Southern New Hampshire University

DAT 325 Project Three
Data Validation

Daniel Palhano
Supermarket Dataset

File Names Table

Perform the following data validation on each of the import, existing, and merge data tables.

- The COUNT of rows in each table. You should reconcile these counts to make sure the number of rows you inserted made it into the merged table.

	Excel Files				
	Source File	**Anomalies**	**Import Data**	**Existing Data**	**Merge Data**
Count	1021	67	205	796	1002

- Use the identified variable for **your chosen data set** as listed below to determine MIN, MAX and AVERAGE for each table:

  Data Set: Supermarket
  Variable: Gross Income by location

  Determine the MIN, MAX, and AVERAGE in your uploaded table, the MIN, MAX, and AVERAGE in the existing table before you inserted the new rows, and the MIN, MAX, and AVERAGE in the existing table after you insert new rows.

	Excel		
	Import Data	**Existing Data**	**Merge Data**
MIN	0.7715	0.5085	0.5085
MAX	49.59	49.65	49.65
AVERAGE	15.17	14.77	14.85

Summary

Distribution Data

From importing the data from Wayne Ent. data sheet we see that we increased the merge data sheet roughly 26% bigger than we did before going from 796 to 1002. This increase in the merge data sheet, we see that we have a slight increase in average in our gross income.

Validation Steps

The steps that I took to clean the data went from simple changes like changing of the column headers to reflect the Bruce data set and changing simple columns like Location, CustomerType, Gender, and thePaymentType anomalies. We changed those columns to reflect the Bruce data set for example Gender from displaying 'Male' and 'Female' to 0 and 1 and the entries where 'Credit card' is to 'CC' instead. Location was changed to the A, B, and C that it should be and member and non-members to 1 and 0 as well. Another small change that was made was dividing the GrossMarginPercentage by 100 to be the 0.04761904762 it should be.

More complex changes that were made was the changing of TimeOfDay of the purchase to 0, 1, 2 which involved a much larger formula to convert the time stamps, including TaxApplied and TotalOrder amount, and lastly the GrossIncome column. The TaxApplied column was made by multiplying CostOfGoodsSold by 5% tax rate and adding those 2 columns together and importing it into the TotalOrder column. The GrossIncome and TaxApplied columns are the same amount so importing it was as easy as copying the data over.