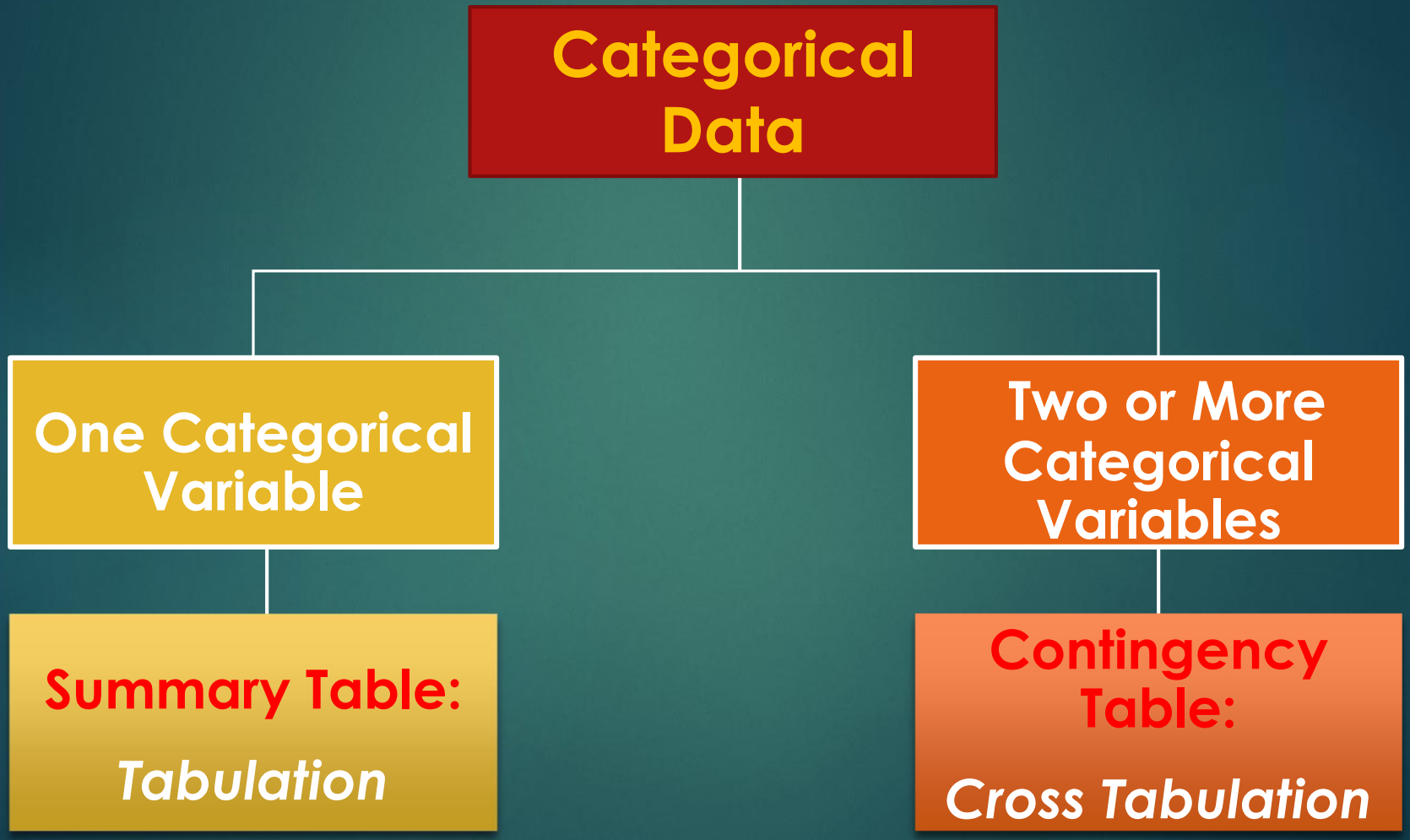
Chapter 2

Organizing and Visualizing Variables

DCOVA

- In Chapter 1, we surveyed different sampling methods that are used to collect the data that are defined by an applied statistician.
  - The D and C in DCOVA
- In this chapter, we survey different methods that can be used to Organize and Visualize data
  - The O and V in DCOVA

Categorical Data Are Organized By Utilizing Tables



Organizing Categorical Data
Summary Table

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

Main Reason Young Adults Shop Online
(USA Today, December 5, 2012)

Reason For Shopping Online?	Percent
Better Prices	37%
Avoiding holiday crowds or hassles	29%
Convenience	18%
Better selection	13%
Ships directly	3%

Organizing Categorical Data
Contingency Table

- A **contingency table** is used to study patterns that may exist between the responses of two or more categorical variables
  - Cross tabulates (i.e., 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

Organizing Categorical Data
Contingency Table

- GIVENCHY's online sales department is asked to report its performance.
- As part of the report, 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.

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

Organizing Categorical Data
Contingency Table

ABSOLUTE VALUES			
Small Amount			
Medium Amount			
Large Amount			
Total			

RELATIVE VALUES			
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

ABSOLUTE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

RELATIVE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

ABSOLUTE VALUES	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount			
Large Amount			
Total			

RELATIVE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

ABSOLUTE VALUES	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount			
Total			

RELATIVE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

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

RELATIVE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

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

RELATIVE VALUES	No Errors	Errors	Total
Small Amount			
Medium Amount			
Large Amount			
Total			

Organizing Categorical Data
Contingency Table

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

RELATIVE VALUES	No Errors	Errors	Total
Small Amount	42.5%		
Medium Amount	25%		
Large Amount	16.2%		
Total	83.7%		

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

Organizing Categorical Data
Contingency Table

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

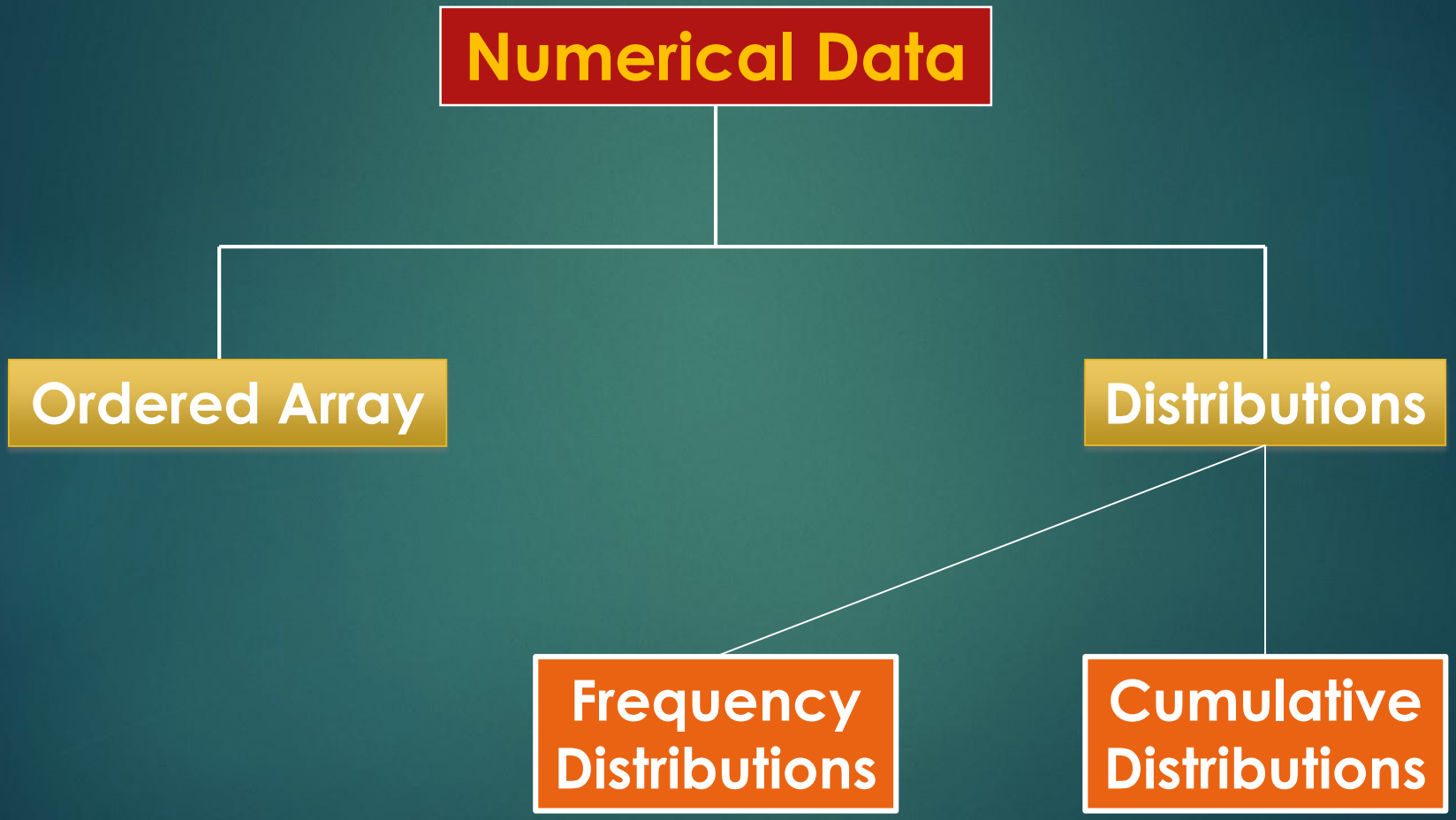
RELATIVE VALUES	No Errors	Errors	Total
Small Amount	42.5%	5%	
Medium Amount	25%	10%	
Large Amount	16.2%	1.2%	
Total	83.7%	16.2%	

Organizing Categorical Data
Contingency Table

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

RELATIVE VALUES	No Errors	Errors	Total
Small Amount	42.5%	5%	47.5%
Medium Amount	25%	10%	35%
Large Amount	16.2%	1.2%	17.5%
Total	83.7%	16.2%	100%

Organizing Numerical Data



Organizing Numerical Data: Ordered Array

- 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)

Meal Costs at 18 City Restaurant and 12 Suburban Restaurants

Prices in the city in dollars

16	17	17	18	18	18
19	19	20	20	21	22
22	25	27	32	38	42

Prices in the suburb in dollars

18	18	19	19	20	21
23	28	32	33	41	45

Organizing Numerical Data: Frequency Distribution

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.
- For this you need to determine:
  1. Class Grouping (i.e., partitions) – typically btw. 5 and 15
  2. Width of a Class

     Divide Range (i.e., Highest value - Lowest value) by the number of class groupings to determine the width.

Organizing Numerical Data: Frequency Distribution

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

VARIABLE: Temperature

DATA:

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

- How to make a Frequency Distribution? A step-by-step guide:

- Step 1.) Sort raw data in ascending order (i.e., from the lowest to the highest):
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Organizing Numerical Data: Frequency Distribution

DATA:

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

- Step 2.) Find the range (i.e., highest value - lowest value):
  Range: 58 - 12 = 46

Organizing Numerical Data: Frequency Distribution

- Step 3.) Select number of classes:
  Let's say 5 (usually btw. 5 and 15)
- Step 4.) Compute the class width:
  Given Step 3, it would be10 (i.e., 46/5, then rounded up)

Organizing Numerical Data: Frequency Distribution

- Step 5.) 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**

Organizing Numerical Data: Frequency Distribution

- Step 6.) Compute class midpoints:
  In this case: 15, 25, 35, 45, 55
- Count observations and assign to classes

Organizing Numerical Data: Frequency Distribution

Here's a tip:

When you put the values in a Frequency Distribution table make sure that:

- Each value is assigned to *only* one class.

and

- Every value is contained in one of the class intervals.

Organizing Numerical Data: Frequency Distribution

DATA:

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 Frequency Distribution

DATA:
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		
20 but less than 30	6		
30 but less than 40	5		
40 but less than 50	4		
50 but less than 60	2		
Total	20		

Organizing Numerical Data:
Relative Frequency Distribution

DATA:
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 (≈ 3/20)	15%
20 but less than 30	6		
30 but less than 40	5		
40 but less than 50	4		
50 but less than 60	2		
Total	20		

Organizing Numerical Data:
Relative Frequency Distribution

DATA:
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 (≈ 3/20)	15%
20 but less than 30	6	.30	30%
30 but less than 40	5		
40 but less than 50	4		
50 but less than 60	2		
Total	20		

Organizing Numerical Data:
Relative Frequency Distribution

DATA:
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 (≈ 3/20)	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 Freq. Distribution

DATA:
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 (=6+3)	45% (≈9/20)
30 but less than 40	5	25%		
40 but less than 50	4	20%		
50 but less than 60	2	10%		
Total	20	100		

Organizing Numerical Data: Frequency Distribution

DATA:
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 (=6+3)	45% (≈9/20)
30 but less than 40	5	25%	14 (=6+3+5)	70% (≈14/20)
40 but less than 50	4	20%	18	90%
50 but less than 60	2	10%	20	100%
Total	20	100	20	100%

Why Do We Use a Frequency Distribution?

- 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 where the data are concentrated

NOTE: Unstacked vs. Stacked Data

UNSTACKED STRUCTURE	City price	Suburb price
Restaurant no 1	.	P1
Restaurant no 2	P2	.
Restaurant no 3	P3	.
Restaurant no 4	.	P4
Restaurant no 5	P5	.

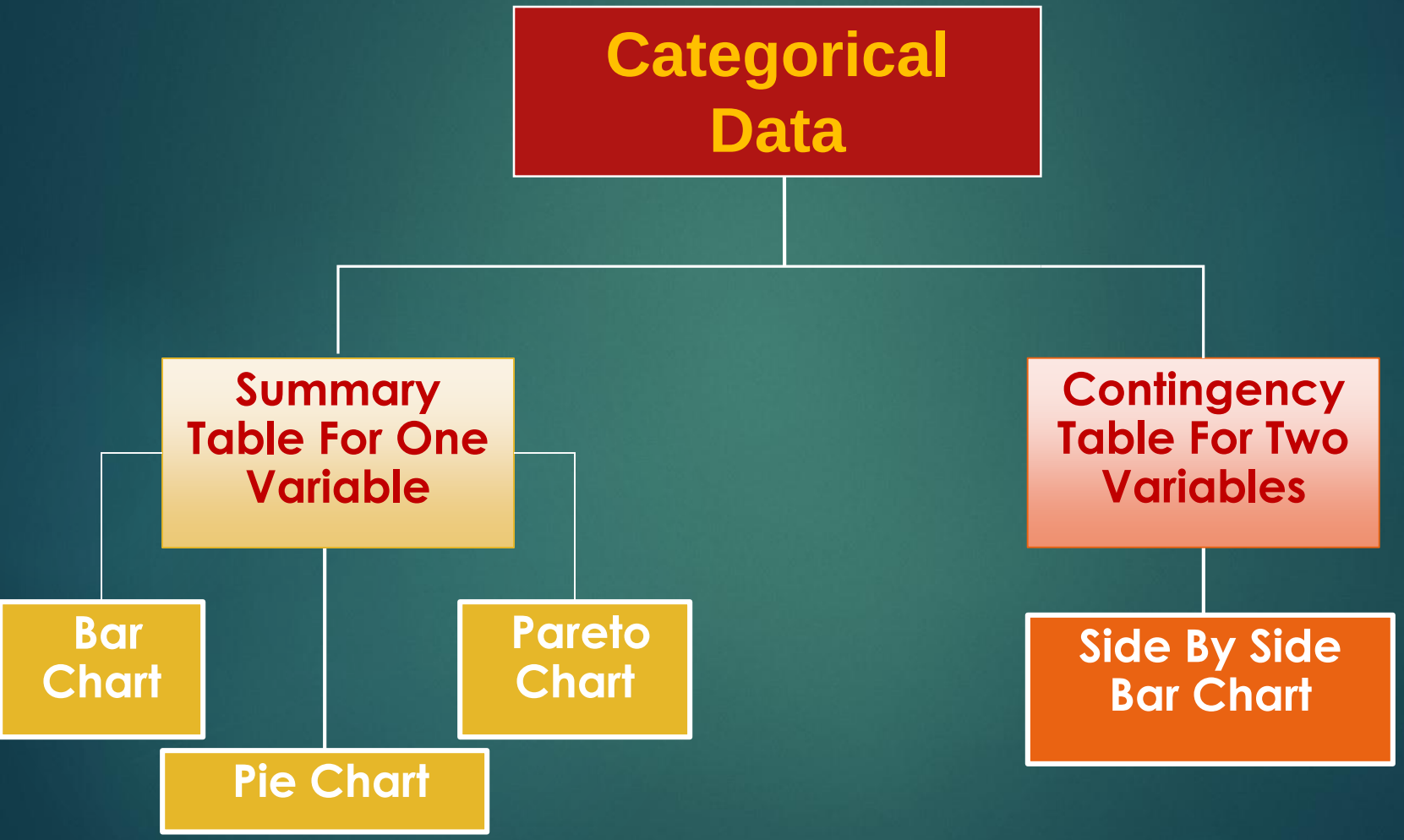
STACKED STRUCTURE	Prices	Location
Restaurant no 1	P1	Suburb
Restaurant no 2	P2	City
Restaurant no 3	P3	City
Restaurant no 4	P4	Suburb
Restaurant no 5	P5	City

Assignment:

MyStatLab, Sections 2.1. and 2.2.

VISUALIZATION

Visualizing Categorical Data Through Graphical Displays



Visualizing Categorical Data: The Bar Chart

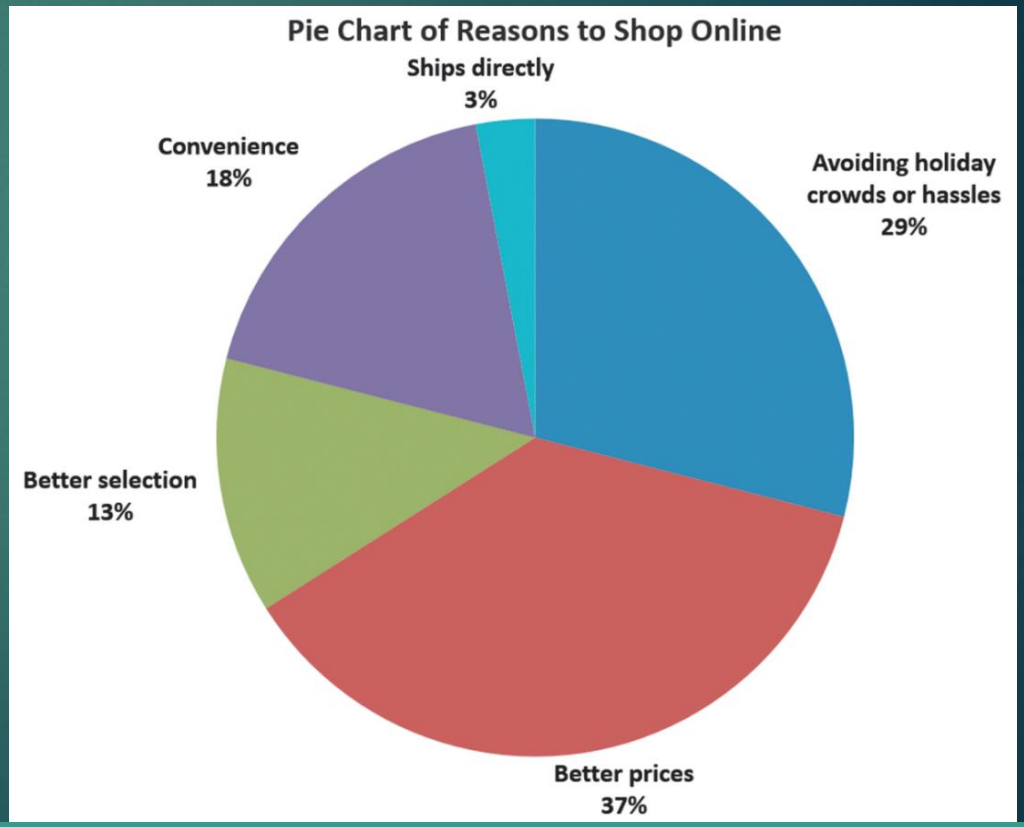
- The **bar chart** visualizes a categorical variable as a series of bars. The length of each bar represents either the frequency or percentage of values for each category. Each bar is separated by a space called a gap.

Reason For Shopping Online?	Percent
Better Prices	37%
Avoiding holiday crowds or hassles	29%
Convenience	18%
Better selection	13%
Ships directly	3%

Visualizing Categorical Data: The Pie Chart

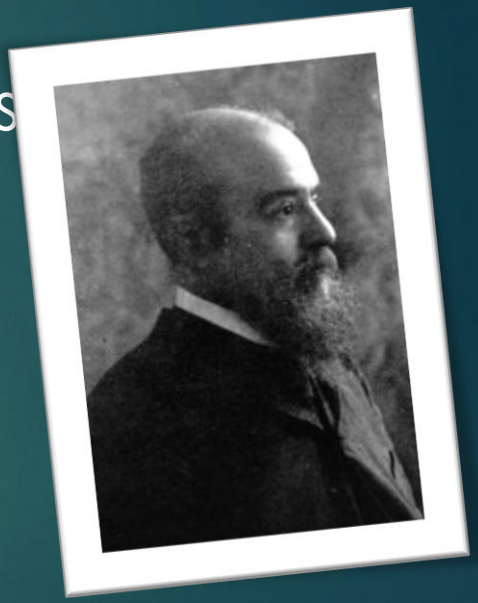
- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

Reason For Shopping Online?	Percent
Better Prices	37%
Avoiding holiday crowds or hassles	29%
Convenience	18%
Better selection	13%
Ships directly	3%



Visualizing Categorical Data: The Pareto Chart

- 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
- The Pareto chart is used to separate the "vital few" from the "trivial many"



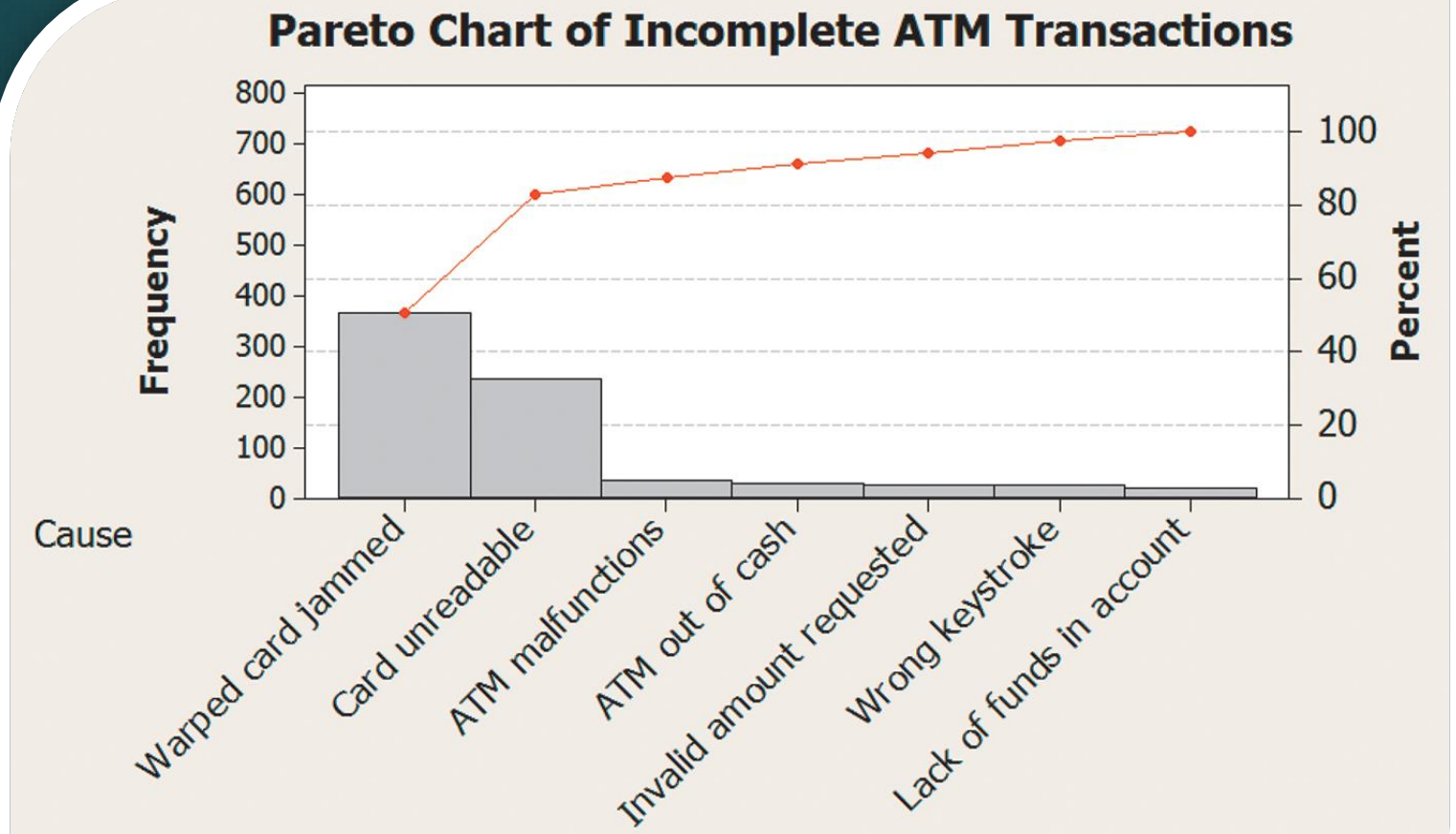
Visualizing Categorical Data: The Pareto Chart

Ordered Summary Table For Causes Of Incomplete ATM Transactions

Cause	Frequency	Percent	Cumulative Percent
Warped card jammed	365	50.41%	50.41%
Card unreadable	234	32.32%	82.73%
ATM malfunctions	32	4.42%	87.15%
ATM out of cash	28	3.87%	91.02%
Invalid amount requested	23	3.18%	94.20%
Wrong keystroke	23	3.18%	97.38%
Lack of funds in account	19	2.62%	100.00%
Total	**724**	**100.00%**	

Source: Data extracted from A. Bhalla, "Don't Misuse the Pareto Principle," *Six Sigma Forum Magazine, May 2009, pp. 15–18.*

Visualizing Categorical Data:
The Pareto Chart

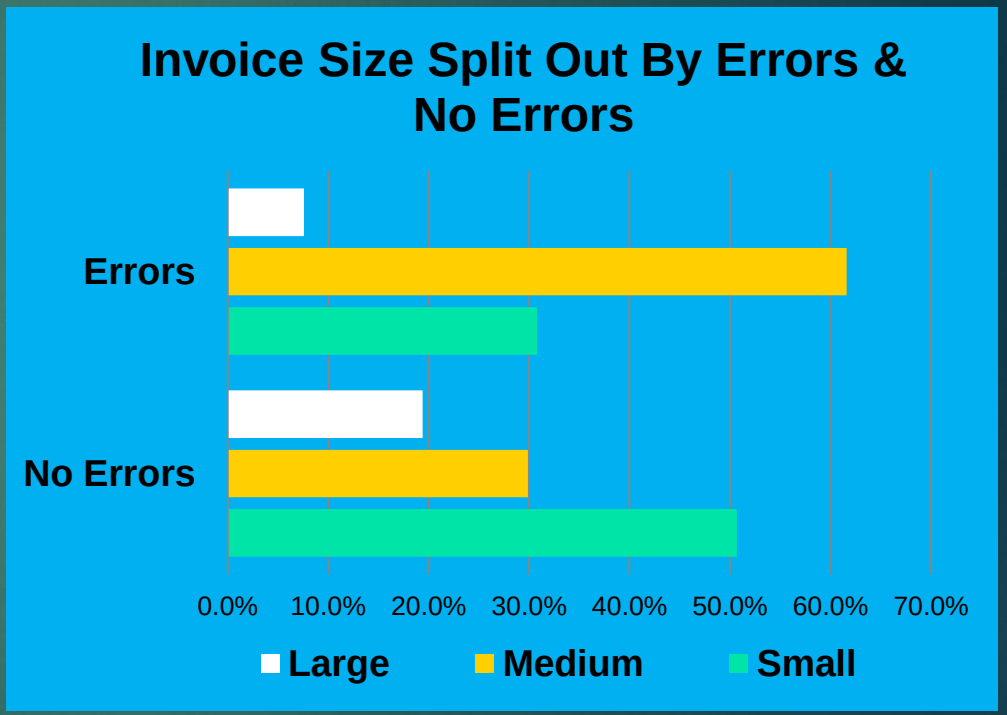


Frequency	365	234	32	28	23	23	19
Percent	50.4	32.3	4.4	3.9	3.2	3.2	2.6
Cum %	50.4	82.7	87.2	91.0	94.2	97.4	100.0

Visualizing Categorical Data: Side By Side Bar Charts

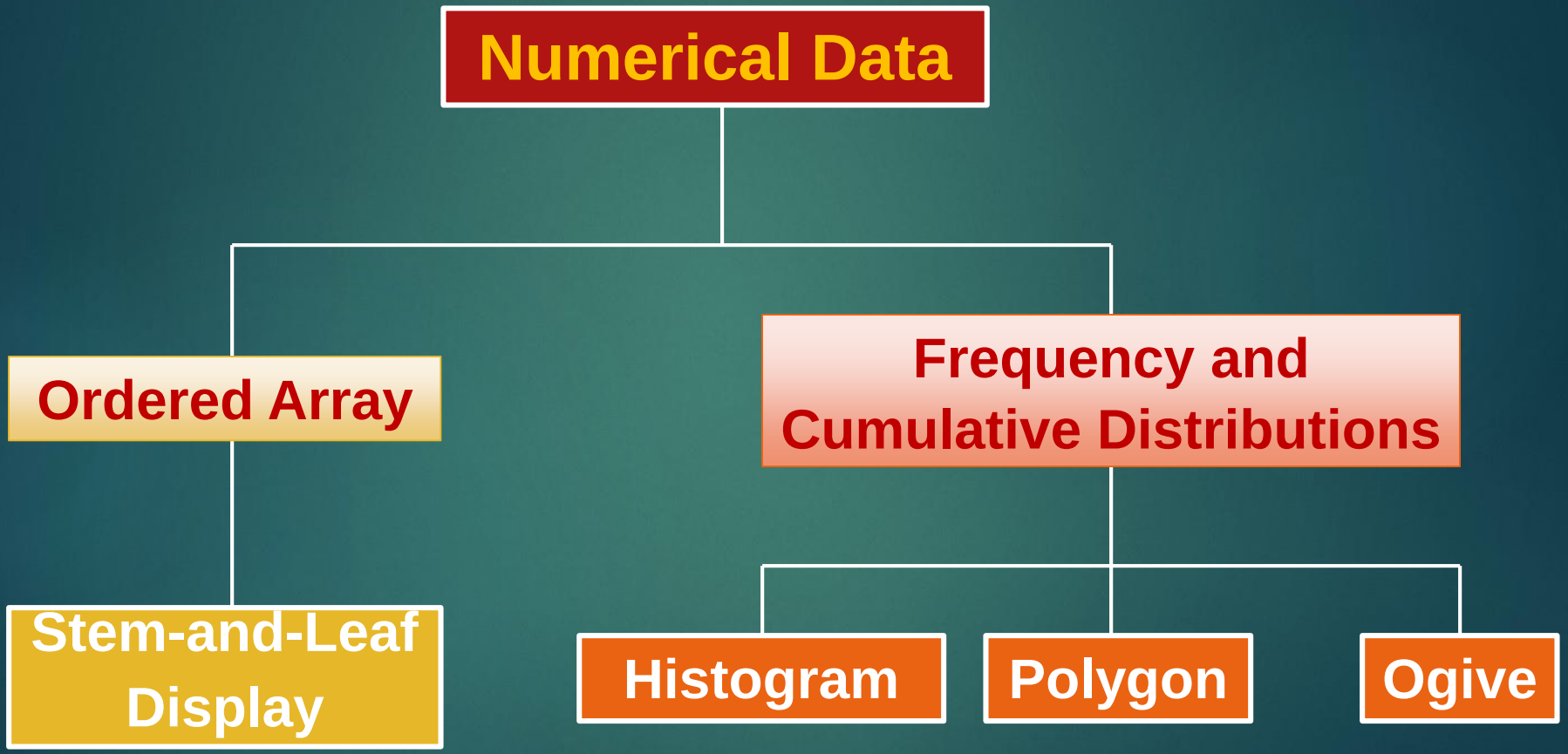
The **side by side bar chart** represents the data from a contingency table.

	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



Stem-and-Leaf Display

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

- Separate the sorted data series

into *leading digits* (the ***stems***)

and

the *trailing digits* (the ***leaves***)

Organizing Numerical Data:
Stem and Leaf Display

In the example below, there are four *leading digits* and multiple *trailing digits*. The Stem and Leaf Display organizes the data using those digits.

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 Student

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

- **Histogram:**

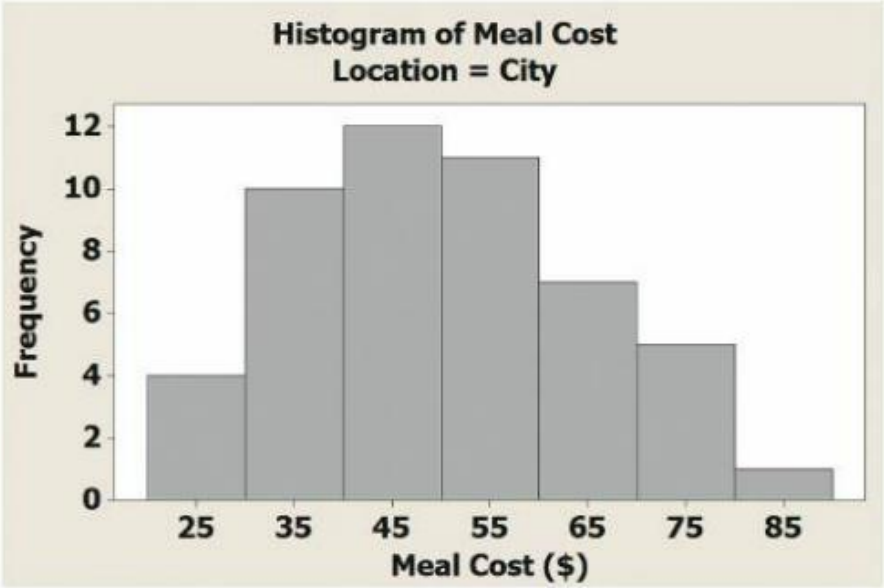
  A vertical bar chart that illustrate the information in a frequency distribution.

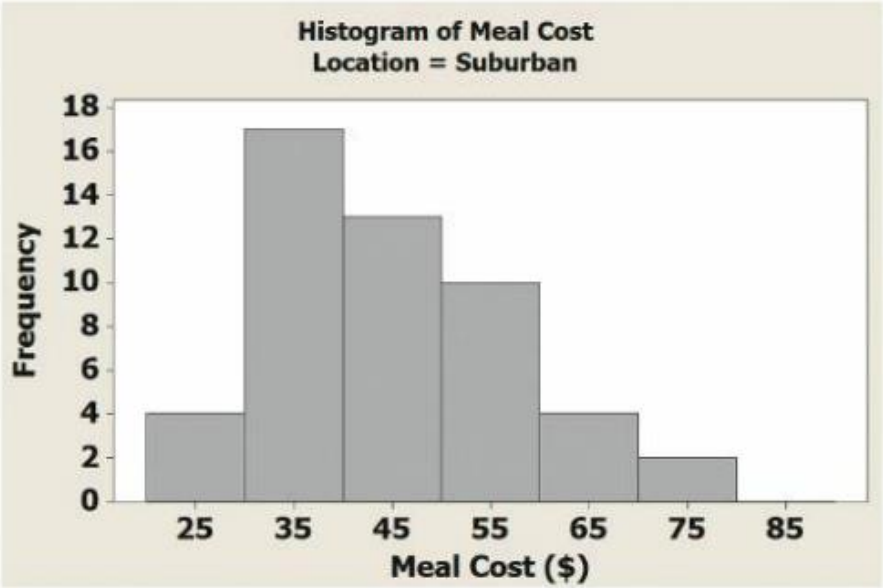
  - In a histogram there are often no gaps between adjacent bars.
  - The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
  - The vertical axis is either **frequency**, **relative frequency**, or **percentage**.

Example:
Frequency Distribution

Meal Cost ($)	City Frequency	Suburban Frequency
20 but less than 30	4	4
30 but less than 40	10	17
40 but less than 50	12	13
50 but less than 60	11	10
60 but less than 70	7	4
70 but less than 80	5	2
80 but less than 90	1	0
Total	50	50

Example:
The Histogram



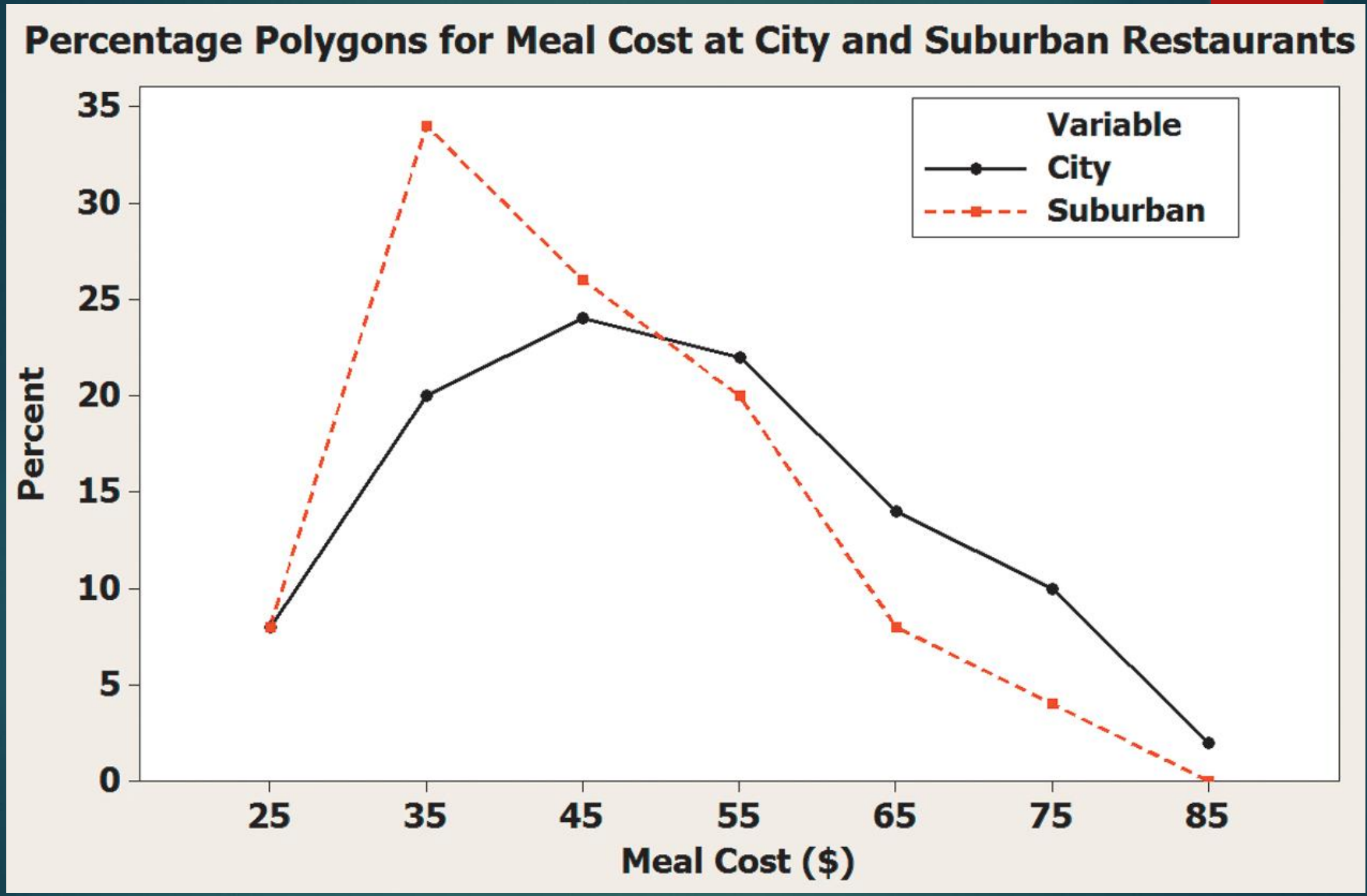


Visualizing Numerical Data: The Polygon and Ogive

- **Percentage Polygon** is formed by having the midpoint of each class represent the *percentage frequency* of observations in that class and then connecting the sequence of midpoints at their respective class percentages.
- **Ogive (Cumulative Percentage Polygon)** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

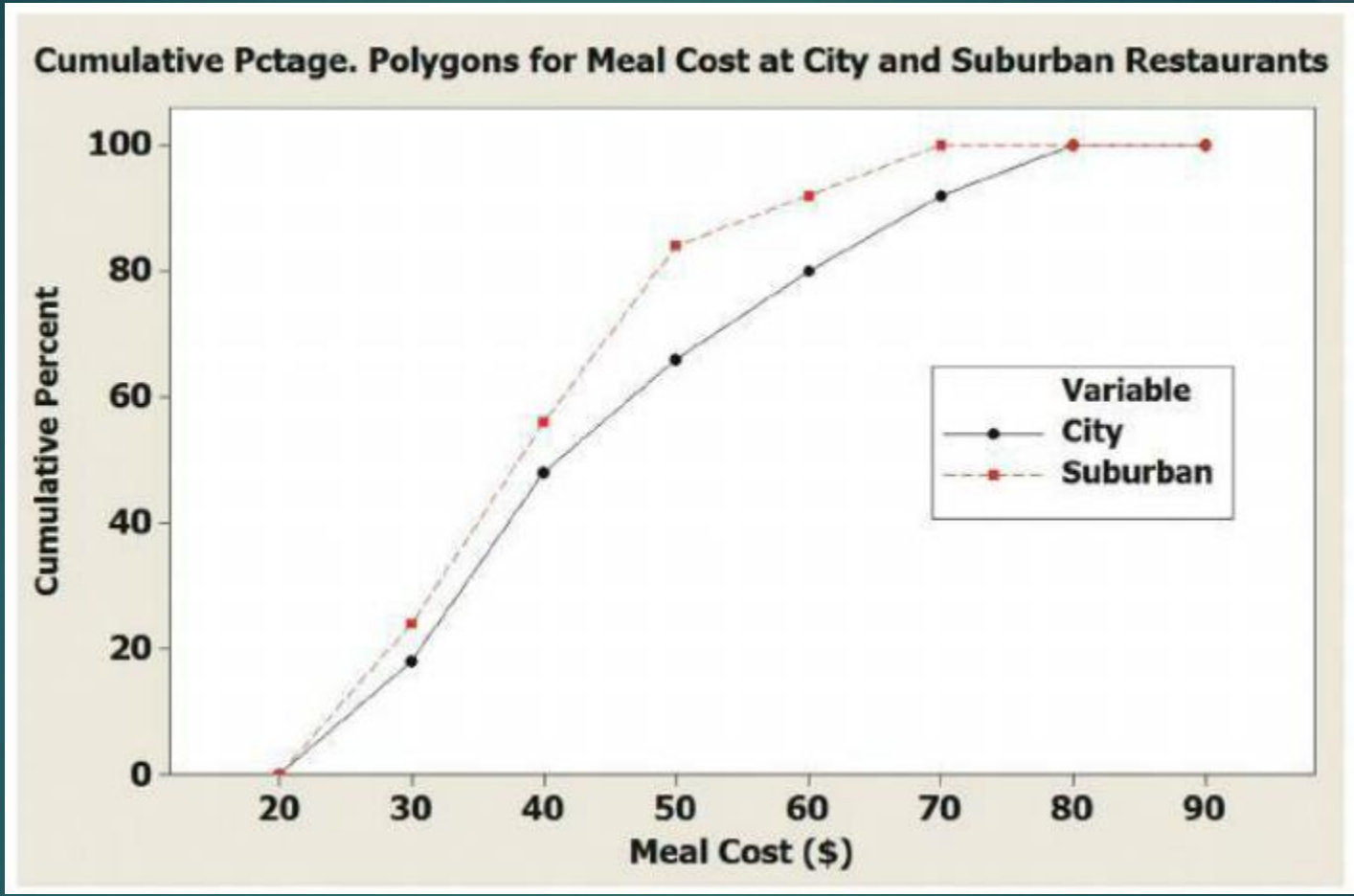
- Useful when there are two or more groups to compare.

MEAL COST ($)	CITY		SUBURBAN	
	Relative Frequency	Percentage	Relative Frequency	Percentage
20 but less than 30	0.08	8.0%	0.08	8.0%
30 but less than 40	0.20	20.0%	0.34	34.0%
40 but less than 50	0.24	24.0%	0.26	26.0%
50 but less than 60	0.22	22.0%	0.20	20.0%
60 but less than 70	0.14	14.0%	0.08	8.0%
70 but less than 80	0.10	10.0%	0.04	4.0%
80 but less than 90	0.02	2.0%	0.00	0.0%
Total	1.00	100.0%	1.00	100.0%



MEAL COST ($)	CITY		SUBURBAN	
	Relative Frequency	Percentage	Relative Frequency	Percentage
20 but less than 30	0.08	8.0%	0.08	8.0%
30 but less than 40	0.20	20.0%	0.34	34.0%
40 but less than 50	0.24	24.0%	0.26	26.0%
50 but less than 60	0.22	22.0%	0.20	20.0%
60 but less than 70	0.14	14.0%	0.08	8.0%
70 but less than 80	0.10	10.0%	0.04	4.0%
80 but less than 90	0.02	2.0%	0.00	0.0%
Total	1.00	100.0%	1.00	100.0%

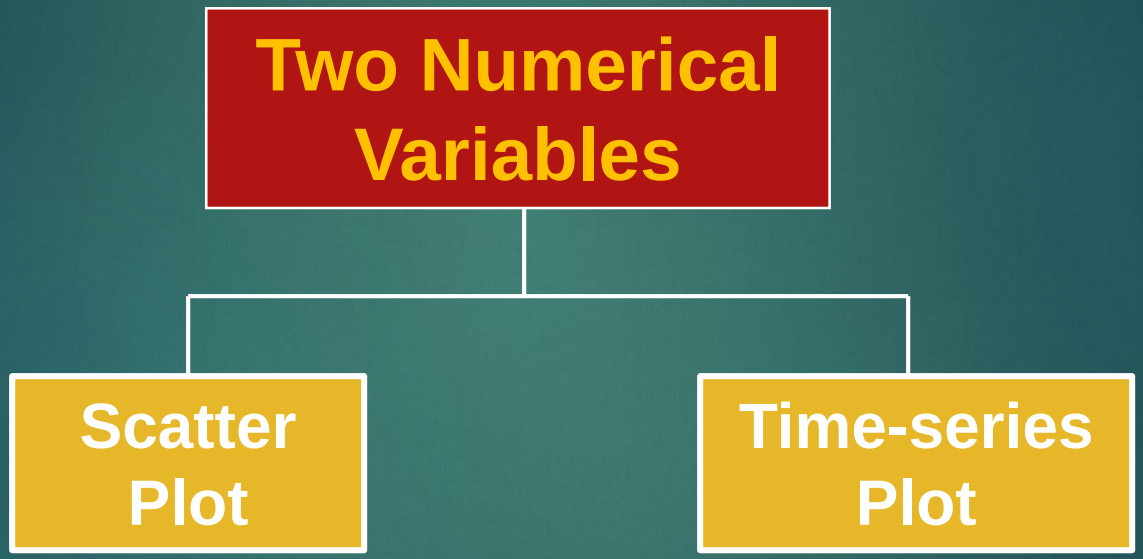
Meal Cost ($)	Percentage of City Restaurants Meals That Cost Less Than Indicated Amount	Percentage of Suburban Restaurants Meals That Cost Less Than Indicated Amount
20	0	0
30	8	8
40	28	42
50	52	68
60	74	88
70	88	96
80	98	100
90	100	100
100	100	100



Assignment:

MyStatLab, Sections 2.3. and 2.4.

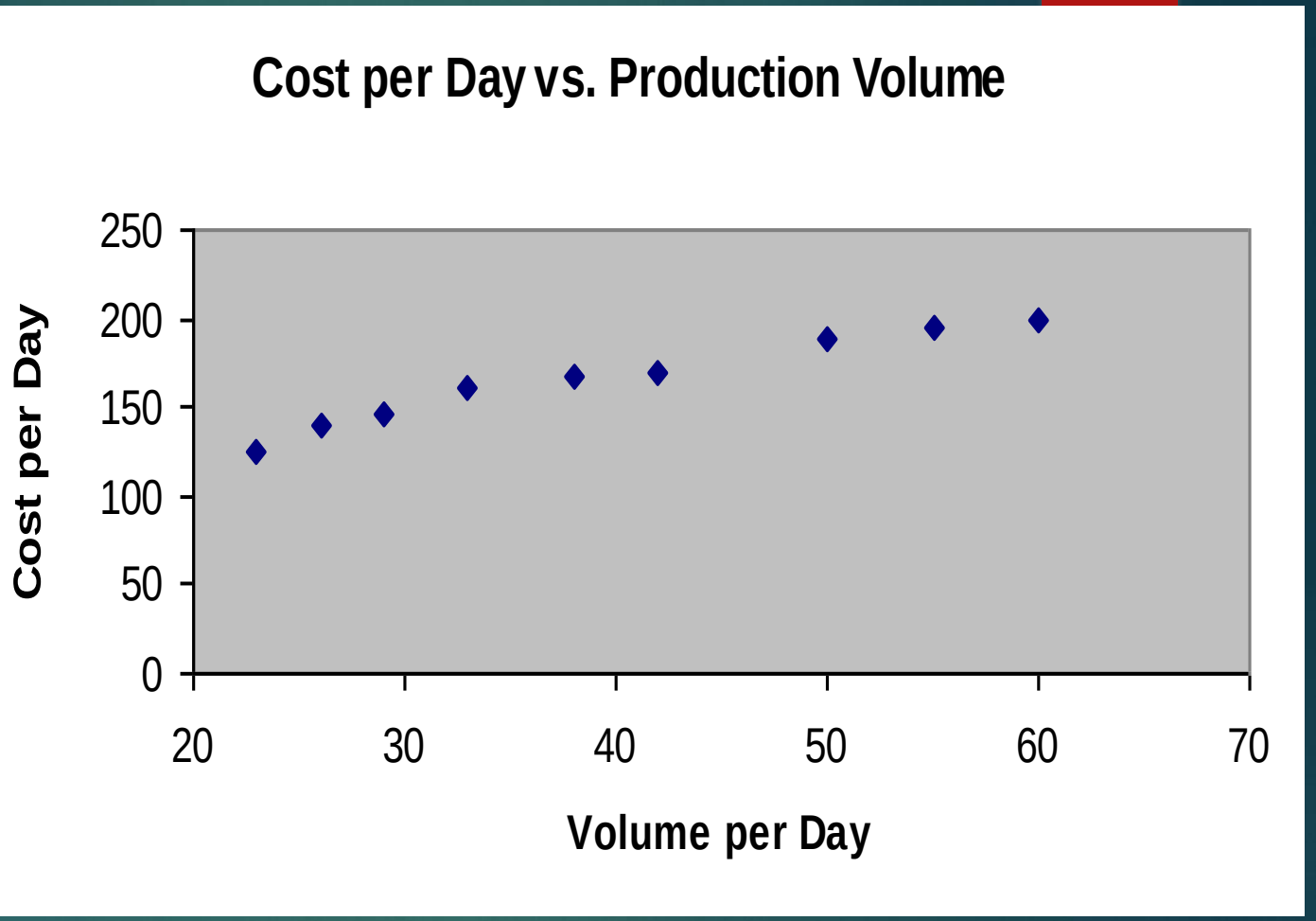
Visualizing Two Numerical Variables By Using Graphical Displays



Visualizing Two Numerical Variables: The Scatter Plot

- **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

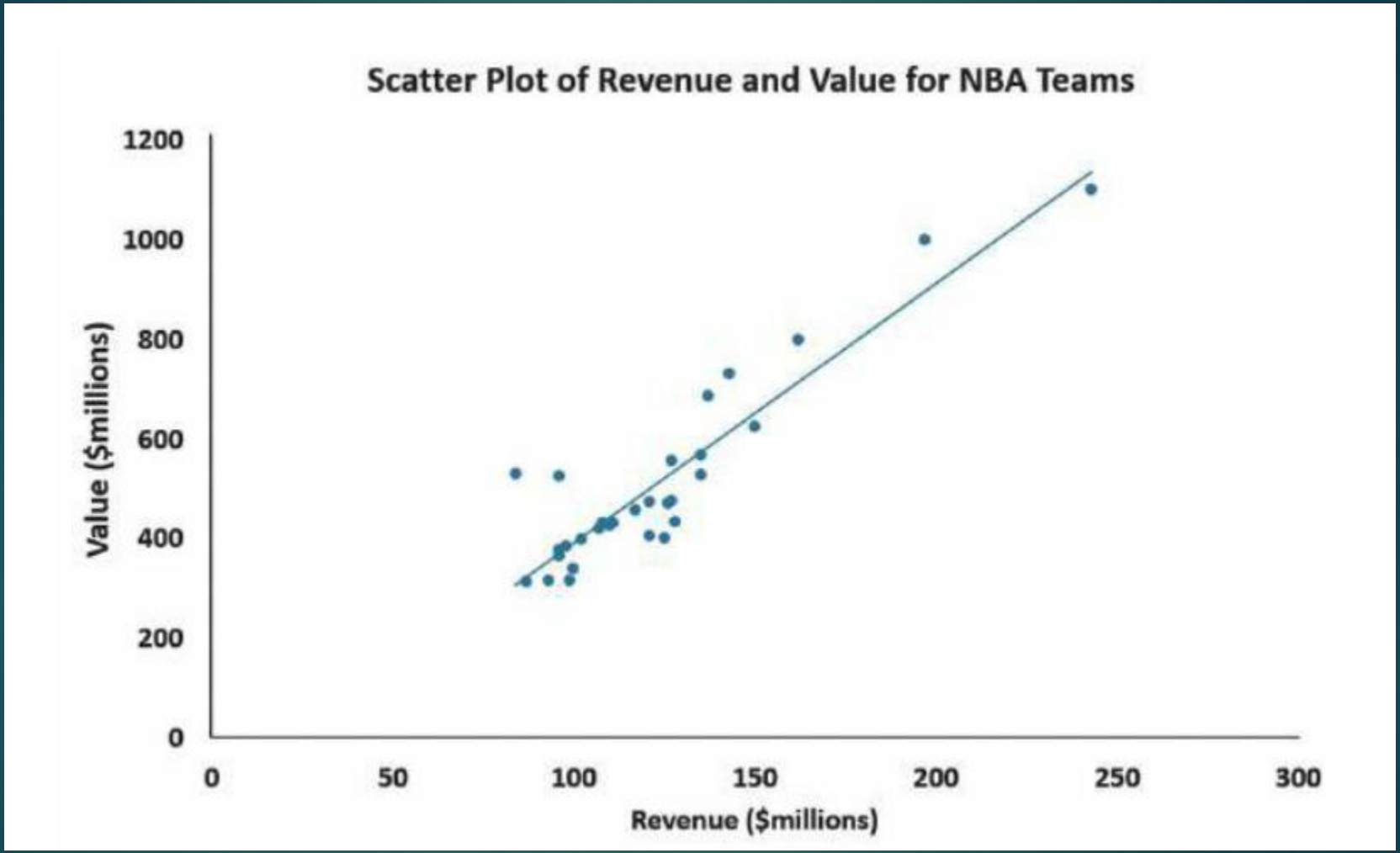
Q per day	C per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Detecting a Relationship using Scatter Plots

Team Code	Revenue ($millions)	Value ($millions)	Team Code	Revenue ($millions)	Value ($millions)	Team Code	Revenue ($millions)	Value ($millions)
ATL	99	316	HOU	135	568	OKC	127	475
BOS	143	730	IND	98	383	ORL	126	470
BRK	84	530	LAC	108	430	PHI	107	418
CHA	93	315	LAL	197	1,000	PHX	121	474
CHI	162	800	MEM	96	377	POR	117	457
CLE	128	434	MIA	150	625	SAC	96	525
DAL	137	685	MIL	87	312	SAS	135	527
DEN	110	427	MIN	96	364	TOR	121	405
DET	125	400	NOH	100	340	UTA	111	432
GSW	127	555	NYK	243	1,100	WAS	102	397

Source: Data extracted from **www.forbes.com/nba-valuations**.



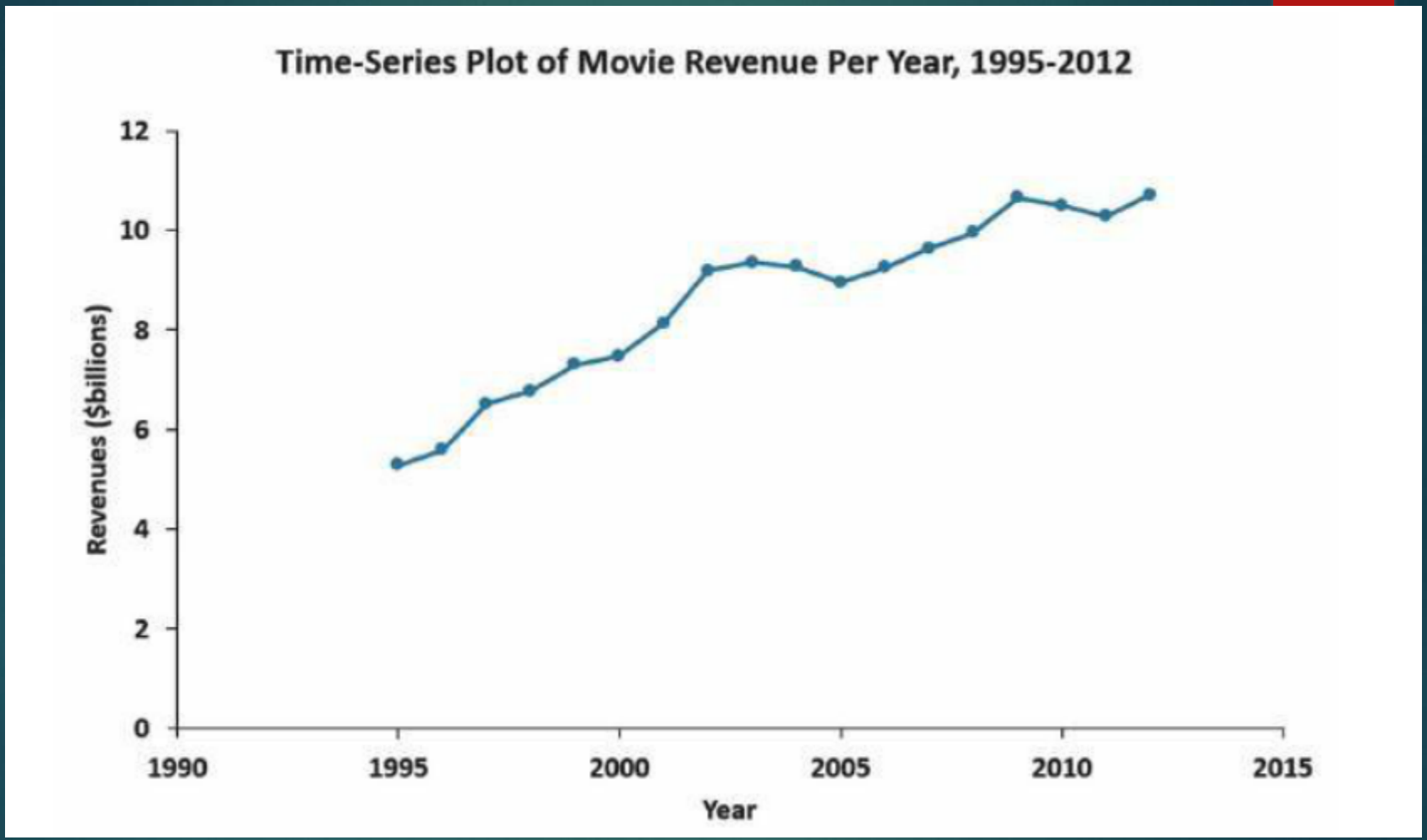
Visualizing Two Numerical Variables: The Time Series Plot

- **Time-Series Plot** is used to study patterns in the values of a numeric variable over time
  - Numeric variable is measured on the vertical axis and the time period is measured on the horizontal axis

Time Series Plot Example: Movie Revenue

Year	Revenue ($billions)	Year	Revenue ($billions)
1995	5.29	2004	9.27
1996	5.59	2005	8.95
1997	6.51	2006	9.25
1998	6.77	2007	9.63
1999	7.30	2008	9.95
2000	7.48	2009	10.65
2001	8.13	2010	10.50
2002	9.19	2011	10.28
2003	9.35	2012	10.71

Source: Data extracted from **www.the-numbers.com/market**, March 18, 2013.



Assignment:

MyStatLab, Sections 2.5.

Read Sections 2.6 and 2.7. (I left some slides below)
Do the homework for those two sections.

Organizing Many Categorical Variables: The Multidimensional Contingency Table

- A multidimensional contingency table is constructed by tallying the responses of three or more categorical variables.
- In Excel creating a *Pivot Table* to yields an interactive display of this type.
- While Minitab will not create an interactive table, it has many specialized statistical & graphical procedures (not covered in this book) to analyze & visualize multidimensional data.

Organizing Many Categorical Variables: The Multidimensional Contingency Table

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.

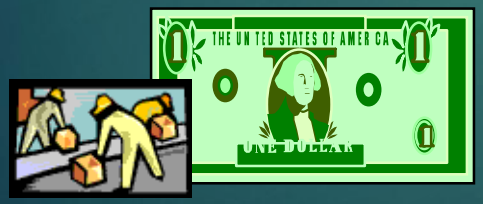
DO NOT OBSCURE THE DATA!

- 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

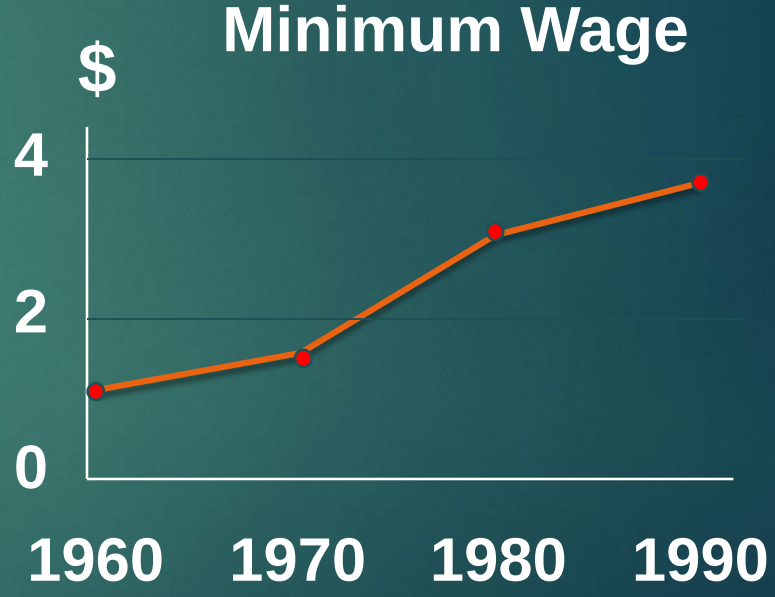
Graphical Errors: Chart Junk

Bad Presentation

Minimum Wage

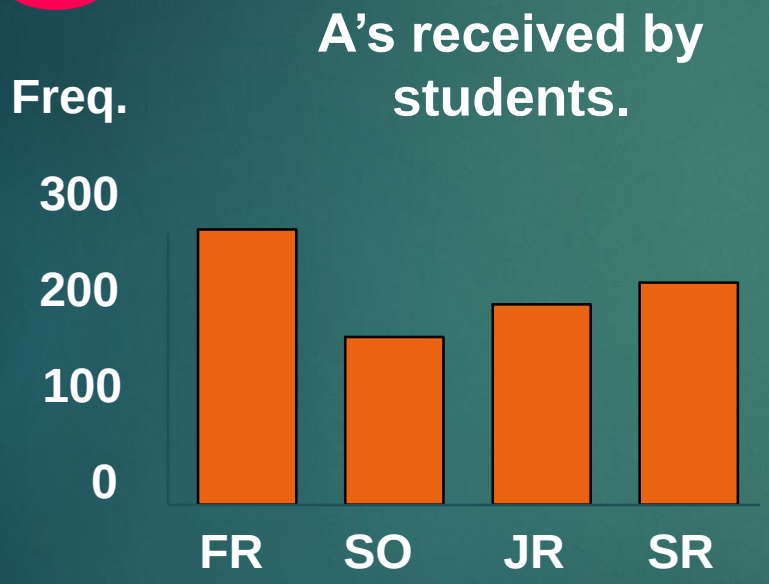


Good Presentation

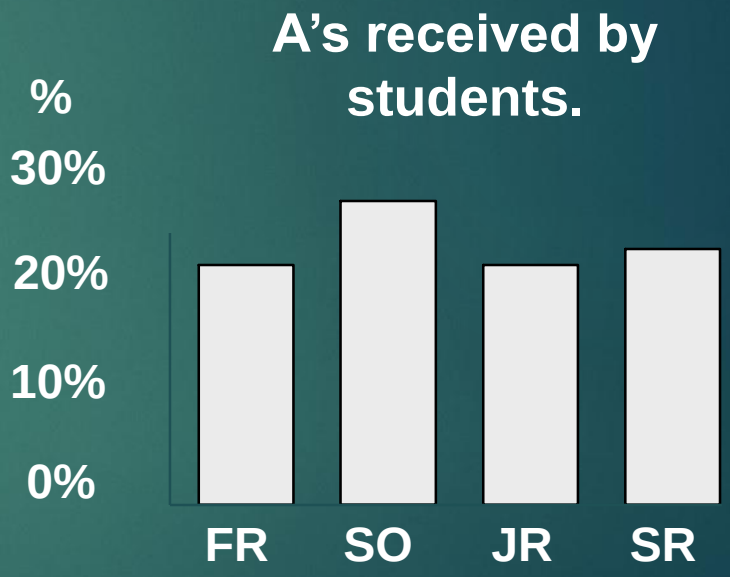


Graphical Errors: No Relative Basis

Bad Presentation



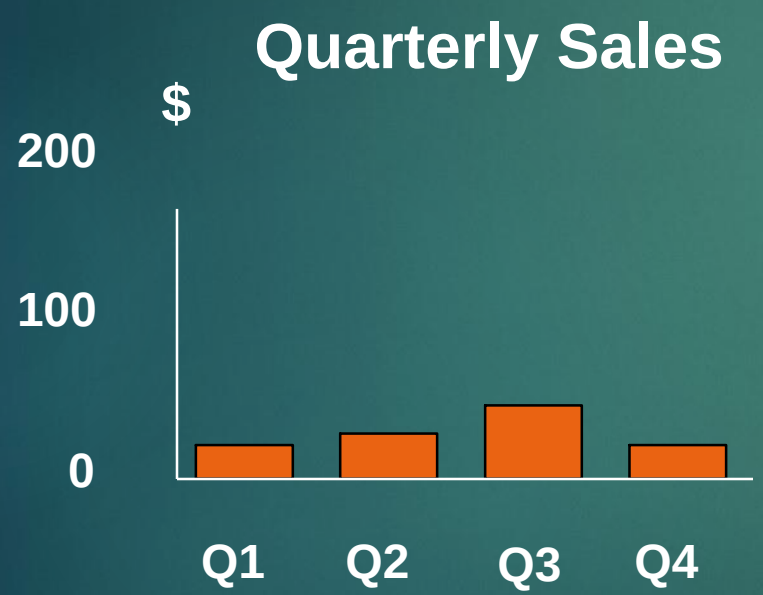
Good Presentation



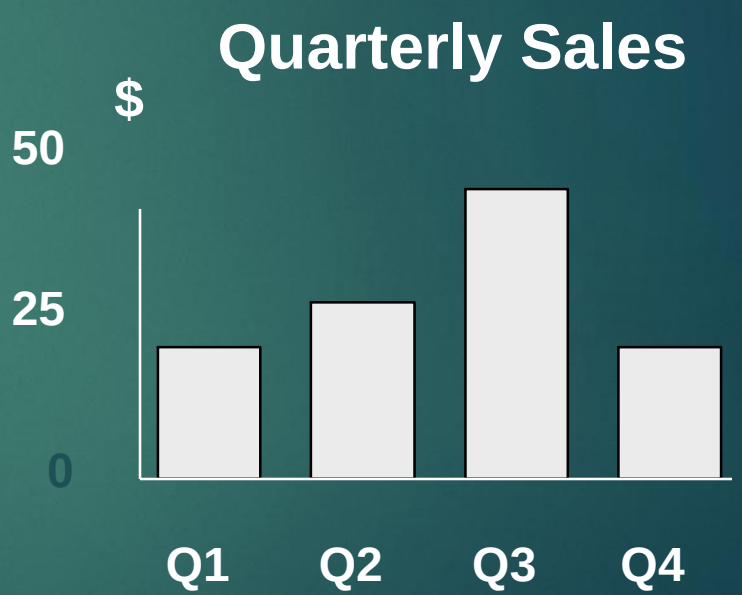
FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Graphical Errors: Compressing the Vertical Axis

Bad Presentation

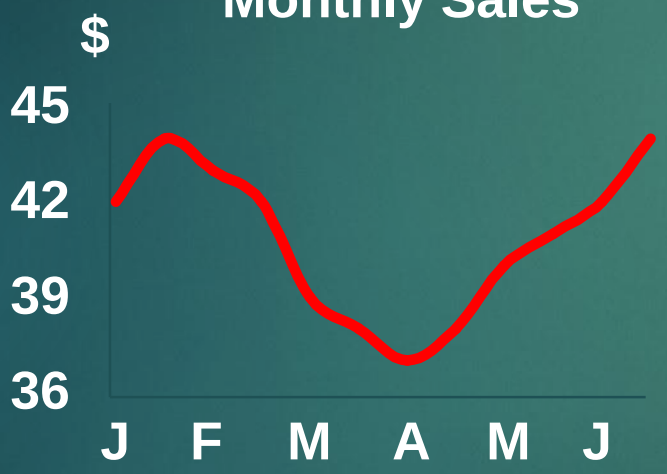


Good Presentation



Graphical Errors: No Zero Point on the Vertical Axis

Bad Presentation



Good Presentations

Graphing the first six months of sales