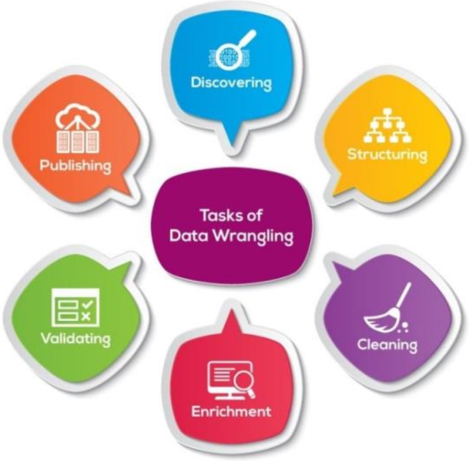
DATA WRANGLING

CCDATSCL | COM221-ML

Data Wrangling

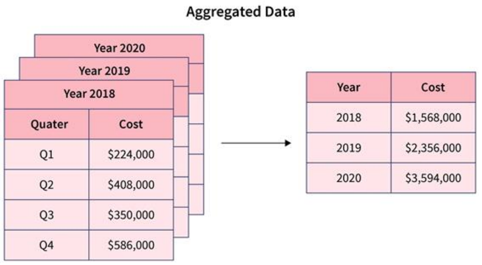
- Data wrangling, sometimes referred to as data munging, is the process of <u>*transforming and mapping data*</u> from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.



Data Wrangling tasks

Data Wrangling
DISCOVERING
STRUCTURING
CLEANING
ENRICHING
VALIDATING
PUBLISHING

1. **Discovering**
   - Discovery refers to the process of <u>*familiarizing yourself with data*</u> so you can conceptualize how you might use it.
   - During discovery, you may identify trends or patterns in the data, along with obvious issues, such as missing or incomplete values that need to be addressed.
   - This is an important step, as it will inform every activity that comes afterward.
2. **Structuring**
   - The data structuring step, sometimes called data transformation, focuses on <u>*organizing the data into a unified format*</u> so that it is suitable for analysis.
   - **Aggregation**: Combining rows of data by using summary statistics and grouping data based on certain variables.

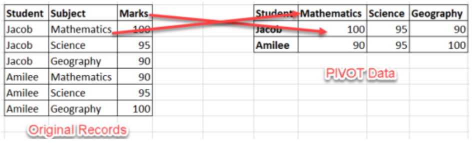

Aggregated Data

Year 2018	
Quater	Cost
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year 2020
Year 2019

Year	Cost
2018	$1,568,000
2019	$2,356,000
2020	$3,594,000

   - **Pivoting**: Shifting data between rows and columns or transforming data into other formats to prepare it for use.



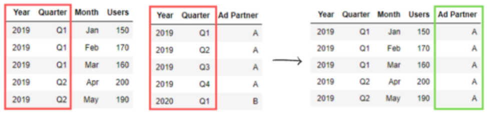
Student	Subject	Marks
Jacob	Mathematics	100
Jacob	Science	95
Jacob	Geography	90
Amilee	Mathematics	90
Amilee	Science	95
Amilee	Geography	100

Original Records

Student	Mathematics	Science	Geography
Jacob	100	95	90
Amilee	90	95	100

PIVOT Data

   - **Joining**: Combining data from multiple tables and combining related information from disparate sources.

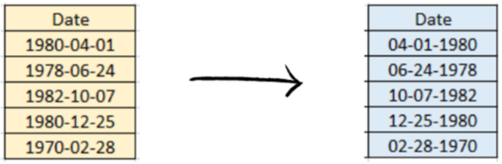


Year	Quarter	Month	Users
2019	Q1	Jan	150
2019	Q1	Feb	170
2019	Q1	Mar	160
2019	Q2	Apr	200
2019	Q2	May	190

Year	Quarter	Ad Partner
2019	Q1	A
2019	Q2	A
2019	Q3	A
2019	Q4	A
2020	Q1	B

Year	Quarter	Month	Users	Ad Partner
2019	Q1	Jan	150	A
2019	Q1	Feb	170	A
2019	Q1	Mar	160	A
2019	Q2	Apr	200	A
2019	Q2	May	190	A

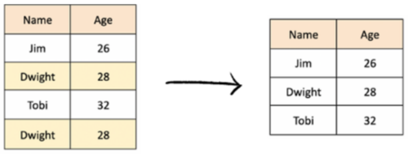
   - **Data type conversion**: Changing the data type of a variable to aid in performing calculations and applying statistical methods.



Date
1980-04-01
1978-06-24
1982-10-07
1980-12-25
1970-02-28

Date
04-01-1980
06-24-1978
10-07-1982
12-25-1980
02-28-1970

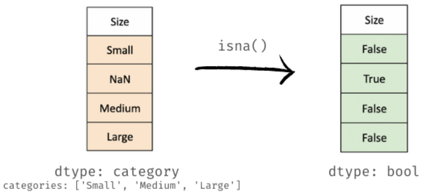
3. **Cleaning**
   - Data cleaning involves handling <u>*missing values, removing duplicates*</u> and correcting errors or inconsistencies.



Name	Age
Jim	26
Dwight	28
Tobi	32
Dwight	28

Name	Age
Jim	26
Dwight	28
Tobi	32

Remove Duplicate Rows



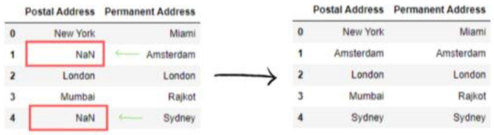
Size
Small
NaN
Medium
Large

Size
False
True
False
False

```
dtype: category
categories: ['Small', 'Medium', 'Large']
```

```
dtype: bool
```

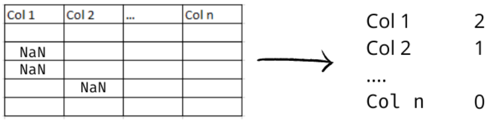
Missing Values in Category Column



	Postal Address	Permanent Address
0	New York	Miami
1	NaN	Amsterdam
2	London	London
3	Mumbai	Rajkot
4	NaN	Sydney

	Postal Address	Permanent Address
0	New York	Miami
1	Amsterdam	Amsterdam
2	London	London
3	Mumbai	Rajkot
4	Sydney	Sydney

Fillna with another column values



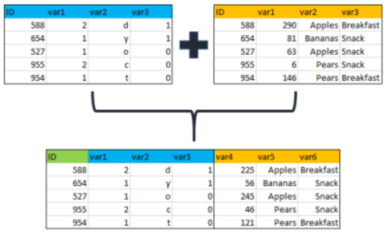
Col 1	Col 2	...	Col n
NaN			
NaN			
	NaN		

Col 1	2
Col 2	1
....	
Col n	0

Missing Values Count

4. **Enriching**

- Data enrichment involves adding new information to existing data sets to enhance their value. Sometimes called *data augmentation*
- It involves assessing what additional information is necessary and where it might come from.
- Then, the additional information must be integrated with the existing data set and cleaned in the same ways as the original data.
- *Merging data* from multiple sources to develop a more comprehensive dataset.



ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

ID	var1	var2	var3
588	290	Apples	Breakfast
654	81	Bananas	Snack
527	63	Apples	Snack
955	6	Pears	Snack
954	146	Pears	Breakfast

ID	var1	var2	var3	var4	var5	var6
588	2	d	1	225	Apples	Breakfast
654	1	y	1	56	Bananas	Snack
527	1	o	0	245	Apples	Snack
955	2	c	0	46	Pears	Snack
954	1	t	0	121	Pears	Breakfast

- *Creating new features* from existing data that can provide additional insights when analyzed.
- *Feature creation* involves generating new features from domain knowledge or by observing patterns in the data.
- It can be:

  Domain-specific: Created based on industry knowledge like business rules.

  Data-driven: Derived by recognizing patterns in data.

  Synthetic: Formed by combining existing features.

5. **Validating**

- This step involves verifying the accuracy and consistency of the wrangled data. First, validation rules must be established based on business logic, data constraints and other issues.
  - ✓ Data type validation: Helping ensure correct data types.
  - ✓ Range or format checks: To verify values fall within acceptable ranges and adhere to certain formats.
  - ✓ Consistency checks: Making sure that there is a logical agreement between related variables.
  - ✓ Uniqueness checks: Confirming that certain variables (such as customer or product ID numbers) have unique values.
  - ✓ Cross-field validation: Checking for logical relationships between variables (for example, age and birthdate).
  - ✓ Statistical analysis: Identifying outliers or anomalies by using descriptive statistics and visualizations.

6. **Publishing**

- Once your data has been validated, you can publish it. This involves making it available to others within your organization for analysis.