E- COMMERCE DATA ANALYSIS

BY – NIDA ZAKI

PROBLEM STATEMENT:

Given a comprehensive dataset encompassing user interactions, product details, order information, and demographic profiles in an e-commerce platform, the objective is to derive actionable insights to enhance business performance and customer experience. The analysis aims to understand user behavior, product performance, sales and revenue trends, order fulfillment efficiency, and demographic trends over time. By leveraging data-driven approaches, the goal is to identify opportunities for optimizing marketing strategies, improving product offerings, streamlining operational processes, and enhancing overall customer satisfaction.

```
users_df.head()
```

	id	first_name	last_name	email	age	gender	state	street_address	postal_code	city	country	latitude	longitude	traffic
0	457	Timothy	Bush	timothybush@example.net	65	M	Acre	87620 Johnson Hills	69917-400	Rio Branco	Brasil	-9.945568	-67.83561	
1	6578	Elizabeth	Martinez	elizabethmartinez@example.com	34	F	Acre	1705 Nielsen Land	69917-400	Rio Branco	Brasil	-9.945568	-67.83561	
2	36280	Christopher	Mendoza	christophermendoza@example.net	13	M	Acre	125 Turner Isle Apt. 264	69917-400	Rio Branco	Brasil	-9.945568	-67.83561	
3	60193	Jimmy	Conner	jimmyconner@example.com	64	M	Acre	0966 Jose Branch Apt. 008	69917-400	Rio Branco	Brasil	-9.945568	-67.83561	
4	64231	Natasha	Wilson	natashawilson@example.net	25	F	Acre	20798 Phillip Trail Apt. 392	69917-400	Rio Branco	Brasil	-9.945568	-67.83561	

```
distribution_centers_df.shape
```

```
(10, 4)
```

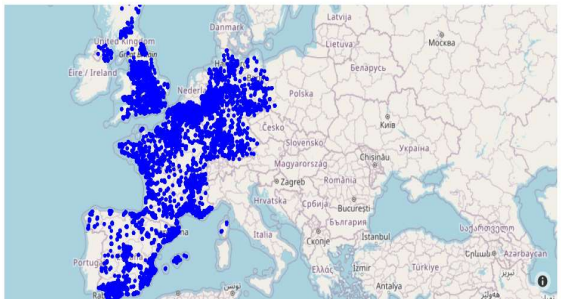
```
distribution_centers_df.head()
```

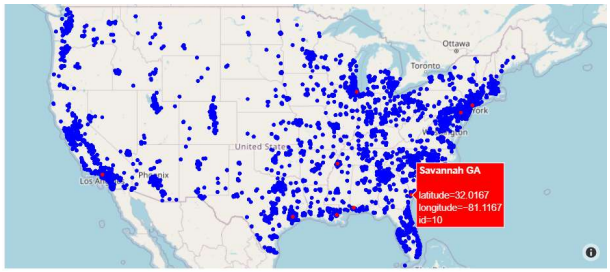
	id	name	latitude	longitude
0	1	Memphis TN	35.1174	-89.9711
1	2	Chicago IL	41.8369	-87.6847
2	3	Houston TX	29.7604	-95.3698
3	4	Los Angeles CA	34.0500	-118.2500
4	5	New Orleans LA	29.9500	-90.0667

Geospatial Analysis:

Utilize distribution_centers.csv and users.csv for mapping and analyzing the geographic distribution of users and distribution centers.

MAP SHOWS DISTRIBUTION OF USERS AND DISTRIBUTION CENTRES



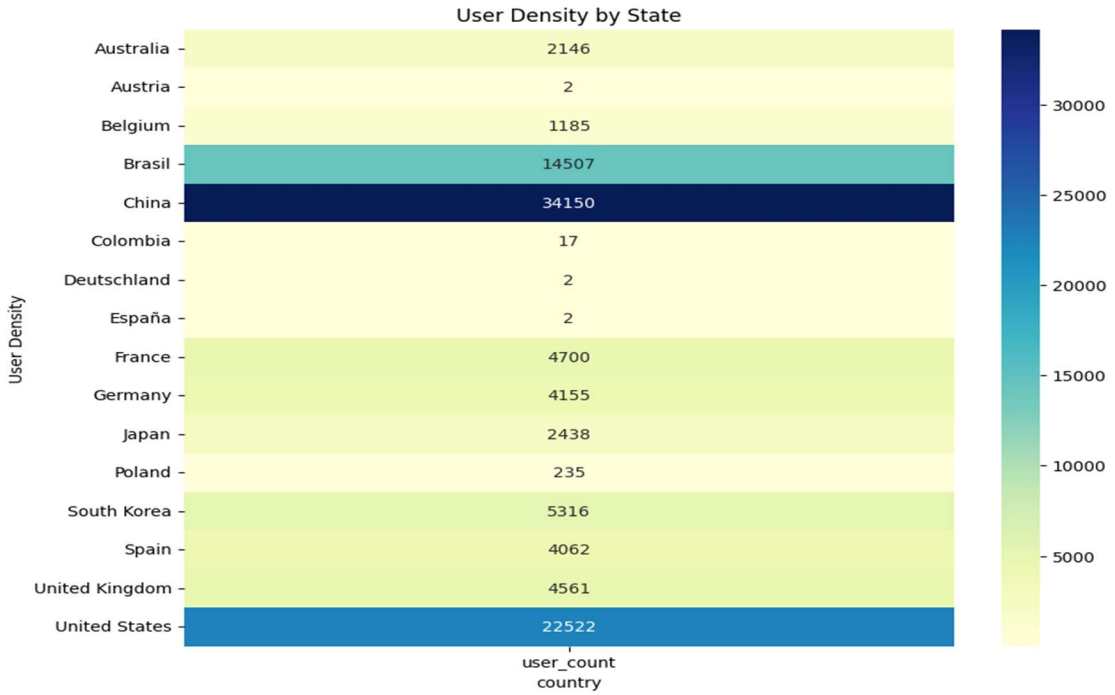


The blue dots show user spread and Red dots shows Distribution centers. Distribution centres are way less than the users in USA and China.

Suggestion:

We can increase distribution centers at high traffic places so as to get timely delivery and customer satisfaction.

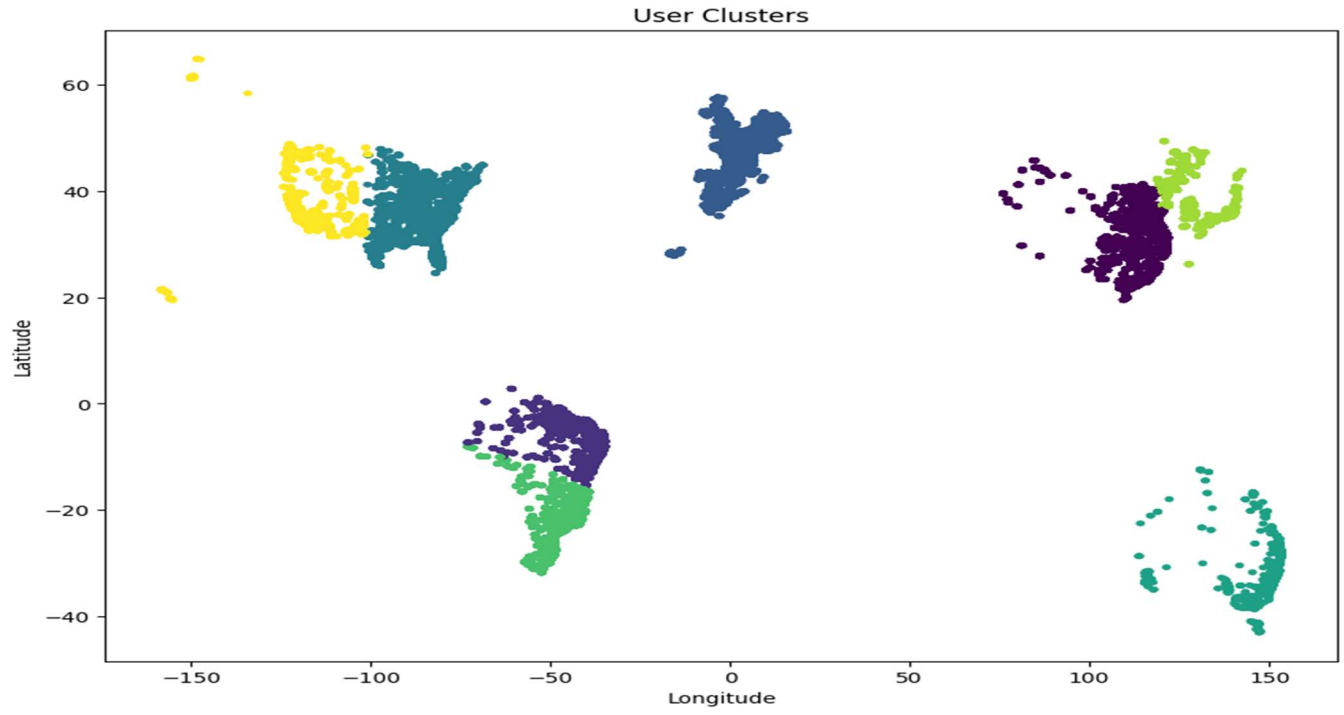
USER DENSITY BY COUNTRY



Inference:

1. China and USA are the biggest market for our e commerce business.
2. Columbia, Deutschland, Espania are our potential markets and we can grow their substantially.
3. South Korea is an emerging market for us and we can tap the opportunity.

SCATTERING OF CLUSTERS USING KMEANS ALGORITHM



INFERENCE:

1.We can identify regions with distinct spatial patterns.
2. Among the continents spatial distribution of users is quite visible and the sparce areas are our potential clients which we can tap if we have more distribution systems there.

nearest_distribution_center_distances

```
array([4875.04189834, 4875.04189834, 4875.04189834, ..., 5896.89144979,
       5896.89144979, 5896.89144979])
```

Inference:

I have found the nearest distribution center for each user and have calculated the distance to it.

Key Points:

- Visualization: Maps are utilized to visualize the spatial distribution of users and distribution centers, providing a clear understanding of their geographic relationship.
- Density Analysis: Heatmaps and clustering techniques helped identify regions with varying user densities, guiding decisions on resource allocation and service expansion.
- Distance Assessment: By calculating distances between users and distribution centers, coverage and accessibility of distribution centers were evaluated, aiding in optimizing logistics and improving service reach.

Overall, these analyses provided valuable insights into the geographic aspects of the dataset, facilitating informed decision-making in areas such as resource allocation, service optimization, and expansion planning.

```
# Explore the data
events_df.head()
```

	id	user_id	sequence_number	session_id	created_at	ip_address	city	state	postal_code	browser	traffic_source	uri	event_type
0	2198523	NaN	3	83889ed2-2adc-4b9a-af5d-154f6998e778	2021-06-17 17:30:00+00:00	138.143.9.202	São Paulo	São Paulo	02675-031	Chrome	Adwords	/cancel	cancel
1	1773216	NaN	3	7a3fc3f2-e84f-44fe-8876-eff76741f7a3	2020-08-07 08:41:00+00:00	85.114.141.79	Santa Isabel	São Paulo	07500-000	Safari	Adwords	/cancel	cancel
2	2380515	NaN	3	13d9b2fb-eee1-43fd-965c-267b38dd7125	2021-02-15 18:48:00+00:00	169.250.255.132	Mairiporã	São Paulo	07600-000	IE	Adwords	/cancel	cancel
3	2250597	NaN	3	96f1d44e-9621-463c-954c-d8deb7fffe7f	2022-03-30 10:56:00+00:00	137.25.222.160	Cajamar	São Paulo	07750-000	Chrome	Adwords	/cancel	cancel
4	1834446	NaN	3	d09dce10-a7cb-47d3-a9af-44975566fa03	2019-09-05 01:18:00+00:00	161.114.4.174	São Paulo	São Paulo	09581-680	Chrome	Email	/cancel	cancel

User Behavior Analysis:

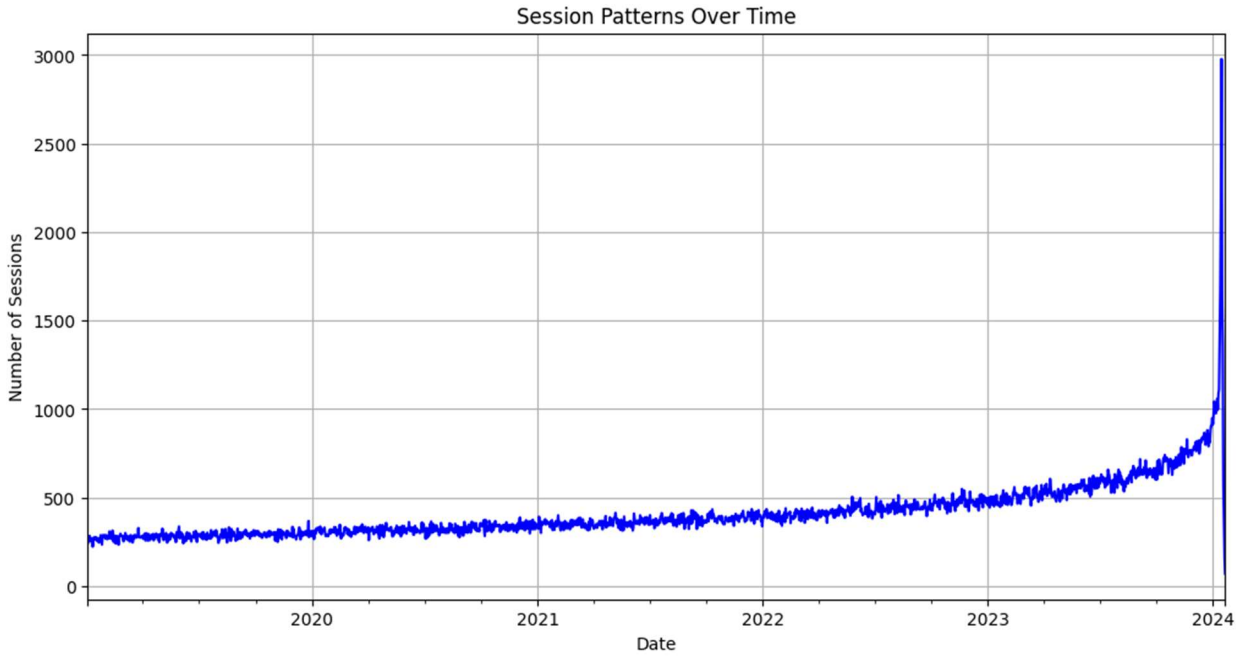
Use events.csv to analyze user behavior, including session patterns, traffic sources, and event types.

- **multiple events have been created on the same date indicating synergy in their sequence.**

Key Points:

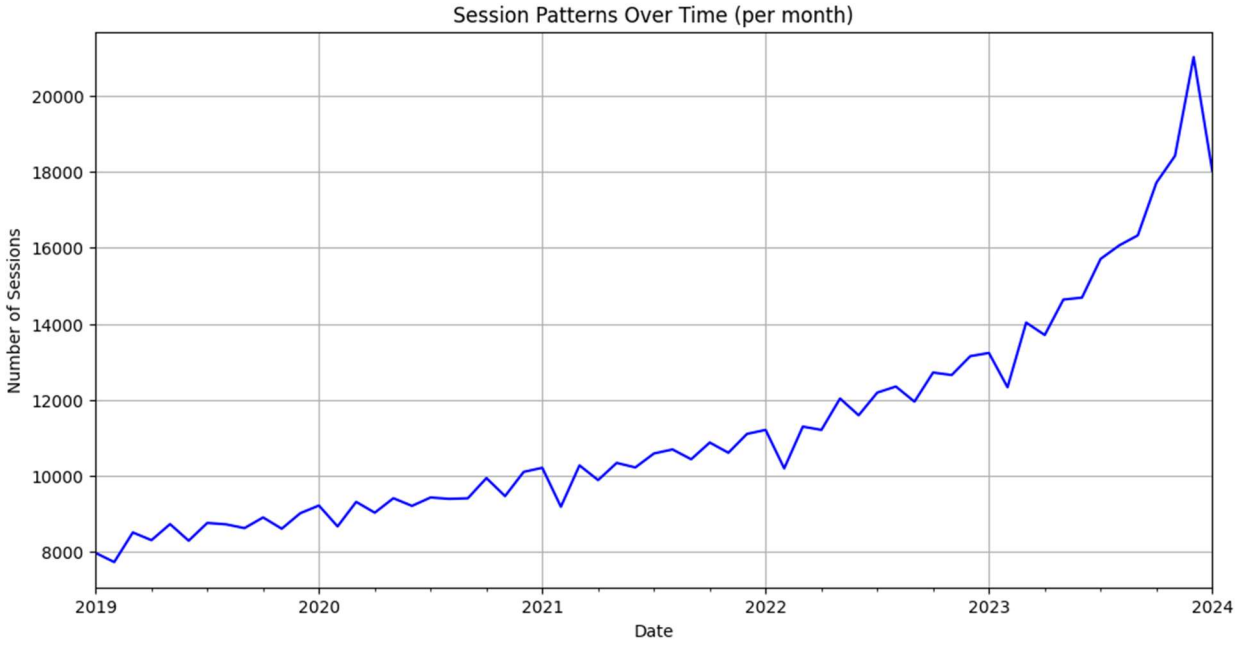
- Session Patterns Analysis: Visualized session patterns over time using line plots to understand trends and fluctuations in user activity. This visualization provided insights into how user activity varies over days, weeks, or months, allowing us to identify patterns such as daily or weekly peaks, seasonal trends, or changes in user behavior over time.
- Traffic Sources Analysis: Analysed correlation between traffic source and event type using heatmap. Allowling us to easily identify which traffic sources associated with certain event types more frequently, helping us understand user behavior patterns and the effectiveness of different traffic sources in driving specific actions
- Event Types - Frequency Analysis: Examined the frequency and distribution of each event type to understand user engagement patterns and preferences.

I have applied Kmeans clustering to group users based on their behavior patterns.



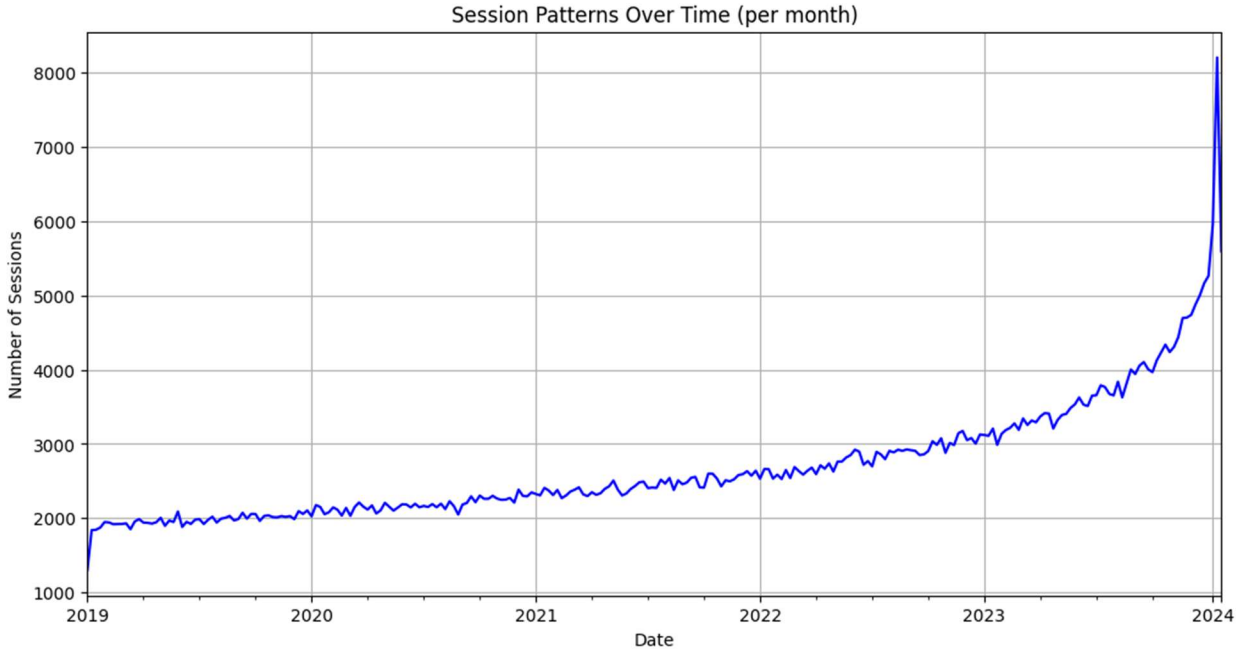
inference:

1. number of unique sessions kept on increasing per day after 2020 but the growth is significant in 2023-2024.
2. no gaps have occurred in occurrence of events, and it has been continuously running.



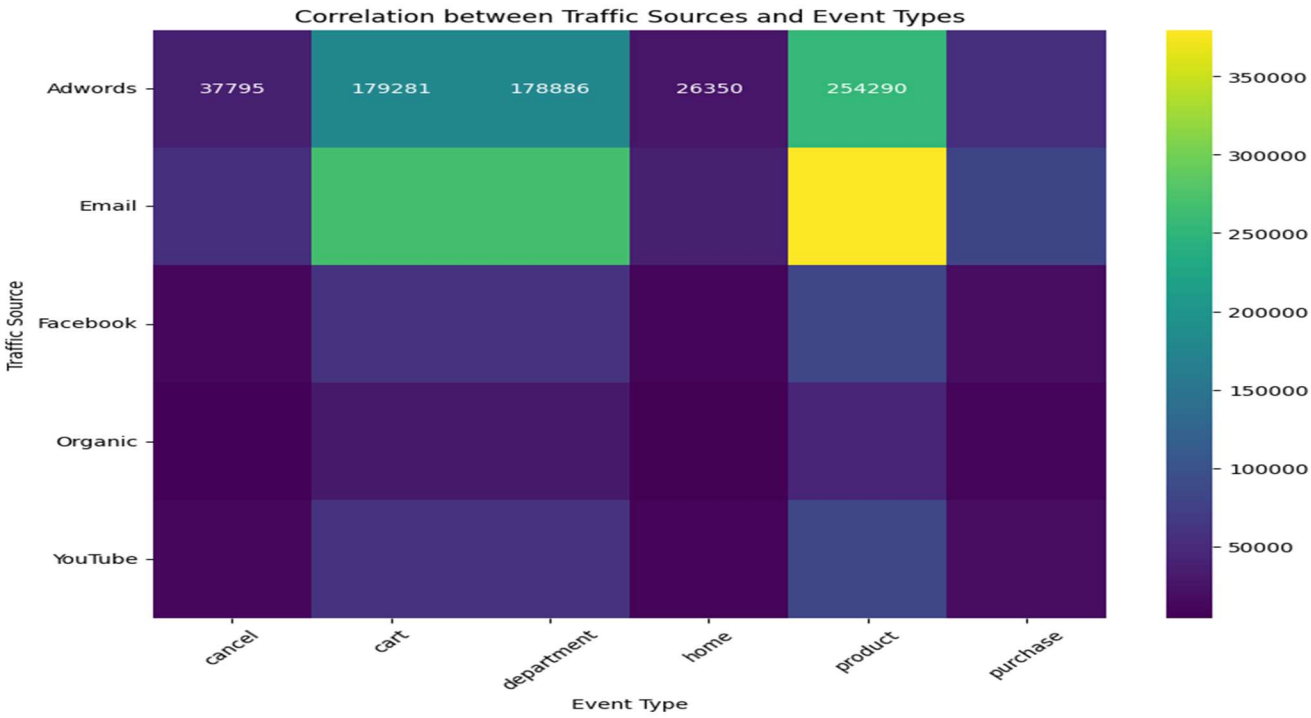
INFERENCE:

1. Every year their is a significant dip in number of unique sessions in first quarter(January to March)
2. From 2023-2024 exponential hike in unique sessions have been observed after first quarter.



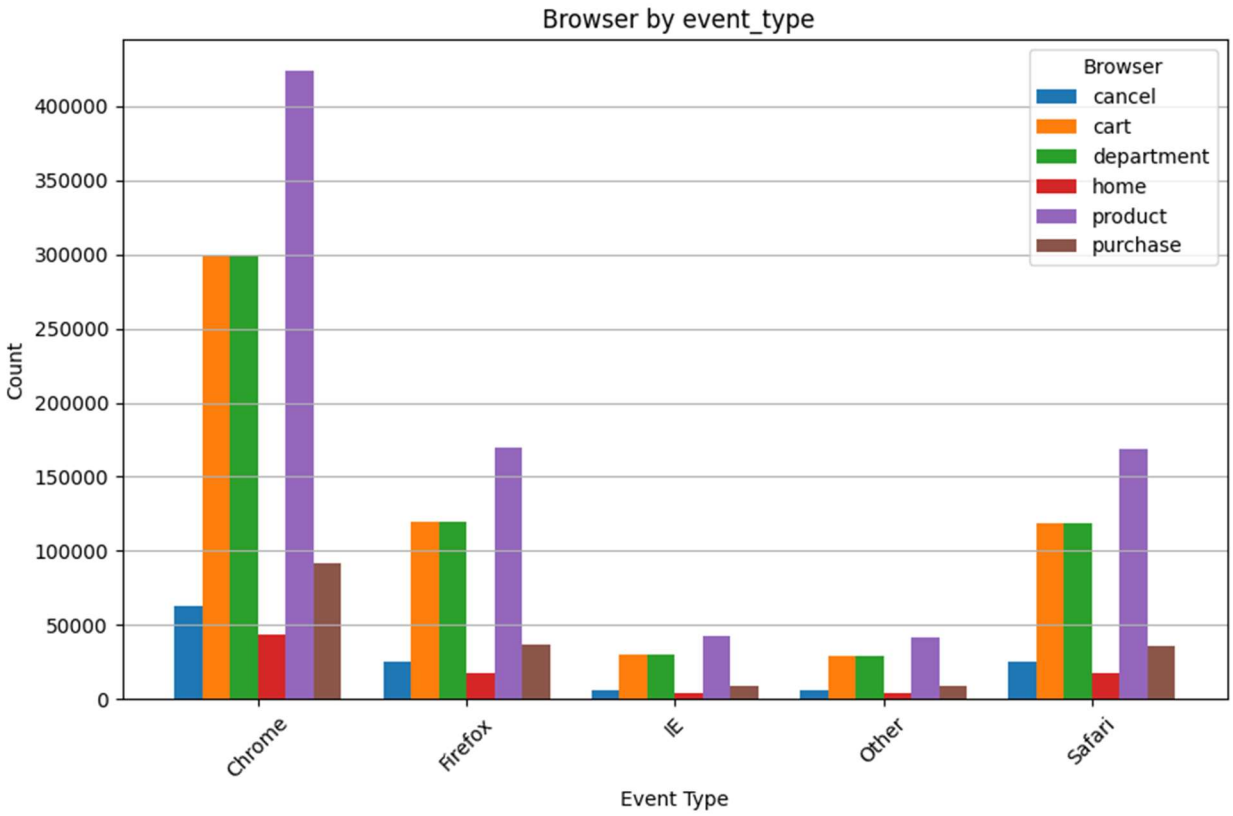
Inference:

in late 2023 and starting 2024 the weekly sessions have shown constant and sharp increase



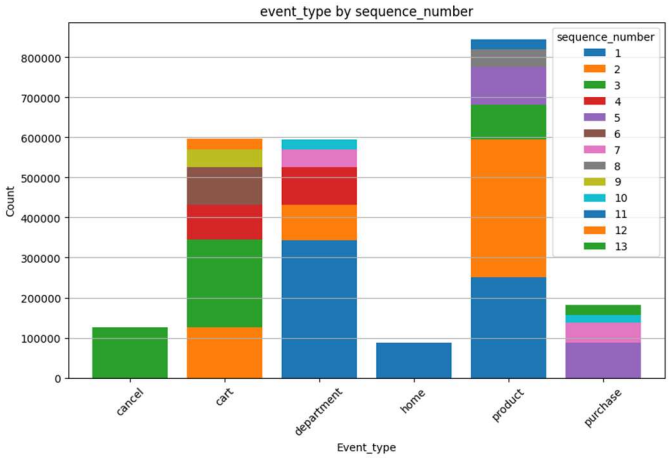
Inference

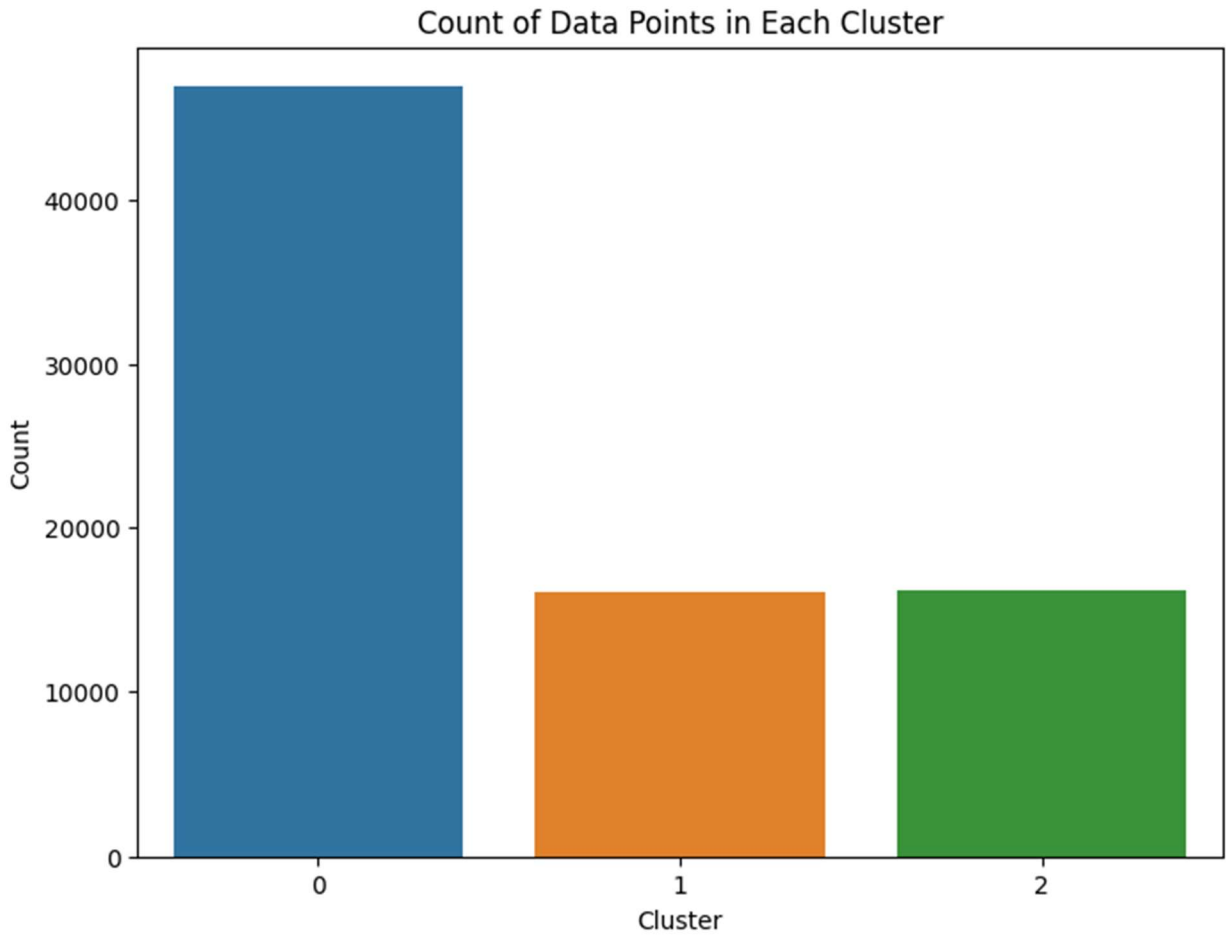
1. purchase is highly likely to be cancelled traffic source is organic.
2. Emails are great source for luring customers directly to access products but the correlation shows that purchase rate through emails are less, this should be handled well as potential customers who are liking the product are not buying them for some reason.



inference:

1. maximum users use chrome browser to access the website
2. the number of users reaching the department section are likely to reach to the desired product butdo not add product in the cart so coupons can be used on products or products can be platform assured or reviews should be good for the products to sell.
3. least number of users are accessing website using Internet Explorer and other mediums.





Inference:

I have performed customer segmentation using Kmeans and got 3 Optimum clusters using Elbow method., parameters I used were,

behavioral_features = ['session_count_per_user', 'total_duration', 'event_type_count']

Segmenting users based on their behavior patterns involves identifying groups of users who exhibit similar behavior in terms of their interactions with the system. This can help tailor marketing strategies, product offerings, and user experiences to different user segments.

SALES AND REVENUE ANALYSIS

Leverage order_items.csv and inventory_items.csv to analyze product sales, revenue, and profitability.

Key points:

Sales Analysis:

- Sales Trends: Examined sales trends over time to identify seasonal patterns, spikes in demand, or fluctuations in sales volumes.
- Product Performance: Evaluated the performance of individual products, categories, or brands based on their sales figures.

Revenue Analysis:

- Total Revenue: Calculated the total revenue generated from product sales by summing up the retail prices of all sold items.
- Profitability: Assessed the profitability of product sales by comparing revenue with associated costs, including inventory costs.

```
order_items_df.head(1)
```

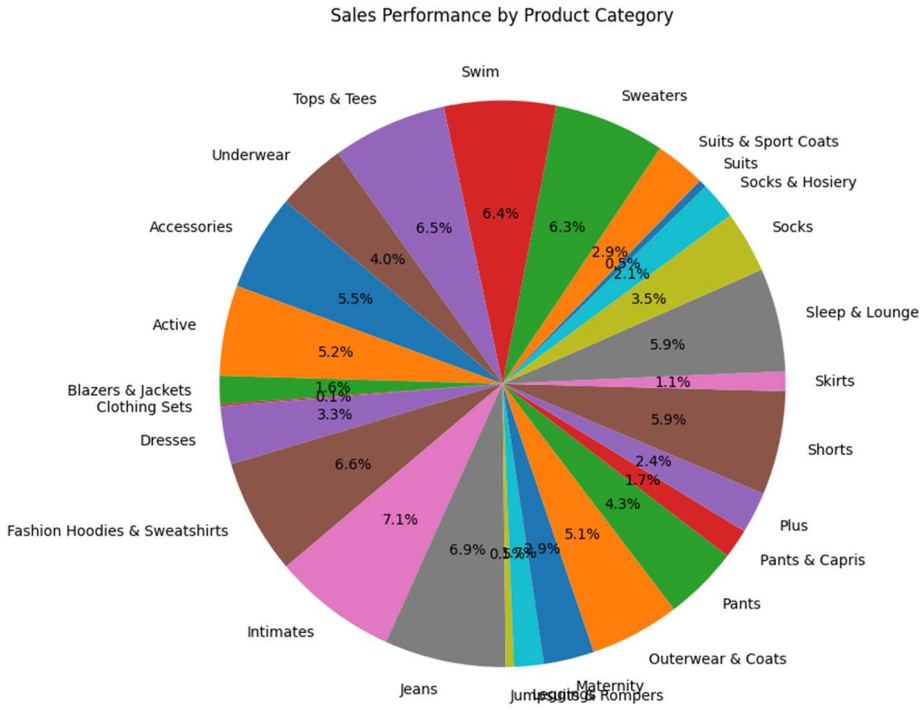
	id	order_id	user_id	product_id	inventory_item_id	status	created_at	shipped_at	delivered_at	returned_at	sale_price
0	152013	104663	83582	14235	410368	Cancelled	2023-05-07 06:08:40+00:00	NaN	NaN	NaN	0.02

```
inventory_items_df.head(1)
```

	id	product_id	created_at	sold_at	cost	product_category	product_name	product_brand	product_retail_price	product_department	
0	67971	13844	2022-07-02 07:09:20+00:00	2022-07-24 06:33:20+00:00	2.76804	Accessories	(ONE) 1 Satin Headband	Funny Girl Designs	6.99	Women	2A3E9

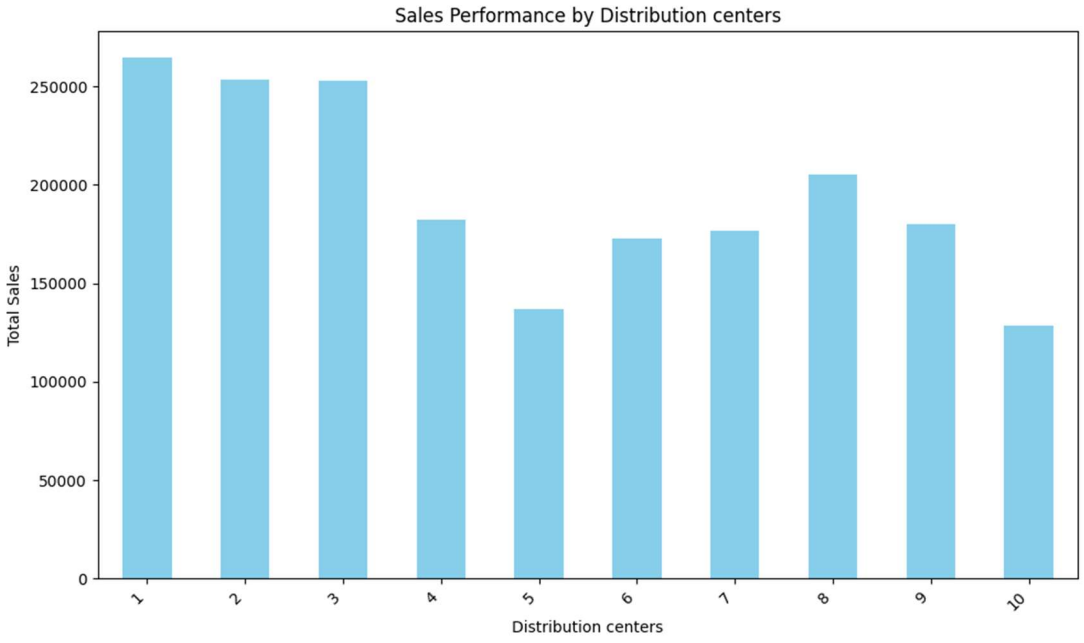
COUNT OF ORDER STATUS in the order_items table

- its a dynamic data and can keep on changing with time , but when the dataset was updated most orders were in transit phase indicating higher logistics involvement.



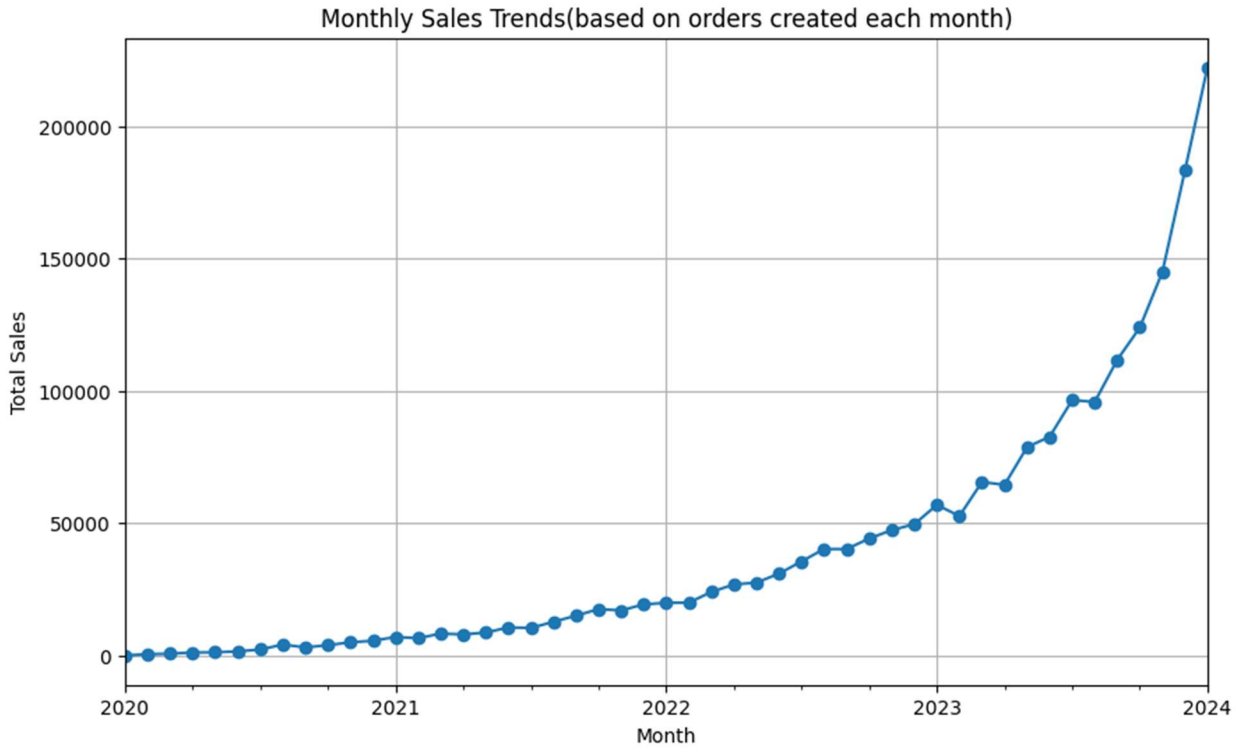
Inference:

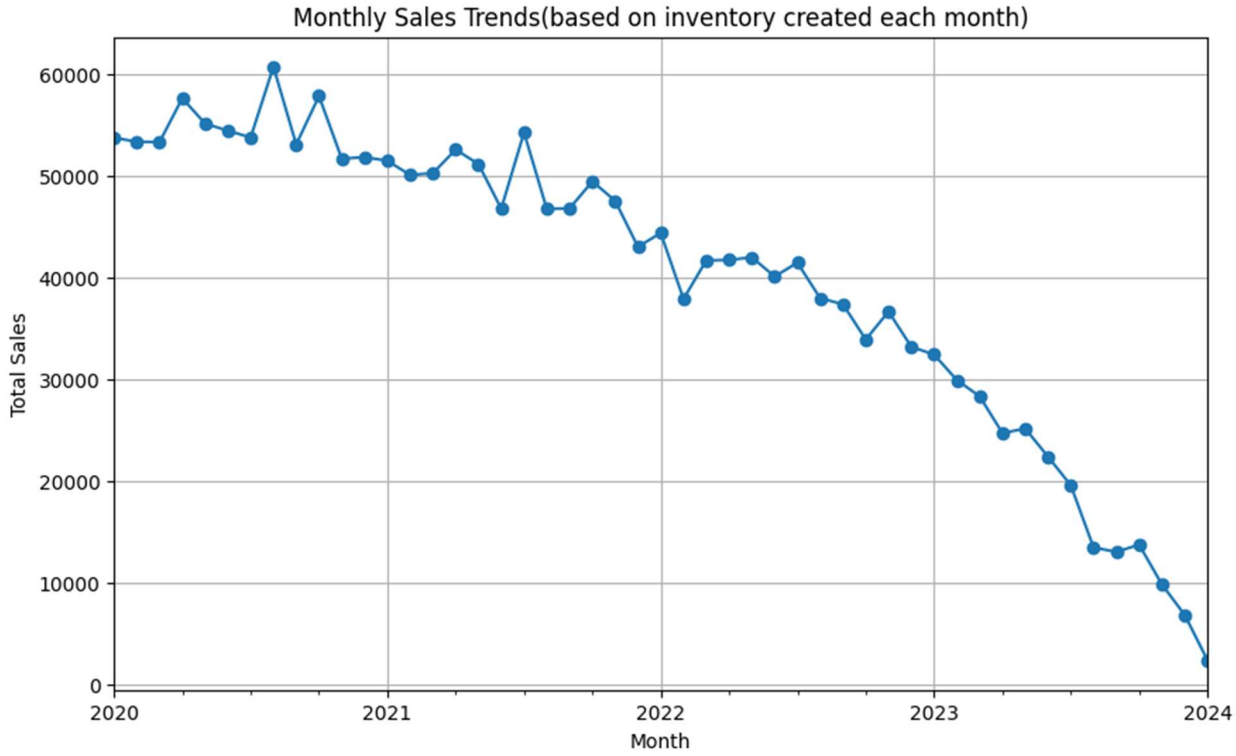
1. 'Intimates' product category has shown highest sales from the available inventory.
2. 'Clothing sets' are least sales generating category.



inferences:

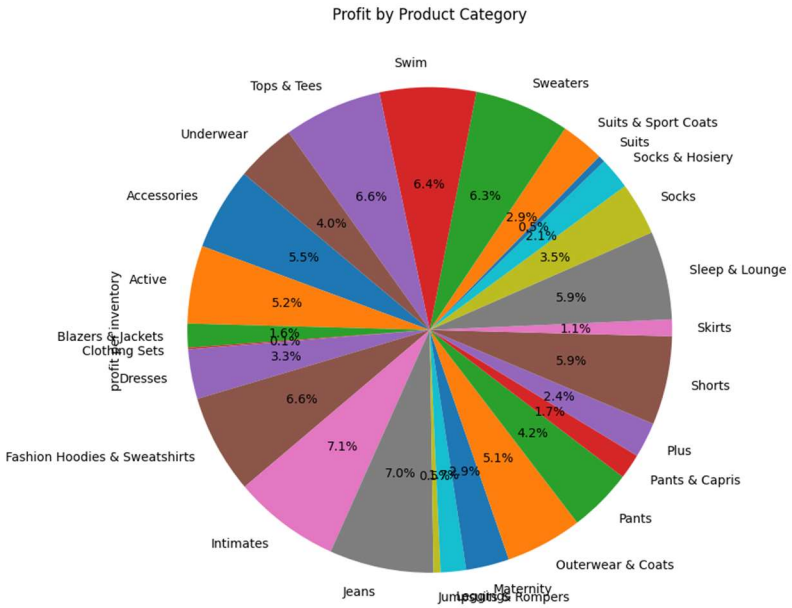
1. Distribution centre number 1 offers maimum inventory sales
2. center no. 5 show the least sales





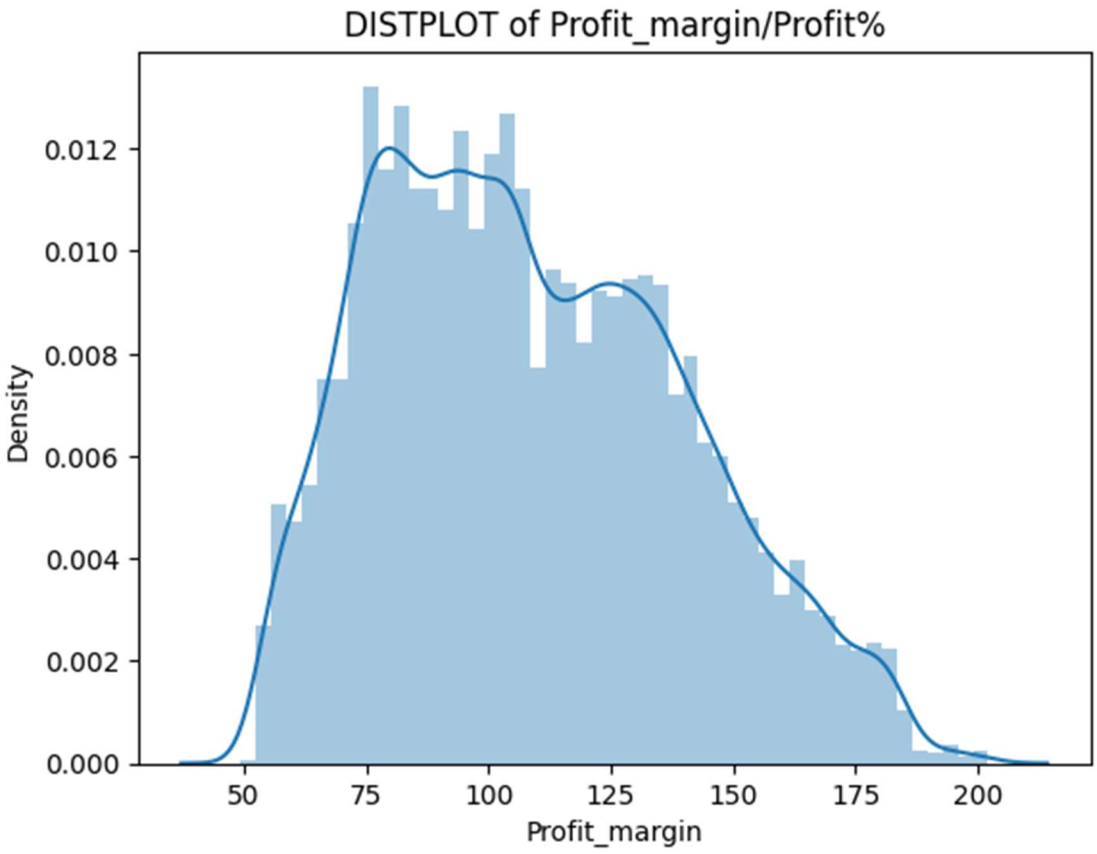
infrence:

1. based on inventory creation the trend shows decline in total sales from inventries with respect to time.
2. based on orders creation the sales trend from each inventory do show a steady increase till 2020 and later on it shows good exponential increase.



Sales performance is directly related to profitability.

Sales and Revenue analysis based on product and its count



inference:

1. on the basis of project the average profit incurred by the ecommerce is 108%.
2. mode < median < mean; the profit margin is right skewed.

Suggestions for Revenue Improvement:

- Product Diversification: Consider expanding the product range to cater to diverse customer preferences and capture new market segments.
- Price Optimization: Analyze pricing strategies to ensure competitiveness while maximizing profitability. Implement dynamic pricing models or discounts to stimulate sales.
- Inventory Management: Optimize inventory levels to prevent stockouts or excess inventory, minimizing carrying costs and maximizing sales potential.
- Customer Engagement: Enhance customer engagement and loyalty through targeted marketing campaigns, personalized offers, and excellent customer service.
- Market Expansion: Explore opportunities for market expansion through geographical expansion, partnerships, or online sales channels to reach new customers and increase sales.

Overall, the Sales and Revenue Analysis provided valuable insights into product performance, revenue generation, and profitability. By implementing the suggested strategies for revenue improvement, businesses can optimize their sales processes, enhance profitability, and drive sustainable growth.

PRODUCT PERFORMANCE ANALYSIS

Product Performance Analysis:

Explore products.csv to analyze product performance, including costs, categories, and popularity.

```
In [577]: products_df.head(3)
```

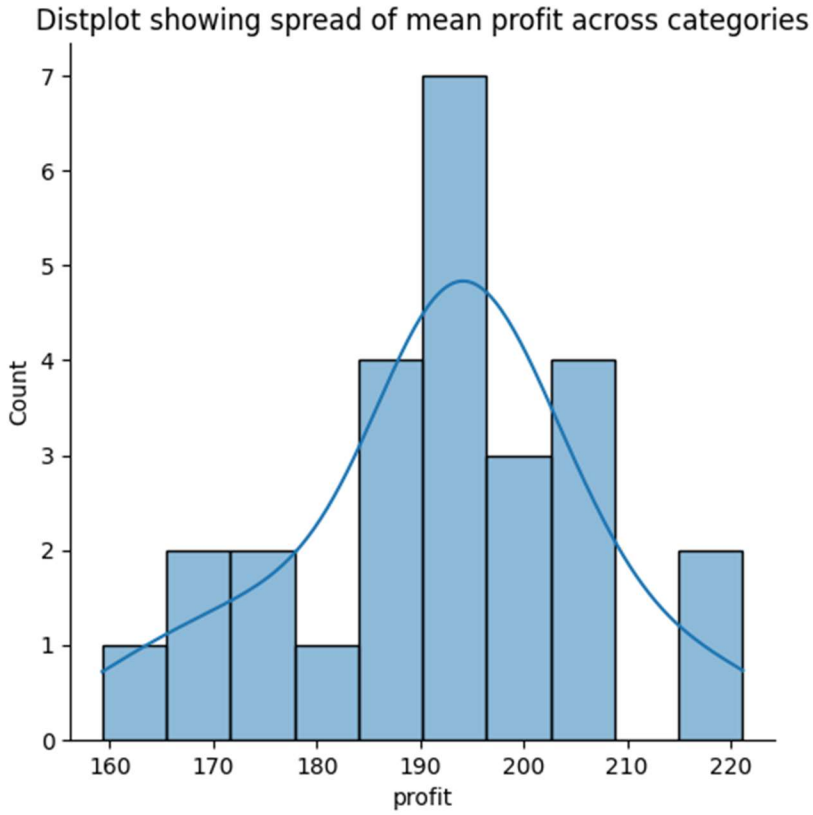
Out[577]:

	id	cost	category	name	brand	retail_price	department	sku	distribution_center_id
0	13842	2.51875	Accessories	Low Profile Dyed Cotton Twill Cap - Navy W39S55D	MG	6.25	Women	EBD58B8A3F1D72F4206201DA62FB1204	1
1	13928	2.33835	Accessories	Low Profile Dyed Cotton Twill Cap - Putty W39S55D	MG	5.95	Women	2EAC42424D12436BDD6A5B8A88480CC3	1
2	14115	4.87956	Accessories	Enzyme Regular Solid Army Caps-Black W35S45D	MG	10.99	Women	EE364229B2791D1EF9355708EFF0BA34	1

KEY POINTS:

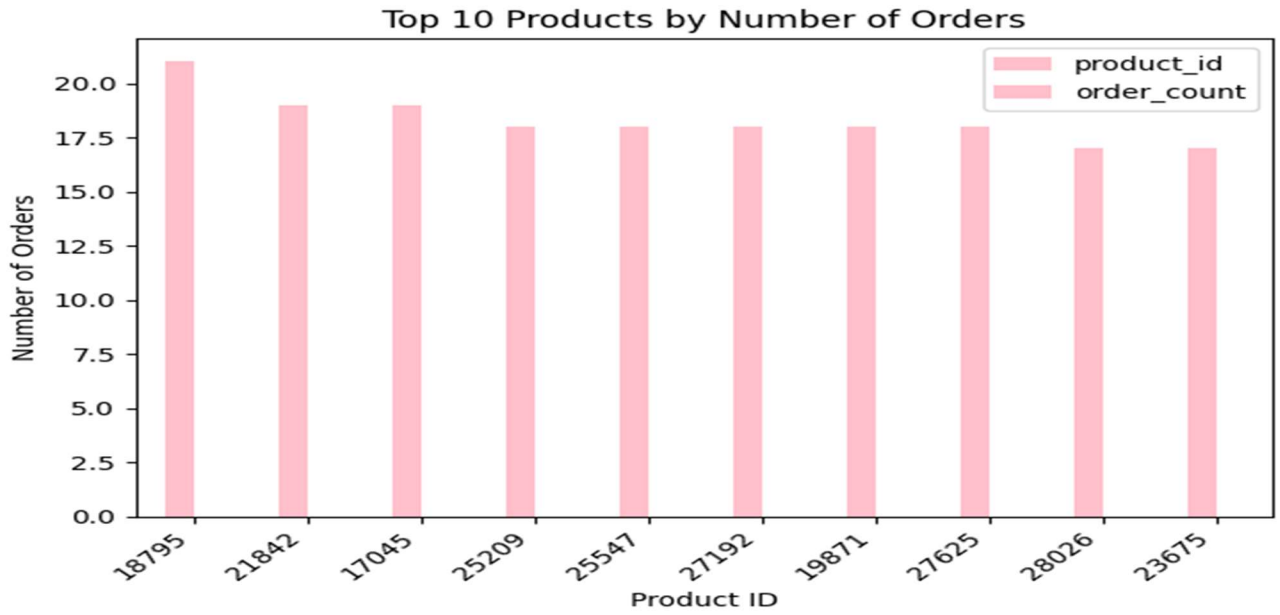
- Top-Selling Products: Identified top-selling products or bestsellers within each category or brand to understand consumer preferences.

- Popularity Trends: Analyzed popularity trends over time to identify emerging trends, seasonal variations, or changes in consumer preferences.



inference:

1. average profit is nearly normally spread across the categories and hence categories do playy vital role in linear variance of total sales.

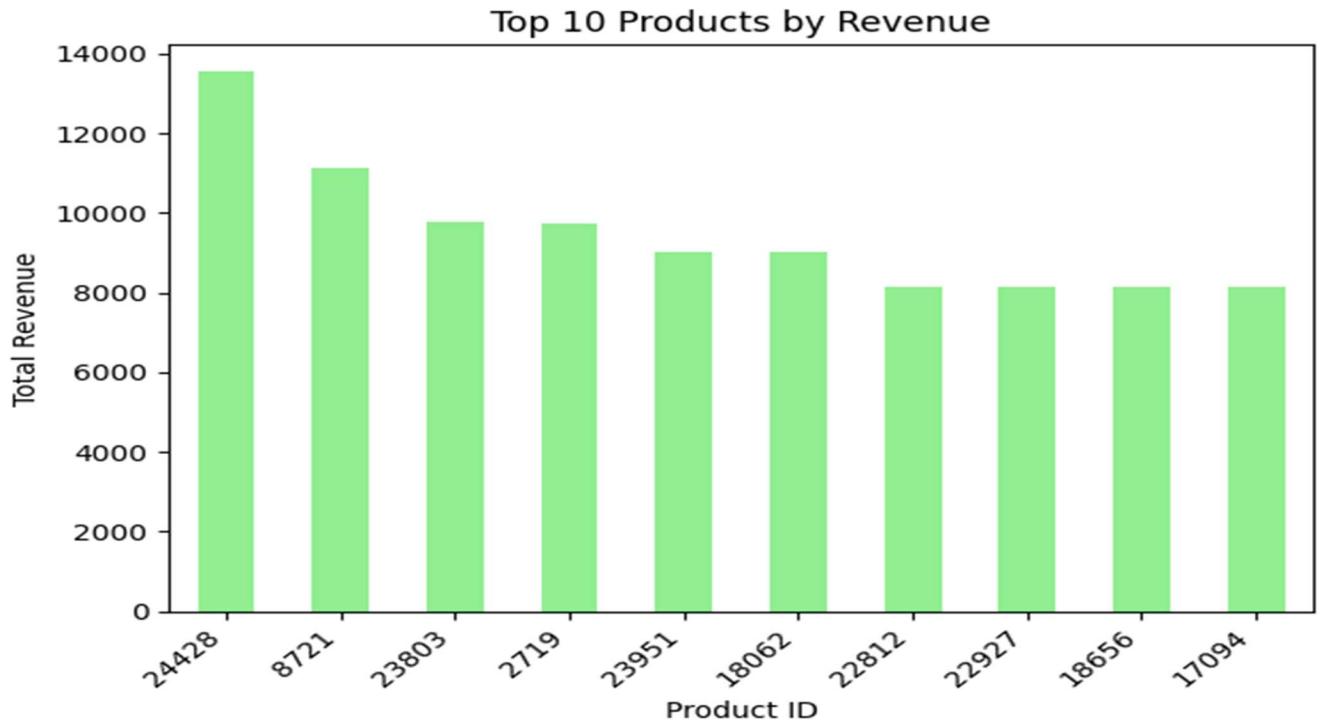


inference:

1. product_id '18795' is the frequent flyer off the carts.
2. The chart gives us the top 10 most ordered products.

inference:

1. Bottom 100 products are only ordered once.



Inference:
Product id 24428 generates most revenue so its inventory shall be constantly updated on priority basis.

Suggestions for Product Performance Improvement:

- Product Portfolio Optimization: Evaluate the product portfolio to identify underperforming or obsolete products that can be discontinued or replaced with more profitable alternatives.
- Price Adjustment: Review pricing strategies to ensure competitiveness and alignment with market trends. Consider implementing dynamic pricing or promotional strategies to stimulate sales.
- New Product Development: Invest in research and development to introduce innovative products or product variations that cater to evolving consumer needs and preferences.
- Marketing and Promotion: Develop targeted marketing campaigns to increase product visibility, enhance brand awareness, and drive customer engagement. Utilize social media, email marketing, and influencer partnerships to reach a wider audience.
- Inventory Management: Implement effective inventory management practices to optimize stock levels, minimize carrying costs, and prevent stockouts or overstock situations.

User Demographics Analysis

User Demographics Analysis:

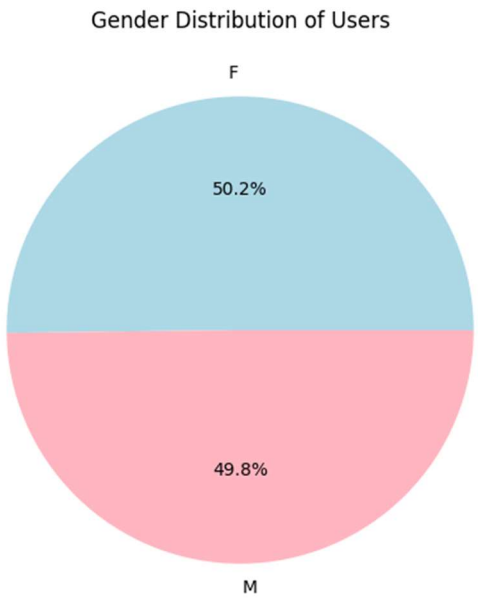
Use users.csv to analyze user demographics, such as age, gender, and location.

Age Analysis:

- Age Distribution: Examined the distribution of user ages to understand the demographic composition of the user base.
- Age Groups: Grouped users into different age brackets or cohorts to analyze age-related patterns and preferences. Gender Analysis:

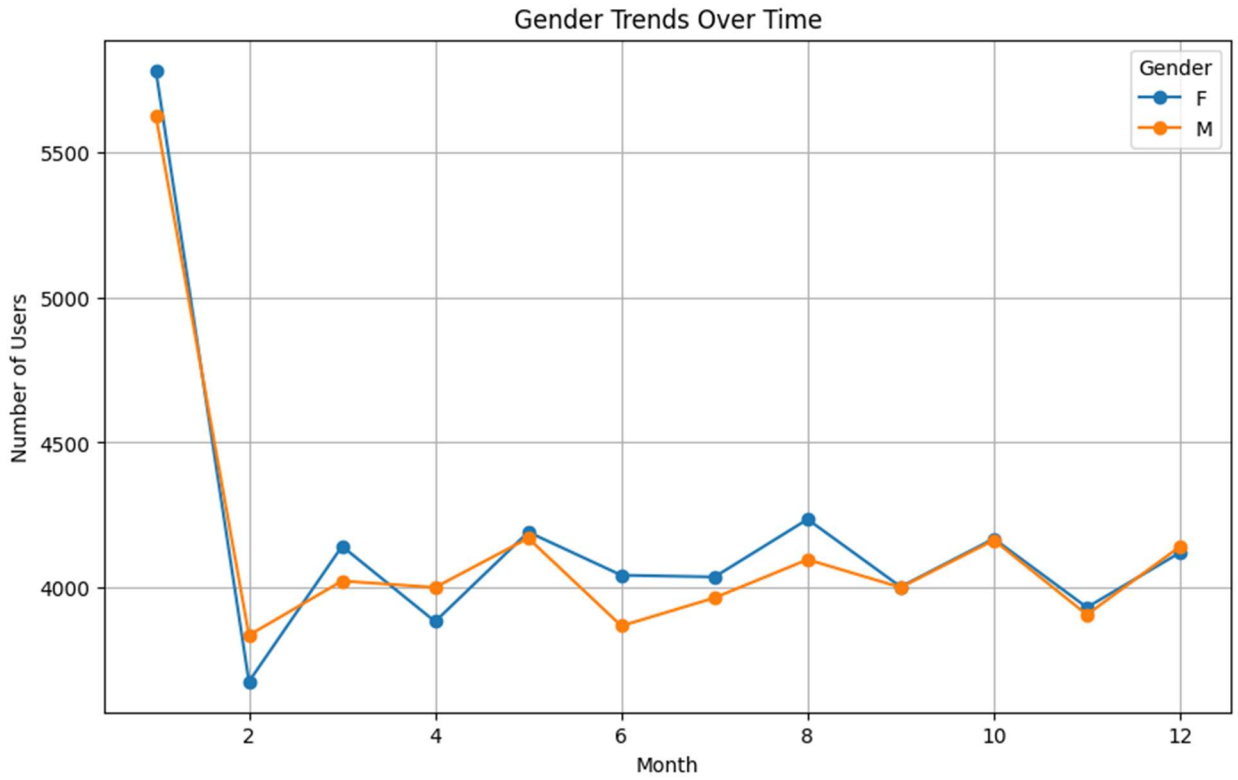
Gender Distribution:

- Analyzed the distribution of user genders to understand gender representation within the user base. User Segmentation:



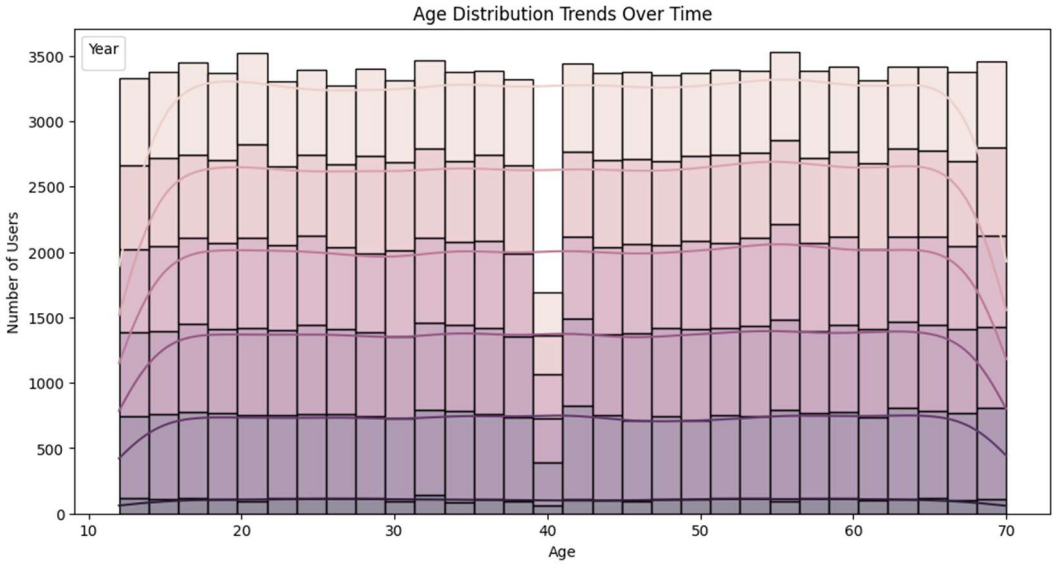
infrence:

1. number of female users is slightly greater than males.



Infrence:

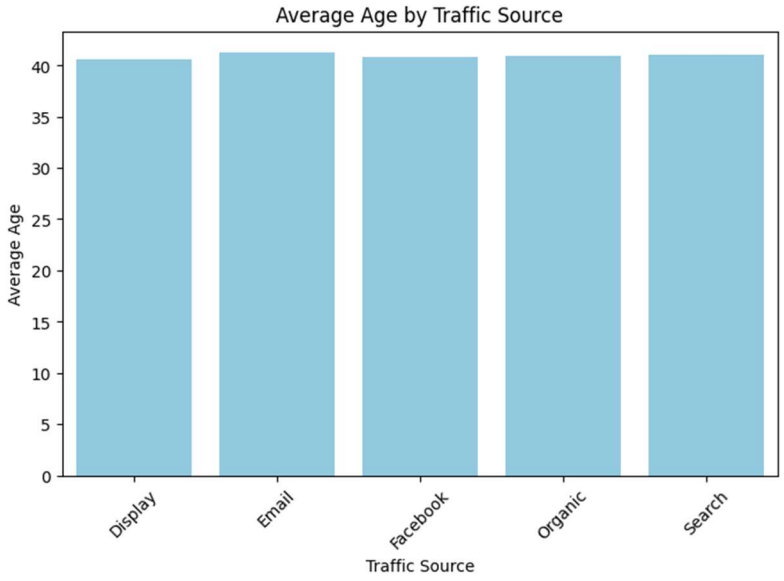
1. Gender trends seem consistent.
2. Female users are purchasing more in months of May to September, indicating their higher purchase for swin wear and intimates in summers.

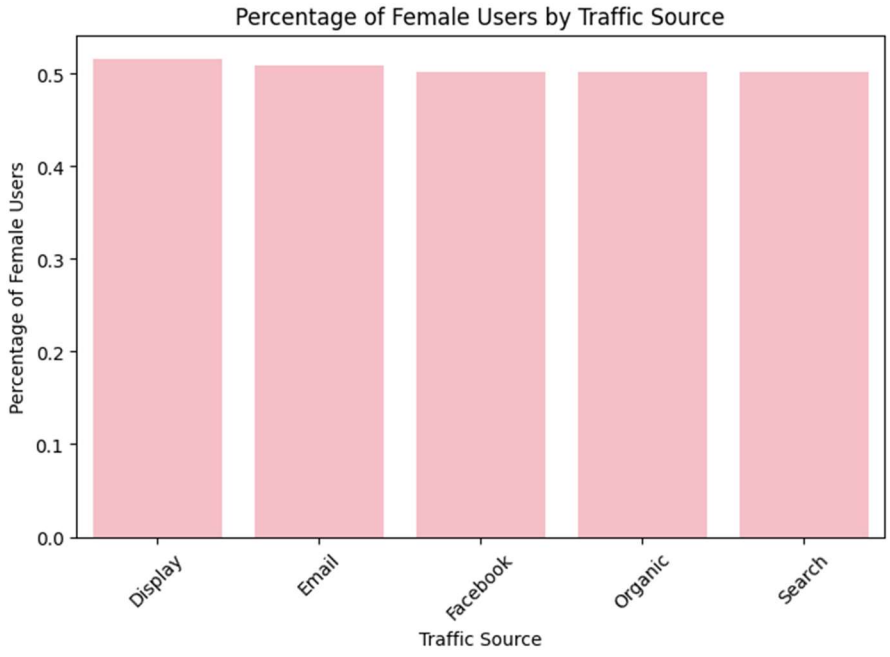


Inference:

1. Though the mean age is 40 but , people amongst that age group are least number of buyers.(38-40)
2. Heavy number of buyers belong to early 40s and we can see that trend change in the graph.

Traffic source analysis to identify what are users favorite sources and improve on them.

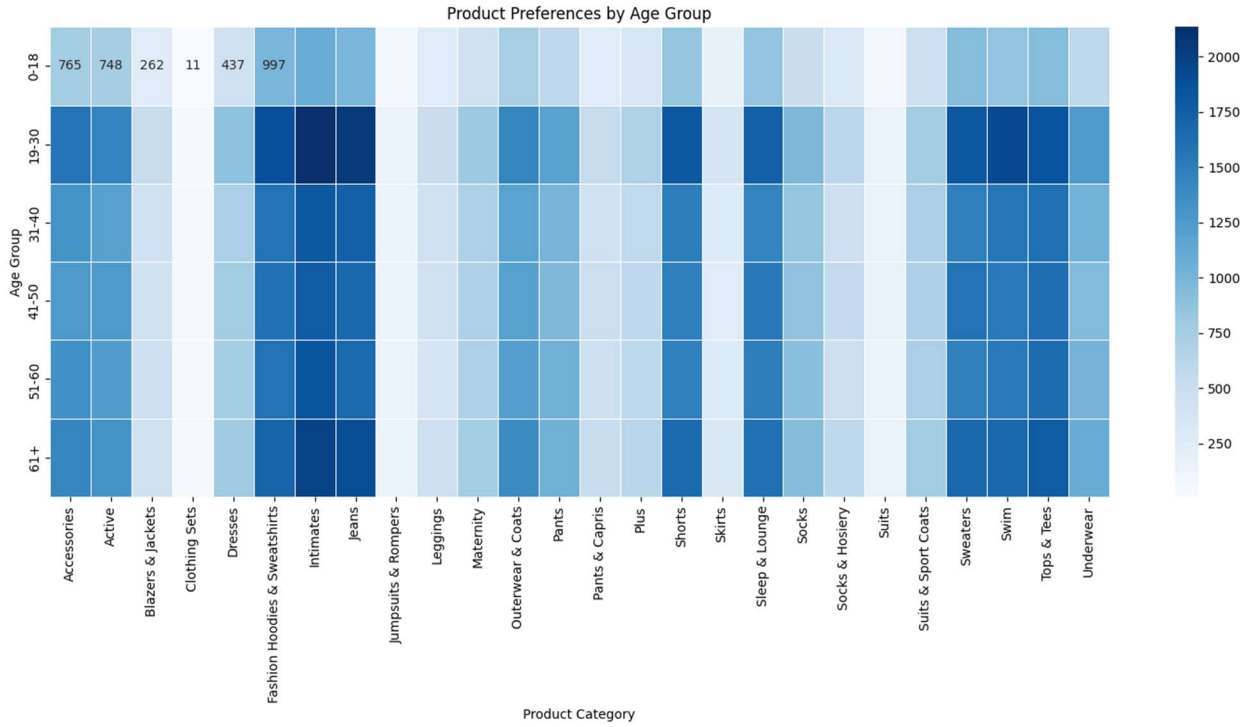




inference:

1. females are certainly using all the traffic sources and almost more than 50% females are using displays and emails to access the website.
2. average age of users through all the traffic sources lies between 40-60, these are our higgh priority target age group.
3. Female users are also our high priority target group.

Product preference based on age group



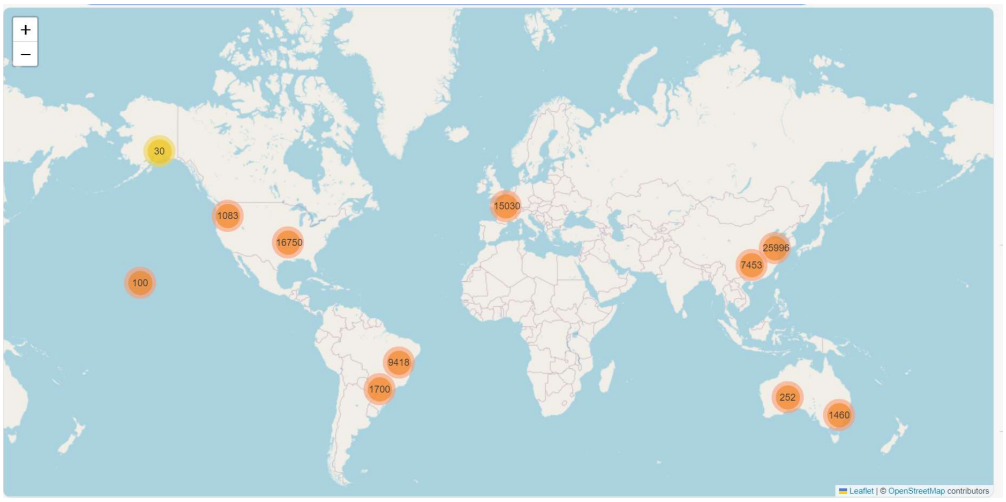

```
Top Products by Age Group:
age_group
0-18     Intimates
19-30    Intimates
31-40    Intimates
41-50    Intimates
51-60    Intimates
61+      Intimates
dtype: object
```

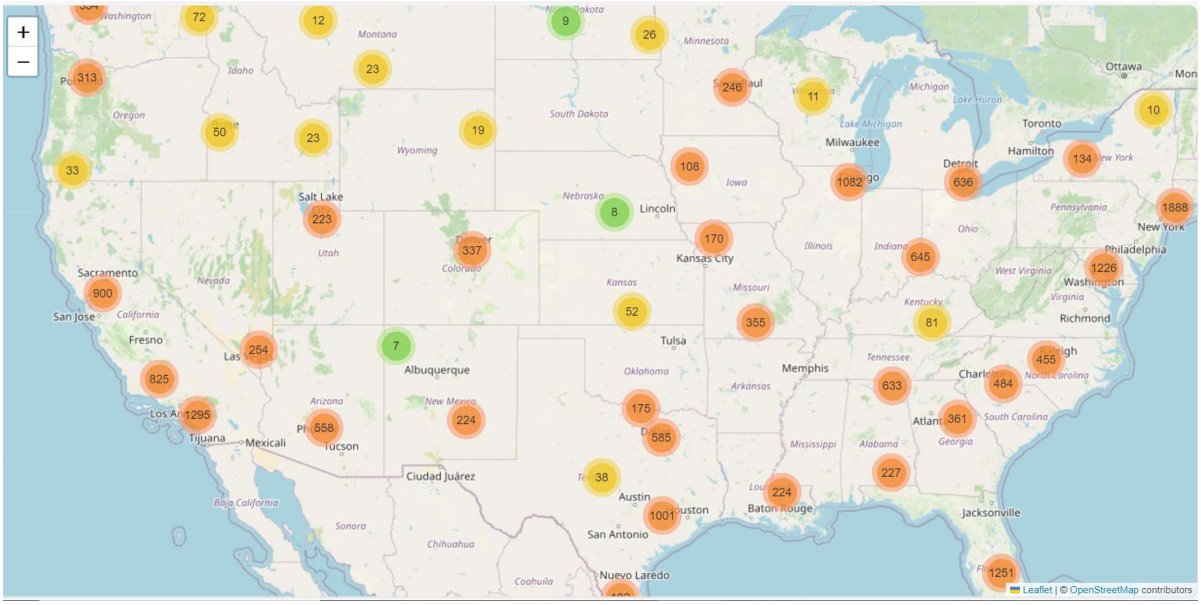
Inference:

1. The young age group of 18-30 invest more in intimates, jeans, swim wear and sweaters.
2. clothing sets are the least favorite amongst all the products.
3. Clearly all the age groups buy most of 'Intimate' clothing from our ecommerce platform.

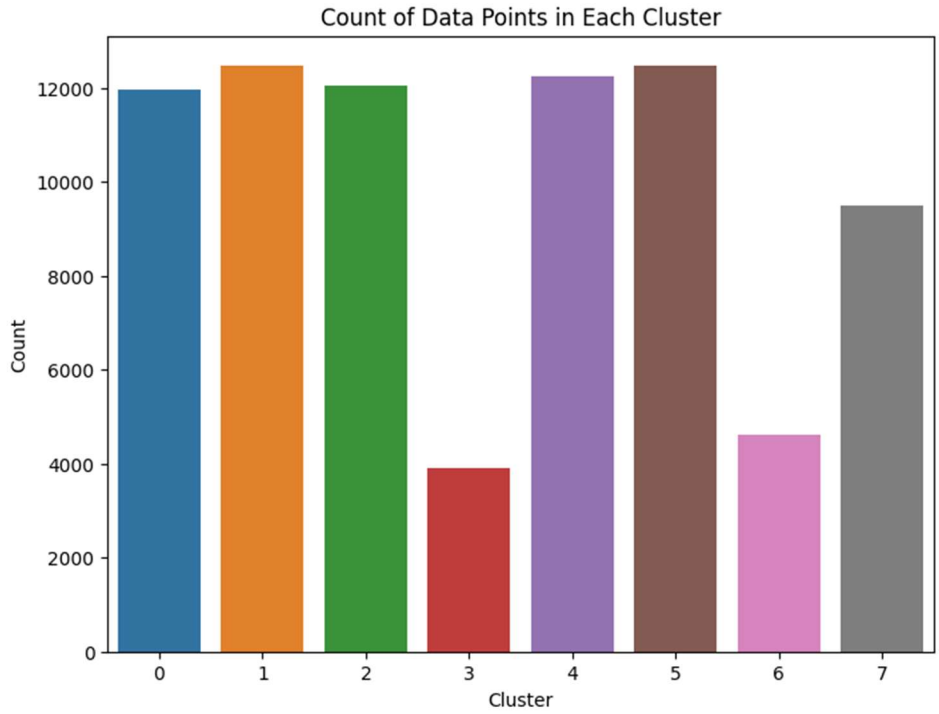
user analysis based on location¶

user_location_map.html





KMEANS CLUSTERING FOR DEMOGRAPHIC ANALYSIS OF EACH CLUSTER



KMeans Analysis: This code provides insights into different user segments based on demographic attributes, allowing us to tailor our marketing strategies and user experiences to specific user groups.

Suggestions for User Experience Improvement:

- Personalized Recommendations: Implement personalized recommendation systems based on user demographics, preferences, and past behavior to enhance user engagement and satisfaction.
- Localized Marketing Campaigns: Develop localized marketing campaigns and promotions tailored to specific regions or demographics to improve relevance and effectiveness.
- Enhanced Communication Channels: Offer multiple communication channels (e.g., live chat, email, social media) to provide personalized support and assistance to users based on their preferences.
- Accessibility and Localization: Ensure that the platform is accessible and localized to accommodate diverse user demographics, languages, and cultural preferences.
- Feedback Mechanisms: Implement feedback mechanisms to collect user feedback and suggestions for continuous improvement, ensuring that the platform evolves in alignment with user needs and expectations.

ORDER FULFILLMENT ANALYSIS

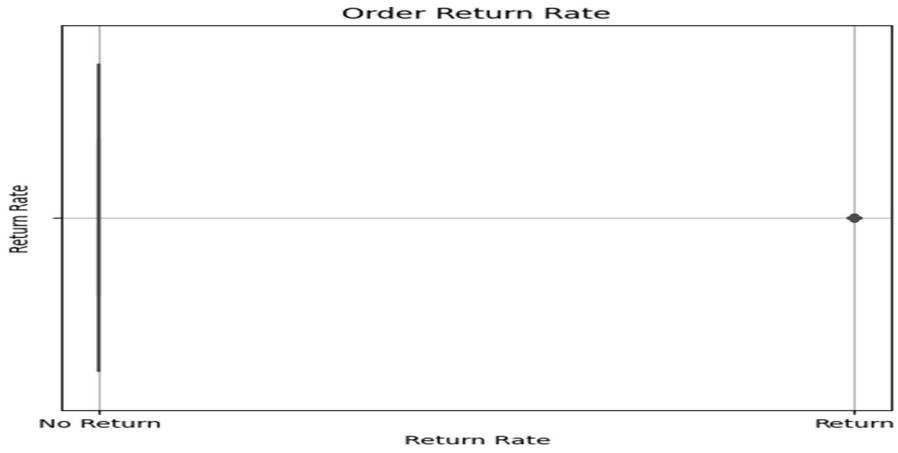
Order Fulfilment Analysis:

Analyze order_items.csv and orders.csv to understand order fulfilment timelines and status.

1. Order Timelines Analysis:

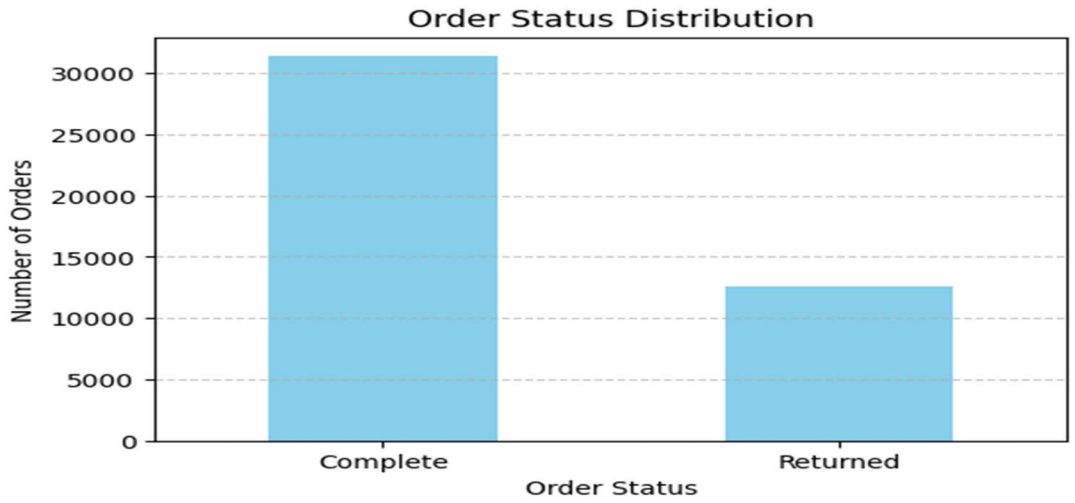
- Order Creation: Examined the time taken from order creation to shipment to understand the efficiency of order processing.
- Shipping Time: Analyzed the duration between order shipment and delivery to assess the effectiveness of shipping operations.
- Return Time: Investigated the time taken for returned orders to understand the return process efficiency and identify potential areas for improvement.

2. Analysed customer retention rate , AOC, Order frequency check, Order fullfilment rate and Delivery variance .



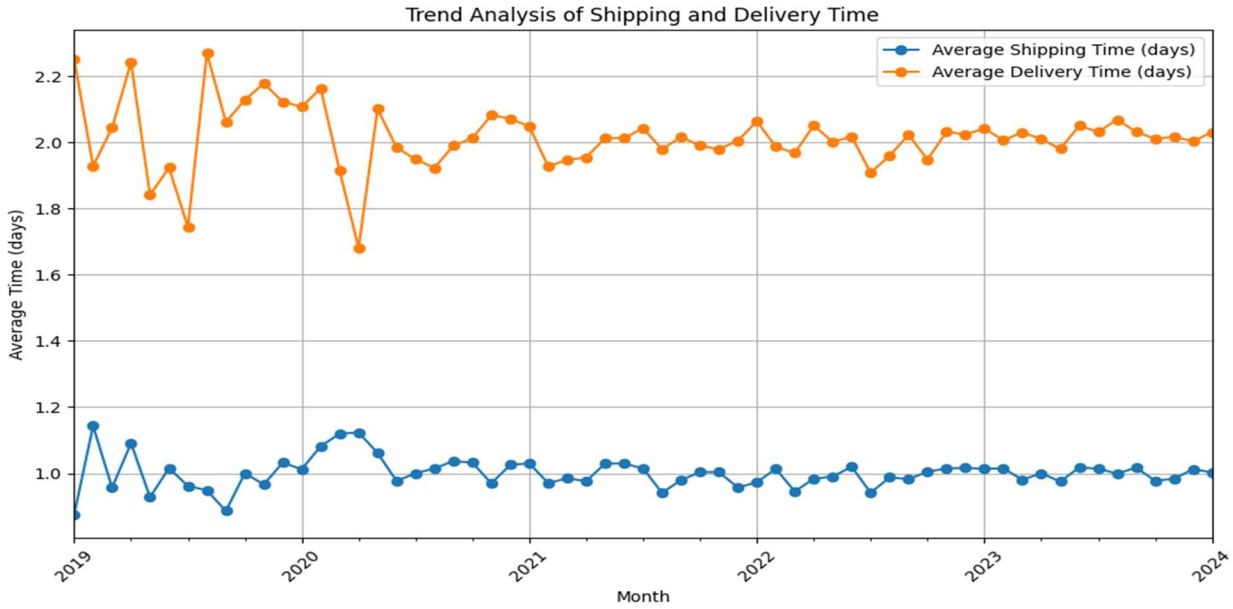
inference:

There are significantly low return rates.



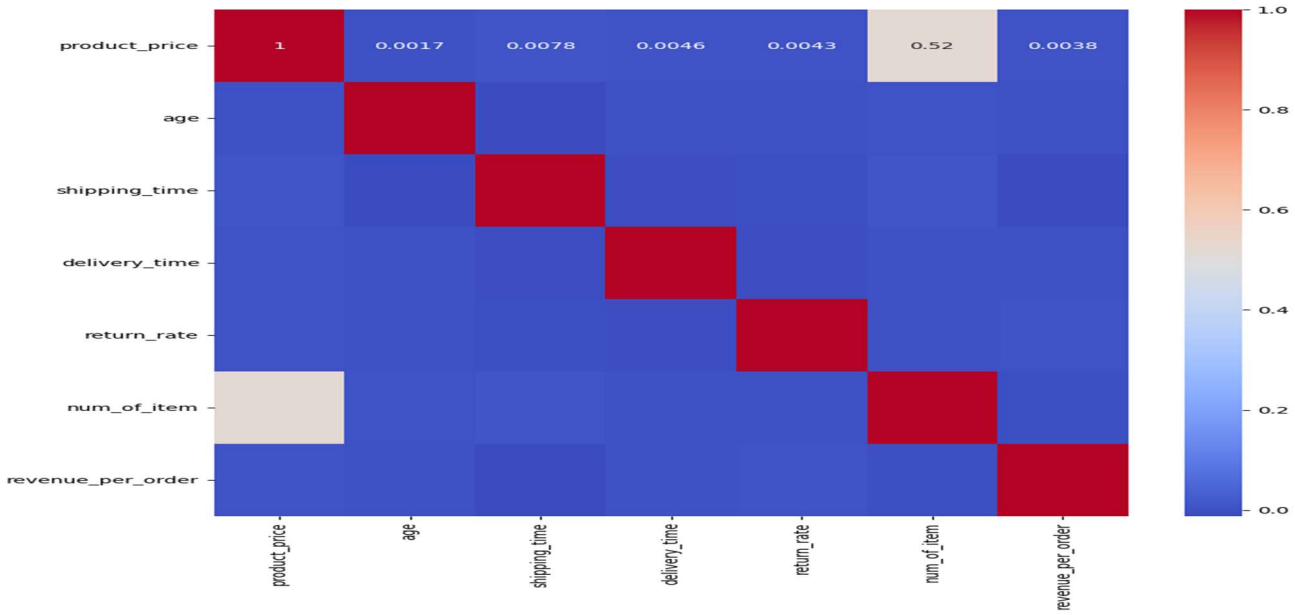
inference:

1. The majority of orders were completed successfully.
2. But the returned orders count is also significant as it reduces the revenue generation and costs the company logistics and delivery cost too.



Inference:

1. The variance in the delivery time was quite high before 2021, it has reduced over time and now the average delivery time shows steady trends.
2. Average delivery time is around 2 days now.
3. The average shipping time has been around 1 day which imply that we are ensuring quick delivery.



inference:

1. none of the features have correlation greater than 7, therefore multicollinearity/dependency is absent.

Suggestions for Improvement:

- Optimize Inventory Management: Implement effective inventory management strategies to ensure adequate stock levels, minimize stockouts, and reduce order processing delays.
- Streamline Order Processing: Automate order processing tasks and workflows to minimize manual errors, streamline operations, and expedite order fulfilment.
- Improve Logistics Efficiency: Enhance logistics and shipping operations by partnering with reliable carriers, optimizing delivery routes, and implementing real-time tracking systems to provide visibility and transparency to customers.