* **ImmverseAI Sonali Umak**

**Data Analytics Challenge**

**Question & Answer Task:**

● **Question 1**: What is Pandas, and why is it commonly used in data cleaning tasks?

**Ans**: **Pandas:**Pandas is an open-source data analysis and manipulation library for the Python programming language.

**Pandas is commonly used in data cleaning tasks for several reason:**

Data import and export,Data Transformation,Data Exploration,Data Integration,Easy data manipulation.

**● Question 2:** Given a DataFrame with missing values, how would you check for missing

values in each column and count the total number of missing values?

**Ans:** In order to check missing values in Pandas DataFrame, we use a function isnull() and notnull(). Both function help in checking whether a value is NaN or not. These function can also be used in Pandas Series in order to find null values in a series.

Example:- df.isnull().sum()

**● Question 3:** How can you remove duplicates from a DataFrame while retaining the first

occurrence of each unique row?

**Ans**:- drop\_duplicates() will keep the first instance of a duplicate row and remove any others. We are then removing the inversion with the second . T property. The DataFrame will then display normally and save it to a new variable no\_dup\_columns.

**● Question 4:** If you have a DataFrame with a column containing string values, how can

you convert all the values in that column to lowercase?

**Ans:** By using the lower function we can convert all the string values in that column to lowercase

Example: he to lower() function converts each uppercase character in a given string to its lowercase equivalent. The str variable is first initialized with the string “HELLO WORLD”, and then each character in the string is passed as an argument to the to lower() function.

**● Question 5:** How do you replace missing values in a DataFrame with a specific value,

like 0, for a particular column?

**Ans:-** With the fillna function, we replace the missing data with the value 0. You can replace it with any kind of value when using the fillna function. For example, I replace the missing values with the string 'zero'.

**● Question 6:** If you have a DataFrame with a datetime column, how can you extract the

year, month, and day into separate columns?

**Ans:-** strftime() Let's create a Pandas DataFrame with the column of Datetime values and use the pd. to\_datetime() function to convert this column into a datetime column and finally, use the strftime() method to extract the month and year from a datetime column.

**● Question 7:** How can you filter rows in a DataFrame where a specific column's values

meet a certain condition (e.g., all rows where 'age' is greater than 30)?

**Ans:-** We will use vectorization to filter out such rows from the dataset which satisfy the applied condition. Let’s use the vectorization operation to filter out all those rows which satisfy the given condition.

# Filter all rows for which the player's

# age is greater than or equal to 30

df\_filtered = df[df['Age'] >= 30]

# Print the new dataframe

print(df\_filtered.head(15))

# Print the shape of the dataframe

print(df\_filtered.shape)

**● Question 8:** What is the purpose of the .apply() function in Pandas, and how would you

use it to create a new column based on values from existing columns?

**Ans:-** The apply() function allows us to apply a function to each row or column of a DataFrame and return a new DataFrame. We can then apply this function to the age column using the apply() function and assign the result to a new column called age\_group .

**● Question 9:** Suppose you want to merge two DataFrames, 'df1' and 'df2,' on a common

column 'key.' How would you perform this merge operation in Pandas?

**Ans:-** Here, df1 and df2 are the two dataframes you want to merge, and the “on” argument defines the column(s) for combining. By default, pandas will perform an inner join, which means that only the rows with matching keys in both dataframes are included in the resulting dataframe.

**● Question 10:** You have a DataFrame with a column containing messy text data. How can

you clean and standardize the text data (e.g., remove punctuation and convert to

lowercase) in that column?

**Ans:-** Clean every punctuations out of the text.Split the messages in seperate words, so that only one word is in one dataframe entry.

If it is possible, one smiley should be considered as a single word. If this it not possible, how to clean them out?

Make every text lower case. There is already a solution for that, but it would be really nice to include it into the "cleaning code".