Report on Data Differences Between Cleaned and Non-Cleaned Datasets

This report outlines the observed discrepancies between the cleaned and non-cleaned datasets for Products, Stores, and Sales. The differences are categorized into three main sections: **Inconsistent Case**, **Different Values**, and **Different Date Values**. Additionally, there is a section on rows present in the non-cleaned dataset but missing in the cleaned dataset.

---

1. Inconsistent Case

These differences arise when the same key (like `Category Name`, `Item Description`, etc.) exists in both datasets but with inconsistent casing (e.g., uppercase vs. capitalized).

Products:

- **Category_Name_Diff:**
  - **Percentage:** 100%
  - **Example:** `TEMPORARY & SPECIALTY PACKAGES / Temporary & Specialty Packages`
  - **Explanation:** Every entry in the `Category Name` field differs between the two datasets because the non-cleaned dataset uses capital case (e.g., `TEMPORARY & SPECIALTY PACKAGES`), while the cleaned dataset uses a capitalized format (e.g., `Temporary & Specialty Packages`).
- **Item_Description_Diff:**
  - **Percentage:** 99%
  - **Example:** `SKREWBALL PEANUT BUTTER WHISKEY / Skrewball Peanut Butter Whiskey`
  - **Explanation:** Almost all item descriptions differ in casing, with the non-cleaned data typically being in uppercase and the cleaned data in a more standardized capitalized format.
- **Vendor_Name_Diff:**
  - **Percentage:** 56%
  - **Example:** `JIM BEAM BRANDS / Jim Beam Brands`
  - **Explanation:** A little over half of the vendor names show differences due to casing, with the non-cleaned dataset using uppercase and the cleaned dataset using capitalized vendor names.

Stores:

- **Name_Diff:**
  - **Percentage:** 99%
  - **Example:** `NEIGHBORHOOD TOBACCO OUTLET / MARION / Neighborhood Tobacco Outlet / Marion`
  - **Explanation:** Nearly all store names have discrepancies due to differences in casing, where the non-cleaned dataset uses uppercase, while the cleaned version is capitalized.
- **Address_Diff:**
  - **Percentage:** 100%
  - **Example:** `1000 73RD ST / 1000 73rd St`

    - **Explanation:** Every address differs because the cleaned dataset uses proper capitalization, while the non-cleaned dataset uses uppercase.

- **City_Diff:**
    - **Percentage:** 100%
    - **Example:** `CEDAR RAPIDS / Cedar Rapids`
    - **Explanation:** All city names differ in case between the datasets, with the non-cleaned data using uppercase and the cleaned data using proper capitalization.

- **Store_Status_Diff:**
    - **Percentage:** 8%
    - **Example:** `I / A (7%) & A / I (1%)`
    - **Explanation:** Some store statuses show a difference in casing or possibly a misunderstanding in the status codes.

Sales:

- **Item_Description_Diff:**
    - **Percentage:** 95%
    - **Example:** `SKREWBALL PEANUT BUTTER WHISKEY / Skrewball Peanut Butter WHISKEY`
    - **Explanation:** The majority of item descriptions differ due to case inconsistencies, with the cleaned dataset generally adhering to a standardized format.

- **Category_Name_Diff:**
    - **Percentage:** 31%
    - **Example:** `cocktails/rtd / cocktails /rtd`
    - **Explanation:** Differences in category names are often due to variations in spacing, punctuation, or pluralization in addition to casing.

- **County_Diff:**
    - **Percentage:** 25%
    - **Example:** `pottawattamie / pottawatta`
    - **Explanation:** Certain county names are different, potentially due to typos or truncation issues in the non-cleaned data.

---

2. Different Values

This section covers cases where the same key exists in both datasets but the values are different.

Products:

- **Age_Diff:**
    - **Percentage:** <1%
    - **Example:** `0 / 3`
    - **Explanation:** Minor differences in age values are found, possibly due to data entry errors or updates in the cleaned dataset.

- **Bottle_Volume_ml_Diff:**

    - **Percentage:** <0.1%
    - **Example:** `700 / 750`
    - **Explanation:** Small differences in bottle volumes could be due to changes in packaging sizes over time or corrections in the cleaned data.

- **Inner_Pack_Diff:**

    - **Percentage:** <0.1%
    - **Example:** `12 / 10`
    - **Explanation:** Discrepancies in inner pack sizes are minimal, likely due to packaging changes or data corrections.

- **Proof_Diff:**

    - **Percentage:** <0.1%
    - **Example:** `60 / 70`
    - **Explanation:** Variations in proof values are rare and could be due to differences in product versions or updates in the cleaned dataset.

- **UPC_Diff & SCC_Diff:**

    - **Percentage:** 41%
    - **Example:** `80432106624 / 89540508818`
    - **Explanation:** These differences indicate significant inconsistencies in product identification codes, potentially pointing to errors in data entry or differences in how products were cataloged.

- **State_Bottle_Retail_Diff & State_Case_Cost_Diff:**

    - **Percentage:** 24% & 23%
    - **Example:** `12.0 / 11.25`
    - **Explanation:** Retail and case cost discrepancies suggest pricing adjustments or errors that were corrected in the cleaned dataset.

Stores:

- **Zip_Diff:**

    - **Percentage:** 0.1%
    - **Example:** `50314 / 50315`
    - **Explanation:** Minor differences in ZIP codes could be due to data entry errors or updates.

- **Report_Date_Diff:**

    - **Percentage:** 100%
    - **Example:** `2024-07-01 / 2022-10-01`
    - **Explanation:** All report dates differ, possibly indicating different data collection periods or corrections.

Sales:

- **County_Number_Diff:**

    - **Percentage:** 50%
    - **Example:** `07 / 7`

    - **Explanation:** Half of the county numbers differ due to leading zeros being dropped in one dataset.

- **Sale_Dollars_Diff:**

    - **Percentage:** 0.1%
    - **Example:** `12.0 / 11.25`
    - **Explanation:** Small discrepancies in sale amounts could indicate rounding errors or pricing corrections.

- **Volume_Sold_Liters_Diff:**

    - **Percentage:** 0.1%
    - **Example:** `1 / 2`
    - **Explanation:** Differences in volume sold are minimal and may result from rounding or data corrections.

---

3. Different Date Values

These discrepancies involve cases where dates differ between the two datasets.

Products:

- **List_Date_Diff:**

    - **Percentage:** 3.5%
    - **Example:** `2023-09-01 / 2022-09-01`
    - **Explanation:** A small percentage of list dates differ, likely due to corrections or updates in the cleaned dataset.

- **Report_Date_Diff:**

    - **Percentage:** 100%
    - **Example:** `2024-07-01 / 2022-10-01`
    - **Explanation:** All report dates differ, indicating different data capture periods or retrospective updates.

---

4. Rows in Non-Cleaned But Not In Cleaned

This section covers rows that exist in the non-cleaned dataset but are missing from the cleaned dataset.

Products:

- **Missing Rows:**
    - **Total:** 6 rows
    - **Example IDs:** `934600`, `933935`, `917640`, `903980`, `21540`, `80438`
    - **Explanation:** These rows could have been removed during the cleaning process due to being duplicates, irrelevant, or containing errors that couldn't be corrected.