Exploratory Data Analysis

Nick Ulle

Exploring Data

What does it mean to “explore” data?

- Get an overview of what’s included
- Look for errors (typos, extreme values, etc.)
- Look for patterns within features
- Look for relationships between features
- Check assumptions (for conclusions, models, etc.)

Use plots, summary statistics, and *with caution*, models.

Choosing a Plot

There isn’t a strict guideline for how to choose a plot.

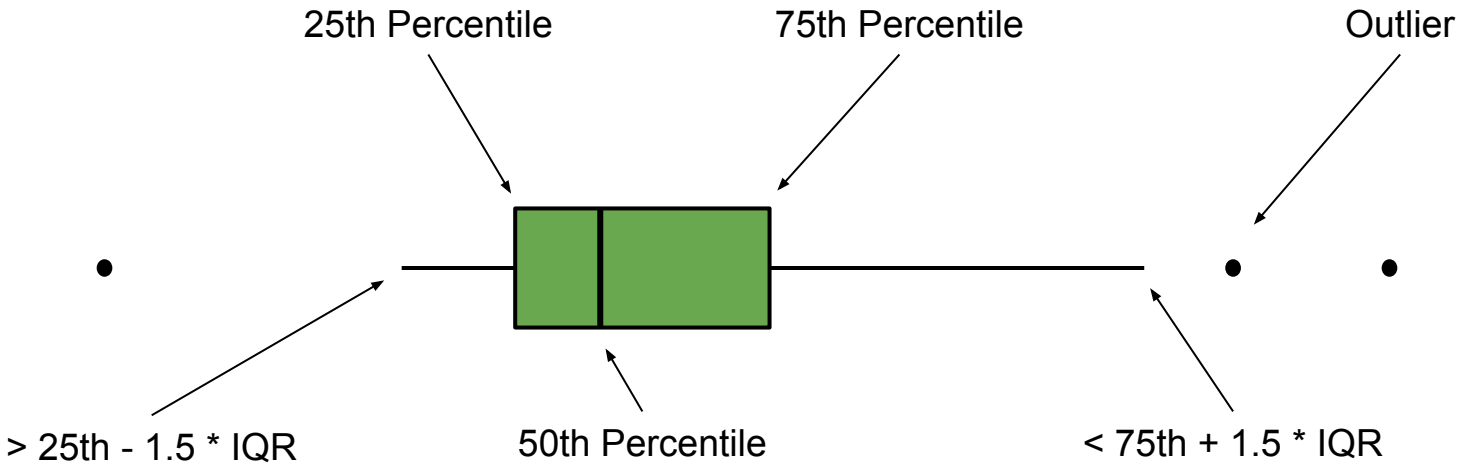
This table has *suggestions:*

First Feature	**Second Feature**	**Suggested Plots**
categorical		bar, dot
categorical	categorical	bar, dot, mosaic
numerical		box, density, histogram
numerical	categorical	box, density
numerical	numerical	line, scatter, smooth scatter

Distribution Plots

Nick Ulle

Anatomy of a Box Plot



Interquartile Range (IQR) is (75th - 25th)

Faceted Plots

Nick Ulle

EDA Strategy

Nick Ulle

EDA Strategy

For a data analysis question, ask yourself:

1. Which rows of the data are relevant?
    - Take a subset if it isn't all of them
2. Which columns of the data are relevant?
3. Does the data need to be aggregated or transformed?
    - For example, computing frequencies or statistics
4. Which plot is appropriate?

EDA Examples

Nick Ulle