GUC
German University in Cairo

Spring 2019

CSEN1083: Data Mining

Data Visualization

Seif Eldawlatly

Statistical Analysis

- Textbook for this Lecture: “Applied Statistics and Probability for Engineers,” by Douglas C. Montgomery and George C. Runger, 6th edition, 2014

- Statistical Data Analysis includes 3 topics:

  - Basic Probability

  - Data Collection and Description

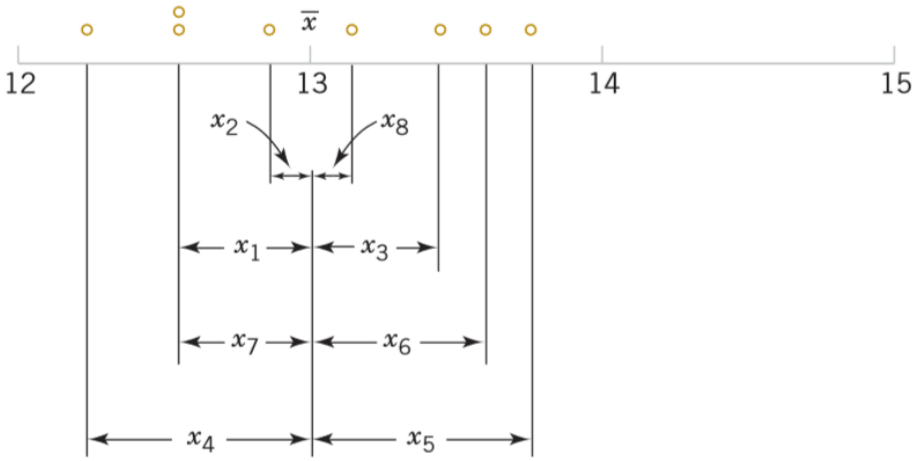
  - Data Presentation:
    - Stem-and-Leaf Diagrams
    - Histograms
    - Box Plots
    - Time Series plots
    - Multivariate Data

  - Inference and Hypothesis Testing

Stem-and-Leaf Diagrams

- Dot Diagram: Represent each sample with a dot
- Example: Find the dot diagram for the variable x

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	–0.4	0.16
2	12.9	–0.1	0.01
3	13.4	0.4	0.16
4	12.3	–0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	–0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60



Stem-and-Leaf Diagrams

- The dot diagram is a useful data display for small samples up to about 20 observations

- However, when the number of observations is large, other graphical displays may be more useful such as Stem-and-Leaf Diagrams

- To construct Stem-and-Leaf Diagram:

  (1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
  (2) List the stem values in a vertical column.
  (3) Record the leaf for each observation beside its stem.
  (4) Write the units for stems and leaves on the display.

Stem-and-Leaf Diagrams

- Example: Construct the Stem-and-Leaf diagram for the data given in the table below

TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Histograms

- The histogram is a visual display of the frequency distribution constructed as follows:

  (1) Label the bin (class interval) boundaries on a horizontal scale.
  (2) Mark and label the vertical scale with the frequencies or the relative frequencies.
  (3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

- Example:

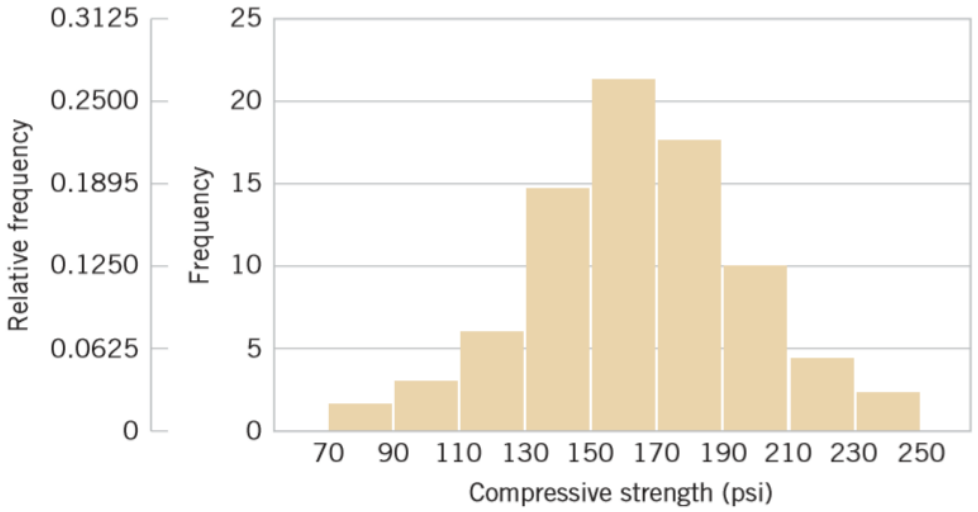
TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

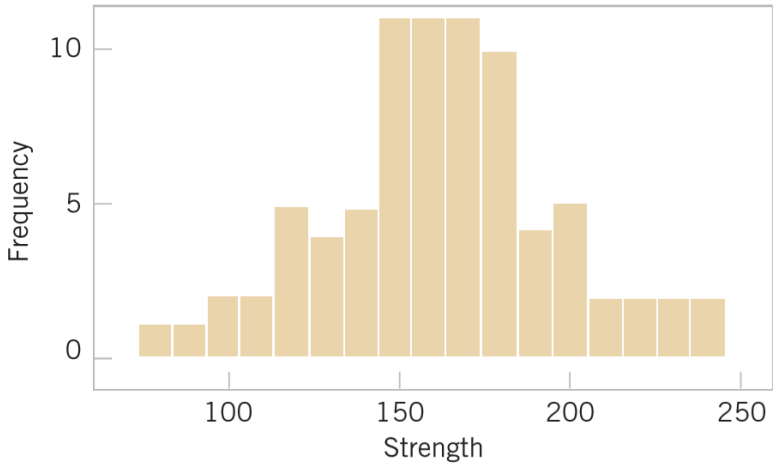
Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Histograms

- A histogram with 9 bins



- A histogram of the same data with 17 bins



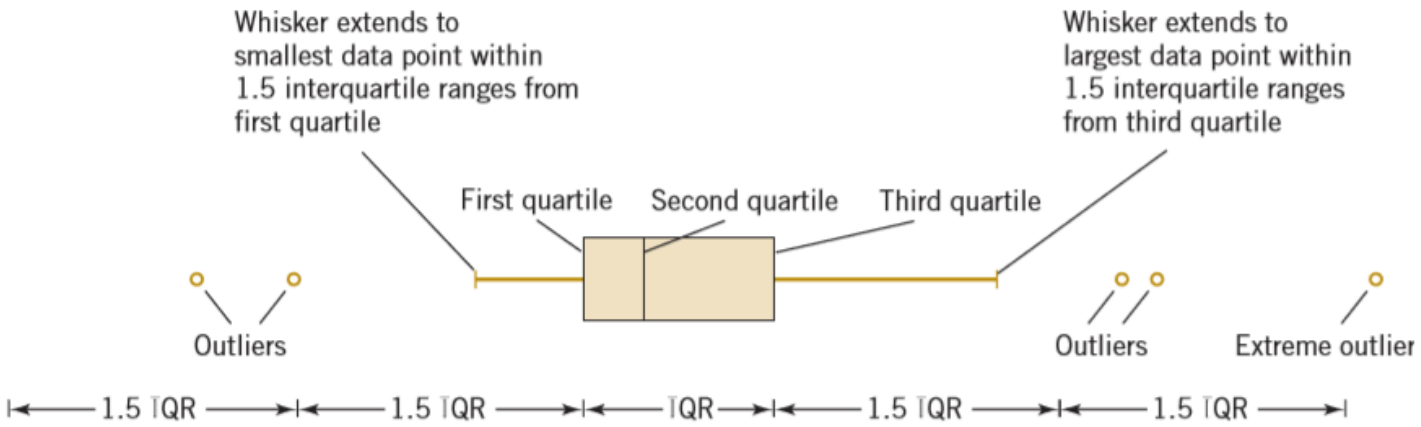
Box Plots

- The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of outliers

- First, we define the concept of dividing the data into quartiles. When an ordered set of data is divided into four equal parts, the division points are called quartiles:
  - The first or lower quartile, q_1, is a value that has approximately 25% of the observations below it and approximately 75% of the observations above
  - The second quartile, q_2, has approximately 50% of the observations below its value. The second quartile is exactly equal to the median
  - The third or upper quartile, q_3, has approximately 75% of the observations below its value

Box Plots

- A sample box plot:



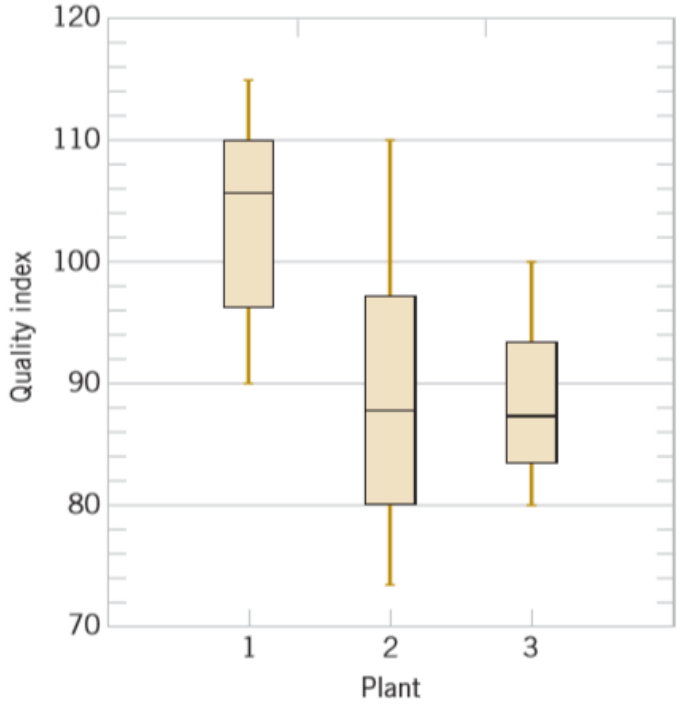
- A point beyond a whisker, but less than three interquartile ranges from the box edge, is called an outlier

- A point more than three interquartile ranges from the box edge is called an extreme outlier

Box Plots

- Box plots are very useful in graphical comparisons among data sets because they have high visual impact and are easy to understand

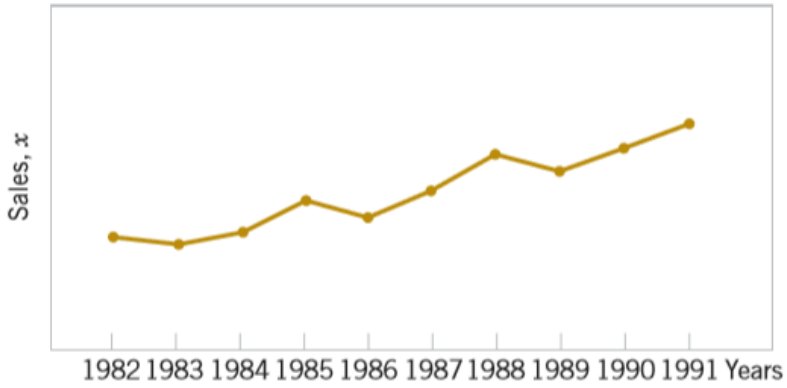
- Example:

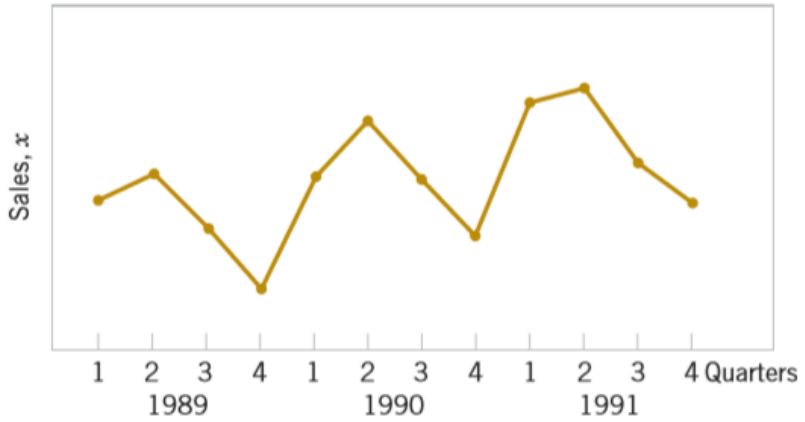


Time Sequence Plots

- A time series or time sequence is a data set in which the observations are recorded in the order in which they occur

- Examples:





Scatter Diagrams

- A scatter diagram is constructed by plotting each pair of observations with one measurement in the pair on the vertical axis of the graph and the other measurement in the pair on the horizontal axis

- Examples:

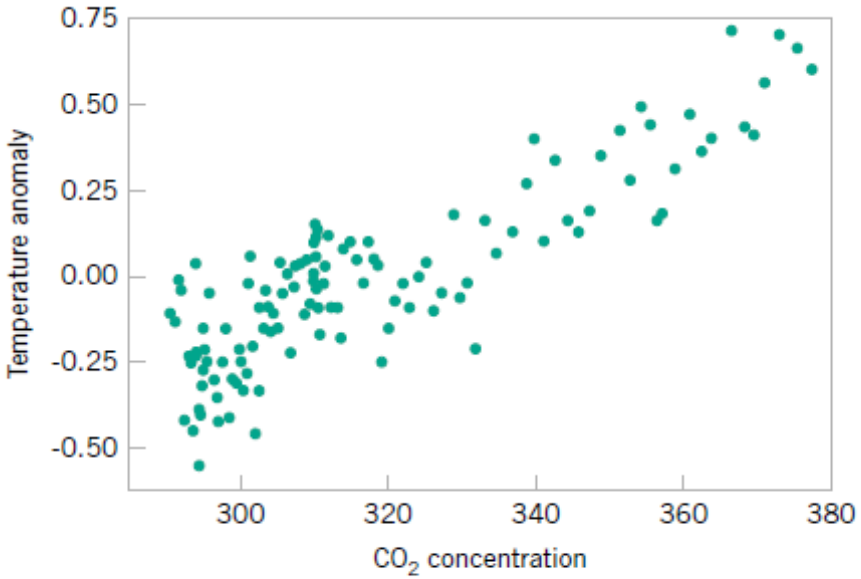


Figure 2-21 Scatter diagram of global mean air temperature anomaly versus global CO_2 concentration.

Scatter Diagrams

- Relationships between the variables can be determined from the scatter diagrams