

# Statistics and Data Analysis

## Descriptive Statistics (1): Visualization

Ling-Chieh Kung

Department of Information Management  
National Taiwan University

## Visualizing the data

- ▶ We will introduce some common ways to **summarize** a set of data.
  - ▶ By **graphs**.
  - ▶ By **statistics**.
- ▶ This is always the **first step** of any data analysis project: To get intuitions that guide our directions.

# Road map

- ▶ **Frequency distributions.**
- ▶ Quantitative data graphs.
- ▶ Qualitative data graphs.
- ▶ Visualizing two variables.

# Descriptive Statistics

- ▶ Consider the column “cnt” in the sheet “Day” of the Excel file “Bike.xlsx”.
  - ▶ Each number is the number of rentals in a day.
  - ▶ 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 1263, 1162, 1406, 1421, 1248, 1204, 1000, 683, 1650, 1927, ..., and 2729.
- ▶ To get more ideas about this data set, we do **Descriptive Statistics**.
  - ▶ Sometimes called **exploratory data analysis**.
  - ▶ Using data graphs to visualize data or using numbers to summarize data.

# Frequency distributions

- ▶ The original 731 numbers form a set of **ungrouped data**.
- ▶ When data are ungrouped, visualizing them is hard.
- ▶ We start by **grouping** them into a **frequency distribution**.
  - ▶ Grouped data presented in the form of class intervals and frequencies.
- ▶ Let's create an intuitive frequency distribution.

## Frequency distributions: an example

- ▶ Let's group the daily bike rental data into a frequency distribution.
  - ▶ Let's label these 731 numbers are  $x_1, x_2, \dots$ , and  $x_{731}$ .
- ▶ Step 1: Find the **range**:

$$\max_{i=1, \dots, 731} \{x_i\} - \min_{i=1, \dots, 731} \{x_i\} = 8714 - 22 = 8692.$$

- ▶ Step 2: Let's divide the range into **classes**:
  - ▶ These classes are intervals with equal lengths.
  - ▶ A typical number of classes is between **5 and 15**.
  - ▶ Let's choose 9, for example.
- ▶ Step 3: **Class width**  $\geq \frac{8692}{9} \approx 965.78$ . Let's try 1000.

## Frequency distributions: an example

- ▶ The resulting classes:

Class	Class interval	(Which means)
1	$[0, 1000)$	$0 \leq x < 1000$
2	$[1000, 2000)$	$1000 \leq x < 2000$
3	$[2000, 3000)$	$2000 \leq x < 3000$
	⋮	
8	$[7000, 8000)$	$7000 \leq x < 8000$
9	$[8000, 9000)$	$8000 \leq x < 9000$

- ▶ How about  $[0, 999]$ ,  $[1000, 1999]$ , etc.?
- ▶ How about  $(0, 1000]$ ,  $(1000, 2000]$ , etc.?

## Frequency distributions: an example

- ▶ Then we **count** to get the frequency distribution at the right.
- ▶ This is a set of **grouped data**.
- ▶ Some remarks:
  - ▶ Typically we have 5 to 15 classes.
  - ▶ Typically all classes have the same width.
  - ▶ Be aware of class endpoints! Classes should NOT overlap with each other.
  - ▶ If there are **outliers**, they should be removed first.

Class interval	Frequency
[0, 1000)	18
[1000, 2000)	80
[2000, 3000)	74
[3000, 4000)	107
[4000, 5000)	166
[5000, 6000)	106
[6000, 7000)	86
[7000, 8000)	82
[8000, 9000)	12



## Outliers

- ▶ An **outlier** in a data set is a value that is “very weird.”
  - ▶ May be due to a very rare case.
  - ▶ May be due to a typo.
- ▶ For examples,
  - ▶ A promotion makes the rental free on December 31, 2012. The number of daily rentals is 34231 (originally 2290).
  - ▶ One mistakenly typed 654 in January 1, 2011, as 6544.
- ▶ Some outliers may be identified with a frequency distribution. Some are not.

Class interval	Frequency
[0, 1000)	<b>17</b>
[1000, 2000)	80
[2000, 3000)	<b>73</b>
[3000, 4000)	107
[4000, 5000)	166
[5000, 6000)	106
[6000, 7000)	<b>87</b>
[7000, 8000)	82
[8000, 9000)	12
⋮	
[34000, 35000)	<b>1</b>

## Something more

- ▶ We may add **class midpoints**, **relative frequencies**, and **cumulative frequencies** into a frequency table:

Class interval	Frequency	Class midpoint	Relative frequency	Cumulative frequency
[0, 1000)	18	500	2.46%	18
[1000, 2000)	80	1500	10.94%	98
[2000, 3000)	74	2500	10.12%	172
[3000, 4000)	107	3500	14.64%	279
[4000, 5000)	166	4500	22.71%	445
[5000, 6000)	106	5500	14.50%	551
[6000, 7000)	86	6500	11.76%	637
[7000, 8000)	82	7500	11.22%	719
[8000, 9000)	12	8500	1.64%	731

- ▶ How about **cumulative relative frequencies**?

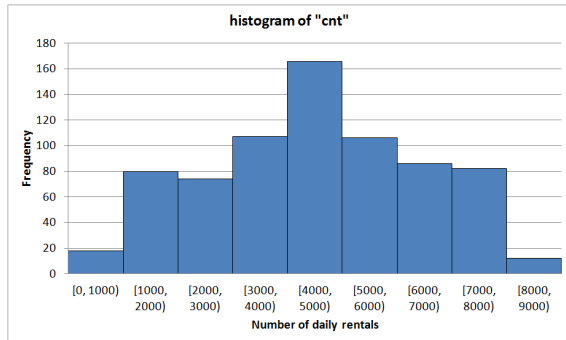
# Road map

- ▶ Frequency distributions.
- ▶ **Quantitative data graphs.**
- ▶ Qualitative data graphs.
- ▶ Visualizing two variables.

# Histograms

- ▶ A frequency distribution may be depicted as a **histogram**.

Interval	Freq.
[0, 1000)	18
[1000, 2000)	80
[2000, 3000)	74
[3000, 4000)	107
[4000, 5000)	166
[5000, 6000)	106
[6000, 7000)	86
[7000, 8000)	82
[8000, 9000)	12



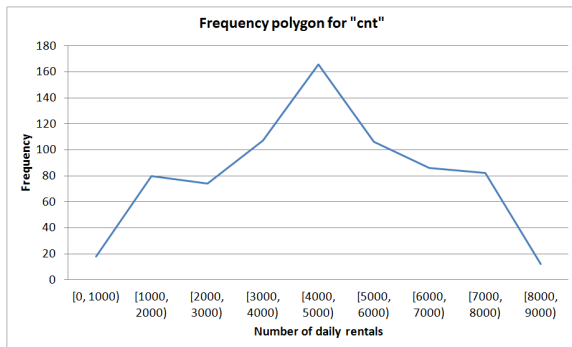
- ▶ It consists of a series of **contiguous** rectangles, each representing the frequency in a class.

# Histograms

- ▶ Histograms may be the most important type of data graphs.
- ▶ One particular reason to draw histograms is to get some ideas about the **distribution**.
  - ▶ Bell shape? M shape? Skewed?
  - ▶ Any outlier?
  - ▶ We will discuss distributions in more details.

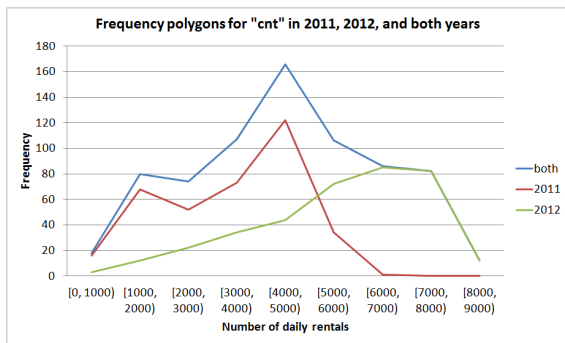
## Frequency polygons

- ▶ Alternatively, we may draw a **frequency polygon** by using **line segments** connecting dots plotted at class **midpoints**.
  - ▶ The information contained in a frequency polygon is quite similar to that contained in a histogram.



## Frequency polygons

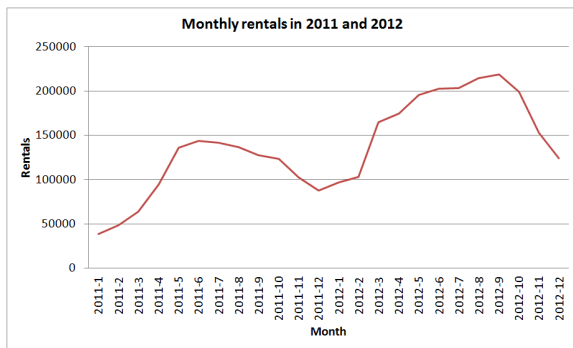
- ▶ It is more convenient to use a frequency polygon to **compare multiple** frequency distributions.



- ▶ Both: Uni-modal and symmetric.
  - ▶ 2011: Bi-modal and skewed to the right (right-tailed).
  - ▶ 2012: Uni-modal and skewed to the left (left-tailed).
- ▶ Warning: People may **misinterpret** a frequency polygon as a **line chart** (for data with a time sequence).

## Line charts

- ▶ A **line chart** is useful in depicting a time series data.
  - ▶ A two-dimensional data set whose first dimension (the  $x$ -axis) is for labels of time points.
- ▶ It visualizes how a quantity changes as time goes by.
- ▶ For our monthly bike rentals:





## Road map

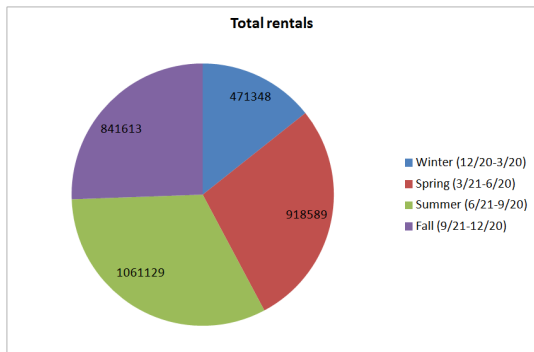
- ▶ Frequency distributions.
- ▶ Quantitative data graphs.
- ▶ **Qualitative data graphs.**
- ▶ Visualizing two variables.

# Pie charts

- ▶ A **pie chart** is a **circular** depiction of data where each slice represents the percentage of the corresponding category.
- ▶ It visualizes **relative frequency distributions** well.
- ▶ For our bike rental data set:
  - ▶ What are the proportions of rentals in the four seasons?
  - ▶ What are the proportions of rentals on the seven days of a week?

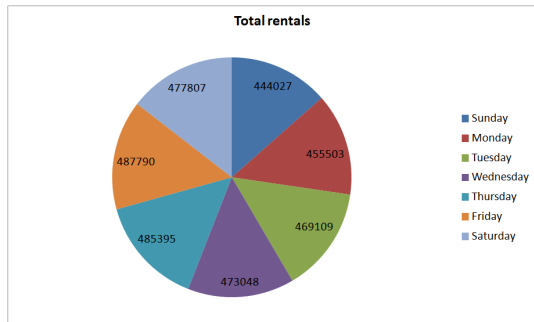
## A pie chart for seasonal rentals

Season	Total rentals	Proportion
Winter (12/20-3/20)	471348	14.3%
Spring (3/21-6/20)	918589	27.9%
Summer (6/21-9/20)	1061129	32.2%
Fall (9/21-12/20)	841613	25.6%



## A pie chart for rentals among weekdays

Day	Total rentals
Sunday	444027
Monday	455503
Tuesday	469109
Wednesday	473048
Thursday	485395
Friday	487790
Saturday	477807



## Data not appropriate for pie charts

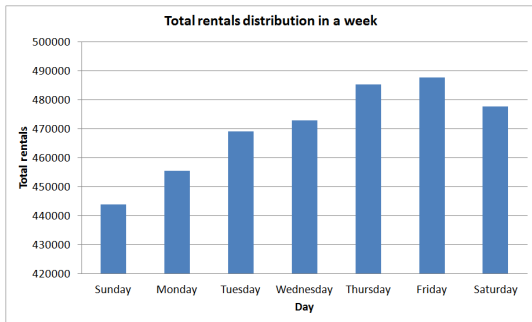
- ▶ Pie charts are used to visualize proportions, i.e., subtotals over the overall total.
- ▶ It should **not** be used to compare **averages**.
  - ▶ The total numbers of rentals made by male and female users are appropriate for a pie chart.
  - ▶ The average numbers of rentals per male and female users are not appropriate for a pie chart.

# Bar charts

- ▶ Pie charts are useful in visualizing the **proportions** of each categories.
- ▶ In demonstrating the **differences** among categories, a **bar chart** is a better choice.
  - ▶ The larger the category, the longer the bar.
  - ▶ Some people draw bars vertically; some horizontally.

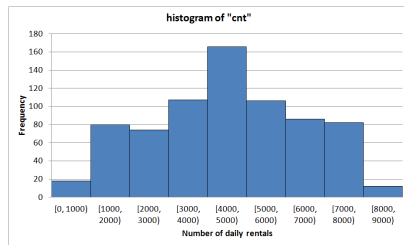
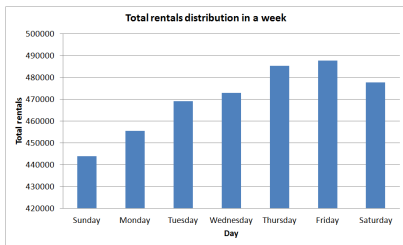
# Bar charts

Day	Total rentals
Sunday	444027
Monday	455503
Tuesday	469109
Wednesday	473048
Thursday	485395
Friday	487790
Saturday	477807



# Bar charts v.s. histograms

- ▶ What are the differences that distinguish a bar chart from a histogram?



- ▶ A bar chart uses **noncontiguous** bars to visualize **categorical** data.
- ▶ A histogram uses **contiguous** bars to visualize **quantitative** data.



# Road map

- ▶ Frequency distributions.
- ▶ Quantitative data graphs.
- ▶ Qualitative data graphs.
- ▶ **Visualizing two variables.**

## Visualizing two variables

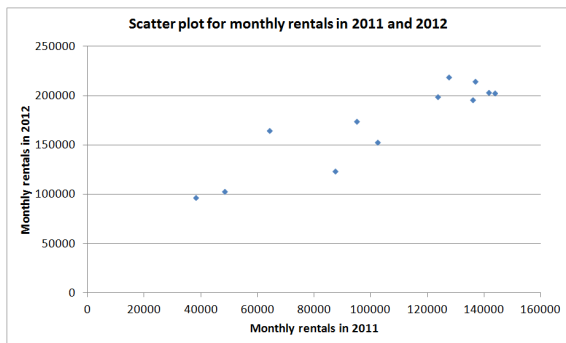
- ▶ When we have data for two variables, typically we want to identify whether there is any **relationship** between them.
- ▶ Visualizing the data in a two-dimensional manner helps.

## Scatter plots

- ▶ Sometimes in an observation there are **two** values recorded.
- ▶ When the two values are both measured in quantitative scales, we may depict each observation as a point on a plane to create a **scatter plot**.
- ▶ For our bike rental example:
  - ▶ How do monthly rentals in 2011 and those in 2012 relate with each other?
  - ▶ How do daily casual and registered rentals relate with each other?

# Monthly rentals in 2011 and 2012

Month	2011	2012
1	38189	96744
2	48215	103137
3	64045	164875
4	94870	174224
5	135821	195865
6	143512	202830
7	141341	203607
8	136691	214503
9	127418	218573
10	123511	198841
11	102167	152664
12	87323	123713



# Daily casual and registered rentals

day	casual	registered
1	331	654
2	131	670
3	120	1229
⋮		
731	439	2290

