80
Questions & Answers
that are discussed
In this course

Statistics

1. What is the difference between Overfitting and Underfitting?

Answer:

Most people think that Model fitting is only a Machine Learning concept but it is not true. Model fitting is a very old concept of Statistics and it is widely used in Machine Learning as well. To create models, we divide data set into two parts train data set and test data set. Test data set is also called new data. So, don't get confused with the term. We create model on train data set and test the model on test data set.

The main difference between overfitting and underfitting is that Overfitting has high accuracy model on training data set but it does not perform well on test data set. and Underfitting has low accuracy model so obviously it won't perform well on test data set.

2. Which one would you choose to select a Linear Regression Model, R-Square or Adjusted R-Square?

Answer:

R Square and Adjusted R Square are two model accuracy measures for linear regression. The R Square will always increase if we add more independent variables to the model but adjusted R Square will only increase if the newly added variable is improving the accuracy of the model. Therefore, we will choose adjusted R square to select a model.

3. What are Type I and Type II Errors?

Answer:

Type I error is the probability of rejecting the null hypothesis when it is true. For example: The chances of not trusting an honest person. Here our null hypothesis is the person is honest and alternative hypothesis is the person liar but we rejected the null hypothesis.

Type II error is the probability of not rejecting the null hypothesis when it is false. For example: The chances of trusting a liar person. Here our null hypothesis is the person is honest and alternative hypothesis is the person liar but we do not reject the null hypothesis.

4. Explain why a continuity correction is needed when a discrete random variable is approximated by a continuous random variable in order to calculate a probability for the discrete random variable?

Answer:

A continuity correction is needed when a discrete random variable is approximated by a continuous random variable in order to calculate a probability for the discrete random variable because a discrete random variable considers integer values while a continuous variable consider real numbers. A continuity correction makes it easy to compute probabilities of each discrete random variable when the sample values are large.

For example, the normal distribution is a continuous distribution which can take any values within an interval and is used as an approximation of binomial distribution.

5. What is a p-value?

Answer:

Definition of p-value is one of the most common questions in Data Science interviews. And most Data Science interviewers complains that people don't know what a p-value is. There are so many people who knows how to use p-value but they don't understand the meaning of p-value.

The p-value is the probability of getting a sample statistic at least as extreme as we got using our sample data if we take more samples of the same size when the null hypothesis is true. If this probability is greater than the level of significance (say 5%) then we do not reject the null hypothesis because that means the chances our sample statistic is more likely to happen under the null hypothesis and hence it is compatible with the null hypothesis. If this probability is less than the level of significance then we reject the null hypothesis because that means the chances our sample statistic is unlikely to happen under the null hypothesis and hence it is not compatible with the null hypothesis.

6. If you have a data set and some variables are highly correlated to each other. You need to run Principal Component Analysis on this data, would you remove correlated variables from the data? If Yes then why and if No then why not?

Answer:

We should remove correlated variables first because if we include correlated variables then the variance of components that have correlated variables will be high. Also, Principal Component Analysis will give more importance to correlated variables therefore the components will be misleading.

7. What does Interpolation and Extrapolation mean in Regression Modeling?

Answer:

Interpolation is estimating new value of dependent variable using the values of independent variables that exists in the data set.

Extrapolation is estimating new value of dependent variable using the values of independent variables that do not exists in the data set.

8. What are confounding variables?

Answer:

The confounding variables are the variables that are correlated with dependent variable and independent variable but not included in the model.

For example, if we want to check the effect of weight gain on blood pressure then fat intake can be a confounding variable which is linearly related to weight gain and blood pressure.

9. What is AIC(Akaike Information Criteria)?

Answer:

AIC is the measure of goodness-of-fit for logistic model. It penalizes the model for more number of variables therefore a model with minimum value of AIC is preferred so that we get the best model with less number of variables. It is also called as analogous measure of adjusted R^2 in logistic regression.

10. What is the use of orthogonal rotation in principal component analysis?

Answer:

Orthogonal rotation maximizes the difference between variance captured by the component so that the overall variance of the data set can be explained by the components. Our objective from PCA is to get a smaller number of components than independent variables which helps to reduce the dimension of the data set and analysis becomes easier. If the components are not rotated then the effect of PCA will be reduced and we will require a greater number of components to explain the variance in the data set. Which will not fulfil our objective from PCA therefore we should rotate the components.

Probability

1. One hundred people line up to board an airplane. Each has a boarding pass with assigned seat. However, the first person to board has lost his boarding pass and takes a random seat. After that, each person takes the assigned seat if it is unoccupied, and one of unoccupied seats at random otherwise. What is the probability that the last person to board gets to sit in his assigned seat?

Answer:

This problem will be easily solved if you focus on the seat assigned to first person and

the seat assigned to last person. First person might sit on the last person's seat or on his seat. That means, we have two only seats available for the last person. His seat and the seat of the first person. Therefore, the probability that the last person will sit on his seat is Total number of seats assigned to him divided by the number of seats he can sit on.

$$\text{Probability}(\text{Last person sit on his seat}) = \frac{\text{Total number of seats assigned to him}}{\text{Total number of seats he can sit on}}$$
$$= \frac{1}{2}$$

Which is equal to 1/2.

2. Teams A and B are playing a game with 7 matches. A has probability p to win a match. What is the probability A winning the game at the 7th match?

Answer:

Team A will win the game at the 7th match if and only if A wins any 3 of the first 6 matches

and wins the last match. The probability that A wins any 3 of the first 6 matches can be calculated by using Binomial formula as combination of 3 out 6 matches multiplied with probability of winning in three matches and probability of not winning in three matches.

$$\text{Prob}(\text{A wins any of the first 3 matches}) = {}^{n}C_{x}p^{x}(1-p)^{n-x}$$
$$= {}^{6}C_{3}p^{3}(1-p)^{6-3}$$

The probability that A wins the last match is p.

$$\text{Prob}(\text{A wins the last match}) = p$$

Therefore, the probability that A wins the game at the 7th match is combination of 3 out 6 matches multiplied with probability of winning in four matches and probability of not winning in three matches.

$$\begin{aligned}\text{Prob}(\text{A wins the game at the 7th match}) &= {}^{6}C_{3}p^{3}(1-p)^{6-3}\times p\\ &= {}^{6}C_{3}p^{4}(1-p)^{3}\end{aligned}$$

3. At a college of Nursing, 89% of incoming freshmen nursing students are female and 11% are male. Recent records indicate that 60% of the entering female students will graduate while 80% of the male students will graduate. If an incoming freshmen nursing student is selected at random, what is the probability that student will graduate and this student is female?

Answer:

We have to use Bayes Probability formula to calculate the required probability.

The Bayes formula is:

$$\text{Prob}\begin{pmatrix}\text{Student will graduate}\\ \text{if}\\ \text{Student is Female}\end{pmatrix} = \frac{\text{Prob}\begin{pmatrix}\text{Student will graduate}\\ \text{and student is Female}\end{pmatrix}}{\text{Prob}(\text{Student is Female})}$$

It is given in the question that the probability that the student will graduate if she is a female is 0.60 and the probability that student is female is 0.89.

$$\text{Prob}\begin{pmatrix}\text{Student will graduate}\\ \text{if}\\ \text{Student is Female}\end{pmatrix} = 0.60$$

$$\text{Prob}(\text{Student is Female}) = 0.89$$

$$\begin{aligned}\text{Prob}\begin{pmatrix}\text{Student will graduate}\\ \text{and student is Female}\end{pmatrix} &= \text{Prob}\begin{pmatrix}\text{Student will graduate}\\ \text{if}\\ \text{Student is Female}\end{pmatrix}\times\text{Prob}(\text{Student is Female})\\ &= 0.60\times 0.89\\ &= 0.534\end{aligned}$$

Therefore, the probability that student will graduate and student is female will be 0.534.

4. If a coin is tossed 10 times. What is the probability of getting exactly 5 heads and 5 tails?

Answer:

Each time we toss a coin there are 2 possible outcomes a head or a tail and the coin is tossed ten times therefore there are 2^10 = 1024 outcomes.

We are interested in 5 heads and 5 tails and we can get 5 heads and tails in any sequence. For example, first four are heads, then 5 are tails and last one is head. So, we will get so many unique sequences but the position of heads and tails will be repeated. Therefore, the number of these sequences will be calculated by using permutation with repeated elements as:

$$\text{Number of unique sequences with repeated elements} = \frac{10!}{5!5!}$$

This is the total number of outcomes to get exactly 5 heads and 5 tails. If 10 coins are tossed the probability of getting exactly 5 heads and 5 tails is:

$$\begin{aligned}\text{Prob}\left(5\ \textit{Heads}\ \text{and}\ 5\ \textit{Tails}\right) &= \frac{\text{Total number of outcomes to get exactly 5 Heads and 5 Tails}}{\text{Total number of possible outcomes}} \\ &= \frac{\frac{10!}{5!5!}}{1024} \\ &= \frac{63}{256}\end{aligned}$$

5. There are an equal number of men and women in a room, 5% of the men are color blind and 2.5% of the women are color blind. If a person comes out from the room, what is the probability that this person is color blind?

Answer:

Since men and women are equal that means they are equal to half. Therefore, the probability that the person who comes out from the room is color blind is:

$$\begin{aligned}\text{Prob}\left(\text{Color blind person comes out}\right) &= \frac{1}{2}\times\text{Prob}\left(\text{Color blind men}\right)+\frac{1}{2}\times\text{Prob}\left(\text{Color blind women}\right) \\ &= \frac{1}{2}\times\frac{5}{100}+\frac{1}{2}\times\frac{2.5}{100} \\ &= 0.0375\end{aligned}$$

We are adding these probabilities because color blindness of men and women is independent of each other.

6. A company manufactures DVDs in lots of 50 and they observed a defective rate of 0.5% so the probability of a DVD being defective is 0.005 therefore the probability of a DVD not being defective is 0.995. What is the probability of getting at least one defective DVD in a lot of 50?

Answer:

The probability that at least one defective DVD in a lot of 50 DVD will be calculated by using complement rule. The Complement Rule says that the sum of the probabilities of an event and its complement must equal 1.

The complement of the event that at least one defective DVD is none of the DVDs are defective.

Therefore, the calculation of the probability is:

$$\begin{aligned}\text{Prob}(\text{at least one defective DVD in 50}) &= 1 - \text{Prob}(\text{None DVD defective in 50}) \\ &= 1 - (0.995)^{50} \\ &= 0.222\end{aligned}$$

7. I have a gun in each hand, one with 3 bullets and other with 2. I fire them together at you, what is the probability that you die?

Answer:

We know that each gun can have 6 bullets and it is given that both the guns are fired together. The firing from guns is not mutually exclusive because both the guns are fired together.

There are three scenarios in this question:

first one is - A person can die with 1st gun shot and 2nd gun shot if both shots are aimed correctly.

second one is - A person can die with 1st gun shot if 2nd gun shot is not aimed correctly.

third one is - A person can die with 2nd gun shot if 1st gun shot is not aimed correctly.

Therefore, the probability will be:

$$\text{Prob}(\text{Die}) = \begin{pmatrix} \text{Prob}(\text{Die with 1st gun}) \times \text{Prob}(\text{Die with 2nd gun}) + \\ \text{Prob}(\text{Die with 1st gun}) \times \text{Prob}(\text{Not Die with 2nd gun}) + \\ \text{Prob}(\text{Not Die with 1st gun}) \times \text{Prob}(\text{Die with 2nd gun}) \end{pmatrix}$$

$$= \frac{3}{6} \times \frac{2}{6} + \frac{3}{6} \times \frac{4}{6} + \frac{3}{6} \times \frac{2}{6}$$

$$= \frac{2}{3}$$

8. If 8 boys are arranged in a row, what is the probability that 3 particular boys will sit together?

Answer:

If 8 boys are arranged in a row then they can sit in 8! ways. But if we want three of them to sit together then they can sit in 6! x 3! ways. Where 6! represents the number of ways 5 boys and 1 group of 3 boys can sit and 3! represents the number of ways 3 boys can sit.

Therefore, the probability that 3 particular boys will sit together is calculated by dividing the total number of ways 3 particular boys can sit together with the total number of ways 8 boys can sit as:

$$\text{Prob}(\text{3 particular boys sit together}) = \frac{\begin{matrix}\text{Total number of ways} \\ \text{3 particular boys can} \\ \text{sit together}\end{matrix}}{\begin{matrix}\text{Total number of ways} \\ \text{8 boys can sit}\end{matrix}}$$

$$= \frac{6! \times 3!}{8!}$$

$$= \frac{6! \times 3 \times 2}{8 \times 7 \times 6!}$$

$$= \frac{3}{28}$$

9. If a 5 digit number is formed using digits 1, 2, ... 9 without repetitions, then what is the probability that it will be an even number?

Answer:

The total numbers that can be formed using 1,2,… 9 without repetition are 9 permutation 5 because order of digits matter. The even digits in 1,2, … 9 are 2,4,6,8. Therefore, there are 4 even digits.

The total even numbers among those total numbers will be 8x7x6x5x4. Because first digit can be taken from any 8 digits expect one even digit, second digit can be taken from any 7 digits except the first digit and one even digit, third digit can be taken from any 6 digits except first, second digits, and one even digit, fourth digit can be taken from any 5 digits except first, second, third, and one even digit, the last digit can be taken any digit from even digits. Therefore, the probability will be calculated by:

$$\begin{aligned}\text{Prob}(5\text{ Digit number is even}) &= \frac{8\times7\times6\times5\times4}{9Permutation5}\\ &= \frac{8\times7\times6\times5\times4}{\frac{9!}{(9-5)!}}\\ &= \frac{8\times7\times6\times5\times4\times4!}{9\times8\times7\times6\times5\times4!}\\ &= \frac{4}{9}\end{aligned}$$

10. A card is drawn randomly from a deck of ordinary playing cards. What is the probability that this card is a spade or an ace?

Answer:

Consider S = the event that the card is a spade and A = the event that the card is an ace.

We know that there are 52 cards in the deck which has 13 spades and 4 aces. And 1 ace is also a spade. Therefore, the probability of S is 13/52. The probability of A is 4/52 And the probability that Spade is an Ace is 1/52.

$$\text{Prob}(S) = \frac{13}{52}$$

$$\text{Prob}(A) = \frac{4}{52}$$

$$\text{Prob}(S \cap A) = \frac{1}{52}$$

The probability that the drawn card is a Spade or an Ace will be calculated by using addition rule as:

$$\begin{aligned}\text{Prob}(S \cup A) &= \text{Prob}(S) + \text{Prob}(A) - \text{Prob}(S \cap A) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{4}{13}\end{aligned}$$

Note: In probability theory, Union sign is used to represent "or" and Intersection sign is used to represent "and".

Machine Learning

1. What do you understand by Accuracy, Recall, and Precision?

Answer:

Consider that we have to predict a dichotomous variable which has Yes and No responses as shown below:

Confusion Matrix	Predicted No	Predicted Yes
No	9800	100
Yes	30	70

There were 9000 No responses in the data and 100 Yes responses.

Accuracy is the percentage of correct predictions of positive and negative responses. Considering Yes as a positive response and No as a negative response. The Accuracy will be (9800+70)/10000 = 98.7%

Recall is the percentage of positive responses that were predicted correctly out of total positive responses. Here our positive response is Yes. Therefore, 70 responses are correctly predicted from total 100 positive responses hence the Recall is 70/100 = 70%

Precision is the percentage of positive predictions that were predicted correctly out of total positive predictions. Therefore, 70 cases are correctly predicted from total 100+70 = 170 positive predictions hence the Precision is 70/170 = 41.18%

2. What is the benefit of reducing the dimension of the data before fitting SVM?

Answer:

Support Vector Machine works better in the reduce space and results in high accuracy of predictions. If the number of independent variables is large as compared to the number of observations then it is beneficial to reduce the dimension of the data by using a dimensionality reduction algorithm then apply Support Vector Machine.

Note: Most of the times you will hear the word "feature" in Data Science but don't get confused by it. It is the synonym for Independent Variable.

3. What are Recommender Systems?

Answer:

Recommender Systems are information filtering systems that suggest products or services to users that will be useful for users based on previous data.

For example, if a visitor on Amazon buys Pizza Bread then Amazon recommends the products that were bought or browsed by this visitor and also recommends other products that were bought by other visitors who bought the same Pizza Bread.

4. What is the difference between kNN and k-Means?

Answer:

The main difference is kNN comes under supervised learning and k-Means is an unsupervised learning algorithm. kNN algorithm tries to classify or regress the data based on k number of neighbours, this k can be any number. In this algorithm we know the values of outcome variable that is used as a dependent variable.

k-Means algorithm is used to find homogeneous clusters and the points within each cluster are close to each other. Here k represents the number of clusters but we do not know what values each cluster will contain.

5. What is “k” in k-Means and how you can choose its value?

Answer:

The value of k in k-Means represent the number of clusters. There is no direct way to find this value. We have to find clusters for a fixed value of k then it is used to find the optimal value of k.

The procedure to find the optimal value for k is:

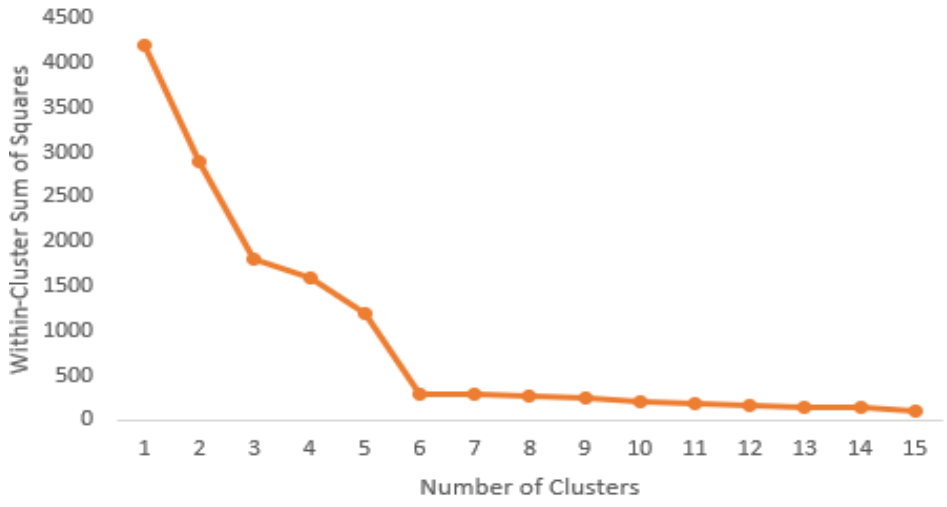
1. Finding clusters for a predefined value of k say 10.

2. Find within-cluster sum of squares for each cluster.

3. Plot within-cluster sum of squares on Y-axis with number of clusters on X-axis.

4. The number of clusters after which there is no significant change in within-cluster sum of squares are considered as optimal number of clusters.

Example:



From the above graph, we can say that the optimal number of clusters will be 6.

6. What is the difference between Euclidean distance & Manhattan distance?

Answer:

Manhattan distance calculates distance horizontally or vertically only therefore it has dimension restrictions.

Euclidean distance calculates the distance in any dimension therefore it does not have any dimension restrictions.

The main difference between the two is Manhattan distance has dimension restrictions and Euclidean distance do not have dimension restrictions.

Therefore, Euclidean distance is a better measure to calculate distance.

7. When do we use regularization in Machine Learning?

Answer:

Regularization is used in Machine Learning when the model suffers from overfitting or underfitting. It introduces a cost term in the model to add more independent variables. Then tries to push the coefficients of many independent variables to become zero and hence reduce the cost term. It is helpful to reduce the complexity of the model so that the predictions using the model becomes better.

8. What do you understand by Random Forest?

Answer:

Random Forest is a decision tree technique used to solve classification as well as regression problems. It grows multiple trees instead of a single tree.

To predict a new value in classification, all trees give their results and the forest of trees chooses the result which gets the highest number of votes then show it as an output.

To predict a new value in regression, all trees give their results and the forest of trees chooses the average of those results then show it as an output.

9. What is Selection Bias?

Answer:

Selection of data in Machine Learning is an initial step of analysis and our objective is that each data point should be randomly selected. If the data points are not randomly selected then we say that there is a selection bias in the data and if selection bias exists in the sample data then it does not represent the population that is intended to be analyzed. Therefore, the results based on this data may not be accurate.

10. What does “Naïve” refer to in Naïve Bayes?

Answer:

Naïve means showing a lack of judgement.

Naïve in Naïve Bayes refers to the lacking to judge the real-world scenarios because Naïve Bayes assumes that all of the independent variables are equally important and they are independent to each other but it is not true in real-world scenarios. Hence this assumption is naive.

SQL

1. What is wrong with the below query?

SELECT UserId, AVG(Total) AS AvgOrderTotal

FROM Invoices

HAVING COUNT(OrderId)>=1

Answer:

This query will get the average order amount by User Id for the customers who have at least 1 order. So, we need to use GROUP BY function to group the customers by User Id. Therefore, the correct query will be:

SELECT UserId, AVG(Total) AS AvgOrderTotal

FROM Invoices

GROUP BY UserId

HAVING COUNT(OrderId)>=1

2. Consider the below tables. Write a query that retrieves all employees that are not recruited by any recruiter.

Employee		
Id	Name	RecruitedBy
1	Ross Taylor	NULL
2	Andy Smith	1
3	Scarlett Berry	NULL
4	Evelyn Depp	3
5	John Lee	3
6	James Dean	NULL

Recruiter	
Id	Name
1	Ross Taylor
2	Evelyn Depp
3	James Dean

Answer:

The query that retrieves all employees that are not recruited by any recruiter is:

SELECT Employee.Name FROM Employee

JOIN Recruiter ON Employee.RecruitedBY = Recruiter.Id

WHERE RecruitedBy = NULL

3. What is the difference between DELETE and TRUNCATE?

Answer:

The DELETE command can be used to remove rows from the table and we can use WHERE clause with it. We can also perform Commit and Rollback after a DELETE command. But TRUNCATE command removes all rows from the table and we cannot use WHERE clause with it. Also, rollback cannot be performed after a TRUNCATE command.

4. Which one would you set as the primary key and which one as unique key in the below table:

Employee ID	FirstName	Salary	Email
0001	Alex	$20000	alex@work.com
0002	John	$35000	john@work.com
0003	Emma	$25000	emma@work.com
0004	Mia	$32000	mia@work.com

Answer:

The primary key in a table is the column which has all the unique values and there is no null value. The unique key in a table is the column which has all the unique values but it can have one null value. Also, there can be many unique keys in a table but there will be only one primary key.

The primary key cannot be NULL therefore, we will set Employee ID as primary key and Email ID as unique key. Because there is a possibility that email ID of an employee is missing so it will be NULL but Employee ID cannot be considered as missing in a database.

5. What is the difference between FULL JOIN and INNER JOIN?

Answer:

The difference between Full Join and Inner Join is they choose the rows from tables in different ways.

Full join Return rows from the two tables even if there are no matching rows between the two. For example, if we want to do a full join for table 1 and table 2 then the full join table will look like as shown below:

Table 1

Company	Revenue($ Billion)
AWS	5
XYZ	6
APS	4.2

Table 2

Company	Revenue($ Billion)
RAP	3.5
APS	4
RAW	5

FULL JOIN table

Company	Revenue($ Billion)	Company	Revenue($ Billion)
APS	4.2	APS	4
AWS	5	-	-
XYZ	6	-	-
-	-	RAP	3.5
-	-	RAW	5

On the other hand, INNER Join Return rows from the two tables if at least one column value of each table matches with the other and it does not return any NULL value.

Suppose we want to join table 1 and table 2 using Company as a matching column to get the revenue and year started. Then it will give us 4.2 and 2009 because company APS is common between the two tables as shown below:

Table 1

Company	Revenue($ Billion)
AWS	5
XYZ	6
APS	4.2

Table 2

Company	Year Started
RAP	2006
APS	2009
RAW	2002

INNER JOIN table

Revenue($Billion)	Year Started
4.2	2009

6. How NULL values and zero or a blank space are different?

Answer:

The NULL value represents a missing value. On the other hand, zero is a number and blank space is a character. We should not get confused if there is a blank space in a database and consider it as a missing value but we should find the reason behind that because missing values are written as NULL in SQL.

7. Consider that you have two tables “sellers” and “sold” and each of them have more than 50000 rows(first five are shown here). Write a query to select top three sellers who sold the most products in total.

Table – sellers

seller_name	product_name
A	Sugar_1kg
A	Butter_100gm
B	Oatmeal_100gm
B	Honey_200gm
C	Pasta_small

Table - sold

product_name	sold_quantity
Sugar_1kg	1000
Butter_100gm	1800
Oatmeal_100gm	2500
Honey_200gm	500
Pasta_small	1200

Answer:

The query to find the top 3 sellers who sold the most products in total is:

```
Select sellers.seller_name, SUM(sold.sold_quantities) AS sold_sum
FROM sellers
JOIN sold
ON sellers.product_name = sold.product_name
GROUP BY sellers.seller_name
ORDER BY sold_sum DESC
LIMIT 3:
```

This query will select sellers name from sellers table whose name matches with the sellers name in sold table then it will sum the total number of quantities sold by the sellers using group by clause because one seller is selling more products. After that the sum will be ordered in descending order to get the seller on top who sold the most products.

8. Identify the mistake in the below query:

```
select case when null = null then 'True' else 'False' end as Result;
```

Answer:

The null value is not compared with an equal sign. We should use IS operator to compare a null value in SQL. So, the correct query will be:

```
select case when null is null then 'True' else 'False' end as Result;
```

9. Suppose you have a table “students”, which has two columns “student_name” and “roll_numbers” of students. Write a query to select student names that start with “S”?

Answer:

The query to select student names that start with “S” is:

```
SELECT * FROM students WHERE student_name like ‘S%'
```

10. Consider that you have two tables “product” and “costs” and each of them have more than 300 rows(first five are shown here). Write a query to print every product name for which the average cost is greater than 1000.

Table – products

product_name	product_id
Smartphone	123
Microwave	511
Smartphone	102
iPhone	319
iPhone	225

Table - costs

cost	product_id
800	123
650	511
700	102
1200	319
1500	225

Answer:

The query to print every product name for which average cost is greater than 1000 will be:

```
SELECT product_name, AVG(costs.cost) AS avg_costs

FROM products

JOIN costs

ON products.product_id = costs.product_id

GROUP BY product_name

HAVING AVG(costs.cost) > 1000;
```

This query will select product name from products table whose product id matches with the product id in costs table then it will average the cost of products by product names using group by clause because one product has different costs. After that it will print only those products for which the average cost is greater than 1000.

R Programming

1. What are the limitations of R?

Answer:

There are many limitations of R but some affect the Data Analysis directly. These limitations are:

It needs to load entire data into memory(RAM), so it is not appropriate for Big Data analysis, processing in R is slower than other programming tools and if the maintainer of the package no longer sustain its package then some R scripts do not work with newer version of R.

2. What is the difference between Inf and NaN?

Answer:

1/0 = Inf

1/0 + 1/0 = Inf

1/0 – 1/0 = NaN

Inf represents the infinity value for example if we divide 1 by 0 then we get infinity.

If we add infinity with infinity we get infinity. But if we subtract infinity from infinity

we don't get infinity there is no value defined for this subtraction. That means if a value that cannot be represented by a number then it is referred to as NaN(Not a number in R).

Note: When you will practice R, you will see NaN if you make a mistake in your code which does not make sense.

3. What is the use of “with” and “by” function in R?

Answer:

with function applies an expression to a dataset. For Example: Suppose we have a data frame called "newdata" which has one group variable and one dependent variable “y” then “with” function to apply one way ANOVA on the data frame can be used as:

with(newdata, aov(y ~ group))

by function applies a function to each level of a factor or factorlist. For example: If we want to find factor wise mean of dependent variable “y” based on “group” factor to a data frame named “newdata” then we can do it as:

by(mydata$y, mydata$group, mean)

4. What is the use of “next” statement in R?

Answer:

next statement is used to skip an iteration in a loop.

For example: if we want to print numbers from 0 to 10 but we don't want 5 then the R code will be:

```
x<-0:10

for (val in x) {

if (val == 5) {

next

}

print(val)

}
```

This code will print values from 0 to 10 and skip 5 therefore we will get 10 values and 5 will not be there.

5. If you want to know all the values in (1, 2, 3, 4, 7, 10) that are not in (2, 5, 10, 11, 17), which in-built function in R can be used to do this? Also, how it can be done without using the in-built function?

Answer:

setdiff function can be used to find the values that are not common between two vectors.

In-built function – setdiff

Code: setdiff(c(1, 2, 3, 4, 7, 10),c(2, 5, 10, 11, 17))

Output: [1] 1 3 4 7

If we don't want to use built-in function then exclamation sign and percentage sign can be used. Exclamation sign represents that "Not values in a vector" and percentage sign combines the two vectors.

Without Inbuilt function – Exclamation sign and Percentage sign

Code: c(1, 2, 3, 4, 7, 10) [!c(1, 2, 3, 4, 7, 10)%in%c(2, 5, 10, 11, 17)]

Output: [1] 1 3 4 7

6. How would you create a function in R? Give an example.

Answer:

The syntax to create a function is:

<object-name>=function(x){

—

—

—

}

We have to give an object-name, this is the name of your function then we write function(x). Here x represents the argument of the function then we write statements within curly brackets these statements will define the functionality of the function.

For example: We can create a function to convert values to 99 if they are less than 3

and 0 if they are greater than 3 then use this function to any variable.

```
CustomFunction<-function(x)
{ifelse(x<3,99,0)
}
```

To check it we created a variable then used CustomFunction to convert the values to 99 if they are less than 3 to 0 if they are greater than 3

Checking CustomFunction

```
Variable<-c(1,2,3,4,5,6,7,8,9)
CustomFunction(Variable)
[1] 99 99  0  0  0  0  0  0  0
```

Our output is correct. Therefore, our function is working.

7. How would you create multiple plots on a single pane?

Answer:

The par(mfrow) function is used to create multiple

plots on a single pane.

For example, if we have four variables (V1, V2, V3, and V4) then their histograms on a single pane can be drawn by the following code:

```
par(mfrow=c(2,2))

hist(V1)

hist(V2)

hist(V3)

hist(V4)
```

8. You have two data sets and you need to combine them. The number of variables is same in both the data sets but the cases are different. How would you combine them?

Answer:

Since the number of variables are same that means we have same columns in both the datasets.

and cases are different so we need to add cases from one data set to another. This can be done using rbind function.

If the question says that number of cases are same but the number of variables is different

then we will be using cbind function.

9. What is the use of “select” function and “filter” function in R?

Answer:

The difference between select function and filter function is that select is applied on columns and filter function is applied on rows. Their use are:

“select” function is used to select some specific columns from the data-set.

“filter” function is used to filter out some rows on the basis of a condition.

10. What will be the output of f(3) for the following function?

```
b<-2
f<-function(a)
{
b<-4
b^2+g(a)
}
g<-function(a)
{
a*b
}
```

Answer:

The value of f(3) will be 22.

Explanation:

The value of “a” passed to the function is 3 and the value for “b” defined in the function

f (a) is 4. So, the output would be 4^2 + g (3). The function g is defined in the global environment and it takes the value of b as 2(due to lexical scoping in R) not 3 returning a value 3*2= 6 to the function f. The result will be 4^2+6= 22.

Python

1. What is Tuple in Python and how it is different from List?

Answer:

Tuple is a sequence of elements and these elements can be a numeric value, date. character etc. Tuple is different from List because it cannot be edited but we can edit a List. We can completely delete a Tuple but we cannot add or delete an element in the Tuple. Also, Tuple uses round brackets and List uses square brackets.

2. What is the main difference between a Pandas series and a single-column Data frame in Python?

Answer:

A single column Data Frame looks similar to a Pandas Series but it has rows and columns

on the other hand, Series are represented by a single column. If we want to apply a Series method on Single column of a Data frame then it needs to be converted into Series. This is the main difference between a single column Data Frame and a Pandas Series.

3. What is the command to sort an array in NumPy by the (n-1)th column?

Answer:

We can sort an array in NumPy by using argsort function. Suppose we have an array X and we want to sort the (n-1)th column of X. Then the command for sorting is as shown here

x[x [: n-2].argsort ()]

4. What is negative index in Python?

Answer:

Negative index in python is used to index starting from the last element of the list,

dictionary, set, or Tuple.

-1 refers to the last element,

-2 refers to the second last element

and so on.

Using negative index on a List to find the value of the last element of the List can be done as shown here. Suppose we have an array with values

>>> array = [15,25,35,45,55,65]

>>> array[-1]

65

Here, 65 is last element in the array. Therefore -1 is used as a negative index to find its value.

It is useful when we don't know the length of the container and want to find the value relative to the last index.

5. What are lambda functions?

Answer:

Lambda functions are used to define very short functions that have only one expression. If we use lambda functions then we don't need to define small functions with a specific name, body, and return statement. Everything can be done using lambda function in one short line of code.

For example:

(lambda x, y, z: (x/y) ** z)(4,2,2)

4

In this example, we've defined a lambda function that has three arguments and takes the division of the first two arguments (4 and 2) to the power of the third argument (2). Therefore, the result is 4.

6. Which library would you use for creating a bar plot in Python?

Answer:

Although, Matplotlib is generally used in python to create plots but these plots require lots of tuning to look professional hence we can use Seaborn library to create a bar plot because it helps to create appealing and meaningful plots with only one line of code.

7. How you can convert the below string to date-time value?

Import time

str = `01/01/2019`

datetime_value = time.strptime(str.date_format)

Answer:

We just need to change the date_format in the given code to change it to date-time value and replace it with %d/%m/%Y under quotation marks as shown below:

Import time

str = `01/01/2019`

datetime_value = time.strptime(str."%d/%m/%Y")

8. How you can fill a missing Admission Date with “02/05/2016” in the following table while reading a csv file with numpy if numpy is imported as np?

Student Name	Age	Admission Date	Fees Due
Sherlyn	20	01/06/2016	1000
Shaun	20		0
John	25	12/06/2016	0
Celina	21	08/06/2016	200

Answer:

Consider that the file name is "filename". The command to fill the missing admission date is:

filling_values = (“-”, 0, “02/05/2016” , 0)

temp = np.genfromtxt(filename, filling_values=filling_values)

9. Suppose you have a data frame df.

df = pd.DataFrame({‘Variable’:[‘X’,’Y’,’Z’,’W'],’Frequency’:[200,300,400,200]})

How can you convert “df” into a dictionary in a way that ‘Variable’ will be the key and ‘Frequency’ will be the value for each key?

Answer:

To convert “df” into a dictionary in a way that ‘Variable’ will be the key and ‘Frequency’ will be the value for each key is:

set_index(‘Variable’)[‘Frequency’].to_dict()

10. Should we use numpy arrays instead of nested Python lists? If Yes, why?

Answer:

Yes, we should use numpy arrays instead of Python lists. Because numpy arrays take less space and reading and writing items in numpy is also faster than doing the same thing in Python lists. If we use Python lists then a costly hardware would be required to deal with the data if data has more than a billion records. So, we should not use Python lists specially for big data sets.

Puzzles

1. You have two sand timers, which can show 2 minutes and 3 minutes respectively. Use both the sand timers (at a time or one after other or any other combination) and measure a time of 6 minutes.

Answer:

Start the 3 minute sand timer.

Once the 3 minute sand timer ends turn it upside down instantly and start the 2 minute sand timer at the same time. Now 3 minutes are over.

Once the 2 minute sand timer ends

Turn the 3 minute sand time upside down

Now the 3 minute sand timer will have 1 minute left and 5 minutes are over because 2 minute sand timer is ended as well.

Now once the 3 minute sand timer will end.

We will have measured 6 minutes of time.

2. There are 25 horses, how many races are needed to find the fastest 3 horses? You don't have a stop-watch and you can race 5 of the horses simultaneously.

Answer:

First do 5 races each with 5 horses group and identify all the winners.

Race 1: 1a, 2a, 3a, 4a, 5a

Race 2: 1b, 2b, 3b, 4b, 5b

Race 3: 1c, 2c, 3c, 4c, 5c

Race 4: 1d, 2d, 3d, 4d, 5d

Race 5: 1e, 2e, 3e, 4e, 5e

Consider the results as follows:

Race 1 Result : 1a(winner), 2a, 3a, 4a, 5a

Race 2 Result : 1b(winner), 2b, 3b, 4b, 5b

Race 3 Result : 1c(winner), 2c, 3c, 4c, 5c

Race 4 Result : 1d(winner), 2d, 3d, 4d, 5d

Race 5 Result: 1e(winner), 2e, 3e, 4e, 5e

Now do the sixth race among the winners of first five races. The winner of this race is the fastest horse.

Now for 2nd and 3rd fastest horses we have the following options:

Option 1: 2a, 3a, (if they are faster than 1b)

Option 2: 2a, 1b, (2a faster than 1b, 1b is faster than 3a)

Option 3: 1b, 2a (1b faster than 2a and 3a, 2a is faster than 1c)

Option 4: 1b, 2b (1b is faster than 2a and 3a, 2b is faster than 2a, 3a and 1c)

Option 5: 1b, 1c (1b is faster than 2a and 3a, 1c is faster than 2a, 3a, 2b and 3b)

The horses found in these options are 2a, 3a, 1b, 2b, and 3a. Therefore, the next race will be among 2a, 3a, 1b, 2b, and 3a. Irrespective of the result of this race, we will find the second and third fastest horse. Hence, 7 races are required to find the fastest three horses.

3. How can you divide a cake into 8 equal pieces in exactly 3 cuts?

Answer:

To divide a cake into 8 equal parts we have to slice the cake horizontally through the center. After that cut it vertically through the centre. Now Again cut the cake vertically through the centre but this cut must also cut the previous vertical cut into half. In this way we will get 8 equal parts.

4. One of the three switches in a room assigned for a bulb in the next room. You cannot see whether the bulb in next room is on or off, until you enter that room. What is the minimum number of times you have to go into the next room to identify which switch corresponds to the bulb in that room?

Answer:

The steps to find the minimum number are:

1. On the first switch for 3 to 4 minutes.

2. Switch off the first switch and quickly on the second switch.

3. Go to the next room and check of the bulb is glowing or not, if it is glowing switch 2 corresponds to bulb in that room, if it is not glowing then touch the bulb if it is warm that means the first switch corresponds to the bulb and if neither the bulb is on nor it is warm then third switch corresponds to the bulb in the next room. Therefore, it can be done by entering the room only one time.

5. There are 100 black socks and 100 white socks mixed up in a drawer. You must pick socks blindly from the drawer. How many socks you need to take out to be sure of having a matching pair?

Answer:

Here, the color of the socks is not necessary hence we can pick a matching pair of any color. At least 3 socks will be required to find a pair because if first and second are not of the same color then we need to pick the third sock and it will surely match with one of the first two.

Guesstimation

How to answer Guesstimation questions:

A lot of people ask that how to answer Guesstimation questions. At the time of the interview, you will not get much time to answer guesstimates. So, the interviewer expects that you are good with numbers. The most important thing you need to understand is that the interviewer is not looking for a correct answer. What they are looking for is a closest value to the actual estimate and they might ask you how you found that value. The best way to prepare for Guesstimation questions is:

1. Study about the company business you're to going to give an interview for. For example, if they are a taxi service provider like uber then they are more likely to ask you guesstimates related to taxi services such as number of cars on the road, number of taxi passengers in a day, number of car accidents etc. Now once you clearly understand the business, the second step is:

2. Search on Google about an approximate value of the whole population that is targeted by the company and round off that value. For example, after searching on Google you get an estimate of 1.32 million then round it to 1 million.

3. After this step start making segments, the number of segments should not be very large because this will take much of your time and it would be difficult to explain how you found the estimate. You should find the simplest way to find the estimated value. For example, to find the number of taxi passengers you can make segments by using age groups and give a percentage to each group then use the whole population and calculate the final estimate.

Any calculations you do to get your estimate, just make sure that you round off the values because if you don't round off. you will run out of time.

1. How many cars are there on New York City roads?

Answer:

Consider that the population of New York City is 100 lakhs, and 50% people use cars, rest of the people uses bus, metro etc. Therefore, we are left with 50 lakhs people. Now suppose that each family in New York City have an average of 5 people therefore there are 10 lakh families in New York City where car is used whether own car, yellow taxi, an uber or any other taxi service. Now suppose that among these 10 lakh families 60% own a car and say 10% of the families have at least 2 cars, therefore the total number of cars are 6 lakhs + 60000 = 6 lakhs 60 thousand.

Rest of the families uses taxi services; therefore, we have 4 lakhs more cars in New York City. Hence the total cars are 10 lakhs and 60 thousand. Now assume that 50% of these people do not travel by car on daily basis. Therefore, we are left with 5 lakhs and 30 thousand. That means 5 lakhs and 30 thousand cars are used on daily basis in New York City. Now if it is office or college starting time or closing time then we can say that there are 5 lakh cars on the roads of the New York. And if it is mid-day time then we can reduce it to 50% and say that there are 2.5 lakh car on the roads of the New York.

2. Can you estimate the total number of English-Speaking Learners in India?

Answer:

Since it talks about India, so the first value we can consider here is India's population that is 1.3 billion. Round it to 1 to make further calculations easy. Now we need to make segments on the basis of education, so the next thing is literacy rate and it is around 70%. Now our population decreased to 70% of 1 billion that is 0.7 billion. There are many English-speaking people in India because almost 50% Indian schools use English language for teaching and students from those schools generally don't take speaking classes. Hence, we can reduce the population to half and the resulting value is 0.35 billion. Round it to 0.4. Now we estimated English Speaking people in India but we want to know English Speaking learners so we are left with 0.4. Among 0.4 consider that 50% are babies, old people, and working people so we are left with 0.2. Again, consider among these 0.2 only 20% are English Speaking learners therefore the target population is 0.04 billion or 40 million. It is a good estimate because India's official language is English and Hindi.

3. Can you estimate the total number of IELTs applications from Germany per year?

Answer:

Germany is a small country assume that the population of Germany is 10 crores and the literacy in Germany is very high so we can expect that almost everyone is a literate. Since most people in Germany are patriotic consider that only 1% people want to leave their country therefore, we are left with 10 lakhs people who want to leave Germany. Again, among these 10-lakh people let's say only 30% are eligible for IELTS. So, we are left with 3 lakh people. We know that IELTS is not an easy exam and many people are afraid of it hence again we can remove 50% people due to difficulty of IELTS exam. Now we have 1.5 lakhs people. Again, consider that most people think about giving IELTS but they do not appear because they are not prepared or they are not interested in giving IELTS in a particular year. Now let's say that 1 lakh people are not prepared or interested hence we are left with 0.5 lakh people or 50 thousand people in Germany who appear in IELTS per year.

4. How many lipsticks are sold on a weekend in Dubai?

Answer:

This can be solved by using the below formula:

$$Pop\ Dubai \times \%Female \times \%Shopping\ on\ a\ weekend\ day \times \%Shopping\ of\ Lipsticks$$

$$30 \times 0.25 \times 0.60 \times 0.01 = 4500$$

Consider that the population of Dubai is 30 lakhs now assume that there are 25% female because most of the people who live in Dubai are from other countries and they are male. Now suppose that 60% females go for shopping on a weekend and 1% buys lipsticks. Hence, we get a value of 4500 that means 4500 lipsticks are sold on a weekend in Dubai.

5. Can you estimate the number of pizza deliveries per day in Paris?

Answer:

Consider that the population of Paris is 30 lakhs because it is not a very big city. Let's say 80% of people in Paris eat pizza. Therefore, there are 30 x 0.80 = 24 lakh pizza eaters in Paris. Let's say 60% pizza orders come from high income group, 30% pizza orders come from middle income group, and 5% come from low income group. Now Consider that 1 in 20 person order pizza on daily basis from high income group, 1 in 50 person order pizza on daily basis from middle income group, and 1 in 80 person order pizza on daily basis.

Therefore, the estimate will be:

$$Daily\ Pizza\ Deliveries = \left(0.60 \times \frac{1}{20} + 0.30 \times \frac{1}{50} + 0.05 \times \frac{1}{80}\right) \times 24$$

$$= \boxed{87900}$$

Hence 87900 people order pizza on daily basis in Paris.

Tricky Questions

1. A census in New York found that about 50% of households consisted of just one person. A TV Channel reported this as "Half the residents of New York live alone". Do you agree?

Answer:

I do not agree with TV channel as they wrongly interpreted census results. Census results found that there are 50% households consisting of just one person that means among 100% households, 50% households contain single person. Hence, if the remaining 50% households are taken into consideration, where each of these households would have more than one person, then it led us to the possibility that the number of residents living in these remaining 50% households is higher than the number of residents in those 50% households consisting of just one person.

Therefore, it can't be possible that half the residents of New York live alone.

2. If 4 clothes can dry in 4 hours, how many hours will be required to dry 8 clothes?

Answer:

If 4 clothes can dry in 4 hours then 8 clothes will also dry in 4 hours because dryness of clothes does not depend on the number of clothes it depends on the time and it will be same for any value.

3. What is the largest 4 digit number that will be divisible by 4444?

Answer:

The largest 4 digit number that will be divisible by 4444 is 8888 because 4444 is itself a four digit number and any number greater than 8888 and less than 10000 cannot be perfectly divisible by 4444.

4. Write down "Eleven Thousand Eleven Hundred and Eleven" in numbers?

Answer:

The value of Eleven Thousand Eleven Hundred and Eleven in numbers can be found by adding them as:

11000+1100+11 = 12111

Hence it is equal to 12111.

5. There are 1000 different students with different heights who are arranged in an array of 10 rows and 100 columns according to their respective heights. Now, the tallest from each row is called out and the shortest one of them is marked as Short. They are asked to go back to their respective position. Then, the shortest from each column is called out and the longest one among them is marked as Tall. Can you find out who between Short and Tall is taller?

Answer:

Both Short and Tall are of the same height because they are the same person.

Suppose that each person is marked with their respective number which also denotes their height and the arrangement starts from rows because it is an array of 10 rows and 100 columns Now when the longest from each row is called out, they will be numbered as 100, 200 ... 1000 and Shortest among them will be 100 also known as Short. When the shortest from each column is called out, they will be 1, 2, 3, ...100 and Longest among them will be 100 also known as Tall. Thus, they both are same person.

Quick Math Questions

1. What is 32% of 33?

Answer:

This can be easily solved by breaking it into parts. We have to calculate 32%, so we can break it into 30% and 2%. Now find 30% of 33 and 2% of 33 then add those percentages.

(30% + 02%) of 33 = 30% of 33 + 02% of 33

= 990/100+ 66/100

= 9.9+6.6

= 10.56

We can also multiply 32 and 33 then divide the multiplication with 100 but if the percentage and number are large then multiplication will take more time. Therefore, breaking the percentage into parts is the easiest way to solve this kind of problems.

2. What is the minimum number of six-pack of coke bottles will be required to fill an order of 99 bottles?

Answer:

To find the number of packs we should divide 99 by 6 and we will get 16 as a quotient and 3 as a remainder that means we have 16 six packs of coke bottles and three more bottles need to be added in 99 so that the number of bottles becomes completely divisible by 6.

Therefore, we will add one more pack to 16 packs and conclude that minimum 17 packs will be required to fill an order of 99 bottles.

3. What is half of 10+10?

Answer:

To solve this question usually people take the sum of two ten's and divide it by 2 which is a wrong approach Because according to BODMAS rule half of 10 should be calculated first then 10 will be added to the resulting value.

Wrong Approach: $(10+10)\times\frac{1}{2}$

Correct Approach: $10\times\frac{1}{2}+10=15$

4. How can you make 1000 by using 8 exactly eight times?

Answer:

We can make 1000 by adding 888, 88, and three 8s as:

$$888+88+8+8+8=1000$$

5. If we divide 100 by half and add 100. What will be the result?

Answer:

To solve this question people multiply 1/2 with 100 which is a wrong Approach. Here we need to divide 100 with 1/2 or for simplicity we can use 0.5 because 0.5 is equal to 1/2. So, we get 200 by dividing 100 with ½ and then we add 100 to 200 therefore the result is 300.

Wrong Approach: $100\times\frac{1}{2}$

Correct Approach: $\frac{100}{\frac{1}{2}}=\frac{100\times 2}{1}=200$

Now add 100: $200+100=300$