DATA CLEANING REPORT

Sprocket Central Pty Ltd

By: Akpoveso Maryrose

Date: May 6, 2025

Table of Contents

1.) Introduction
- Company Overview
- Dataset Overview

2.) Before and After Cleaning Summary
- Dataset Structure
- Row and Column Changes

3.) Key Data Issues Identified
- Missing/Null Values
- Inconsistent Entries & Standardization
- Format Mismatches

4.) Data Cleaning Process
- Transactions Sheet
  - Removed columns
  - Data Format fixes
  - Renaming columns
  - Handling Missing Values
  - New Column
  - Column Standardization
- Customer Demographic Sheet
  - Handling Missing Values
  - Column Standardization
  - Data Format fixes
  - New Column created
  - Dropped Columns
- Customer Address Sheet
  - State Standardization
  - Property Valuation notes
- New Customer List Sheet
  - Handling Missing Values
  - Column Standardization
  - Data Format fixes

- New Columns created
- Tenure Column Review
- Rank & Value Column Notes

5.) Tools Used
6.) Conclusion
7.) Additional Resources

Introduction

Sprocket Central Pty Ltd is a bicycle sale business located in Australia and operating in 3 states.

Dataset Overview

The dataset contains 4 sheets: Transactions, New Customer List, Customer Demographic and Customer Address.

Before and After Cleaning Summary

Dataset Structure

Transactions sheet: 13 columns, 20001 rows with headers.

New Customer List: 18 columns, 1001 rows with headers.

Customer Demographic: 13 columns, 4001 rows with headers.

Customer Address: 6 columns, 4000 rows with headers.

Row and Column Changes

Clean transactions sheet: 15 columns, 19804 rows with headers.

Full transactions sheet: 12 columns, 20001 rows with headers.

New Customer List: 19 columns, 1001 rows with headers.

Customer Demographic: 13 columns, 4001 rows with headers.

Customer Address: 6 columns, 4000 rows with headers.

State name & Abbreviation: 2 columns, 9 rows with headers.

Key Data Issues Identified

- Missing/Null data
- Inconsistent entries & Standardization issues
- Data format mismatches

Data Cleaning Process

Transactions Sheet

1) **Removed columns:** The product id column was removed from analysis due to inconsistency in associated attributes(e.g. brand and product details) making it an unreliable identifier. For example, a single product id not having consistent entries in either the brand or product details columns.
2) **Data Format fixes:** Some data formats were changed during the cleaning process. They included:
    - **Transaction date column:** The format was changed from custom to date.
    - **Product first sold date:** The format was changed from general to date.
    - **Standard cost column:** The format was changed from currency to number in order to avoid issues when exporting the dataset to other tools.
    - **List price column:** The format was changed from general to number.
3) **Renaming columns:** To enhance clarity and improve interpretability, the original "list_price" column was changed to "sale_price"
4) **Handling missing values:** There are missing values in numerous columns. Columns and their treatment include:
    - **Online_Order column:** This column had 360 missing values which accounted for 1.8% of the dataset. Given the negligible number of missing values, the decision was made to fill the missing cells with "Unknown".
    - **Multiple columns were affected by missing values:** The "brand" column had 197 missing values, which extended to other critical product attributes such as product_line, product_class, product_size, standard cost and product_first_sold_date. While the percentage of missing values may seem negligible in isolation, their combines absence rendered the records non-actionable. Therefore, the decision was made to exclude these rows from the analysis.
5) **New columns:** An order combination column was created to aid deeper analysis. The column was a combination of the brand, product_line,

product_class and product_size columns. The formula used to achieve this was =TRIM(CONCATENATE(I2, " ",J2," ",K2," ",L2)).

6) **Column Standardization:** To ensure uniformity in the online_order column, True was replaced by "Online" and False was replaced by "Offline".

Customer Demographic Sheet

1) **Missing values:** There are missing values in some columns. The columns and their treatment include:
    - **Last Name Column:** The Last_name had 125 missing values. As a result, the First_name and Last_name columns were merged into a single Full_name column to better distinguish between. The First_name and Last_name columns were deleted thereafter. This approach ensures that entries with partial name information are not lost or misinterpreted. The formula used to achieve this was : =TRIM(CONCATENATE(B4," ", C4))
    - **DOB column:** This column had 87 missing values. The decision was made to leave the blank columns intact. Reason being that the column would be used to create a new column for analysis.
    - **Age column:** This column had 87 missing values. The decision was made to replace missing values with "Missing".
    - **Job_Industry column:** This column had 506 missing values. The decision was made to replace missing values with "Unknown".
2) **Column Standardization:** To ensure uniformity in some columns, they underwent a standardization process:
    - **Gender column:** This column contained a variety of unstandardized values and was cleaned accordingly; "F" and "Femal" were rewritten as "Female", "M" was rewritten as "Male" and "U" was rewritten as "Unknown"
    - **Job_title column:** This column included variations of the same role differentiated by Roman numerals (e.g., "Budget/Accounting Analyst I", "Budget/Accounting Analyst IV"). Since there was no consistent or meaningful distinction between these levels in the dataset, and no additional data to support their hierarchy, all Roman numerals were

removed to standardize job titles for clearer grouping and analysis. This was achieved using this formula: =TRIM(REGEXREPLACE(G2,"\s+(I{1,3}|IV|V)$", "")).

- **Job_Industry column:** Entries labeled “n/a” in this column were standardized to “Unknown” for clarity and consistency, avoiding ambiguity in interpretation.

3) **Data Format fixes:** Some data formats were changed during the cleaning process. They included:
    - DOB column: The format was changed from custom to date.
4) **New Columns:** Some new columns were created to aid the analysis process. They included:
    - **Age column:** The data was extracted from the DOB column using the formula =IF(ISBLANK(E2), " ", 2017-YEAR(E2)) and formatting the column to Number. A single row returned a #VALUE error because excel does not recognize dates pre1990’s. This was resolved using the formula =2017 - VALUE(RIGHT(E35,4))
    - **Age category column:** The data was extracted from the age column using the formula =IFS(F2<18, "Under 18",F2<=29, "Young Adults",F2<=44, "Adults",F2<=59, "Middle-Aged Adults",F2>=60, "Seniors"). For columns where the age column was missing (87 values), they were filled with “Unknown”.
    - **Year column:** The data was extracted from the transaction_date column using the formula =YEAR(C2).
    - **Month number column:** The data was extracted from the transaction_date column using the formula =MONTH(C2).
    - **Month name column:** The data was extracted from the transaction_date column using the formula =TEXT(C2,"mmmm").
5) **Dropped Columns:** The default column contained data of multiple symbols. It was dropped due to the data being incomprehensible and therefore unusable for analysis.

Customer Address sheet

1) **State Column Standardization:** The state column contains both full names (e.g., "New South Wales") and abbreviations (e.g., "NSW") for Australian states. Since the dataset pertains to Australia, the assumption was made that all abbreviations relate to Australian states. A new sheet named "State name & Abbreviation" that includes a full list of Australian states and their corresponding abbreviations was sources and created from Wikipedia.*(see sources section)*. Abbreviated entries were manually replaced with their full names using a filter. The reference sheet serves as a guide to ensure accuracy during the standardization process.
2) **Property Valuation column:** This column contains values from 1-12 but there is no supporting information to clarify what these values represent. Given its presence in an address focused sheet and lack of metadata, the column was retained for completeness but excluded from any key analysis or dashboard visuals.

New Customer List Sheet

1) **Missing Values:** There are missing values in some columns. The columns and their treatment include:
    - Last Name Column: This column contained 29 missing values. The decision made to address this issue was the combination of the First_name and Last_name column to create a Full_name column. The formula used to accomplish this was =TRIM(CONCATENATE(A2," ",B2)). The First_name and Last_name columns were deleted thereafter.
    - Age Column: This column had 17 missing values. The decision was made to replace missing values with "Missing".
    - Job Title column: This column had 106 missing values. The decision was made to replace missing values with "Unknown".
2) **Column Standardization:** To ensure uniformity in some columns, they underwent a standardization process. The columns and their treatment include:

- **Gender column:** To standardize this column, entries written as “U” were replaced with “Unknown”.
- **Job Title column:** This column included variations of the same role differentiated by Roman numerals (e.g., “Budget/Accounting Analyst I”, “Budget/Accounting Analyst IV”). Since there was no consistent or meaningful distinction between these levels in the dataset, and no additional data to support their hierarchy, all Roman numerals were removed to standardize job titles for clearer grouping and analysis. This was achieved using this formula =TRIM(REGEXREPLACE(F2,"\s+(I{1,3}|IV|V)$","")).
- **Job Industry column:** Entries labeled “n/a” in this column were standardized to “Unknown” for clarity and consistency, avoiding ambiguity in interpretation.
- **State column:** The state column was standardized using the values from the State name & Abbreviation column. The formula used to achieve this was =XLOOKUP(NewCustomerList!N2,'State name & Abbreviation'!$F$3:$F$10,'State name & Abbreviation'!$E$3:$E$10).

3) **Data Format fixes:** The data format of some columns were changed during the cleaning process. The columns and their treatment include:
    - Past 3 Years Bike Related Purchases: The format was changed from text to general
    - DOB Column: The format was changed from text to date using this formula =IF(ISNUMBER(D2),D2,DATEVALUE(LEFT(D2,4) & "-" & MID(D2,6,2) & "-" & RIGHT(D2,2)))
    - Postcode column: The format was changed from text to general format.
4) **New Columns:** Some New columns were created to aid the analysis process. They included:
    - Age column: The data was extracted from the DOB column using this formula =2019-YEAR(D3) and converting the data format to number.
    - Age category column: The data was extracted from the age column using the formula =IFS(E2<18, "Under 18",E2<=29, "Young Adults",E2<=44, "Adults",E2<=59, "Middle-Aged Adults",E2>=60, "Seniors"). For columns where the age column was missing (87 values), they were filled with “Unknown”.

5) **Tenure column:** The column initially appeared to represent the length of time customers have been with the company. However, further inspection showed that nearly all values were greater than zero, which contradicts expectations for new customers. This suggests either a mislabeling or a different undocumented meaning for the column. As a result, the column was retained for record keeping but excluded from all key analysis and dashboards due to uncertainty around its validity and relevance
6) **Rank & Value columns:** The rank column appears to be based on the value column which lacks contextual information about what the values represent. Due to this ambiguity, both columns were retained in the dataset but excluded from final dashboards and strategic analysis.

TOOLS USED

- Excel

CONCLUSION

The data cleaning process for the Sprocket Central Pty Ltd dataset involved careful handling of inconsistencies, null values, and structural issues across four data sheets. These cleaning efforts have made the dataset more reliable and analysis-ready. By retaining key data where useful and excluding unreliable entries where necessary, the integrity of both customer and product insights is preserved. The cleaned dataset now supports more accurate segmentation, trend analysis and business decision-making.

SOURCES

- ISO 3166-2:AU on Wikipedia https://en.wikipedia.org/wiki/ISO_3166-2%3AAU?utm