Statistic for Business

Week 1-2

Collecting, Organizing and Visualizing Data

Agenda

Time	Activity
90 minutes	Collecting and Organizing Data
60 minutes	Break
90 minutes	Visualizing Data

Objectives

By the end of this class, students will:

- Understand how to collect data in statistic
- Be able to organize categorical and numerical data
- Understand how to read and interpret an organized data (table)
- Be able to visualize categorical and numerical data
- Understand how to make conclusion based on the data visualizations (charts and graphs)

REVIEW

1.4

For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the measurement scale.

a. Number of telephones per household
b. Length (in minutes) of the longest telephone call made in a month
c. Whether someone in the household owns aWi-Fi-capable cell phone
d. Whether there is a high-speed Internet connection in the household

1.20

In 2008, a university in the midwestern United States surveyed its full-time first-year students after they completed their first semester. Surveys were electronically distributed to all 3,727 students, and responses were obtained from 2,821 students. Of the students surveyed, 90.1% indicated that they had studied with other students, and 57.1% indicated that they had tutored another student. The report also noted that 61.3% of the students surveyed came to class late at least once, and 45.8% admitted to being bored in class at least once.

a. Describe the population of interest.

b. Describe the sample that was collected.

COLLECTING AND ORGANIZING DATA

Content

Data Collection

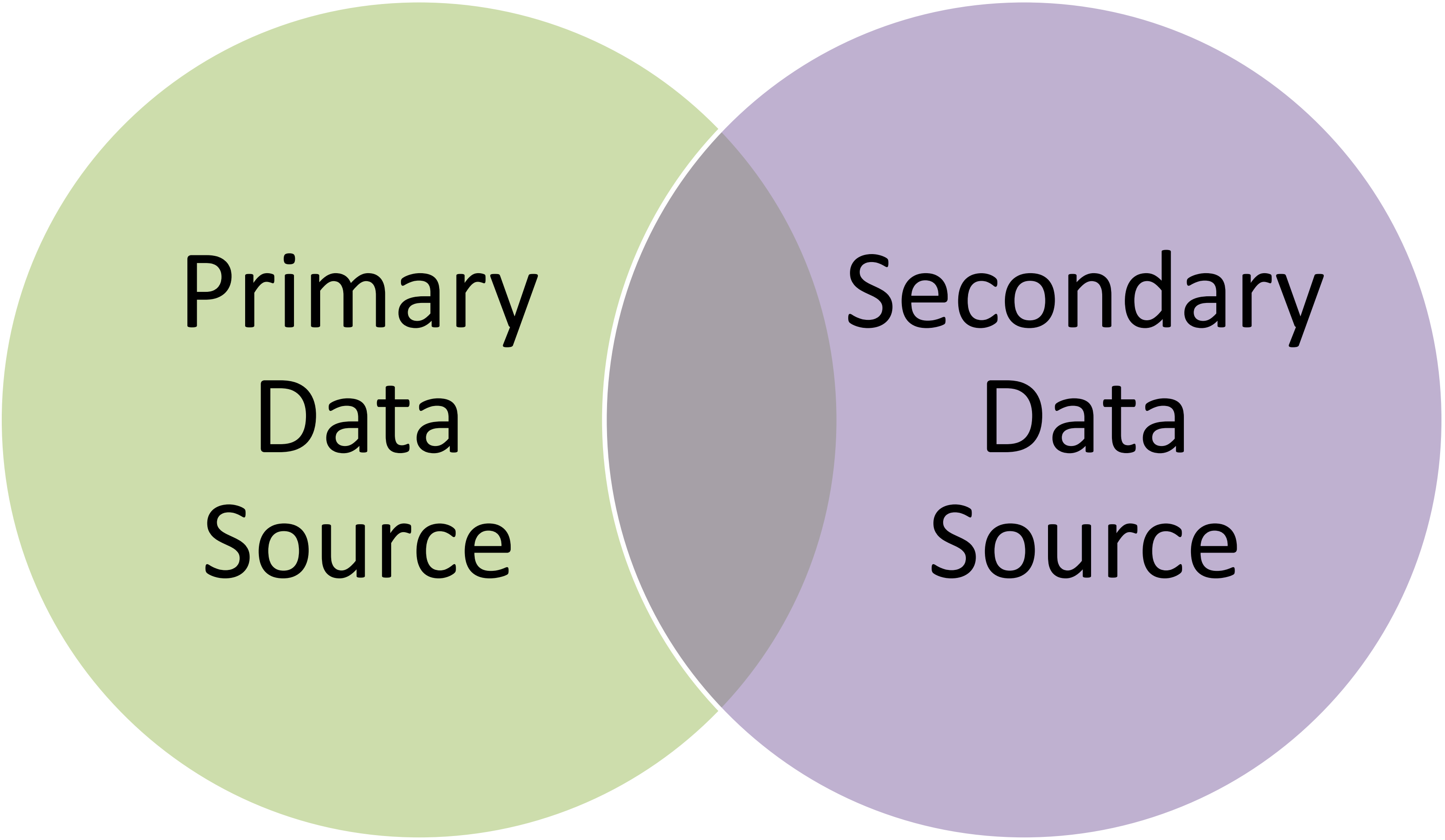
Organizing Data

- Categorical Data
- Numerical Data

Visualizing Data

- Categorical Data
- Numerical Data
- Two Numerical Data

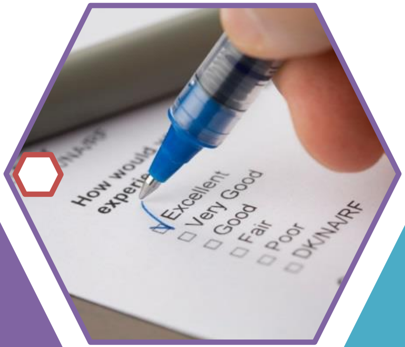
Data Collection



Primary Data Source

Secondary Data Source

Data Source



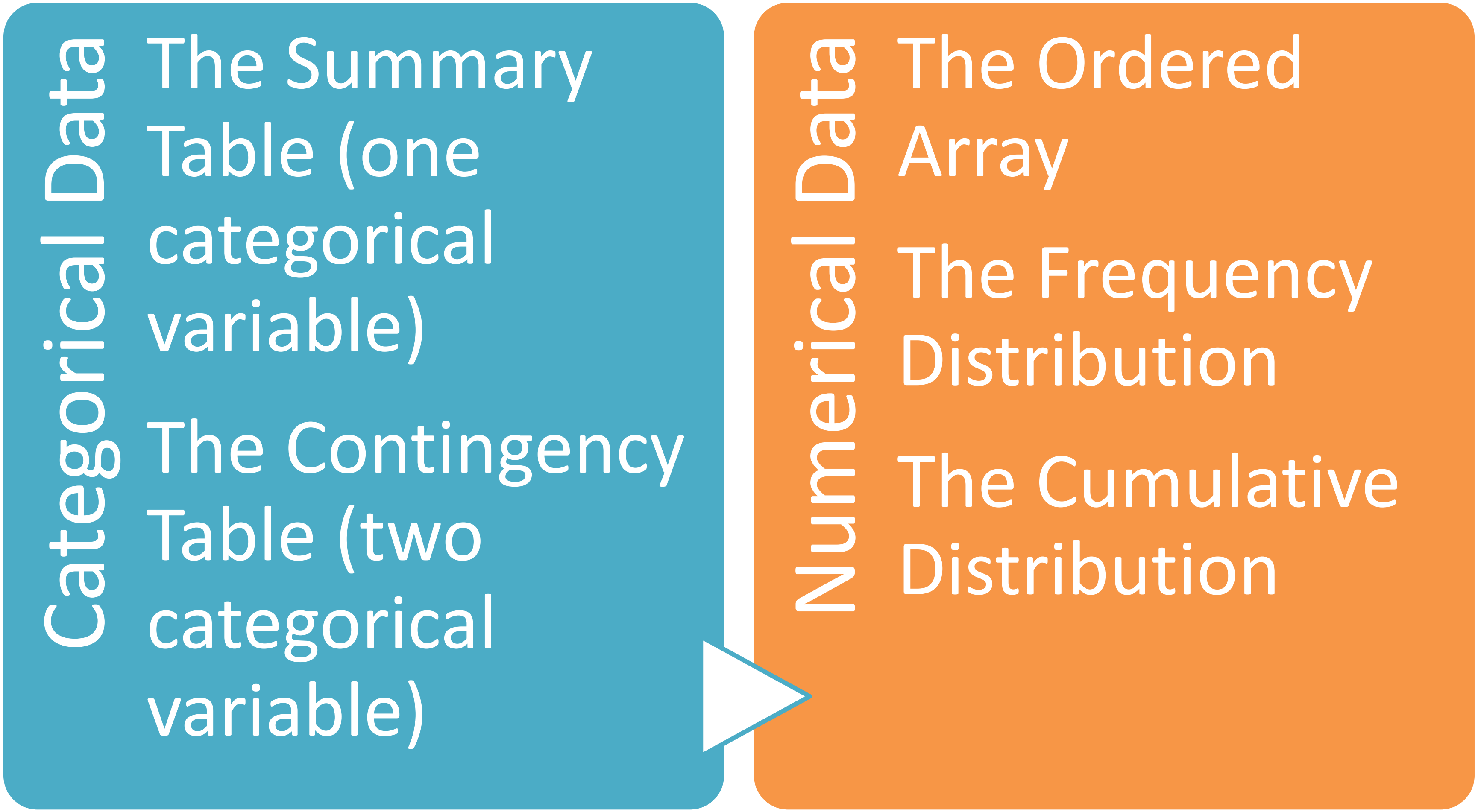
As responses from a survey

As data distributed by an organization or individual

As outcomes of a designed experiment

As a result of conducting an observational study

Organizing Data



Categorical Data

- The Summary Table (one categorical variable)
- The Contingency Table (two categorical variable)

Numerical Data

- The Ordered Array
- The Frequency Distribution
- The Cumulative Distribution

CATEGORICAL DATA

Class Survey

What is your hand phone brand?

What is your phone carrier?

The Summary Table

Students' Home Province of Statistic for Business 1 Year 2014

Province	Frequency	Percentage
West Java	13	46.43%
South Sulawesi	5	17.86%
Jakarta	2	7.14%
East Java	2	7.14%
North Sumatera	1	3.57%
South Sumatera	1	3.57%
Central Sulawesi	1	3.57%
Banten	1	3.57%
Bali	1	3.57%
West Sumatera	1	3.57%
Total	**28**	**100.00%**

The Contingency Table

Student of Statistic for Business 1 Year 2014
Based on Gender and Sibling Status

Gender	Has sibling(s)?		Total
	Yes	No	
Male	6	1	7
Female	18	2	20
Total	24	3	27

The Contingency Table

Overall Percentage

Student of Statistic for Business 1 Year 2014
Based on Gender and Sibling Status

Gender	Has sibling(s)?		Total
	Yes	No	
Male	22%	4%	26%
Female	67%	7%	74%
Total	89%	11%	100%

The Contingency Table

Row Percentage

Student of Statistic for Business 1 Year 2014
Based on Gender and Sibling Status

Gender	Has sibling(s)?		Total
	Yes	No	
Male	86%	14%	100%
Female	90%	10%	100%
Total	89%	11%	100%

The Contingency Table

Column Percentage

Student of Statistic for Business 1 Year 2014
Based on Gender and Sibling Status

Gender	Has sibling(s)?		Total
	Yes	No	
Male	25%	33%	26%
Female	75%	67%	74%
Total	100%	100%	100%

NUMERICAL DATA

Class Survey

How tall are you?

What is your shoe size?

The Ordered Array

150 155 155 155 155 156 156 156 156 157
157 160 160 160 160 162 168 168 168 170
170 171 173 173 174 174 175

The Frequency Distribution

Sort raw data in ascending order:

150 155 155 155 155 156 156 156 156 157 157 160 160 160 160 162
168 168 168 170 170 171 173 173 174 174 175

- Find range: **175 - 150 = 25**
- Select number of classes: **5 (usually between 5 and 15)**
- Compute class interval (width): **5 (25/5 then round up)**
- Determine class boundaries (limits):
    - **Class 1: 150 to less than 155**
    - **Class 2: 155 to less than 160**
    - **Class 3: 160 to less than 165**
    - **Class 4: 165 to less than 170**
    - **Class 5: 170 to less than 175**
    - **Class 6: 175 to less than 180**
- Compute class midpoints: **152.5, 157.5, 162.5, 167.5, 172.5, 177.5**
- Count observations & assign to classes

The Frequency Distribution

The Height of Statistic for Business 1's student Year 2014

Height	Frequency
150 but less than 155	1
155 but less than 160	10
160 but less than 165	5
165 but less than 170	3
170 but less than 175	7
175 but less than 180	1
Total	**27**

The Relative Frequency Distribution and the Percentage Distribution

The Height of Statistic for Business 1's student Year 2014

Height	Relative Frequency	Percentage
150 but less than 155	0.04	4%
155 but less than 160	0.37	37%
160 but less than 165	0.19	19%
165 but less than 170	0.11	11%
170 but less than 175	0.26	26%
175 but less than 180	0.04	4%
Total	**1**	**100.00%**

Developing the Cumulative Percentage Distribution

The Height of Statistic for Business 1’s student Year 2014

Height	Percentage (%)	Percentage of Meals Less Than Lower Boundary of Class Interval (%)
150 but less than 155	4	0
155 but less than 160	37	4
160 but less than 165	19	41=4+37
165 but less than 170	11	50=4+37+19
170 but less than 175	26	70=4+37+19+11
175 but less than 180	4	96=4+37+19+11+26

The Cumulative Distribution

The Height of Statistic for Business 1's student Year 2014

Height	Cumulative Percentage less than indicated value
150	0
155	4%
160	41%
165	59%
170	70%
175	96%
180	**100%**

VISUALIZING DATA

Visualizing Data

Categorical Variable

- Visualizing one variable
  - Bar chart, Pie chart an Pareto chart
- Visualizing two variables
  - Side-by-side bar chart

Numerical Variable

- Visualizing one variable
  - Stem-and-leaf display
  - Histogram, polygon and ogive
- Visualizing two variables
  - Scatter plot and time-series plot

Visualizing Data

Categorical Variable

- Visualizing one variable
  - Bar chart, Pie chart an Pareto cha
- Visualizing two varia
  - Side-by-side bar

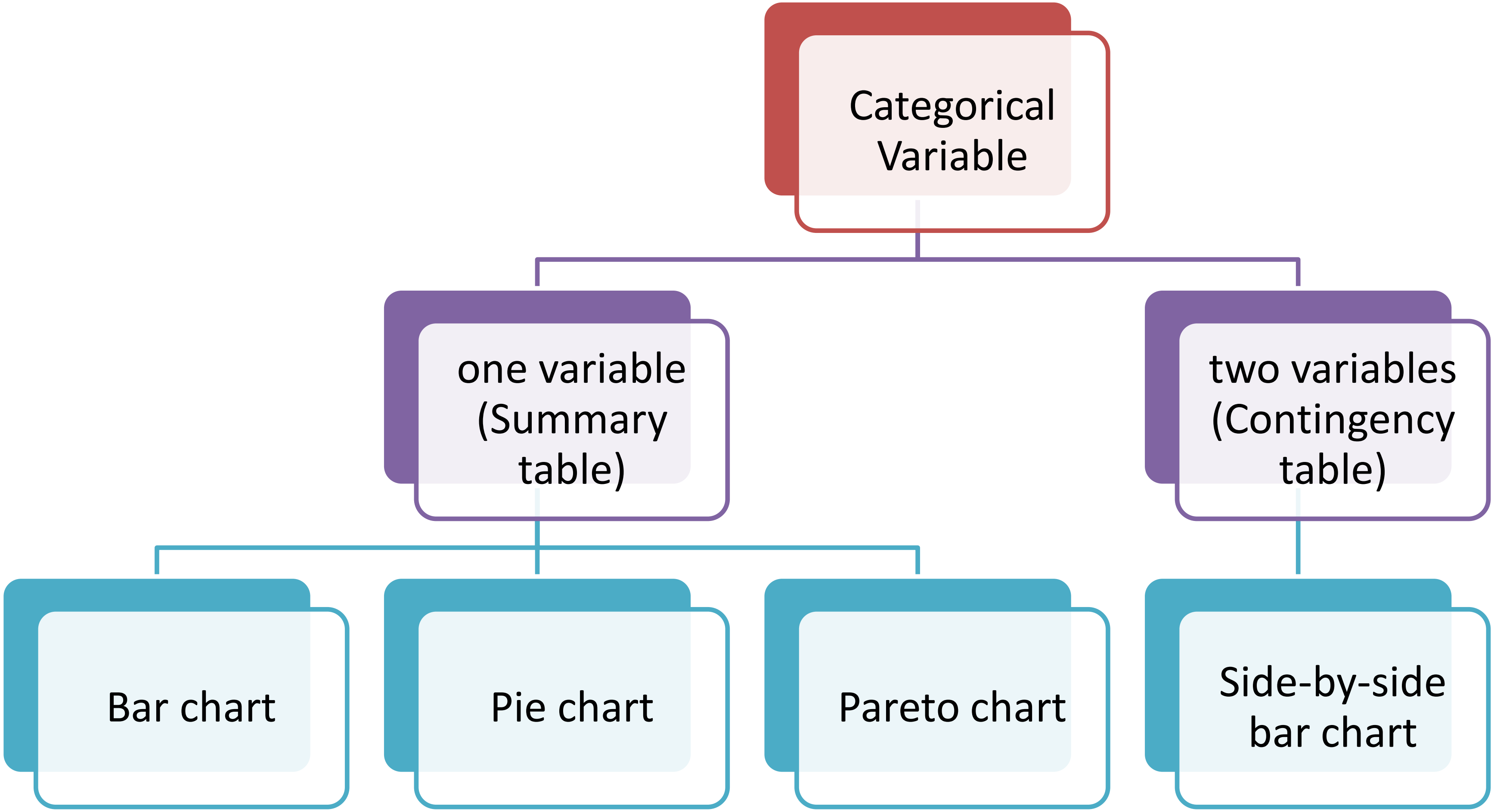
Numerical Var

- Visualizing one
  - Stem-and-le
  - Histogram, polygo
- Visualizing two variables
  - Scatter plot and time-series plot

Graphical Errors

CATEGORICAL VARIABLE

Visualizing Data



Categorical Variable

one variable (Summary table)

two variables (Contingency table)

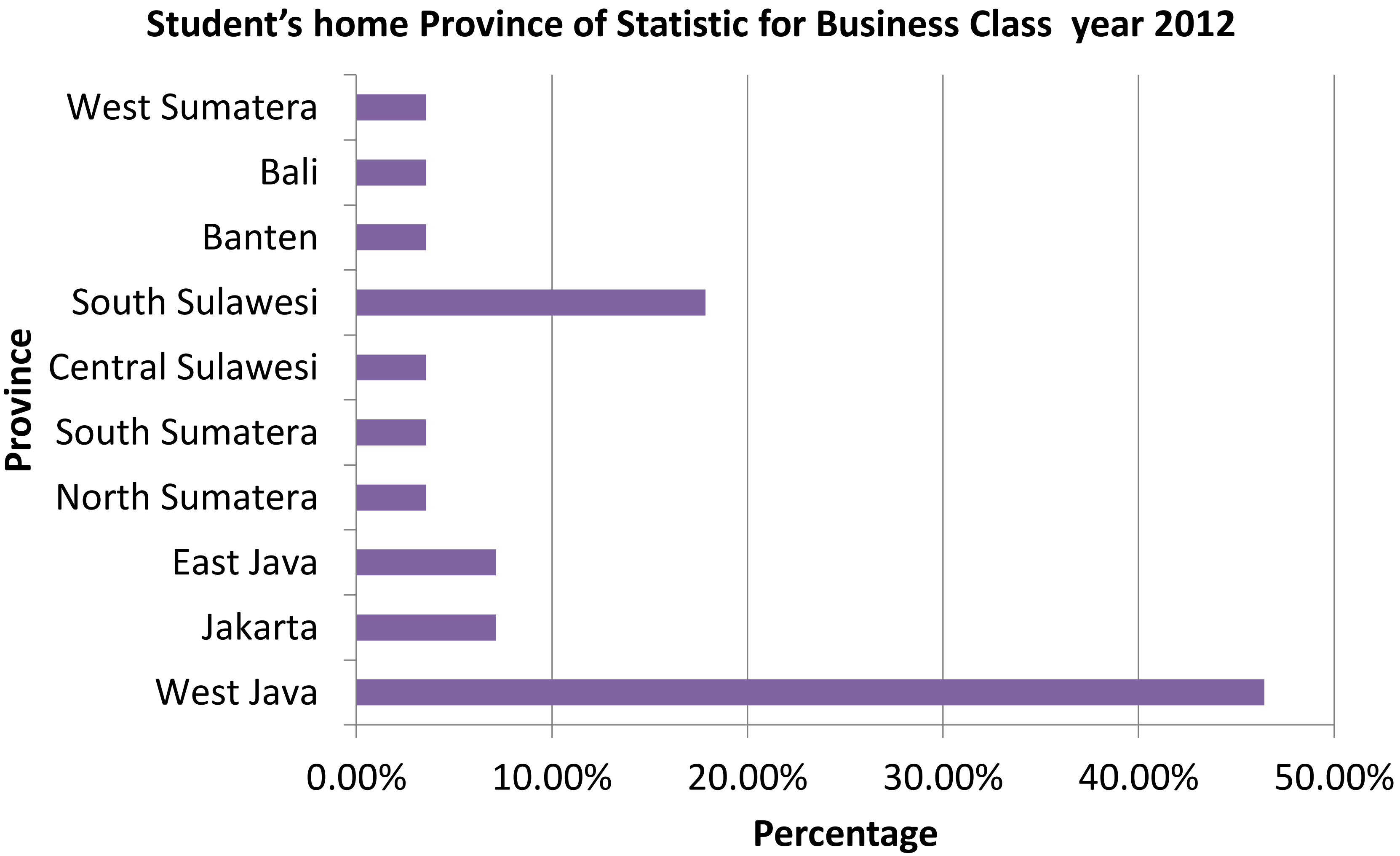
Bar chart

Pie chart

Pareto chart

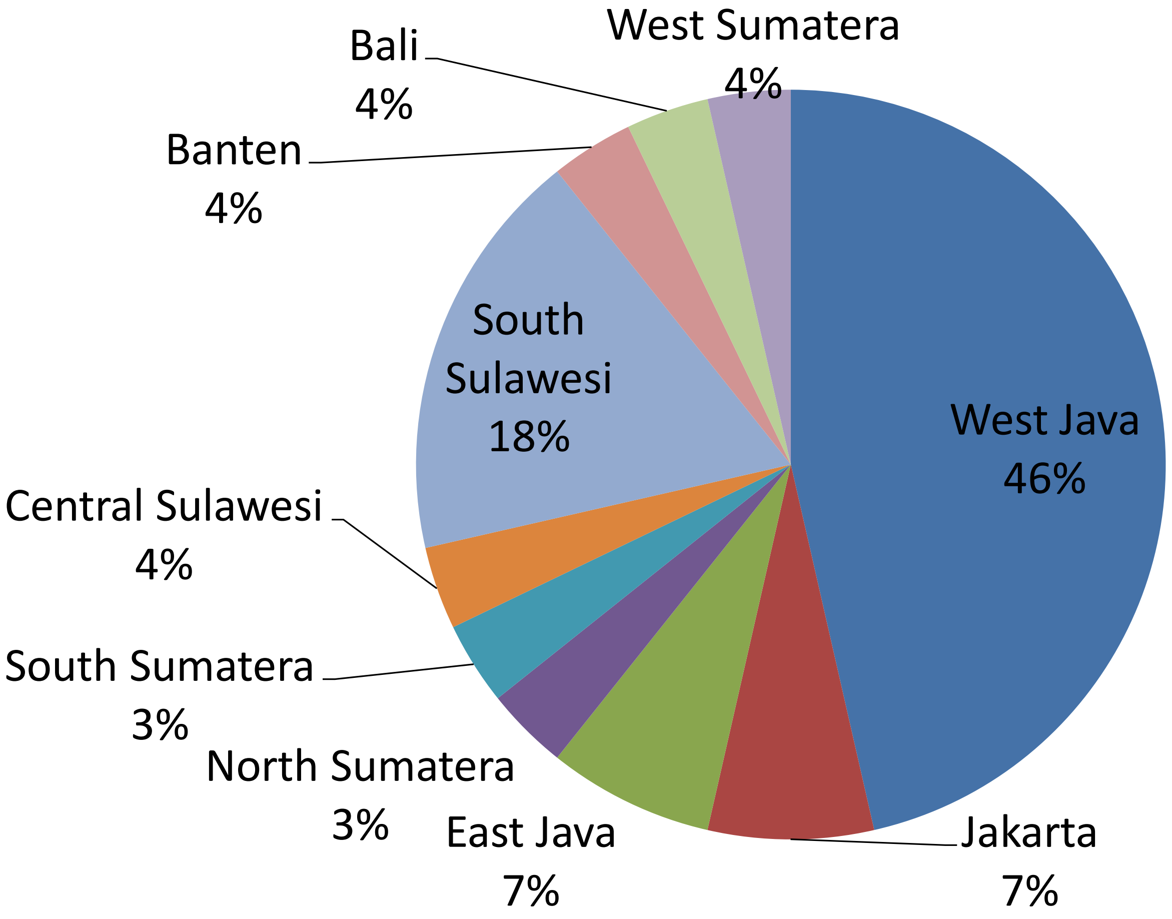
Side-by-side bar chart

Bar Chart


Student’s home Province of Statistic for Business Class year 2012
West Sumatera
Bali
Banten
South Sulawesi
Central Sulawesi
South Sumatera
North Sumatera
East Java
Jakarta
West Java
Province
0.00%
10.00%
20.00%
30.00%
40.00%
50.00%
Percentage

Pie Chart

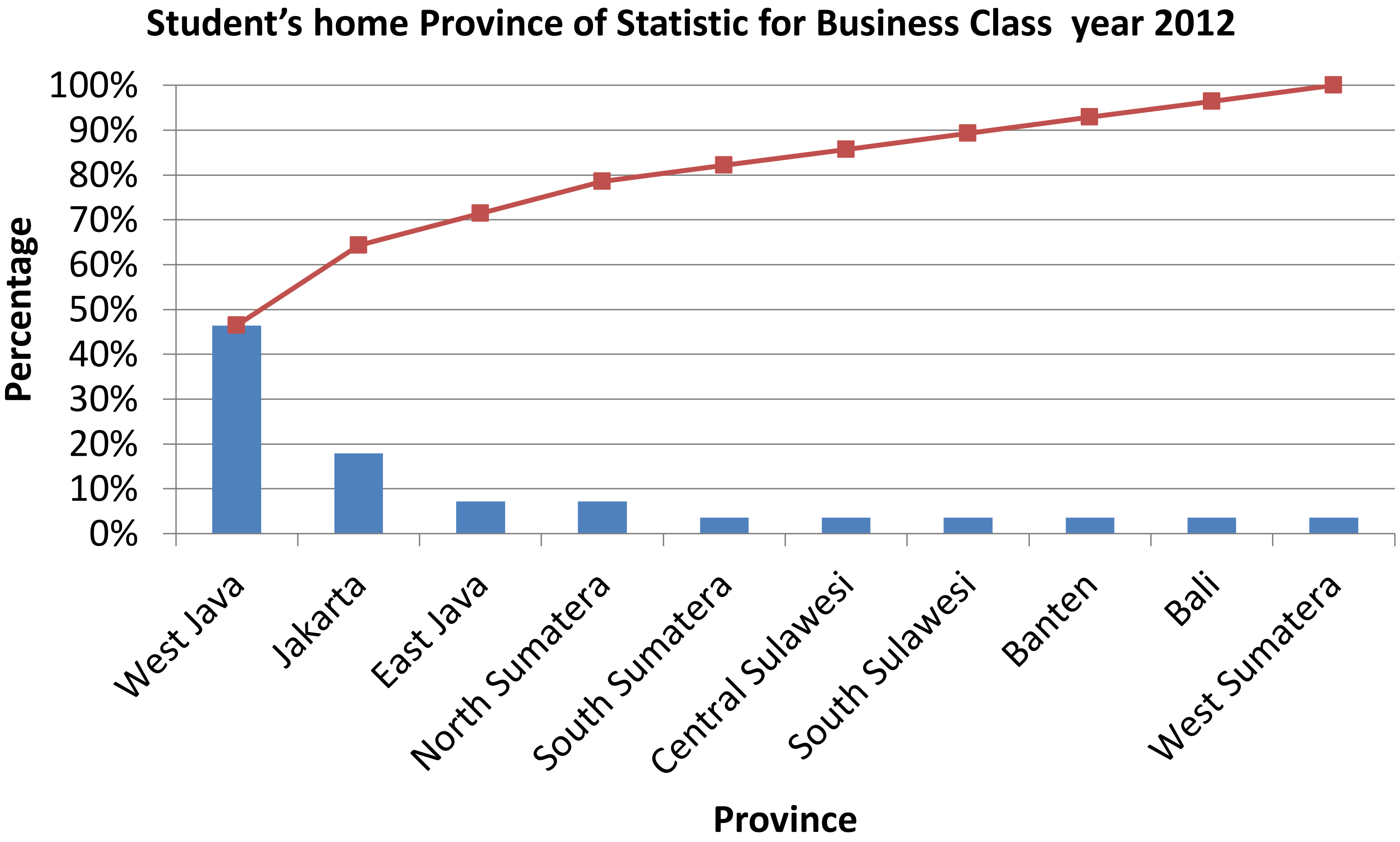
Student's home Province of Statistic for Business Class year 2014



Pareto Chart

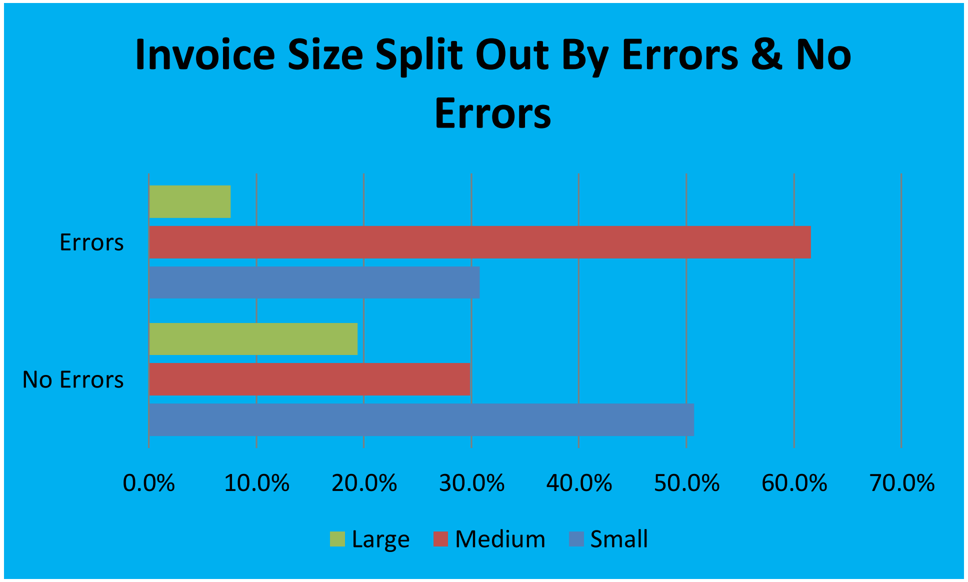
- A Pareto chart has the capability to separate the “vital few” from the “trivial many,” enabling you to focus on the important categories.
- In situations in which the data involved consist of defective or nonconforming items, a Pareto chart is a powerful tool for prioritizing improvement efforts.

Pareto Chart

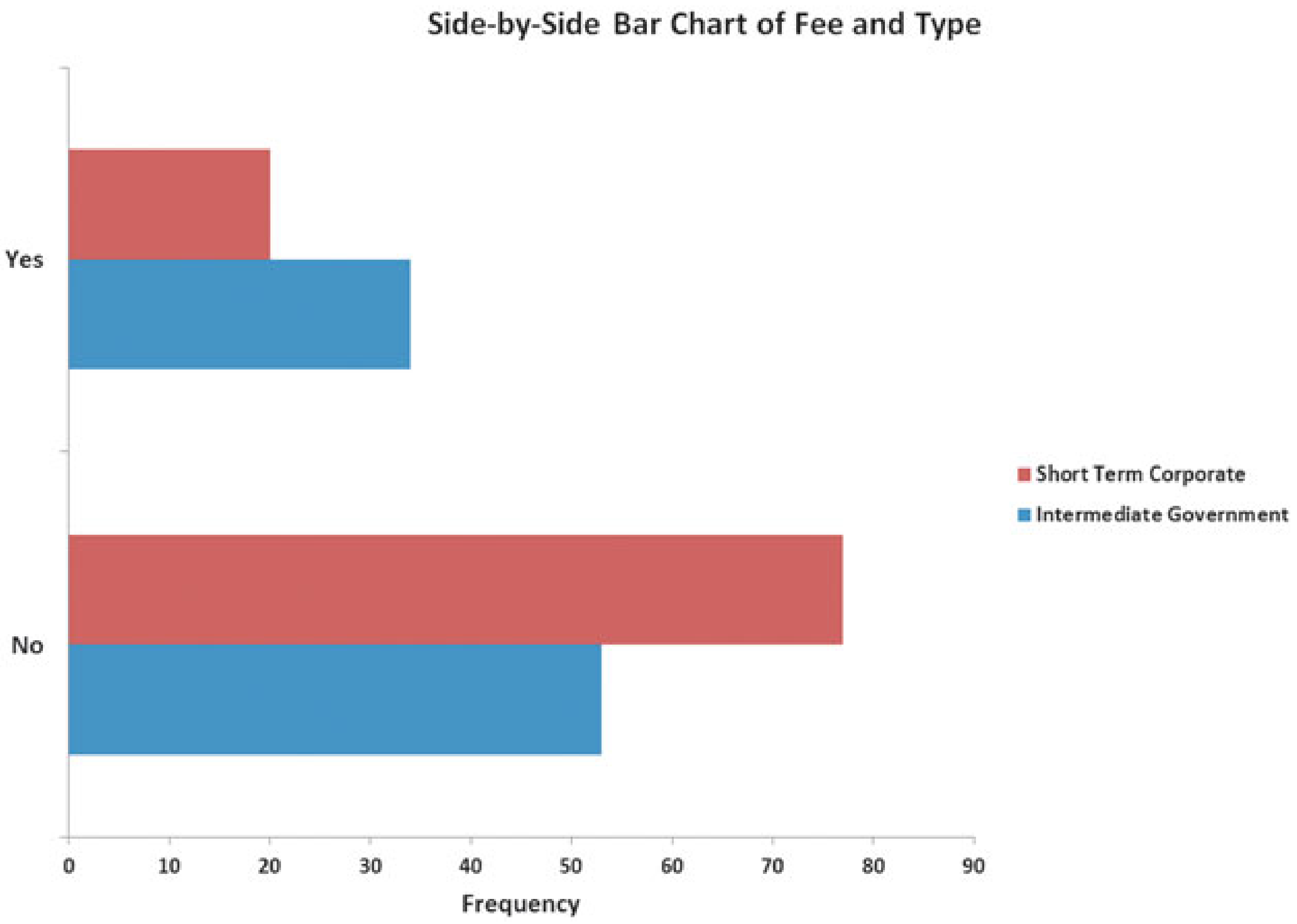


Side-By-Side Bar Chart

	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%

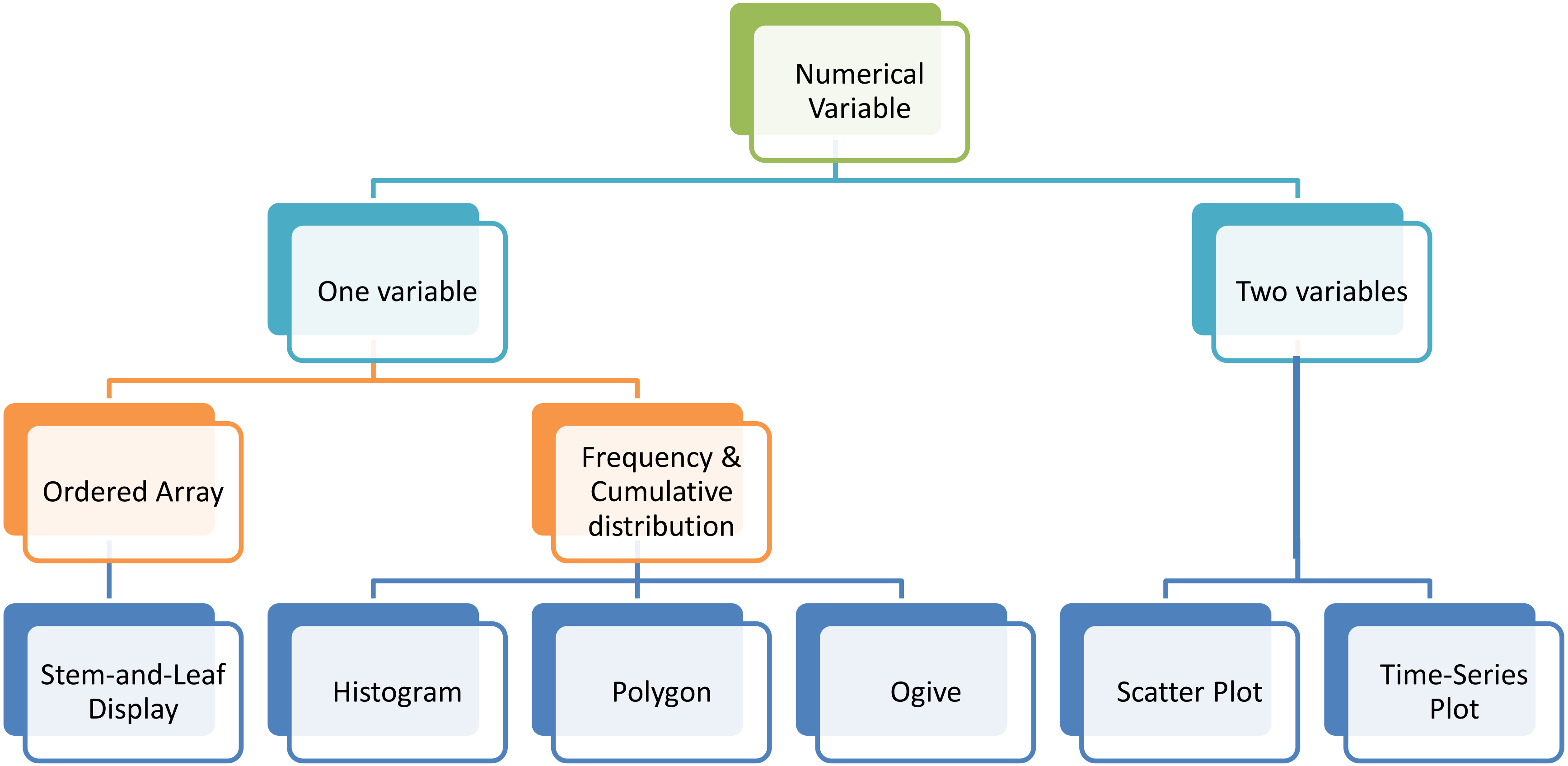


Side-By-Side Bar Chart



NUMERICAL VARIABLE

Visualizing Data



Numerical Variable

One variable

Two variables

Ordered Array

Frequency & Cumulative distribution

Stem-and-Leaf Display

Histogram

Polygon

Ogive

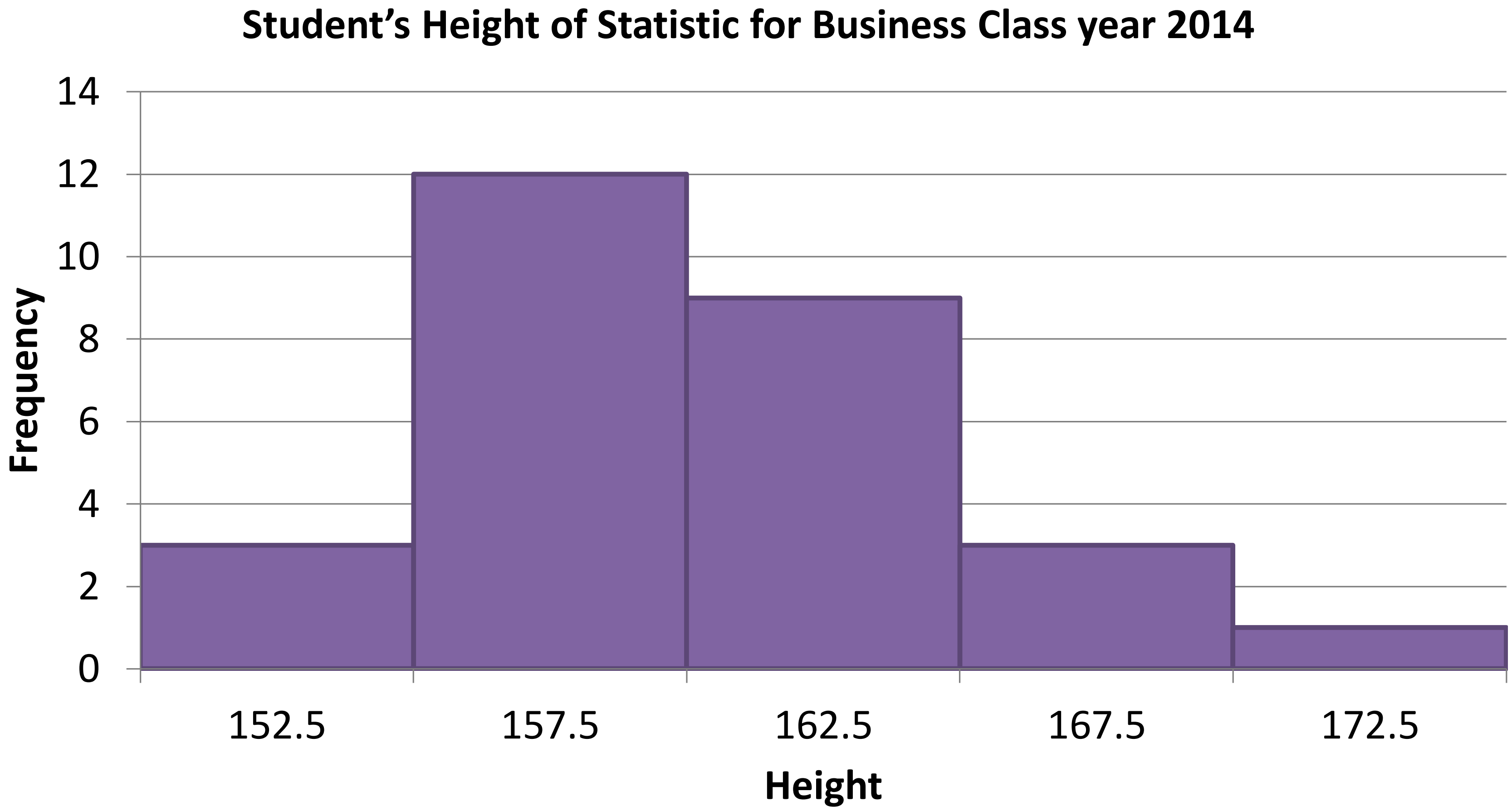
Scatter Plot

Time-Series Plot

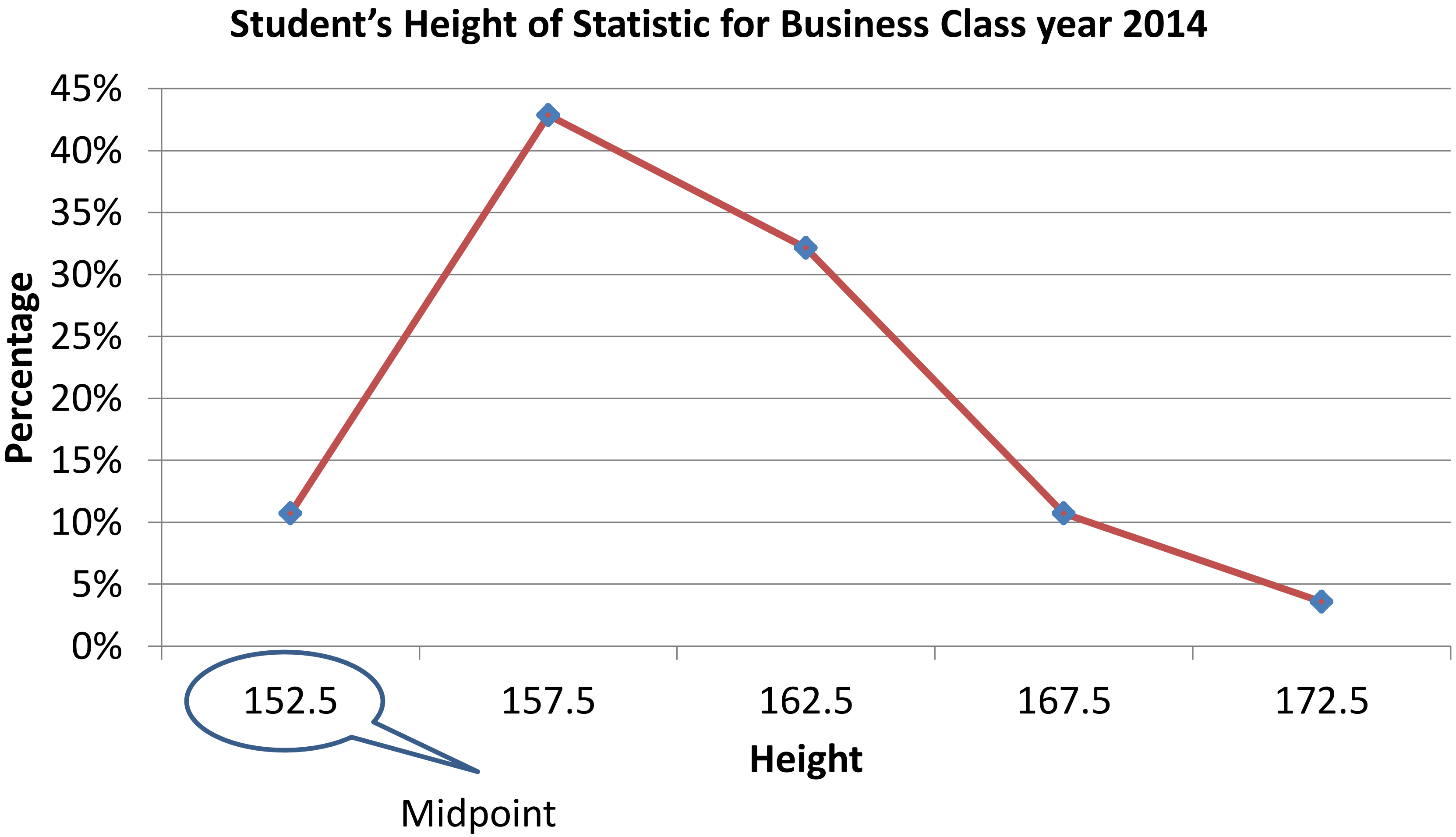
Stem-and-Leaf Display

Stem	Leaf
15	0 2 4 5 5 5 5 5 5 7 8 8 8 9 9
16	0 0 0 0 0 0 1 2 3 5 5 5
17	0

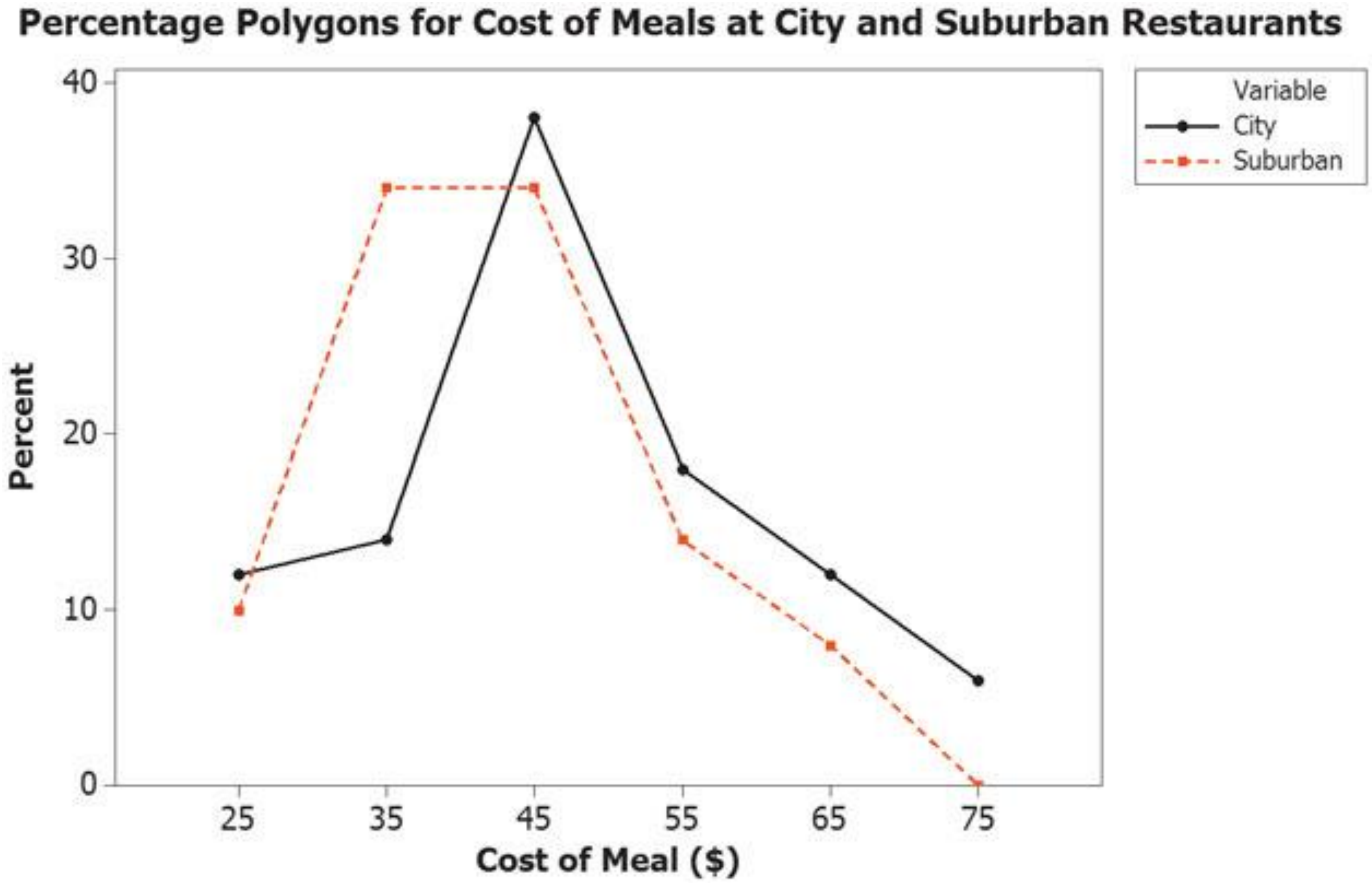
Histogram


Student’s Height of Statistic for Business Class year 2014
Frequency
14
12
10
8
6
4
2
0
152.5
157.5
162.5
167.5
172.5
Height

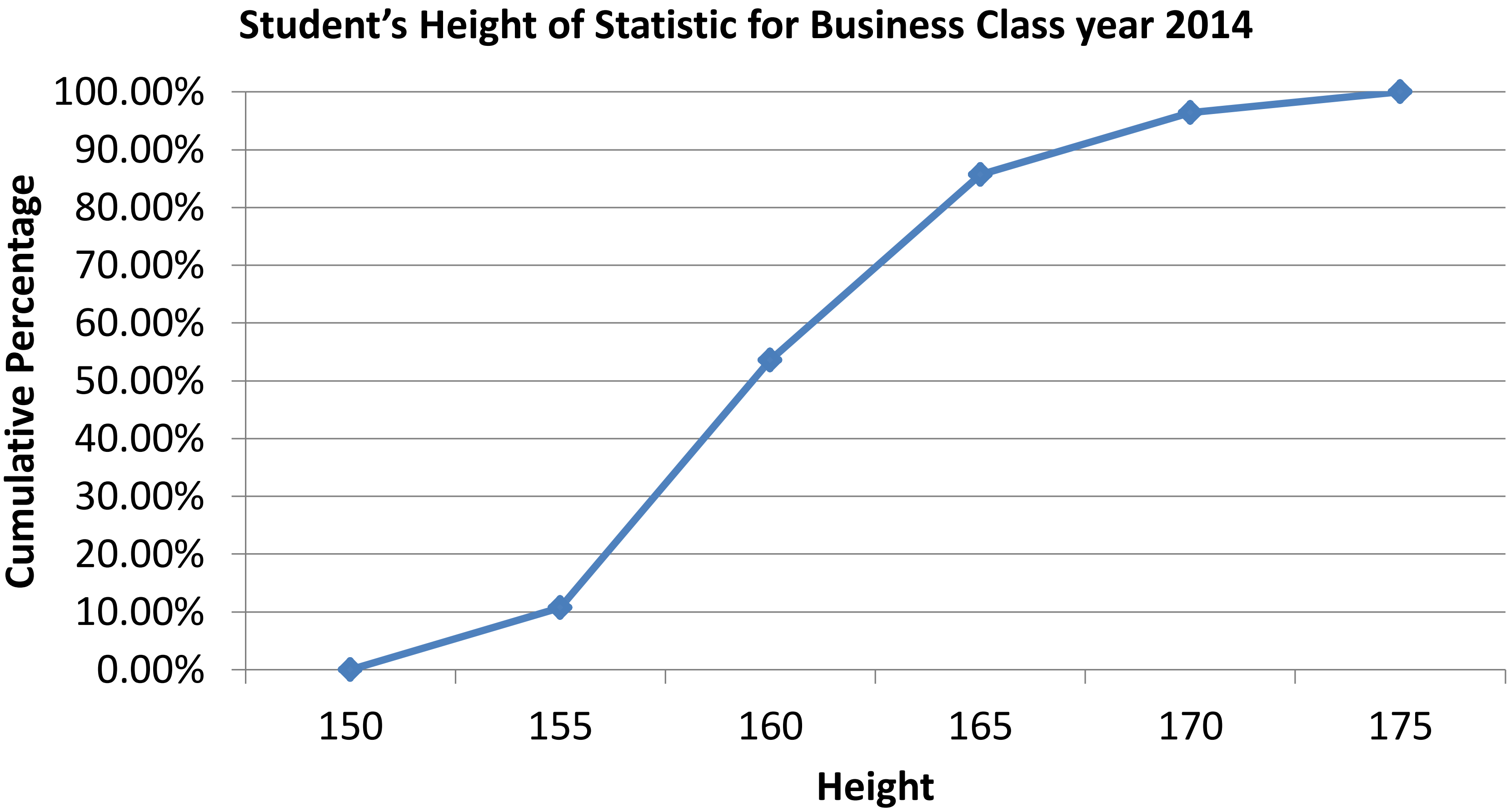
Percentage Polygon



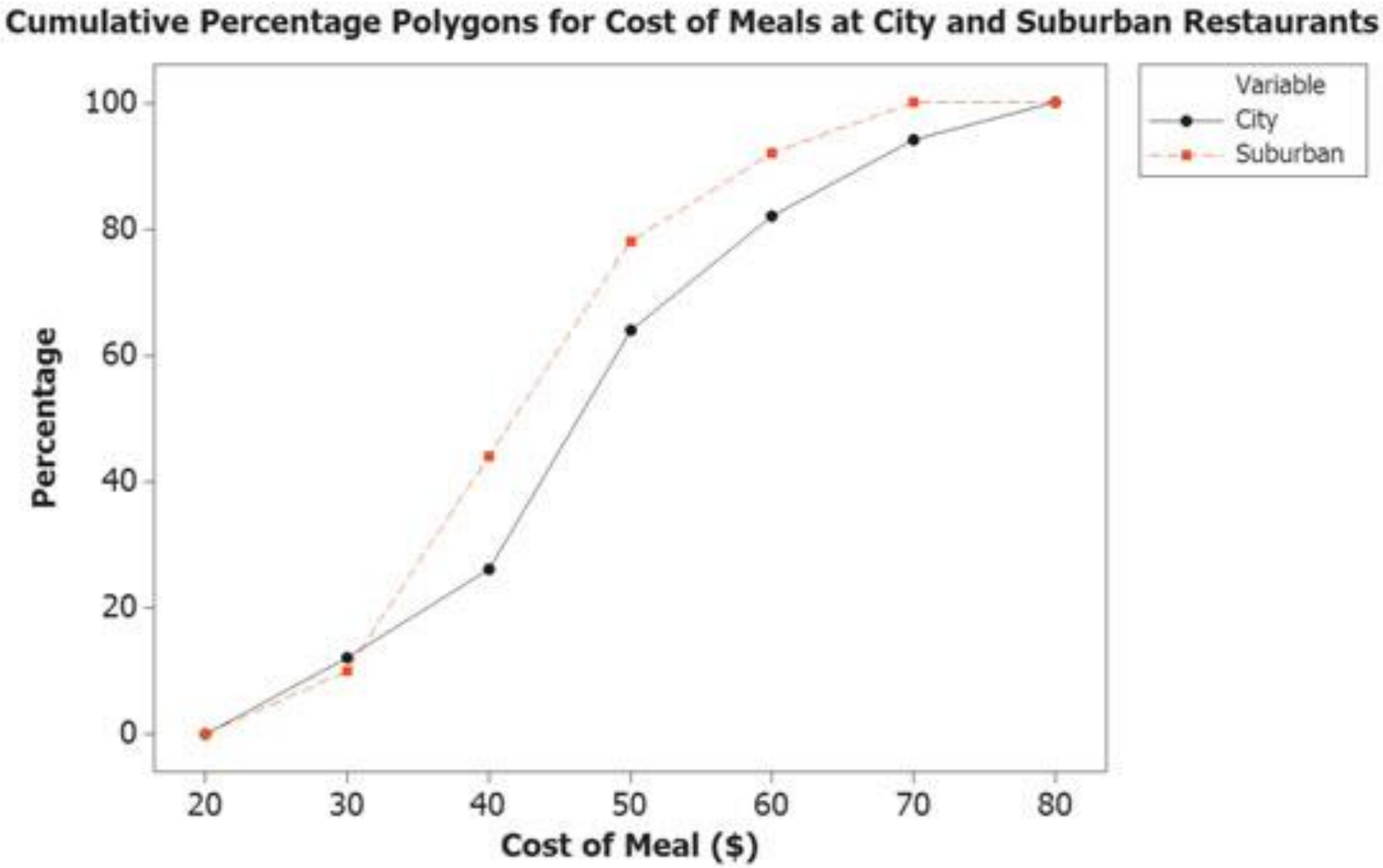
Percentage Polygon



Cumulative Percentage Polygon (Ogive)



Cumulative Percentage Polygon (Ogive)

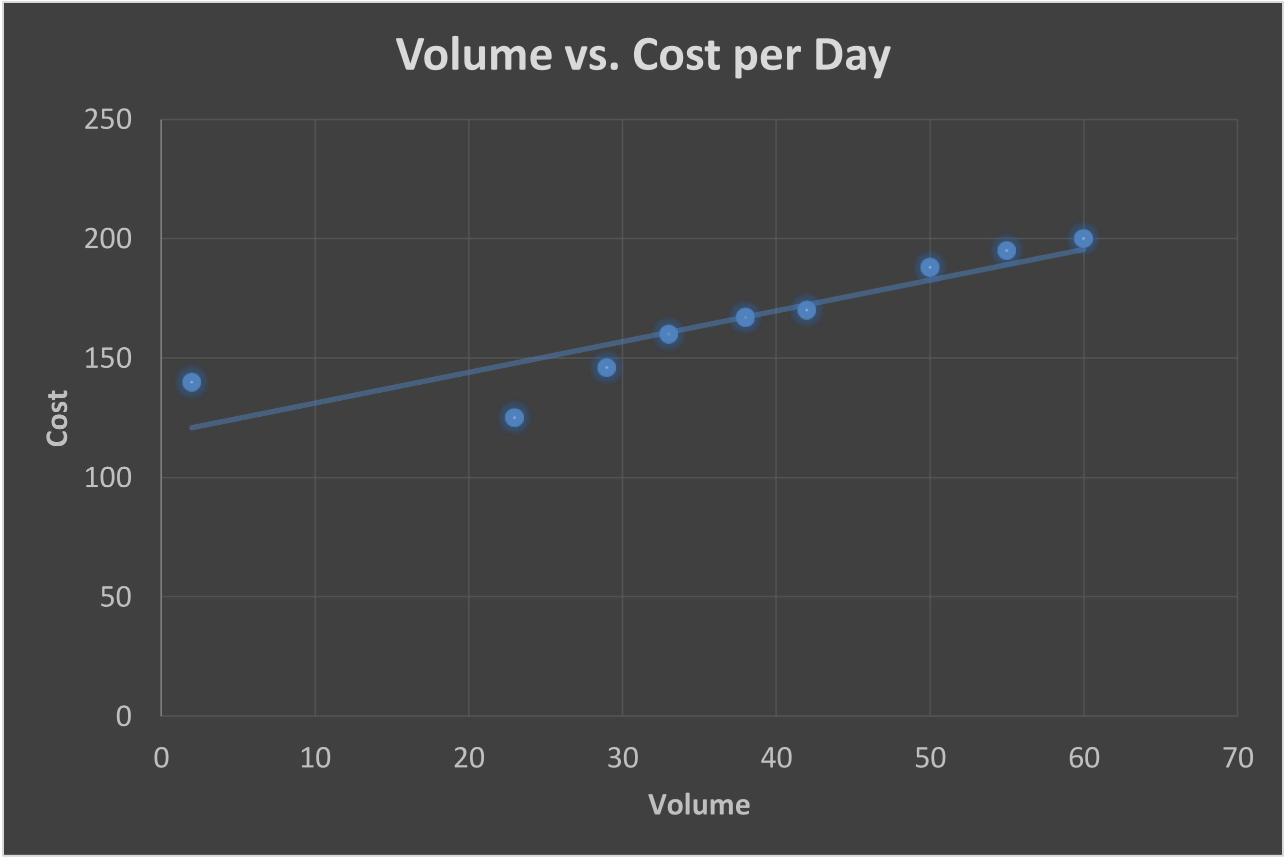


Note!

When you construct polygons or histograms, the vertical (*Y*) axis should show the true zero, or “origin,” so as not to distort the character of the data.

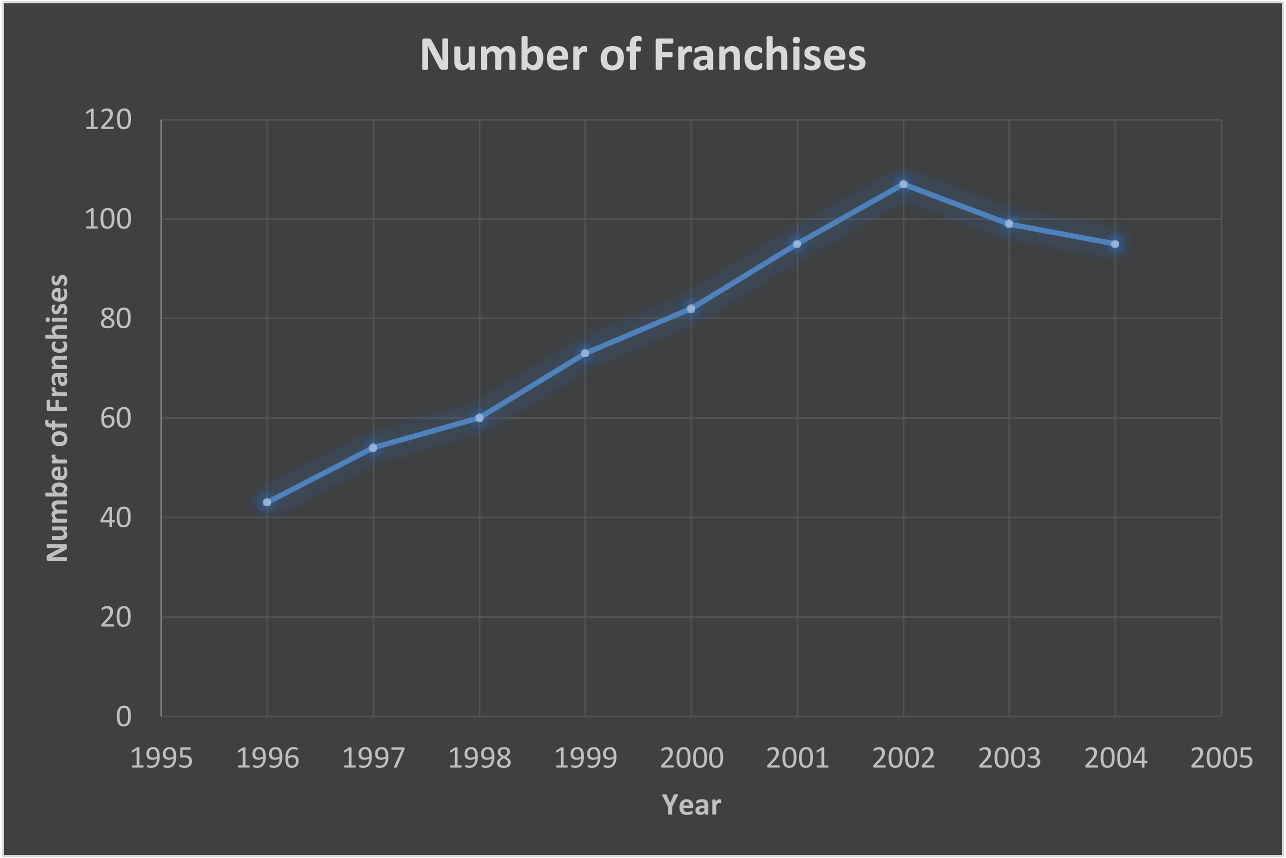
Scatter Plot

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



Time Series Plot

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



Principles of Excellent Graphs

The graph should not distort the data.

The graph should not contain unnecessary adornments (sometimes referred to as chart junk).

The scale on the vertical axis should begin at zero.

All axes should be properly labeled.

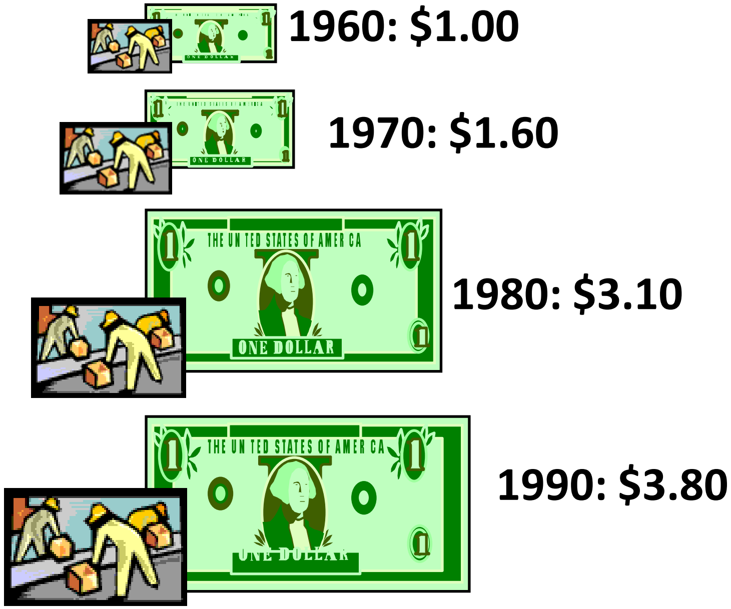
The graph should contain a title.

The simplest possible graph should be used for a given set of data.

Graphical Errors: Chart Junk

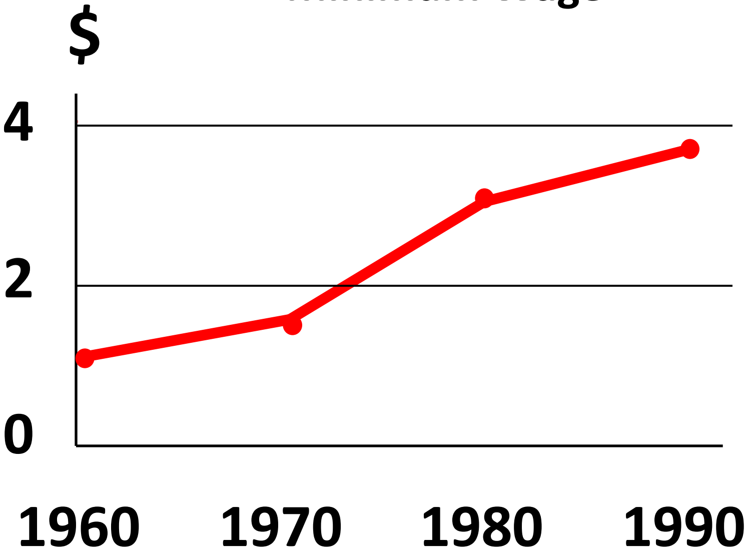
🚫 **Bad Presentation**

Minimum Wage



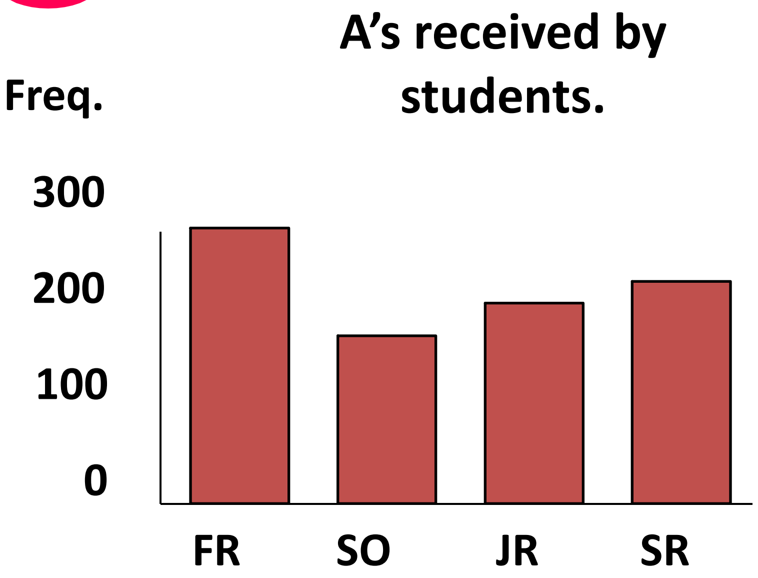
✓ **Good Presentation**

Minimum Wage



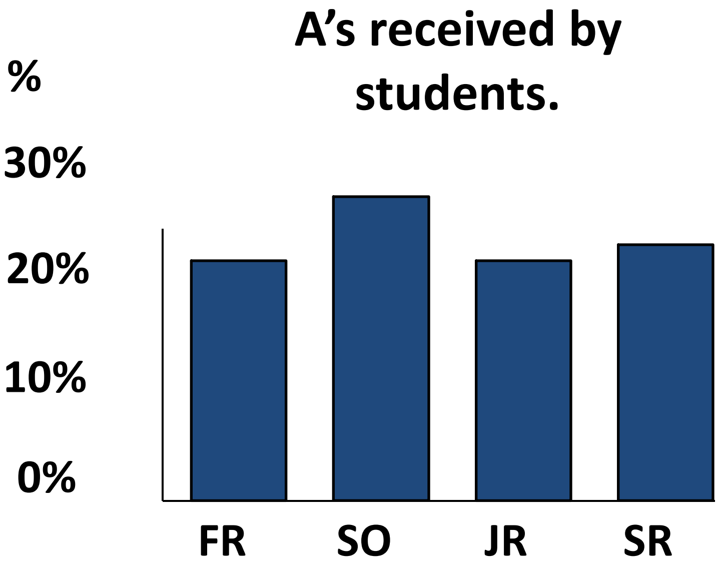
Graphical Errors: No Relative Basis

🚫 Bad Presentation



A's received by students.

✓ Good Presentation

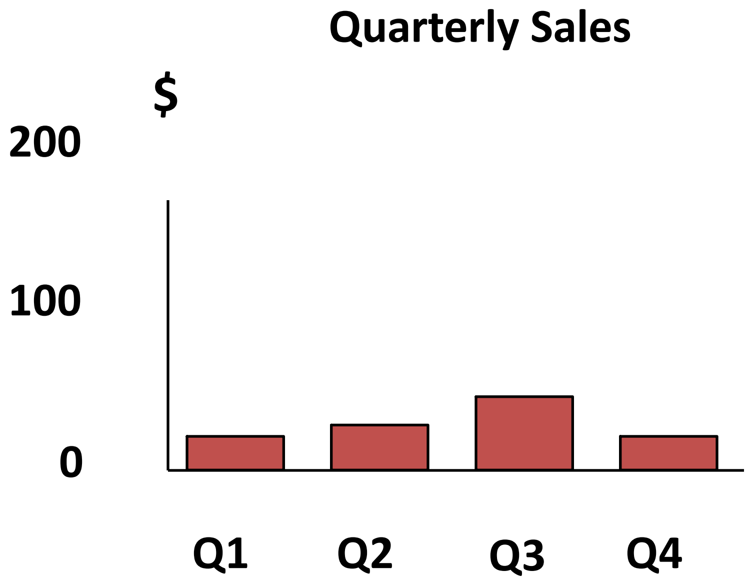


A's received by students.

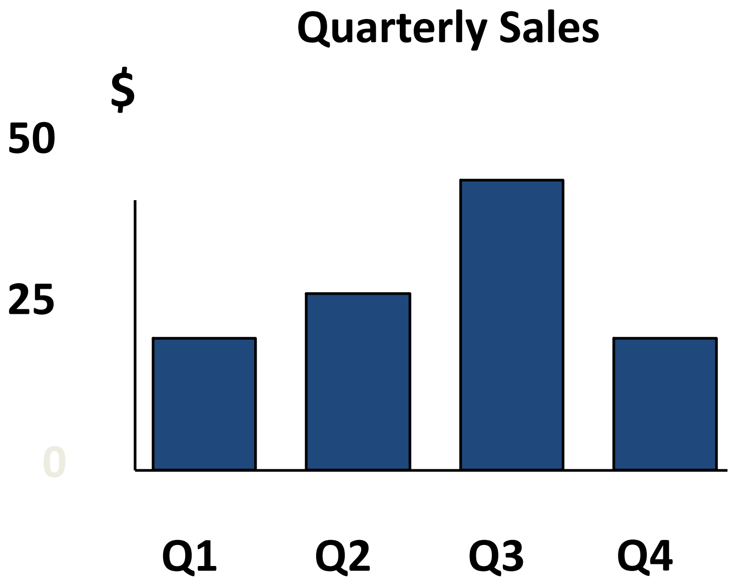
FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Graphical Errors: Compressing the Vertical Axis

🚫 **Bad Presentation**



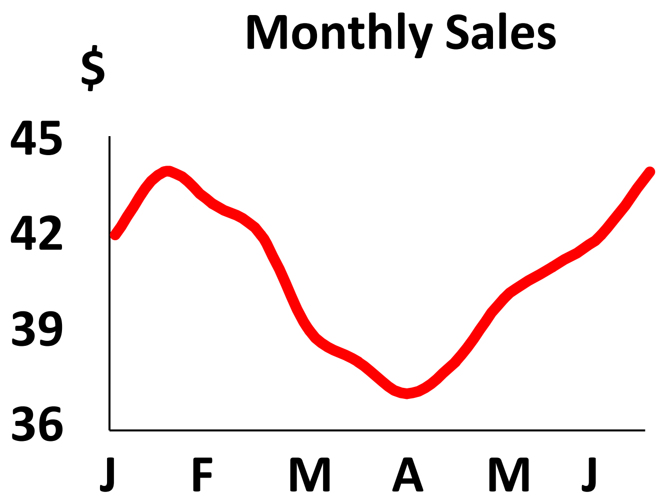
✓ **Good Presentation**

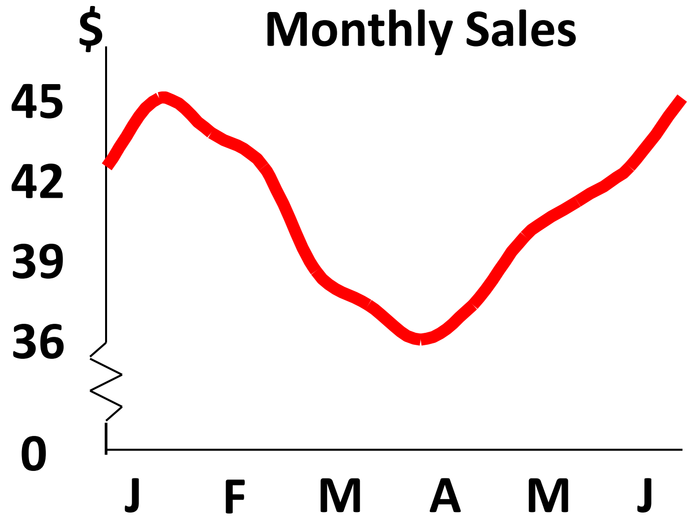


Graphical Errors: No Zero Point on the Vertical Axis

🚫 **Bad Presentation**

✓ **Good Presentations**





Graphing the first six months of sales

EXERCISE

2.28

The following table indicates the percentage of residential electricity consumption in the United States, organized by type of appliance in a recent year:

2.28

Type of Appliance	Percentage (%)
Air conditioning	18
Clothes dryers	5
Clothes washers/other	24
Computers	1
Cooking	2
Dishwashers	2
Freezers	2
Lighting	16
Refrigeration	9
Space heating	7
Water heating	8
TVs and set top boxes	6

2.28

a. Construct a bar chart, a pie chart, and a Pareto chart.

b. Which graphical method do you think is the best for portraying these data?

2.37

This data contains the cost per ounce ($) for a sample of 14 dark chocolate bars:

0.68	0.72	0.92	1.14	1.42	0.94	0.77
0.57	1.51	0.57	0.55	0.86	1.41	0.90

a. Construct an ordered array.

b. Construct a stem-and-leaf display.

c. Does the ordered array or the stem-and-leaf display provide more information? Discuss.

d. Around what value, if any, is the cost of dark chocolate bars concentrated? Explain.

2.38

The following data is about the cost of electricity during July 2010 for a random sample of 50 one-bedroom apartments in a large city:

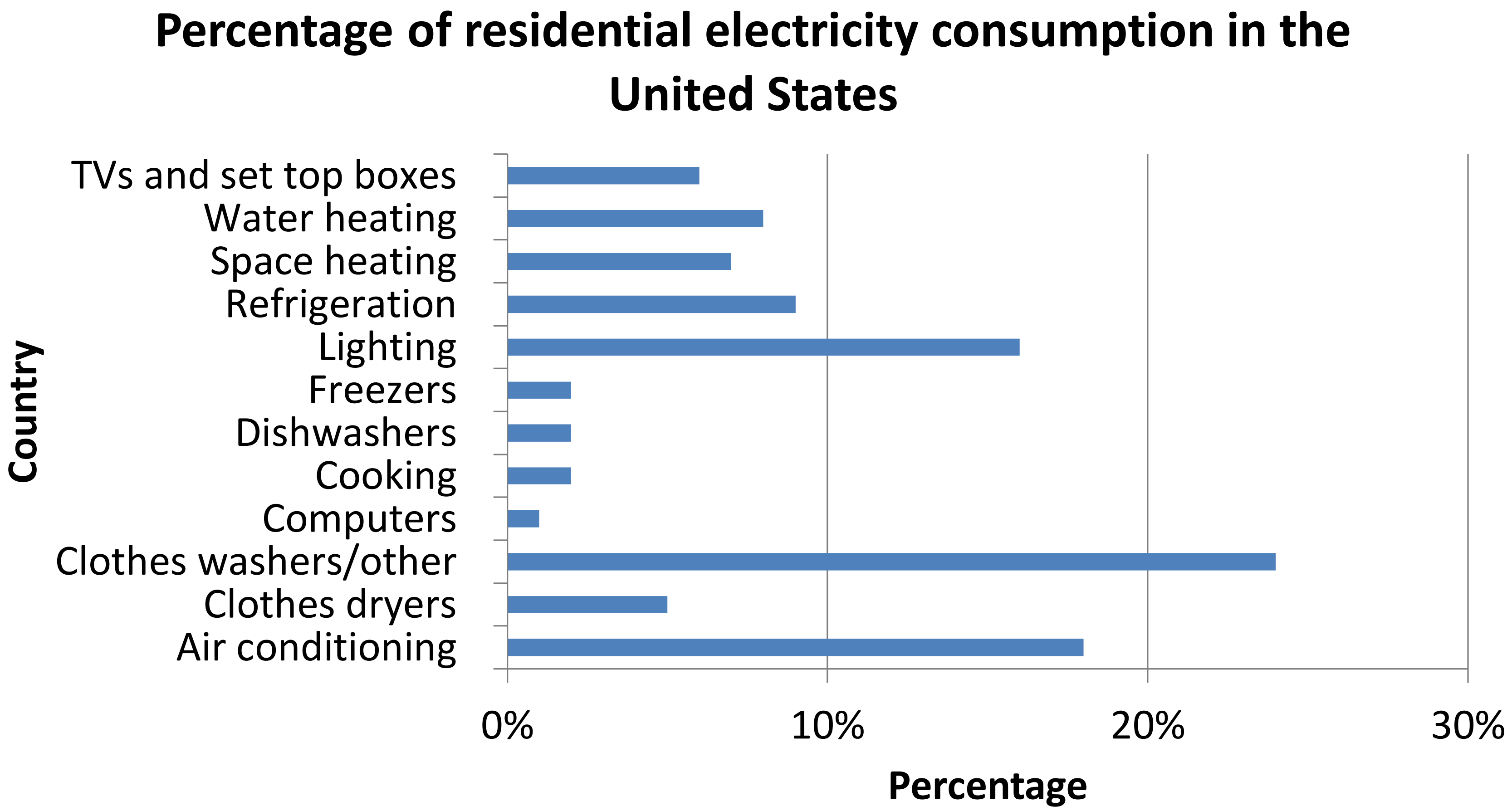
96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

2.38

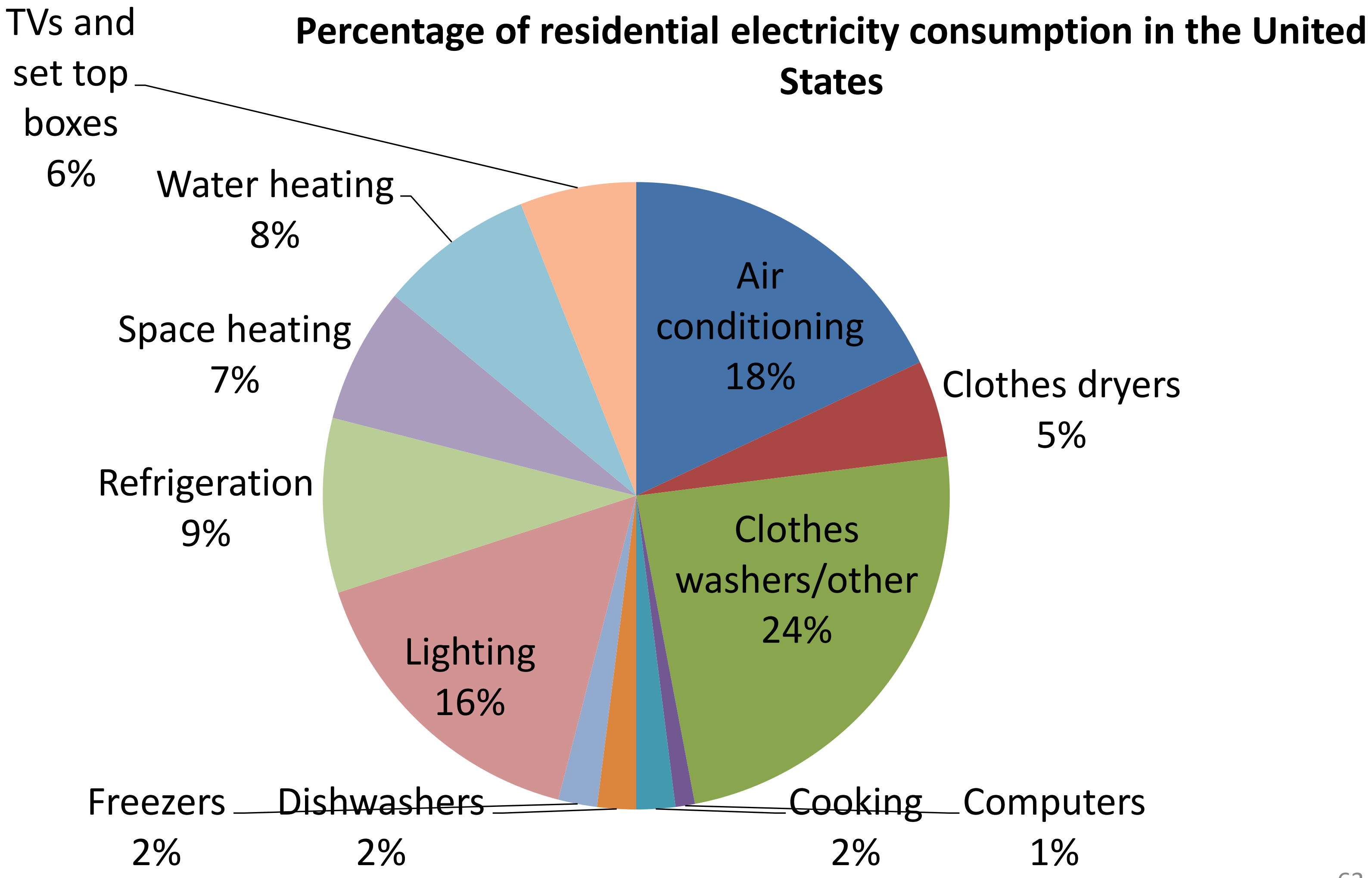
a. Construct a histogram and a percentage polygon.
b. Construct a cumulative percentage polygon.
c. Around what amount does the monthly electricity cost seem to be concentrated?

ANSWER

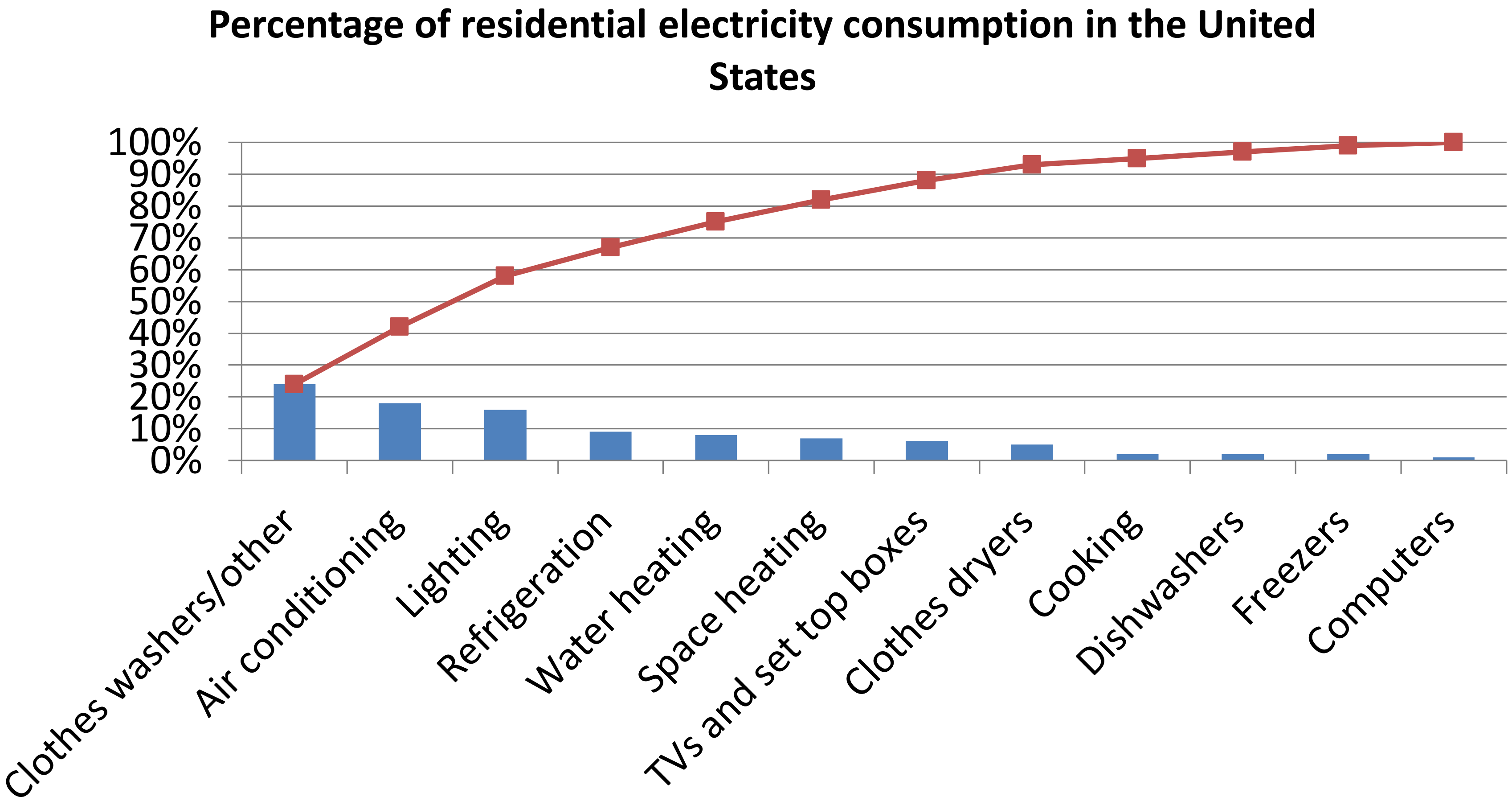
2.28


Percentage of residential electricity consumption in the United States
Country
TVs and set top boxes
Water heating
Space heating
Refrigeration
Lighting
Freezers
Dishwashers
Cooking
Computers
Clothes washers/other
Clothes dryers
Air conditioning
0%
10%
20%
30%
Percentage

2.28


Percentage of residential electricity consumption in the United States
Air conditioning 18%
Clothes dryers 5%
Clothes washers/other 24%
Computers 1%
Cooking 2%
Dishwashers 2%
Freezers 2%
Lighting 16%
Refrigeration 9%
Space heating 7%
Water heating 8%
TVs and set top boxes 6%

2.28


Percentage of residential electricity consumption in the United States
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Clothes washers/other
Air conditioning
Lighting
Refrigeration
Water heating
Space heating
TVs and set top boxes
Clothes dryers
Cooking
Dishwashers
Freezers
Computers

2.37

Ordered array:

0.55 0.57 0.57 0.68 0.72 0.77 0.86
0.90 0.92 0.94 1.14 1.41 1.42 1.51

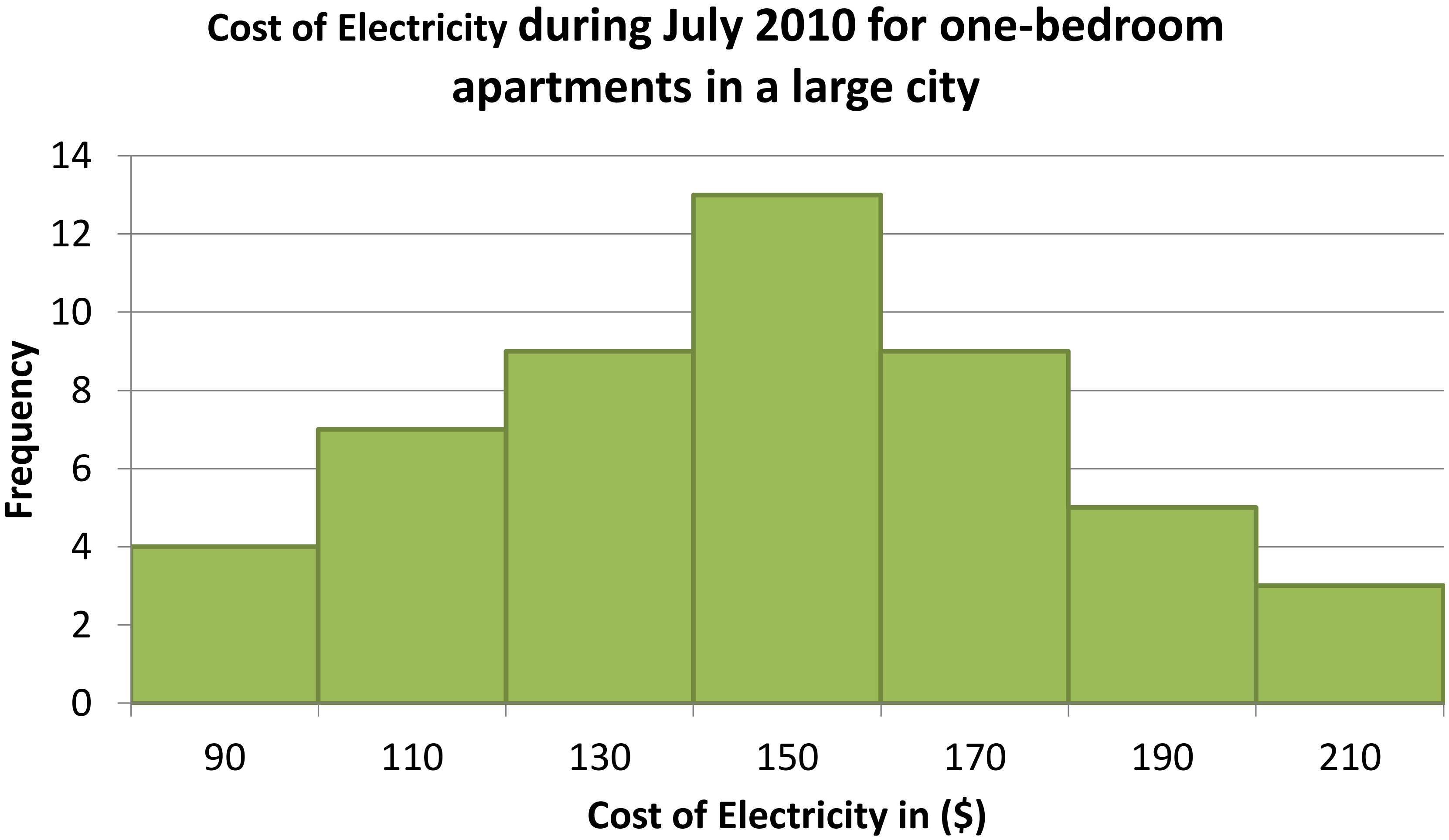
2.37

Stem-and-Leaf Display:

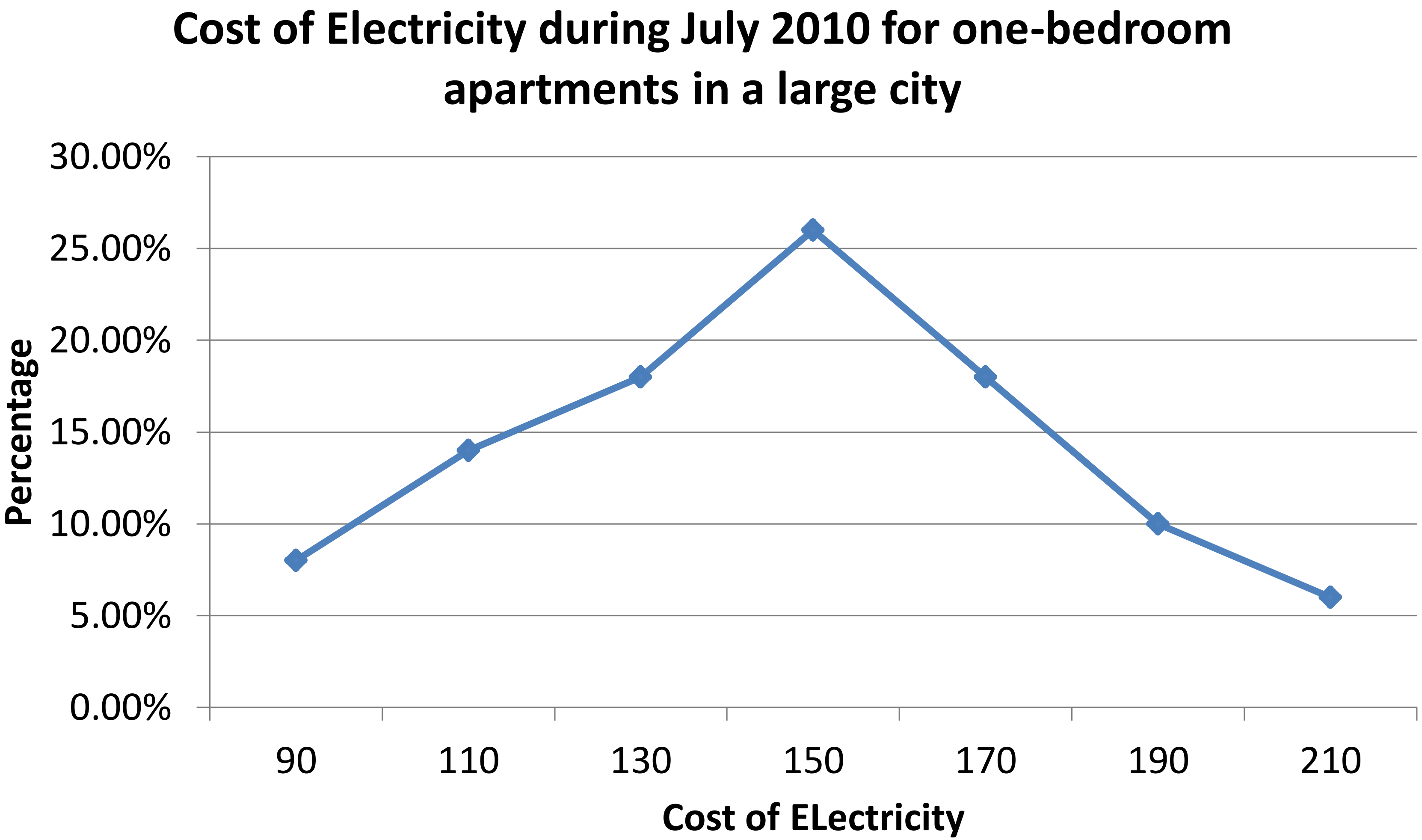
5	5 7 7
6	8
7	2 7
8	6
9	0 2 4
1	
11	4
12	
13	
14	1 2
15	1

Key: 5|7 means: 0.57

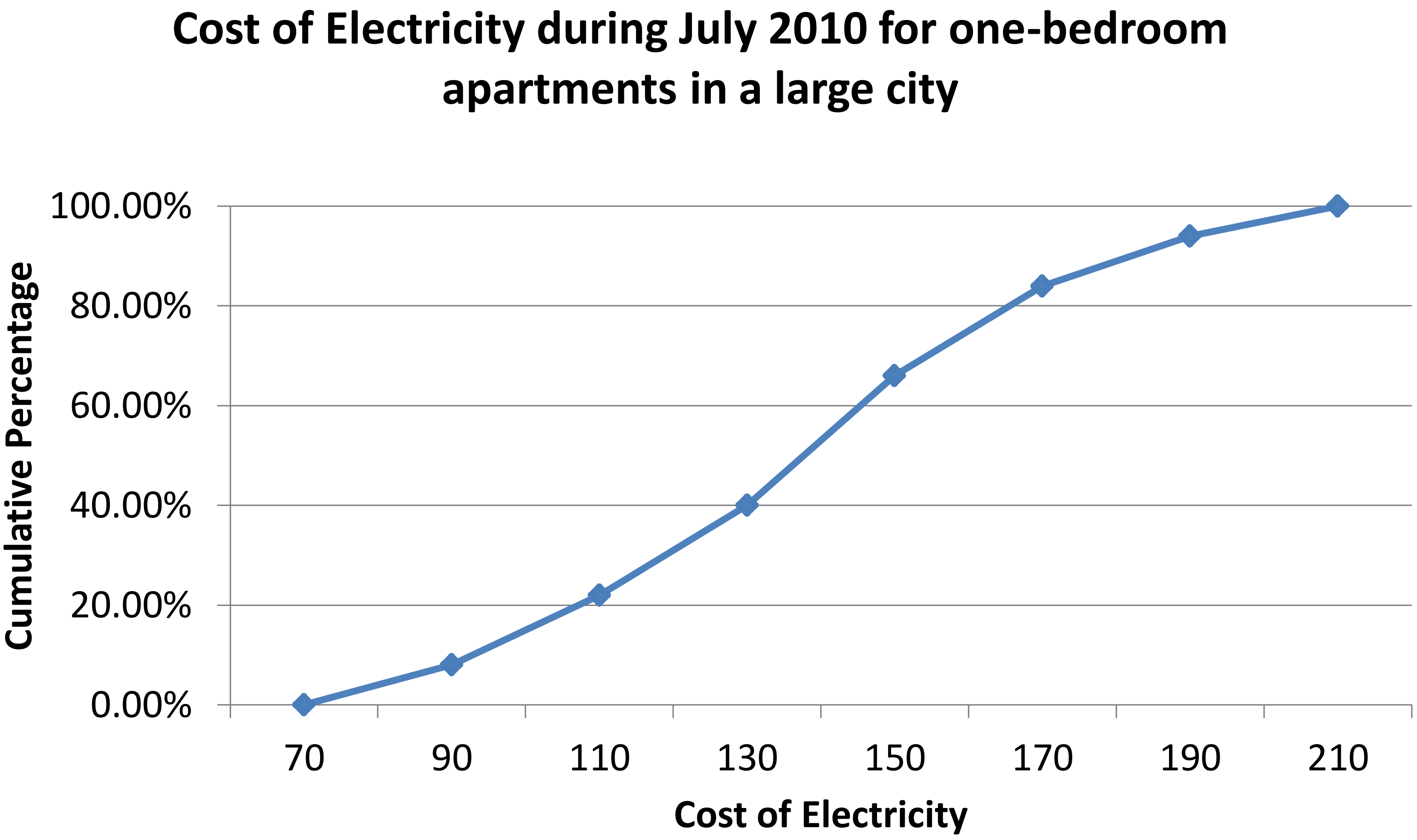
2.38


Cost of Electricity during July 2010 for one-bedroom apartments in a large city
14
12
10
8
6
4
2
0
Frequency
90
110
130
150
170
190
210
Cost of Electricity in ($)

2.38


Cost of Electricity during July 2010 for one-bedroom apartments in a large city
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Percentage
90
110
130
150
170
190
210
Cost of ELectricity

2.38


Cost of Electricity during July 2010 for one-bedroom apartments in a large city
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Cumulative Percentage
70
90
110
130
150
170
190
210
Cost of Electricity

HOMEWORK

1

- Open any online shop/mall (amazon, lazada.com, etc.)
- Collect data on ONE categorical AND ONE numerical variables. Store those raw data accordingly (min. 20 data).
- Organize and Visualize those data into its appropriate table and display.
- Pay attention on how to make an excellent graph (page 49)
- Use Microsoft Excel in storing the data, organizing it and making the graphs.

THANK YOU