BIG DATA - BUAN - 6346

Semester Project

Group Members:

Mourlaye Traore (*mxt230013*)

Lal Reddy Indla(lxi220003)

Kiran Kumar Reddy Kanchani

Project Instructions Highlights

Per our instructions, the goal of this project is to identify actionable business insights through the formulation of a minimum of three key questions based on the dataset we found. These questions will serve as the guiding framework for our analysis, enabling us to derive meaningful conclusions. As directed, the datasets will be loaded into Hadoop using either Flume or Sqoop, and Hive will be used for query execution. While the use of Hadoop or Spark is permitted, the use of Jupiter is explicitly prohibited. Furthermore, visualization tools like Tableau can be utilized to enhance the presentation of our findings. It is specified that the combined size of the datasets should ideally be 1GB or more, with a minimum threshold of 200MB to ensure adequate data for analysis. Through adherence to these instructions, we aim to conduct a thorough examination of the data, uncovering valuable insights and offering informed recommendations to drive business growth and efficiency.

About Dataset

Context

Our dataset "***Liquor_sales***" for this project is from Kaggle, and comprises a comprehensive record of spirits purchases made by Iowa Class "E" liquor licenses from January 2021 to January 2022. It offers detailed insights into the procurement patterns at the store level, with each entry capturing the date of purchase, product specifics (such as type, brand, and alcohol content), and the purchasing store's information. Additionally, the dataset includes financial data, detailing the cost to the retailer alongside the quantity purchased. This rich dataset is ideal for analyzing trends in spirits sales across different regions and periods, enabling stakeholders to understand consumer preferences and seasonal demand fluctuations in the Iowa liquor market. Such analysis can aid retailers in optimizing their stock levels and developing targeted marketing strategies, while policymakers can utilize this information to better understand the economic aspects of liquor sales in the state.

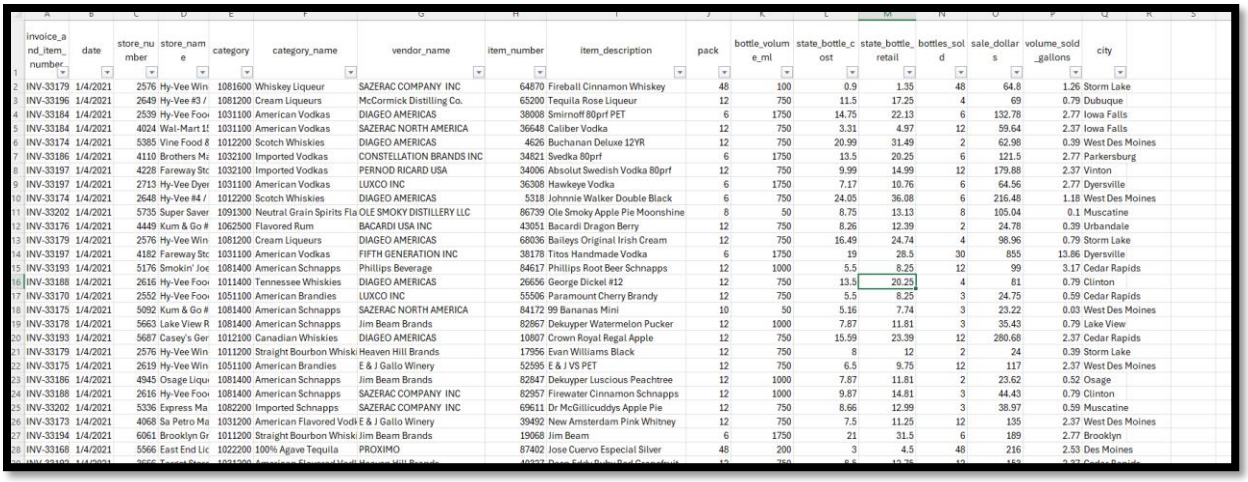
Data Description

This comprehensive dataset consists of approximately 3 million rows and 24 columns, providing an extensive overview of liquor orders made by Iowa Class "E" liquor licenses from January 2021 to January 2022. Each record is uniquely identified by a concatenated **invoice_and_item_number**, which links directly to the specific liquor product ordered. The dataset captures detailed transactional information, including the **date** of order, **store_number**, **store_name**, and precise **store_location** with geographic coordinates derived from the store's **address**, **city**, **zip_code**, and **county**. It also details the product specifics such as **category_name**, **item_description**, **vendor_name**, and packaging details (**pack**, **bottle_volume_ml**). Financial aspects are thoroughly documented, showing **state_bottle_cost**, **state_bottle_retail**, **bottles_sold**, **sale_dollars**, as well as the total volume of liquor ordered in

both liters and gallons. This rich dataset is pivotal for detailed analytics on liquor sales trends, store performance, and consumer behavior across different geographic locations in Iowa.
Columns:

1. **invoice_and_item_number**: Concatenated invoice and line number providing a unique identifier for individual liquor products within an order.
2. **date**: Date of the order.
3. **store_number**: Unique identifier assigned to the store placing the liquor order.
4. **store_name**: Name of the store placing the liquor order.
5. **address**: Address of the store placing the liquor order.
6. **city**: City where the store placing the liquor order is located.
7. **zip_code**: Zip code of the store placing the liquor order.
8. **store_location**: Geographic coordinates of the store placing the liquor order, derived from address, city, state, and zip code.
9. **county_number**: Iowa county number for the county where the store placing the liquor order is located.
10. **county**: County where the store placing the liquor order is located.
11. **category**: Category code associated with the liquor ordered.
12. **category_name**: Category name of the liquor ordered.
13. **vendor_number**: Vendor number of the company for the brand of liquor ordered.
14. **vendor_name**: Vendor name of the company for the brand of liquor ordered.
15. **item_number**: Item number for the individual liquor product ordered.
16. **item_description**: Description of the individual liquor product ordered.
17. **pack**: Number of bottles in a case for the liquor ordered.
18. **bottle_volume_ml**: Volume of each liquor bottle ordered in milliliters.
19. **state_bottle_cost**: Cost that the alcoholic beverages division paid for each bottle of liquor ordered.
20. **state_bottle_retail**: Price that the store paid for each bottle of liquor ordered.
21. **bottles_sold**: Number of bottles of liquor ordered by the store.
22. **sale_dollars**: Total cost of the liquor order (number of bottles multiplied by the state bottle retail).
23. **volume_sold_liters**: Total volume of liquor ordered in liters (calculated as bottle volume in ml multiplied by bottles sold and divided by 1000).
24. **volume_sold_gallons**: Total volume of liquor ordered in gallons (calculated as bottle volume in ml multiplied by bottles sold and divided by 3785.411784).

Here is a simple of the dataset:



invoice_and_item_number	date	store_number	store_name	category	category_name	vendor_name	item_number	item_description	pack	bottle_volume_ml	state_bottle_cost	state_bottle_retail	bottles_sold	sale_dollars	volume_sold_gallons	city
INV-33179	1/4/2021	2576	Hy-Vee Win	1081600	Whiskey Liqueur	SAZERAC COMPANY INC	64870	Fireball Cinnamon Whiskey	48	100	0.9	1.35	48	64.8	1.26	Storm Lake
INV-33196	1/4/2021	2649	Hy-Vee #3 /	1081200	Cream Liqueurs	McCormick Distilling Co.	65200	Tequila Rose Liqueur	12	750	11.5	17.25	4	69	0.79	Dubuque
INV-33184	1/4/2021	2539	Hy-Vee Foo	1031100	American Vodkas	DIAGEO AMERICAS	38008	Smirnoff 80prf PET	6	1750	14.75	22.13	6	132.78	2.77	Iowa Falls
INV-33184	1/4/2021	4024	Wal-Mart 1	1031100	American Vodkas	SAZERAC NORTH AMERICA	36648	Caliber Vodka	12	750	3.31	4.97	12	59.64	2.37	Iowa Falls
INV-33174	1/4/2021	5385	Vine Food &	1012200	Scotch Whiskies	DIAGEO AMERICAS	4626	Buchanan Deluxe 12YR	12	750	20.99	31.49	2	62.98	0.39	West Des Moines
INV-33186	1/4/2021	4110	Brothers Ma	1032100	Imported Vodkas	CONSTELLATION BRANDS INC	34821	Svedka 80prf	6	1750	13.5	20.25	6	121.5	2.77	Parkersburg
INV-33197	1/4/2021	4228	Fareway Stc	1032100	Imported Vodkas	PERNOD RICARD USA	34006	Absolut Swedish Vodka 80prf	12	750	9.99	14.99	12	179.88	2.37	Vinton
INV-33197	1/4/2021	2713	Hy-Vee Dyer	1031100	American Vodkas	LUXCO INC	36308	Hawkeye Vodka	6	1750	7.17	10.76	6	64.56	2.77	Dyersville
INV-33174	1/4/2021	2648	Hy-Vee #4 /	1012200	Scotch Whiskies	DIAGEO AMERICAS	5318	Johnnie Walker Double Black	6	750	24.05	36.08	6	216.48	1.18	West Des Moines
INV-33202	1/4/2021	5735	Super Saver	1091300	Neutral Grain Spirits Fla	OLE SMOKY DISTILLERY LLC	86739	Ole Smoky Apple Pie Moonshine	8	50	8.75	13.13	8	105.04	0.1	Muscatine
INV-33176	1/4/2021	4449	Kum & Go #	1062500	Flavored Rum	BACARDI USA INC	43051	Bacardi Dragon Berry	12	750	8.26	12.39	2	24.78	0.39	Urbandale
INV-33179	1/4/2021	2576	Hy-Vee Win	1081200	Cream Liqueurs	DIAGEO AMERICAS	68036	Baileys Original Irish Cream	12	750	16.49	24.74	4	98.96	0.79	Storm Lake
INV-33197	1/4/2021	4182	Fareway Stc	1031100	American Vodkas	FIFTH GENERATION INC	38178	Titos Handmade Vodka	6	1750	19	28.5	30	855	13.86	Dyersville
INV-33193	1/4/2021	5176	Smokin' Joe	1081400	American Schnapps	Phillips Beverage	84617	Phillips Root Beer Schnapps	12	1000	5.5	8.25	12	99	3.17	Cedar Rapids
INV-33188	1/4/2021	2616	Hy-Vee Foo	1011400	Tennessee Whiskies	DIAGEO AMERICAS	26656	George Dickel #12	12	750	13.5	20.25	4	81	0.79	Clinton
INV-33170	1/4/2021	2552	Hy-Vee Foo	1051100	American Brandies	LUXCO INC	55506	Paramount Cherry Brandy	12	750	5.5	8.25	3	24.75	0.59	Cedar Rapids
INV-33175	1/4/2021	5092	Kum & Go #	1081400	American Schnapps	SAZERAC NORTH AMERICA	84172	99 Bananas Mini	10	50	5.16	7.74	3	23.22	0.03	West Des Moines
INV-33178	1/4/2021	5663	Lake View R	1081400	American Schnapps	Jim Beam Brands	82867	Dekuyper Watermelon Pucker	12	1000	7.87	11.81	3	35.43	0.79	Lake View
INV-33193	1/4/2021	5687	Casey's Ger	1012100	Canadian Whiskies	DIAGEO AMERICAS	10807	Crown Royal Regal Apple	12	750	15.59	23.39	12	280.68	2.37	Cedar Rapids
INV-33179	1/4/2021	2576	Hy-Vee Win	1011200	Straight Bourbon Whiski	Heaven Hill Brands	17956	Evan Williams Black	12	750	8	12	2	24	0.39	Storm Lake
INV-33175	1/4/2021	2619	Hy-Vee Win	1051100	American Brandies	E & J Gallo Winery	52595	E & J VS PET	12	750	6.5	9.75	12	117	2.37	West Des Moines
INV-33186	1/4/2021	4945	Osage Liqu	1081400	American Schnapps	Jim Beam Brands	82847	Dekuyper Luscious Peachtree	12	1000	7.87	11.81	2	23.62	0.52	Osage
INV-33188	1/4/2021	2616	Hy-Vee Foo	1081400	American Schnapps	SAZERAC COMPANY INC	82957	Firewater Cinnamon Schnapps	12	1000	9.87	14.81	3	44.43	0.79	Clinton
INV-33202	1/4/2021	5336	Express Ma	1082200	Imported Schnapps	SAZERAC COMPANY INC	69611	Dr McGillicuddys Apple Pie	12	750	8.66	12.99	3	38.97	0.59	Muscatine
INV-33173	1/4/2021	4068	Sa Petro Ma	1031200	American Flavored Vodk	E & J Gallo Winery	39492	New Amsterdam Pink Whitney	12	750	7.5	11.25	12	135	2.37	West Des Moines
INV-33194	1/4/2021	6061	Brooklyn Gr	1011200	Straight Bourbon Whiski	Jim Beam Brands	19068	Jim Beam	6	1750	21	31.5	6	189	2.77	Brooklyn
INV-33168	1/4/2021	5566	East End Lic	1022200	100% Agave Tequila	PROXIMO	87402	Jose Cuervo Especial Silver	48	200	3	4.5	48	216	2.53	Des Moines

```
>>> liquor_sales_df.printSchema()
root
 |-- invoice_and_item_number: string (nullable = true)
 |-- date: string (nullable = true)
 |-- store_number: integer (nullable = true)
 |-- store_name: string (nullable = true)
 |-- category: double (nullable = true)
 |-- category_name: string (nullable = true)
 |-- vendor_name: string (nullable = true)
 |-- item_number: integer (nullable = true)
 |-- item_description: string (nullable = true)
 |-- pack: integer (nullable = true)
 |-- bottle_volume_ml: integer (nullable = true)
 |-- state_bottle_cost: double (nullable = true)
 |-- state_bottle_retail: double (nullable = true)
 |-- bottles_sold: integer (nullable = true)
 |-- sale_dollars: double (nullable = true)
 |-- volume_sold_gallons: double (nullable = true)
 |-- city: string (nullable = true)

>>>
```

```
>>> summary_stats = df_clean.describe(['state_bottle_retail', 'bottles_sold', 'sale_dollars'])
>>> summary_stats.show()
+-------+-------------------+------------------+------------------+
|summary|state_bottle_retail|      bottles_sold|      sale_dollars|
+-------+-------------------+------------------+------------------+
|  count|            2805307|           2805307|           2805307|
|   mean|  16.98823169442807| 11.85836915531883|162.44968686135388|
| stddev| 16.546276842554224|35.668168779373794| 587.1846395910108|
|    min|               0.99|                 1|              1.34|
|    max|              99.75|               990|            999.54|
+-------+-------------------+------------------+------------------+
```

Business Questions

Questions 1 *(Mourlaye)*

How store order volumes and liquor product sales volumes correlate with total dollar sales at both the store and product levels?

To answer this question, we will walk through a series of questions in Hadoop + pyspark. Then for visualization, we can use some of the tools Tableau allowed for this project.
After loading our dataset in our VM, we are ready to run queries and provide some answers.
To analyze these correlations, we can:

- Compute the total number of orders (invoices) per store.
- Calculate the total volume sold and revenue for each store.
- Assess the correlation between the number of orders a store places and its total revenue.
- Investigate if there's a relationship between the volume sold of individual products and the revenue they generate.

Due to the volume of the dataset, we will focus only on the top 20 stores based on revenue to answer our question. After running the appropriate queries in the pyspark session in Hadoop, here is our first results.

store_name	formatted_total_sales_dollars	formatted_total_volume_sold
Hy-Vee #3 / BDI / Des Moines	13,266,416.86	187,461.240
Central City 2	12,901,621.86	181,086.020
Hy-Vee Wine and Spirits / Iowa City	5,821,589.25	97,310.230
Costco Wholesale #788 / WDM	5,000,957.79	89,863.030
Benz Distributing	4,592,046.20	63,074.290
Wilkie Liquors	4,278,719.74	71,816.990
Sam's Club 8162 / Cedar Rapids	3,887,888.06	66,024.810
I-80 Liquor / Council Bluffs	3,680,217.43	48,315.680
Sam's Club 6344 / Windsor Heights	3,546,451.17	60,694.570
Lot-A-Spirits	3,402,977.63	48,028.730
Hy-Vee Food Store / Urbandale	3,298,320.24	39,712.200
Sam's Club 6979 / Ankeny	3,233,622.07	56,069.310
Another Round / DeWitt	3,076,358.85	43,279.730
Hy-Vee Food Store / Coralville	2,885,988.44	41,627.080
Hy-Vee / Waukee	2,831,663.98	37,254.900
Central City Liquor, Inc.	2,799,272.61	26,774.170
Sam's Club 6514 / Waterloo	2,733,032.37	52,845.640
Costco Wholesale #1111 / Coralville	2,711,095.10	50,398.880
Happy's Wine & Spirits	2,707,078.75	37,635.400
Hy-Vee Wine and Spirits / WDM	2,631,598.90	34,964.360

This result provides a clear perspective on the sales performance of various stores. Hy-Vee #3 / BDI in Des Moines emerges as the leading store with total sales amounting to over $13 million and an impressive volume of approximately 187,461 gallons sold. This indicates not only a high turnover but also a significant market share in terms of liquor sales in its location.

The data reveals a variety of performances across different stores, with Hy-Vee Wine and Spirits / Iowa City and Costco Wholesale #788 / WDM also demonstrating strong sales, each surpassing $5 million in revenue. The sales figures show a high volume of liquor sold, with the top stores selling tens of thousands of gallons, reflecting their success in meeting consumer demand.

Understanding these dynamics can help the business to make informed decisions on inventory management, distribution, and promotional efforts to maximize sales and profitability across all stores. In addition to the dollar and volume of each of these top stores, we also want to check the ratio of dollar to volume sold for each. We then modified our query to include the dollar to volume ratio. Here are the results.

```
... )
>>> top_stores.show(20, False)
+----------------------------------+---------------------------+-------------------------+----------------------------
|store_name                        |formatted_total_sales_dollars|formatted_total_volume_sold|formatted_dollar_to_volume_ra
+----------------------------------+---------------------------+-------------------------+----------------------------
|Hy-Vee #3 / BDI / Des Moines      |13,266,416.86              |187,461.240              |70.7689
|Central City 2                    |12,901,621.86              |181,086.020              |71.2458
|Hy-Vee Wine and Spirits / Iowa City|5,821,589.25               |97,310.230               |59.8250
|Costco Wholesale #788 / WDM       |5,000,957.79               |89,863.030               |55.6509
|Benz Distributing                 |4,592,046.20               |63,074.290               |72.8038
|Wilkie Liquors                    |4,278,719.74               |71,816.990               |59.5781
|Sam's Club 8162 / Cedar Rapids    |3,887,888.06               |66,024.810               |58.8853
|I-80 Liquor / Council Bluffs      |3,680,217.43               |48,315.680               |76.1703
|Sam's Club 6344 / Windsor Heights |3,546,451.17               |60,694.570               |58.4311
|Lot-A-Spirits                     |3,402,977.63               |48,028.730               |70.8530
|Hy-Vee Food Store / Urbandale     |3,298,320.24               |39,712.200               |83.0556
|Sam's Club 6979 / Ankeny          |3,233,622.07               |56,069.310               |57.6719
|Another Round / DeWitt            |3,076,358.85               |43,279.730               |71.0808
|Hy-Vee Food Store / Coralville    |2,885,988.44               |41,627.080               |69.3296
|Hy-Vee / Waukee                   |2,831,663.98               |37,254.900               |76.0078
|Central City Liquor, Inc.         |2,799,272.61               |26,774.170               |104.5512
|Sam's Club 6514 / Waterloo        |2,733,032.37               |52,845.640               |51.7173
|Costco Wholesale #1111 / Coralville|2,711,095.10               |50,398.880               |53.7928
|Happy's Wine & Spirits            |2,707,078.75               |37,635.400               |71.9291
|Hy-Vee Wine and Spirits / WDM     |2,631,598.90               |34,964.360               |75.2652
+----------------------------------+---------------------------+-------------------------+----------------------------

>>>
```

An overview of the ratio of dollars to volume gives us a better way of measuring the performance and effectiveness of each store. The results show that even though ***Hy-Vee #3 / BDI in Des Moines*** leads the pack with impressive total sales exceeding $13 million and a substantial volume of liquor sold at over 187,000 gallons, it is not necessary the store with the highest ratio. Interestingly, stores like ***Central City Liquor, Inc***. have a notably higher dollar to volume ratio of **104.55,** with only $2.8M revenue hinting at either a premium product mix or a more effective pricing strategy. On the other end of the spectrum, stores like ***Costco Wholsesale #788 / WDM*** despite a strong sales performance $5M revenue, show a ratio of $55.6 $/g, which might suggest a focus on volume sales with potentially lower margins per unit.

This analysis serves as a foundation for further investigation into each store's product offerings, customer demographics, and market positioning to refine their sales and pricing strategies. The goal is to strike an optimal balance that maximizes revenue while sustaining or increasing the volume of sales. To investigate the correlation between sales and revenue, it’s also useful to look at the statistics for the revenue per product. For this part, we will run the query that displays the top 20 products with the highest revenue. Along with the dollar sales, we are also interested in the total volume of sales and the ratio of dollars to volume. Below are the results.

```
>>> top_categories.show(20, False)
+------------------------------+---------------------------+-------------------------+-------------------------------+
|category_name                 |formatted_total_sales_dollars|formatted_total_volume_sold|formatted_dollar_to_volume_ratio|
+------------------------------+---------------------------+-------------------------+-------------------------------+
|American Vodkas               |65,582,684.00              |1,603,247.150            |40.9062                        |
|Canadian Whiskies             |50,322,883.27              |846,501.210              |59.4481                        |
|Straight Bourbon Whiskies     |36,948,852.92              |377,571.990              |97.8591                        |
|Whiskey Liqueur               |26,664,116.55              |327,699.950              |81.3675                        |
|100% Agave Tequila            |25,359,203.78              |171,832.120              |147.5813                       |
|Spiced Rum                    |24,549,501.08              |415,171.880              |59.1309                        |
|Tennessee Whiskies            |17,219,118.57              |146,473.460              |117.5579                       |
|Imported Brandies             |16,646,780.08              |90,520.270               |183.9011                       |
|Imported Vodkas               |15,239,358.28              |224,434.090              |67.9013                        |
|Blended Whiskies              |13,477,222.83              |238,734.640              |56.4527                        |
|Mixto Tequila                 |11,258,758.49              |178,015.010              |63.2461                        |
|American Flavored Vodka       |11,213,426.84              |199,830.150              |56.1148                        |
|Imported Cordials & Liqueurs  |10,903,561.22              |93,719.370               |116.3427                       |
|Irish Whiskies                |9,880,194.07               |84,860.790               |116.4283                       |
|Cream Liqueurs                |9,545,889.06               |106,311.010              |89.7921                        |
|Flavored Rum                  |9,514,014.85               |161,823.640              |58.7925                        |
|Cocktails /RTD                |9,079,479.92               |302,664.870              |29.9985                        |
|Temporary & Specialty Packages|8,631,775.17               |75,375.250               |114.5174                       |
|Imported Schnapps             |7,113,738.03               |98,804.280               |71.9983                        |
|White Rum                     |7,097,933.79               |170,746.380              |41.5700                        |
+------------------------------+---------------------------+-------------------------+-------------------------------+
```

The results show ***American Vodkas*** as the dominant category in both sales revenue and volume. With over $65 million in sales and around 1.6 million gallons sold. However, it seems to have a lower dollar to volume ratio, $41/g compared to most of the categories.

Comparatively**, *Imported Brandies***, despite a lower revenue, ***16M***, boast the highest dollar-to-volume ratio at ***$183.90/g***. This is in stark contrast to American Vodkas, which, while generating substantial total revenue, **65M**, do so at a much lower price point per unit, **$41/g.**

Considering categories like **100% Agave Tequila** with a moderate ratio of **$147.38/g**, it is apparent that this category has found a sweet spot between the volume sold and the revenue per unit, which could reflect a strong market position with potential growth opportunities.

For business strategy, leveraging the popularity and volume sales of **American Vodkas** is key, but also capitalizing on the high-profit margins of categories like Imported Brandies can enhance overall profitability. Balancing the product mix to cater to both high-volume sales and high-margin categories can help diversify revenue streams and stabilize market position.

Conclusion:

The analysis reveals a clear correlation between the volume of sales and the total revenue across stores, with top performers like **Hy-Vee #3 / BDI** displaying significant volumes and revenues. A more volume of sale technically implies a more revenue with a good or average ratio. Categories like ***American Vodkas*** lead in both volume and revenue, indicating regular ordering contributes to higher revenue, whereas high-value categories like ***Imported Brandies*** command significant revenue with less volume. This suggests a nuanced approach to inventory and ordering can optimize revenue across stores.

Business recommendations:

Leverage High-Performing Stores for Market Insights:

We have seen that **Hy-Vee #3 / BDI** shows the highest total sales, which suggests that they may have a better approach that helps them maximize their sales, ranging from product selection to customer service, is highly effective. However, since the dollar to volume ratio seems to be lower than the average, the store should think about merchandising strategies that could help them increase or leverage that ratio without necessarily affecting their volume sales.

Strategic Inventory Management:
Given the varying dollar-to-volume ratios, stores could optimize inventory levels based on sales performance and profitability. For instance, ***Hy-Vee Food Store / Urbandale*** and ***Central City Liquor, Inc.***, which show relatively high dollar-to-volume ratios, suggest that customers there may prefer premium products. The business could increase the stock of higher-end liquors in these stores.

Expand Premium Selections:
With **Imported Brandies** having the highest dollar-to-volume ratio, there's an indication that consumers are willing to pay more for premium products. The business should consider expanding its selection of premium brands and specialty products in other high-margin categories, such as **Whiskey Liqueur** and **100% Agave Tequila**, to capitalize on this trend.

Strategic Pricing and Promotions for Volume Drivers:
American Vodkas drive both volume and revenue, signaling a significant market share. The business could develop strategic promotions for this category to increase customer loyalty and draw in new customers.

Question 2 *(Kiran Kumar Reddy)*

"How does pricing affect the volume and profitability of liquor sales, and what are the optimal price points for various categories to enhance revenue?"

Introduction
This document aims to analyze the relationship between pricing strategies and their effects on the sales volume and profitability of liquor. Understanding this correlation will aid in identifying optimal price points across various liquor categories to maximize revenue.

Objective:
To determine how pricing adjustments can influence the sales volume and profitability of different liquor categories and establish the optimal pricing points that enhance revenue generation.

Table 1: Sample Data Overview
Placeholder for Table 1: A table summarizing the initial data characteristics, including sample sizes and missing values.

```
>>> df_categories = spark.read.csv("file:///home/kiran/Downloads/selected_liquor_columns.csv", header=True, inferSchema=True)
>>> df_categories
DataFrame[category: double, state_bottle_retail: double, bottles_sold: int, sale_dollars: double, item_description: string, category_name:
 string]
>>> df_categories.show()
+--------+------------------+------------+------------+--------------------+--------------------+
|category|state_bottle_retail|bottles_sold|sale_dollars|    item_description|       category_name|
+--------+------------------+------------+------------+--------------------+--------------------+
|1011400.0|             11.25|           3|       33.75|Jack Daniels Old ...|  Tennessee Whiskies|
|1081600.0|              1.35|          48|        64.8|Fireball Cinnamon...|     Whiskey Liqueur|
|1051100.0|              19.5|           1|        19.5|           E & J VS|   American Brandies|
|1041100.0|             10.38|           6|       62.28|         Caliber Gin|   American Dry Gins|
|1011200.0|             13.49|           3|       40.47| Makers Mark Replica|Straight Bourbon ...|
|1012100.0|             15.68|           6|       94.08|  Black Velvet Apple|   Canadian Whiskies|
|1012100.0|             14.25|           6|        85.5|Windsor Canadian PET|   Canadian Whiskies|
|1012100.0|             21.75|           1|       21.75|Canadian Club Whisky|   Canadian Whiskies|
|1012100.0|             11.03|           1|       11.03|    Crown Royal Mini|   Canadian Whiskies|
|1012100.0|              7.85|          12|        94.2|Black Velvet Toas...|   Canadian Whiskies|
|1081600.0|               8.0|          24|       192.0|Fireball Cinnamon...|     Whiskey Liqueur|
|1081600.0|               8.0|          24|       192.0|Fireball Cinnamon...|     Whiskey Liqueur|
|1042100.0|             10.49|           3|       31.47|       Tanqueray Gin|  Imported Dry Gins|
|1022200.0|              21.0|           2|        42.0|  Cazadores Reposado|100% Agave Tequila|
|1011400.0|             13.59|           3|       40.77|Jack Daniels Tenn...|  Tennessee Whiskies|
|1011100.0|             15.75|           1|       15.75|Red Stag Black Ch...|    Blended Whiskies|
|1011100.0|              3.14|          24|       75.36|           Five Star|    Blended Whiskies|
|1081400.0|              7.11|           1|        7.11|Arrow Mcdales But...|   American Schnapps|
|1012100.0|              2.34|          48|      112.32|        Black Velvet|   Canadian Whiskies|
|1012100.0|             49.49|          12|      593.88|Crown Royal Regal...|   Canadian Whiskies|
+--------+------------------+------------+------------+--------------------+--------------------+
only showing top 20 rows

>>>
```

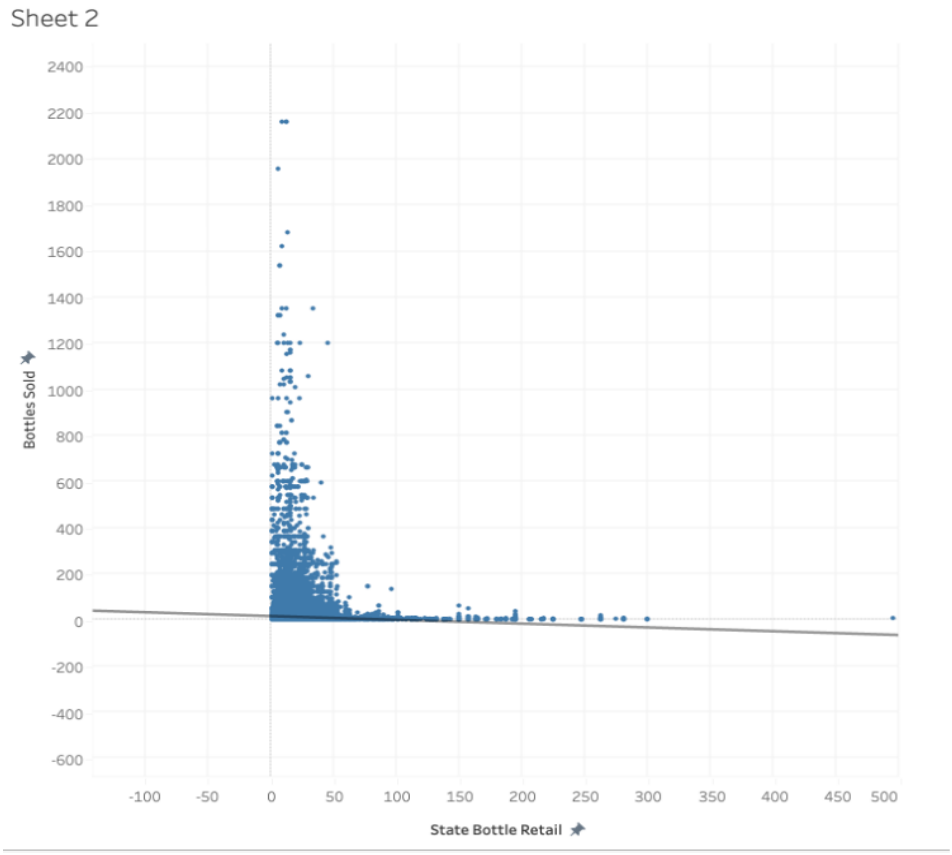
Regression Analysis

The regression analysis to quantitatively predict the effects of different pricing on sales and profit margins. This method will allow for a more nuanced understanding of the causal relationships within the data.

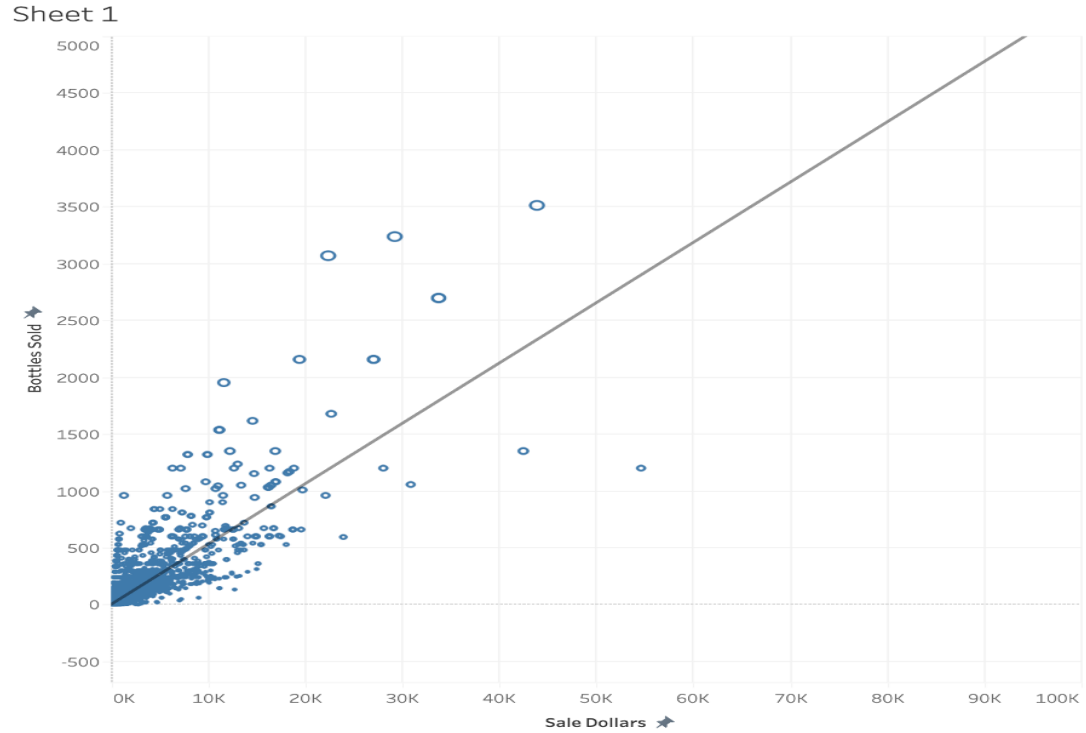
Table 2: Regression Model Results

```
>>> correlations.show()
+--------------------+--------------------+
|   price_volume_corr|  price_profit_corr|
+--------------------+--------------------+
|-0.06570430284499904|0.09645475239875996|
+--------------------+--------------------+
```

State Bottle Retail vs Bottles Sold



Sale Dollar's vs Bottles Sold



Output Explanation: Table 2 presents the results of the regression model. These results help in understanding how different pricing strategies affect sales volume and profitability, thereby illustrating the causal relationships within the data.

Price Elasticity Calculation: Calculate the price elasticity of demand for each category by analyzing the percentage change in sales volume relative to the percentage change in pricing.

Table 3: Price Elasticity Results

```
>>> sampled_df_no_duplicates.show(5)
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+---
----------------+-----------------+---------+--------------------+
|state_bottle_retail|total_bottles_sold|       total_sales|prev_bottles_sold|    ΔQ|             ΔQ/Q|prev_price|                  ΔP|
           ΔP/P|  Price_Elasticity| category|       category_name|
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+---
----------------+-----------------+---------+--------------------+
|              7.01|           48790| 342017.8999999994|           20336| 28454| 58.31932773109244|      6.99|0.019999999999999574|  0
.2853067047075546|204.40924369748333|1701100.0|Temporary & Speci...|
|             46.02|           41313|1901224.2599999884|             935| 40378| 97.73678987243724|     45.96|0.060000000000002274| 0.
1303780964797963 4| 749.6411783215652|1011400.0|  Tennessee Whiskies|
|             47.01|              89|          4183.89|              40|    49|  55.0561797752809|      47.0| 0.009999999999999801|0.0
21272069772384623|  2588.19101123647|1901200.0| Special Order Items|
|             10.13|            7532| 76299.16000000009|            4441|  3091| 41.03823685608072|     10.11| 0.020000000000000135| 0.
19743336623890767|207.85866967603485|1062400.0|          Spiced Rum|
|             13.53|           19553|         261770.87|           30827|-11274|-57.65867130363627|     13.52|0.009999999999999787|  0
.0739098300073894|-780.1218227382153|1011300.0|Single Barrel Bou...|
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+---
----------------+-----------------+---------+--------------------+
only showing top 5 rows

>>>
```

Placeholder for Table 3: A table displaying calculated price elasticity values for different liquor categories.

Price Elasticity Calculation: Calculate the price elasticity of demand for each category by analyzing the percentage change in sales volume relative to the percentage change in pricing.

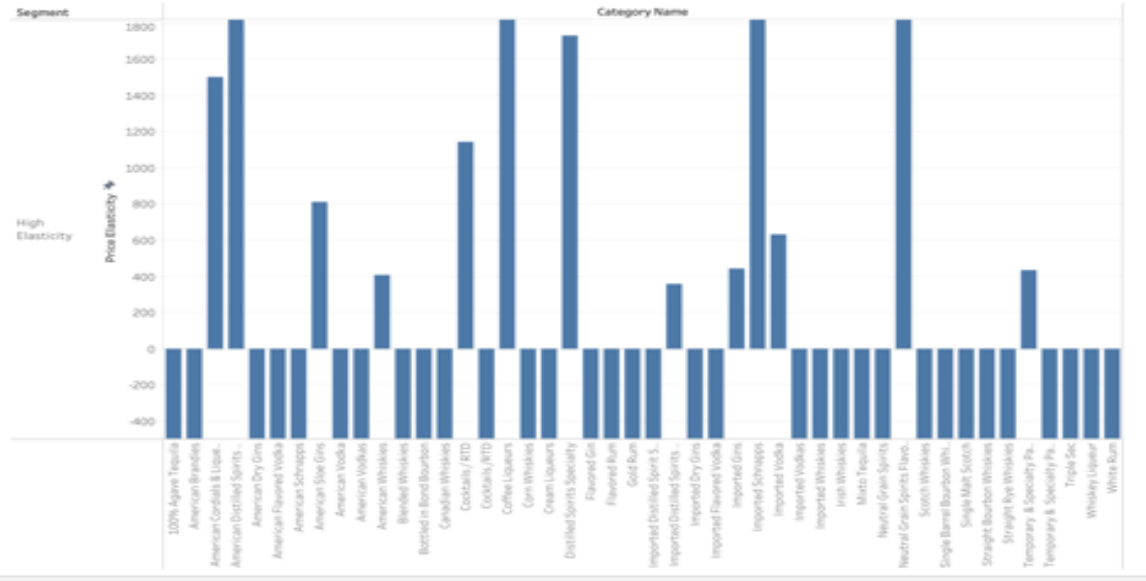
Table 4: Price Elasticity Results

Placeholder for Table 4: A table displaying calculated price elasticity values for different liquor categories.

```
>>> top_rows = segmented_df.orderBy("segment").show(5)
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+----
--------------+-----------------+--------------+
|state_bottle_retail|total_bottles_sold|       total_sales|prev_bottles_sold|    ΔQ|             ΔQ/Q|prev_price|                  ΔP|
           ΔP/P|  Price_Elasticity|       segment|
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+----
--------------+-----------------+--------------+
|               2.7|          505920|1365983.9999999672|            7244|498676| 98.5681530676787|      2.69|0.010000000000000231| 0.3
703703703703789|266.1340132827263 6|High Elasticity|
|              5.01|          263142|1318341.4199999475|           18569|244573|92.94335377856822|       5.0|0.009999999999999787|0.19
960079840318934| 465.6462024306368|High Elasticity|
|               3.2|           31442| 100614.3999999999|             485| 30957|98.45747725971631|      3.17| 0.03000000000000025| 0.9
375000000000078|105.02130907702987|High Elasticity|
|              4.83|          114074| 550977.4199999977|            2385|111689| 97.9092518891246|      4.82|0.009999999999999787|0.20
703933747411568| 472.9016866244819|High Elasticity|
|              4.97|          116068| 576857.9599999968|           30920| 85148|73.36044387772685|      4.95|0.019999999999999574| 0.4
024144869215206| 182.3007030361551|High Elasticity|
+------------------+----------------+-----------------+----------------+------+-----------------+----------+--------------------+----
--------------+-----------------+--------------+
only showing top 5 rows

>>>
```

Output Explanation:



Findings

Initial correlations indicate a weak negative relationship between price increases and volume sold, suggesting that higher prices slightly reduce the number of units sold.

A weak positive correlation exists between price increases and profitability, implying that revenue might increase with price hikes, though the effect is not strong.

Strategic Recommendations

Price Adjustment: Based on the analysis, selectively adjust prices in categories where the elasticity suggests increased profitability with minimal volume loss.

Regression Analysis: Employ regression techniques to refine understanding of price impacts on revenue and establish clearer pricing strategies.

Segment-Specific Analysis: Further analyze the data by segments to tailor pricing strategies to specific consumer groups and conditions, considering factors like promotions and seasonal demand.

Conclusion

High Elasticity: Categories with bars above the horizontal axis have positive elasticity, meaning sales increase when the price drops and decrease when the price rises. Taller bars indicate greater sensitivity. For these categories, even a small change in price could lead to a large change in the quantity sold.

Low or Negative Elasticity: Categories with bars below the axis have low or negative elasticity. Here, changes in price have a less significant impact on the quantity sold, or sales might even go up when the price increases.

Some of the High Elasticity Categories

1. American Cordials & Liqueurs
2. American Distilled Spirits Specialty
3. Cocktails / RTD
4. Coffee Liqueurs
5. Distilled Spirits Specialty

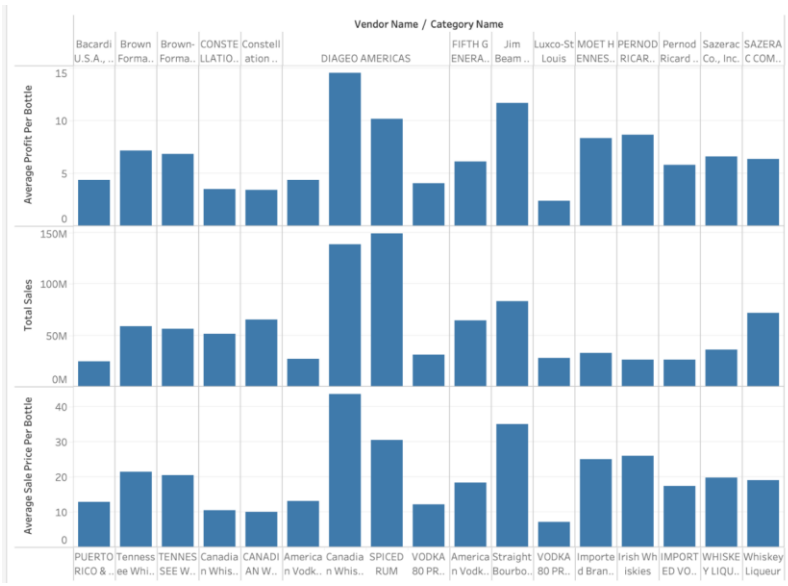
Question 3 *(Lal Reddy)*

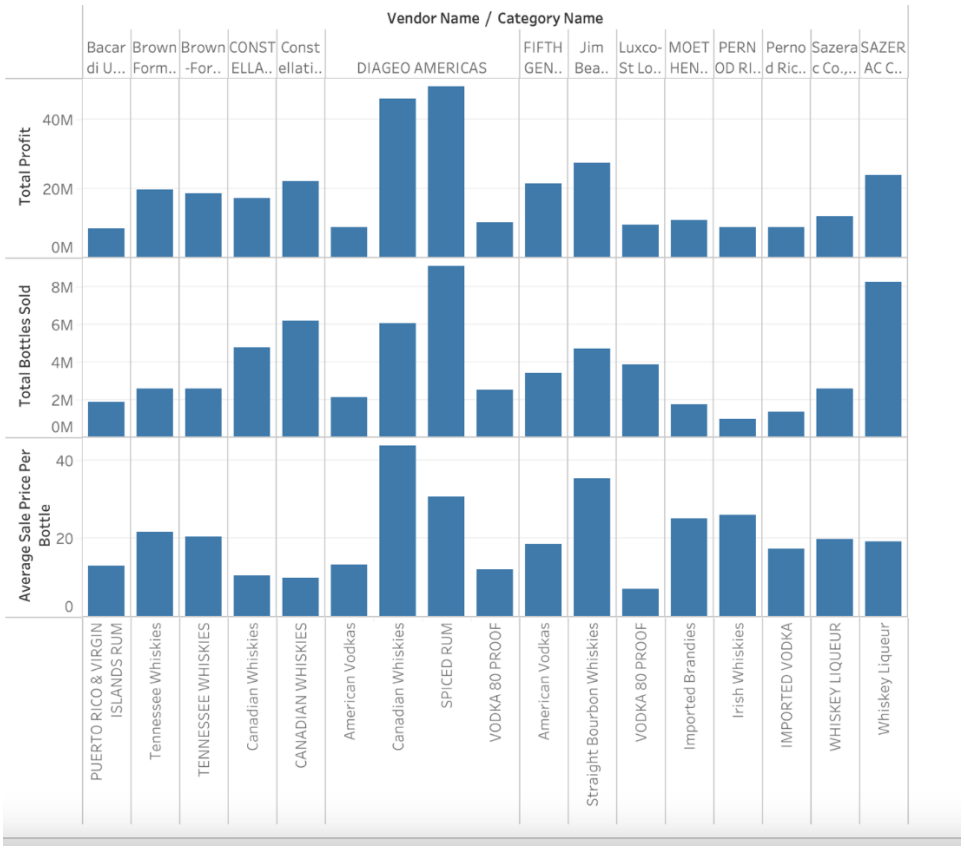
Which vendors and product categories contribute most to the store's profitability, and how can this information improve vendor negotiations and inventory decisions?

To answer the query, we did:

- Compute the sales metrics: Total Sales, Average Sale price per bottle.
- Compute the Gross Margin by Vendor and Category.
- Assess the relation between Sales metrics and Cost details corresponding to varied Categories of liquor and the associated vendors.

Vendor Name	Category Name	Total Profit	Total Sales	Average Profit Per Bottle	Average Sale Price Per Bottle
DIAGEO AMERICAS	Canadian Whiskies	2.6804918950000834E7	8.02488490299996E7	7.143848396334013	21.424394272627705
Diageo Americas	SPICED RUM	2.5952129039998062E7	7.775624414999816E7	5.1133499153228	15.31912275558262
SAZERAC COMPANY INC	Whiskey Liqueur	2.387872845000009E7	7.174731151000008E7	6.373573813244071	19.117334218285077
DIAGEO AMERICAS	Spiced Rum	2.3628504130001575E7	7.096484845000362E7	5.069567079424951	15.20599687699381
Constellation Win...	CANADIAN WHISKIES	2.2026625379998032E7	6.469269810000175E7	3.3845670638528884	10.042223192482565
FIFTH GENERATION INC	American Vodkas	2.1627639500000138E7	6.442162636000207E7	6.150680331593812	18.449947073815423
Brown Forman Corp.	Tennessee Whiskies	1.958391525000071E7	5.874403819000196E7	7.183007817679969	21.543973012614366
Diageo Americas	CANADIAN WHISKIES	1.9366998640000068E7	5.806601404999848 5E7	7.404055269845713	22.202664198667147
Brown-Forman Corp...	TENNESSEE WHISKIES	1.873814374000024E7	5.620814805000143E7	6.819289520872533	20.452059417471514
CONSTELLATION BRA...	Canadian Whiskies	1.7244171029998474E7	5.155494673000377 4E7	3.4788128994573055	10.427228472269473
Jim Beam Brands	Straight Bourbon ...	1.5640158669999797E7	4.697402999999916E7	5.825280035162035	17.472580338466685
Sazerac Co., Inc.	WHISKEY LIQUEUR	1.2055908609999834E7	3.615643032999957E7	6.610578343158289	19.828460102107012
Jim Beam Brands	STRAIGHT BOURBON ...	1.2031368129999982E7	3.59614809899989E7	5.906379084629999	17.66844211543308
MOET HENNESSY USA	Imported Brandies	1.0963034169999948E7	3.258828296999804 7E7	8.349207574082298	25.040803564273254
Diageo Americas	VODKA 80 PROOF	1.0264779020000178E7	3.068146009000004 5E7	4.074191538863648	12.18520342723558
Luxco-St Louis	VODKA 80 PROOF	9330587.69999944	2.79671475200023E7	2.4213715272187644	7.257971501464157
DIAGEO AMERICAS	American Vodkas	8882327.290000418	2.664116447000229 7E7	4.408409892135771	13.21718030171438
Pernod Ricard USA...	IMPORTED VODKA	8786084.850000104	2.628494331999779 E7	5.842784648613119	17.48440154825601
PERNOD RICARD USA	Irish Whiskies	8774770.189999944	2.603133397999879E7	8.700731483325647	26.097863307443273
Bacardi U.S.A., Inc.	PUERTO RICO & VIR...	8308912.209999895	2.462493531000018 5E7	4.3719666005355	12.980464924783734





Contributions to Profitability:

1. Vendors:

- **DIAGEO AMERICAS** shows a strong contribution to profitability, particularly in the categories of **Canadian Whiskies** and **Spiced Rum**. Their products are leading in both total profit and average profit per bottle, indicating their significance to the store's revenue.

2. Product Categories:

- Canadian Whiskies, Spiced Rum stand out with the highest total profit and a substantial average profit per bottle. **Whiskey Liquor from SAZERAC** also shows strong profitability metrics. These categories appear to be particularly lucrative for the stores.

Key Findings:

Store Performance:

The correlation between volume of sales and profit is evidential with DIAGEO AMERICAS is leading the pack in total sales.

Profitability by Category:

American Vodkas dominate sales volume but offer lower profit margins, while Imported Brandies show a higher dollar-to-volume ratio, suggesting premium pricing strategies might be effective in increasing profitability.

Vendor Analysis:

DIAGEO AMERICAS's Canadian Whisky, Spiced Rum and SAZERAC COMPANY INC's Whiskey Liqueurare top contributors to profitability.

Recommendations:

1.Strategic Inventory Management:

Prioritize stocking high-margin products such as Imported Brandies to boost profitability. Maintain an optimal balance of high-volume, lower-margin products like American Vodkas to drive traffic and complement revenue.

2.Vendor Negotiations:

Engage with top-performing vendors like DIAGEO AMERICAS for improved purchase terms, given their significant contribution to store profitability.

3.Data-Driven Promotional Strategies:

Utilize the profit ratio and volume metrics to design targeted promotional campaigns aimed at both increasing customer base and enhancing average profit per sale.

4.Market Insights Utilization:

Leverage data from high-performing stores to understand the successful tactics they employ, such as product selection, pricing strategy, and customer service.

Appendix