Data Science

Data Cleaning
Lecture 2

KAGGLE DATASETS

https://www.kaggle.com/datasets?fileType=csv

- Data cleaning is a crucial step in the machine learning (ML) pipeline.
- Involves identifying and removing any missing, duplicate, or irrelevant data.
- The goal of data cleaning is to ensure that the data is accurate, consistent, and free of errors, as incorrect or inconsistent data can negatively impact the performance of the ML model.
- Professional data scientists usually invest a very large portion of their time in this step because of the belief that “Better data beats fancier algorithms”.
- Data cleaning involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability
- Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.

Characteristics of Quality Data

Validity: Valid data adheres to the rules and constraints set for the specific data type or field. Ensuring validity means checking that the data falls within the acceptable range of values and follows the correct format.

Accuracy: Accurate data is free from errors and closely represents the true value. To ensure accuracy, data cleaning should involve identifying and correcting any incorrect or misleading information.

Completeness: Complete data contains all the necessary information and does not have any missing or null values. Data cleaning should involve filling in missing values or addressing incomplete records to ensure a comprehensive dataset.

Characteristics of Quality Data

Consistency: Consistent data maintains the same format, units, and terminology across the dataset. Data cleaning should involve identifying and resolving any discrepancies or inconsistencies to ensure uniformity and comparability.

Uniformity: Uniform data follows a standard format, making it easier to analyze and compare. Data cleaning should involve converting data into a common format or structure, ensuring that it is consistent and easy to work with.

Id	Name	Date of Birth	Blood Pressure
123434	John Doe	1985-05-15	120/80
234523	Jane Smith	1990-10-23	130/85
987567	Alice Brown	1978-05-12	115/7

Validity

- **Patient ID**: Should be a unique identifier within the dataset, following a consistent format (e.g., numeric). Example: “123456” is a valid numeric ID.

- **Date of Birth**: Should be in a valid date format (YYYY-MM-DD) and represent a realistic date. Example: “1985-05-15” is a valid date format.

- **Blood Pressure**: Should follow the format “Systolic/Diastolic” and fall within realistic physiological ranges. Example: “120/80” is a valid blood pressure reading.

Accuracy

- **Name**: Should accurately reflect the patient's true name without spelling errors. Example: "John Doe" is accurately spelled.

- **Blood Pressure**: Should be a precise reading taken using a reliable method. Example: "120/80" accurately reflects a standard blood pressure reading.

Completeness

- **All Fields**: Should be filled in with no missing or null values. Example: The table shows no missing values; every field for each patient is complete.

Consistency

- **Date of Birth**: Consistently formatted as “YYYY-MM-DD” for all patients. Example: All dates of birth follow the same format.

- **Blood Pressure**: Consistently formatted as “Systolic/Diastolic” across the dataset. Example: “120/80” and “130/85” follow the same format.

Uniformity

- **Patient ID**: Follows a standard numeric format. Example: All IDs are numeric.

- **Blood Pressure**: Follows the same unit and format, making it easy to compare. Example: All blood pressure readings are in the “Systolic/Diastolic” format.

Data cleaning techniques

Here are Some Important data-cleaning techniques:

- Remove duplicates
- Detect and remove Outliers
- Remove irrelevant data
- Standardize capitalization
- Convert data type
- Fixing Structure errors
- Fix errors
- Language translation
- Handle missing values

Remove duplicates

It is likely that you will have duplicate entries if you scrape your data or get it from a variety of sources. These duplication may result from human error on the part of the individual entering the data or completing a form.

Detect and Remove Outliers

Outliers are data points that fall significantly outside the expected range for a particular variable. They can be caused by errors in data collection or measurement, or they may represent genuine but unusual cases. Leaving outliers in your data set can skew your analysis and lead to misleading results.

There are a number of statistical methods for detecting outliers, and the best approach will depend on the specific nature of your data. Once outliers have been identified, you can decide whether to remove them from your data set or to investigate them further.

Remove Irrelevant Data

Any analysis you wish to perform will be slowed down and confused by irrelevant data. Thus, before you start cleaning your data, you must determine what is and is not significant. For example, you do not need to provide your customers' email addresses if you are studying the range of ages of your consumers.

Standardize Capitalization

You must ensure that the text in your data is consistent. Different incorrect categories may be formed if your capitalization is inconsistent.

Since capitalization can alter meaning, it could also be problematic if you had to translate something before processing. For example, a bill or to bill is something else entirely, yet Bill is a person's name.

Convert Data Types

When cleaning your data, numbers are the most frequent data type that needs to be converted. Numbers are frequently imputed as text, but they must appear as digits in order to be processed.

They are categorized as strings and cannot be used by your analytical algorithms to solve mathematical equations if they are shown as text.

Fix Errors

It should go without saying that you must take great care to eliminate any inaccuracies from your data. Typographical errors are just as prone to error and might cause you to overlook important insights from your data. Something as easy as a fast spell check can help prevent some of them.

Errors in spelling or excessive punctuation in data, such as an email address, may prevent you from reaching out to customers. Additionally, you can end yourself sending unsolicited emails to recipients who never requested them.

Language Translation

You will want everything in the same language if you want consistent data.

The majority of Natural Language Processing (NLP) models that underpin data analysis tools are monolingual, which means they cannot process more than one language. Thus, everything will have to be translated into a single language.

Handle Missing Values

Eliminating the absent value entirely could lead to the loss of valuable information from your data. You intended to extract this information in the first place for a reason, after all.

Thus, it could be preferable to fill in the blanks by looking up the appropriate information for that field. You might use the word missing in its place if you’re not sure what it is. You can enter a zero in the blank box if it is numerical.

Fixing Structure errors:

Address structural issues in the dataset, such as inconsistencies in data formats, naming conventions, or variable types.

Standardize formats, correct naming discrepancies, and ensure uniformity in data representation.

Example: To clean data for an e-commerce and retail store, follow these steps:

1. Identify Data Discrepancies Using Data Observability Tools
- Use tools to find errors, inconsistencies, and anomalies, like phone numbers in the "email" field. Prioritize areas needing cleaning based on these insights.

2. Remove Unnecessary Values
- Eliminate irrelevant fields, such as "preferred store location" if only online purchases are analyzed. Streamline the dataset for better focus and accuracy.

3. Remove Duplicate Data
- Delete repeated records to avoid skewed results, ensuring unique and accurate information. This prevents overestimation and inaccuracies in analysis.

4. Fix Structural Errors
- Correct inconsistencies in formats, like standardizing date formats to MM/DD/YYYY. Align mislabeled fields to maintain consistent data structure.

5. Address Missing Values

- Fill or remove missing values using interpolation or regression techniques for completeness. This ensures the dataset is comprehensive and accurate.

6. Standardize Data Entry and Formatting

- Enforce consistent data entry rules, like using title case for names. Standardizing formats minimizes errors and inconsistencies.

7. Validate and Correct Values Against a Known List of Entities

- Cross-check data against predefined lists, such as verifying ZIP codes, to ensure accuracy. This maintains data integrity and reliability.