Data Challenge

1 Theoretical

As a company involved in Competitive Retail Intelligence, we collect data from different retailers by sending auditors into physical stores to record pricing and assortment information. This data is then aggregated and validated using several different models and outlier identification techniques. One of these methods is a quick pricing model for price checking that assumes a store price is equal to a base price for the product/UPC multiplied by a store-specific scaling factor and a region-specific scaling factor. This can be represented as

$$P_{i,j,k} = P_i \times A_j \times B_k$$

for observed price P for product i in store j and region k, where

P_i = base price for product/upc i, can simply be the median price
A_j = the store multiplier (range from 0 to infinity) for store j
B_k = the region multiplier (range from 0 to infinity) for region k

For example, Whole Foods is usually more expensive than Safeway, so we might expect

$$A_{WholeFoods} > A_{Safeway}$$

As another example, if we assume the Northern California region is more expensive than the Kansas, then we'd represent this as

$$B_{NorCal} > B_{Kansas}$$

In this exercise, we'll aggregate some sample data and look through for anything that should be called out. It will not be necessary to apply this model to the sample data, but you may choose to use it as a guideline to help you in the following questions.

2 Questions

1. The file *prices.csv* describes prices collected for products, represented as *UPC*, at specific physical store locations, represented as *Store ID*. The auditors who collected prices at each store are represented as *Auditor ID*. Store attribute information is described in *stores.json*, and auditor information is shown in *auditors.csv*. Can you transform these sources into a cross-tabulation of regional prices alongside each other, broken down by banner, and write this out to a spreadsheet (CSV or XLSX)? Note that a given product is not guaranteed to be found in all markets at a given banner.

Banner	UPC	Northern California	New York	Kansas	Texas
Safeway	4011	0.69	0.79	0.59	0.69
Safeway	94011	0.99	0.99	–	0.89
Whole Foods	4011	0.89	1.09	0.69	0.69
...	...	...	...	...	...

Table 1: Example view of final output

2. Do you notice anything that seems off with the data we've collected? Call out anything you find noteworthy. Again, it is not necessary to use the model to find the anomalies we're looking for, but you may use it as a tool to assist you if you wish.