Maren Rieker &
Kabir Sandrolini

Outline

1. Brief history of tsibble
2. Quick recap of time series data
3. Basic structure and usage
4. Working with tsibble
5. The Tidyverts ecosystem and other related packages

Intro to tsibble

- tsibble first emerged around 4 years ago
- Built on top of tibble, tsibble extends the tidyverse to temporal-context data, creating a new data infrastructure for time series data
- Main improvements over previous approaches:
  - Time is an explicitly declared variable rather than an implicit attribute
  - Allows for heterogeneous data structures, implicit missing values, multiple variables, and more
  - Handles irregularly spaced time series well
  - Adheres to tidyverse logic: it is data-centric rather than model-centric, and is designed for a human readable pipeline

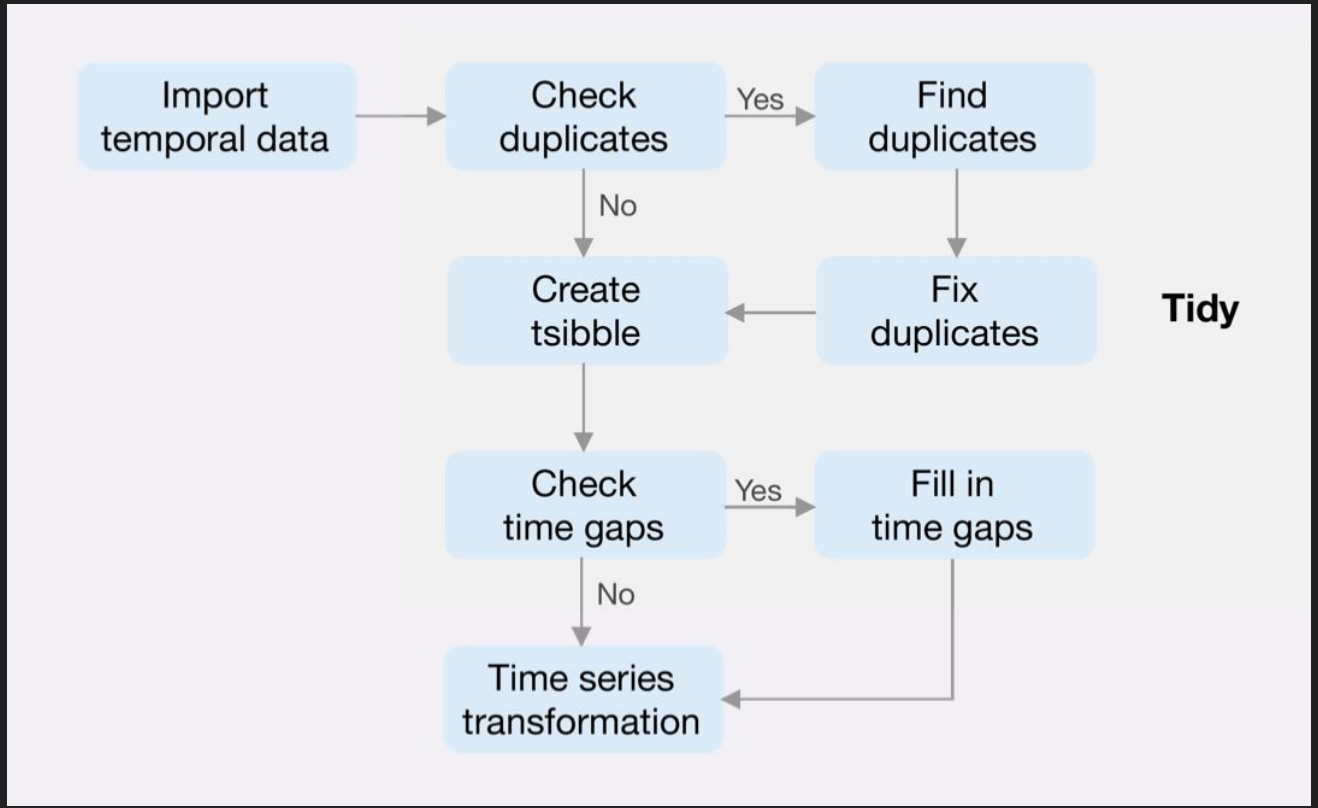
Time series data

- A collection of observations obtained from repeated measurements over time
  - Useful for identifying trends, cycles, and seasonal variances
- Time series data can be:
  - Regular: measurements obtained at a constant time interval
  - Irregular: measurements obtained at inconsistent time intervals (‘event data’)

Basic structure

- Each observation should be uniquely identified by a **key** and an **index**
  - **Key** corresponds to a set of variables that uniquely identify observational units over time, by referencing the index timestamps.
    - It can consist of empty, one, or more variables, meaning you can **nest** and **cross** variables.
  - **Index** contains time indices. tsibble will default to regularly spaced intervals (for irregular data you must specify otherwise).
- tsibbles are sorted by keys, and then by index chronologically.

Working with tsibble



```r
weather <- nycflights13::weather # extract hourly weather dataset

weather_tsibble <- as_tsibble(weather, key = origin, index = time_hour) # coerce to tsibble

has_gaps(weather_tsibble) # check whether time gaps exist
```

```
## # A tibble: 3 × 2
##   origin .gaps
##   <chr>  <lgl>
## 1 EWR    TRUE
## 2 JFK    TRUE
## 3 LGA    TRUE
```

```r
scan_gaps(weather_tsibble) # report all implicit missing values
```

```
## # A tsibble: 75 x 2 [1h] <America/New_York>
## # Key:       origin [3]
##    origin time_hour
##    <chr>  <dttm>
##  1 EWR    2013-01-01 12:00:00
##  2 EWR    2013-02-17 23:00:00
##  3 EWR    2013-02-20 14:00:00
##  4 EWR    2013-02-21 00:00:00
##  5 EWR    2013-07-02 07:00:00
##  6 EWR    2013-07-02 09:00:00
##  7 EWR    2013-07-31 02:00:00
##  8 EWR    2013-08-19 17:00:00
##  9 EWR    2013-08-22 18:00:00
## 10 EWR    2013-08-22 20:00:00
## # … with 65 more rows
```

```r
count_gaps(weather_tsibble) # summarize the time ranges that are absent from data
```

```
## # A tibble: 45 × 4
##    origin .from               .to                   .n
##    <chr>  <dttm>              <dttm>             <int>
##  1 EWR    2013-01-01 12:00:00 2013-01-01 12:00:00     1
##  2 EWR    2013-02-17 23:00:00 2013-02-17 23:00:00     1
##  3 EWR    2013-02-20 14:00:00 2013-02-20 14:00:00     1
##  4 EWR    2013-02-21 00:00:00 2013-02-21 00:00:00     1
##  5 EWR    2013-07-02 07:00:00 2013-07-02 07:00:00     1
##  6 EWR    2013-07-02 09:00:00 2013-07-02 09:00:00     1
##  7 EWR    2013-07-31 02:00:00 2013-07-31 02:00:00     1
##  8 EWR    2013-08-19 17:00:00 2013-08-19 17:00:00     1
##  9 EWR    2013-08-22 18:00:00 2013-08-22 18:00:00     1
## 10 EWR    2013-08-22 20:00:00 2013-08-22 21:00:00     2
## # … with 35 more rows
```

```r
fill_gaps(weather_tsibble) # turn these to explicit NA's
```

```r
weather_tsibble %>%
   group_by_key() %>%
   index_by(date = ~ as_date(.)) %>%
   summarise(
      temp_high = max(temp, na.rm = TRUE),
      temp_low = min(temp, na.rm = TRUE)
   )
```

```
## # A tsibble: 1,092 x 4 [1D]
## # Key:       origin [3]
##    origin date       temp_high temp_low
##    <chr>  <date>         <dbl>    <dbl>
##  1 EWR    2013-01-01      41       28.0
##  2 EWR    2013-01-02      34.0     24.1
##  3 EWR    2013-01-03      34.0     26.1
##  4 EWR    2013-01-04      39.9     28.9
##  5 EWR    2013-01-05      44.1     32
##  6 EWR    2013-01-06      48.0     33.1
##  7 EWR    2013-01-07      46.9     32
##  8 EWR    2013-01-08      48.9     28.9
##  9 EWR    2013-01-09      50       34.0
## 10 EWR    2013-01-10      50       39.0
## # … with 1,082 more rows
```

Tidyverts & other packages

- The **tidyverts** ecosystem is built around the tsibble object for tidy time series analysis
- feasts - visualizing data and extracting time series features
- Fable: a collection of commonly used univariate and multivariate time series forecasting models
- tsibbledata - a range of tsibble dataset examples

References

tidyverts/tsibble: Tidy Temporal Data Frames and Tools. Retrieved 30 October 2021, from https://github.com/tidyverts/tsibble

tsibble: Tidy data structures to support exploration and modeling of temporal-context data. (2018). Retrieved 30 October 2021, from https://www.youtube.com/watch?v=v6yRmbulxUM

Wang, E. (2019). Reintroducing tsibble: data tools that melt the clock. Retrieved 30 October 2021, from https://blog.earo.me/2018/12/20/reintro-tsibble/

Wang, E., Cook, D., Hyndman, R., & O'Hara-Wild, M. (2021). Tsibble documentation. Retrieved 30 October 2021, from https://cran.r-project.org/web/packages/tsibble/tsibble.pdf

Wang, E. Introduction to tsibble. Retrieved 30 October 2021, from https://mran.microsoft.com/snapshot/2018-01-26/web/packages/tsibble/vignettes/intro-tsibble.html

Wang, E. Introduction to tsibble. Retrieved 30 October 2021, from https://cran.rstudio.com/web/packages/tsibble/vignettes/intro-tsibble.html

What is time series data?. Retrieved 30 October 2021, from https://www.influxdata.com/what-is-time-series-data/

Wang, E., Cook, D., & Hyndman, R. J. (2020). A new tidy data structure to support exploration and modeling of temporal data. *Journal of Computational and Graphical Statistics, 29(3)*, 466-478.