Data Cleaning

MIE223

Winter 2025

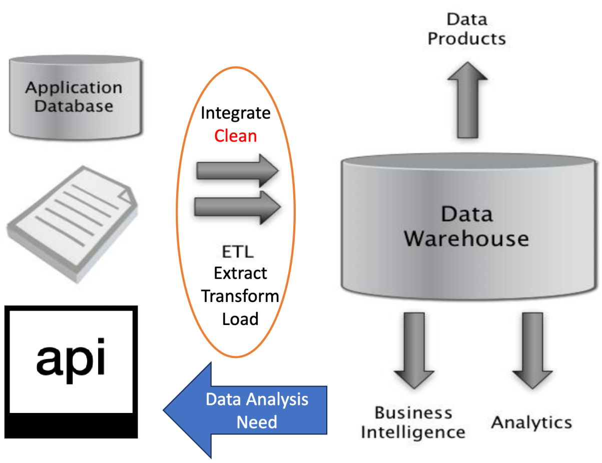
1 Why Clean Data?

1.1 Example

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210
Intl. Business Machines	Armonk, NY	$200
Microsoft	Redmond, WA	$250
Sally's Lemonade Stand	Alameda,CA	$260,000

- Apple: Missing Data
- IBM: Entity Resolution / Unnormalized Naming
- Sally's Lemonade Stand: Unit Mismatch

1.2 Where is Data Cleaning in the Organization?



- Data comes in all shapres and forms
- ETL: Extract, Transform, Load
- Data science exists in business intelligence and analytics
- Data engineers work in application database

1.3 Cost of Bad Quality Data Over Time

- Cost for preventing bad data: $1
- Cost for correcting bad data: $10
- Cost for fixing problem resulting from bad data: $100

2 Perspectives on Data Science and Data Cleaning

2.1 Data Science is an Intersectional Discipline

- Computer Science
- Math and Statistics
- Domain Knowledge

2.2 Dirty Data

- The Statistics View:
  - There is a process that produces data
  - Any dataset is a sample of the output of that process
  - Results are probabilistic
  - You can correct bias in your sample
- The Database View:
  - I got my hands on this data set
  - Some of the values are missing, corrupted, wrong, duplicated
  - Results are absolute (relational model)
  - You get a better answer by improving the quality of the values in your dataset
- The Domain Expert's View
  - This Data Doesn't look right
  - This Answer Doesn't look right
  - What happened?
- The data scientist's view is a combination of the above.

Note 1. A Data scientist knows more programming than a statistician and more statistics than a programmer. A Data scientist intimately understands their domain!

2.3 Data Quality Problems

- Data is dirty on its own
  - Human collection or data entry (generally lazy and want to see their children)
- Data sets are clean but integration (i.e., combining them) screws them up (such as different units of measurement)
- Data sets are clean but suffer “bit rot” (lose data from cutting)
  - Constant data transformations introduce errors, noise, data loss
  - Meanings of labels change over time
- Any combination of the above... and much, much more!

2.4 Some Data Issues you will Encounter

1. Parsing text into fields (separator issues)
2. Naming conventions and entity resolution D: NYC vs New York
3. Missing required field (birthdate)
4. Gaps in time series
5. Different representations (“2” vs 2), Unicode lookalikes
6. Fields too long (get truncated)
7. Mixed data types (feet, meters)
8. Redundant Records (exact match or other)
9. Formatting issues – especially dates
10. Licensing issues/privacy/cost prevent access to all data

2.5 Key Types of Data Quality Issues

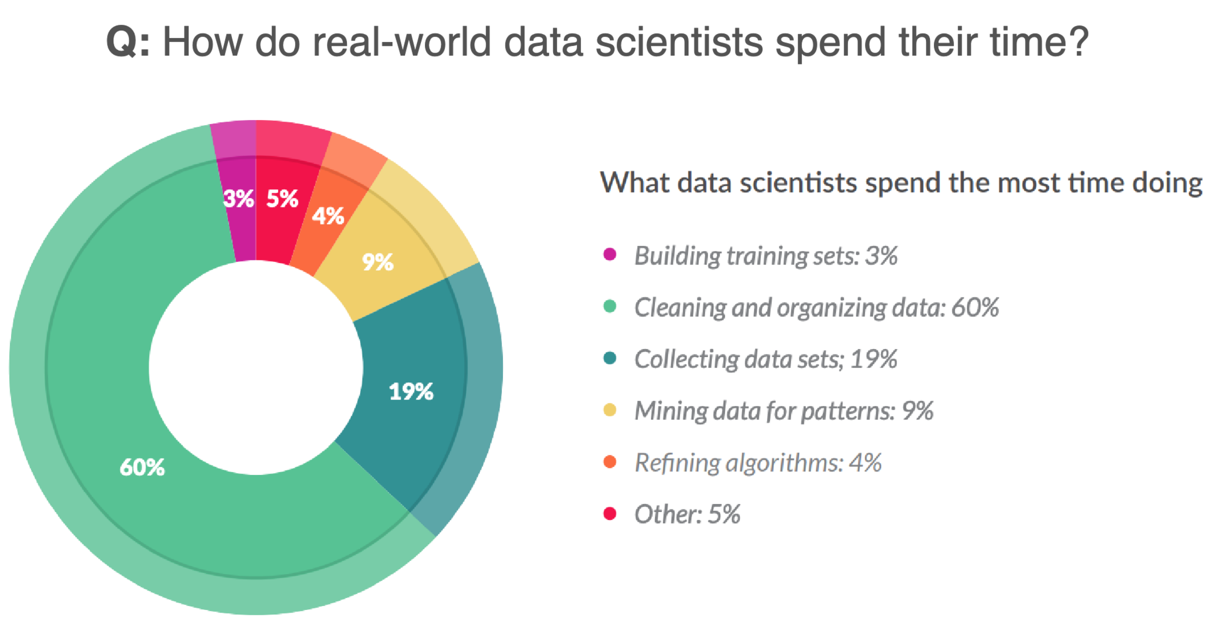
- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”
- **noisy**: containing errors or outliers
  - e.g., Salary=“-10”
- **inconsistent**: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records
- **redundancy**: duplicated content

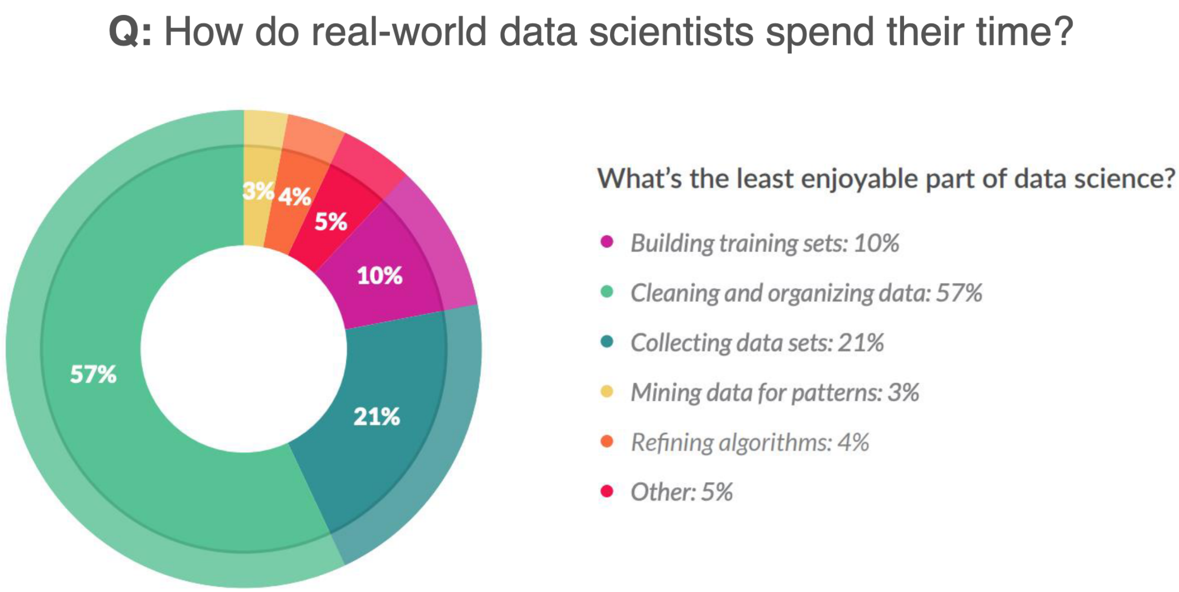
2.6 Why Is Data Dirty?

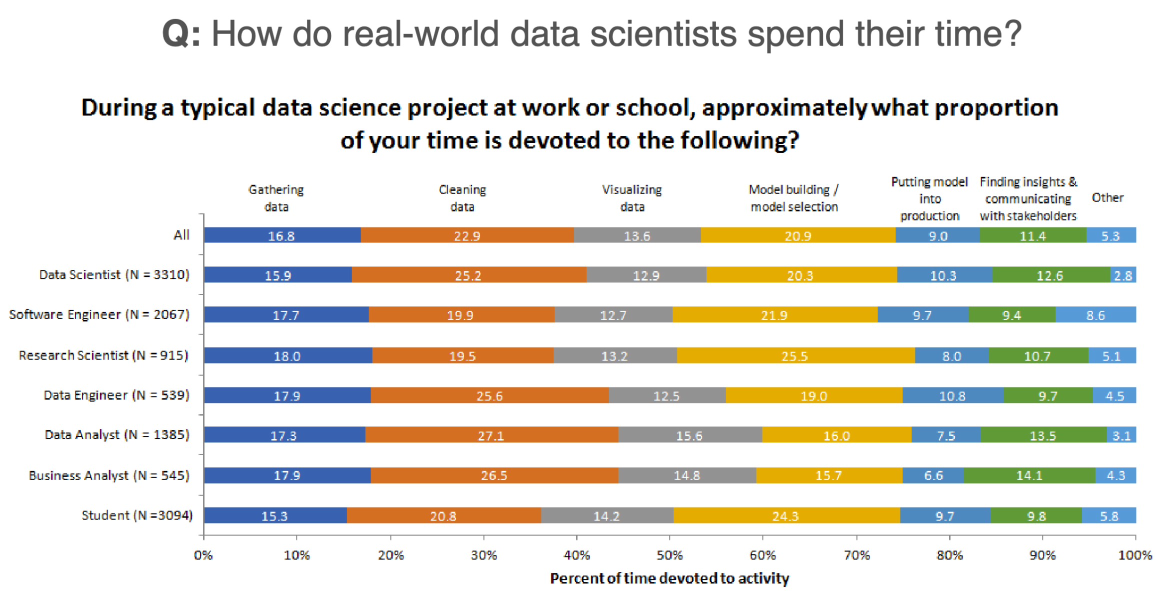
- Incomplete data may come from
  - "Not applicable" data value when collected
  - Changes in the data collected over time
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments, undocumented API changes
  - Human or computer error at data entry, UI changes!
  - Errors in data transmission, discretization, conversion (losing precision)
  - Typing errors (meters, feet, km mixed in same column)
- Inconsistent data may come from
  - Different data sources (data integration)
  - Changes in data collection practices over time
- Redundancy
  - Human error (could not find previous record), data integration

3 The Role of Data Cleaning in Data Science

3.1 Data Science in the Real World







3.2 Data Cleaning Makes Everything Okay?

Note 2. "The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning." - National Center for Atmospheric Research

The data was rejected as unreasonable by data quality control algorithms.

4 Principles of Data Cleaning

4.1 Data Cleaning Tasks

- Data Summary
- Consider filling in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy

5 Data Cleaning: Data Summary

5.1 Data Summary Procedure

- Look at data types of columns
  - Beware of ”Object” columns, indicator that you’ve mixed types (e.g., ”2” and 2)
- For high cardinality strings and categorical (e.g,. ”country”) columns
  - How many unique elements are there?
  - Are any missing?
  - Count frequencies of string
    * Look at top-10 most frequent and least frequent values
    * Look at strings near the mean and median frequency
- For low cardinality strings and categorical (e.g,. “province”) columns
  - Look at frequency of each unique value
- For ordinal numeric (integers) or floating point values
  - Compute summary statistics
  - View a histogram (but before you do this, hypothesize what you will see)

5.2 Summary Statistics and for Each Variable (Column)

- Range
  - Minimum
  - Maximum
- Central Tendency
  - Mean
  - Median
  - Mode

    * Value that occurs most frequently in discrete data
    * Value with highest frequency (or value range with highest frequency) in continuous data
    * Informally: # peaks in numeric data – unimodal, bimodal, trimodal, multimodal, ...

- If Data is non-numeric, consider frequencies below (e.g., mean frequency of jobs in "Job")
- If Data is a string, consider sorting it and looking at nearby values

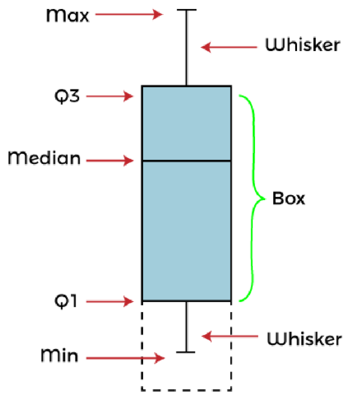
5.3 Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: IQR = Q3 – Q1
  - Five number summary: min, Q1, M, Q3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers (min / max), and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
  - $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^{\ 2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\ 2} - \mu^2$$
  - Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

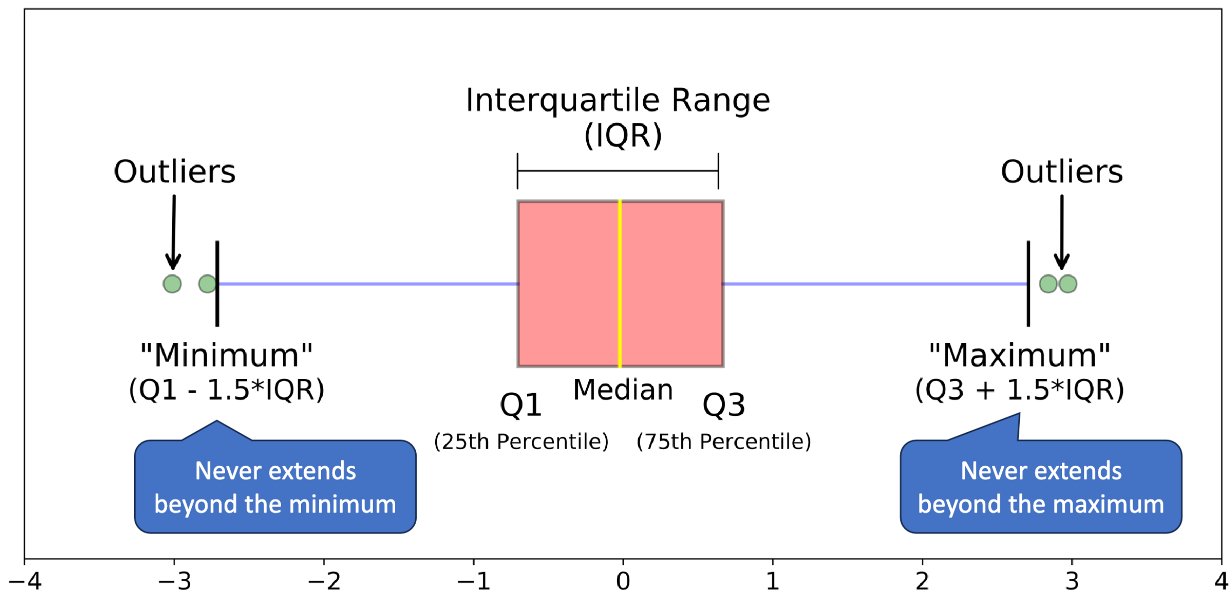
5.4 Boxplot Analysis

- Five-number summary of a distribution:
  - Minimum, Q1, M, Q3, Maximum (IQR=Q3-Q1)
- Historical Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles (Q1,Q3), i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
  - Shows asymmetry unlike mean ± std

Note 3. Beware boxplots can hide multimodality.



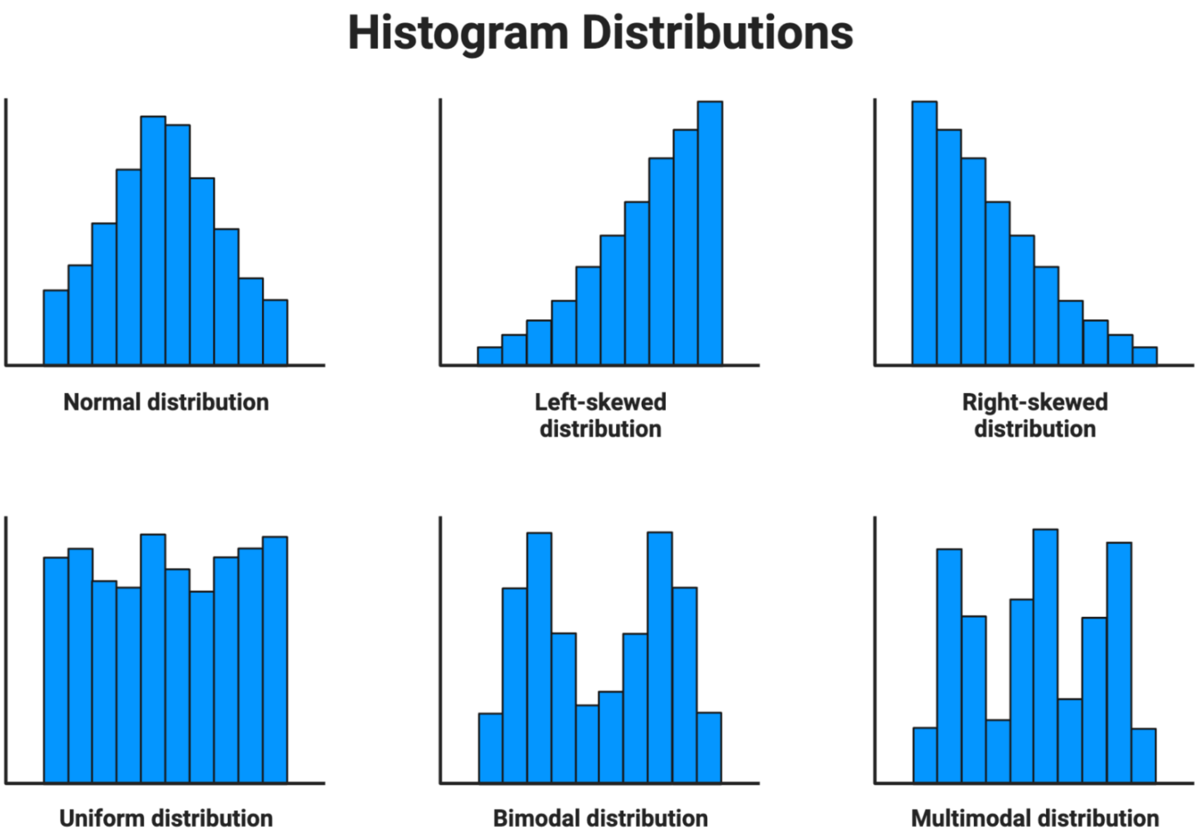
5.5 Python Boxplot (not all Boxplots are the same)



5.6 Understanding the Data Distribution: Histograms

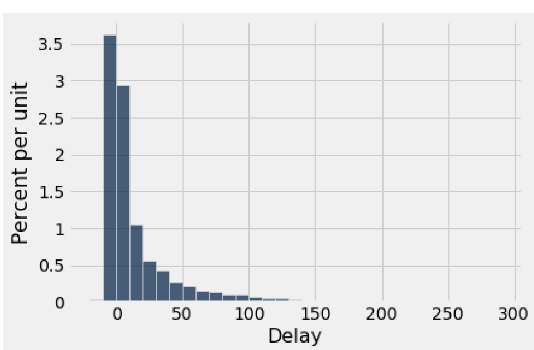
Note 4. The normal distribution is unimodal.

- Summary statistics like standard deviation can be very misleading if you don't know the distribution
- A Histogram shows an empirical density
  - I.e., the frequency of binned value ranges of your variable (column)
  - Use a bar chart of frequencies if variable is discrete

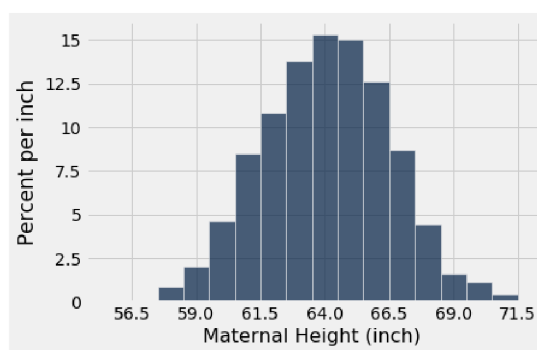


5.7 Processes Generate Different Distributions

- Power Law Distributed of Flight Delays



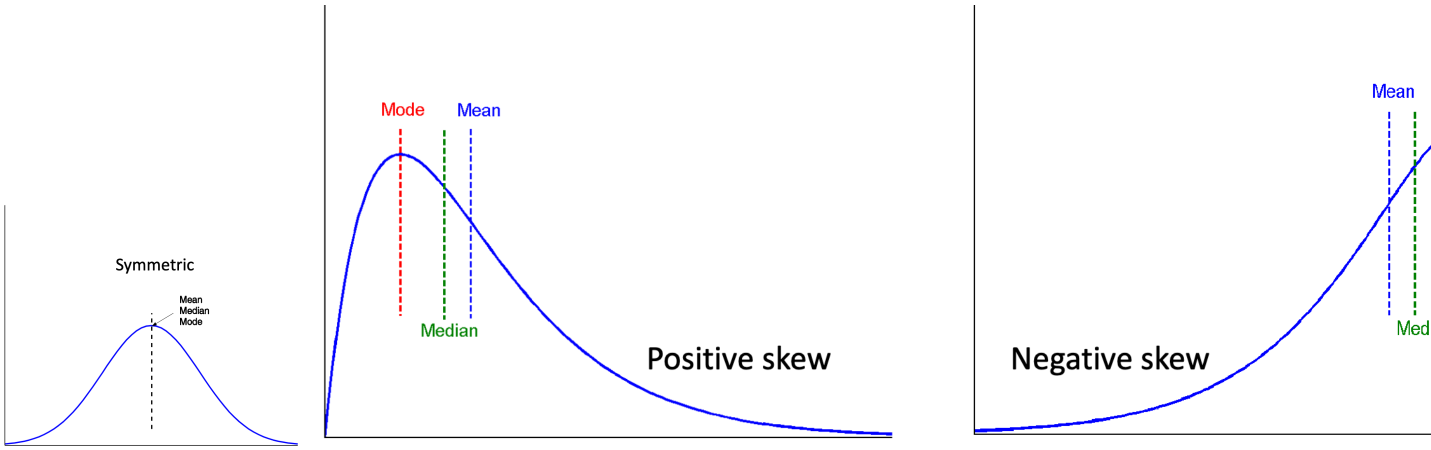
- Normally Distributed Data of Height of New Mothers



- Mean, median, and mode are same for Gaussian data, but (very) different for power law data.

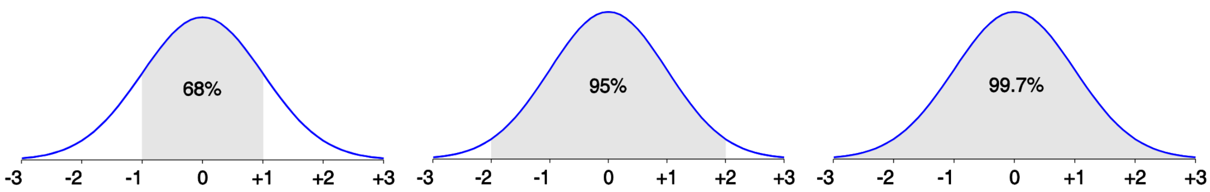
5.8 Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



5.9 Properties of Normal Distribution Curve

- The normal (distribution) curve
- From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
- From μ–2σ to μ+2σ: contains about 95% of it
- From μ–3σ to μ+3σ: contains about 99.7% of it
- Useful for interpreting z-scores (cf. transformation slides)



6 Data Cleaning: Missing Values

6.1 Missing Values

- Missing Completely at Random (MCAR): the random process by which data is missing does not depend on any other observed (i.e., column) nor latent (i.e., unobserved) variable
  - e.g., for column ”income” uniformly randomly pick a row and with probability P, delete the ”income” entry at that row
- Missing at Random (MAR): the random process by which the data is missing depends on another variable (column), but not on a latent variable
  - e.g., for column ”income” uniformly randomly pick a row and with probability P if that row has ”gender=Male” or probability Q if that row has ”gender=Female”, delete ”income” at that row
- Missing Not at Random (MNAR): the random process by which the data is missing depends on a latent variable

- e.g., for column ”income” delete the ”income” at a row *if* the day the row is added to the table is on a Monday (where the timestamp of when a row was added is not recorded in any column and is hence latent)
- e.g., for column ”income”, the ”income” at a row is missing if it was below \$10,000 – here, the value that made it missing is not recorded in the table and therefore a latent cause

6.2 Example Customer Data

Name	Age	Sex	Income	Job
Mike	20	Male	40k	Uber Driver
Jenny	40	Female	?	Neurosurgeon
...				

6.3 Imputation

- Imputation is risky (makes assumptions and “creates” missing values)
  - If very few rows are missing data, it may be better to delete those rows
  - But if high rate of missingness, cannot delete most of data!
- Some downstream analysis requires complete data in rows
  - Correlation between all pairs of variables
  - Classification and regression in most of machine / deep learning
- Impute if you must, but imputer beware
  - As they said in Ancient Rome: caveat imputor

6.4 Imputation Methods

- String or categorical data
  - Most frequent value (or mode)
- Integer data
  - Median makes more sense than mean (because it is an integer value)
  - Mode?
- Continuous (floating point) data
  - Mean
  - Median (can differ significantly from Mean for asymmetric distribution)
- Consider a specific value if MNAR
  - No response to a survey could mean “No” be default

6.5 Conditional Imputation

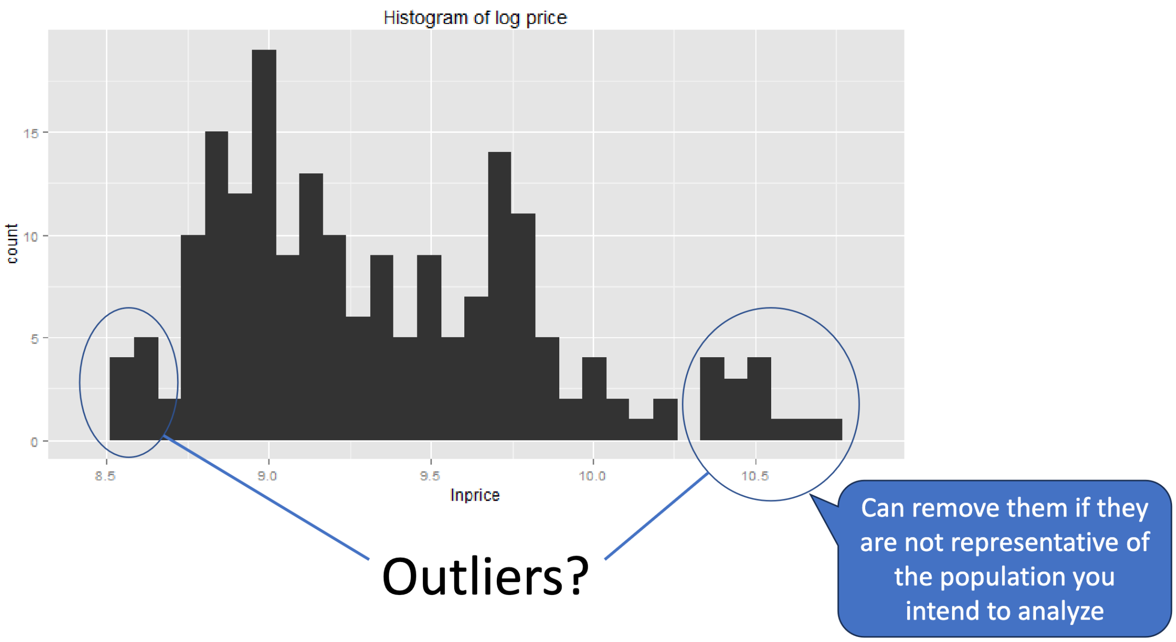
- But wait, we can also conditionally impute!
  - Can use "groupby" and aggregation on one or more other columns
    - E.g., Impute "Income" conditioned on median income grouped by "Job"
  - Could train a classifier or predictor from other columns (be careful)
- If Data is MCAR
  - Consider how other columns may impact prediction
    - E.g., "Sex=male" gives a strong hint about "Is Pregnant?"
- If Data is MAR
  - Consider if the missingness condition should influence the prediction
    - E.g., all stores in "North Yorkville" did not report client income
  - Note: columns that influence missingness do not necessarily influence the prediction
    - E.g., "Income" missing for "Location=Scarborough" but may be most influenced by "Age" and "Job"

7 Data Cleaning: Noise, Inconsistency, Redundancy

7.1 Reality Check from Summary Statistics and Domain Knowledge

- Range
  - Minimum: can height be -10m?
  - Maximum: can height be 3m?
- Central Tendency:
  - Mean: can mean height be 1.8m? (may be skewed by some outliers)
  - Median: can median height be 1.8m? (says that half of the people are ¿= 1.8m)
  - Mode: can mode height be 1.8m?
    - Can height be multimodal?
- If Data is non-numeric, consider frequencies or percentages
  - Can the most frequent word be "computer"?
  - Can 75% of the jobs in your database be neurosurgeon?
- For sorted strings, do both "Microsft" and "Microsoft" appear?

7.2 Identify Outliers and Potential Errors



7.3 Data Cleaning Task: Entity Matching

Customers1

FullName	Age	City	Sate
Aisha Williams	27	San Diego	CA

Customers2

LastName	FirstName	MI	Age	Zipcode
Williams	Aisha	R	27	92122

Q: Are these the same person ("entity")?

- Duplicates often occur in data integrated from multiple sources
- Or duplicates can simply result from redundant data entry
- Without merging / deduplication: will overcount records
  - Deduplication methods provided by record linkage or entity linkage

8 Data Transformation: Improve Interpretability

8.1 Data Transformation

- Smoothing: remove noise from data

- Aggregation: summarization
- Generalization: back off labels to a concept hierarchy
- Attribute/feature construction
  - New attributes derived from the given ones
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - subtract mean (preserves magnitude)
  - z-score normalization
  - log normalization

8.2 Data Transformation: Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score normalization (μ: mean, σ: standard deviation):

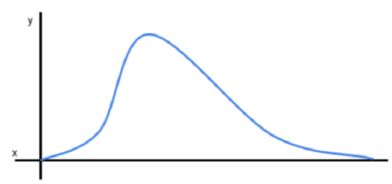
$$v' = \frac{v - \mu_A}{\sigma_A}$$

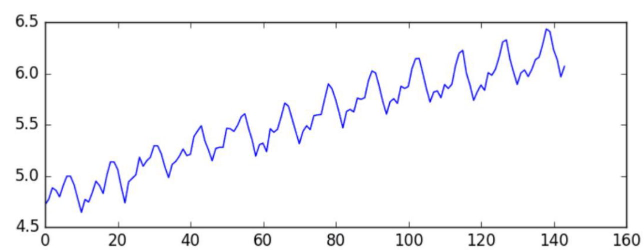
  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
  - Why useful? Because now |Z|≥ 1 is 1 std, |Z|≥ 2 is 2 std's (5% outlier)
    - Recall Gaussian distribution slide (earlier)

Claim: use min/max normalization if data has a natural known range, otherwise z-score is more stable… why?

- Log normalization: $v' = \log(v)$
  - Data must be positive, if contains 0's use $v' = \log(v + 1)$

- Why log normalization?
  - Positive skewness (long-tail)
  - Outliers
  - Multiplicative effects and growth
    - Time series exhibiting growth





- More generally log is a special case of Box-Cox power transforms
  - Includes square root