Homework_3_wi19b004

Michael Scheidl

26 09 2021

Contents

1 Assignment

Choose a web site of your choice that contains tabular data worth being harvested. You could look at sites like Wikipedia, or official sites like Statistik Austria offering information. Retrieve the data using tidiyverse and rvest packages so it becomes a "tidy" tibble. If the data is in "wide" format, transform into long format. If it is in long format, transform into wide format. Document your steps in a Notebook.

2 Solution

2.1 Load Libraries

```
library(tidyverse)
library(magrittr)
library(rvest)
```

2.2 Load Website Source Code

```
page <- read_html("https://de.wikipedia.org/wiki/Liste_von_Erdbeben_in_%C3%96sterreich")
html_nodes(page, "*") # Inspect Page Source
```

2.3 Parse Table from HTML Code

```
tmp <-
  page %>%
  html_nodes("table") %>% # Search for table HTML-Nodes
  first() %>% # Use the First Node
  html_table() # Scrape table into tibble
tmp
```

```
## # A tibble: 55 x 9
##    `Datum(UTC)`  `Zeit(UTC)` `ZeitMEZ /MESZ` Epizentrum Beschreibung T     M
##    <chr>         <chr>       <chr>           <chr>      <chr>        <chr> <chr>
##  1 4. Mai 1201   10 Uhr      11 Uhr          Katschber~ Beim frühes~ 8     6,1
##  2 8. Mai 1267   02 Uhr      03 Uhr          Kindberg ~ Über das Be~ 8     5,4
##  3 25. Jän. 1348 16 Uhr      17 Uhr          Friaul / ~ Auch bekann~ 8     6,8
##  4 1. Nov. 1571  -           -               Innsbruck~ Ein in der ~ -     -
##  5 4. Jän. 1572  18:45       19:45           Innsbruck~ Laut einer ~ 6     4,2
##  6 15. Sep. 1590 17 Uhr      18 Uhr          Riederber~ Auch bekann~ 6     5,2
##  7 15. Sep. 1590 23:50       00:50           Riederber~ Vergleiche ~ 6     5,75
##  8 27. Aug. 1668 -           -               Wr. Neust~ Ein in der ~ -     -
##  9 17. Juli 1670 01:15       02:15           Hall / Ti~ Das Erdbebe~ 6     5,2
## 10 22. Dez. 1689 01 Uhr      02 Uhr          Innsbruck~ In einstürz~ 6     4,8
## # ... with 45 more rows, and 2 more variables: I <chr>, Q <chr>
```

2.4 Clean Table

```
tmp <- select(tmp,-last_col())
tmp <- tmp %>% mutate(across(where(is.character), ~na_if(., "-"))) %>% drop_na() # Convert - to NA and
tmp <- tmp %>% mutate_all(str_replace_all, " Uhr", ":00") # Unify Time
tmp <- tmp %>% separate(`Zeit(UTC)`, c("Hour", "Minute"), sep = ":") %>%
  mutate(Hour = str_pad(Hour, "0", width = 2, side = "left")) %>%
  unite(Hour,Minute, sep = ":", col = "Zeit(UTC)") %>%
  separate(`ZeitMEZ /MESZ`, c("Hour", "Minute"), sep = ":") %>%
  mutate(Hour = str_pad(Hour, "0", width = 2, side = "left")) %>%
  unite(Hour,Minute, sep = ":", col = "ZeitMEZ /MESZ") #Unify Time Columns
# Unify Month names -> Unifying the Complete Date didnt work´because of an issue with a " "-Error
tmp <- tmp %>% mutate_all(str_replace_all, "März", "Mrz.")
tmp <- tmp %>% rename(Intentency = I, Magnitude = M, Depth = T) # Give the columns a speaking Name
tmp
```

```
## # A tibble: 48 x 8
##    `Datum(UTC)`  `Zeit(UTC)` `ZeitMEZ /MESZ` Epizentrum  Beschreibung      Depth
##    <chr>         <chr>       <chr>           <chr>       <chr>             <chr>
##  1 4. Mai 1201   10:00       11:00           Katschberg~ Beim frühesten b~ 8
##  2 8. Mai 1267   02:00       03:00           Kindberg /~ Über das Beben i~ 8
##  3 25. Jän. 1348 16:00       17:00           Friaul / I~ Auch bekannt als~ 8
##  4 4. Jän. 1572  18:45       19:45           Innsbruck ~ Laut einer Chron~ 6
##  5 15. Sep. 1590 17:00       18:00           Riederberg~ Auch bekannt als~ 6
##  6 15. Sep. 1590 23:50       00:50           Riederberg~ Vergleiche Beben~ 6
##  7 17. Juli 1670 01:15       02:15           Hall / Tir~ Das Erdbeben im ~ 6
```

```
##  8 22. Dez. 1689 01:00       02:00           Innsbruck ~ In einstürzenden~ 6
##  9 4. Dez. 1690  14:45       15:45           Friaul ? /~ Bei dieser auch ~ 8
## 10 27. Feb. 1768 01:45       02:45           Wr. Neusta~ Für die damalige~ 9
## # ... with 38 more rows, and 2 more variables: Magnitude <chr>,
## #   Intentency <chr>
```

2.5 In Long Format

```
tmp_long <- tmp %>%
  # Combine Intentency,Magnitude,Depth to column "Stat" and show the corresponding value in column "Val
gather(key = "Stat", value = "Value", Intentency,Magnitude,Depth)
# Print Tibble sorted and redundant/unimportant columns hidden
tmp_long %>% arrange(`Datum(UTC)`) %>% select(-`ZeitMEZ /MESZ`,-Beschreibung)
```

```
## # A tibble: 144 x 5
##    `Datum(UTC)`  `Zeit(UTC)` Epizentrum         Stat       Value
##    <chr>         <chr>       <chr>              <chr>      <chr>
##  1 1. Mai 1916   10:24       Judenburg / Stmk.  Intentency 7
##  2 1. Mai 1916   10:24       Judenburg / Stmk.  Magnitude  4,7
##  3 1. Mai 1916   10:24       Judenburg / Stmk.  Depth      7
##  4 11. Juli 2000 02:49       Ebreichsdorf / NÖ  Intentency 6
##  5 11. Juli 2000 02:49       Ebreichsdorf / NÖ  Magnitude  4,8
##  6 11. Juli 2000 02:49       Ebreichsdorf / NÖ  Depth      13
##  7 12. Apr. 1888 05:10       Siegendorf / Bgld. Intentency 7
##  8 12. Apr. 1888 05:10       Siegendorf / Bgld. Magnitude  4,6
##  9 12. Apr. 1888 05:10       Siegendorf / Bgld. Depth      6
## 10 13. Juli 1841 12:30       Wr. Neustadt / NÖ  Intentency 6
## # ... with 134 more rows
```