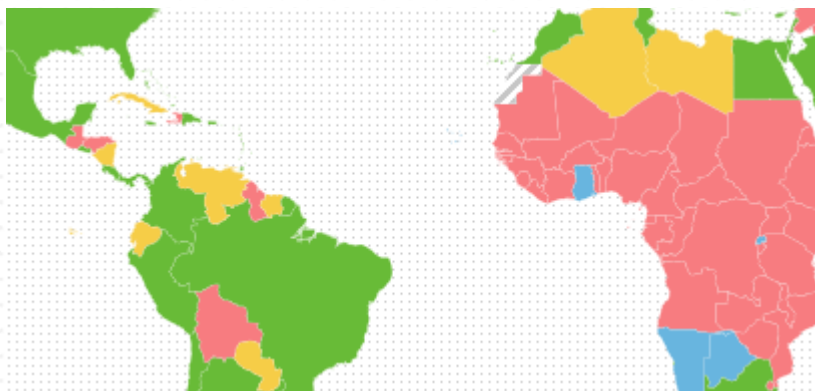
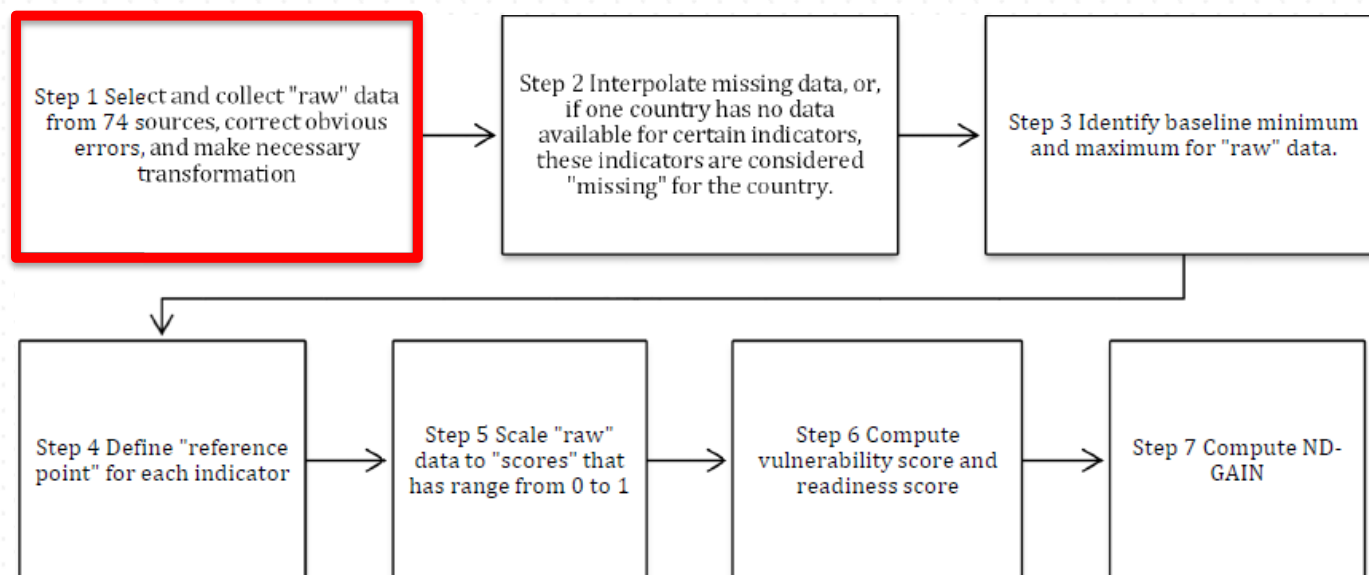


Chapter 3. Data
Processing:
Data Cleaning

Meng Jiang
Data Science

The Notre Dame-Global Adaptation Index (ND-GAIN)

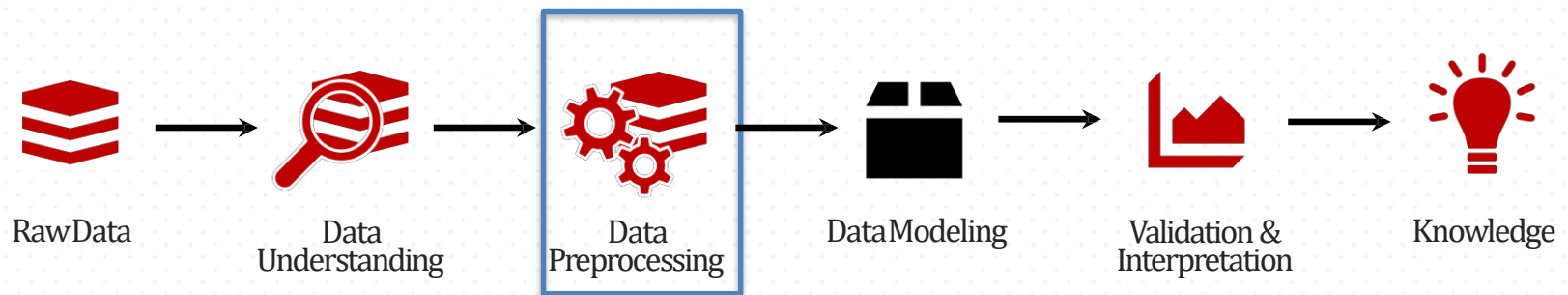
Step Number One



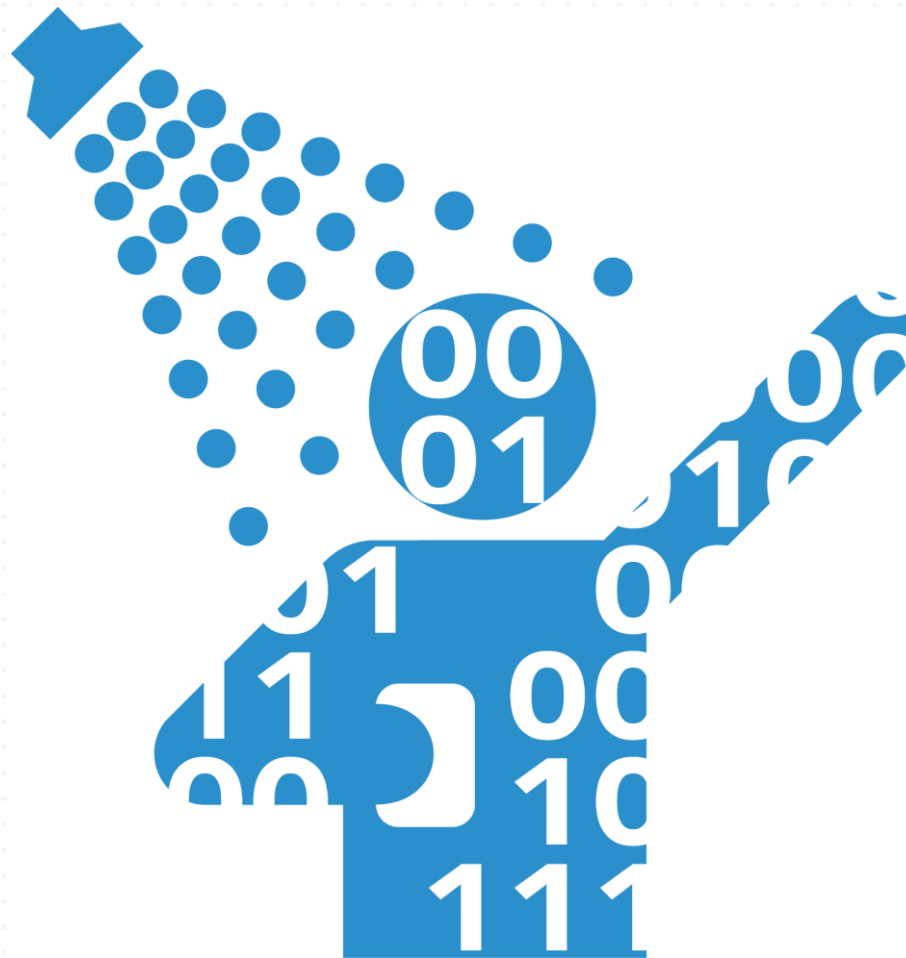
Water

Projected change of annual runoff	0.304
Projected change of annual groundwater recharge	0.738
Fresh water withdrawal rate	0.009
Water dependency ratio	0.477
Dam capacity	0.339
Access to reliable drinking water	0.042

Chapter 2. Getting to Know Your Data



Why Do We Clean Data?



Write Down Your Answers

- Name
- Dorm
- Height
- Hometown
- Major
- Intended Job after Graduation

Why Do We Clean Data?

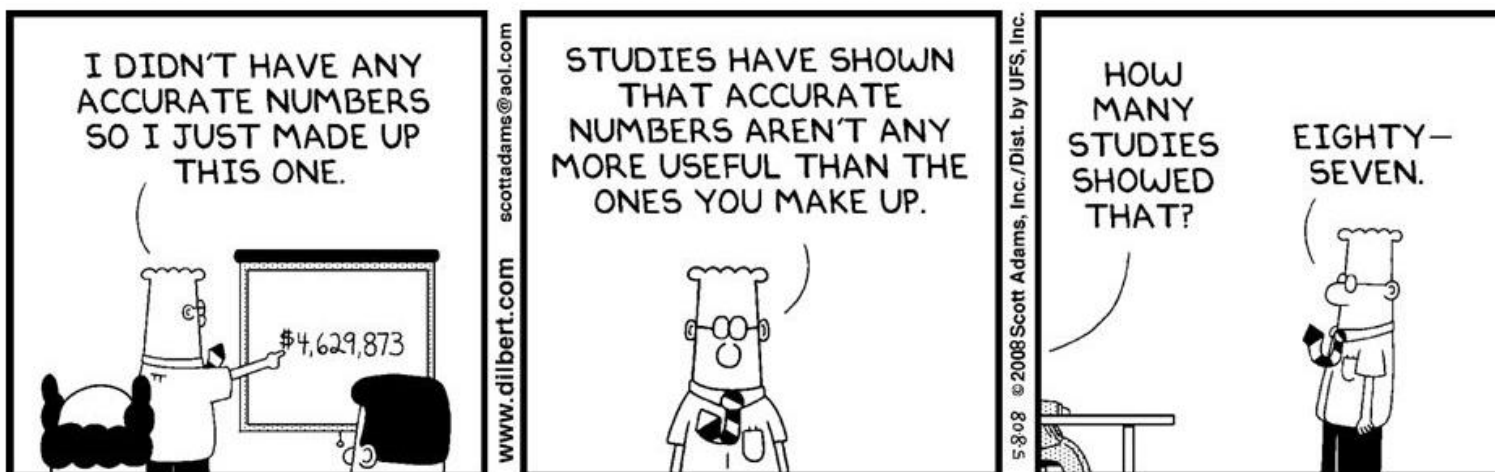
- Measures for data quality: A multidimensional view
 - _____
 - _____
 - Completeness: not recorded, unavailable, ...
 - _____
 - _____
 - _____

Why? Data Quality Issues

- Measures for data quality: A multidimensional view
 - Accuracy: correct or wrong, accurate or not
 - Believability: how trustable the data are correct?
 - Completeness: not recorded, unavailable, ...
 - Consistency: some modified but some not, ...
 - Timeliness: timely update?
 - Interpretability: how easily the data can be understood?

GIGO

Garbage In – Garbage Out



DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc.

Data Preprocessing

- **Data cleaning**
- **Data integration**
- Data reduction
- Dimensionality reduction

Data Cleaning

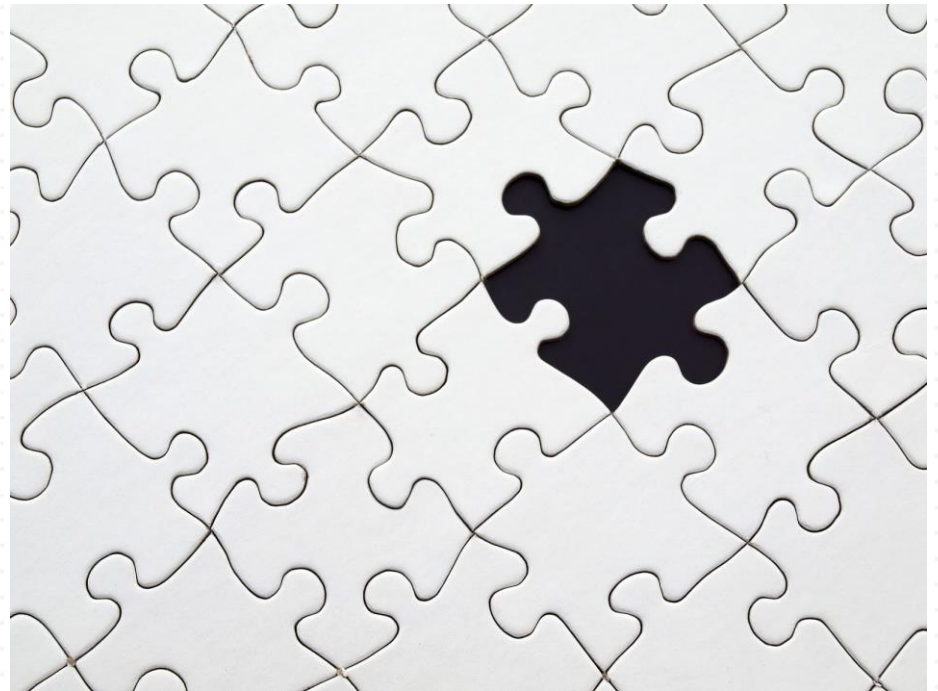
- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
 - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., *Occupation* = " " (missing data)
 - Noisy: containing noise, errors, or outliers
 - e.g., *Salary* = "-10" (an error)
 - Inconsistent: containing discrepancies in codes or names, e.g.,
 - *Age* = "42", *Birthday* = "03/07/2010"
 - Was rating "1, 2, 3", now rating "A, B, C"

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
 - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., *Occupation* = " " (missing data)
 - Specical Case: Intentional (e.g., *disguised missing data*)
 - *Jan. 1 as everyone's birthday?*
 - Noisy: containing noise, errors, or outliers
 - e.g., *Salary* = "-10" (an error)
 - Inconsistent: containing discrepancies in codes or names, e.g.,
 - *Age* = "42", *Birthday* = "03/07/2010"
 - Was rating "1, 2, 3", now rating "A, B, C"

Incomplete (Missing) Data

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - Equipment malfunction
 - Inconsistent with other recorded data and thus deleted
 - Data were not entered due to misunderstanding
- Missing data may need to be inferred



Types of Missing Data

There are three types of missing data

Data may be:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Missing Completely At Random

- Missingness does not depend on any values of any variables in the dataset.
- Missingness instead depends on neither the values of the observed variables, nor on those of unobserved variables.

Example: The accidental dropping a test tube leading to missing lab test result

Missing At Random

- Missingness does not depend on the values of any of the missing or unobserved variables.
- Instead, missingness might depend on values of the observed variables.
- This means that the pattern of missing values is identifiable.

Example: suppose males are less likely to respond to their income question in general, but the likelihood of responding is independent of their actual income. In this case, unbiased sex-specific income estimates can be made if we have data on the sex variable (by replacing the missing value with the sex-specific median income, for example)

Missing Not At Random

- Missingness depends on the values of the missing or unobserved variables.
- This means that the pattern is non-random, non-ignorable, and typically arises due to the variable on which the data is missing.

Example: A certain question on a questionnaire tend to be skipped deliberately by participants with certain characteristics

Example: Missing Value Types

A

Customer	Age	Account Balance
Customer 1	25	<i>Missing</i>
Customer 2	25	100,000
Customer 3	25	<i>Missing</i>
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	150,000

B

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	<i>Missing</i>
Customer 3	25	15,000
Customer 4	60	50,000
Customer 5	60	<i>Missing</i>
Customer 6	60	<i>Missing</i>

C

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	100,000
Customer 3	25	<i>Missing</i>
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	<i>Missing</i>

1. *Missing Completely At Random*
2. *Missing At Random*
3. *Missing Not At Random*

Example: Missing Value Types

MAR

Customer	Age	Account Balance
Customer 1	25	<i>Missing</i>
Customer 2	25	100,000
Customer 3	25	<i>Missing</i>
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	150,000

The account balance is observed only for *Age = 60*, thus the missingness can be modeled on age

MNAR

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	<i>Missing</i>
Customer 3	25	15,000
Customer 4	60	50,000
Customer 5	60	<i>Missing</i>
Customer 6	60	<i>Missing</i>

P Balance Missing Balance < 100,000 = 0
P Balance Missing Balance > 100,000 = 1

MCAR

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	100,000
Customer 3	25	<i>Missing</i>
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	<i>Missing</i>

How to Handle Missing Data?



How to Handle Missing Data?

Ignore the tuple

Customer	Age	Account Balance
1	25	Missing
2	25	100,000
3	25	Missing
4	60	50,000
5	60	120,000
6	60	150,000

Customer	Age	Account Balance
2	25	100,000
4	60	50,000
5	60	120,000
6	60	150,000

How to Handle Missing Data?

Manually Fill The Data

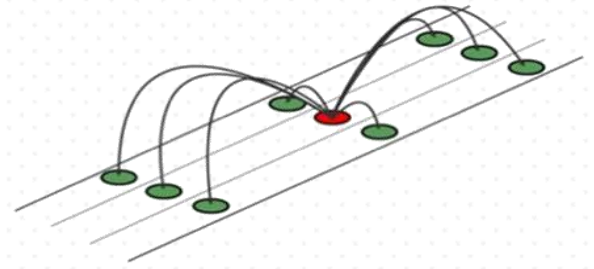
Customer	Age	Sex	Account Balance	Job
1	25	M	Missing	Student
2	25	Missing	100,000	Student
3	25	F	Missing	Missing
4	60	M	50,000	Staff
5	60	Missing	120,000	Student
6	60	Missing	150,000	Student
7	25	Missing	Missing	Student
8	25	F	100,000	Missing
9	25	M	Missing	Staff
10	60	M	50,000	Student
11	60	Missing	120,000	Student
12	60	F	150,000	Missing
13	25	F	Missing	Missing
14	25	F	100,000	Student
15	25	M	Missing	Student
16	60	F	50,000	Missing
17	60	Missing	120,000	Missing
18	60	F	150,000	Student
19	25	F	Missing	Staff
20	25	F	100,000	Missing
21	25	F	Missing	Staff
22	60	M	50,000	Student
23	60	Missing	120,000	Missing
24	60	Missing	150,000	Student

How to Handle Missing Data?

Automatically Fill The Data



Imputation



Basic

- Global Constant
- Attribute Statistics

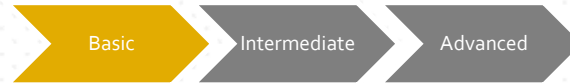
Intermediate

- Similar Instances
- Regression / Classification Imputation

Advanced

- Model Based Inference
- Multiple Imputation

Global Constant



- A global constant
 - “unknown”, -1, etc.

Customer	Age	Account Balance
1	25	Missing
2	25	100,000
3	25	Missing
4	60	50,000
5	60	120,000
6	60	150,000

Customer	Age	Account Balance
1	25	-1
2	25	100,000
3	25	-1
4	60	50,000
5	60	120,000
6	60	150,000

Attribute Statistics



- Value derived from attributes non-missing instances
 - **Mean/Median/Mode**

Customer	Age	Account Balance
1	25	Missing
2	25	100,000
3	25	Missing
4	60	50,000
5	60	120,000
6	60	150,000

Customer	Age	Account Balance
1	25	105,000
2	25	100,000
3	25	105,000
4	60	50,000
5	60	120,000
6	60	150,000

- ❖ *May result in underestimates of the standard deviation*
- ❖ *"Pull" estimates of correlation to zero*

Utilize Similar Instances



- Value derived from similar non-missing instances

Customer	Age	Account Balance
1	25	Missing
2	25	100,000
3	25	Missing
4	60	50,000
5	60	120,000
6	60	150,000

Customer	Age	Account Balance
1	25	100,000
2	25	100,000
3	25	100,000
4	60	50,000
5	60	120,000
6	60	150,000

❖ *What comprises Similarity?*

Regression / Classification Imputation

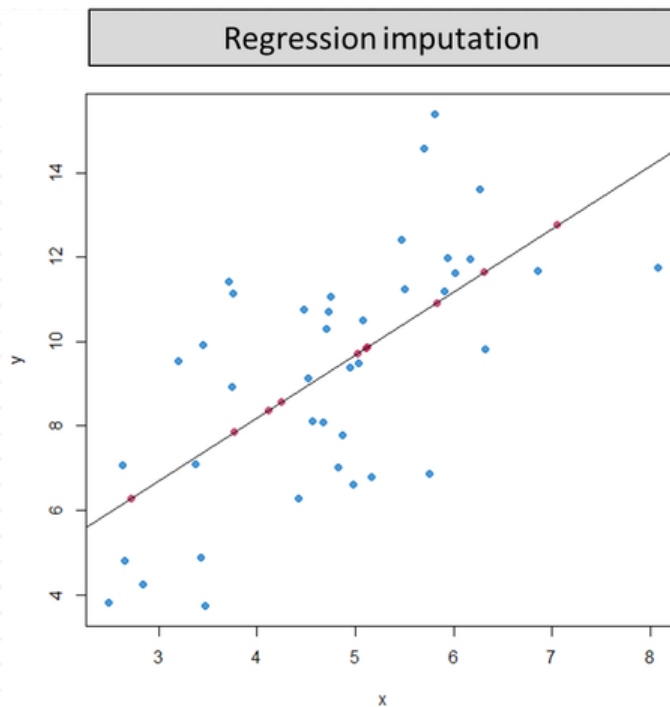
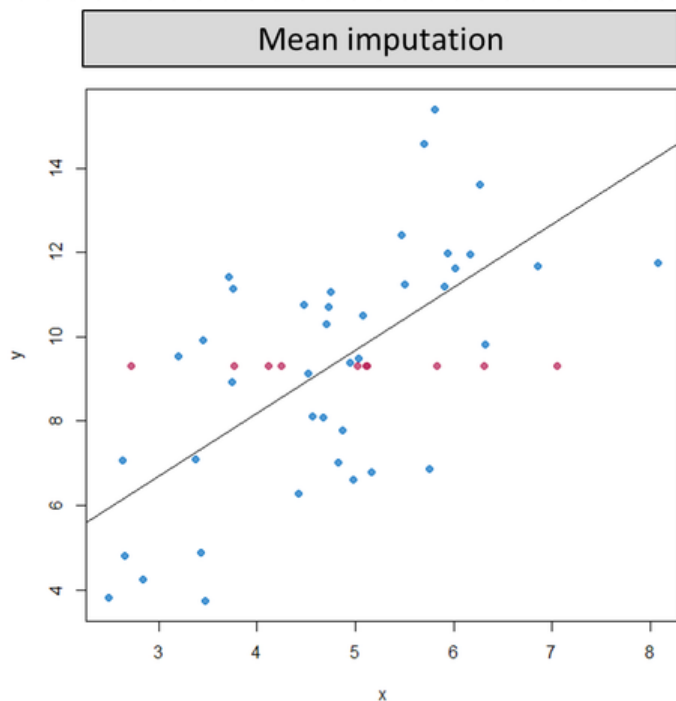


- Impute with the incomplete attribute with a model of the complete attributes.
 - Imputes the value based on other feature values.

Customer	Age	Account Balance
1	25	Missing
2	25	100,000
3	25	Missing
4	60	50,000
5	60	120,000
6	60	150,000

Customer	Age	Account Balance
1	25	100,000
2	25	100,000
3	25	100,000
4	60	50,000
5	60	120,000
6	60	150,000

Constant vs Model Based Imputation



Model Based Inference

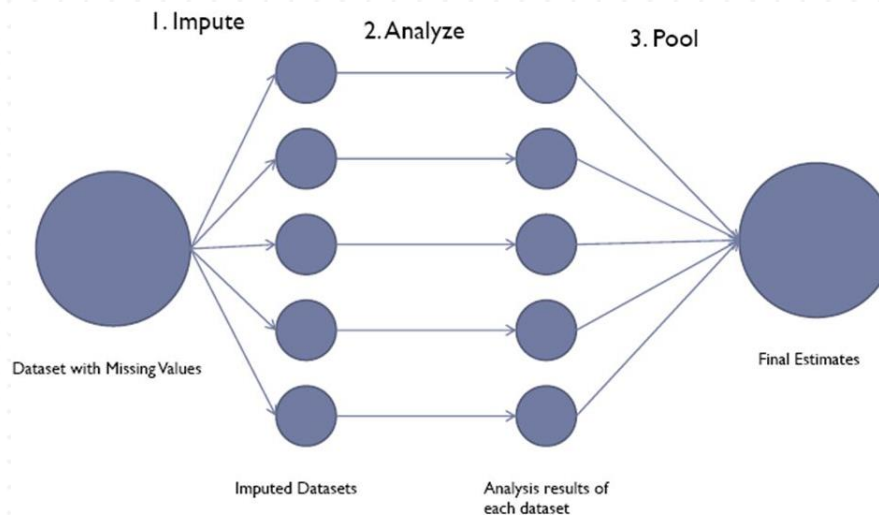


- Assign value based on the probability distribution of the non-missing data.
 - Fill with the value that is most likely to have resulted in the observed data
 - Tries to capture the “observed” empirical distribution of data.
- Common examples, EM, Maximum Likelihood estimation

Multiple Imputation



- Replace with several imputed values that reflect uncertainty in imputation.
 - Simulate uncertainty associated with the parameters of the probability distribution of the data.



iPython Examples



Is a Full Data Set Enough For Success?

100	0	0	0	0	0
0	100	0	0	0	0
0	0	92	2	6	0
0	1	0	97	0	2
0	0	5	0	95	0
0	0	0	4	0	96
88	2	0	0	0	10
0	100	0	0	0	0
0	0	49	2	49	0
0	0	0	47	0	53
0	0	4	0	96	0
0	0	0	1	0	99
78	14	0	0	0	8
0	94	0	6	0	0
0	0	11	1	88	0
0	0	0	3	1	96
0	0	3	0	97	0
0	0	0	1	1	98



Case 1: Directly Evaluated

Heart Failure Last Week		
Keith	0	
Louis	0	
Xian	1	
Chao	0	

Case 2: Didn't Happen / Wasn't Measured

	Heart Failure Last Week	High Blood Pressure
Keith	0	1
Louis	0	0*
Xian	1	0*
Chao	0	1

Case 3: Was Measurement Accurate?

	Heart Failure Last Week	High Blood Pressure	Swine Flu
Keith	0	1	0*
Louis	0	0*	1*
Xian	1	0*	0*
Chao	0	1	0*

Measurement Variation

TABLE 1. COMPARISON OF STEPS TAKEN MEASURED USING HAND COUNTING COMPARED TO STEPS TAKEN FROM THE ACTIVITY DEVICES.

Devices	Treadmill Walking	Treadmill Running	Elliptical	Agility
Actual	2425±177.9	3182±173.9	2631±371.5	805±51.9
Jawbone UP	2403±176.6	3186±171.5	2627±359.0	783±110.1
Nike Fuelband	2273±154.8*	3169±171.2	2580±458.7	533±70.4*
Fitbit Ultra	2425±177.2	2990±313.0*	2630±370.6	645±90.0*
NL-2000i	2425±178.0	2869±247.1*	2477±471.1*	671±106.9*

Values represent means ± standard deviation.

*Significantly different than actual steps ($p < .05$).

TABLE 3. COMPARISON OF CALORIC EXPENDITURE MEASURED USING THE PORTABLE METABOLIC GAS ANALYZER COMPARED TO KCAL VALUES OBTAINED FROM THE ACTIVITY DEVICES.

Devices	Treadmill Walking (n=19)	Treadmill Running (n=18)	Elliptical (n=20)	Agility (n=20)
Actual	109±19.6	240±47.3	161±25.6	90±20.7
Jawbone UP	123±25.2	288±63.6*	161±74.1	63±23.5*
Nike Fuelband	107±24.2	275±56.4*	118±38.0*	77±18.0*
Fitbit Ultra	111±22.8	230±50.5	154±34.1	75±19.2*
Adidas MiCoach	146±18.2*	261±52.4	--	36±6.8*
BodyMedia FIT Core	112±16.2	210±37.2	129±19.5*	74±19.2*

Values represent means ± standard deviation.

*Significantly different than portable metabolic gas analyzer kcal ($p < .05$).

Broadly: Noisy Data

Noise: random error or variance in a measured variable

- Attribute values may vary be due to
 - Faulty data collection instruments
 - Data transmission problems
 - Technology limitation
 - Inconsistency in naming convention

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
 - Faulty data collection instruments
 - Data transmission problems
 - Technology limitation
 - Inconsistency in naming convention
- Other data problems
 - Duplicate records
 - Incomplete data
 - Inconsistent data

How to Handle Noisy Data?

- Binning
 - First sort data and partition into bins to get a more general grouping
 - Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Discretization: Equal Width Binning

- Divides the range into N intervals of equal size
- If A and B are the lowest and highest values of the attribute, the width of intervals will be:

$$W = \frac{(B - A)}{N}$$

- The most straight-forward
- But outliers may dominate presentation
- Skewed data is not handled well.

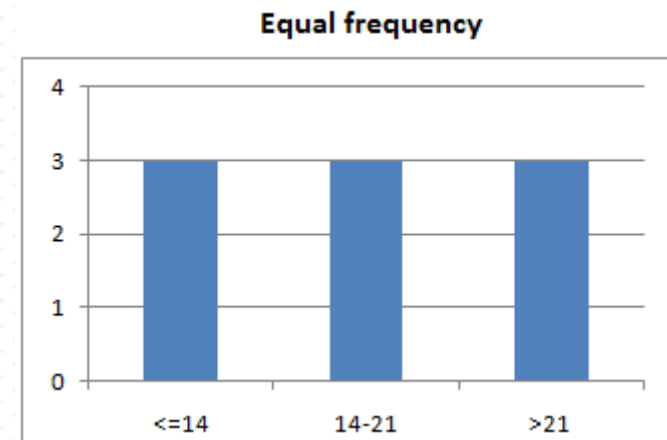
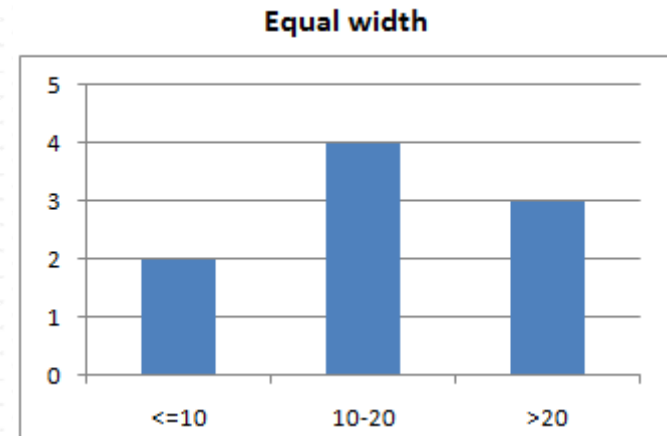
Discretization: Equal Frequency Binning

- It divides the range of size N into K intervals, each containing approximately the same number of samples
- Good data scaling.
- Managing categorical features can be tricky.

Binning Example

Data = 0,4,12,16,16,18,24,28

- Equal Width
 - Bin 1: 0,4 $[-\infty, 10)$
 - Bin 2: 12,16,16,18 $[10, 20)$
 - Bin 3: 24,26,28 $[20, \infty)$
- Equal Frequency
 - Bin 1: 0,4,12 $[-\infty, 14)$
 - Bin 2: 16,16,18 $[14, 21)$
 - Bin 3: 24,26,28 $[21, \infty)$



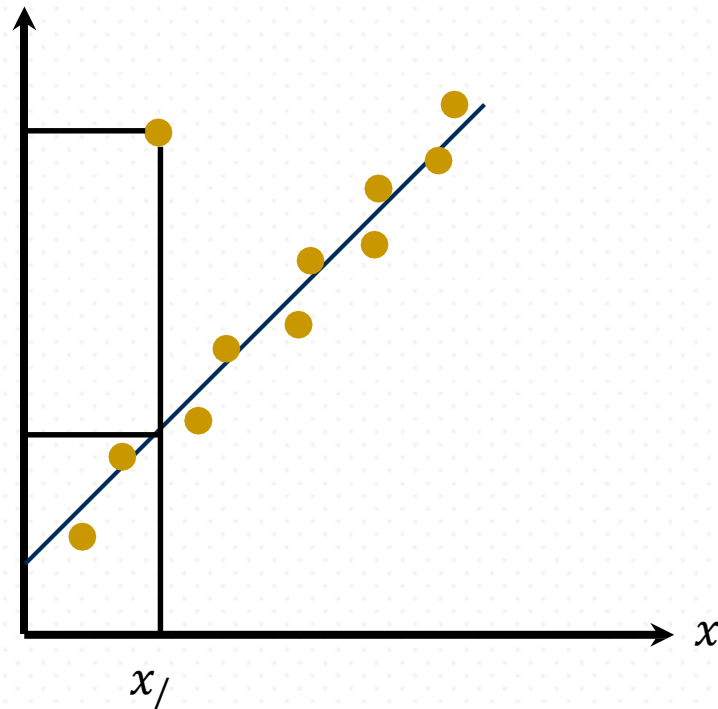
Binning Smoothing Example

Data = 4, 8, 15, 21, 21, 24, 25, 28, 34

- Equal Frequency
 - Bin 1: 4, 8, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 28, 34
- *Smoothing by Bin Mean*
 - Bin 1: 9, 9, 9
 - Bin 2: 22, 22, 22
 - Bin 3: 29, 29, 29
- *Smoothing by Boundaries*
 - Bin 1: 4, 4, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 25, 34

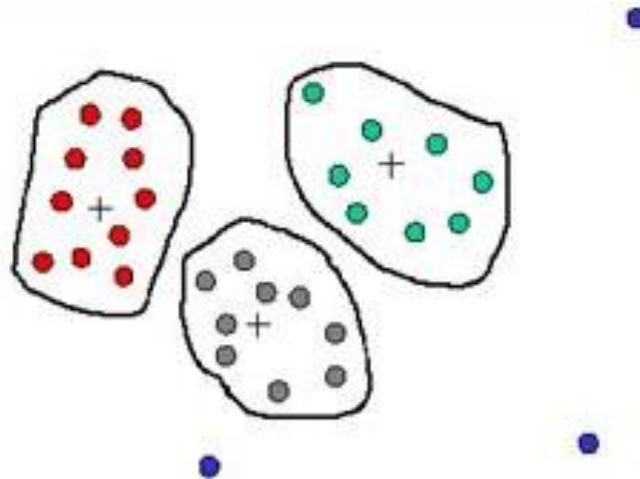
Regression

Smooth by fitting the data into regression functions



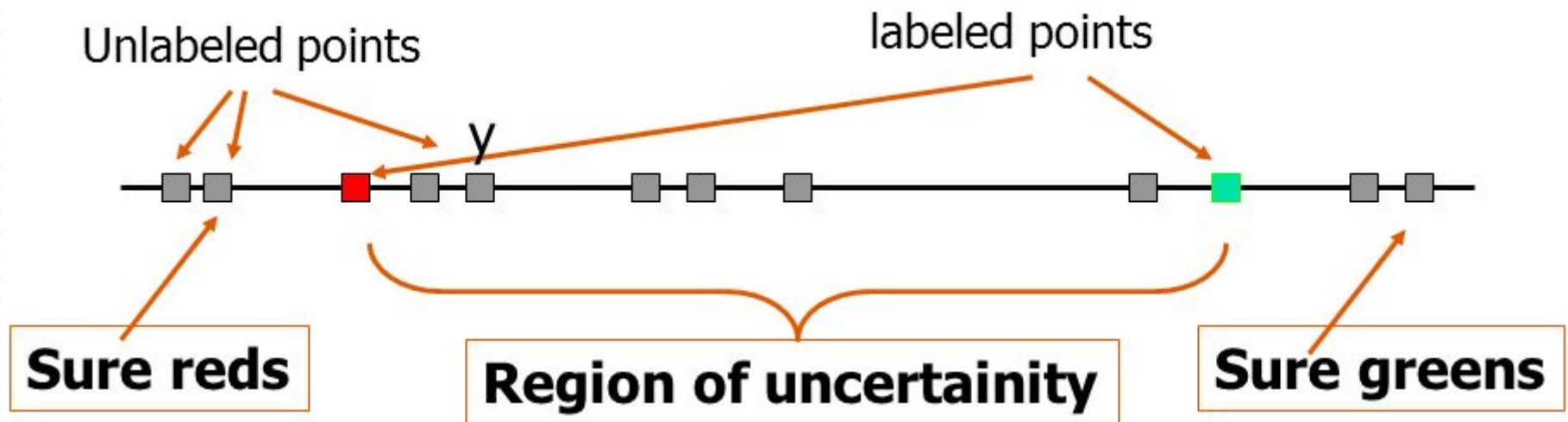
Clustering

Group Data
Detect and remove outliers

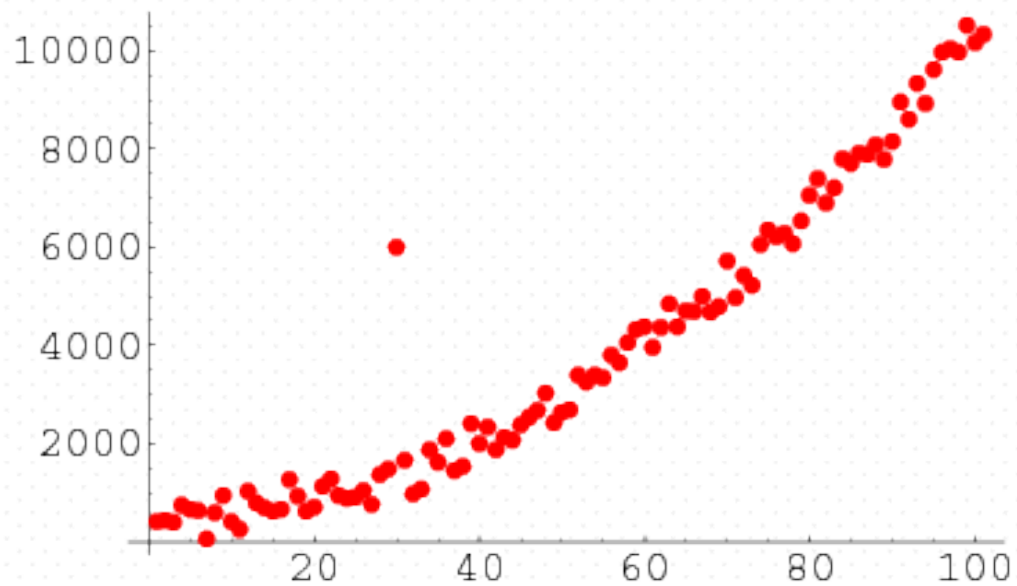


Semi-supervised

- Combined computer and human inspection
 - Detect suspicious values and check by human (e.g., deal with possible outliers)

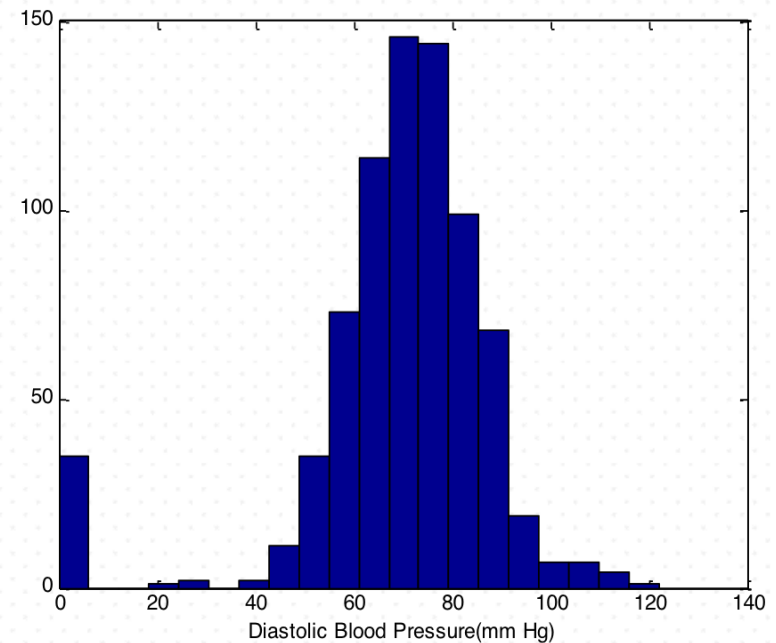


Outliers!



Removing Outliers

- Proximity-based techniques:
 - Define a proximity measure between instances, with outliers being distant from most other instances.



Removing Outliers

- Statistical techniques:
 - Evaluate the value of the instance in relation to an expected distribution

1.5-IQR



Z-Score

$$Z_i = \frac{X_i - \bar{X}}{s}$$

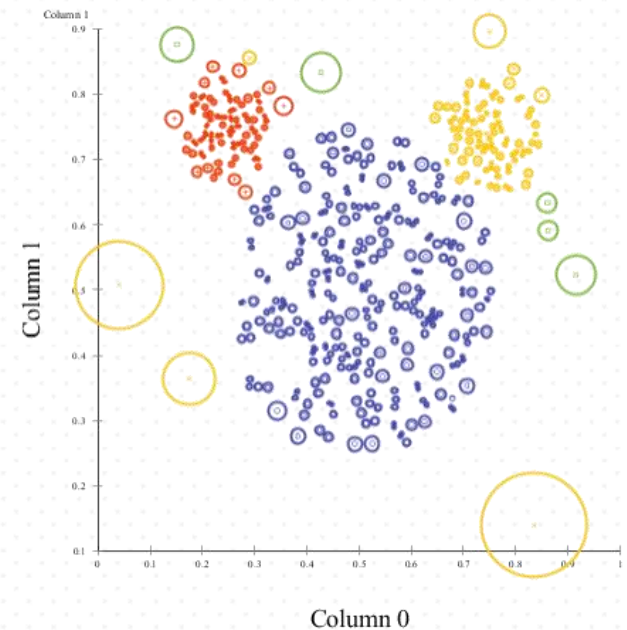
Median Average
Distribution

$$MAD = \text{Median}(|X_i - \text{median}(X)|)$$

$$\text{OutlierScore} = \frac{0.6745(x_i - \bar{x})}{MAD}$$

Removing Outliers

- Density-based techniques:
 - Define outliers as instances that have a local density significantly less than that of neighbors.

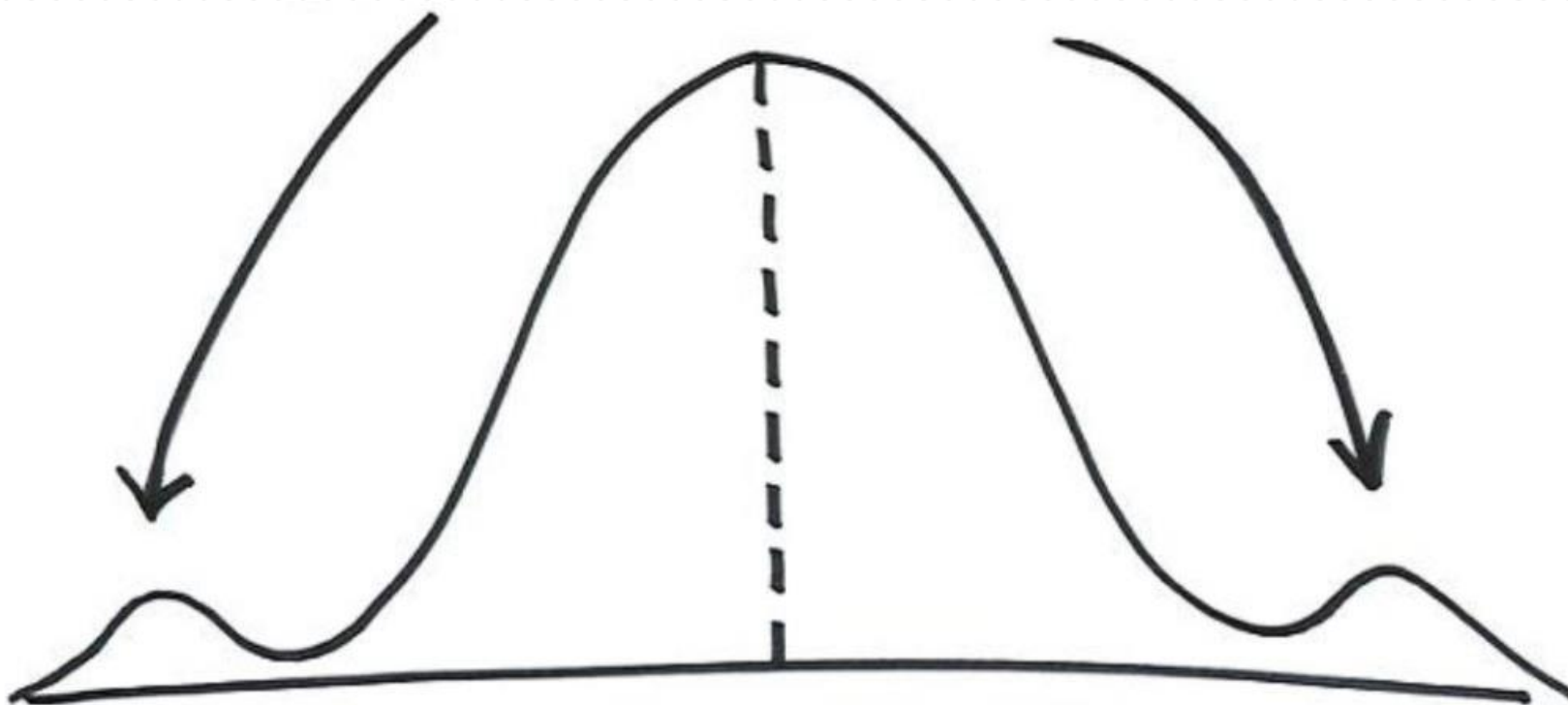


Outlier Removal Methods

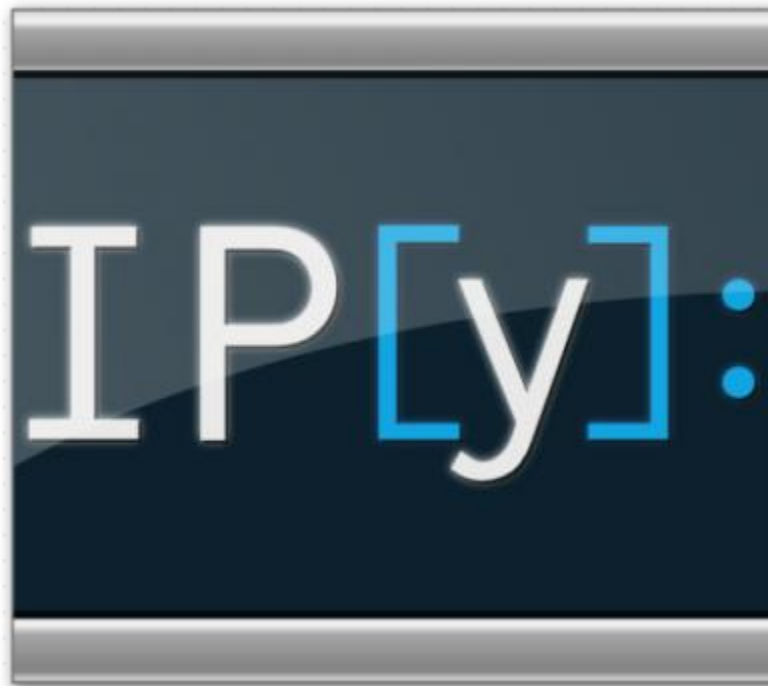
- Over 100 outlier tests
 - Data distribution
 - Distributional parameters
 - Number of expected outliers
 - Types of outliers

Be Careful...

Outliers are often considered error or noise, but may carry important information.



iPython Examples



Data Preprocessing

- Data cleaning
- **Data integration**
- Data reduction
- Dimensionality reduction

Data Integration

- Data integration
 - Combining data from **multiple sources** into a coherent store
- Schema integration: e.g., $A.cust-id \equiv B.cust-\#$
 - Integrate metadata from different sources
- **Entity identification:**
 - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton



Inconsistent Data

- Data can contain inconsistent values.
 - e.g., an address field with both ZIP code and city, but where the specified ZIP code area is not in the specified city.
- Some inconsistencies are easy to detect; some may require consulting an external source.



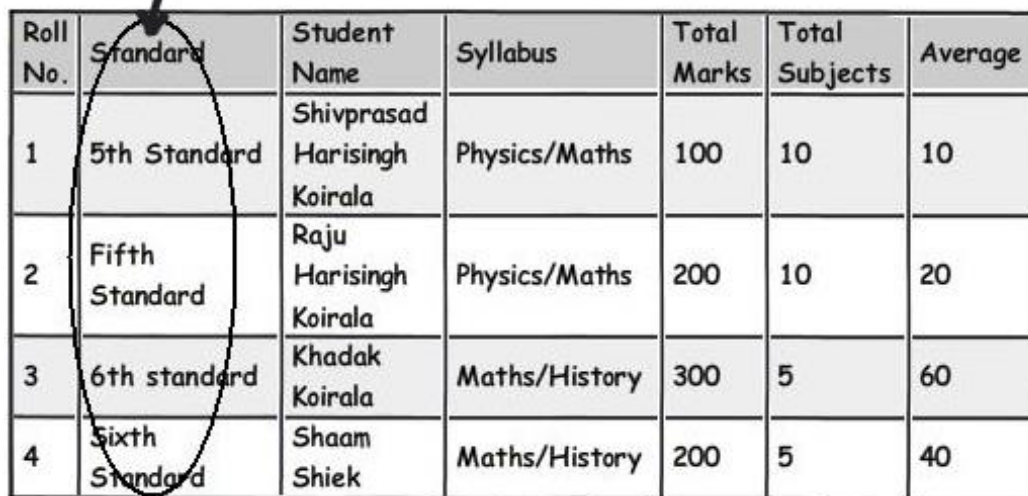
Some Are Simple to Find

	Age	Sex	BP
Keith	26	M	120/80
Louis	25	M	130/85
Xian	25	F	13/85
Chao	26	M	125/90

Data Redundancy

For the same real world entity, attribute values from different sources are different

Duplicate Data



Roll No.	Standard	Student Name	Syllabus	Total Marks	Total Subjects	Average
1	5th Standard	Shivprasad Harisingh Koirala	Physics/Maths	100	10	10
2	Fifth Standard	Raju Harisingh Koirala	Physics/Maths	200	10	20
3	6th standard	Khadak Koirala	Maths/History	300	5	60
4	Sixth Standard	Shaam Shiek	Maths/History	200	5	40

Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
 - *Object identification*: The same attribute or object may have different names in different databases
 - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis (often for categorical attributes)* and *covariance analysis (often for numerical attributes)*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Redundancy Identification: Categorical Data

Can Be Achieved with Correlation Analysis

	Play chess	Don't play chess
Like science fiction		
Don't like science fiction		

Correlation Analysis: Observed Counts

	Play chess	Don't play chess	Sum (row)
Like science fiction	250	200	450
Don't like science fiction	50	1000	1050
Sum(col.)	300	1200	1500

Correlation Analysis: Expected Value

	Play chess	Don't play chess	Sum (row)
Like science fiction	90	360	450
Don't like science fiction	210	840	1050
Sum(col.)	300	1200	1500

Like Science Fiction – Play Chess = $(450/1500)*300 = 90$

Like Science Fiction – Don't Play Chess = $(450/1500)*1200 = 360$

Don't Like Science Fiction – Play Chess = $(1050/1500)*300 = 210$

Don't Like Science Fiction – Don't Play Chess = $(1050/1500)*1200 = 840$

Correlation Analysis

- X² (chi-square) test:**

$$\chi^2 = \sum_i^n \frac{\overset{\text{observed}}{\downarrow} (O_i - E_i)^2}{\underset{\text{expected}}{E_i}}$$

- Null hypothesis:** The two distributions are independent
- The cells that contribute the most to the X² value are those whose actual count is different from the expected count
 - The larger the X² value, the more the null hypothesis of independence is rejected, and the more likely the variables are related

	Play chess	Don't play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Don't like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

Example: Chi-Square Calculation

	Play chess	Don't play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Don't like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

We can reject the null hypothesis of independence at a confidence level of 0.001.

- It shows that like_science_fiction and play_chess are correlated.

Example: Chi-Square Calculation

Degrees of freedom (df)	χ^2 value ^[19]											
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83	
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82	
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27	
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47	
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52	
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46	
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32	
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12	
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88	
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59	
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001	

Correlation does not imply causality

- Ex. # of hospitals and # of car-theft in a city are correlated
 - Both are causally linked to the third variable: population
- Causal analysis

Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach by K. Kuang, M. Jiang, P. Cui, J. Sun, S. Yang. IEEE Transactions on Big Data (TBD), 2017.

Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing by K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2017.

Variance for Single Variable (Numerical Data)

- The variance of a random variable X provides a measure of how much the value of X deviates from the mean or expected value of X :

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- where σ^2 is the variance of X , σ is called *standard deviation*
 μ is the mean, and $\mu = E[X]$ is the expected value of X
- That is, variance is the expected value of the square deviation from the mean
- It can also be written as:

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2
$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1X_2] - \mu_1\mu_2 = E[X_1X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1 ; similarly for μ_2
- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance:** If $\sigma_{12} < 0$
- **Independence:** If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
 - Some pairs of random variables may have a covariance 0 but are not independent
 - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
 - $(2, 5), (3, 8), (5, 10), (4, 11), (6, 14)$
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
 - $(2, 5), (3, 8), (5, 10), (4, 11), (6, 14)$
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula
$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$
- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
 - $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
 - $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
 - $\sigma_{12} = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

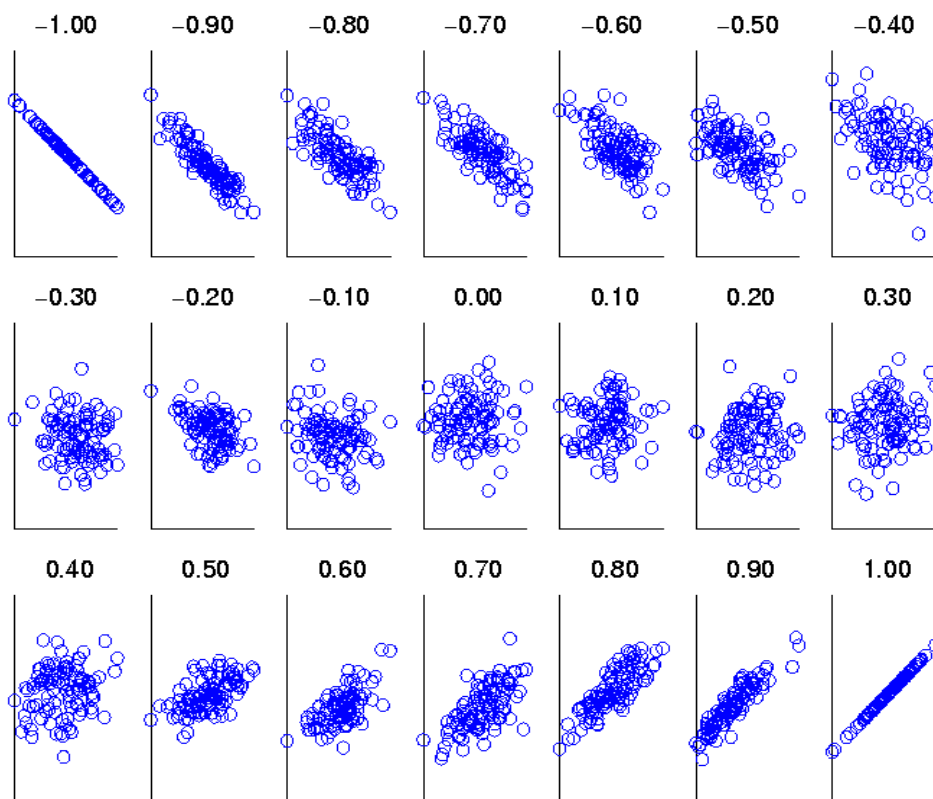
- **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- If $\rho_{12} > 0$: A and B are positively correlated (X_1 's values increase as X_2 's)
 - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient

- Correlation coefficient value range: $[-1, 1]$ Can you prove the range?
- A set of scatter plots shows sets of points and their correlation coefficients changing from -1 to 1



Covariance Matrix

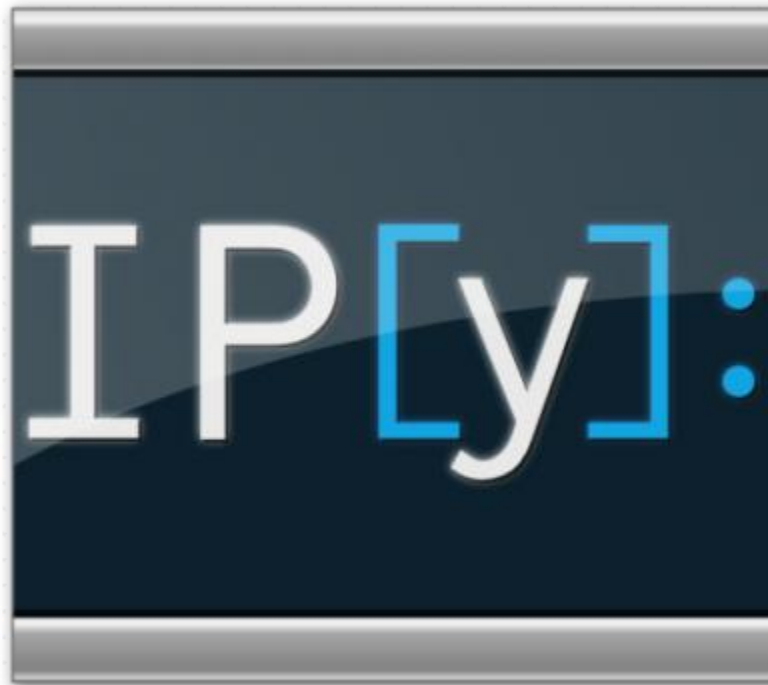
- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2×2 covariance matrix as

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & X_2 - \mu_2 \end{pmatrix}\right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}$$

- Generalizing it to d dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

iPython Examples



Discussion

- Can you use Chi-Square or p-value (doing correlation analysis) to select meta paths (as features) for relationship prediction?

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Comm. of ACM*, 42:73-78, 1999
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. [Mining Database Structure; Or, How to Build a Data Quality Browser](#). SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995