Dear Sprocket Central Pty Ltd,

I have looked through the dataset provided and determined the data quality could be improved. Some of the major issues are that in the Customer Demographic dataset there is an unnecessary column named default. This column contains useless data that makes no sense. In addition, in the Transactions dataset the date column is empty for all entries. To help mitigate issues with invalid columns we can just ignore them. The default does not seem to contain any data that would be useful for analysis and date column contains no data at all so it would be safe to remove.

Furthermore, using Python I’ve found all the rows in each dataset that contain empty values. I’ve linked a text file for Transactions and Customer Demographic containing the ID’s, empty cell values and the total number of rows which contain empty values. To help counteract this, we can either choose to remove or ignore the empty cell. The optimal choice is to do both, ideally we can go through the dataset and if the empty cell value is essential to the data analysis we can ignore that entire row and if the empty cell value was nonessential we can choose to ignore that empty cell and continue with the rest of row data. Customer Addresses had no empty values at all. In total transactions had 555 rows with empty cell values and customer demographic had 689.

On a positive note, none of the datasets contained any duplicate rows and all the transactions are less than 30 years old so they can be considered up to date. In customer demographic, only 1 customer has an age over 110. Their ID is 34. We can remove this person safely since it is a false entry.

If you follow the guidelines above the data should be relatively clean and values can be analysed. If you would like to see the invalid data or more information on analysis I’ve linked the jupyter notebook with python code and text files containing rows with invalid entries.

Link to customer demographic invalid entries: 

Link to transaction demographic invalid entries:

Link to jupyter notebook on github:

From,

Anirudh Lakra