Retail Data Analysis

Anthony Medrano

Investigating Anomalies with Box Plots

Simple exploratory data analysis using box plots give a glance at the distribution of a data set and can help identify anomalies or outliers. Figure 1 depicts two graphs: The price of all products according to store (shown on the left) and the price of all products according to region (shown on the right).

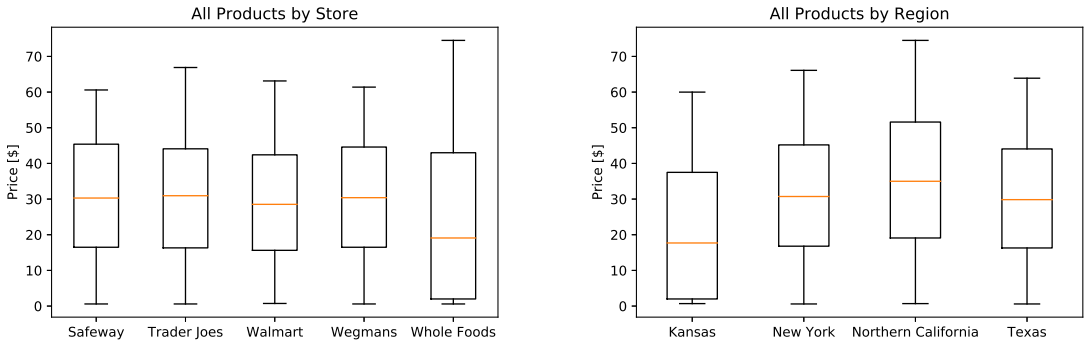

Figure 1: *Whole Foods and Kansas distributions are biased towards the minimum. The clustering of many points near the minimum greatly minimize the distance between the minimum and first quartile. The median is also affected by a large portion of data points near the minimum ($1.99) and is significantly less than that of other stores and regions.*

The left box plot in Figure 1 negates the assumption that Whole Foods products are more expensive than Safeway products prompting further investigation. The right box plot in Figure 1, however, does not negate the assumption that Kansas products are less than Northern California products. Nonetheless, given the abnormal distribution, further investigation is also required. Prior to examining Whole Foods and Kansas distributions, box plots of Texas products and Trader Joes products are displayed in Figure 2 to examine the shape of the data. These two categories were chosen because their distributions in Figure 1 are consistent with expectations and do not deviate significantly from neighboring distributions.

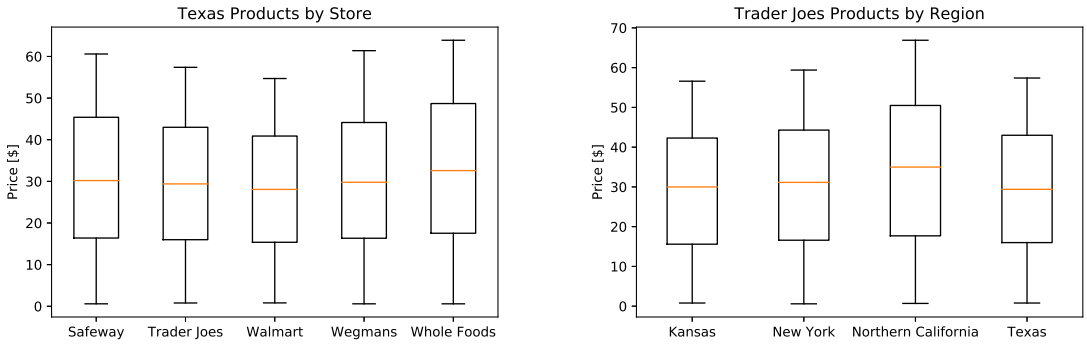

Figure 2

It is clear that the distribution patterns in Figure 2 resemble the distribution patterns of Figure 1, with the exception of Whole Foods, suggesting that these distributions are reliable. Next, Figure 3 displays the problematic distributions: Kansas and Whole Foods.

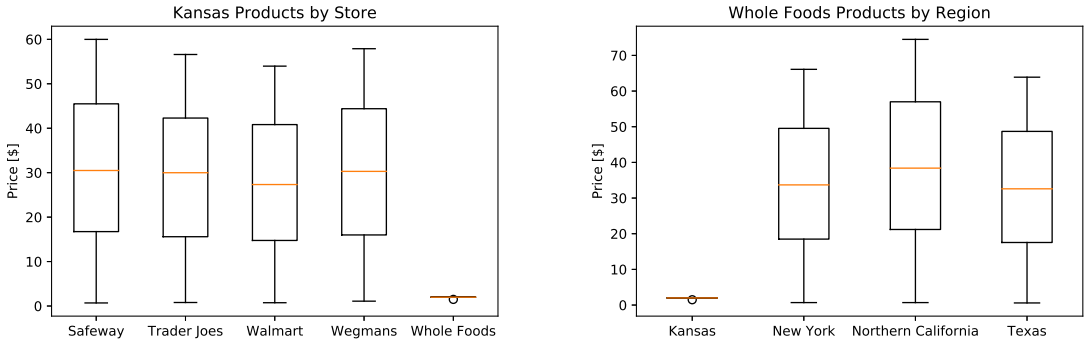

Figure 3

From Figure 3 it is evident that the price of Whole Foods products in Kansas are reported at or near a single value far below the median price of products in Kansas or other Whole Foods stores. This set of product prices is clearly an anomaly and suggests an error in price data collection at Whole Foods in Kansas. Alternatively, producing a five-number summary or computing the mean and standard deviation of each distribution would lead to the same conclusion.