510 DATA SCIENCE

Lecture 03

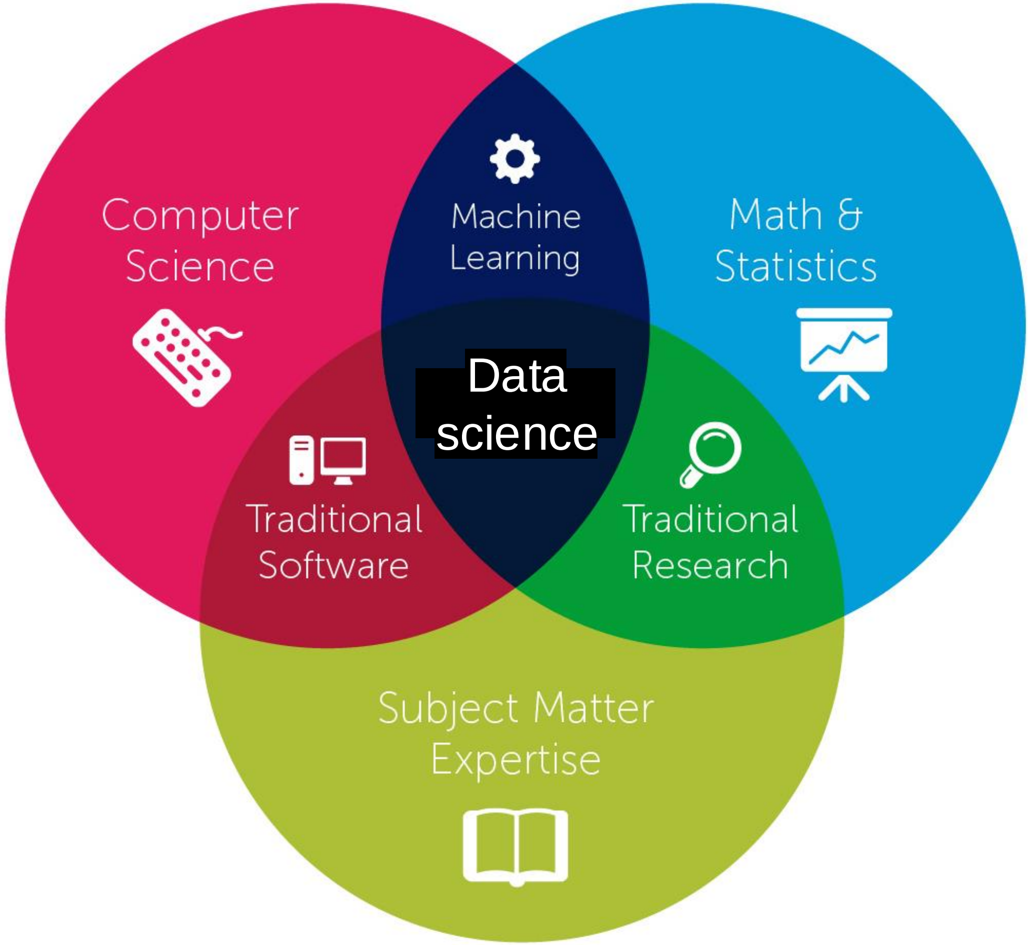
Fall 2024

Instructor: Assoc. Prof. Şener Özönder

Email: sener.ozonder@bogazici.edu.tr

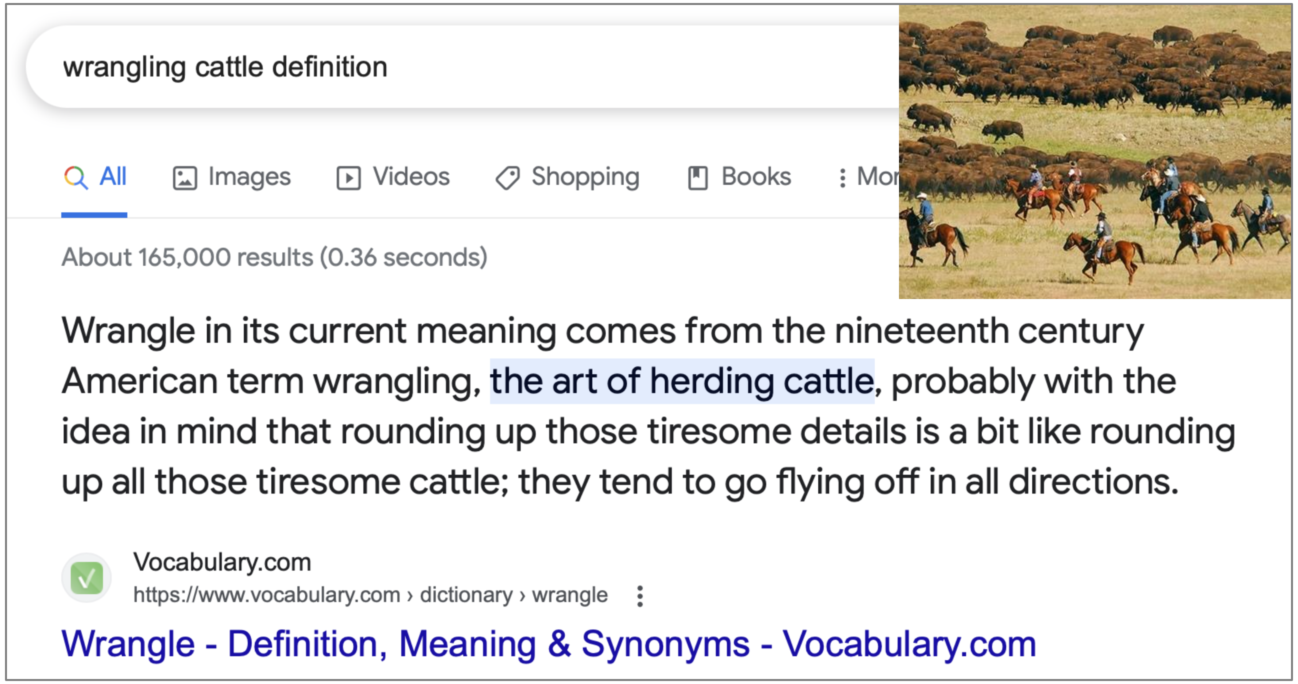
Institute for Data Science & Artificial Intelligence

Boğaziçi University



Data wrangling (data munging)

- Real-world data often comes with challenges such as being dirty, disorganized, scattered, and containing missing or incorrect entries. Data wrangling addresses these issues, serving as both an art and a science to prepare data for analysis, modeling, and visualization.
- The word “wrangling” is about cattle herding:



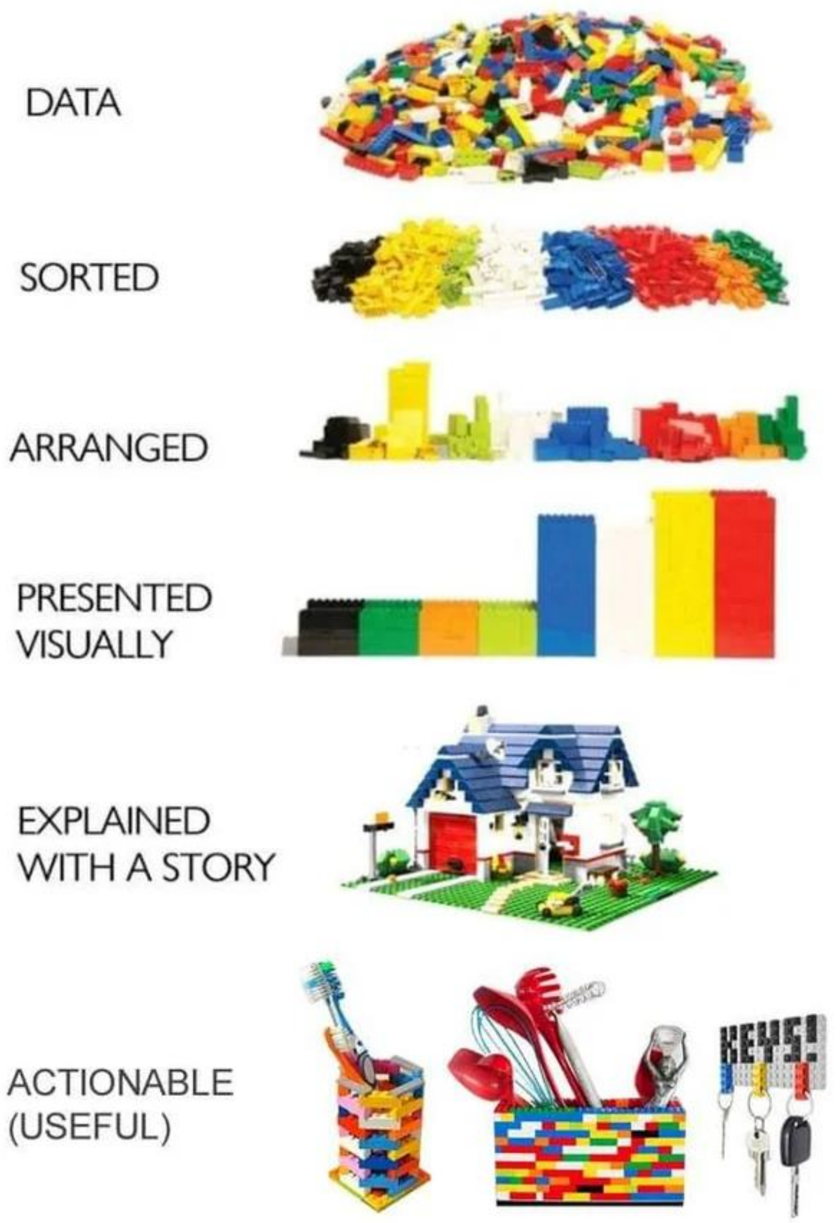
wrangling cattle definition

About 165,000 results (0.36 seconds)

Wrangle in its current meaning comes from the nineteenth century American term wrangling, the art of herding cattle, probably with the idea in mind that rounding up those tiresome details is a bit like rounding up all those tiresome cattle; they tend to go flying off in all directions.

Vocabulary.com
https://www.vocabulary.com › dictionary › wrangle

Wrangle - Definition, Meaning & Synonyms - Vocabulary.com



Data wrangling steps

1. Data Discovery: Understanding the nature of the data, its source, and its relevance.

2. Data Structuring: Structure data with reshaping data frames and converting data types.

3. Data Cleaning: Identifying and correcting errors, inconsistencies, and inaccuracies in data. This can include handling missing values, removing duplicates, and correcting typos.

4. Data Enrichment: Enhancing data with additional variables or attributes that can be derived from the existing dataset or by integrating with other data sources.

5. Data Validation: Ensuring that the dataset meets the required standards and quality benchmarks. This can involve setting up rules or constraints on the data.

6. Data Transformation: Converting data into a suitable format or structure for analysis. This can involve normalization, scaling, or encoding categorical variables.

7. Data Integration: Combining data from different sources and providing a unified view. This can involve tasks like database merging, concatenation, or joining tables based on common keys.

pandas functions

- For data wrangling, we’ll use pandas library in Python.

1. Data Import/Export:

- read_csv(): Read a comma-separated values (csv) file into DataFrame.
- read_excel(): Read an Excel file into DataFrame.
- read_sql(): Read SQL query or database table into DataFrame.
- read_json(): Read a JSON string/file into DataFrame.
- to_csv(): Write DataFrame to a CSV file.
- to_excel(): Write DataFrame to an Excel file.
- to_sql(): Write DataFrame to a SQL database.
- to_json(): Convert DataFrame to JSON format.

2. Data Inspection:

- head(): Return the first n rows of the DataFrame.
- tail(): Return the last n rows of the DataFrame.
- info(): Provide a concise summary of the DataFrame's columns, data types, and non-null values.
- describe(): Generate descriptive statistics of the DataFrame's columns.
- shape: Return the dimensions of the DataFrame (rows, columns).
- dtypes: Return the data types of each column.

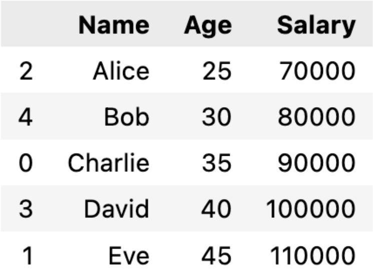
pandas functions

3. Data Cleaning:

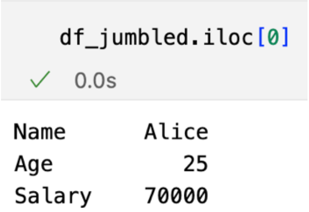
- dropna(): Remove missing values.
- fillna(): Fill missing values using specified method.
- replace(): Replace values in the DataFrame.
- drop(): Drop specified labels from rows or columns.
- rename(): Rename columns or index.
- astype(): Convert data type of one or more columns.

4. Data Filtering:

- loc[]: Access a group of rows and columns by labels.
- iloc[]: Access a group of rows and columns by integer location.
- query(): Query the DataFrame using a string expression.

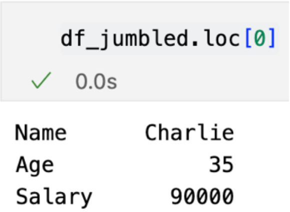


	Name	Age	Salary
2	Alice	25	70000
4	Bob	30	80000
0	Charlie	35	90000
3	David	40	100000
1	Eve	45	110000



```
df_jumbled.iloc[0]
✓ 0.0s

Name      Alice
Age          25
Salary    70000
```

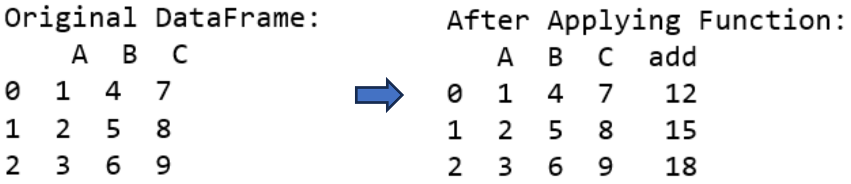


```
df_jumbled.loc[0]
✓ 0.0s

Name      Charlie
Age            35
Salary      90000
```

5. Data Transformation:

- apply(): Apply a function along an axis (row/column) of the DataFrame.
- map(): Map values of a Series using a function or dictionary.
- transform(): Transform the DataFrame using a function.
- cut(): Segment and sort data values into bins.
- qcut(): Quantile-based discretization function.



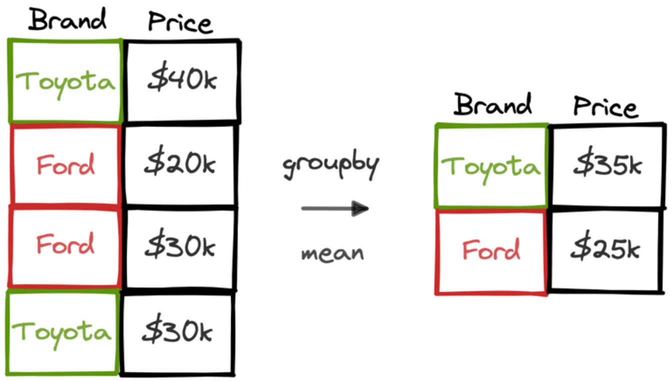
```
Original DataFrame:
   A  B  C
0  1  4  7
1  2  5  8
2  3  6  9
```

```
After Applying Function:
   A  B  C  add
0  1  4  7   12
1  2  5  8   15
2  3  6  9   18
```

pandas functions

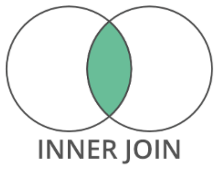
6. Data Aggregation:

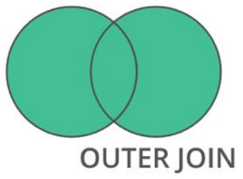
- groupby(): Group DataFrame using a column or columns.
- agg(): Aggregate data using one or more operations.
- pivot_table(): Create a spreadsheet-style pivot table.



7. Data Merging, Joining, and Concatenation:

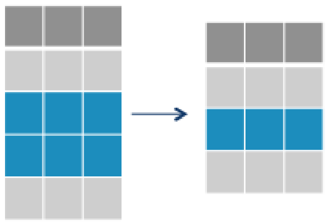
- merge(): Merge DataFrame objects by column or index.
- join(): Join columns of another DataFrame.
- concat(): Concatenate pandas objects along a specified axis.





8. Handling Duplicates:

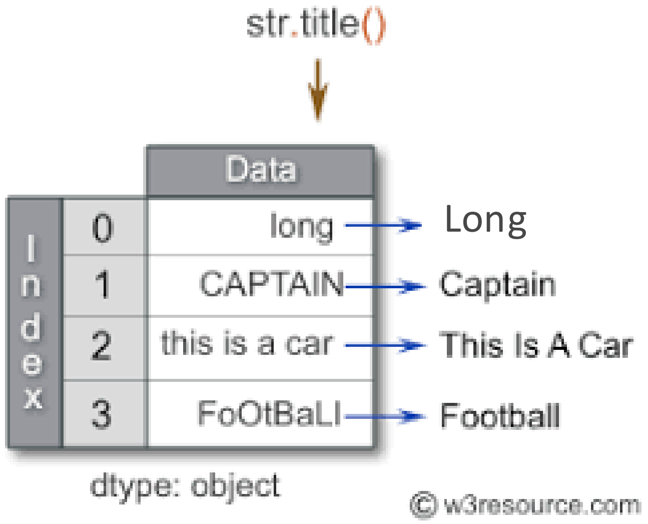
- duplicated(): Indicate duplicate rows.
- drop_duplicates(): Remove duplicate rows.



pandas functions

9. Handling Text Data:

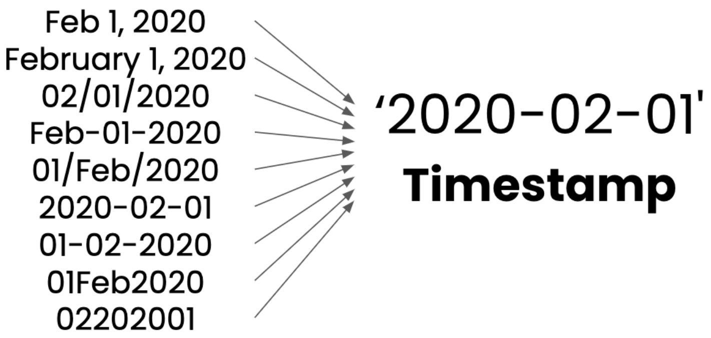
- str.split(): Split strings around given separator/delimiter.
- str.contains(): Check if string contains a pattern.
- str.replace(): Replace occurrences of a pattern.
- str.extract(): Extract groups from strings.
- str.cat(): Concatenate strings.
- str.lower(): Convert strings to lowercase.
- str.upper(): Convert strings to uppercase.
- str.strip(): Remove leading and trailing whitespace.



© w3resource.com

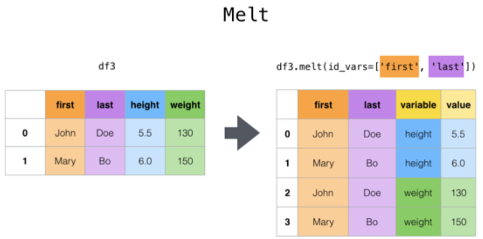
10. Handling Date and Time Data:

- to_datetime(): Convert argument to datetime format.
- date_range(): Create a fixed frequency DatetimeIndex.
- DatetimeIndex(): Immutable ndarray of datetime64 data.



11. Reshaping Data:

- pivot(): Reshape data based on column values.
- melt(): Unpivot a DataFrame from wide to long format.
- stack(): Stack a DataFrame or Series in multi-level columns.
- unstack(): Pivot a level of column labels.

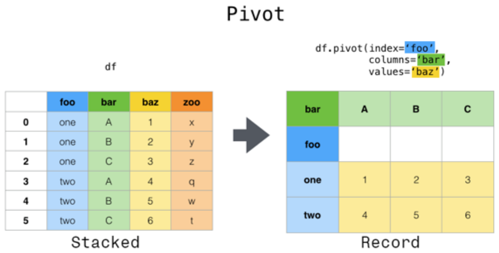

Melt

df3

	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

df3.melt(id_vars=['first', 'last'])

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150


Pivot

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

Stacked

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

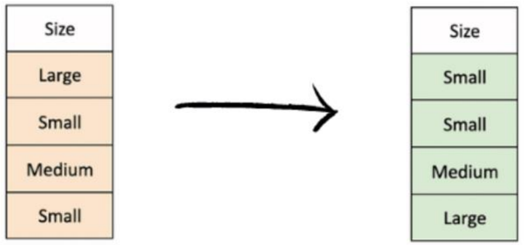
bar	A	B	C
foo			
one	1	2	3
two	4	5	6

Record

pandas functions

12. Sorting Data:

- sort_values(): Sort by values along either axis.
- sort_index(): Sort DataFrame by index.

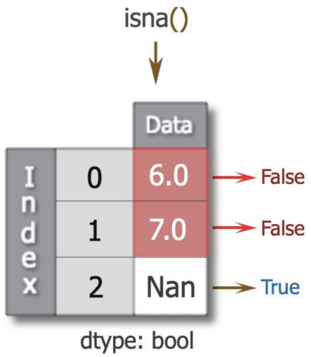


Size
Large
Small
Medium
Small

Size
Small
Small
Medium
Large

13. Data Validation:

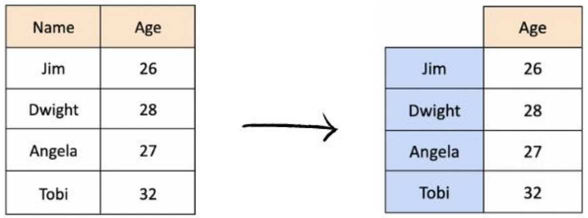
- isna(): Detect missing values.
- notna(): Detect non-missing values.
- isnull(): Alias for isna().
- notnull(): Alias for notna().



Index		Data
	0	6.0
	1	7.0
	2	Nan

14. Setting and Resetting Index:

- set_index(): Set the DataFrame index using existing columns.
- reset_index(): Reset the index of the DataFrame.

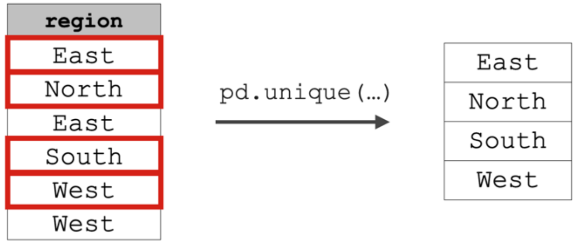


Name	Age
Jim	26
Dwight	28
Angela	27
Tobi	32

	Age
Jim	26
Dwight	28
Angela	27
Tobi	32

15. Unique Values and Value Counts:

- unique(): Find unique values in a Series.
- nunique(): Count distinct observations in a Series.
- value_counts(): Compute a histogram of a categorical variable.



region
East
North
East
South
West
West

East
North
South
West

Data validation

- Make sure to validate your data: Data validation is the process of ensuring that the data being input or processed adheres to a set of predefined criteria or rules.
- *Ensuring Data Consistency and Integrity:* If a dataset has a column for “Country” and a user inputs both “USA” and “U.S.A.”, it's inconsistent.
- *Validation Rules and Checks:*

  i. Range Checks: For example, the age of a person should be between 0 and 120.

  ii. List Checks: Checking data against a list of valid inputs. For instance, a “Sex” field might only accept “Male” or “Female”.

  iii. *Format Checks:* Ensuring data is in a specific format. For example, a date might need to be in the "YYYY-MM-DD" format.

  iv. *Uniqueness Checks:* For instance, in a database, each person might need to have a unique ID.

  v. *Consistency Checks:* For example, a person's birth date should be earlier than their date of employment.

- Continue with Lecture 03.ipynb