BigMart Data Analysis

Data Cleaning Process:

Import Data:

- Data from the CSV file named 'Train.csv' is retrieved into a blank Excel sheet for further data cleaning process.

Identify Missing Values:

- The entire data is formatted into a table and all the values of each column are inspected using the filter dropdown and is checked for any blank values or irregular texts.

- The *item_weight* column has some null values and they are filled with the median weight value. The *Item_Fat_Content column* has two values 'Regular' and 'Low Fat' but some of the rows consisted of 'reg' and 'LF'. So they are replaced with 'Regular' and 'Low Fat' using find and replace.

- The *Item_visibility* column also has some null values. To handle this first the data is sorted according to the *Item_Type* and then according to *Outlet_Identifier.* As the products of similar type are placed almost in the same location at a particular store, they have approximately same visibility. So, the missing values in the visibility column is filled with the average visibility of each item type in each store. This way, the data that is filled have more probability to be close to the actual value.

- The *Outlet_Size* column has null values too which belong to 3 outlets particularly. But there is not enough information to predict the sizes of these outlets so they are replaced with NA so that they don't interfere with the analysis.

Remove Duplicates:

- The Whole table is selected and "Remove duplicates" function in the data tab is applied to it. But no duplicates are found.

Handling Outliers:

- All the columns are checked for outliers by comparing their mean(50% IQR) and median. They are close values for all the columns suggesting that there are no outliers.

Data Type Conversion:

- The data type of the data in each column is checked and all the datatypes match the datatype requirement of the attributes.

Text Data Cleaning:

- The inconsistencies present in the text such as unnecessary spaces special characters variations in capitalization unwanted symbols are eliminated.

Clean data set is stored as a new file so that the original data is preserved.

Dashboard manual

- Preliminary insights like ‘Total sales ’ and ‘Average sales’ are displayed in a straight forward way as numerical values.

- A pie chart representing share of total sales among different types of stores is displayed. In that chart we can clearly observe that ‘Supermarket Type1’ has a majority (69.48%) share.

- From the line chart between visibility and items sold, it can be observed that the most products sold has a visibility 0.03 and 0.11 and very less products are sold that has visibility of 0.2 and above. From this chart, we can understand that more visibility does not necessarily mean more sales.

- From the bar chart showing sales of different products based on fat content it is understood that products with low fat content are bought more by the customers.

- From the bar chart describe sales different types of products it is understood that products in the category of fruits and vegetables the most sales food products have the least sales.

- In the line chart showing total sales of each outlet established in different years, more sales can be seen in the outlet which are older than the relatively newer outlets.

- Total sales of the items are plotted with their respective MRPs (Maximum Retail Price) in a line chart. As the MRP is increased, the sales increased. So a large part of the sales is contributed by the costlier items. But the count of items vs MRP chart reveals that all the items are bought by the customers in equal quantity.

- The distribution of outlet sales across different location types show Tier 1 cities bring in highest number of sales.

- The line chart between sales and outlet size tells that there is no relation between store size and sales. So we can conclude that sales does not depend on the outlet size.