Questions about your work in python as data analyst:

1. **Question:** Can you tell me what type of projects have you worked in Python?
   Answer : In Python I have implemented a project for FMCG sector, where I have written a code to connect python programs to different data sources like SQL server and Oracle databases. The raw data was ingested to this databases from different sources like ERP systems. As a data analyst I have performed different operations on this data, like data cleaning, data transformations and in turn the final dataset was being used to drive the decisions for business.
   There are many operations I have performed extensively in python and used different libraries like Numpy, seaborn, pandas, matplotlib. I am very confident in implementing python libraries for data analysis.
2. **Question:** Tell me what type of data cleaning operations you have performed in python?

   Answer : yes, data cleaning is an important activity in data analytics project and I have used following data cleaning operations:

   - Removing null/missing values
   - Removing duplicate values
   - Cleaning unexpected characters
   - Changing data types of columns
   - Renaming column header names to proper names
3. **Question:** In project you implemented, How you handled the data discrepancies?
   Answer : Yes, data discrepancies is one of the common problems in data analytics project, and finding the root cause of this important, so I generally try to go to the root of that data, where I will try to check the source file itself, whether it comes clearly from the source is not that we need to identify, sometimes the data is not coming correctly from the source. So at that point we need to inform the business about the possible loss of data in the process. The other case is, from source if it is coming correctly then finding the problem in our code is next step, debugging the code, step by step is something we need to do. In Database as well we can run some queries to check the data accuracy and then track it in the programs.
   These are some of the steps I followed as a part of finding data discrepancies.
4. **Question:** In your opinion which is the most important step in the data analytics project?

Answer : The most important part of data analytics projects are “Data” and the step we perform at first place is “Data cleaning” that from my perspective is important, this should be very careful implementation, because if this step goes wrong, all the next steps is going to carry the wrong data and final visualizations will give wrong results.

5. **Question:** tell me how you connect Python to SQL database:
   Answer : yes I have quite extensively connected python to various databases and this is the code, I generally used to connect python to SQL Server:

```
import pyodbc
import pandas as pd

cnxn_str = ("Driver={SQL Server Native Client 11.0};"
        "Server=<Server-Name>;"
        "Database=AdventureWorks2019;"
        "Trusted_Connection=yes;")

cnxn = pyodbc.connect(cnxn_str)
```

6. **Question:** Can you tell me how you can create a calendar in Python? Lets suppose I want to create a calendar dates for a year?
   Answer : date_range is one function we can use of Pandas library as below:

```
[3]: import pandas as pd

[19]: calendar_df= pd.date_range(start="01/01/2023", end="12/31/2023",freq="D")

[21]: print(calendar_df)

DatetimeIndex(['2023-01-01', '2023-01-02', '2023-01-03', '2023-01-04',
               '2023-01-05', '2023-01-06', '2023-01-07', '2023-01-08',
               '2023-01-09', '2023-01-10',
               ...
               '2023-12-22', '2023-12-23', '2023-12-24', '2023-12-25',
               '2023-12-26', '2023-12-27', '2023-12-28', '2023-12-29',
               '2023-12-30', '2023-12-31'],
              dtype='datetime64[ns]', length=365, freq='D')
```

7. **Question:** In your understanding what is EDA? And can you explain what steps we generally follow to perform EDA? OR What is the purpose of EDA and what’s the use of it in business?
   Answer: EDA means, Exploratory Data Analysis, which is a process to perform the initial investigations on the data, which helps to:
   - Discover patterns
   - Spot anomalies
   - Test hypotheses
   - Check assumptions

generally we follow the following steps for EDA, when comes to doing practicals

- Connecting to the data
- Identifying the variables - Numerical & Categorical
- Analyzing the variables
- Relationships between variables
- Scatter plot - Numerical variables
- box plot - one numerical value and one categorical values

In a business terms, when it comes to identify/analyze the central tendency, variability & distribution of data, EDA is performed.

8. **Question:** Which tools you have used to write the python codes:
   Answer: Based on your comfort of using a tool, you should answer this question. Jupyter notebook or jupyter lab or visual studio code or any other tool.

9. **Question:**I have one file with the data and I wanted you to start analyzing it with python, what should be your first step?
   Answer: basically interviewer here trying to understand your understanding or thought process once you get the data. So it is very important that you say that looking at the data at first place is my first priority, so I will get the data in python(via connectivity method) and then once it is in python, I will read it first and then try to analyze the data, there are various things I need to identify like:
   - Null values
   - Duplicate values
   - Any junk character
   - Verifying the data types etc.

10. **Question:** Have you used ChatGPT or any other AI tool to generate any code for your projects?
    Answer : Now this is very tricky question, which you should answer diplomatically, now while working on the office laptops, generally we are not allowed to use the AI tool, so you can say something like this:
    “In my work, I am aware about the AI tool, however in my office projects, we were not allowed to us the AI tool for code generation or any other work. So for any help I definitely used any references online but the code I have not generated online, and this surely has helped me to write my code, and build proficiency in my code”.

11. **Question:** You are working in a project and you have tight deadlines, your manager has asked you to work on weekends as well, how you will respond on this?
Answer: yes, this has happened with my projects earlier where I had to work extra hours in weekdays and sometimes in weekends, so I understand when situation demands, I will definitely work over weekends, however if it is continuous things, then I would definitely wanted my manager to thing about the workloads, because I believe there should be some work life balance as well.

12. **Question:** In Python, can you tell me how you join the two files? And what are different types of joins?
Answer: with the help of pandas library we can implement joins in python and based on the common id column between two tables/files, we can join them, some common join types are:
    - Inner join
    - Left join
    - Right join
    - Outer join

    Below example shows about implementation of joins:
    Reading the different files with pandas library:

```
import pandas as pd
week1  = pd.read_csv('Restaurant - Week 1 Sales.csv')
week2  = pd.read_csv('Restaurant - Week 2 Sales.csv')
customers = pd.read_csv('Restaurant - Customers.csv',index_col="ID")
foods  = pd.read_csv('Restaurant - Foods.csv',index_col="Food ID")
satisfaction= pd.read_csv('Restaurant - Week 1 Satisfaction.csv')
```

    Inner join :

```
week1.merge(week2,how="inner",on="Customer ID",suffixes =[' - Week 1',' - Week 2'])
```

	Customer ID	Food ID - Week 1	Food ID - Week 2
0	537	9	5
1	155	9	3
2	155	1	3
3	503	5	8
4	503	5	9
...	...	...	...
57	945	5	4
58	343	3	5
59	343	3	2

Left join on Customer ID column:

```
sales=week1.merge(customers,how="left",left_on="Customer ID",right_index=True).head()
sales
```

	Customer ID	Food ID	First Name	Last Name	Gender	Company	Occupation
0	537	9	Cheryl	Carroll	Female	Zoombeat	Registered Nurse
1	97	4	Amanda	Watkins	Female	Ozu	Account Coordinator
2	658	1	Patrick	Webb	Male	Browsebug	Community Outreach Specialist
3	202	2	Louis	Campbell	Male	Rhynoodle	Account Representative III
4	155	9	Carolyn	Diaz	Female	Gigazoom	Database Administrator III

Similarly we can use right join as well.

13. Question: Can you brief about differences between power BI and Python
    Answer: in brief you should be able to answer this question in terms of data connectivity, how Power BI uses different tools for data connectivity and analysis and then you should be able to give some details about flexibility, data visualization in tools.
    Understand more details about this in the below link:
    https://github.com/AtulKadlag/PythonDataAnalyst/blob/main/PowerBIvsPython.pdf

I. Python Fundamentals & Data Structures

1. **Question:** Explain the difference between lists and tuples in Python. When would you use each?
    - **Answer:**
        - Lists are mutable (changeable) sequences, while tuples are immutable.
        - Lists are typically used when you need to store a collection of items that might change during the program's execution (e.g., storing a list of user inputs).

        - Tuples are used when you need to store a collection of items that should not be modified (e.g., representing coordinates or storing configuration settings).
        - Lists use [] brackets, while tuples use () parenthesis.

2. **Question**: What are dictionaries in Python, and how are they useful for data analysis?

   Answer:

   - Dictionaries are key-value pairs, allowing you to store and retrieve data efficiently using keys.
   - In data analysis, dictionaries are useful for:
     - Storing data in a structured format (e.g., representing a record with field names as keys).
     - Counting occurrences of items.
     - Mapping values (e.g., mapping category names to numerical codes).
     - Creating lookup tables.

3. **Question:** How do you handle missing values in a list or a dictionary?

- **Answer:**
    - For lists: You can use None, NaN (from NumPy), or remove the elements.
    - For dictionaries: You can check if a key exists using key in dict, use dict.get(key, default_value) to provide a default, or remove the key-value pair using del dict[key].
    - When using Pandas, using NaN and fillna is very common.

II. **NumPy & Pandas**

1. **Question:** What is NumPy, and what are its advantages over Python lists for numerical computations?
    - **Answer:**
        - NumPy is a Python library for numerical computing, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
2. **Question:** Explain the difference between a Pandas Series and a DataFrame.

- **Answer:**
    - A Pandas Series is a one-dimensional labeled array, similar to a column in a spreadsheet or SQL table.
    - A Pandas DataFrame is a two-dimensional labeled data structure, similar to a spreadsheet or SQL table, consisting of rows and columns. It can be thought of as a collection of Series objects sharing the same index.

3. **Question:** How do you handle missing values in a Pandas DataFrame?

- **Answer:**
    - Use df.isnull() or df.isna() to identify missing values.
    - Use df.fillna(value) to replace missing values with a specific value, mean, median, or other strategies.
    - Use df.dropna() to remove rows or columns with missing values.
    - Use df.interpolate() to fill missing values using interpolation methods.

4. **Question:** How do you filter rows in a Pandas DataFrame based on a condition?

- **Answer:**
    - Use boolean indexing. For example, df[df['column_name'] > 10] selects rows where the 'column_name' column is greater than 10.
    - You can also use .loc and .iloc for more complex filtering.

5. **Question:** Explain how to group data in a Pandas DataFrame and calculate aggregate statistics.

- **Answer:**
    - Use the df.groupby('column_name') method to group rows based on the values in a specified column.
    - Then, use aggregation functions like mean(), sum(), count(), min(), max(), std(), etc., to calculate statistics for each group.
    - Example: df.groupby('category')['sales'].sum() calculates the total sales for each category.

III. Data Analysis & Visualization

1. **Question:** What are some common data cleaning tasks you would perform on a dataset?

- **Answer:**
    - Handling missing values (imputation or removal).
    - Removing duplicate rows.
    - Correcting data type inconsistencies.
    - Standardizing or normalizing numerical data.
    - Removing outliers.
    - Parsing and formatting dates and strings.

2. **Question:** How would you visualize the distribution of a numerical variable?
    - **Answer:**
        - Use a histogram (e.g., with matplotlib.pyplot.hist() or seaborn.histplot()).
        - Use a box plot (e.g., with seaborn.boxplot()) to show quartiles and outliers.
        - Use a density plot (e.g., with seaborn.kdeplot()).

3. **Question:** How would you visualize the relationship between two categorical variables?

- **Answer:**
    - Use a stacked bar chart or grouped bar chart (e.g., with matplotlib.pyplot.bar() or seaborn.countplot()).
    - Use a heatmap (e.g., with seaborn.heatmap()) to show the frequency of combinations.
    - Use a mosaic plot.

4. **Question:** What is the purpose of a correlation matrix, and how do you create one in Python?

- **Answer:**
    - A correlation matrix shows the pairwise correlations between numerical variables in a dataset.
    - It helps identify relationships and potential multicollinearity.
    - In Python, you can create a correlation matrix using df.corr() in Pandas, and visualize it with seaborn.heatmap().

5. **Question:** What is the purpose of using libraries like Seaborn or Matplotlib?

- **Answer:**
    - These libraries are used for data visualization.
    - Matplotlib provides a low-level, flexible framework for creating various plots.
    - Seaborn is built on top of Matplotlib and offers a higher-level interface with attractive statistical plots, simplifying common visualization tasks.

IV. SQL & Database Interaction (Often required for Data Analysts)

1. **Question:** How would you connect to a SQL database from Python?
    - **Answer:**
        - Use libraries like psycopg2 (for PostgreSQL), mysql.connector (for MySQL), or sqlite3 (for SQLite).
        - Establish a connection using connection parameters (hostname, username, password, database name).
        - Create a cursor object to execute SQL queries.
        - Fetch the results and process them.
        - Close the cursor and connection.
2. **Question:** How would you execute a SQL query from python and put the result into a Pandas DataFrame?
    - **Answer:**
        - Establish a connection to the database.
        - Create a cursor.
        - Execute a SQL SELECT query.
        - Use pandas.read_sql_query(sql_query, connection) to put the result directly into a DataFrame.

V. General Data Analysis Concepts

1. **Question:** What are some common data analysis techniques you've used?
    - **Answer:**
        - Descriptive statistics (mean, median, standard deviation).

        - Data visualization (histograms, scatter plots, box plots).
        - Correlation analysis.

2. **Question:** How do you handle outliers in your data?
    - **Answer:**
        - Identify outliers using visualization techniques (box plots, scatter plots) or statistical methods (z-scores, IQR).
        - Remove outliers if they are due to errors or anomalies.
        - Transform the data (e.g., log transformation) to reduce the impact of outliers.
        - Use robust statistical methods that are less sensitive to outliers.
        - Impute the outliers with acceptable values.
3. **Question:** Explain the concept of data normalization or standardization.
    - **Answer:**
        - Normalization and standardization are techniques used to rescale numerical data to a common range.
        - Normalization typically scales data to a range of [0, 1].
        - Standardization scales data to have a mean of 0 and a standard deviation of 1.
        - These techniques are used to improve the performance of machine learning algorithms and to make data comparable.

Scenario based Questions:

Scenario 1: E-commerce Sales Analysis

Scenario: You are a data analyst for an e-commerce company. You have a dataset containing sales transactions, including order IDs, product names, prices, quantities, and order dates.

Question 1: How would you use Pandas to load and clean this dataset, handling potential missing values and data type inconsistencies?

Possible answer : Discuss using

pd.read_csv(), df.isnull().sum(),df.fillna(), df.dropna(), df.dtypes, pd.to_datetime(), df['column'].astype()

Question 2: *How would you calculate the total revenue for each product category?*

Possible Answer: *Explain using* `df.groupby('category')['price'].sum()`

Question 3: *How would you visualize the monthly sales trends over the past year?*

Possible Answer: Describe converting the order date to a datetime object, extracting the month and year, grouping by month and year, calculating the total sales, and using matplotlib or seaborn to create a line plot.

Question 4: *How would you identify the top 5 customers with the highest total spending?*

Possible Answer: *Explain grouping by customer ID, calculating the total spending per customer, sorting the results in descending order, and selecting the top 5.*

Scenario 2: Social Media Engagement Analysis

Scenario *: You are analyzing social media data for a marketing campaign. You have a dataset containing post IDs, user IDs, post content, number of likes, number of comments, and post timestamps.*

Question 1: *How would you extract relevant features from the post content, such as the number of hashtags or mentions?*

Possible answer : *Discuss using regular expressions (re module) or string manipulation techniques to count hashtags and mentions.*

Question 2: *How would you calculate the average engagement rate (likes + comments / number of views) for each user? (Assume views are also in the dataset).*

Possible Answer : *Explain creating a new 'engagement' column, calculating the engagement rate, grouping by user ID, and calculating the mean engagement rate.*

Question 3: *How would you visualize the distribution of post engagement rates?*

Possible Answer: *Describe using a histogram or box plot with seaborn or matplotlib.*

Question 4: *How would you identify posts that received an unusually high number of likes or comments (outliers)?*

Possible Answer: *Explain using box plots, z-scores, or IQR to identify outliers.*

Scenario 3: Website Traffic Analysis

Scenario: You are analyzing website traffic data. You have a dataset containing timestamps, user IDs, page URLs, and session durations.

Question 1: *How would you calculate the average session duration for each page?*

Possible Answer: *Explain grouping by page URL and calculating the mean session duration.*

Question 2: *How would you identify the most popular entry and exit pages?*

Possible Answer: *Describe identifying the first and last page URLs in each session, and counting the occurrences of each page URL.*

Question 3: *How would you visualize the daily website traffic over a specific period?*

Possible Answer: *Explain converting timestamps to datetime objects, extracting the date, grouping by date, counting the number of sessions, and creating a line plot.*

Question 4: *How would you identify users who exhibit unusual browsing behavior (e.g., extremely long or short sessions)?*

Possible Answer: *Discuss using box plots or scatter plots to visualize session durations and identify outliers.*

Scenario 4: Customer Churn Analysis

Scenario:*You are analyzing customer churn data for a subscription-based service. You have a dataset containing customer IDs, subscription start dates, subscription end dates (if applicable), and customer demographics.*

Question 1: *How would you calculate the customer churn rate over time?*

Possible answer : *Explain calculating the number of churned customers and the total number of customers for each time period (e.g., monthly), and dividing the churned customers by the total customers.*

***Question 2**: How would you identify customer demographics that are associated with higher churn rates?*

***Possible answer** : Describe grouping by demographic variables and comparing the churn rates between different groups.*

***Question 3** : How would you visualize the distribution of customer subscription durations?*

***Possible answer** : Explain calculating the subscription duration for each customer, and using a histogram or density plot.*

***Question 4** : How would you determine the average customer lifetime value for different customer segments?*

***Possible answer** : Explain calculating the customer lifetime value for each customer, grouping by customer segments, and calculating the average lifetime value for each segment.*