**Product Cluster Analysis**

**Warehouse and Retail Sales**

**Project Overview**

**Problem Statement:**

The company is facing problems with the management of its inventory, which leads to overstocking of some products and understocking of others.

There is also a need to customize the products offered to suit the tastes of the customers and stay ahead in a dynamic market.

Without actionable insights, decisions on purchasing, marketing, and sales strategies are made based on intuition rather than data, thus limiting the company's ability to attain optimal performance.

**What questions must be asked of stockholders to gain more information on business and datasets?**

Regarding the overstocking/understocking issue, what are the current processes for forecasting demand and managing inventory levels, and *why* are they leading to imbalances?

What systems are in place to track inventory levels and movement, and *why* are they not providing accurate real-time visibility?

How does the company currently gather and utilize customer feedback regarding desired product features and customization options, and *why* is it insufficient?

How are decisions about purchasing, marketing, and sales strategies currently made, and *why* is intuition prioritized over data analysis?

What data is currently collected and how is it stored, and *why* isn't it being effectively used for insights?

What tools and technologies are currently used for data analysis, and *why* aren't they sufficient for addressing the company's challenges?

What are the estimated costs associated with the current inventory inefficiencies (overstocking, lost sales), and *why* haven't these costs driven change?

How does the company measure the success of its marketing and sales strategies, and *why* are these metrics not effectively linked to inventory management?

What are the potential benefits (e.g., increased sales, reduced costs) of implementing a more data-driven approach to inventory management and product customization, and *why* hasn't this been prioritized?

**Data Collection**

**Data Sources:**

* **The company has provided the dataset.**
* **Data Volume:** The dataset contains **307,645 rows** and **9 columns**, indicating a substantial size suitable for clustering analysis.
* **Memory Usage:** The dataset requires approximately **21.1 MB**, which is manageable for most data analysis workflows.
* **Data Types:**
  + **Numeric Data:** 5 columns, including **float64 (3)** and **int64 (2)**.
  + **Categorical Data:** 4 columns of object type, capturing textual or categorical information.
* **Null Values:**
  + Missing values are present in 4 columns:
    - SUPPLIER: 167 missing entries.
    - ITEM TYPE: 1 missing entry.
    - RETAIL SALES: 3 missing entries.
  + All other columns are complete.
* **Duplicates:** No duplicate rows were found in the dataset.
* **Outliers:**
  + Approximately **1.52%** of rows contain outlier attributes.
  + Outliers are present in RETAIL TRANSFERS (**3560** rows) and WAREHOUSE SALES (**1749** rows).

### **Feature Explanation**

1. **YEAR (int64):**
   * Represents the year of the transaction or record.
   * **Example**: 2024.
   * No missing values or outliers.
2. **MONTH (int64):**
   * Represents the month of the transaction or record.
   * **Example**: 1 for January.
   * No missing values or outliers.
3. **SUPPLIER (object):**
   * Represents the name or identifier of the supplier.
   * **Example**: Supplier\_A.
   * Contains **167** missing values, possibly due to incomplete data capture or unrecorded suppliers.
4. **ITEM CODE (object):**
   * Unique identifier for each product.
   * **Example**: A12345.
   * No missing values or duplicates.
5. **ITEM DESCRIPTION (object):**
   * Provides a detailed description of the product.
   * **Example**: Wireless Headphones.
   * Complete and descriptive.
6. **ITEM TYPE (object):**
   * Represents the category or type of the item.
   * **Example**: Electronics.
   * Contains 1 missing value.
7. **RETAIL SALES (float64):**
   * Represents the revenue generated from retail sales for each item.
   * **Example**: 500.75.
   * Contains 3 missing values but no outliers.
8. **RETAIL TRANSFERS (float64):**
   * Represents the quantity of items transferred between retail stores.
   * **Example**: 20.0.
   * Outliers identified in 3560 rows.
9. **WAREHOUSE SALES (float64):**
   * Represents the quantity of items sold from the warehouse.
   * **Example**: 100.0.
   * Outliers identified in 1749 rows.

**Data Preprocessing:**

**Handling Missing Values:**

* + **SUPPLIER Column:**
    - The **167** null values were replaced with a default value of "**YEAR**".
  + **ITEM TYPE Column:**
    - The single null value was replaced with "**NON-ALCOHOL**".
  + **RETAIL SALES, RETAIL TRANSFERS, WAREHOUSE SALES Columns:**
    - Null values were replaced with 0 before rounding off values.

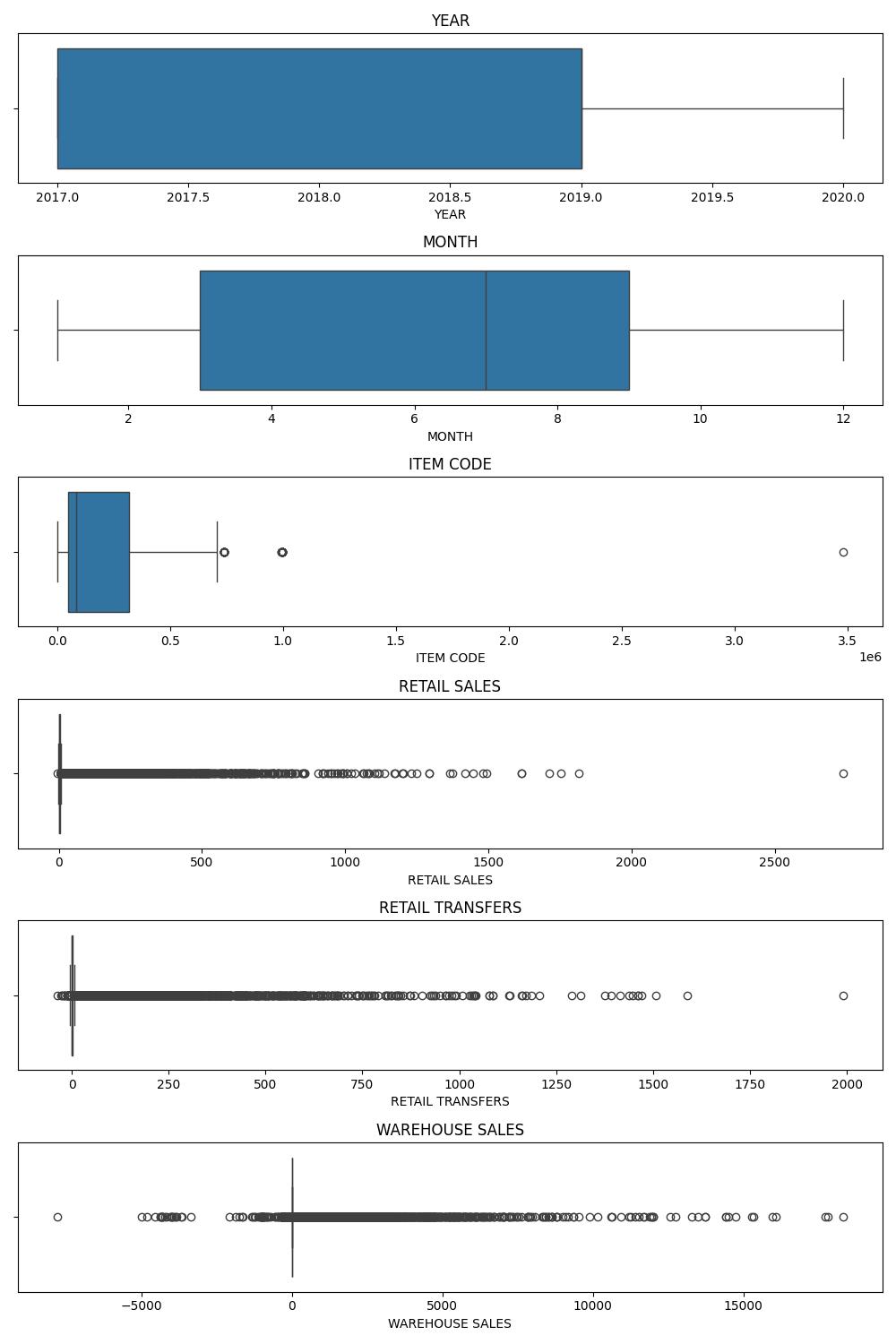
**Data Transformation:**

* + **ITEM CODE Column:**
    - Converted categorical values to numerical representations:
      1. "**BC**" converted to 11.
      2. "**WC**" converted to 22.
  + **ITEM DESCRIPTION Column:**
    - The column contained non-alphanumeric, non-space, and non-point characters.
    - It was split into 8 separate columns, and the contents were merged according to a defined pattern.
    - Extracted and separated numerical and alphabetical strings into two new columns:
      1. **Brand Name** (Alphabetical part).
      2. **Units** (Numerical part).
    - Standardized the Units column to ensure consistent formatting.

**Rounding Off Values:**

* The values in RETAIL SALES, RETAIL TRANSFERS, and WAREHOUSE SALES columns were rounded to the nearest whole number.

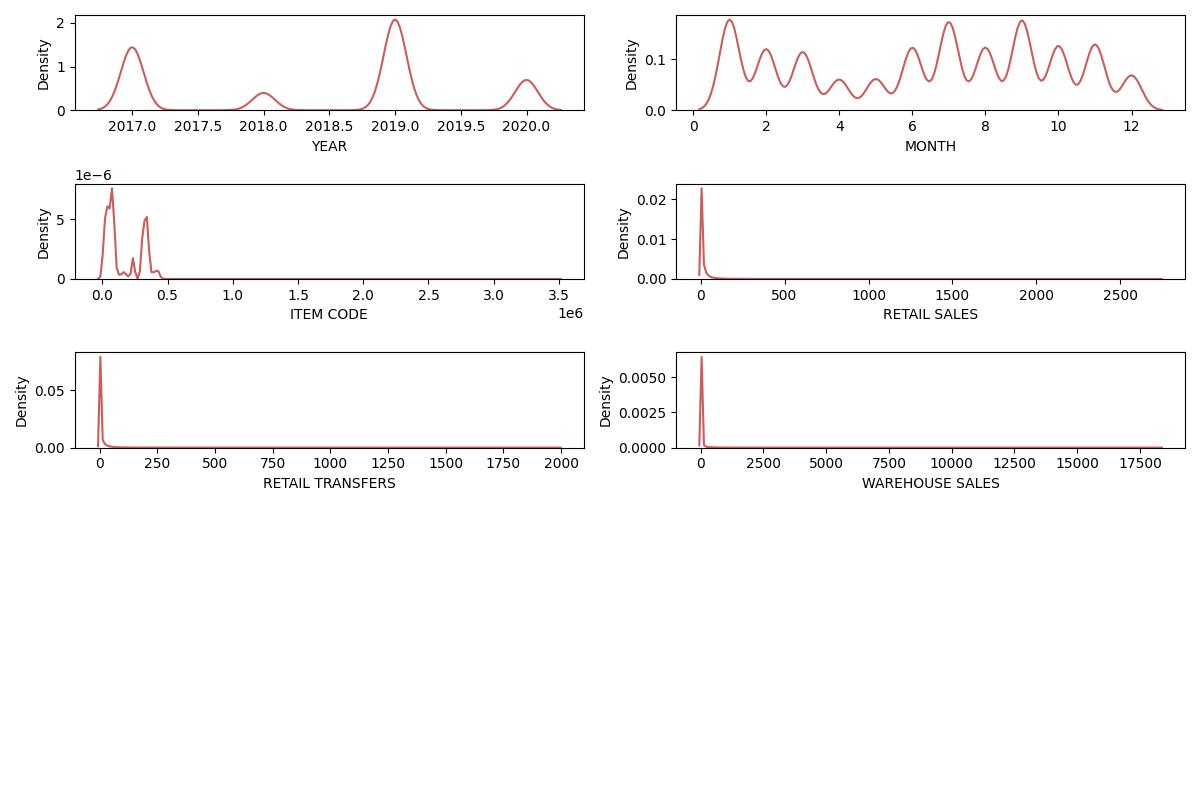
**Renaming Columns for Readability:**

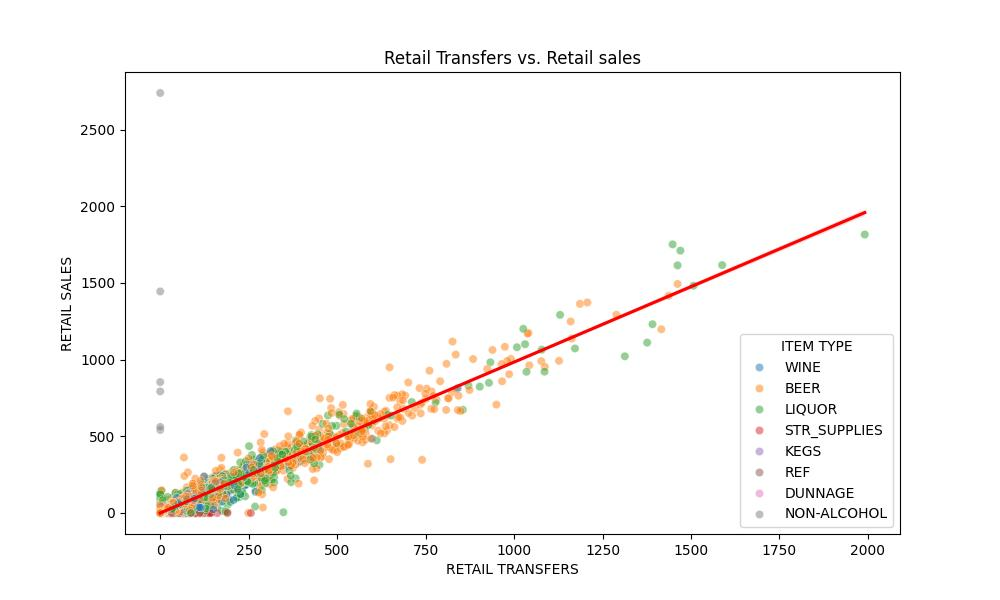
* **Joined\_column** → Renamed to **BRAND** **NAME** for the sake of its brand.
* **number\_column** → Renamed to **QUANTITY**, which represents the numerical part of item descriptions.
* **Standardized unit** → **STD UNIT** - renamed for standardization in units of measurement.
* **number\_and\_unit** → Renamed to **NUMBER UNIT**, combining quantity and standardized unit for clarity.
* **Summary of Outliers:**

Visualization (**checked\_outlier.png**): This barplot summary identifies features with potential outliers, thus helping to prioritize which attributes require treatment

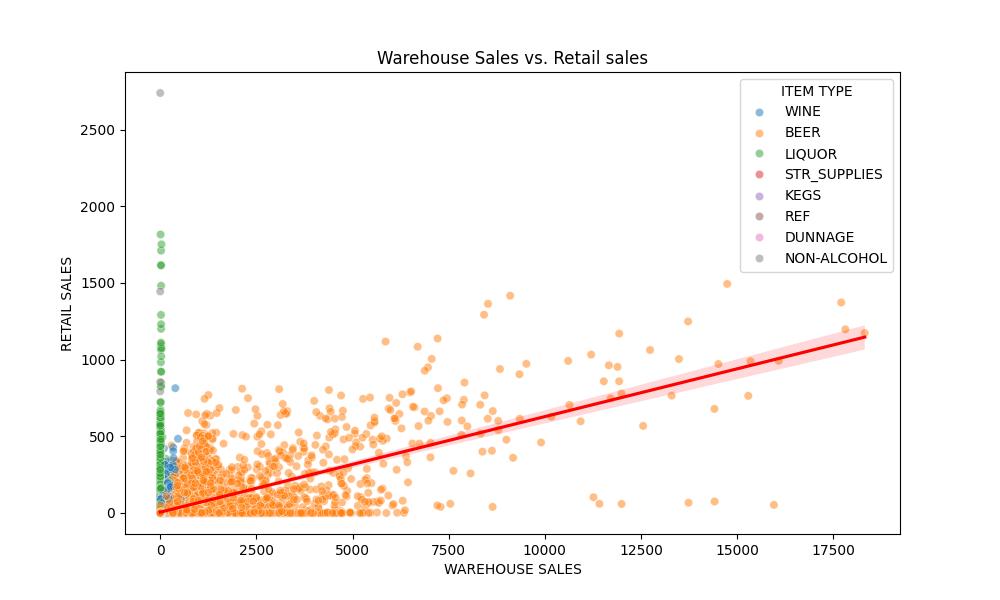
**Key highlights:**

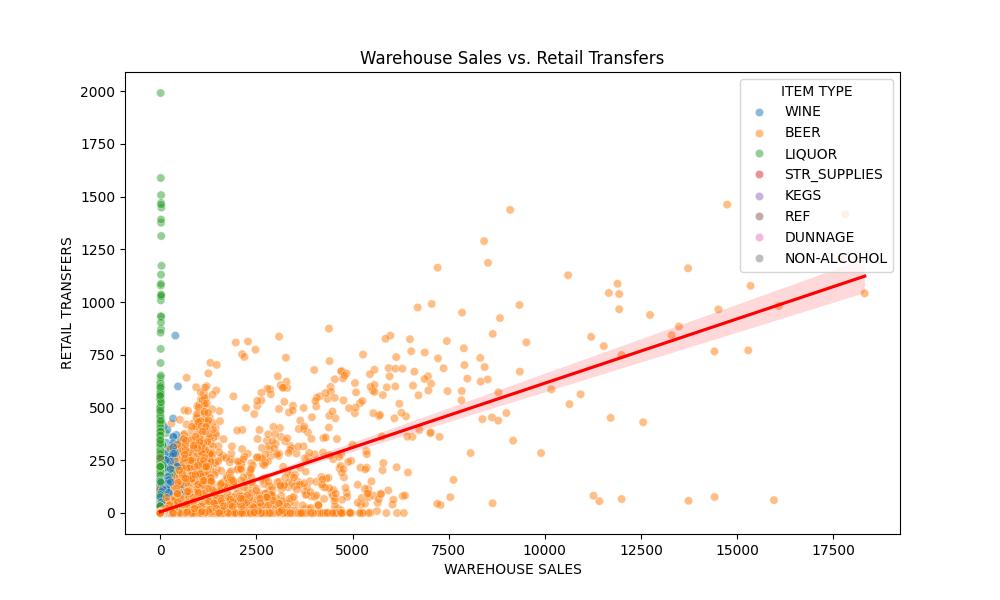
* **Retail Transfers** and **Warehouse Sales, Retail Sales** exhibit strong outlier presence suggesting anomalous data that may bias some analyses.
* No outliers were found in features such as **YEAR**, **MONTH**.

**Distribution of data before scaling**

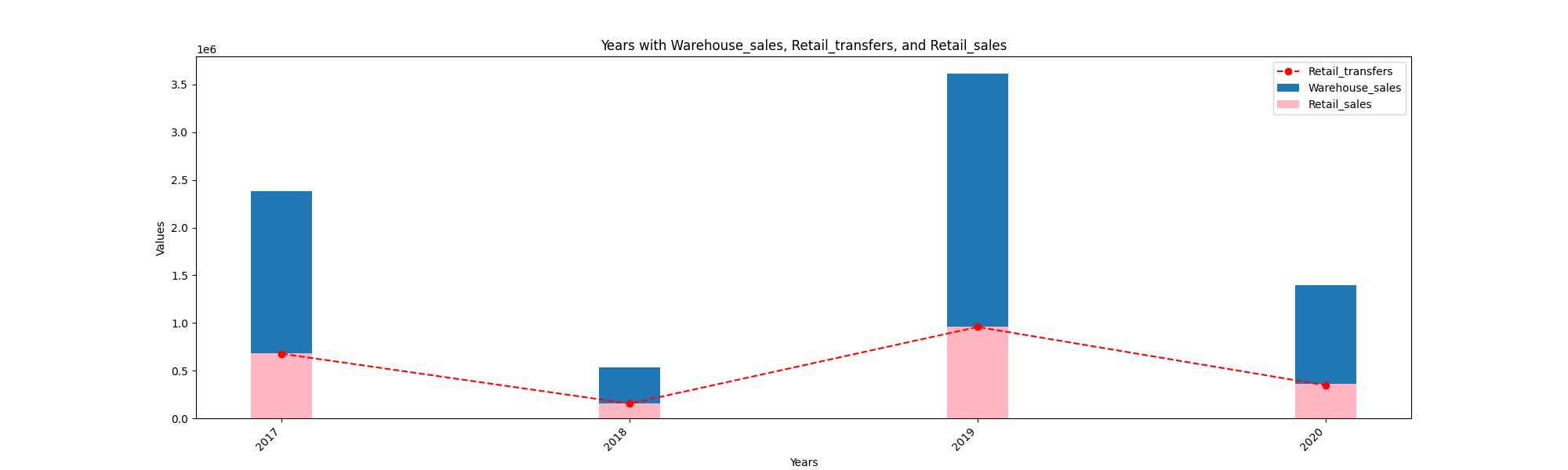
****

**Before Treat outliers Retail Transfers VS Retail Sales**

****

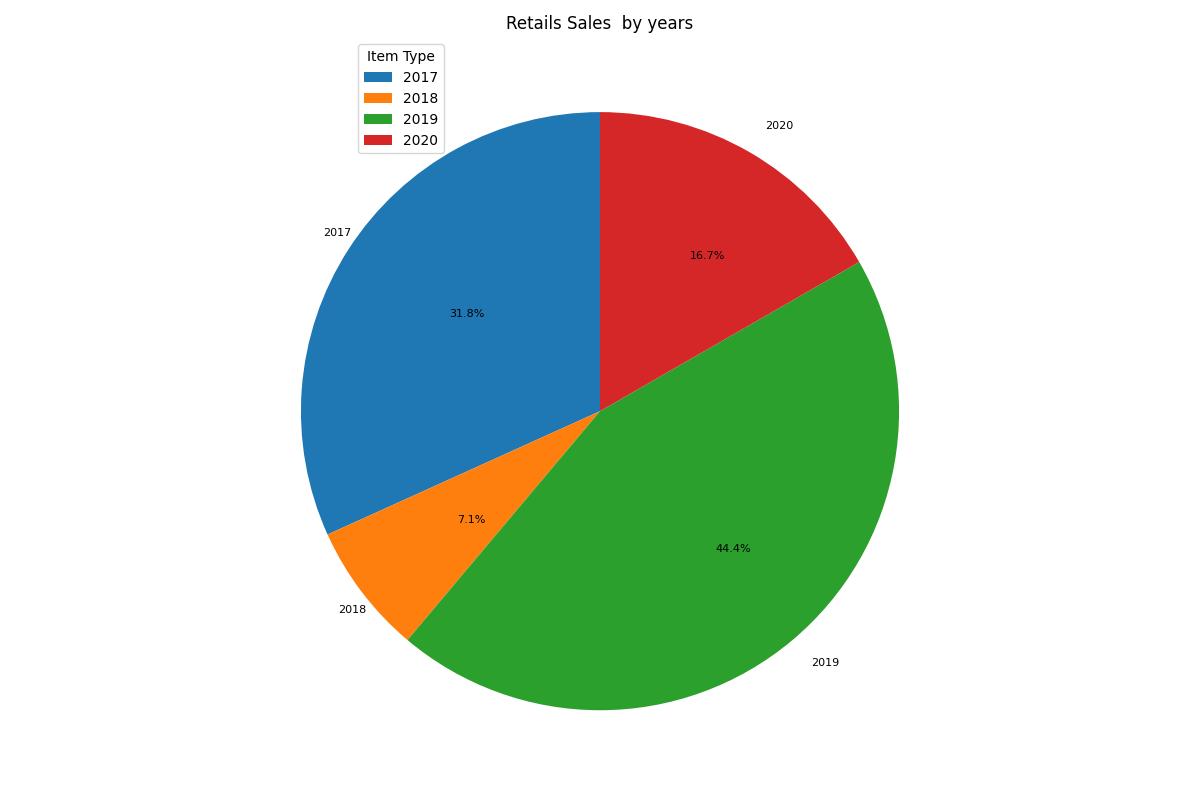
**Before Treat Outlier Warehouse Sales Vs Retail Sales**

**Before Treat Outlier Warehouse Sales Vs Retail Transfer**

1. **Exploratory Data Analysis (EDA)**

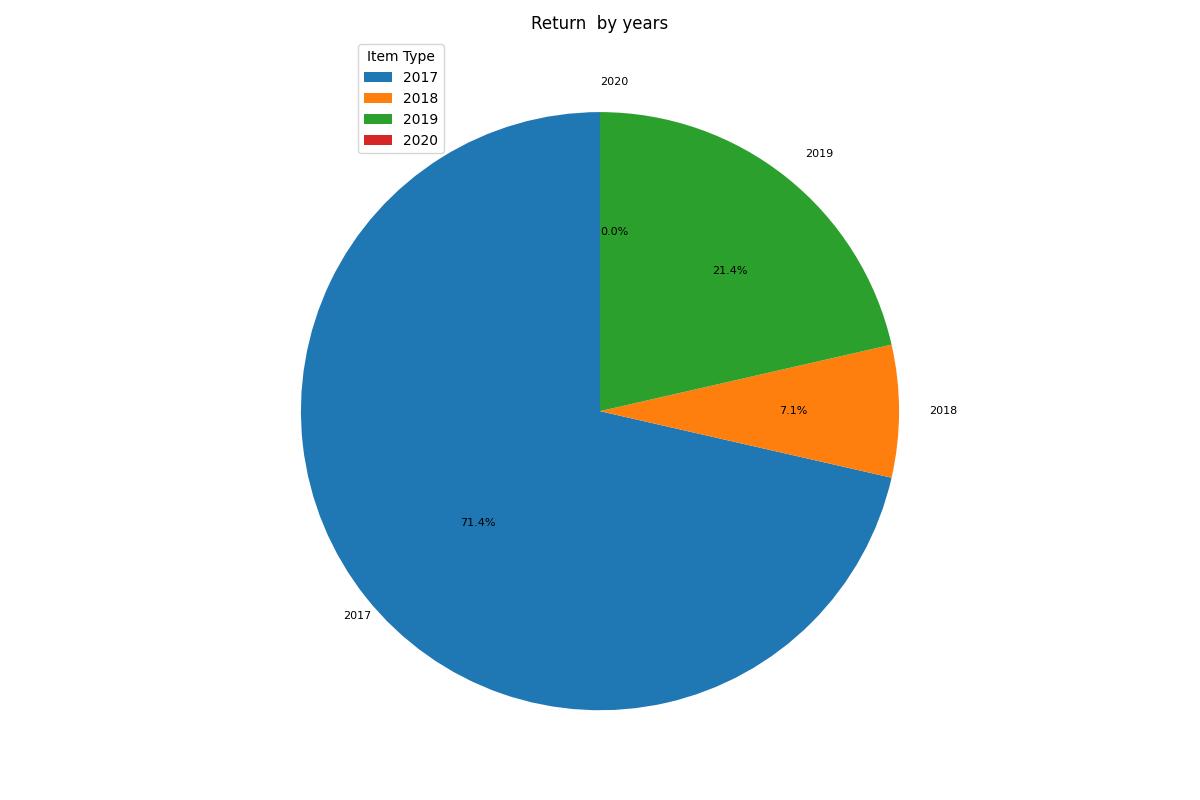
**Years\_with\_warehouse\_sales\_and\_retail\_transfers\_and\_Retail\_sales**

In the graph, we observe that retail sales, represented by the pink bar, follow the trend of retail transfers each year, particularly after the transfer stocks have sold out. Additionally, when we examine warehouse sales in **2019**, we notice that the stock levels were higher compared to previous years. This trend is influenced by the global pandemic caused by **COVID-19**, which resulted in widespread lockdowns. This information is based on my research findings.

****

**Retails Sales By Years**

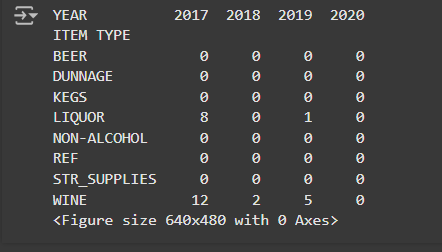
According to **Retail Sales** by Year, we can see which year sold out: **2019** saw sales of **44.4%**, followed by **2017** with **31.8%**. We had a bad year in **2018** compared to others. We had a fantastic year in **2019**, but in **2020** we sold barley at a rate of **16.7%**. We need to concentrate next and develop a plan for the upcoming years.

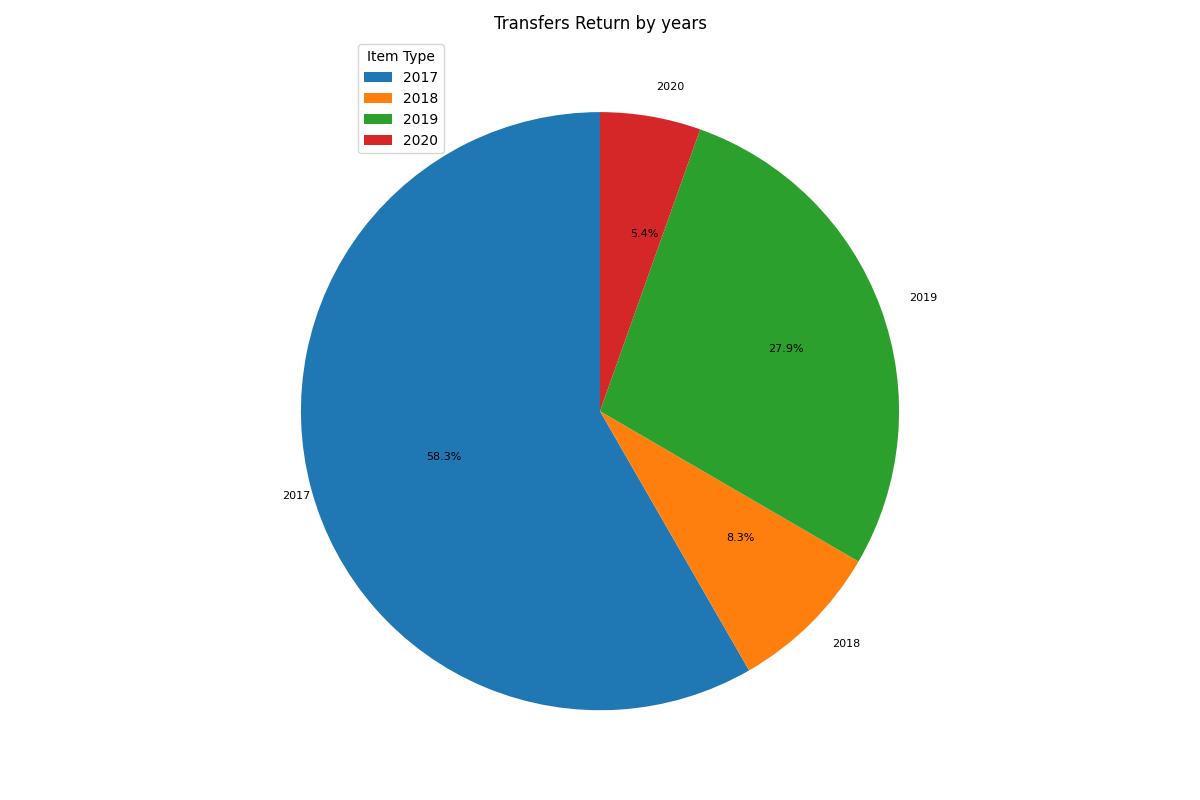
****

**Return By Years**

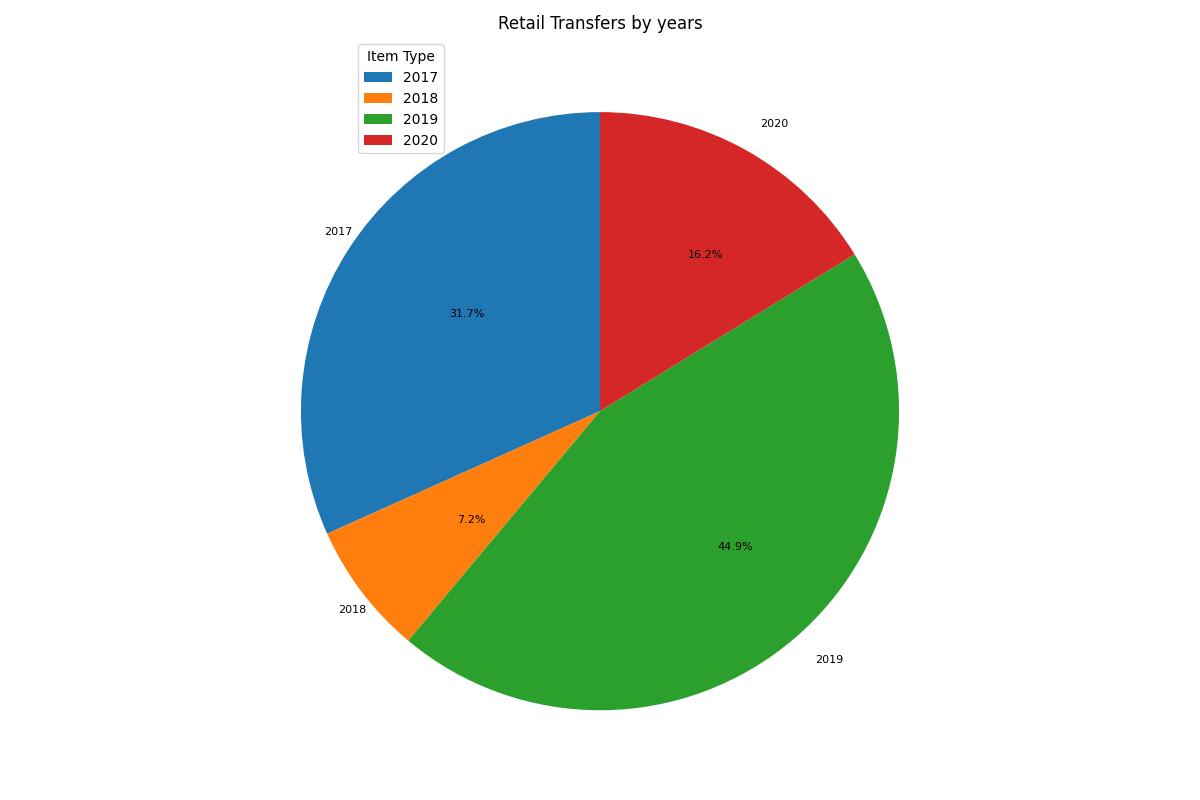
In **2017**, **71.4%** of products were returned by customers, compared to **31.8%** of sold-out items. We received returned goods, and now there are some questions.

The product type that customers return is displayed in the pivot table below.



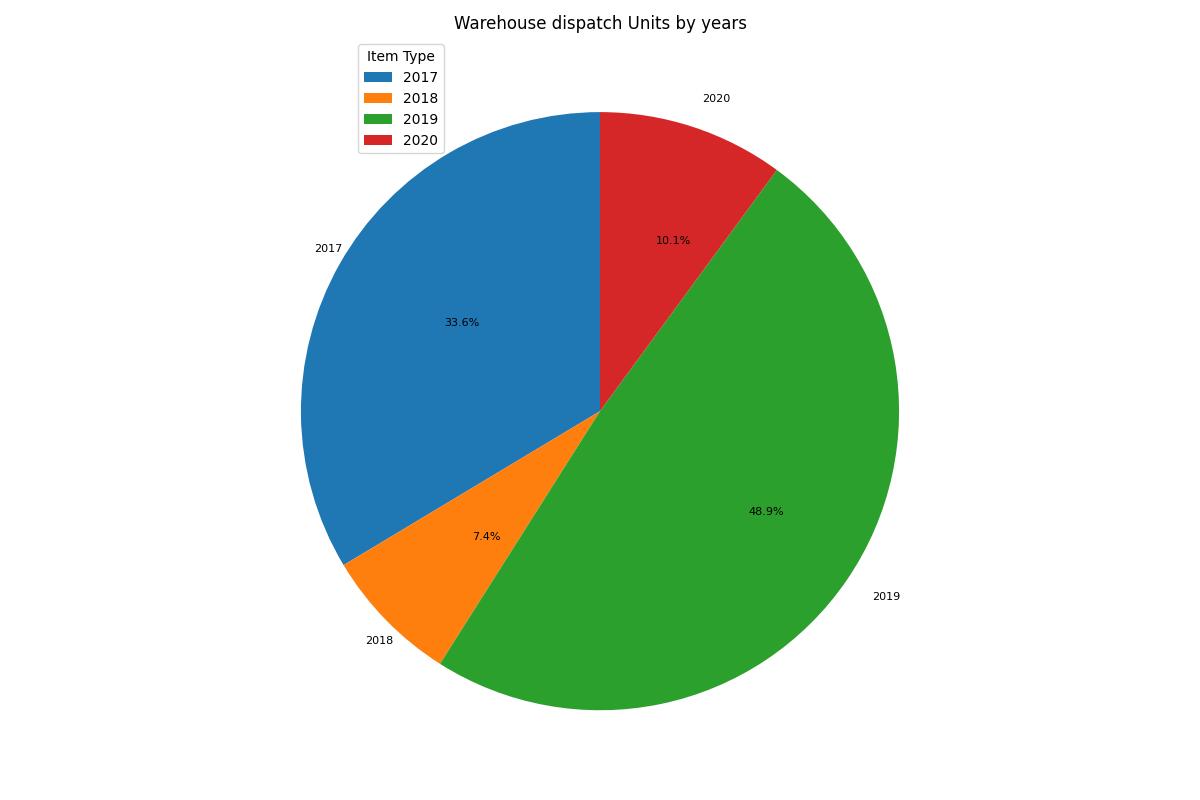
****

**Transfers Return By Years**

****The graph displays the percentage of products returned from retail stores to the transfer station. According to the pie chart, in **2017**, **58.3%** of products were returned. However, this percentage dropped significantly in **2020**, reaching its lowest point at **5.4%**. Notably, in **2019**, we experienced our best sales year, but returns accounted for **27.9%** in that same year.

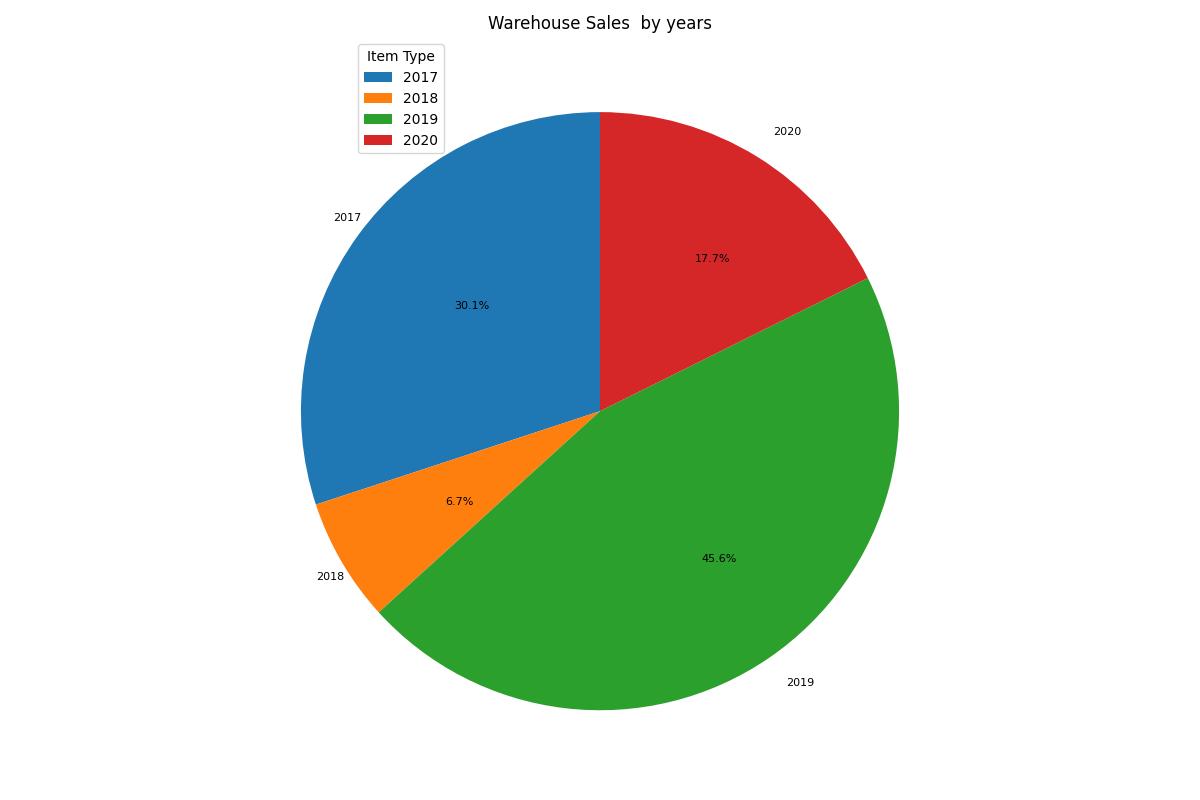
**Retail Transfers By Years**

In **2019**, we transferred **44.9%** of our products from the **Warehouse**, while in **2018** we did not require extra stock, only transferring **7.2%** compared to **2019**.

****

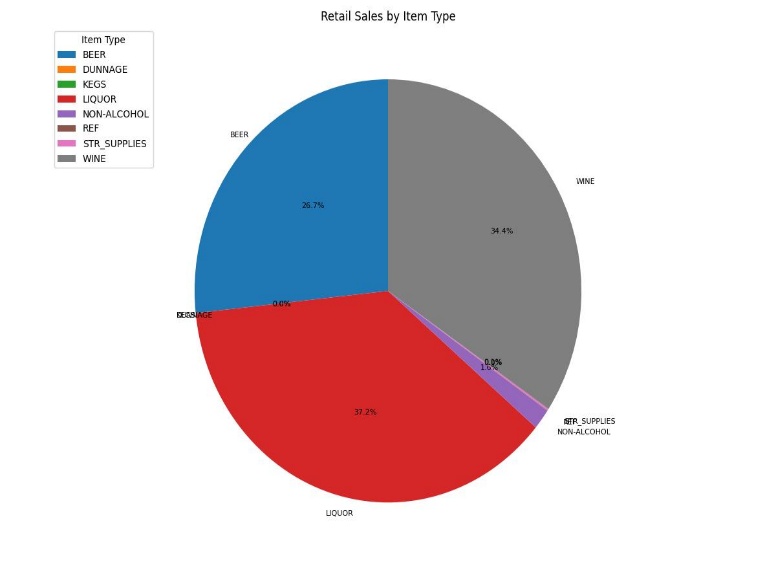
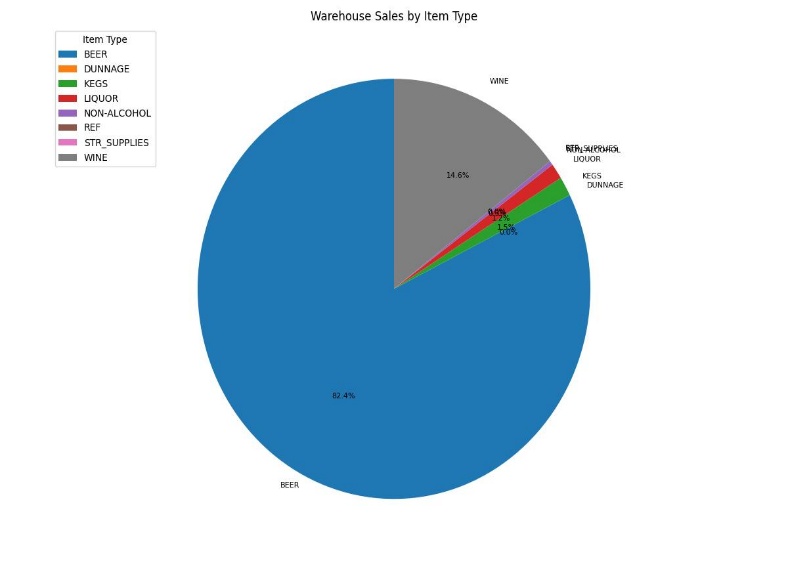
**Warehouse dispatch Units By Years**

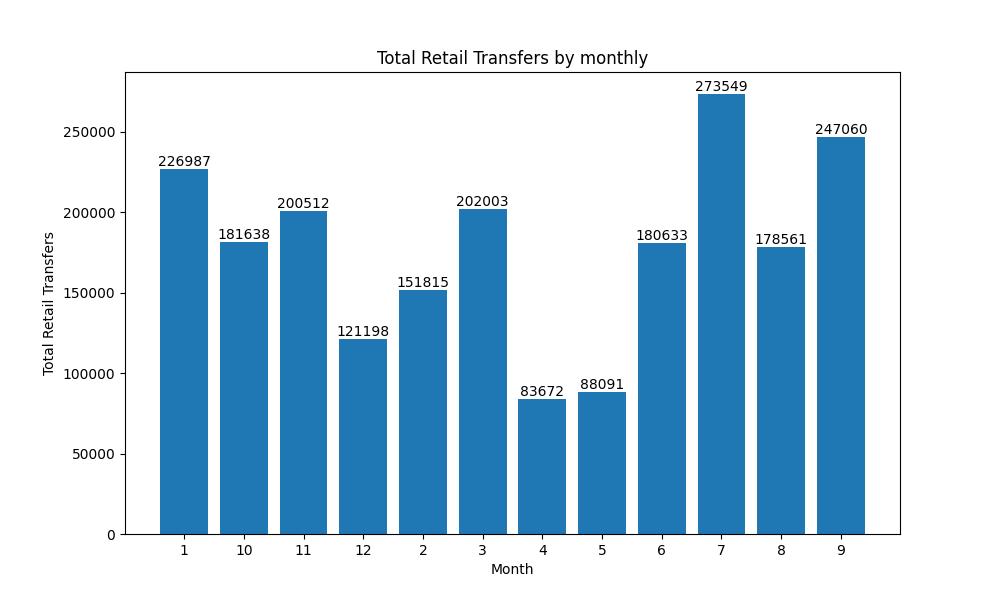
The **Warehouse** and **Retail Transfers** have the same units of products, but there are slight differences between them.

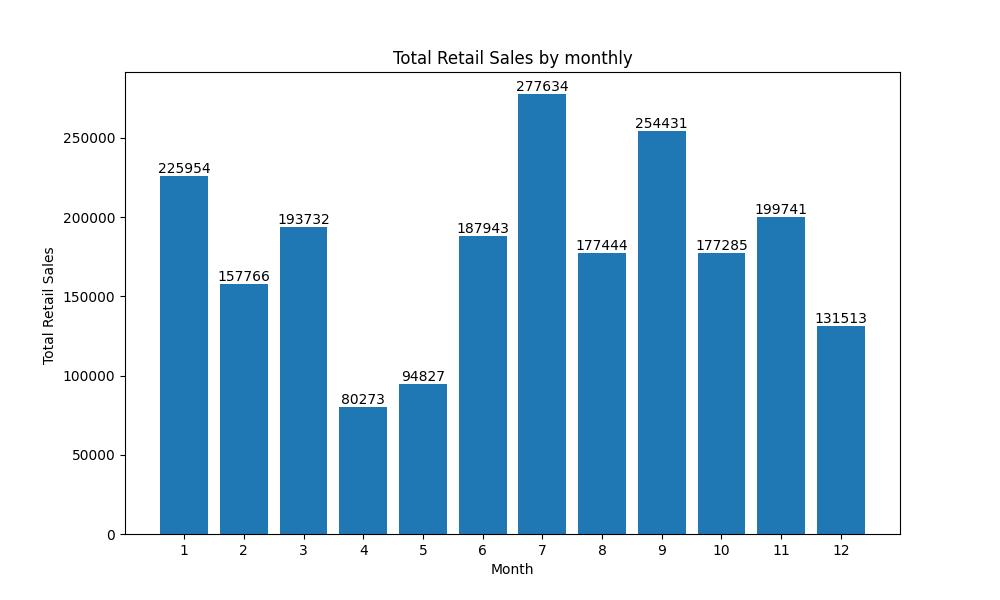
****

**Warehouse Sales By Years**

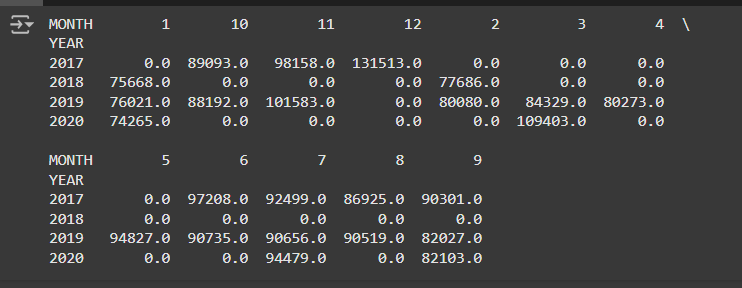
**Warehouses** have the right to sell directly, but customers must buy products in bulk. In **2019**, **45.9%** of warehouse units were sold directly, which is an increase compared to the units sold in the past three years.

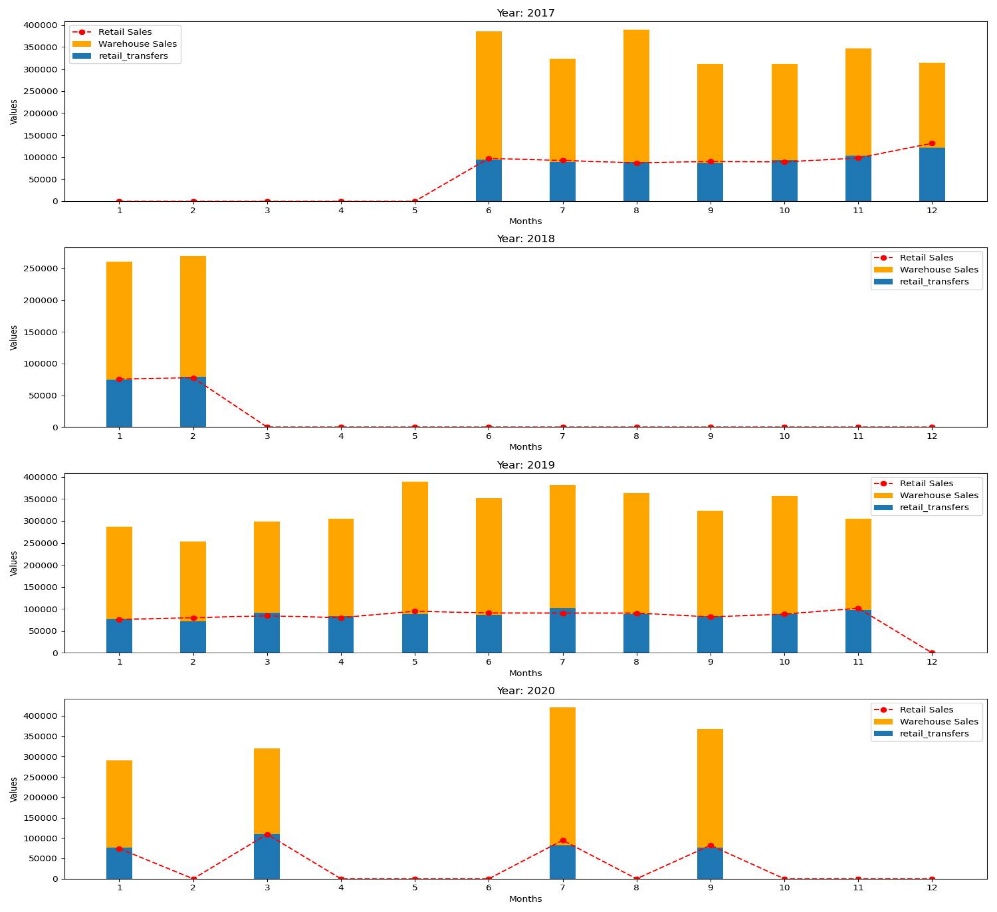
****

****When comparing **Retail Shops** and **Warehouses**, we found that most of the beer sold out from the warehouse. Customers were purchasing beer directly from the warehouse, whereas wine had a different trend. More customers were buying wine from **Retail Shops**, along with liquor.

****

This data on **Sales** and **Stock Transfers** in **2019** shows how much stock has been transferred to retail shops each month.

****



1. The graphs display **Retail Sales, Warehouse Sales, and Retail Transfers** from **2017** to **2020**, organized by month.

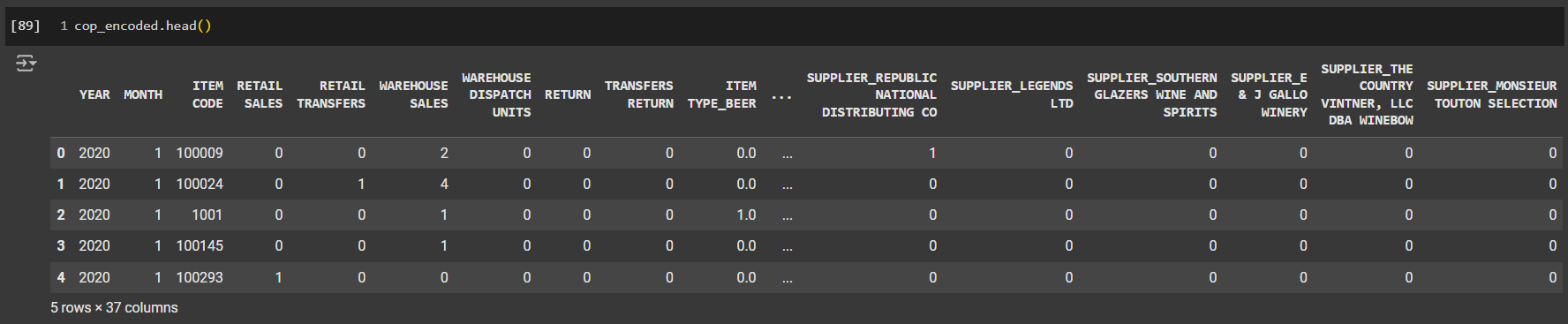
2. Warehouse Sales consistently dominate the figures for most months, while **Retail Transfers** contribute smaller but steady amounts.

3. **Retail Sales** are generally low; however, noticeable surges occur in certain months, particularly in **2017, 2019, and 2020**.

4. In some years, such as **2018**, there are significant gaps in most months, indicating missing or minimal sales data.

**Model Development**

**Feature Engineering**

* + **Machine Learning Models Used**
    - Before inserting into the model, I made changes and fine-tuned the data by using feature engineering techniques like **one-hot encoding** and **one-hot encoding with multiple categorical** variables for columns such as **BRAND NAME** and **SUPPLIER**.
    - Next, use a dimensionality reduction technique to obtain the explained variance ratio: [9.99996546e-01 3.29623296e-06 8.25174599e-08].
  + **Algorithms Selection Process**
* I chose the **K-Means Clustering model** because it efficiently handles large datasets, making it suitable for real-world problems. K-Means is a widely used and well-studied algorithm that is easy to understand, implement, and computationally efficient.
* **Other model Need to use now.**
  + **Model Evaluation Metrics**

For model evaluation, I used the **Calinski**-**Harabasz** Score and the **Davies**-**Bouldin** **Index**, both of which indicate that higher scores are better. A higher **Calinski**-**Harabasz** Score suggests that the clusters are well-defined, meaning they are dense and well-separated from one another. Conversely, for the **Davies**-**Bouldin** **Index**, a lower score indicates better clustering quality.

These metrics provide quantitative measures to assess how effectively your K-Means algorithm has clustered the data based on a specific K value. By comparing the scores (with a higher **Calinski**-**Harabasz** Score and a lower **Davies**-**Bouldin** **Index**) across different K values, I was able to determine the optimal number of clusters for the dataset. This process ensures that the resulting clusters are meaningful and represent distinct groups within the customer data.

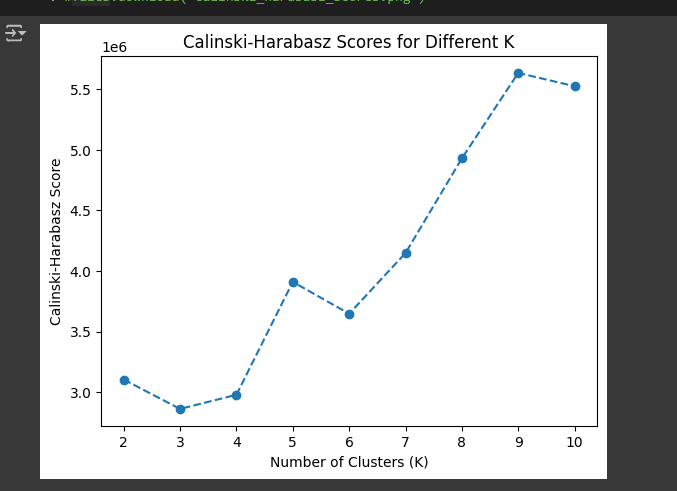
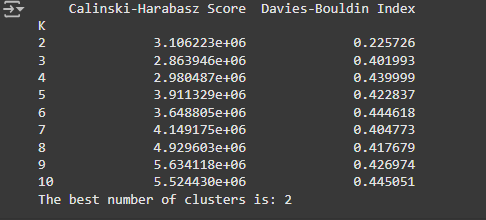
* **Hyperparameter Tuning:**

Techniques used (Grid Search, Random Search, Bayesian Optimization, etc.)

**Model Evaluation**

Performance of Model

The Best number we found is 2 for K-Means cluster.



**Deployment**

* Deployment tools (Streamlit, Flask, FastAPI, etc.)
* Steps taken for deployment
* Link to deployed model (if available)

**Challenges and Solutions**

I received a dataset that lacked domain knowledge, which made it challenging to work with. I tried several strategies to approach the last observation, but I got mixed up with some techniques and became entangled in them.

The most problematic column was the description, as it contained a lot of information that confused me during the data cleaning and Pre-Processing stages. I then applied correlation techniques and used Pearson correlation.

**Conclusion**

* + **Summary**

The clustering analysis optimized inventory management, identified sales trends, and provided actionable insights for tailored product offerings, helping the company improve operational efficiency, address high return rates, and make data-driven decisions to maintain a competitive edge.

* + **Future Work or Recommendations**

Future work involves enhancing data quality, adopting advanced clustering techniques, enabling real-time analysis, integrating predictive analytics, and focusing on customer-centric insights. Additionally, it will include deploying models into operational systems, continuously refining clusters, and taking external market factors into account to improve inventory management and strategic decision-making.