MATH 189
Data Analysis & Inference

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

Challenges in Multivariate Analysis

Objectives: What is the scientific or social problem of interest? How can it be drafted into a multivariate analysis problem? Does the explanation meet the “common sense” to the domain knowledge?

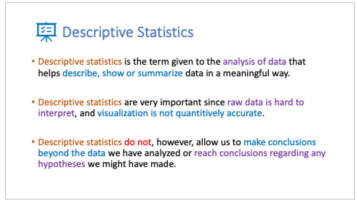
Dataset: Data is noisy (randomness). Measurement error. Missing data. Normality versus Heavy tailedness. Homogeneity vs Heterogeneity.

Methodology: Which method should we use? What are the assumptions? Statistically accurate? Computationally efficient?

Exploratory Data Analysis

- An important step in data analysis is exploratory data analysis (EDA).

- EDA methods include visualization of data, descriptive statistics and more.



- The benefits of EDA:
  1. Check the quality of data: cleaned or not, missing data, outliers, …
  2. Gain the first impression of data: data type, distribution, symmetry, …
  3. Illustrate analysis results: gain intuition, collaboration, general audience, …

Visualization of Multivariate Data

- Why do we look at graphical displays of the data?

One difficulty of understanding complex multivariate data is the human perceptional system. Visualization tools can help!

- Visualization may:
  1. suggest a plausible model for the data,
  2. assess validity of model assumptions,
  3. detect outliers or suggest plausible normalizing transformation,
  4. and more.

Scatter Plot

- A scatter plot is a data visualization tool that uses dots to represent the values obtained for two different variables.

- Plotted on Cartesian coordinates: x-axis is the value of the first variable and y-axis is the value of the second variable.

- Used to check the relationship between two variables:
  1. Correlation
  2. Linear or nonlinear
  3. Joint normality
  4. Groups

Example 1: USDA Women's Health Survey

- In 1985, the USDA commissioned a study of women's nutrition. Nutrient intake was measured for a random sample of 737 women aged 25-50 years.

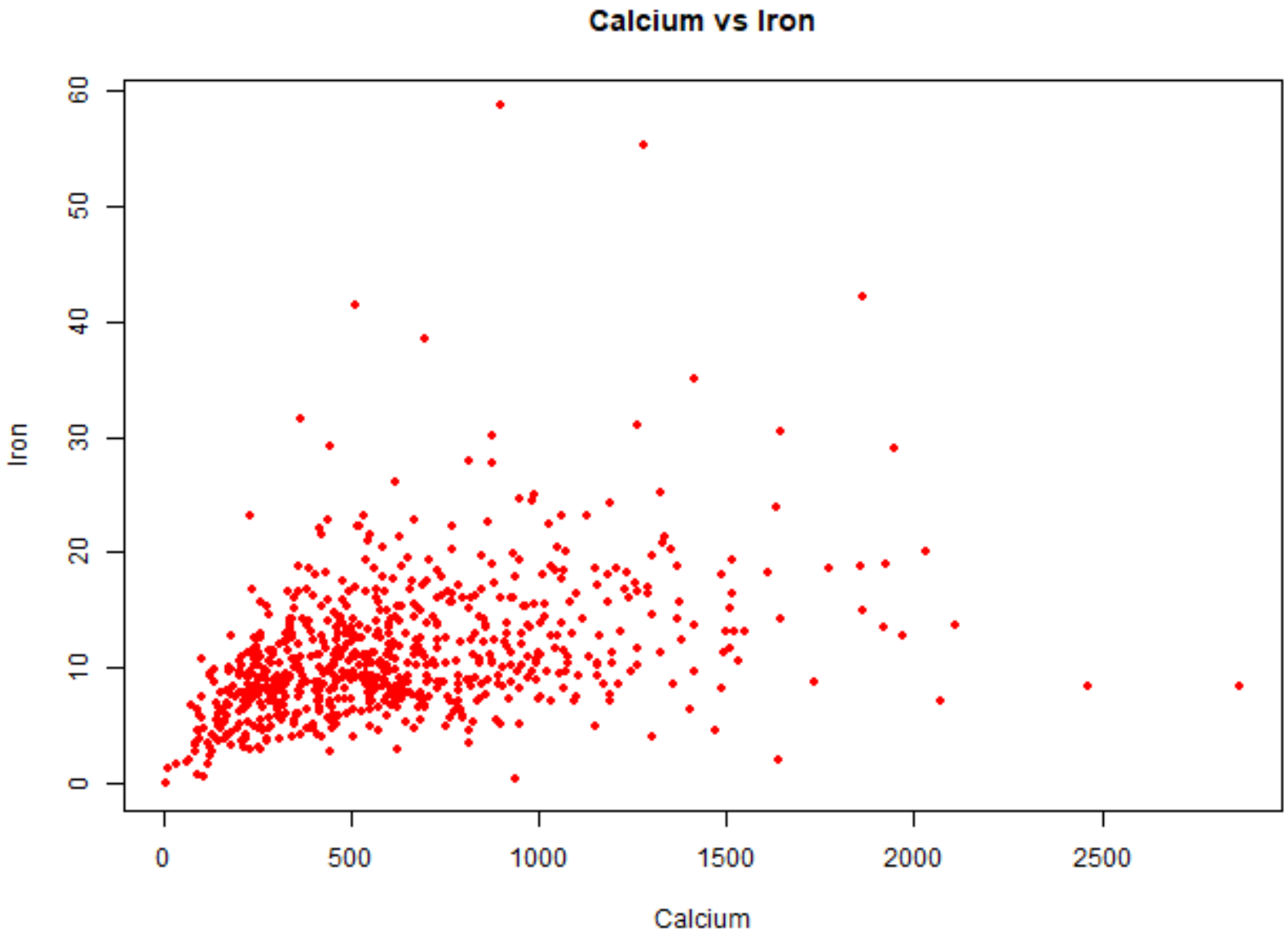
- The following variables were measured:
  1. Calcium (mg)
  2. Iron (mg)
  3. Protein (g)
  4. Vitamin A (μg)
  5. Vitamin C (mg)

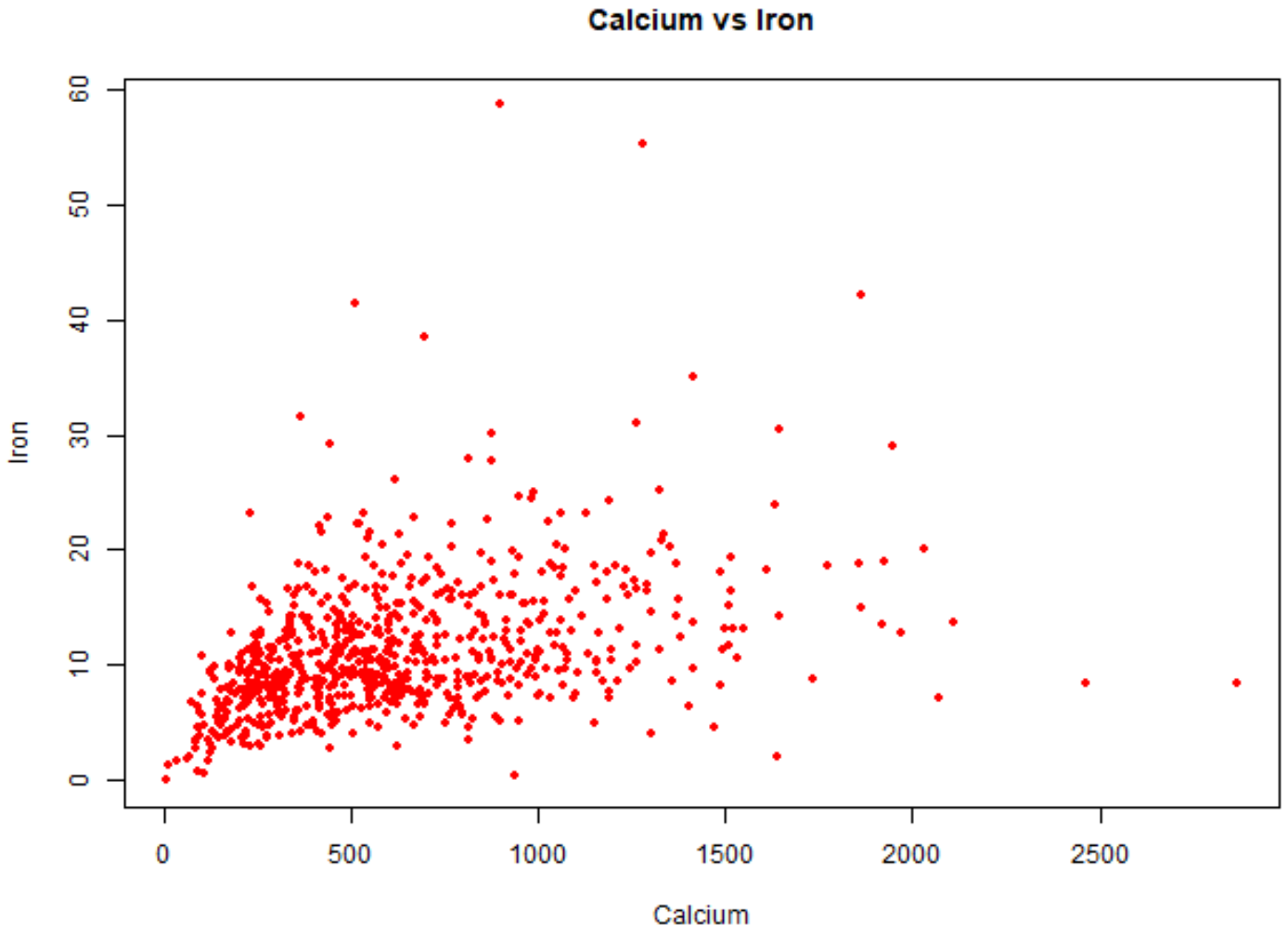
A Peek at the Data

- Dataset contains 5 variables and 737 observations.
- Table of first five observations

	Calcium	Iron	Protein	Vitamin A	Vitamin C
1	522.29	10.188	42.561	349.13	54.141
2	343.32	4.113	67.793	266.99	24.839
3	858.26	13.741	59.933	667.90	155.455
4	575.98	13.245	42.215	792.23	224.688
5	1927.50	18.919	111.316	740.27	80.961

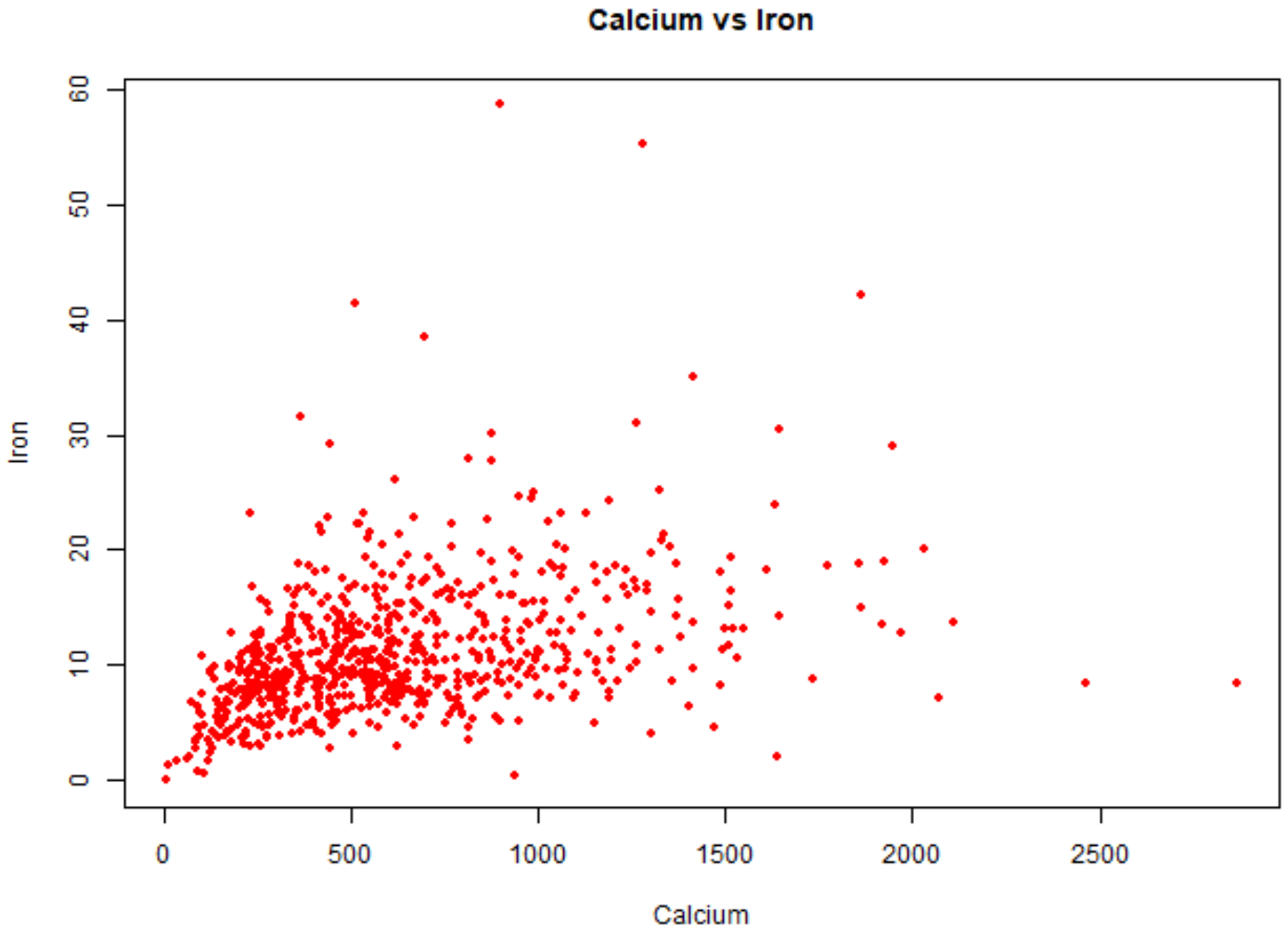
Scatter Plot between Calcium and Iron





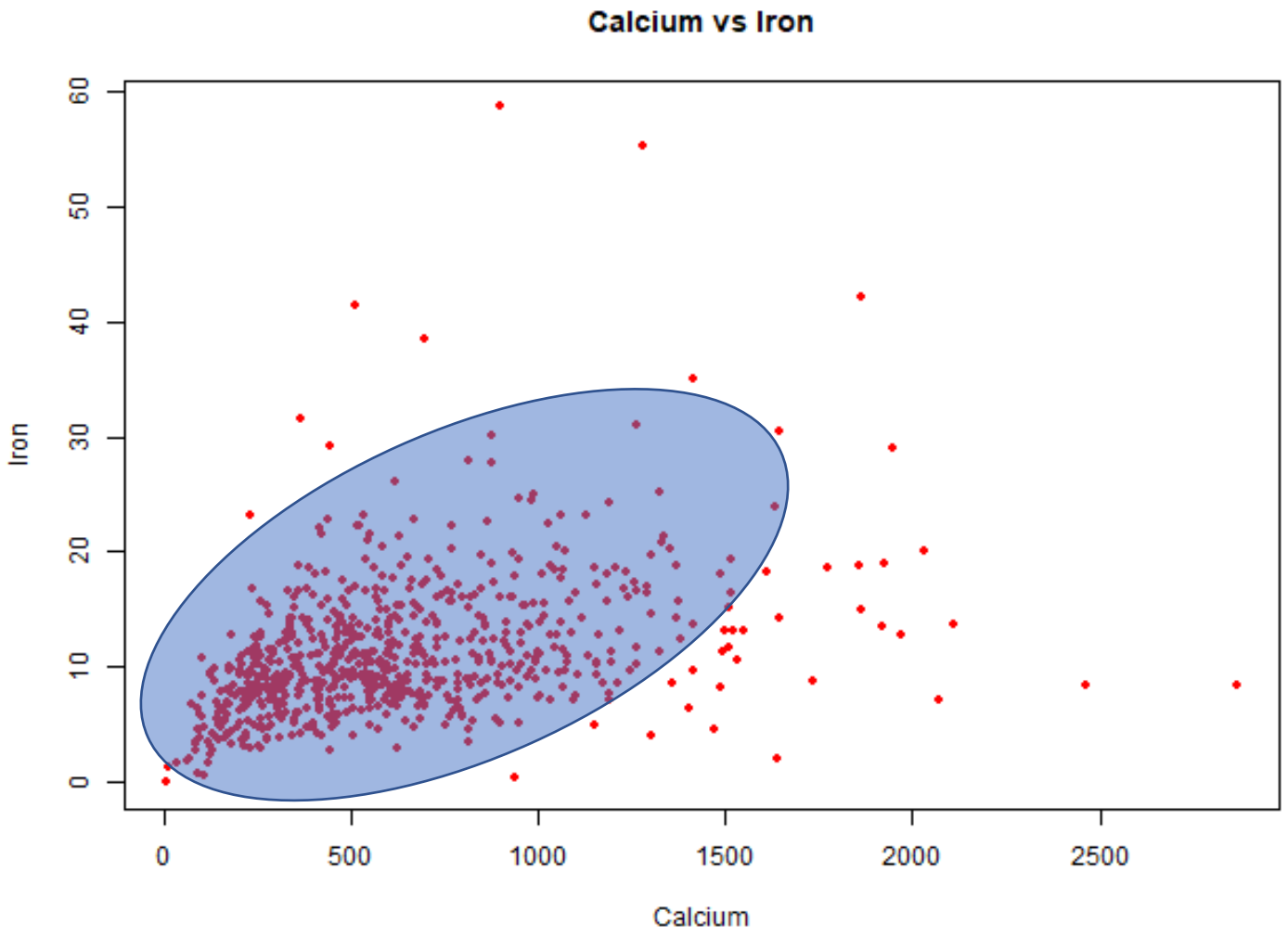
- Each red dot represents an observation
- Calcium on x-axis
- Iron on y-axis

What can we say about this plot?



Some Questions

- Are there outliers?
- What model should we use to fit the data?
- Can we model the data using bivariate normal?



- Majority part of data are clustered in the shaded area.
- Are other points outliers?
- How large is large?
- How to compare 500 in Calcium and 30 in Iron?

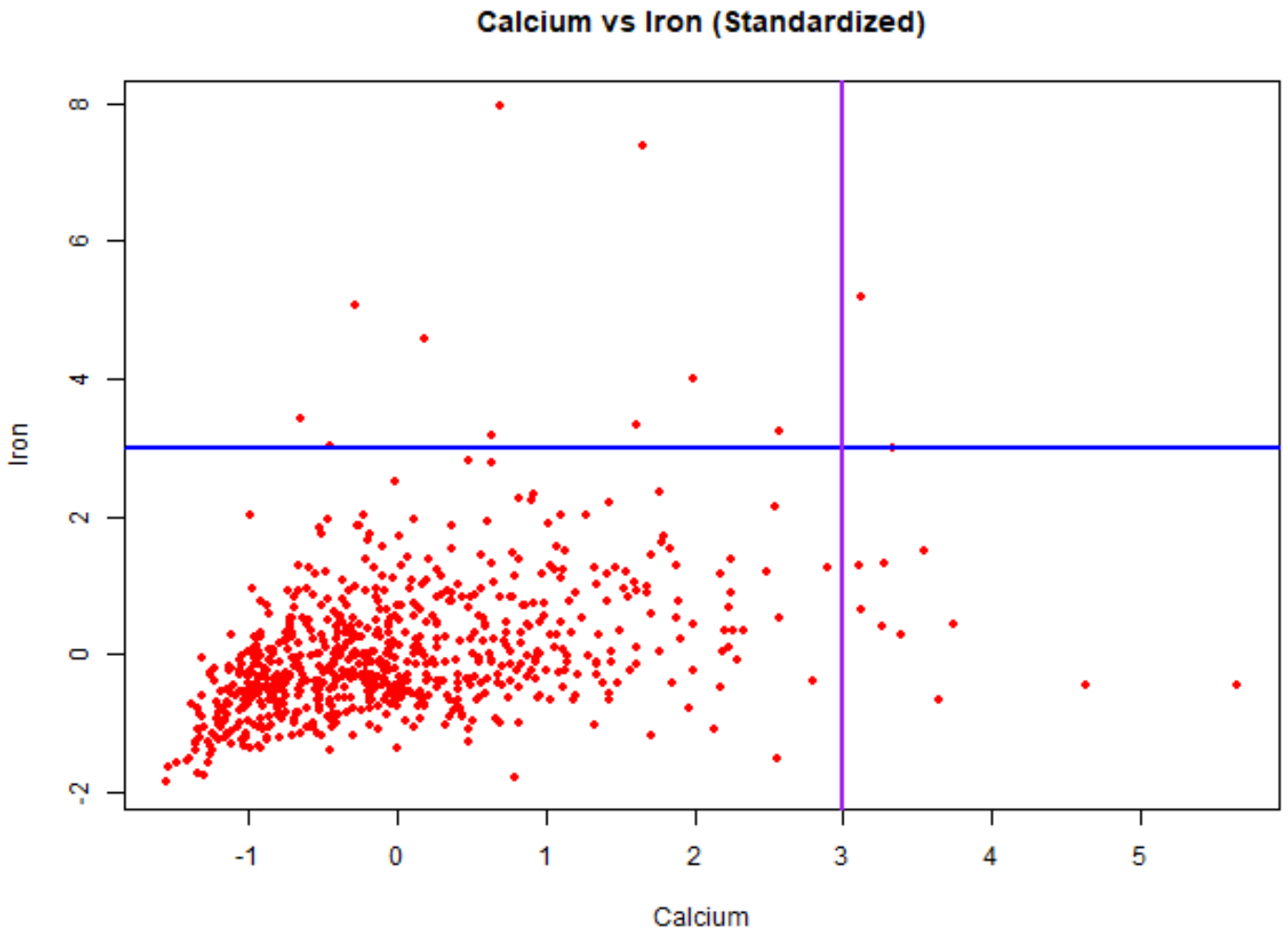
Standardize Data

- Rescale data from different sources and measures to a "standard" scale
- Avoid compare apple to pear
- A common standardization is called **Z-score scaling** which scales a random sample to have zero sample mean and unit sample variance.
- For a random sample $(x_1, \dots, x_n)$, the Z-score scaling transforms each observation by

$$x_i^* = \frac{x_i - \mu}{s},$$

where μ is the sample mean and s is sample standard deviation.

Scatter plot after Z-score standardization



- Zero sample mean and unit sample variance.
- Solid lines stand for 3 standard deviations.
- If data is normal, we should expect about 99% of data in the bottom-left box.

Transformation Methods

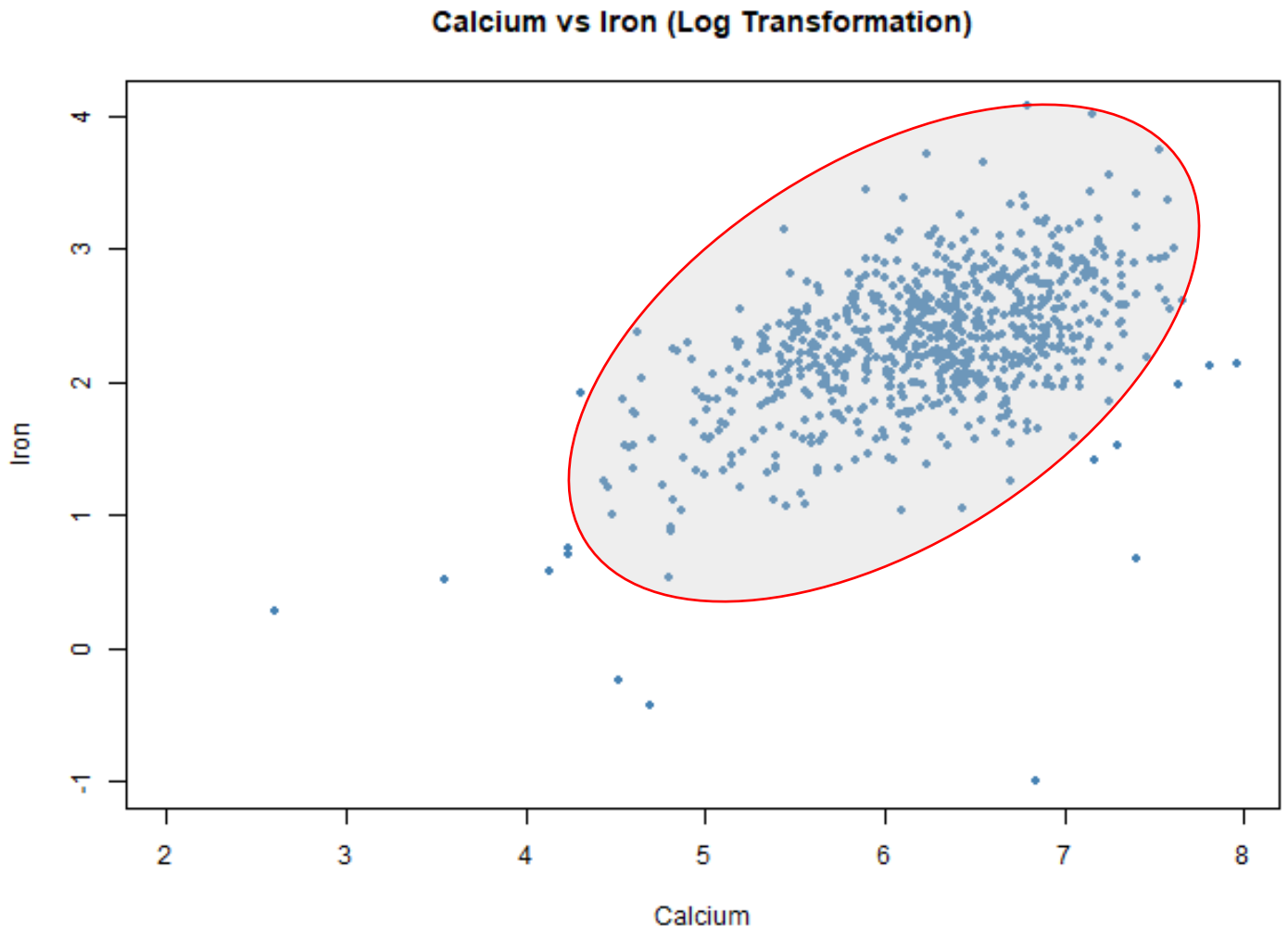
- Sometimes data is "irregular": non-normal, outliers, skewed, heavy-tailed, …
- Data transformation techniques can be used to stabilize variance, make the data more normal-like, improve the validity of measures of association

- Power transformation:

$$y = x^{\alpha}, 0 < \alpha < 1.$$

- Log transformation:

$$y = \ln(x), 0 \leq x.$$



Scatter plot after Log transformation

- More Normal-like.
- Reduced variance
- Less outliers

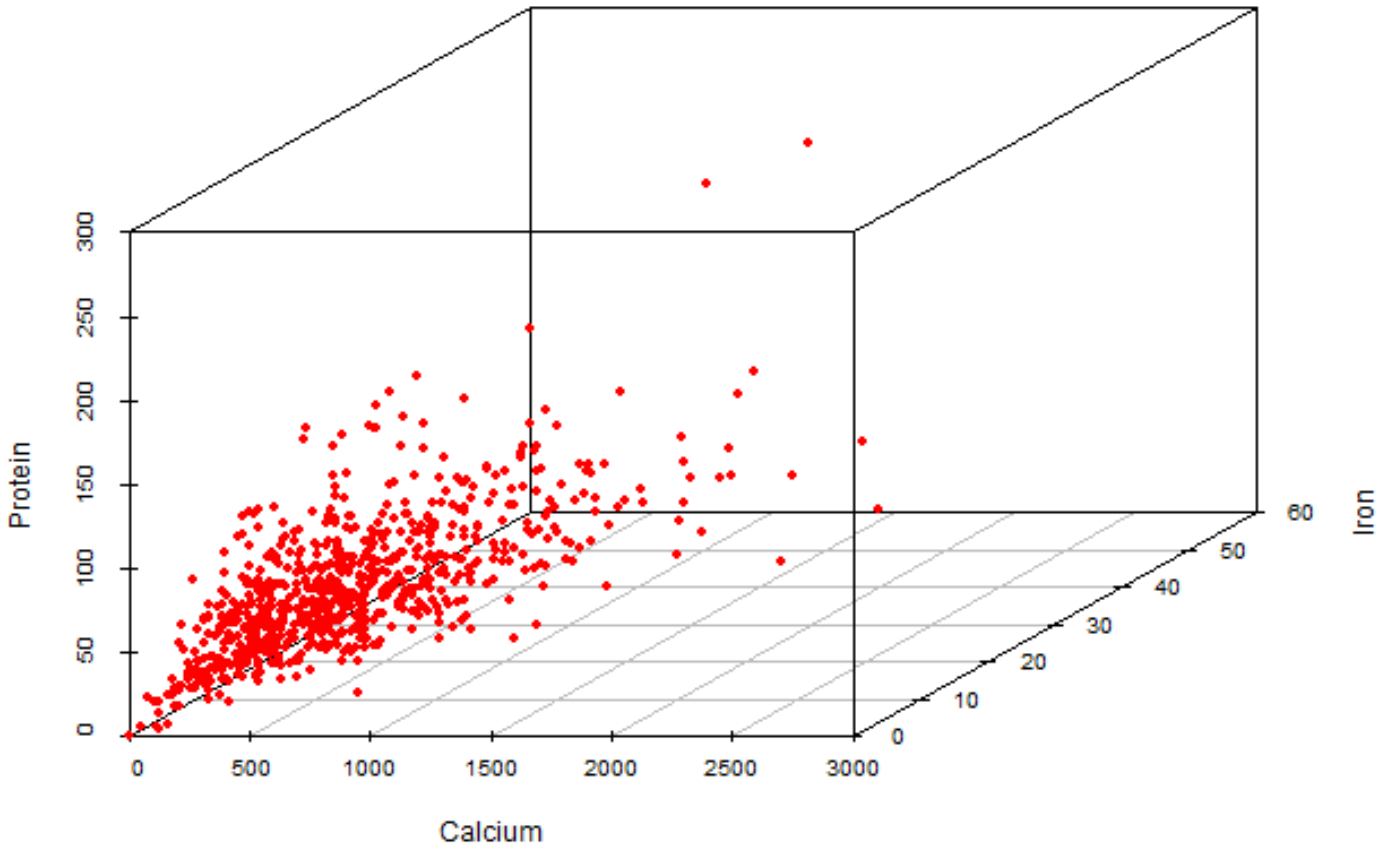
Scatter Plot for Three Variables

- The scatter plot can be extended to visualize the relationship among three different variables which is called 3D scatter plot.

- Plotted on Cartesian coordinates: x-axis is the value of the first variable, y-axis is the value of the second variable and z-axis is the value of the third variable.

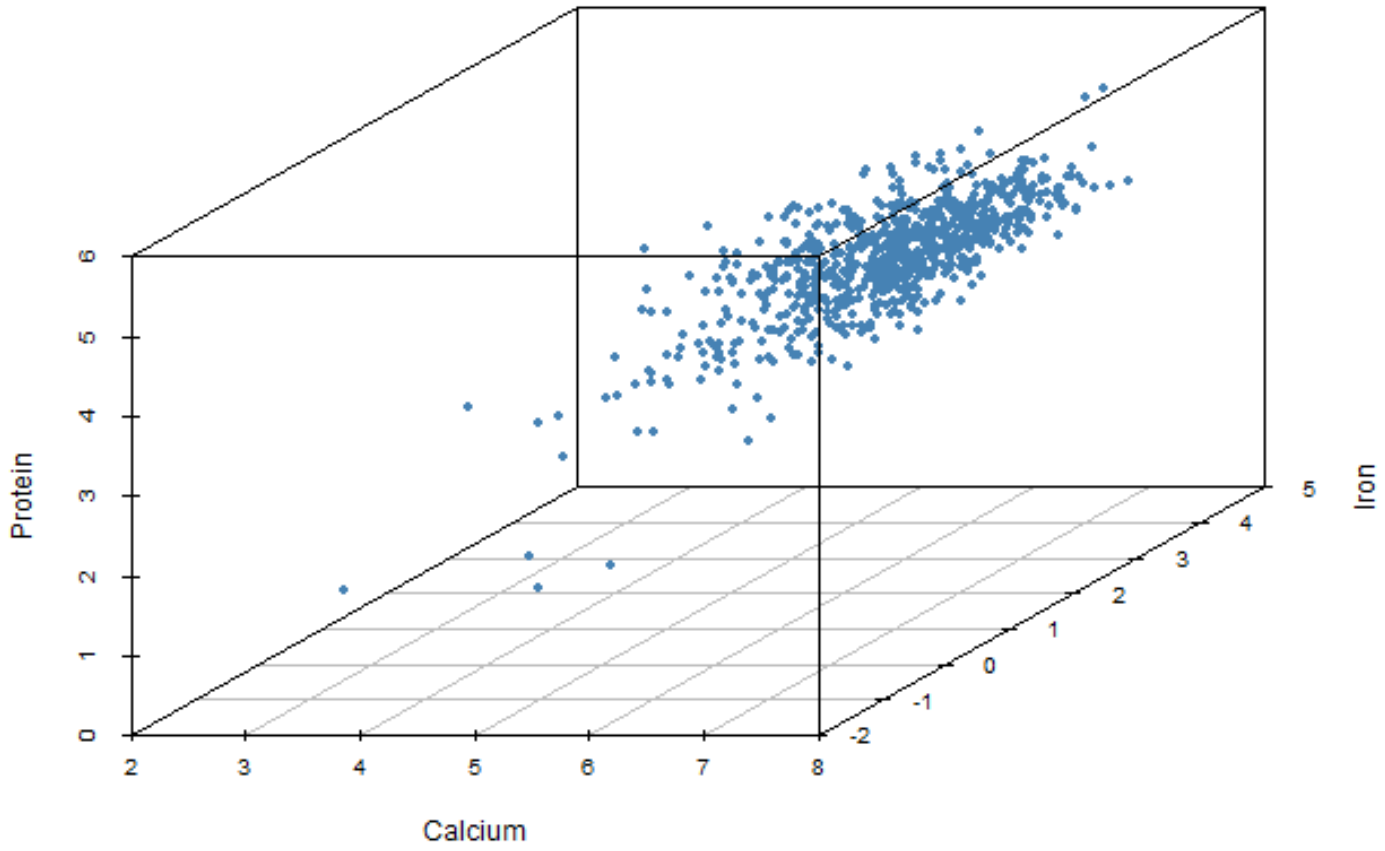
- A fourth variable can be set to denote the color or size of the markers.

3D scatter plot for Calcium, Iron and Protein



- Each red dot represents an observation
- Calcium on x-axis
- Iron on y-axis
- Protein on z-axis

3D scatter plot after Log transformation



- Each blue represents an observation
- More clustered in a “ball” or “ellipse”
- Reduced variance
- Less outliers

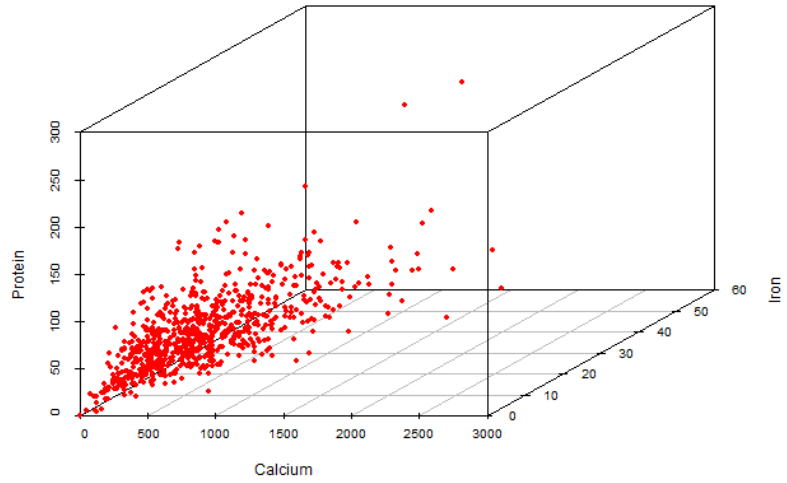
Pros and Cons of 3D Scatter Plot

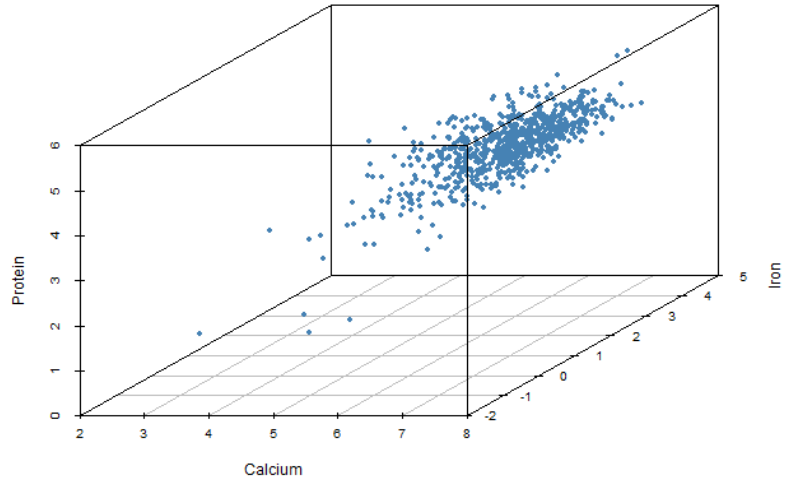
Pros

- Visualization for 3 or 4 variables.
- Complex relationship rather than pair wise
- Joint sample distribution

Cons

- Not friendly to bare eyes (angle dependent)
- Hard to interpret
- Not working for more (≥ 5) variables



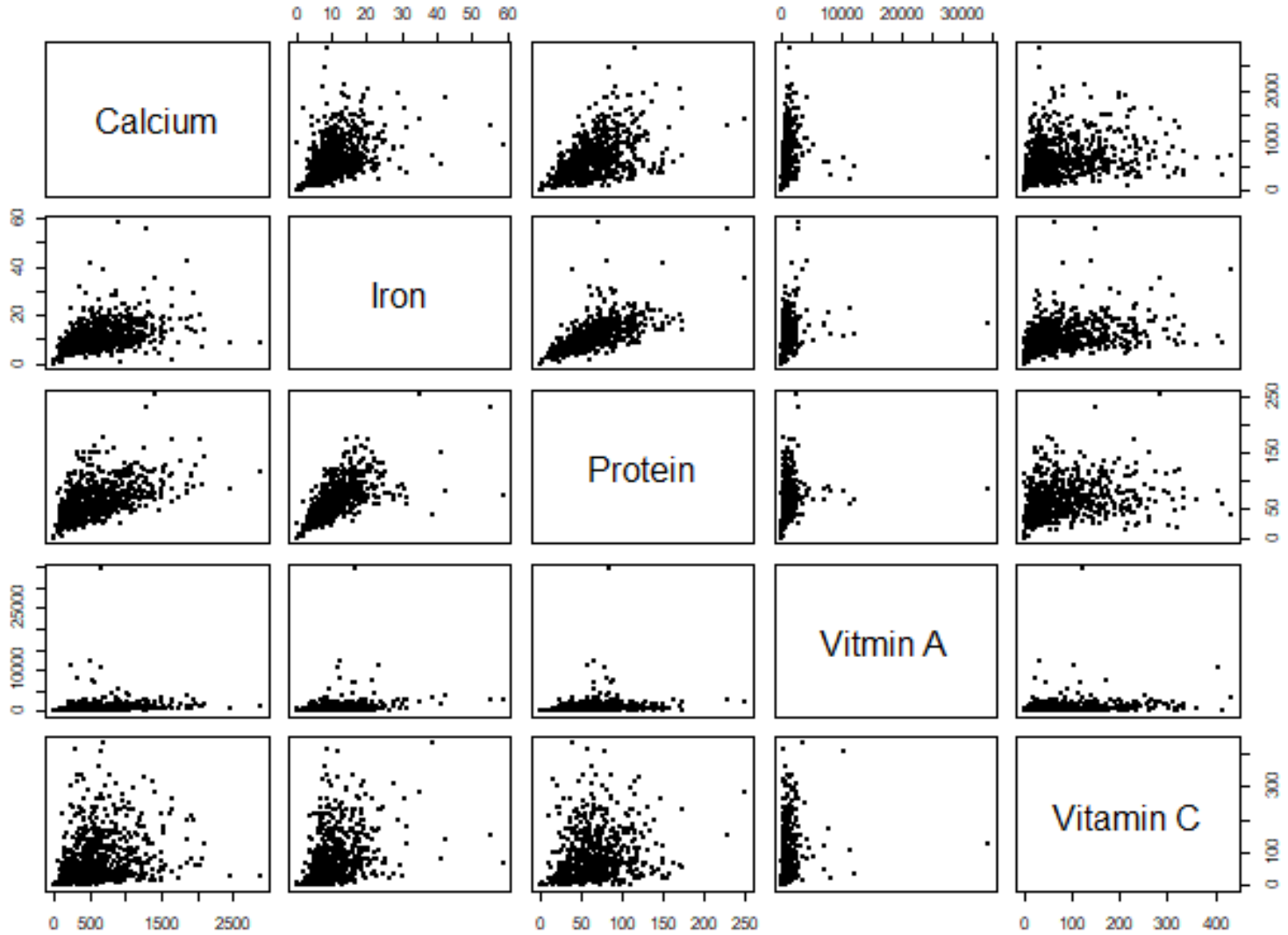


Pairwise Scatter Plot for More Variables

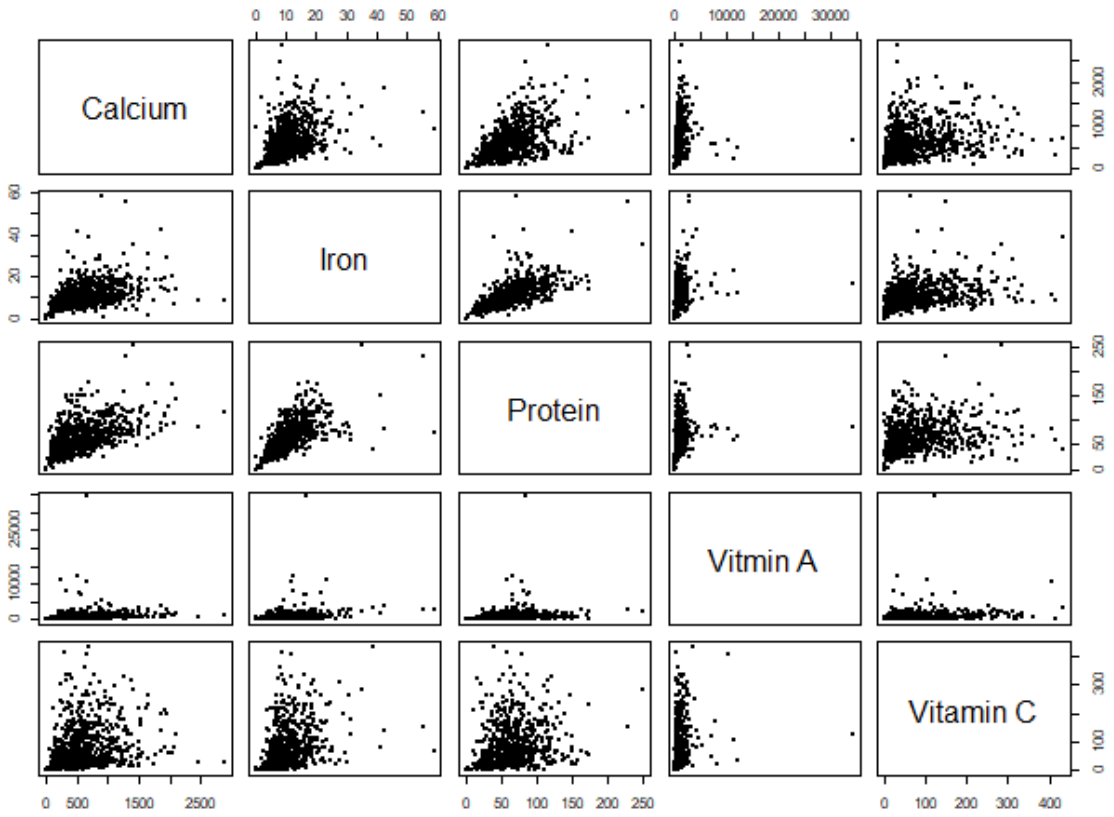
- The pairwise scatter plot aims to visualize the relationship for each pair of variables in a multivariate dataset.

- A pairwise scatter plot is an array of scatter plots, the (i, j)-th plot in the array is the scatter plot between the i-th and j-th variables.

Pairwise Scatter Plot for USDA Women's Health Survey



Pros and Cons of Pairwise Scatter Plot



Pros

- Visualization for many variables simultaneously
- Interpret pairwise relationships

Cons

- No joint relationship for more than 2 variables
- Huge array when the number of variables is large

Time Plot

- A time plot (sometimes called a time series graph) displays values versus time. It is similar to scatter plot, but x-axis is chosen to be time (or age, survival time …).

- A time plot is useful to compare the "growth" of multiple variables with respect to time (or some other common index).

- Application of time plot:
  1. Finance: compare multiple stock returns vs time
  2. Clinical: compare multiple patients vs survival time
  3. Biology: compare multiple genome sequences vs positions
  4. Physics: multiple measurements vs time

Time Plot for Financial Time Series

- Time series are one of the most common data types encountered in finance and weather forecasting

- One powerful yet simple visualization tool in financial analysis is to draw the time plot for multiple assets.

- Things to check in a time plot:
  1. Co-movement of variables
  2. Trend (increasing or decreasing …)
  3. Periodical patterns (weekly, seasonal, long term …)
  4. Black swan event (huge gap, financial crisis …)

Example 2: Stock Prices of High-tech Companies

- We collect daily stock prices (closing price) of four leading high-tech companies between January 2018 and January 2019.

- The following variables were included:
  1. Daily stock price of Apple
  2. Daily stock price of Facebook
  3. Daily stock price of IBM
  4. Daily stock price of Microsoft

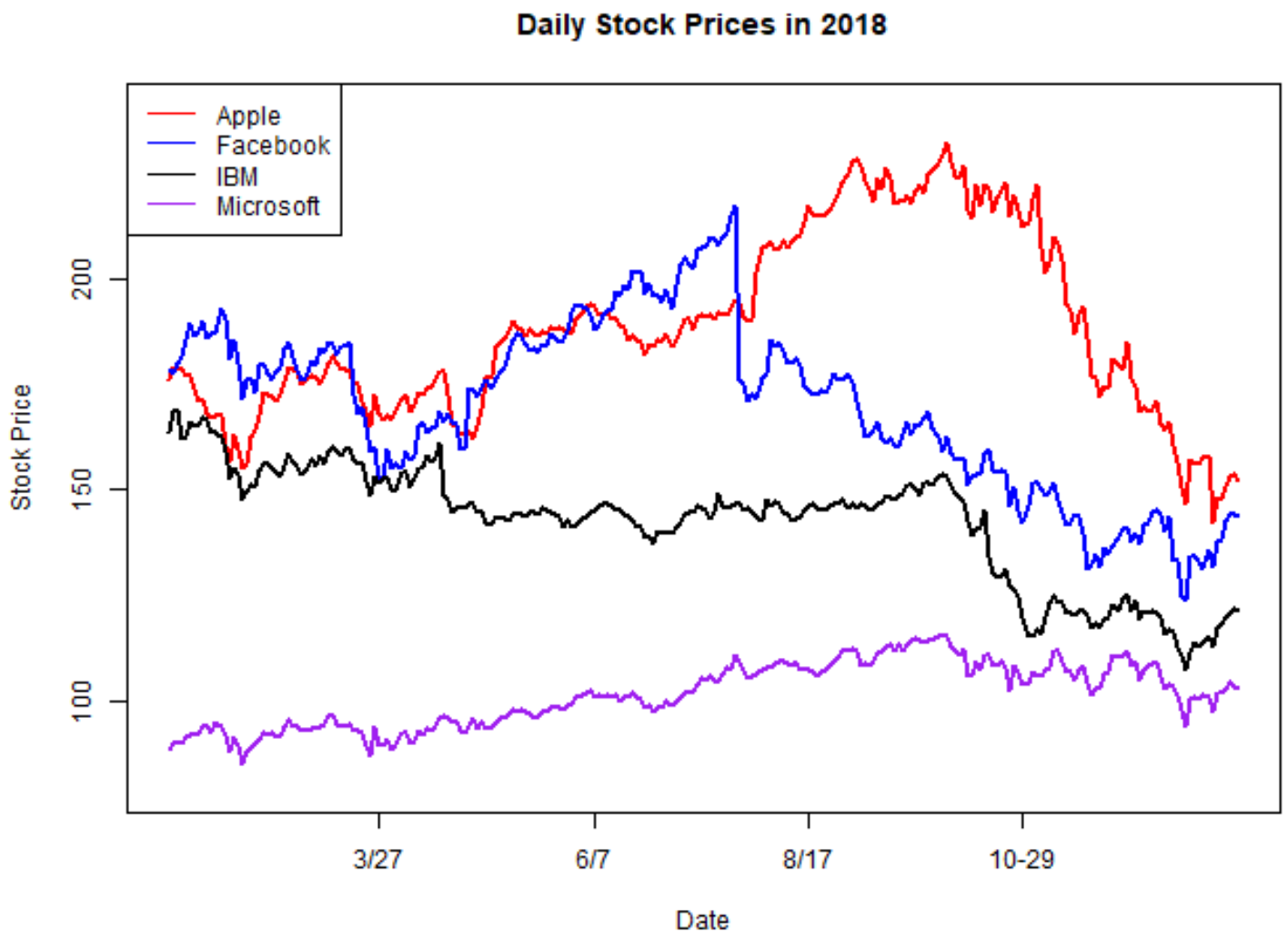
A Peek at the Data

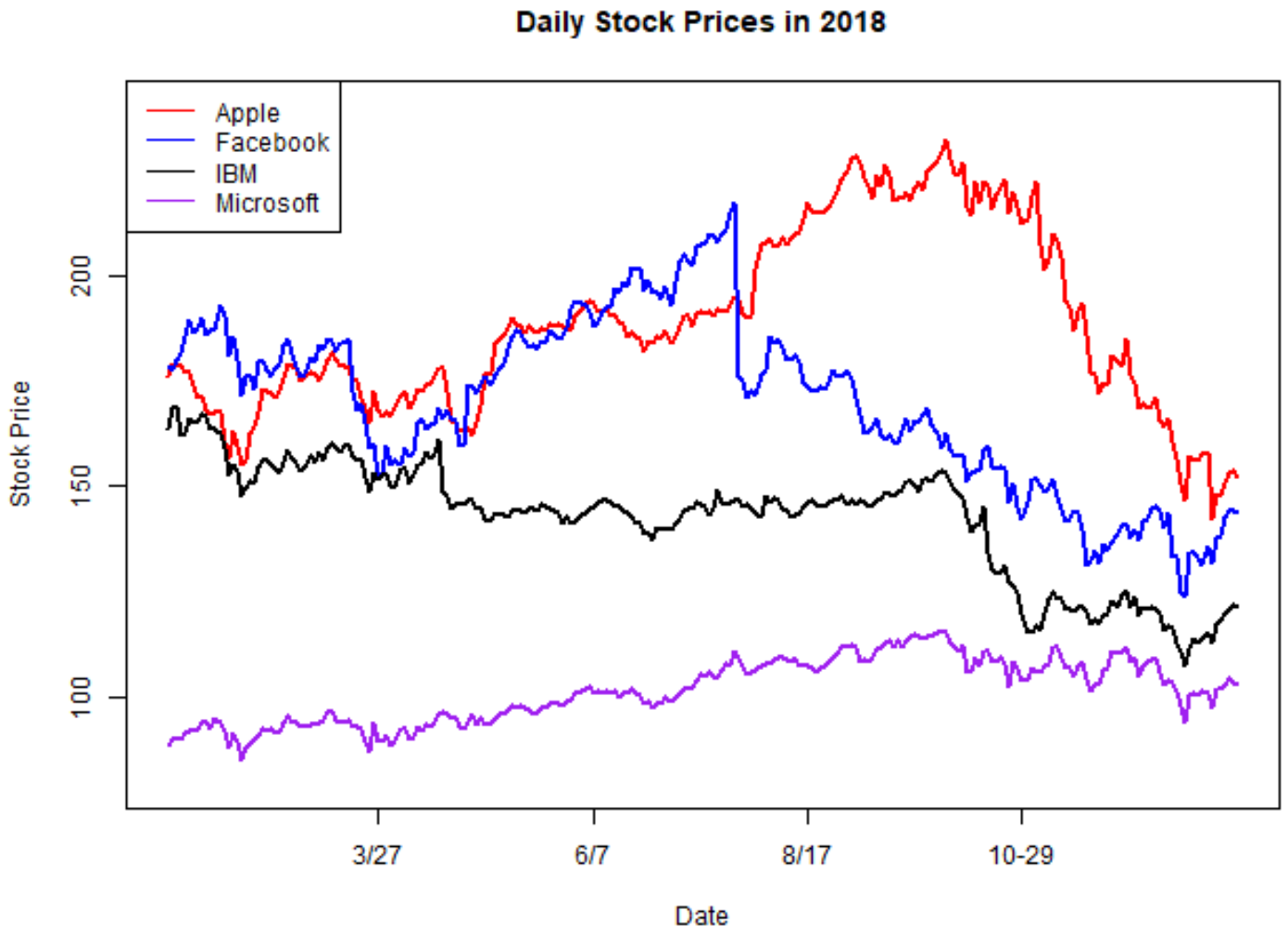
- Dataset contains 4 variables and 250 observations (250 trading days).
- Table of first five observations

Date	Apple	Facebook	IBM	Microsoft
1/16/2018	176.19	178.39	163.85	88.35
1/17/2018	179.10	177.60	168.65	90.14
1/18/2018	179.26	179.80	169.12	90.10
1/19/2018	178.46	181.29	162.37	90
1/22/2018	177	185.37	162.60	91.61

Time Plot of Stock Prices

What can you say about this plot?

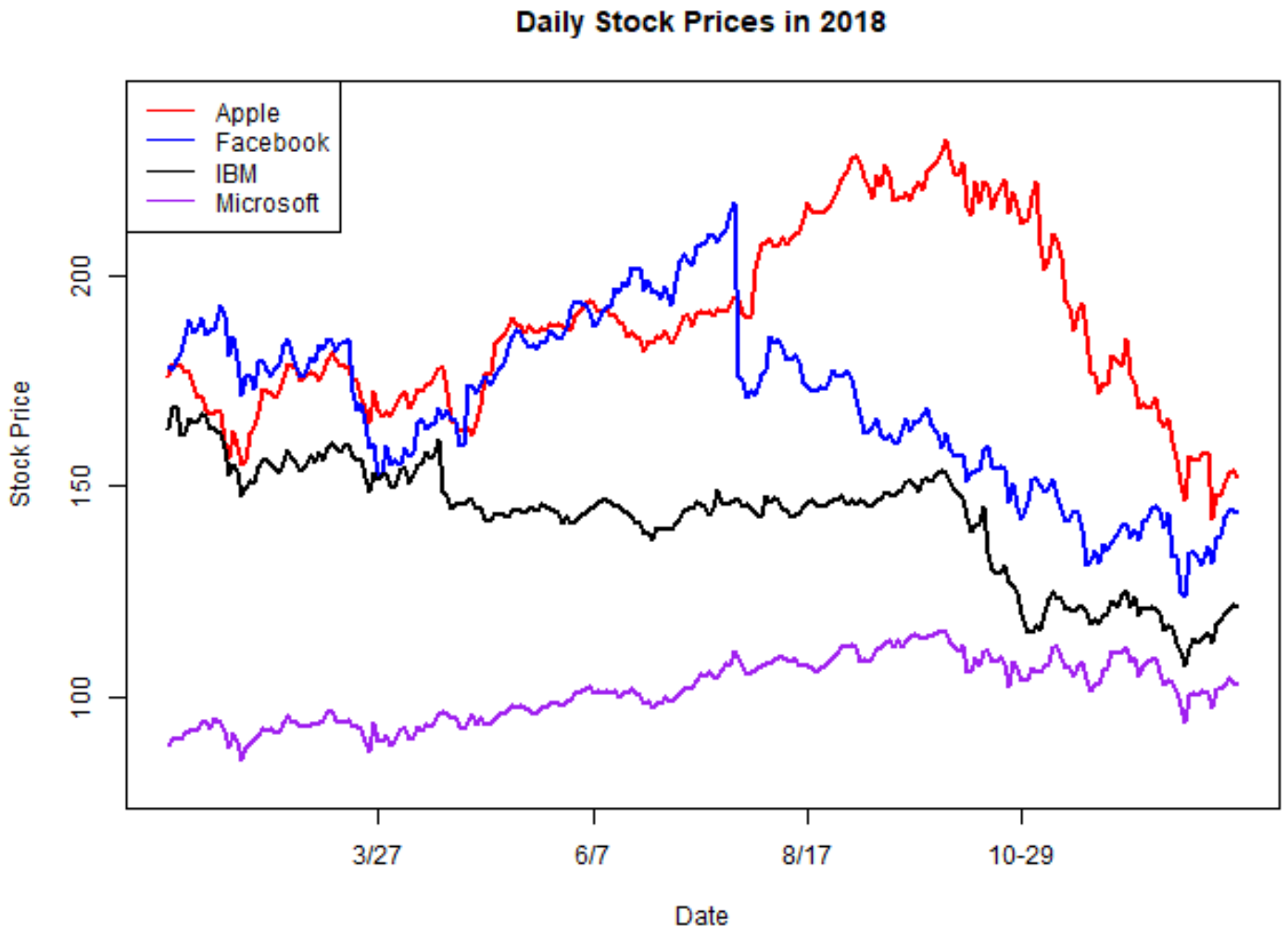




Some Questions?

- Can you observe any co-movement of these four stocks?
- Any trends?
- Which asset is most risky?

Pros and Cons of Price Data



Pros

- Straightforward
- Easy to check trends and high/low prices

Cons

- Compare apple to pear
- No relative gain/loss
- Non-stationary data

Log-return of Financial Assets

- In financial analysis, logarithm of returns is more popular than prices or raw returns.

- For an asset (e.g. stock, bond, gold, bitcoin …), log-return at time t is

$$r_t = \log(1 + R_t) = \log\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) = \log(P_t) - \log(P_{t-1}),$$

where P_t and R_t are the price and simple return at time t .

Why Log-return ?

- Log-return is favored for multiple reasons:
  - More normal-like (recall log transformation)
  - More stationary time series (zero mean and fixed variance)
  - Log additivity

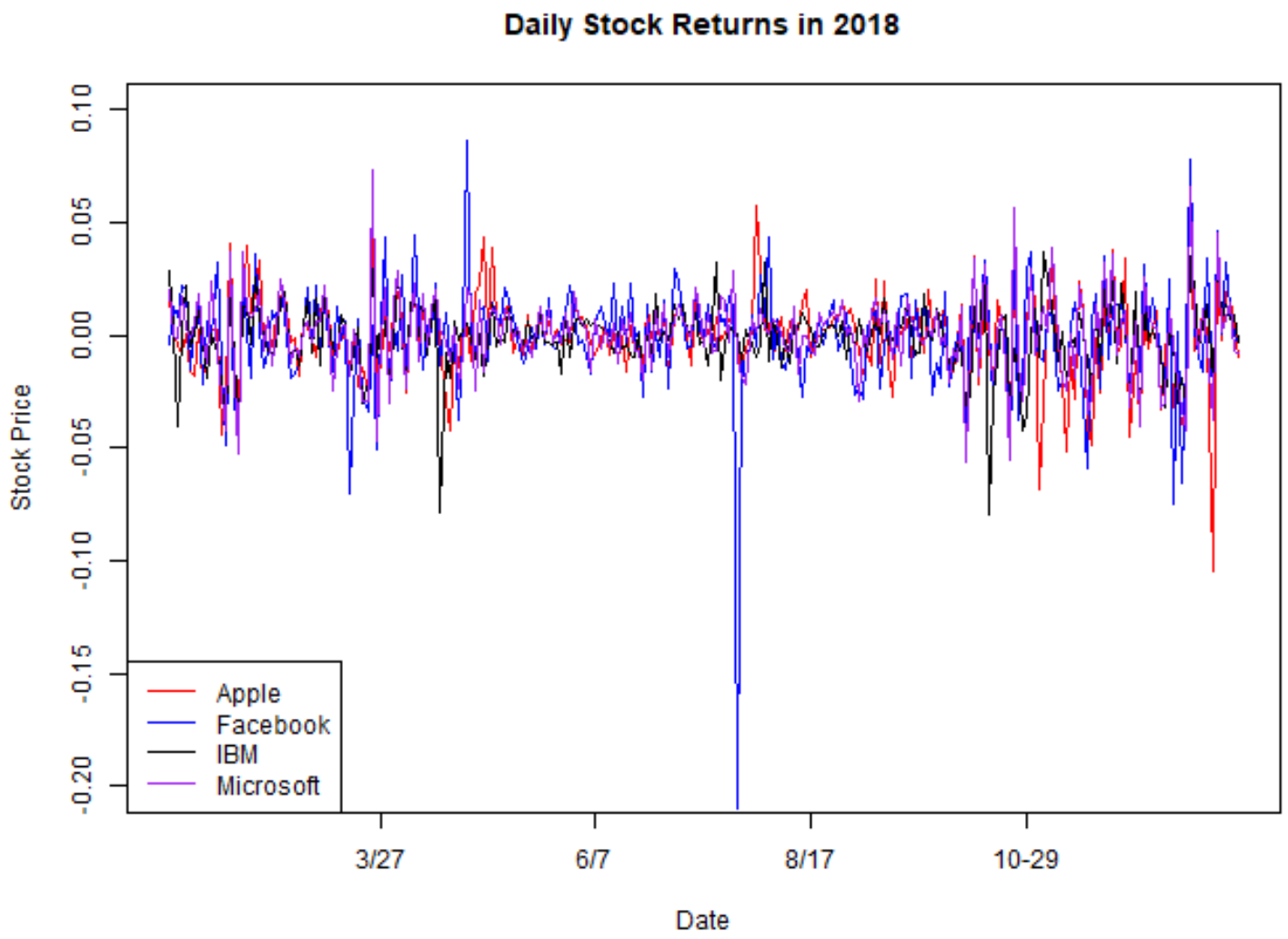
$$\sum_{1}^{T} r_t = \log(P_t) - \log(P_0)$$

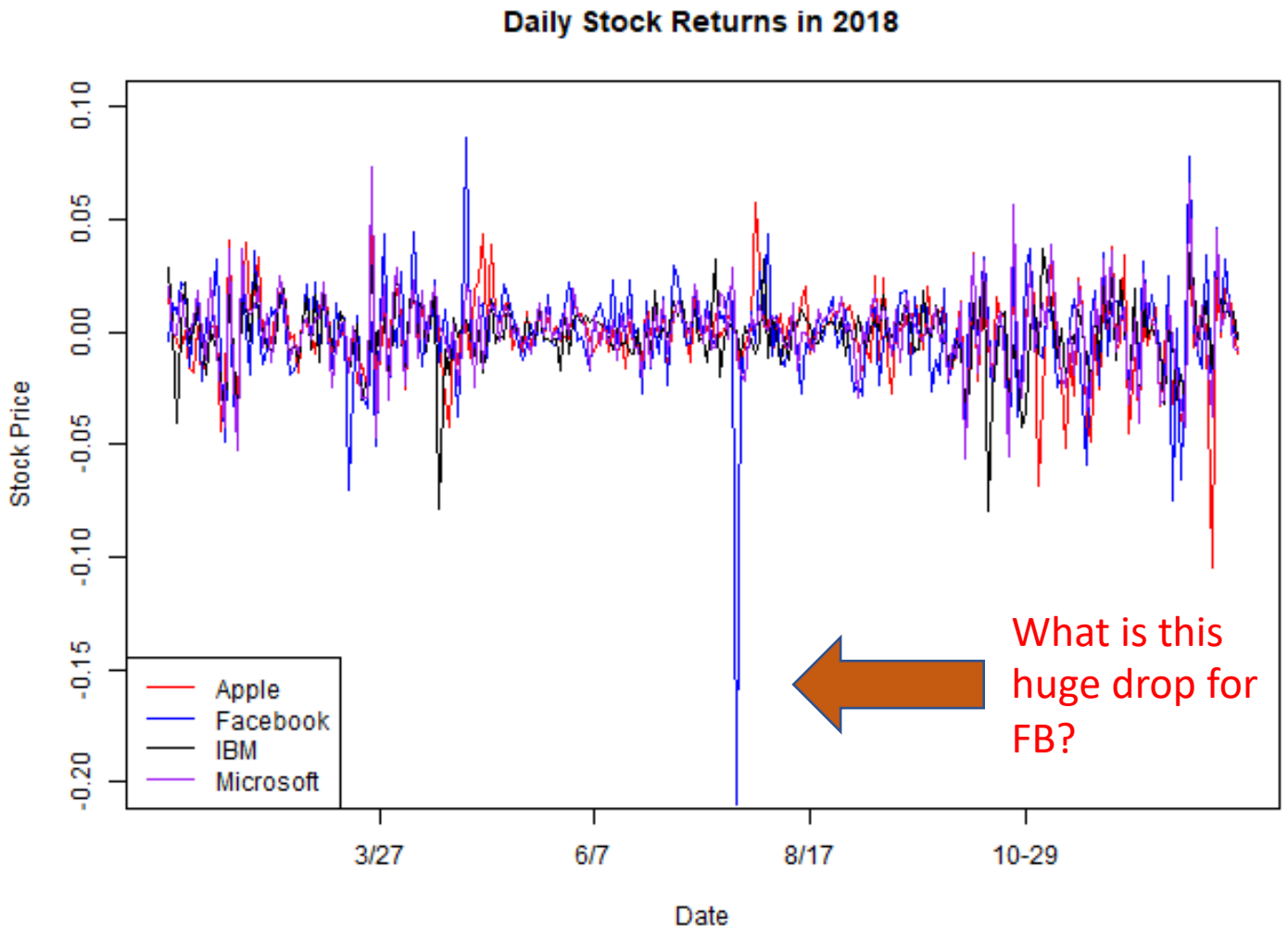
  - Easy calculus

$$e^x = \int e^x dx = \frac{d}{dx} e^x$$

Time Plot of Stock Returns

What can you say about this plot?





Some Questions?

- Which asset is the most risky (largest variance)?
- Can you observe any black swan event?

What happened?

Tech stocks fall with Facebook seeing its biggest drop ever

- The tech-heavy Nasdaq closed lower Thursday after disappointing quarterly results from Facebook.
- The social media giant saw its biggest share price decline ever and roughly $119 billion in market value.

John Melloy | @johnmelloy

Published 6:16 PM ET Wed, 25 July 2018 | Updated 4:44 PM ET Thu, 26 July 2018

CNBC

Facebook stock drops roughly 20%, loses $120 billion in value after warning that revenue growth will take a hit

Published: July 26, 2018 6:59 p.m. ET

195

Aa

Facebook earnings include 'nightmare guidance'

Radar Chart

- Radar chart (also known as spider, web, polar, star chart) is a graphical method of comparing multivariate data in the form of a two-dimensional chart of three or more quantitative variables.

- A radar chart is useful to:
  1. Find similar observations
  2. Find observation with high/low scores
  3. Find observations, clusters
  4. Find outliers

Example 3: High School Final Scores

- We generate a toy dataset. The dataset contains the final scores of some students in a hypothetical high school.

- Suppose the following 8 subjects are tested:

  1. Math, 2. English, 3. Biology, 4. Music,

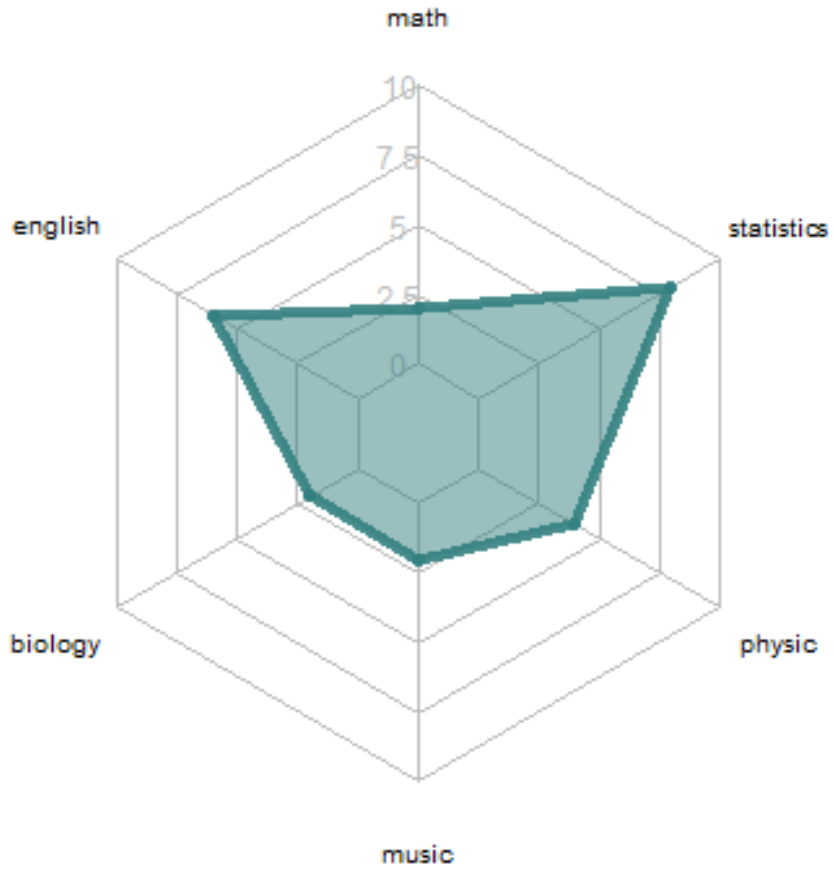
  5. Programming, 6. French, 7. Physic and 8. Statistics

- Each subject is scored from 1 to 10.

A Peek at the Data

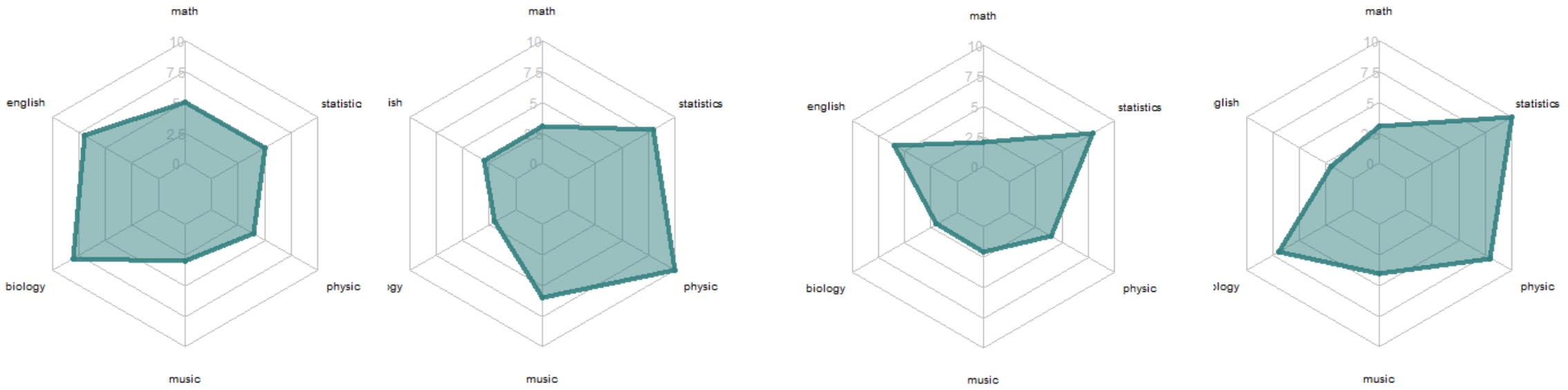
- We want to compare the performance of students
- Table of first four observations

ID	Math	English	Biology	Music	Prog.	French	Physics	Stat.
1	4	6	9	3	9	2	6	8
2	5	5	3	5	3	3	8	5
3	3	2	3	5	4	8	6	8
4	9	2	6	9	4	2	7	6

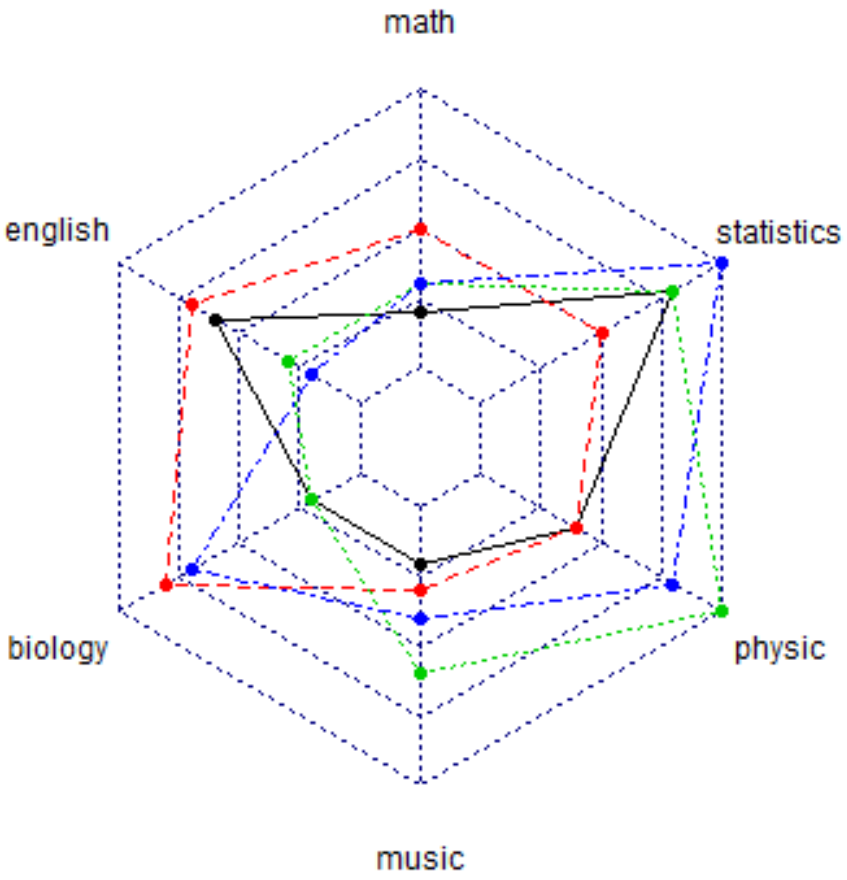


Radar Chart for 1st Student

- Each axis represents the score in a subject
- Require standardized variables
- The area covered can be considered as a score for overall performance



Comparison on parallel charts



Comparison on One Chart

- Draw multiple observations on the same chart
- Easy to compare areas, pros and cons
- Not ideal given a large number of observations

Chernoff Face

- Chernoff faces, invented by Herman Chernoff in 1973, display multivariate data in the shape of a human face.

- The individual parts, such as eyes, ears, mouth and nose represent values of the variables by their shape, size, placement and orientation.

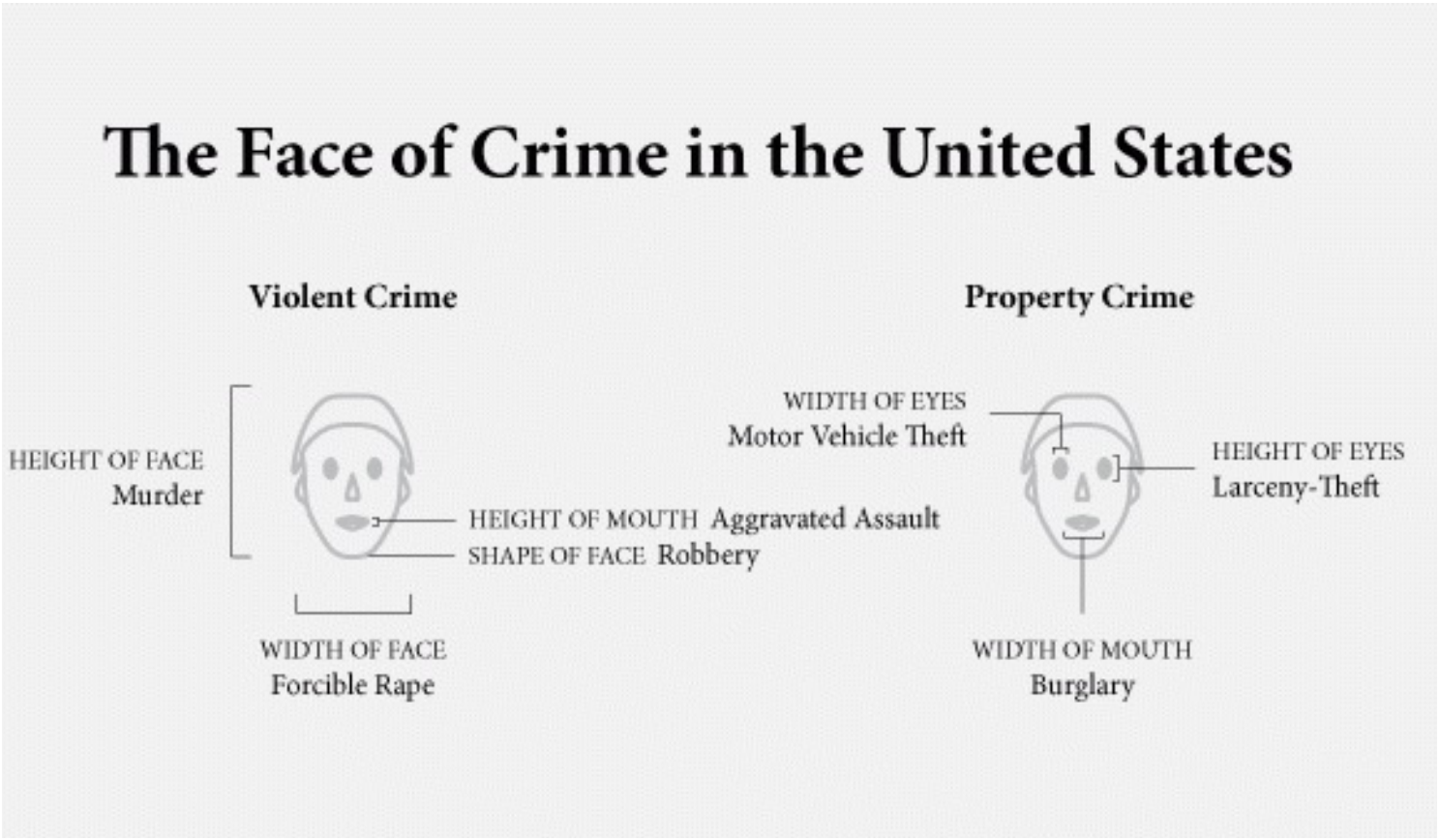
- The idea behind using faces is that people easily recognize faces and notice small changes without difficulty.

Example 4: Crime Rates by State in 2008

- The data contains the rates of various types of crimes in different states. The data source is Table 301 of the 2008 US Statistical Abstract.

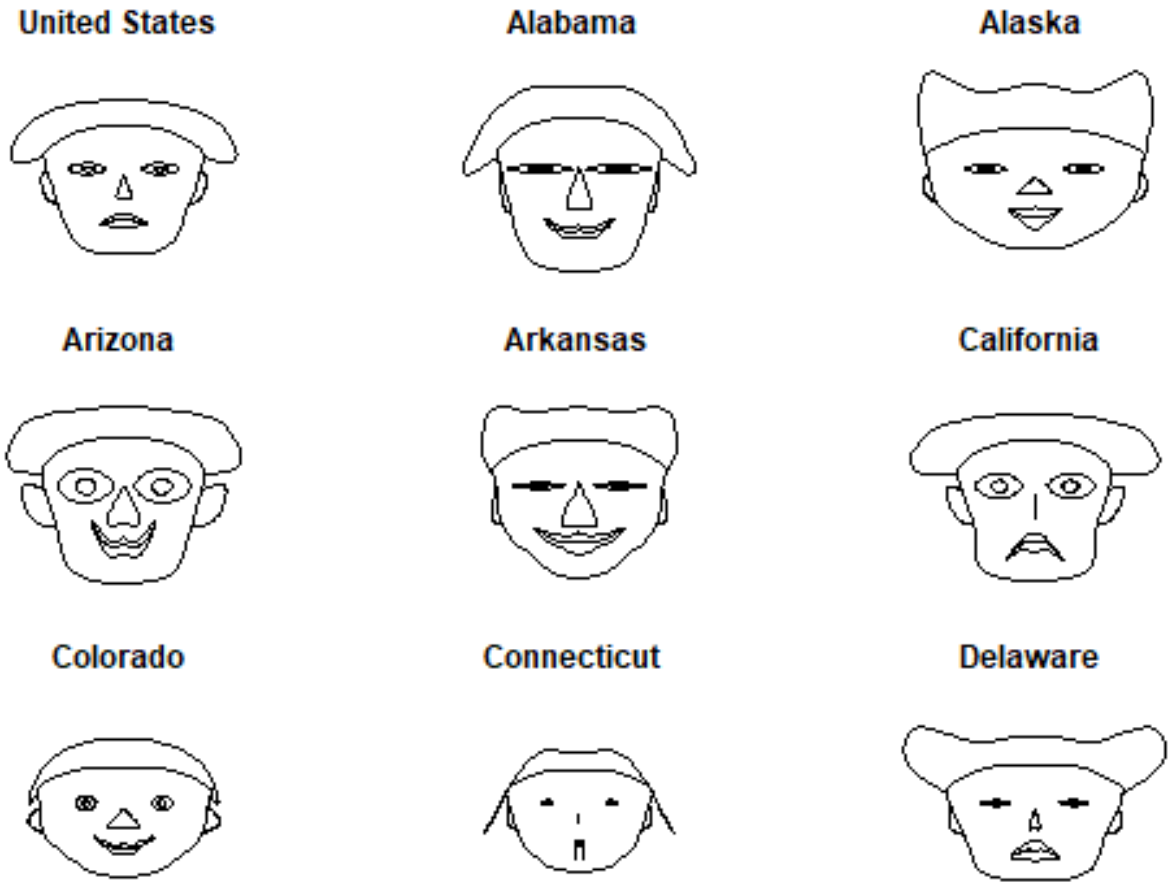
- Rates of the following crime types are recorded:

1. Murder
2. Forcible rape
3. Robbery
4. Aggravated assault
5. Burglary
6. Larceny theft
7. Motor vehicle theft



How to make a face?

- Each type of crime corresponds to a character in face.
- Shape of the character depends on the value of variable



Let's make some faces

Pros

- Easy to tell and remember the differences between states

Cons

- Hard to translate faces back to the value of variables

More Faces

United States

Alabama

Alaska

Arizona

Arkansas

California

Colorado

Connecticut

Delaware

District of Columbia

Florida

Georgia

Hawaii

Idaho

Illinois

Indiana

Iowa

Kansas

Kentucky

Louisiana

Maine

Maryland

Massachusetts

Michigan

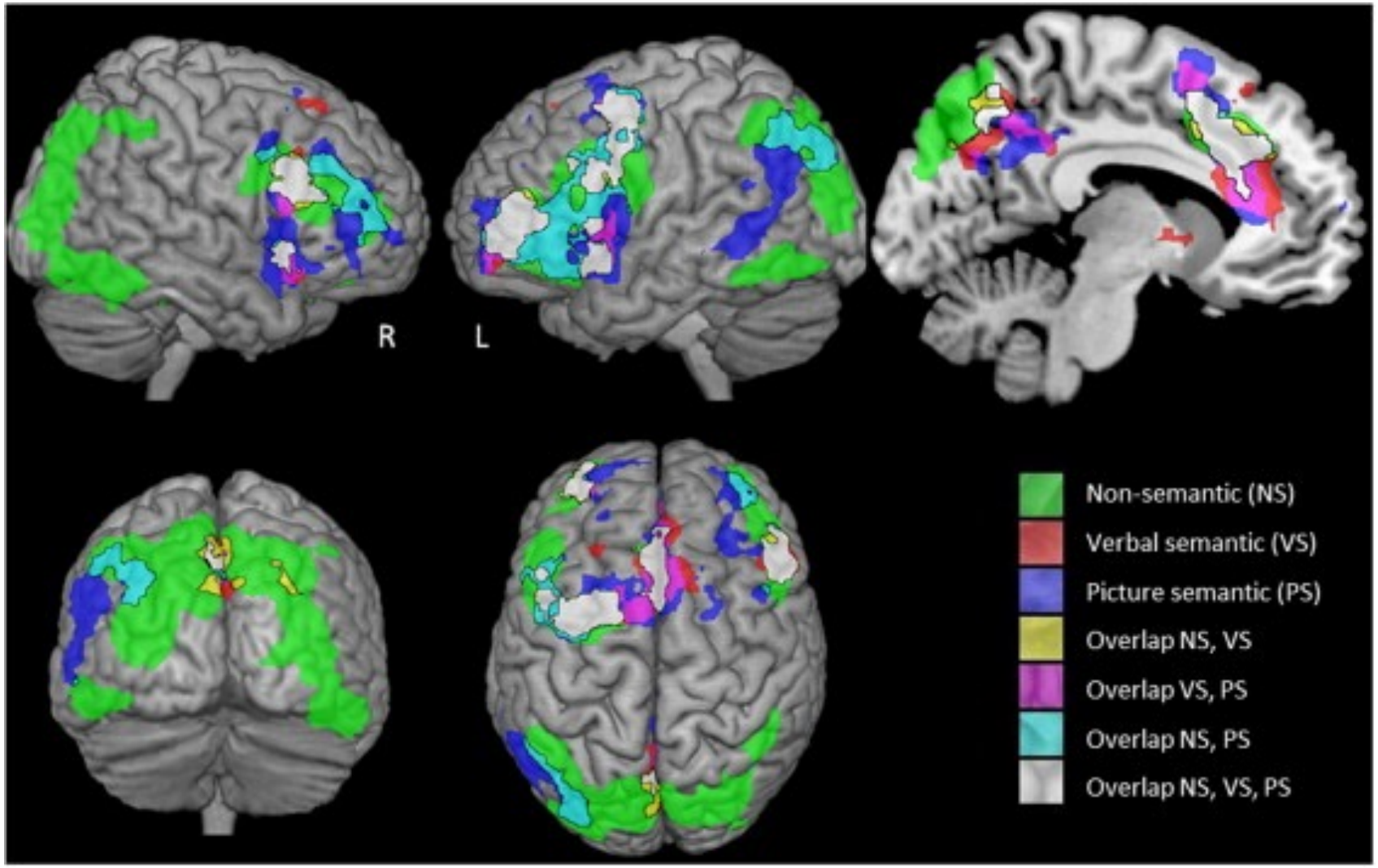
Minnesota

Heat Map

- A heat map (or heatmap) is a visualization tool which represent values in a data matrix by colors in a 2D graph

- Applications of heat map:
  1. Molecular biology
  2. Neural science
  3. Physics
  4. Density plot

Heatmap for Whole Brain Analysis



Network Map

- Network Map is a visualization tool to study of the physical connectivity of networks.
- Each node in the map represents a variable (e.g. users, characters, features).
- Two nodes are connected if there is an edge between them.

- An example: Network map for marvel cinematic universe

Network Map

- Network Map is a visualization tool to study of the physical connectivity of networks.
- Each node in the map represents a variable (e.g. users, characters, features).
- Two nodes are connected if there is an edge between them.

- An example: Network map for marvel cinematic universe