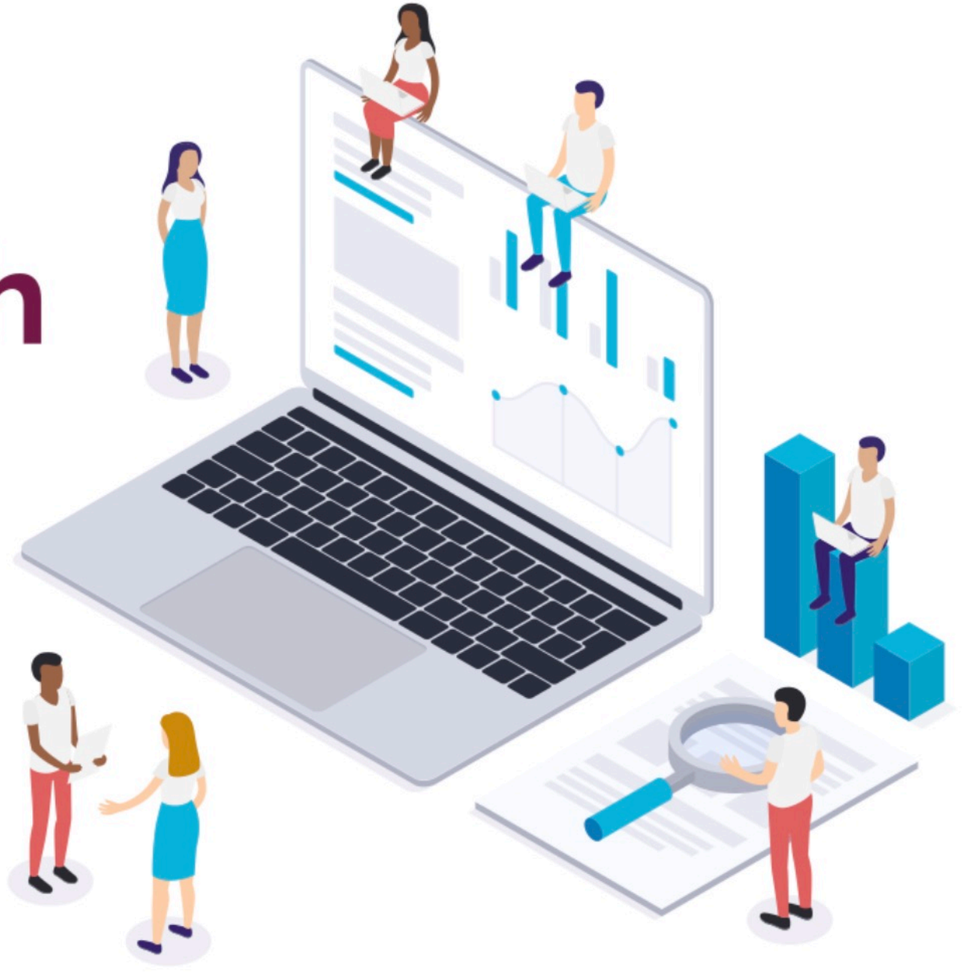
Data Analysis with Python

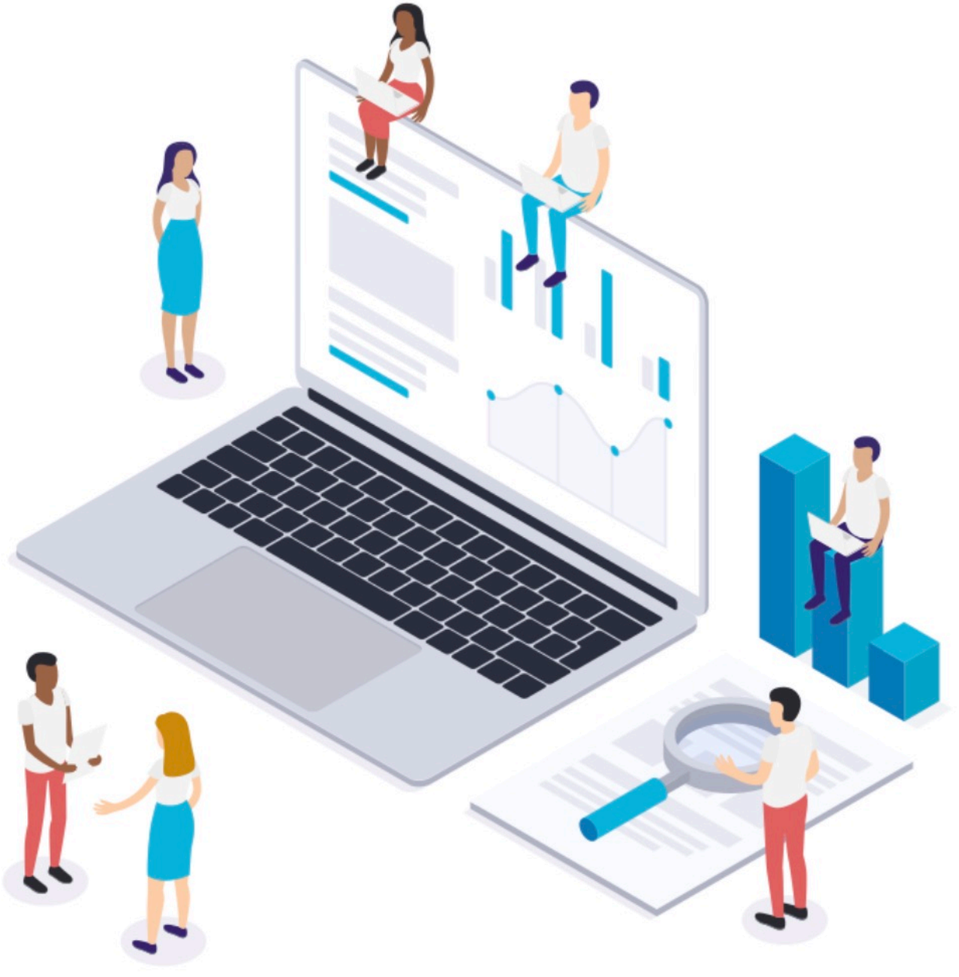
Session-7



CLARUSWAY©
WAY TO REINVENT YOURSELF

pandas

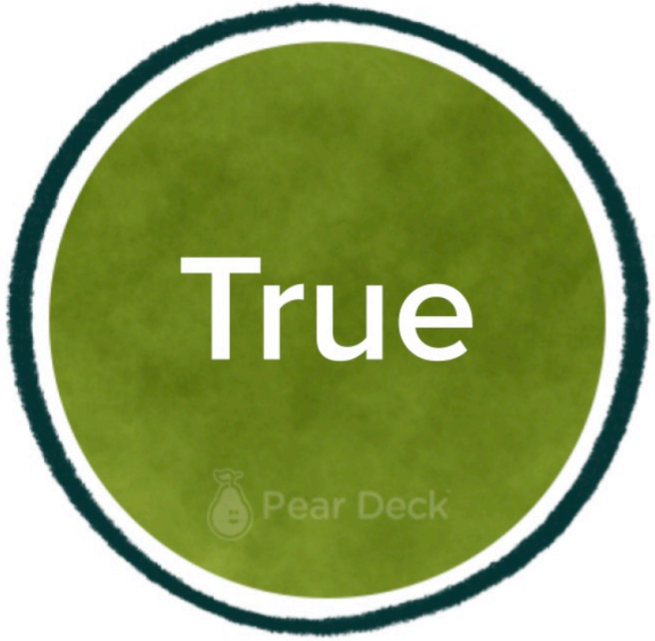
Missing Values

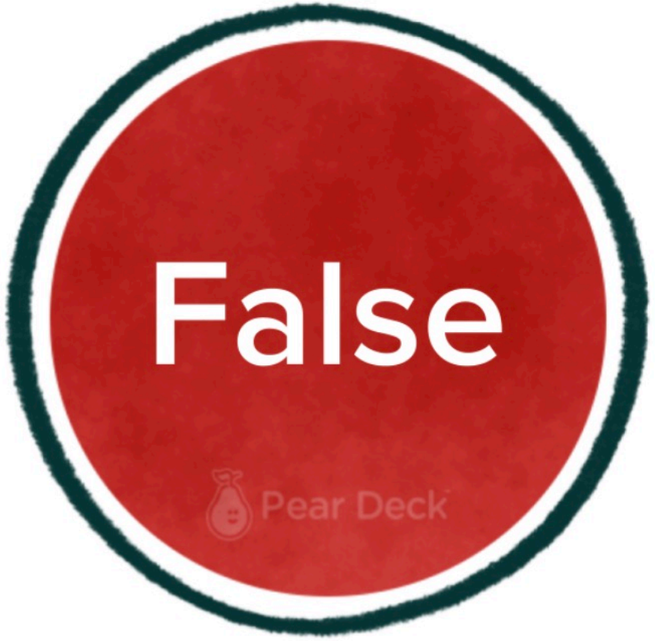


CLARUSWAY©
WAY TO REINVENT YOURSELF

Table of Contents

I've completed the **pre-class** content?





Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

What is Missing Value?

- Missing data is common in many different areas of data science and machine learning.

- Unfortunately, it can be challenging to handle effectively, and often there is no best solution.

What is Missing Value?

What does the **NaN** stand for?

Students choose an option

No Multiple Choice Response
You didn't answer this question

What is Missing Value?

- Used Audi A-3 Car prices (£) in UK

index	car_price
1	22.000
2	24.000
3	NaN
4	28.000
5	NaN

No value (`car_price`)

- **NaN** :Not a Number

Such values are called ***missing values***

What is Missing Value?

- Missing data occurs because of variety of reasons, including;
  - Manual data entry techniques,
  - Equipment faults,
  - Wrong measurements.

Types of Missing Values?

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)
- Structurally missing

Types of Missing Values

- Missing completely at random (MCAR)
  - Follow no discernable pattern
  - Cannot be predicted from the remaining know variables
  - Example; data generated explicitly at random or survey data using a random subset of questions from a pre-defined list.

Types of Missing Values

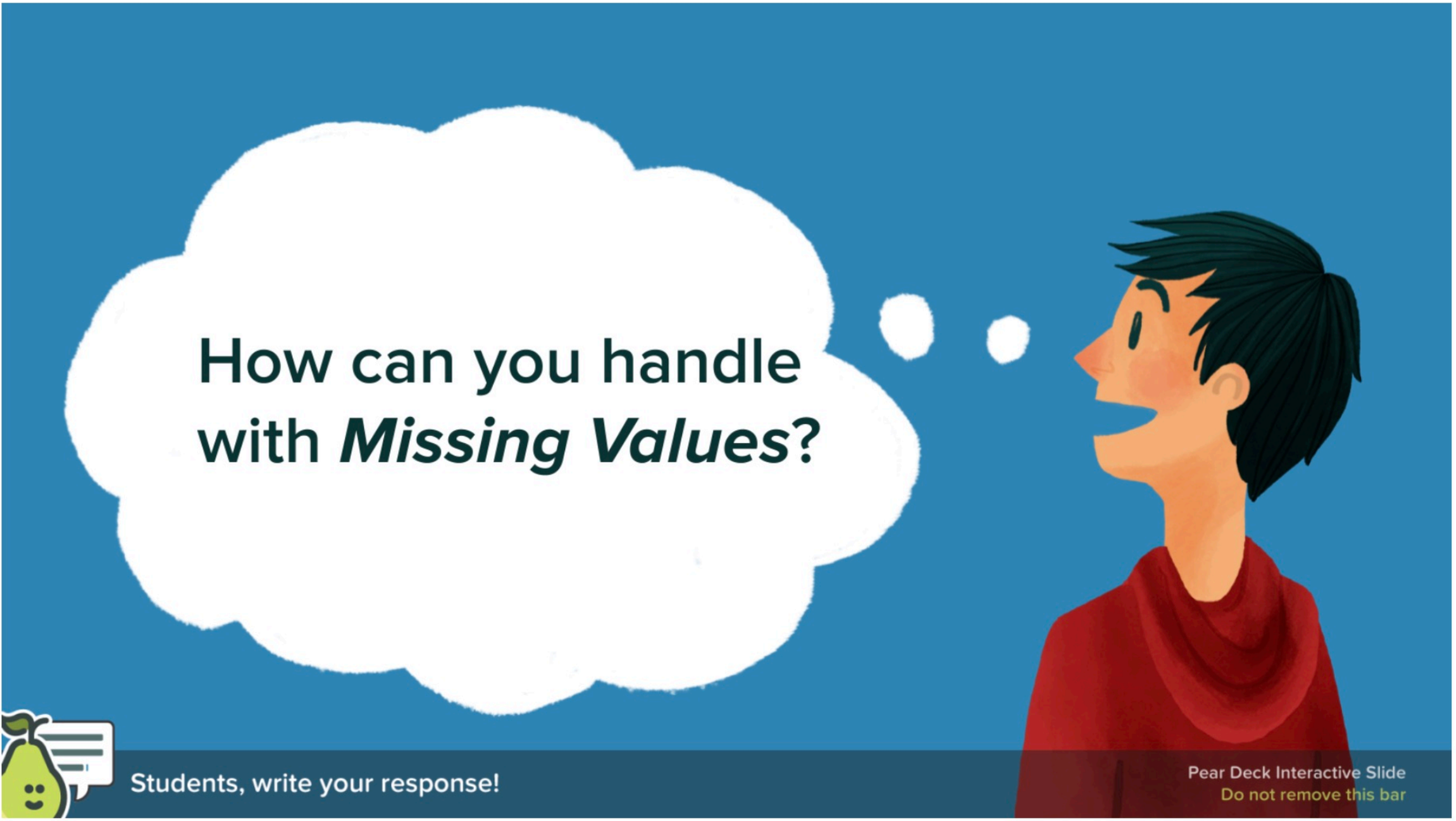
- Missing at random (MAR)
  - Errors with recording the data correctly
  - Can roughly be interpolated from the remaining values to a reasonable degree of accuracy.
  - Example; A sensor that misses a particular minute's measurement

Types of Missing Values

- Missing not at random (MNAR)
  - Why the data is missing is known
  - Can not effectively be inferred or predicted
  - Example; people in a certain age/income bracket refuse to answer how many vehicles or houses they own

Types of Missing Values

- Structurally missing
  - The missing data is missing for an apparent reason.
  - Mechanism that caused the missing data is easily inferred
  - Example; a survey that asks for income from employment would have missing values for those who do not have a job



No Text Response

You didn't answer this question

Handling with Missing Values

- Remove the missing data instances. (This method should be acceptable if there are few missing values and you have a lot of data.)
- Imputation methods. (This is a common approach it allows most models to function as usual without any modifications.)
- Keep the missing values and use a model which incorporates them. (This method limits the models available.)

Handling with Missing Values

Imputing Methods

Continuous Variable	Categorical Variable	Other Methods	Prediction of Missing Values	Using DL (Datawig)
mean median mode	mode	ffill bfill interpolate		

We'll focus on these methods

Handling with Missing Values

There are several methods for handling with Missing Values.

index	car_price	model
1	22.000	2012
2	18.000	2005
3	NaN	**2005**
4	28.000	2012
5	NaN	**2012**

We should consider the *group* (**`model`**) of the missing values (**`prices`**)

Handling with Missing Values

The most important point when handling with missing value

Some Useful Methods

- `isnull()`
- `isna()`
- `notnull()`
- `notna()`
- `drop()`
- `dropna()`
- `any()`
- `all()`
- `fillna()`
- `where()`
- `map()`
- `replace()`
- `interpolate()`

Draw lines to match the attributes/methods to their definitions:

`df.unique()`

`df.dropna()`

`df.value_counts()`

Used for imputing in a missing data.

Sort by the values along either axis.

Return unique values of Series object.

Return object with labels on given axis omitted where alternately any or all of the data are missing.

Count distinct observations over requested axis.

Apply a function along an axis of the DataFrame.

`df.apply()`

`df.sort_values()`

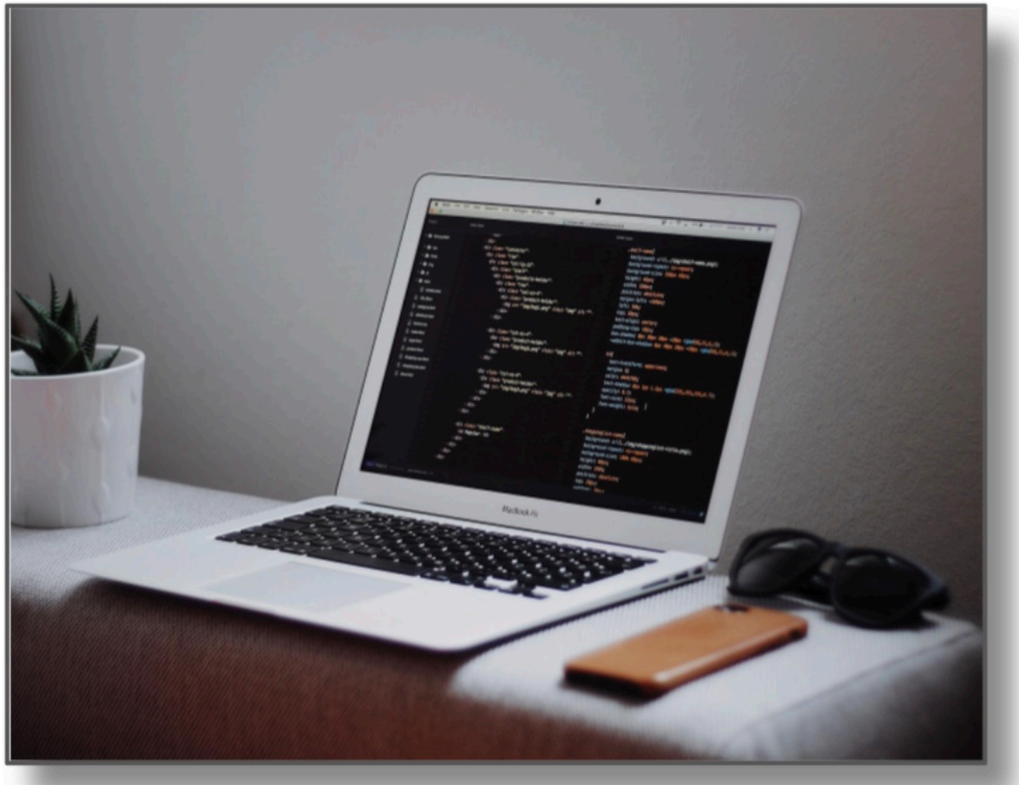
`df.fillna()`

Students, draw anywhere on this slide!

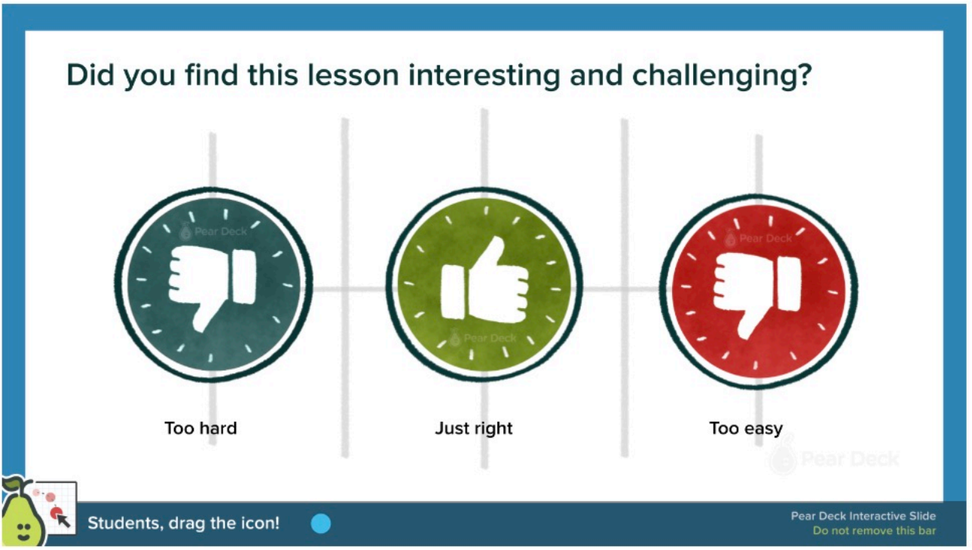
No Drawing Response

You didn't answer this question

Data Analysis with Python



let's start the hands-on phase



No Draggable™ Response

You didn't answer this question