

Resping Data with tidyr

Saul Diaz Infante Velasco

3/16/23

Table of contents

Preface	3
Introduction	4
1 Tidy Data	5
1.1 Multiple variables per column	5
Instructions 100 XP	5
1.2 Internall phone numbers	6
Instructions 100 XP	6
1.3 Extracting observations from values	6
Instructions 100 XP	6
1.4 Separating into columns and rows	6
Instructions 100 XP	6
1.5 And the Oscar tfor best director goet to	7
Instructions 100 XP	7
1.6 Imputing sales data	7
Instructions 100 XP	7
1.7 Nuclear bombs per continent	7
Instructions 100 XP	7
2 From Wide to Long and Back	8
3 Expanding Data	9
4 Rectangling Data	10
5 Summary	11
References	12

Preface

Preface for this part

Introduction

Data in the wild can be scary—when confronted with a complicated and messy dataset you may find yourself wondering, where do I even start? The `tidyr` package allows you to wrangle such beasts into nice and tidy datasets. Inaccessible values stored in column names will be put into rows, JSON files will become data frames, and missing values will never go missing again. You'll practice these techniques on a wide range of messy datasets, learning along the way how many dogs the Soviet Union sent into space and what bird is most popular in New Zealand. With the `tidyr` package in your tidyverse toolkit, you'll be able to transform almost any dataset in a tidy format which will pay-off during the rest of your analysis.

1 Tidy Data

You'll be introduced to the concept of tidy data which is central to this course. In the first two lessons, you'll jump straight into the action by separating messy character columns into tidy variables and observations ready for analysis. In the final lesson, you'll learn how to overwrite and remove missing values.

1.1 Multiple variables per column

Being a busy person, you don't want to spend too much time on Netflix, so you decide to crunch some numbers on TV show and movie durations before deciding what to watch. You've managed to obtain a dataset named `netflix_df`, but its `duration` column has an issue. It contains strings with both a value and unit of duration ("`min`" or "`Season`").

You'll tidy this dataset so that each variable gets its own column.

As will always be the case in this course, load the `tidyr` package.

Instructions 100 XP

- Inspect `netflix_df` by typing its name directly in the R console and hitting Enter to see what string separates the value from the unit in the `duration` column.
- Separate the `duration` column over two variables named `value` and `unit`. Pass the string separating the number from the unit to the `sep` argument.

`ex_001.R`

```
netflix_df %>%  
  # Split the duration column into value and unit columns  
  separate(duration, into =c("value","unit"),sep = " ", convert = TRUE)
```

1.2 Internall phone numbers

You work for a multinational company that uses auto-dialer software to contact its customers. When new customers subscribe online they are asked for a phone number but they often forget to add the country code needed for international calls. You were asked to fix this issue in the database. You've been given a data frame with national numbers and country codes named `phone_nr_df`. Now you want to combine the `country_code` and `national_number` columns to create valid international numbers.

Instructions 100 XP

Use the `unite()` function to create a new `international_number` column, using an empty string as the separator.

ex__002.R

1.3 Extracting observations from values

Instructions 100 XP

ex__003.R

1.4 Separating into columns and rows


Instructions 100 XP

ex__004.R

1.5 And the Oscar tfor best director goet to ..

Instructions 100 XP


ex__005.R



1.6 Imputing sales data

Instructions 100 XP


ex__006.R



1.7 Nuclear bombs per continent

Instructions 100 XP

ex__007.R



2 From Wide to Long and Back

3 Expanding Data

4 Rectangling Data

5 Summary

In summary, this book has no content whatsoever.

References