Welcome to:

Unit 2 – Data understanding and preparation

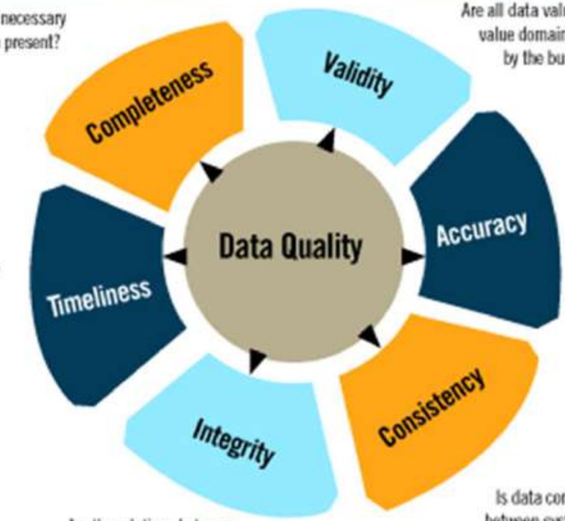
Unit objectives

After completing this unit, you should be able to:

- Know the steps involved in data understanding and data preparation.
- Read data from various sources.
- Visualize the data in different forms.
- Understand the issues related to data quality.
- Know the different outlier detection methods.
- Combine data files.
- Understand how to partition the data.
- Aggregate the data.

Data Preparation

Figure 1: Data Quality Dimensions



Is all necessary data present?

Are all data values within the value domains specified by the business?

Does data reflect the real-world objects or a verifiable source?

Is data consistent between systems? Do duplicate records exist?

Are the relations between entities and attributes consistent? Within tables and between?

Is the data available at the time needed?

Data Understanding: Quality

- Data quality is an essential characteristic that determines the reliability of data for making decisions. High-quality data is:

- **Complete:** All relevant data —such as accounts, addresses and relationships for a given customer—is linked.

- **Accurate:** Common data problems like misspellings, typos, and random abbreviations have been cleaned up.

- **Available:** Required data is accessible on demand; users do not need to search manually for the information.

- **Timely:** Up-to-date information is readily available to support decisions.

Data Understanding: Quality Dimensions

Completeness

- There are no missing values were completeness is required
- The number of records present is the appropriate amount of data
- All necessary fields are present
- Primary keys are present, unique and in good format
- All foreign key fields are present and in good format

Duplicates

- Duplicate records are not present
- Redundant fields are not present
- Duplicate records across distinct datasets are not present

Data Understanding: Quality Dimensions

Rules

- All rules have been identified and are accurate
- Data has been tested and follows data rules
- All field data is formatted correctly for the representative data type

Usability

- Metadata is available
- Data is easy to interpret
- Data is representative of intended objectives

Data Understanding

- Steps involved in data understanding are:
  - A. Collect the data
  - B. Describe the data
  - C. Explore the data
  - D. Verify the data quality

A. Data Collection Method: Sampling

- Sampling: obtaining a small sample s to represent the whole data set N

- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data

- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:

- Note: Sampling may not reduce database I/Os (page at a time)

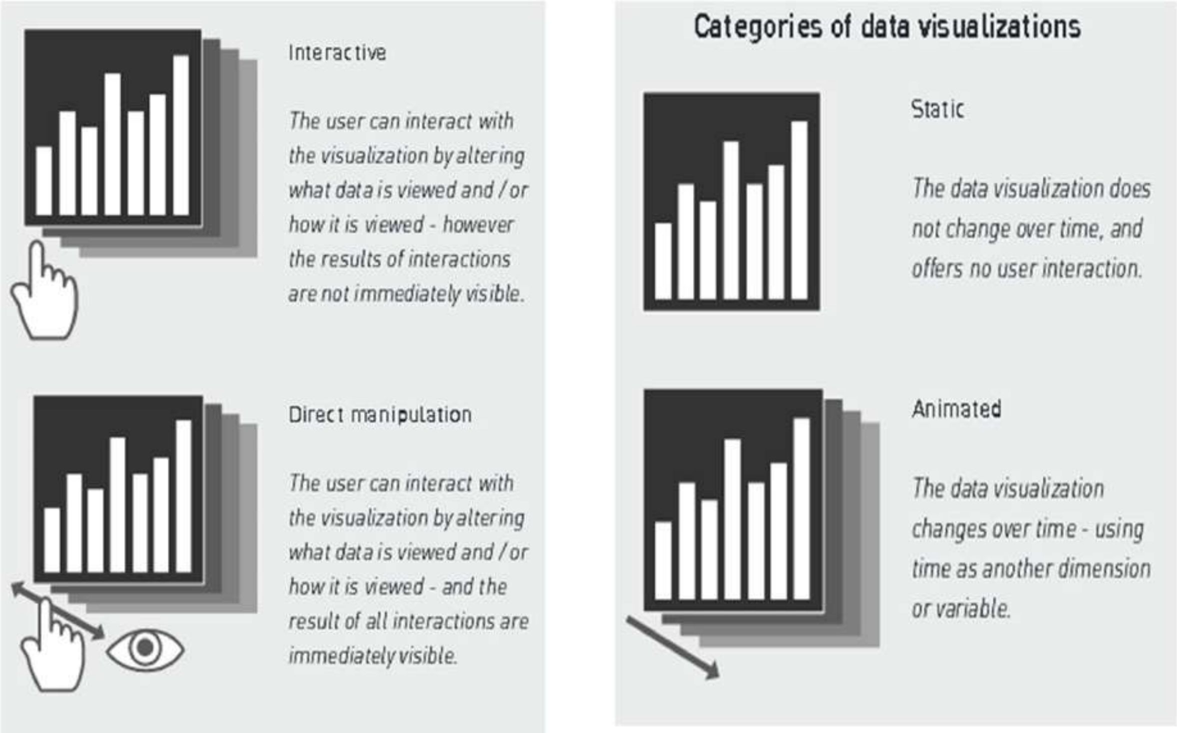
Types of Sampling

- Simple random sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - Once an object is selected, it is removed from the population
- Sampling with replacement
  - A selected object is not removed from the population
- Stratified sampling:
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

B. Data Description

- Total Number or "N", Mean, Median, Mode and Standard Deviation are used to describe your data.
- The Total Number or "N" is the number of observations made.
- Mean: This is the average of the data. Adding the values of all of the observations and dividing the total by the total number of observations or "N".
- Median: This is the middle value of the observations.
- Mode: This is the most frequent observation.
- The Standard Deviation is a description of how tightly the observed data points are clustered around the mean. One standard deviation should include approximately 68% of the data points. Two Standard Deviations should include approximately 95% of the data points and three Standard Deviations should include approximately 99% of the data points.

C. Data Exploration: Categories of data visualization



Categories of data visualizations

Interactive

The user can interact with the visualization by altering what data is viewed and / or how it is viewed - however the results of interactions are not immediately visible.

Direct manipulation

The user can interact with the visualization by altering what data is viewed and / or how it is viewed - and the result of all interactions are immediately visible.

Static

The data visualization does not change over time, and offers no user interaction.

Animated

The data visualization changes over time - using time as another dimension or variable.

D. Verification of data quality: Outlier Detection

- Outlier detection also known as anomaly detection.
- Process of finding data objects with behaviors that are very different from expectation.
- Outlier detection methods
- Supervised
- Semi-Supervised
- Unsupervised

Data Understanding: Relevance

- What data is available for the task?
- Is this data relevant?
- Is additional relevant data available?
- How much historical data is available?
- Who is the data expert ?

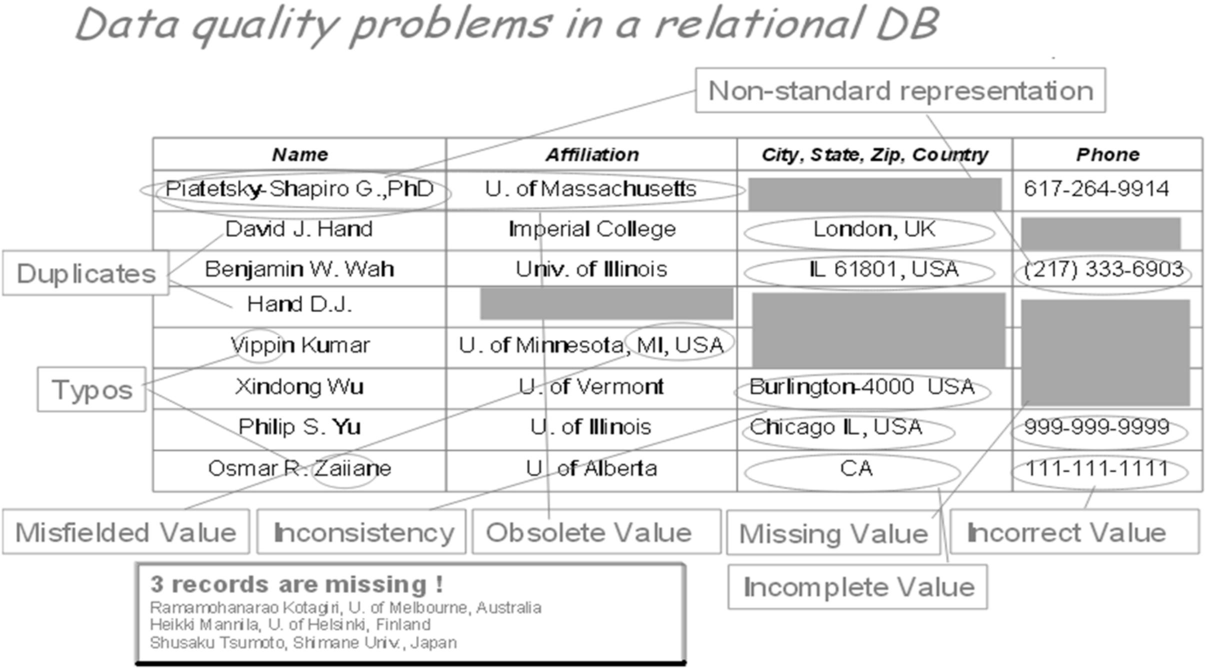
Data Understanding: Thumb Rule

- Number of instances (records)
  - Rule of thumb: 5,000 or more desired
  - if less, results are less reliable; use special methods (boosting, …)

- Number of attributes (fields)
  - Rule of thumb: for each field, 10 or more instances
  - If more fields, use feature reduction and selection

- Number of targets
  - Rule of thumb: >100 for each class
  - if very unbalanced, use stratified sampling

Data Understanding



Data quality problems in a relational DB

Name	Affiliation	City, State, Zip, Country	Phone
Piatetsky-Shapiro G.,PhD	U. of Massachusetts		617-264-9914
David J. Hand	Imperial College	London, UK	
Benjamin W. Wah	Univ. of Illinois	IL 61801, USA	(217) 333-6903
Hand D.J.			
Vippin Kumar	U. of Minnesota, MI, USA		
Xindong Wu	U. of Vermont	Burlington-4000 USA	
Philip S. Yu	U. of Illinois	Chicago IL, USA	999-999-9999
Osmar R. Zaiiane	U. of Alberta	CA	111-111-1111

Non-standard representation

Duplicates

Typos

Misfielded Value

Inconsistency

Obsolete Value

Missing Value

Incorrect Value

Incomplete Value

3 records are missing !

Ramamohanarao Kotagiri, U. of Melbourne, Australia
Heikki Mannila, U. of Helsinki, Finland
Shusaku Tsumoto, Shimane Univ., Japan

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, 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”
    - discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone’s birthday?

Data Cleaning

- Identify invalid data: Use your standards of data quality and your key necessities to identify all the invalid or inaccurate data.

- Investigate the reasons for the bad data. Having this understanding will assist you in taking the necessary actions to correct the data.

- Determine how the dirty data should be cleaned. Whenever possible, invalid data should be corrected so it can be used for your project.

- Perform accuracy tests to ensure the data were properly cleaned. Accuracy tests are a physical comparison of the data collected with the actual event/object.

Steps involved in Data Cleaning

- 1.1. Data acquisition and metadata
- 1.2. Missing values
- 1.3.Unified date format
- 1.4.Converting nominal to numeric
- 1.5.Discretization of numeric data
- 1.6.Data validation and statistics

1.1. Data Cleaning: Acquisition (Reading Data)

- You can read the data from various sources like:
  - Query-based data extracts from the database to flat files
  - High-level query languages for direct access to the database
  - Low-level connections for direct access to the database
  - Programming languages to extract data from files(text file or excel-sheet or XML)
  - Data can be in DBMS
  - ODBC, JDBC protocols

- Data in a flat file
  - Fixed-column format
  - Delimited format: tab, comma “,” , other
  - E.g. C4.5 and Weka “arff” use comma-delimited data
  - Attention: Convert field delimiters inside strings

- Verify the number of fields before and after

1.1. Data Cleaning: Example

- Original data (fixed column format)
  -
    ```
    000000000130.06.19971979-10-3080145722      #000310 111000301.01.0001000000000004
    0000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.00000
    0000000000. 000000000000000.000000000000000.0000000......
    000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000
    000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000
    000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000
    000000.000000000000000.000000000000000.000000000000000.00 000000000300.00 0
    000000000300.00
    ```

- Clean data
  -
    ```
    0000000001,199706,1979.833,8014,5722   ,   ,#000310  ....
    ,111,03,000101,0,04,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,0300.00
    ```

1.1. Data Cleaning: Metadata

- Field types:
  - binary, nominal (categorical), ordinal, numeric, …
  - For nominal fields: tables translating codes to full descriptions

- Field role:
  - input : inputs for modeling
  - target : output
  - id/auxiliary : keep, but not use for modeling
  - ignore : don’t use for modeling
  - weight : instance weight
  - …

- Field descriptions

1.2. Data Cleaning: Reformatting

- Convert data to a standard format (e.g. arff or csv)
- Missing values
- Unified date format
- Binning of numeric data
- Fix errors and outliers
- Convert nominal fields whose values have order to numeric.
  - Q: Why?
  - Convert nominal fields whose values have order to numeric to be able to use ">" and "<" comparisons on these fields.

1.2.Missing Data

- Data is not always available
  - E.g., many shops 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
  - certain 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

1.2.Missing Data

- Ignore the Missing data: usually done when class label is missing (when doing classification)—not effective when the % 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", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

1.2.Data Cleaning: Missing Values

- Missing data can appear in several forms:
  - <empty field> “0” “.” “999” “NA” …
- Standardize missing value code(s)
- How can we dealing with missing values?
- Dealing with missing values:
  - ignore records with missing values
  - treat missing value as a separate value
  - Imputation: fill in with mean or median values

1.3.Data Cleaning: Unified Date Format

- We want to transform all dates to the same format internally
- Some systems accept dates in many formats
  - e.g. “Sep 24, 2003” , 9/24/03, 24.09.03, etc
  - dates are transformed internally to a standard value
- Frequently, just the year (YYYY) is sufficient
- For more details, we may need the month, the day, the hour, etc
- Representing date as YYYYMM or YYYYMMDD can be OK, but has problems
- Q: What are the problems with YYYYMMDD dates?
- Problems with YYYYMMDD dates:
  - YYYYMMDD does not preserve intervals:
  - 20040201 - 20040131 /= 20040131 – 20040130
  - This can introduce bias into models

1.4. Categorical Variables: Ordinal Variables

- Ordinal variable—a categorical variable with some intrinsic order or numeric value

- Examples of ordinal variables:
  - Education (no high school degree, HS degree, some college, college degree)
  - Agreement (strongly disagree, disagree, neutral, agree, strongly agree)
  - Rating (excellent, good, fair, poor)
  - Frequency (always, often, sometimes, never)
  - Any other scale (“On a scale of 1 to 5...”)

1.4.Categorical Variables: Nominal Variables

- Nominal variable – a categorical variable without an intrinsic order

- Examples of nominal variables:
  - Where a person lives in the U.S. (Northeast, South, Midwest, etc.)
  - Sex (male, female)
  - Nationality (American, Mexican, French)
  - Race/ethnicity (African American, Hispanic, White, Asian American)
  - Favorite pet (dog, cat, fish, snake)

1.4.Categorical Variables: Dichotomous Variables

- Dichotomous (or binary) variables – a categorical variable with only 2 levels of categories
  - Often represents the answer to a yes or no question

- For example:
  - "Did you attend the church picnic on May 24?"
  - "Did you eat potato salad at the picnic?"
  - Anything with only 2 categories

1.4.Coding

- Coding – process of translating information gathered from questionnaires or other sources into something that can be analyzed

- Involves assigning a value to the information given—often value is given a label

- Coding can make data more consistent:
  - Example: Question = Sex
  - Answers = Male, Female, M, or F
  - Coding will avoid such inconsistencies

1.4.Coding Systems

- Common coding systems (code and label) for dichotomous variables:
  - 0=No 1=Yes
  - (1 = value assigned, Yes= label of value)
  - OR: 1=No 2=Yes

- When you assign a value you must also make it clear what that value means
  - In first example above, 1=Yes but in second example 1=No
  - As long as it is clear how the data are coded, either is fine

- You can make it clear by creating a data dictionary to accompany the dataset

1.4.Coding: Dummy Variables

- A “dummy” variable is any variable that is coded to have 2 levels (yes/no, male/female, etc.)

- Dummy variables may be used to represent more complicated variables
  - Example: # of cigarettes smoked per week--answers total 75 different responses ranging from 0 cigarettes to 3 packs per week
  - Can be recoded as a dummy variable:
  - 1=smokes (at all) 0=non-smoker

- This type of coding is useful in later stages of analysis

1.4.Coding:Attaching Labels to Values

- Many analysis software packages allow you to attach a label to the variable values
  - Example: Label 0’s as male and 1’s as female

- Makes reading data output easier:
  - Without label:

Variable SEX	Frequency	Percent
0	21	60%
1	14	40%

  - With label:

Variable SEX	Frequency	Percent
Male	21	60%
Female	14	40%

1.4.Coding- Ordinal Variables

- Coding process is similar with other categorical variables

- Example: variable EDUCATION, possible coding:
  - 0 = Did not graduate from high school
  - 1 = High school graduate
  - 2 = Some college or post-high school education
  - 3 = College graduate

- Could be coded in reverse order (0=college graduate, 3=did not graduate high school)

- For this ordinal categorical variable we want to be consistent with numbering because the value of the code assigned has significance

1.4.Coding – Ordinal Variables (cont.)

- Example of bad coding:
  - 0 = Some college or post-high school education
  - 1 = High school graduate
  - 2 = College graduate
  - 3 = Did not graduate from high school

- Data has an inherent order but coding does not follow that order—NOT appropriate coding for an ordinal categorical variable

1.4.Coding: Nominal Variables

- For coding nominal variables, order makes no difference

- Example: variable RESIDE
  - 1 = Northeast
  - 2 = South
  - 3 = Northwest
  - 4 = Midwest
  - 5 = Southwest

- Order does not matter, no ordered value associated with each response

1.4.Coding: Continuous Variables

- Creating categories from a continuous variable (ex. age) is common

- May break down a continuous variable into chosen categories by creating an ordinal categorical variable

- Example: variable = AGECAT
  - 1 = 0–9 years old
  - 2 = 10–19 years old
  - 3 = 20–39 years old
  - 4 = 40–59 years old
  - 5 = 60 years or older

1.4.Coding:Continuous Variables (cont.)

- May need to code responses from fill-in-the-blank and open-ended questions
  - Example: “Why did you choose not to see a doctor about this illness?”

- One approach is to group together responses with similar themes
  - Example: “didn’t feel sick enough to see a doctor”, “symptoms stopped,” and “illness didn’t last very long”
  - Could all be grouped together as “illness was not severe”

- Also need to code for “don’t know” responses”
  - Typically, “don’t know” is coded as 9

1.4.Conversion: Nominal to Numeric

- Some tools can deal with nominal values internally
- Other methods (neural nets, regression, nearest neighbor) require only numeric inputs
- To use nominal fields in such methods need to convert them to a numeric value
  - Q: Why not ignore nominal fields altogether?
  - A: They may contain valuable information
- Different strategies for binary, ordered, multi-valued nominal fields

1.4.Conversion

- How would you convert binary fields to numeric?
  - E.g. Gender=M, F
- How would you convert ordered attributes to numeric?
  - E.g. Grades
- Binary fields
  - E.g. Gender=M, F
- Convert to Field_0_1 with 0, 1 values
  - e.g. Gender = M → Gender_0_1 = 0
  - Gender = F → Gender_0_1 = 1
- Ordered attributes (e.g. Grade) can be converted to numbers preserving natural order, e.g.
  - A → 4.0
  - A- → 3.7
  - B+ → 3.3
  - B → 3.0
- Q: Why is it important to preserve natural order?

1.4.Conversion: Nominal, Few Values

- Multi-valued, unordered attributes with small (*rule of thumb* < *20*) no. of values
  - e.g. Color=Red, Orange, Yellow, …, Violet
  - for each value *v* create a binary “flag” variable C_*v* , which is 1 if Color=*v*, 0 otherwise

ID	Color	…
371	red	
433	yellow	

→

ID	C_red	C_orange	C_yellow	…
371	1	0	0	
433	0	0	1	

1.4.Coding Tip

- Though you do not code until the data is gathered, you should think about how you are going to code while designing your questionnaire, before you gather any data. This will help you to collect the data in a format you can use.

1.5. Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers

- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by Discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

1.5.Data Discretization Methods

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)
  - Decision-tree analysis (supervised, top-down split)
  - Correlation (e.g., $\chi 2$) analysis (unsupervised, bottom-up merge)

1.5.Simple Discretization: Binning

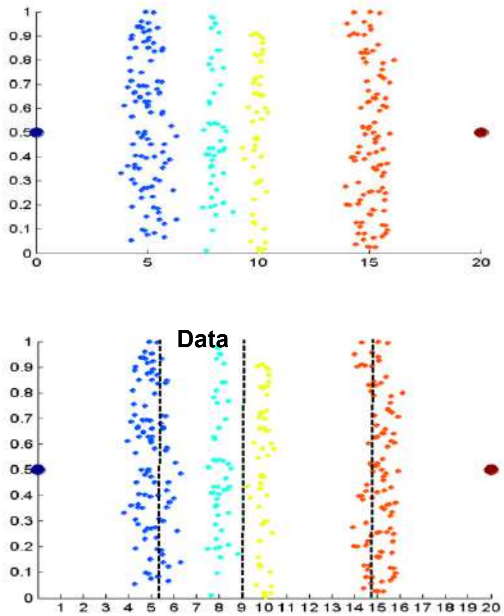
- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B – A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

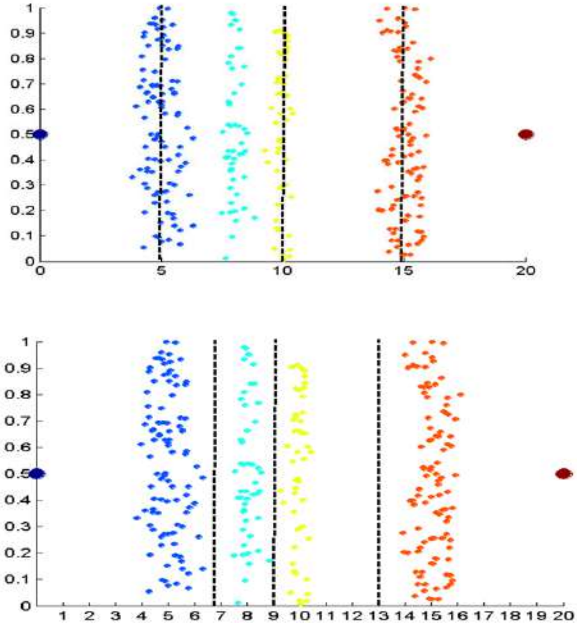
1.5.Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 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
- * Smoothing by bin boundaries:
- - Bin 1: 4, 4, 4, 15
- - Bin 2: 21, 21, 25, 25
- - Bin 3: 26, 26, 26, 34

1.5.Discretization Without Using Class Labels



Equal frequency (binning)



K-means clustering leads to better results

1.5.Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using entropy to determine split point (Discretization point)
  - Top-down, recursive split
  - Details to be covered in Chapter “Classification”

- Correlation analysis (e.g., Chi-merge: χ2-based Discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

1.6. Statics: Univariate Data Analysis

- Univariate data analysis-explores each variable in a data set separately
  - Serves as a good method to check the quality of the data
  - Inconsistencies or unexpected results should be investigated using the original data as the reference point

- Frequencies can tell you if many study participants share a characteristic of interest (age, gender, etc.)
  - Graphs and tables can be helpful

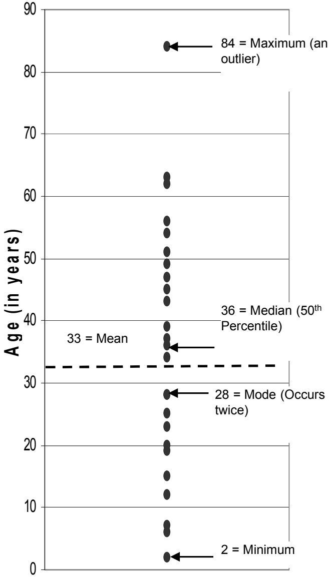
1.6. Univariate Data Analysis (cont.)

- Examining continuous variables can give you important information:
  - Do all subjects have data, or are values missing?
  - Are most values clumped together, or is there a lot of variation?
  - Are there outliers?
  - Do the minimum and maximum values make sense, or could there be mistakes in the coding?

1.6. Univariate Data Analysis (cont.)

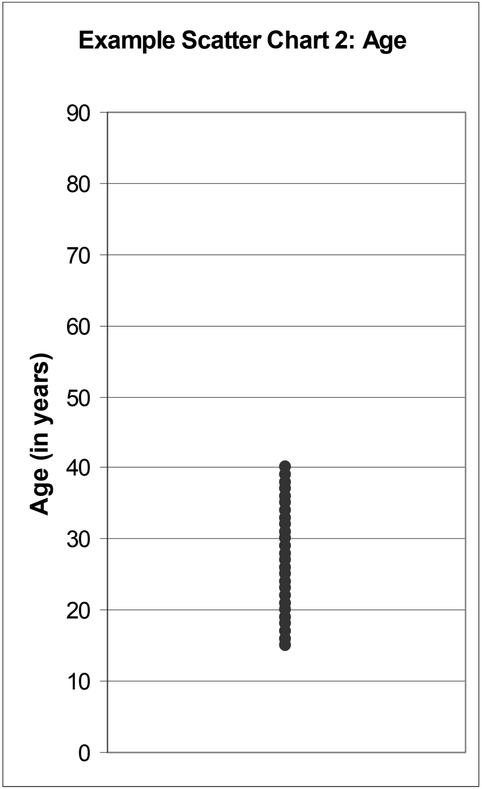
- Commonly used statistics with univariate analysis of continuous variables:
  - Mean – average of all values of this variable in the dataset
  - Median – the middle of the distribution, the number where half of the values are above and half are below
  - Mode – the value that occurs the most times
  - Range of values – from minimum value to maximum value

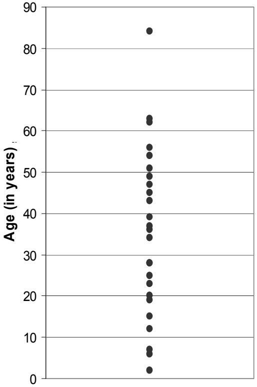
1.6. continuous variable distribution



1.6. Standard Deviation

- Figure left: narrowly distributed age values (SD = 7.6)
- Figure right: widely distributed age values (SD = 20.4)

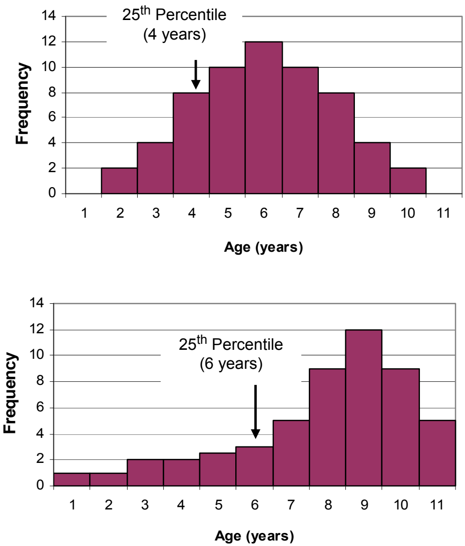




1.6. Distribution and Percentiles

- ***Distribution*** – whether most values occur low in the range, high in the range, or grouped in the middle
- ***Percentiles*** **–** the percent of the distribution that is equal to or below a certain value

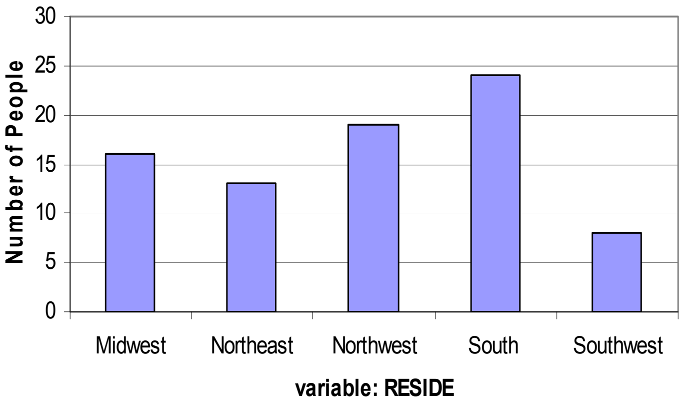
Distribution curves for variable AGE



1.6. Analysis of Categorical Data

- Distribution of categorical variables should be examined before more in-depth analyses
  - Example: variable RESIDE

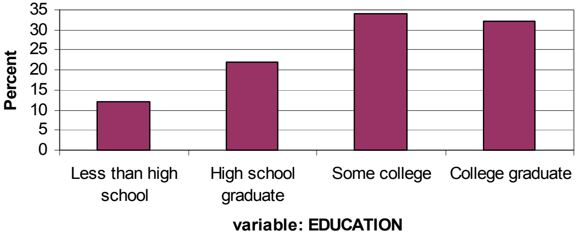
Number of people answering example questionnaire who reside in 5 regions of the United States



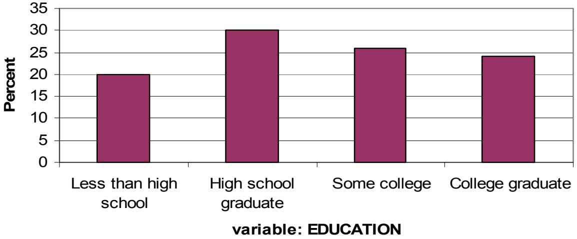
1.6. Analysis of Categorical Data (cont.)

- Another way to look at the data is to list the data categories in tables
- Table shown gives same information as in previous figure but in a different format

Observed data on level of education from a hypothetical questionnaire



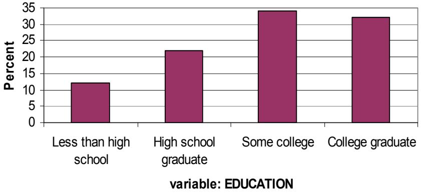
Data on the education level of the US population aged 20 years or older, from the US Census Bureau



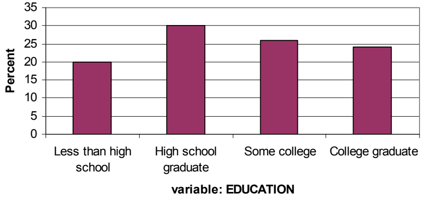
1.6. Observed vs. Expected Distribution

- Education variable
  - Observed distribution of education levels (top)
  - Expected distribution of education (bottom) (1)
  - Comparing graphs shows a more educated study population than expected
- Are the observed data really that different from the expected data?
- Answer would require further exploration with statistical tests

Observed data on level of education from a hypothetical questionnaire



Data on the education level of the US population aged 20 years or older, from the US Census Bureau



Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values

- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to
  - $$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max(|v'|) < 1

Checkpoint (1 of 2)

1)Which of the following is not an aggregate function

a) Sum
b) Average
c) Minimum
d) Ceil

2)Which dataset is used to train or build a model?

a) Training set
b) Validation set
c) Test set

3)Which sampling is also called systematic sampling

a) First N Sampling
b) Cluster Sampling
c) N-th Record Sampling
d) Simple random sampling

Checkpoint (2 of 2)

4. ___________ Segmentation is the process of grouping a customer market using variables such as age, marital status, sex, family life cycle, and family type/size.

   a) Geographic
   b) Demographic
   c) Psycho-graphic
   d) Behavioral

5. _________ Data visualization changes over time.

   a) Static
   b) Animated
   c) Interactive
   d) Direct manipulation

6. Outlier detection also known as anomaly detection.

   a) True
   b) False

Unit summary

Having completed this unit, you should be able to:

- Know the steps involved in data understanding and data preparation.
- Read data from various sources.
- Visualize the data in different forms.
- Understand the issues related to data quality.
- Know the different outlier detection methods.
- Combine data files.
- Understand how to partition the data.
- Aggregate the data.