Southern New Hampshire University

DAT 325 Project Two
Executive Summary Report

Data Set Anomalies

Daniel Palhano
February 9, 2025
Supermarket Dataset

Key Value	Description of Anomaly	Plan for Resolution
12/31/1899 10:16:00 PM (Time Column)	Time column displaying a date while there already is a date column next to it.	Delete the date portion of the time column and just leave time of the transaction.
Mandalay, Myanmar (City Column)	The only country the supermarket operates in is Myanmar, there is no need for the country's name.	Delete the country name behind the city name. Add a separate column for the country's name if needed.
Credit card (Payment column)	Credit card should be abbreviated as CC as per the dictionary.	Change credit card values to CC as they should be.
Bitcoin (Payment column)	Bitcoin is not in the dictionary as a payment type. Only 1 bitcoin entry.	Should be changed to Ewallet.
'cogs' column name	Does not reflect the dictionary column name.	Column name should be changed to 'CostOfGoodsSold' or change the dictionary column name to 'cogs'.
Myanmar entry for invoice ID 704-48-3927	Does not mention which city the invoice was from.	Find out which city the invoice was made in and edit the entry with the city name.
Missing date entry for invoice ID 778-71-5554	Does not have a date for the invoice entry.	Find date in which the invoice was made and add it to missing cell.
-43.19 unit price entry for invoice ID 252-56-2699	False negative unit price entry.	Delete negative sign and double check unit price is correct.

Data Types

Header Name From File	Data Types Note
InvoiceID	Ok
Location	Change change city names to A, B, C to add both sets together
CustomerType	Change customer type from “Member” and “Normal” to 1 and 0 like dictionary
Gender	Change “Male” to 0 and “Female” to 1 and N/A to 2
ProductLine	Ok
UnitPrice	Ok
Quantity	Ok
TaxApplied	Ok
TotalOrder	Ok
DateOfPurchase	Ok
TimeOfDay	Change data type to int and adding where time of purchase is according to dictionary
PaymentType	Change where Credit card is written to CC
CostOfGoodsSold	Ok
GrossMarginPercentage	Change from 4.76…. to 0.0476…
GrossIncome	Ok

Specific Transformations Needed to Join the Data

Header Name From File	Excel Function One	Excel Function Two	Excel Function Three
Location	IF(A2="Yangon", "A")	IF(A2="Mandalay", "B")	IF(A2="Naypyitaw", "C")
TimeOfDay	IF(AND(A2>=TIME(7,0, 0), A2<=TIME(11,0,0)), 0	IF(AND(A2>TIME(11,0, 0), A2<=TIME(3, 0, 0)), 1, 2	
PaymentType	SUBSTITUTE(A2, "Credit card", "CC")		
CustomerType	IF(A2="Normal", 0)	IF(A2="Member", 1)	
Gender	IF(A2="Male", 0)	IF(A2="Female", 1)	IF(A2="", 2)
GrossMarginPercentage	A2 / 100		

Executive Summary

For our datasets, the changes we have made from the Wayne Enterprises’ set to transform it into the Bruce Inc. dataset was on the Location, TimeOfDay, PaymentType, CustomerType, Gender, and GrossMarginPercentage. The plan was to look at the data sets and find inconsistencies between the two and change the Wayne dataset to incorporate them together.

Southern New Hampshire University

Cleaning of the dataset involved looking at the anomalies within the dataset in preparation for the final cleaning and merge. First this involved changing certain aspects like the TimeOfDay column which involved a poorly written time system compared to the Bruce dictionary. Another column that required changes was the City column which included the country name as well which served no purpose since both companies only operate within Myanmar. The last couple of changes that have been made are “Credit card” entries being changed to “CC” as it should and “cogs” column needing to be changed to “CostOfGoods”.

For the final cleaning, some excel functions were needed to make the final changes needed for the merger. First was the big changes of TimeOfDay and Location columns which needed to be made to reflect the Bruce dictionary. These changes involved changing city names to A, B, and C, and times of day to be within the number ranges that they are in Bruce set as well. The easier changes are Gender from written out genders to numeric values they should be, 0 for male, 1 for female, and 2 for any null or missing values. CustomerType was also given the same treatment with 0 indicating normal customers and 1 being for members. Lastly, for the “Credit card” entries we substituted them to “CC” as they are to be and GrossMarginPercentage from 4.76 to 0.0476 by dividing all those values by 100.