[Academic Script]

[Use of Spreadsheets for Data Analysis]

Subject:	Business Economics
Course:	B. A. (Hons.), 5th Semester, Undergraduate
Paper No. & Title:	Paper – 502 Computational Techniques Business Economics
Unit No. & Title:	Unit – 2 Use of Spreadsheets for Data Analysis
Lecture No. & Title:	1 (One) Use of Spreadsheets for Data Analysis

USE OF SPREADSHEETS FOR DATA ANALYSIS

0. Objective

To understand various Excel's statistical tools and its implementations and provides illustrative experience in the use of Excel for data summary, presentation, and for other basic statistical analysis.

1. Introduction

Excel is the widely used statistical package, which serves as a tool to understand statistical concepts and computation to check hand-worked calculation in solving homework problems. Most of Excels statistical procedures are part of the Data Analysis tool pack, which is in the Tools menu. It includes a variety of choices including simple descriptive statistics, t-tests, correlations, 1 or 2-way analysis of variance, regression, etc. If excel spreadsheet does not have a Data Analysis item on the Tools menu, we need to install the **Data Analysis ToolPak add-in**. Pivot Table in the Data menu can be used to generate summary tables of means, standard deviations, counts, etc. We can also use functions to generate some statistical measures such as average, standard deviation and correlation coefficient.

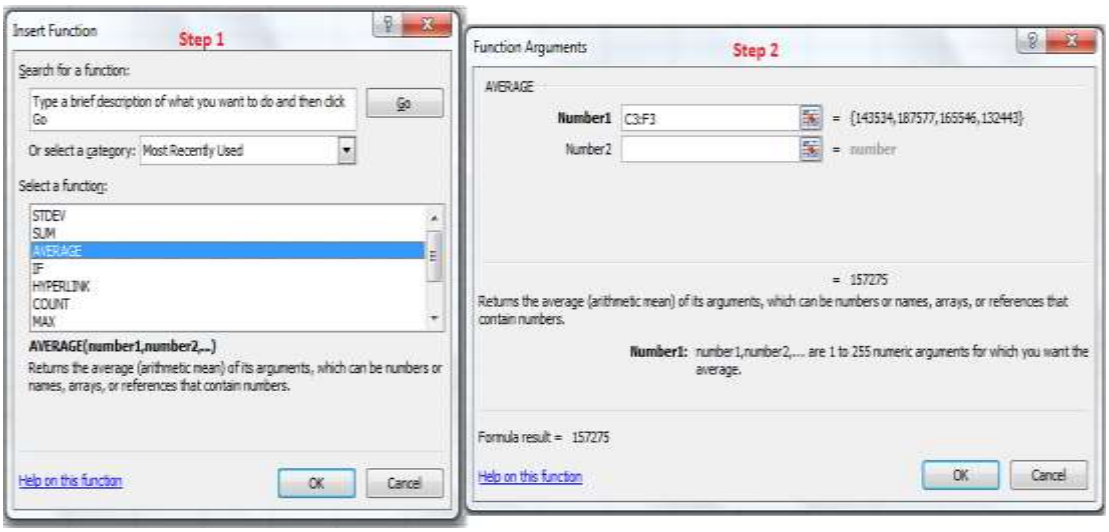
2. Basic quantitative data analysis tools in Excel

There are four sets of tools particularly used in Excel for data analysis, which is discussed below:

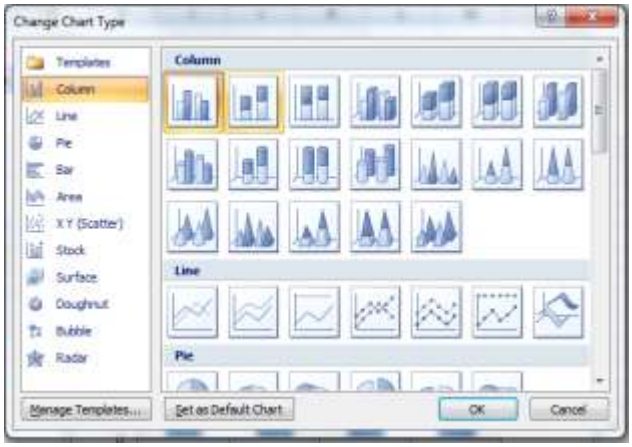
2.1 Statistical functions: Excel offers a broad range of built-in statistical functions. These are used to carry out specific data manipulation tasks. For example:

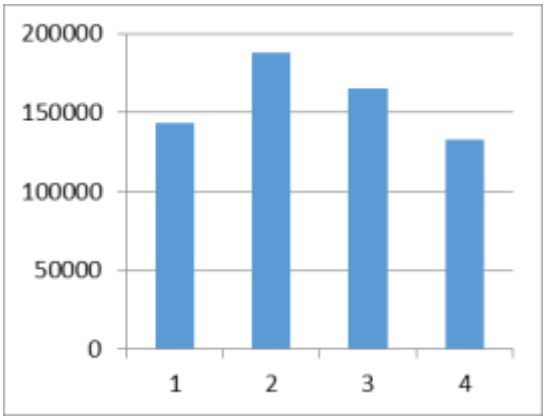
Average Function: It calculates the arithmetic mean of the cells in a specified range.

- Select the cell in which the calculation to be placed. Go to Formulas → More Functions → Statistical → AVERAGE.
- Select the range the cells to which the function should be applied.
- Press OK key, the formula will be calculated as "=AVERAGE (C3:F3)"



2.2 Charts: Excel's in-built charts (graphs) cover most of the chart types and are invaluable in data exploration and presentation.

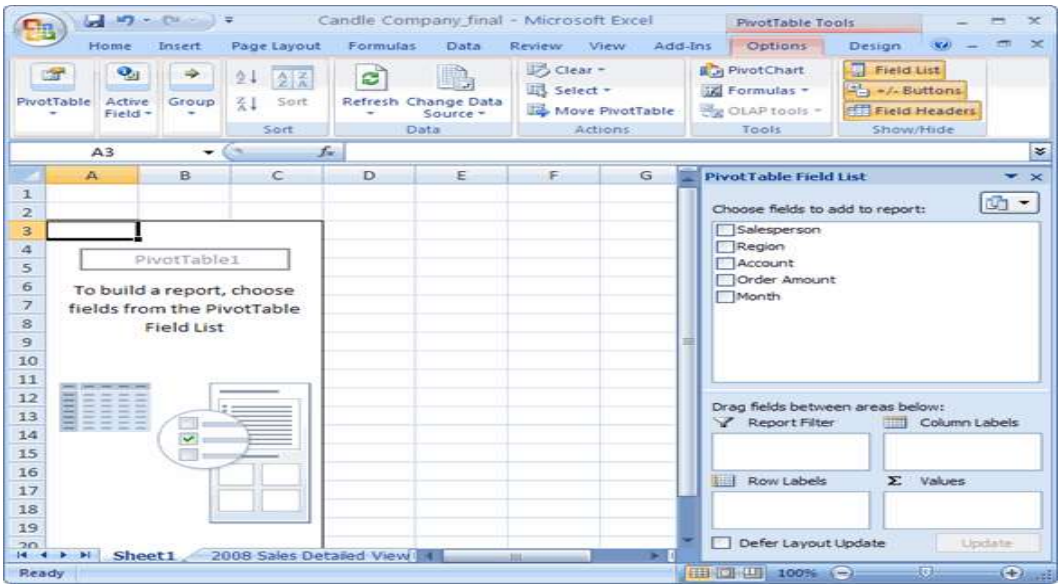


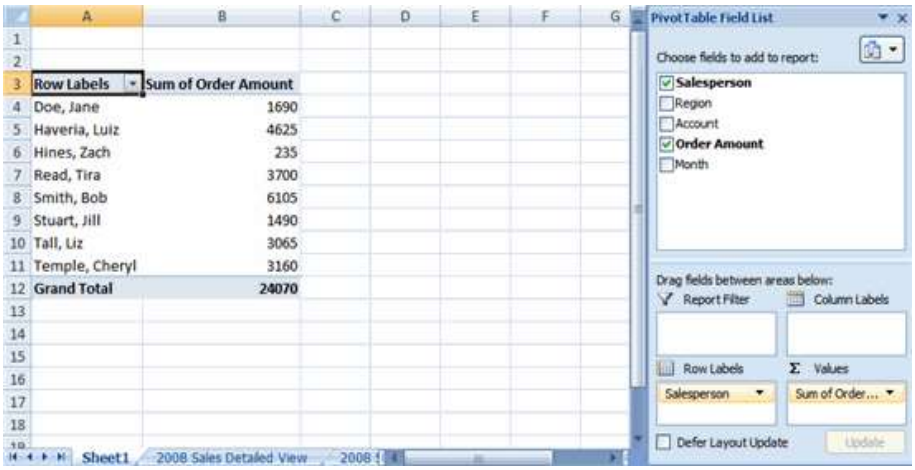


2.3 Pivot tables: Pivot tables provide a way of generating summaries and organising data in ways that are more useful for particular tasks. These are extremely useful for creating contingency tables, cross-tabulations and tables of means and other summary statistics. Presenting concise, attractive, and annotated online or printed reports, they are also called as univariate analysis.

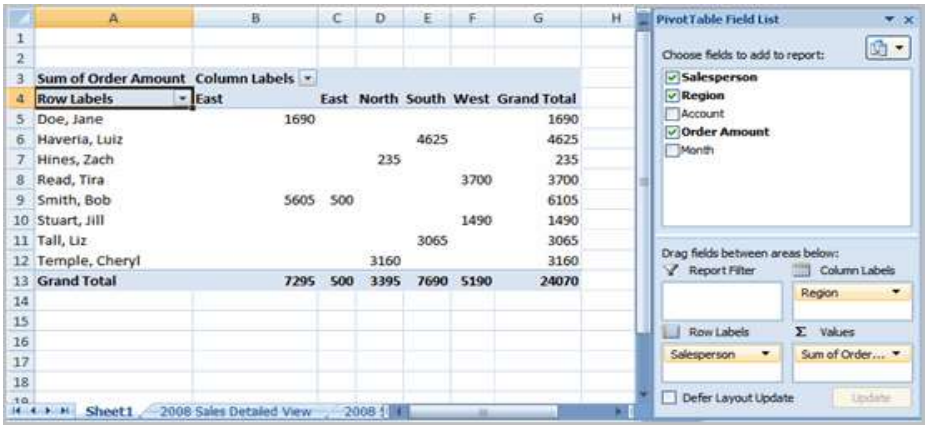
While creating a PivotTable, Excel adds a blank grid for the new pivot table and displays a PivotTable Field List task pane on the right side of the worksheet area while the layout area appears on the left. Each column label in our data becomes a field that can be used in the report.

For example, a sample spreadsheet prepared with the column headings: salesperson, region, account, order amount, and month. If we want to know which salesperson sold the highest amount, pivot table can easily create this report. Drag-and-drop the fields in the right side layout area of the worksheet. The order amount data appears on the right. All of the salesperson data appears on the left side as rows. This is a default setting in Excel-data with numbers will always appear on the right. Now we can see **report** on the left of the work sheet.

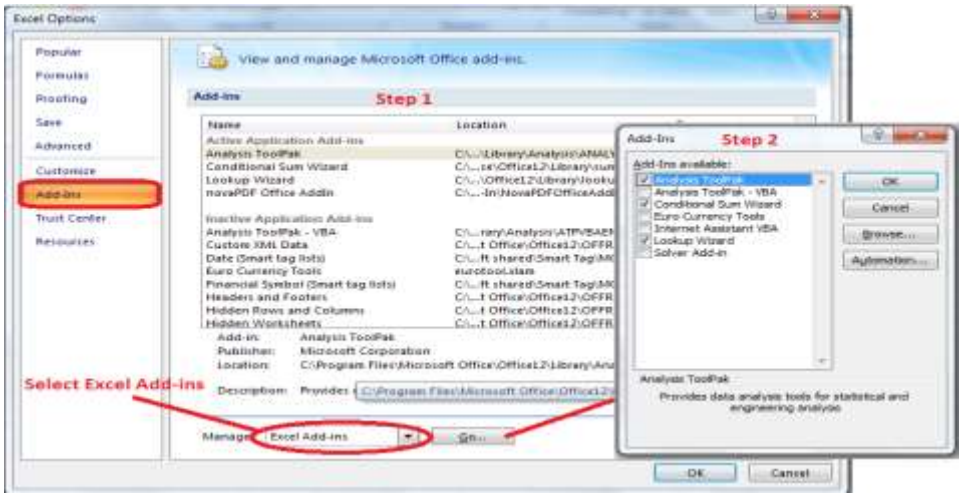




We can also change the report by dragging the other fields into the **Report Filter section** and **Column Level.** For showing the data for a specific region, just drag **region** filed into the **Column Level**, and then click OK. The PivotTable report will be changed.



2.4 Data Analysis ToolPak: The Data Analysis ToolPak is an Excel add-in, gives access to a number of helpful tools for running statistical analysis in our workbooks, more specifically, undertaking a variation analysis. It is installed when we install Excel but needs to be loaded prior to use. To do so: Select File → Excel Options → Add-Ins → Analysis ToolPak



Once the Data Analysis ToolPak is loaded (in Excel 2007), it will be available via the Data > Data Analysis,

Statistical Functions available in the Data Analysis ToolPak are:

Function name	Description
ANOVA: Single Factor	Performs a ANOVA to compare the equality of three or more means.
Correlation	Creates a correlation matrix showing the Pearson correlation Coefficient (r) for each pair of variables of N cases selected.
Descriptive statistics	Calculates a range of univariate descriptive statistics, including measures of central tendency, dispersion, skewness and kurtosis for a variable.
Histogram	Generates a histogram for a range of data (this function also generates a table of the data on which the histogram is based and can be used to generate data for the number of occurrences of a value in a data set).
Regression	Performs linear regression analysis by using the "least squares" method to fit a line through a set of observations and analyze how a single dependent variable is affected by the values of one or more independent variables.
T-Test: Paired Two-Sample for Means	Performs a t-test to compare the means of a paired sample.
T-Test: Two-Sample Assuming Equal Variances	Performs a t-test to compare the means of two independent samples, assuming equal variances.
T-Test: Paired Two-Sample Assuming Unequal Variances	Performs a t-test to compare the means of two independent samples, assuming unequal variances.
F-Test: Two Sample for Variances	Performs a f-test to compare two population variances.
Z-Test: Two Sample for Means	Performs a z-test to Compare of the means of two independent groups of samples, taken from two populations with known variance.

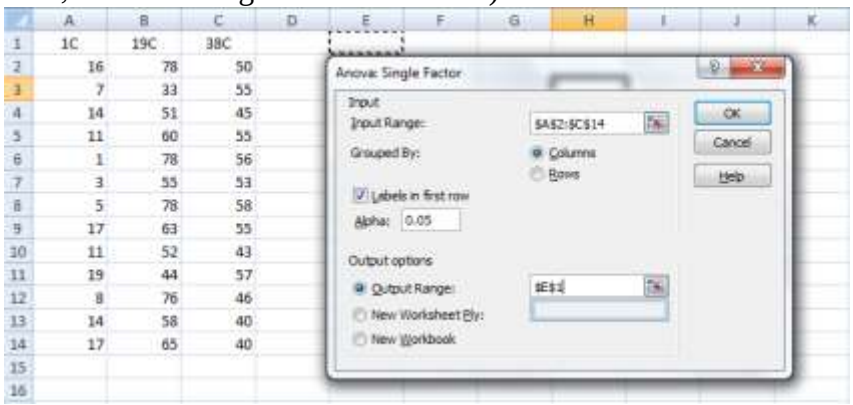
3. Data Analysis using Excel Statistical Functions

3.1 ANOVA

The ANOVA function in Excel is the analytical tool used for variance analysis. A form of hypothesis testing, it will determine whether two or more factors have the same mean. Currently, it has three different variations depending on the test that can be performed: Single factor, two-factor with replication and two factors without replication. In this chapter, we will look at single factor ANOVA where we want to compare the results for different levels (treatments) of the factor.

- **ANOVA: Single-factor –** This tool performs a simple analysis of variance on data for two or more samples. If there are only two samples, we can use the function t-Test. If you are comparing the means of three or more groups then an *ANOVA test* is used. A single factor or one-way ANOVA is used to test the null hypothesis that the means of several populations are all equal.

For example, take the measurements of water temperatures in three columns (one for each water temperature). With ANOVA (Analysis of Variance), we are testing different groups to see if there's a significant difference between them. (if $p < 0.05$ there is a significant difference. If $p > 0.05$, there is no significant difference)



- Select ANOVA: Single Factor and click OK.
- Input range by selecting data in all three columns recorded for all temperatures in the example.
- Define the output range so that the output doesn't overwrite any of the data in the worksheet.
- Clicking on OK, the output of ANOVA test will look like this:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	13	143	11	33.66667
Column 2	13	791	60.84615	200.9744
Column 3	13	653	50.23077	43.52564

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	17924.31	2	8962.154	96.65594	3.36E-15	3.259446
Within Groups	3338	36	92.72222			
Total	21262.31	38				

From this output, it is obvious that $p > 0.05$, it means there is not a significant difference between the three groups, hence are not all equal ($F > F$ crit, $p = 3 * 10^{-15}$).

- **ANOVA: Two-Factor with Replication -** A two way ANOVA with replication is performed when we have two groups and individuals within that group are doing more than one thing (i.e. taking two tests). For example, there are 12 individuals in a group having scores in two tests. Here, we want to know, is there any significant difference between results and scores in Group A and Group B. We will figure out if we are going to reject the null hypothesis or not, we'll basically be looking at two factors:
  - If the F-value (f)is larger than the f critical value (f crit)
  - If the p-value is smaller than the chosen alpha level (0.05).

Steps: Select "ANOVA two factor with replication" from "Data Analysis" Dialog Box then click "OK." The two way ANOVA window will open. Type an Input Range into the Input Range box, and type a number in the "Rows per sample" box, in the example it is 12. Then choose the output range and click "OK". We will get following result.

	A	B	C	D	E	F	G	H	I	J
1	Groups	Math	English	Anova: Two-Factor With Replication						
2	Group A	90	67							
3		87	89	SUMMARY	Math	English	Total			
4		78	84	*Group A*						
5		77	86	Count	12	12	24			
6		89	98	Sum	1013	1009	2022			
7		98	91	Average	84.417	84.083	84.25			
8		88	92	Variance	118.45	118.08	113.15			
9		81	99							
10		84	86	*Group B*						
11		92	77	Count	12	12	24			
12		56	71	Sum	740	978	1718			
13		93	69	Average	61.667	81.5	71.583			
14	Group B	81	99	Variance	411.52	173.55	382.43			
15		75	77							
16		54	71	*Total*						
17		55	61	Count	24	24				
18		45	98	Sum	1753	1987				
19		56	68	Average	73.042	82.792				
20		79	98	Variance	388.48	141.22				
21		66	77							
22		67	80							
23		89	89	ANOVA						
24		12	91	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
25		61	69	Sample	1925.3	1	1925.3	9.3737	0.0037	4.0617
26				Columns	1140.8	1	1140.8	5.5539	0.023	4.0617
27				Interaction	1220.1	1	1220.1	5.9401	0.0189	4.0617
28				Within	9037.5	44	205.4			
29										
30				Total	13324	47				

From the ANOVA test, we can see that F > F crit. P-value, P-value greater than the exact alpha value 0.05. That means, there is significant difference between the scores in two groups. If we compare the Average and Variances in two groups, there is definitely significant difference between two groups.

- **ANOVA: Two-Factor without Replication -** A Two way ANOVA in Excel without replication compares a group of individuals performing more than one task. For example, there is group or 6 students' scores across a series of tests. We want to know, what is the difference between each groups, how do the student compare each other and how do the exams compare each other.

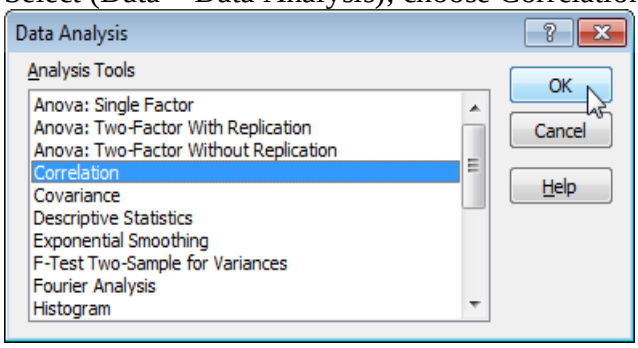
	A	B	C	D	E	F	G	H	I	J	K
4	Student	Math	English	Science	Anova: Two-Factor Without Replication						
5	1	89	68	89							
6	2	87	74	90	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*		
7	3	99	89	99	1	3	246	82	147		
8	4	100	90	85	2	3	251	83.66667	72.33333		
9	5	96	84	96	3	3	287	95.66667	33.33333		
10	6	100	82	100	4	3	275	91.66667	58.33333		
11					5	3	276	92	48		
12					6	3	282	94	108		
13											
14					Math	6	571	95.16667	33.36667		
15					English	6	487	81.16667	74.56667		
16					Science	6	559	93.16667	36.56667		
17											
18											
19					ANOVA						
20					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
21					Rows	476.5	5	95.3	3.873984	0.032594	3.325835
22					Columns	688	2	344	13.98374	0.001267	4.102821
23					Error	246	10	24.6			
24											
25					Total	1410.5	17				

Looking primarily two things, we can compare the F-value to the F critical value. The first row of ANOVA is rows; it comparing each student in each different row, as F-value is larger than F critical value that means there is difference in each student's performance. The second row of ANOVA is columns; if we compare F-value to the F critical value there is huge difference in performance on three different tests that means there is difference in each student's performance. The second factor is p-value. P-values are smaller than the chosen alpha level, so it determines that the differences are statistical significant and we would likely cheeked the null hypothesis.

3.2 Correlation (r)

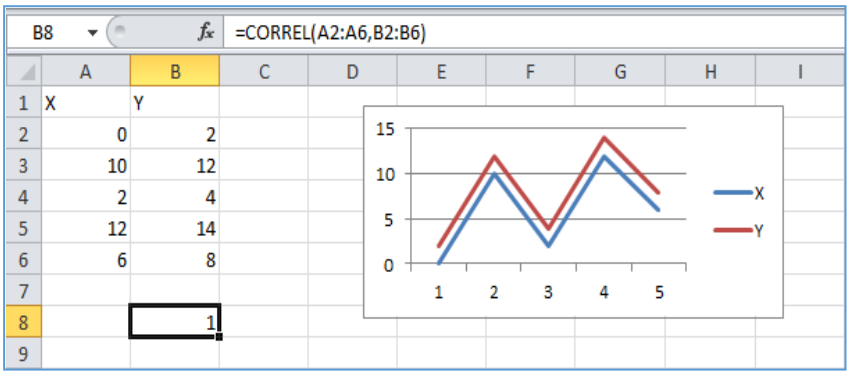
The correlation coefficient (a value between -1 and +1) tells how strongly two variables are related to each other. We can use the **CORREL** function or the **Analysis ToolPak add-in** to find the correlation coefficient between two variables.

Select (Data > Data Analysis), choose Correlation in the Data Analysis dialog box.

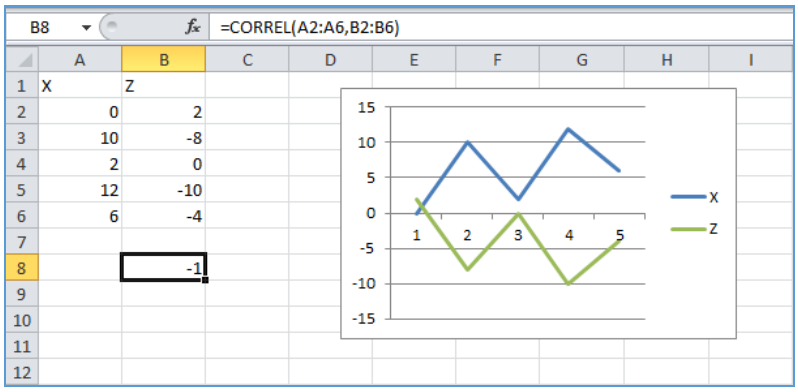


In correlation, when values of one variable increases with the increase in another variable, it is supposed to be a **positive correlation**. On the other hand, if the values of one variable decrease with the decrease in another variable, then it would be a **negative correlation**. There might be the case when there is no change in a variable with any change in another variable. In this case, it is defined as **no correlation** between the two.

- A correlation coefficient of +1 indicates a perfect positive correlation. In the example, as variable X increases, variable Y increases. As variable X decreases, variable Y decreases.

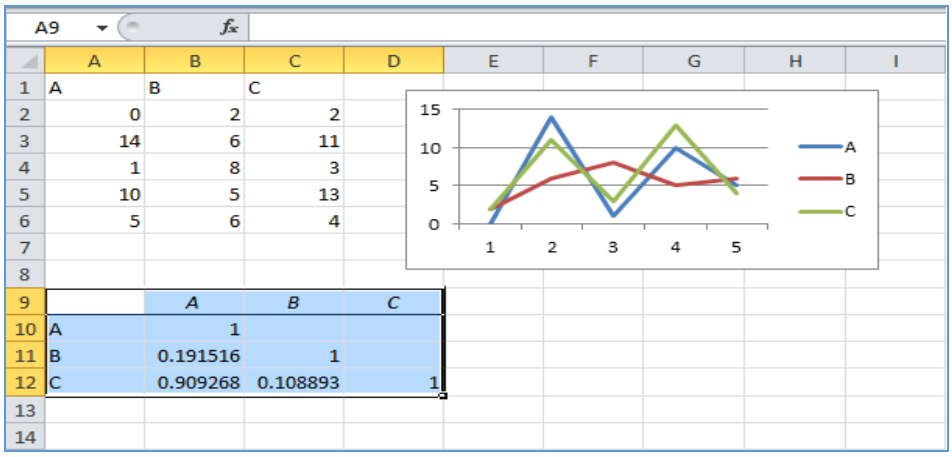


- A correlation coefficient of -1 indicates a perfect negative correlation. As variable X increases, variable Z decreases. As variable X decreases, variable Z increases.



- A correlation coefficient near 0 indicates no correlation.

We can also quickly generate correlation coefficients between multiple variables. For example, select the range A1:C6 as the Input Range and cell A9 as the Output Range. We can find that variables A and C are positively correlated (0.91). Variables A and B are not correlated (0.19). Variables B and C are also not correlated (0.11).



➢ **Linear Correlation -** When the change in one variable results in the constant change in the other variable, then we say the correlation is linear.

Example**:**

X:	10	20	30	40	50
Y:	20	40	60	80	100

Here, there is a linear relationship between the variables. There is a ratio 1:2 at all points. Also, if we plot them they will be in a straight line.

➢ **Non Linear Correlation -** When the amount of change in one variable is not in a constant ratio to the change in the other variable, we say that the correlation is non linear.

Example:

X:	10	20	30	40	50
Y:	10	30	70	90	120

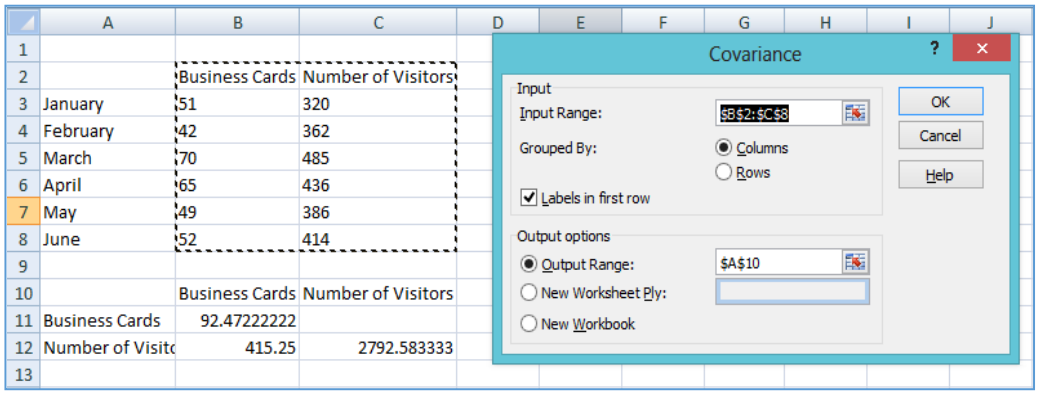
Here there is a non linear relationship between the variables. The ratio between them is not fixed for all points. Also if we plot them on the graph, the points will not be in a straight line. It will be a curve.

3.3 Covariance

Covariance is the measure of how much two sets of data vary. It determines the degree to which the two variables are related or how they vary together. The Covariance is the average of the product of deviations of data points from their respective means.
It is important to remember the difference between variance and covariance:

- Variance - This is a measure of how much a single variable changes.
- Covariance - This is a measure of how much two variables change together.

Example: We want to find the covariance for the number of business cards we give out and the number of people who visit our website.
Select (Tools > Data Analysis) and choose Covariance in the Data Analysis dialog box and press OK. Specify the input range to be B2:C8 and specify the output to be pasted in cell "A10" and press OK. The result of covariance analysis is as follows.



	A	B	C
1			
2		Business Cards	Number of Visitors
3	January	51	320
4	February	42	362
5	March	70	485
6	April	65	436
7	May	49	386
8	June	52	414
9			
10		Business Cards	Number of Visitors
11	Business Cards	92.47222222	
12	Number of Visit	415.25	2792.583333
13			

We can use the covariance tool to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

3.4 Descriptive Statistics

The Descriptive Statistics analysis tool generates a report of univariate statistics for data in the input range, providing information about the central tendency and variability of data including:

- The mean, mode, median and range.
- Variance and standard deviation.
- Skewness and Kurtosis
- Count, maximum and minimum.

Go to Data → Data Analysis → Descriptive Statistics. Select the input and output range in the respective boxes and click "OK". A summery result will be displayed.

	A	B	C	D
1	**Descriptive Data Analysis Tool**			
2				
3	Scores		*Scores*	
4	23			
5	38		Mean	30.81818182
6	45		Standard Error	4.931933388
7	21		Median	23
8	17		Mode	21
9	21		Standard Deviation	16.35737254
10	8		Sample Variance	267.5636364
11	61		Kurtosis	-0.547131093
12	21		Skewness	0.625138751
13	52		Range	53
14	32		Minimum	8
15			Maximum	61
16			Sum	339
17			Count	11
18				

- The **mean** is the sum of the scores divided by the total number of scores. The excel function is: = AVERAGE(A2:A11)
- The **Standard Error (SE)** indicates how close the sample mean is from the true sores mean. It is calculated by dividing the standard deviation of the score (or the sample) by the square root of the total number of scores. The function is = STDEV(A2:A11) / SQRT(COUNT(A2:A11)
- The **median** is another measure of central tendency. To get the median we need to order the data from lowest to highest. The median is the number in the middle. If the number of cases is odd the median is the single value, for an even number of cases the median is the average of the two numbers in the middle. The function is = MEDIAN(A2:A11)
- The **mode** refers to the most frequent, repeated or common number in the data. The excel function is = MODE(A2:A11)
- The **sample variance** (SV) measures the dispersion of the data from the mean. It is the simple mean of the squared distance from the mean. It is calculated by:

  SV = sum of (X-mean of X)^2 / Number of observation -1

  The excel formula is = VAR(A2:A11)
- The **standard deviation** (S) is the squared root of the variance. It indicates how close the data is to the mean. The standard deviation formula is similar to the variance formula. It is given by:

  $$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

  The excel formula is = STDEV(A2:A11)
- **Skewness** characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values. A normal distribution has a skew of 0. Zero indicates perfect symmetry. The equation for skewness is defined as:

  $$\text{Skewness} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \overline{X}}{s}\right)^3$$

  The excel formula is = SKEW(A2:A11)
- Kurtosis is a measure of flatness of the distribution. Heavier tailed distributions have larger kurtosis measures. The normal distribution has a kurtosis of 3. A normal distribution has a kurtosis of 0 (otherwise it will have a kurtosis of 3). Excel uses the following equation to calculate kurtosis:

  $$\text{Kurtosis} = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i - \overline{X}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

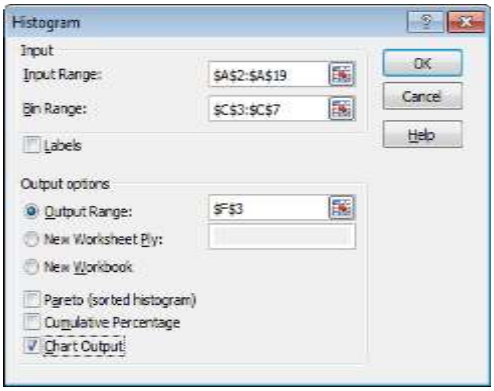
  The excel function for kurtosis is: = KURT(A2:A11)
- **Range** is a measure of dispersion. The difference between the largest and smallest value, max, min.

3.5 Histogram

Histogram is a column chart that displays frequency data. To create a histogram, we must organize the data in two columns on the worksheet. These columns must contain input data and bin numbers. Input data is the data that to be analyzed by using the Histogram tool. Bin numbers are the numbers that represent the intervals that we use for measuring the input data in the data analysis.

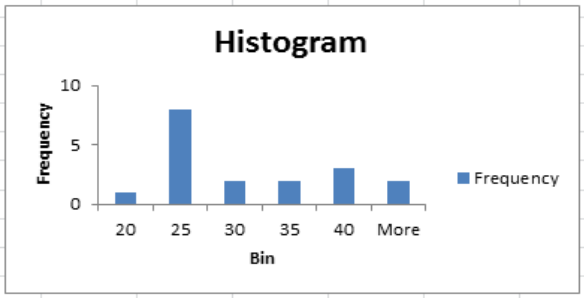
For example: the column contains the data in the range A2:A19 and bin numbers in the range C3:C7, then select **Histogram** from the **Data Analysis Tool** and click OK.

	A	B	C	D
1	Number of students			
2	22			
3	29		20	
4	40		25	
5	30		30	
6	48		35	
7	24		40	
8	21			
9	19			
10	24			
11	22			
12	25			
13	52			
14	35			
15	40			
16	31			
17	37			
18	21			
19	23			
20				
21				



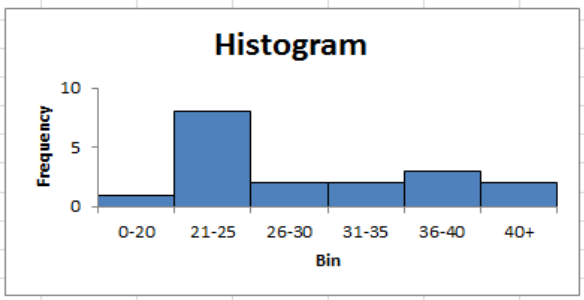
- Select the range A2:A19.
- Click in the Bin Range box and select the range C3:C7.
- Click the Output Range option button, click in the Output Range box and select cell F3.
- Check Chart Output and click OK.

Bin	*Frequency*
20	1
25	8
30	2
35	2
40	3
More	2



We can remove the space between the bars, to do this, right click a bar, select Format Data Series and change the Gap Width to 0%.

Bin	*Frequency*
0-20	1
21-25	8
26-30	2
31-35	2
36-40	3
40+	2



When we use the Histogram tool, Excel counts the number of data points in each data bin. A data point is included in a particular data bin if the number is greater than the lowest bound and equal to or less than the largest bound for the data bin. If we omit the bin range, Excel creates a set of evenly distributed bins between the minimum and maximum values of the input data.

The output of the histogram analysis is displayed on a new worksheet (or in a new workbook) and shows a histogram table and a column chart that reflects the data in the histogram table.

Summary

In the first part, we have discussed various quantitative tools used in spreadsheet for data analysis such as AVARAGE (one of the statistical functions), Charts, Pivot Tables and "Data Analysis ToolPak" functions such as ANOVA, Correlation, Covariance, Descriptive Statistics, and Histogram with examples.