

# ECON 0150 | Spring 2026 | Homework 2.3

*Due: Tuesday, February 24 at 11:59 PM*

Homework is designed to both test your knowledge and challenge you to apply familiar concepts in new applications. Answer clearly and completely. You are welcomed and encouraged to work in groups so long as your work is your own. Use the provided datasets to answer the following questions. Then submit your figures and answers to Gradescope.

## Q1. Explore the Data

The dataset `amazon_book_sales.csv` contains the 50 bestselling Amazon books each year between 2009 and 2021. Prices are expressed in US dollars and rounded to the nearest dollar.

- a) How many unique books are in the dataset (*hint: use `.unique()` to count a book only once even if it shows up in multiple years*)?
- b) How many books are Fiction? How many are Non-Fiction? (*hint: use a filter*)

## Q2. Ratings by Genre

The column `Genre` describes whether the bestseller is Fiction or Non-Fiction.

- a) Create a boxplot of `User Rating` by `Genre`. Add a stripplot overlay.
- b) Calculate the mean and standard deviation of `User Rating` by `Genre`.
- c) Which genre has higher average ratings? Is the difference large or small relative to the variation within each group?

## Q3. Reviews and Ratings

- a) Create a scatter plot of `Reviews` (x-axis) vs `User Rating` (y-axis) for all books.
- b) Is there a relationship between reviews and ratings? If so, is it positive, negative, or unclear?

c) How might you use a transformation to improve this figure?

#### **Q4. Does the Relationship Differ by Genre?**

a) Create a scatter plot of Reviews vs User Rating, colored by Genre. Use any transformations you suggested in Q3.c).

b) Does the relationship between reviews and ratings differ between Fiction and Non-Fiction? Describe the pattern.

c) In one sentence, what hypothesis might explain any differences you observe?