Data Findings – Cereals data set

By : Gaurav Sharma

Correlation Findings

Correlation Table		
Independent	Dependent	Correlation
calories	rating	-0.68937603
protein		0.470618465
fat		-0.40928366
sodium		-0.4012952
fiber		0.58416042
carbo		0.052054661
sugars		-0.75967466
potass		0.380165369
vitamins		-0.24054361
shelf		0.025158816
weight		-0.29812398
cups		-0.20316006

Calories , protein , sugars and Fiber are highly correlated with the response variable

Regression Findings

Linear Regression - All variables	
	Coefficients
Intercept	54.92718423
calories	-0.222724163
protein	3.273173861
fat	-1.691408004
sodium	-0.054492702
fiber	3.443479785
carbo	1.092450944
sugars	-0.72489514
potass	-0.033993351
vitamins	-0.051211969

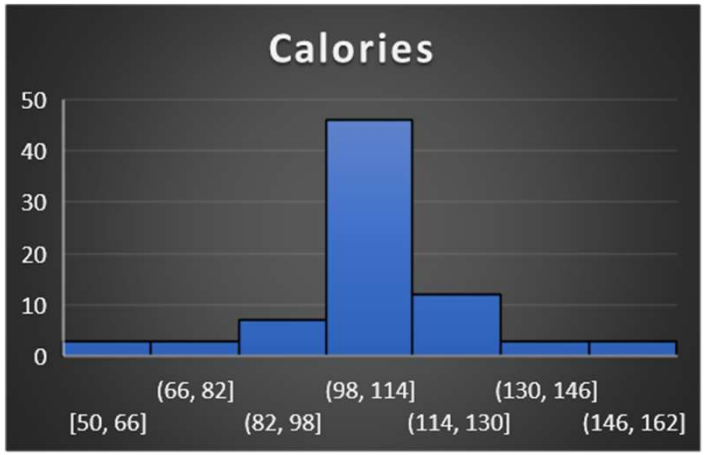
After performing linear regression on all variables ,Coefficients of protein and fibers comes out to be higher which means that these factors have strong influence on the ratings

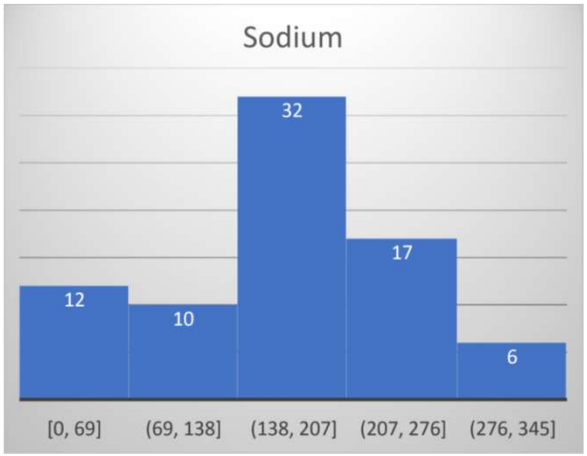
Regression Findings

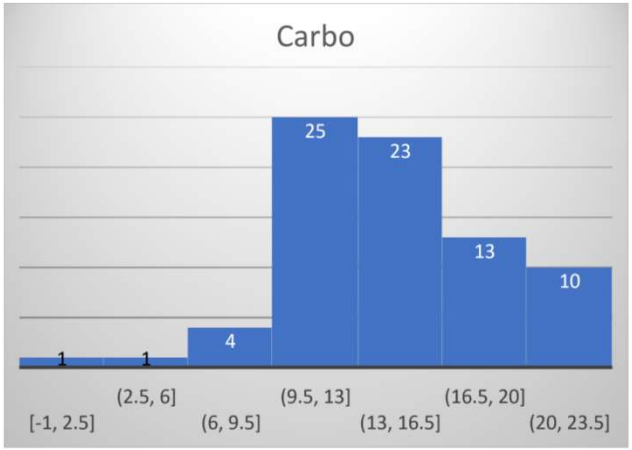
	Coefficients
Intercept	68.51718988
calories	-0.226771116
protein	1.79097183
sugars	-1.538985228
fiber	2.082139988

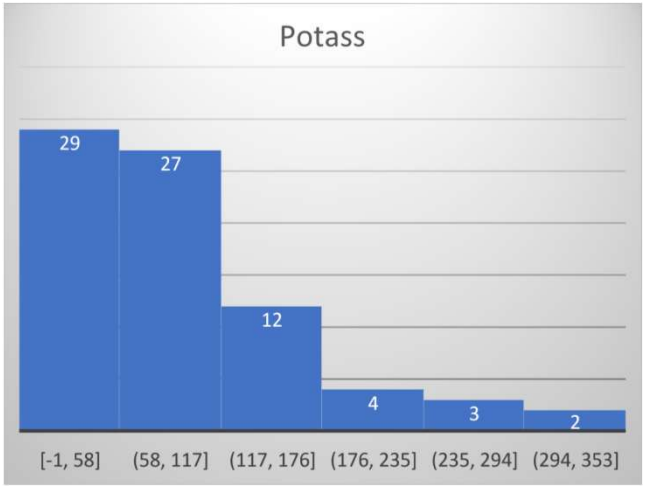
Linear regression was on run on 4 variables who have high correlation with output. Result shows that proteins and fiber have strong effect on the output variable.

Histogram Findings







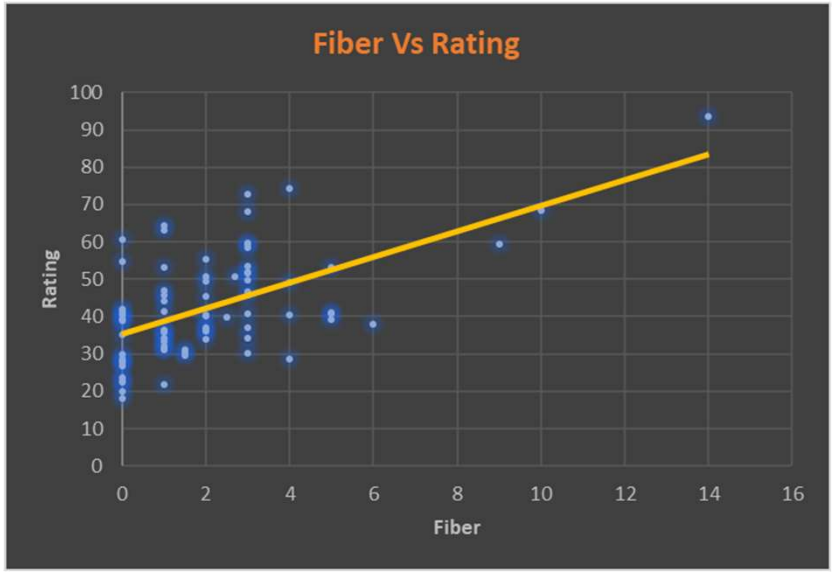


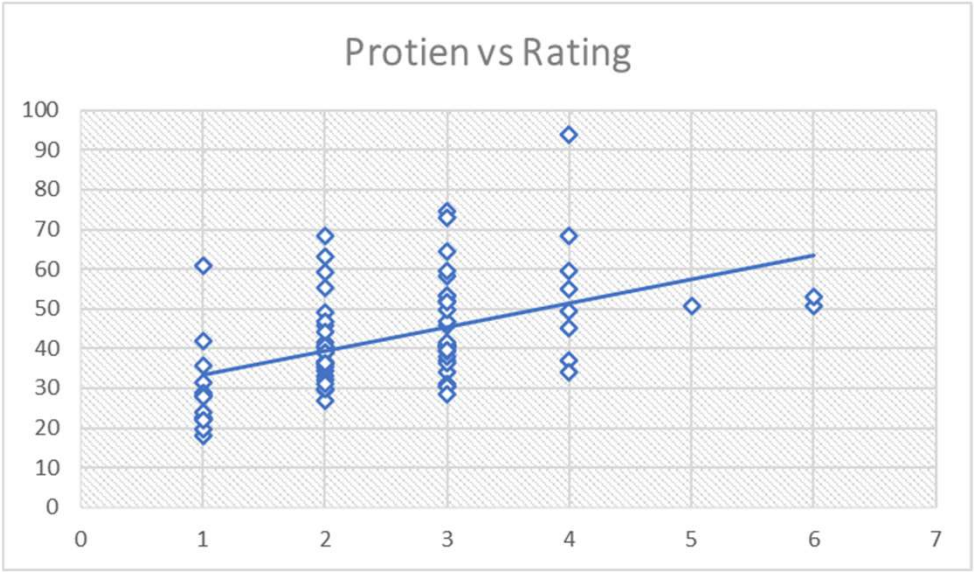
Histogram of Calories and Sodium shows a normal distribution trend with high positive kurtosis

Histogram of potass shows a positive skew while carbo shows a negative skew

Overall Impression

- proteins and fiber have come out to be the most influential variables among all in the data set





Thank You