Data Cleaning Documentation

Indian Food Cuisine Dataset

## Introduction:

Data cleaning is a crucial step in the data preparation process, ensuring that the dataset is accurate, consistent, and usable for analysis. This documentation outlines the steps taken to clean the sample dataset, including examples for reference.

## 1. Dataset Overview

**Dataset Name:** Indian Food Cuisine Dataset

**Source:** [Indian Food Cuisine Dataset.xlsx](https://ripetizi-my.sharepoint.com/:x:/g/personal/satyaveer_ripetizi_onmicrosoft_com/EW-2-feZRPtGrOU9f99x5HwBqjYgF3U2HhviXGloA4yDGA?rtime=I7jbrQ7c3Ug)

**Content Type:** Excel File(.xlsx)

**Purpose:** To clean and prepare recipe data for accurate analysis of Indian and international cuisines.

**Total Records:** 6056

**Total Columns**: 13

**Columns Present:** name\_of\_Dish, Diet\_Type, Course\_name, Discrption\_of\_Dish, Cuisine\_name, Ratings\_of\_Dish, Similar\_Dishes, Ingredients\_of\_Dish, Prepration\_tie, Cooking\_tie, Total\_tie, Validations, Makes, Recipe\_Instructions

## 2. Initial Data Issues Identified

1. Column names contain inconsistent spelling (e.g., 'Prepration\_tie' instead of 'Preparation\_time').
2. Special characters / encoding issues ('Â' present in Discrption\_of\_Dish).
3. Data type mismatches (e.g., cooking time stored as text in larger dataset).
4. Inconsistent casing (e.g., 'Vegetarian' vs 'vegetarian').
5. Possible missing values in ***Similar\_Dishes*** column (empty lists).
6. Long text fields without proper wrapping (***Recipe\_Instructions***).
7. Duplicate Records Found in ***name\_of\_Dish*** column

## 3. Cleaning Steps Taken

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Step**  **No** | **Column**  **Name** | **Cleaning Action** | **Method/Formula**  **Excel** | **Before**  **Example** | **After**  **Example** | **Records**  **Affected** |
| 1 | name\_of\_Dish | Identify the Duplicate | =COUNTIF(A2:$A$6056,A2) & Conditional Formatting Duplicate Values | Avocado Mint Soup Recipe  Avocado Mint Soup Recipe | Avocado Mint Soup Recipe | 69 |
| 2 | Diet\_Type | Removed the “Diet:” | Replace the value with empty/blank value | Diet: Vegetarian | Vegetarian | 5987 |
| 3 | Course\_name | Removed the “Course:” | Replace the value with empty/blank value | Course: Lunch | Lunch | 5987 |
| 4 | Discrption\_of\_Dish | Remove encoding issues | =SUBSTITUTE(D2,"Â","") |  |  | 2656 |
| 5 | Cuisine\_name | Remove ['Cuisine, \ufeff'] characters issues | Replace the value with Space and Empty/Blank | ['Cuisine: Gujarati Recipes\ufeff'] | Gujarati Recipes | 5987 |
| 6 | Ratings\_of\_Dish | Consider the left value & average of those values | Average value of Cleaned numbers | "56, "53", "57", "57", 45, "54", "53" & "55" | 4.9 | 8 |
| 7 | Prepration\_time, Cooking\_time, Total\_time | Convert data types & Remove ‘M’ | Data → Text to Columns | 10M | 10 | 3 |
| 8 | All Columns | Trim extra spaces | =TRIM(A1) | Biryani | Biryani | 1 |
| 9 | Entire Data Set all columns | Wrap text for readability | Format → Wrap Text | To begin making... | Same (wrapped) | 4 |

## 4. Data Validation Rules Applied

* Diet\_Type: Allowed values = {Vegetarian, Non-Vegetarian, Vegan, High Protein Vegetarian}
* Preparation\_time, Cooking\_time, Total\_time: Must be positive integers only
* Prepration\_time + Cooking\_time = Total\_time
* Ratings\_of\_Dish: Must be between 0 and 5
* Cuisine\_name: Dropdown list with standardized cuisine names
* No blank name\_of\_Dish allowed

## 5. Final Summary

- Total records after cleaning: 5987

- Duplicates removed: 69

- Missing values handled: 0

- Columns renamed: 0

- Text encoding fixed: 2656 instances

- Outcome: Dataset is ready for analysis with consistent formatting, correct data types, and improved readability.