Consumer Products

I will explore the relationship between grocery products and detergents and paper. I posit that the more consumers buy groceries, the more they must buy detergents and paper. If proven true, I could then track my grocery spending and use it to predict my detergents and paper products needs. This dataset tracks consumer spending in Portugal. The data is divided amongst regions. Lisbon and Porto are both on the coast. With a population of 550,000, Lisbon is more than double the size of Porto. The preponderance of data comes from other areas. It tracks the spending on varied consumer products. The data, as is evident from the graphs, is clearly not normally distributed. The Jarque-Bera test statistic is 1988.121 with a prob = 0.00. The skewness values of the variables are generally greater than 3, and the kurtosis exceeds 10.

At first glance, the Peason's correlation between grocery products and detergent and paper products shows a high correlation of 0.925. However, the data is not normally distributed. The boxplots, scatterplots, cdfs, pdfs, and summary statistics clearly show a positive, right skew. Spearman's rank still shows a strong correlation of 0.801. Kendall's tau confirms this with a correlation of 0.632 and a p-value of 2.2* 10^-87. The assorted consumer product spending follows a general pattern: most of the data points of spending are bunched towards 0, and singular points are manifest as one goes further to the right, the higher spending. There are no null values to contend with. The predominance of outliers throughout the dataset exists in the right tails. There is no suggestion that these outliers are mistaken input values. They seem to tell the narrative that consumers will sometimes spend a disproportionate amount of money on any of the items. Each product category displays this behaviour, albeit to varying degrees.

A partial correlation controls for other variables. Running a partial correlation between grocery and detergents and paper with milk as a control and using the 'spearman' method, produces an output of r=0.592583 and a p-value=5.746998e-43. Controlling for all the other variables produces r=0.58576 and p-value=1.616e-41, which is slightly lower than controlling for only one variable. A simple linear regression produces an r-squared of 0.855, or 86% of the variance in grocery is caused by detergents and paper. A multiple regression model using all product categories improves the r-squared to 0.886, or 89%. It is questionable whether to add such complexity for a benefit of 3%.

There is a paucity of information about the dataset collection and its influence on the data. Does the dataset represent a random sample of wholesalers per area? What are the demographics of the consumers? Were there large families with many young children buying enormous amounts of milk products? Did seasonality affect the data? Do people consume less frozen items during the winter? Were different products bought at the same time? Did specific products necessarily precede other specific products? This could potentially prove causation. It may definitively disprove causation if the variable in question does not precede the independent variable.

One can confidently assert that he would spend a certain amount of money over the year on detergents and paper given a known amount spent on groceries. The missing bits of information present a major impediment to our leveraging the data. Were one to prove causation, one could target customers with marketing to encourage increased sales of groceries. This would then perforce increase sales of detergents and paper, thus increasing sales in two categories while only marketing to one. While linear regression performs well, it would be appropriate to experiment with different ML models to optimize performance and potentially gain more insight into the data.