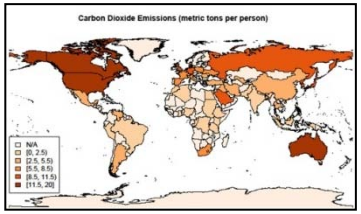
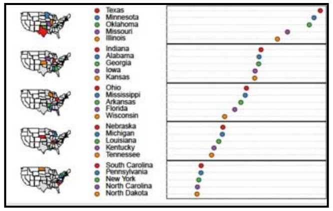
Ignoring the Spatial Context in Intro Stats Classes - And Some Simple Graphical Remedies

JÜRGEN SYMANZIK*, NATHAN VOGE,
UTAH STATE UNIVERSITY, LOGAN, UT, USA

*e-mail: symanzik@math.usu.edu,
nvoge@hotmail.com
Web: http://www.math.usu.edu/~symanzik



Content



- Context and Motivation
- Introduction
- Background
- Examples of Spatial Context and Simple Graphical Remedies
- Conclusion

Motivation

Can YOU identify the pattern?

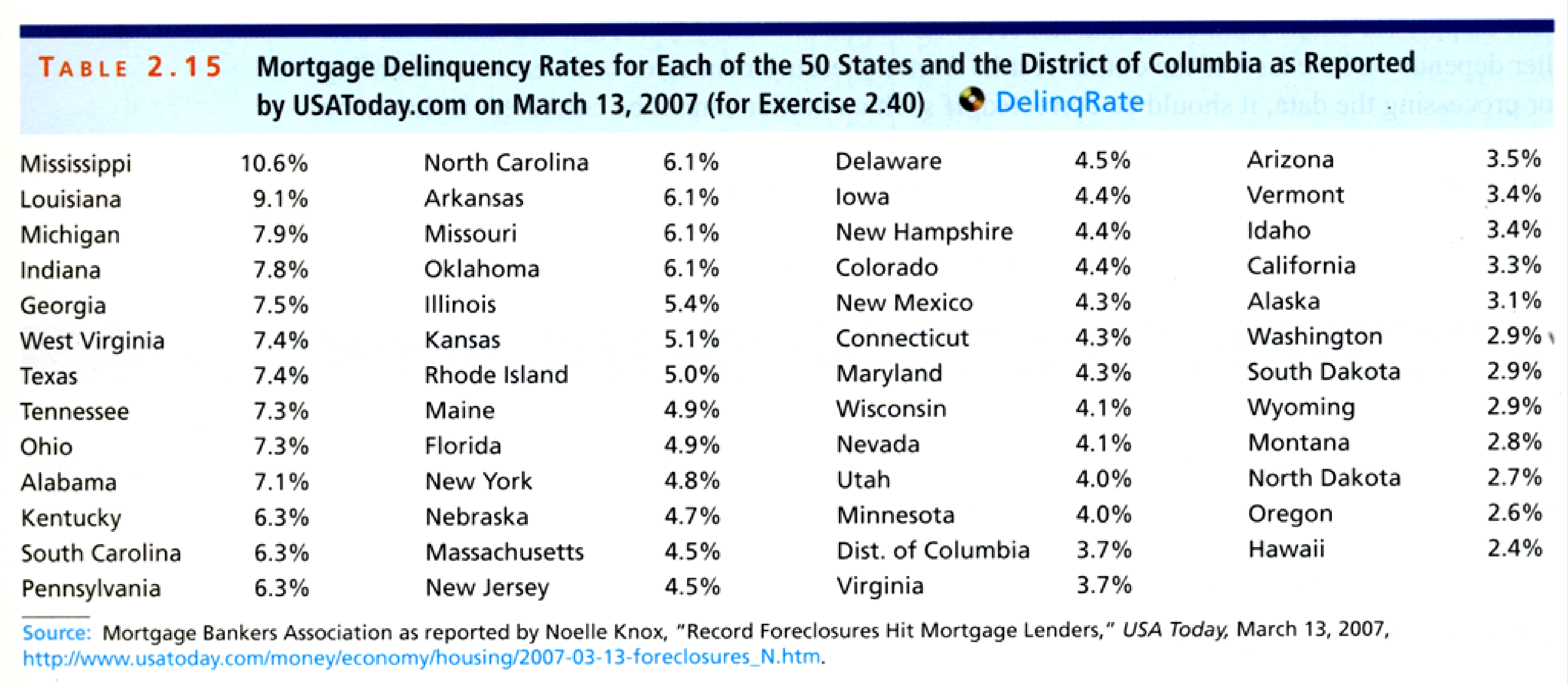


TABLE 2.15 Mortgage Delinquency Rates for Each of the 50 States and the District of Columbia as Reported by USAToday.com on March 13, 2007 (for Exercise 2.40) DelinqRate

Mississippi	10.6%	North Carolina	6.1%	Delaware	4.5%	Arizona	3.5%
Louisiana	9.1%	Arkansas	6.1%	Iowa	4.4%	Vermont	3.4%
Michigan	7.9%	Missouri	6.1%	New Hampshire	4.4%	Idaho	3.4%
Indiana	7.8%	Oklahoma	6.1%	Colorado	4.4%	California	3.3%
Georgia	7.5%	Illinois	5.4%	New Mexico	4.3%	Alaska	3.1%
West Virginia	7.4%	Kansas	5.1%	Connecticut	4.3%	Washington	2.9%
Texas	7.4%	Rhode Island	5.0%	Maryland	4.3%	South Dakota	2.9%
Tennessee	7.3%	Maine	4.9%	Wisconsin	4.1%	Wyoming	2.9%
Ohio	7.3%	Florida	4.9%	Nevada	4.1%	Montana	2.8%
Alabama	7.1%	New York	4.8%	Utah	4.0%	North Dakota	2.7%
Kentucky	6.3%	Nebraska	4.7%	Minnesota	4.0%	Oregon	2.6%
South Carolina	6.3%	Massachusetts	4.5%	Dist. of Columbia	3.7%	Hawaii	2.4%
Pennsylvania	6.3%	New Jersey	4.5%	Virginia	3.7%		

Source: Mortgage Bankers Association as reported by Noelle Knox, "Record Foreclosures Hit Mortgage Lenders," *USA Today*, March 13, 2007, http://www.usatoday.com/money/economy/housing/2007-03-13-foreclosures_N.htm.

Bowerman , O'Connell, Orris, Murphree, (2009) , p. 74

Motivation (cont.)

- Wainer (1997), states:

  “*The aim of good data graphics is to display data accurately and clearly*.” (p. 12)

- Ignoring spatial context lies within his 5th rule of displaying data badly, “*Graph data out of context*” (p. 25) and his 2nd rule“*Hide what data you do show*” (p. 16)
- Ignoring spatial context is a way to display data badly!

Introduction

- We will look at examples of spatial association in the four books used in introductory statistics classes at Utah State University

1. Probablity and Statistics for Engineers and Scientists

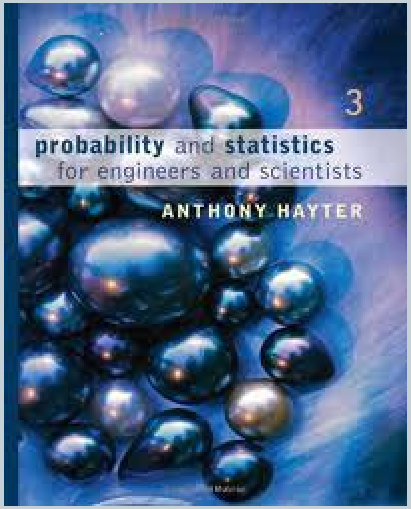
   By A. Hayter
2. Introduction to the Practice of Statistics

   By D. S. Moore, G. P. McCabe, and B. Craig
3. Essentials of Business Statistics

   By B. L. Bowerman, R. T. O'Connell, J. B. Orris, E. S. Murphree
4. Statistics

   By D. Freedman, R. Pisani, R. Purves

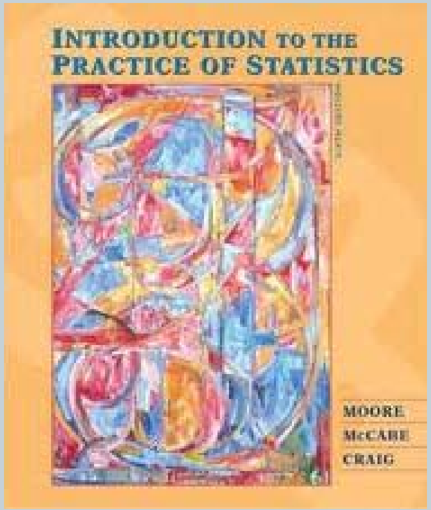
Introduction (cont.) - Textbooks



1. Probablity and Statistics for Engineers and Scientists; Third Edition; A. Hayter; Duxbury Press, Belmont, CA (2006)

- Did not contain any spatial data sets
- Examples include data from engineering, manufacturing, and medical backgrounds

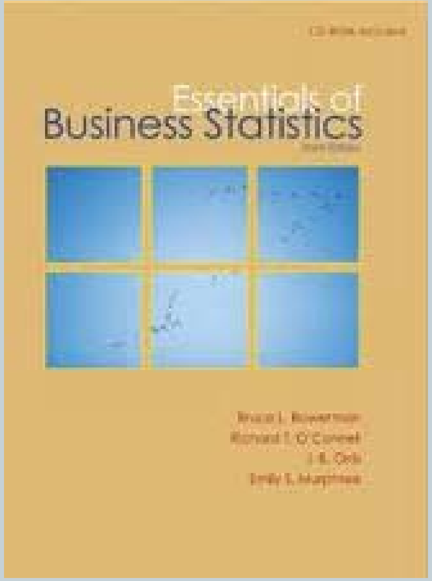
Introduction (cont.) - Textbooks



2. Introduction to the Practice of Statistics; Sixth Edition; D. S. Moore, G. P. McCabe, and B. Craig; W. H. Freeman, New York (2007)

- Contains seven major data sets with spatial components
- Examples include data from environmental, economic, social, and health backgrounds

Introduction (cont.) - Textbooks



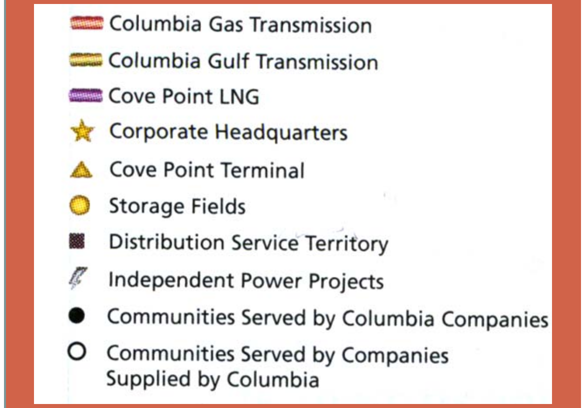
3. Essentials of Business Statistics; Third Edition; B. L. Bowerman, R. T. O'Connell, J. B. Orris, E. S. Murphree; McGraw-Hill/Irwin, New York (2009)

- Contains four major data sets with spatial components
- Examples include data from mainly economic and social backgrounds
- Contains a map

Introduction (cont.) - Textbooks

- Illustrates the pipelines of and the cities served by the Columbia Gas System

- Natural gas companies use prediction models when ordering to avoid transmission and storage fees aligned with over/under consumption



- 🟥 Columbia Gas Transmission
- 🟨 Columbia Gulf Transmission
- 🟪 Cove Point LNG
- ★ Corporate Headquarters
- ▲ Cove Point Terminal
- ○ Storage Fields
- ■ Distribution Service Territory
- Independent Power Projects
- ● Communities Served by Columbia Companies
- ○ Communities Served by Companies Supplied by Columbia

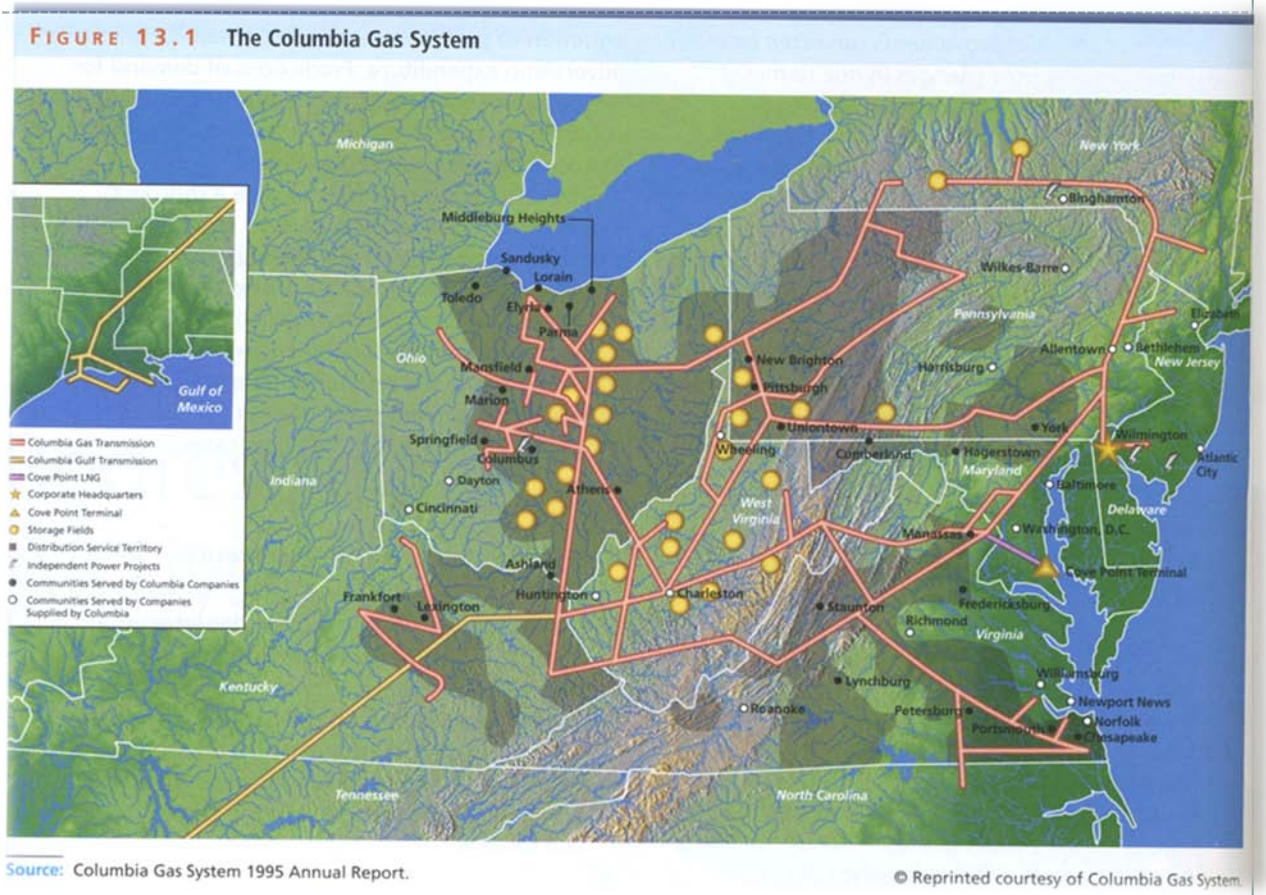


FIGURE 13.1 The Columbia Gas System

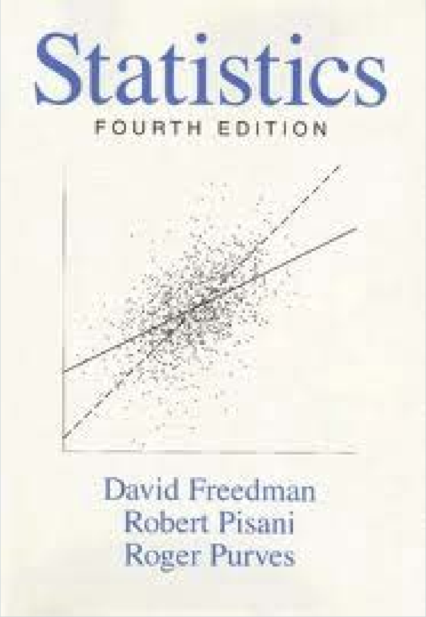
Source: Columbia Gas System 1995 Annual Report.

© Reprinted courtesy of Columbia Gas System.

Part 1: The natural gas transmission problem When the natural gas industry was deregulated in 1993, natural gas companies became responsible for acquiring the natural gas needed to heat the homes and businesses in the cities they serve. To do this, natural gas companies purchase natural gas from marketers (usually through long-term contracts) and periodically (perhaps daily, weekly, or monthly) place orders for natural gas to be transmitted by pipeline transmission systems to their cities. There are hundreds of pipeline transmission systems in the United States, and many of these systems supply a large number of cities. For instance, the map in Figure 13.1 illustrates the pipelines of and the cities served by the Columbia Gas System.

Bowerman , O’Connell, Orris, Murphree, (2009) , p. 74

Introduction (cont.) - Textbooks

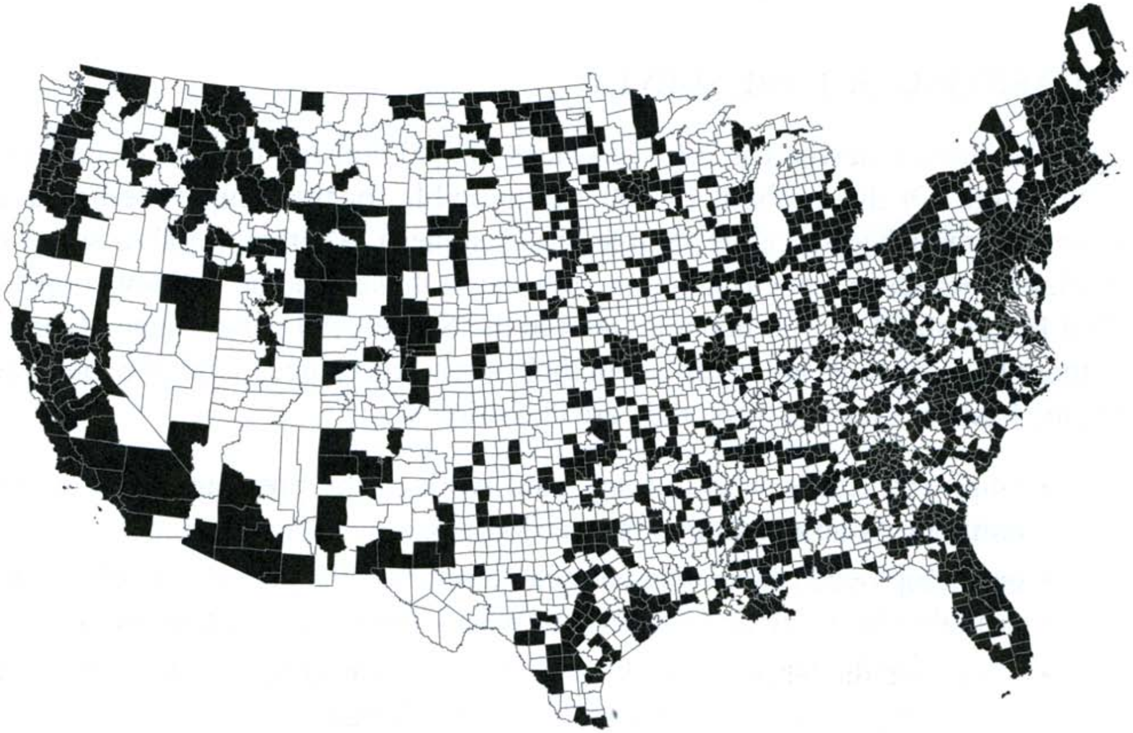


4. Statistics; Fourth Edition; D. Freedman, R. Pisani, R. Purves; W. W. Norton & Company, New York (2007)
   - Contains two major data sets with spatial components
   - Examples include data from mainly health related backgrounds
   - Contains a map

Introduction (cont.) - Textbooks

- Example of a map found in an introductory statistics class textbook

Figure 2. Primary Sampling Units for the Current Population Survey: the 1995 sample design with 792 PSUs.



Note: Alaska and Hawaii not shown.
Source: Bureau of the Census, Statistical Methods Division.

Freedman, Pisani, Purves (2007), p. 397

Background

- Maps showing spatial data make it easier to:

  - Know the data
  - Identify patterns
  - Identify outliers in space

Choropleth Map – Mortgages

- Numerically sorted list of Mortgage Delinquency Rates for all 50 states and D.C.
- Asked to describe the distribution and find outliers
- Create a stem and leaf plot

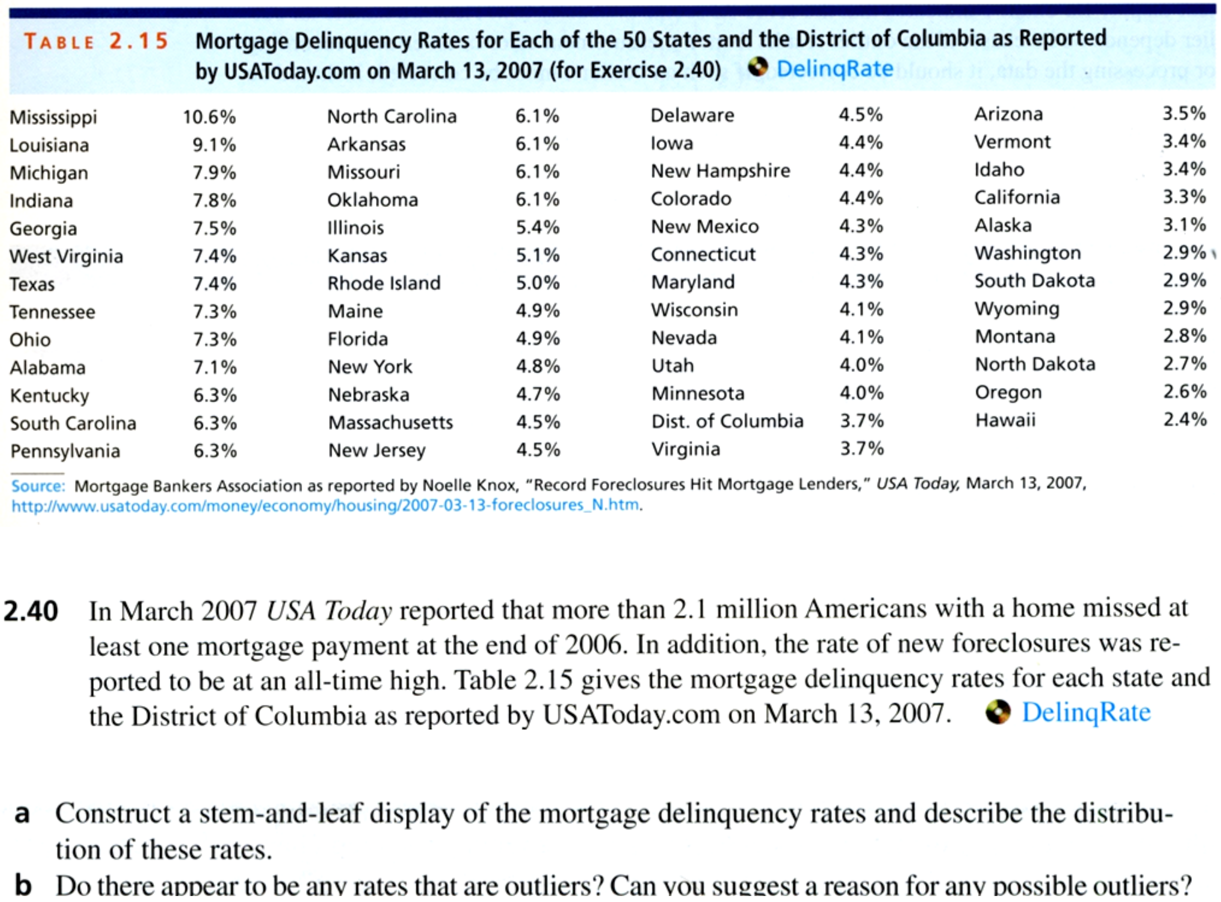


TABLE 2.15 Mortgage Delinquency Rates for Each of the 50 States and the District of Columbia as Reported by USAToday.com on March 13, 2007 (for Exercise 2.40) DelinqRate

Mississippi	10.6%	North Carolina	6.1%	Delaware	4.5%	Arizona	3.5%
Louisiana	9.1%	Arkansas	6.1%	Iowa	4.4%	Vermont	3.4%
Michigan	7.9%	Missouri	6.1%	New Hampshire	4.4%	Idaho	3.4%
Indiana	7.8%	Oklahoma	6.1%	Colorado	4.4%	California	3.3%
Georgia	7.5%	Illinois	5.4%	New Mexico	4.3%	Alaska	3.1%
West Virginia	7.4%	Kansas	5.1%	Connecticut	4.3%	Washington	2.9%
Texas	7.4%	Rhode Island	5.0%	Maryland	4.3%	South Dakota	2.9%
Tennessee	7.3%	Maine	4.9%	Wisconsin	4.1%	Wyoming	2.9%
Ohio	7.3%	Florida	4.9%	Nevada	4.1%	Montana	2.8%
Alabama	7.1%	New York	4.8%	Utah	4.0%	North Dakota	2.7%
Kentucky	6.3%	Nebraska	4.7%	Minnesota	4.0%	Oregon	2.6%
South Carolina	6.3%	Massachusetts	4.5%	Dist. of Columbia	3.7%	Hawaii	2.4%
Pennsylvania	6.3%	New Jersey	4.5%	Virginia	3.7%		

Source: Mortgage Bankers Association as reported by Noelle Knox, "Record Foreclosures Hit Mortgage Lenders," *USA Today*, March 13, 2007, http://www.usatoday.com/money/economy/housing/2007-03-13-foreclosures_N.htm.

2.40 In March 2007 *USA Today* reported that more than 2.1 million Americans with a home missed at least one mortgage payment at the end of 2006. In addition, the rate of new foreclosures was reported to be at an all-time high. Table 2.15 gives the mortgage delinquency rates for each state and the District of Columbia as reported by USAToday.com on March 13, 2007. DelinqRate

a Construct a stem-and-leaf display of the mortgage delinquency rates and describe the distribution of these rates.

b Do there appear to be any rates that are outliers? Can you suggest a reason for any possible outliers?

Bowerman , O'Connell, Orris, Murphree, (2009) , p. 74

Choropleth Map – Mortgages

- Skewed
- Likely couple of outliers
- Possibility of 2-3 modes
- No consistent pattern

Stem and Leaf Plot of Mortgage Delinquency Rates

- The decimal point is at the |

```
 2 | 4678999
 3 | 134457
 4 | 00113334445557899
 5 | 014
 6 | 1111333
 7 | 13344589
 8 |
 9 | 1
10 | 6
```

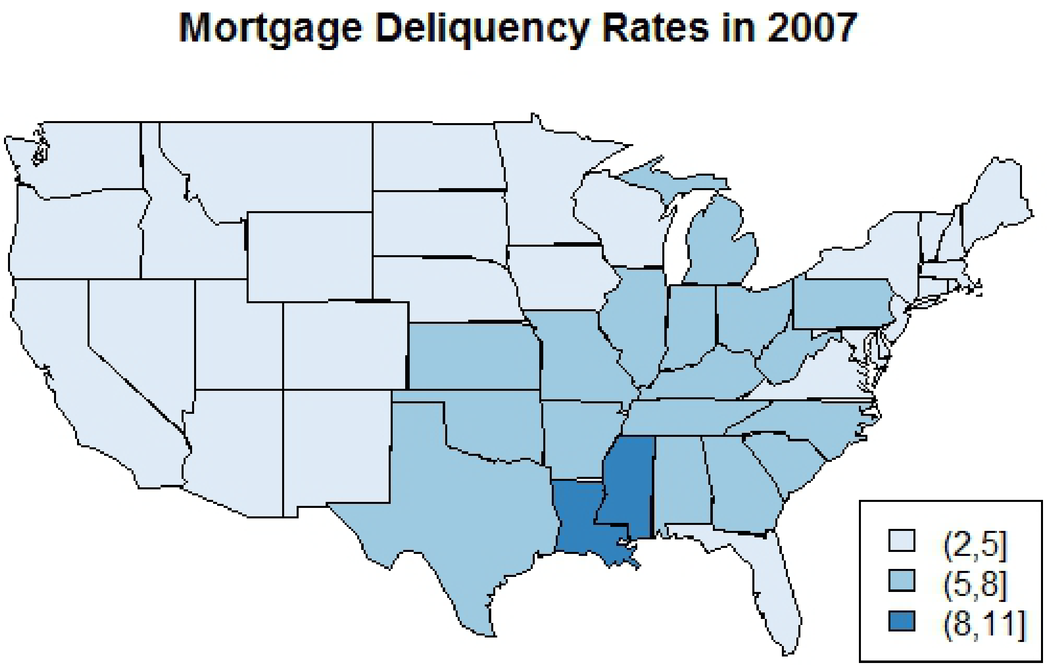
Choropleth Map – Mortgages

- Symanzik & Carr (2008), p. 270, state:

“Choropleth maps use the color or shading of regions in a map to represent region values.”

- Several different breaks may need to be considered

- Makes it clear south-eastern states have higher percentage of foreclosures



*D.C., Alaska, Hawaii not on map.

R code - Mortgages

- Maps can be created fairly simply
  - Only about 10 lines of code in this example

```
# load required packages
library(RColorBrewer)
library(maps)

# read in data
## Note, D.C was in our data set, but not on the map we are using
## so it is deleted from the data set
mort = read.csv(“DelinqRate_edit.csv”, header=T)

# set how to divide up the data
breaks = c(2,5,8,11)
```

R code - Mortgages

- Maps can be created fairly simply

```
# apply the breaks to the data set
m.class = cut(mort[,2], breaks)

# pick a color pallette
brewer.pal(3, "Blues")

# assign colors to breaks
m.col = brewer.pal(3, "Blues")[m.class]

# save map as a .pdf
pdf("Mortagages_Map3.1.pdf", width = 11, paper = "USr")
```

R code - Mortgages

- Maps can be created fairly simply

```
# match states in map to states in the data set
## Note, states in the map are “characters”, so we must make the
## states in our data set “characters” as well
map.m.col = m.col[match.map("state", as.character(mort[,1]))]

# create the map
map("state", fill = T, col = map.m.col)
legend("bottomright", legend = levels(m.class), fill = brewer.pal(3, "Blues"))
title("Mortgage Deliquency Rates in 2007")
dev.off()
```

Choropleth Map - CO2

- CO2 emissions from countries with population at least 20 million
- Sorted alphabetically
- Asked to describe the distribution of the data and to identify outliers

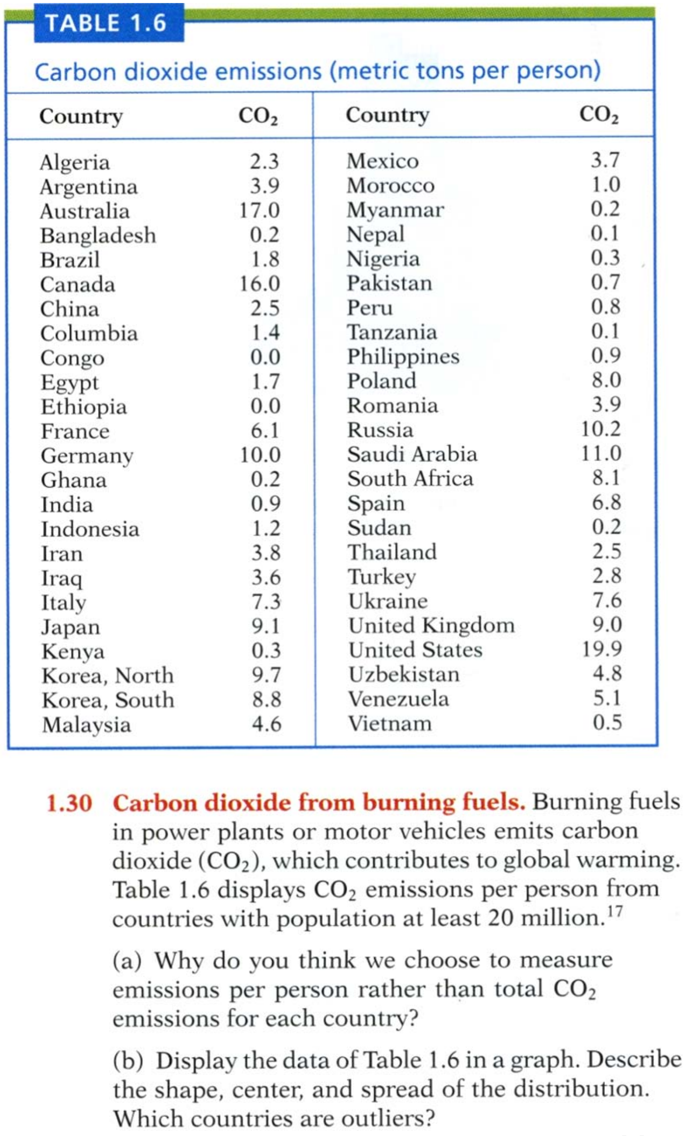


TABLE 1.6

Carbon dioxide emissions (metric tons per person)

Country	CO_2	Country	CO_2
Algeria	2.3	Mexico	3.7
Argentina	3.9	Morocco	1.0
Australia	17.0	Myanmar	0.2
Bangladesh	0.2	Nepal	0.1
Brazil	1.8	Nigeria	0.3
Canada	16.0	Pakistan	0.7
China	2.5	Peru	0.8
Columbia	1.4	Tanzania	0.1
Congo	0.0	Philippines	0.9
Egypt	1.7	Poland	8.0
Ethiopia	0.0	Romania	3.9
France	6.1	Russia	10.2
Germany	10.0	Saudi Arabia	11.0
Ghana	0.2	South Africa	8.1
India	0.9	Spain	6.8
Indonesia	1.2	Sudan	0.2
Iran	3.8	Thailand	2.5
Iraq	3.6	Turkey	2.8
Italy	7.3	Ukraine	7.6
Japan	9.1	United Kingdom	9.0
Kenya	0.3	United States	19.9
Korea, North	9.7	Uzbekistan	4.8
Korea, South	8.8	Venezuela	5.1
Malaysia	4.6	Vietnam	0.5

1.30 **Carbon dioxide from burning fuels.** Burning fuels in power plants or motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. Table 1.6 displays CO_2 emissions per person from countries with population at least 20 million.[17]

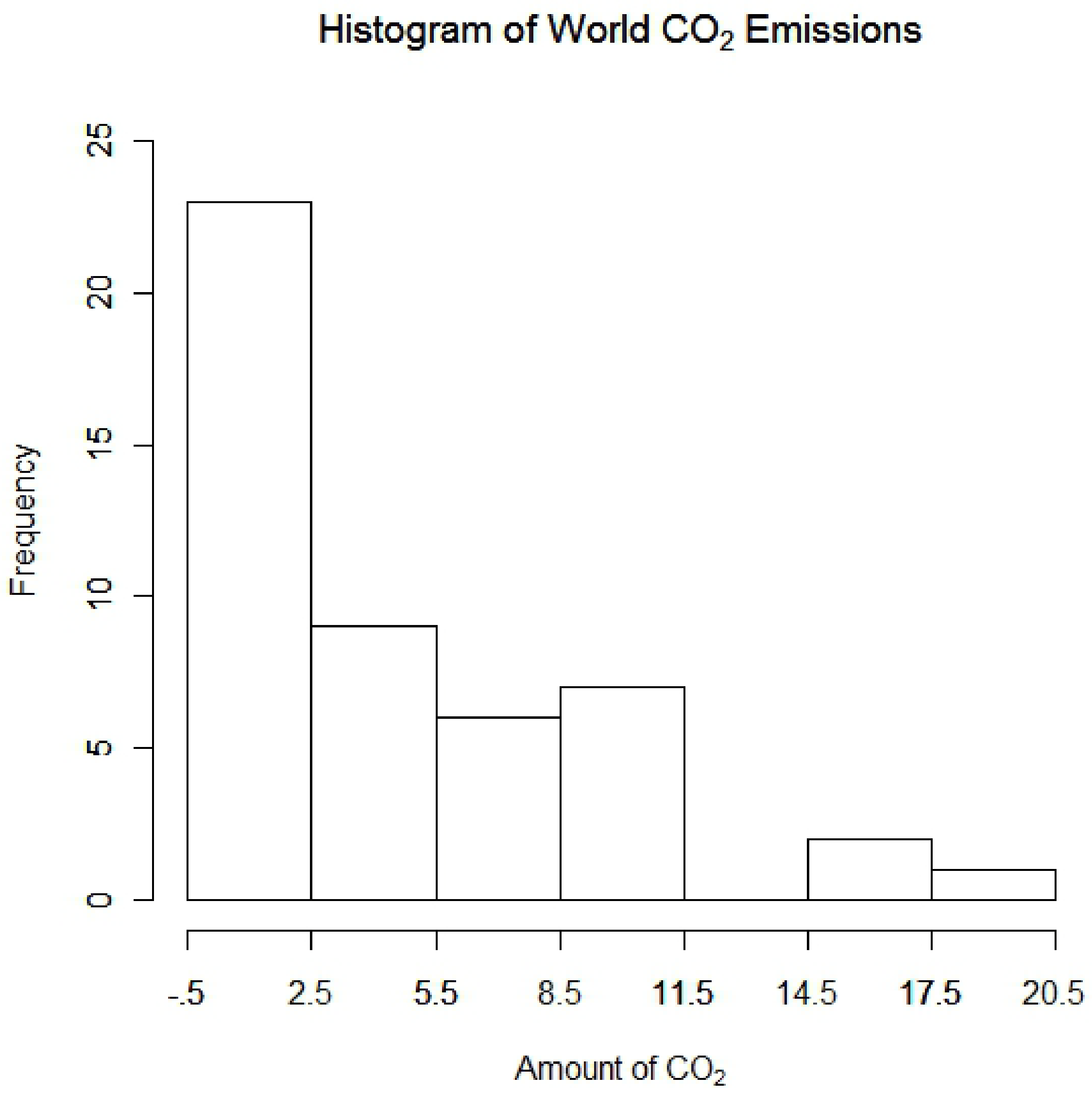
(a) Why do you think we choose to measure emissions per person rather than total CO_2 emissions for each country?

(b) Display the data of Table 1.6 in a graph. Describe the shape, center, and spread of the distribution. Which countries are outliers?

Moore, McCabe, Craig (2007), p. 26

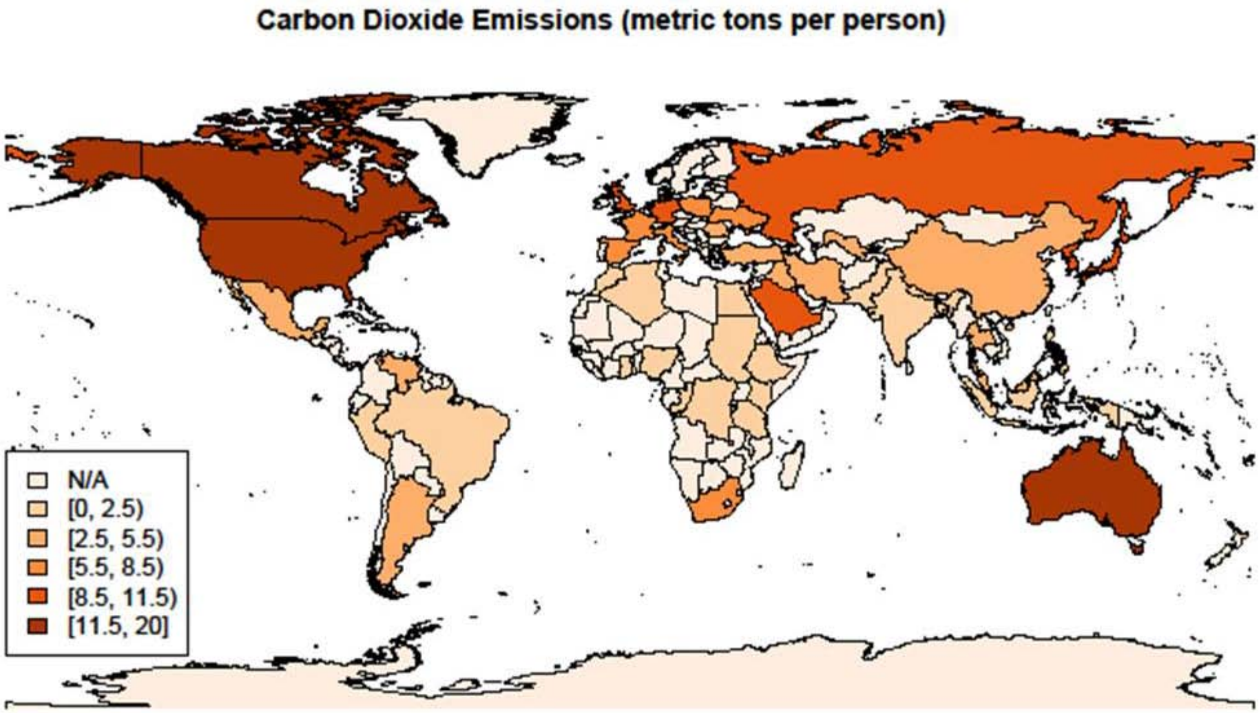
Choropleth Map - CO2

- Histogram tells us the data are right-skewed
- There are possibly three outliers
- Median: 3.200
- Mean: 4.596
- SD: 4.822



Choropleth Map - CO2

- Able to see spatial outliers and grouping easily



Google Map - Tuition

- Comparing in-state tuition and fees for 32 universities between the year 2000 and 2005

- Asked to plot data and describe relationship

- Asked to identify outliers and obtain the residuals

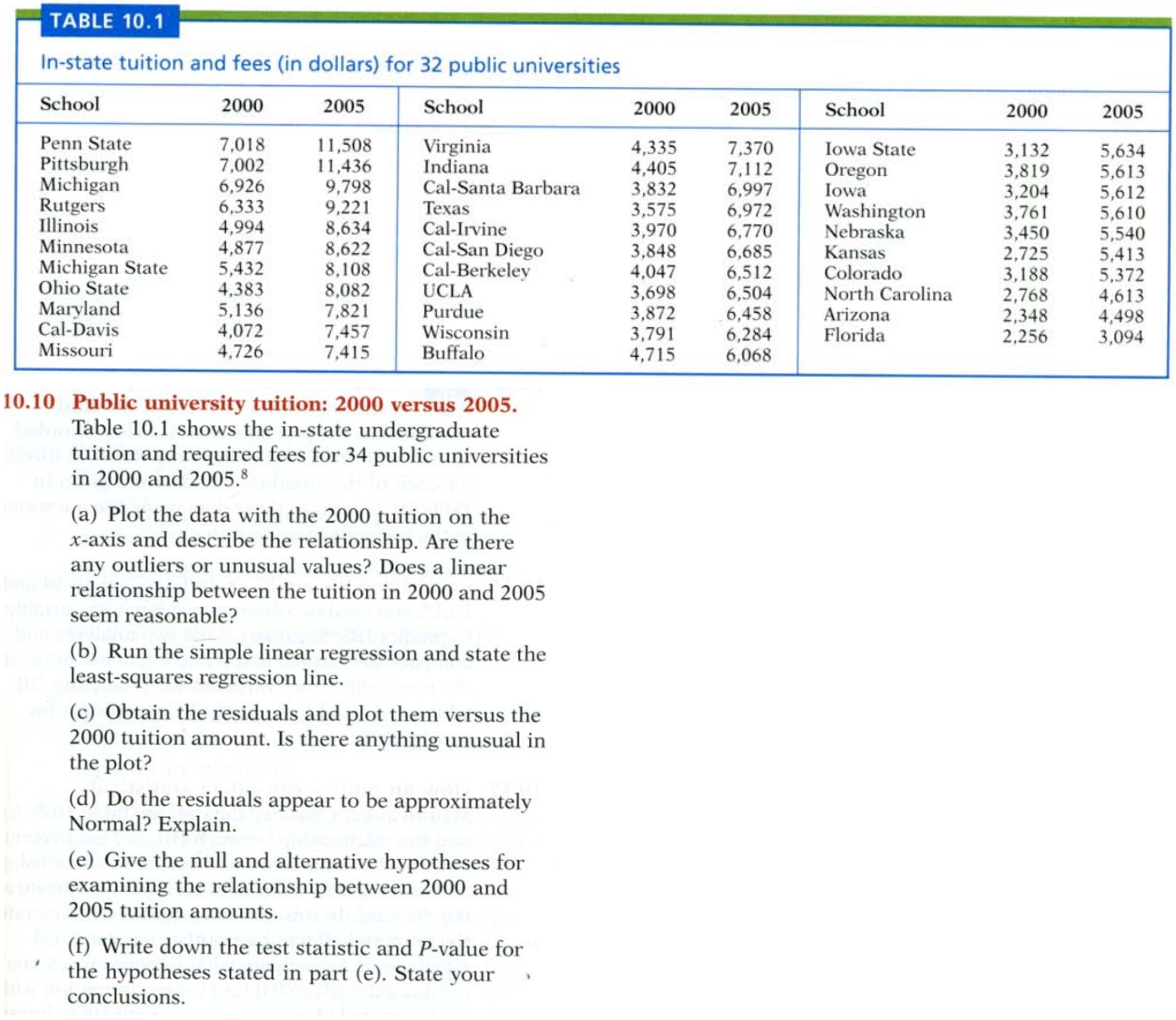


TABLE 10.1

In-state tuition and fees (in dollars) for 32 public universities

School	2000	2005	School	2000	2005	School	2000	2005
Penn State	7,018	11,508	Virginia	4,335	7,370	Iowa State	3,132	5,634
Pittsburgh	7,002	11,436	Indiana	4,405	7,112	Oregon	3,819	5,613
Michigan	6,926	9,798	Cal-Santa Barbara	3,832	6,997	Iowa	3,204	5,612
Rutgers	6,333	9,221	Texas	3,575	6,972	Washington	3,761	5,610
Illinois	4,994	8,634	Cal-Irvine	3,970	6,770	Nebraska	3,450	5,540
Minnesota	4,877	8,622	Cal-San Diego	3,848	6,685	Kansas	2,725	5,413
Michigan State	5,432	8,108	Cal-Berkeley	4,047	6,512	Colorado	3,188	5,372
Ohio State	4,383	8,082	UCLA	3,698	6,504	North Carolina	2,768	4,613
Maryland	5,136	7,821	Purdue	3,872	6,458	Arizona	2,348	4,498
Cal-Davis	4,072	7,457	Wisconsin	3,791	6,284	Florida	2,256	3,094
Missouri	4,726	7,415	Buffalo	4,715	6,068			

10.10 **Public university tuition: 2000 versus 2005.** Table 10.1 shows the in-state undergraduate tuition and required fees for 34 public universities in 2000 and 2005.[8]

(a) Plot the data with the 2000 tuition on the *x*-axis and describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the tuition in 2000 and 2005 seem reasonable?

(b) Run the simple linear regression and state the least-squares regression line.

(c) Obtain the residuals and plot them versus the 2000 tuition amount. Is there anything unusual in the plot?

(d) Do the residuals appear to be approximately Normal? Explain.

(e) Give the null and alternative hypotheses for examining the relationship between 2000 and 2005 tuition amounts.

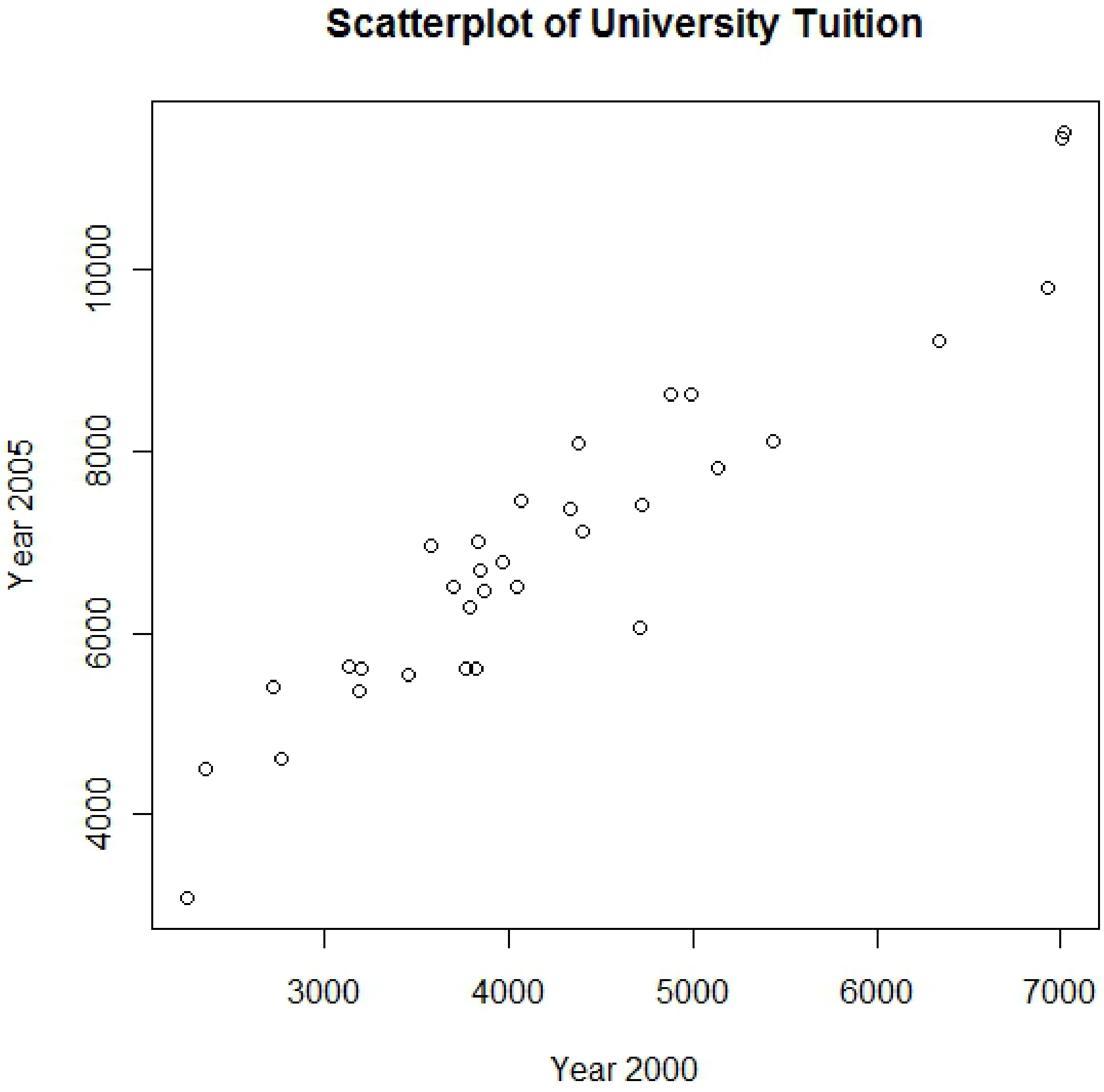
(f) Write down the test statistic and *P*-value for the hypotheses stated in part (e). State your conclusions.

Moore, McCabe, Craig (2007), p. 26

Google Map - Tuition

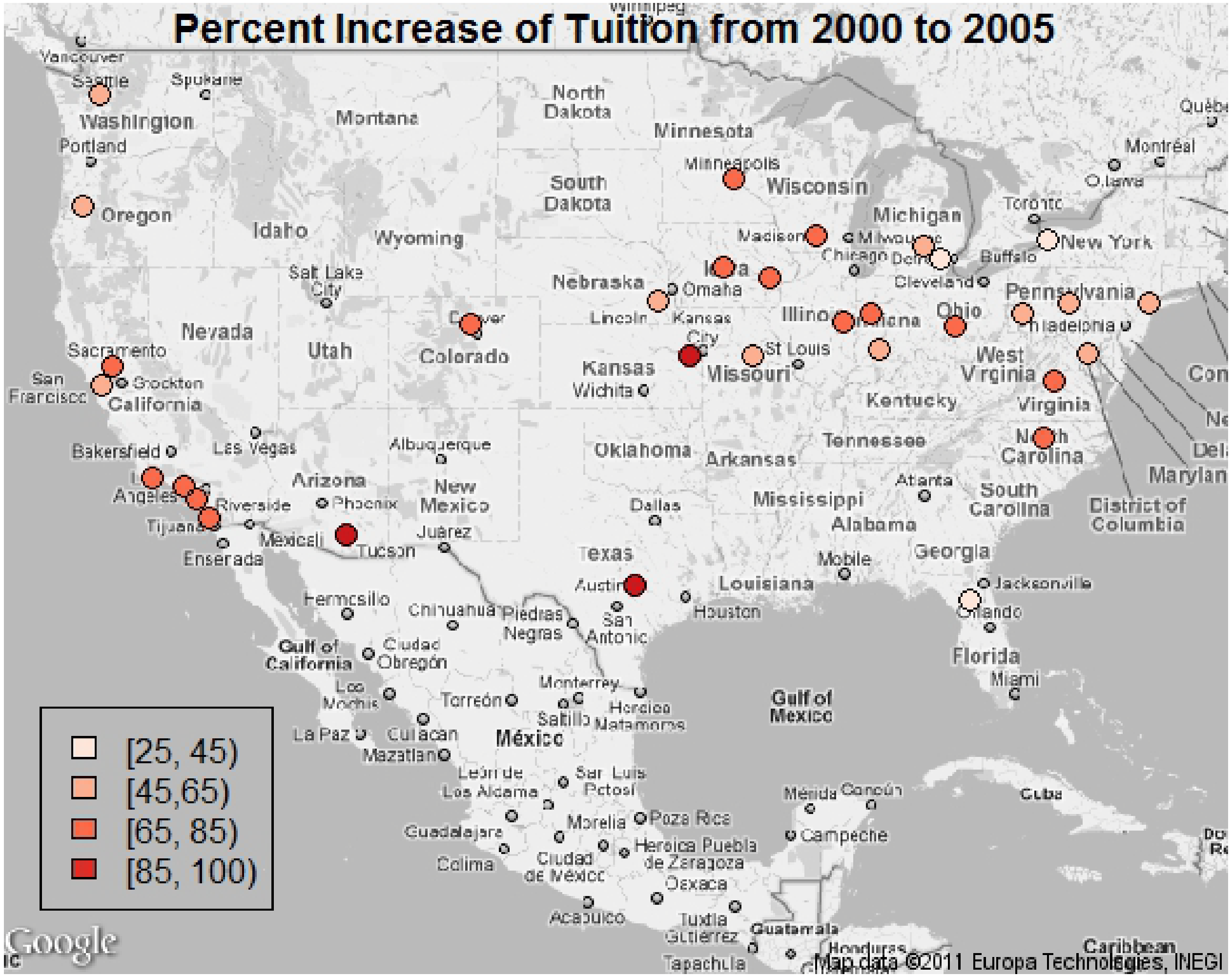
- Linear Regression Model:

$Yr_{2005} = 1059 + (1.4)*Yr_{2000}$



Google Map - Tuition

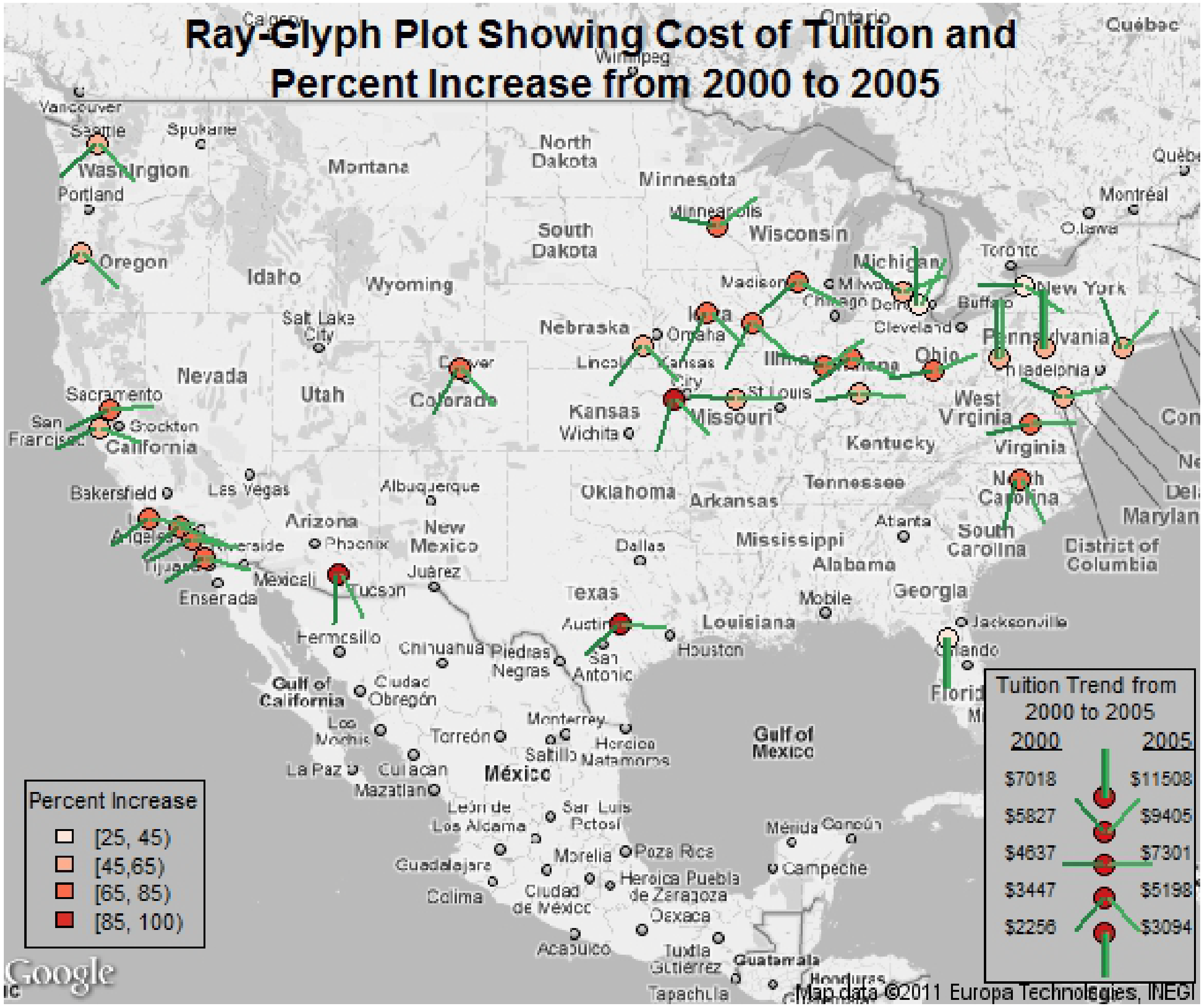
- Data with latitude and longitude points can be represented on a Google Map
- Makes more sense to look at the percent increase for each university from 2000 to 2005
- Able to see grouping
- We can identify outliers



Percent Increase of Tuition from 2000 to 2005

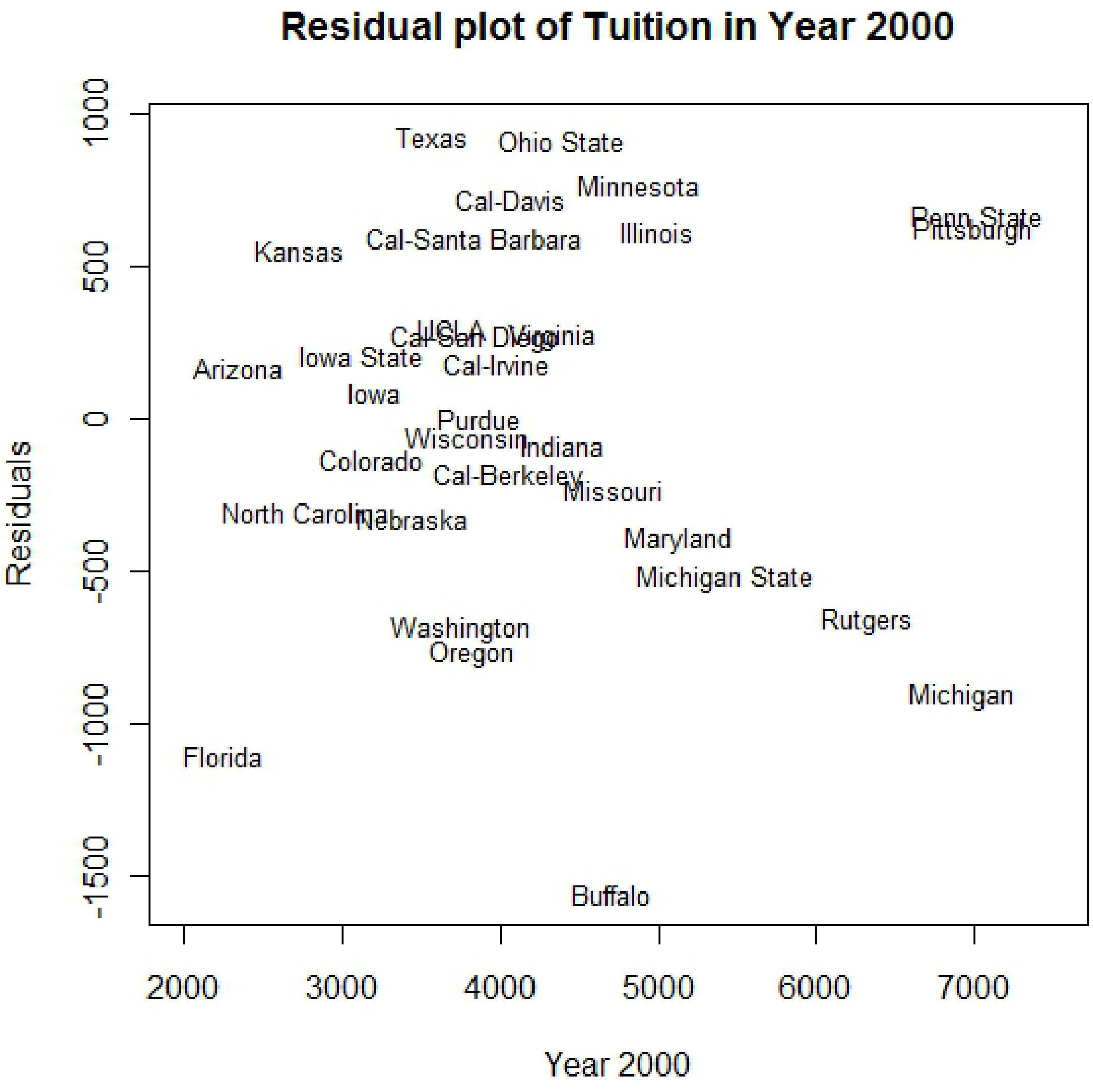
Google Map - Tuition

- If we add a Ray-Glyph plot we can look at multiple variables on one map
- The green lines represent the tuition for 2000 and 2005 respectively with a line pointing down as the lowest and up as the highest tuition in that year
- Makes it easy to see outliers and identify trends



Ray-Glyph Plot Showing Cost of Tuition and Percent Increase from 2000 to 2005

Google Map - Tuition



Micromaps - Tornado

- From 1950 to 1999, adjusted for inflation
- Table is sorted alphabetically
- Asked to identify the top 5 and bottom 5 states
- Asked to make a histogram using software with classes in increments of 10
- Asked to identify outliers

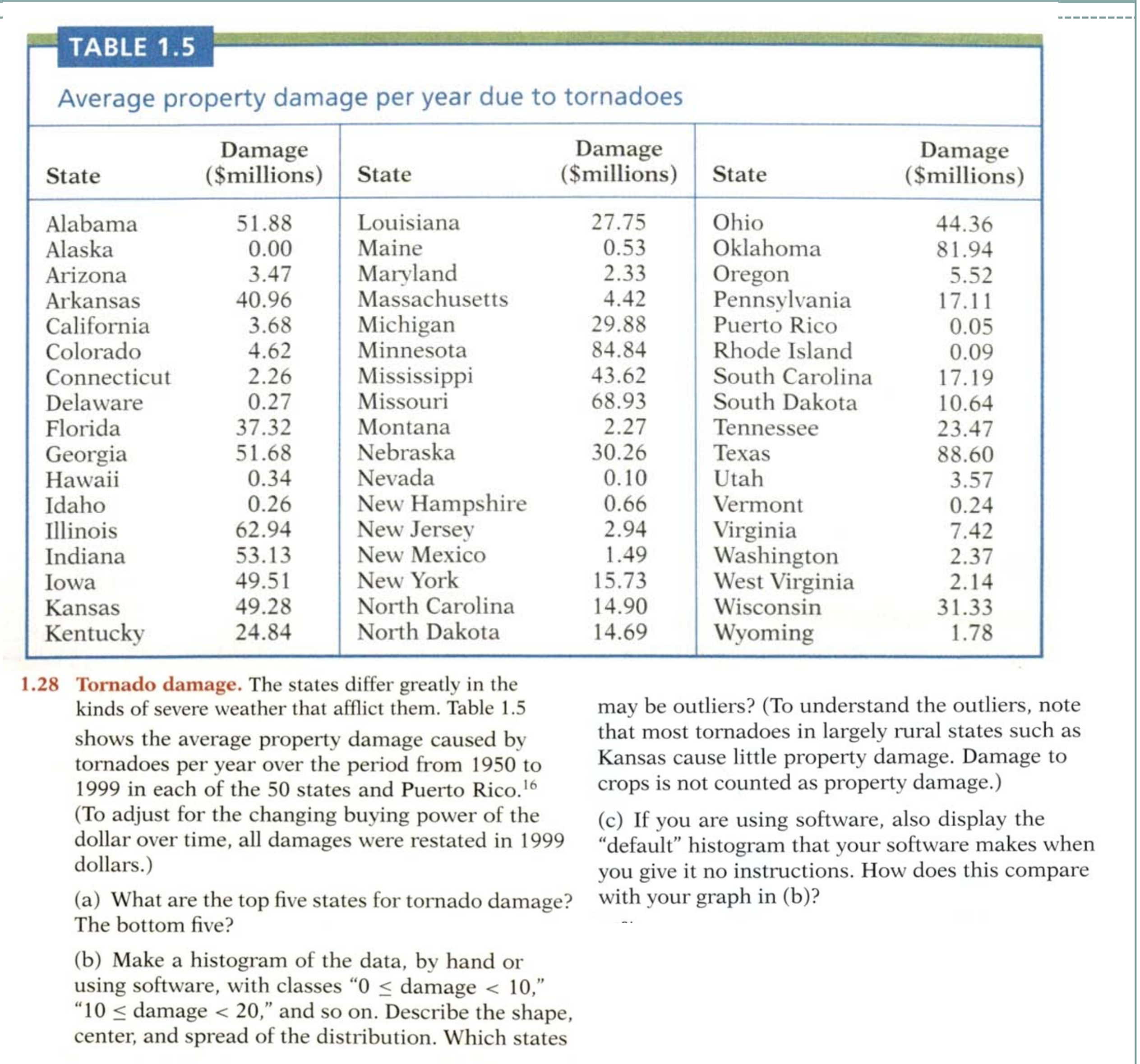


TABLE 1.5

Average property damage per year due to tornadoes

State	Damage ($millions)	State	Damage ($millions)	State	Damage ($millions)
Alabama	51.88	Louisiana	27.75	Ohio	44.36
Alaska	0.00	Maine	0.53	Oklahoma	81.94
Arizona	3.47	Maryland	2.33	Oregon	5.52
Arkansas	40.96	Massachusetts	4.42	Pennsylvania	17.11
California	3.68	Michigan	29.88	Puerto Rico	0.05
Colorado	4.62	Minnesota	84.84	Rhode Island	0.09
Connecticut	2.26	Mississippi	43.62	South Carolina	17.19
Delaware	0.27	Missouri	68.93	South Dakota	10.64
Florida	37.32	Montana	2.27	Tennessee	23.47
Georgia	51.68	Nebraska	30.26	Texas	88.60
Hawaii	0.34	Nevada	0.10	Utah	3.57
Idaho	0.26	New Hampshire	0.66	Vermont	0.24
Illinois	62.94	New Jersey	2.94	Virginia	7.42
Indiana	53.13	New Mexico	1.49	Washington	2.37
Iowa	49.51	New York	15.73	West Virginia	2.14
Kansas	49.28	North Carolina	14.90	Wisconsin	31.33
Kentucky	24.84	North Dakota	14.69	Wyoming	1.78

1.28 Tornado damage. The states differ greatly in the kinds of severe weather that afflict them. Table 1.5 shows the average property damage caused by tornadoes per year over the period from 1950 to 1999 in each of the 50 states and Puerto Rico.[16] (To adjust for the changing buying power of the dollar over time, all damages were restated in 1999 dollars.)

(a) What are the top five states for tornado damage? The bottom five?

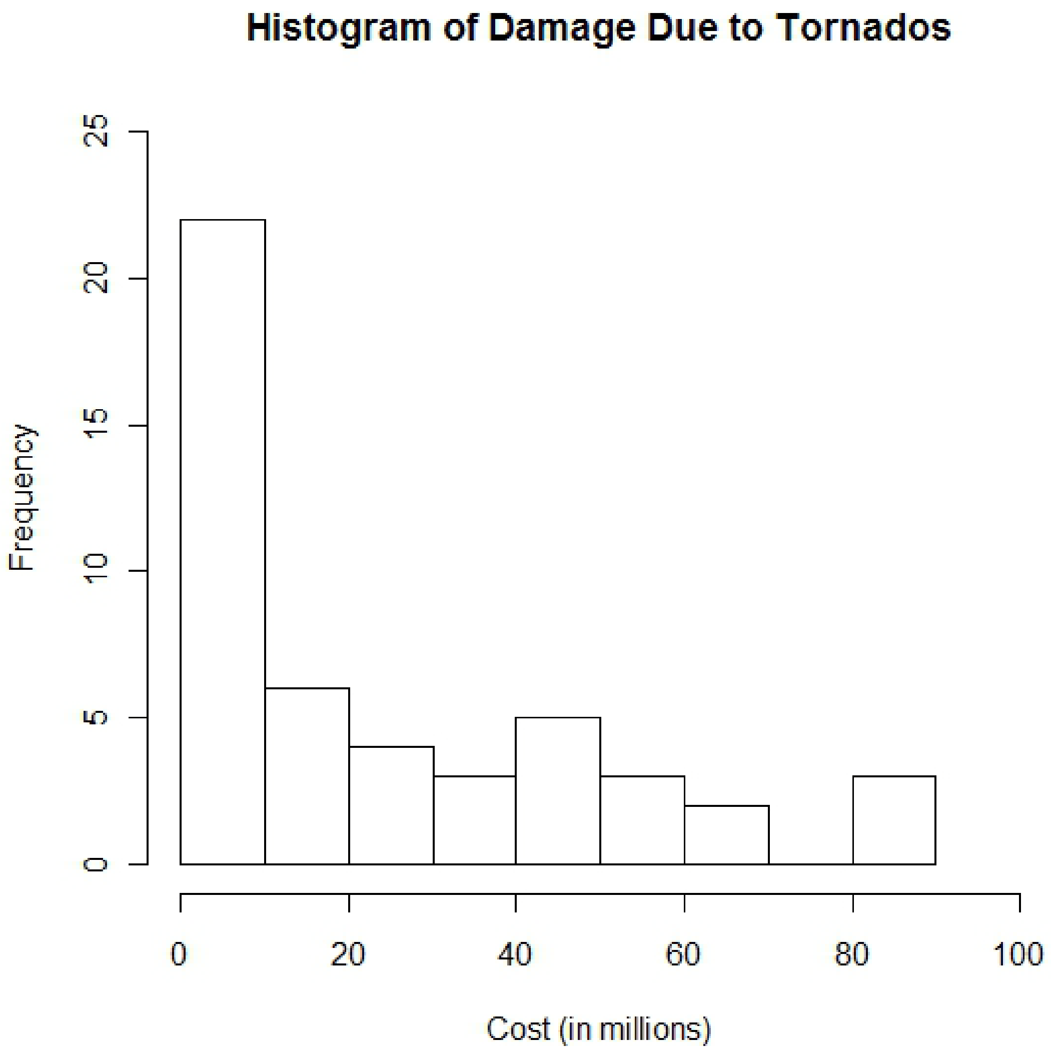
(b) Make a histogram of the data, by hand or using software, with classes "$0 \le$ damage < 10," "$10 \le$ damage < 20," and so on. Describe the shape, center, and spread of the distribution. Which states may be outliers? (To understand the outliers, note that most tornadoes in largely rural states such as Kansas cause little property damage. Damage to crops is not counted as property damage.)

(c) If you are using software, also display the "default" histogram that your software makes when you give it no instructions. How does this compare with your graph in (b)?

Moore, McCabe, Craig (2007), p. 26

Micromaps - Tornado

- Histogram tells us the data are right-skewed
- There are possibly three outliers



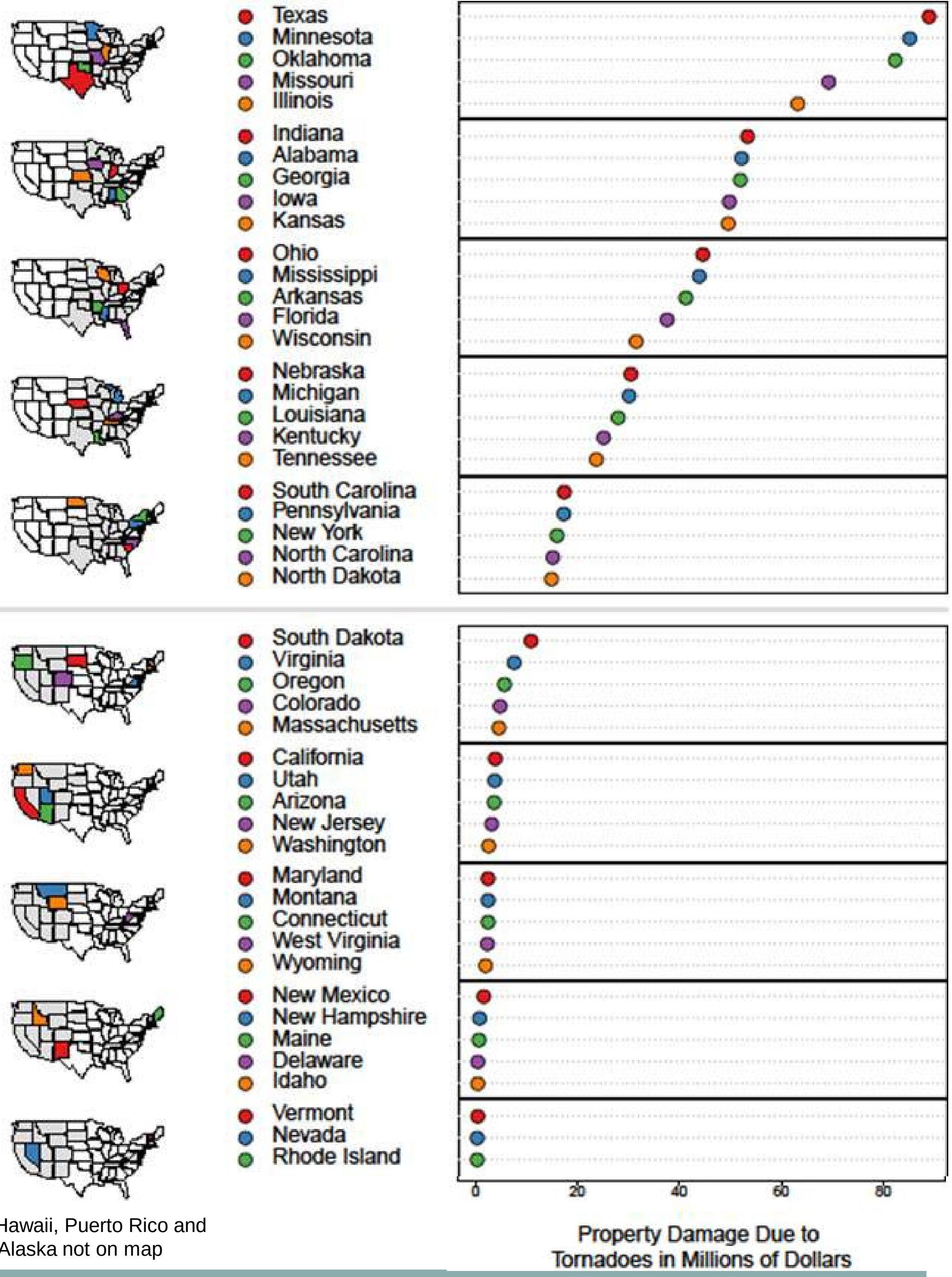
Micromaps - Tornado

- Able to visually see where and how the data are distributed

- Able to easily indentify the top and bottom 5 states and see where they are located

- Helps see outliers

- For continental U.S. only



*Hawaii, Puerto Rico and Alaska not on map

Conclusion

- Displaying spatial data enriches understanding
- Outliers can be explained by spatial context
- Maps can reveal spatial outliers
- Textbook authors are encouraged to include maps
- Students are encouraged to create maps to answer textbook questions

Software Used

- R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
- Packages used in R: RColorBrewer, maps, RgoogleMaps, PBSmapping, maptools