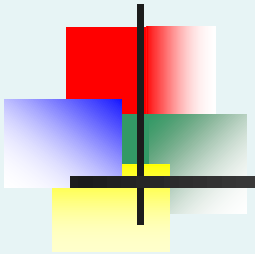
Statistics for Managers Using Microsoft Excel
7th Edition



Chapter 2

Organizing and Visualizing Data

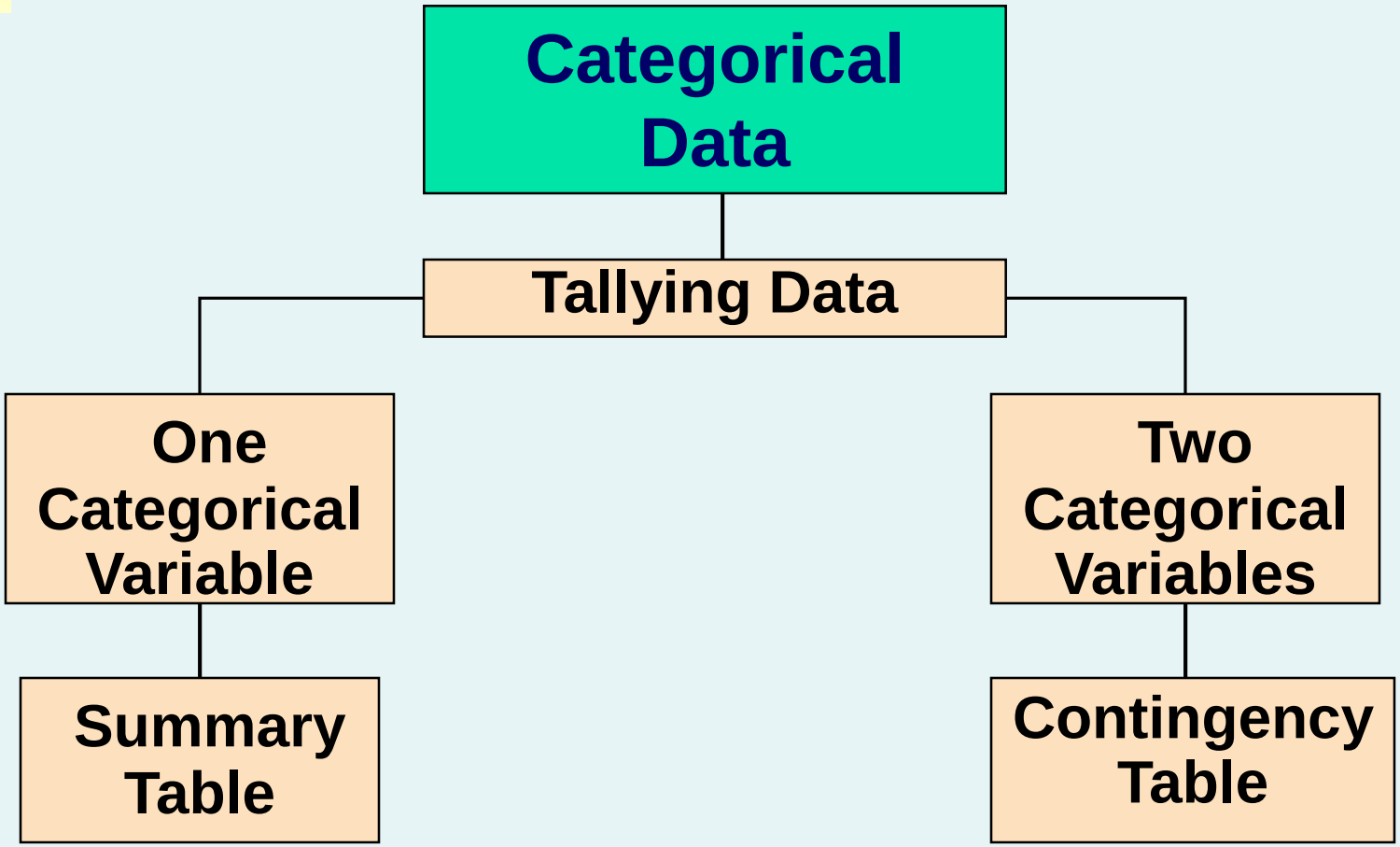
Learning Objectives

In this chapter you learn:

- To construct tables and charts for categorical data
- To construct tables and charts for numerical data
- The principles of properly presenting graphs
- To organize and analyze many variables

Categorical Data Are Organized By Utilizing Tables

DCOVA



Organizing Categorical Data: Summary Table

DCOVA

- A **summary table** tallies the frequencies or percentages of items in a set of categories so that you can see differences between categories.

Summary Table From A Survey of 1000 Banking Customers

Banking Preference?	Percent
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

A Contingency Table Helps Organize Two or More Categorical Variables

DCOVA

- Used to study patterns that may exist between the responses of two or more categorical variables
- Cross tabulates or tallies jointly the responses of the categorical variables
- For two variables the tallies for one variable are located in the rows and the tallies for the second variable are located in the columns

Contingency Table - Example

DCOVA

- A random sample of 400 invoices is drawn.
- Each invoice is categorized as a small, medium, or large amount.
- Each invoice is also examined to identify if there are any errors.
- This data are then organized in the contingency table to the right.

Contingency Table Showing Frequency of Invoices Categorized By Size and The Presence Of Errors

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

Contingency Table Based On Percentage Of Overall Total

DCOVA

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

42.50% = 170 / 400
25.00% = 100 / 400
16.25% = 65 / 400

	No Errors	Errors	Total
Small Amount	42.50%	5.00%	47.50%
Medium Amount	25.00%	10.00%	35.00%
Large Amount	16.25%	1.25%	17.50%
Total	83.75%	16.25%	100.0%

83.75% of sampled invoices have no errors and 47.50% of sampled invoices are for small amounts.

Contingency Table Based On Percentage of Row Totals

DCOVA

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

89.47% = 170 / 190
71.43% = 100 / 140
92.86% = 65 / 70

	No Errors	Errors	Total
Small Amount	89.47%	10.53%	100.0%
Medium Amount	71.43%	28.57%	100.0%
Large Amount	92.86%	7.14%	100.0%
Total	83.75%	16.25%	100.0%

Medium invoices have a larger chance (28.57%) of having errors than small (10.53%) or large (7.14%) invoices.

Contingency Table Based On Percentage Of Column Totals

DCOVA

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

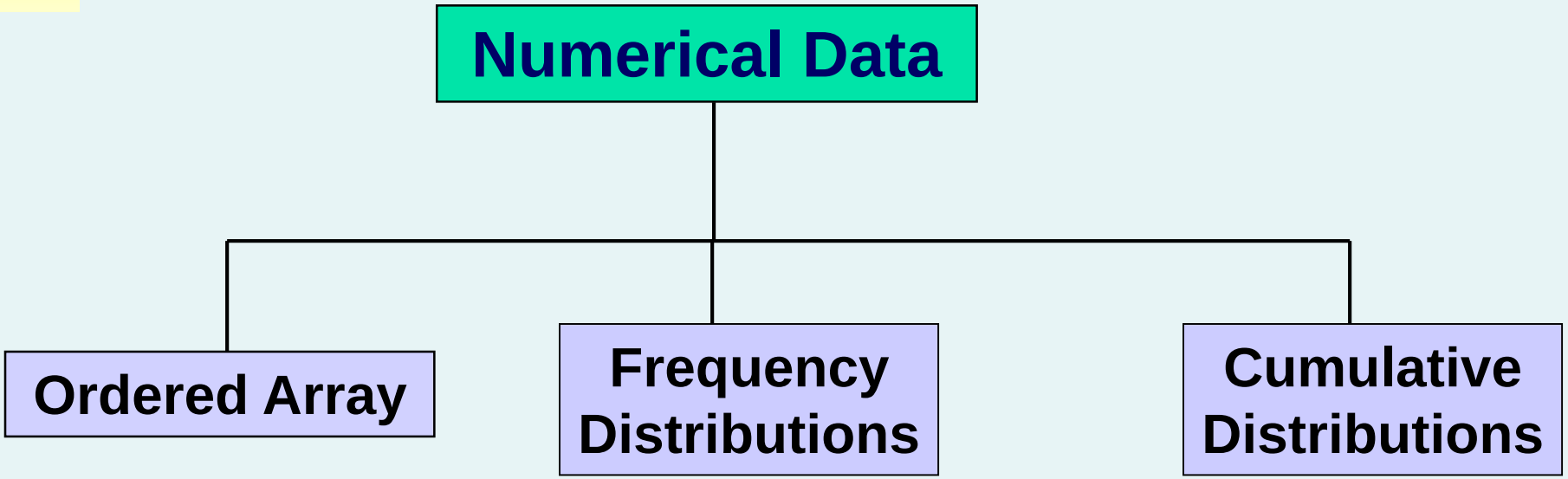
50.75% = 170 / 335
30.77% = 20 / 65

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

There is a 61.54% chance that invoices with errors are of medium size.

Tables Used For Organizing Numerical Data

DCOVA



Stacked Or Unstacked Format

- This is an issue when you have a categorical variable that may be used group your numerical variable for analysis.
- Stacked format is when your numerical variable is in one column and a second column identifies the value of the categorical variable.
- Unstacked format is when the values of the numerical variable in each group (unique value of the categorical variable) are in different columns.

Example of Stacked & Unstacked Format

Different Programs & different analyses may require a specific format

Stacked Format		Unstacked Format	
Age Of Students	Day or Night Student	Age Of Day Students	Age Of Night Students
16	D	16	18
19	D	19	23
22	D	22	18
18	N	17	28
23	N	19	19
17	D	25	32
19	D	17	19
25	D	20	33
18	N	27	
28	N	18	
17	D	20	
20	D	32	
27	D		
19	N		
32	N		
18	D		
20	D		
32	D		
19	N		
33	N		

Organizing Numerical Data: Ordered Array

DCOVA

- An **ordered array** is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value)
- May help identify outliers (unusual observations)

Age of Surveyed College Students	**Day Students**					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Organizing Numerical Data: Frequency Distribution

DCOVA

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.

- You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.

- The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.

- To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

Organizing Numerical Data: Frequency Distribution Example

DCOVA

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

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

Organizing Numerical Data: Frequency Distribution Example

DCOVA

- Sort raw data in ascending order:
  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 15)**
- Compute class interval (width): **10 (46/5 then round up)**
- Determine class boundaries (limits):
  - **Class 1: 10 to less than 20**
  - **Class 2: 20 to less than 30**
  - **Class 3: 30 to less than 40**
  - **Class 4: 40 to less than 50**
  - **Class 5: 50 to less than 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

Organizing Numerical Data: Frequency Distribution Example

DCOVA

Data in ordered array:

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

Class	Midpoints	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2
Total		20

Organizing Numerical Data: Relative & Percent Frequency Distribution Example

DCOVA

Data in ordered array:

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

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15%
20 but less than 30	6	.30	30%
30 but less than 40	5	.25	25%
40 but less than 50	4	.20	20%
50 but less than 60	2	.10	10%
Total	20	1.00	100%

Organizing Numerical Data: Cumulative Frequency Distribution Example

DCOVA

Data in ordered array:

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

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15%	3	15%
20 but less than 30	6	30%	9	45%
30 but less than 40	5	25%	14	70%
40 but less than 50	4	20%	18	90%
50 but less than 60	2	10%	20	100%
Total	20	100	20	100%

Why Use a Frequency Distribution?

DCOVA

- It condenses the raw data into a more useful form
- It allows for a quick visual interpretation of the data
- It enables the determination of the major characteristics of the data set including where the data are concentrated / clustered

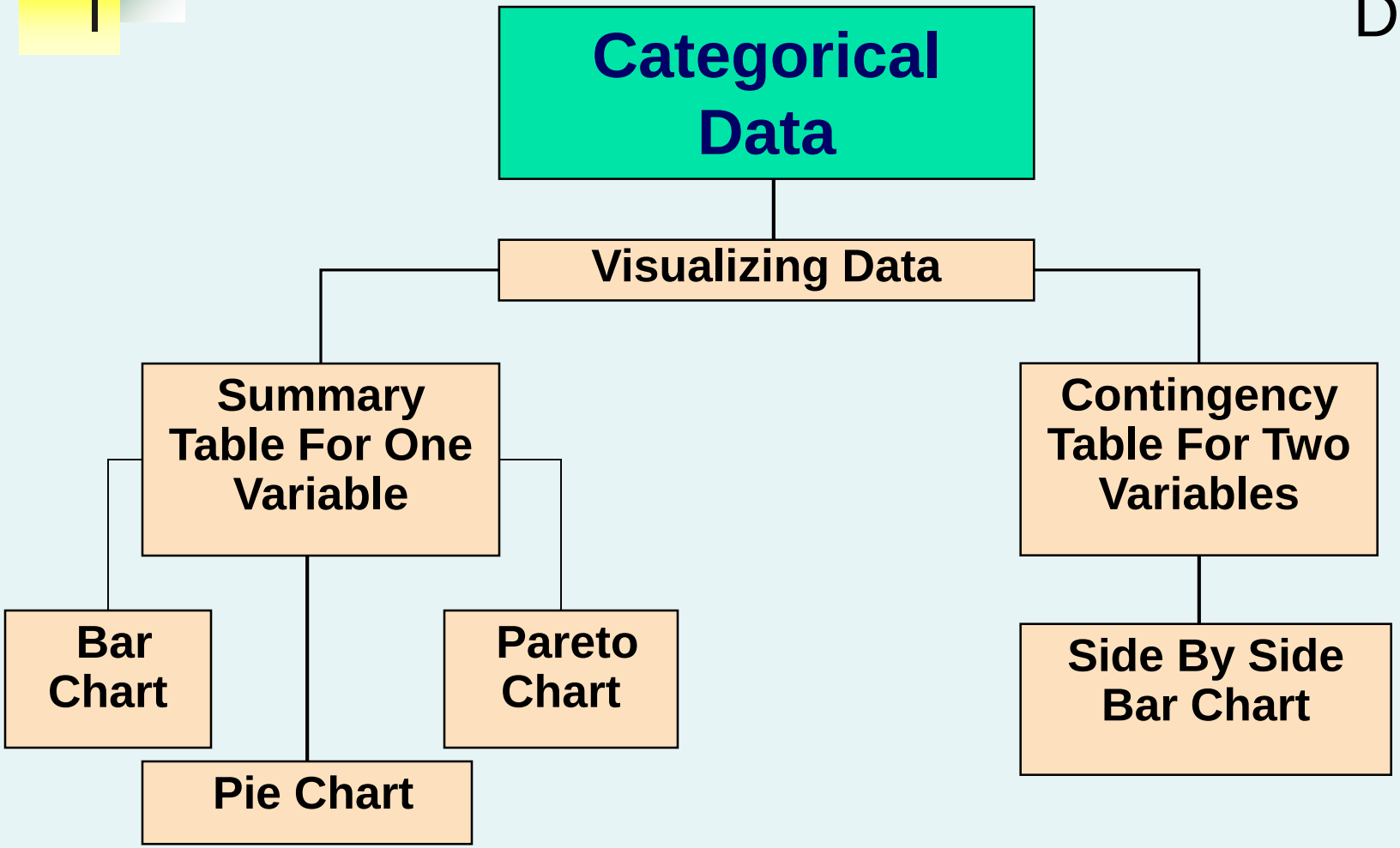
Frequency Distributions: Some Tips

DCOVA

- Different class boundaries may provide different pictures for the same data (especially for smaller data sets)
- Shifts in data concentration may show up when different class boundaries are chosen
- As the size of the data set increases, the impact of alterations in the selection of class boundaries is greatly reduced
- When comparing two or more groups with different sample sizes, you must use either a relative frequency or a percentage distribution

Visualizing Categorical Data Through Graphical Displays

DCOVA

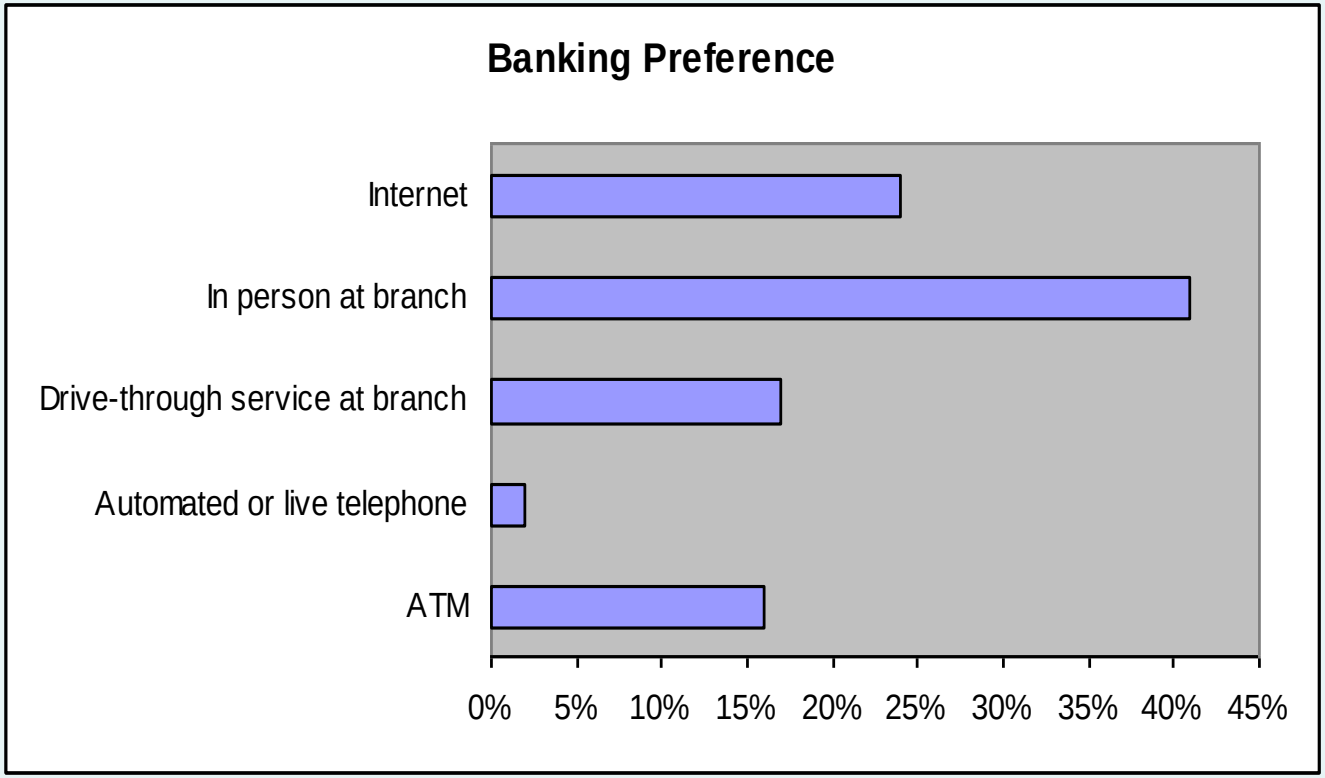


Visualizing Categorical Data: The Bar Chart

DCOVA

- In a **bar chart,** a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category which come from the summary table of the variable.

Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

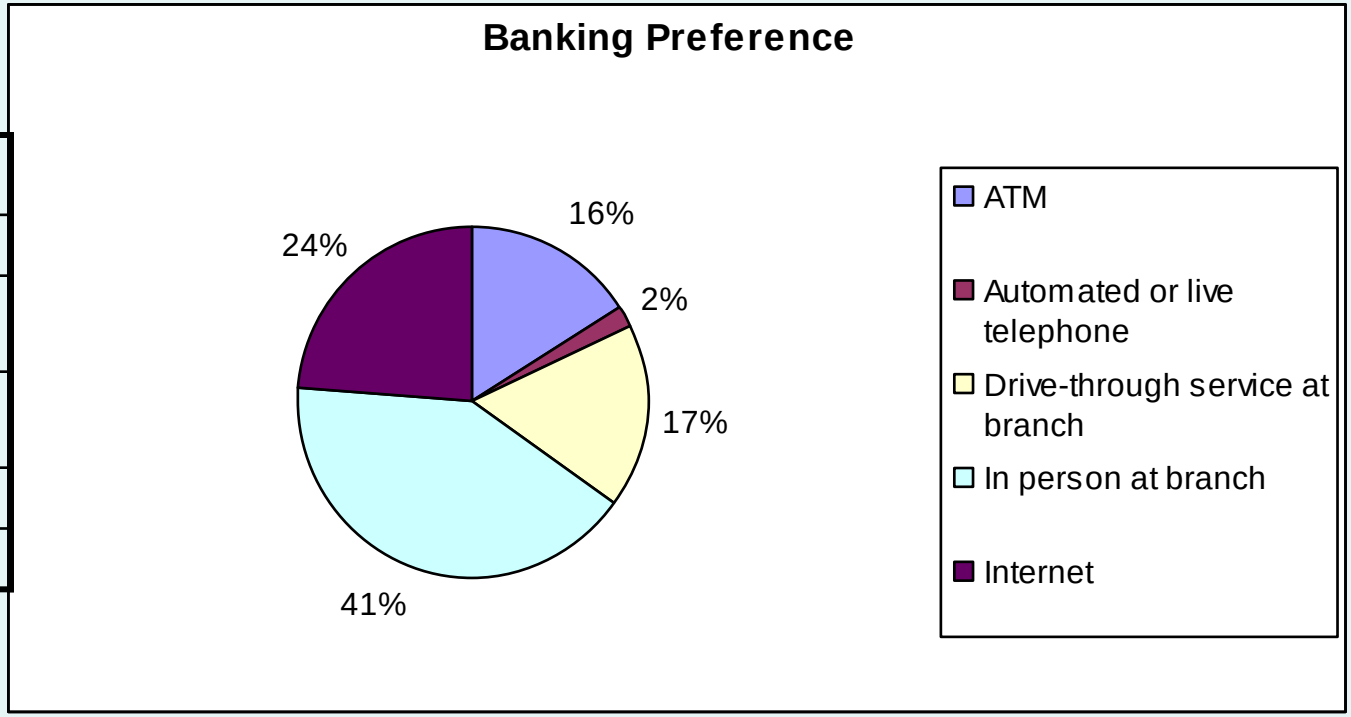


Visualizing Categorical Data: The Pie Chart

DCOVA

- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%



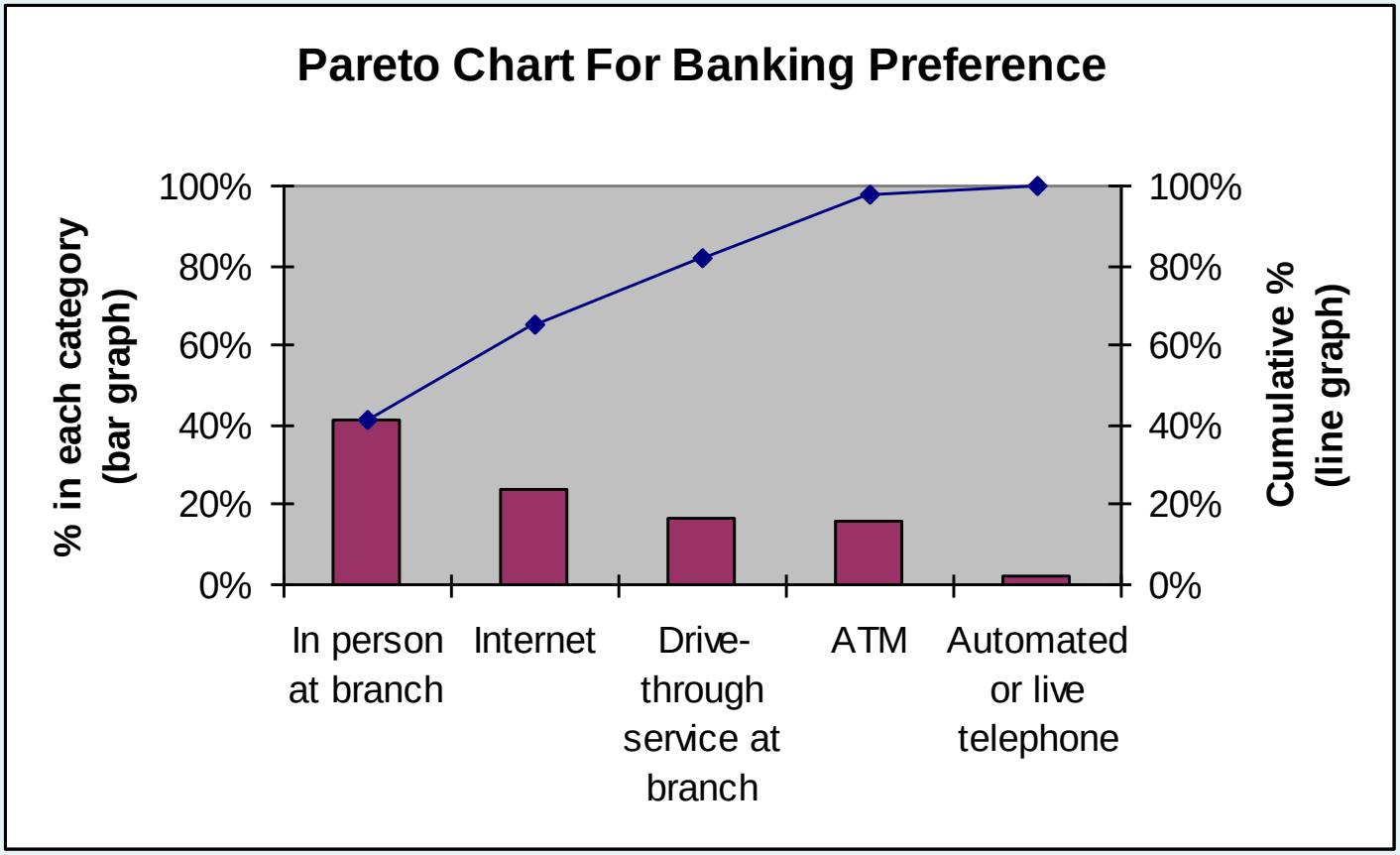
Visualizing Categorical Data: The Pareto Chart

DCOVA

- Used to portray categorical data (nominal scale)
- A vertical bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is shown in the same graph
- Used to separate the “vital few” from the “trivial many”

Visualizing Categorical Data: The Pareto Chart (con't)

DCOVA

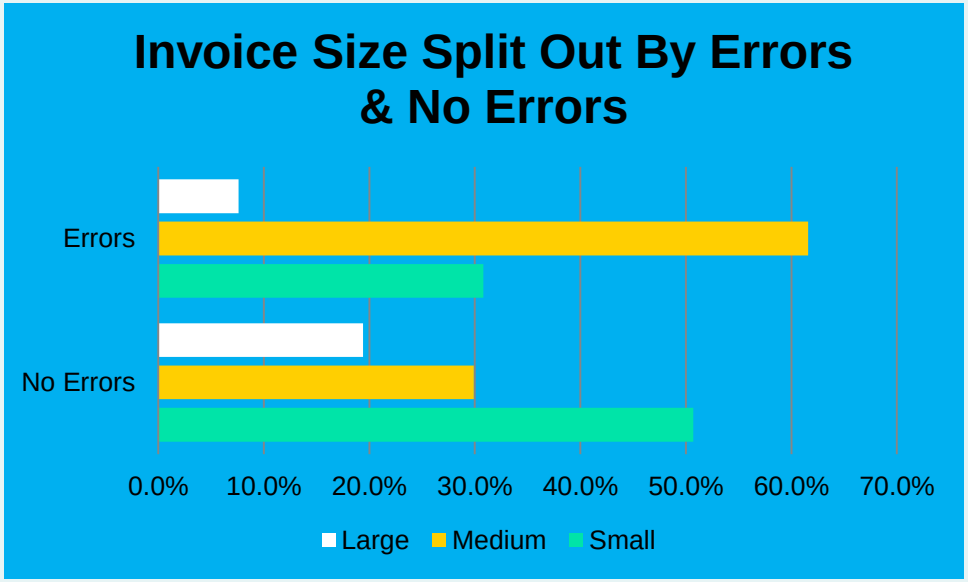


Visualizing Categorical Data: Side By Side Bar Charts

DCOVA

- The **side by side bar chart** represents the data from a contingency table.

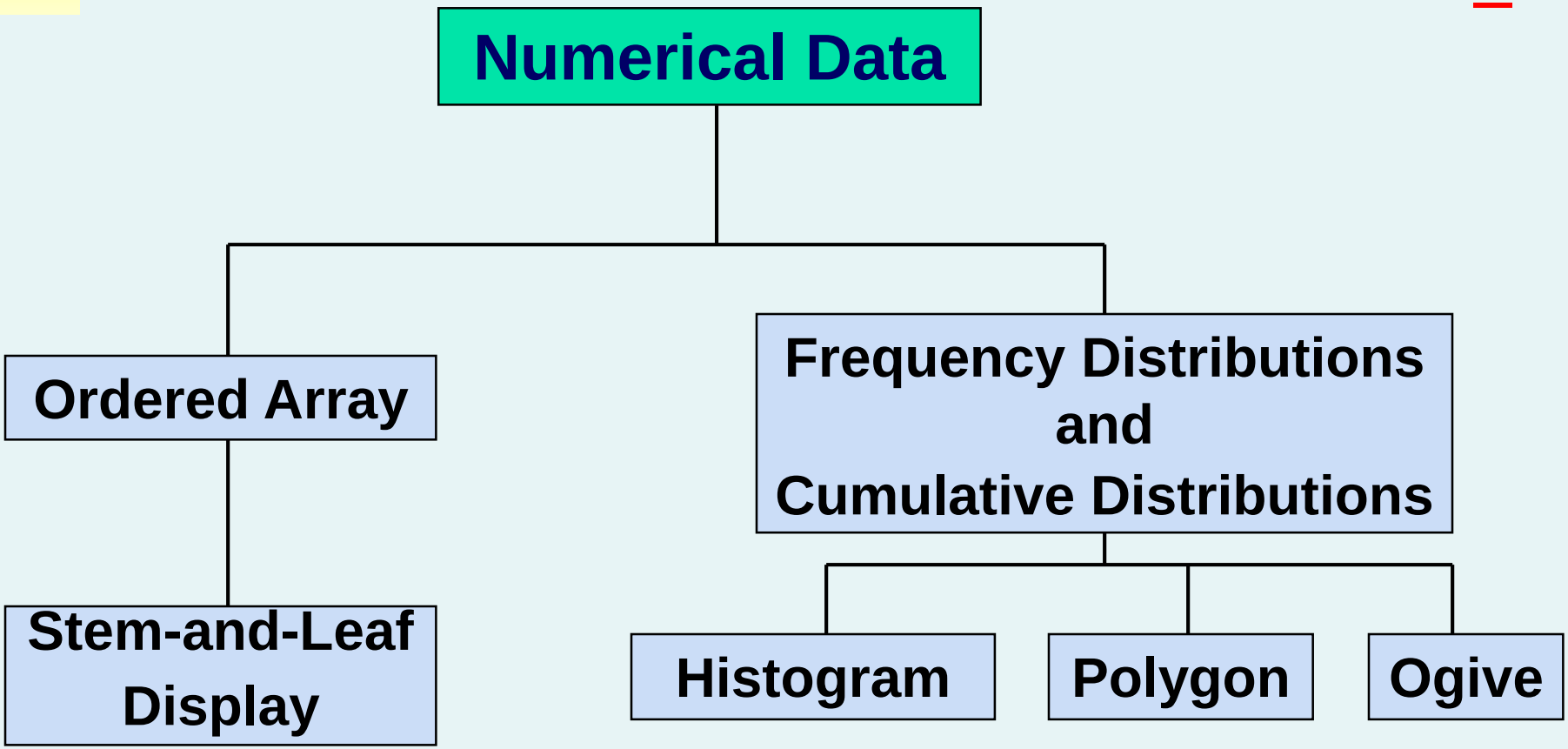
	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%



Invoices with errors are much more likely to be of medium size (61.54% vs 30.77% and 7.69%)

Visualizing Numerical Data By Using Graphical Displays

DCOVA



Stem-and-Leaf Display

DCOVA

- A simple way to see how the data are distributed and where concentrations of data exist

METHOD: Separate the sorted data series into leading digits (the **stems**) and the trailing digits (the **leaves**)

Organizing Numerical Data: Stem and Leaf Display

DCOVA

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of Surveyed College Students						
	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Age of College Students

Day Students

Stem	Leaf
1	67788899
2	0012257
3	28
4	2

Night Students

Stem	Leaf
1	8899
2	0138
3	23
4	15

Visualizing Numerical Data: The Histogram

DCOVA

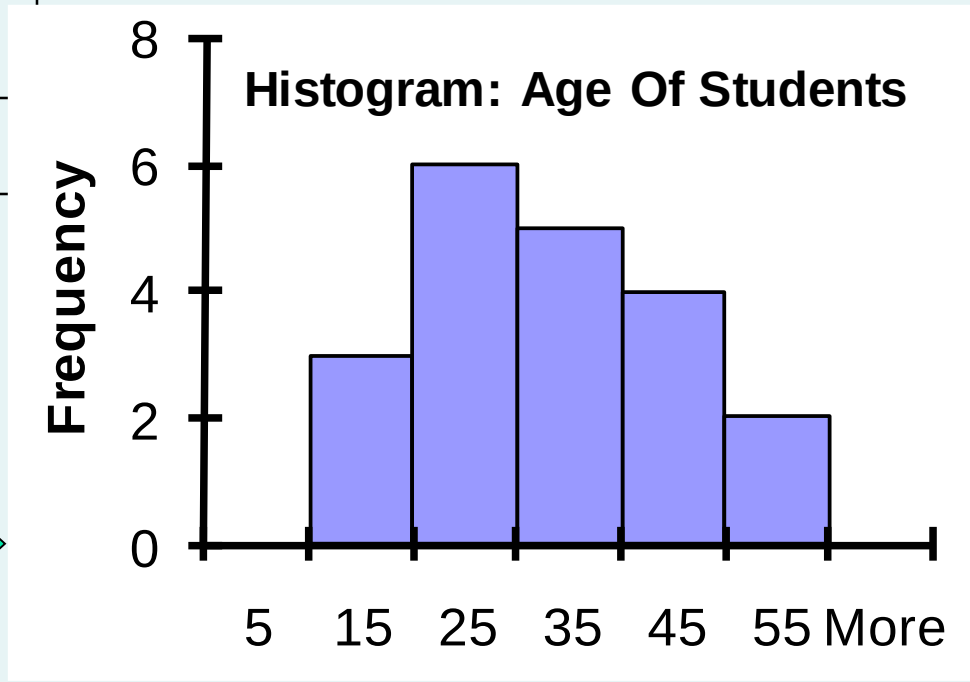
- A vertical bar chart of the data in a frequency distribution is called a **histogram.**
- In a histogram there are no gaps between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency, relative frequency,** or **percentage**.
- The height of the bars represent the frequency, relative frequency, or percentage.

Visualizing Numerical Data: The Histogram

DCOVA

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



Visualizing Numerical Data: The Polygon

DCOVA

- A **percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.

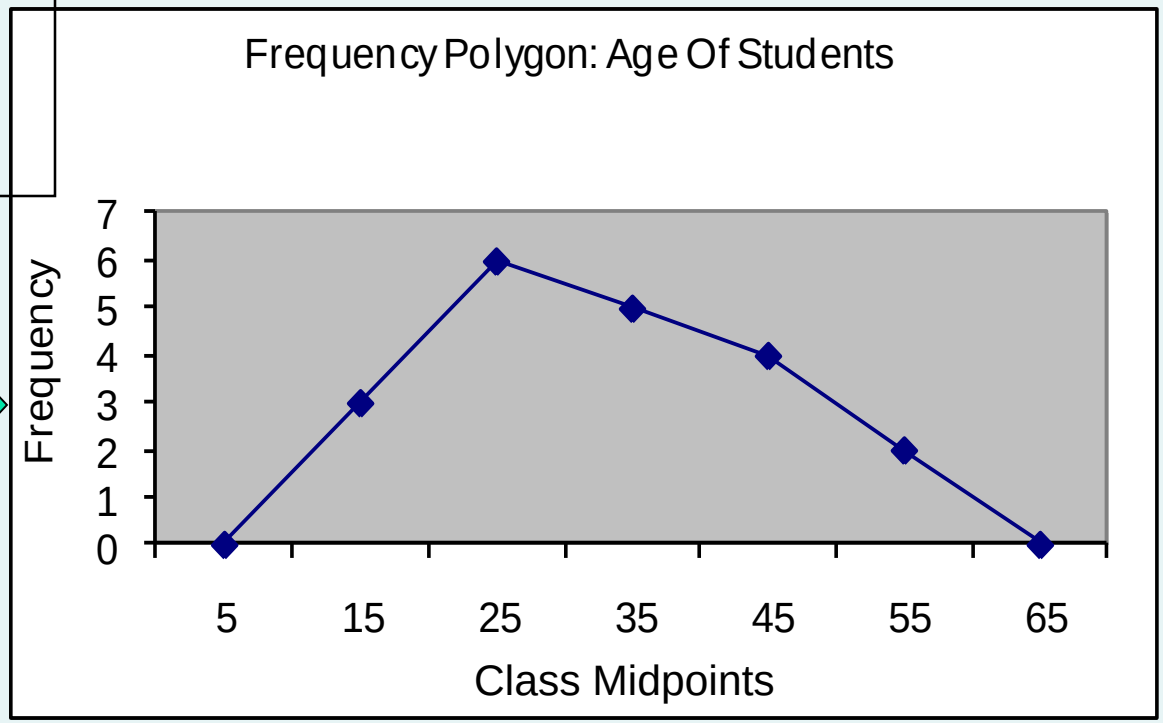
- The **cumulative percentage polygon,** or **ogive,** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

- Useful when there are two or more groups to compare.

Visualizing Numerical Data: The Frequency Polygon

DCOVA

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2



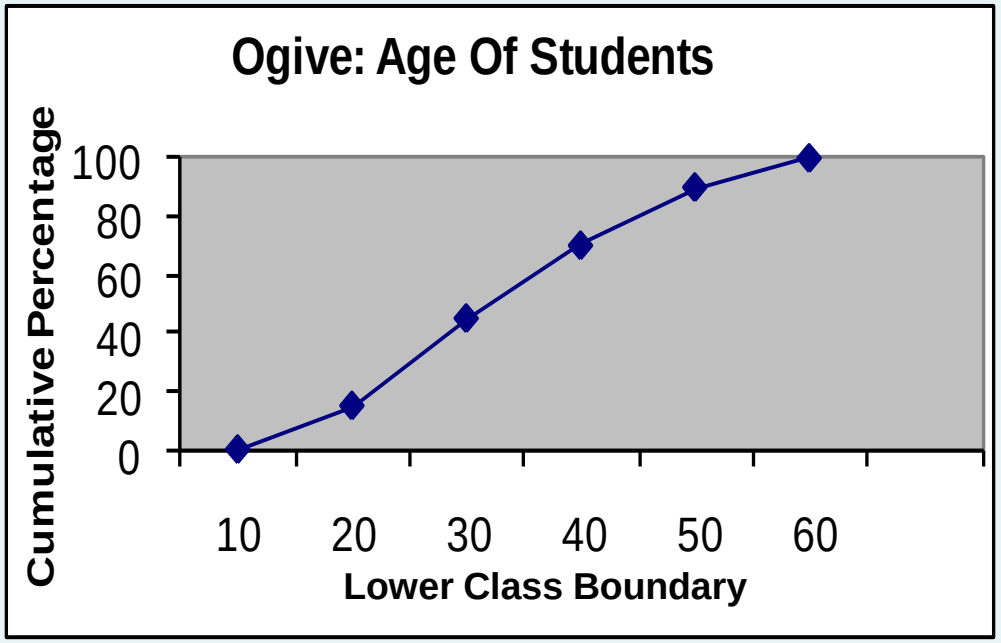
(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)

Visualizing Numerical Data: The Ogive (Cumulative % Polygon)

DCOVA

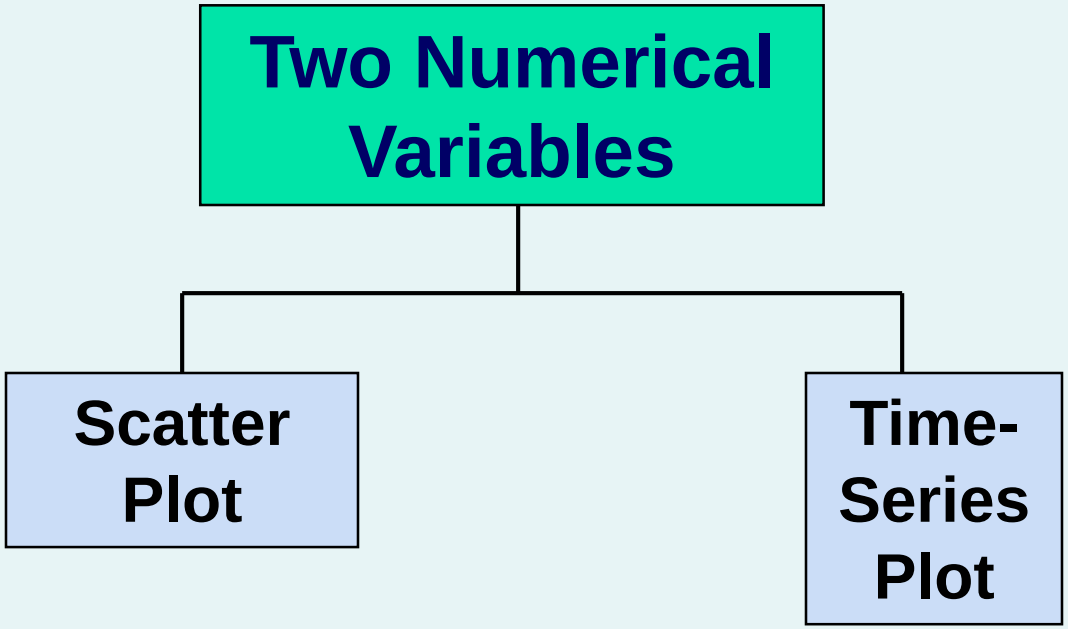
Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100

(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.



Visualizing Two Numerical Variables By Using Graphical Displays

DCOVA



Visualizing Two Numerical Variables: The Scatter Plot

DCOVA

- **Scatter plots** are used for numerical data consisting of paired observations taken from two numerical variables

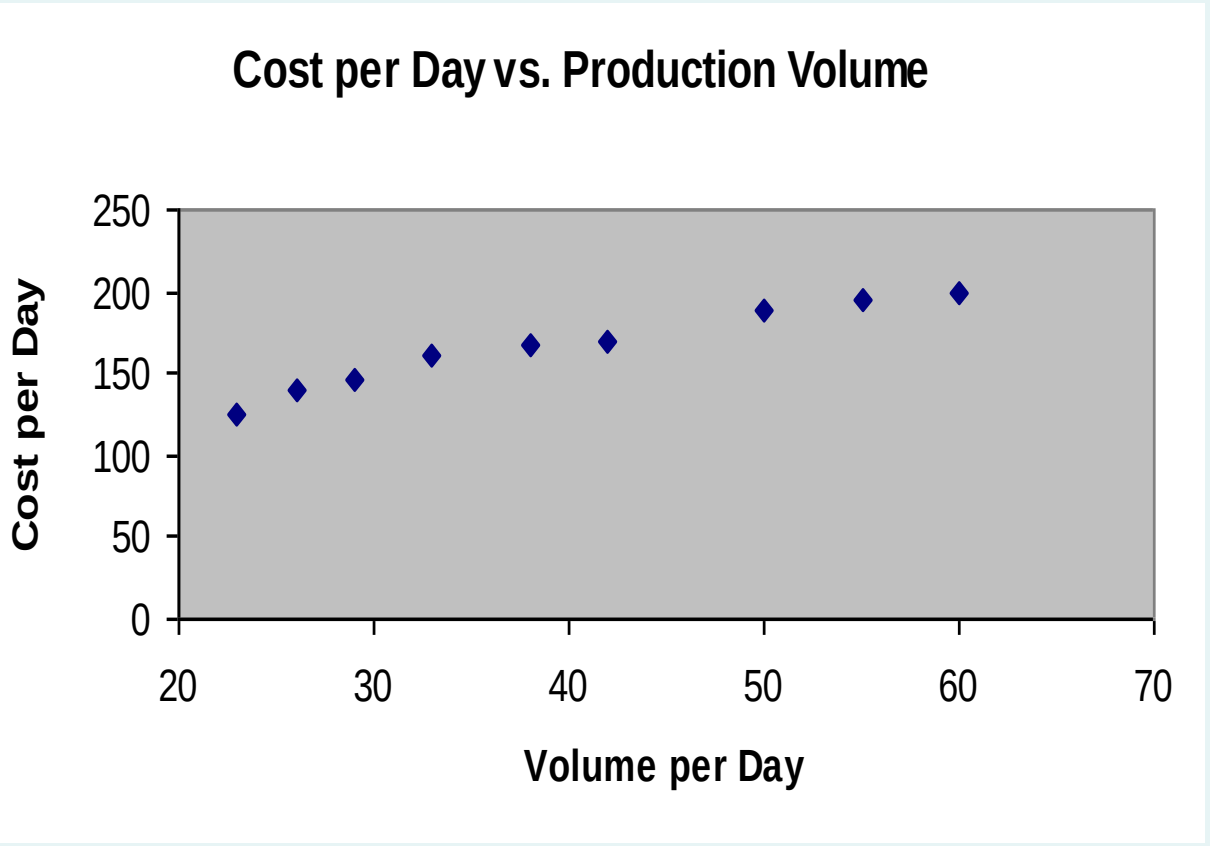
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis

- Scatter plots are used to examine possible relationships between two numerical variables

Scatter Plot Example

DCOVA

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



Visualizing Two Numerical Variables: The Time Series Plot

DCOVA

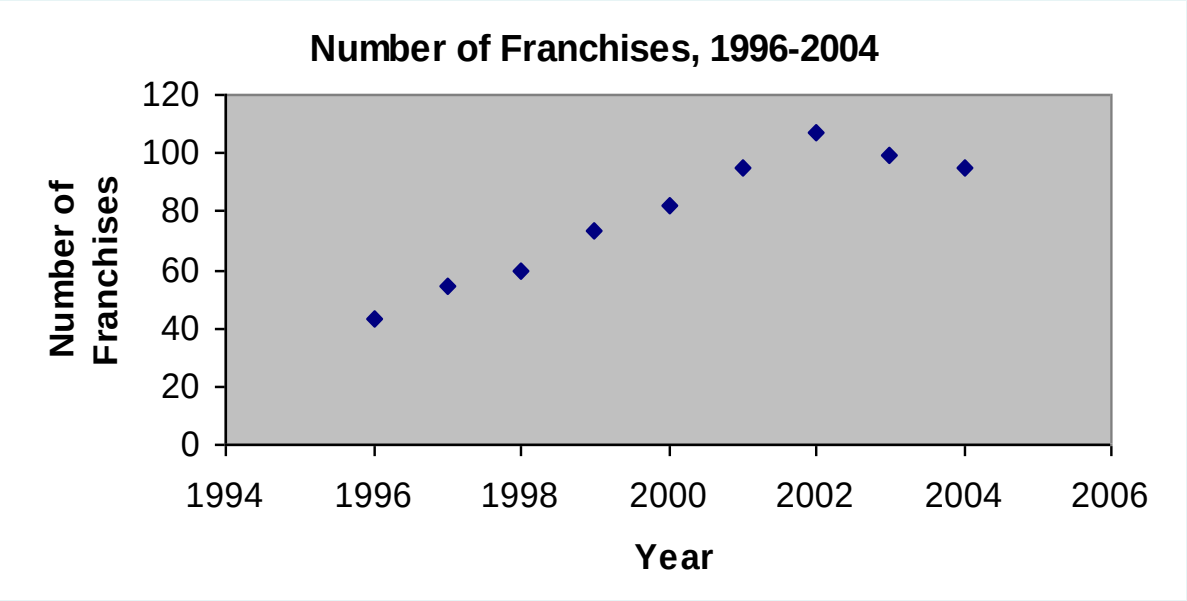
- A Time-Series Plot is used to study patterns in the values of a numeric variable over time

- The Time-Series Plot:
  - Numeric variable is measured on the vertical axis and the time period is measured on the horizontal axis

Time Series Plot Example

DCOVA

Year	Number of Franchises
1996	43
1997	54
1998	60
1999	73
2000	82
2001	95
2002	107
2003	99
2004	95



Guidelines For Developing Visualizations

DCOVA

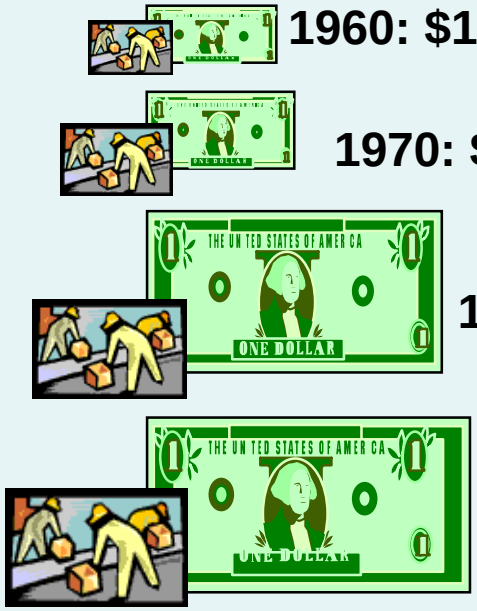
- Avoid chartjunk
- Use the simplest possible visualization
- Include a title
- Label all axes
- Include a scale for each axis if the chart contains axes
- Begin the scale for a vertical axis at zero
- Use a constant scale

Graphical Errors: Chart Junk

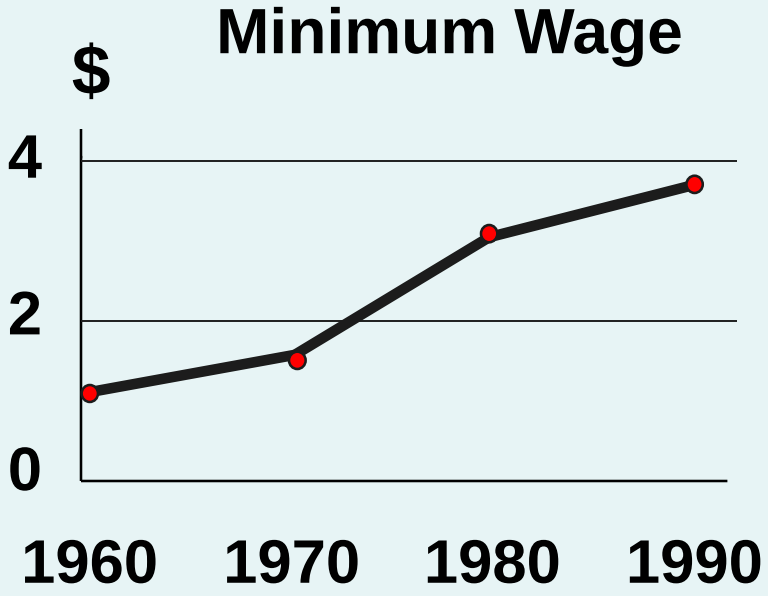
DCOVA

Bad Presentation

Minimum Wage



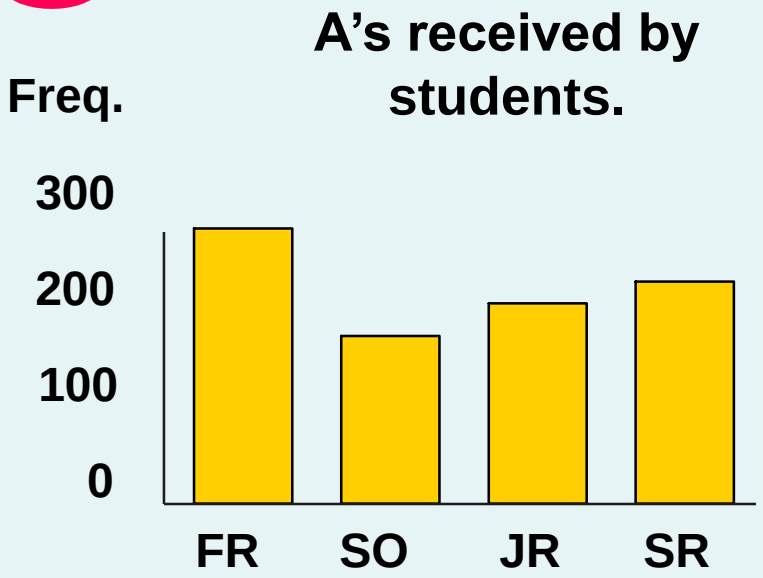
Good Presentation



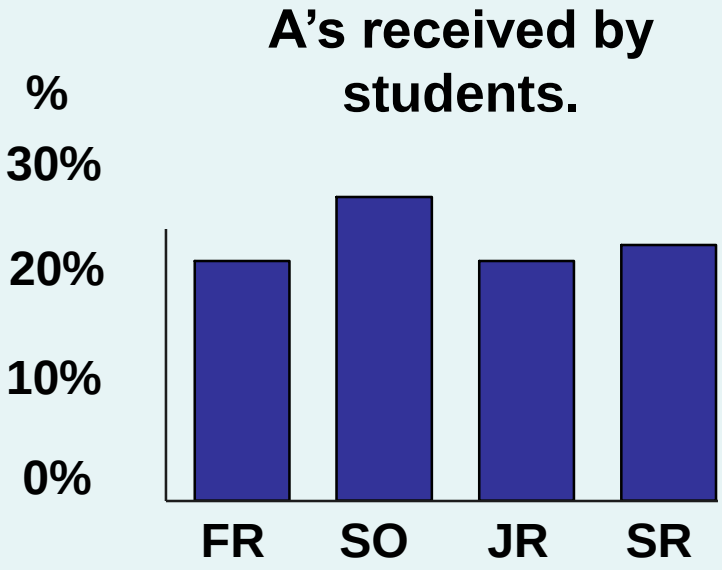
Graphical Errors: No Relative Basis

DCOVA

Bad Presentation



Good Presentation

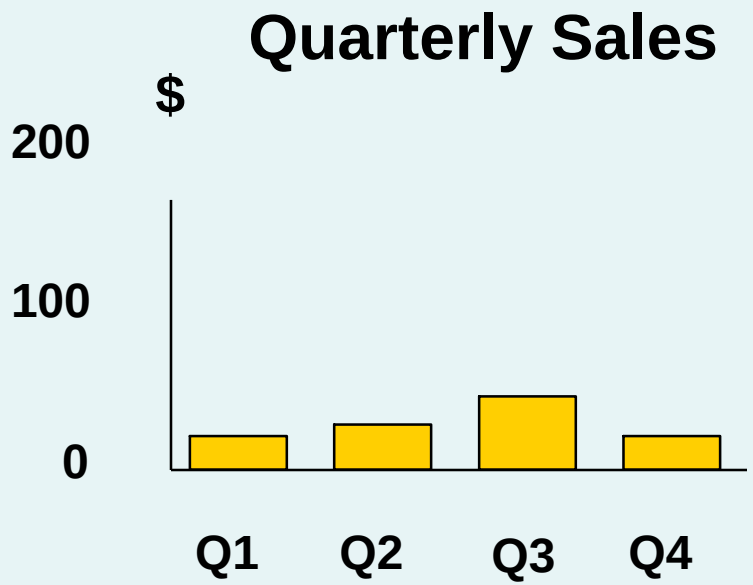


FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Graphical Errors: Compressing the Vertical Axis

DCOVA

Bad Presentation

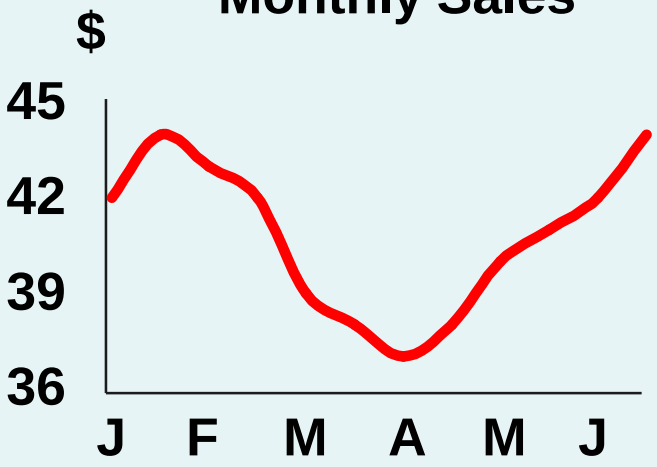


Good Presentation

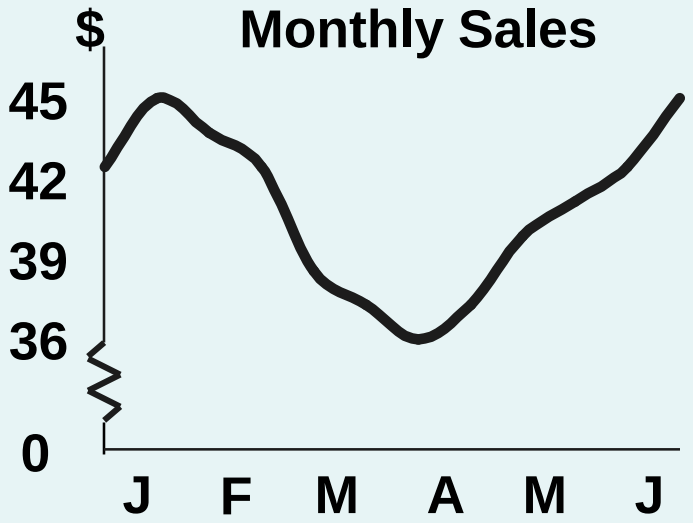
Graphical Errors: No Zero Point on the Vertical Axis

DCOVA

Bad Presentation



Good Presentations



Graphing the first six months of sales

In Excel It Is Easy To Inadvertently Create Distortions

- Excel often will create a graph where the vertical axis does not start at 0
- Excel offers the opportunity to turn simple charts into 3-D charts and in the process can create distorted images
- Unusual charts offered as choices by excel will most often create distorted images

Using Excel Pivot Tables To Organize & Visualize Many Variables

DCOVA

A pivot table:

- Summarizes variables as a multidimensional summary table
- Allows interactive changing of the level of summarization and formatting of the variables
- Allows you to interactively “slice” your data to summarize subsets of data that meet specified criteria
- Can be used to discover possible patterns and relationships in multidimensional data that simpler tables and charts would fail to make apparent.

A Two Variable Contingency Table For The Retirement Funds Data

DCOVA

There are many more growth funds of average risk than of low or high risk

	A	B	C	D	E
1	Contingency Table of Fund Type and Risk				
2					
3		RISK			
4	TYPE	Low	Average	High	Grand Total
5	Growth	19.50%	35.53%	15.09%	70.13%
6	Value	11.64%	10.06%	8.18%	29.87%
7	Grand Total	31.13%	45.60%	23.27%	100.00%

A Multidimensional Contingency Table Tallies Responses Of Three or More Categorical Variables

DCOVA

	A	B	C	D	E
1	Contingency Table of Fund Type, Market Cap, and Risk				
2					
3		RISK			
4	TYPE	Low	Average	High	Grand Total
5	Growth	19.50%	35.53%	15.09%	70.13%
6	Large	15.09%	14.78%	2.52%	32.39%
7	Mid-Cap	3.77%	13.84%	3.14%	20.75%
8	Small	0.63%	6.92%	9.43%	16.98%
9	Value	11.64%	10.06%	8.18%	29.87%
10	Large	9.43%	7.86%	0.00%	17.30%
11	Mid-Cap	1.57%	1.57%	2.83%	5.97%
12	Small	0.63%	0.63%	5.35%	6.60%
13	Grand Total	31.13%	45.60%	23.27%	100.00%

- Growth funds risk pattern depends on market
- Value funds risk risk pattern is different from that of growth funds.

Multidimensional Contingency Tables Can Include Numerical Variables

DCOVA

This table displays average 10-year return with the market cap collapsed or hidden from view

	A	B	C	D	E
1	Contingency Table of Fund Type, Market Cap, and Risk				
2					
3	Average of 10YrReturn%	RISK			
4	TYPE	Low	Average	High	Grand Total
5	Growth	4.12	5.07	4.72	4.73
6	Value	5.14	4.71	6.87	5.47
7	Grand Total	4.50	4.99	5.48	4.95

Value funds with low or high risk have a higher average 10 year return than growth funds with those risk levels

The Same Table With Market Cap Expanded Shows A More Complicated Pattern

DCOVA

	A	B	C	D	E
1	Contingency Table of Fund Type, Market Cap, and Risk				
2					
3	Average of 10YrReturn%	RISK			
4	TYPE	Low	Average	High	Grand Total
5	Growth	4.12	5.07	4.72	4.73
6	Large	3.69	3.65	1.26	3.48
7	Mid-Cap	5.62	6.04	5.77	5.92
8	Small	5.38	6.15	5.30	5.65
9	Value	5.14	4.71	6.87	5.47
10	Large	4.52	4.13		4.34
11	Mid-Cap	6.62	6.27	5.52	6.01
12	Small	10.77	8.12	7.58	7.94
13	Grand Total	4.50	4.99	5.48	4.95

Growth funds with large market capitalizations are the poorest performers and depress the average for growth fund category

Double-clicking A Cell Drills Down & Displays The Underlying Data

DCOVA

Double-clicking in the cell where the joint response "value fund and high risk" is tallied creates a new worksheet where the details for all the funds that meet this criteria are displayed

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fund Number	Market Cap	Type	Assets	Turnover Ratio	Beta	SD	Risk	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%	Expense Ratio	Star Rating
2	RF318	Small	Value	83.6	124	0.85	23.62	High	-3.77	19.06	2.16	9.13	1.6	Three
3	RF316	Small	Value	9	131	0.85	25.14	High	-1.34	20.13	-0.94	6.45	1.87	Two
4	RF315	Small	Value	22.3	127	0.68	24.86	High	4.63	20.9	-0.92	4.44	1.96	Five
5	RF314	Small	Value	1698.9	123	0.95	23.68	High	5.64	21.74	1.35	8.43	1.26	Three
24	RF228	Mid-Cap	Value	5926.3	95	1.38	25.91	High	-3.26	25.33	-1.41	6.41	0.6	Five
25	RF227	Mid-Cap	Value	1352.3	38	1.44	28.42	High	0.57	29.83	4.82	10.09	1.29	Four
26	RF226	Mid-Cap	Value	28	381	1.57	32.05	High	0.44	30.04	-2.87	2.03	1.54	Five
27	RF225	Mid-Cap	Value	18.3	26	1.46	28.97	High	0.81	35.01	2.73	4.46	2.07	Three

Pivot Tables, Slicers & Business Analytics

DCOVA

- Many analytics processes start with many variables and let you explore the data by use of filtering
- In Excel, using slicers is one way to mimic this filtering operation
- Slicers can be used to filter any variable that is associated with a Pivot Table
- By clicking buttons in slicer panels you can subset and filter data and visually see answers to questions

Chapter Summary

In this chapter we have:

- Constructed tables and charts for categorical data
- Constructed tables and charts for numerical data
- Examined the principles of properly presenting graphs
- Examined methods to organize and analyze many variables in Excel

This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.