Short Answer type question (5 marks)

1. What is data science? Explain the application areas of data science with example.

ANS:

Data science is the field of applying advanced analytics techniques and scientific principles to extract valuable information from data for business decision-making, strategic planning and other uses. It's increasingly critical to businesses: The insights that data science generates help organizations increase operational efficiency, identify new business opportunities and improve marketing and sales programs, among other benefits. Ultimately, they can lead to competitive advantages over business rivals.

Data science incorporates various disciplines -- for example, data engineering, data preparation, data mining, predictive analytics, machine learning and data visualization, as well as statistics, mathematics and software programming. It's primarily done by skilled data scientists, although lower-level data analysts may also be involved. In addition, many organizations now rely partly on citizen data scientists, a group that can include business intelligence (BI) professionals, business analysts, data-savvy business users, data engineers and other workers who don't have a formal data science background.

2. Briefly discuss structured data, unstructured data and semi structured data with example.

ANS:

Structured data – Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. *Example:* Relational data.

Semi-Structured data – Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. *Example*: XML data.

Unstructured data – Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

3. Explain the process of ETL in brief.

ANS:

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

As the databases grew in popularity in the 1970s, ETL was introduced as a process for integrating and loading data for computation and analysis, eventually becoming the primary method to process data for data warehousing projects.

ETL provides the foundation for data analytics and machine learning workstreams. Through a series of business rules, ETL cleanses and organizes data in a way which addresses specific business intelligence needs, like monthly reporting, but it can also tackle more advanced analytics, which can improve back-end processes or end user experiences. ETL is often used by an organization to:

- Extract data from legacy systems
- Cleanse the data to improve data quality and establish consistency
- Load data into a target database

4. Discuss feature extraction and feature reduction in the light of feature engineering.

ANS:

Feature extraction is usually used when the original data was very different. In particular when you could not have used the raw data.

E.g. original data were images. You extract the redness value, or a description of the shape of an object in the image. It's lossy, but at least you get some result now.

Feature engineering is the careful preprocessing into more meaningful features, even if you could have used the old data. Feature Reduction is a pert of Feature engineering.

E.g. instead of using variables x, y, z you decide to use log(x)-sqrt(y)*z instead, because your engineering knowledge tells you that this derived quantity is more meaningful to solve your problem. You get better results than without.

5. Explain data pre- processing and data cleaning with example.

ANS:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Regression:

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

6. Draw and explain Box Plot and Scatter Plot.

ANS:

A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data. Box limits indicate the range of the central 50% of the data, with a central line marking the median value. Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers.

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

7. Explain Histogram and Multi-vari chart with example.

ANS:

Multi-vari charts are used to investigate the stability or consistency of a process. The chart consists of a series of vertical lines, or other appropriate schematics, along a time scale. The length of each line or schematic shape represents the range of values found in each sample set.

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

8. Differentiate between data science and data analytics. What is bias in Data Science?

ANS:

Data Analytics	**Data Science**
Data Analytics is a subset of Data Science.	Data Science is a broad technology that includes various subsets such as Data Analytics, Data Mining, Data Visualization, etc.
The goal of data analytics is to illustrate the precise details of retrieved insights.	The goal of data science is to discover meaningful insights from massive datasets and derive the best possible solutions to resolve business issues.
Requires just basic programming languages.	Requires knowledge in advanced programming languages.
It focuses on just finding the solutions.	Data Science not only focuses on finding the solutions but also predicts the future with past patterns or insights.
A data analyst's job is to analyse data in order to make decisions.	A data scientist's job is to provide insightful data visualizations from raw data that are easily understandable.

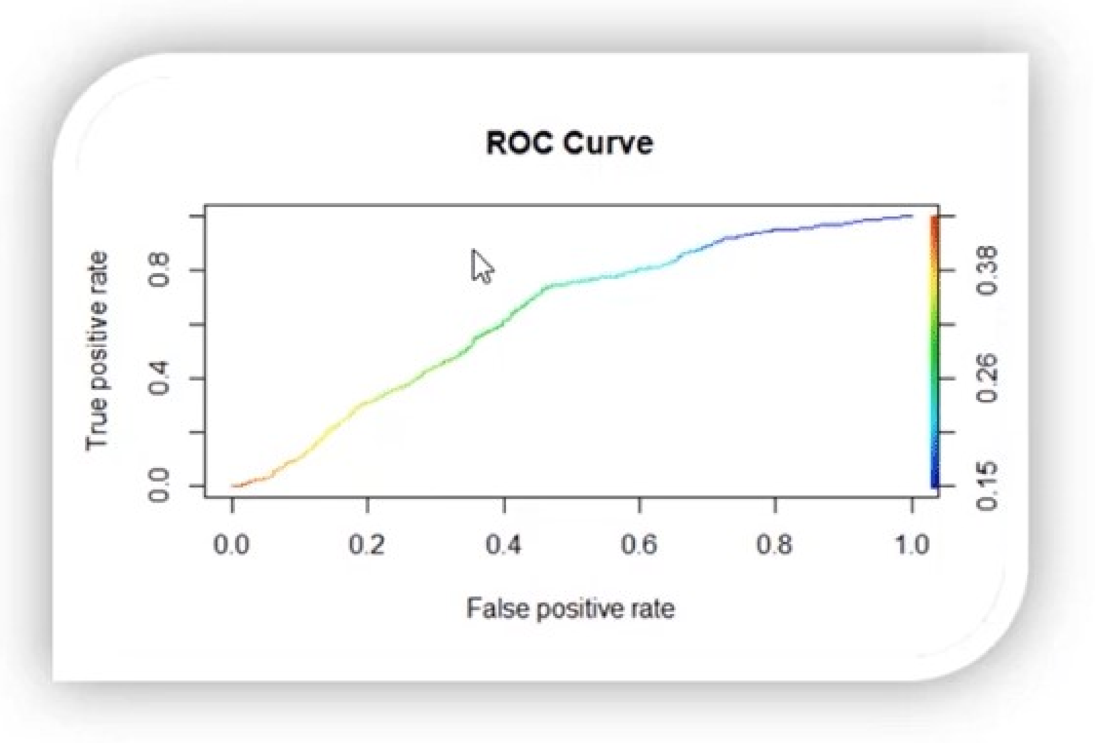
Bias is a type of error that occurs in a Data Science model because of using an algorithm that is not strong enough to capture the underlying patterns or trends that exist in the data. In other words, this error occurs when the data is too complicated for the algorithm to understand, so it ends up building a model that makes simple assumptions. This leads to lower accuracy because of underfitting. Algorithms that can lead to high bias are linear regression, logistic regression, etc

9. Explain ROC curve with diagram.

ANS:

It stands for Receiver Operating Characteristic. It is basically a plot between a true positive rate and a false positive rate, and it helps us to find out the right tradeoff between the true positive rate and the false positive rate for different probability thresholds of the predicted values. So, the closer the curve to the upper left corner, the better the model is. In other words, whichever curve has greater area under it that would be the better model. You can see this in the below

graph:



10. What does the word 'Naïve' mean in Naïve Bayes algorithm?

Two candidates Suman and Rana appear for a data science job interview. The probability of Suman cracking the interview is 1/8 and that of Rana is 5/12. What is the probability that at least one of them will crack the interview?

ANS:

Naive Bayes is a Data Science algorithm. It has the word 'Bayes' in it because it is based on the Bayes theorem, which deals with the probability of an event occurring given that another event has already occurred.

It has 'naive' in it because it makes the assumption that each variable in the dataset is independent of the other. This kind of assumption is unrealistic for real-world data. However, even with this assumption, it is very useful for solving a range of complicated problems, e.g., spam email classification, etc.

The probability of Aman getting selected for the interview is 1/8

P(A) = 1/8

The probability of Mohan getting selected for the interview is 5/12

P(B)=5/12

Now, the probability of at least one of them getting selected can be denoted at the Union of A and B, which means

P(A U B) =P(A)+ P(B) – (P(A ∩ B)) ………………………(1)

Where P(A ∩ B) stands for the probability of both Aman and Mohan getting selected for the job.

To calculate the final answer, we first have to find out the value of P(A ∩ B)

So, P(A ∩ B) = P(A) * P(B)

1/8 * 5/12

5/96

Now, put the value of P(A ∩ B) into equation (1)

P(A U B) =P(A)+ P(B) – (P(A ∩ B))

1/8 + 5/12 -5/96

So, the answer will be 47/96.

11. What is stacking in data science? Find out the values of mean, median and range from the given numbers: 11,7,11,18,9,7,6,23,7.

ANS:

Just like bagging and boosting, stacking is also an ensemble learning method. In bagging and boosting, we could only combine weak models that used the same learning algorithms, e.g., logistic regression. These models are called homogeneous learners.

However, in stacking, we can combine weak models that use different learning algorithms as well. These learners are called heterogeneous learners. Stacking works by training multiple (and different) weak models or learners and then using them together by training another model, called a meta-model, to make predictions based on the multiple outputs of predictions returned by these multiple weak models.

12. Explain categorical and Numerical data with example. What is outlier? 3+2

ANS:

Categorical data is also called qualitative data while numerical data is also called quantitative data. This is because categorical data is used to qualify information before classifying them according to their similarities.

Numerical data is used to express quantitative values and can also perform arithmetic operations which is a quantitative characteristic. Both numerical and categorical data have other names that depict their meaning. But the names are however different from each other.

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

13. What is Machine Learning? Explain the differences between unsupervised and supervised learning. 2+3

ANS:

Machine Learning is the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

Supervised Learning	Unsupervised Learning
Works on the data that contains both inputs and the expected output, i.e., the labeled data	Works on the data that contains no mappings from input to output, i.e., the unlabeled data
Used to create models that can be employed to predict or classify things	Used to extract meaningful information out of large volumes of data
Commonly used supervised learning algorithms: Linear regression, decision tree, etc.	Commonly used unsupervised learning algorithms: K-means clustering, Apriori algorithm, etc.

14. What are population and sample in statistics? What is the importance of central tendency of data? 3+2

ANS:

Two basic but vital concepts in statistics are those of population and sample. We can define them as follows.

- Population is the entire group that you wish to draw data from (and subsequently draw conclusions about). While in day-to-day life, the word is often used to describe groups of people (such as the population of a country) in statistics, it can apply to any group from which you will collect information. This is often people, but it could also be cities of the world, animals, objects, plants, colors, and so on.

- A sample is a representative group of a larger population. Random sampling from representative groups allows us to draw broad conclusions about an overall population. This approach is commonly used in polling. Pollsters ask a small group of people about their views on certain topics. They can then use this information to make informed judgments about what the larger population thinks. This saves time, hassle, and the expense of extracting data from an entire population (which for all practical purposes is usually impossible).

Central tendency is the name for measurements that look at the typical central values within a dataset. This does not just refer to the central value within an entire dataset, which is called the median. Rather, it is a general term used to describe a variety of central measurements. For instance, it might include central measurements from different quartiles of a larger dataset. Common measures of central tendency include:

- **The mean:** The average value of all the data points.

- **The median:** The central or middle value in the dataset.

- **The mode:** The value that appears most often in the dataset.

15. Find linear regression equation for the following set of data.

x	2	4	6	8
y	3	7	5	10

ANS:

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
∑x = 20	∑y = 25	∑x2 = 120	∑xy = 144

b

=

n∑xy−(∑x)(∑y)n∑x2−(∑x)2

b

=

4×144–20×254×120–400

b = 0.95

a=∑y∑x2–∑x∑xyn(∑x2)–(∑x)2

a=25×120–20×1444(120)–400

a = 1.5

Linear regression is given by:

$y = a + bx$

$y = 1.5 + 0.95\ x$

16. Explain confusion matrix with example. Define classification accuracy.

ANS: Refer to any standard book

17. What are the data types in Python? Explain mutable and immutable objects in Python

In programming, data type is an important concept.

ANS:

Variables can store data of different types, and different types can do different things.

Python has the following data types built-in by default, in these categories:

Text Type:	str
Numeric Types:	int, float, complex
Sequence Types:	list, tuple, range
Mapping Type:	dict
Set Types:	set, frozenset
Boolean Type:	bool
Binary Types:	bytes, bytearray, memoryview
None Type:	NoneType.

18. Explain list slicing and list comprehension in Python with example.

ANS:

List comprehension is an elegant way to define and create a list in python. We can create lists just like mathematical statements and in one line only. The syntax of list comprehension is easier to grasp.

A list comprehension generally consists of these parts :

1. Output expression,
2. Input sequence,
3. A variable representing a member of the input sequence and
4. An optional predicate part.

In Python, list slicing is a common practice and it is the most used technique for programmers to solve efficient problems. Consider a python list, In-order to access a range of elements in a list, you need to slice a list. One way to do this is to use the simple slicing operator i.e. colon(:)
With this operator, one can specify where to start the slicing, where to end, and specify the step. List slicing returns a new list from the existing list.

Syntax:
Lst[Initial : End : IndexJump]

19. List down the conditions for overfitting and underfitting. What is meant by imbalanced data?

ANS:

Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has low bias and high variance.

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

20. How can we select an appropriate value of k in k-means clustering algorithm? Define precision and recall.

ANS:

There is a popular method known as **elbow method** which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster

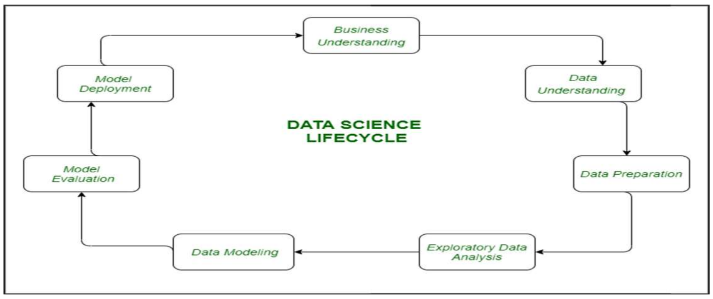
Precision is calculated by dividing the true positives by anything that was predicted as a positive. Recall. Recall (or True Positive Rate) is calculated by dividing the true positives by anything that should have been predicted as positive.

Long Answer Type Questions (15)

1. a) Explain the data science life cycle with diagram.

ANS:

Data Science Lifecycle revolves around the use of machine learning and different analytical strategies to produce insights and predictions from information in order to acquire a commercial enterprise objective. The complete method includes a number of steps like data cleaning, preparation, modelling, model evaluation, etc. It is a lengthy procedure and may additionally take quite a few months to complete. So, it is very essential to have a generic structure to observe for each and every hassle at hand. The globally mentioned structure in fixing any analytical problem is referred to as a Cross Industry Standard Process for Data Mining or CRISP-DM framework.



b) What are the different techniques in data engineering?

ANS:

Data that is not properly organized and processed leads to long delays and failed analytics projects. Our disciplined multi-phase data engineering process organizes and prepares your data to provide the highest potential for success from data science initiatives.

Phases of Data Engineering

The phases of data engineering are represented by a data pipeline. Our team supports you during each step in the process, from IDS Assessment to Data Visualization and Reporting.

IDS Assessment

We refer to source data stores as Immutable Data Stores (IDS). From both a process and a technology standpoint, source data is usually not changed, or mutated, by the data engineer. Our data engineers ensure your IDS is developed appropriately. Non-traditional, non-relational “append only” data stores are often used as the size of the data you collect and manage grows. These technologies, often referred to as NoSQL, are optimized to efficiently read and write large volumes of data, but sacrifice performance on, or in some cases don’t allow, updates and deletes.

ETL Development

Extract, Load, and Transform (ETL) is a process to extract source data from the IDS, transform it for use in analytics, and then load it to the Analytic Data Store (ADS). Data that is too big for traditional data stores is usually too big for most modeling and analytic algorithms. Our data engineers use tools and techniques such as down-sampling or aggregation to transform the IDS into a new mutable data store used for analytics.

ADS Development

Once the IDS is transformed, the resultant mutable data is loaded to the Analytic Data Store. This secondary store facilitates data fusion (data that has been enriched by other data sources) and rendering data for visualization and reporting. The ADS holds the Analytic Base Table (ABT) used to build the model and represents the dividing line between data engineering and data science. The ABT drives the modelling process and is owned by the data scientist.

Modeling continues to transform the data through a process called scoring. Examples include generating predictions, automating a business process, or “online learning” where subject matter experts make corrections to the training data to improve the next model iteration. The modeling/scoring results are stored in the ADS for use in the visualization and reporting phase.

When additional data is required from the IDS to support the modelling process, the data scientist asks the data engineer to enhance the ETL process to make the new data available in the ADS.

Visualization and Reporting

Data engineers use the results stored in the ADS to build focused reports and visuals that enable end users to make informed business decisions. We work with your team to make results accessible and understandable to stakeholders so they can take action and inform decisions. This can range from spreadsheet or email delivery to enterprise-class visualization tools. We have experience delivering results using our own proprietary tools (such as RADR) and most off-the-shelf visualization tools and libraries.

c) Differentiate between covariance and correlation of data.

ANS:

Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes

2. a) Explain Nominal, Ordinal, Interval and Ratio variables with example.

ANS:

Nominal Scale: 1st Level of Measurement

Nominal Scale, also called the categorical variable scale, is defined as a scale used for labeling variables into distinct classifications and doesn't involve a quantitative value or order. This scale is the simplest of the four variable measurement scales. Calculations done on these variables will be futile as there is no numerical value of the options.

There are cases where this scale is used for the purpose of classification – the numbers associated with variables of this scale are only tags for categorization or division. Calculations done on these numbers will be futile as they have no quantitative significance.

For a question such as:

Where do you live?

1- Suburbs

2- City

3- Town

Nominal scale is often used in research surveys and questionnaires where only variable labels hold significance.

For instance, a customer survey asking "Which brand of smartphones do you prefer?" Options : "Apple"- 1 , "Samsung"-2, "OnePlus"-3.

In this survey question, only the names of the brands are significant for the researcher conducting consumer research or netnography. There is no need for any specific order for these brands. However, while capturing nominal data, researchers conduct analysis based on the associated labels.

In the above example, when a survey respondent selects Apple as their preferred brand, the data entered and associated will be "1". This helped in quantifying and answering the final question – How many respondents selected Apple, how many selected Samsung, and how many went for OnePlus – and which one is the highest.

This is the fundamental of quantitative research, and nominal scale is the most fundamental research scale.

Nominal Scale Data and Analysis

There are two primary ways in which nominal scale data can be collected:

By asking an open-ended question, the answers of which can be coded to a respective number of label decided by the researcher.

The other alternative to collect nominal data is to include a multiple choice question in which the answers will be labeled.

In both cases, the analysis of gathered data will happen using percentages or mode,i.e., the most common answer received for the question. It is possible for a single question to have more than one mode as it is possible for two common favorites can exist in a target population.

Nominal Scale Examples

Gender

Political preferences

Place of residence

What is your Gender? What is your Political preference? Where do you live?

M- Male

F- Female 1- Independent

2- Democrat

3- Republican 1- Suburbs

2- City

3- Town

Create a free account

Nominal Scale SPSS

In SPSS, you can specify the level of measurement as scale (numeric data on an interval or ratio scale), ordinal, or nominal. Nominal and ordinal data can be either string alphanumeric or numeric.

Upon importing the data for any variable into the SPSS input file, it takes it as a scale variable by default since the data essentially contains numeric values. It is important to change it to either nominal or ordinal or keep it as scale depending on the variable the data represents.

Ordinal Scale: 2nd Level of Measurement

Ordinal Scale is defined as a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables. These scales are generally used to depict non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc. It is quite straightforward to remember the implementation of this scale as ‘Ordinal’ sounds similar to ‘Order’, which is exactly the purpose of this scale.

Ordinal Scale maintains descriptional qualities along with an intrinsic order but is void of an origin of scale and thus, the distance between variables can’t be calculated. Descriptional qualities indicate tagging properties similar to the nominal scale, in addition to which, the ordinal scale also has a relative position of variables. Origin of this scale is absent due to which there is no fixed start or “true zero”.

Ordinal Data and Analysis

Ordinal scale data can be presented in tabular or graphical formats for a researcher to conduct a convenient analysis of collected data. Also, methods such as Mann-Whitney U test and Kruskal–Wallis H test can also be used to analyze ordinal data. These methods are generally implemented to compare two or more ordinal groups.

In the Mann-Whitney U test, researchers can conclude which variable of one group is bigger or smaller than another variable of a randomly selected group. While in the Kruskal–Wallis H test, researchers can analyze whether two or more ordinal groups have the same median or not.

Learn about: Nominal vs. Ordinal Scale

Ordinal Scale Examples

Status at workplace, tournament team rankings, order of product quality, and order of agreement or satisfaction are some of the most common examples of the ordinal Scale. These scales are generally used in market research to gather and evaluate relative feedback about product satisfaction, changing perceptions with product upgrades, etc.

For example, a semantic differential scale question such as:

How satisfied are you with our services?

Very Unsatisfied – 1

Unsatisfied – 2

Neutral – 3

Satisfied – 4

Very Satisfied – 5

Here, the order of variables is of prime importance and so is the labeling. Very unsatisfied will always be worse than unsatisfied and satisfied will be worse than very satisfied.

This is where ordinal scale is a step above nominal scale – the order is relevant to the results and so is their naming.

Analyzing results based on the order along with the name becomes a convenient process for the researcher.

If they intend to obtain more information than what they would collect using a nominal scale, they can use the ordinal scale.

This scale not only assigns values to the variables but also measures the rank or order of the variables, such as:

Grades

Satisfaction

Happiness

How satisfied are you with our services?

1- Very Unsatisfied

2- Unsatisfied

3- Neural

4- Satisfied

5- Very Satisfied

Interval Scale: 3rd Level of Measurement

Interval Scale is defined as a numerical scale where the order of the variables is known as well as the difference between these variables. Variables that have familiar, constant, and computable differences are classified using the

Interval scale. It is easy to remember the primary role of this scale too, 'Interval' indicates 'distance between two entities', which is what Interval scale helps in achieving.

These scales are effective as they open doors for the statistical analysis of provided data. Mean, median, or mode can be used to calculate the central tendency in this scale. The only drawback of this scale is that there no pre-decided starting point or a true zero value.

Interval scale contains all the properties of the ordinal scale, in addition to which, it offers a calculation of the difference between variables. The main characteristic of this scale is the equidistant difference between objects.

For instance, consider a Celsius/Fahrenheit temperature scale –

80 degrees is always higher than 50 degrees and the difference between these two temperatures is the same as the difference between 70 degrees and 40 degrees.

Also, the value of 0 is arbitrary because negative values of temperature do exist – which makes the Celsius/Fahrenheit temperature scale a classic example of an interval scale.

Interval scale is often chosen in research cases where the difference between variables is a mandate – which can't be achieved using a nominal or ordinal scale. The Interval scale quantifies the difference between two variables whereas the other two scales are solely capable of associating qualitative values with variables.

The mean and median values in an ordinal scale can be evaluated, unlike the previous two scales.

In statistics, interval scale is frequently used as a numerical value can not only be assigned to variables but calculation on the basis of those values can also be carried out.

Even if interval scales are amazing, they do not calculate the "true zero" value which is why the next scale comes into the picture.

Interval Data and Analysis

All the techniques applicable to nominal and ordinal data analysis are applicable to Interval Data as well. Apart from those techniques, there are a few analysis methods such as descriptive statistics, correlation regression analysis which is extensively for analyzing interval data.

Descriptive statistics is the term given to the analysis of numerical data which helps to describe, depict, or summarize data in a meaningful manner and it helps in calculation of mean, median, and mode.

Interval Scale Examples

There are situations where attitude scales are considered to be interval scales.

Apart from the temperature scale, time is also a very common example of an interval scale as the values are already established, constant, and measurable.

Calendar years and time also fall under this category of measurement scales.

Likert scale, Net Promoter Score, Semantic Differential Scale, Bipolar Matrix Table, etc. are the most-used interval scale examples.

The following questions fall under the Interval Scale category:

What is your family income?

What is the temperature in your city?

Create a free account

Ratio Scale: 4th Level of Measurement

Ratio Scale is defined as a variable measurement scale that not only produces the order of variables but also makes the difference between variables known along with information on the value of true zero. It is calculated by assuming that the variables have an option for zero, the difference between the two variables is the same and there is a specific order between the options.

With the option of true zero, varied inferential, and descriptive analysis techniques can be applied to the variables. In addition to the fact that the ratio scale does everything that a nominal, ordinal, and interval scale can do, it can also establish the value of absolute zero. The best examples of ratio scales are weight and height. In market research, a ratio scale is used to calculate market share, annual sales, the price of an upcoming product, the number of consumers, etc.

Ratio scale provides the most detailed information as researchers and statisticians can calculate the central tendency using statistical techniques such as mean, median, mode, and methods such as geometric mean, the coefficient of variation, or harmonic mean can also be used on this scale.

Ratio scale accommodates the characteristic of three other variable measurement scales, i.e. labeling the variables, the significance of the order of variables, and a calculable difference between variables (which are usually equidistant).

Because of the existence of true zero value, the ratio scale doesn't have negative values.

To decide when to use a ratio scale, the researcher must observe whether the variables have all the characteristics of an interval scale along with the presence of the absolute zero value.

Mean, mode and median can be calculated using the ratio scale.

Ratio Data and Analysis

At a fundamental level, Ratio scale data is quantitative in nature due to which all quantitative analysis techniques such as SWOT, TURF, Cross-tabulation, Conjoint, etc. can be used to calculate ratio data. While some techniques such as SWOT and TURF will analyze ratio data in such as manner that researchers can create roadmaps of how to improve products or services and Cross-tabulation will be useful in understanding whether new features will be helpful to the target market or not.

Ratio Scale Examples

The following questions fall under the Ratio Scale category:

What is your daughter's current height?

Less than 5 feet.

5 feet 1 inch – 5 feet 5 inches

5 feet 6 inches- 6 feet

More than 6 feet

What is your weight in kilograms?

Less than 50 kilograms

51- 70 kilograms

71- 90 kilograms

91-110 kilograms

More than 110 kilograms

b) Discuss the different techniques for dimensionality reduction.

ANS:

Common techniques of Dimensionality Reduction

- Principal Component Analysis.
- Backward Elimination.
- Forward Selection.
- Score comparison.
- Missing Value Ratio.
- Low Variance Filter.
- High Correlation Filter.
- Random Forest.

c) What is Vector Space Model (VSM)?

ANS:

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers (such as index terms). It is used in information filtering, information retrieval, indexing and relevancy rankings.

3. a) Discuss Normal, t-student and Chi-squared distribution of data.

ANS:

Any standard text/reference books

b) What is the role of Eigen Value and Eigen Vector in PCA?

ANS:

Eigenvalues are coefficients applied to eigenvectors that give the vectors their length or magnitude. So, PCA is a method that: Measures how each variable is associated with one another using a Covariance matrix. Understands the directions of the spread of our data using Eigenvectors.

4. a) Given the following dataset use PCA to reduce the dimension from 2 to 1. Draw the 1-dimensional data distribution after dimensionality reduction.

ANS:

Any standard text/reference books

b) What is covariance matrix?

ANS:

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 matrix of this from:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

Since the covariance of a variable with itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. And since the covariance is commutative (Cov(a,b)=Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

c) What is LDA?

ANS:

Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes. It is used to project the features in higher dimension space into a lower dimension space.

For example, we have two classes and we need to separate them efficiently. Classes can have multiple features. Using only a single feature to classify them may result in some overlapping as shown in the below figure. So, we will keep on increasing the number of features for proper classification.

5. a) Write down the steps of PCA algorithm.

ANS:

Step 1: Standardize the dataset.

Step 2: Calculate the covariance matrix for the features in the dataset.

Step 3: Calculate the eigenvalues and eigenvectors for the covariance matrix.

Step 4: Sort eigenvalues and their corresponding eigenvectors.

Step 5: Pick k eigenvalues and form a matrix of eigenvectors.

Step 6: Transform the original matrix.

(Explain each step briefly)

b) What are applications of PCA algorithm?

ANS:

Some of the applications of Principal Component Analysis (PCA) are:

- Spike-triggered covariance analysis in Neuroscience
- Quantitative Finance
- Image Compression
- Facial Recognition
- Other applications like Medical Data correlation
- Feature reduction
- Dimensionality reduction & Feature extraction

c) Name the graphical techniques used in EDA?

ANS:

Typical graphical techniques used in EDA are:

Box plot

Histogram

Multi-vari chart

Run chart

Pareto chart

Scatter plot (2D/3D)

Stem-and-leaf plot

Parallel coordinates

Odds ratio

Targeted projection pursuit

Heat map

Bar chart

Horizon graph

6. a) Discuss the different techniques used in EDA in details.

ANS:

There are four exploratory data analysis techniques that data experts use, which include:

Univariate Non-Graphical

This is the simplest type of EDA, where data has a single variable. Since there is only one variable, data professionals do not have to deal with relationships.

Univariate Graphical

Non-graphical techniques do not present the complete picture of data. Therefore, for comprehensive EDA, data specialists implement graphical methods, such as stem-and-leaf plots, box plots, and histograms.

Multivariate Non-Graphical

Multivariate data consists of several variables. Non-graphic multivariate EDA methods illustrate relationships between 2 or more data variables using statistics or cross-tabulation.

Multivariate Graphical

This EDA technique makes use of graphics to show relationships between 2 or more datasets. The widely-used multivariate graphics include bar chart, bar plot, heat map, bubble chart, run chart, multivariate chart, and scatter plot.

b) Explain normalization and orthogonality of variables.

ANS:

Refer to any standard text book

7. a) Discuss the different techniques of Descriptive and Inferential Statistics.

ANS:

Descriptive statistics are used to describe the characteristics or features of a dataset. The term ‘descriptive statistics’ can be used to describe both individual quantitative observations (also known as ‘summary statistics’) as well as the overall process of obtaining insights from these data. We can use descriptive statistics to describe both an entire population or an individual sample. Because they are merely explanatory, descriptive statistics are not heavily concerned with the differences between the two types of data.

So what measures do descriptive statistics look at? While there are many, important ones include:

Distribution

Central tendency

Variability

Let’s briefly look at each of these now.

What is distribution?

Distribution shows us the frequency of different outcomes (or data points) in a population or sample. We can show it as numbers in a list or table, or we can represent it graphically. As a basic example, the following list shows the number of those with different hair colors in a dataset of 286 people.

Brown hair:	130
Black hair:	39
Blond hair:	91
Auburn hair:	13
Gray hair:	13

We can also represent this information visually, for instance in a pie chart.

Generally, using visualizations is common practice in descriptive statistics. It helps us more readily spot patterns or trends in a dataset.

What is central tendency?

Central tendency is the name for measurements that look at the typical central values within a dataset. This does not just refer to the central value within an entire dataset, which is called the median. Rather, it is a general term used to describe a variety of central measurements. For instance, it might include central measurements from different quartiles of a larger dataset. Common measures of central tendency include:

The mean: The average value of all the data points.

The median: The central or middle value in the dataset.

The mode: The value that appears most often in the dataset.

Once again, using our hair color example, we can determine that the mean measurement is 57.2 (the total value of all the measurements, divided by the number of values), the median is 39 (the central value) and the mode is 13 (because it appears twice, which is more than any of the other data points). Although this is a heavily simplified example, for many areas of data analysis these core measures underpin how we summarize the features of a data sample or population. Summarizing these kinds of statistics is the first step in determining other key characteristics of a dataset, for example, its variability. This leads us to our next point…

What is variability?

The variability, or dispersion, of a dataset, describes how values are distributed or spread out. Identifying variability relies on understanding the central tendency measurements of a dataset. However, like central tendency, variability is not just one measure. It is a term used to describe a range of measurements. Common measures of variability include:

Standard deviation: This shows us the amount of variation or dispersion. Low standard deviation implies that most values are close to the mean. High standard deviation suggests that the values are more broadly spread out.

Minimum and maximum values: These are the highest and lowest values in a dataset or quartile. Using the example of our hair color dataset again, the minimum and maximum values are 13 and 130 respectively.

Range: This measures the size of the distribution of values. This can be easily determined by subtracting the smallest value from the largest. So, in our hair color dataset, the range is 117 (130 minus 13).

Kurtosis: This measures whether or not the tails of a given distribution contain extreme values (also known as outliers). If a tail lacks outliers, we can say that it has low kurtosis. If a dataset has a lot of outliers, we can say it has high kurtosis.

Skewness: This is a measure of a dataset's symmetry. If you were to plot a bell-curve and the right-hand tail was longer and fatter, we would call this positive skewness. If the left-hand tail is longer and fatter, we call this negative skewness. This is visible in the following image.

Inferential Statistics:

We've established that descriptive statistics focus on summarizing the key features of a dataset. Meanwhile, inferential statistics focus on making generalizations about a larger population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.

Unsurprisingly, the accuracy of inferential statistics relies heavily on the sample data being both accurate and representative of the larger population. To do this involves obtaining a random sample. If you've ever read news coverage of scientific studies, you'll have come across the term before. The implication is always that random sampling means better results. On the flipside, results that are based on biased or non-random samples are usually thrown out. Random sampling is very important for carrying out inferential techniques, but it is not always straightforward!

1. Defining a population

This simply means determining the pool from which you will draw your sample. As we explained earlier, a population can be anything—it isn't limited to people. So it could be a population of objects, cities, cats, pugs, or anything else from which we can derive measurements!

2. Deciding your sample size

The bigger your sample size, the more representative it will be of the overall population. Drawing large samples can be time-consuming, difficult, and expensive. Indeed, this is why we draw samples in the first place—it is rarely feasible to draw data from an entire population. Your sample size should therefore be large enough to give you confidence in your results but not so small that the data risk being unrepresentative (which is just shorthand for inaccurate). This is where using descriptive statistics can help, as they allow us to strike a balance between size and accuracy.

3. Randomly select a sample

Once you've determined the sample size, you can draw a random selection. You might do this using a random number generator, assigning each value a number and selecting the numbers at random. Or you could do it using a range of similar techniques or algorithms (we won't go into detail here, as this is a topic in its own right, but you get the idea).

4. Analyze the data sample

Once you have a random sample, you can use it to infer information about the larger population. It's important to note that while a random sample is representative of a population, it will never be 100% accurate. For instance, the mean (or average) of a sample will rarely match the mean of the full population, but it will give you a good idea of it. For this reason, it's important to incorporate your error margin in any analysis (which we cover in a moment). This is why, as explained earlier, any result from inferential techniques is in the form of a probability.

However, presuming we've obtained a random sample, there are many inferential techniques for analyzing and obtaining insights from those data. The list is long, but some techniques worthy of note include:

Hypothesis testing

Confidence intervals

Regression and correlation analysis

b) Explain the following: Standard Deviation, Kurtosis and Skewness of data.

ANS:

Standard deviation: This shows us the amount of variation or dispersion. Low standard deviation implies that most values are close to the mean. High standard deviation suggests that the values are more broadly spread out.

Kurtosis: This measures whether or not the tails of a given distribution contain extreme values (also known as outliers). If a tail lacks outliers, we can say that it has low kurtosis. If a dataset has a lot of outliers, we can say it has high kurtosis.

Skewness: This is a measure of a dataset's symmetry. If you were to plot a bell-curve and the right-hand tail was longer and fatter, we would call this positive skewness. If the left-hand tail is longer and fatter, we call this negative skewness. This is visible in the following image.

8. a) Describe the steps of obtaining a random sample of data.

ANS:

Random sampling can be a complex process and often depends on the particular characteristics of a population. However, the fundamental principles involve:

1. Defining a population

This simply means determining the pool from which you will draw your sample. As we explained earlier, a population can be anything—it isn't limited to people. So it could be a population of objects, cities, cats, pugs, or anything else from which we can derive measurements!

2. Deciding your sample size

The bigger your sample size, the more representative it will be of the overall population. Drawing large samples can be time-consuming, difficult, and expensive. Indeed, this is why we draw samples in the first place—it is rarely feasible to draw data from an entire population. Your sample size should therefore be large enough to give you confidence in your results but not so small that the data risk being unrepresentative (which is just shorthand for inaccurate). This is where using descriptive statistics can help, as they allow us to strike a balance between size and accuracy.

3. Randomly select a sample

Once you've determined the sample size, you can draw a random selection. You might do this using a random number generator, assigning each value a number and selecting the numbers at random. Or you could do it using a range of similar techniques or algorithms (we won't go into detail here, as this is a topic in its own right, but you get the idea).

4. Analyze the data sample

Once you have a random sample, you can use it to infer information about the larger population. It's important to note that while a random sample is representative of a population, it will never be 100% accurate. For instance, the mean (or average) of a sample will rarely match the mean of the full population, but it will give you a good idea of it. For this reason, it's important to incorporate your error margin in any analysis (which we cover in a moment). This is why, as explained earlier, any result from inferential techniques is in the form of a probability.

However, presuming we've obtained a random sample, there are many inferential techniques for analyzing and obtaining insights from those data. The list is long, but some techniques worthy of note include:

Hypothesis testing

Confidence intervals

Regression and correlation analysis

b) What is hypothesis testing?

Hypothesis testing involves checking that your samples repeat the results of your hypothesis (or proposed explanation). The aim is to rule out the possibility that a given result has occurred by chance. A topical example of this is the clinical trials for the covid-19 vaccine. Since it's impossible to carry out trials on an entire population, we carry out numerous trials on several random, representative samples instead.

The hypothesis test, in this case, might ask something like: 'Does the vaccine reduce severe illness caused by covid-19?' By collecting data from different sample groups, we can infer if the vaccine will be effective. If all samples show similar results and we know that they are representative and random, we can generalize that the vaccine will have the same effect on the population at large. On the flip side, if one sample shows higher or lower efficacy than the others, we must investigate why this might be. For instance, maybe there was a mistake in the sampling process, or perhaps the vaccine was delivered differently to that group. In fact, it was due to a dosing error that one of the Covid vaccines actually proved to be more effective than other groups in the trial… Which shows how important hypothesis testing can be. If the outlier group had simply been written off, the vaccine would have been less effective!

c) Briefly exemplify confidence interval.

Confidence intervals are used to estimate certain parameters for a measurement of a population (such as the mean) based on sample data. Rather than providing a single mean value, the confidence interval provides a range of values. This is often given as a percentage. If you've ever read a scientific research paper, conclusions drawn from a sample will always be accompanied by a confidence interval.

For example, let's say you've measured the tails of 40 randomly selected cats. You get a mean length of 17.5cm. You also know the standard deviation of tail lengths is 2cm. Using a special formula, we can say the mean length of tails in the full population of cats is 17.5cm, with a 95% confidence interval. Essentially, this tells us that we are 95% certain that the population mean (which we cannot know without measuring the full population) falls within the given range. This technique is very helpful for measuring the degree of accuracy

9. a) Write down the differences between Descriptive and Inferential Statistics.

ANS:

Descriptive statistics:

- Describe the features of populations and/or samples.
- Organize and present data in a purely factual way.
- Present final results visually, using tables, charts, or graphs.

- Draw conclusions based on known data.
- Use measures like central tendency, distribution, and variance.

Inferential statistics:

- Use samples to make generalizations about larger populations.
- Help us to make estimates and predict future outcomes.
- Present final results in the form of probabilities.
- Draw conclusions that go beyond the available data.
- Use techniques like hypothesis testing, confidence intervals, and regression and correlation analysis.

b) Explain null and alternate hypothesis.

ANS:

Definition of Null Hypothesis

A null hypothesis is a statistical hypothesis in which there is no significant difference exist between the set of variables. It is the original or default statement, with no effect, often represented by H_0 (H-zero). It is always the hypothesis that is tested. It denotes the certain value of population parameter such as μ, s, p. A null hypothesis can be rejected, but it cannot be accepted just on the basis of a single test.

Definition of Alternative Hypothesis

A statistical hypothesis used in hypothesis testing, which states that there is a significant difference between the set of variables. It is often referred to as the hypothesis other than the null hypothesis, often denoted by H_1 (H-one). It is what the researcher seeks to prove in an indirect way, by using the test. It refers to a certain value of sample statistic, e.g., x¯, s, p

The acceptance of alternative hypothesis depends on the rejection of the null hypothesis i.e. until and unless null hypothesis is rejected, an alternative hypothesis cannot be accepted.

c) What is the importance of p-value in statistics?

p-value typically ≤ 0.05

This indicates strong evidence against the null hypothesis; so you reject the null hypothesis.

p-value typically > 0.05

This indicates weak evidence against the null hypothesis, so you accept the null hypothesis.

p-value at cutoff 0.05

This is considered to be marginal, meaning it could go either way.

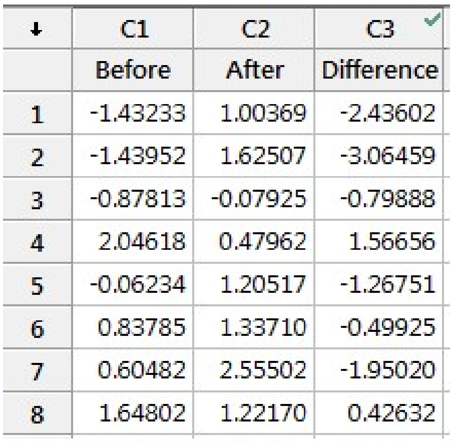
10. a) Explain 1-sample, 2-sample, and Paired t-Tests .

ANS:

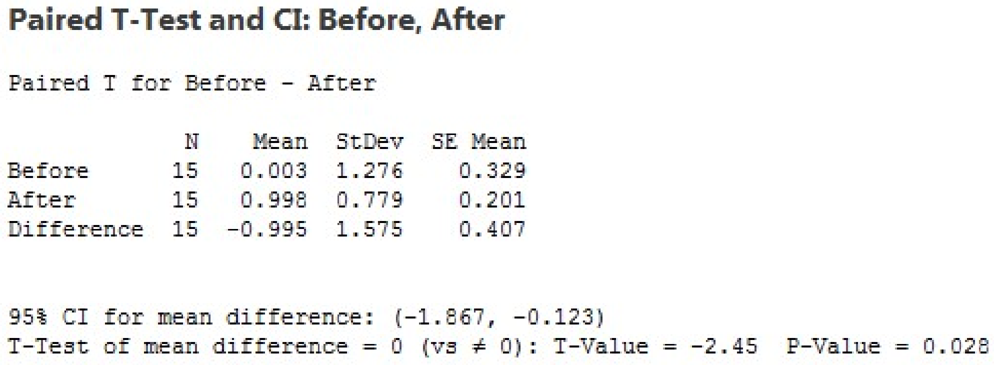
A PAIRED T-TEST IS JUST A 1-SAMPLE T-TEST

Many people are confused about when to use a paired t-test and how it works. I'll let you in on a little secret. The paired t-test and the 1-sample t-test are actually the same test in disguise! As we saw above, a 1-sample t-test compares one sample mean to a null hypothesis value. A paired t-test simply calculates the difference between paired observations (e.g., before and after) and then performs a 1-sample t-test on the differences.

You can test this with this data set to see how all of the results are identical, including the mean difference, t-value, p-value, and confidence interval of the difference.



↓	C1	C2	C3 ✓
	Before	After	Difference
1	-1.43233	1.00369	-2.43602
2	-1.43952	1.62507	-3.06459
3	-0.87813	-0.07925	-0.79888
4	2.04618	0.47962	1.56656
5	-0.06234	1.20517	-1.26751
6	0.83785	1.33710	-0.49925
7	0.60482	2.55502	-1.95020
8	1.64802	1.22170	0.42632

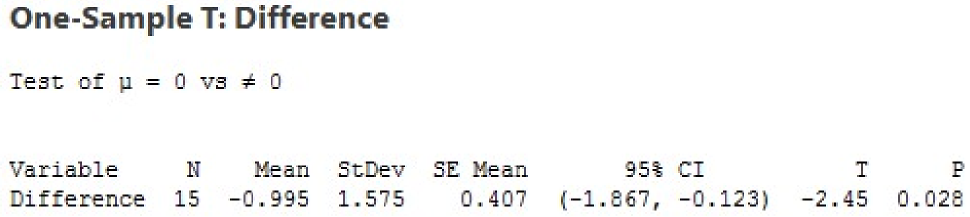


Paired T-Test and CI: Before, After

```
Paired T for Before - After

              N    Mean  StDev  SE Mean
Before       15   0.003  1.276    0.329
After        15   0.998  0.779    0.201
Difference   15  -0.995  1.575    0.407


95% CI for mean difference: (-1.867, -0.123)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -2.45  P-Value = 0.028
```



One-Sample T: Difference

```
Test of μ = 0 vs ≠ 0


Variable     N    Mean  StDev  SE Mean        95% CI             T      P
Difference  15  -0.995  1.575    0.407  (-1.867, -0.123)  -2.45  0.028
```

Understanding that the paired t-test simply performs a 1-sample t-test on the paired differences can really help you understand how the paired t-test works and when to use it. You just need to figure out whether it makes sense to calculate the difference between each pair of observations.

For example, let's assume that "before" and "after" represent test scores, and there was an intervention in between them. If the before and after scores in each row of the example worksheet represent the same subject, it makes sense to calculate the difference between the scores in this fashion—the paired t-test is appropriate. However, if the scores in each row are for different subjects, it doesn't make sense to calculate the difference. In this case, you'd need to use another test, such as the 2-sample t-test, which I discuss below.

Using the paired t-test simply saves you the step of having to calculate the differences before performing the t-test. You just need to be sure that the paired differences make sense!

When it is appropriate to use a paired t-test, it can be more powerful than a 2-sample t-test. For more information, go to Overview for paired t.

HOW TWO-SAMPLE T-TESTS CALCULATE T-VALUES

The 2-sample t-test takes your sample data from two groups and boils it down to the t-value. The process is very similar to the 1-sample t-test, and you can still use the analogy of the signal-to-noise ratio. Unlike the paired t-test, the 2-sample t-test requires independent groups for each sample.

The formula is below, and then some discussion.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s}$$

For the 2-sample t-test, the numerator is again the signal, which is the difference between the means of the two samples. For example, if the mean of group 1 is 10, and the mean of group 2 is 4, the difference is 6.

The default null hypothesis for a 2-sample t-test is that the two groups are equal. You can see in the equation that when the two groups are equal, the difference (and the entire ratio) also equals zero. As the difference between the two groups grows in either a positive or negative direction, the signal becomes stronger.

In a 2-sample t-test, the denominator is still the noise, but Minitab can use two different values. You can either assume that the variability in both groups is equal or not equal, and Minitab uses the corresponding estimate of the variability. Either way, the principle remains the same: you are comparing your signal to the noise to see how much the signal stands out.

Just like with the 1-sample t-test, for any given difference in the numerator, as you increase the noise value in the denominator, the t-value becomes smaller. To determine that the groups are different, you need a t-value that is large.

b) What is meant by t-value?

ANS:

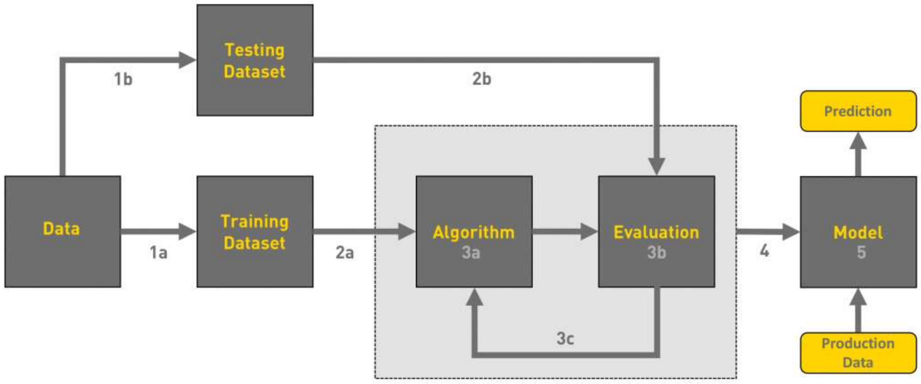
Each type of t-test uses a procedure to boil all of your sample data down to one value, the t-value. The calculations compare your sample mean(s) to the null hypothesis and incorporates both the sample size and the variability in the data. A t-value of 0 indicates that the sample results exactly equal the null hypothesis. In statistics, we call the difference between the sample estimate and the null hypothesis the effect size. As this difference increases, the absolute value of the t-value increases.

That's all nice, but what does a t-value of, say, 2 really mean? From the discussion above, we know that a t-value of 2 indicates that the observed difference is twice the size of the variability in your data. However, we use t-tests to evaluate hypotheses rather than just figuring out the signal-to-noise ratio. We want to determine whether the effect size is statistically significant.

11. a) Draw the generic workflow of any machine learning algorithm.

Ans: The ML workflow

The diagram below gives a high-level overview of the stages in an ML workflow.



b) Discuss the applications of Machine Learning in retail, banking and healthcare sector.

Ans:

Applications in Banking:

- ❑ Plenty of Hedge funds across the globe are using high end systems to deploy artificial intelligence models which learn by taking input from several sources of variation in financial markets and sentiments about the entity to make investment decisions on the fly. Reports claim that more than 70% of the trading today is actually carried out by automated artificial intelligence systems.

- ❑ A few hedge funds active in AI space are: Two Sigma, PDT Partners, DE Shaw, Winton Capital Management, Ketchum Trading, LLC, Citadel, Voleon, Vatic Labs, Cubist, Point72, Man AHL.

- ❑ Fraud detection is one of the fields which has received massive boost in providing accurate and superior results with the intervention of artificial intelligence. It's one of the key areas in banking sector where artificial intelligence systems have excelled the most.

- ❑ Starting from the early example of successful implementation of data analysis techniques in the banking industry is the FICO Falcon fraud assessment system, which is based on a neural network shell to deployment of sophisticated deep learning based artificial intelligence systems.

Applications in Retail:

Walmart's competitor Target was in the news for a data-driven decision-making case of its own. Like most retailers, Target cares about consumers' shopping habits, what drives them, and what can influence them. Consumers tend to have inertia in their habits and getting them to change is very difficult. Decision makers at Target knew, however, that the arrival of a new baby in a family is one point where people do change their shopping habits significantly. In the Target analyst's words, "As soon as we get them buying diapers from us, they're going to start buying everything else too."

Applications in Healthcare

- Identifying Diseases and Diagnosis. ...
- Drug Discovery and Manufacturing. ...
- Medical Imaging Diagnosis. ...
- Personalized Medicine. ...
- Machine Learning-based Behavioral Modification. ...
- Smart Health Records. ...
- Clinical Trial and Research. ...
- Crowdsourced Data Collection.
-

c) What is meant by Discretization and Binarization of data?

ANS:

Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. In contrast, data binarization is used to transform the continuous and discrete attributes into binary attributes. As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.

12. a) Explain parametric and non-parametric statistical learning methods.

ANS:

Parametric Methods: The basic idea behind the parametric method is that there is a set of fixed parameters that uses to determine a probability model that is used in Machine Learning as well. Parametric methods are those methods for which we priory knows that the population is normal, or if not then we can easily approximate it using a normal distribution which is possible by invoking the Central Limit Theorem. Parameters for using the normal distribution is as follows:

- Mean
- Standard Deviation

Eventually, the classification of a method to be parametric is completely depends on the presumptions that are made about a population. There are many parametric methods available some of them are:

- Confidence interval used for – population mean along with known standard deviation.
- The confidence interval is used for – population means along with the unknown standard deviation.
- The confidence interval for population variance.
- The confidence interval for the difference of two means, with unknown standard deviation.

Nonparametric Methods: The basic idea behind the parametric method is no need to make any assumption of parameters for the given population or the population we are studying. In fact, the methods don't depend on the population. Here there is no fixed set of parameters are available, and also there is no distribution (normal distribution, etc.) of any kind is available for use. This is also the reason that nonparametric methods are also referred to as distribution-free methods. Nowadays Non-parametric methods are gaining popularity and an impact of influence some reasons behind this fame is:

- The main reason is that there is no need to be mannered while using parametric methods.
- The second important reason is that we do not need to make more and more assumptions about the population given (or taken) on which we are working on.
- Most of the nonparametric methods available are very easy to apply and to understand also i.e. the complexity is very low.

There are many nonparametric methods are available today but some of them are as follows:

- Spearman correlation test
- Sign test for population means
- U-test for two independent means

b) What is regression analysis? Write down the different regression analysis techniques.

ANS:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression. Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

The different regression analysis techniques are:

1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Polynomial Regression
6. Bayesian Linear Regression

c) Discuss the differences between classification and regression.

ANS:

The most significant difference between regression vs classification is that while regression helps predict a continuous quantity, classification predicts discrete class labels. There are also some overlaps between the two types of machine learning algorithms.

- A regression algorithm can predict a discrete value which is in the form of an integer quantity
- A classification algorithm can predict a continuous value if it is in the form of a class label probability

Let’s consider a dataset that contains student information of a particular university. A regression algorithm can be used in this case to predict the height of any student based on their weight, gender, diet, or subject major. We use regression in this case because height is a continuous quantity. There is an infinite number of possible values for a person’s height.

On the contrary, classification can be used to analyse whether an email is a spam or not spam. The algorithm checks the keywords in an email and the sender’s address is to find out the probability of the email being spam. Similarly, while a regression model can be used to predict temperature for the next day, we can use a classification algorithm to determine whether it will be cold or hot according to the given temperature values.

13. a) Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations: $2Y–X–50 = 0$ and $3Y–2X–10 = 0$.

ANS:

To find:

Mean of the variables X and Y

Correlation coefficient

Solution:

$2y - x - 50 = 0$

$2y - x = 50$ (i)

$3y - 2x - 10 = 0$

$3y - 2x = 10$ (ii)

Solving equation (i) and (ii) simultaneously

$2y - x = 50$ ×2

$3y - 2x = 10$

So, we get

$4y - 2x = 100$

$3y - 2x = 10$

(-) (+) (-)

$y = 90$

Putting value of y in equation (i)

$2y - x = 50$

$2(90) - x = 50$

$180 - x = 50$

$x = 180 - 50$

$x = 130$

So, we get X' = 130 and Y' = 90

Assume equation (i), regression equation of Y on X

$2y - x = 50$

$2y = x + 50$

$$y = \frac{1}{2}x + 25$$

So, $b_{yx} = \frac{1}{2}$

Consider equation (ii), regression equation of X on Y

$3y - 2x = 10$

$2x = 3y - 10$

$$x = \frac{3}{2}y - 5$$

So, $b_{xy} = \frac{3}{2}$

$$r = \sqrt{b_{xy} * b_{yx}}$$

$$r = \sqrt{\frac{1}{2} * \frac{3}{2}}$$

$r = 0.866$

b) What are the assumptions that must hold for a regression model?

ANS:

Important assumptions in regression analysis:

There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

The error terms must be normally distributed.

c) What do you understand by logistic regression?

ANS:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

14. a) Mention the different techniques used for sampling.

ANS:

There are two types of sampling methods:

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random samplingYou want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic samplingAll employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified samplingThe company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling: The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can’t produce generalizable results.

Example: Convenience samplingYou are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

Example: Voluntary response sampling. You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can’t be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria.

Example: Purposive samplingYou want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people.

Example: Snowball samplingYou are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn’t possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

b) What is bias in data science?

ANS:

Bias is a type of error that occurs in a Data Science model because of using an algorithm that is not strong enough to capture the underlying patterns or trends that exist in the data. In other words, this error occurs when the data is too complicated for the algorithm to understand, so it ends up building a model that makes simple assumptions. This leads to lower accuracy because of underfitting. Algorithms that can lead to high bias are linear regression, logistic regression, etc.

c) What is k-fold cross validation?

ANS:

In k-fold cross-validation, we divide the dataset into k equal parts. After this, we loop over the entire dataset k times. In each iteration of the loop, one of the k parts is used for testing, and the other k − 1 parts are used for training. Using k-fold cross-validation, each one of the k parts of the dataset ends up being used for training and testing purposes.

15. a) What is F1 score and how to calculate it?

ANS:

When using classification models in machine learning, a common metric that we use to assess the quality of the model is the F1 Score.

This metric is calculated as:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

where:

- Precision: Correct positive predictions relative to total positive predictions
- Recall: Correct positive predictions relative to total actual positives

b) Define the similarity functions: Jaccard Index, Cosine based similarity and Manhattan distance similarity.

ANS:

i) Cosine Similarity:

Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

ii) **Manhattan distance:**

Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.

iii) The **Jaccard index**, also known as **Intersection over Union** and the **Jaccard similarity coefficient** is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

Formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

c) What is the difference between an error and residual error?

ANS:

An error occurs in values while the prediction gives us the difference between the observed values and the true values of a dataset. Whereas, the residual error is the difference between the observed values and the predicted values. The reason we use the residual error to evaluate the performance of an algorithm is that the true values are never known. Hence, we use the observed values to measure the error using residuals. It helps us get an accurate estimate of the error.

d) What is bias-variance trade-off in data science?

ANS:

When building a model using Data Science or Machine Learning, our goal is to build one that has low bias and variance. We know that bias and variance are both errors that occur due to either an overly simplistic model or an overly complicated model. Therefore, when we are building a model, the goal of getting high accuracy is only going to be accomplished if we are aware of the tradeoff between bias and variance.
Bias is an error that occurs when a model is too simple to capture the patterns in a dataset. To reduce bias, we need to make our model more complex. Although making the model more complex can lead to reducing bias, and if we make the model too complex, it may end up becoming too rigid, leading to high variance. So, the tradeoff between bias and variance is that if we increase the complexity, the bias reduces and the variance increases, and if we reduce complexity, the bias increases and the variance reduces. Our goal is to find a point at which our model is complex enough to give low bias but not so complex to end up having high variance.

16. a) What is RMSE?

ANS:

RMSE stands for the root mean square error. It is a measure of accuracy in regression. RMSE allows us to calculate the magnitude of error produced by a regression model. The way RMSE is calculated is as follows:

First, we calculate the errors in the predictions made by the regression model. For this, we calculate the differences between the actual and the predicted values. Then, we square the errors.

After this step, we calculate the mean of the squared errors, and finally, we take the square root of the mean of these squared errors. This number is the RMSE, and a model with a lower value of RMSE is considered to produce lower errors, i.e., the model will be more accurate.

b) How to calculate binary classification error using its confusion matrix?

ANS:

In a binary classification algorithm, we have only two labels, which are True and False. Before we can calculate the accuracy, we need to understand a few key terms:

- True positives: Number of observations correctly classified as True
- True negatives: Number of observations correctly classified as False
- False positives: Number of observations incorrectly classified as True
- False negatives: Number of observations incorrectly classified as False

To calculate the accuracy, we need to divide the sum of the correctly classified observations by the number of total observations.

c) What is ensemble learning?

ANS:

When we are building models using Data Science and Machine Learning, our goal is to get a model that can understand the underlying trends in the training data and can make predictions or classifications with a high level of accuracy.

However, sometimes some datasets are very complex, and it is difficult for one model to be able to grasp the underlying trends in these datasets. In such situations, we combine several individual models together to improve performance. This is what is called ensemble learning.

d) What is a kernel function in SVM?

ANS:

In the SVM algorithm, a kernel function is a special mathematical function. In simple terms, a kernel function takes data as input and converts it into a required form. This transformation of the data is based on something called a kernel trick, which is what gives the kernel function its name. Using the kernel function, we can transform the data that is not linearly separable (cannot be separated using a straight line) into one that is linearly separable.

17. a) Explain univariate, bivariate, and multivariate analyses.

ANS:

When we are dealing with data analysis, we often come across terms such as univariate, bivariate, and multivariate. Let’s try and understand what these mean.

- Univariate analysis: Univariate analysis involves analyzing data with only one variable or, in other words, a single column or a vector of the data. This analysis allows us to understand the data and extract patterns and trends out of it. Example: Analyzing the weight of a group of people.

- Bivariate analysis: Bivariate analysis involves analyzing the data with exactly two variables or, in other words, the data can be put into a two-column table. This kind of analysis allows us to figure out the relationship between the variables. Example: Analyzing the data that contains temperature and altitude.

- Multivariate analysis: Multivariate analysis involves analyzing the data with more than two variables. The number of columns of the data can be anything more than two. This kind of analysis allows us to figure out the effects of all other variables (input variables) on a single variable (the output variable).

Example: Analyzing data about house prices, which contains information about the houses, such as locality, crime rate, area, the number of floors, etc.

b) How can we handle missing data?

ANS:

To be able to handle missing data, we first need to know the percentage of data missing in a particular column so that we can choose an appropriate strategy to handle the situation.

For example, if in a column the majority of the data is missing, then dropping the column is the best option, unless we have some means to make educated guesses about the missing values. However, if the amount of missing data is low, then we have several strategies to fill them up.

One way would be to fill them all up with a default value or a value that has the highest frequency in that column, such as 0 or 1, etc. This may be useful if the majority of the data in that column contains these values.

Another way is to fill up the missing values in the column with the mean of all the values in that column. This technique is usually preferred as the missing values have a higher chance of being closer to the mean than to the mode.

Finally, if we have a huge dataset and a few rows have values missing in some columns, then the easiest and fastest way is to drop those columns. Since the dataset is large, dropping a few columns should not be a problem anyway.

c) How can we deal with outliers?

ANS:

Outliers can be dealt with in several ways. One way is to drop them. We can only drop the outliers if they have values that are incorrect or extreme. For example, if a dataset with the weights of babies has a value 98.6-degree Fahrenheit, then it is incorrect. Now, if the value is 187 kg, then it is an extreme value, which is not useful for our model.

In case the outliers are not that extreme, then we can try:

- A different kind of model. For example, if we were using a linear model, then we can choose a non-linear model
- Normalizing the data, which will shift the extreme values closer to other data points
- Using algorithms that are not so affected by outliers, such as random forest, etc.

18. a) Explain bagging in data science.

ANS:

Bagging is an ensemble learning method. It stands for bootstrap aggregating. In this technique, we generate some data using the bootstrap method, in which we use an already existing dataset and generate multiple samples of the *N* size. This bootstrapped data is then used to train multiple models in parallel, which makes the bagging model more robust than a simple model.

Once all the models are trained, when it's time to make a prediction, we make predictions using all the trained models and then average the result in the case of regression, and for classification, we choose the result, generated by models, that have the highest frequency.

b) What are the popular Python libraries used in data science?

ANS:

Below are the popular libraries used for data extraction, cleaning, visualization, and deploying DS models:

- TensorFlow: Supports parallel computing with impeccable library management backed by Google.
- SciPy: Mainly used for solving differential equations, multidimensional programming, data manipulation, and visualization through graphs and charts.
- Pandas: Used to implement the ETL(Extracting, Transforming, and Loading the datasets) capabilities in business applications.
- Matplotlib: Being free and open-source, it can be used as a replacement for MATLAB, which results in better performance and low memory consumption.
- PyTorch: Best for projects which involve Machine Learning algorithms and Deep Neural Networks.

c) What are the Support Vectors in SVM?

ANS:

The data/vector points closest to the hyperplane (black line) are known as the support vector (SV) data points because only these two points are contributing to the result of the algorithm (SVM), other points are not.

d) What is the basic principle of a Support Vector Machine?

ANS:

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.

19. a) What are the hard margin and soft margin SVMs?

ANS:

Soft Margin – As most of the real-world data are not fully linearly separable, we will allow some margin violation to occur which is called soft margin classification. It is better to have a large margin, even though some constraints are violated. Margin violation means choosing a hyperplane, which can allow some data points to stay on either the incorrect side of the hyperplane and between the margin and correct side of the hyperplane.

Hard Margin – If the training data is linearly separable, we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible.

b) What is the role of C hyper-parameter in SVM?

ANS:

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable.

c) Explain different types of kernel functions in SVM.

ANS:

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

20. a) What affects the decision boundary in SVM?

ANS:

According to the SVM algorithm we find the points closest to the line from both the classes.These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane. Optimal Hyperplane using the SVM algorithm. Thus SVM tries to make a decision boundary in such a way that the separation between the two classes(that street) is as wide as possible.

b) What is bootstrapping in Bagging and Random Forest?

ANS:

RF is a techniques of ensemble learning through Bagging.: Bagging = Bootstrap + Aggregation

Bootstrap means that instead of training on all the observations, each tree of RF is trained on a subset of the observations. The chosen subset is called the bag, and the remaining are called Out of Bag samples.

Multiple trees are trained on different bags, and later the results from all the trees are aggregated. The aggregation step helps reduce Variance.

c) Write down the steps of Random Forest Algorithm.

ANS:

The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.