Data Cleaning Documentation

Overview

This document outlines the data cleaning process, detailing the corrections made for spelling mistakes and inconsistencies in the dataset, as well as handling missing values.

1. Standardization

Whitespace was stripped from column names and data entries, and case was standardized for uniformity.

2. Fixing Typos and Inconsistencies

a. Brand Names

- **Typos Identified:**
    - "Caprce" was corrected to "Caprice"
    - "Royal Stallion " (with extra space) was corrected to "Royal Stallion"
    - "GoldenHarvest" was corrected to "Golden Harvest"
    - "Mamma Gold" was corrected to "Mama Gold"
    - "Uncl Sam" was corrected to "Uncle Sam"
    - "Royal Stallon" was corrected to "Royal Stallion"
    - "King's Pride" remained unchanged as it was already correct.

b. Grain Types

- **Typos Identified:**
    - "basmati" (lowercase) was corrected to "Basmati" (capitalized)
    - "BASMATI" (uppercase) was corrected to "Basmati"
    - "long grain" (lowercase) was corrected to "Long Grain"
    - "LONG GRAIN" (uppercase) was corrected to "Long Grain"
    - "local" (lowercase) was corrected to "Local"
    - "LOCAL" (uppercase) was corrected to "Local"
    - "Long grain" (capitalized) was corrected to "Long Grain"

c. City Names

- **Typos Identified:**
    - "Kumaci" was corrected to "Kumasi"

- "Acccra" was corrected to "Accra"
- All other city names were verified and found to be correct.

d. Month Sold

- **Typos Identified:**
  - Month names were verified, and no specific typos were found. All entries were standardized to ensure consistency.

e. Country Names

- **Typos Identified:**
  - The entries were verified, and no discrepancies were found. All country names were confirmed as correct.

3. Handling Missing Values

An initial check revealed 43 missing values in the "Price per Bag (Naira)" column. Missing prices were filled using the average price based on brand and grain type, followed by using the overall average for any remaining missing values.

4. Feature Engineering

New features were created to enhance the dataset, including revenue calculations, indicators for premium grains, categorization by quarter, log transformations, and brand presence indicators.

Conclusion

This documentation details the corrective actions taken to ensure the dataset is accurate and consistent, facilitating reliable analysis and insights.