Marketing Campaign Dataset - Data Cleaning & Preparation Summary

Project Overview

Objective: Clean and prepare a customer marketing dataset for analysis to understand customer behaviour and improve marketing strategies. Dataset: 2,240 customer records with 29 features including demographics, purchasing behaviour, and campaign responses.

Made with GAMMA

1. Initial Data Assessment

What We Started With:

- File Format: Tab-separated values (TSV) file

Initial Issues:

- All data was read as text (object data type)
- No proper data type recognition
- Potential missing values
- Data not ready for analysis

Initial Data Structure:

Raw Data: 2,240 rows × 29 columns All columns showed as 'object' (text) type

Made with GAMMA

2. Data Cleaning Steps Performed

Step 1: Data Loading & Structure Fix

Problem: Data wasn't loading properly - all columns merged into one Solution: Used proper tab separator to split data into correct columns

Python code :-

```
df = pd.read_csv('marketing_campaign.csv', sep='\t')
```

Result: Successfully separated 29 distinct columns

Step 2: Data Type Conversion

A. Numerical Columns Conversion

Converted to Whole Numbers (Integers):

- Customer ID, Birth Year, Number of Kids/Teens
- Spending amounts (Wines, Fruits, Meat, Fish, Sweets, Gold)
- Purchase counts (Web, Store, Catalog, Deals)
- Campaign responses (AcceptedCmp1-5)
- Recency, Complaints, Response

Made with GAMMA

2. Data Cleaning Steps Performed (Continued)

Converted to Decimal Numbers (Float):

- Income (may have decimal values)

B. Date Column Conversion

Problem: Dates were stored as text

Solution: Converted to proper date format

Python code :-

```
df['Dt_Customer'] = pd.to_datetime(df['Dt_Customer'], format='%d-%m-%Y')
```

C. Category Columns Conversion

Converted to Categories:

- Education (Graduation, PhD, Master, etc.)
- Marital Status (Married, Single, Divorced, etc.)

Why Categories? More efficient storage and faster analysis

Made with GAMMA

2. Data Cleaning Steps Performed (Continued)

Step 3: Missing Value Treatment

Found: Missing values in Income column

Solution: Filled missing income with average income

Python code :-

```
df['Income'] = df['Income'].fillna(df['Income'].mean())
```

Step 4: Duplicate Check

Result: No duplicate rows found - data quality was good

Step 5: Feature Engineering (Created New Columns)

New Business Metrics Created:

Total Spending

Sum of all product category spending

Formula: Wines + Fruits + Meat + Fish + Sweets + Gold

Total Children

Combined count of kids and teens at home

Formula: Kidhome + Teenhome

Made with GAMMA

2. Data Cleaning Steps Performed (Continued)

Total Purchases

Total number of purchases across all channels

Formula: Web + Store + Catalog + Deal purchases

Customer Tenure

How long customer has been with company (in days)

Formula: Today's date - Customer join date

Made with GAMMA

3. Final Dataset Structure

Cleaned Data Specifications:

- Rows: 2,240 customers
- Columns: 33 (29 original + 4 new features)
- Memory Usage: Optimized with proper data types
- Missing Values: 0 (completely clean)

Data Types After Cleaning:

- Integers: 25 columns (ID, counts, amounts)
- Float: 1 column (Income)
- Datetime: 1 column (Customer join date)
- Categories: 2 columns (Education, Marital Status)
- New Features: 4 columns (business metrics)

Made with GAMMA

4. Business Value Created

Ready for Analysis:

Customer Segmentation

By demographics, spending patterns

Campaign Effectiveness

Which campaigns got best response

Purchase Behavior

Channel preferences, product preferences

Customer Lifetime Value

Based on spending and tenure

Made with GAMMA

4. Business Value Created (Continued)

New Insights Enabled:

- Who are our high-value customers?

  (Total Spending)

- What are the preferred purchase channels?

  (Total Purchases breakdown)

- How does family size affect purchasing?

  (Total Children)

- How does customer loyalty evolve over time?

  (Customer Tenure)

Made with GAMMA

5. Key Technical Achievements

Data Quality Improvements:

- 100% clean data - no missing values
- Proper data types - optimized for analysis
- No duplicates - unique customer records
- Enhanced features - added business metrics

Process Excellence:

- Reproducible: Complete script for re-running
- Documented: Clear step-by-step process
- Scalable: Can handle larger datasets
- Business-focused: Created actionable metrics

6. Next Steps for Analysis

With this cleaned dataset, we can now:

- Perform customer segmentation (clustering)
- Build predictive models for campaign response
- Analyze purchasing patterns by demographic
- Calculate customer lifetime value
- Optimize marketing campaign targeting

Made with GAMMA