String Manipulation and Relational Data: Takeaways

Syntax

FINDING NUMBERS IN A STRING

- Extracting numbers in a string:

```
`Total Grads - % of cohort` <- parse_number(`Total Grads - % of cohort`)
```

- Extracting numbers from multiple variables at once:

```
data_frame %>%
mutate_at(vars(`first column in range`:`last column in range`), parse_number)
```

SUBSETTING STRINGS

- Subsetting a character string from left to right:

```
Vector_2  <- Vector_1 %>%
str_sub(5, 7)
```

- Subsetting a character string from right to left:

```
Vector_2  <- Vector_1 %>%
str_sub(-4, -6)
```

JOINING DATA FRAMES

- Combining two data frames using an inner join:

```
sat_results %>%
inner_join(class_size, by = "DBN")
```

- Combining two data frames using a left join:

```
sat_results %>%
left_join(class_size, by = "DBN")
```

- Combining two data frames using a right join:

```
sat_results %>%
right_join(class_size, by = "DBN")
```

- Combining two data frames using a full join:

```
sat_results %>%
full_join(class_size, by = "DBN")
```

Concepts

- Data is optimally organized for use with tidyverse tools when it is "tidy":
    - Variables in columns
    - Observations in rows
    - Values in cells
- Relational data is data that has a relation to some data in another table.
- A key refers to the variable that connects pairs of tables.
    - Mutating joins add new variables to a data frame based on matching observations in another data frame.
        - Inner joins match pairs of variables in two data frames if their values of the key are the same.
        - Outer joins keep observations that appear in at least one of the two tables you're combining. Outer joins can be divided into three types:
            - Left joins
            - Right joins
            - Full joins
- Performing a left join keeps all observations in the data frame on the left and drops observations from the data frame on the right that have no key match.
- Performing a right join keeps all observations in the data frame on the right and drops observations from the data frame on the left that have no key.

- Performing a full join keeps all observations from both data frames and fills in missing variables with `NA`.

Resources

- Documentation for parse_number
- Cheat sheet for dplyr join functions