CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Cardiff Business School
Ysgol Busnes Caerdydd

Forecasting in R

Prepare and transform data

Bahman Rostami-Tabar

Outline

1. Learning outcomes
2. Time series in R
3. Example: create and work with `tsibble`
4. Lab Session 1

Outline

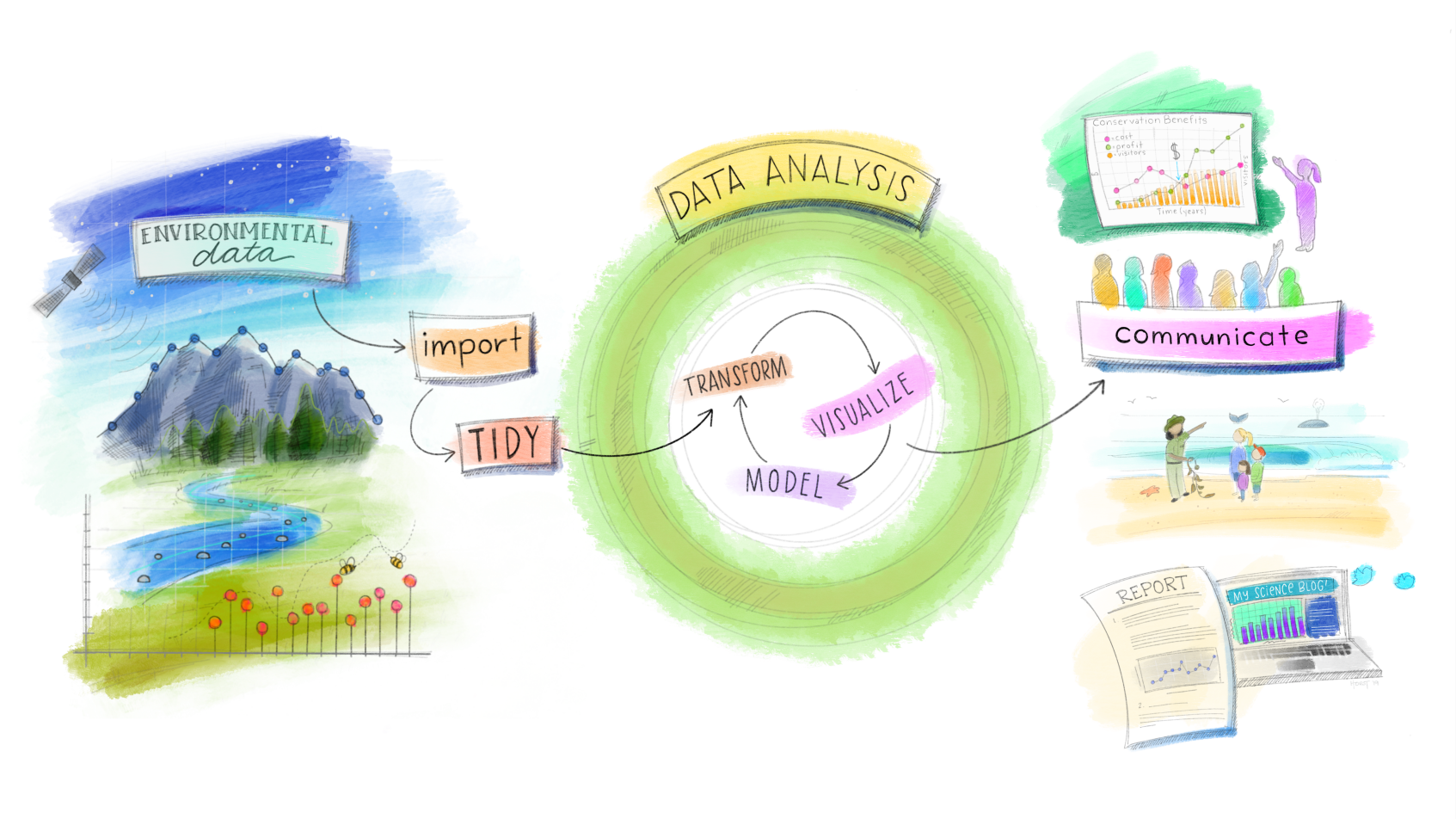
Learning outcomes

You should be able to:

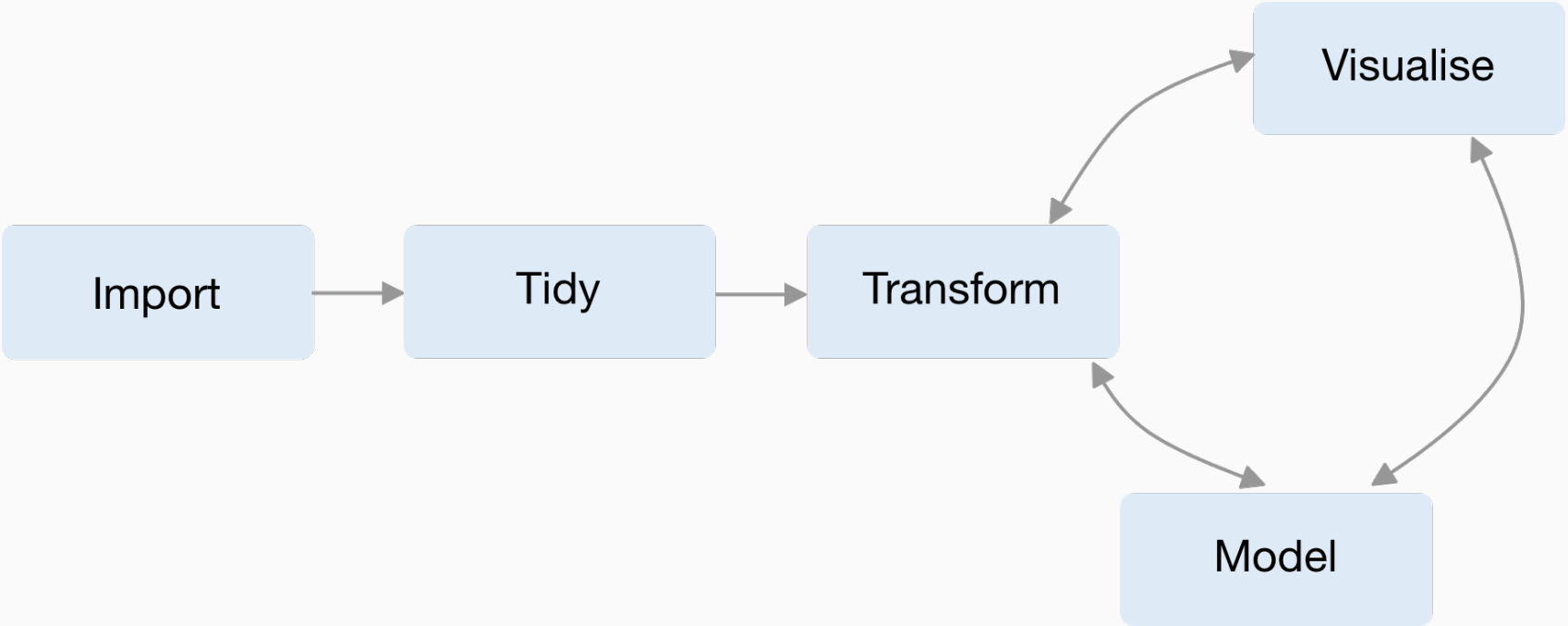
1. Convert any given data into a `tsibble` object
2. Prepare data for analysis using `tsibble` functions
3. Work with `tsibble` using tidyverse functions

Outline

Tidyverse


ENVIRONMENTAL data
import
TIDY
DATA ANALYSIS
TRANSFORM
VISUALIZE
MODEL
Conservation Benefits
Time (years)
Communicate
REPORT
My Science BLOG!

Tidyverse

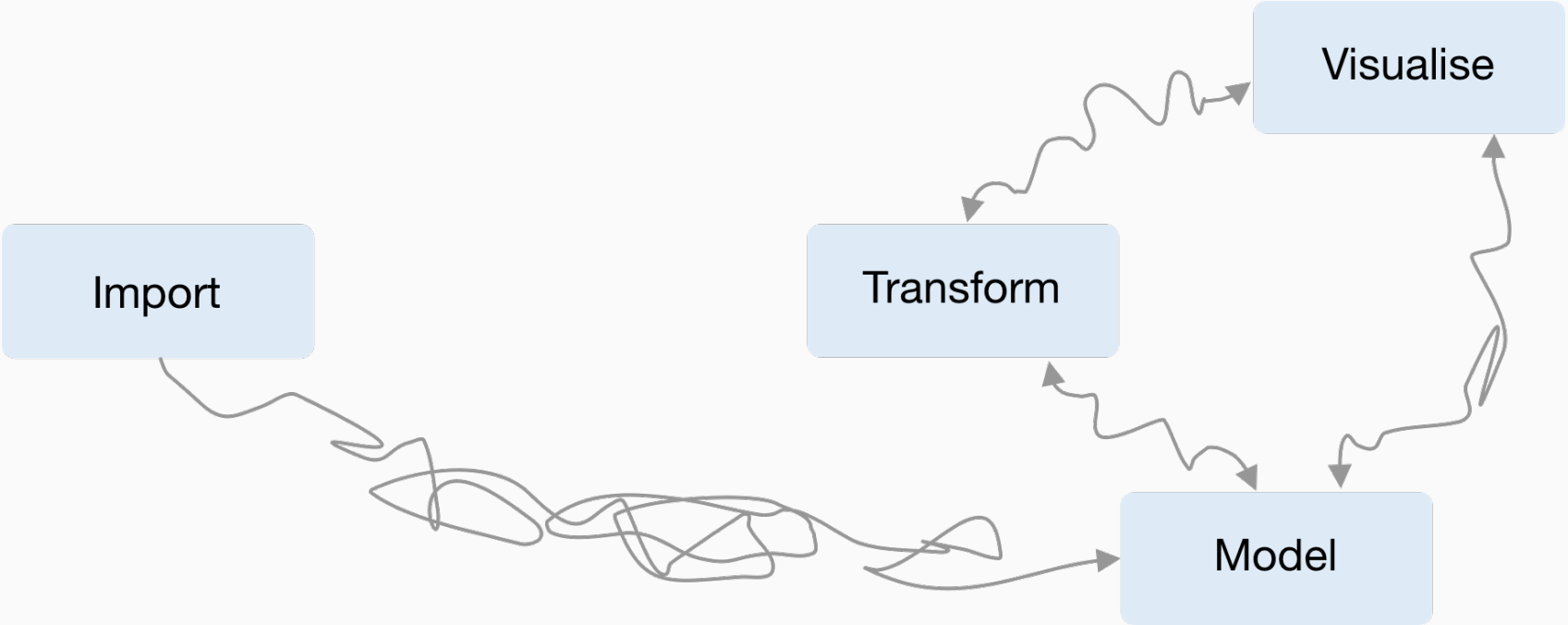


Features of data

- heterogeneous data types
- irregular time interval
- multiple measured variables
- multiple grouping variables

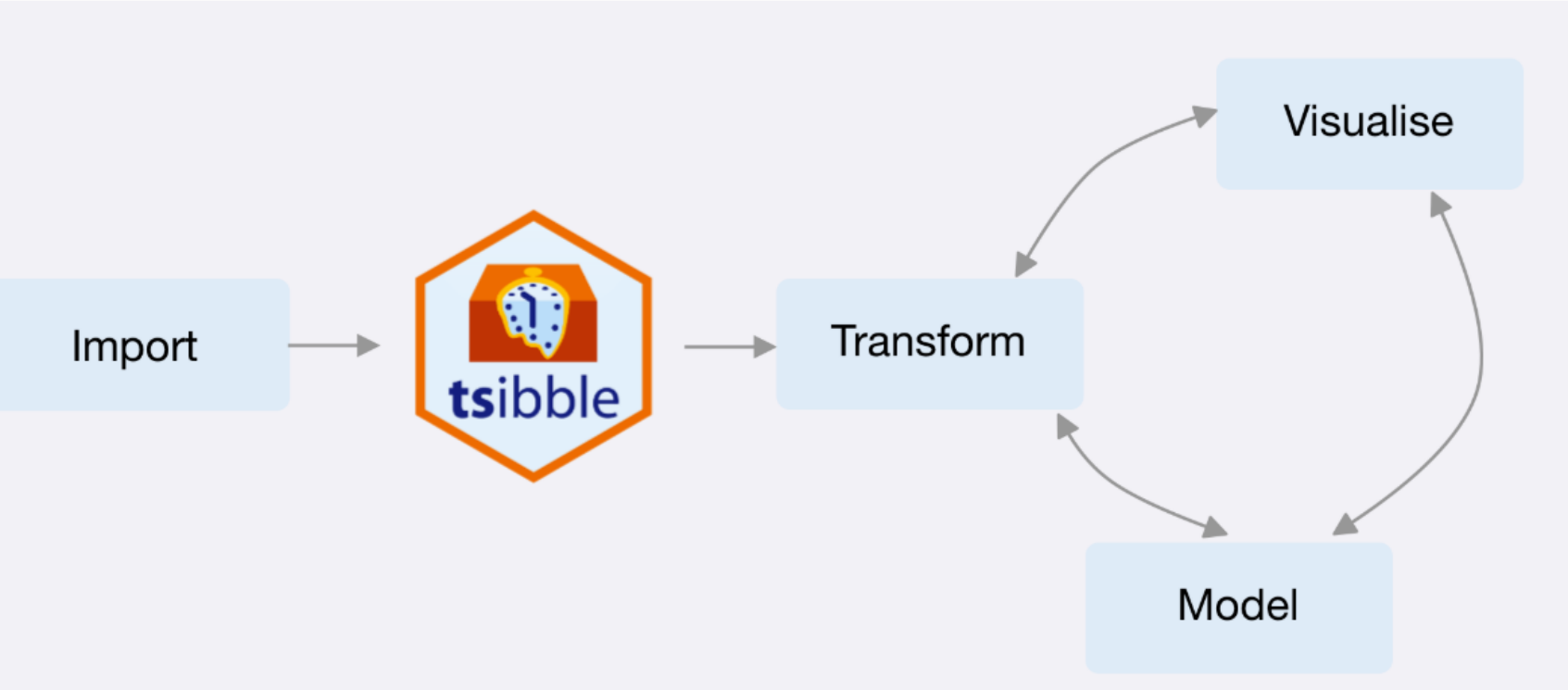
Time series verse

- does not work with `ts()`, `zoo()`, `xts()`, `etc`
- difficult to work with tidyverse



Tsibble

It defines tidier data for temporal analysis



Time series

A `time series` can be thought of as a list of numbers (the measurements), along with some information about what times those numbers were recorded (the index). This information can be stored as a tsibble object in R.

`tsibble` objects

In tsibble:

- An index: time information about the observation
- Measured variable(s): numbers of interest
- Key variable(s): set of variables that define observational units over time
- It works with tidyverse functions.

The `tsibble` index

Common time index variables can be created with these functions:

Frequency	Function
Annual	start:end
Quarterly	yearquarter()
Monthly	yearmonth()
Weekly	yearweek()
Daily	as_date(), ymd()
Sub-daily	as_datetime()

Example: ts object

```
USAccDeaths
```

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul
## 1973  9007  8106  8928  9137 10017 10826 11317
## 1974  7750  6981  8038  8422  8714  9512 10120
## 1975  8162  7306  8124  7870  9387  9556 10093
## 1976  7717  7461  7767  7925  8623  8945 10078
## 1977  7792  6957  7726  8106  8890  9299 10625
## 1978  7836  6892  7791  8192  9115  9434 10484
##        Aug   Sep   Oct   Nov   Dec
## 1973 10744  9713  9938  9161  8927
## 1974  9823  8743  9129  8710  8680
## 1975  9620  8285  8466  8160  8034
## 1976  9179  8037  8488  7874  8647
## 1977  9302  8314  8850  8265  8796
## 1978  9827  9110  9070  8633  9240
```

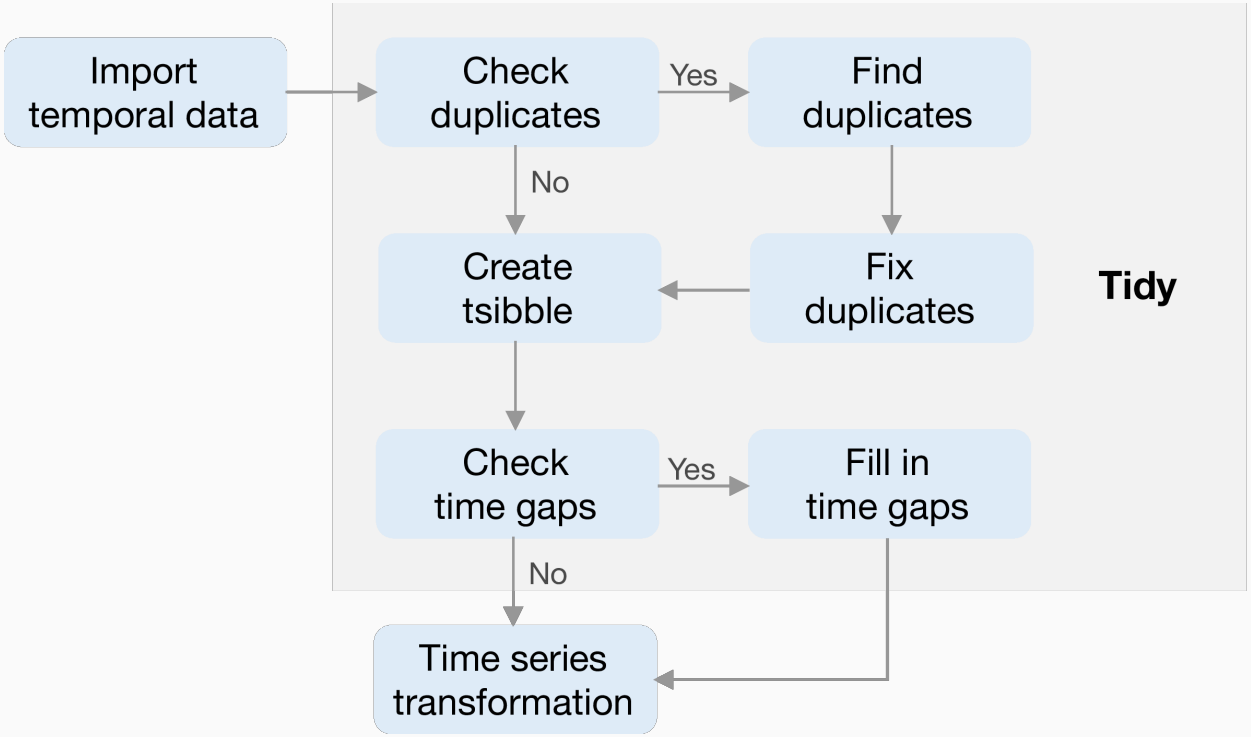
Convert `ts` to `tsible` object

```r
USAccDeaths_ts <- USAccDeaths %>% as_tsibble()
```

```
## # A tsibble: 72 x 2 [1M]
##       index value
##       <mth> <dbl>
##  1 1973 Jan  9007
##  2 1973 Feb  8106
##  3 1973 Mar  8928
##  4 1973 Apr  9137
##  5 1973 May 10017
##  6 1973 Jun 10826
##  7 1973 Jul 11317
##  8 1973 Aug 10744
##  9 1973 Sep  9713
## 10 1973 Oct  9938
## # ... with 62 more rows
```

Outline

Steps to create a tsibble



Read a csv file

quarterly overnight trips across Australia

```
tourism <- readxl::read_excel("data/tourism.xlsx")
```

```
## # A tibble: 24,320 x 5
##    Quarter    Region   State         Purpose Trips
##    <chr>      <chr>    <chr>         <chr>   <dbl>
##  1 1998-01-01 Adelaide South Austra~ Busine~  135.
##  2 1998-04-01 Adelaide South Austra~ Busine~  110.
##  3 1998-07-01 Adelaide South Austra~ Busine~  166.
##  4 1998-10-01 Adelaide South Austra~ Busine~  127.
##  5 1999-01-01 Adelaide South Austra~ Busine~  137.
##  6 1999-04-01 Adelaide South Austra~ Busine~  200.
##  7 1999-07-01 Adelaide South Austra~ Busine~  169.
##  8 1999-10-01 Adelaide South Austra~ Busine~  134.
##  9 2000-01-01 Adelaide South Austra~ Busine~  154.
## 10 2000-04-01 Adelaide South Austra~ Busine~  169.
## # ... with 24,310 more rows
```

Check duplicates

```r
tourismd <- tourism %>% duplicated()
sum(tourismd)
```

```
## [1] 0
```

```r
#are_duplicated()
#tourism %>% distinct()
```

Change index to `yearquarter`

```r
tourism <- tourism %>%
  mutate(Quarter = yearquarter(Quarter))
```

```
## # A tibble: 24,320 x 5
##    Quarter Region   State           Purpose  Trips
##      <qtr> <chr>    <chr>           <chr>    <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

Craete a tsibble

```r
tourism <- tourism %>%
  as_tsibble(
    index = Quarter,
    key = c(Region, State, Purpose)
  )
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
##      <qtr> <chr>    <chr>           <chr>    <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
```

Check gaps

```r
tourism %>% has_gaps()
tourism %>% count_gaps()
tourism %>% scan_gaps()
tourism %>% fill_gaps(Trips=0L)
```

tsibble objects

```
tourism
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
##      <qtr> <chr>    <chr>           <chr>    <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

tsibble objects

```
tourism
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
##    Index   <chr>    <chr>           <chr>    <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

tsibble objects

```
tourism
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
##    Index   Keys                                <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

tsibble objects

```
tourism
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
```

Index | Keys | Measure

```
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

tsibble objects

```
tourism
```

```
## # A tsibble: 24,320 x 5 [1Q]
## # Key:       Region, State, Purpose [304]
##    Quarter Region   State           Purpose  Trips
##    Index   Keys                                Measure
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Busin
##  4 1998 Q4 Adelaide South Australia Busin
##  5 1999 Q1 Adelaide South Australia Busin
##  6 1999 Q2 Adelaide South Australia Business
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 24,310 more rows
```

Domestic visitor nights in thousands by state/region and purpose.

Working with `tsibble` objects

We can use the `filter()` function to select rows.

```
tourism %>%
  filter(Purpose == "Business")
```

```
## # A tsibble: 6,080 x 5 [1Q]
## # Key:       Region, State, Purpose [76]
##    Quarter Region   State           Purpose  Trips
##      <qtr> <chr>    <chr>           <chr>    <dbl>
##  1 1998 Q1 Adelaide South Australia Business  135.
##  2 1998 Q2 Adelaide South Australia Business  110.
##  3 1998 Q3 Adelaide South Australia Business  166.
##  4 1998 Q4 Adelaide South Australia Business  127.
##  5 1999 Q1 Adelaide South Australia Business  137.
##  6 1999 Q2 Adelaide South Australia Business  200.
##  7 1999 Q3 Adelaide South Australia Business  169.
##  8 1999 Q4 Adelaide South Australia Business  134.
##  9 2000 Q1 Adelaide South Australia Business  154.
## 10 2000 Q2 Adelaide South Australia Business  169.
## # ... with 6,070 more rows
```

Working with `tsibble` objects

We can use the `select()` function to select columns.

```
tourism %>%
  filter(Purpose == "Business") %>%
  select(Region,Trips)
```

```
## Selecting index: "Quarter"

## # A tsibble: 6,080 x 3 [1Q]
## # Key:       Region [76]
##    Region   Trips Quarter
##    <chr>    <dbl>   <qtr>
##  1 Adelaide  135. 1998 Q1
##  2 Adelaide  110. 1998 Q2
##  3 Adelaide  166. 1998 Q3
##  4 Adelaide  127. 1998 Q4
##  5 Adelaide  137. 1999 Q1
##  6 Adelaide  200. 1999 Q2
##  7 Adelaide  169. 1999 Q3
##  8 Adelaide  134. 1999 Q4
```

Working with `tsibble` objects

- We can use `group_by()` function to group over keys.
  - We can also do it with: `group_by_key()`
- We can use the `summarise()` function to summarise over keys.

```
tourism %>%
  group_by(Region, Purpose) %>%
  summarise(Trips = mean(Trips)) %>%
  ungroup()
```

```
## # A tsibble: 24,320 x 4 [1Q]
## # Key:       Region, Purpose [304]
##    Region   Purpose  Quarter Trips
##    <chr>    <chr>      <qtr> <dbl>
##  1 Adelaide Business 1998 Q1  135.
##  2 Adelaide Business 1998 Q2  110.
##  3 Adelaide Business 1998 Q3  166.
##  4 Adelaide Business 1998 Q4  127.
```

Working with `tsibble` objects

- We can use `index_by()` function to group over index
- We can use the `summarise()` function to summarise over index.

```
tourism %>%
index_by(Quarter) %>%
  summarise(total_trips = sum(Trips))
```

```
## # A tsibble: 80 x 2 [1Q]
##    Quarter total_trips
##      <qtr>       <dbl>
##  1 1998 Q1      23182.
##  2 1998 Q2      20323.
##  3 1998 Q3      19827.
##  4 1998 Q4      20830.
##  5 1999 Q1      22087.
##  6 1999 Q2      21458.
##  7 1999 Q3      19914.
##  8 1999 Q4      20028.
```

Working with `tsibble` objects

We can use the `mutate()` function to create new variables.

```
tourism %>%
  mutate(year = year(Quarter)) -> m1
```

```
## # A tsibble: 24,320 x 6 [1Q]
## # Key:       Region, State, Purpose [304]
##      Quarter Region  State     Purpose Trips  year
##        <qtr> <chr>   <chr>     <chr>   <dbl> <dbl>
##  1   1998 Q1 Adelai~ South Au~ Busine~  135.  1998
##  2   1998 Q2 Adelai~ South Au~ Busine~  110.  1998
##  3   1998 Q3 Adelai~ South Au~ Busine~  166.  1998
##  4   1998 Q4 Adelai~ South Au~ Busine~  127.  1998
##  5   1999 Q1 Adelai~ South Au~ Busine~  137.  1999
##  6   1999 Q2 Adelai~ South Au~ Busine~  200.  1999
##  7   1999 Q3 Adelai~ South Au~ Busine~  169.  1999
##  8   1999 Q4 Adelai~ South Au~ Busine~  134.  1999
##  9   2000 Q1 Adelai~ South Au~ Busine~  154.  2000
## 10   2000 Q2 Adelai~ South Au~ Busine~  169.  2000
## #     with 24,310 more rows
```

Outline

Before the lab

- Open RStudio and Create a project
- you can create a new RScript

You may also want to use RMarkdown:

- Create a new RMarkdown file, save it
- Delete the template after the setup r chunk
- Create the first section using `## Prepare data`
- create your first **r chunk**

Lab Session 1

1. Read [`ae_uk.csv`] into R
2. Check duplications!
3. Create a tsibble object! Is the `index` a regular interval?
4. Create a new tsibble which has a regular interval of `1 hour`, and has total attendance per hour for the combination of gender and injury_type.
5. Is there any gap in data? you can use `has_gaps()`, `count_gaps()` and `scap_gaps()`
6. How can we regularise an irregular index in tsibble?