Data Analysis with Python

Session-8

CLARUSWAY©
WAY TO REINVENT YOURSELF

pandas

Outliers

CLARUSWAY©
WAY TO REINVENT YOURSELF

Table of Contents

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

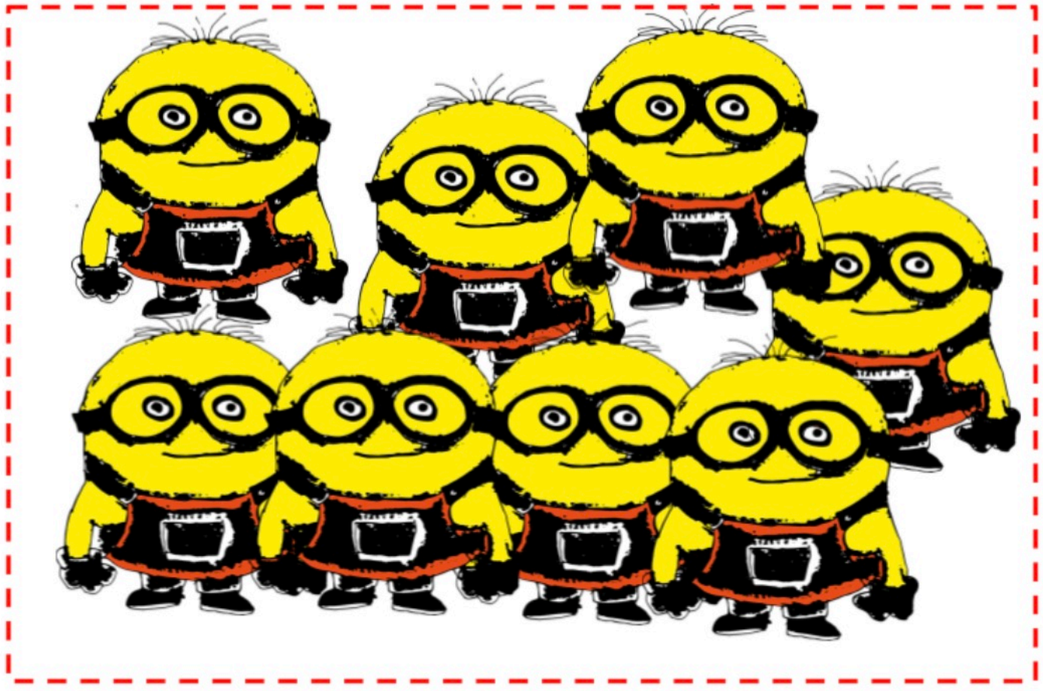
Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

What is the Outlier?

- Outliers can be unusually and extremely different from most of the data points existing in our sample.



What is the Outlier?

- Outliers can create biased results while calculating the stats of the data due to its extreme nature, thereby affecting further statistical/ML models.

Index	car_price
1	22.000
2	24.000
3	1050
4	28.000
5	149.000

The abnormal values of given variable (`car_price`)

Such values are called ***outliers***

What is the Outlier?

Causes of Outliers

- Data entries errors
- Measurement errors or instrument errors
- Sampling errors
- Data processing errors
- Natural novelties in data

What is the Outlier?

Types of Outliers

Univariate Outliers

- generally referred to as extreme points on **a variable**

Multivariate Outliers

- generally combination of unusual data points for **two or more variables**

An assumption of many multivariate statistical analysis, such as Multiple linear regression, is that there are no multivariate outliers.

Detecting Outliers

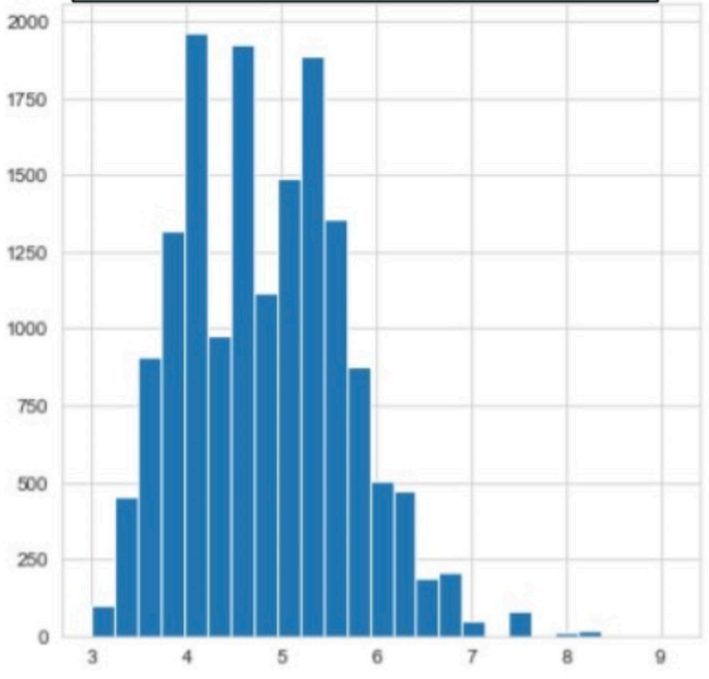
Methods for Detecting Outliers

Graphs	InterQuartile range (IQR) technique	Statistical Tests
▸ Scatter plot ▸ Box plot ▸ Histogram		▸ Grubbs' test ▸ Chi-square test ▸ Dixon's Q test

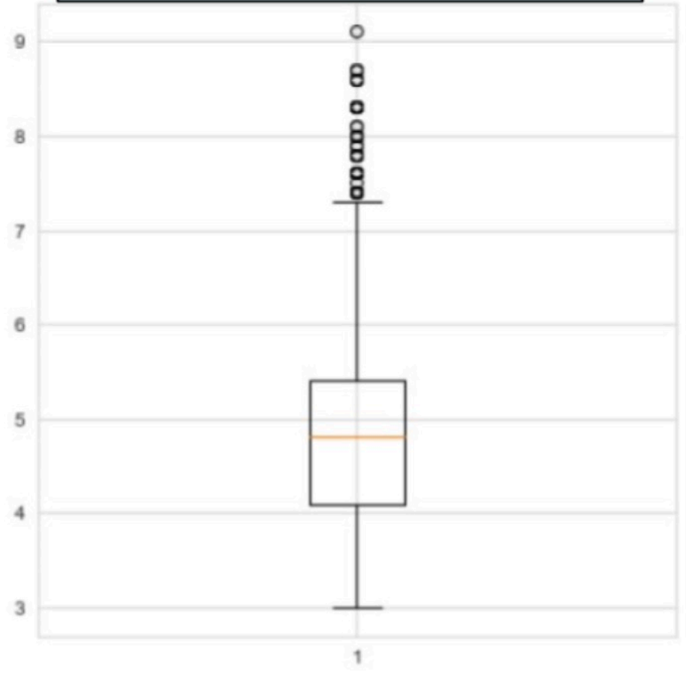
Detecting Outliers

Graphs

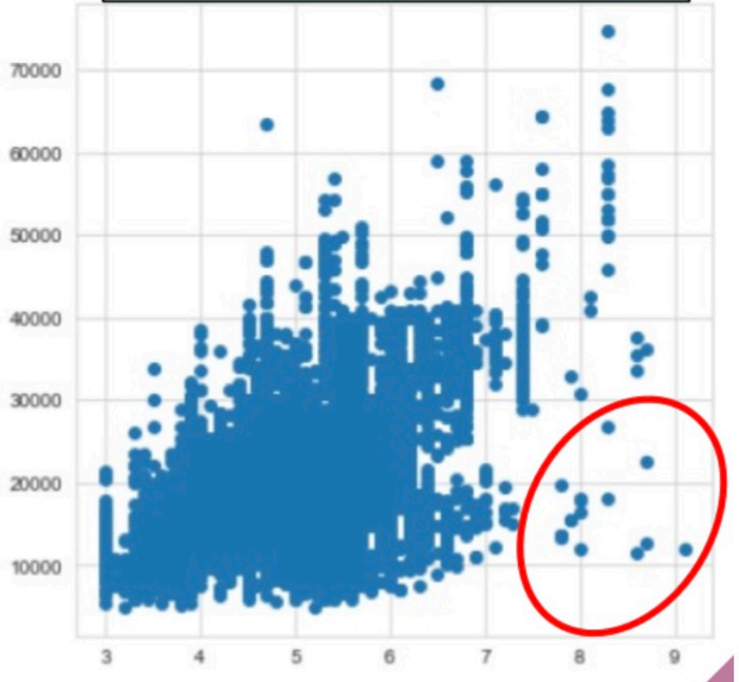
histogram



box plot

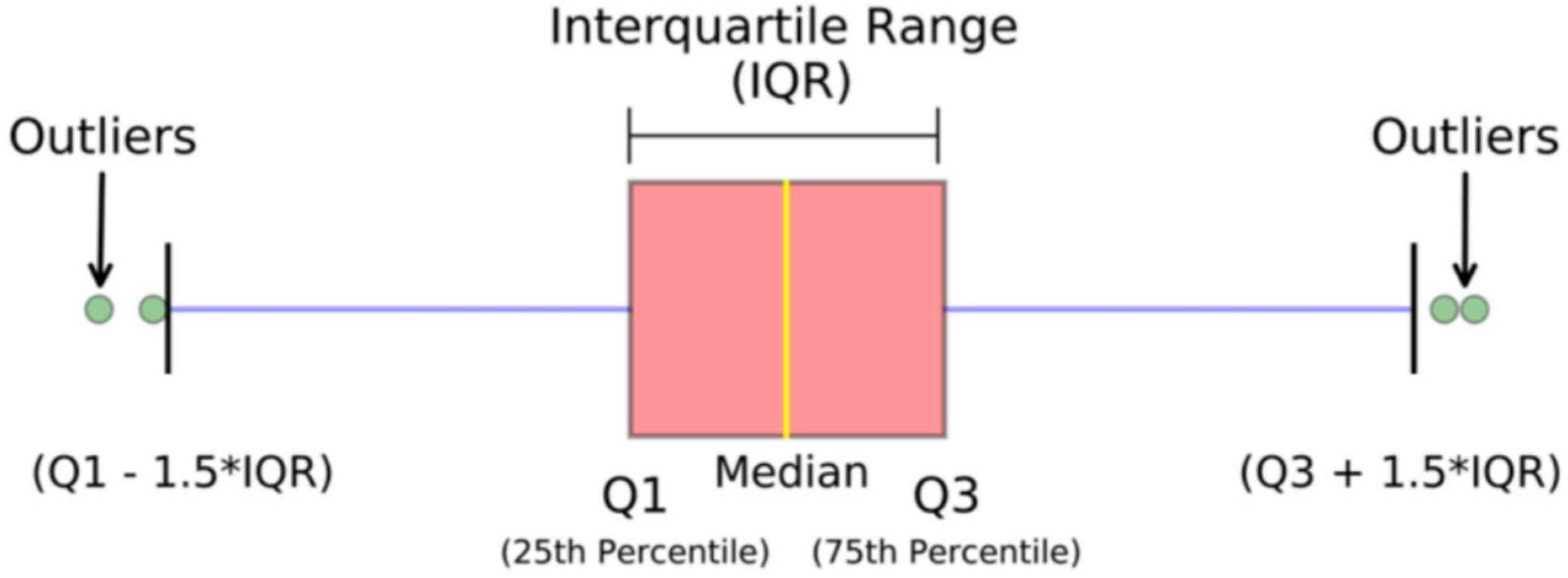


scatter plot



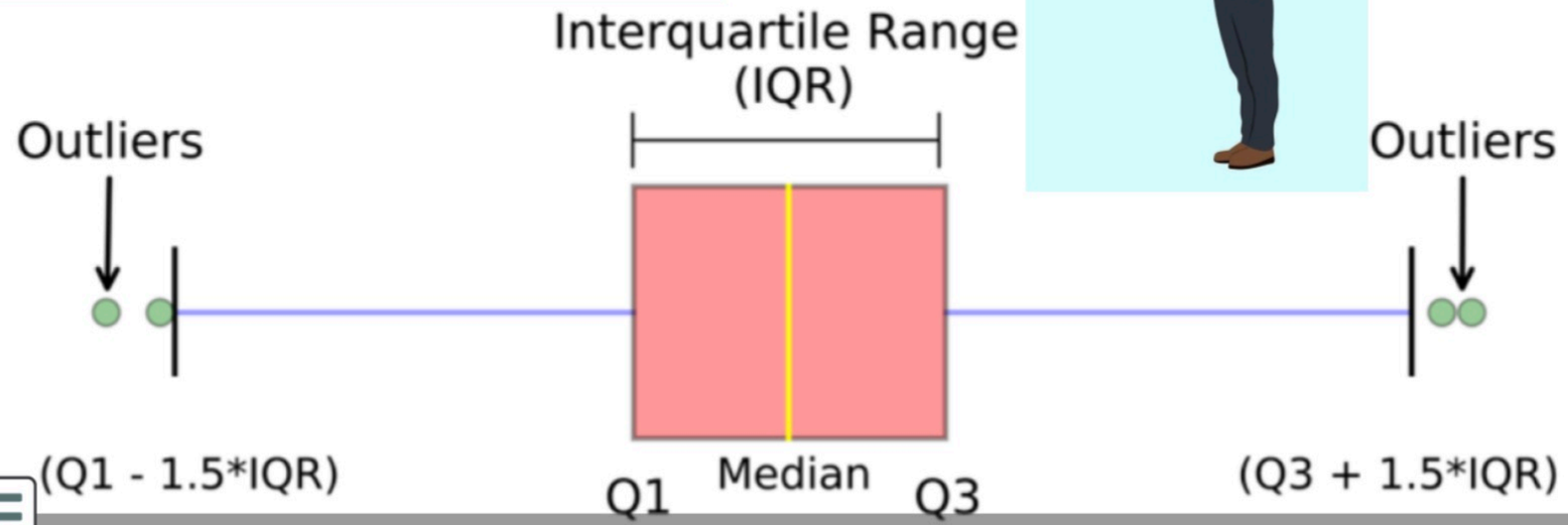
Detecting Outliers

InterQuartile range (IQR) technique



Handling with Outliers

How can you handle with the **outliers**?



Students, write your response!

No Text Response
You didn't answer this question

Handling with Outliers

Methods for Handling Outliers

- Removing the outliers.
- Limitation the outliers. (winsorize)
- Data transformation. (log, square root, exponentiating)
- Replacing the outliers. (mean, median, mode)
- Using different analysis methods. (statistical/nonparametric tests)
- Valuing the outliers. (valid reason for the outlier to exist)

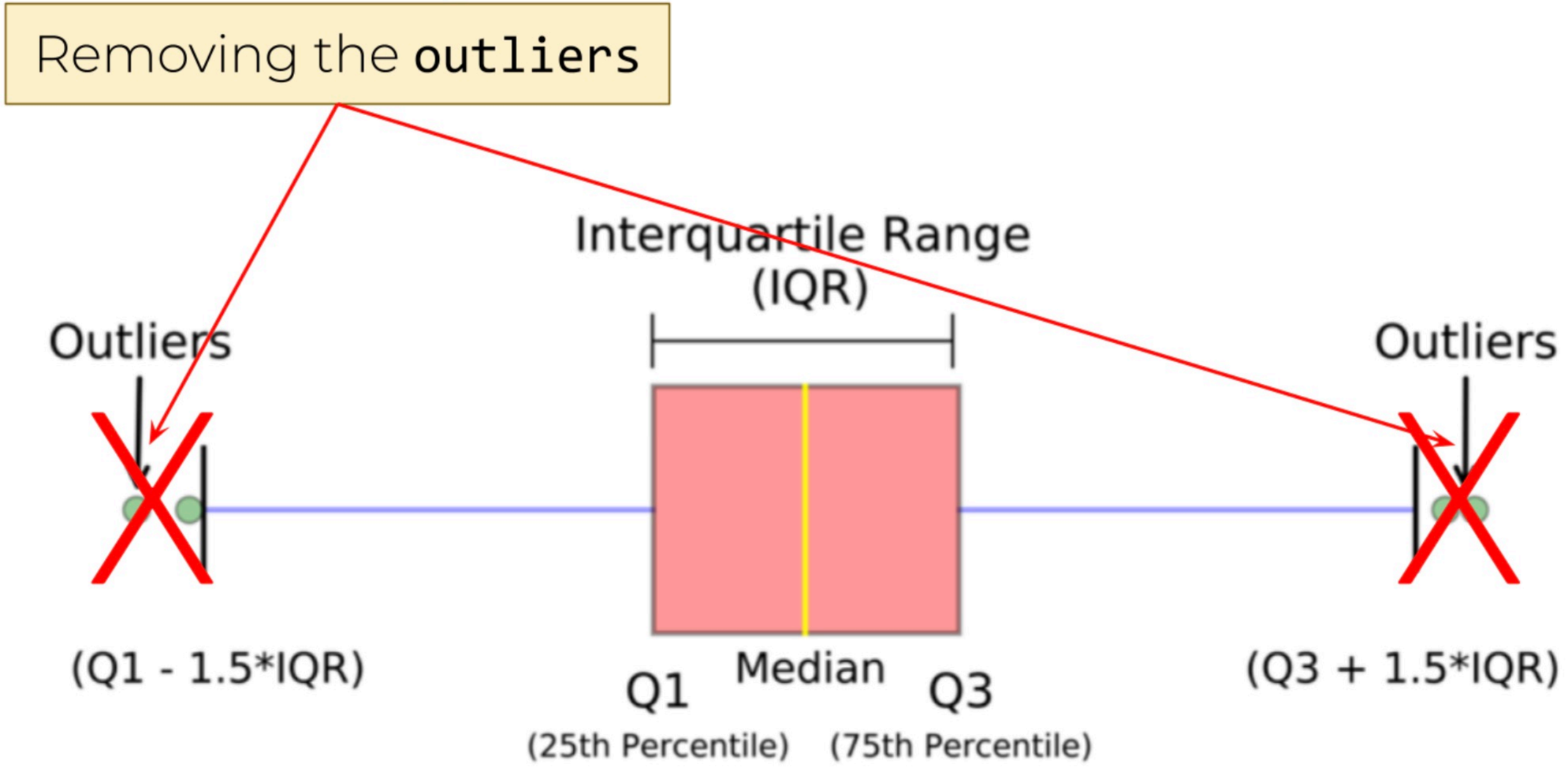
Handling with Outliers

Guideline for Handling Outliers

If the outlier in question is:

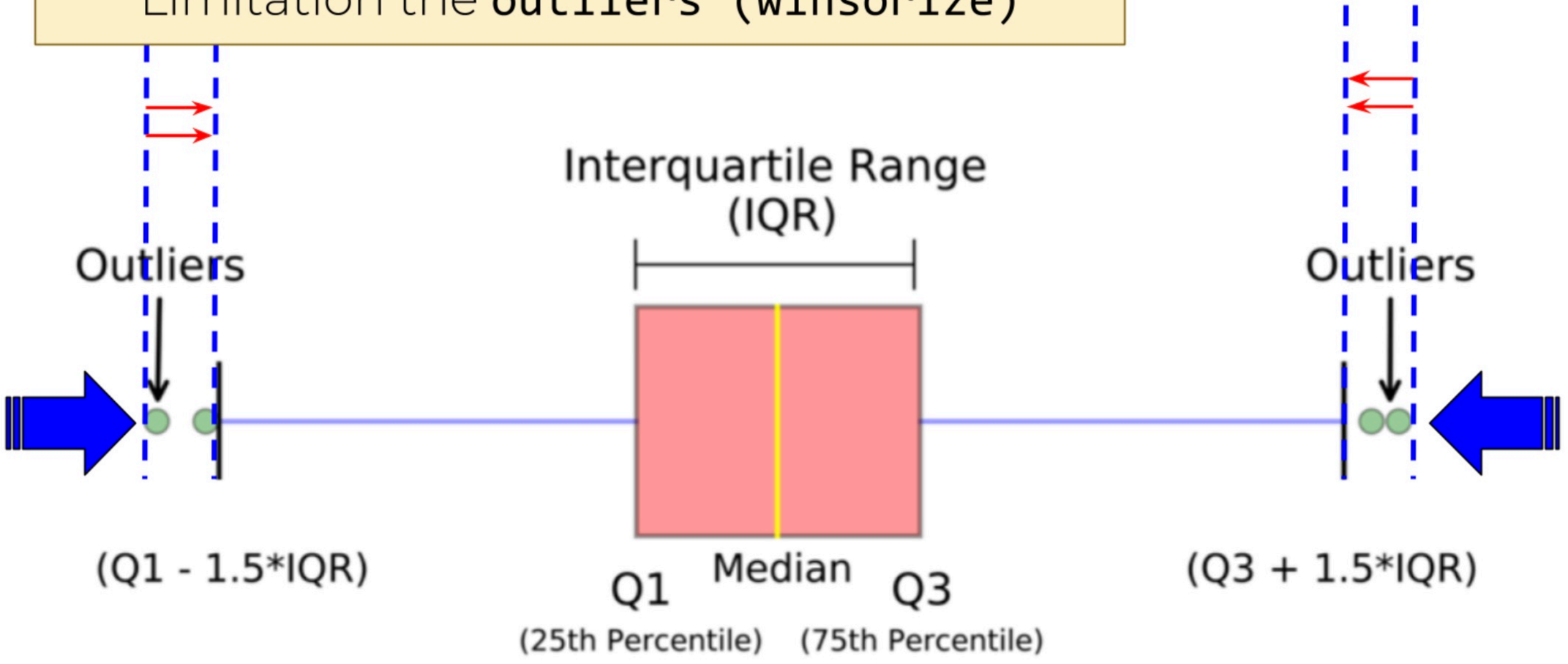
- A measurement error or data entry error, correct the error if possible. If you can't fix it, remove that observation because you know it's incorrect.
- Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.
- A natural part of the population you are studying, you should not remove it.

Handling with Outliers



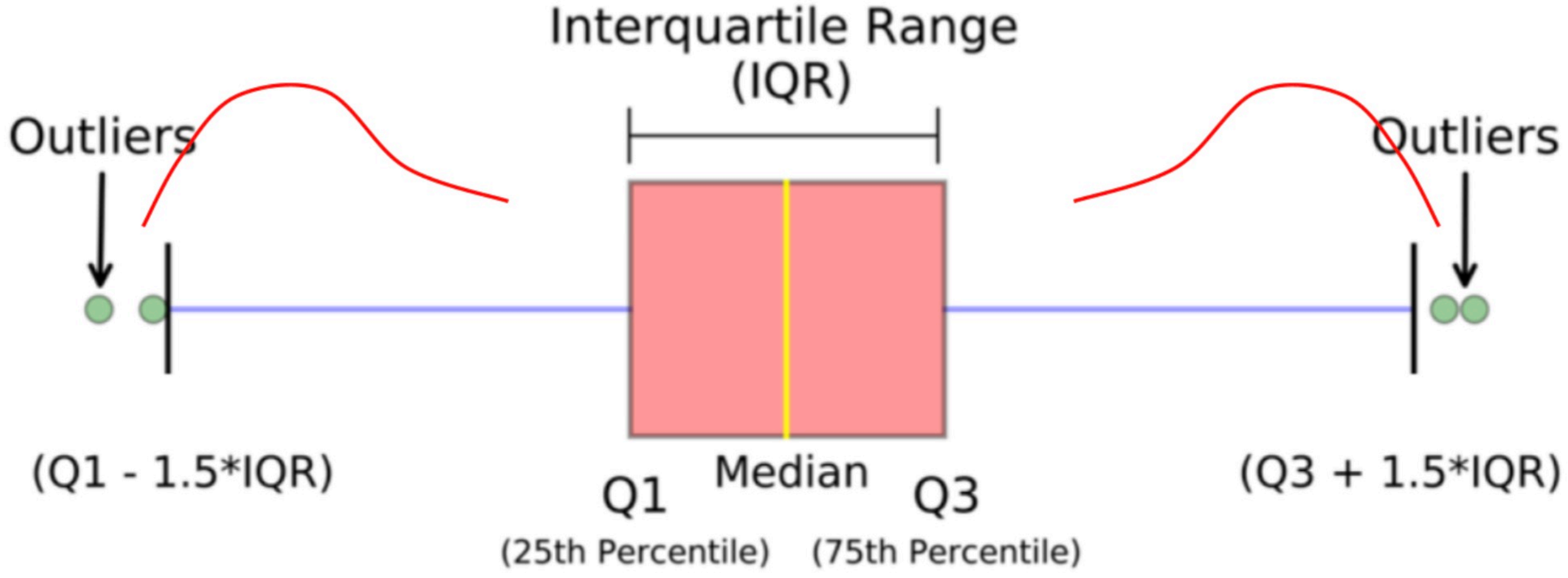
Handling with Outliers

Limitation the `outliers (winsorize)`



Handling with Outliers

Transformation of the **`outliers`**



Make connections

What are the advantages and disadvantages of **dropping** & **limiting** the outliers?

Students, write your response!

No Text Response

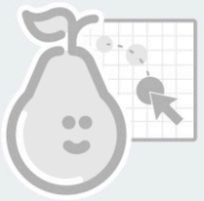
You didn't answer this question

Some Useful Methods

- `quantile()`
- `winsorize()`
- `log()`

Data Analysis with Python

let's start the hands-on phase



No Draggable™ Response

You didn't answer this question