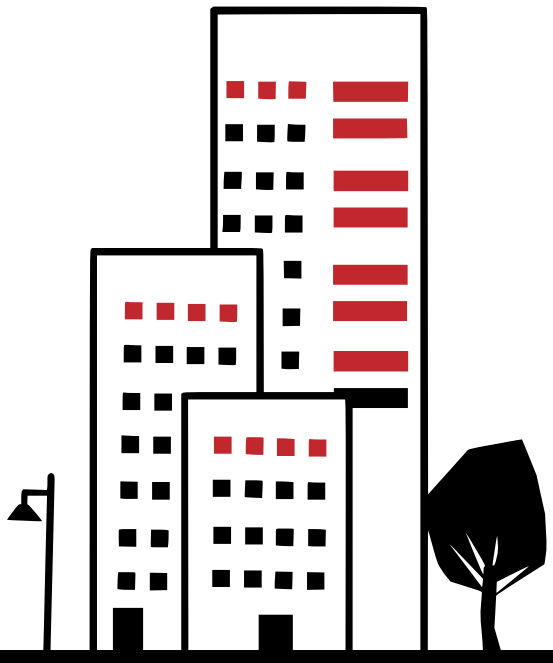
DataCo - Smart Supply Chain Dataset

PYTHON PROJECT

Project Team:

Tejasvi Bhasker
J. Sai Maansi
Lakshmipriya Anil
Anmol Agnihotri



INTROUCTION

This project focuses on analyzing a Supply Chain dataset, exploring key parameters such as transaction types, shipping details, customer information, and product characteristics. Our aim to extract actionable insights that can enhance business operations and contribute to the optimization of supply chain efficiency.

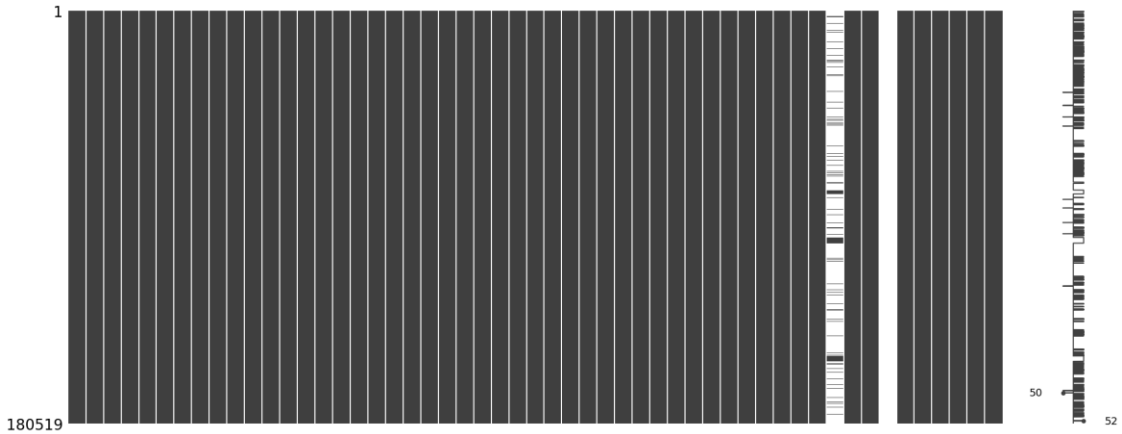
Type	Days for s	Days for s	Benefit pe	Sales per c	Delivery S	Late_deliv	Category I	Category I	Customer	Customer	Customer	Customer
DEBIT	3	4	91.25	314.64	Advance s	0	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Cally
TRANSFER	5	4	-249.09	311.36	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Irene
CASH	4	4	-247.78	309.72	Shipping c	0	73	Sporting G	San Jose	EE. UU.	XXXXXXXX	Gillian
DEBIT	3	4	22.86	304.81	Advance s	0	73	Sporting G	Los Angele	EE. UU.	XXXXXXXX	Tana
PAYMENT	2	4	134.21	298.25	Advance s	0	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Orli
TRANSFER	6	4	18.58	294.98	Shipping c	0	73	Sporting G	Tonawanc	EE. UU.	XXXXXXXX	Kimberly
DEBIT	2	1	95.18	288.42	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Constance
TRANSFER	2	1	68.43	285.14	Late delive	1	73	Sporting G	Miami	EE. UU.	XXXXXXXX	Erica
CASH	3	2	133.72	278.59	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Nichole
CASH	2	1	132.15	275.31	Late delive	1	73	Sporting G	San Ramo	EE. UU.	XXXXXXXX	Oprah
TRANSFER	6	2	130.58	272.03	Shipping c	0	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Germane
TRANSFER	5	2	45.69	268.76	Late delive	1	73	Sporting G	Freeport	EE. UU.	XXXXXXXX	Freya
TRANSFER	4	2	21.76	262.2	Late delive	1	73	Sporting G	Salinas	EE. UU.	XXXXXXXX	Cassandra
DEBIT	2	1	24.58	245.81	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Natalie
TRANSFER	2	1	16.39	327.75	Late delive	1	73	Sporting G	Peabody	EE. UU.	XXXXXXXX	Kimberley
DEBIT	2	1	-259.58	324.47	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Sade
PAYMENT	5	2	-246.36	321.2	Late delive	1	73	Sporting G	Canovana	Puerto Ric	XXXXXXXX	Brynne
CASH	2	1	23.84	317.92	Late delive	1	73	Sporting G	Paramoun	EE. UU.	XXXXXXXX	Ciara
DEBIT	2	1	102.26	314.64	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Bo
PAYMENT	0	0	87.18	311.36	Shipping c	0	73	Sporting G	Mount Pro	EE. UU.	XXXXXXXX	Kim
TRANSFER	0	0	154.86	309.72	Shipping c	0	73	Sporting G	Long Beac	EE. UU.	XXXXXXXX	Kellie
TRANSFER	5	4	82.3	304.81	Late delive	1	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Alma
TRANSFER	4	2	22.37	298.25	Late delive	1	73	Sporting G	Rancho Cc	EE. UU.	XXXXXXXX	Yeo
TRANSFER	3	2	17.7	294.98	Shipping c	0	73	Sporting G	Caguas	Puerto Ric	XXXXXXXX	Lucy

We reviewed the column names and their corresponding descriptions, carefully selecting the key columns essential for our analysis. This allowed us to filter out the important data fields that form the basis of our work.

FIELDS	DESCRIPTION
Type	: Type of transaction made
Days for shipping (real)	: Actual shipping days of the purchased product
Days for shipment (schedu	: Days of scheduled delivery of the purchased product
Benefit per order	: Earnings per order placed
Sales per customer	: Total sales per customer made per customer
Delivery Status	: Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time
Late_delivery_risk	: Categorical variable that indicates if sending is late (1), it is not late (0).
Category Id	: Product category code
Category Name	: Description of the product category
Customer City	: City where the customer made the purchase
Customer Country	: Country where the customer made the purchase
Customer Email	: Customer's email
Customer Fname	: Customer name
Customer Id	: Customer ID
Customer Lname	: Customer lastname
Customer Password	: Masked customer key
Customer Segment	: Types of Customers: Consumer , Corporate , Home Office
Customer State	: State to which the store where the purchase is registered belongs
Customer Street	: Street to which the store where the purchase is registered belongs
Customer Zipcode	: Customer Zipcode

Initial Data Exploration

We initiated our analysis by conducting a preliminary examination of the dataset to identify any missing values. To visually represent the distribution of missing data across different columns, we utilized the ‘missingno’ library and generated a matrix plot using the matrix function. In the resulting image, each row corresponds to a data point, and missing values are depicted as white lines, allowing for a quick overview of the completeness of our dataset.

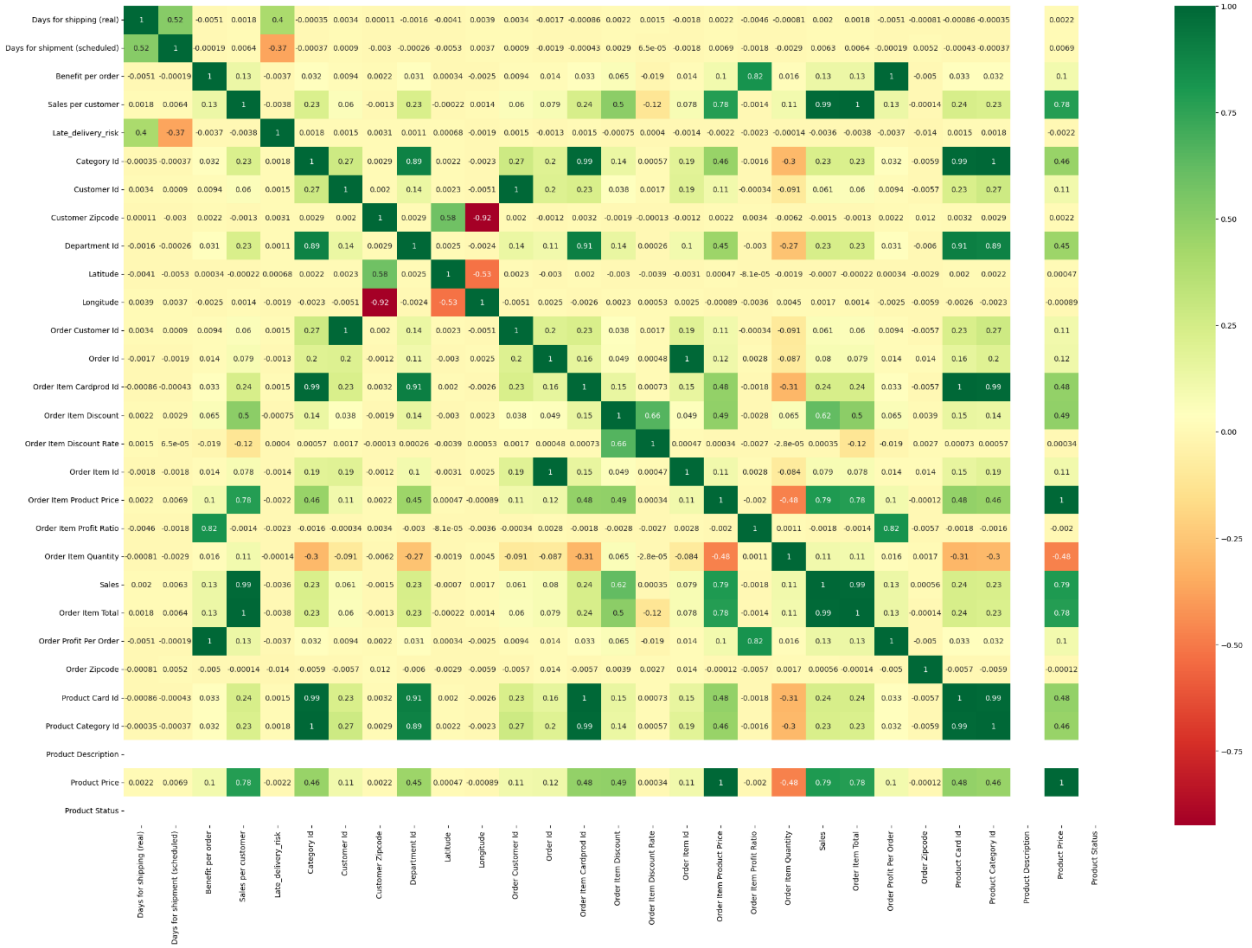


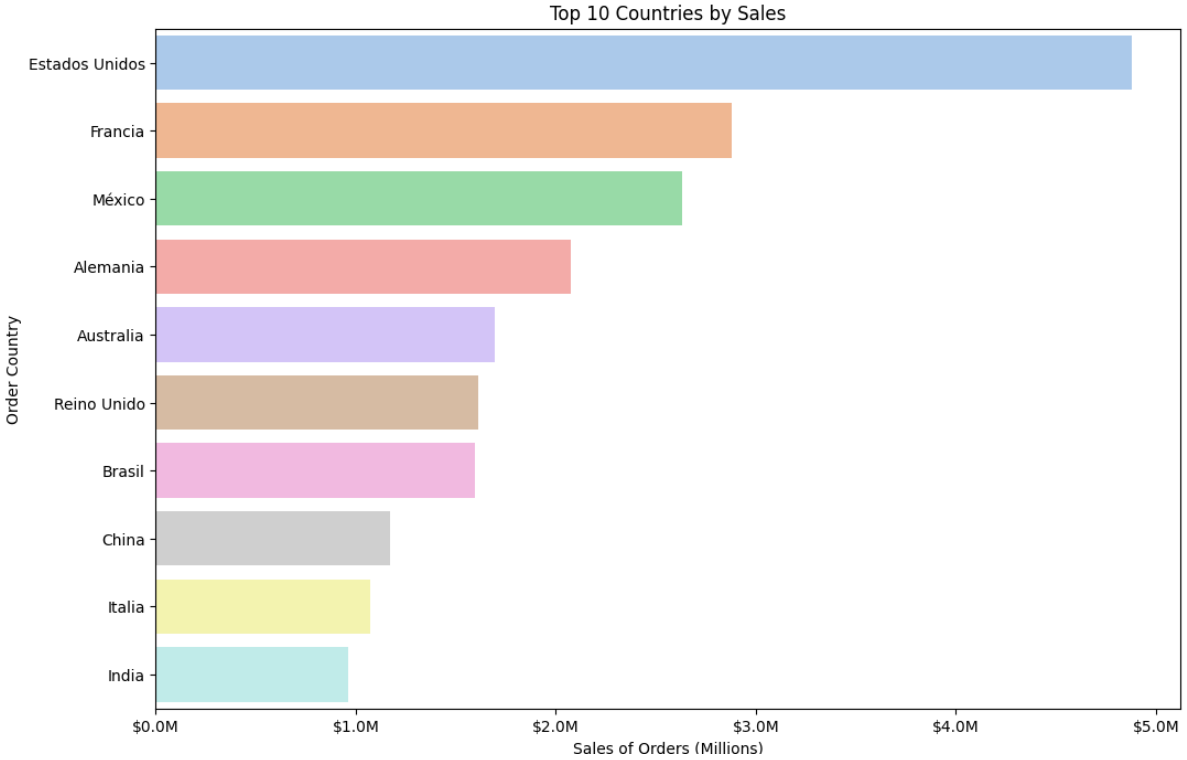
Upon noticing missing data in columns like 'Customer Lname,' 'Customer Zipcode,' 'Order Zipcode,' and 'Product Description,', we decided to remove the 'Order Zipcode' and 'Product Description' columns. And with only a few missing values in 'Customer Lname' and 'Customer Zipcode,' we chose to improve data completeness by removing the corresponding rows.

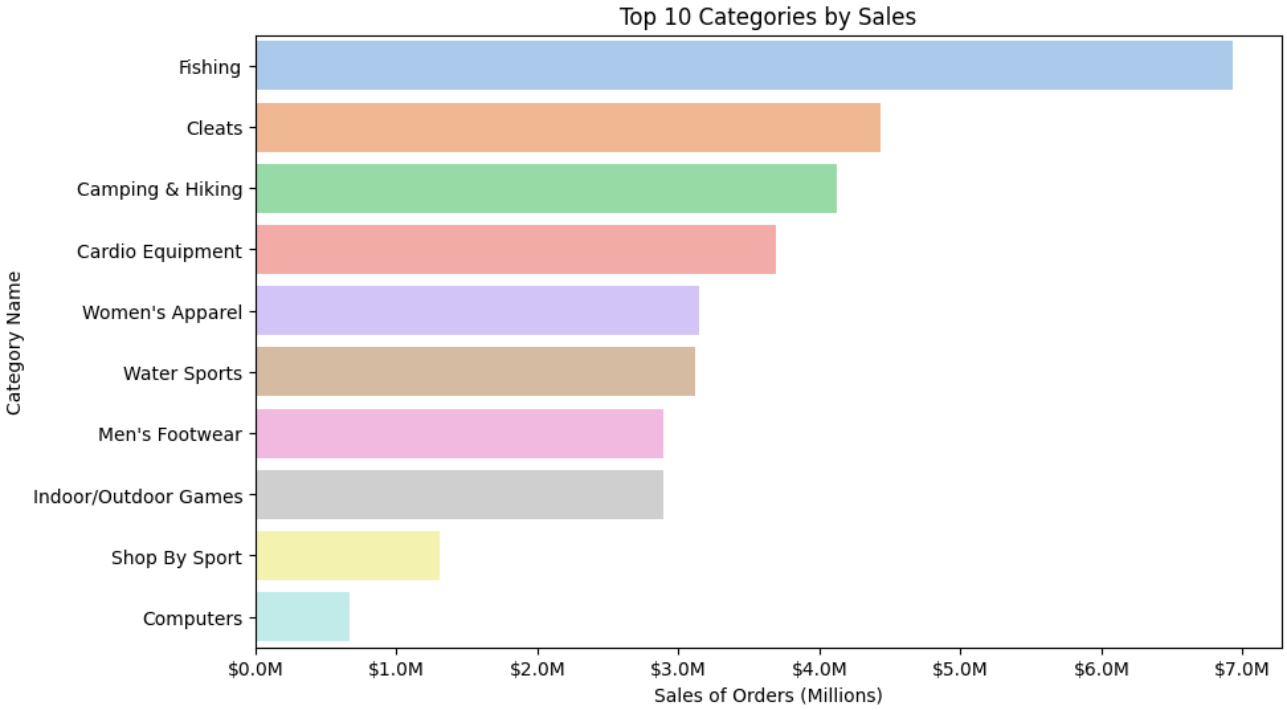
```
Customer Lname           8
Customer Zipcode         3
Order Zipcode       155679
Product Description 180519
dtype: int64
```

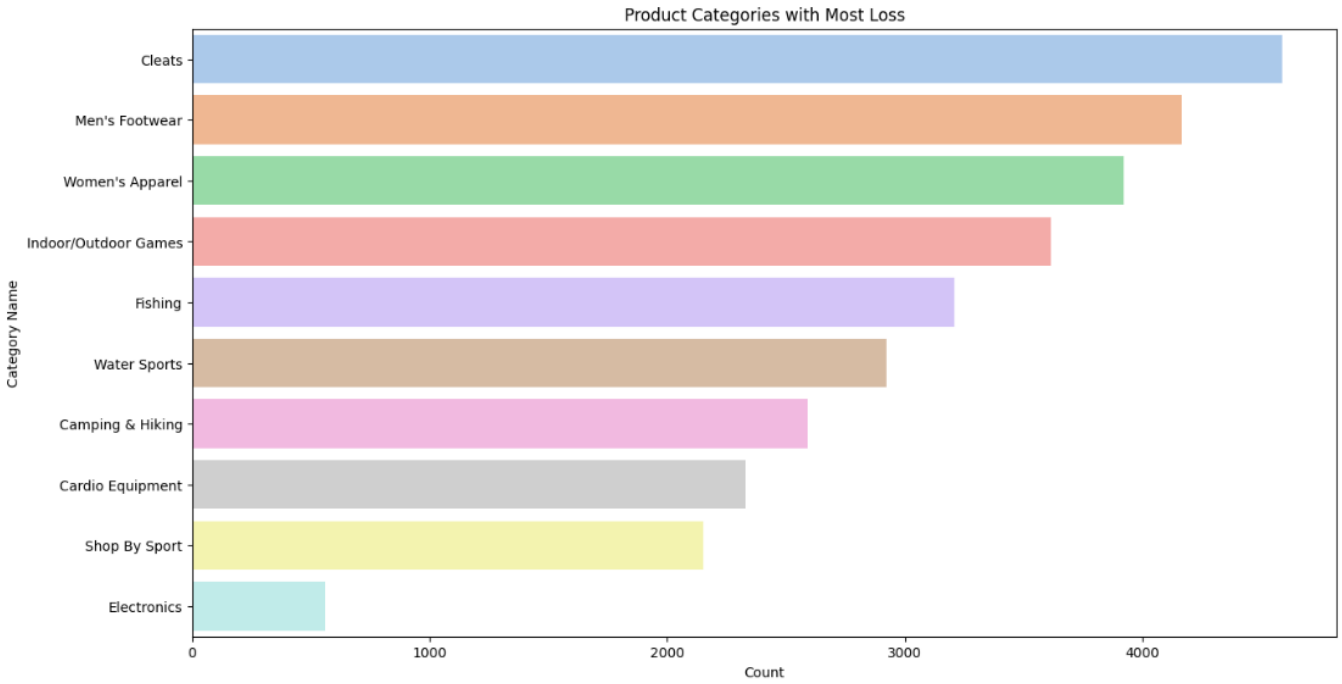
In our initial analysis, we employed key graphs to gain a comprehensive understanding of the dataset. The 'Top 10 Countries by Sale' and 'Top 10 Categories by Sale' charts were instrumental in identifying strong performers, shedding light on critical aspects of sales distribution. Simultaneously, we delved into potential areas for improvement through visualizations such as 'Products with Most Loss' and 'Regions with Most Loss,' pinpointing specific products and

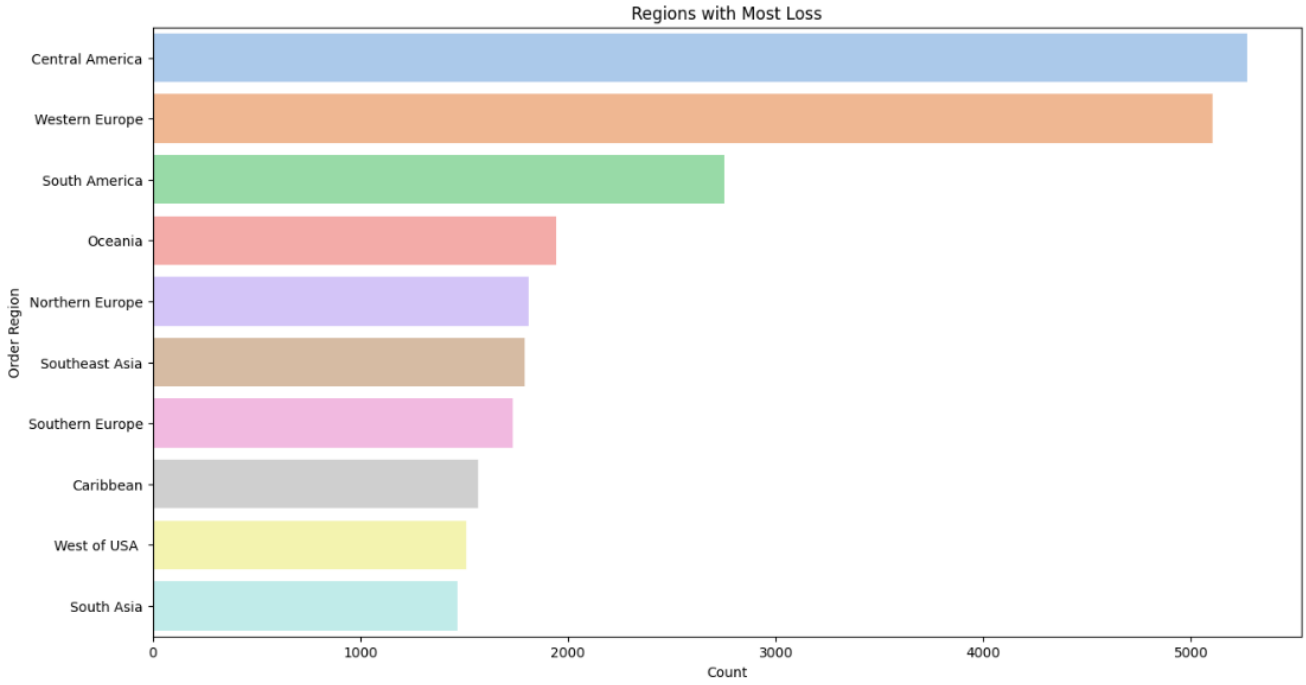
regions that demanded attention. To enhance the reliability of our findings, we meticulously examined the data for discrepancies, addressing issues like null values and incorrect data types. With 53 columns and 180,519 rows, we identified and rectified correlations among columns, successfully navigating through approximately 3.5 percent null values, primarily concentrated in the order zip code and product description columns. Moreover, we harmonized language inconsistencies in the country column, translating names from Spanish to English. This comprehensive approach ensures that our data is now primed for in-depth analysis, enabling us to draw meaningful insights and strategically improve overall performance




Top 10 Countries by Sales
Estados Unidos
Francia
México
Alemania
Australia
Reino Unido
Brasil
China
Italia
India
Order Country
$0.0M
$1.0M
$2.0M
$3.0M
$4.0M
$5.0M
Sales of Orders (Millions)

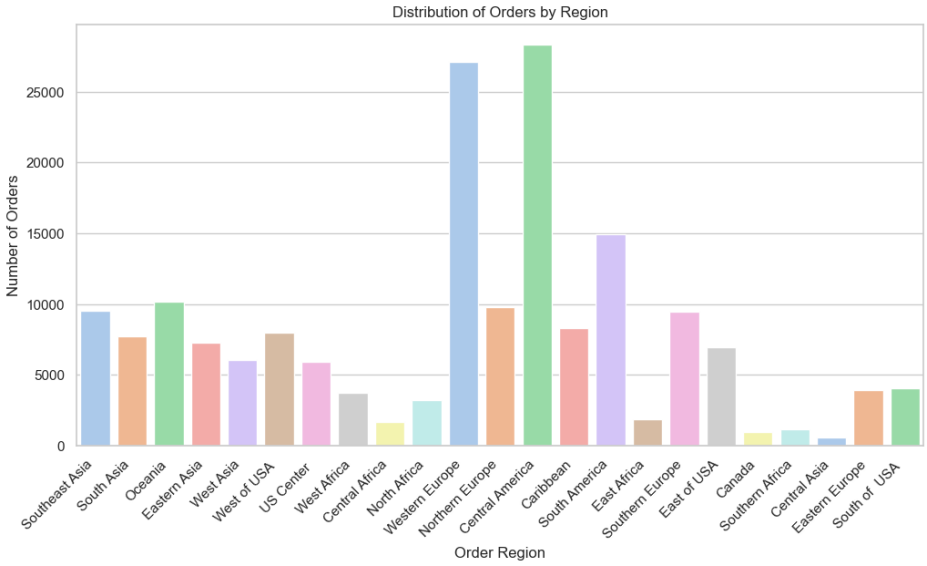

Top 10 Categories by Sales
Fishing
Cleats
Camping & Hiking
Cardio Equipment
Women's Apparel
Water Sports
Men's Footwear
Indoor/Outdoor Games
Shop By Sport
Computers
Category Name
$0.0M
$1.0M
$2.0M
$3.0M
$4.0M
$5.0M
$6.0M
$7.0M
Sales of Orders (Millions)


Product Categories with Most Loss
Cleats
Men's Footwear
Women's Apparel
Indoor/Outdoor Games
Fishing
Water Sports
Camping & Hiking
Cardio Equipment
Shop By Sport
Electronics
Category Name
0
1000
2000
3000
4000
Count


Regions with Most Loss
Central America
Western Europe
South America
Oceania
Northern Europe
Southeast Asia
Southern Europe
Caribbean
West of USA
South Asia
Order Region
0
1000
2000
3000
4000
5000
Count

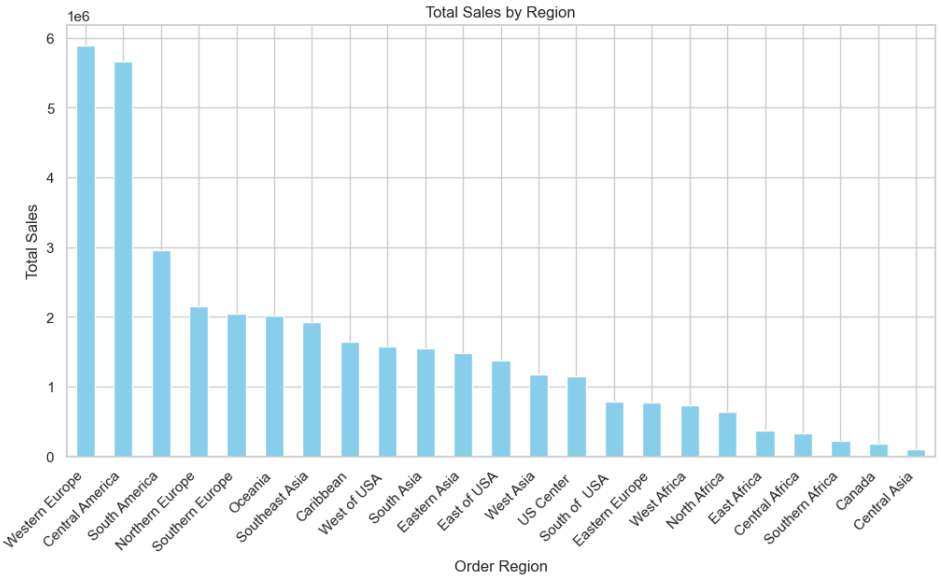
Distribution of Orders by Region:

Utilizing a count plot, we examined the distribution of orders across different regions. This visual provided a quick overview of order frequency in each region.



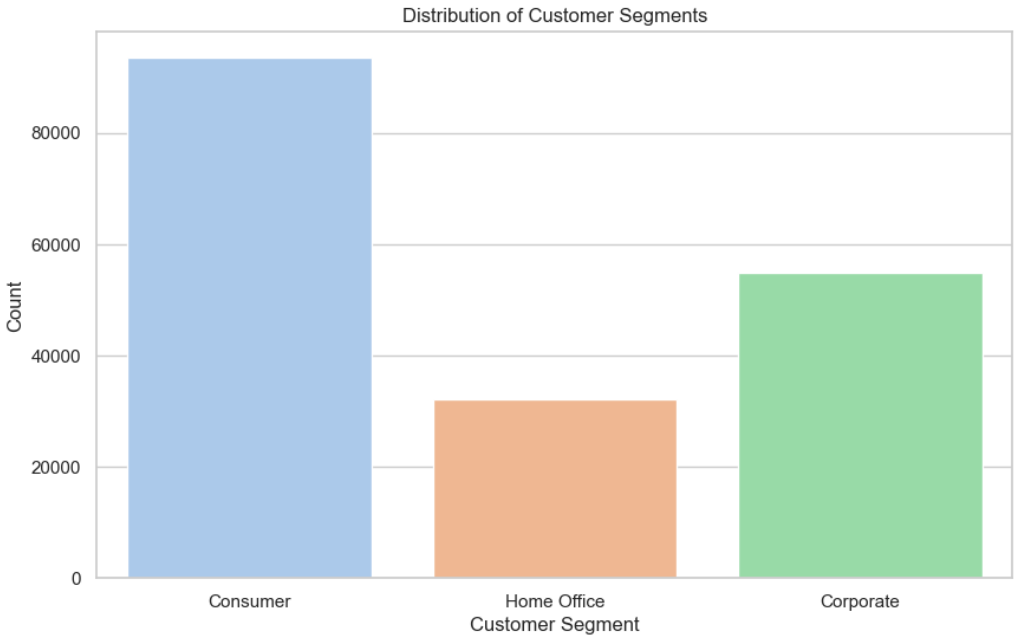
Total Sales by Region:

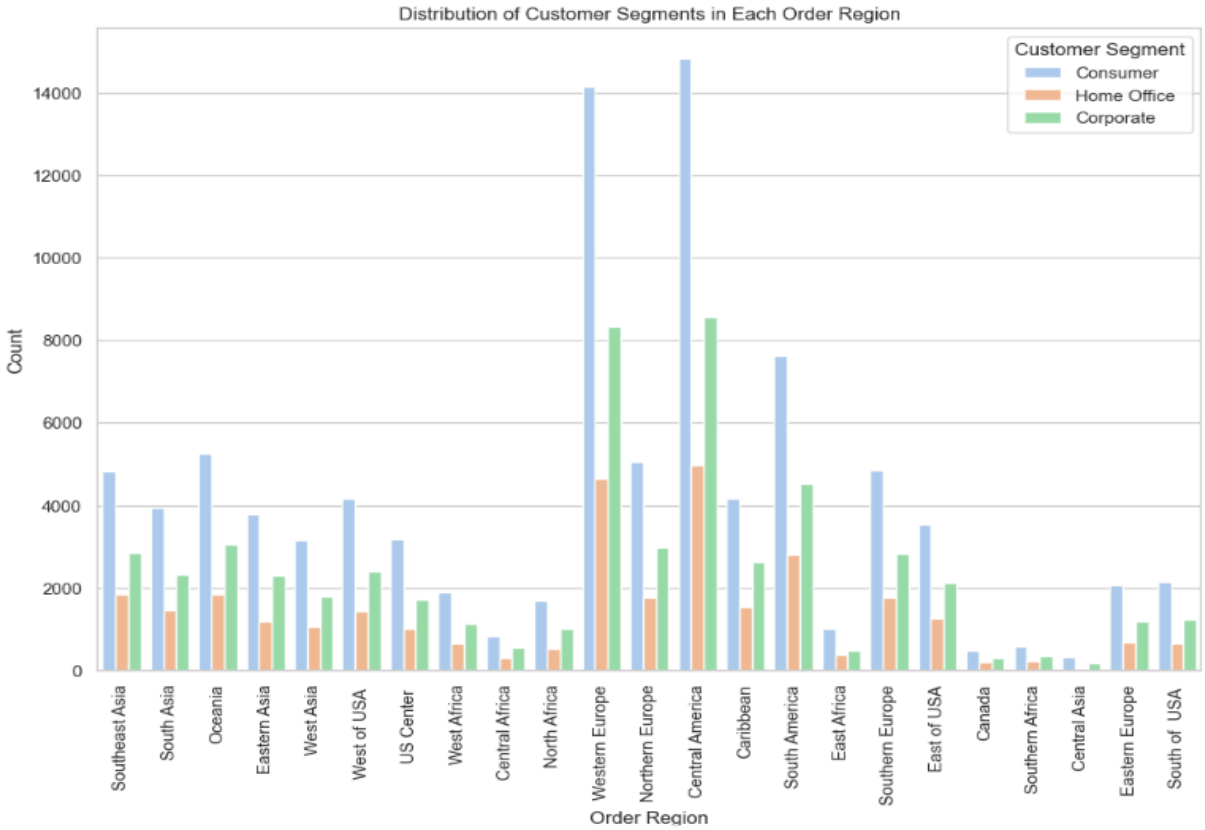
Through a bar chart, we delved into the total sales across various regions. This analysis highlighted the regions contributing significantly to overall sales.



Distribution of Customer Segments/ By Order Region:

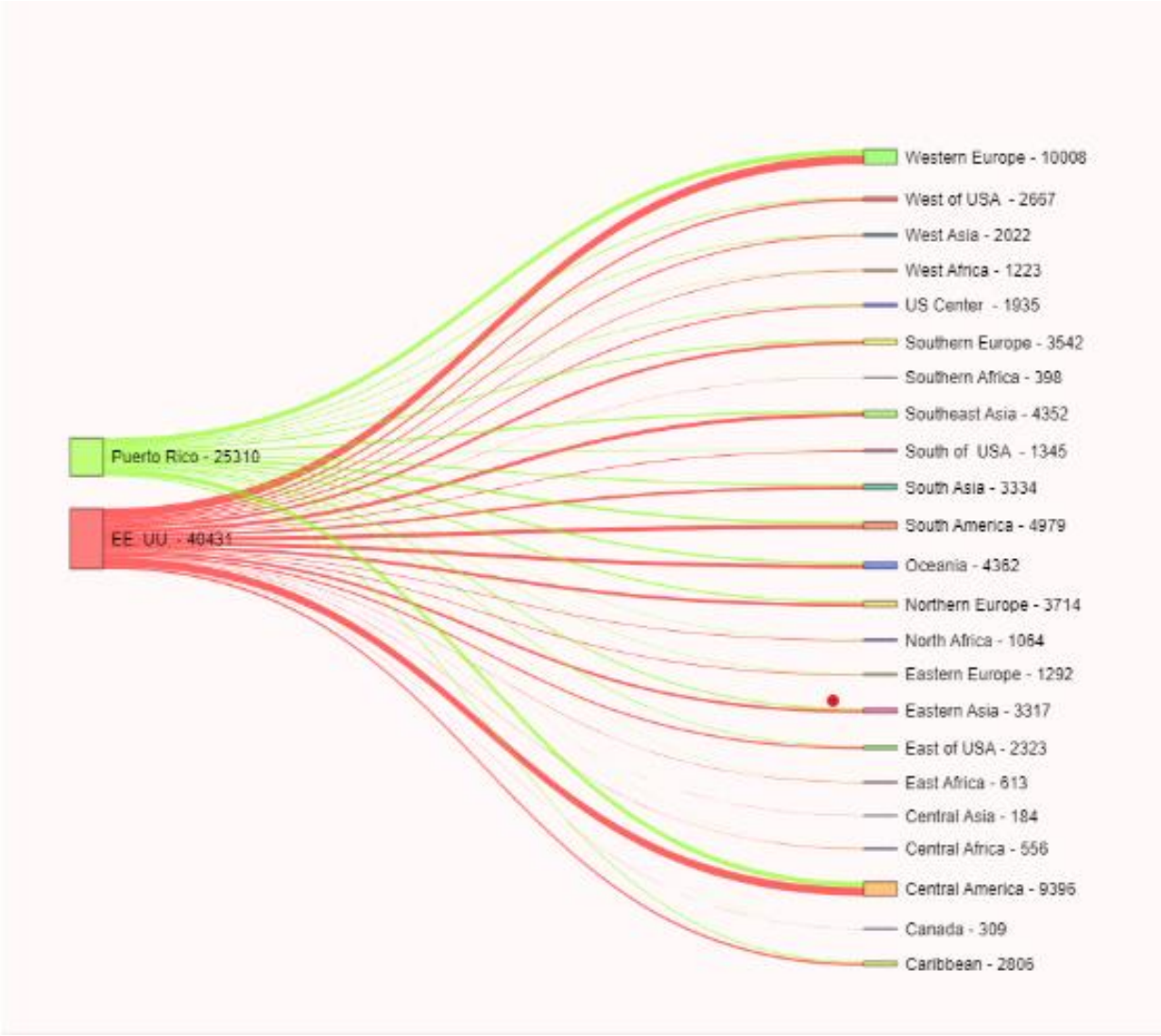
Another count plot focused on understanding the distribution of customer segments, shedding light on the prevalence of different customer categories and distribution of customer segments within each order region.





Sankey Flow of Orders:

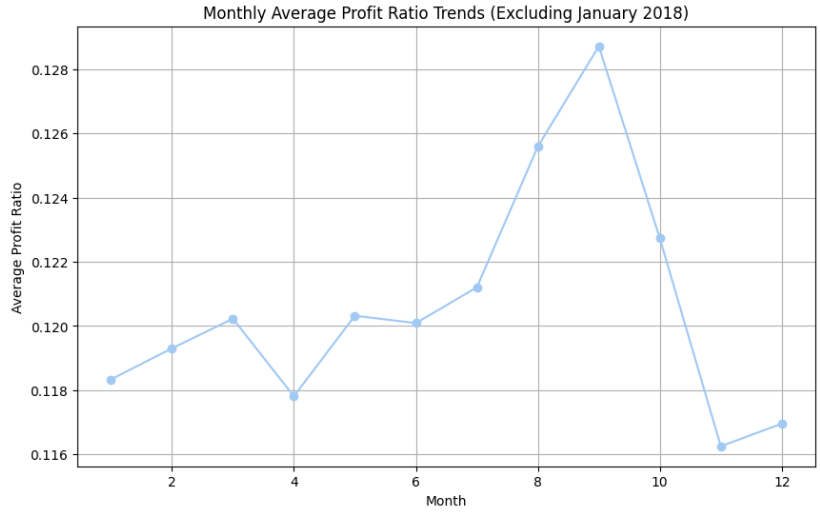
Using a Sankey diagram, we visualized the flow of orders from customers' countries to delivery regions. This representation offered a comprehensive understanding of the pathways through which orders traverse various regions.

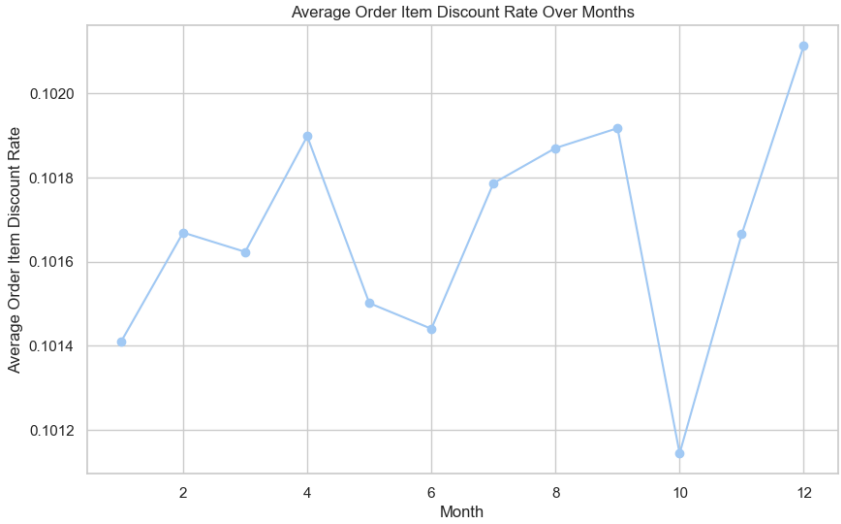


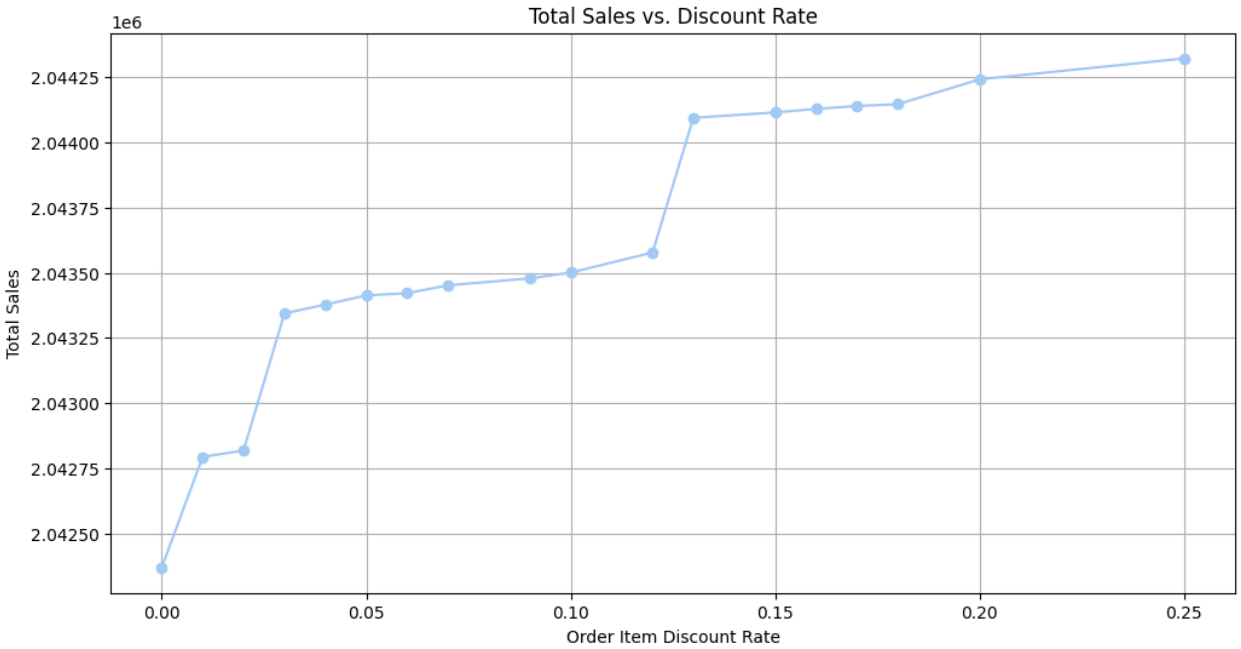
INSIGHTS

Financial Analysis

In our analysis, we carefully looked at how profits, sales, and discount rates changed over the months across several years. We wanted to find patterns that could help us understand how our business performs over time. We noticed that profit ratios were highest in month 9 but dipped between months 11 and 12. At the same time, discount rates were highest in month 12, showing a big increase in discounts towards the end of the year. Sales stayed steady, but we saw a dip towards the end of the year. Likewise, the sales vs discount rate also showed that sales tended to increase with discount rate offered but only up to a limit, post which it tends to flatten despite discount increase.



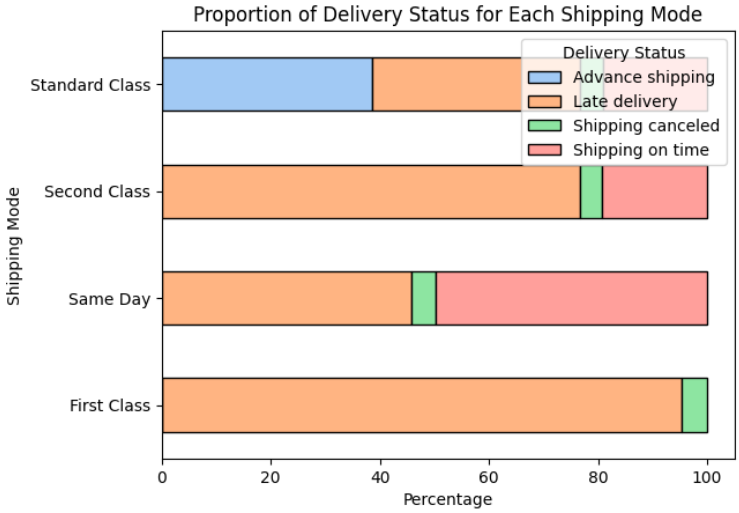


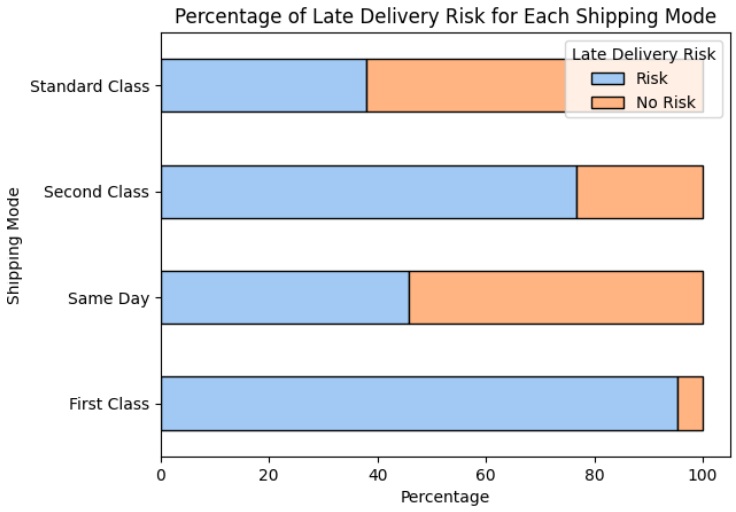


Actionable Insight: Based on the data analysis, it is recommended that sellers focus on maintaining discounts within the range of 10% to 15% to effectively enhance both profit ratio and sales. This suggests a strategic approach to pricing that can be implemented to achieve the best results. Sellers should carefully evaluate their discounting strategies, ensuring they fall within this identified sweet spot to strike the right balance between driving sales and maximizing profitability. Regularly monitoring and adjusting discounts within this range can lead to a sustainable and optimal pricing strategy, allowing sellers to capitalize on the positive correlation between discounts, profit ratio, and sales without risking potential diminishing returns.

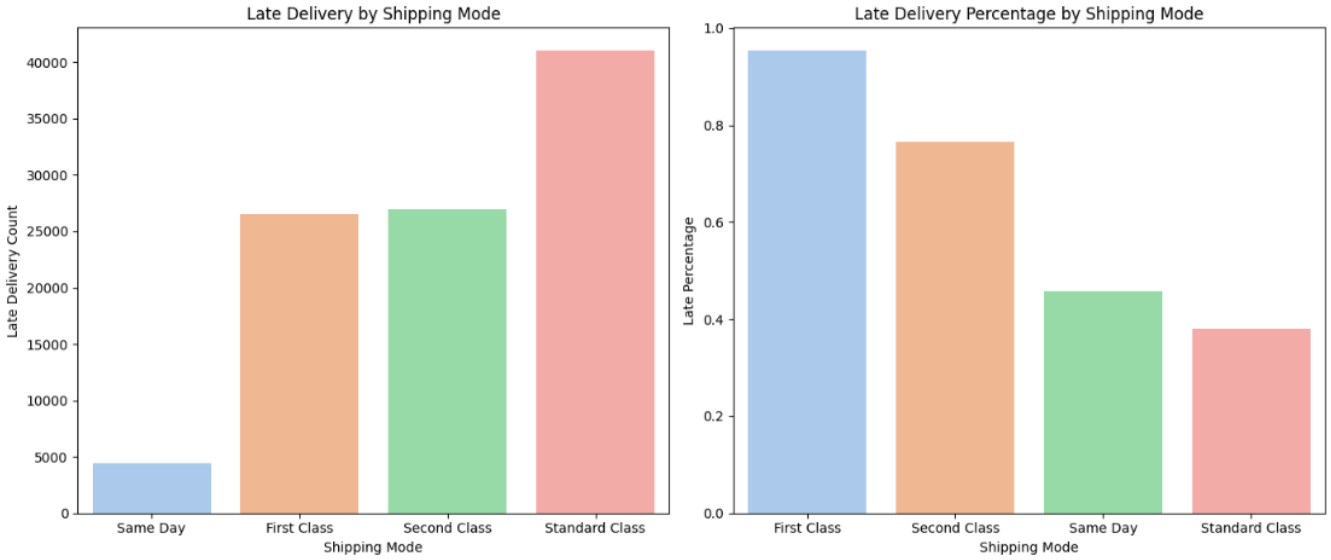
Analysis of Late Deliveries

Given the critical importance of timely deliveries in the supply chain domain, we conducted a detailed examination to assess the variability in the risk of late deliveries across different shipping methods. Our findings unveiled those first-class shipments exhibited the highest likelihood of late deliveries, contrasting with the standard class, which demonstrated the lowest risk. Notably, the standard class not only emerged as the most reliable in terms of on-time deliveries but also stood out as the only class incorporating advanced delivery practices. To further visualize and comprehend these insights, we plotted graphs depicting the correlation between late deliveries and shipping modes, as well as the relationship between delivery status and various shipping methods. These graphical representations provide a comprehensive view of the performance disparities among shipping classes and serve as valuable tools for optimizing supply chain logistics and enhancing overall delivery efficiency.





Actionable Insight: Given the alarming 95 percent late delivery rate associated with First Class shipments, businesses should promptly investigate the underlying causes and strategically shift focus to alternative methods, particularly Standard Class, which shows a lower risk of delays. Adopting successful advanced delivery practices from Standard Class across all shipping methods is recommended.



Continuous monitoring and adaptation of shipping strategies based on real-time data are crucial for an efficient and resilient supply chain. While certain items may require First Class shipping, a strategic evaluation of the trade-off between expedited delivery benefits and associated risks should guide item allocation for a balanced approach to meeting customer delivery expectations.

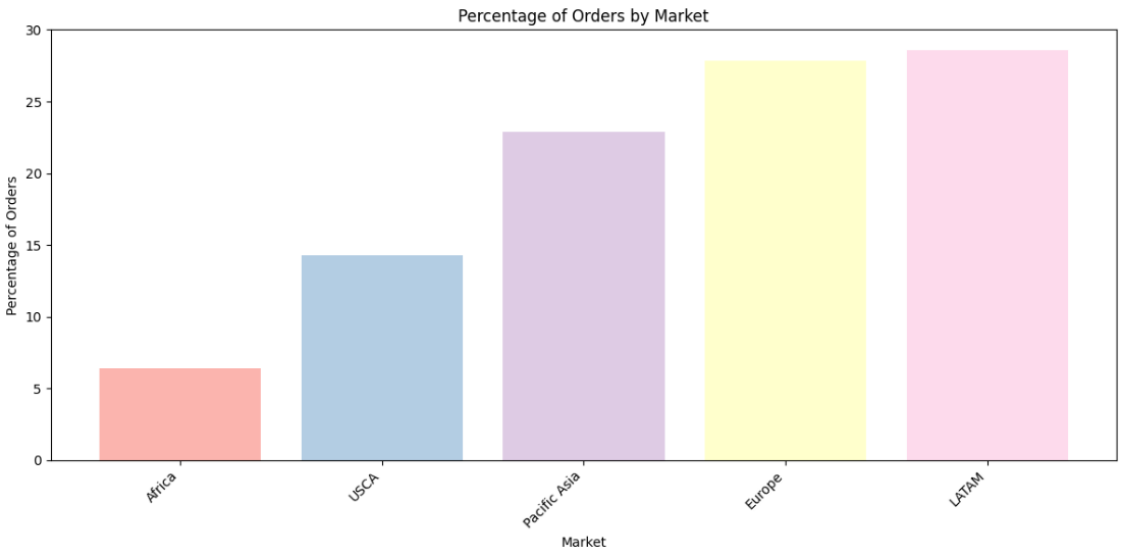
Market Analysis

Market and Order Region analysis plays a crucial role in understanding and optimizing various aspects of business operations. It provides valuable insights into the geographical distribution of customer activities and sales transactions thereby offering businesses the opportunity to make informed decisions that can impact overall performance, profitability, and customer satisfaction.

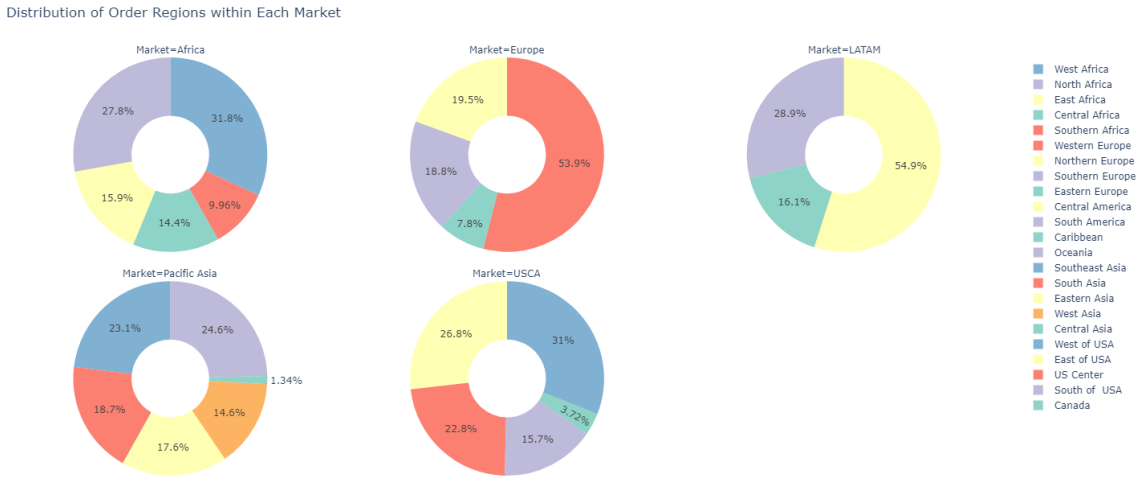
By examining the specific markets or geographical regions where a business operates, the business can also derive valuable insight into the current market penetration as well as identify regions for future expansion.

Order Region analysis also allows businesses to track and understand where their products or services are in demand thus enabling effective inventory management, logistics optimization, and customer service enhancements.

From the given data, it can be seen that 28.58 % of purchases are from LATAM Market closely followed by Europe at 27.84 %. These values are considerably higher that the USCA Market which consists of the domestic purchases as well. It can also be noted that Africa is the least performing market as of now, standing at only 6.43 % and the causation of this should be identified by the Business to improve future sales.



When the distribution of Order Regions within each Market is analyzed, then it can be seen that West and North Africa together constitute 59.6 % of the purchases from African Market.



Similarly, Western Europe and Northern Europe constitute 53.9 % and 19.5 % respectively within the European Market. 54.9 % of the purchases from LATAM can be attributed to Central America alone.

When the Asia-Pacific Market is analyzed in detail, then it can be observed that the purchases from the upcoming developing regions such as South and Southeast Asia accumulate a relatively lower market share of 23.1 % and 18.7 % respectively when compared to Oceania with a slightly higher average purchasing power and comprising of Commonwealth countries such as Australia and New Zealand. It can also be observed at trade to Central Asia stands at an abysmally low value of 1.34 %.

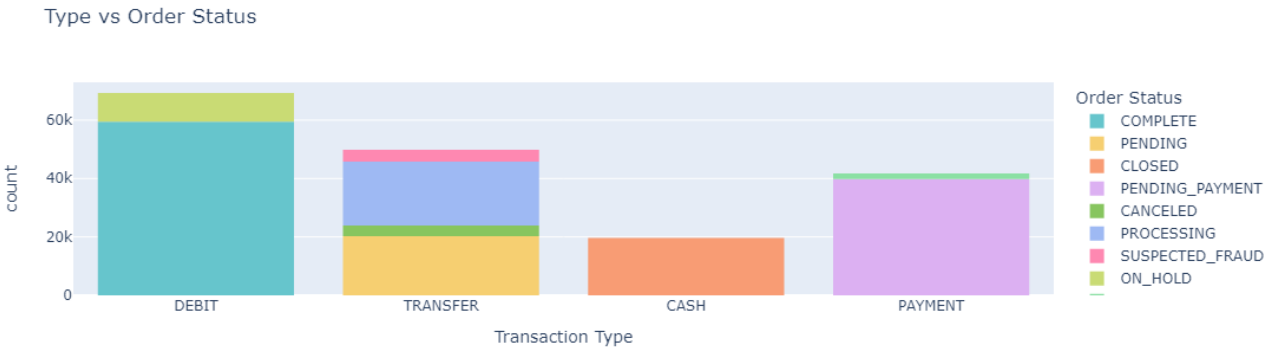
In the USCA market, the domestic trade cumulates to a total of 96.3 % with West, East, Central and South USA totaling 31 %, 26.8 %, 22.8 % and 15.7 % respectively. The international trade to Canada scores a pitiful 3.72 % despite being the closest neighbour of the source country.

Fraud Analysis

The impact of fraud in supply chain can be monumental. Long-term repercussions can include a loss of business, and potentially, a loss of market share. Payment fraud may involve unauthorized transactions made using either false or stolen payment information and it is imperative to come up with strategies to counter or minimize the repercussions due to fraud.

In order to take targeted actions to achieve that, we need to identify the source of potential fraud transactions so that they can be flagged or put under surveillance.

For the given data, we first analyzed the Transaction Type and the Order Payment Status and identified that that all the occurrences of Suspected Payment Fraud for the given timeline were found to have been conducted during “Transfer” mode of payment.

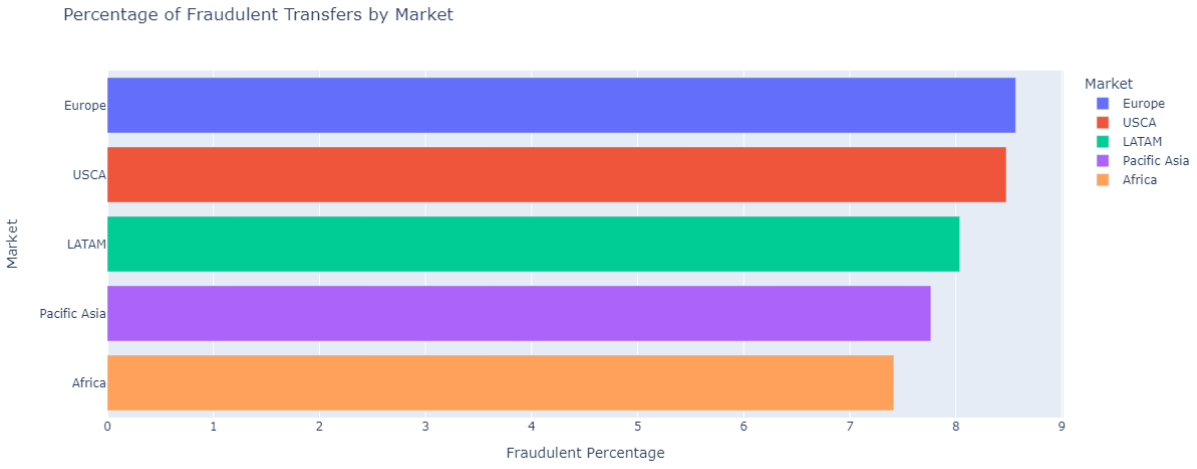


On summarizing transfer transactions across different market regions, it can be observed that Europe has the highest the percentage of fraudulent transfers is at 8.56%. While the USCA Market exhibits a nearly half the transfer volume as Europe, the percentage of fraudulent transfers is still comparable to that of Europe, standing at 8.48%.

	Market	Total Transfers	Fraud Transfers	Fraudulent Percentage
1	Europe	13685	1172	8.564121
4	USCA	7103	602	8.475292
2	LATAM	14707	1182	8.036989
3	Pacific Asia	11192	869	7.764475
0	Africa	3196	237	7.415519

Latin America has substantial total transfer count at 14707 and the percentage of fraudulent transfers is slightly lower than Europe and USCA but still considerable at 8.04%. The same can be said about Pacific Asia.

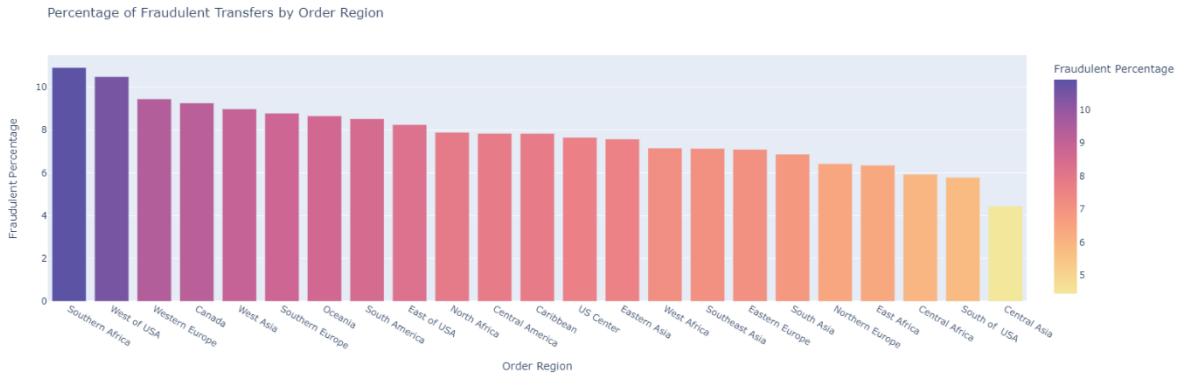
Africa has the lowest total transfer count among the analyzed regions but the percentage of fraudulent transfers is relatively high at 7.42%. This suggests that despite a lower transaction volume, there is a noteworthy risk of fraud in the African market.



Actionable Insight: While Europe, USCA, and LATAM have higher transaction volumes, they also exhibit higher percentages of fraudulent transfers warranting a focused approach to fraud prevention. Pacific Asia and Africa, despite having comparatively lower transaction volumes, still face a considerable risk of fraudulent activities emphasizing the need for robust security measures in these regions as well.

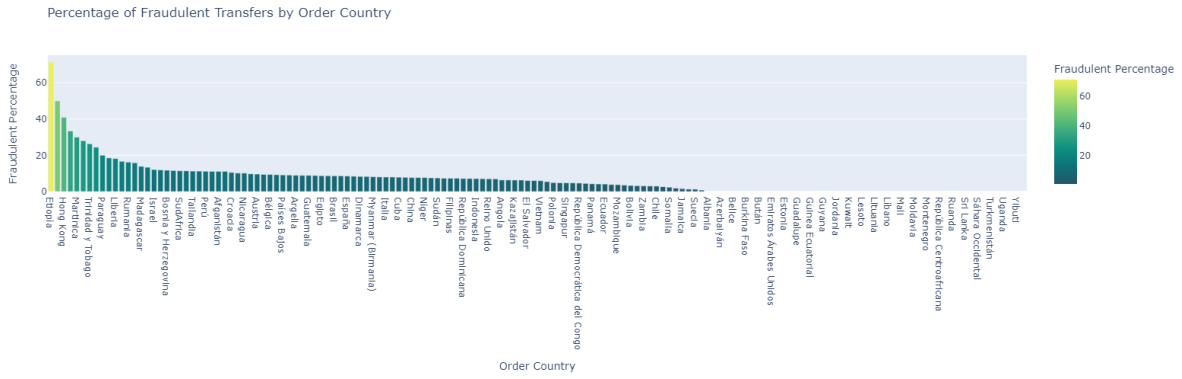
	Order Region	Total Transfers	Fraud Transfers	Fraudulent Percentage
16	Southern Africa	293	32	10.921502
21	West of USA	2249	236	10.493553
22	Western Europe	7457	705	9.454204
0	Canada	313	29	9.265176
20	West Asia	1636	147	8.985330
17	Southern Europe	2540	223	8.779528
11	Oceania	2644	229	8.661120
12	South America	4236	361	8.522191
6	East of USA	1818	150	8.250825
9	North Africa	950	75	7.894737
3	Central America	8047	631	7.841432
1	Caribbean	2424	190	7.838284
18	US Center	1581	121	7.653384
7	Eastern Asia	2058	156	7.580175
19	West Africa	993	71	7.150050
15	Southeast Asia	2622	187	7.131960
8	Eastern Europe	1043	74	7.094919
13	South Asia	2097	144	6.866953
10	Northern Europe	2645	170	6.427221

Going a level deeper and analyzing the percentage of fraud transfer transactions across Order Regions reveals that Southern Africa and West of USA exhibit double-digit percentages of fraudulent transfers indicating significant fraud risks. This is the case for Western Europe despite having a high transfer volume. Canada and West Asia also demonstrates a relatively high percentage risk of fraudulent transfers.



Using these region-specific insights, businesses should tailor their fraud prevention to mitigate risks and ensure the integrity of financial transactions. Utilizing Order Region-specific analysis provides a broader perspective thus enabling the identification of trends and risks that may span multiple countries within a geographic region.

This analysis can be complemented by doing a country-specific scrutiny of the fraudulent transfers to gain a comprehensive understanding of the varied fraud landscape. Country-specific analysis is essential to comprehend the unique socio-economic, regulatory and cultural factors contributing to fraud risks within individual nations. It allows businesses to tailor their fraud prevention strategies based on localized insights, considering factors such as payment infrastructure, legal frameworks and specific fraud patterns prevalent in each country. While Country-specific analysis allows for targeted interventions, Order Region-specific analysis helps in identifying regional trends and implementing broader strategies.

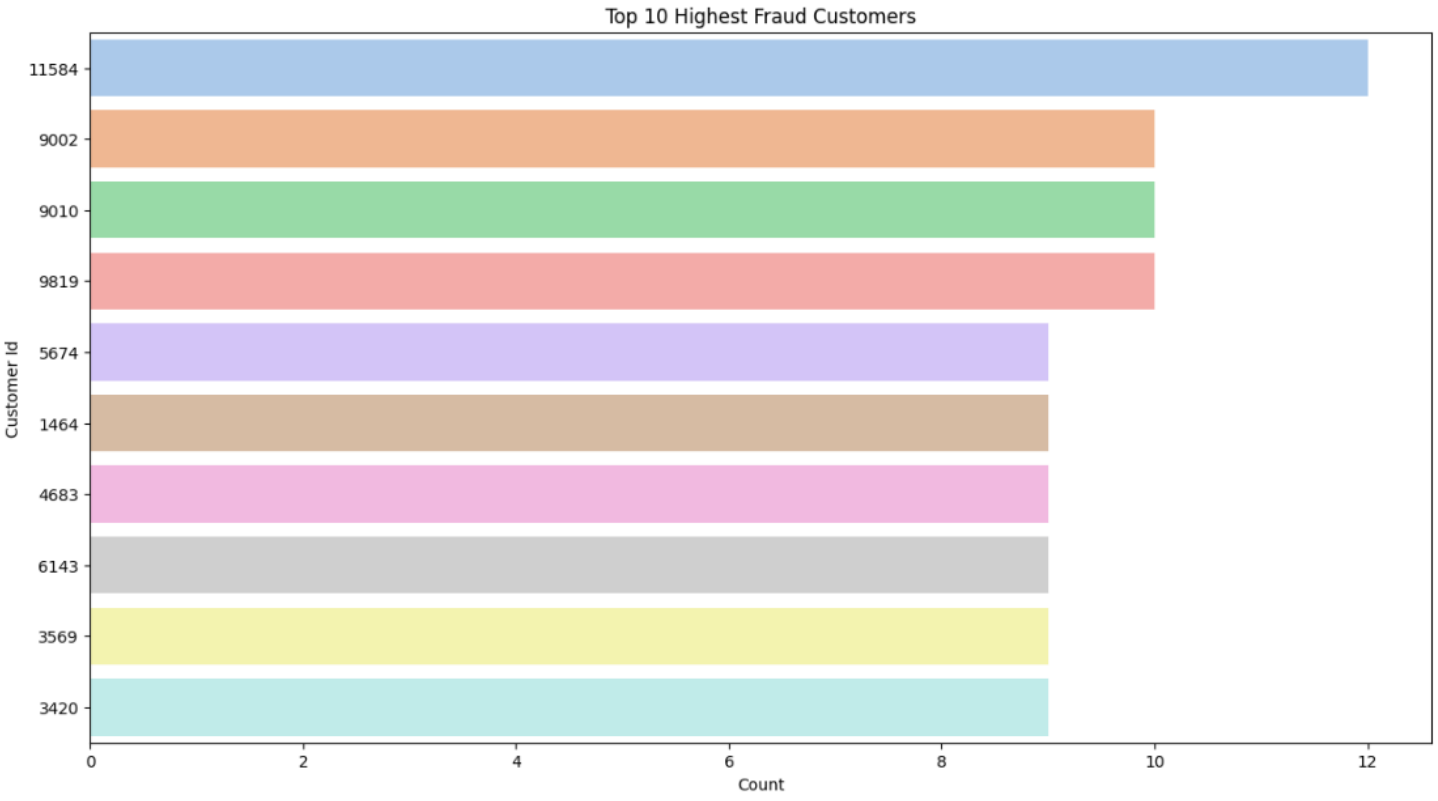


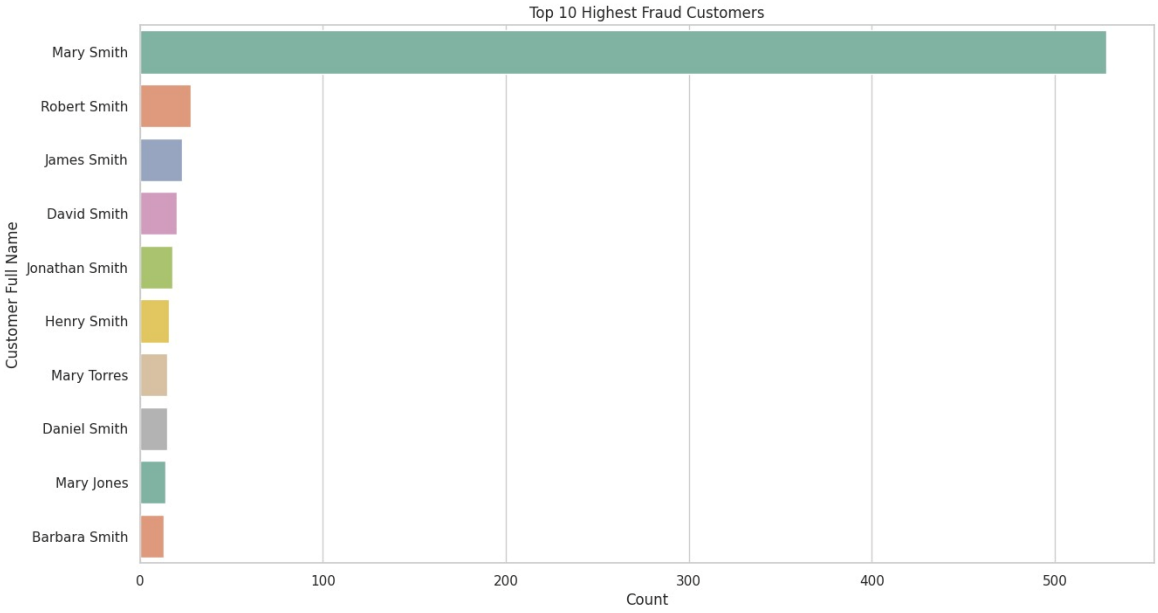
Actionable Insight: From the above graph, it is evident that Ethiopia, Guinea, and Hong Kong exhibit particularly high percentages of fraudulent transfers with Ethiopia at an alarming 71.43% fraudulent percentage.

Several countries such as Papua New Guinea, Martinique, Syria and Trinidad and Tobago show moderate fraud percentages warranting attention to fraud prevention measures. Portugal, Yemen, Qatar and Romania demonstrate varied but noteworthy fraud percentages indicating potential challenges in these European and Middle Eastern regions. Israel and Iraq in the Middle East, Venezuela in Latin America as well as Kenya, South Africa and Cameroon in Africa stand out suggesting challenges that may require focused countermeasures.

Customer Fraud Detection

Fraud is a significant risk in business, leading to potential financial losses. Our examination of fraudulent activities, focusing on Customer ID and Name, revealed noteworthy patterns. Specific Customer IDs displayed recurring instances of fraud, surpassing 10 occurrences, indicating a concerning trend. Particularly striking was the discovery of Mary Smith, associated with over 500 instances of fraud, amounting to total sales exceeding 100,000 dollars. To visualize these findings, we created graphs highlighting the Top 10 fraud customers by both Customer ID and Name. These visuals offer a clear overview of prolific fraud perpetrators.



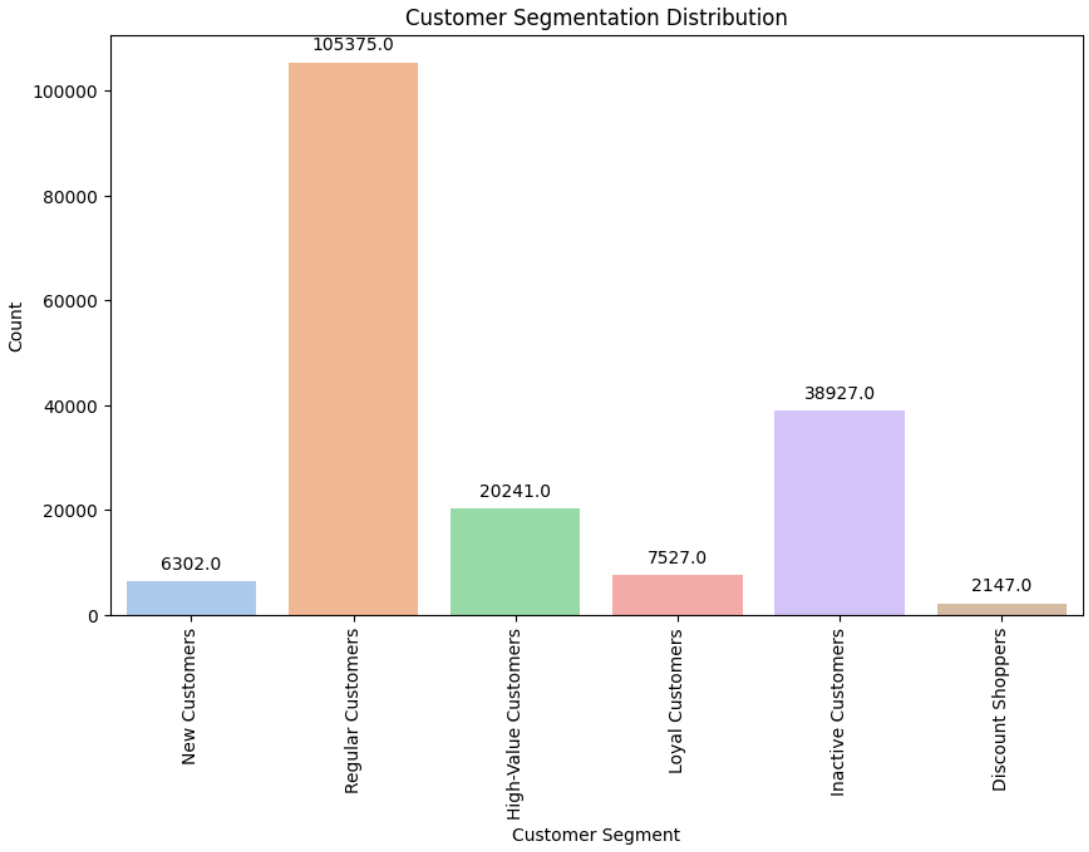


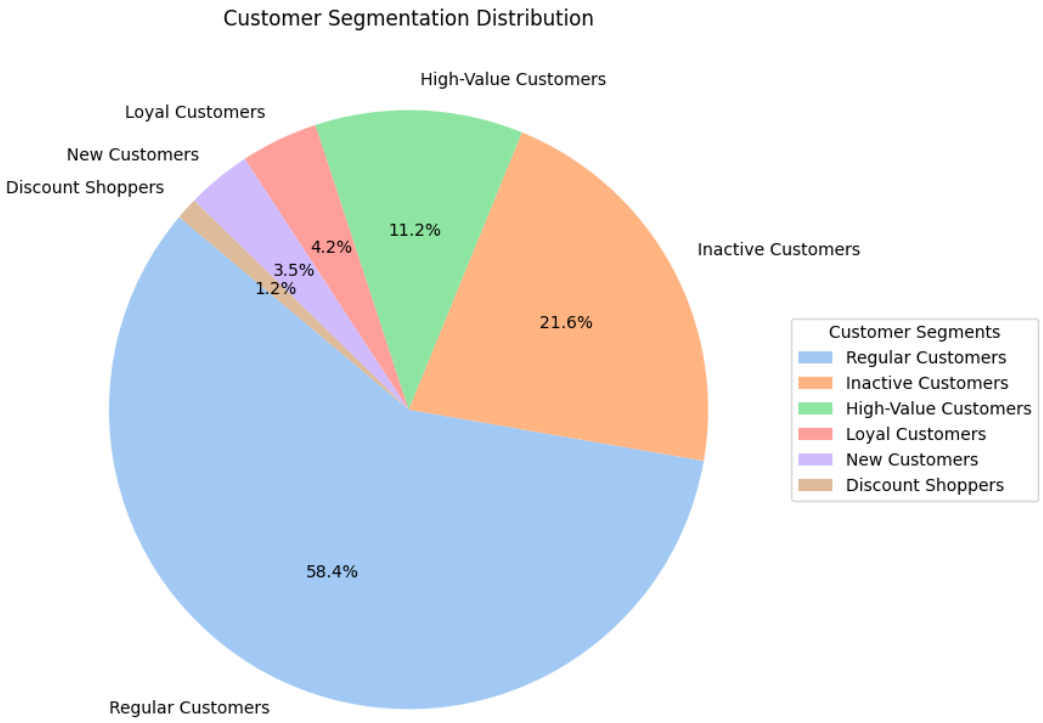
Actionable Insight: Since the name Mary Smith is associated with fraud more than 500 times, only two scenarios arise. The first is that Mary Smith are all different people, the probability of which will be fairly low since this name is associated with 500+ counts of fraud. The second is the far more probable one, which is that Mary Smith is a single person committing fraud by registering orders from different locations (thereby having different Customer Ids). The businesses should thus have more stringent measures while verifying identity to prevent fraud of this type. Note that Mary Smith alone is responsible for being involved in fraudulent sales of 100k, thus doing so is paramount for reducing fraud in the future.

RFM ANALYSIS

Incorporating RFM analysis into our project was crucial for businesses as it provided a nuanced understanding of customer behavior, enabling effective segmentation based on recency, frequency, and monetary factors.

The business faces a notable risk with 21.6% of customers classified as inactive. Re-engaging this segment is crucial to prevent revenue loss and maintain brand loyalty. Additionally, focusing on converting the 1.2% discount shoppers and 3.5% new customers into regular or high-value patrons presents growth opportunities. The 4.2% of loyal customers should not be overlooked, as they contribute to a stable revenue stream. Allocating resources strategically to these customer segments is vital for mitigating potential risks and ensuring sustainable business growth.





Actionable Insight: To tackle the risk linked to the 21.6% inactive customer segment, deploying targeted re-engagement strategies is necessary. Introducing personalized incentives, such as exclusive discounts or tailored promotions, is a good way to draw dormant customers back. Simultaneously, gaining insights into their inactivity via surveys or feedback mechanisms enables a nuanced approach in customizing these incentives to address specific concerns. Consistent communication channels, like well-crafted email campaigns and exclusive promotions, are also crucial in reminding inactive customers of the distinct value and benefits associated with continued engagement.

Furthermore, an enhanced overall customer experience, achieved through improved services or refined products, not only adds intrinsic value but also contributes significantly to reigniting interest. By actively addressing the unique needs and concerns of the inactive customer base, the business can work towards preserving this segment, effectively mitigating the identified risk of potential revenue erosion and customer attrition.