Student Activity 4: Can shoe size predict reading?

A Course on Data Visualization · Group Activity

Livio Finos, University of Padova

Contents

Your task **1**

Part 1: Design a study **1**

Part 2: Analyse the data **2**

Study context and data 2

Check the observations 3

First look: ignore school year 4

Second look: explore other categorical variables 5

Third look: compare children within school years 5

Explain the change 6

What to submit **6**

Your task

Work in groups to design a study and prepare a **reproducible exploratory report**. Investigate two familiar measurements, then ask whether their relationship holds when comparing similar children.

1. **Design and plan:** agree on a study design and analysis plan before opening the data or asking an AI assistant for an analysis.
2. **Analyse and verify:** implement the plan, check calculations and graphs, and record any changes to the original plan.
3. **Discuss and compare:** compare interpretations with other groups and explain what the evidence supports.

Part 1: Design a study

Do children with larger feet tend to obtain higher reading scores?

Imagine that a school asks your group to investigate this question among primary-school children. No measurements are available yet.

Write a short study proposal addressing the following questions.

1. **Question and population.** Which children would you like your conclusion to describe? Are you interested in prediction, association, or a causal effect? Explain the distinction for this question.
2. **Data collection.** How would you select children and obtain consent? How would you measure shoe size and reading performance consistently? What would one row represent?
3. **Sample size and sampling.** How many children would you recruit, and why? How would you ensure enough observations in the groups you plan to compare? A reasoned proposal is sufficient; no power calculation is required.
4. **Other variables.** What else could explain an association between shoe size and reading? Propose **categorical** background variables and explain why each might relate to both measurements. Do not propose continuous adjustment variables or collect every variable you can imagine.
5. **Analysis plan.** Which descriptive summaries, scatterplots, correlations, and simple regression lines would answer your question? How would you compare an overall pattern with patterns within groups? Specify axes, units, and grouping variables before writing code.
6. **Limitations.** Which biases or unmeasured differences might remain? What could an observational study establish, and what remains uncertain?

> **Plan before prompting.** Save your proposal before opening the CSV. If you use AI later, supply your plan and verify its implementation. Keep a record of what you accepted, changed, or rejected and why.

Pause here for a class discussion. The instructor will then release the data and Part 2. Keep the original proposal alongside any revisions.

Part 2: Analyse the data

Study context and data

The supplied CSV is a **simulated teaching dataset**, not observations of real children. The setting is a fictional primary-school study with one measurement per child. Children in Years 1, 3, and 5 complete the same reading assessment, scored from 0 to 100, and have their shoe size recorded using the same European sizing scale. Higher scores indicate better performance on this assessment.

Use `data/shoe_size_reading.csv`, in the `data` folder next to this Rmd. The CSV is the fixed input for your report; do not simulate replacement data. The examples use base R. Knitting requires `knitr` and `rmarkdown`.

Variable	Type	Meaning
`child_id`	Identifier	Anonymous label; one row per fictional child
`shoe_size`	Quantitative	European shoe size, recorded in half-size steps
`reading_score`	Quantitative	Common reading assessment, 0–100 points
`school_year`	Categorical	Year 1, Year 3, or Year 5; use as a factor
`handedness`	Categorical	Preferred writing hand: Left or Right
`favourite_colour`	Categorical	Preferred colour: Blue, Green, Red, or Yellow

Variable	Type	Meaning
`birth_season`	Categorical	Season of birth: Winter, Spring, Summer, or Autumn (Northern Hemisphere)

Shoe size is quantitative despite its discrete steps. Treat school year as **three categories**, not a numeric predictor with an assumed linear effect. Treat the other background variables as factors too. Their labels provide context, but do not establish whether they explain an association.

```
data_file <- "data/shoe_size_reading.csv"
if (!file.exists(data_file)) stop("Missing file: ", data_file)
dat <- read.csv(data_file, stringsAsFactors = FALSE)
dat$school_year <- factor(dat$school_year,
                          levels = c("Year 1", "Year 3", "Year 5"))
other_categories <- c("handedness", "favourite_colour", "birth_season")
dat[other_categories] <- lapply(dat[other_categories], factor)
head(dat)
```

```
##   child_id shoe_size reading_score school_year handedness favourite_colour
## 1     C001        28            29      Year 1      Right             Blue
## 2     C002        31            43      Year 1      Right              Red
## 3     C003        34            35      Year 1      Right              Red
## 4     C004        31            41      Year 1       Left            Green
## 5     C005        30            33      Year 1      Right             Blue
## 6     C006        28            40      Year 1      Right           Yellow
##   birth_season
## 1       Autumn
## 2       Summer
## 3       Winter
## 4       Winter
## 5       Autumn
## 6       Autumn
```

```
str(dat)
```

```
## 'data.frame':    300 obs. of  7 variables:
##  $ child_id        : chr  "C001" "C002" "C003" "C004" ...
##  $ shoe_size       : num  28 31 34 31 30 28 28.5 28.5 28.5 28.5 ...
##  $ reading_score   : int  29 43 35 41 33 40 39 26 35 44 ...
##  $ school_year     : Factor w/ 3 levels "Year 1","Year 3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ handedness      : Factor w/ 2 levels "Left","Right": 2 2 2 1 2 2 2 2 2 2 ...
##  $ favourite_colour: Factor w/ 4 levels "Blue","Green",..: 1 3 3 2 1 4 4 4 3 1 ...
##  $ birth_season    : Factor w/ 4 levels "Autumn","Spring",..: 1 3 4 4 1 1 3 4 2 3 ...
```

Check the observations

1. Check row counts, uniqueness of child identifiers, missing values, and the ranges and coding of all measurements.
2. Count children in each school year. Compare these counts with your original sampling proposal.

3. Summarise shoe size and reading score overall and by school year. Inspect the distributions and note the range of shoe sizes shared by different years. Do not delete observations merely because they look unusual.

```
nrow(dat)
```

```
## [1] 300
```

```
anyDuplicated(dat$child_id)
```

```
## [1] 0
```

```
colSums(is.na(dat))
```

```
##         child_id        shoe_size    reading_score      school_year
##                0                0                0                0
##       handedness favourite_colour     birth_season
##                0                0                0
```

```
table(dat$school_year, useNA = "ifany")
```

```
##
## Year 1 Year 3 Year 5
##    100    100    100
```

```
summary(dat)
```

```
##       child_id     shoe_size     reading_score   school_year  handedness
##  Length   :300   Min.   :24.00   Min.   :17.00   Year 1:100   Left : 30
##  N.unique :300   1st Qu.:29.50   1st Qu.:40.00   Year 3:100   Right:270
##  N.blank  :  0   Median :31.50   Median :53.50   Year 5:100
##  Min.nchar:  4   Mean   :31.93   Mean   :55.41
##  Max.nchar:  4   3rd Qu.:34.12   3rd Qu.:71.25
##                  Max.   :41.50   Max.   :93.00
##  favourite_colour birth_season
##  Blue  :75        Autumn:75
##  Green :75        Spring:75
##  Red   :75        Summer:75
##  Yellow:75        Winter:75
##
##
```

First look: ignore school year

1. Draw a scatterplot of reading score against shoe size, with clear units.
2. Calculate Pearson's correlation and fit one simple regression line.
3. Describe the direction and strength of the association. Interpret the slope in reading-score points per one-unit increase in shoe size.
4. Write a two-sentence provisional conclusion. Would it justify advising parents to buy larger shoes? Explain your reasoning.

These commands may help; add the graph and your interpretation.

```
cor(dat$shoe_size, dat$reading_score)
```

```
## [1] 0.720209
```

```
pooled_fit <- lm(reading_score ~ shoe_size, data = dat)
coef(pooled_fit)
```

```
## (Intercept)   shoe_size
##  -77.162698    4.151867
```

Second look: explore other categorical variables

Before focusing on school year, compare `handedness`, `favourite_colour`, and `birth_season`. Do not assume that every recorded variable is a confounder.

For each variable:

1. Count children in each category. Cross-tabulate it with school year and compare category proportions within each year, not just raw counts.
2. Compare average shoe size and reading score between its categories.
3. Draw scatterplots with one panel per category and common axis limits. Add a separate simple regression line within each category.
4. Calculate the sample size, correlation, and slope within each category. Does this grouping change the interpretation of the pooled association?
5. Explain whether the variable looks like a plausible confounder of this relationship. Use its relationship with school year and both measurements, together with your study explanation; do not choose variables only because they make a correlation smaller.

Third look: compare children within school years

1. Redraw the scatterplot using colour **and** plotting symbols for school year. Add a separate fitted line for each year.
2. Make one panel per year, with the **same axis limits** in every panel. Fit a separate simple regression within each year. Avoid extending its line beyond that year's observed shoe sizes.
3. Prepare a table with sample size, Pearson correlation, and simple regression slope for each year. Compare these with the pooled results.
4. Does the overall association describe the relationship within each year? Support your answer with numerical results and features of the graphs.

Split the observations into groups and repeat the calculations. Complete this outline in your own report.

```
by_year <- split(dat, dat$school_year)
lapply(by_year, function(d) {
  # Return n, the correlation, and the slope from lm() for this year.
})
```

Use **stratified correlations and separate simple regressions**. Multiple regression and continuous adjustment variables are not required. Focus on effect sizes and graphs; formal significance tests are not required.

Explain the change

- Which categorical grouping changes the association most? How does this compare with the patterns for handedness, favourite colour, and birth season?
- How do average shoe size and average reading score differ between years?
- How can differences between groups produce an overall pattern that is weak or absent within the groups?
- Why is school year a plausible confounder? Draw arrows showing your proposed explanation, or describe it in words.
- Does a correlation close to zero prove independence or the absence of every possible relationship? What does the scatterplot add?
- Does stratifying by school year establish a causal effect or exclude all remaining confounding in a real study?
- How would you revise your original sampling and analysis plan?

What to submit

Submit a short reproducible Rmd report and its knitted HTML containing:

1. Your original study proposal and analysis plan, with later revisions marked.
2. Data checks and overall and within-year descriptive summaries.
3. The pooled scatterplot and grouped and faceted views, with captions.
4. Cross-tabulations with school year and comparison tables of sample sizes, correlations, and slopes for all four categorical background variables.
5. A final interpretation distinguishing overall association, within-year association, and causal claims, plus limitations and a revised study proposal.
6. If applicable, the main AI prompt, relevant follow-ups, and at least one calculation independently checked by your group.

Write in English. Keep numerical precision sensible, label axes and units, and connect each conclusion to a calculated quantity or a visible pattern.

Note on AI assistance

These teaching materials were prepared with assistance from ChatGPT (OpenAI) in drafting explanations, developing examples, and generating R code. The author retains responsibility for the final content.