2. Data exploration

DSCI 5240 Data Mining and Machine Learning in Business

1

Today's agenda

01

Understand data – where they come from

02

Understand data – measurement levels

03

Explore data using Python

2

Customer data - Framework

- Demographics: age, gender, zip code, education, income…
- Psychographics: interests, lifecycle stage, attitudes, beliefs…
- Technographics: how long they are online every week, data usage, equipment, apps/tools
- Transactions: plan, how much they spend on the plan
- Consumption and usage: data usage, voice usage
- Interactions: complaints, number of times to call service, network issues; marketing promotions; events

3

Sample customer data of a bank

ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Accoun	CD Account	Online	CreditCard
1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
3	39	15	11	94720	1	1	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
5	35	8	45	91330	4	1	2	0	0	0	0	0	1
6	37	13	29	92121	4	0.4	2	155	0	0	0	1	0
7	53	27	72	91711	2	1.5	2	0	0	0	0	1	0
8	50	24	22	93943	1	0.3	3	0	0	0	0	0	1
9	35	10	81	90089	3	0.6	2	104	0	0	0	1	0
10	34	9	180	93023	1	8.9	3	0	1	0	0	0	0
11	65	39	105	94710	4	2.4	3	0	0	0	0	0	0
12	29	5	45	90277	3	0.1	2	0	0	0	0	1	0
13	48	23	114	93106	2	3.8	3	0	0	1	0	0	0
14	59	32	40	94920	4	2.5	2	0	0	0	0	1	0
15	67	41	112	91741	1	2	1	0	0	1	0	0	0
16	60	30	22	95054	1	1.5	3	0	0	0	0	1	1
17	38	14	130	95010	4	4.7	3	134	1	0	0	0	0
18	42	18	81	94305	4	2.4	1	0	0	0	0	0	0
19	46	21	193	91604	2	8.1	3	0	1	0	0	0	0
20	55	28	21	94720	1	0.5	2	0	0	1	0	0	1
21	56	31	25	94015	4	0.9	2	111	0	0	0	1	0
22	57	27	63	90095	3	2	3	0	0	0	0	1	0
23	29	5	62	90277	1	1.2	1	260	0	0	0	1	0
24	44	18	43	91320	2	0.7	1	163	0	1	0	0	0
25	36	11	152	95521	2	3.9	1	159	0	0	0	0	1

4

Where do they come from?

- Internal
  - Customer relationship management systems: Relational databases
  - Transaction systems: relational databases
  - Surveys: e.g., customer satisfaction survey – data in files
  - Server logs: customer usage data
  - Voice files/NoSQL: e.g., customer calls to service centers
- External
  - Purchased from data providers
  - Social media: web scraping

5

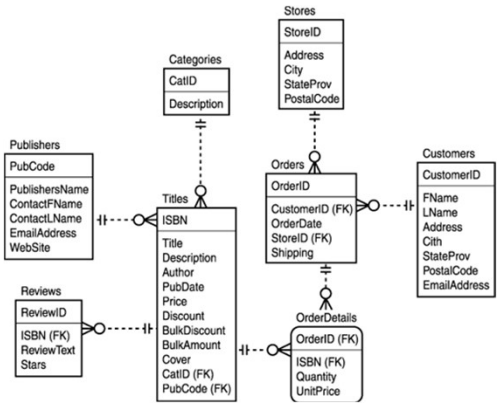
Data storage

- Files
  - Text files: .csv (comma separated values)
  - Log files: machine data
  - Web files: HTML, XML, JSON
  - Application-specific files: Word, PowerPoint, Excel
- Databases: organized collection of data
  - Relational databases
  - Other types of databases: object-oriented, key-value databases, etc.

6

A relational Database

- The most common type of database system
- Information stored in two dimensional tables with defined set of relationships among them
- Efficient and intuitive



7

Measurement levels of variables

8

Measuring Objects

- Customers, products, companies, and any other "**object**" are described by their **attributes** (satisfaction, price, innovativeness…)
- **Attribute**, **dimension**, **feature**, and **variable** are often used interchangeably
- Attributes may vary from one object to another (cross-sectional) or from one time to another (longitudinal).
- To **measure** attributes, we assign numbers or symbols to them

9

Nominal Measure

- Symbols or names of things
- Valid operations: =, ≠
- No intrinsic ordering

Color of horses

	Battleship	Black Gold	Bushranger
Color	White	Black	Brown

10

Binary Measure

- Nominal attribute with only two categories or states
- 0 or 1
- Examples:
  - Smokers: yes/no
  - Medical test: positive/negative
  - Gender: male/female

Gender of the horse

	Battleship	Black Gold	Bushranger
Gender	M	M	F

11

Ordinal Measure

- Meaningful order or ranking
- But the distance between ordinal measures has no accurate meaning: the difference between 1st and 2nd ≠ the difference between 2nd and 3rd
- Valid operations: =, ≠, >, ≥ ,<, ≤
- You cannot perform arithmetic operations on them

Position results in race

	Battleship	Black Gold	Bushranger
Race 1	1st	2nd	3rd
Race 2	2nd	3rd	1st
Race 3	3rd	1st	2nd
Race 4			

12

Interval Measure

- Measured on a scale with meaningful difference
- We can perform meaningful arithmetic operations on them
  - In Race 1, Black Gold is 122.4/120 = 1.02 times slower than Battleship (or, 2% slower).
- Valid operations: =, ≠, >, ≥ ,<, ≤, +, -, x, /,…

Total time used

	Battleship	Black Gold	Bushranger
Race 1	120 s	122.4s	122.7s
Race 2	118.2s	121.3s	118s
Race 3	122.3s	122s	122.2s
Race 4	121.1s	121s	123.7s

13

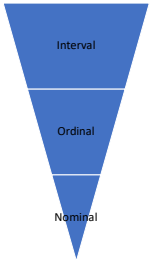
Levels of Measurement

- An attribute can be measured at a certain level or any lower level, e.g.,
  - horse-speed intrinsically interval-scaled
  - Can be measured as ordinal

So

- An **interval-scaled attribute** can be measured by **interval, ordinal or nominal** measures
- An **ordinal-scaled attribute** can be measured by **ordinal or nominal** measures, but NOT interval
- **Nominal attributes** can only have **nominal** measures

- Caution: Measuring at a lower level loses information and limits possible analyses



14

Variable types

Categorical variables

- With categorical variable
  - Binary
  - Ordinal
  - Nominal

Interval variables

- Different names for interval
  - Numeric variables
  - Continuous variables

15

What is the measurement scale?

- During which season of the year were you born?
  _____winter _____spring _____summer _____fall
- What is your total household income? _____
- Which are your three most preferred brands of beer? Rank them from 1 to 3 with 1 being most preferred.
  Tsing Tao___ Heineken____ Corona____ Budweiser____ Carlsberg_____
- How satisfied are you with the labor day shopping experience?
  ___very satisfied ___satisfied ____neutral ___dissatisfied ___very dissatisfied

16

Data exploration

- To present data in a form that makes sense to people so that we can have a general idea about the data and find directions for further analysis
  - **single variable**
  - **relationships between variables**
- Methods to explore:
  - **Statistics**
    - Descriptive statistics: counts, percentages, averages, and measures of variability, etc.
    - Tables
  - **Graphs**: bar chart, line chart, histograms, scatterplots, box plots and time series graphs, etc.

17

Single variable

18

Single variable statistics

Categorial variables	Interval variables

• Count the number of observations in each category
• Frequency distribution: frequency of observations in each category
	• Measures of **central tendency**
	  • Mean, Median
	  • Minimum, Maximum, Percentiles, and Quartiles
	• Measures of **dispersion/variability**
	  • Variance
	  • Standard deviation
	• Measures of distribution **shape**
	  • Skewness: occurs when the sample is lack of symmetry
	  • Kurtosis: this is all about the extreme observations.

19

Measures of Central Tendency

Arithmetic Mean

- Affected by unusually large or small observations (outliers)

Median

- Middle value when data are ordered from smallest to largest.
- Not affected by extremes

20

20

Minimum, Maximum, Percentiles, and Quartiles

- Minimum and Maximum
- For any percentage *p*, the *p*th **percentile** is the value such that a percentage *p* of observations are smaller than it.
- The **quartiles** divide the data into four groups, each with (approximately) a quarter of all observations.
  - The first, second and third quartiles are the percentiles corresponding to *p* = 25%, *p* = 50%, and *p* = 75%.
  - By definition, the second quartile (*p* = 50%) is equal to the median.

21

21

Measures of Dispersion

- Variance
  - Population
  - Sample

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

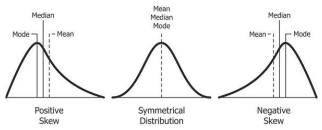
- Standard Deviation
  - = Square root of variance
  - The standard deviation has the same units of measurement as the original data, unlike variance

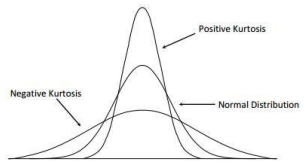
22

22

Measures of Distribution Shape (optional)

- Distribution of numeric variables
- Skewness (CS)
  - -0.5 < CS < 0.5 indicates relative symmetry
- Kurtosis
  - Refers to the peakedness or flatness of a distribution, The lower the kurtosis, the flatter the distribution.
  - CK < 3: more flat with wide degree of dispersion
  - CK >3 more peaked with less dispersion





23

23

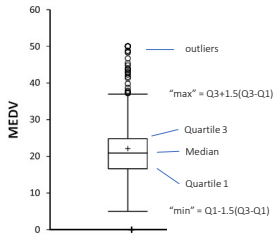
Distribution graphs

Categorial variables	Interval variables
• **Bar chart**: Display frequency distribution - "how many" observations in each category	• **Histogram**: Display "how many" observations in each of the bins occur in a data set • **Box plot**: Quartiles

24

Box Plots

- A **box plot** (or **box-whisker plot**) is an alternative type of chart for showing the distribution of a variable
  - The elements of a generic box plot:



- ✦ Top outliers defined as those above Q3+1.5(Q3-Q1).
- ✦ "max" = maximum of non-outliers
- ✦ "min" = minimum of non-outliers
- ✦ IQR = Q3-Q1: Interquartile range

25

25

Outliers

- An **outlier** is a value or an entire observation (row) that lies well outside of the norm.
  - Some statisticians define an outlier as any value more than three standard deviations from the mean, but this is only a rule of thumb
- Domain knowledge is required to decide whether an "outlier" is truly an error, an abnormal or a special case
- What to do with outliers:
  - Careful review for more information to make decisions
  - Remove if confirmed to be errors or unexplained abnormal

26

26

Multiple variables

27

Two interval variables

Statistics	Graphs
•Covariance •Correlation	•Scatterplot •Correlation heatmap

28

Relationship: two continuous variables - Covariance

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- A measure of the linear association between two (continuous) variables, X and Y.
- For a population, COVARIANCE.P:
- For a sample, COVARIANCE.S:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

29

29

Relationship: two continuous variables - Correlation

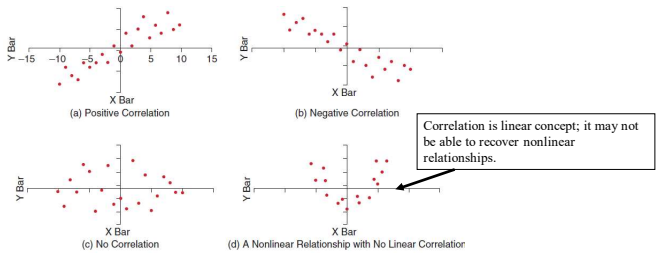
$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

- A measure of the linear association between two variables, X and Y.
- The scale of measurement is normalized
- The correlation value is always between -1 and +1.

30

30

Correlation



31

31

One interval vs. one categorical

Statistics	Graphs
• Compare the distribution of the interval variable in each category	• Side-by-side plots across categories • Or subplots of each category

32

Take-away for today

- Where data come from
- Measurement levels
- Single variable exploration
  - Categorical variables: frequency count, bar chart
  - Continuous variables: summary statistics, histogram, boxplot
- Multiple variable exploration
  - Two continuous variable: covariance, correlation, scatterplot, heatmap
  - A continuous variable vs. a categorical variable: comparison of statistics, boxplots across categories