Bivariate Relationships - Correlation and Scatterplots: Takeaways

Syntax

EVALUATING BIVARIATE RELATIONSHIPS

- Exploratory scatterplot showing commas instead of scientific notation:

```
ggplot(data = df,
   aes(x = predictor, y = response)) +
   scale_y_continuous(labels = scales::comma) +
   geom_point()
```

- Highlighting outliers as defined by the interquartile method:

```
quartiles <- quantile(df$col)
iqr <- quartiles[[4]] - quartiles[[2]]
lower_bound <- quartiles[[2]] - (1.5 * iqr)
upper_bound <- quartiles[[5]] + (1.5 * iqr)
outliers <- df %>%
  filter(col > upper_bound | col < lower_bound)
ggplot(data = df,
       aes(x = predictor, y = response)) +
  geom_point(data = outliers, aes(predictor, response),
             color = "orangered3", size = 4) +
  geom_point() +
  scale_y_continuous(labels = scales::comma) +
  geom_smooth(method = "lm", se = FALSE)
```

- Generating correlation coefficients between numeric variables:

```
correlation_matrix <- df %>%
  select_if(is.numeric) %>%
  cor(use = "pairwise.complete.obs")
correlation_df <- correlation_matrix %>%
  as_tibble(rownames = "variable") %>%
  select([variables of interest]) %>%
  tidyr::drop_na()
```

Concepts

- *Bivariate linear regression* can be performed between pairs of quantitative data measured on an interval or ratio scale..
- Problems with a quantitative response where there exists a dependency relationship are known as *regression problems*.
- *Bivariate linear regression* is a type of model that uses a single quantitative input to explain a single quantitative response.
- Categorical variables and non-numeric variables can be useful for subsetting or filtering data into groups of similar observations.
- When we generate a scatterplot to explore bivariate relationships, we can evaluate the relationships between the variables by examining the following characteristics: direction, linearity, and strength.
- *Direction:* With a *positive relationship*, an increase in value along the x-axis results in an increase in value along the y-axis. A *negative relationship* is when an increase in one variable is associated with the decrease in value for another variable.
- *Linearity:* A bivariate relationship is linear when the scatter, or spread of data points generally follows a linear pattern. Examples of non-linear patterns include situations where the data exibits curvature, spread, or clustering.
- *Strength:* A bivariate relationship is strong when the spread of the data is narrow. With a strong bivariate relationship, there will not be a lot of scatter, or noise, present. Scatterplots are useful for visualizing bivarite relationships. With a strong relationship, the residuals will be lower.
- *Correlation* is a statistic that quantifies the strength of the relationship between two variables.
- Correlation analysis is called for when there is no dependence between two variables. Regression analysis can be performed when the magnitude of one variable (dependent variable) is a function of the magnitude of the other variable (independent variable).

- In the context of linear regression, an *outlier* is an observation for which the response value is far from the value predicted by our model. Outliers identified using the interquartile range for a given variable may be useful for investigating potential outliers in linear regression, but it is possible that an outlier in this context will not be considered an outlier with linear regression, and vice versa.

Resources

- Wikipedia entry on linear regression.
- Wikipedia entry on regression analysis.
- Wikipedia entry on outliers.
- Pearson correlation coefficient Wikipedia entry.