Data Management II – ETL and Smart Data Project Submission Template

Name of the Student: Rishabh Garg

Matriculation Number: 11011875

Dataset Considered for Project: Walmart_Detail_DataSet

Note:

- All the 3 questions mentioned below must be answered by the dataset considered.

Question 1:
Is Star Schema or Snowflake schema or Fact Constellation Schema possible with the dataset? If yes, then which schema? If no, then drop a mail regarding the issue with reason.

Answer: Star schema is possible with the dataset. There may be snowflake schema that I will explore during further analysis.

Question 2:
Which all quality dimensions are covered by the dataset? Mention the data quality dimensions for the relevant columns and also provide a summary of data profiling for the dataset?

Answer: The following are the data dimensions covered by the dataset and the technique to clean it are given below.

Column Name	Dimension	Technique
Item_Identifer	Completeness	Text facet operation is done to check for any blank records in open refine. There are no blank records presents in the item_identifier.
	Consistency	The complete column is following a standard format, and this was verified by doing a toupper operation on the column. It is observed everything is in upper case.
	Validity	To check whether the id is following a definite pattern of 3 alphabets and 2 numerical value format or not, trifacta tool is used. Using the extract function, a regular expression to written in the Text to extract option /[A-Z]{3}/. The result is a new column is made which extracts the first three alphabets of the id. It is observed all the ids are in upper case and follows a definite pattern.
Item_Weight	Completeness	It has been observed a total of 1510 records have no information about weight. The specific ids are pin pointed and for each pin pointed identifier, a manual fill down operation is done in the openrefine to complete the dataset
	Consistency	Mass edit operation was performed on the column to change all the values with more than

		2 decimal points. These values are rounded off to values with maximum 2 decimal points. Each value is chosen manually and edited accordingly
Item_Fat_Content	Consistency	Many abbreviations have used in the column and inconsistent formats are being used to identify records such as LF or low fat instead of the correct proper case format which is Low Fat and reg for Regular. A text facet operation on the column is done in the open refine tool and inconsistent format records are selected and a mass edit is performed correspondingly to ensure a consistency check throughout the data. There approximately 307 records with LF value, 108 with low fat value and 115 records with reg values that are converted into proper case consistent value. Also there are no null values present in the column.
Item_visibility	Consistency	The column can be interpreted as a percentage value, which can be defined as the area occupied by the product divided by total area of the outlet. Thus for ease of understanding the column, the decimal point values are converted into percentage values with maximum 2nd decimal point position in the trifacta tool. The formula applied is NUMFORMAT($col, ##.## %). As a result all the values are converted into a proper format as per the use case.
Item_MRP	Consistency	As per the use case defined the item_MRP column has inconsistent values. Since the lowest denomination is 1 cent which can be represented at the 2nd decimal point. Beyond that, practically it doesnot make any sense to have any values in the 3rd decimal point and above. All the values are rounded off to the nearest 2nd decimal point using the trifacta tool. A Formula NUMFORMAT($col, ###.##) is applied which converts all the values in the column into consistent values as per the business rule.
Outlet_Identifier	Completeness	There are no null values in the column. It is checked by doing a facet operation on the column using the open refine tool.
	Consistency	The complete column is following a standard format, and this was verified by doing a toupper operation on the column. It is observed everything is in upper case.
	Validity	To check whether the id is following a definite pattern of 3 alphabets and 3 numerical value format or not, trifacta tool is used. Using the extract function, a regular expression to written in the Text to extract option /[A-Z]{3}/. The result is a new column is made which extracts the first three alphabets of the id. It is observed

		all the ids are in upper case and follows a definite pattern.
Outlet_Establishment_Year	Validity	A text facet operation is performed to identify any invalid outlet establishment year i.e. to identify if any establishment year is greater than 2019 which would be invalid. It is observed none of the year are beyond 2019 and all the records have a valid establishment year
Outlet_Size	Completeness	A text facet operation is performed on the column to identify any blank record. There are 2410 blank records corresponding to the outlet_identifier ids **OUT010, OUT017 and OUT045**. For schema completeness, "Null" is being updated to all the records for completeness of the dataset.
Item_Outlet_Sales	Consistency	As per the use case defined the item_Outlet_Sales column has inconsistent values. Since the lowest denomination is 1 cent which can be represented at the 2nd decimal point. Beyond that, practically it does not make any sense to have any values in the 3rd decimal point and above. All the values are rounded off to the nearest 2nd decimal point using the trifacta tool. A Formula NUMFORMAT($col, ###.##) is applied which converts all the values in the column into consistent values as per the business rule.

Question 3:
Question that will be answered from the dataset using data analytics techniques?
Explain in 2-3 lines the question that will be answered and also specify under which data mining model the question will be classified.

Answer:

The following are the questions that will be answered from the dataset using data analytics techniques. Further new analysis will also be explored on the dataset for new insights.

1. What is the impact of weight, visibility & MRP of a product on product sales?

   The above question would be solved using multiple linear regression.

2. Is there a dependency of Outlet size and Outlet type on outlet sales for outlet in the location type Tier 1?

   The Above question would be solved using Anova.