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 is 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.

   Please see the attachment.

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.

a. Data Overview

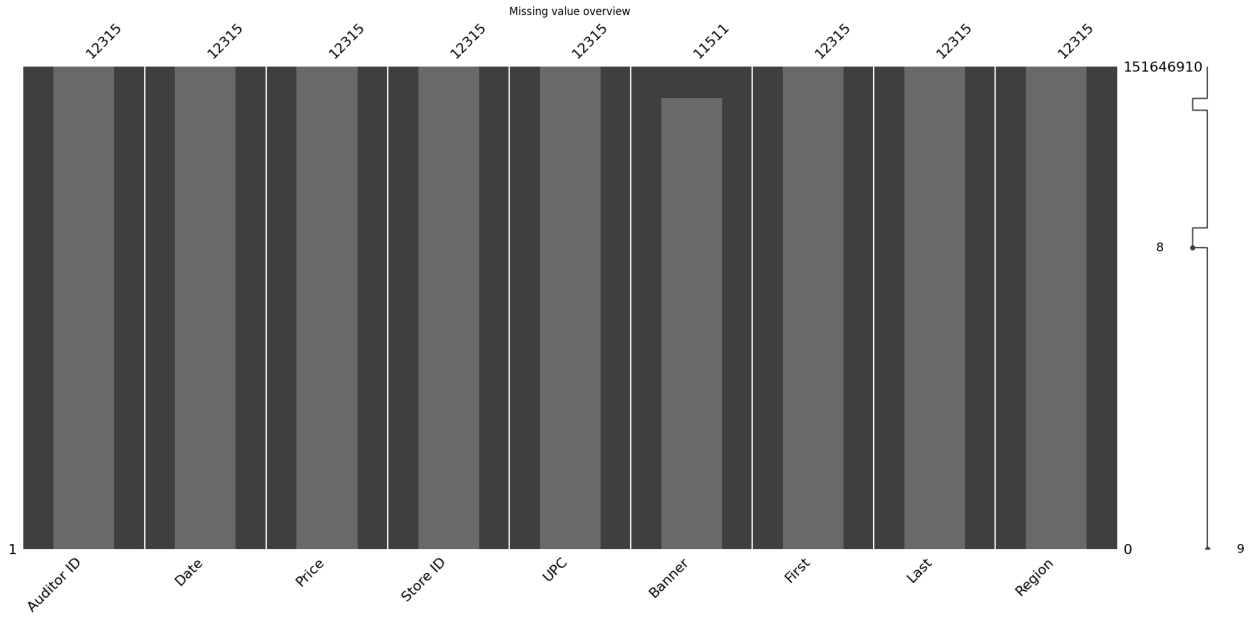

Figure 1 Missing value distribution

The 'Price.csv' file contains a total of 12,315 price datasets. The 'Auditor.csv' file includes data for 8 auditors, and the 'Stores.json' file comprises 28 datasets. As shown in Figure 1, the data is largely

complete, except for the Banner feature, which has 804 missing values.

b. Observation of Abnormal data

Given the same UPC code, we can calculate the price difference and the ratio of the price difference between two regions using the following formulas:

1. **Price Difference ($\Delta P_{Reg(j)-Reg(i)}$):**

$$\Delta P_{Reg(j)-Reg(i)} = P_{Reg(j)} - P_{Reg(i)}$$

2. **Ratio of Price Difference ($r(\Delta P_{Reg(j)-Reg(i)})$):**

$$r(\Delta P_{Reg(j)-Reg(i)}) = \frac{\Delta P_{Reg(j)-Reg(i)}}{P_{Reg(i)}} \times 100\%$$

Where $Reg(i)$ and $Reg(j)$ denote the ith and jth regions, respectively. The regions considered are Northern California, New York, Texas, and Kansas.

Given a threshold of 50%, all ratios $r \geq 50\%$ are considered abnormal prices. This threshold can be optimized for better results.

Based on the given threshold of 50%, 854 out of 4931 data points were observed as abnormal. The abnormal ratio is calculated as follows:

$$\text{Abnormal Ratio} = \frac{854}{4931} \times 100\% = 17.32\%$$

e.g.

Table 1 Example of anormal data

	Banner	UPC	Kansas	New York	Northern California	Texas	v_ny-ks	v_nc-ks	v_tx-ks	v_nc-ny	v_tx-ny	v_tx-nc	r_ny-ks	r_nc-ks	r_tx-ks	r_nc-ny	r_tx-ny	r_tx-nc
268	Safeway	286906735	10.99	16.89	NaN	11.09	5.90	NaN	0.10	NaN	-5.80	NaN	53.69	NaN	0.91	NaN	-34.34	NaN
3934	Whole Foods	11873171	1.99	5.69	NaN	5.49	3.70	NaN	3.50	NaN	-0.20	NaN	185.93	NaN	175.88	NaN	-3.51	NaN
3935	Whole Foods	15052612	1.99	59.49	NaN	57.79	57.50	NaN	55.80	NaN	-1.70	NaN	2889.45	NaN	2804.02	NaN	-2.86	NaN
3936	Whole Foods	16482322	1.99	19.69	NaN	NaN	17.70	NaN	NaN	NaN	NaN	NaN	889.45	NaN	NaN	NaN	NaN	NaN
3939	Whole Foods	16900911	1.99	32.19	36.19	NaN	30.20	34.20	NaN	4.00	NaN	NaN	1517.59	1718.59	NaN	12.43	NaN	NaN

Table 1 shows the example of anormal data. Here, features starting with v_ denote price differences, while features starting with r_ denote ratios of price differences. The ratio of price differences can highlight significant discrepancies, which may indicate pricing anomalies.

The analysis indicates that the abnormally high ratios originate from inconsistent prices in the Kansas region.

c. Anomalous Data in the Kansas Region

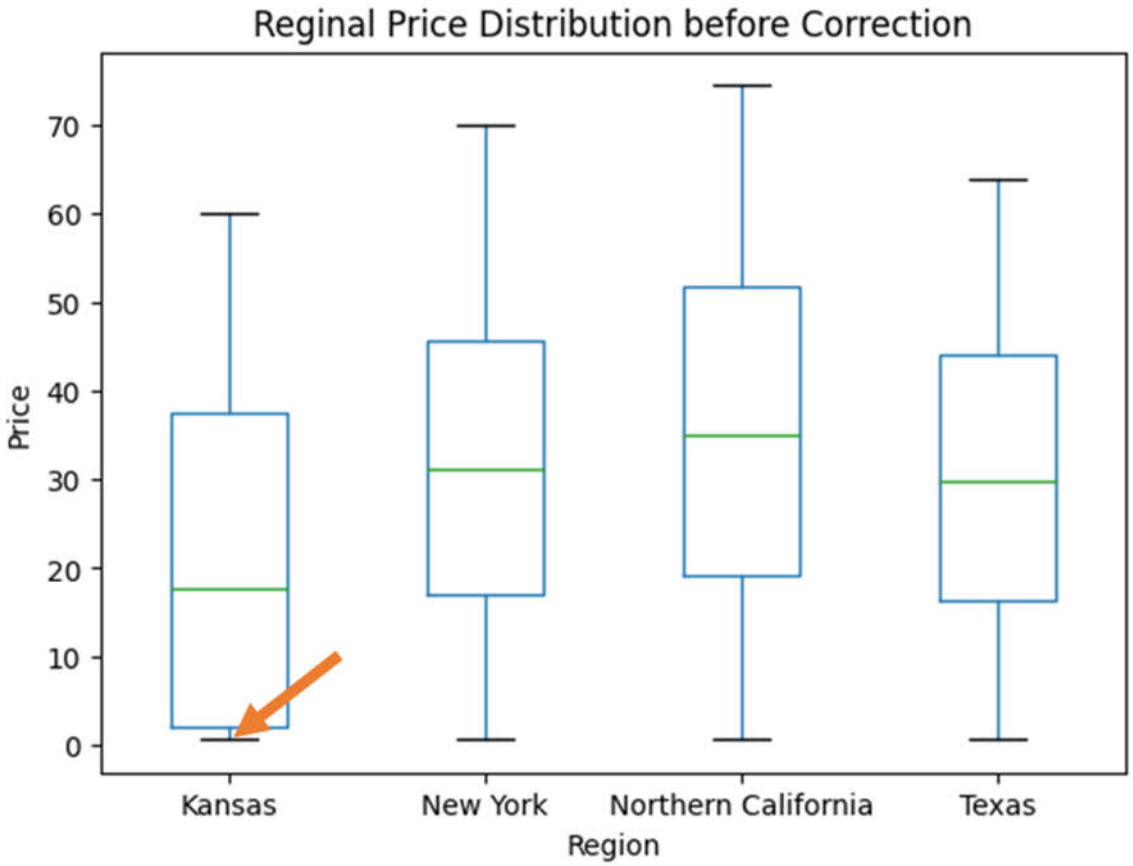

Figure 2 Reginal price distribution before correction

In Figure 2, boxes represent the interquartile range (IQR), which is the range between the first quartile (25th percentile) and the third quartile (75th percentile). Whiskers extend to the minimum and maximum values within 1.5 times the IQR from the quartiles. Green Line represents the median (50th percentile) price for each region.

The median prices, in Figure 2, for Northern California, New York, Texas, and Kansas decrease respectively. This observation aligns with real-world data. New York and Northern California have similar price distributions, characterized by moderate to high prices and less variability compared

to Kansas. Texas also shows a wide price range, but not as extreme as Kansas, indicating a more moderate distribution. However, Kansas displays a distinct pattern with a cluster of low prices, as highlighted by the arrow. This could indicate potential pricing anomalies or unique market conditions in this region.

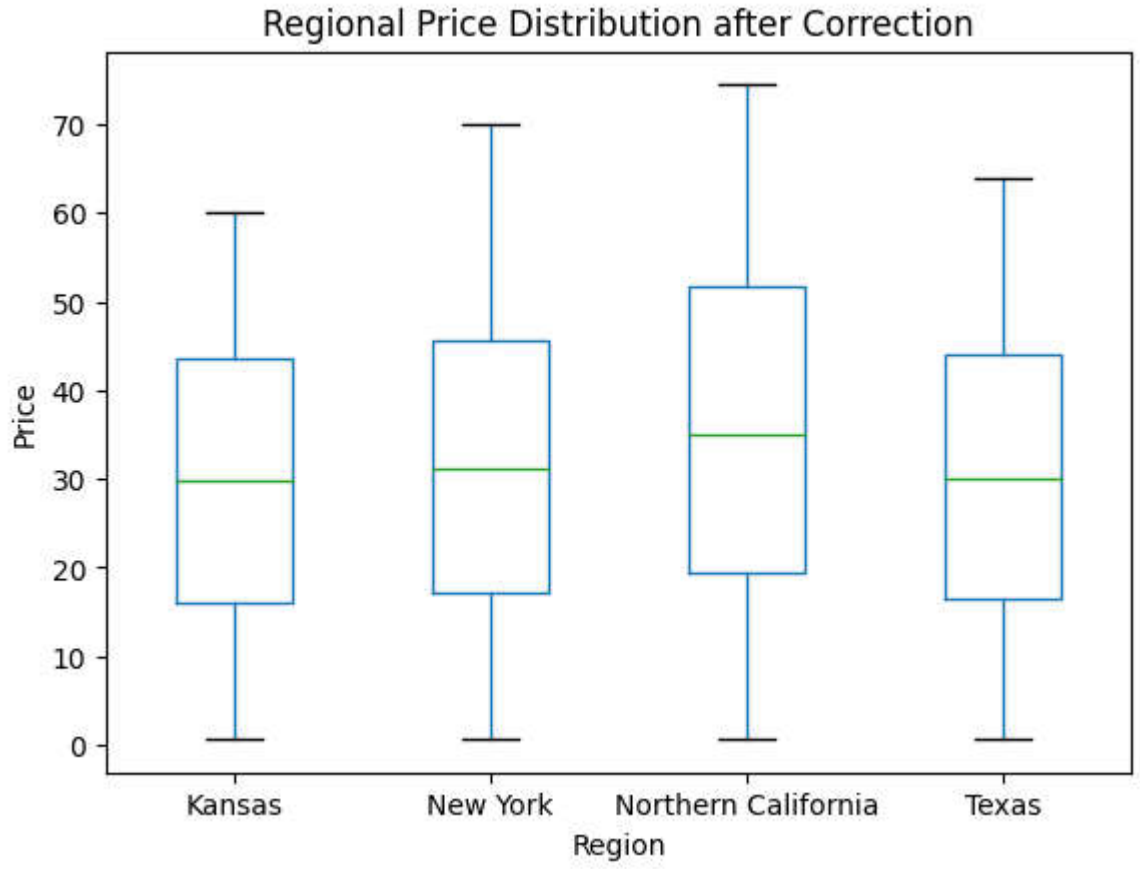

Figure 3 Regional Price Distribution after Correction

After correction, the median prices for all regions are approximately equal, around 30 shown in Figure 3. Kansas shows a balanced distribution with prices ranging from 0 to 60. New York and Northern California display a similar price distribution with a wider range, up to 70. Texas exhibits a narrower price range, up to 50.

Overall, the Figure 3 indicates that after correction, the price distributions across the regions have become more consistent, with similar median prices. However, some variability still exists, particularly in the range of prices in New York and Northern California.

d. Comparison of Regional Price Level

Table 2 Regional Price Index Comparison

	Region	Price_index
0	Kansas	29.64
3	Texas	29.96
1	New York	31.24
2	Northern California	35.14

Table 2 and Figure 4 provide a clear comparison of the price indices across four different regions, with Northern California having the highest prices and Kansas having the lowest. Despite the close values, Northern California stands out with a noticeably higher price index, suggesting that it might have unique market conditions or higher living costs contributing to the higher prices.

In Table 2, Northern California has the highest price index at 35.14. New York follows with a price index of 31.24. Texas has a price index of 29.96. Kansas has the lowest price index at 29.64. The price index ranges from 29.64 to 35.14, indicating a relatively narrow variation in price indices among the regions.

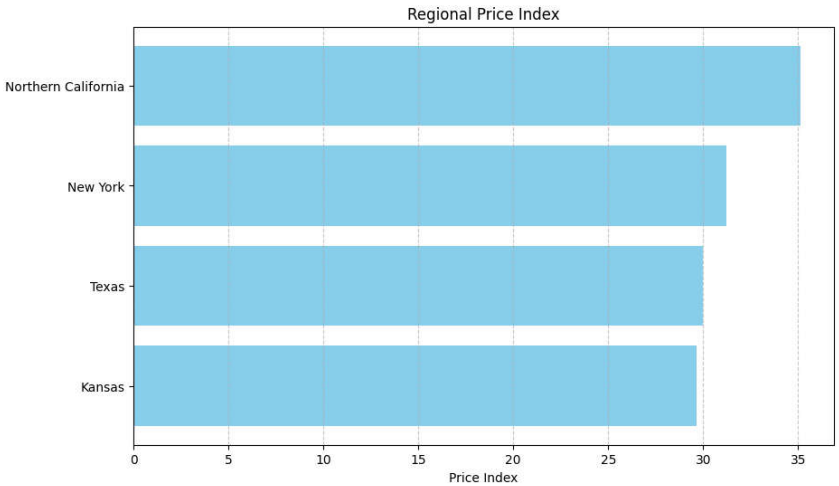

Figure 4 Regional Price Index

The bar chart effectively highlights the difference in price indices across the four regions. The

higher price index in Northern California could be attributed to several factors such as higher cost of living, regional demand, and supply chain differences. On the other hand, Kansas, with the lowest price index, might benefit from lower living costs and different economic conditions.

To gain a deeper understanding of these differences, further analysis could be conducted on:

- **Factors Influencing Price Indices**: Investigate what drives the higher prices in Northern California compared to the other regions.
- **Price Composition**: Break down the price index into categories (e.g., food, housing, services) to see which areas contribute most to the index.
- **Market Conditions**: Analyze the supply and demand conditions in each region to understand the economic dynamics at play.

This figure serves as a useful tool for identifying regional disparities in pricing, which can inform policy decisions, business strategies, and economic studies.

e. Average price per UPC by Region and Store

In Table 3, Northern California exhibits the highest overall price index, with Whole Foods reaching 38.42. This suggests higher living costs or unique market conditions in that region. Among the four regions, Walmart in Kansas has the lowest price index at 27.65. Both Kansas and Texas have relatively lower and more consistent price indices compared to New York and Northern California.

Table 3 Price Index Comparison by Store and Region

	Region	Store	Price Index
2	Kansas	Walmart	27.65
1	Kansas	Trader Joes	29.25
3	Kansas	Wegmans	30.35
0	Kansas	Safeway	30.91
6	New York	Walmart	28.43
5	New York	Trader Joes	30.52
7	New York	Wegmans	30.89
8	New York	Whole Foods	33.83
4	New York	Safeway	35.29
10	Northern California	Walmart	32.85
9	Northern California	Trader Joes	34.38
11	Northern California	Whole Foods	38.42
14	Texas	Walmart	27.96
13	Texas	Trader Joes	29.32
15	Texas	Wegmans	30.12
12	Texas	Safeway	30.62
16	Texas	Whole Foods	32.82

Figure 5 and Figure 6 highlight significant regional differences in price indices across various stores. Walmart consistently offers the lowest prices in all regions, whereas Whole Foods and Safeway generally have higher price indices. Northern California stands out with the highest price indices, reflecting potentially higher living costs or different market dynamics. Kansas and Texas show relatively lower and more uniform price indices, suggesting more stable pricing in these regions. This data can be useful for understanding regional price variations and developing targeted pricing strategies.

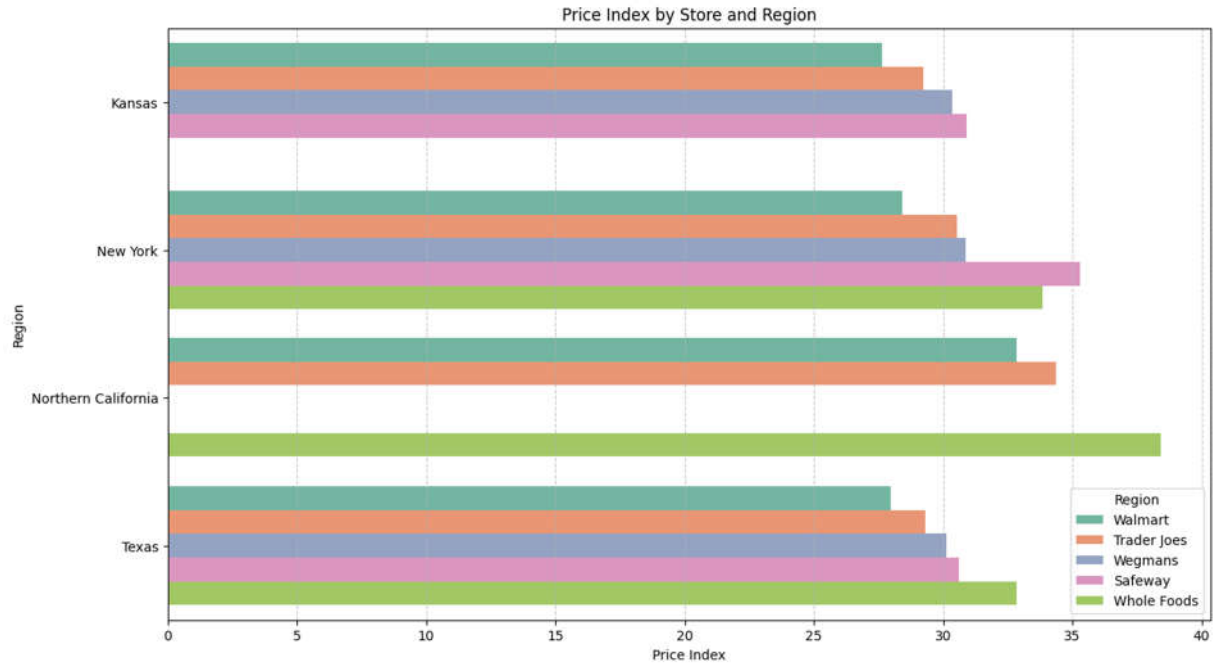

Figure 5 Comparison of Price Index by Store Across Region

Figure 6 Heatmap of Price Index by Region and Store

f. Analysis of Anomalous Data in Kansas Region

Out of a total of 3,087 data points collected in Kansas, 924 data points are identified as anomalies. These data were collected by two auditors: Mike Johnson (Audit_ID: 1326) and Dave Johnson (Audit_ID: 713). Among the 924 anomalous data points, 9 were collected by Mike Johnson, while a significantly higher number, 915, were collected by Dave Johnson.

This discrepancy raises concerns about the consistency and accuracy of the data collected by Dave Johnson. It is crucial to investigate the reasons behind the high number of anomalies in Dave's data collection. Possible factors to consider include differences in data collection methods, external conditions, or potential errors.

To address this issue, we need to determine whether Dave Johnson requires retraining to align his data collection practices with the standards. A thorough review and comparison of his methods with those of Mike Johnson and other auditors will help in identifying any gaps or areas for improvement.

Additionally, it would be beneficial to conduct a detailed analysis of the anomalous data to uncover any patterns or specific issues that may have contributed to the discrepancies. This analysis will provide insights into whether the anomalies are due to procedural errors, environmental factors, or other underlying causes.

Ultimately, ensuring the accuracy and reliability of data collection is essential for maintaining the integrity of our analysis and decision-making processes. Addressing these concerns promptly will help in enhancing the overall quality of our data.