4/8/2024

Data Preparation and Exploration

Change or transforming raw data prior to processing and analysis
May involve reformatting data
Making corrections on data
Combine data sets to enrich the data

Data preparation is frequently the lengthiest part of a data analyst's job/project
Put the data in context
Eliminate bias resulting from poor data quality

Data preparation steps:

- Gathering the data
- Discover and access the data
- Cleanse and validate the data
    - Remove extraneous data, and outliers
    - Filling in missing values
    - Conform data to standard pattern
    - Mask private or sensitive data entries
- Transform or enrich the data

Sometimes this process is called data wrangling
    Is more important for unstructured data, less so for structured

Look closely at data dictionaries or metadata or codebooks that may come with the data that explain the variables, how they were collected, what they measure, units, encoding keys, other encoding schemes, any imputed values, any other transformations that the data has already undergone, etc.

This is a good stage to determine how data will be published and in what format and made available to whom.

Different analysis tools may require different preparations

Data cleaning – missing values, duplicate values, outliers/noise, inconsistencies/conflicting data
Data integration – combining data sets
Data transformation – generalizing of data, normalizing or scaling, aggregation, feature construction
Data reduction – dimensionality reduction, aggregation, data compression, numerocity reduction, generalization
Data discretization – supervised and unsupervised (classification)
Feature engineering – deriving or updating feature

Dealing with missing data:

Option #1 – delete to the rows with missing values
    Hazards include reducing the size of your dataset, or could introduce bias
    Works best if the missing data represents 30% or less of your dataset

Option #2 – impute values into missing spaces (using mean, median, mode, etc.)

Only works if the data is numeric

There may be no mode

Can increase the variance in the data

Variations of imputing data can involve using nearby values in the dataset (preceding or following values) to replace a missing value—works best if the data is sorted, linear. Or time series data.

Preferred method to do with small data sets where loss of data can be problematic.

Option #3 – (for categorical data) is assign the missing values into another category.

Option #4 – predict the missing values with regression

You can split the data into two sets: one without the missing values as the train set, and the other as the "test" set. Use another value in the data set (both data sets) to predict the missing value

Can be easily done in Excel

You can experiment with different combinations of values to get the best replacement value

Option #5 – use algorithm that supports the use of missing values (random forest)

Can be time consuming/computationally intensive especially for large data sets

Does not depend on the relationships between variables

Data Preparation Activities	What to do?	How to do?
Data Cleaning	Dealing with Missing Values/Features	• Ignore respective records having missing values or features • Substitute with dummy value, mean, mode, regressed values or values predicted by an algorithm
	Dealing with Duplicate values/ Redundant Data	• Deletion of duplicate or redundant records
	Dealing with Outliers and Noise	• Binning • Regression (smoothing or curve fitting) • Clustering (grouping values in cluster to identify and eliminate outliers)
	Dealing with Inconsistent/ Conflicting Data	• Use of domain expertise, business understanding, human discretion to correct the data
Data Integration (Integrate multiple sources)	Dealing with issues like Schema integration, entity identification and redundancy	• Joining data sets • Editing metadata to handle data inconsistencies like naming, type etc.
Data Transformation	• Generalization of data	• Concept hierarchy climbing to replace low level attributes with high level concepts or attributes (ex. 'Street' can be generalized to 'country')
	• Normalization/ Scaling of attribute values to a specified range	• Z-score method • Min-Max method • Decimal scaling
	• Aggregation	• Applying summary or aggregation operators to data (ex. Using daily sales to compute annual sales)
	• Feature Construction	• Add or replace with new features derived from existing ones
Data Reduction (Reducing data to make it easy to handle and produce similar analytical results)	• Dimensionality Reduction to eliminate insignificant features	• Feature Selection • Attribute Sampling • Heuristic Methods
	• Aggregation	• Use of aggregation techniques (as above)
	• Data Compression	• Reducing data size by using methods like wavelet transform, PCA etc.
	• Numerousity reduction to have smaller data representations	• Record Sampling, Clustering, Regression etc.
	• Generalization	• Concept hierarchy generation (as above)
Data Discretization (cont. features into discrete)	• Unsupervised (no label is used)	• Binning (equal-width and equal-depth)
	• Supervised (uses labels)	• Entropy-based
Feature Engineering	• Using or deriving the right features to improve accuracy of your analytical model	• Feature Selection • Validation & improvement of features • Brainstorming to create and test more features

Resources:

1. https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation
2. https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281
3. https://blogs.oracle.com/analytics/post/what-is-data-preparation-and-why-is-it-important
4. https://www.upgrad.com/blog/data-cleaning-techniques/
5. https://monkeylearn.com/blog/data-cleaning-techniques/
6. https://www.digitalvidya.com/blog/data-cleaning-techniques/
7. https://technologyadvice.com/blog/information-technology/data-cleaning/
8. https://www.javatpoint.com/data-cleaning-in-data-mining
9. https://coresignal.com/blog/data-transformation
10. https://www.geeksforgeeks.org/data-integration-in-data-mining/
11. https://www.talend.com/resources/data-transformation-defined/
12. https://www.purestorage.com/knowledge/what-is-data-reduction.html
13. https://www.geeksforgeeks.org/data-reduction-in-data-mining/
14. https://www.javatpoint.com/discretization-in-data-mining
15. https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2
16. https://medium.com/codex/data-discretization-b5faa2b77f06
17. https://www.geeksforgeeks.org/discretization-by-histogram-analysis-in-data-mining/
18. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10
19. https://aws.amazon.com/what-is/feature-engineering/
20. https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423
21. https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/
22. https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e
23. https://medium.com/bycodegarage/a-comprehensive-guide-on-handling-missing-values-b1257a4866d1
24. https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html

Extended commentary:

Data preparation and exploration are critical stages in the data analysis lifecycle. Data preparation is the process of cleaning, transforming, and shaping raw data to make it suitable for analysis. This involves identifying and fixing data quality issues, such as missing values, incorrect data types, outliers, and inconsistencies. Data preparation also involves creating new variables, aggregating data, and reducing data dimensionality.

Data exploration, on the other hand, involves visually and statistically analyzing the prepared data to gain insights and discover patterns, trends, and relationships. This involves creating summary statistics, charts, graphs, and other visualizations to explore the data. Exploratory data analysis (EDA) is a common technique used in data exploration, which involves systematically examining and visualizing the data to identify interesting patterns or anomalies.

The goal of data preparation and exploration is to ensure that the data is accurate, complete, and relevant, and that it can be used to answer the research questions or business objectives. By preparing and exploring the data thoroughly, data analysts can gain a deeper understanding of the data and uncover important insights that can guide decision-making.

Data cleaning: This involves identifying and correcting errors or inconsistencies in the data. For example, removing duplicates, correcting misspellings, and filling in missing values.

Data transformation: This involves converting the data into a format that is more suitable for analysis. For example, converting dates into a standard format or converting categorical variables into numerical variables.

Data integration: This involves combining data from multiple sources into a single dataset. For example, merging data from different databases or combining data from different file formats.

Data reduction: This involves reducing the amount of data to be analyzed. For example, aggregating data at a higher level (e.g., summarizing monthly data to yearly data) or using sampling techniques to select a representative subset of the data.

Data formatting: This involves formatting the data in a way that is suitable for analysis. For example, converting text to lowercase, removing special characters, or standardizing units of measurement.

Data wrangling, also known as data munging or data cleaning, is the process of transforming and mapping raw data from one form into another format with the intent of making it more appropriate and valuable for downstream processes like analysis. Data wrangling typically involves several steps, including data cleaning, data transformation, data aggregation, and data enrichment.

During data wrangling, data is cleaned to remove inconsistencies, errors, and inaccuracies, transformed to convert data into a standard format or structure, aggregated to combine data from different sources, and enriched by adding additional data to fill in gaps or provide more context.

Data wrangling is an important step in the data analysis process because it helps ensure that the data being used is accurate, consistent, and relevant to the analysis being performed. It can also help identify data quality issues and provide insights into potential data biases or errors.

Metadata, data dictionaries, and codebooks provide crucial information that can help with data preparation and cleaning. Here's how:

Metadata: Metadata is "data about data," and it provides information about the properties and characteristics of a dataset. Metadata can include information about the data's source, format, structure, content, and quality. This information can help analysts understand the data they are working with, identify any issues or inconsistencies, and plan appropriate cleaning and preparation procedures.

Data dictionaries: A data dictionary is a document or database that contains information about the variables, fields, and columns in a dataset. It provides a detailed description of the data, including the variable name, data type, range of values, and any associated metadata. A data dictionary can help analysts understand the structure of a dataset, identify any missing or invalid data, and standardize data across multiple sources.

Codebooks: A codebook is a document that provides information about how data was collected, including the survey questions and response categories. Codebooks can also include information about data cleaning procedures, such as how missing data was handled and how variables were recoded. A codebook can help analysts understand the context in which the data was collected, identify any issues or inconsistencies, and plan appropriate cleaning and preparation procedures.

By providing this information, metadata, data dictionaries, and codebooks can help analysts ensure the accuracy and consistency of their data, identify any issues or inconsistencies, and plan appropriate cleaning and preparation procedures.

Missing data is a common problem in data analysis, and there are several methods for dealing with it. Some of the most common methods are:

Deletion: In this method, the rows or columns with missing values are simply deleted from the dataset. This method is easy to implement but can result in a loss of valuable information.

Imputation: In this method, the missing values are replaced with estimated values based on other information in the dataset. There are several ways to impute missing values, including mean imputation, mode imputation, and regression imputation.

Multiple imputation: In this method, missing values are imputed multiple times, creating multiple complete datasets. These datasets are then analyzed separately, and the results are combined to produce a final result.

Prediction modeling: In this method, a model is created to predict the missing values based on other variables in the dataset.

Expert knowledge: In some cases, missing values can be filled in by expert knowledge or judgment. For example, if a dataset contains missing values for a person's occupation, an expert in that field may be able to provide an estimate based on other information about the person.

Each of these methods has its strengths and weaknesses, and the choice of method will depend on the specific situation and the goals of the analysis.