KASPER
ANALYTICS

kasperanalytics.com

9311386511

Data Analyst Interview QnA

hr@kasperanalytics.com

kasper-analytics

Q1. What is the difference between Data Mining and Data Analysis?

Data Mining	Data Analysis
Used to recognize patterns in data stored.	Used to order & organize raw data in a meaningful manner.
Mining is performed on clean and well-documented data.	The analysis of data involves Data Cleaning. So, data is not present in a well-documented format.
Results extracted from data mining are not easy to interpret.	Results extracted from data analysis are easy to interpret.

Table 1: Data Mining vs Data Analysis – Data Analyst Interview Questions

So, if you have to summarize, Data Mining is often used to identify patterns in the data stored. It is mostly used for Machine Learning, and analysts have to just recognize the patterns with the help of algorithms. Whereas, Data Analysis is used to gather insights from raw data, which has to be cleaned and organized before performing the analysis.

Q2. What is the process of Data Analysis?

Data analysis is the process of collecting, cleansing, interpreting, transforming and modeling data to gather insights and generate reports to gain business profits. Refer to the image below to know the various steps involved in the process.

Fig 1: Process of Data Analysis – Data Analyst Interview Questions

- **Collect Data**: The data gets collected from various sources and is stored so that it can be cleaned and prepared. In this step, all the missing values and outliers are removed.
- **Analyse Data**: Once the data is ready, the next step is to analyze the data. A model is run repeatedly for improvements. Then, the mode is validated to check whether it meets the business requirements.
- **Create Reports**: Finally, the model is implemented and then reports thus generated are passed onto the stakeholders.

Q3. What is the difference between Data Mining and Data Profiling?

Data Mining: Data Mining refers to the analysis of data with respect to finding relations that have not been discovered earlier. It mainly focuses on the detection of unusual records, dependencies and cluster analysis.

Data Profiling: Data Profiling refers to the process of analyzing individual attributes of data. It mainly focuses on providing valuable information on data attributes such as data type, frequency etc.

Q4. What is data cleansing and what are the best ways to practice data cleansing?

Data Cleansing or Wrangling or Data Cleaning. All mean the same thing. It is the process of identifying and removing errors to enhance the quality of data. You can refer to the below image to know the various ways to deal with missing data.

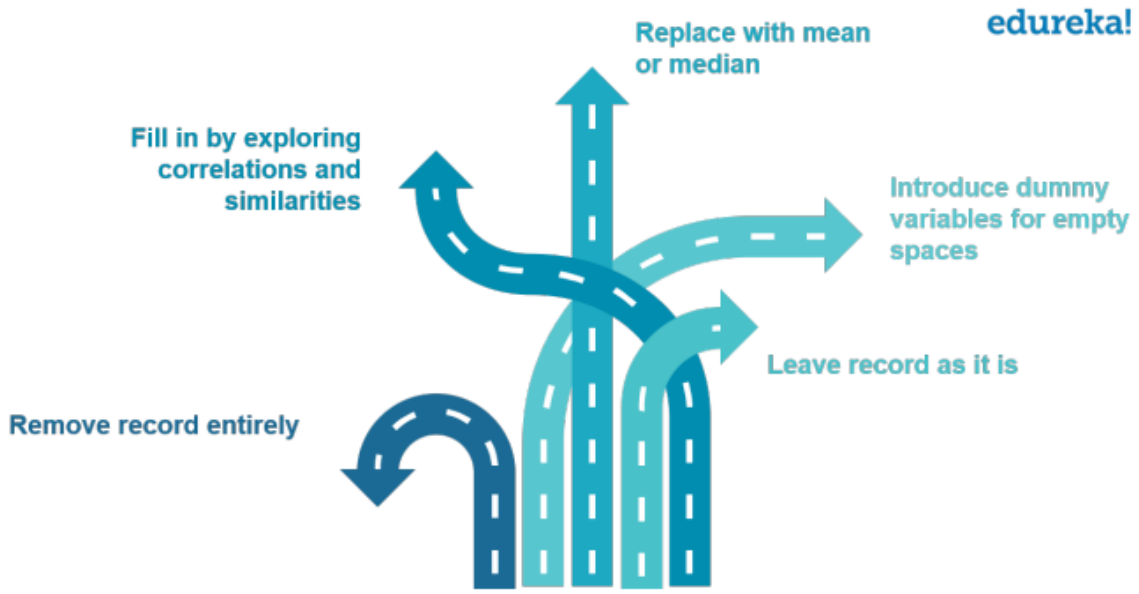


Fig 2: Ways of Data Cleansing – Data Analyst Interview Questions

Q5. What are the important steps in the data validation process?

As the name suggests Data Validation is the process of validating data. This step mainly has two processes involved in it. These are Data Screening and Data Verification.

- **Data Screening**: Different kinds of algorithms are used in this step to screen the entire data to find out any inaccurate values.
- **Data Verification**: Each and every suspected value is evaluated on various use-cases, and then a final decision is taken on whether the value has to be included in the data or not.

Q6. What do you think are the criteria to say whether a developed data model is good or not?

Well, the answer to this question may vary from person to person. But below are a few criteria which I think are a must to be considered to decide whether a developed data model is good or not:

- A model developed for the dataset should have predictable performance. This is required to predict the future.
- A model is said to be a good model if it can easily adapt to changes according to business requirements.
- If the data gets changed, the model should be able to scale according to the data.
- The model developed should also be able to easily consumed by the clients for actionable and profitable results.

Q7. When do you think you should retrain a model? Is it dependent on the data?

Business data keeps changing on a day-to-day basis, but the format doesn't change. As and when a business operation enters a new market, sees a sudden rise of opposition or sees its own position rising or falling, it is recommended to retrain the model. So, as and when the business dynamics change, it is recommended to retrain the model with the changing behaviors of customers.

Q8. Can you mention a few problems that data analyst usually encounter while performing the analysis?

The following are a few problems that are usually encountered while performing data analysis.

- Presence of Duplicate entries and spelling mistakes, reduce data quality.
- If you are extracting data from a poor source, then this could be a problem as you would have to spend a lot of time cleaning the data.
- When you extract data from sources, the data may vary in representation. Now, when you combine data from these sources, it may happen that the variation in representation could result in a delay.
- Lastly, if there is incomplete data, then that could be a problem to perform analysis of data.

Q9. What is the KNN imputation method?

This method is used to impute the missing attribute values which are imputed by the attribute values that are most similar to the attribute whose values are missing. The similarity of the two attributes is determined by using the distance functions.

Q10. Mention the name of the framework developed by Apache for processing large dataset for an application in a distributed computing environment?

The complete Hadoop Ecosystem was developed for processing large dataset for an application in a distributed computing environment. The Hadoop Ecosystem consists of the following Hadoop components.

- HDFS -> Hadoop Distributed File System
- YARN -> Yet Another Resource Negotiator
- MapReduce -> Data processing using programming
- Spark -> In-memory Data Processing
- PIG, HIVE-> Data Processing Services using Query (SQL-like)
- HBase -> NoSQL Database
- Mahout, Spark MLlib -> Machine Learning
- Apache Drill -> SQL on Hadoop
- Zookeeper -> Managing Cluster
- Oozie -> Job Scheduling
- Flume, Sqoop -> Data Ingesting Services
- Solr & Lucene -> Searching & Indexing
- Ambari -> Provision, Monitor and Maintain cluster

Now, moving on to the next set of questions, which is the Excel Interview Questions.

Microsoft Excel is one of the simplest and most powerful software applications available out there. It lets users do quantitative analysis, statistical analysis with an intuitive interface for data manipulation, so much so that its usage spans across different domains and professional requirements. This is an important field that gives a head-start for becoming a Data Analyst. So, now let us quickly discuss the questions asked with respect to this topic.

Q1. Can you tell what is a waterfall chart and when do we use it?

The waterfall chart shows both positive and negative values which lead to the final result value. For example, if you are analyzing a company's net income, then you can have all the cost values in this chart. With such kind of a chart, you can visually, see how the value from revenue to the net income is obtained when all the costs are deducted.

Q2. How can you highlight cells with negative values in Excel?

You can highlight cells with negative values in Excel by using the conditional formatting. Below are the steps that you can follow:

- Select the cells which you want to highlight with the negative values.
- Go to the Home tab and click on the Conditional Formatting option
- Go to the Highlight Cell Rules and click on the Less Than option.
- In the dialog box of Less Than, specify the value as 0.

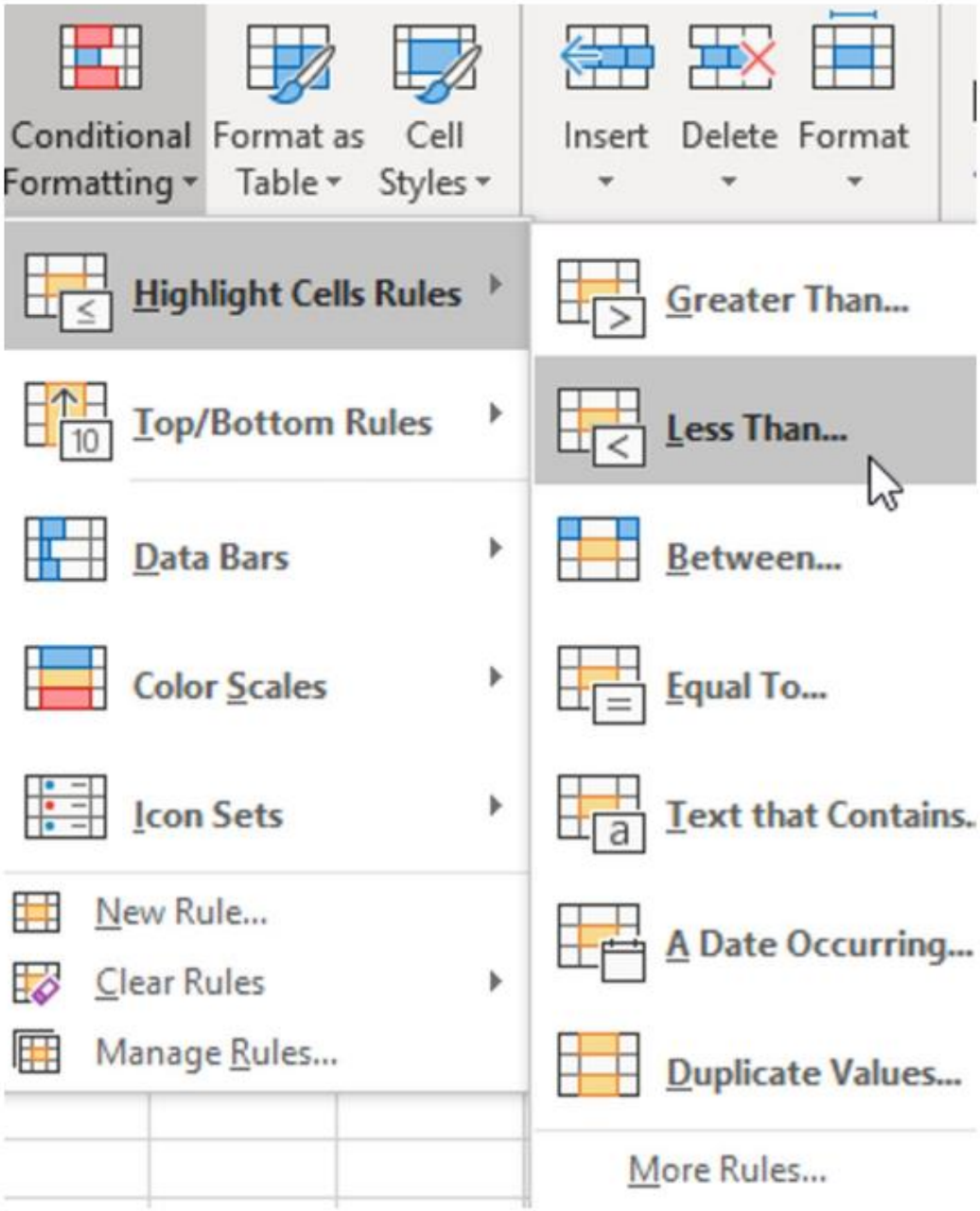


Fig 3: Snapshot of Highlighting cells in Excel – Data Analyst Interview Questions

Q3. How can you clear all the formatting without actually removing the cell contents?

Sometimes you may want to remove all the formatting and just want to have the basic/simple data. To do this, you can use the 'Clear Formats' options found in the

Home Tab. You can evidently see the option when you click on the 'Clear' drop down.

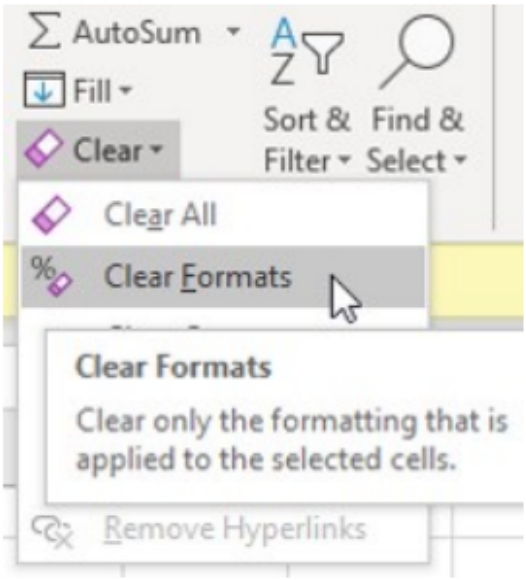


Fig 4: Snapshot of clearing all formatting in Excel – Data Analyst Interview Questions

Q4. What is a Pivot Table, and what are the different sections of a Pivot Table?

A Pivot Table is a simple feature in Microsoft Excel which allows you to quickly summarize huge datasets. It is really easy to use as it requires dragging and dropping rows/columns headers to create reports.

A Pivot table is made up of four different sections:

- **Values Area**: Values are reported in this area
- **Rows Area**: The headings which are present on the left of the values.

- **Column Area**: The headings at the top of the values area makes the columns area.
- **Filter Area**: This is an optional filter used to drill down in the data set.

Q5. Can you make a Pivot Table from multiple tables?

Yes, we can create one Pivot Table from multiple different tables when there is a connection between these tables.

Q6. How can we select all blank cells in Excel?

If you wish to select all the blank cells in Excel, then you can use the Go To Special Dialog Box in Excel. Below are the steps that you can follow to select all the blank cells in Excel.

- First, select the entire dataset and press F5. This will open a Go To Dialog Box.
- Click the 'Special' button which will open a Go To special Dialog box.
- After that, select the Blanks and click on OK.

The final step will select all the blank cells in your dataset.

Q7. What are the most common questions you should ask a client before creating a dashboard?

Well, the answer to this question varies on a case-to-case basis. But, here are a few common questions that you can ask while creating a dashboard in Excel.

- Purpose of the Dashboards
- Different data sources
- Usage of the Excel Dashboard
- The frequency at which the dashboard needs to be updated
- The version of Office the client uses.

Q8. What is a Print Area and how can you set it in Excel?

A Print Area in Excel is a range of cells that you designate to print whenever you print that worksheet. For example, if you just want to print the first 20 rows from the entire worksheet, then you can set the first 20 rows as the Print Area.

Now, to set the Print Area in Excel, you can follow the below steps:

- Select the cells for which you want to set the Print Area.
- Then, click on the Page Layout Tab.
- Click on Print Area.
- Click on Set Print Area.

Q9. What steps can you take to handle slow Excel workbooks?

Well, there are various ways to handle slow Excel workbooks. But, here are a few ways in which you can handle workbooks.

- Try using manual calculation mode.
- Maintain all the referenced data in a single sheet.
- Often use excel tables and named ranges.
- Use Helper columns instead of array formulas.
- Try to avoid using entire rows or columns in references.
- Convert all the unused formulas to values.

Q10. Can you sort multiple columns at one time?

Multiple sorting refers to the sorting of a column and then sorting the other column by keeping the first column intact. In Excel, you can definitely sort multiple columns at a one time.

To do multiple sorting, you need to use the Sort Dialog Box. Now, to get this, you can select the data that you want to sort and then click on the Data Tab. After that, click on the Sort icon. In this Dialog box, you can specify the details for one column, and then sort to another column, by clicking on the Add Level button.

Moving onto the next set of questions, which is questions asked related to Statistics.

Statistics is a branch of mathematics dealing with data collection and organization, analysis, interpretation, and presentation. Statistics can be divided into two categories: Differential and Inferential Statistics. This field is related to mathematics and thus gives a kickstart to Data Analysis career.

Q1. What do you understand by the term Normal Distribution?

This is one of the most important and widely used distributions in statistics. Commonly known as the Bell Curve or Gaussian curve, normal distributions, measure how much values can differ in their means and in their standard deviations. Refer to the below image.

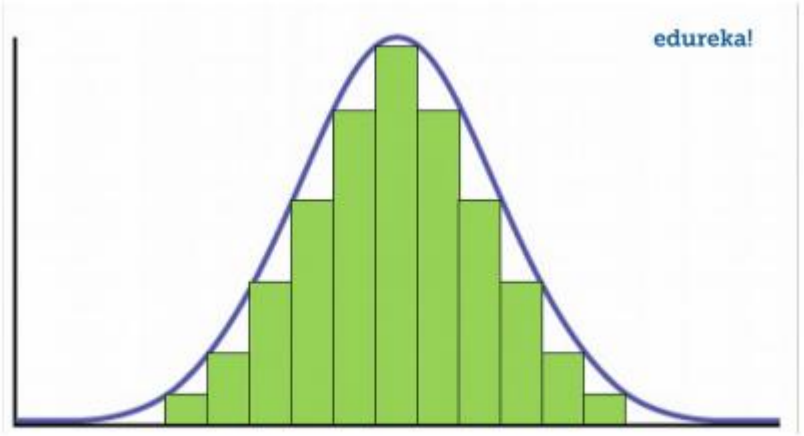


Fig 5: Normal Distribution – Data Analyst Interview Questions

As you can see in the above image, data is usually distributed around a central value without any bias to the left or right side. Also, the random variables are distributed in the form of a symmetrical bell-shaped curve.

Q2. What is A/B Testing?

A/B testing is the statistical hypothesis testing for a randomized experiment with two variables A and B. Also known as the split testing, it is an analytical method that estimates population parameters based on sample statistics. This test compares two web pages by showing two variants A and B, to a similar number of visitors, and the variant which gives better conversion rate wins.

The goal of A/B Testing is to identify if there are any changes to the web page. For example, if you have a banner ad on which you have spent an ample amount of money. Then, you can find out the return of investment i.e. the click rate through the banner ad.

Q3. What is the statistical power of sensitivity?

The statistical power of sensitivity is used to validate the accuracy of a classifier. This classifier can be either Logistic Regression, Support Vector Machine, Random Forest etc.

If I have to define sensitivity, then sensitivity is nothing but the ratio of Predicted True Events to Total Events. Now, True Events are the events which were true and the model also predicts them as true.

$$\text{Seasonality} = \frac{\text{True Positives}}{\text{Positives in Actual Dependent Variable}}$$

Fig 6: Seasonality Formula – Data Analyst Interview Questions

Q4. What is the Alternative Hypothesis?

To explain the Alternative Hypothesis, you can first explain what the null hypothesis is. Null Hypothesis is a statistical phenomenon that is used to test for possible rejection under the assumption that result of chance would be true.

After this, you can say that the alternative hypothesis is again a statistical phenomenon which is contrary to the Null Hypothesis. Usually, it is considered that the observations are a result of an effect with some chance of variation.

Q5. What is the difference between univariate, bivariate and multivariate analysis?

The differences between univariate, bivariate and multivariate analysis are as follows:

- **Univariate**: A descriptive statistical technique that can be differentiated based on the count of variables involved at a given instance of time.
- **Bivariate**: This analysis is used to find the difference between two variables at a time.
- **Multivariate**: The study of more than two variables is nothing but multivariate analysis. This analysis is used to understand the effect of variables on the responses.

Q6. Can you tell me what are Eigenvectors and Eigenvalues?

Eigenvectors: Eigenvectors are basically used to understand linear transformations. These are calculated for a correlation or a covariance matrix.

For definition purposes, you can say that Eigenvectors are the directions along which a specific linear transformation acts either by flipping, compressing or stretching.

Eigenvalue: Eigenvalues can be referred to as the strength of the transformation or the factor by which the compression occurs in the direction of eigenvectors.

Q7. What is the difference between 1-Sample T-test, and 2-Sample T- test?

You can answer this question, by first explaining, what exactly T-tests are. Refer below for an explanation of T-Test.

T-Tests are a type of hypothesis tests, by which you can compare means. Each test that you perform on your sample data, brings down your sample data to a single value i.e. T-value. Refer below for the formula.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Fig 7: Formula to calculate t-value – Data Analyst Interview Questions

Now, to explain this formula, you can use the analogy of the signal-to-noise ratio, since the formula is in a ratio format.

Here, the numerator would be a signal and the denominator would be the noise.

So, to calculate 1-Sample T-test, you have to subtract the null hypothesis value from the sample mean. If your sample mean is equal to 7 and the null hypothesis value is 2, then the signal would be equal to 5.

So, we can say that the difference between the sample mean and the null hypothesis is directly proportional to the strength of the signal.

Now, if you observe the denominator which is the noise, in our case it is the measure of variability known as the standard error of the mean. So, this basically indicates how accurately your sample estimates the mean of the population or your complete dataset.

So, you can consider that noise is indirectly proportional to the precision of the sample.

Now, the ratio between the signal-to-noise is how you can calculate the T-Test 1. So, you can see how distinguishable your signal is from the noise.

To calculate, 2-Sample Test, you need to find out the ratio between the difference of the two samples to the null hypothesis.

So, if I have to summarize for you, the 1-Sample T-test determines how a sample set holds against a mean, while the 2-Sample T-test determines if the mean

between 2 sample sets is really significant for the entire population or purely by chance.

Q8. What are different types of Hypothesis Testing?

The different types of hypothesis testing are as follows:

- T-test: T-test is used when the standard deviation is unknown and the sample size is comparatively small.
- Chi-Square Test for Independence: These tests are used to find out the significance of the association between categorical variables in the population sample.
- Analysis of Variance (ANOVA): This kind of hypothesis testing is used to
- analyze differences between the means in various groups. This test is often used similarly to a T-test but, is used for more than two groups.
- Welch's T-test: This test is used to find out the test for equality of means between two population samples.

Q9. How to represent a Bayesian Network in the form of Markov Random Fields (MRF)?

To represent a Bayesian Network in the form of Markov Random Fields, you can consider the following examples:

Consider two variables which are connected through an edge in a Bayesian network, then we can have a probability distribution that factorizes into a probability of A and then the probability of B. Whereas, the same network if we mention in Markov Random Field, it would be represented as a single potential function. Refer below:

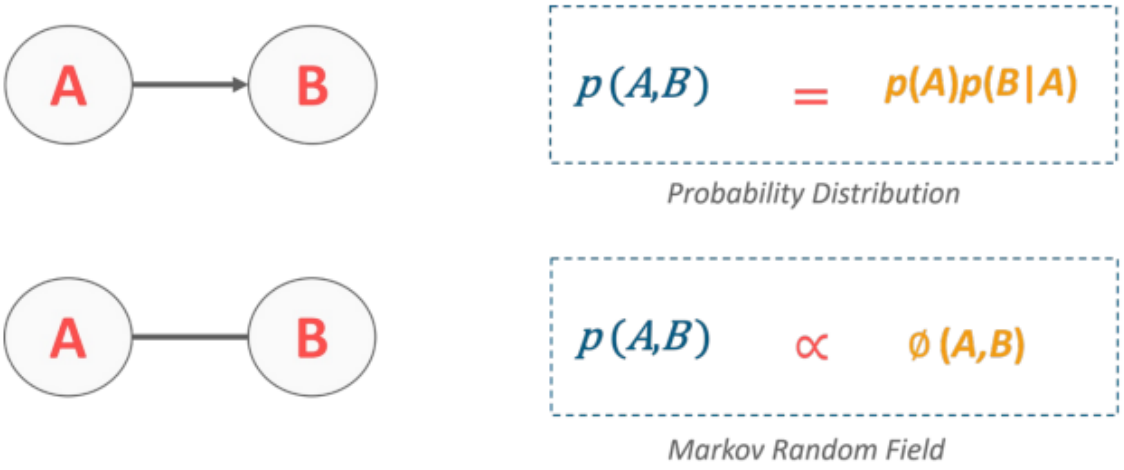


Fig 7: Representation of Bayesian Network in MRF – Data Analyst Interview Questions

Well, that was a simple example to start with. Now, moving onto a complex example where one variable is a parent of the other two. Here A is the parent variable and it points down to B and C. In such a case, the probability distribution would be equal to the probability of A and the conditional probability of B given A and C given A. Now, if you have to convert this into Markov Random Field, the factorization of the similarly structured graph, where we have the potential function of A/B edge and a potential function for A/C edge. Refer to the image below.

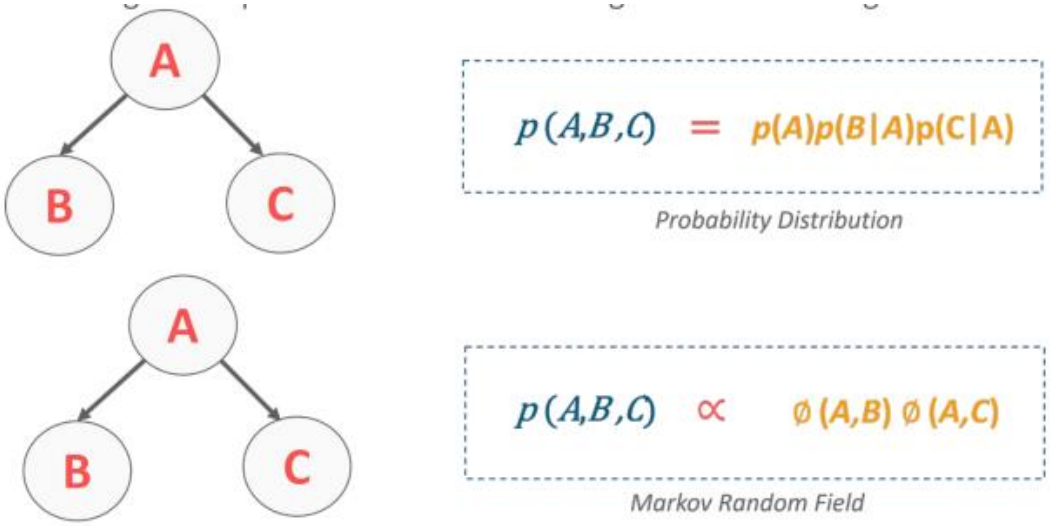


Fig 8: Representation of Bayesian Network in MRF – Data Analyst Interview Questions

Q10. What is the difference between variance and covariance?

Variance and Covariance are two mathematical terms which are used frequently in statistics. Variance basically refers to how apart numbers are in relation to the mean. Covariance, on the other hand, refers to how two random variables will change together. This is basically used to calculate the correlation between variables.

In case you have attended any Data Analytics interview in the recent past, do paste those interview questions in the comments section and we'll answer them ASAP. You can also comment below if you have any questions in your mind, which you might have faced in your Data Analytics interview.

Now, let us move on to the next set of questions which is the SAS Interview Questions.

Statistical Analysis System(SAS) provided by SAS Institute itself is the most popular Data Analytics tool in the market. In simple words, SAS can process complex data and generate meaningful insights that would help organizations make better decisions or predict possible outcomes in the near future. So, this lets you mine, alter, manage and retrieve data from different sources and analyze it.

Q1. What is interleaving in SAS?

Interleaving in SAS means combining individual sorted SAS data sets into one sorted data set. You can interleave data sets using a SET statement along with a BY statement.

In the example that you can see below, the data sets are sorted by the variable Age.

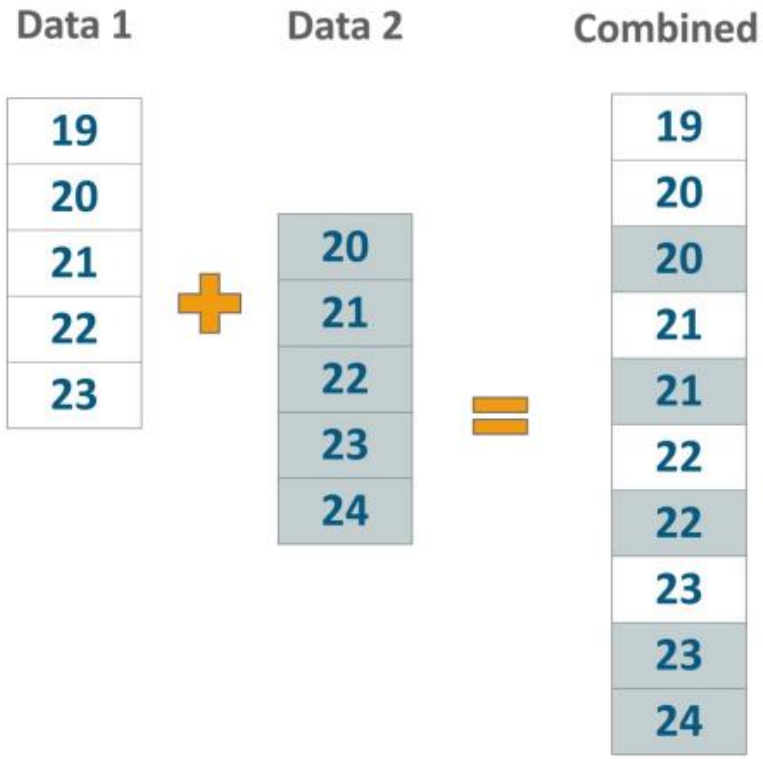


Fig 9: Example for Interleaving in SAS – Data Analyst Interview Questions

Q2. What is the basic syntax style of writing code in SAS?

The basic syntax style of writing code in SAS is as follows:

1. Write the DATA statement which will basically name the dataset.
2. Write the INPUT statement to name the variables in the data set.
3. All the statements should end with a semi-colon.
4. There should be a proper space between word and a statement.

Q3. What is the difference between the Do Index, Do While and the Do Until loop? Give examples.

To answer this question, you can first answer what exactly a Do loop is. So, a Do loop is used to execute a block of code repeatedly, based on a condition. You can refer to the image below to see the workflow of the Do loop.

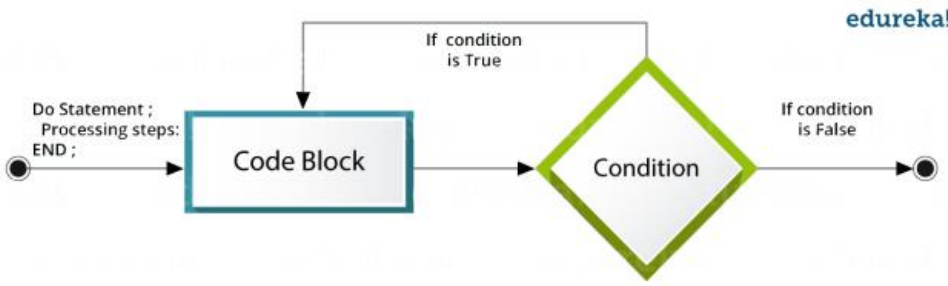


Fig 10: Workflow of Do Loop – Data Analyst Interview Questions

- Do Index loop: We use an index variable as a start and stop value for Do Index loop. The SAS statements get executed repeatedly till the index variable reaches its final value.
- **Do While Loop:** The Do While loop uses a WHILE condition. This Loop executes the block of code when the condition is true and keeps executing it, till the condition becomes false. Once the condition becomes false, the loop is terminated.
- **Do Until Loop:** The Do Until loop uses an Until condition. This Loop executes the block of code when the condition is false and keeps executing it, till the condition becomes true. Once the condition becomes true, the loop is terminated.

If you have to explain with respect to the code, then let us say we want to calculate the SUM and the number of variables.

For the loops you can write the code as follows:

Do Index

```
DATA ExampleLoop;
SUM=0;
Do VAR = 1 = 10;
SUM = SUM + VAR;

END;
PROC PRINT DATA = ExampleLoop;
Run;
```

The output would be:

Obs	SUM	VAR
1	55	11

Table 2: Output of Do Index Loop – Data Analyst Interview Questions

Do While

```
DATA ExampleLoop;
SUM = 0;
VAR = 1;
Do While(VAR<15);
SUM = SUM + VAR;
VAR+1;
END;
PROC PRINT DATA = ExampleLoop;
Run;
```

Obs	SUM	VAR
1	105	15

Table 3: Output of Do While Loop – Data Analyst Interview Questions

Do Until

```
DATA ExampleLoop;
   SUM = 0;
   VAR = 1;
Do Until(VAR>15);
  SUM=SUM+VAR;
    VAR+1;
     END;
 PROC PRINT;
    Run;
```

Obs	SUM	VAR
1	120	16

Table 4: Output of Do Until Loop – Data Analyst Interview Questions

Q4. What is the ANYDIGIT function in SAS?

The ANYDIGIT function is used to search for a character string. After the string is found it will simply return the desired string.

Q5. Can you tell the difference between VAR X1 – X3 and VAR X1 — X3?

When you specify sing dash between the variables, then that specifies consecutively numbered variables. Similarly, if you specify the Double Dash between the variables, then that would specify all the variables available within the dataset.

For Example:

Consider the following data set:

Data Set: ID NAME X1 X2 Y1 X3

Then, X1 – X3 would return X1 X2 X3

and X1 — X3 would return X1 X2 Y1 X3

Q6. What is the purpose of trailing @ and @@? How do you use them?

The trailing @ is commonly known as the column pointer. So, when we use the trailing @, in the Input statement, it gives you the ability to read a part of the raw data line, test it and decide how can the additional data be read from the same record

- The single trailing @ tells the SAS system to "hold the line".
- The double trailing @@ tells the SAS system to "hold the line more strongly".

An Input statement ending with @@ instructs the program to release the current raw data line only when there are no data values left to be read from that line. The @@, therefore, holds the input record even across multiple iterations of the data step.

Q7. What would be the result of the following SAS function (given that 31 Dec 2017 is Saturday)?

Weeks = intck ('week','31 dec 2017'd,'01jan2018'd);

Years = intck ('year','31 dec 2017'd,'01jan2018'd);

Months = intck ('month','31 dec 2017'd,'01jan2018'd);

Here, we will calculate the weeks between 31st December 2017 and 1st January 2018. 31st December 2017 was a Saturday. So 1st January 2018 will be a Sunday in the next week.

- Hence, Weeks = 1 since both the days are in different weeks.
- Years = 1 since both the days are in different calendar years.
- Months = 1 since both the days are in different months of the calendar.

Q8. How does PROC SQL work?

PROC SQL is nothing but a simultaneous process for all the observations. The following steps occur when a PROC SQL gets executed:

- SAS scans each and every statement in the SQL procedure and checks the syntax errors.
- The SQL optimizer scans the query inside the statement. So, the SQL optimizer basically decides how the SQL query should be executed in order to minimize the runtime.
- If there are any tables in the FROM statement, then they are loaded into the data engine where they can then be accessed in the memory.
- Codes and Calculations are executed.
- The Final Table is created in the memory.
- The Final Table is sent to the output table described in the SQL statement.

Q9. If you are given an unsorted data set, how will you read the last observation to a new dataset?

We can read the last observation to a new dataset using end = dataset option. For example:

```
1    data example.newdataset;
2  set example.olddataset end=last;
3       If last;
4         run;
```

Where newdataset is a new data set to be created and olddataset is the existing data set. last is the temporary variable (initialized to 0) which is set to 1 when the set statement reads the last observation.

Q10. What are the differences between the sum function and using "+" operator?

The SUM function returns the sum of non-missing arguments whereas "+" operator returns a missing value if any of the arguments are missing. Consider the following example.

Example:

```
data exampledata1;
   input a b c;
      cards;
      44 4 4
      34 3 4
      34 3 4
       . 1 2
      24 . 4
      44 4 .
      25 3 1
        ;
       run;
data exampledata2;
 set exampledata1;
  x = sum(a,b,c);
     y=a+b+c;
       run;
```

In the output, the value of y is missing for 4th, 5th, and 6th observation as we have used the "+" operator to calculate the value of y.

x y

52 52

41 41

41 41

3 .

28 .

48 .

29 29

Now, let us move on to the next set of questions which is the SQL Interview Questions.

RDBMS is one of the most commonly used databases till date, and therefore SQL skills are indispensable in most of the job roles such as a Data Analyst. Knowing Structured Query Language, boots your path on becoming a data analyst, as it will be clear in your interviews that you know how to handle databases.

Q1. What is the default port for SQL?

The default TCP port assigned by the official Internet Number Authority(IANA) for SQL server is 1433.

Q2. What do you mean by DBMS? What are its different types?

A Database Management System (DBMS) is a software application that interacts with the user, applications and the database itself to capture and analyze data. The data stored in the database can be modified, retrieved and deleted, and can be of any type like strings, numbers, images etc.

There are mainly 4 types of DBMS, which are Hierarchical, Relational, Network, and Object-Oriented DBMS.

- Hierarchical DBMS: As the name suggests, this type of DBMS has a style of predecessor-successor type of relationship. So, it has a structure similar to that of a tree, wherein the nodes represent records and the branches of the tree represent fields.
- Relational DBMS (RDBMS): This type of DBMS, uses a structure that allows the users to identify and access data in relation to another piece of data in the database.
- Network DBMS: This type of DBMS supports many to many relations wherein multiple member records can be linked.
- Object-oriented DBMS: This type of DBMS uses small individual software called objects. Each object contains a piece of data and the instructions for the actions to be done with the data.

Q3. What is ACID property in a database?

ACID is an acronym for Atomicity, Consistency, Isolation, and Durability. This property is used in the databases to ensure whether the data transactions are processed reliably in the system or not. If you have to define each of these terms, then you can refer below.

- Atomicity: Refers to the transactions which are either completely successful or failed. Here a transaction refers to a single operation. So, even if a single transaction fails, then the entire transaction fails and the database state is left unchanged.
- Consistency: This feature makes sure that the data must meet all the validation rules. So, this basically makes sure that the transaction never leaves the database without completing its state.
- Isolation: Isolation keeps transactions separated from each other until they're finished. So basically each and every transaction is independent.
- Durability: Durability makes sure that your committed transaction is never lost. So, this guarantees that the database will keep track of pending changes in such a way that even if there is a power loss, crash or any sort of error the server can recover from an abnormal termination.

Q4. What is Normalization? Explain different types of Normalization with advantages.

Normalization is the process of organizing data to avoid duplication and redundancy. There are many successive levels of normalization. These are

called normal forms. Each consecutive normal form depends on the previous one. The first three normal forms are usually adequate.

- First Normal Form (1NF) – No repeating groups within rows
- Second Normal Form (2NF) – Every non-key (supporting) column value is dependent on the whole primary key.
- Third Normal Form (3NF) – Dependent solely on the primary key and no other non-key (supporting) column value.

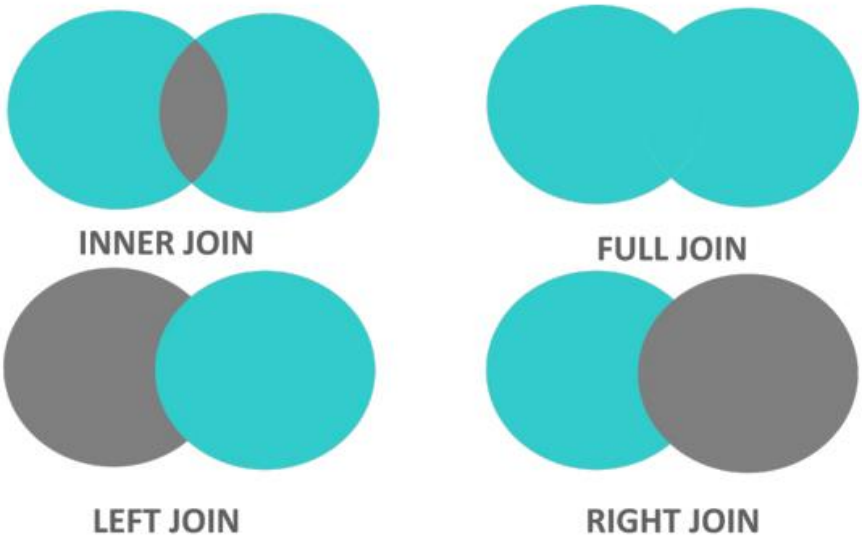
- Boyce- Codd Normal Form (BCNF) – BCNF is the advanced version of 3NF. A table is said to be in BCNF if it is 3NF and for every X ->Y, relation X should be the super key of the table.

Some of the advantages are:

- Better Database organization
- More Tables with smaller rows
- Efficient data access
- Greater Flexibility for Queries
- Quickly find the information
- Easier to implement Security
- Allows easy modification
- Reduction of redundant and duplicate data
- More Compact Database
- Ensure Consistent data after modification

Q5. What are the different types of Joins?

The various types of joins used to retrieve data between tables are Inner Join, Left Join, Right Join and Full Outer Join. Refer to the image on the right side.



- Inner join: Inner Join in MySQL is the most common type of join. It is used to return all the rows from multiple tables where the join condition is satisfied.
- Left Join: Left Join in MySQL is used to return all the rows from the left table, but only the matching rows from the right table where the join condition is fulfilled.
- Right Join: Right Join in MySQL is used to return all the rows from the right
- table, but only the matching rows from the left table where the join condition is fulfilled.
- Full Join: Full join returns all the records when there is a match in any of the tables. Therefore, it returns all the rows from the left-hand side table and all the rows from the right-hand side table.

Q6. Suppose you have a table of employee details consisting of columns names (employeeId, employeeName), and you want to fetch alternate records from a table. How do you think you can perform this task?

You can fetch alternate tuples by using the row number of the tuple. Let us say if we want to display the employeeId, of even records, then you can use the mod function and simply write the following query:

```
Select employeeId from (Select rownumber, employeeId from employee) where
mod(rownumber,2)=0
```

where '***employee***' is the table name.

Similarly, if you want to display the employeeId of odd records, then you can write the following query

```
Select employeeId from (Select rownumber, employeeId from employee) where
mod(rownumber ,2)=1
```

Q7. Consider the following two tables

Q7. Consider the following two tables.

Customer_Id	CustomerName
1	Julia
2	Alice
3	Johnathan
4	Bob
5	Stacy

Customer_Course_Id	Customer_Id	Course_Id	Course_Date
1	1	1	2018-09-03
2	2	1	2018-09-04
3	3	2	2018-09-03
4	4	2	2018-09-05
5	4	2	2018-09-04
6	2	1	2018-09-03
7	1	2	2018-09-05
8	3	3	2018-09-04
9	1	4	2018-09-04
10	3	1	2018-09-03
11	4	2	2018-09-05
12	3	2	2018-09-03
13	1	1	2018-09-03
14	2	2	2018-09-04
15	3	4	2018-09-04

Table 5: Example Table – Data Analyst Interview Questions

Now, write a query to get the list of customers who took the course more than once on the same day. The customers should be grouped by customer, and course and the list should be ordered according to the most recent date.

```
SELECT
c.Customer_Id,
CustomerName,
Course_Id,
Course_Date,
count(Customer_Course_Id) AS count
FROM customers c JOIN course_details d ON d.Customer_Id = c.Customer_Id
GROUP BY c.Customer_Id,

CustomerName,
Course_Id,
Course_Date
HAVING count( Customer_Course_Id ) > 1
ORDER BY Course_Date DESC;
```

Customer_Id	CustomerName	Course_Id	Course_Date	Count
4	Bob	2	September, 05 2018	2
3	Johnathan	2	September, 03 2018	2
1	Julia	1	September, 03 2018	2

Table 6: Output Table – Data Analyst Interview Questions

Q8. Consider the below Employee_Details table. Here the table has various features such as Employee_Id, EmployeeName, Age, Gender, and Shift. The Shift has m = Morning Shift and e = Evening Shift. Now, you have to swap the 'm' and the 'e' values and vice versa, with a single update query.

Employee_Id	EmployeeName	Age	Gender	Shift
1	Bob	23	Male	m
2	Julia	25	Female	e
3	Johnathan	31	Male	m
4	Alice	20	Female	e
5	Rick	32	Male	m
6	Stacy	27	Female	e

Table 7: Example Table – Data Analyst Interview Questions

You can write the below query:

Employee_Id	EmployeeName	Age	Gender	Shift
1	Bob	23	Male	e
2	Julia	25	Female	m
3	Johnathan	31	Male	e
4	Alice	20	Female	m
5	Rick	32	Male	e
6	Stacy	27	Female	m

Q9. Write a SQL query to get the third highest salary of an employee from Employee_Details table as illustrated below.

Employee_Id	EmployeeName	Age	Gender	Salary
1	Bob	23	Male	25000
2	Julia	25	Female	15000
3	Johnathan	31	Male	40000
4	Alice	20	Female	10000
5	Rick	32	Male	45000
6	Stacy	27	Female	27000

Table 9: Example Table – Data Analyst Interview Questions

```
SELECT TOP 1 Salary
FROM(
SELECT TOP 3 Salary
FROM Employee_Details
ORDER BY salary DESC) AS emp
ORDER BY salary ASC;
```

Q10. What is the difference between NVL and NVL2 functions in SQL?

NVL(exp1, exp2) and NVL2(exp1, exp2, exp3) are functions which check whether the value of exp1 is null or not.

If we use NVL(exp1,exp2) function, then if exp1 is not null, then the value of exp1 will be returned; else the value of exp2 will be returned. But, exp2 must be of the same data type of exp1.

Similarly, if we use NVL2(exp1, exp2, exp3) function, then if exp1 is not null, exp2 will be returned, else the value of exp3 will be returned.

If you wish to know more questions on SQL, then refer a full-fledged article on SQL Interview Questions.

Now, moving onto the next set of questions asked i.e. the Tableau Interview Questions.

Tableau is a business intelligence software which allows anyone to connect to the respective data. It visualizes and creates interactive, shareable dashboards. knowing Tableau will enhance your understanding of Data Analysis and Data Visualization.

Q1. What are the differences between Tableau and Power BI?

Parameters	Tableau	Power BI
Cost	Tableau may costs you around $1000 for a yearly subscription	$100 for a yearly subscription
Licensing	Tableau is not free	3 months trial period
Ease of use	Tableau offers variety when it comes to implementation and consulting services.	Power BI is easier to implement.
Visualization	scales better to larger datasets	Power BI it is easier to upload data sets
Year Of Establishment	2003	2013
Cost	High	Low
Application	AD-Hoc Analysis	Dashboard
Users	Analysts	Technical / Non-technical People
Support Level	High	Low
Scalability (Large Data-Sets)	Very Good	Good
Licensing	Flexible	Rigid

If you wish to learn more about the Differences between Power BI and Tableau, you can check out the following video:

Power BI vs Tableau

Q2. What is a dual axis?

Dual Axis is a phenomenon provided by Tableau. This helps the users to view two scales of two measures in the same graph. Websites such as Indeed.com make use of dual axis to show the comparison between two measures and the growth of these two measures in a septic set of years. Dual axes let you compare multiple measures at once, having two independent axes layered on top of one another. Refer to the below image to see how it looks.

Fig 11: Representation of Dual Axis – Data Analyst Interview Questions

Q3. What is the difference between joining and blending in Tableau?

The Joining term is used when you are combining data from the same source, for example, worksheet in an Excel file or tables in an Oracle database. While blending requires two completely defined data sources in your report.

Q4. How to create a calculated field in Tableau?

To create a calculated field in Tableau, you can follow the below steps:

- Click the drop down to the right of Dimensions on the Data pane and select "Create > Calculated Field"to open the calculation editor.
- Name the new field and create a formula.

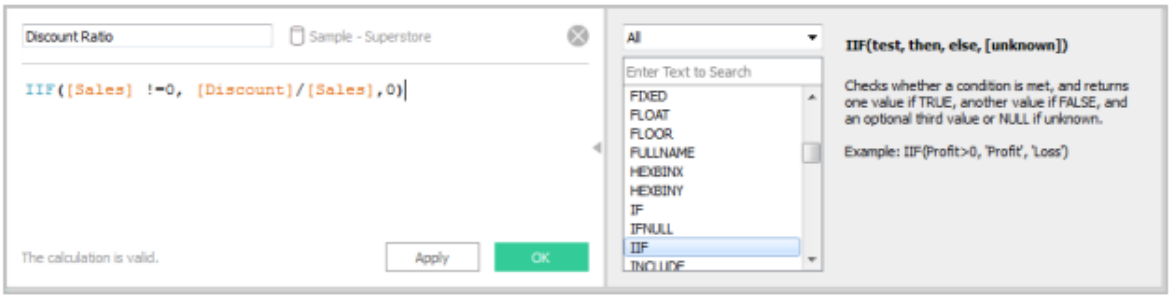


Fig 12: Snapshot of calculated fields – Data Analyst Interview Questions

Q5. How to view underlying SQL Queries in Tableau?

To view the underlying SQL Queries in Tableau, we mainly have two options:

- Use the Performance Recording Feature: You have to create a Performance Recording to record the information about the main events you interact with the workbook. Users can view the performance metrics in a workbook created by Tableau. Help -> Settings and Performance -> Start Performance Recording. Help -> Setting and Performance -> Stop Performance Recording.
- Reviewing the Tableau Desktop Logs: You can review the Tableau Desktop Logs located at C:UsersMy DocumentsMy Tableau Repository. For live connection to the data source, you can check log.txt and tabprotosrv.txt files. For an extract, check tdeserver.txt file.

Q6. Design a view in a map such that if a user selects any country, the states under that country has to show profit and sales.

According to your question, you must have a country, state, profit and sales fields in your dataset.

- Double-click on the country field.
- Drag the state and drop it into Marks card.
- Drag the sales and drop it into size.
- Drag profit and drop it into color.
- Click on size legend and increase the size.
- Right-click on the country field and select show quick filter.
- Select any country now and check the view.

Q7. What is the difference between heat map and tree map?

A heat map is used for comparing categories with color and size. With heat maps, you can compare two different measures together. A treemap is a powerful visualization that does the same as that of the heat map. Apart from that, it is also used for illustrating hierarchical data and part-to-whole relationships.

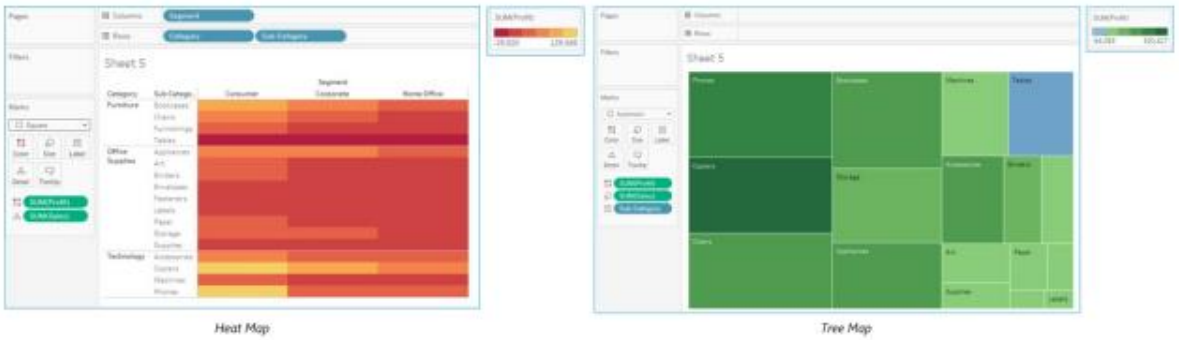


Fig 13: Difference Between Heat Map and Tree Map – Data Analyst Interview Questions

Q8. What is aggregation and disaggregation of data?

Aggregation of data: Aggregation of data refers to the process of viewing numeric values or the measures at a higher and more summarized level of data. When you place a measure on a shelf, Tableau will automatically aggregate your data. You can determine whether the aggregation has been applied to a field or not, by simply looking at the function. This is because the function always appears in front of the field's name when it is placed on a shelf.

Example: Sales field will become SUM(Sales) after aggregation.

You can aggregate measures using Tableau only for relational data sources. Multidimensional data sources contain aggregated data only. In Tableau, multidimensional data sources are supported only in Windows.

Disaggregation of data: Disaggregation of data allows you to view every row of the data source which can be useful while analyzing measures.

Example: Consider a scenario where you are analyzing results from a product satisfaction survey. Here the Age of participants is along one axis. Now, you can aggregate the Age field to determine the average age of participants, or you can disaggregate the data to determine the age at which the participants were most satisfied with their product.

Q9. Can you tell how to create stories in Tableau?

Stories are used to narrate a sequence of events or make a business use-case. The Tableau Dashboard provides various options to create a story. Each story point can be based on a different view or dashboard, or the entire story can be based on the same visualization, just seen at different stages, with different marks filtered and annotations added.

To create a story in Tableau you can follow the below steps:

- Click the New Story tab.
- In the lower-left corner of the screen, choose a size for your story. Choose from one of the predefined sizes, or set a custom size, in pixels.

- By default, your story gets its title from its sheet name. To edit it, double-click
- the title. You can also change your title's font, color, and alignment. Click Apply to view your changes.
- To start building your story, drag a sheet from the Story tab on the left and
- drop it into the center of the view.
- Click Add a caption to summarize the story point.
- To highlight a key takeaway for your viewers, drag a text object over to the story worksheet and type your comment.
- To further highlight the main idea of this story point, you can change a filter or sort on a field in the view, then save your changes by clicking Update above the navigator box.

Q10. Can you tell how to embed views onto Web pages?

You can embed interactive Tableau views and dashboards into web pages, blogs, wiki pages, web applications, and intranet portals. Embedded views update as the underlying data changes, or as their workbooks are updated on Tableau Server. Embedded views follow the same licensing and permission restrictions used on Tableau Server. That is, to see a Tableau view that's embedded in a web page, the person accessing the view must also have an account on Tableau Server.

Alternatively, if your organization uses a core-based license on Tableau Server, a Guest account is available. This allows people in your organization to view and interact with Tableau views embedded in web pages without having to sign in to the server. Contact your server or site administrator to find out if the Guest user is enabled for the site you publish to.

You can do the following to embed views and adjust their default appearance:

- Get the embed code provided with a view: The Share button at the top of each view includes embedded code that you can copy and paste into your webpage. (The Share button doesn't appear in embedded views if you change the showShareOptions parameter to false in the code.)

- Customize the embed code: You can customize the embed code using parameters that control the toolbar, tabs, and more. For more information, see Parameters for Embed Code.
- Use the Tableau JavaScript API: Web developers can use Tableau JavaScript objects in web applications. To get access to the API, documentation, code examples, and the Tableau developer community, see the Tableau Developer Portal.

Now, moving onto something more interesting, I have planned up a set of 5 puzzles, that are most commonly asked in the Data Analyst Interviews.

The analytics industry predominantly relies on professionals who not only excel in various Data Analyzing tools available in the market but also on those professionals who have excellent problem-solving skills. The most important skill that you need to possess is the approach to the problem. Oh yes, your approach should also be in such a way that you should be able to explain to the interviewer.

So, let's get started!

Q1. There are 3 mislabeled jars with Black and White balls in the first and the second jar respectively. The third jar contains a mixture of white and black balls. Now, you can pick as many balls as required to label each jar correctly.

Tell the minimum number of balls to be picked up in this process of labeling the jars.

If you notice the condition in the question, you will observe that there is a circular misplacement. By which I mean that, if Black is wrongly labeled as Black, Black cannot be labeled as White. So, it must be named as Back + White. If you consider that all the 3 jars are wrongly placed, that is, Black + White jar contains either the Black balls or the White balls, but not the both. Now, just assume you pick one ball from the Black + White jar and let us assume it to be a Black ball. So, obviously, you

will name the jar as Black. However, the jar labeled Black cannot have Black + White. Thus, the third jar left in the process should be labeled Black + White. So, if you just pick up one ball, you can correctly label the jars.

Q2. Pumpkin must be equally divided into 8 equal pieces. You can have only 3 cuts.

How do you think, will you make this possible?

The approach to answering this question is simple. You just must cut the pumpkin horizontally down the center, followed by making 2 other cuts vertically intersecting each other. So, this would give you your 8 equal pieces.

Q3. There are 5 lanes on a race track. One needs to find out the 3 fastest horses among the total of 25.

Determine the minimum number of races to be conducted in order to find the fastest three cars.

Now, you can start solving the problem by considering the number of cars racing. Since there are 25 cars racing with 5 lanes, there would be initially 5 races conducted, with each group having 5 cars. Next, a sixth race will be conducted between the winners of the first 5 races to determine the 3 fastest cars(let us say X1, Y1, and Z1).

Now, suppose X1 is the fastest among the three, then that means A1 is the fastest car among the 25 cars racing. But the question is how to find the 2nd and the 3rd fastest? We cannot assume that Y1 and Z1 are 2nd and 3rd since it may happen that the rest cars from the group of X1s' cars could be faster than Y1 and Z1. So, to determine this a 7th race is conducted between cars Y1, Z1, and the cars from X1's group(X2, X3), and the second car from Y1's group Y2.

So, the cars that finish the 1st and 2nd is the 7th race are actually the 2nd and the 3rd fastest cars among all cars.

Q4. Consider 10 stacks of 10 coins each, where each coin weighs 10 grams. But, one of the 10 stacks is defective, and this defective stack contains the coins of 9 grams each.

Find the minimum number of weights needed to identify the defective stack.

The solution to this puzzle is very simple. You just must pick 1 coin from the 1st stack, 2 coins from the 2nd stack, 3 coins from the 3rd stack and so on till 10 coins from the 10th stack. So, if you add the number of coins then it would be equal to 55.

So, if none of the coins are defective then the weight would 55*10 = 550 grams.

Yet, if stack 1 turns out to be defective, then the total weight would be 1 less then 550 grams, that is 549 grams. Similarly, if stack 2 was defective then the total weight would be equal to 2 less than 50 grams, that is 548 grams. Similarly, you can find for the other 8 cases.

So, just one measurement is needed to identify the defective stack.

Q5. Two buses running towards each other on the same track are moving at a speed of 40km/hr and are separated by 80km. A bird takes it flight from the bus A and flies towards bus B at a constant speed of 100km/hr. Once it reaches bus Y, it turns and starts flying back towards bus X. The bird keeps flying to and forth till both the buses collide.

Find the distance traveled by the bird.

The solution to the above problem can be as follows:

- The velocity of the two buses approaching towards each other = (40 + 40)km/hr
- The time taken for the buses to collide = 80km/hr = 1 hour.
- The total distance traveled by the bird = 100km/hr * 1 hr = 100 km.