Statistics for non-statisticians with practical applications on Microsoft Excel

Ahmad Th. AL Sultan

PH.D. in Operations Research & Statistics

2024

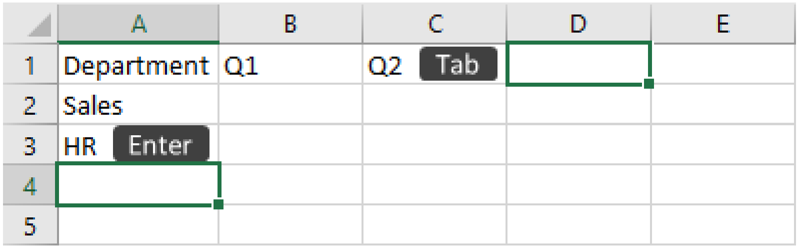
Contents Table

Chapter 1
Writing formulas

1.1 Enter data

To manually enter data:

1. Select an empty cell, such as A1, and then type text or a number.
2. Press Enter or Tab to move to the next cell.

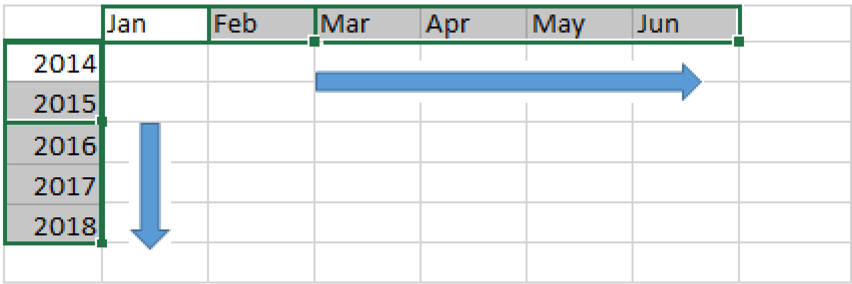


	A	B	C	D	E
1	Department	Q1	Q2		
2	Sales				
3	HR				
4					
5					

To fill data in a series:

1. Enter the beginning of the series in two cells: such as Jan and Feb; or 2014 and 2015.

2. Select the two cells containing the series, and then drag the fill handle Fill handle across or down the cells.



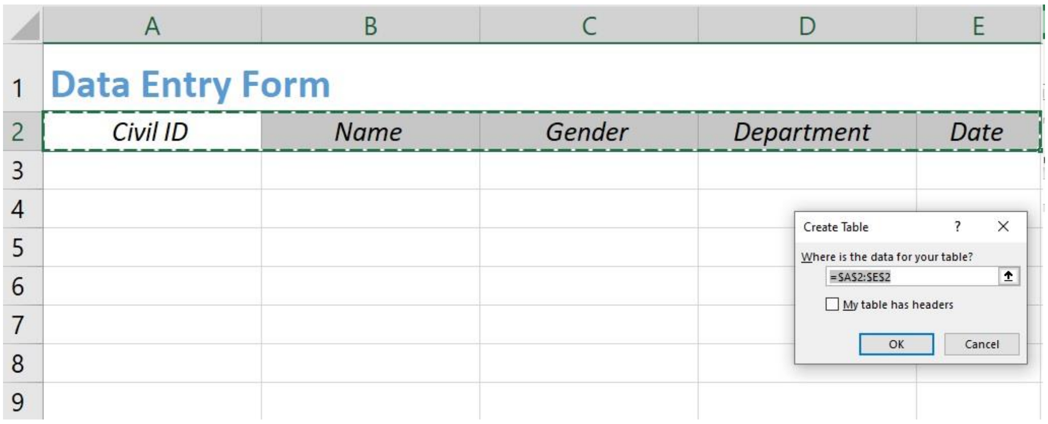
	Jan	Feb	Mar	Apr	May	Jun
2014						
2015						
2016						
2017						
2018						

1.2 Data Entry Form

We can enter data using entry form

<u>*Steps*</u>:

- Insert → table
- quick access toolbar → more commends → commands not in the ribbon.
- Select “form”.
- Enter the data and press “enter” to enter new record.



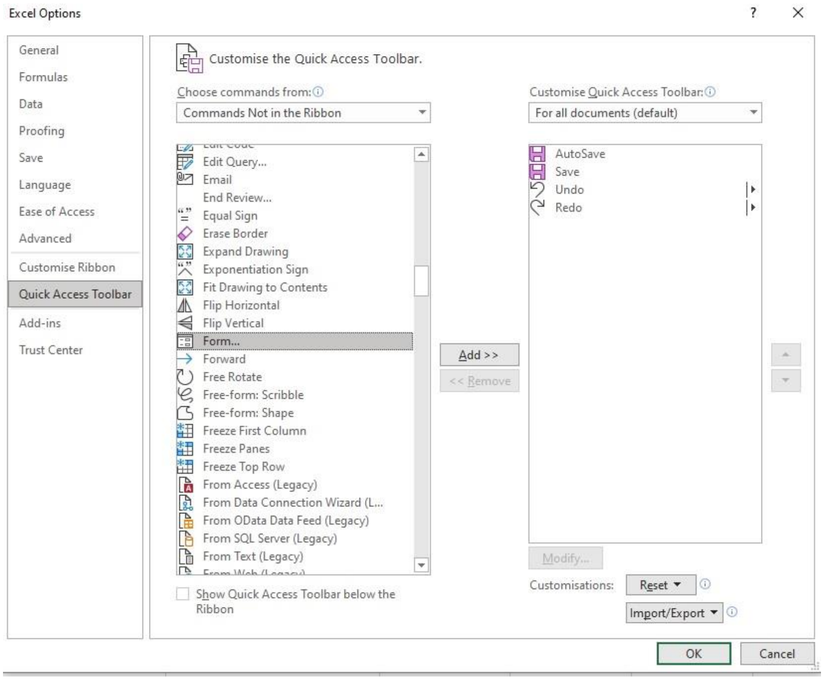
	A	B	C	D	E
1	Data Entry Form				
2	Civil ID	Name	Gender	Department	Date
3					
4					
5					
6					
7					
8					
9					

Create Table

Where is the data for your table?

=$A$2:$E$2

☐ My table has headers



Excel Options

Customise the Quick Access Toolbar.

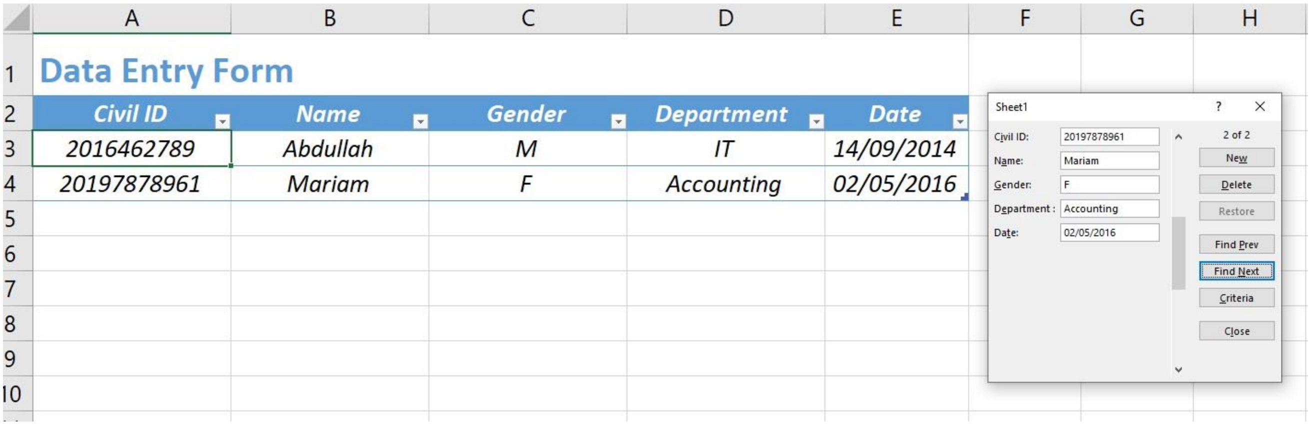
Choose commands from: Commands Not in the Ribbon

Customise Quick Access Toolbar: For all documents (default)

Commands: Edit Code, Edit Query..., Email, End Review..., Equal Sign, Erase Border, Expand Drawing, Exponentiation Sign, Fit Drawing to Contents, Flip Horizontal, Flip Vertical, Form..., Forward, Free Rotate, Free-form: Scribble, Free-form: Shape, Freeze First Column, Freeze Panes, Freeze Top Row, From Access (Legacy), From Data Connection Wizard (L..., From OData Data Feed (Legacy), From SQL Server (Legacy), From Text (Legacy), From Web (Legacy)

Quick Access Toolbar: AutoSave, Save, Undo, Redo

☐ Show Quick Access Toolbar below the Ribbon



Data Entry Form

Civil ID	Name	Gender	Department	Date
2016462789	Abdullah	M	IT	14/09/2014
20197878961	Mariam	F	Accounting	02/05/2016

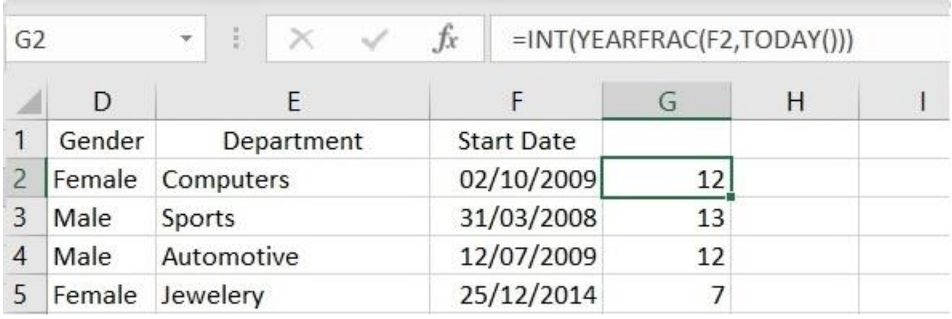
Sheet1

2 of 2

Civil ID:	20197878961
Name:	Mariam
Gender:	F
Department :	Accounting
Date:	02/05/2016

1.3 Some useful formulas

1.3.1 Calculate Age Based on Date of Birth.



G2 =INT(YEARFRAC(F2,TODAY()))

	D	E	F	G
1	Gender	Department	Start Date	
2	Female	Computers	02/10/2009	12
3	Male	Sports	31/03/2008	13
4	Male	Automotive	12/07/2009	12
5	Female	Jewelery	25/12/2014	7

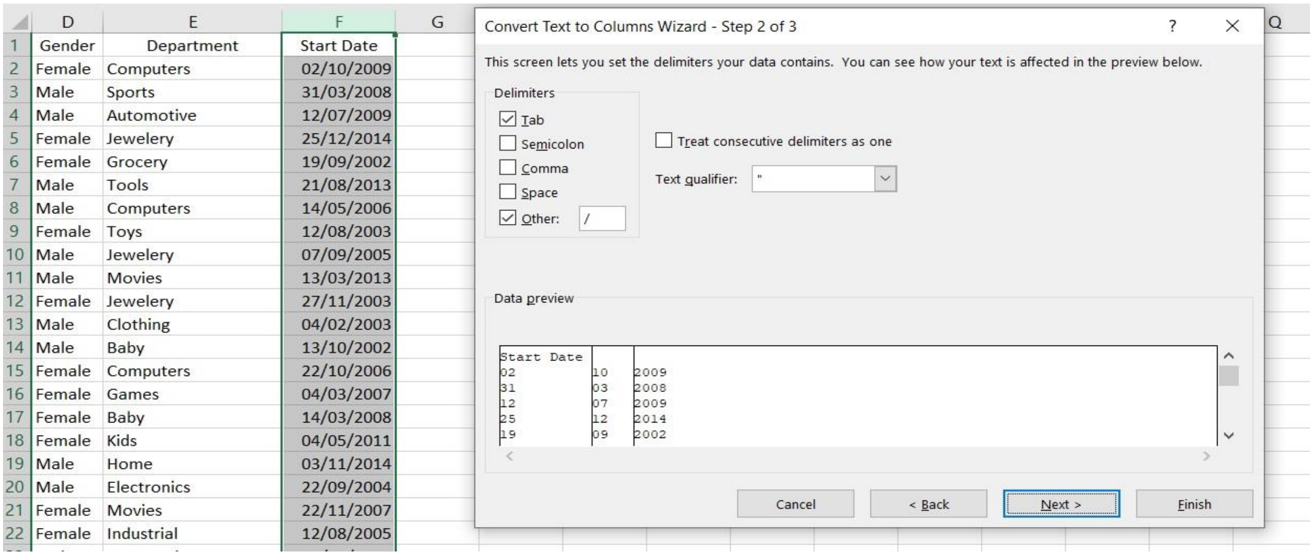
=INT(YEARFRAC(f2,TODAY()))

1.3.2 Split Data from One Cell into Multiple Columns.

Q: For each employee, we want to separate the start "day", "month", and "year" in different column.

<u>Steps</u>:

- Select a column.
- Data → text to column.
- Delimited → other “ / ”.



	D	E	F	G
1	Gender	Department	Start Date	
2	Female	Computers	02/10/2009	
3	Male	Sports	31/03/2008	
4	Male	Automotive	12/07/2009	
5	Female	Jewelery	25/12/2014	
6	Female	Grocery	19/09/2002	
7	Male	Tools	21/08/2013	
8	Male	Computers	14/05/2006	
9	Female	Toys	12/08/2003	
10	Male	Jewelery	07/09/2005	
11	Male	Movies	13/03/2013	
12	Female	Jewelery	27/11/2003	
13	Male	Clothing	04/02/2003	
14	Male	Baby	13/10/2002	
15	Female	Computers	22/10/2006	
16	Female	Games	04/03/2007	
17	Female	Baby	14/03/2008	
18	Female	Kids	04/05/2011	
19	Male	Home	03/11/2014	
20	Male	Electronics	22/09/2004	
21	Female	Movies	22/11/2007	
22	Female	Industrial	12/08/2005	

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab

□ Semicolon

□ Comma

□ Space

☑ Other: /

□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
Start Date
02          10    2009
31          03    2008
12          07    2009
25          12    2014
19          09    2002
```

Or you can use the following formulas to get "day", "month", and "year" separately:

=day()   =month()   =year()

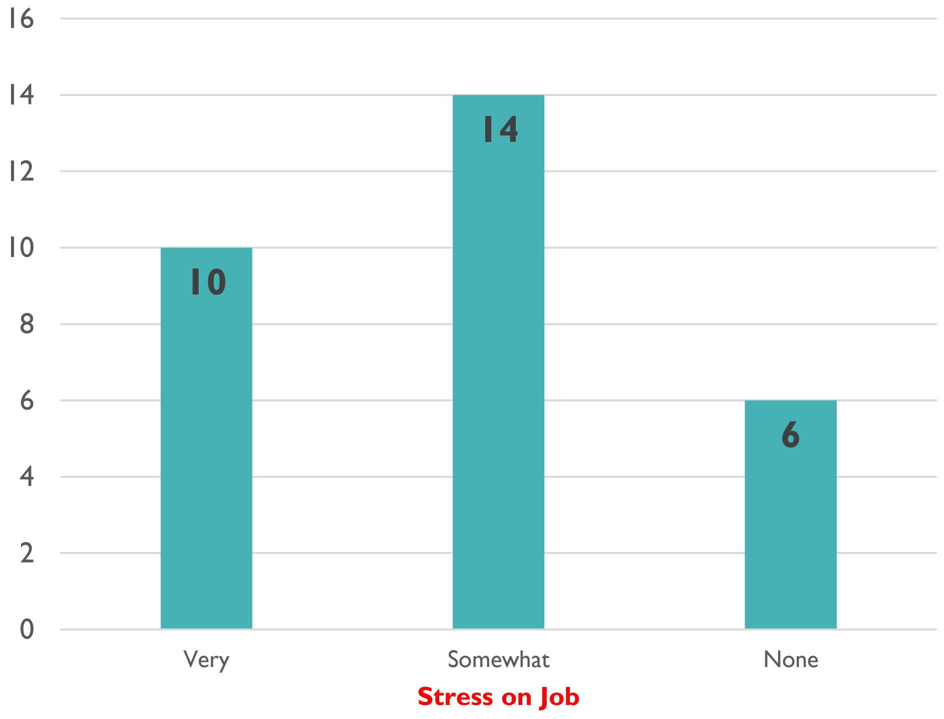
Chapter 2
Presentation of Data

2.1 Categorical Data

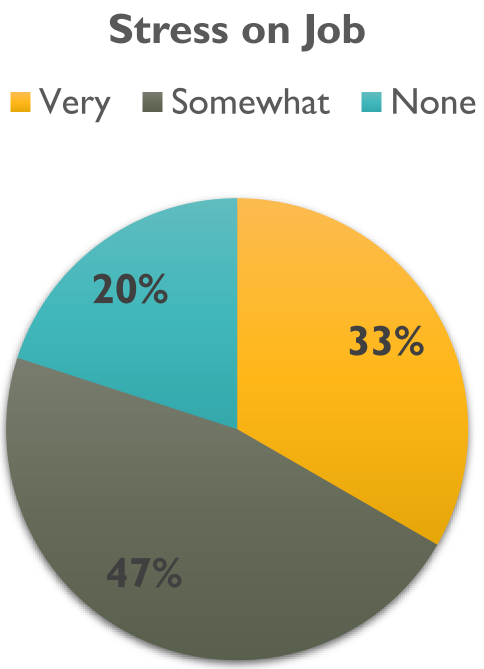
- *Bar Chart*
  Bar charts are familiar to most people. They conveniently summarize information from a table in a clear and illustrative manner.

  Example 2.1: Frequency distribution for sample of 30 employees from large companies was selected, and these employees were asked how stressful their jobs were. The responses of these employees are recorded below, where very represents very stressful, somewhat means somewhat stressful, and none stands for not stressful at all.

Stress on Job	Frequency	Percentage%
Very	10	33%
Somewhat	14	47%
None	6	20%
Total	**30**	**100%**

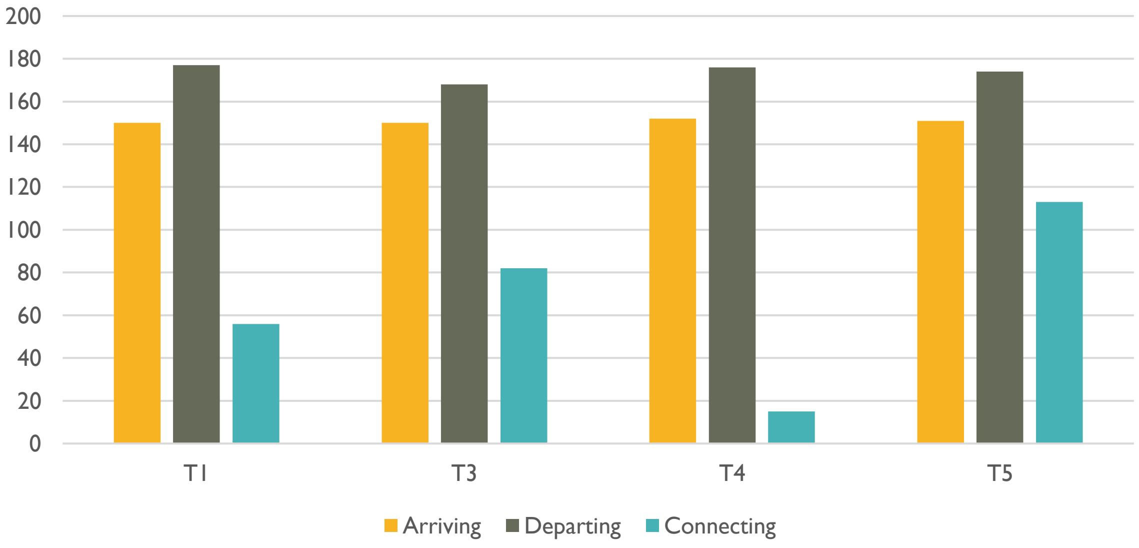


- *Pie Chart*



Example 2.2: Number of Passenger flow (Heathrow).

Passenger Flow	**LHR**				**Total**
	T1	**T3**	**T4**	**T5**	
Arriving	150	150	152	151	**603**
Departing	177	168	176	174	**695**
Connecting	56	82	15	113	**266**
Total	**383**	**400**	**343**	**438**	**1,564**

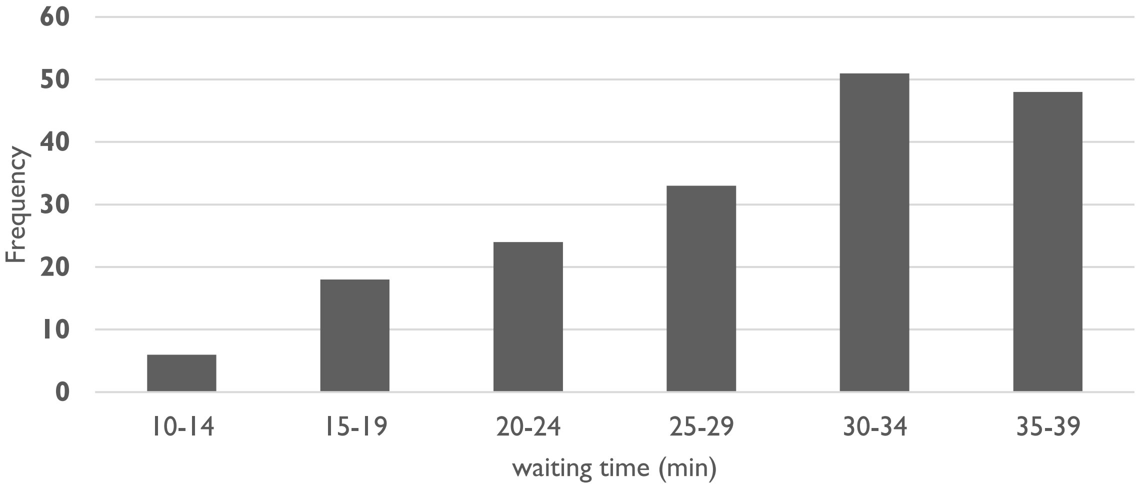


2.2 Numerical Data

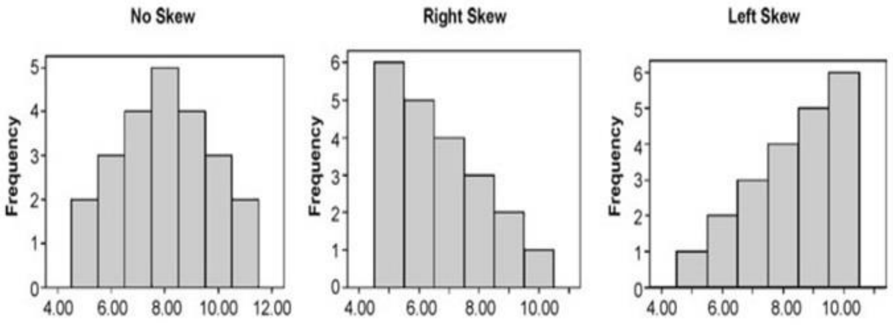
- *Histograms*

Example 2.3: sample for the waiting time (min) at the airport.

Waiting Time (min)	Frequency	Percentage%
10-14	6	3%
15-19	18	10%
20-24	24	13%
25-29	33	18%
30-34	51	29%
35-39	48	27%
Total	**180**	**100%**



The density curve is “Skewed to the left” which implies that the passengers with long waiting times are more frequent than those with short Waiting times. The following graphs shows some other density curve skewness:



Three histograms illustrating skewness.

The first decision is: How many bars should you use?

– The histogram should normally have 3–13 bars.

– The more the observations, the more the bars.

Technical Note

To determine the number of bars, we can use the following formula:

$$No. of\ bars = \frac{\log(n)}{\log(2)}$$

Here *n* is the number of values and log is the logarithmic function.

For example, if $n = 30$. Here we use logarithms of base 10:

$$No. of\ bars = \frac{\log(n)}{\log(2)} = \frac{\log(30)}{\log(2)} = \frac{1.48}{0.30} = 4.9\ (app. 5)$$

The next question is: How wide should the bars be?

Having determined the number of bars, we can easily find out how wide each bar must be:

1. Interval length = $\frac{(Maximum\ value - Minimum\ value)}{Number\ of\ bars}$

2. Round off the result to an appropriate number, if necessary.

For example, if the maximum value = 198, the minimum value = 97 and the number of bars = 5. Then,

Interval length = $\frac{(198-97)}{5} = 20.2\ (app. 20)$

Problem 2.1: Write a frequency table on Excel for a sample of 25 passengers waiting time (min) at the airport. Draw the histogram.

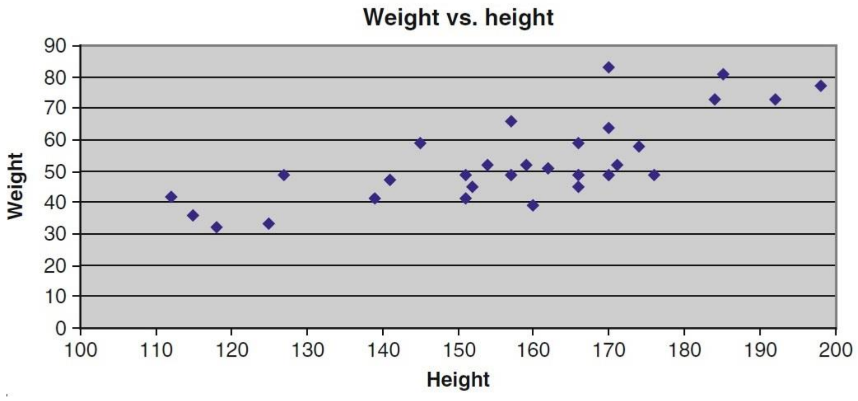
15	21	15	32	21
17	22	20	29	19
18	23	24	27	30
19	24	26	28	38
19	16	26	31	40

- *Scatter Plots*

Scatter plots are well suited to show relationships between two variables.

Example 2.4: Suppose in a “Fitness Club”, we assume that there is a relationship between height and weight: the taller a kid is, the heavier it is. A scatter plot is illustrated in Fig. 2.1.

Fig. 2.1 relationship between height and weight Suppose in a Fitness Club



– Weight is the Y variable or the dependent variable. Height is the X-variable or the independent variable. We imagine that weight depends on the height, i.e., there is a “cause” and an “effect.”
– In other cases, it is more arbitrary, as to which variable we choose as X and Y. We simply imagine that there must be a relationship (or correlation), without necessarily a “cause” and an “effect.”
– Chapter 5 gives tools to investigate whether there is indeed a statistical correlation between the two variables.

Problem 2.2: An auto manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The research department at the company took a sample of eight cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

Problem 2.3: Open “education” file in excel and answer the followings:

1- How many male and female students?
2- Frequency table for level of education. (Draw the pie chart).
3- Frequency table for level of education by gender. (Draw the bar chart).

Chapter 3
Description of Data

3.1 Measures of Location

Average

The average is a measure of the center in the distribution of data values. The average is calculated as the sum of all the data values divided by their number.

The average is highly influenced by “extreme values” (i.e., very large or very small values). If for example, there are many very large data values, the average becomes “excessively” large. An alternative is to use the “median” rather than the average.

To calculate the average, use the statistical function AVERAGE.

Example: we want to calculate the average waiting time (by minuets) for five passengers: 20, 25, 19, 23, and 50.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{20+25+19+23+50}{5} = \frac{137}{5} = 27.4 \text{ (Min)}$$

Median

The median is the data value “in the middle”, i.e., a number that divides the data values into two parts with an equal number of values.

The median can be found by first sorting the data values in ascending order.

– In case of an odd number of data values, the median is the middle value.

– In case of an even number of data values, there is no single data value dividing data values into two equally large parts; we then define the median as the average of the two middle values.

The median is not as sensitive to extreme values as the average! Often, you will therefore complement the average with the median.

The statistical function is called MEDIAN.

Example: we want to calculate the median for the waiting time (by minuets) for five passengers: 20, 25, 19, 23, and 50.

First, we sort the data in ascending order: 19, 20, **23**, 25, 50.

Therefore, the median is 23.

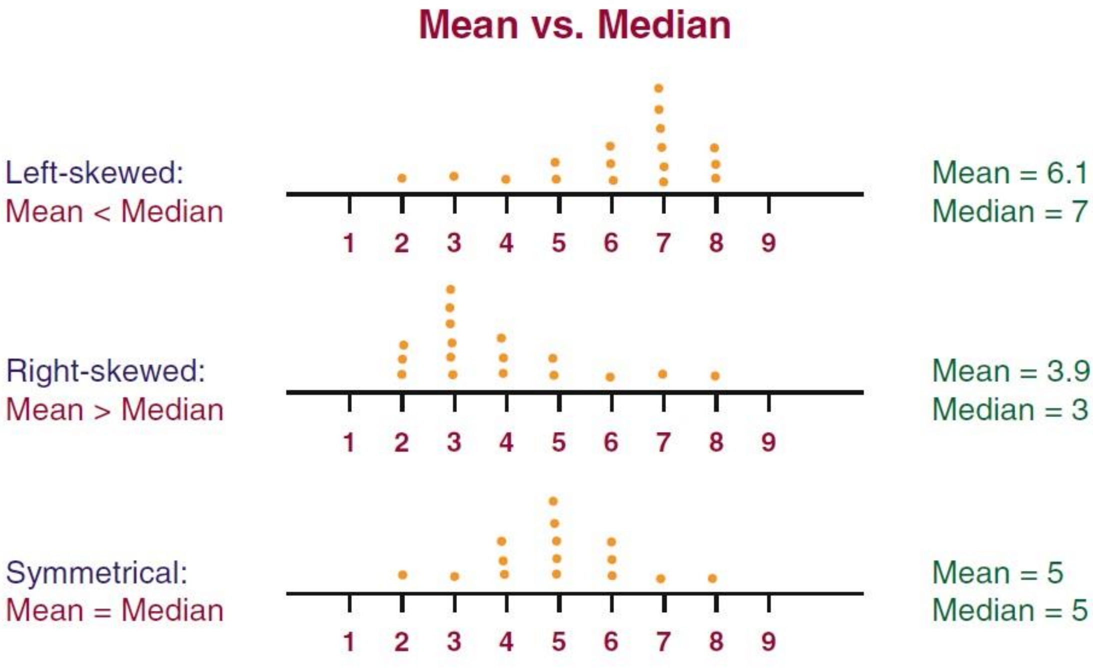
Choosing a Measure of Location

If the distribution is symmetrical, i.e., there are an equal number of large and small data values, you will usually use the average, but the median provides virtually the same result.

In a skewed (i.e., non-symmetrical) distribution, the average and the median are not identical:

– For a right-skewed distribution, i.e., a distribution with many large data values, the average is larger than the median.

– For a left-skewed distribution, i.e., a distribution with many small data values, the average is smaller than the median. (See Figure. 3.4).

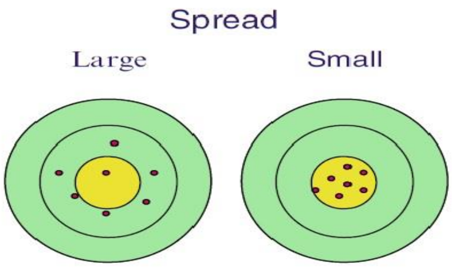


3.2 Measures of Dispersion.

In this chapter, we have reviewed the main measures of location, i.e., the center of a distribution. Now we look at various measures of dispersion, i.e., the spread of a distribution.

Standard Deviation.

The standard deviation is “the average distance” between the data values and the average. The larger the standard deviation, the larger is the spread of the distribution.



To calculate the standard deviation, we use the function STDEV.

Example: calculate the standard deviation for the waiting time (by minuets) for five passengers: 20, 25, 19, 23, and 50.

12.9 (=STDEV.S())

Interquartile Range.

Another important measure of dispersion is the Interquartile Range (IQR), explained below:

When the median is calculated, you can further divide the two parts of the data values into two parts each. Thus the entire set of data values is divided into four parts, with (roughly) the same number of data values. The new points of division are called the “quartiles”. The difference between the quartiles is called the interquartile range, often denoted by the abbreviation IQR. The interpretation of IQR is that it is the length of the interval with the “middle 50%” of the data values and it is often used when you use the median as a measure of location.

When we find the quartiles, we sort the data values in ascending order, as when calculating the median.

The lower quartile (or first quartile) Q_1 is a number that divides the data values into two parts so that one-fourth of the data values are smaller than the lower quartile, and three-fourths of the data values are larger.

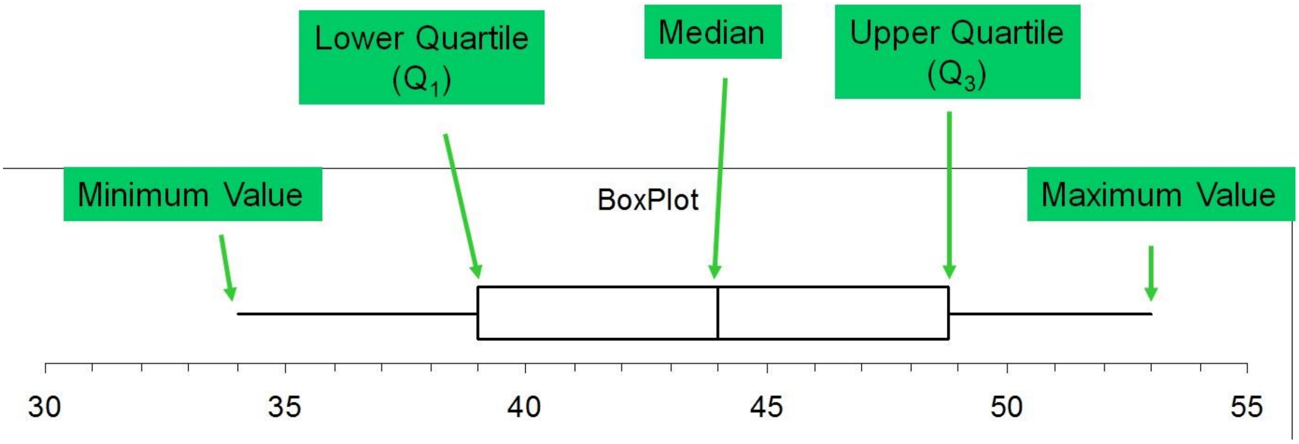
Often, the median is considered to be the middle or second quartile and is sometimes denoted by Q_2.

The upper quartile (or third quartile) Q_3 is a figure that divides the data values into two parts so that three-fourths of the data values are smaller than the upper quartile and one-fourth of the data values are larger.

The interquartile range IQR is then the difference between the upper and lower quartiles:

$$IQR = Q_3 – Q_1$$

The quartiles are calculated using the function QUARTILE.EXC().



Example: calculate the lower and upper quartile then find the Interquartile range for the waiting time (by minuets) for five passengers: 20, 25, 19, 23, and 50.

Q_1=19.5

Q_3= 37.5

IQR = 18

3.3 Outlier

An outlier is an observation whose value, x, either exceeds the value of the third quartile by a magnitude greater than 1.5(IQR) or is less than the value of the first quartile by a magnitude greater than 1.5(IQR).

That is, an observation of $x > Q_3 + 1.5(\mathrm{IQR})$ or an observation of $x < Q_1 - 1.5(\mathrm{IQR})$ is called an outlier.

Let us name the value for $Q_3 + 1.5(\mathrm{IQR})$ the upper limit and the value for $Q_1 - 1.5(IQR)$ the lower limit.

***Example*:** calculate the lower and upper limit for outlier observation for the following data which represents the number of passengers complaints recorded on Saturday for the last 16 weeks.

12	15	12	12	25	9	14	15
11	10	13	12	10	12	11	10

***Solution*:**

$Q_1 = 10.25\,, Q_3 = 13.75\,,\ \mathrm{IQR} = 3.5$

The lower limit: $Q_1 - 1.5(IQR) = 5$

The upper limit: $Q_3 + 1.5(\mathrm{IQR}) = 19$

Then we conclude that from the given data, the value 25 is considered an outlier.

Note: By using Excel, we could find the Descriptive statistics for data set.

The steps: *Data* ⟶ *Data Analysis* ⟶ *Descriptive statistics*

***Example*:** Find the Descriptive statistics for the previous passenger's complaints record example.

***Solution*:**

passenger's complaints	
Mean	12.6875
Standard Error	0.92969
Median	12
Mode	12
Standard Deviation	3.718759
Sample Variance	13.82917
Kurtosis	8.482342
Skewness	2.628153
Range	16
Minimum	9
Maximum	25
Sum	203
Count	16

Problem 3.1: The following data represent the number of customer's peak hour for 20 days. Use Excel to find the Descriptive statistics and explain the outliers (if any).

1,723	1,499
1,741	1,542
1,567	1,929
1,564	1,787
1,554	1,737
1,135	1,679
1,911	1,802
1,780	1,873
1,559	1,617
1,771	1,631

Problem 3.2: Open "education" file in excel and answer the followings:

Q1. Draw the box plot for the math, reading and writing score.

Q2. Complete the missing values on the following table:

	math score	**reading score**	**writing score**
Average			
Median			
Standard Deviation			
IQR			

What are the appropriate measures of location and dispersion for Math score, reading score and writing score?

Problem 3.3: Open “employees” file in excel and answer the followings:

Q1. Add the department location to the data (hint: VLOOKUP).

Q2. What are the total salaries for each department?

Q3. Sort the data according to employee salary from highest to lowest.

Q.4 add filters to the columns (variables).

Q.5 do the following steps:

Step1: create new column to salary and add a 100 KD “bonus” to the salary for Married employee.

Step2: create another new column to salary and add a 100 KD “bonus” to the salary for Married male employee only.

Finally, calculate the change percentage in the total salary for the above two steps.

Q.6 complete the following table (find the number of employees in each cell)

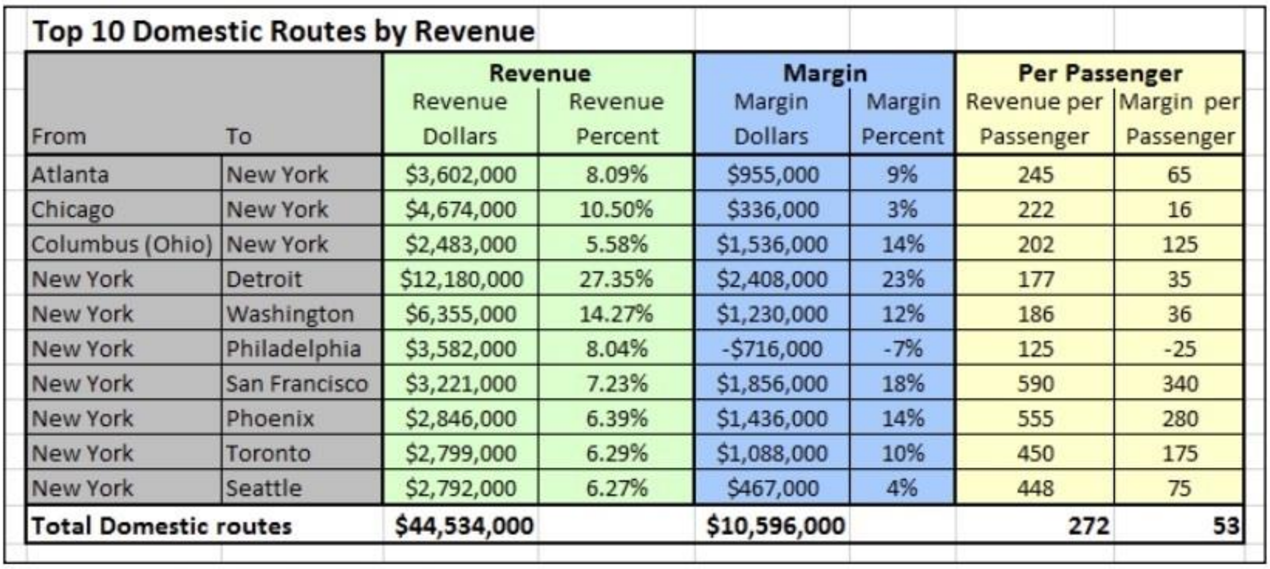
Department	**Gender**		**Total**
	M	**F**	
administration			
engineering			
Finance			
information technology			
planning			
Total			

Q.7 What is the percentage of male?

Q.8 What is the percentage of female in Finance department?

Chapter 4
Reports with Conditional Formatting

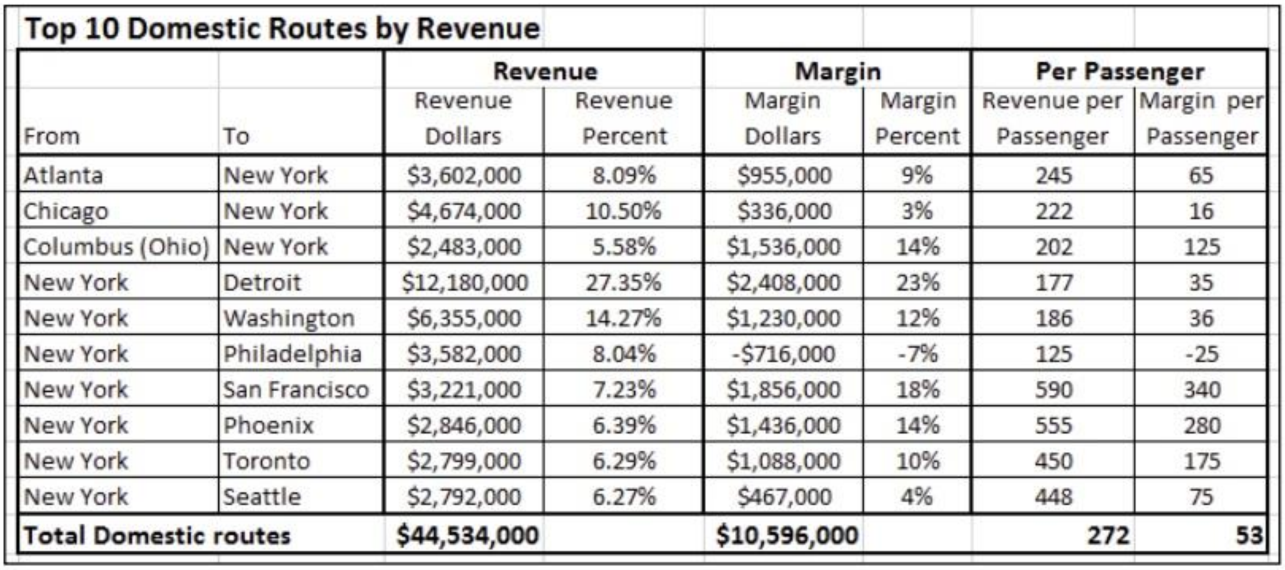
This table is similar to many found in Excel reports (The thick borders, the different colors)



Top 10 Domestic Routes by Revenue

		Revenue		Margin		Per Passenger	
From	To	Revenue Dollars	Revenue Percent	Margin Dollars	Margin Percent	Revenue per Passenger	Margin per Passenger
Atlanta	New York	$3,602,000	8.09%	$955,000	9%	245	65
Chicago	New York	$4,674,000	10.50%	$336,000	3%	222	16
Columbus (Ohio)	New York	$2,483,000	5.58%	$1,536,000	14%	202	125
New York	Detroit	$12,180,000	27.35%	$2,408,000	23%	177	35
New York	Washington	$6,355,000	14.27%	$1,230,000	12%	186	36
New York	Philadelphia	$3,582,000	8.04%	-$716,000	-7%	125	-25
New York	San Francisco	$3,221,000	7.23%	$1,856,000	18%	590	340
New York	Phoenix	$2,846,000	6.39%	$1,436,000	14%	555	280
New York	Toronto	$2,799,000	6.29%	$1,088,000	10%	450	175
New York	Seattle	$2,792,000	6.27%	$467,000	4%	448	75
Total Domestic routes		**$44,534,000**		**$10,596,000**		**272**	**53**

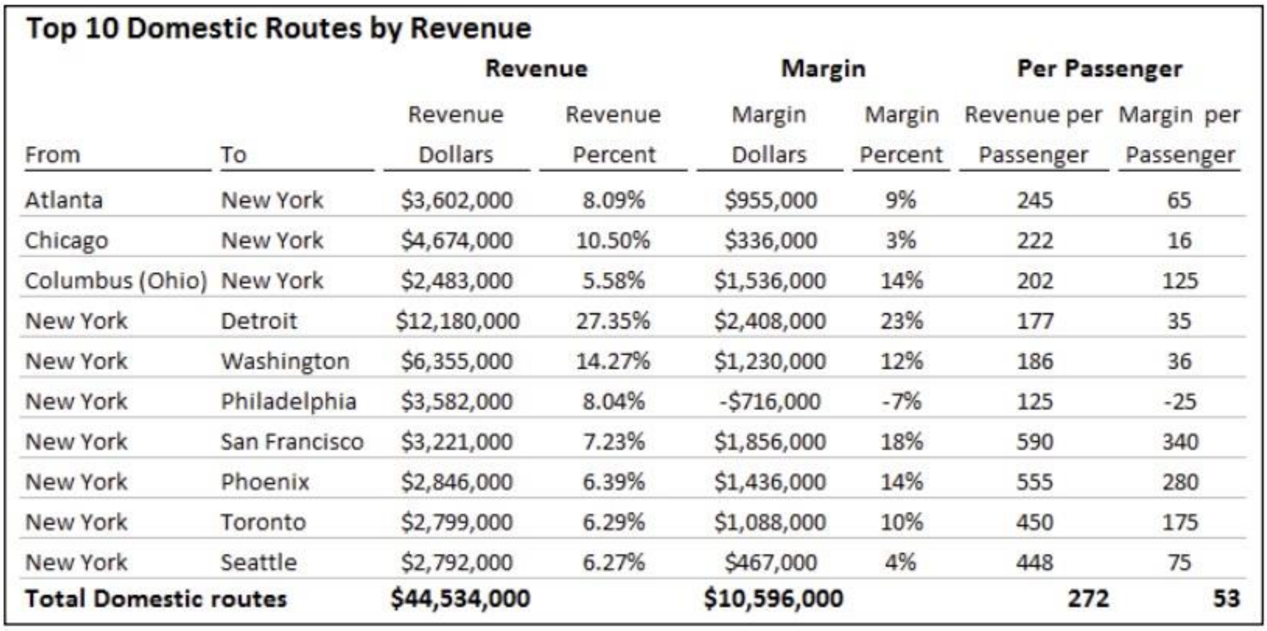
A table with removed colors. As you can see, it is already easier to read.



Top 10 Domestic Routes by Revenue

		Revenue		Margin		Per Passenger	
From	To	Revenue Dollars	Revenue Percent	Margin Dollars	Margin Percent	Revenue per Passenger	Margin per Passenger
Atlanta	New York	$3,602,000	8.09%	$955,000	9%	245	65
Chicago	New York	$4,674,000	10.50%	$336,000	3%	222	16
Columbus (Ohio)	New York	$2,483,000	5.58%	$1,536,000	14%	202	125
New York	Detroit	$12,180,000	27.35%	$2,408,000	23%	177	35
New York	Washington	$6,355,000	14.27%	$1,230,000	12%	186	36
New York	Philadelphia	$3,582,000	8.04%	-$716,000	-7%	125	-25
New York	San Francisco	$3,221,000	7.23%	$1,856,000	18%	590	340
New York	Phoenix	$2,846,000	6.39%	$1,436,000	14%	555	280
New York	Toronto	$2,799,000	6.29%	$1,088,000	10%	450	175
New York	Seattle	$2,792,000	6.27%	$467,000	4%	448	75
Total Domestic routes		**$44,534,000**		**$10,596,000**		**272**	**53**

Notice the following table how the numbers are no longer caged in gridlines. Also, headings now jump out at you with the addition of Single Accounting underlines. (Minimize the use of borders and use the Single Accounting underlines)



Top 10 Domestic Routes by Revenue

		Revenue		**Margin**		**Per Passenger**	
From	To	Revenue Dollars	Revenue Percent	Margin Dollars	Margin Percent	Revenue per Passenger	Margin per Passenger
Atlanta	New York	$3,602,000	8.09%	$955,000	9%	245	65
Chicago	New York	$4,674,000	10.50%	$336,000	3%	222	16
Columbus (Ohio)	New York	$2,483,000	5.58%	$1,536,000	14%	202	125
New York	Detroit	$12,180,000	27.35%	$2,408,000	23%	177	35
New York	Washington	$6,355,000	14.27%	$1,230,000	12%	186	36
New York	Philadelphia	$3,582,000	8.04%	-$716,000	-7%	125	-25
New York	San Francisco	$3,221,000	7.23%	$1,856,000	18%	590	340
New York	Phoenix	$2,846,000	6.39%	$1,436,000	14%	555	280
New York	Toronto	$2,799,000	6.29%	$1,088,000	10%	450	175
New York	Seattle	$2,792,000	6.27%	$467,000	4%	448	75
Total Domestic routes		**$44,534,000**		**$10,596,000**		**272**	**53**

Subduing headers and labels. Note how the data now becomes the focus of attention, whereas the muted labels work in the background.

4.1 Creating Sparklines

Example 4.1

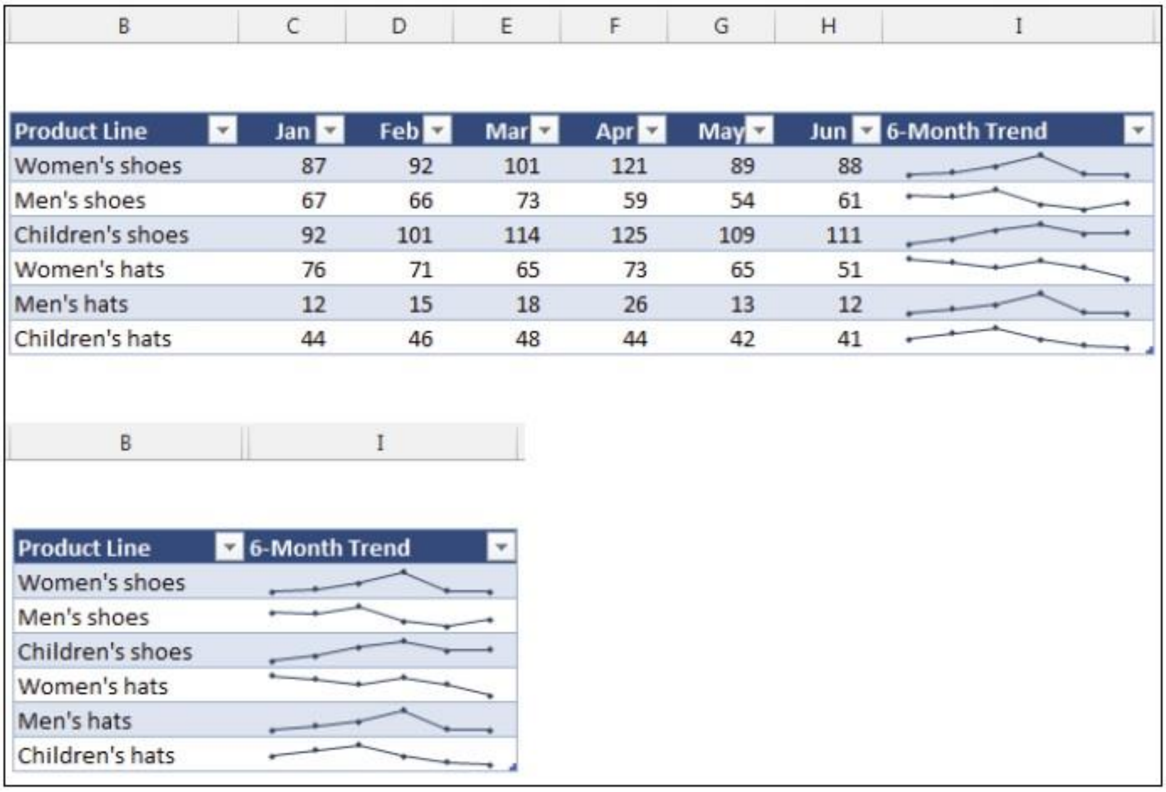
Steps:

1- Select the data range that you want to summarize. In this example, select B2:G7. If you are creating multiple sparklines, select all the data.

2- With the data selected, click the Insert tab on the Ribbon and find the Sparklines group. There you can select any one of the three sparkline types: Line, Column, or Win/Loss. In this case, select the Column option.

3- Excel displays the Create Sparklines dialog box, as shown

4- Specify the data range and the location for the sparklines. For this example, specify H2:H7 as the Location Range.



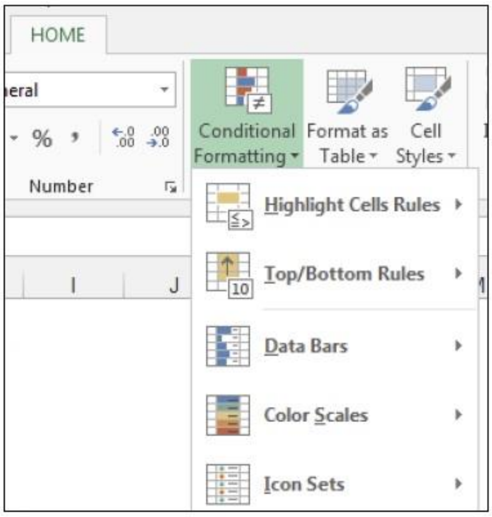
Product Line	Jan	Feb	Mar	Apr	May	Jun	6-Month Trend
Women's shoes	87	92	101	121	89	88	
Men's shoes	67	66	73	59	54	61	
Children's shoes	92	101	114	125	109	111	
Women's hats	76	71	65	73	65	51	
Men's hats	12	15	18	26	13	12	
Children's hats	44	46	48	44	42	41	

Product Line	6-Month Trend
Women's shoes	
Men's shoes	
Children's shoes	
Women's hats	
Men's hats	
Children's hats	

4.2 Applying basic conditional formatting

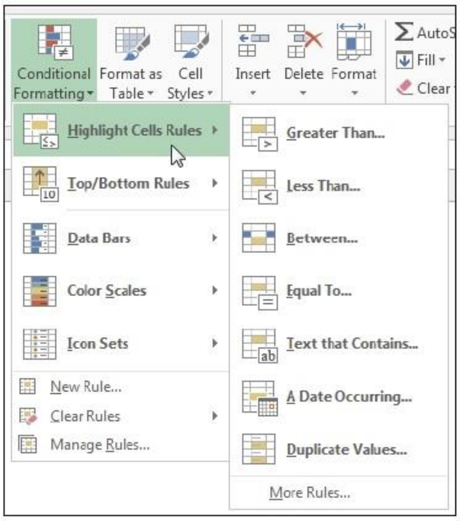
As we can see, five categories of predefined options are available:

- Highlight Cells Rules
- Top/Bottom Rules
- Data Bars
- Color Scales
- Icon Sets



Using Highlight Cells Rules

The formatting options in the Highlight Cells Rules category allow you to highlight those cells whose values meet a specific condition.



The options in the Highlight Cells Rules category are pretty self-explanatory:

➤ Greater Than: Allows you to conditionally format a cell whose value is greater than a specified amount.

For instance, you can tell Excel to format those cells that contain a value greater than 50.

➤ Less Than: Allows you to conditionally format a cell whose value is less than a specified amount.

For instance, you can tell Excel to format those cells that contain a value less than 100.

➤ Between: Allows you to conditionally format a cell whose value is between two given amounts.

For example, you can tell Excel to format those cells that contain a value between 50 and 100.

➤ Text That Contains: Allows you to conditionally format a cell whose contents contain any form of a given text you specify as a criterion.

For example, you can tell Excel to format the cells that contain the text North.

➤ A Date Occurring: Allows you to conditionally format a cell whose contents contain a date occurring in a specified period relative to today’s date.

For example, Yesterday, Last Week, Last Month, Next Month, Next Week, and so on.

➤ Duplicate Values: Allows you to conditionally format both duplicate values and unique values in a given range of cells.

Example 4.2: how to apply one of these options. To highlight all values greater than a certain amount, follow these steps:

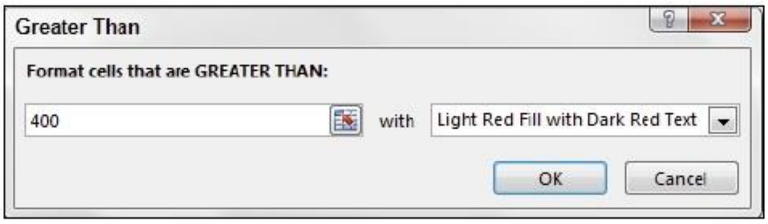
1. Select the range of cells to which you need to apply the conditional formatting.

2. In the Highlight Cells Rules category, choose the Greater Than option. The Greater Than dialog box opens.

● Type the value (400 in this example).

● Reference a cell that contains the trigger value.

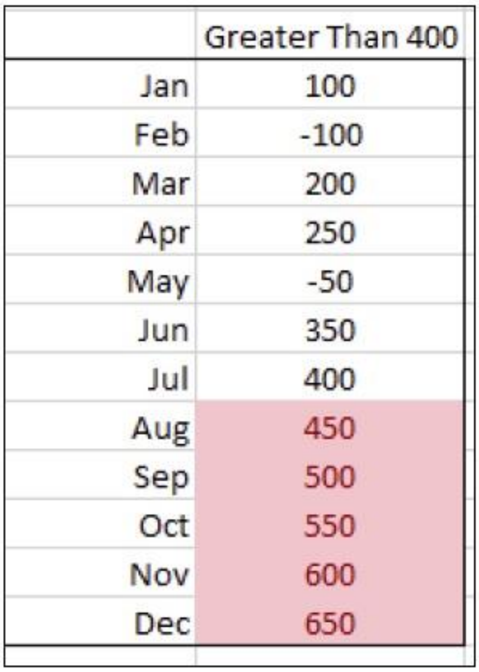
Also in this dialog box, you can use the drop-down menu to specify the format you want applied.



Greater Than

Format cells that are GREATER THAN:

400 with Light Red Fill with Dark Red Text

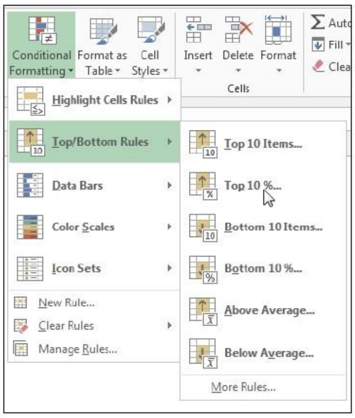


	Greater Than 400
Jan	100
Feb	-100
Mar	200
Apr	250
May	-50
Jun	350
Jul	400
Aug	450
Sep	500
Oct	550
Nov	600
Dec	650

Cells greater than 400 are formatted (Aug to Dec).

Applying Top/Bottom Rules

The formatting options in the Top/Bottom Rules category allow you to highlight those cells whose values meet a given threshold**.**



You can select from these options:

➤ Top 10 Items: Allows you to specify any number of cells to highlight based on individual cell values (not just 10 cells).

For example, you can highlight the cells whose values are the 5 largest numbers of all the cells selected.

➤ Top 10%: Allows you to specify any percentage of cells to highlight based on individual cell values (not just 10 percent) option.

For instance, you can highlight the cells whose values make up the top 20 percent of the total values of all the selected cells.

➤ Bottom 10 Items: Allows you to specify the number of cells to highlight based on the lowest individual cell values (not just 10 cells).

For example, you can highlight the cells whose values are within the 15 smallest numbers among all the cells selected.

➤ Bottom 10%: Allows you to specify any percentage of cells to highlight based on individual cell values (not just 10 percent).

For instance, you can highlight the cells whose values make up the bottom 15 percent of the total values of all the selected cells.

➤ Above Average: Allows you to conditionally format each cell whose value is above the average of all cells selected.

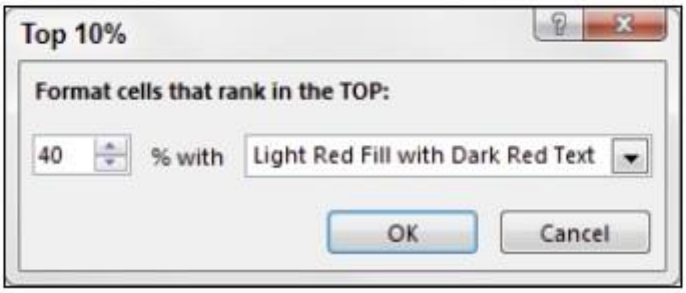
➤ Below Average: Allows you to conditionally format each cell whose value is below the average of all cells selected.

Example 4.3: you conditionally format all cells whose values are within the top 40 percent of the total values of all cells.

1. Select the range of cells to which you need to apply the conditional formatting.

2. In the Top/Bottom Options category, choose Top 10%. The Top 10% dialog box opens, here you define the threshold that that will trigger the conditional formatting.

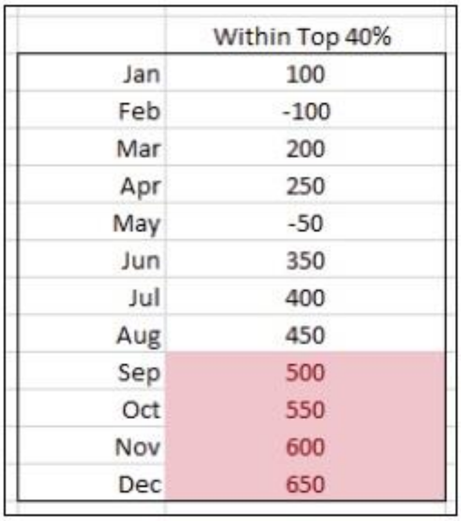
3. In this example, enter 40. Here you can also use the drop-down menu to specify the format you want to apply.



Top 10%

Format cells that rank in the TOP:

40 % with Light Red Fill with Dark Red Text



	Within Top 40%
Jan	100
Feb	-100
Mar	200
Apr	250
May	-50
Jun	350
Jul	400
Aug	450
Sep	500
Oct	550
Nov	600
Dec	650

With conditional formatting, you can easily see that September through December makes up 40 percent of the total value in this dataset.

Creating Data Bars

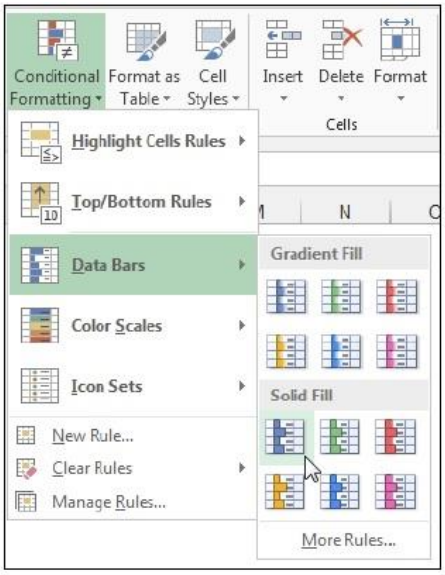
Data Bars fill each cell you're formatting with mini-bars in varying length, indicating the value in each cell relative to other formatted cells. Excel essentially takes the largest and smallest values in the selected range and calculates the length for each bar.

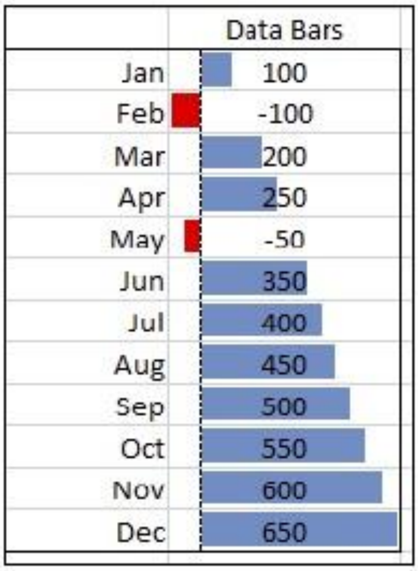
To apply Data Bars to a range, do the following:

1. Select the target range of cells to which you need to apply the conditional formatting.

2. Click the Home tab and choose Conditional Formatting ➜ Data Bars.

Example 4.4:





	Data Bars
Jan	100
Feb	-100
Mar	200
Apr	250
May	-50
Jun	350
Jul	400
Aug	450
Sep	500
Oct	550
Nov	600
Dec	650

Applying Color Scales

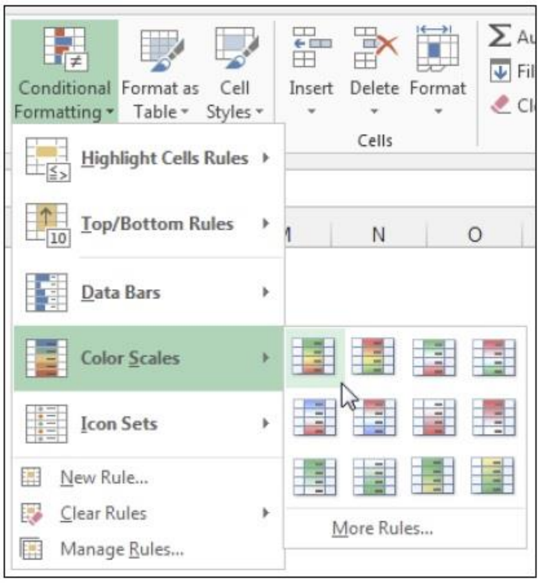
Color Scales fill each cell you're formatting with a color, varying in scale based on the value in each cell relative to other formatted cells.

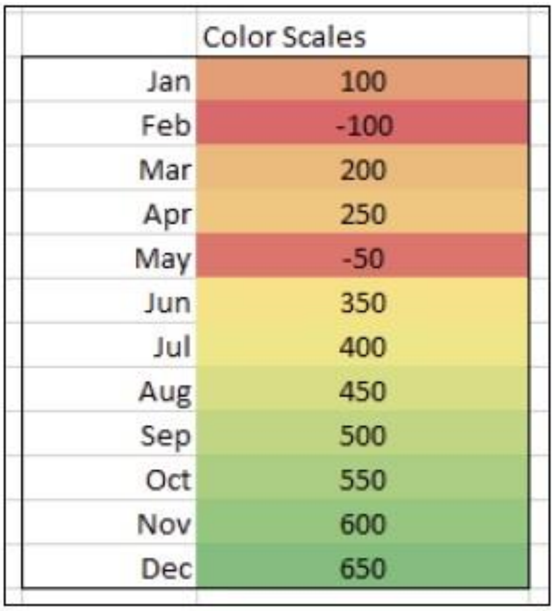
To apply Color Scales to a range, do the following:

1. Select the target range of cells to which you need to apply the conditional formatting.

2. Click the Home tab and choose Conditional Formatting ➜ Color Scales.

Example 4.5:





	Color Scales
Jan	100
Feb	-100
Mar	200
Apr	250
May	-50
Jun	350
Jul	400
Aug	450
Sep	500
Oct	550
Nov	600
Dec	650

Using icon sets

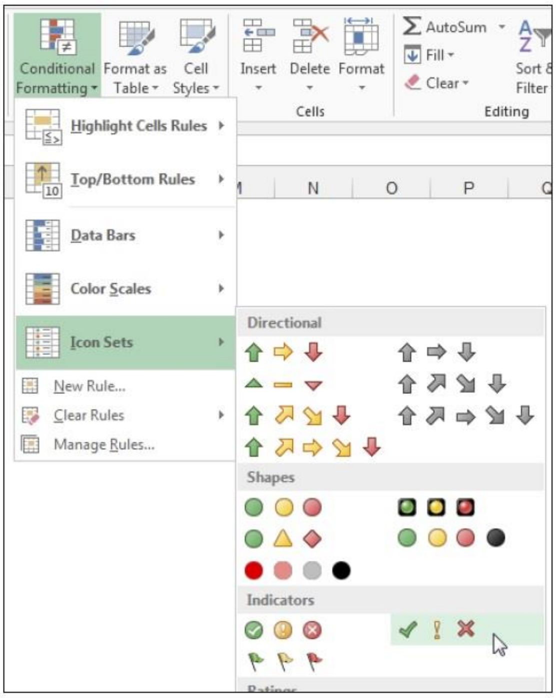
Icon sets are sets of symbols that are inserted in each cell you're formatting. Excel determines which symbol to use based on the value in each cell relative to other formatted cells.

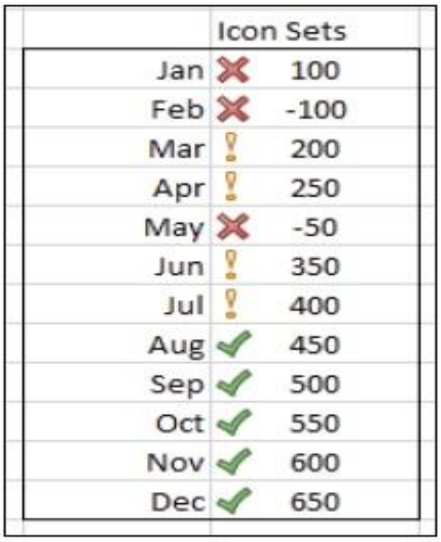
To apply an icon set to a range, do the following:

1. Select the target range of cells to which you need to apply the conditional formatting.

2. Click the Home tab and choose Conditional Formatting ➜ Icon Sets.

Example 4.6:



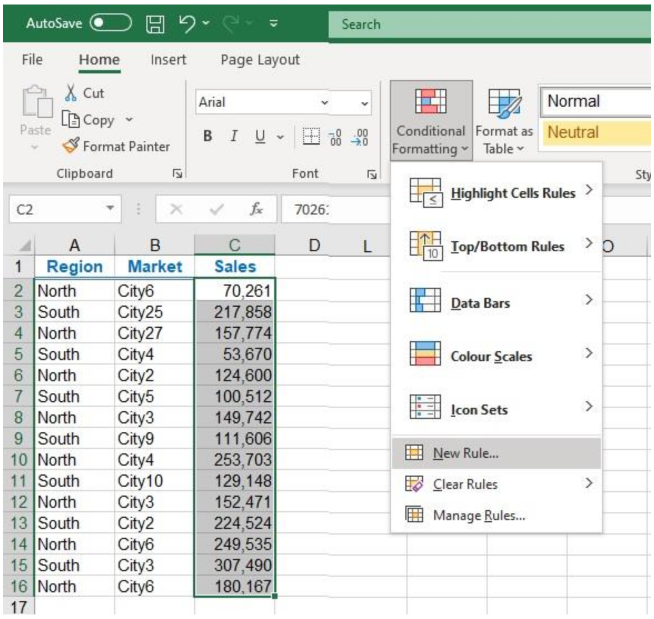


	Icon Sets	
Jan	✗	100
Feb	✗	-100
Mar	!	200
Apr	!	250
May	✗	-50
Jun	!	350
Jul	!	400
Aug	✓	450
Sep	✓	500
Oct	✓	550
Nov	✓	600
Dec	✓	650

Adding your own formatting rules manually

Example 4.7:

Select the target range of cells to which you need to apply the conditional formatting and select New Rule.

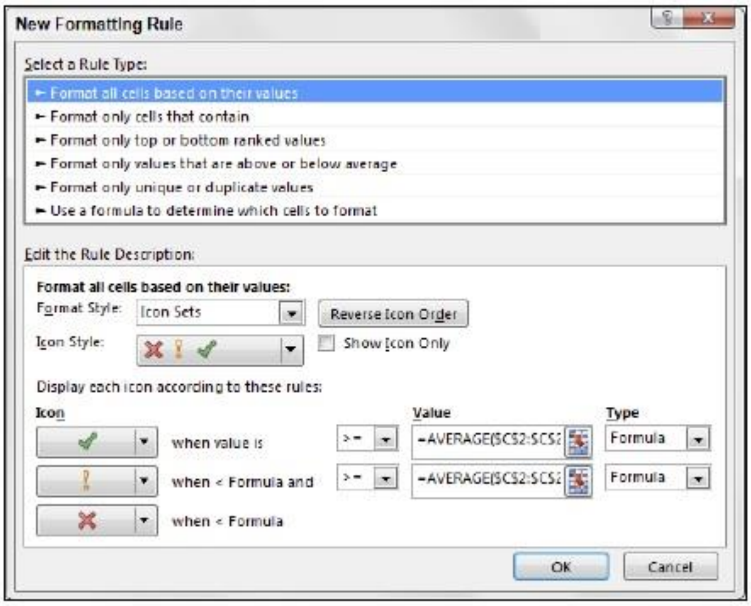


	A	B	C
1	Region	Market	Sales
2	North	City6	70,261
3	South	City25	217,858
4	North	City27	157,774
5	South	City4	53,670
6	North	City2	124,600
7	South	City5	100,512
8	North	City3	149,742
9	South	City9	111,606
10	North	City4	253,703
11	South	City10	129,148
12	North	City3	152,471
13	South	City2	224,524
14	North	City6	249,535
15	South	City3	307,490
16	North	City6	180,167
17			

- **Format All Cells Based on Their Values:** Measures the values in the selected range against each other. This selection is handy for finding general anomalies in your dataset.
- **Format Only Cells That Contain**: Applies conditional formatting to those cells that meet specific criteria you define. This selection is perfect for comparing values against a defined benchmark.
- **Format Only Top or Bottom Ranked Values:** Applies conditional formatting to those cells that are ranked in the top or bottom nth number or percent of all the values in the range.
- **Format Only Values That Are Above or Below the Average:** Applies conditional formatting to those values that are mathematically above or below the average of all values in the selected range.
- **Format Only Unique or Duplicate Values:** Applies conditional formatting to cells that either contain values that are duplicated within the selected range or contain values are unique (not duplicated) within the selected range.
- **Use a Formula to Determine Which Cells to Format:** Evaluates values based on a formula you specify. If a particular value evaluates to true, then the conditional

formatting is applied to that cell. This selection is typically used when applying conditions based the results of an advanced formula or mathematical operation.

1. Ensure that the Format All Cells Based on Their Values rule is selected; then use the Format Style drop-down menu to switch to icon sets.
2. Click the Icon Style drop-down menu to select your desired icon set.
3. In the Type drop-down boxes, change both types to Formula.
4. In each Value box, enter **=Average($C$2:$C$16).**
   a. This tells Excel that the value in each cell must be greater than the average of the entire dataset in order to get the Check icon.
5. Click OK to apply your conditional formatting.


New Formatting Rule
Select a Rule Type:
► Format all cells based on their values
► Format only cells that contain
► Format only top or bottom ranked values
► Format only values that are above or below average
► Format only unique or duplicate values
► Use a formula to determine which cells to format
Edit the Rule Description:
Reverse Icon Order
OK
Cancel

Format all cells based on their values:

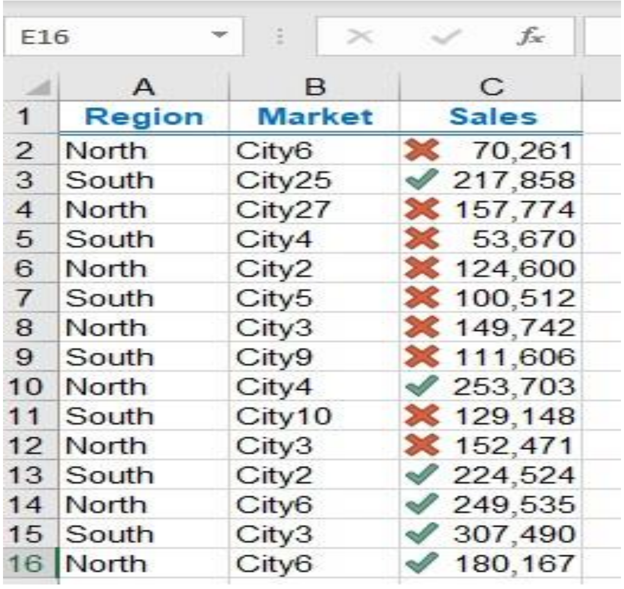
Format Style: Icon Sets

Icon Style: ✗ ! ✓

□ Show Icon Only

Display each icon according to these rules:

Icon			Value	Type
✓	when value is	>=	=AVERAGE($C$2:$C$:	Formula
!	when < Formula and	>=	=AVERAGE($C$2:$C$:	Formula
✗	when < Formula			



	A	B	C
1	Region	Market	Sales
2	North	City6	✗ 70,261
3	South	City25	✓ 217,858
4	North	City27	✗ 157,774
5	South	City4	✗ 53,670
6	North	City2	✗ 124,600
7	South	City5	✗ 100,512
8	North	City3	✗ 149,742
9	South	City9	✗ 111,606
10	North	City4	✓ 253,703
11	South	City10	✗ 129,148
12	North	City3	✗ 152,471
13	South	City2	✓ 224,524
14	North	City6	✓ 249,535
15	South	City3	✓ 307,490
16	North	City6	✓ 180,167

Show Data Bars and icons outside of cells

The Data Bars are, by default, placed directly inside each cell, almost obfuscating the data. From a dashboarding perspective, this is less than ideal for two reasons:

➤ The numbers can get lost in the colors of the Data Bars, making them difficult to read especially when printed in black and white.

➤ It's difficult to see the ends of each bar.

Example 4.8:

The answer to this issue is to show the Data Bars outside the cell that contains the value. Here's how:

1. To the right of each cell, enter a formula that references the cell that contains your data value.

For example, if your data is in C2, go to cell D2 and enter =C2.

2. Apply the Data Bar conditional formatting to the formulas you just created.

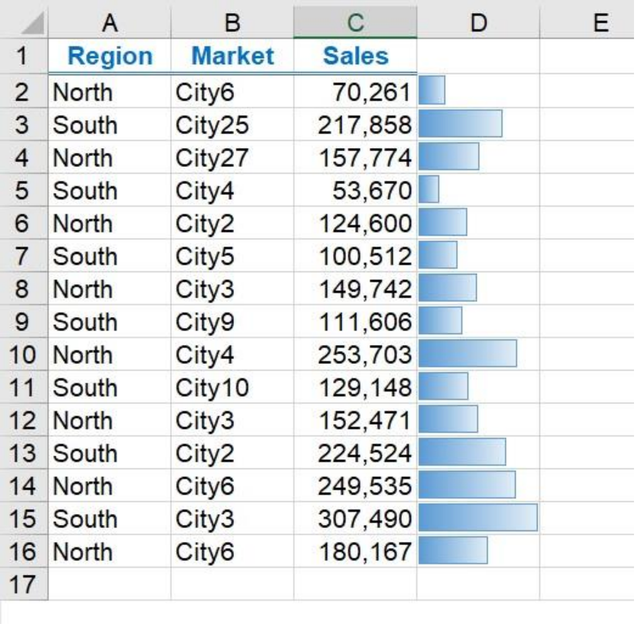
3. Select the formatted range of cells; then click the "Home" tab and select Conditional Formatting ➜ Manage Rules.

The Conditional Formatting Rules Manager dialog box opens.

4. Click the Edit Rule button.

5. Place a check in the Show Bar Only option, as demonstrated in Figure.

6. Click OK to apply your change.



	A	B	C	D	E
1	Region	Market	Sales		
2	North	City6	70,261		
3	South	City25	217,858		
4	North	City27	157,774		
5	South	City4	53,670		
6	North	City2	124,600		
7	South	City5	100,512		
8	North	City3	149,742		
9	South	City9	111,606		
10	North	City4	253,703		
11	South	City10	129,148		
12	North	City3	152,471		
13	South	City2	224,524		
14	North	City6	249,535		
15	South	City3	307,490		
16	North	City6	180,167		
17					

Chapter 5
Tests of Hypotheses

What is a Hypothesis?

A hypothesis is an educated guess or proposition about something in the world around you. It should be testable, either by experiment or observation.

- Average weight of new-born baby in Kuwait is 2750 gram
- New-born babies born to smoker mothers weigh on average smaller than those born to non-smoker mothers.

Statistical Hypothesis and Testing

- Statistical hypothesis is an assumption about the population parameter which may or may not be true.
- Hypothesis testing is the formal procedure by which we conduct a test of the hypothesis to determine if it is plausible or not.
- Goal of statistical hypothesis is to determine if what we see from the sample is just happened by chance or it is real.

Type of Hypothesis

- **Null hypothesis:** Represents the normal or usual or default situation. For example, normal body temperature is 36.7.
- **Alternative hypothesis:** represents the proposed research question, the opposite of the null hypothesis.
- The test is for the null, so we either reject the null or don't reject the null. Not rejecting the null doesn't mean it is true, rather there is no evidence against it

Note: We reject the Null hypothesis if the p-value of a test is less than 0.05. otherwise, we cannot reject the Null hypothesis.

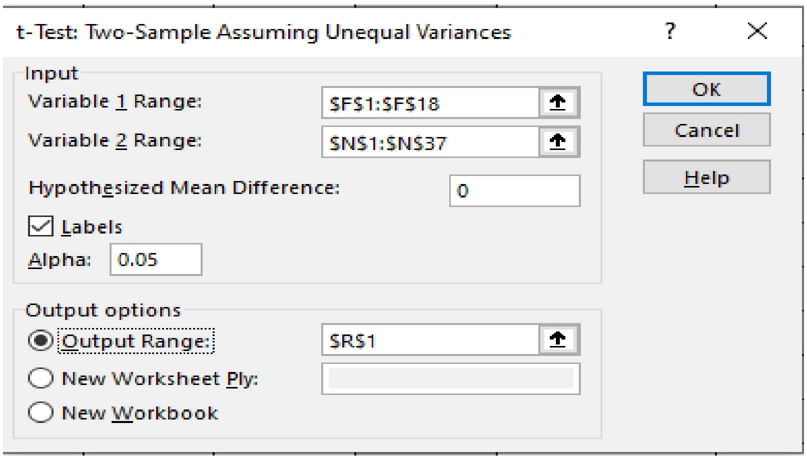
Example 5.1: (open "employees") test the following Hypothesis.

A. Is there a difference in mean salary between male and female?

Null hypothesis: There is no difference between the mean salary in the two groups.
Alternative hypothesis: there is significant difference between the mean salary in the two groups.

B. Is the mean salary for male is higher than female?

Null hypothesis: There is no difference between the mean salary in the two groups.
Alternative hypothesis: mean salary for male is significantly higher than mean salary for female.



t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $F$1:$F$18

Variable 2 Range: $N$1:$N$37

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

● Output Range: $R$1

○ New Worksheet Ply:

○ New Workbook

t-Test: Two-Sample Assuming Unequal Variances

	Salary-Male	*Salary-Female*
Mean	1589.470588	1560.5
Variance	53050.88971	46945.62857
Observations	17	36
Hypothesized Mean Difference	0	
df	30	
t Stat	0.435528014	
P(T<=t) one-tail	0.333148435	
t Critical one-tail	1.697260887	
P(T<=t) two-tail	0.66629687	
t Critical two-tail	2.042272456	

A. No significant difference (p-value = 0.67)

B. Mean salary for male is not significantly higher than female (p-value = 0.33)

Problem 5.1 (open "employees") test the following Hypothesis.

Is there a difference in mean age between male and female?

Chapter 6
Assessment of Relationship

6.1 Association between two quantitative variables

In many different disciplines you need to assess, whether there is a relationship between two variables. This can be in administration, social sciences, economics, industry, and science.

The purpose could be one of the following:

– To get a basic understanding of a subject area.

– To find reasons or explanations of phenomena.

– To try to predict future developments.

We study some techniques to assess a relationship and assess whether an apparent relationship is real or just a statistical coincidence. The technique is called regression analysis. We will only consider the basic technique, linear regression, which assumes that there is a linear relationship between two variables, i.e., a plot of Y against X shows a number of points scattered around a straight line.

One of the variables is the Y-variable or dependent variable. The other variable is the X-variable or the independent variable.

Regression: Regression analysis is helpful in assessing specific forms of the relationship between variables, and the ultimate objective when this method of analysis is employed usually is to predict or estimate the value of one variable corresponding to a given value of another variable.

Correlation: Correlation analysis is concerned with measuring the strength of the relationship between variables.

Main Goals:

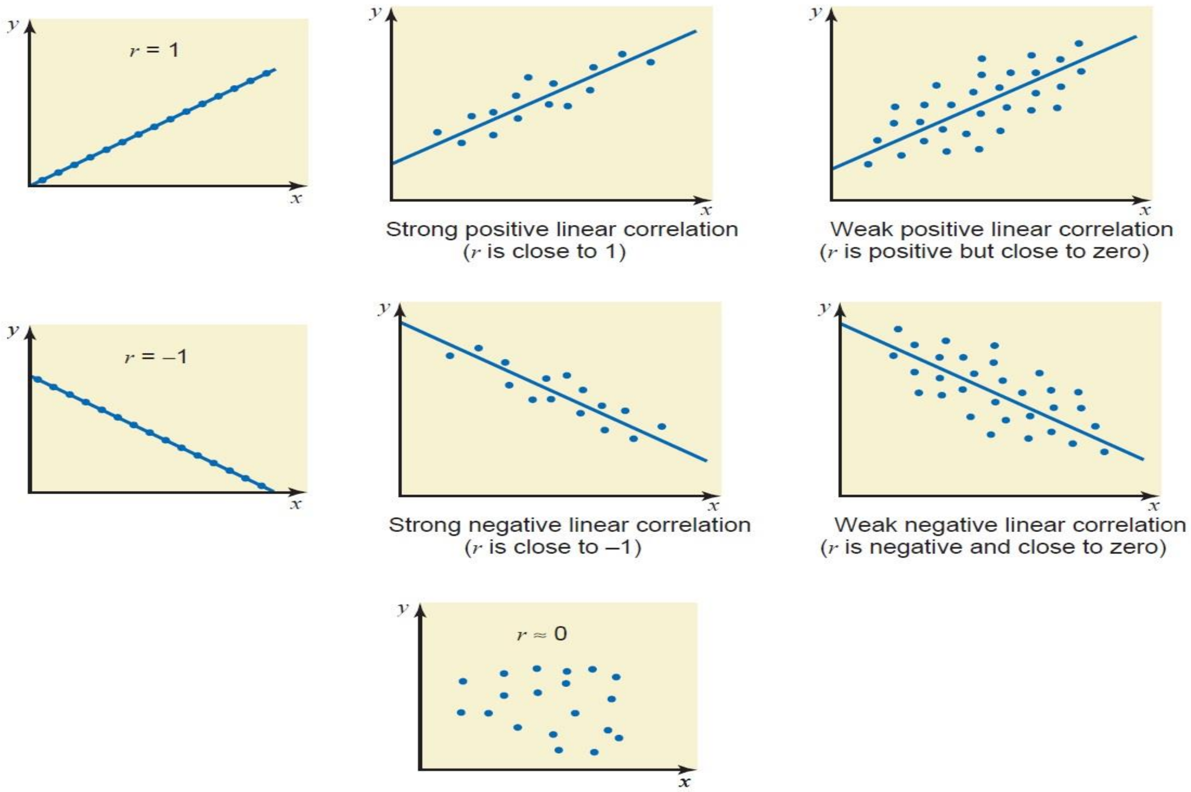
(1) Estimate the strength of such relationship – correlation

(2) Find the equation of the straight line that best fits this relationship – linear regression

Simple correlation coefficient (Pearson)

R=0	“No relation between x and y “
R=1	“Complete relation between x and y “
0≤R≤0.25	“Week relation between x and y”
0.25≤R≤0.75	“Moderate relation between x and y”
0.75≤R≤1	“Strong relation between x and y”

Note that for above cases, negative sign means reflexive relation and positive sign means direct positive relation.



Strong positive linear correlation (r is close to 1)

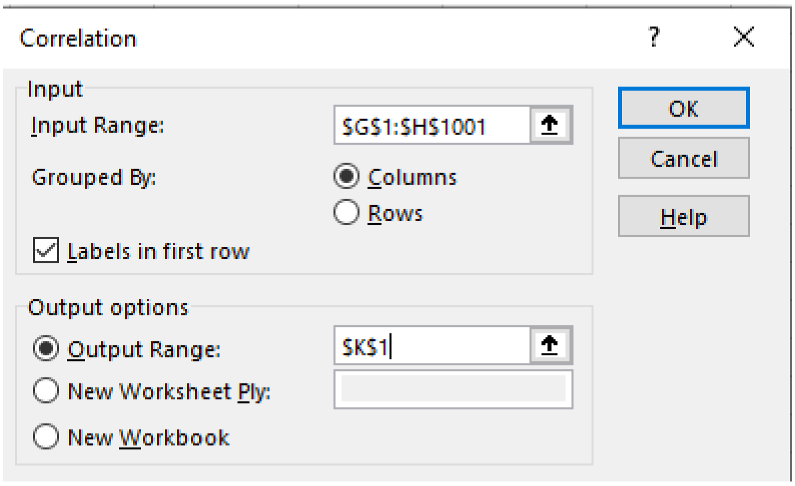
Weak positive linear correlation (r is positive but close to zero)

Strong negative linear correlation (r is close to –1)

Weak negative linear correlation (r is negative and close to zero)

Example 6.1: (open "education file")

- Measure the strength of the relationship (association) between the three scores (math score, reading score, and writing score).
- Scoter plot between reading score (x), and writing score (y).



Correlation

Input

Input Range: $G$1:$H$1001

Grouped By: ● Columns ○ Rows

☑ Labels in first row

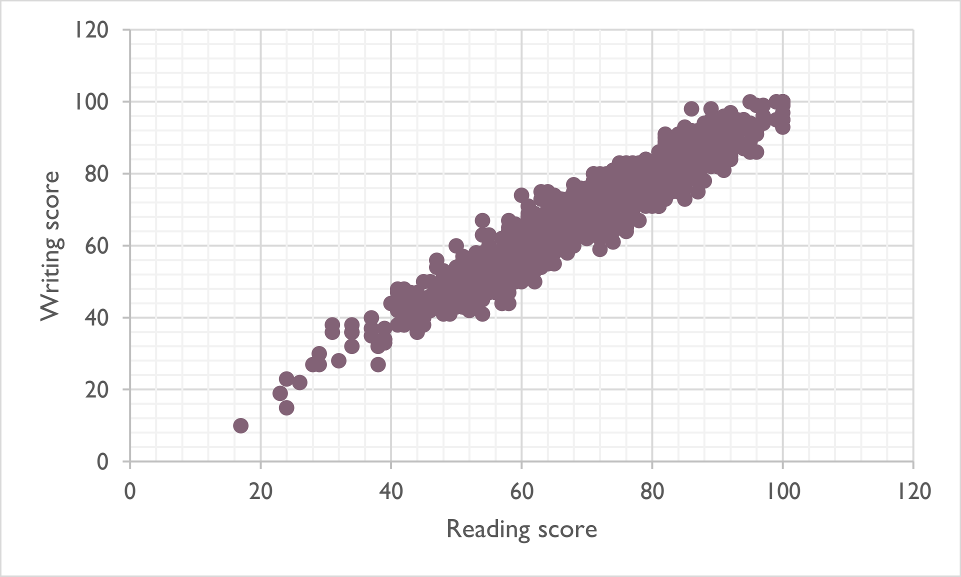
Output options

● Output Range: $K$1

○ New Worksheet Ply:

○ New Workbook

	math score	*reading score*	*writing score*
math score	1		
reading score	0.817579664	1	
writing score	0.802642046	0.954598077	1



Simple linear regression:

Simple linear regression model: $\mathbf{y = \beta_0 + \beta_1 x + \epsilon}$

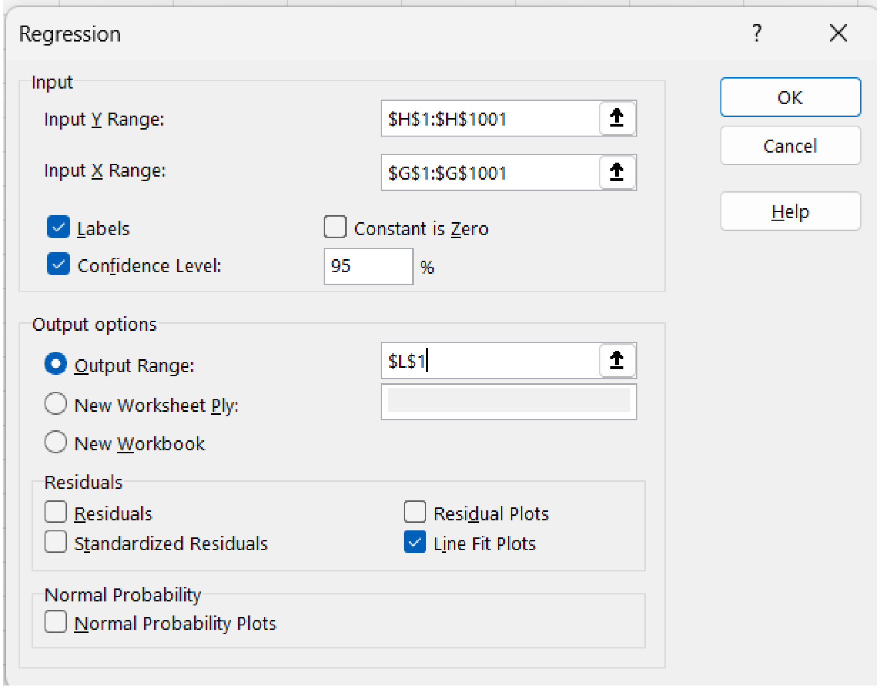
y : Dependent variable which we want to estimate.
β_0: y-intercept.
β_1: Slope.
x: Independent variable.
ϵ: Error term.

Linear equation for the least-squares line

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$ Where $\widehat{\beta_1} = \frac{\sum xy - n(\bar{x})(\bar{y})}{\sum x^2 - n(\bar{x})^2}$ & $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

Example 6.2: (open "education file")

- Find the linear regression equation to assess the relation between reading score (x) and writing score (y).
- Predict the mean score for writing score if the reading score = 30



Regression

Input

Input Y Range: $H$1:$H$1001

Input X Range: $G$1:$G$1001

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

● Output Range: $L$1

○ New Worksheet Ply:

○ New Workbook

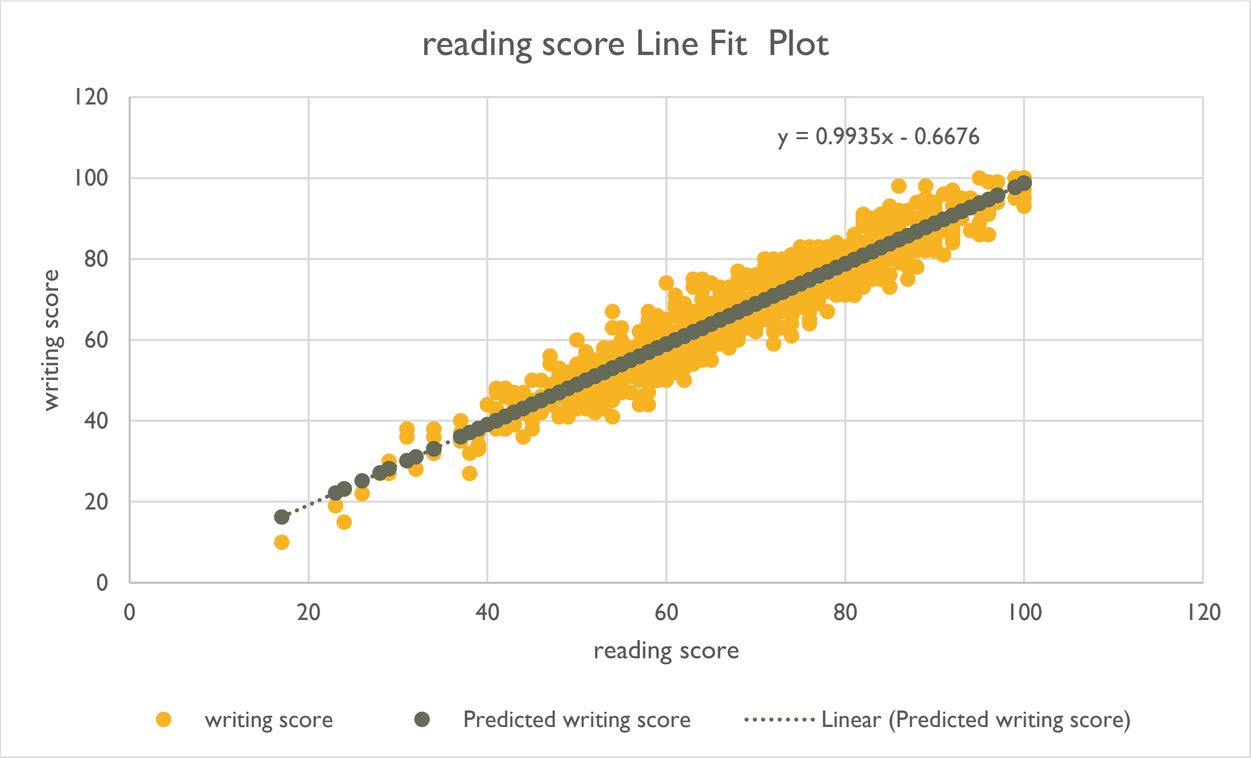
Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☑ Line Fit Plots

Normal Probability

☐ Normal Probability Plots



$$y = 0.9953(x) - 0.6676$$

$$y = 0.9953(30) - 0.6676 = 29.14$$

Problem 6.1

Find the regression line for the data on incomes and food expenditures on the seven households given in the following table. Use income as an independent variable and food expenditure as a dependent variable.

Income	55	83	38	61	33	49	67
Food expenditure	14	24	13	16	9	15	17

- Construct a scatter plot.
- What is the correlation coefficient between incomes and food expenditures?
- What is the predicted value of food expenditure for a household with $7,000 income?

Chapter 7
Time Series and forecasting

A time series is a sequence of observations of any random phenomenon measured a different point of time.

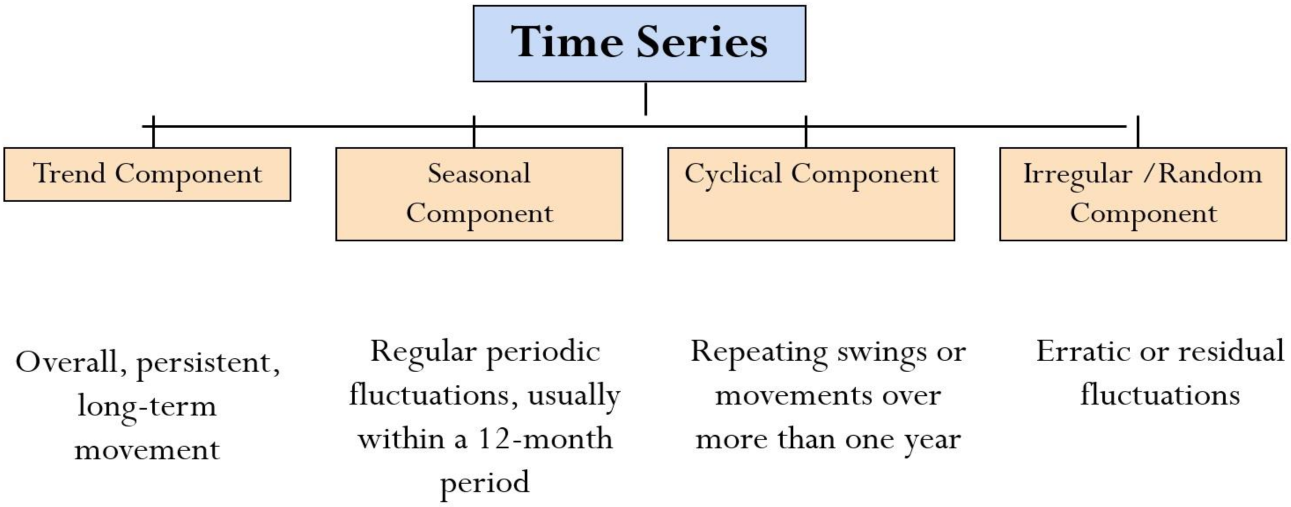
Data that are obtained from observations of a phenomenon over time are extremely common. In business and economics, we observe weekly interests' rates, daily closing stocks prices, monthly price indices, yearly sales figures.

The purposes of time series analysis are generally twofold: to understand or model the random mechanism that gives rise to an observed series and to predict or forecast future values of a series on the knowledge of the past.

Example:

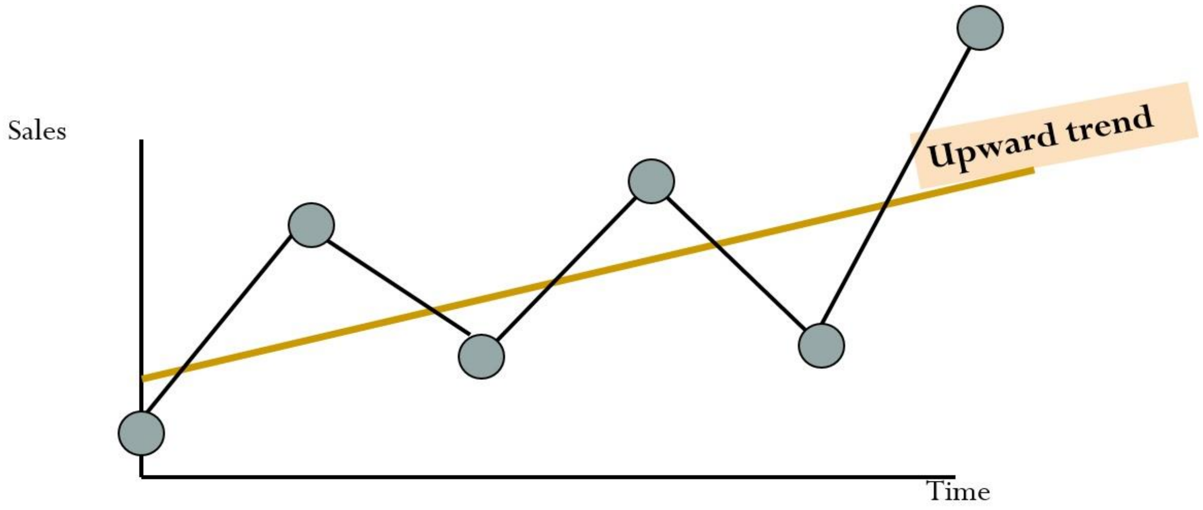
Year:	2005	2006	2007	2008	2009
Sales	75.3	74.2	78.5	79.7	80.2

Time-Series Components

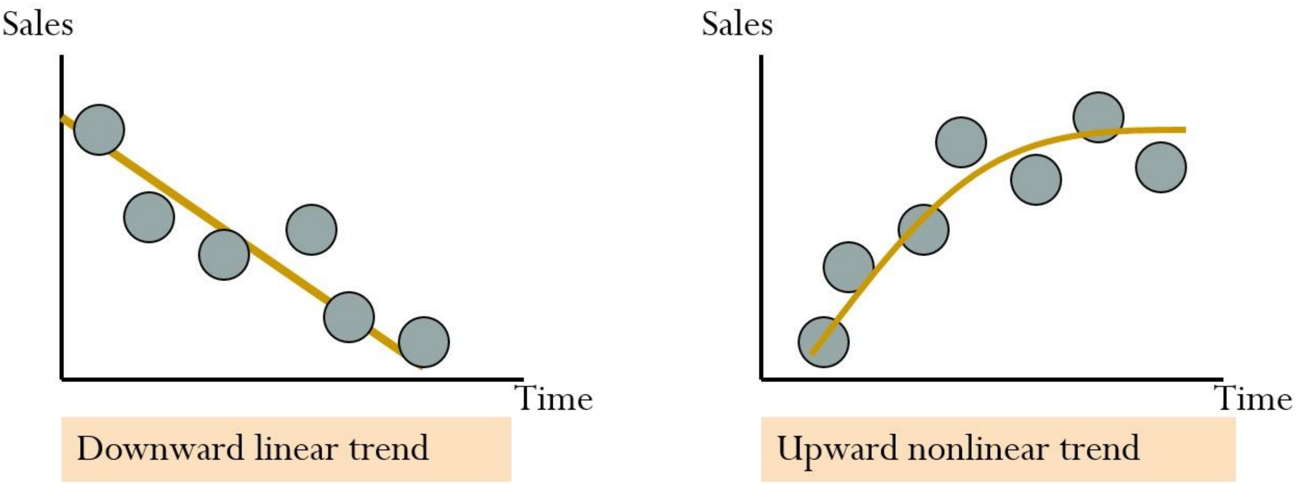

Time Series
Trend Component
Seasonal Component
Cyclical Component
Irregular /Random Component
Overall, persistent, long-term movement
Regular periodic fluctuations, usually within a 12-month period
Repeating swings or movements over more than one year
Erratic or residual fluctuations

7.1 Trend Component

- Long-run increase or decrease over time (overall upward or downward movement)
- Data taken over a long period of time.

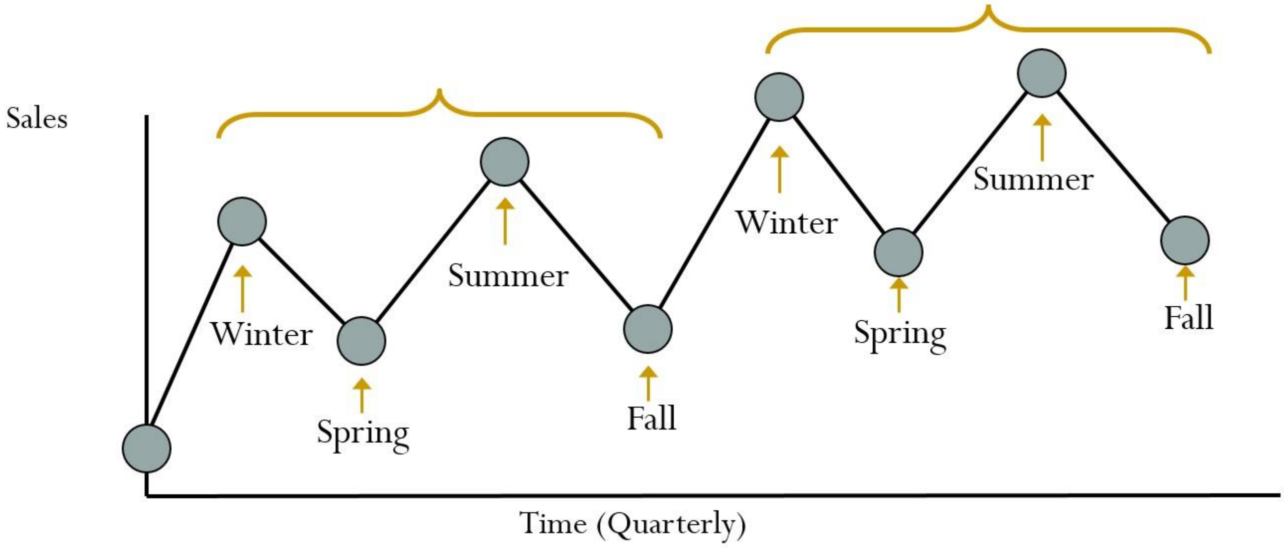


- Trend can be upward or downward.
- Trend can be linear or non-linear.



7.2 Seasonal Component

- Short-term regular wave-like patterns.
- Observed within 1 year.
- Often monthly or quarterly.



“ To Construct the Trend line”

Insert → Line
Layout → Trend line

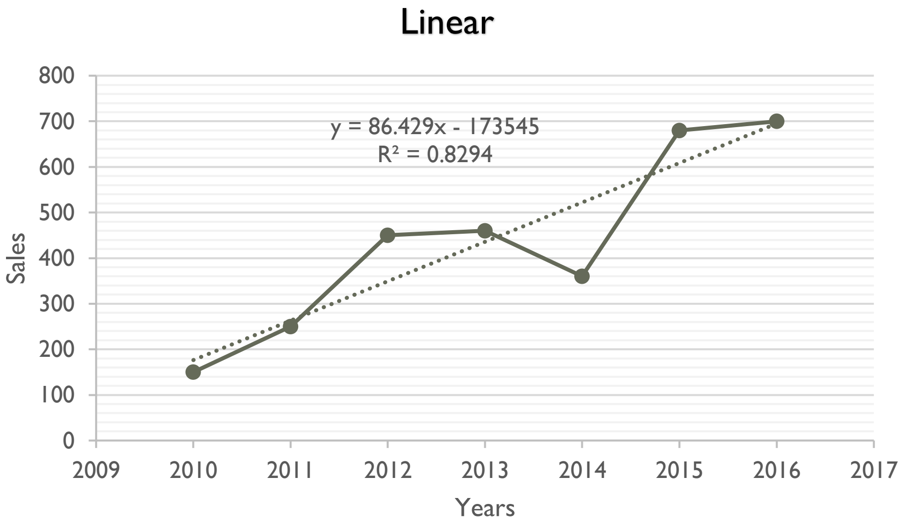
“Some forecast formulas in Excel”

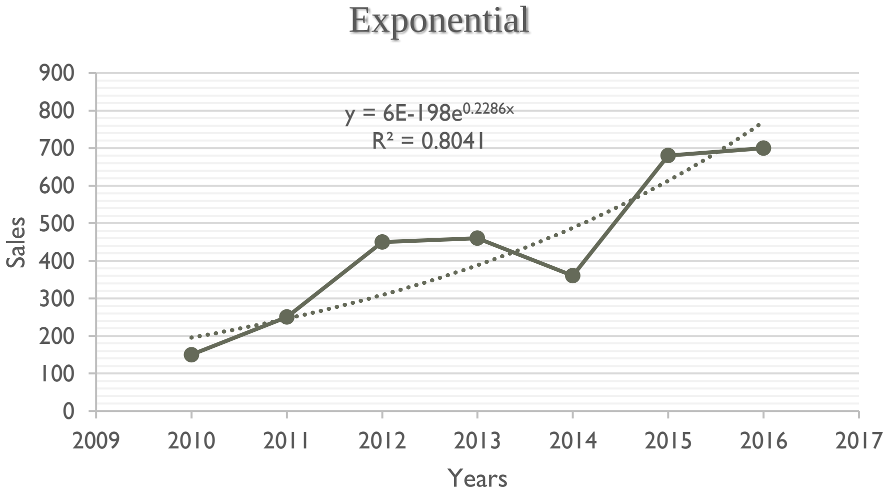
=FORECAST “Linear trend”

=GROWTH “Exponential trend”

Example 7.1: give a forecast for the tickets sales until 2026. (try different trend types and give your recommended model that fits your data)

Years	2018	2019	2020	2021	2022	2023	2024
Sales	150	250	450	460	360	680	700

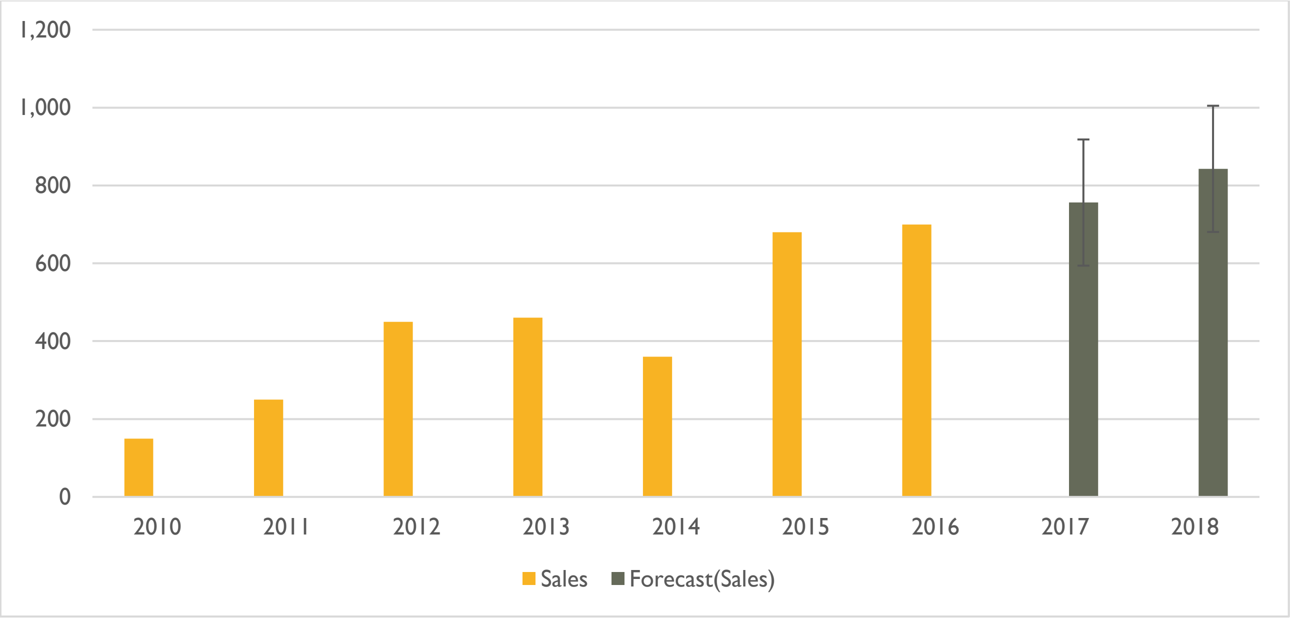




We choose the linear since it has the highest R^2

Years	2018	2019	2020	2021	2022	2023	2024	2025	2026
Sales	150	250	450	460	360	680	700	781	868

Also, by using the (forecast sheet) by going to Data → forecast sheet:



Problem 7.1**:** give a forecast for the Total U.S. food imports until 2025. (file Total U.S. food imports)

Problem 2.1:

$$No.of\ bars = \frac{\log(25)}{\log(2)} = (app.5)$$

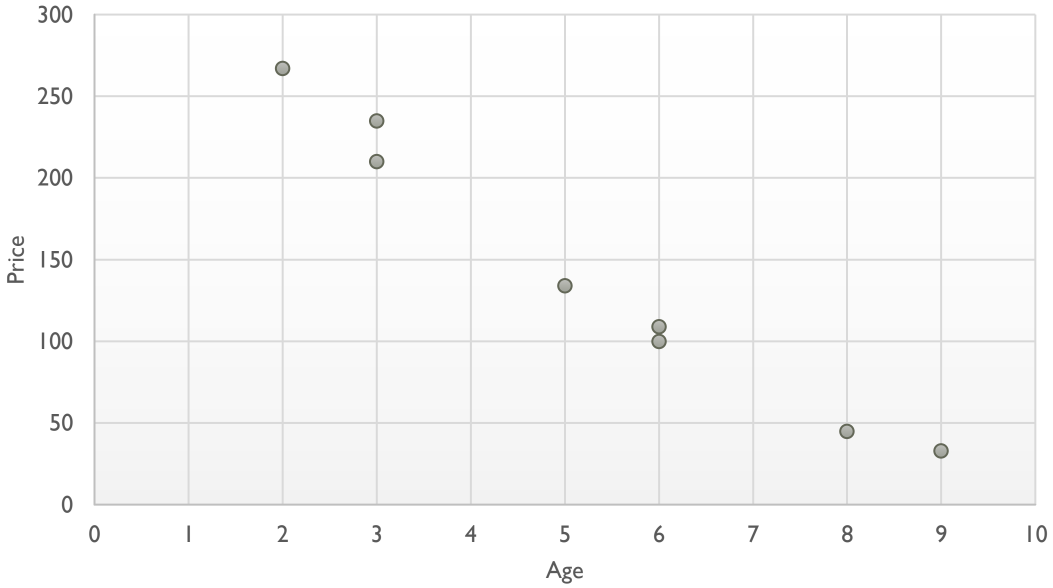
Interval length = $\frac{(40-15\,)}{5} = 5$

Waiting Time (min)	Frequency	Percentage%
15-19	8	32%
20-24	7	28%
25-29	5	20%
30-34	3	12%
35-40	2	8%
Total	**25**	**100%**

Data ⟶ Data analysis ⟶ Histogram

Input Range: (select data), Bin Range: (select 19,24,29,34,40)

Problem 2.2



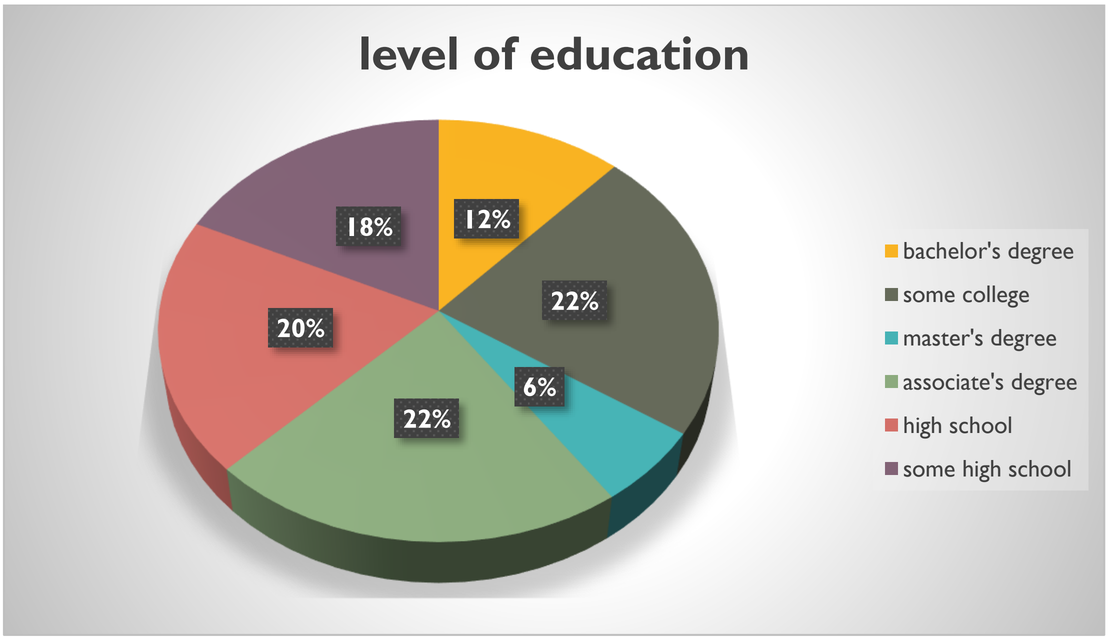
Problem 2.3

Q.1

gender	# of students
male	482
female	518
Total	**1000**

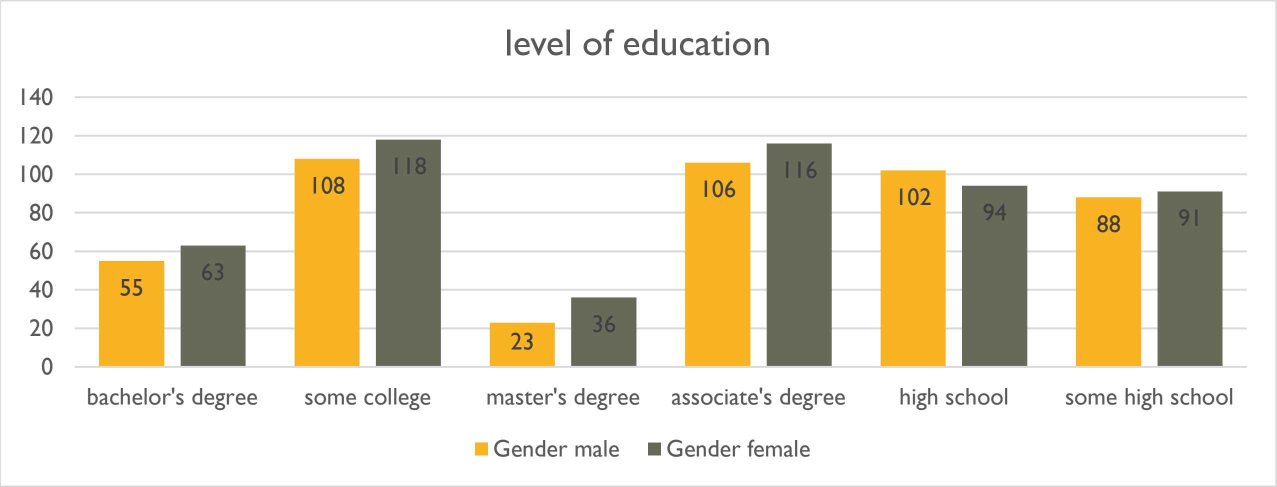
Q.2

level of education	Frq
bachelor's degree	118
some college	226
master's degree	59
associate's degree	222
high school	196
some high school	179
Total	**1000**



Q.3

level of education	Gender		Total
	male	female	
bachelor's degree	55	63	118
some college	108	118	226
master's degree	23	36	59
associate's degree	106	116	222
high school	102	94	196
some high school	88	91	179
Total	**482**	**518**	**1000**



Problem 3.1:

Peak hours	
Mean	1670.05
Standard Error	40.26253
Median	1701
Mode	#N/A
Standard Deviation	180.0595
Sample Variance	32421.42
Kurtosis	2.922258
Skewness	-1.1914
Range	794
Minimum	1135
Maximum	1929
Sum	33401
Count	20

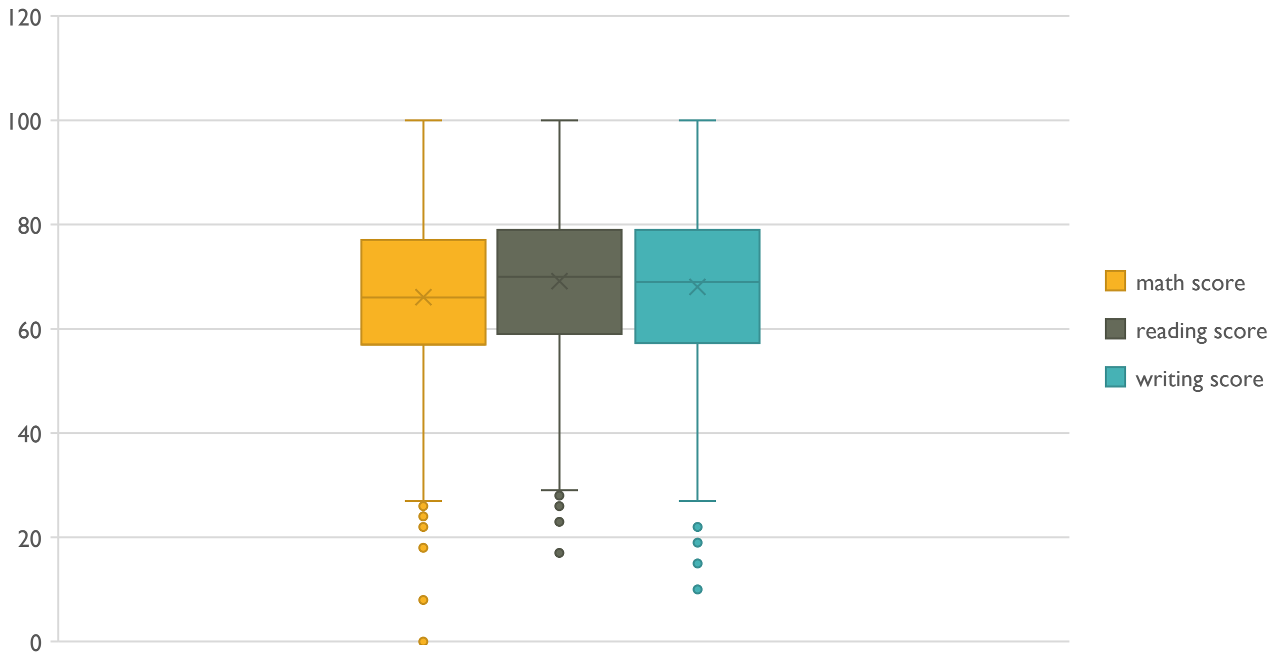
$Q_1 = 1560.25$, $Q_3 = 1785.25$, IQR $= 225$

The lower limit: $Q_1 - 1.5(IQR) = 1222.75$

The upper limit: $Q_3 + 1.5(\text{IQR}) = 2122.75$

Then we conclude that from the given data, the value 1135 is considered an outlier.

Problem 3.2



	math score	reading score	writing score
Average	66.09	69.17	68.05
Median	66.00	70.00	69.00
Standard Deviation	15.16	14.60	15.20
IQR	20.00	20.00	21.75

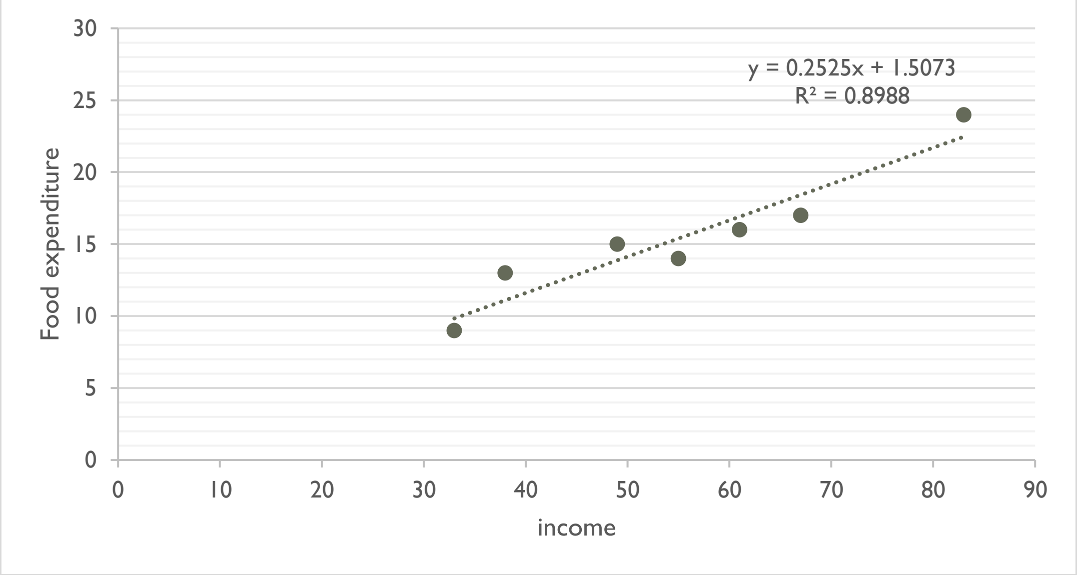
Since there are outlies, the median and the IQR are the appropriate measures of location and dispersion for Math score, reading score and writing.

Problem 5.1

t-Test: Two-Sample Assuming Unequal Variances

	Age-Male	*Age-Female*
Mean	31.11764706	32.02777778
Variance	19.11029412	25.57063492
Observations	17	36
Hypothesized Mean Difference	0	
df	36	
t Stat	-0.671975065	
P(T<=t) one-tail	0.252944753	
t Critical one-tail	1.688297714	
P(T<=t) two-tail	0.505889506	
t Critical two-tail	2.028094001	

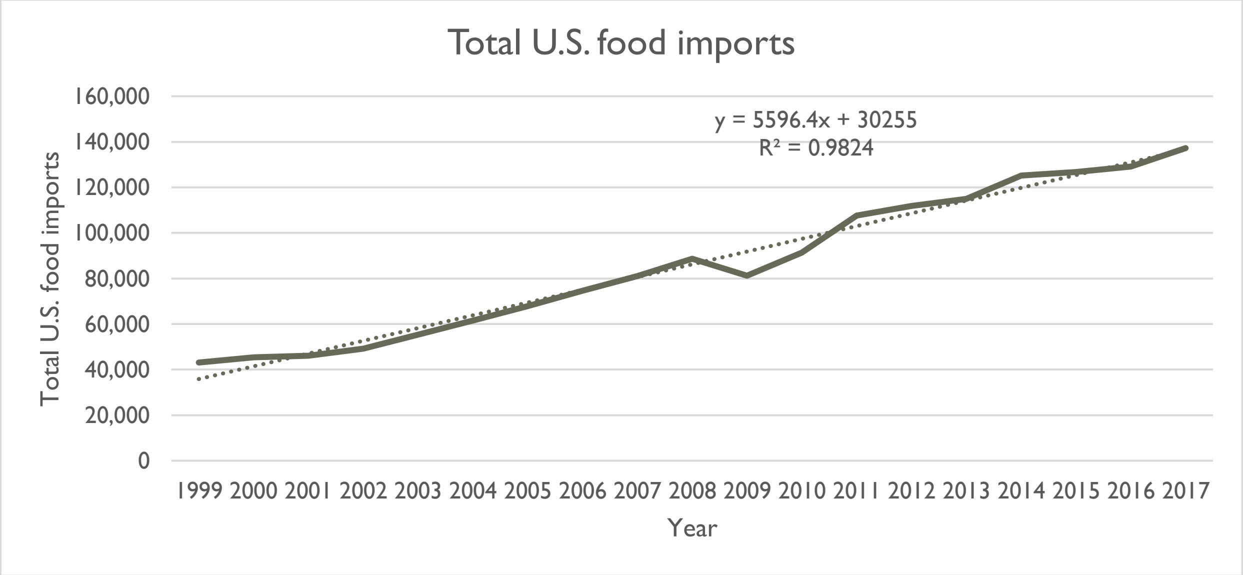
Problem 6.1

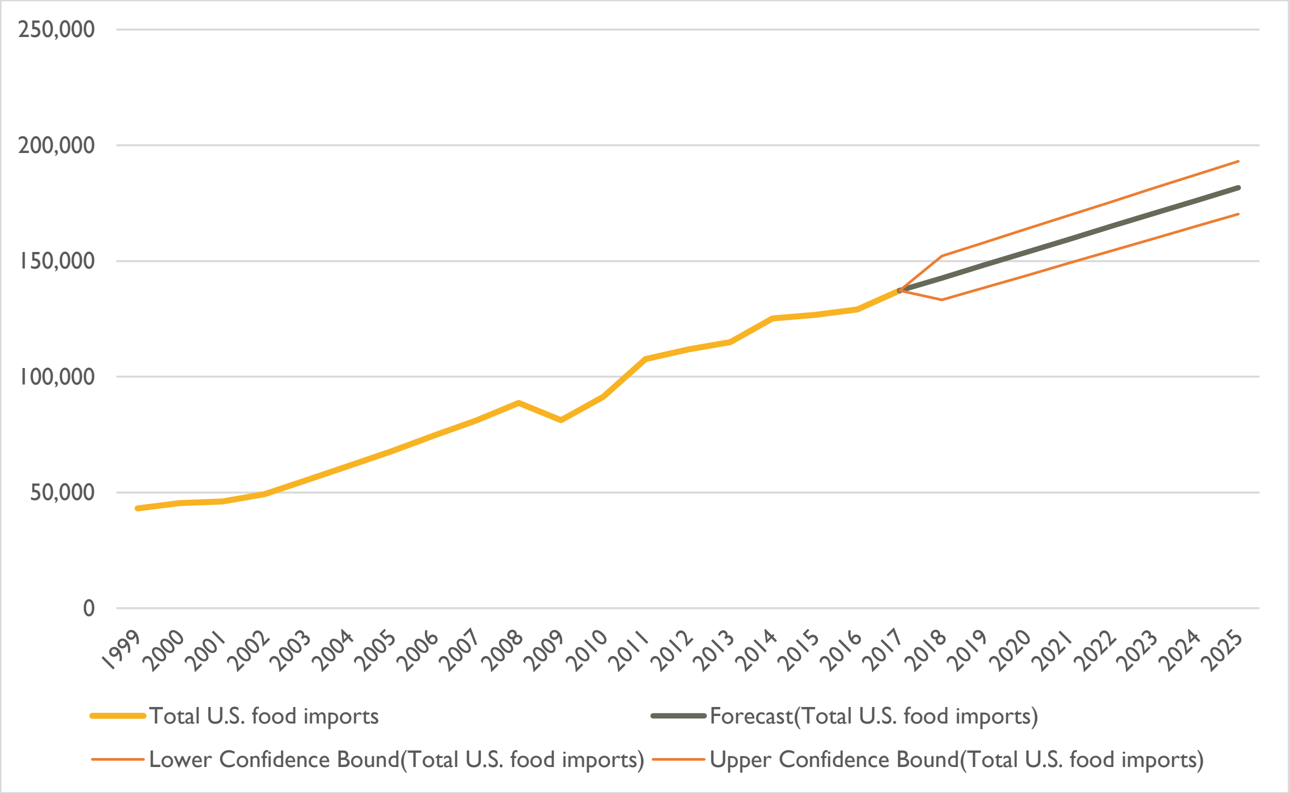


R= 0.95 (Strong and positive relation).

$$y = 0.2525(70) + 1.5073 = 19.18$$

Problem 7.1:




250,000
200,000
150,000
100,000
50,000
0
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Total U.S. food imports
Forecast(Total U.S. food imports)
Lower Confidence Bound(Total U.S. food imports)
Upper Confidence Bound(Total U.S. food imports)