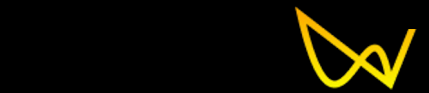
DATA PREPROCESSING

TREVOR YU & CARTER DEMARS

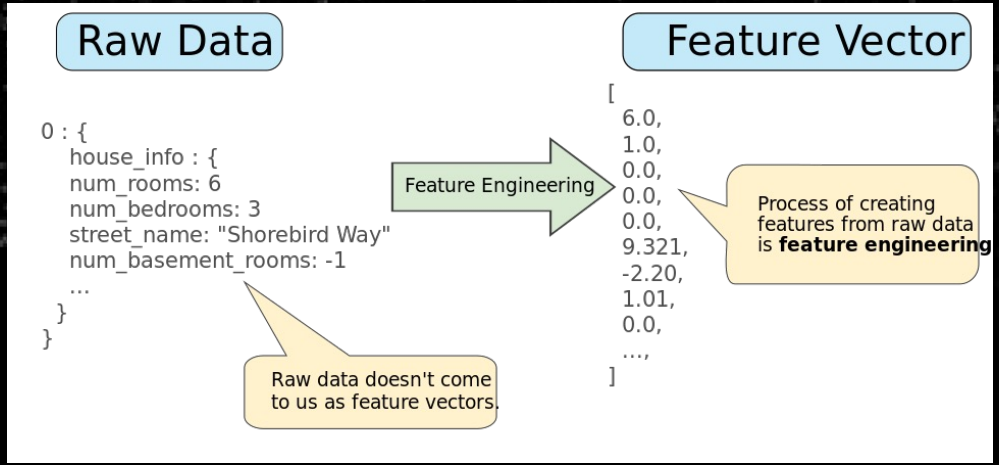


TODAY'S AGENDA

- Learn to perform basic exploratory data analysis (EDA) and data visualization
- Identify outliers, handle missing values, and perform other common data operations such as normalization, interpolation, and filtering
- Understand the intuition behind various preprocessing techniques for both categorical and continuous features
- Apply EDA and data preprocessing techniques to a novel data set without context

ML PRACTITIONERS NEED GOOD DATA

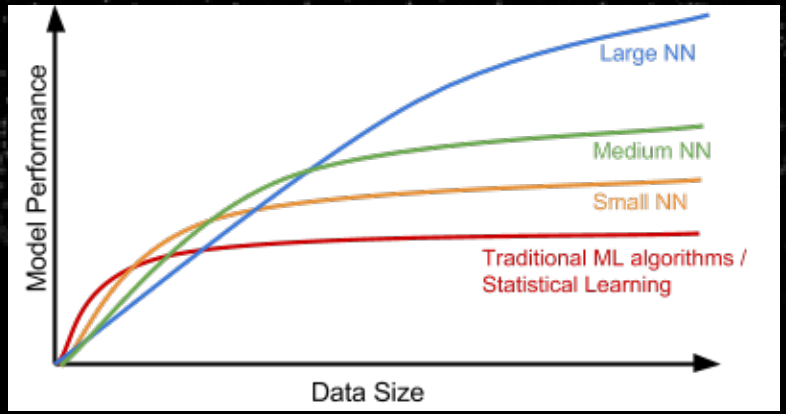
- Most machine learning applications require clean data in the form of vectors
- Most models expect data inputs to be passed in a consistent way
- Different modalities of data (tabular data, images, text) require different preprocessing techniques



ML PRACTITIONERS NEED GOOD DATA

- Machine learning algorithms adjust their parameters based on the patterns they observe.
- For some algorithms, such as neural networks, increasing the size of the training dataset can have a huge impact on the effectiveness of the algorithm.

- On the other hand, insufficient data or poor data quality will often result in an underperforming model



THE IMPORTANCE OF GOOD DATA

- ML Practitioners need to leverage domain knowledge during data preprocessing and feature selection
- Not all data is useful and redundant features can hurt model performance
- Data quality vs. data quantity

USEFUL PYTHON LIBRARIES

NUMPY (NUMERICAL PYTHON)

- Built on top of C
- Library for working with arrays and matrices and python, with the associated high-level functions to operate on these arrays

SCIPY (SCIENTIFIC PYTHON)

- Scientific and technical computing
- Optimization, linear algebra, signal processing, integration, eigenvalue problems

USEFUL PYTHON LIBRARIES

PANDAS

- Open-source data analysis & manipulation tool
- Reads data into a series/DataFrame
- Features for dealing with missing data, changing data format, aggregating data, slicing, sorting and applying transformations

SCIKIT-LEARN

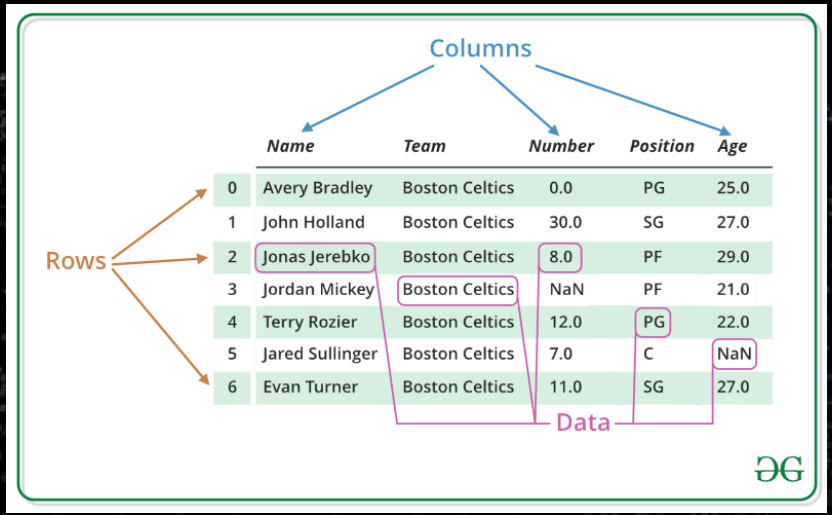
- Tools for data analysis and machine learning
- Highly popular library with classification, regression, and clustering algorithms
- Built on top of NumPy and SciPy

WAT.ai

CODE-ALONG ACTIVITY

DATAFRAME BASICS

- A DataFrame is a 2-dimensional tabular data representation
- Uniquely labeled axes, called rows and columns
- Each row is typically one collected data point
- Each column contains values of a "feature" across many examples



DATAFRAME BASICS

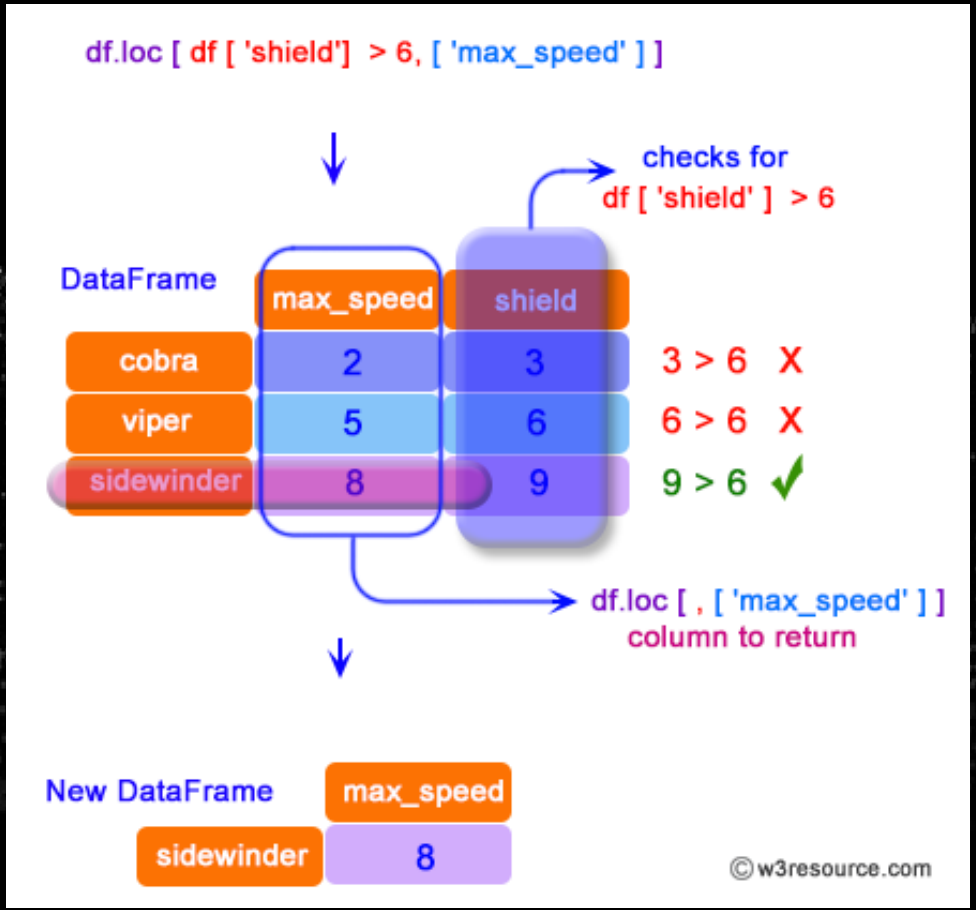
- By convention, refer to dataframes as `df` in code
- `df.index, df.columns, df.values` can access these aspects of the dataframe
- `df.shape` gives the number of (rows, columns)

RENAMING COLUMNS

- Sometimes, datasets will need to be combined or analyzed together but will have inconsistent naming conventions
- Consistent dataset structure = higher quality data
  - Column names
  - Data types
  - Representation of invalid values
- Familiarity with the datasets and domain specific knowledge will inform how to best formulate a workable dataset structure

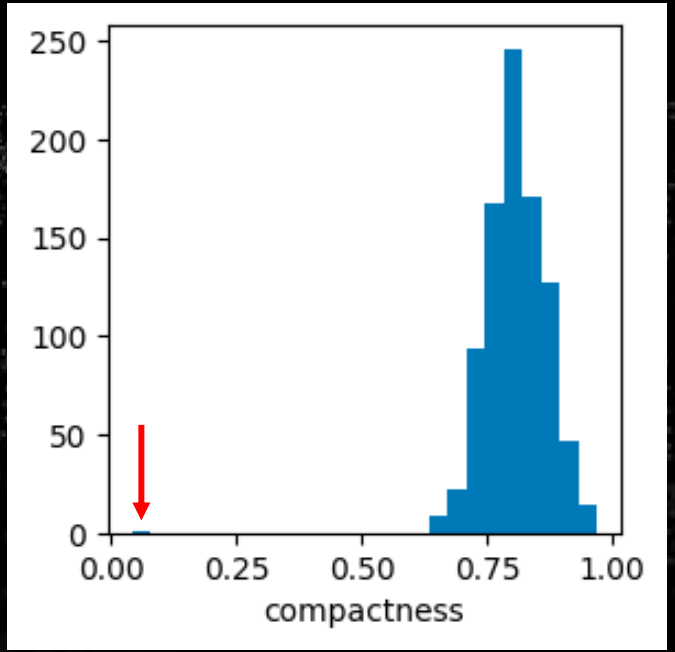
VIEWING SUBSETS

- Common operations:
  - Selecting columns
  - Selecting rows by index
  - Selecting rows by condition
  - Sorting results by value
- Selections do not happen in-place, results must be assigned to variable to persist



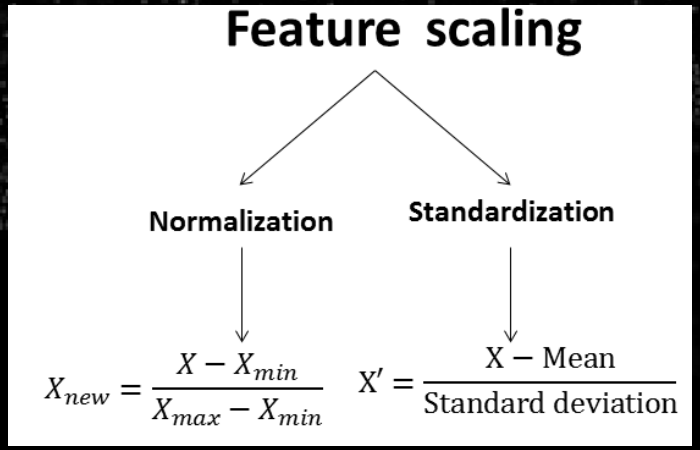
OUTLIER REMOVAL

- Outliers are data points that deviate significantly from the mean distribution of the data
  - Often represent incorrect measurements
  - Domain knowledge is often required to interpret the meaning of outliers
- Heuristic methods like thresholds or statistical methods like Z-scores
- Removing too many data points could adversely affect a model's ability to generalize



FEATURE NORMALIZATION

- Many ML models assume that the input data (as vectors) are roughly normally distributed with 0 mean and unit standard deviation
  - Sometimes, scaling data between [-1, 1] is also used
- Model performance generally improves when features are normalized
- Typically, this is the last step before passing the data to a model



WORKING WITH DIFFERENT DATA TYPES

- **Ordinal categorical variables** have discrete categories whose order matters (ex: small, medium, and large)
- **Nominal categorical variables** have discrete categories without order, so concepts such as the mean have no interpretation (ex: gender)

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

HANDLING MISSING VALUES

TYPES OF MISSING VALUES

Missing Not At Random - when a value is missing for a reason related to the true value. (Ex: if a survey responding chooses not to disclose their income, this could be because they have an abnormally high or low income)

Missing at Random - when a value is missing for a reason related to another observed variable. (Ex: many age values are missing for survey respondents of a particular gender)

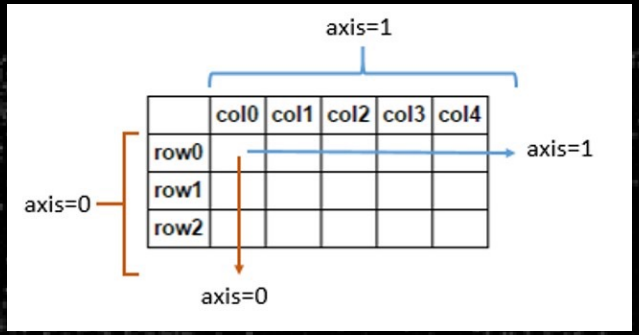
Missing Completely at Random - when there's no patterns in the missing values.

HANDLING MISSING VALUES

DELETION

Column deletion: removing a column that has too many missing values and is non-essential for your model

Row deletion: removing rows with missing values, ideally if the missing values are Missing At Random, to avoid biasing your model



IMPUTATION

- Fill missing values with their defaults (empty string, zero, etc...)
- Fill missing values with the mean, median, or mode
- Backward or forward fill

INTERACTIVE ACTIVITY

SPOTIFY EMOTION CLASSIFICATION

- The full dataset was split using a stratified train test split: the proportion of labels in the test set it the same as in the training set

UPCOMING EDUCATION SESSIONS

- Classical Machine Learning
- Neural Networks for Novices
- Dive into Deep Learning

EXIT SURVEY – ATTENDANCE!