DATA CLEANING AND ESSENTIAL FUNCTIONS:-

QUESTION -1 What is data cleaning ,why Is it important in data analysis?

. what are the potential consequences of analyzing unclean or messy data?

.Explain the common steps involved in cleaning and organizing data.

ANSWER-1

Data cleaning is the process of finding and correcting errors, inconsistencies, or missing values in a dataset to ensure that the data is accurate , complete, and reliable for analysis.

IT TYPICALLY INCLUDES STEPS LIKE:-

1- Removing duplicates- deleting repeated entries
2- Handling missing values- filling in, estimating, or removing missing data
3- Correcting error- fixing typos, wrong formats, or incorrect entries
4- Standarlizing data- ensuring consistency (e.g; all dates in YYYY-MM-DD format)
5- Filtering outliers- identifying and handling unusual or extreme data points.
6- Validating data- checking that data meets defined rules or constraints.

WHY DATA CLEANING IS IMPORTANT:-

1- Accuracy of results- dirty data leads to incorrect conclusions and misleading in insights.
2- Better model performance -in machine learning, clean data improves model accuracy and generalization.
3- Efficiency-clean datasets reduce processing time and simplify analysis.
4- Consistency- ensure data from multiple sources aligns correctly
5- Better decision -making -Reliable data leads to trustworthy business or research decisions.

Analyzing unclean or messy data can lead to serious problem that affect the accuracy , reliability, and usefulness of your analysis.

(1) Inaccurate Reaults

From

. Errors, duplicate ,or missing values can distort averages , total ,and trends.

. Example: if sales data has duplicate entries , your total sales woll be overestimated.

(2) Misleading insights
. Faulty data can lead you grow the wrong conclusions.
.Example : you might believe a product is performing poorly due to incorrect data, leading to bad business decisions .

(3) Poor decision- making
. Decisions based on incorrect analysis can cause financial losses, wasted resources , or strategic errors.

. Example: A company might spend money on the wrong marketing campaign because of inaccurate customer data.

(4) Reduced model performance (in machine learning)
.Unclean data confuses models, causing low accuracy and biased predictions.
.Example : missing or mislabeled data leads to weak training and poor real – world performance.

(5) Loss of credibility
. if your analysis produces wrong or inconsistent results, people lose trust in your reports and skill.
.This is especially serious in scientific research ,finance ,or healthcare

(6) increased time and cost
.cleaning data after analysis often takes more efforts than cleaning it properly from the start.
.Analysis spend extra time fixing issue and rerunning analyses.

(7) Compliance and legal risks
. in regulated industries, inaccurate or incomplete data can lead to legal penalties or non – compliance with standards (like GDPR or HIPAA)

From

STEPS IN CLEANING AND ORGANISISNG DATA ARE :-

(1) Important and inspecting the data

. Load data from sources such as CSV files ,databases , APIs, or spreadsheet

. use summary function (e. g;) head() , info(), describe () to understand structure, types , and potential issues.

(2) Handling missing data

. Identify missing values using tools like isnull () or visualization.

. Decide a strategy , such as:

- Removing rows or columns with too many missing values .
- Filling (imputing) with mean/median/mode
- Forward/backward fill (for time series)
- Using mode model -based imputation if necessary .

(3) Removing or correcting outliers

- Detect outliers using statistical methods (Z- score, IQR) or visualizations (boxplots, scatterplots)
- Depending on context:

  . Remove them

  .cap values (Winsorization)

  .Investigate and correct if caused by data entry error.

(4) Fixing data types

. Convert numerical text data into numeric format

. parse data and times

. convert categories/ labels to categorical types

. Ensure consistent units (e.g., meters vs feet)

(5) Handling duplicate

. identify duplicate rows

.Remove or consolidate them depending or domain knowledge.

(6) Standardizing and cleaning text

. Trim whitespace

. Fix inconsistent capitalization

. Remove special characters if necessary

From

. Standardize formats (e.g., country name, product IDS)

(7)Features engineering (organizing for analysis)

* create new variables from existing ones

. Extract year /month / day from date

. combine related features

. Encode categorical variables (one -hot encoding , label encoding)

- Normalize or scale numerical features if required .

(8) Restructuring data

- Reshape data using :
  . pivot tables
  .melt/unpivot functions
  .Merging or joining datasets

(9) Validating cleaned data

.Double -check distributions , summary statistics , and relationships
.Validate against domain expectations (e.g., ages shouldn't be negative)

(10) Documenting the cleaning process

. keep notes or code (such as a Jupyter notebook
. Record decisions or results can be replicated .

QUESTION -2

How would you sort the following data sets first "DEPARTMENT" (A-Z) and then by "salary" (largest to smallest)? Write step by step approach

EMPLOYEE	DEPARTEMENT	SALARY
SONU	IT	4000
PRANAV	HR	5000

From

RAHUL	IT	2500

ANSWER:-

(1) Select the entire data table (include columns : employee, department , salary)

(2) Go to the data tab on the excel ribbon

(3) Click sort(no filter)

(4) A sort dialog box will open

(5) A sort dialog box will open

(6) In the first “sort by “ option , choose Department and set the order to A -Z

(7) Click on add level to apply another sorting condition

(8) In the “then by” option ,choose salary and set the order to largest to smallest

(9) Click OK to apply the multi -level sorting.

(10) The data will now be sorted first by department alphabetically and then by salary in descending order within each department.

QUESTION -3

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

ANSWER-3

1- TRIM()

PURPOSE:-

Removes extra spaces from text

HOW IT HELPS IN DATA CLEANING:

. Eliminates leading , trailing , and multiple spaces between words.

. prevents errors when matching , comparing , or merging text fields.

EXAMPLE-

TRIM (“ JOHN DOE”)-“ JOHN DOE”

2- LEFT ()

PURPOSE:

From

Extract characters from the beginning (left side) of a text string.

HOW IT HEPLS IN DATA CLEANING:

- Extract prefixes, codes , or initial segments.
- Useful for standardizing IDs or pulling out specific components.

EXAMPLE:

LEFT (“ A12345” , 1)- “A”

(3) Right ()

PURPOSE :

Extract characters from the end (right side) of a text string .

HOW IT HELPS IN DATA CLEANING :

- Retrieves suffixes, last digits , or category indicators.
- Helps isolate the end part of formatted text.

EXAMPLE:-

RIGHT (“INV- 2024 “ 4)- “2024”

(4) MID()

PURPOSE :

Extract from the middle of string , starting at a specified position for a specified number of characters .

HOW IT HELPS IN DATA CLEANING :

- Extract structured component like data segments , codes , or identifiers .
- Useful when the needed data is buried inside a longer string

EXAMPLE:-

MID(“ 2025- 11 – 15” , 6 , 2) – “11” (The month)

(6) CONCAT () OR CONCATENATE ()

PURPOSE :-

Joins two or more strings together

HOW IT HELPS IN DATA CLEANING –

- Combine fields into one (e.g., first name+ last name)
- Formats addresses , IDs, labels , and standardized entries .
- Helps reconstruct text after splitting and cleaning

EXAMPLE :-

CONCAT (“ JOHN “ , “ “ , “ DOE”) – “ JOHN DOE “

From

QUESTION -4

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

ANSWER-

1- AUTOMATING DATE – BASED CALCULATIONS
Ussing today () , you can automatically calculating things like :

- AGE
- Days until a deadline
- Days since an event
- Renewal or expiration dates

EXAMPLE :
=TODAY() -A2
(Finds how many days have passed since the date in cell A2)
This keeps your dataset always up to date without manual input

2- TRACKING TIME – SENSITIVE INFORMATION
Date functions help manage :

- Project progress
- Employee tenure
- Product shelf life
- Subscription or contact periods

This improve data accuracy and reduce the chance of outdated information .

3- CREATING DYNAMIC FILTERS AND REPORTS
Today () allows dashboards and reports to update automatically .
EXAMPLE :

- Show tasks due this week
- Highlight item that are overdue
- Filter transactions from the past 30 days

This is often used in :

- Sales reports
- Inventory management
- HR dashboards
- Finance and accounting summaries

4- ENCHANCING DATA VALIDATION AND QUALITY

Date functions help detect incorrect or impossible dates .

EXAMPLE :-

- Flag a future date in birthdate column
- Prevent entry of dates older than a specific threshold
- Validate date ranges automatically

This improves data integrity .

5- SUPPORTING TIME -BASED ANALYSIS

Using today() with formulas allows for :

- Trend analysis
- Forecasting
- Time series evaluations
- Period comparisons (e.g., month -to- date , year – to date)

These analyses depend on accurate and dynamic date calculations .

QUESTION -5

Apply data validation to restrict quantity values to only whole numbers between 1and 10

- Configure an input message that appears when a user selects a cell in the " QUANTITY" column explain " please enter a whole number between 1 and 10 "
- Set up an error alert message that triggers if the user enters a number less than 1 or greater than 10 ,

" 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 Davis	Sofa	clothing		239
Alex Moore	Shoes	Electronics		500
Michael Johnson	Table lamp	Home décor		423
Daniel Johnson	Backpack	Electronics		160
Isabella Davis	Headphones	Electronics		348
jane Davis	Headphones	Electronics		152
Alex Wilson	T-shirt	Home décor		369

From

ANSWER:-

1. Select the quantity column

- Click and drag to select all the cells under the quantity column

2. Open data validation

- Go to the data tab
- Click data validation (data tools group)

3. Set validation criteria

- In the data validation window

  . under settings tab:

  - Allow : whole number
  - Data : between
  - Minimum :1
  - Maximum:10

    This ensure that users can only enter whole numbers from 1 to 10

4 . configure input message

- Click the input message tab
- Tick " show input message when cell is selected "

  . fill in :-

- Title : quantity rule
- Input message : please enter a whole number between 1 and 10

5- configure Error Alert

- Click the error alert tab
- Tick " show error alert
- Choose style : stop

  . fill in :

  - Title : invalid input
  - Error message : invalid amount quantity must be a whole number between 1 and 10

From

6- Click ok

- The data validation is now applied to all selected cells .

QUESTION -6

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

- Extract the first 5 character from the string " ExcelTipsAreGreat" using the left fun
- Extract the last 4 character 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 character from "ExcelTipsAreGREAT" using left
   FORMULA: LEFT(ExcelTipsAreGreat",5)
   Result : Excel
2. Extract the last 4 characters from "DataAnalysis.xlsx" using RIGHT
   FORMULA: RIGHT("DataAnalysis.xlsx",4)
   Result: xlsx
   (if counting only letters , last 4 letters =" xlsx")
3. Extract the substring "Tips" from " ExceltipsAreGreat" using MID "Tips" start from characters number 6 and has 4 character
   FORMULA:=MID("ExcelTipsAreGreat",6,4)
   Result:Tips
4. Count total number of characters in "hello world!" using LEN

FORMULA:LEN(“hello world!”)
Result:12
(11 letters + 1 exclamation mark)

5. Extract the middle 6 characters from “12345-67890-ABCDE”
The string length is 17 character.
Middle starts at :(17/2)-2 =6(approx.)
But the middle 6 characters given visually are :
67890-
FORMULA:
MID(“12345-67890-ABCDE”,6,6)
Result :67890

QUESTION – 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 cells 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 everyday . what keyboard shortcut should you use ?

Answer

a) Difference between TODAY() and NOW()

1- TODAY()

- Returns only the current date
- Does not include time
- Example use: When you need age ,days remaining ,or deadlines.

2.NOW()

* Returns current date+ current time

* Example use : When you need time – based calculation like timestamps, attendance ,time ,billing ,etc

(B) if cell A1 contain the date 2025-06 -10 , find days left using TODAY()

FORMULA : A1-TODAY()

This subtracts today's date from the future date and gives days remaining

© Formula using NOW() to display date & time in format

MM/DD/YYYY HH:MM AM/PM**

FORMULA:

NOW()

Then apply custom format :

MM/DD/YYYY HH:MM AM/PM

(D) If a cell contain = TODAY() ,what happen when you reopen the worksheet the next day ?

* the value will automatically update to the new day's date .

* Reason : TODAY() is a volatile function that refreshes every time the file is opened

(E) To store a static date (not changing every day) , what keyboard shortcut should you use ?

Ctrl+;

From

This inserts today’s date once and it will not change later .

From

Customer's Name	product Name	Category	Quantity
Jane smith	Shoes	Electronics	
Isabella Moore	Laptop	Electronics	
Daniel Davis	Sofa	Clothing	
Alex Moore	Shoes	Electronics	
Michael Johnson	Table Lamp	Home décor	
Daniel Johnson	Backpack	Electronics	
Isabella Davis	Headphone	Electronics	
Jane Davis	Headphone	Electronics	
Alex Wilson	T-Shirt	Home décor	

Question -6

1- EXCEL

2- XLSX

3- TIPS

4- 12

5- 67890-

Question-7

1- HELLO WORLD

2- APPLE,BANANA,CHERRY

3- 25 : EXCEL FUNCTIONS

4-

APPLE
BANANA
CHERRY
MANGO
ORANGE

APPLE,BANANA,

5-

FIRST NAME	LAST NAME	RESULT
RIYA	SHARMA	RIYA SHARMA
KUNAL	MEHTA	KUNAL MEHTA
AISHA	RATHORE	AISHA RATHORE
TANU	CHAUHAN	TANU CHAUHAN

Unit price ($)
81
121
239
500
423
160
348
152
369

CHERRY,MANGO,ORANGE