

Lecture 2.1 - Association activity

Student

2025-03-27



Planning

Expectations

You should develop some expectations based on some pre-existing ideas about how you think the data is related. Pick **Calories** and another quantitative variable and write down how you expect the two to be related.

Write down your expectation and provide a brief reason for the expectation. Remember to describe your expectation about the relationship between the two variables in terms of direction, form, strength, and outliers.

Direction of the relationship

Make a decision about which variable is the outcome variable you care about and which is the variable that predicts it. Write down your choice.

Investigation

Distribution displays

Histograms

- Make histograms of your two variables using `ggplot()` and `geom_histogram()`. Do you see anything unusual? What do you think these distributions indicate about the possible relationship between the two?

Scatterplot

- Make a high-quality scatterplot of the two variables using `geom_point()`. You can add a smoother to the scatterplot by adding a `geom_smooth()` layer to your `ggplot`.

Describe the relationship between the two variables using the terms we learned in class.

- Direction
- Form
- Strength
- Outliers

Does the relationship match your expectations? Why or why not? Write some notes about comparing expectations vs. reality.

Correlation

Find correlation result – does it match the scatterplot? Your expectations? Why or why not?

The RStudio code for correlations is:

```
cor(x, y, use="complete.obs")
```

- Note: you will need to replace `x` and `y` in the above line of code with the variables of interest. Remember to directly specify a variable, it should be in the `<dataset name>$<variable>` format.

Analysis

Outliers

If you have any outliers, identify them. Do you think they should be excluded from your analysis? Why or why not? If you remove the outliers (via the `filter()` verb), does it change your correlation? The shape of the smoother?

Transformations

- Could the relationship of your two chosen variables benefit from being re-expressed? Try a few re-expressions and see how it affects the relationship and correlation.
 - Hint: remember, you can use the `mutate()` verb in a piped command, such as:
`mutate(logcals = log(Calories))`

Conclusion

Overall, summarize what you have learned about the relationship of your two variables.

Extra

If you have time, complete the above steps with additional variables.