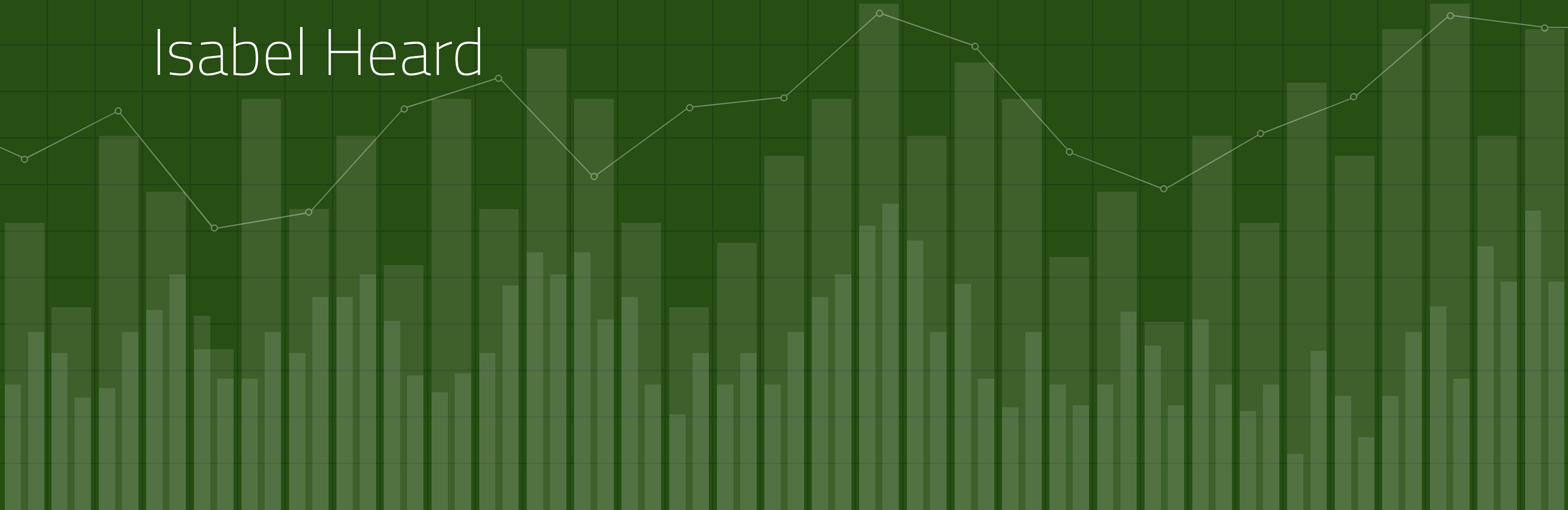
Missing Data

Isabel Heard



Agenda

01 Motivation

02 Importance

03 Types of Missing Data

04 Techniques

05 Example

06 Conclusion & Considerations

Motivation

What do we do with missing data?

Why is this Topic Important?

- When not handled appropriately, missing data can:
  - Reduces statistical power
  - Have biased results
  - Inaccurate insights
  - Raise ethical issues

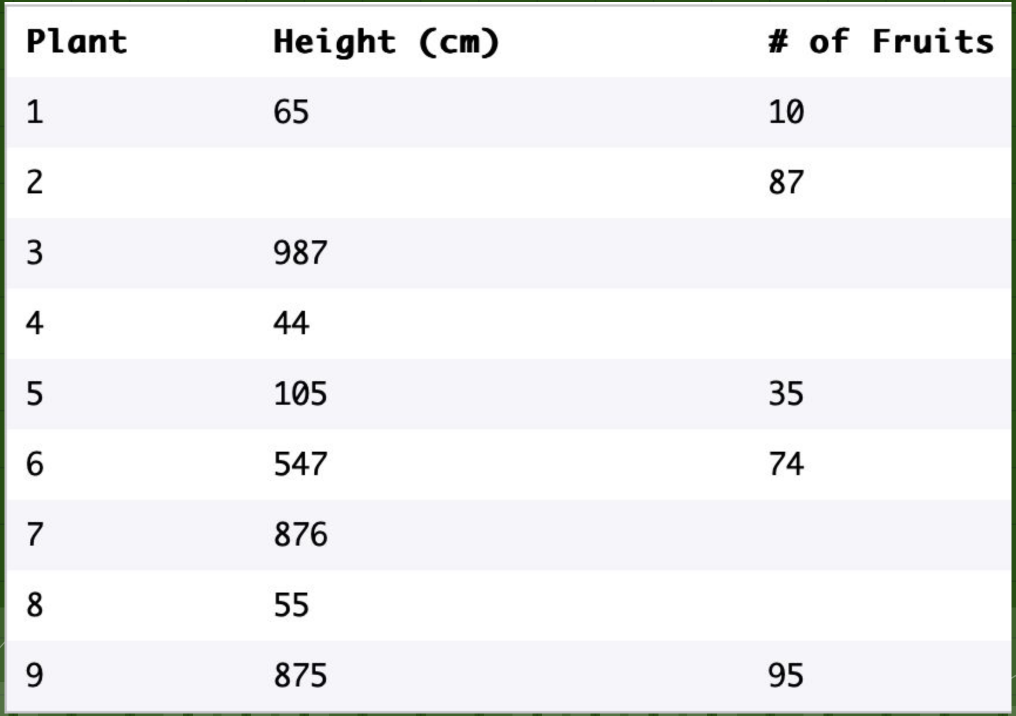
Different Types of Missing Data

Missing Completely at Random

Missing at Random

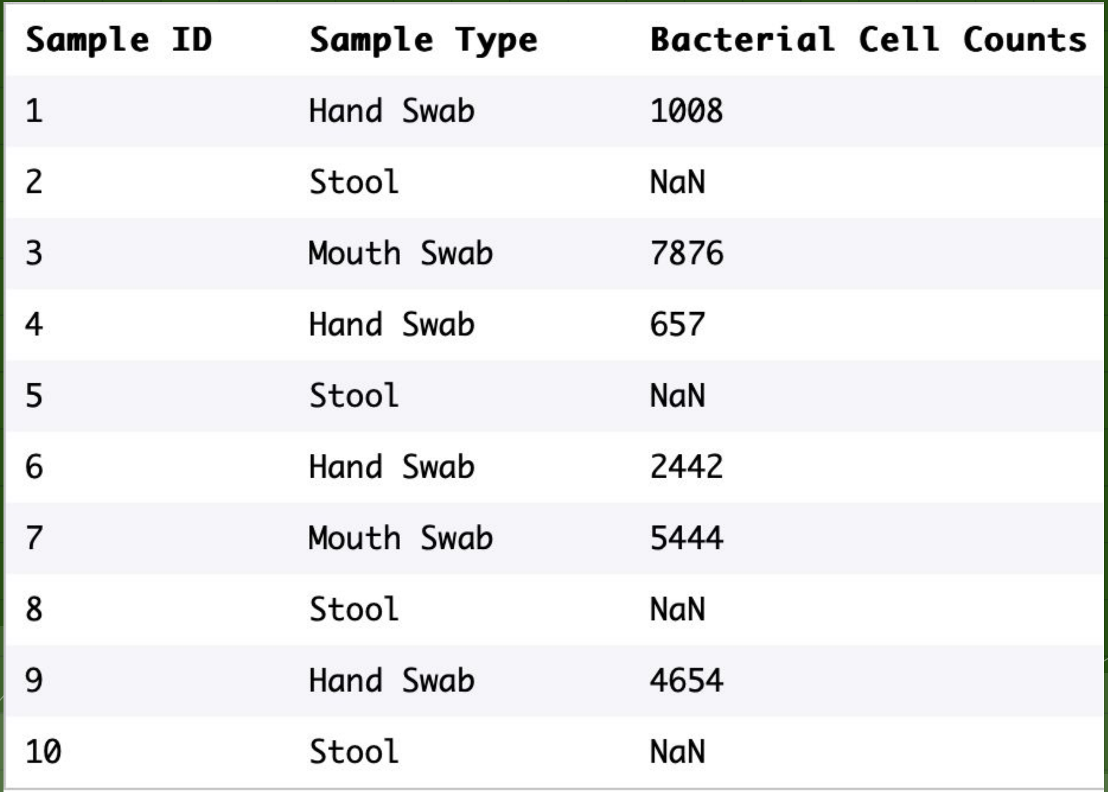
Missing Not at Random

MCAR – The missingness of data is completely random



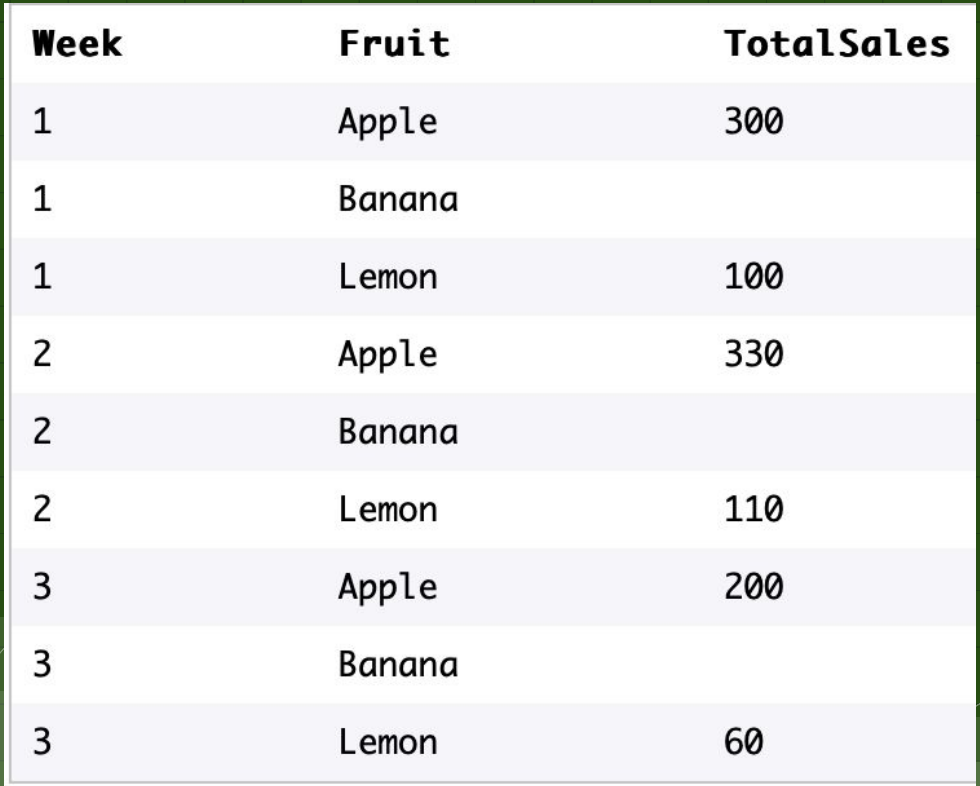
Plant	Height (cm)	# of Fruits
1	65	10
2		87
3	987	
4	44	
5	105	35
6	547	74
7	876	
8	55	
9	875	95

MAR – The probability of the value being missing is related to the value of the other variables in the dataset



Sample ID	Sample Type	Bacterial Cell Counts
1	Hand Swab	1008
2	Stool	NaN
3	Mouth Swab	7876
4	Hand Swab	657
5	Stool	NaN
6	Hand Swab	2442
7	Mouth Swab	5444
8	Stool	NaN
9	Hand Swab	4654
10	Stool	NaN

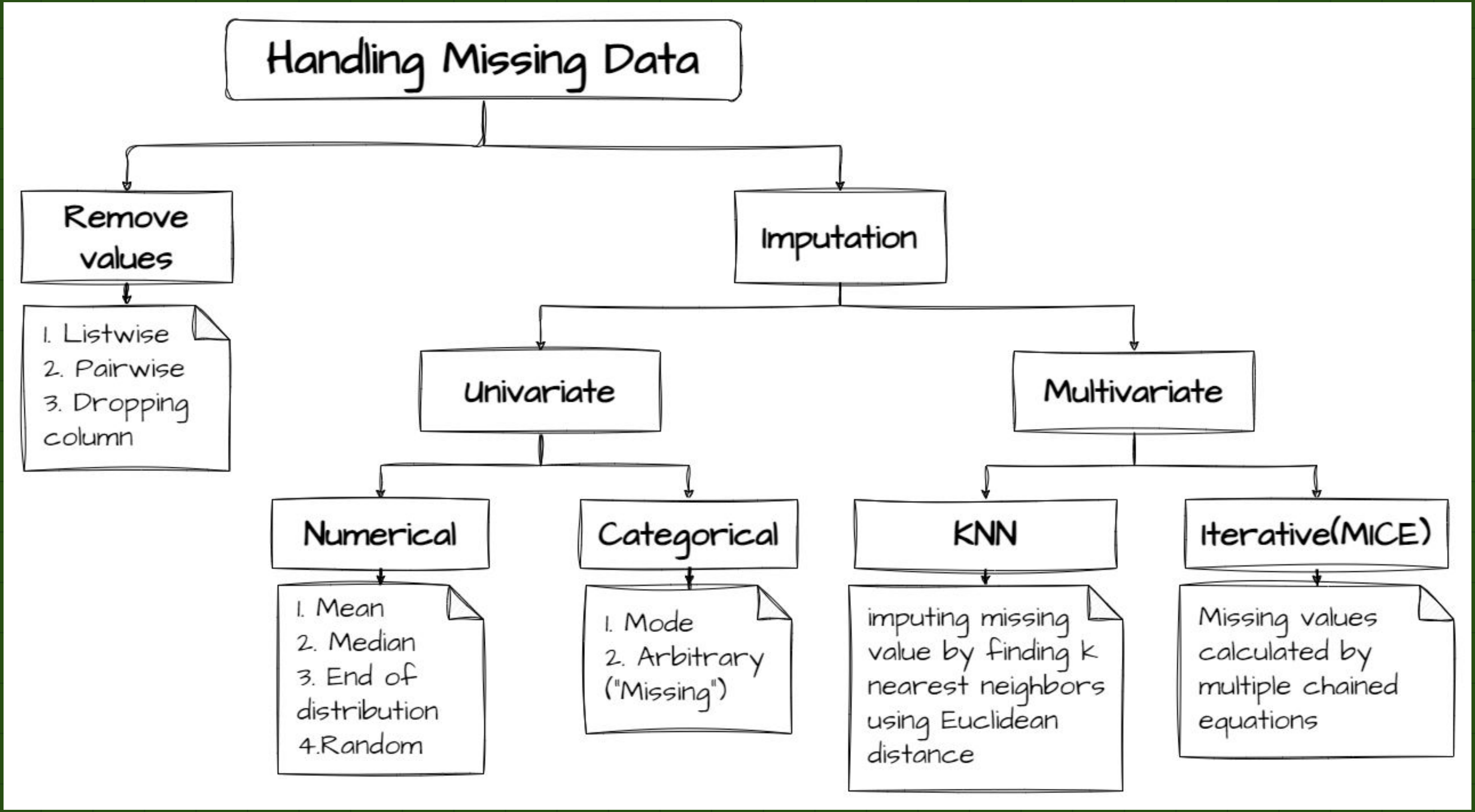
MNAR – The probability of being missing is completely different for different values of the same variable



Week	Fruit	TotalSales
1	Apple	300
1	Banana	
1	Lemon	100
2	Apple	330
2	Banana	
2	Lemon	110
3	Apple	200
3	Banana	
3	Lemon	60

Missing Data Diagnostics

1. Make reasonable guesses about the type of missing data you have
   a. Why the data is missing (if possible)?
   b. Distribution of missing data
2. Decide on the best analysis strategy to yield the least biased estimates



Handling Missing Data

Remove values

1. Listwise
2. Pairwise
3. Dropping column

Imputation

Univariate

Numerical

1. Mean
2. Median
3. End of distribution
4. Random

Categorical

1. Mode
2. Arbitrary ("Missing")

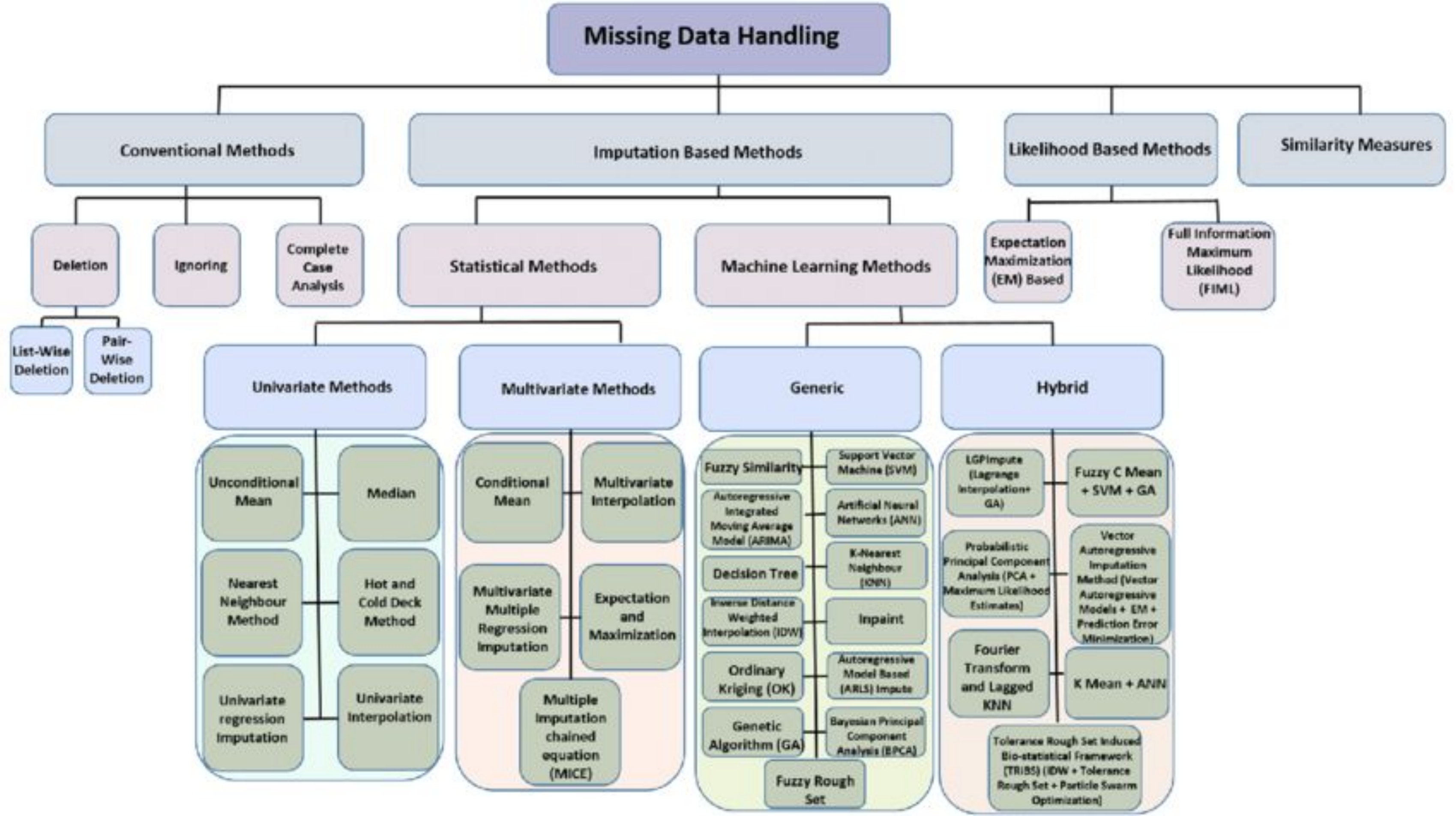
Multivariate

KNN

imputing missing value by finding k nearest neighbors using Euclidean distance

Iterative(MICE)

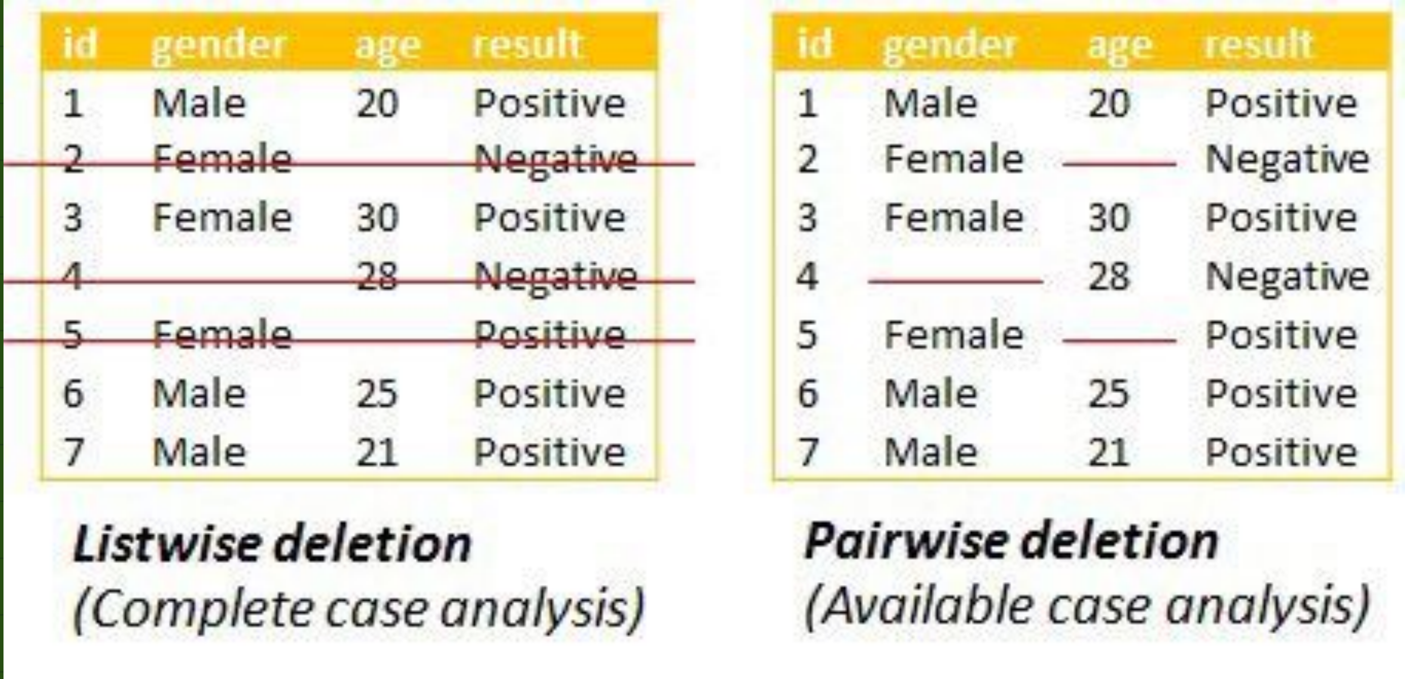
Missing values calculated by multiple chained equations



Missing Data Handling

- Conventional Methods
  - Deletion
    - List-Wise Deletion
    - Pair-Wise Deletion
  - Ignoring
  - Complete Case Analysis
- Imputation Based Methods
  - Statistical Methods
    - Univariate Methods
      - Unconditional Mean
      - Median
      - Nearest Neighbour Method
      - Hot and Cold Deck Method
      - Univariate regression imputation
      - Univariate Interpolation
    - Multivariate Methods
      - Conditional Mean
      - Multivariate Interpolation
      - Multivariate Multiple Regression Imputation
      - Expectation and Maximization
      - Multiple Imputation chained equation (MICE)
  - Machine Learning Methods
    - Generic
      - Fuzzy Similarity
      - Support Vector Machine (SVM)
      - Autoregressive Integrated Moving Average Model (ARIMA)
      - Artificial Neural Networks (ANN)
      - Decision Tree
      - K-Nearest Neighbour (KNN)
      - Inverse Distance Weighted Interpolation (IDW)
      - Inpaint
      - Ordinary Kriging (OK)
      - Autoregressive Model Based (ARLS) Impute
      - Genetic Algorithm (GA)
      - Bayesian Principal Component Analysis (BPCA)
      - Fuzzy Rough Set
    - Hybrid
      - LGPImpute (Lagrange Interpolation+ GA)
      - Fuzzy C Mean + SVM + GA
      - Probabilistic Principal Component Analysis (PCA + Maximum Likelihood Estimates)
      - Vector Autoregressive Imputation Method (Vector Autoregressive Models + EM + Prediction Error Minimization)
      - Fourier Transform and Lagged KNN
      - K Mean + ANN
      - Tolerance Rough Set Induced Bio-statistical Framework (TRIBS) (IDW + Tolerance Rough Set + Particle Swarm Optimization)
- Likelihood Based Methods
  - Expectation Maximization (EM) Based
  - Full Information Maximum Likelihood (FIML)
- Similarity Measures

Deletion Methods



id	gender	age	result
1	Male	20	Positive
~~2~~	~~Female~~		~~Negative~~
3	Female	30	Positive
~~4~~		~~28~~	~~Negative~~
~~5~~	~~Female~~		~~Positive~~
6	Male	25	Positive
7	Male	21	Positive

Listwise deletion

(Complete case analysis)

id	gender	age	result
1	Male	20	Positive
2	Female	—	Negative
3	Female	30	Positive
4	—	28	Negative
5	Female	—	Positive
6	Male	25	Positive
7	Male	21	Positive

Pairwise deletion

(Available case analysis)

Imputation Methods

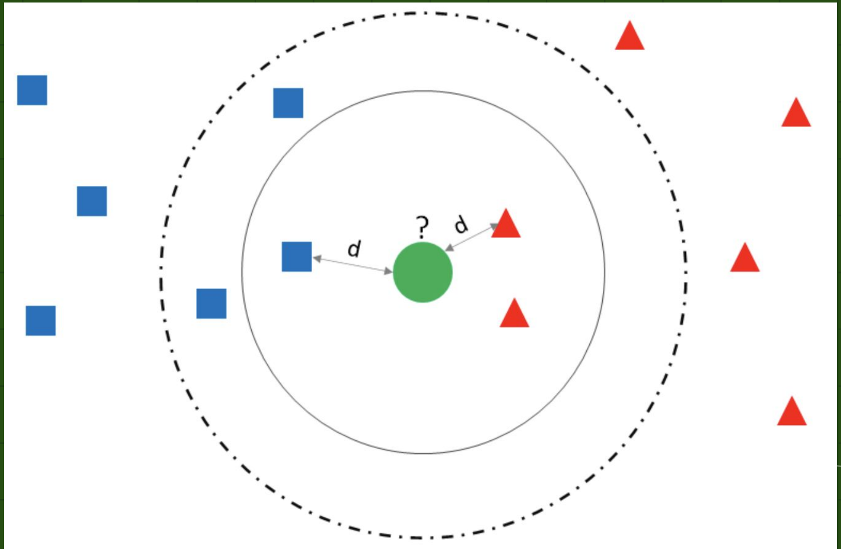
- Mean/Median/Mode
- Linear Regression
- Forward/Backward Fill
- Hot Deck
- Multiple Imputation

ML Techniques

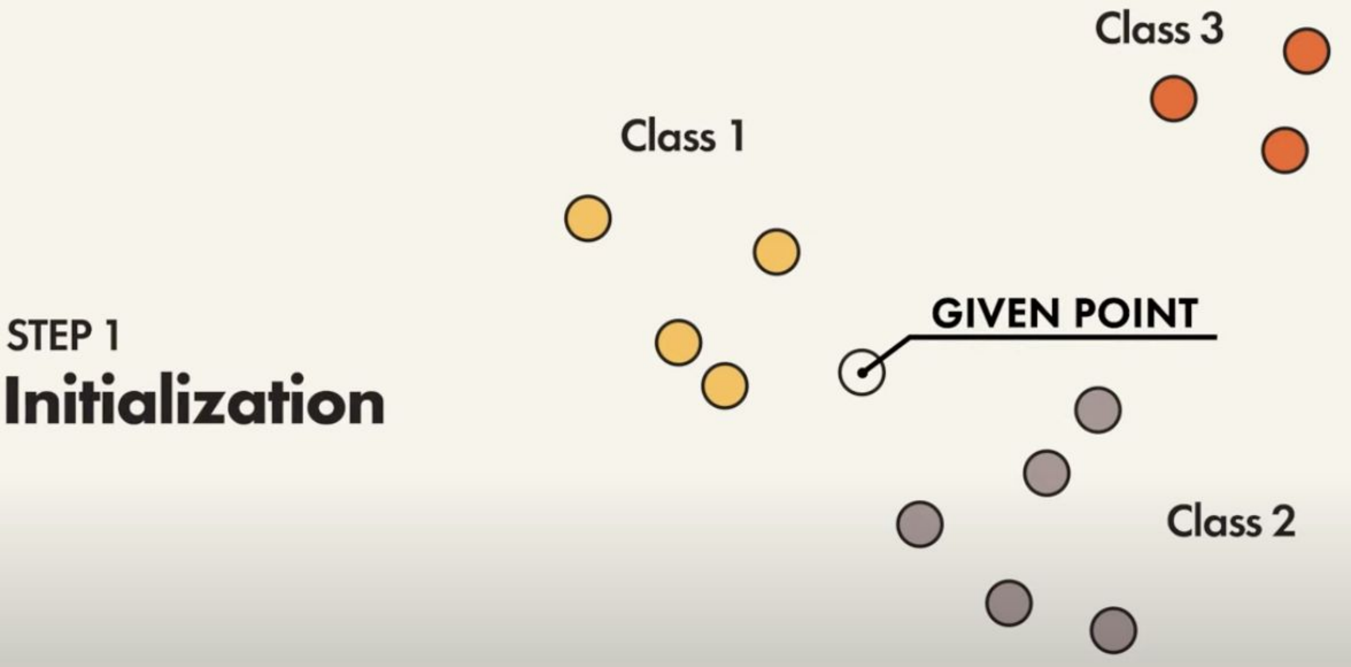
- Deep Learning
- Random Forest
- Decision Trees
- MLP
- KNN

K-Nearest Neighbors

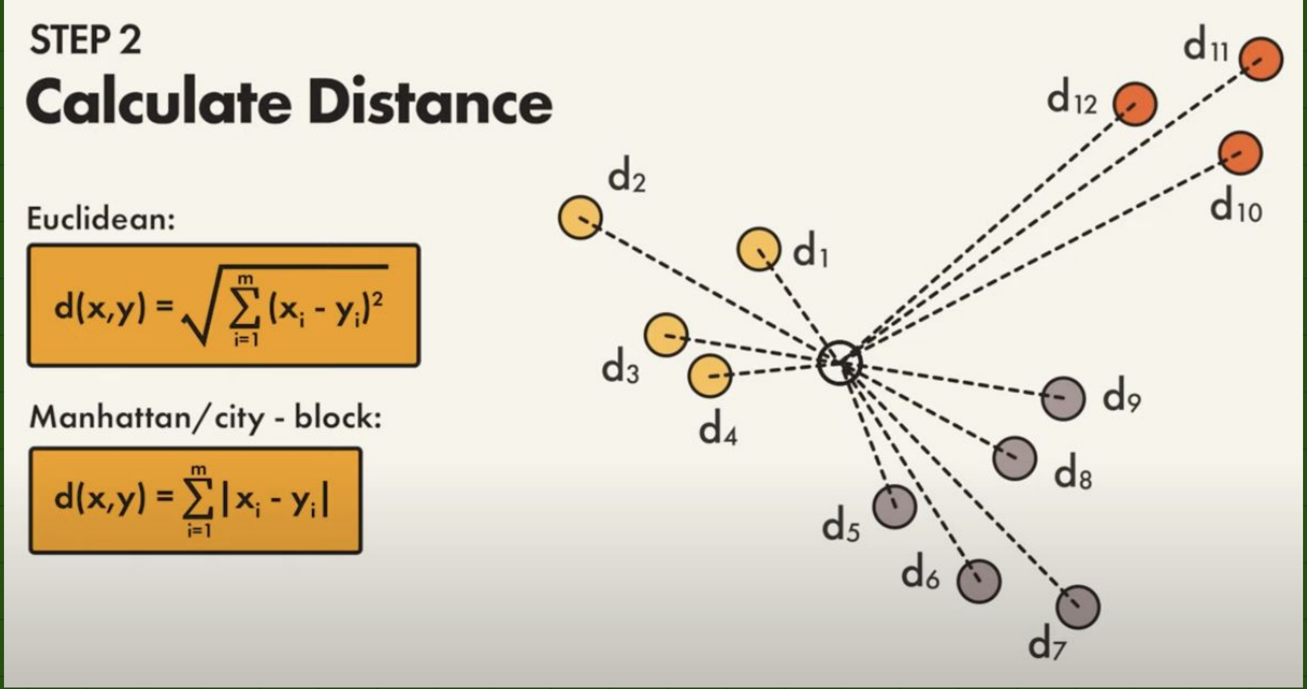
KNN is a imputation technique used to fill in missing values in a dataset by leveraging information from neighboring observations.



K-Nearest Neighbors



K-Nearest Neighbors



STEP 2

Calculate Distance

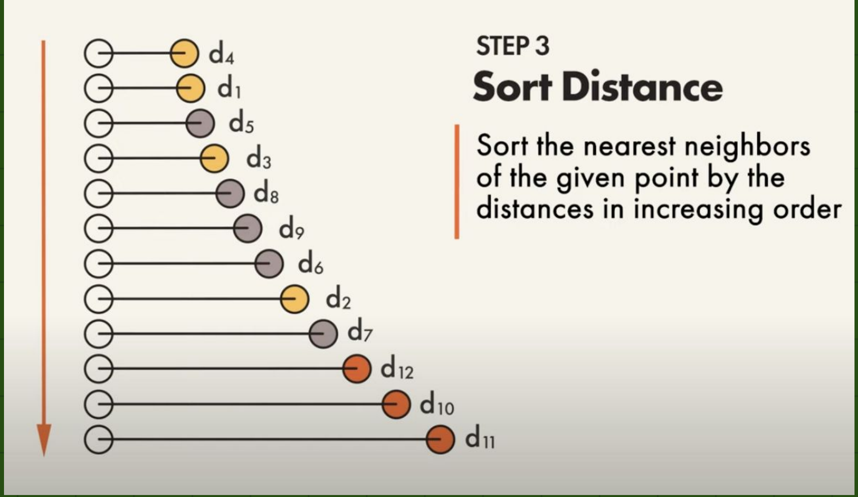
Euclidean:

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan/city - block:

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$$

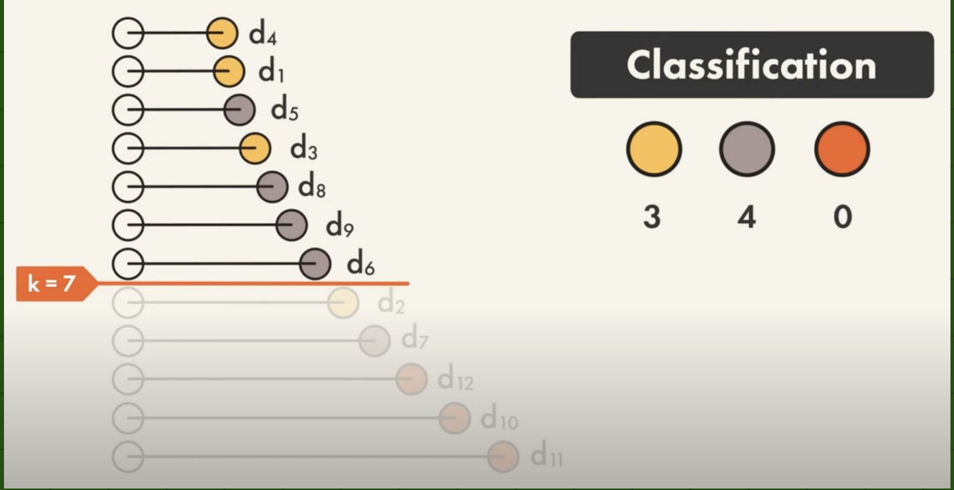
K-Nearest Neighbors



STEP 3

Sort Distance

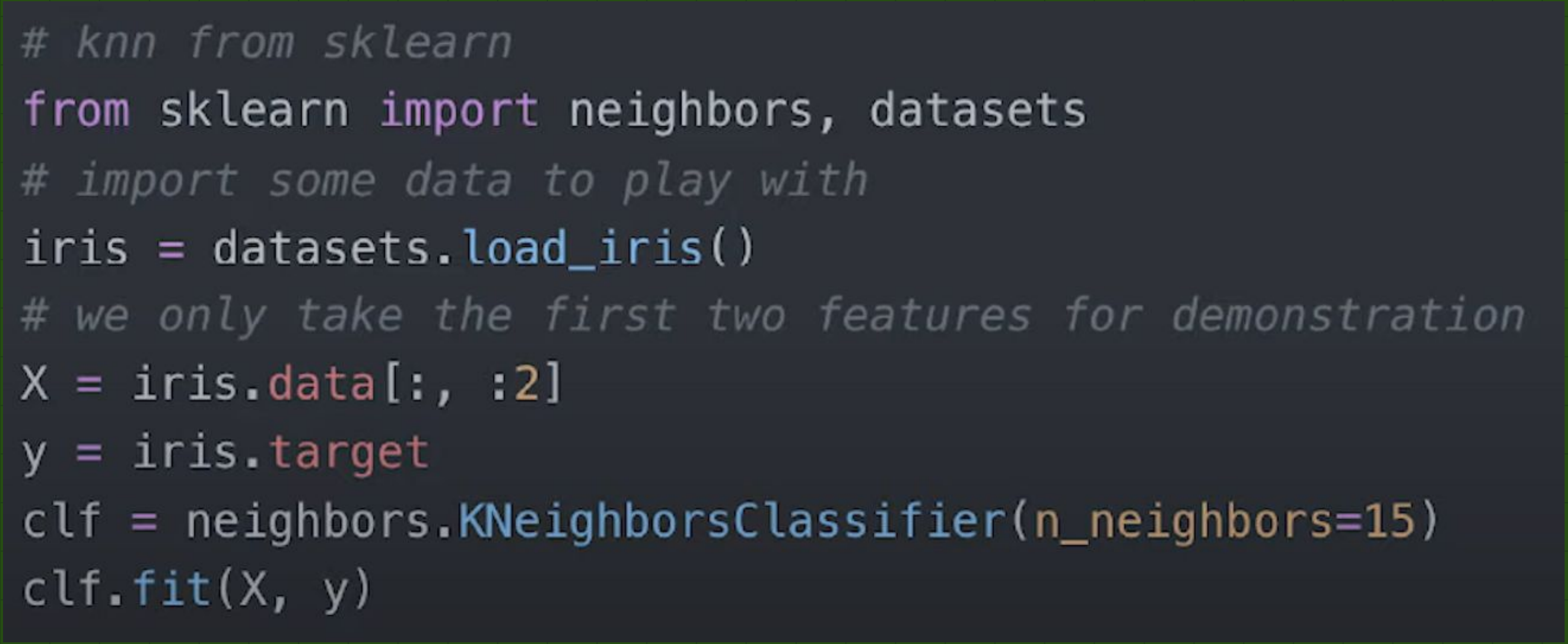
Sort the nearest neighbors of the given point by the distances in increasing order



Classification

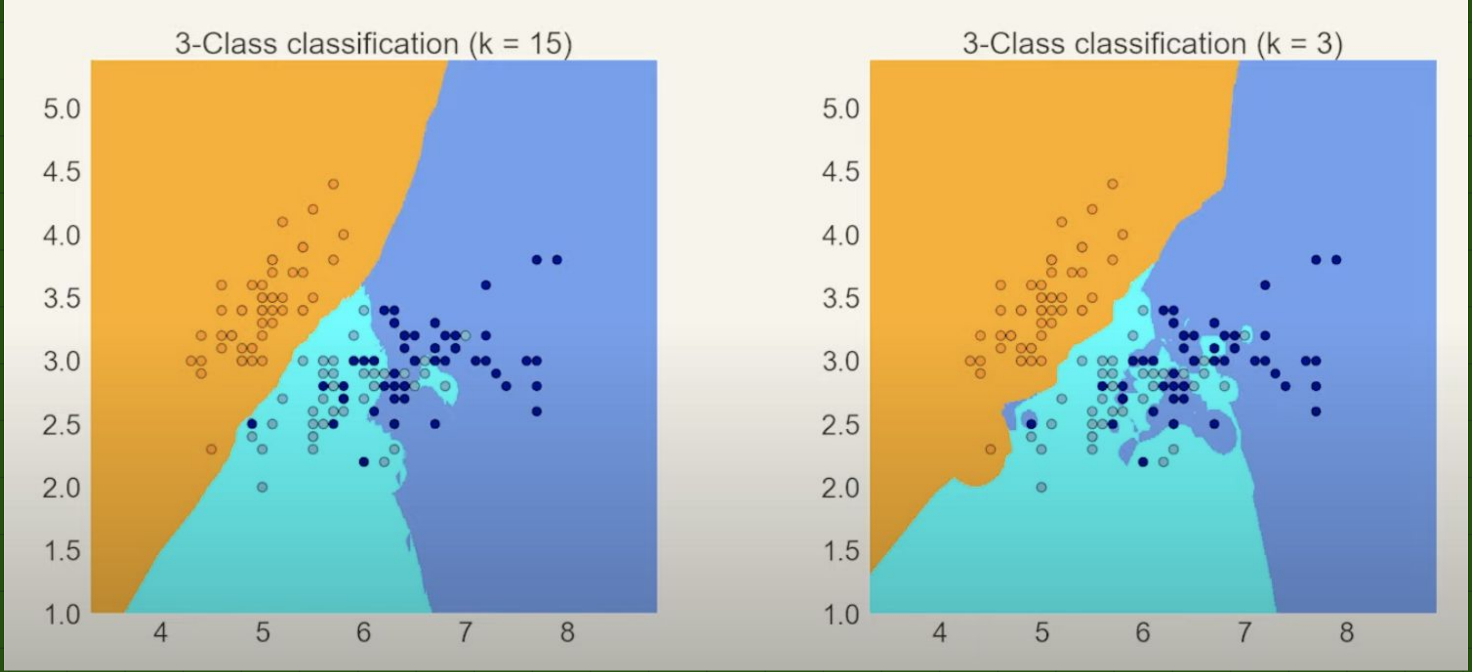
●	●	●
3	4	0

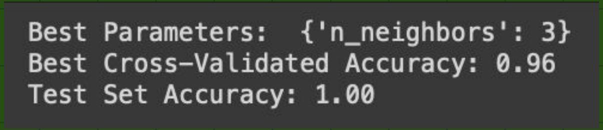
K-Nearest Neighbors



```
# knn from sklearn
from sklearn import neighbors, datasets
# import some data to play with
iris = datasets.load_iris()
# we only take the first two features for demonstration
X = iris.data[:, :2]
y = iris.target
clf = neighbors.KNeighborsClassifier(n_neighbors=15)
clf.fit(X, y)
```

K-Nearest Neighbors





```
Best Parameters:  {'n_neighbors': 3}
Best Cross-Validated Accuracy: 0.96
Test Set Accuracy: 1.00
```

Domain-Specific Methods

- Use expert judgment to input missing values based on knowledge about the dataset and the subject matter
  - Substitution by constants
  - Develop imputation rules for specific domains
  - Leverage known relationships among variables

Example - Census

- They have long established procedures used in previous cases
    - Have used characteristic imputation since the 1960s
- Collect missing data from outside sources
- **“We use respondent’s first name to try to fill in missing sex. We also assign sex to maintain household consistency. For example, if sex is missing for the householder’s opposite-sex spouse or unmarried partner, we assign the sex that fits with that response.”**

Conclusion & Considerations

- When you have missing data, always think about why they are missing.
- Missing data handled improperly can bias your conclusions.
- It can be helpful to summarize or visualize patterns of missingness.
- Question a dataset that has no missing data.

Sources

- https://www.researchgate.net/figure/Different-methods-for-handling-missing-data-Hierarchical-tree-depicting-the_fig1_333304659
- https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values#
- https://harvard-iacs.github.io/2020-CS109A/lectures/lecture19/slides/Lecture19_Missingdata.pdf
- https://www-users.york.ac.uk/~mb55/intro/typemiss4.htm
- https://towardsdatascience.com/missing-data-effects-on-the-correlation-between-ice-cream-sales-and-temperature-f4bb2b3fcde1
- https://www.codecademy.com/learn/handling-missing-data/modules/handling-missing-data-intro/cheatsheet
- https://www.montecarlodata.com/blog-bad-data-quality-examples/
- https://www.youtube.com/watch?v=0p0o5cmgLdE
- https://www.census.gov/programs-surveys/sipp/methodology/data-editing-and-imputation.html

Questions?