Sales Prediction for E-commerce Product Categories to Increase the Revenue and Profit.

***Abstract*—This paper presents the Design Documentation for the project of subject Domain Application of Predictive Analytics. The Domain of this project is the e-commerce industry. The Ecommerce industry is growing exponentially every year; this has increased the competition among the businesses. Thus, it has become necessary to make use of technology to perform predictive analysis for Sales prediction. The project aims at accurately predicting the Sales of each Category. The dataset taken for this project is made available by Data Co Organization. The project design presents a detailed view about the Domain Background, Scope, Business Value, Ethics also it has documented the benefits of performing such analysis. The exploratory data analysis is performed using Python and Power BI which gives key insights from historical data. A list of Machine Learning algorithms is given which are applicable for predicting continuous variable, which is Sales.**

Keywords—Sales Prediction, Visualizations, Machine Learning, Predictive Analysis, Revenue.

I. INTRODUCTION

A. Background of the Domain

The availability of the Internet, Computers and Mobile devices have revolutionized many day-to-day tasks of human beings. Out of those tasks, one of them is Shopping. Now anything in the world can be bought from anywhere in the world within the blink of an eye. Over the past two decades, the e-Commerce industry has evolved a lot. Now it is a 26 Trillion industry. Currently, 2 billion people use e-commerce websites for shopping online worldwide. In comparison to the world population, it is 25% of the world's population. These numbers are forecasted to be increasing year by year. By the year 2040, it is forecasted that more than 95% of retail sales would be done through e-commerce. There will be 2.20 billion e-commerce shopping users by the year 2021 [1].

As the e-commerce market started boosting many companies were established to fulfill the growing demand. From Electronics to Grocery every type of product is started selling through online websites. This resulted in increased competition among the sellers. To achieve the goals of the business and profit e-commerce companies need to understand the market trend and demand. Analysis needs to be carried out not only on the market and competitors but also on the company's past data available. A business should be completely aware of its consumers, what are the products which customers want to buy? What are the products which customers like? What are the most preferred payment methods? Orders trend need to be identified category-wise, most popular categories need to be identified [2].

It is important to predict the orders for the coming month, a quarter, year which would help the company ready to accommodate the consumer demand which will lead to better customer satisfaction, increased profit and edge over competitors [3]. Currently, the Covid-19 has disrupted the global supply chain and e-commerce operations. The effect of that was also seen on buying habits of the consumers, they were more likely to buy essential goods in such a situation than non-essential goods. This project aims to successfully predict the number of orders category-wise by implementing a suitable machine learning algorithm.

B. Overview of the Data:

Column Name	Data Type	Description
Customer ID	Integer	Unique ID to identify each customer.
Order ID	Integer	Unique ID to identify each Order.
Category Id	Integer	Unique ID to identify each Category.
Category Name	String	Name of the Category.
Department Id	Integer	Unique ID to identify each Department.
Department Name	String	Name of the Department.
Market	String	Name of the Markets.
Order City	String	Name of the Cities.
Order Country	String	Name of the Countries
Order Date	DateTime	Date of Order placed.
Sales	Integer	Amount of Sales per Order.
Order Discount	Integer	Amount of Discount given.
Discount Rate	Integer	Percentage of Discount given.
Product Price	Integer	Price of the product.
Order Quantity	Integer	Quantity of the product purchased in each order.
Order Total	Integer	Total Order amount after discount.
Order Profit per Order	Integer	Profit earned per Order.
Order Region	String	Name of the Order Region
Order Status	String	Status of the Order.
Shipping Mode	String	Type of Order Shipping Selected.

Table 1. Column Names and Description.

The dataset is taken from *Kaggle for exploratory data analysis. The data is prepared by **Data Co organization. The Dataset contains details of 1,80,520 orders. Table 1. represent the List of columns in the dataset, their data types and description. Here we will be predicting Order Quantity and Sales for the individual category.

*https://www.kaggle.com/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis
** http://dataco-global.com/

C. *Scope*

The dataset contains the historical data from the year 2015-2017. There are internal and external factors that would be significantly helpful in predicting sales. Website hits, Conversion Rate, Traffic Sources, Products, Promotions, Stocks are the internal factors that can help us accurately predict Sales. Competitors, High Demand Period, Trends are the external factors. The dataset doesn't comprise this information. Using category-wise sales data, we would be predicting the Sales.

Product Price, Order Price, Order Discount Amount, Discount Percentage, Order Profit Percentage, Profit per Order, Order Item Quantity are the factors that are present in the dataset. Using these factors, we can predict the weekly, monthly, quarterly sales for all the product categories available.

II. Goal

The objective of this project is to implement an appropriate machine learning algorithm/technique which will successfully predict the Sales for the individual product category which is being sold by the company.

The secondary objective of this project is to find out key insights from the available data which will help us improve sales, revenue, profit for the organization. Insights will be found using exploratory data analysis methods and data visualizations.

III. Ethical Concerns

One of the important factors in Data Analytics is data itself. Data is gathered from various sources using multiple techniques. Data gathering requires effort and time due to which is not viable to gather data every time for research or analysis [4]. Therefore, it is preferred to reuse the available datasets to complete the research or analysis but there are some ethical concerns when we are using publicly available datasets. Below are the ethical concerns considered for using a publicly available dataset related to e-commerce sales data [5].

A. *Level of Data Trustworthiness:*

Reliability and trustworthiness are important factors when reusing the data for analysis. Data could be obtained from unknown repositories. There is no commonly accepted measure or standards to verify the credibility and quality of the data. The dataset is obtained from the *Kaggle repository due to which we are considering data to be credible, reliable and trustworthy.

B. *Informed Consent:*

It is necessary to avail the consent of the concerned people whose data is getting gathered and making their data available publicly for the research. The data is obtained by Data Co global from a supply chain company. Therefore, we are considering that consent of the users was taken before making the dataset available openly.

C. *Anonymity:*

Anonymity is an important factor when reusing the shared data. No information must be gathered which reveals the identity of the user whose data is gathered. The dataset doesn't contain any field which exposes the user's identity.

D. *Online Privacy:*

While making purchases online, users do use their vital banking details to complete the transaction. Also, electronic payments store the identity information of a buyer. It is the responsibility of e-commerce companies to not making identity and banking details available publicly. The dataset doesn't contain any information regarding identity and banking details.

E. *Stakeholder Identification:*

The Important stakeholders involved in this project are Retailers, Transport suppliers, Brands. Data related to these stakeholders is available in the dataset. The analysis done in this project will either help them or will affect them in a negative context.

IV. Business Value

It is important to adopt appropriate and advanced business strategies that are data-driven to improve sales and revenue. The exploratory analysis and visualizations conducted in this project significantly benefit the creation of business strategies that will improve the sales, order count and revenue of the e-commerce business.

A. *Competitive Advantage:*

1) ***Seasonal campaigns and Trends*****:** If we look at different regions of the world the majority of sales occur in different shopping seasons. E.g. In the majority of European and Western countries the Summer and Winter vacation leads to the biggest shopping sales. In India, festivals and wedding season are the biggest sales season. By analyzing the data of this period, we can find the most popular and sold products which will help us build a shopping campaign that would help us increase sales and revenue. This would also increase brand awareness.

2) ***SMS, Email and Social Media Marketing:*** When users visit the webpage of a company, the information regarding their searches, view gets stored in the cookies also when user registers and make first purchase information such as email id, phone number address gets captured. This information is useful in creating customized, dedicated ads, recommendations for individual customer

3) **Customer acquisition and Churn Prevention*:***

During the start of the e-commerce business, it is important to create and gather a large customer base that will create brand awareness. This can be done using different Marketing and Advertisement strategies. Later when the customer base is created it is necessary to analyze Churn.

Churn is defined as a user or customer who has stopped consuming the service or products. On an e-commerce platform, it is difficult to categorize which user has turned into a churn or no. Churn identification helps an organization find flaws in the service provided which later can be improved to regain the customer. After churn identification next step is to use digital marketing to convert the churn into repeating consumer [6].

B. *New Revenue Opportunities:*

The e-commerce business can adopt multiple different revenue generation methods apart from sales of the product.

1) **Advertising Revenue:** The E-commerce platform can allow other businesses to display advertisement which is not direct competitors of own business. This becomes an important alternative income source that can earn revenue from Cost per Click and Cost per Action [7].

2) **Subscription-based Model:** The e-commerce business can use a subscription-based model which would help them generate revenue monthly, quarterly, yearly by providing subscriptions. This can include providing Priority deliveries, Free deliveries, No Cost EMI Options, additional discounts, Early access to discounted products [7].

3**) Transaction Fee:** The sellers on the e-commerce website can be charged with the transaction fee on each order payment. As the fee is charged to the seller and not the buyer it doesn't impact the sales and creates an additional way of revenue generation [7].

4) **Loyalty Programs and Referrals:** Normal customer base gets converted into loyal customer base due to implementation of loyalty programs in which user is encouraged to buy repetitively to obtain additional discounts and benefits. Referrals are one of the important methods to expand the customer base which helps increasing sales and brand awareness as well [8].

C. *Increased Efficiency:*

The prediction of Order count and Sales category-wise in advance helps the organization to improve the efficiency of the process from Product Stocking, Order Confirmation to Order Labeling, Billing, Packaging then Shipping and Delivery. The process from Stock Arrival to Product Delivery is very complex, First Labeling, barcoding of the stock is done then they are reflected on the Website Looking at the Humungous quantity of the products this process needs to be automated and efficiently managed. Therefore, prediction done gives more time frame to complete these tasks does increase the efficiency of the further processes.

D. *Better Customer Service:*

Customer satisfaction is important for every business, in the e-commerce business there are several factors on which customer satisfaction dependent. Simple, Easy to use the platform, Advance and Correct Product Filtering, Better and Intelligent Recommendations, hassle-free Payment, Accurate Order details and Payment Tracking, Good Packaging and Product Quality. By Analyzing and Understanding customer preferences, the use of predictive analysis would lead to better customer satisfaction as the organization is doing what customer is expecting.

E. *Increased Profitability:*

Predicting the Sales and Order Quantity category-wise would lead to improved and efficient inventory management. By predicting the upcoming demand and trend the stocks can be built earlier which would cost less than paying more for the same stock when it is trending thus resulting in increased profit. Also, the trending products will be available so that buyers will not be found a shortage of them also it gives an edge over competitors as they might not have the stocks available, or it might be available at a higher price.

V. Preliminary Visualizations

Now we will look at the Key Insights from the dataset with the help of Visualization. Visualizations help us find out Patterns, Trends, Relationships easily.

A. *Preferred Shipping Mode:*

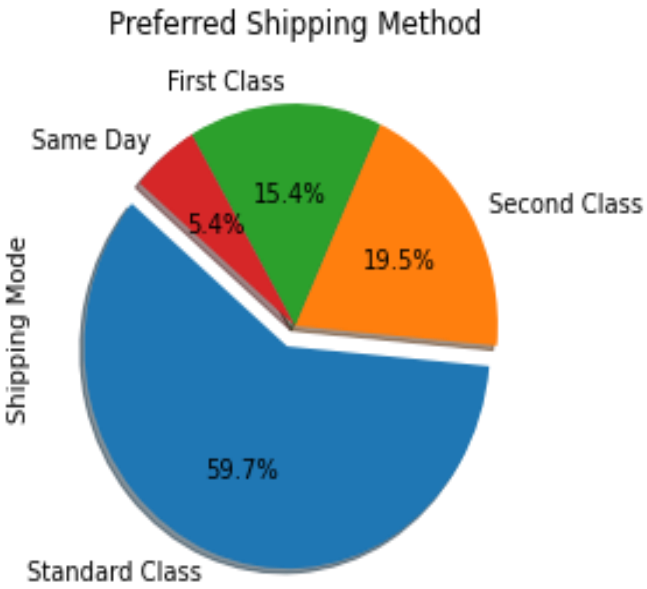

Fig 1. Preferred Shipping Method

Here we have done the analysis based on preferred shipping mode by customers. Out of 1,80,520 orders 60% of orders have Selected Standard Shipping mode. Strategies need to be implemented so that customers will start selecting Same Day, First Class, Second Class shipping mode more.

B. *Preferred Payment Mode.*

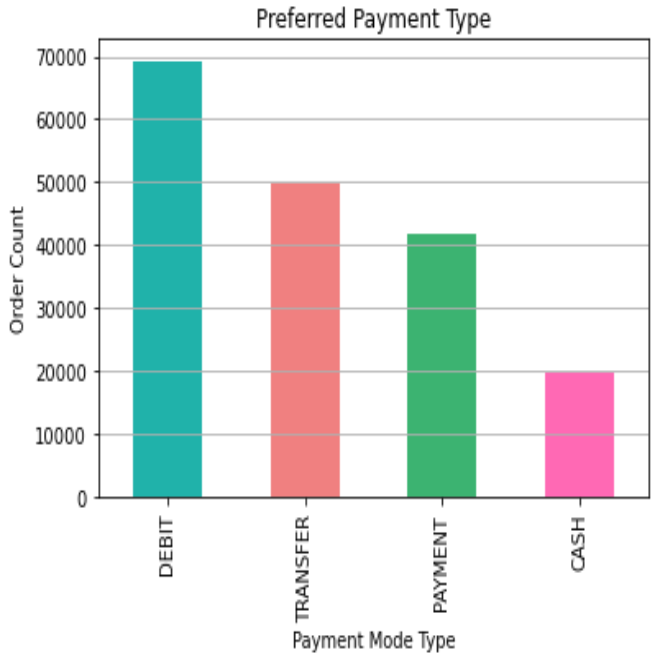

Fig 2. Preferred Payment Mode

Fig 2. Indicated the preferred Payment mode by the customers of the e-commerce company. This is necessary because when the customer selects the electronic payments method then the order amount gets added to the bank account within no time but when the customer selects the cash on delivery as a payment method the Order amount takes time to reach the bank account as the customer pays when delivery is completed and after that, the cash gets deposited. So, allowing electronic payment modes helps the organization to maintain the cash flow.

C. *OrderCount as per Delivery Status:*

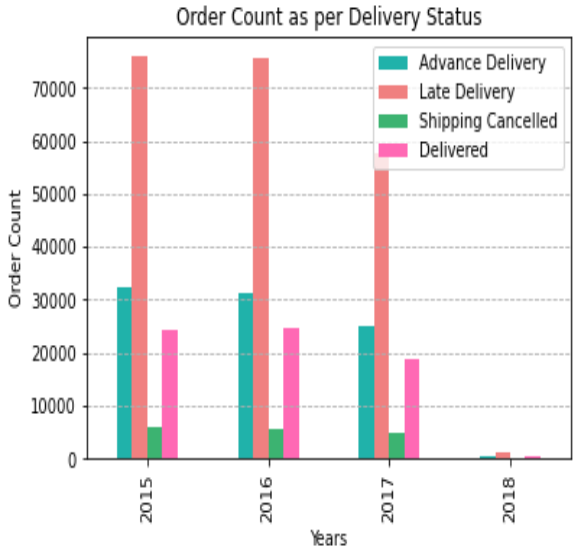

Fig 3. Order Count as per Delivery Status.

For an e-commerce organization, customer satisfaction is important for business growth. To achieve that orders placed by customers need to get delivered within the promised delivery timeline. From Fig 3. We can interpret that majority of orders are getting delivered late. This could lead to bad customer satisfaction and can harm business growth.

D. *Sales Trend from Year 2015-2017 Monthwise:*

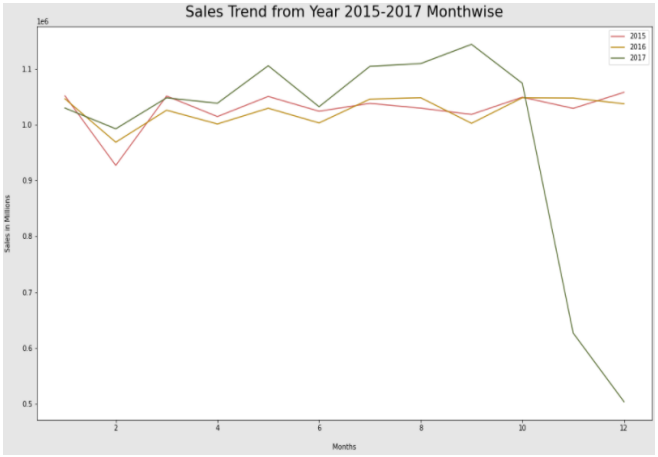

Fig 4. Sales Trend from Year 2015-2017 Month wise

The Fig 4. Visualizes the amount of sales for a particular month and year. This helps us analyze how sales of e-commerce performing every year and what are the months in which sales have increased or decreased. Here X-axis represents the Months of the Year while Y-axis Represents Sales in Millions. We can also identify that sales in the last quarter of 2017 have drastically decreased.

E. *Profit Trend from Year 2015-2017 Monthwise:*

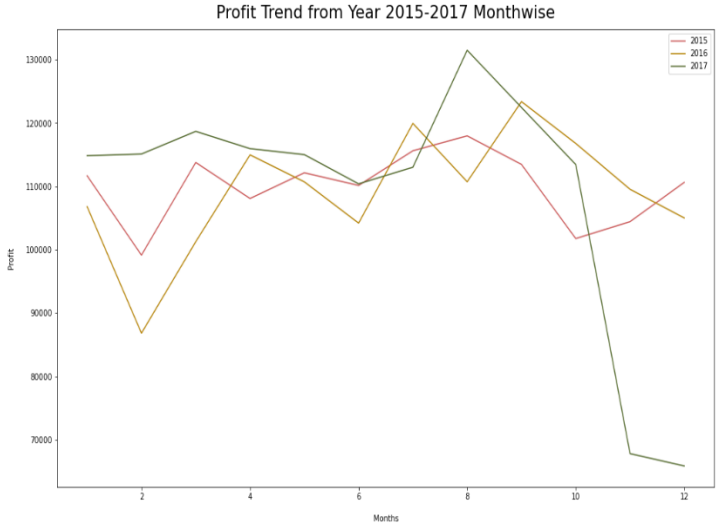

Fig 5. Profit Trend from Year 2015-2017 Month wise.

Profit is important for any business and if the business is not profitable it will not sustain for long enough. Here we can see that for February and June profit has decreased in all three years. Analysis needs to be done to find out the reason for decreasing sales in those months.

F. *Sales by Product Category:*

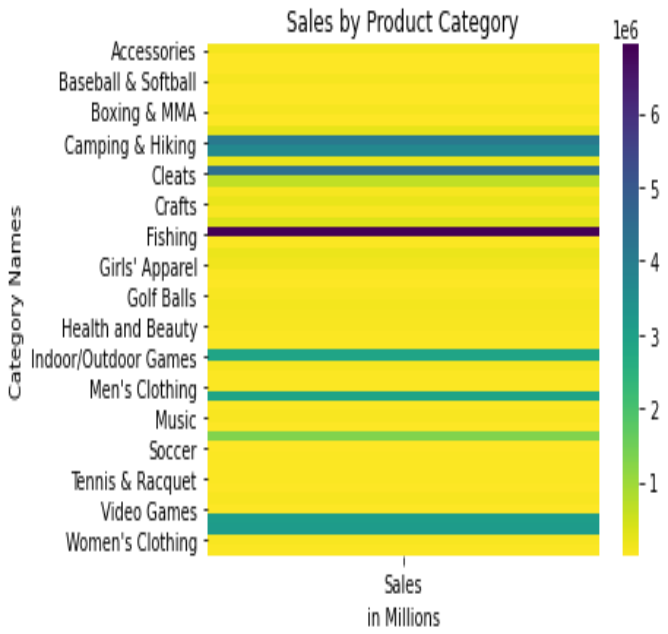

Fig 6. Sales by Product Category

For an e-commerce company, it is necessary to understand which product category is getting sold highly and which product category is selling less. Here Fishing, Cleats, Camping & Hiking, Men's Clothing, Indoor / Outdoor Games, Video Games, Women's Clothing were sold most but the rest of the other categories are not getting sold. Analysis and Strategy need to be done to improve the sales from other categories as well. Here Yellow color represents the categories with Fewer sales. The green and blue color represent the Intermediate and Good sales respectively.

G. *Order Count by Product Category:*

Fig 7. Order Count by Product Category.

Fig 7. Shows the Order Count for a particular order category. This is a vital insight for an organization to know which product categories are getting ordered more and which are getting ordered less. According to the products, demand company can plan for inventory management in which products with high demand or stocked more than products with less demand. Red color represents the categories that are ordered 10,000 or less than 10,000. Yellow, Orange color represents the categories with order count from 40,000-20000. The green color represents the categories with an order count from 7,0000 to 50,000.

H. *Order Count by Department:*

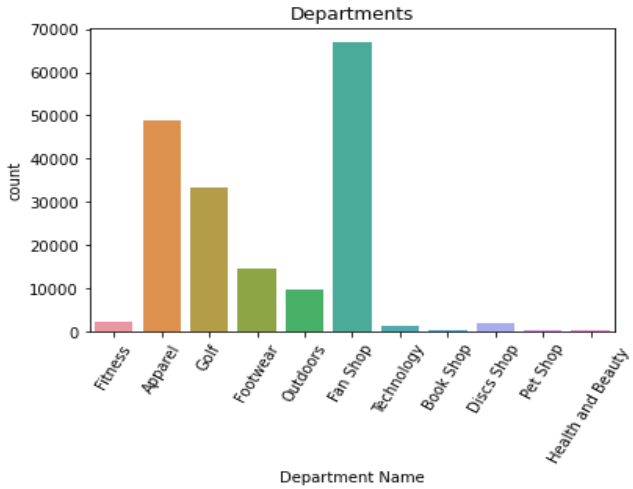

Fig 8. Order Count by Department.

The product categories get generalized into product departments. All the Departments need to generate similar order purchases for the growth of the business but from Fig 8. We can see that this not the case here. Only a few departments are having a good order count.

I. *Sales by Country:*

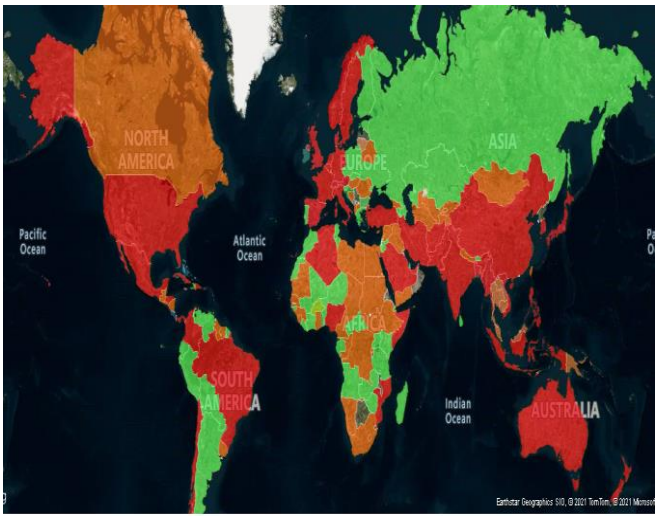

Fig 9. Sales by Country

Fig 9. represents the world map which indicates the amount of sales generated from every country from years 2015-2017. After hovering on a particular country, it provides the exact amount of sales. Here Green color represents the countries with significant sales ranging from 71-90 Million. Orange color represents the countries with sales 41-70 Million and Red color represent the countries with 10-40 million.

J. *Sales by Market List:*

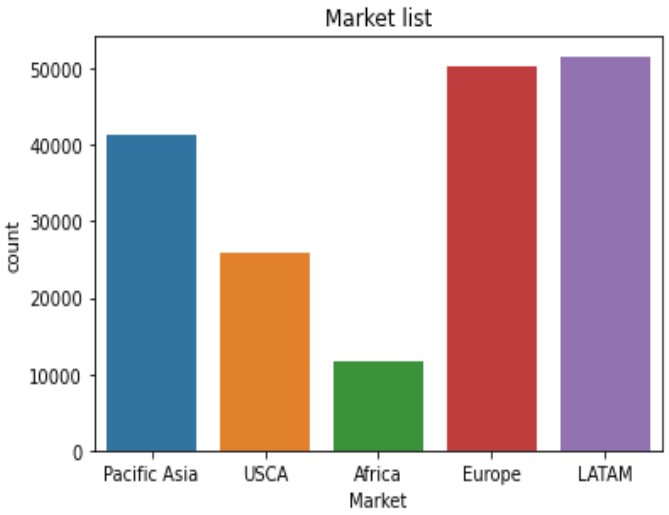

Fig 10. Sales by Market List:

The above Fig 10. Represents the Sales by respective Market. Businesses must capture the entire market to be the market leader. That's why it is necessary to analyzes the sales by Market. If sales in a particular market is less then it could affect the other market regions as well. Also, the market with fewer sales becomes the growth potential.

K. *Order Count by Region:*

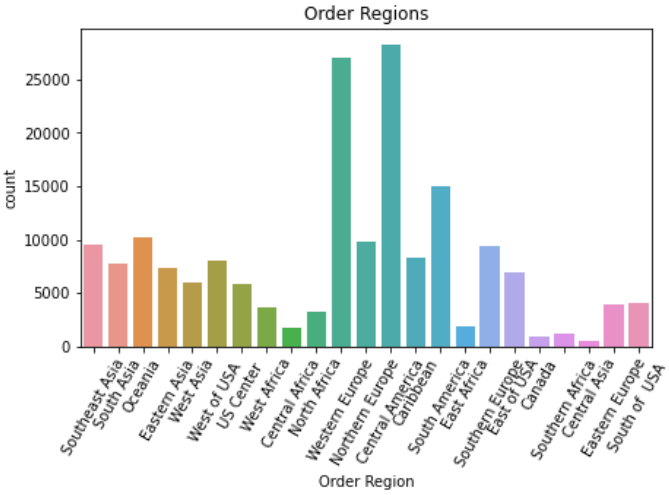

Fig 11. Order Count by Region

The world is divided into different regions and every region is having different countries which show similar kinds of culture and trends. To be the market leader, the organization first needs to be the focus on these regions. Fig 11. Shows the order count by the regions. Regions with low order count become potential regions where sales and orders could be increased more.

VI. Applicable Techniques

In this project, we are predicting the Sales amount which is continuous variable. Here we can use Supervised and Unsupervised methods to predict the target variable. We can consider different regression and time series algorithms for prediction purposes. The implemented model can be evaluated using R Square, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE).

1) Support Vector Regression.
2) Lasso Regression (least absolute shrinkage and selection operator.
3) Multiple Linear Regression
4) Long Short-Term Memory (LSTM)
5) ARIMA & SARIMA
6) KNN Regression,
7) Partial Least Squares & Principal component regression (PCR & PLS).

VII. References

[1] UNCTAD, "Global e-Commerce hits $25.6 trillion – latest UNCTAD estimates," UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT, 27 04 2020. [Online]. Available: https://unctad.org/press-material/global-e-commerce-hits-256-trillion-latest-unctad-estimates. [Accessed 25 03 2021].

[2] V. K. S. B. S. B. S. K. a. B. S. A. Jain, "Demand Forecasting for E-Commerce Platforms," IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020 .

[3] J. Y. Y. a. Y. J. F. L. Wu, "Data Mining Algorithms and Statistical Analysis for Sales Data Forecast," Fifth International Joint Conference on Computational Sciences and Optimization, Harbin, China, 2012.

[4] J. &. T. M. Bote, "Reusing Data: Technical and Ethical Challenges," DESIDOC Journal of Library & Information Technology., Barcelona,Spain, (2019).

[5] C. &. L. R. &. G. I. &. S. J. Kopp, "Ethical Considerations When Using Online Datasets for Research Purposes," DESIDOC Journal of Library & Information Technology, Barcelona,Spain, 2016.

[6] M. S. H. a. R. M. R. N. Forhad, "Churn analysis: Predicting churners," Ninth International Conference on Digital Information Management (ICDIM 2014), Phitsanulok, Thailand, 2014.

[7] M. Demchenko, "Types of eCommerce Business and Revenue Models," ncube.com, 23 June 2020. [Online]. Available: https://ncube.com/blog/types-of-ecommerce-business-and-revenue-models. [Accessed 25 03 2021].

[8] i. Logistics, "A Detailed Guide on E-commerce Revenue Model 2020," iThink Logistics, 21 July 2020. [Online]. Available: https://ithinklogistics.com/blog/a-detailed-guide-on-e-commerce-revenue-model-2020/. [Accessed 25 March 2021].