# Data Preprocessing and Transformation

- Data Collection
- Issues with Data
- Data Cleaning, dealing with missing values, noise and outliers
- Data Integration, removing inconsistencies, and deduplication
- Data Reduction - Sampling and Feature Selection
- Data Transformation - Scaling and Standardization, Numeric Transformation

Imdad ullah Khan

# Data Collection

## Data Collection

Data collection is the first step in the data anlysis pipeline

▷ Often from multiple sources

**Importance:** The quality and quantity of collected data directly influence the insights derived from big data analytics

**Challenges:** Ensuring data accuracy, dealing with large volumes, and integrating diverse data formats

## Issues in Data Collection and Techniques

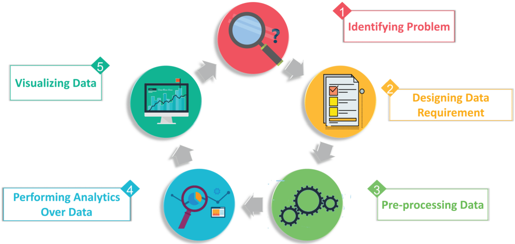
Identifying and addressing common issues in data collection is essential for ensuring the integrity of data

- Incomplete data collection
- Biases in data due to collection methods
- Collection of irrelevant or redundant data

To overcome common issues, several techniques can be employed:

- **Automation:** Use scripts and APIs to collect data systematically
- **Validation:** Implement real-time data validation to catch errors early
- **Sampling:** Employ statistical sampling techniques to manage large volumes

# Data Preprocessing



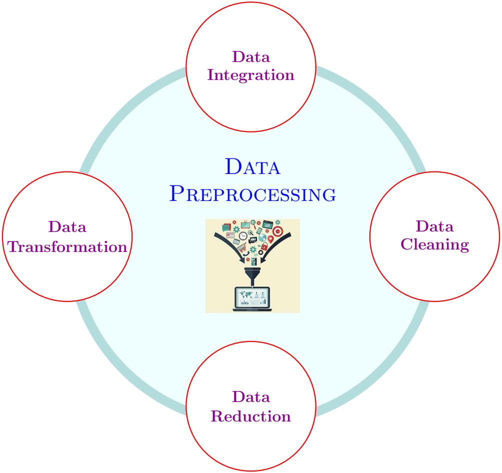
- Data preprocessing is a very important step
- It helps improve quality of data
- Makes the data ready and more suitable for analytics
- Should be followed and guided by a thorough EDA
- EDA helps identify quality issues in data that are dealt with in this step

## Issues with data

- Bad Formatting: Grade 'A' vs. 'a'
- Trailing Space: Extra spaces in commentary, white font ',' to avoid plagiarism detection
- Duplicates and Redundant Data: A ball repeated could be confused with a wide/No ball, a grade repeated confused with repetition
- Empty Rows: Could cause a lot of troubles during programming
- Synonyms, Abbreviations: rhb, right hand batsman
- Skewed Distribution and Outliers: Outliers could be points of interest or could be just noise, errors, extremities
- Missing Values: Missing grades, missing score
- Different norms, units, and standards: miles vs. kilometers
  - 1999: NASA lost equipment worth $125m because of an engineering mistake of not converting English to Metric unit

# Steps in Preprocessing

Steps and processes are performed when necessary



# Data Cleaning

## Data Cleaning

Data cleaning is a critical process that ensures the accuracy and completeness of data in analytics

It involves correcting or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

- **Objective:** Enhance data quality to produce reliable analytics
- **Common Issues:** Inconsistencies, missing values, noise, and outliers.

Also called data scrubbing, data munging, data wrangling

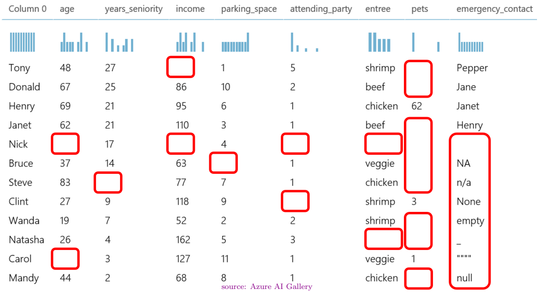
- Dealing with Missing values
- Noise Smoothing
- Correcting Inconsistencies
- Identifying Outliers

# Data Cleaning: Missing Values

Missing data is very common and generally significantly consequential

**Causes:**

- Changes in experiments
- human/data entry error
- measurement impossible
- hardware failure
- human bias
- combined datasets



| Column 0 | age | years_seniority | income | parking_space | attending_party | entree | pets | emergency_contact |
|---|---|---|---|---|---|---|---|---|
| Tony | 48 | 27 | | 1 | 5 | shrimp | | Pepper |
| Donald | 67 | 25 | 86 | 10 | 2 | beef | | Jane |
| Henry | 69 | 21 | 95 | 6 | 1 | chicken | 62 | Janet |
| Janet | 62 | 21 | 110 | 3 | 1 | beef | | Henry |
| Nick | | 17 | | 4 | | | | |
| Bruce | 37 | 14 | 63 | | 1 | veggie | | NA |
| Steve | 83 | | 77 | 7 | 1 | chicken | | n/a |
| Clint | 27 | 9 | 118 | 9 | | shrimp | 3 | None |
| Wanda | 19 | 7 | 52 | 2 | 2 | shrimp | | empty |
| Natasha | 26 | 4 | 162 | 5 | 3 | | | _ |
| Carol | | 3 | 127 | 11 | 1 | veggie | 1 | """" |
| Mandy | 44 | 2 | 68 | 8 | 1 | chicken | | null |

source: Azure AI Gallery

- Missing values can have a meaning, e.g. absence of a medical test could mean that it was not conducted for a reason
- Knowing why and how data is missing could help in data imputation

# Data Cleaning: Missing Values

Knowing why and how data is missing could help in data imputation

- Missing Completely at Random (MCAR)
  - Missingness independent of any observed or unobserved variables
- Missing at Random (MAR)
  - Missingness independent of missing values or unobserved variables
  - Missingness depend on observed variables with complete info
- Missing Not at Ranodm (MNAR)
  - Missingness depends on the missing values or unobserved variable

# Data Cleaning: Missing Values - MCAR

## Missing Completely at Random (MCAR)

- Missingness independent of any observed or unobserved variables
- Values of a variable being missing is completely unsystematic
- This assumption can somewhat be verified by examining complete and incomplete cases
- Data is likely representative sample and analysis will be unbiased

| Age | 25 | 26 | 29 | 30 | 30 | 31 | 44 | 46 | 48 | 51 | 52 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IQ |  | 121 | 91 |  | 110 |  | 118 | 93 |  |  | 116 |  |

Note that values of age variable are roughly the "same" when IQ value is missing and when it is not

# Data Cleaning: Missing Values - MAR

## Missing at Random (MAR)

- Missingness independent of missing values or unobserved variables
- Missingness depend on observed variables with complete info
- The event that a value for Variable 1 is missing depends only on another observed variables with no missing values
- Not statistically verifiable (rely on subjective judgment)

| Age | 25 | 26 | 29 | 30 | 30 | 31 | 44 | 46 | 48 | 51 | 52 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IQ | | | | | | | 118 | 93 | 116 | 141 | 97 | 104 |

- Note that only young people have missing values for IQ
- Shouldn't be the case that only high IQ people have missing values
- Or that only males have IQ values missing (unobserved variable)

# Data Cleaning: Missing Value - MNAR

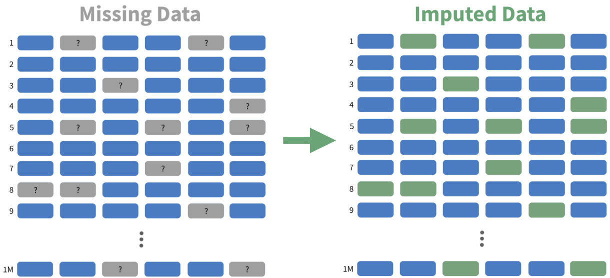
## Missing Not at Random (MNAR)

- Missingness depends on the missing values or unobserved variable(s)
- Pattern is non-random, non-ignorable, and typically arises due to the variable on which the data is missing
- Generally very hard to ascertain the assumption
- e.g. only low IQ people have missing values
- Or only males have missing IQ values

| Age | 25 | 26 | 29 | 30 | 30 | 31 | 44 | 46 | 48 | 51 | 52 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IQ | 133 | 121 |  |  | 110 |  | 118 |  | 116 | 141 |  | 104 |

# Data Cleaning: Dealing with missing values

- Ignore the objects with missing attributes
  - May lose many objects
- Ignore the attribute which has “many” missing values
  - May lose many meaningful attributes  what if class label is missing?
- Impute Data
  - Domain knowledge and understanding of missing values help



source: towards data science

# Data Cleaning: Data Imputation

- Manually fill in, works for small data and few missing values
- Use a global constant, e.g. MGMT Major, or Unknown, or $\infty$
- Substitute a measure of central tendency, e.g. mode, mean or median
  - Missed Quiz: student mean, class mean, class mean in this or all quizzes, the student mean in remaining quizzes
  - Cricket DLS system
- Use class-wise mean or median
  - for missing players score in a match, use player's average, average of Pak batsmen, average of Pak batsmen against India, average of middle order Pak batsmen again India in Summer in Sharjah
- Use average of top $k$ similar objects ▷ based on non-missing attributes
  - can be weighted by similarity average of all other data objects

# Data Cleaning: Data Imputation

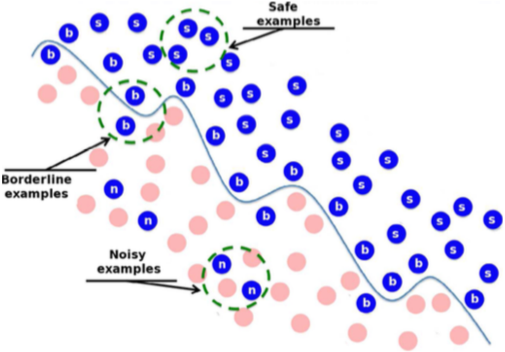
Advanced techniques for imputing missing values

- Expectation Maximization Imputation
- Regression based Imputation

# Data Cleaning: Noise

Noise: Random error or variation in measured data

- Elimination is generally difficult
- Analytics should be robust to have acceptable quality despite presence of noise



# Data Cleaning: Handling Noise and Outliers

Noise and outliers can distort the true picture of data insights and must be managed carefully

| Age | Salary |
|---|---|
| 25 | 50,000 |
| 30 | 55,000 |
| 35 | 60,000 |
| 40 | 650,000 |

Table: Data with Outlier in Salary

# Data Cleaning: Noise

## Dealing with noise

- Smoothing by Binning
  - Essentially replace each value by the average of values in the bin
  - Could be mean, median, midrange etc. of values in the bin
  - Could use equal width or equal depth (sized) bins
- Smoothing by local neighborhoods
  - *k*-nearest neighbors, blurring, boundaries
  - Smoothing is also used for data reduction and discretization
- Smoothing Time Series
  - Moving Average
  - Divide by variance of each period/cycle

# Data Cleaning: Correcting Inconsistencies

Inconsistencies in data can arise from various sources such as human error, data migration, or integration of multiple datasets

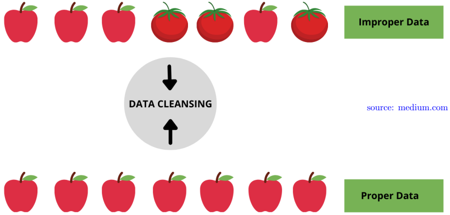
| ID | Product Name | Price |
|---|---|---|
| 1 | Product-A | 20 |
| 2 | product-a | 20 |
| 3 | PRODUCT-A | 19 |

Table: Inconsistent Data Entries

# Data Cleaning: Correcting Inconsistencies

## Data can contain inconsistent values

- e.g. an address with both ZIP code and city, but they don't match



source: medium.com

- Some are easy to detect, e.g. negative age of a person
- Some require consulting an external source
- Correcting inconsistencies may requires additional information

# Data Cleaning: Identifying Outliers

## Outliers are either

- Objects that have characteristics substantially different from most other data

  ▷ the object is an outlier

- Value of a variable that is substantially different than the variable's typical values

  ▷ the feature value is an outlier

- Unlike noise, outliers can be legitimate data or values
- Outliers could be points of interest
- Consider students record in Zambeel, what values of age could be
  - noise
  - inconsitency
  - outlier

# Data Integration

# Data Integration

Data integration involves combining data from different sources to provide a unified view. This process is crucial for comprehensive analysis but comes with challenges

- **Objective:** To merge diverse datasets into a coherent whole
- **Common Issues:** Inconsistencies, entity resolution, duplication

Inconsistencies arise when data from different sources conflict in format, scale, or interpretation

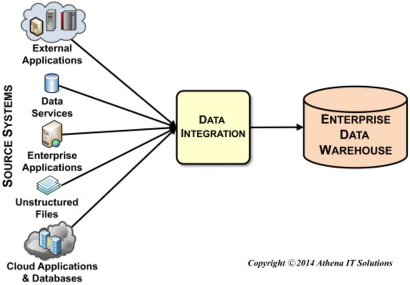
| Date (Source 1) | Date (Source 2) |
|---|---|
| 2024-04-14 | 14/04/2024 |
| 2024-04-15 | 15/04/2024 |

Table: Format inconsistencies in date fields from two sources.

# Data Integration

Merging data from multiple sources

- e.g. RO and Admissions Data Cricinfo and PCB Data



Copyright ©2014 Athena IT Solutions

Data merging causes or require

- Entity identification problem
- Data duplication and redundancy
- Data conflict & inconsistencies

# Data Integration

Entity Identification Problem: Objects do not have same IDs in all sources

e.g. Sentiment analysis on cricket match tweets to assess player contribution

Network Reconciliation Project

- Schema Integration
- Object Matching
  - Make sure that player ID in cricinfo dataset is the same as player code in PCB data (source of domestic games)
- Check metadata, names of attributes, range, data types and formats

# Data Integration

Object Duplication: instance/object etc. may be duplicated

- Occasionally two or more object can have all feature values identical, yet they could be different instances
  - e.g. two students with the same grades in all courses

# Data Integration

## Redundancy and Correlation Analyses

- Redundant (not necessarily duplicate) features
- Sometimes caused by data integration ▷ Data duplication
- An attribute is redundant if it can be derived from one or more others
  - e.g. if runs scored and balls faced are given, then no need to store strike rate
  - If aggregate score in course is given in absolute grading, then no need to store letter grade
- Covariance/Correlation and $\chi^2$-statistics are used for pairs of numerical or ordinal/categorical attributes

# Data Integration

## Data Value Conflict Detection and Resolution

- Sometimes there are two conflicting values in different sources
- e.g. name is spelled differently in educational and NADRA's record
- This might require expert knowledge

# Entity Resolution

Entity resolution is the process of linking and merging records that correspond to the same entities from different databases.

| **Name (Source 1)** | **Email (Source 1)** | **Email (Source 2)** |
|---|---|---|
| John Doe | johndoe@example.com | doe.john@example.com |
| Jane Smith | janesmith@example.com | jane.smith@example.com |

Table: Different email formats for the same individuals across sources.

# Data Integration: Data Duplication

Duplication occurs when identical or nearly identical records exist across datasets, leading to redundancy and possible errors in analysis.

| Customer ID | Name |
|---|---|
| 1 | John Doe |
| 1 | John Doe |

Table: Duplicate records in customer data.

# Data Reduction

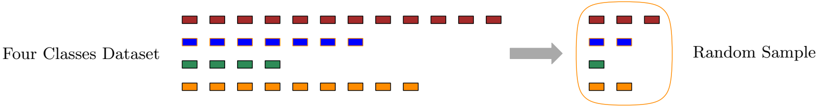
# Data Reduction

Sometime we do not need all the data

We reduce the data in either direction

- Reduce instances
- Reduce dimensions

- Helps reduce computational complexity
- Reduces storage requirements
- Make data visualization more effective
- Get a representative sample of data
- Potentially enhanced model performance



# Data Reduction: Sampling

Equal probability sampling of $k$ out of $n$ objects

- select objects from an ordered sampling window
- first select an object, then every $(n/k)th$ element (going circular)
- If there is some peculiar regularity in the how the objects are ordered, there is a risk of getting a very bad sample

Random Sampling of $k$ out of $n$ objects

- Randomly permute objects (shuffle)
- Select the first $k$ in this order
- Deals with the above regularity issue, but if there is big imbalance among classes or groups, we can get very bad sample

# Data Reduction: Sampling

Stratified Sampling of $k$ out of $n$ objects

- Suppose data is grouped into groups (strata)
- Randomly sample $k/n$ fraction from each stratum
- New sample will exhibits the distribution of population
- Works for imbalanced classes but is computationally expensive

Clustered Sampling of $k$ out of $n$ objects

- Cluster data items based on some ‘similarity’ (details later)
- Randomly sample $k/n$ fraction from each cluster
- Efficient but not necessarily optimal, similarity definition is crucial
- Underlying assumption is that similarity captures the classes

## Data Reduction: Sampling

Imbalanced Classes: Classes or groups have huge difference in frequencies and the target class is rare

Class imbalance is a common issue where some classes are significantly underrepresented in the data, potentially leading to biased models.

- Attrition prediction: 97% stay, 3% attrite (in a month)
- Medical diagnosis: 95% healthy, 5% diseased
- eCommerce: 99% do not buy, 1% buy
- Security: $>$ 99.99% of people are not terrorists
- Similar situation with multiple classes
- Predictions can be 97% correct, but useless
- Requires special sampling methods, oversampling, undersampling

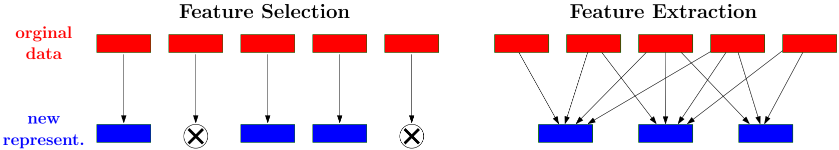
# Data Reduction: Feature Selection

- More importantly, one does dimensionality reduction
- We will study in quite detail the Curse of Dimensionality (problems associated with high dimensions and difficulties in dealing with higher dimensional vectors)
- We will discuss these techniques for dimensionality reduction (time permitting)
    - Locality Sensitive Hashing
    - Johnson-Lindenstrauss Transform
    - AMS Sketch
    - PCA and SVD

# Data Reduction: Feature Selection & Extraction

## Represent data by fewer (and “better”) attributes

- The new features should be so that the probability distribution of class is roughly the same as the one obtained from original features



# Data Reduction: Feature Selection and Correlation Analysis

Feature selection reduces the number of input variables by selecting only the relevant features, often using statistical tests for association like correlation coefficients or chi-square tests.

- High correlation between two features might mean redundancy.
- Chi-square tests are used to determine the independence of two categorical variables.

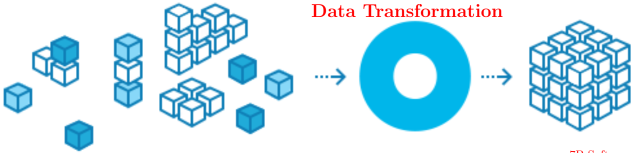
# Data Transformation

# Data Transformation

Data transformation involves converting raw data into a format that is more appropriate for analysis.

Values in original data is transformed via a mathematical function so that

- Compatibility with machine learning algorithms
- Analytics is more efficient - improved data consistency
- Analytics is more meaningful - Enhanced model accuracy
- Visualization is more meaningful and easier



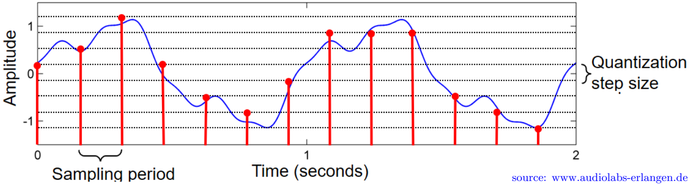
source: 7B Software

# Data Transformation

Values in original data is transformed via a mathematical function

Depending on given data and requirements of analytics, this include

- Ordinal to Numeric ▷ We will discuss it later
- Smoothing ▷ e.g. by binning see dealing with noise
- Aggregation (e.g. GPA from grades)
- Discretization and Quantization ▷ needed e.g. for decision trees



source: www.audiolabs-erlangen.de

- Standardization, scaling and normalization

## Standardization and Scaling

- The goal is to make an entire set of values have a particular property
    - e.g. variables to have the same range, same unit (or lack thereof)
    - to shift the data to a manageable range e.g. shifting to positive
- Variety of possibilities for different applications

# Standardization and Scaling

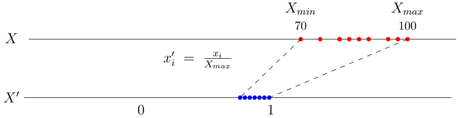
Scaling data so it falls in a smaller, comparable or manageable range

- Data could be in different units e.g. kilometers and miles
- Units might not be known
- Small units means larger values and larger ranges
- In values of "norms" and many distance measures, attributes of smaller units get more weights than attributes with larger units
- All attributes will get the same weight
  - Huge implications in distance values (see clustering & recommenders)

# MAX-MIN Scaling

Transform the data (values of an attribute $X$) to the $\leq 1$

$$x_i' = \frac{x_i}{X_{max}}$$



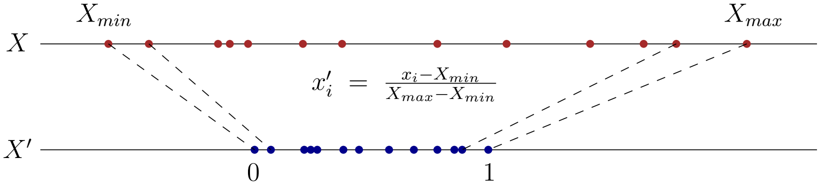
- new max is 1 ▷ new min could be negative
- Preserves relationships among original objects
  - max, min, median and all quantiles are the same objects
- May get very narrow range within $[0, 1]$

| Original Value | Scaled Value |
|---|---|
| 10 | 0 |
| 20 | 0.5 |
| 30 | 1 |

## MAX-MIN Scaling

Transform the data (values of an attribute $X$) to the interval $[0,1]$

$$x'_i = \frac{x_i - X_{min}}{X_{max} - X_{min}}$$



- First shift everything to $[0, sth]$ by subtracting $X_{min}$
- We get different (scaled) std-dev, can suppress effect of outliers
- If attribute $Y$ is also scaled similarly, then $X$ and $Y$ are comparable
- Two sections one with harsh and lenient grading, GIKI and LUMS GPA

# *z*-score Normalization

Transform the data to a scale with mean 0 and std-dev 1

$$x_i' = \frac{x_i - \overline{x}}{\sigma_x}$$

- Good, if we don't know min/max (no full data) or outliers are dominant
    - in such cases MAX-MIN scaled data is harder to interpret
- Stable data, common scale, all variables are unit-less and scalar
- Resulting data have properties of standard normal $\triangleright \mu = 0, \sigma = 1$
- Again the relative order of points is maintained
- It makes no difference to the shape of a distribution

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Sec1 | 90 | 10 | 50 | 30 | 40 | 80 | 74 | 68 | 61 |
| Sec2 | 63 | 40 | 35 | 38 | 21 | 18 | 28 | 19 | 30 |
| Sec1 | 1.4 | −1.9 | −.24 | −1.07 | −.65 | .99 | 0.75 | .5 | .21 |
| Sec2 | 2.3 | .3 | −.14 | .13 | .3 | −1.6 | −.74 | .04 | −.57 |

# Other families of transformation

In statistical analysis we often transform a variable $X$ by a function $f(X)$ of that variable

- It changes the distribution of $X$ or the relationship of $X$ with another variable $Y$
- "Transformations are needed because there is no guarantee that the world works on the scales it happens to be measured on"
- Often it helps and is needed to transform the results back to the original scale by taking the inverse transform
- Mathematical transformations are applied to data to improve its properties for analysis, which includes enhancing normality, linear relationships, and uniformity across features
- **Objectives:** Address skewness, improve model performance, and simplify relationships between variables

# Reasons for Transformation

In statistical analysis we often transform a variable $X$ by a function $f(X)$ of that variable

- Convenience
- Improve the statistical properties of the data
  - Reduced skew
  - Equal Spreads - homogeneity of variance
- Linear relationship: Normalize relationships between features for better correlation analysis
- Additive relations
- Enhance algorithm convergence speeds and accuracy
- For one variable the first three reasons apply

# Reasons for Transformation

In statistical analysis we often transform a variable $X$ by a function $f(X)$

## Convenience

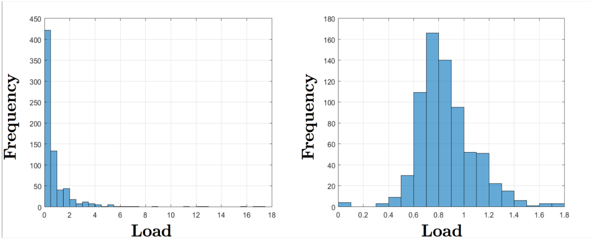
- The transformed scale may be as natural as the original and more convenient for a specific purpose
- Since transformation often change units, one can transform the data to a unit that is easier to think about
- $z$-score normalization is extremely useful for comparing variables expressed in different units
- Rather than $^{101}/_{120}$, $^{130}/_{140}$, and $^{10}/_{73}$, easier to work with percentages. We might want to work with sines rather than degrees

# Reasons for Transformation

In statistical analysis we often transform a variable $X$ by a function $f(X)$

## Reucing Skew

- Many statistical model assume data is from certain distribution with fixed parameters ▷ Generally the (easiest) normal distribution
- Needed to say something like the probability to get a max/mean etc.
- Assumption doesn't have to be true ▷ Data might have skew

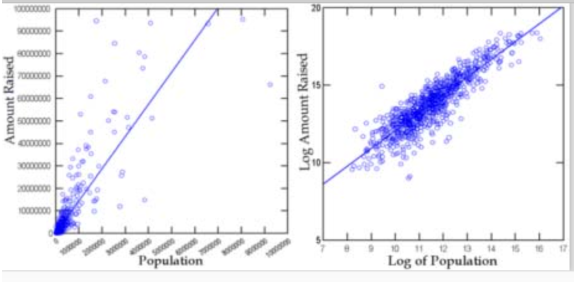


## Reasons for Transformation

In statistical analysis we often transform a variable $X$ by a function $f(X)$

### Equal Spread, Homoskedasticity

- Data is transformed to achieve approximately equally spread across the regression line (marginals)
- Homoskedasticity: Subsets of data having roughly equal spread
- Its opposite property is heteroskedasticity



# Common Transformations

In statistical analysis we often transform a variable $X$ by a function $f(X)$

- All the following transformations improves normality
- Some reduce the relative distance among values while still preserving the relative order
- They reduce the relative distance of values on the right sides (larger values) more than the values on the left side
- They are used to reduce right skew of data
- Issue of dealing with left skew of data is discussed afterwards

# Transformations to Reduce Right Skew

Right skew in data can be handled effectively using transformations that compress large values more than smaller ones

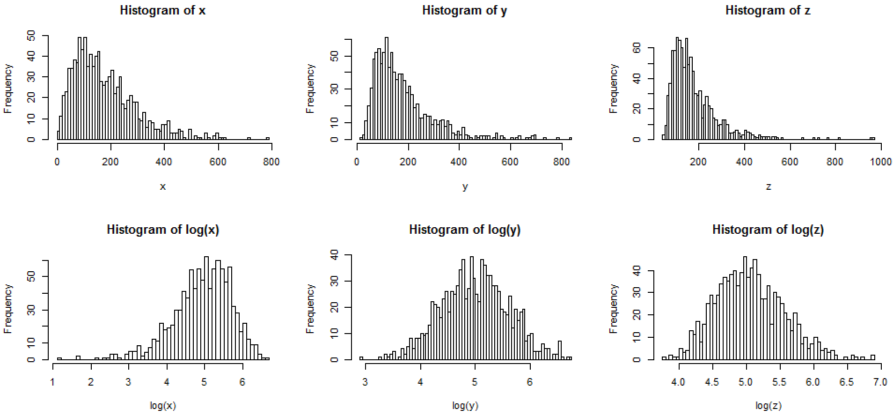
- **Logarithmic Transformation:** Reduces multiplicative relationships to additive.
- **Square Root Transformation:** Mildly reduces skew and is useful for count data.

# Common Transformations: Logarithms

$$x' = \log x$$

- It has major effect on the shape of the distribution
- Commonly used to reduce right skewness
- Often appropriate for measured variables (real numbers)
- Since log of negative numbers are not defined and that of numbers $0 < x < 1$ are negatives, we must shift values to a minimum of 1.00
- Can use different bases (commonly used: natural log, base 2, base 10)
  - One often tries multiple first to settle on one
- Higher bases pull larger values drastically

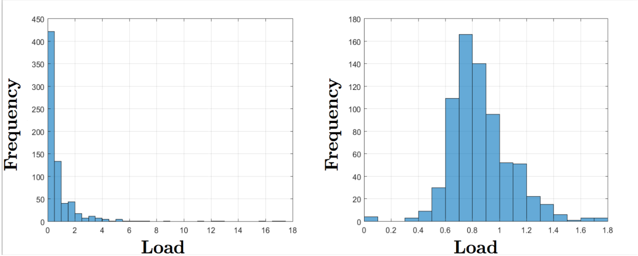
# Common Transformations: Logarithms



# Common Transformations: Cube-root

$$x' = x^{1/3}$$

- Has significant effect on shape of distribution ▷ weaker than log
- Reduces right skew
- Can be applied to 0 and negative numbers
- Cube root of a volume has the units of a length



# Common Transformations: Square-root

$$x' = \sqrt{x}$$

- Reduces right skew,
- square root of an area has unit of a length
- Commonly applied to counted data
- Negative values must first be shifted to positives
- Important consideration: roots of $x \in (0,1)$ is $\geq x$, while roots of $x \in [1,\infty)]$ decreases ($\leq x$), so we must be careful
- Might not be desirable to treat some number differently than others, though the relative order of values will be maintained

# Reciprocal and Negative Reciprocal Transformations

$$x' = \frac{1}{x} \qquad \text{OR} \qquad x' = -\frac{1}{x}$$

- Cannot be applied to 0 ▷ used when all data is positive or negative
  - population density (people per unit area) becomes area/person
  - persons per doctor becomes doctors per person
  - rates of erosion become time to erode a unit depth
- Reciprocal reverses order among values of the same sign
- Makes very large number very small and very small numbers very large
- Negative reciprocal preserves order among values of the same sign, this is commonly used
- This has the strongest effect

# Left Skewed Data: Squares and higher powers

All the above transformation essentially deal with right skew

Left skew (or negative skew) can be reduced by applying transformations that expand smaller values more significantly.

For left skew first reflect the data (multiply $-1$) and then apply these transformations

Generally one needs to shift the data to a new minimum of 1.0 after reflection and then apply the transform

- **Squaring:** Amplifies larger values disproportionately compared to smaller ones, suitable for data with negative values after adjustment.
- **Cubing:** Stronger effect than squaring, can also handle zero and negative values.

$$x' = x^2$$

- moderate affect on shape of distribution
- can be used to reduce left skew

## Transformation to make linear relationship

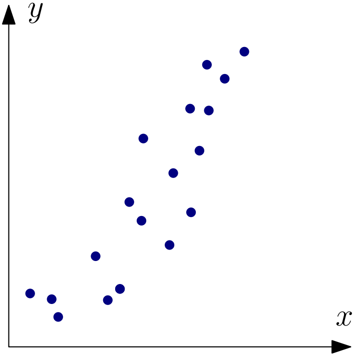
- Suppose we want to describe a variable $Y$ in terms of $X$
- We want to express it as linear relationship

$$Y = aX + b$$

- Transformation in many cases helps us fit a good line

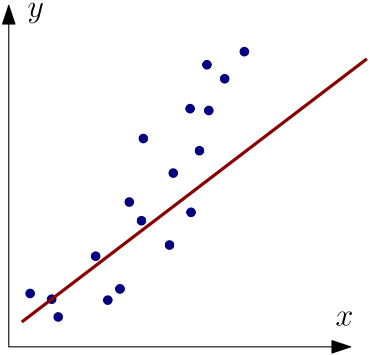
# Transformation to make linear relationship

$$Y = aX + b$$



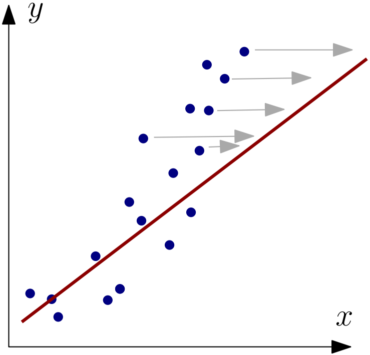
# Transformation to make linear relationship

$$Y = aX + b$$



# Transformation to make linear relationship

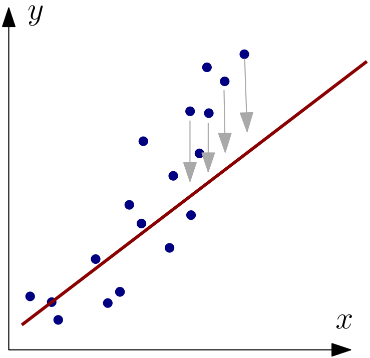
$$Y = aX + b$$



Instead, express as $Y = aX^2 + b$

# Transformation to make linear relationship

$$Y = aX + b$$



Can also do $\log Y = aX + b$