Lecture 2

Describing the data graphically:

Frequency distributions, histograms, and

other types of graphs.

2.1 Frequency Distributions and Histograms

- Frequency Distribution
  - A summary of a set of data that displays the number of observations in each of the distribution’s distinct categories or classes
  - Is a list or a table
  - Contains the values of a variable (or a set of ranges within which the data fall)
  - Also contains the corresponding frequencies with which each value occurs (or frequencies with which data fall within each range)

Discrete Data

- Data that can take on a countable number of possible values
  - Example: An advertiser asks 200 customers how many days per week they read the daily newspaper

Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	**200**

Relative Frequency

- The proportion of total observations that are in a given category.

$$\boxed{\text{Relative frequency} = \frac{f_i}{n}}$$

f_i - Frequency of the ith value of the discrete variable

$n = \sum_{i=1}^{k} f_i$ - Total number of observations

k - The number of different values for the discrete variable

Example

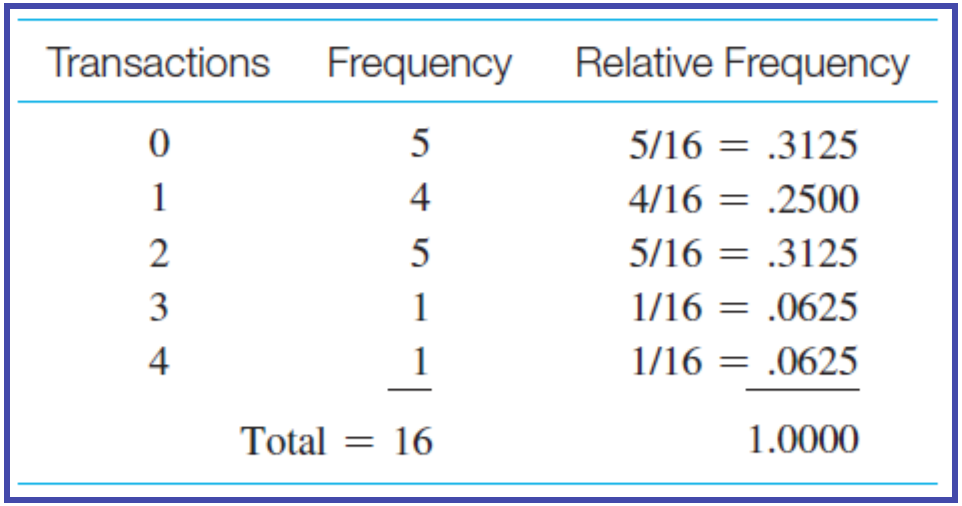
Number of days read	Frequency	Relative Frequency
0	44	.22
1	24	.12
2	18	.09
3	16	.08
4	20	.10
5	22	.11
6	26	.13
7	30	.15
Total	**200**	**1.00**

$$\frac{44}{200} = .22$$

22% of the people in the sample report that they read the newspaper 0 days per week

Developing Frequency Distribution for Discrete Data

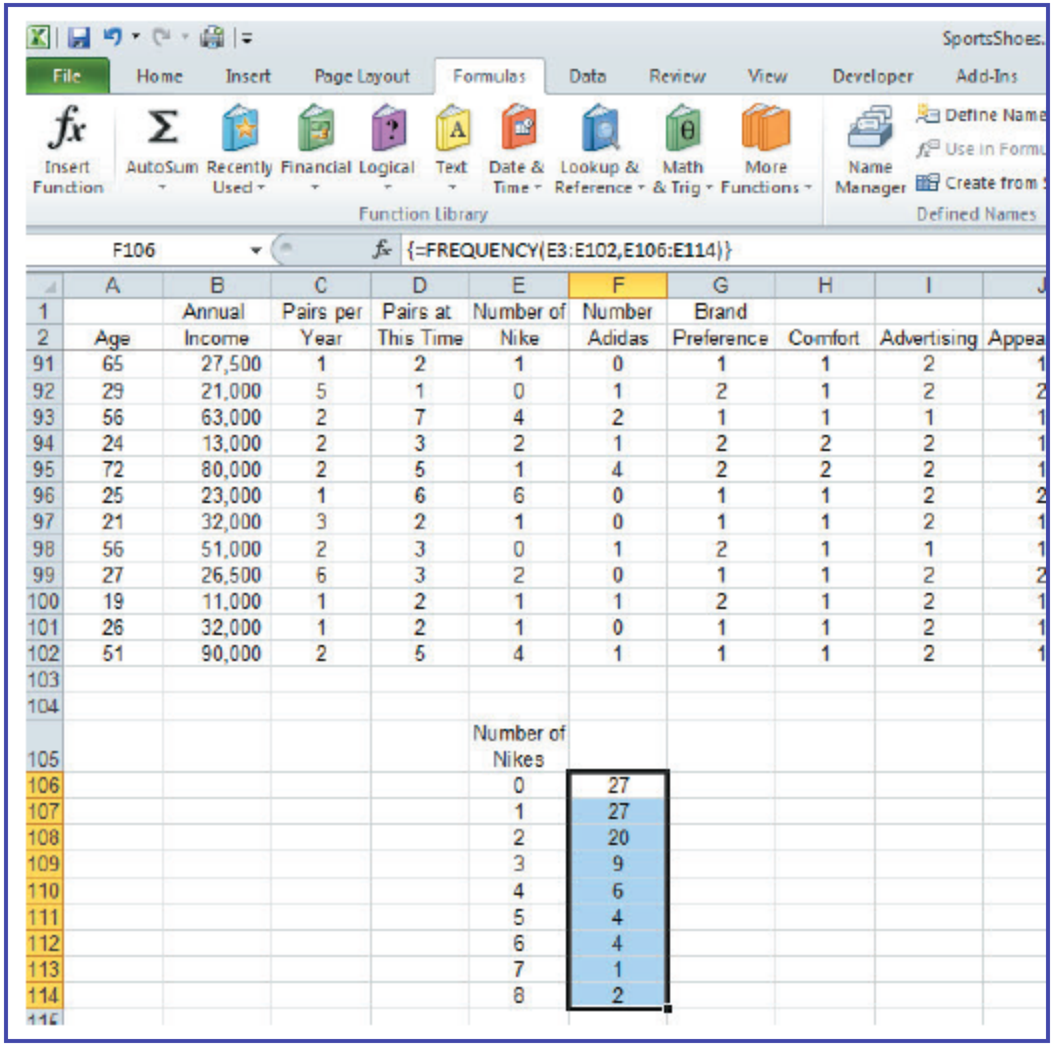
- Step 1: List the possible values
- Step 2: Count the number of occurrences at each value
- Step 3: Determine the relative frequencies



Transactions	Frequency	Relative Frequency
0	5	5/16 = .3125
1	4	4/16 = .2500
2	5	5/16 = .3125
3	1	1/16 = .0625
4	1	1/16 = .0625
	Total = 16	1.0000

How to Do It in Excel?

1. Open File.
2. Enter the Possible Values for the Variable; i.e., 0, 1, 2, 3, 4, etc.
3. Select the cells to contain the Frequency values.
4. Select the **Formulas** tab.
5. Click on the f_x button.
6. Select the **Statistics—FREQUENCY** function.
7. Enter the range of data and the bin range (the number of shoes).
8. Press **Ctrl-Shift-Enter** to determine the frequency values.



F106 {=FREQUENCY(E3:E102,E106:E114)}

	A	B	C	D	E	F	G	H	I	J
1		Annual	Pairs per	Pairs at	Number of	Number	Brand			
2	Age	Income	Year	This Time	Nike	Adidas	Preference	Comfort	Advertising	Appea
91	65	27,500	1	2	1	0	1	1	2	1
92	29	21,000	5	1	0	1	2	1	2	2
93	56	63,000	2	7	4	2	1	1	1	1
94	24	13,000	2	3	2	1	2	2	2	1
95	72	80,000	2	5	1	4	2	2	2	1
96	25	23,000	1	6	6	0	1	1	2	2
97	21	32,000	3	2	1	0	1	1	2	1
98	56	51,000	2	3	0	1	2	1	1	1
99	27	26,500	6	3	2	0	1	1	2	2
100	19	11,000	1	2	1	1	2	1	2	1
101	26	32,000	1	2	1	0	1	1	2	1
102	51	90,000	2	5	4	1	1	1	2	1
103										
104										
105					Number of Nikes					
106					0	27				
107					1	27				
108					2	20				
109					3	9				
110					4	6				
111					5	4				
112					6	4				
113					7	1				
114					8	2				

Grouped Data

- Continuous data
  - data whose possible values are uncountable and that may assume any value in an interval (weight, length, time)
- Discrete data with many possible outcomes (age, income, stock price)
- Summarized in a grouped data frequency distribution
- Data are organized in classes

Criteria for Building Classes

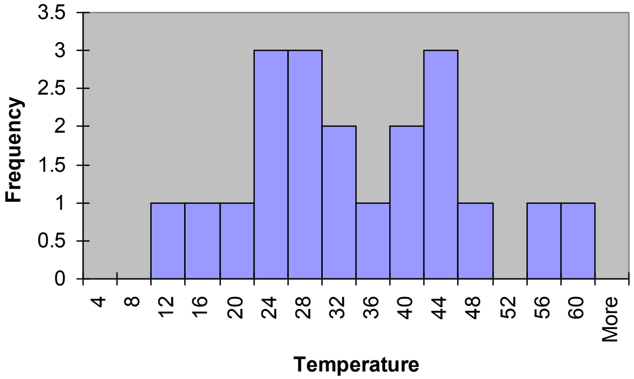
- Classes must be mutually exclusive
  - classes do not overlap
- Classes must be all-inclusive
  - a set of classes contains all possible data values
- Classes should be of equal width, if possible
  - the distance between the lowest and the highest possible values in each class is equal for all classes.
- Empty classes should be avoided

Developing Frequency Distribution for Continuous Data

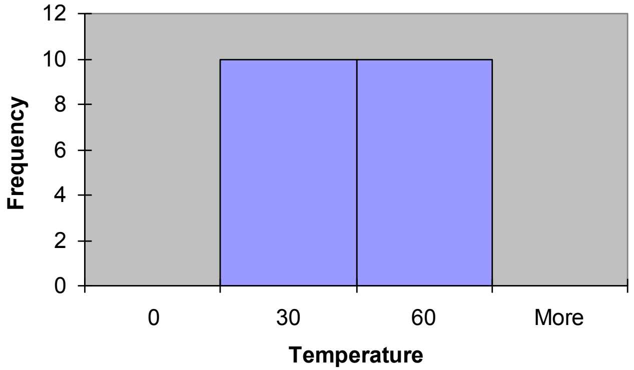
- Step 1: Determine the number of classes
- Step 2: Establish the class width
- Step 3: Determine the class boundaries for each class
  - the upper and lower values of each class
- Step 4: Determine the class frequency for each class
  - number of data points in each class

How Many Classes?

- **Many (Narrow class intervals)**
  - May yield a very jagged distribution with gaps from empty classes
  - Can give a poor indication of how frequency varies across classes



- **Few (Wide class intervals)**
  - May compress variation too much and yield a blocky distribution
  - Can obscure important patterns of variation



How Many Classes?

- Rule of thumb: between 5 and 20 classes
- $2^k \geq n$ rule,
  - where k is the number of classes and is defined to be the smallest integer so that $2^k \geq n$, where n is the number of data values
- The minimum class width:

$$\boxed{W = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}}$$

Example

- Sort raw data from low to high:

  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Find range: 58 - 12 = 46
- Select number of classes: 5 (usually between 5 and 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries: 10, 20, 30, 40, 50

  (Sometimes class midpoints are reported: 15, 25, 35, 45, 55)

- Count the number of values in each class

Example (continue)

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

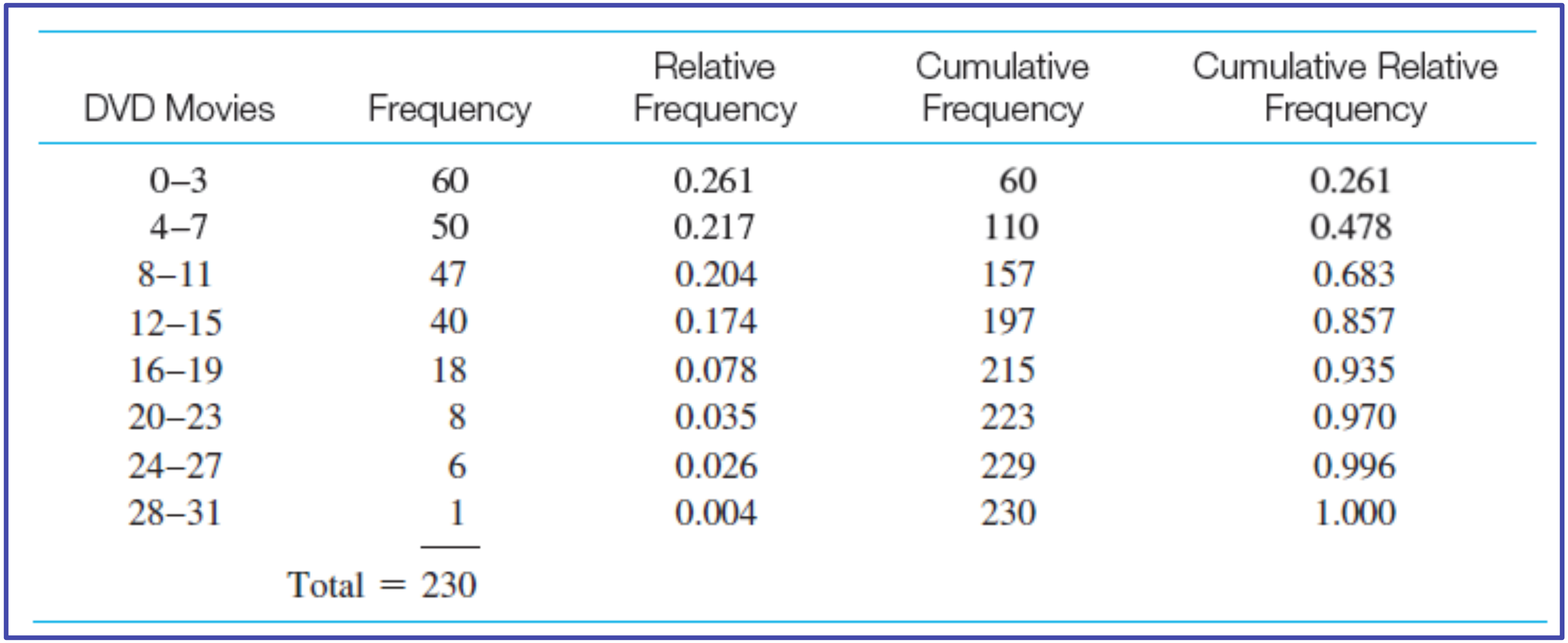
Frequency Distribution

Class	Frequency	Relative Frequency
10 but under 20	3	.15
20 but under 30	6	.30
30 but under 40	5	.25
40 but under 50	4	.20
50 but under 60	2	.10
Total	20	1.00

More on Frequency Distributions

- Cumulative Frequency Distribution
  - a summary of a set of data that displays the number of observations with values less than or equal to the upper limit of each of its classes
- Cumulative Relative Frequency Distribution
  - A summary of a set of data that displays the proportion of observations with values less than or equal to the upper limit of each of its classes

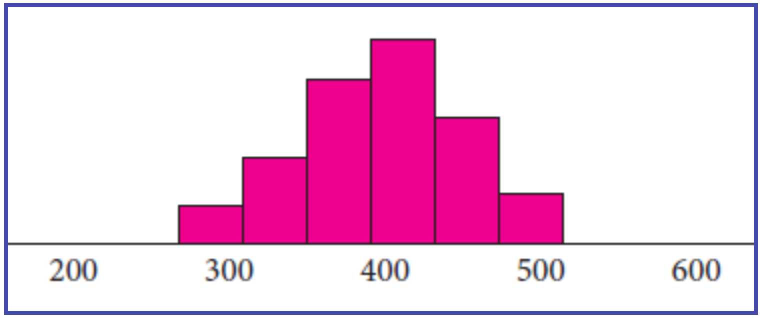
Example



DVD Movies	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0–3	60	0.261	60	0.261
4–7	50	0.217	110	0.478
8–11	47	0.204	157	0.683
12–15	40	0.174	197	0.857
16–19	18	0.078	215	0.935
20–23	8	0.035	223	0.970
24–27	6	0.026	229	0.996
28–31	1	0.004	230	1.000
	Total = 230			

Frequency Histograms

- A graph of a frequency distribution with the horizontal axis showing the classes, the vertical axis showing the frequency count, and (for equal class widths) the rectangles having a height equal to the frequency in each class.



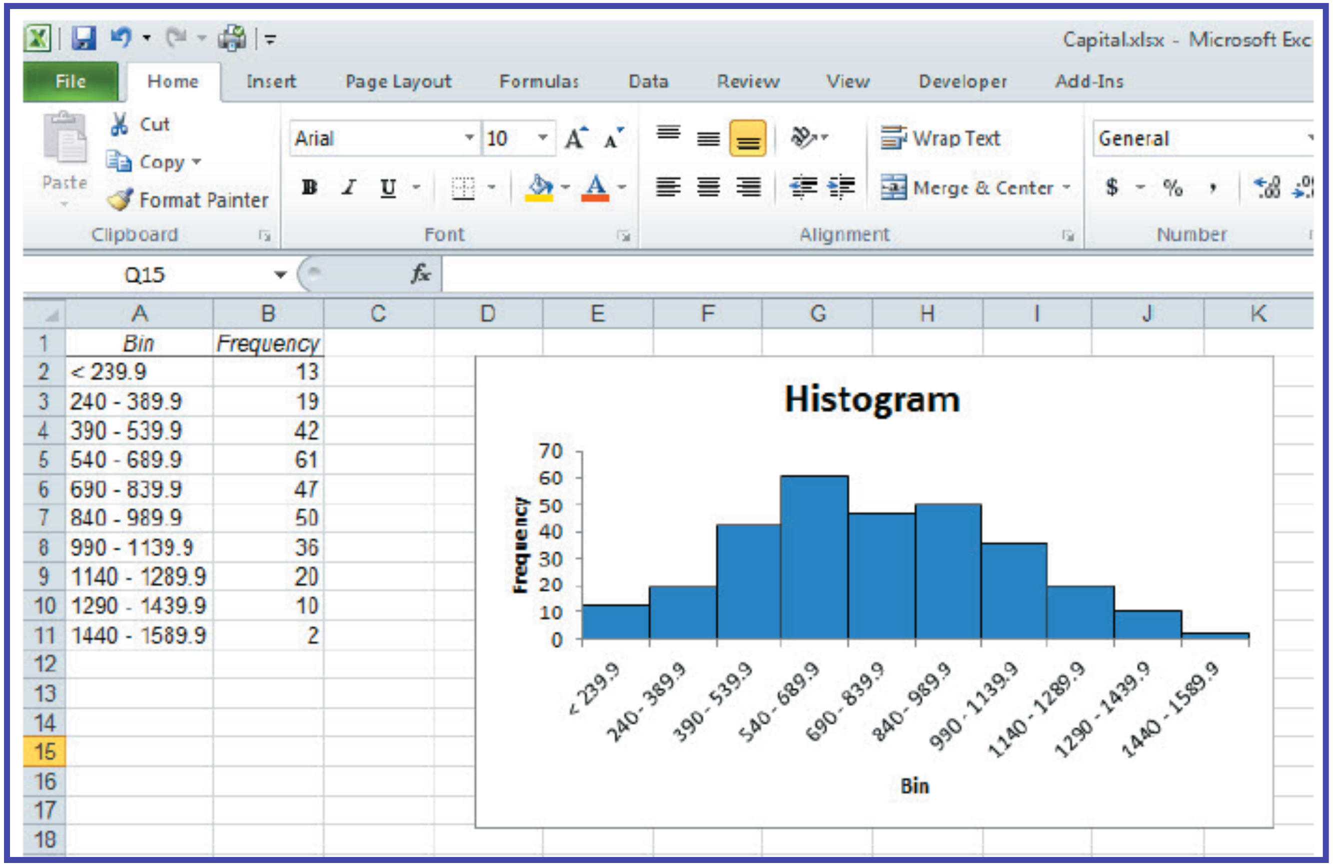
Constructing Frequency Histograms

- Steps 1-4: Construct a frequency distribution
- Step 5: Construct the axes for the histogram
- Step 6: Construct bars with heights corresponding to the frequency of each class
- Step 6: Label the histogram appropriately

How to Do It in Excel?

1. Open file.
2. Set up an area on the worksheet for the bins.
3. On the **Data** tab, click **Data Analysis**.
4. Select **Histogram**.
5. Input Range specifies the actual data values and the bin range as the area defined in 2.
6. Put on a new worksheet ply and include the Chart Output.
7. Right mouse click on the bars and use the **Format Data Series Options** to set gap width to zero and add lines to the bars.
8. Convert the bins to actual class labels by typing labels in Column A.

Note: The bin 239.99 is labeled < 239.99.



	A	B
1	*Bin*	*Frequency*
2	< 239.9	13
3	240 - 389.9	19
4	390 - 539.9	42
5	540 - 689.9	61
6	690 - 839.9	47
7	840 - 989.9	50
8	990 - 1139.9	36
9	1140 - 1289.9	20
10	1290 - 1439.9	10
11	1440 - 1589.9	2

Relative Frequency Histogram and Ogive

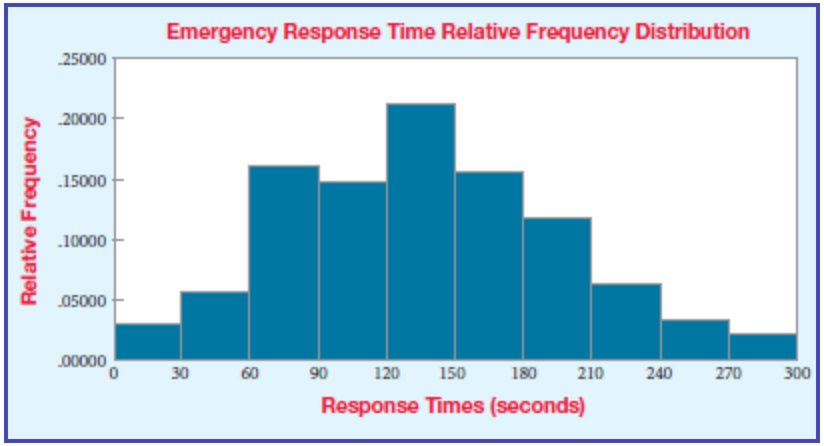
- Step 1: Convert the frequency distribution into relative frequencies and cumulative relative frequencies
- Step 2: Construct the relative frequency histogram
  - Place the quantitative variable on the horizontal axis and the relative frequencies on the vertical axis. The vertical bars are drawn to heights corresponding to the relative frequencies of the classes.

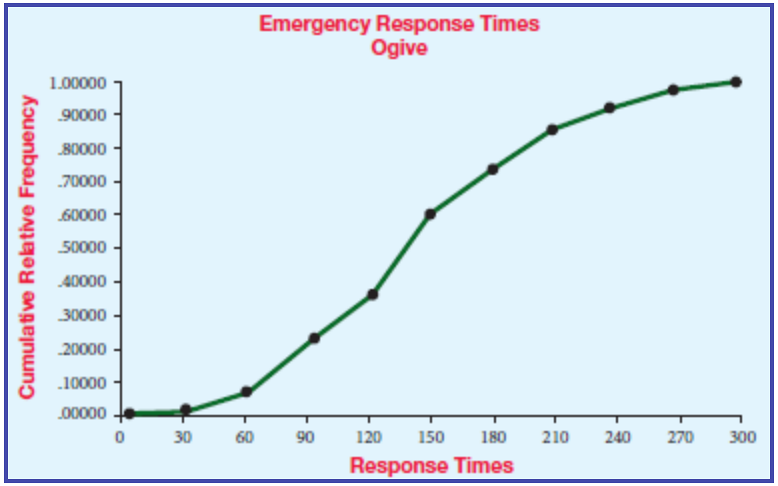
Relative Frequency Histogram and Ogive

- Step 3: Construct the ogive
- Ogive
  - The graphical representation of the cumulative relative frequency. A line is connected to points plotted above the upper limit of each class at a height corresponding to the cumulative relative frequency.

Example

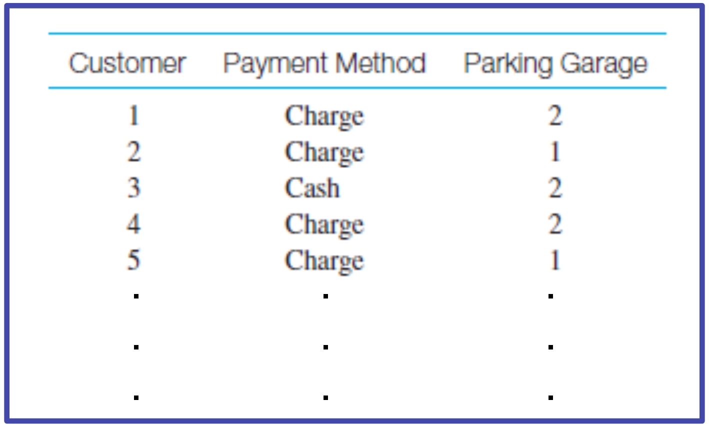
Response Time	Frequency	Relative Frequency	Cumulative Relative Frequency
0 and under 30	36	36/1220 = 0.0295	0.0295
30 and under 60	68	68/1220 = 0.0557	0.0852
60 and under 90	195	195/1220 = 0.1598	0.2451
90 and under 120	180	180/1220 = 0.1475	0.3926
120 and under 150	260	260/1220 = 0.2131	0.6057
150 and under 180	182	182/1220 = 0.1492	0.7549
180 and under 210	145	145/1220 = 0.1189	0.8738
210 and under 240	80	80/1220 = 0.0656	0.9393
240 and under 270	43	43/1220 = 0.0352	0.9746
270 and under 300	31	31/1220 = 0.0254	1.0000
	1,220	1.0000	





Joint Frequency Distribution

- Data are characterized by more than one variable
- Can be constructed for qualitative and quantitative variables
- Step 1: Obtain the data
  - – Example:

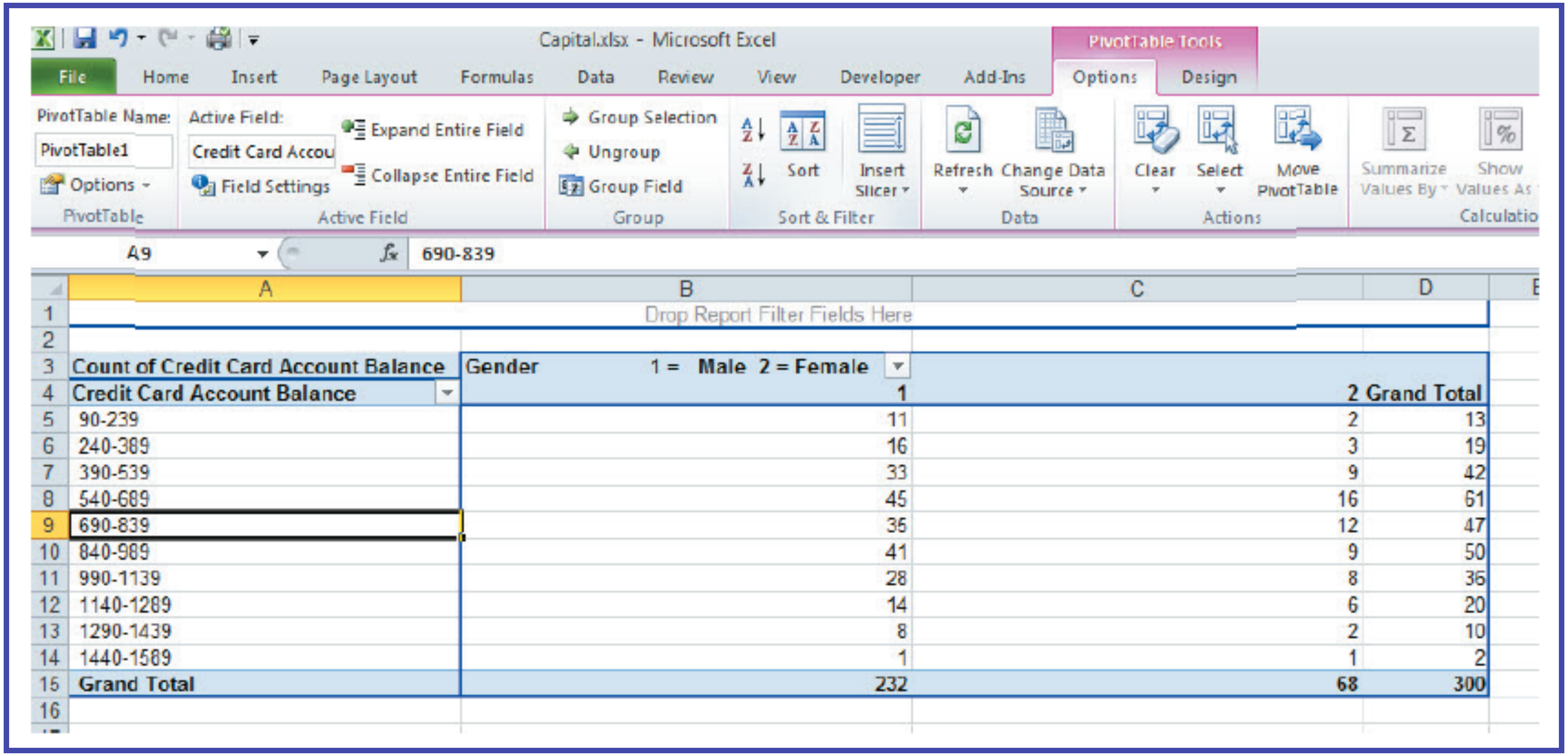


Customer	Payment Method	Parking Garage
1	Charge	2
2	Charge	1
3	Cash	2
4	Charge	2
5	Charge	1
.	.	.
.	.	.
.	.	.

Joint Frequency Distribution

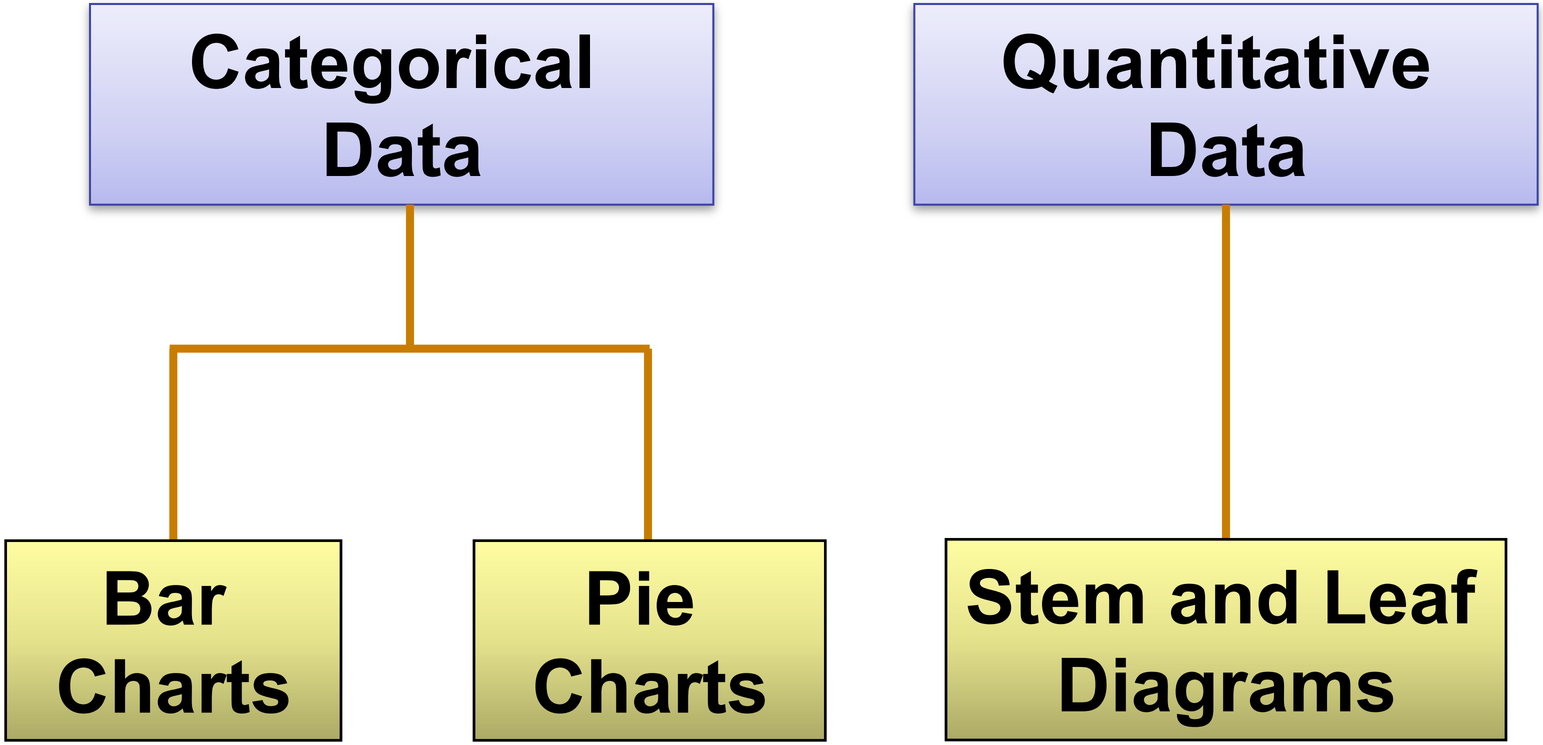
- Step 2: Construct the rows and columns of the joint frequency table
- Step 3: Count the number of joint occurrences at each row level and each column level for all combinations of row and column values and place these frequencies in the appropriate cells
- Step 4: Calculate the row and column totals

Cross-tabulation Table Example



Count of Credit Card Account Balance	Gender 1 = Male 2 = Female		
Credit Card Account Balance	1	2	Grand Total
90-239	11	2	13
240-389	16	3	19
390-539	33	9	42
540-689	45	16	61
690-839	35	12	47
840-989	41	9	50
990-1139	28	8	36
1140-1289	14	6	20
1290-1439	8	2	10
1440-1589	1	1	2
Grand Total	232	68	300

2.2 Bar Charts, Pie Charts, and Stem and Leaf Diagrams



Categorical Data

Quantitative Data

Bar Charts

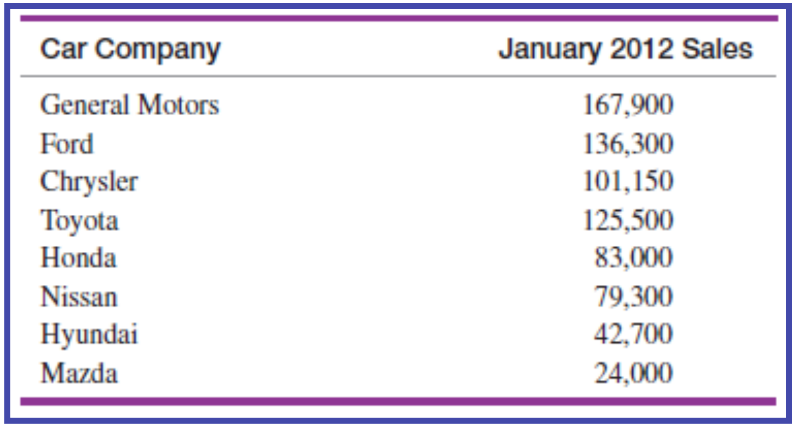
Pie Charts

Stem and Leaf Diagrams

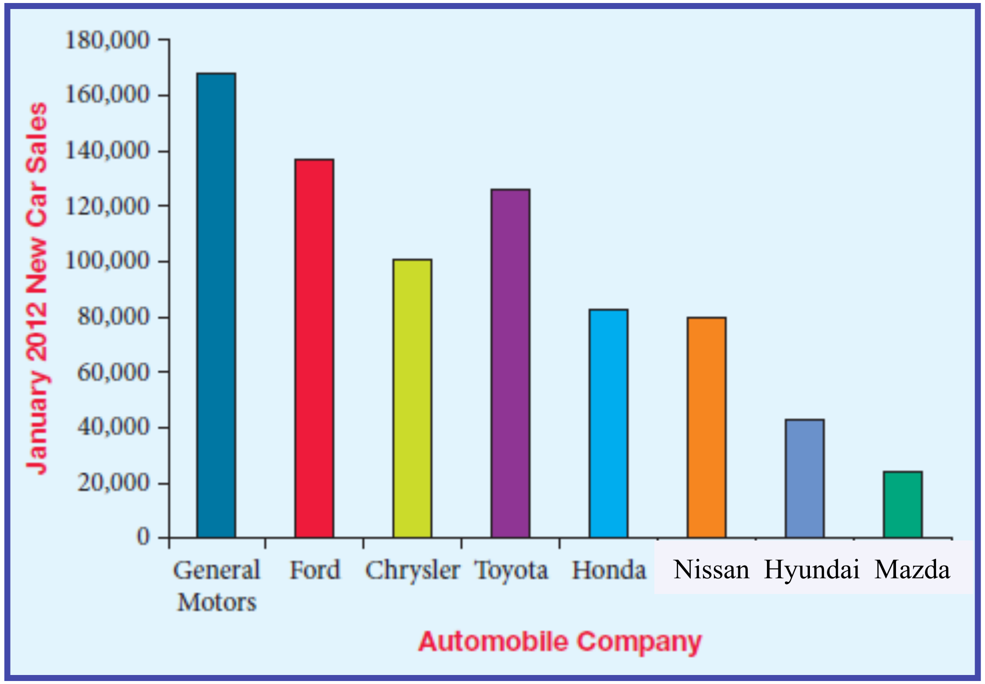
Bar Charts

- A graphical representation of a categorical data set in which a rectangle or bar is drawn over each category or class
- The length or height of each bar represents the frequency or percentage of observations or some other measure associated with the category
- The bars may be vertical or horizontal

Bar Chart Example



Car Company	January 2012 Sales
General Motors	167,900
Ford	136,300
Chrysler	101,150
Toyota	125,500
Honda	83,000
Nissan	79,300
Hyundai	42,700
Mazda	24,000



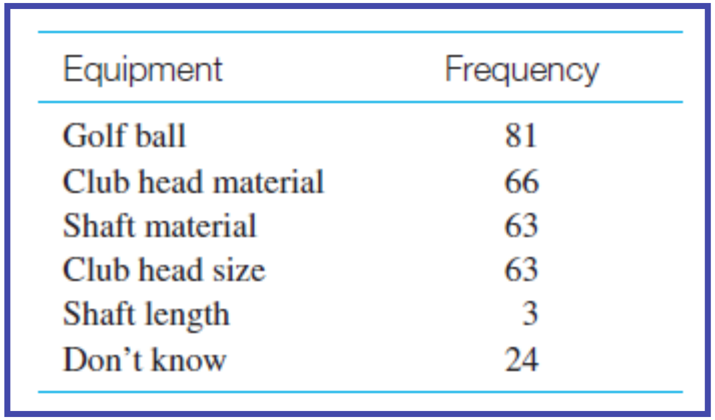
Constructing Bar Chart

- Step 1: Define the categories for the variable of interest
- Step 2: For each category, determine the appropriate measure or value
- Step 3: For a column bar chart, locate the categories on the horizontal axis. For a horizontal bar chart, place the categories on the vertical axis. Then construct bars, either vertical or horizontal, for each category such that the length or height corresponds to the value for the category.
- Step 4: Interpret the results

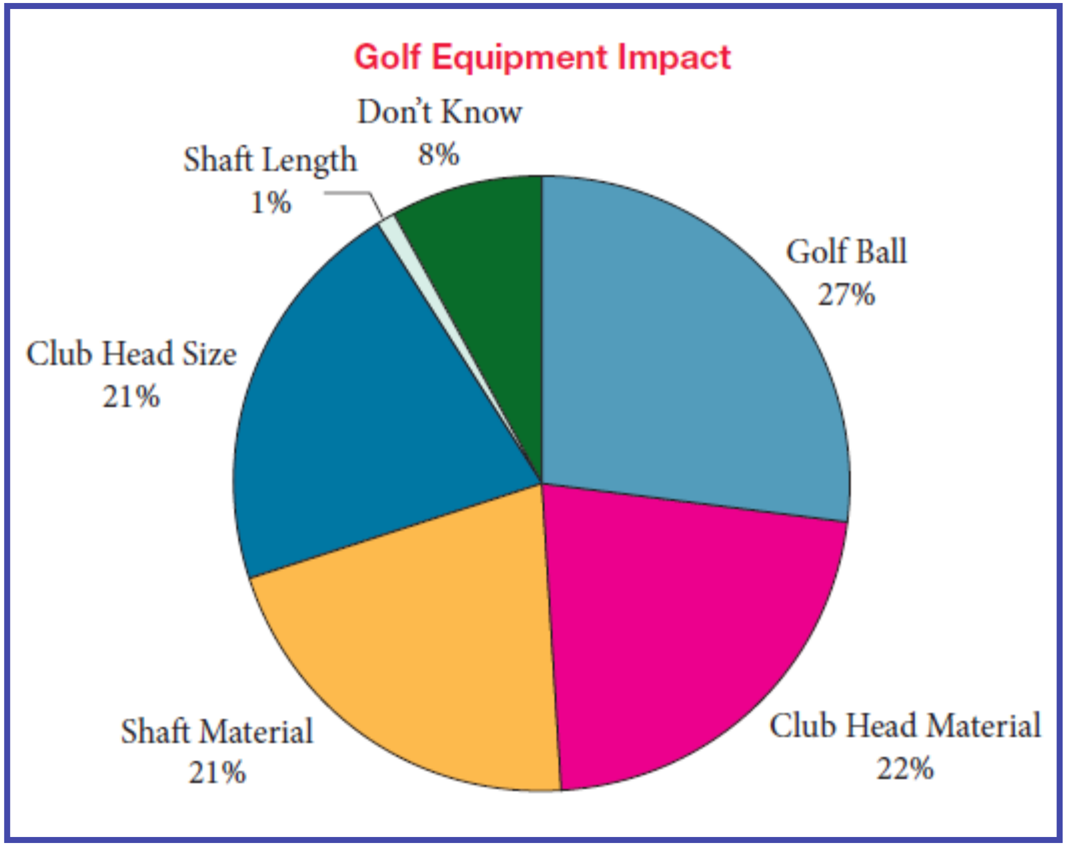
Pie Charts

- A graph in the shape of a circle.
- The circle is divided into “slices” corresponding to the categories or classes to be displayed.
- The size of each slice is proportional to the magnitude of the displayed variable associated with each category or class.

Pie Chart Example



Equipment	Frequency
Golf ball	81
Club head material	66
Shaft material	63
Club head size	63
Shaft length	3
Don't know	24



Constructing Pie Chart

- Step 1: Define the categories for the variable of interest.
- Step 2: For each category, determine the appropriate measure or value. The value assigned to each category is the proportion the category is to the total for all categories.
- Step 3: Construct the pie chart by displaying one slice for each category that is proportional in size to the proportion the category value is to the total of all categories.

Constructing Stem and Leaf Diagram

- Step 1: Sort the data from low to high.
- Step 2: Analyze the data for the variable of interest to determine how you wish to split the values into a stem and a leaf.
- Step 3: List all possible stems in a single column between the lowest and highest values in the data.
- Step 4: For each stem, list all leaves associated with the stem.

Stem and Leaf Diagram Example

Scores: 81 86 78 80 81 82 92 90
79 83 84 95 85 88 80 78
84 79 80 83 79 87 84 80

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

Step 1: The lowest value is 78, the highest – 95

Step 2: Stem is tens place, leaf is unit place

Step 3: List possible stems: 7, 8, and 9

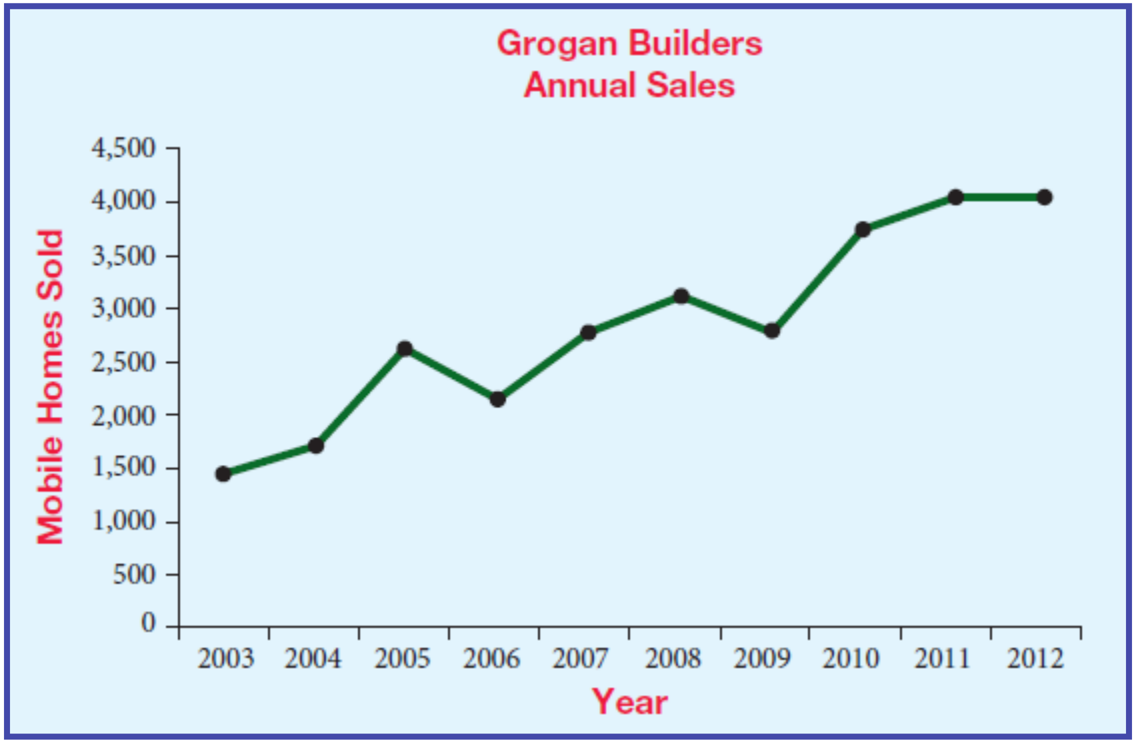
Step 4: Itemize the leaves from lowest to highest and place next to the appropriate stem

2.3 Line Charts and Scatter Diagrams

- Line Chart
  - A two-dimensional chart showing time on the horizontal axis and the variable of interest on the vertical axis
- Scatter Diagram
  - A two-dimensional graph of plotted points in which the vertical axis represents values of one quantitative variable and the horizontal axis represents values of the other quantitative variable

Line Chart Example

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1,426	1,678	2,591	2,105	2,744	3,068	2,755	3,689	4,003	3,997



Constructing Line Charts

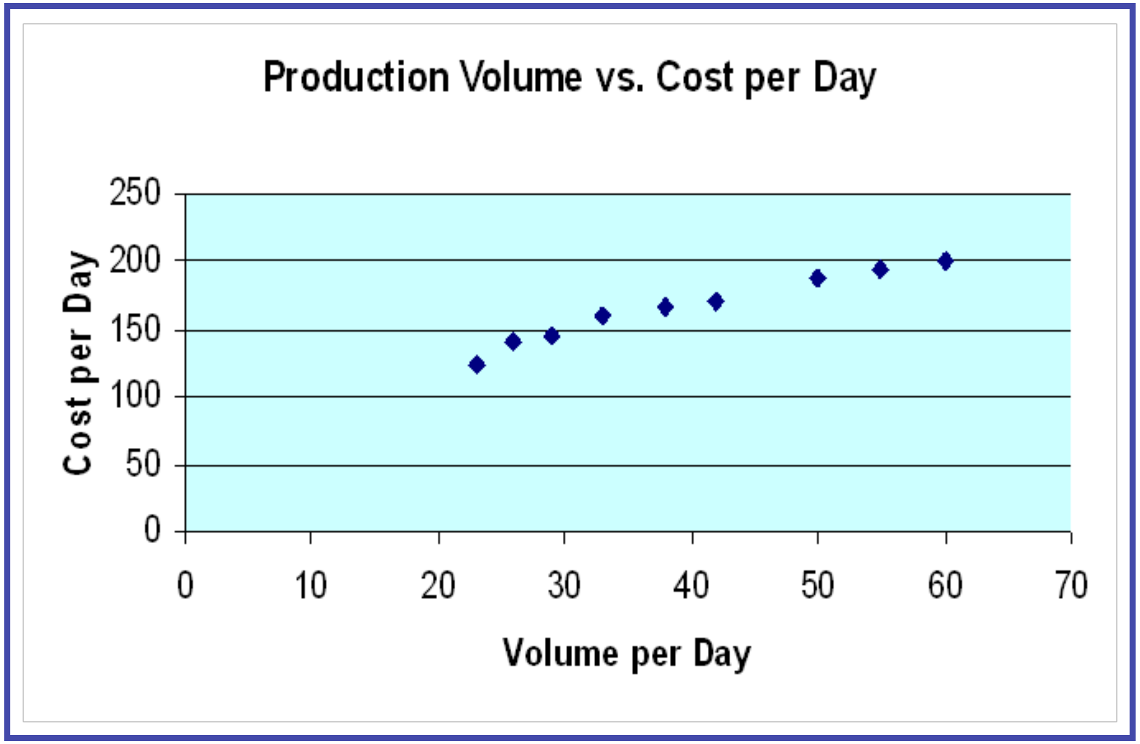
- Step 1: Identify the time-series variable of interest and determine the maximum value and the range of time periods covered in the data
- Step 2: Construct the horizontal axis for the time periods. Construct the vertical axis with a scale appropriate for the range of values
- Step 3: Plot the points of the graph and connect them with straight lines

Scatter Diagram

- Also called the scatter plot
- Shows the relationship between two quantitative variables
- Dependent Variable
  - Values are thought to be a function of another variable
- Independent Variable
  - Values are thought to impact the values of the dependent variable

Scatter Diagram Example

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Constructing Scatter Diagram

- Step 1: Identify the two quantitative variables and collect paired responses for the two variables.
- Step 2: Determine which variable will be placed on the vertical axis (*y*) and which variable will be placed on the horizontal axis (*x*)
- Step 3: Define the range of values for each variable and define the appropriate scale for the *x* and *y* axes
- Step 4: Plot the joint values for the two variables by placing a point in the *x,y* space