NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

3.0 Descriptive Analytics

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

@thephdataminer

Module 1 of the Business Intelligence and Analytics Track of UP NEC and the UP Center of Business Intelligence

Module 1 Outline

1. Intro to Business Intelligence
   – Case Study on Selecting BI Projects
2. Data Warehousing
   – Case Study on Data Extraction and Report Generation
3. **Descriptive Analytics**
   – **Case Study on Data Analysis**
4. Visualization
   – Case Study on Dashboard Design
5. Classification Analysis
   – Case Study on Classification Analysis
6. Regression and Time Series Analysis
   – Case Study on Regression and Time Series Analysis
7. Unsupervised Learning and Modern Data Mining
   – Case Study on Text Mining
8. Optimization for BI

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- Data Transformations
- Pivot Tables
- Case Study

eljalao@up.edu.ph

What is Data?

Definition 3.1: Data

- Collection of **objects** and their **attributes**

Definition 3.2: Objects

- An object is physical/conceptual **entity of interest**
  - Examples: customers, orders, accounts
  - Object is also known as record, tuple, case, sample, or instance

Definition 3.3: Attributes

- An attribute is a **property** or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature

eljalao@up.edu.ph

What is Data?

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Figure 3.1: A Dataset

What is Data?

Definition 3.4: Attribute Values

- Attribute values are **numbers or symbols** assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to **different attribute values**
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the **same set of values**
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

What is Data?

- There are different **types of attributes**
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temp in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

What is Data?

- Properties of Attribute Values
  - The type of an attribute depends on which of the following **properties** it possesses:
    - Distinctness: $=$ $\neq$
    - Order: $<$ $>$
    - Addition: $+$ $-$
    - Multiplication: $*$ $/$
    - Nominal attribute: distinctness
    - Ordinal attribute: distinctness & order
    - Interval attribute: distinctness, order & addition
    - Ratio attribute: all 4 properties

What is Data?

Definition 3.5: Discrete Attributes

- A discrete attribute can only have a **finite set** of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as **integer variables**.
  - Note: binary attributes are a special case of discrete attributes

What is Data?

Definition 3.6: Continuous Attributes

- Continuous Attributes have **real numbers** as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - **Continuous attributes** are typically represented as floating-point variables.

Outline for this Session

- What is Data?
- **Types of Datasets**
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- Data Transformations
- Pivot Tables
- Case Study

Types of Datasets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
  - Time Series

Types of Datasets: Record Data

Definition 3.7: Record Data

- Data that consists of a **collection of records**, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 3.2: Example of a Record Dataset

Types of Datasets: Record Data

Definition 3.8: Record Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a **multi-dimensional space**, where each dimension represents a distinct attribute
- Such data set can be represented by an $\boldsymbol{m \times n}$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Figure 3.3: Example of a Data Matrix Dataset

Types of Datasets: Record Data

Definition 3.9: Term by Document Dataset

- Dataset the where value of each attribute is the **number of times** the corresponding term occurs in the object.
  - Each term is a component (attribute) of the vector,
  - Extension of a data matrix

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Figure 3.4: Example of a Document Dataset

Types of Datasets: Record Data

Definition 3.10: Market Basket/Transaction Dataset

- A special type of record data, where
  - each record (transaction) involves **a set of items**.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

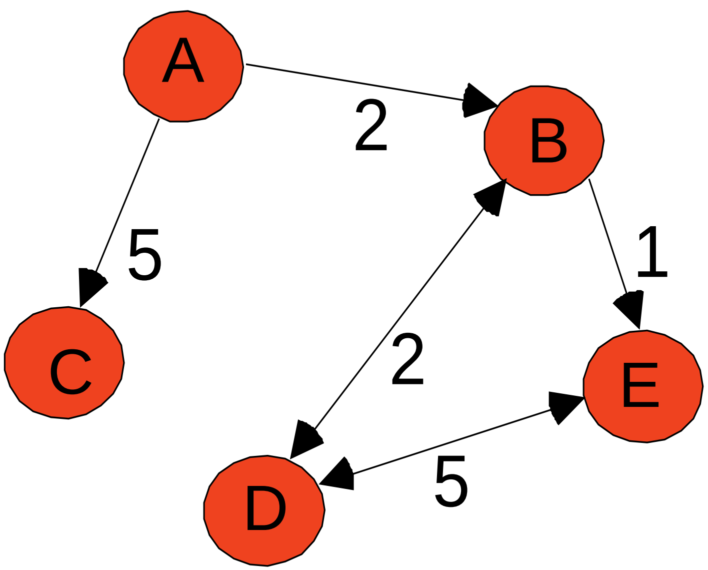
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Figure 3.5: Example of a Transaction Dataset

Types of Datasets: Graph Data

Definition 3.11: Graph Dataset

- Dataset that shows the **interactions** of multiple entities



	A	B	C	D	E
A	0	2	5		
B		0		2	1
C			0		
D		2		0	5
E				5	0

Figure 3.6: Example of a Graph Dataset

Types of Datasets: Ordered Data

Definition 3.12: Sequence Transactions Dataset

- Dataset transactions **over time** grouped by elements

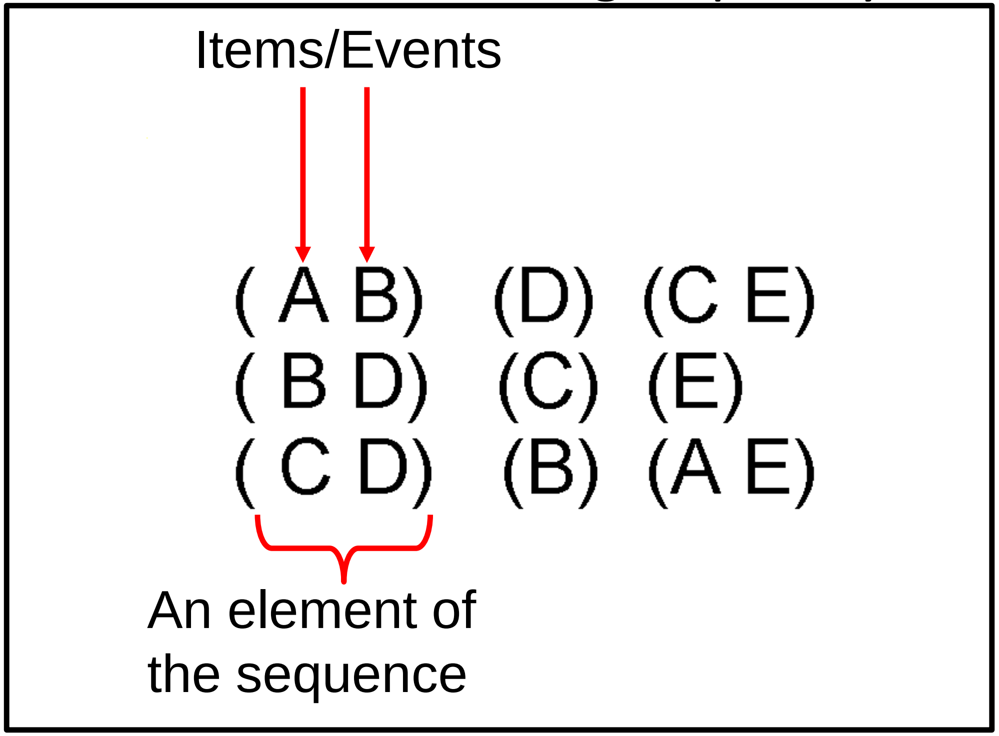


Figure 3.7: Example of a Sequence Dataset

Types of Datasets: Ordered Data

Definition 3.13: Time Series Dataset

- A single attribute of interest **over time**

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2001	48	58	57	65
2002	50	61	59	68
2003	52	62	59	69
2004	52	64	60	73
2005	53	65	60	75

Figure 3.8: Example of a Time Series Dataset

Outline for this Session

- What is Data?
- Types of Datasets
- **Descriptive Statistics**
- Data Preprocessing
- Data Cleaning
- Data Transformations
- Pivot Tables
- Case Study

Descriptive Statistics

Definition 3.14: Descriptive Statistics

- Descriptive Statistics are used by analysts to report on **populations and samples**
- Descriptive statistics **speed up and simplify** comprehension of a group's characteristics
- Descriptive statistics are a collection of measurements of two things: **location and variability.**
  - Location tells you the central value of your variable (the mean is the most common measure).
  - Variability refers to the spread of the data from the center value (i.e. variance, standard deviation).

Sample vs. Population

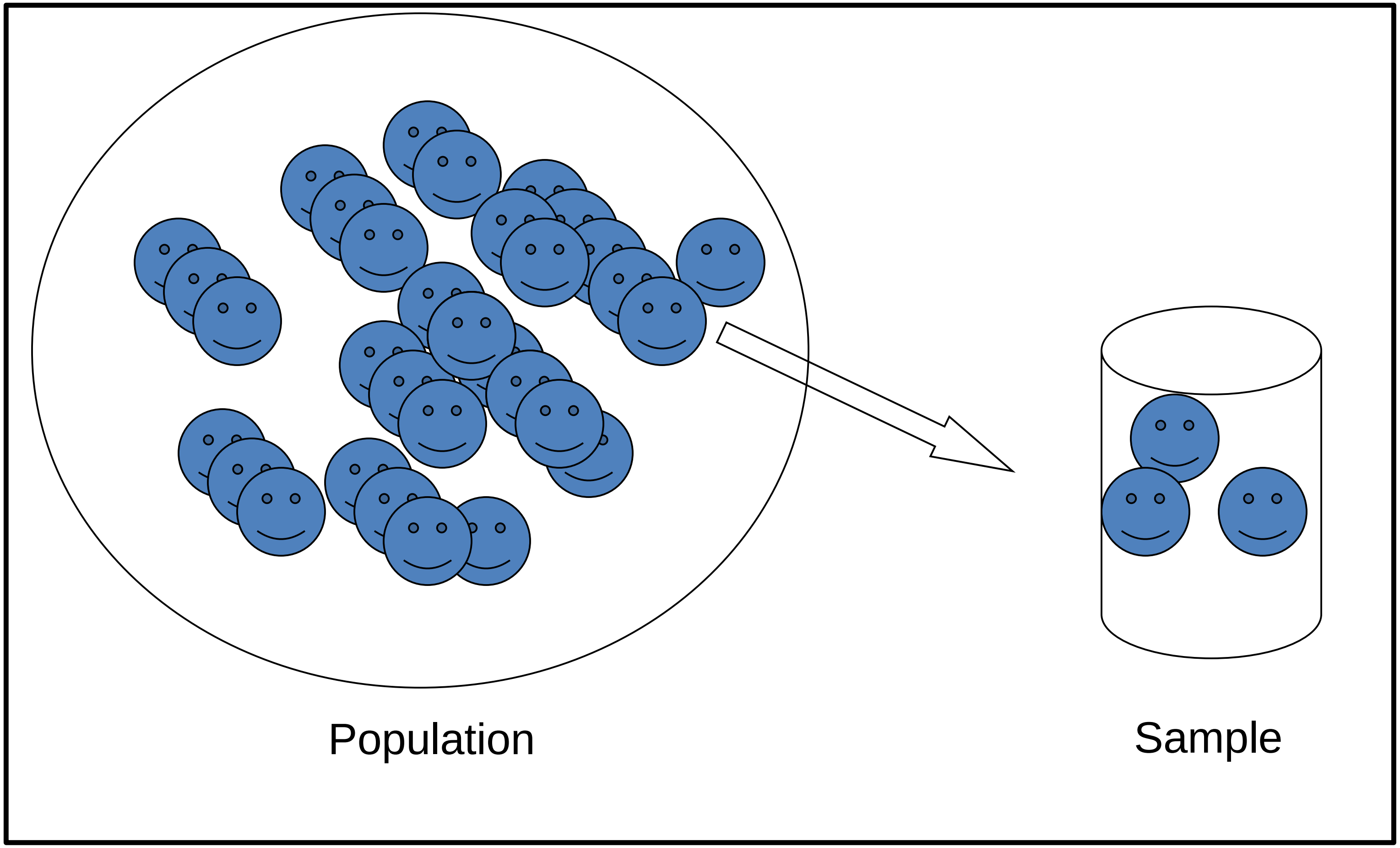


Figure 3.9: Sample versus Population

Descriptive Statistics

- **Types** of descriptive statistics:
  - **Summarizing Data:**
    - Central Tendency (or Groups' "Middle Values")
      - Mean
      - Median
      - Mode
    - Variation (or Summary of Differences Within Groups)
      - Range
      - Interquartile Range
      - Variance
      - Standard Deviation
  - **Organize Data**
    - Graphs
      - Summary Charts
      - Bar Chart or Histogram
      - Box Plots and Dot Plots
      - Scatter Plots

Descriptive Statistics

Example 3.1 Which Group is smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Descriptive Statistics

Definition 3.15: Mean

- Most commonly called the "**average**."
- Calculated using equation 3.1

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

- Add up the values for each case and divide by the total number of cases.

Mean

Example 3.1: (Cont.) Which Group is smarter?

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1437$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1437}{13} = 110.54$$

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1433$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1433}{13} = 110.23$$

Descriptive Statistics

- Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

- With a summary descriptive statistic, it is much easier to answer generalization questions.

Descriptive Statistics

- The mean is the "**balance point**."
- Each person's score is like 1 pound placed at the score's position on a see-saw. Below, on a 200 cm see-saw, the mean equals 110, where a fulcrum finds balance:

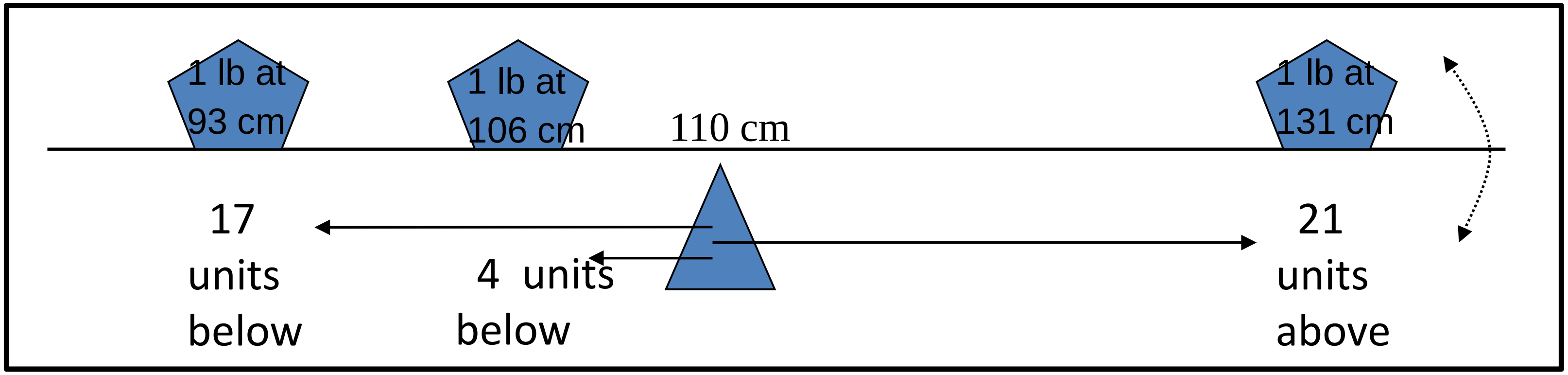


Figure 3.10: Mean Balance Point

- The scale is **balanced** because 17 + 4 on the left = 21 on the right

Descriptive Statistics

- Means can be badly affected by **outliers** (data points with extreme values unlike the rest)
- Outliers can make the mean a **bad measure** of central tendency or common experience

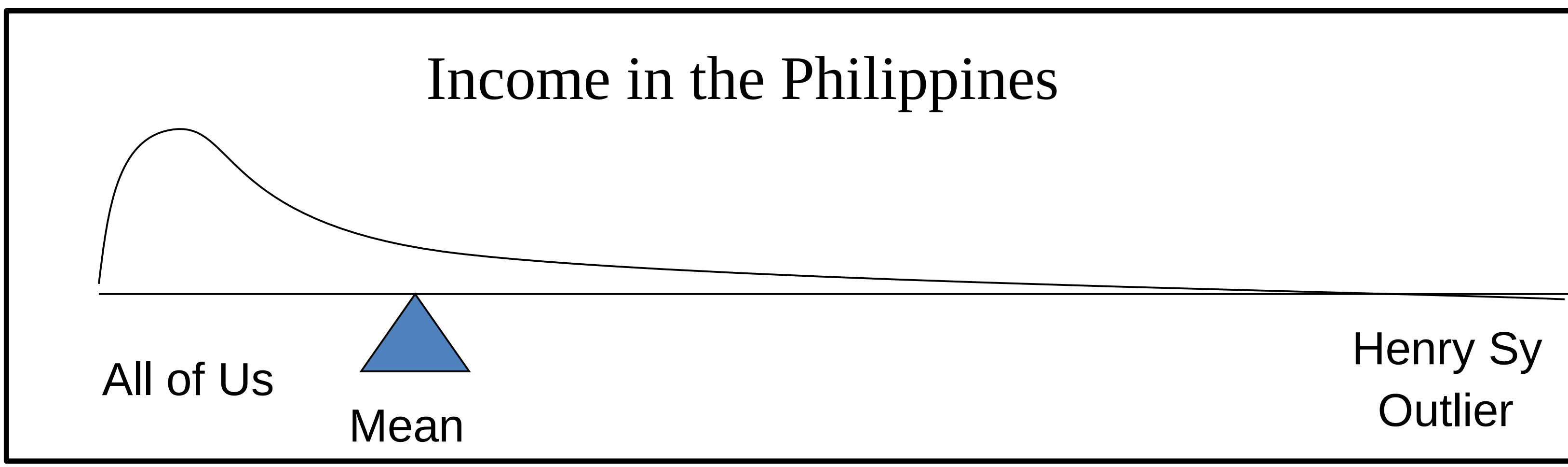


Figure 3.11: Outliers

Descriptive Statistics

Definition 3.16: Median

- The **middle value** when a variable's values are ranked in order; the point that divides a distribution into two equal halves.
- When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.
- Median is the 50th percentile.

Descriptive Statistics

Example 3.1: (Cont.) Which Group is smarter?

- Class A--IQs of 13 Students

89
93
97
98
102
106
109
110
115
119
128
131
140

Median = 109

(six cases above, six below)

Descriptive Statistics

- If the first student were to drop out of Class A, there would be a new median:

~~89~~
93
97
98
102
106
109
110
115
119
128
131
140

Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

Descriptive Statistics

- The median is **unaffected by outliers**, making it a better measure of central tendency, better describing the typical person than the mean when data are skewed.

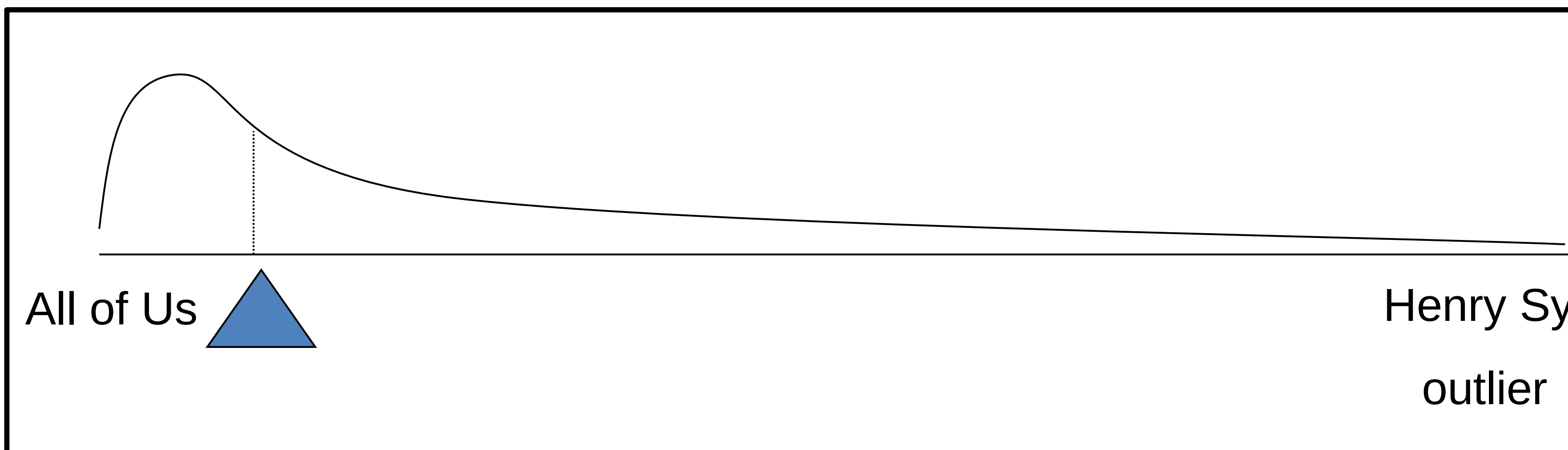


Figure 3.12: Outliers

Descriptive Statistics

- If the recorded values for a variable form a **symmetric** distribution, the **median and mean are identical.**
- In skewed data, the mean lies further toward the skew than the median.

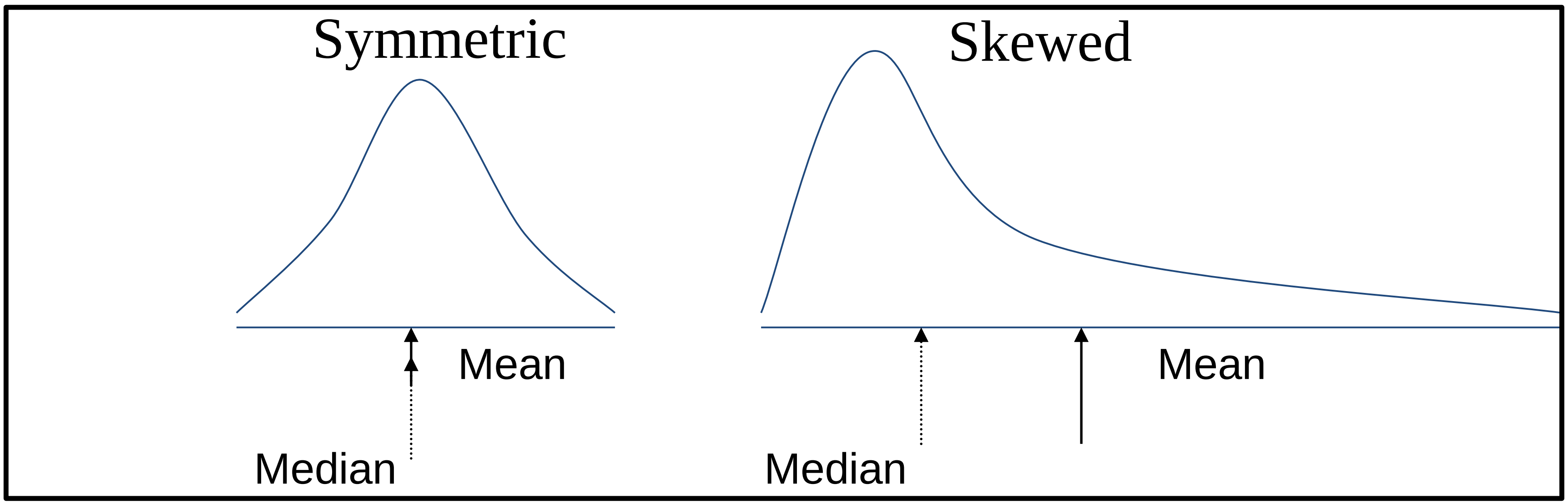


Figure 3.13: Skewed Data

Descriptive Statistics

Definition 3.17: Mode

- The **most common** data point is called the mode.
- The combined IQ scores for Classes A & B:

80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120 127 128 131 131 140 162

mode

- It is **possible** to have more than one mode

Descriptive Statistics

- It may give the most likely experience rather than **the typical or central** experience.
- In **symmetric data**, the mean, median, and mode are the same. In skewed data, the mean and median lie **further toward the skew** than the mode.

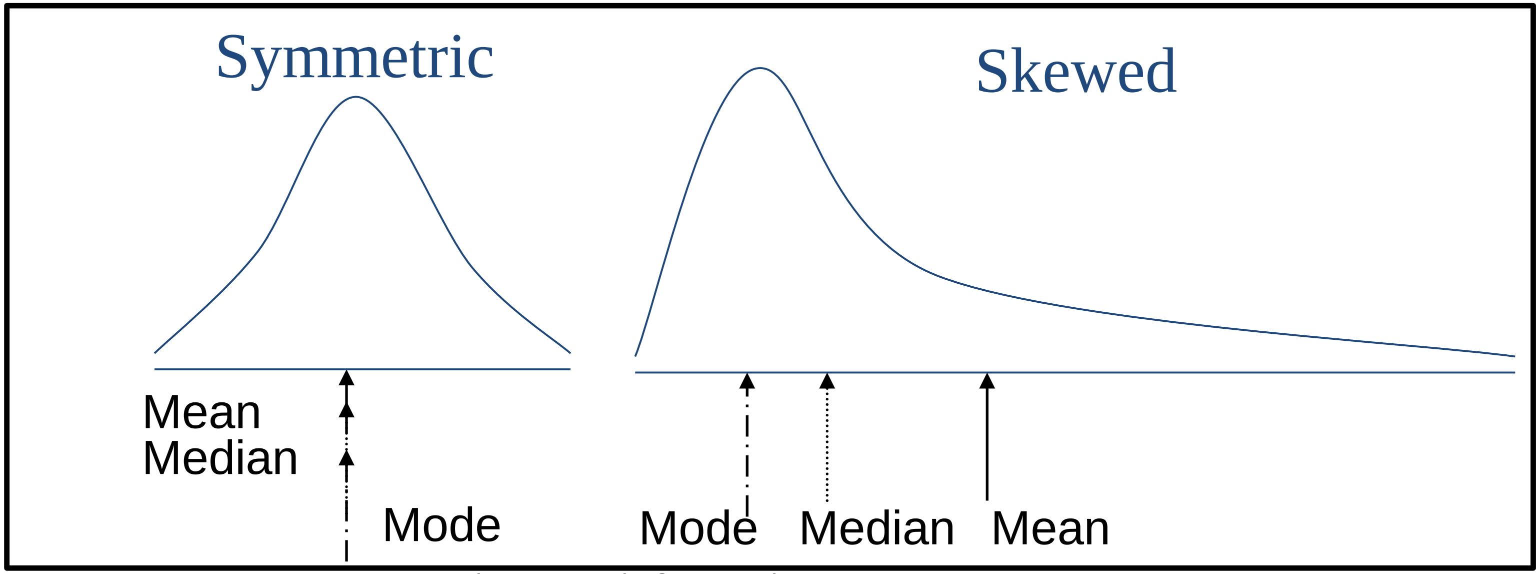


Figure 3.14: Skewed Data

Descriptive Statistics

Definition 3.18: Range

- The **spread or the distance**, between the lowest and highest values of a variable.

$$Range = \max(x) - \min(x) \quad (3.2)$$

- To get the range for a variable, you subtract its lowest value from its highest value.

Descriptive Statistics

Example 3.1: (Cont.) Which Group is smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	**162**
128	109	131	103
131	**89**	96	111
98	106	**80**	109
140	119	93	87
93	97	120	105
110		109	

Class A Range = 140 - 89 = 51

Class B Range = 162 - 80 = 82

Descriptive Statistics

Definition 3.19: Quartile

- **A quartile** is the value that marks **one of the divisions** that breaks a series of values into four equal parts.
- **25^{th} percentile** is a quartile that divides the first ¼ of cases from the latter ¾. **75^{th} percentile** is a quartile that divides the first ¾ of cases from the latter ¼.

Definition 3.20: Interquartile Range

- **The interquartile range** is the distance or range between the 25^{th} percentile and the 75^{th} percentile.

Descriptive Statistics

Example 3.1: (Cont.) Which Group is smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	
Q1 = 97.50		**Q1 = 94.50**	
Q3 = 123.5		**Q3 = 123.5**	
IQR = 26		**IQR = 29**	

Descriptive Statistics

Definition 3.21: Variance

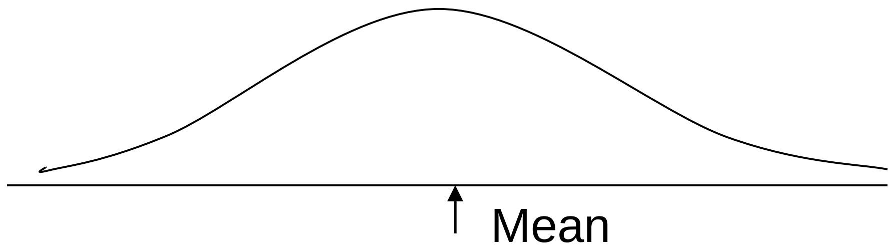
- A measure of the **spread of the recorded values** on a variable.

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad (3.3)$$

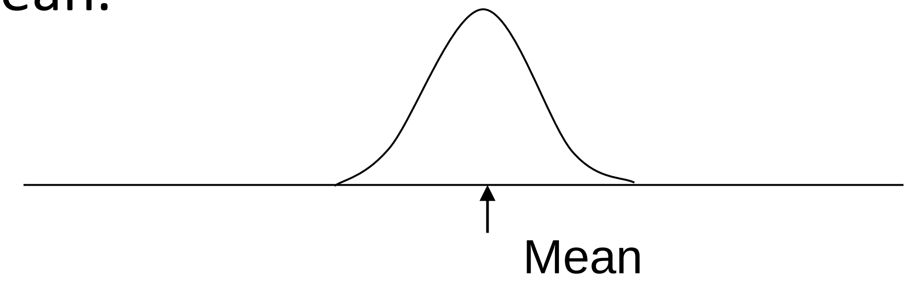
- It's a measure of dispersion.

Descriptive Statistics

- The larger the variance, the **further** the individual cases are from the mean.



- The smaller the variance, the **closer** the individual scores are to the mean.



Descriptive Statistics

- Variance is a number that at first seems complex to calculate.
- Calculating variance starts with a **deviation**.
- A deviation is the distance **away from the mean** of a record's score.
- $x_i - \bar{x}$

If the average person's car costs $20,000, my deviation from the mean is - $14,000!

6K - 20K = -14K

Descriptive Statistics

Example 3.1: (Cont) Which Group is smarter?

- The deviation of 102 from 110.54 is?
  - **102 - 110.54 = -8.54**
- Deviation of 115?
  - **115 - 110.54 = 4.46**

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Descriptive Statistics

- Squaring the deviations will eliminate negative signs
- Deviation Squared: $(\mathrm{x}_i - \bar{x})^2$
- Back to the IQ example,
- Deviation squared for 102 is:
  - **$(102 - 110.54)^2 = (-8.54)^2 = 72.93$**
- Deviation squared for 115 is:
  - **$(115 - 110.54)^2 = (4.46)^2 = 19.89$**

Descriptive Statistics

- If you were to add all the squared deviations together, you'd get what we call the **Sum of Squares**.
- Sum of Squares

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (3.4)$$

Descriptive Statistics

Class A, sum of squares:

$(102 - 110.54)^2 + (115 - 110.54)^2 +$

$(126 - 110.54)^2 + (109 - 110.54)^2 +$

$(131 - 110.54)^2 + (89 - 110.54)^2 +$

$(98 - 110.54)^2 + (106 - 110.54)^2 +$

$(140 - 110.54)^2 + (119 - 110.54)^2 +$

$(93 - 110.54)^2 + (97 - 110.54)^2 +$

$(110 - 110.54)$ **= SS = 2825.39**

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar = 110.54

Descriptive Statistics

- The last step…
- The approximate **average** sum of squares is the variance.
- $\frac{SS}{N}$ = Variance for a population.
- $\frac{SS}{n-1}$ = Variance for a sample.
- Variance = $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
- For Class A, Variance = 2825.39 / n - 1

  = 2825.39 / 12 = 235.45

Descriptive Statistics

Definition 3.22: Standard Deviation

- The square root of the variance reveals the average deviation of the observations from the mean. It's the spread of the data in the **original unit of measure** of the variable.
- Standard Deviation $= \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$
- For Class A, the standard deviation is: $\sqrt{235.45} = 15.34$
- The **average of person's deviation** from the mean IQ of 110.54 is 15.34 IQ points.

Descriptive Statistics

Definition 3.23: Correlation

- Correlation measures the **linear relationship** between attributes x and y.

$$\rho(x, y) = \frac{\sum_{i=1}^{n} x_i * y_i}{n - 1} \tag{3.5}$$

  - $\rho(x, y) = 1$, Positively Correlated,
  - $\rho(x, y) = 0$, no correlation
  - $\rho(x, y) = -1$, Negatively Correlated

Descriptive Statistics

Definition 3.24: Standardization

- To compute correlation, standardization of data attributes must be **done first**
- Process of **removing the unit of measure** of the variable

$$x' = \frac{\left(x_i - mean(x)\right)}{std(x)} \tag{3.6}$$

Descriptive Statistics

Example 3.2: Correlation Example

x	y
0.80	1.59
0.48	0.21
0.66	0.07
0.73	1.05
0.70	0.21
0.20	0.39
0.20	0.25
0.06	0.06
0.07	0.07
0.91	0.99

Mean	0.48	0.49
Std Dev	0.32	0.53

x'	y'
0.99	2.08
0.01	-0.53
0.56	-0.78
0.77	1.05
0.68	-0.53
-0.87	-0.19
-0.86	-0.44
-1.32	-0.82
-1.29	-0.78
1.33	0.95

$$\rho(x,y) = \frac{\sum_{i=1}^{n} x_i' * y_i'}{n-1}$$

$$\rho(x,y) = \frac{5.94}{10-1}$$

$$\rho(x,y) = 0.66$$

eljalao@up.edu.ph

Descriptive Statistics

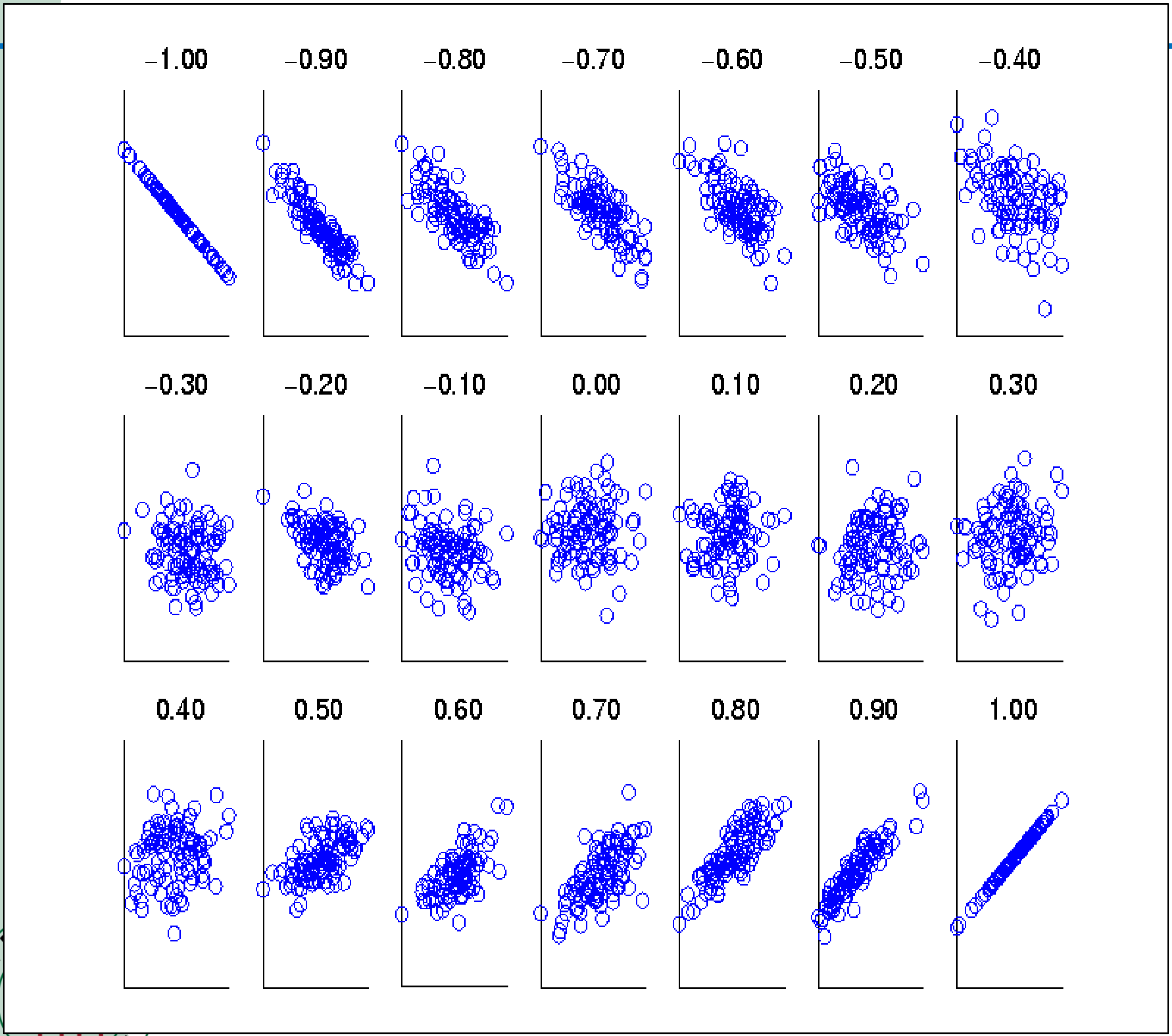


Figure 3.15: Scatter plots showing the similarity from -1 to 1.

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- **Data Preprocessing**
- Data Cleaning
- Data Transformations
- Pivot Tables
- Case Study

Data Preprocessing

Definition 3.25: Data Preprocessing

- A step in the business analytics framework wherein data is **transformed** from its **raw state** as an input for various business analytics algorithms
- Data in the real world is **dirty**
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - **noisy**: containing errors or outliers
    - e.g., Salary=“-10”
  - **inconsistent**: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”

Data Preprocessing

- Why is data dirty?
  - **Incomplete data** may come from
    - "Not applicable" data value when collected
    - Different considerations between the time when the data was collected and when it is analyzed.
    - Human/hardware/software problems
  - **Noisy data** (incorrect values) may come from
    - Faulty data collection instruments
    - Human or computer error at data entry
    - Errors in data transmission
  - **Inconsistent data** may come from
    - Different data sources
    - Functional dependency violation (e.g., modify some linked data)
  - **Duplicate records** also need data cleaning

Data Preprocessing

- No quality data, no quality mining results!
  - **Quality decisions** must be based on **quality data**
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the **majority of the work** of building a data warehouse

Data Preprocessing

- Major Tasks in Data Preprocessing:
  - **Data cleaning**
    - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
  - **Data integration**
    - Integration of multiple databases, data cubes, or files
  - **Data transformation**
    - Normalization and aggregation
  - **Data reduction**
    - Obtains reduced representation in volume but produces the same or similar analytical results
  - **Data discretization**
    - Part of data reduction but with particular importance, especially for numerical data

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- **Data Cleaning**
- Data Transformations
- Pivot Tables
- Case Study

eljalao@up.edu.ph

Data Cleaning

Definition 3.26: Data Cleaning

- Data cleaning is a **data preprocessing** task that constitute the following:
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration
- Importance
  - "Data cleaning is one of the **three biggest problems** in data warehousing"—Ralph Kimball

Data Cleaning

- **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 **not entered** due to misunderstanding
  - data may **not be considered** important at the time of entry
  - not register **history or changes** of the data
- Missing data may need to be **inferred**.
- Note that zero values **are not equal to** missing values

Data Cleaning

- How to Handle Missing Data?
  - **Ignore the tuple:** usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.)
  - Fill in the missing value **manually**: tedious + infeasible?
  - Fill in it automatically with
    - **a global constant** : e.g., “unknown”,
    - the attribute **mean (numerical) or mode (categorical)**
    - the attribute mean for all samples belonging **to the same class**: smarter

Data Cleaning

Example 3.3: Missing Data Example

- Tax Income
- Avg = 93.6 K

- Yes Tax Income
- Avg = 87.5 K

- No Tax Income
- Avg = 96 K

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	?	No
4	Yes	Married	120K	No
5	?	Divorced	?	Yes
6	?	Married	60K	No
7	Yes	Divorced	?	No
8	No	Single	85K	Yes
9	?	Married	75K	No
10	No	Single	90K	Yes

Data Cleaning

Definition 3.27: Noise

- Noise: **random error or variance** in a measured variable
- Incorrect attribute values may due to
  - **faulty data** collection instruments
  - **data entry** problems
  - **data transmission** problems
  - **technology** limitation
  - **inconsistency** in naming convention
- Other data problems which requires data cleaning
  - **duplicate** records
  - **incomplete** data
  - **inconsistent** data

Data Cleaning

Definition 3.28: Binning

- The process of data cleaning to reduce **dataset noise**.
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Two Types
  - Equal Width Binning
  - Equal Depth Binning

Data Cleaning

Definition 3.29: Equal-Width Binning

- Type of binning wherein the dataset is divided into a range of into N intervals of **equal size** forming a uniform grid
  - if A and B are the lowest and highest values of the variable, the width of intervals will be:

$$W = (B - A)/N \tag{3.7}$$

  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

Data Cleaning

Example 3.4: Equal-Width Binning

- Raw data for price (in PhP): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into 3 Bins: $W = \frac{B-A}{N} = \frac{34-4}{3} = 10$
- 3 Bins: [4,14), [14,24),[24,34]
  - Partition into equal-width bins:
    - Bin 1: 4, 8, 9
    - Bin 2: 15, 21, 21
    - Bin 3: 24, 25, 26, 28, 29, 34
- Smoothed data for price (in PhP): 7, 7, 7, 19, 19, 19, 27.6, 27.6, 27.6, 27.6, 27.6, 27.6

Data Cleaning

Definition 3.30: Equal-Depth Binning

- Type of binning wherein the dataset where the range is divided into N intervals, each containing approximately **same number of samples**
  - Good data scaling
  - Managing categorical attributes can be tricky

Data Cleaning

Example 3.5: Equal-Depth Binning

- Sorted data for price (in PhP): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partition into equal-frequency (equi-depth) bins:
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Smoothing by bin means:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
- Sorted data for price (in PhP): 9, 9, 9, 9, 23, 23, 23, 23, 29, 29, 29, 29

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- **Data Transformations**
- Pivot Tables
- Case Study

Data Transformation

- Some Data Transformation Tasks
  - Normalization: scaled to fall within **a small, specified range**
    - min-max normalization
    - z-score standardization
  - Attribute/feature **construction**
    - New attributes constructed from the given ones
      - May 2015 -> 201505

eljalao@up.edu.ph

Data Transformation: Normalization

Definition 3.31: Min-Max Normalization

- Data transformation task in which the transformed data will **fall between** a user specified $newmin$ and $newmax$ value.
- New data x' is computed from the old value x as follows:

$$x' = \frac{x - min}{max - min}(newmax - newmin) + newmin \qquad (3.9)$$

Data Transformation

Example 3.6: Min-Max Normalization

- Ex. Sorted data for Ages (in Years): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. We normalize it to a new range from 0 to 1.
- For example, 8 is transformed to:

$$x' = \frac{x - min}{max - min}(newmax - newmin) + newmin$$

$$x' = \frac{8 - 4}{34 - 4}(1 - 0) + 0 = 0.13$$

x	Transformed x
4	0.00
8	0.13
9	0.17
15	0.37
21	0.57
21	0.57
24	0.67
25	0.70
26	0.73
28	0.80
29	0.83
34	1.00

Data Transformation

Definition 3.32: z-Score Standardization

- Data transformation task in which the transformed data will follow a **standard normal distribution** of mean 0 and standard deviation of 1.

$$x' = \frac{(x_i - mean(x))}{std(x)} \tag{3.10}$$

Data Transformation

Example 3.6: Min-Max Normalization

- Ex. Sorted data for Age (in Year): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.
- For example, 8 is transformed to:

$$mean(x) = 20.33$$
$$std(x) = 9.36$$

$$x_i' = \frac{(x_i - mean(x))}{std(x)}$$

$$x_i' = \frac{8 - 20.33}{9.36} = -1.32$$

x	Transformed x
4	-1.74
8	-1.32
9	-1.21
15	-0.57
21	0.07
21	0.07
24	0.39
25	0.50
26	0.61
28	0.82
29	0.93
34	1.46

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- Data Transformations
- **Pivot Tables**
- Case Study

Analyzing Data with PivotTables

Definition 3.33: Pivot Tables

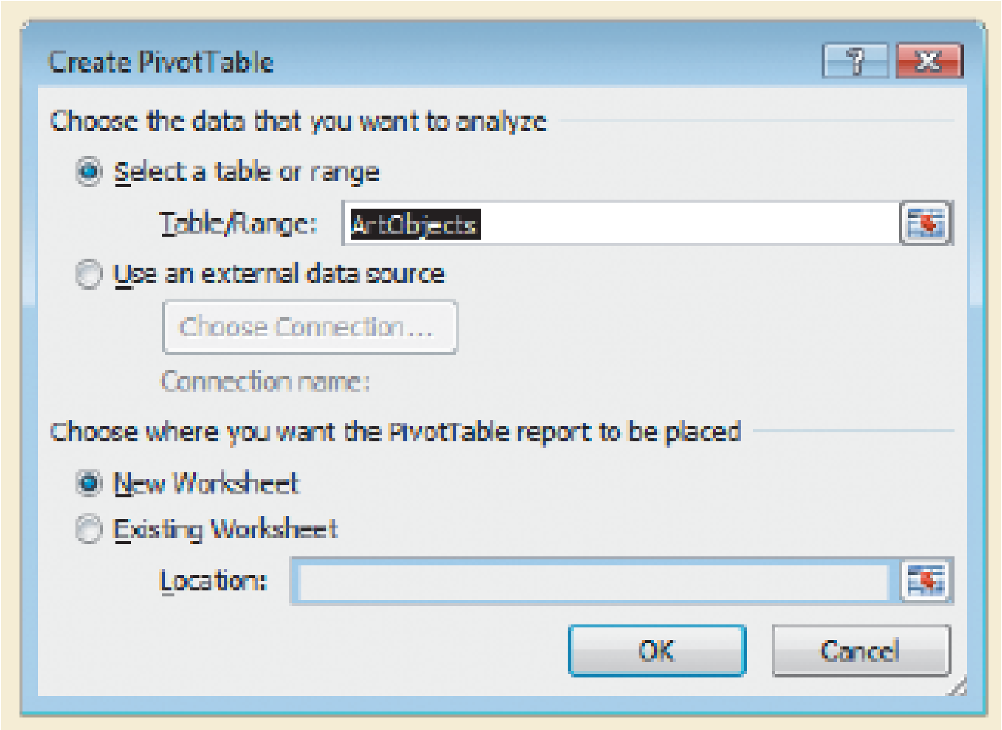
- A **PivotTable** is an interactive table that enables you to group and summarize either a range of data or an Excel table into a concise, tabular format for easier reporting and analysis

	A	B	C	D	E	F
1	Location	(All)				
2						
3	Sum of Appraised Value	Column Labels				
4	Row Labels	Excellent	Good	Fair	Poor	Grand Total
5	Installation	$185,000	$2,500			$187,500
6	Painting	$611,520	$41,669	$10,500	$18,450	$682,139
7	Sculpture	$194,292	$16,300	$3,942	$3,950	$218,484
8	Textile	$7,400	$18,100	$27,500		$53,000
9	Grand Total	$998,212	$78,569	$41,942	$22,400	$1,141,123

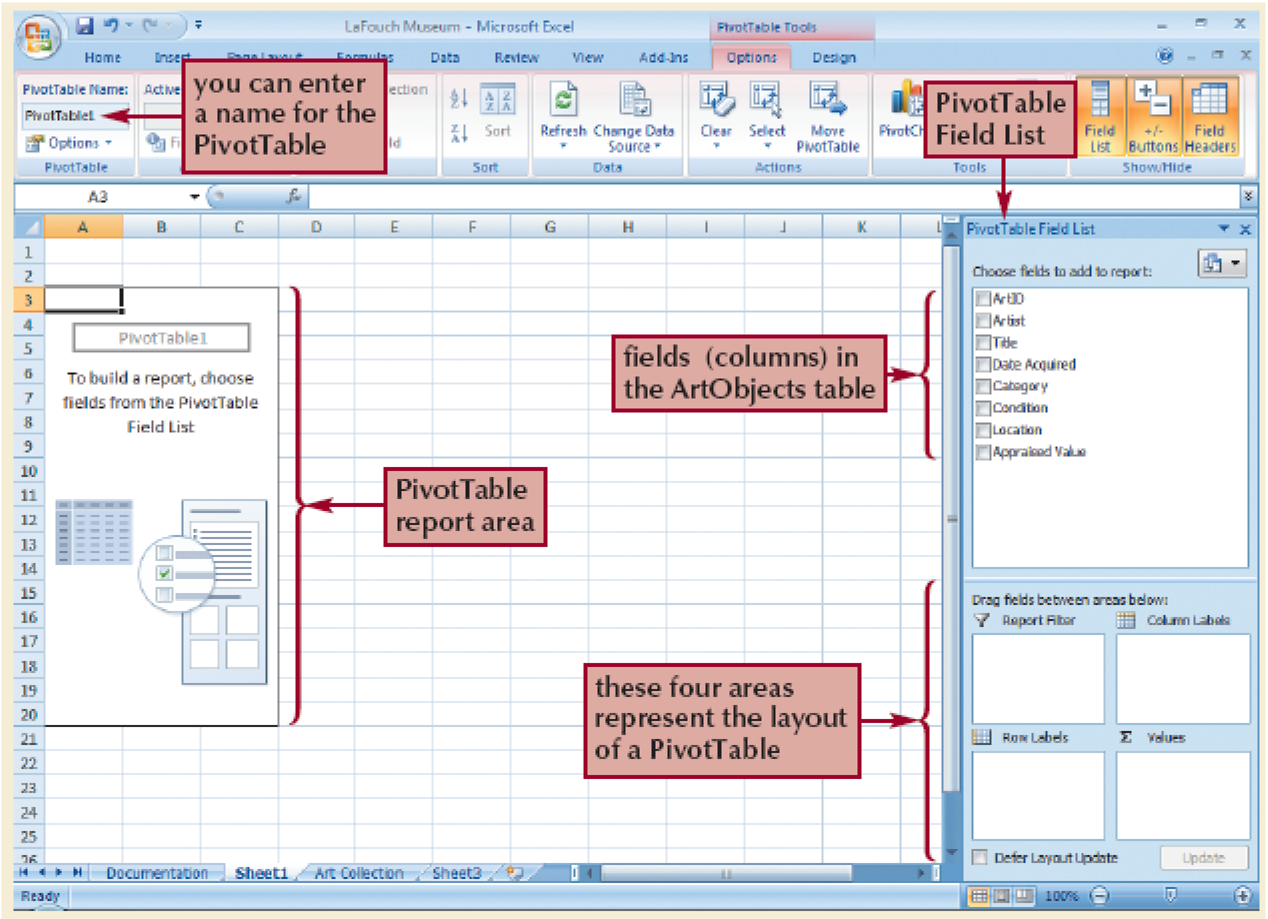
Creating a PivotTable

- Click in the Excel table or select the range of data for the PivotTable
- In the Tables group on the Insert tab, click the PivotTable button
- Click the Select a table or range option button and verify the reference in the Table/Range box
- Click the New Worksheet option button or click the Existing worksheet option button and specify a cell
- Click the OK button
- Click the check boxes for the fields you want to add to the PivotTable (or drag fields to the appropriate box in the layout section)
- If needed, drag fields to different boxes in the layout section

Creating a PivotTable

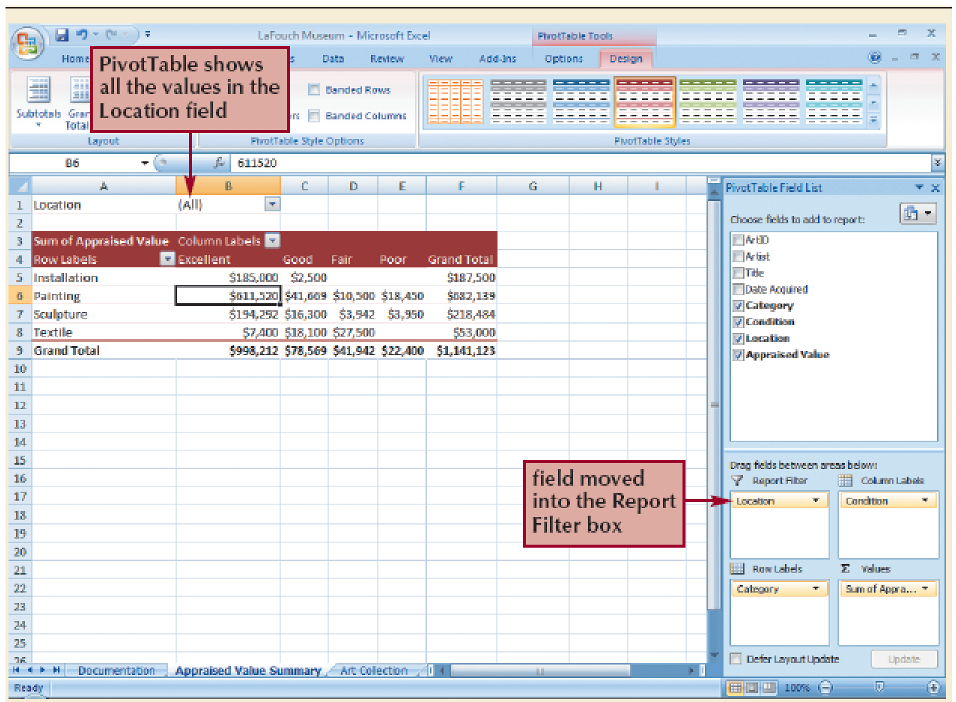


Creating a PivotTable



Adding a Report Filter to a PivotTable

- A report filter allows you to filter the PivotTable to display summarized data for one or more field items or all field items in the Report Filter area



Filtering PivotTable Fields

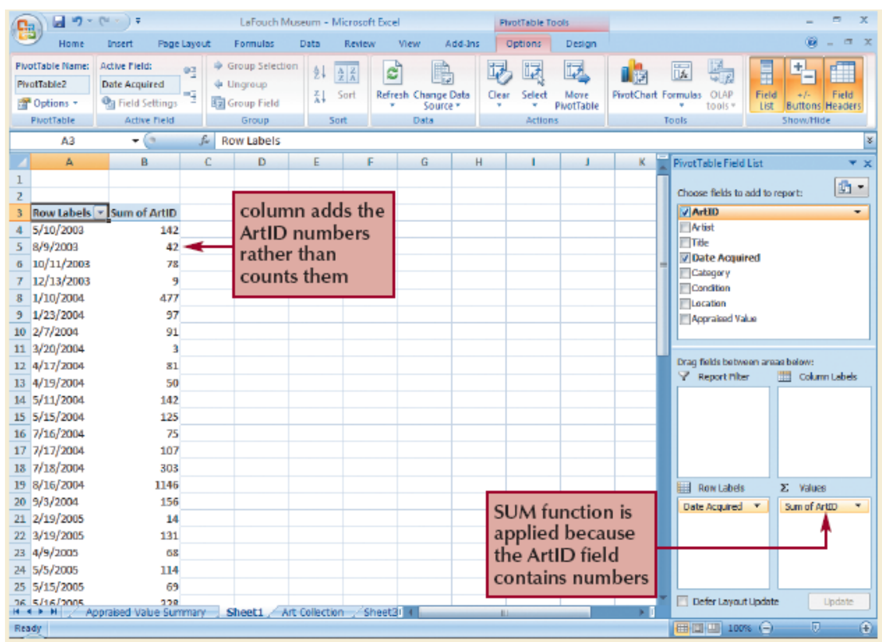
- Filtering a field lets you **focus** on a subset of items in that field
- You can filter field items in the PivotTable by clicking the field arrow button in the PivotTable that represents the data you want to hide and then uncheck the check box for each item you want to hide

Refreshing a PivotTable

- You cannot change the data directly in the PivotTable. Instead, you must edit the Excel table, and then **refresh**, or update, the PivotTable to reflect the current state of the art objects list
- Click the **PivotTable Tools Options** tab on the Ribbon, and then, in the **Data** group, click the **Refresh** button

Grouping PivotTable Items

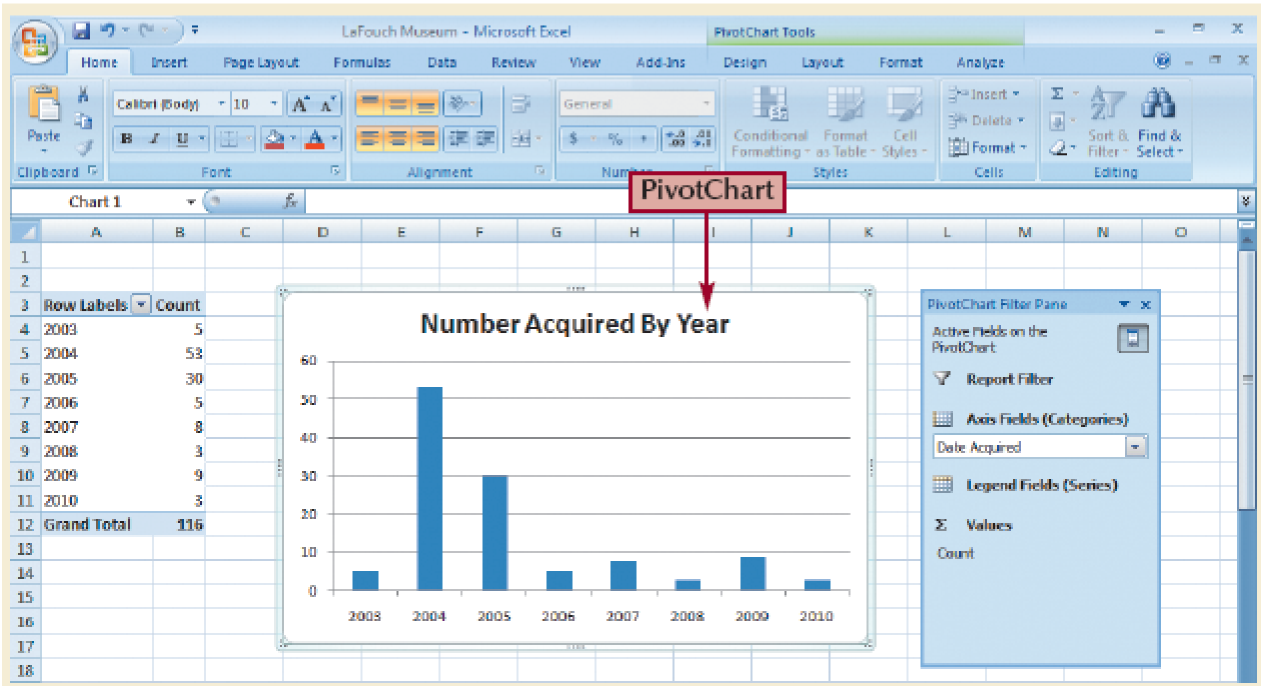
- When a field contains numbers, dates, or times, you can combine items in the rows of a PivotTable and combine them into groups automatically



Creating a PivotChart

- A **PivotChart** is a graphical representation of the data in a PivotTable
- A PivotChart allows you to interactively add, remove, filter, and refresh data fields in the PivotChart similar to working with a PivotTable
- Click any cell in the PivotTable, then, in the Tools group on the PivotTable Tools Options tab, click the **PivotChart button**

Creating a PivotChart



Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- Data Transformations
- Pivot Tables
- **Case Study**

Case Study 3

- Analyzing Art
  - Generate Descriptive Statistics
  - Generate Graphs

Outline for this Session

- What is Data?
- Types of Datasets
- Descriptive Statistics
- Data Preprocessing
- Data Cleaning
- Data Transformations
- Pivot Tables
- Case Study

References

- James Lee Notes From: http://www.sjsu.edu/people/james.lee/courses/102/s1/asDescriptive_Statistics2.ppt
- Section on Pivot Tables from: New Perspectives: Working with Excel Tables, PivotTables, and PivotCharts Tutorial
- Tan et al. Intro to Data Mining Notes
- www.cs.gsu.edu/~cscyqz/courses/dm/slides/ch02.ppt