MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY (MUST), MIRPUR

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DATA MINING

BCS-3605

Lecture 08

Dr Yasir Mehmood

(Assistant Professor)

Agenda of Today's Lecture

- ***Graphic Displays of Basic Statistical Descriptions***

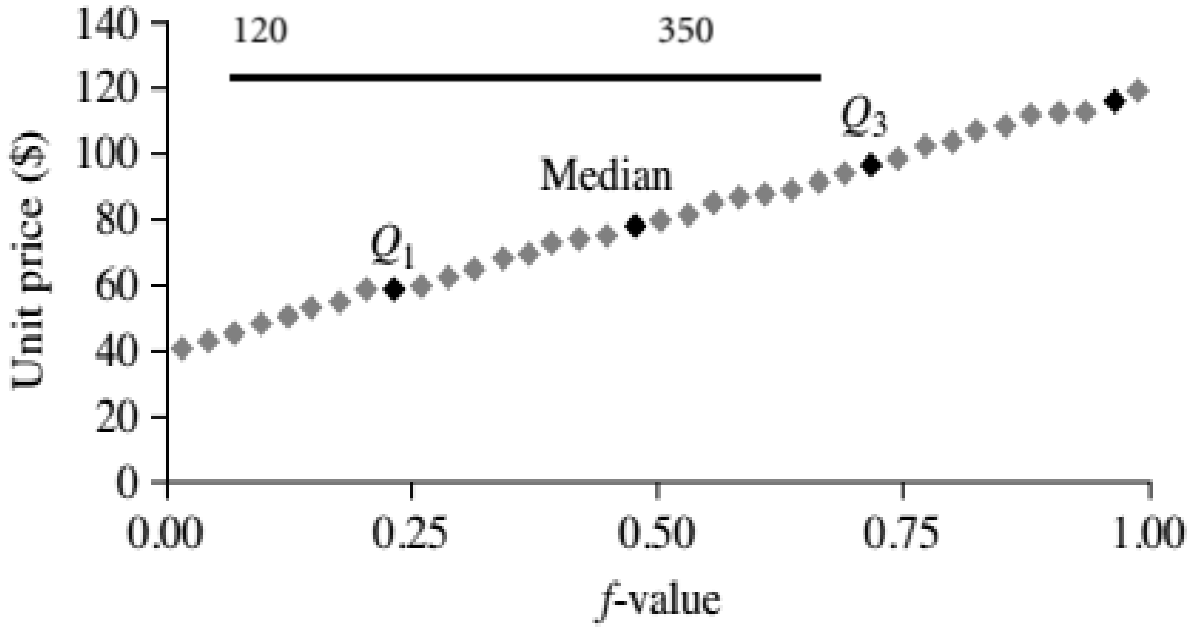
Graphic Displays of Basic Statistical Descriptions

- Visual inspection of data for preprocessing
- **Boxplot**: graphic display of five-number summary
- **Quantile plot**: each value x_i is paired with % f_i indicating that approximately f_i 100% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Histogram**: x-axis are values, y-axis represents frequencies
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Quantile Plot

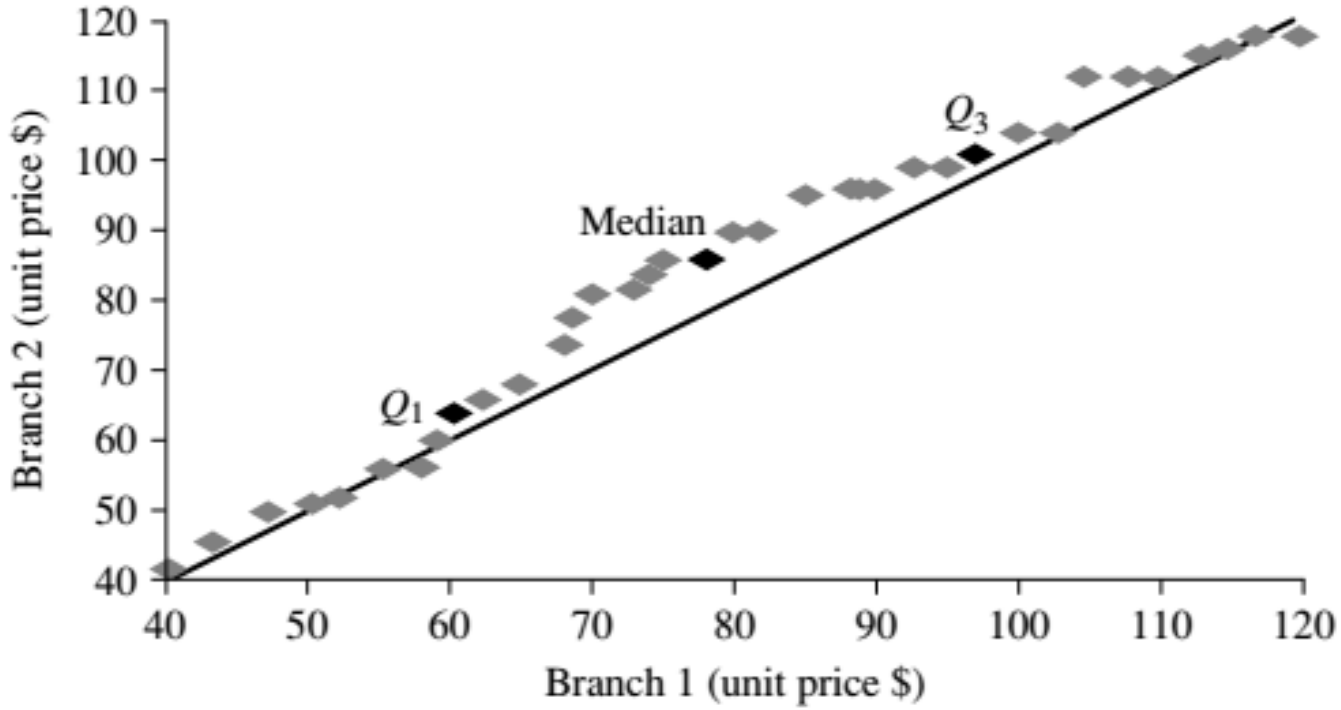
- Common graphic displays of univariate data distributions
- First, it displays all of the data for the given attributes
- Second, it plots **quantile** information
  - Let *xi*, for $i = 1$ to N, be the data sorted in increasing order so that $x1$ is the smallest observation and xN is the largest for some ordinal or numeric attribute X.
    Each observation, *xi*, is paired with a percentage, *fi*, which indicates that approximately $fi \times 100\%$ of the data are below the value, *xi*

Unit price ($)	Count of items sold
40	275
43	300
47	250
–	–
74	360
75	515
78	540
–	–
115	320
117	270
120	350



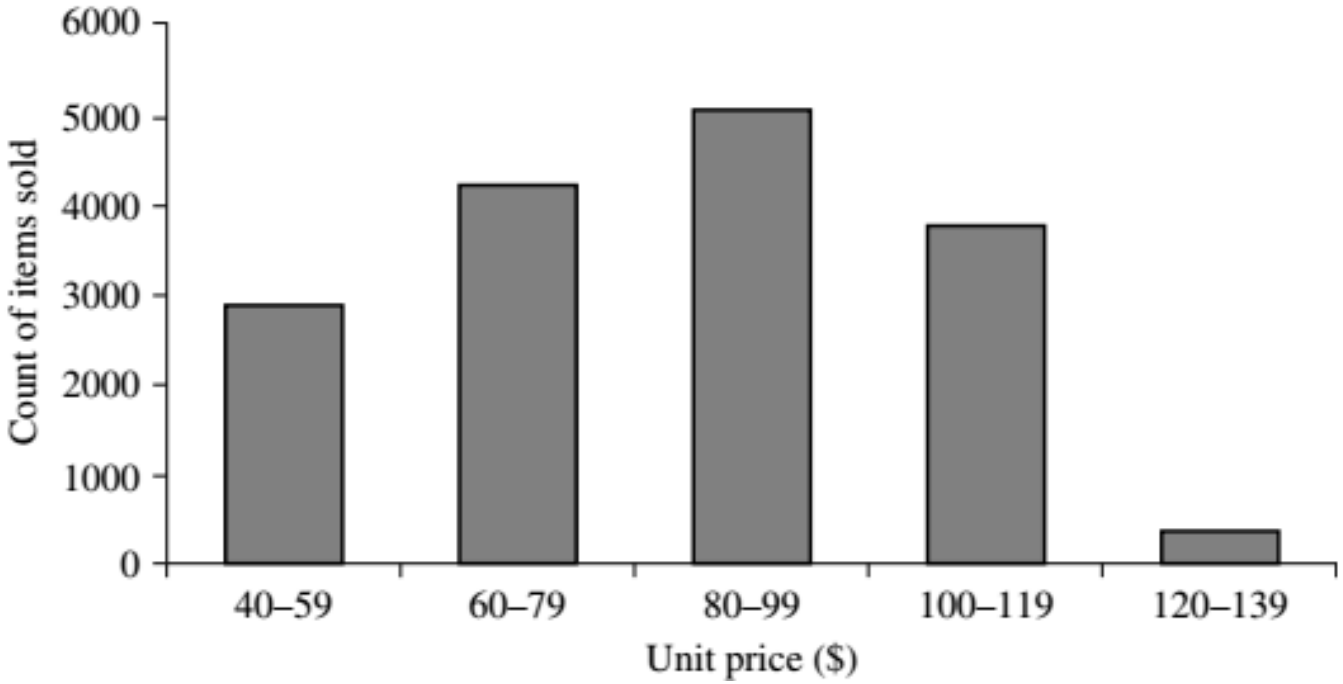
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Visualize whether there is a shift from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those Branch 2.

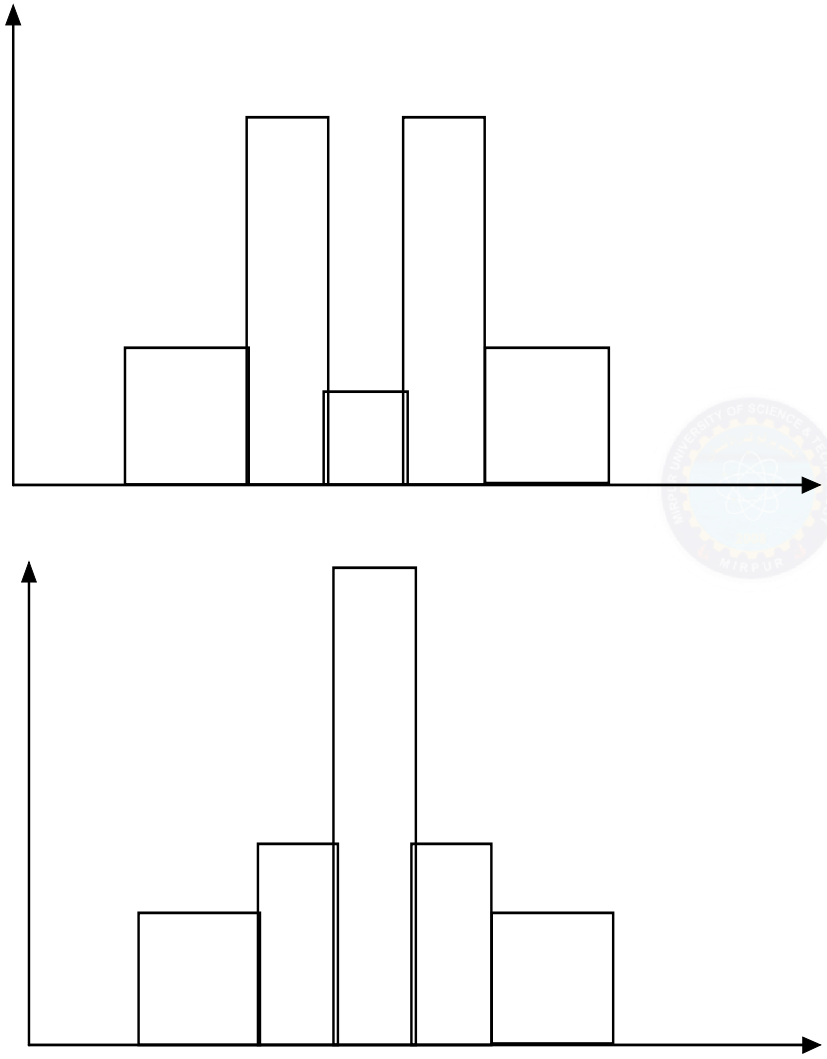


Histogram Analysis

- Histogram: Chart of plots
- Summarize the distribution of attribute values
- Height indicates the frequencies
- Bucket or bin represents the sub range of values on numeric attributes
- It shows what proportion of cases fall into each of several categories



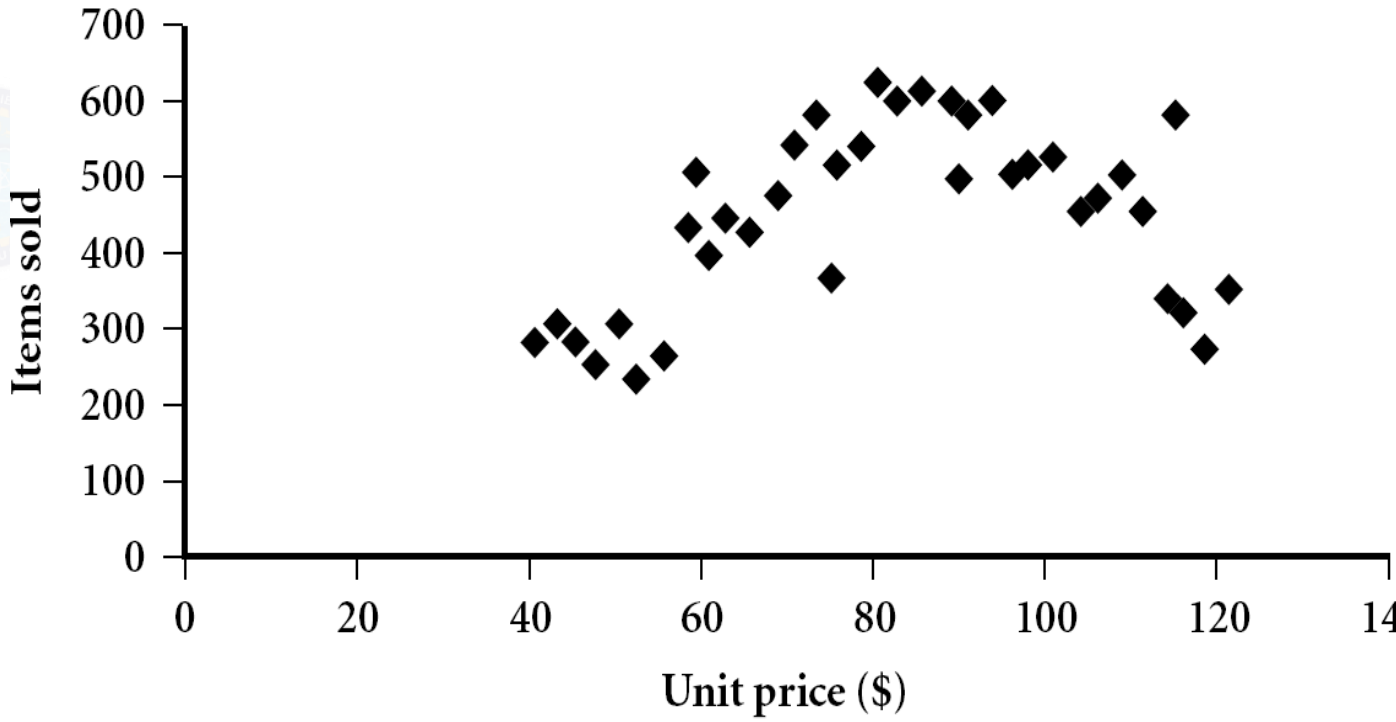
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

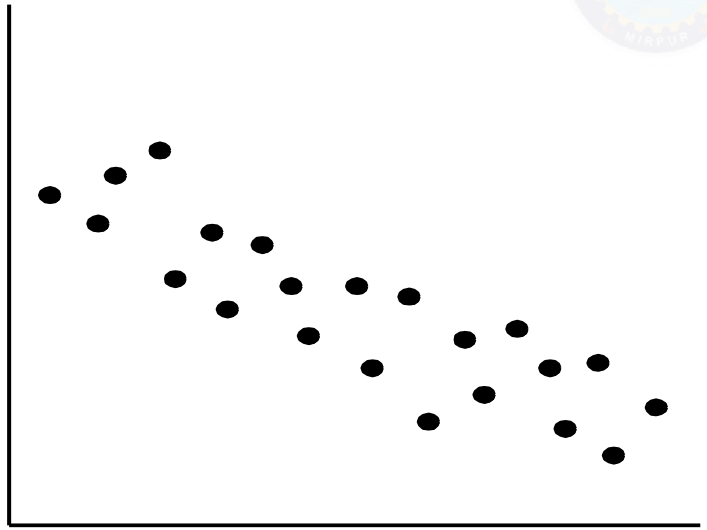
Scatter plot

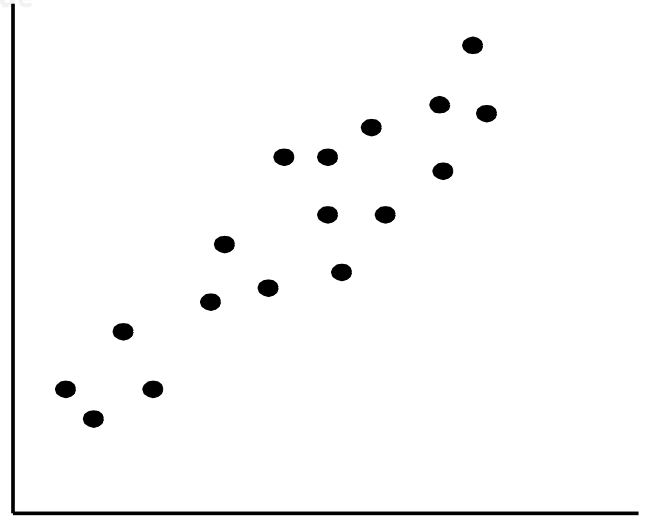
- It determines relationship, pattern, and trend between two attributes
- Provides a first look at bivariate data to see clusters of points, outliers, correlationship
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane
- Correlations can be positive, negative, or null (uncorrelated



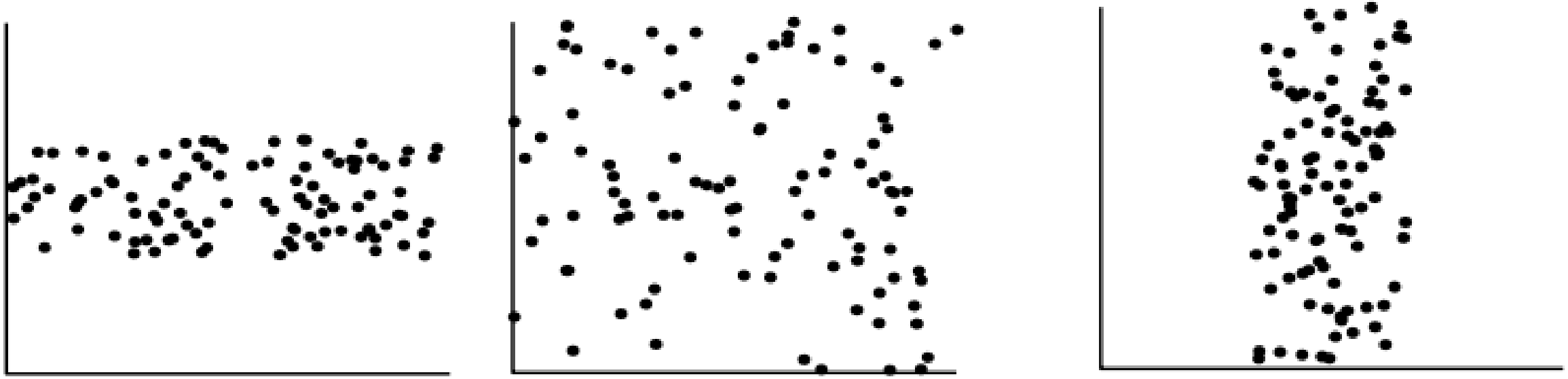
Positively and Negatively Correlated Data

- If the plotted points pattern slopes from lower left to upper right, this means that the values of *X* increase as the values of *Y* increase, suggesting a *positive correlation*
- If the pattern of plotted points slopes from upper left to lower right, the values of *X* increase as the values of *Y* decrease, suggesting a *negative correlation*





Un Correlated Data



Reference

Data Mining Concepts and Techniques Third Edition

2.2.3 Graphic Displays of Basic Statistical Descriptions of Data

THANKS