MXB107 Week 1
Introduction

the university
for the real world

Statistical Modelling

- What is *statistics*?
  - Statistics is the science of extracting meaning from data
  - This is often done through statistical modelling

- What is *statistical modelling*?
  - Description of a real process in mathematical terms
  - Often driven by the real questions we want to answer

- What is *data*?
  - Data are a collection of facts that describe some characteristic(s) that can be ranked, counted, or measured.
  - The way data are collected will often depend on the questions we want to answer.

Randomness and Probability

- Almost every variable is *random*
  - How long does it take you to get to campus?
  - Who will win the next game of football?
  - Even our measurements of constants are random

- *Probability* is a mathematical construct for dealing with randomness and uncertainty
  - Provides us a set of rules for calculating the uncertainty associated with our data

Experimental Units and Measurements

- An *Experimental Unit* is an individual that generates information for the data collection process. Must be made to ensure that it aligns with the questions of interest
- A *measurement* is the information we collect on experimental units
- Examples: What are the experimental units and measurements for these research questions?
  - How does fertilizer type affect plant growth?

Experimental Unit: plant
Measurement: plant height, volume, etc

  - How consistent is the measurement of temperature using different types of thermometers?

Experimental Unit: thermometers
Measurement: temperature.

Sample versus Population

- We might have questions about a very large collection of things called a *population*
  - It is usually not feasible to collect data from the entire population

- A *sample* is a subset of a population that we take measurements from to infer something about the population.
  - We would like our sample to be *representative* of the population
  - A *random sample* is one where the sample members are selected from the population by chance
  - Sometimes we might need a more sophisticated sampling strategy

Data Types

- If we collect multiple measurements on experimental units, we can investigate relationships between measurements
  - *Univariate* data corresponds to a single measurement
  - *Bivariate* data corresponds to two measurements
  - *Multivariate* data corresponds to > two measurements

- *Experimental data* is data collected in a controlled manner. Designed with a specific question in mind

- *Observational data* have been collected and curated without any specific analyses or modelling in mind

Examples – Experimental vs Observational

- Label these as experimental or observational studies:

  - A researcher gives one group of mice a new drug and another group a placebo, then measures how active they are over 24 hours. experimental
  - Doctors review past medical records to compare recovery times between patients who took two different medications for the same illness. observational
  - A health researcher surveys people about how many hours of exercise they get per week and tracks their weight over six months. observational
  - In a clinical trial, participants are randomly assigned to receive either a vaccine or a placebo, and infection rates are compared. experimental

Data Types (Cont…)

- *Quantitative data* are things that are naturally represented numerically
  - *Discrete* – are observations that occur as natural or whole numbers
  - *Continuous* - are measurements on a continuum or measures that may be subdivided infinitely
  - *Ordinal* – is data where the order or ranking of values (discrete or continuous) is important

- *Qualitative or Categorical* data - where the variable of interest is membership in a group or category

Examples – Data Types

- Label these data types as discrete/continuous/ordinal/categorical:
  - A survey asks students how many books they read last month discrete
  - A stopwatch is used to measure how long it takes runners to complete a 100-meter dash continous
  - A teacher assigns letter grades (A, B, C, D, E, F) to students ordinal
  - Students are asked to select their favourite colour categorical
  - Pain is rated on a scale from 1 (no pain) to 10 (worst pain imaginable) ordinal
  - The number of siblings each student has discrete
  - Students select their birth month categorical

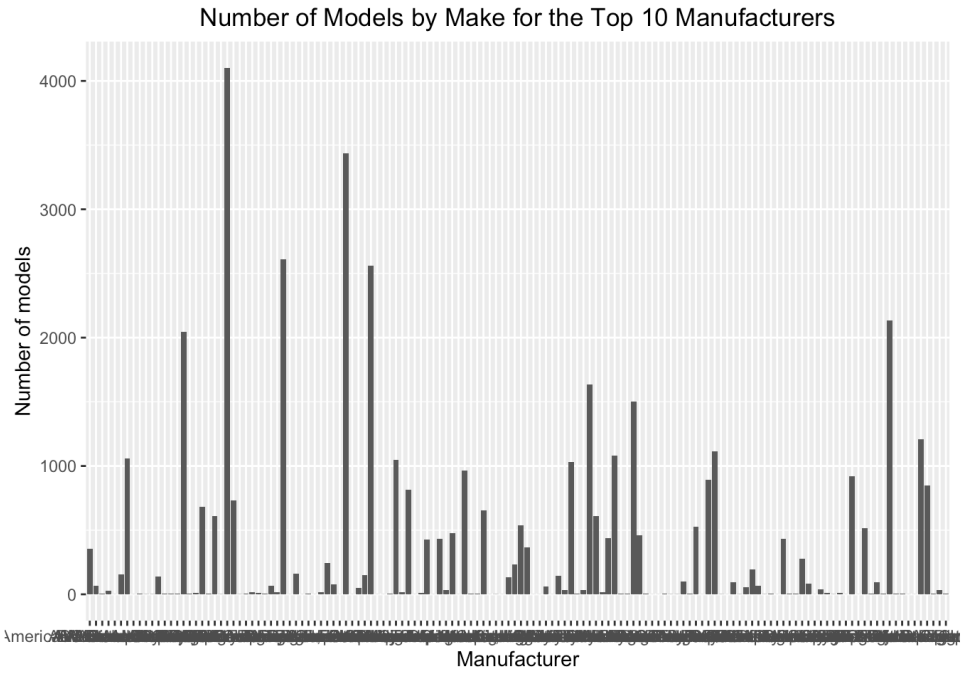
Summarising Data - Table

- A common way to summarise a dataset is via **Tables**
  - Each row corresponds to a measurement unit
  - Each column corresponds to a measurement

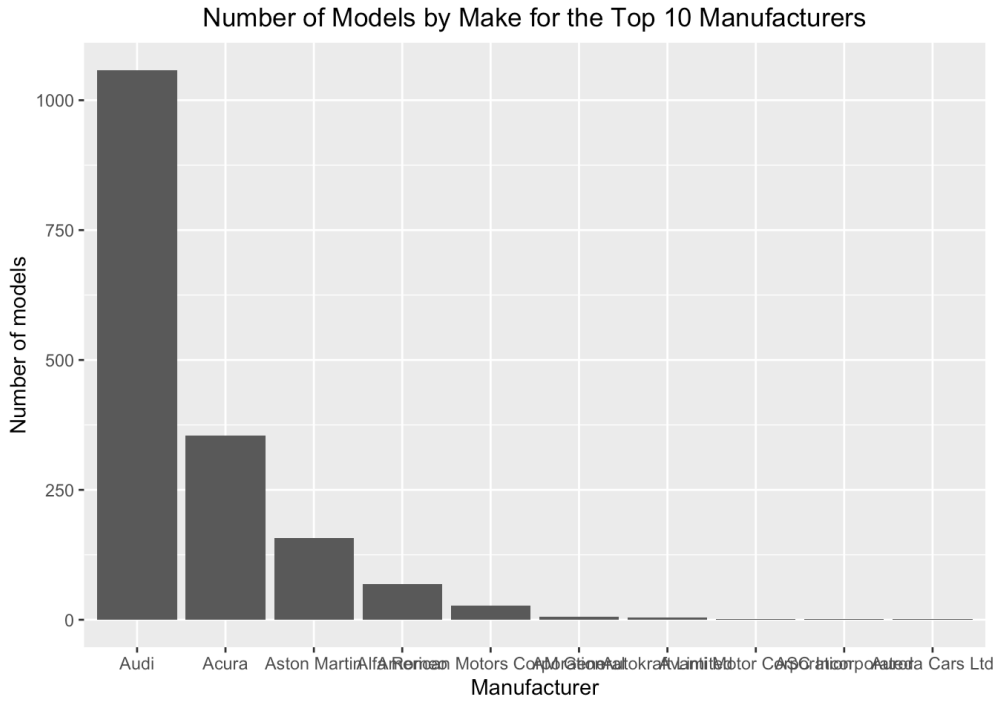
city	hwy	cy	disp	drive	make	model	trans	year
19	25	4	2.0	Rear-Wheel Drive	Alfa Romeo	Spider Veloce 2000	Manual	1985
9	14	12	4.9	Rear-Wheel Drive	Ferrari	Testarossa	Manual	1985
23	33	4	2.2	Front-Wheel Drive	Dodge	Charger	Manual	1985
10	12	8	5.2	Rear-Wheel Drive	Dodge	B150/B250 Wagon 2WD	Automatic	1985
17	23	4	2.2	4-Wheel or All-Wheel Drive	Subaru	Legacy AWD Turbo	Manual	1993

Summarising Data – Bar Chart

- A **Bar Chart** is most useful for visualising counts in categorical data

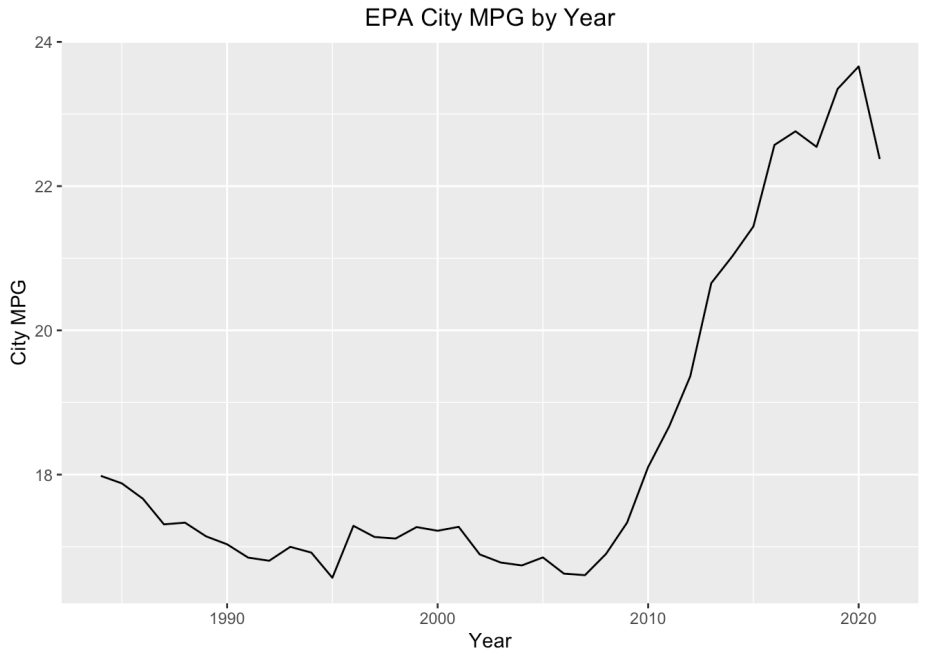


- A bar chart ordered is descending order is sometimes called a **Pareto Plot**



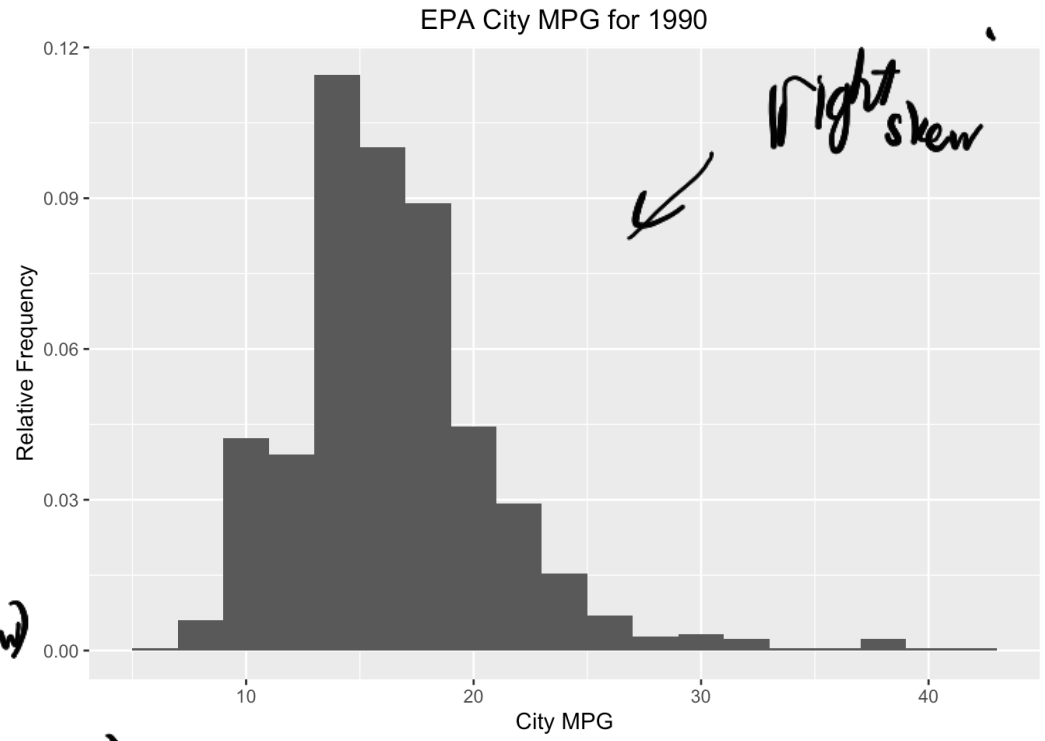
Summarising Data – Line Chart

- A **Line Chart** illustrates trend based on two quantitative variables



Summarising Data – Histogram

- A **Histogram**
  - Involves "binning" or grouping data into data ranges
  - Gives us an idea about the shape or distribution of continuous data
- Gives us an idea about the **centrality** of the data – ie where the data are "centred"
- Gives us an idea of the **skew** in the data – ie the deviation from symmetry about the centre of the data.
  - "right" skew means heavier tail to the right
  - "left" skew means heavier tail to the left



(positive skew)

(negative skew)

Examples – Summarising data

- Choose the most appropriate graph for these scenarios:
  - A class survey asks students to choose their favourite fruit: apples, bananas, grapes, or watermelon. bar chart
  - A scientist records the daily temperature in a city for 30 days. line chart (histogram if needed)
  - A researcher measures the ages of 100 people. Histogram (bar chart if needed)
  - A fitness tracker records a person's heart rate every minute during a 30-minute workout line chart
  - A film reviewer counts how many movies fall into each genre: action, drama, comedy, horror bar chart.