DATA CO SUPPLY CHAIN EDA

MAIb 23

yash patkar AJ23SYD006

2023

**Contents:**

|  |  |
| --- | --- |
| **Topic** | **Page No.** |
| 1.Problem statement | 2 |
| 2.Data Set | 2 |
| 3.Cleaning Data and feature Engineering | 2-3 |
| 4.Exploratory Data Analysis | 4 |
| 4.1 Sales and Profit Analysis For The Top 10 Regions | 4-6 |
| 4.2 Shipment Delay Analysis | 6 |
| 4.3 Top 5 Customers overall and region wise | 7-9 |
| 4.4 Fraud Analysis | 10 |
| 4.5 Customer segment Analysis | 10-11 |
| 5. Conclusion | 12 |
| 6. Reference | 12 |

1. **Problem Statement:**

We have been given a supply chain data set which has had some work done on it. The goal is to get as much information out of the data set as much possible and implement it on the supply chain as to maximize the profits and/or minimize the loss. The end goal of this project is to pick out at least five features from the data set and perform Exploratory Data Analysis on it using univariant, bivariant or multivariant analysis depending on the problem statement.

1. **Dataset:**

Dataset Link: https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis

This dataset consists of **53 features and 180519 rows**. The features in the dataset include:

'Type', 'Days for shipping (real)', 'Days for shipment (scheduled)','Benefit per order', 'Sales per customer', 'Delivery Status','Late\_delivery\_risk', 'Category Id', 'Category Name', 'Customer City’, ‘Customer Country', 'Customer Email', 'Customer Fname', 'Customer Id', 'Customer Lname', 'Customer Password', 'Customer Segment', 'Customer State', 'Customer Street', 'Customer Zipcode', 'Department Id', 'Department Name', 'Latitude', 'Longitude', 'Market', 'Order City', 'Order Country', 'Order Customer Id', 'order date (DateOrders)', 'Order Id', 'Order Item Cardprod Id', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Id', 'Order Item Product Price', 'Order Item Profit Ratio’, ‘Order Item Quantity', 'Sales', 'Order Item Total’, ‘Order Profit Per Order', 'Order Region', 'Order State', 'Order Status', 'Order Zipcode', 'Product Card Id', 'Product Category Id', 'Product Description', 'Product Image', 'Product Name', 'Product Price', 'Product Status', 'shipping date (DateOrders)', 'Shipping Mode'

Reason for selecting this dataset:

The above data set has had some work done on it so it is a lot easier to work on it and there’s very few missing values that are not going to be dropped. The number of rows is on the higher side but it can be worked with as it is thorough .

1. **Data Cleaning and feature Engineering**

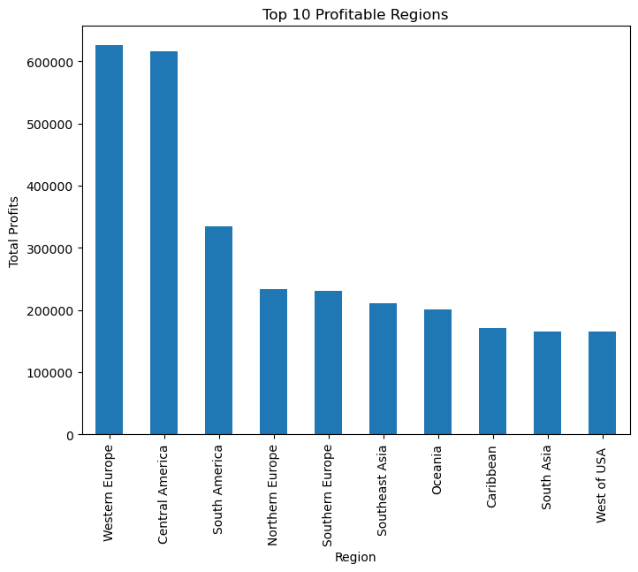
After performing initial data exploration and analyzing the features in-depth, the columns features were cleaned:

* Order Zipcode – This Column has too many null values and doesn’t add much value to the Analysis
* Product Description – This Column has too many null values and doesn’t add value to the Analysis
* Product Image – This Column has links to the Picture of the product, doesn’t hold any vale
* Customer Email – This Column has all the email ID’s of the customers which are of no use when it comes to analysis
* Customer Fname – This Column has the first name of the customer hence of no value
* Customer Lname – This Column has the Last name of the customer hence of no value
* Customer Password – This Column has the password which we don’t need for analysis
* Product Status – This column has all duplicate values that is zero which is of no use
* Customer Street – This column hold the street name which won’t be used in the EDA as we have the state and country
* Order Item Product Price – This column matches value with another column hence dropping it
* Days for shipping (real) – Dropping this column because it has no use as we have a separate column indicating the status of shipment in a more accurate manner
* Days for shipment (scheduled) – Dropping this column because it has no use as we have a separate column indicating the status of shipment in a more accurate manner.
* Latitude – Dropping this column as we don’t need the coordinates of the order/shipping.
* Longitude – Dropping this column as we don’t need the coordinates of the order/shipping.
* Benefit per order – This column matches value with another column, hence dropping it.
* Order Customer Id – This column matches value with another column, hence dropping it.
* Order Item Cardprod Id – This column matches value with another column hence dropping it
* Product Category Id – This column matches value with another column, hence dropping it.
* Order Item Id – This column matches value with another column, hence dropping it.
* Sales per customer – This column matches value with another column, hence dropping it.
* Order Item Profit Ratio – This Column is not needed as a new column with profit or loss is created.
* Customer Zipcode – This column is dropped as it add no value to the EDA
* Profits – A bracket of Profit Loss or None is created to make it easy to categorize.

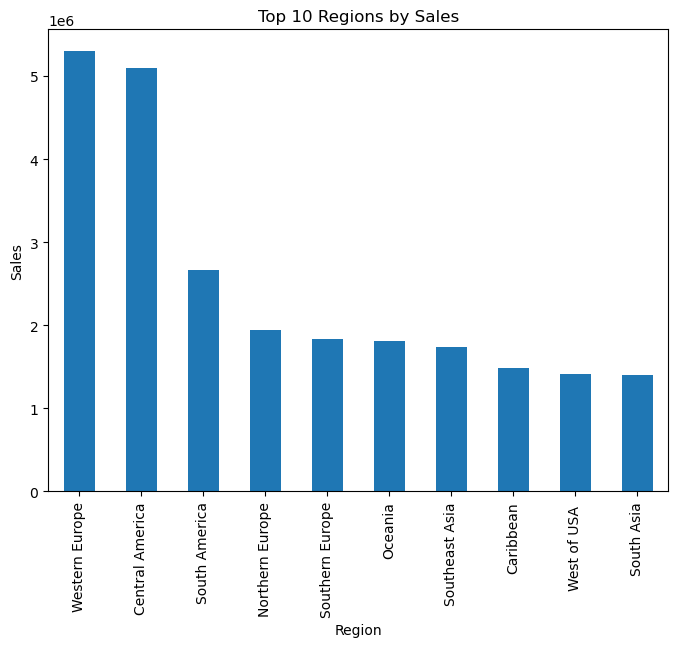
1. **Exploratory Data Analysis:**

Exploratory Data Analysis helps us understand how the features of the dataset vary for the different aspects of the Supply Chain Data .

1. **Sales and Profit Analysis For The Top 10 Regions**

The top 10 regions sorted by profits are Western Europe, central America, South America, Northern Europe, Southern Europe, Southeast Asia, Oceania, Caribbean, South Asia, West of USA in that particular order.

The top 10 regions sorted by Sales are Western Europe, central America, South America, Northern Europe, Southern Europe, Oceania, Southeast Asia,

Caribbean, West of USA, South Asia in that order.

Across all 10 of these regions the highest exported products belong to the Fan Shop department and the next after that is Apparel which have the most sales as well as the most profit and after that comes footwear and golf which we can se consistently across all the regions but the profit and sales vary among them in some regions golf sells more than footwear and in some the other way around the only two other consistent Category to exported across all the top 10 regions are Fitness and outdoor.

1. **Western Europe**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Technology, Outdoor , fitness ,Book Shop and Discs Shop Goods are along exported but in lesser quantity
2. **Central America**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Outdoor and fitness goods are along exported but in lesser quantity.
3. **South America**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Outdoor and fitness goods are along exported but in lesser quantity
4. **Northern Europe**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go while Technology goods bring in more profits than Footwear. Technology, Outdoor , fitness ,Book Shop and Discs Shop Goods are along exported but in lesser quantity
5. **Southern Europe**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go while Technology goods bring in more profits than Footwear. Technology, Outdoor , fitness ,Book Shop and Discs Shop Goods are along exported but in lesser quantity
6. **Oceania**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Technology, Outdoor , fitness ,Pet Shop ,Health and Beauty ,Book Shop and Discs Shop Goods are along exported but in lesser quantity. Where technology goods perform very poorly when it comes to making profits compared to other goods selling as much as technology
7. **Southeast Asia**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Technology , Outdoor , fitness ,Pet Shop ,Health and Beauty ,Book Shop and Discs Shop Goods are along exported but in lesser quantity.
8. **Caribbean**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Outdoor and fitness goods are along exported but in lesser quantity
9. **West of USA**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Outdoor and fitness goods are along exported but in lesser quantity
10. **South Asia**Highest exported goods here are Fan Shop goods and Apparel then after that comes the golf and Footwear being the second and third highest and so does the profit list for them go. Technology, Outdoor , fitness ,Pet Shop ,Health and Beauty ,Book Shop and Discs Shop Goods are along exported but in lesser quantity.
11. **Shipment Delay Analysis**­­­­Off all the deliveries made 55% of them are late which affects the price of the product. Maximum number of late deliveries occur for Fan Shop and Apparel after these two categories the most delayed shipping happens for golf products and footwear products. The standard Deliveries are the one that are getting delayed the most especially for Fan Shop. The regions that are experiencing the most amount of late deliveries for Fan Shop are Central America, Western Europe and South America, this problem retains with all the top five categories . A lot of the First-class shipments are also being delayed. The major issue is that some of the same day deliveries are being delayed.  
      
    The major delays are experienced by the Consumer Segment, with the highest delay being In standard class and the next being in First class shipment which is a major problem as Second Class shipments have a lower number of delays.
12. **Top 5 Customers overall and region wise**The Top 5 Customers by profits to the company are Customer ID 2641,1657,9833,2626,5004.  
    2641 = Betty Spears  
    1657 = Betty Phillips

9833 = Jacob Smith

2626 = Laura Smith  
5004 = Dorothy Friedman

1. **Betty Spears**51% of the deliveries made to Betty were late deliveries ,37% of them were advanced deliveries and 12 % of them were shipped on time. Betty often bought products from fan shop and apparels which made about 82% of the purchases with the remaining 18% split between golf footwear and fitness. The orders from Betty are 23% from western Europe and the rest is equally distributed across Caribbean, Central America ,East of USA ,North Africa ,Oceania and South Asia. 30% of the Payment are completed and 12% are still processing. 30% of the payment are pending and 5% are under review with 23% still pending. Off all the orders 88% percent result in profit for the company and 12% result in loss
2. **Betty Phillips**67% of the deliveries made to Betty were late deliveries ,21% of them were advanced deliveries and 12 % of them were shipped on time. Betty often bought products from fan shop and apparels which made about 65% of the purchases with the remaining 12% being Footwear ,10% on golf and the remaining on outdoor products. The orders from Betty are 21% from South America 19% from Eastern Europe ,17% from Central America ,12% Eastern Asia, 10% from South of USA and Southeast Asia each and the remaining 12% from southern Europe and West Of USA . 26% of the Payment are completed. 38% of the payment are pending and 12% are closed with 24% still pending. Off all the orders 90% percent result in profit for the company and 12% result in loss
3. **Jacob Smith**54% of the deliveries made to Jacob were late deliveries ,33% of them were advanced deliveries and 12 % of them were shipped on time. Jacob often bought products from fan shop and apparels which made about 83% of the purchases with the remaining 17% split between golf footwear and fitness. The orders from Jacob are 17% from Western Europe, North Africa, Northern Europe and Eastern Europe each and12% from Caribbean and Eastern Asia each and the remaining 8% from Oceania. 50% of the Payment are completed and 29% are still processing. 8% of the payment are pending and 12% are Closed. Off all the orders 96% percent result in profit for the company and 4 % result in loss
4. **Laura Smith**41% of the deliveries made to Laura were late deliveries ,15% of them were advanced deliveries and 44 % of them were shipped on time. Lauren often bought products from fan shop and apparels which made about 70% of purchases ,19% were golf related purchases with the remaining 11% split between footwear and Outdoors. The orders from Laura are 30% from Central America , 26% East of USA, 19% Southern America, 15% US Center and remaining 11% from Caribbean. 26% of the Payment are completed and 22% are still processing. 15% of the payment are pending and 37% are Closed. Off all the orders 100% percent result in profit for the company but the Total sum still falls short of Betty.
5. **Dorothy Friedman**38% of the deliveries made to Betty were late deliveries ,42% of them were advanced deliveries and 20 % of them were shipped on time. Dorothy often bought products from fan shop, apparels and Golf which made about 89% of purchases and the remaining 11% were fitness footwear and Outdoors combined. The orders from Dorothy are 22% from Central America and South America Combined , 11% East of USA and Eastern Asia each, 16% Southern Europe, 9% from West Asia, 7% from Western Europe and 2% from Eastern Europe .47 % of the Payment are completed and 16% are still processing. 27% of the payment are pending and 11% are on Hold. Off all the orders 82% percent result in profit for the company , 16% result in loss and 2% breaks even .

**The top Customers from Puerto Rico are :**

**Chart, bar chart

Description automatically generated**

**The Top Customers from EE. UU. Are:**

**Chart, bar chart

Description automatically generated**

1. **Fraud Analysis**Since the data has already been processed before there is a column that has done the fraud analysis and determined weather an order is suspected fraud or not, hence we will work backwards and try to understand how it works. We can see that all the flagged orders have their payment type as direct transfer. But we can’t completely say that all transfer payment are fraudulent as we can see from the unstacked graph a lot of pending orders and processing do fall under the direct transfer category. The pie chat gives us the idea about the profit to the company and even that doe not help in understanding as it shows that 80% of these transactions the company would have profited only 20% would have been a loss which is not such a huge margin. Next we can see majority of these Fraudulent orders are being placed for products from Western Europe, Central America, South America and southern Europe. The graph for the distribution of goods that are flagged doesn’t help us much as the items are from Fan Shop , Apparel , golf and Footwear which is also the top selling goods outside of fraud. Now as we go to the graph that shows the Customer city we can see the feature that rises the flag as the suspected fraud in Caguas is much higher compared to any other city. As we go further in, we see the customer segment which does not help much as we know most of the orders are from consumers so nothing to tip off the algorithm or raise flags and we can see that majority of the orders have standard shipping.as we look at the product name it raises the flags as there about nine products there suspected fraudsters frequently order and those are the Perfect Fitness Perfect Rip Deck, Nike Men's CJ Elite 2 TD Football Cleat Nike Men's Dri-FIT Victory Golf Polo, O’Brien Men's Neoprene Life Vest, Field & Stream Sportsman 16 Gun Fire Safe, Pelican Sunstream 100 Kayak, Diamondback Women's Serene Classic Comfort Bi, Nike Men's Free 5.0+ Running Shoe, Under Armour Girls' Toddler Spine Surge Running which do raise a flag.
2. **Customer segment Analysis**

The overall order status distribution and the profit across all the segments are same  **1. Consumer**59% of all orders are standard deliveries with 19% second class and 17% being First Class and 5% being same day.39% of all the transactions are done using debit, 27% using transfer ,22% from Payment and the remaining using cash. 54% of the consumers receive their deliveries late , 24% receive them before delivery date,18% receive them on the day of delivery and 4% have their order cancelled. The top most selling Items for these category of Customers is Fishing, Cleats, Cardio Equipment, Camping and Hiking, Women’s Apparel, Indoor/Outdoor Games, Water Sports, Men’s Footwear. Most of the orders are placed in the Europen, LATAM and Pacific Asis region.

**2. Home Office**

60% of all orders are standard deliveries with 18% second class and 16% being First Class and 5% being same day.36% of all the transactions are done using debit, 28% using transfer ,25% from Payment and the remaining using cash. 55% of the consumers receive their deliveries late , 23% receive them before delivery date,18% receive them on the day of delivery and 4% have their order cancelled. The top most selling Items for these category of Customers is Fishing, Cleats, Cardio Equipment, Camping and Hiking, Women’s Apparel, Indoor/Outdoor Games, Water Sports, Men’s Footwear. Most of the orders are placed in the LATAM, Europen and Pacific Asis region.

**3. Corporate**61% of all orders are standard deliveries with 20% second class and 14% being First Class and 5% being same day.40% of all the transactions are done using debit, 28% using transfer ,21% from Payment and the remaining using cash. 53% of the consumers receive their deliveries late , 24% receive them before delivery date,20% receive them on the day of delivery and 4% have their order cancelled. The top most selling Items for these category of Customers is Fishing, Cleats, Cardio Equipment, Camping and Hiking, Women’s Apparel, Indoor/Outdoor Games, Water Sports, Men’s Footwear. Most of the orders are placed in the Europen, LATAM and Pacific Asis region.

1. **Conclusion:**This project helped me to think outside the box and able to come up with creative ways to explore the relations with the data . The exploration of data helped me better understand data and correlation it even help me understand the necessity of feature engineering in creation relationship between feature .
2. **Reference**Data: <https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis>

Information  
<https://www.w3schools.com/python/default.asp>  
<https://www.python.org/about/gettingstarted/>

https://stackoverflow.com/  
<https://chat.openai.com/chat>  
<https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python>  
<https://towardsdatascience.com/exploratory-data-analysis-in-python-a-step-by-step-process-d0dfa6bf94ee>  
https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/

.