

Exploratory Data Analysis

Grinnell College

September 30, 2026

What is exploratory data analysis

1. Describe distributions
2. Explore relationships and differences between groups
3. Refine questions, generate new ones

The goal of EDA is to be able to say something affirmative about the data, suggest questions, or identify potential issues

Typically, we start with a set of questions

- ▶ Do I need all of my observations or only a subset
- ▶ What variables are needed to answer this? Quantitative or categorical
- ▶ What types of summaries do these involve?

While graphs and tables alone cannot establish a relationship between variables, they can suggest forms for us to test

Match the summary to the question

Variables	Numerical summary	Graph
One quantitative variable	Mean and SD; median and IQR	Histogram or boxplot
Quantitative and Categorical	Center, spread, and count within each group	Boxplots
Two categorical variables	Counts and proportions	Bar chart
Two quantitative variables	Correlation, regression	Scatterplot

- ▶ Center alone does not describe a distribution.
- ▶ When group sizes differ, compare proportions rather than raw counts.

Organize and identify your summaries

Typically each question will involve the creation of a data summary

- ▶ Create descriptive names for datasets
- ▶ What does one row constitute?
- ▶ Understand which variables are included

Summarizing data can involve numerical summaries, graphical summaries, or even operations and mutations of our data frames

Knowing the correct plot or summary is first step
Being able to identify relevant information is second

- ▶ Skewed or symmetric distributions
- ▶ Outliers or unusual values
- ▶ Direction, form, and strength
- ▶ Identify patterns across groups

A Successful EDA

- ▶ The main goal is communication
- ▶ Typically 2-3 main ideas, do not include everything you found
- ▶ Concise summaries with written explanations
- ▶ Only provide output and text relevant to the problem at hand
- ▶ Never include R code or raw output













