Data Preparation

Dr. Teena Sharma
University of Quebec at Chicoutimi, Canada

Summer 2024

What?

The standard definition of data preparation is:

"The process of gathering, combining, structuring, and organizing data."

Why?

- Good data is essential for producing efficient models of any type.
- Data should be formatted according to required software tool.
- Data need to be made adequate for given method.
- Data in the real world is dirty.

Data Cleaning

What?

Data cleaning, also known as *data cleansing* or *data scrubbing*, is the process of identifying and correcting or removing inaccuracies, inconsistencies, and irrelevant data in a dataset.

The goal of data cleaning is to ensure that the data is accurate, consistent, and usable for analysis and modeling.

Common Challenges: Missing Values

In many real-world datasets, some values may be missing or incomplete.

This can occur due to various reasons such as incorrect data entry, missing data, or technical problems during data collection.

Handling missing values is an important step in data cleaning, as it can affect the accuracy and reliability of the results if not properly addressed.

Common Challenges: Outliers

Outliers are values that are significantly different from other values in the dataset.

These can arise due to measurement errors, data entry errors, or other anomalies (rare items, events or observations which deviate significantly from the majority of the data).

Outliers can greatly impact the results of data analysis and modeling, so it's important to detect and handle them appropriately.

Common Challenges: Inconsistent Data Formats

In a dataset, different columns or fields may have different data formats.

This can cause problems when trying to perform data analysis or modeling, as the data must be in a consistent format.

Data cleaning can involve converting data into a consistent format, such as converting date strings into date objects or converting string values into numerical values.

Common Challenges: Duplicates

Duplicate records can occur in a dataset for various reasons, such as repeated data entry or merging of datasets.

Duplicates can greatly impact the results of data analysis and modeling, so it's important to identify and remove them.

Common Challenges: Invalid Data

Some records in a dataset may contain invalid or irrelevant data that does not belong in the dataset.

This can arise due to incorrect data entry, measurement errors, or other anomalies.

Invalid data must be identified and removed during the data cleaning process.

Demo: Titanic Passenger Dataset

Common Challenges: Techniques to handle?

1. Remove rows with problematic values: This method is suitable for datasets with a small amount of missing data, as removing too many rows can greatly reduce the size of the dataset and affect the results of the analysis.

2. Impute missing values: This method involves replacing missing values with estimated or predicted values. There are various imputation methods, including mean imputation, and median imputation.

Common Challenges: how to handle?

3. Interpolation: This method involves estimating missing values based on the values of other observations in the dataset. For example, linear interpolation can be used to estimate missing values based on a linear relationship between two known values.

4. Use a model to predict missing values: This method involves using machine learning or statistical models to predict missing values based on the values of other variables in the dataset. This method can be more accurate than imputation methods, but requires more computational resources and a deeper understanding of the data and the relationships between variables.

Demo: Titanic Passenger Dataset

Data Transformation

What?

- Data transformation is a process of converting data from one format or structure to another format or structure to make it usable for analysis or modeling.

- It is a crucial step in the data preparation process as it helps to transform raw data into a format that is suitable for further analysis and modeling.

Normalization

- A data preprocessing technique used to transform the values in a dataset into a common scale.

- The purpose of normalization is to ensure that all variables in a dataset are on a similar scale, so that they can be compared and analyzed meaningfully.

Normalization Techniques

1. Min-Max normalization: This technique scales the values in the dataset to a range between 0 and 1.
2. Z-Score normalization: This technique standardizes the values in the dataset to have a mean of zero and a standard deviation of one.
3. Decimal scaling normalization: This technique scales the values in the dataset by dividing each value by a power of 10, so that all values have a maximum absolute value of 1.
4. L2 normalization: This technique scales the values in the dataset to have a Euclidean norm of 1.

Encoding

- the process of converting categorical data into a numerical representation that can be processed by machine learning algorithms.

- Encoding is necessary because many machine learning algorithms are based on mathematical operations and require numerical data to work properly.

Encoding Techniques

1. Label encoding: This technique assigns a numerical value to each unique category in the data, usually starting from 0.

2. Ordinal encoding: This technique assigns a numerical value to each unique category in the data based on the rank or order of the categories.

3. One-hot encoding: This technique creates a binary column for each unique category in the data and assigns a 1 to the corresponding column if the category is present, and a 0 otherwise.
   Example?

Encoding Techniques

4. Count encoding: This technique replaces each category in the data with the count of the number of times it appears in the data.

5. Target encoding: This technique replaces each category in the data with the average target value for that category.

Aggregation

- A technique in data analysis that summarizes data by combining multiple values into a single one.

- can be thought of as a way of reducing the dimensionality of the data.

- The most common aggregation functions include summing, averaging, counting, and finding the minimum or maximum value.

- Aggregation is often used in business intelligence, data warehousing, and other data analysis applications to get a high-level view of the data.

Example?

Transformation

- The process of transforming variables to remove skewness or outliers and make the data more normally distributed.

Transformation Techniques

- Log transformation: The logarithmic transformation is used to reduce the skewness in data by transforming the data into a more normally distributed format. The log transformation is applied to data that is heavily skewed to the right.
- Particularly useful in dealing with data that spans several orders of magnitude, where some values are significantly larger than other.

- Square root transformation: The square root transformation is used to reduce the skewness in data by transforming the data into a more normally distributed format. The square root transformation is applied to data that is heavily skewed to the right.

Transformation Techniques

- Box-Cox transformation: The Box-Cox transformation is a family of transformations used to transform skewed data into a more normally distributed format. The Box-Cox transformation can be applied to both left- and right-skewed data.
- It can handle broader range of data types and offer a family of power transformations, allowing more flexibility in transforming data.

- Yeo-Johnson transformation: The Yeo-Johnson transformation is a family of transformations used to transform data into a more normally distributed format. The Yeo-Johnson transformation can be applied to both left- and right-skewed data, and it can handle zero and negative values.

Scaling

- a technique used in data preprocessing to adjust the range and distribution of features in a dataset.

- It helps to standardize the data by transforming it so that each feature has a similar range of values.

- Sounds similar to normalization, right?
  - Think of normalization as a specific type of scaling where we adjust the data to a specific range (0 to 1), and is used for specific purposes such as in deep learning networks and computer vision.

Scaling Techniques

1. Min-Max normalization: This technique scales the values in the dataset to a range between 0 and 1.
2. Standardization: This technique scales the data so that it has a mean of 0 and a standard deviation of 1. It is done by subtracting the mean of the feature from each data point, and then dividing the result by the standard deviation of the feature ($_{\text{Z score}}$).
3. Robust Scaling: This technique is similar to Min-Max scaling, but it uses the median and the interquartile range instead of the mean and standard deviation. This makes it more robust to outliers.

Demo: Iris Dataset

This dataset can be used to illustrate data transformation by transforming the variables in the dataset.

For example,

- normalizing the sepal and petal length and width to bring them to the same scale,
- encoding the species names into numerical values,
- aggregating the data by species to get summary statistics,
- transforming variables like sepal length and width to meet the assumptions of statistical models, and
- scaling variables to prepare them for machine learning algorithms.

Data Integration

What?

- the process of combining data from multiple sources into a single, unified data set.

- This is typically done to enable a more comprehensive analysis of the data and to support more informed decision-making.

- Data integration can involve combining data from databases, spreadsheets, or other sources, and may involve data cleaning, data transformation, and data enrichment to ensure that the integrated data set is accurate and consistent.

Example

Customer information dataset

customer_id	name	address	phone_number
1	John Doe	123 Main St	555-555-5555
2	Jane Doe	456 Oak Ave	555-555-5556
3	John Smith	789 Birch Rd	555-555-5557

Purchase information dataset

purchase_id	customer_id	purchase_date	product
1	1	2022-01-01	T-Shirt
2	2	2022-02-01	Hat
3	1	2022-03-01	Shoes
4	3	2022-04-01	Pants

Merged dataset

customer_id	name	address	phone_number	purchase_id	purchase_date	product
1	John Doe	123 Main St	555-555-5555	1	2022-01-01	T-Shirt
1	John Doe	123 Main St	555-555-5555	3	2022-03-01	Shoes
2	Jane Doe	456 Oak Ave	555-555-5556	2	2022-02-01	Hat
3	John Smith	789 Birch Rd	555-555-5557	4	2022-04-01	Pants

Data Reduction

What?

- A process in data preparation that involves reducing the size or complexity of a dataset without losing significant information.

- The goal of reduction is to make the dataset easier to analyze, understand, and process.

- This can be achieved by either removing redundant or irrelevant data, or by summarizing or aggregating data into fewer variables.

Dimensionality Reduction

- Involves reducing the number of variables or features in a dataset, while retaining as much information as possible.

- Important because high dimensional data is difficult to visualize, process, and analyze.

- One example of dimensionality reduction is using PCA (Principal Component Analysis).

PCA

- a linear transformation technique that transforms a set of correlated variables into a set of uncorrelated variables called principal components.

- The first principal component accounts for the largest amount of variance in the data, the second principal component accounts for the second-largest amount of variance, and so on.

- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

Feature Selection

- Feature selection is the process of identifying a subset of features (or columns) from a larger set of features in a dataset that are most relevant and contribute the most to the target variable or the task at hand.

- The goal of feature selection is to reduce the dimensionality of the data and remove irrelevant, redundant, or noisy features that can negatively impact the performance of machine learning models and make them more complex, slower, and harder to interpret.

Feature Selection Techniques

Feature selection can be performed based on various methods such as

- univariate statistical tests,
- recursive feature elimination, or
- by using machine learning algorithms to estimate feature importance.

Data Compression

- Data compression refers to techniques for reducing the size of a data set.

- The goal of data compression is to store or transmit data using as little storage or bandwidth as possible, without sacrificing the quality of the original data.

Data Compression: Techniques

- This is often achieved through the use of algorithms that identify and remove redundant information from the data, as well as through the use of lossless or lossy compression methods (see next slide).

- Examples of data compression techniques include run-length encoding, Huffman coding, and wavelet compression.

Lossless vs Lossy

- Lossless compression methods preserve the original data exactly without any loss of information.
  - For example, ZIP is a common lossless compression method. Lossless compression methods are mainly used in scenarios where preserving the original data is important, like in medical images, scientific data, and text files.
- Lossy compression methods, on the other hand, discard some of the information in the original data to achieve a higher level of compression.
  - For example, JPEG is a common lossy compression method used in image and video files. Lossy compression methods are mainly used in scenarios where data quality is not as important, like in photos, music, and video files.

Run-length encoding (RLE)

- A lossless data compression method that is used to compress repeating patterns of data.

- In RLE, a sequence of repeating values is represented by a count of the number of repetitions followed by the repeated value. For example, the sequence "AAAABB" could be compressed to "4A2B".

Data Sampling

- Data sampling is the process of selecting a representative subset of a larger dataset for analysis, modeling, or visualization purposes.

- This technique is used when dealing with large datasets as it can save time and computing resources by only processing a portion of the data.

- The representative subset should accurately reflect the underlying distribution and characteristics of the full dataset.

Data Sampling Techniques

- Simple random sampling: This method involves selecting a random sample from the entire population of data. The sample size is determined prior to the selection process and each data point has an equal chance of being selected.
- Stratified sampling: In this method, the population is divided into strata (homogeneous subgroups) and a random sample is taken from each stratum. This is done to ensure that each stratum is represented in the sample in proportion to its size in the population.
- Cluster sampling: In this method, the population is divided into groups (clusters) and a random sample of clusters is selected. Then, all the data points within the selected clusters are included in the sample.