Data Cleaning & Essential Functions Assignment

1. What is data cleaning, and why is it important in data analysis? What are the potential consequences of analysing unclean or messy data? Explain the common steps involved in cleaning and organising data.

Answer: **Data cleaning** (also called **data cleansing**) is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset so that the data is accurate, complete, consistent, and usable for analysis.

It ensures that the data truly represents real-world information before applying statistical analysis, machine learning, or business decision-making.

Data analysis is only as good as the data used. Clean data helps to:

- Improve **accuracy of results**
- Reduce **bias and errors**
- Ensure **reliable insights and conclusions**
- Save time during analysis and modelling
- Improve **model performance** in machine learning

Analysing messy data can cause serious problems, such as:

1. Incorrect Results

Missing values or wrong entries can distort averages, trends, and predictions.

2. Misleading Insights

Duplicate or inconsistent data can lead to false patterns and wrong business decisions.

3. Poor Model Performance

Machine learning models trained on dirty data give inaccurate predictions.

4. Increased Costs and Time

Fixing issues later is more expensive and time-consuming.

5. Loss of Trust

Stakeholders may lose confidence in reports and analytics.

Common Steps Involved in Cleaning and Organising Data:

Step 1: Understanding the Data

- Check data types, formats, ranges, and structure
- Identify what each variable represents

Step 2: Handling Missing Values

- Remove rows or columns with too many missing values
- Fill missing values using mean, median, mode, or appropriate methods

Step 3: Removing Duplicates

- Identify and delete repeated records

Step 4: Correcting Errors

- Fix incorrect values (e.g., negative age, wrong dates)
- Standardize spellings and categories (e.g., “M”, “Male” → “Male”)

Step 5: Handling Outliers

- Detect unusually high or low values
- Decide whether to keep, transform, or remove them

Step 6: Standardising and Formatting Data

- Convert data into consistent units (e.g., INR, dates, measurements)
- Ensure uniform data types

Step 7: Validating Data

- Recheck accuracy after cleaning
- Ensure data meets required rules and constraints

Simple Example

Unclean Data

Age: 25, 30, -5, 200, NULL

Clean Data

Age: 25, 30, NULL (or removed)

Summary

Aspect	Explanation

Data Cleaning	Process of fixing errors and inconsistencies
Importance	Ensures accuracy, reliability, and trust
Messy Data Impact	Wrong insights, poor decisions
Key Steps	Missing values, duplicates, errors, formatting

2. How would you sort the following dataset first by "Department" (A-Z) and then by "Salary" (Largest to Smallest)? Write a step-by-step approach.

Employee	Department	Salary
Sonu	IT	4000
Pranav	HR	5000
Rahul	IT	2500

Answer: Step 1: Identify the Sorting Criteria

- **Primary sort:** Department (A–Z)
- **Secondary sort:** Salary (Largest → Smallest)

Step 2: Sort by Department (A–Z)

Alphabetical order of departments:

- **HR**
- **IT**

After sorting by Department:

Employee	Department	Salary
Pranav	HR	5000
Sonu	IT	4000
Rahul	IT	2500

Step 3: Sort by Salary within Each Department (Descending)

- HR has only one employee → no change
- IT department salaries:
  - Sonu → 4000
  - Rahul → 2500
    (Already in descending order)

Step 4: Final Sorted Dataset

Employee	Department	Salary
Pranav	HR	5000
Sonu	IT	4000
Rahul	IT	2500

First, the dataset is sorted alphabetically by **Department**. Then, within each department, employees are arranged by **Salary from highest to lowest**.

3. Explain the use of text functions such as TRIM , LEFT, RIGHT, MID, and CONCAT in data cleaning.

Answer: Use of Text Functions in Data Cleaning-

Text functions help clean, format, and standardise text data so it can be **analysed correctly and consistently**.

1. TRIM()

Use:

- Removes **extra spaces** from text (leading, trailing, and extra spaces between words)

Why important:

- Prevents mismatches during sorting, filtering, or matching

Example:

TRIM(" Data   Analysis ")

Result:

"Data Analysis"

2. LEFT()

Use:

- Extracts a specified number of characters from the **left side** of a text string

Why important:

- Useful for extracting codes, initials, or prefixes

Example:

LEFT("EMP12345", 3)

Result:

"EMP"

3. RIGHT()

Use:

- Extracts a specified number of characters from the **right side** of a text string

Why important:

- Helps extract year, ID numbers, or suffixes

Example:

RIGHT("Invoice2025", 4)

Result:

"2025"

4. MID()

Use:

- Extracts text from the **middle** of a string by specifying the start position and length

Why important:

- Useful when required data is located inside a text value

Example:

MID("ORD-56789-IN", 5, 5)

Result:

"56789"

5. CONCAT()

Use:

- Combines multiple text strings into **one single string**

Why important:

- Helps create full names, addresses, or combined IDs

Example:

CONCAT("Data", " ", "Cleaning")

Result:

"Data Cleaning"

Summary Table:-

Function	Purpose	Data Cleaning Use
TRIM	Removes extra spaces	Fixes spacing errors

LEFT	Extracts from left	Gets prefixes/codes
RIGHT	Extracts from right	Gets suffixes/years
MID	Extracts from middle	Gets internal values
CONCAT	Joins text	Combines fields

Text functions are essential in data cleaning because they **remove inconsistencies, extract useful information, and standardise text data**, ensuring accurate analysis.

4. What is the role of date functions like TODAY in managing datasets?

Answer: Role of Date Functions like TODAY() in Managing Datasets-

Date functions help handle, track, and analyse **time-based data** accurately. The TODAY() function is especially useful because it **automatically returns the current date**, which updates whenever the dataset is opened or recalculated.

- Returns the **current system date**
- Updates **automatically** (no manual entry needed)

Example (Excel / Google Sheets):

=TODAY()

If today is 19-Dec-2025, the result will be:

19-12-2025

<u>Key Roles of TODAY() in Dataset Management-</u>

1. Tracking Current Status

- Identifies **overdue tasks**, expired contracts, or pending payments
- Example: Check if a due date is before today

=IF(A2 < TODAY(), "Overdue", "On Time")

2. Calculating Time Differences

- Helps calculate **age, tenure, delivery time, or project duration**

=TODAY() - A2

(Where A2 is a start date)

3. Automating Reports

- Reports update **daily without manual changes**
- Useful in dashboards and performance tracking

4. Data Validation & Cleaning

- Detects **future or invalid dates**
- Helps ensure data consistency

=IF(A2 > TODAY(), "Invalid Date", "Valid Date")

5. Time-Based Analysis

- Enables grouping and filtering of data by:
  - Today
  - Last 7 days
  - Last month

Benefit	**Explanation**
Automation	Updates automatically
Accuracy	Reduces manual date errors
Efficiency	Saves time in recurring reports
Consistency	Uses a single reference date

Date functions, such as TODAY(), play a crucial role in managing datasets by automating date tracking, enabling real-time analysis, validating data, and enhancing **decision-making**.

5. Apply Data Validation to restrict Quantity values to only whole numbers between 1 and 10.

   a. Configure an input message that appears when a user selects a cell in the "Quantity" column, explaining: "Please enter a whole number between 1 and 10."

   b. Set up an error alert message that triggers if the user enters a number less than 1 or greater than 10, showing: "Invalid input! The quantity must be a whole number between 1 and 10."

Write a step-by-step approach for this question

Customer Name	**Product Name**	**Category**	**Quantity**	**Unit Price ($)**
Jane Smith	Shoes	Electronics		81
Isabella Moore	Laptop	Electronics		121
Daniel Davisz	Sofa	Clothing		239
Alex Moore	Shoes	Electronics		500
Michael Johnson	Table Lamp	Home Decor		423
Daniel Johnson	Backpack	Electronics		160
Isabella Davis	Headphones	Electronics		348
Jane Davis	Headphones	Electronics		152
Alex Wilson	T-shirt	Home Decor		369

Answer: Applying Data Validation on the Quantity Column-

Objective:

- Allow **only whole numbers between 1 and 10** in the **Quantity** column Show:
  - An **input message** when the cell is selected
  - An **error alert** when invalid data is entered

Step-by-Step Approach:

Step 1: Select the Quantity Column

- Open the dataset in **Excel**
- Select all cells under the **Quantity** column
  (Example: D2:D10, excluding the header)

Step 2: Open Data Validation

1. Go to the **Data** tab on the Excel ribbon
2. Click on **Data Validation**
3. The **Data Validation** dialog box will open

Step 3: Set Validation Rules (Settings Tab)

1. In the **Settings** tab:
    - **Allow:** Select Whole number
    - **Data:** Select between
    - **Minimum:** Enter 1
    - **Maximum:** Enter 10

This restricts entries to **whole numbers from 1 to 10 only**.

Step 4: Configure Input Message (Input Message Tab)

1. Click on the **Input Message** tab
2. Check **"Show input message when cell is selected"**
3. Enter:
    - **Title:** Quantity Input
    - **Input Message:**

Please enter a whole number between 1 and 10.

This message appears when the user clicks a Quantity cell.

Step 5: Set Error Alert (Error Alert Tab)

1. Click on the **Error Alert** tab
2. Check **"Show error alert after invalid data is entered"**
3. Set:
    - **Style:** Stop
    - **Title:** Invalid Input
    - **Error Message:**

Invalid input! The quantity must be a whole number between 1 and 10.

This prevents invalid values from being entered.

Step 6: Click OK

- Click **OK** to apply the data validation rules

Final Result-

- ✔ Only whole numbers **1–10** are allowed
- ✔ Input message guides the user
- ✔ Error alert blocks invalid entries

Data Validation ensures data accuracy by restricting Quantity values to valid whole numbers between 1 and 10, while providing clear guidance and error messages to users.

6. Understand and apply fundamental text functions like LEFT, RIGHT, MID, and LEN.

   · Extract the first 5 characters from the string "ExcelTipsAreGreat" using the LEFT function.

   · Extract the last 4 characters from "DataAnalysis.xlsx" using the RIGHT function.

   · Extract the substring "Tips" from "ExcelTipsAreGreat" using the MID function.

   · Count the total number of characters in the string "Hello World!" using the LEN function.

   · Create a formula to extract the middle 6 characters from "12345-67890-ABCDE".

Answer: 1. Extract the first 5 characters using LEFT

Task: Extract the first 5 characters from "ExcelTipsAreGreat"

Formula:

=LEFT("ExcelTipsAreGreat", 5)

Result:

Excel

Explanation:

- LEFT(text, num_chars) takes characters from the **start (left)** of the string.

2. Extract the last 4 characters using RIGHT

Task: Extract last 4 characters from "DataAnalysis.xlsx"

Formula:

=RIGHT("DataAnalysis.xlsx", 4)

Result:

xlsx

Explanation:

- RIGHT(text, num_chars) takes characters from the **end (right)** of the string.

3. Extract the substring "Tips" using MID

Task: Extract "Tips" from "ExcelTipsAreGreat"

Formula:

=MID("ExcelTipsAreGreat", 6, 4)

Result:

Tips

Explanation:

- MID(text, start_num, num_chars) extracts a substring starting at **start_num** and spanning **num_chars** characters.
- Here, "Tips" starts at **6th character** and has **4 characters**.

4. Count total number of characters using LEN

Task: Count characters in "Hello World!"

Formula:

=LEN("Hello World!")

Result:

12

Explanation:

- LEN(text) returns the total number of characters **including spaces and punctuation**.

5. Extract middle 6 characters from "12345-67890-ABCDE"

Task: Extract middle 6 characters

Formula:

=MID("12345-67890-ABCDE", 6, 6)

Result:

-67890

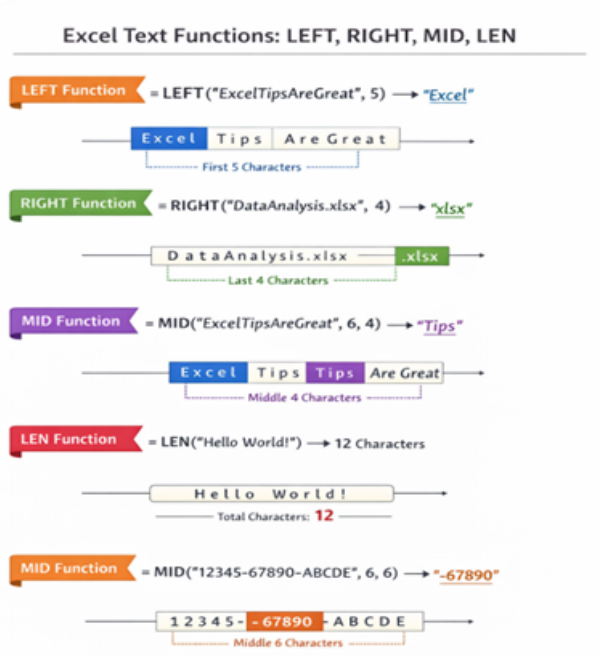
Explanation:

- Start at **6th character** ("-") and take **6 characters**: "-67890"

Summary Table-

Function	Formula	Result
LEFT	`=LEFT("ExcelTipsAreGreat",5)`	Excel
RIGHT	`=RIGHT("DataAnalysis.xlsx",4)`	xlsx
MID	`=MID("ExcelTipsAreGreat",6,4)`	Tips
LEN	`=LEN("Hello World!")`	12
MID (middle)	`=MID("12345-67890-ABCDE",6,6)`	-67890

A visual diagram is also stated below:



7. Understand how to combine text using CONCAT, TEXTJOIN, and the & operator.

   a. Use CONCAT to combine "Hello" and "World" with a space in between.

   b. Combine "Apple", "Banana", and "Cherry" into a single string separated by commas using TEXTJOIN.

   c. Use the & operator to create the string "2025: Excel Functions" by combining "2025", ": ", and "Excel Functions".

   d. Create a comma-separated list from the range A1:A5 using TEXTJOIN

   e. Combine first names in column A with last names in column B to create full names in column C.

Answer: 1. Using CONCAT-

Task: Combine "Hello" and "World" with a space in between.

Formula:

=CONCAT("Hello", " ", "World")

Result:

Hello World

Explanation:

- CONCAT merges multiple text strings.
- You can manually add spaces or other separators.

2. Using TEXTJOIN with a comma-

Task: Combine "Apple", "Banana", "Cherry" into a single string separated by commas.

Formula:

=TEXTJOIN(", ", TRUE, "Apple", "Banana", "Cherry")

Result:

Apple, Banana, Cherry

Explanation:

- TEXTJOIN(delimiter, ignore_empty, text1, text2, …)
- delimiter = ", " adds a comma and space between items
- ignore_empty = TRUE ignores blank cells

3. Using & Operator-

Task: Create "2025: Excel Functions" by combining "2025", ": ", and "Excel Functions".

Formula:

="2025" & ": " & "Excel Functions"

Result:

2025: Excel Functions

Explanation:

- & joins strings directly
- You can add spaces, punctuation, or other text manually

4. Creating a comma-separated list from a range using TEXTJOIN-

Task: Combine values in cells A1:A5 into one string separated by commas.

Formula:

=TEXTJOIN(", ", TRUE, A1:A5)

Result (example):

Value1, Value2, Value3, Value4, Value5

Explanation:

- Works for ranges
- Ignores empty cells if TRUE is used as the second argument

5. Combining first and last names into full names-

Task: Column A = First Name, Column B = Last Name → Column C = Full Name

Formula (for C1):

=CONCAT(A1, " ", B1)

Or using &:

=A1 & " " & B1

Result:

- If A1 = John, B1 = Doe → C1 = John Doe

Explanation:

- Adds a space between first and last name
- Can drag the formula down the column for all rows

Summary Table:

Function / Operator	Formula Example	Output	Notes
CONCAT	`=CONCAT("Hello"," ","World")`	Hello World	Combines multiple strings
TEXTJOIN	`=TEXTJOIN(", ",TRUE,"Apple","Banana","Cherry")`	Apple, Banana, Cherry	Allows delimiter & ignore blanks
& Operator	`"2025" & ": " & "Excel Functions"`	2025: Excel Functions	Simple concatenation
TEXTJOIN (Range)	`=TEXTJOIN(", ",TRUE,A1:A5)`	Value1, Value2,...	Combine multiple cells

CONCAT / &	`=CONCAT(A1," ",B1)` or `=A1 & " " & B1`	John Doe	Combine columns

8. Understanding TODAY() and NOW()

    a. What is the difference between TODAY() and NOW() in Excel? Provide an example of when you would use each function.

    b. If cell A1 contains the date 2025-06-10, write a formula using TODAY() to determine how many days are left until that date.

    c. Write an Excel formula using NOW() to display the current date and time in the format MM/DD/YYYY HH:MM AM/PM'.

    d. If a cell contains =TODAY(), what will happen when the worksheet is reopened the next day? Explain

    e. You want to store a static date (today's date) in a cell without it changing every day. What keyboard shortcut should you use.

Answer: a. Difference between TODAY() and NOW()

- **TODAY()**
    - Returns only the current date (no time).
    - Updates automatically whenever the worksheet recalculates.
    - Example use: Tracking deadlines, calculating age, or showing "Days until due date."
    - Formula example: =TODAY() → If today is 19-Dec-2025, it will show **19-Dec-2025**.

    - **NOW()**

    - Returns the current date and time.
    - Updates automatically whenever the worksheet recalculates.
    - **Example use:** Logging current time for reports, timestamps, or time-sensitive calculations.
    - Formula example: =NOW() → If the current date and time is 19-Dec-2025 8:20 PM, it will show 19-Dec-2025 20:20.
    - **Summary:** TODAY() = Date only; NOW() = Date + Time.

b. Days left until a specific date

- If A1 = 2025-06-10, the formula is:
- =A1-TODAY()
- This will return the number of days left from today until 10-Jun-2025.

- If today is 19-Dec-2025, the result will be negative because the date has already passed.

c. Display current date and time in a specific format

- Formula:
- =TEXT(NOW(),"MM/DD/YYYY HH:MM AM/PM")
- TEXT() lets you format the date/time as you want.
- Example output: 12/19/2025 08:20 PM.

d. What happens to =TODAY() the next day

- =TODAY() updates automatically each day.
- If you reopen the worksheet tomorrow, the cell will show tomorrow's date.
- **Explanation:** TODAY() is a volatile function, meaning it recalculates every time the sheet is recalculated or opened.

e. Storing a static date (does not change)

- Keyboard shortcut:
- Ctrl + ; (semicolon) → Inserts the current date as a fixed/static value.
- Ctrl + Shift + ; (semicolon) → Inserts the current time as a fixed/static value.