FHS

Sales Analysis of Filtration Company

Data Analysis and Visualization

Emmanuel Masindi
6-8-2023

Contents

INTRODUCTION

This document will serve as a summary to the project undertaken on the Jupyter Notebook file '20230609 ASSESSMENT'.

In this project we have been hired as a Data Analyst for a company that sells filtration devices. The aim is to gain insights into customer behaviour, identify target customer segments, provide recommendation for improving sales and customer satisfaction and optimize the business's data model for improved efficiency.

This project will be completed using Python and Jupyter Notebook. We will focus on the datasets given where we have SALES, INVENTORY and CUSTOMER tables. We will start by cleaning data in each of these tables, then merge them using appropriate columns. This Notebook will detail the Analysis and the thought process behind every visualization. It will also include discussions of insights and possible recommendations

The Notebook will conclude with a discussion about the given data model and the recommendations to improve its efficiency.

At the conclusion of this project, to improve sales, recommendations were identified as follows:

- The business should look to maximize the peak in sales during the months of July, August, September and October. Marketing should be intensified before and during these months.
- Mining is showing a steady increase in sales. Efforts need to put in place to consolidate this customer base.
- Products recorded as 'Other' have not recovered to pre-pandemic levels and are mostly to blame for why we don't see an even higher increase in overall sales. These need to be identified and marketing addressed.
- Customers recorded as 'Other' generate the most sales and they seem to be interested in filters for vehicles and hardware equipment. These are probably walk in individuals looking to buy service parts for vehicles. Efforts to market directly to them in stores should be adopted.

PART 1: DATA CLEANING

SALES TABLE

First impressions:

- Null values in most columns
- Date columns are in the incorrect data types
- We may struggle to understand the meaning in some column names

Dealing with missing values:

Null values will be handled by replacing them with appropriate placeholders depending on data type and column description.

- column 'NatureofBus' replace with 'other'
- column 'Country' replace with 'other'. We could try to look for clues in other columns but data in this category is minimal
- column 'Description_1' replace with 'unknown'.
- column 'ItemClass' replace with 'other'
- column 'PublicGPPer' replace with 0
- column 'repdivision' replace with 'not_specified'
- column 'ExtOrderNum' replace with 'unknown'
- column 'Brand' replace with 'other'
- column 'WeightValue' replace with 0
- column 'udIILastGRVDate' replace with 'unknown'
- column 'ucIDSOrdCreationAgent' replace with 'unknown'
- column 'Order_No' replace with 'unknown'
- column 'warehousecode' replace with 'unknown'
- column 'warehousename' replace with 'unknown'

We will remove columns which are duplicated on other tables and which are irrelevant to this analysis. We will later merge all tables, we don't want to deal with duplicated columns. Columns to be excluded:

- Customer_name
- Description
- ItemGroup
- ItemGroupname
- Period
- FinPeriod
- RepName
- repdivision
- WeightValue

INVENTORY TABLE

We will assess dataset to ensure it only has unique products for each line item

Dealing with missing values:

- column 'Code' replace with 'no code'
- column 'Description_1' replace with 'unknown'

CUSTOMER TABLE

We will keep columns which are relevant to the analysis

Dealing with missing values:

- column 'Contact_Person' replace with 'unknown'
- column 'Physical1' replace with 'unknown'
- column 'DCBalance' replace with 0.
- column 'fForeignBalance' replace with 0
- column 'ulARCountry' replace with 'unknown'
- column 'ulARProvinces' replace with 'unknown'

PART 2: DATA INTEGRATION

We used the pandas pd.merge function for this section of the project.

We use a left join to merge the inventory table to the sales table (sales table on the left). Then we merge the resultant table with the customer table (with the merged table on the left).

We can confirm that data Integrity has been maintained. The total number of rows of the resultant merged dataset is equal to the total row count of the sales table. Only unique keys were used to join the tables; that is we ensured the joining keys are indeed the primary keys for the table on the right of the join.

PART 3: EXPLORATORY DATA ANALYSIS

Statistical Summary

Categorical columns of interest

- NatureofBus
- Country
- ItemClass
- slsQUARTER
- warehousecode
- ulARProvinces
- Brand
- ucIDSOrdCreationAgent
- Contact_Person'
- SlsYear

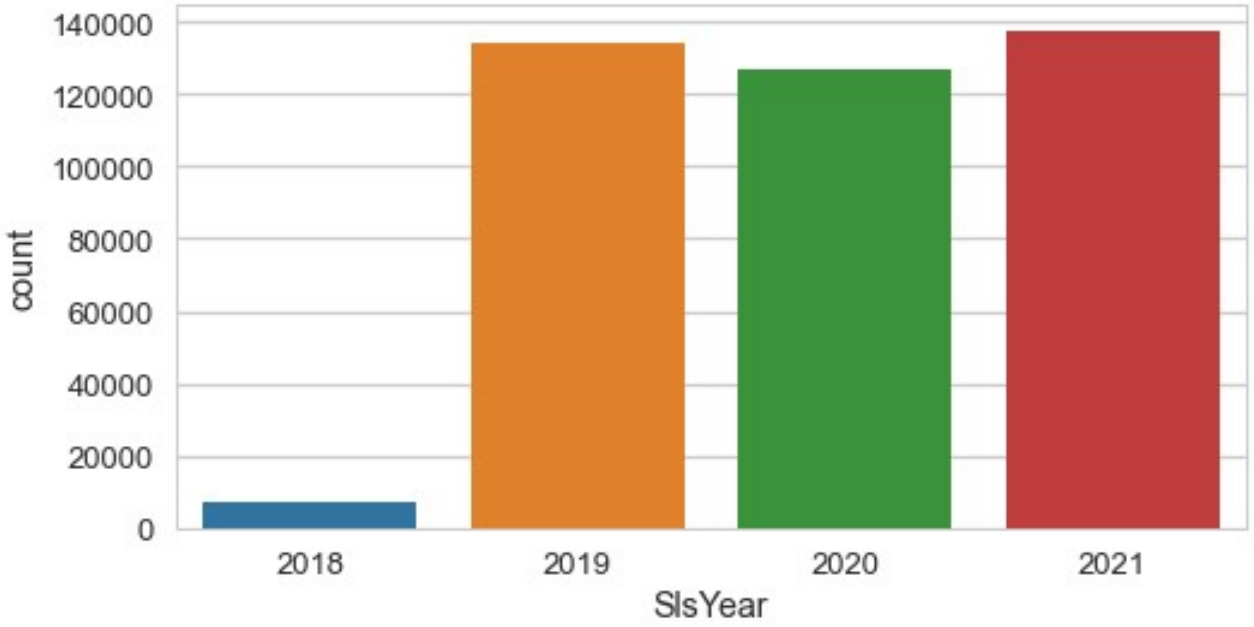


Figure 1: Year Distribution

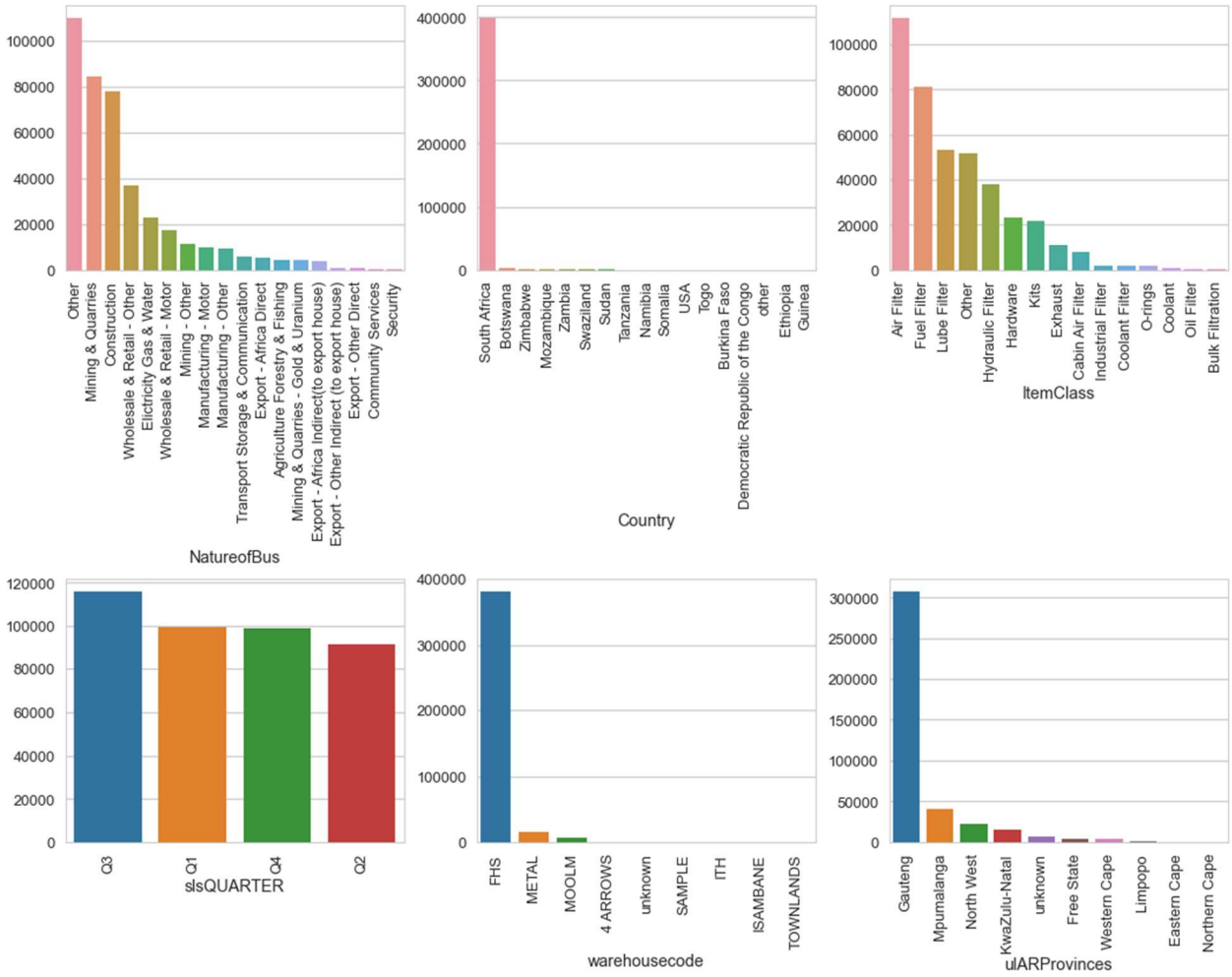


Figure 2: Categorical features distribution

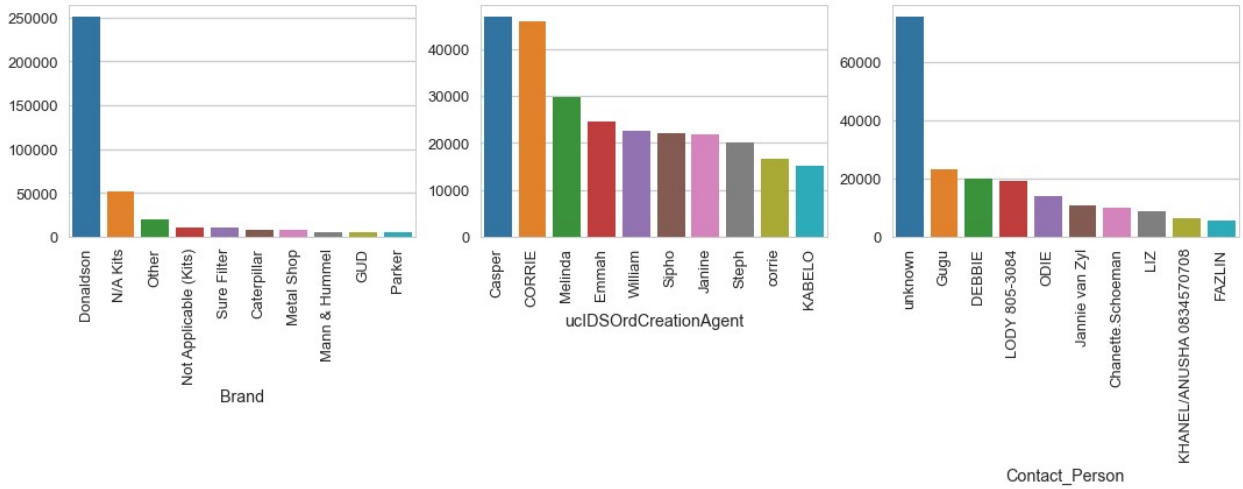


Figure 3: Categorical features distribution

Numerical Columns of interest

- UnitPrice
- DCBalance

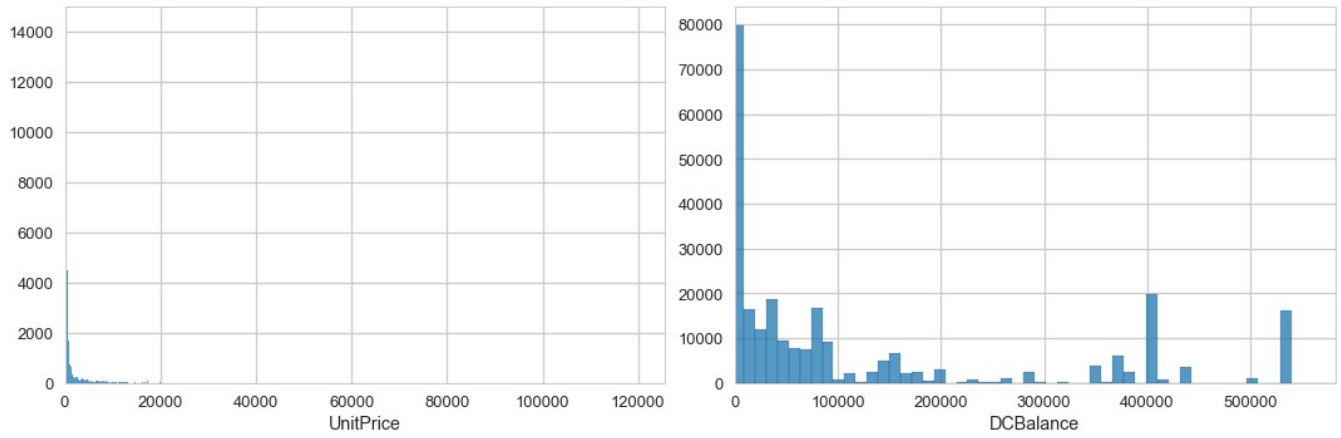


Figure 4:

Histograms for Numerical columns

Discussion:

- It seems most customers purchase items with a unit price of up to R20,000
- The histplot on the left suggests that the business has sold items even significantly higher than R120,000 even though these purchases are few and far between
- The histplot on the right suggests that the business sells items to customers having a widely distributed DC Balance. The data is fairly skewed to the left. We should assess later the correlation between DC Balance and Purchase price.

Customer Sales distribution by Date

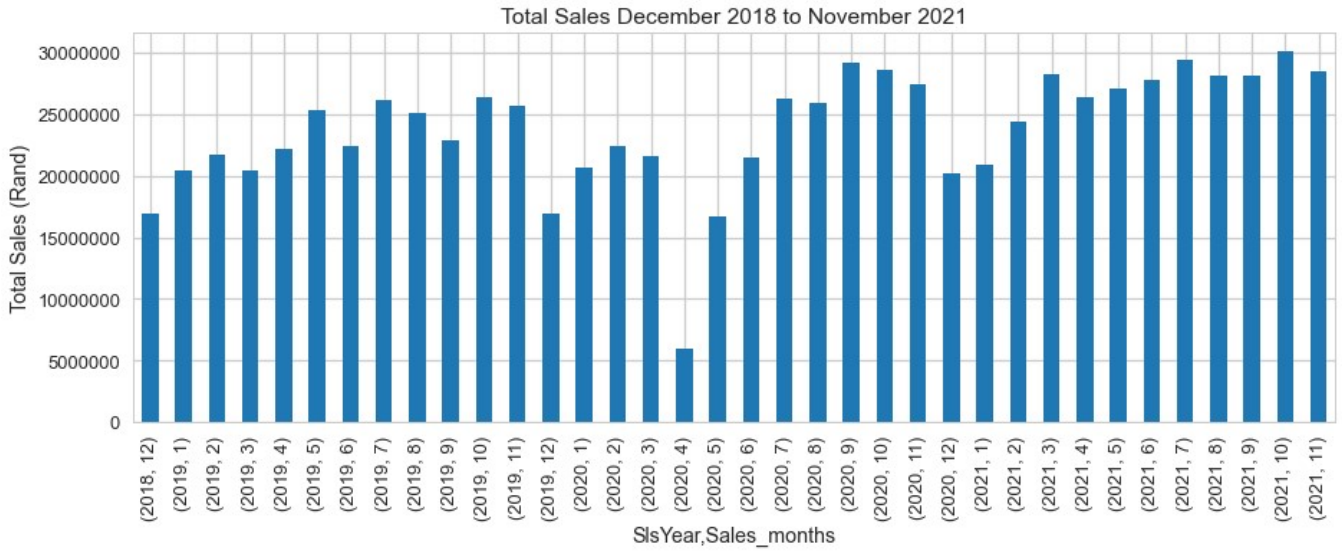
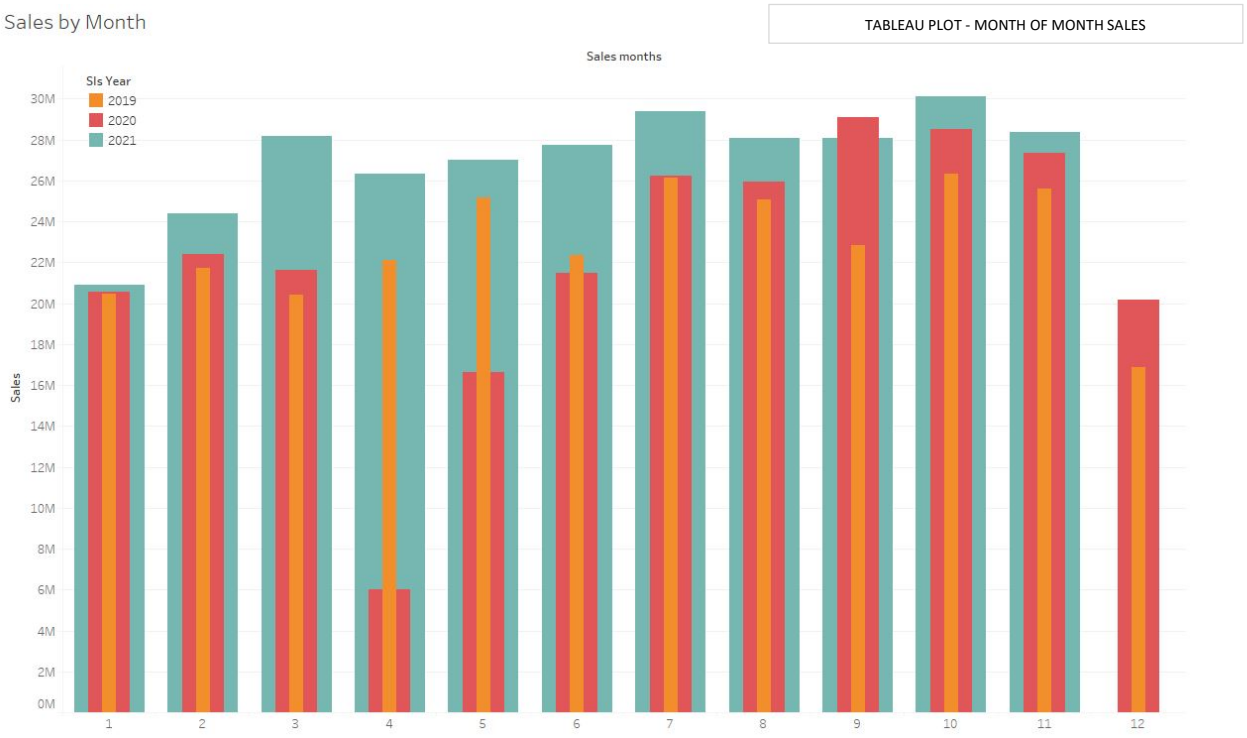


Figure 4: Total Sales December 2018 to November 2021

Discussion:

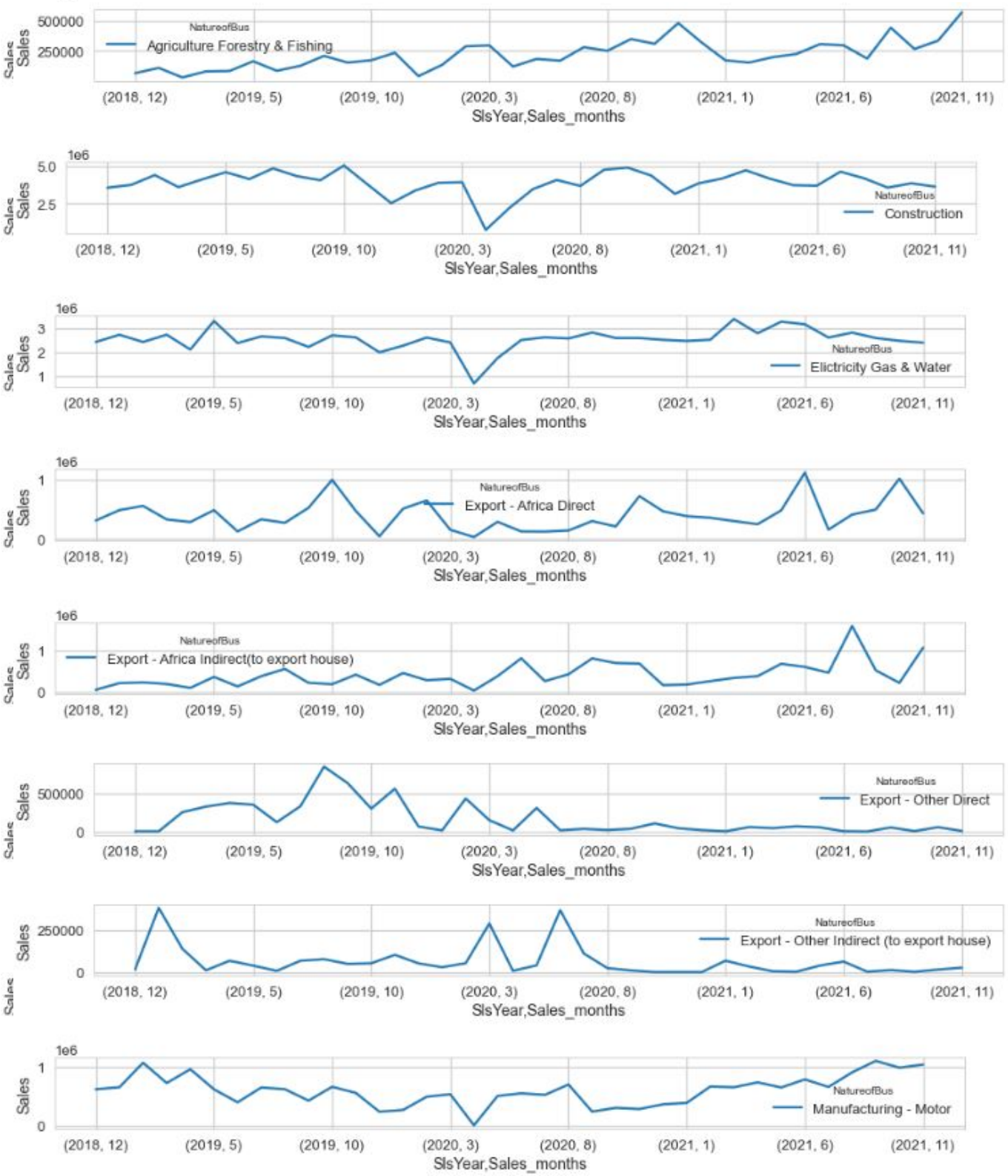
- Sales Peak between July and October year on year.
- December and January record the lowest sales for the year due to obvious reasons.
- The significant deep on April 2020 would most likely been caused by the start of the lockdown period.

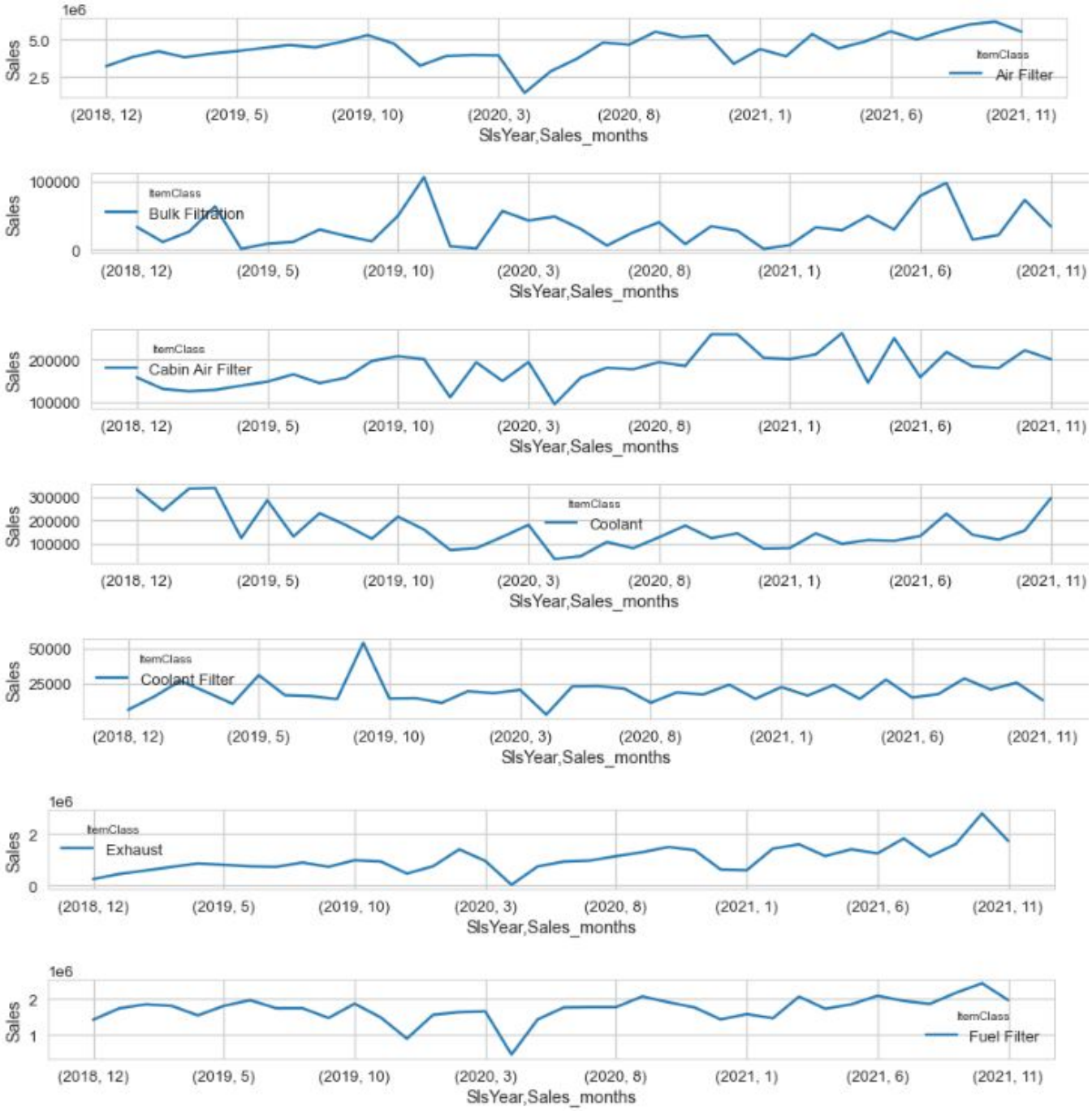
To analyze sales by customers, we need a suitable categorical feature with which to link customers who have a similar feature such as their location, or their gender, or age. The dataset does not breakdown our customers into categories of interest. However we have a column under the SALES table specifying the nature of business the customer is involved in. We will use this in our analysis.



There are a total of 19 categories under this column. We will focus on the customers who lie within the top 15 of these.

Figure 5: Total Sales December 2018 to November 2021



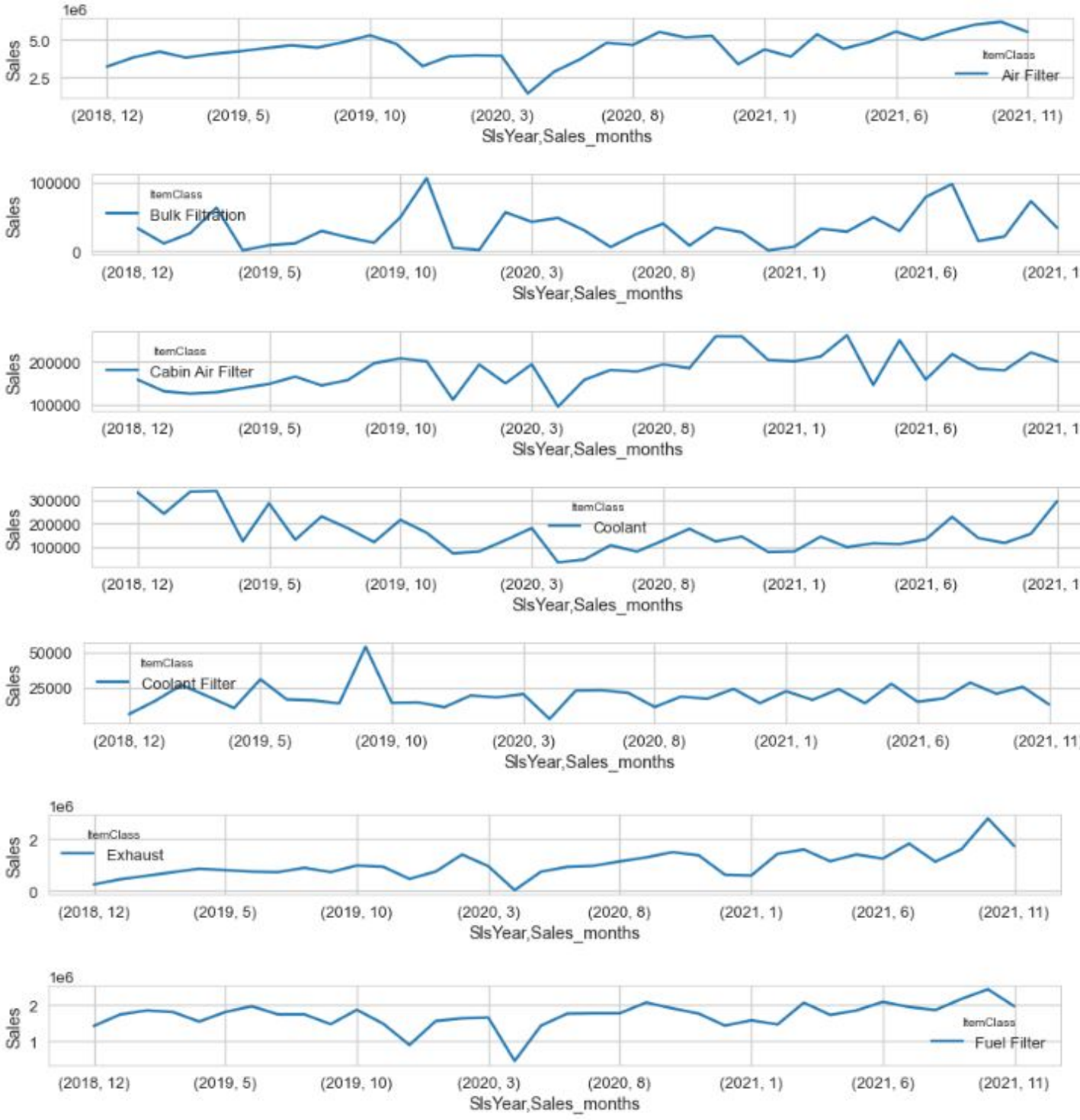


Discussion:

- All customer segments (excluding Export) recorded sales consistently over the time period.
- The customer segment recorded as 'Other' is showing an increase in sales over the entire time period, potentially showing signs of even more growth potential in the sector. It would be advantageous to have an actual description of the type of work involved in this segment rather than having it labelled 'Other'.
- The same can be said for Mining. The market is showing steady signs of growth.
- Sales from Export customers (Exluding Africa) are low and have not recovered to pre-lockdown levels.

Sales by Product

To analyze this, we will repeat the exercise from above with the only difference being that each graph will now represent a different product category. More specifically the Item Class.



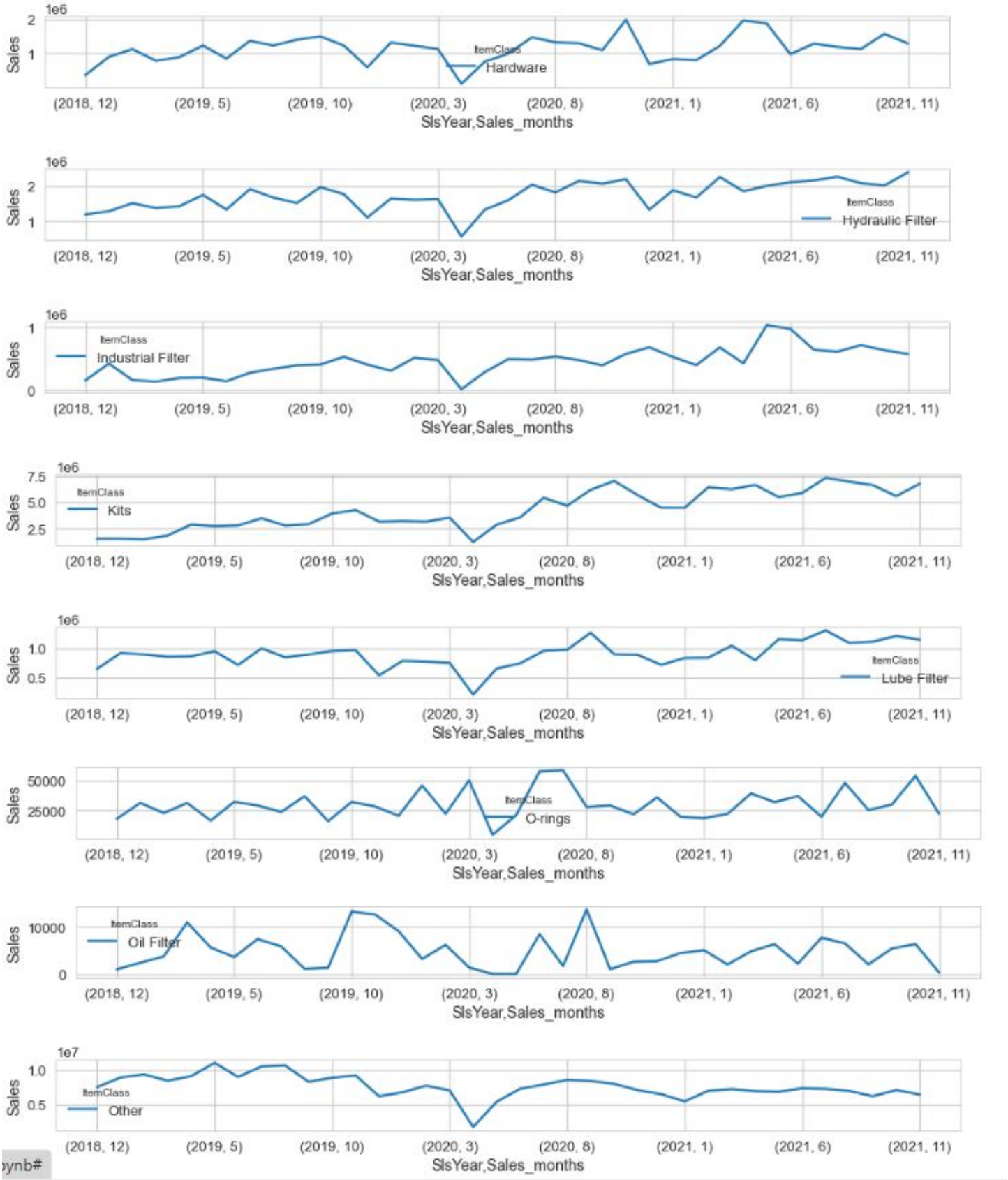


Figure 7: SALES BY PRODUCT

We will plot the above line graphs on the same axis to better analyze how each product's sales compares with the rest.

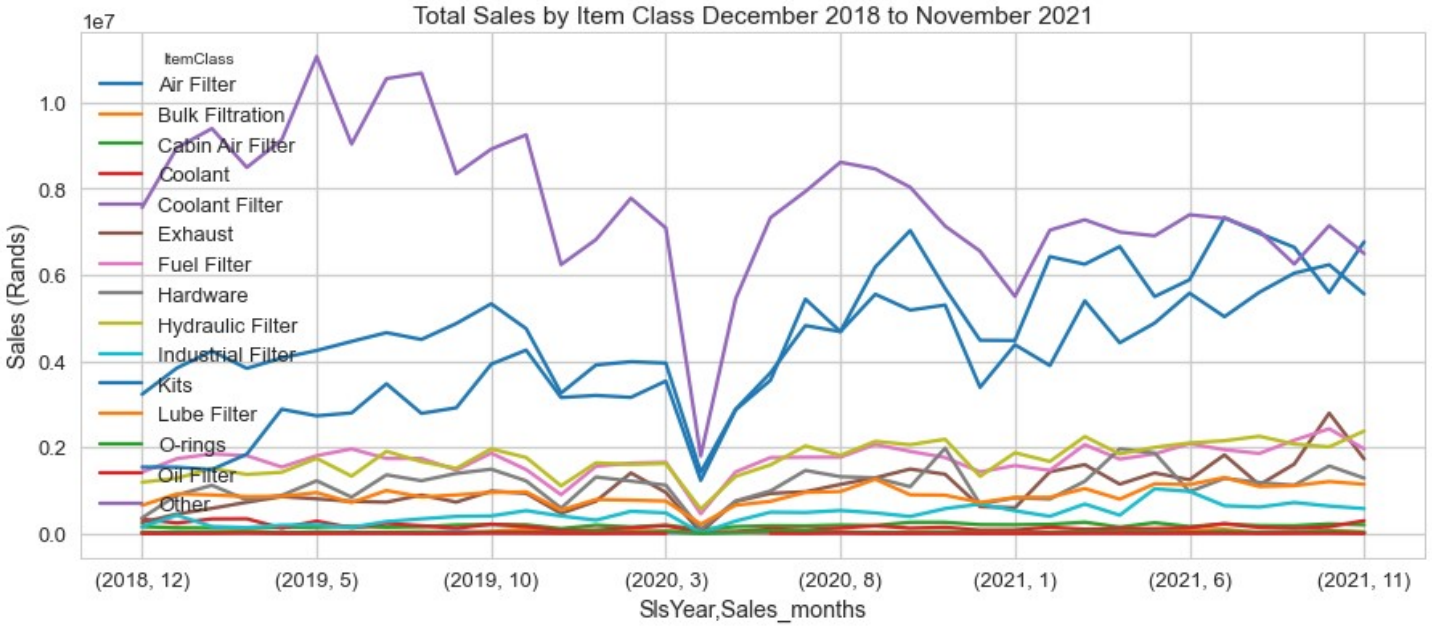


Figure 8: SALES BY PRODUCTS:

Discussion:

- Air, Lube, Fuel, Hydraulic, Industrial filters have shown an upward growth trajectory. This is also true for Exhausts, Kits and Hardware.
- Items currently being recorded as 'Other' have shown domination however they have not been able to recover to pre-pandemic levels.
- All other Items are showing a general recovery to pre-pandemic levels and are also improving.
- The decrease of Sales of items currently recorded as 'Other' is the main reason why we do not see a higher increase in overall sales in the business for post pandemic periods.

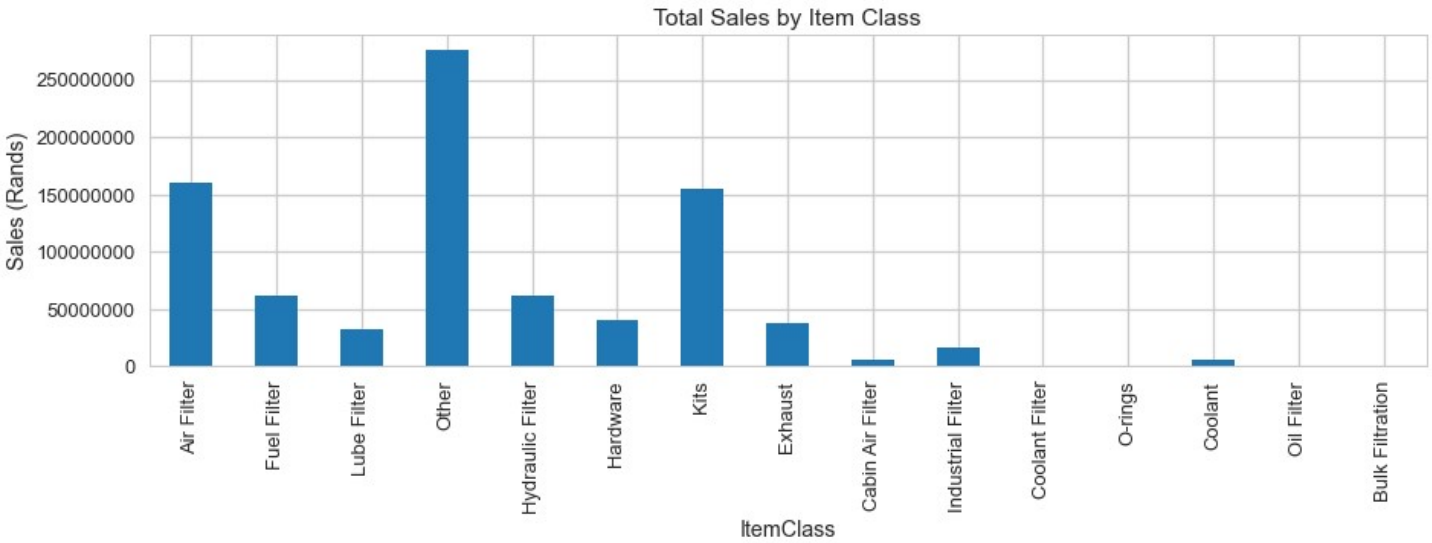


Figure 9: Total Sales by Item Class

Next, we will analyze our product sales and see how our different customers mainly shop in the business.

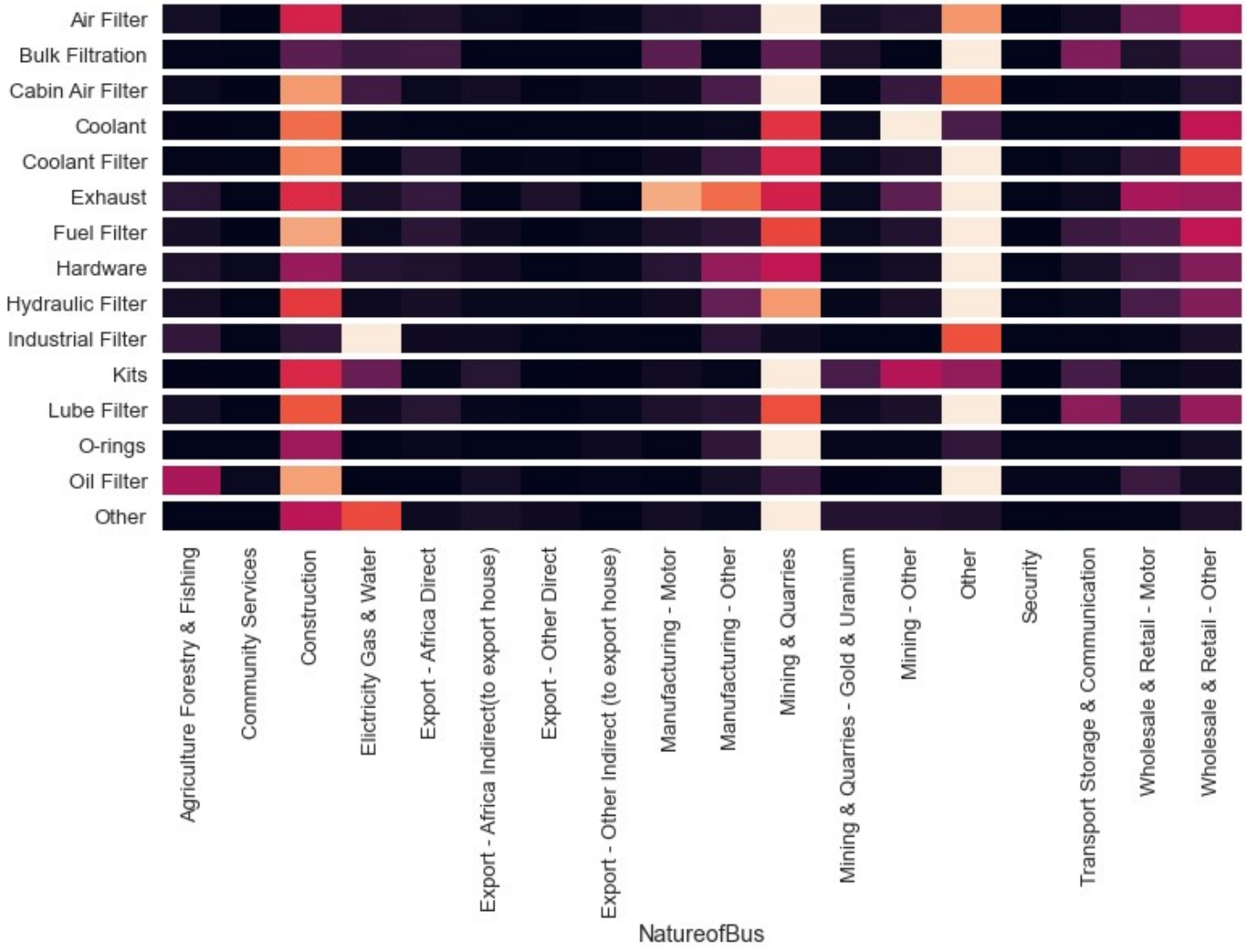


Figure 10: Customers vs Product

Discussion:

- Items recorded as 'Other' (these represent a significant portion of the sales in all items) are mostly being sold to Mining & Quarries.
- Customers who are recorded as 'Other' generate the most Sales. They moistly seem to buy filters for vehicles and hardware equipment.
- Construction customers generally purchase all item categories.
- This is also true for mining customers however they are more invested into filters and kits.

Next, we will investigate how the number of units sold, affects the overall total sales. We will plot a scatter plot to try and visualize this.

Figure 11: Year vs Units Sold (Size = Sales)

Discussion:

- The number of units sold seems to be correlated to sales.
- The business is showing an overall steady increase in sales. The dip n performance for year 2020 is due to obvious reasons, however the upward trend is still ongoing.

PART 4: CUSTOMER SEGMENTATION

We will now investigate Sales based on the 'Nature of Business' column to see if we can uncover any insights in our date.

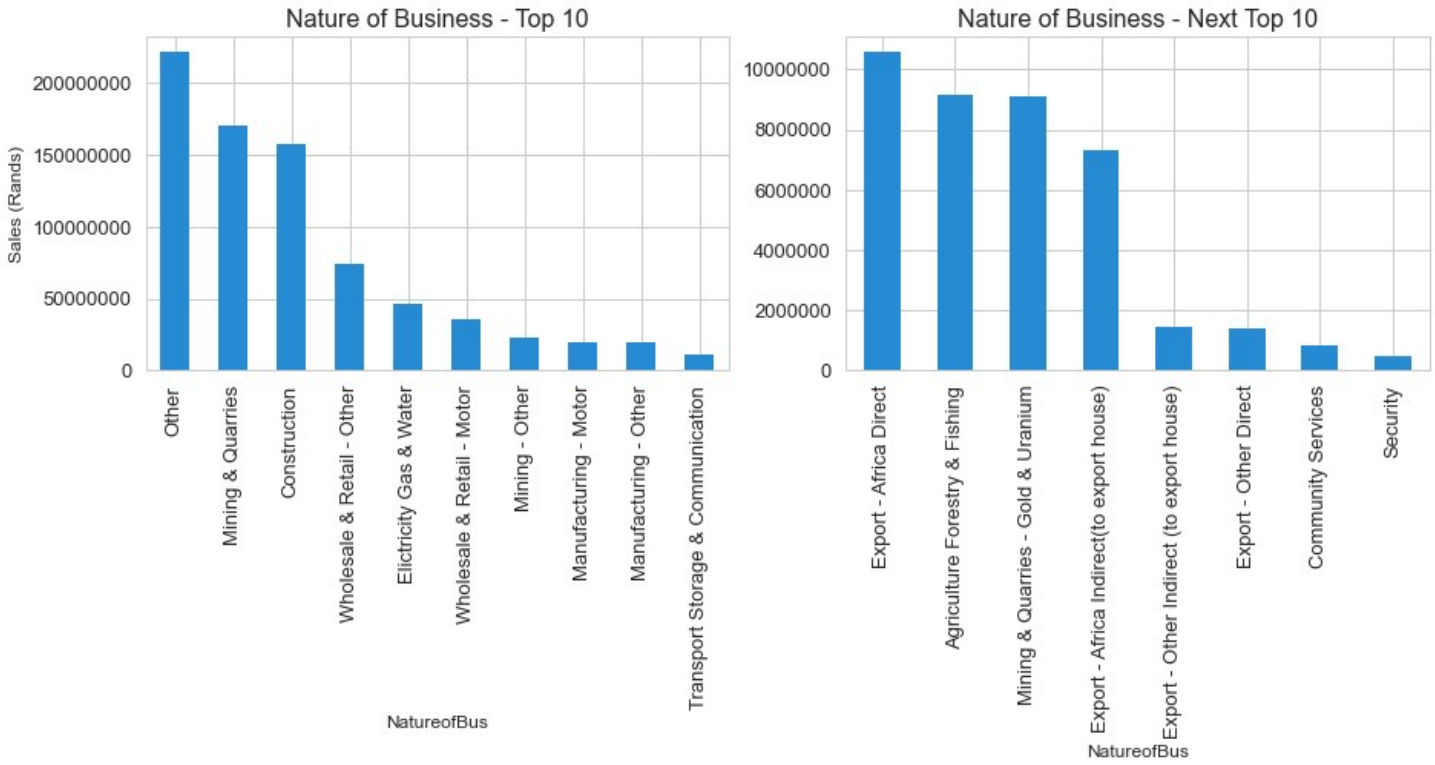


Figure 12: Sales by Nature of Business

We should follow the same convention above to plot the next visualizations. I want to analyze the way our customers engage with the business based on month. We may be able to uncover patterns of spending.

Figure 13: Sales by Customer vs Month

Discussion:

- Sales to customers recorded as 'Other' peak from July to October.
- This is also true for mining customers however sales from these customers are uniform for the rest of the year.
- All customers record their lowest sales in December and January.

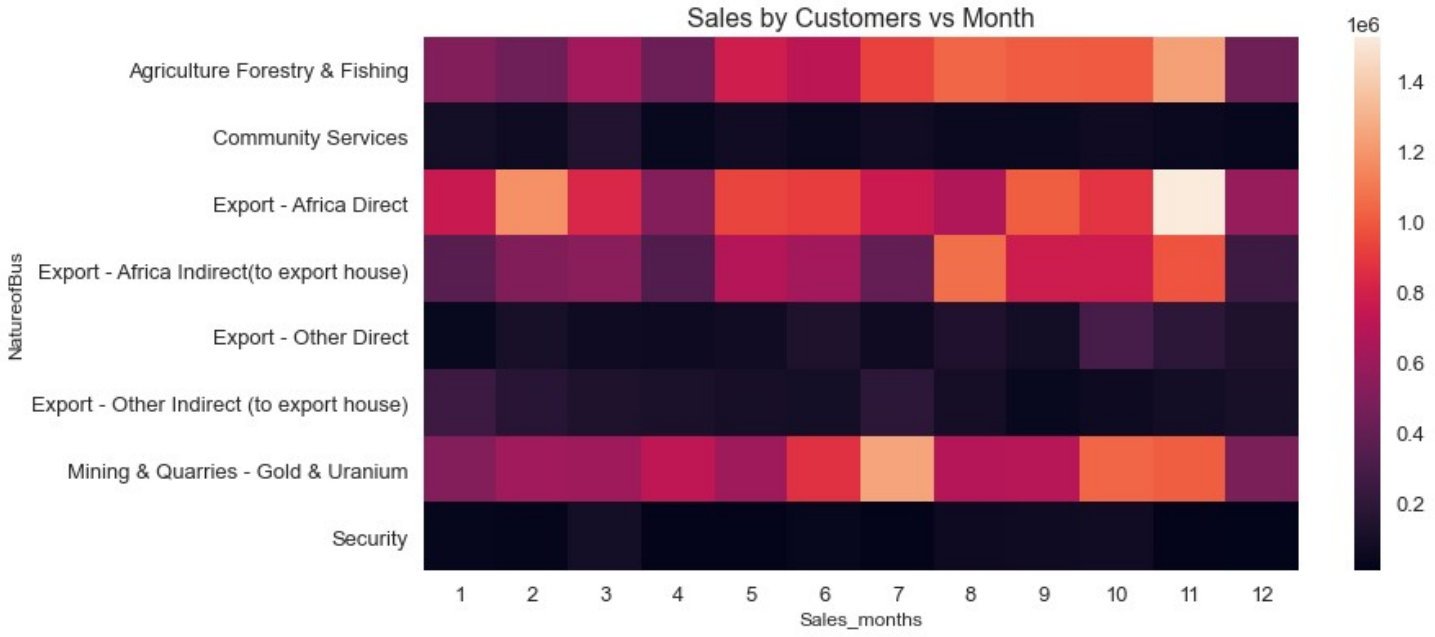


Figure 14: Sales by Customer vs Month

Next, we will assess the correlation of the column names 'DCBalance', 'fForeignBalance', and 'Credit_Limit' to overall Sales.

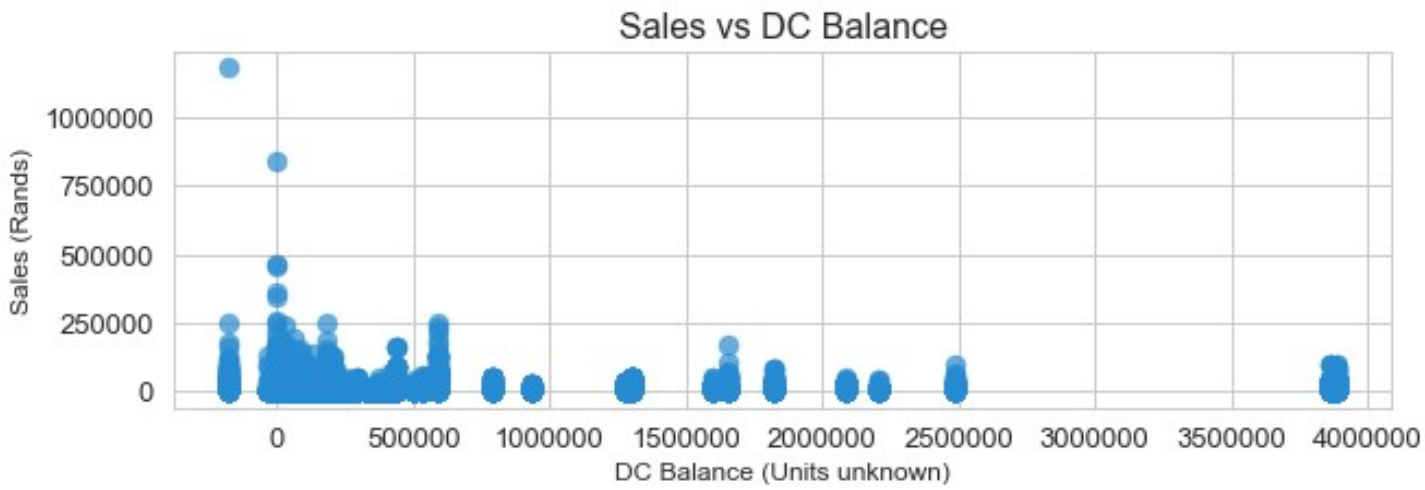


Figure 15: Sales vs DC Balance

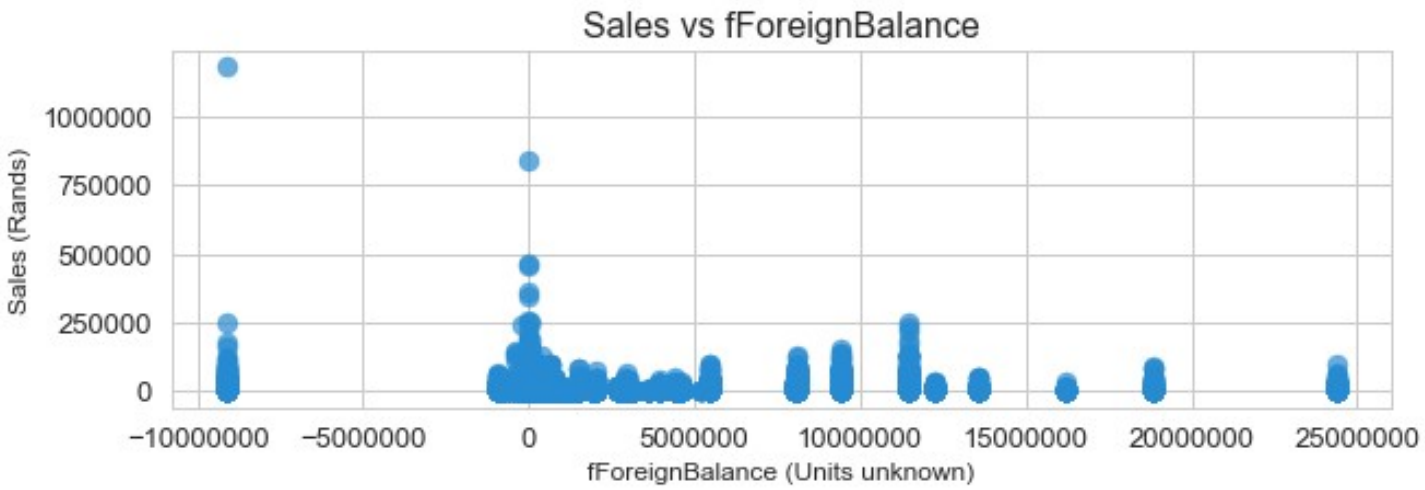


Figure 15: Sales vs Foreign Balance

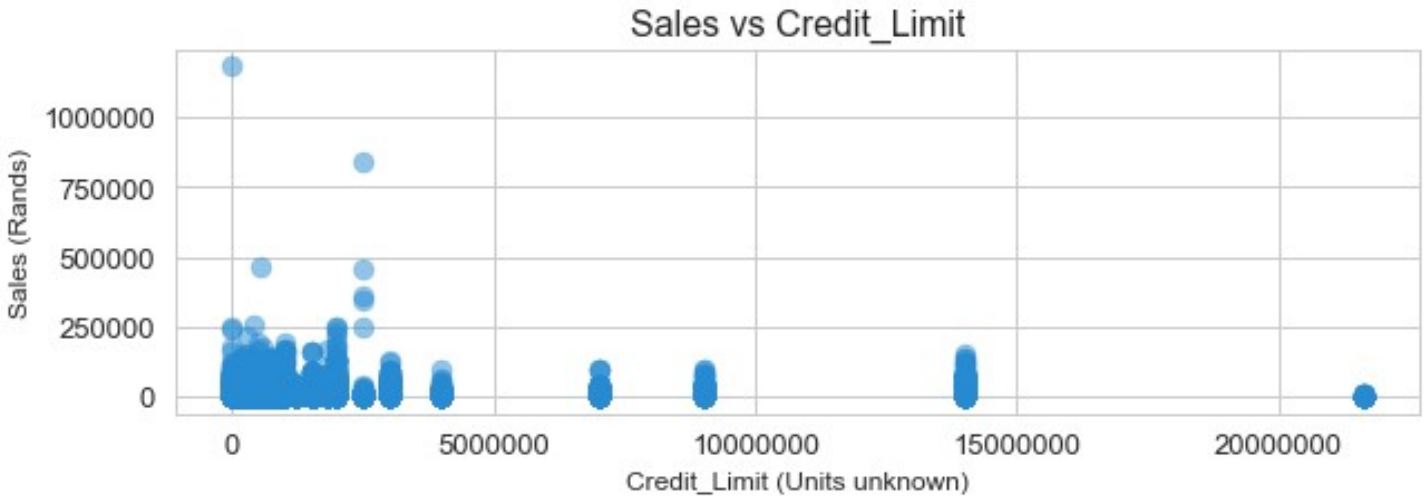


Figure 16: Sales vs Credit Limit

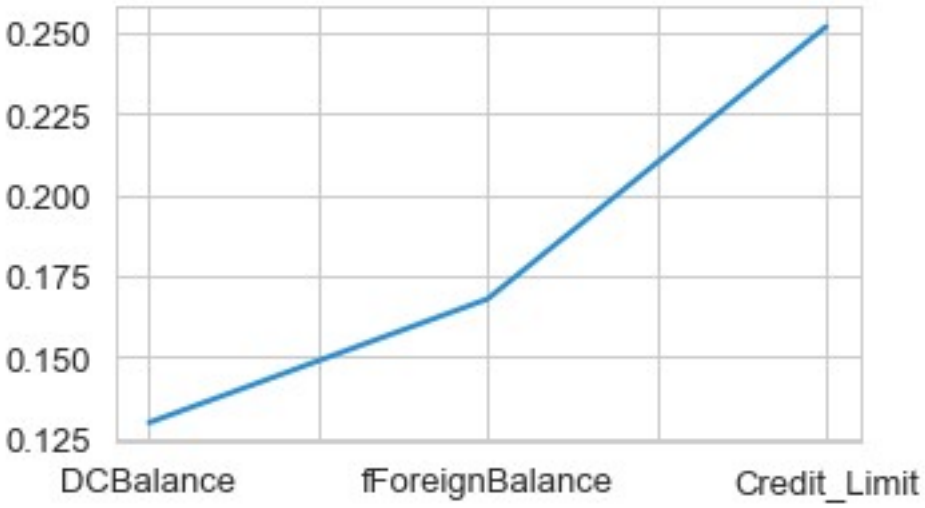


Figure 17: Correlation to Sales

Discussion:

- There doesn't appear to be a strong correlation between the 3 variables to overall sales.
- The Credit limit is the most important variable. It seems we can expect higher sales from customers with a higher Credit limit.

PART 5: PRODUCT ANALYSIS

Next, we will be analyzing the 'Brand' column to assess the how different brands are sold.

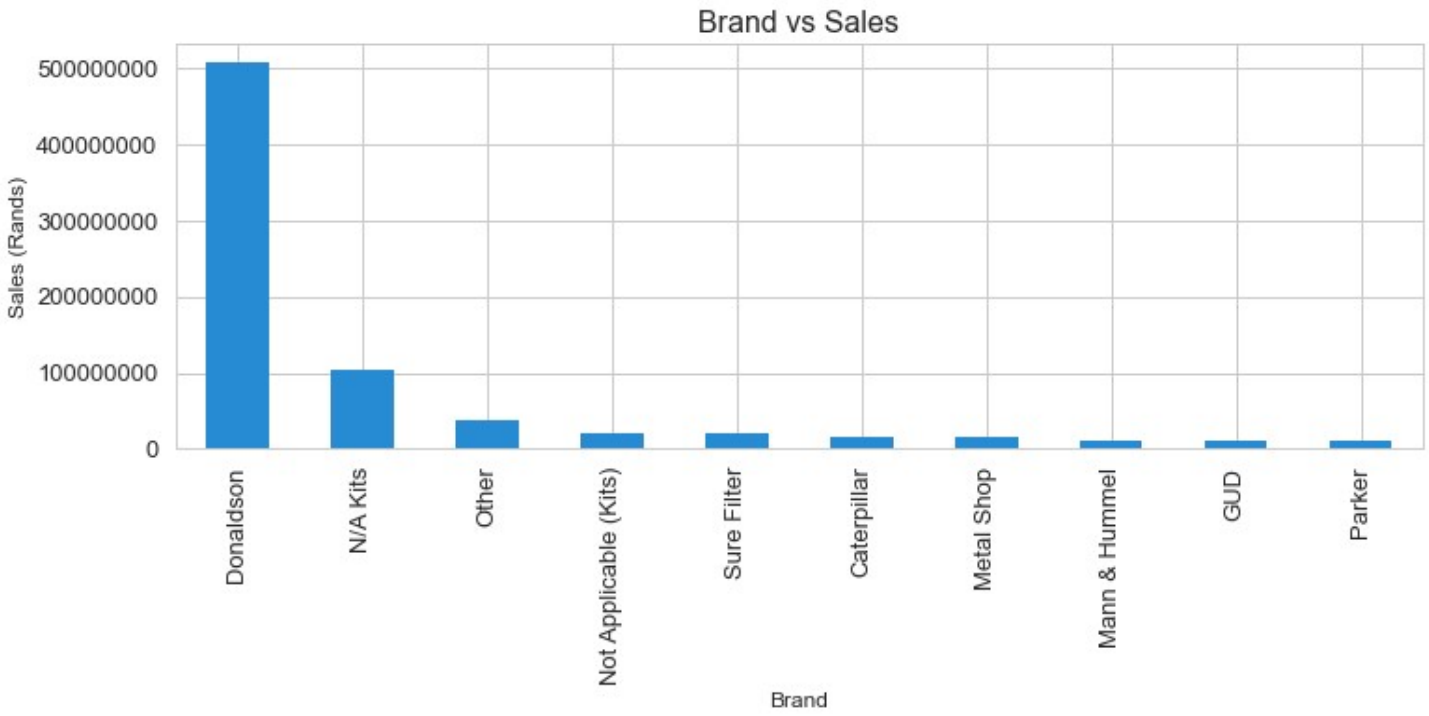


Figure 18: Sales by Product brands

Next, we will look at which brands customers engage with regards to the products sold in the business.

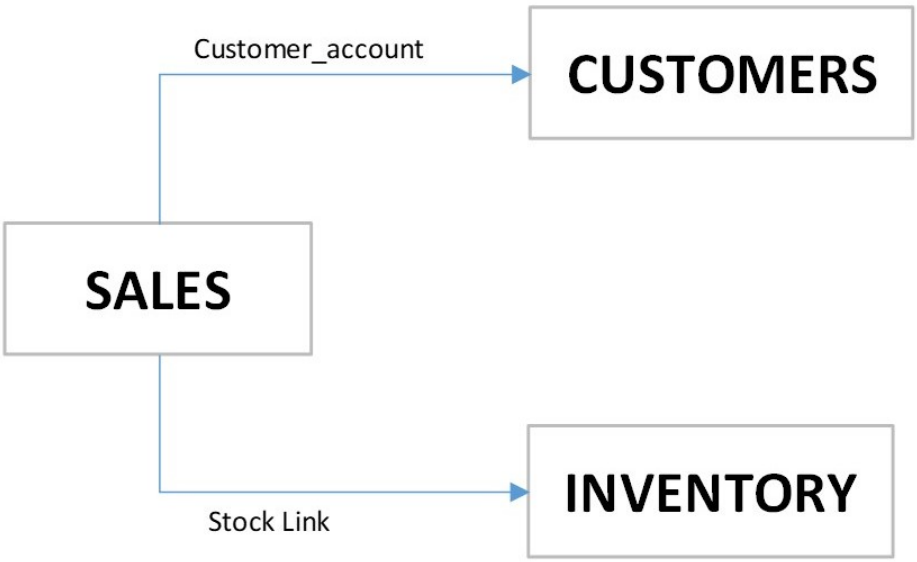
Figure 19: SALES BY BRANDS VS PRODUCT

Discussion:

- It seems customers either trust 1 brand for each item or the business only sells 1 kind of brand for each item.
- The brand name Doosan seems to dominate sales for most products.

PART 6: DATA MODELLING AND OPTIMIZATION

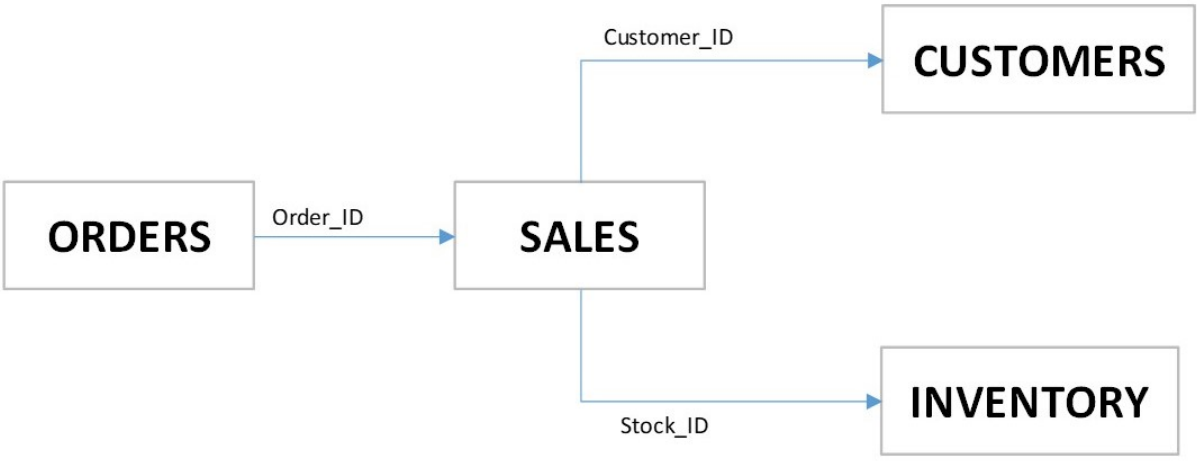
Current data model:



The following steps detail what I would do to improve the data model to improve efficiency with how we engage with the database.

- The tables in the database are cluttered due to their many columns. We need to drastically decrease the number of columns in the **SALES** table by removing columns which are already in **INVENTORY** and **CUSTOMERS** table. These include Customer_name, Description_1, ItemGroup, ItemGroupname, Period, FinPeriod, RepName, repdivision, WeightValue and all the date columns (to be reasoned below).
- For the **INVENTORY** table, it would be wise to address the 3 'description' columns. It is bad practice to add a new column for data that already has a dedicated column. There are also way too many columns. With background knowledge, these can be addressed. For this exercise, I advise collapsing the three descritpion columns into 1.
- For the **CUSTOMERS** table, the same issue as above is seen where we have 3 'address' columns. These appear to be related. Therefore the only sensible option would be to collapse the 3 columns into 1. There are too many columns in this table and it seems they are to blame for why the table has so many NULL values. These need to be removed.
- We need to eliminate confusion, Primary keys need to be easily spotted on every table. We need to rename all Primary_key columns to include the ending 'ID'. So for the **CUSTOMERS** table, we need a 'Customer_ID' instead of 'Customer_account'. For the **INVENTORY** table, we need a 'Stock_ID' instead of 'StockLink' and so on.
- In conclusion I would advise there to be an **'ORDERS'** table. This will contain all the unique orders processed in the business. This table will be linked to the **SALES** table using an Order_ID. The most important columns in this table would be the Order_ID and date of transaction. We will then be able to delete the columns pertaining to date from the **SALES** tables.

Proposed model:



--end--