Joe Paul, Bartley Cai

AUE6930 Data Science and Machine Learning

5/2/2022

Case Study- Data Analysis and Recommendations for The Velo Fellow

Introduction

The Velo Fellow is a small, single location bar located in Greenville, SC. The bar was opened in 2011 and has grown a reputation as a relaxed, comfortable space with great people. Greenville has experienced a population growth of nearly 30% since the bar opened ("Greenville, South Carolina Population 2022 (Demographics, Maps, Graphs)."), and competition has also increased simultaneously. Greenville was named one of the fastest growing cities in the US (Vasel), and this growth has expanded the potential customer base and increased the importance of remaining competitive with newer, growing businesses.

For The Velo Fellow (Velo) to remain relevant, it must adjust to the times and begin intelligent processing of its available data to understand shifts in customer preference, the competitive landscape, and how to price its products. Small businesses, like Velo, typically do not have the resources and staff with data analysis experience to deeply analyze their available data and coupled with typically aged point-of-sale (POS) systems, often makes small businesses reliant on relatively basic processing methodologies. Businesses like Velo often use standardized markups to price their products, and order products based on intuition from the management staff about what is selling well. We, as students of the AUE 6390 Data Science class, therefore, believe that we can utilize this data to illustrate the potential benefits of machine learning and data processing techniques to illustrate the changes Velo has experienced and make business recommendations accordingly. Due to the availability of information provided by Velo, this report will primarily focus on beer, wine, and liquor sales, despite Velo also operating a full kitchen and food sales making up roughly 1/3 of its overall sales.

In this report, we will discuss our analysis of data provided by Velo and make recommendations based on the available data. We will also recommend future data management strategies that would allow for a better analysis and understanding of the bar's business. The report will first discuss the industry background, our original problem description, our approach, the results of our analysis, and finally conclusions and recommendations.

Industry Background

Very limited information is available in terms of formal research studies of beverage pricing. The Beverage Information Group, a branch of Magazine ("Cheers Online, About Cheers."), issues pricing and analysis guides for this industry, however, these guides are primarily intended for use by businesses and therefore are quite costly (roughly $1,400). Due to our project only being funded in the sense of the businesses time, these reports were not accessible. Other research is available in relation to beer, wine, and liquor (BWL) pricing, however this research is primarily centered around health and public policy implications rather than profitability (Sharma, Sinha, and Vandenberg 149).

Some research is available on pricing elastics, the study of how price increases effect product demand levels. This research has been conducted by multiple groups, (Gruenewald et al. 96) (Jiang et al. 222-228) (Ornstein and Levy 303-345), and primarily focuses on the complexity of alcohol pricing due to the categorical nature of the involved goods (Gruenewald et al. 96). This categorization typically divides alcohol into the categories of beer, liquor, and wine, and differentiates between pricing categories within (top shelf, well). These categories are important because even subtle changes between them can lead to demand decreases in more profitable products and increases in less profitable products.

Velo, like many bars and restaurants, therefore, instead prices their goods based on cost rather than seeking profitability optimization. This strategy primarily involves pricing goods as a fixed percentage increase over their cost ("2nd Kitchen, Liquor Pricing ."). This process is a good heuristic to ensure that products are not being sold at a loss and attempts to roughly include labor costs and a profitability margin into the price of each drink.

Many bars in Greenville offer similar product ranges in part to Velo, but Velo is unique by combining the aspects of in-house beer production, alcohol sales, and a full restaurant. This is relatively atypical. A very geographically close competitor, Vault and Vator, only offers a very limited array of beers without a complete kitchen ("Vault and Vator Homepage."). Another competitor, Society Sandwich Club, has a more complete bar restaurant but does not make any of its own beer ("Society Homepage."). Breweries are also very popular in the area with over 20 being present in the small city ("Your Ultimate Guide to Greenville Breweries.").

These factors make a few important items obvious. Competition is steep in the area, but Velo has a unique service offering relative to most competitive businesses. Additionally, price elasticity is a tool that Velo could utilize to intelligently approach their pricing and sales strategy.

Problem Description

Every business has the goal of maximizing revenue while minimizing costs. Revenue can be increased by attracting more sales, which can be accomplished by different pricing models, discounting, marketing towards a target audience, etc. Costs can be reduced by determining the minimum amount of labor needed for the business' processes, recycling or reusing materials, and efficient usage of inventory. Many of these are methods often utilized in supply chain management.

In the context of Velo, management orders and maintains an inventory of wine, liquor, and beer. Management subsequently determines pricing and promotions but rather than make data-driven decisions, pricing is decided by "feel". Thus, Velo may be losing out on revenue on products that customers may be willing to pay more for. Additionally, products that do not drive demand also incur holding costs which are costs due to storing unsold inventory and can include labor, insurance, and spoiled products. Accordingly, the problem or goal for Velo is to increase their revenue by making changes to their pricing while reducing costs.

Throughout the course of the project, other opportunities for analysis were identified based on the available data. Given that sales data was provided for a period spanning from 2019 to 2022, an opportunity for evaluation of sales mix changes was allowed. Additionally, some competitor pricing data was also compiled to conduct competitive pricing analysis.

The factors influencing these problems are primarily price and cost. The relation of these two key characteristics determines product profitability and aggregates to total profitability. Additionally, product categorization is important to understand the sales landscape and the types of products that customers are buying. These factors also are key in understanding strategic changes in ordering and pricing that may affect profitability long term.

Approach

Data Collection

Data was collected from Velo's POS system. The collected data was a per-product aggregate sales summary for each year from 2019-2021, with partial year data collected for 2022. This data contained pricing, sales volume, total sales, product name, and product categorization information. The product categorization, a 2-letter code, was determined to be relatively unreliable and unmeaningful, with codes showing no obvious correlation to the product type or otherwise. These codes were therefore largely not considered in analysis. Additionally, purchasing information was only obtained for the year of 2021. This dataset contained information on the product name, purchase unit size, purchase unit price, units purchased, total cost, as well as supplier information. The data available from each of the two data sources is summarized in table 1.

Data Source	**Sales Data**	**Purchasing Data**
Columns	Product Name	Order Data
	Product Price	Product Name
	Units Sold	Product Type
	Sales	Unit Cost
	Category	Units Ordered
		Total Cost

Table 1. Summary of provided data.

This data required significant cleaning and preprocessing to be useful for analysis. One of the lengthiest requirements was the updating of dissimilar product names between the two datasets such that the two sets could be combined for further analysis. Additionally, the ordering size of products was often not the same as the sales unit size. For example, kegs of beer are typically sold as an entire unit, where they are sold by the pour to customers.

Data Supplementation

The data was supplemented through the addition of several calculated columns. These columns utilized matches between the 2021 sales and ordering datasets. These calculated columns included cost per sale unit, cost percentage relative to purchase price, profit per unit, and total profit per product. The calculation of these columns assumed that the 2021 ordering dataset encompassed all the products sold within 2021. This is not necessarily true as many products, especially top shelf products, typically sell at low enough sales volumes that they are held in inventory over several months. This process was useful at reducing the effects of the extraneous variable of keg yield, where a portion of a keg is lost as foam or waste and not sold to customers. Additionally, kegs typically are sold with a refundable deposit attached for the keg itself, although the credit for returning the kegs was not considered in analysis.

Finally, to utilize a price elasticity model, the addition of a pricing categorization was necessary for each of the products within their respective BWL landscapes. This categorization was estimated based on product knowledge and relative pricing within the set.

Exploratory Data Analysis

Two separate analyses were done, one for the summary sales from 2019 to 2022, and the other for the 2021 sales combined with ordering data. For the summary sales data, the numerical variables include product price, units sold, and sales. Various metrics were computed to gain a sense of the data's shape and outliers, which can be seen in Table 2 below.

All three variables exhibited high skewness and kurtosis, indicating a heavily right skewed dataset with large positive outliers; boxplots showing the large number of positive outliers are shown below in Figure 1. Two products in particular, 'Coors Banquet, 18pk' and 'George Dickel' are massive sales outliers. In fact, these two products are responsible for the top 6 outliers, one each for each year between 2019 and 2021.

Metrics	Price	Units	Sales
Max	40	12964	3.49E+04
Min	1	1	0
Mean	6.78	141.17	7.06E+02
Median	6	31	1.96E+02
Stdev	4.32	528.42	1.91E+03
Var	18.7	279235.9	3.67E+06
IQR	2	84	4.94E+02
Skew	4.27	13.12	8.30E+00
Kurtosis	22.03	243.61	1.00E+02

Table 2. Metrics of Sales Summary 2019-2022

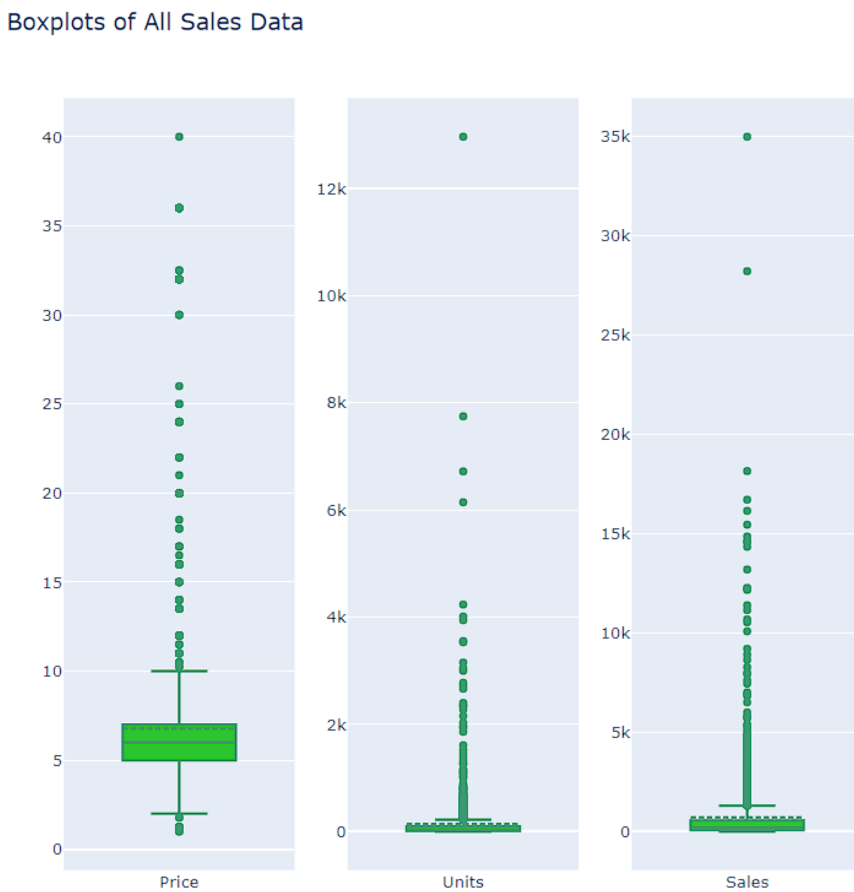

Figure 1. Boxplots of Summary Sales

Categorical analysis was also performed on the data to visualize the popularity of product types. A histogram of the product categories is shown below in Figure 2. Beer is the most popular, followed by spirits, and then wine.

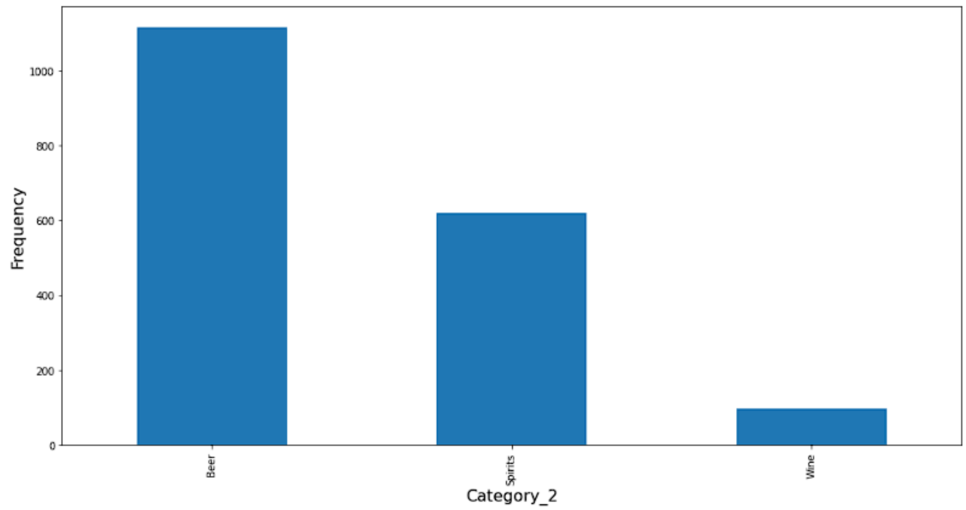

Figure 2. Histogram of Product Category

The summary sales data was also used to evaluate the projected performance for The Velo Fellow for 2022. To uphold confidentiality, relative performance was gauged, rather than explicit sales numbers, by taking the sum of all sales, setting that equivalent to 100%, and determining the contributing percentage of each year. This resulted in 33% for 2019, 22% for 2020, 35% for 2021, and 10% for 2022. The sales data for 2022 is collected up to mid-April. Assuming sales are linear throughout the year, the projected 2022 sales should be about the same as 2021. Since there are likely seasonal demand changes that affect the final output of 2022, for example sales ramping up through the summer before waning through fall and the start of winter, an accurate prediction cannot be determined without historical day-to-day sales to reference. A chart of the relative sales is shown below in Figure 3.

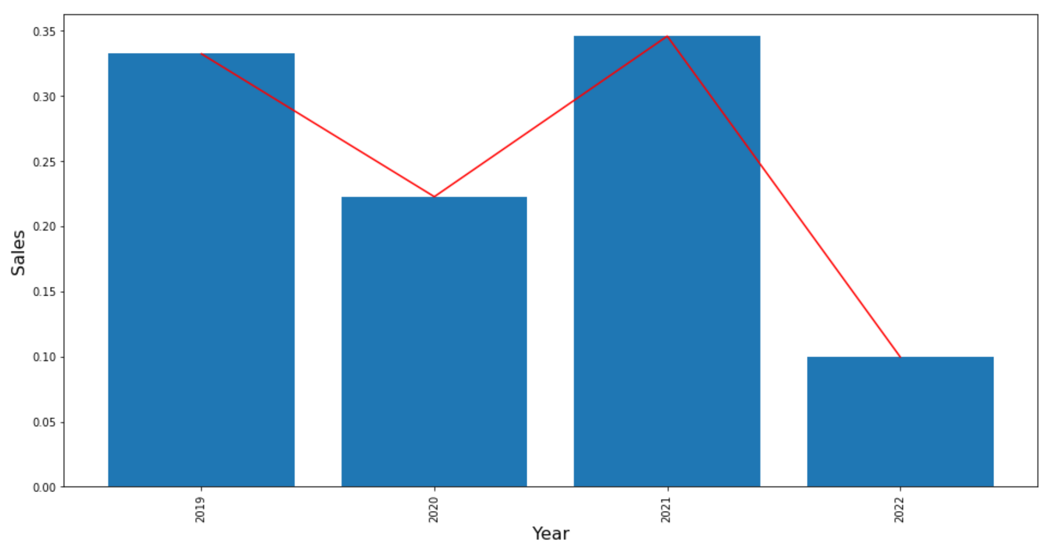

Figure 3. Relative Sales Between 2019 and 2022

A heatmap using Pearson's coefficient and pair plots were also constructed on price, units, and sales. As expected, units and sales were highly correlated, whereas price had no correlation with either variable, perhaps very weakly negatively correlated. With a higher sample size, the negative correlation could potentially become stronger, which suggests that the higher the product price, the lower the units sold, though the effect on the correlation with sales is indeterminate. The heatmap and pair plots are shown below in Figure 4.

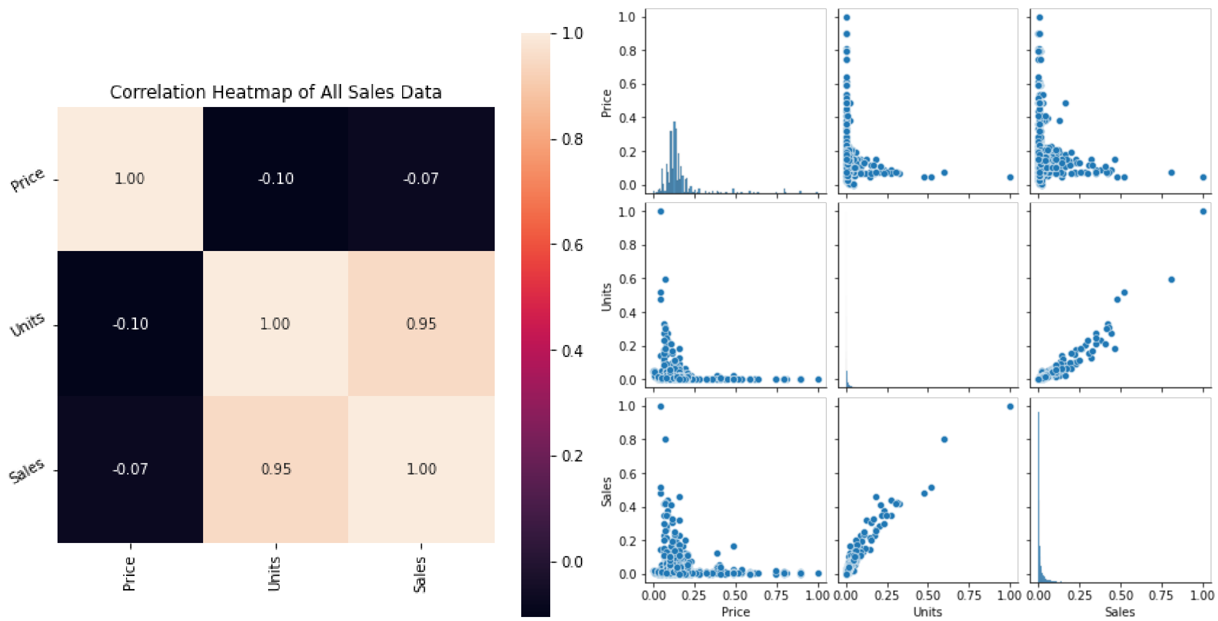

Figure 4. Correlation Heatmap and Pair Plots

A similar approach was completed for the data analysis for the combined 2021 sales and ordering data, which includes the same columns of product price, units, and sales, and is supplemented with cost per sale unit, cost percentage relative to purchase price, profit per unit, and total profit per product. Metrics were also computed for this data set, shown below in Table 3.

Metrics	**Price**	**Units**	**Sales**	**Unit Cost**	**TOTAL**	**COST**	**Quantity**	**Profit Per Unit**	**Cost %**	**Sale Unit Cost**
Maximum	22.36	2.80E+04	7.65E+04	190.00	9139.74	455.62	13.95	67378.16	4.06	35.00
Minimum	3.33	2.00	15.00	5.99	16.80	1.00	-19.90	-457.88	0.00	0.01
Mean	6.56	556.64	2671.53	66.03	291.73	7.49	3.84	2379.80	0.38	2.46
Median	6.00	75.00	485.00	72.00	138.00	1.72	3.84	305.51	0.32	1.94
Stdev	2.69	2206.64	7710.85	34.08	654.01	30.27	2.95	7267.43	0.41	3.21

Variance	7.23	4.87E+06	5.95E+07	1161.47	427735.50	916.44	8.71	5.28E+07	0.17	10.28
IQR	1.49	192.25	1227.14	49.84	171.02	3.72	2.23	887.85	0.39	2.50
Skewness	3.39	8.98	6.32	0.96	9.53	11.91	-3.00	6.12	4.30	5.72
Kurtosis	14.03	95.58	47.60	2.14	117.71	166.31	20.97	43.59	29.65	47.15

Table 3. Metrics of Combined Sales and Ordering Data 2019-2022

Like the summary sales data, this data set also exhibits high skewness and kurtosis except for ‘profit per unit’, which has a negative skew of about –3.00 indicating a left skewed dataset with, in this case, numerous negative outliers. ‘Profit’ and ‘profit per unit’ are also the only columns with negative values in their range. Visual boxplots of the data are shown below in Figure 5.

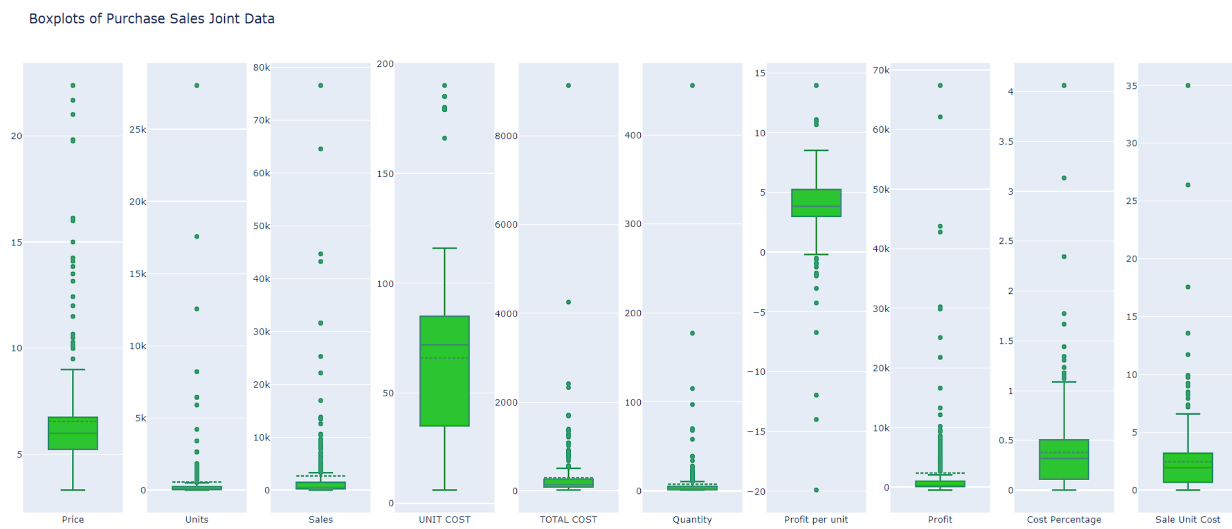

Figure 5. Boxplots of Combined 2021 Sales and Order Data

A heatmap was performed on the data, however since most of the data columns exhibit codependence, for example profit is the difference between cost and sales, the variables are multicollinear and there is no meaningful knowledge to be gained from the variables indicated correlation. Nonetheless, the heatmap is shown below in Figure 6.

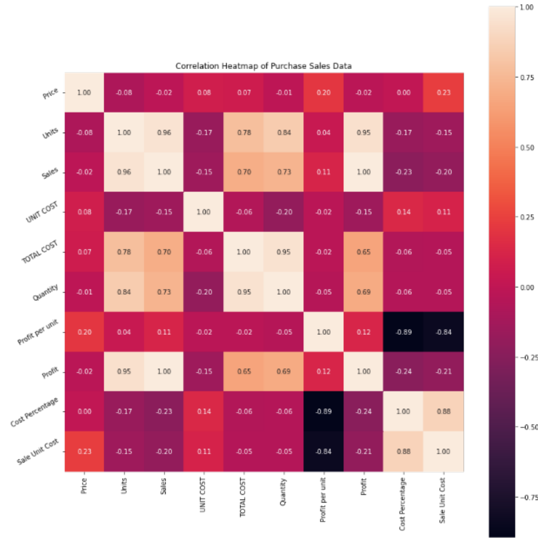

Figure 6. Correlation Heatmap

Pricing Elastics

To determine Velo's most profitable product pricing scheme, optimization was conducted through utilizing an alcohol price-demand elasticity model (Gruenewald et al. 96). The PED model generated by Grunewald et al. utilizes a set of coefficients that determine the effect of a segmental price increase on the demand of all other segments. To utilize optimization, an objective profitability function was formulated. This function is defined as the following:

$$max \sum_{n=1}^{I} p_{i,} * n_i$$

i represents the product in the set of all products I. p_i is the price of the product, and n_i is the sales volume of the product. n_i is related to p_i in the following fashion:

$$n_i = n_{i,original} * n_{increase}$$

$$n_{increase} = \sum_{m=1}^{C} b(c) * \sum_{l=1}^{I} \frac{p_i - p_{i,original}}{p_{i,original}}$$

$b(c)$ is the categorical price increase PED effect coefficients, c is the pricing category (ex. liquor, high price), $p_{i,original}$ is the original price, and $n_{i,original}$ is the original demand. This function is evaluated by aggregating products in each product category and can be evaluated on the basis of a per-product price increase.

This objective function was used to optimize markups in two ways, first through unconstrained optimization, and second by constrained optimization. In unconstrained optimization, price increase percentage was allowed to be any value. The unconstrained optimizer yielded unrealistic markup and markdowns for categories and therefore will not be further considered.

The constrained optimization limited the price increases to a range between 0 and 30%, the same as those reflected in the model's base paper. This optimization yielded a potential revenue increase of $629,000 a 31% increase, indicating that price increases to certain products are likely to increase profitability. This is a massive and significant increase, and likely needs to be additionally bound by factors considered outside those in the initial model paper.

Linear Regression

Linear regression was attempted to understand whether product profitability could be predicted from product characteristics. This regression was attempted through matrix optimization based linear regression and machine learning based linear regression through SKLearn's linear regression function. Both methods yielded similar results. The characteristics available of the products were generally bad predictors of the item's profitability. This is clearly illustrated in the error plots for each regression method:

Figure 7. Error plots for both regression methods.

These error plots indicated that linear models were unable to predict profitability especially for certain products that illustrate very high outlying error values within the chart.

Competitor Analysis

To assess Velo's beer pricing relative to its direct competitors, the Velo pricing dataset was filtered to only cans and bottles. This data was then directly compared to a dataset composed of roughly 40 competitive products. The comparison of the pricing distribution is visualized:

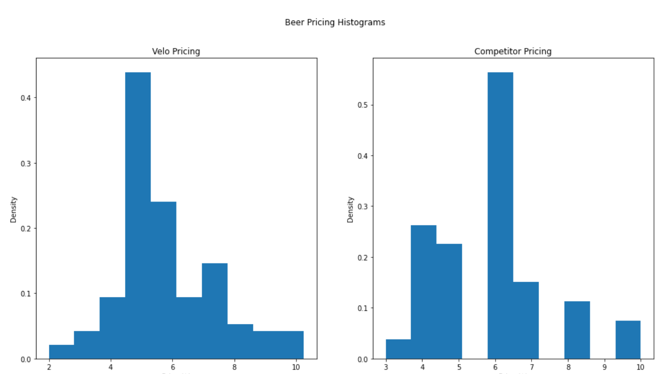

Figure 8. Histograms for Velo and Competitors beer pricing.

This analysis yielded a mean can beer price of $5.80 for Velo, with a mean price of $5.87 for its competitors. This indicated that Velo's pricing is very similar to its direct competition, and this is likely not a competitive concern for Velo.

Additionally, to reflect Velo's beer pricing more accurately, prices were weighted as a percentage of their products contribution to the unit sales. This weighting more accurately reflects the mean price based on what customers are paying for, rather than assuming that each product is sold evenly. After this weighting, the histogram shifted to be more heavily left skewed as displayed:

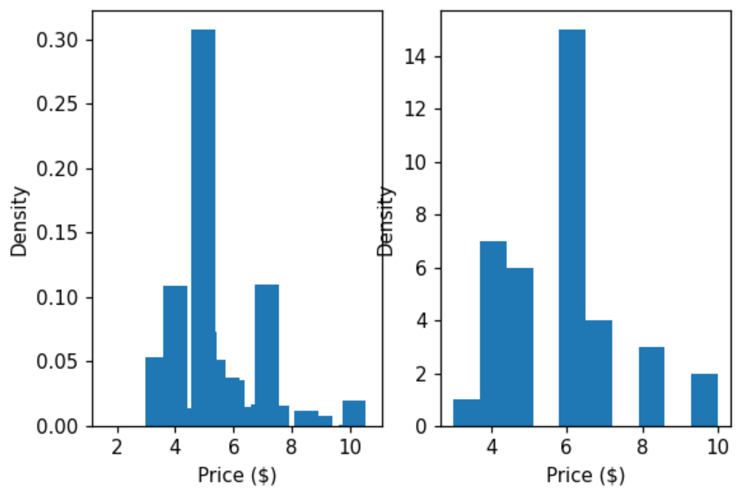

Figure 9. Sales weighted pricing histogram for Velo compared to competitor pricing.

This analysis yielded a mean beer price of $4.96, nearly $1 less than the unweighted prices. This indicates that Velo's beer pricing is very competitive, however, estimation of competitor sales weights for each product would make a direct comparison significantly more viable.

Product Mix Analysis

Finally, Velo's sale product mix was analyzed to understand trends in the sales of certain products relative to all sales. Since data was available from 2019 to 2022, the dataset was viable for such a comparison. The product mix was condensed to the top 10 selling products for each year with the rest of the products aggregated into an "Other" category. These aggregations were then used to generate a trend chart for the sales mix across the four years.

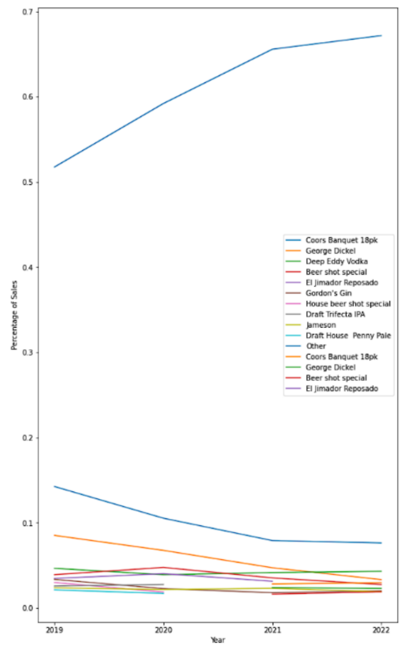

Figure 10. Product sales mix of Velo for 2019-2022.

This chart makes it apparent that Velo's customers are shifting away from some of the previously very high selling products and towards other products. Additionally, pie charts for each year (figure 5) reveal that the addition of a low-end beer quickly captured a significant portion of Velo's sales. This indicates that Velo should explore the opportunity to offer more down-market options that may sell at very high volumes. This can be further evaluated by visiting competitive bars and observing the percentage of sales that down-market products (mostly domestic beer) capture.

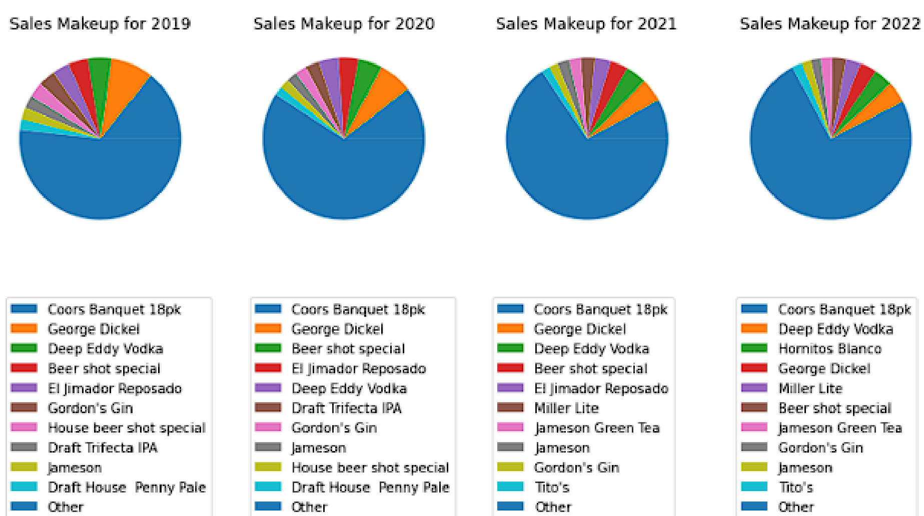

Figure 11. Sales product composition for 2019-2022.

Classification

To evaluate the predictive ability of models generated from the dataset, a classifier was also fit to the data. This KNN based classifier was fit to the cost percentage, profit, and units sold variables to predict the product type. Although this classifier is not necessarily useful in forming predictions, it is a useful indication of the relationships between the variables within the dataset. The resulting classifier was able to predict the product type with 62% accuracy. This low accuracy indicates a very loose relationship between the variables.

Upon further analysis, the classifier additionally was not predicting based on variable relationships but was instead simply guessing a single category for all variables. This preempts accuracy as being a determinant of the usefulness of the predictor and illustrates that categorical classification is not a useful exercise in better understanding the relationships within the dataset.

This process was kept in the report and code as indication of our ability to implement and interpret the results of a KNN classifier.

Recommendations/Conclusions

This section summarizes the recommendations and conclusions constructed from the analyses detailed within the approach section.

Analysis Based Conclusions

1. Exploratory data analysis evaluated the composition of SKUs and predicted that 2022 sales were on track to be similar to 2021 levels. Heatmapping indicated no obvious correlations between variables within the dataset outside of those forced by calculated columns.

2. Classification and regression further indicated the lack of relationships between variables within the dataset due to one-category prediction and high error values within regression.

3. Price Demand Elasticity optimization indicated that increasing the prices of certain products to increase demand in different levels of products is likely a viable strategy to increase overall profitability. We recommend that this option be explored further by Velo.

4. Competitive analysis indicated that Velo's canned and bottled beer pricing is in line of that of its competitors, and possible even lower if sales-based weighting is considered. We recommend that Velo attempt to evaluate its competitor's product sales mix to compare product pricing more directly. This exercise would also be valuable with other product lines, especially liquor, although this information is generally less easily available and may require more direct market research.

5. Product sales mix analysis indicates that Velo's product sales mix is clearly changing with time, and we recommend that they actively consider targeting products that more accurately reflect the desires of their customers.

Data Recommendations

1. Collection and maintenance of an ingredient list for cocktails would better allow analysis of pricing and costs of mixed drinks. Mixed drinks represent a significant portion of Velo's sales and incorporating their component products directly for analysis would

provide a more transparent sales overview.

2. Subcategorizing beer by the type (IPA, Light, Belgian, etc.) would allow for analysis of trends in beer sales and shifting style popularities. Deep analysis is heavily limited by the lack of categorization, and name-based categorization relies on intense manual work to yield meaningful results. This categorization only needs to take place in the ordering dataset to be usefully applied to the sales dataset as well.

3. Consistent order reporting would alleviate some of the assumptions required in the costing and profitability analysis. Better maintenance of this dataset would allow for trends in costs and availability to be more easily observed.

4. Evaluation of inventory holding costs would allow for optimization of products that take up valuable shelf real estate.

5. Reporting available at shorter time periods in an automated fashion would allow for the possibility of automatic ordering and more accurate ordering recommendation calculations.

6. Gathering additional information and suggestions from customers on how the restaurant or bar can be improved through customer satisfaction and engagement surveys.

References

"2nd Kitchen, Liquor Pricing." Web. 5/4/2022 <https://2ndkitchen.com/breweries/liquor-pricing/>.

"Cheers Online, About Cheers
." Web. 5/4/2022 <https://cheersonline.com/about/>.

"Greenville, South Carolina Population 2022 (Demographics, Maps, Graphs)." *World Population Review.* 2022. Web. 5/4/2022 <https://worldpopulationreview.com/us-cities/greenville-sc-population>.

Gruenewald, Paul J., et al. *Alcohol Prices, Beverage Quality, and the Demand for Alcohol: Quality Substitutions and Price Elasticities*. 30 Vol. Wiley, 2006a. Web.

---. *Alcohol Prices, Beverage Quality, and the Demand for Alcohol: Quality Substitutions and Price Elasticities*. 30 Vol. Wiley, 2006b. Web.

Jiang, Heng, et al. "Price Elasticity of on- and Off-Premises Demand for Alcoholic Drinks: A Tobit Analysis." *Drug And Alcohol Dependence* 163 (2016): 222-8. *MEDLINE.* Web.

Ornstein, S. I., and D. Levy. "Price and Income Elasticities of Demand for Alcoholic Beverages." *Recent developments in alcoholism : an official publication of the American Medical Society on Alcoholism, the Research Society on Alcoholism, and the National Council on Alcoholism* 1 (1983): 303-45. *MEDLINE.* Web.

Sharma, Anurag, Kompal Sinha, and Brian Vandenberg. *Pricing as a Means of Controlling Alcohol Consumption*. 123 Vol. Oxford University Press (OUP), 2017. Web.

"Society Homepage
." Web. 5/4/2022 <https://www.societygvl.com/>.

Vasel, Kathryn. "These are the fastest-growing cities in the U.S." *CNN Money.* 6/7/ 2017. Web. 5/4/2022 <https://money.cnn.com/gallery/real_estate/2017/06/02/fastest-growing-cities-census/4.html>.

"Vault and Vator Homepage
." Web. 5/4/2022 <https://www.vaultandvator.com/>.

"Your Ultimate Guide to Greenville Breweries." Web. 5/4/2022 <https://www.visitgreenvillesc.com/food-drink/stories-guides/breweries/>.