# Data Analysis

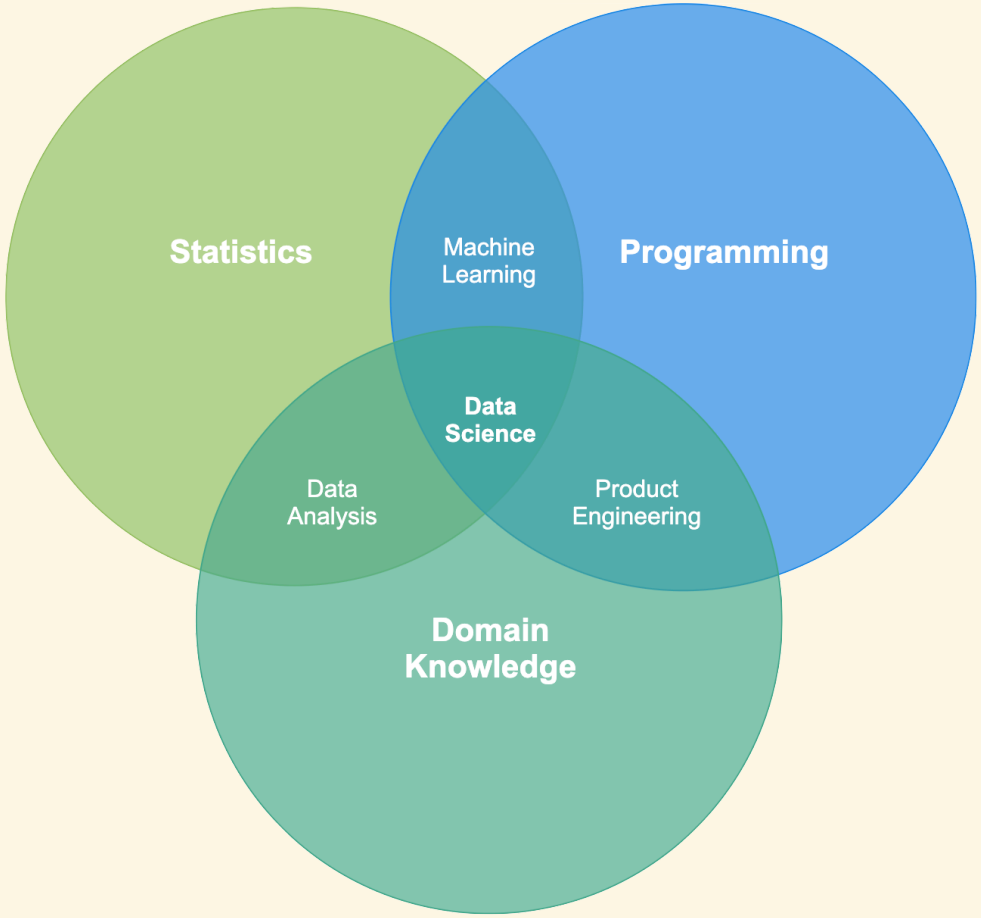
Analyzing data to extract insights and inform decisions

# Data Science

- Math (but mostly Statistics)
- Programming
- Domain Knowledge

# Data Analysis

The confluence of Math and Domain Knowledge.



# Types of Analysis

We can categorize data analysis into three main types.

## Descriptive

Summarizing or explaining trends within data.

## Predictive

Making predictions based on existing data.

## Prescriptive

Recommending actions based on predictions.

# Descriptive Analysis

Alice, a Data Scientist, pulls data on visits to her LA/NYC news-focused website inspecting:

- Times of user arrivals
- IP addresses (to identify user locations)
- Screen sizes

After aggregating by `hour`, `location`, and `screen_size`, she notices spikes in traffic from small-screen devices at:

- New York at 1pm (EDT)
- Los Angeles at 10am (PDT).

She can **describe** these traffic patterns by the times, screen sizes, and city.

# From Descriptive to Predictive

Now, Alice is planning a launch for the same site in Chicago. She uses her descriptive findings (peak times, device usage, traffic spikes) to **predict** how new Chicago readers might behave.

For example, she might guess:

- Chicago traffic will also peak around **lunch hours** on small-screen (likely mobile) devices.
- There may be similarities in local time-based usage patterns (e.g., a spike around 12pm local time).

# Prescriptive Analysis

Alice's boss, Bob, wants to **increase traffic** based on her predictions. What should they do?

- **Timing**: Schedule targeted notifications or new content releases around local midday in each city to capture users during their break.
- **Mobile Optimization**: Ensure the site is well-optimized for small screens, because spikes are seen on mobile devices.

These recommendations (or **prescriptions**) are guided by the prior descriptive and predictive insights.

# What did Alice need?

## Data Access

Alice needed web traffic logs (timestamps, locations, screen sizes).

## Aggregation & Trend Identification

Tools (like Pandas) to group the data by time/location and find peaks.

## Domain Knowledge

Interpreting results in context (e.g., understanding lunch breaks, mobile usage, local times).

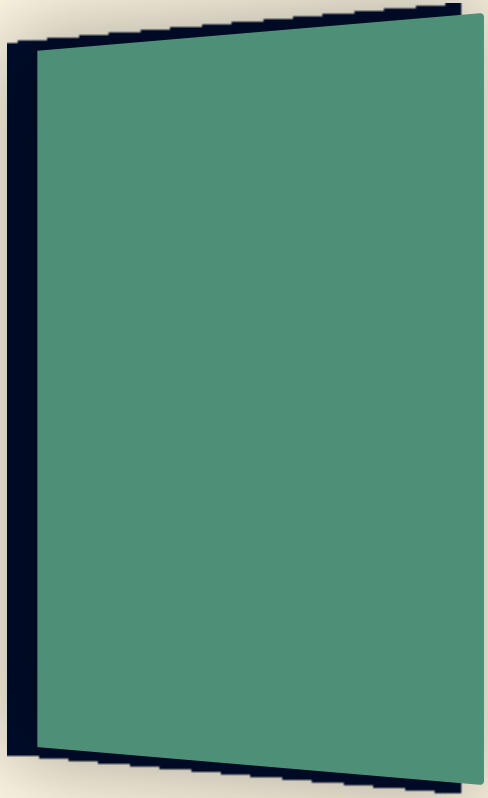
## Testable Predictions

Ability to validate guesses (e.g., comparing Chicago's actual launch metrics to predicted metrics).

# What Is the Role of a Data Analyst?

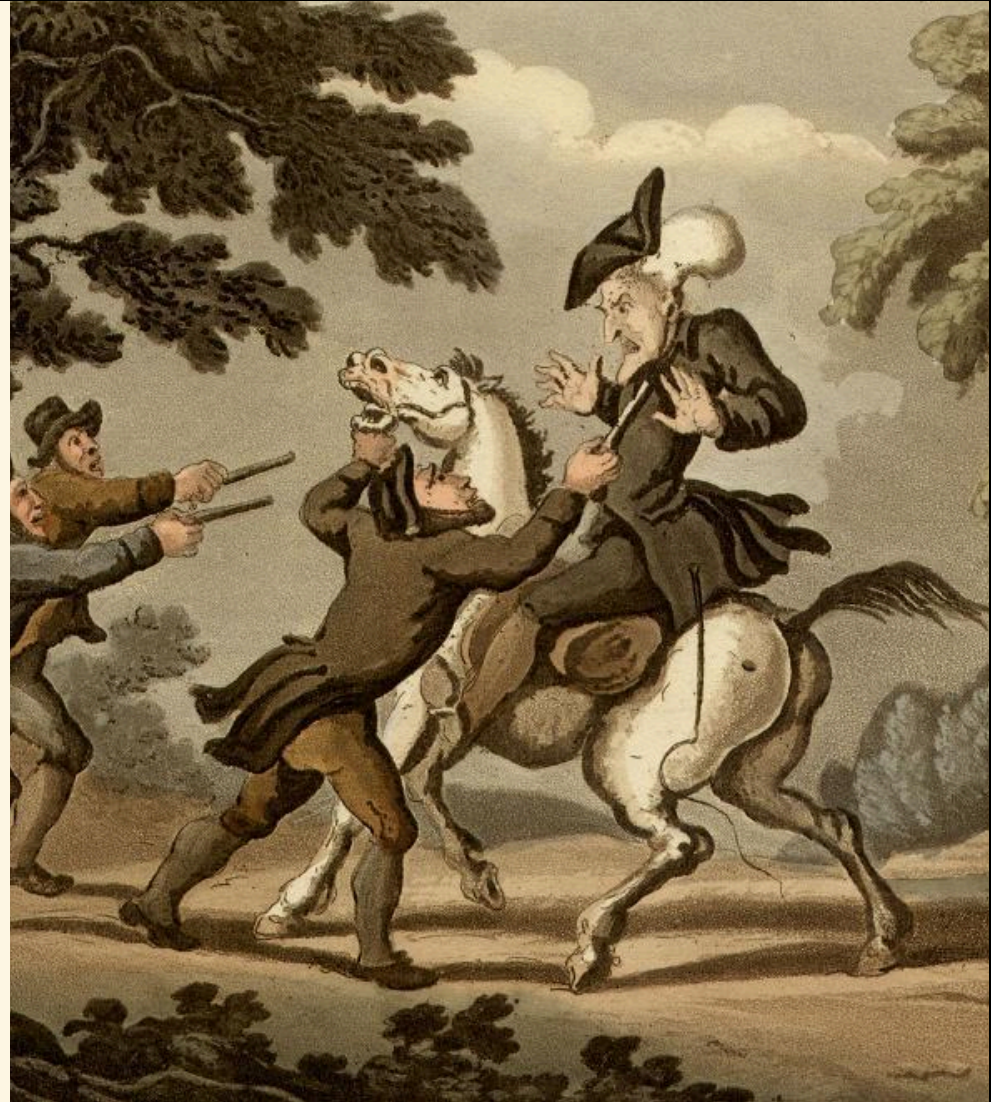
- **Data Gathering:** Like Alice, a data analyst locates, collects, and integrates data from various sources (e.g., server logs, APIs, databases).
- **Data Cleaning and Preparation:** Alice ensures the data is accurate and consistent by handling missing values, correcting errors, and transforming formats.
- **Analysis and Modeling:** Using statistical methods or machine learning, Alice explores the data to uncover patterns and build predictive or descriptive models.
- **Domain Knowledge:** Just like Alice uses her understanding of web traffic behavior, domain expertise helps interpret results meaningfully.
- **Communication and Storytelling:** A crucial part of Alice's role is explaining insights—often through visualizations, reports, or presentations—to guide decision-making.
- **Prescriptive Action:** Finally, based on findings, Alice recommends actions or strategies (e.g., the best time to launch a campaign) to stakeholders like her boss, Bob.

# Data



# Statistics: A Matter of State

- In the 18th century, as empires expanded, governments collected data on population, land, and resources for administration.
- The word "statistics" initially meant **knowledge of the state** — counting births, deaths, taxes, etc.
- In the beginning, statistics was really just about collecting **datasets**.



# What Is a Dataset?

- A **dataset** is a structured collection of data, often presented in tabular form with rows representing different observations (e.g., users, timestamps) and columns representing features or variables (e.g., IP address, screen size, city).
- The goal of **describing** a dataset is to understand its key characteristics—like typical values, spread, and the relationships among variables.

Taxes and Home Prices
http://lib.stat.cmu.edu/DASL/Stories/hometax.html

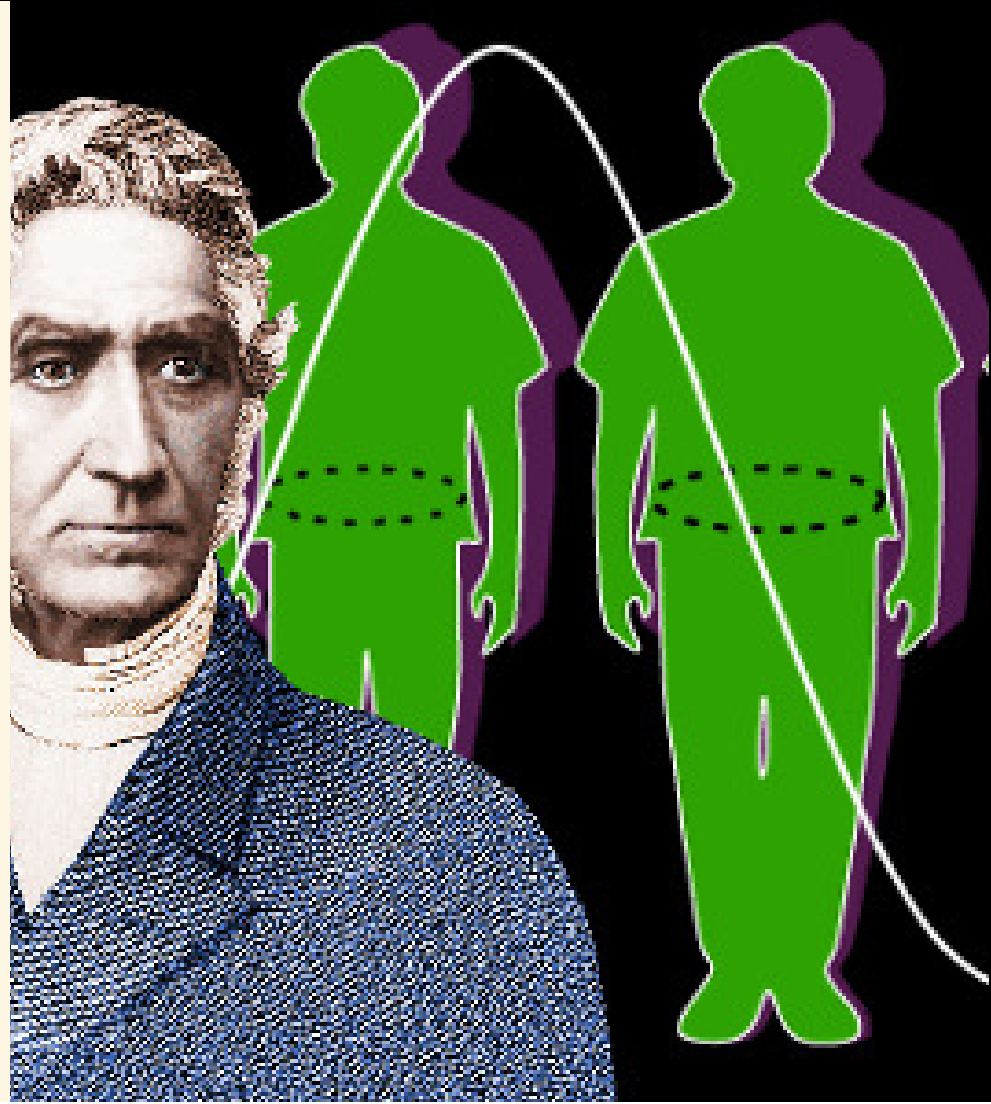
| House | Sale price (100$) | Size (sqft) | Age (years) |
|---|---|---|---|
| Avalon | 2050 | 2650 | 13 |
| Cross Winds | 2080 | 2600 | * |
| The White House | 2150 | 2554 | 6 |
| The Rectory | 2150 | 2921 | 3 |
| Larchwood | 1999 | 2580 | 4 |
| Orchard House | 1900 | 2580 | 4 |
| Shangri-La | 1800 | 2774 | 2 |
| The Stables | 1560 | 1920 | 1 |
| Cobweb Cottage | 1450 | 2150 | * |
| Nairn House | 1449 | 1710 | 1 |

1 2 3 4 5

# Adolphe Quetelet

(1796–1874)

- Adolphe Quetelet was a Belgian astronomer.
- In 1835, he published "Sur l’homme", applying astr. methods to human data.
- He discovered that human characteristics (height) followed a "normal distribution"
- This inspired Quetelet to apply the same averaging techniques to human and social data, such as crime rates and physical measurements, laying the groundwork for ideas like the "Average Man" and BMI.



The determination of the average man is not merely a matter of speculative curiosity; it may be of the most important service to the science of man and the social system. It ought necessarily to precede every other inquiry into social physics, since it is, as it were, the basis.

The average man, indeed, is in a nation what the centre of gravity is in a [celestial] body; it is by having that central point in view that we arrive at the apprehension of all the phenomena of equilibrium and motion.

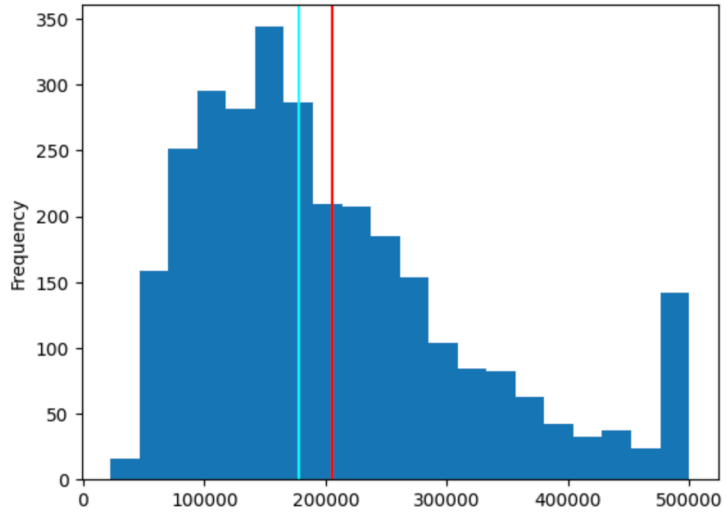
— Adolphe Quetelet, A Treatise on Man and the Development of His Faculties

# Measures of Central Tendency

1. **Mean** The arithmetic average. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. **Median** The middle value in an ordered set. Often more robust to outliers than the mean.
3. **Mode** The most frequently occurring value.

```
[23]  1 plt = df["median_house_value"].plot(kind="hist", bins=20)
      2 plt.axvline(df["median_house_value"].mean(), color="red")
      3 plt.axvline(df["median_house_value"].median(), color="cyan")
```

```
<matplotlib.lines.Line2D at 0x7a7afc5b13f0>
```

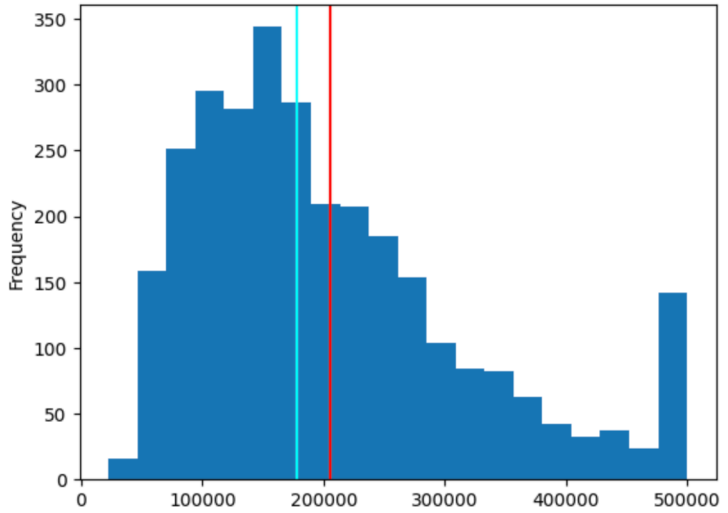


# Mean vs. Median

- Outliers or skewed data can heavily distort the mean. The median is more stable against extreme values.
- Notice how a few large values can **pull the mean** to the right, whereas the median stays near the center of most data points.
- Generally, **median** is preferred for skewed distributions.

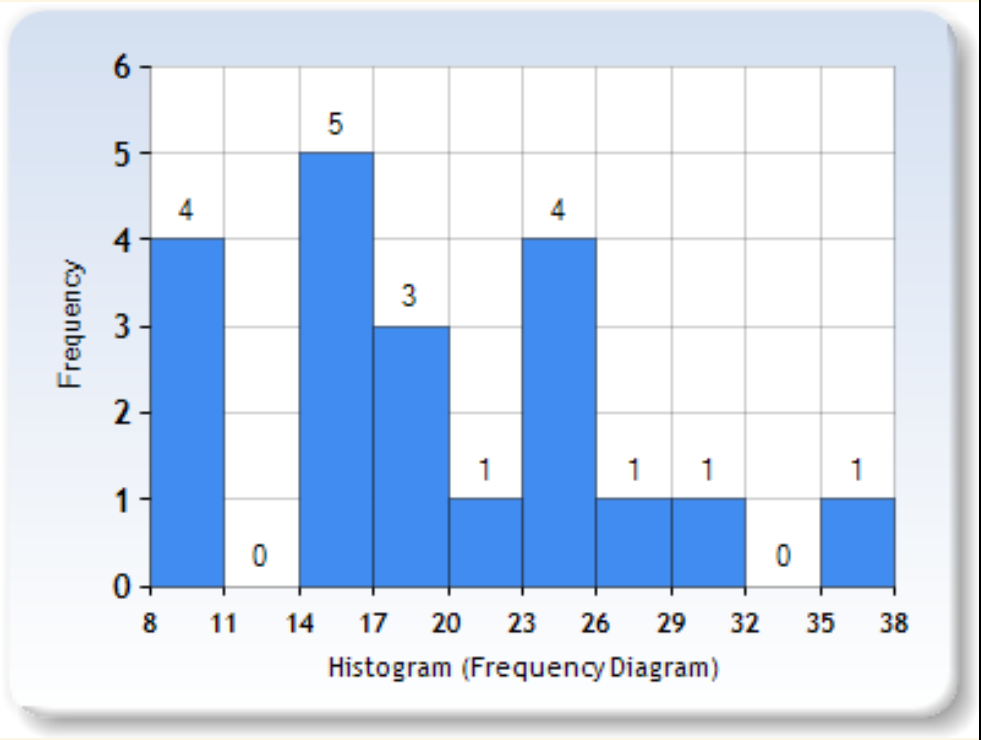
```
[23]  1 plt = df["median_house_value"].plot(kind="hist", bins=20)
      2 plt.axvline(df["median_house_value"].mean(), color="red")
      3 plt.axvline(df["median_house_value"].median(), color="cyan")
```

```
<matplotlib.lines.Line2D at 0x7a7afc5b13f0>
```



# Forever Age Distribution

- Mean: 18.8
- Median: 17.5
- Standard Deviation: 7.06064
- Lowest Score: 8
- Highest Score: 35
- Distribution Range: 27

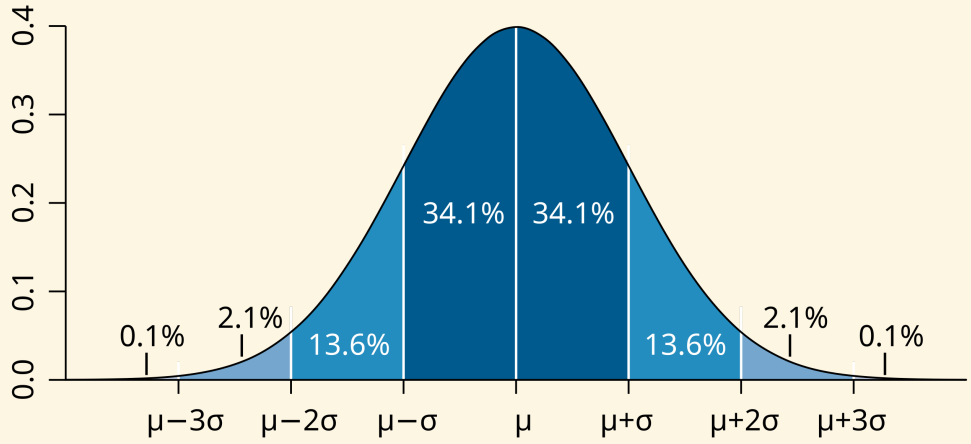


# Standard Deviation

A measure of how spread out the data is around the mean.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

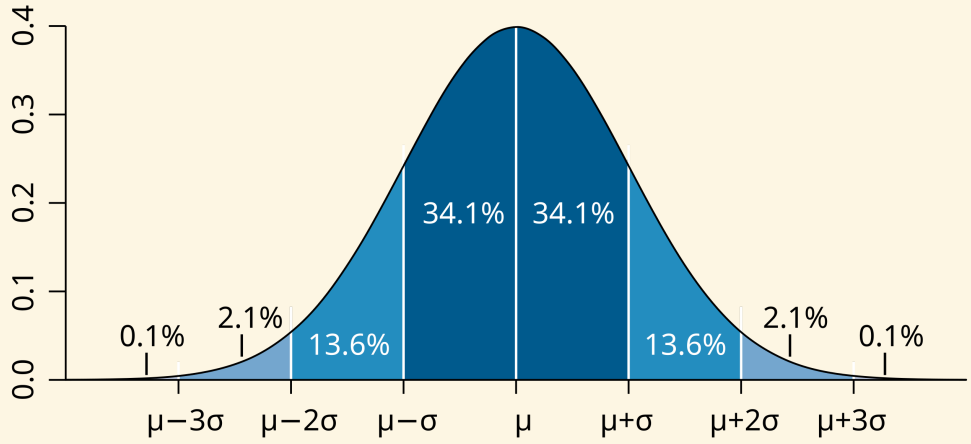
- **Low** std dev: Data points are close to the mean.
- **High** std dev: Data points are spread out over a wider range.



# The Normal Distribution

Also known as the **Bell Curve** or **Gaussian Distribution**:

- Symmetrical about the mean.
- Mean = Median = Mode.
- Approximately 68% of data within 1σ, 95% within 2σ, and 99.7% within 3σ from the mean.



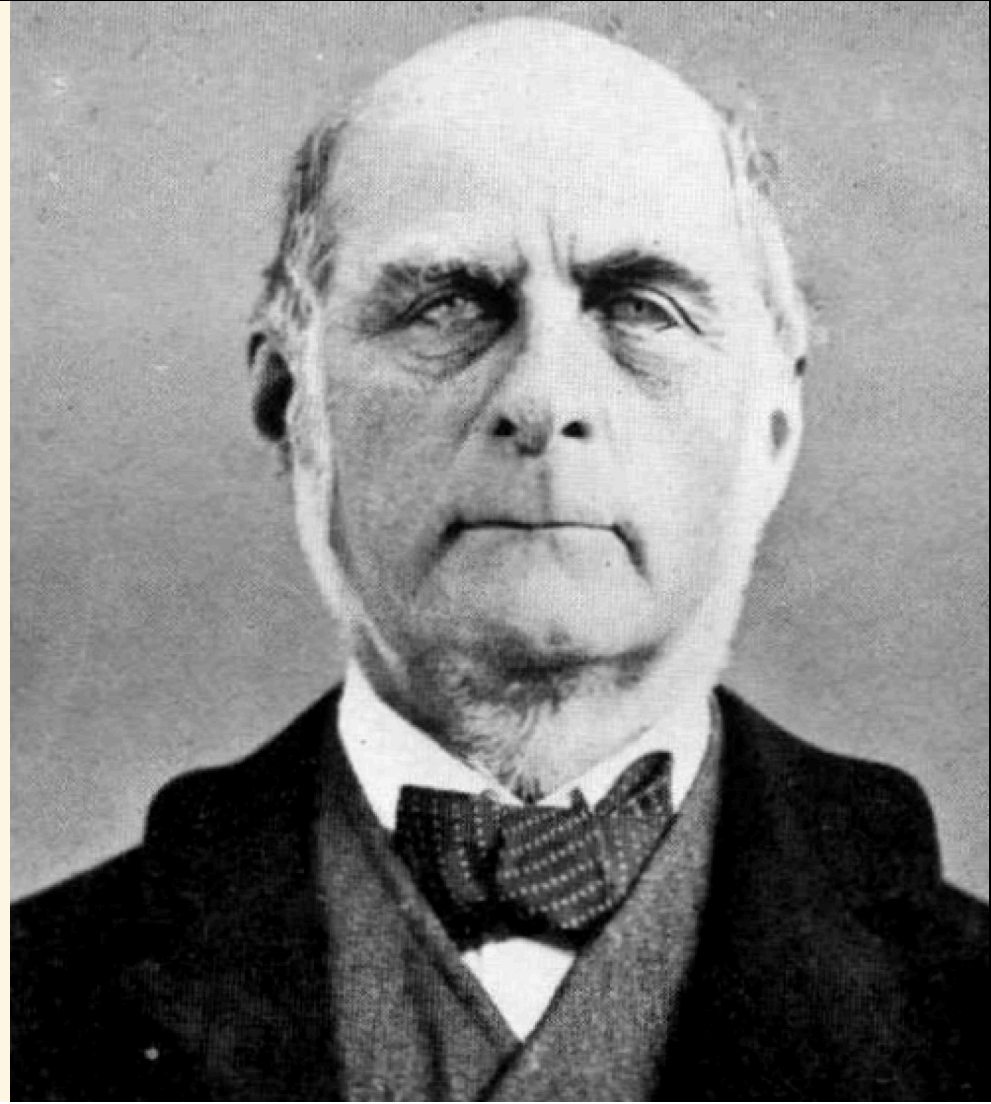
# Francis Galton (1822–1911)

## The Good

- Expanded on Quetelet's ideas.
- Introduced **regression** and **correlation**.
- Created many statistical tools we still use.

## The Ugly

- Influenced by his cousin, Charles Darwin.
- Aimed to rank individuals within distributions (blame him for standardized tests).
- Believed intelligence and ability were primarily inherited.
- Coined the term **"eugenics"** to describe



"We want abler commanders, statesmen, thinkers, inventors, and artists.

The natural qualifications of our race are no greater than they used to be in semi-barbarous times, but the conditions amid which we are born are vastly more complex than of old."

~ Francis Galton

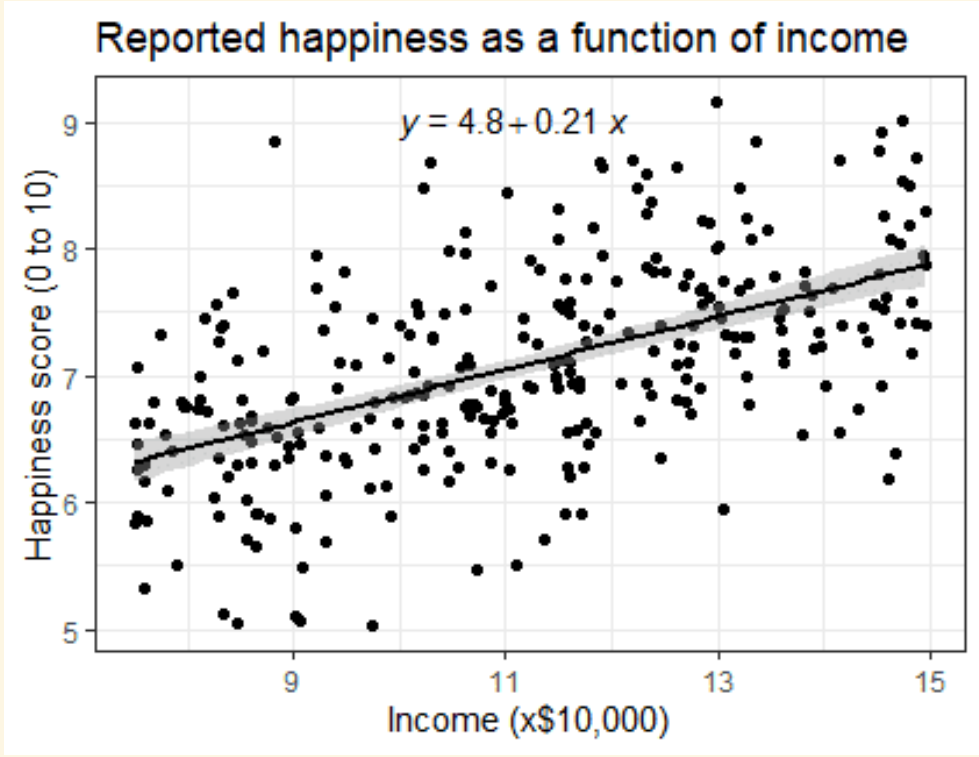
# Linear Regression

A statistical method for modeling relationships between a **dependent variable** and one or more **independent variables**. The simplest form is:

$$y = mx + b$$

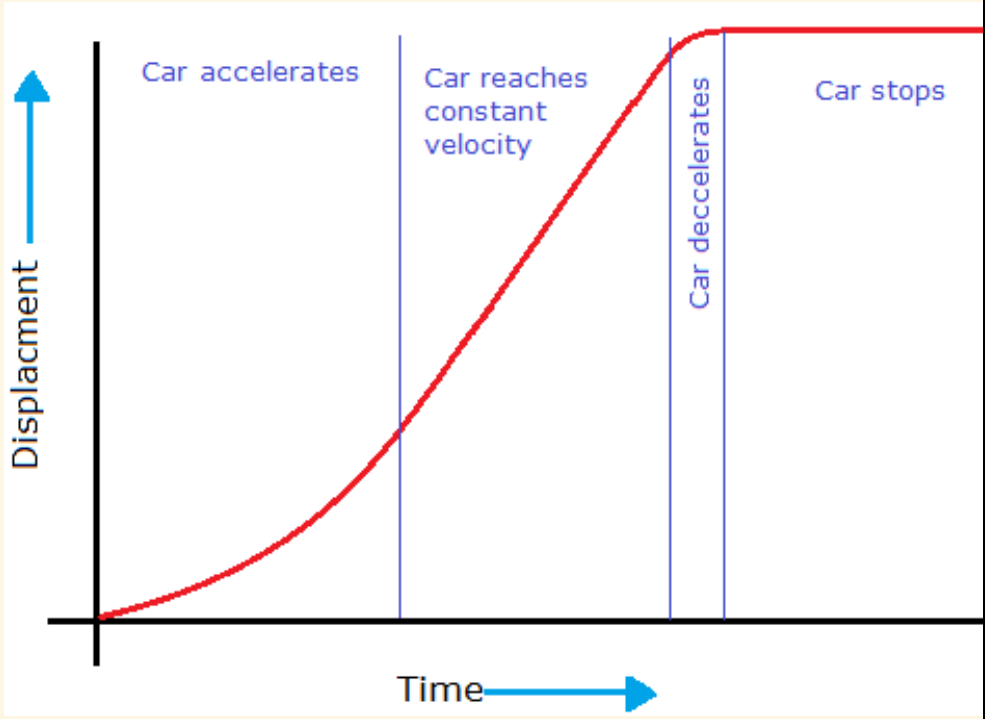
Where $m$ is the slope and $b$ is the intercept.

*Galton originally studied how children's heights regressed toward the mean height of the population.*



# IVs vs. DVs

- **Independent Variable (IV):** The factor you believe influences or causes changes (x-axis).
- **Dependent Variable (DV)**: The outcome that responds to changes in the IV(s) (y-axis).
- **Regression** is a way to model and measure the relationship between the IV(s) and DV.




## Example: Car Travel

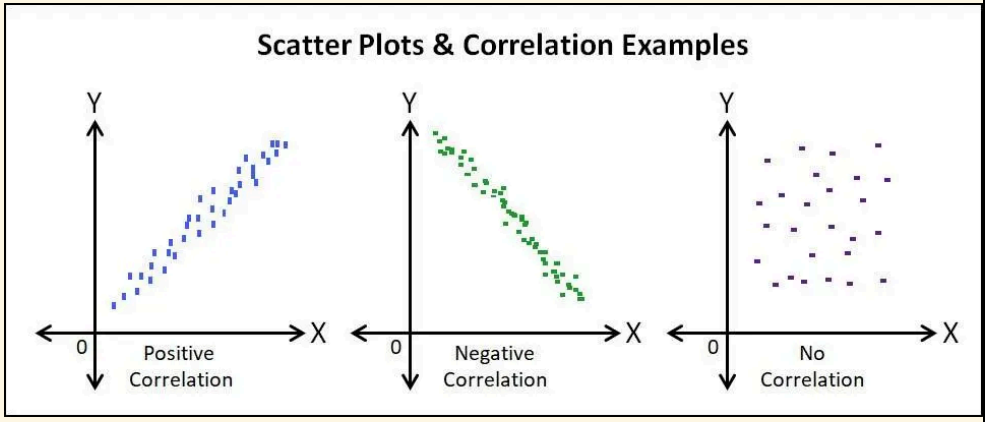
- **IV:** Time driving (minutes, hours)
- **DV:** Distance from start point
- **Regression:** Estimated speed (slope) – how fast distance changes over time.

# Correlation

Indicates the strength and direction of a linear relationship between two variables.
**Pearson's correlation coefficient** ($r$):

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

- **+1**: Perfect positive correlation.
- **0**: No linear correlation.
- **-1**: Perfect negative correlation.




## Covariance

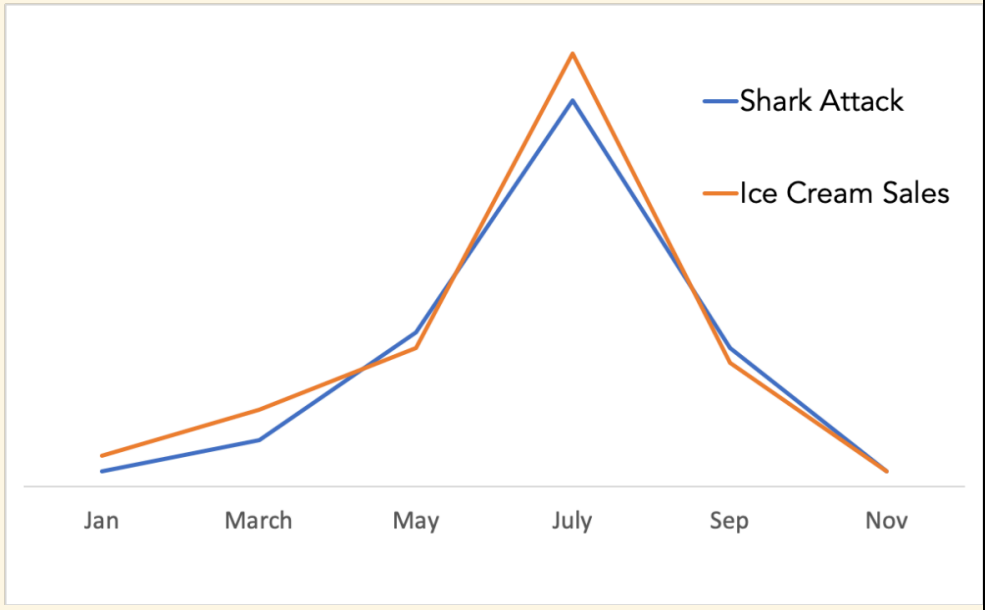
How two variables change together:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

# Correlation DOES NOT IMPLY Causation

- **Correlation**: Two variables move together.
- **Causation**: One variable causes the other to change.

It's easy to find **spurious correlations** that are not meaningful.



# spurious correlations

*correlation is not causation*

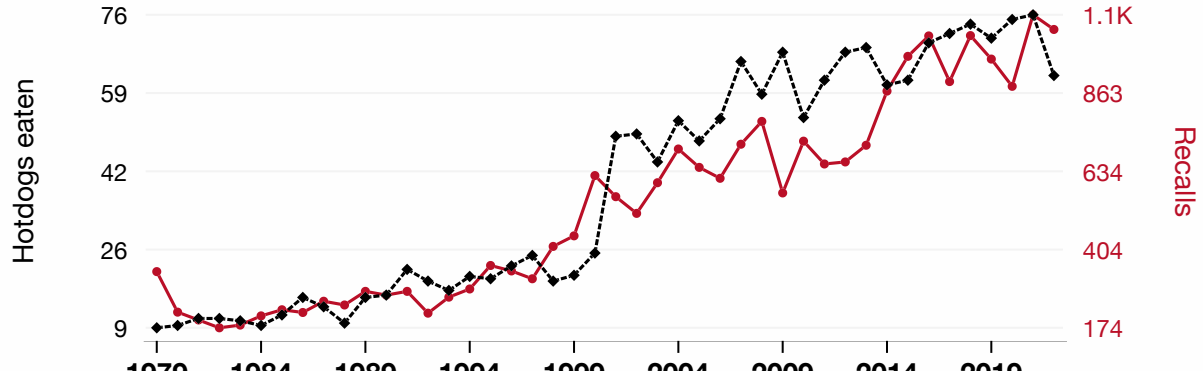
random · discover · *next page →*

don't miss spurious scholar,
where each of these is an academic paper

## Hotdogs consumed by Nathan's Hot Dog Eating Competition Champion

correlates with

## Total number of automotive recalls



# Making Sense of a Dataset

# Analyzing a New Dataset

## 1. **What Kind of Data?**

- Nominal, ordinal, continuous, discrete?

## 2. **Central Tendency**

- Mean, median, mode.

## 3. **Spread**

- Range, standard deviation, IQR, percentiles.

## 4. **Inspect Shape**

- Is it normal? Is there skew or kurtosis?

## 5. **Identify Outliers**

- 5th/95th percentiles, IQR method, z-scores.

## 6. **Look for Relationships**

- Correlation, covariance, scatter plots.