Modifying data in Excel

Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)
1	Honda	Accord	33616	29120
2	Nissan	Altima	32289	24649
3	Toyota	Camry	31464	36251
4	Honda	Civic	31213	22463
5	Toyota	Corolla/Matrix	30234	29623
6	Ford	Fusion	27566	22773
7	Hyundai	Sonata	22894	18935
8	Hyundai	Elantra	19255	8225
9	Toyota	Prius	18605	11786
10	Chevrolet	Cruze/Cobalt	18101	10316
11	Chevrolet	Impala	18063	15594
12	Nissan	Sentra	17851	8721
13	Ford	Focus	17178	19500
14	Volkswagon	Jetta	16969	9196
15	Chevrolet	Malibu	15551	17750
16	Mazda	3	12467	11353
17	Nissan	Versa	11075	13811
18	Subaru	Outback	10498	7619
19	Kia	Soul	10028	5106
20	Ford	Fiesta	9787	0

File: Top20Cars.xlsx

Sorting data in excel

- to sort the automobiles by March 2010 sales

  Step 1: Select cells A1:F21

  Step 2: Click the **DATA** tab in the Ribbon

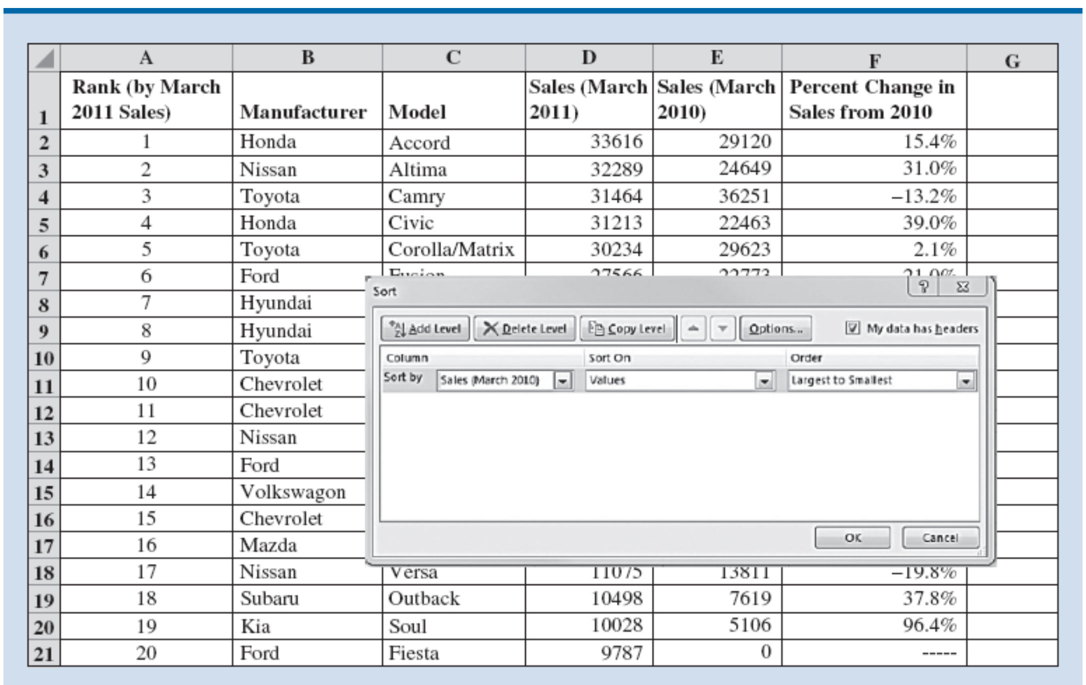
  Step 3: Click **Sort** in the **Sort & Filter** group

  Step 4: Select the check box for **My data has headers**

  Step 5: In the first **Sort by** dropdown menu, select **Sales (March 2010)**

  Step 6: In the **Order** dropdown menu, select **Largest to Smallest**

  Step 7: Click **OK**



	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	3	Toyota	Camry	31464	36251	–13.2%
3	5	Toyota	Corolla/Matrix	30234	29623	2.1%
4	1	Honda	Accord	33616	29120	15.4%
5	2	Nissan	Altima	32289	24649	31.0%
6	6	Ford	Fusion	27566	22773	21.0%
7	4	Honda	Civic	31213	22463	39.0%
8	13	Ford	Focus	17178	19500	–11.9%
9	7	Hyundai	Sonata	22894	18935	20.9%
10	15	Chevrolet	Malibu	15551	17750	–12.4%
11	11	Chevrolet	Impala	18063	15594	15.8%
12	17	Nissan	Versa	11075	13811	–19.8%
13	9	Toyota	Prius	18605	11786	57.9%
14	16	Mazda	3	12467	11353	9.8%
15	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
16	14	Volkswagon	Jetta	16969	9196	84.5%
17	12	Nissan	Sentra	17851	8721	104.7%
18	8	Hyundai	Elantra	19255	8225	134.1%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

The result of using Excel's Sort function for the March 2010 data is shown above. Although the Honda Accord was the best-selling automobile in March 2011, both the Toyota Camry and the Toyota Corolla/Matrix outsold the Honda Accord in March 2010. Note that while Sales (March 2010), which is in column E, is sorted, the data in all other columns are adjusted accordingly.

Filtering data in excel

- Using Excel's Filter function to see the sales of models made by Toyota.

  Step 1: Select cells A1:F21

  Step 2: Click the **DATA** tab in the Ribbon

  Step 3: Click **Filter** in the **Sort & Filter** group

  Step 4: Click on the **Filter Arrow** in column B, next to **Manufacturer**

  Step 5: Select only the check box for **Toyota**. You can easily deselect all choices by unchecking (**Select All**)

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	3	Toyota	Camry	31464	36251	–13.2%
3	5	Toyota	Corolla/Matrix	30234	29623	2.1%
13	9	Toyota	Prius	18605	11786	57.9%

The figure above displays of only the data for models made by Toyota, of the 20 top-selling models in March 2011, Toyota made three of them. Further filter the data by choosing the down arrows in the other columns. All data can be made visible again by clicking on the down arrow in column B and checking (**Select All**) or by clicking **Filter** in the **Sort & Filter** Group again from the **DATA** tab.

Conditional formatting

- Makes it easy to identify data that satisfy certain conditions in a data set.
- To identify the automobile models in Table 2.2 for which sales had decreased from March 2010 to March 2011.

Step 1: Starting with the original data of the file Top20Cars.xlsx, select cells F1:F21

Step 2: Click on the **HOME** tab in the Ribbon

Step 3: Click **Conditional Formatting** in the **Styles** group

Step 4: Select **Highlight Cells Rules**, and click **Less Than** from the dropdown menu

Step 5: Enter *0%* in the **Format cells that are LESS THAN:** box Step 6: Click **OK**

	A	B	C	D	E	F
1	**Rank (by March 2011 Sales)**	**Manufacturer**	**Model**	**Sales (March 2011)**	**Sales (March 2010)**	**Percent Change in Sales from 2010**
2	1	Honda	Accord	33616	29120	15.4%
3	2	Nissan	Altima	32289	24649	31.0%
4	3	Toyota	Camry	31464	36251	–13.2%
5	4	Honda	Civic	31213	22463	39.0%
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%
7	6	Ford	Fusion	27566	22773	21.0%
8	7	Hyundai	Sonata	22894	18935	20.9%
9	8	Hyundai	Elantra	19255	8225	134.1%
10	9	Toyota	Prius	18605	11786	57.9%
11	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
12	11	Chevrolet	Impala	18063	15594	15.8%
13	12	Nissan	Sentra	17851	8721	104.7%
14	13	Ford	Focus	17178	19500	–11.9%
15	14	Volkswagon	Jetta	16969	9196	84.5%
16	15	Chevrolet	Malibu	15551	17750	–12.4%
17	16	Mazda	3	12467	11353	9.8%
18	17	Nissan	Versa	11075	13811	–19.8%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

Here, the models with decreasing sales (Toyota Camry, Ford Focus, Chevrolet Malibu, and Nissan Versa) are now clearly visible.

	A	B	C	D	E	F
1	**Rank (by March 2011 Sales)**	**Manufacturer**	**Model**	**Sales (March 2011)**	**Sales (March 2010)**	**Percent Change in Sales from 2010**
2	1	Honda	Accord	33616	29120	15.4%
3	2	Nissan	Altima	32289	24649	31.0%
4	3	Toyota	Camry	31464	36251	–13.2%
5	4	Honda	Civic	31213	22463	39.0%
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%
7	6	Ford	Fusion	27566	22773	21.0%
8	7	Hyundai	Sonata	22894	18935	20.9%
9	8	Hyundai	Elantra	19255	8225	134.1%
10	9	Toyota	Prius	18605	11786	57.9%
11	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
12	11	Chevrolet	Impala	18063	15594	15.8%
13	12	Nissan	Sentra	17851	8721	104.7%
14	13	Ford	Focus	17178	19500	–11.9%
15	14	Volkswagon	Jetta	16969	9196	84.5%
16	15	Chevrolet	Malibu	15551	17750	–12.4%
17	16	Mazda	3	12467	11353	9.8%
18	17	Nissan	Versa	11075	13811	–19.8%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

We can choose **Data Bars** from the **Conditional Formatting** dropdown menu in the **Styles** Group of the **HOME** tab in the Ribbon.

Data bars are essentially a bar chart input into the cells that show the magnitude of the cell values. The width of the bars in this display are comparable to the values of the variable for which the bars have been drawn; a value of 20 creates a bar twice as wide as that for a value of 10. **Negative values** are shown to the left side of the axis; positive values are shown to the right. Cells with negative values are shaded in a color different from that of cells with positive values.

Creating Distributions from Data

Frequency distributions for categorical data

Frequency distribution: A summary of data that shows the number (frequency) of observations in each of several nonoverlapping classes, typically referred to as **bins**, when dealing with distributions.

Coca-Cola	Sprite	Pepsi
Diet Coke	Coca-Cola	Coca-Cola
Pepsi	Diet Coke	Coca-Cola
Diet Coke	Coca-Cola	Coca-Cola
Coca-Cola	Diet Coke	Pepsi
Coca-Cola	Coca-Cola	Dr. Pepper
Dr. Pepper	Sprite	Coca-Cola
Diet Coke	Pepsi	Diet Coke
Pepsi	Coca-Cola	Pepsi
Pepsi	Coca-Cola	Pepsi
Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coca-Cola	Coca-Cola
Coca-Cola	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coca-Cola	Pepsi	Sprite
Coca-Cola	Diet Coke	

File: Softdrinks.xlsx

Each purchase is for one of five popular soft drinks, which define the five bins: Coca-Cola, Diet Coke, Dr. Pepper, Pepsi, and Sprite.

To get the frequency we can use the five bins and tally. To use excel to get the frequency we will use the function **countif**. COUNTIF is a function to count cells that meet a single criterion. COUNTIF can be used to count cells with dates, numbers, and text that meet specific criteria. The COUNTIF function supports logical operators (>,<,<>,=) and wildcards (*,?) for partial matching.

```
=COUNTIF (range, criteria)
```

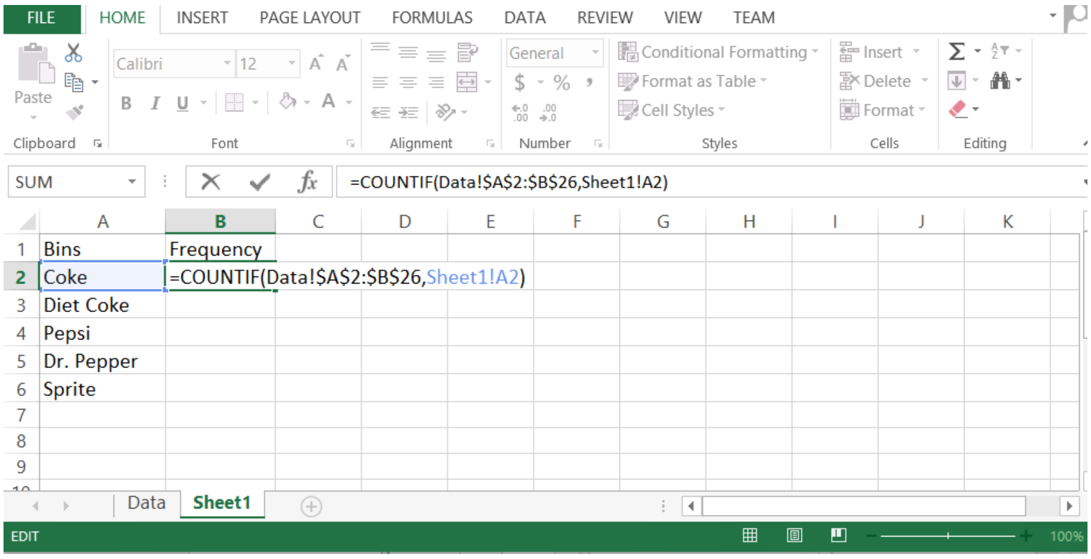
range - The range of cells to count.

criteria - The criteria that controls which cells should be counted.

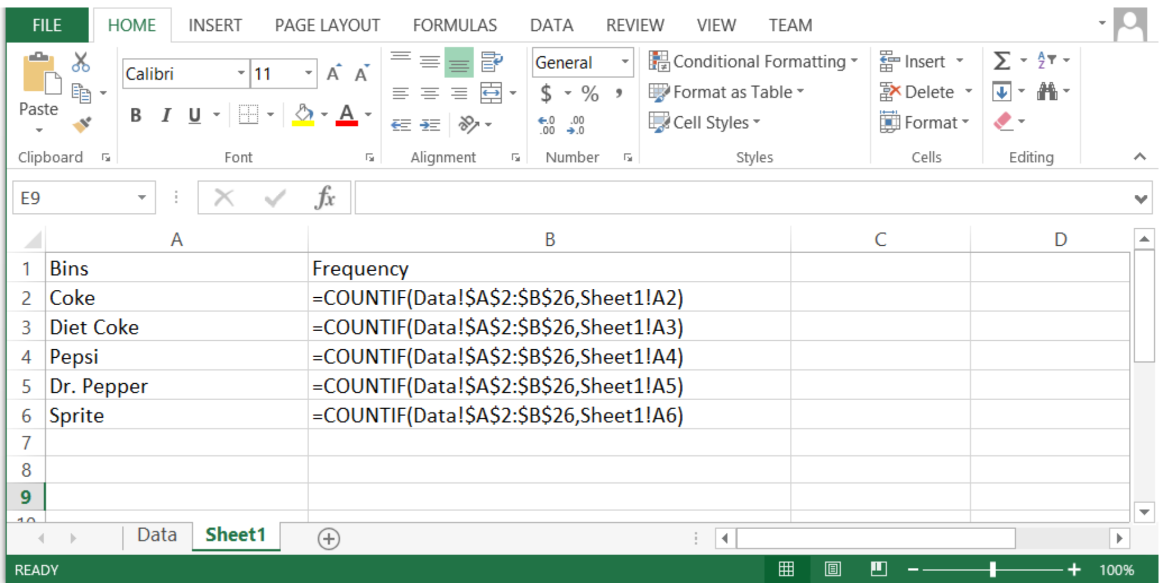
Step 1: Select the cell in which you want to see the count. These are your bins

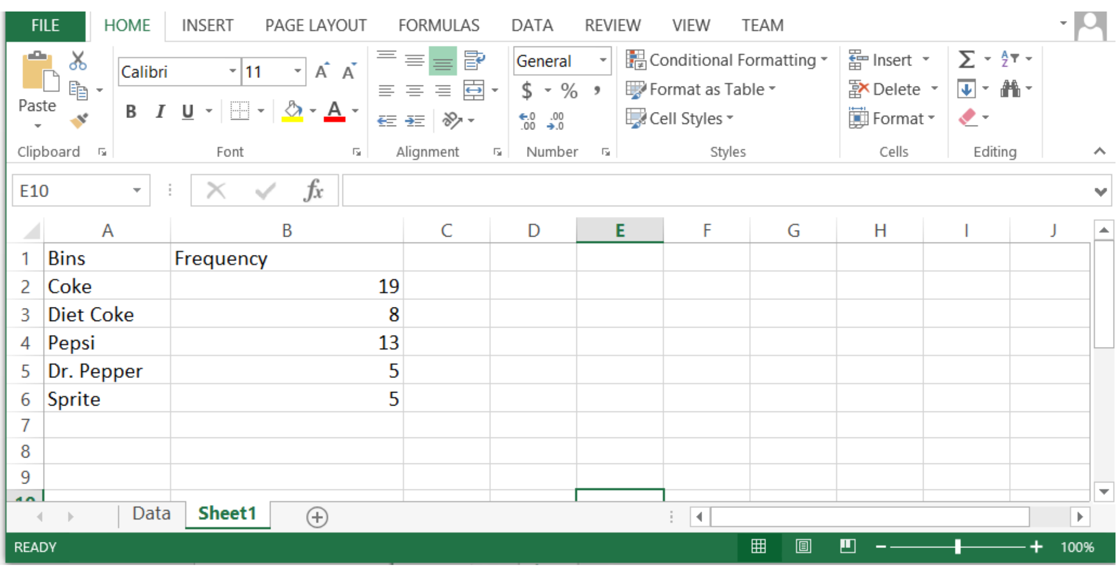
Steps 2: Select the cell where you want to displays the frequency.

Step 3: Type your formula.



Step 4: Copy and paste the formula





Frequency Distribution of Soft drinks Purchases

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

The frequency distribution summarizes information about the popularity of the five soft drinks: Coca-Cola is the leader, Pepsi is second, Diet Coke is third, and Sprite and Dr. Pepper are tied for fourth.

Relative frequency and percent frequency distributions

Relative frequency distribution: It is a tabular summary of data showing the relative frequency for each bin.

Percent frequency distribution: Summarizes the percent frequency of the data for each bin.

Used to provide estimates of the relative likelihoods of different values of a random variable.

Soft Drink	Relative Frequency	Percent Frequency (%)
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

The table shows that the relative frequency for Coca-Cola is 19/50 = 0.38, the relative frequency for Diet Coke is 8/50 = 0.16, and so on.

From the percent frequency distribution, it is seen that 38 percent of the purchases were Coca-Cola, 16 percent of the purchases were Diet Coke, and so on.

Frequency distributions for quantitative data

Three steps necessary to define the classes for a frequency distribution with quantitative data:

1. Determine the number of nonoverlapping bins. Using the $2^n \geq n$

2. Determine the width of each bin.

- It should be the same for each bin.
- Thus the choices of the number of bins and the width of bins are not independent decisions.
- A larger number of bins means a smaller bin width and vice versa.

$$\text{Approximate bin width} = \frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of bins}}$$

NOTE: You must *round up*

3. Determine the bin limits.

- The starting point for each class should be divisible by the width or less than the lowest data value
- For the data in our example, the minimum is 65 and the maximum is 114, a range of about 50. We can therefore choose intervals of size 5, and have ten of them. Our classes are 65 - 70, 70 - 75, etc.

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

File: AuditData.xlsx

Audit Times (days)	Frequency	Relative Frequency	Percent Frequency
10–14	4	0.20	20
15–19	8	0.40	40
20–24	5	0.25	25
25–29	2	0.10	10
30–34	1	0.05	5

Step 1: Determine the number of nonoverlapping bins

$2^n \geq n$

$2^1 \geq 20 = 2 \geq 20$ (False)

$2^2 \geq 20 = 4 \geq 20$ (False)

$2^4 \geq 20 = 16 \geq 20$ (False)

$2^5 \geq 20 = 32 \geq 20$ (True)

Number of bins: 5

Step 2: Determine the width of each bin.

$$\text{Approximate bin width} = \frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of bins}}$$

The largest data value is 33, and the smallest data value is 12.

= (33 – 12)/5

= 4.2 (ROUND up)

= 5

Step 3: Determine the bin limits.

- Multiple of width (5,10,15 and so on)
- Since the lowest data is 12, multiple of 5 lower than 12 is 10. The smallest data value, 12, is included in the 10–14 bin. We then selected 15 days as the lower bin limit and 19 days as the upper bin limit of the next class.
- We continued defining the lower and upper bin limits to obtain a total of five classes: 10–14, 15–19, 20–24, 25–29, and 30–34.

Using Excel function

	A	B	C	D
1	Year-End Audit Times (in Days)			
2	12	14	19	18
3	15	15	18	17
4	20	27	22	23
5	22	21	33	28
6	14	18	16	13
7				
8				
9	Bin	Frequency		
10	14	=FREQUENCY(A2:D6,A10:A14)		
11	19	=FREQUENCY(A2:D6,A10:A14)		
12	24	=FREQUENCY(A2:D6,A10:A14)		
13	29	=FREQUENCY(A2:D6,A10:A14)		
14	34	=FREQUENCY(A2:D6,A10:A14)		

	A	B	C	D
1	Year-End Audit Times (in Days)			
2	12	14	19	18
3	15	15	18	17
4	20	27	22	23
5	22	21	33	28
6	14	18	16	13
7				
8				
9	Bin	Frequency		
10	14	4		
11	19	8		
12	24	5		
13	29	2		
14	34	1		

The sample of 20 audit times is contained in cells A2:D6.

The upper limits of the defined bins are in cells A10:A14.

We can use the FREQUENCY function in Excel to count the number of observations in each bin:

Step 1. Select cells B10:B14

Step 2. Enter the formula =FREQUENCY(A2:D6, A10:A14). The range A2:D6 defines the data set, and the range A10:A14 defines the bins

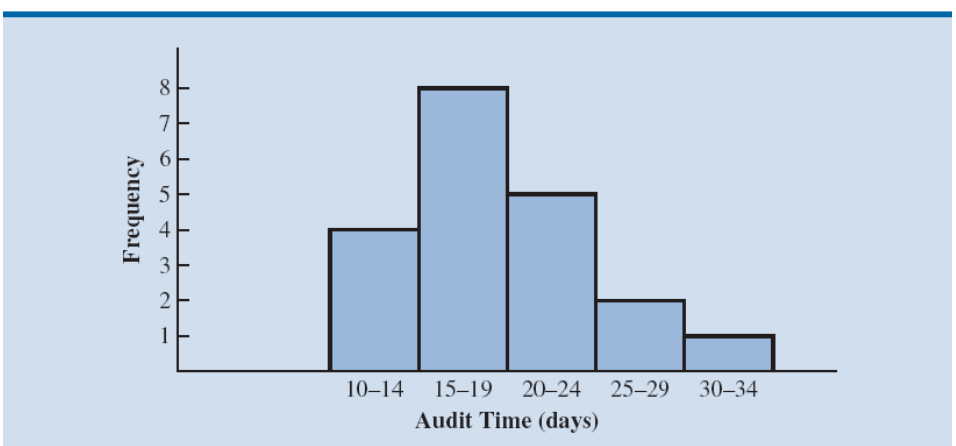
Step 3. Press CTRL+SHIFT+ENTER

Excel will then fill in the values for the number of observations in each bin in cells B10 through B14 because these were the cells selected in Step 1 above.

Histogram

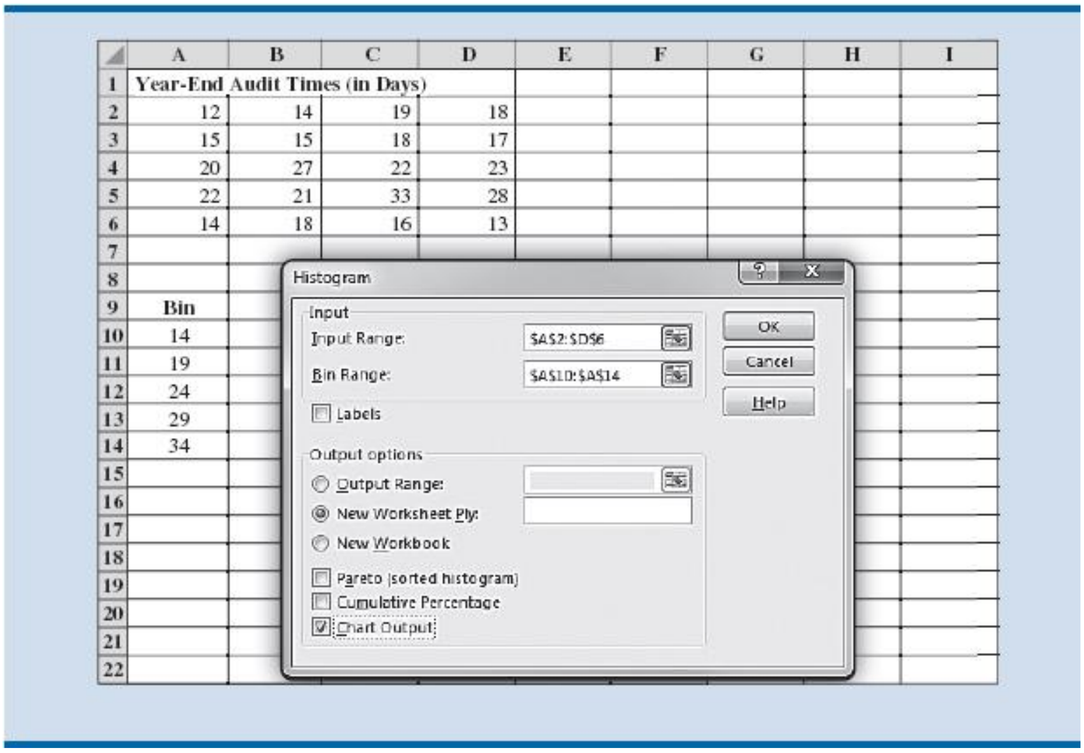
Histogram: A common graphical presentation of quantitative data

- Constructed by placing the variable of interest on the horizontal axis and the selected frequency measure (absolute frequency, relative frequency, or percent frequency) on the vertical axis.
- The frequency measure of each class is shown by drawing a rectangle whose base is determined by the class limits on the horizontal axis and whose height is the corresponding frequency measure.



The class with the greatest frequency is shown by the rectangle appearing above the class of 15–19 days. The height of the rectangle shows that the frequency of this class is 8.

Creating Histogram using Excel



- Histograms can be created in Excel using the Data Analysis ToolPak. Following are the steps to create histogram in Excel.

  Step 1. Click the **DATA** tab in the Ribbon

  Step 2. Click **Data Analysis** in the **Analysis** group

  Step 3. When the **Data Analysis** dialog box opens, choose **Histogram** from the list of **Analysis Tools**, and click **OK**

  In the **Input Range:** box, enter *A2:D6*

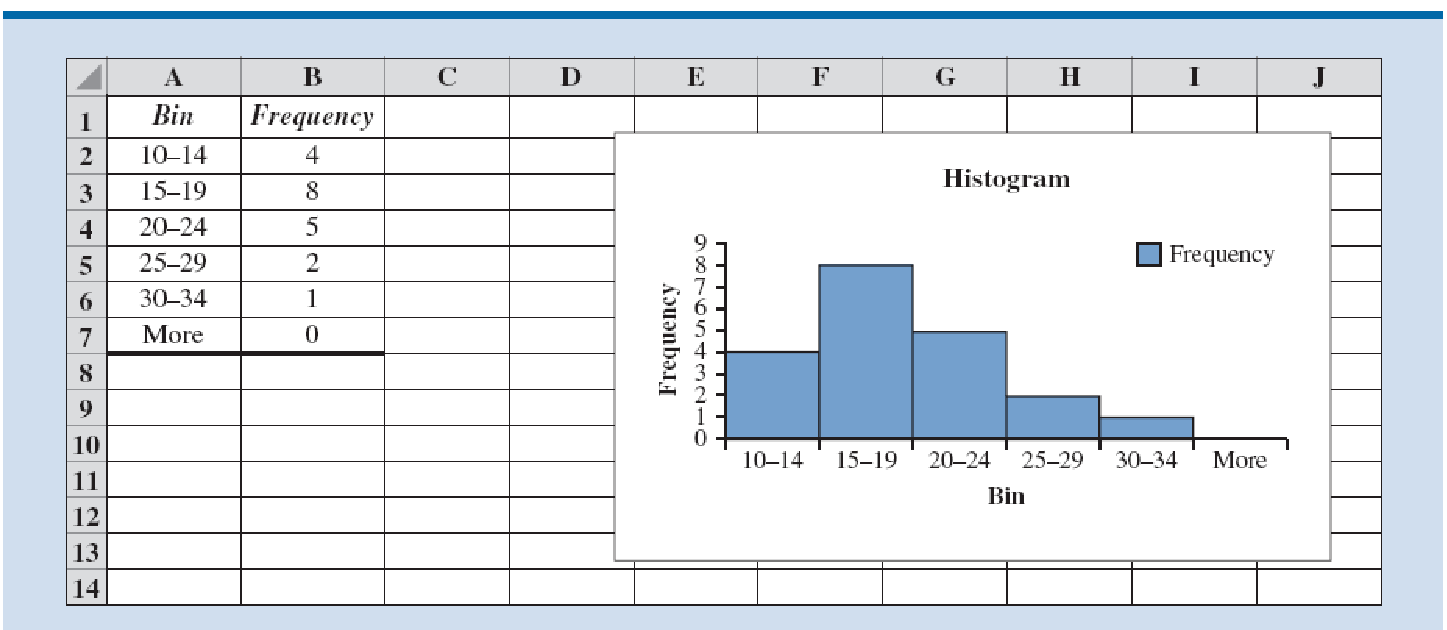
  In the **Bin Range:** box, enter *A10:A14*

  Under **Output Options:**, select **New Worksheet Ply:**

  Select the check box for **Chart Output**

  Click **OK**

  Note: Data Analysis (File-> Options-> Add-Ins-> Go-> Ok)



➢ Modify the bin ranges in column A by typing the values shown in Figure 2.13 into cells A2:A6 so that the chart created by Excel shows both the lower and upper limits for each bin.

➢ We have also removed the gaps between the columns in the histogram in Excel to match the traditional format of histograms.

➢ To remove the gaps between the columns in the Histogram created by Excel, follow these steps:

Step 1. Right-click on one of the columns in the histogram

Select **Format Data Series…**

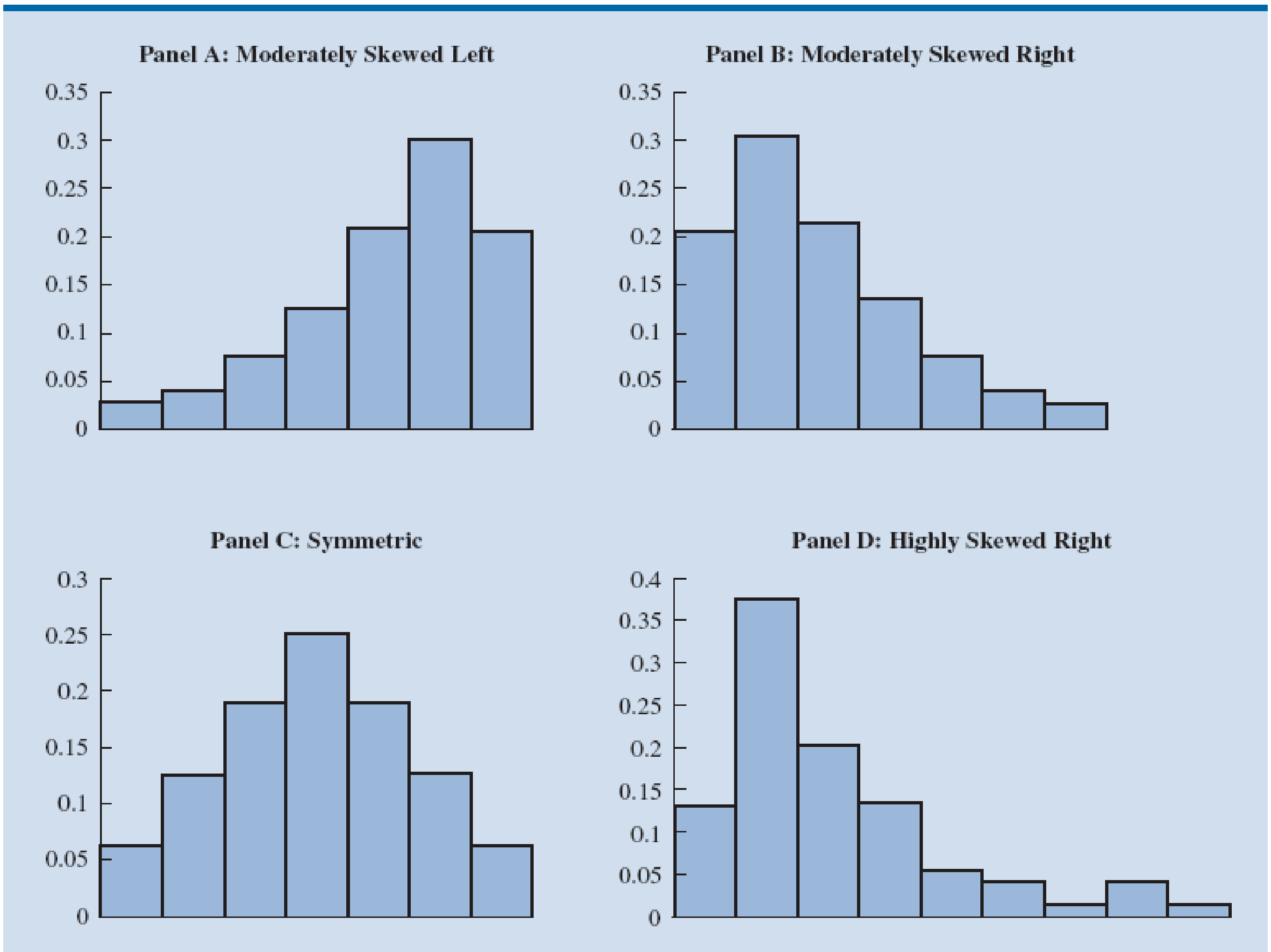
Step 2. When the **Format Data Series** pane opens, click the **Series Options** button.

Set the **Gap Width** to 0%

Histogram provides information about the shape, or form, of a distribution.

- **Skewness**: Lack of symmetry

  Important characteristic of the shape of a distribution



- Panel A: Moderately skewed to the left
- Here, tail extends farther to the left than to the right.
  - Example: Exam scores, with no scores above 100 percent, most of the scores above 70 percent, and only a few really low scores.
  - Panel B: Moderately skewed to the right
- Tail extends farther to the right than to the left.
  - Example: Housing prices; a few expensive houses create the skewness in the right tail.
  - Panel C: Symmetric
- The left tail mirrors the shape of the right tail.

- Example: Data for SAT scores, the heights and weights of people, and so on lead to histograms that are roughly symmetric.
- Panel D: Highly skewed to the right
- Example: Data on housing prices, salaries, purchase amounts, and so on often result in histograms skewed to the right.

Cumulative frequency distribution

- **Cumulative frequency distribution**: A variation of the frequency distribution that provides another tabular summary of quantitative data.

  Uses the number of classes, class widths, and class limits developed for the frequency distribution.

  Shows the number of data items with values less than or equal to the upper class limit of each class.

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

- The cumulative frequency for this class is simply the sum of the frequencies for all classes with data values less than or equal to 24.
- The sum of the frequencies for classes 10–14, 15–19, and 20–24 indicates that 4 + 8 + 5 = 17 data values are less than or equal to 24. Hence, the cumulative frequency for this class is 17.
- In addition, the cumulative frequency distribution in Table 2.8 shows that four audits were completed in 14 days or less and that 19 audits were completed in 29 days or less.
- The cumulative relative frequency distribution can be computed either by summing the relative frequencies in the relative frequency distribution or by dividing the cumulative frequencies by the total number of items.
- Using the latter approach, we found the cumulative relative frequencies in column 3 of Table 2.8 by dividing the cumulative frequencies in column 2 by the total number of items (n = 20).
- The cumulative percent frequencies were again computed by multiplying the relative frequencies by 100.
- The cumulative relative and percent frequency distributions show that 0.85 of the audits, or 85 percent, were completed in 24 days or less, 0.95 of the audits, or 95 percent, were completed in 29 days or less, and so on.

Measures of Location

These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode.

Arithmetic Mean

- Average value for a variable.
- The mean is denoted by $\bar{x}$.
- Denoted by $\bar{x}$ for sample data.
- Denoted by μ for population data.
  - n = sample size

    x_1 = value of variable x for the first observation

    x_2 = value of variable x for the second observation

    x_n = value of variable x for the nth observation

$$\text{Sample mean, } \bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

Value in the middle when the data are arranged in ascending order.

Middle value, for an odd number of observations

Average of two middle values, for an even number of observations

Mode

Value that occurs most frequently in a data set.

Consider the class size data:

32 42 46 46 54

Observe - 46 is the only value that occurs more than once.

Mode is 46.

Multimodal data - Data contain at least two modes.

Bimodal data - Data contain exactly two modes.

Home Sale	Selling Price ($)
1	138,000
2	254,000
3	186,000
4	257,500
5	108,000
6	254,000
7	138,000
8	298,000
9	199,500
10	208,000
11	142,000
12	456,250

File: Homesales.xlsx

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	=AVERAGE(B2:B13)
3	2	254,000		Median:	=MEDIAN(B2:B13)
4	3	186,000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257,500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	$ 219,937.50
3	2	254,000		Median:	$ 203,750.00
4	3	186,000		Mode 1:	$ 138,000.00
5	4	257,500		Mode 2:	$ 254,000.00
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

- The Excel MODE.SNGL function will return only a single most-often-occurring value.
- For multimodal distributions, we must use the MODE.MULT command in Excel to return more than one mode.
  - To find both of the modes in Excel, we take these steps:

    Step 1. Select cells E4 and E5

    Step 2. Enter the formula =MODE.MULT(B2:B13)

    Step 3. Press **CTRL+SHIFT+ENTER**
  - Excel enters the values for both modes of this data set in cells E4 and E5: $138,000 and $254,000.

Measure of Variability

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. How spread out are the values? While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center.

Range

Range: Found by subtracting the smallest value from the largest value in a data set.

Variance

Variance: Measure of variability that utilizes all the data.

It is based on the deviation about the mean, which is the difference between the value of each observation (x_i) and the mean.

The deviations about the mean are squared while computing the variance.

Sample variance, $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$

Population variance , $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

- Computation of Sample Variance:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64$$

Standard Deviation; Coefficient of Variation

Standard deviation: Positive square root of the variance

It tells you how spread out the data is. It is a measure of how far each observed value is from the mean. In any distribution, about 95% of values will be within 2 **standard deviations** of the mean

Measured in the same units as the original data.

For sample , $s = \sqrt{s^2}$

For population, $\sigma = \sqrt{\sigma^2}$

Coefficient of variation:

It is the ratio of the standard deviation to the mean. The higher the **coefficient of variation**, the greater the level of dispersion around the mean. It is generally expressed as a percentage. ... The lower the value of the **coefficient of variation**, the more precise the estimate.

$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$

Measures the standard deviation relative to the mean.

Expressed as a percentage.

<u>Illustration</u>:

Consider the class size data:

46 54 42 46 32

Mean, $\bar{x} = 44$

Standard deviation, $s = 8$

Coefficient of variation = $\left(\frac{8}{44} \times 100\right)\% = 18.2\%$

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	=AVERAGE(B2:B13)
3	2	254000		Median:	=MEDIAN(B2:B13)
4	3	186000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108000			
7	6	254000		Range:	=MAX(B2:B13)-MIN(B2:B13)
8	7	138000		Variance:	=VAR.S(B2:B13)
9	8	298000		Standard Deviation:	=STDEV.S(B2:B13)
10	9	199500			
11	10	208000		Coefficient of Variation:	=E9/E2
12	11	142000			
13	12	456250		85th Percentile:	=PERCENTILE.EXC(B2:B13,0.85)

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	$ 219,937.50
3	2	254,000		Median:	$ 203,750.00
4	3	186,000		Mode 1:	$ 138,000.00
5	4	257,500		Mode 2:	$ 254,000.00
6	5	108,000			
7	6	254,000		Range:	$ 348,250.00
8	7	138,000		Variance:	9037501420
9	8	298,000	Standard Deviation:		$ 95,065.77
10	9	199,500			
11	10	208,000	Coefficient of Variation:		43.22%
12	11	142,000			
13	12	456,250	85th Percentile:		$ 305,912.50

Percentile

Percentile: Value of a variable at which a specified (approximate) percentage of observations are below that value.

The *p*th percentile tells us the point in the data where:

- Approximately *p* percent of the observations have values less than the *p*th percentile;
- Approximately (100 – *p*) percent of the observations have values greater than the *p*th percentile.

Steps to calculate the *p*th percentile:

Step 1: Arrange the data in ascending order (smallest to largest value).

Step 2: Compute $k = (n + 1) \times p$.

Step 3: Divide k into its integer component, i, and its decimal component, d.

- If $d = 0$, find the *k*th largest value in the data set. This is the *p*th percentile.
- If $d > 0$, the percentile is between the values in positions i and $i + 1$ in the sorted data. To find this percentile, we must interpolate between these two values.
    - Calculate the difference between the values in positions i and $i + 1$ in the sorted data set. We define this difference between the two values as m.
    - Multiply this difference by d: $t = m \times d$.

- To find the *p*th percentile, add *t* to the value in position *i* of the sorted data.

Illustration: To determine the 85th percentile for the home sales data in Table 2.9.

Step 1: Arrange the data in ascending order.

108,000 138,000 138,000 142,000 186,000 199,500 208,000 254,000 254,000 257,500 298,000 456,250

Step 2: Compute $k = (n + 1) \times p = (12 + 1) \times 0.85 = 11.05$.

Step 3: Dividing 11.05 into the integer and decimal components gives us $i = 11$ and $d = 0.05$.

- $d > 0$, interpolate between the values in the 11th and 12th positions in the sorted data.
  - The value in the 11th position is 298,000, and
  - The value in the 12th position is 456,250.
  - $m = 456{,}250 - 298{,}000 = 158{,}250$
  - $t = m \times d = 158{,}250 \times 0.05 = 7912.5$
  - *p*th percentile = 298,000 + 7912.5 = 305,912.5

$305,912.50 represents the 85th percentile of the home sales data.

Quartiles

Quartiles : When the data is divided into four equal parts:

Each part contains approximately 25% of the observations.

Division points are referred to as quartiles.

Q_1 = first quartile, or 25th percentile

Q_2 = second quartile, or 50th percentile (also the median)

Q_3 = third quartile, or 75th percentile

Z-scores

z-score:Measures the relative location of a value in the data set.

Helps to determine how far a particular value is from the mean relative to the data set's standard deviation.

Standardized value

If $x_1, x_2, \ldots, x_n$ is a sample of *n* observations

$$z_i = \frac{x_i - \bar{x}}{s}$$

- z_i = *z*-score for x_i
- $\bar{x}$ = sample mean
- *s* = sample standard deviation

Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = .25
54	10	10/8 = 1.25
42	−2	−2/8 = −.25
46	2	2/8 = .25
32	−12	−12/8 = −1.50

z_1 = .25 would indicate that x_1 is .25 standard deviations greater than the sample mean.

For class size data, $\bar{x}$ = 44 and s = 8.

For observations with a value > mean, *z*-score > 0.

For observations with a value < mean, *z*-score < 0.

	A	B	C
1	Home Sale	Selling Price ($)	z-Score
2	1	138000	=STANDARDIZE(B2,B15,B16)
3	2	254000	=STANDARDIZE(B3,B15,B16)
4	3	186000	=STANDARDIZE(B4,B15,B16)
5	4	257500	=STANDARDIZE(B5,B15,B16)
6	5	108000	=STANDARDIZE(B6,B15,B16)
7	6	254000	=STANDARDIZE(B7,B15,B16)
8	7	138000	=STANDARDIZE(B8,B15,B16)
9	8	298000	=STANDARDIZE(B9,B15,B16)
10	9	199500	=STANDARDIZE(B10,B15,B16)
11	10	208000	=STANDARDIZE(B11,B15,B16)
12	11	142000	=STANDARDIZE(B12,B15,B16)
13	12	456250	=STANDARDIZE(B13,B15,B16)
14			
15	Mean:	=AVERAGE(B2:B13)	
16	Standard Deviation:	=STDEV.S(B2:B13)	

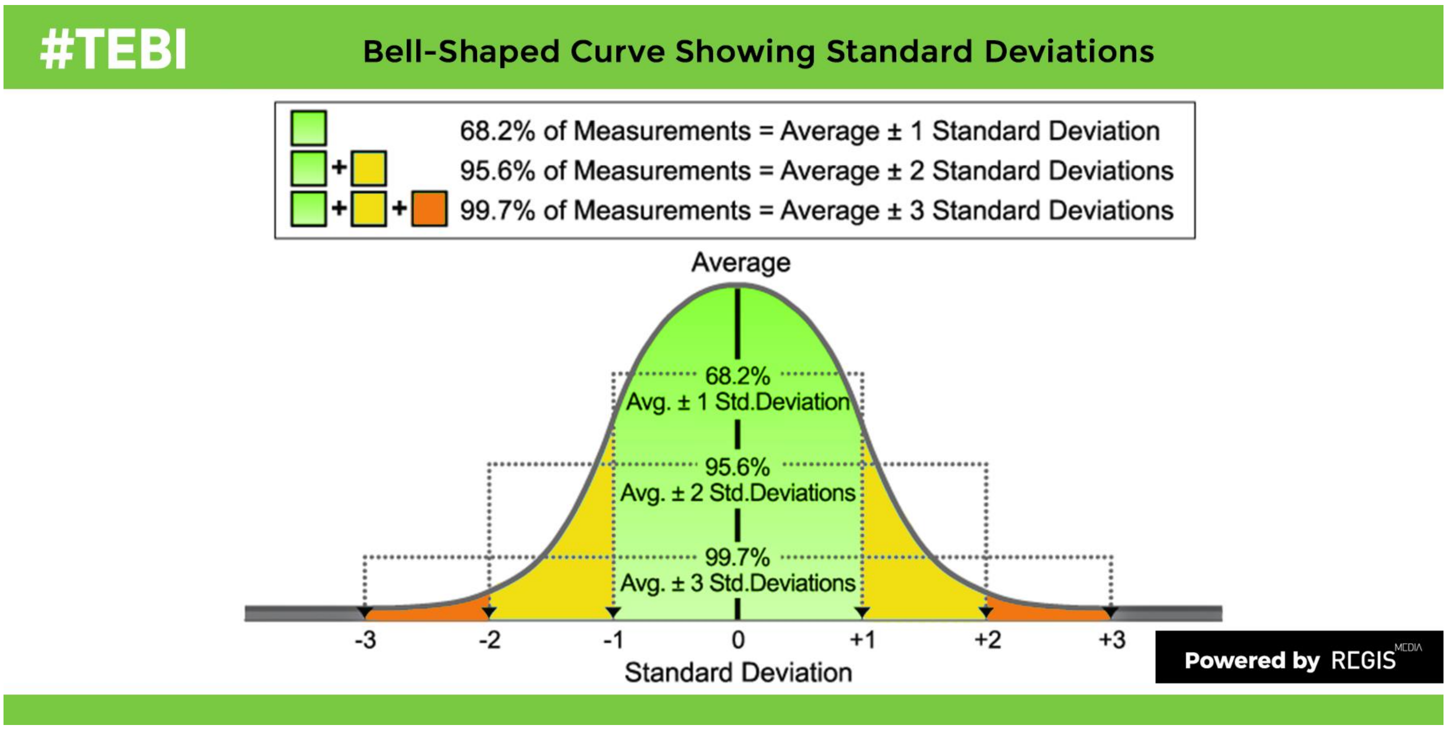
	A	B	C
1	Home Sale	Selling Price ($)	z-Score
2	1	138,000	–0.862
3	2	254,000	0.358
4	3	186,000	–0.357
5	4	257,500	0.395
6	5	108,000	–1.177
7	6	254,000	0.358
8	7	138,000	–0.862
9	8	298,000	0.821
10	9	199,500	–0.215
11	10	208,000	–0.126
12	11	142,000	–0.820
13	12	456,250	2.486
14			
15	Mean:	$ 219,937.50	
16	Standard Deviation:	$ 95,065.77	

- To calculate the *z*-scores, the mean and standard deviation for the data set must be provided in the arguments of the STANDARDIZE function.
- For instance, the *z*-score in cell C2 is calculated with the formula =STANDARDIZE(B2, B15, B16), where cell B15 contains the mean of the home sales data and cell B16 contains the standard deviation of the home sales data.
- Then, this formula can be copy and pasted into cells C3:C13.

Empirical rule

- For data having a bell-shaped distribution:
  - Within 1 standard deviation – approximately 68% of the data values.

- Within 2 standard deviations – approximately 95% of the data values.
- Within 3 standard deviations – almost all the data values.



Identifying outliers:

- **Outliers**: Extreme values in a data set.
- It can be identified using standardized values (z-scores).
    - Any data value with a z-score less than –3 or greater than +3 is an outlier.