Locating Extreme Outliers: Z-Score

$$Z = \frac{X - \overline{X}}{S}$$

For a bell-shape (Normal) distribution,

$|Z| < 1$ for 68% of data

$|Z| < 2$ for 95% of data

$|Z| < 3$ for 99.7% of data

So, values X with large $|Z|$ can be **<u>outliers</u>**.

Locating Extreme Outliers: Z-Score

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

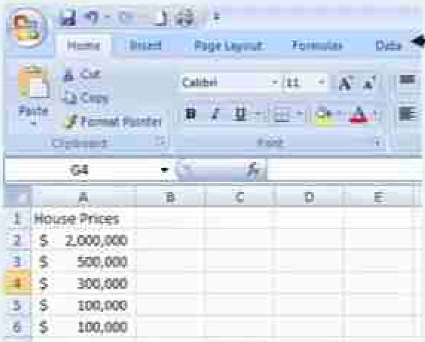
$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

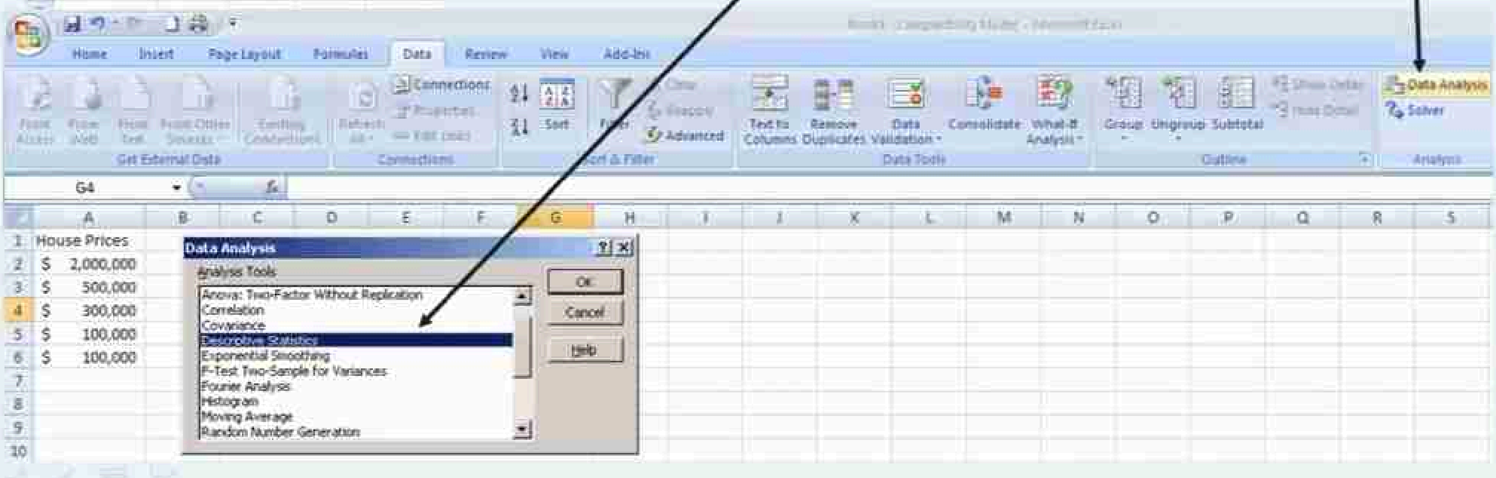
General Descriptive Stats Using Microsoft Excel Functions

House Prices		Descriptive Statistics			
$	2,000,000	Mean	$	600,000	=AVERAGE(A2:A6)
$	500,000	Standard Error	$	357,770.88	=D6/SQRT(D14)
$	300,000	Median	$	300,000	=MEDIAN(A2:A6)
$	100,000	Mode	$	100,000.00	=MODE(A2:A6)
$	100,000	Standard Deviation	$	800,000	=STDEV(A2:A6)
		Sample Variance		640,000,000,000	=VAR(A2:A6)
		Kurtosis		4.1301	=KURT(A2:A6)
		Skewness		2.0068	=SKEW(A2:A6)
		Range	$	1,900,000	=D12 - D11
		Minimum	$	100,000	=MIN(A2:A6)
		Maximum	$	2,000,000	=MAX(A2:A6)
		Sum	$	3,000,000	=SUM(A2:A6)
		Count		5	=COUNT(A2:A6)

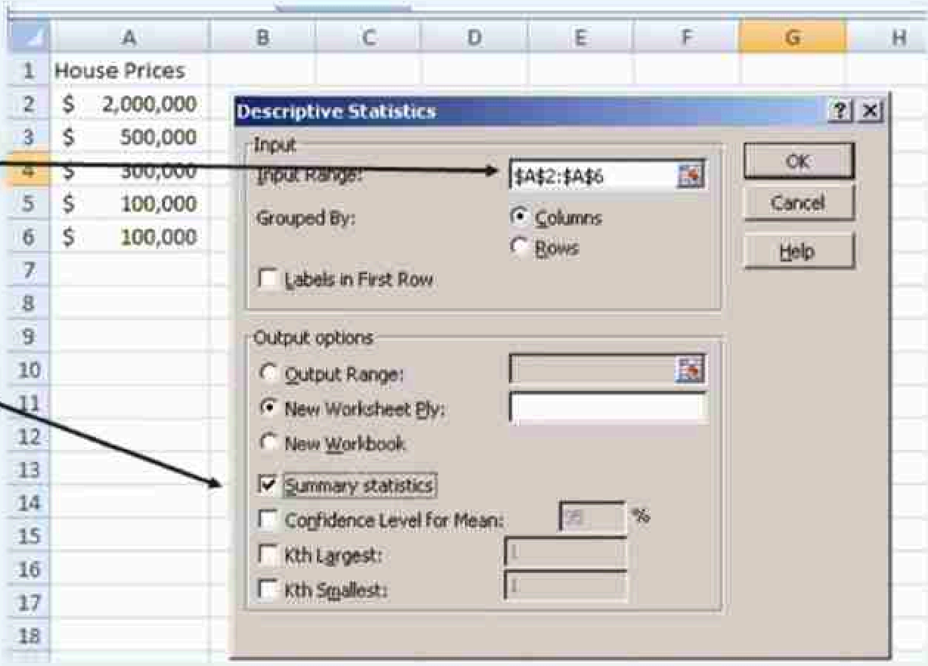
General Descriptive Stats Using Microsoft Excel Data Analysis Tool



1. Select Data.
2. Select Data Analysis.
3. Select Descriptive Statistics and click OK.



General Descriptive Stats Using Microsoft Excel



4. Enter the cell range.

5. Check the Summary Statistics box.

6. Click OK

Excel output

Microsoft Excel descriptive statistics output, using the house price data:

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

House Prices	
Mean	600000
Standard Error	357770.8764
Median	300000
Mode	100000
Standard Deviation	800000
Sample Variance	640,000,000,000
Kurtosis	4.1301
Skewness	2.0068
Range	1900000
Minimum	100000
Maximum	2000000
Sum	3000000
Count	5

Numerical Descriptive Measures for a Population

- Descriptive statistics discussed previously described a *sample*, not the *population*.
- Summary measures describing a population, called **parameters**, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

Numerical Descriptive Measures for a Population: The mean μ

- The population mean is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \Lambda + X_N}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population: The Variance σ^2

- Average of squared deviations of values from the mean

  - Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population: The Standard Deviation σ

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the population variance
- Has the same units as the original data

  - Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Sample statistics versus population parameters

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment

- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Measures: Locating Quartiles

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = (n+1)/4$ ranked value

Second quartile position: $Q_2 = (n+1)/2$ ranked value

Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

where n is the number of observed values

Quartile Measures: Calculation Rules

- When calculating the ranked position use the following rules
  - If the result is a whole number then it is the ranked position to use
  - If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
  - If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

Quartile Measures
Calculating The Quartiles: Example

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

($n = 9$)

Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data,

so $\mathbf{Q_1 = (12+13)/2 = 12.5}$

Q_2 is in the $(9+1)/2 = 5^{th}$ position of the ranked data,

so $\mathbf{Q_2 = median = 16}$

Q_3 is in the $3(9+1)/4 = 7.5$ position of the ranked data,

so $\mathbf{Q_3 = (18+21)/2 = 19.5}$

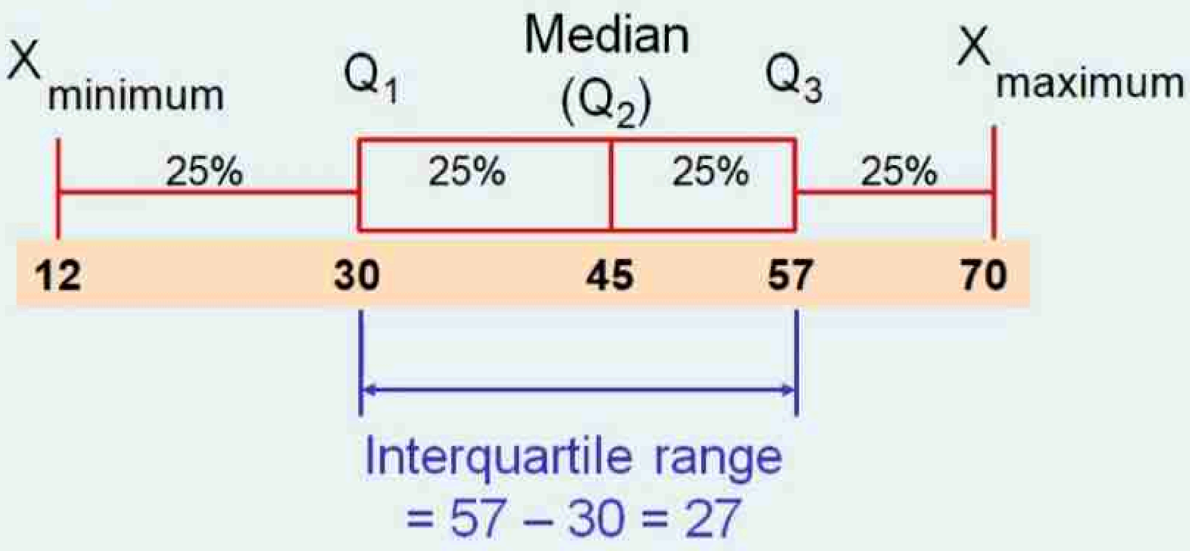
Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Interquartile Range (IQR)

- The IQR is $Q_3 - Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread because it covers the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like Q_1, Q_3, and IQR that are not influenced by outliers are called resistant measures

Calculating The Interquartile Range

Example:



The Five-Number Summary

The five numbers that help describe the center, spread and shape of data are:

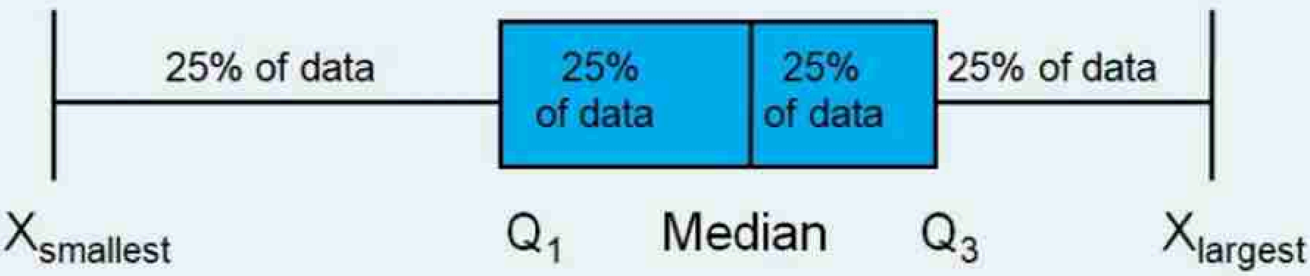
- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Five Number Summary and The Boxplot

- **The Boxplot**: A Graphical display of the data based on the five-number summary:

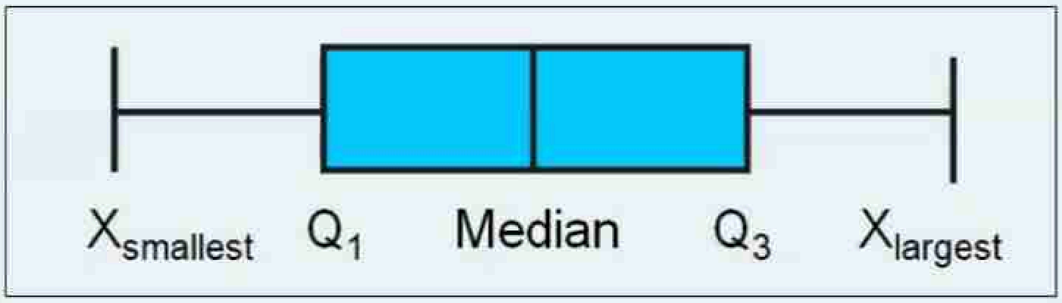
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



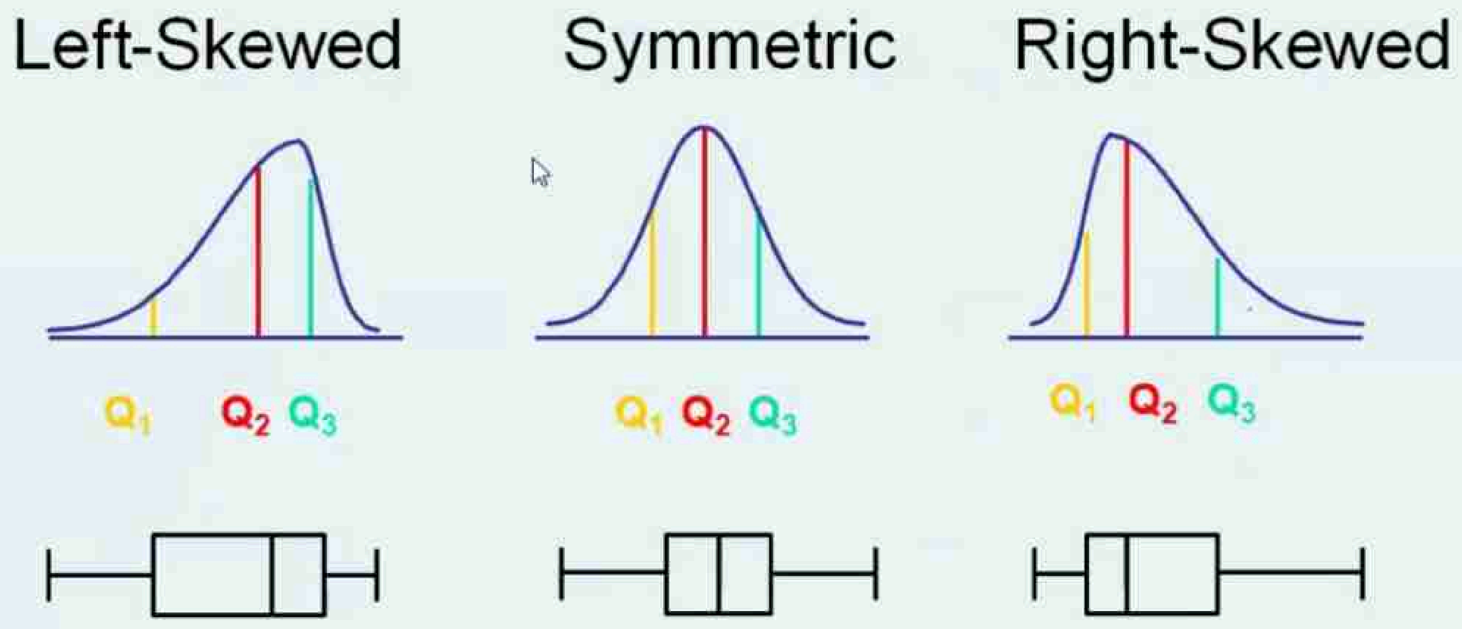
Five Number Summary: Shape of Boxplots

- If data are symmetric around the median then the box and central line are centered between the endpoints



- A Boxplot can be shown in either a vertical or horizontal orientation

Distribution Shape and The Boxplot



Measures Of The Relationship Between Two Numerical Variables

- Scatter plots allow you to visually examine the relationship between two numerical variables and now we will discuss two quantitative measures of such relationships.

- The Covariance
- The Coefficient of Correlation

The Covariance

- The covariance measures the strength of the linear relationship between **two numerical variables** (X & Y)

- The sample covariance:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Only concerned with the strength of the relationship
- No causal effect is implied

Interpreting Covariance

- **Covariance** between two variables:

cov(X,Y) > 0 → X and Y tend to move in the same direction

cov(X,Y) < 0 → X and Y tend to move in opposite directions

cov(X,Y) = 0 → X and Y are independent

- The covariance has a major flaw:
  - It is not possible to determine the relative strength of the relationship from the size of the covariance

Coefficient of Correlation

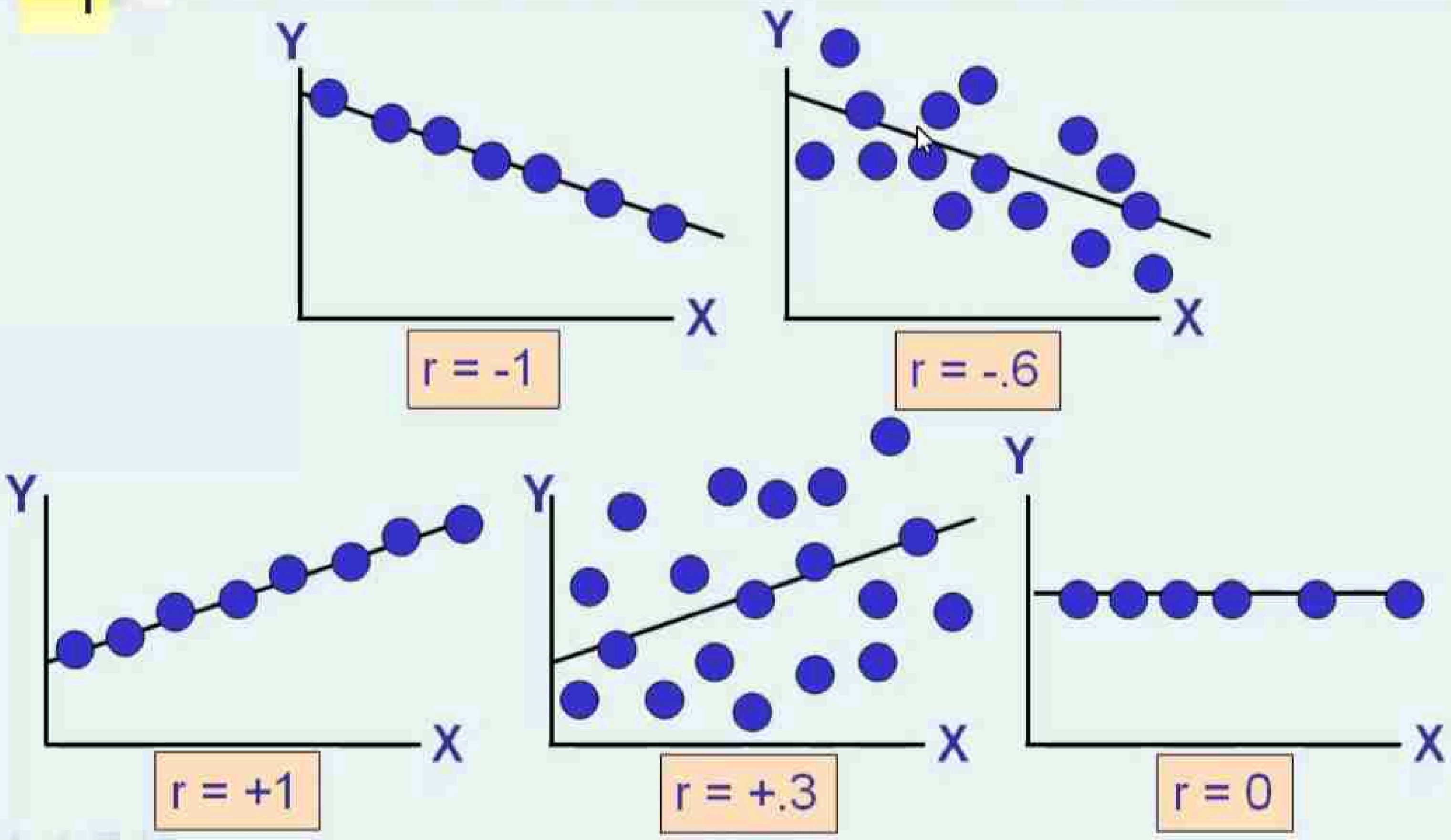
- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

where

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \qquad S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

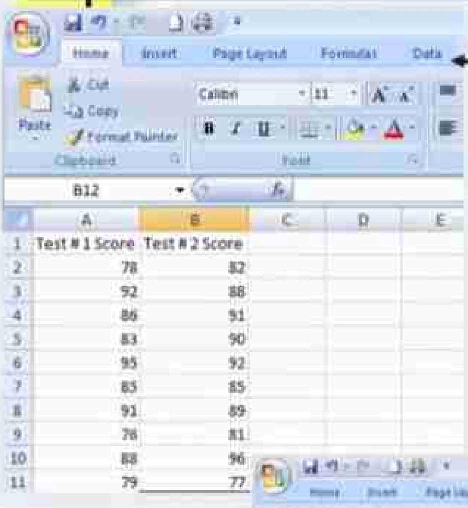
Scatter Plots of Sample Data with Various Coefficients of Correlation



The Coefficient of Correlation Using Microsoft Excel Function

Test #1 Score	Test #2 Score		**Correlation Coefficient**	
78	82		0.7332	=CORREL(A2:A11,B2:B11)
92	88			
86	91			
83	90			
95	92			
85	85			
91	89			
76	81			
88	96			
79	77			

The Coefficient of Correlation Using Microsoft Excel Data Analysis Tool



1. Select Data
2. Choose Data Analysis
3. Choose Correlation & Click OK