# Data Preparation

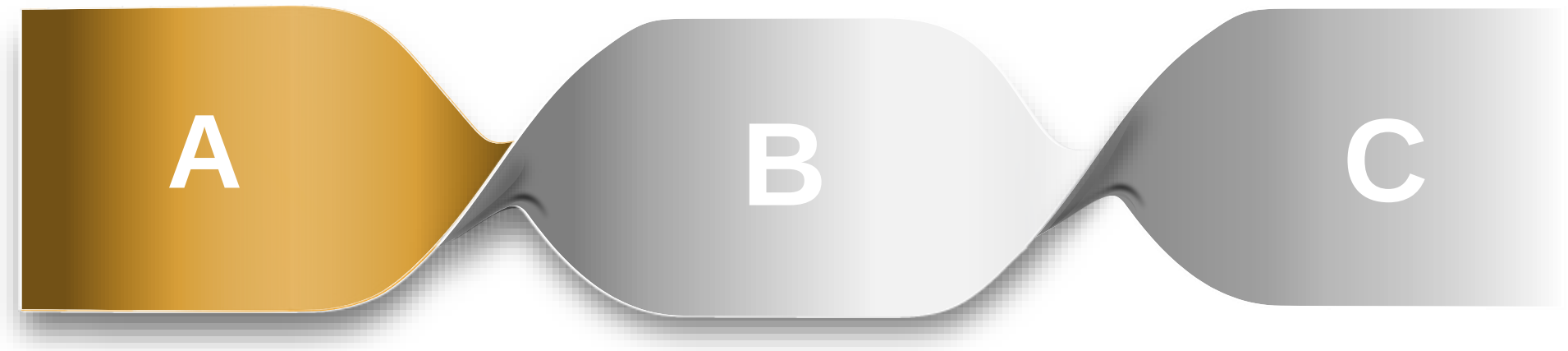
## Improve your data


**Standardization**

**Missing value imputation**

**Outlier Detection and treatment**

## Standardization



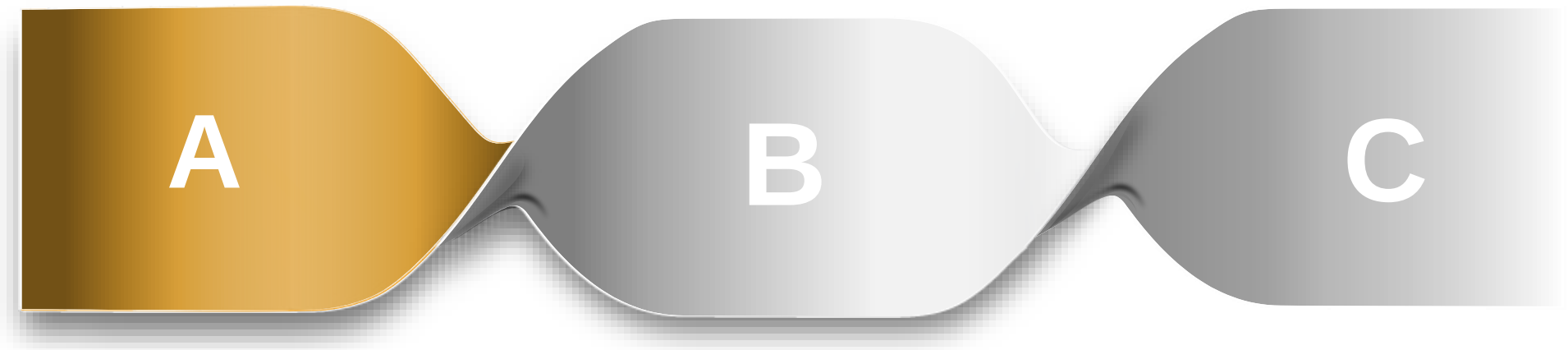

**Data Standardization-**

- Brings data into a **common format** across multiple system.
- Ensures **validation** of standard data elements such as if city, state, pin codes, street addresses are validated.
- Helps in **de-duplication and meaningful segmentation**

**Parsing-**

- It helps in data standardization
- Parsing process **splits** a customer data record into pre-defined data components so that comparisons can be correctly made with other records & external data sources.

# Standardization




Matthew Barn
8 Sheffield Appts.
Tel: 239838352
Email ID: matb@gamil.com

- Apartment is abbreviated
- Country code and area code not included in the telephone number
- Email is incorrect
- Street name and post code missing

Mr. Matthew Barn
8 Sheffield Apartments.
Palos Verdes Peninsula street, 90274
Tel: (+44) (0121) 239838352
Email ID- matb@gmail.com

- Apartment is corrected
- Country code and area code added
- Correct Email id
- Street name and post code are added

***PARSING***

| | |
|---|---|
| **Salutation** | Mr. |
| **First Name** | Matthew |
| **Middle Name** | NA |
| **Last Name** | Barn |
| **House Number** | 8 |
| **House Name** | Sheffield Apartments |
| **Street Name** | Palos Verdes Peninsula street |
| **Country Code** | +44 |
| **Area Code** | 0121 |
| **Telephone No.** | 239838352 |
| **Postcode** | 90274 |

# Missing value imputation


## Kinds of missing data:

- **Missing completely at random (MCAR):** Missing ness is nothing to do with the person being studied.
  E.g.- A questionnaire might be lost in the post, or a blood sample might be damaged in the lab

- **Missing at random (MAR):** When the missing ness is not random, but where missing ness can be fully accounted for by variables where there is complete information.
  E.g. Males are less likely to fill in a depression survey but this has nothing to do with their level of depression

- **Missing not at random (MNAR):** Not MCAR or MAR

Mean Imputation

Mean imputation= 20.2

Median Imputation

Median imputation= 19

| Age |
|---|
| 23 |
| |
| 19 |
| 32 |
| 12 |
| 15 |

**Numerical values -> Mean, Median substitution**

**Categorical values -> Mode substitution**

Mode Imputation

Mode imputation= M

| Sex |
|---|
| M |
| M |
| |

# Missing value imputation


## Methods of Deletion

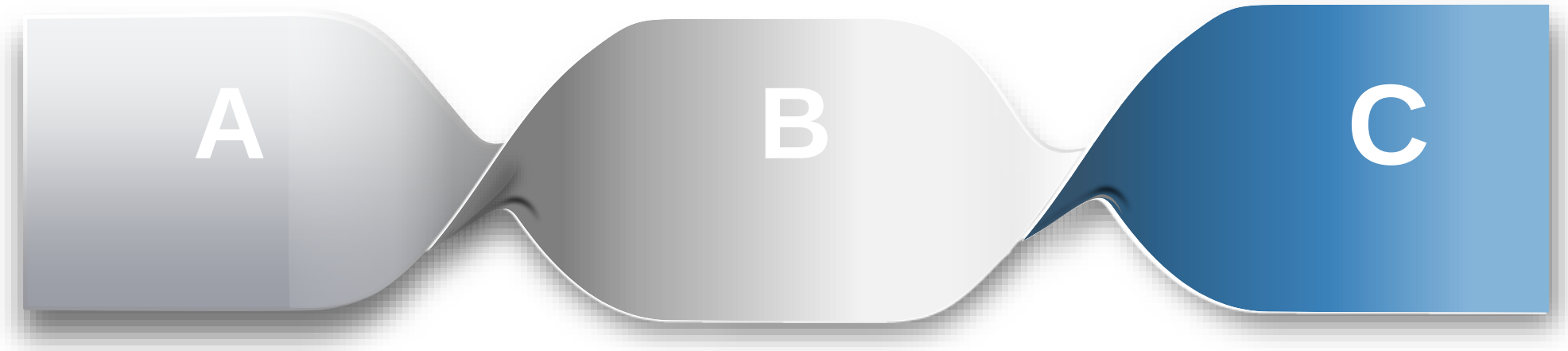
### List wise deletion

It removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations

### Dropping the value

If a variable has only 1% data, it is better to drop it then to use any imputation techniques

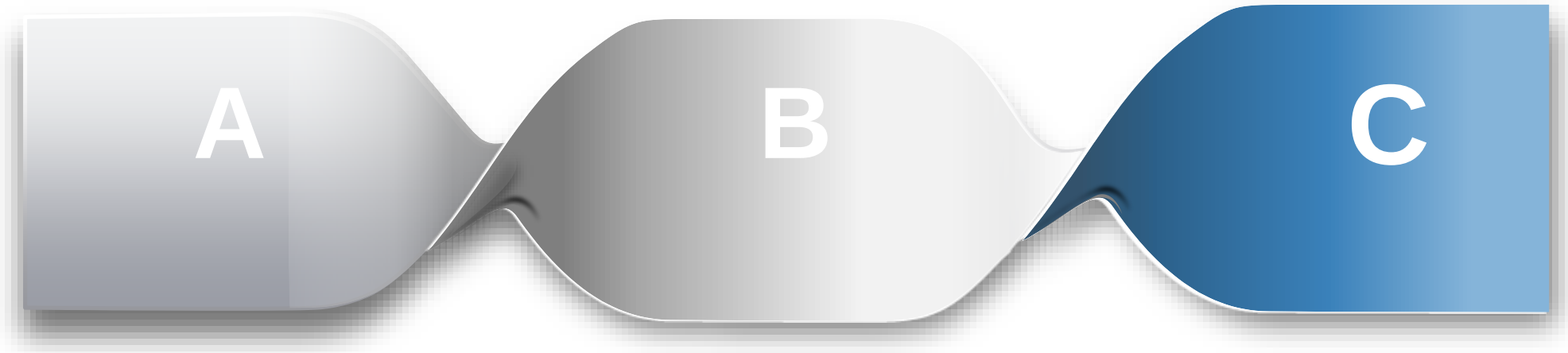
# Outlier detection and treatment




An outlier or extreme value is very large or very small compared with the rest of the data.

An example is if you were looking at the income of people in a training class and Bill Gates took the class. His income would be an outlier compared to the rest of the class.

# Outlier detection and treatment

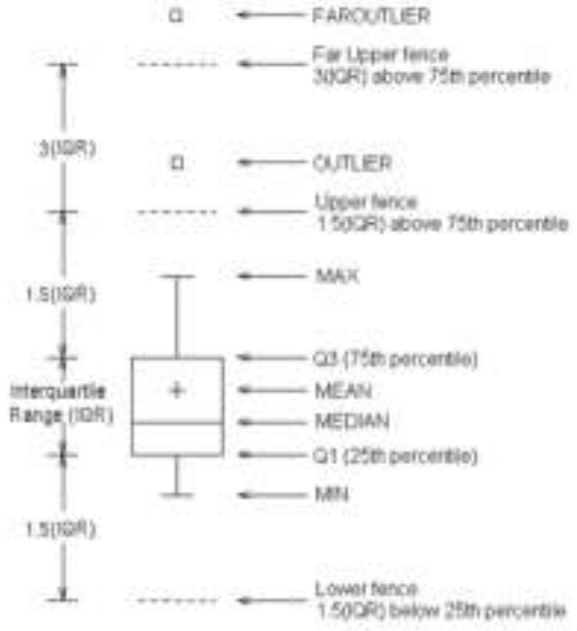



**APPROACH**

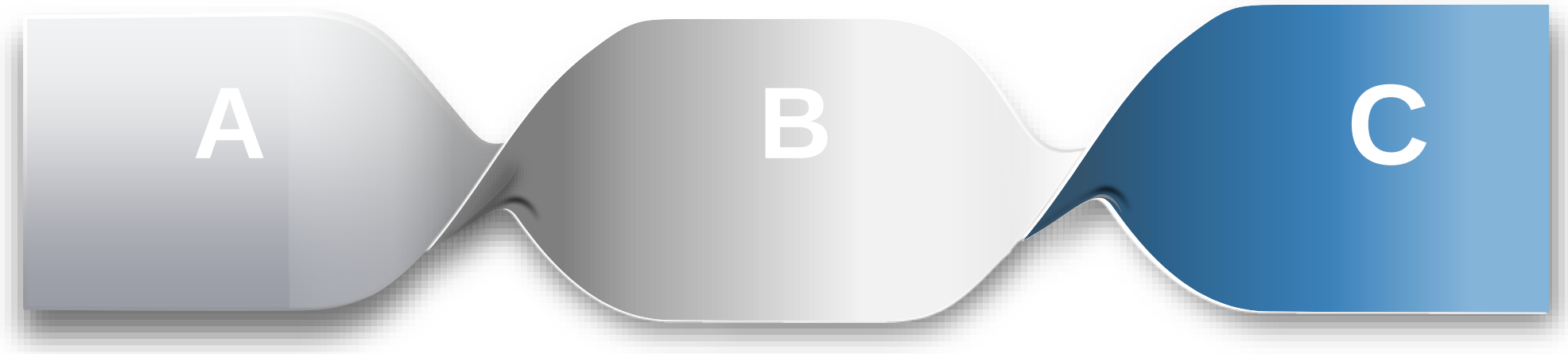
- Calculate Q1 (25%ILE) , Q3 (75%ILE) , IQR (Q3-Q1)
- Calculate inner & outer fences. Anything outside the fences are outliers
- Lower inner fence = Q1 – 1.5*IQR
- Lower outer fence = Q1 – 3*IQR
- Upper inner fence = Q3 + 1.5*IQR
- Upper outer fence = Q3 + 3*IQR
- In normally distributed data, you'd see about 1 in every 100 points outside the inner fence, but only 1 in every 500,000 points outside the outer fence

**LIMITATIONS**

- Will work only if Normally distributed data, 1-dimensional data



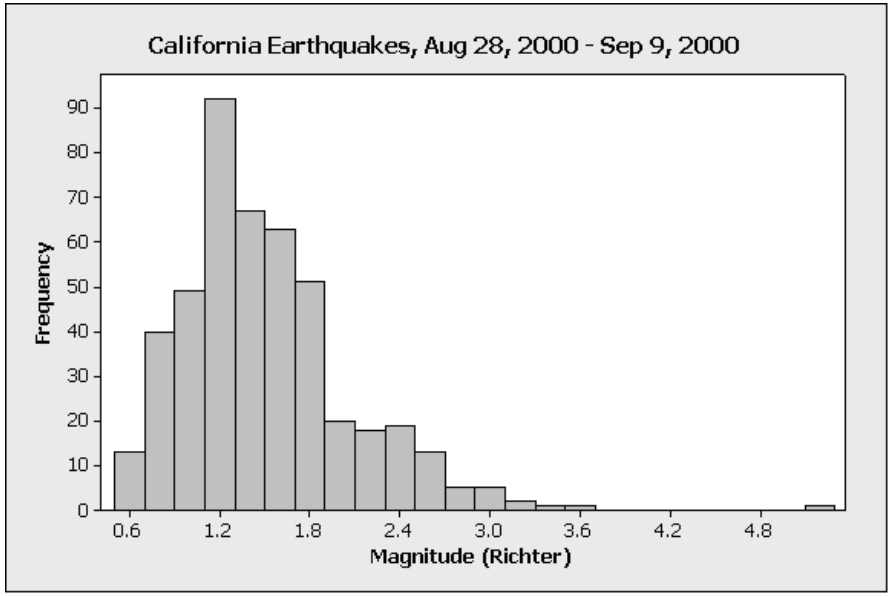
# Outlier detection and treatment




**Just looking at the histogram will help you realize the outliers (have you noticed the long tail, it can happen on either side)**

- Another simple classical approach to screen outliers is to use the Mean & SD (Standard Deviation) method.
- Calculate the mean & SD.

Any value less than Mean - 3*SD
OR
Any value more than Mean + 3*SD
Are considered as outliers



# Thanks!

***Any questions ?***

Next steps

- Attempt quiz -
- Share feedback -