DATA 3464: Fundamentals of Data Processing

Basic machine learning models

Charlotte Curtis

January 13, 2026

This week's topics

- Exploring and understanding your data
- Splitting your data
- Assignment 1: Exploring Calgary traffic data

Resources used:

Basic things to look at

- Data source - File? Database? API?
- Structured/unstructured
- Assumption 1: relatively small (fits in memory) tabular dataset
  - Data types - numeric/categorical, text, other
  - Assumption 2: numeric data
    - Ranges
    - Summary statistics
    - Missing values
- Next week: categorical data, after reading week unstructured

Example: Anscombe's Quartet

- Very small dataset, constructed by hand in 1973 by Francis Anscombe

> Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:
>
> (1) numerical calculations are exact, but graphs are rough;
>
> (2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
>
> (3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

- Not known exactly how he made it, but Drs. Roberta La Haye and Peter Zizler proposed a compelling argument for linear algebra

Useful starting points

- `pandas.DataFrame.info`: data type, number of non-null, names, dimensions
- `pandas.DataFrame.head`: return the first `n` rows (default 5)
- `pandas.DataFrame.describe`: Compute a bunch of summary statistics

Types of exploratory visualizations

- I will not provide an exhaustive list of visualizations!
- Pandas provides a handy wrapper around matplotlib
- Some of my favourites:
  - Histograms
  - Scatter plots/hexbin plots
  - Box plots/violin plots