

WS #7 - Strings

Monday, September 28, 2026

DS 002R - Jo Hardin

Name: _____

Names of people you worked with: _____

Describe your ideal sandwich.

Task: Consider the following made up data on enrollments. The dataset is called `classes`. Try as many of the following tasks as possible in a tidy pipeline. Use `str_*`() functions.

```
# A tibble: 3 x 4
  sem    area  enroll instructor
<chr> <chr> <chr>    <chr>
1 SP2023 History 30 - people Ernesto Capello
2 FA2023 Math    20 - people Lori Ziegelmeier
3 SP2024 Anthro  25 - people Arjun Guneratne
```

1. Change the areas to `history`, `math`, `anthro` instead of `History`, `Math`, `Anthro`. (Hint: there is an easier way to do it than `str_replace()`.)
2. Create a TRUE / FALSE variable that identifies which courses were taught in spring.
3. Change the semester labels to “fall2023”, “spring2024”, “spring2023”.
4. In the `enroll` variable, change all e’s to 3’s (just because?).
5. Use `sem` to create 2 new variables: 1 with the semester (SP/FA) and 1 with the year.
6. Define a new variable `num` that adds up the number of characters in the `area` column. (Hint: one of the `str_*`() functions does *exactly* this!)

Solution:

```
# Change the areas to "history", "math", "anthro"
classes |>
  mutate(area = str_to_lower(area))
```

```
# A tibble: 3 x 4
  sem      area  enroll  instructor
<chr> <chr> <chr>      <chr>
1 SP2023 history 30 - people Ernesto Capello
2 FA2023 math    20 - people Lori Ziegelmeier
3 SP2024 anthro  25 - people Arjun Guneratne
```

```
# Create a variable that identifies courses taught in spring
classes |>
  mutate(spring = str_detect(sem, "SP"))
```

```
# A tibble: 3 x 5
  sem      area  enroll  instructor  spring
<chr> <chr> <chr>      <chr>      <lgl>
1 SP2023 History 30 - people Ernesto Capello TRUE
2 FA2023 Math    20 - people Lori Ziegelmeier FALSE
3 SP2024 Anthro  25 - people Arjun Guneratne TRUE
```

```
# Change the semester labels to "fall2023", "spring2024", "spring2023"
classes |>
  mutate(sem = str_replace(sem, "SP", "spring")) |>
  mutate(sem = str_replace(sem, "FA", "fall"))
```

```
# A tibble: 3 x 4
  sem          area  enroll  instructor
<chr>      <chr> <chr>      <chr>
1 spring2023 History 30 - people Ernesto Capello
2 fall2023    Math    20 - people Lori Ziegelmeier
3 spring2024 Anthro  25 - people Arjun Guneratne
```

```
# or

classes |>
  mutate(
```

```
sem = str_replace(
  sem,
  pattern = c("SP", "FA", "SP"),
  replacement = c("spring", "fall", "spring")
)
```

```
# A tibble: 3 x 4
  sem      area  enroll      instructor
<chr>    <chr>  <chr>      <chr>
1 spring2023 History 30 - people Ernesto Capello
2 fall2023   Math   20 - people Lori Ziegelmeier
3 spring2024 Anthro  25 - people Arjun Guneratne
```

```
# In the enroll variable, change all e's to 3's (just because?)
classes |>
  mutate(enroll = str_replace_all(enroll, "e", "3"))
```

```
# A tibble: 3 x 4
  sem      area  enroll      instructor
<chr>    <chr>  <chr>      <chr>
1 SP2023 History 30 - p3opl3 Ernesto Capello
2 FA2023 Math   20 - p3opl3 Lori Ziegelmeier
3 SP2024 Anthro  25 - p3opl3 Arjun Guneratne
```

```
# Use sem to create 2 new variables, one with only the semester (SP/FA) and 1 with the year
classes |>
  mutate(semester = str_sub(sem, 1, 2), year = str_sub(sem, 3, 6))
```

```
# A tibble: 3 x 6
  sem      area  enroll      instructor  semester year
<chr>    <chr>  <chr>      <chr>          <chr>   <chr>
1 SP2023 History 30 - people Ernesto Capello SP      2023
2 FA2023 Math   20 - people Lori Ziegelmeier FA      2023
3 SP2024 Anthro  25 - people Arjun Guneratne SP      2024
```

```
# Define a new variable "num" that adds up the number of characters in the area label
classes |>
  mutate(num = str_length(area))
```

```
# A tibble: 3 x 5
  sem    area  enroll    instructor    num
<chr> <chr> <chr>      <chr>      <int>
1 SP2023 History 30 - people Ernesto Capello      7
2 FA2023 Math    20 - people Lori Ziegelmeier      4
3 SP2024 Anthro 25 - people Arjun Guneratne      6
```