



CSC 3303 BIG DATA ANALYTICS

1

DR. RAINI HASSAN

Department of Computer Science, KICT, IIUM

Office Room: C2-14 (Block C, Level 2)

Email: hrai@iium.edu.my

Office Telephone No: +603 6196 5655



Dua Before Studying

2

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Dua Before Studying

Oh Allah! Make useful for me what you have taught me and teach me knowledge that will be useful to me. Oh Allah! I ask you for the understanding of the prophets and the memory of the messengers, and those nearest to you. Oh Allah! Make my tongue full of your remembrance and my heart with awe of you. Oh Allah! You do whatever you wish, and you are my availer and protector and best of aid.

اللَّهُمَّ انْفَعْنِي بِمَا عَلَّمْتَنِي وَ عَلَّمْنِي مَا يَنْفَعُنِي
اللَّهُمَّ إِنِّي أَسْأَلُكَ فَهَمَ النَّبِيِّينَ وَ حِفْظَ الْمُرْسَلِينَ الْمُقَرَّبِينَ
اللَّهُمَّ اجْعَلْ لِسَانِي عَامِرًا بِذِكْرِكَ وَ قَلْبِي بِخَشْيَتِكَ .
إِنَّكَ عَلَى مَا تَشَاءُ قَدِيرٌ وَ أَنْتَ حَسْبُنَا اللَّهُ وَ نِعْمَ الْوَكِيلُ

*Allahumma infan'ni bimaa 'allamtanii wa'allimnii maa
yanfa'uunii. Allahumma inii as'aluka fahmal-nabiyyen wa hifzal mursaleen
al-muqarrabeen. Allahumma ij'al leesanee 'aiman bi dhikrika wa qalbi bi
khashyatika. Innaka 'ala ma-tasha'u qadeer wa anta hasbun-allahu wa
na'mal wakeel.*



Data Collection, Sampling and Pre-Processing

3

| DATA COLLECTION, SAMPLING, AND PREPROCESSING, IN ANALYTICS
IN A BIG DATA WORLD: THE ESSENTIAL GUIDE TO DATA SCIENCE AND
ITS APPLICATIONS | BAESENS B. | 2012 | JOHN WILEY & SONS |
AND
| MY OWN ADDITIONAL OUTLINES AND CONTENTS, PREPARED FROM
MULTIPLE RESOURCES |



Outlines

4

1. Introduction
2. Types of Data Sources
3. Sampling
4. Types of Data Elements
5. Visual Data Exploration
6. Missing Values
7. Outlier Detection and Treatment
8. Variable Selection
9. Standardizing Data (*additional information*)
10. Categorization (*additional information*)
11. Weights of Evidence Coding (*additional information*)
12. Segmentation (*additional information*)



1. Introduction

5



Introduction [1]

6

- Data are the key ingredients for any analytical exercise.
- The general rule is **the more data, the better**.
- However, in real life the data can be dirty because of:
 - inconsistencies
 - incompleteness
 - duplication
 - merging problems

Just observe what happens when you copy/export your contacts from your old smartphone into a new one, or from one mobile Operating System into a different one.



Introduction [2]

7

- Therefore, the data need to be:
 - cleaned up
 - reduced to a manageable and relevant size
- There are various data filtering mechanisms to do (*more details later*)
- Messy data will yields messy analytical models...
 - Garbage-In, Garbage-Out (GIGO)



Introduction [3]

8

- The role of pre-processing is to avoid mistakes that can make the data useless for further analysis.
- Therefore, all steps in pre-processing must be:
 - carefully justified
 - carried out
 - validated
 - documented
- ... before proceeding with further analysis.



2. Types of Data Sources



Types of Data Source [1]

10

- Data can originate from a variety of different sources.
 1. Transactions data
 2. Unstructured data
 3. Qualitative/expert-based data
 4. Publicly available data



Types of Data Source [2]

11

1. Transactions Data:

- It consist of **structured, low-level, detailed information** capturing the key characteristics of a customer transaction (e.g., *purchase, claim, cash transfer, credit card payment*).
- It is usually stored in massive online transaction processing (OLTP) relational databases.
- It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.



Types of Data Source [3]

12

2. Unstructured Data [1]:

- It refers to the **data embedded in text documents** (e.g., emails, web pages, files, PDFs or documents such as claim forms) or multimedia content.
- ***Sensitive unstructured data** is usually data that was first created in a protected structured system such as SAP Financials for example, and then exported into an Excel spreadsheet for easier consumption by audiences who are not SAP users.

****SAP** stands for Systems, Applications and Products in data processing. It is and Enterprise Resource Planning (ERP) system by SAP AG, a company based in Walldorf, Germany.

* <https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#a888658493a3>

** <https://www.saponlinetutorials.com/what-is-sap-erp-system-definition/>



Types of Data Source [4]

13

2. Unstructured Data [2]:

- *However, there are a few existing problems with unstructured data:
 - A lack of tools that easily manage unstructured data. Tools need to provide efficient text parsing and analytics, taxonomy and metadata management.
 - Difficulty integrating unstructured data with existing information systems. The two are often seen as apples and oranges when it comes to analytics and decision making.
 - Shortage of skills in existing staff.
 - Missing sense of urgency for managing unstructured data.
- Therefore, this kind of data require **extensive pre-processing** before they can be successfully included in an analytical exercise.

* <https://www.wired.com/insights/2013/09/whats-the-big-deal-with-unstructured-data/>



Types of Data Source [5]

14

3. Qualitative/expert-based data [1]:

- An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., *credit portfolio manager, brand manager*).
- The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run.
- This will steer the modeling in the right direction and allow you to interpret the analytical results from the right perspective.



Types of Data Source [6]

15

3. Qualitative/expert-based data [2]:

- A popular example of applying expert-based validation is checking the univariate signs of a regression model.
- For example, one would expect **a priori** that higher debt has an adverse impact on credit risk, such that it should have a negative sign in the final scorecard.
- If this turns out not to be the case (e.g., *due to bad data quality, multicollinearity*), the expert/business user will not be tempted to use the analytical model at all, since it contradicts prior expectations.



Types of Data Source [7]

16

4. Publicly available data:

- **Examples:**
 - Government data – macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on.
 - Social media data – Facebook, Twitter, Instagram, etc.
- However, since it is a public data, before being used, all data gathering respects both local and international privacy regulations are adhered to.



3. Sampling

17



Sampling [1]

18

- The aim of sampling is **to take a subset of past customer data and use that to build an analytical model.**
- It means we need to do sampling.
- On the other hand...
 - with the availability of high performance computing facilities (e.g., grid/cloud computing), one could also directly analyse the full data set.
- In order to avoid biased sample, a key requirement for a good sample is that it should be representative of the future customers on which the analytical model will be run.



Sampling [2]

19

- The timing aspect becomes important because customers of today are more similar to customers of tomorrow than customers of yesterday.
- Choosing the optimal time window for the sample involves:
 1. A trade-off between lots of data (*and hence a more robust analytical model*)
 2. Recent data (*which may be more representative*).
- The sample should also be taken from an average business period to get a picture of the target population that is as accurate as possible.



Sampling [3]

20

*A national survey of 1,520 adults conducted March 7 April 4, 2016, finds:

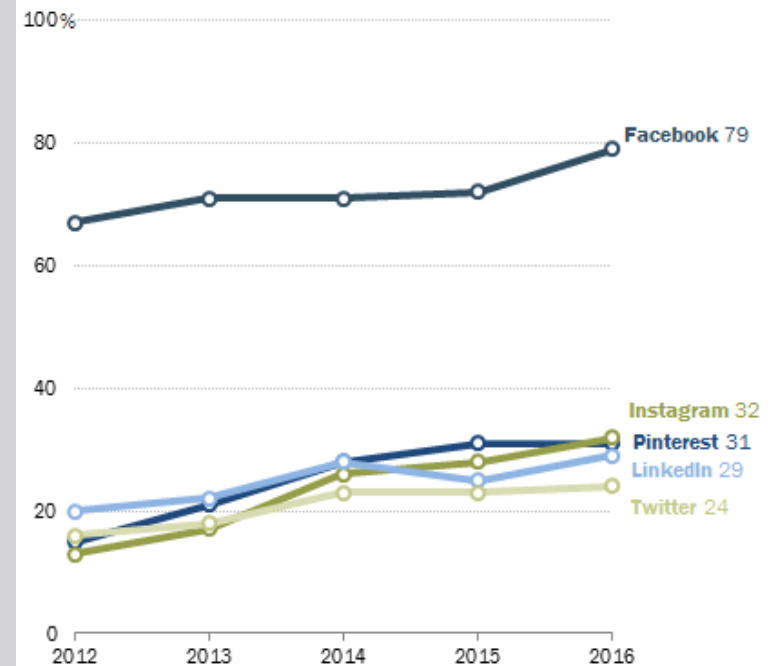
1. Facebook (79%)
2. Twitter (24%)
3. Pinterest (31%)
4. Instagram (32%)
5. LinkedIn (29%)

Will Facebook still dominate the online users 5 years from now?

* <http://www.pewinternet.org/2016/11/11/social-media-update-2016/>

Facebook remains the most popular social media platform

% of online adults who use ...



Note: 86% of Americans are currently internet users
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER



4. Types of Data Elements



Types of Data Elements [1]

22

- It is important to appropriately consider the different types of data elements at the start of the analysis.
- There are 2 types:
 1. Continuous
 2. Categorical



Types of Data Elements [2]

23

1. Continuous:

- These are data elements that are defined on an interval that can be limited or unlimited.
- Examples include:
 - Income
 - Sales
 - Temperature



Types of Data Elements [3]

24

2. Categorical:

- Can be further divided into the following:
 1. **Nominal**
 - Can only take on a limited set of values with no meaningful ordering in between.
 - Examples include marital status, profession, purpose of loan.
 2. **Ordinal**
 - Can only take on a limited set of values with a meaningful ordering in between.
 - Examples include credit rating; age coded as young, middle aged, and old.
 3. **Binary**
 - can only take on two values.
 - Examples include gender, employment status.



Types of Data Elements [4]

25

- It is of key importance to appropriately distinguishing between these different data elements before when importing the data into an analytics tool.
- For example, if marital status were to be incorrectly specified as a continuous data element, then the software would calculate its mean, standard deviation, and so on, which is obviously meaningless.



5. Visual Data Exploration

26



Visual Data Exploration [1]

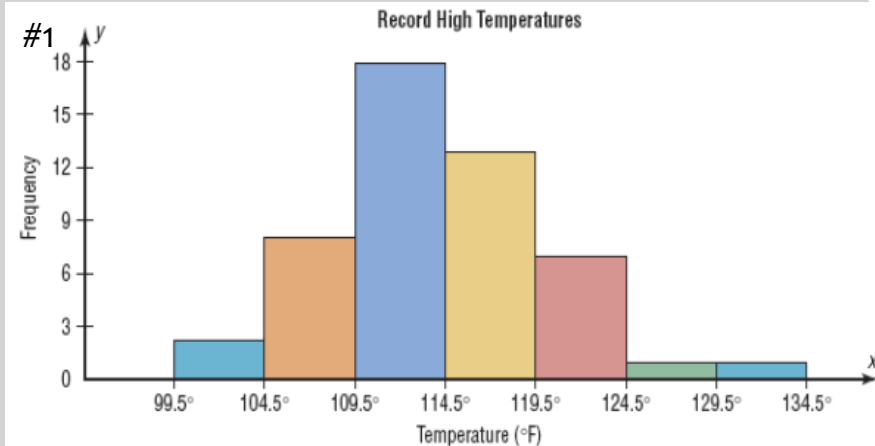
27

- It is a very important part of getting to know data in an “informal” way.
- It allows you to get some initial insights into the data, which can then be usefully adopted throughout the modeling.
- Different plots/graphs can be useful here.
- For examples:
 - Histogram
 - Ogives
 - Frequency polygon
 - Pie chart

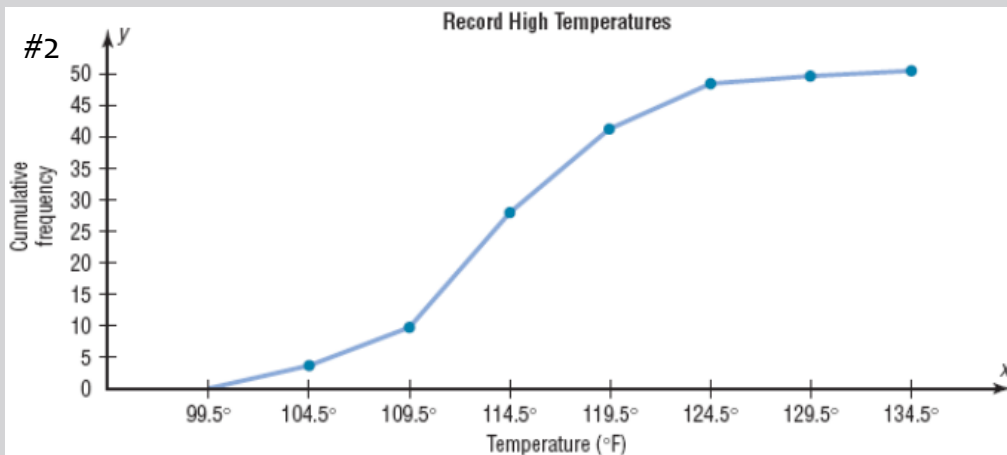
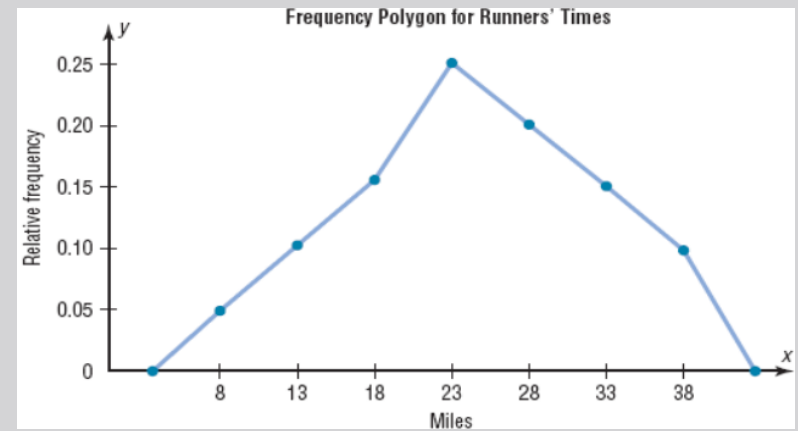


Visual Data Exploration [2]

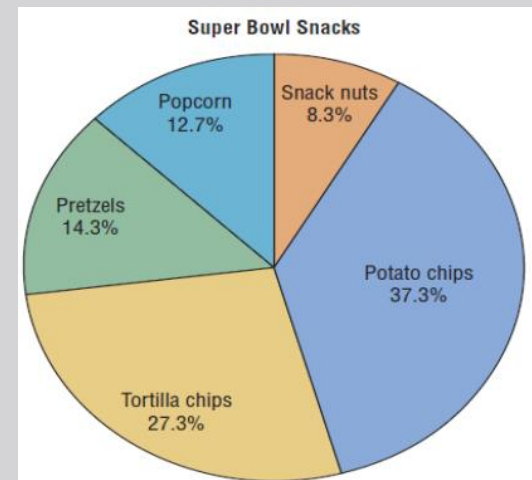
28



#3



#4



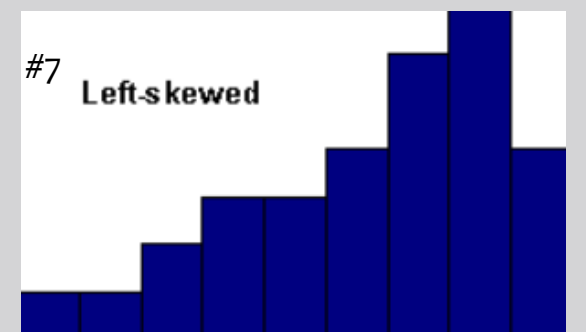
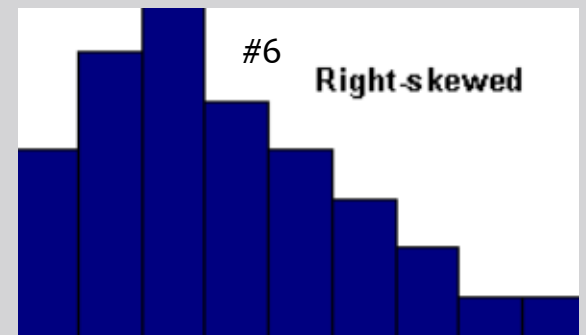
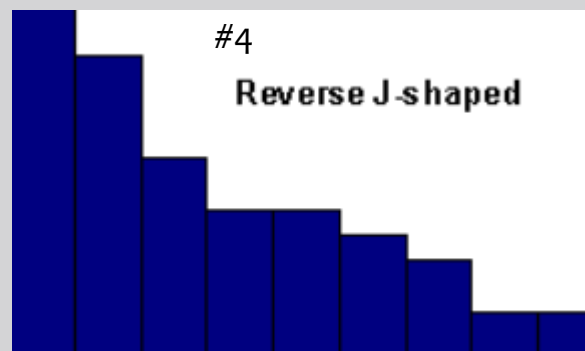
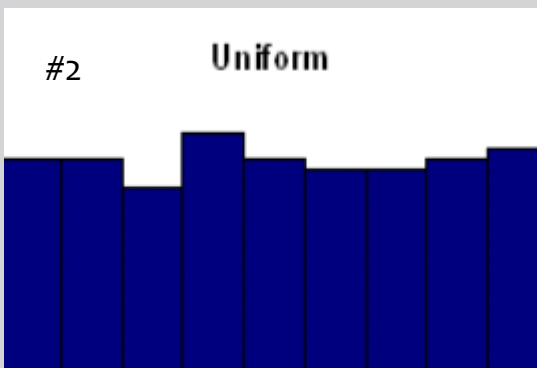
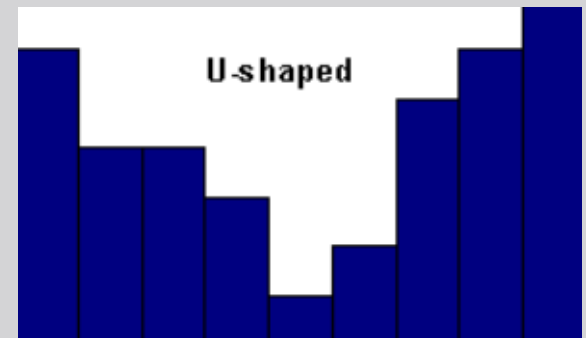
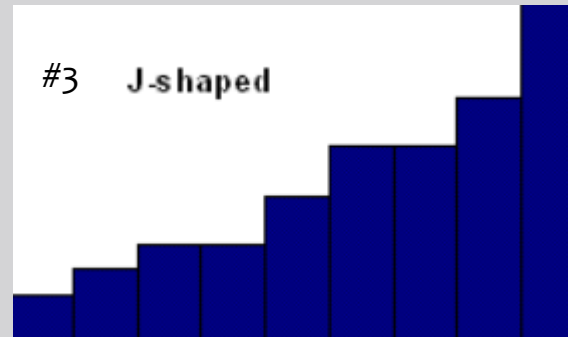
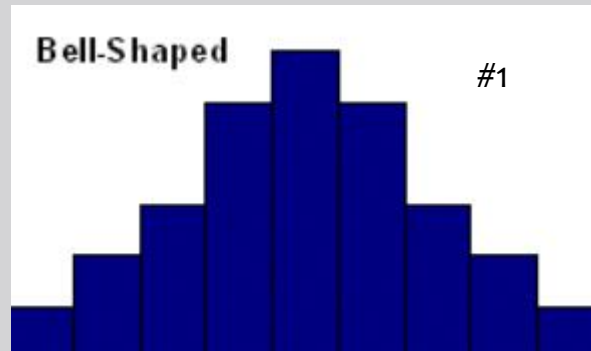
Sources #1 to 4: Alan G. Bluman, Chapter 2, Elementary Statistics: A Step by Step Approach (10th edition), 2017, McGraw-Hill Education



Visual Data Exploration [3]

29

- The shapes of the plots/graphs can also give some initial insights:
- Examples for histogram:



Sources #1 to 7: Alan G. Bluman, Chapter 2, Elementary Statistics: A Step by Step Approach (10th edition), 2017, McGraw-Hill Education

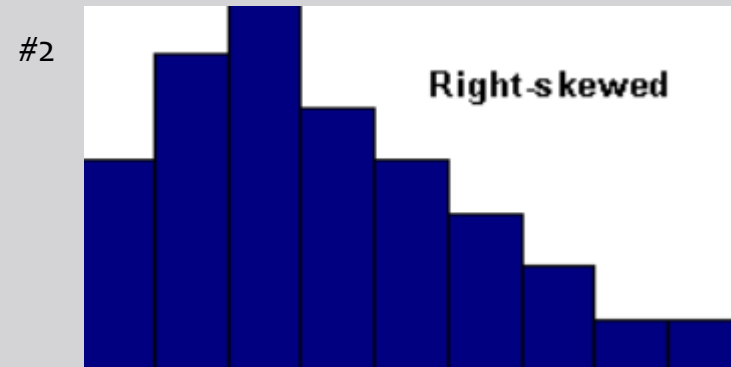
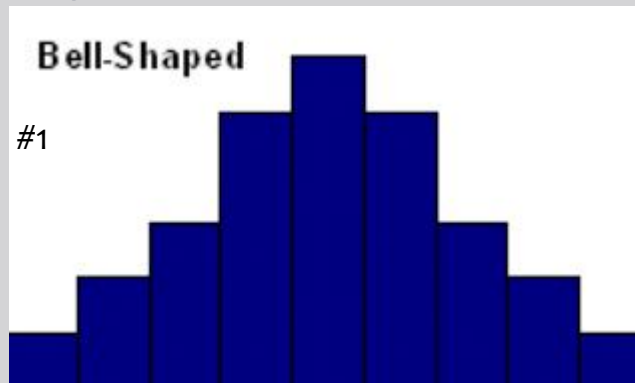


Visual Data Exploration [4]

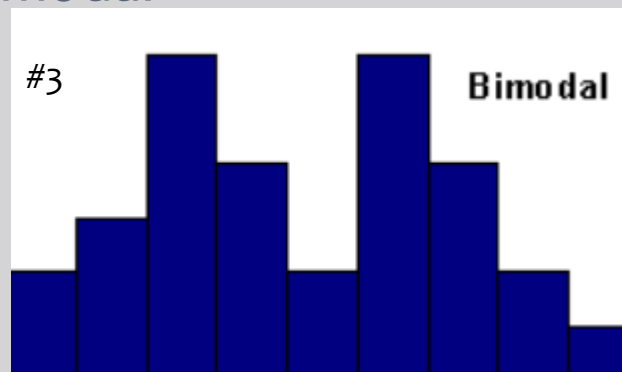
30

- Other than that, we can take a look at the peak(s)

- Unimodal



- Bimodal



Sources #1 to 3: Alan G. Bluman, Chapter 2, *Elementary Statistics: A Step by Step Approach* (10th edition), 2017, McGraw-Hill Education



Visual Data Exploration [5]

31

- A next step after visual analysis could be inspecting some basic statistical measurements, such as:
 - averages
 - standard deviations
 - 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*).



6. Missing Values

32



Missing Values [1]

33

- Missing values can occur because of various reasons:
 - The information can be non-applicable.

For example, “salary” is only available for those data with subjects that are employed.
 - The information can also be undisclosed.

For example, a customer decided not to disclose his or her income because of privacy.
 - Missing data can also originate because of an error during merging (e.g., *typos in name or ID*).



Missing Values [2]

34

- Some analytical techniques (e.g., *decision trees*) can directly deal with missing values.
- Other techniques need some additional pre-processing.
- The following are the most popular schemes to deal with missing values:
 1. Replace (impute)
 2. Delete
 3. Keep



Missing Values [3]

35

1. Replace (impute):

- This implies replacing the missing value with a known value.
- Example 1:
 - For missing total sales, can be replaced with the average or median of the known values.

Total Sales (RM)
(R1) 15000
(R2) 17500
(R3) ?
(R4) 22500
(R5) 50000
(R6) 14275


$$\bar{x} / R3 = (R1 + R2 + R4 + R5 + R6) / 5 = \text{RM}23855$$



Missing Values [4]

36

1. Replace (impute):

- Example 2:

- For marital status, can be replaced with the mode of the known values.

Marital Status
(R1) Married
(R2) Divorced
(R3) Married
(R4) Widowed
(R5) ?
(R6) Single

→ **Mode or R5 = Married**



Missing Values [5]

37

2. Delete:

- This is the most straightforward option and consists of deleting observations or variables with lots of missing values.
- This, of course, assumes that information is missing at random and has no meaningful interpretation and/or relationship to the target.



Missing Values [6]

38

3. Keep:

- Missing values can be meaningful (e.g., a customer did not disclose his or her income because he or she is currently unemployed).
- Obviously, this is clearly related to the target (e.g., good/bad risk) and needs to be considered as a separate category.



Missing Values [7]

39

What to do initially with the missing value?:

- As a practical way of working, one can first start with statistically testing whether missing information is related to the target variable (using, for example, a chi-squared test).
- If yes, then we can adopt the keep strategy and make a special category for it.
- If not, one can, depending on the number of observations available, decide to either delete or impute.



7. Outlier Detection and Treatment

40



Outlier Detection and Treatment [1]

41

- Outliers are extreme observations that are very dissimilar to the rest of the population.
- In statistics, they are referring to the data values that are extremely high or extremely low.
- Two important steps in dealing with outliers are:
 1. Detection
 2. Treatment



Outlier Detection and Treatment [2]

42

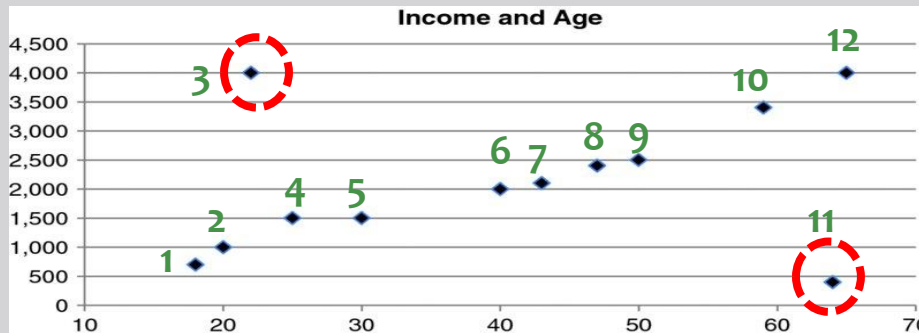
- Two types of outliers can be considered:
 - ❖ Valid observations (e.g., salary of boss is \$1 million)
 - ❖ Invalid observations (e.g., age is 300 years)
- Both are univariate outliers in the sense that they are outlying on one dimension/variable – “Salary” only.
- However, outliers can be hidden in unidimensional views of the data.
- Multivariate outliers are observations that are outlying in multiple dimensions.




Outlier Detection and Treatment [3]

43

- One example is “Income” and “Age” – 2 dimensions, 2 variables.



 Potential outliers

3:
Income = RM4000 & Age = 22

11:
Income = RM400 & Age = 64

- Could be coming from this table:

No.	Income (RM)	Age
1.	700	18
2.	1000	20
⋮	⋮	⋮
12.	4000	65



Outlier Detection and Treatment [4]

44

- For univariate data: by glancing through the data, if the data seems to be off, perform a general check:
 - ❖ I would start by checking the original/raw data; that could be from the questionnaire and somehow being recorded inaccurately.
 - ❖ If there is nothing wrong with the record, then the min and max values need to be checked – I would choose to determine the min and max range by using the Five-Number Summary and IQR.
 - ❖ If each of the data value or observations are within the range, then all is good, else, there is/are outlier(s).
- Various graphical tools also can be used to detect outliers;
 - ❖ can be applied to either or both univariate and multivariate data.



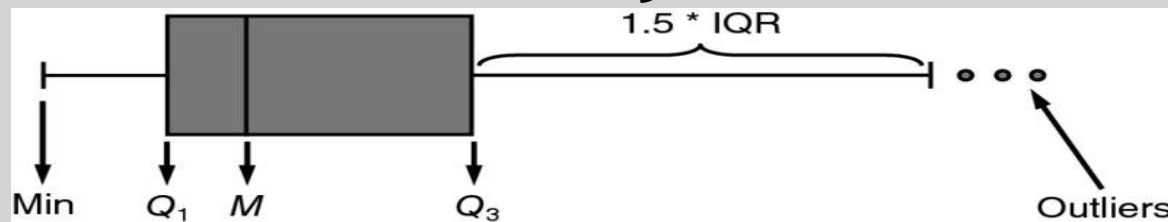
Outlier Detection and Treatment [5]

45

Graphical Tools - Box Plots:

Univariate outliers

- The 5 Numbers Summary: Low (min), Q_1 , Q_2 or MD, Q_3 , and High (max).
- The Q_1 , Q_2 or MD, Q_3 are in the box.
- 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 $1.5 * \text{Interquartile Range}$ ($\text{IQR} = Q_3 - Q_1$).



Outlier Detection and Treatment [6]



46

CLASS ACTIVITY:

Let's process this data as examples for the previous 2 slides:

1. Find the min and max range of data.
2. Create a box plot, to identify outliers

30 39 47 48 78 89 138 164 215 296

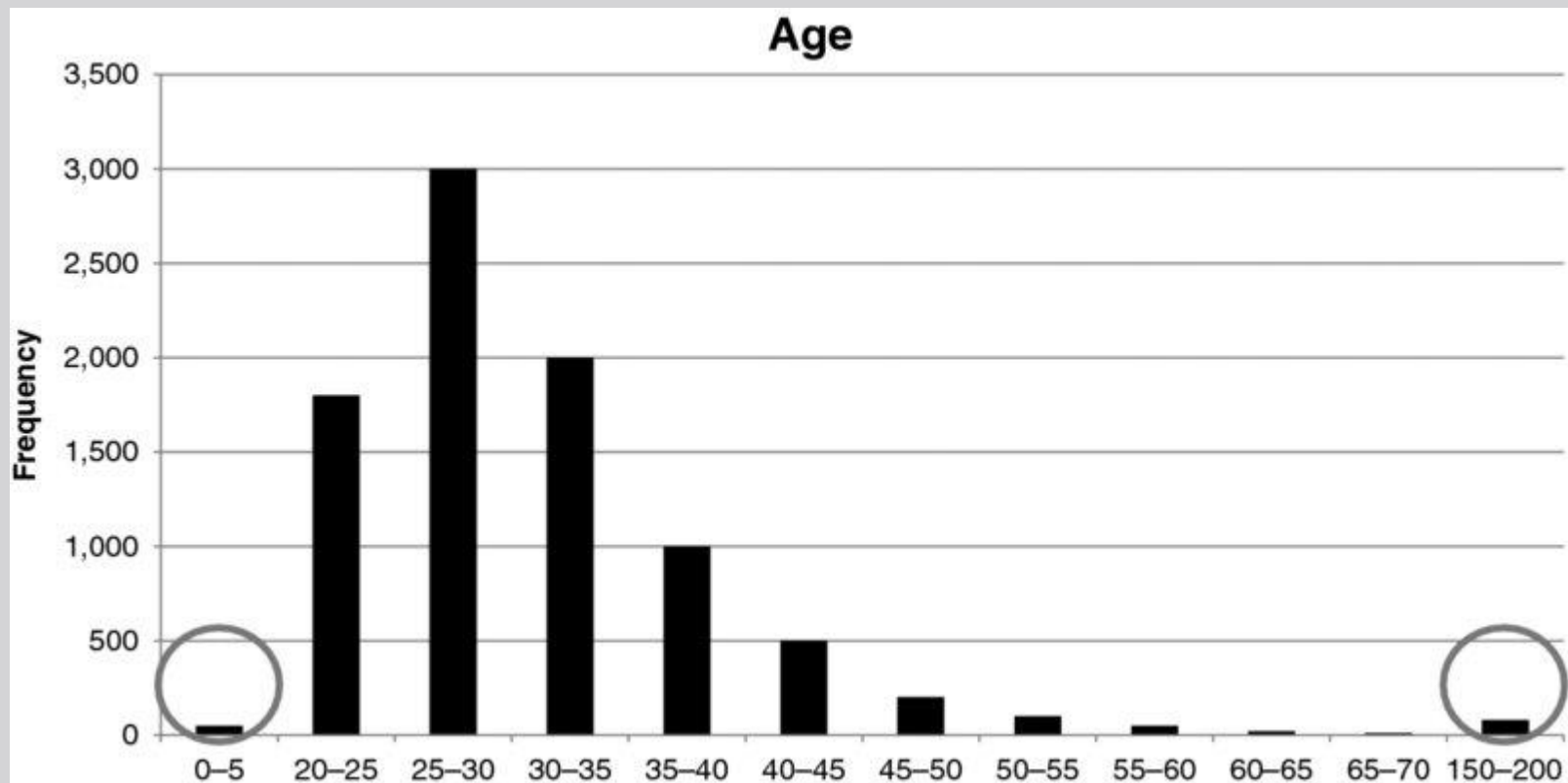


Outlier Detection and Treatment [7]

47

Graphical Tools - Histograms:

Univariate outliers





Outlier Detection and Treatment [8]

48

Statistical Method – z-score:

Univariate outliers

- z-score measures how many standard deviations an observation lies away from the mean, as follows:

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

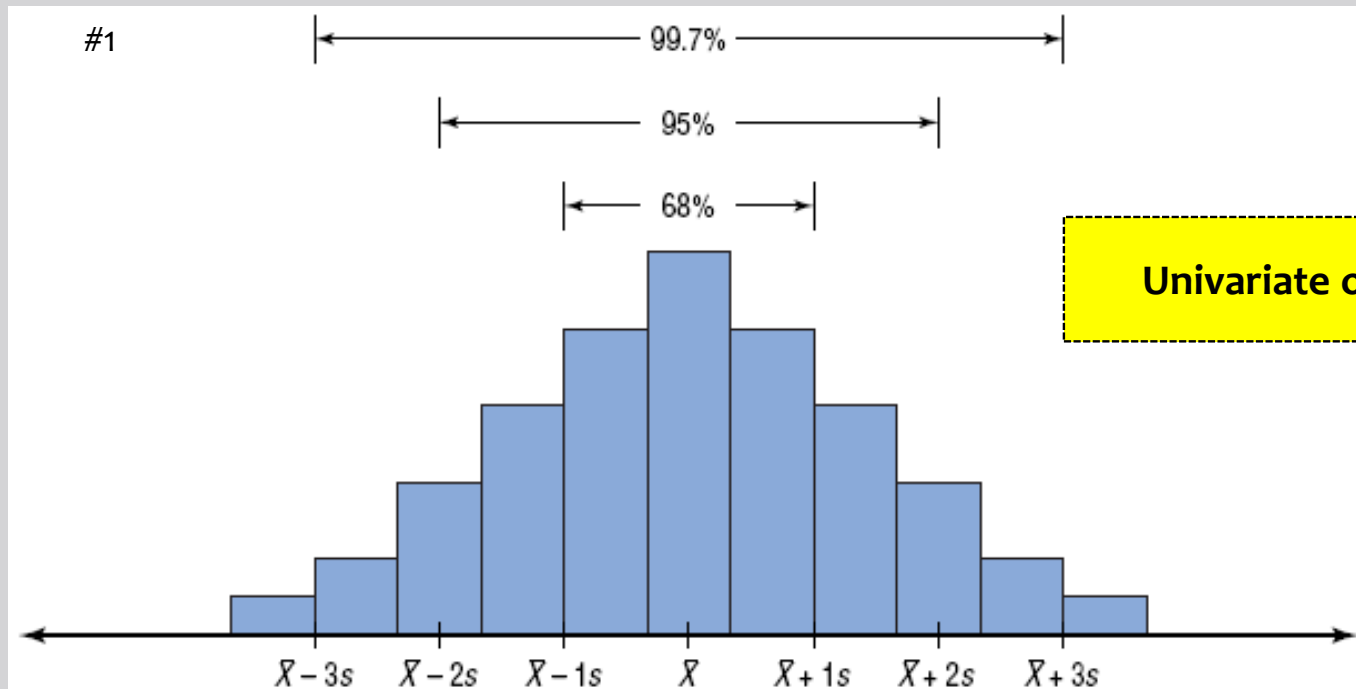
- where μ represents the average of the variable, σ represents its standard deviation, and x represents an observation or a data value.
- A practical rule of thumb then defines outliers when the absolute value of the z-score $|z|$ is bigger than 3.
- Note that the z-score relies on the normal distribution.



Outlier Detection and Treatment [9]

49

Statistical Method – z-score:



#1: Alan G. Bluman, Chapter 3, *Elementary Statistics: A Step by Step Approach* (10th edition), 2017, McGraw-Hill Education

Outlier Detection and Treatment [10]



50

Multivariate Outliers:

- There are many methods to detect multivariate outliers:
 - ❖ By fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).
 - ❖ By clustering or calculating the Mahalanobis distance.
 - ❖ By applying some analytical/machine learning techniques (e.g., decision trees, neural networks, Support Vector Machines (SVMs)) that are fairly robust with respect to outliers.
 - ❖ Or by applying others techniques (e.g., linear/logistic regression) that are more sensitive to them.



Outlier Detection and Treatment [11]

51

- Various schemes exist to deal with outliers.
- It highly depends on whether the outlier represents a valid or invalid observation.
- For invalid observations (e.g., age is 300 years), one could treat the outlier as a missing value using any of the schemes discussed in the previous slides.
- For valid observations (e.g., income is \$1 million), other schemes are needed.



Outlier Detection and Treatment [12]

52

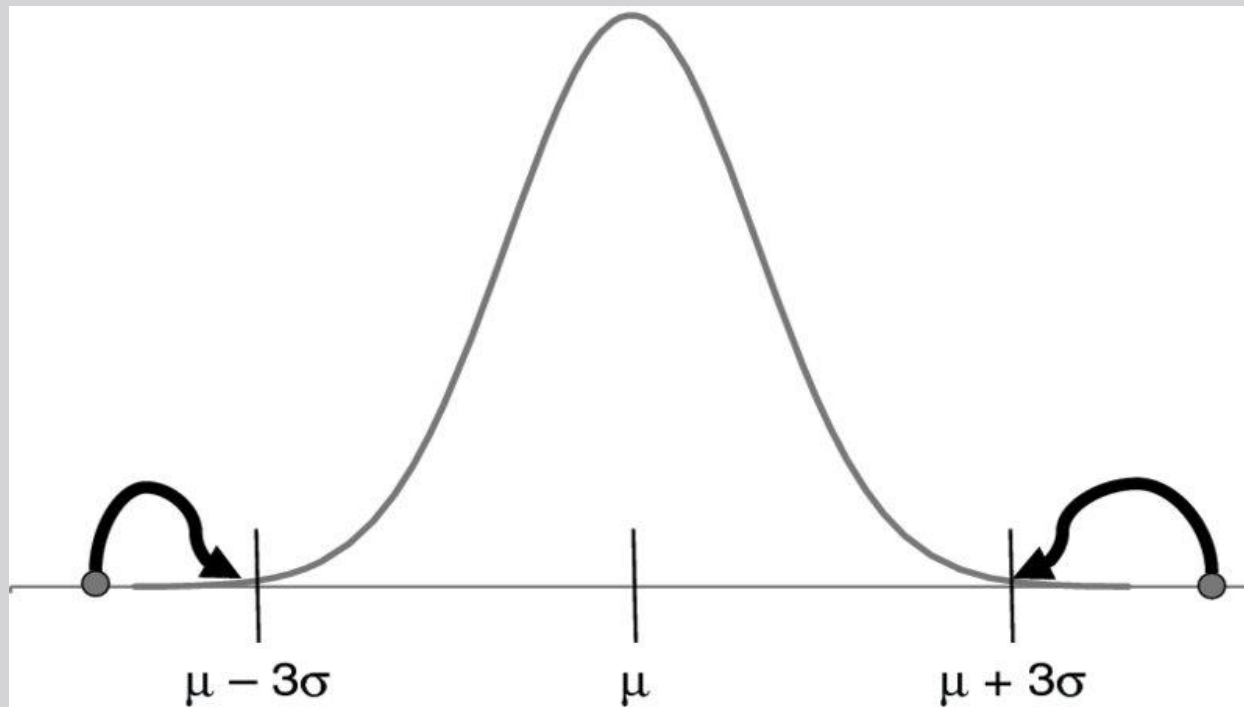
- A popular scheme is truncation/capping/winsorizing.
- It imposes both a lower and upper limit on a variable and any values below/above are brought back to these limits.
- The limits can be calculated using the z-score or the IQR (which is more robust than the z-scores).



Outlier Detection and Treatment [13]

53

Example on using z-score for truncation:



Upper/lower limit = $M \pm 3s$, with M = median and $s = \text{IQR}/(2 \times 0.6745)$.³



Outlier Detection and Treatment [14]

54

- In any cases, of course the expert-based limits based on business knowledge and/or experience can also be imposed.



8. Variable Selection

55



Variable Selection [1]

56

- Many analytical modeling exercises start with tons of variables, of which typically only a few actually contribute to the prediction of the target variable.
- For example, the average application/behavioural scorecard in credit scoring has somewhere between 10 and 15 variables.
- The key question is how to find these variables.



Variable Selection [2]

57


- Filters are a very handy variable selection mechanism.
- They work by measuring univariate correlations between each variable and the target.
- As such, they allow for a quick screening of which variables should be retained for further analysis.



Variable Selection [3]

58

Example of my own work:



الجامعة الإسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
بوتري باسا

INPUT SIGNIFICANCE ANALYSIS: FEATURE SELECTION THROUGH SYNAPTIC WEIGHTS MANIPULATION FOR EFUNNS CLASSIFIER

International Symposium on Computational Intelligence
& Applications (ISCIA2017), 14 to 15 July 2017

Dr. Raini Hassan ^{#1}
Prof. Dr. Imad Fakhri Taha Al-Shaikhli ^{#2}
Dept. of Computer Science, Faculty of ICT, Gombak

Assoc. Prof. Dr. Salmiah Ahmad ^{*}
** Dept. of Mechatronics Engineering, Faculty of Engineering, Gombak*

FS stands for Feature Selection, has the same role with Variable Selection: reducing the data size/dimension.

ISCIA2017, The Hatten Hotel, 15 July 2017 24

5. Findings (Cont'd.)

Group 2 Experiments: To test the FS on different dataset

Holdout Validation Method

Original Data (without FS)

Training time = 176.5
Testing time = 0.86
Created Node = 714
RMSE = 2.51
Error Rate = 47.76

GA-Ranked Data (with FR)

Training time = 198.37
Testing time = 1.28
Created Node = 695
RMSE = 2.35
Error Rate = 46.21

GA-Ranked Data (with FS)

Training time = 121.32
Testing time = 0.85
Created Node = 646
RMSE = 2.27
Error Rate = 44.98

Improved results

My own simple FS algorithm reduced the data size.

The “more than small-sized” data.

The algorithm that I used to rank all variables/features.



Variable Selection [4]

59

- Various filter measures have been suggested in the literature.
- One can categorize them as depicted in the following table:

	Continuous Target (e.g., CLV, LGD)	Categorical Target (e.g., churn, fraud, credit risk)
Continuous variable	1. Pearson correlation	2. Fisher score
Categorical variable	2. Fisher score/ANOVA	3. Information value (IV) 4. Cramer's V 5. Gain/entropy



Variable Selection [5]

60

1. Pearson correlation:

- The Pearson correlation ρ_P is calculated as follows:

$$\rho_P = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- It measures a linear dependency between two variables and always varies between -1 and +1.
- To apply it as a filter, one could select all variables for which the Pearson correlation is significantly different from 0 (according to the p -value), or, for example, the ones where $|\rho_P| > 0.50$.



Variable Selection [6]

61

2. Fisher score/Anova:

- The Fisher score can be calculated as follows:

$$\frac{|\bar{X}_G - \bar{X}_B|}{\sqrt{s_G^2 + s_B^2}}$$

- Where \bar{X}_G (\bar{X}_B) represents the average value of the variable for the Goods (Bads) and...
- s_G^2 (s_B^2) the corresponding variances.
- High values of the Fisher score indicate a predictive variable.
- To apply it as a filter, for example, keep the top 10 percent.
- Fisher score may generalize to a well-known analysis of variance (ANOVA) in case a variable that has multiple categories.



Variable Selection [7]

62

3. Information value (IV):

- It is based on WOE and is calculated as follows:

$$IV = \sum_{i=1}^k (Dist\ Good_i - Dist\ Bad_i) * WOE_i$$

- where k represents the number of categories of the variable.
- The following rules of thumb apply for the IV:
 - < 0.02 : unresponsive
 - $0.02-0.1$: weak responsive
 - $0.1-0.3$: medium responsive
 - > 0.3 : strong responsive



Variable Selection [8]

63

4. Cramer's V

- Followed from a chi-squared distribution with $k - 1$ degrees of freedom, with k being the number of classes of the characteristic, the Cramer's V measure can be calculated as follows:

$$\text{Cramer's } V = \sqrt{\frac{\chi^2}{n}} = 0.10,$$

- With n being the number of observations in the data set. Cramer's V is always bounded between 0 and 1 and higher values indicate better predictive power.
- As a rule of thumb, a cutoff of 0.1 is commonly adopted. One can then again select all variables where Cramer's V is bigger than 0.1, or consider the top 10 percent.



Variable Selection [9]

64

- Filters are very handy because they allow you to reduce the number of dimensions of the data set early in the analysis in a quick way.
- Their main drawback is that they work univariately and typically do not consider, for example, correlation between the dimensions individually.
- Hence, a follow-up input selection step during the modeling phase will be necessary to further refine the characteristics.



Variable Selection [10]

65

- Other criteria may play a role in selecting variables.
 - For example, from a regulatory compliance viewpoint, some variables may not be used in analytical models
 - (e.g., the U.S. Equal Credit Opportunities Act states that one cannot discriminate credit based on age, gender, marital status, ethnic origin, religion, and so on, so these variables should be left out of the analysis as soon as possible).
- Different regulations may apply in different geographical regions and hence should be checked.
- Also, operational issues could be considered (e.g., trend variables could be very predictive but may require too much time to be computed in a real-time online scoring environment).



9. Standardizing Data

66



Standardizing Data [1]

67

- Standardizing data is a data pre-processing activity targeted at scaling variables to a similar range.
- Consider, for example, two variables:
 - gender (*coded as 0/1*)
 - income (*ranging between \$0 and \$1 million*)
- When building logistic regression models using both information elements, the coefficient for income might become very small.
- Hence, it could make sense to bring them back to a similar scale.



Standardizing Data [2]

68

- The following standardization procedures could be adopted:
 1. Min/max standardization
 2. z-score standardization
 3. Decimal scaling
- Again note that standardization is especially useful for regression-based approaches...
 - but is not needed for decision trees, for example.



Standardizing Data [3]

69

1. 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).



Standardizing Data [4]

70

1. Min/max standardization:

#1

Input	Connection Weights Products		Sum of Hidden A & Hidden B	Normalized Importance
	Hidden A	Hidden B	(Importance)	
1	-0.00393	-0.14105	-0.144982423	0
2	0.003751	-0.07686	-0.073109952	0.239225809
3	-0.00477	0.09565	0.090880787	0.785065081
4	-0.00349	0.115376	0.111884541	0.854975576
5	-0.00393	-0.04478	-0.048710869	0.320437576
6	0.000907	0.025872	0.026779196	0.571704461
7	0.001628	0.04496	0.046588235	0.637638376
8	-0.00065	-0.10942	-0.110062086	0.116231511
9	0.002447	-0.07843	-0.075979804	0.229673575
10	0.001461	0.108544	0.110005321	0.848720637

#2

121	0.00147	0.07402	0.075490124	0.733837625
122	-0.00192	-0.12465	-0.126569239	0.061287845
123	0.003166	-0.03354	-0.030369898	0.381485062
124	0.002674	-0.09738	-0.094710746	0.16732808
125	-0.00379	0.123058	0.119263985	0.87953788
126	0.003655	0.151801	0.155455357	1
127	0.003537	-0.14066	-0.137125485	0.026151631
128	0.000434	0.15114	0.15157343	0.987079098

Sources for #1 & #2: Raini Hassan, Appendix B, PhD Computer Science thesis, 2016, IIUM



Standardizing Data [5]

71

2. z-score standardization:

Calculate the z-score

(see 7. Outlier Detection and Treatment section).



Standardizing Data [6]

72

3. 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.



10. Categorization

73



Categorization [1]

74

- Categorization (*also known as coarse classification, classing, grouping, binning, etc.*) can be done for various reasons.
- Generally:
 - Dimension reduction
 - Simplifies the analysis



Categorization [2]

75

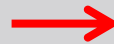
- For categorical variables, it is needed to reduce the number of categories.
- Consider, for example, the variable “purpose of loan” having 50 different values.
- When this variable would be put into a regression model, 49 dummy variables are needed ($50 - 1$ because of the collinearity), which would necessitate the estimation of 49 parameters for only one variable.
- With categorization, categories of values can be created with fewer parameters will have to be estimated and a more robust model is obtained.



Categorization [3]

76

Cat. No.	Purpose of Loan
1.	To buy a new house
2.	To buy a new car
3.	To enter college
4.	Always dreamed of a country house
5.	To pay for holidays
6.	To buy new furniture
7.	College's fees
8.	Buying dream car
9.	Travel to exotic places
10.	Need a second car



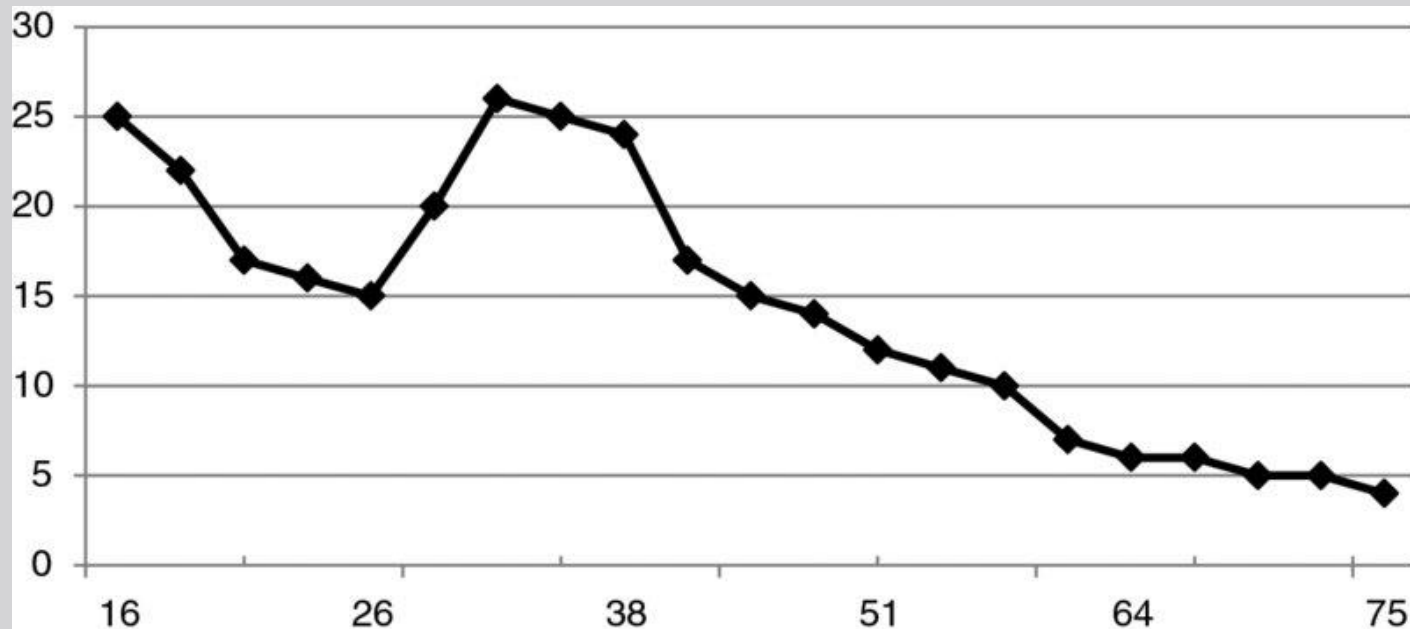
Cat. No.	Purpose of Loan
1.	To buy a new house
2.	To buy a new car
3.	To enter college
4.	Always dreamed of a country house
5.	To pay for holidays
6.	To buy new furniture
7.	College's fees
8.	Buying dream car
9.	Travel to exotic places
10.	Need a second car



Categorization [4]

77

- For continuous variables, categorization may also be very beneficial.
- Consider, for example, the age variable and its risk.





Categorization [5]

78

- Clearly, there is a non-monotonous (2-ways) relation between risk and age.
- If a nonlinear (*one way/one line*) model (e.g., *neural network, support vector machine*) were to be used, then the nonlinearity can be perfectly modeled.



Categorization [6]

79

- However, if a regression model were to be used (*which is typically more common because of its interpretability*), then since it can only fit a line, it will miss out on the non-monotonicity.
- By categorizing the variable into ranges, part of the non-monotonicity can be taken into account in the regression.
- Hence, categorization of continuous variables can be useful to model nonlinear effects into linear models.



Categorization [7]

80

- Various methods can be used to do categorization.
- Two very basic methods are:
 1. Equal interval binning
 2. Equal frequency binning.
- However, both methods are quite basic and do not take into account a target variable (e.g., fraud, credit risk).



Categorization [8]

81

1. Equal interval binning:

- Consider, for example, the income values 1,000, 1,200, 1,300, 2,000, 1,800, and 1,400.
- Equal interval binning would create two bins with the same range – Bin:
 - Bin 1: 1,000, 1,500
 - Bin 2: 1,500, 2,000.



Categorization [9]

82

2. Equal frequency binning:

- Consider, for example, the income values 1,000, 1,200, 1,300, 2,000, 1,800, and 1,400.
- Equal frequency binning would create two bins with the same number of observations:
 - Bin 1: 1,000, 1,200, 1,300
 - Bin 2: 1,400, 1,800, 2,000.



Categorization [10]

83

Chi-squared analysis:

- It is the more sophisticated way to do coarse classification.
- Consider the example depicted in the following table for coarse classifying a residential status variable.

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1

* L. C. Thomas, D. Edelman, and J. N. Crook, *Credit Scoring and its Applications* (Society for Industrial and Applied Mathematics, Philadelphia, Penn., 2002).



Categorization [11]

84

Chi-squared analysis:

- Suppose we want three categories and consider the following options:
 - Option 1: owner, renters, others
 - Option 2: owner, with parents, others
- Both options can now be investigated using chi-squared analysis.
- The purpose is to compare the empirically observed with the independence frequencies.



Categorization [12]

85

Chi-squared analysis:

- For option 1: owner, renters, others:

Original Data

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1



Empirically Observed Data

Attribute	Owner	Renters	Others	Total
Goods	6,000	1,950	1,050	9,000
Bads	300	540	160	1,000
Total	6,300	2,490	1,210	10,000



Categorization [13]

86

Chi-squared analysis:

- For option 2: owner, with parents, others:

Original Data

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1



Empirically Observed Data

Attribute	Owner	With Parents	Others	Total
Goods	6,000	950	2050	9,000
Bads	300	100	600	1,000
Total	6,300	1050	2650	10,000



Categorization [14]

87

Chi-squared analysis:

- For option 1 - the independence frequencies:
 - The number of good owners, given that the odds are the same as in the whole population, is $6,300/10,000 \times 9,000/10,000 \times 10,000 = 5,670$.

Empirically Observed Data

Attribute	Owner	Renters	Others	Total
Goods	6,000	1,950	1,050	9,000
Bads	300	540	160	1,000
Total	6,300	2,490	1,210	10,000



Independence Frequencies Data

Attribute	Owner	Renters	Others	Total
Goods	5,670	2,241	1,089	9,000
Bads	630	249	121	1,000
Total	6,300	2,490	1,210	10,000



Categorization [15]

88

Chi-squared analysis:

- For option 2 - the independence frequencies:
 - The number of good owners, given that the odds are the same as in the whole population, is $6,300/10,000 \times 9,000/10,000 \times 10,000 = 5,670$.

Empirically Observed Data

Attribute	Owner	With Parents	Others	Total
Goods	6,000	950	2050	9,000
Bads	300	100	600	1,000
Total	6,300	1050	2650	10,000



Independence Frequencies Data

Attribute	Owner	With Parents	Others	Total
Goods	5670	945	2385	9,000
Bads	630	105	265	1,000
Total	6,300	1050	2650	10,000



Categorization [16]

89

Chi-squared analysis:

- For option 1 – chi-squared distance:

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

- For option 2 – chi-squared distance:

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



Categorization [17]

90

Chi-squared analysis:

- So, based upon the chi-squared values, **option 2 is the better categorization.**
- Note that formally, one needs to compare the value with a chi-squared distribution with $k - 1$ degrees of freedom with k being the number of values of the characteristic.
- Many analytics software tools have built-in facilities to do categorization using chi-squared analysis.
- A very handy and simple approach (*available in Microsoft Excel*) is pivot tables.



11. Weights of Evidence Coding

91



Weights of Evidence Coding [1]

92

- Categorization reduces the number of categories for categorical variables.
- For continuous variables, categorization will introduce new variables.



Weights of Evidence Coding [2]

93

- Consider a regression model with the following characteristics:
 - ❖ age (4 categories, so 3 parameters)
 - ❖ purpose (5 categories, so 4 parameters)
- The model then looks as follows:

$$Y = \beta_0 + \beta_1 \text{Age}_1 + \beta_2 \text{Age}_2 + \beta_3 \text{Age}_3 + \beta_4 \text{Purp}_1 + \beta_5 \text{Purp}_2 + \beta_6 \text{Purp}_3 + \beta_7 \text{Purp}_4$$



Weights of Evidence Coding [3]

94

- Despite having only two characteristics, the model still needs 8 parameters to be estimated.
- It would be handy to have a monotonic transformation $f(\cdot)$ such that our model could be rewritten as follows:

$$Y = \beta_0 + \beta_1 f(\text{Age}_1, \text{Age}_2, \text{Age}_3) + \beta_2 f(\text{Purp}_1, \text{Purp}_2, \text{Purp}_3, \text{Purp}_4)$$

- The transformation should have a monotonically increasing or decreasing relationship with Y .



Weights of Evidence Coding [4]

95

- Weights-of-evidence (WOE) coding is one example of a transformation that can be used for this purpose.

Age	Count	Distr. Count	Goods	Distr. Good	Bads	Distr. Bad	WOE
Missing	50	2.50%	42	2.33%	8	4.12%	-57.28%
18-22	200	10.00%	152	8.42%	48	24.74%	-107.83%
23-26	300	15.00%	246	13.62%	54	27.84%	-71.47%
27-29	450	22.50%	405	22.43%	45	23.20%	-3.38%
30-35	500	25.00%	475	26.30%	25	12.89%	71.34%
35-44	350	17.50%	339	18.77%	11	5.67%	119.71%
44+	150	7.50%	147	8.14%	3	1.55%	166.08%
	2,000		1,806		194		



Weights of Evidence Coding [5]

96

- The WOE is calculated as:
 - ❖ $\ln(\text{Distr. Good} / \text{Distr. Bad})$.
- Because of the logarithmic transformation, a positive (negative) WOE means $\text{Distr. Good} > (<) \text{Distr. Bad}$.
- The WOE transformation thus implements a transformation monotonically related to the target variable.



Weights of Evidence Coding [6]

97

- The model can then be reformulated as follows:

$$Y = \beta_0 + \beta_1 \text{WOE}_{\text{age}} + \beta_2 \text{WOE}_{\text{purpose}}$$

- This gives a more concise model than the model with which we started in this section.



12. Segmentation

98



Segmentation [1]

99

- Sometimes the data is segmented before the analytical modeling starts, due to the following reasons:
 - Strategic motivation
(e.g., banks might want to adopt special strategies to specific segments of customers).
 - Operational viewpoint motivation
(e.g., new customers must have separate models because the characteristics in the standard model do not make sense operationally for them).



Segmentation [2]

100

- It could also be needed to take into account significant variable interactions
 - *(e.g., if one variable strongly interacts with a number of others, it might be sensible to segment according to this variable).*
- The segmentation can be conducted using:
 - the experience and knowledge from a business expert
 - Or it could be based on statistical analysis using, for example, decision trees, k-means, or self-organizing maps.



Segmentation [3]

101

- It is a very useful pre-processing activity because one can now estimate different analytical models each tailored to a specific segment.
- However, one needs to be careful with it because by segmenting, the number of analytical models to estimate will increase, which will obviously also increase the production, monitoring, and maintenance costs.



Dua After Studying

102

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Dua After Studying

Oh Allah! I entrust you with what I have read and I have studied. Oh Allah! Bring it back to me when I am in need of it. Oh Allah! You do whatever you wish, you are my availer and protector and the best of aid.

اللَّهُمَّ إِنِّي أَسْتَوْدِعُكَ مَا قَرَأْتُ وَمَا حَفَظْتُ، فَرُضْهُ عَلَيَّ عِنْدَ حَاجَتِي إِلَيْهِ، إِنَّكَ عَلَى مَا تَشَاءُ قَدِيرٌ وَأَنْتَ حَسْبِي وَنِعْمَ الْوَكِيلُ

Allahumma inni astaodeeka ma qara'tu wama hafaz-tu. Farudduhu 'allaya inda hajati elahi. Innaka 'ala ma-tasha'-u qadeer wa anta hasbeeya wa na'mal wakeel.