Data Preprocessing

By: Matt Collins, Surya Maddali

Attendance: Meeting Name “PreProcess”

Table of contents

Why Preprocess Data?

Data Quality

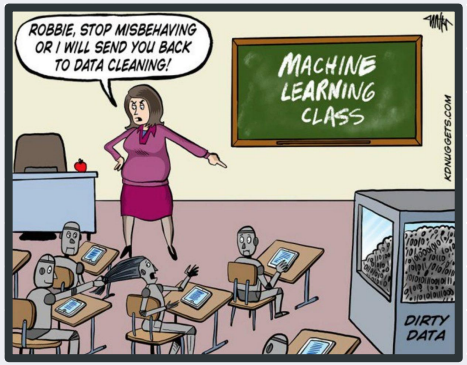
- Incomplete Data
- Noisy Data
- Data Discrepancy

Incompatibility

- Data Transformation
- Normalization

Interpretability

- Pattern recognition
- Bias removal



Colab

Datasets and Libraries

Datasets

- *Ideally imported in **csv** form*
  - Comma-separated values
- Python - *Pandas*
- AirBnB open dataset

Datasets and Libraries

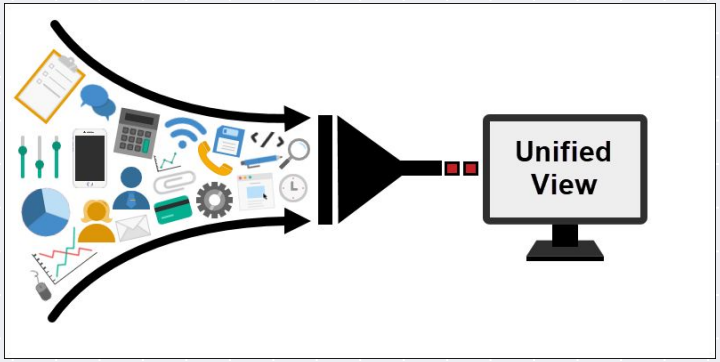
Libraries

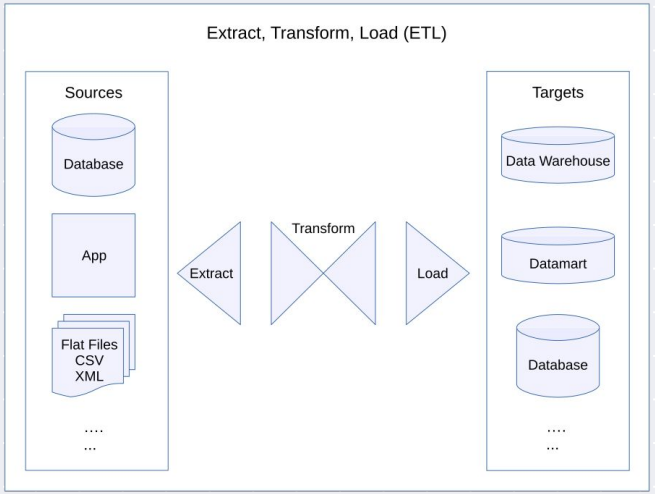
- *Pandas* library
- *Numpy* library
- *Sci-kit learn*
- *matplotlib*

Data Preparation

Data Integration

- Different sources → **One View**
- Timing varies
- Extract, Transform, Load
- Data Lakes/Warehouse

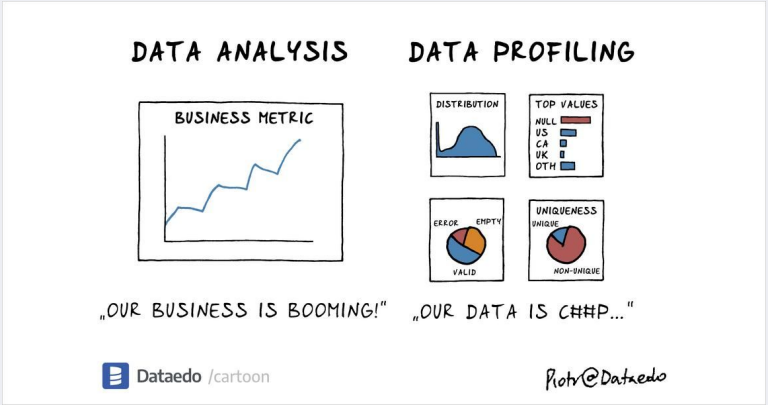


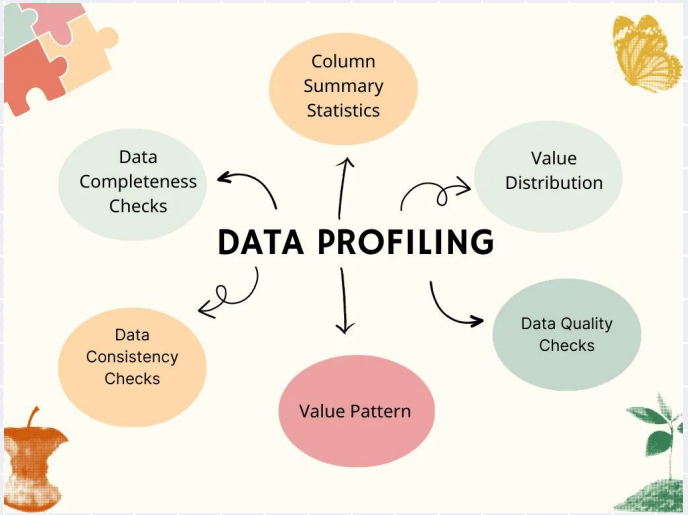


Data Preparation

Data Profiling

- Assess data quality
- Structure, content, relationship discovery
- Repeated throughout process

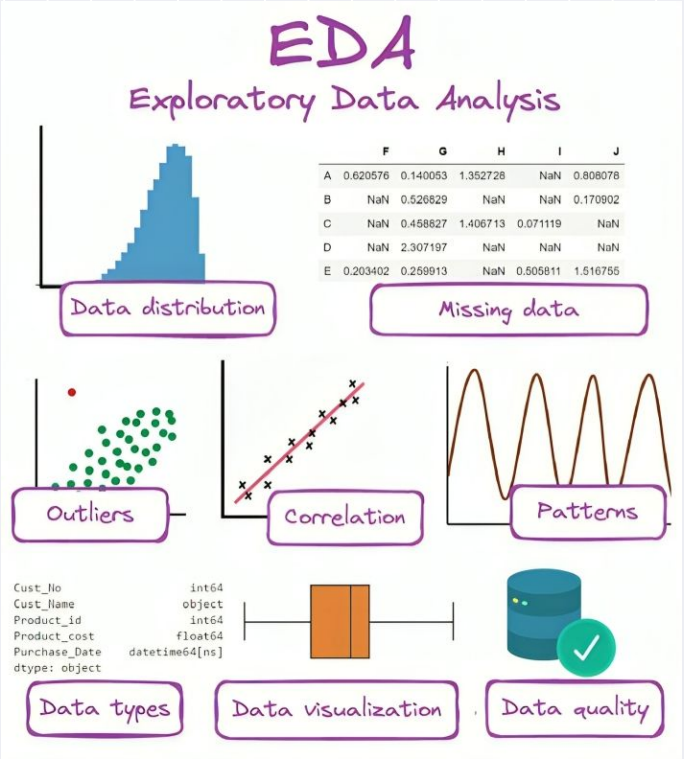


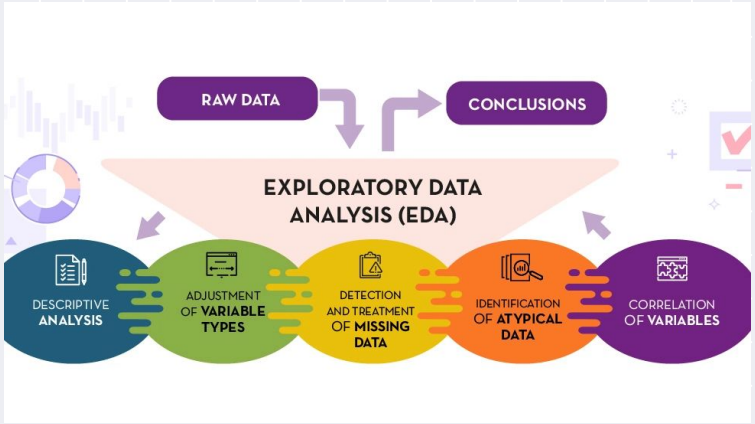


Data Cleaning

- Correcting inaccurate, incomplete, or irrelevant information from a dataset
- Nulls
- Noisy Data
- Inconsistent Data

Exploratory Data Analysis

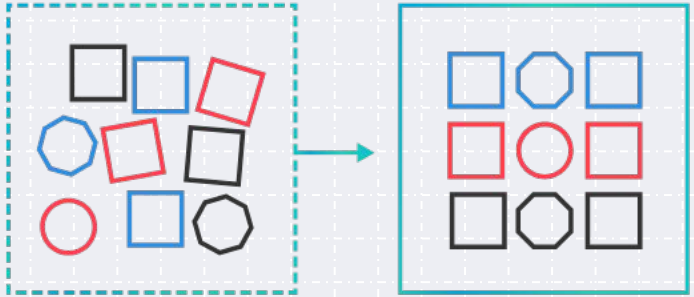


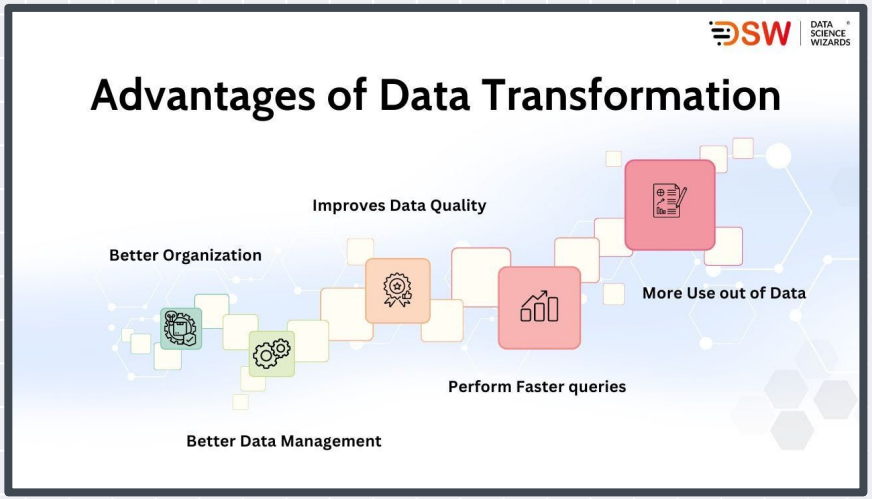


- Visualizations to analyze data
- Reveals patterns, checks assumptions
- More detailed than profiling
- For scientists and stakeholders

Data Transformation

- Changing structure or format
- Goal to improve usability
- Various techniques
- Feature extraction
- Encoding and normalization





Data Transformation

Encoding

- Categorical data → Quantitative
- Label encoding vs. one hot encoding
- One number per unique entry
- Columns of unique entries

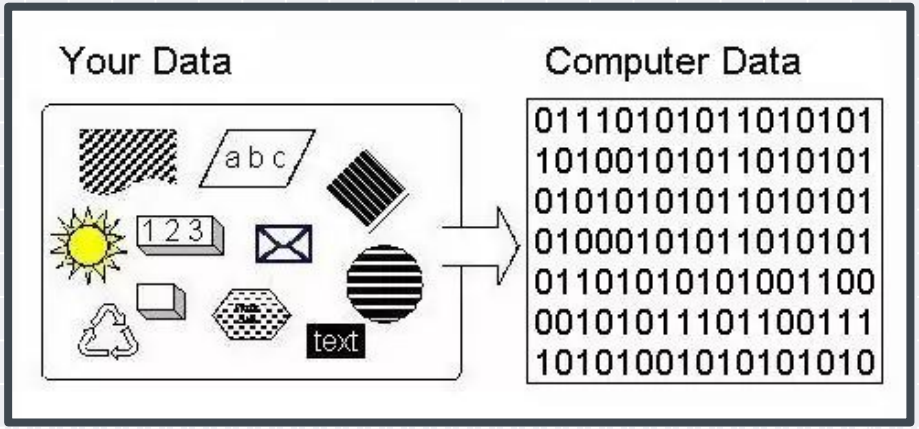
Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

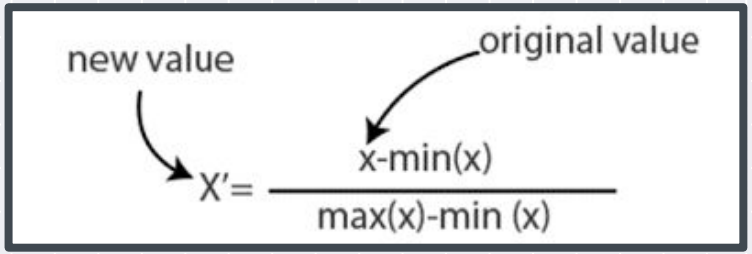
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50



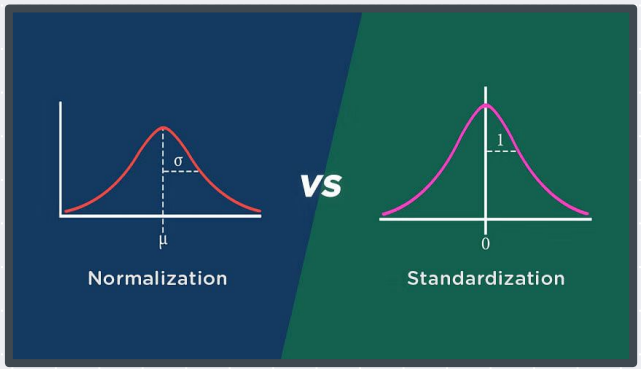
Data Transformation

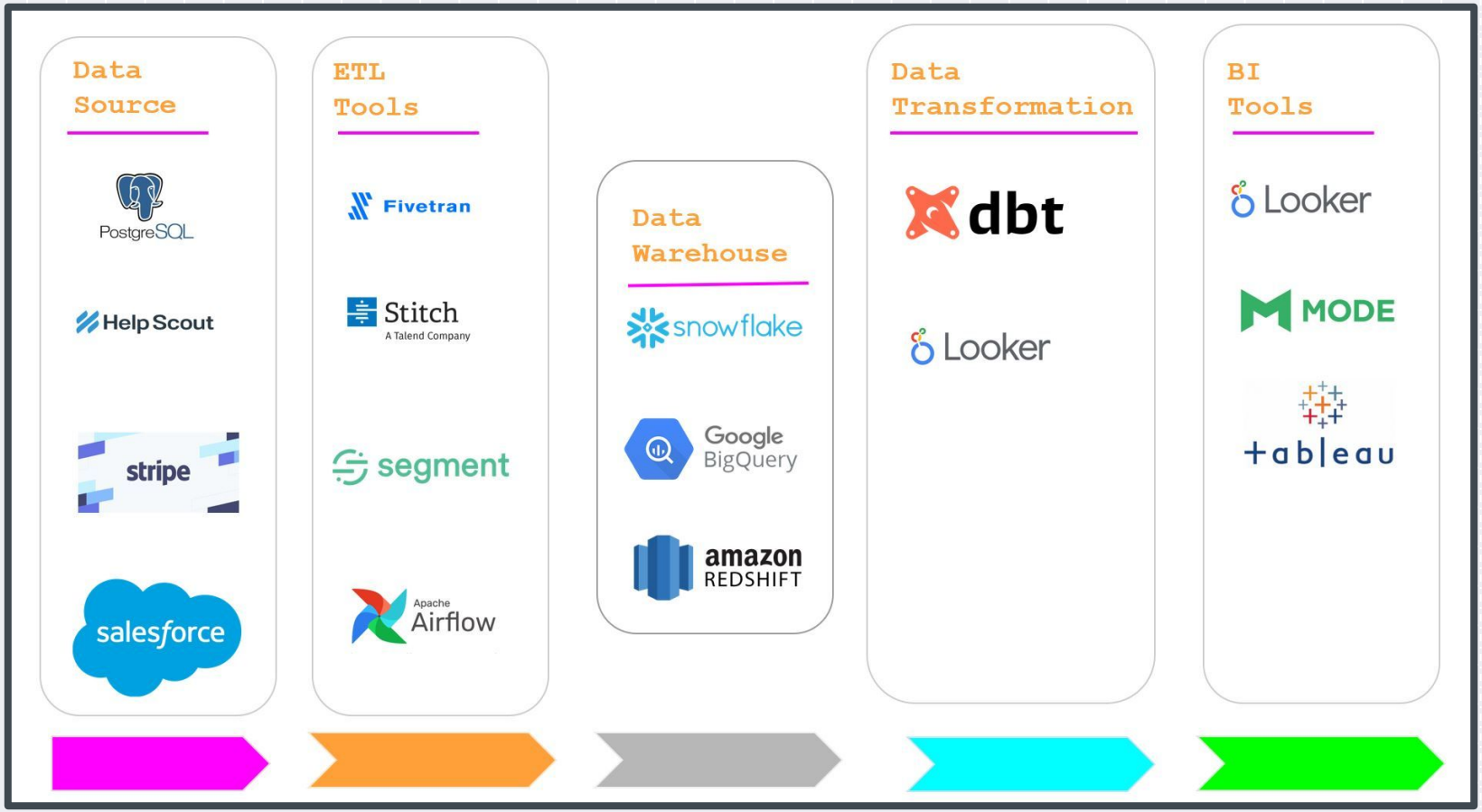
Normalization/Standardization

- Quantitative → Quantitative
- Protection against outliers
- Plug data into formulas
- Normalization vs. Standardization



$$Z = \frac{X - \mu}{\sigma}$$




Data Source
PostgreSQL
Help Scout
stripe
salesforce
ETL Tools
Fivetran
Stitch
A Talend Company
segment
Apache Airflow
Data Warehouse
snowflake
Google BigQuery
amazon REDSHIFT
Data Transformation
dbt
Looker
BI Tools
Looker
MODE
+ableau

Thanks NDL!

Feel free to stay and ask questions!