Analyzing Cereal data

Top 6 rows of the cereal data

Table

Description automatically generated

Used histograms to evaluate our variables, the diagram below is a histogram of the fat column, which shows that the column is skewed to the right.

Chart

Description automatically generated

Did a log transformation for the variables skewed to the right, this is what the histogram above looked like after applying the log transformation

Chart, histogram

Description automatically generated

After checking for missing values and fixing the skewed variables, I removed the variables that are not of interest and then compared the variables left to the response variable (rating). From the diagram below, fat generally has a negative correlation to rating. Cereals with more fat have lower rating.

Chart, scatter chart

Description automatically generated

Building a predictive model for nutritional rating

I divide the data into test and training data, 80% of the data was for the training and the 20% left was for testing. The training data was used to fit a linear model and the test data was used to test the accuracy of the linear model.

Table

Description automatically generated

To test if our linear model is a good fit, we use the results shown above. The R^2 value is near 1 which generally means that our model fits the data well. We can also see from the p -values which variables were a good addition to our model. Variables with a p -value less than 0.05 are not very meaning for our model.

I also calculated the Mean square error (MSE) which measures how far our model is to our actual data points, and the MSE is 7.270.