SDS Fellow Workshop: Introduction to R

Casey Crary and Tyler McCord

Introduction to R

Welcome (back) to R! R is a programming language used for statistical computing and data analysis, and is the primary tool used by the Statistics Department at Amherst. This workshop is designed to introduce you to R and R-Studio, and to get you familiar with how to interact with it for your courses. You can refer to the slides in this workshop for some tips to get you started.

We will run through a small example of data analysis using R.

EXAMPLE: The relationship between foot length and foot width in children

The first step in any data analysis is to load the necessary packages and libraries. The `tidyverse` package is the most commonly used package in R, and it contains many useful functions for data manipulation and visualization.

The `KidsFeet` dataset gives us information on 39 kids and their foot measurements. We can use the `data()` function to load the dataset into R.

Run the code chunk below, and notice the output in the environment tab in the top-right corner of R-Studio.

```
data("KidsFeet")
```

We can use the `head()` function to see the first few rows of the dataset, and get an idea of what the data looks like. This is useful to get familiar with data that you might not have seen before to learn what it's telling us.

Run the code chunk below and notice the output below the code chunk.

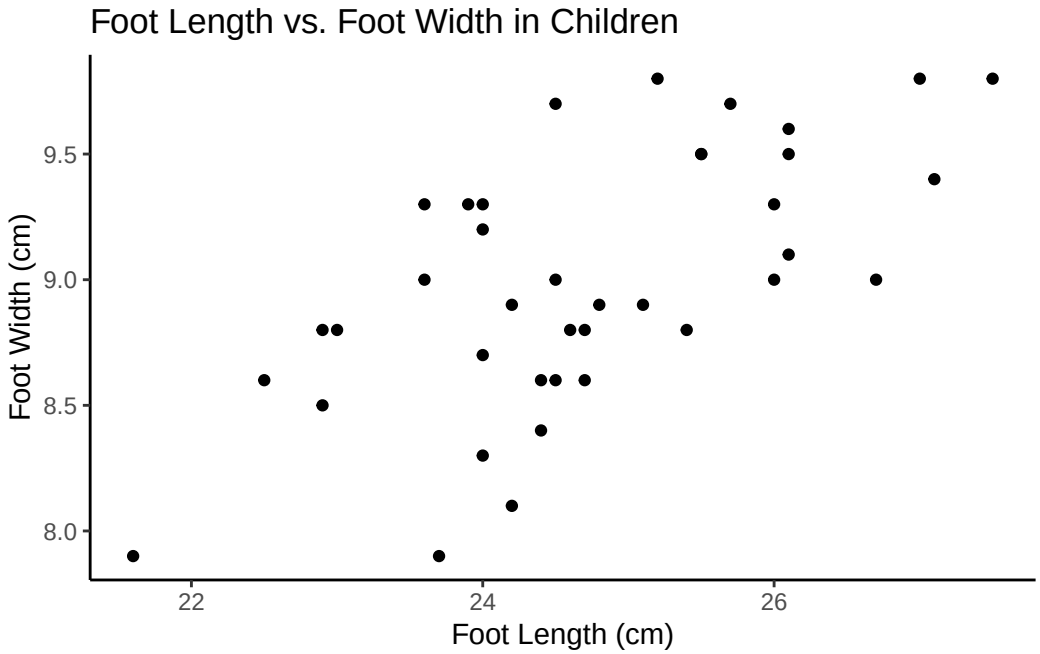
```
head(KidsFeet)
```

```
    name birthmonth birthyear length width sex biggerfoot domhand
1  David          5        88   24.4   8.4   B          L       R
2   Lars         10        87   25.4   8.8   B          L       L
3   Zach         12        87   24.5   9.7   B          R       R
4   Josh          1        88   25.2   9.8   B          L       R
5   Lang          2        88   25.1   8.9   B          L       R
6 Scotty          3        88   25.7   9.7   B          R       R
```

Now that we are getting familiar with the data, we can start to explore it. Visualizations are key to understanding data, so let's start by making a scatterplot of foot length vs. foot width. Don't worry about what the code means right now, just run the code chunk below and look at the output.

Take note of what some key takeaways from this visualization are. Also, notice the green text that appears after the # symbol. These are comments; they are not code, but they are used to describe what the code is doing to help readers understand.

```
# visualizing the relationship between foot length and width
ggplot(data = KidsFeet) +
  geom_point(mapping = aes(x = length, y = width)) +
  labs(x = "Foot Length (cm)", y = "Foot Width (cm)", title = "Foot Length vs. Foot Width in
  theme_classic()
```



Now, we can use Simple Linear Regression using the `lm()` function to quantify the relationship between foot length and foot width. The `<-` symbol is used to assign the output of the `lm()` function to a new object called `model`. Then we can use the `summary()` function to see the results of the regression.

```
# linear regression of foot width on foot length
model <- lm(width ~ length, data = KidsFeet)
summary(model)
```

```
Call:
lm(formula = width ~ length, data = KidsFeet)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83864 -0.31056 -0.00892  0.27622  0.76300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8623     1.2081   2.369   0.0232 *
length        0.2480     0.0488   5.081  1.1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3963 on 37 degrees of freedom
Multiple R-squared:  0.411, Adjusted R-squared:  0.3951
F-statistic: 25.82 on 1 and 37 DF,  p-value: 1.097e-05
```