Professional Report on Data Cleaning Process for "restaurants.csv", "orders.csv", and "users.csv" Datasets

---

Objective:

This report outlines the cleaning and preparation steps undertaken for three datasets ("restaurants.csv," "orders.csv," and "users.csv") in Google Sheets. The goal was to standardize, organize, and clean the data to facilitate further analysis. The process involved several key actions such as renaming columns, cleaning data, and applying formulas to ensure consistency, accuracy, and readiness for analysis.

Cleaning the Spreadsheets

Professional Report on Cleaning the "restaurants.csv" Spreadsheet

Objective:

To clean and prepare the "restaurants.csv" dataset for analysis by organizing and standardizing the data.

Actions:

1. **Downloading and Preparing the Spreadsheet**
    a. **Action:** Download the "restaurants.csv" file and upload it to Google Sheets for easy editing.
    b. **Objective:** Facilitate further cleaning and manipulation using Google Sheets functions.
2. **Reformatting the "Entry Number" Column**
    a. **Action:** Move Column A by one row, so the numbers start from #1.
    b. **Objective:** Establish a standardized and sequential order for each restaurant entry, simplifying sorting and allowing easy reversion to the original form.
    c. **Implementation:** Rename the first column to "Entry Number" for clarity.
3. **Renaming the "id" Column**
    a. **Action:** Rename the "id" column to "restaurant_id" to ensure consistency with other datasets.
    b. **Objective:** Ensure uniformity across datasets and improve clarity by providing a consistent identifier.
    c. **Implementation:** Update the column header to "restaurant_id."
4. **Cleaning the "Cost" Column (Column G)**

a. **Action:** Remove the "₹" symbol from the values in Column G to isolate numeric values.
b. **Objective:** Standardize the cost data for more straightforward numerical analysis.
c. **Formula:**
    i. `=ARRAYFORMULA(SUBSTITUTE(G2:G, "₹ ", ""))`
d. **Implementation:** Place this formula in Column M, renaming it to "clean_cost."

5. **Cleaning the "Address" Column (Column K)**
    a. **Action:** Capitalize the first letter of each word in the address column to ensure consistency.
    b. **Objective:** Standardize the address format for uniformity.
    c. **Formula:**
        i. `=ARRAYFORMULA(PROPER(K2:K))`
    d. **Implementation:** Place this formula in Column N, renaming it to "clean_address."
6. **Cleaning the "City" Column (Column D)**
    a. **Action:** Remove any leading or trailing white spaces in city names.
    b. **Objective:** Prevent errors due to whitespace and standardize city names.
    c. **Formula:**
        i. `=ARRAYFORMULA(TRIM(D2:D))`
    d. **Implementation:** Place this formula in Column O, renaming it to "clean_city."
7. **Cleaning the "Cuisine" Column (Column H)**
    a. **Action:** Split the cuisine data into two columns for primary and secondary cuisines.

b. **Objective:** Standardize cuisine data for more straightforward analysis.

c. **Formula:**

i. `=ARRAYFORMULA(SPLIT(H2:H, ","))`

d. **Implementation:** Place this formula in Column P and Column Q, renaming them to "clean_cuisine_main" and "clean_cuisine_secondary."

8. **Cleaning the "Rating" Column (Column E)**

a. **Action:** Change the value of "--" to 1 and convert the column to number format.

b. **Objective:** Provide a consistent baseline value for missing ratings.

c. **Formula:**

i. `=ARRAYFORMULA(IF(E2:E = "--", 1, E2:E))`

d. **Implementation:** Place this formula in Column R, renaming it to "clean_rating."

9. **Cleaning the "Rating Count" Column (Column F)**

a. **Action:** Replace "Too Few Ratings" with 1 and extract numeric values from strings like "50+ ratings."

b. **Objective:** Standardize the rating count data for numerical analysis.

c. **Formula:**

i. `=ARRAYFORMULA(IF(F2:F = "Too Few Ratings", 1, VALUE(REGEXEXTRACT(F2:F, "\d+"))))`

d. **Implementation:** Place this formula in Column S, renaming it to "clean_rating_count."

10. **Creating the "Popularity Score" Column**

a. **Action:** Multiply the rating by the logarithm of the rating count to generate the "popularity_score."

b. **Objective:** Provide a more balanced weighting to ratings with more data.

c. **Formula:**

   i. =ARRAYFORMULA(R2:R * LOG(S2:S))

d. **Implementation:** Create a new column for the "popularity_score."

Summary of Actions Taken:

- **Reformatted Columns:**
  - Adjusted the "Entry Number" column.
- **Standardized Data:**
  - Removed currency symbols.
  - Capitalized the first letter of each word in addresses.
  - Trimmed extra spaces in city names.
  - Split and standardized cuisine data.
  - Replaced missing or erroneous ratings.
  - Extracted numeric values from rating counts.
- **Created New Columns:**
  - "clean_cost"
  - "clean_address"
  - "clean_city"
  - "clean_cuisine_main"
  - "clean_cuisine_secondary"
  - "clean_rating"
  - "clean_rating_count"

- "popularity_score."

Professional Report on Cleaning the "orders.csv" Spreadsheet

Objective:

To clean and prepare the "orders.csv" dataset for analysis by standardizing key columns and ensuring correct formatting.

Actions:

1. **Downloading and Preparing the Spreadsheet**
    a. **Action:** Download the "orders.csv" file and upload it to Google Sheets for easy editing.
    b. **Objective:** Facilitate further cleaning and manipulation using Google Sheets functions.
2. **Reformatting the "Entry Number" Column**
    a. **Action:** Move Column A up by one row so that the numbers start from #1.
    b. **Objective:** Establish a standardized and sequential order for each order.
    c. **Implementation:** Rename the first column to "Entry Number."

3. **Renaming the "r_id" Column**
    a. **Action:** Rename the "r_id" column to "restaurant_id" for consistency.
    b. **Objective:** Ensure uniformity across datasets.
    c. **Implementation:** Update the column header to "restaurant_id."

4. **Ensuring the "Order Date" is Properly Formatted**
    a. **Action:** Format the "order_date" column to YYYY-MM-DD.
    b. **Objective:** Standardize the date format for easier time-based analysis.
    c. **Implementation:** Use the Google Sheets date formatting feature.
5. **Ensuring Sales Quantity ("sales_qty") is Numeric**
    a. **Action:** Ensure that the "sales_qty" column is numeric.
    b. **Objective:** Ensure proper numerical analysis and calculations.
    c. **Implementation:** Change the format to "Number."
6. **Ensuring Sales Amount ("sales_amount") is Numeric**
    a. **Action:** Ensure that the "sales_amount" column is numeric.
    b. **Objective:** Standardize sales data for accurate calculations.
    c. **Implementation:** Change the format to "Number."

Summary of Actions Taken:

- **Reformatted Columns:**
    - Adjusted the "Entry Number" column.
- **Renamed Columns:**
    - Changed "r_id" to "restaurant_id."
- **Standardized Date Format:**
    - Formatted "order_date" to YYYY-MM-DD.
- **Ensured Numeric Formatting:**
    - Changed "sales_qty" and "sales_amount" to "Number" format.

Professional Report on Cleaning the "users.csv" Spreadsheet

Objective:

To clean and prepare the "users.csv" dataset for analysis, focusing on standardizing the "monthly_income" column and renaming columns for consistency.

Actions:

1. **Downloading and Preparing the Spreadsheet**
    a. **Action:** Download the "users.csv" file and upload it to Google Sheets for easy editing.
    b. **Objective:** Facilitate further data manipulation and cleaning.
2. **Reformatting the "Entry Number" Column**
    a. **Action:** Move Column A up by one row so the numbers start from #1.
    b. **Objective:** Establish a standardized and sequential order for each user.

    Implementation: Rename the first column to "Entry Number."

3. **Renaming Columns for Clarity and Consistency**
    a. **Action:** Rename the columns to ensure clarity and consistency.
    b. **Objective:** Improve readability and standardization of the dataset.
    c. **Implementation:**
        i. "marital status" → "marital_status"
        ii. "family size" → "family_size"

iii. "monthly income" → "monthly_income"

iv. "educational qualifications" → "educational_qualifications"

4. **Cleaning the "Monthly Income" Column (Column E)**
    a. **Action:** Convert text categories into numerical ranges for standardization.
    b. **Objective:**
    c. **Standardization for Analysis:** Facilitate statistical analysis by converting text to numeric values.
    d. **Improved Data Consistency:** Ensure consistent and uniform representation of income data.

    Formula:

    i. =ARRAYFORMULA(IF(J2:J = "", "",

    IF(J2:J = "No Income", 0,

    IF(J2:J = "Below Rs.10000", 5000,

    IF(J2:J = "10001 to 25000", 17500,

    IF(J2:J = "25001 to 50000", 37500,

    IF(J2:J = "More than 50000", 60000, NA()))))))

    e. **Implementation:** Place the formula in Column M and rename it "clean_monthly_income."

Summary of Actions Taken:

- **Reformatted Columns:**

    - Adjusted the "Entry Number" column.
- **Renamed Columns:**
    - Standardized column names for clarity.
- **Standardized "monthly_income":**
    - Converted income categories into numerical values for more straightforward analysis.
- **Created "clean_monthly_income":**
    - Generated a new column for the cleaned data.

Conclusion:

The cleaning process for the "restaurants.csv," "orders.csv," and "users.csv" datasets involved a series of steps to ensure data consistency and improve their usability for analysis. By renaming columns, standardizing data formats, and applying necessary transformations, the datasets are now ready for detailed analysis, ensuring accuracy and providing a solid foundation for future insights.