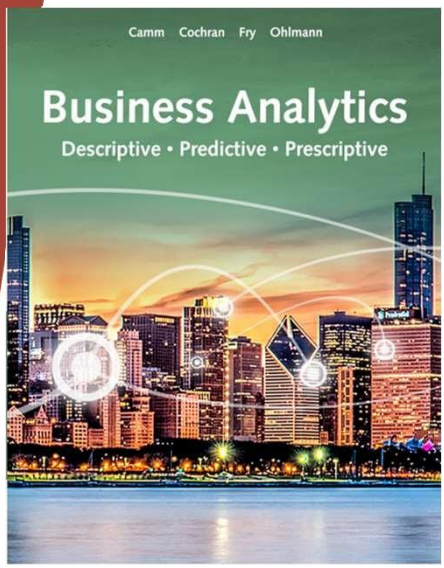


Business Analytics

Chapter 3 Data Visualization

HAMLINE UNIVERSITY
SCHOOL OF BUSINESS

Introduction

- Data visualization involves:
  - Creating a summary table for the data.
  - Generating charts to help interpret, analyze, and learn from the data.
- Uses of data visualization:
  - Helpful for identifying data errors.
  - Reduces the size of your data set by highlighting important relationships and trends in the data.

Why is Data Visualization Important?

It Can Save Lives!

London 1854

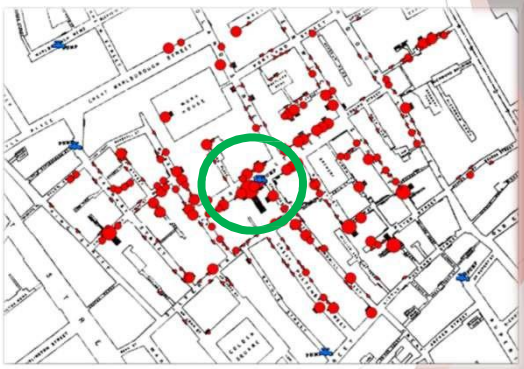
- John Snow (Not that one!)
- Mapped Cholera cases

What do you notice about the data?

- Clustering around water pump!
- Cholera was water-borne!

Results?

- Transformed the medical profession
- Saved Lives!



1. Know Your Audience

- What is the purpose?
- Who will use the data?
- What are their objectives?
- How will they interact with the data?
- What business questions do users need answered?

2. Use Space and Colors Well

Use of Color is Powerful!

- ☐ Aesthetics
- ☐ Higher Memorability
- ☐ Draws Attention

Used For:

- ☐ Sequential Identification
- ☐ Show Divergence
- ☐ Categories

Avoid Clutter

- ☐ Lose Actionable Information

Figure: U.S. Regional Composition

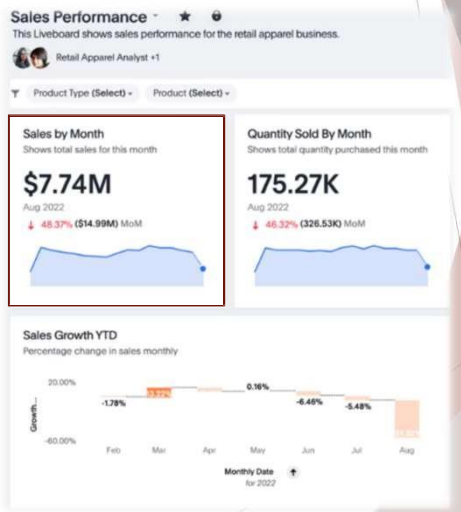
3. Highlight Important Information

Tell a Story

- ☐ **Point out what is important**
- ☐ **Make it easy to follow**

Placement

- ☐ **For Western Audiences:**
- ☐ **Top Left = Most Important**



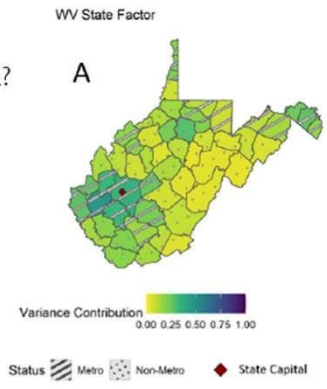
4. Keep it Simple and Straightforward (K.I.S.S)

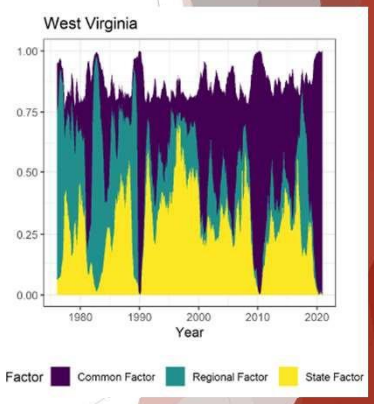
Use Correct Figure Types

- ☐ What story are you trying to tell?
- ☐ What type of data do you have?

Make It Easy For Viewers

- ☐ Less Distractions
- ☐ Less Conscious Thought





Overview of Data Visualization

Effective Design Techniques:

- **Data-ink ratio:**
  - Measures the proportion of what Tufte terms "data-ink" to the total amount of ink used in a table or chart.
- Helpful for creating effective tables and charts for data visualization:
  - Data-ink: Ink used in a table or chart that is necessary to convey the meaning of the data to the audience.
  - Non-data-ink: Ink used in a table or chart that serves no useful purpose in conveying the data to the audience.

Which One is Better?

A

Scarf Sales			
Day	Sales (units)	Day	Sales (units)
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

B

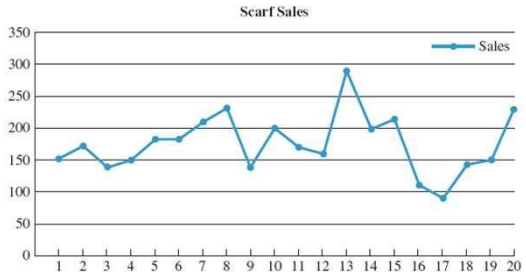
Increasing the Data-Ink Ratio by Removing Unnecessary Gridlines

Scarf Sales			
Day	Sales (units)	Day	Sales (units)
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

Which One is Better?

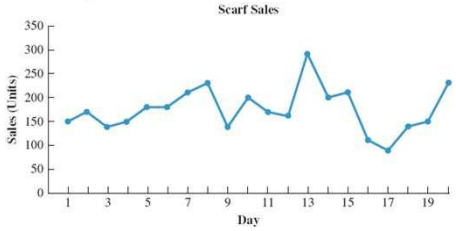
A

Example of a Low Data-Ink Ratio Chart



B

Increasing the Data-Ink Ratio by Adding Labels to Axes and Removing Unnecessary Lines and Labels



What's Wrong With This?

Cluster Specific Summary Statistics from June 2017 FooDS for Clustering and Censored Regression Variables

	Cluster=1 (High Income)		Cluster=2 (Older and More Concerned)		Cluster=3 (Younger and Less Concerned)		Cluster=4 (Households with Children)		Cluster=5 (Food Stamp Reciepients)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Level of Concern for e. coli	3.18	1.24	4.18	0.81	2.21	1.17	3.40	1.29	3.41	1.31
Household Size	2.74	1.08	1.87	0.72	2.22	1.16	3.91	0.91	2.10	1.05
Household Income										
<$20,000	0.00%		13.38%		21.99%		5.56%		68.22%	
$20,000 - $39,999	0.00%		26.76%		28.63%		14.65%		20.56%	
$40,000 - $59,999	0.49%		19.40%		21.99%		15.15%		5.61%	
$60,000 - $79,999	3.92%		25.08%		14.94%		17.68%		1.87%	
$80,000 - $99,999	14.71%		9.70%		7.88%		13.64%		0.93%	
$100,000 - $119,999	21.08%		5.02%		4.56%		11.11%		0.93%	
$120,000 - $139,999	19.61%		0.67%		0.00%		7.07%		0.93%	
$140,000 - $159,999	17.16%		0.00%		0.00%		7.58%		0.93%	
>$159,999	23.04%		0.00%		0.00%		7.58%		0.00%	
Age										
18-24	7.84%		0.00%		29.05%		13.13%		8.41%	
25-34	12.75%		4.68%		28.63%		31.31%		11.21%	
35-44	11.27%		6.69%		17.01%		27.78%		19.63%	
45-54	24.51%		13.38%		10.79%		20.20%		29.91%	
55-64	23.53%		27.76%		10.79%		5.56%		18.69%	
65-73	17.16%		40.13%		3.73%		2.02%		9.35%	
>73	2.94%		7.36%		0.00%		0.00%		2.80%	
Current Food Stamp Recipient	0.00%	0.00	0.00%	0.00	0.00%	0.00	16.16%	0.37	100.00%	0.00
Children in the Household	0.00%	0.00	0.00%	0.00	0.41%	0.06	99.49%	0.07	14.95%	0.36
Cluster size	204		299		241		198		107	
Percentage of total Sample	19.45%		28.50%		22.97%		18.88%		10.20%	

Choosing FSMA Exempt Tomatoes when ONLY Food Safety Label is Present		
Variable	Estimate	SE
Intercept	-2.0587***	0.4773
Importance of Food Safety	-0.1234	0.1405
Female	**0.2467***	**0.1486**
Vegan and/or Vegetarian	**-0.9259****	**0.4724**
Spanish Origin	0.3266	0.236
Liberal	-0.0143	0.1521
Food at home expenditure	**0.3326***	**0.1738**
Food at home expenditure squared	**-0.0261***	**0.0152**

***, **, *** indicated significance at the 10%, 5%, and 1% levels

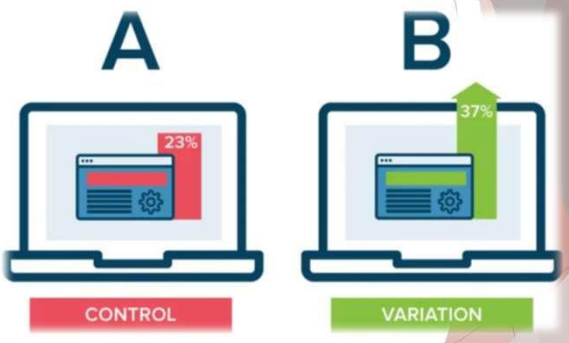
5. Get Other Opinions

Ask someone!

- Ask a lot of people
- What looks better
- What performs better

Example: A/B Testing

- Compare 2 versions of something to figure out which performs better
- Typically used for websites and apps
  - Started 1920



Tables

Table Design Principles

Crosstabulation

PivotTables in Excel

Recommended PivotTables in Excel

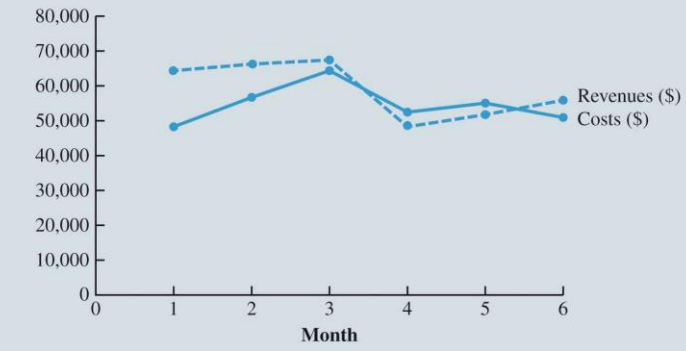
Should I Use Tables or Charts?

Tables should be used when:

1. The reader needs to refer to specific numerical values.
2. The reader needs to make precise comparisons between different values and not just relative comparisons.
3. The values being displayed have different units or very different magnitudes.

Tables (4 of 18)

Figure 3.6: Combined Line Chart and Table for Monthly Costs and Revenues at Gossamer Industries



	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

Tables

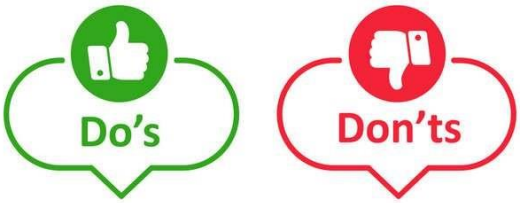
Table 3.4: Table Displaying Head Count, Costs, and Revenues at Gossamer Industries

	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Total
Head Count	8	9	10	9	9	9	
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

Tables

Table Design Principles:

- Avoid using vertical lines in a table unless they are necessary for clarity.
- Horizontal lines are generally necessary only for separating column titles from data values or when indicating that a calculation has taken place.



Tables

Figure 3.7: Comparing Different Table Designs

Design A:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design B:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design C:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design D:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Tables

Table 3.5: Larger Table Showing Revenues by Location for 12 Months of Data

Revenues by Location ($)	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
Temple	8,987	8,595	8,958	6,718	8,066	8,574
Killeen	8,212	9,143	8,714	6,869	8,150	8,891
Waco	11,603	12,063	11,173	9,622	8,912	9,553
Belton	7,671	7,617	7,896	6,899	7,877	6,621
Granger	7,642	7,744	7,836	5,833	6,002	6,728
Harker Heights	5,257	5,326	4,998	4,304	4,106	4,980
Gatesville	5,316	5,245	5,056	3,317	3,852	4,026
Lampasas	5,266	5,129	5,022	3,022	3,088	4,289
Academy	4,170	5,266	7,472	1,594	1,732	2,025
Total Costs ($)	64,124 48,123	66,128 56,458	67,125 64,125	48,178 52,158	51,785 54,718	55,687 50,985

Tables

- **Crosstabulation:**
  - A useful type of table for describing data of two variables.
- **PivotTable:**
  - A crosstabulation in Microsoft Excel.

Pivot Table 1

Sales	Sep	Oct	Nov	Total
Apples	250	590		840
John		180		180
Mike		120		120
Pete		290		290
Sally	250			250
Bananas		430	600	1030
John			400	400
Mike			200	200
Pete		180		180
Sally		250		250
Cherries	580	910		1490
John		250		250
Mike	250	330		580
Pete		330		330
Sally	330			330
Oranges		120	720	840
John			120	120
Mike			400	400
Pete		120		120
Sally			200	200
Total	830	2050	1320	4200

Pivot Table 2

Month	(All)				
Sales	Product				
Reseller	Apples	Bananas	Cherries	Oranges	Total
John	$180	$400	$250	$120	$950
Mike	$120	$200	$580	$400	$1,300
Pete	$290	$180	$330	$120	$920
Sally	$250	$250	$330	$200	$1,030
Total	$840	$1,030	$1,490	$840	$4,200

Pivot Table 3

Product	(All)			
Sales	Month			
Reseller	Sep	Oct	Nov	Total
John		$430	$520	$950
Mike	$250	$450	$600	$1,300
Pete		$920		$920
Sally	$580	$250	$200	$1,030
Total	$830	$2,050	$1,320	$4,200

Tables

Table 3.6: Quality Rating and Meal Price for 300 Los Angeles Restaurants

Restaurant	Quality Rating	Meal Price ($)	Wait Time (min)
1	Good	18	5
2	Very Good	22	6
3	Good	28	1
4	Excellent	38	74
5	Very Good	33	6
6	Good	28	5
7	Very Good	19	11
8	Very Good	11	9
9	Very Good	23	13
10	Good	13	1

Tables

Table 3.7: Crosstabulation of Quality Rating and Meal Price for 300 Los Angeles Restaurants

	Meal Price				
Quality Rating	$10-19	$20-29	$30-39	$40-49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

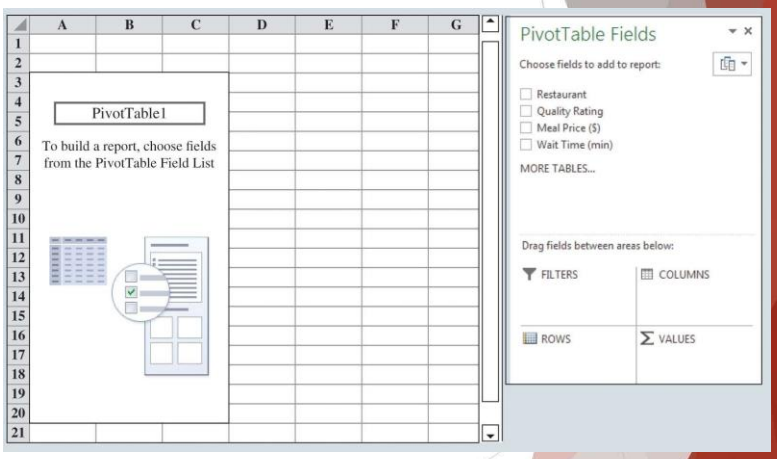
Tables

Figure 3.8: Excel Worksheet Containing Restaurant Data

	A	B	C	D
1	Restaurant	Quality Rating	Meal Price ($)	Wait Time (min)
2	1	Good	18	5
3	2	Very Good	22	6
4	3	Good	28	1
5	4	Excellent	38	74
6	5	Very Good	33	6
7	6	Good	28	5
8	7	Very Good	19	11
9	8	Very Good	11	9
10	9	Very Good	23	13
11	10	Good	13	1
12	11	Very Good	33	18
13	12	Very Good	44	7
14	13	Excellent	42	18
15	14	Excellent	34	46
16	15	Good	25	0
17	16	Good	22	3
18	17	Good	26	3
19	18	Excellent	17	36
20	19	Very Good	30	7
21	20	Good	19	3
22	21	Very Good	33	10
23	22	Very Good	22	14
24	23	Excellent	32	27
25	24	Excellent	33	80
26	25	Very Good	34	9

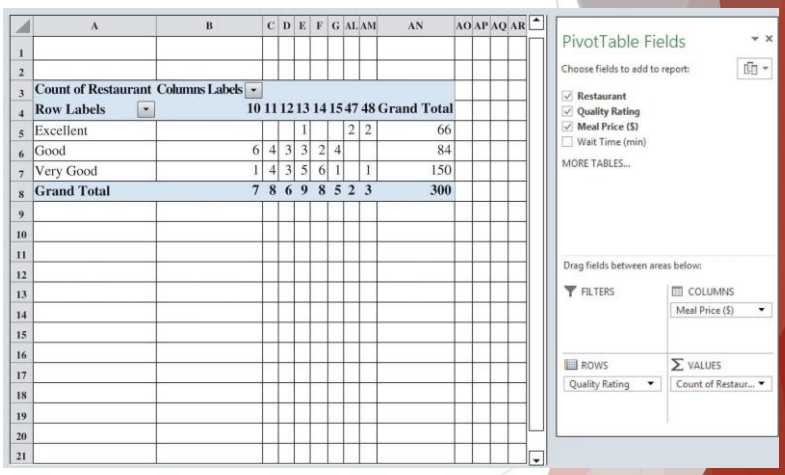
Tables

Figure 3.9: Initial PivotTable Field List and PivotTable Field Report for the Restaurant Data



Tables

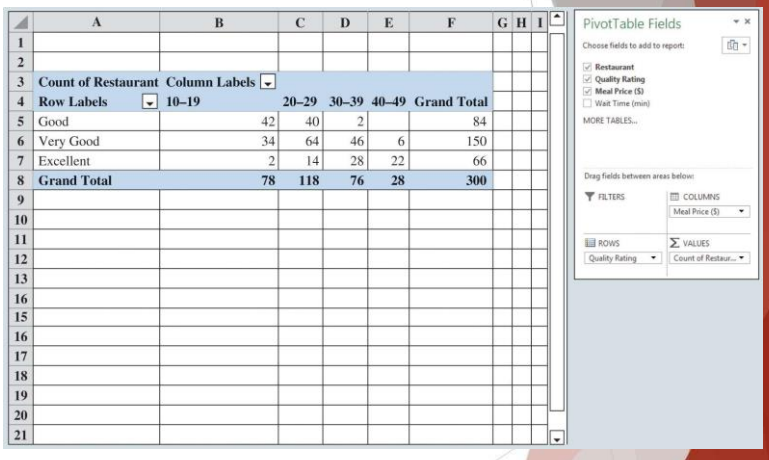
Figure 3.10: Completed PivotTable Field List and a Portion of the PivotTable Report for the Restaurant Data (Columns H:AK Are Hidden)



Count of Restaurant	Columns Labels							
Row Labels	10	11	12	13	14	15	47	48	Grand Total
Excellent				1			2	2	66
Good	6	4	3	3	2	4			84
Very Good	1	4	3	5	6	1		1	150
Grand Total	7	8	6	9	8	5	2	3	300

Tables

Figure 3.11: Final PivotTable Report for the Restaurant Data

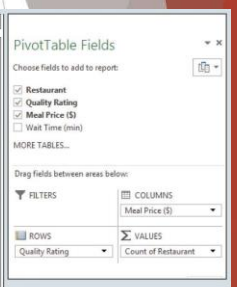


Count of Restaurant	Column Labels				
Row Labels	10–19	20–29	30–39	40–49	Grand Total
Good	42	40	2		84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Grand Total	78	118	76	28	300

Tables

Figure 3.12: Percent Frequency Distribution as a PivotTable for the Restaurant Data

	A	B	C	D	E	F	G
1							
2							
3	**Count of Restaurant**	**Column**					
4	**Row Labels**	**Labels 10–19**	**20–29**	**30–39**	**40–49**	**Grand Total**	
5	Good	14.00%	13.33%	0.67%	0.00%	28.00%	
6	Very Good	11.33%	21.33%	15.33%	2.00%	50.00%	
7	Excellent	0.67%	4.67%	9.33%	7.33%	22.00%	
8	**Grand Total**	**26.00%**	**39.33%**	**25.33%**	**9.33%**	**100.00%**	
9							
10							
11							
12							
13							
16							
15							
16							
17							
18							



Tables

Figure 3.13: PivotTable Report for the Restaurant Data with Average Wait Times Added

	A	B	C	D	E	F	G
1							
2							
3	**Average of Wait Time (min)**	**Column**					
4	**Row Labels**	**Labels 10–19**	**20–29**	**30–39**	**40–49**	**Grand Total**	
5	Good	2.6	2.5	0.5		2.5	
6	Very Good	12.6	12.6	12.0	10.0	12.3	
7	Excellent	25.5	29.1	34.0	32.3	32.1	
8	**Grand Total**	**7.6**	**11.1**	**19.8**	**27.5**	**13.9**	
9							
10							
11							
12							
13							
16							
15							
16							
17