Data Pre-processing

CSCI 5622: Machine Learning

Ayoub Ghriss (Teaching Assistant)

Acknowledgment:
Dr. Ami Gates Notes https://gatesboltonanalytics.com/ , Nidhin Harilal (Spring 2024 TA)

Overview

- Data cleaning is used to describe anything that needs to be done to prepare data for analysis, visualization, and/or modeling.

- For this reason, "**cleaning**" is a broad term that can and will mean different things depending on the **nature of the data** (such as text, record, biological, image, etc.), the **methods or models** you plan to apply, and the **programming language** you plan to use.

- **This is not a one-size-fits-all**, but rather an active and situation-specific process.

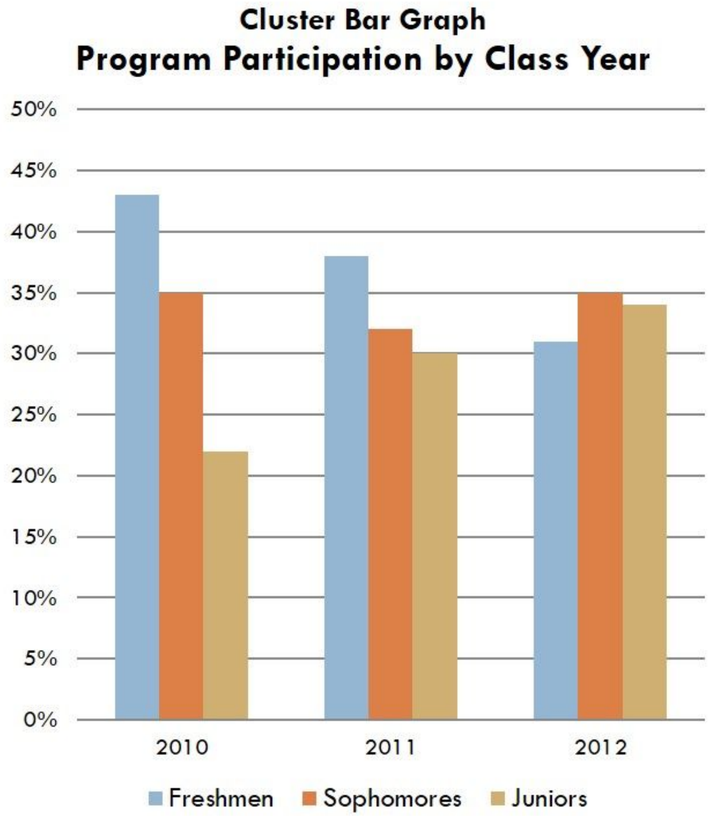
- There are some common elements that are employed when generally cleaning record data, such as managing incorrect, missing, or improperly formatted data, as well as dealing with outliers.

Outline

1. Data Types and Formats
2. Introduction to “Cleaning”
    - Missing Values
    - Incorrect Values
    - Duplications
    - Outliers
3. Formatting for Models
    - Transformation
    - Normalization
    - Aggregation and Handling Noise
    - Dimensionality Reduction

Data Types and Formats

- Data can be:
  - Quantitative (**numeric**)
    - Discrete or continuous, interval or ratio etc.
  - Qualitative (**descriptive and conceptual**)
    - Categorical, Text, Image/Video, Audio etc.
  - And…can be temporal (**time-based** like dates) or geographical (**locations**).



Introduction to Cleaning

Data is the beginning of information discovery.

- **Data cleaning is a cyclic and iterative** set of processes includes exploring, visualizing, updating, formatting, transforming, normalizing, discretization, and correcting data. It is different depending on the data and the goals.

- **Data preparation (cleaning)** often also includes steps such as the identifying and management of outliers, and the generation of new features.

- **These processes are repeated** until the data is prepared for analysis, or clean, and, **are different for different goals and data types.**

Cleaning Record Data

Basic Steps

- While data cleaning can be **unique** for each dataset, each model or method, and each goal, there are some **core commonalities when cleaning/preparing** record style data.

- Recall that record data is organized into rows and columns. Steps:

    1. Managing Missing Values
    2. Managing Incorrect (or incorrectly formatted) Values
    3. Dealing with duplicates
    4. Managing outliers

Missing Values

- Generally – **Finding** missing values can be straightforward.
- However, **correcting** missing values can be difficult or impossible. Why?
- Let’s try finding and correcting missing values in Titanic Dataset!

Handling Missing Values: Options

- Eliminate data objects
- Fill in each missing value manually
- Estimate missing value with global constant (mean or mode)
- Think of all possible values (e.g. use most probably value)
- Randomly select based on regression line

Linear Estimation & More

- In a linear relationship, if the value of one variable changes a certain amount, the value of another variable changes by another certain amount in a specific direction.
- In practice, assuming a linear relationship for missing data determination introduces very little bias.
- Other Methods:
  - Nonlinear submodels, Neural networks, Nearest neighbor estimators
- The purpose of replacing missing values is not to use the values themselves, but to make available the information contained in the other variable-values that are present. **If the missing values are hard to be replaced, the whole instance may be ignored.**

To Note:

- **The missing value estimate** depends as much on which characteristic is to be unbiased as it does on the actual value. Therefore, we need to determine **which relationships need to be preserved**, both within and between variables.

- If many missing values are replaced with the mean, the **confidence level** for statistical inference will be **overoptimistic** since the spread of the data will be reduced.

- It is better to sometime replace the value with **random draws** from the **variable distribution observed**. This means that the values will draw proportionally to the distribution and the center and spread should remain close to the original.

Incorrect Values

- **Incorrect values** are often more challenging to locate within a dataset than missing values.
- Unlike missing values, which tend to adhere to a given format, such as blank or NaN, **incorrect values can be anything**.
- Therefore, incorrect values must first be **discovered,** which requires knowledge about the dataset and the domain.

Duplicate Values

- Datasets can sometimes have **duplicated data.**
- These duplicates may be **true duplicates** - errors caused by the same exact data being repeated by accident.
- **Duplicates may NOT be true duplicates -**
  - For example, in transaction data, two transactions may be identical, but made by different people.

Outliers

How do you Define “too different” or “far”

- An outlier, while noisy and perhaps incorrect, is in its own data –cleaning category.

- An outlier is a value that is significantly “far” or “different” from the other values and from what is expected.
- Is an Outlier:

    1. A mistake?

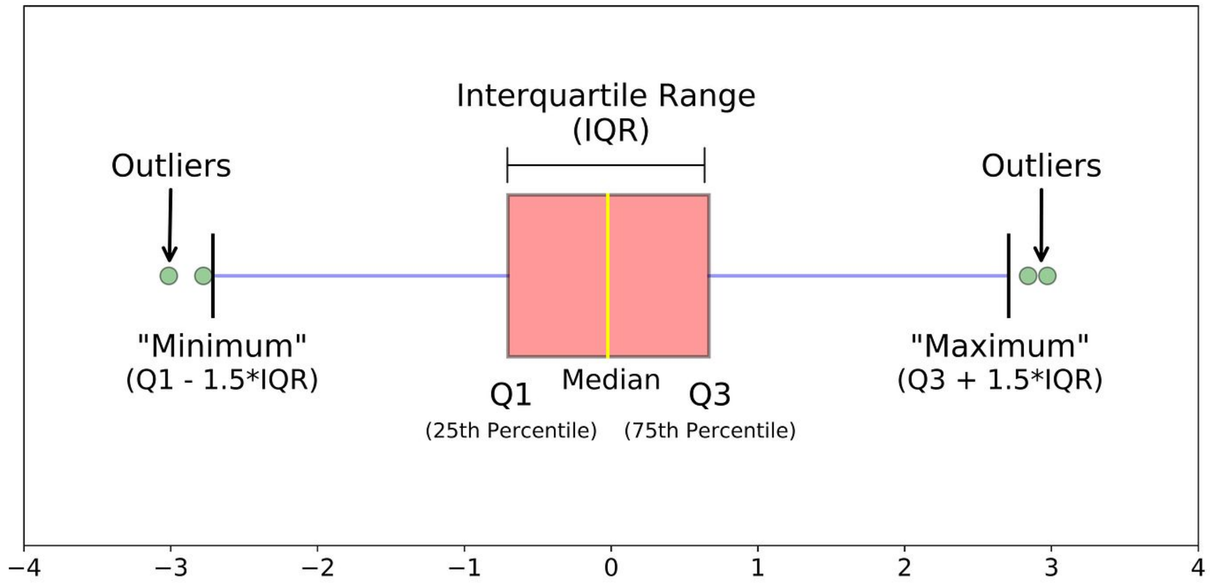
    2. An extreme or different value that is unusual but correct?

The concept of significantly different **requires a measure of similarity.

Outlier Definition

- **There is no universal definition for an outlier.**
- Common definitions include:
  - an observation that differs **so much** from other values that it raises suspicion (Hawkins, 1980)
  - an observation that appears to be **very inconsistent** with the other data (Barnett and Lewis, 1994).
- In both of these definitions, the notion of an outlier is that it is **rare** (so very few or just one data point), and **very different (far)** from all the other data points.
- However, even the idea of “different” is based on a measure of similarity or distance, and **different distance measures can offer different results**.

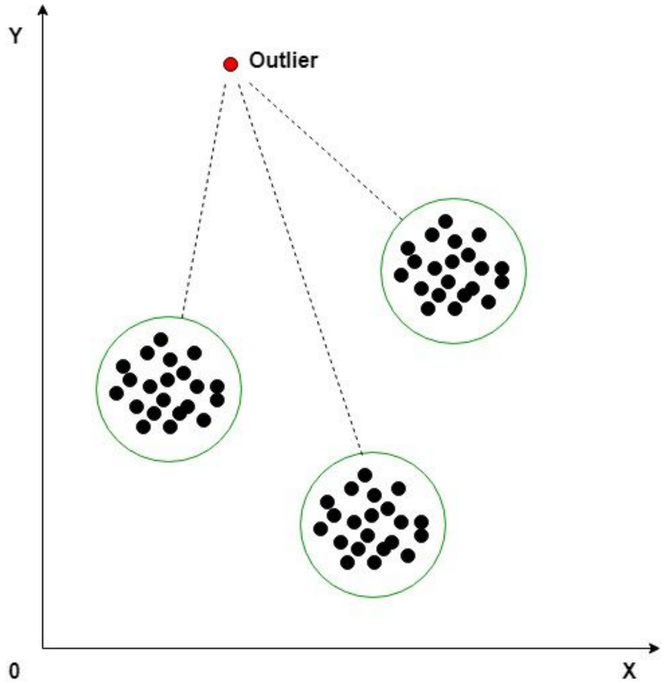
Visualizing Outliers

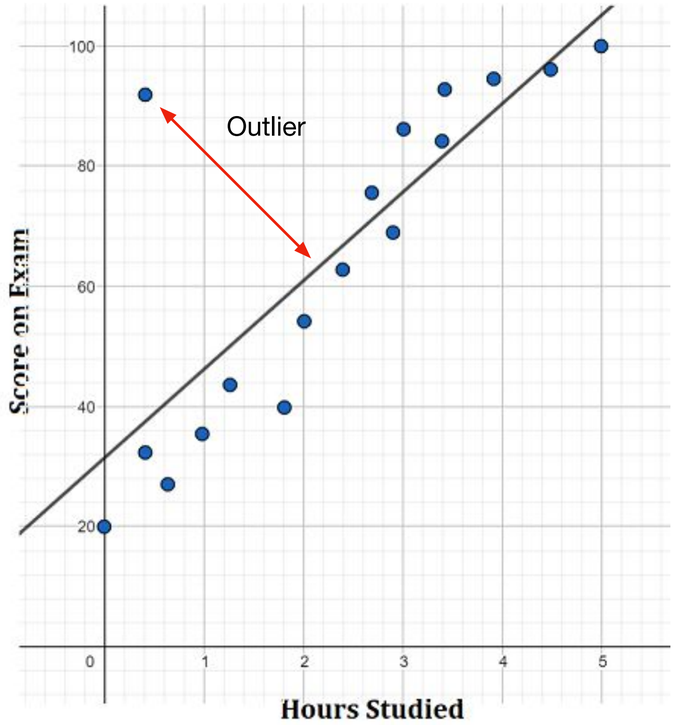


- Now, let's try doing this on the Titanic Dataset!

Image credits: https://towardsdatascience.com/how-to-detect-handle-and-visualize-outliers-ad0b74af4af7

Visualizing Outliers





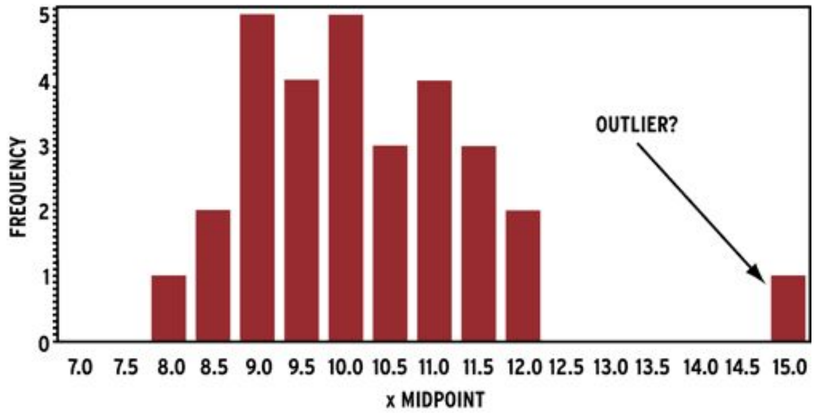


Image credits:
https://study.com/skill/learn/determining-outliers-in-a-scatterplot-explanation.html
https://www.geeksforgeeks.org/challenges-of-outlier-detection-in-data-mining/

The NASA oops of Outliers

- http://www.realclimate.org/index.php/archives/2017/12/what-did-nasa-know-and-when-did-they-know-it/
- A machine collecting data on the ozone removed extremely low values because it logged them as outliers.
- In fact, there were correct and represented a hole in the ozone layer.

The NASA oops of Outliers

- http://www.realclimate.org/index.php/archives/2017/12/what-did-nasa-know-and-when-did-they-know-it/

- A machine collecting data on the ozone removed extremely low values because it logged them as outliers.

- In fact, there were correct and represented a hole in the ozone layer.

- Outliers:

    1. Is it OK to remove outliers? Sometimes.

    2. Are there rules for removing outliers? No.

    3. Can an outlier represent a true datapoint? Yes.

Formatting for Models

Transformation

- The look or expression of data can be altered without destroying its relative integrity.
- Math Transformations:
  - Sqrt, Log, Quadratic (Square, cube), Kernel-based
- Discretization:
  - Binning, Conversion to Binary, Aggregation

Data Remapping

- Words to numbers:
  - Example 1: Text data to tokenized and vectorized frequency count dataframe.
  - Example 2: Male and Female mapped to 0 and 1.
  - Example 3: Theatre Ticket cost (quantitative) mapped to an ordinal variable (such as Group1, Group2, and Group3 – this is discretization).

Normalization

- Why do we need normalization?
- ***Example:*** Consider a dataset with two features: age (ranging from 18 to 90 years) and income (ranging from $20,000 to $200,000).
  - Without normalization, the "income" feature will dominate the "age" feature because of its larger values, which can lead to a **biased model** that doesn't accurately consider the age.
- Every dimension (attribute) is constructed so that its maximum and minimum values are the same. **Linear scaling (Min-Max Normalization):**

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Value re-assignment: Normalization cont.

- Z-score Normalization/Standardization:

$$x_{\text{z-normalized}} = \frac{x - \min(x)}{\text{std}(x)}$$

- Decimal-scaling:

$$x_{\text{decimal-scaling}} = \frac{x}{10^j}$$

Noise in Data

- Examples:
  - Audio (voice poor quality on phone)
  - Video (“snow on TV)
  - Sensor noisy
  - Survey data collection by different people
- Noisy values may be due to:
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems

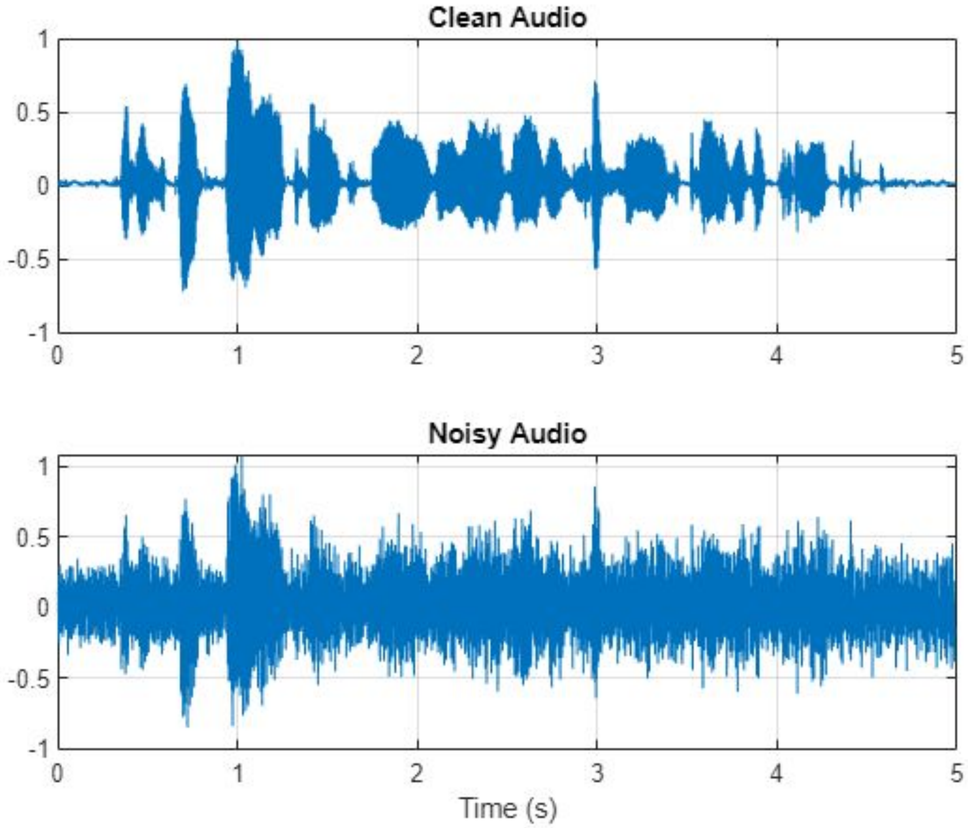


Image Credits: https://www.mathworks.com/help/audio/ug/denoise-speech-using-deep-learning-networks.html

Handling Noisy data

- **Binning**: Look at neighborhood and locally smooth data.
- **Regression**: Smooth by fitting the data into regression functions.
- **Clustering/Anomaly** Detection algorithms.
- Combined computer and **human** inspection.

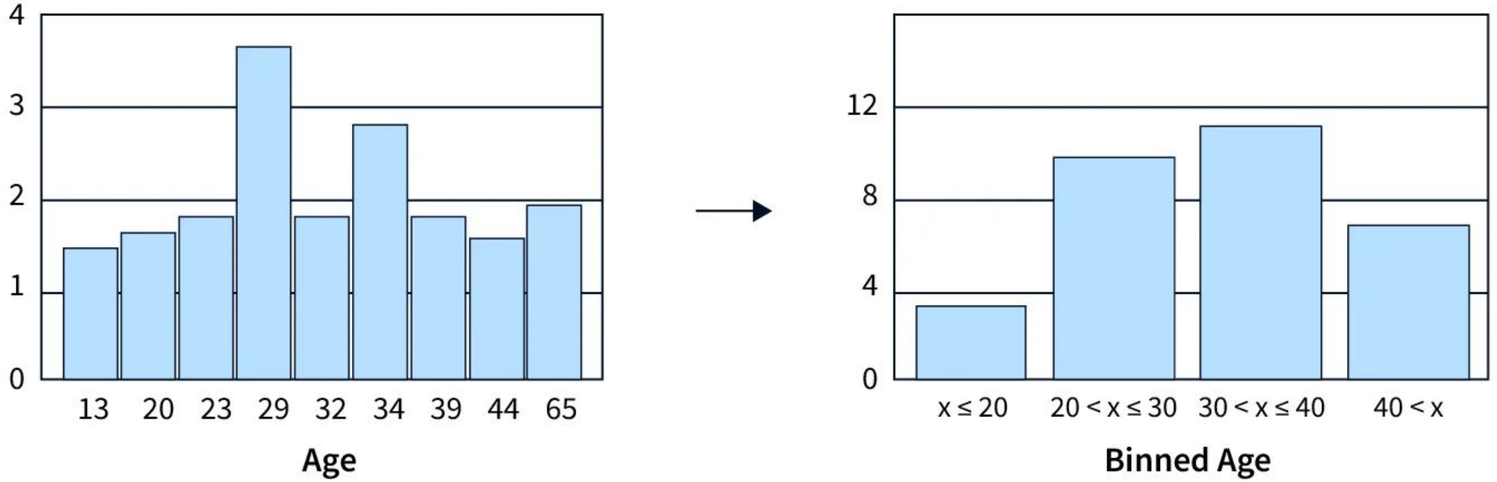


Image Credits: https://www.scaler.com/topics/binning-in-data-mining/
https://www.dailyfx.com/education/technical-analysis-tools/moving-average.html

Data Aggregation

- Aggregation is the process of **gathering** and/or **summarizing** information from different sources into a cohesive, structured format.
- Aggregation can involve combining datasets, summarizing data points, or calculating aggregate statistics (e.g., averages, sums, counts).

Dataset A:

Customer ID	Name	Age

Dataset B:

Customer ID	Purchase Date	Purchase Amount

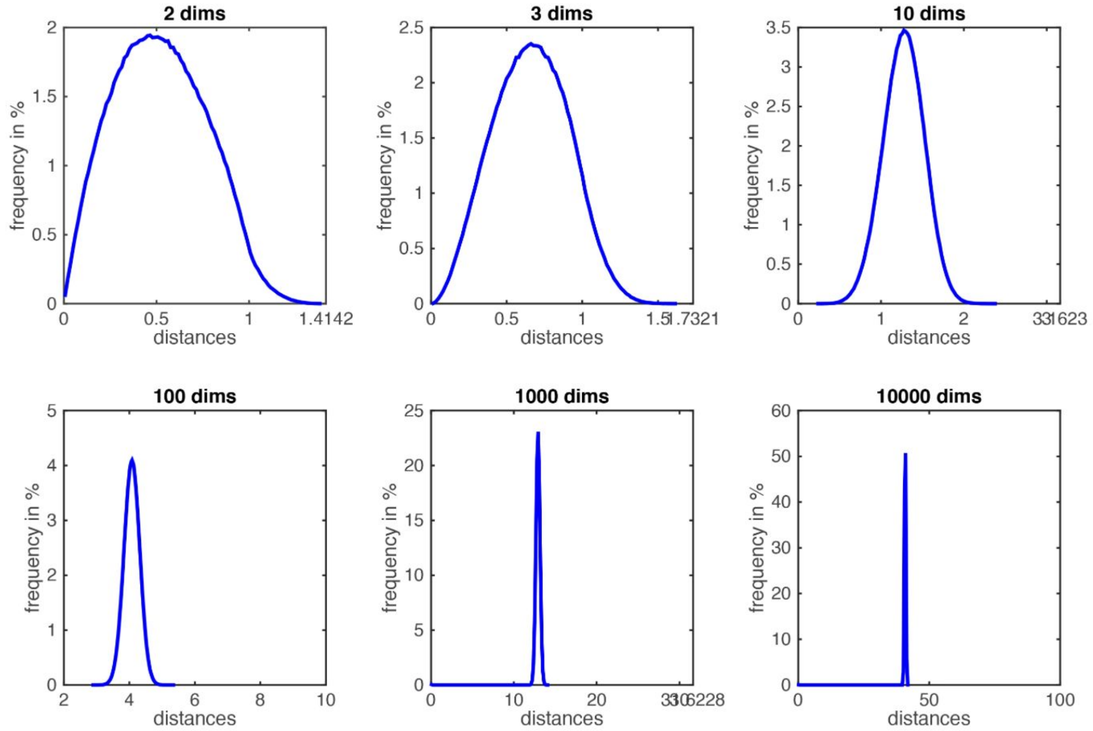
Dataset C:

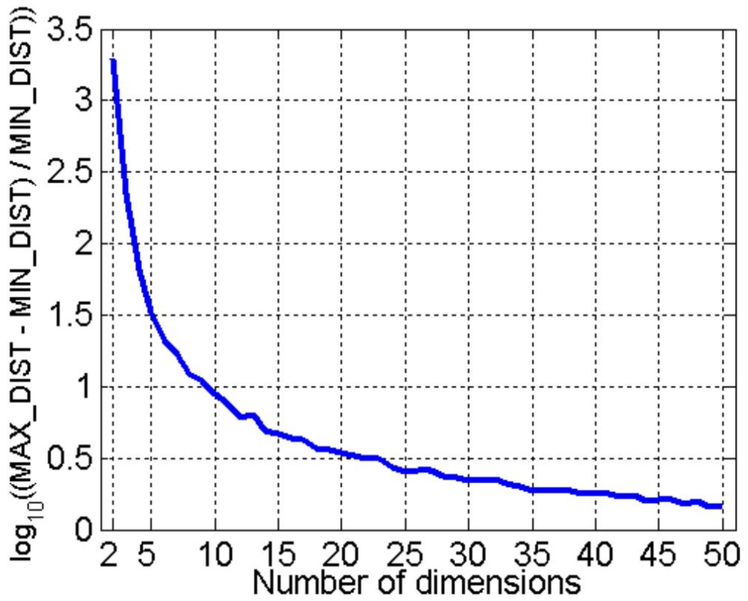
Customer ID	Purchase Amount 14th Jan	Purchase Amount 22nd Jan	Purchase Amount 5th Feb

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful.





- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Image Credits: https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html

Dimensionality Reduction

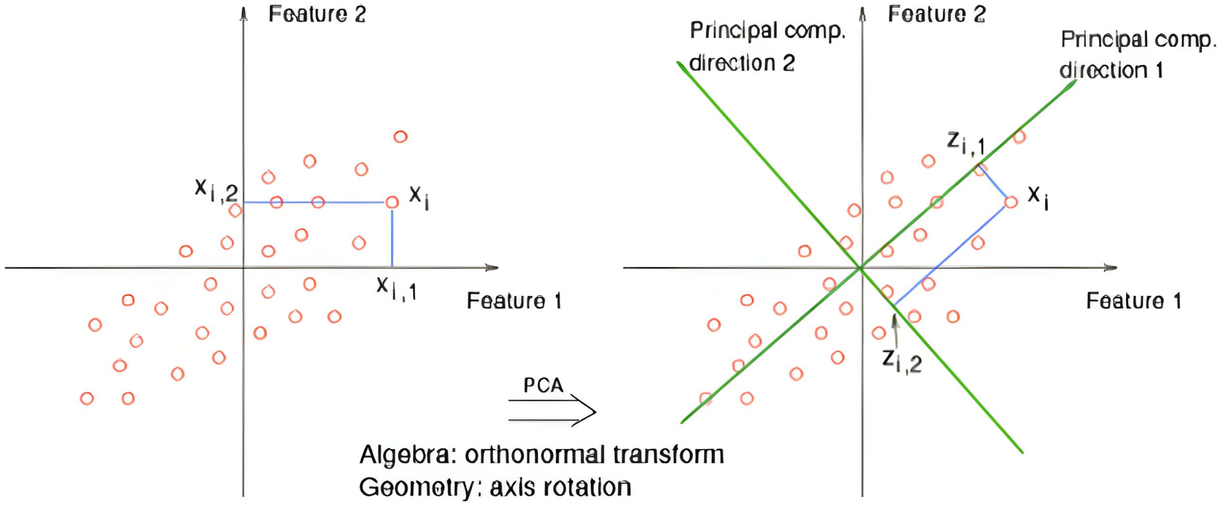
- **Purpose:**
  - Avoid curse of dimensionality.
  - Reduce amount of time and memory required by data mining algorithms.
  - Allow data to be more easily visualized.

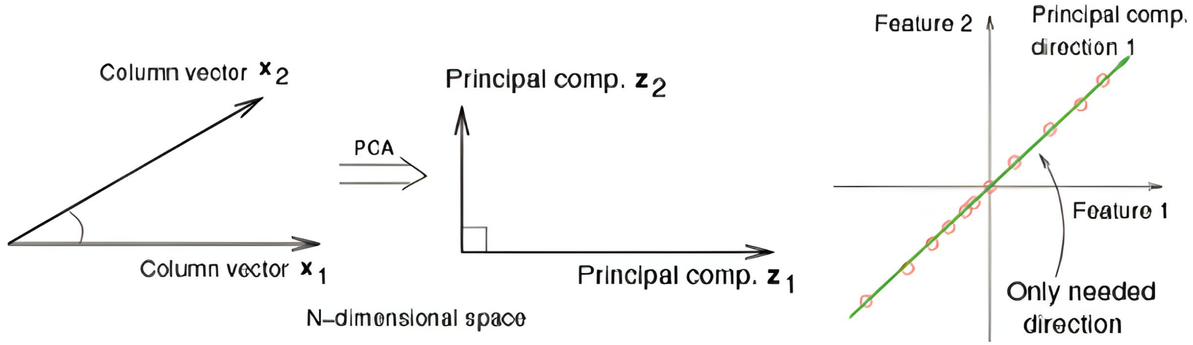
- **Techniques:**
  - Principle Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques.

Principle Component Analysis (PCA)

Intuition

- Goal is to find a projection that captures the largest amount of variation in data.





You will study PCA in detail later in class!

Image Credits:
https://medium.com/@brijesh_soni/topic-14-principal-component-analysis-or-pca-53dfb15b8b9e

End Note

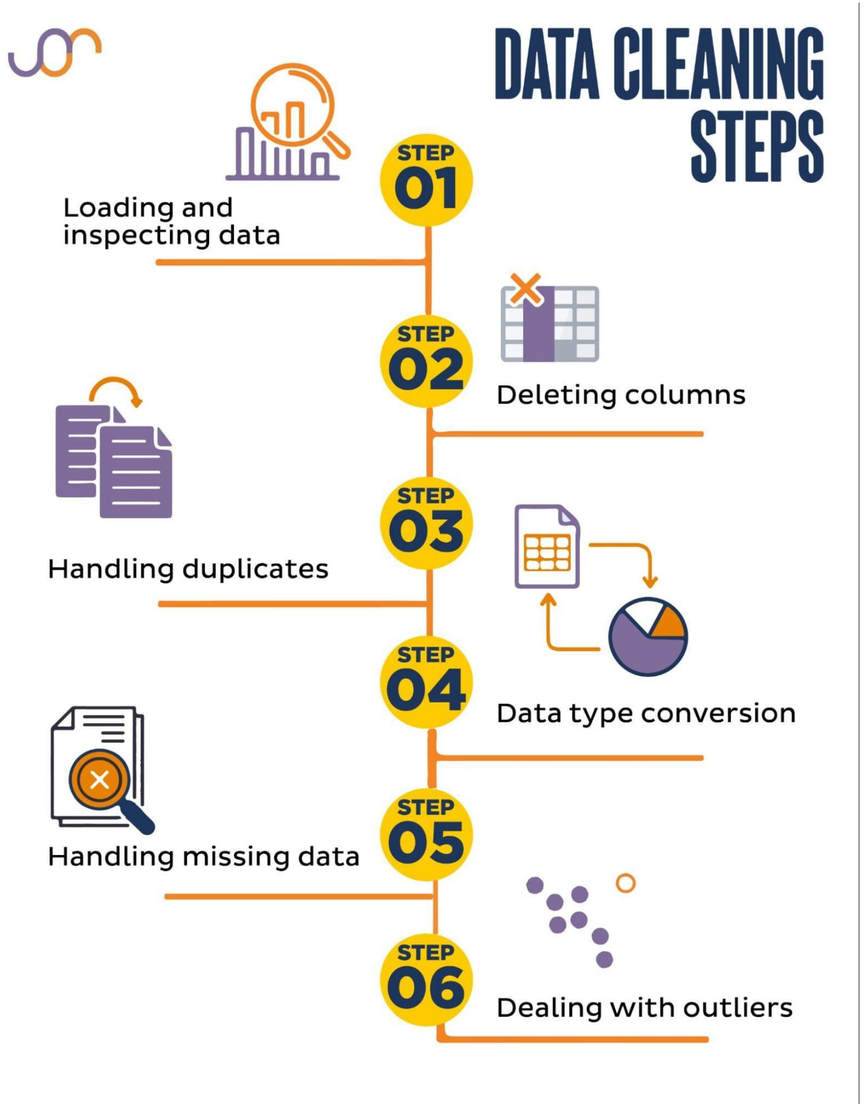


Image Credits: https://www.scaler.com/topics/binning-in-data-mining/
https://www.dailyfx.com/education/technical-analysis-tools/moving-average.html

Thank you!

Credits: https://gatesboltonanalytics.com