**Data analyst internship Task-01**

**Data cleaning & pre-processing:**

The Analysis is performed on "Mall Customer Segmentation Data" Dataset.

1. There was no need for identifying and handling the missing values as every column was fully filled in the Dataset.

2. There are no duplicate rows, every row is unique.

The rest of the points or steps that were needed to be performed on the Dataset were done in Google Colab and the link is give below,

**Google Colab link:** <https://colab.research.google.com/drive/1nFR0EElxTWCPdFCif71V4M2RDGQVNIOC?usp=sharing>

Reference and Tools Used,

Google, Youtube, ChatGPT, Deepseek, Google Colab

**Interview Questions Related To Above Task:**

**1. What are Missing Values and How to Handle Them?**

Missing values are empty or blank cells in a dataset. They happen due to errors or unavailable data. If not handled, they can lead to wrong results.

How to handle them:

* Remove them using .dropna().
* Fill them using .fillna() with mean, median, or a default value like 0 or "unknown".
* Predict them using machine learning models for advanced cases.

**2. How do you treat duplicate records?**

Duplicate records are rows that are exactly the same across all columns. They can happen due to data entry mistakes or merging datasets.

How to handle them:

* Use df.duplicated() to find duplicates
* Use df.drop\_duplicates() to remove them and keep the first one

**3. Difference between dropna() and fillna() in Pandas?**

Both are used to handle missing values:

* **dropna()** removes rows (or columns) with missing values.
* **fillna()** fills missing values with a value like the mean, median, or a fixed number.

**Example:**  
If a row has a missing age:

* dropna() deletes the row.
* fillna(30) fills it with 30.

**4. What is Outlier Treatment and Why is it Important?**

Outliers are values that are very different from most of the data — like someone earning 10 million when others earn 30k–100k. They can distort averages and harm the accuracy of analysis or machine learning models.

How to handle outliers:

* Remove**:** Delete them if they’re clear errors.
* Cap/Floor**:** Limit them to a max or min value.
* Transform**:** Apply log or square root to reduce their impact.

**5. Explain the process of standardizing data?**

Standardizing data means making values consistent and uniform for easier analysis.

Steps include:

* Convert text to lowercase (e.g., "MALE" → "male")
* Remove extra spaces in text
* Format dates consistently (e.g., dd-mm-yyyy)
* Rename columns to lowercase with underscores (e.g., "Annual Income" → "annual\_income")

**6. How do you handle inconsistent data formats (e.g., date/time)?**

Inconsistent formats happen when the same data appears in different styles, e.g. 01-01-2024, 2024/01/01, or "Jan 1, 2024".

How to fix:

* Use pd.to\_datetime() to convert all to one standard format.
* Set a consistent format (like dd-mm-yyyy) for the whole column.
* Remove or correct entries that don't follow the format.

**7. What are common data cleaning challenges?**

Data cleaning often faces many issues that make the process time-consuming and complex.

Common challenges include:

* Missing or incomplete data
* Duplicates and redundancy
* Inconsistent formats (date, text, currency)
* Outliers and invalid values
* Mixed data types in a single column
* Typographical and spelling errors
* Hidden characters or extra spaces

**8. How can you check data quality?**

Data quality means how accurate, complete, and reliable the data is. You must check data quality before analyzing it to ensure results are valid.

Quick ways to check data quality:

* df.info() – See data types and missing values.
* df.describe() – Spot unusual or extreme values.
* df.duplicated().sum() – Count duplicate rows.
* Use histograms/boxplots – Detect outliers visually.
* Check formats – Make sure dates, text, and numbers are consistent.
* Review column names and data types – Keep them clean and uniform.