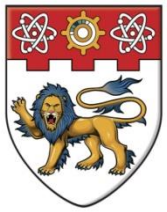


NANYANG TECHNOLOGICAL UNIVERSITY

BC2406 Business Analytics I: Predictive Techniques

Seminars 4 and 5
Basic Knowledge about Data

Instructor: Prof. Lee Gun-woong

Nanyang Business School

Important things about data

- **What kinds of data can be used in analytics?**
- Where can we obtain these data?
- Data Structures and Data Types
- How much data is needed?
- Data Treatment

Potentially Useful Data

- Point-of-sale data (including coupons proffered, discounts applied, …)
- Credit card charge records
- Medical insurance claims data
- Web log data
- E-commerce server application logs
- Direct mail response records
- Call-centre records, including memos written by the call-center reps
- Survey response data

Types of Data

- Demographic data
- Transaction data
- Click Stream data
- Network Data

Important things about data

- What kinds of data can be used in analytics?
- **Where can we obtain the data?**
- Data Structures and Data Types
- How much data is needed?
- Data Treatment

Data Sources

- Log/cookie information
- Survey/Interview
- Customer service/phone call
- Scanner data
- Documents

Example: Target

The desire to collect information on customers is not new for Target or any other large retailer, of course. For decades, Target has collected vast amounts of data on every person who regularly walks into one of its stores. Whenever possible, Target assigns each shopper a unique code — known internally as the Guest ID number — that keeps tabs on everything they buy. "If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we've sent you or visit our Web site, we'll record it and link it to your Guest ID," Pole said. "We want to know everything we can."

Example: Target

Also linked to your Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you've moved recently, what credit cards you carry in your wallet and what Web sites you visit. Target can buy data about your ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own. (In a

Important things about data

- What kinds of data can be used in analytics?
- Where can we obtain these data?
- **Data Structures and Data Types**
- How much data is needed?
- Data Treatment

Data Structure

- Record data
- Transaction data

Record Data

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

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

Transaction Data

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

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

Organizing data for analytics

- All data mining algorithms want their inputs in tabular format
  - All data should be in a single table.
  - Each row should correspond to an entity, such as a customer, that is relevant to the business.
    - is also known as record, point, case, sample, entity, or instance
  - Each column stands for each attribute
    - An attribute is a property or characteristic of an entity
    - A collection of attributes describe an object
    - Is also known as variables.
  - Relational database

Sales Ranks on Mobile Apps (19 July, 2013)

Variables

Records (Obs.)

AppId	Category	Rank	Name	Price	Seller	Screenshots	Rating_Score	Rating_Volumn	Release_Date	Data_Date
342548956	Business	1	TurboScan: quickly scan multipage documents into high-quality	1.99	Piksoft Inc.	5	4.5	11302	7-Dec-09	19-Jul-13
294934058	Business	2	HotSchedules	2.99	HotSchedules	4	3.5	2392	31-Oct-08	19-Jul-13
428974099	Business	3	Mail+ for Outlook	5.99	iKonic Apps LLC	5	3.5	5124	18-Apr-11	19-Jul-13
347803339	Business	4	CamCard - Business card scanner & Business card reader & sca	2.99	IntSig Information Co.,Ltd	5	4.5	2623	29-Dec-09	19-Jul-13
307868751	Business	5	JotNot Scanner Pro: scan multipage documents to PDF	0.99	MobiTech 3000 LLC	5	4	7039	27-Mar-09	19-Jul-13
577499909	Business	6	TapeACall Pro - Record Calls	9.99	Epic Enterprises LLC	5	4	148	22-Jan-13	19-Jul-13
437818260	Business	7	SayHi Translate	0.99	SayHI, LLC	4	4.5	6961	26-May-11	19-Jul-13
561712083	Business	8	Boxer - Your Inbox for Outlook, Gmail, Exchange, Hotmail, iClou	3.99	Bodkin Software Inc.	5	4	398	26-Sep-12	19-Jul-13
539943615	Business	9	Voice Translate Pro	0.99	Intellectual Flame Co., Ltd	4	4.5	690	1-Aug-12	19-Jul-13
333710667	Business	10	Scanner Pro by Readdle	6.99	Igor Zhadanov	5	4.5	6355	9-Oct-09	19-Jul-13
333211045	Business	11	WorldCard Mobile - business card reader & business card scan	6.99	Penpower Inc.	5	4.5	3104	3-Nov-09	19-Jul-13
335047649	Business	12	ScanBizCards Business Card Reader	4.99	ScanBiz Mobile Solutions L	5	4	1921	28-Oct-09	19-Jul-13
401818935	Business	13	Genius Scan+ - PDF Scanner	2.99	The Grizzly Labs	5	4.5	462	16-Dec-10	19-Jul-13
468081771	Business	14	Secret Folder Pro: Secure Photo Gallery & Wifi Transfer App	2.99	chen kaiqian	5	4.5	1284	13-Oct-11	19-Jul-13
561386772	Business	15	Splashtop Personal - Remote Desktop for iPhone & iPod	2.99	Splashtop Inc.	5	4.5	1615	16-Oct-12	19-Jul-13
556500145	Business	16	TinyScan Pro - PDF scanner to scan multipage documents	4.99	Blue Tags	5	4.5	1582	18-Oct-12	19-Jul-13
317107309	Business	17	Documents To GoÃ‚Â® Premium - Office Suite	17	DataViz, Inc	5	3	5087	14-Jun-09	19-Jul-13
570779598	Business	18	NADA MarketValues	1.99	N A D A SERVICES CORPOF	5	0	0	13-Dec-12	19-Jul-13
373045717	Business	19	Voice Recorder HD for Audio Recording, Playback, Trimming ar	1.99	eFUSION	5	4	507	28-May-10	19-Jul-13
323133888	Business	20	PDF Expert (professional PDF documents reader)	9.99	Igor Zhadanov	5	4	1153	18-Jul-09	19-Jul-13
338550388	Business	21	Audio Memos - The Voice Recorder	0.99	Imesart S.a.r.l.	5	3.5	1143	25-Nov-09	19-Jul-13
597820271	Business	22	Color Effects FX HD - ReColor And Splash Photo Effect Editor Sh	1.99	Tao Lin	5	4	175	5-Feb-13	19-Jul-13
498174936	Business	23	Voice Commands.	2.99	Component Studios LLC	5	3	612	17-Mar-12	19-Jul-13
285877935	Business	24	QuickVoice2Text Email (PRO Recorder)	2.99	nFinity Inc	4	2.5	3187	14-Apr-09	19-Jul-13
598955472	Business	25	Photo Editor HD-Edit,Sticker,Rotate,Filter&Enhance Image Effe	0.99	Tao Lin	5	4	40	12-Feb-13	19-Jul-13
382013715	Business	26	SuperCam_Pro	1.99	Shenzhen TVT Digital Tech	2	2	1113	20-Jul-10	19-Jul-13
595865489	Business	27	Avocado Scanner Deluxe - Scan and Fax Documents, Receipts, I	2.99	Avocado Hills, Inc.	5	4	76	6-Mar-13	19-Jul-13

NANYANG TECHNOLOGICAL UNIVERSITY

Types of Variables

- Different Dimensions
  - Numerical & Text
  - Discrete & Continuous
- Numerical
  - Sales, Age, #days…
- Text
  - Comments, messages

Discrete and Continuous Attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as floating-point values (1.0, 2.34, 100.2)

- Discrete Attribute
  - Has only a finite or countable set of values
  - Examples: zip codes, counts…
  - Integer (-2 , 0 , 2) or Categorical values (Low, Mid, High)
  - Categorical Variable
    - Ordered (Low, Mid, High) vs. Unordered (Male, Female)

Variable Handling

- Numerical Variables
  - Most algorithms can handle numeric data
  - May occasionally need to “bin” into categories
    - Continuous Variable to Categorical Variable

Income	Category
- $20,000	Low (=1)
$20,001 - $50,000	Mid (=2)
$50,001 -	High (=3)

- Categorical Variables
  - Continuous Values to Ordered Values: such as 1,2,3
    - Treat them as continuous values
  - Categorical variable to Binary dummies
    - Number of dummies = Number of categories – 1

ID	Female	Male
C1	0	1
C2	1	0
C3	1	0

ID	Female
C1	0
C2	1
C3	1

Important things about data

- What kinds of data can be used in analytics?
- Where can we obtain these data?
- Data Structures and Data Types
- **How much data is needed?**
- Data Treatment

How many variables

- Sometime, analysts may throw out variables that seem unlikely to be interesting, keeping only a few carefully chosen variables they expect to be important.
- The data mining approach calls for letting the data itself reveal what is and is not important.
- At the beginning, we should use many variables, and let the model decide

How many data observations?

- Depends on
  - the particular algorithms employed
  - the complexity of the data
  - the relative frequency of possible outcomes
  - # of variables used
- When data *is* scarce, data mining is less effective and it is less likely to be useful.
  - 1 outcome with 1 variable (15 observations)
  - 1 outcome with 15 variables (1 observation)

ID	Purchase?	Gender
C1	*NA*	M
C2	*NA*	F
C3	Yes	M
…	…	…
C14	*NA*	F
C15	*NA*	M

ID	Purchase?	Gender	Age	…	VAR15
C3	Yes	M	34	…	184

Size and Density

- Density refers to the dominance of certain outcome values in the target variable (5% Yes, 95%No).
- Often the target variable represents something relatively rare.
  - It is *rare* for prospects to respond to a direct mail offer (2% responses)
- It is desirable for the model set to be balanced with equal numbers of each of the outcomes (50% Yes, 50% No) during the model-building process
- A smaller, *balanced sample* is preferable to a larger one with a very low proportion of rare outcomes

Size and Time

- When the model set is large enough to build good, stable models, making it larger is counterproductive because everything will take longer to run on the larger dataset.

Advice 1

- A simple test for whether the sample used for modelling is large enough is to try doubling it and measure the improvement in the model's accuracy.
- If the model created using the larger sample is significantly better than the one created using the smaller sample, then the smaller sample is not big enough.
- If there is no improvement, or only a slight improvement, then the original sample is probably adequate.

Advice 2

- A good rule of thumb:
  - Have 10 *complete records* for each predictive variable
  - At least 6*m*n
    - m: number of outcome classes
    - n: number of variables
    - by Delmaster and Hancock (2001, p.68)

ID	Purchase?	Gender	Age	…	VAR15
C3	Yes	M	34	…	184

    - Number of outcome classes: 2 (Yes or No)
    - Number of variables: 15
    - Minimum number of observations: 6*2*15 = 180

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

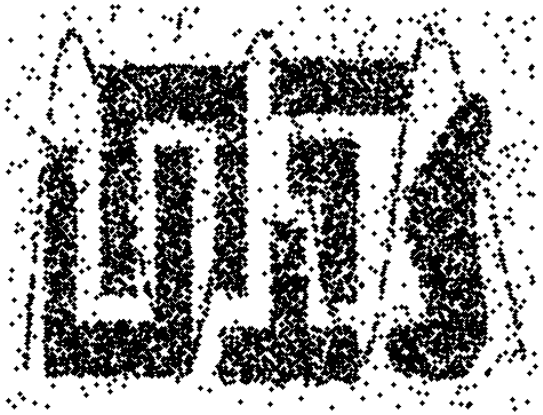
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time-consuming.

- Sampling is sometimes used in data mining because processing the entire set of data of interest is too expensive or time-consuming.

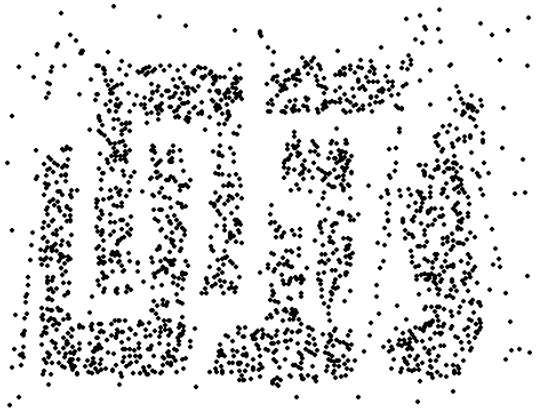
Sampling …

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is *representative*
  - A sample is representative if it has approximately the same property (of interest) as the original set of data
    - Random Sampling

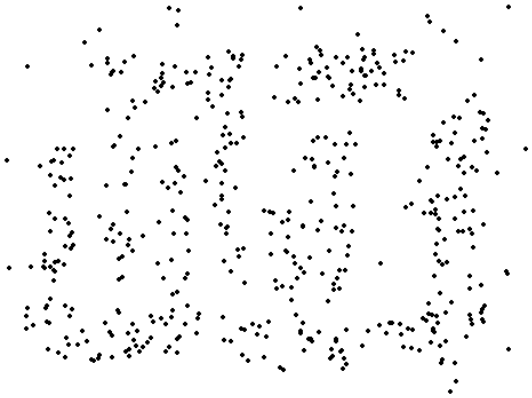
Sample Size



8000 points



2000 Points



500 Points

Rare Event Oversampling

- A sample of 100 people (70 Men, 30 Women)
- Need to correct for a bias in the sample
- Often the event of interest is rare
- Examples: response to a mailing, fraud in taxes, …
- Sampling may yield too few “interesting” cases to effectively train a model
- A popular solution: oversample the rare cases to obtain a more balanced training set
- Later, need to adjust results for the oversampling

Important things about data

- What kinds of data can be used in analytics?
- Where can we obtain these data?
- Data Structures and Data Types
- How much data is needed?
- **Data Treatment**

Data Treatment

- Exploring Data
  - Statistic summary
  - Correlation
  - Visualization
- Data Cleaning
- Data Preparation

Statistical summary

- common metrics
  - Mean
  - Median
  - Minimum
  - Maximum
  - Standard deviation
  - Counts & percentages

Measures of Location: Mean and Median

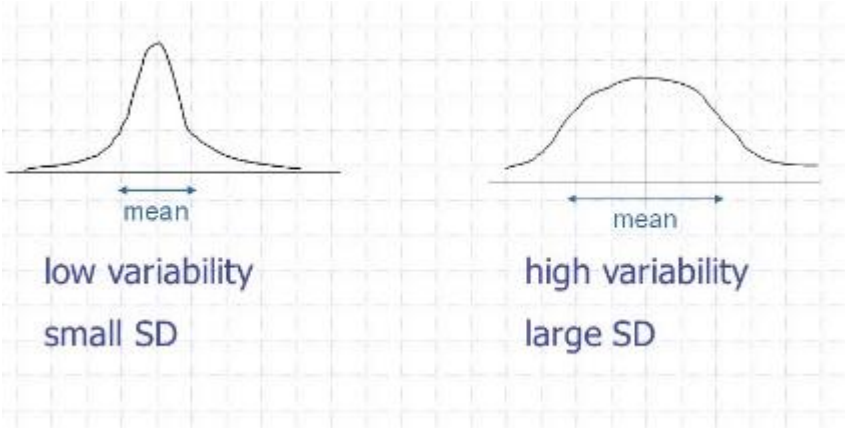
- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the **max** and **min**
- The variance or standard deviation is the most common measure of the spread of a set of points.



$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

Frequency

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.

- The notions of frequency is typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the *p*th percentile is a value x_p of x such that *p*% of the observed values of x are less than x_p.

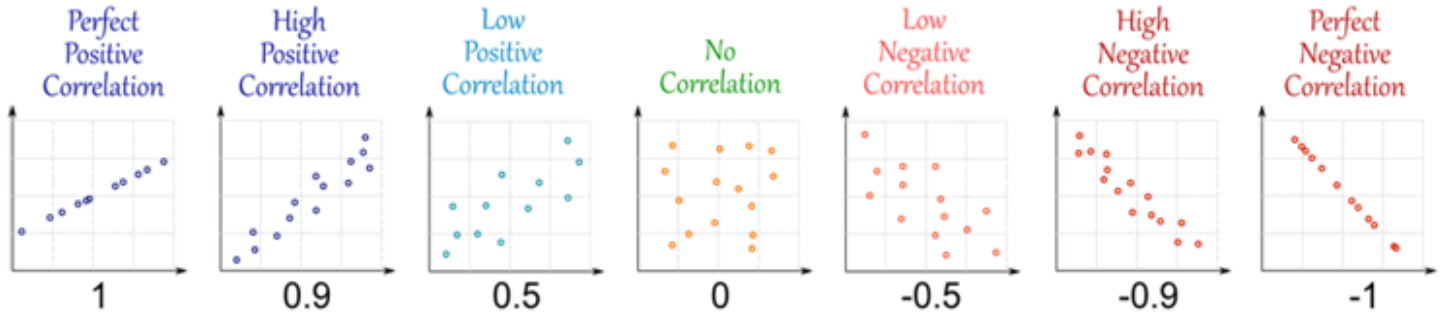
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

Summary Statistics – Mobile Apps

Variable	Obs	Mean	Std. Dev.	Min	Max
Rank	1,200	150.5	86.63817	1	300
Price	1,200	3.1575	4.071436	.99	89.99
Screenshots	1,200	4.655	.844533	0	5
Rating_Score	1,200	3.785	1.047667	0	5
Rating_Vol~n	1,200	8196.536	45469.04	0	823547

Correlation Analysis

- In statistics, dependence is any statistical relationship between two random variables or two sets of data.
- Correlation refers to any of a broad class of statistical relationships involving dependence.
- The most familiar measure of dependence between two quantities is *the Pearson's correlation coefficient* (the correlation coefficient).
- It is obtained by dividing the covariance of the two variables by the product of their standard deviations .



Source: www.mathisfun.com, Sep 5, 2016

NANYANG TECHNOLOGICAL UNIVERSITY

Correlation Analysis

- Below: Correlation matrix for Mobile Apps Data
- Shows correlation between variable pairs

	Rank	Price	Screen~s	Rating~e	Rating~n
Rank	1.0000				
Price	0.0115	1.0000			
Screenshots	-0.1275	-0.0009	1.0000		
Rating_Score	-0.1996	-0.0856	0.1932	1.0000	
Rating_Vol~n	-0.1392	-0.0515	0.0363	0.1064	1.0000

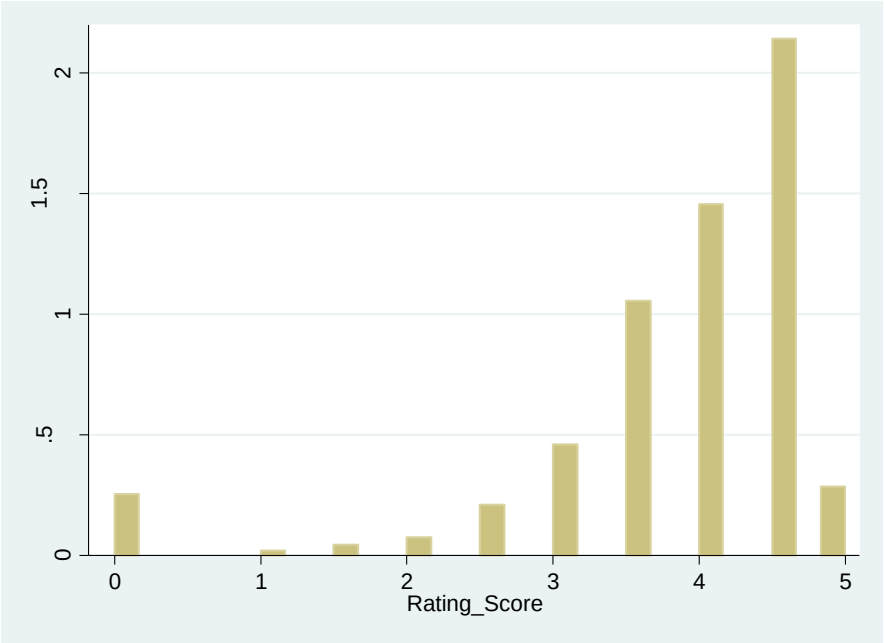
Visualization Techniques

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins

Histogram

- Example: Frequencies of Rating Scores

Histogram shows the distribution of Rating Scores (0 – 5) on Apps



NANYANG TECHNOLOGICAL UNIVERSITY

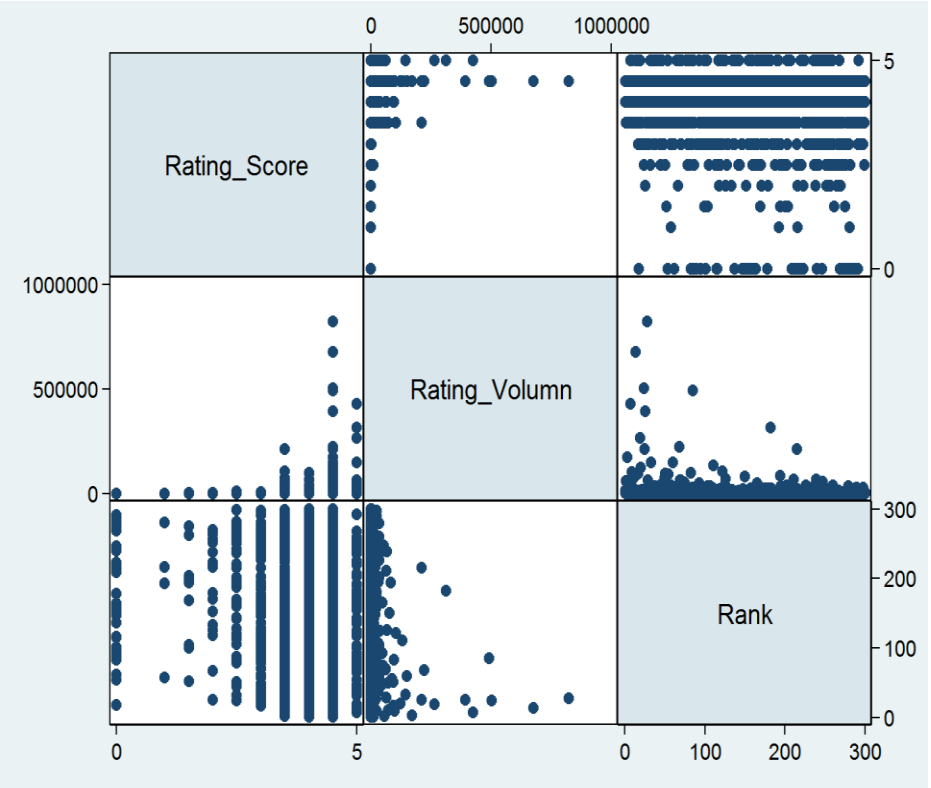
Matrix Plot

- Variable selection
- A straight line would be an indicator that one variable is exactly correlated with another
- Other kinds of relationships?
- Weed out irrelevant and redundant variables

Matrix Plot

Shows scatterplots for variable pairs

Example: scatterplots for the three mobile App variables



Data Quality and Data Cleaning

- Data quality problems
  - Duplicate data
  - Outliers
  - Missing values
- What can we do about these problems?

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning method
  - Keep the newest one
  - Merge some
  - Delete

Missing Values

- **Reasons for missing values**
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- **Special case of "0"**
  - The value is missing (*NA*)
  - The value is actually zero
  - "0" in the credit card past due
    - Situation 1: paid up (Zero remaining balance)
    - Situation 2: with no credit history (*NA*)
  - Human judgement

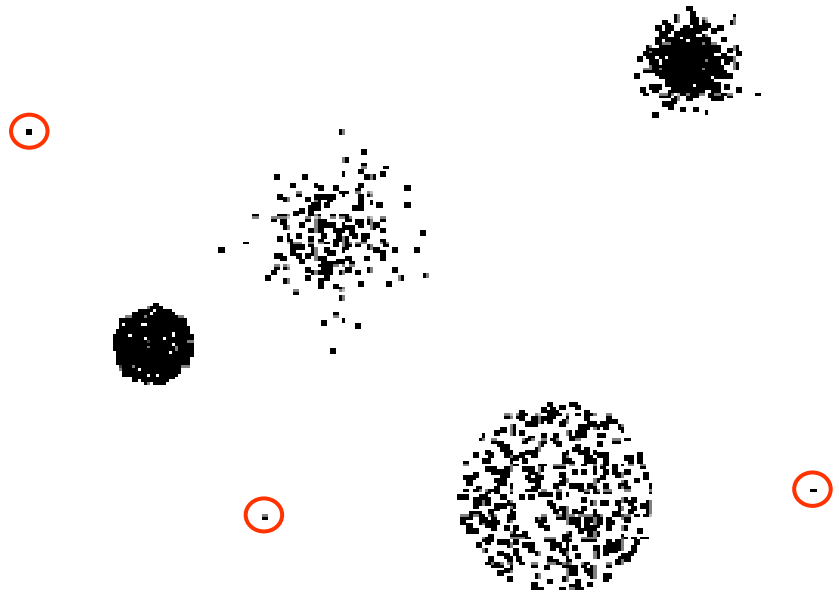
- **Most algorithms will not process records with missing values.**

Handling Missing Data

- **Eliminate Data Records (omissions)**
  - If a small number of records have missing values, can omit them
- **Imputation**
  - 1 variable with many missing values, and 29 variables without missing values
  - Replace missing values with reasonable substitutes (mean)
  - Lets you keep the record and use the rest of its (non-missing) information
- **Eliminate Data Variables**
  - If many records are missing values on a small set of variables, can drop those variables If many records have missing values, omission is not practical

Outliers 1

- Outliers are data objects with characteristics that are considerably different than most of the other data obje



Outliers 2

- An outlier is an observation that is “extreme”, being distant from the rest of the data (definition of “distant” is deliberately vague)
- Rules of thumb
  - Anything over 3 standard deviations from the mean is an outliner
    - Extreme_Low < Mean – 3*SD
    - Extreme_High < Mean – 3*SD
- Outliers can have disproportionate influence on models
- An important step in data pre-processing is detecting outliers

Data Cleaning for Outlier

- Once detected, domain knowledge is required to determine if it is *an error*, or truly *extreme*.
  - Statistical procedures can do little beyond identifying the outliers
  - Manual review is needed
    - $ 0 for Paid Apps
    - $ 100 for Paid Apps
  - Temperature of 60 in Singapore
  - If the number of records with outliers is very small, treat them as missing data.

Detecting Outliers

- In some contexts, finding outliers is the purpose of the DM exercise (airport security screening). This is called “anomaly detection”.

Special Cases of Distributions

- ***Columns with One Value***
- ***Columns with Almost Only One Value***
- ***Columns with Unique Values***

Data Preprocessing

- Normalization
- Attribute Transformation
- Dimensionality Reduction
  - Principle Component Analysis
  - Factor Analysis

Normalizing (Standardizing) Data

- Used in some techniques when variables with the largest scales would dominate and skew results
- Normalizing function: Subtract mean and divide by standard deviation

$$\frac{x - Mean(X)}{SD(X)}$$

- Alternative function: scale to 0-1 by subtracting minimum and dividing by the range

$$\frac{x - Min(X)}{Max(X) - Min(X)}$$

Attribute Transformation

- Continues Variable to Binary Variable
  - Churn/ stay (1 or 0)
- Continuous Variable to Categorical Variable
  - *Do higher incomes (from Low to High) influence the sales of a product A?*
- Categorical variables to binary dummies
  - *Do low-income customers ('Low') influence the sales of a product A?*

ID	Income
C1	$10,000
C2	$40,000
C3	$70,000
C4	$30,000
C5	$90,000

ID	Income
C1	1
C2	2
C3	3
C4	2
C5	3

ID	Low	Mid	High
C1	1	0	0
C2	0	1	0
C3	0	0	1
C4	0	1	0
C5	0	0	1

Income	Category
- $20,000	Low (=1)
$20,001 - $50,000	Mid (=2)
$50,001 -	High (=3)

NANYANG TECHNOLOGICAL UNIVERSITY

Reducing Categories

- A single categorical variable with *m* categories is typically transformed into *m-1* dummy variables
- Each dummy variable takes the values 0 or 1

  0 = “no” for the category

  1 = “yes”
- Problem: Can end up with too many variables
- Solution: Reduce by combining categories that are close to each other

ID	Income
C1	Low
C2	Mid
C3	High
C4	Mid
C5	High

ID	Low	Mid	High
C1	1	0	0
C2	0	1	0
C3	0	0	1
C4	0	1	0
C5	0	0	1

ID	Mid	High
C1	0	0
C2	1	0
C3	0	1
C4	1	0
C5	0	1

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.

ID	Gender	Height	Weight
C1	1	170	64
C2	0	NA	45
C3	NA	165	45
C4	0	155	NA
C5	1	187	70
C6	NA	190	NA
C7	1	NA	80

- Definitions of distance between points, which is critical for clustering and outlier detection, become less meaningful
- Reduce dimension: principle component analysis