Guided Internship Notebook: Reading & Subsetting Data in Python

1. Loading & Previewing Data

- **CSV Data**: Loaded using pandas.read_csv(), displaying first five rows.
- **JSON Data**: Read using json.load(), flattened with pandas.json_normalize().

2. Data Exploration Tasks

- Checking column names using df_csv.columns.
- Identifying missing values with df_csv.isnull().sum().
- Aggregating sales by product to identify the top five highest-selling items.

3. Subsetting Techniques

- **Position-Based**: Extracting specific rows and columns using iloc.
- **Label-Based**: Filtering rows based on outlet identifier using loc.
- **Value-Based**: Selecting products with sales over 5000 in the "Snack Foods" category.

4. Data Modifications

- **New Column**: Calculating estimated profit as Sales - MRP.
- **Renaming Columns**: Changing 'Item_Weight' to 'Weight_kg'.
- **Handling Missing Data**: Dropping null values for clean analysis.

5. Additional Insights

- Identifying items sold in the highest number of outlets.
- Detecting sales outliers using a **boxplot** visualization.

Final Summary Report

- Loaded data files and previewed structure.
- Performed three subsetting operations for analysis.
- Modified dataset by adding a new column, renaming fields, and handling missing values.
- Extracted insights on top-performing products and key metrics.
- Visualized trends using **Seaborn** and **Matplotlib**.

Now, you can copy and paste this text into **Microsoft Word** or **Google Docs**, format it to fit one page, and save it as a PDF. Let me know if you need any refinements!