



DATA PREPARATION 101

FOR MACHINE LEARNING MODEL BUILDING

LIVE SESSION ON 5TH DECEMBER 2020

AS A PART OF DPHI BOOTCAMP

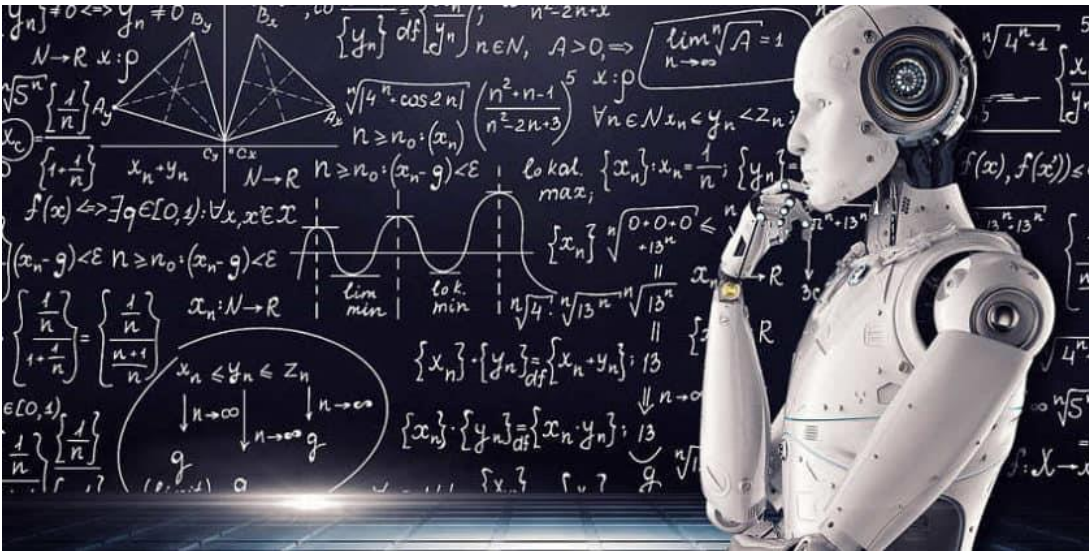


www.linkedin.com/in/aarthikumar

- Accountant + Data Scientist
- 10+ years of experience in DS
- Lead Data Scientist, Allianz Benelux
- Ambassador for Women in Data Science, a Stanford initiative

DATA SCIENTISTS

What people think I do?

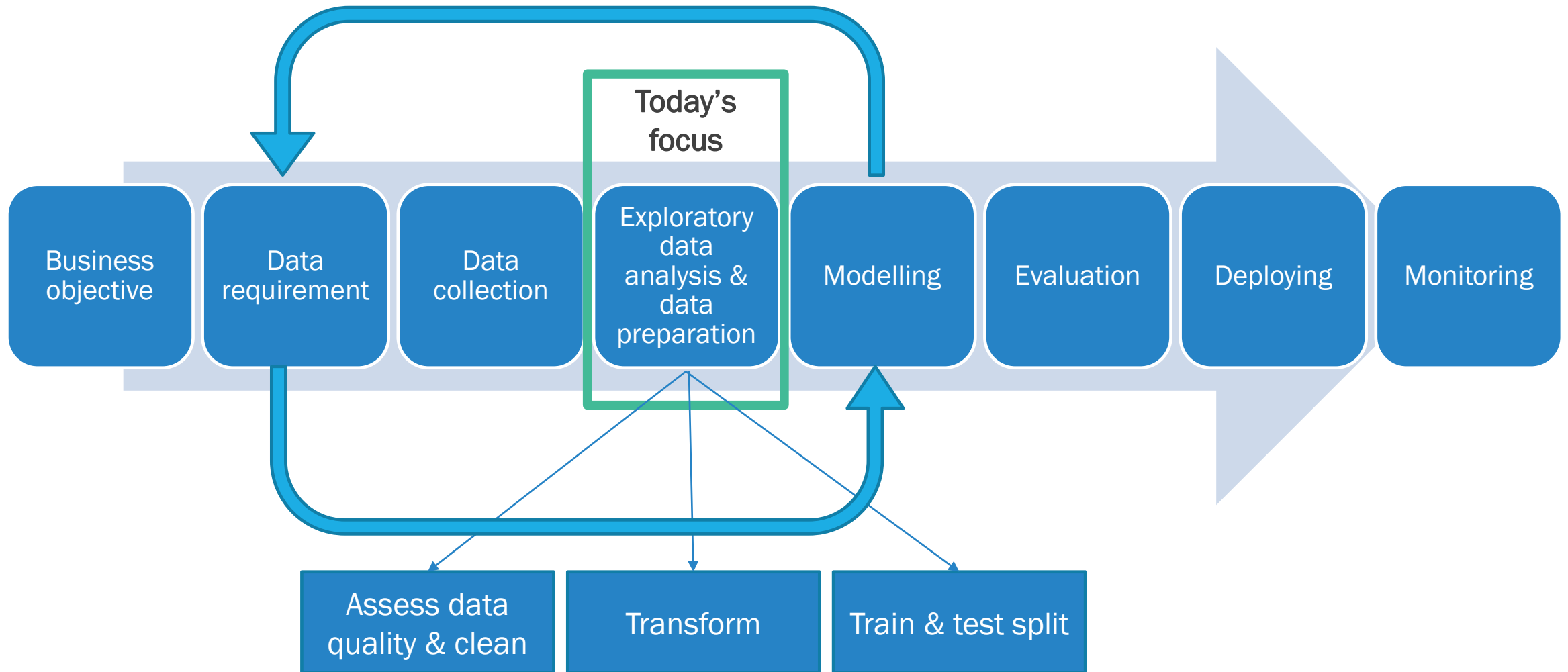


What I actually do?

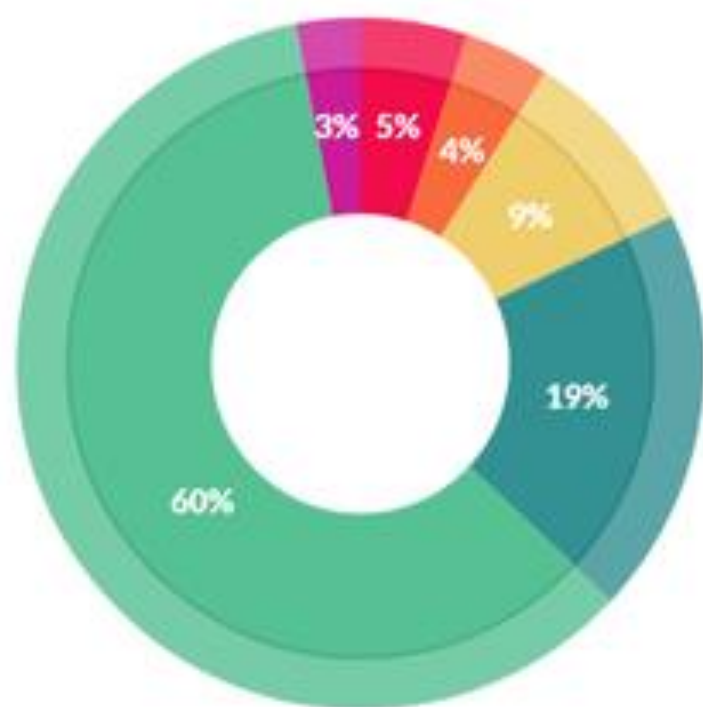
```
In [1]: import numpy as np  
import pandas as pd
```

```
In [ ]:
```

MACHINE LEARNING PROCESS



Perception Vs Reality



What data scientists spend the most time doing

● Building training sets: 3%

● Cleaning and organizing data: 60%

● Collecting data sets: 19%

● Mining data for patterns: 9%

● Refining algorithms: 4%

● Other: 5%

79%

79% of Machine Learning is Working with Data



WHAT IS DATA PREPARATION?

“Data preparation is the process of collecting data from several (usually disparate) data sources, and then profiling, cleansing, enriching, and combining those into a derived data set for use in analytics process”



NEED FOR DATA PREPARATION

- “Garbage in, Garbage out” - The models are only as good as the data you feed in
- Poor models lead to poor decisions
- Poor decisions costs a lot. Poor data quality is also hitting organizations where it hurts – to the tune of \$15 million as the average annual financial cost in 2017*

*Source - Gartner's Data Quality Market Survey



UNDERSTAND THE DATA

- Understand your data from the perspective of solving the business problem – what each variable means for the prediction
- Separate data from noise (this will also be covered in the model evaluation)
 - Are you using relevant variables only? Due to a shift to digital native environment, collecting and storing data is very easy. To use the right data is the actual challenge.
 - Are the variables you are using has an impact on the business decision making?
 - Using all the columns could cause discrimination – e.g., Gender; geo location, economic status, etc.



ASSESSING THE DATA QUALITY

A quick summary of the data can reveal

- Data type error
- Data format error
- Missing value – numerical and categorical
- Outlier
- Duplicates

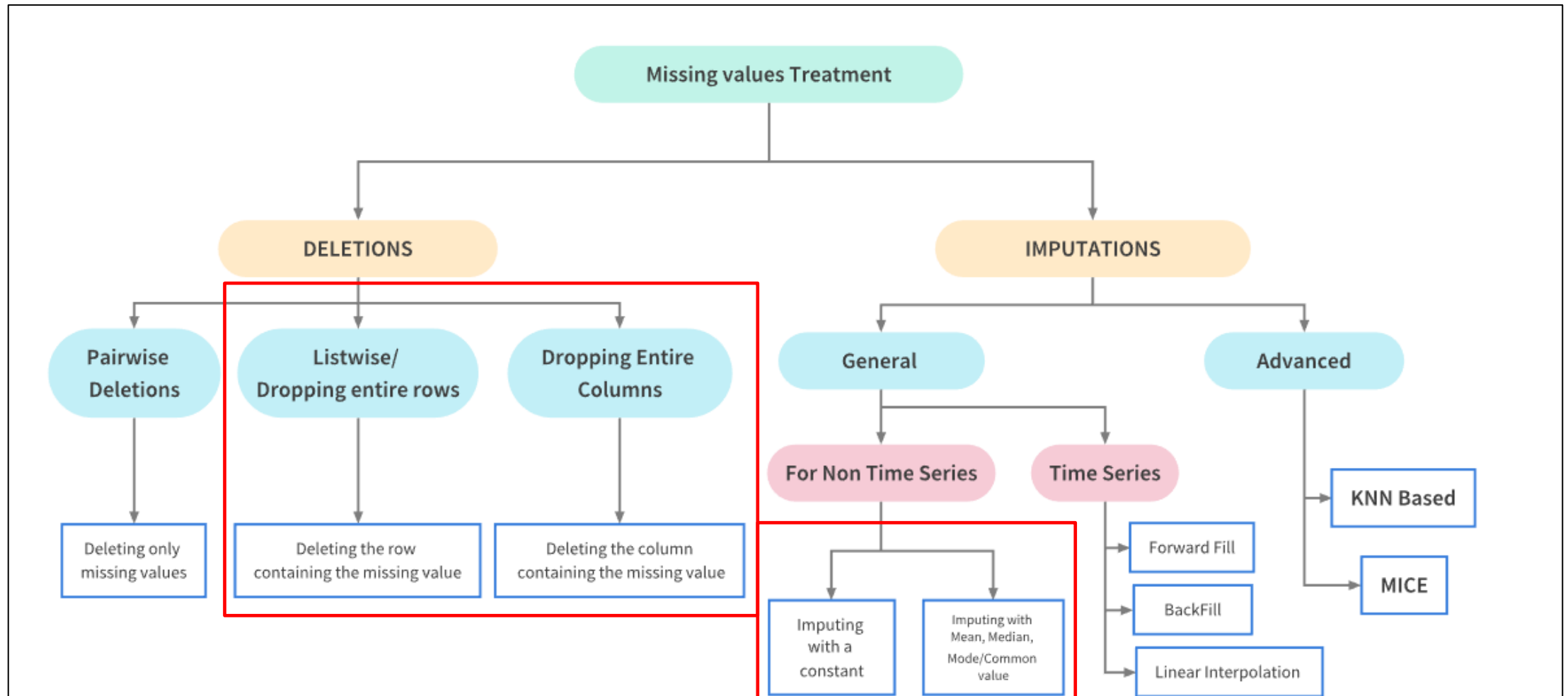
TITANIC DATASET – UNDERSTAND AND ASSESS THE DATA

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

DATA CLEANING – TYPE & FORMAT ISSUE

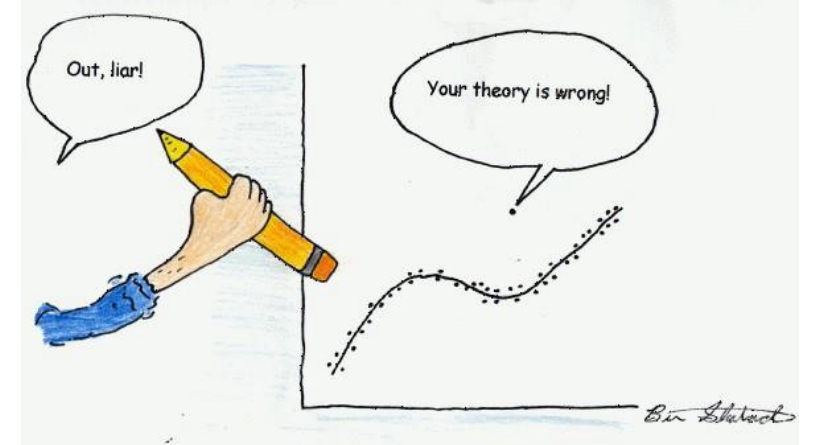
- How to fix data type error ? E.g., Numeric value in a string column or String value in a date column
 - If the error is at the source data level, it is difficult to fix in the analysis – Data quality check is an important activity at data collection stage itself
 - Treat them as missing values and impute
- How to fix data format error – did you see a DD/MM/YY format in a MM/DD/YYYY?
 - Keep the date column as a string and specify the date format of the variable
 - **`pd.to_datetime('01-10-2014 00:00:00', format='%d-%m-%Y %H:%M:%S')`**

DATA CLEANING – MISSING VALUES



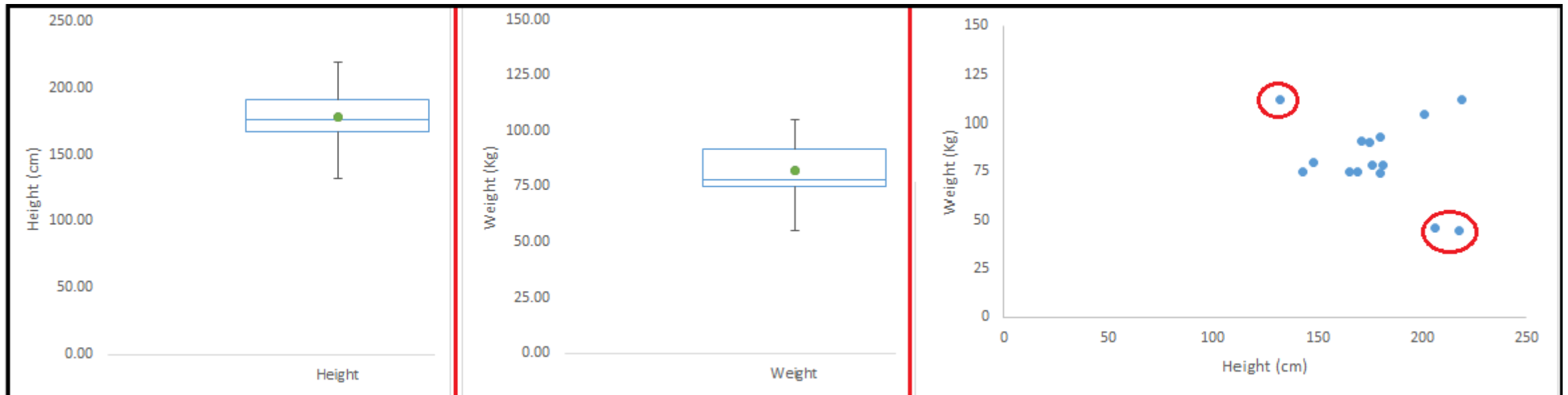
DATA CLEANING - OUTLIERS

- What is an Outlier? – an anomaly – We will generally define outliers as samples that are exceptionally far from the mainstream of the data.
- Outliers can have many causes, such as:
 - Measurement or input error
 - Data corruption
 - True outlier observation (e.g., Michael Jordan in basketball)
- Outlier can be in numeric, string, and date
 - String - E.g., City name is mentioned in a Country or gibberish data
 - Date – E.g., date appearing outside the period of analysis in scope



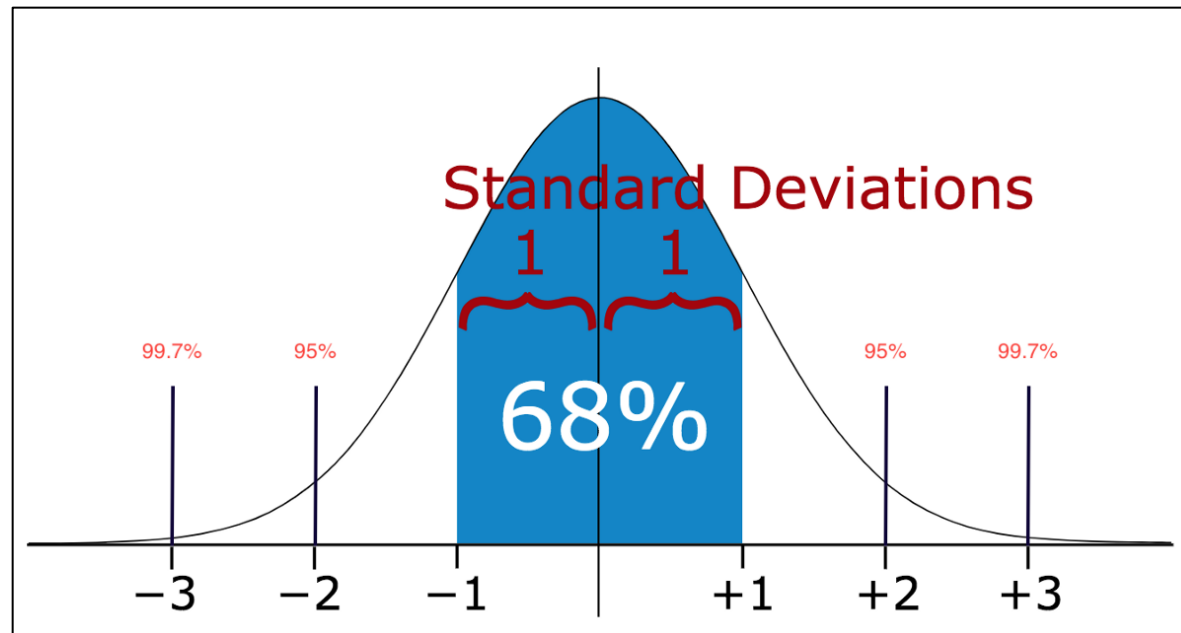
DATA CLEANING - OUTLIERS

- Outlier can be
 - Univariate outliers - can be found when we look at distribution of a single variable
 - Multi-variate outliers - are found in n-dimensional space. E.g., understanding the relationship between height and weight



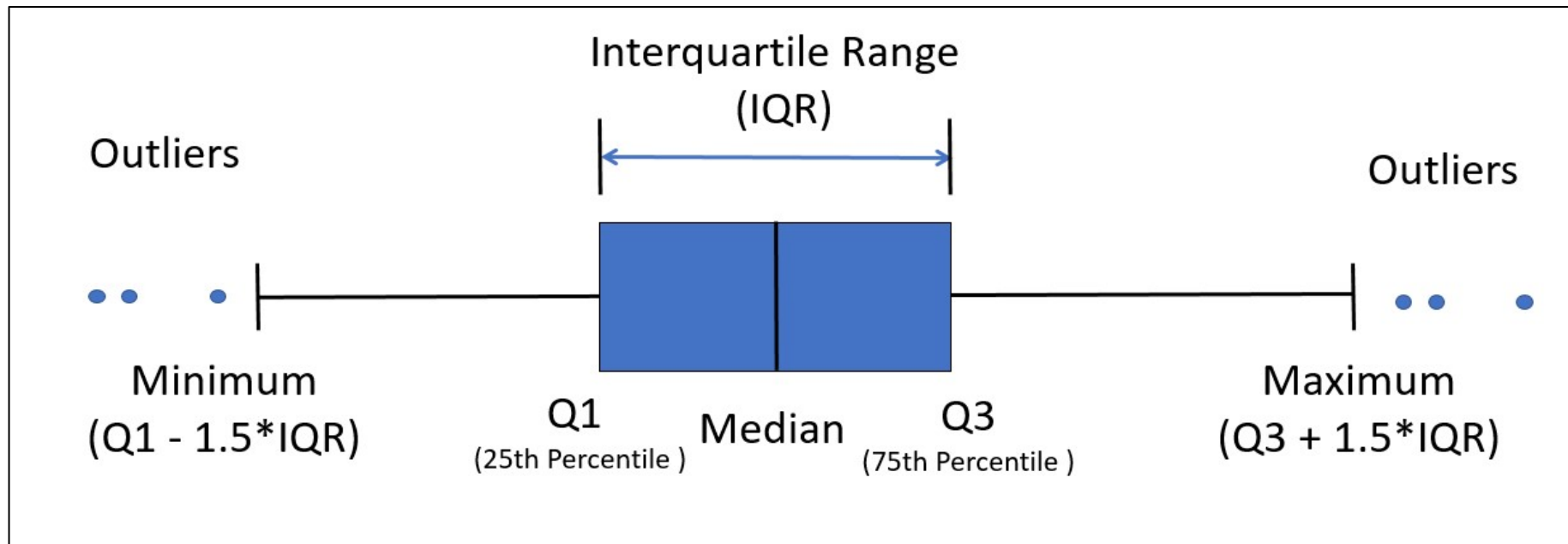
DATA CLEANING - OUTLIER DETECTION

- Outlier detection methods – Numeric
 - Standard Deviation - If we know that the distribution of values in the sample is Gaussian or Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers



DATA CLEANING - OUTLIER DETECTION

- Outlier detection methods – Numeric
 - Inter-Quartile Range (IQR) / Box-plot – IQR is a concept in statistics that is used to measure the statistical dispersion and data variability by dividing the dataset into quartiles.



DATA CLEANING - OUTLIER DETECTION

- Some of the complex outlier removal methods (not discussed today)
 - Isolation Forest
 - Minimum Covariance Determinant
 - Local Outlier Factor
 - One-Class SVM
 - DBScan Clustering



DATA CLEANING – OUTLIER TREATMENT

- Mean/Median or random imputation – same as missing value treatment
- Dropping values – remove the observations with outliers - same as missing value treatment
- Top, Bottom, and Zero Coding
- Discretization (Binning)




DATA CLEANING – DUPLICATE DATA

- Before removing duplicates from the data make sure
 - be sure they are not real data that coincidentally have values that are identical
 - try to figure why you have duplicates in your data (is it due to class imbalance?)

FEATURE ENGINEERING

- Feature encoding
 - Ordinal encoding - Ranking based numeric (e.g., Pclass)
 - One hot encoding – Survived; Embarked

id	color
1	red
2	blue
3	green
4	blue

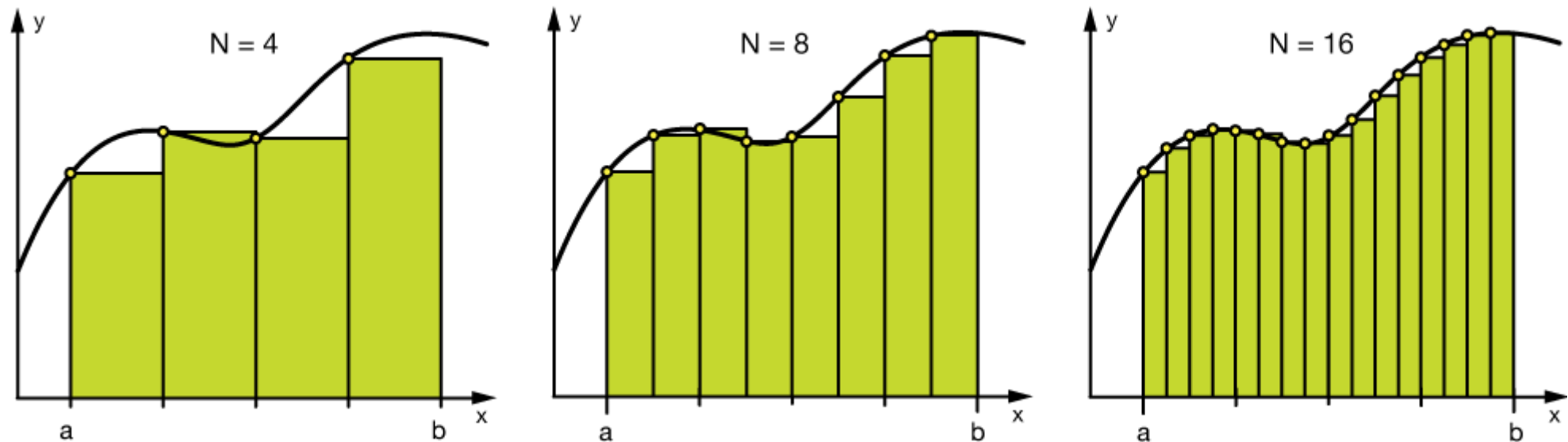


id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

- Dummy Variable Encoding (in case of linear regression)

FEATURE ENGINEERING

- Binning – can be done for both categorical and numeric variable
- The main motivation of binning is to make the model more **robust** and prevent **overfitting**, however, it has a cost to the performance. Every time you bin something, you sacrifice information and make your data more regularized.



FEATURE ENGINEERING

- Log transforms - Logarithm transformation (or log transform) is one of the most used mathematical transformations in feature engineering

```
#Log Transform Example
data = pd.DataFrame({'value':[2,45, -23, 85, 28, 2, 35, -12]})

data['log+1'] = (data['value']+1).transform(np.log)

#Negative Values Handling
#Note that the values are different
data['log'] = (data['value']-data['value'].min()+1)
               .transform(np.log)
```

	value	log(x+1)	log(x-min(x)+1)
0	2	1.09861	3.25810
1	45	3.82864	4.23411
2	-23	nan	0.00000
3	85	4.45435	4.69135
4	28	3.36730	3.95124
5	2	1.09861	3.25810
6	35	3.58352	4.07754
7	-12	nan	2.48491

FEATURE ENGINEERING

Scaling

- **Normalization**

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Standardization**

$$z = \frac{x - \mu}{\sigma}$$

the **mean** is shown as μ and
the **standard deviation** is shown as σ .



BEFORE RUNNING THE MODEL

- Input / Target variable – store in separate DataFrame
- Split into train and validation set
 - Train – what we use to train the model
 - Validation – what we use to evaluate the model
 - Test – unexposed data to the model

CLASS IMBALANCE

- An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed (let us look only at binary classification)
- Imbalance can occur due to:
 - Biased sampling – E.g., Sampling only from a single geographic location
 - Nature of the problem statement – E.g., any fraudulent transactions like credit card frauds, etc.
- The imbalance could be
 - Slight Imbalance (gender distribution – 60% male; 40% female)
 - Severe Imbalance (claims prediction in insurance)
- Terms (Minority class - that has few examples; Majority class - that has many examples)

CLASS IMBALANCE

- How to deal with class imbalance?
 - **Resampling Techniques** –
 - **Oversample Minority Class:** Oversampling can be defined as adding more copies of the minority class. In other words, we are creating artificial/synthetic data of the minority class (or group). Oversampling could be a good choice when you don't have a lot of data to work with.
 - **Undersample Majority Class:** Undersampling can be defined as removing some observations of the majority class. Undersampling can be a good choice when you have a ton of data -think millions of rows. But a drawback is that we are removing information that may be valuable. This could lead to underfitting and poor generalization to the test set
 - **Generate Synthetic Samples:** Here we will use imblearn's SMOTE or Synthetic Minority Oversampling Technique. SMOTE uses a nearest neighbors' algorithm to generate new and synthetic data we can use for training our model.

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.



Questions?