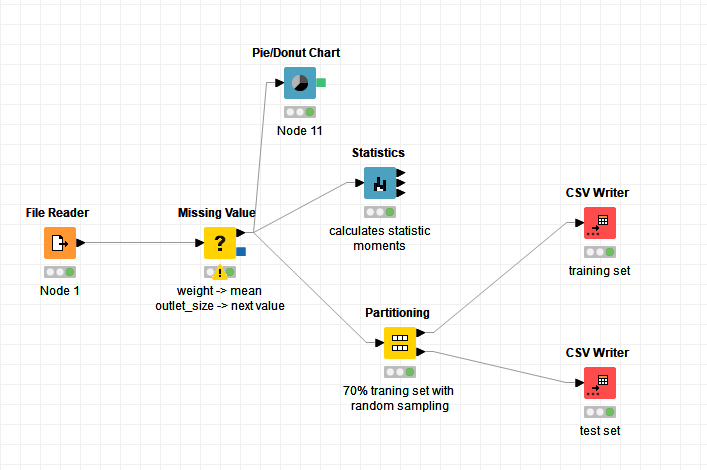
KNIME PROCESS LOG

****

For the pre-processing workflow, the BigMart dataset is imported as a CSV file into the KNIME Analytics Platform through the ‘File Reader’ node. The ‘Missing Value’ node takes the dataset from the ‘File Reader’ node and implement changes to missing or incorrect values. Any missing values of the product weight are replaced with the mean of labeled values in the same column. Similarly, any missing values of the size of outlet locations are filled with the next labeled value in its column. Other variables that do not have missing values are excluded from this node.

Next, the ‘Statistics’ and ‘Donut Chart’ nodes are linked from the ‘Missing Value’ node. Specifically, the ‘Statistics’ node computes data points such as mean, median, standard deviation and illustrates histograms for each possible variable. On the other hand, the ‘Donut Chart’ node creates a donut chart based on a column input. The ‘Partitioning’ node splits the dataset formed from the ‘Missing Value’ node into two datasets, a training dataset, and a test dataset. For this assignment, the split method is constructed by randomly selecting rows from the original dataset, in which 70% of that dataset goes into the training data, leaving the other 30% for the test data. The training dataset is then extracted to a CSV file by connecting the upper output port of the ‘Partitioning’ node with the ‘CSV Writer’ node. After that, to retrieve the test dataset without the prediction target attribute, it is required to configure the ‘Missing Value’ node and remove every column customization done before. Then, using a fixed seed when randomly generating the training dataset in the ‘Partitioning’ node, the test dataset will be created. To extract the test dataset into a CSV file, the lower output port is linked with a different ‘CSV Writer’ node.