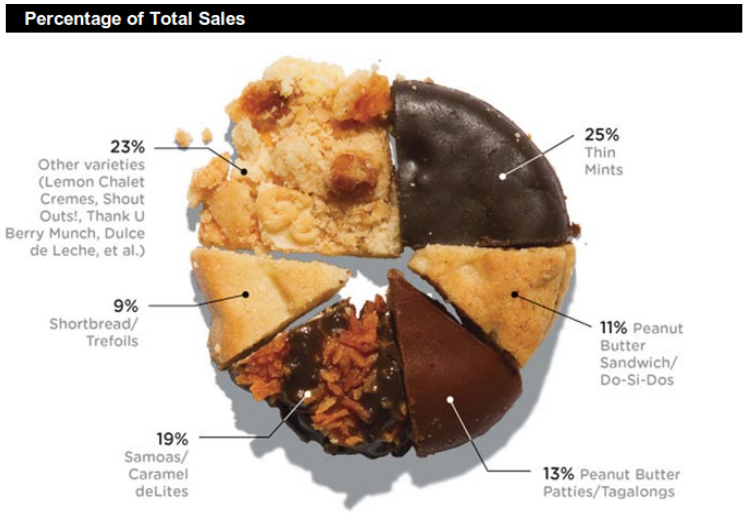
Chapter 2
Part A – Descriptive Statistics: Tabular and Graphical Displays

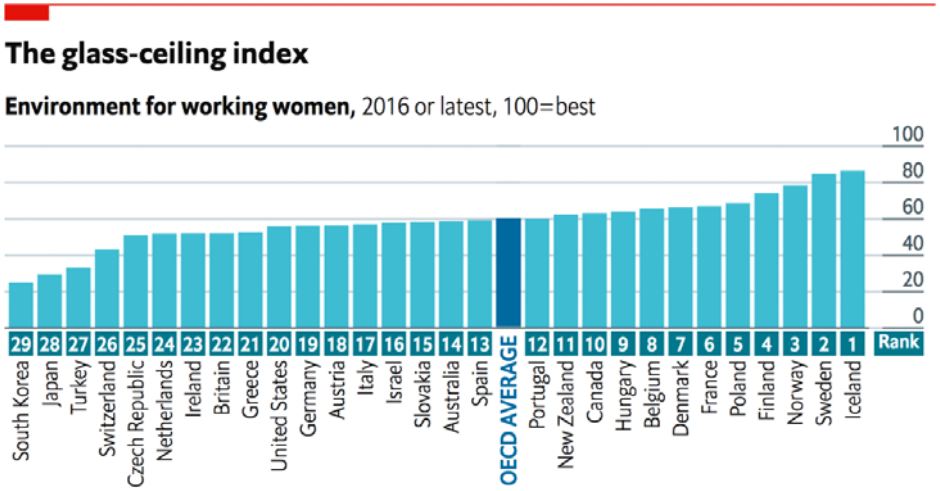
- Summarizing Data for a Categorical Variable
  - Categorical data use labels or names to identify categories of like items.
- Summarizing Data for a Quantitative Variable
  - Quantitative data are numerical values that indicate how much or how many.

Summarizing Categorical Data

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Chart
- Pie Chart

Rating	Frequency	Relative Frequency	Percent Frequency
Poor	2	.10	10
Below Average	3	.15	15
Average	5	.25	25
Above Average	9	.45	45
Excellent	1	.05	5
Total	20	1.00	100





Example: Marada Inn

- Guests staying at Marada Inn were asked to rate the quality of their accommodations as being *excellent*, *above average*, *average*, *below average*, or *poor*.
- The ratings provided by a sample of 20 guests are:

Below Average	Average	Above Average	Above Average
Above Average	Above Average	Above Average	Average
Above Average	Below Average	Below Average	Average
Average	Poor	Poor	Above Average
Above Average	Excellent	Above Average	Average

Frequency, Relative Frequency, and Percent Frequency Distributions

Example: Marada Inn

Rating	Frequency	Relative Frequency	Percent Frequency
Poor	2	.10	10
Below Average	3	.15	15
Average	5	.25	25
Above Average	9	.45	45
Excellent	1	.05	5
Total	20	1.00	100

Frequency Distribution

- A frequency distribution is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.

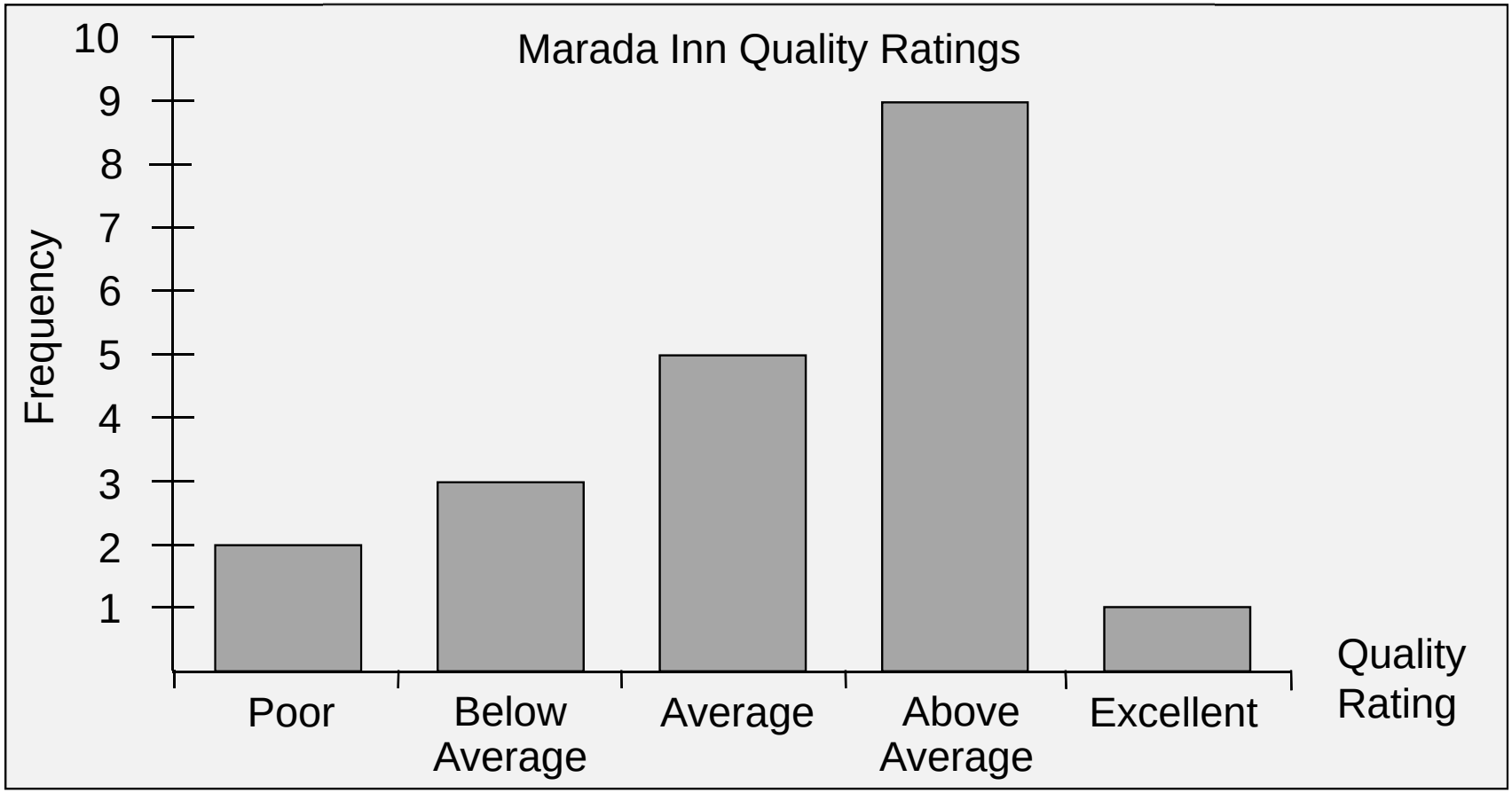
Relative Frequency Distribution

- $\text{Relative frequency of a class} = \dfrac{\text{Frequency of the class}}{n}$

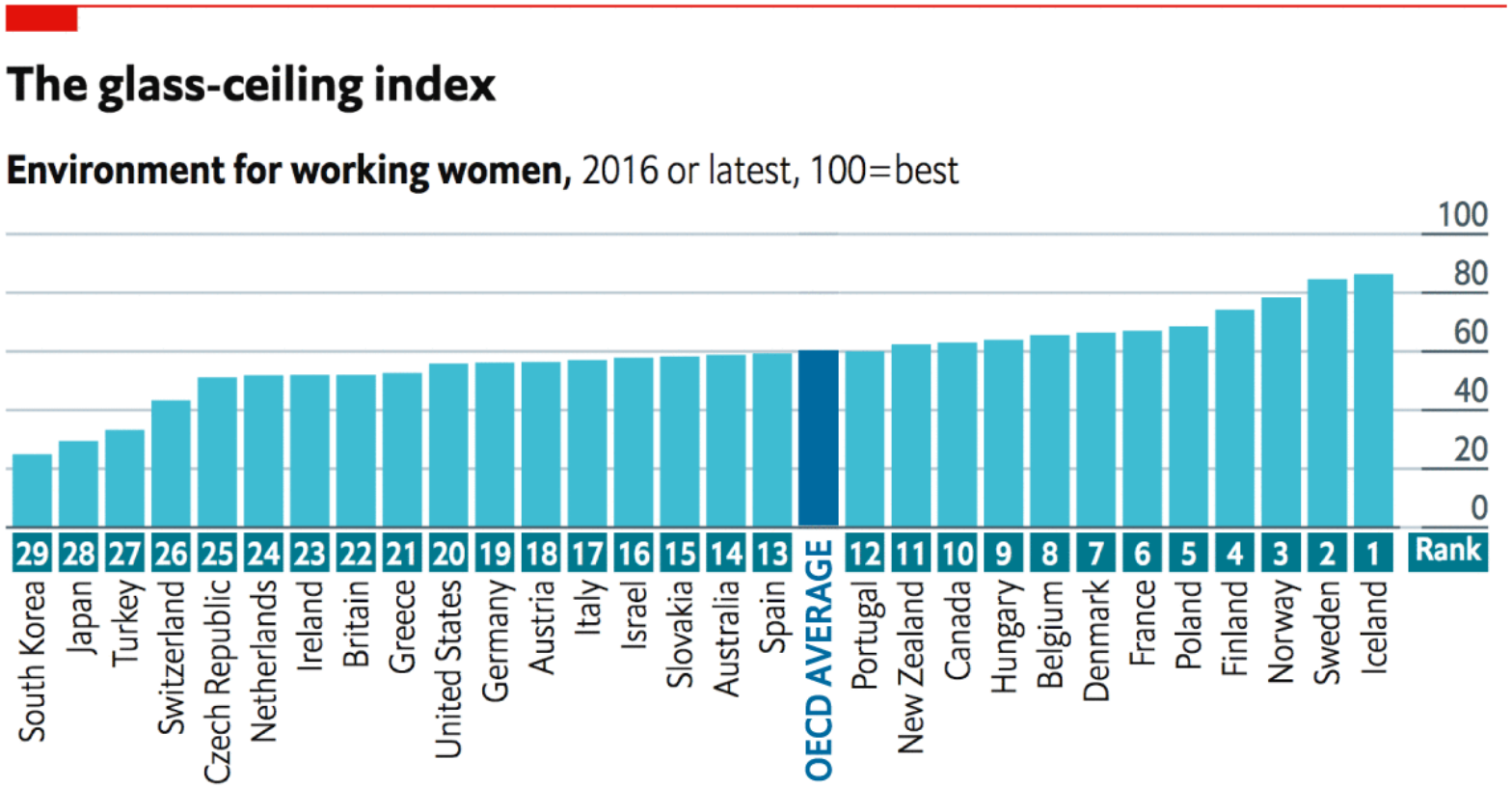
Percent Frequency Distribution

- The percent frequency of a class is the relative frequency multiplied by 100.

Bar Chart


Marada Inn Quality Ratings
Frequency
10
9
8
7
6
5
4
3
2
1
Poor
Below Average
Average
Above Average
Excellent
Quality Rating

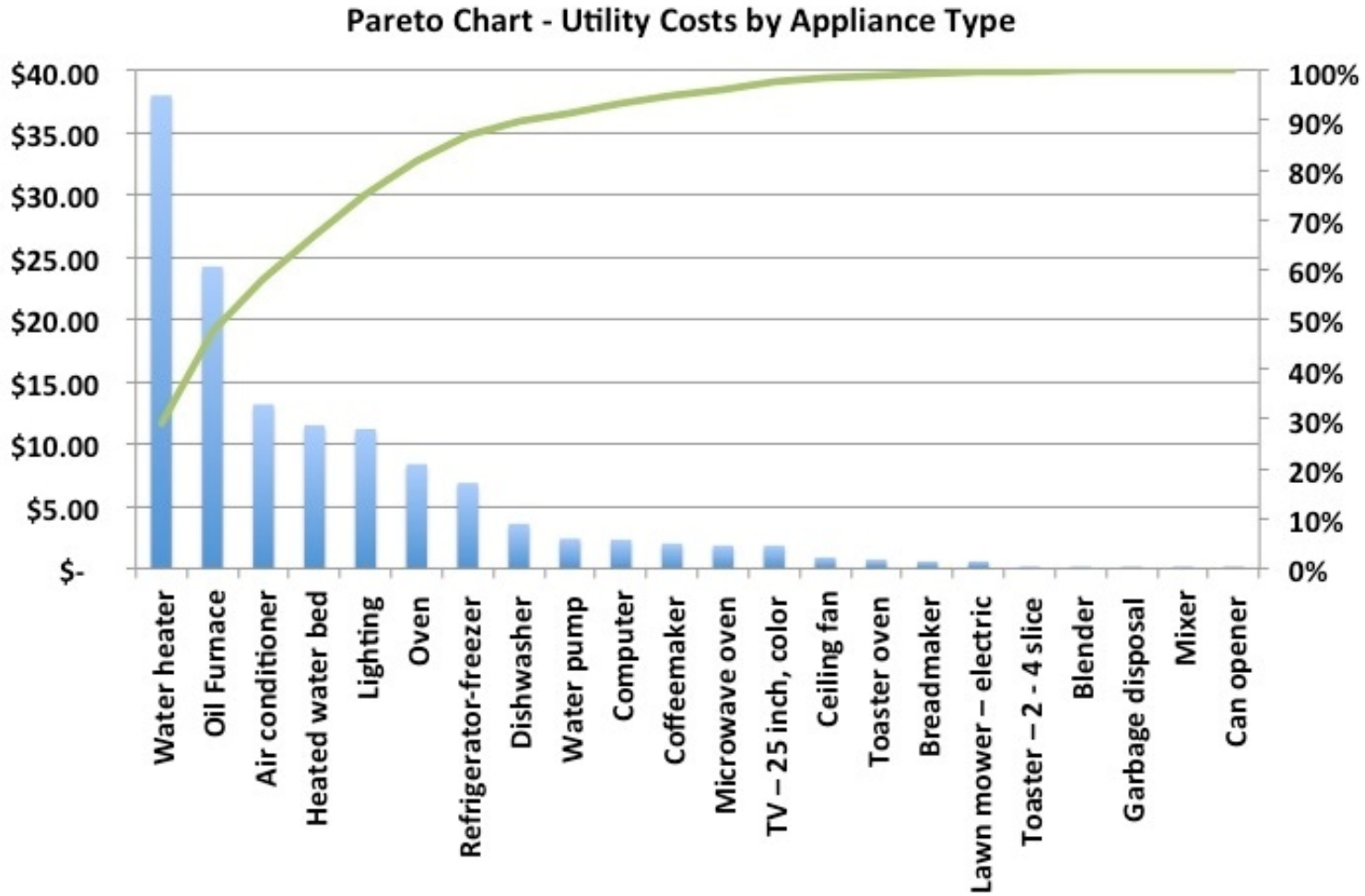
The best and worst places to be a working woman



Mar 8th 2017 by THE DATA TEAM (Economist) https://www.economist.com/blogs/graphicdetail/2017/03/daily-chart-0?fsrc=scn/fb/te/bl/ed/

Bar Chart

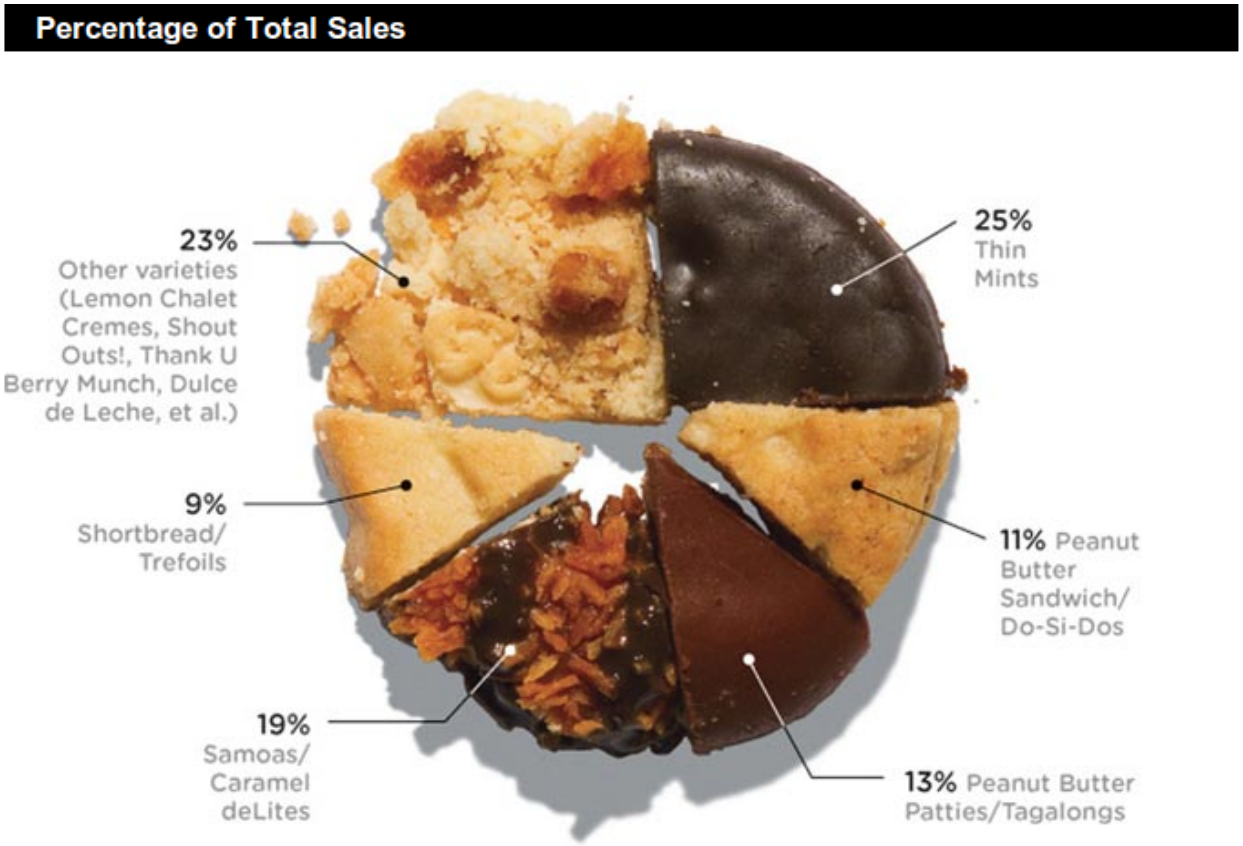
- A bar chart is a graphical display for depicting **qualitative** data.
- On one axis (usually the horizontal axis), we specify the labels that are used for each of the classes.
- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).
- Using a bar of fixed width drawn above each class label, we extend the height appropriately.
- The bars are **separated** to emphasize the fact that each class is a separate category.



kwvs=22z z z 1fthdfdghp |1frp 2fth0erg|0ri0nqrz dngjh2frqwlqxrxv0lp suryhp hqw2txdolw|0frqwuro0wrrov2sduhwr0fkduw2

Pareto Diagram

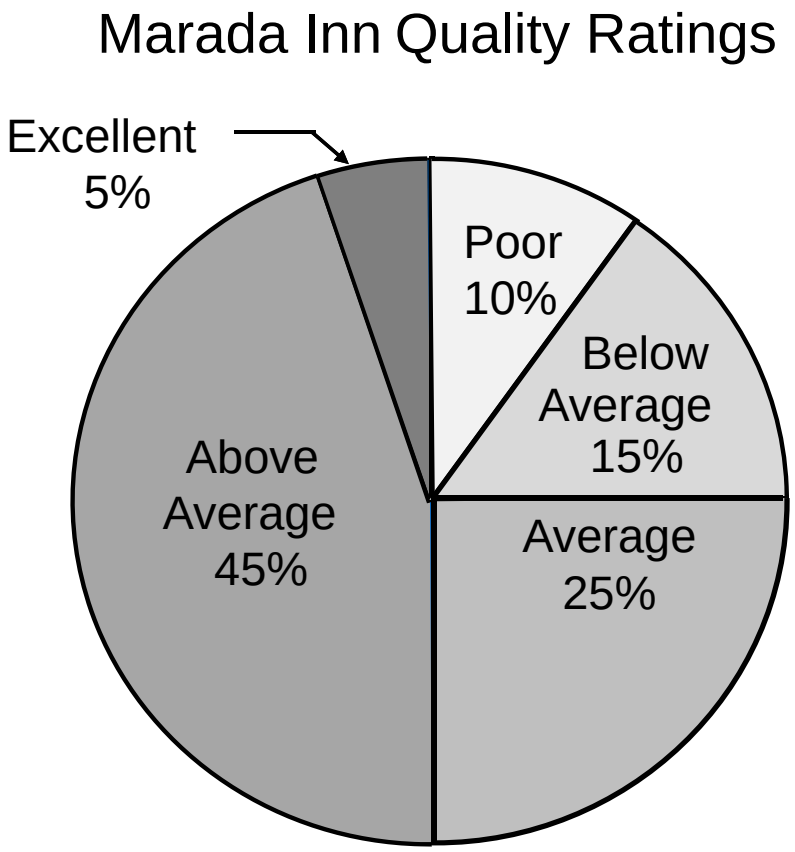
- In quality control, bar charts are used to identify the most important causes of problems.
- When the bars are arranged in descending order of height from left to right (with the most frequently occurring cause appearing first) the bar chart is called a Pareto diagram.
- This diagram is named for its founder, Vilfredo Pareto, an Italian economist.



Vrxufh=Z luhg
XUO=kwws=22z z z 1z luhg1frp 253 4423 ; 2vwbgdwdjluovfrxwfrrnlhv2

Photo: Celine Grouard

Pie Chart



- One-half of the customers surveyed gave Marada a quality rating of "above average" or "excellent" (looking at the left side of the pie). This might please the manager.
- For each customer who gave an "excellent" rating, there were two customers who gave a "poor" rating (looking at the top of the pie). This should displease the manager.

Pie Chart

- The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for **categorical** data.
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.
- Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume .25(360) = 90 degrees of the circle.

Summarizing Quantitative Data

- Frequency Distribution
- Relative Frequency and Percent Frequency Distributions
- Dot Plot
- Histogram
- Cumulative Distributions
- Stem-and-Leaf Display

Example: Hudson Auto Repair

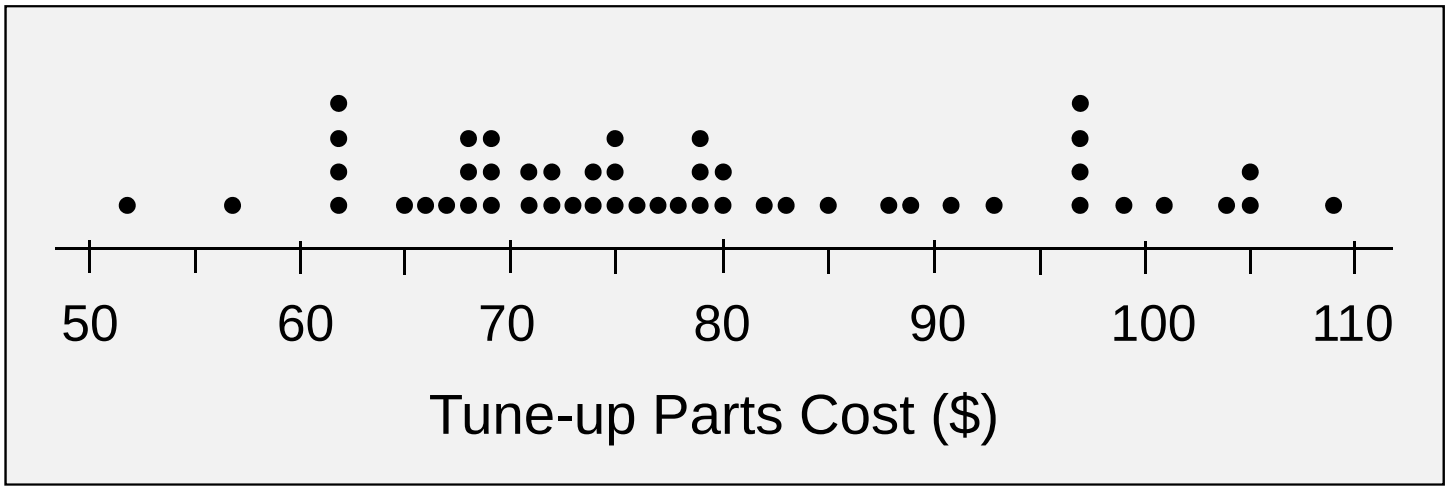
Sample of Parts Cost($) for 50 Tune-ups

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

The manager of Hudson Auto would like to gain a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups.

Dot Plot

- Example: Hudson Auto Repair



- One of the simplest graphical summaries of data is a dot plot.
- A horizontal axis shows the range of data values.
- Then each data value is represented by a dot placed above the axis.

Frequency Distribution

The three steps necessary to define the classes for a frequency distribution with quantitative data are:

1. Determine the number of non-overlapping classes.
2. Determine the width of each class.
3. Determine the class limits.

Frequency Distribution

- Guidelines for Determining the Number of Classes
  - Use between 5 and 20 classes.
  - Data sets with a larger number of elements usually require a larger number of classes.
  - Smaller data sets usually require fewer classes.
  - The goal is to use enough classes to show the variation in the data, but not so many classes that some contain only a few data items.

Frequency Distribution

- Guidelines for Determining the Width of Each Class
  - Use classes of equal width.
  - Approximate Class Width =

$$\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$$

  - Making the classes the same width reduces the chance of inappropriate interpretations.

Frequency Distribution

- Note on Number of Classes and Class Width
  - In practice, the number of classes and the appropriate class width are determined by trial and error.
  - Once a possible number of classes is chosen, the appropriate class width is found.
  - The process can be repeated for a different number of classes.
  - Ultimately, the analyst uses judgment to determine the combination of the number of classes and class width that provides the best frequency distribution for summarizing the data.

Frequency Distribution

- Guidelines for Determining the Class Limits
  - Class limits must be chosen so that each data item belongs to one and only one class.
  - The lower class limit identifies the smallest possible data value assigned to the class.
  - The upper class limit identifies the largest possible data value assigned to the class.
  - The appropriate values for the class limits depend on the level of accuracy of the data.
  - An open-end class requires only a lower class limit or an upper class limit.

Frequency Distribution

- Class Midpoint
  - In some cases, we want to know the midpoints of the classes in a frequency distribution for quantitative data.
  - The class midpoint is the value halfway between the lower and upper class limits.

Frequency, Relative Frequency, and Percent Frequency Distributions

- Example: Hudson Auto Repair

  If we choose six classes:

  Approximate Class Width = (109 - 50)/6 = 9.83≅ 10

Parts Cost ($)	Frequency	Relative Frequency	Percent Frequency
50-59	2	2/50 = .04	.04(100) = 4
60-69	13	.26	26
70-79	16	.32	32
80-89	7	.14	14
90-99	7	.14	14
100-109	5	.10	10
Total	50	1.00	100

- Rqo 7(riwkh sduw frvw duh lq wkh ’8308< fodvv1
- 63(riwkh sduw frvw duh xqghu’:31
- Wkh juhdwhvw shufhqwdjh +65(rudqp rvwrqh0wklug, riwkh sduw frvw duh lq wkh ’:30:< fodvv1
- 43(riwkh sduw frvw duh ’433 rup ruh1

Cumulative Distributions

- Hudson Auto Repair

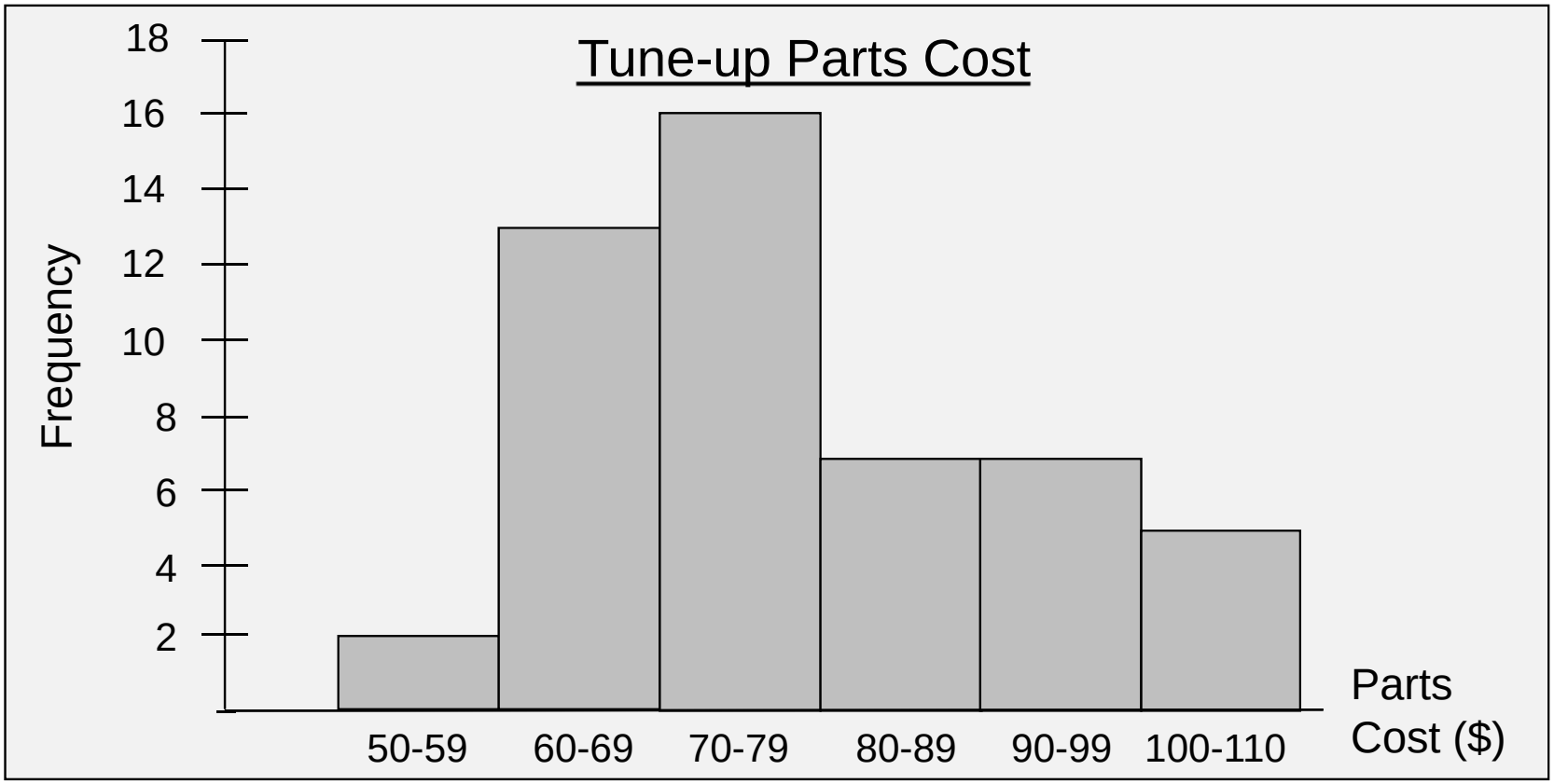
Cost ($)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
≤ 59	2	.04	4
≤ 69	15 = 2+13	.30 = 15/50	30 = .30(100)
≤ 79	31	.62	62
≤ 89	38	.76	76
≤ 99	45	.90	90
≤ 109	50	1.00	100

資料剖析

- 排序
  - 由小到大，由大到小
  - 特定目的
- 找極端值
  - 找Max與Min
  - 判斷正常值域
- 計算全距
  - R=Max-Min
- 計算組數
  - 組數太少，看不出各組的特性
  - 組數太多，不易掌握資料分佈型態與變化
  - k=1+3.322*log n
  - 一般在5~20組之間
- 計算組距
  - W=R ÷k
  - 各組組距最好相等
  - 最好為2 、5 、10的倍數
- 決定組限
  - 由最小值或選定數為第一組的組下限(Lower)
  - 間斷性v.s連續性
  - 依分析目的可任意變換
- 分組次數表
  - 留意組距、組數、組限的決定是有彈性的
  - 其中之一修改，分析出來的效果會不一樣

Histogram

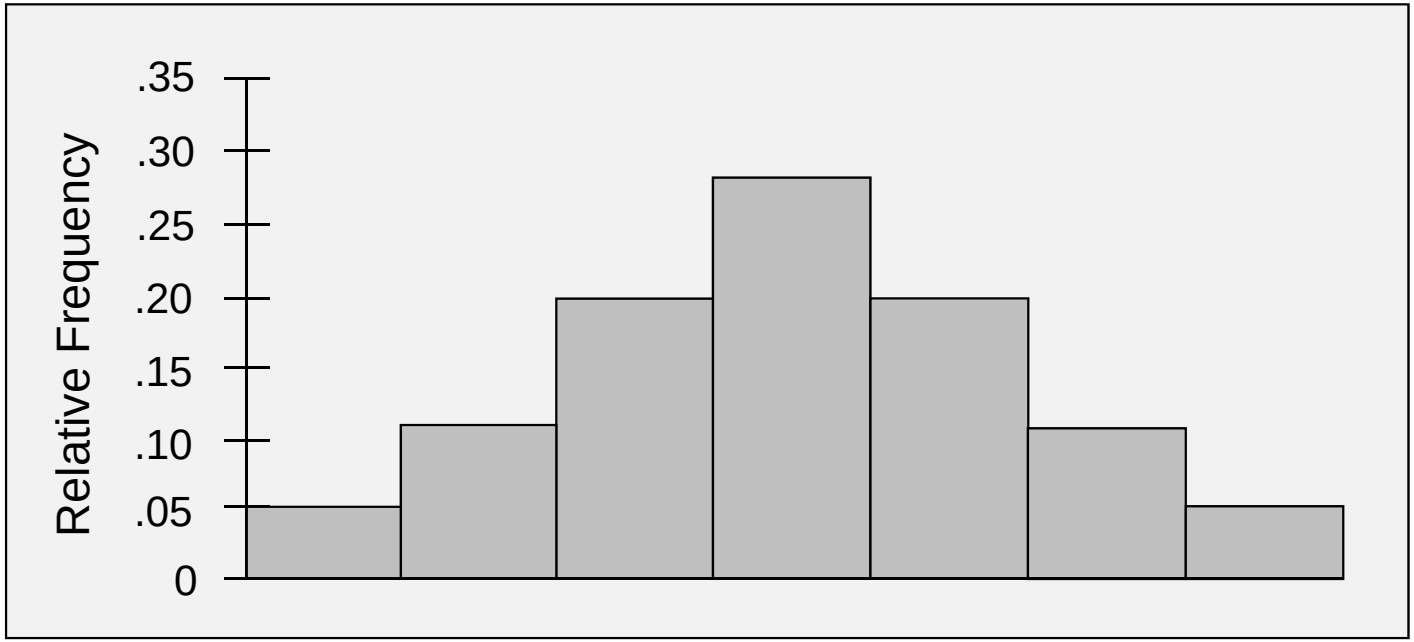
- Example: Hudson Auto Repair



- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes.

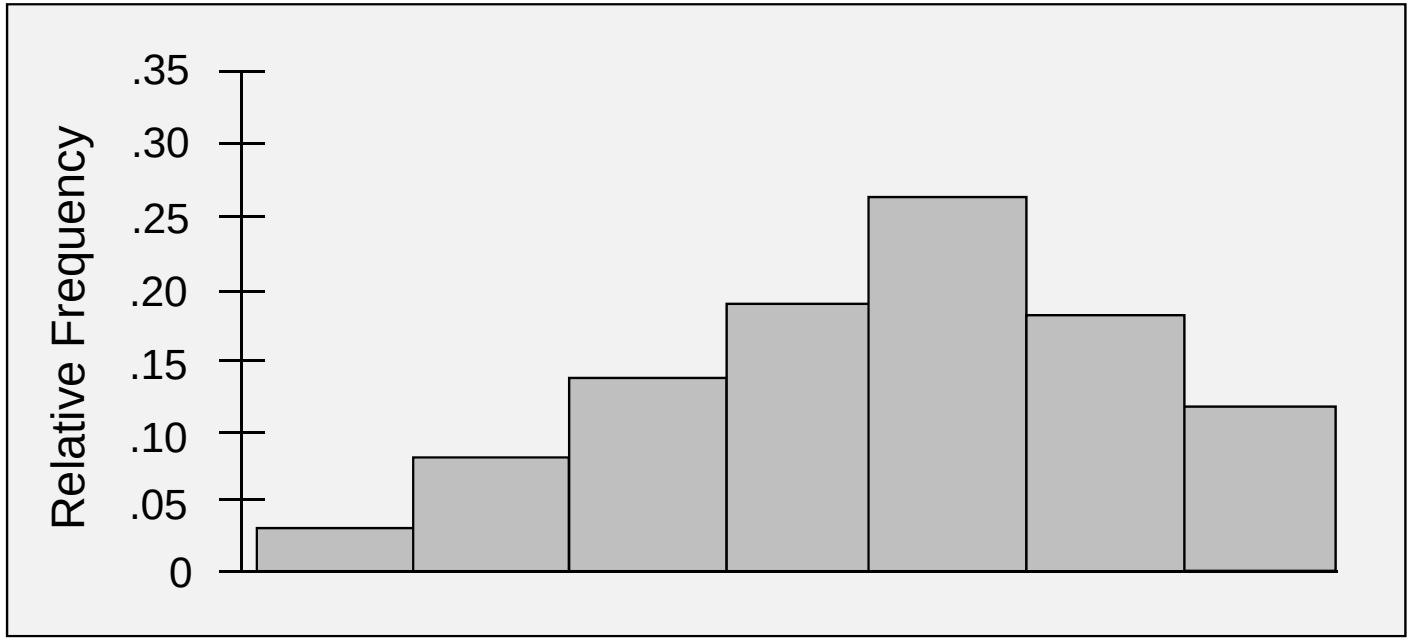
Histograms Showing Skewness (偏態)

- Symmetric (對稱)
  - Left tail is the mirror image of the right tail
  - Example: Heights of People



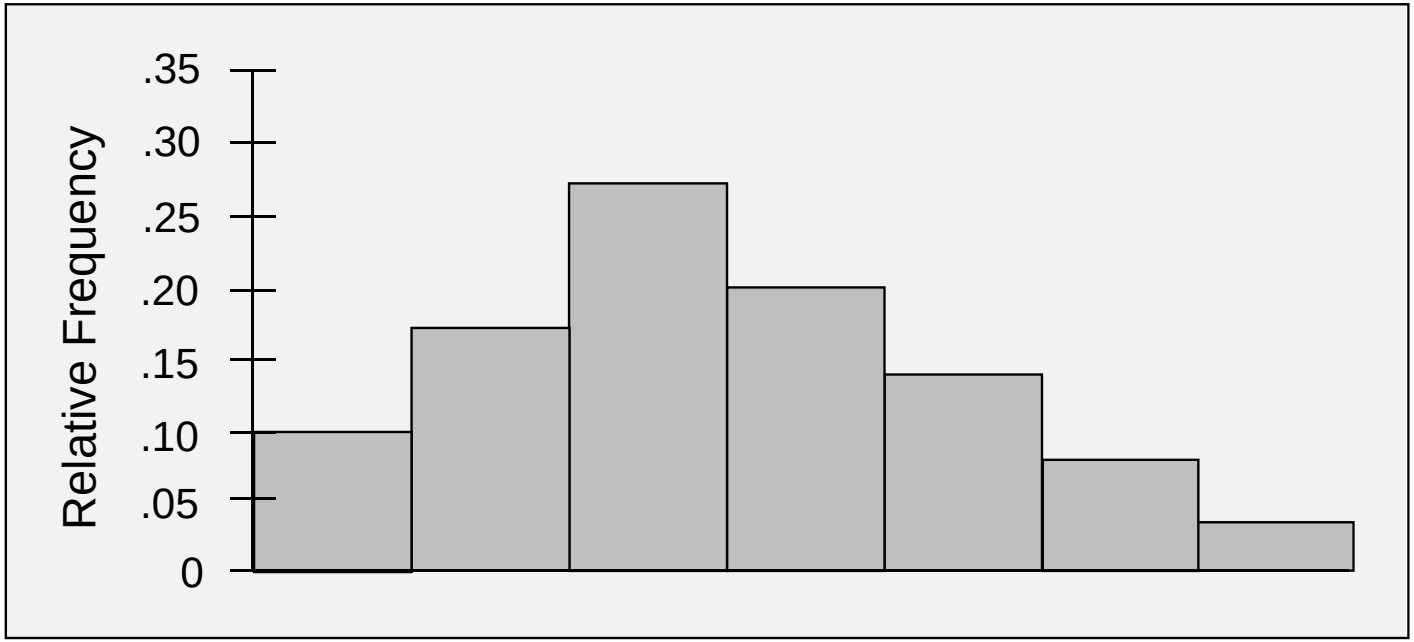
Histograms Showing Skewness

- Moderately Skewed Left (左偏)
  - A longer tail to the left
  - Example: Exam Scores



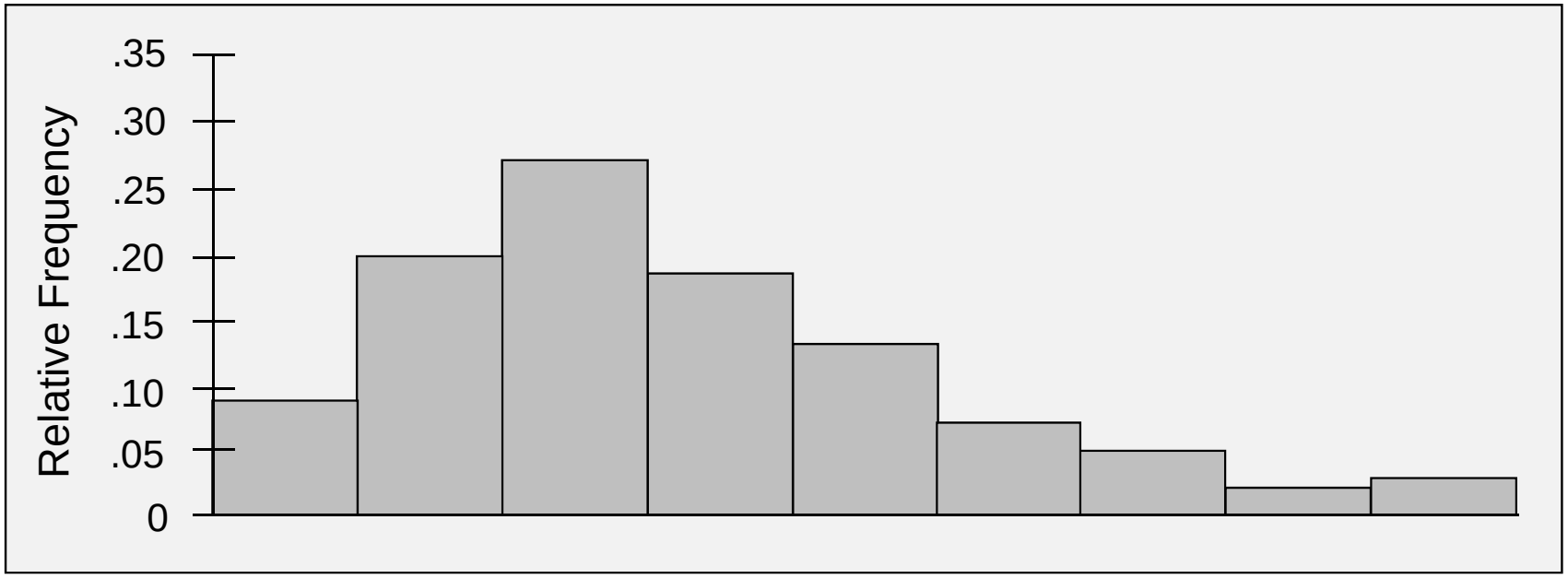
Histograms Showing Skewness

- Moderately Right Skewed (右偏)
  - A Longer tail to the right
  - Example: Housing Values



Histograms Showing Skewness

- Highly Skewed Right
  - A very long tail to the right
  - Example: Executive Salaries



The London School of Economics and the Harvard Business School conducted a study of how chief executive officers (CEOs) spend their day. The study found that CEOs spend on average about 18 hours per week in meetings, not including conference calls, business meals, and public events (*The Wall Street Journal,* February 14, 2012). Shown below is the time spent per week in meetings (hours) for a sample of 25 CEOs.

14	15	18	23	15
19	20	13	15	23
23	21	15	20	21
16	15	18	18	19
19	22	23	21	12

a. What is the least amount of time spent per week on meetings? The highest?

b. Use a class width of two hours to prepare a frequency distribution and a percent frequency distribution for the data.

c. Prepare a histogram and comment on the shape of the distribution.

Entrepreneur magazine ranks franchises using performance measures such as growth rate, number of locations, startup costs, and financial stability. The number of locations for the top 20 U.S. franchises follow (*The World Almanac,* 2012).

Franchise	No. U.S. Locations	Franchise	No. U.S. Locations
Hampton Inns	1,864	Jan-Pro Franchising Intl. Inc.	12,394
ampm	3,183	Hardee's	1,901
McDonald's	32,805	Pizza Hut Inc.	13,281
7-Eleven Inc.	37,496	Kumon Math & Reading Centers	25,199
Supercuts	2,130	Dunkin' Donuts	9,947
Days Inn	1,877	KFC Corp.	16,224
Vanguard Cleaning Systems	2,155	Jazzercise Inc.	7,683
Servpro	1,572	Anytime Fitness	1,618
Subway	34,871	Matco Tools	1,431
Denny's Inc.	1,668	Stratus Building Solutions	5,018

Use classes 0–4999, 5000–9999, 10,000–14,999 and so forth to answer the following questions.

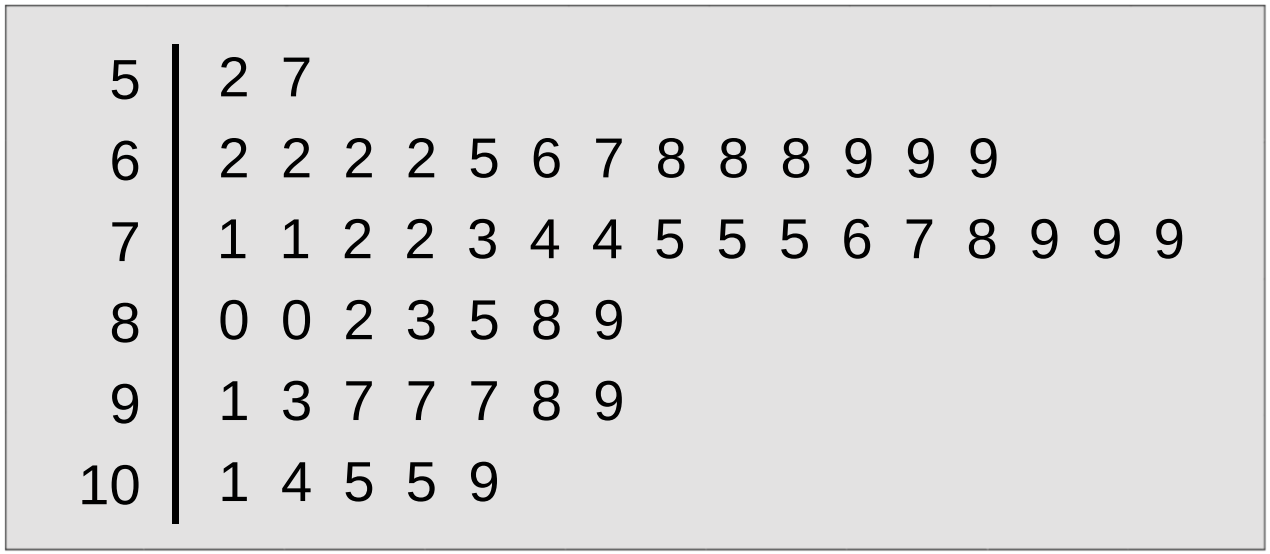
a. Construct a frequency distribution and a percent frequency distribution of the number of U.S. locations for these top-ranked franchises.

b. Construct a histogram of these data.

c. Comment on the shape of the distribution.

Stem-and-Leaf Display

- Example: Hudson Auto Repair



Stem-and-Leaf Display

- A stem-and-leaf display shows both the rank order and shape of a distribution of data.
- It is similar to a histogram on its side, but it has the advantage of showing the actual data values.
- The leading digits of each data item are arranged to the left of a vertical line.
- To the right of the vertical line we record the last digit for each item in rank order.
- Each line (row) in the display is referred to as a stem.
- Each digit on a stem is a leaf.

Construct a stem-and-leaf display for the data

- A psychologist developed a new test of adult intelligence. The test was administered to 20 individuals, and the following data were obtained.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Stretched Stem-and-Leaf Display

- Example: Hudson Auto Repair

5	2
5	7
6	2 2 2 2
6	5 6 7 8 8 8 9 9 9
7	1 1 2 2 3 4 4
7	5 5 5 6 7 8 9 9 9
8	0 0 2 3
8	5 8 9
9	1 3
9	7 7 7 8 9
10	1 4
10	5 5 9

- If we believe the original stem-and-leaf display has condensed the data too much, we can stretch the display vertically by using two stems for each leading digit(s).
- Whenever a stem value is stated twice, the first value corresponds to leaf values of 0 - 4, and the second value corresponds to leaf values of 5 - 9.

Stem-and-Leaf Display

- Leaf Units
  - A single digit is used to define each leaf.
  - In the preceding example, the leaf unit was 1.
  - Leaf units may be 100, 10, 1, 0.1, and so on.
  - Where the leaf unit is not shown, it is assumed to equal 1.
  - The leaf unit indicates how to multiply the stem-and-leaf numbers in order to approximate the original data.

Stem-and-Leaf Display

- Example: Leaf Unit = 0.1

If we have data with values such as

8.6 11.7 9.4 9.1 10.2 11.0 8.8

Leaf Unit = 0.1

8	6 8
9	1 4
10	2
11	0 7

Stem-and-Leaf Display

- Example: Leaf Unit = 10

If we have data with values such as

1806 1717 1974 1791 1682 1910 1838

Leaf Unit = 10

16	8
17	1 9
18	0 3
19	1 7

The 82 in 1682 is rounded down to 80 and is represented as an 8.

Construct a stem-and-leaf display for the following data. Use a leaf unit of 10.

1161	1206	1478	1300	1604	1725	1361	1422
1221	1378	1623	1426	1557	1730	1706	1689