Data Visualization: Uncovering the Hidden Pattern in Data using Basic and New Quality Tools

Amar Sahay, Ph.D.

American Society for Quality (ASQ)

Section 1:
Introduction to Data Visualization

Overview:

This presentation examines the graphical (or visual) tools as well as the information visualization tools widely used in data analysis, visualization and quality improvement to analyze, enhance and improve the quality of products and services.

Overview…cont.

- Visual tools are an easy way to gain a first look at your data.
- They have been used to gain an insight into the data before applying more complex analysis.
- We examine a collection of visual and graphical tools.
- Visual tools are commonly referred to as graphical tools.
- In Quality and Data Science, several charts and graphs are commonly used to create visuals that provide a quick summary, trends and patterns in the data which are not usually apparent from the data in raw form.

Graphical/Visual Techniques

- It is said that a picture is worth a thousand words; this is particularly true when a large set of data is effectively presented using charts and graphs that quickly reveal important features.
- Visual displays of the data are easily recognizable and found ubiquitously in business periodicals, financial magazines, on the internet, and televisions.

Visual Techniques: Benefits

- The techniques will help us gain insight into the way the variable or variables seem to behave.
- The visual techniques enable one to understand how the values of a random variable under study are distributed.
- The shapes produced using the graphical techniques help select an appropriate theoretical distribution for the random variable in question. These are critical in drawing conclusions from the data.
- The charts and graphs help us visualize the important characteristics of data which are usually not apparent from the raw data, for example identifying the trends or patterns.
- Some of the visual representations of data provide excellent means of comparing data from processes, checking the variation, and taking corrective actions when deviations from stable conditions occur.

Examine the techniques of :

- Summarizing and describing data using charts graphs.
- Constructing a frequency distribution from data
- Constructing different types of graphs using quantitative data including histograms, frequency polygons, ogives, stem-and-leaf plots, dot plots and interpret these plots
- Constructing charts and pie charts using qualitative (categorical) data and learn their applications
- Constructing other types of charts and graphs including time series plots and scatter plots

Use of computer packages to construct visualizations

Quality Tools

*These are a set of **graphical and information visualization tools.***

Developed and used over the years in quality improvement and Lean Six Sigma programs.

Tools are also widely used in:

1. **Data science & data analysis,**
2. **Healthcare, Finance,**
3. **Product and process design,**
4. **Big Data Analysis, and others**
5. **Process improvement,**
6. **Manufacturing, Engineering,**
7. **Business Data Processing,**
8. **Lean Six Sigma**

These are powerful decision-making tools.

Basic tools of Quality

(1) Process Maps (2) Check sheets

(3) Histograms (4) Scatter Diagrams

(5) Run/Control Charts (6)Cause-and-Effect (Ishikawa)/Fishbone Diagrams

(7) Pareto Charts/Pareto Analysis

- We will discuss the construction and applications of the above graphical tools with numerous examples along with many other visualization tools.

Seven new tools of Quality

These tools are referred to as graphical & *information visualization* tools. They have wide applications in decision making and quality improvement programs.
We will discuss the following visualization tools :

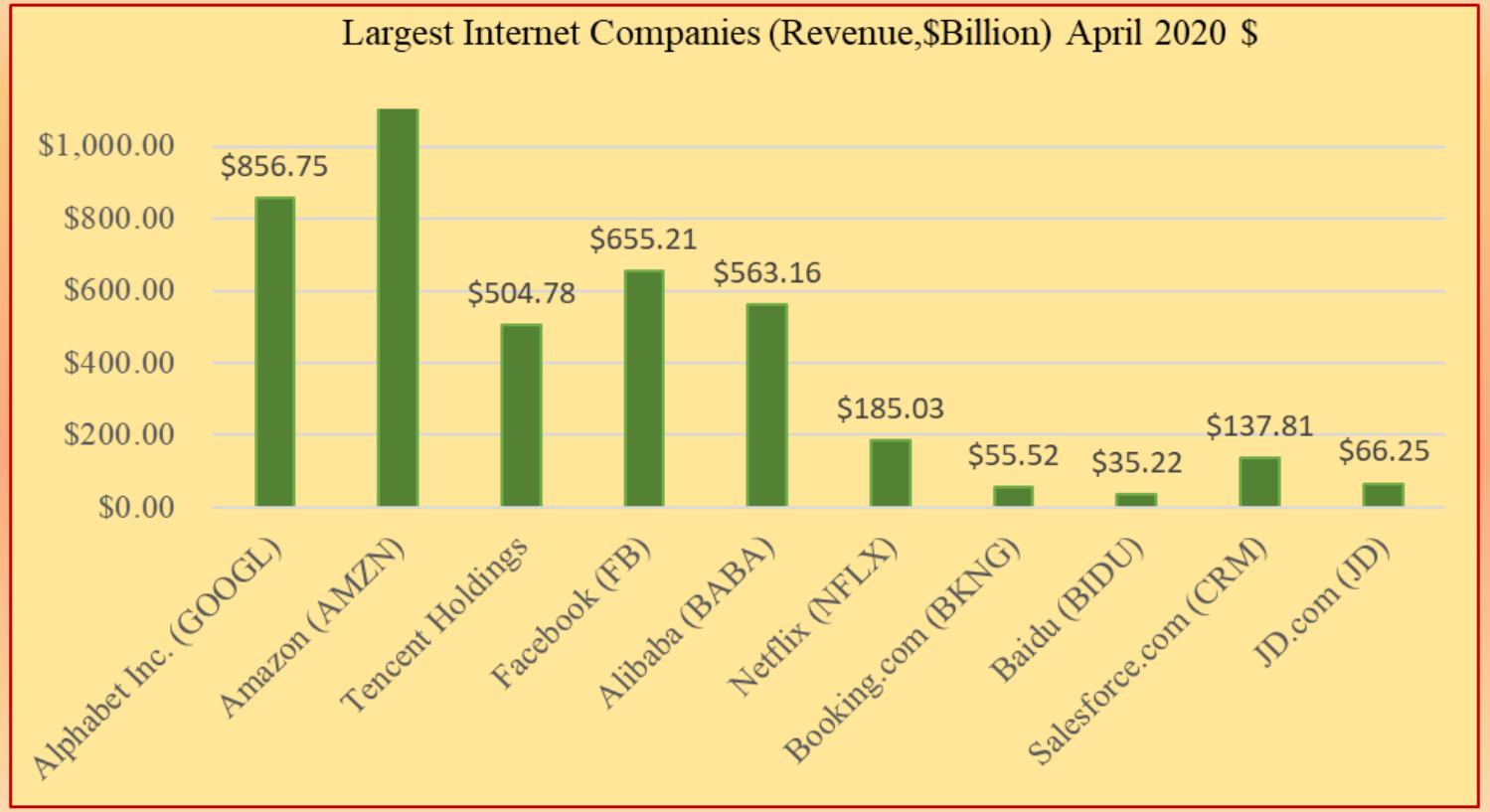
(1) Affinity Diagram
(2) Interrelationship Digraph
(3) Tree Diagram
(4) Prioritizing Matrices
(5) Matrix Diagram
(6) Process Decision Program Ch.
(7) Activity Network Diagram

Section 2:
Commonly used Visualization Tools

Before we discuss the basic and new quality tools, we present some commonly used visualization tools with current data.

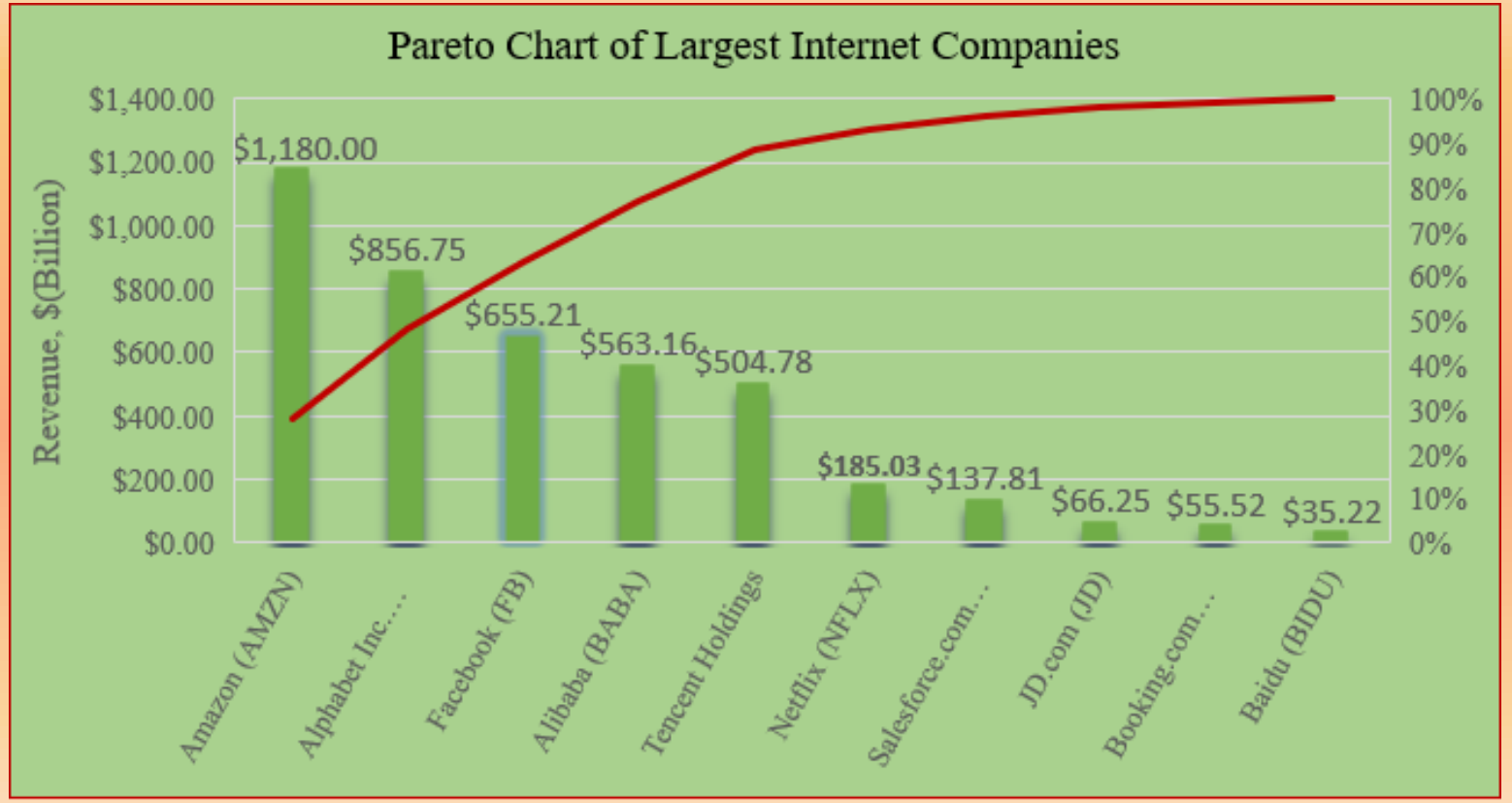
Real World Examples of Graphs_Bar Chart

Graphs summarizing the market value of largest Internet Companies April 2020

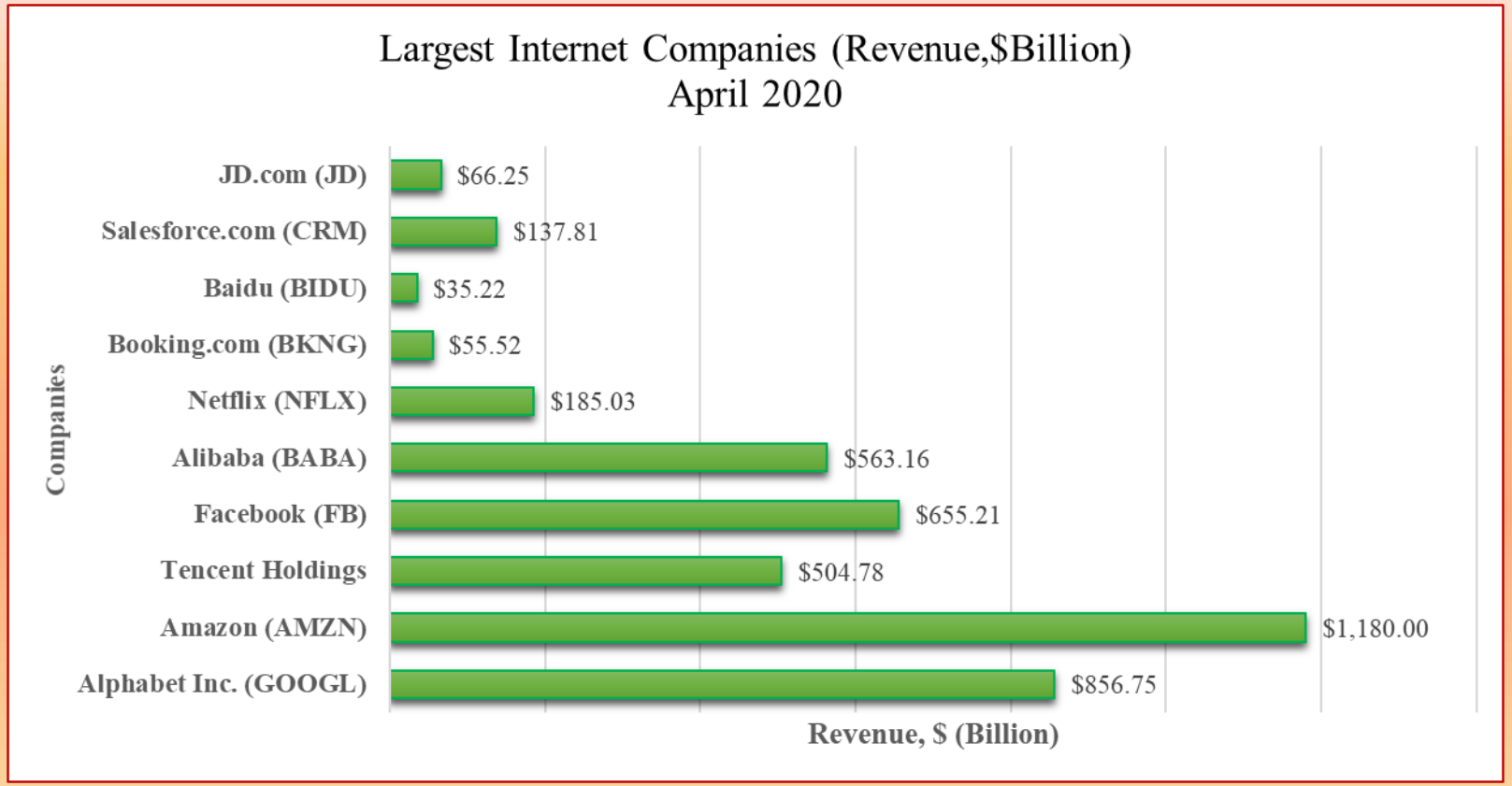


Real World Examples of Graphs_Pareto Chart

Graphs summarizing the market value of Internet Companies (Largest to Smallest) April 2020

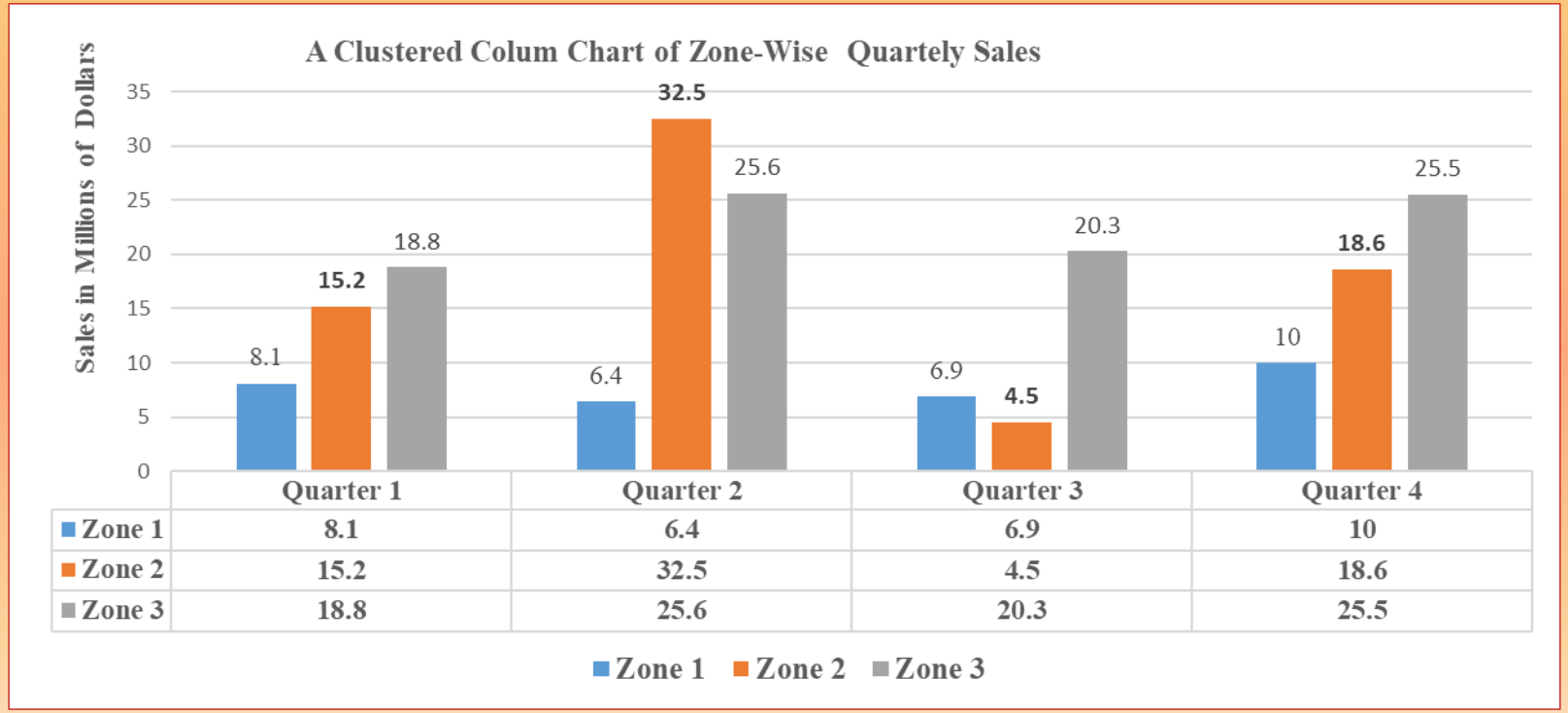


Horizontal Bar Chart



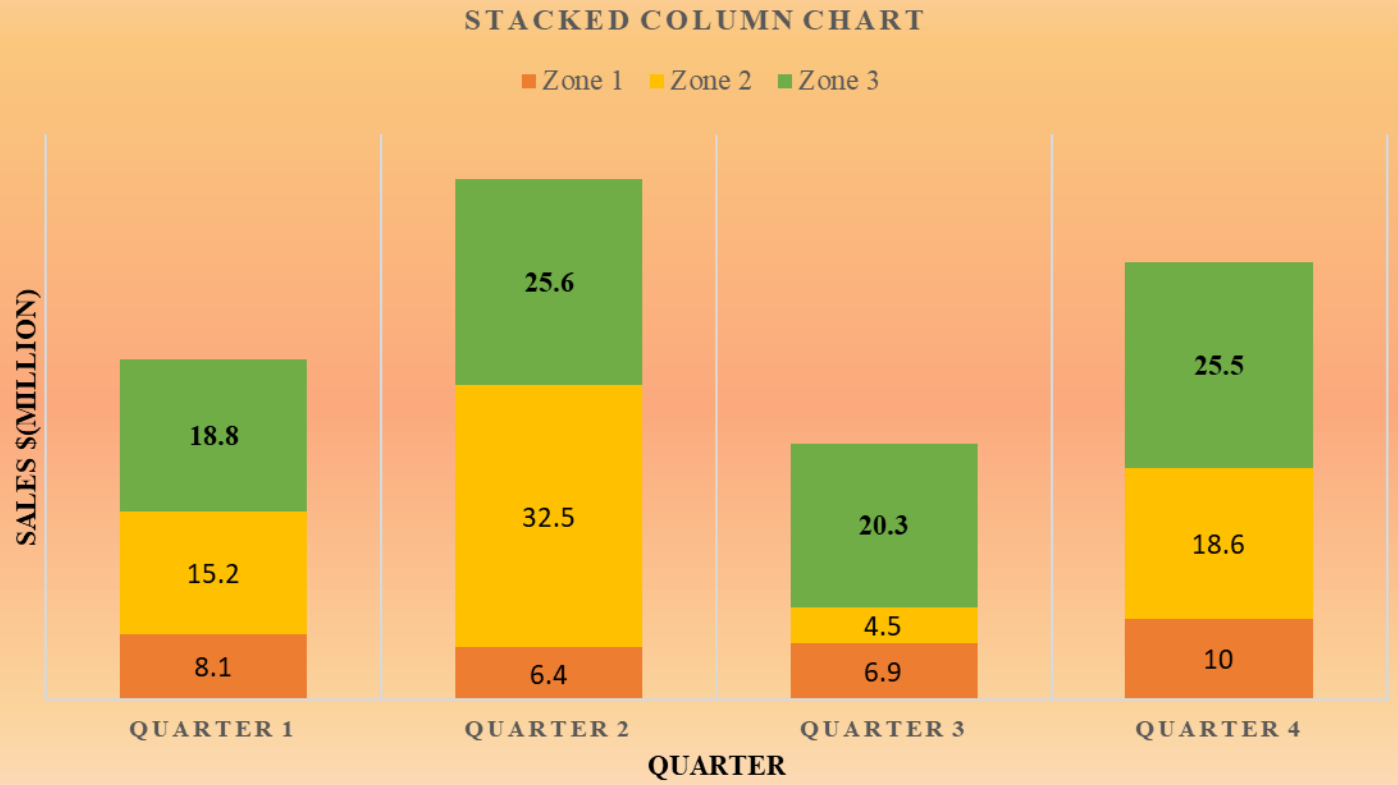
Variations of Bar Chart_A Clustered Column Chart

A clustered column chart is useful in comparing values across a few categories when the order of categories is not important.



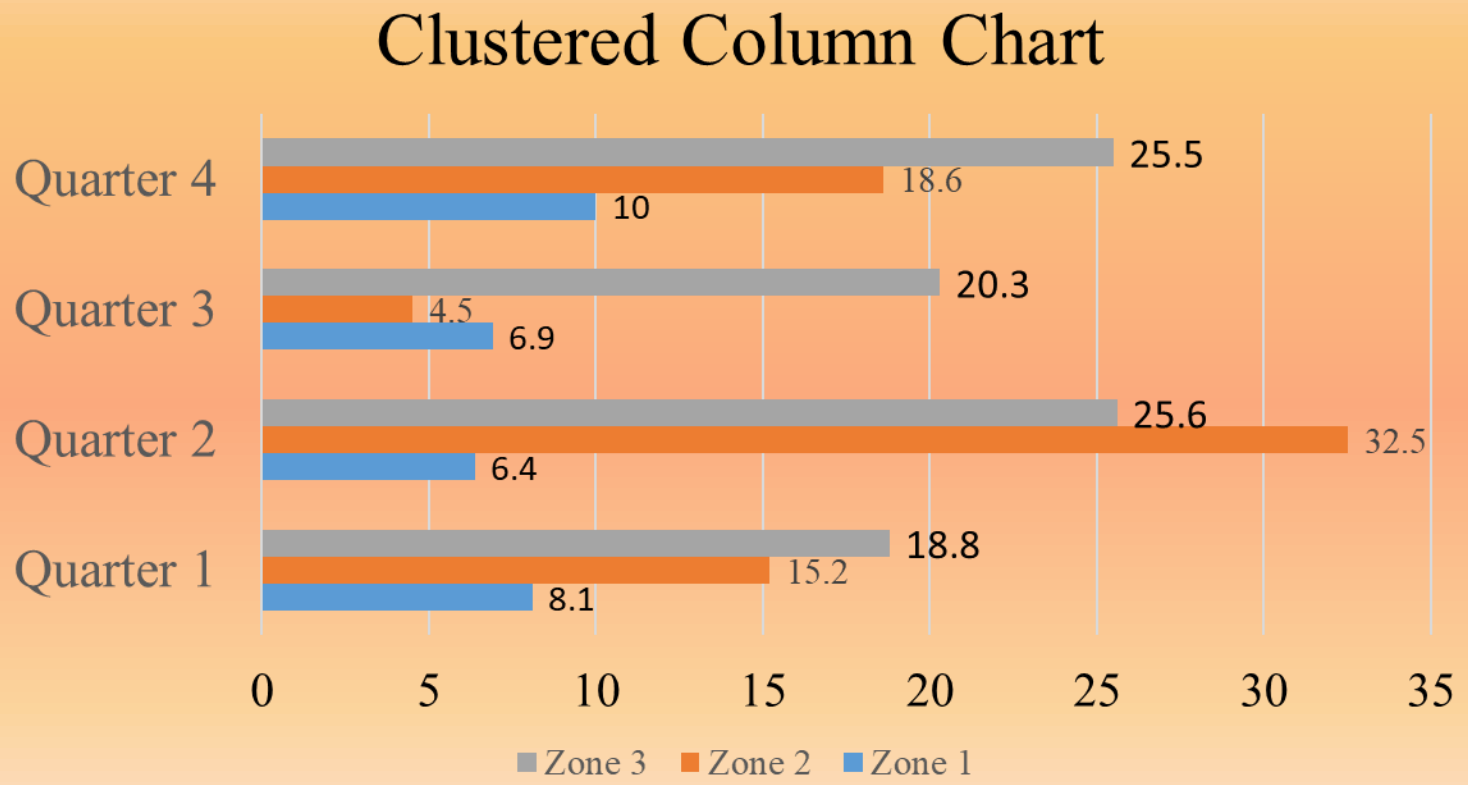
A Stacked Column Chart

A stacked column or chart shown below is used to compare parts of a whole. The chart is used to show how the parts of a whole change over some period.



Clustered Bar Column Chart

This chart is used to compare the values in a variable (sales) across different categories.

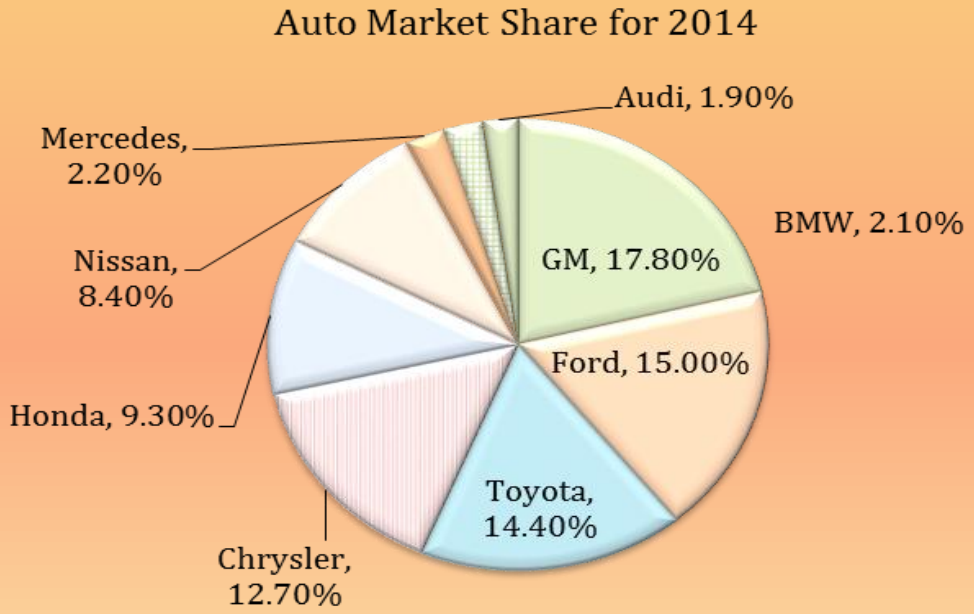


The Pie Chart

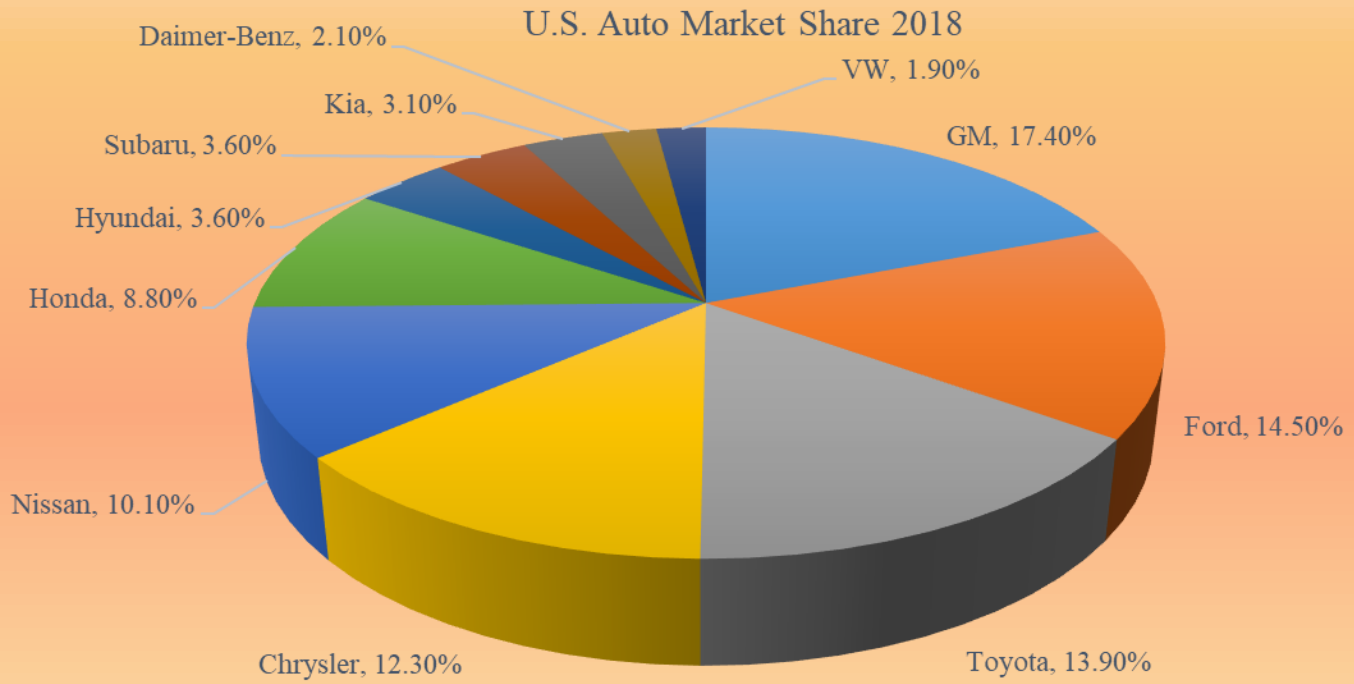
- A pie chart is used to show the relative magnitudes of parts to a whole.
- In this chart relative frequencies of each group of data are plotted.
- A circle is constructed and is divided into distinct sections. Each section represents one category of data.
- The area of each section is determined by multiplying the relative frequency of each section by the angle of a circle. Since there are 360 degrees in a circle, each section is multiplied by 360 degrees to obtain the correct number of degrees to represent each section.

Examples of Pie Chart

Graphs summarizing the Auto Market Share for 2014

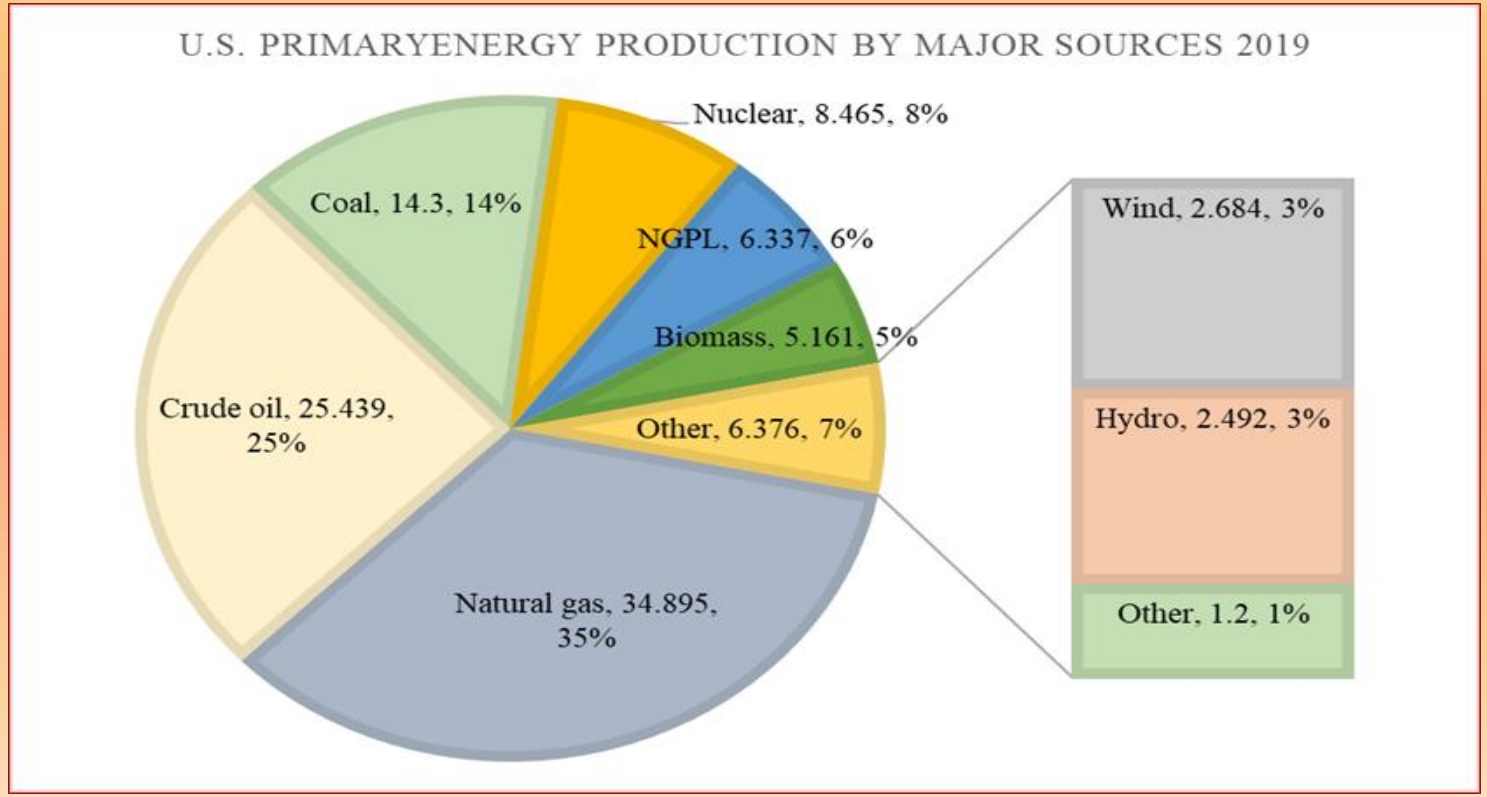


Graphs summarizing the Auto Market Share for 2018

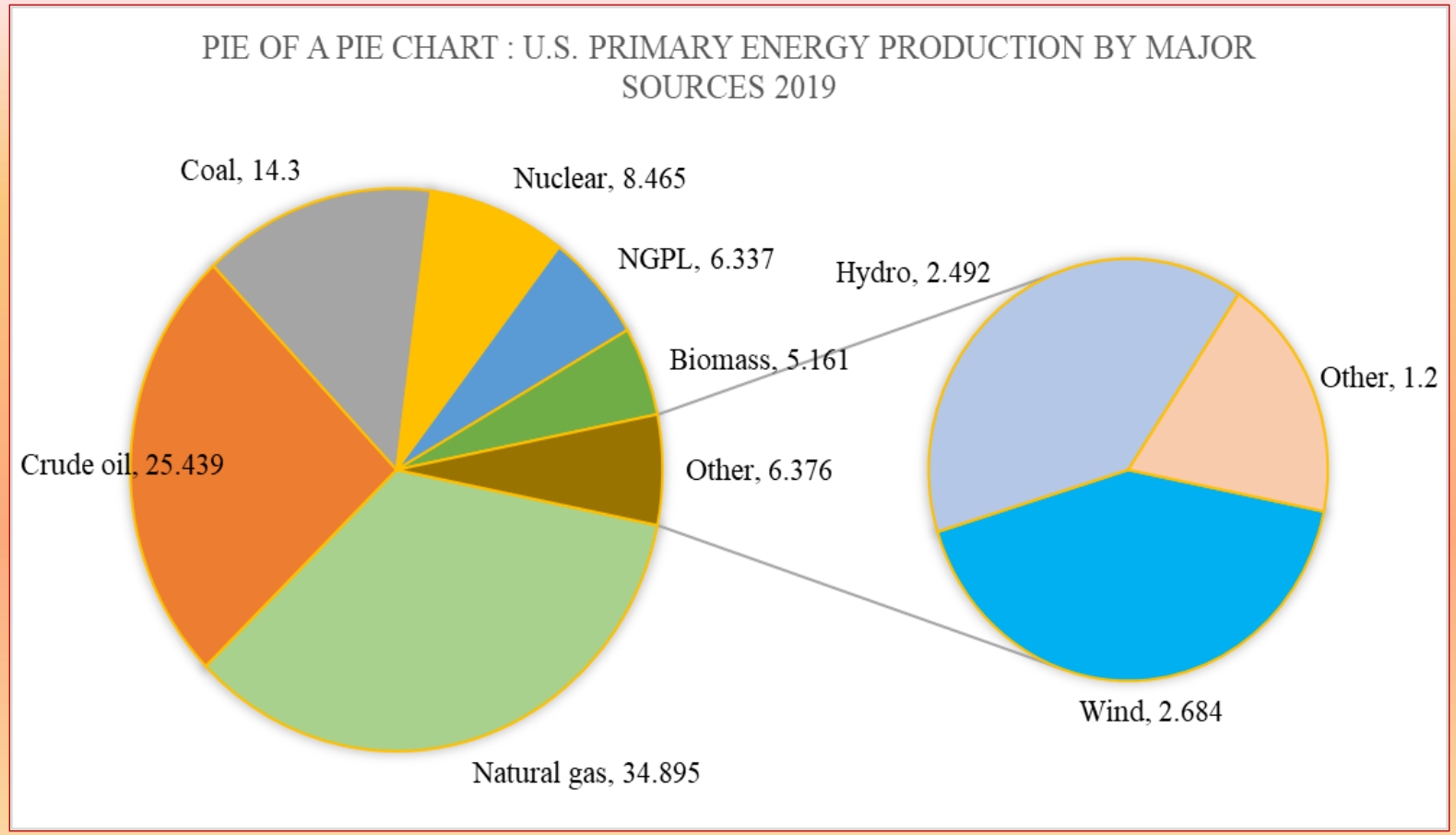


Variations of Pie Chart_Bar of a Pie Chart

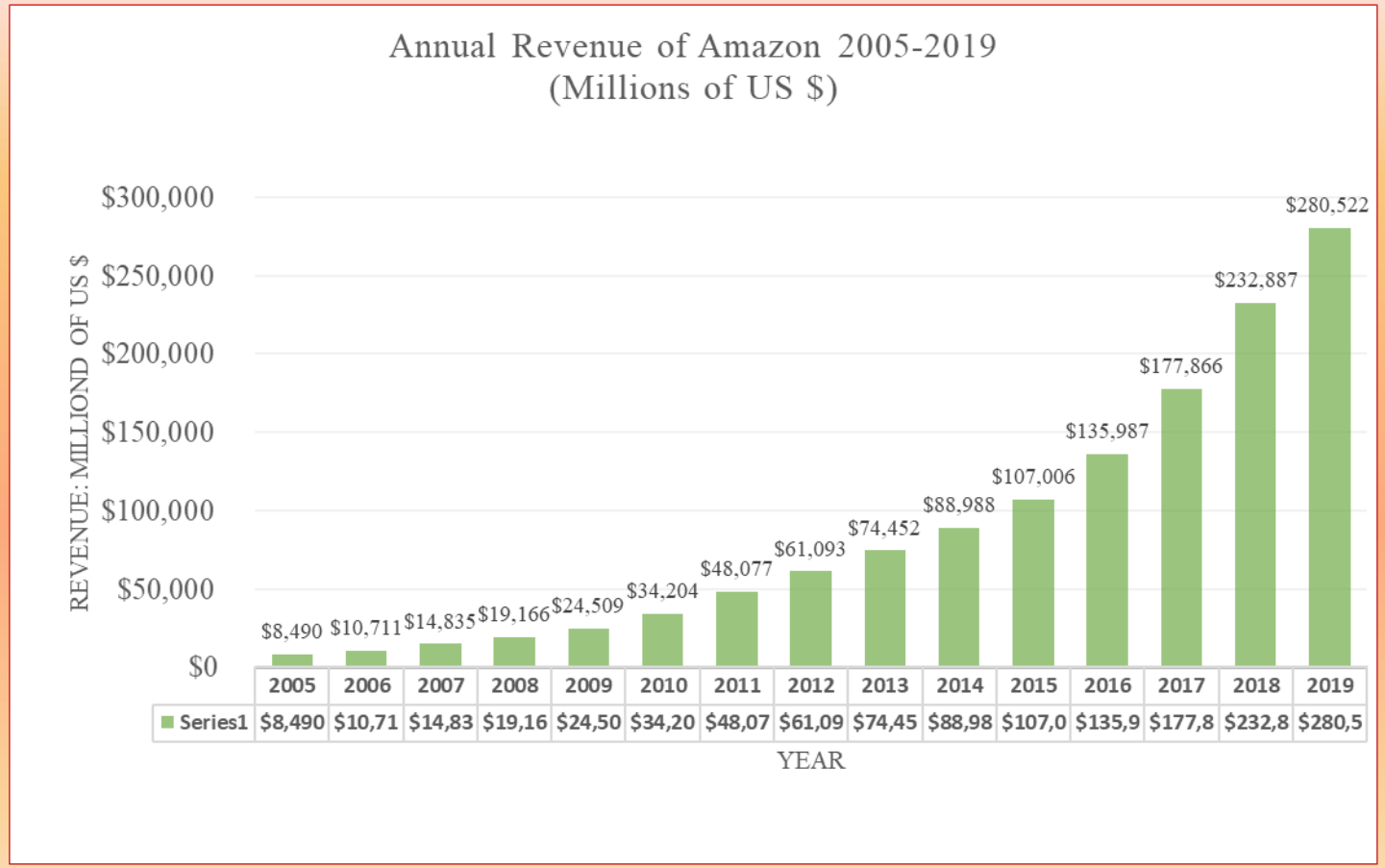
This chart is useful in displaying the categories with small percentages as a separate bar or a pie chart that is created as an extension of the main chart.



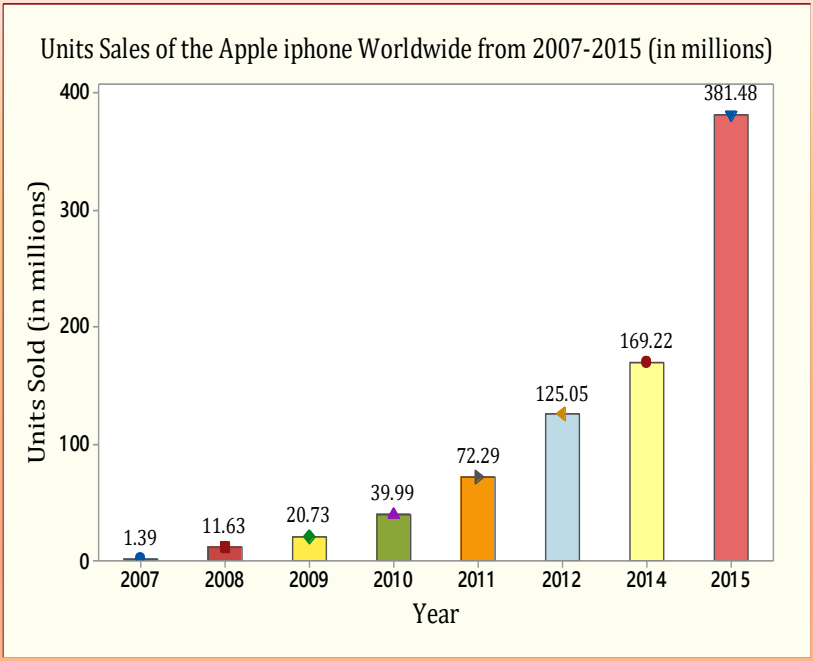
Variations of Pie Chart_Pie of a Pie Chart

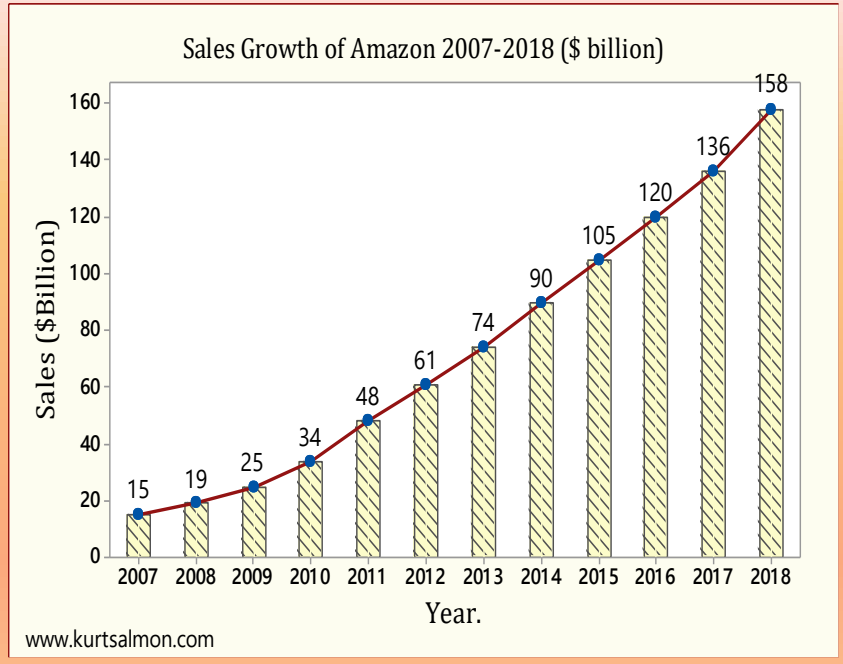


Bar Chart_ Annual Revenue of Amazon (2005-2019)



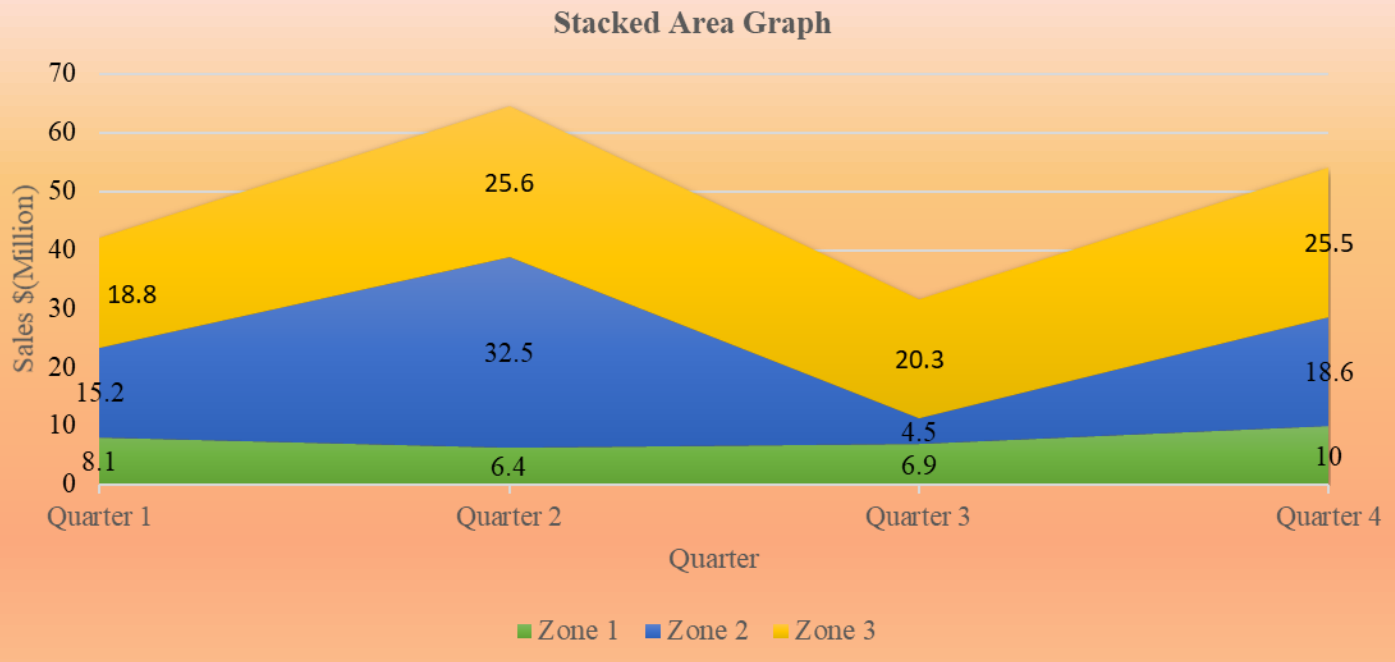
More Examples of Graphs





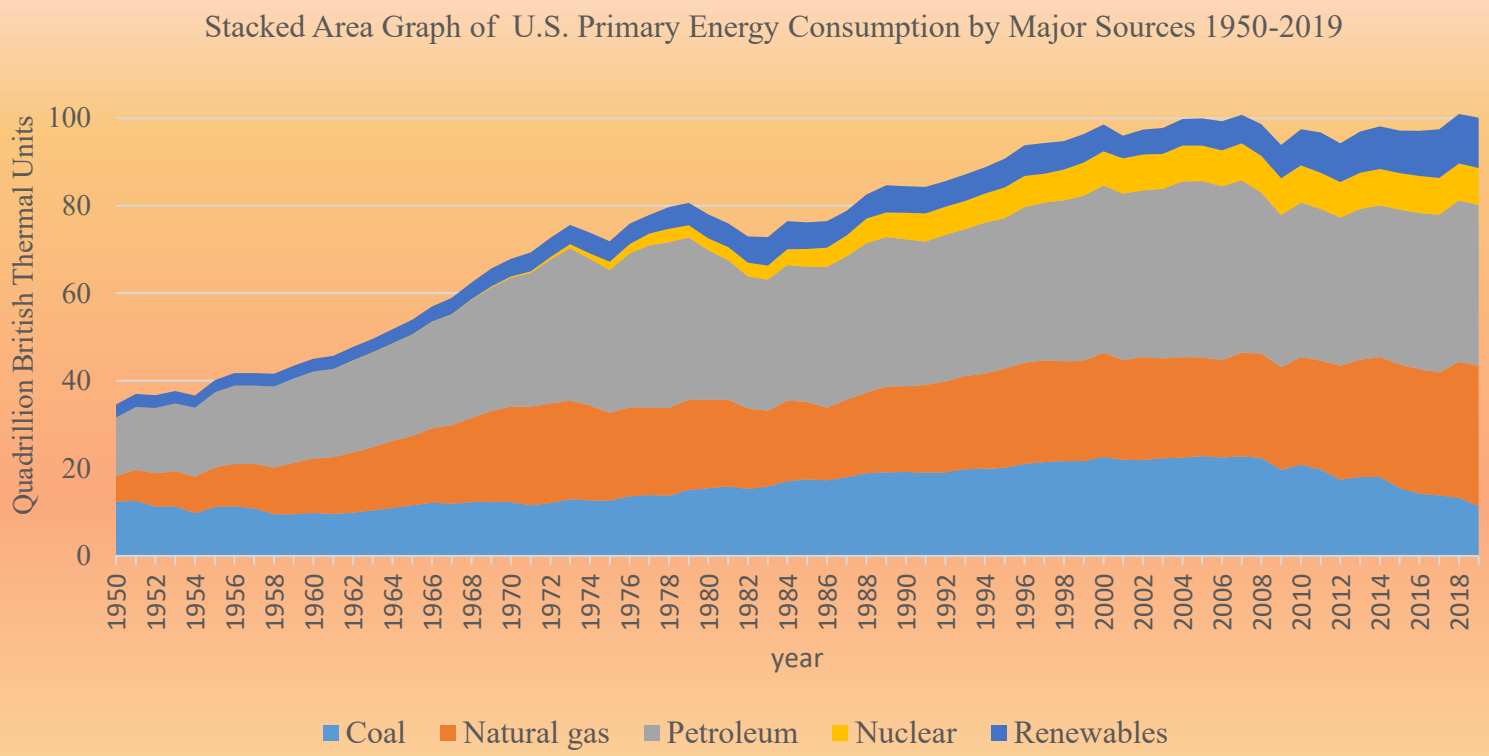
A number of charts and graphs can be found in in reports of financial periodicals like ***The Economist***, ***Business Week***, ***Fortune*** and many others. Almost every issue of ***USA Today*** and ***The Wall Street Journal*** contain a number of visual displays in their articles.

Stacked Area Chart

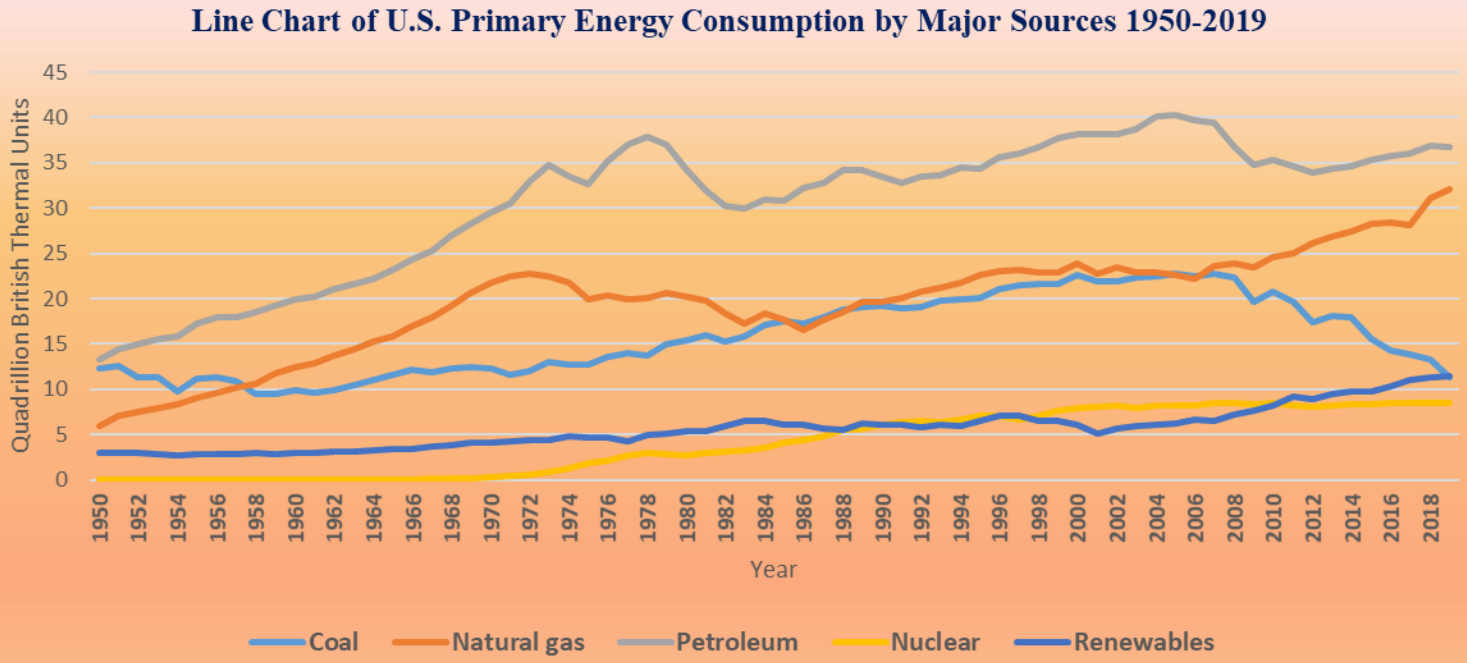


A stacked area chart is used to show the relationship of parts to whole over time or different categories.

Stacked Area Chart – another example



Line Chart

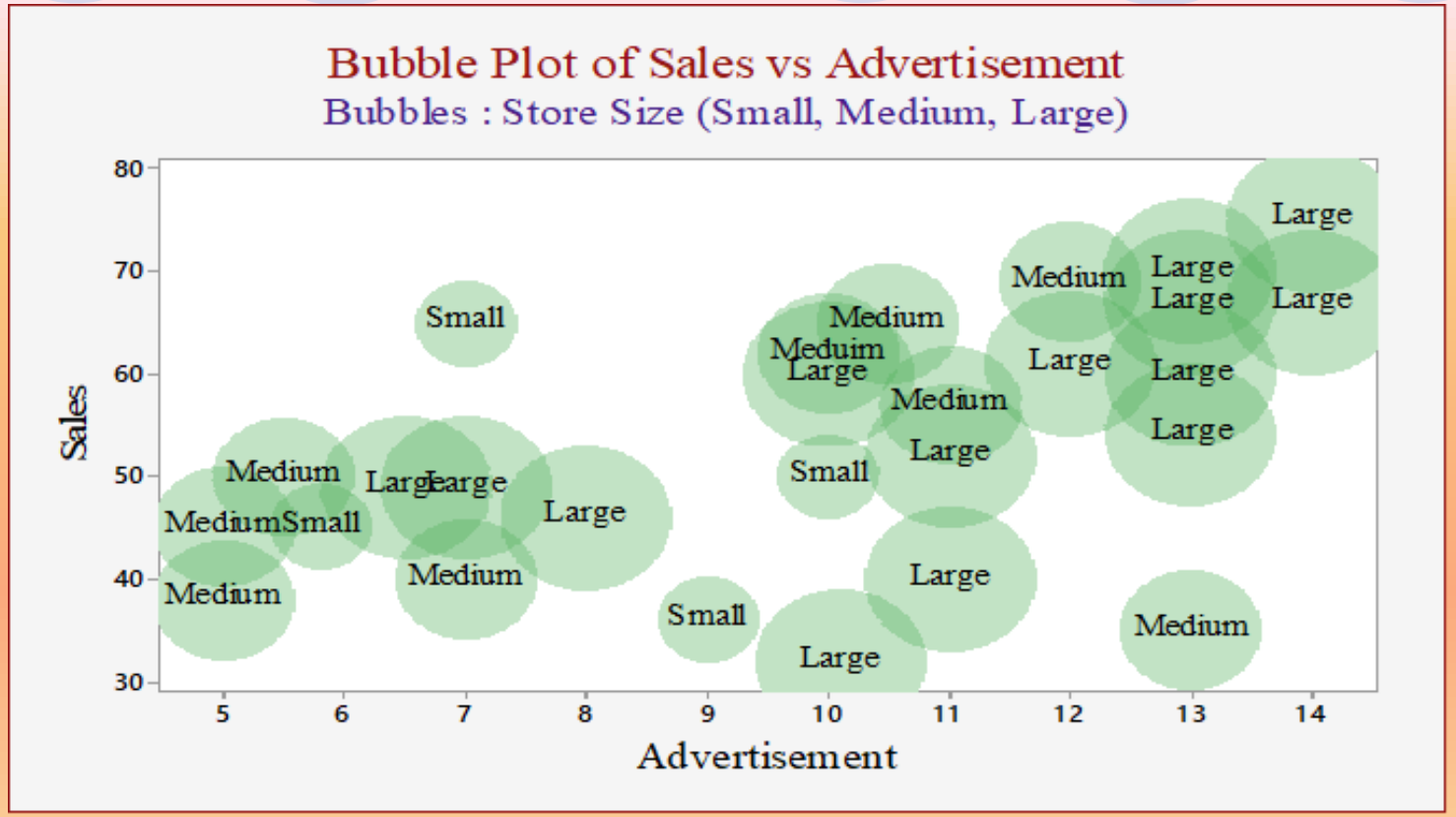


Line Chart of U.S. primary energy consumption by major sources from 1950 to 2019. This chart shows the trends over time (days, months, years, etc.) or categories when the order is important. The chart can be used for large data sets when the order of data is important as in time-series.

Bubble Graph_ Describing three variables

The bubble graph or chart is like the scatter plot but instead of two variables, it shows the relationship between three variables. The value of the third variable is determined by the relative size of the bubble. Data for advertisement, sales for different store sizes (small, medium, large labeled 3,2, and 1 respectively).

Bubble Graph_ Another Variation



The third variable is store sizes (small, medium, large) shown by the size of the bubble.

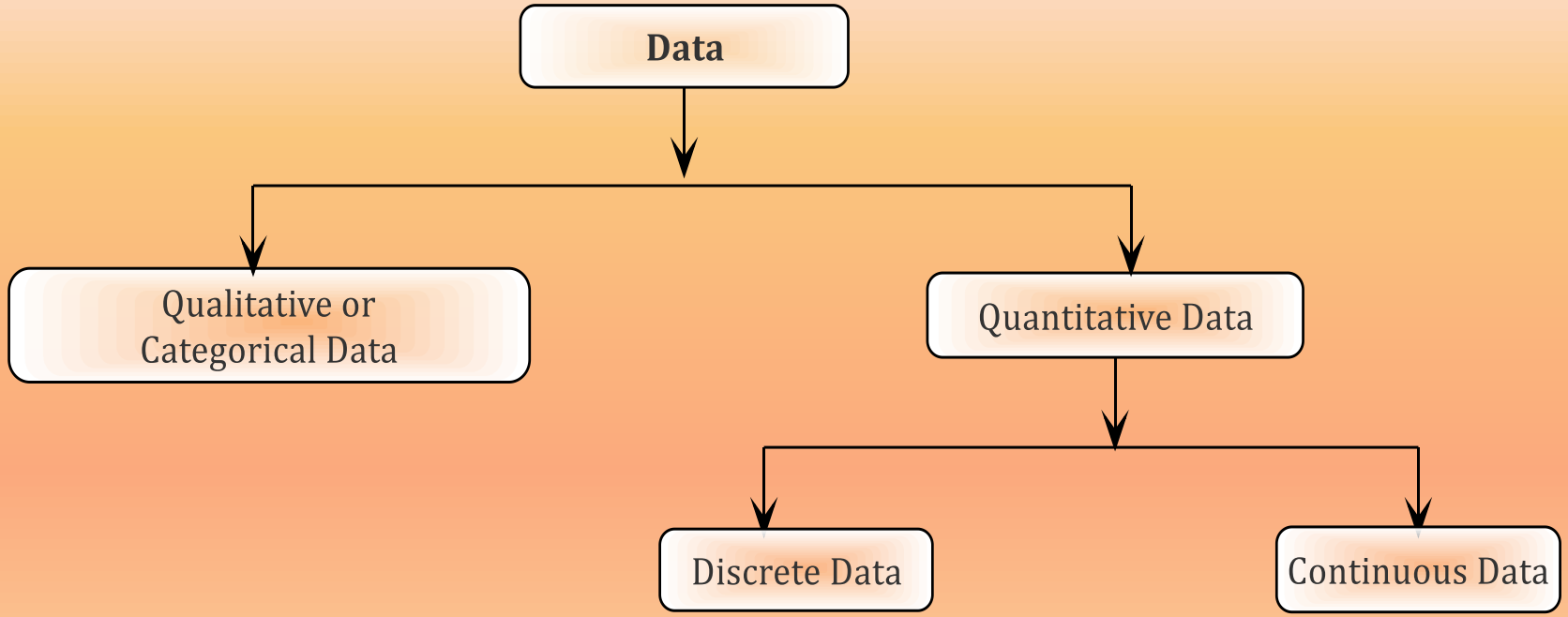
Section 3:
Construction and Application of other Types of Charts

Application of other Types of Graphs

- Charts and graphs that are used to describe the key features of data.
- How these graphs enable one to summarize a large sets of data.
- Create Graphs of both the quantitative and categorical data
- Computer software including Excel and MINITAB to construct and interpret the charts and graphs.

A review of major data categories is described.

Classification of Data: Review



Different Forms of Data

Lifetime of 200 Television Components (in Hours)
Ungrouped and Raw Data

314	330	371	365	267	307	371	297	291	398
276	253	286	344	385	349	269	304	319	283
430	253	378	306	376	308	339	368	289	344
340	298	330	311	318	358	354	406	369	254
322	242	331	236	344	418	328	393	267	305
325	282	315	328	319	353	336	384	298	398
343	203	373	297	276	333	257	367	296	349
322	325	252	345	373	317	307	289	363	340
309	246	302	260	292	231	338	372	226	365
271	302	331	374	355	336	312	354	329	345
276	329	379	288	356	302	263	364	337	361
416	360	337	273	298	390	215	382	329	306
306	279	414	262	372	303	346	331	362	366
387	304	302	280	287	368	281	329	309	310
375	346	413	309	283	299	335	330	376	260
277	366	345	409	312	266	383	289	294	370
359	363	243	339	323	297	333	299	302	384
370	357	314	348	257	291	358	409	337	347
215	277	313	300	322	304	282	410	390	332
373	280	339	349	363	297	274	334	359	330

Data Array: Data Arranged in Increasing Order
(Read row-wise)

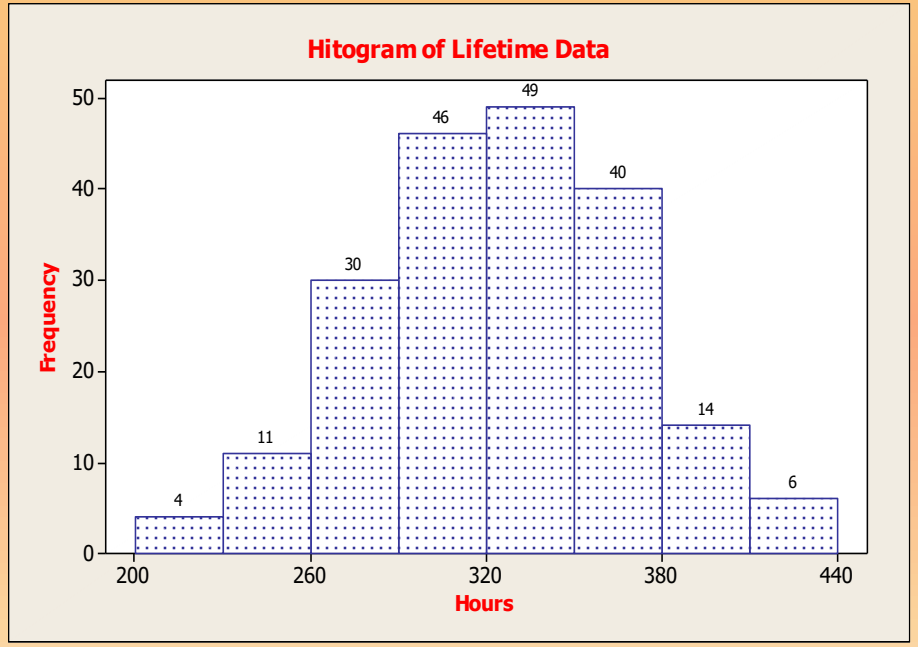
203	215	215	226	231	236	242	243	246	252
253	253	254	257	257	260	260	262	263	266
267	267	269	271	273	274	276	276	276	277
277	279	280	280	281	282	282	283	283	286
287	288	289	289	289	291	291	292	294	296
297	297	297	297	298	298	298	299	299	300
302	302	302	302	302	303	304	304	304	305
306	306	306	307	307	308	309	309	309	310
311	312	312	313	314	314	315	317	318	319
319	322	322	322	323	325	325	328	328	329
329	329	329	330	330	330	330	331	331	331
332	333	333	334	335	336	336	337	337	337
338	339	339	339	340	340	343	344	344	344
345	345	345	346	346	347	348	349	349	349
353	354	354	355	356	357	358	358	359	359
360	361	362	363	363	363	364	365	365	366
366	367	368	368	369	370	370	371	371	372
372	373	373	373	374	375	376	376	378	379
382	383	384	384	385	387	390	390	393	398
398	406	409	409	410	413	414	416	418	430

Frequency Distribution – Grouping and Graphical Presentation – Example 1

Frequency Distribution (Grouped Data)

Class- interval	Frequency (f)
200 - 230	4
230 - 260	11
260 - 290	30
290 - 320	46
320 - 350	49
350 - 380	40
380 - 410	14
410 - 440	6
	$\sum f = 200$

Graph of frequency Distribution or Histogram



Frequency Distribution

Frequency distribution provides a compact representation of data.

This is another name for grouping

- Compact representation is obtained by ***arranging the data into groups*** or ***class intervals*** usually of ***equal width*** , and
- recording or counting the number of observations in each interval. Counting the number of observations in each group is called the ***class frequency***

A grouped data or frequency distribution would look like:

Class- interval	***Frequency***
200 – 230	4
230 - 260	11
260 – 290	30
:	
and so forth	

The class interval 200 – 230 means 200 ≤ x < 230 (includes the value 200 but not 230). The value 200 is known as the ***lower-class boundary or lower class limit*** and the value 230 is known as the ***upper class boundary or upper class limit.***

Frequency Distribution: Example 2

Table shows data for speed of 100 cars in miles per hour (mph) passing through a highway intersection with a 60-mph speed limit. These cars were randomly selected and represent a sample of n=100.

Table 3.1: Driving Speed (mph) **Raw data**

51	46	62	70	54	59	59	57	61	66	49	57	57	65	61	62	51	63	62	65	55	55	65
64	60	55	70	61	63	55	70	65	51	53	49	62	56	61	64	54	60	63	69	72	69	60
57	63	60	56	60	61	57	57	61	54	58	55	69	63	55	58	58	62	59	59	62	53	69
56	59	57	60	63	60	56	52	65	58	60	62	54	57	60	53	56	60	71	59	64	58	71
68	62	61	61	67	59	58	49															

Table 3.2: Driving Speed (mph) - (Sorted Data) **Data array or Ordered array**

46	49	49	49	51	51	51	52	53	53	53	54	54	54	54	55	55	55	55	55	55	56	56
56	56	56	57	57	57	57	57	57	57	57	58	58	58	58	58	58	59	59	59	59	59	59
59	60	60	60	60	60	60	60	60	60	60	61	61	61	61	61	61	61	62	62	62	62	
62	62	62	62	63	63	63	63	63	63	64	64	64	65	65	65	65	65	66	67	68	69	69
69	69	70	70	70	71	71	72															

Summarizing Quantitative Data: Frequency Distribution

- A *frequency distribution* provides a compact representation of data.
- This is also known as grouping. Compact representation is obtained by arranging the data into groups or ***class intervals*** usually of ***equal width*** and then recording or counting the number of observations in each interval.
- Number of observations in each group is called the ***class frequency***.
- For example, examine the ordered data in previous slide. We can divide this data into 10 class intervals with a width of 3 and tabulate the results as shown below.

Class- interval	***Frequency (f)***
45 – 48	1
48 – 51	4
51– 54	9….. and so on.

The above class frequency is an example of a frequency distribution. The class interval of 45 - 48 means that this interval contains all the values from 45 to 48 (not including 48). If we count the number of observations between

Frequency Distribution…cont.

- There is no unique frequency distribution for a given set of data
- Several frequency distributions are possible for the same set of data
- When dividing the data into class intervals, 5 to 15 class intervals are recommended
- If there are ***too many class intervals***, the class frequency (count) is low and the saving in computational effort is small
- If there are too few class intervals, the true characteristic of the distribution may be obscured and some information may be lost.

Frequency Distribution…cont.

- How many class-intervals or groups?

The ***number of class intervals*** should be governed by the ***amount*** and ***scatter*** of data present.

An ***estimate*** for the number of class intervals can be calculated using the following formula:

where, K = number of class intervals

$$K = 1 + 3.33 \log_{10} n$$

- If there are n=200 observations then using the above formula,

$K = 1 + 3.33 \log_{10} 200$

$= 1 + 3.33 (2.3010)$

$= 8.66$ or approximately 9 classes

Frequency Distribution…cont.

- Class intervals should be chosen so that no result falls on a class boundary.
- Class intervals may be written two different ways

Class Intervals 1 (upper boundary inclusive)

200 - 229

230 - 259

260 – 289

:

and so on (In this case, the upper boundary is inclusive)

Class Intervals 2 (upper boundary exclusive)

200 - 230

230 – 260

260 - 290

:

and so forth (In this case, the upper boundary is exclusive)

Frequency Distribution…cont.

- What should be the width of each class?

The class width is of equal size, the number of classes determines the width of each class. Use the following formula to determine the class width:

$$\textit{Width of class interval} = \frac{\textit{(Largest value in the data - Smallest value)}}{\textit{Number of Class Intervals}}$$

If the number of observations, n= 200 and K= number of classes is 9. Using these values, the minimum and maximum values in the data are 203 and 430 then class width,

$$= \frac{430 - 203}{8} = 28.375$$

Note: The number of class intervals and the class width using the above formulas are approximate and are not exact. They can be adjusted to get desired class intervals and width.

Summarizing Quantitative Data: Frequency Distribution – another example

Table 2.1:Driving Speed (mph), n=100

51	46	62	70	54	59	59	57	61	66	49	57	57	65	61	62	51	63	62	65	55	55	65
64	60	55	70	61	63	55	70	65	51	53	49	62	56	61	64	54	60	63	69	72	69	60
57	63	60	56	60	61	57	57	61	54	58	55	69	63	55	58	58	62	59	59	62	53	69
56	59	57	60	63	60	56	52	65	58	60	62	54	57	60	53	56	60	71	59	64	58	71
68	62	61	61	67	59	58	49															

Table 2.2: Driving Speed (mph) - (Sorted Data)

46	49	49	49	51	51	51	52	53	53	53	54	54	54	54	55	55	55	55	55	55	56	56
56	56	56	57	57	57	57	57	57	57	57	58	58	58	58	58	58	59	59	59	59	59	59
59	60	60	60	60	60	60	60	60	60	60	61	61	61	61	61	61	61	61	62	62	62	62
62	62	62	62	63	63	63	63	63	63	64	64	64	65	65	65	65	65	66	67	68	69	69
69	69	70	70	70	71	71	72															

Forming Frequency Distribution

1. Approximate number of classes can be found using the formula:

$$K = 1 + 3.33 \log_{10} n$$
$$K = 1 + 3.33 \log_{10} 100$$
$$K = 1 + 3.33(2.0)$$

= 7.66 or approximately 8 classes

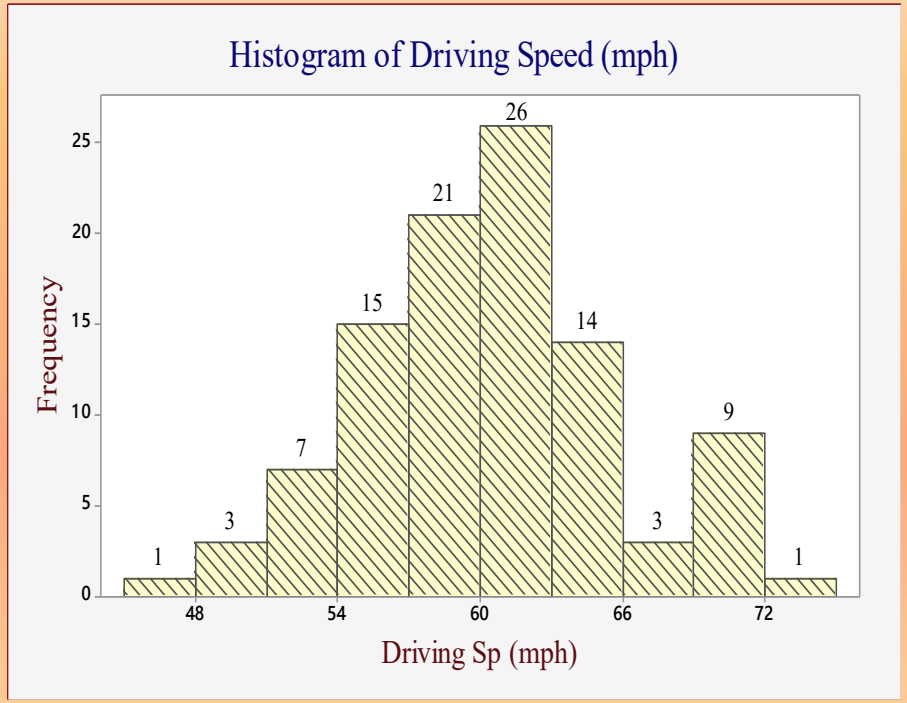
2. Class- width using 10 class intervals:

$$Class\ width = \frac{72 - 46}{10} = 2.6$$

Frequency Distribution and Histogram

Frequency distribution of 100 drivers with 60 miles per hour (mph) speed limit

Class-interval (mph)	Frequency (f)
45- 48	1
48 - 51	3
51 - 54	7
54 - 57	15
57 - 60	21
60 - 63	26
63 - 66	14
66 - 69	3
69 - 72	9
72 - 75	1
Total	$\sum f_i = 100$



Histogram of Driving Speed (mph)
(10 class-intervals)

Different Frequency Distribution for the Same Data Set

Table 2.4: Utility Charges of 50 Customers (rounded to nearest dollar)

Utility Charge ($)												
98	173	204	180	149	104	155	199	129	84	159	187	92
118	174	113	150	215	132	167	143	151	208	177	125	130
146	170	111	169	97	165	152	156	132	145	189	168	141
151	110	121	185	153	116	137	193	139	131	160		

Table 2.5: Sorted Data

Utility Charge (Sorted) – Read Row wise												
84	92	97	98	104	110	111	113	116	118	121	125	129
130	131	132	132	137	139	141	143	145	146	149	150	151
151	152	153	155	156	159	160	165	167	168	169	170	173
174	177	180	185	187	189	193	199	204	208	215		

Form a frequency distribution having: [a] Five class intervals, [b] Six class intervals, and [c] Seven class intervals.

Different Frequency Distribution for the Same Data Set …cont.

- Frequency Distribution with Five Class-intervals

  Determine the class-width using

Width = Range / Number of classes = (215-84)/5 = 26.2

Use a width of 30

Table 2.6: Frequency Distribution with Five Class Intervals

Class Intervals	Frequency(f_i)
80 - 110	5
110 - 140	14
140 -170	18
170 - 200	10
200 - 230	3
	$\sum f_i = 50$

Different Frequency Distribution for the Same Data Set …cont.

Frequency Distribution with Six and Seven Class-intervals

Width = (Maximum value – Minimum value) / Number of classes = (215-84)/6 = 21.8

Use a class-width of 25 in this case.

Table 2.7: Frequency Distribution with Six Class Interval

Class Intervals	Frequency (f_i)
75-100	4
100-125	7
125-150	13
150-175	16
175-200	7
200-225	3
	$\sum f_i = 50$

Table 2.8: Frequency Distribution with Seven Class Intervals

Class Intervals	Frequency (f_i)
80-100	4
100-120	6
120-140	9
140-160	13
160-180	9
180-200	6
200-220	3
	$\sum f_i = 50$

Types of Histogram

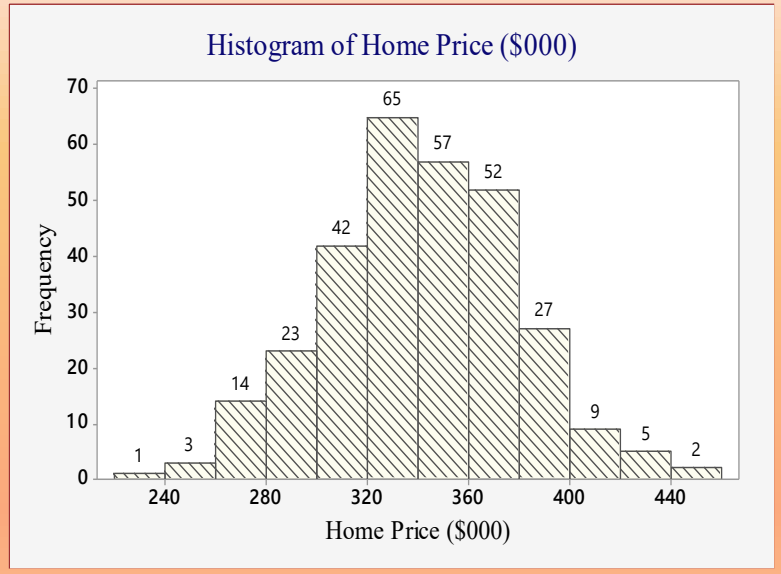
- Frequency Histogram
- Relative Frequency Histogram
- Percent Frequency Histogram

All of the above histograms have the same shape. The difference is in plotting the values on the y-axis.

Histogram: Summarizing the data and examining the distribution (shape)

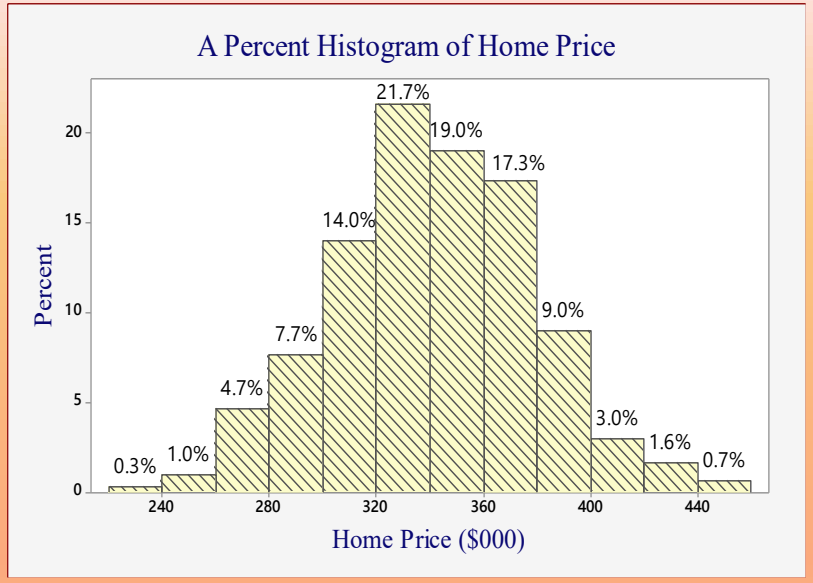
Frequency Distribution of Home Price

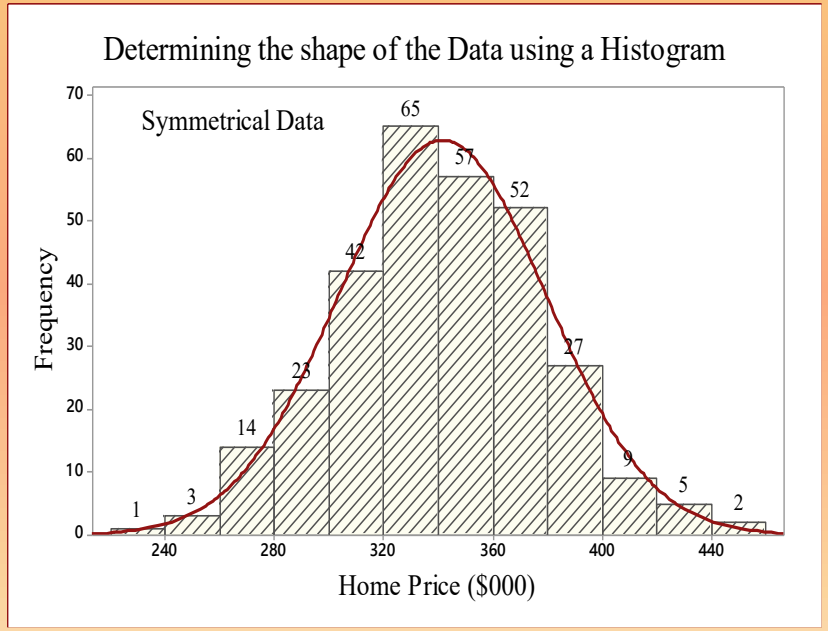
Class-interval Selling Price($000)	Frequency (f) No. of Houses	Relative Frequency
220 – 240	1	0.003
240 – 260	3	0.010
260 – 280	14	0.047
280 – 300	23	0.077
300 – 320	42	0.140
320 – 340	65	0.217
340 – 360	57	0.190
360 – 380	52	0.173
380 – 400	27	0.090
400 – 420	9	0.030
420 – 440	5	0.017
440 – 460	2	0.007
	$\sum f = 300$	Sum=1.00



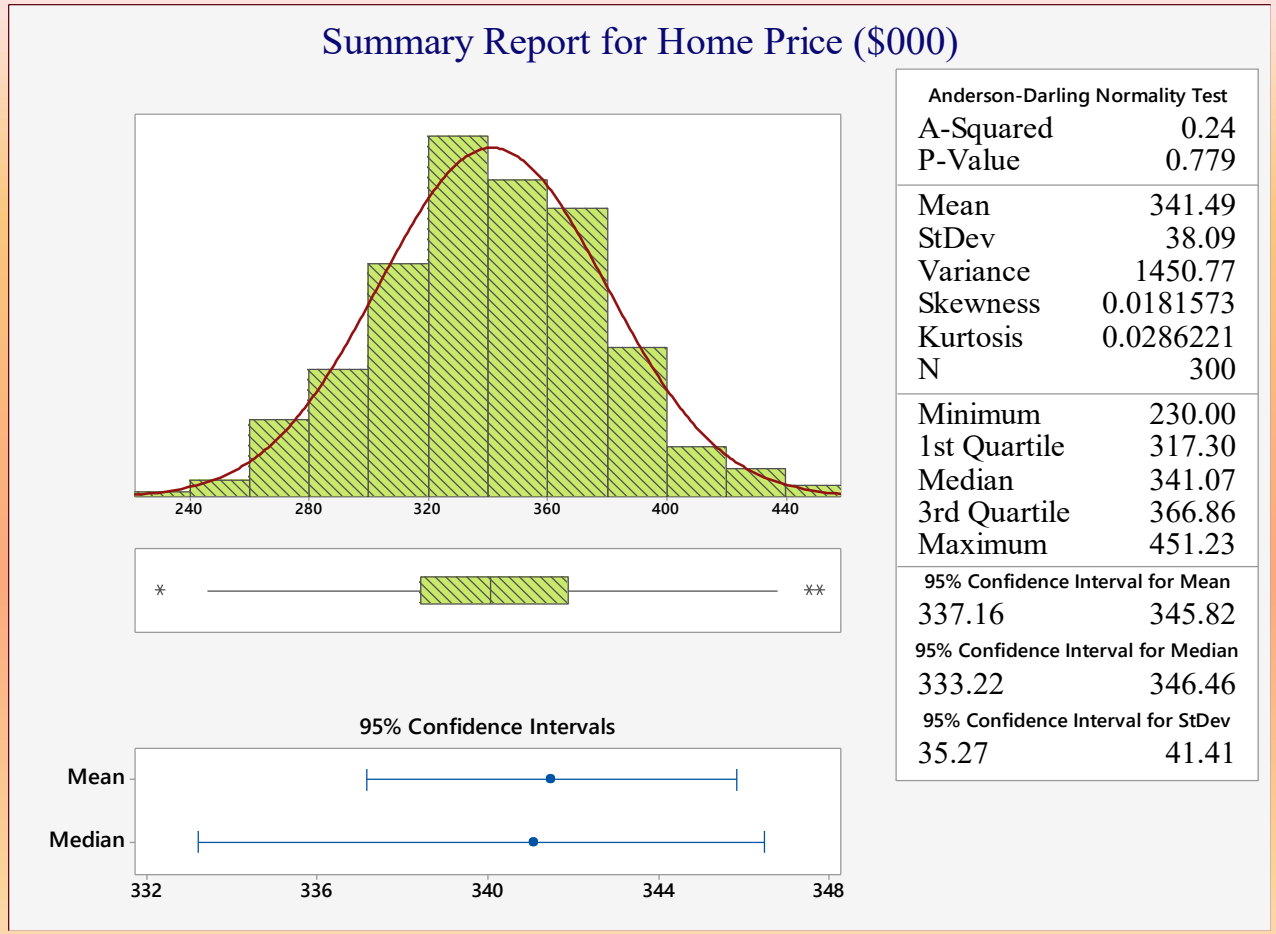
Frequency Histogram of Home Price

Determining the shape or distribution of the data using a histogram

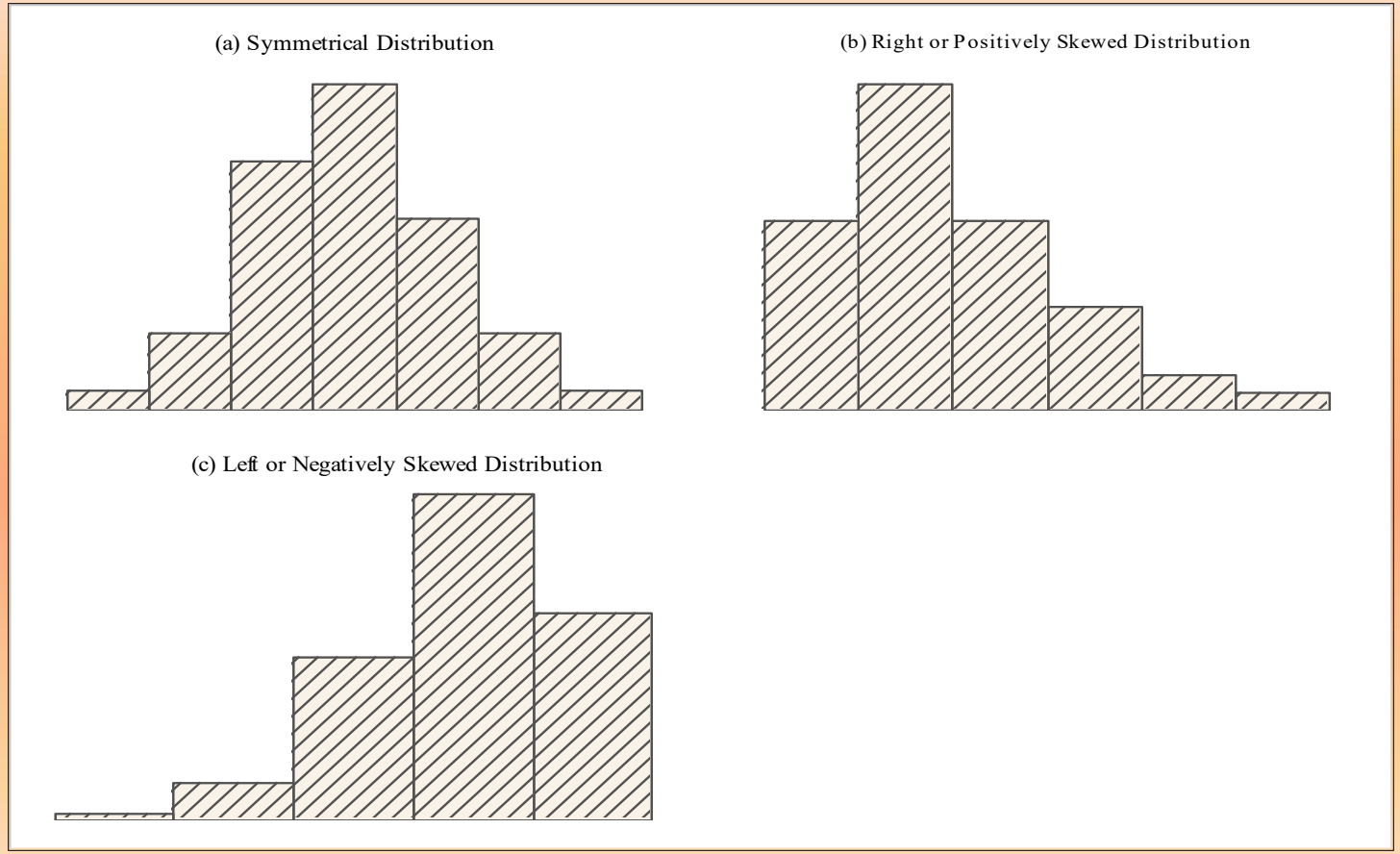




Graphical Summary of Data



(a) Symmetrical Distribution (b) Right or Positively Skewed Distribution (c) Left or Negatively Skewed Distribution



Example

A Frequency Distribution of 100 Drivers with 60 Miles per Hour Speed Limit is given below. Calculate the relative frequency, cumulative frequency, and relative cumulative frequency.

Class-interval (mph)	Frequency (f)
45- 48	1
48 - 51	4
51 - 54	9
54 - 57	14
57 - 60	25
60 - 63	24
63 - 66	10
66 - 69	5
69 - 72	7
72 - 75	1
Total	$\sum f_i = 100$

Solution

Class-interval, Frequency, Relative Frequency, Cumulative and Relative Cumulative Frequency

Class-interval (mph)	Frequency (f)	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
45- 48	1	0.01	1	0.01
48 - 51	4	0.04	5	0.05
51 - 54	9	0.09	14	0.14
54 - 57	14	0.14	28	0.28
57 - 60	25	0.25	53	0.53
60 - 63	24	0.24	77	0.77
63 - 66	10	0.10	87	0.87
66 - 69	5	0.05	92	0.92
69 - 72	7	0.07	99	0.99
72 - 75	1	0.01	100	1.00
Total	$\sum f_i = 100$	1.00		

(a) Find the percentage of drivers who were speeding (driving 66 mph or above).

13%

(c) Find the percentage of the drivers whose speed does not exceed 60 mph.

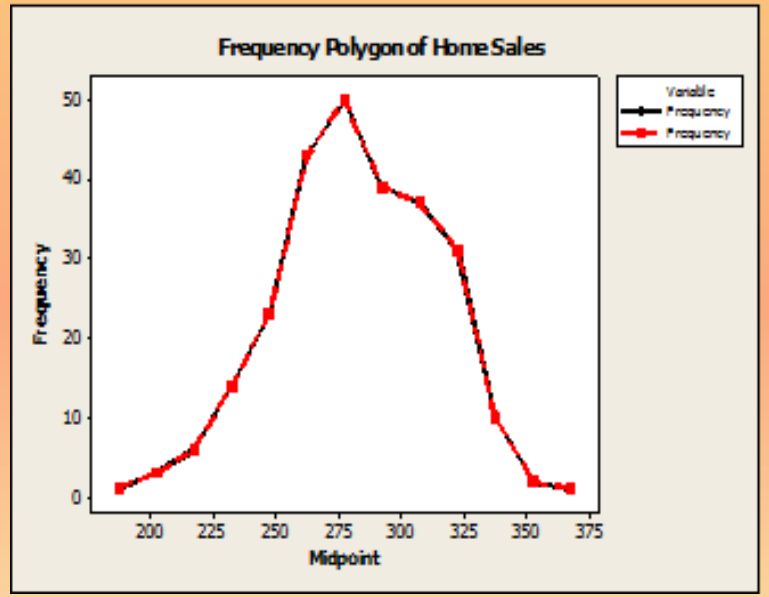
53%

Frequency Polygon

A frequency polygon is the plot of class frequencies and the midpoint – this plot is helpful in visualizing the pattern from the data (that is, if the data are symmetrical or skewed. See the example below

Midpoint for the Frequency Distribution and the Midpoints

Class-interval Selling Price($000)	Midpoint	Frequency No. of Houses
180 – 195	187.5	1
195 – 210	202.5	3
210 – 225	217.5	6
225 – 240	232.5	14
240 – 255	247.5	23
255 - 270	262.5	43
270 - 285	277.5	50
285 - 300	292.5	39
300 - 315	307.5	37
315 - 330	322.5	31
330 - 345	337.5	10
345 - 360	352.5	2
360 - 375	367.5	1
		$\sum f = 260$



Calculating Relative Frequency, Cumulative frequency, and Relative Cumulative Frequency

Table 2.9: Frequency Distribution with Seven Class Intervals

(1) Class Intervals ($)	(2) Frequency (f_i) No. of Customers	(3) Relative Frequency	(4) Cumulative Frequency	(5) Relative Cumulative Frequency
80-100	4	4/50=0.08	4	4/50=0.08
100-120	6	6/50=0.12	6+4= 10	10/50=0.20
120-140	9	9/50=0.18	9+6+4=19	19/50=0.38
140-160	13	13/50=0.26	13+9+6+4=32	32/50=0.64
160-180	9	9/50=0.18	9+13+9+6+4=41	41/50=0.82
180-200	6	6/50=0.12	6+9+13+9+6+4=47	47/50=0.94
200-220	3	3/50=0.06	3+6+9+13+9+6+4=50	50/50=1.00
	$\sum f_i = 50$	1.00		

$$\text{Relative Frequency of a Class} = \frac{\text{Frequency of that class}}{\text{Total Number of Observations}}$$

$$\text{Cumulative Relative Frequency of a Class} = \frac{\text{Cumulative Frequency of that class}}{\text{Total Number of Observations}}$$

Ogive: The Plot of Cumulative or Relative Cumulative Frequency

- An ogive is a very useful plot in estimating the data values between the class intervals when the data are grouped into a frequency distribution. It also can be used in estimating different percentiles from the data
- **Example:**

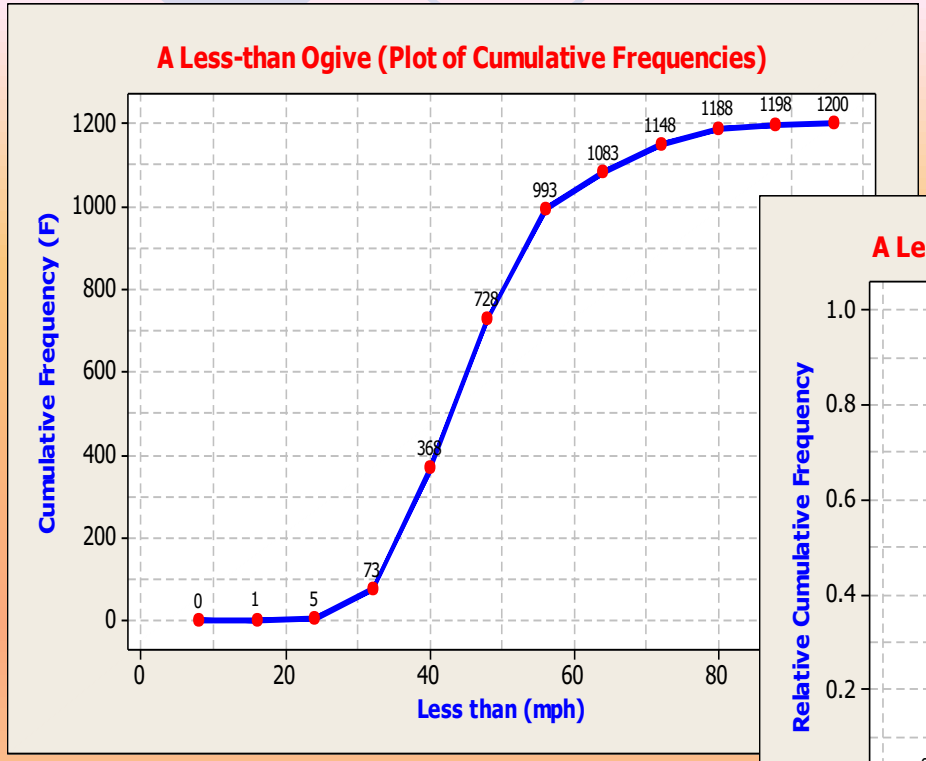
Table 2.23: Spot Speed for 1200 Cars

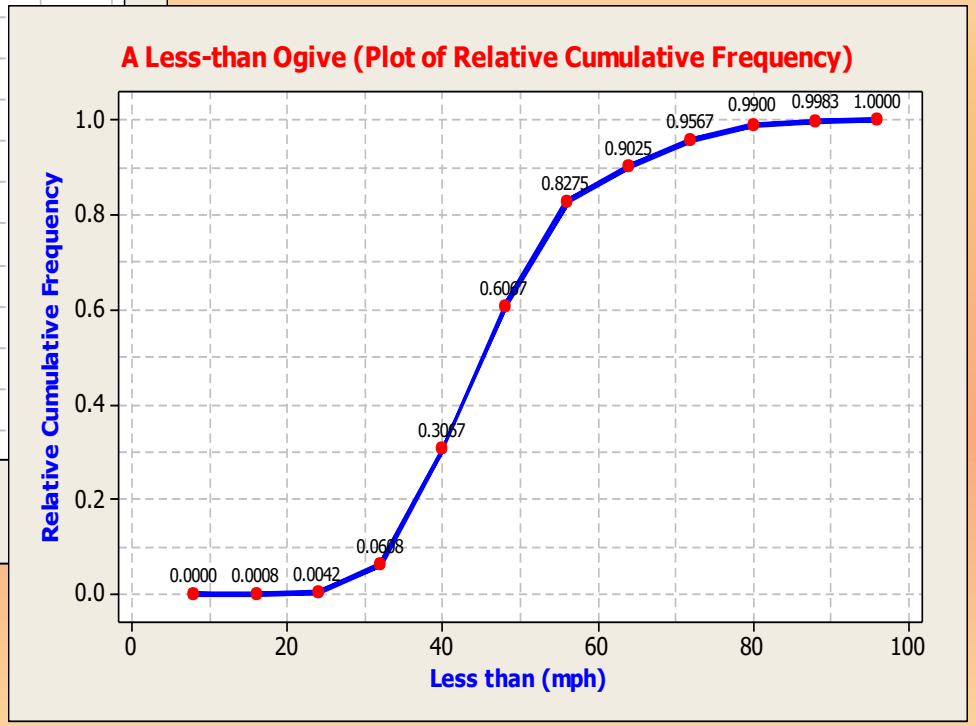
Class Interval (speed in mph)	Frequency (f) (No. of cars)
8 - 16	1
16 - 24	4
24 - 32	68
32 - 40	295
40 - 48	360
48 - 56	265
56 - 64	90
64 - 72	65
72 - 80	40
80 - 88	10
88 - 96	2
	$\sum f = 1200$

Table 2.24: Less than, Cumulative and Relative Cumulative Freq. of Data in Table 2.23

Less than (mph)	Cumulative Frequency (F)	Relative Cumulative Frequency
8	0	0.0000
16	1	0.0008
24	5	0.0042
32	73	0.0608
40	368	0.3067
48	728	0.6067
56	993	0.8275
64	1083	0.9025
72	1148	0.9567
80	1188	0.9900
88	1198	0.9983
96	1200	1.0000

Ogives

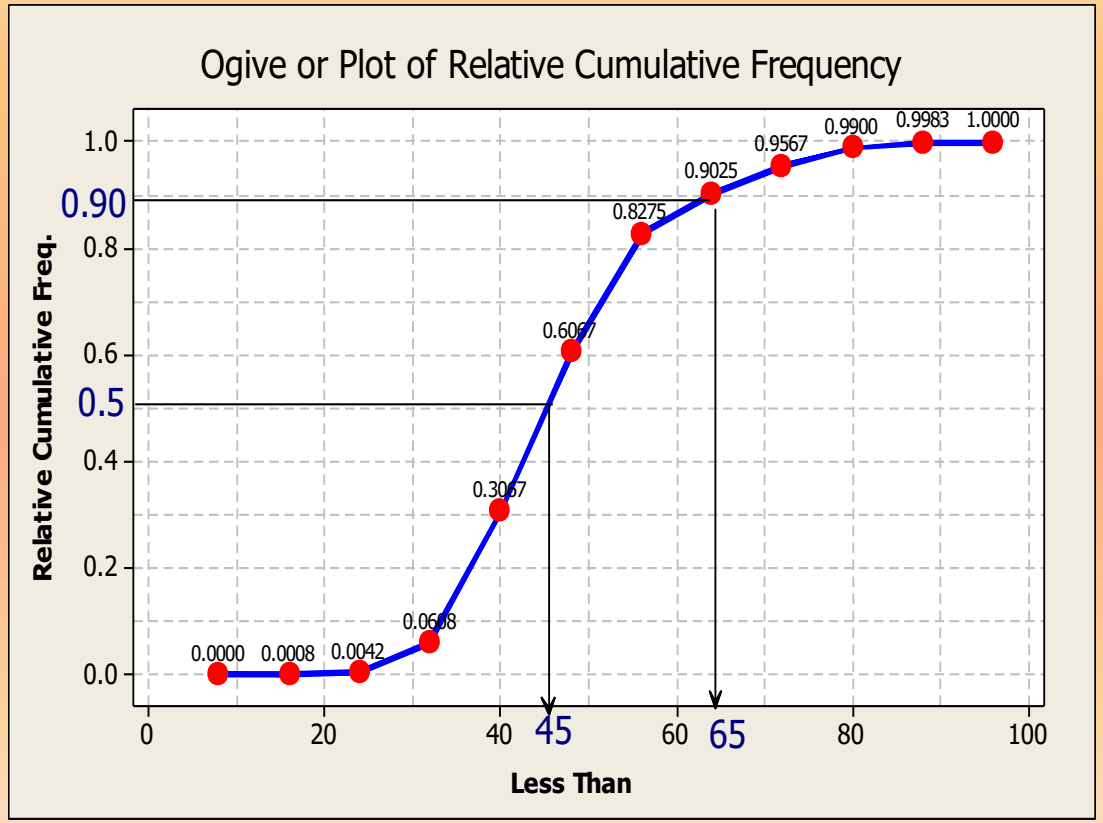

A Less-than Ogive (Plot of Cumulative Frequencies)
Cumulative Frequency (F)
Less than (mph)
0
200
400
600
800
1000
1200
0
20
40
60
80
0
1
5
73
368
728
993
1083
1148
1188
1198
1200


A Less-than Ogive (Plot of Relative Cumulative Frequency)
Relative Cumulative Frequency
Less than (mph)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
0.0000
0.0008
0.0042
0.0608
0.3067
0.6097
0.8275
0.9025
0.9567
0.9900
0.9983
1.0000

WHAT INFORMATION CAN BE OBTAINED FROM THE OGIVE?

- What is the median speed of the cars sampled? 45
- Determine the speed of 90% of the cars crossing the intersection. 65 mph

See the figure



Stem-and-Leaf Plot

- ***Stem-and-leaf*** plot is a very efficient way of displaying data, checking the variation and shape of the distribution.
- Stem-and-leaf plots are obtained by dividing each data value into two parts; stem and leaf.
- For example, if the data are two-digit numbers, e.g., 34, 56, 67, etc., then the first number (the tens digit) is considered the stem value, and the second number (the ones digit) is considered the leaf value. Thus, in data value 56, 5 is the stem and 6 is the leaf. In a three-digit data value, the first two digits are considered as the stem and the third digit as the leaf.
- A data set with values 80, 82, 85, 87, and 89 have the common stem of 8
- A stem-and-leaf can be constructed easily using sorted data

Stem-and-Leaf Plot

Table 2.27: Number of Defective Products

No. of Defects (out of 1000)												
42	14	55	54	30	58	76	44	28	20	60	68	26
93	35	81	82	38	27	83	52	82	71	54	62	73
68	64	66	66	75	63	56	34	88	39	49	79	76
54	46	44	86	58	45	10	41	34	65	53	54	37
57	46	30	55	40	46	38	25	12	105	59	95	103

Table 2.28: Sorted Data from Table 2.27

No. of Defects (Sorted— read row-wise)

10	**12**	**14**	**20**	**25**	**26**	**27**	**28**	**30**	**30**	**34**	**34**	**35**
37	**38**	**38**	**39**	**40**	**41**	**42**	**44**	**44**	**45**	**46**	**46**	**46**
49	**52**	**53**	**54**	**54**	**54**	**54**	**55**	**55**	**56**	**57**	**58**	**58**
59	**60**	**62**	**63**	**64**	**65**	**66**	**66**	**68**	**68**	**71**	**73**	**75**
76	**76**	**79**	**81**	**82**	**82**	**83**	**86**	**88**	**93**	**95**	**103**	**105**

Stem-and-leaf of No. of Defects (out of 1000)
N = 65
Leaf Unit = 1.0

1 2 3

3	1	024
8	2	05678
17	3	004457889
27	4	0124456669
(13)	5	2344445567889
25	6	023456688
16	7	135669
10	8	122368
4	9	35
2	10	35

Stem-and-Leaf Plot : Example

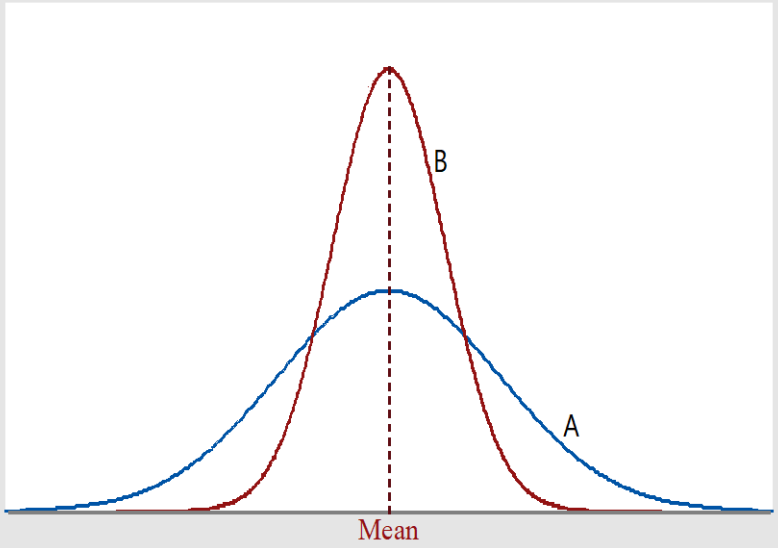
The stem-and leaf plot shows the number of orders received per day by a company.

1	9	2
2	10	3
5	11	245
7	12	78
8	13	2
11	14	137
15	15	1229
22	16	2266778
27	17	01599
(11)	18	00013346799
17	19	03346
12	20	4679
8	21	0177
4	22	45
2	23	18

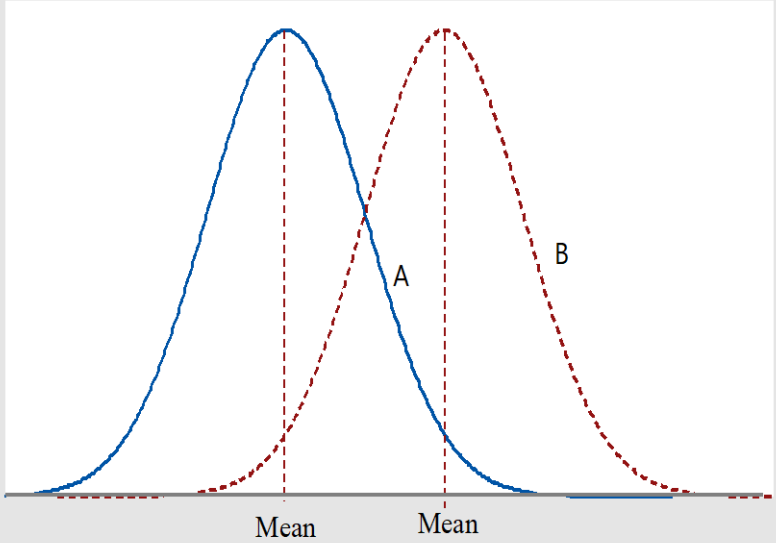
- [a] How many days were studied? **55**
- [b] How many observations are in the fourth class? ***2***
- [c] What are the smallest and largest orders? ***92, 238***
- [d] List the actual values in the sixth class? ***141, 143, 147***
- [e] How many days did the firm receive less than 140 orders? ***8***
- [f] How many days did the firm receive 200 or more orders? ***12***
- [g] How many days did the firm receive 180 orders? ***3***
- [h] What is the middle value? ***180***
- [i] What can you say about the shape of the data? ***Left or negatively skewed***

Graphical Display of Variation

Two Data Sets with Same Mean but different Standard Deviations



Two Data Sets with Different Means but Same Standard deviation



Boxplots

- The box-plot displays the smallest and the largest values in the data along with the three quartiles: Q1, Q2, and Q3.
- These five numbers (known as five measure summary) may be used to study the shape of the distribution and draw conclusion from the data.

Waiting Time(min.)

6.8	9.9	11.0	11.8	12.6	14.0	16.0	8.0	10.1	11.1	11.8	12.6	14.0	16.6	8.2
10.2	11.3	11.9	12.6	14.0	8.8	10.4	11.4	12.0	12.7	14.2	9.0	10.5	11.5	12.(
13.0	14.3	9.1	10.7	11.6	12.1	13.1	14.4	9.3	10.8	11.7	12.2	13.1	14.5	9
10.8	11.7	12.5	13.3	14.5										

Descriptive Statistics of Waiting Time

Variable	N	N*	Mean	SE Mean	StDev	**Minimum**	**Q1**	**Median**
Waiting Time	50	0	11.784	0.289	2.045	6.800	10.475	11.800

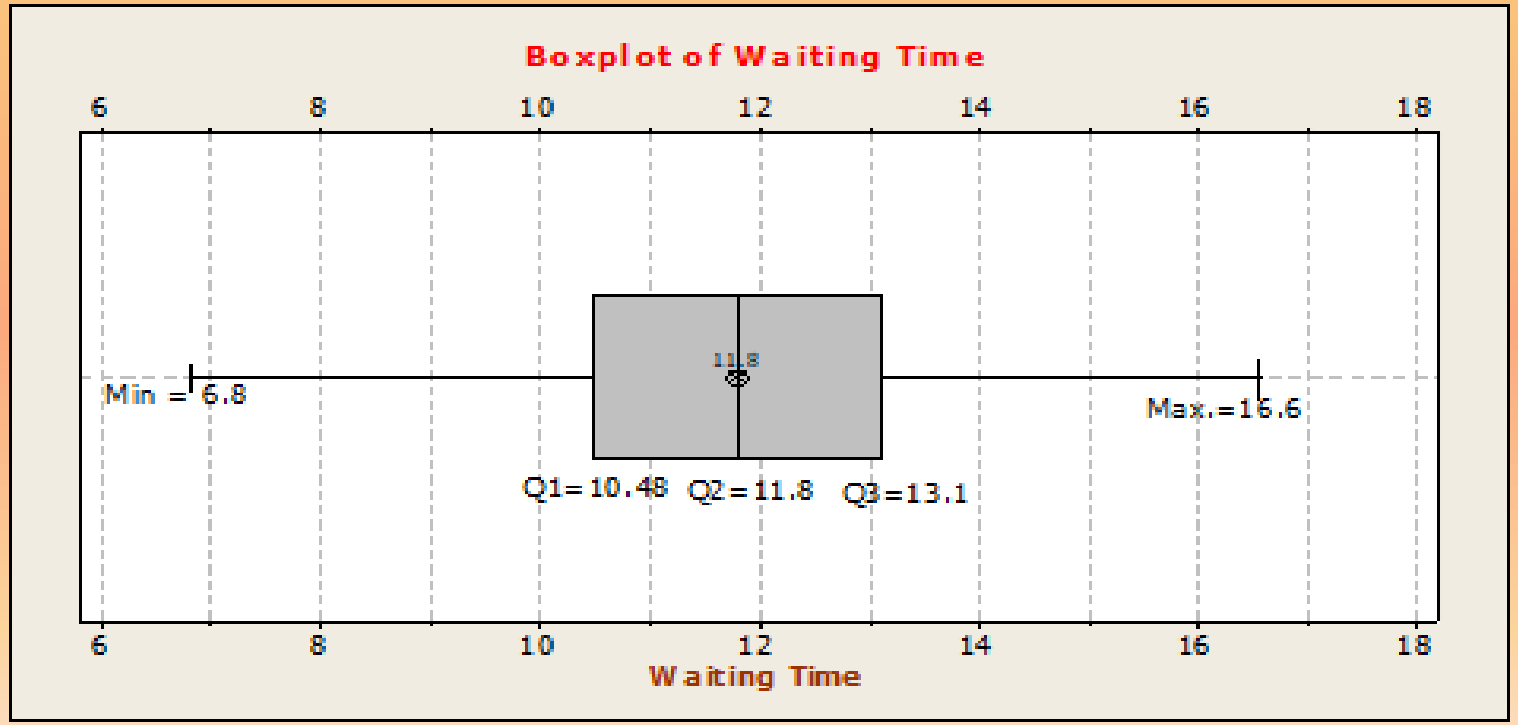
Maximum
16.600

Box Plot …Cont.

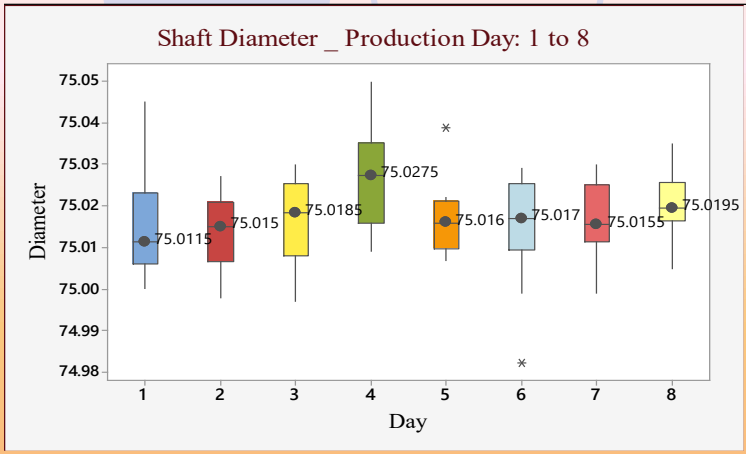
Descriptive Statistics of Waiting Time

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
Waiting Time	50	0	11.784	0.289	2.045	6.800	10.475	11.800

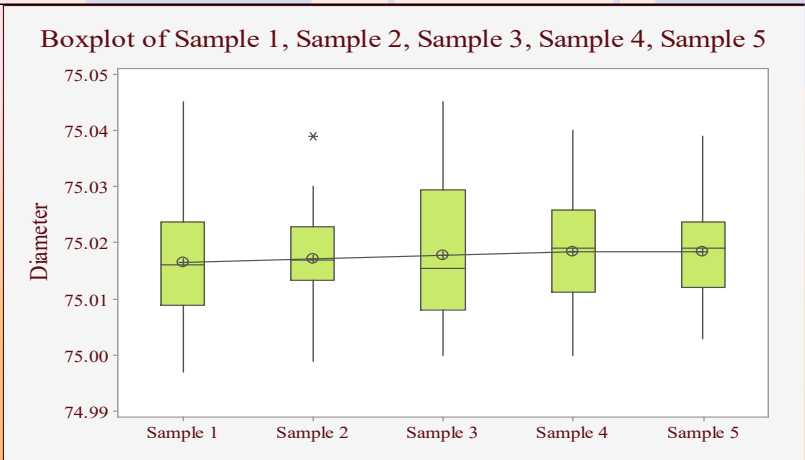
Maximum
16.600



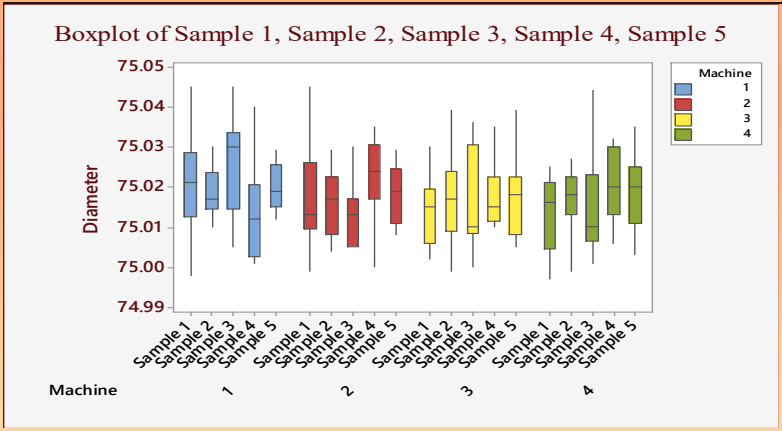
Variation of Box Plot



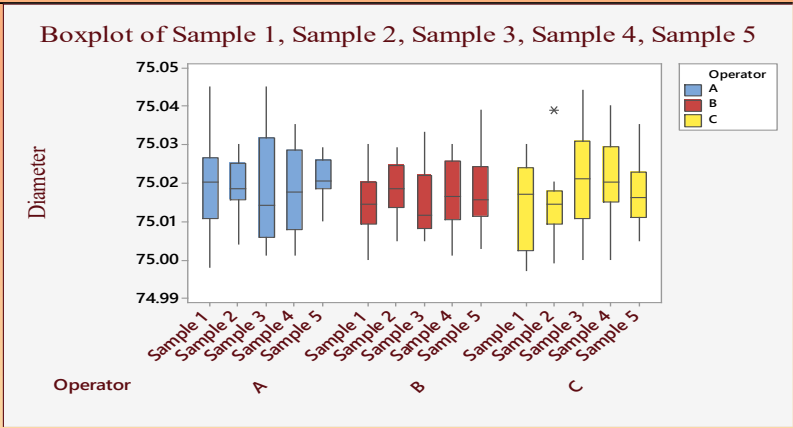
Box Plots of Shaft Diameter Over a Period of 8 Days



Box-plots for 5 Samples of Same Product



Box plots of Samples vs. Machines



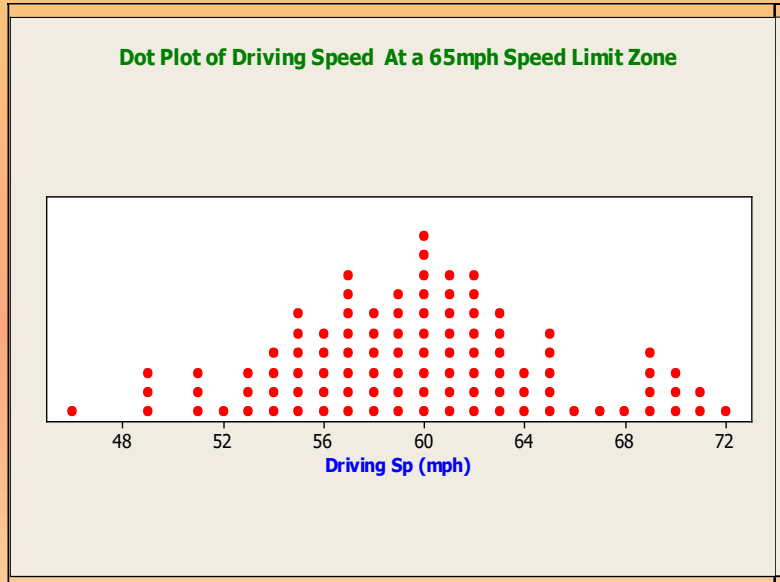
Box plots of Samples vs. Operators

Dot Plot

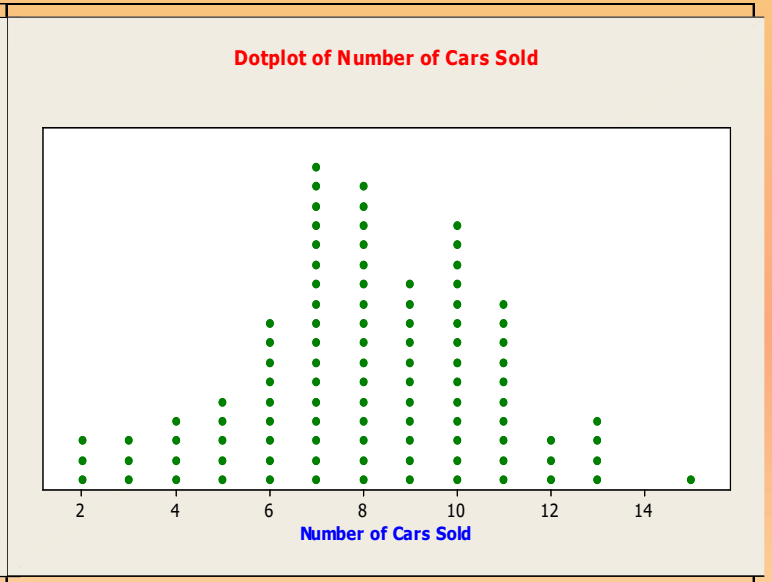
- A ***dot plot*** may be used to study the shape of the distribution or to compare two or more than two sets of data.
- In a dot plot, the horizontal axis shows the range of values in the data. Each observation is represented by a dot placed above the axis.
- If the data value repeats, the dots are piled up at that location, showing a dot for each repeated value.
- A dot in this plot may also represent multiple observations.
- These plots are particularly useful when the data sets are small

Dot Plot

Plot below shows the spot speed of 100 cars at a 65 mph speed limit zone. The plot shows that most of the cars were at or below the speed limit. There were 13 cars over the speed limit of 65 mph. The shape of the data is approximately symmetrical.



Dot plot of 100 Cars at 65 mph Speed Zone



Dot plot of Number of Cars Sold

Describing Categorical Data

Tallies

A tally is a count of numbers in each category. It provides a table like a frequency distribution displaying the counts and percentages for each category.

A product was rated by 200 customers using a scale of 1 to 10 where 1 is unacceptable, 10 is outstanding, and 5 is average

Product Rating														
5	5	10	5	5	2	4	7	3	1	10	8	8	2	1
5	8	4	5	1	8	3	5	2	2	3	8	2	5	2
3	3	4	8	6	3	2	7	6	6	5	3	4	6	6
6	2	5	2	4	1	2	6	1	9	5	5	6	1	7
5	6	4	7	6	6	3	7	3	6	10	4	5	1	5
6	4	2	5	6	3	4	6	4	6	4	4	6	7	6
1	3	4	3	4	4	7	3	7	4	10	3	1	7	5
6	8	4	4	2	5	5	1	3	5	4	4	9	5	7
6	3	2	7	2	7	1	5	4	5	10	6	7	10	1
2	5	5	2	6	2	1	6	9	8	5	3	8	2	1
10	3	2	10	5	5	1	6	5	2	2	2	2	6	6
6	9	1	8	2	7	2	3	2	2	5	4	6	6	1
10	2	6	3	3	2	10	5	5	10	4	2	10	5	4
1	7	8	2	6										

Tallies

Tally of Product Rating

Tally for Discrete Variables: Product Rating

Product Rating	Count	Percent
1	18	9.00
2	31	15.50
3	21	10.50
4	24	12.00
5	33	16.50
6	31	15.50
7	15	7.50
8	11	5.50
9	4	2.00
10	12	6.00
N=	200	

Describing, Summarizing and Graphing Categorical Variables

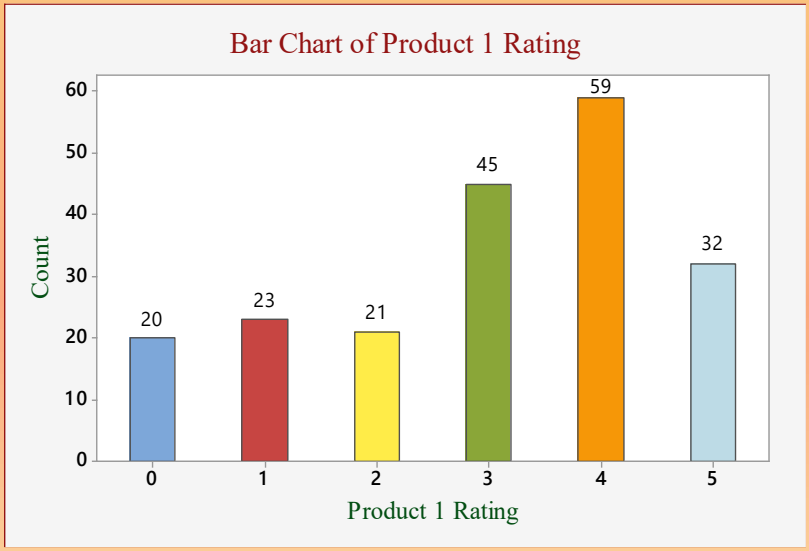
- Categorical data are the data arranged in classes or categories.
- Categorical data may result from a counting process
- Examples of categorical data are the number of orders received by a company per day, the number of companies belonging to a certain industrial classification,etc.

Creating a Bar Chart from a Simple Tally

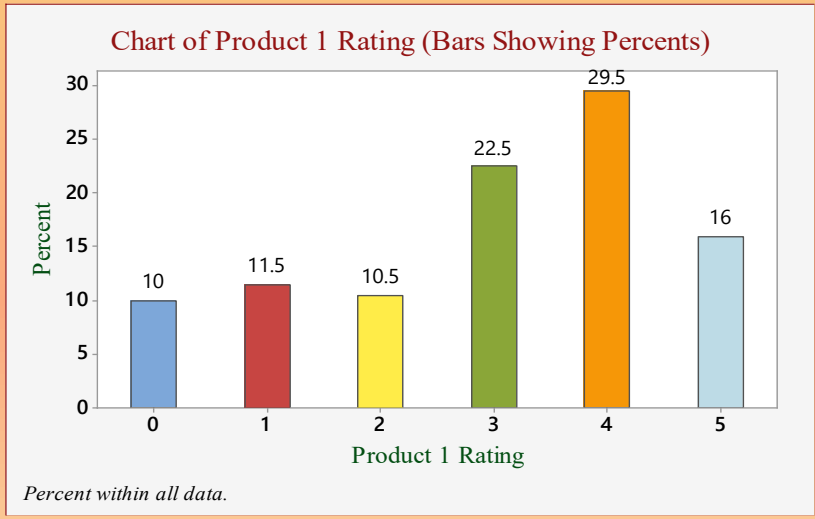
A tally is a count or percentage of number of cases in a category. Table in the next slide contains the ratings for Product 1 for a sample of 200 product users. The variable Product Rating is a categorical variable with a scale ranging from 0 to 5 (0= Unacceptable, 1=Fair, 2=Poor, 3=Satisfactory, 4=Good, 5=Excellent).

Creating a Bar Chart from a Simple Tally

Product	Rating																				
0	1	3	3	4	5	1	4	3	3	4	5	1	0	3	4	5	3	4	3	5	0	1
3	4	5	4	3	2	1	4	3	0	0	0	3	3	1	1	1	1	4	4	4	5	4
5	5	3	2	3	3	4	4	4	4	4	5	3	2	4	5	3	1	4	5	5	0	0
0	0	3	3	4	2	1	5	2	2	4	4	1	1	4	4	2	2	4	4	5	5	5
3	3	1	1	0	0	4	5	4	4	5	5	5	3	3	1	5	3	4	4	3	3	4
2	3	5	4	0	0	0	3	4	3	2	4	4	4	4	4	5	3	3	0	4	4	3
3	5	4	4	5	3	3	2	2	5	4	3	2	1	1	2	3	4	5	4	3	2	1
0	5	4	2	3	1	0	0	4	4	4	4	5	4	3	2	1	5	4	3	2	2	1
0	3	3	4	5	4	2	4	4	5	3	4	4	3	2	1							



Bar Chart of Product 1 Rating



Bar Chart of Product 1 Rating (Bars showing Percent)

Cross Tabulation

Cross tabulation is used to summarize the data for two variables. Suppose that a manufacturing firm is in the process of implementing a Lean Six Sigma Quality program. The company hired a consulting firm to provide training to 100 executives at different levels. At the conclusion of the training, the management asked the executives at different levels to rate the training program as effective, somewhat effective, useless, or very effective. The response can be shown as a cross tabulation.

Two-way Table of Product Rating

Tabulated Statistics: Executive Level, Rating

Rows: Executive Level Columns: Rating

	Effective	Somewhat Effective	Useless	Very Effective	All
Manager	11	3	11	13	38
Senior Manager	7	4	8	7	26
Supervisor	7	9	4	16	36
All	25	16	23	36	100

Cell Contents: Count

Cross Tabulation with Two and Three Categorical Variables

The data for variables: **Gender** (male, female);
Degree major (1=computer science, 2=engineering, 3=social science, 4=business, 5=other); and
Employment status (employed, self-employed) are summarized in Table.
Using cross tabulation, we construct bar charts to show the employment status and degree major for the male and female respondents.

Cross-table_ Employment Status, Degree Major and Gender

Tabulated statistics: Employment Status, Major, Gender

Results for Gender = Female

Rows: Employment Status Columns: Major

	1	2	3	4	5	All
Employed	5	7	26	26	16	80
Self-employed	1	4	8	6	3	22
All	6	11	34	32	19	102

Cell Contents: Count

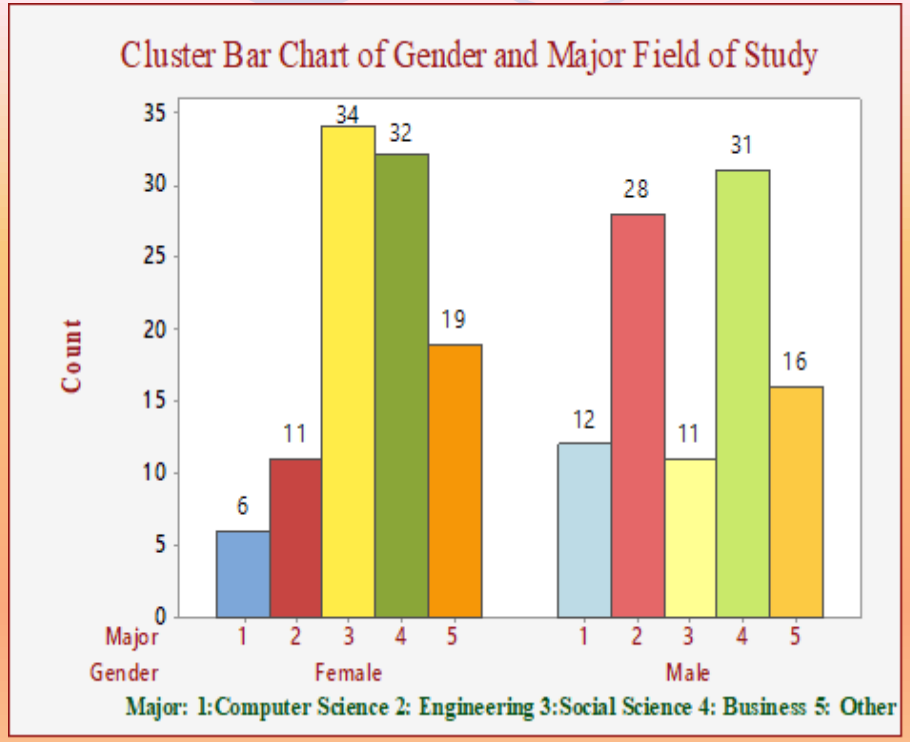
Results for Gender = Male

Rows: Employment Status Columns: Major

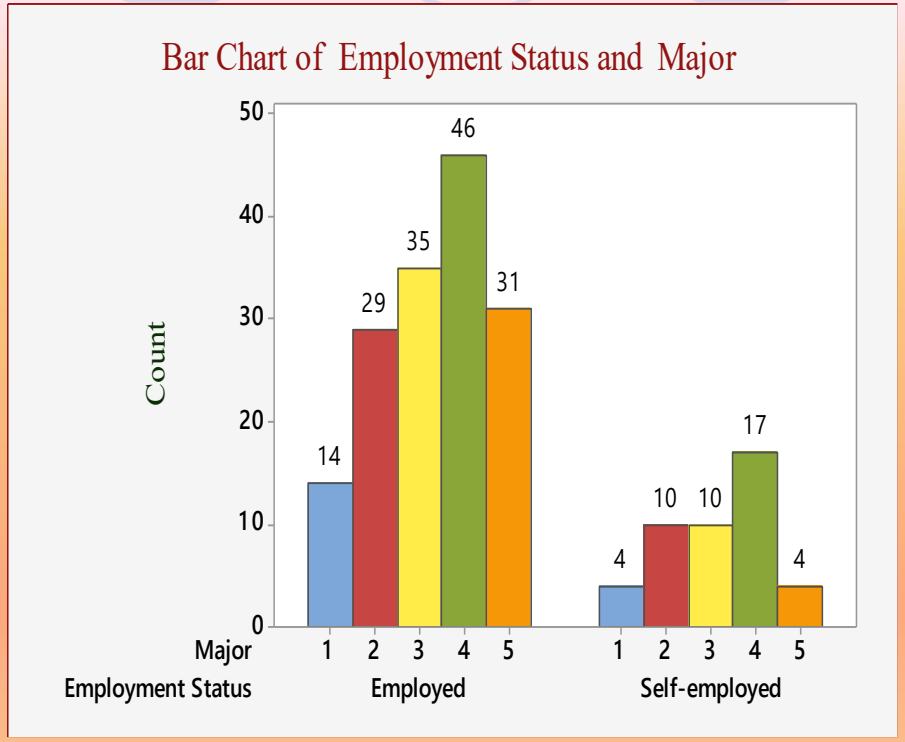
	1	2	3	4	5	All
Employed	9	22	9	20	15	75
Self-employed	3	6	2	11	1	23
All	12	28	11	31	16	98

Cell Contents: Count

A Bar Chart of Gender and Major

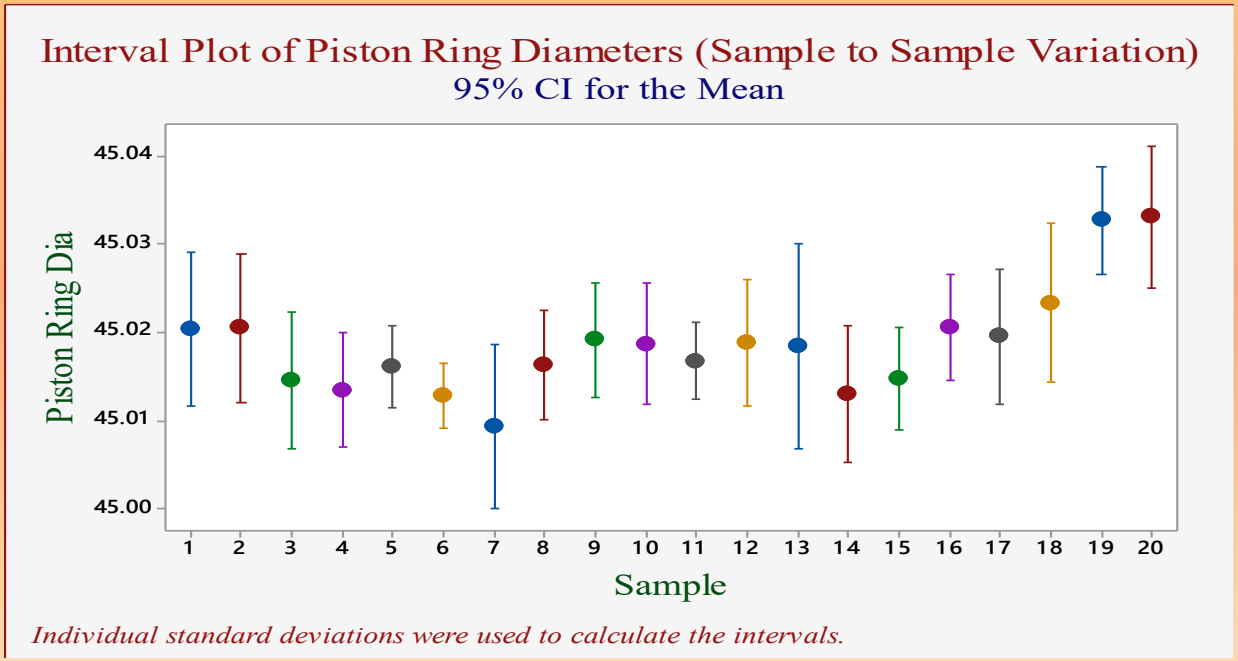


A Bar Chart of Employment Status and Major



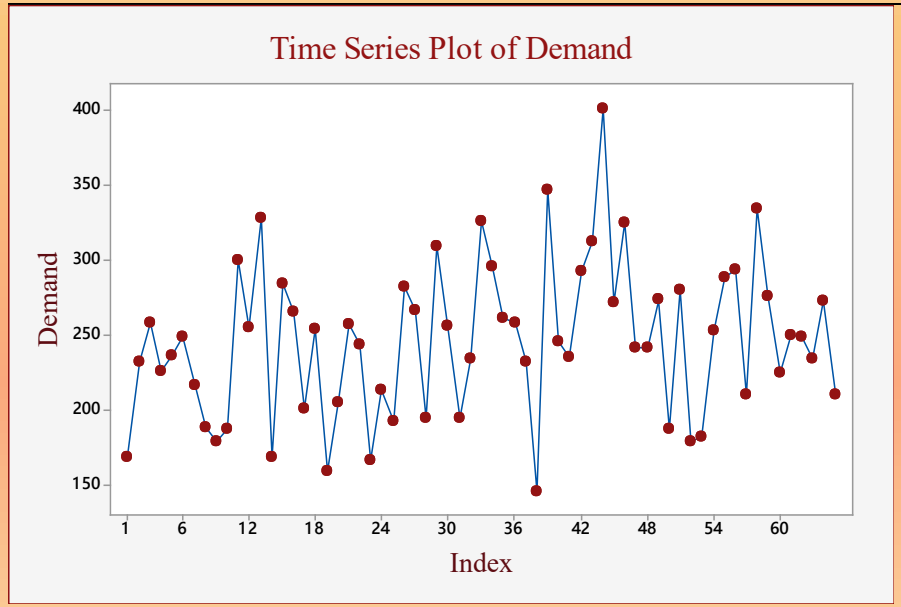
Interval Plot showing the Variation in Sample Data

- Interval plot is also useful in visualizing the variation in samples.
- The data plotted shows 20 samples each of size 10 of finished inside diameter of piston ring (in mm).
- Investigate sample to sample variation and the mean for each sample by constructing an interval plot

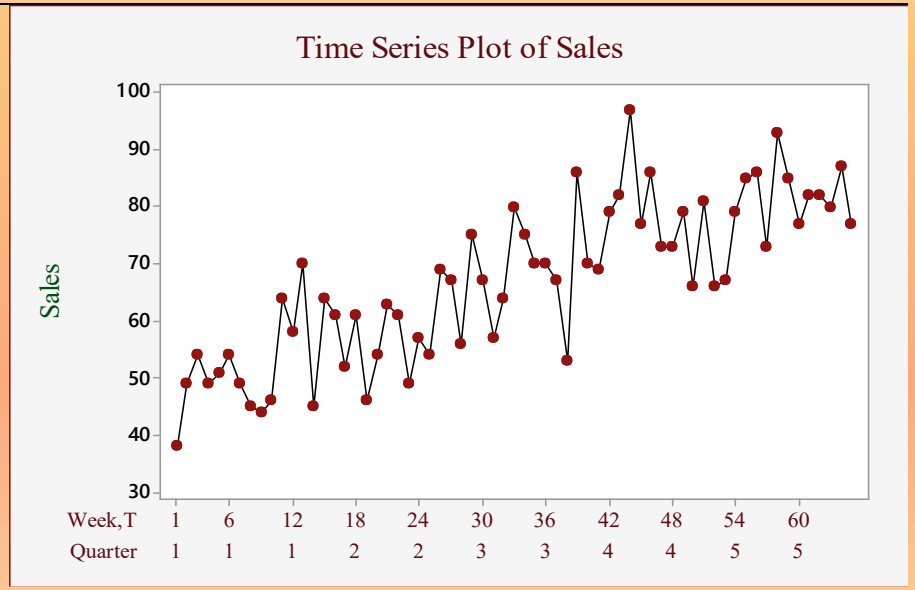


Time Series Plots

A time series plots the data over time.
The plot is helpful in visualizing a trend or pattern in a data set.

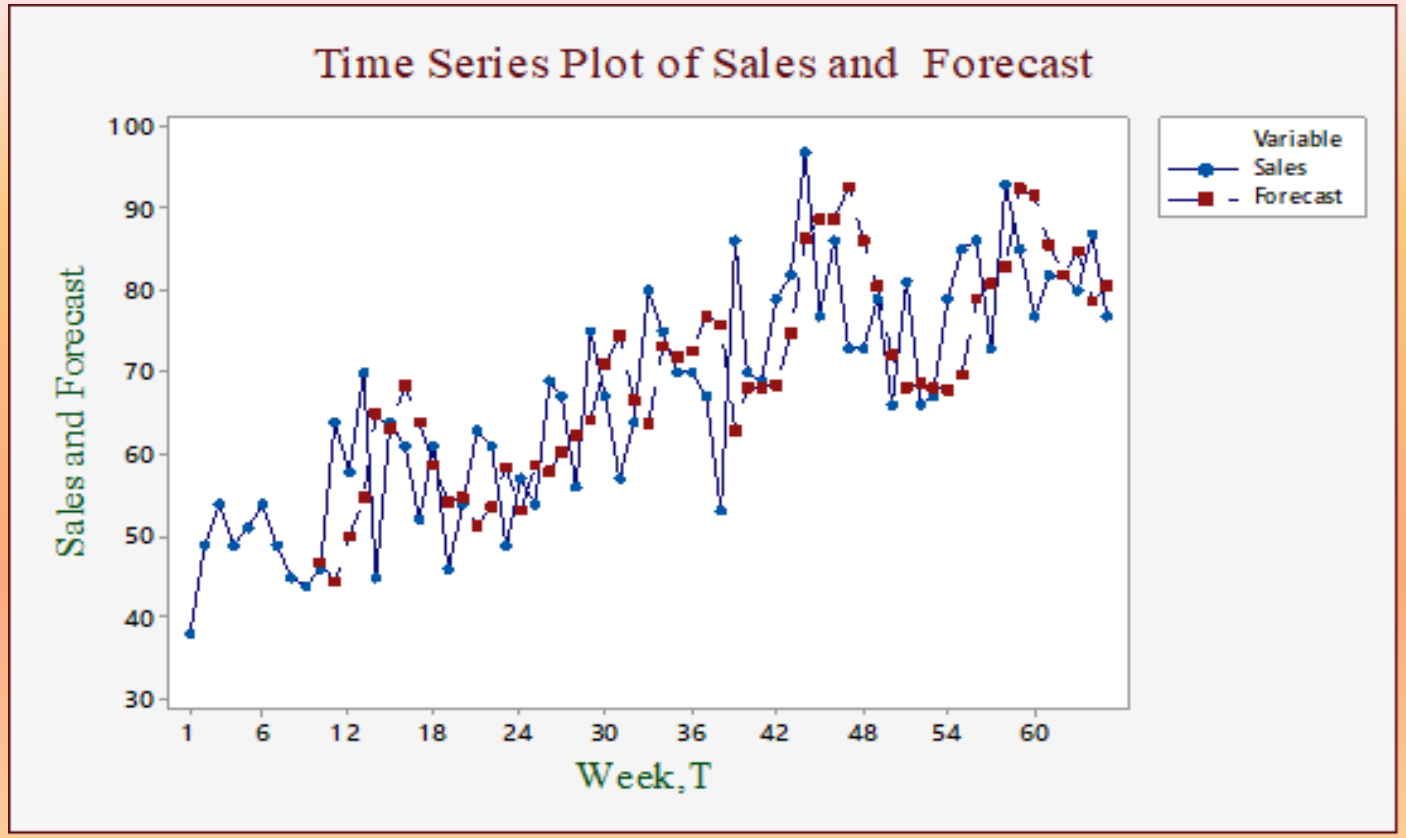


Simple Time Series Plot of Weekly Demand

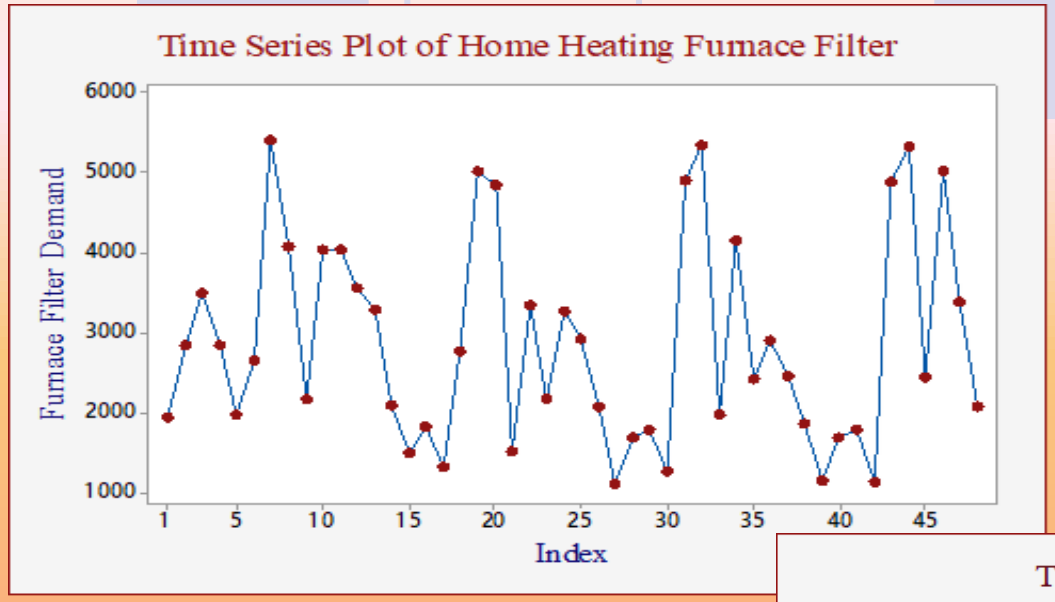


A Simple Time Series Plot of Sales Data

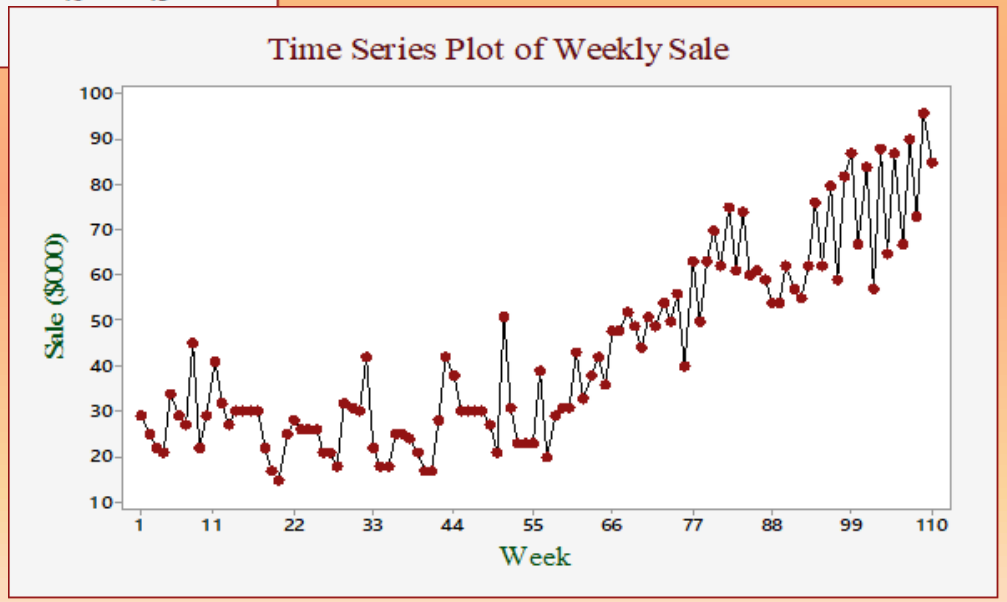
A Time Series Plot Showing Sales and Forecast



A Time Series Plot Showing a Seasonal Pattern



A Time Series Plot Showing a Trend



Sequence Plot: Plot of Process Data

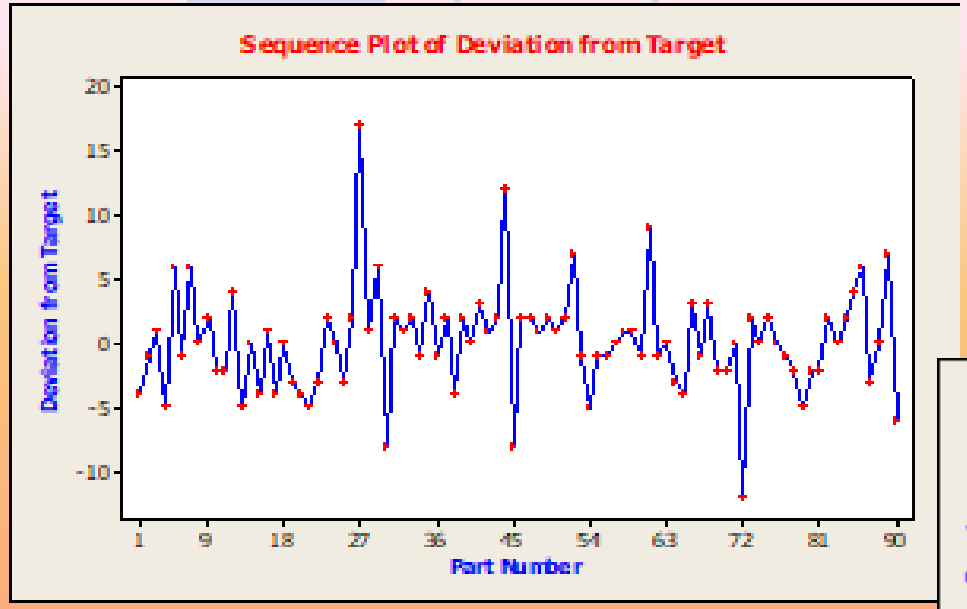
- Sequence plot shows the evolution of a measured characteristic over time.
- Similar to a time-series plot with the time plotted on the horizontal axis and the corresponding process characteristic on the vertical axis.
- Shows the behavior of the process over time.
- The variation or the trend in the process can be seen easily from this plot.
- The plot shows the deviation of a process from a specified target value.

The data in Table list the deviation (in 0.00025-inch units) of the diameter of 90 machined shafts from the target value. In these data, 0 means that the measured diameter was right on target, 2 means that the measured diameter was 0.0005 inch above the target value; whereas, a 3 means that the measured diameter was 0.00075 above the target value. We construct a sequence plot of the data and interpret the results.

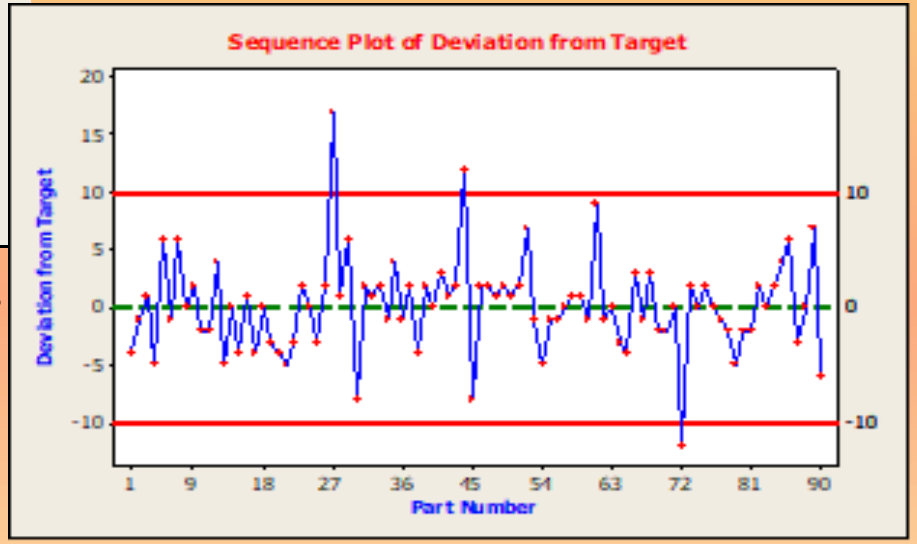
Data

Diameter Deviation From Target [Coded in 0.00025 inch deviation from targ										
-4	-1	1	-5	6	-1	6	0	2	-2	-2	4
0	-4	1	-4	0	-3	-4	-5	-3	2	0	-3
17	1	6	-8	2	1	2	-1	4	-1	2	-4
0	3	1	2	12	-8	2	2	1	2	1	2
-1	-5	-1	-1	0	1	1	-1	9	-1	0	-3
3	-1	3	-2	-2	0	-12	2	0	2	0	-1
-5	-2	-2	2	0	2	4	6	-3	0	7	-6

Variations of the sequence plot



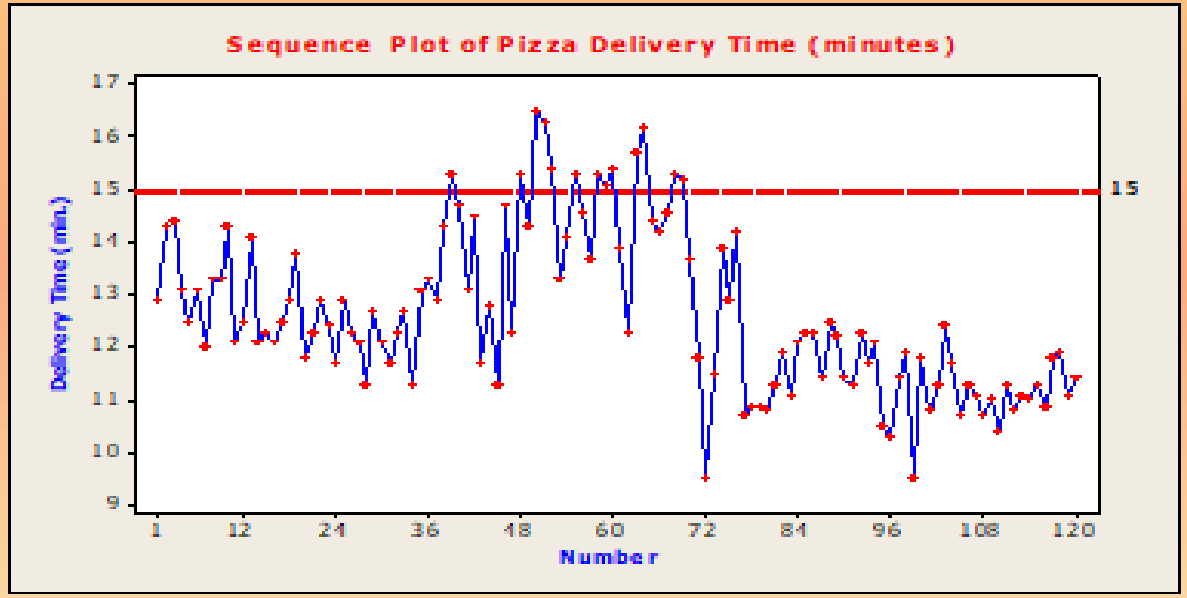
Sequence Plot of the Measurements on Machined Parts



Sequence Plot with Specification Limits

Sequence Plot of Pizza Delivery Time

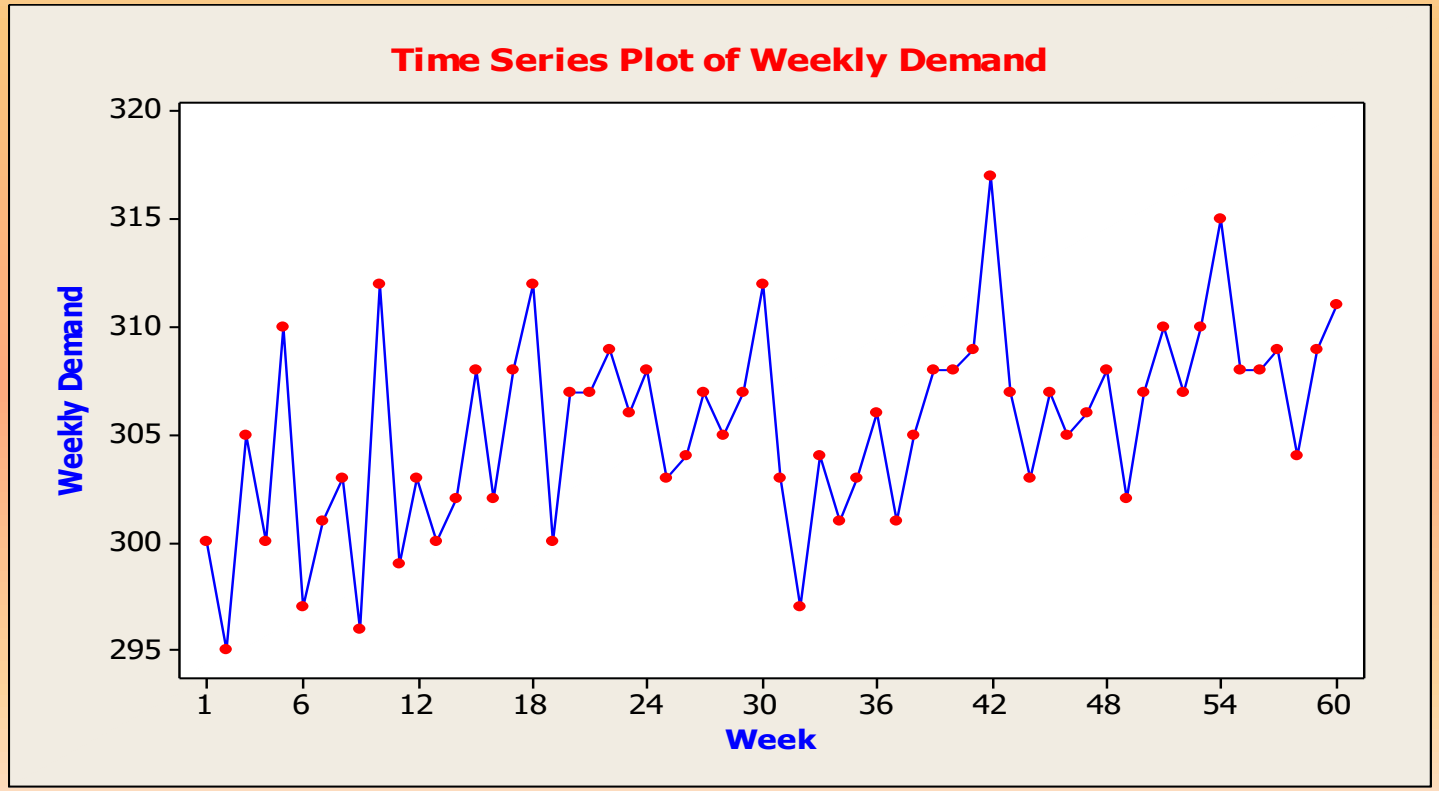
A pizza chain is going to launch a campaign with a target delivery time 0f 15 min. or less for their customer orders. Before they launch the campaign, the pizza chain would like to study the current delivery process. If the current process indicates large variations in the delivery time, the causes of variation will be studied, and corrective actions be taken to meet the target delivery time of 15 minutes or less. The data for the delivery time (in minutes) of 120 deliveries by different carriers were collected. A sequence plot of the delivery time data is shown.



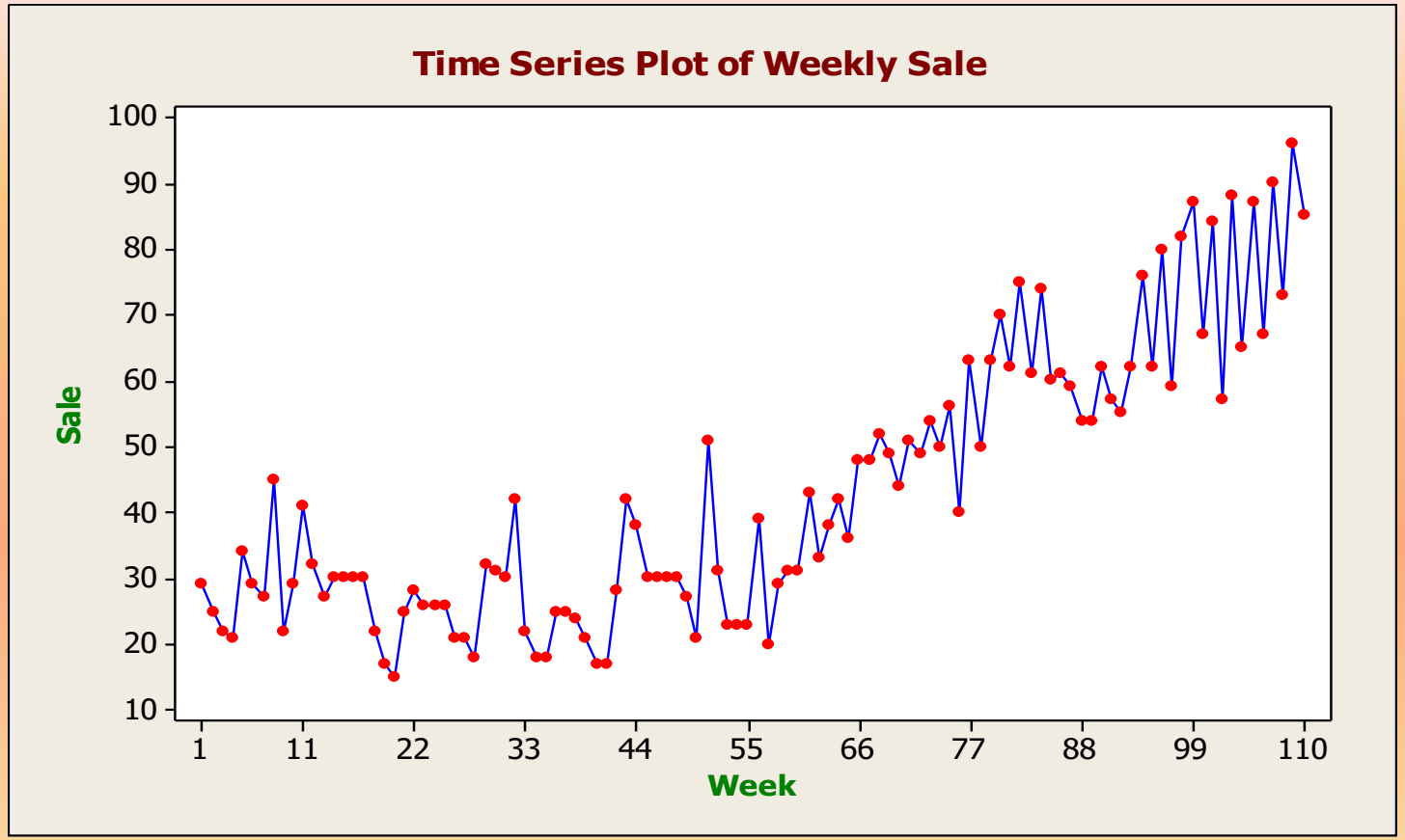
A line is drawn at 15 minutes to show the target value.

Time Series Plots

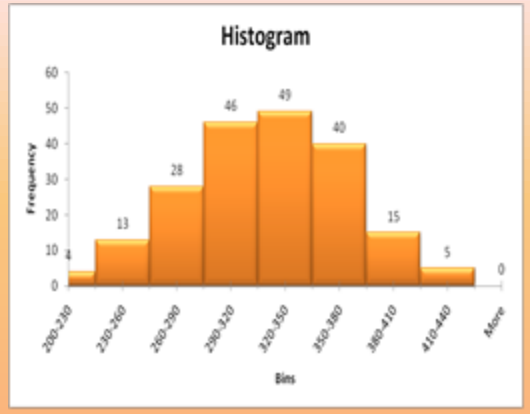
A time series plots the data over time. The graph plots the pairs of points and connects these points using a straight line. In a time series plot, the x values are time. The plot is helpful in visualizing a trend or a pattern in a data set over time.

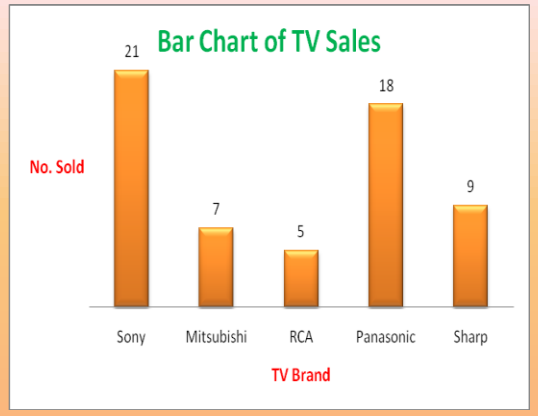


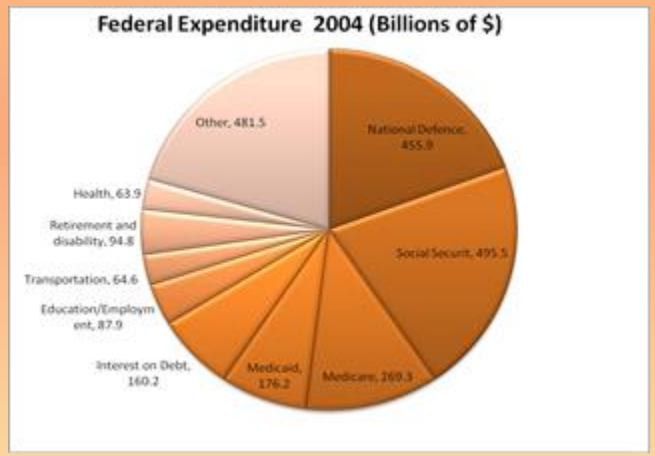
A Time Series Plot Showing Trend

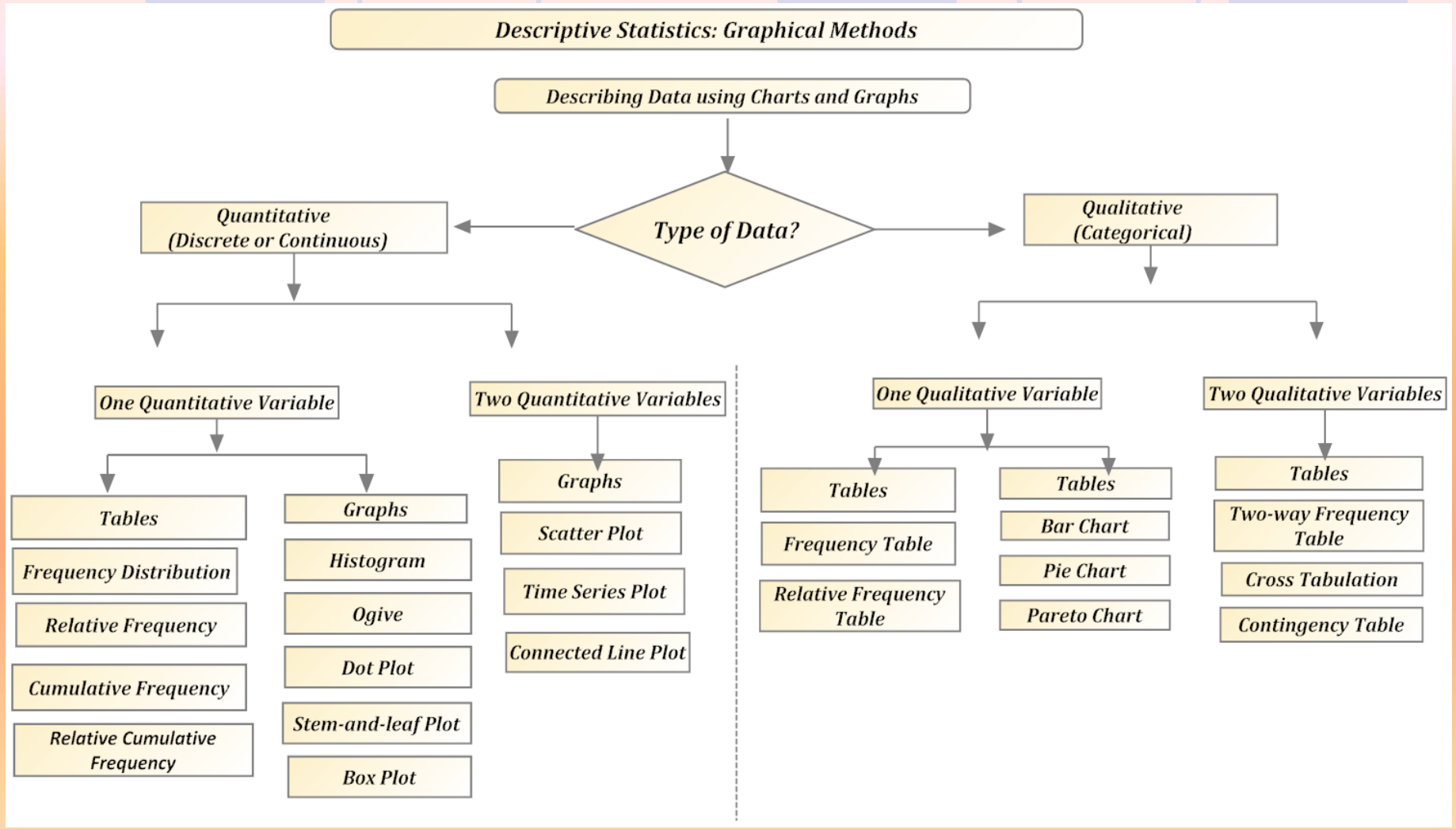


Widely used Charts and Graphs: Overview





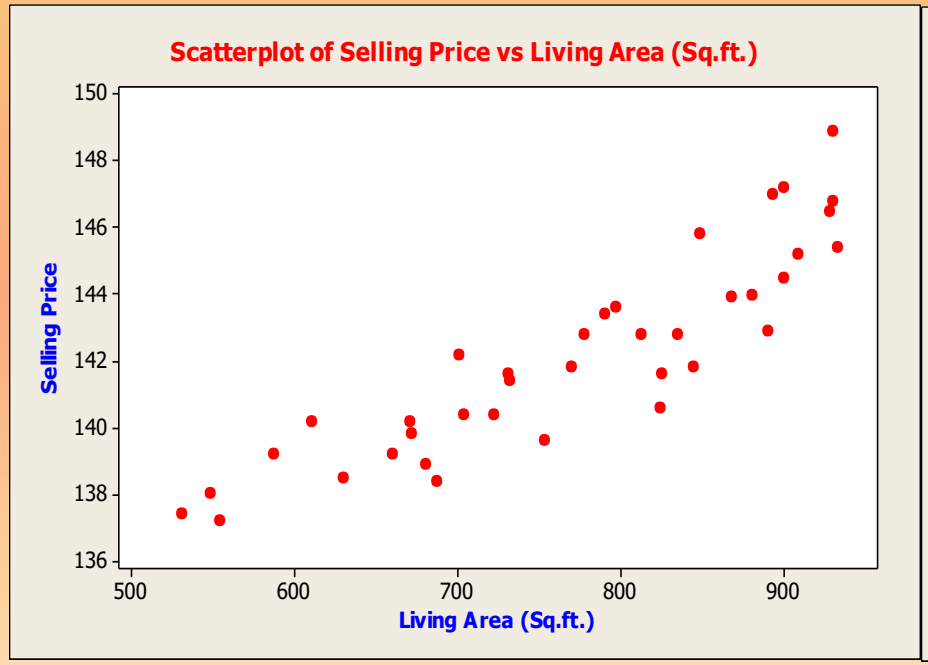


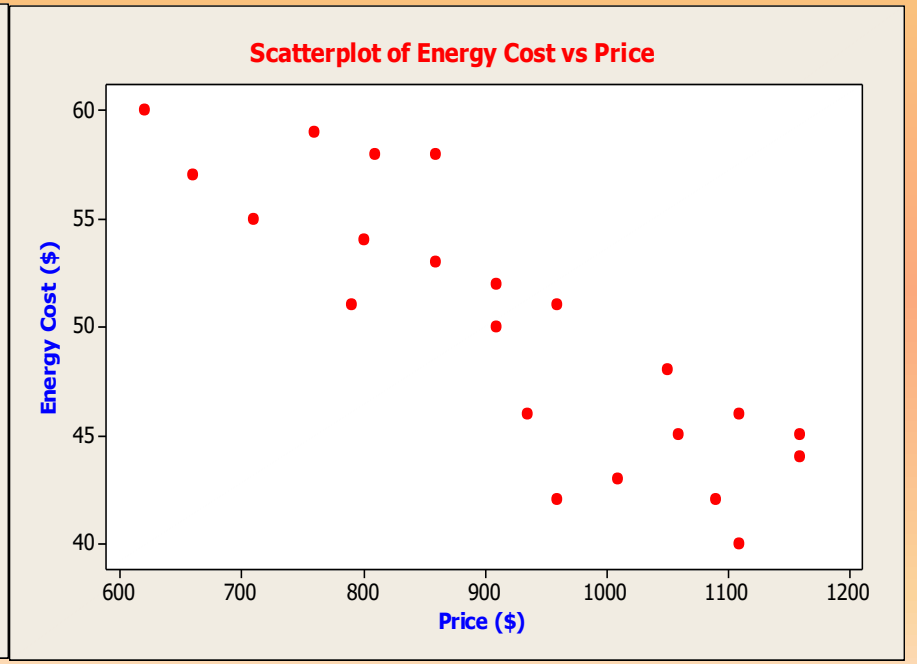

Descriptive Statistics: Graphical Methods
Describing Data using Charts and Graphs
Type of Data?
Quantitative (Discrete or Continuous)
Qualitative (Categorical)
One Quantitative Variable
Two Quantitative Variables
One Qualitative Variable
Two Qualitative Variables
Tables
Frequency Distribution
Relative Frequency
Cumulative Frequency
Relative Cumulative Frequency
Graphs
Histogram
Ogive
Dot Plot
Stem-and-leaf Plot
Box Plot
Graphs
Scatter Plot
Time Series Plot
Connected Line Plot
Tables
Frequency Table
Relative Frequency Table
Tables
Bar Chart
Pie Chart
Pareto Chart
Tables
Two-way Frequency Table
Cross Tabulation
Contingency Table

Section 4 :
Describing Two or More Variables Visually

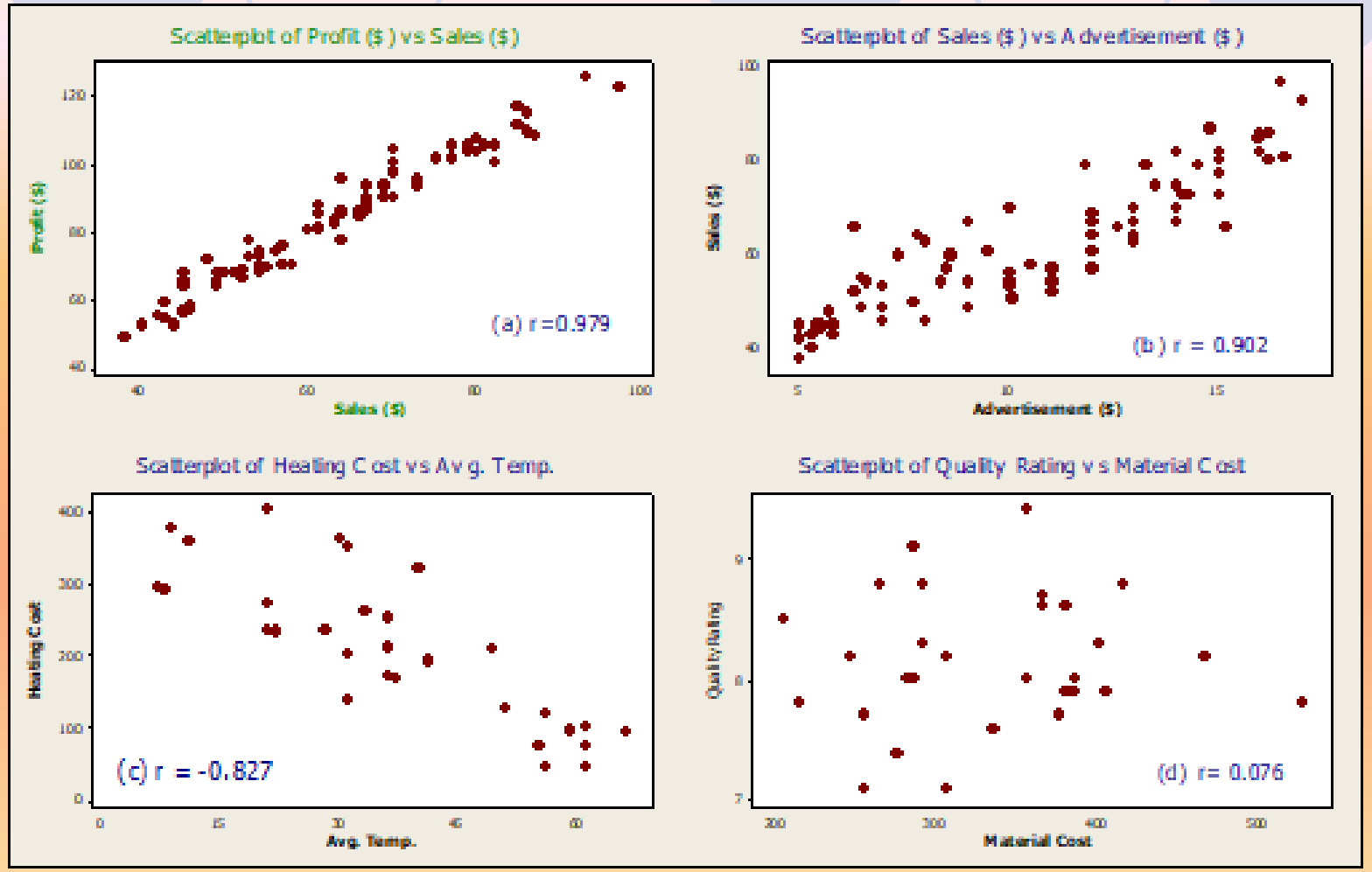
Scatter Plots

Scatter plots are helpful in investigating the relationship between two variables. One of these variables is considered a dependent variable and the other an independent variable. The data value is thought of as having a x value and a y value. Thus, we have (x_i, y_i), i=1,2,3,.....,n pairs. The relationships between the variables can be seen from scatter plots.



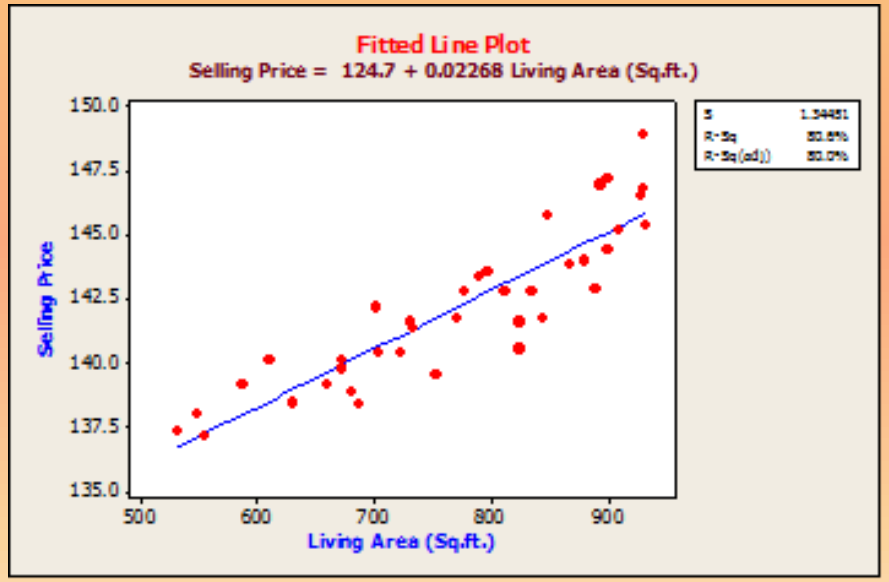


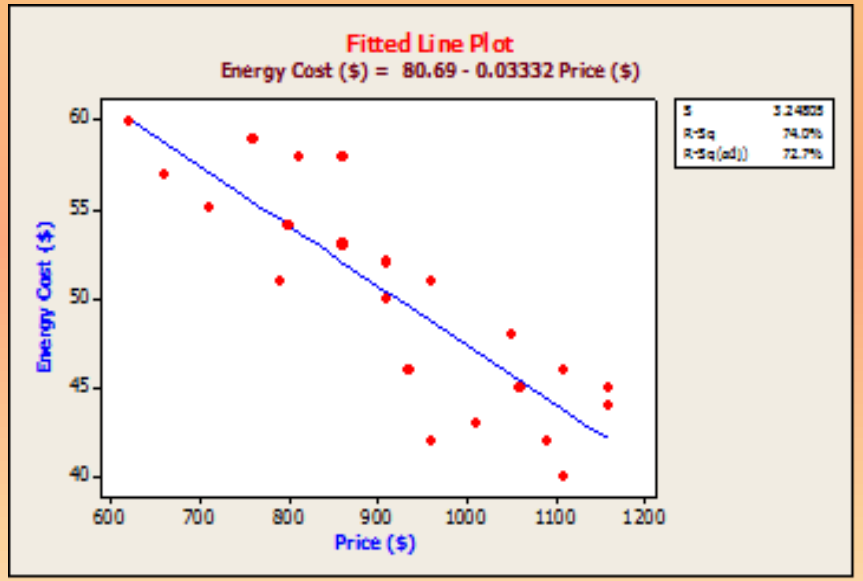
Scatter Plots_ Examples



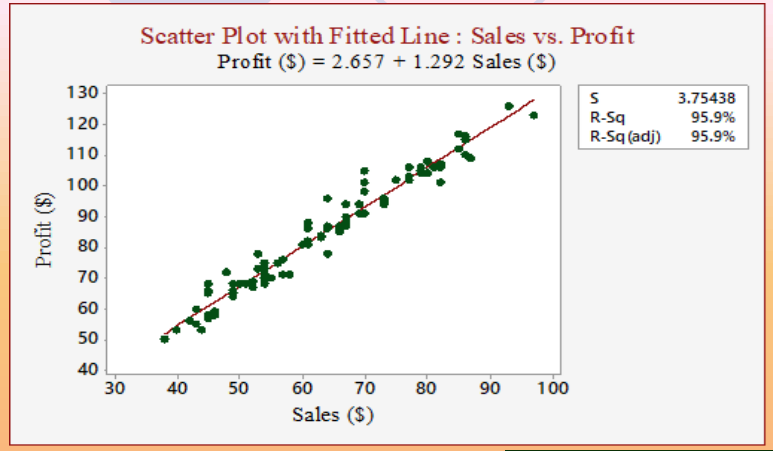
Fitted Line Plots

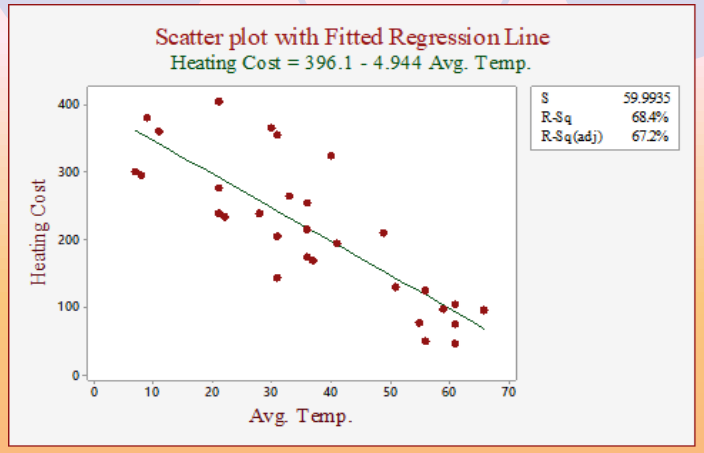
The scatter plot shows a linear or non-linear relationship or no relationship between the two variables. If the relationship between the two variables is linear a best fitting line over the scatter plot can be drawn using a computer package. Some packages also provide the equation of the best fitting line that can be used to predict one variable using the other.

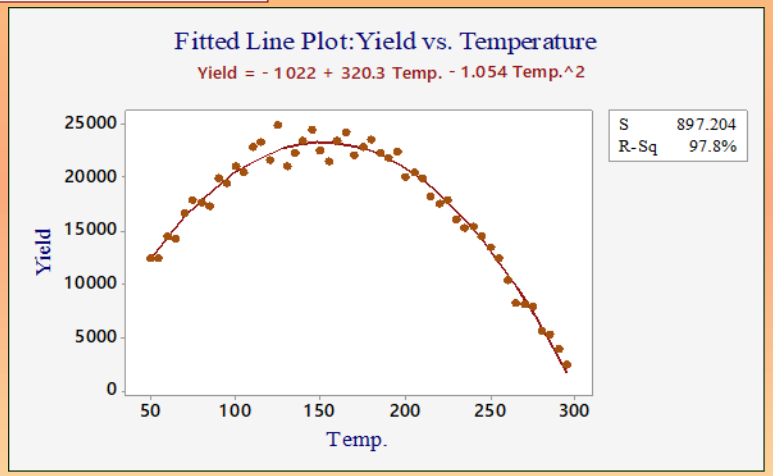




Scatterplot with Regression







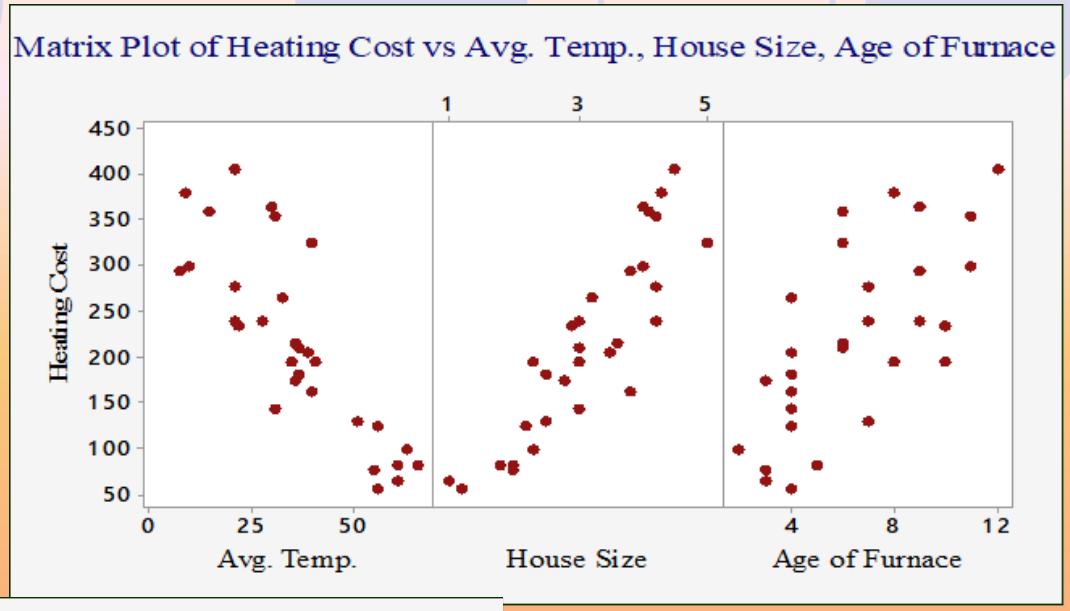
Scatterplot with the Best-fit it Line or Curve

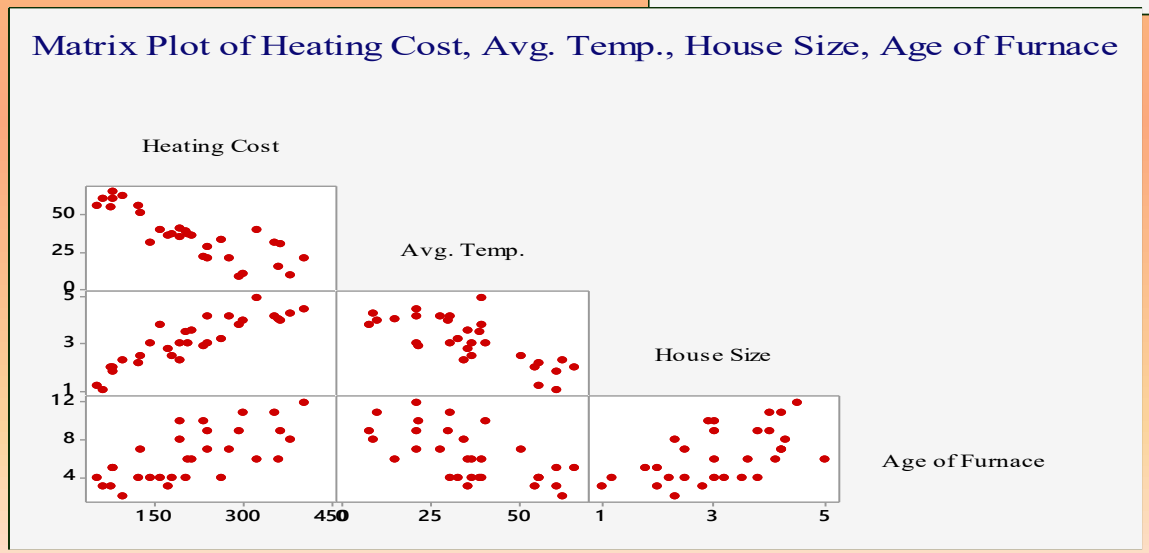
Matrix Plots

- A matrix plot is a useful graphical tool to investigate the relationships between pairs of variables by creating an array of scatterplots.
- In regression analysis and modeling, often the relationship between multiple variables is of interest.
- Matrix plots can be created to visually investigate the relationship between the response variable and each of the independent variables or predictors. Matrix plots can also be created to display the relationship between the response variable and one or many independent variables simultaneously.
- The visual displays in the form of matrix plots can show whether there is a linear or non-linear relationship between the response and each of the independent variables or the predictors.
- They also display whether there is a direct or indirect relationships between the response and the independent variables.
- This information obtained from the matrix plots is very helpful in building the correct model and prediction equation.

Examples of Matrix Plot

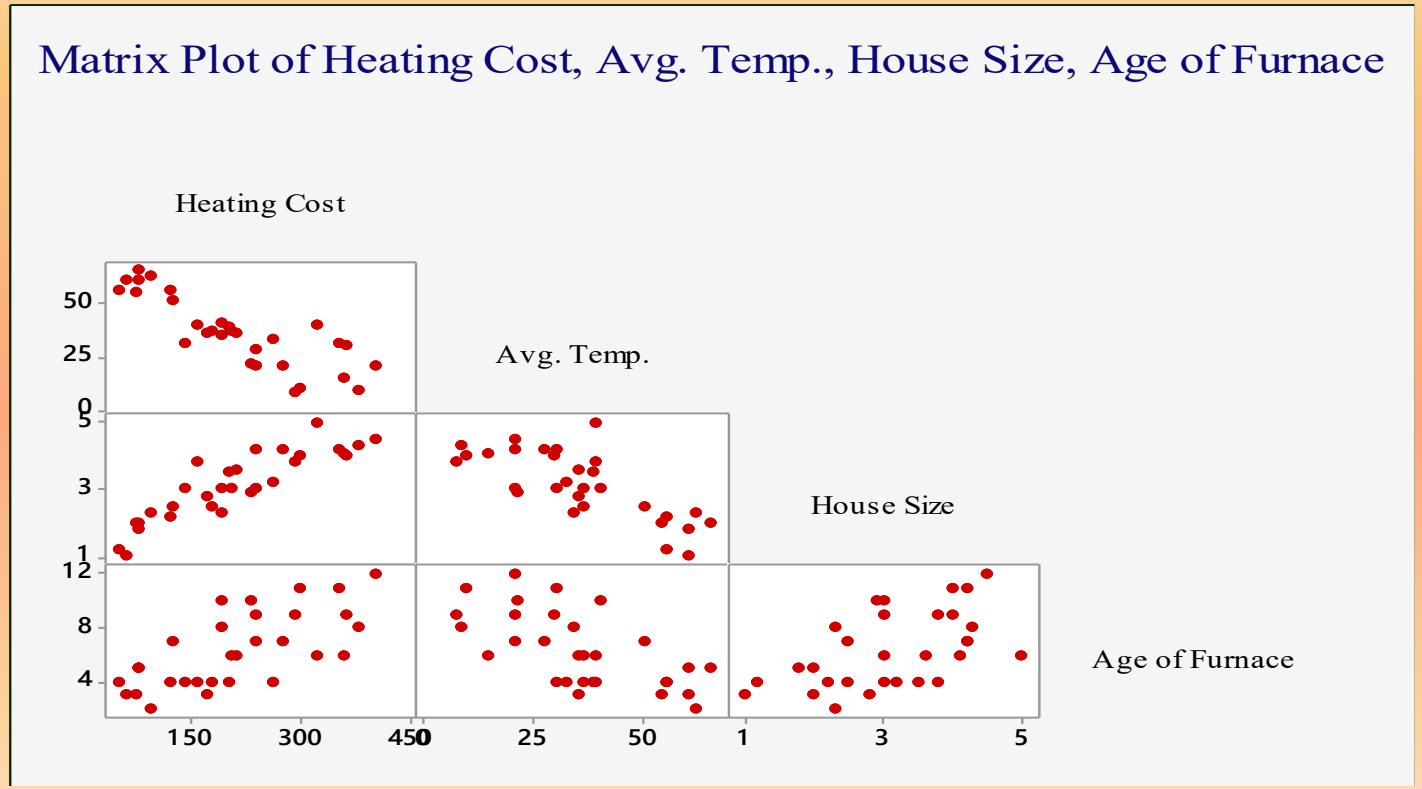
Matrix plot of the dependent variable –Heating Cost (y) with each of the independent variables, Average Temperature , House Size and Age of the Furnace



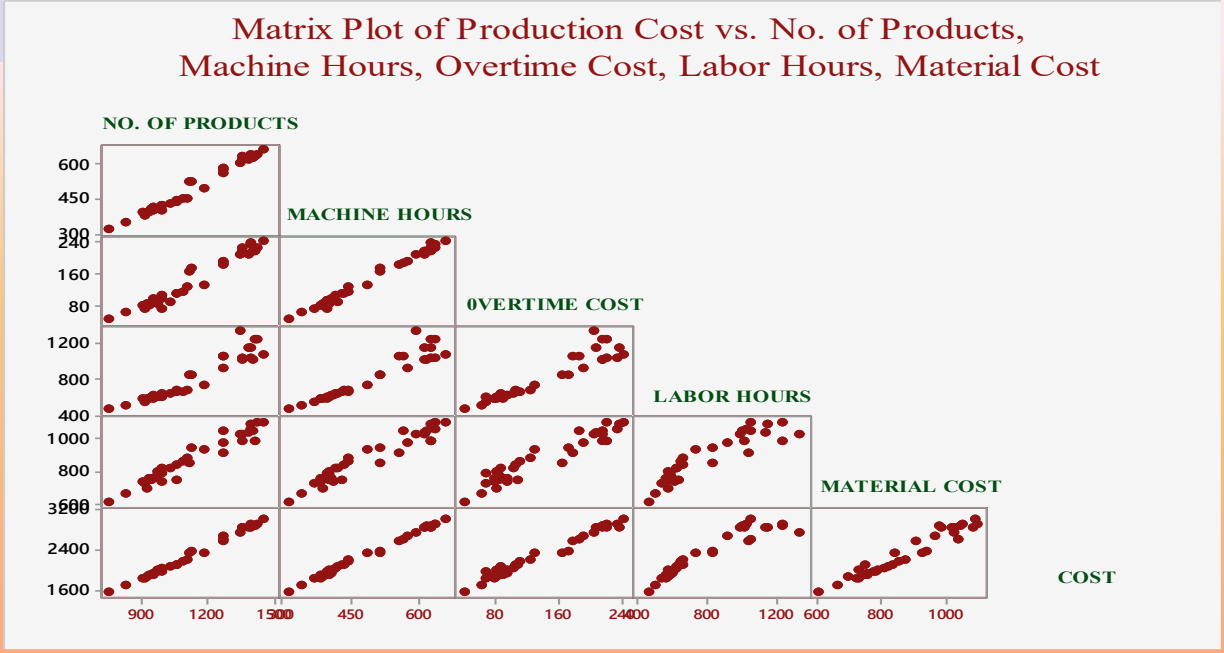


Matrix Plot_ Examples

Figure shows another form of matrix plot depicting the relationship between the home heating cost based on the average outside temperature, size of the house (x1000 square feet), and the life of the furnace (years) by creating an array of scatterplots.

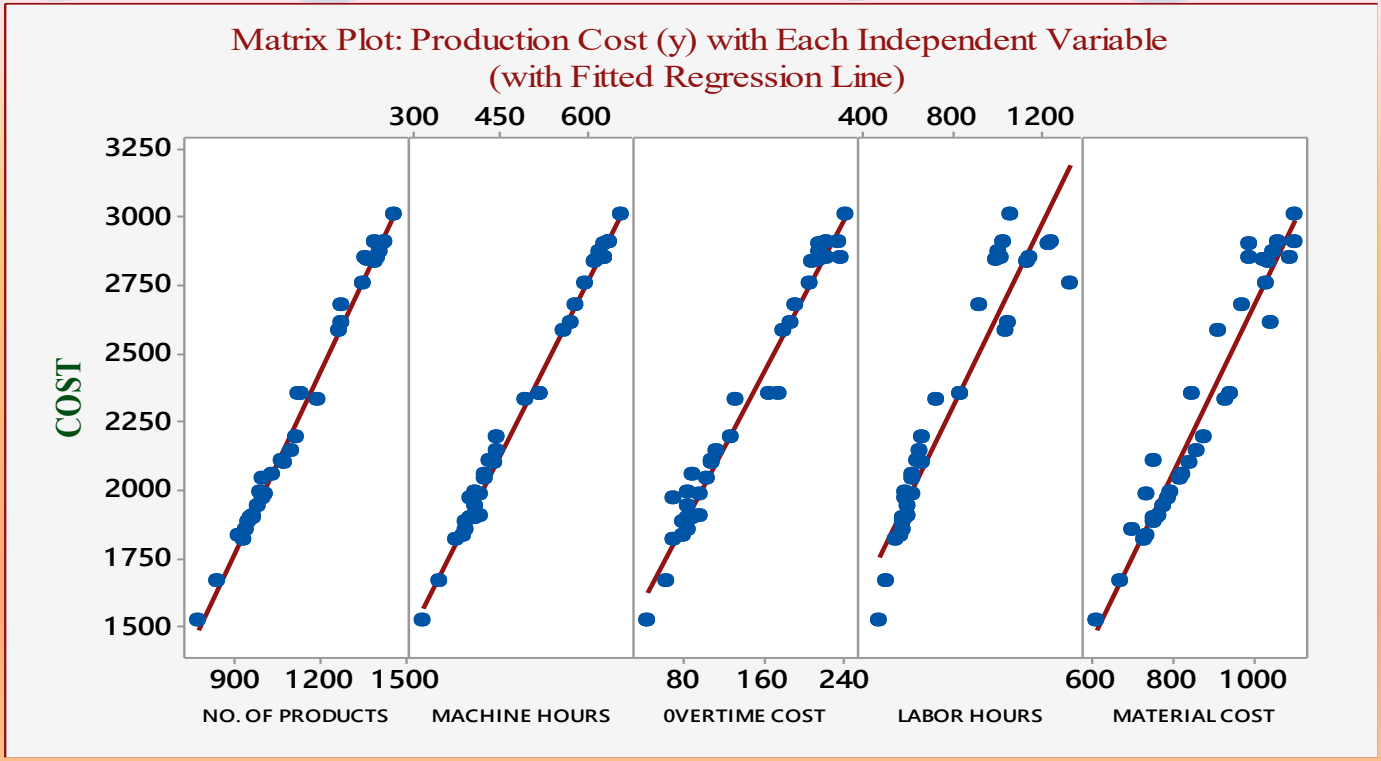


Matrix Plot

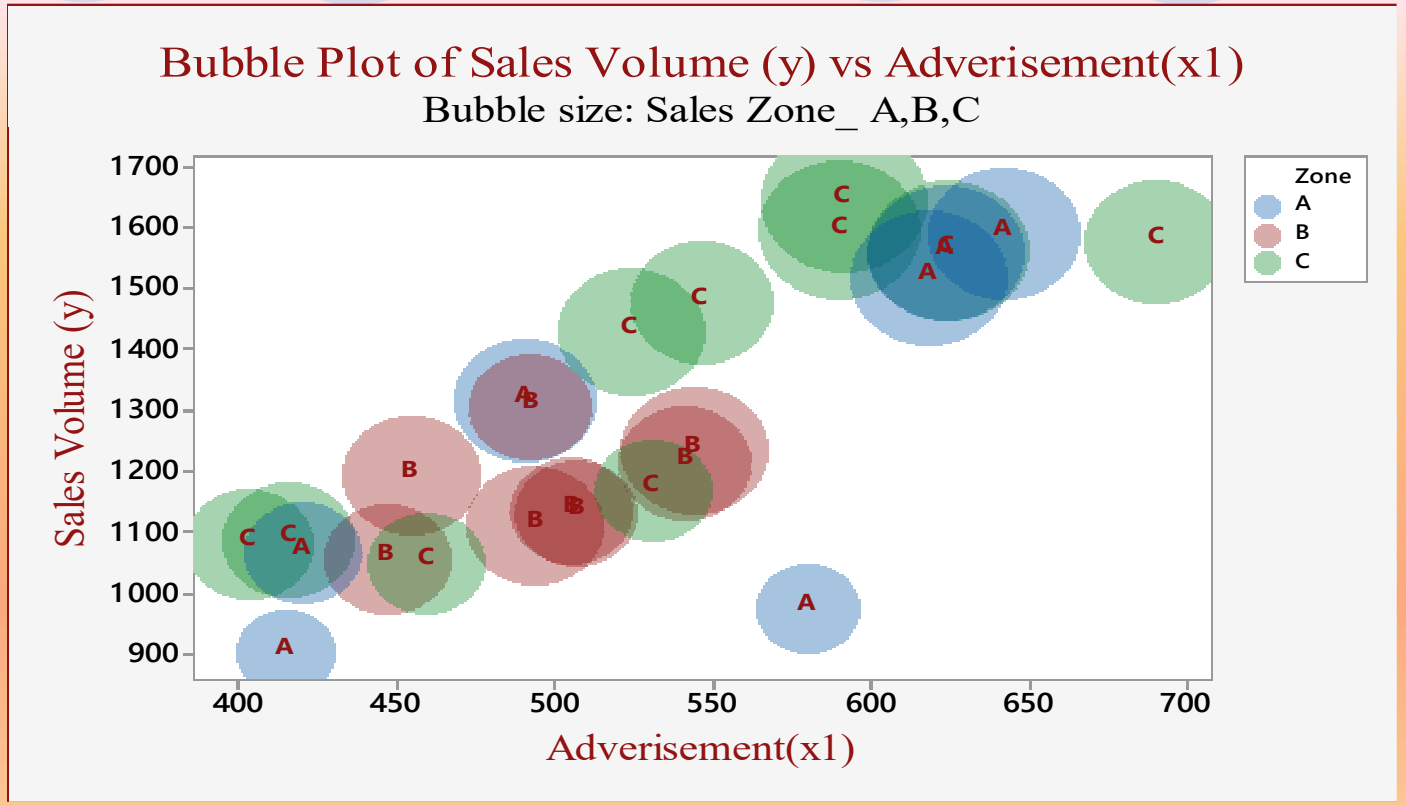


The bottom row shows production cost (COST) (response var.) versus each of the predictor variables in various columns, with production cost (COST) on the vertical axis. COST versus NO. OF PRODUCTS is shown in the lower left corner. Next to that is the plot of COST versus MACHINE HOURS followed by the plots of COST versus OVERTIME COST, COST versus LABOR HOURS, and COST versus MATERIAL COST. The other plots show the relationships among predictor variables. For example, the top left plot shows the relationship between NO. OF PRODUCTS and MACHINE HOURS (as the number of products increases so does the machine hours).

Matrix Plot

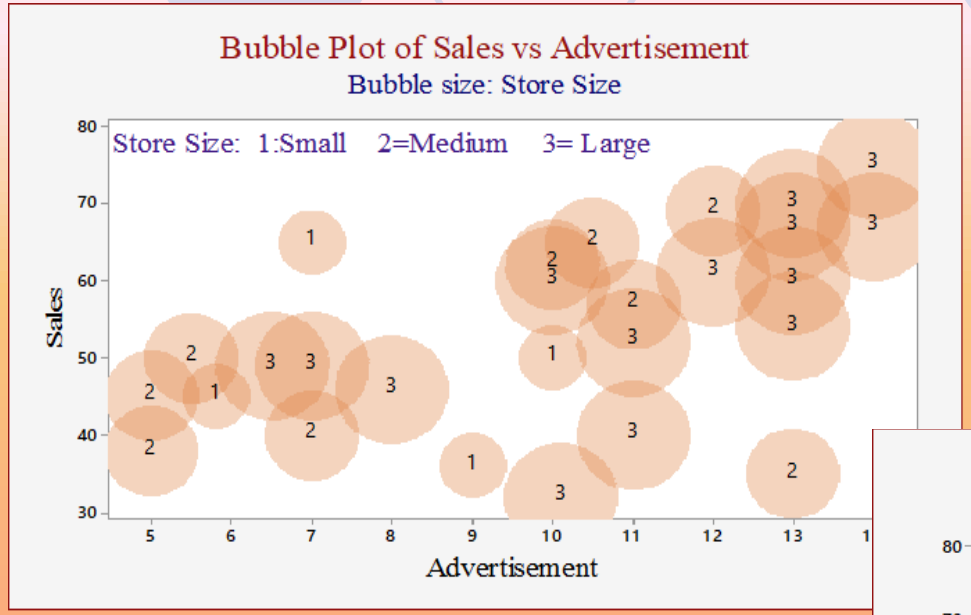

Matrix Plot: Production Cost (y) with Each Independent Variable
(with Fitted Regression Line)
COST
3250
3000
2750
2500
2250
2000
1750
1500
300
450
600
400
800
1200
900
1200
1500
80
160
240
600
800
1000
NO. OF PRODUCTS
MACHINE HOURS
0VERTIME COST
LABOR HOURS
MATERIAL COST

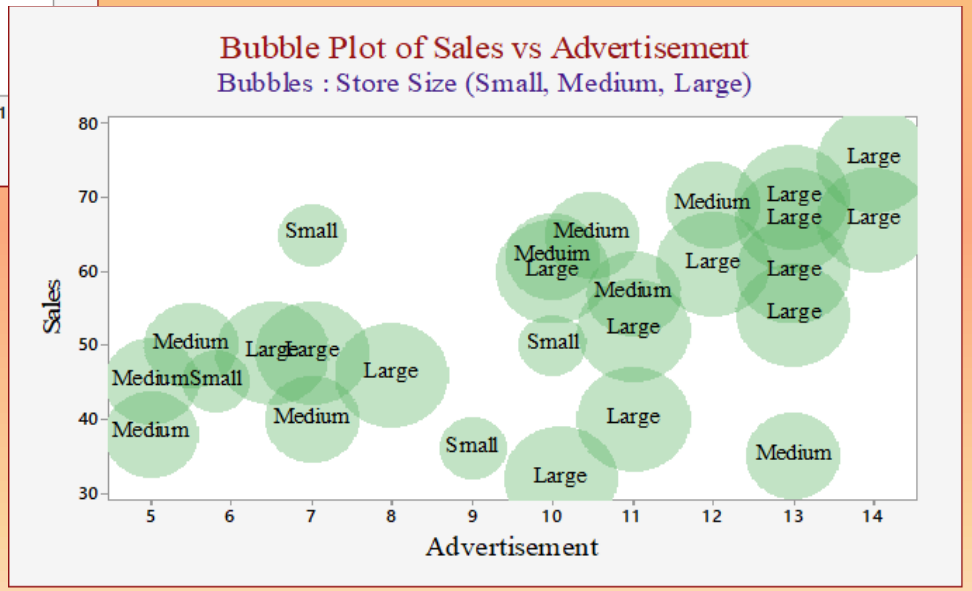
Bubble Plot _Relationship between Three Variables



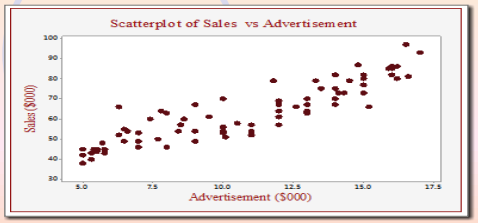
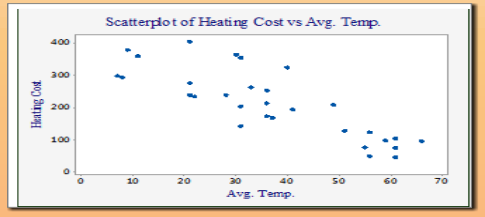
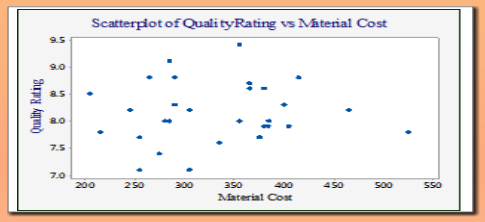
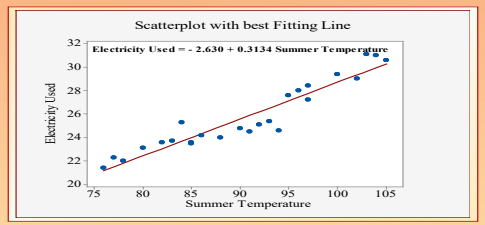
Bubble Plot Showing the relationship between Sales, Advertisement, and Sales Zones

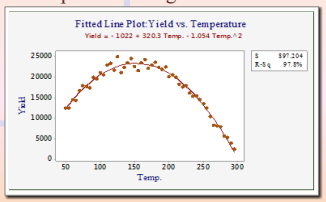
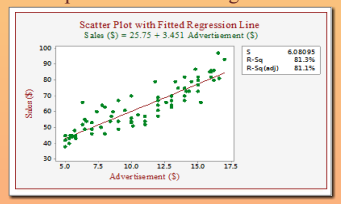
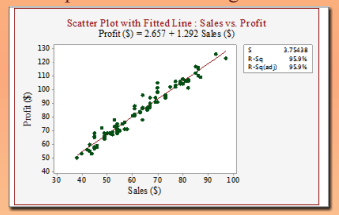
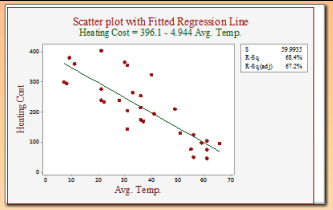
Bubble Plot _ variation

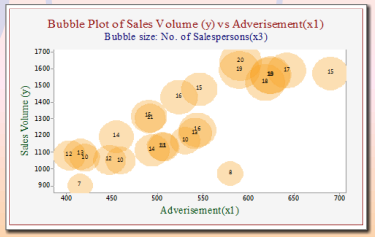
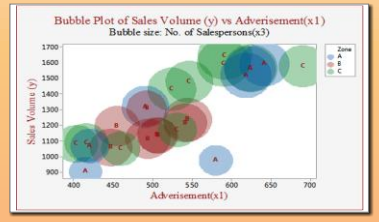
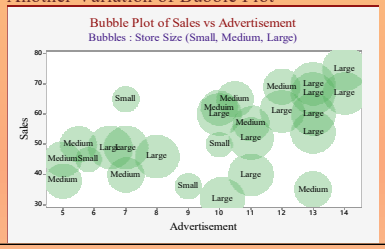
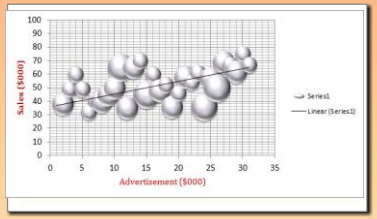


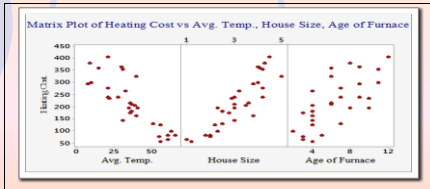
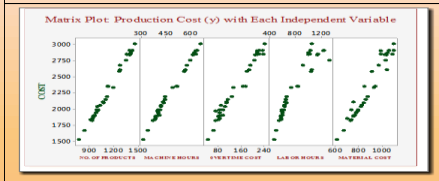
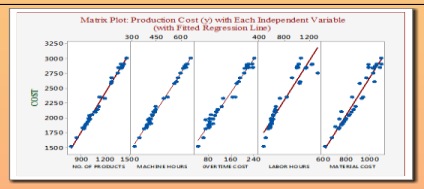
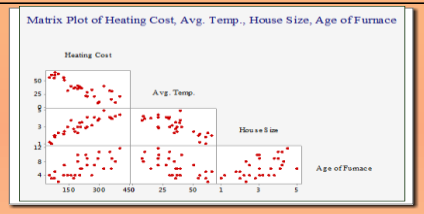
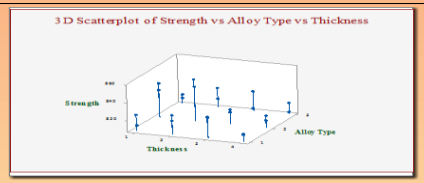


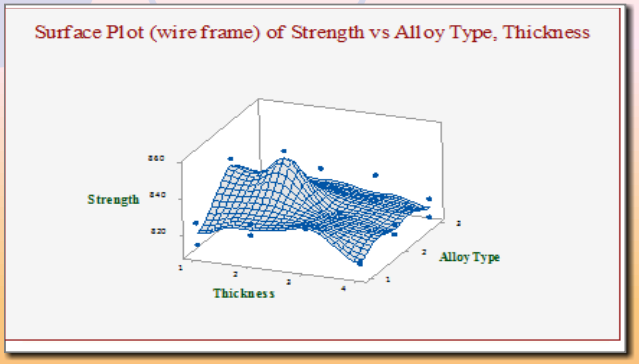
Summary

Types of Graphs/Charts	Description/Application
Scatter Plot 	**A scatterplot is a two-dimensional plot The pairs of points (x_i, y_i) plotted on the scatterplot are helpful in *visually* examining the relationship between the two variables. A scatterplot showing a *positive relationship* between the two variables is shown here.**
A Variation of Scatter Plot 	**A scatterplot showing an *inverse relationship* between the two variables. This means that an increase in x variable leads to a decrease in the y-variable.**
Another Variation of Scatter Plot 	**A scatterplot showing a weak or no relationship between the two variables.**
Fitted Line Plot 	**A line fitted over the scatter plot. The line is known as the regression line and is known as the line of "best fit." This line is uniquely determined using a mathematical technique known as the *least squares method*.**

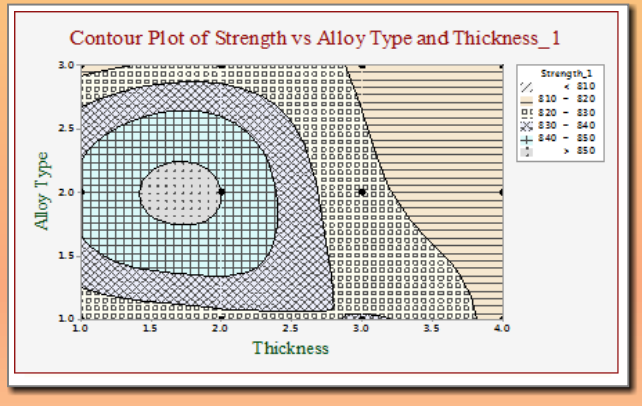
Scatterplot Showing Nonlinear Relationship 	**Non-linear Relationship: The graph on the left shows that the relationship between the two variables under study may be non-linear.** **The scatterplot of the variables temperature (x) and the yield (y) shows a nonlinear relationship that can be best approximated by a quadratic equation. The equation of the fitted curve obtained using a computer package is $y = -1022 + 320.3x - 1.054x^2$.** **This equation can be used to predict the yield (y) for a particular temperature (x).**
Scatterplot with Fitted Regression Line 	**A fitted line over the scatter plot provides the best fitting line through the data points. The equation of this line is used to predict the dependent variable, y using the independent variable, x .**
Scatterplot with Fitted Regression Line 	**A variation of fitted regression line with regression equation. The points are very close to the fitted line. This is an indication of smaller error and better fit.**
	**A variation of fitted regression line with negative slope. This is an indication of inverse relationship between the two variables.**

Bubble Plot 	**The bubble plot is used to explore the relationships among three variables on a single plot. The plot uses bubbles to plot the third variable hence the name *bubble plot*. This plot uses bubbles of different sizes to represent the third variable. The area of the bubble represents the value of the third variable. The plot shows the relationship between the sales, advertisement and the number of salespersons which is the third variable represented using bubbles.**
A Variation of Bubble Plot 	**Bubble plot showing the relationship between the sales, advertisement dollars spent, and the sales zone.**
Another Variation of Bubble Plot 	**In this plot, the bubbles show the store sizes – small, medium, and large.**
A Bubble Plot with a Trend Line 	**A bubble plot with a trend line showing an increasing trend between the sales and advertisement. The trend line helps to visualize the relationship between the variables.**

	**A matrix plot is used to investigate the relationships between pairs of variables by creating an array of scatterplots. In regression analysis and modeling, often the relationship between multiple variables is of interest. In such cases, matrix plots can be created to visually investigate the relationship between the response variable and *each* of the independent variables or predictors.**
	**A variation of the matrix plot - Plot of response variable (cost) on the y-axis with each of the independent variables.**
	**A variation of the matrix plot - This matrix plot shows the fitted regression lines on each plot. The response variable is the cost and the variables on the x-axis are the independent variables.**
	**The plot shows another form of matrix plot depicting the relationship between the home heating cost (the response variable) based on the average outside temperature, size of the house (x1000 square feet), and the life of the furnace (years) by creating an array of scatterplots.**
	**A 3D scatter plot with projected lines can sometimes be useful in visualizing the relationships between three variables.**



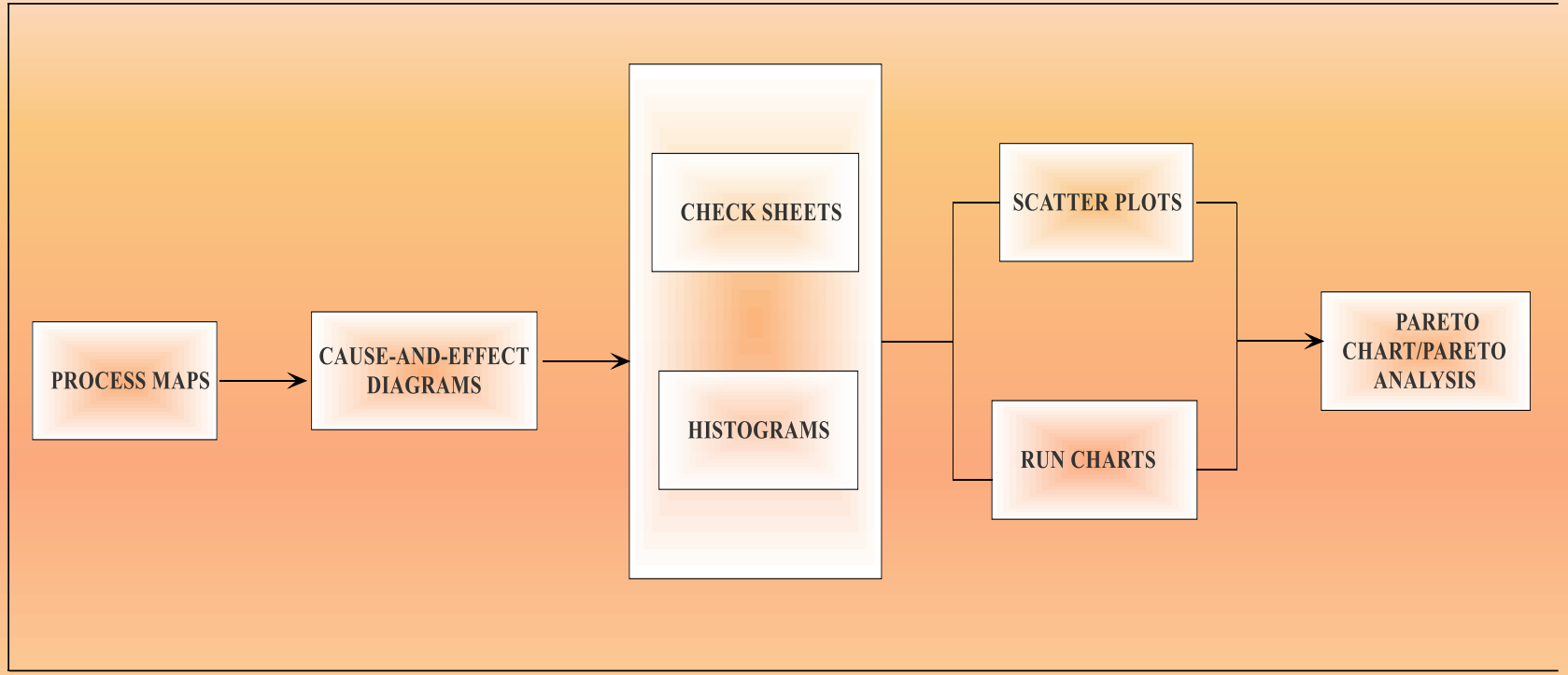
The 3D surface plot is used to explore the relationships between three variables. The Surface Plot uses interpolation to produce a continuous surface (surface plot) or grid (wireframe plot) of z-values to fit the data.



A contour plot shows the values for two variables on the x- and y-axes, while the values of the third variable are represented by shaded regions, called contours. This plot is like a topographical map in which x-, y-, and z-values are plotted instead of longitude, latitude, and altitude. The contour plot shown here is constructed from the surface plot above. The highest z-values are at the intersection of x and y values and are shown in the center as a circular region.

Section 5:
Seven Basic Tools of Quality

(1)Process Maps **(2) Check sheets**
(3) Histograms **(4) Scatter Diagrams**
(5) Run/Control Charts
(6)Cause-and-Effect (Ishikawa)/Fishbone Diagrams (7) Pareto Charts/Pareto Analysis


PROCESS MAPS
CAUSE-AND-EFFECT DIAGRAMS
CHECK SHEETS
HISTOGRAMS
SCATTER PLOTS
RUN CHARTS
PARETO CHART/PARETO ANALYSIS

Process mapping is flow charting a production or service process. The chart provides a model of an operation that facilitates communication about the various steps in the process. Any organization, or any of its parts, can be viewed as a process.

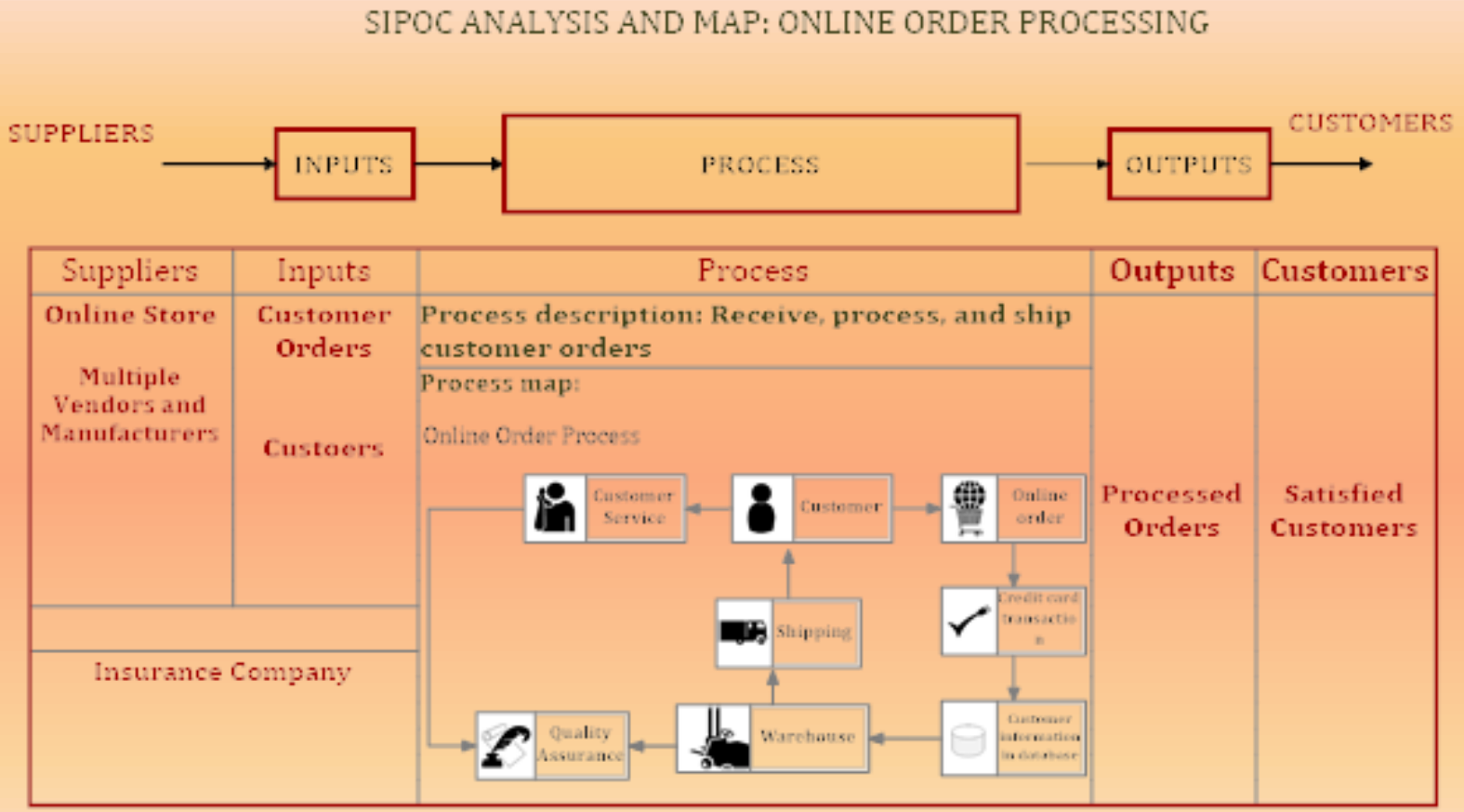
Process Maps

- Process mapping is flow charting a production or service process.
- The chart provides a model of an operation that facilitates communication about the various steps in the process.
- Any organization, or any of its parts, can be viewed as a process.

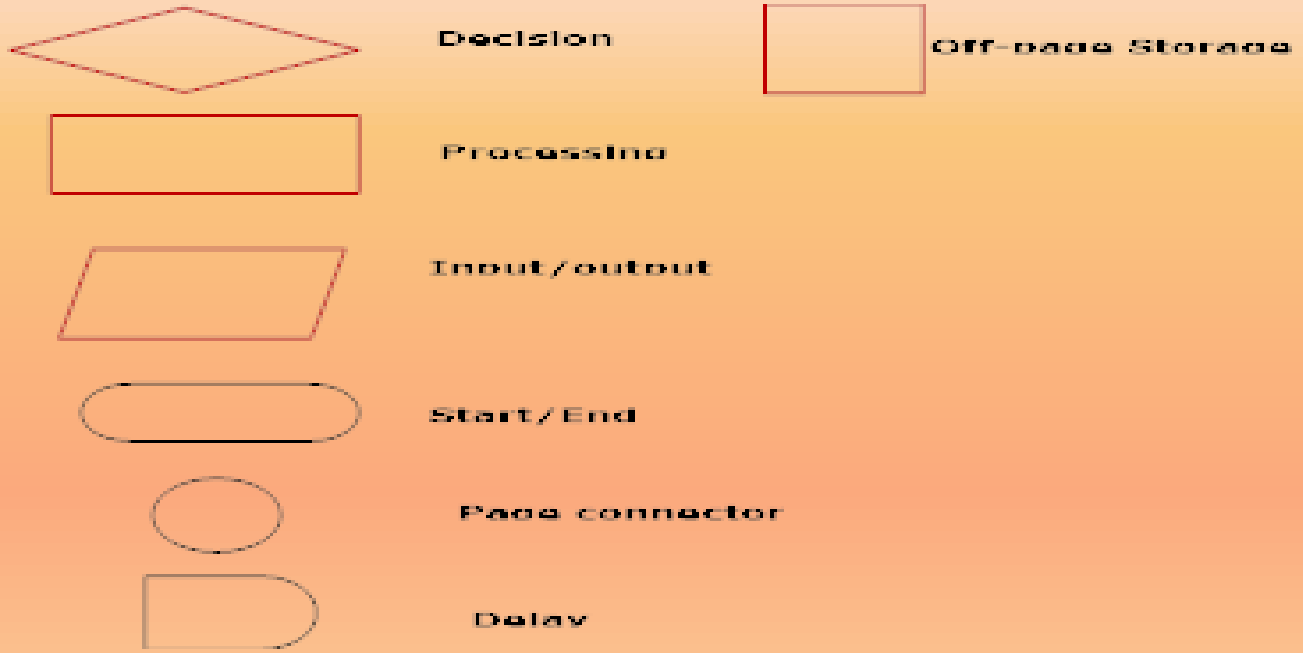
SIPOC Process Map: A high-level process map.

- The SIPOC (supplier, input, process, output, and customer) Identifies the inputs, outputs, suppliers, the process under investigation, and the customers of the process.

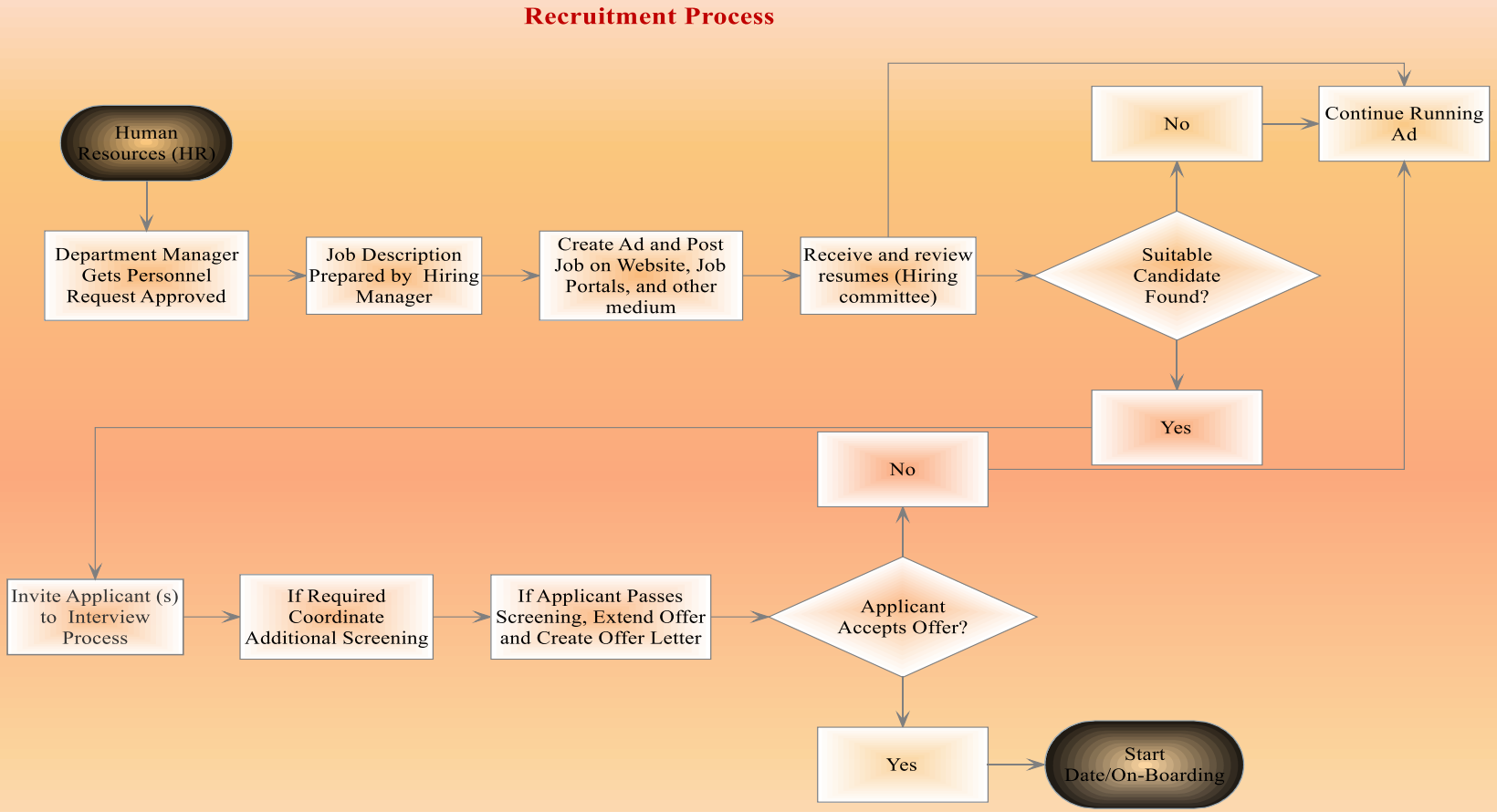
An example of a simple Process Map .



Symbols and their Meaning in Process Mapping



Examples of a Flow Process Charts



Histograms

- A Histogram is created using a frequency distribution of data.
- Helpful in many ways in data visualization applications.

In quality, and data analysis, histograms are useful in

- determining the shape of the distribution, i. e., whether the shape is symmetrical or skewed,
- determining the concentration of data points that is, which intervals or classes have more values in them,
- detecting process problems-including a shift in the process,
- evaluating process capability (ability of the process to be within its specification limits),
- determining how well centered the process is, or how close the data values are to the target value and determining the process variation.

Histogram: Detecting a Shift and Variation in the Process

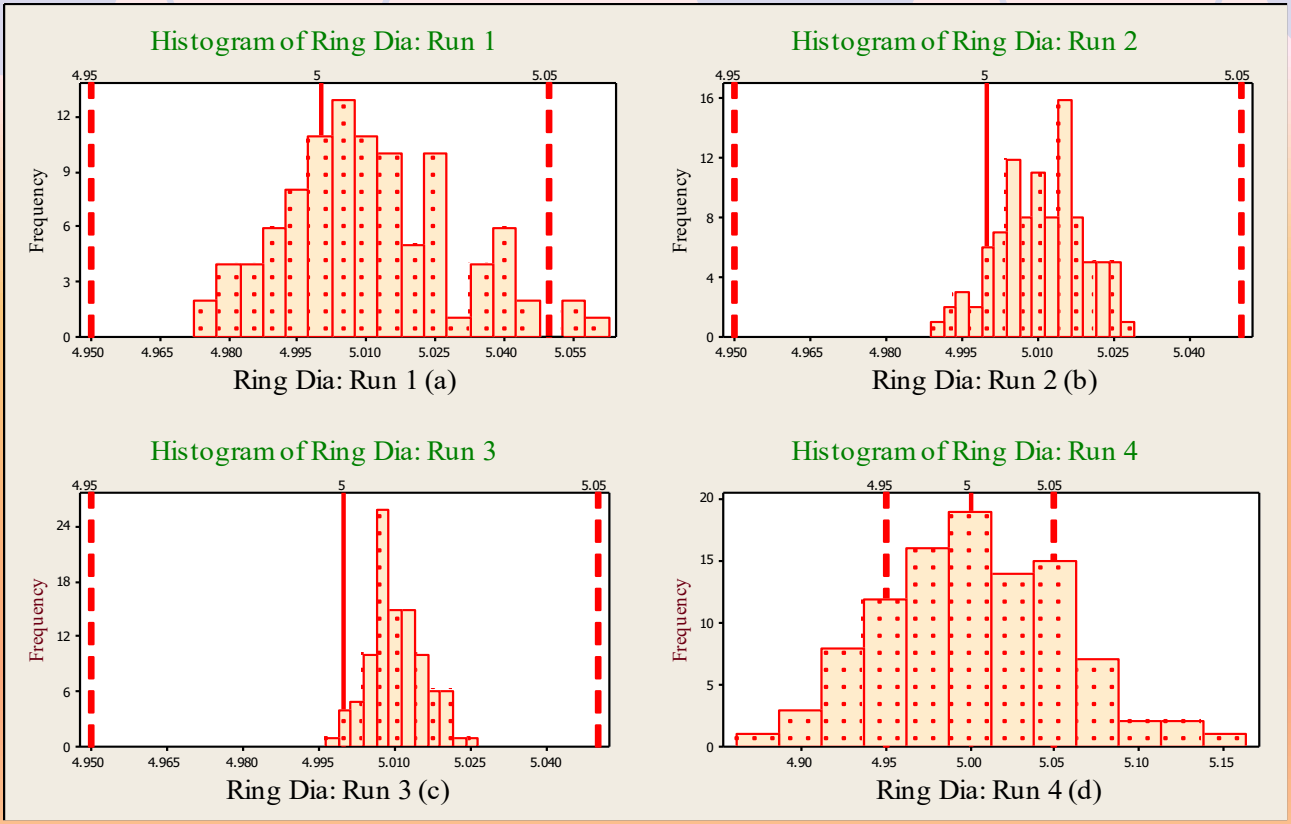


Figure (a) Most values within specification, some assignable causes may be present (b) Process within control but has slightly shifted to the right (c) Process variation has reduced compared to (a) and (b) above but there is a shift to the right, (d) Process out of control and has large variation

Histogram: Detecting a Shift and Variation in the Process

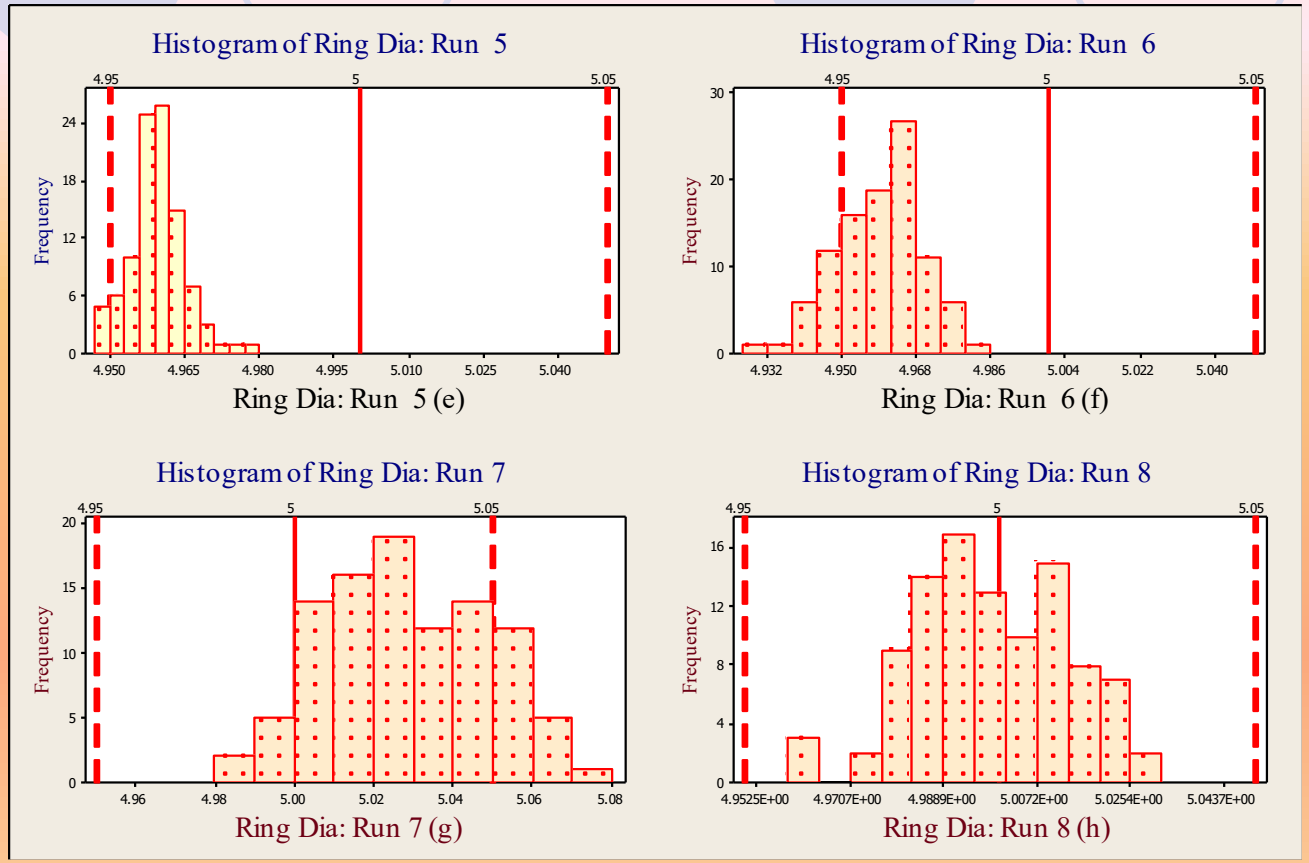
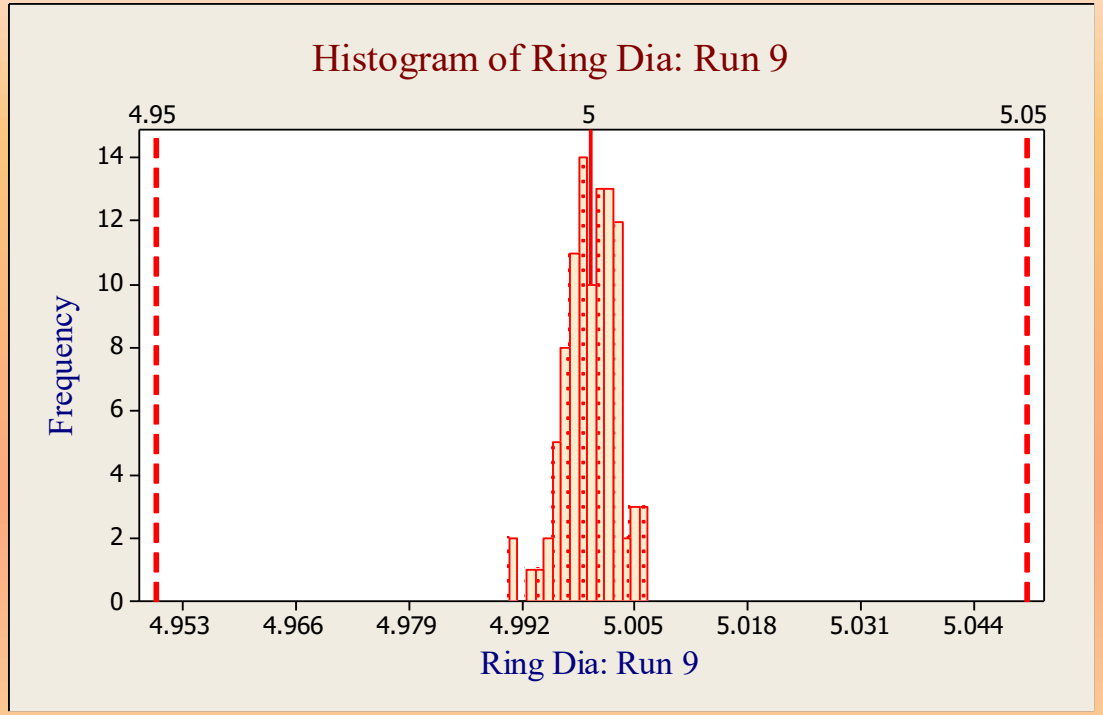


Figure (e) The process has shifted to the left; products out of specification, (f) Process shift to the left; more variation compared to (e), (g) Process out of control and has large variation, (h) Process within control but has large variation,

Histogram: Detecting a Shift and Variation in the Process

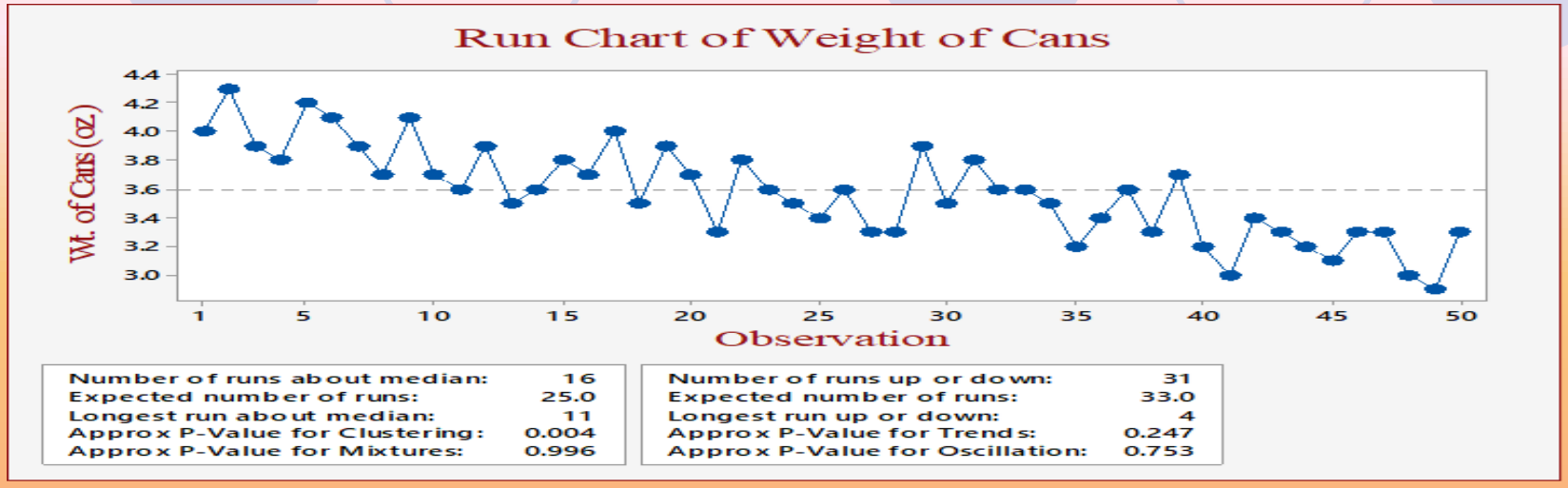


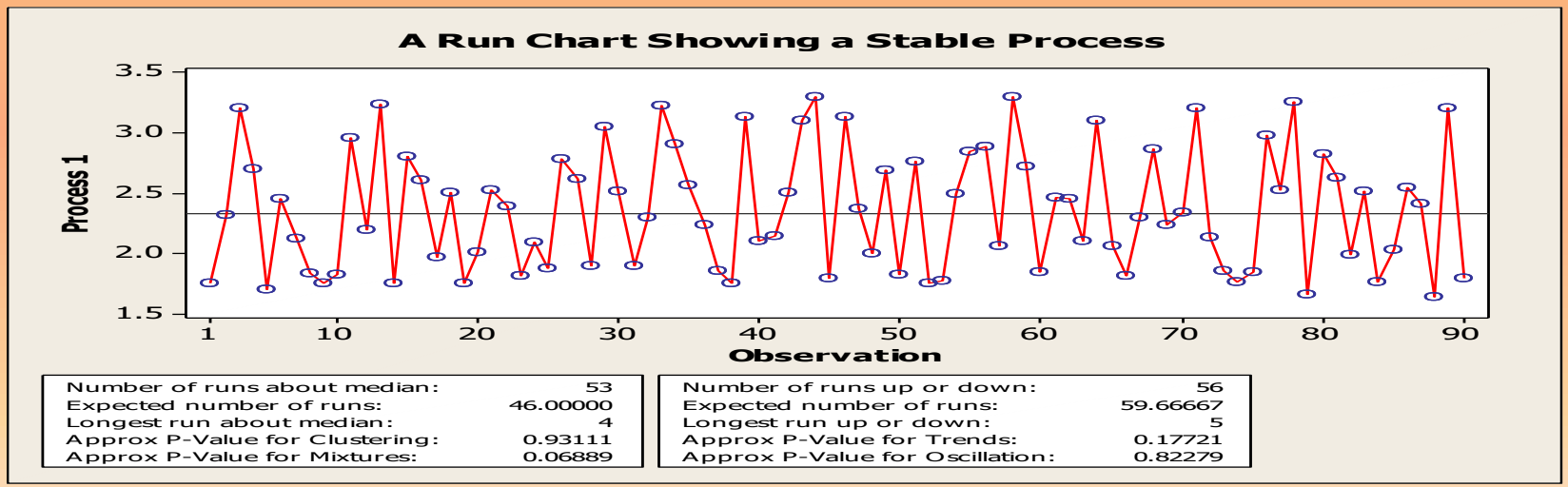
Process stable and close to the target (desirable)

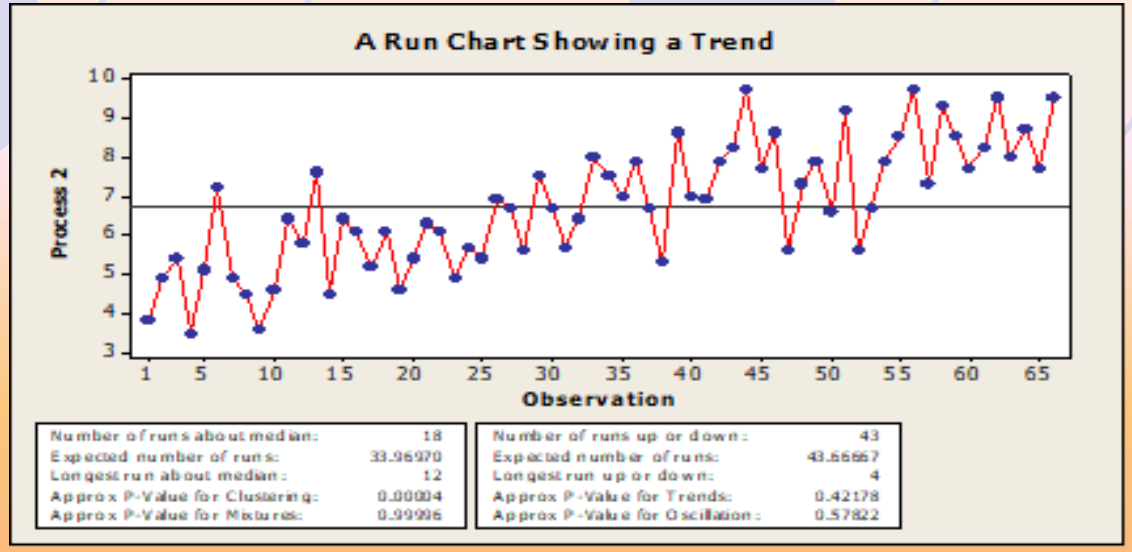
Run Chart

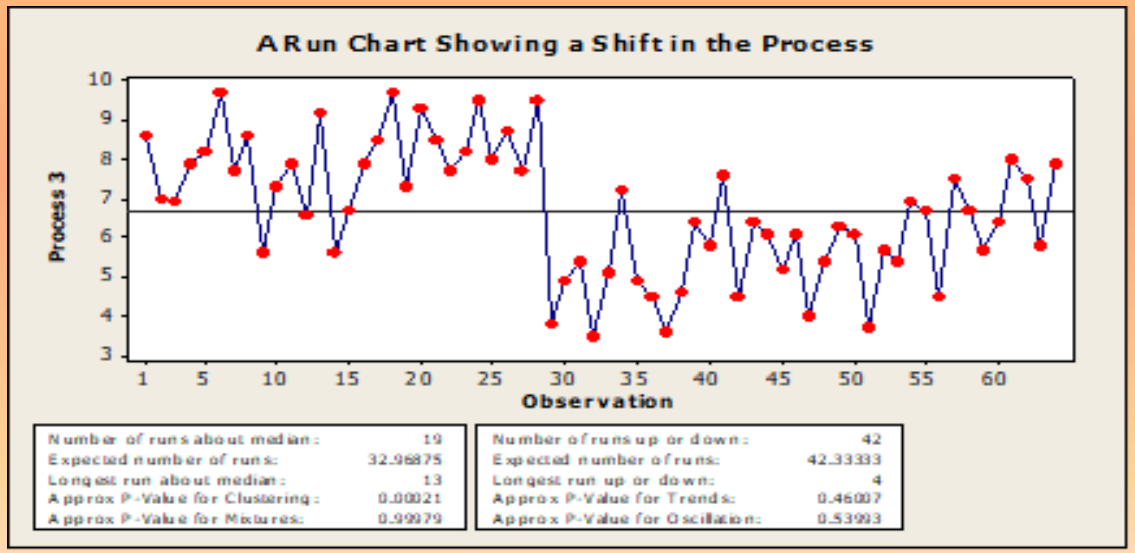
- A run chart is used in quality to analyze the data either in the development stage of a product or before the state of statistical control.
- As explained , histograms are very useful in assessing the stability and capability of a process.
- However, histograms do not plot the data over time.
- To determine the stability and capability, plotting the data over time is very helpful.
- A histogram usually leaves out an essential element: time.
- A run chart is a very simple and helpful tool in showing the stability and variation in the process over some time period.
- It is one of the simple ways to investigate changes in the process over time.

Run Charts

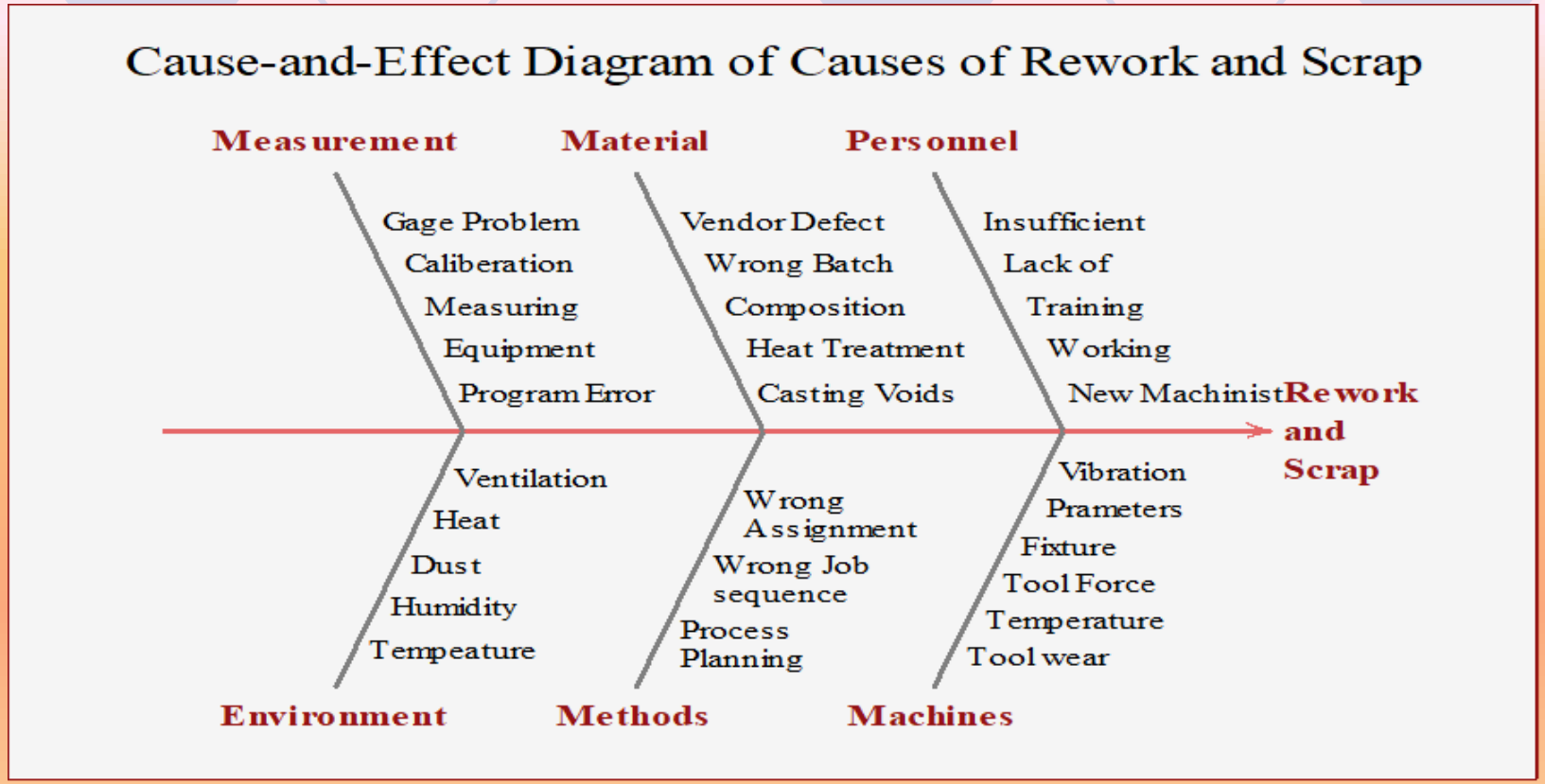

Run Chart of Weight of Cans
Wt. of Cans (oz.)
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
1
5
10
15
20
25
30
35
40
45
50
Observation
Number of runs about median: 16
Expected number of runs: 25.0
Longest run about median: 11
Approx P-Value for Clustering: 0.004
Approx P-Value for Mixtures: 0.996
Number of runs up or down: 31
Expected number of runs: 33.0
Longest run up or down: 4
Approx P-Value for Trends: 0.247
Approx P-Value for Oscillation: 0.753


A Run Chart Showing a Stable Process
Process 1
3.5
3.0
2.5
2.0
1.5
1
10
20
30
40
50
60
70
80
90
Observation
Number of runs about median: 53
Expected number of runs: 46.00000
Longest run about median: 4
Approx P-Value for Clustering: 0.93111
Approx P-Value for Mixtures: 0.06889
Number of runs up or down: 56
Expected number of runs: 59.66667
Longest run up or down: 5
Approx P-Value for Trends: 0.17721
Approx P-Value for Oscillation: 0.82279


A Run Chart Showing a Trend
Process 2
Observation
Number of runs about median: 18
Expected number of runs: 33.96970
Longest run about median: 12
Approx P-Value for Clustering: 0.00004
Approx P-Value for Mixtures: 0.99996
Number of runs up or down: 43
Expected number of runs: 43.66667
Longest run up or down: 4
Approx P-Value for Trends: 0.42178
Approx P-Value for Oscillation: 0.57822

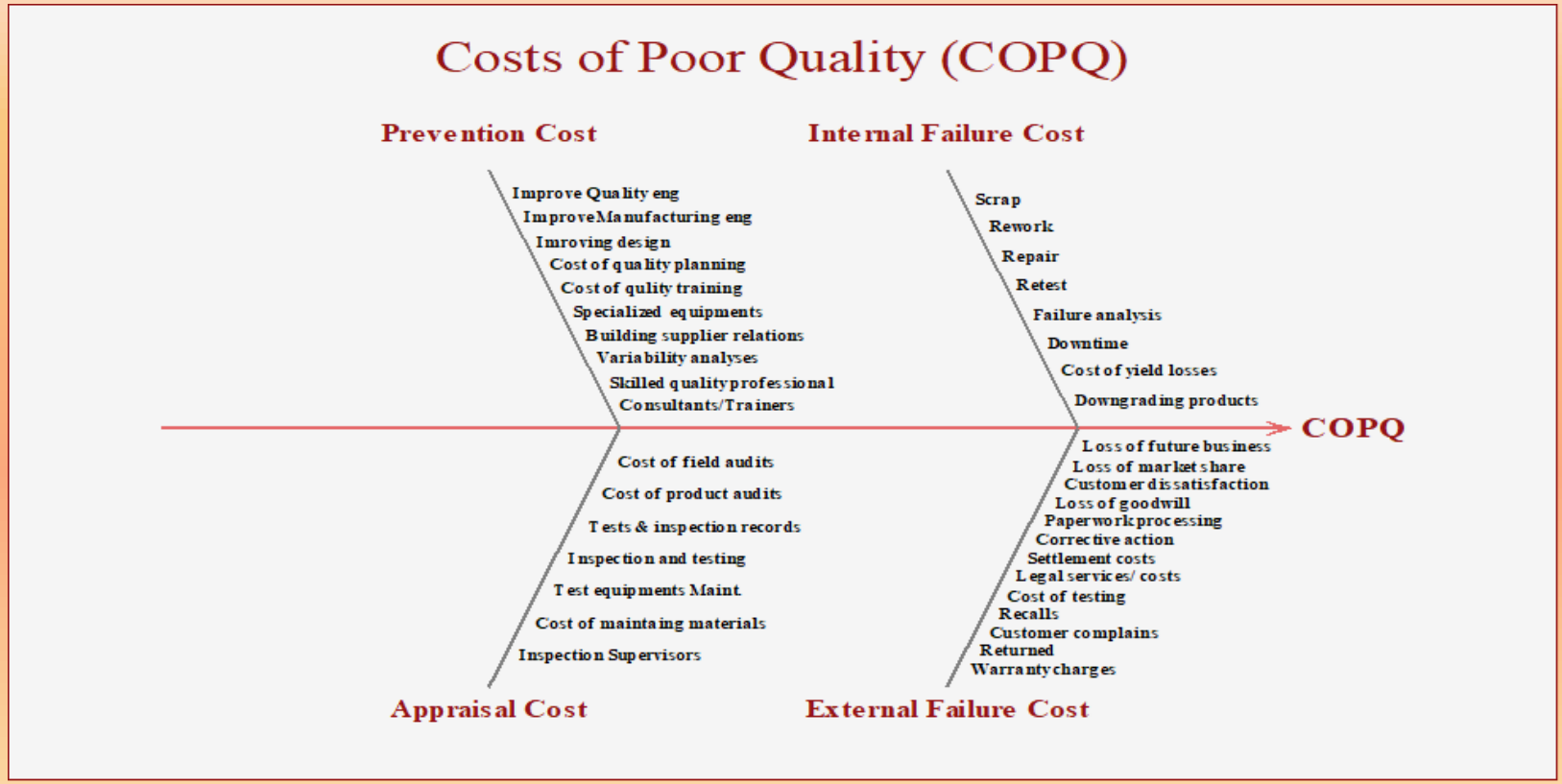

A Run Chart Showing a Shift in the Process
Process 3
Observation
Number of runs about median: 19
Expected number of runs: 32.98875
Longest run about median: 13
Approx P-Value for Clustering: 0.00021
Approx P-Value for Mixtures: 0.99979
Number of runs up or down: 42
Expected number of runs: 42.33333
Longest run up or down: 4
Approx P-Value for Trends: 0.46007
Approx P-Value for Oscillation: 0.53993

Cause-and-effect Diagram (1)



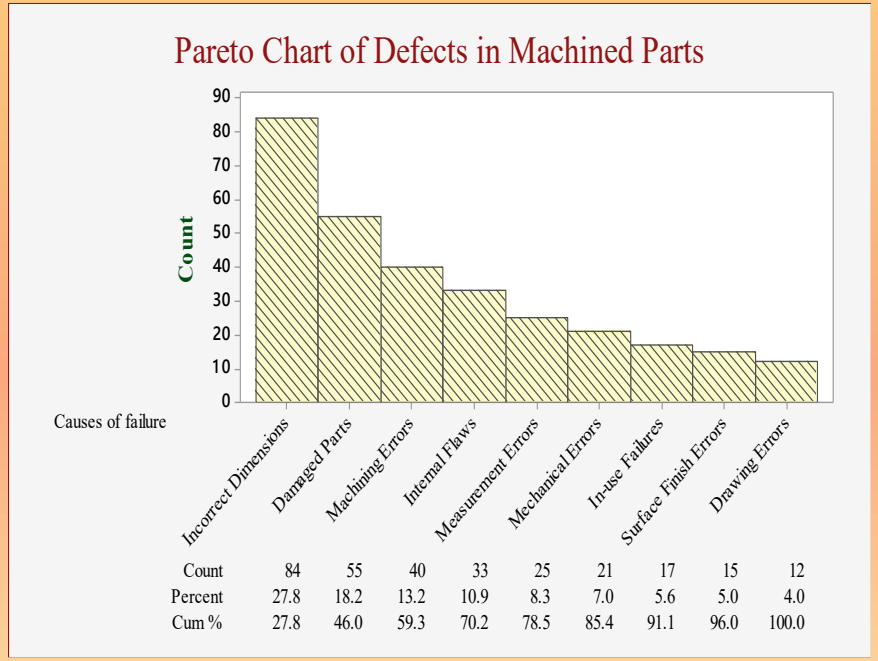
A cause-and-effect diagram is a very useful tool in establishing the relationship between the causes and their effects. Once a problem has been identified, it is necessary to analyze potential cause or causes of the problem. In such cases, the cause-and-effect diagram is a very helpful tool.

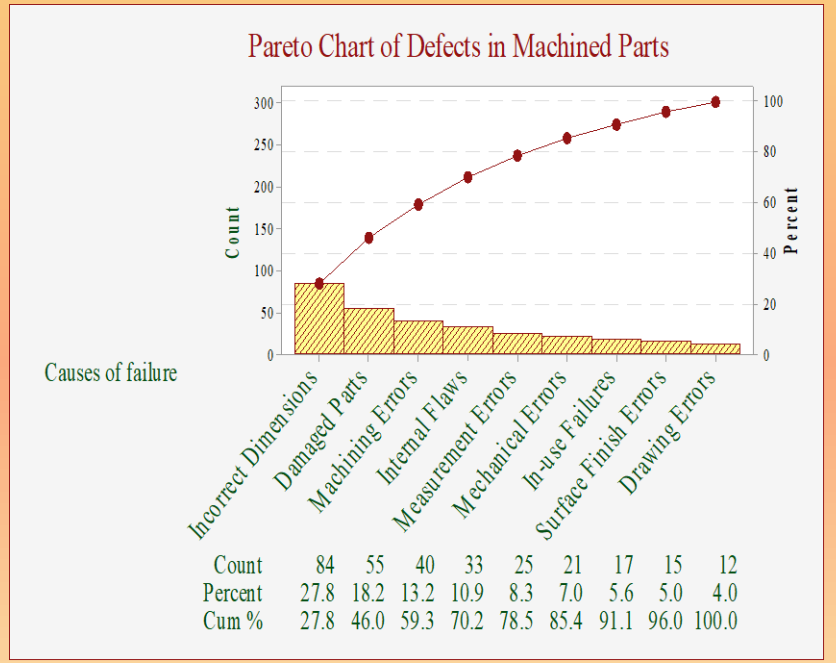
Cause-and-effect Diagram (2)



Pareto Charts

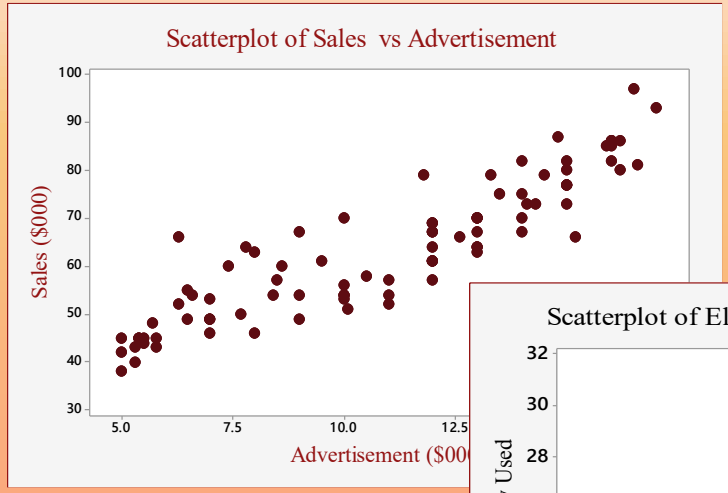
A Pareto chart is very similar to a histogram or a frequency distribution of attribute data where the bars are arranged by categories from largest to smallest with a line that shows the cumulative percentage and count of the bars.

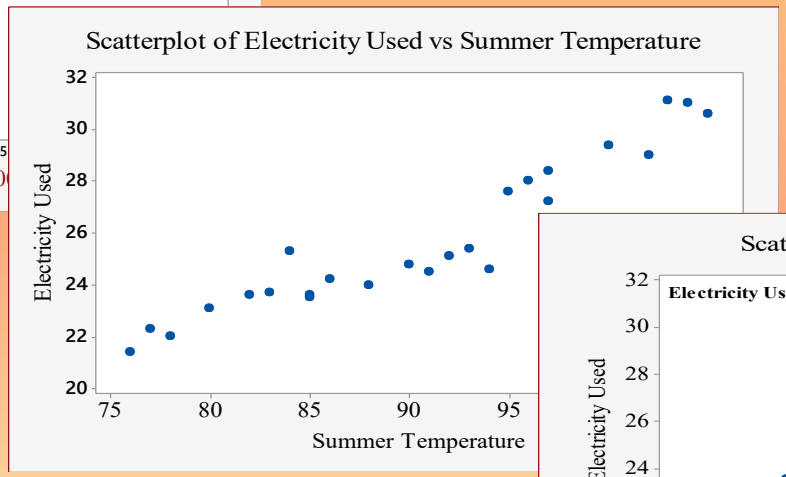


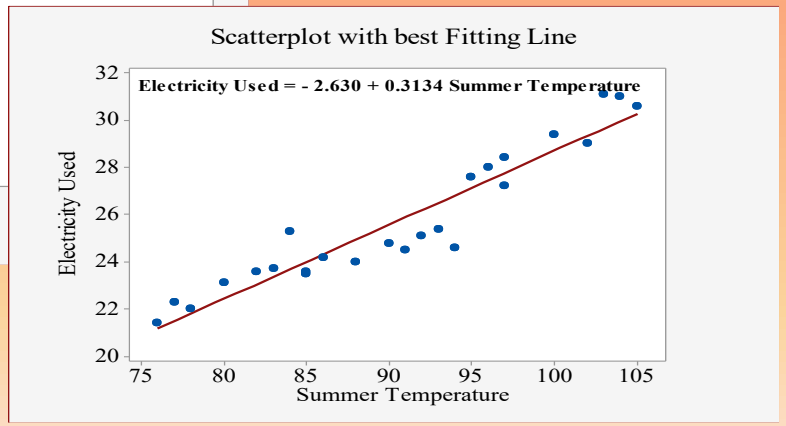


Scatter Plot

Describing the relationship between two quantitative variables is called a ***bivariate relationship***. One way of investigating this relationship is to construct a ***scatter plot***.







Check Sheets

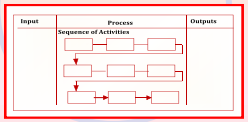
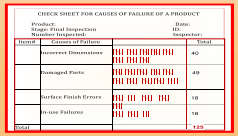
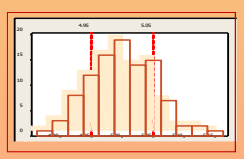
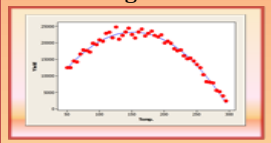
- Check sheets are data gathering tools.
- They help organize the data in a form that can be used easily for further analysis (creating histograms).

CHECK SHEET FOR CAUSES OF FAILURE OF A PRODUCT

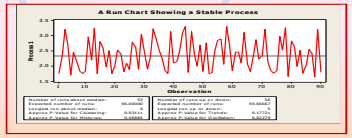
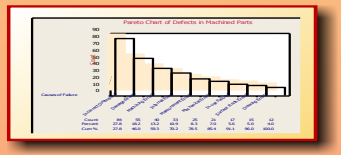
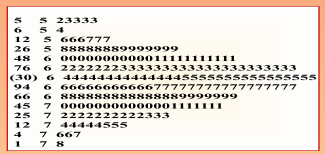
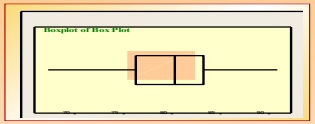
Product: Date:
Stage: Final Inspection ID:
Number Inspected: Inspector:

Item#	Causes of Failure		Total
	Incorrect Dimensions	𝍸 𝍸 𝍸𝍸 𝍸 𝍸 𝍸	40
	Damaged Parts	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	49
	Surface Finish Errors	𝍸 𝍸 𝍸 𝍸	18
	In-use Failures	𝍸 𝍸 𝍸 𝍸 𝍸	18
Total			125

Summary

Type of Chart/Graph	Description/Application	Number of Variables Plotted
Process Maps 	Process mapping is flow charting a production or service process. The chart provides a model of an operation that facilitates communication about the various steps in the process. Any organization, or any of its parts, can be viewed as a process. *A process is a transformation of inputs into outputs. A process can also be viewed as a sequence of activities performed by resources that transform inputs into outputs.* Variations of process maps are SIPOC diagram and VSM or value stream mapping.	
Check sheets 	Check sheets are data gathering tools. The purpose of check sheets is to ensure that the data are collected properly and accurately for analysis. Check sheets help organize the data in a form that can be used easily for further analysis (for example, creating histograms). Check sheets have many variations; they can be individualized for a situation.	
Histograms 	• Determining the shape and location of data measured on one characteristic. Also used for detecting process problems including a shift in the process either to the left or right evaluating process capability (ability of the process to be within its specification limits) • Determining how well centered the process is or how close the data values are to the target value • Determining the process variation	One variable plotted: Univariate data
Scatter Diagrams 	Scatter diagrams investigate the relationship between two quantitative variables of interest. Describing the relationship between two quantitative variables is called a *bivariate relationship*. One way of investigating this relationship is to construct a *scatterplot or a scatter diagram.* A scatterplot is a two-dimensional plot where one variable is plotted along the vertical axis and the other along the horizontal axis. The pair of points plotted on the scatterplot is	Two variables plotted: bivariate relationship

Summary

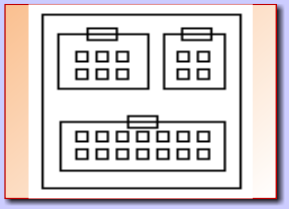
Type of Chart/Graph	Description/Application	No. of Variables Plotted
Run Chart 	A run chart is a very simple and helpful tool for visualizing the stability and variation in the process over some time period. It is one of the simpler ways to investigate changes in the process over time. In this chart, the measurements are plotted over time or in the order of production. The chart can also be used to identify the trend or shift in the process.	One variable plotted: Univariate data
Cause-and-Effect (Ishikawa)/Fish-bone Diagrams 	A cause-and-Effect diagram is a useful tool in establishing the relationship between the causes and their effects. Once a problem has been identified, this diagram can be used to analyze potential cause or causes of the problem. This is one of the critical tools in solving quality problems.	
Pareto Chart 	The chart shows (in descending order) the contribution of the vital few versus the trivial many. Used to identify the problems, causes, sources, or defects that should be considered first in the problem-solving process.	One variable divided into different categories: Univariate data.
Stem-and-leaf Plot 	A simple and useful way for summarizing and presenting data. The stem-and-leaf plot displays the range and concentration of the data. Easy to assess the distribution and estimate the percentiles from the data. Plots individual data points; no need for constructing frequency distribution.	One variable plotted: Univariate data
Box-Plot 	Plot of five measures: the minimum, first quartile, median, third quartile, and the maximum. Displays the distribution of underlying data. Used to detect the outliers in data.	One variable plotted: Univariate data

Section 6:
Seven New Tools of Quality

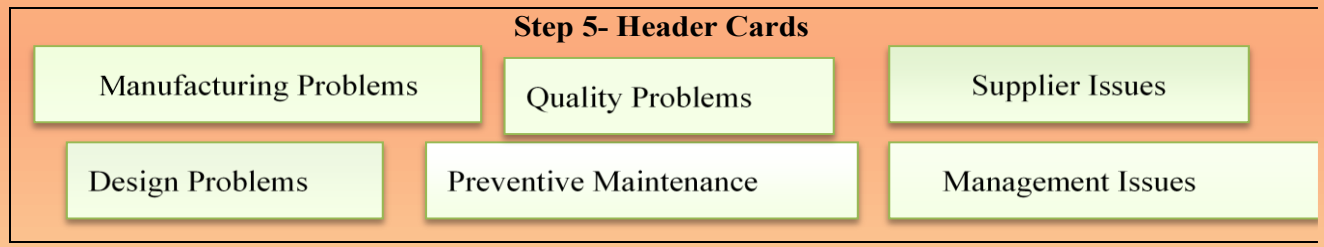
Seven new tools of Quality

These tools are referred to as graphical & *information visualization* tools. They have wide applications in decision making and quality improvement programs.
We will discuss the following visualization tools :

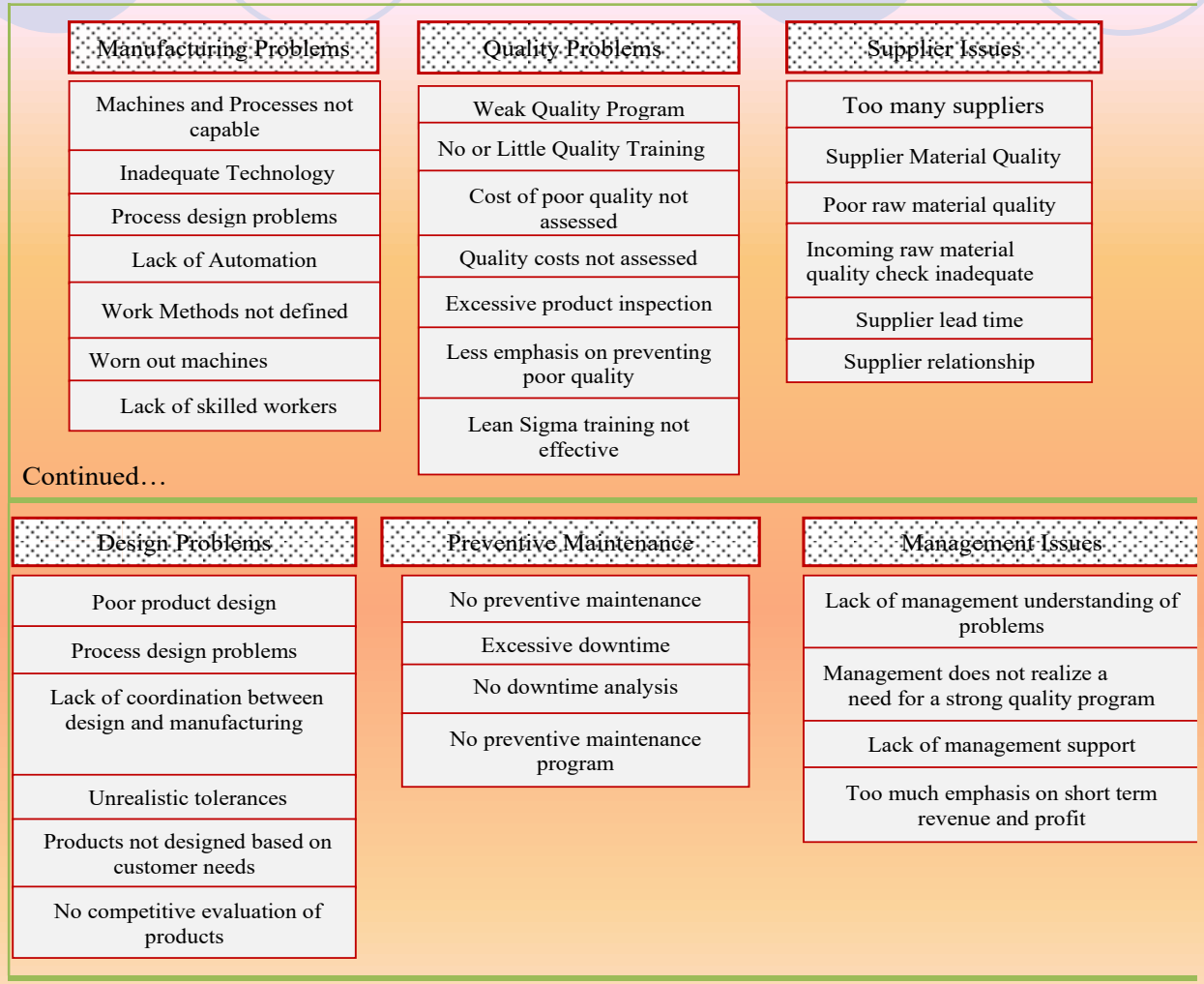
(1) Affinity Diagram
(2) Interrelationship Digraph
(3) Tree Diagram
(4) Prioritizing Matrices
(5) Matrix Diagram
(6) Process Decision Program Chart
(7) Activity Network Diagram

Affinity Diagram	
	Affinity diagram is a visual tool that gathers large amounts of data and organizes them into groups based on natural relationships. This is a special type of brainstorming tool.

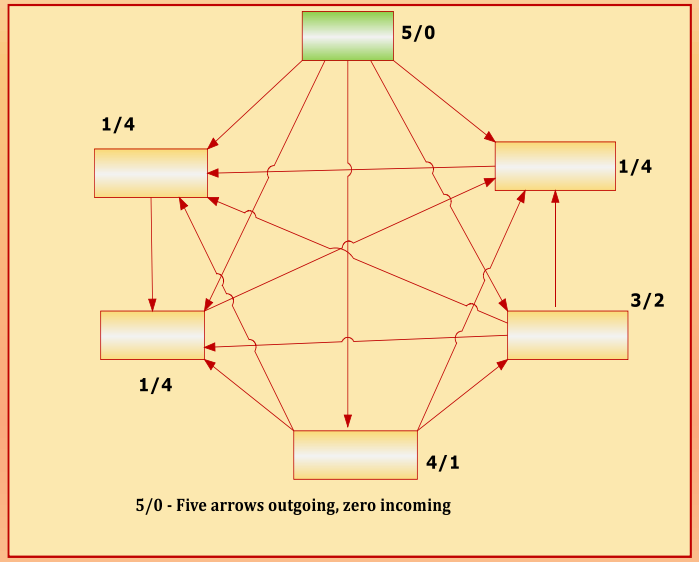
Issues Related to Poor Quality and Excessive Defective Products



Finished Affinity Diagram



Interrelationship Digraph

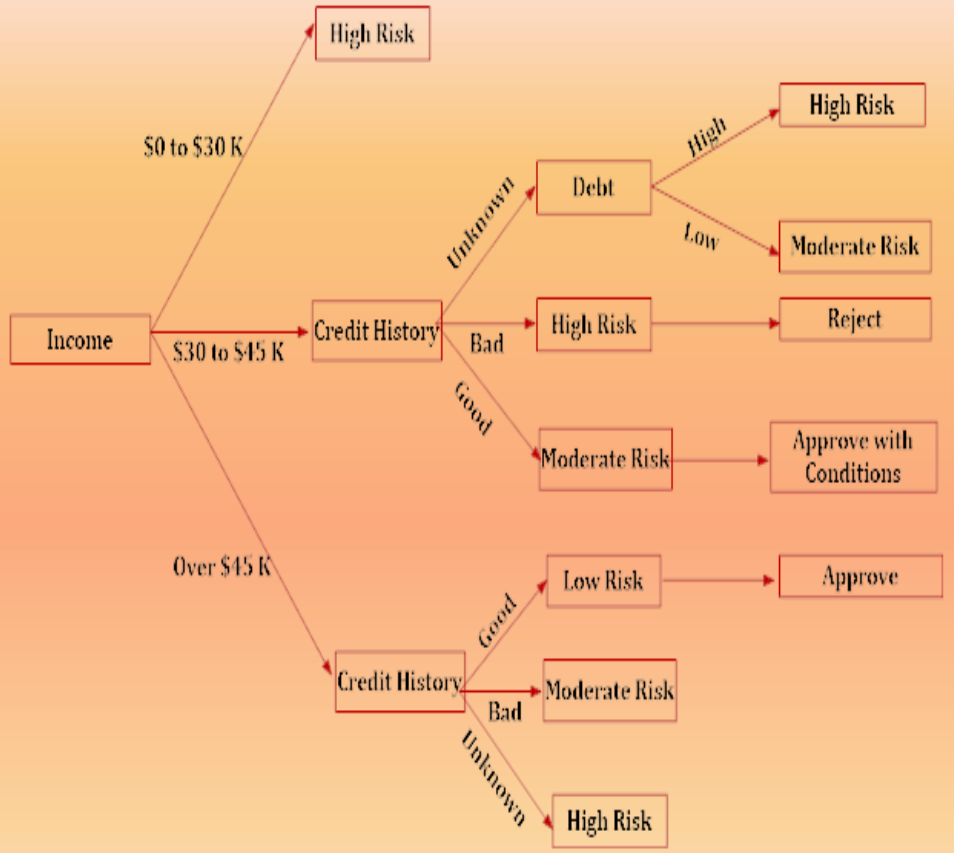


The interrelationship digraph is used to identify the relationships between different issues relating to a problem. This graph may be used as an extension to the affinity diagram and often used in conjunction with the affinity diagram. The diagraph displays all the interrelated cause-and-effect relationships and issues involved in a complex problem in their order of importance. Identifying the most important issues relating to a problem helps diverting the efforts in the right direction in seeking the solution.

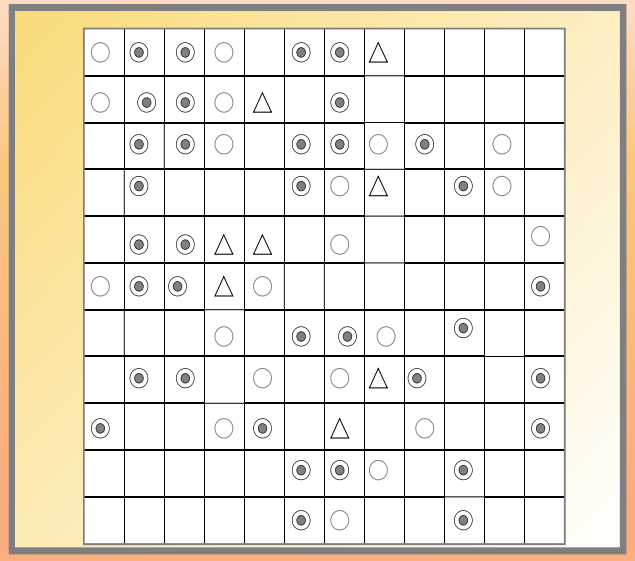
Tree Diagrams

- The tree diagram is used to break down broad categories into different levels of detail.
- It starts with one item or issue that branches into two or more forming new branches that can be viewed as the next level.
- Each of the new level branches into two or more, and so on.
- The tree diagram helps to break down and get to the details of a general idea.

Decision Tree for a Loan Application Process



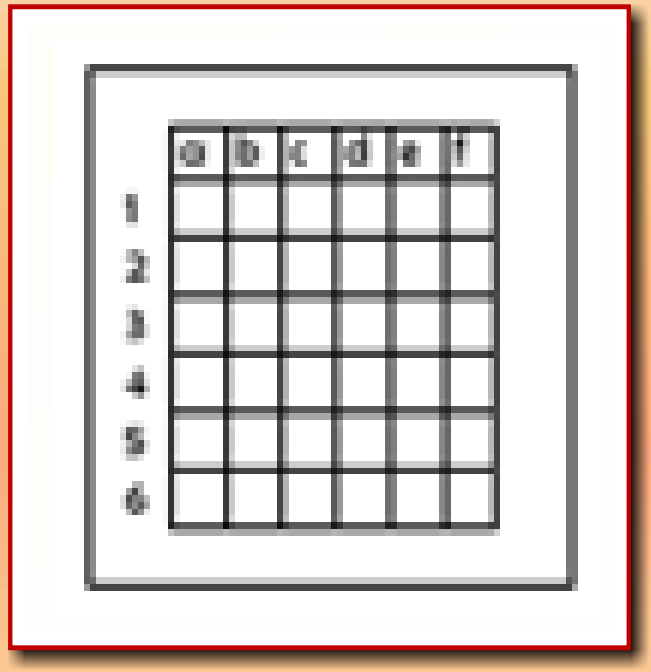
Matrix Diagram



The matrix diagram is a tool used to identify, analyze, and rate the relationship among two or more variables. It defines the relationship between pairs of variables by identifying the importance rating between the variables or factors within the lists. The factors with higher ratings or a higher relationship are given high priority.

Prioritizing Matrix

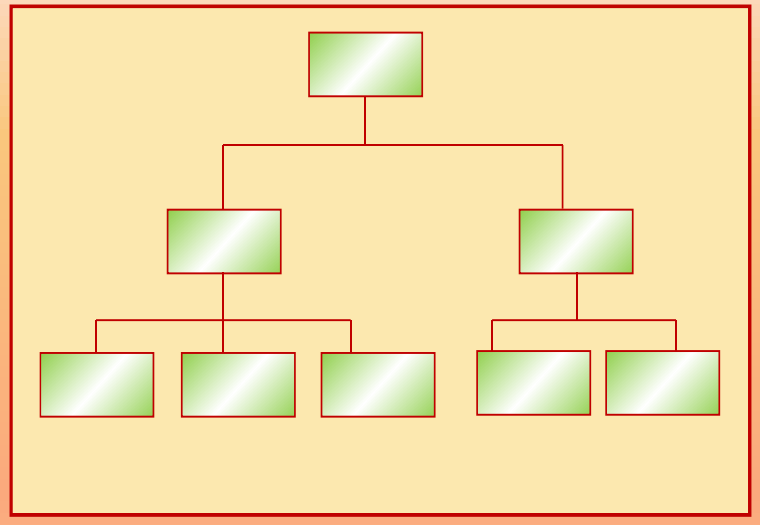
- A prioritization grid is used to make decisions involving multiple criteria and multiple alternatives.
- This tool prioritizes issues based on weights by performing a pair-wise evaluation based on a weighted scale.
- The problem involving prioritization grid usually has several options or alternatives that need to be compared and several criteria that need to be evaluated.



Process Decision Program Chart (PDPC)

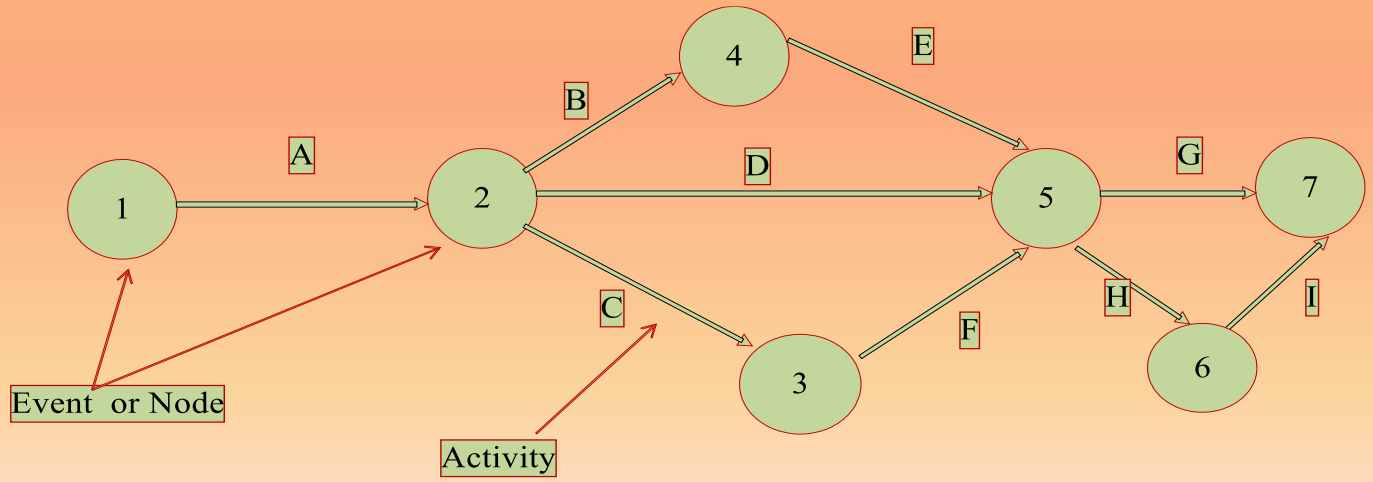
- **The Process Decision Program Chart (PDPC) is a tool for contingency planning. This tool can be used to implement a plan or a program or improve a program.**
- **It helps to brainstorm possible problems and identify what could go wrong with the implementation of that program.**
- **It breaks down the tasks into a hierarchy using a tree diagram and extends the tree diagram to different levels to identify risks and countermeasures for the tasks at the bottom level.**

General Structure



Activity Network Diagram

- The Activity Network Diagram is also known as the PERT (Program Evaluation and Review Technique) network, and CPM - the Critical Path Method (CPM) network.
- The CPM and PERT networks are widely used in planning, scheduling, and controlling projects.
- These network diagrams are used to evaluate the time it takes from the beginning of a project to the end and determine the critical path – the project completion time.



Section 7:
Data Visualization with Big Data

Overview:

- Data visualization is a form of visual communication that presents the data in graphical form.
- It involves creating charts, graphs, and other visual tools that also include flow charts to communicate the information in the data effectively

- One of the major objectives of data visualization is to reveal the essential characteristics of data that may not be apparent otherwise.
- ***Data visualization*** makes complex and large data understandable.
- ***Visual analytics*** is an added feature in data visualization. Several data visualization software are equipped with interactive visualization that aid in the processing and analysis of data by drilling down into the graphical displays.
- A number of charts and graphs can be created from the same data set that help to visualize different features of the data and at the same time can answer "what if" questions.
- ***Visual analytics*** can also be used interactively with continuously changing data to identify new patters.

Big Data:

- ***Big Data*** refers to data set that is massive and range from terabytes to many petabytes.
- The processing and analysis of such data is beyond the capability of traditional software.
- Recently, specialized software and computer application have been designed to store, analyze, visualize, and share large amounts of data.
- ***Big data*** is different from the conventional data in many aspects.
- Often, a ***"3Vs"*** model is used to distinguish big data from the conventional data. The 3Vs refer to **volume**, **variety** and **velocity**.

Big Data…cont.

- **According to Gartner (2012): "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."**
- **The 3Vs are further expanded to include the following characteristics: the volume refers to the huge volume of data.**
- **It is important to note that big data does not use sampling as in statistical analysis; it looks into the entire data set to visualize and observes what happens.**
- **The big data may be available in real time that may continuously change (velocity). These data are drawn from text, images, audio, videos, etc. (variety).**
- **It is important to note that besides descriptive statistics, the applications of big data involve inductive statistics that uses inferential statistics and predictive modeling tools to make predictions of future behavior of the variables.**

Characteristics of Big Data

According to **Martin** big data possess the following characteristics:
[*"Big Data for Development: A Review of Promises and Challenges. Development Policy Review."*. *martinhilbert.net. Retrieved 2015-10-07]*

Volume

The volume refers to the quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety

This is the type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity

The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development

Variability

Inconsistency of the data set can hamper processes to handle and manage it.

Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

Current State of Big Data Software

- Due to the increasing need for storing, processing, and analyzing big data, there has been a significant increase in the software applications in this area.
- Software firms including Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP, Dell, Tableau Software, Dundas BI, Google and others have spent more than $15 billion on software firms specializing in data management and analytics.
- In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.[2]

Big Data Application Areas

- Advancements in data visualization, visual analytics, business analytics, and big data analysis are helping to detect the areas where improvement efforts can be directed.

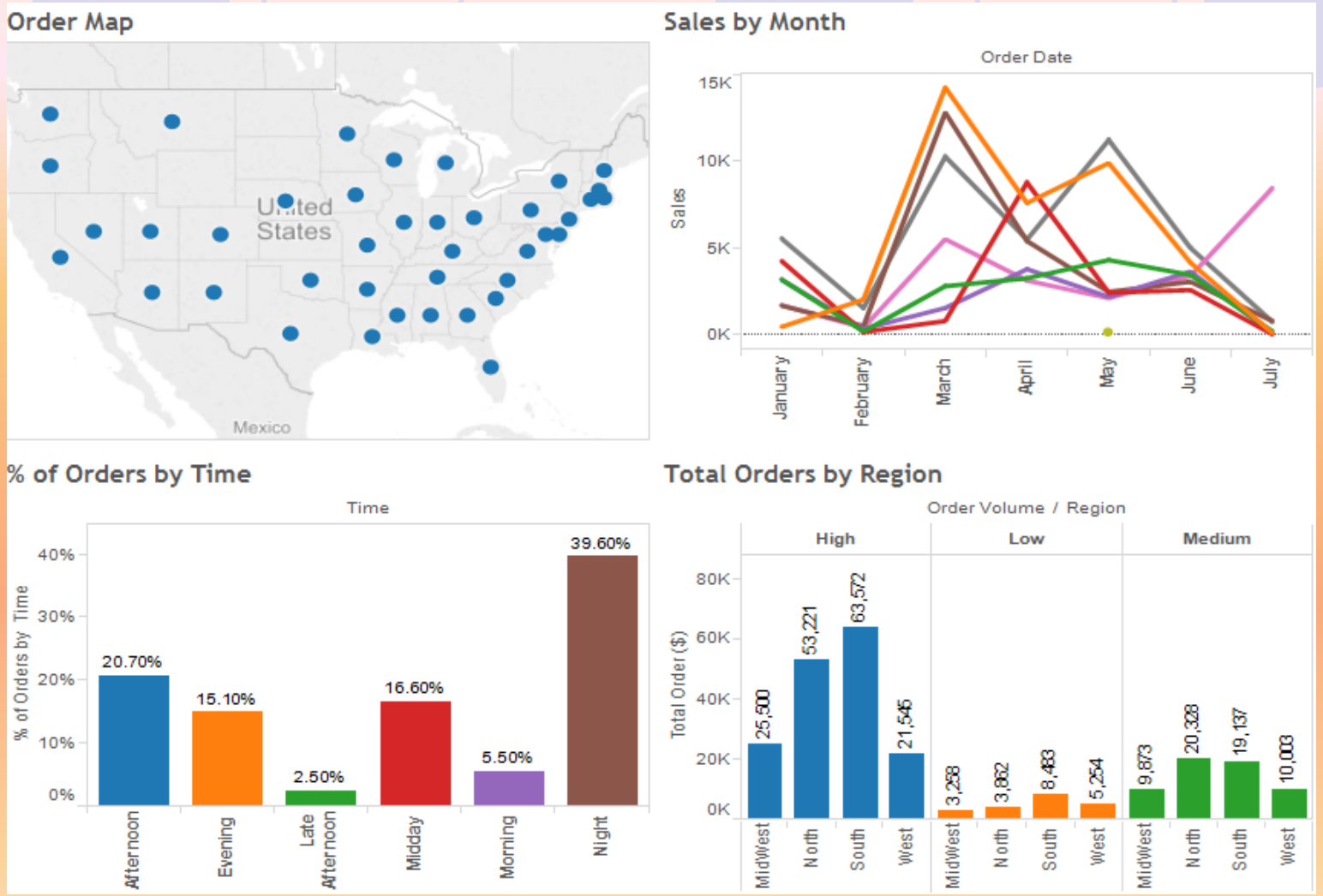
- Big Data and Data Analytics are also helping to improve decision-making in critical areas, such as health care, employment, economic out look and development/productivity, crime, security, education, natural disaster, and resource management.[61][62][63] to name a few.

- However, they come with a number of challenges. These include inadequate technological infrastructure, privacy, appropriate methodology, and data preparation and management .[61]

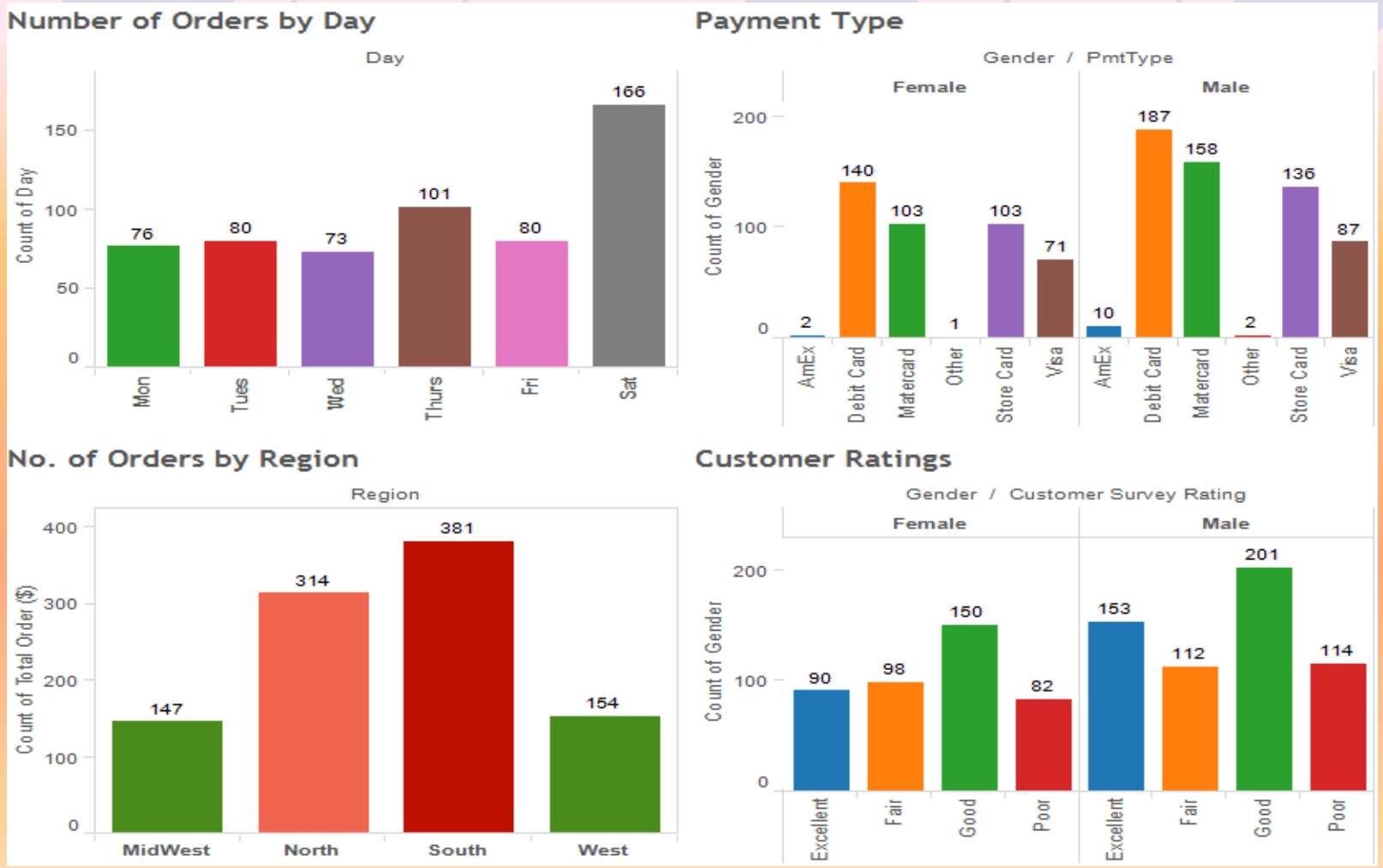
Some Big Data Software Application

- Data Visualization and business intelligence software can create interactive and customizable dashboards that allow one to measure performance in real time.
- These software provide drag and drop design where data can be easily added and removed; visualization can be added and arranged in different ways. The built-in design also allows performing simple to advanced calculations.
- With ever increasing demand for managing and handling big data, the data visualization and business intelligence software are critical in visualizing and analyzing data from across the organization to gain valuable insight and to make accurate and timely decisions.
- Figures on next two slides are examples of simple dashboards showing different scenarios of the same data collected from an on-line ordering process of a business. These dashboards are very helpful in simultaneously visualizing key business performance that can be used in effective and timely decision making.

An Example of a Dashboard using Big Data Software Tableau (1)



Example of a Dashboard using Big Data Software Tableau (2)



We provided a comprehensive overview of Data Visualization and Quality tools, and techniques with applications. Data Visualization and business intelligence software are used to create interactive and customizable dashboards that allow a company to study the current state, measure, and predict performance in real time and in future .