Correlations and Reshaping Data: Takeaways

Syntax

RESHAPING DATA FOR VISUALIZATION

- Reshape a dataframe so that variable names are values of a new variables:

```
combined_socio_longer <- combined %>%
  pivot_longer(cols = c(frl_percent, ell_percent, sped_percent),
           names_to = "socio_indicator",
           values_to = "percent")
```

- Reshape a dataframe so that a variable values are variable names of a new variables:

```
combined_socio_wider <- combined %>%
  pivot_wider(names_from = socio_indicator,
              values_from = percent)
```

CALCULATING PEARSON'S CORRELATION COEFFICIENT

- Calculate Pearson's correlation coefficient for a pair of variables:

```
cor(combined$avg_sat_score, combined$asian_per, use = "complete.obs")
```

- Create a correlation matrix to calculate Pearson's correlation coefficient for multiple pairs of variables:

```
cor_mat <- combined %>%
  select(when(is.numeric)) %>%
  cor(use = "pairwise.complete.obs")
```

- Convert a correlation matrix to a tibble:

```
cor_tib <- cor_mat %>%
  as_tibble(rownames = "variable")
```

- Index a tibble to identify moderate to strong correlations:

```
apscore_cors <- cor_tib %>%
  select(variable, high_score_percent) %>%
  filter(high_score_percent > 0.25 | high_score_percent < -0.25)
```

Concepts

- Reshaping data is a common task we'll need to perform as we clean and analyze data.
- Most tidyverse functions work best when data are organized according to "tidy data principles":
    - Variables in columns.
    - Observations in rows.
    - Values in cells.

- The `pivot_longer()` function takes multiple columns and collapses them into key-value pairs, duplicating all other columns as needed, so that the dataframe is "tidy."
- The `pivot_wider()` function takes two parameters (`names_from` and `values_from`) and expands `names_from` variable into distinct columns containing values from `values_from` parameter.
- Calculating correlation coefficients (Pearson's r) allows us to measure the strength of a relationship between a pair of variables.
    - Correlation coefficients have a value between +1 and −1.
    - The closer a correlation coefficient is to zero, the weaker the relationship is between the two variables.
    - The closer a correlation coefficient is to -1 or 1, the stronger the relationship.
    - Positive values indicate a relationship where both variables' values increase.
    - Negative values indicate a relationship where one variable decreases as another one increases.
    - Values above 0.25 or below -0.25 are enough to qualify a correlation as potentially interesting and worthy of further investigation.
    - Values above 0.75 or below -0.75 indicate strong relationships.

Resources

- Documentation for tidyr
- Tidyr package's release changelog
- Article on Pearson's correlation coefficient