

Business Analytics

Subject code:AB207

Module-1

(Introduction to data analytics)

Introduction to Data:

Since the invention of computers, people have used the term data to refer to computer information, and this information was either transmitted or stored. But that is not the only data definition; there exist other types of data as well. So, what is the data? Data can be texts or numbers written on papers, or it can be bytes and bits inside the memory of electronic devices, or it could be facts that are stored inside a person's mind.

What is Data?

Now, if we talk about data mainly in the field of science, then the answer to "what is data" will be that data is different types of information that usually is formatted in a particular manner. All the software is divided into two major categories, and those are programs and data. Programs are the collection made of instructions that are used to manipulate data.

Types and Uses of Data:

Growth in the field of technology, specifically in smartphones has led to text, video, and audio is included under data plus the web and log activity records as well. Most of this data is unstructured. The term Big Data is used in the data definition to describe the data that is in the petabyte range or higher. Big Data is also described as 5V's (variety, volume, value, veracity, and velocity)

Nowadays, web-based e-Commerce has spread vastly, business models based on Big Data have evolved, and they treat data as an asset itself. And there are many benefits of Big Data as well, such as reduced costs, enhanced efficiency, enhanced sales, etc. The meaning of data expands beyond the processing of data in computing applications. When it comes to what data science is, a body made of facts is called data science. Accordingly, finance, demographics, health, and marketing also have different meanings of data, which ultimately make up different answers for what is data.

How to Analyze Data?

Ideally, there are two ways to analyze the data:

1. Data Analysis in Qualitative Research
2. Data Analysis in Quantitative Research

Data Analysis in Qualitative Research:

Data analysis and research in subjective information work somewhat better than numerical information as the quality information consists of words, portrayals, pictures, objects, and sometimes images. Getting knowledge from such entangled data is a confounded procedure; thus, it is usually utilized for exploratory research as well as data analysis.

Data Analysis in Quantitative Research

The primary stage in research and analysis of data is to do it for the examination with the goal that the nominal information can be changed over into something important. The preparation of data comprises the following.

- a) Data Validation
- b) Data Editing
- c) Data Coding

Top 5 Jobs in Data:

Mentioned below are the names of a few job titles that are high in- demand.

1. **Data Scientist:** This is one of the most in-demand jobs right now.
2. **BIA:** Business Intelligence Analysts help the companies to make fruitful decisions with the help of using data and making the required recommendations.
3. **Database Developer:** Database developer mainly focus on improving the databases and developing new applications for better use of data.
4. **Database Administrator:** The job of a Database administrator is to setup the databases then maintain and secure them at all times.
5. **Data Analytics Manager:** Nowadays, more and more companies are starting to rely on data managers to extract out the most useful information from massive amounts of data.

BUSINESS ANALYTICS

What are analytics?

You've probably heard the term thrown around a lot, but it's always good to start at the beginning and define exactly what it means. As defined by Wikipedia: "Analytics is the discovery, interpretation, and communication of meaningful patterns in data. In the context of Internet marketing, this means things like user demographics, how users interact with your website, mapping out conversion paths, and so on. When we talk about measuring analytics, this almost always refers to the use of Google Analytics, which is Google's free analytics platform. Marketers sometimes distinguish between "web analytics" and "marketing analytics" as well. Web analytics refers to website-specific metrics like page load times, time spent on a certain page, page views per visit, and so on. Marketing analytics refers to things like overall traffic, leads acquired, leads converted, your sales funnel, and so on. Both web and marketing analytics are necessary to make fully informed decisions regarding your web presence and performance.

Why are they important?

Internet marketing without analytics is like pointing and saying "the building is somewhere in there." Internet marketing with analytics is asking to pinpoint the exact location of a specific building using satellite technology. You can't really make informed marketing decisions without proper analytics. Throwing mud at a wall and seeing what sticks might work, but how can you replicate that success if you don't know why it stuck in the first place? Analytics allow you to quantify the effects of making a change to your marketing strategy, and that's invaluable to the process of improving and optimizing online marketing campaigns. The biggest benefit of utilizing proper analytics is being able to identify strengths and weaknesses. For example, let's say you run a blog for your car detailing business. You're just starting out, and aren't sure what kinds of posts will bring you the most traffic, or provide the most value to your readers. If you're using analytics, you'll be able to measure which blog posts attract the most traffic, which get the least traffic, which have a high bounce rate, a low bounce rate, and so on. It will be easy to tell which blog posts are performing better or worse than others. You can then take this information to further improve your blogging process. The more accurate your perception of what works and what doesn't, the less time you'll waste experimenting with blog topics that end up flopping. Analytics essentially allow for a defined path of optimization that leads to better results on all fronts.

BUSINESS ANALYTICS

What is Business Analytics:

Business analytics or simply analytics is the use of data, information technology, statistical analysis, quantitative methods and mathematical or computer based models to help managers gain improved insight about their business operation and make better, fact based decisions. Business analytics is a process of transforming data into actions through analysis and insights in the context of organizational decision making and problem solving Business analytics is supported by various tools such as microsoft excel and various excel add-ins ,commercial statistical software packages such as SAS or minitab and more complex business intelligence suites that integrate data with analytical software

Data for Business Analytics:

Since the emerging of the electronic age and the internet has increased both individuals and organizations have had access to an enormous wealth of data and information .Data is the combination of numerical facts and figures that are collected through some type of measurement process. Information comes from analyzing data i.e.. Extracting meaning from data to support evaluation and decision making Data is used virtually in every major function in a business. Some examples of how data is used in business includes the following:

1. Annual reports summarize data about companies profit and market share both in numerical form and in charts and graphs to communicate with shareholders
2. Accounts conduct audits to determine whether figures reported on a firm's balance sheet fairly represent the actual data by examining samples of accounting data such as accounts receivable.
3. Financial analysts collect and analyze a variety of data to understand the contribution that a business provides to its shareholders .These typically include profitability ,revenue growth, return on investment, asset utilization ,operating margins, earnings per share ,economic value added(EVA),shareholder value and other relevant measures.
4. Economists use data to help companies understand and predict population trends, interest rates, industry performance, consumer spending and international trade .such data are often obtained from external sources such as standard and poor datasets, industry trade associations or government databases .
5. Marketing researchers collect and analyze extensive customer data. These data often consists of demographics preferences and opinions transactions and payment history shopping behavior and a lot more. such data may be collected by surveys personal interviews, focus groups or from shopper loyalty cards

BUSINESS ANALYTICS

6. Operations manager use data on production performance, manufacturing quality, delivery times, order accuracy, supplier performance, productivity, costs and environmental compliance to manage their operations.

7. Human resource managers measure employee satisfaction, training costs, turnover, market innovation, training effectiveness and skills development

Data sets and Databases:

Data set is simply a collection of data

Ex: Marketing survey data, Table of stock prices etc

Database is collection of related files containing records on people, place or things. The people, place or things for which we store and maintain information are called Entities.

Big Data:

Big data refers to massive amount of business data from a wide variety of sources, much of which is available in real time, and much of which is uncertain or unpredictable. Big data has five characteristics: 5v's

1. Volume
2. Velocity
3. Variety
4. Veracity
5. value

Most often, big data revolves around customer behavior and customer experience .Big data provides an opportunity for organizations to gain a competitive advantage.

BUSINESS ANALYTICS

Metrics:

A metric is a unit of measurement that provides a way to objectively quantify performance. Ex: Senior managers might assess overall business performance using such metrics as net profit, return on investment, market share and customer satisfaction. Metrics can be either discrete or continuous .A discrete metrics is one that is derived from counting something.

Ex: A delivery is either on time or not

Ex: An order is complete or incomplete

A continuous metrics are based on a continuous scale of measurement.

Ex: Dollars, length, time, volume, weight.

Data Classification:

Classification of data is by the type of measurement scale .Data may be classified into four groups:

1. Categorical(nominal) data: This data is sorted into categories according to specified characteristics Ex: Geographical region
2. Ordinal data: This data is ordered or ranked according to some relationship to one another Ex: Rating a service as poor, average, good, very good, excellent.
3. Interval data: The data which have constant differences between observations and have arbitrary zero points. Ex: Time and Temperature
4. Ratio data: This data is continuous and have a natural zero. Most business and economic data such as dollars fall into this category. Ex: Dollars

Data Reliability and validity:

Data used in business decisions need to be reliable and valid. Reliability means data should be accurate and consistent. Validity means data should measure correctly what it suppose to measure. Ex: A tire pressure gauge that consistently read several pounds of pressure below the true value is not reliable but it is valid because it does measure tier pressure. Ex: The number of calls to a customer service desk might be counted correctly each day which is a reliable measure but not valid because if it assess to customer dissatisfaction.

BUSINESS ANALYTICS

Problem solving with analytics:

The fundamental purpose of analytics is to help managers solve problems and make decisions. Problem solving is the activity associated with defining, analyzing, and solving a problem and selecting an appropriate solution that solves a problem. Problem solving consists of several phases:

1. Recognizing a problem
2. Defining the problem
3. Structuring the problem
4. Analyzing the problem
5. Interpreting results and making a decision
6. Implementing the solution

Business analytics in practice:

Hewlett packard(HP) uses analytics extensively .Many applications are used by managers with little knowledge of analytics .These require that analytical tools be easily understood .Based on years of experience , HP analysts compiled some key lessons . Before creating an analytical decision tool, HP asks three questions:

1. Will analytics solve the problem?
2. Can we leverage an existing solution?
3. Is a decision model really needed?

Once a decision is made to develop an analytical tool, they use several guidelines to increase the chances of successful implementation:

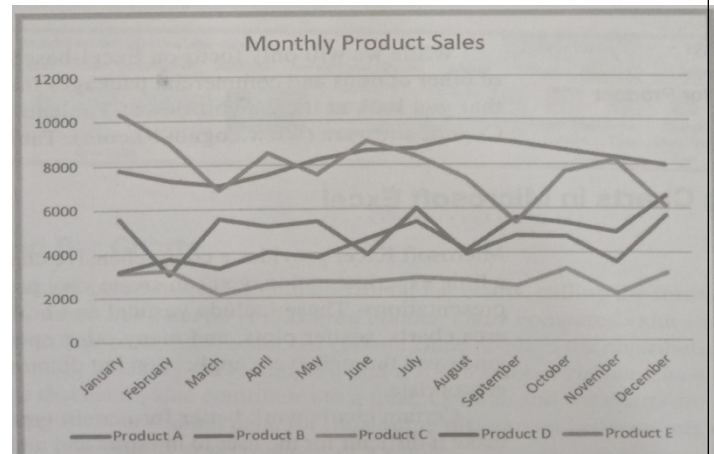
- Use prototyping
- Build insight, not black boxes
- Remove unneeded complexity
- Partner with end users in discovery and design
- Develop an analytic champion

BUSINESS ANALYTICS

Data Visualization:

Data visualization is the process of displaying data in a meaningful fashion to provide insights that will support better decisions. Researchers have observed that data visualization improves decision making and provides managers with better analysis capabilities that reduce reliance on IT professionals and improves collaboration and information sharing.

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088



Monthly product sales data

Visualization of Monthly product sales data

Data visualization tools:

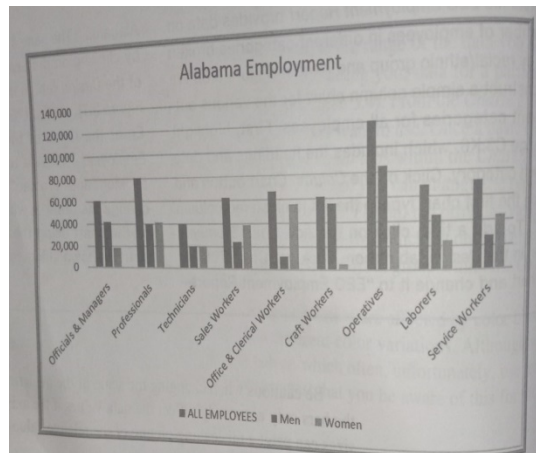
Data visualization ranges from simple excel charts to more advanced interactive tools and software that allow users to easily view and manipulate data with a few clicks, not only on computers but on i-pads and other devices as well. Now we discuss basic tools available in excel.

1. Column and bar charts
2. Line charts
3. Pie charts
4. Area charts
5. Scatter charts
6. Bubble charts

BUSINESS ANALYTICS

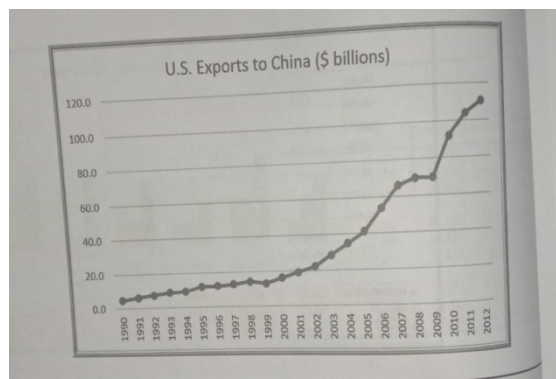
Column and bar charts:

A clustered column charts compare values across categories using vertical rectangles and stacked column chart displays the contribution of each value to the total by stacking the rectangles. Column and bar charts are useful for comparing categorical or ordinal data for illustrating differences b/w sets of values and for showing proportions or percentage of a whole.



Line charts:

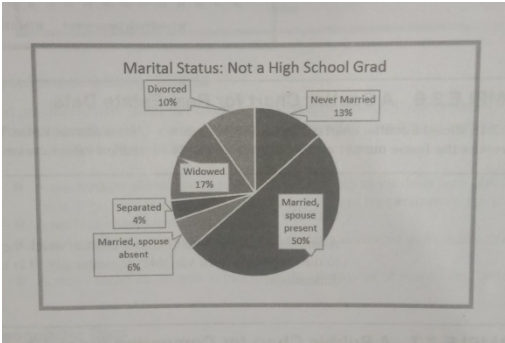
Line charts provide a useful means for displaying data over time. You may plot multiple data series in line chart, however it can be difficult to interpret if the magnitude of the data values differs greatly. In that case, it would be advisable to create separate charts for each data series.



BUSINESS ANALYTICS

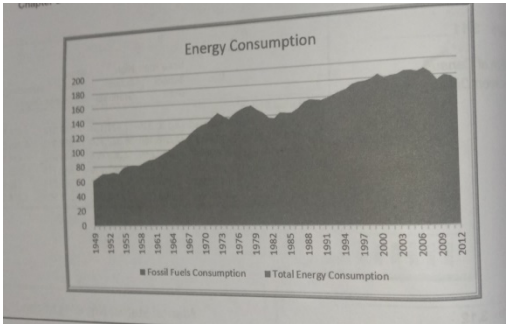
Pie charts:

A pie chart displays data by partitioning into a circle with areas showing the relative proportion. Data visualization professionals don't recommend pie charts because pie charts can be drawn only when data is 100% so that data is divided according to their proportion.



Area charts:

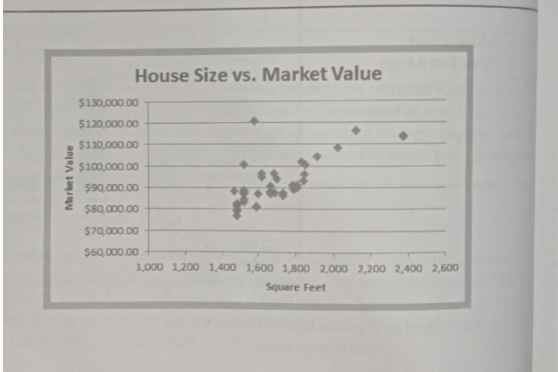
An area chart combines the feature of a pie chart with those of line charts. Area charts present more information than pie or line charts alone but may clutter the observer mind with too many details if too many data series are used. They should be used with care.



BUSINESS ANALYTICS

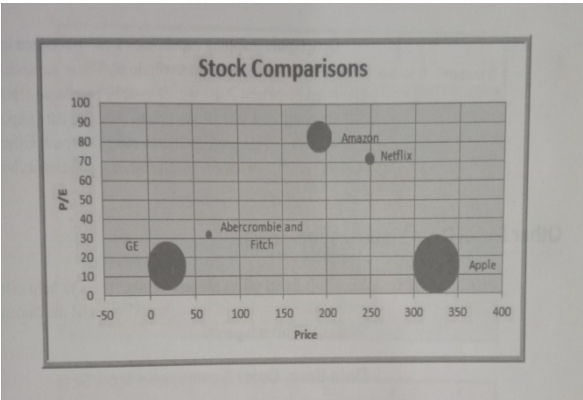
Scatter charts:

Scatter charts show the relationship between two variables. To construct a scatter chart, we need observation that consists of pairs of variables. For example, Students in a class might have grades for midterm and a final exam. A scatter chart would show whether high or low grades on the midterm correspond strongly to high or low grades on the final exam or whether the relationship is weak or nonexistent.



Bubble charts:

A bubble chart is a type of scatter chart in which the size of the data marker corresponds to the value of a third variable consequently it is a way to plot three variables in two dimensions.



Miscellaneous Excel charts:

Excel provides several additional charts for special applications. These additional types of charts can be selected and created from the other charts button in the Excel ribbon. These include the following:

1. A Stock chart allows you to plot stock prices, such as the daily high, low, and close. It may also be used for scientific data such as temperature changes.
2. A Surface chart shows 3-d data
3. A Doughnut chart is similar to a pie chart but can contain more than one data series.
4. A Radar chart allows you to plot multiple dimensions of several data series.

Other Excel data visualization tools:

Microsoft Excel offers numerous other tools to help visualize data.

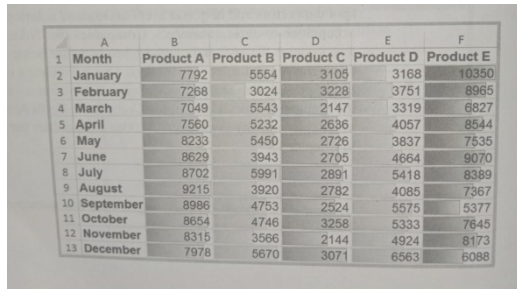
These include the following:

1. Data bars
2. color scales
3. icon sets
- 4 .spark lines
5. excel camera tool.

BUSINESS ANALYTICS

Data bars:

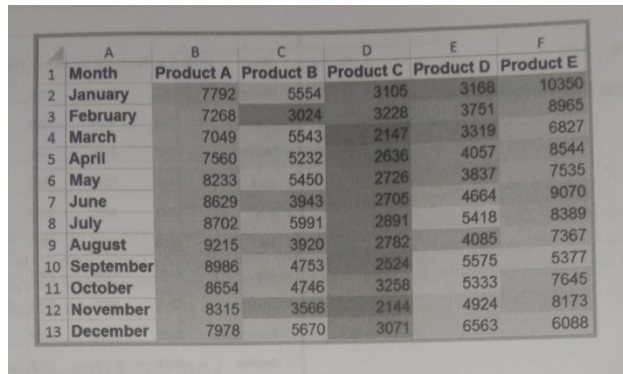
Data bars display color bars that are scaled to the magnitude of the data values but placed directly within the cells of a range. Highlight the data in the each column then click conditional formatting button in the styles group within the home tab and then select data bars and choose the fill option and color.



	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Color scales:

Color scales shade cells based on their numerical value using a color palette. This is another option in the conditional formatting menu.

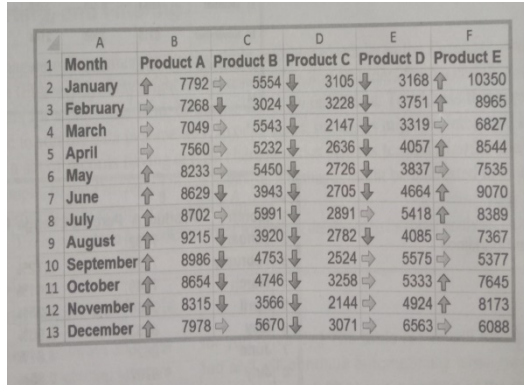


	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

BUSINESS ANALYTICS

Icon sets:

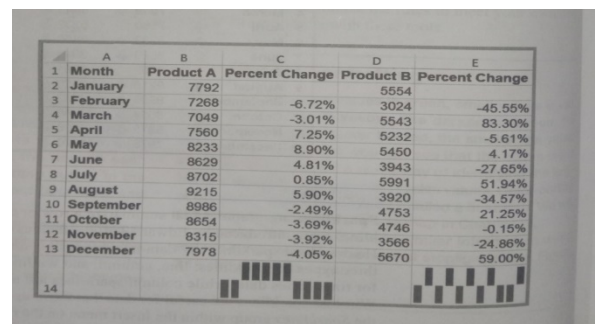
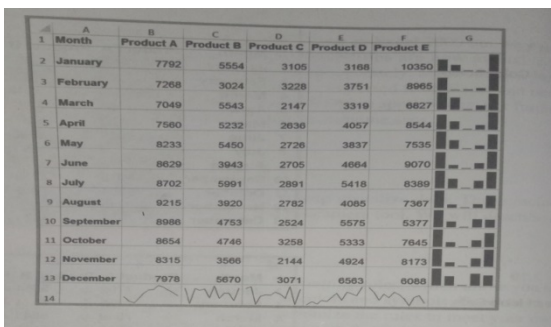
Icon sets provide similar information using symbols such as arrows or stoplight colors.



	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	↑ 7792 →	5554 ↓	3105 ↓	3168 ↑	10350
3	February	→ 7268 ↓	3024 ↓	3228 ↓	3751 ↑	8965
4	March	→ 7049 →	5543 ↓	2147 ↓	3319 →	6827
5	April	→ 7560 →	5232 ↓	2636 ↓	4057 ↑	8544
6	May	↑ 8233 →	5450 ↓	2726 ↓	3837 →	7535
7	June	↑ 8629 ↓	3943 ↓	2705 ↓	4664 ↑	9070
8	July	↑ 8702 →	5991 ↓	2891 →	5418 ↑	8389
9	August	↑ 9215 ↓	3920 ↓	2782 ↓	4085 →	7367
10	September	↑ 8986 ↓	4753 ↓	2524 →	5575 →	5377
11	October	↑ 8654 ↓	4746 ↓	3258 →	5333 ↑	7645
12	November	↑ 8315 ↓	3566 ↓	2144 →	4924 ↑	8173
13	December	↑ 7978 →	5670 ↓	3071 →	6563 →	6088

Spark lines:

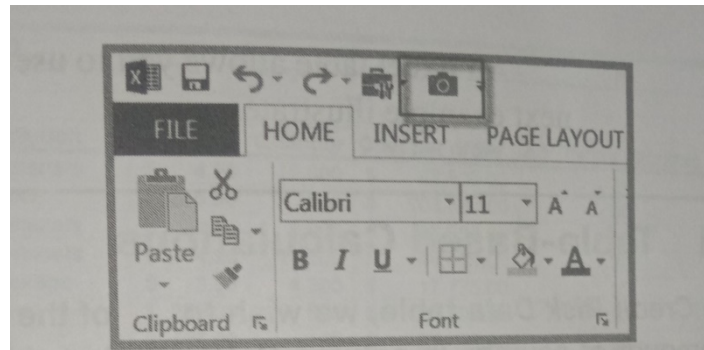
Spark lines are graphics that summarize a row or column of data in a single cell. Excel has three types of spark lines they are line, column and win/loss spark lines. Line spark lines are clearly useful for time series data and column spark lines are more appropriate for categorical data and win/loss spark lines are useful for data that moves up or down over time. They are found in the spark lines group within the Insert menu or button.



BUSINESS ANALYTICS

Excel camera tool:

A little known feature of the excel is the camera tool. This allows you to create live pictures of various ranges from different worksheets that you can place on a single page size them and arrange them easily



Data Queries:

Managers make numerous queries about data. For example, in the purchase order database they might be interested in finding all orders from a certain supplier, all orders for a particular item or tracing orders by order data. To address these queries we need to sort the data in some way. In other cases, managers might be interested in extracting a set of records having certain characteristic. This is termed filtering the data. Excel provides a convenient way of formatting databases to facilitate analysis called Tables.

1. Creating an Excel table
2. Sorting data in Excel
3. Filtering data

BUSINESS ANALYTICS

Statistical methods for summarizing data:

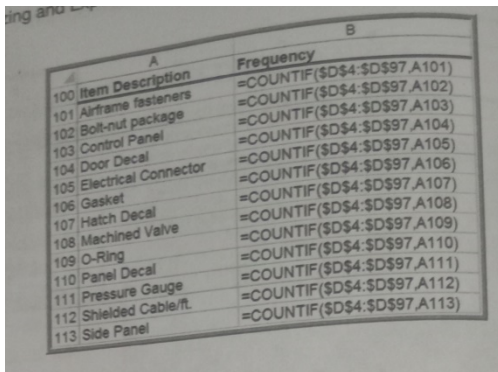
Statistics is defined as science of uncertainty and the technology of extracting information from data. A Statistics is a summary measure of data. Statistics involve collecting, analyzing, interpreting and presenting data. Statistical methods are essential to business analytics and are used throughout the topic. Microsoft Excel supports statistical analysis in two ways:

1. Statistical functions
2. Excel Analysis toolpak

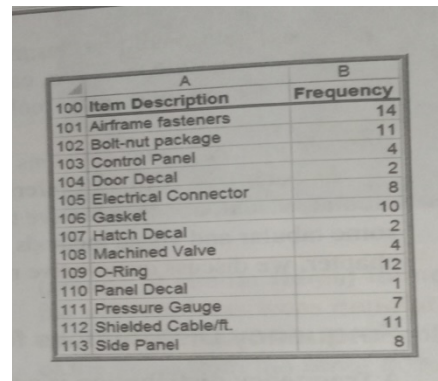
Descriptive statistics refers to methods of describing and summarizing data using tabular, visual and quantitative techniques. In this topic we mostly concentrate on numerical and categorical data.

Frequency distribution for categorical data:

A frequency distribution is a table that shows the number of observations in each of several non-overlapping groups. Categorical variables naturally define the groups in a frequency distribution. Here we use the countif function to count the number of orders placed for each item which means frequency `COUNTIF(D4:D97, cell reference)` where cell reference is the cell containing the item name.



A	B
100 Item Description	=COUNTIF(\$D\$4:\$D\$97,A101)
101 Airframe fasteners	=COUNTIF(\$D\$4:\$D\$97,A102)
102 Bolt-nut package	=COUNTIF(\$D\$4:\$D\$97,A103)
103 Control Panel	=COUNTIF(\$D\$4:\$D\$97,A104)
104 Door Decal	=COUNTIF(\$D\$4:\$D\$97,A105)
105 Electrical Connector	=COUNTIF(\$D\$4:\$D\$97,A106)
106 Gasket	=COUNTIF(\$D\$4:\$D\$97,A107)
107 Hatch Decal	=COUNTIF(\$D\$4:\$D\$97,A108)
108 Machined Valve	=COUNTIF(\$D\$4:\$D\$97,A109)
109 O-Ring	=COUNTIF(\$D\$4:\$D\$97,A110)
110 Panel Decal	=COUNTIF(\$D\$4:\$D\$97,A111)
111 Pressure Gauge	=COUNTIF(\$D\$4:\$D\$97,A112)
112 Shielded Cable/ft.	=COUNTIF(\$D\$4:\$D\$97,A113)
113 Side Panel	



A	B
100 Item Description	Frequency
101 Airframe fasteners	14
102 Bolt-nut package	11
103 Control Panel	4
104 Door Decal	2
105 Electrical Connector	8
106 Gasket	10
107 Hatch Decal	2
108 Machined Valve	4
109 O-Ring	12
110 Panel Decal	1
111 Pressure Gauge	7
112 Shielded Cable/ft.	11
113 Side Panel	8

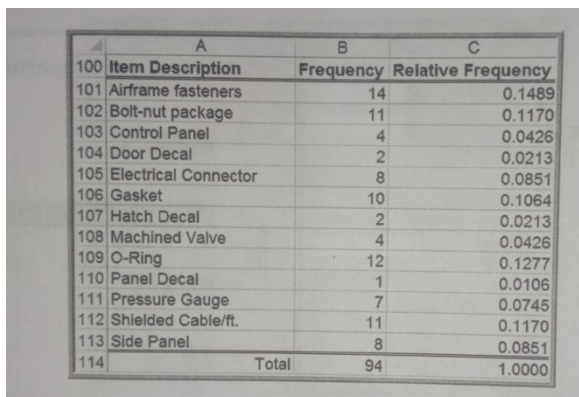
BUSINESS ANALYTICS

Relative frequency distribution:

We may express the frequencies as a fraction or proportion of the total this is called as relative frequency. If a data set has 'n' observations the relative frequency of category 'i' is computed as

Relative frequency of category i = Frequency of category i/n

We often multiply the relative frequency by 100 to express them as percentages. A relative frequency distribution is a tabular summary of the relative frequencies of all categories.



	A	B	C
100	Item Description	Frequency	Relative Frequency
101	Airframe fasteners	14	0.1489
102	Bolt-nut package	11	0.1170
103	Control Panel	4	0.0426
104	Door Decal	2	0.0213
105	Electrical Connector	8	0.0851
106	Gasket	10	0.1064
107	Hatch Decal	2	0.0213
108	Machined Valve	4	0.0426
109	O-Ring	12	0.1277
110	Panel Decal	1	0.0106
111	Pressure Gauge	7	0.0745
112	Shielded Cable/ft.	11	0.1170
113	Side Panel	8	0.0851
114	Total	94	1.0000

Frequency distribution for numerical data:

Here numerical data means discrete values. We construct a frequency distribution similar to categorical data using countif method to count the discrete values. For numerical data that have many discrete values with little repetition or continuous, frequency distribution requires that we define by specifying

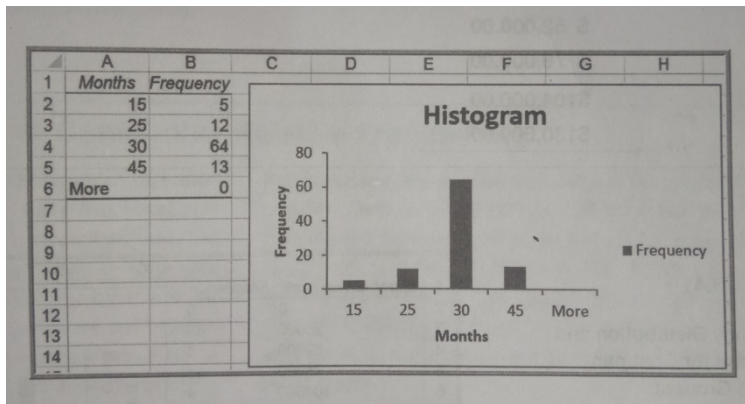
1. The number of groups
2. Width of each group
3. Upper and lower limits of each group

$$\text{Group width} = \frac{\text{Upper limit} - \text{lower limit}}{\text{number of groups}}$$

BUSINESS ANALYTICS

Excel histogram tool:

A graphical depiction of a frequency distribution for numerical data in the form of a column chart is called as Histogram.



Cumulative relative frequency distribution:

For numerical data, we may also compute the relative frequency of observations in each group. By summing all the frequencies at or below each upper limit we will obtain the cumulative relative frequency. The cumulative relative frequency represents the proportion of the total number of observations that fall at or below the upper limit of each group. A tabular summary of relative frequencies is called as cumulative relative frequency distribution.

Percentiles and Quartiles:

Data are often expressed as percentiles and quartiles. Generally speaking, the k^{th} percentile is a value at or below which at least k percent of the observations lie. However, the way by which percentiles are calculated is not standardized. The most common way to compute the k^{th} percentile is to order the data values from smallest to largest and calculate the rank of the k^{th} percentile using the formula

$$nk/100+0.5.$$

Quartiles break the data into four parts. The 25th percentile is called the first quartile(Q1), the 50th percentile is called the second quartile (Q2), the 75th percentile is called the third quartile(Q3), the 100th percentile is called the fourth quartile(Q4).The one-fourth of the data fall below the first quartile, one-half are below the second quartile, and three-fourth are below the third quartile. We may compute quartiles using the Excel function QUARTILE.INC(array,quart) where array specifies the range of the data and quart is a whole number between 1 and 4, designating the desired quartile.

Cross Tabulations:

One of the most basic statistical tools used to summarize categorical data and examine the relationship between two categorical variables is cross-tabulation. A cross-tabulation is a tabular method that displays the number of observations in a data set for different sub-categories of two categorical variables. A cross-tabulation table is often called as contingency table.

Exploring data using pivot tables:

Excel provides a powerful tool for distilling a complex data set into meaningful information. Pivot table allows you to create custom summaries and charts of key information in the data. Pivot tables can be used to quickly create cross-tabulations and to drill down into a large set of data in numerous ways. To apply pivot tables, you need a data set and select any cell in the data set and choose pivot table from the tables group under the insert tab.

1. Creating pivot table
2. Using pivot table report filter
3. Pivot charts
4. Using slicers

Module-2

(Descriptive statistical measures)

WHAT IS DESCRIPTIVE STATISTICS?

Descriptive statistics describes or summarizes the basic features or characteristics of the data. It assigns numerical values to describe the trend of the samples collected. It converts large volumes of data and presents it in a simpler, more meaningful format that is easier to understand and interpret. It is paired with graphs and tables; descriptive statistics offer a clear summary of the data's complete collection. Descriptive statistics indicate that interpretation is the primary purpose, while inferential statistics make future predictions for a larger set of data based on descriptive values obtained. Hence, descriptive statistics form the first step and the basis of quantitative data analysis.

In statistics as well as in quantitative methodology, the set of data are collected and selected from a statistical population with the help of some defined procedures. There are two different types of data sets namely, population and sample. So basically when we calculate the mean deviation, variance and standard deviation, it is necessary for us to know if we are referring to the entire population or to only sample data. Suppose the size of the population is denoted by 'n' then the sample size of that population is denoted by $n-1$. Let us take a look of population data sets and sample data sets in detail.

Population:

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a parameter. For example, All people living in India indicates the population of India.

There are different types of population. They are:

1. Finite Population
2. Infinite Population
3. Existent Population
4. Hypothetical Population

BUSINESS ANALYTICS

Let us discuss all the types one by one.

1. **Finite Population:** The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite. For statistical analysis, the finite population is more advantageous than the infinite population. Examples of finite populations are employees of a company, potential consumer in a market.

2. **Infinite Population:** The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

3. **Existent Population:** The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

4. **Hypothetical Population:** The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. In some situations, the populations are only hypothetical. Examples are an outcome of rolling the dice, the outcome of tossing a coin.

Sample:

It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

Probability sampling

Non-probability sampling

1. Probability Sampling: In probability sampling, the population units cannot be selected at the discretion of the researcher. This can be dealt with following certain procedures which will ensure that every unit of the population consists of one fixed probability being included in the sample. Such a method is also called random sampling. Some of the techniques used for probability sampling are:

a) Simple random sampling

- b) Cluster sampling
- c) Stratified Sampling
- d) Disproportionate sampling
- e) Proportionate sampling
- f) Optimum allocation stratified sampling
- g) Multi-stage sampling

2. Non Probability Sampling:

In non-probability sampling, the population units can be selected at the discretion of the researcher. Those samples will use the human judgement for selecting units and has no theoretical basis for estimating the characteristics of the population. Some of the techniques used for non-probability sampling are:

- a) Quota sampling
- b) Judgement sampling
- c) Purposive sampling

Measures of Location:

Measures of location describe the central tendency of the data. They include the mean, median and mode and midrange.

Mean (or) Average:

The arithmetic mean or average of n observations (pronounced “x bar”) is simply the sum of the observations divided by the number of observation

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{no of observations}}$$

Median:

The median is defined as the middle point of the ordered data. It is estimated by first ordering the data from smallest to largest, and then counting upwards for half the observations. The estimate of the median is either the observation at the center of the ordering in the case of an odd number of observations, or the simple average of the middle two observations if the total number of observations is even. More specifically, if there are an odd number of observations,

BUSINESS ANALYTICS

it is the $[(n+1)/2]$ th observation, and if there are an even number of observations, it is the average of the $[n/2]$ th and the $[(n/2)+1]$ th observations.

BUSINESS ANALYTICS

Mode:

A third measure of location is the mode. This is the value that occurs most frequently, or, if the data are grouped, the grouping with the highest frequency. It is not used much in statistical analysis, since its value depends on the accuracy with which the data are measured; although it may be useful for categorical data to describe the most frequent category.

Midrange:

A fourth measure of location that is used occasionally is the midrange. This is simply the average of the greatest and least values in the data set.

Using measures of location in business decisions:

Because everyone is so familiar with the concept of the average in daily life, managers often use the mean inappropriately in business when other statistical information should be considered.

Measures of dispersion or variability:

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. How spread out are the values While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center. We talk about variability in the context of a distribution of values. A low dispersion indicates that the data points tend to be clustered tightly around the center. High dispersion signifies that they tend to fall further away. In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. Just as there are multiple measures of central tendency, there are several measures of variability

Range:

Let's start with the range because it is the most straightforward measure of variability to calculate and the simplest to understand. The range of a dataset is the difference between the largest and smallest values in that dataset

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

The Interquartile Range (IQR):

The interquartile range is the middle half of the data. To visualize it, think about the median value that splits the dataset in half. Similarly, you can divide the data into quarters. Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, and Q3. The lowest quartile (Q1) contains the quarter of the dataset with the smallest values. The upper

BUSINESS ANALYTICS

quartile (Q4) contains the quarter of the dataset with the highest values. The interquartile range is the middle half of the data that is in between the upper and lower quartiles. In other words, the interquartile range includes the 50% of data points that fall between Q1 and Q3. The interquartile range is a robust measure of variability in a similar manner that the median is a robust measure of central tendency. Neither measure is influenced dramatically by outliers because they don't depend on every value. Additionally, the interquartile range is excellent for skewed distributions, just like the median.

As you'll learn, when you have a normal distribution, the standard deviation tells you the percentage of observations that fall specific distances from the mean. However, this doesn't work for skewed distributions, and the IQR is a great alternative.

Variance:

Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean. To calculate this statistic, you calculate a set of squared differences between the data points and the mean, sum them, and then divide by the number of observations. Hence, it's the average squared difference. There are two formulas for the variance depending on whether you are calculating the variance for an entire population or using a sample to estimate the population variance.

Population variance: The formula for the variance of an entire population is the following:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

In the equation, σ^2 is the population parameter for the variance, μ is the parameter for the population mean, and N is the number of data points, which should include the entire population.

Sample variance: To use a sample to estimate the variance for a population, use the following formula. Using the previous equation with sample data tends to underestimate the variability. Because it's usually impossible to measure an entire population, statisticians use the equation for sample variances much more frequently.

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

BUSINESS ANALYTICS

In the equation, s^2 is the sample variance, and M is the sample mean. $N-1$ in the denominator corrects for the tendency of a sample to underestimate the population variance.

Standard Deviation:

The standard deviation is the standard or typical difference between each data point and the mean. When the values in a dataset are grouped closer together, you have a smaller standard deviation. On the other hand, when the values are spread out more, the standard deviation is larger because the standard distance is greater. Conveniently, the standard deviation uses the original units of the data, which makes interpretation easier. Consequently, the standard deviation is the most widely used measure of variability. The standard deviation is just the square root of the variance. Recall that the variance is in squared units. Hence, the square root returns the value to the natural units.

The symbol for the standard deviation as a population parameter is σ while s represents it as a sample estimate. To calculate the standard deviation, calculate the variance as shown above, and then take the square root of it. Voila! You have the standard deviation!

Coefficient of variation:

The coefficient of variation provides a relative measure of the dispersion in data relative to the mean and is defined as

$$CV = \text{Standard deviation} / \text{mean}$$

Measures of association:

When two variables appear to be related you might suspect a cause and effect relationship such relation is called as association and its types are:

1. Covariance
2. Correlation
3. outliers

Discrete Probability Distribution:

A discrete probability distribution is made up of discrete variables. Specifically, if a random variable is discrete, then it will have a discrete probability distribution. For example, let's say you had the choice of playing two games

Game 1: Roll a die. If you roll a six, you win a prize.

BUSINESS ANALYTICS

Game 2: Guess the weight of the man. If you guess within 10 pounds, you win a prize.

One of these games is a discrete probability distribution and one is a continuous probability distribution. Which is which? For game 1, you could roll a 1,2,3,4,5, or 6. All of the die rolls have an equal chance of being rolled (one out of six, or $1/6$). This gives you a discrete probability distribution. For the guess the weight game, you could guess that the man weighs 150 lbs. Or 210 pounds. Or 185.5 pounds. Or any fraction of a pound (172.566 pounds). Even if you stick to, say, between 150 and 200 pounds, the possibilities are endless: In reality, you probably wouldn't guess 160.111111 lbs...that seems a little ridiculous. But it doesn't change the fact that you could (if you wanted to), so that's why it's a continuous probability distribution. The following are examples of discrete probability distributions commonly used in statistics:

- Binomial distribution
- Geometric Distribution
- Hyper geometric distribution
- Multinomial Distribution
- Negative binomial distribution
- Poisson distribution

Continuous Probability Distribution:

Probability distributions are either continuous probability distributions or discrete probability distributions. A continuous distribution has a range of values that are infinite, and therefore uncountable. For example, time is infinite: you could count from 0 seconds to a billion seconds...a trillion seconds...and so on, forever. A discrete distribution has a range of values that are countable.

Module-3

(predictive analytics)

Karl Pearson's Coefficient of Correlation:

There are many situations in our daily life where we know from experience, the direct association between certain variables but we can't put a certain measure to it. For example, you know that the chances of you going out to watch a newly released movie is directly associated with the number of friends who go with you because the more the merrier! But there are many other factors too, like your interest in that movie, your budget etc. Thus to analyze the situation in detail, you need to note down your similar past experiences and form a sort of distribution from that data. It is at this point that you require a Correlation Coefficient, which will now provide you with a value, based on which you can calculate the possibility of you not going for the movie this time if your friends don't turn up! Karl Pearson's Coefficient of Correlation is one such type of parameter which we'll be studying in this section.

Introduction to Coefficient of Correlation:

The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r or r_{xy} (x and y being the two variables involved). This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, r , indicates how far away all these data points are to this line of best fit.

Karl Pearson Correlation Coefficient Formula:

The coefficient of correlation r_{xy} between two variables x and y , for the bivariate dataset (x_i, y_i) where $i = 1, 2, 3, \dots, N$; is given by

$$r(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

$\text{cov}(x,y)$: the covariance between x and y

σ_x and σ_y are the standard deviations of the distributions x and y .

Multiple correlation:

It is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables. The coefficient of multiple correlation takes values between 0 and 1. Higher values indicate higher predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.

The coefficient of multiple correlation is known as the square root of the coefficient of determination, but under the particular assumptions that an intercept is included and that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases, including those of nonlinear prediction and those in which the predicted values have not been derived from a model-fitting procedure. The coefficient of multiple correlation, denoted R , is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.

Spearman's Rank-Order Correlation:

This guide will tell you when you should use Spearman's rank-order correlation to analyse your data, what assumptions you have to satisfy, how to calculate it, and how to report it. If you want to know how to run a Spearman correlation in SPSS Statistics, go to our Spearman's correlation in SPSS Statistics guide.

When should you use the Spearman's rank-order correlation?

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ , also signified by r_s) measures the strength and direction of association between two ranked variables.

What are the assumptions of the test?

You need two variables that are either ordinal, interval or ratio (see our Types of Variable guide if you need clarification). Although you would normally hope to use a Pearson product-moment correlation on interval or ratio data, the Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated.

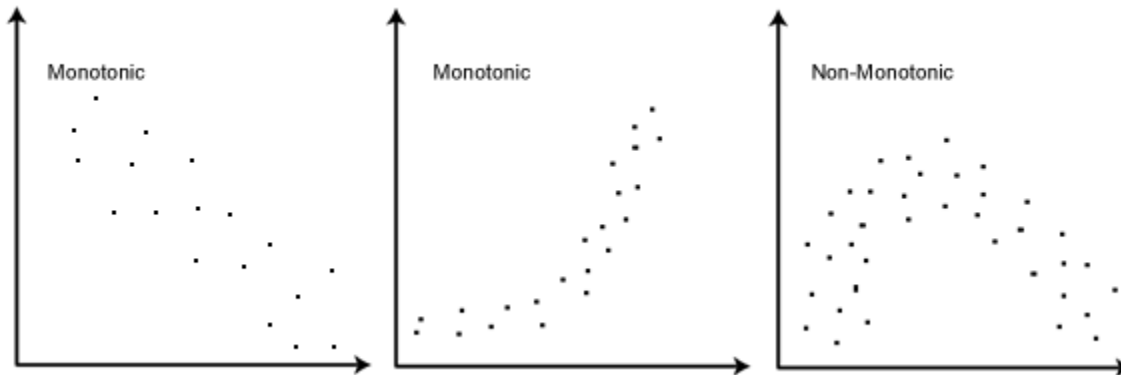
BUSINESS ANALYTICS

However, Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following:

- (1) As the value of one variable increases, so does the value of the other variable; or
- (2) As the value of one variable increases, the other variable value decreases. Examples of monotonic and non-monotonic relationships are presented in the diagram below:



Why is a monotonic relationship important to Spearman's correlation?

Spearman's correlation measures the strength and direction of monotonic association between two variables. Monotonicity is "less restrictive" than that of a linear relationship. For example, the middle image above shows a relationship that is monotonic, but not linear. A monotonic relationship is not strictly an assumption of Spearman's correlation. That is, you can run a Spearman's correlation on a non-monotonic relationship to determine if there is a monotonic component to the association. However, you would normally pick a measure of association, such as Spearman's correlation, that fits the pattern of the observed data. That is, if a scatterplot shows that the relationship between your two variables looks monotonic you would run a Spearman's correlation because this will then measure the strength and direction of this monotonic relationship. On the other hand if, for example, the relationship appears linear you would run a Pearson's correlation because this will measure the strength and direction of any linear relationship. You will not always be able to visually check whether you have a monotonic relationship, so in this case, you might run a Spearman's correlation anyway.

What is the definition of Spearman's rank-order correlation?

There are two methods to calculate Spearman's correlation depending on whether:

- (1) Your data does not have tied ranks or
- (2) Your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Simple vs. Multiple Linear Regression:

Linear regression is a model that captures the linear relationship between two (simple) or more (multiple) variables, one labeled as the dependent variable and the other(s) labeled as the independent variable(s). A linear relationship exists when increasing or decreasing the independent variable(s) results in a corresponding increase or decrease of the dependent variable. In Simple Linear Regression (SLR), our goal is to predict the value of the dependent variable y based on the independent variable x . We examine the relationship between these two variables only. For example, we would utilize SLR to predict house prices based only on the square footage of living. You will most likely be dealing with more than one independent variable when creating a regression model. In these cases, we will use Multiple Linear Regression (MLR).