# STATS 415: Exploring Data, Part 1

Prof. Liza Levina

Department of Statistics, University of Michigan

# Data

- Collection of data objects and their attributes
- Data object: aka observation, record, data point, sample point, case, entity, instance
- Attribute is a property or characteristic of an object: aka variable, characteristic, feature, predictor
- A collection of attributes describes a data object

# Example: tax returns

| ID | Refund | Marital Status | Taxable Income | ⋯ |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | ⋯ |
| 2 | No | Married | 100K | ⋯ |
| 3 | No | Single | 70K | ⋯ |
| 4 | Yes | Married | 120K | ⋯ |
| 5 | No | Divorced | 95K | ⋯ |
| ⋮ | | | | |

# Possible operations with variables

- Distinctness: $=, \neq$
- Order: $<, >$
- Addition: $+, -$
- Multiplication: $*, /$

Different types of variable possess different properties

# Types of Variables

- Categorical, or nominal
  - Property: distinctness
  - Examples: ID numbers, eye color, zip code
- Ordinal
  - Property: distinctness, order
  - Examples: pain on a scale 1-10, t-shirt size (XS, S, M, L, XL)
- Interval: has values in equal intervals
  - Property: distinctness, order, addition
  - Examples: calendar dates

can say three days later but doesn't make sense to say three times later

- Ratio
  - Property: distinctness, order, addition, multiplication
  - Examples: distance, speed, weight

# Discrete and Continuous Variables

- Discrete variable
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, particle counts, set of words in a collection of documents
  - Often represented as integer variables
  - Binary variables are a special case of discrete variables
- Continuous variable
  - Has real numbers as variable values
  - Examples: temperature, height, weight
  - Continuous variables are typically represented as floating-point variables

# Types of Data Sets

Data matrix: $n$ observations (rows), $p$ variables (columns)

- Record data
- Document data
- Graph data
- Spatial data
- Temporal or sequential data

# Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of variables
- Data matrix conversion: simply arrange records in rows

| ID | Refund | Marital Status | Taxable Income |
|---|---|---|---|
| 1 | Yes | Single | 125K |
| 2 | No | Married | 100K |
| 3 | No | Single | 70K |
| 4 | Yes | Married | 120K |
| 5 | No | Divorced | 95K |

# Document/Text Data

- Term: words, usually stripped of endings and common "stop words" (e.g. go, going, goes → go, no "the", "and", etc)
- Variables: terms
- Observations: documents
- Value of the variable: the number of times the corresponding term occurs in the document (often normalized)

| | team | coach | play | ball | game |
|---|---|---|---|---|---|
| Document1 | 3 | 0 | 5 | 0 | 2 |
| Document2 | 0 | 7 | 0 | 2 | 1 |
| Document3 | 0 | 1 | 0 | 0 | 1 |

# Transaction Data

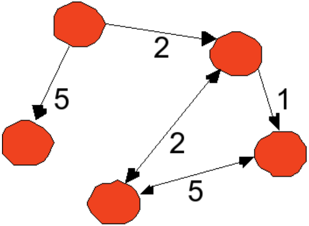
- A special type of record data
- Each record (transaction) involves a set of items
- Example: a grocery store purchase constitutes a transaction, and the individual products purchased are the items.

| ID | Items | Bread | Coke | Milk | Beer | Diaper |
|---|---|---|---|---|---|---|
| 1 | Bread, Coke, Milk | 1 | 1 | 1 | 0 | 0 |
| 2 | Beer, Bread | 1 | 0 | 0 | 1 | 0 |
| 3 | Beer, Coke, Diapers, Milk | 0 | 1 | 1 | 1 | 1 |
| 4 | Beer, Bread, Diapers, Milk | 1 | 0 | 1 | 1 | 1 |
| 5 | Coke, Diapers, Milk | 0 | 1 | 1 | 0 | 1 |

How would you represent this as a data matrix?

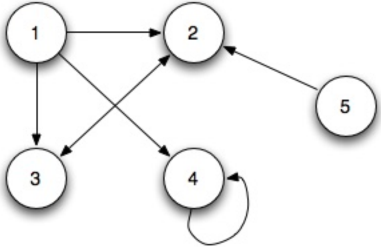
# Graph Data

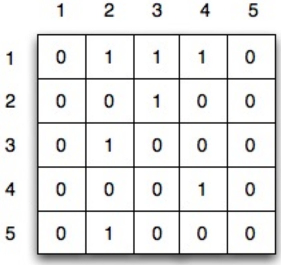
- Consists of objects (nodes) and connections between them (edges)
  - Internet (the Web, social networks)
  - Computer / mobile / electric grid networks
  - Transportation
  - Ecosystems (predator / prey networks)



- Edges can be directed, undirected, have weights and/or signs

# Graph data: adjacency matrix





|   | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 |

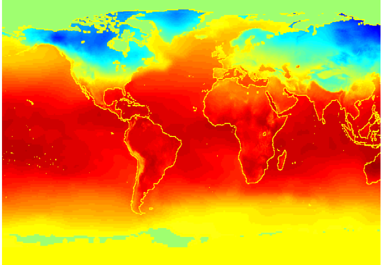
If $n$ is the number of nodes, the adjacency matrix is $n \times n$

# Ordered Data

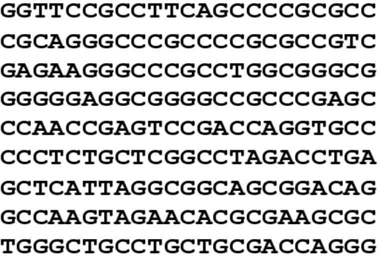
- Spatial data (e.g. temperature at various weather stations in the US)
- Temporal data (time series): (e.g. stock prices over a year)
- Functional data (e.g. spectral measurements at different wavelengths)
- Sequential data (e.g. human genome)

# Ordered data examples

Spatial



Sequential



```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGCTGCCTGCTGCGACCAGGG
```

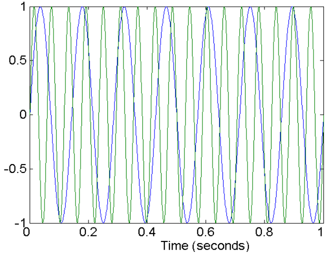
# Data Quality

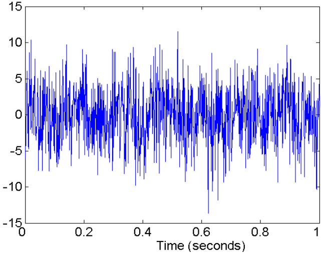
Possible problems:

- noise and outliers
- missing values
- sampling not representative of the target population

# Noise

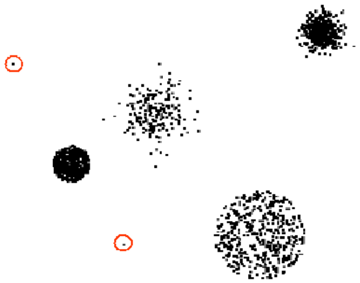
- Perturbation of original data values due to recording errors, interference, etc
- Example: distortion of a person’s voice on a bad phone line





# Outliers

- Outliers are data points “considerably different” from most of the other data points



# Outliers

- Different statistical methods have different levels of sensitivity to outliers – robust methods are less sensitive
- Sometimes outliers are obvious errors and can be removed
- Sometimes removing outliers might make you miss out on a Nobel prize...
- “Big data” tends to have more outliers than “small data”; a difficult problem.

# Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their weight; equipment malfunction)
  - Variables may not be applicable to all cases (e.g., annual income is not applicable to children)
  - Records combined from multiple studies which measured different sets of variables
  - The Big Question of missing values: is the missing status correlated with the value of the variable?
- Handling missing values
  - Eliminate data points with missing values
  - Eliminate variables with a lot of missing values entirely
  - Estimate (impute) missing values

# Sampling distortion

A mismatch between the sample and the population of interest.

- Selection bias
- Response bias
- Convenience samples
- Hard to reach populations

# Exploratory Data Analysis (EDA)

Preliminary exploration of the data to better understand its characteristics

- Helping to select the right tool for preprocessing or analysis
- Making use of humans' abilities to recognize patterns
- The term Exploratory Data Analysis (EDA) was coined by statistician John Tukey in a seminal book
- EDA is not data snooping, when done properly... but it can be hard to keep them separate.

# Techniques Used in Data Exploration

- In EDA, as originally defined by Tukey, the focus was on visualization
- In our discussion of data exploration, we focus on
    - Summary statistics
    - Visualization

# Summary Statistics

- Summary statistics are numbers that summarize properties of the data.
  - Location or center: e.g., mean, median
  - Frequency: e.g. mode
  - Spread: e.g. standard deviation
- Most summary statistics can be calculated in a single pass through the data.

# Frequency and Mode

- The frequency of a variable value is the percentage of time the value occurs in the data set.
  - Example: for the variable “gender” in the population of the US, the value “female” occurs about 50% of the time.
- The mode of a variable is the most frequent variable value.
- The notions of frequency and mode are typically used with categorical data.

# Location: Mean and Median

- The mean is the most common measure of the location of a set of points

$$\text{mean}(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- However, the mean is sensitive to outliers.
- The more robust median or a trimmed mean are also commonly used

$$\text{median}(x) = m(x) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

- Trimmed mean: drop the largest and the smallest $x$% of data points (e.g. 5% or 10%), average the rest

# Percentiles

- For ordinal or continuous data
- Given an ordinal or continuous variable $x$ and a number $q$ between 0 and 100, the $q$th percentile is the smallest value $x_q$ such that $q$% of the observed values of $x$ are $\leq x_q$.
- The median is the 50th percentile

# Measures of spread: Range and Variance

- Range is the difference between the max and min

$$\text{range}(x) = \max(x) - \min(x)$$

- The variance and standard deviation are the most common measures of spread of a set of points

$$\text{Var}(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{SD}(x) = s_x = \sqrt{\text{Var}(x)}$$

- However, variance depends on the mean and is also sensitive to outliers

More robust measures of spread:

- Mean Absolute Deviation

$$\mathrm{AAD}(x) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

- Median Absolute Deviation

$$\mathrm{MAD}(x) = \text{median}(|x_1 - m(x)|, \ldots, |x_n - m(x)|)$$

- Inter-Quartile Range

$$\mathrm{IQR}(x) = x_{75\%} - x_{25\%}$$

# Multivariate Summary Statistics

- Location: compute the mean or median separately for each variable

$$\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)$$

- Spread: covariance matrix

i=1,…n

j,j'=1,…,p

Cov_jj'

$$\mathrm{Cov}(x_j, x_{j'}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Cov_jj'=Var(Xj)

- Correlation matrix

$$\mathrm{Cor}(x_j, x_{j'}) = \frac{\mathrm{Cov}(x_j, x_{j'})}{\mathrm{SD}(x_j)\mathrm{SD}(x_{j'})}$$

-1<=Cor<=1

X~N(0,1) y=x^2 Cor(x,y)=0

# Other Measures

- Skewness: measures the degree to which the values are symmetrically distributed around the mean

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

- Some characteristics are not easy to measure quantitatively: e.g. whether a distribution is unimodal or multi-modal