- A time variable has special structure:
  - it can have *cyclical* patterns, eg seasonality (summer, winter), an over in cricket
  - the cyclical patterns can be *nested*, eg postcode within state, over within innings
- Measurements are also NOT independent - yesterday may influence today.
- It still likely has non-cyclical patterns, trends and associations with other variables, eg temperature increasing over time, over is bowled by Elise Perry or Sophie Molineaux

Case study 1 Melbourne pedestrian traffic

learn R

Pedestrian counts at Southern Cross in Feb 2016

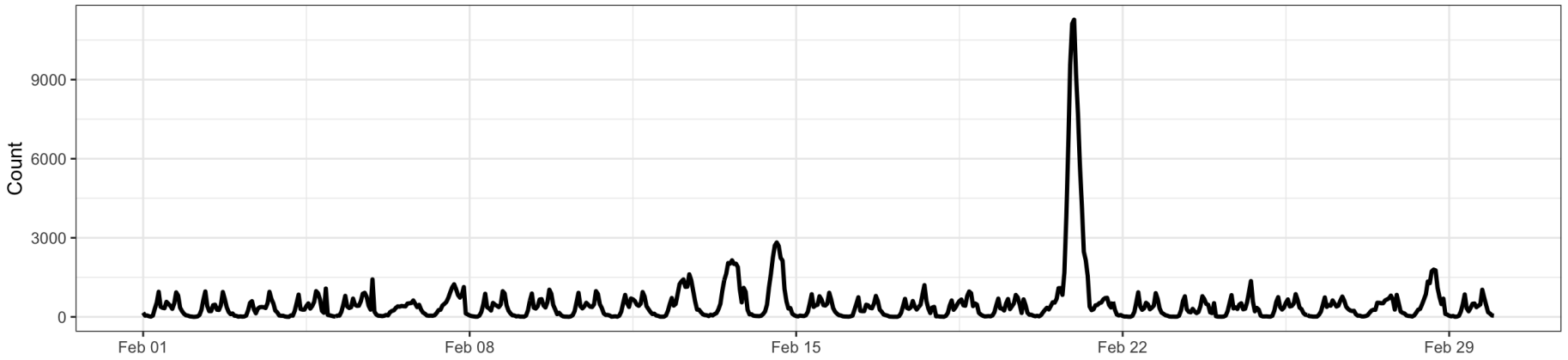


This is interesting!

Case study 1 Melbourne pedestrian traffic

learn R

Pedestrian counts at Birrarung Marr in Feb 2016

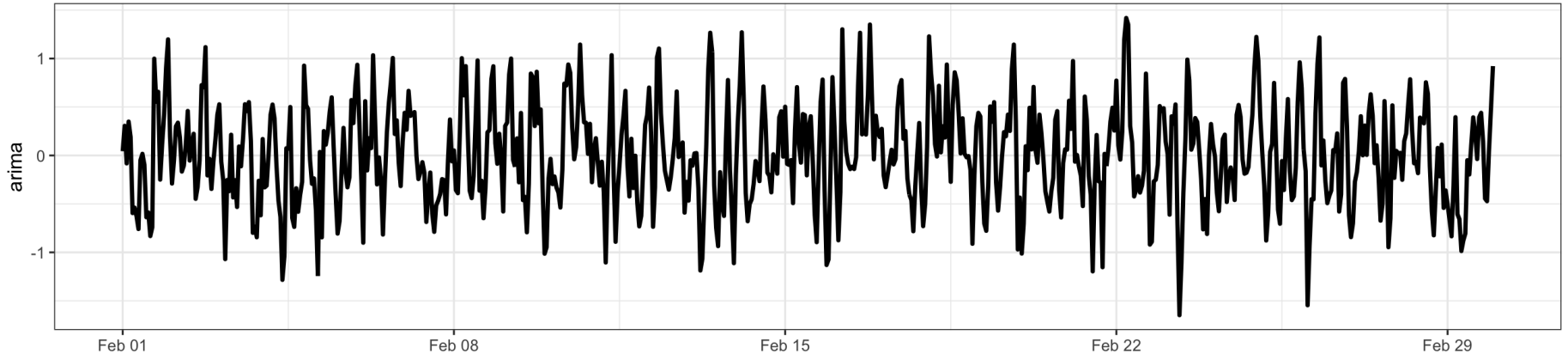


This is interesting!

Case study 1 Melbourne pedestrian traffic

learn R

What does Heike mean?



This is a little bit boring! It is important for fitting a model that accounts for dependencies between measurements, though.

Exploratory analysis of temporal data is interested in extracting the trend and general patterns.

What is exploratory analysis of time series?

Exploratory analysis of time series investigates trends, patterns, cyclical, nested cyclical, temporal outliers, and temporal dependence.

For the pedestrian sensor data this is:

- work day vs holiday pattern
- daily patterns
- weather and season related changes
- event related patterns

Regular vs irregular

The Melbourne pedestrian sensor data has a regular period. Counts are provided for every hour, at numerous locations.

In contrast, the US flights data, below, is irregular.

```
## # A tsibble: 336,776 x 20 [!] <UTC>
## # Key:       origin, dest, carrier, tailnum [52,807]
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1    30     2224           2000       144     2316           2101
##  2  2013     2    17     2012           2010         2     2120           2114
##  3  2013     2    26     2356           2000       236       41           2104
##  4  2013     3    13     1958           2005        -7     2056           2109
##  5  2013     5    16     2214           2000       134     2307           2112
##  6  2013     5    30     2045           2000        45     2141           2112
##  7  2013     9    11     2254           2159        55     2336           2303
##  8  2013     9    12       NA           2159        NA       NA           2303
##  9  2013     9     8     2156           2159        -3     2250           2303
## 10  2013     1    26     1614           1620        -6     1706           1724
## # … with 336,766 more rows
```

question discussion

Is pedestrian traffic regular, really?

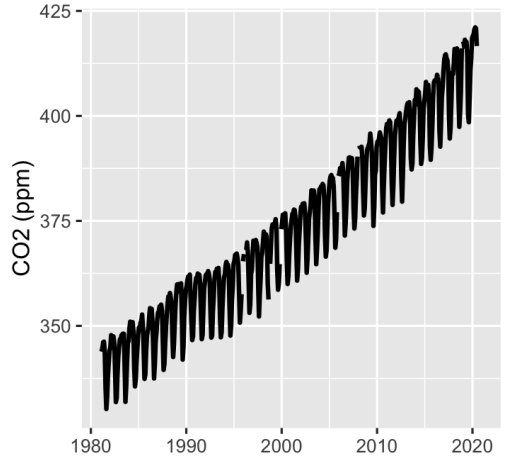
Plotting temporal data

- lines: connecting sequential time points indicates the temporal dependence is important
- aspect ratio: wide or tall? Cleveland, McGill, McGill (1988) argue the average line slope in a line chart should be 45 degrees, which is called banking to 45 degrees. But this is refuted in Talbot, Gerth, Hanrahan (2012) that the conclusion was based on a flawed study. Nevertheless, aspect ratio is an inescapable skill for designing effective plots. For time series, typically a wide aspect ratio is good.
- conventions:
  - time on the horizontal axis,
  - ordering of elements like week day, month.

Aspect ratio

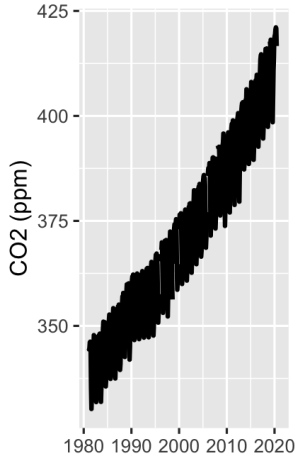
learn R

1 to 1 (may be useless)

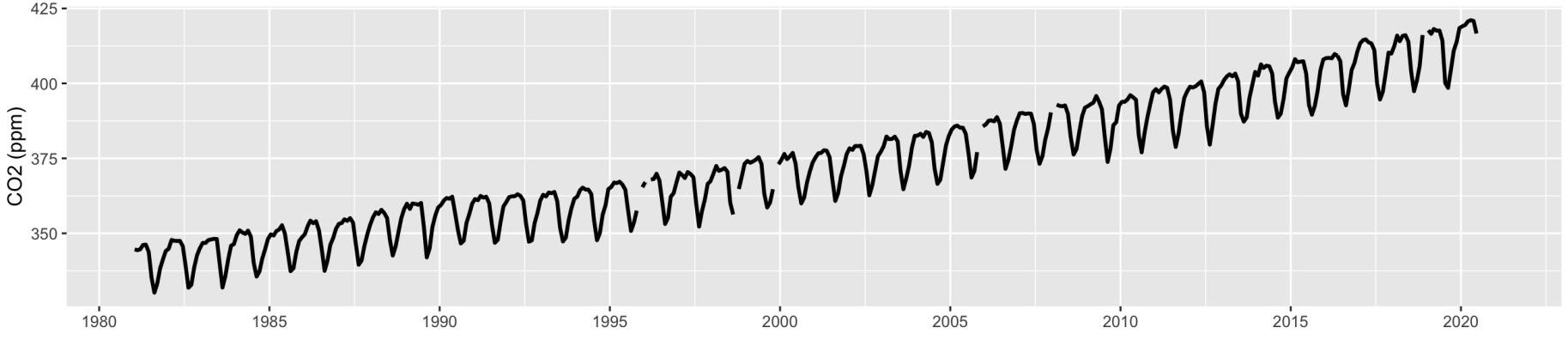


CO2 at
Point Barrow,
Alaska

tall & skinny: trend



short & wide: seasonality



Case study 2 nycflights13 Part 1/7

```
library(nycflights13)
```

What is a useful time element to use, in order to study traffic over time?

Hour, 15 minutes, day, month?

Possibly, all of these.

Let's start with hourly.

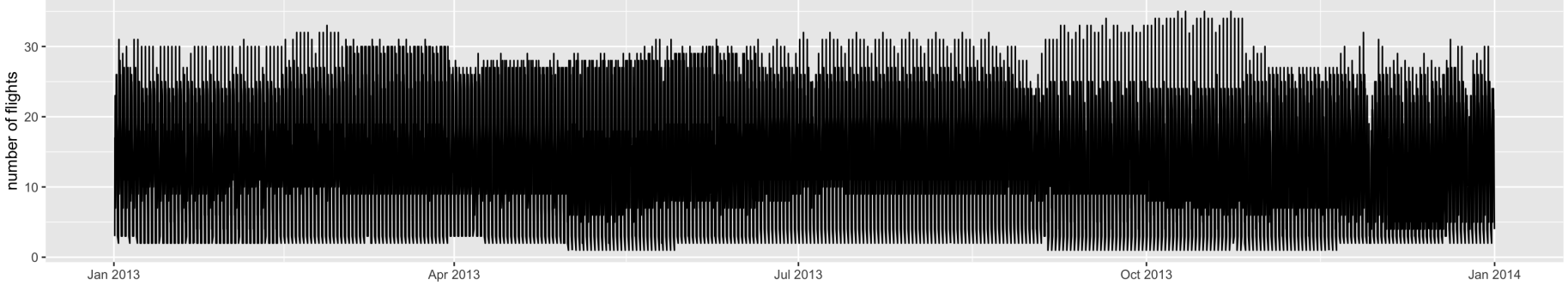
```
flights_hourly <- flights %>%
  group_by(time_hour, origin) %>%
  summarise(count = n(),
    dep_delay = mean(dep_delay,
                     na.rm = TRUE)) %>%
  ungroup() %>%
  as_tsibble(index = time_hour,
             key = origin)
flights_hourly
```

```
## # A tsibble: 19,486 x 4 [1h] <America/New_York>
## # Key:       origin [3]
##    time_hour           origin count dep_delay
##    <dttm>              <chr>  <int>     <dbl>
##  1 2013-01-01 05:00:00 EWR        2     -1
##  2 2013-01-01 06:00:00 EWR       18      3.06
##  3 2013-01-01 07:00:00 EWR       12     14.2
##  4 2013-01-01 08:00:00 EWR       20      0.75
##  5 2013-01-01 09:00:00 EWR       19      9.05
##  6 2013-01-01 10:00:00 EWR       18      2.06
##  7 2013-01-01 11:00:00 EWR       11      0
```

Case study 2 nycflights13 Part 2/7

Pick one airport, and examine the hourly number of flights.

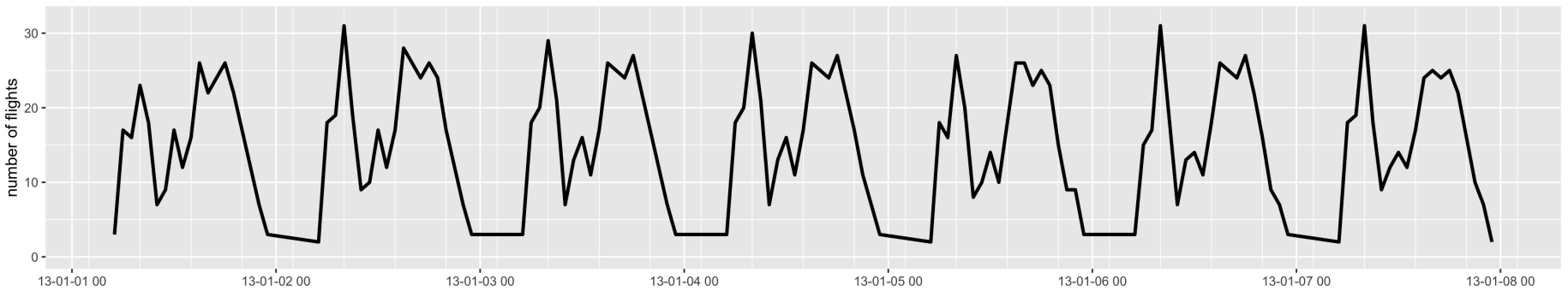
```
flights_hourly %>%
  filter(origin == "JFK") %>%
  ggplot(aes(x=time_hour, y=count)) +
  geom_line() +
  xlab("") + ylab("number of flights")
```



No, that's too much information, too much time. There's no overall trend. Not an interesting plot.

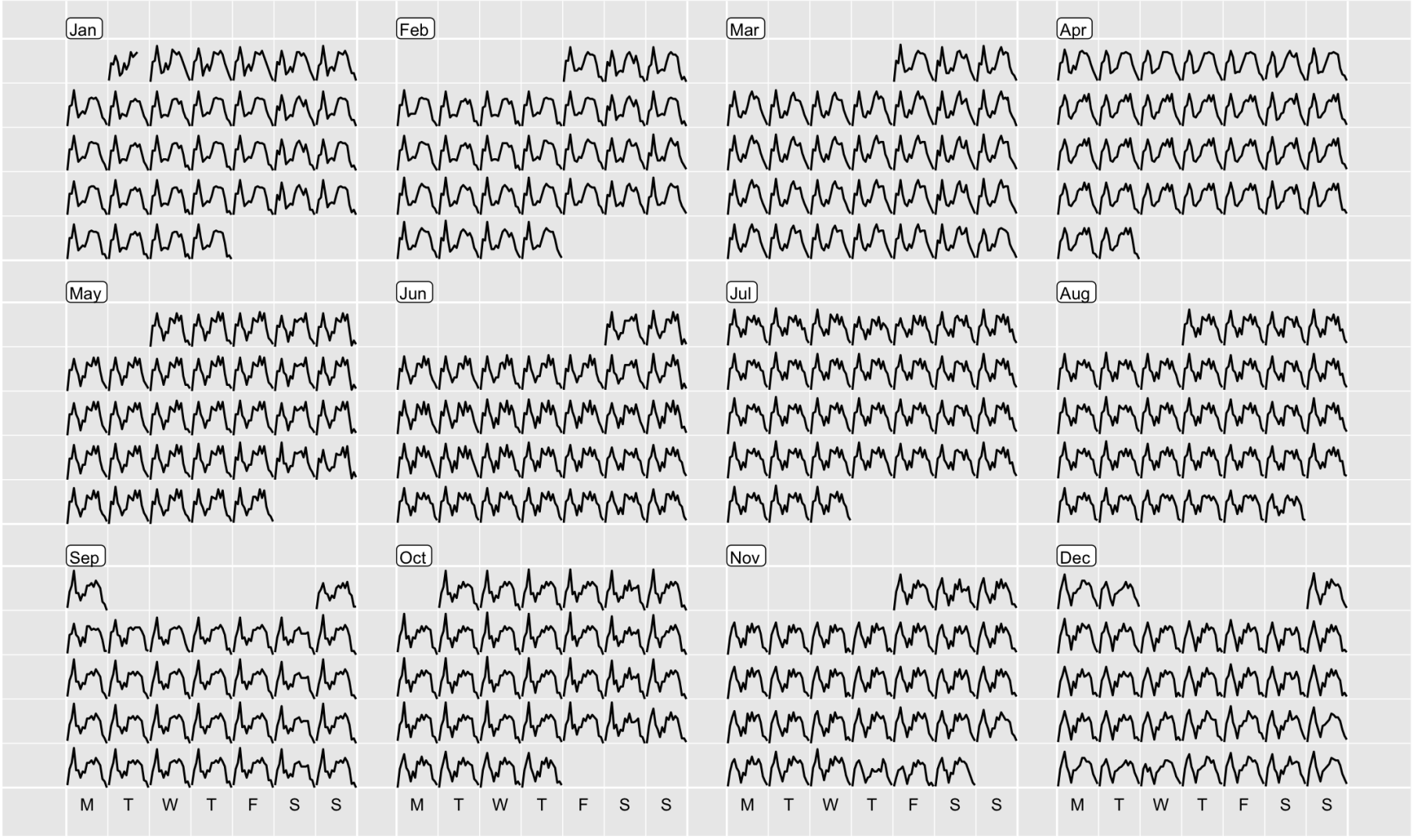
Reduce the time frame to check for periodicities

```
flights_hourly %>%
  filter(origin == "JFK",
         time_hour < ymd("2013-01-08")) %>%
  ggplot(aes(x=time_hour, y=count)) +
    geom_line(size=1.1) +
    scale_x_datetime("",
                     date_breaks = "1 day",
                     date_labels = "%y-%m-%d %H",
                     date_minor_breaks = "6 hours") +
    ylim(c(0, 32)) +
    xlab("") + ylab("number of flights")
```



Case study 2 nycflights13 Part 4/7

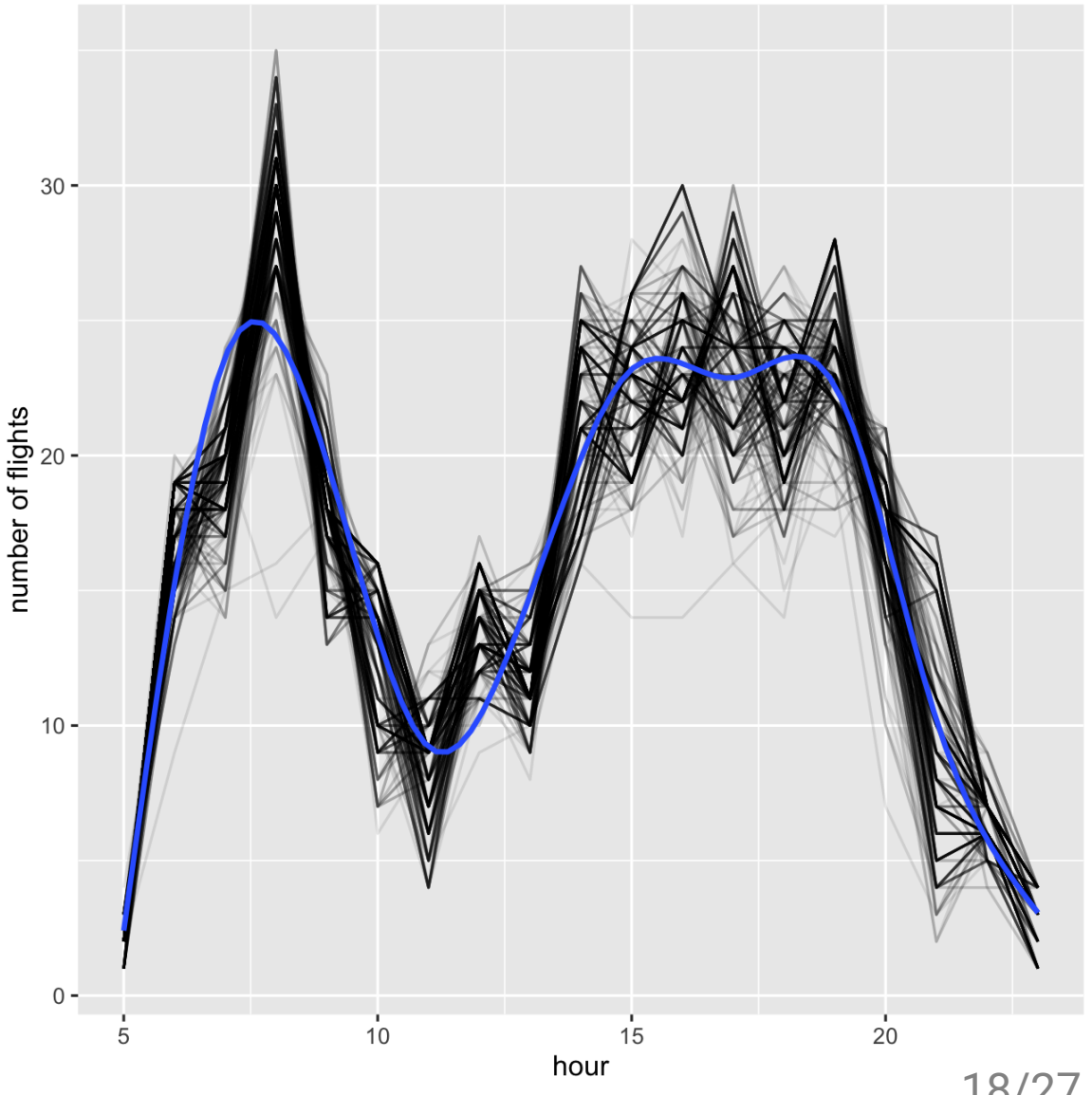
learn R



Case study 2 nycflights13 Part 5/7

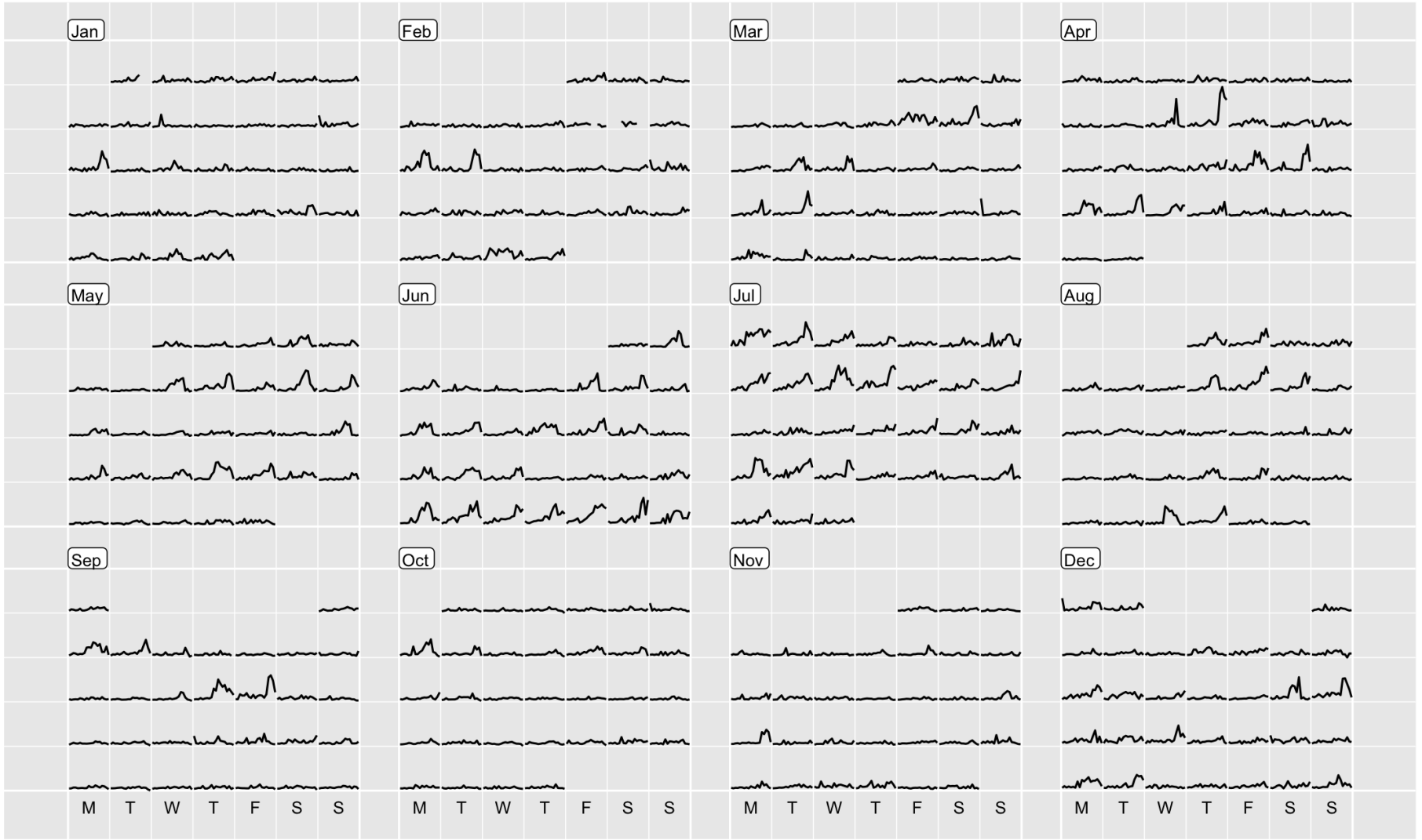
```
flights_hourly %>%
  filter(origin == "JFK") %>%
  mutate(month = month(time_hour),
         hour = hour(time_hour),
         date = as.Date(time_hour)) %>%
  ggplot(aes(x=hour, y=count)) +
    geom_line(aes(group=date), alpha = 0.1) +
    geom_smooth(se = FALSE) +
    xlab("hour") + ylab("number of flights")
```

Its very regular. Does this make sense?



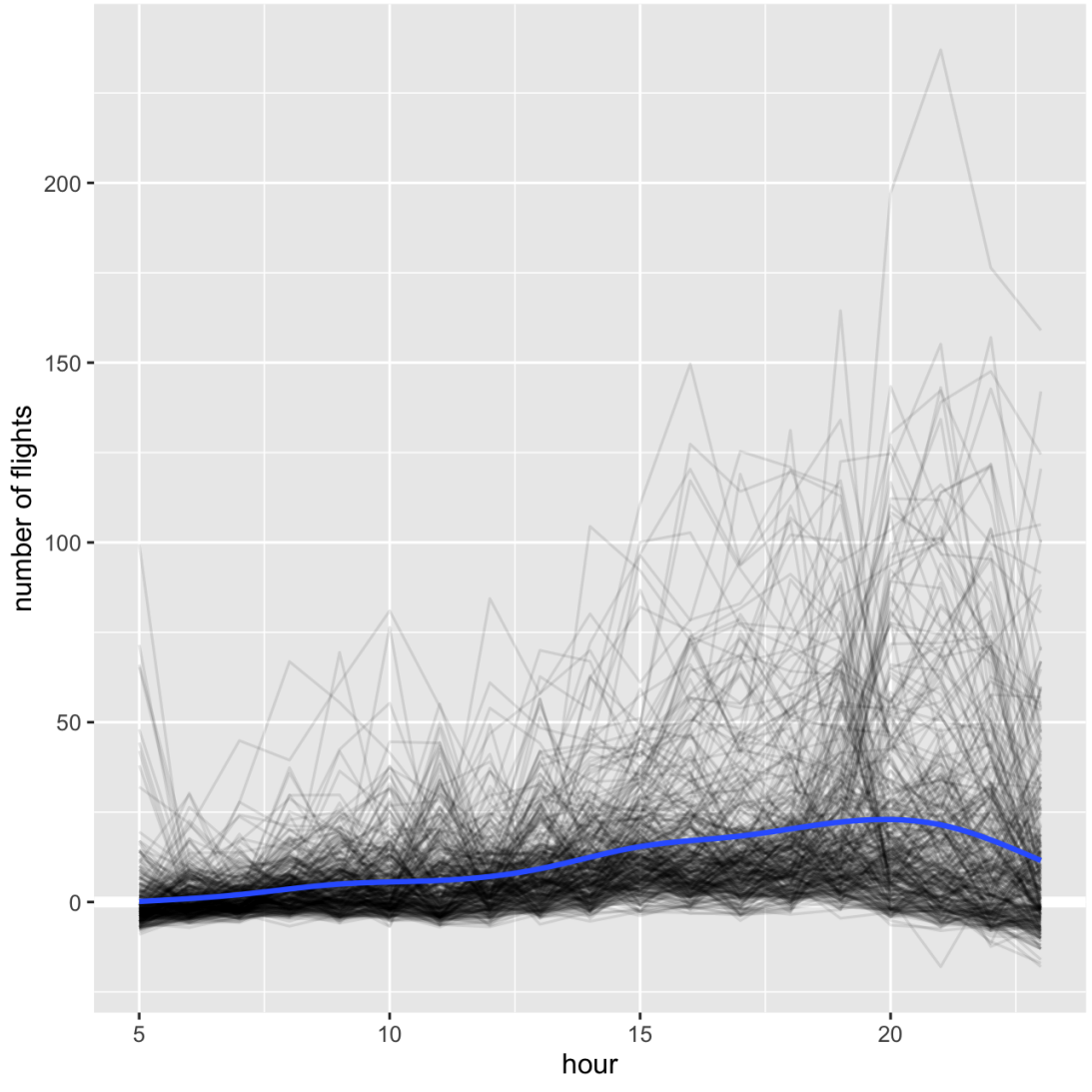
Case study 2 nycflights13 Part 6/7

learn R



Case study 2 nycflights13 Part 7/7

```
flights_hourly %>%
  filter(origin == "JFK") %>%
  mutate(month = month(time_hour),
         hour = hour(time_hour),
         date = as.Date(time_hour)) %>%
  ggplot(aes(x=hour, y=dep_delay)) +
    geom_hline(yintercept=0, colour="white", size=2) +
    geom_line(aes(group=date), alpha = 0.1) +
    geom_smooth(se=FALSE) +
    xlab("hour") + ylab("number of flights")
```



- Delays tend to pile up as the day goes on.
- Also notice a lot of flights depart a few minutes early.
- This data is harder to model and forecast.

Summary: Melting time

- The structure of the airlines data is very useful. Date-time has already been broken out into: `year`, `month`, `day`, `hour`, `minute`.
- There are also several possible key variables: `origin`, `carrier`, `tailnum`.

Why isn't `dest` considered a key variable? Why not have `air_time` as a key variable?

- Aggregate by temporal components, in different ways to explore different patterns of variables in relation to elements of time.

01:00

Interactive exploration with tsibbletalk

🔧 Your turn, cut and paste the code into your R console. Click on the tree, click on a point, line, ...

```
tourism_feat <- tourism_shared %>%
  features(Trips, feat_stl)

p1 <- tourism_shared %>%
  ggplot(aes(x = Quarter, y = Trips)) +
  geom_line(aes(group = Region), alpha = 0.5) +
  facet_wrap(~ Purpose, scales = "free_y")
p2 <- tourism_feat %>%
  ggplot(aes(x = trend_strength, y = seasonal_strength_year)) +
  geom_point(aes(group = Region))

library(plotly)
subplot(p0,
  subplot(
    ggplotly(p1, tooltip = "Region", width = 900),
    ggplotly(p2, tooltip = "Region", width = 900),
    nrows = 2),
  widths = c(.4, .6)) %>%
  highlight(dynamic = TRUE)
```

🔧 Your turn, cut and paste the code into your R console. Drag the scroll bar to wrap the series on itself.

```
p <- fill_gaps(pedestrian) %>%
  filter_index(~ "2015") %>%
  ggplot(aes(x = Date_Time, y = Count, colour = Sensor)) +
  geom_line(size = .2) +
  facet_wrap(~ Sensor, scales = "free_y") +
  theme(legend.position = "none")

library(shiny)
ui <- fluidPage(tsibbleWrapUI("tswrap"))
server <- function(input, output, session) {
  tsibbleWrapServer("tswrap", p, period = "1 day")
}
shinyApp(ui, server)
```

A step back in time, classic almost periodic data

See if you can get the peaks to match up with wrapping.

Annual numbers of lynx trappings for 1821–1934 in Canada. Almost 10 year cycle.

```
lynx_tsb <- as_tsibble(lynx) %>%
  rename(count = value)
pl <- ggplot(lynx_tsb, aes(x = index, y
  geom_line(size = .2)

ui <- fluidPage(tsibbleWrapUI("tswrap")
server <- function(input, output, sessi
  tsibbleWrapServer("tswrap", pl, peric
}
shinyApp(ui, server)
```

Monthly mean relative sunspot numbers from 1749 to 1983. Almost 10 year cycle.

```
sunspots_tsb <- as_tsibble(sunspots) %>
  rename(count = value)
pl <- ggplot(sunspots_tsb, aes(x = inde
  geom_line(size = .2)

ui <- fluidPage(tsibbleWrapUI("tswrap")
server <- function(input, output, sessi
  tsibbleWrapServer("tswrap", pl, peric
}
shinyApp(ui, server)
```

That's it, for this lecture!

Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu