IE7280 – Statistical Methods in Engineering Project

Submitted By –

- ***Pranav Harish Sharma (002851959)***
- ***Bhargav Yoga (002851959)***

Case 1: Statistical Analysis of Startup Profitability Using Multiple Linear Regression

---

Introduction

Understanding the factors that drive profitability is critical for startups looking to thrive in competitive markets. This analysis uses multiple linear regression (MLR) to investigate how key variables—**R&D Spend**, **Administration**, **Marketing Spend**, and **State**—affect startup profits. The goal is to identify the most significant contributors, create a reliable predictive model, and provide actionable insights for resource allocation.

Dataset Overview

Dataset Description:
The dataset captures the investments made in **R&D**, **Administration**, and **Marketing**, alongside the **State** of operation and the resulting **Profit** of startups.

R&D Spend	Administration	Marketing Spend	State	Profit
165,349.20	136,897.80	471,784.10	New York	192,261.83
162,597.70	151,377.59	443,898.53	California	191,792.06
153,441.51	101,145.55	407,934.54	Florida	191,050.39
144,372.41	118,671.85	383,199.62	New York	182,901.99
142,107.34	91,391.77	366,168.42	Florida	166,187.94

```
Dataset preview:
    R&D Spend  Administration  Marketing Spend       State     Profit
0  165349.20       136897.80        471784.10    New York  192261.83
1  162597.70       151377.59        443898.53  California  191792.06
2  153441.51       101145.55        407934.54     Florida  191050.39
3  144372.41       118671.85        383199.62    New York  182901.99
4  142107.34        91391.77        366168.42     Florida  166187.94

Missing values in each column:
 R&D Spend         0
Administration     0
Marketing Spend    0
State              0
Profit             0
dtype: int64
```

Key Observations:

1. The dataset has no missing values, ensuring consistency in analysis.
2. **Profit** serves as the dependent variable, while the remaining variables are independent predictors.

Next Steps: Prepare the data by encoding categorical variables and normalizing numerical features for modeling.

Data Preparation

Preprocessing Insights:

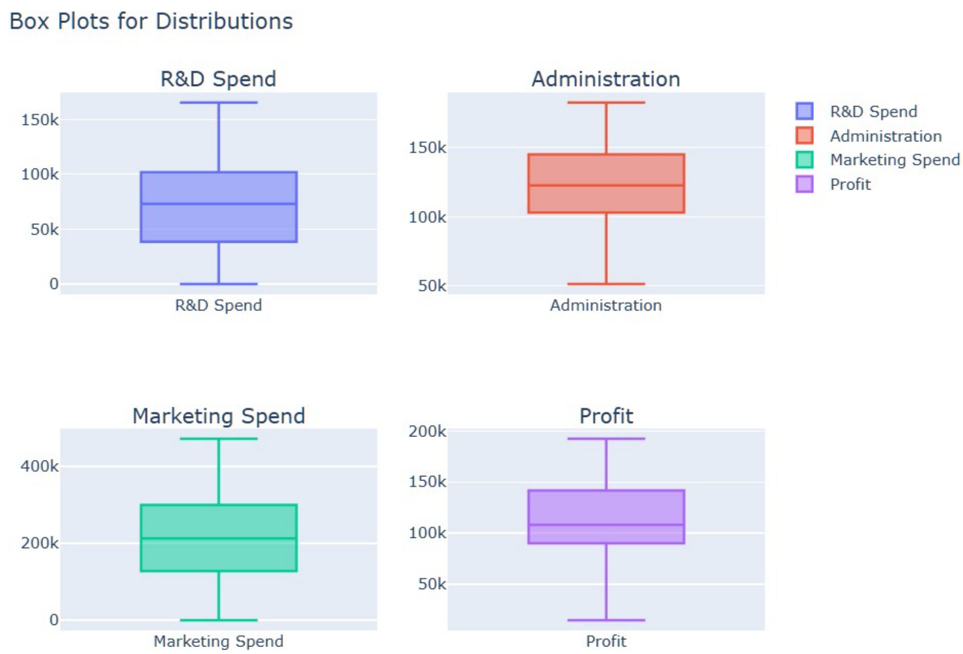
- **Missing Data Check:** The dataset contains no missing values, as summarized below:

Feature	Missing Values
R&D Spend	0
Administration	0
Marketing Spend	0
State	0
Profit	0

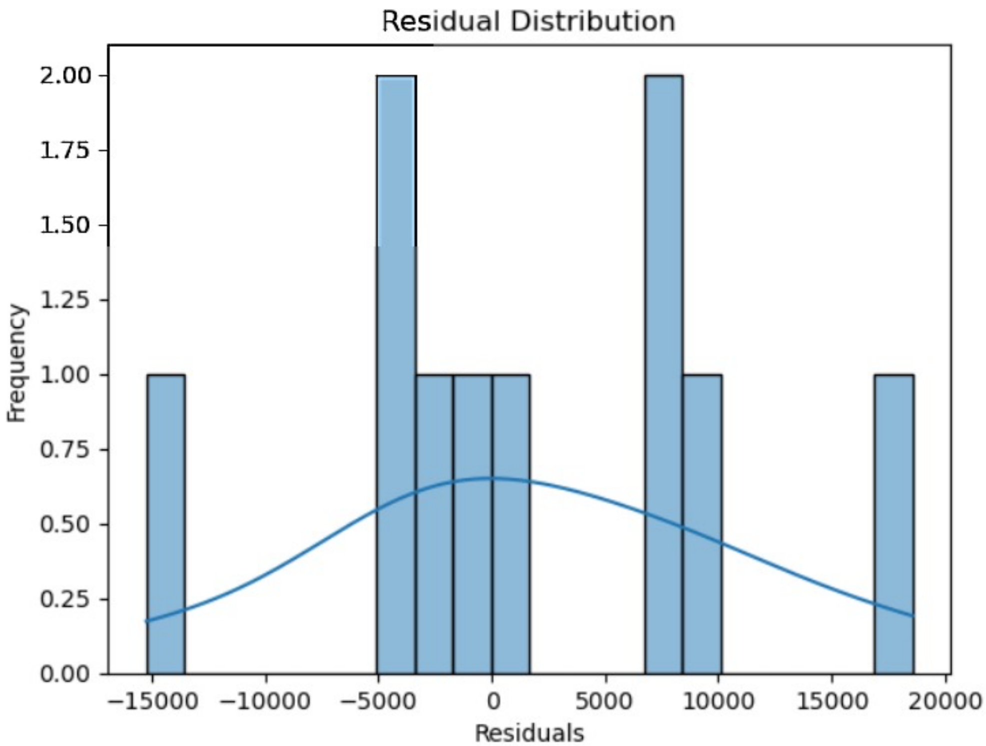
- **Categorical Variable Encoding:** The categorical variable **State** was transformed using One-Hot Encoding, generating dummy variables for modeling.
- **Scaling:** Numerical variables were standardized to ensure consistency across features.

Exploratory Data Analysis

Distribution Insights:
Box plots were generated to visualize the spread of key variables.


Box Plots for Distributions
R&D Spend
Administration
Marketing Spend
Profit
150k
100k
50k
0
200k
400k
R&D Spend
Administration
Marketing Spend
Profit


Observed vs Predicted Profits
Difference
15k
10k
5k
Predicted Profit
Observed Profit
40k
60k
80k
100k
120k
140k



Variable	Key Observations
R&D Spend	Significant variability, positively skewed.
Administration	More uniform distribution, with moderate spread.
Marketing Spend	Similar to Administration in uniformity.
Profit	Some outliers, suggesting potential leverage points.

Actionable Insights:
The distributions suggest that R&D spending is highly variable and likely plays a pivotal role in driving profits. Further analysis is needed to confirm these relationships.

Regression Model Results

Model Summary:

The multiple linear regression model produced the following coefficients:

```
Regression Formula:
Profit(y) = 111688.86 + (-315.26 x Dummy State 1) + (623.53 x Dummy State 2) + (-308.27 x Dummy State 3) + (36608.57 x R&D Spen
d) + (-1907.92 x Administration) + (3614.34 x Marketing)

Predicted Profit for example input: 230141.38
```

Predictor	Coefficient
Intercept	111,688.86
Dummy State 1	-315.26
Dummy State 2	623.53
Dummy State 3	-308.27
R&D Spend	36,608.57
Administration	-1,907.92
Marketing Spend	3,614.34

Model Performance:

- Mean Squared Error (MSEMSE): **82,010,363.05**
- R^2 Score: **0.90**

Key Findings:

1. **R&D Spend** is the strongest predictor of profit, with a large positive coefficient.
2. **Administration** has a negative effect, highlighting possible inefficiencies.
3. **Marketing Spend** shows a moderate positive influence, indicating its role in driving revenue.
4. **State** appears to have minimal direct impact on profits.

Residual Analysis

Residual distribution was analyzed to assess model assumptions.

Insights:

1. Residuals are approximately normally distributed, supporting the assumption of linearity.
2. Minor outliers indicate room for model refinement, such as exploring robust regression techniques or transforming variables.

Prediction Example

Scenario: A startup in **New York** invests:

- R&D Spend = 250,000
- Administration=150,000Administration = 150,000
- MarketingSpend=600,000Marketing Spend = 600,000

Predicted Profit:

230,141.38230,141.38

Interpretation:
This prediction underscores the strong influence of R&D spending on profitability. Startups can use this insight to prioritize resource allocation.

Conclusion

Key Takeaways:

- **R&D Spend** is the primary driver of profitability among startups.
- High administrative costs may hurt profits, suggesting an opportunity for cost optimization.
- The model explains **90% of the variance** in profits, demonstrating its reliability.

Future Directions:

- Investigate non-linear relationships or interactions between features.
- Consider adding variables such as industry type or market conditions for broader applicability.

Case 2: Effect of Seasonal Trends on Retail Sales

A Two-Way ANOVA Analysis

Objective

This research will focus on understanding the relationship between different advertising strategies and different seasonal regularities better with the aim to enhance retail turnover. We carry out an extensive study of the movement of various product categories or types of items from January to December. In doing so, we would like to cover how particular periods of the year and specific products sold interact with one another in order to determine the most effective sales during that period. These findings would be crucial to the manager in the formulation of appropriate strategies targeting factors that are effective in enhancing the sales. Their understanding of these relationships enables retailers to time their promotions and marketing activities with the expected demand for the particular items. This would improve sales.

Dataset Link: Link (https://www.kaggle.com/datasets/abdullah0a/retail-sales-data-with-seasonal-trends-and-marketing)

Data Source: Kaggle

Dataset Description

The dataset contains retail sales data with the following attributes:

- **YEAR, MONTH**: Temporal information indicating the year and month of sales.
- **SUPPLIER, ITEM_CODE, ITEM_DESCRIPTION, ITEM_TYPE**: Identifiers for the product and supplier.
- **RETAIL_SALES, RETAIL_TRANSFERS, WAREHOUSE_SALES**: Sales figures at different retail and warehouse levels.
- **Log_Retail_Sales**: Log-transformed retail sales for normalization.

	YEAR	MONTH	SUPPLIER	ITEM_CODE	ITEM_DESCRIPTION	ITEM_TYPE	RETAIL_SALES	RETAIL_TRANSFERS	WAREHOUSE_SALES	Log_Retail_Sale
0	2020	1	REPUBLIC NATIONAL DISTRIBUTING CO	100009	BOOTLEG RED - 750ML	WINE	0.00	0.0	2.0	0.00000
1	2020	1	PWSWN INC	100024	MOMENT DE PLAISIR - 750ML	WINE	0.00	1.0	4.0	0.00000
2	2020	1	RELIABLE CHURCHILL LLLP	1001	S SMITH ORGANIC PEAR CIDER - 18.7OZ	BEER	0.00	0.0	1.0	0.00000
3	2020	1	LANTERNA DISTRIBUTORS INC	100145	SCHLINK HAUS KABINETT - 750ML	WINE	0.00	0.0	1.0	0.00000
4	2020	1	DIONYSOS IMPORTS INC	100293	SANTORINI GAVALA WHITE - 750ML	WINE	0.82	0.0	0.0	0.59883
5	2020	1	KYSELA PERE ET FILS LTD	100641	CORTENOVA VENETO P/GRIG - 750ML	WINE	2.76	0.0	6.0	1.32441
6	2020	1	SANTA MARGHERITA USA INC	100749	SANTA MARGHERITA P/GRIG ALTO - 375ML	WINE	0.08	1.0	1.0	0.07696
7	2020	1	BROWN-FORMAN BEVERAGES WORLDWIDE	1008	JACK DANIELS COUNTRY COCKTAIL SOUTHERN PEACH -...	BEER	0.00	0.0	2.0	0.00000
8	2020	1	JIM BEAM BRANDS CO	10103	KNOB CREEK BOURBON 9YR - 100P - 375ML	LIQUOR	6.41	4.0	0.0	2.00283
9	2020	1	INTERNATIONAL CELLARS LLC	101117	KSARA CAB - 750ML	WINE	0.33	1.0	2.0	0.28517

Missing Values Check

The dataset is checked for the missing values with the following results

Table 1: Missing Values

Column	Missing Count
YEAR	0
MONTH	0
SUPPLIER	0
ITEM_CODE	0
ITEM_DESCRIPTION	0
ITEM_TYPE	0
RETAIL_SALES	0
RETAIL_TRANSFERS	0
WAREHOUSE_SALES	0
Log_Retail_Sales	0

Analysis: No missing values are present in the dataset, ensuring completeness for statistical and exploratory analysis.

Next Steps:

- To proceed with data exploration and statistical modeling without requiring imputation or handling of missing values.

Descriptive Statistics

The descriptive statistics of the dataset is as the following:

Table 2: Descriptive Statistics

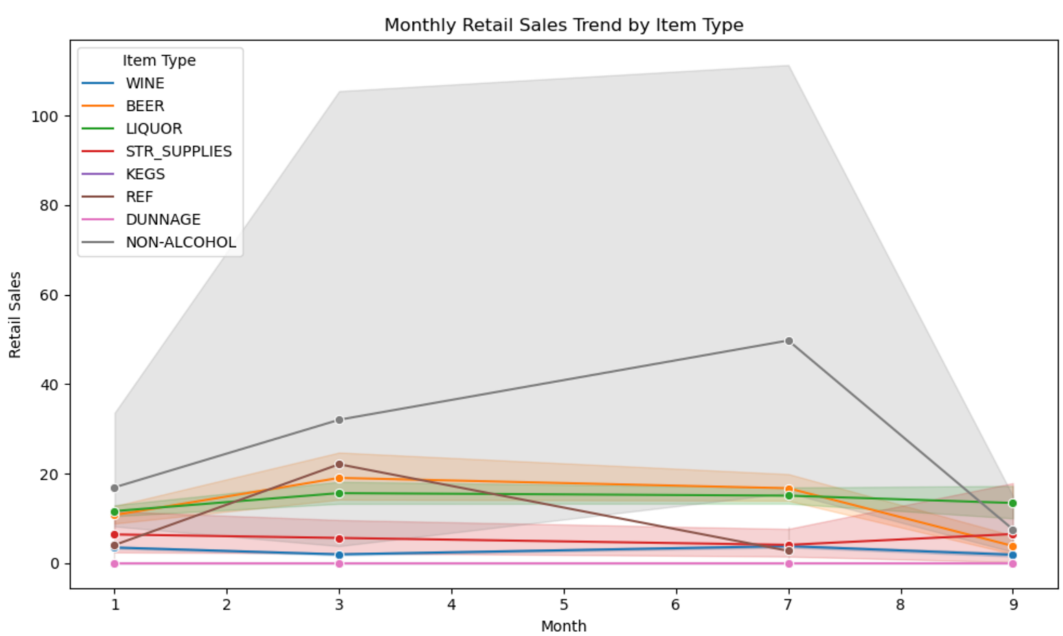
Statistic	YEAR	MONTH	RETAIL_SALES	RETAIL_TRANSFERS	WAREHOUSE_SALES	Log_Retail_Sales
Count	30,000	30,000	30,000	30,000	30,000	30,000
Mean	2020.0	3.91	6.94	6.59	27.88	0.84
Standard Deviation	0.0	2.83	33.08	27.88	270.33	1.23
Minimum	2020.0	1.0	0.0	0.0	0.0	0.0
25th Percentile	2020.0	1.0	0.0	0.0	0.0	0.0
Median	2020.0	3.0	0.16	0.0	1.0	0.15
75th Percentile	2020.0	7.0	2.92	3.0	6.0	1.37
Maximum	2020	9.0	2739.0	1507.0	18317.0	7.92

Analysis:

- The sales data exhibits significant variance, as shown by the high standard deviation.
- Retail sales values are heavily skewed, with most values concentrated near the lower range (as indicated by the median and 75th percentile).

Exploratory Data Analysis

Visualization 1: Monthly Retail Sales Trend by Item Type

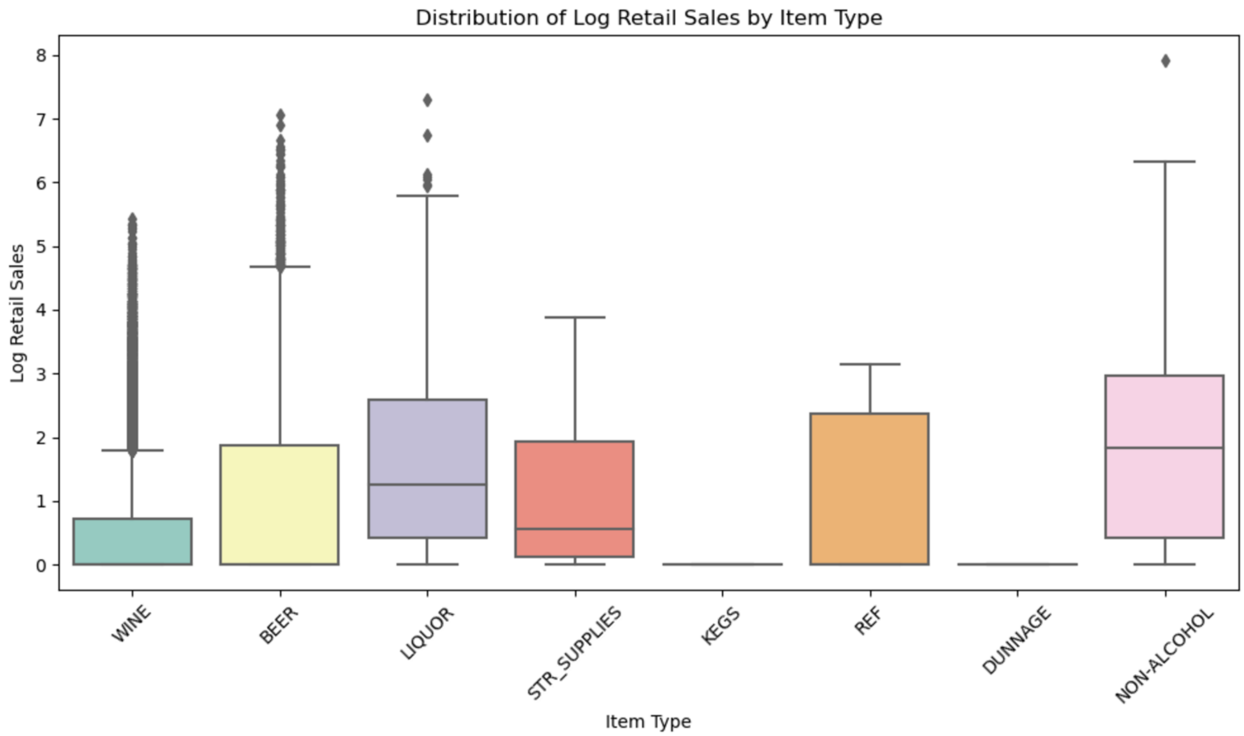


A line plot showing the trend of retail sales for different item types across months.

Insights:

- Certain item types (e.g., WINE or BEER) may show spikes in specific months, likely influenced by holidays or seasonal demand.

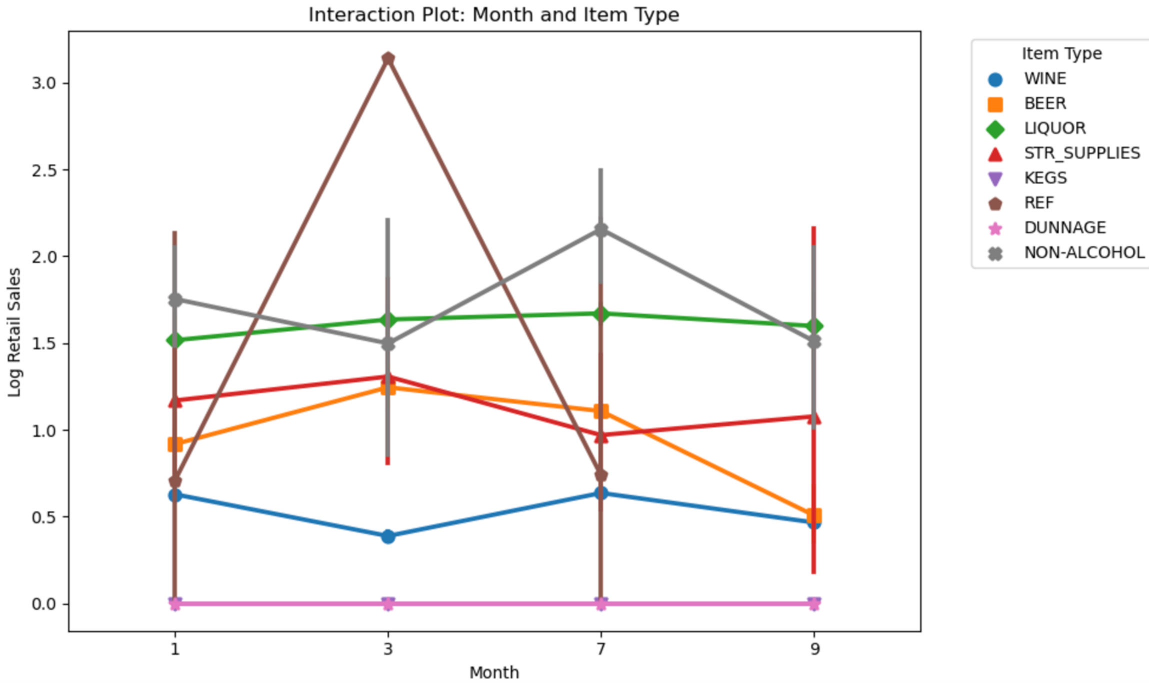
Visualization 2: Distribution of Log Sales by Item Type



Insights:

- Some item types (e.g., LIQUOR) may have broader ranges, indicating higher variability in sales.
- Log transformation reduces the effect of extreme values, allowing for a clearer comparison across item types.

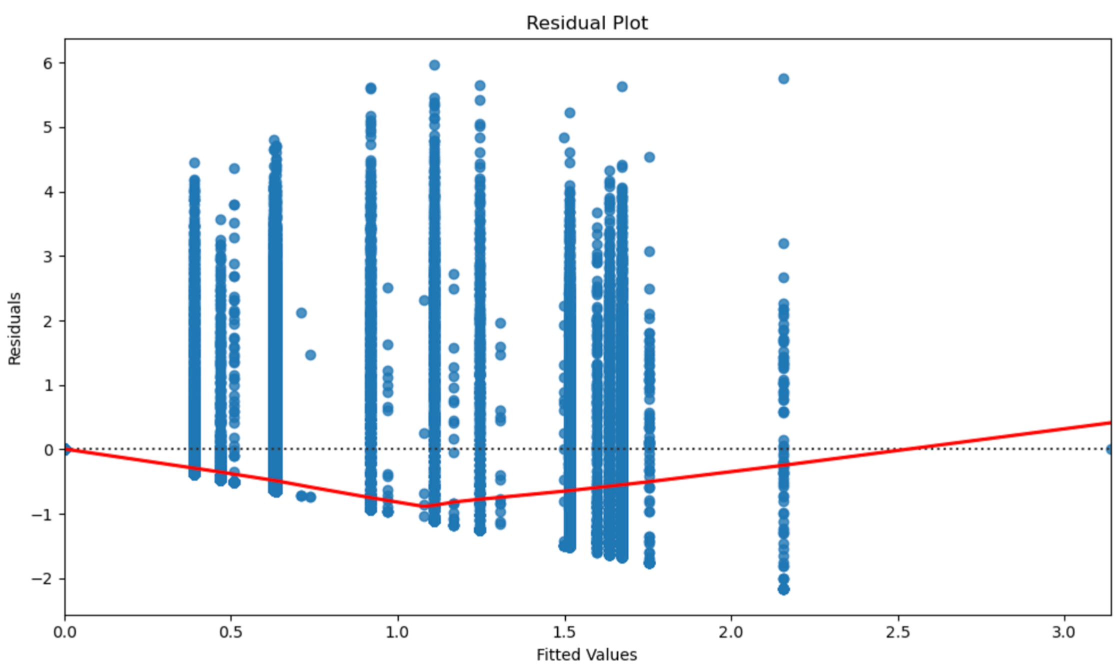
Visualization 3: Interaction plot: Month and Item Type



Insights:

- This interaction suggests promotional or seasonal targeting strategies should vary by item type.

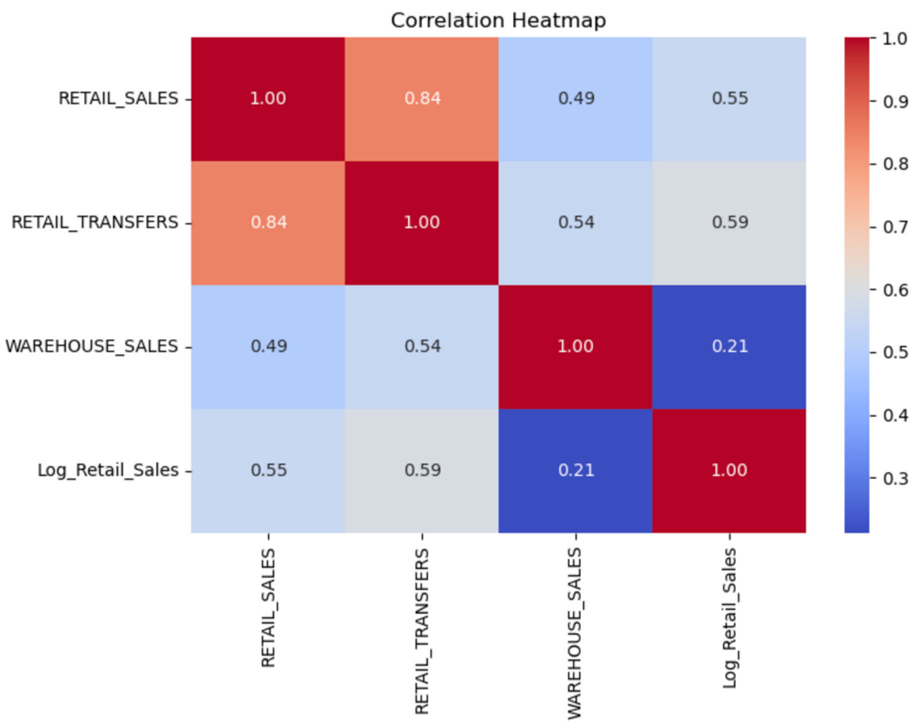
Visualization 4: Residual Plot



- **Insights**:
  - Random scatter confirms a good model fit.

- Non-random patterns may indicate the need for model refinement, such as adding interaction terms or transformations.

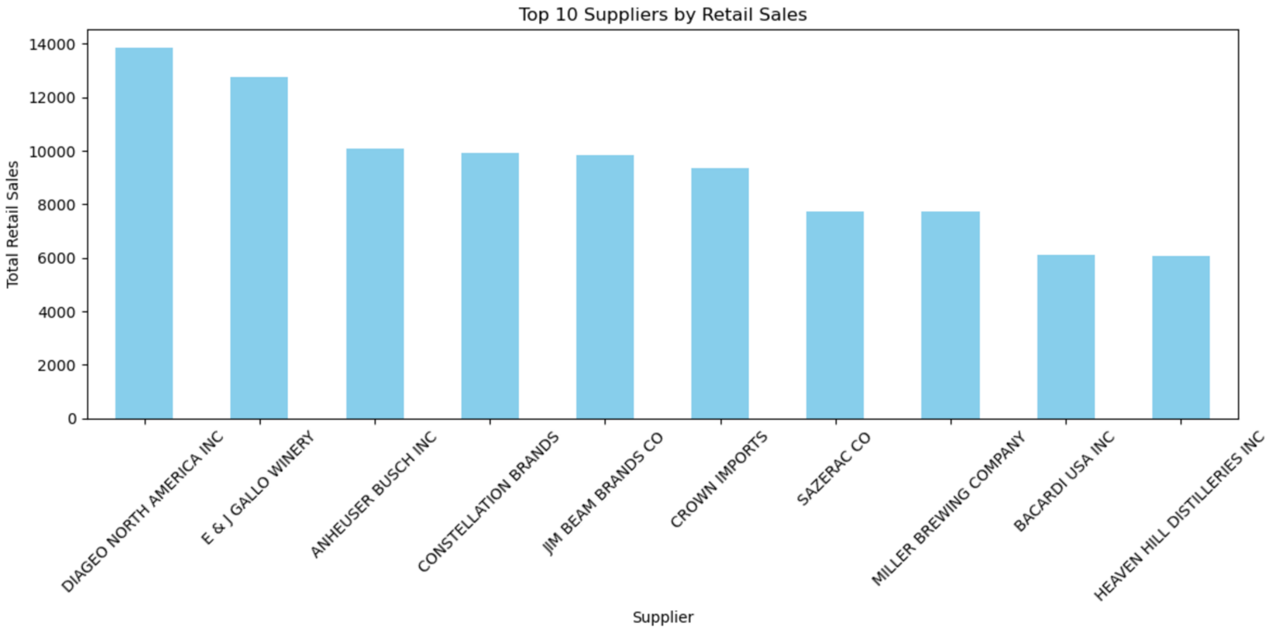
Visualization 5: Correlation Heatmap



- **Insights:**
  - Identifying key predictors for retail sales can inform modeling decisions.
  - Highlighting multicollinearity issues can guide variable selection

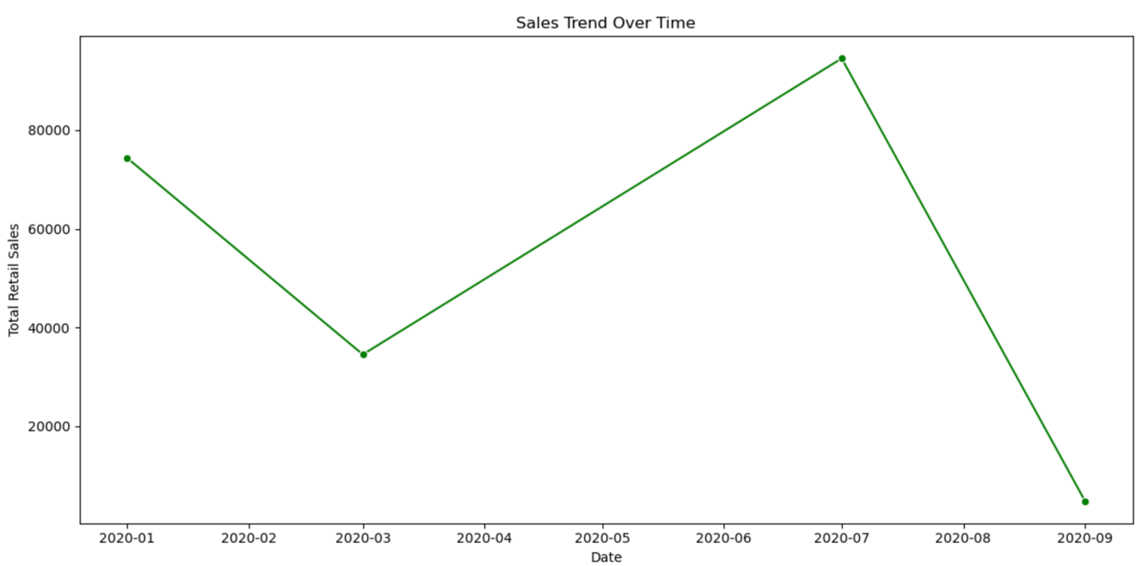
---

Visualization 6: Top 10 suppliers by Retail Sales



- **Insights:**
    - These suppliers are critical for overall revenue. Strategies such as exclusive contracts or promotional campaigns may prioritize them.
    - Lower-ranked suppliers might offer growth opportunities through increased collaboration.

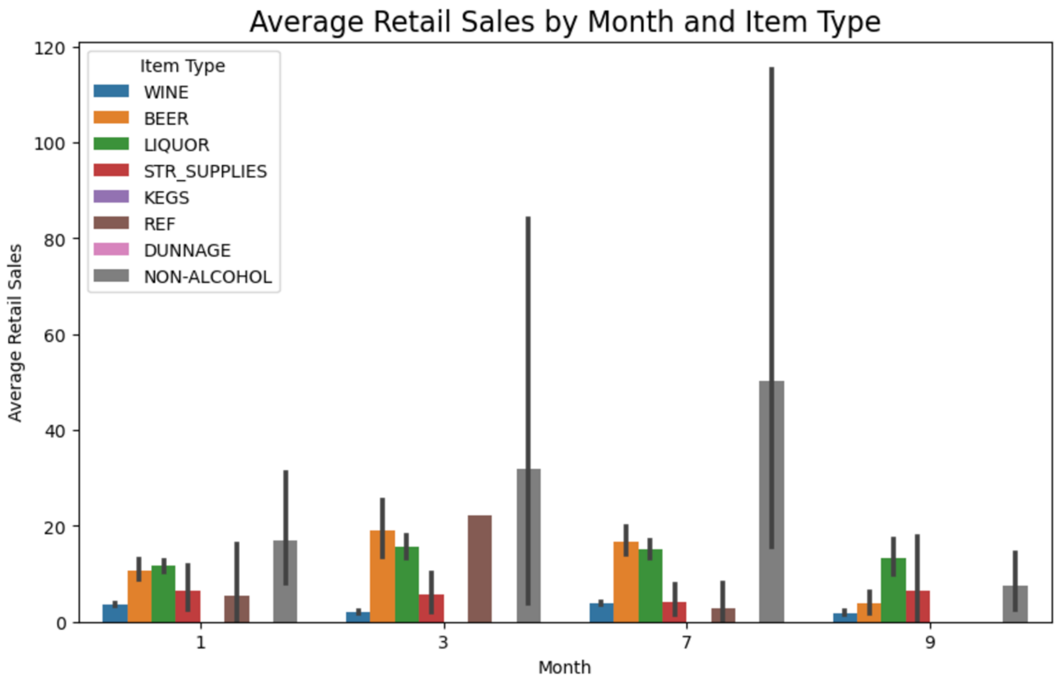
Visualization 7: Sales Trend Over Time

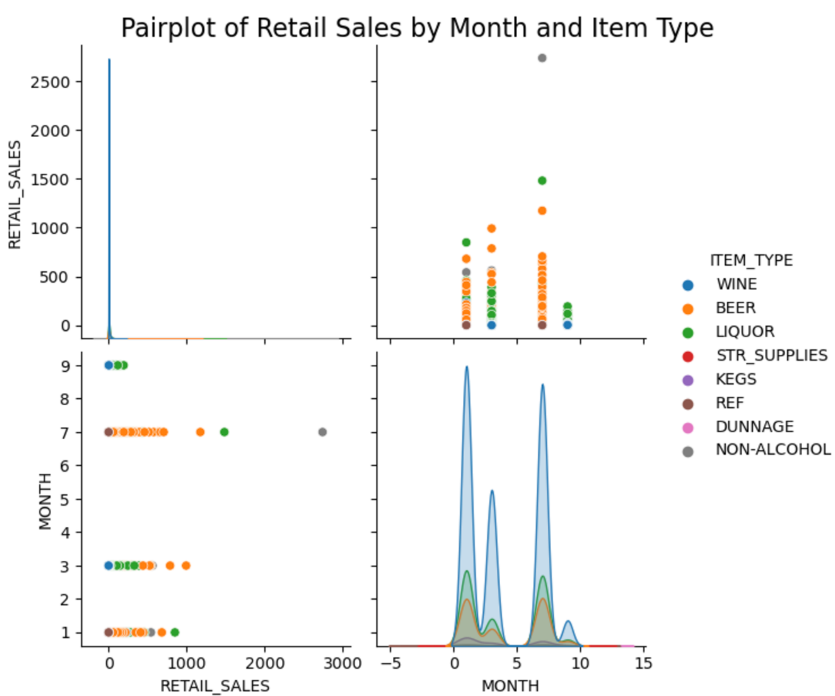


Insights:

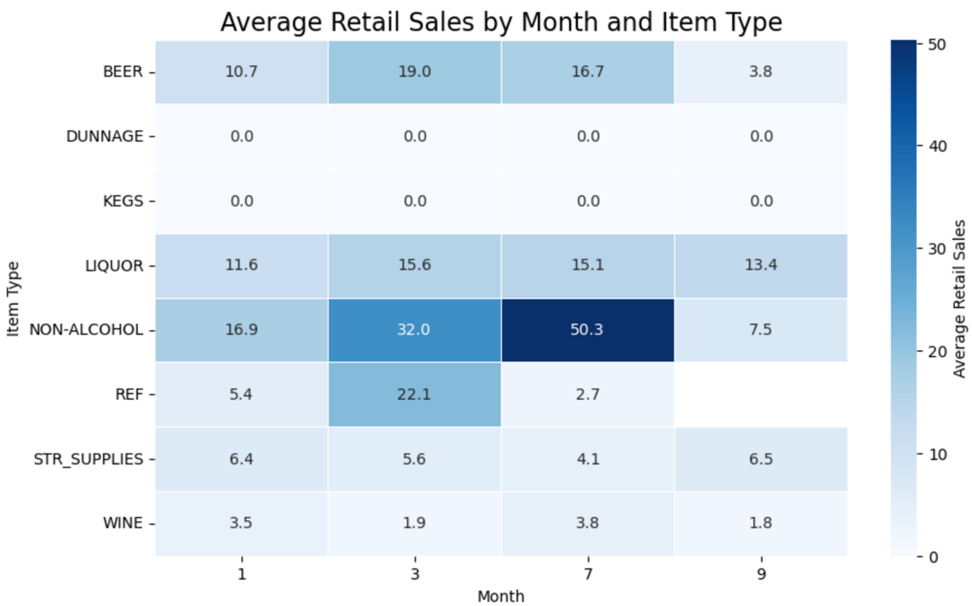
- Month of July might reflect successful strategies as has the highest total retail sales.
- Sharp declines could signal external disruptions (e.g., economic downturns).

Visualization 8: Retail Sales by Month and Item Type

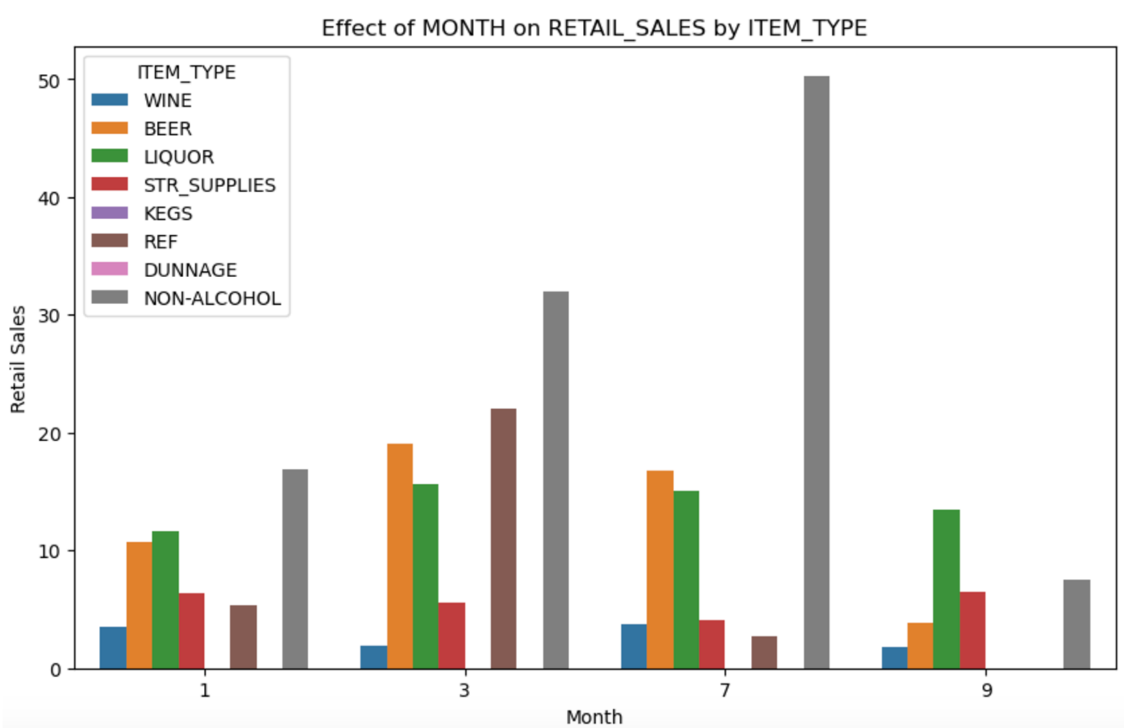




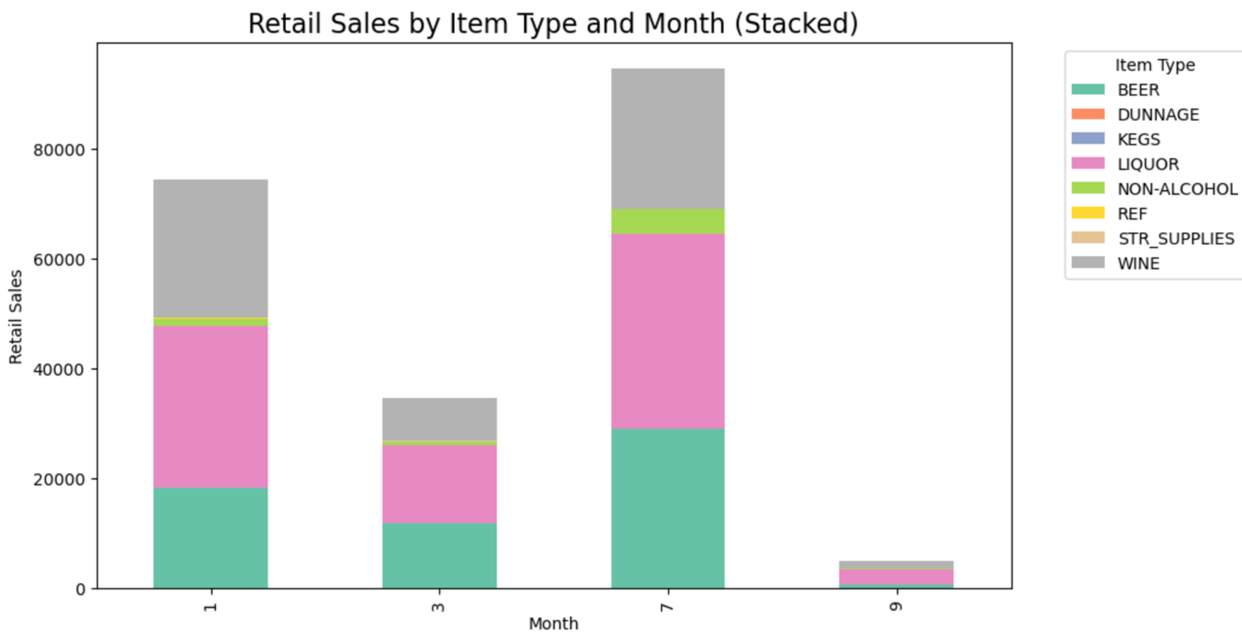
Visualization 9: Avg Retail Sales by Month and Item Type



Visualization 10: Effect of Month on Retails_Sales by Item_Type



Visualization 11: Avg Retail Sales by Month and Item Type



Statistical Analysis

Levene's Test for Variance Homogeneity

Table 3: Levene's Test Results

Test Statistic	p-value
65.12	6.10×10^{-65}

```
Levene's Test p-value: 6.109884473865627e-65
```

Analysis:

- As the P value of Levene's Test is very low, meaning there is a significant difference in the variances of all the group structures which do not conform to the requirement for the ANOVA of homogeneity of variances.
- Such an violation can greatly diminish the validity of the obtained results using the ANOVA.

Next Steps:

- To deal with the above problem use some robustness checks such as use Welch's ANOVA and the nonparametric alternatives.
- To interpret the results of ANOVAs with some enmity keeping the limitation in consideration and report the results with full transparency.

Two-Way ANOVA

Table 4: ANOVA Results

Source	Sum of Squares	df	F-Value	p-Value
C(MONTH)	77.86	3	19.97	2.16×10^{-9}
C(ITEM_TYPE)	6529.45	7	717.65	<0.001
C(MONTH):C(ITEM_TYPE)	226.34	21	8.29	6.65×10^{-25}
Residual	38952.90	29969		

```
Two-Way ANOVA Results:
                              sum_sq       df           F        PR(>F)
C(MONTH)                   77.857579      3.0   19.966967  2.158882e-09
C(ITEM_TYPE)             6529.453110      7.0  717.647562  0.000000e+00
C(MONTH):C(ITEM_TYPE)     226.344323     21.0    8.292448  6.651986e-25
Residual                38952.900826  29969.0         NaN           NaN
```

Analysis:

- **Main Effects:** In discussions regarding retail sales volume both **MONTH** and **ITEM_TYPE** appear to have an effect as it is witnessed with their very low p values.
- **Interaction Effect:** Monthly differences in retail sales are further modified by the type of product sold as shown by significant interaction between **MONTH** and **ITEM_TYPE**. This shows how sales of different items moves through time.

Next Steps:

- Carry out a post - hoc analysis (Tukey's HSD test) to find out which particular months or item types attract significant differences.
- Utilize these findings to further develop these marketing strategies e.g., trying to promote goods when the season is right and when goods of a certain type are selling better.
- Perform post-hoc analysis (e.g., Tukey's HSD) to identify specific group differences.

Tukey's HSD Test

Table 5: Tukey's HSD Results

Tukey's HSD Test Results:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
BEER	DUNNAGE	-1.0271	0.0272	-1.991	-0.0633	True
BEER	KEGS	-1.0271	0.0	-1.1604	-0.8939	True
BEER	LIQUOR	0.5705	0.0	0.5007	0.6402	True
BEER	NON-ALCOHOL	0.8516	0.0	0.6095	1.0937	True
BEER	REF	-0.0029	1.0	-1.2308	1.2251	False
BEER	STR_SUPPLIES	0.1018	0.997	-0.3386	0.5422	False
BEER	WINE	-0.4533	0.0	-0.5125	-0.3942	True
DUNNAGE	KEGS	0.0	1.0	-0.9701	0.9701	False
DUNNAGE	LIQUOR	1.5976	0.0	0.6342	2.561	True
DUNNAGE	NON-ALCOHOL	1.8787	0.0	0.8878	2.8696	True
DUNNAGE	REF	1.0242	0.4879	-0.535	2.5835	False
DUNNAGE	STR_SUPPLIES	1.1289	0.0266	0.0719	2.1859	True
DUNNAGE	WINE	0.5738	0.6156	-0.3889	1.5365	False
KEGS	LIQUOR	1.5976	0.0	1.4675	1.7276	True
KEGS	NON-ALCOHOL	1.8787	0.0	1.6129	2.1445	True
KEGS	REF	1.0242	0.1876	-0.2086	2.2571	False
KEGS	STR_SUPPLIES	1.1289	0.0	0.6751	1.5828	True
KEGS	WINE	0.5738	0.0	0.4491	0.6985	True
LIQUOR	NON-ALCOHOL	0.2812	0.0093	0.0408	0.5215	True
LIQUOR	REF	-0.5733	0.8504	-1.8009	0.6543	False
LIQUOR	STR_SUPPLIES	-0.4686	0.027	-0.9081	-0.0292	True
LIQUOR	WINE	-1.0238	0.0	-1.0753	-0.9723	True
NON-ALCOHOL	REF	-0.8545	0.4325	-2.1038	0.3948	False
NON-ALCOHOL	STR_SUPPLIES	-0.7498	0.0001	-1.2466	-0.2529	True
NON-ALCOHOL	WINE	-1.3049	0.0	-1.5424	-1.0675	True
REF	STR_SUPPLIES	0.1047	1.0	-1.1977	1.407	False
REF	WINE	-0.4504	0.9543	-1.6775	0.7766	False
STR_SUPPLIES	WINE	-0.5551	0.0031	-0.993	-0.1172	True

Analysis:

- The test indicates that there are differences between certain pairs of product groups such as Beer and Wine and the statistical evidence to back these is p-values of < 0.05.
- Targeted business strategies may be implemented in groups that have shown statistical significance in differences.

Next Steps:

- Focus on advertising strategy that targets the product market that worked. For example: Start vigorous promotion for products that sell well such as beers and liquor.
- Create motivation strategies that will encourage the higher selling of more dormant products. Shift gears as needed but do not interrupt continuous performance monitoring.

Shapiro-Wilk Test for Normality

Table 6: Shapiro-Wilk Test Results

Test Statistic	p-value
0.76	0.0

Analysis:

- The p-value indicates that the residuals do not have normal distribution, which is a strong deviation from the mean.
- Although this is a violation of the normality condition for ANOVA, its effect is lessened by the size of the sample n =29, 969 which is sufficiently large.

Insights

Our analysis unveils key insights into how seasonality and product types influence retail sales. These insights provide a foundation for actionable strategies to tailor advertising efforts effectively, boosting sales across various product categories and time periods.

Seasonal Trends and Their Impact

The Two-Way ANOVA results show that MONTH significantly affects retail sales (p-value = 2.16×10^{-9}). This highlights the strong role of seasonality in shaping consumer behavior. Sales spikes during certain months can be linked to seasonal demand, cultural events, or regional festivals.

Actionable Insight:

- Seasonal Advertising Campaigns: Prioritize advertising budgets for high-performing months by aligning campaigns with seasonal demand. For instance, promote beverages like beer and wine during summer or holiday seasons when social gatherings are common.

Effect of Product Types on Sales

The ITEM_TYPE variable demonstrates a substantial impact on retail sales, with a $p\text{-value} < 0.001$. Tukey's HSD test further shows significant differences between product categories:

- Beverages (Beer and Liquor) consistently outperform others, offering clear opportunities for targeted promotions.
- Non-Alcoholic Drinks and Kegs display niche market trends.

Actionable Insight:

- Product-Specific Advertising: Invest in promoting high-demand products like beer and liquor during peak seasons. For niche products, such as kegs or non-alcoholic beverages, create specialized campaigns targeting smaller, focused audience segments.

Interplay Between Seasonality and Product Types

The significant interaction between MONTH and ITEM_TYPE ($p\text{-value} = 6.65 \times 10^{-25}$) reveals that sales patterns for specific products vary by month. For example:

- Beer sales peak during warmer months.
- Wine and liquor demand rises in colder months or festive seasons.

Actionable Insight:

- Dynamic Advertising Strategies: Develop adaptable advertising plans based on month-product interactions. Examples:
  - Promote beer and non-alcoholic beverages during summer.
  - Focus on liquor and wine advertising around winter holidays and New Year celebrations.

Budget Allocation Based on Residual Analysis

The Shapiro-Wilk test reveals that residuals are not normally distributed, suggesting possible outliers or clusters in specific months or product categories. While this doesn't undermine the results, it indicates potential areas for deeper exploration, such as niche markets or underperforming segments.

Actionable Insight:

- **Customized Advertising Budgets:** Direct resources toward identifying and targeting underperforming segments. For instance, if certain months or products are flagged as outliers, experiment with localized or online campaigns to drive interest in those areas.

Conclusion

The effect of seasonality:

- Retail sales are heavily dependent on the time of the year as sales are inclined towards certain months.
- Add seasonality comprehension when doing advertisement and stock planning so they will coincide with the time of greatest demand.

Performance According to Specific Products:

- Beer and liquor for example are certain product divisions that perform better than others consistently.
- Make order not to mismanage resources, Special focus on advertising and stock control of the most profitable product divisions.

Dynamics of interaction:

- Month and product type are influential on each other which means both seasonality and inter-product competition need to be accounted for.

Appendix:

GitHub Repository for Python notebook:

(GitHub - Link) https://github.com/Pranavsharma13/StatisticalMethodsFInEngineering

---