BACS3013 Data Science

Chapter 2: Data collection, sampling and preprocessing

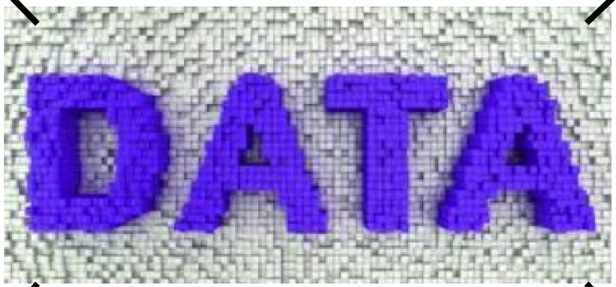
Content

- Types of data sources and data elements
- Data collection
- Populations and Samples of Big Data
- Data munging/wrangling
- Data pre-processing
- Visual data and exploratory statistical analysis
- Data storage and management of Big Data

Types of Data Sources and Data Elements

Garbage in, garbage out (GIGO)

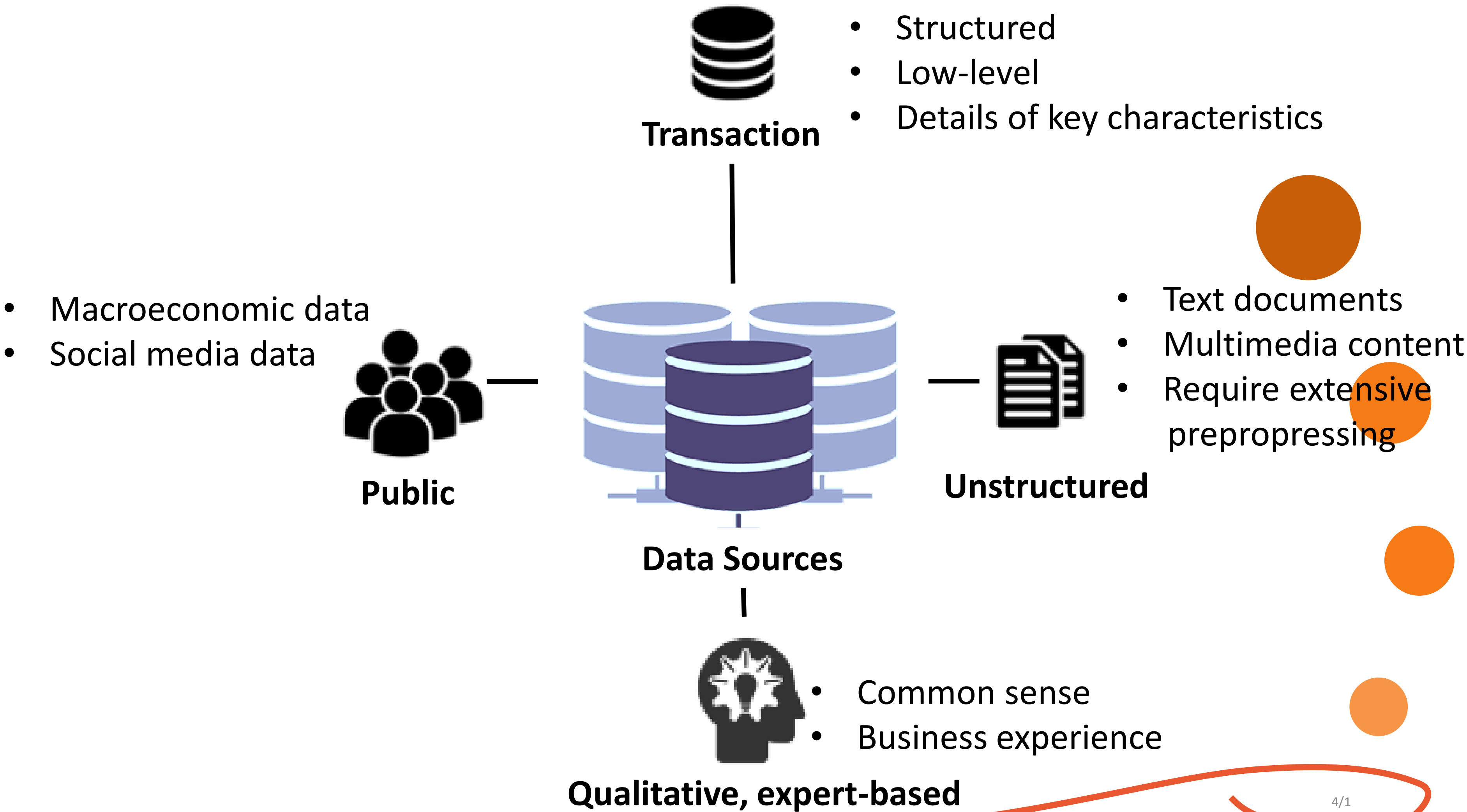
Key Ingredient



Messy

The more the better

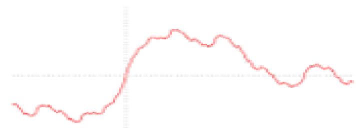
Types of Data Sources

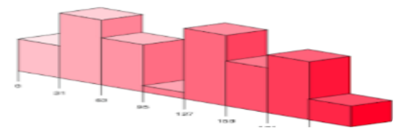


Discussion

- For each type of data source, provide an example of analytic that can be done. Example
  - –**publicly available data**: The most popular food in Malaysia based on Facebook data.
  - –Transaction data?
  - –Unstructured data?
  - –Qualitative/expert-based data?
  - –Publicly available data?

Types of Data Elements





interval

Limit/without

e.g. sales

Continuous

Categorical

Nominal

Ordering is not meaningful

e.g. job title

Ordinal

Ordering is meaningful

e.g. age-group

Binary

2 values only

e.g. gender

Question

- Provide other examples for the following data elements:

1. Continuous data
2. Nominal data
3. Ordinal data
4. Binary data

Data collection

The activity of collecting information that can be used to find out about a particular subject.

(Cambridge Dictionary)

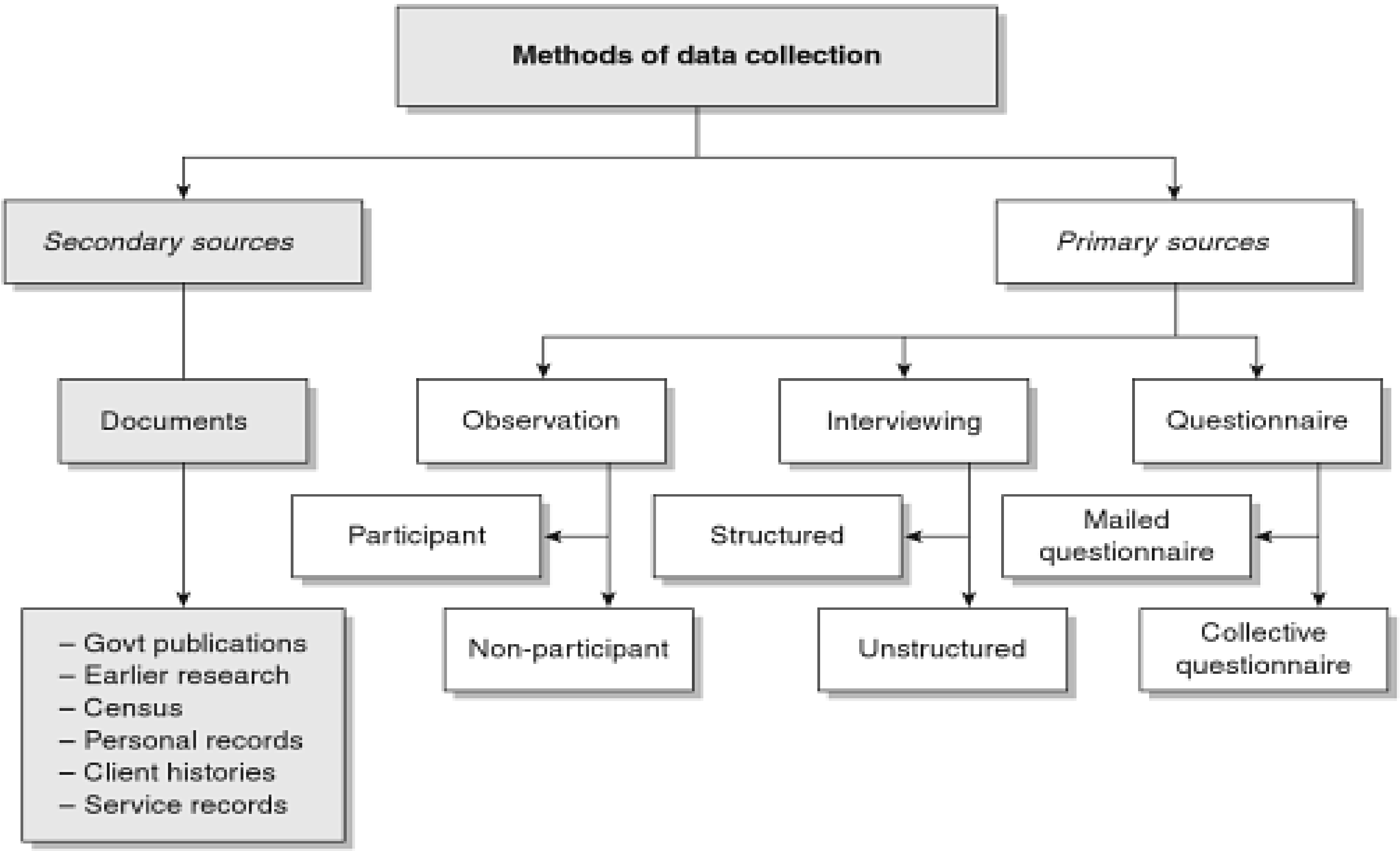
Data collection is the process of **gathering** and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.

(Wikipedia)

Methods of Collecting Data

- Direct Observation
- Interview
- Experiment
- Surveys
- And others...

Methods of Data Collection



Sampling of Big Data

Big data is a term that describes the **large volume of data** – both **structured and unstructured** – that inundates a business on a day-to-day basis. But **it's not the amount of data that's important**. It's **what organizations do with the data that matters**. Big data can be analyzed for **insights** that lead to better decisions and strategic business moves.

Sampling

- The aim of sampling is to take a subset of past customer data and use that to build an analytical model.

Question: Given high performance of computer ability nowadays, why do we need sampling while we could also directly analyze the full data set?

Why Sampling of Big Data

Storing the *full* data may not be feasible

- Your application may not keep *everything*

Work with data in full is inconvenient

- What is the need to analyze the *full* data?

Work with a compact summary is faster

- Would you rather exploring data with a PC than a supercomputer/cluster?

Why Sample?

intuitive semantics	• We obtain a smaller data set with the same structure
straightforward	• Run the analysis on the sample that you would on the full data • Some rescaling/reweighting may be necessary
general and agnostic	• Other summary methods only work for certain computations • Though sampling can be tuned to optimize some criteria
easy to understand	• So prevalent that we have an intuition about sampling

Sampling

- Choosing a sample

The sample should be taken from an average business period to get a picture of the target population that is as accurate as possible.

Sampling Bias

- Sampling bias should be avoided as much as possible. However, this is not straight-forward ☹
- Example:

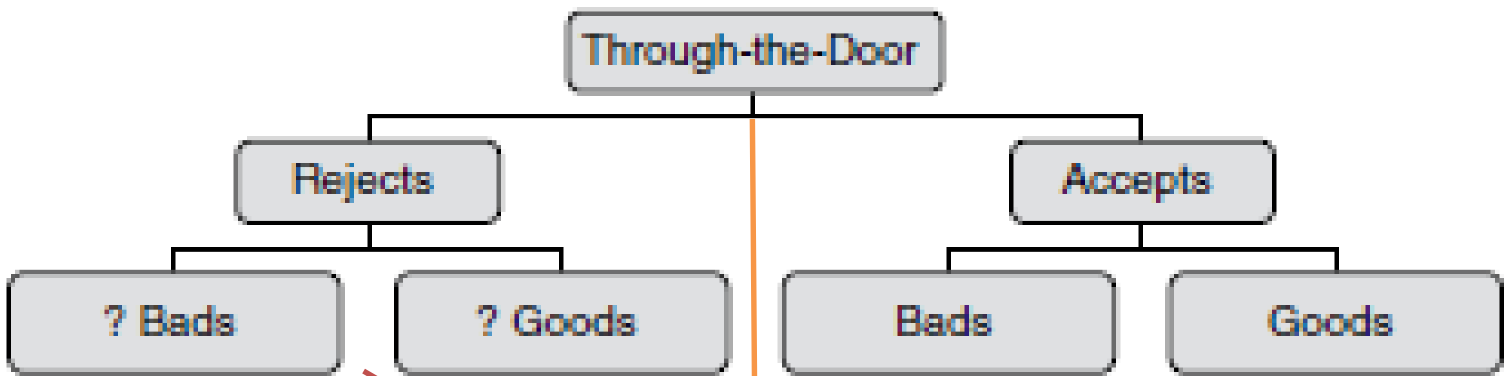


Figure 2.1 The Reject Inference Problem in Credit Scoring

And what about those who withdraw?

Sampling Bias, another example with stratified sampling

In stratified sampling, a sample is taken according to predefined strata.

strata

Stratified sample

Stratified Sampling

Considering in a fraud detection context, which data sets are typically very skewed (e.g., 99% Non-fraud vs 1% fraud).

Non-fraud

Fraud

Stratified sample

Discussion:
Why is there a bias?

®

Next

DATA PROCESSING

Data munging/wrangling

RAW DATA =>
Messy / noisy data

CLEANED DATA =>
Data that can be analyzed

Process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data.

Data Wrangling - Steps

1. Obtain
2. Understand
3. Transform
4. Augment
5. Shape

Dealing with Missing Values

No information of those students who withdraw.

Private data , such as salary, may not be disclosed.

Nonapplicable information

Undisclosed information

Why values are missing?

Error

Outlier values

Human factor – question was skipped by respondent, typo
Technical issue

The values have to be treated missing. E.g. extremely low or extremely high values

Dealing with Missing Values

Replace (Impute)

- Replacing the missing values with a known value (e.g. mean, median, mode)

Delete

- Deleting observations or variables with lots of missing values as the data may not be meaningful

Keep

- If the data with missing values are meaningful. Needs to be considered as a separate category.

Choosing the Right Way to Deal with Missing Values

Statistical test

- Test whether the missing information is related to the target variable
- If yes, then choose **keep**

Observe the number of available observations

- If available observations are high, then consider **delete**
- Else, consider **impute**

Dealing with Missing Values (Example)

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800		620	Churner
2	28	1,200	Single		Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married		Nonchurner
6	44				Nonchurner
7	22	1,200	Single		Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34		Single		Churner
10	50	2,100	Divorced		Nonchurner

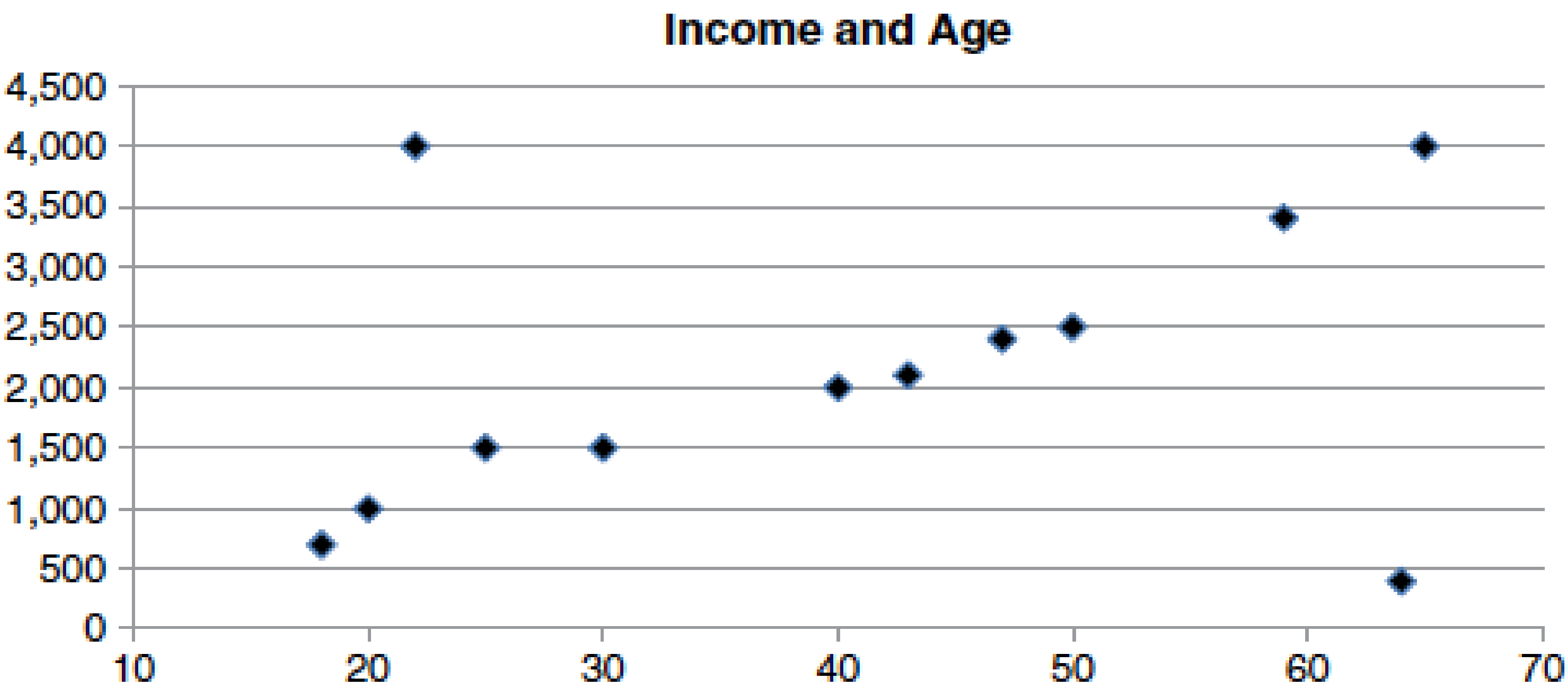
Suggest a way to deal with the missing values of Record 1, 6 and 10.

Dealing with Outliers

Valid observations	Invalid Observations
Salary of CEO is $1 million	Age = 300 years
?	?
?	?

Provide some examples of each type of outliers.

Multivariate Outliers



Multivariate outliers are observations that are outlying in multiple dimensions (e.g. age and income)

Steps to deal with Outliers

Detection

- Calculate the min and max
- Use visual tools, e.g. histograms, boxplots
- Z-scores

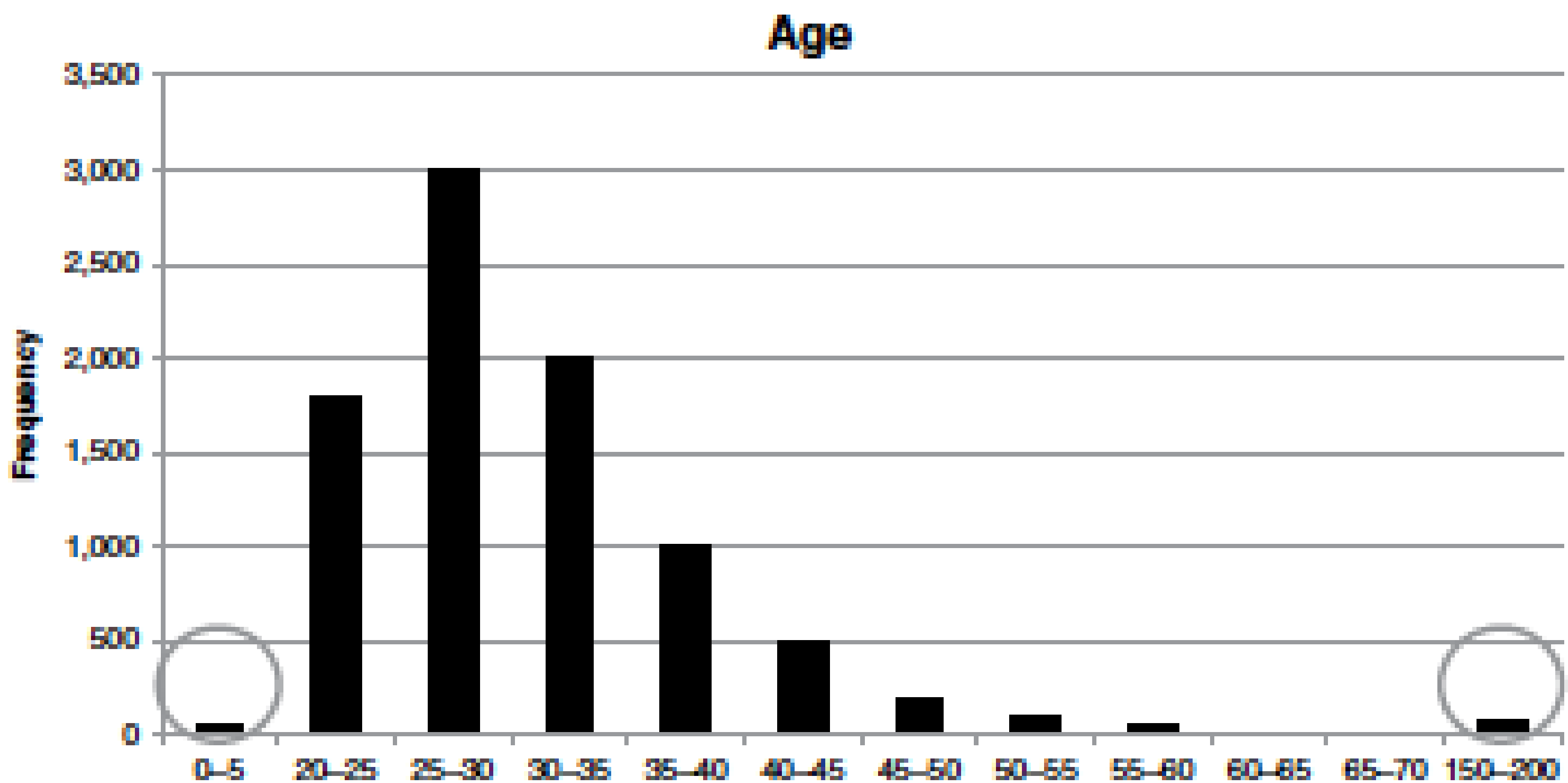
For univariate outliers

- regression

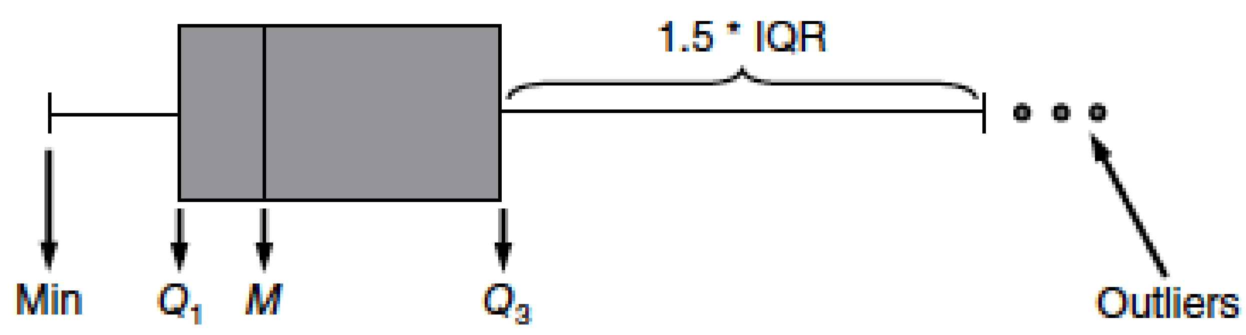
For multivariate outliers

Treatment

Using a histogram for Outliers Detection



Using a Boxplot for Outliers Detection



A box plot represents three key quartiles of the data: the first quartile (25% of the observations have a lower value), the median (50% of the observations have a lower value), and the third quartile (75% of the observations have a lower value).
The minimum and maximum values are then also added unless they are too far away from the edges of the box.

Too far away is then quantified as more than 1.5 * Interquartile Range ($IQR = Q_3 - Q_1$).

Using z-scores for Outliers Detection

- z-scores measures how many standard deviations, σ, an observation lies away from the mean, μ.

$$z_i = \frac{x_i - \mu}{\sigma}$$

A practical rule of thumb then defines outliers when the absolute value of the z -score $|z|$ is bigger than 3. Note that the z -score relies on the normal distribution.

Example of z-scores Calculation

ID	Age	Z-Score
1	30	$(30 - 40)/10 = -1$
2	50	$(50 - 40)/10 = +1$
3	10	$(10 - 40)/10 = -3$
4	40	$(40 - 40)/10 = 0$
5	60	$(60 - 40)/10 = +2$
6	80	$(80 - 40)/10 = +4$
...	...	...
	$\underline{\mu = 40}$ $\underline{\sigma = 10}$	$\underline{\mu = 0}$ $\underline{\sigma = 1}$

Based on the table above, which record could be an outlier?

Dealing with Multivariate Outliers

Multivariate outliers can be detected by

- fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).
- clustering or calculating the Mahalanobis distance.

Multivariate outlier detection is typically not considered in many modeling exercises due to the typical marginal impact on model performance.

Other useful methods that deal with Outliers

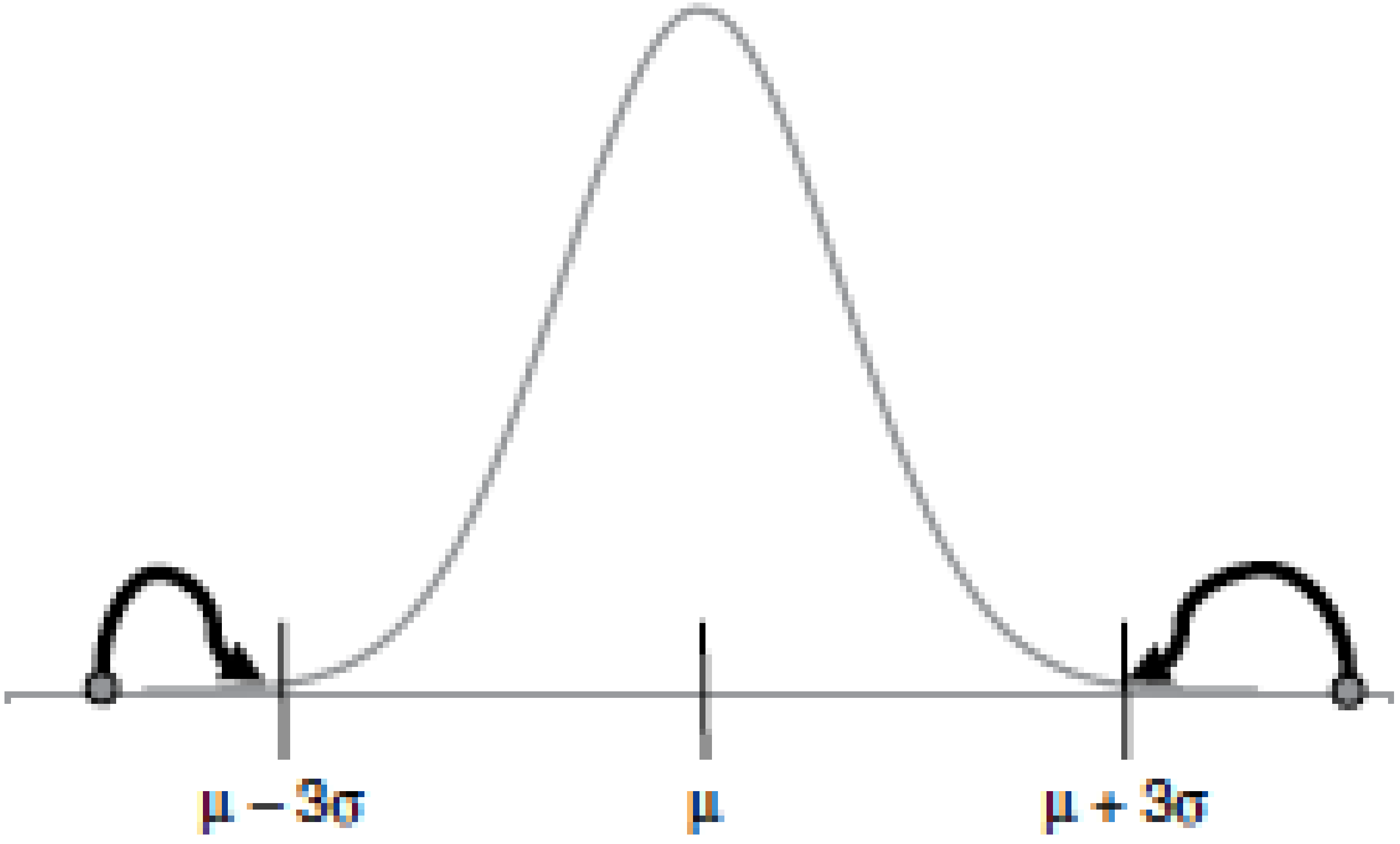
- Some analytical techniques (e.g., decision trees, neural networks, Support Vector Machines (SVMs)) are fairly robust with respect to outliers.
- A popular scheme is truncation/capping/winsorizing. One hereby imposes both a lower and upper limit on a variable and any values below/above are brought back to these limits. The limits can be calculated using the z -scores or the IQR.

Upper/lower limit = M ± 3s, with M = median and s = IQR/(2 x 0.6745)

- A sigmoid transformation ranging between 0 and 1 can also be used for capping

$$f(x) = \frac{1}{1+e^{-x}}$$

Using the Z-Scores for Truncation



Standardizing Data

- To scale variables to a similar range.
- E.g. Gender (0/1) vs. income ($0 - $1 million) – the coefficient of the relation of gender to income may be too small as their ranges are wide.
- Standardization is especially useful for regression-based analysis, but not needed for decision trees.

Standardizing Data

- Use Min/max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})}(newmax - newmin) + newmin,$$

whereby newmax and newmin are the newly imposed maximum and minimum (e.g., 1 and 0).

- Use z-score standardization
- Decimal scaling

Dividing by a power of 10 as follows: $X_{new} = \frac{X_{old}}{10^n}$, with n the number of digits of the maximum absolute value.

Categorization

- Categorization (also known as coarse classification, classing, grouping, binning, etc.) can be done for various reasons.
- For categorical variables, it is needed to reduce the number of categories. E.g. purpose of loan that may have 50 values initially should be reduced to fewer parameters.
- For continuous variables, categorization may also be very beneficial. E.g. age vs. hours using mobile phone each day

Methods of Categorization

Income = [1000, 1200, 1300, 2000, 1800, 1400]

Equal Interval Binning

- Bin 1: 1000, 1500
- Bin 2: 1500, 2000
- Do not take into account a target variable

Equal Frequency Binning

- Bin 1: 1000, 1200, 1300
- Bin 2: 1400, 1800, 2000
- Do not take into account a target variable

Exercise

- Suggest two suitable categories for the following age values using each of the following methods
  – (a) equal interval binning
  – (b) equal frequency binning

Income = [10, 12, 13, 17, 20, 18, 14, 24]

Using Chi-Squared Analysis in Coarse Classification

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6000	1600	350	950	90	10	9000
Bads	300	400	140	100	50	10	1000
Total	6300	2000	490	1050	140	20	10000

- Suggest 3 categories you would like to group the data above

Empirical and Individual Frequencies for Coarse Classifying Residential Status

Attribute	Owner	Renter	Others	Total
Goods	6000	1950	1050	9000
Bads	300	540	160	1000
Total	6300	2490	1210	10000

The more the numbers in both tables differ, the less independence, hence better dependence and a better coarse classification.

Attribute	Owner	Renter	Others	Total
Goods	5670	2241	1089	9000
Bads	630	249	121	1000
Total	6300	2490	1210	10000

Table 2: Individual Frequencies for Coarse Classifying Residential Status

The independence frequencies can be calculated as follows:

6,300/10,000 x 9,000/10,000 x 10,000 = 5,670

Empirical Frequencies for Coarse Classifying Residential Status

- Formally, one can calculate the chi-squared distance as follows:

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(1950-2241)^2}{2241} + \frac{(540-249)^2}{249} + \frac{(1050-1089)^2}{1089} + \frac{(160-121)^2}{121} = 583$$

Exercise

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6000	1600	350	950	90	10	9000
Bads	300	400	140	100	50	10	1000
Total	6300	2000	490	1050	140	20	10000

- Based on Option 2: [Owner, With Parents, Others]
- Compute the

1. Empirical frequencies for Coarse Classifying Residential Status
2. Individual frequencies for Coarse Classifying Residential Status
3. The Chi-squared distance

Discussion

- The Chi-squared distance for Option 2 is shown below:

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(950-945)^2}{945} + \frac{(100-105)^2}{105} + \frac{(2050-2385)^2}{2385} + \frac{(600-265)^2}{265} = 662$$

- Reviewing that the Chi-squared distance of Option 1 is 583, discuss which categorization is better, Option 1 or Option 2?

®

Next section

VISUAL DATA AND EXPLORATORY STATISTICAL ANALYSIS

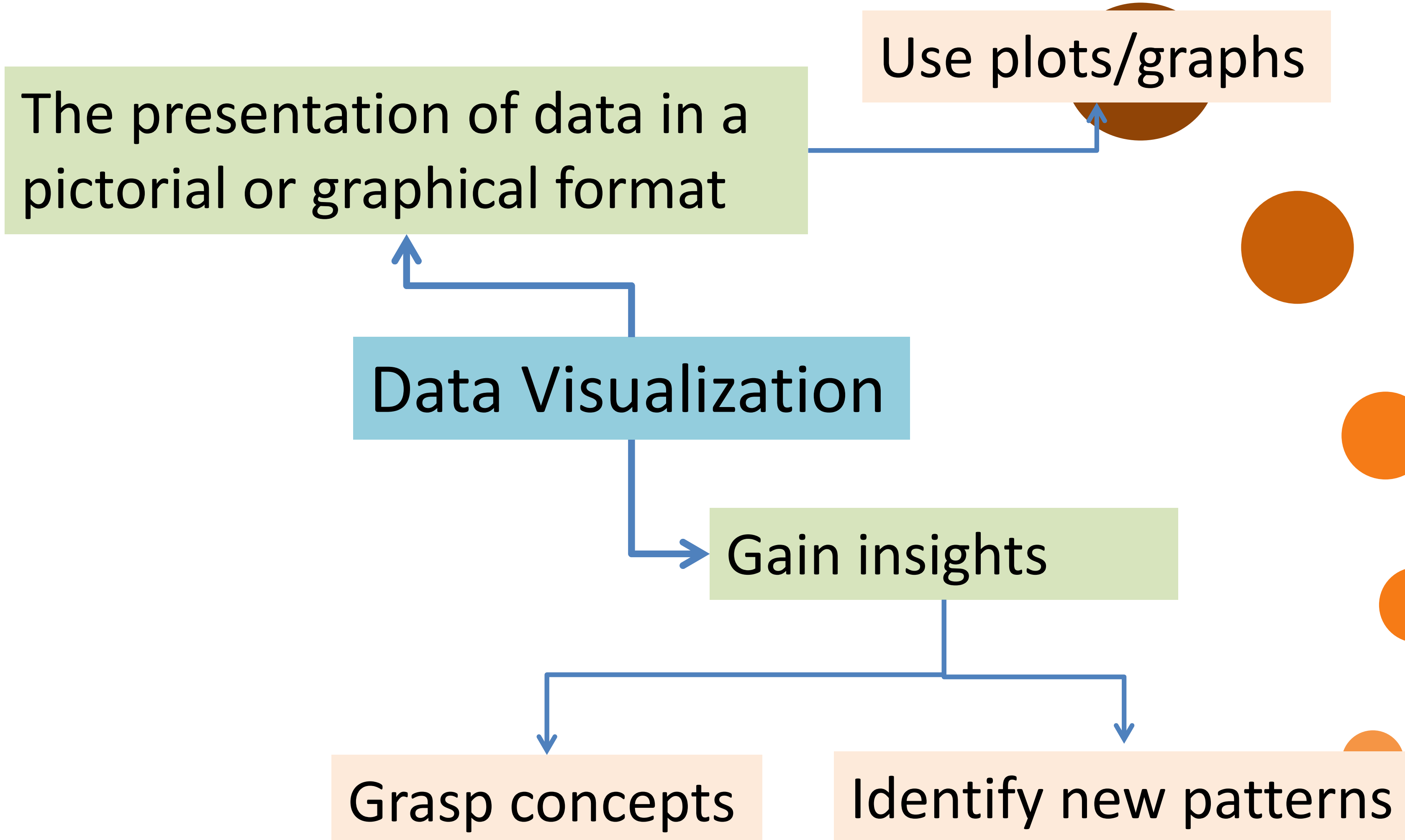
Exploratory Statistical Analysis

Some basic statistical measurements

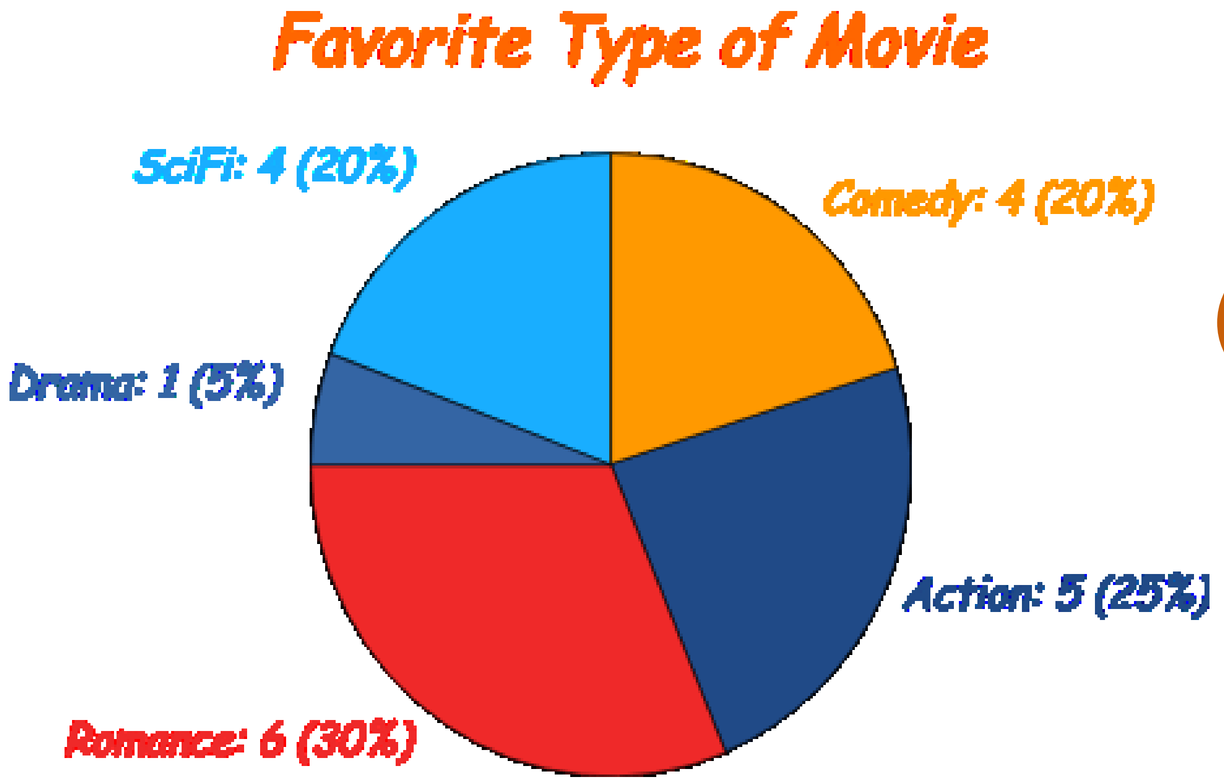
- averages,
- standard deviations,
- minimum,
- maximum,
- percentiles,
- confidence intervals.

One could calculate these measures separately for each of the target classes (e.g., good versus bad customer) to see whether there are any interesting patterns present (e.g., whether bad payers usually have a lower average age than good payers).

Data Visualization



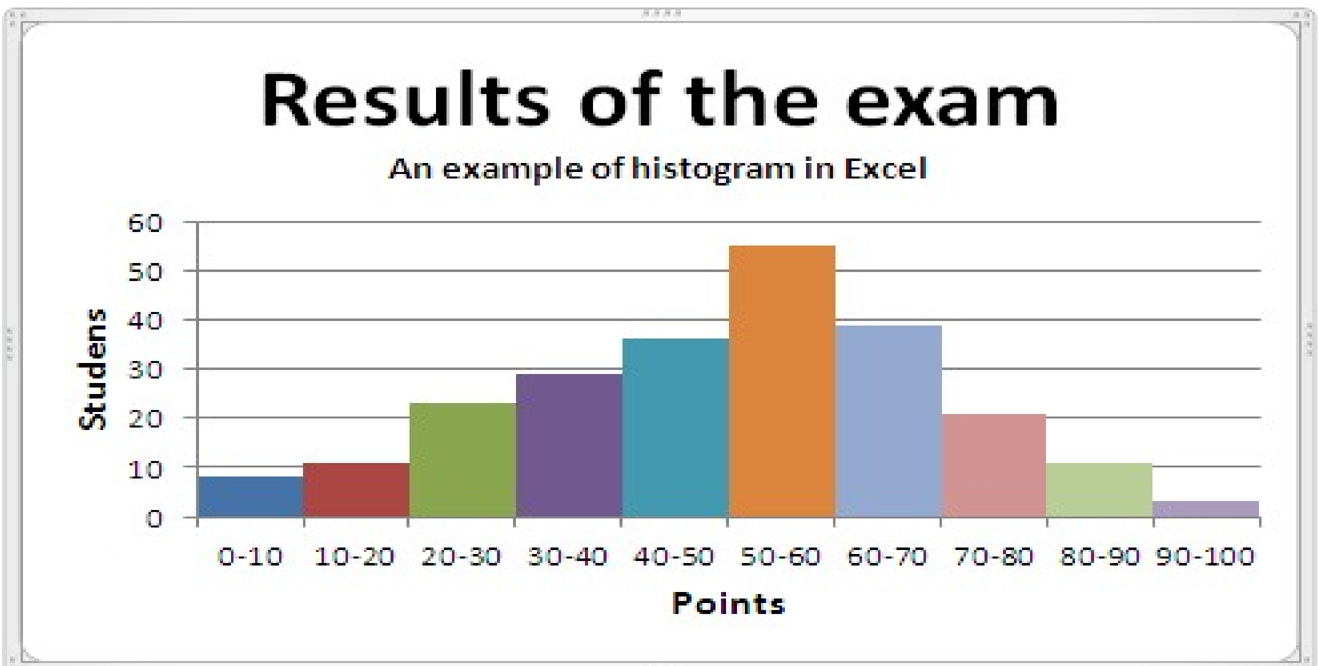
Example 1: Pie Chart



1. What can you tell from this chart?
2. As a decision maker, why is this important for you?

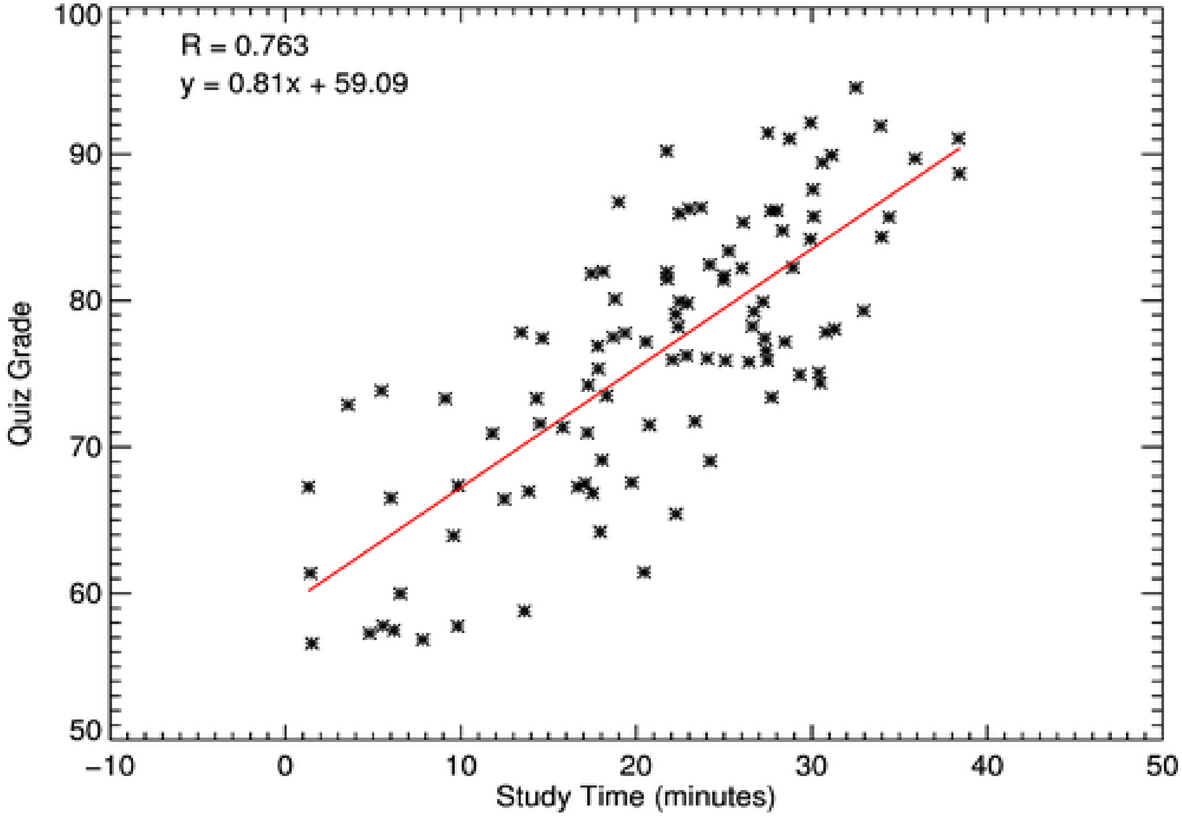
Example 2: Histogram

- A histogram provides an easy way to visualize the central tendency and to determine the variability or spread of the data. It also allows you to contrast the observed data with standard known distributions (e.g., normal distribution).



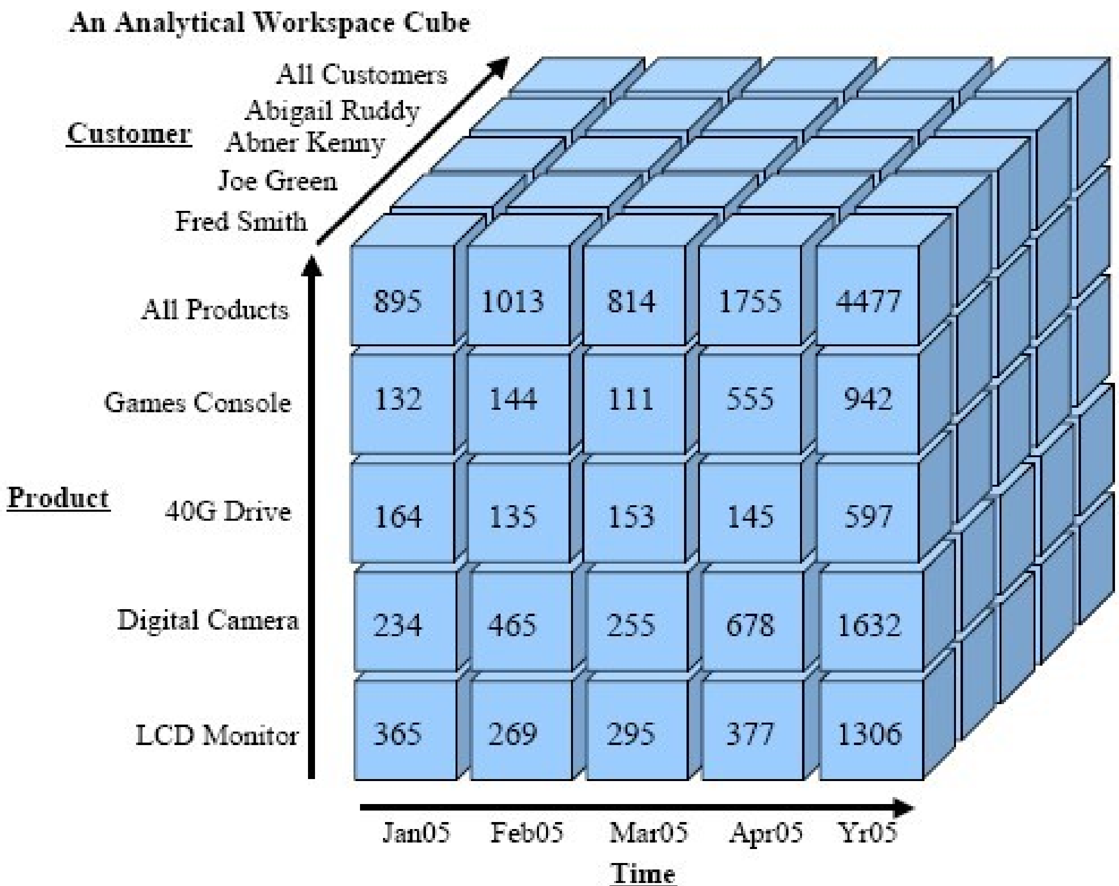
Example 3: Scatter Plot

- Scatter plots allow you to visualize one variable against another to see whether there are any correlation patterns in the data.



Example 4: OLAP

- OLAP-based multidimensional data analysis can be usefully adopted to explore patterns in the data



®

Next

DATA STORAGE AND MANAGEMENT OF BIG DATA

Data storage and management of Big Data

capacity

speed

volume

Increasing unstructured data

Big data storage requirements

performance

Throughput

cost

access time and data transfer rate

Input-Output-Operations-Per Second (IOPS)

scalability

Reliability

Challenges in Big Data Storage

Storage Mediums Issues

- Mechanical disk drives (HDD) - overheating and magnetic faults, and disk access overhead
- Solid State Disk (SSD) – more reliable but price per gigabyte is high

Comparisons of Storage Mediums

Table 1. Storage Mediums Characteristics

	MAGNETIC STORAGE	OPTICAL STORAGE	SOLID STATE	HYBRID STORAGE
CAPACITY	Up to 6TB	Up to 50GB	Up to 2TB	Up to 6TB
ACCESS TIME	Relatively low latency	Slow access time	Very low latency	Low latency
COST	Less expensive	Relatively cheaper	Very expensive	Less expensive
DATA TRANSFER RATE	Relatively low transfer rate	Good transfer rate	High transfer rate	High transfer rate

Rajeev Agrawal & Christopher Nyamful, 2016

Current Storage Architectures for Big Data Storage

Cloud providers offers **Storage-as-a-Service (StaaS)**

- e.g. Drop-box, Google Drive, Google Docs, and Microsoft. StaaS allows organizations and individuals to rent a storage space at a relatively minimal cost.

Network Attached Storage (NAS)

- designed for file sharing. It uses a high level abstraction that enables cross-platform data sharing. NAS comes with a processor and software for management and backup of data. The massive increase of data and widespread of mobile devices means limited connectivity to files on this system

Current Storage Architectures for Big Data Storage

Storage Area Network over IP (IP-SAN)

- provides the platform where thousands of computers connect to share a large amount of storage devices that range from simple disks to large, high-performance, high-functioning storage system.
- It is less expensive.
- can span over a wide geographical area.

Object-based Storage

- storage object is said to be collection of bytes on a storage device, with methods for accessing data, and security policies to prevent unauthorized access. . Because of the variable-length nature of object, it makes it ideal to store the different types of data. Objects can be seen as the union of both file and block technologies.

References

- Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons.
- Agrawal, R. & Nyamful, C. (2016). Challenges of big data storage and management. *Global Journal of Information Technology*. *6*(1), 01-10.
- Ashwini Kuntamukkala (2016) https://www.slideshare.net/AshwiniKuntamukkala/data-wrangling-62017599