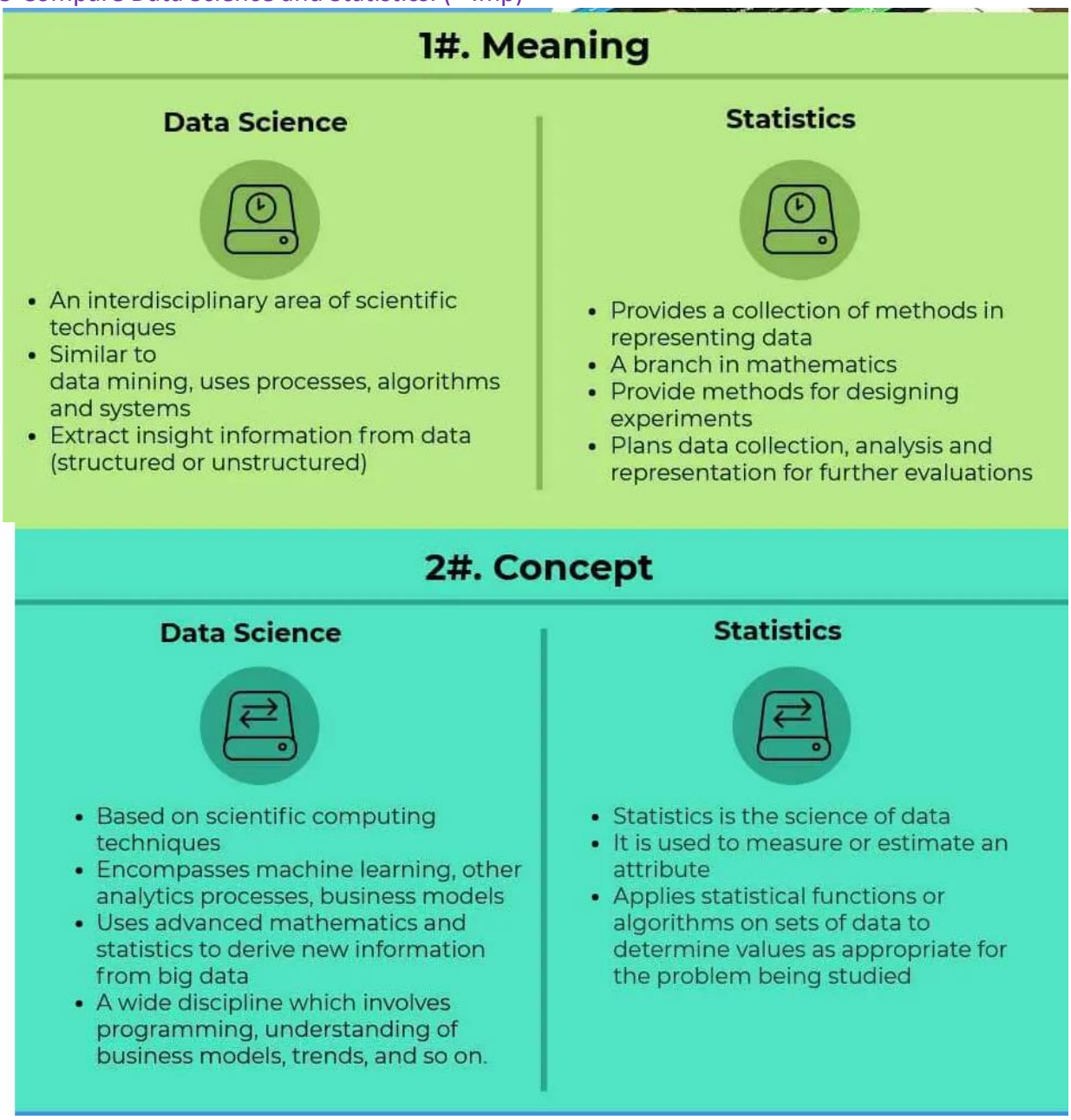
Data Science

Notes <----- click

Question

3-Compare Data Science and Statistics. (* Imp)



3#. Basis of formation

Data Science

- To solve data related problems
- Model big data for analysis towards understanding trends, patterns, behaviors and business performance
- Supports in decision making

Statistics

- To design and formulate real world questions based on data.
- Represent data in the form of tables, charts, graphs.
- Understand techniques in data analysis
- Support for decision making.

4#. Application areas

Data Science

- Healthcare systems
- Finance
- Fraud and intrusion detection
- Manufacturing, engineering
- Market analysis, etc.

Statistics

- Commerce and trade
- Industry
- Population studies, economics
- Psychology
- Biology and physical sciences
- Astronomy, etc.

5#. Approach

Data Science

- Apply scientific methods in problem solving using random data.
- Identifies data requirements for a given problem.
- Identify techniques to obtain desired results.
- Provide value to organizations using data.

Statistics

- Use of mathematical formulas, models and concepts
- Analysis of random data
- Estimate values for different data attributes
- To determine behaviors based on data

7-List out the areas in which Data Science can be applied. (* Imp)

Data science has found its applications in almost every industry.

1. Healthcare

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

2. Gaming

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

3. Image Recognition

Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.

4. Recommendation Systems

Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.

5. Logistics

Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

6. Fraud Detection

Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

7. Internet Search

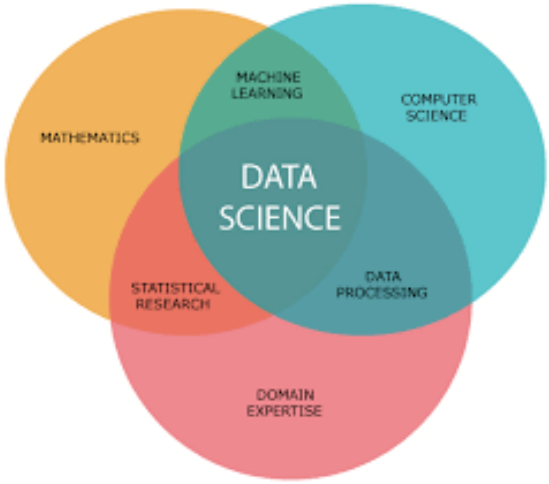
8. Speech recognition

9. Targeted Advertising

10. Airline Route Planning

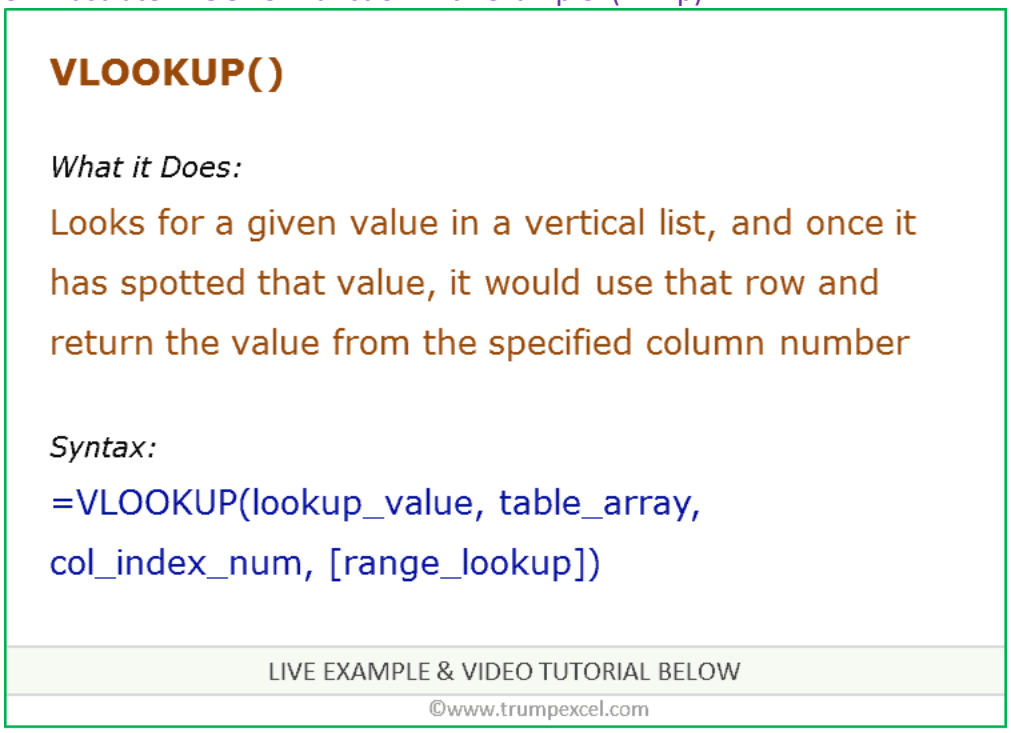
11. Augmented Reality

44-Analyze various data Science components. (* Imp)



- **Statistics**- Descriptive and Inferential- both help in organizing & generalizing large data sets and applying probability before concluding, thereby focusing on characteristics of data providing parameters.
- **Data Visualization**- It absorbs information quickly, improves insights, helps make faster decisions, improves the ability to maintain audiences' interest, eliminates the role of data scientists & eases out the distribution of information hence collected.
- **Machine Learning**- Frequently used in fraud detection & client retention and eases the process of making predictions with unforeseen/ future data easy.
- **Deep Learning**- These algorithms and multi-layered ANN require very powerful machines & are very useful in detecting patterns from input data.
- **Domain Expertise**- A high level of expertise in the area can vastly improve the accuracy of the model you want to build.
- **Data Engineering**- It leads to acquiring, storing, retrieving, and transforming the data, collecting them into a single warehouse representing the data uniformity as a single source of truth.
- **Advanced Computing**- The said capabilities are used to handle growing range of challenging science & engineering problems, most of which are data-intensive.
- **Mathematics & Programming**- Widespread usage of the most popular programming languages- Python, R, Java, and NoSQL ensures improved performance in storing huge data.

57-illustrate VLOOKUP function with example. (* Imp)



VLOOKUP function is best suited for situations when you are looking for a matching data point in a column, and when the matching data point is found, you go to the right in that row and fetch a value from a cell which is a specified number of columns to the right.

Syntax

=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

Input Arguments

- **lookup_value –** this is the look-up value you are trying to find in the left-most column of a table. It could be a value, a cell reference, or a text string. In the score sheet example, this would be your name.
- **table_array –** this is the table array in which you are looking for the value. This could be a reference to a range of cells or a named range. In the score sheet example, this would be the entire table that contains score for everyone for every subject
- **col_index –** this is the column index number from which you want to fetch the matching value. In the score sheet example, if you want the scores for Math (which is the first column in a table that contains the scores), you'd look in column 1. If you want the scores for Physics, you'd look in column 2.
- **[range_lookup] –** here you specify whether you want an exact match or an approximate match. If omitted, it defaults to TRUE – approximate match *(see additional notes below).*

Example 1 – Finding Brad's Math Score

In the VLOOKUP example below, I have a list wth student names in the left-most column and marks in different subjects in columns B to E.

	A	B	C	D	E
1		***Subject***			
2	***Name***	**Math**	**Physics**	**Chemistry**	**Biology**
3	Matt	38	58	66	49
4	Bob	88	92	74	90
5	Tom	57	77	91	91
6	Brad	82	56	45	95
7	Jenny	55	55	65	75
8	Maria	44	69	80	90
9	Jill	75	51	57	84
10	Josh	38	37	51	56

Now let's get to work and use the VLOOKUP function for what it does best. From the above data, I need to know how much Brad scored in Math.

From the above data, I need to know how much Brad scored in Math.

Here is the VLOOKUP formula that will return Brad's Math score:

=VLOOKUP("Brad",$A$3:$E$10,2,0)

The above formula has four arguments:

- **"Brad:** – this is the lookup value.
- **$A$3:$E$10** – this is the range of cells in which we are looking. Remember that Excel looks for the lookup value in the left-most column. In this example, it would look for the name Brad in A3:A10 (which is the left-most column of the specified array).
- **2** – Once the function spots Brad's name, it will go to the second column of the array, and return the value in the same row as that of Brad. The value 2 here indicated that we are looking for the score from the second column of the specified array.
- **0** – this tells the VLOOKUP function to only look for exact matches.

61-Discuss about statistics and different types of statisitics. (* Imp)

Meaning of Statistics

Basically, the statistical analysis is meant to collect and study the information available in large quantities. Statistics is a branch of mathematics, where computation is done over a bulk of data using charts, tables, graphs, etc.

The data collected for analysis here is called measurements. Now, if we have to measure the data based on a scenario, a sample is taken out of a population. Then the analysis or calculation is done for the following measurement. Learn mathematical statistics in detail at BYJU'S.

Types of Statistics in Maths

Statistics have majorly categorised into two types:

- Descriptive statistics
- Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Statistics Example

In a class, the collection of marks obtained by 50 students is the description of data. Now when we take out the mean of the data, the result is the average of marks of 50 students. If the average mark obtained by 50 students is 88 out of 100, then we can reach to a conclusion or give a judgment on the basis of the result.

52. Explain the concept of correlation and illustrate the different steps required for calculate it in excel.

Correlation measures the relationship between two variables. A correlation coefficient of 0 means that variables have no impact on one another — increases or decreases in one variable have no consistent effect on the other.

A correlation coefficient of +1 indicates a “perfect positive correlation”, which means that as variable X increases, variable Y increases at the same rate. A correlation value of -1, meanwhile, is a “perfect negative correlation”, which means that as variable X increases, variable Y decreases at the same rate. Correlation analysis may also return results anywhere between -1 and +1, which indicates that variables change at similar but not identical rates.

How to Calculate Correlation Coefficient in Excel

- Open Excel.
- Install the Analysis Toolpak.
- Select “Data” from the top bar menu.
- Select “Data Analysis” in the top right-hand corner.
- Select Correlation.
- Define your data range and output.
- Evaluate your correlation coefficient.

53. Explain the concept of Histogram and illustrate the different steps required for calculate it in excel.

It's a column chart that shows the frequency of the occurrence of a variable in the specified range.

A simple example of a histogram is the distribution of marks scored in a subject. You can easily create a histogram and see how many students scored less than 35, how many were between 35-50, how many between 50-60 and so on.

Here are the steps to create a Histogram chart in Excel 2016:

- Select the entire dataset.
- Click the Insert tab.
- In the Charts group, click on the 'Insert Static Chart' option.
- In the HIstogram group, click on the Histogram chart icon.

The above steps would insert a histogram chart based on your data set

Now you can customize this chart by right-clicking on the vertical axis and selecting Format Axis.

Creating a Histogram Using Data Analysis Tool pack

To install the Data Analysis Toolpak add-in:

- Click the File tab and then select 'Options'.
- In the Excel Options dialog box, select Add-ins in the navigation pane.
- In the Manage drop-down, select Excel Add-ins and click Go.
- In the Add-ins dialog box, select Analysis Toolpak and click OK.

This would install the Analysis Toolpak and you can access it in the Data tab in the Analysis group.

Once you have the Analysis Toolpak enabled, you can use it to create a histogram in Excel.

To create a histogram using this data, we need to create the data intervals in which we want to find the data frequency. These are called bins.

With the above dataset, the bins would be the marks intervals.

You need to specify these bins separately in an additional column

Now that we have all the data in place, let's see how to create a histogram using this data:

- Click the Data tab.
- In the Analysis group, click on Data Analysis.
- In the 'Data Analysis' dialog box, select Histogram from the list.
- Click OK.
- In the Histogram dialog box:
    - Select the Input Range (all the marks in our example)
    - Select the Bin Range (cells D2:D7)
    - Leave the Labels checkbox unchecked (you need to check it if you included labels in the data selection).
    - Specify the Output Range if you want to get the Histogram in the same worksheet. Else, choose New Worksheet/Workbook option to get it in a separate worksheet/workbook.
    - Select Chart Output.
- Click OK.

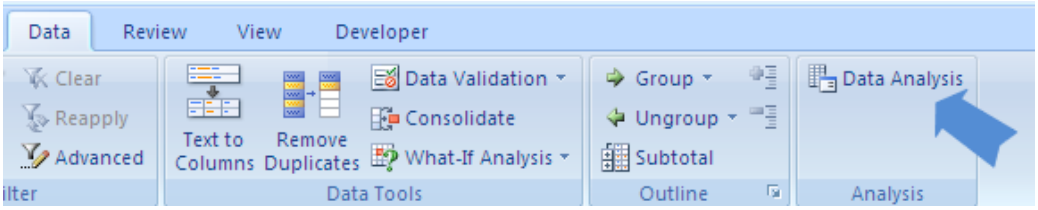
This would insert the frequency distribution table and the chart in the specified location.

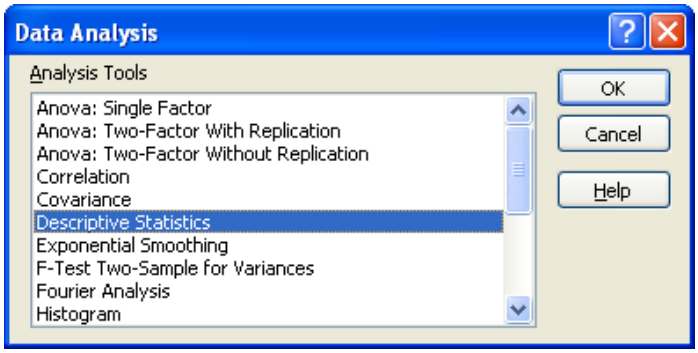
54. Explain the concept of descriptive statistics and illustrate the different steps required for calculate it in excel

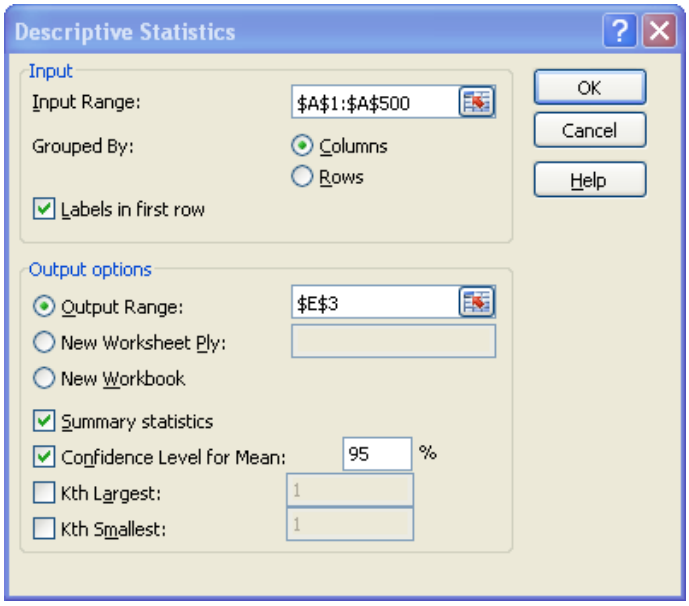
It provides information on summary statistics that includes *Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count*

1. If you haven't already installed the **Analysis ToolPak** , Click the **Microsoft Office** button, then click on the **Excel Options** , and then select **Add-Ins** , Click **Go**, check the **Analysis ToolPak** box, and click **Ok**.

2. Select **Data** tab, then click on the **Data Analysis** option, then selects **Descriptive Statistics** from the list and Click **Ok**. ***[Data tab >> Data Analysis >> Descriptive Statistics]***







3. In the **Input Range** we select the data, and then select **Output Range** where you want the output to be stored. *If you don't specify the output range it will throw output in the new worksheet.*

4. Check **Summary Statistics** and **Confidence Level for Mean** options. By default the confidence level is 95%. You can change the level as per the hypothesis standard of study.

5. When you click **Ok**, you will see the result in the selected output range.

Column1	
Mean	5.533066
Standard Error	0.131332
Median	6
Mode	8
Standard Deviation	2.933741
Sample Variance	8.606836
Kurtosis	-1.27785
Skewness	-0.03386
Range	9
Minimum	1
Maximum	10
Sum	2761
Count	499
Confidence Level(95.0%)	0.258034

Interpretation:
The average value is 5.533. The middle value is 6 and the most frequent value is 8. Negative skewness indicates a left skewed data. Negative kurtosis indicates a flat distribution. The 95% confidence level indicates you can be 95% sure that the true percentage of the population lies between 5.275 (5.533 – 0.258) and 5.791 (5.533 + 0.258).

55. Explain the concept of Moving avearge and illustrate the different steps required for calculate it in excel

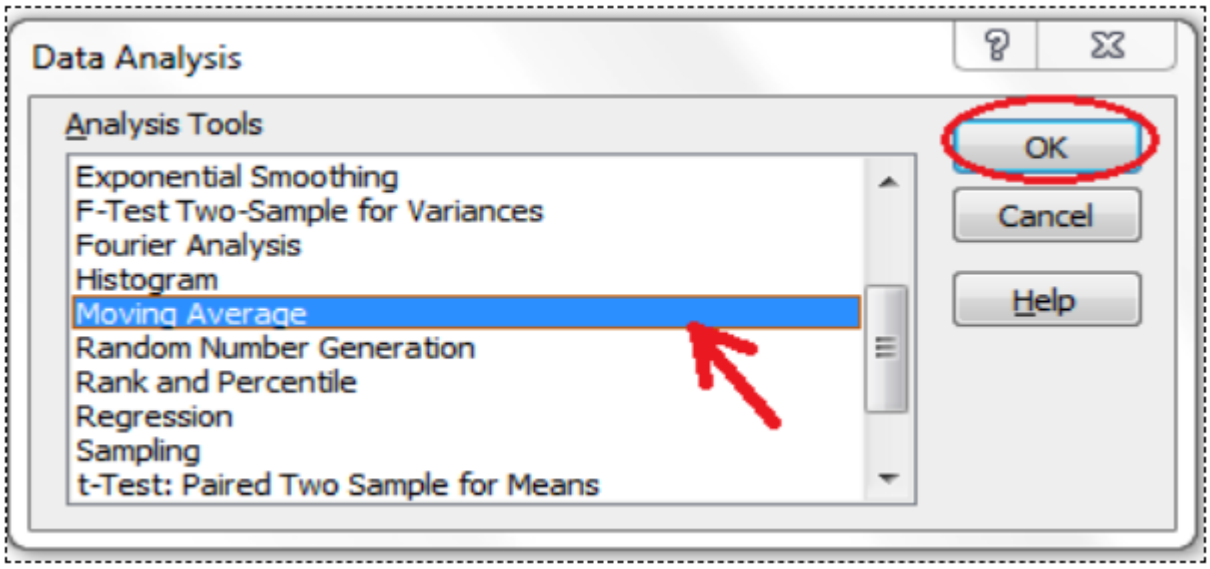
A moving average, also called a moving mean or a rolling mean, is a calculation that relies on a series of averages from data subsets within an entire data set. It's a term statisticians, technical analysts and financial analysts use to describe changes to averages as new data becomes available. It explains how a data series changes over a set period. The moving

average also updates to include recent data along with data points from pre-determined intervals.

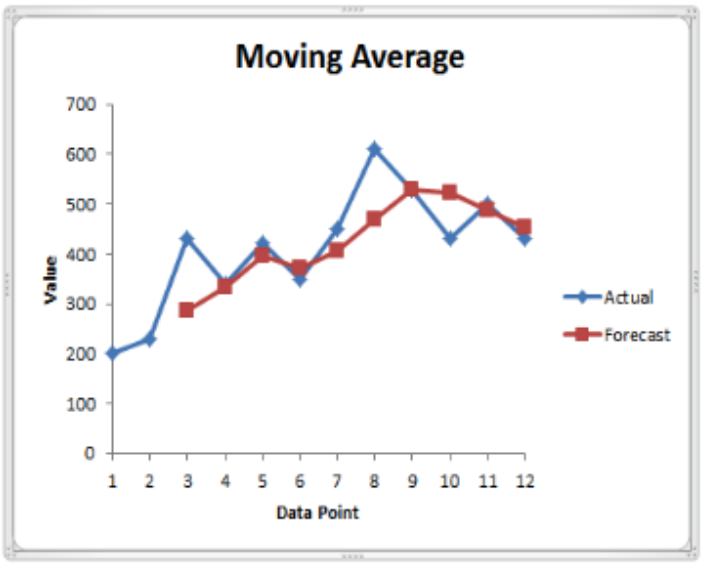
Moving average can be a helpful metric for tracking price trends because it won't reflect short-term fluctuations or infrequent outliers as drastically. In stock trading, experts often rely on 200-day moving averages, but short-, medium- and long-term averages can all be useful metrics to track.
In Excel, Analysis ToolPak add-in has a built-in option to calculate moving average for the range of data. For this purpose, you need to first install this add-in from available add-ins in Excel Options dialog box.

After installing Analysis ToolPak add-in, you need to go back to the main Excel interface, click on the Data tab and click on Data Analysis button in Analysis section.

A Data Analysis dialog box appears, click on Moving Average option from Analysis Tools and click on OK.



	A	B	C
1	Month	Sales	Interval =3
2	Jan	200	#N/A
3	Feb	230	#N/A
4	Mar	430	286.6666667
5	Apr	340	333.3333333
6	May	420	396.6666667
7	Jun	350	370
8	July	450	406.6666667
9	Aug	610	470
10	Sept	530	530
11	Oct	430	523.3333333
12	Nov	500	486.6666667
13	Dec	430	453.3333333
14			



Now, Moving Analysis dialog box appears to make calculations. You need to do followings in this dialog box;

- First, put a cursor in the Input Range section and select the range of sales data B2:B13.
- Second, go to Interval section and insert 3 as an interval period.
- Third, insert the data range to show the result of the moving average in the Output Range section as C2:C13.
- Fourth, you can select Chart Output checkbox optionally to generate chart as well showing a trend of sales.
- Finally, press OK button to calculate the moving average of sales data for the last 3 months.

You will get a series of moving averages in output cells range, and a Moving Average chart will also be created, showing actual and forecast trend based on the last 3 months sales figures.

Conclusion: The larger the interval, the more the peaks and valleys are smoothed out. The smaller the interval, the closer the moving averages are to the actual data points.

56. Explain the concept of exponential smoothing and illustrate the different steps required for calculate it in excel

Exponential Smoothing in Excel is an inbuilt smoothing method used for Forecasting, Smoothing the data, trend projection. To access, Exponential Smoothing in Excel, go to the Data menu tab and, from the Data Analysis option, choose Exponential Smoothing. Select the input range which we want to smooth and then choose the dumping factor, which should be between 0 and 1 (1 – α) and then select the output range cell. This will smoothen the select input range number by the percentage of dumping factor we choose.

Where is the Exponential Smoothing found in Excel?

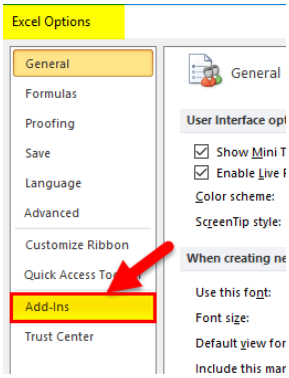
It is found under Analysis ToolPak in Excel. The Analysis ToolPak is a Microsoft Excel data analysis add-in. This add-in is not loaded automatically on excel. Before using this first, we need to load it.

Steps to load the Analysis ToolPak add-in:

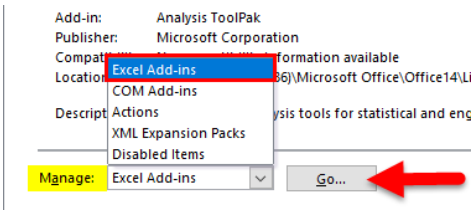
We need to add this feature in Excel for analyzing business by using Excel Add-Ins. To add this feature in Excel, follow the below steps:

- Go to the **FILE** tab. Click on the **OPTIONS** tab in the left pane window. Refer to the below screenshot.

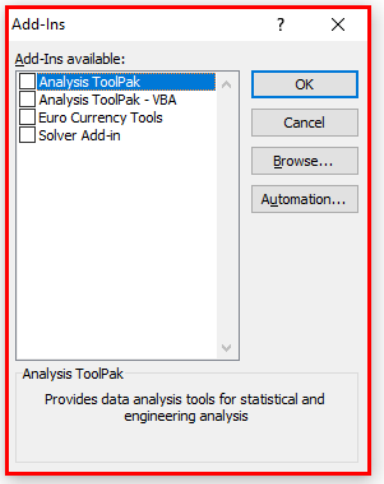
- It will open a dialog box for Excel Options. Click on the **Add-Ins** tab, as shown in the below screenshot.



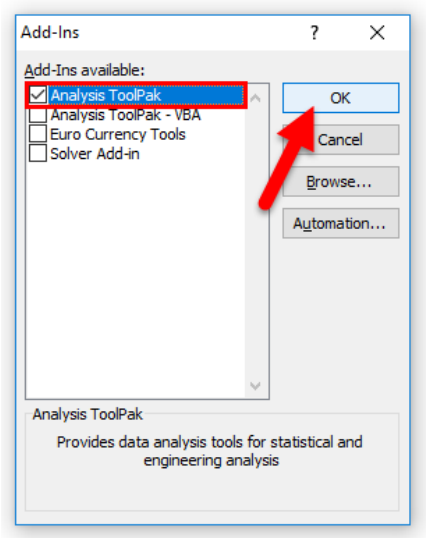
- It will again display some options.
- Select the **Excel Add-Ins** options under Manage Box and click on the **Go** button as shown in the below screenshot. (However, Excel Add-Ins is by default selected)



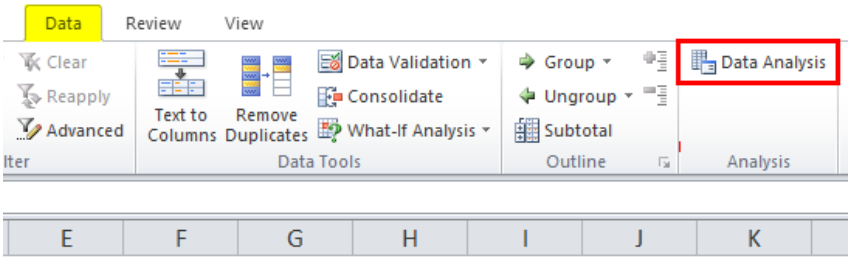
- It will open an **Add-Ins** dialog box.



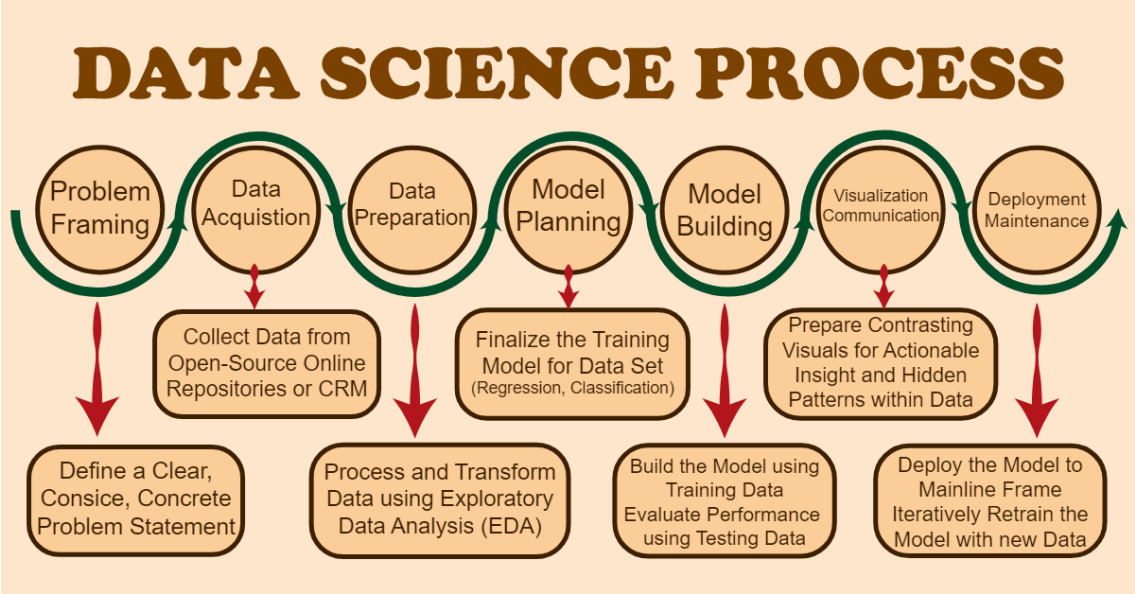
- Tick on the **Analysis Toolpak** checkbox and then click on **OK,** as shown in the below screenshot.



- The above steps will add the **Data Analysis** section for statistical analysis under the DATA tab.



1. Explain Data Science process.



2. Differentiate Business Intelligence (BI) and Data Science.

Business Intelligence *vs* Data Science

Factors	Business Intelligence	Data Science
Concept	Deals with data analysis on the business platform.	Consists of several data operations in various domains.
Scope	BI analyzes past data	Past data is analyzed for future predictions.
Data	Handling static and structured data	Both structured & unstructured data that is also dynamic.
Data Storage	Data stored mostly in data-warehouses	Data utilized is distributed in real time clusters.
Procedure	BI helps companies to solve questions.	Questions are both curated and solved by data scientists.
Tools	MS Excel, SAS BI, Sisense, Microstrategy	Python, R, Hadoop/Spark, SAS, TensorFlow.

3. Compare Data Science and Statistics. (* Imp) done
4. Define Data Science.

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

5. Define the role of data scientist

Data scientist roles and responsibilities include:

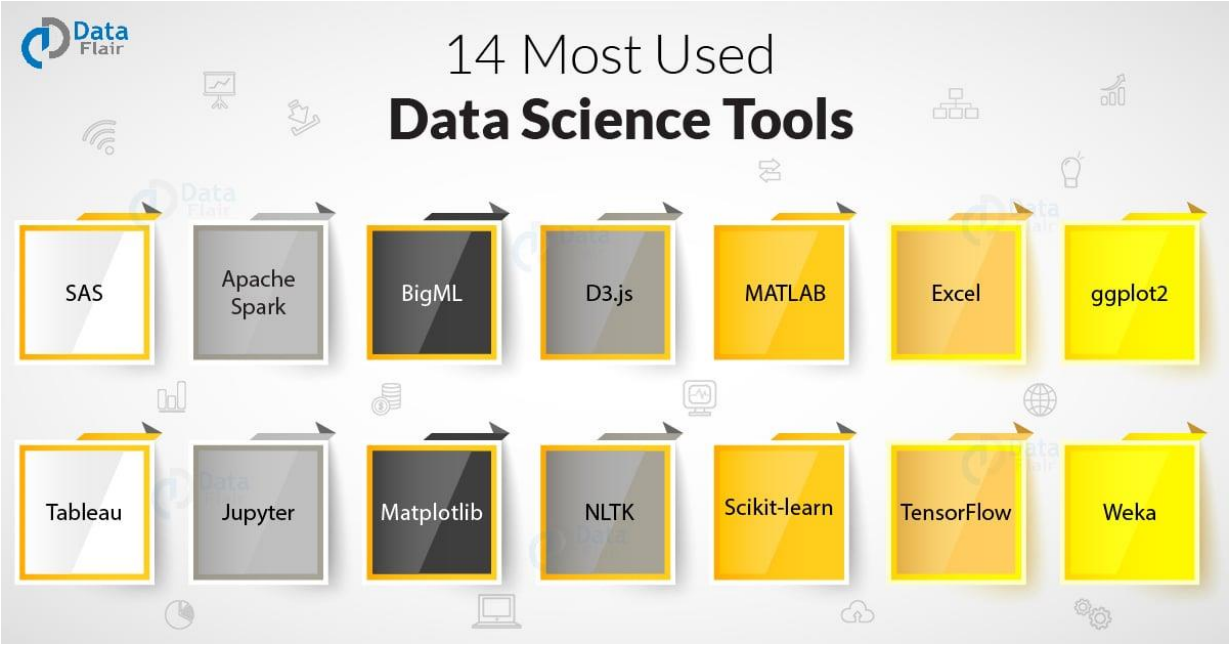
- Data mining or extracting usable data from valuable data sources
- Using machine learning tools to select features, create and optimize classifiers
- Carrying out preprocessing of structured and unstructured data
- Enhancing data collection procedures to include all relevant information for developing analytic systems
- Processing, cleansing, and validating the integrity of data to be used for analysis
- Analyzing large amounts of information to find patterns and solutions
- Developing prediction systems and machine learning algorithms
- Presenting results in a clear manner
- Propose solutions and strategies to tackle business challenges
- Collaborate with Business and IT teams

6. Analyze the different essential skills required for a data scientist

Key skills needed to become a data scientist:

- Programming Skills – knowledge of statistical programming languages like R, Python, and database query languages like SQL, Hive, Pig is desirable. Familiarity with Scala, Java, or C++ is an added advantage.
- Statistics – Good applied statistical skills, including knowledge of statistical tests, distributions, regression, maximum likelihood estimators, etc. Proficiency in statistics is essential for data-driven companies.
- Machine Learning – good knowledge of machine learning methods like k-Nearest Neighbors, Naive Bayes, SVM, Decision Forests.
- Strong Math Skills (Multivariable Calculus and Linear Algebra) - understanding the fundamentals of Multivariable Calculus and Linear Algebra is important as they form the basis of a lot of predictive performance or algorithm optimization techniques.
- Data Wrangling – proficiency in handling imperfections in data is an important aspect of a data scientist job description.
- Experience with Data Visualization Tools like matplotlib, ggplot, d3.js., Tableau that help to visually encode data

Question-different popular software tools for data science



1. SAS

It is one of those data science tools which are specifically designed for statistical operations. **SAS is a closed source proprietary software** that is used by large organizations to analyze data. SAS uses base SAS programming language which for performing statistical modeling.

2. Apache Spark

Apache Spark or simply Spark is an all-powerful analytics engine and it is the most used Data Science tool. Spark is specifically designed to handle batch processing and **Stream Processing.**

3. BigML

BigML, it is another widely used Data Science Tool. It provides a fully interactable, cloud-based GUI environment that you can use for processing **Machine Learning Algorithms**. BigML provides standardized software using cloud computing for industry requirements.

4. D3.js

Javascript is mainly used as a client-side scripting language. D3.js, a Javascript library allows you to make interactive visualizations on your web-browser.

5. MATLAB

MATLAB is a multi-paradigm numerical computing environment for processing mathematical information. It is a closed-source software that facilitates matrix functions, algorithmic implementation and statistical modeling of data.

6. Excel

Probably the most widely used Data Analysis tool. Microsoft developed Excel mostly for spreadsheet calculations and today, it is widely used for data processing, visualization, and complex calculations.

7. ggplot2

ggplot2 is an advanced data visualization package for the **R programming language**. The developers created this tool to replace the native graphics package of R and it uses powerful commands to create illustrious visualizations.

8. Tableau

Tableau is a Data Visualization software that is packed with powerful graphics to make interactive visualizations. It is focused on industries working in the field of business intelligence.

9. Jupyter

Project **Jupyter** is an open-source tool based on IPython for helping developers in making open-source software and experiences interactive computing. Jupyter supports multiple languages like Julia, **Python**, and R.

10. Matplotlib

Matplotlib is a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.

Question-what is data analysis?

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it. For that, we gather memories of our past or dreams of our future. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

Types of Data Analysis: Techniques and Methods

There are several **types of Data Analysis** techniques that exist based on business and technology. However, the major Data Analysis methods are:

- Text Analysis
- Statistical Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

Text Analysis

Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover a pattern in large data sets using databases or data mining tools. It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.

Statistical Analysis

Statistical Analysis shows "What happen?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data. It analyses a set of data or a sample of data. There are two categories of this type of Analysis – Descriptive Analysis and Inferential Analysis.

Descriptive Analysis

analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.

Inferential Analysis

analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

Diagnostic Analysis

Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar

patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

Predictive Analysis

Predictive Analysis shows “what is likely to happen” by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course it’s not easy like this because you have to think about other circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house!

So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

Prescriptive Analysis

Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

Ques-Statistics fundamentals for data science?

1. Descriptive Statistics

It is used to describe the basic features of data that provide a summary of the given data set which can either represent the entire population or a sample of the population. It is derived from calculations that include:

Mean: It is the central value which is commonly known as arithmetic average.

Mode: It refers to the value that appears most often in a data set.

Median: It is the middle value of the ordered set that divides it in exactly half.

2. Variability

Variability includes the following parameters:

Standard Deviation: It is a statistic that calculates the dispersion of a data set as compared to its mean.

Variance: It refers to a statistical measure of the spread between the numbers in a data set. In general terms, it means the difference from the mean. A large variance indicates that numbers are far apart from the mean or average value. Small variance indicates that the numbers are closer to the average values. Zero variance indicates that the values are identical to the given set.

Range: This is defined as the difference between the largest and smallest value of a dataset.

Percentile: It refers to the measure used in statistics that indicates the value below which the given percentage of observation in the dataset falls.

Quartile: It is defined as the value that divides the data points into quarters.

Interquartile Range: It measures the middle half of your data. In general terms, it is the middle 50% of the dataset.

3. Correlation

It is one of the major statistical techniques that measure the relationship between two variables. The correlation coefficient indicates the strength of the linear relationship between two variables.

A correlation coefficient that is more than zero indicates a positive relationship.

A correlation coefficient that is less than zero indicates a negative relationship.

Correlation coefficient zero indicates that there is no relationship between the two variables.

4. Probability Distribution

It specifies the likelihood of all possible events. In simple terms, an event refers to the result of an experiment like tossing a coin. Events are of two types dependent and independent.

Independent event: The event is said to be an Independent event when it is not affected by the earlier events. For example, tossing a coin, let us consider a coin is tossed the first outcome is head when the coin is tossed again the outcome may be head or tail. But this is entirely independent of the first trial.

Dependent event: The event is said to be dependent when the occurrence of the event is dependent on the earlier events. For example when a ball is drawn from a bag that contains red and blue balls. If the first ball drawn is red, then the second ball may be red or blue; this depends on the first trial.

The probability of independent events is calculated by simply multiplying the probability of each event and for a dependent event is calculated by conditional probability.

5. Regression

It is a method that is used to determine the relationship between one or more independent variables and a dependent variable. Regression is mainly of two types:

Linear regression: It is used to fit the regression model that explains the relationship between a numeric predictor variable and one or more predictor variables.

Logistic regression: It is used to fit a regression model that explains the relationship between the binary response variable and one or more predictor variables.

6. Normal Distribution

Normal is used to define the probability density function for a continuous random variable in a system. The standard normal distribution has two parameters – mean and standard deviation that are discussed above. When the distribution of random variables is unknown, the normal distribution is used. The central limit theorem justifies why normal distribution is used in such cases.

7. Bias

In statistical terms, it means when a model is representative of a complete population. This needs to be minimized to get the desired outcome.

The three most common types of bias are:

Selection bias: It is a phenomenon of selecting a group of data for statistical analysis, the selection in such a way that data is not randomized resulting in the data being unrepresentative of the whole population.

Confirmation bias: It occurs when the person performing the statistical analysis has some predefined assumption.

Time interval bias: It is caused intentionally by specifying a certain time range to favor a particular outcome.

Question-Use of formulae to calculate the values in excel?

There are various functions present in the Excel 2019 version to calculate cell values:

- SUM
- PRODUCT
- AVERAGE
- COUNT & COUNTA
- IF
- MAX & MIN
- TRIM
- DEC2BIN

SUM Function

The SUM function is used to add values from multiple cells. Here's the syntax:

```
1=SUM(n1, n2, n3, ...)
```

A	B	C	D	E	F	G
Sr. Number	Value1	Value2	Value3	Value4	Formula	Result
1	103	53	21	3423	=SUM(B1:E1)	3600
2	122	25	51	321	=SUM(B2, C2, D2, E2)	519
3	88	50	15	12	=SUM(88, 50, 15, 12)	165
4	62	57	17	343	=SUM(B4, D4)	79
5	15	51	11	87	=SUM(B2:E2, B5:E5)	683

PRODUCT Function

As the name suggests, all the numbers passed inside the PRODUCT function gets multiplied. Here's the syntax:

```
1=PRODUCT(n1, n2, n3, ...)
```

AVERAGE Function

If you want to calculate the arithmetic mean of given numbers, use the AVERAGE function. Here's the syntax:

```
1=AVERAGE(n1, n2, n3, ...)COUNT & COUNTA Function
```

The COUNT function is used "to get the number of entries in a number field that is in a range or array of numbers," whereas the COUNTA function "counts the number of cells that are not empty in a range
The COUNT and COUNTA functions have the following syntax:

```
1=COUNT(n1, n2, n3, ...)
2=COUNTA(n1, n2, n3, ...)
```

IF Function

When you have a logical condition that can either be true or false, use the IF function.
The IF function has the following syntax:

```
1=IF(condition, response_on_true, response_on_false)
```

MAX & MIN Function

You can rely on the MAX and the MIN functions to return the maximum and minimum value from a range of values respectively. Here's the syntax for both the functions:

```
1=MAX(n1, n2, n3, ...)
2=MIN(n1, n2, n3, ...)
```

TRIM Function

The TRIM function "removes all spaces from text except for single spaces between words," according to Excel's documentation.
The TRIM function has the following syntax:

```
1=TRIM(text_with_unwanted_spaces)
```

DEC2BIN Function

The DEC2BIN function is used "to convert a decimal number to binary," according to Excel's documentation. The DEC2BIN function has the following syntax:

```
1=DEC2BIN(num, char_required)
```

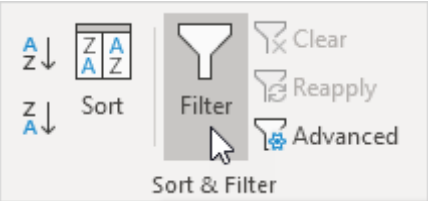
7. List out the areas in which Data Science can be applied. (* Imp)

Question-fliter in excel?

Filter your Excel data if you only want to display records that meet certain criteria.

1. Click any single cell inside a data set.

2. On the Data tab, in the Sort & Filter group, click Filter.

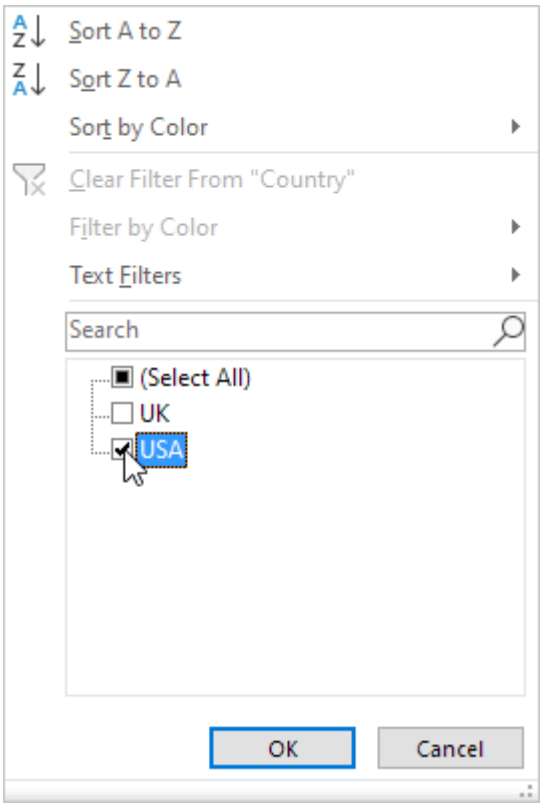


Arrows in the column headers appear.

	A	B	C	D	E
1	Last Nan	Sales	Count	Quart	
2	Smith	$16,753.00	UK	Qtr 3	
3	Johnson	$14,808.00	USA	Qtr 4	
4	Williams	$10,644.00	UK	Qtr 2	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
7	Williams	$12,438.00	UK	Qtr 1	
8	Johnson	$9,339.00	UK	Qtr 2	
9	Smith	$18,919.00	USA	Qtr 3	
10	Jones	$9,213.00	USA	Qtr 4	
11	Jones	$7,433.00	UK	Qtr 1	
12	Brown	$3,255.00	USA	Qtr 2	
13	Williams	$14,867.00	USA	Qtr 3	
14	Williams	$19,302.00	UK	Qtr 4	
15	Smith	$9,698.00	USA	Qtr 1	
16					

3. Click the arrow next to Country.

4. Click on Select All to clear all the check boxes, and click the check box next to USA.



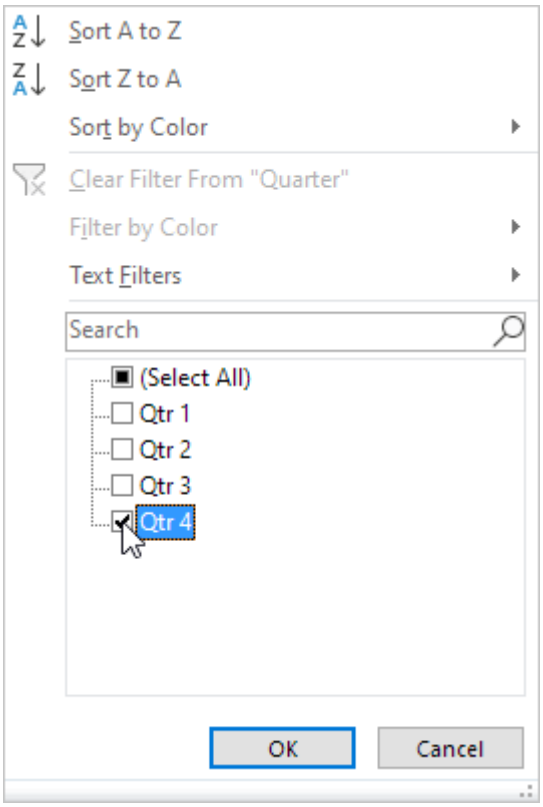
5. Click OK.

Result. Excel only displays the sales in the USA.

	A	B	C	D	E
1	Last Nan	Sales	Count	Quart	
3	Johnson	$14,808.00	USA	Qtr 4	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
9	Smith	$18,919.00	USA	Qtr 3	
10	Jones	$9,213.00	USA	Qtr 4	
12	Brown	$3,255.00	USA	Qtr 2	
13	Williams	$14,867.00	USA	Qtr 3	
15	Smith	$9,698.00	USA	Qtr 1	
16					

6. Click the arrow next to Quarter.

7. Click on Select All to clear all the check boxes, and click the check box next to Qtr 4.

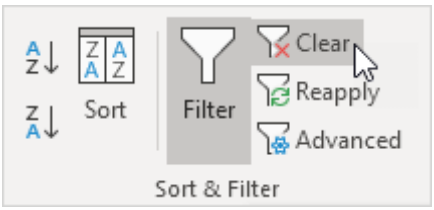


8. Click OK.

Result. Excel only displays the sales in the USA in Qtr 4.

	A	B	C	D	E
1	Last Nan	Sales	Count	Quart	
3	Johnson	$14,808.00	USA	Qtr 4	
6	Brown	$4,865.00	USA	Qtr 4	
10	Jones	$9,213.00	USA	Qtr 4	
16					

9. To remove the filter, on the Data tab, in the Sort & Filter group, click Clear. To remove the filter and the arrows, click Filter.

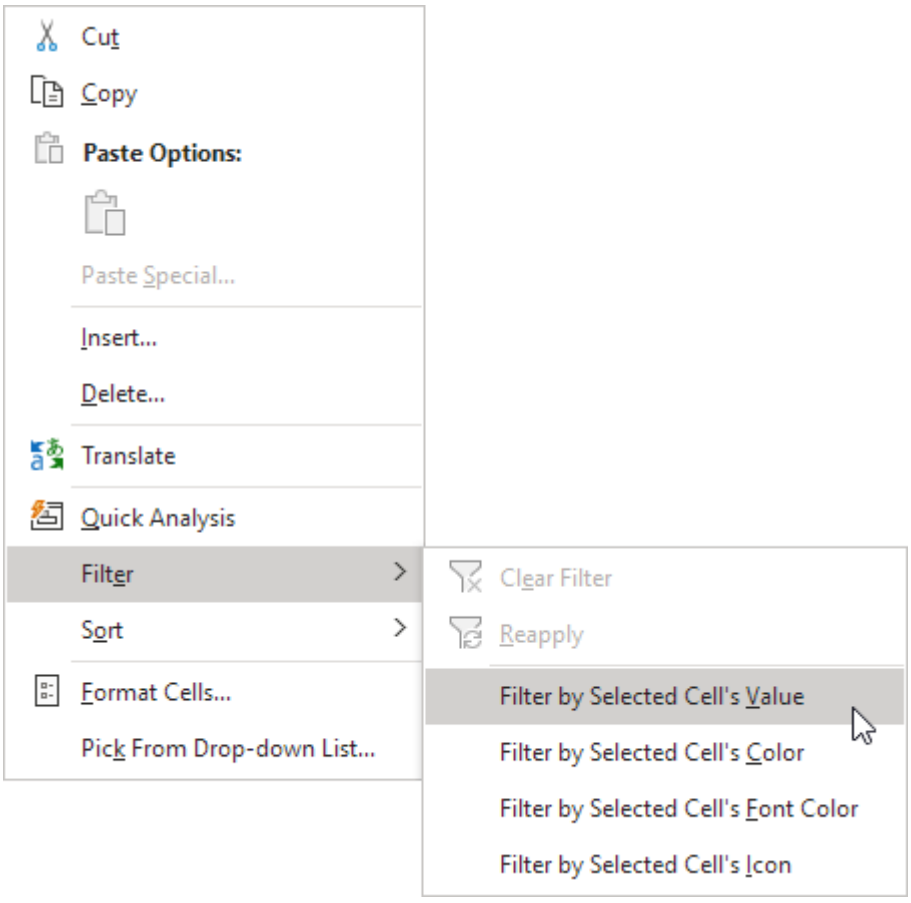


There's a quicker way to filter Excel data.

10. Select a cell.

	A	B	C	D	E
1	**Last Name**	**Sales**	**Country**	**Quarter**	
2	Smith	$16,753.00	UK	Qtr 3	
3	Johnson	$14,808.00	USA	Qtr 4	
4	Williams	$10,644.00	UK	Qtr 2	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
7	Williams	$12,438.00	UK	Qtr 1	
8	Johnson	$9,339.00	UK	Qtr 2	
9	Smith	$18,919.00	USA	Qtr 3	
10	Jones	$9,213.00	USA	Qtr 4	
11	Jones	$7,433.00	UK	Qtr 1	
12	Brown	$3,255.00	USA	Qtr 2	
13	Williams	$14,867.00	USA	Qtr 3	
14	Williams	$19,302.00	UK	Qtr 4	
15	Smith	$9,698.00	USA	Qtr 1	
16					

11. Right click, and then click Filter, Filter by Selected Cell's Value.



Result. Excel only displays the sales in the USA.

	A	B	C	D	E
1	Last Nan	Sales	Count	Quart	
3	Johnson	$14,808.00	USA	Qtr 4	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
9	Smith	$18,919.00	USA	Qtr 3	
10	Jones	$9,213.00	USA	Qtr 4	
12	Brown	$3,255.00	USA	Qtr 2	
13	Williams	$14,867.00	USA	Qtr 3	
15	Smith	$9,698.00	USA	Qtr 1	
16					

8. Give the significant of normal distribution.
9. Compare Big Data with Data Science.
10. Analyze Data Science ethics.
11. Discuss about the structure data
12. Discuss about the unstructure data
13. Discuss about the semi-structure data
14. Differentiate Data Mining and Data Science.
15. Can Data Science Predict the Stock Market? Examine.
16. Discuss how data science can be applied for fraud detection
17. Show the ways in which decision making and predictions are made in Data Science.
18. Discuss on categorical, ordinal, interval and ratio data with example
19. Specify the life cycle of Data Science.
20. Illustrate the use of Data Science with an example.
21. Develop a general algorithm for Data Science process.
22. Given single set of data, explain central tendencies of the data.
23. Demonstrate the concept of variance and standard deviation with an example
24. Brief any four statistical measure with example.
25. Analyze the roles of Data Science.
26. Classify the different distribution of values of random variables.
27. Relate probability with respect to Data Science with your own illustration.
28. Compare variance and covariance.
29. Explain about various data wrangling techniques.
30. Describe the features of a big data in detail.
31. Explain the 5 K's of big data
32. Describe life cycle of Data Science with neat diagram.

33. List any four realtime applications of Big Data.
34. Discuss various types of data.
35. Give detail description of applications of data science.
36. Give the difference between Traditional Business Intelligence (BI) versus Big Data.
37. Give the various drawbacks of using Traditional system approach.
38. Demonstrate the ETL (Extract, Transform and Load) system?
39. Analyze and write short notes on the following.
    i. Hadoop Distributed File System (HDFS). (3)
    ii. YARN(2)
40. Analyze different roles of business analyst
41. Discuss the importance of big data analytics?
42. Extrapolate summary of various applications of Data science in the real world scenario.
43. Describe the roles and stages in data science project.
44. Analyze various data Science components. (* Imp)
45. Illustrate Barchart, piechart with neat diagram and variants of COUNT operation in excel.
46. Explain the data science classification and illustrate data science tasks.
47. Analyze different challenges of data science technology
48. List out the various challenges faced in big data in detail.
49. Explain storage consideration in Big Data.
50. Discuss Data Cleaning and Sampling.
51. Define data science and exemplify the need of data science?
52. Explain the concept of correlation and illustrate the different steps required for calculate it in excel.
53. Explain the concept of Histogram and illustrate the different steps required for calculate it in excel.
54. Explain the concept of descriptive statistics and illustrate the different steps required for calculate it in excel
55. Explain the concept of Moving avearge and illustrate the different steps required for calculate it in excel
56. Explain the concept of exponential smoothing and illustrate the different steps required for calculate it in excel
57. Illustrate VLOOKUP function with example. (* Imp)
58. Describe the challenges with real time data
59. Explain the different types of data.
60. Describe variables and its types. Also describe on Population and Sample.
61. Discuss about statistics and different types of statisitics. (* Imp)
62. Discuss on measuring variables using different scales of measurement.
63. Write a python program to get the data type
64. Write a python program to get the shape of the array
65. Write a python program to find min, max, average
66. Write a python program find variance and standard deviation

67. Write a python program to find corelation coefficient
68. Write a python program to convert a Series to DataFrame.
69. Define Numpy broadcasting
70. Describe Time Series, Time Offset and Time periods in Pandas with suitable program.
71. Write a python program to convert String to date.
72. Explain the filtering operation in Excel
73. Explain in detail about conditional formatting in Excel.
74. Describe the following excel analysis: Concatenation, Len and Trim with example.
75. Discuss about different variants of count operation in excel
76. Describe the following excel analysis: CountA, Averageifs and Find/Search with example.
77. Describe the following excel analysis: Sumifs, Countifs and VLookup with example. (* Imp)
78. Describe the following excel analysis: Left/Right, If() and HLookup with example.
79. Write a python program that uses numpy and explain it.
80. Illustrate an array creation using Numpy.
81. List any three different types of charts in excel
82. Discuss various Toolkits in Python in detail.