Data Analyst Interview QnA

1. What are the steps involved in data analysis?

There are several steps: -

**problem statement** > defining the problem statement/ need

**Collecting the data** > Collection of the data by using the different tools.

**Data cleaning**/ **pre-processing** > remove errors, outliers , remove unwanted data point

**Data exploration** > visualize data to uncover insights and identifying patterns

**Modeling**> analysing and defining all the different data that business collects and produces.

**Data** **validation** > checking the integration, accuracy and structure of the data

**Implementation**.> Implementing the business operations.

1. What are the key differences between the Data Analysis and the Data Mining?

Data analysis involves the process of cleaning and organizing the data and produce the meaningful insights whereas Data mining is used to search for hidden patterns in the data.

1. What is data profiling?

Data profiling is a methodology that involves analysing all entities present in data to a greater depth.

Attributes like – data type, frequency of occurance and more.

1. What are the top tools used to perform the data analysis?

There is a wide spectrum of tools that can be used in the field of data analysis:-

Microsoft Office | Power BI | Tableau | RapidMiner | KNMIE | Pandas | Numpy | Matplotlib | Seaborn

1. What is an outlier?

Outliers are data points or observations that are significantly different from the other data points. There are two types of outliers:

1.Univariate

2.multivariate

1. What are some of the popular tools used in Big data?

There are multiple tools for the Big data :

Apache Hadoop | Apache Spark | Apache Storm | Flink | MongoDB | RapidMiner

1. What is a use of Pivot table? - EXCEL

Pivot tables are one of the key features of Excel. They allow user to view quickly summarize large amount of data. Most of the operations of the pivot table includes the drag and drop operations that aid in the quick creation of reports.

1. Explain different types of data validation techniques ?

Some of the data validation techniques are given below:

1.**Field level validation**: it is done across each of the field to ensure that there are no errors in the data entered by the user.

2.**Form level validation**: Here validation is done when the user completes working with the form but before the information is saved.

3. **Data saving validation**: Validation done when the file or a record is being saved.

4. **Search criteria validation**: used to check whether valid results are returned when the user is looking for something.

1. What are the top Apache frameworks used in the distributed computing environment?

MapReduce and Hadoop are considered as the top Apache frameworks when the situation calls for working with a huge dataset in the distributed working environment.

1. What are the ideal situation where the t-test or z-test can be used?

t-test is used when a sample size is less than 30 and the z-test is consider when the sample size exceeds 30 in most cases.

t-test – sample size < 30

z-test – sample size > 30

1. What are the disadvantages of data analytics?

There are few disadvantages of the data analytics;

* Data Analytics can cause a breach of customer privacy.
* Some of the tools are complex and require prior training
* It takes a lot of skills and expertise to select the right analytics tool every time.

1. How can one handle the suspicious or missing data in a dataset while performing analysis?

* Applying several strategies together to find the missing values and using approximation if needed.
* Replacing the invalid data with corresponding valid and up-to-date data.
* Creation of validation report with details about the data in discussion.

1. What is the difference between the recall and the true positive rate?

Recall = True Positive / True Positive + False Negative

Recall and True positive rate are totally same.

1. What is the difference between the standardize and unstandardized coefficients?

In the case of standardized coefficients, they are interpreted on their standard deviation values. While the unstandardized coefficient is measured based on the actual value present in the dataset.

1. What are the ways or techniques used to detect the outliers?

Below are some of the techniques of detecting outliers:

* Box plots
* Z-score
* Interquartile range (IQR)

1. What are Z test, Chi-square test, F test and T-test?

**T-test**: Hypothesis testing for small sample size.

**Z-test**:Hypothesis testing for large sample.

**Chi-square test**: determine the difference observed and expected frequencies of certain observations.

**F test**: Hypothesis of interest are about rhe difference between population means.

1. What are the advantages of using the version control ?

-Creates an easy way to track the life cycle of an application build, including every stage in it such as development, production, testing etc.

-Brings about a good way to establish a collaborative work culture

-Ensures that every verion and variant of code is kept safe and secure.

1. What is collaborative filtering?

Collaborative filtering is an algorithm used to create recommendation systems mainly considering the behavioural data of a customer or user.

E commerce site – recommendations ( based on the previous purchase)

1. What are the properties of the clustering algorithms?

Clustering algorithms , when implemented will have the following properties:

* Flat or hierarchical
* Iterative
* Disjunctive

1. Explain the term normal distribution.

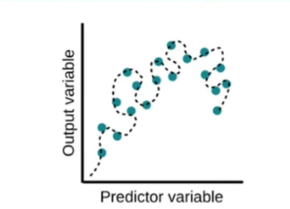
Normal distribution also known as a Gaussian distribution is a type of probability distribution that is symmetric about its mean. (The frequency of occurrences of the data are far from the mean).

* In normal distribution mean, median and mode are equal.

1. What is the difference between overfitting and underfitting?

Overfitting occurs when our machine learning model attempts to cover all or more than the required datapoints present in the given dataset.

* The chances of increase of the overfitting of the data as we much provide the training to our model.

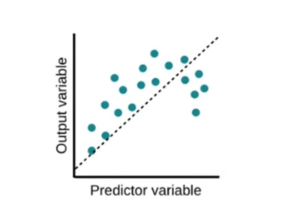
In this graph the model is trying to cover all the datapoints hence it results into overfitting.

* How to avoid the overfitting in our model
* There are several ways by which we can reduce the occurance of overfitting in our model.
* 1.cross -validation
* 2.Training with more data
* 3.Removing features

Other techniques such as ensembling , regularization and many more.

Underfitting:

Underfitting occurs when our machine learning model is unable to capture the underlying trend present in the data.

model is not able to learn enough from the training data.

Underfitted model is high biased and low variance.

* How to avoid the underfitting in model.

1.Increase the model training’s time

2.Increase the number of features.  
22. Mention the difference between the data mining and data profiling?

**Data mining** :- Data mining is a process of discovering relevant information that has not yet been identified before.

- In data mining raw data is converted into the valuable information.

Ex. Suppose a retail company wants to increase their sales. It can use data mining techniques to analyze the sales data and customer’s information to analyze the trends and patterns. By using the market basket analysis, to create the targeted marketing campaigns and optimize the product placement in stores.

**Data Profiling** :- Data profiling is done to evaluate a dataset for its uniqueness , logic and consistency. It involves gathering metadata and statistics about the data to assesses its completeness, consistency and accuracy.

It cannot identify inaccurate or incorrect values.

Ex. A healthcare organization collected the patient data from the various sources such as electronics health records (HER) and medical claims. Before using this data for analyzing and reporting the organization conducts data profiling to ensure its quality and integrity. Data profiling tools can identify the missing values, outliers, duplicates and inconsistencies. This helps the organization understands the data better, and take corrective actions to improve its quality such as cleaning the data.

1. Define the term data wrangling?

Data wrangling is the process of transforming and mapping data from one data form into the another.

* Raw data is cleaned, structured and enriched with intent of making it more valuable for analytics.
* Turned and map out large amount of data from the various sources by using the techniques such as merging, grouping and concatenating , joining and sorting with another data for the finalized data and thereafter it is used as the analytical dataset.

1. What are the common problems that the data analysts encounter during analysis?

* Handling the redundant data.
* Collecting the right data at the right time.
* Handling data purging and Storage Problems.
* Keeping the data secure
* Handling Complaince issues

1. List out some of the common tools that you have worked with?

* **DataBase System** – MySQL, MongoDB, Apache Cassandra , CouchDB
* **Reports and Dasboards** – Excel , Tableau and Power BI
* **Programming Languages** – Python, R, SPSS
* **Presentations and Reports** – Powerpoint and keynotes.

1. What is the significance of the Exploratory data analysis?

Helps you to better understand of the data

* Have more confidence in your decisions. EDA is based on the statistics and the decisions make are based on that.
* EDA can helps in feature selection during modelling. EDA can help to choose the right model.
* EDA can helps to discover unique trends in data.

1. Explain descriptive, predictive and prescriptive analytics.

**Descriptive**: It provides insights into the past to answer “What was happened?” . It uses the data aggregation and data mining techniques.- Hypothesis testing (Data Aggregation and data mining techniques)

**Predictive**:Understands the future to answer “What could happen?”. Uses the statistical methods and forecasting techniques.- (statistical modelling and forecasting techniques)

**Prescriptive**: Suggests various courses of action to answer “What should you do?”. Uses simulation algorithms and optimization techniques to advise possible outcomes.(Simulation algorithm and Optimization techniques)

1. What are different types of sampling techniques used by data analyst?

Sampling is a statistical method to select a subset of data from an entire dataset(population) to estimate the characteristics of the whole population.

* Take the part of the entrie dataset and analyze that.

1.Simple Random sampling

2.Systematic sampling

3.Cluster sampling

4.Stratified sampling

5.Judgemental sampling

1. Describe Univariate , Bivariate and multivariate analysis?

**Univariate analysis** is the simplest and easiest form of data analysis where the data being analyzed contains only one variable.

**Bivariate analysis** can be explained using correlation coefficients, Linear regression, Logistic regression, Scatter plots and Box plots (Two variables)

**Multivariate analysis** involves the analysis of three or more variables to understand the relationship of each variable with the other variable.

1. What are the best methods for the data cleaning?

Firstly create the data cleaning plan by understanding where the common errors take place and keep all the communications open.

Before working with the data, identify and remove the duplicates. This will lead to an easy and effective data analysis process.

Focus on the accuracy of the data. Set the cross-field validation, maintain the value types of the data, and provide mandatory constraints.

Normalize the data at the entry points so that is is less chaotic. You will be able to ensure that all information is standardized, leading the fewer errors at entry.

Statistics QnA

1. How can you handle the missing values in the dataset?

1.Listwise Deletion : Entire record will be deleted from the dataset if any single value is been missing.

2.Average Imputation : Average value will be filled in place of the missing value.

3.Regression Substitution

4.Multiple Imputation

1. Explain the term Normal Distribution ?

Normal distribution refers to continuous probability distribution that is symmetric about the mean.

* Mean , Median and Mode of the Normal Distribution are equal and all located at the centre of the distribution.

1. What is time series analysis?

Time series analysis is a statistical procedure that deals with the ordered sequence of values of a variable at equally spaced time intervals.

* Time series data is data that is collected at the adjacent time periods.

1. Explain the difference between overfitting and underfitting?

**Overfitting** : Model trains well using the training dataset. Performance takes a deep dive on testing dataset.

Happens when your model learns random fluctuations and noise in training dataset.

**Underfitting** : Model does not train well on training dataset. Performs poorly on both training and testing dataset. The training dataset is small. Developing linear model with the non-linear dataset.

1. How do you treat outliers in the dataset?

An outlier is a datapoint that is distant from other similar points. They may be due to variability in the measure or may indicate experimental errors.

Drop Outliers

Cap Outliers data

Assigning a new value

Try a new transformation.

1. What are the different types of Hypothesis testing?

Hypothesis testing is the procedure used by the statisticians and scientist to accept or reject the statistical hypothesis.

Null Hypothesis

Alternative Hypothesis

1. What is the difference between the Type I error and Type II error?

**Type I error** : Occurs when the null hypothesis is rejected even if it is true.

**Type II error** : Occurs when the null hypothesis is not rejected, even if it is false.

Python QnA

1. What is the correct syntax for the reshape function in python?

Reshape(array, shape)

Import numpy as np

Arr = np.arange(12)

Arr3\_4 = **arr.reshape([3,4])**

1. What are the different ways to create a dataframe in Pandas?

There are two ways to create a dataframe in pandas:

1. Initializing a list
2. Initializing a dictionary

To create the dataframe from a list = pd.DataFrame(data=[list], columns=[‘column\_names’])

To create the dataframe from a dictionary = pd.DataFrame(dictionary)

1. There are two arrays ‘A’ and ‘B’. Stack the arrays A and B using the Numpy library in python.

There are two ways to do it:

1. By concatenating method : np.**concatenate**([a,b], axis=1)
2. By using the hstack method: np.**hstack**([a,b])
3. How can you add column to the pandas dataframe?

We can add the column to the pandas dataframe by the : df[‘column\_name’] = list

1. How will you print 4 random numbers between 1 and 15 using numpy?

Import numpy as np

We will use the randint in the numpy : **np.random.randint(1,15,4)**

1. Suppose there is an array that has the values: [0,1,2,3,4,5,6,7,8,9]. How will you display the following values from the array. [1,3,5,7,9].

we will use the function to display the numbers:

arr = np.arange(10)

arr = ([0,1,2,3,4,5,6,7,8,9])

**arr[arr % 2 ==1]** = [1,3,5,7,9] – logic for the odd numbers.

1. Supose there is an array . [[1,2,3],[4,5,6],[7,8,9]]. Extract the value 8 using 2D indexing.

Suppose arr = . [[1,2,3],[4,5,6],[7,8,9]].

We can extract the value 8 by : **arr[2][1]**

1. How do you select specific columns from a dataframe?

We can select the specific columns by : df[[‘**column1’**, ‘**column2’**]]

**SQL- Questions**

1. What is the difference between the where clause and having clause in SQL?

**WHERE** :

* WHERE clause operates on row data
* In the WHERE clause, filter occurs before any grouping are made
* Aggregate functions cannot be used with the where clause.

**HAVING :**

* The HAVING clause operates on aggregated data.
* HAVING is used to filter the values from a group.
* Aggregate functions can be used.

1. Is the following SQL query correct?

SELECT custid, YEAR(order\_date) AS order\_year FROM Orders WHERE order\_year >= 2023;

INCORRECT

SELECT custid, YEAR(order\_date) AS order\_year FROM Orders WHERE YEAR(order\_date) >= 2023; - correct

1. What is subquery in SQL?

A subquery in SQL is a query within a query. It is also known as a nested query or an inner query.

* Subqueries are used to enhance the data which is queried by the main query.
* Correlated and non-correlated

1. What is the difference between the DELETE and TRUNCATE statement?

* DELETE command is used to delete the rows in the table.
* You can rollback the data after using the DELETE command.
* It is a DML command
* It is slower than truncate statement.

TRUNCATE

-TRUNCATE is used to delete all rows from a table

-You cannot rollback the data

- It is a DDL command.

-It is faster than the DELETE command.

1. What do you understand by the query optimization

The phase that identifies a plan for evaluation query which has the least estimated cost is known as query optimization.

* The output is provided faster
* A large number of queries can be executed in less time
* Reduces time and space complexity