ISSN:0975-9646

DATA ANALYSIS and BUSINESS MODELLING in Microsoft Excel using Analysis ToolPak

Abhishek Kanal
Department of Computer Engineering
Thadomal Shahani Engineering College

Aishwarya Raman
Department of Computer Engineering
Thadomal Shahani Engineering College

***Abstract*— Analysis toolpak is a Microsoft excel add-in that can be used for data analysis and business modeling. Analysis toolpak can be used for predicting trends, finding optimal solutions, etc. This paper demonstrates various tools and techniques provided by the toolpak such as creating histograms, analyzing descriptive statistics, analysis of variance (anova), F-test, T-test, moving averages, exponential smoothening and correlation of data sets. These features of the toolpak are explained with the help of different examples.**

***Index Terms*—Analysis toolpak, Microsoft Excel, statistical analysis, correlation, descriptive statistics, anova, f-test, moving averages, exponential smoothening, t-test.**

INTRODUCTION

Today, one of the most essential component of a business process is data. Data can be defined as a set of facts and statistics which may have been collected together for reference or analysis. Data could describe information that has been collected over the past experiences of the organization or it could be a projection looking at the past trends. Data is an important driving force in paving the way for an optimized business approach irrespective of the size of the organization. If studied appropriately, analyzing of data can prove to bring about a drastic improvement in the organizations' methodology of performing business processes. One of the most viable analysis of data can be performed by the simple software MS-EXCEL. Microsoft Excel is a spreadsheet software used to maintain chunks of data in an organized way. Apart from acting as a hoard for precious data items, excel provides for multiple tools to study and visualize data. One such tool is the Analysis Toolpak offered by excel which in itself includes a range of sub tools or techniques which permit the analysis of data items from various point of views. This paper, delves into the explanations and examples of the different techniques under the analysis toolpak.

Analysis Toolpak

Microsoft Excel provides for analyzing of data with the help of a special tool called analysis toolpak. The analysis toolpak is an excel add-in that provides for statistical, financial and engineering analysis upon the data [3]. The data and criteria is provided for each of the analysis options that we have under the analysis toolpak option. It involves the use of macro functions to perform operations and produce outputs in the form of tables or charts. Firstly, to use the analysis toolpak, it needs to be loaded into excel. In order to do so, under the file drop down, the options button is clicked.

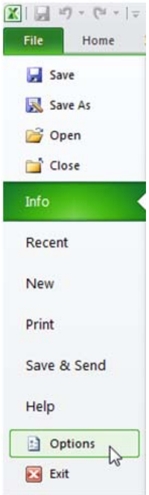
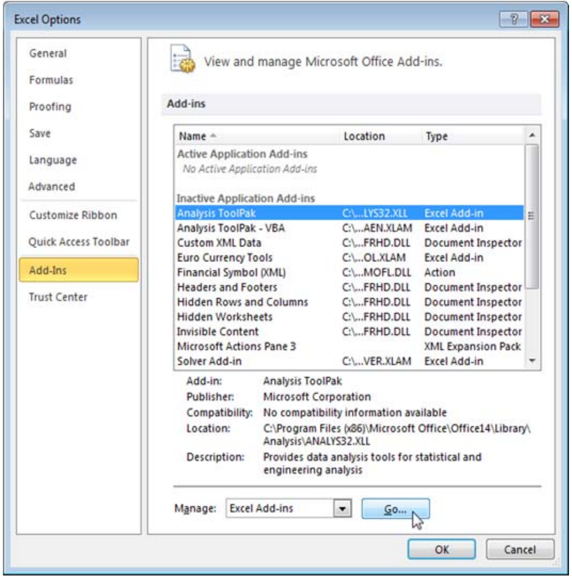


Figure 1: Options in File drop down

This opens the Excel Options window, under which the add-ins tab is clicked which shows the list of active and inactive add-ins. The Analysis ToolPak option is selected and go is pressed.



Name	Location	Type
Active Application Add-ins		
No Active Application Add-ins		
Inactive Application Add-ins		
Analysis ToolPak	C:\...LYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...AEN.XLAM	Excel Add-in
Custom XML Data	C:\...FRHD.DLL	Document Inspector
Euro Currency Tools	C:\...OL.XLAM	Excel Add-in
Financial Symbol (XML)	C:\...MOFL.DLL	Action
Headers and Footers	C:\...FRHD.DLL	Document Inspector
Hidden Rows and Columns	C:\...FRHD.DLL	Document Inspector
Hidden Worksheets	C:\...FRHD.DLL	Document Inspector
Invisible Content	C:\...FRHD.DLL	Document Inspector
Microsoft Actions Pane 3		XML Expansion Pack
Solver Add-in	C:\...VER.XLAM	Excel Add-in

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Microsoft Office\Office14\Library\Analysis\ANALYS32.XLL
Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins

Figure 2: Add-Ins tab in Excel Options

Another add-in window opens up from which Analysis ToolPak again is checked OK is hit.

Now, the Analysis ToolPak is loaded and excel is ready to perform analysis of the data. This can be cross checked with the presence of the data analysis button under the data tab of the excel window.

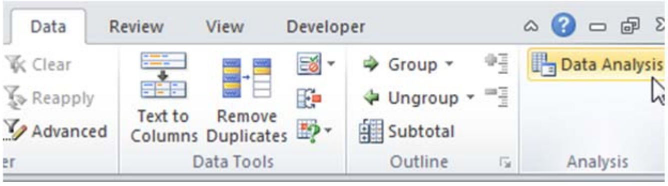


Figure 3: Data analysis in Data tab

IMPLEMENTATION

Now, upon clicking of the data analysis option under the data tab, the data analysis pop-up window appears allowing to choose from multiple techniques with different criterions to perform the different types of analysis on the data as per the need.

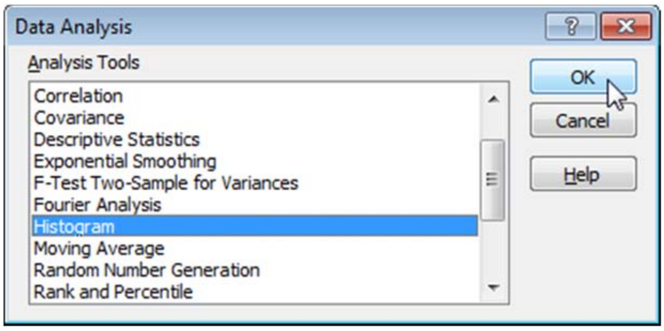
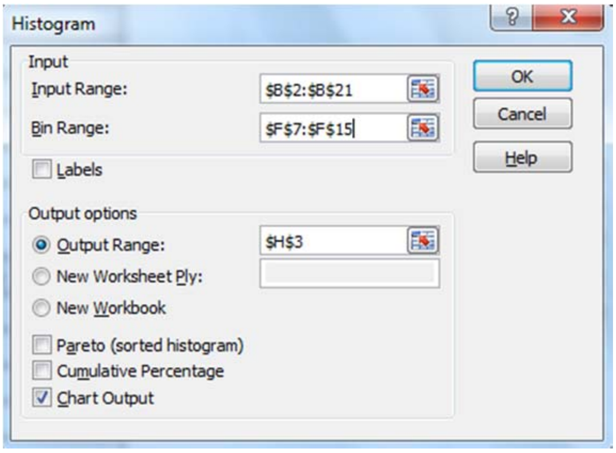


Figure 4: Data analysis dialog box

1. Histogram

Histogram is an important statistical tool which is a graphical representation of the distribution of numeric data over a bin range. From a histogram, it is possible to get an estimate of the probability distribution of a continuous variable. The histogram can depict information about a data set such as the current state of a particular system along with the scope of improvement. Before the optimizations, for a particular system, histograms can be constructed for references post optimization. After the optimizations to a particular system have been made, histograms can be constructed again and the two histograms, i.e. the one before optimization and the one after optimization can be compared to study the improvements made, the discrepancies which may have crept in during the optimization and the differences in the level of stability of the two systems.

To construct a histogram, histogram is selected from the data analysis window. Now, the histogram window appears on the screen which asks for the necessary fields for the construction of the histogram. Specify the values that have to be studied as the input range, secondly, specify the bin range in accordance to the input range. Select where the output would like to be seen by specifying the Output range and check the Chart output option.



Input
Input Range: $B$2:$B$21
Bin Range: $F$7:$F$15
□ Labels
Output options
● Output Range: $H$3
○ New Worksheet Ply:
○ New Workbook
□ Pareto (sorted histogram)
□ Cumulative Percentage
☑ Chart Output

Figure 5: Histogram dialog box

To understand the use of this from a practical point of view, consider an organization which is worried about its low employee quorum due to health problems. To solve this, an analysis of the current age scenario at the organization is made with the help of a histogram. The following was the result obtained,

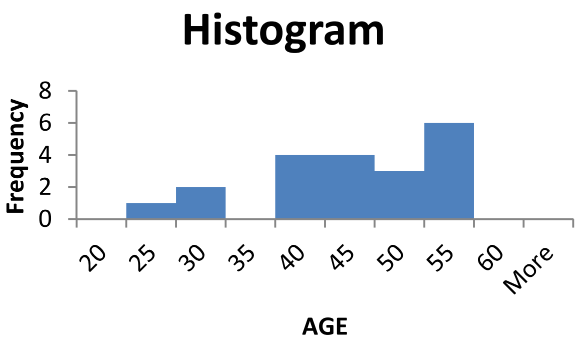
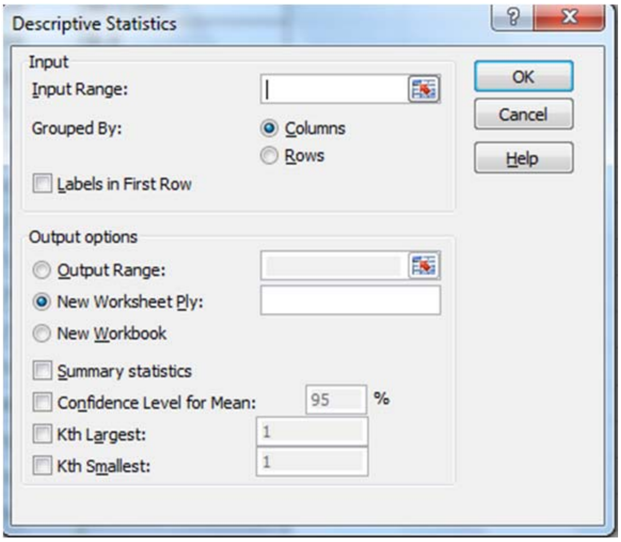


Figure 6: Histogram of age and no. of health problems

Upon studying the above histogram, it was observed that the employee age is more on the above 40 side, making it obvious to account for the reason of health problems. The organization could now make necessary amendments to recruit more people on the younger side.

2. Descriptive Statistics

Excel already provides with the utilities of formulae like sum, difference etc. These formulae could be used for statistical functions like mean, median etc. by combining the results produced by the different formulae. However, Excel provides for a simpler and faster way to perform these functions. They are called descriptive statistics and fall under the data analysis option under the data tab. Descriptive statistics are a set of terminologies that summarize a given set of data along with measures of distributiveness or dispersion. However, it must be understood that the descriptive statistics are just descriptive in nature and do not involve generalization after a certain limit.



Descriptive Statistics

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Figure 7: Descriptive statistics dialog box

The column or data set to be analyzed is specified in the input range. The output range points to the location where the output has to be specified, and the summary statistics needs to be checked.

To understand this in practical sense, we shall consider the same situation as the previous one where an organization is looking at analyzing the age of a sample set of its employees and in the process obtain the descriptive statistics.

Mean	43.65
Standard Error	2.122901441
Median	44.5
Mode	54
Standard Deviation	9.493903861
Sample Variance	90.13421053
Kurtosis	-0.683380139
Skewness	-0.562701947
Range	30
Minimum	25
Maximum	55
Sum	873
Count	20

Table 1: Descriptive statistics of Age of employees

The above table provides a general description of the data set of age of the employees.

a. Mean: It depicts the average age of the employees.
b. Standard Error: It points to the expected error in the mean if we consider other samples from the same population.
c. Median: It renders the middle value of the age of the employees.
d. Mode: It gives the value of the age which occurs the most number of times in the given data set.
e. Standard Deviation: This expresses the amount by which the members of the age data set differ from the mean.
f. Sample Variance
g. Kurtosis: It depicts whether the data is heavy-tailed or light-tailed.
h. Skewness: It is the measure of symmetry or the lack of symmetry of a data set.
i. Range: It points to the range of the data set over which the values are spread.
j. Maximum: Largest value of age in the data set.
k. Minimum: Least value of age in the data set.
l. Sum: The addition of all the ages.
m. Count**:** The number of ages in the data set.

3. ANOVA

There are often situations involved with data, where different set of data are exposed to different set of conditions and then their means need to be checked if they are equal. A null hypothesis is proposed and depending upon the output of the Anova, this hypothesis can be accepted or rejected. This could be understood better with the help of an example of a hospital which needs to carry out some research on the effects of a new drug in the market and another drug which it used earlier for the same medical purposes. In this situation, the null hypothesis would be the fact that when the two types of drugs were used on two different set of patients and they produced similar results for both the set of patients. On the other hand, the null hypothesis would be rejected if they produced different results.

Thus, in these cases where the means of the different groups of the same data need to be studied, ANOVA or the Analysis of Variance is of great help.

Excel provides for 3 techniques of ANOVA:-

1. Anova: Single Factor
2. Anova: Two Factor without Replication
3. Anova: Two Factor with Replication

To understand the difference between the different types of ANOVA, we shall consider the example of an international university which has students admitted locally as well as globally and from both the genders.

Considering a situation where a professor conducts quizzes at three occasions during a semester. Once, before beginning a topic. Second, after the completion of the topic. Third, 4 weeks after the completion of the topic. In this situation, a Single Factor Anova would be useful is seeing how the performance of the candidate varies over the course of the topic being taught.

Now, if the professor wishes to study the performance of local and international students in the same test apart from studying their individual performances, then the Anova: Two Factor without Replication comes handy.

A third situation, which involves study of the grasping power of different genders at the different times in addition to the analysis of individual student performances, then it makes sense to make use of Anova: Two factor with Replication.

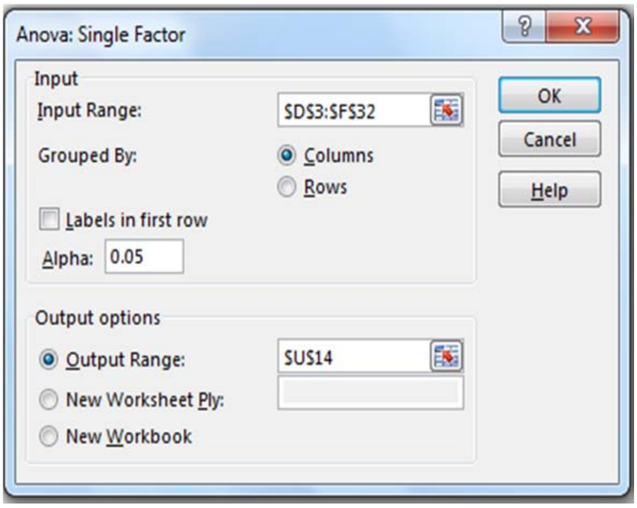
Anova can be implemented by selecting the required Anova technique in the data analysis window which pops up on clicking the data analysis window under the data tab.

Consider, the following sample data.

NAME	**QUIZ 1**	**QUIZ 2**	**QUIZ 3**
Norah Resler	78	69	63
Becky Chitwood	93	83	76
Magnolia Sletten	73	76	75
Lavenia Devaul	64	86	66
Raelene Kincheloe	92	95	85
Jed Franko	94	81	69
Joella Coloma	71	75	79
Donetta Hiner	38	79	58
Siu Kempton	88	85	95
Alison Shupp	47	53	43
Kym Sleeper	43	41	78
Toshia Canchola	92	47	46
Augusta Principato	67	90	61
Sumiko Seibold	85	93	70
Buena Treacy	85	91	75
Adelaide Sartin	92	92	100
Nathanial Lagarde	76	62	57
Hildegard Hornick	59	65	47
Elenore Jess	61	79	56
Mauricio Nguyen	69	92	95
Corrie Mines	78	69	82

Table 2: Quiz scores of students

To perform ANOVA, analysis on selecting the anova: single Factor option. We encounter the Anova window.



Anova: Single Factor

Input

Input Range: $D$3:$F$32

Grouped By: ● Columns ○ Rows

□ Labels in first row

Alpha: 0.05

Output options

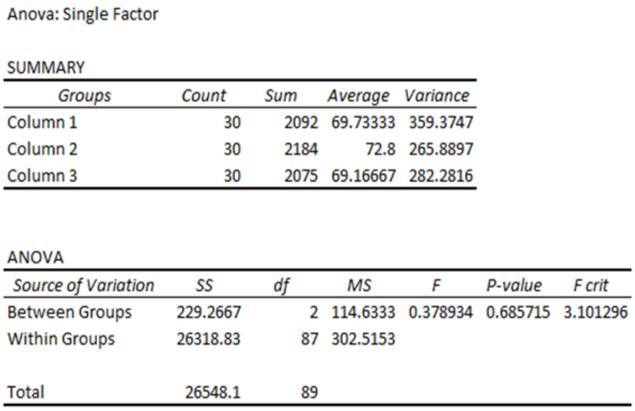
● Output Range: $U$14

○ New Worksheet Ply:

○ New Workbook

Figure 8: Anova: Single factor dialog box

Upon selecting, the range of the input values which includes the scores in the three columns, the location where output has to be seen and clicking OK, we observe the following.



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	30	2092	69.73333	359.3747
Column 2	30	2184	72.8	265.8897
Column 3	30	2075	69.16667	282.2816

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	229.2667	2	114.6333	0.378934	0.685715	3.101296
Within Groups	26318.83	87	302.5153			
Total	26548.1	89				

Figure 9: Results of anova: single factor

Depending on the values obtained in the ANOVA table, we either accept the hypothesis or reject it. If the value of F is greater than the value of F CRIT then, we reject the hypothesis, otherwise, accept it. In the case above, we can see that F<F-crit, therefore we accept the hypothesis at a confidence level of 95% that there is no significant difference between the means of the different groups.

4. F-Test

Another type of analysis that can be performed on the set of data includes the F-test. This test is useful to analyze the difference in the variances between different data sets. We initially propose a hypothesis and then accept or reject it on the basis of the F and the F-crit values that are obtained. Variance is basically the quality of being divergent or different from the other members of the data set. With the help of F-tests, we can compare the variances between two data sets, thereby studying which one is more diverge in nature.

To implement, F-test, the F-test option is selection from the data analysis window by clicking on the data analysis button under the data tab. In the window, which opens up, we enter select the two data sets, specify the output location and click on OK.

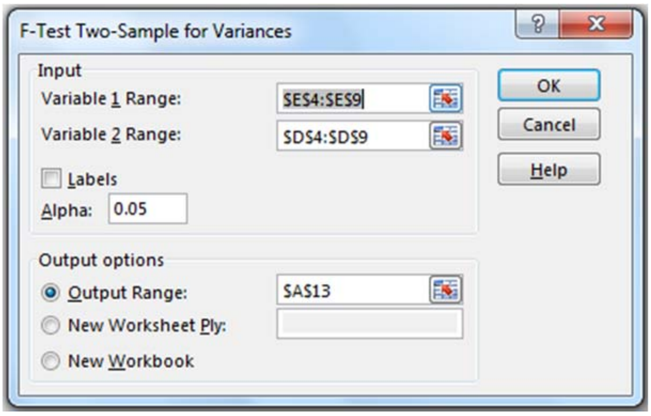
The practical application of this test can be elaborated by the following example.

The table consists of ages of a set of male and a set of female customers of a grocery store.

Male	Female
10	1
16	7
25	9
26	34
30	47
37	51

Table 3: No. of customers at a grocery store

To study the difference of the variances, we implement the F-test.



F-Test Two-Sample for Variances

Input

Variable 1 Range: $E$4:$E$9

Variable 2 Range: $D$4:$D$9

□ Labels

Alpha: 0.05

Output options

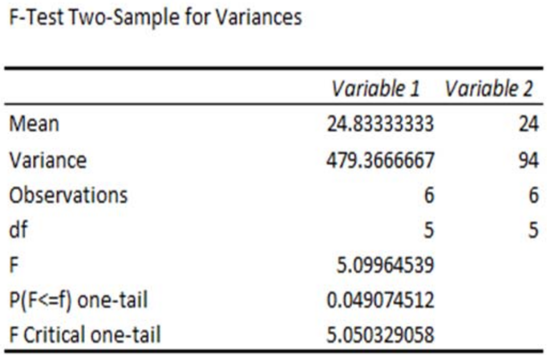
● Output Range: $A$13

○ New Worksheet Ply:

○ New Workbook

Figure 10: F-Test dialog box

The output obtained is as shown.



F-Test Two-Sample for Variances

	Variable 1	Variable 2
Mean	24.83333333	24
Variance	479.3666667	94
Observations	6	6
df	5	5
F	5.09964539	
P(F<=f) one-tail	0.049074512	
F Critical one-tail	5.050329058	

Figure 11: Results of F-Test

It must be taken care of that the Variance of variable 1 must be greater than the Variance of variable 2. If not, then the two data sets must be swapped. This is because F is a ratio of variance of variable 1 to the variance of variable 2.

If F>F-crit, then we reject the hypothesis otherwise we accept it. In the above case, we reject the hypothesis that the variances of the two age sets do not differ significantly since 5.099>5.050(F>F-critical)

5. Moving Averages

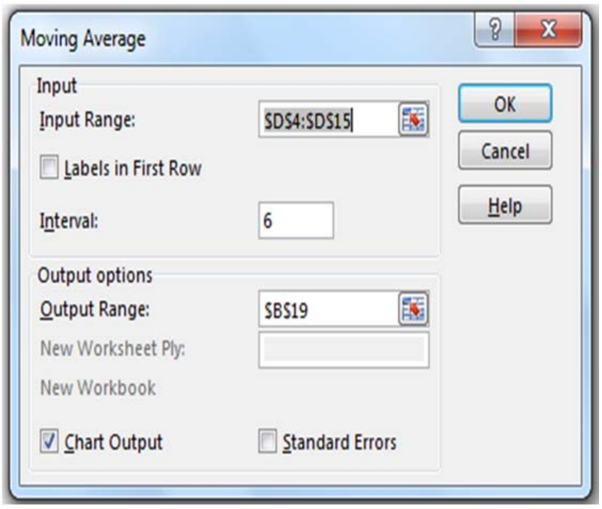
A major type of analysis in organizations involves the study of some kind of a data over a period of time. For example, a company's average profit study for the last 5 years, or any kind of trends which change over time. Another example could be the supply chain organizations, to understand the trend of demands and prepare itself with expected quantities of goods that are going to be needed well in advance. To visualize this kind of data in the appropriate way, excel provides for the construction of a moving average chart. The moving average chart depicts the changes in the trend along with the changes in the average as the trends change. Another aspect of the moving averages is the fact that it is used to smooth out the peaks and valleys to easily recognize trends. To understand the moving averages better, consider the following data,

Year	Profits
1	1.21
2	1.25
3	1.45
4	1.64
5	1.53
6	1.58
7	1.61
8	1.89
9	1.94
10	2.02
11	2.12
12	1.8

Table 4: Yearly profits of a company

The above table gives a relation between the year number (Since the organization started) and the profits under by the organization in million dollars.

In the Moving Average window, we specify the input range as the values whose average trend over time has to be observed. Depending on the interval needed, the moving average graph is shown. For example, for an interval of 6, the moving average is the average of the 5 previous points and the current data point. Therefore, for an interval of 6, we can also conclude that excel will not be able to calculate the moving average for the first 5 data points since they would not be enough, considering the interval is 6.



Moving Average

Input

Input Range: $D$4:$D$15

□ Labels in First Row

Interval: 6

Output options

Output Range: $B$19

New Worksheet Ply:

New Workbook

☑ Chart Output □ Standard Errors

Figure 12: Moving average dialog box

On similar lines, the charts can be obtained for different intervals.

Below are the moving averages for the same data but intervals of 2, 4 and 6.

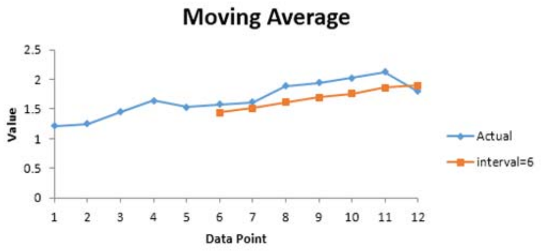


Figure 13: Moving Average for Interval=6

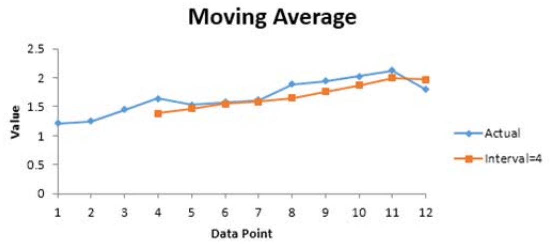


Figure 14: Moving Average for Interval=4

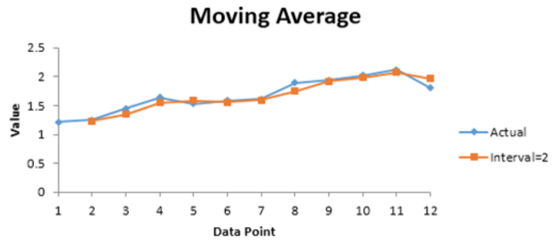


Figure 15: Moving Average for Interval=2

Another important observation to be noticed in case of moving averages is the fact that larger the interval is, smother are the valleys and the peaks. This is because, shorter the interval, closer are the moving average points to the data points. Thus, the moving average is less smooth in nature.

6. Exponential Smoothing

There exists another technique used to study the average of trends over times which is called as the exponential smoothing. It is very similar to the moving averages. The only difference which exists between these two is the fact that exponential weighs the latest trends more than the earlier ones. The moving averages on the other hand, gives all the values an equal weightage. Both of them are highly similar because they are interpreted in similar ways and are commonly used by the technical analysts to smooth out fluctuations. Since Exponential smoothing lays more emphasis on the recent data than the earlier ones, they are more reactive towards latest modifications in data which makes the results more at par with the recent changes, making it a popular technique among the analysts. If we consider the same table as specified previously,

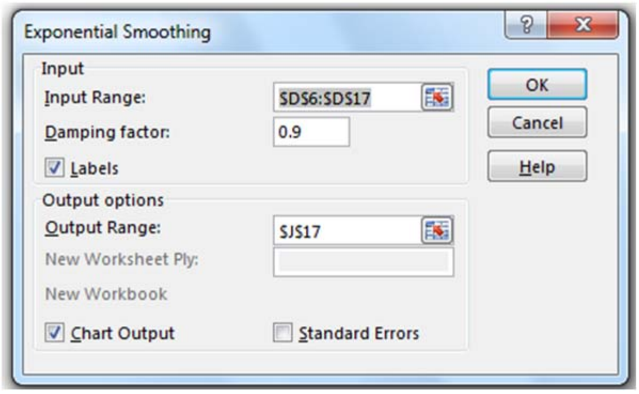
Year	Profits
1	1.21
2	1.25
3	1.45
4	1.64
5	1.53
6	1.58
7	1.61
8	1.89
9	1.94
10	2.02
11	2.12
12	1.8

Table 5: Yearly Profits of a company

The exponential smoothing window needs to be fed in with the input range of values, the output location and a damping factor. This concept of damping factor is useful for determining the weighing factor for the latest values. The relation between α and damping factor is:

Damping factor + Smoothing factor (α) =1

If we consider a situation where we set the damping factor to 0.9, α becomes equal to 0.1. Thereby, giving the previous data points a relatively smaller weight (0.1) as compared to the weight of the recent data point (0.9).


Exponential Smoothing

Input

Input Range: $D$6:$D$17

Damping factor: 0.9

☑ Labels

Output options

Output Range: $J$17

New Worksheet Ply:

New Workbook

☑ Chart Output

□ Standard Errors

Figure 16: Exponential Smoothing dialog box

The output of the chart obtained for the above set of conditions is:

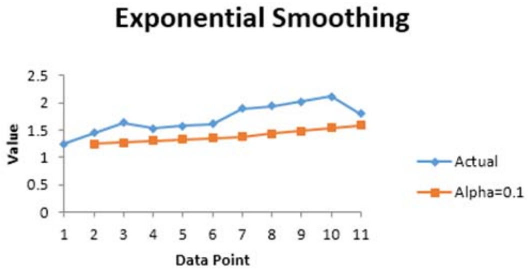


Figure 17: Exponential smoothing chart

Yet again, for this technique, it can be observed that smaller the value of the damping factor is, more the valleys and peaks are smoothed out.

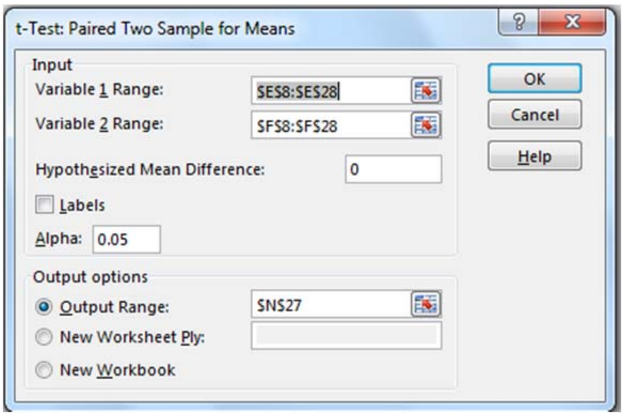
7. T-test

The T-test is quite similar to the anova technique of data analysis. The only difference which exists between t-test and the anova is that anova deals with testing for equal means among multiple data sets, t-test can perform the check only among two data sets. The t-test is used to test the null hypothesis that the mean of two sets of data are the same. The aspect here that needs to be kept in mind is that only 2 sets of data can be checked for equal means. Apart from this, the hypothesis can be accepted or rejected on the basis of the output obtained by performing the T-test. This is on similar lines with the anova. Consider the same set of data as mentioned under the Anova subheading but we shall consider the test marks of only two quizzes, considering the constraint of t-tests to work with a maximum of two sets of data.

NAME	**QUIZ 1**	**QUIZ 2**
Norah Resler	78	69
Becky Chitwood	93	83
Magnolia Sletten	73	76
Lavenia Devaul	64	86
Raelene Kincheloe	92	95
Jed Franko	94	81
Joella Coloma	71	75
Donetta Hiner	38	79
Siu Kempton	88	85
Alison Shupp	47	53
Kym Sleeper	43	41
Toshia Canchola	92	47
Augusta Principato	67	90
Sumiko Seibold	85	93
Buena Treacy	85	91
Adelaide Sartin	92	92
Nathanial Lagarde	76	62
Hildegard Hornick	59	65
Elenore Jess	61	79
Mauricio Nguyen	69	92
Corrie Mines	78	69

Table 6: Quiz scores of students at a University

The T-test window needs to be fed in with the both the input ranges separately. There exists a hypothesized mean difference field which needs to be filled in with the value 0 since the hypothesis we are considering states that there is no significant difference in the means of the two data sets.


t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $E$8:$E$28

Variable 2 Range: $F$8:$F$28

Hypothesized Mean Difference: 0

□ Labels

Alpha: 0.05

Output options

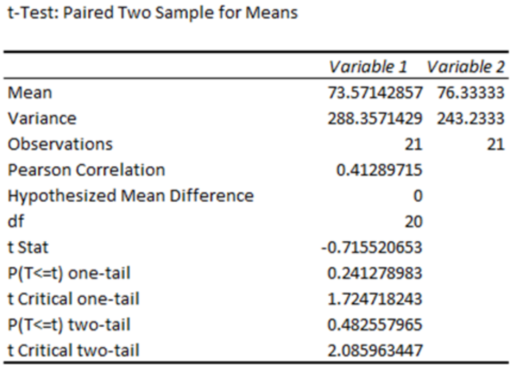
● Output Range: $N$27

○ New Worksheet Ply:

○ New Workbook

Figure 18: T-test dialog box

The output obtained for the above set of conditions:


t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	73.57142857	76.33333
Variance	288.3571429	243.2333
Observations	21	21
Pearson Correlation	0.41289715	
Hypothesized Mean Difference	0	
df	20	
t Stat	-0.715520653	
P(T<=t) one-tail	0.241278983	
t Critical one-tail	1.724718243	
P(T<=t) two-tail	0.482557965	
t Critical two-tail	2.085963447	

Figure 19: Result of T-test

This technique involves a two-tail test, if t-Stat < -t critical two tail or t Stat > t Critical two tail, then, we reject the hypothesis. For the case we have considered, this condition is not satisfied, therefore, we cannot reject the hypothesis and the difference between the means observed is too small to be considered significant.

8. Correlation

Often, there are situations where there is a need to understand how are two data sets related to each other. In other words, when the values of one data set increase, the other data set may increase or decrease. Statistically, the correlation factor always varies between -1 and 1. A practical application of the correlation could be in the field of technical stock market analysis where we would want to identify the correlation between market indicators and specific stocks. For example, the relation between customer spending and the price of a stock. If customers spend more on the products of that company, its stock price will rise, and if he customer spending goes down, the stock price will also decrease. A correlation coefficient of -1 indicates negative correlation, which indicates that if the value of one data set increases, the value of the other data set will decrease. A correlation coefficient of +1 indicates positive correlation, which implies that if the values of one data set is increasing, the values of the other data set also increase.

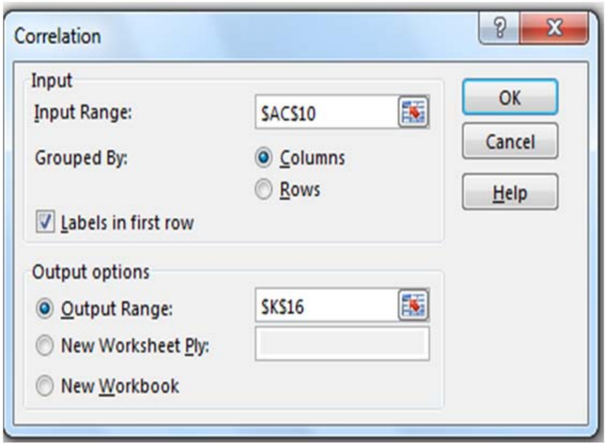
Correlation must not be confused with causation i.e. one data set is not causing the other. For better understanding on practical purposes, the following data sets are considered.

AGE	SALARY	TIME SINCE DOJ
23	25000	6
22	22000	3
21	20000	1
25	30000	36
24	27000	24
41	50000	12
35	40000	120
28	34000	17
38	45000	36
51	55000	20
21	19000	1

Table 7: Age, salary and time since date of joining of employees

For the data mentioned above, we find out the correlation between the three data sets.

The correlation window opens up when we select the correlation option in the data analysis window.



Correlation

Input

Input Range: $AC$10

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

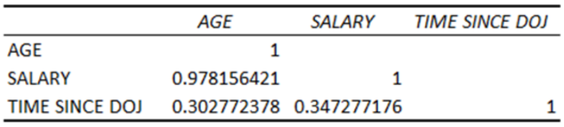
● Output Range: $K$16

○ New Worksheet Ply:

○ New Workbook

Figure 20: Correlation dialog box

The output obtained according to the conditions given above is shown below.



	AGE	SALARY	TIME SINCE DOJ
AGE	1		
SALARY	0.978156421	1	
TIME SINCE DOJ	0.302772378	0.347277176	1

Figure 21: Result of Correlation

The correlation between age and salary can be referred to as a positive one since the value is approximately 1. Thus, it would be safe to conclude for the organization that with age, the salary of its employees would increase.

On the other hand, the correlation between the Time since date of joining and salary, and between time since date of joining and age are approximately 0.3, which clearly indicates that there is little or no correlation between these fields.

LIMITATIONS

Even though MS excel is a convenient tool for data analysis, it does have some limitations. It lacks certain tools like boxplots, which are widely used in statistical analysis. There is concern over the format of output of some specific functions. [1]. Missing values are sometimes handled inconsistently and incorrectly. Different analyses require the data to be arranged in various ways. If a variety of different tests are to be performed, data might need to be rearranged multiple times. Output may be scattered in many different worksheets, or all over one worksheet and it may be incomplete or may not be properly labelled, increasing possibility of misidentifying output. It does not maintain a record of what was done to generate the results, making it difficult to document the analysis, or to repeat it at a later time. [2]

CONCLUSION

Various tools and techniques of analysis toolpak were explained and demonstrated. Analysis toolpak was used to create histograms to solve the problem of low employee quorum due to health problems, obtain and analyze descriptive statistics of ages of a sample set of employees in an organization, analyze variance between set of scores of a test quiz conducted by a university which enrolls both local and global students of both genders using anova, F-test and t-test techniques, study changes in profits of an organization over a period of time using moving averages, exponential smoothening and study correlation between age and salary of employees in an organization.

REFERENCES

[1] Management Research: Applying the Principles, Susan Rose, Nigel Spinks & Ana Isabel Canhoto.

[2] Using Excel for Statistical Data Analysis – Caveats**,** Eva Goldwater, Biostatistics Consulting Center, University of Massachusetts School of Public Health.

[3] Excel easy, Analysis toolpak [Online]. Available at http://www.excel-easy.com/data-analysis/analysis-toolpak.html.