# Homework 2

Using Knime and/or Excel on the cereal.csv dataset...

* Explore whether there are missing values for any of the variables
* Determine whether there are any outliers among the "Sodium" values
* Normalize the variables “Calories", "Sodium" and "Potassium" (for example using Min-Max transformation)
* Analyze and interpret the correlations between each one of the variables and the variable "Rating"

Output Report:

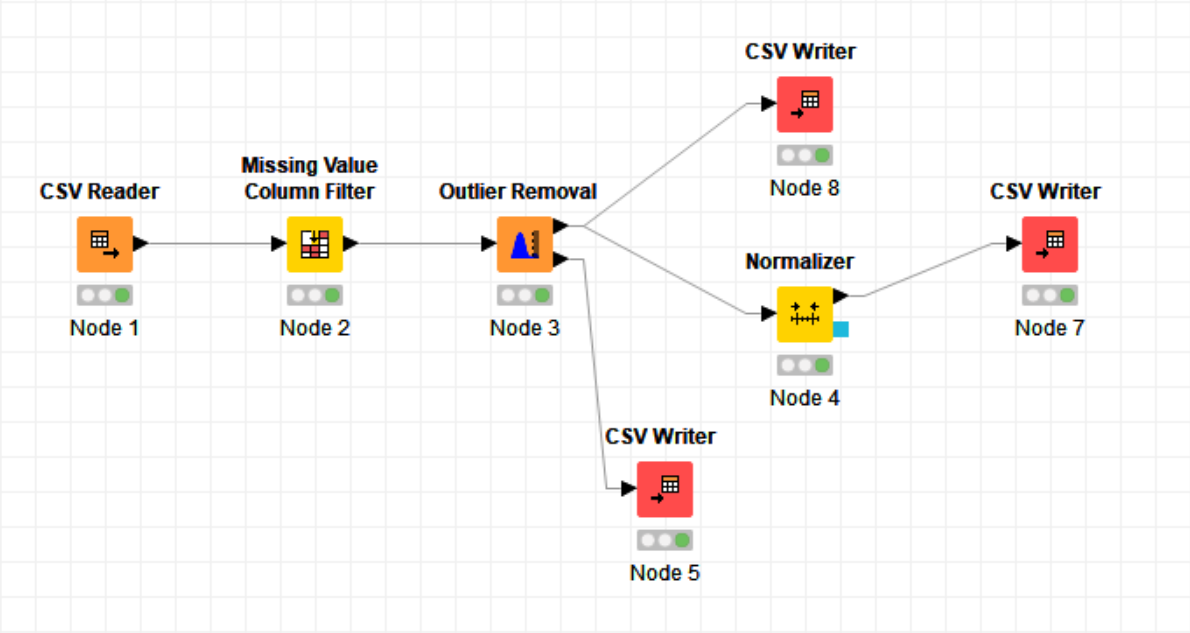


Figure Project Work Board

The figure 1 shows the project work board for this assignment. I imported the ‘cereal.csv’ file as an input to ‘CSV Reader’ and was passed through a ‘Missing Value Filter’ with a 90% threshold. After that the csv file was passed through an ‘Outlier Removal’ process for only the column of Sodium. The Outlier were recorded in a CSV file and are shown as in figure 2. There were no recorded outliers from ‘Sodium’.

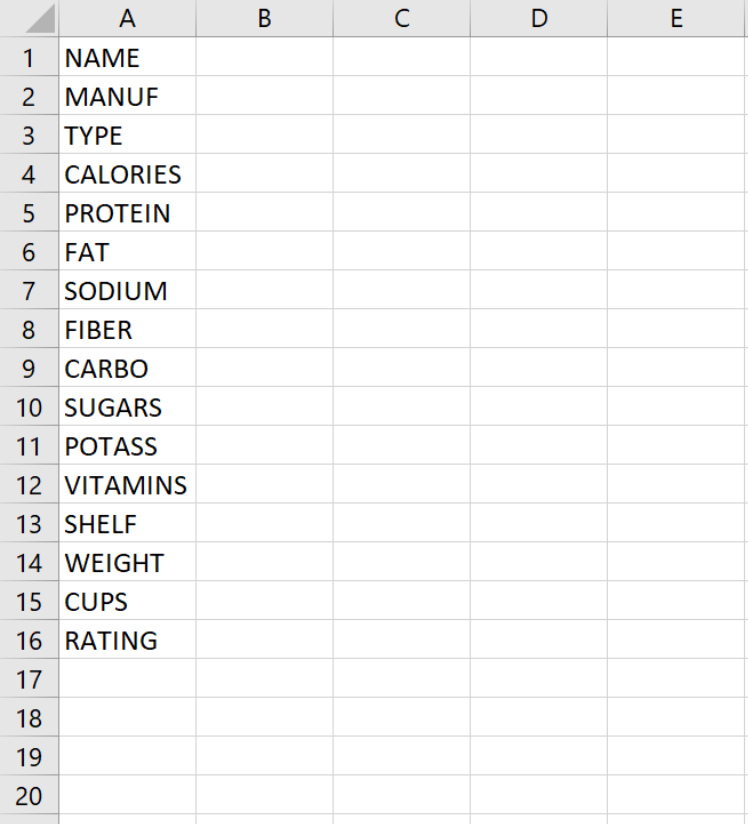


Figure Outliers

The filtered output from ‘Outlier Remover’ was also recorded and a part of it is shown in figure 3.

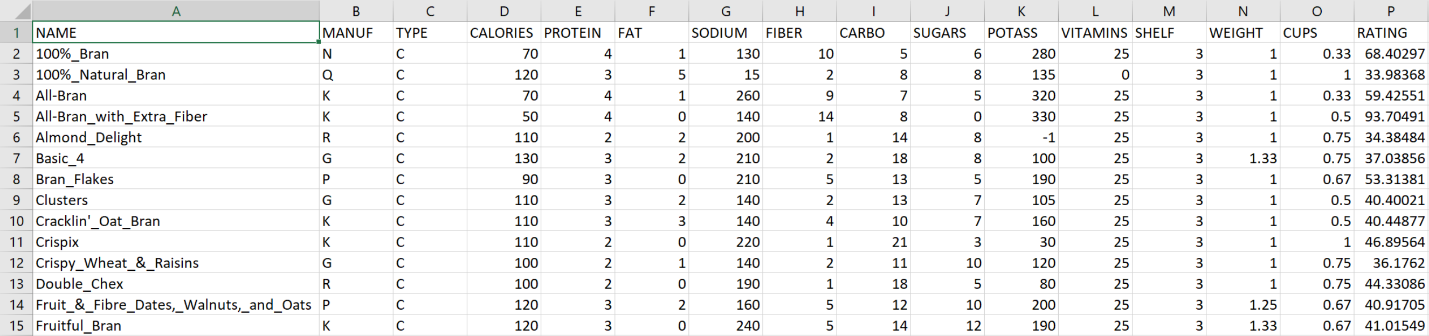


Figure Part of Filtered Output

After filtering the output from through an outlier filter, it was passed to a ‘Normalizer’ for only the columns of ‘Sodium’, ‘Calories’ and ‘Potass’ using Min-Max Transformation. The normalized output were in the range of 0 to 1. A part of the output is shown in figure 3.

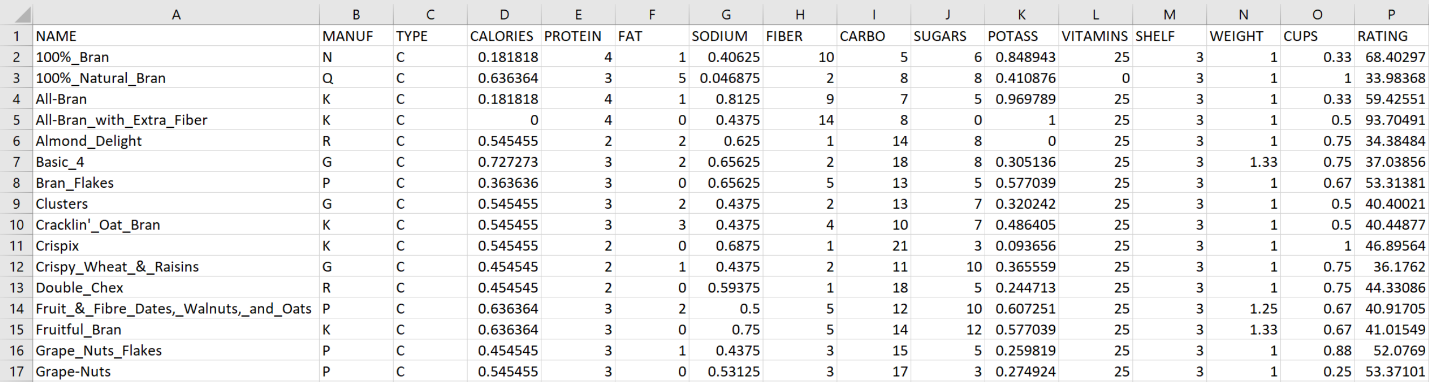


Figure Normalized Output

To analyze the relation for ratings, ‘Rank Correlation’ node was added [shown in figure 5] and a correlation matrix was developed [shown in figure 6]. From the graph it is clear that the ratings are inversely proportional to ‘calories’ and ‘sugar’ and is directly proportional to ‘protein’ and ‘fiber’. There are other relations as well, detailed in the figure 6.

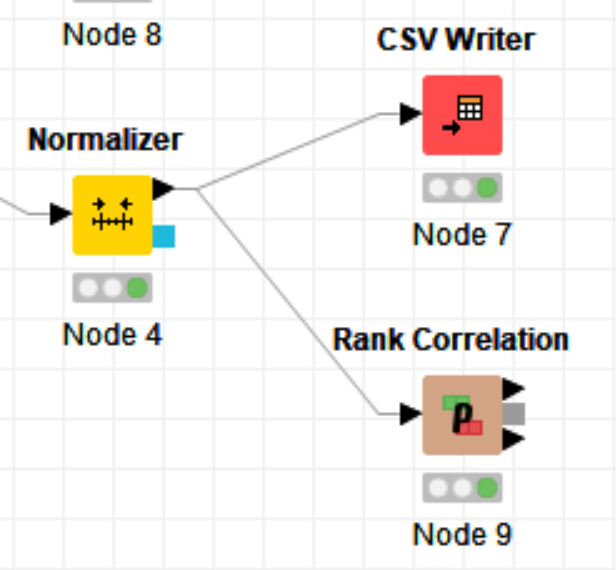


Figure Rank Correlation Node

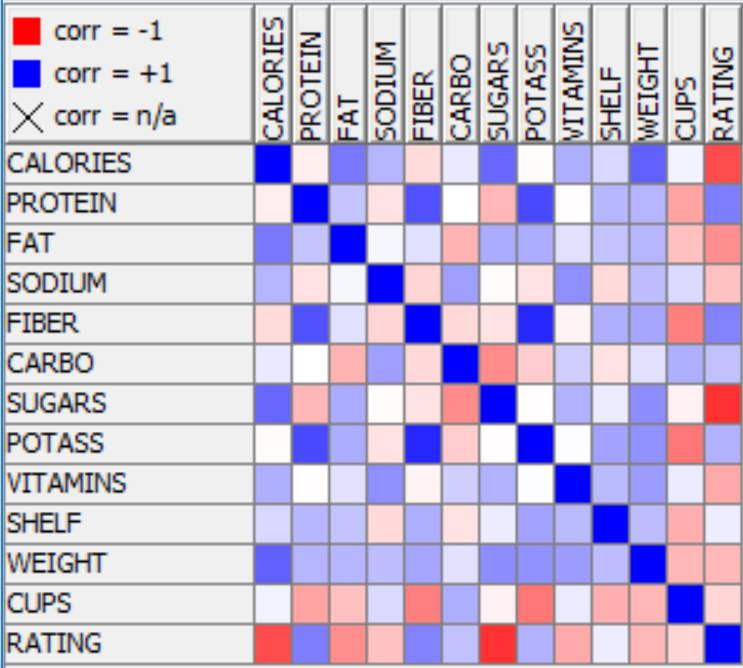


Figure Correlation Matrix