NAME: SNEHA K

TASK 1—INTEVIEW QUESTION

1. WHAT ARE MISSING VALUES AND HOW DO YOU HANDLE THEM ?

   - Missing values are the null values present in the dataset. It is usually shown as NaN, None, null, etc.
   - They usually arises due to :
     1. Data entry errors
     2. Corrupted files and
     3. Information not available at that time
   - To handle missing values we generally fill or drop the null values using 'fillna' or 'dropna' functions in python(pandas)

   Eg: df.fillna(method='ffill')

   df.fillna(method='bfill')

   df .dropna(inplace=True)

   - apart from this we can also fill the missing values using mean , median and mode of specific columns in a datasets.
   - Generally, we use mean and median to fill the null values in numerical columns

   And mode and unknown to fill the null values in categorical columns.

2. HOW DO YOU TREAT DUPLICATE RECORDS.
   - If duplicates are exact and not useful remove that to clean the data.
   - Keep the first and last occurences and remove all the duplicate rows using "drop_duplicates()". If they have different information they need to be merged (using pd.merge()) instead of deleting.

3. Difference between dropna() and fillna() in pandas?

Dropna():

- Removes the null values from the dataset.
- Delete rows with NaNs
- Data is lost removing rows and columns.

Fillna():

- Fills the null values in the dataset.
- Replace rows with NaNs.
- Data is not lost.

4. WHAT IS OUTLIER TREATMENT AND WHY IS IT IMPORTANT?

⇨ It is the process of detecting and managing extreme values in the datasets.

Importance of outliers treatment:

- It improves the accuracy : when outliers are in a dataset it distorts the statistical analyses affecting mean , median and mode therefore treating it can improve the accuracy of the dataset.
- Better insights
- Good model performance
- Consistent
- Reliable and robust

5. EXPLAIN THE PROCESS OF STANDARDIZING THE DATA?

Text standardization is the process of cleaning and transforming text data into consistent format, making it suitable for analysing and modeling .

Process:

- Convert to lower or proper case
- Strip whitespace
- Replace inconsistent labels
- Remove special characters

6. HOW DO YOU HANDLE INCONSISTENT DATA FORMAT?

Handling inconsistent data formats is a key step in data cleaning and preprocessing . Inconsistent formats can lead to errors in analysis, incorrect model results and difficulty in visualization .

To handle inconsistent data formats:

- Understand the data
- Convert datatypes
- Standardize text fields
- Creating mapping for unknown variants
- Fill or remove unconvertible values
- Unify units of measurements

7. WHAT ARE COMMON DATA CLEANING CHALLENGES?
    - Missing values: Deciding whether to impute , remove or leave based on the impact .
    - Duplicate records : Identifying subtle duplicates when names or IDs are slightly different.
    - Inconsistent Formatting: Standardizing values without losing meaning.
    - Outliers and noise: Distinguishing between extreme valid and actual errors
    - Incorrect datatype: Converting types safely for large datasets.
    - Human errors: Automation and validation rules can reduce but not eliminate these issues.
8. HOW CAN YOU CHECK DATA QUALITY?

    To check data quality the dataset can be evaluated on several key factors:

    - Completeness: check for missing values using ".isnull()" in pandas and fill or drop the null values.
    - Accuracy: Ensure entries reflect real world truth.
    - Consistency: standard formatting, naming and units.
    - Uniqueness: check for duplicate values and remove or merge the duplicate values.
    - Timeliness: Check whether the data is current and updated by comparing the versioning of records , it should be relevant and recent for the task.
    - Validity: check that the data follows defined formats, rules or ranges .

TOOLS USED: EXCEL, PYTHON(PANDAS) , POWER BI , TABLEAU, SQL QUERIES.