Ekene Michael Ahuche
Personal EDA Project

EDA Project – Sales Analysis (Data Analysis with Python)

Project Overview:

Sales Analysis

- Key Findings:
- - Top Supplier: "REPUBLIC NATIONAL DISTRIBUTING CO".
- - Negative Sales: Found 716 entries with negative warehouse sales (likely returns).
- - Conclusions: Wine is the top-selling category accounting for 61.0% and Liquor comes second at 21.8%.
- - From the monthly sales data, we can infer that:
- The highest single-month sale appears in July 2020 with WareHouse Sales accounting for 70.2% of total sales.
- However, when comparing aggregate totals, the majority of the sales occurred in 2019.
- Overall Total Sales**:**
- -Retail Sales: $2,160,899.37
- -Retail Transfers: $2,133,968.63
- -Warehouse Sales: $7,781,756.28
- Python and Pandas Library were used to read and manipulate the csv dataset, While Seaborn and Matplotlib were used for plotting graphs, Pie Charts, plots and other forms of visualizations.

Data Overview:

The dataset tracks sales and inventory transactions for beverages(liquor, wine, beer) and related supplies across retail and warehouse channels. It helps analyze sales performance, supplier relationship and inventory adjustment. The dataset was sourced from data.gov with the name '**Warehouse_and_Retail_Sales.csv**'. The dataset possesses 307,645 rows and 9 key columns(Year, Month, Supplier, Item Code, Item Description, Item Type, Retail Sales, Retail Transfers, WareHouse Sales).

Code snippets like `data.head()`, `data.info()`, `data.shape` and `data.describe()` were used for initial data inspection.

Data Cleaning and Transformation:

For this dataset, the Item Type column had 1 missing value, Retail Sales had 3 while SUPPLIER column had missing values of 167 (empty strings) which were safely

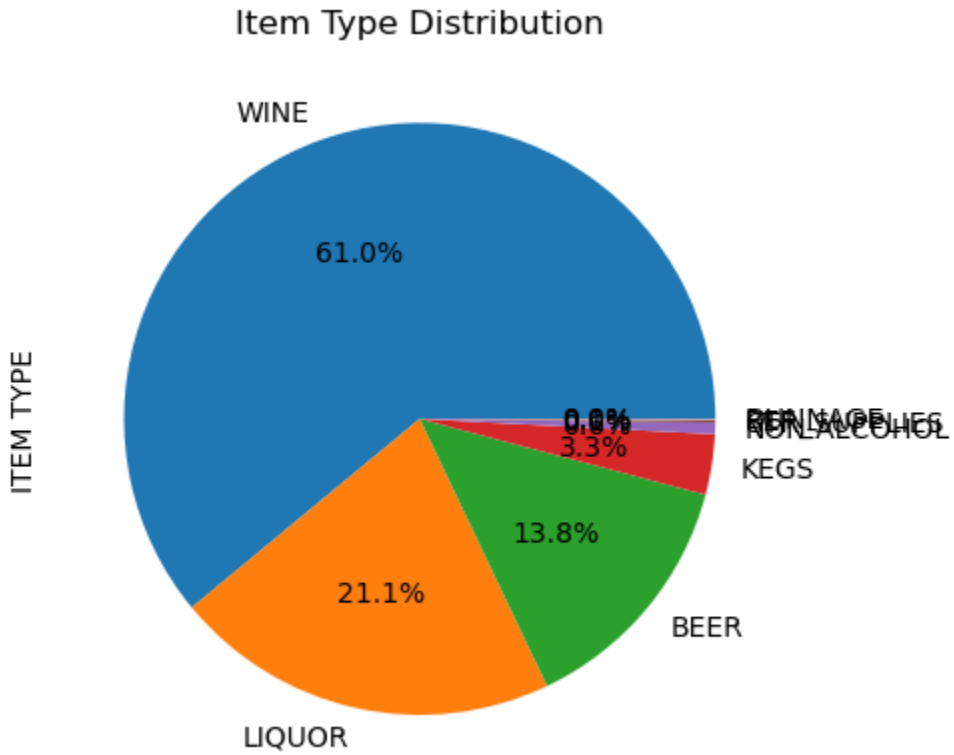
replaced with “Unknown’” without losing any other valuable data. Code snippets like `data['SUPPLIER'].replace('', 'Unknown')`) were used to perform the replacement. There seemed to be 0 duplicate rows in the dataset.

In the dataset, there are 716 records where the warehouse sales value is less than zero, meaning that they were negative. This likely means that these negative values are probably recording returns, cancellations, or corrections, because the items associated with negative entries include descriptions like "EMPTY WINE KEG" and "EMPTY 1/2 KEG," which might suggest that these entries are related to returns or adjustments for container returns or similar corrections.

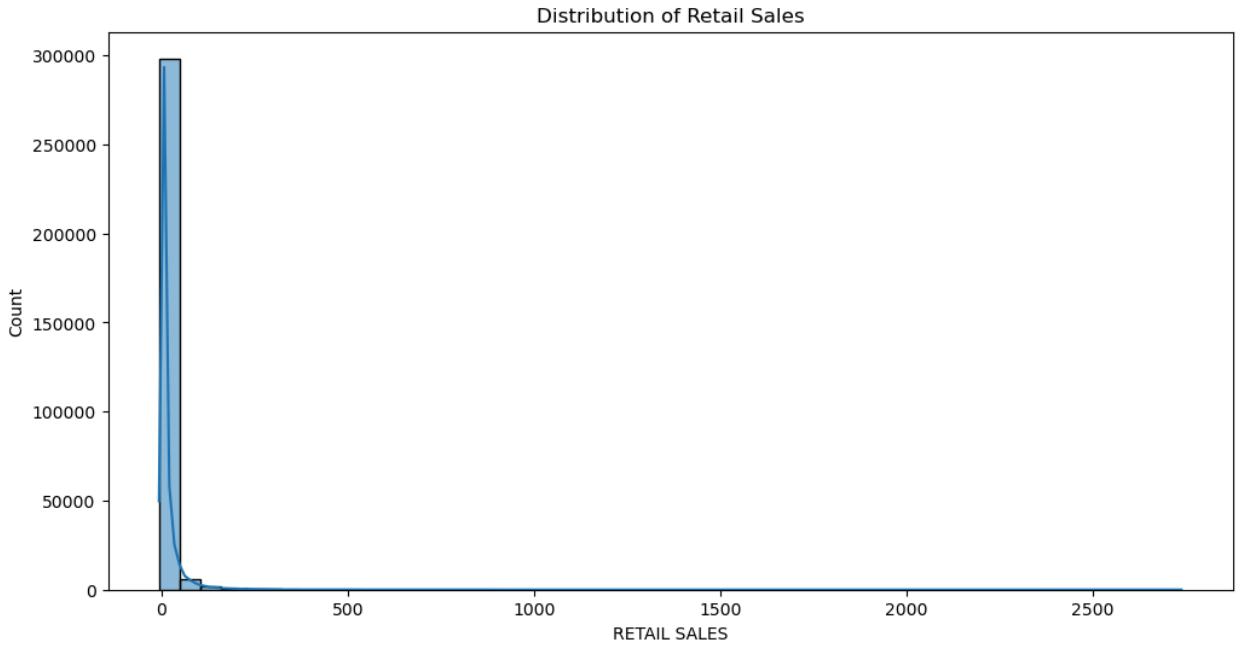
EDA Questions and Answers:

Question	Code Snippet	Purpose
What does the dataset look like?	data.head(), data.info(), data.describe(), data.shape	Initial inspection of data structure, types, and summary stats.
Are there missing values?	data['SUPPLIER'].replace('', 'Unknown')	Fix empty supplier names.
Are there negative sales?	data[data['WAREHOUSE SALES'] < 0]	Identify returns/adjustments (e.g., -12.00 for empty kegs).
Who are the top suppliers?	data['SUPPLIER'].value_counts().head(10)	Find most frequent suppliers.
What are the most common item types?	data['ITEM TYPE'].value_counts()	See distribution of LIQUOR, WINE, BEER.
What are total sales by category?	data[['RETAIL SALES', 'RETAIL TRANSFERS', 'WAREHOUSE SALES']].sum()	Calculate total sales for retail, transfers, and warehouse.
How do sales vary by month?	data.groupby(['YEAR', 'MONTH'])[sales_columns].sum()	Analyze monthly trends (from June 2017 to Sep 2020 in the data).
What’s the distribution of retail sales?	sns.histplot(df['RETAIL SALES'], bins=50, kde=True)	Visualize sales frequency (most are small, a few large).
What’s the correlation between sales metrics?	sns.heatmap(df[sales_columns].corr(), annot=True)	Check relationships (e.g., retail vs. warehouse sales).

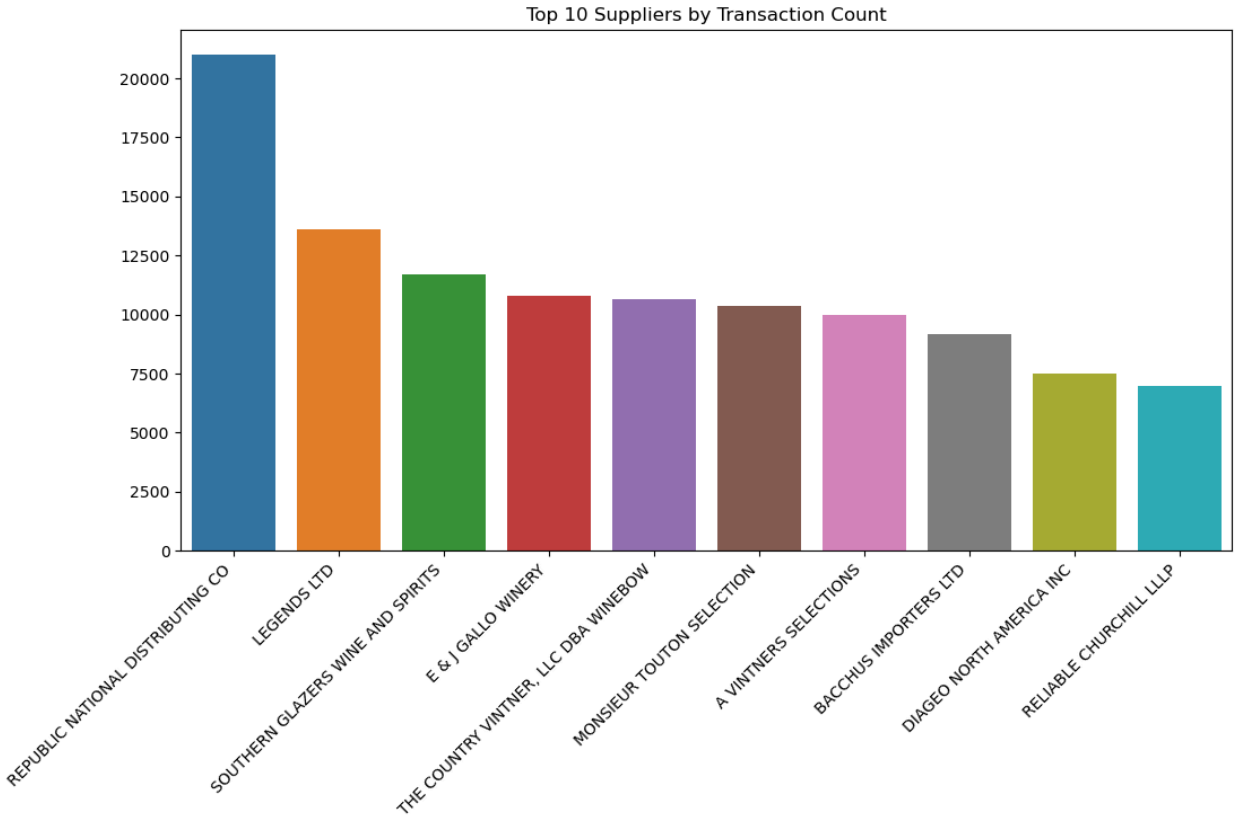
Visualizations:



- This Pie Chart visualizes the proportion of WIne, Liquor, Beer, Kegs and other small categories. Wine is the top-selling category accounting for 61.0%, Liquor comes in second with 21.8%, Beer accounts for 13.8% while Kegs at 3.3% and other Item type categories at 0.0%. **note: the very small categories (e.g., NON-ALCOHOL, DUNNAGE) have such a low percentage that when rounded, they might display as 0.0% (or very close to 0).**

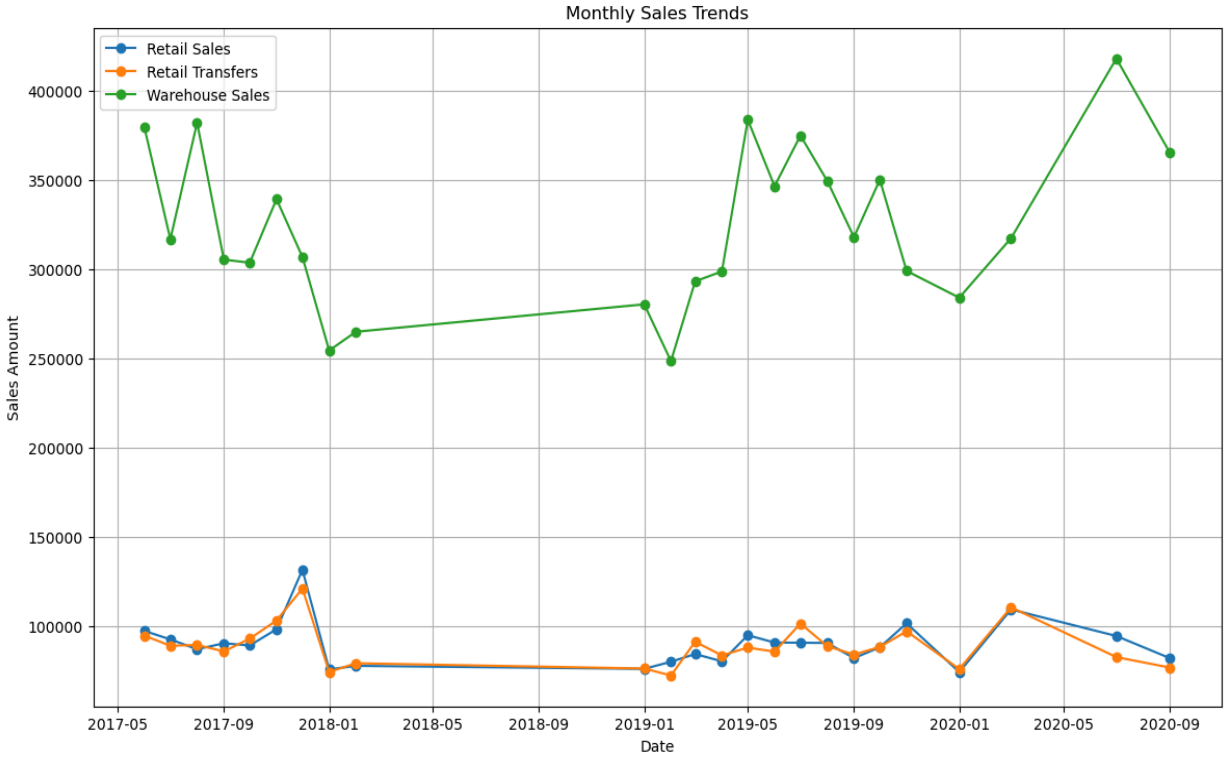


- The Histogram above shows that Retail Sales are heavily right-skewed (most transactions are small, with a few large sales). The right-skewed nature of the Retail Sales distribution implies that while most sales are on the lower side, there are occasional months or transactions that result in very high sales. This is typical in many real-world sales datasets like this one where many transactions are small, but a few are exceptionally large—possibly due to bulk orders or seasonal surges.

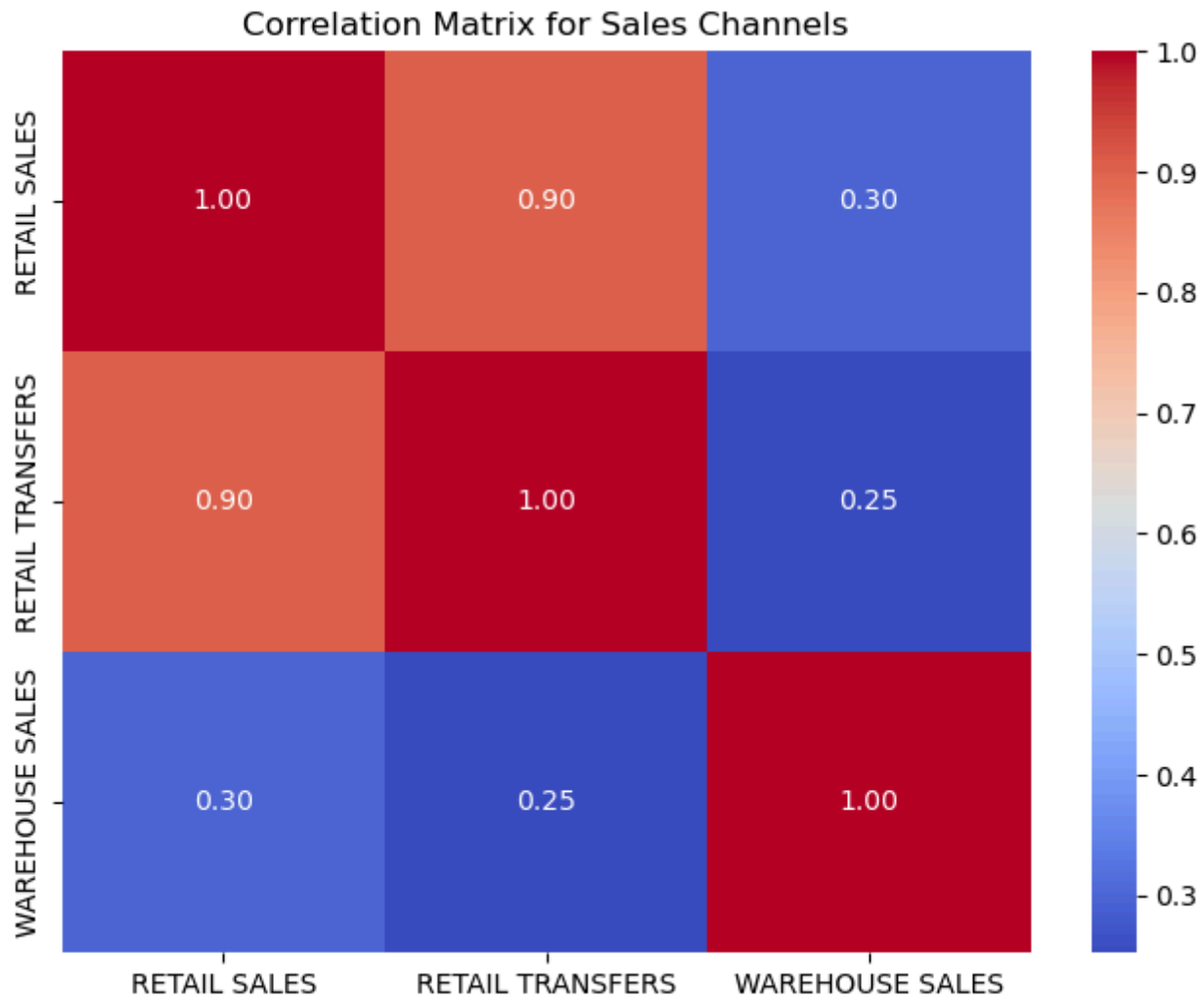


- The Bar Chart above shows suppliers who have the most transactions with the top 5 suppliers being: REPUBLIC NATIONAL DISTRIBUTING CO, LEGENDS LTD, SOUTHERN GLAZERS WINE AND SPIRITS, E & J GALLO WINERY, THE COUNTRY VINTNER.

  Item Types: Dominated by Wine (61.0%), Liquor (21.1%), and BEER (13.8%).



- Plotting the line chart of retail sales, retail transfers, and warehouse sales over the years is to clearly visualize and show trends, spikes, and seasonal patterns in the dataset. For example we can infer from the line graph that in every month, warehouse sales are much higher than both retail sales and retail transfers and for most months, the values for retail sales and retail transfers are quite similar. This could imply that the transfers perhaps representing inventory movements or internal reallocations closely mirror actual retail transactions. There are also clear seasonal spikes for retail sales in December 2017 and March 2020, while warehouse sales peaked in July 2020.



- The heat map above shows a high value between Retail Sales and Retail Transfers, it confirms that these two aspects of the sales operations are very closely related. However, there's a Weak correlation between RETAIL SALES and WAREHOUSE SALES.

Insights and Next Steps:

Key Findings

- **Negative Warehouse Sales:**
    - The dataset contains 716 negative entries in the "WAREHOUSE SALES" column. These likely represent returns, refunds, or inventory adjustments.

- **Top Suppliers:**
    - Analysis of supplier frequencies reveals a set of top suppliers that consistently appear, indicating their key role in the supply chain.
- **High-Selling Categories:**
    - The data shows that wine and liquor are among the most dominant product categories, contributing significantly to overall sales.
- **Sales Distribution and Data Consistency:**
    - The histogram of Retail Sales reveals a heavily right-skewed distribution—most transactions are small, but a few are very large.
    - Retail Sales and Retail Transfers are nearly equal in magnitude, which could imply that the data is consistently recorded or that these two categories represent related aspects of the same overall sales process (e.g., direct sales vs. internal transfers).

Business Implications

- **Operational Focus:**
    - With Warehouse Sales being the largest contributor, the company might be more focused on bulk or wholesale operations. This suggests that a significant portion of business operations is driven by large-scale warehouse transactions.
- **Resource Allocation:**
    - Since warehouse operations generate much more revenue, management may need to allocate more resources or monitor warehouse performance more closely. This might include investing in logistics, inventory management, or process improvements in warehouse operations.
- **Data Consistency:**
    - The near-equality between Retail Sales and Retail Transfers suggests a high level of data consistency. It might also indicate that these two metrics are closely related—possibly representing different facets of the same sales process (for example, direct sales versus internal inventory reallocations).

Recommendations & Next Steps

1. **Investigate Negative Warehouse Sales:**
    - **Action:** Dive deeper into the 716 negative entries. Determine whether they are due to returns, refunds, or inventory adjustments.
    - **Next Step:** Collaborate with operations or finance teams to verify the reason behind these adjustments and understand their impact on net sales and profitability.
2. **Focus on Top Suppliers for Inventory Optimization:**

    - **Action:** Analyze the performance of your top suppliers to understand their reliability and impact on overall sales.
    - **Next Step:** Evaluate opportunities to negotiate better terms, optimize ordering patterns, and reduce holding costs by focusing on high-performing suppliers.
3. **Explore Promotional Strategies for High-Selling Categories (Wine and Liquor):**
    - **Action:** Since wine and liquor drive a substantial portion of revenue, consider targeted marketing and promotional strategies for these categories.
    - **Next Step:** Design and test promotional campaigns (e.g., bundle deals, seasonal discounts) to boost sales further, particularly retail sales and attract more customers in this and other high-revenue areas like warehouse sales.