# Visualization (EDA)

# Graphic Displays of Basic Statistical Descriptions of Data

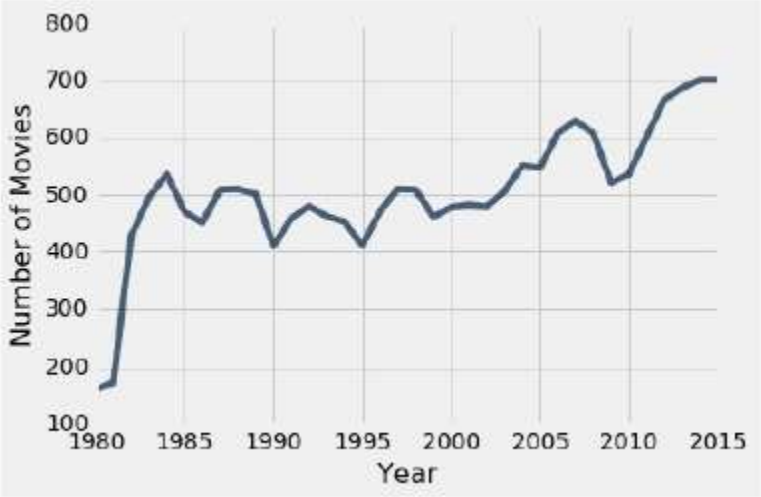
- Graphic displays of basic statistical descriptions. These include
  - *Line Plots*
  - *scatter plots.*
  - *Histograms,*
  - *quantile plots,*
  - *quantile–quantile plots,*

- *Helpful* for the visual inspection of data, which is useful for data preprocessing.
  - scatter plots show bivariate distributions (i.e., involving two attributes).

- A **quantile plot is a simple and effective way to have a first look at a univariate** data distribution.
  - First, it displays all of the data for the given attribute (allowing the user to assess both the overall behavior and unusual occurrences).
  - Second, it plots quantile information.

# Line Plots

- Line graphs are among the most common visualizations and are often used to study chronological trends and patterns.

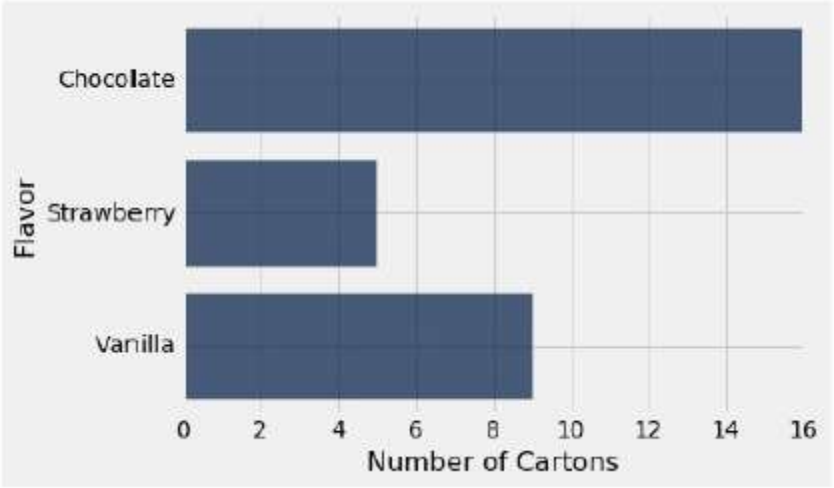


# Visualizing Categorical Distributions

- Data come in many forms that are not numerical.
- Data can be pieces of music, or places on a map.
- They can also be categories into which you can place individuals. Here are some examples of *categorical variables*.
  - The individuals are cartons of ice-cream, and the variable is the flavor in the carton.
  - The individuals are professional basketball players, and the variable is the player's team.
  - The individuals are years, and variable is movies of the year.

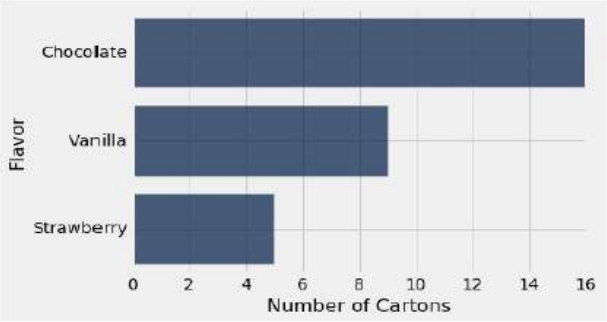
# Bar Chart

- The bar chart is a familiar way of visualizing categorical distributions. It displays a bar for each category.
- The bars are equally spaced and equally wide. The length of each bar is proportional to the frequency of the corresponding category.
- We draw bar charts with horizontal bars because it's easier to label the bars that way.



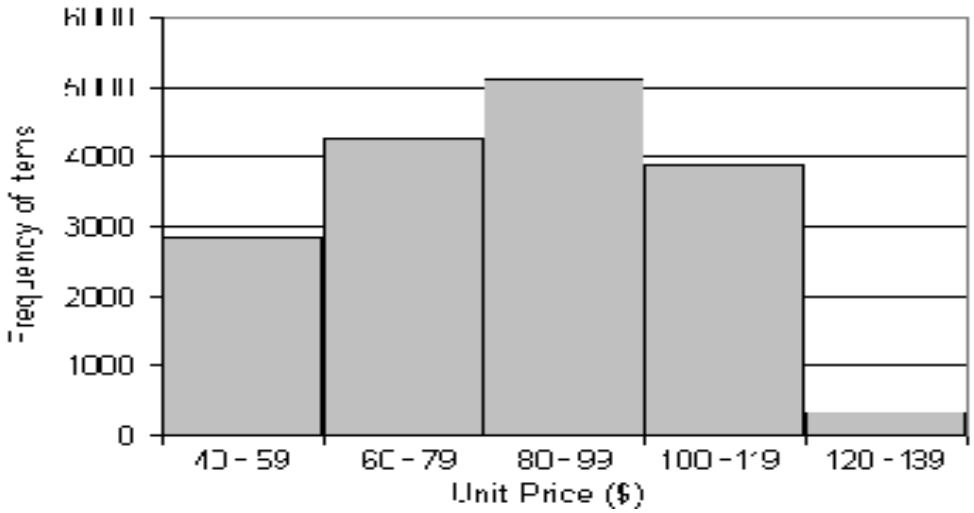
# Bar Chart

- Fundamental distinction between bar charts and the scatter plot and the line plot
  - Scatter/ Line plots display two numerical variables – the variables on both axes are numerical.
- In contrast, the bar chart has categories on one axis and numerical frequencies on the other.
- The width of each bar and the space between consecutive bars is entirely up to the person who is producing the graph
- Most importantly, the bars can be drawn in any order. The categories "chocolate," "vanilla," and "strawberry" have no universal rank order,
- This means that we can draw a bar chart that is easier to interpret, by rearranging the bars in decreasing order.



# Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data

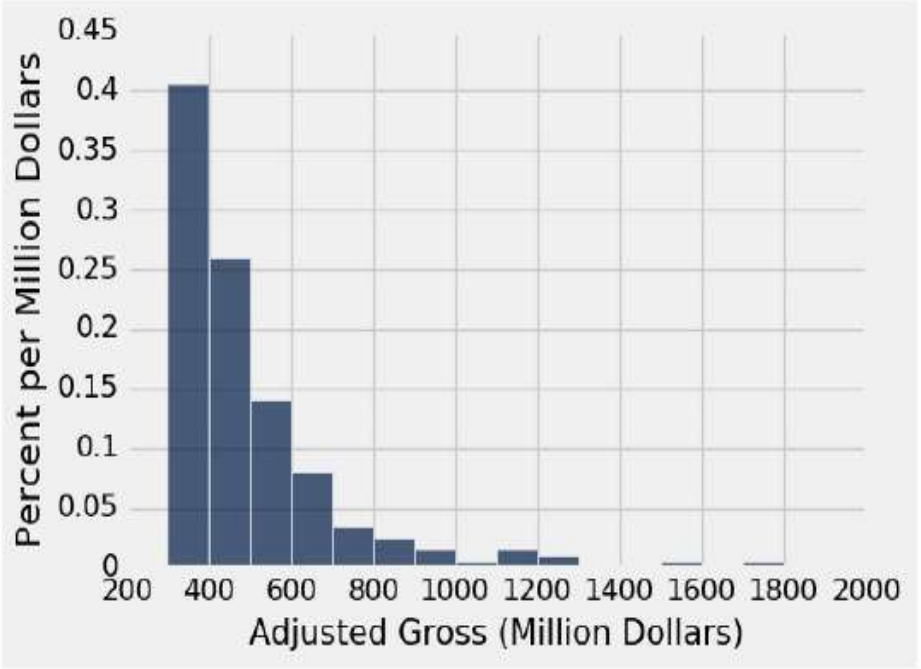



  - A relative frequency histogram uses the same information as a frequency histogram but compares each class interval to the total number of items

# Histogram Analysis

- There are 200 movies in the dataset. The [300, 400) bin contains 81 movies.That is 40.5% of all the movies:

| bin | Count | Percent | Height |
|---|---|---|---|
| 300 | 81 | 40.5 | 0.405 |
| 400 | 52 | 26 | 0.26 |
| 500 | 28 | 14 | 0.14 |
| 600 | 16 | 8 | 0.08 |
| 700 | 7 | 3.5 | 0.035 |
| 800 | 5 | 2.5 | 0.025 |
| 900 | 3 | 1.5 | 0.015 |
| 1000 | 1 | 0.5 | 0.005 |
| 1100 | 3 | 1.5 | 0.015 |
| 1200 | 2 | 1 | 0.01 |




$$\text{Percent} = \frac{81}{200} \cdot 100 = 40.5$$

The width of the [300, 400) bin is $400 - 300 = 100$. So

$$\text{Height} = \frac{40.5}{100} = 0.405$$

# Differences Between Bar Charts and Histograms

- Bar charts display one quantity per category. They are often used to display the distributions of categorical variables.
- Histograms display the distributions of quantitative variables.
- All the bars in a bar chart have the same width, and there is an equal amount of space between consecutive bars.
- The bars of a histogram can have different widths, and they are contiguous.
- The lengths (or heights, if the bars are drawn vertically) of the bars in a bar chart are proportional to the value for each category.
- The heights of bars in a histogram measure densities; the *areas of bars in a histogram are proportional to the numbers of entries in* the bins.

# Stem-and-leaf plots

| | |
|---|---|
| 6 | 2355689 |
| 7 | 00111223444456677789 9 |
| 8 | 0123 |

| | |
|---|---|
| 6 | 23 |
| 6 | 55689 |
| 7 | 00111223444 4 |
| 7 | 566777899 |
| 8 | 0123 |
| 8 | |

Stem-and-leaf plots display data in a structured list.

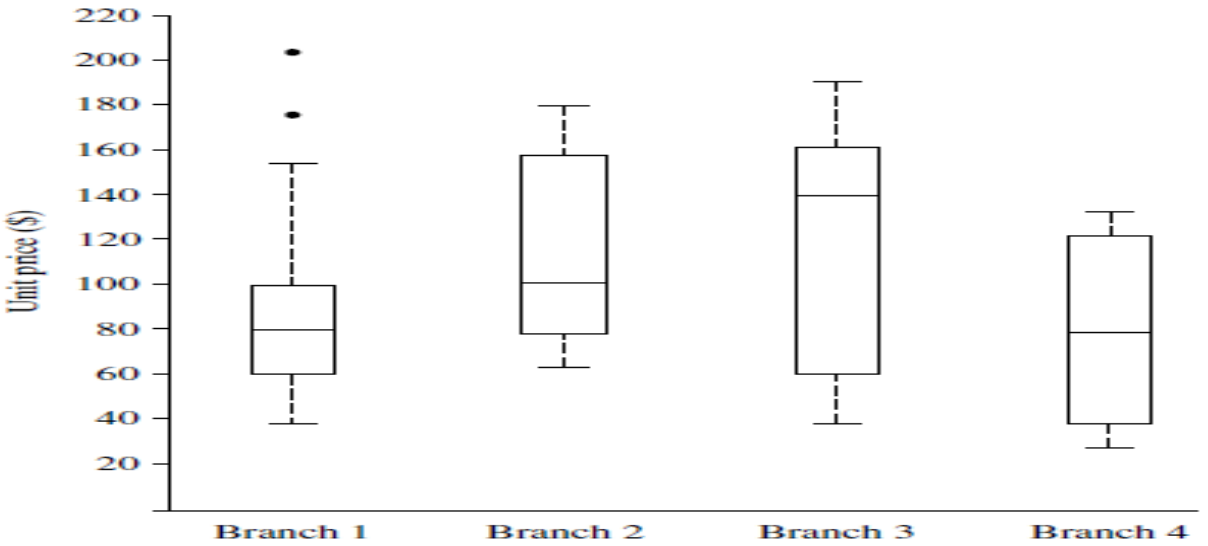
Presenting data in a table or an ordered list does not readily convey information about how the data are distributed, as is the case with histograms.

Eg: Use of Stem chart while grading

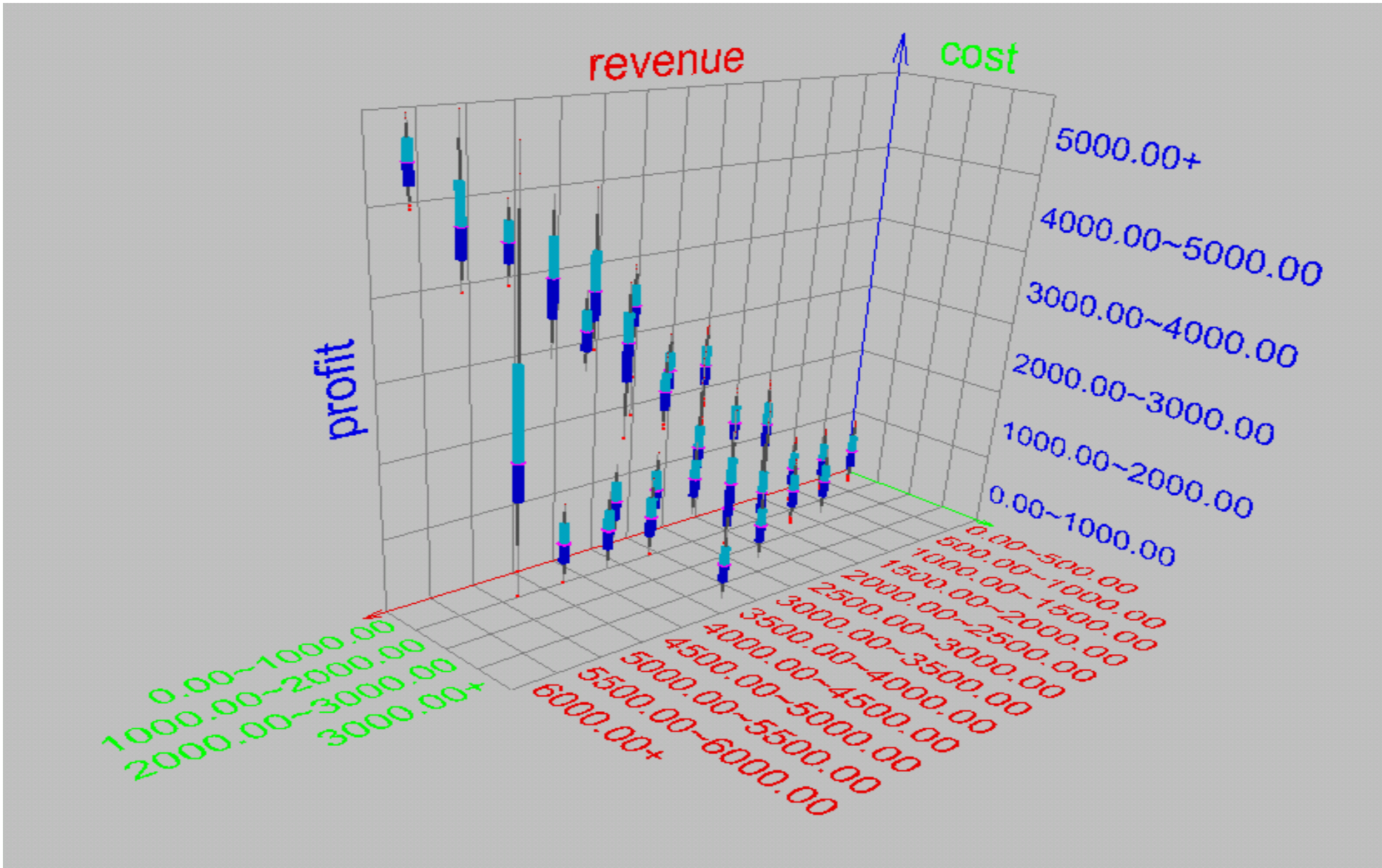
# Measuring the Dispersion of Data **Boxplots, and Outliers**

**Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the** five-number summary as follows:

- The ends of the box are at the quartiles so that the box length is the IQR.
- The median is marked by a line within the box.
- Two lines (called *whiskers) outside the box extend to the smallest (Minimum) and* largest (*Maximum) observations.*
- **For branch 1, we see that the median price of items sold is $80, Q1 is $60, and Q3 is $100. Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.**

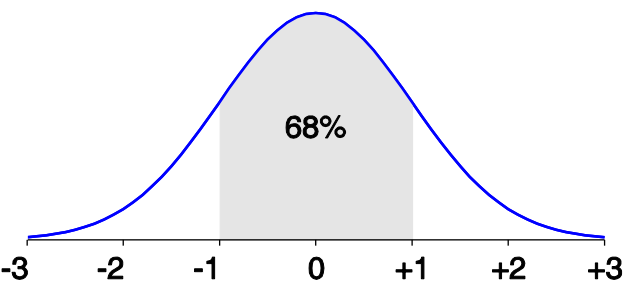


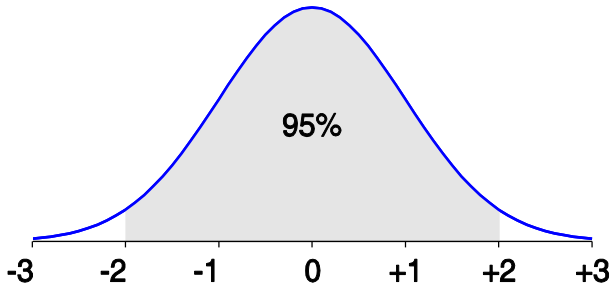
# Visualization of Data Dispersion: Boxplot Analysis

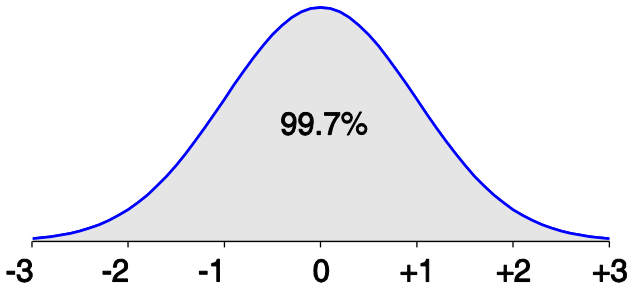


# Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

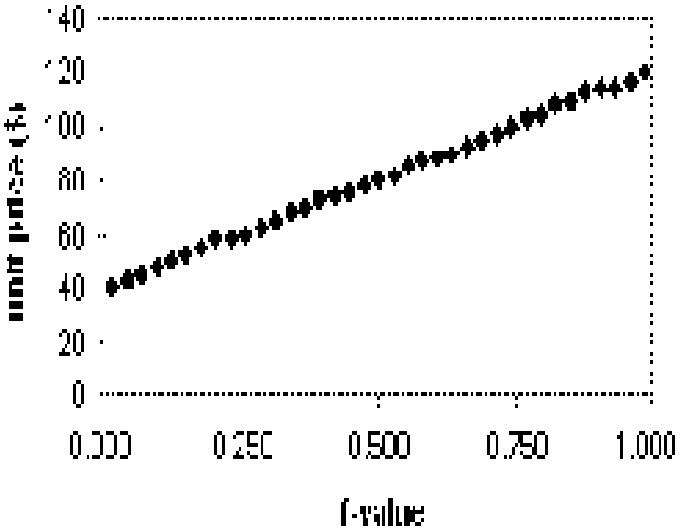


# Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data $x_i$ data sorted in increasing order, $f_i$ indicates that approximately 100 $f_i$% of the data are below or equal to the value $x_i$




| Unit price ($) |
|---|
| 40 |
| 43 |
| 47 |
| — |
| 74 |
| 75 |
| 78 |
| — |
| 115 |
| 117 |
| 120 |

# Quantile Plot

- Simple way -1st look at a univariate data distribution.
- Displays the given attribute to assess both
  - the overall behavior and
  - unusual occurrences.
- It plots quantile information.
  - Let $x_i$ , for i = 1 to N, be the data sorted in increasing order
  - $x_1$ is the smallest observation and
  - $x_N$ is the largest for some attribute X.
- Each observation, $x_i$ , is paired with a percentage, $f_i$ , which => approx $f_i$ 100% of the data are below the value, $x_i$
- Note: 0.25 percentile corresponds to quartile Q1, the 0.50 percentile is the median, and the 0.75 percentile is Q3.

# Quantile Plot

- *Let*

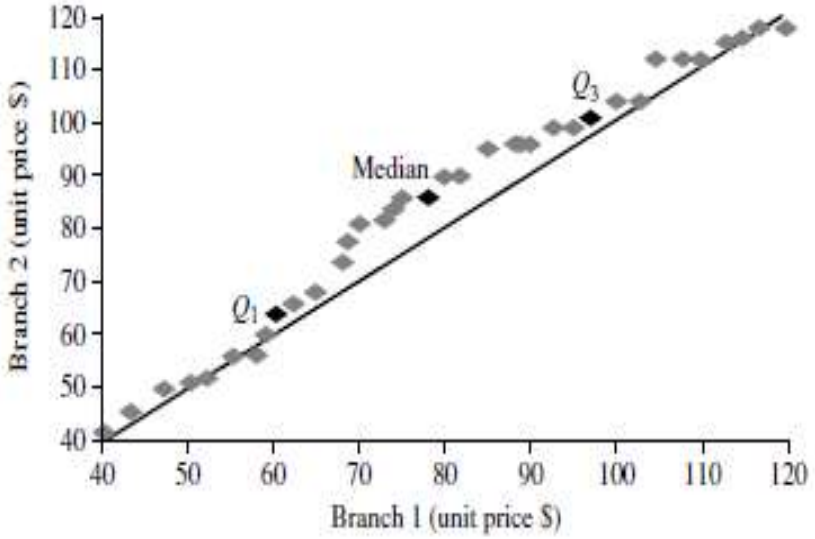
$$f_i = \frac{i - 0.5}{N}.$$

These numbers increase in equal steps of $1/N$, ranging from $\frac{1}{2N}$ (which is slightly above 0) to $1 - \frac{1}{2N}$ (which is slightly below 1). On a quantile plot, $x_i$ is graphed against $f_i$. This allows us to compare different distributions based on their quantiles. For example, given the quantile plots of sales data for two different time periods, we can compare their $Q_1$, median, $Q_3$, and other $f_i$ values at a glance.

# Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution Vs corresponding quantiles of another

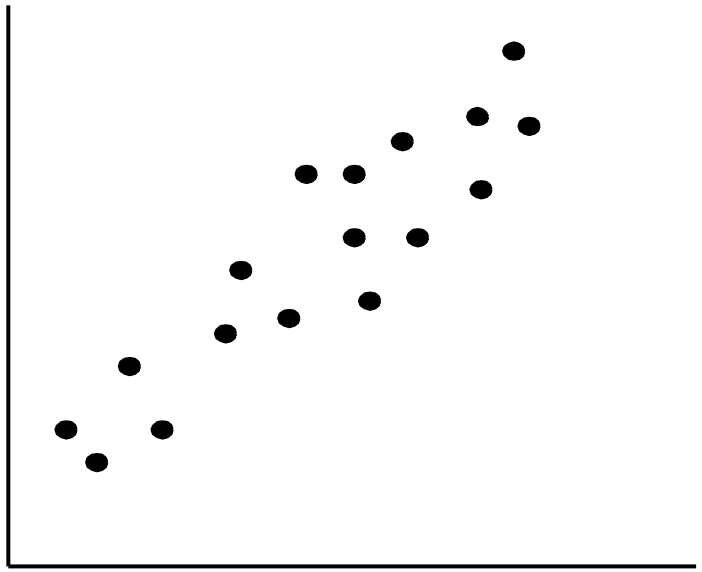
- Allows the user to view whether there is a shift in going from one distribution to another



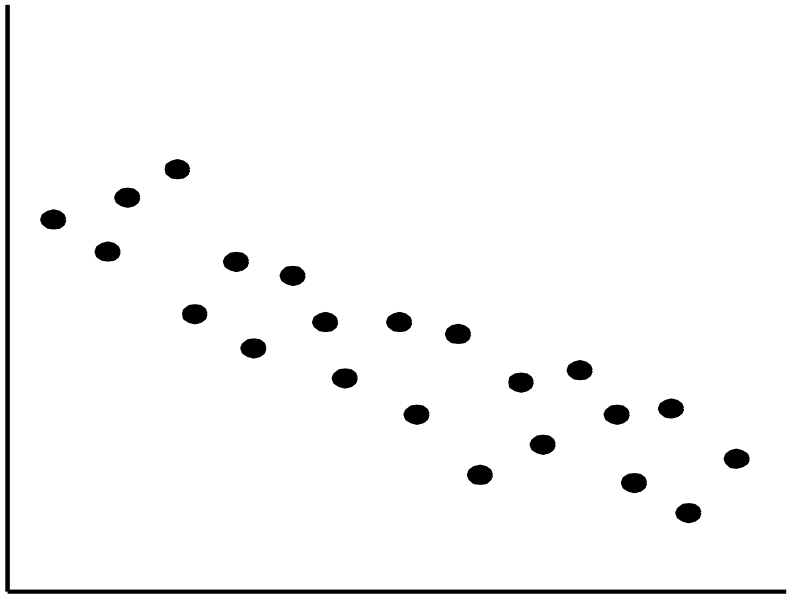
# Quantile-Quantile (Q-Q) Plot

- Each point corresponds to the same quantile for each data set and shows the unit price of items sold at branch 1 Vs 2 for that quantile.
- For comparison, the straight line represents => for each given quantile, the unit price at each branch is the same.
- The darker points - data for *Q1, the median, and Q3.*
- At Q1, the unit price of items sold at branch 1 < at branch 2. ie, 25% of items sold at branch 1 were <= $60, Vs 25% of items at branch 2 <= $64.
- At Q2, the 50th percentile (marked by the median), 50% of items sold at branch 1 <= $75, Vs branch 2 <= $85.
- ***In general, a shift in the distribution of branch 1 Vs 2 in that the unit prices of items sold at branch 1 < at branch 2.***

# Positively and Negatively Correlated Data

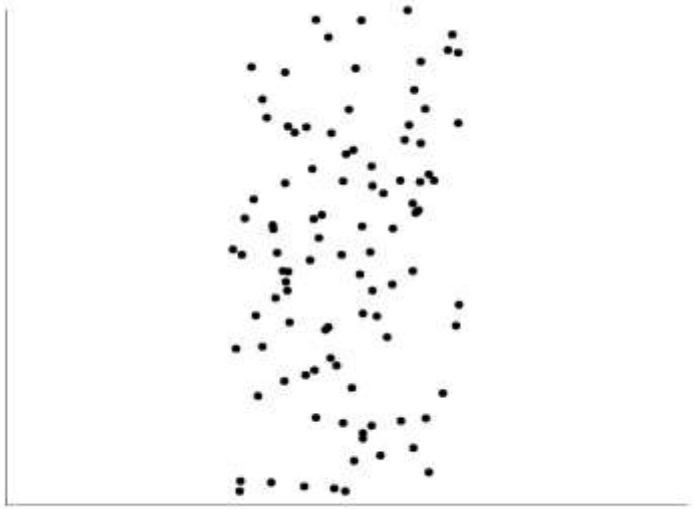


Positively correlated



Negatively correlated

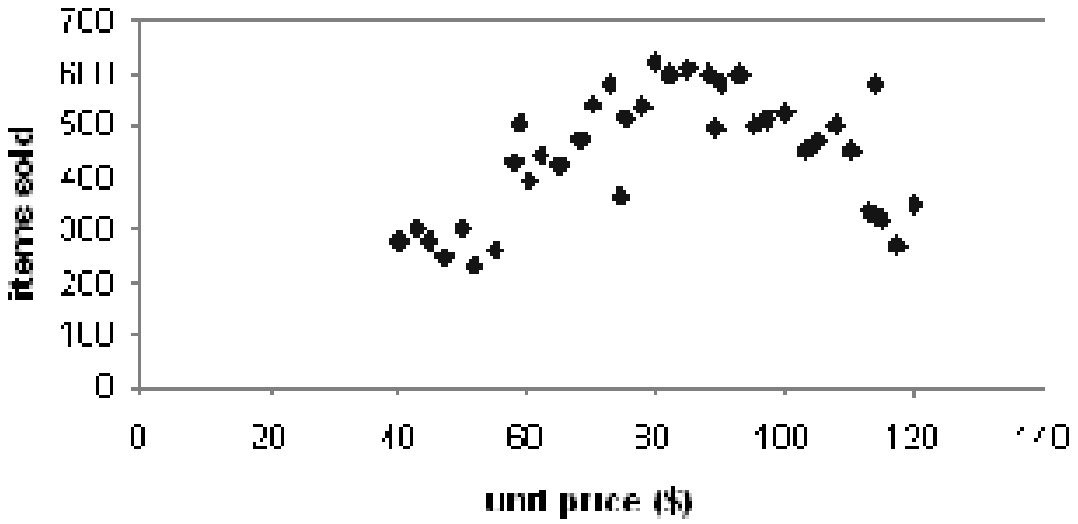
# Not Correlated Data



# Exploring Bivariate Data

# Scatter plot

- A scatter plot - effective graphical methods for determining if there appears to be a relationship/ pattern/ or trend between two numeric attributes.
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



# Scatter plot

- Formally, we say that the plot shows an *association between the variables.*

- The association is

  - Positive: high values of one variable tend to be associated with high values of the other, and

  - Negative: low values of one with low values of the other.

# Surface Plots

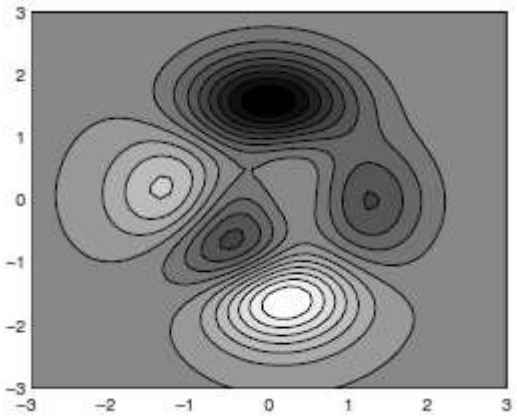
- If data that represents a function defined over a bivariate domain,

$$z = f(x,y)$$

then view values for z as a surface.

# Contour Plots

- Use contour plots to view surface.
- Contour plots show lines of constant surface values, similar to topographical maps



# Summary: Graphic Displays of Basic Statistical Descriptions

- Histogram
- Boxplot
- Quantile plot: each value $x_i$ is paired with $f_i$ indicating that approximately 100 $f_i$% of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

- **Thank You**