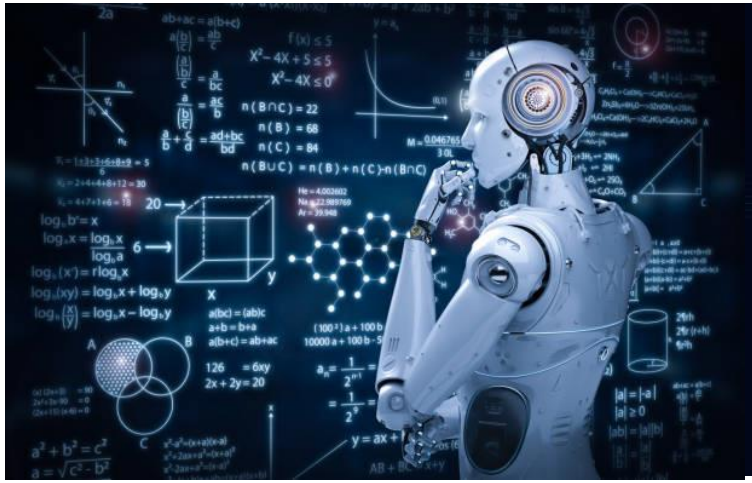
Welcome…

Data Cleaning

CS 797Q
Fall 2024

Sept 16,2024

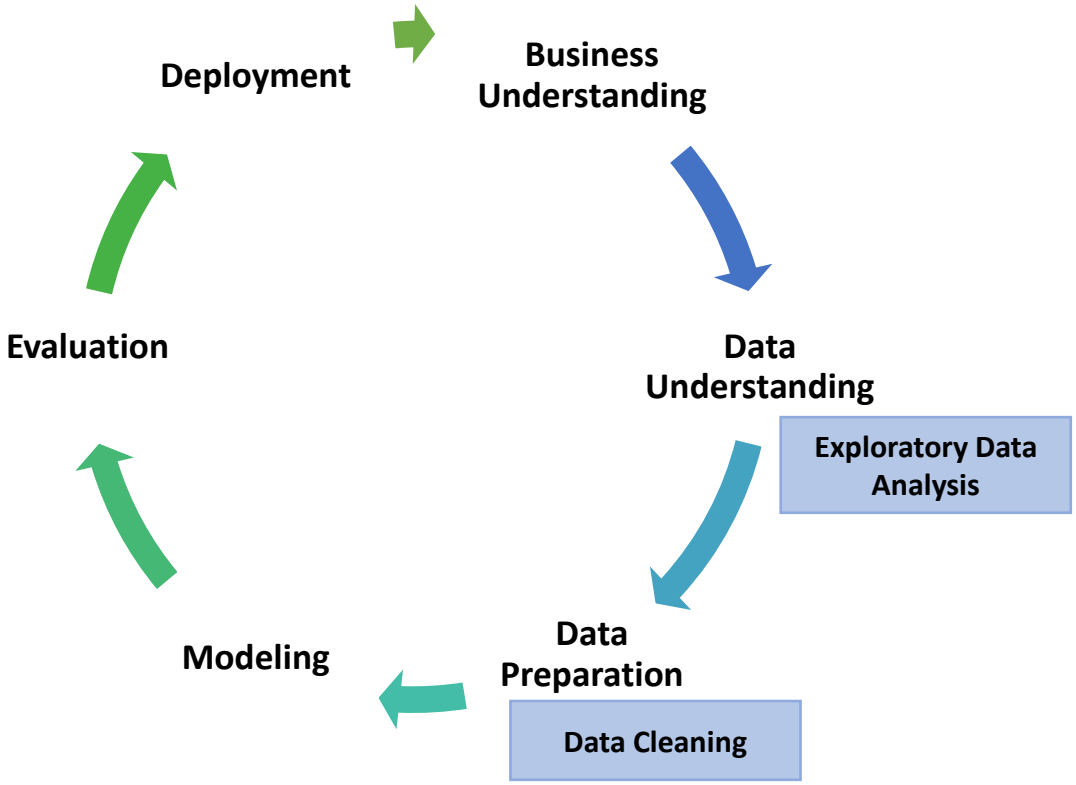


Outline

- Data Cleaning
  - Understanding the Dataset
  - Handling Missing Data
  - Handling Duplicates
  - Handling Outliers
  - Feature Scaling

EDA and Data Cleaning

- The vast majority of your work will be cleaning and exploring data
- Data cleaning and exploration go hand in hand
- This is where you will spend 80% of your time as a data scientist



Data Cleaning

- Data cleaning is the process of identifying and correcting (or removing) inaccurate records, incomplete information, and errors in a dataset.
- Cleaning ensures that the data is reliable, consistent, and accurate for analysis.
- Steps typically include:

- Handling missing data

- Detecting and removing outliers

- Correcting data types

- Dealing with duplicate entries

- Standardizing or normalizing features

Understanding the Dataset

- Before performing any cleaning operations, it is essential to understand the dataset. The following checks help in the initial examination:

- Use `.head()` to inspect the first few rows

- Use `.info()` to check the data types and identify missing values

- Use `.describe()` to understand the distribution of numerical columns

- These steps give insights into the structure of the data and help in identifying potential issues.

Handling Missing Data

- Missing data is common in real-world datasets and needs to be addressed properly.
- Some strategies for handling missing data include:

- **Dropping rows or columns**: Use this when missing values are not numerous and can be removed without losing significant data.

- **Imputation**: Fill in missing values based on certain criteria:

- Numerical columns: Use mean, median, or mode.

- Categorical columns: Fill with the mode (most frequent value).

- **Advanced methods**: Use models or algorithms to predict missing values based on other data points (e.g., KNN, regression).

Handling Duplicates

- Duplicates can occur due to repeated data entry or merging datasets improperly. Identifying and removing duplicates is essential to avoid skewed analysis.

- Use `.duplicated()` to identify duplicated rows.

- Use `.drop_duplicates()` to remove these rows from the dataset.

- Duplicates may not always be exact copies, so care is needed when identifying them.

Handling Outliers

- Outliers are data points that significantly differ from others in the dataset.
- They can distort the analysis if not handled properly. Several techniques to handle outliers include:

- **IQR method**: Identify and remove values that fall outside 1.5 times the interquartile range (Q1 - Q3).

- **Z-score method**: Outliers can also be detected if their Z-score exceeds a certain threshold.

- **Capping or clipping**: Instead of removing, we can cap outliers to a specified maximum or minimum value to reduce their impact.

Feature Scaling

- Feature scaling brings all numerical features onto the same scale, which is critical for many machine learning algorithms:

- **Standardization**: Centers the data around the mean (Z-score normalization).

- **Normalization**: Scales data to fit within a fixed range, such as [0, 1].

- Scaling ensures that larger magnitude features do not dominate algorithms that are sensitive to feature magnitude.

Data Type Conversion

- Incorrect data types can lead to errors during analysis and model training.
- Converting columns to the correct data type ensures accurate processing:

- **Numerical columns**: Ensure integers and floats are correctly assigned.

- **Date and time columns**: Convert strings to datetime objects for proper handling.

- **Categorical data**: Convert text-based columns to categorical data types for efficiency in memory and processing.