Chase Nuzum
NAZ Beer Data Exercise and Results

See repository provided as appendix and logic.
Readme.md includes instructions to run and pip install

Solve These questions

1. Recommend three products which were not sold in the store (retailer) for the last 6 months.

Based on the analytics, here are three products that have not been sold in the store for the last six months, which could optimize overall revenue:

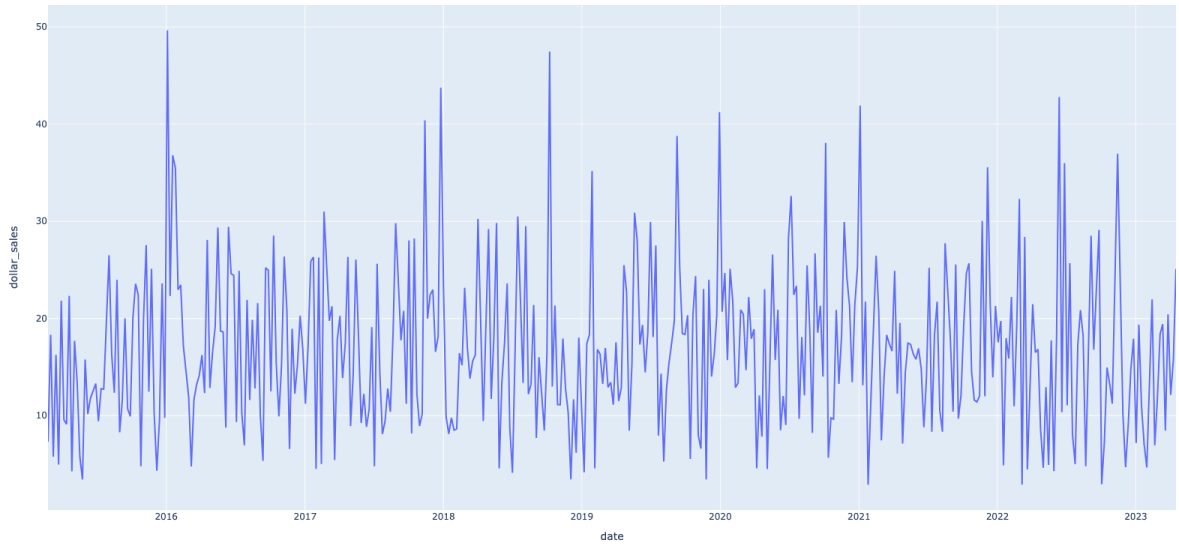
Product Key	Brand	Package	Log Revenue
1402066	Samuel Adams Boston Lager	12-Pack 11.2-13 oz Glass	$7,022.19
188169323	Mike's Hard Party Kit	12-Pack 11.2-13 oz Glass	$6,658.61
609521221	Truly Spiked & Sparkling Citrus Mix Pack Variety Pack	12-Pack 12 oz Can	$6,156.02

What we see here are 3 products that ended production sometime in the last six months that had solid sales and seasonality. Why they were discontinued need further investigation and can be digested by various procurement stakeholders.

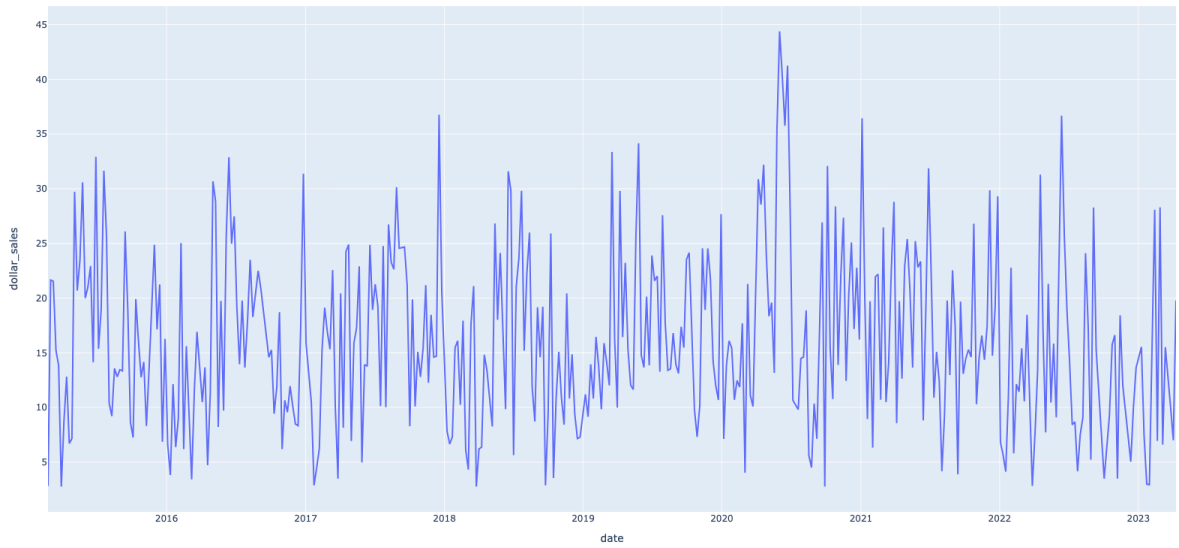
These products have the potential to diversify the product offering and attract new customers to the store, thus contributing to revenue growth.

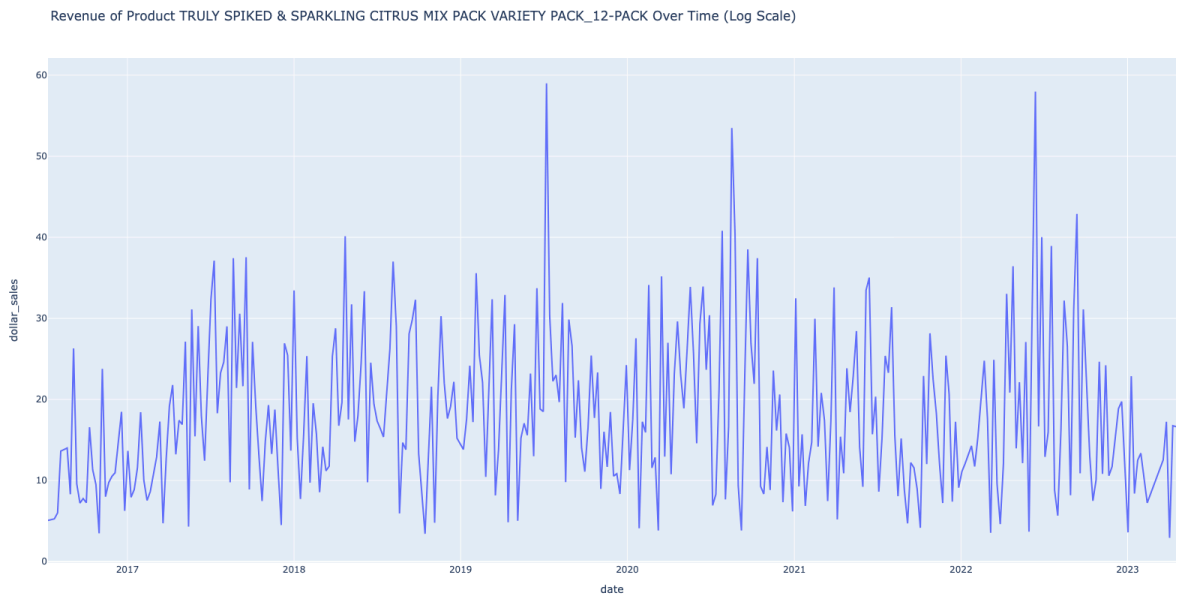
Here is a chart over time to show the performance of these products:

Revenue of Product SAMUEL ADAMS BOSTON LAGER_12-PACK 11.2-1 Over Time (Log Scale)



Revenue of Product MIKES HARD PARTY KIT_12-PACK 11.2-1 Over Time (Log Scale)





As we can see these were constant quantities demanded over time and their discontinuation needs to be investigated further, as they were solid sellers in the markets we serve.

2. Recommend two products which were not sold in the store for the last 6 months to optimize the overall revenue at the store. However, if the store accepts these two recommended products, you have to withdraw two existing products at that store to keep the number of products in that store constant.

Two to drop can be viewed as ‘Non-movers’ or stale inventory. We use the log historical dollar sales to derive the conclusion in this space.

Based on historical data 2 products to bring back (not sold in last 6 months):

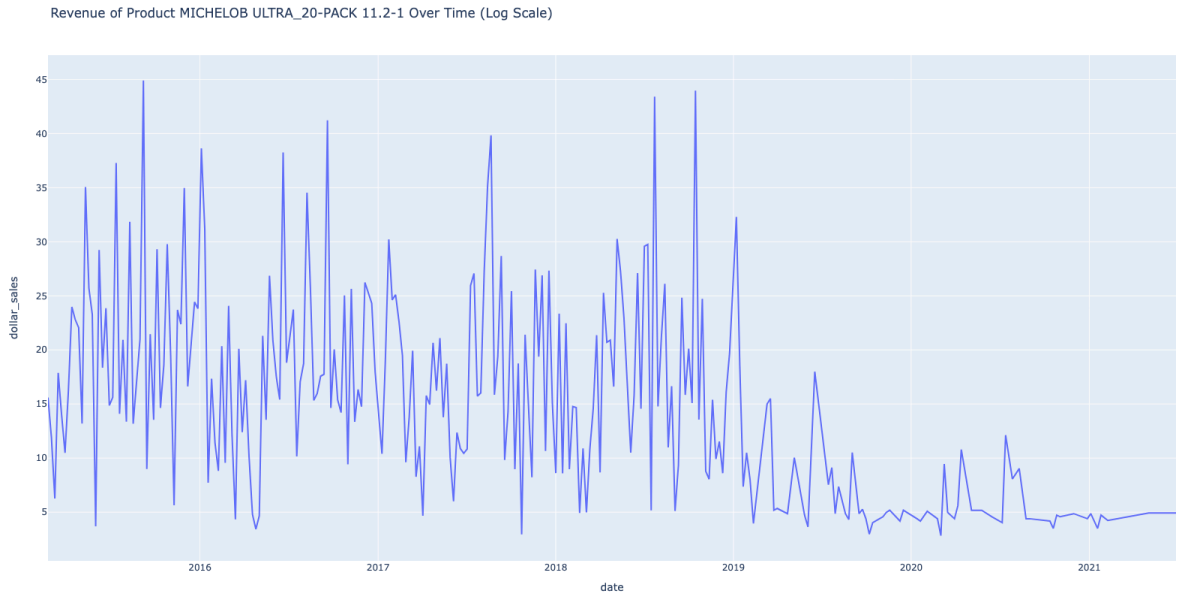
Product Key	Brand	Package	Log Revenue
1401391	Michelob Ultra	20-Pack 11.2-13 oz Glass	$4026.29
32389632	Stella Artois Lager	12-Pack 11.2 oz Can	$4562.63

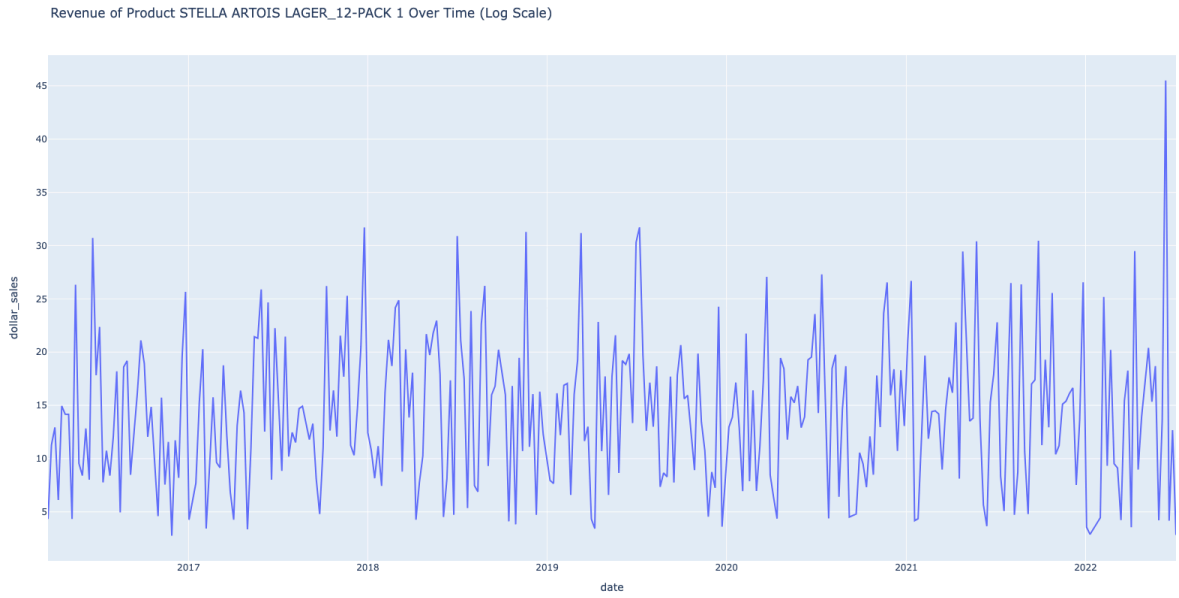
And two products to drop:

Product Key	Brand	Package	Log Revenue
950729043	Omission Ultimate Light Golden Ale	6-Pack 12 oz Can	$2.52
1401002	Busch Light	12-Pack 11.2-13 oz Glass	$3.13

As we can see, the two products to drop are a fraction of the dollar volume that the additions are. Examining these products more closely with marketing, procurement, and leadership will help drive more revenue in the long run.

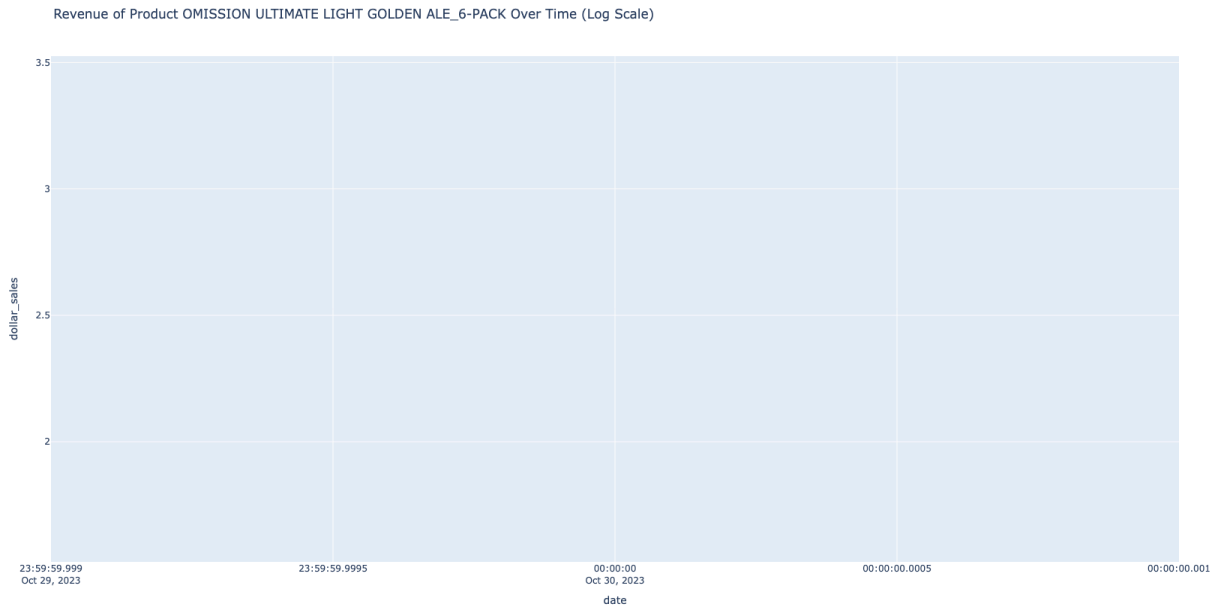
Performance over time for evidence:
Additions:

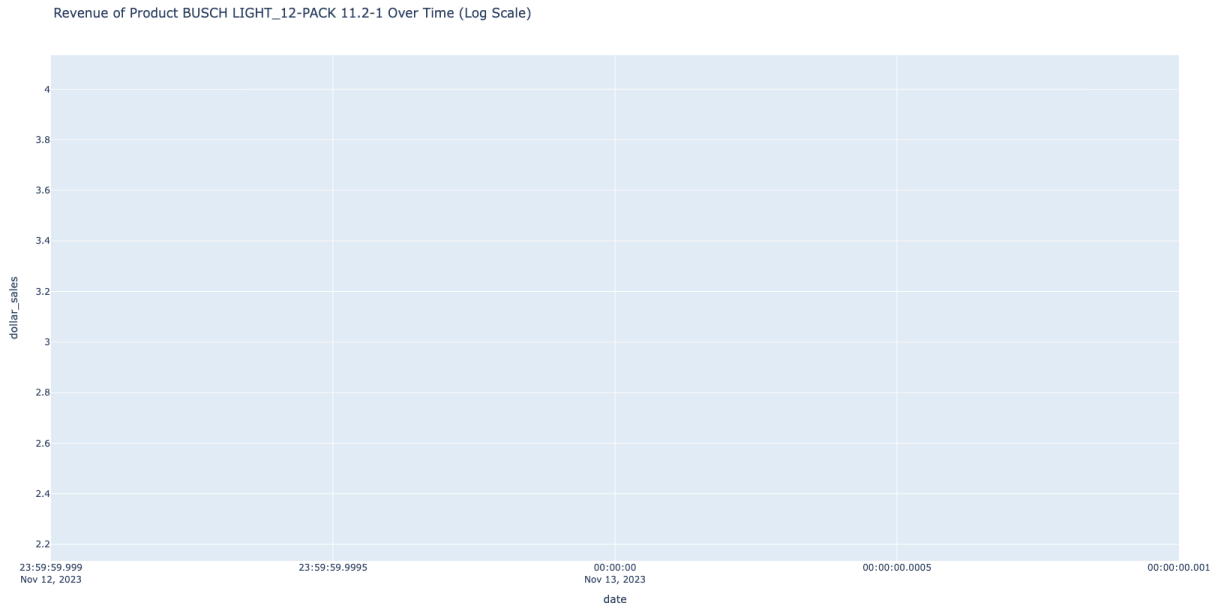




- Good performance, but discontinued before last 6 months. Michelob Ultra might be a candidate to drop still as the trend is definitely in the wrong direction.
- Based on what we see here and second iteration is to test the slope (trend) of the last 2 years before the cutoff as opposed to all historical dollar volume data. The historical performance still constitutes consideration to bring back, based on it once being a solid seller.

Withdraw:





- 1 sale each, if 6 months = in circulation, these would be drops as they are close to zero dollar sales and only purchased based on this evidence.

Next steps for questions 1 and 2 would be:

- Automate notebook functionality in Airflow or another service like Databricks using spark.
- Upload artifacts to delta live table or blob for cloud storage
- Refresh and display graphics that stakeholders want to see in Power BI, Looker, or a Plotly Dash web app (already created plotly artifacts in notebooks)

3. Cluster/segment stores based on their characteristics such as volume level, price level, discounting pattern, geo info, and product portfolio composition (not limited only to these).

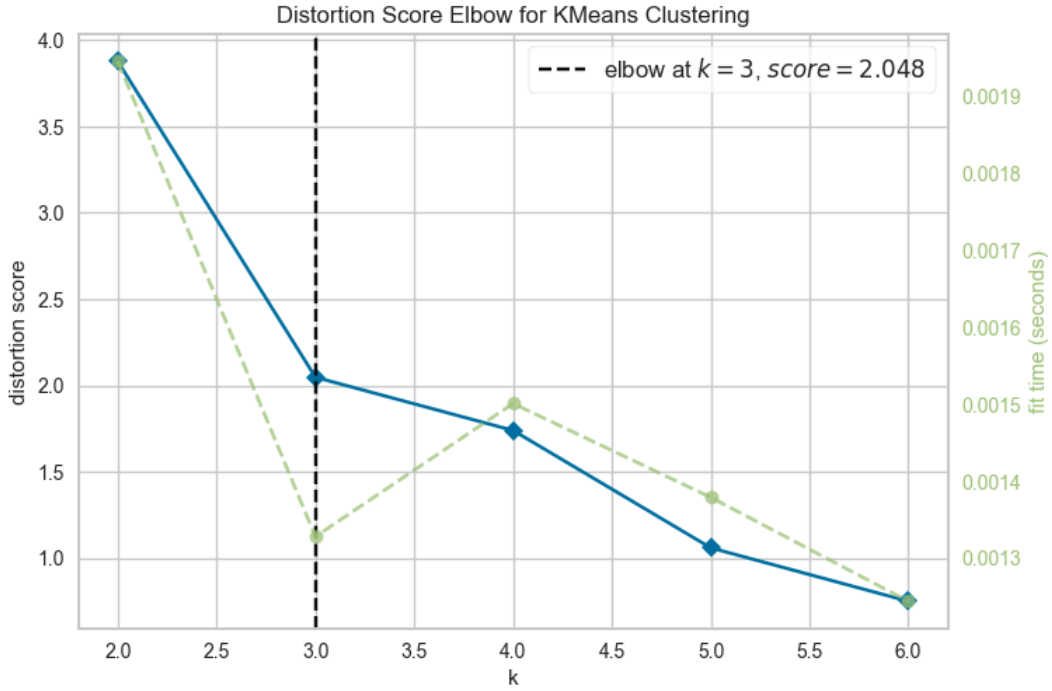
For clustering I ended for the 23 stores (23 cities with 10 retailer IDs):
Full table: store_clusters_with_stats.csv (in /data folder)

Top Brand	Unit Sales	Elasticity	Dollar per Unit	Cluster	Store ID
Bud Light	0.439	0.404	0.542	2	1_LIVERPOOL
Bud Light	0.070	0.174	0.497	0	1_PATTERSON
Bud Light	0.174	0.083	0.477	0	1_SCHENECTADY
Coors Light	0.241	0.154	0.493	0	3_TROY
Coors Light	0.471	0.469	0.510	2	3_WEBSTER
Bud Light	0.711	0.641	0.440	2	30_FAYETTEVILLE
Bud Light	0.155	0.025	0.480	0	30_OGDENSBURG
Bud Light	0.408	0.304	0.476	2	38_AUBURN
Bud Light	0.144	0.082	0.547	0	38_LOCKPORT

Bud Light	0.230	0.140	0.473	0	38_SCOTIA
Bud Light	0.127	0.035	0.586	0	39_NEWFANE
Bud Light	0.268	0.164	0.420	0	39_SARATOGA SPRINGS
Bud Light	0.492	0.506	0.518	2	39_SYRACUSE
Labatt Blue Light	0.572	0.566	0.504	2	91_BUFFALO
Bud Light	0.214	0.146	0.468	0	220_BINGHAMTON
Bud Light	0.128	0.032	0.524	0	220_CHAMPLAIN
Bud Light	0.204	0.087	0.485	0	220_DEPEW
Bud Light	0.109	0.034	0.0	0	220_GENEVA
Bud Light	0.104	0.001	0.378	1	226_OLD FORGE
Bud Light	0.170	0.024	0.558	1	226_OSWEGO

Bud Light	0.157	0.032	0.435	1	226_TONAWANDA
Corona Extra	1.000	1.000	0.337	2	2251_WESTBURY
Corona Extra	0.000	0.000	1.000	1	2431_BROOKLYN

To derive cluster we used the python Kmeans helper Yellowbrick (https://pypi.org/project/yellowbrick/), this gives a good estimation for the number of clusters that are optimal, we ended up with three seen here:



How Yellowbrick works to find optimal number of cluster:

- Distortion score, also known as inertia, is the sum of squared distances of samples to their closest cluster center. It measures how tightly grouped the samples are within each cluster. The distortion score decreases as the number of clusters increases because the clusters become smaller and tighter. However, at some point, adding more clusters will not significantly decrease the distortion score, resulting in an "elbow" in the plot.

With the following features we derived three clusters and added in qualitative features such as ‘Top Brand’ which represents the top brand for retailer (data/top_brands.csv):

Retailer Store Number	Top Brand	Top Brand Log Sales
460	BUD LIGHT	4841.74
3723	COORS LIGHT	4149.78
6764	BUD LIGHT	3650.23
10081	BUD LIGHT	5509.27
12811	BUD LIGHT	5924.27
17118	LABATT BLUE LIGHT	2466.53
19075	BUD LIGHT	7869.80
20686	BUD LIGHT	4063.50
22546	CORONA EXTRA	1894.03
25397	CORONA EXTRA	176.20

To derive clusters with qualitative data, as per requirements, we add them in with K-Prototypes which allows for a mix of qualitative and quantitative features. For more information on K-Means and K-Modes see:
https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

To further analyze the clusters, we want to be sure we can reference pricing information. To do this we used the following model and features:

Model specification:

- Number clusters: 3 (combo visual and elbow analysis)
- Numerical Features:
    - Unit Sales: This represents the volume of sales for each product in the store. It helps understand the popularity and demand for different products.
    - Product Age: The age of a product in the store could indicate its market maturity or relevance. Older products might need price adjustments to increase sales.

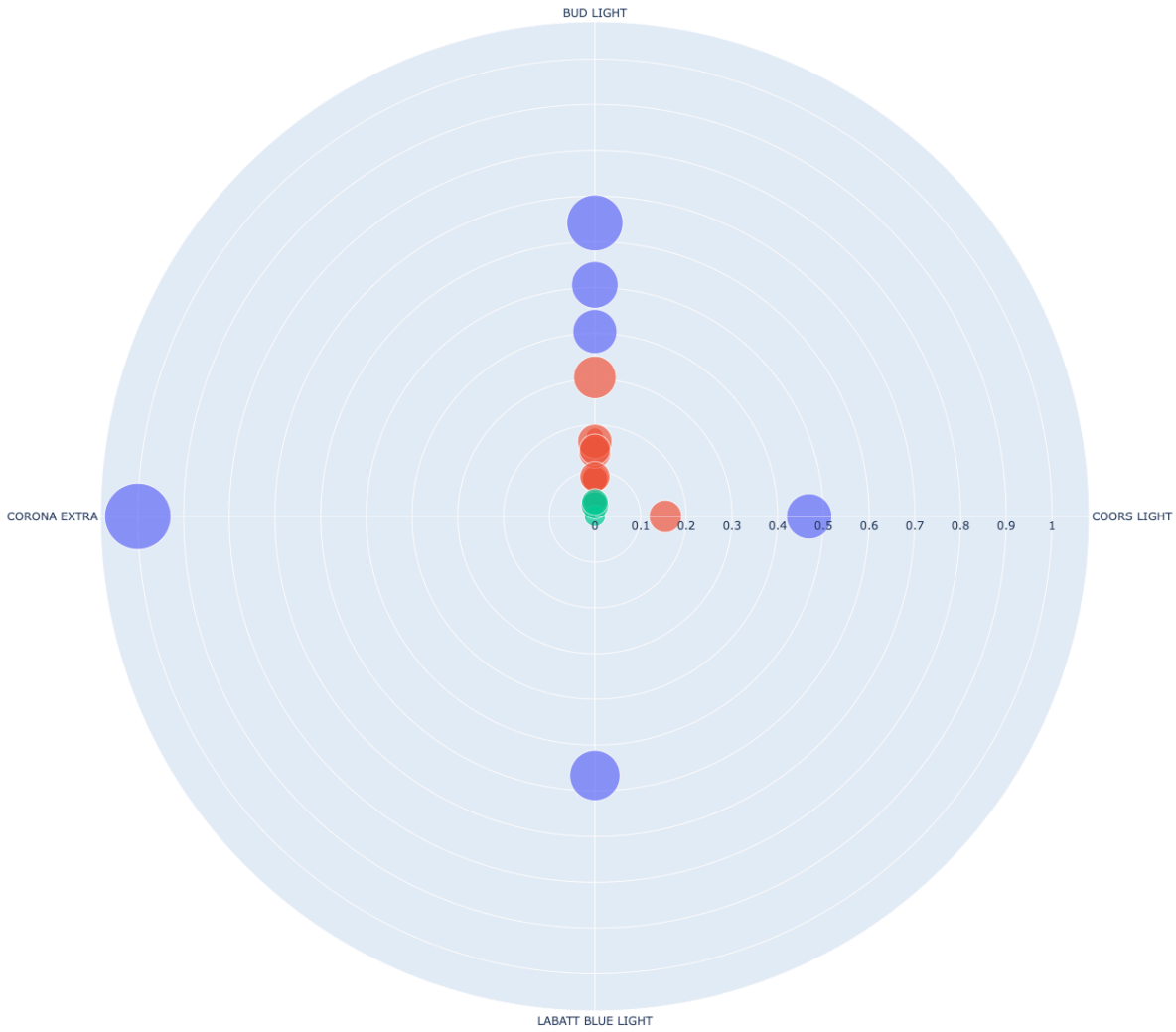
    - Elasticity: This measures how sensitive the unit sales are to changes in price. Understanding elasticity helps in setting optimal prices for maximum revenue.
    - Retailer Sales: This indicates the total sales generated by the retailer. Stores with higher sales might have different pricing strategies compared to lower-performing stores.
    - Dollar per Unit: This represents the average price of each product sold. It helps understand the pricing structure and potential areas for optimization.
- Qualitative Feature:
    - Top Brand: This categorical feature identifies the top-selling brand in each store. Clustering based on the top brand can help understand how brand preferences vary across different store segments and how pricing strategies can be tailored accordingly.
- Used MinMaxScaler so that metrics can be thought of ‘percent’ performance of top performer as opposed to total. (All stats available /data folder).

Note: With this basket of features we start to see separation of ‘centroids’ of interest. When clustering it is important to remember (especially with only 23 entities) that we can run out of degrees of freedom quickly. So using the ‘kitchen-sink’ approach of all features is not applicable. Also we need to be cognizant of our purpose which is grouping stores for pricing.

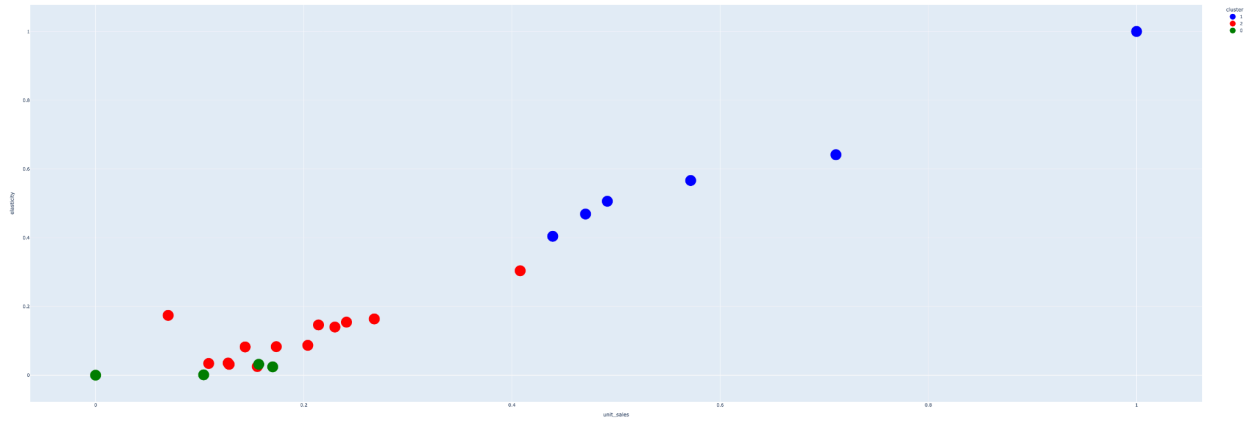
Cluster results:

- **Cluster 0 (Green)**: This cluster exhibits the lowest performance in terms of tracked sales metrics and displays the least elastic demand or most inelastic pricing. This implies that consumers at these retailers are less sensitive to price changes. The top brands in this cluster are Corona and Bud Light. To optimize this cluster's performance, a deeper understanding of the cost structure, total portfolio of products, and the target consumer is necessary. Notably, some stores within this cluster demonstrate a high price per item, suggesting a potential classification as boutique retailers. The hypothesis to test is as follows: High price per item correlates with boutique status, while low price suggests a low performer.
- **Cluster 1 (Blue):** This cluster stands out for its high volume, dollar sales, and total sales. It also exhibits the highest elasticity, indicating a price-sensitive customer base. Positioned in the upper right quadrant for most tracked metrics, it boasts exposure to top brands such as Bud Light, Coors, Labatt, and Corona, with prices falling within the mid-range.
- **Cluster 2 (Red):** Characterized as mid-tier in both associated sales and pricing metrics, these sites can be viewed as "median" sellers, constituting the bulk count of retailers at 12. While their retail performance aligns with that of the Blue Cluster, they demonstrate less price sensitivity and lower volume in units. The top brands in this cluster are Coors and Bud Light, with prices falling within the mid-range.

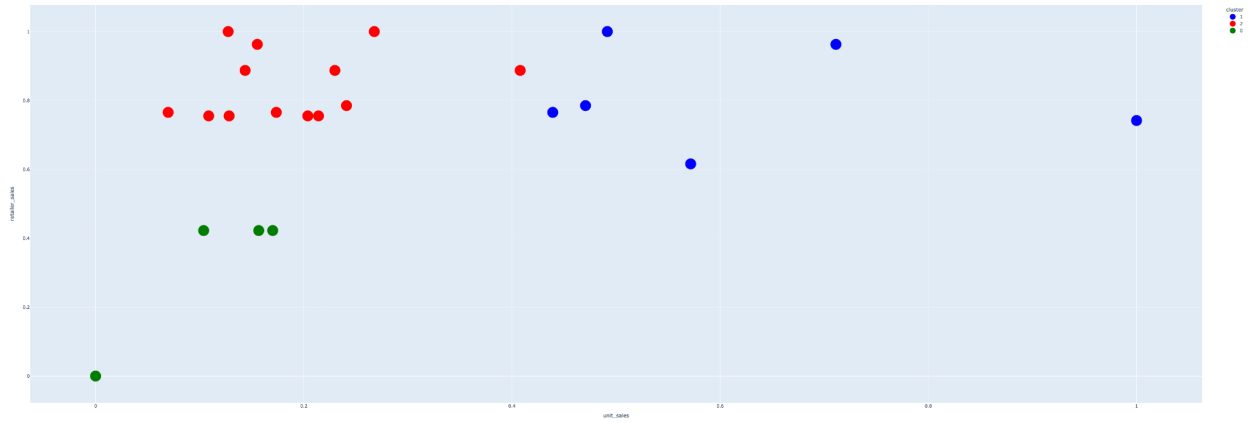
Cluster Polar Chart: Size = Volume Sales, Radial = Elasticity (Price Sensitivity)



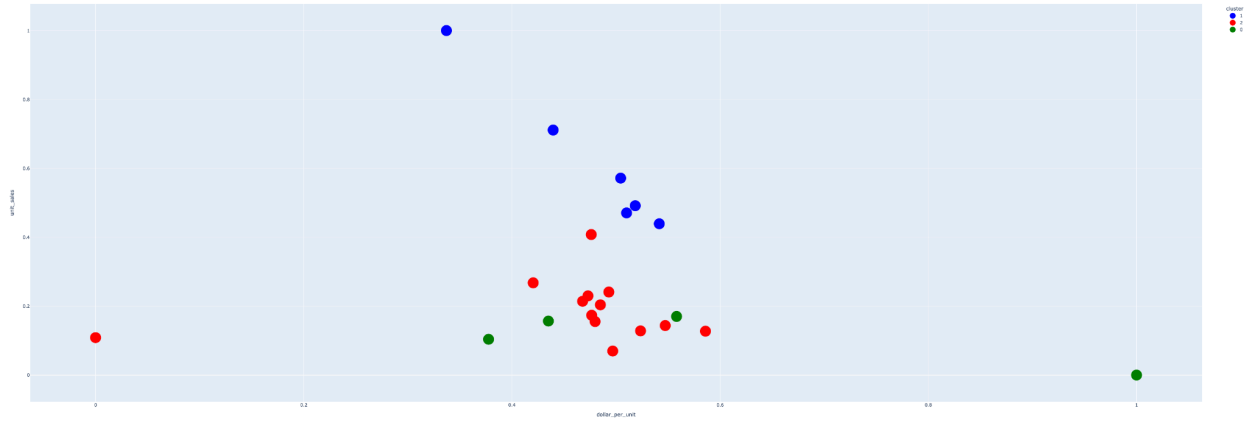
More cluster metrics and centroids for drawing analysis of clusters:
Elasticity over Unit Sales (Volume)



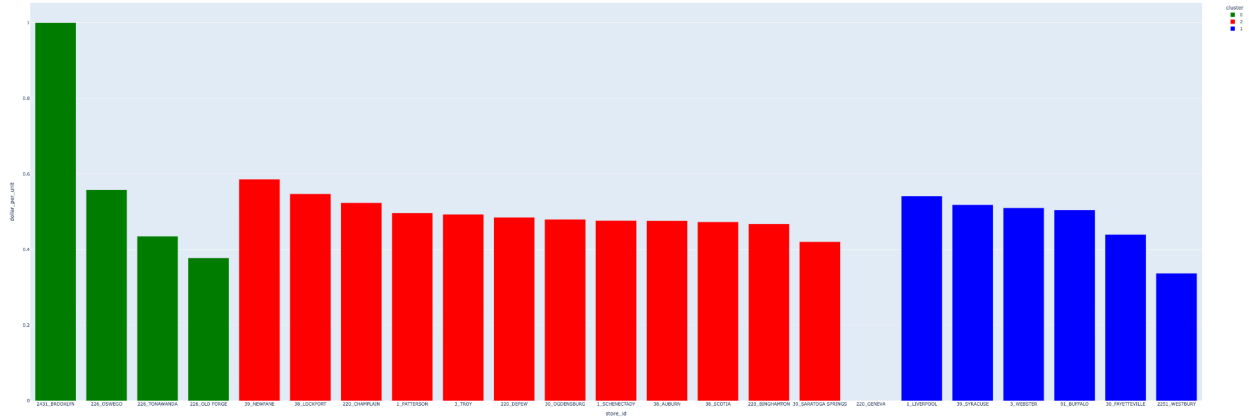
Total Retailer Sales (Dollars) over Unit Sales



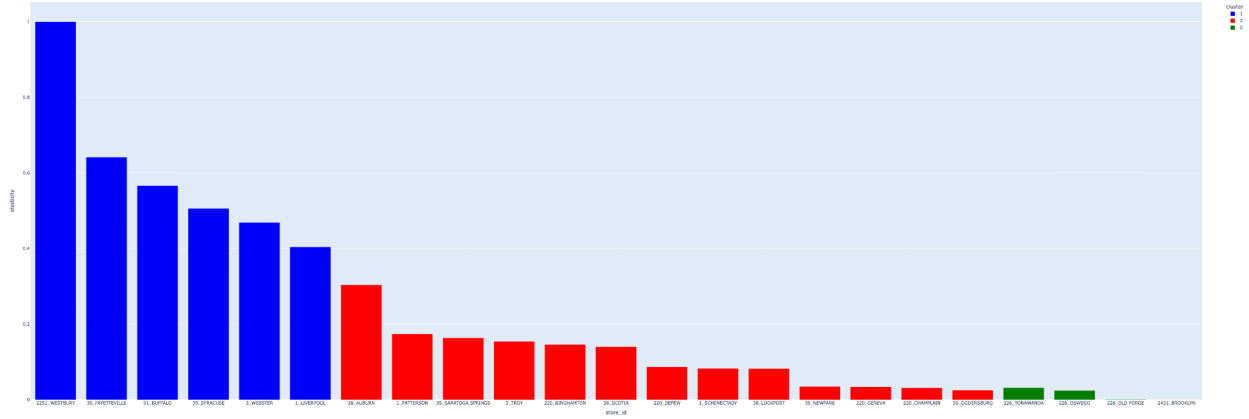
Unit Sales over Dollar Per Item (Price)



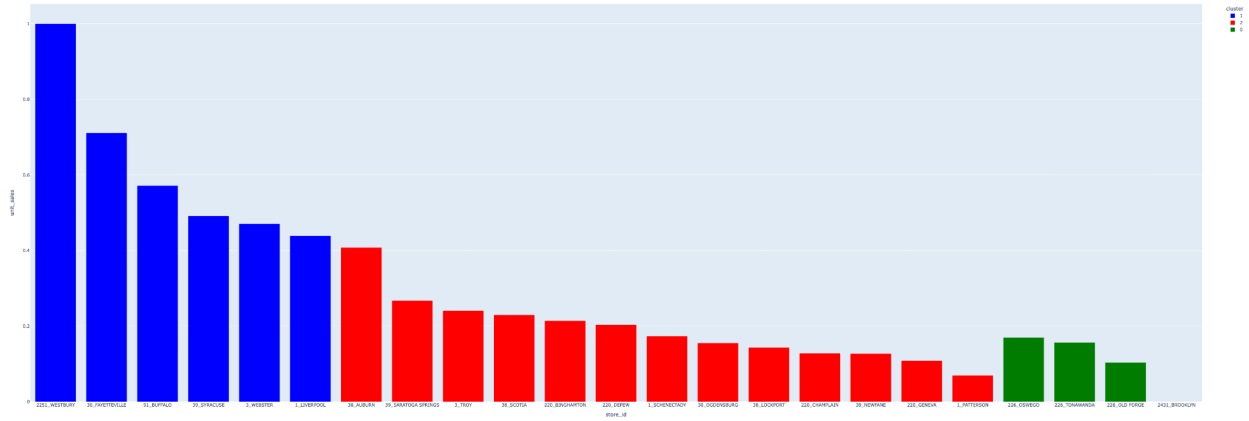
Mean price chart:



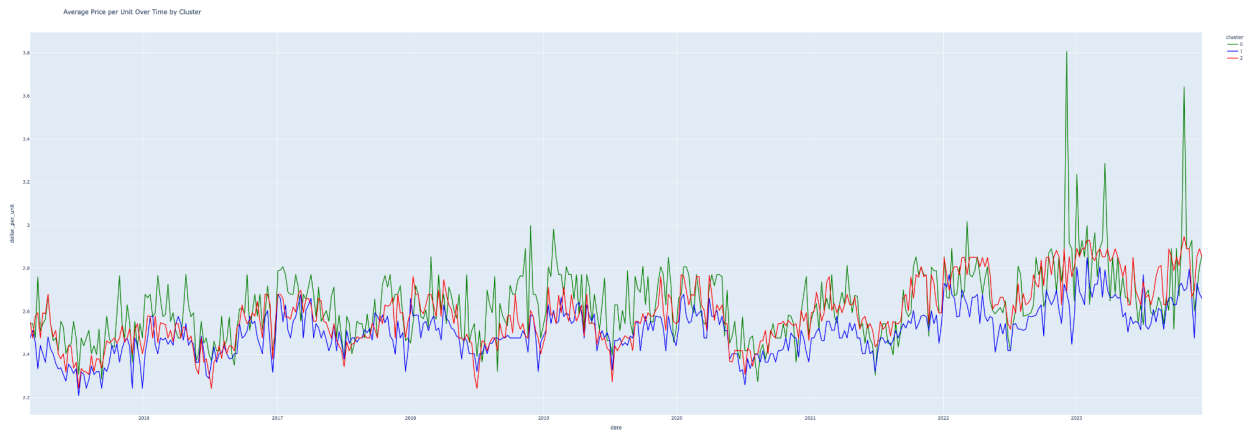
Elasticity chart:



Volume chart:

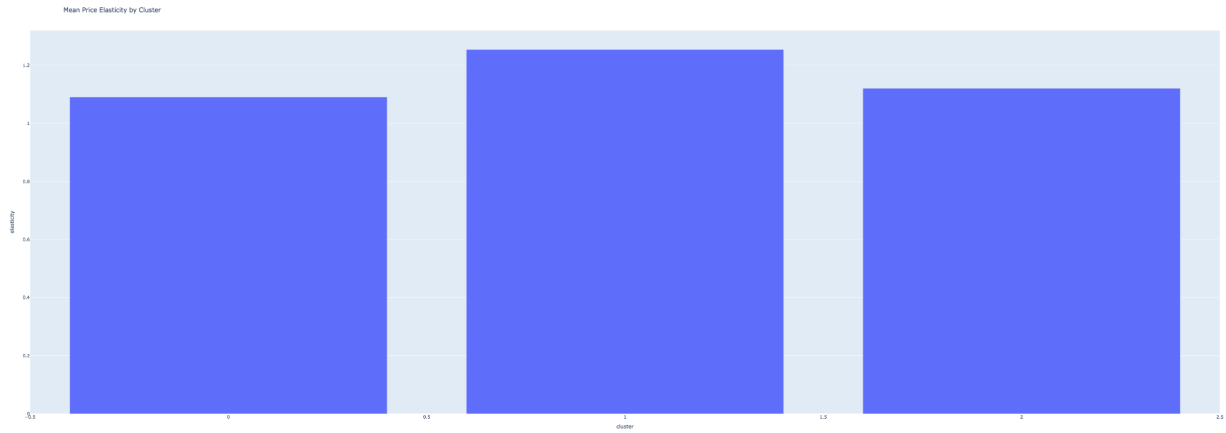


Prices over time by Cluster:



Green = Highest, Red = Median, Blue = Lowest

Mean elasticity by cluster:



Follow the same evidence… Lowest least price sensitive, highest most.

4. How can we differentiate pricing strategy across these store clusters/segments from 3. to optimize the overall revenue?

Cluster 0 (Green):

a. Characterized by low sales performance and least elastic demand, indicating consumers are less sensitive to price changes.
b. Top brands are Corona and Bud Light.
c. It's essential to gain a better understanding of the cost structure, portfolio of products, and target consumer.
d. Given the high price per item in some stores, there's a hypothesis that these may be boutique retailers.
e. Strategy:
    i. Conduct a detailed analysis of the cost structure to determine pricing thresholds.
    ii. Explore opportunities to optimize the product portfolio to cater to the specific preferences of the target consumer.
    iii. Consider implementing premium pricing strategies for boutique retailers while ensuring competitiveness in low-performing stores.
    iv. Focus on brand loyalty and experiential marketing to attract and retain consumers who are less price-sensitive.

Cluster 1 (Blue):

f. High volume, dollar sales, and total sales with the highest elasticity of price-sensitive customers.
g. Top brands include Bud Light, Coors, Labatt, and Corona.
h. Positioned in the upper right quadrant for most metrics, indicating strong performance.
i. Strategy:
    i. Implement dynamic pricing strategies to capitalize on the high volume of price-sensitive customers.
    ii. Leverage promotions and discounts strategically to drive sales without sacrificing profitability.
    iii. Monitor competitor pricing closely to ensure competitiveness while maintaining margins.
    iv. Focus on value propositions and promotions to appeal to price-sensitive consumers while retaining brand loyalty.

Cluster 2 (Red):

j. Mid-tier in sales and pricing metrics, with moderate price sensitivity and volume.
k. Top brands are Coors and Bud Light.
l. Represents a median level of performance, with potential for optimization.
m. Strategy:
    i. Fine-tune pricing strategies to balance profitability and competitiveness within this segment.

ii. Explore bundle pricing or value-added promotions to stimulate sales without significant price reductions.
iii. Invest in marketing efforts to differentiate brands from competitors and capture consumer interest.
iv. Monitor consumer feedback and market trends to adapt pricing strategies accordingly.

Differentiated Pricing Strategy Across Clusters:

- For Cluster 0 (Green), focus on premium pricing for boutique retailers while maintaining competitive pricing in low-performing stores.
- In Cluster 1 (Blue), prioritize dynamic pricing and promotions to appeal to price-sensitive consumers while maximizing volume.
- For Cluster 2 (Red), adopt a balanced approach to pricing, leveraging value propositions and promotions to drive sales while preserving profitability.

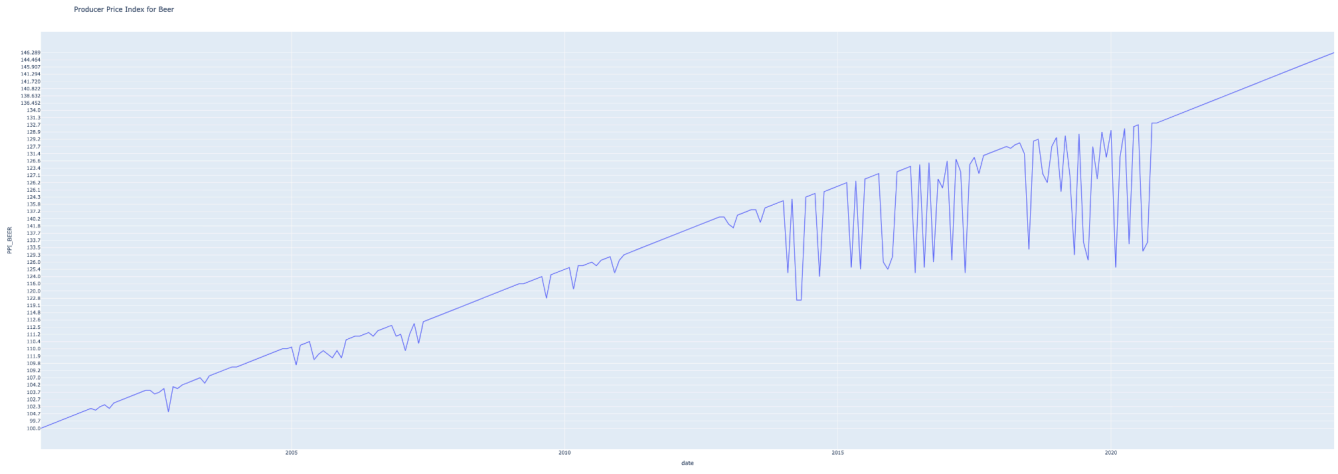
4. (Bonus) Dynamic Pricing with Random Forest Regressor:

To extend this exercise, my pitch for further price optimization would be to create a regression model that we and retailers could use to predict the price per unit. With a model of deterministic inputs that we control we can make a compelling and deployable model that has multiple use cases:

- REST API pricing for sales teams based on characteristics.
- Web app for retailers with predictions to make sure prices are in line for inflation.
- Would lead to transparency and pricing team could use for 'unbiased opinion'
- Peg pricing to unbiased index such Producer Price Index (PPI) (see below) to make sure pricing stays in line with industry cost and price expectations. (could use more correlation analysis here to make sure we select the correct basket of goods from FRED)

Model specifications:

- **Continuous Features:**
    1. Elasticity: Reflects consumer sensitivity to price changes, aiding in tailored pricing strategies across market segments.
    2. Unit Sales: Offers insights into demand patterns, guiding price adjustments to maximize revenue and market share.
    3. Unit Size: Influences pricing decisions; larger sizes may command higher prices, while smaller sizes appeal to budget-conscious consumers.
    4. Retailer Sales: Provides valuable insights into consumer preferences and market trends, facilitating tailored pricing decisions.
    5. Producer Price Index (PPI) for Beer: Indicates industry price trends, helping adapt pricing strategies to remain competitive.
        a. U.S. Bureau of Labor Statistics, Producer Price Index by Industry: Beer, Wine, and Liquor Retailers [PCU4453144531], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PCU4453144531, March 27, 2024.

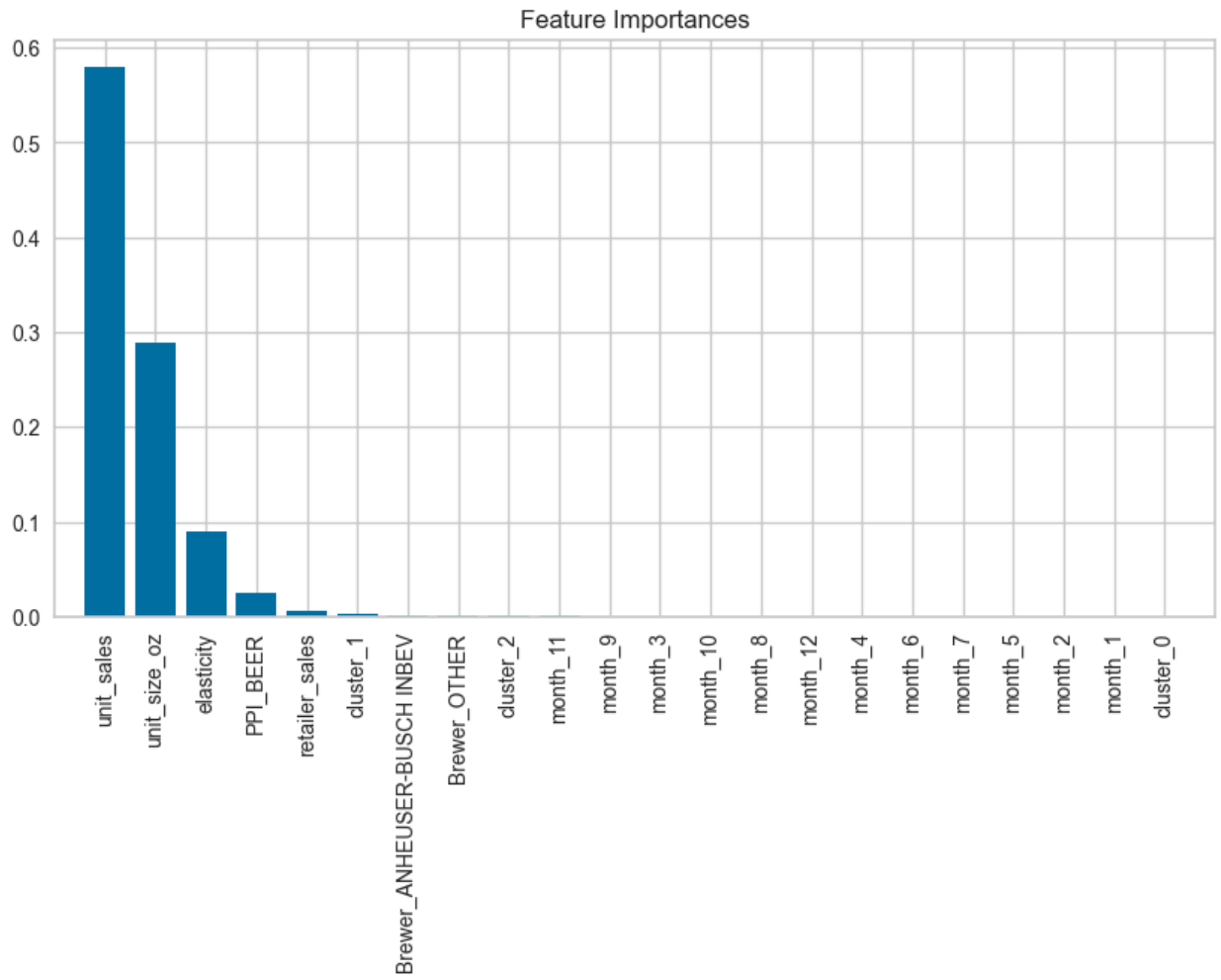


* constant trend and forecast future price level based on FRED.

- **Qualitative Features:**
    1. Cluster: Captures variations in consumer behavior and preferences across segments, enabling personalized pricing strategies.
    2. Month: Accounts for seasonal demand fluctuations due to factors like weather and holidays, optimizing pricing strategies accordingly.
    3. Brewer: Identifies beer brand and its impact on pricing, allowing for differentiation in pricing strategies among various brands and competitors.

Model Results:

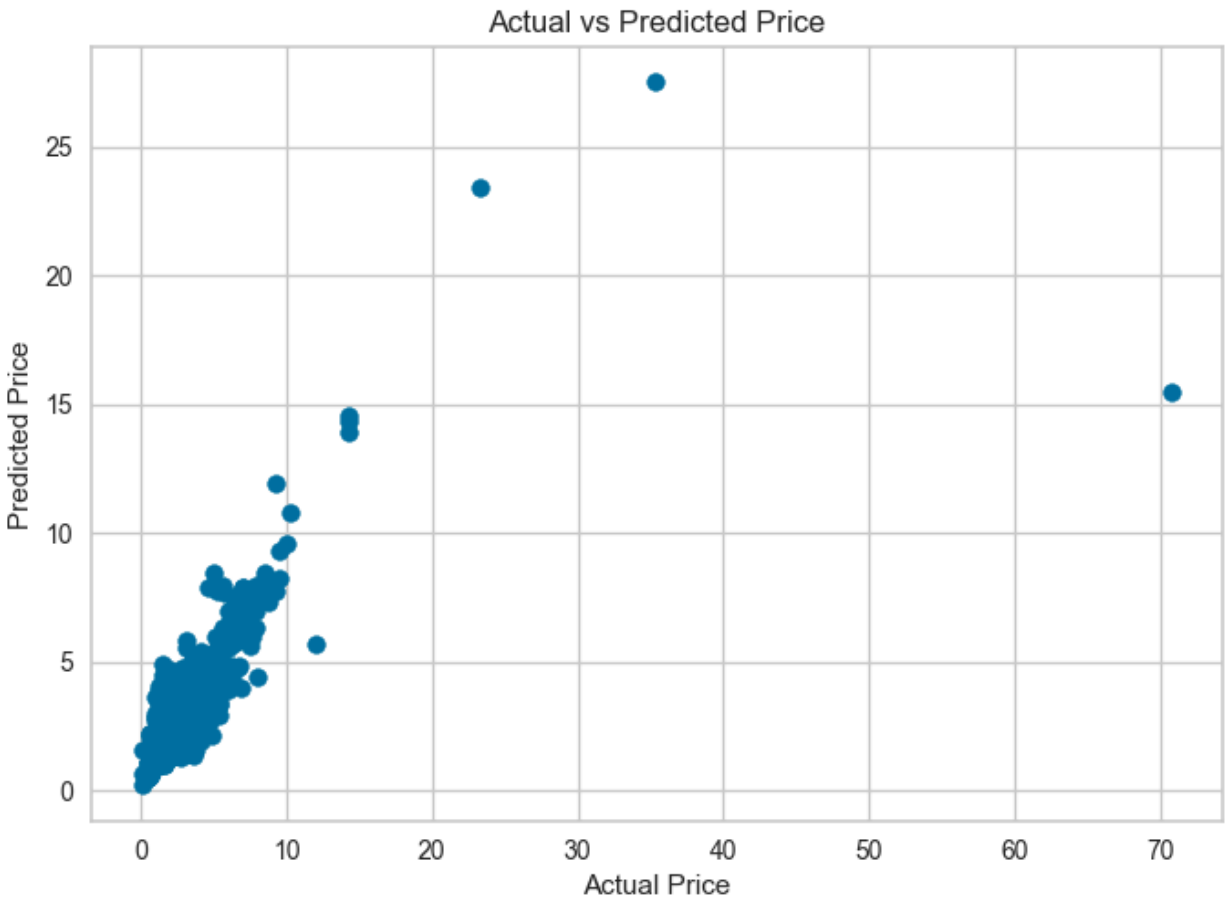
Feature Importance:

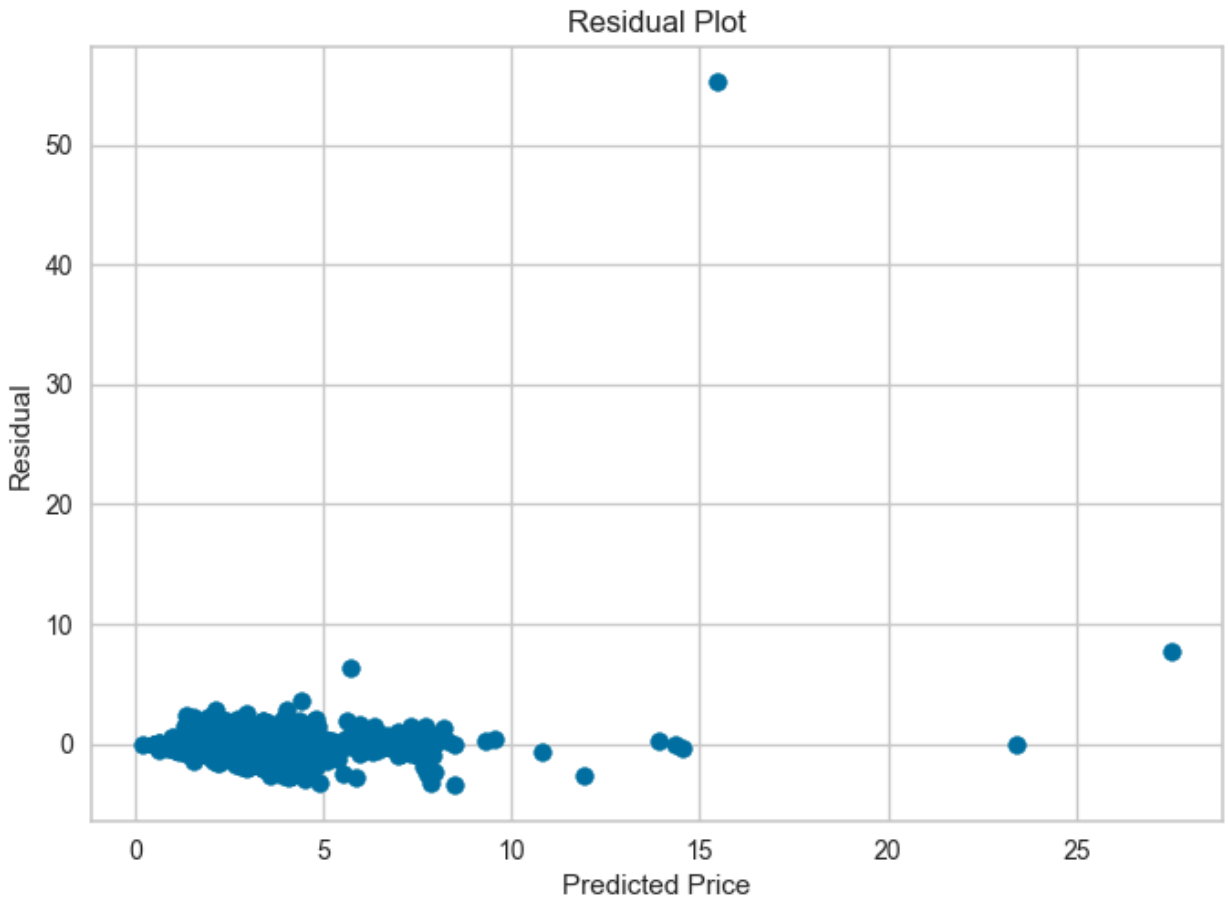


Best features or entropy reducers (randomness) are unit sales, unit size, elasticity, and PPI (indexed cost). Less importance shown as well.

Metrics

- R-squared: .96 (highly predictive, could be overfitting, need more diagnostic, but out of scope at this time).
- Mean squared error: .02 or roughly $0.02 cents off.


Actual vs Predicted Price
Predicted Price
Actual Price
0
5
10
15
20
25
0
10
20
30
40
50
60
70



* mostly well configured regression… one outlier error, needs investigation.
Artifacts in /models in repo:

- Rf.pkl (random forest object)
- Onehot.pkl (one hot encoder)
- Scaler.pkl (standard scaler)
- Kprototypes.pkl (clustering model)

Next steps would include:

Implementation of Decision Support Tools:

- Integrate the pricing model with user-friendly decision support tools like dashboards and reports for easy access and interpretation of model predictions.
- Provide stakeholders with actionable insights and recommendations to support pricing decisions effectively.
- Ensure seamless integration of these tools into existing workflows for streamlined decision-making processes.

Continuous Monitoring and Optimization:

- Implement robust monitoring systems to track the pricing model's performance in real-time, flagging any deviations or anomalies.
- Continuously optimize the model based on feedback, market changes, and evolving business requirements to enhance its accuracy and relevance.

- Ensure proactive management of model updates and adjustments to maintain its effectiveness over time.

Change Management and Governance:

- Establish clear governance processes and policies to govern the use of the pricing model, ensuring compliance with data privacy and regulatory requirements.
- Implement change management practices to address any organizational resistance or challenges associated with adopting the model.
- Foster a collaborative culture that encourages transparency, accountability, and alignment with business objectives in the use of the pricing model.

More to explore!

From EDA and modeling other compelling pieces of data and evidence for business intelligence:

Artifacts of note for appendix:

- data/<feature>_values_counts.csv
    - Includes data related to ratios of object or string features (qualitative)
- data/correlation.csv
    - Correlation between continuous features
- data/seller_data.csv
    - Seller related datetime statistics
- data/top_brands.csv
    - Top brands for each seller
- data/elasticity_key.csv
    - For each product a price-demand elasticity (great data)
- data/processed_data.csv
    - Cleaned data with plethora of intriguing features
- data/store_clusters_with_stats.csv
    - Cluster and percentile store characteristics