CRACKING THE DATA ANALYST INTERVIEW QUESTIONS

WITH FREE RESUME TEMPLATE

Expert strategies and practice questions for success in technical interviews

Curated by DATA WITH VIVIDUS

TABLE OF CONTENTS

QUESTION 1

What are the responsibilities of a Data Analyst?

Some of the responsibilities of a data analyst include:

- Collects and analyzes data using statistical techniques
- Interpret and analyze trends or patterns in complex data sets.
- Establishing business needs together with business teams or management teams.
- Find opportunities for improvement in existing processes or areas.
- Data set commissioning and decommissioning.
- Follow guidelines when processing confidential data or information.
- Provide end-users with training on new reports and dashboards.
- Assist in the data storage structure, data mining, and data cleansing.

QUESTION 2

What is the data analysis process?

Data analysis generally refers to the process of assembling, cleaning, interpreting, transforming, and modeling data to gain insights or conclusions and generate reports to help businesses become more profitable. The process of data analysis is as follows;

- **Collect Data:** The data is collected from a variety of sources and is then stored to be cleaned and prepared. This step involves removing all missing values and outliers.
- **Analyse Data:** As soon as the data is prepared, the next step is to analyze it. Improvements are made by running a model repeatedly. Following that, the model is validated to ensure that it is meeting the requirements.
- **Create Reports:** In the end, the model is implemented, and reports are generated as well as distributed to stakeholders.

QUESTION 3

Explain data cleansing.

Data cleaning, also known as data cleansing or wrangling, is basically a process of identifying and then modifying, replacing, or deleting the incorrect, incomplete, inaccurate, irrelevant, or missing portions of the data as the need arises. This fundamental element of data science ensures data is correct, consistent, and usable.

QUESTION 4

What are the tools useful for data analysis?

Some of the tools useful for data analysis include:

- Microsoft Excel
- Microsoft Power BI
- Tableau
- Python
- R Programme e.t.c

QUESTION 5

Difference between data mining and data profiling.

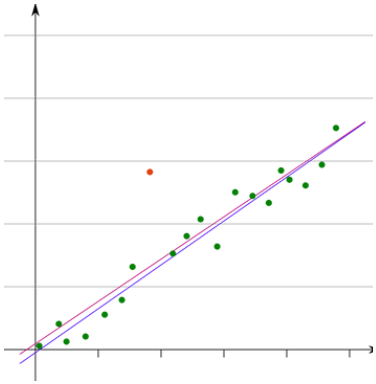
Data mining Process: It generally involves analyzing data to find relations that were not previously discovered. In this case, the emphasis is on finding unusual records, detecting dependencies, and analyzing clusters. It also involves analyzing large datasets to determine trends and patterns in them.

Data Profiling Process: It generally involves analyzing the data's individual attributes. In this case, the emphasis is on providing useful information on data attributes such as data type, frequency, etc. Additionally, it also facilitates the discovery and evaluation of enterprise metadata.

QUESTION 6

Explain Outlier.

In a dataset, Outliers are values that differ significantly from the mean of characteristic features of a dataset. With the help of an outlier, we can determine either variability in the measurement or an experimental error. There are two kinds of outliers i.e., Univariate and Multivariate.

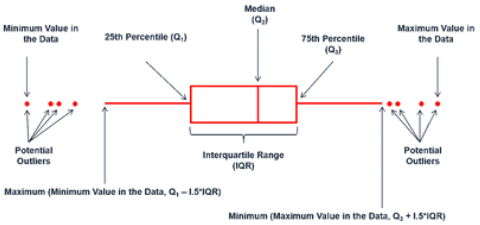


QUESTION 7

What are the ways to detect outliers? Explain different ways to deal with it.

Outliers are detected using two methods:

- **Box Plot Method:** According to this method, the value is considered an outlier if it exceeds or falls below 1.5*IQR (interquartile range), that is if it lies above the top quartile (Q3) or below the bottom quartile (Q1).



- **Standard Deviation Method:** According to this method, an outlier is defined as a value that is greater or lower than the mean ± (3*standard deviation).

QUESTION 8

Difference between data analysis and data mining.

Data Analysis: It generally involves extracting, cleansing, transforming, modeling, and visualizing data in order to obtain useful and important information that may contribute towards determining conclusions and deciding what to do next. Analyzing data has been in use since the 1960s.

Data Mining: In data mining, also known as knowledge discovery in the database, huge quantities of knowledge are explored and analyzed to find patterns and rules. Since the 1990s, it has been a buzzword.

QUESTION 9

What is KNN imputation method.

A KNN (K-nearest neighbor) model is usually considered one of the most common techniques for imputation. It allows a point in multidimensional space to be matched with its closest k neighbors. By using the distance function, two attribute values are compared. Using this approach, the closest attribute values to the missing values are used to impute these missing values.

QUESTION 10

What are the different challenges one faces during data analysis?

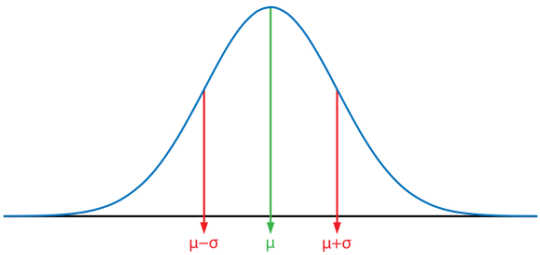
While analyzing data, a Data Analyst can encounter the following issues:

- Duplicate entries and spelling errors. Data quality can be hampered and reduced by these errors.
- The representation of data obtained from multiple sources may differ. It may cause a delay in the analysis process if the collected data are combined after being cleaned and organized.
- Another major challenge in data analysis is incomplete data. This would invariably lead to errors or faulty results.
- You would have to spend a lot of time cleaning the data if you are extracting data from a poor source.
- Business stakeholders' unrealistic timelines and expectations
- Data blending/ integration from multiple sources is a challenge, particularly if there are no consistent parameters and conventions
- Insufficient data architecture and tools to achieve the analytics goals on time.

QUESTION 11

Explain Normal Distribution.

Known as the bell curve or the Gauss distribution, the Normal Distribution plays a key role in statistics and is the basis of Machine Learning. It generally defines and measures how the values of a variable differ in their means and standard deviations, that is, how their values are distributed.



Known as the bell curve or the Gauss distribution, the Normal Distribution plays a key role in statistics and is the basis of Machine Learning. It generally defines and measures how the values of a variable differ in their means and standard deviations, that is, how their values are distributed.

QUESTION 12

What do you mean by data visualization?

The term data visualization refers to a graphical representation of information and data. Data visualization tools enable users to easily see and understand trends, outliers, and patterns in data through the use of visual elements like charts, graphs, and maps. Data can be viewed and analyzed in a smarter way, and it can be converted into diagrams and charts with the use of this technology

QUESTION 13

How does data visualization help you?

Data visualization has grown rapidly in popularity due to its ease of viewing and understanding complex data in the form of charts and graphs. In addition to providing data in a format that is easier to understand, it highlights trends and outliers. The best visualizations illuminate meaningful information while removing noise from data.

QUESTION 14

Write disadvantages of Data analysis.

The following are some disadvantages of data analysis:

- Data Analytics may put customer privacy at risk and result in compromising transactions, purchases, and subscriptions.
- Tools can be complex and require previous training.
- Choosing the right analytics tool every time requires a lot of skills and expertise.
- It is possible to misuse the information obtained with data analytics by targeting people with certain political beliefs or ethnicities.

QUESTION 15

What do you mean by Time Series Analysis? Where is it used?

In the field of Time Series Analysis (TSA), a sequence of data points is analyzed over an interval of time. Instead of just recording the data points intermittently or randomly, analysts record data points at regular intervals over a period of time in the TSA. It can be done in two different ways: in the frequency and time domains. As TSA has a broad scope of application, it can be used in a variety of fields. TSA plays a vital role in Statistics, Signal processing, Econometrics, Weather forecasting, Earthquake prediction, Astronomy, Applied science

QUESTION 16

Can you explain the concept of data normalization and why it is used in data analysis?

Data normalization is the process of scaling data to a common range, typically between 0 and 1. It is used to ensure that different variables with varying scales can be compared and analyzed together without one variable dominating the analysis. Normalization helps in reducing bias and making data more interpretable.

QUESTION 17

What is the difference between structured and unstructured data, and how does this difference impact data analysis?

Structured data is well-organized and typically found in databases, spreadsheets, and tables. Unstructured data, on the other hand, lacks a predefined structure and can include text, images, audio, and video. Structured data is easier to analyze, while unstructured data requires advanced techniques like natural language processing and image recognition, making it more challenging to work with.

QUESTION 18

What is the ETL process, and why is it essential in data analysis?

ETL stands for Extract, Transform, Load. It is a critical process in data analysis where data is extracted from various sources, transformed to fit the desired format and structure, and then loaded into a data repository for analysis. ETL ensures data quality, consistency, and accessibility, making it suitable for analysis.

QUESTION 19

How can data analysis assist businesses in making informed decisions and improving their operations? Can you provide an example?

Data analysis empowers businesses by revealing patterns and trends in their data. For example, a retail company can use sales data analysis to identify which products are selling well and in which regions. This information can help optimize inventory management, marketing strategies, and product offerings to increase profitability. Data analysis enables data-driven decision-making and continuous improvement.

QUESTION 20

What do you mean by univariate, bivariate, and multivariate analysis?

- **Univariate Analysis:** The word uni means only one and variate means variable, so a univariate analysis has only one dependable variable. Among the three analyses, this is the simplest as the variables involved are only one.
- **Bivariate Analysis:** The word Bi means two and variate mean variables, so a bivariate analysis has two variables. It examines the causes of the two variables and the relationship between them. It is possible that these variables are. dependent on or independent of each other.
- **Multivariate Analysis:** In situations where more than two variables are to be analyzed simultaneously, multivariate analysis is necessary. It is similar to bivariate analysis, except that there are more variables involved.

MORE RESOURCES

NEXT EXCEL DATA ANALYST INTERVIEW QUESTIONS

QUESTION 1

What do you mean by Relative cell referencing and Absolute cell referencing in MS Excel?

- Relative cell referencing in Excel means that when you copy a formula to another cell, the cell references within the formula adjust relative to the new location.
- Absolute cell referencing means that the cell references within a formula do not change when you copy the formula to another cell. They remain fixed to specific cells.
- Relative referencing uses no special symbols. Absolute referencing uses a dollar sign ($) before the column and row identifiers (e.g., A1).

QUESTION 2

Excel Function vs. Formula: What Is the Difference?

- A formula is a written instruction for a calculation in Excel. All calculations within a spreadsheet will be written as formulas (e.g. =C3+C4+C5+C7+C8).
- Functions are prewritten formulas and a feature of Excel. The software has over 500 built-in functions that allow users to achieve complicated calculations without having to create the formula themselves or type it out in full.

QUESTION 3

What is the difference between COUNT, COUNTA, and COUNTBLANK?

- **COUNT:** Counts the number of cells containing numerical values in a range.
- **COUNTA:** Counts the number of non-empty cells in a range, including both numerical and text values.
- **COUNTBLANK:** Counts the number of empty cells in a range.

QUESTION 4

How do you remove duplicates in Excel?

To remove duplicates in Excel:

1. **Select the range of data.**
2. **Go to the "Data" tab.**
3. **Click on "Remove Duplicates."**
4. **Choose the columns to check for duplicates.**
5. **Click "OK."**

QUESTION 5

What is the purpose of the CONCATENATE function in Excel?

The CONCATENATE function in Excel is used to combine (concatenate) multiple text strings into one.

QUESTION 6

Explain the VLOOKUP function and its usage.

VLOOKUP is an Excel function used to search for a specific value in a table and return a corresponding value from the same row. It's commonly used for data analysis and allows data analysts to quickly find and extract information from large datasets.

To perform a VLOOKUP in Excel:

1. Select a cell where you want the result to appear.
2. Use the formula: =VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup]).
   - lookup_value: The value you want to find.
   - table_array: The range where you want to search.
   - col_index_num: The column number from which to return a value.
   - [range_lookup]: Set to `TRUE` for an approximate match or **FALSE** for an exact match.
3. Press Enter to get the result.

For example: =VLOOKUP(A2, B2:C10, 2, FALSE) searches for the value in cell A2 in the range B2:C10 and returns the corresponding value from the second column.

QUESTION 7

How do you split text from one cell into multiple cells using a delimiter?

To split text from one cell into multiple cells using a delimiter, you can use the "Text to Columns" feature in Excel.

QUESTION 8

What is the purpose of the COUNTIF and SUMIF functions?

- **COUNTIF:** Counts the number of cells in a range that meet a specified condition.
- **SUMIF:** Adds the values in a range that meet a specified condition.

QUESTION 9

How can you create a pivot table in Excel, and why is it useful?

To create a pivot table in Excel, select the data range, go to the "Insert" tab, and choose "PivotTable." A pivot table is useful for summarising and analysing large datasets, allowing data analysts to quickly generate insights by summarising and visualising data in a structured format.

QUESTION 10

How do you create a bar chart in Excel?

To create a bar chart in Excel:

1. Select the data you want to represent.
2. Go to the "Insert" tab.
3. Choose "Bar Chart" from the chart options.
4. Select the specific bar chart type (e.g., clustered or stacked).

QUESTION 11

Explain the difference between filtering and sorting in Excel.

Filtering in Excel involves displaying a subset of data that meets specific criteria, hiding the rest. Sorting, on the other hand, arranges data in a specific order (e.g., alphabetical or numerical) without hiding any data.

QUESTION 12

What is conditional formatting, and how can it be used for data visualisation?

Conditional formatting is a data visualization technique that applies specific formatting rules to data based on user-defined conditions. It's used to highlight trends, patterns, and anomalies in data, making it easier for data analysts to identify key insights at a glance.

QUESTION 13

Explain the concept of data validation in Excel.

Data validation in Excel refers to the process of defining rules and restrictions for cell entries. It ensures that data entered meets specific criteria, such as numeric values within a range or selecting from a predefined list, enhancing data accuracy and integrity.

QUESTION 14

What is the INDEX-MATCH function combination, and when is it useful?

The INDEX-MATCH function combination is used in Excel to search for and retrieve specific data from a table. It's useful when you need to perform more flexible and powerful lookups than what VLOOKUP or HLOOKUP can provide.

QUESTION 15

What is the purpose of the Text-to-Columns feature in Excel?

The purpose of the Text-to-Columns feature in Excel is to split and separate text in a single cell into multiple columns based on a delimiter, such as a comma or space.

NEXT POWER BI DATA ANALYST INTERVIEW QUESTIONS

QUESTION 1

What are data sources supported by Power BI?

Power BI supports various data sources, including:

1. Excel files
2. SQL databases
3. SharePoint
4. Web services
5. Cloud-based sources (Azure, AWS)
6. On-premises data (using a gateway)
7. Many other connectors and APIs.

QUESTION 2

Explain the concept of data modelling in Power BI?

Data modeling in Power BI involves shaping and transforming data to create a structured model for analysis and visualization. It includes defining relationships between tables, creating calculated columns, and measures to facilitate meaningful insights.

QUESTION 3

What are calculated columns and measures in Power BI, and when would you use them?

- **Calculated Columns:** Calculated columns are columns added to a table in Power BI, and their values are computed row by row based on a DAX formula. They are used for creating new data based on existing data.
- **Measures:** Measures are calculations performed on data at a visual, aggregate level. They are used for aggregations, summaries, and calculations like totals, averages, or ratios in Power BI visuals.
- Use Calculated Columns for row-level computations and Measures for aggregations and summaries in Power BI.

QUESTION 4

How do you create a relationship between tables in Power BI?

In Power BI, you create a relationship between tables by defining a common field as the relationship key in both tables.

QUESTION 5

What is the DAX language, and how is it used in Power BI?

DAX (Data Analysis Expressions) is a formula language used in Power BI to create custom calculations and expressions for data analysis, including measures, calculated columns, and calculated tables. It helps data analysts perform calculations and aggregations on data within Power BI reports and dashboards.

QUESTION 6

Explain the difference between a slicer and a filter in Power BI.

In Power BI, a slicer is a visual element that allows users to interactively filter data within a report, while a filter is a static condition applied to data, typically set by the report designer. Slicers enable dynamic filtering, whereas filters are predefined and do not change based on user interaction.

QUESTION 7

What is the importance of data cleansing and transformation in Power BI?

Data cleansing and transformation in Power BI are crucial for ensuring data accuracy and consistency, ultimately enabling more reliable and meaningful insights.

QUESTION 8

How can you implement row-level security in Power BI?

Row-level security in Power BI can be implemented by defining security roles with DAX expressions in Power BI Desktop, restricting data access based on user attributes, such as username or role.

Click HERE for an article on creating row-level security.

QUESTION 9

What is the role of the Query Editor in Power BI?

The Query Editor in Power BI is used to transform, clean, and shape data from various sources before loading it into a data model. It helps data analysts prepare and manipulate data for analysis.

QUESTION 10

What is Power Query, and how does it help in data transformation?

Power Query is a data transformation tool in Microsoft Excel and Power BI. It helps data analysts clean, reshape, and combine data from various sources for analysis through a user-friendly interface.

QUESTION 11

How can you handle date and time data in Power BI effectively?

In Power BI, you can handle date and time data effectively by using date tables, creating calculated columns and measures, and leveraging built-in time intelligence functions for various time-based analyses.

QUESTION 12

Explain the concept of drill-through in Power BI.

Drill-through in Power BI allows users to navigate from a summarized report to a more detailed report by clicking on a specific data point. It helps analysts explore underlying data for insights.

QUESTION 13

What is the purpose of bookmarks in Power BI?

Bookmarks in Power BI allow users to save and recall specific views, selections, and interactions within a report.

QUESTION 14

Can you describe the difference between DirectQuery and Import modes in Power BI?

- **DirectQuery:**In this mode, data remains in the source, and queries are sent to the source system in real time. It's suitable for large datasets but may have slower performance.
- **Import:** In this mode, data is loaded into Power BI's internal data model. It's suitable for faster analysis but may not be ideal for very large datasets.

QUESTION 15

How can you optimize the performance of a Power BI report?

Optimize Power BI report performance by:

1. Reducing unnecessary visuals.
2. Using summarized data.
3. Applying filters selectively.
4. Using query folding.
5. Reducing custom visuals.
6. Minimizing DAX complexity.
7. Proper data modeling.
8. Using DirectQuery mode.
9. Optimizing visuals' interactions.

NEXT TABLEAU DATA ANALYST INTERVIEW QUESTIONS

QUESTION 1

How can you improve the performance of a Tableau dashboard with a large dataset?

To improve the performance of a Tableau dashboard with a large dataset:

1. Data Extraction: Use data extracts (TDE) to optimize query speed.
2. Data Source Filters: Apply filters at the data source level to reduce the data loaded.
3. Aggregation: Aggregate data where possible to reduce granularity.
4. Limit Visualizations: Limit the number of visualizations on the dashboard.
5. Optimize Calculations: Simplify complex calculations.
6. Dashboard Layout: Optimize layout and use actions for interactivity.
7. Server Performance: Ensure the Tableau server has enough resources.

QUESTION 2

What are the primary data sources that Tableau can connect to?

Tableau can connect to a wide range of primary data sources, including databases like SQL Server, MySQL, Oracle, and cloud platforms like AWS, as well as data in spreadsheets, web data connectors, and more.

QUESTION 3

What is the difference between a live connection and an extract in Tableau?

A live connection in Tableau queries the data source in real-time, while an extract is a static snapshot of the data stored locally for faster performance.

QUESTION 4

How do you create a calculated field in Tableau?

To create a calculated field in Tableau:

1. Right-click in the "Data" pane.
2. Select "Create Calculated Field."
3. Enter the calculation using Tableau's formula syntax.
4. Click "OK" to create the calculated field.

QUESTION 5

Explain the concept of sorting in Tableau.

Sorting in Tableau refers to arranging data in a specific order, either ascending (A to Z, 0 to 9) or descending (Z to A, 9 to 0), based on one or more dimensions or measures. Sorting helps data analysts present information in a more organized and meaningful way for effective data visualization.

QUESTION 6

What is a quick filter in Tableau, and how does it work?

A quick filter in Tableau is a user-friendly filtering option that allows viewers of a Tableau dashboard to easily change the data displayed by selecting values from a list. It works by dynamically filtering the visualizations based on the user's selections without the need for extensive configuration or setup.

QUESTION 7

What is the difference between a table calculation and an LOD (Level of Detail) expression in Tableau?

What is the difference between a table calculation and an LOD (Level of Detail) expression in Tableau?

QUESTION 8

What is the purpose of the Tableau Data Engine?

The Tableau Data Engine is used for in-memory data processing and acceleration of queries, allowing for faster data retrieval and visualization in Tableau.

QUESTION 9

Describe the process of blending data in Tableau.

Blending data in Tableau involves combining data from different sources by matching common fields, allowing for analysis and visualization of integrated data.

QUESTION 10

How can you join data from multiple data sources in Tableau?

In Tableau, you can join data from multiple data sources using the "Data Source" tab and the drag-and-drop interface to define relationships between tables.

QUESTION 11

What is the difference between a discrete and a continuous field in Tableau?

In Tableau, a discrete field contains distinct, separate values, often used for categorical data, while a continuous field contains a range of values and is typically used for numerical data.

QUESTION 12

How do you create a calculated field in Tableau?

To create a calculated field in Tableau, you can right-click on a blank space in the Data pane, then select "Create Calculated Field." Enter the calculation formula and give it a name.

QUESTION 13

What is a dimension and a measure in Tableau, and how are they different?

In Tableau, a "dimension" is a categorical or qualitative field used for grouping and segmenting data, while a "measure" is a quantitative field used for numerical calculations and aggregations. Dimensions are discrete, while measures are continuous.

QUESTION 14

Can you explain the concept of "Marks" in Tableau?

In Tableau, "Marks" refer to the individual data points or visual elements (e.g., bars, points, labels) displayed on a visualization. They represent the level of detail in the view and can be customized to convey specific information in the data visualization.

QUESTION 15

Explain the difference between a worksheet and a dashboard in Tableau.

A worksheet in Tableau is a single view or visualization, while a dashboard is a collection of multiple worksheets and objects combined on a single page for a consolidated view.

NEXT SQL DATA ANALYST INTERVIEW QUESTIONS

QUESTION 1

How do you read and write data to/from various file formats (e.g., CSV, Excel) in Python?

To read and write data in Python:

1. Use `pandas.read_csv('file.csv')` to read data from a CSV file.
2. Use `df.to_csv('file.csv', index=False)` to write data to a CSV file.
3. Use `pandas.read_excel('file.xlsx')` to read data from an Excel file.
4. Use `df.to_excel('file.xlsx', index=False)` to write data to an Excel file.

QUESTION 2

Explain the concept of correlation and how it is calculated.

Correlation measures the statistical relationship between two variables. It is calculated using a correlation coefficient, often Pearson's correlation coefficient (r), which ranges from -1 to 1. A positive value indicates a positive correlation, while a negative value indicates a negative correlation. The closer the coefficient is to 1 or -1, the stronger the correlation, while values near 0 suggest little to no correlation.

QUESTION 3

What are the differences between the .loc and .iloc methods in Pandas, and when would you use each of them?

The main differences between the .loc and .iloc methods in Pandas are:

- .loc uses label-based indexing, while ***.iloc*** uses integer-based indexing.
- .loc includes the end index, while ***.iloc*** excludes it.

Use .loc when you want to select data by label and use `.iloc` when you want to select data by integer position.

QUESTION 4

Explain the use of the apply function in Pandas and provide an example of when it might be helpful.

The ***apply*** function in Pandas is used to apply a given function to each element or row of a data frame or Series. It is helpful for custom data transformations. For example, you can ***apply*** to calculate the square root of each element in a DataFrame column.

QUESTION 5

What is the purpose of the matplotlib library in Python, and how can it be used in data analysis and visualisation?

The purpose of the Matplotlib library in Python is to create data visualizations, including charts, graphs, and plots. Data analysts use Matplotlib to visually represent and analyze data, making it easier to understand and communicate insights from datasets.

QUESTION 6

Describe the purpose of the `merge` and `join` functions in Pandas.

The purpose of the `merge` and `join` functions in Pandas is to combine data from different DataFrames based on common columns, allowing data analysts to perform operations like data aggregation and consolidation.

QUESTION 7

What is a pivot table, and how is it created using Pandas?

A pivot table is a data summarization tool in Pandas. It's created using the **pivot_table()** function, specifying the index, columns, and values to aggregate.

QUESTION 8

What is the difference between a DataFrame and a Series in Pandas?

A DataFrame is a 2-dimensional, tabular data structure in pandas that can store data of different data types in rows and columns. A Series is a 1-dimensional data structure representing a single column or row of data in a DataFrame.

QUESTION 9

What is the purpose of the `groupby` function in Pandas?

The ***groupby*** function in Pandas is used to group and aggregate data based on one or more columns, allowing for summary statistics, data transformations, and analysis by specific categories or criteria.

QUESTION 5

How would you handle missing data in a dataset using Python?

You can handle missing data in Python using the Pandas library by either dropping the missing values using **dropna()** or filling them with a specific value using **fillna()**.

QUESTION 6

Describe the purpose of the `merge` and `join` functions in Pandas.

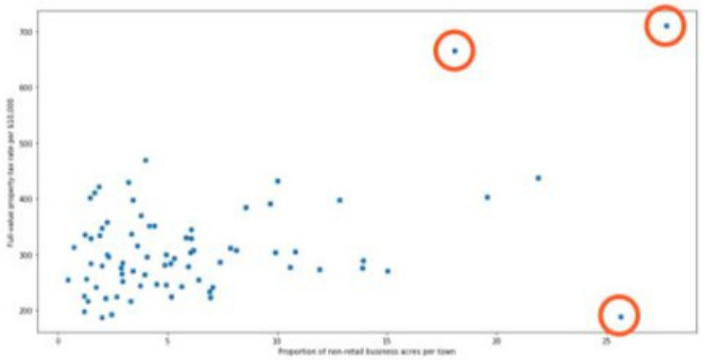
The purpose of the `merge` and `join` functions in Pandas is to combine data from different DataFrames based on common columns, allowing data analysts to perform operations like data aggregation and consolidation.

QUESTION 7

How do you treat outliers in a dataset?

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

The graph depicted below shows there are three outliers in the dataset.



To deal with outliers, you can use the following four methods:

- Drop the outlier records
- Cap your outlier's data
- Assign a new value
- Try a new transformation

QUESTION 8

Suppose there is an array that has 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]?

```
import numpy as np

arr = np.arange(10)
arr
```

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```

Since we only want the odd number from 0 to 9, you can perform the modulus operation and check if the remainder is equal to 1.

```
arr[arr % 2 == 1]
```

```
array([1, 3, 5, 7, 9])
```

QUESTION 9

Suppose there is an array, what would you do?

num = np.array([[1,2,3],[4,5,6],[7,8,9]]). Extract the value 8 using 2D indexing.

```
import numpy as np

num = np.array([[1,2,3],[4,5,6],[7,8,9]])
print(num)
```

```
[[1 2 3]
 [4 5 6]
 [7 8 9]]
```

Since the value eight is present in the 2nd row of the 1st column, we use the same index positions and pass it to the array.

```
num[2,1]
```

```
8
```

NEXT SQL DATA ANALYST INTERVIEW QUESTIONS

QUESTION 1

What are the different types of SQL constraints, and how are they used to maintain data integrity?

The different types of SQL constraints are:

1. Primary Key: Ensures unique and non-null values in a column, often used to identify records.

2. Unique Key: Guarantees uniqueness but allows for null values.

3. Foreign Key: Enforces referential integrity by linking a column to another table's primary key.

4. Check Constraint: Defines conditions that data must meet for insertion or updating.

5. Default Constraint: Provides a default value for a column if one is not specified.

QUESTION 2

How can you optimize the performance of SQL queries when working with large datasets?

To optimize SQL query performance with large datasets:

1. Use indexes on frequently queried columns.
2. Minimize the use of SELECT *.
3. Write efficient WHERE clauses to filter data.
4. Limit the use of subqueries.
5. Normalize your database to reduce redundancy.
6. Optimize joins and use appropriate join types.
7. Consider denormalization for reporting.

QUESTION 3

What is a subquery, and how is it different from a JOIN operation?

A subquery is a nested SQL query that is used within another query. It returns a single value or a set of values to be compared in the outer query.

In contrast, a JOIN operation combines rows from two or more tables based on a related column to create a result set with columns from all the tables involved.

QUESTION 4

What is the difference between UNION and UNION ALL?

1. UNION: The UNION command is used to select related information from two tables, much like the JOIN command. However, when using the UNION command all selected columns need to be of the same data type. With UNION, only distinct values are selected.

2. UNION ALL: The UNION ALL command is equal to the UNION command, except that UNION ALL selects all values.

QUESTION 5

What is the SQL server query execution sequence?

FROM -> goes to Secondary files via the primary file
WHERE -> applies filter condition (non-aggregate column)
SELECT -> dumps data in temp DB system database
GROUP BY -> groups data according to grouping predicate
HAVING -> applies filter condition (an aggregate function)
ORDER BY -> sorts data ascending/descending

QUESTION 6

What is a view in SQL? How to create one

A view is a virtual table based on the result-set of an SQL statement. We can create using the create view syntax.

```
CREATE VIEW view_name AS
SELECT column_name(s)
FROM table_name
WHERE condition
```

QUESTION 7

What is the difference between “Primary Key” and “Unique Key”?

1. We can have only one Primary Key in a table whereas we can have more than one Unique Key in a table.

2. The Primary Key cannot have a NULL value whereas a Unique Key may have only one null value.

3. By default, a Primary Key is a Clustered Index whereas by default, a Unique Key is a unique non-clustered index.

QUESTION 8

What is the difference between the “WHERE” clause and the “HAVING” clause?

1. WHERE clause can be used with a Select, Update and Delete Statement Clause but the HAVING clause can be used only with a Select statement.

2. We can't use aggregate functions in the WHERE clause unless it is in a sub-query contained in a HAVING clause whereas we can use an aggregate function in the HAVING clause. We can use a column name in the HAVING clause but the column must be contained in the group by clause.

QUESTION 9

What is the difference between the “DELETE” and “TRUNCATE” SQL commands?

1. The DELETE command is used to remove rows from a table based on a WHERE condition whereas TRUNCATE removes all rows from a table.

2. So we can use a where clause with DELETE to filter and delete specific records whereas we cannot use a Where clause with TRUNCATE.

3. DELETE is executed using a row lock, each row in the table is locked for deletion whereas TRUNCATE is executed using a table lock and the entire table is locked for removal of all records.

QUESTION 10

What are the differences between OLTP and OLAP?

OLTP stands for Online Transactional Processing
OLAP stands for Online Analytical Processing

OLTP:
- Normalization Level: highly normalized
- Data Usage: Current Data (Database)
- Processing: fast for delta operations (DML)
- Operation: Delta operation (update, insert, delete) aka DML Terms Used: table, columns, and relationships

OLAP:
- Normalization Level: highly denormalized Data Usage: historical Data (Data warehouse)
- Processing: fast for read operations Operation: read operation (select)
- Terms Used: dimension table, fact table

QUESTION 11

What is a stored procedure in SQL, and what are its advantages in data analysis and database management?

What is a stored procedure in SQL, and what are its advantages in data analysis and database management?

QUESTION 12

What is data warehousing, and how does it differ from traditional database systems?

Data warehousing is a centralised repository that stores, integrates, and manages data from various sources for efficient querying and reporting. It differs from traditional databases by focusing on historical and analytical data, while traditional databases are more oriented toward transactional data and real-time operations.

QUESTION 13

What is the purpose of the GROUP BY clause in SQL, and when is it used?

The GROUP BY clause in SQL is used to group rows with similar values in a specific column. It is used to perform aggregate functions on grouped data, such as SUM, COUNT, AVG, etc. It's commonly used to summarize data and perform analysis on subsets of a dataset.

QUESTION 14

Explain the difference between INNER JOIN and LEFT JOIN in SQL. Provide an example for each.

INNER JOIN:
Returns only the rows with matching values in both tables.

Example:
```
SELECT customers.name, orders.product
FROM customers
INNER JOIN orders
ON customers.id = orders.customer_id;
```

LEFT JOIN:
Returns all rows from the left table and the matched rows from the right table. If there's no match, NULL values are returned.

Example:
```
SELECT customers.name, orders.product
FROM customers
LEFT JOIN orders
ON customers.id = orders.customer_id;
```

QUESTION 15

Explain the concept of an index in a database. How does it improve query performance?

An index in a database is a data structure that enhances query performance by allowing the database management system to quickly locate and retrieve specific rows from a table. It works like the index of a book, enabling faster data retrieval and reducing the need for scanning the entire table.

RESUME TIPS FOR YOUR ANALYTICS JOB SEARCH

In today's competitive job market, a strong resume can land you a data job, as companies rely on data.

Here are five resume tips and AI tools for creating an impressive resume. Included also is a free resume template you can use.

1. Create a master resume
2. Organize your skills into categories
3. List your major projects and accomplishments under each role
4. Remember that career summary is optional
5. When to include relevant projects on your resume

DOWNLOAD THE FREE TEMPLATE AND READ FULL ARTICLE HERE:

bit.ly/dwv-resume-tips

Resume Tips For Your Analytics Job Search (With Free Template)

Craft a Standout Resume and Boost Your Chances in the Competitive Data Job Market

UPSKILL YOUR SKILL WITH

DATA WITH VIVIDUS

Join a Thriving Community of 11,000+ Data Scientists, Analysts, and Machine Learning Pros by Subscribing to Our Newsletter and Video Content!

EXPLORE MORE

Subscribe to our newsletter bit.ly/subscribe-to-dwv

Subscribe to our YouTube bit.ly/YT-DWV

Read More on Medium @aareadegboyega

Follow DVW on

DatawithVividus | DatawithVividus | DatawithVividus

CLICK ON THE LINKS ABOVE TO VISIT

Why Data with Vividus?

Navigating the complex world of data can be overwhelming with lots of information, so our newsletter and video content aim at delivering complex topics in the simplest way possible through valuable tips, insightful articles, and more, all geared toward boosting your data science, analysis, and machine learning professional skills.

So, whether you're trying to transition into a data career, a beginner, or leveling your skills – You're home.