Supply Chain Data Set

**Contents:**

|  |  |
| --- | --- |
| **Topic** | **Page Number** |
| 1. Problem Statement | 2 |
| 1. Dataset | 2 |
| 1. Data Cleaning and Feature Engineering | 3 |
| 1. Exploratory Data Analysis   4.1 Shipment Delay Analysis  4.2 Payment Fraud Ana;ysis  4.3 Order analysis  4.4 Customer Analysis  4.5 Market Analysis  4.6 Profit Analysis | 3  4  7  9  14  17  21 |
| 1. Conclusion | 23 |
| 1. Refrences | 23 |

1. **Problem Statement:**

We have been given a dataset of a global supply chain management company that operates in multiple regions of the world. The company has collected varying amounts of data on its supply chain activities for the past several years. The goal is to clean, analyze and provide insight into the data. We are to come up with at least 5 Visual Statistical summaries and insights on 5 different aspects of the data.

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, Sales per customer, Delivery Status, Late\_delivery\_risk, Category Id, Category Name, Customer City, Customer Country, Customer Id, Customer Segment, Customer State, Department Name, Market, Order City, Order Country, order date (DateOrders), Order Id, Order Item Discount, Order Item Id, Order Item Product Price, Order Item Profit Ratio, Order Item Quantity, Sales, Order Profit Per Order, Order Region, Order State, Order Status, Product Card Id, Product Name, shipping date (DateOrders), Shipping Mode, CustomerFullName, Delay, Sales Range

Reason for selecting this dataset:

The dataset has a good collection of features. The data set is based on real-world transactions and can help us explore different aspects of the supply chain industry. The data set is complex with a vast amount of variables which can help us develop our skills and knowledge in data analysis.

1. **Data Cleaning and Feature Engineering:**

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

* Days for Shipping (real) and Days for Shipping (scheduled)- Don’t provide any useful insights instead I used Delay for more insights.
* Benefit per order – The benefit per order column is the repetition of Order Profit Per Order.
* Customer Fname and Customer Lname- I have removed these 2 columns and have created a new column called CustomerFullName also Customer Id can be used to correlate with other features.
* Latitude and Longitude – This doesn’t give any insights and also it can raise concerns for customers as it can reveal their locations.
* Product Description- NAN Values.
* Customer Zipcode – This doesn’t give any insights. Customer Country provides enough information about the location.
* Customer Street- Same reason as Customer Zipcode.
* Customer Email- It is a piece of sensitive information and also it doesn’t have its use in the analysis part.
* Customer Password- Same reason as above(Customer Email).
* Order Zipcode- It contains sensitive information and it doesn’t have its use in the analysis.
* Department Id- Doesn’t provide any insights.
* Order Item Discount Rate – I have decided to use Order Profit Per Order instead of this feature for my analysis.
* Order Customer Id- it is redundant with other customer-related columns such as ‘Customer id’ or ‘CustomerFullName’.
* Order Item Cardprod Id- It contains a unique value that is of no importance. I have used product names for my analysis.
* Product Image- The image and description have no value for the analysis.
* Product Category Id- This column doesn’t provide much information about the product, we have a Product Name for analysis which contains more understandable information, and also this feature is a repeat of Customer Id.
* Product Status- NAN Values.
* Product Price- It has a similar value as Order Item Product Price.
* CustomerFullName- Created a new column, which is a concatenation of 2 columns Customer Fname and Customer Lname.
* Order date(DateOrders)- I have removed time data and also converted the data type into data time type.
* Shipping date(DateOrders)- The same operation I have done as above.
* Delay- Created a new column by subtracting the order date from the shipping date.
* Sale Range – Created in less than 500, 500-1000, 1000-1500,1500-2000.

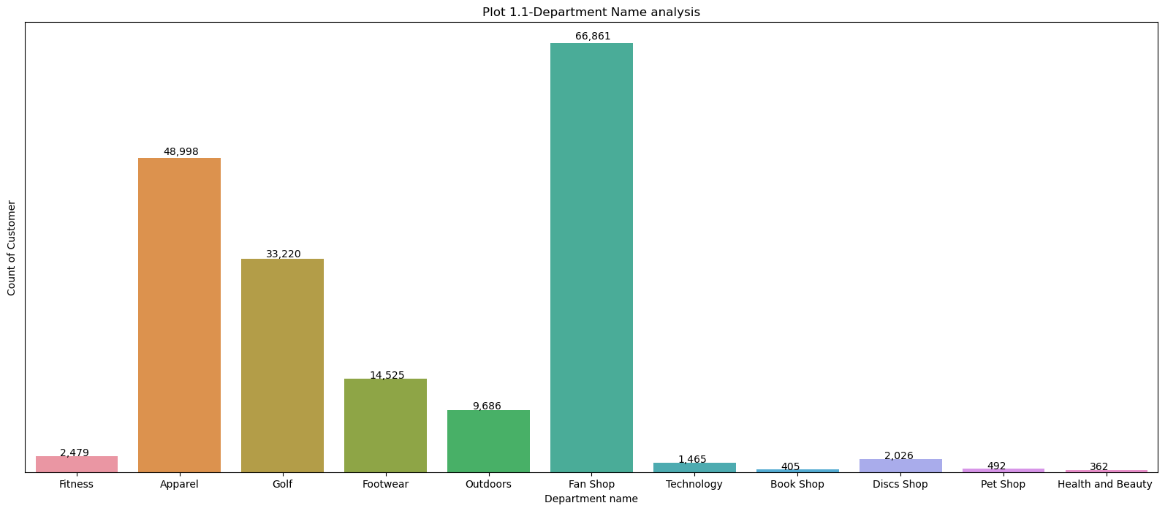
1. **Exploratory Data Analysis:**

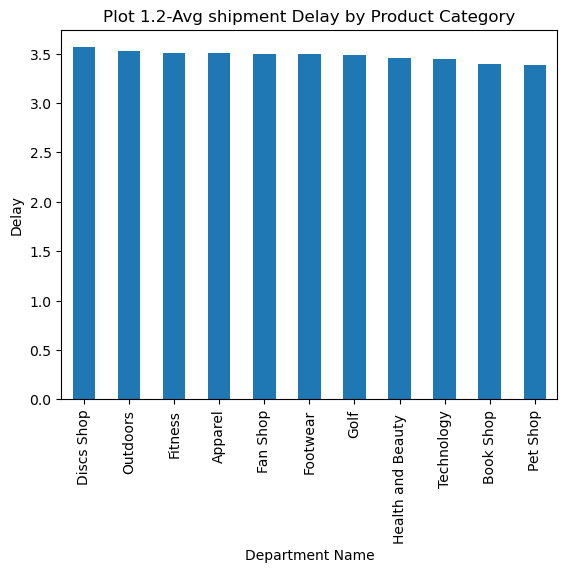
Exploratory Data Analysis helps us understand how the features of the dataset vary for the fraudulent cases of vehicle insurance claims. We first start by understanding the features alone (Univariate Analysis) and then perform bivariate and multivariate analysis to understand the data and relations between the features better.

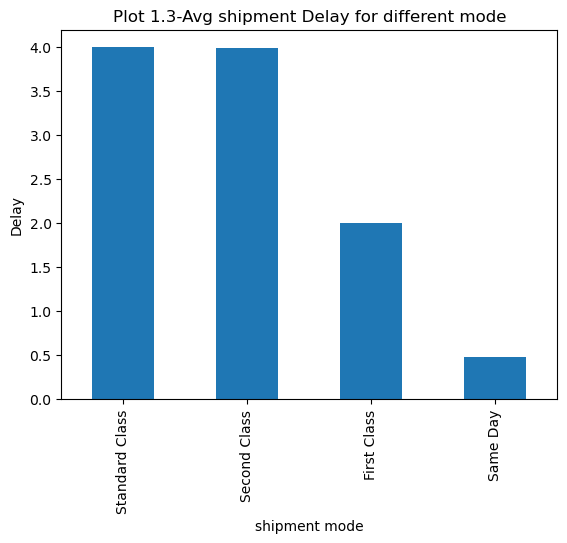
Different Exploratory Data I have done for this data set are**- Shipment Delay Analysis, Payment Fraud Analysis, Order Analysis, Customer Analysis, Market Analysis, and Profit Analysis.**

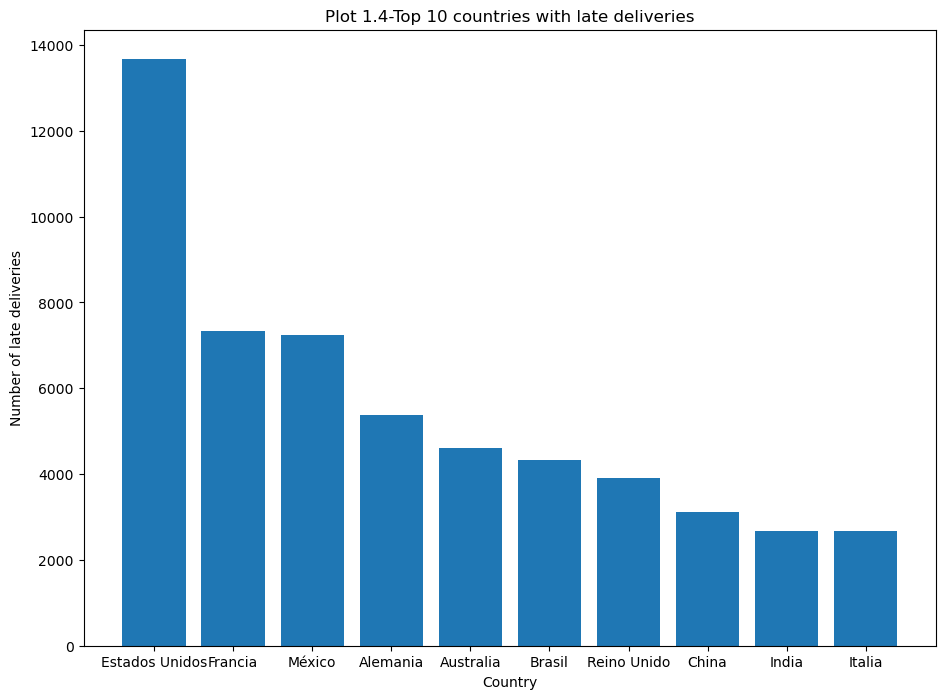
* 1. **Shipment Delay Analysis:**

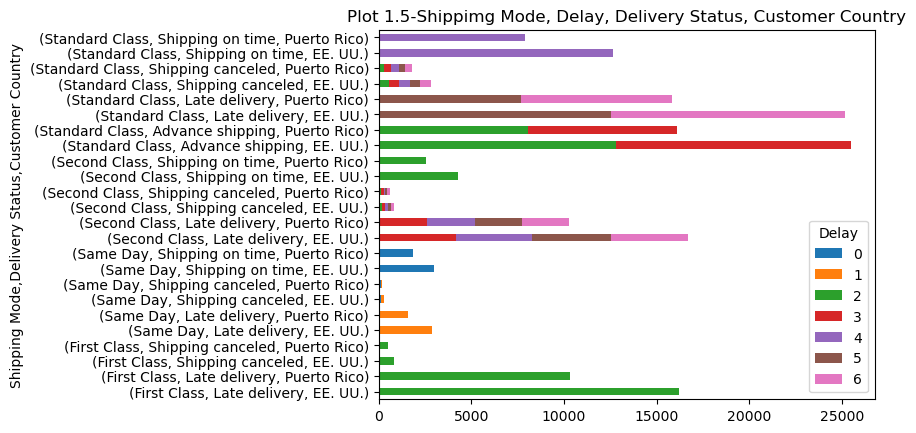
This analysis is done to optimize shipping routes, improve packaging, enhancing the delivery process. It is also used to mitigate the impact of delays.

****

****

****

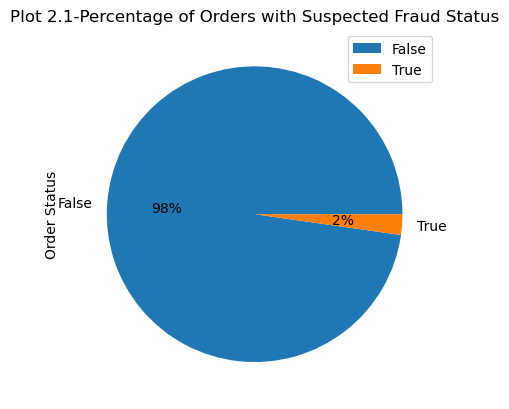
****

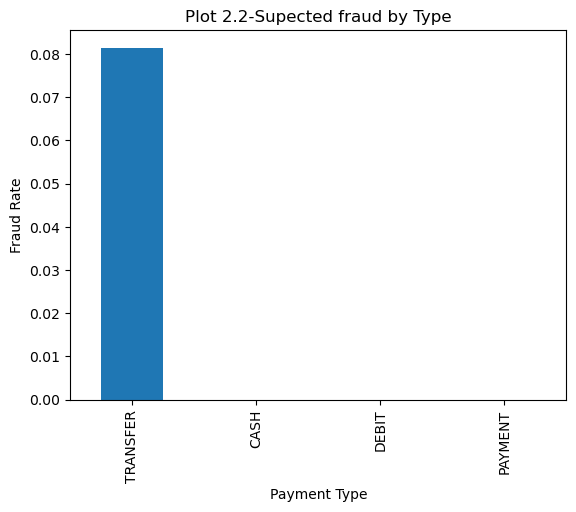
****

**Observations-**

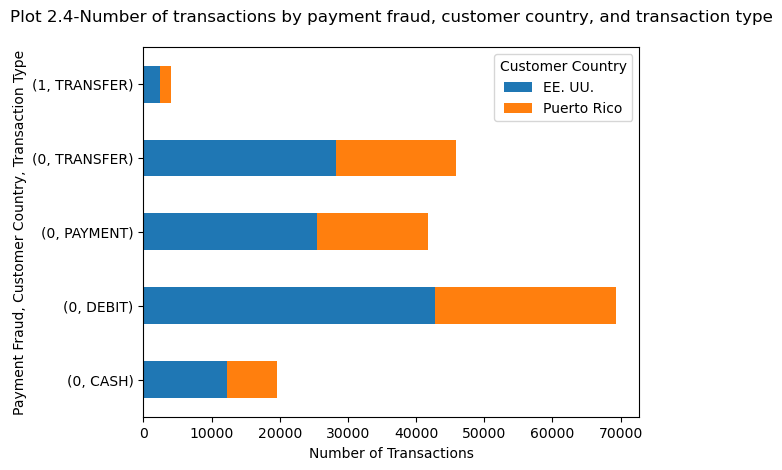
* **Plot 1.1-** In this plot, each bar represents a department, and its height corresponds to the number of customers in that department. This can help businesses to identify which departments are more popular among customers and which will give insights to a company regarding resource allocation. In our data **fans** shop is the most popular department with 66,861 counts of customers.
* **Plot 1.2 -** This department indicates which department has the most shipment delays. The discs shop has the most average shipment delay of more than 3.5 days. Whereas fans ship which is the most popular department has a shipment delay of 3.3 days
* **Plot 1.3-** The graph shows the average shipment delay for different modes (First class, Same day, Standard class, Second class). Standard class and Second class have the almost same amount of shipment delays i.e around 4 days and same-day mode has been the one with the least delay..
* **Plot 1.4-** The graph shows the top 10 countries with the highest number of late deliveries. Estados Unidos has the highest number of deliveries.
* **Plot 1.5-** his plot can help find the order status and delivery status of EE UU and Puerto Rico. We can observe that EE UU country has received more orders when compared to Puerto Rico. Most delays can be observed in EE UU when a customer's order status is second-class. his graph can give us insights into identifying areas where improvements can be made in order fulfillment and delivery processes which can increase customer satisfaction.
  1. **Payment Fraud Analysis:**

Payment fraud analysis is the process of analyzing and identifying fraudulent transactions in payment systems.





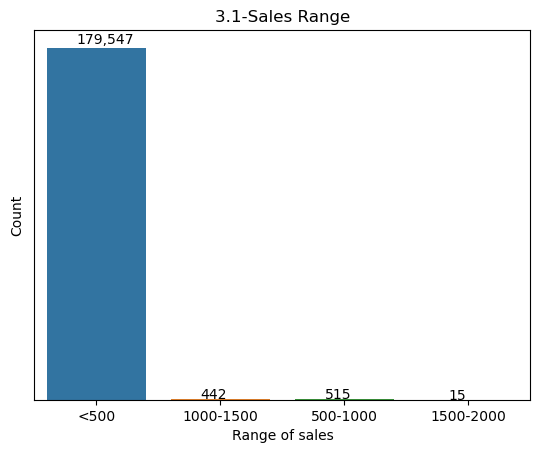


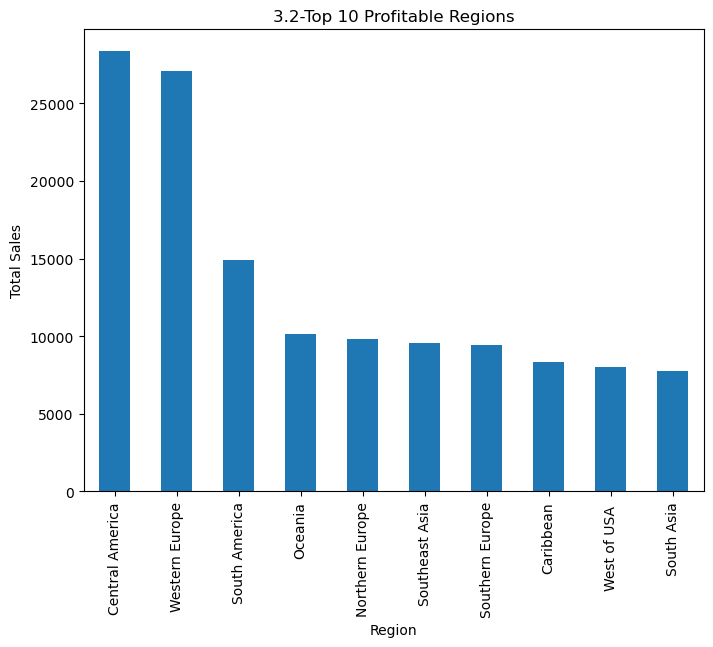


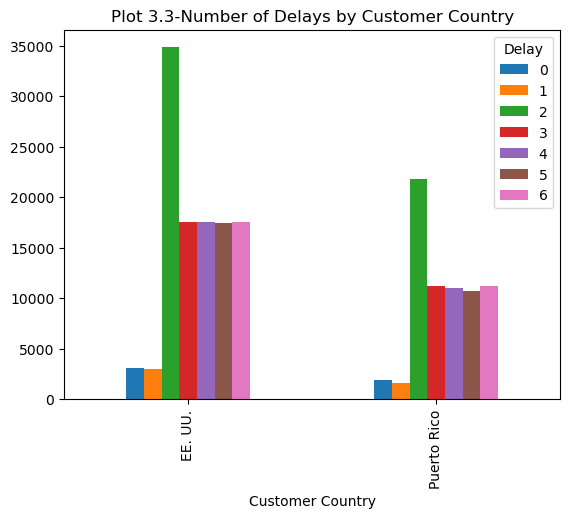
**Observations:**

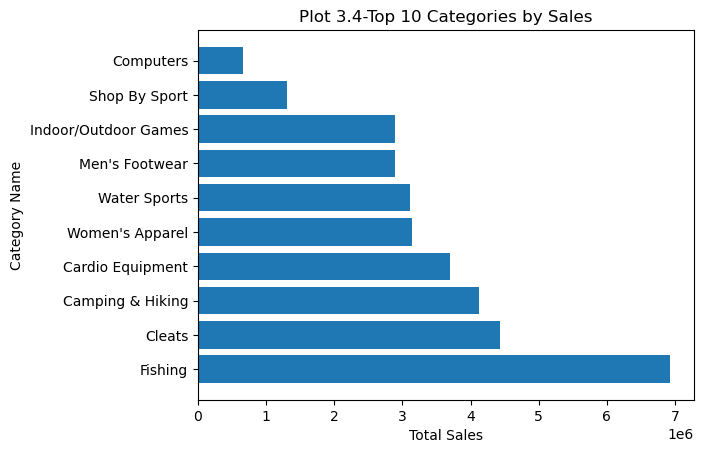
* **Plot 2.1-** The pie chart shows the percentage of orders with suspected fraud status. From the graph, we can observe that there have been 2% fraudulent cases. We can use this graph to calculate fraudulent activity in the business.
* **Plot 2.2-** The second graph shows the fraud rate by payment type. It is observed that when fraud status is only seen when the transfer-type payment method is used. The company can identify which payment methods are most vulnerable to fraud using this graph.
* **Plot 2.3 -** This graph shows the number of customers by payment fraud and customer segment. From the graph, we can observe that the consumer segment has the most fraudulent case. So, the consumer segment is most vulnerable to fraud.
* **Plot 2.4 -** The fourth plot shows the number of transactions by payment fraud, customer country, and transaction type. The most fraudulent case is been observed in EE. UU country.
  1. **Order Analysis:**

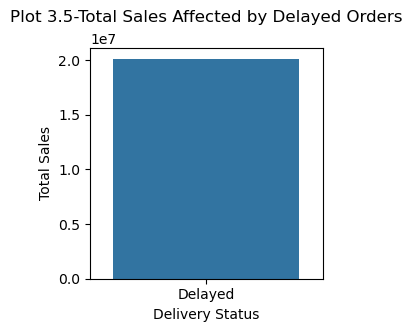
It is used to find trends and insight into orders placed by customers.

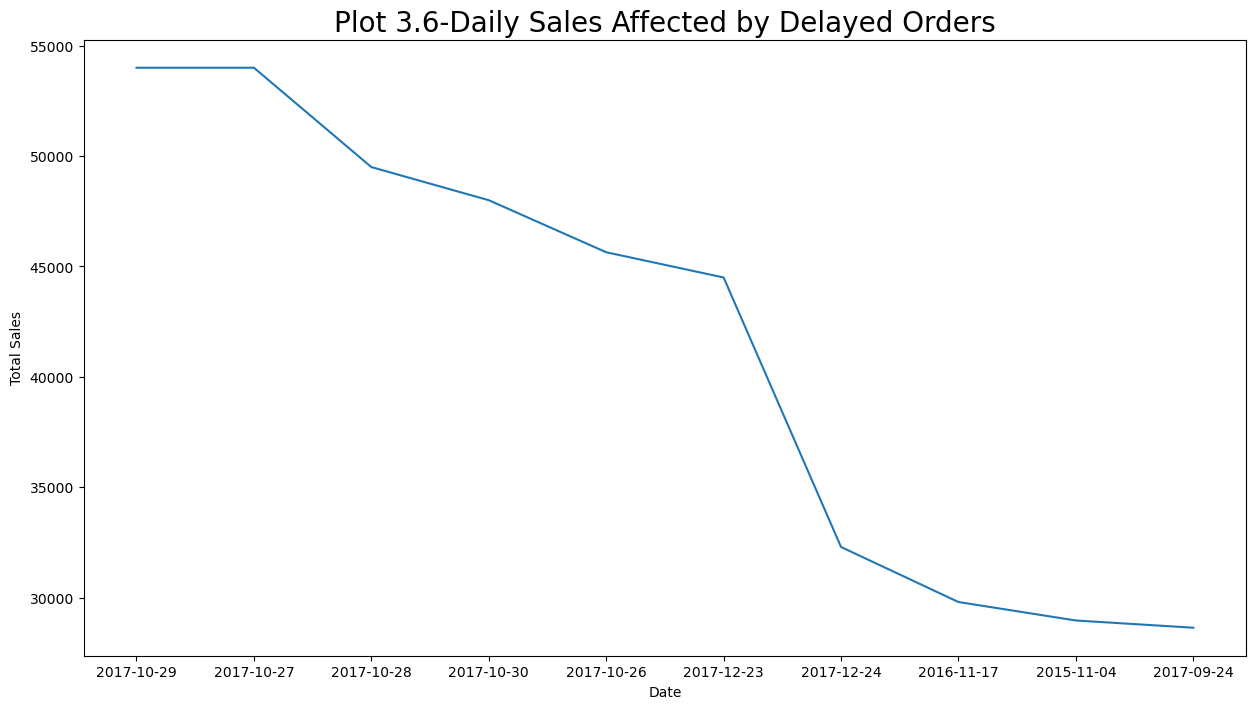
****

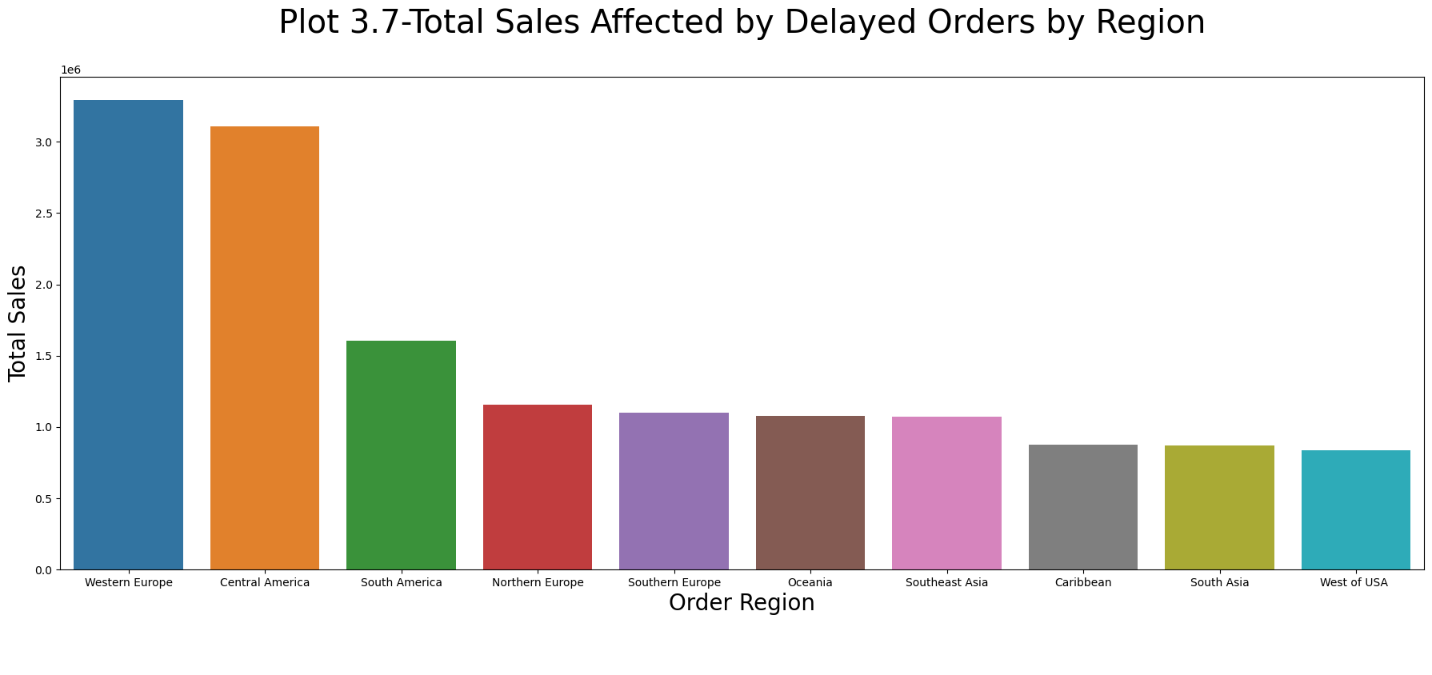
****







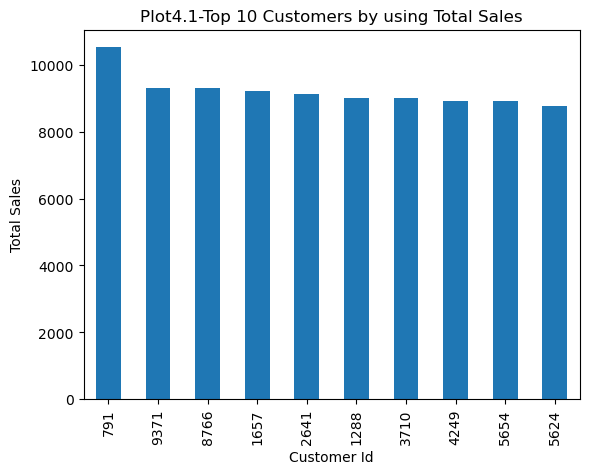


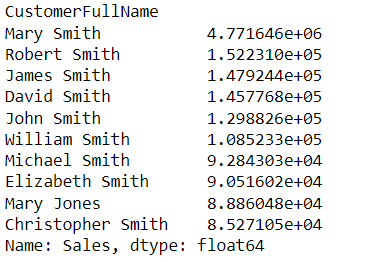


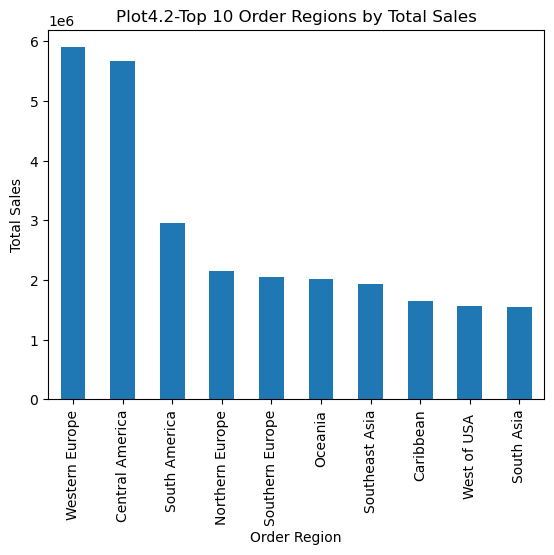
**Observations:**

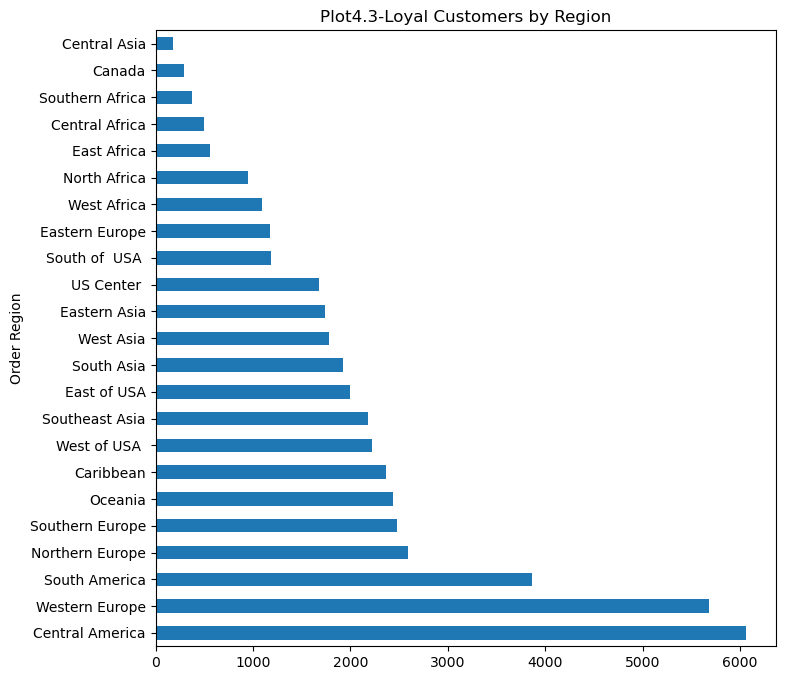
* **Plot 3.1-** The graph shows the count of sales in various sales ranges. From this graph, we can observe different counts of sales for a given sales range. The sales for less than 500 count is the highest with the value of 179547.
* **Plot 3.2-** We can see the results of the top 10 regions with the highest amount of sales with Central America leading as the top contender for sales.
* **Plot 3.3-** The graph helps us distinguish the country with the most delays. EE UU has the highest number of delays.
* **Plot 3.4-** This graph helps us infer the top 10 categories which were sold to customers. Fishing is the most sold category.
* **Plot 3.5-** By looking at the plot, we can observe how much sales were affected due to delay. The overall impact on sales can be found using this graph.
* **Plot 3.6**- The graph indicates how much the total sales were impacted by delayed orders on each day. You can observe a drastic delay from the 23rd to the 24th of December 2017. we can observe how much money is affected by delays. These insights can help in identifying areas for improvement in delivery mode and process.
* **Plot 3.7**- The graph gives us an observation about total sales affected by delayed orders in the top 10 regions. We can see that Western Europe was most affected by the delays with central America falling behind it.
  1. **Customer Analysis:**

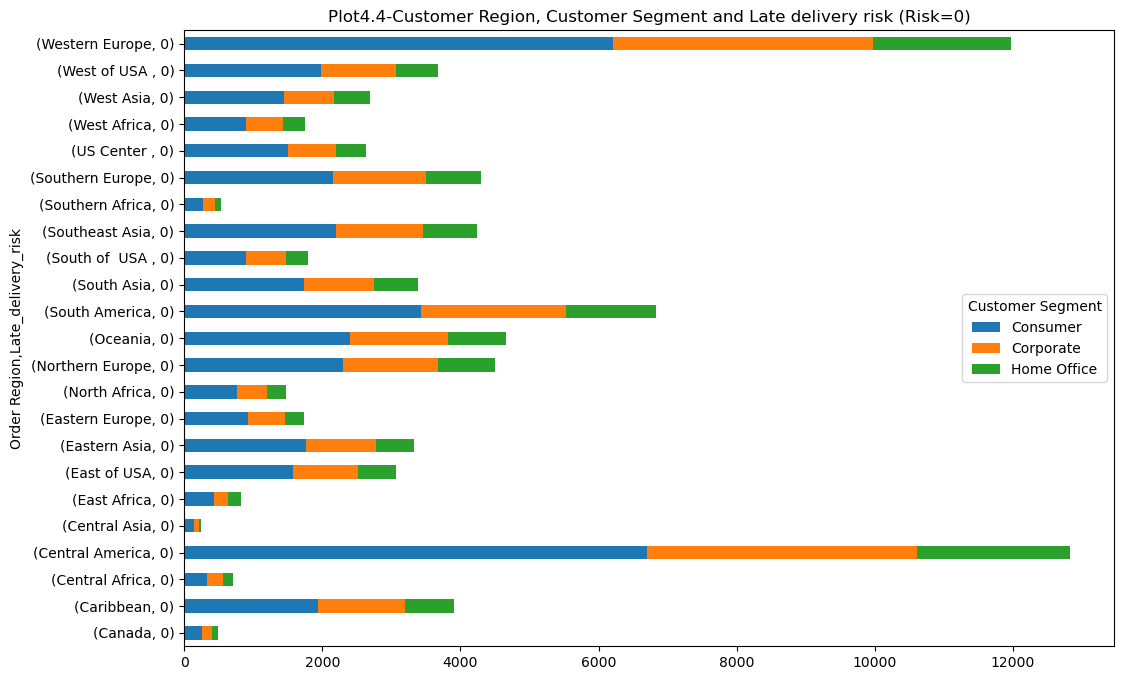
This analysis is used to improve customer experience, loyalty and sales.







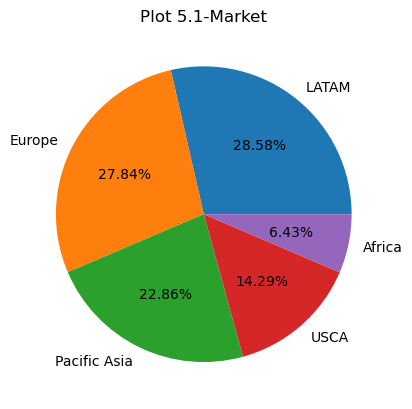


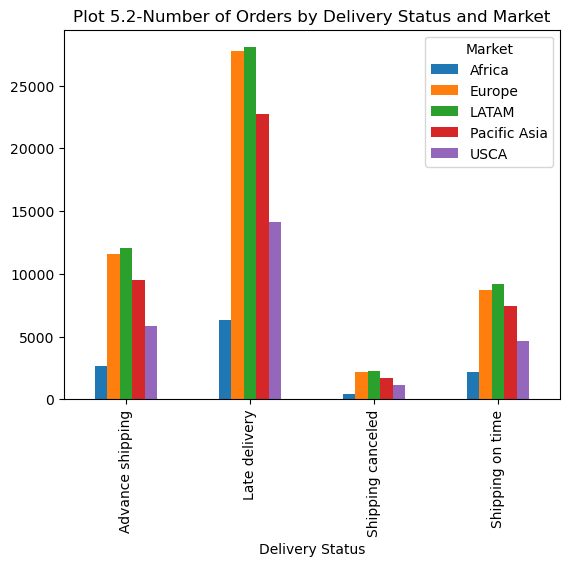


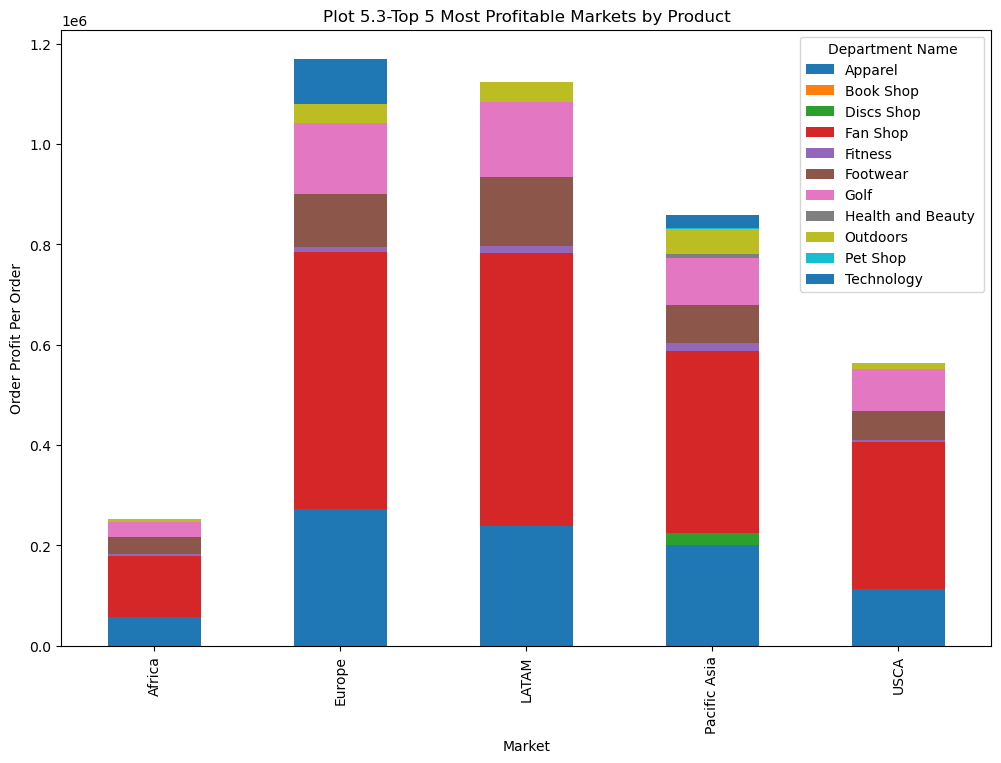
**Observations:**

* **Plot 4.1-** This graph shows the top 10 customers based on total sales. From this graph, we can observe which customer has the highest sales. Customer ID 719, Mary Smith, has the highest sales count.
* **Plot 4.2-** This graph shows the top 10 order regions on total sales. From this graph, we can observe the top 10 regions with the highest sales. We can observe that Western Europe has the highest sales.
* **Plot 4.3-** This is an important graph that shows the number of loyal customers by region. We can observe that Central America has the most loyal customers.
* **Plot 4.4-** This graph helps us find the distribution of orders across different regions and customer segments with a late delivery risk of 0.Consumer product has the highest sales in Central America with a late delivery risk of 0 followed by Western Europe.
  1. **Market Analysis:**

It is used to find insights into different markets to understand the dynamics of the market.

****

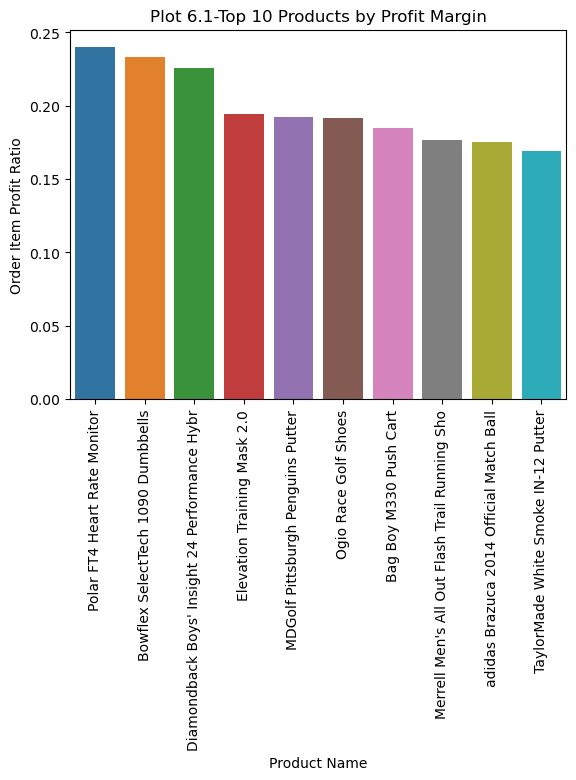
****

****

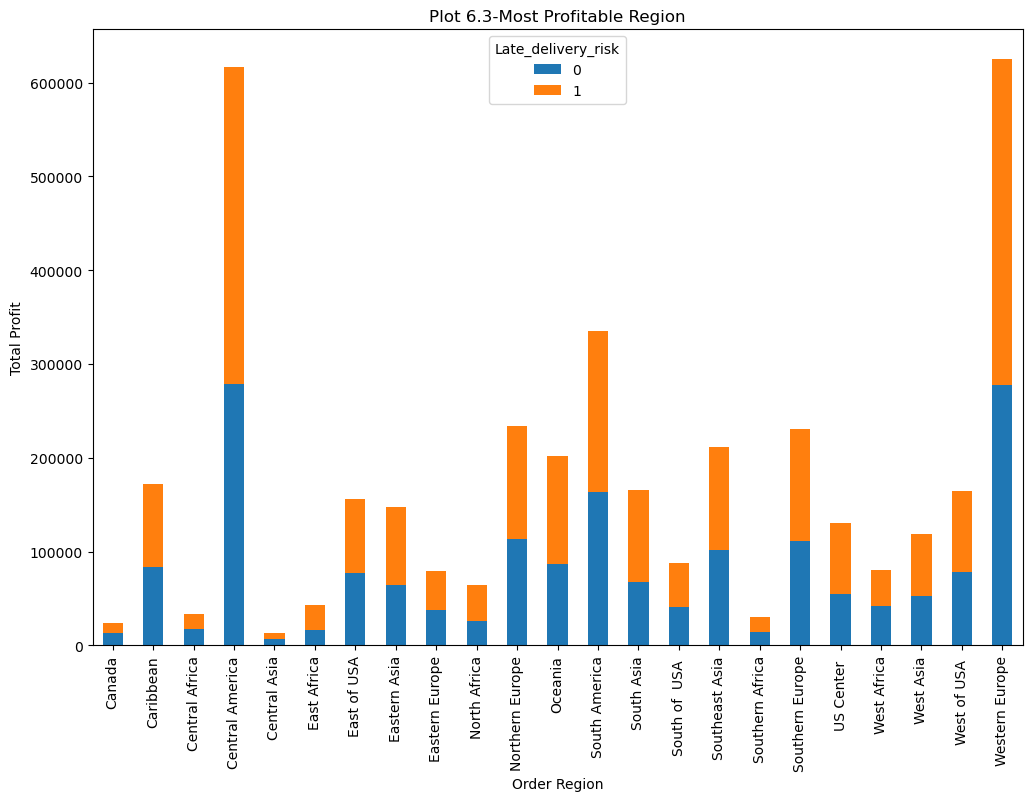
**Observations:**

* **Plot 5.1-** The graph shows the percentage of each market of the dataset. Latam has the highest market when compared with others.
* **Plot 5.2**- We can observe that there are more late deliveries. Latam market has the latest delivery follower by the Europe market. We can also observe that Africa has the lowest amount of orders of all.
* **Plot 5.3-** The graph helps us analyze the most profitable market by department name. You can observe from the graph that Europe has the most profitable market with fan shops being the most marketable option.
  1. **Profit Analysis:**

In profit analysis, we find the reason for incurring profits so that in the future we can use these insights to get a clear picture of how to proceed.







**Observations:**

* **Plot 6.1-** Using this graph we can find the top 10 products with the highest profit ratio. We can observe that the most profitable product is Polar FT4 Heart Rate Monitor.
* **Plot 6.2-** This graph shows which segment has the most profitability with consumer been the segment with highest profit ratio.
* **Plot 6.3-** The chart shows which regions are the most profitable, and whether late delivery risks affect profitability in those regions.

1. **Conclusion :**
2. **Shipment Delay Analysis:** The insights from the analysis can help the company to improve its operations, reduce costs and increase the satisfaction of customers.
3. **Payment Fraud Analysis:** The insights from the analysis can help company take necessary steps to prevent fraud activities. It can also be used to investigate payment method and take necessary steps toward increasing the security of the method.
4. **Order fraud Analysis :** The insights can help prevent fraud, can identify high rish countries and improve customer segmentation and at last can improve payment processing.
5. **Customer Analysis:** The insights from the analysis can improve customer statisfaction, prevent fraud, avoid risk and adjust pricing strategy.
6. **Market Analysis:** The insights from the analysis can identify top performing markets, improve supply chain management, optimize pricing and expand into new markets.
7. **Profit Analysis:** The insights from the analysis can be used to make strategic decisions regarding their product offerings, pricings and markets. The analysis can also help the company identify areas where they can reduce costs or increase efficiencies to improve profit margins
8. **References:**

<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/>

<https://www.kaggle.com/code/shashankmkadiwal/dataco-smart-supply-chain-eda>

<https://www.kaggle.com/code/divyeshardeshana/supply-chain-shipment-price-data-analysis>

<https://www.kaggle.com/code/jhurley/walmart-supplychain-eda-kk/notebook>

<https://stackoverflow.com/>

https://www.w3schools.com/

Top of Form

Bottom of Form