Developing revenue increasing strategies for a Database Engineer of a computer hardware company using Customer Segmentation, Demand Forecasting and Customer Lifecyle Value Analysis

A Final Submission report for the BDM capstone Project

Submitted by

Name: Manoj

Roll number: 23ds2000018

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1 Executive Summary:

This project focuses on the sales data of computer, printing, and other hardware components of the multinational technology company. The data, which I would end up procuring, will primarily focus on transaction details between distributors, retailers, and customers of inventories around the world. And for market security reasons, the names of the retailers, distributors, and customers would be changed or fabricated.

The main business problem here is to basically develop methods to increase the sales and revenue of these products that can be easily interpreted by a backend database engineer. By developing these methods, the backend database team could procure data in a more effective way and make the company concentrate more resources, inventory space, and marketing on a particular group. The objective here is to procure the historical transaction data, clean it, and analyze it based on attributes such as units sold, revenue generated, propensity of buyer, expiration date of products, etc. based on different hierarchies. Some of these hierarchies include product type/category wise and geographical wise.

Using the results of this analysis, customer segmentation, demand forecasting using Machine Learning models, and Customer lifecycle value analysis are done to focus marketing and resources on a particular product group or a particular geographical location.

<u>Going over the key attributes of the data:</u>

1) Data From - denotes whether the sale is sell-to (From Retailer to customer) or sell-through (From Distributor to Retailer)

2) Quarter - Denotes the quarter at which the sale happened. It could either be a calendar year (January to December) or a Fiscal Year (November to October)

3) The next set of attributes are a hierarchy denoting the category of the products, the product sold belongs to. The hierarchy from top to bottom goes: Business Unit -> Chai Mapping -> Product Segment -> Product group -> Product Categories -> Product line. The number of unique categories of each attribute reduces as you go down the hierarchy.

4) The next set of attributes are a hierarchy denoting the geographical location of the customers. They are structured in 3 levels - Partner Market (Continent), Partner cluster (Cluster of countries) and Partner country.

5) The next set of Attributes are regarding the partner customers, which have been replaced with dummy values for data security. As this list contains only the sales data of power customers, all are marked as power customers.

6) The data also shows the total revenue of each of the purchase, in both the local currency and United States Dollars and the number of units sold of the product.

7) Finally, many other information about the purchase like, 'Other Party Site Instance Identifier', if the purchase was made online or not etc. were also procured. But the number of entries for these columns are missing for majority of the entries.

8) I was also able to procure two more supporting databases with valuable information like, Propensity of the customer to buy the same product again, expiration date of already purchased products and end of warranty dates.

2 Analysis and Process:

- Data Pre-Processing: The procured data, will have a lot of useless attributes/columns, missing values, data in unintended formats, etc. So, the first step was the remove all the useless columns, which contributed nothing to actual analysis of sales such as IDs, currency types etc. Next, we saw two hierarchies present in the data – For the product and the Geographical location of the customers. So, only one attribute from each of these hierarchies was kept. Next, columns with more than 50% of its row value missing were removed. Finally, for time-series demand/revenue forecasting, the quarters in which the sale took place was converted to datetime format. Next, for the missing date values, the revenue and units sold columns were interpolated such that, each and every single day from 2018 to 2023 has a separate value. This interpolation was done based on the corresponding quarter revenue and units sold values.

1) Customer and product segmentation: For this Mid-Term submission, the pre-processed data was plotted in graphs and analyzed, for customer and product segmentation analysis. While the required graphs have been plotted, the final inference and takeaway from these graphs will be done thoroughly and presented in the final submission, along with some more detailed graphs such as Pareto Charts for Gross Revenue and Units Sold. The process of this was described in detail in the project proposal.

   Customer segmentation can help in understanding the buying behavior of different customer groups, enabling targeted marketing campaigns and personalized product offerings. For this analysis on the sales data to evaluate the performance of different product lines, groups, or segments will be conducted. Identification of the top-performing products in terms of units sold and revenue generation, as well as their contribution to overall profitability will be done. This analysis can help in determining which products should be prioritized for marketing efforts or potential product line expansions.

2) Demand and Revenue Forecasting: So, for this analysis, the pre-processed data was first split into training and test dataset. Next, the time-series forecasting method ARIMA was used to develop a model using the training set and was tested upon the test set. The error percentage was also calculated for each of the forecasting done by the ARIMA model.

   Accurate demand forecasting can assist in optimizing inventory management, production planning, and supply chain operations, ultimately reducing costs and improving customer satisfaction.

3) Customer Lifetime Value Analysis (CLV): This analysis will be done and its results will be presented in the final submission. By understanding the value of each customer over

their lifetime, you can identify high-value customers for retention strategies, allocate marketing resources effectively, and make informed decisions regarding customer acquisition and loyalty programs. For this, the data on customer's products' expiry date and warranty periods will be useful.

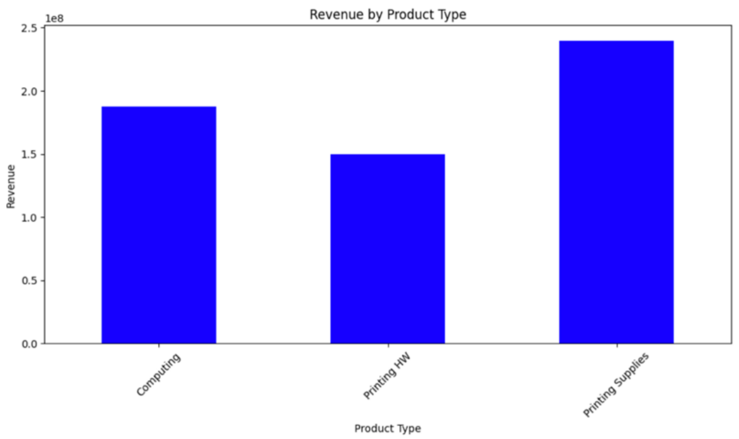
3 Results and Findings:

Python Notebook Link:
https://colab.research.google.com/drive/1TpKE2BwaR_1GWhzidGYQO6w57oTGb_4G?usp=sharing

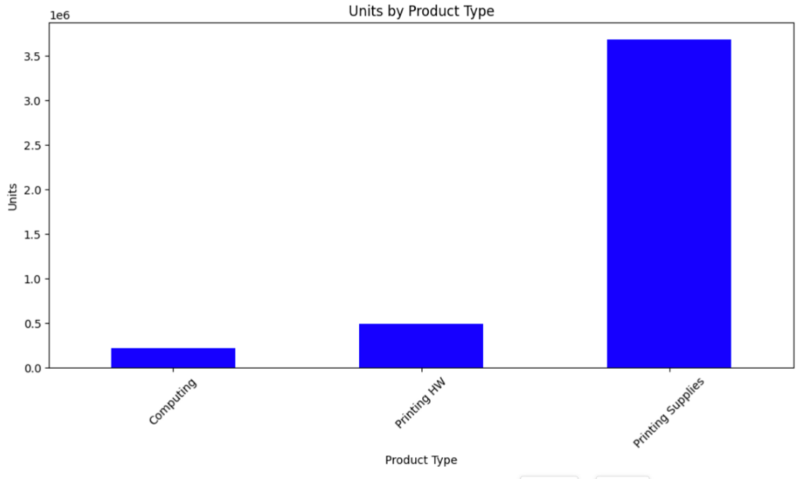
I. <u>Customer and product segmentation:</u>

For Customer segmentation and product analysis, I'm going to choose one major attribute from the two different hierarchies available in the data to compare against both number of units sold and revenue. From the product type hierarchy, I'll be choosing the 'BUSINESS_UNIT' attribute which majorly divides all the available products into three main categories, namely: Computing, Printing Hardware and Printing Supplies. From the geographical location of customers hierarchy, I'll be choosing the attribute 'PARTNER_MARKET', which broadly divides the customers into seven different regions/continents.

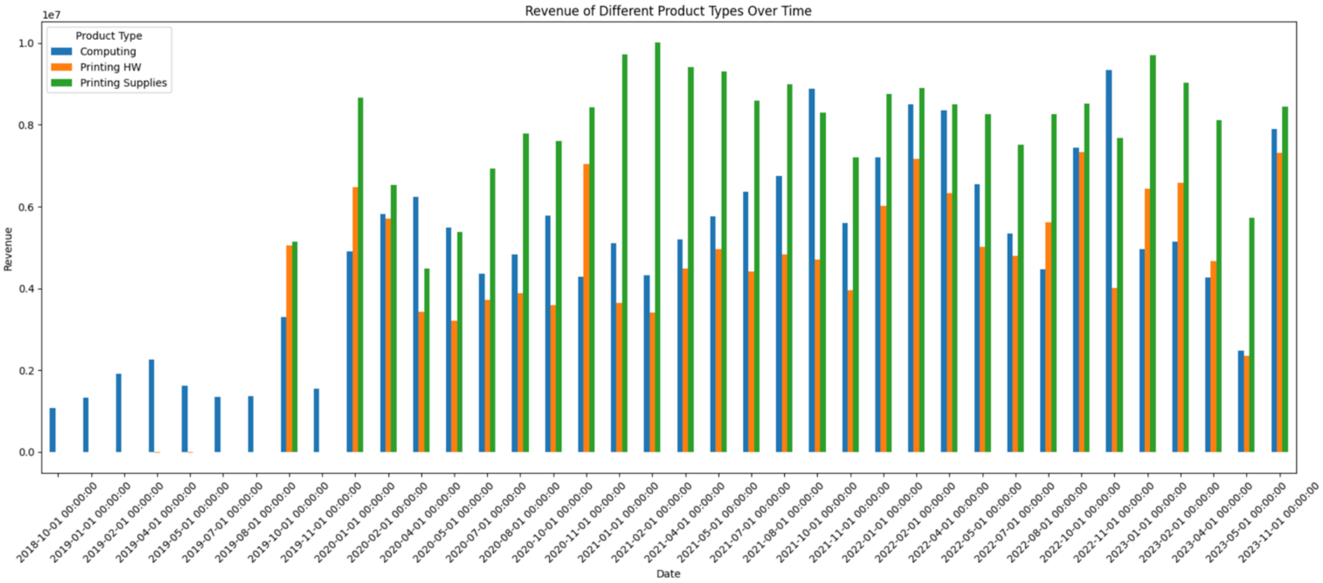
(i) Visualizing how sum of Revenue of the Products and Number of Units sold are distributed among the product types:



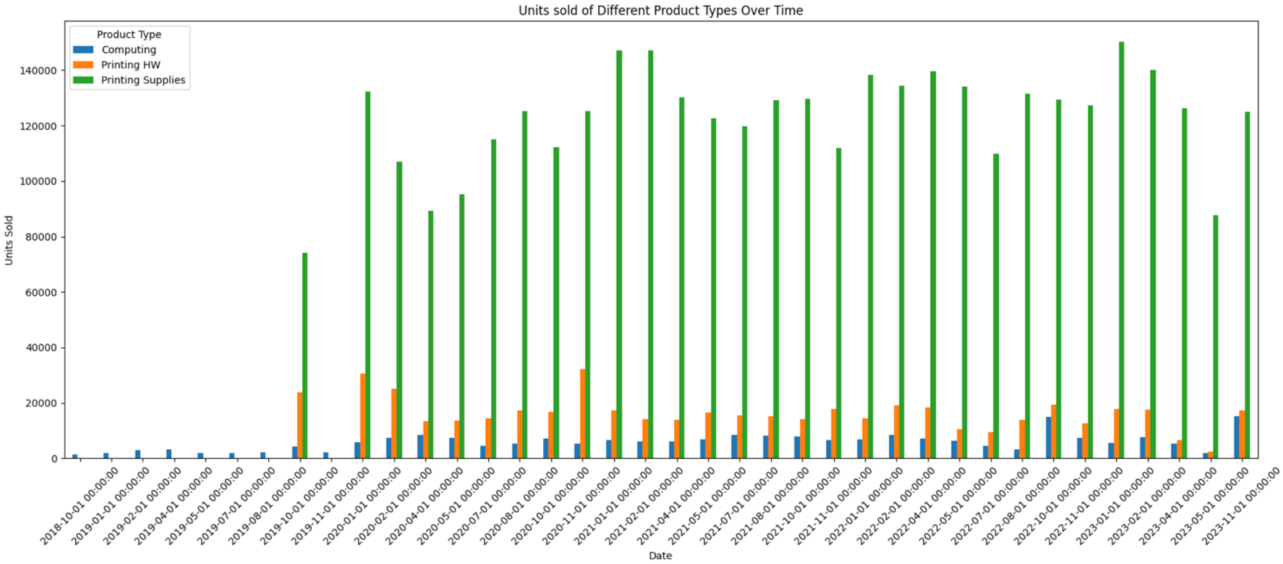
- We can observe here that Printing supplies procure the most revenue (About 230 million dollars), but the revenues of computing and printing hardware aren't far behind lagging.



- We can observe here that Printing supplies sell the greatest number of units, way higher than printing hardware or computing. From these two graphs, we can kind of infer that the cost of one unit of a printing supply is way lesser when compared to the per unit price of printing hardware or computing product.

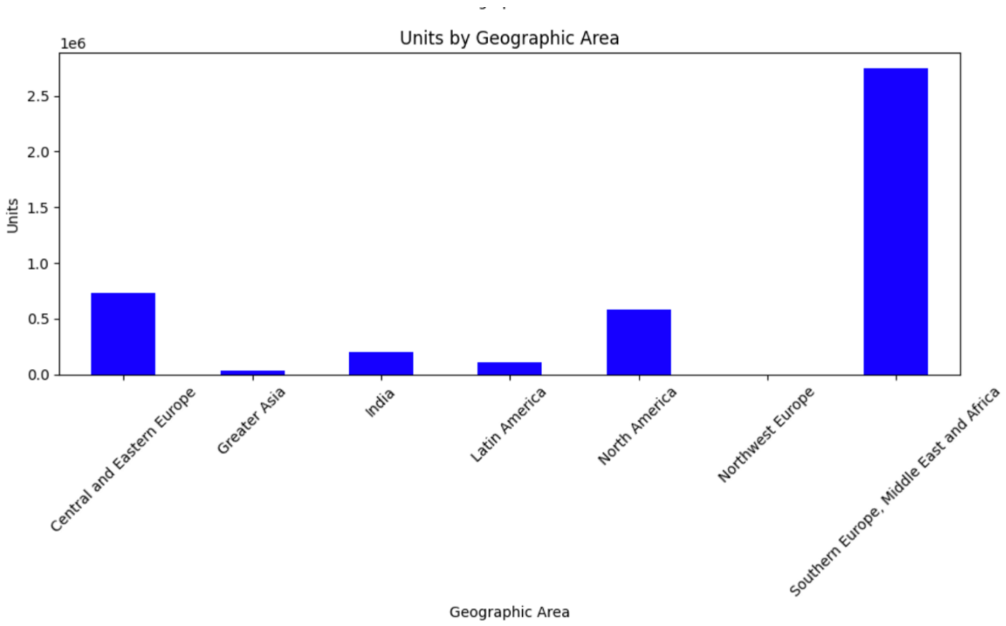


- This bar graph shows, how much revenue each type of product generated over a three-year time period from October 2018, till November 2023. Apart from showing the trend noticed from the previous bar graph, we can observe here that, computing has been generating revenue throughout the entire timeline whereas, printing supplies and hardware start generating revenue a year later.

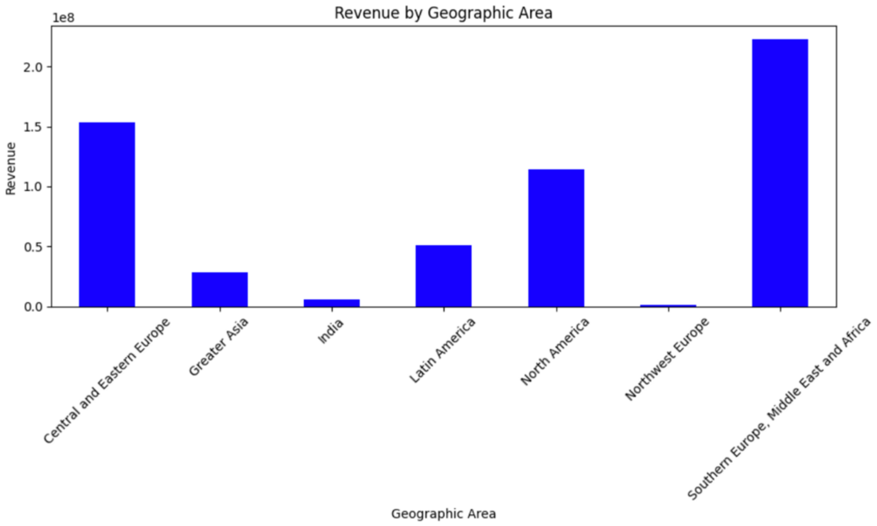


- This bar graph shows, how may units were sold for each type of product over a three-year time period from October 2018 till November 2023. Apart from showing the trend noticed from the previous bar graph, where number of printing supplies sold were high, we can observe the same trend mentioned from the previous graph.

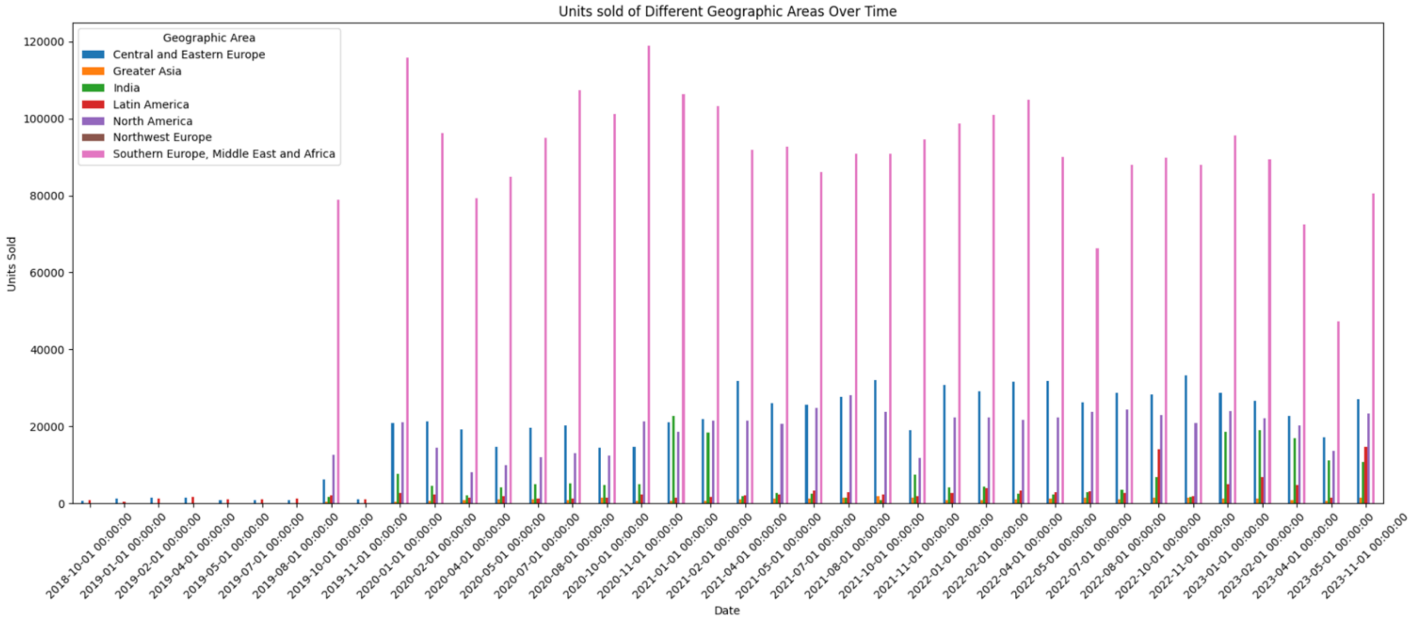
(ii) *Visualizing how sum of Revenue of the Products and Number of Units sold are distributed among the geographical areas:*



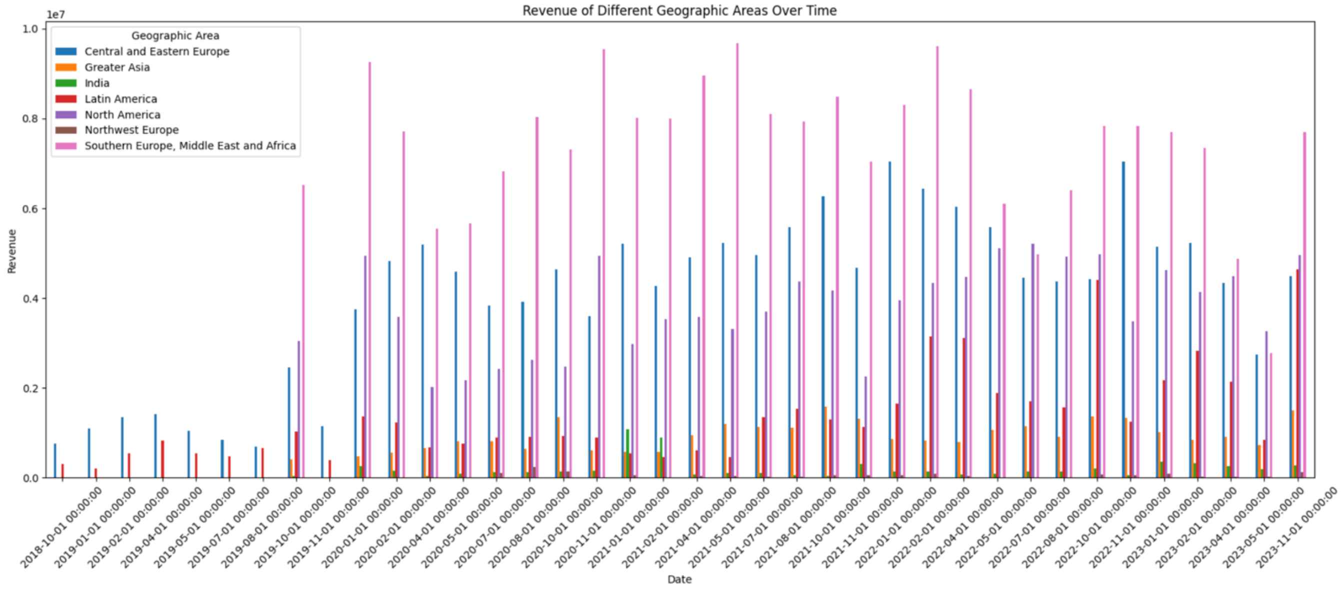
- We can observe here that, Southern Europe/Middle East and Africa (SEMEA) region has sold the greatest number of units, towering over other regions. Central and Eastern Europe has also sold some significant number of units as well, but others have sold very less.



- The revenue generated however is a very different story, while SEMEA region still dominates this list due to number of units sold being high, the second and third highest aren't far away. A bold prediction now would be that Central and Easter Europe must have customers who buy a lot of computing and printing hardware supplies.



- This bar graph shows, how many numbers of units sold for each type of geographical area over a three-year time period from October 2018, till November 2023. Not much of new trend can be recognized here, other than SEMEA region selling more units at all times.



- This bar graph shows, how much revenue each geographical area generated over a three-year time period from October 2018, till November 2023. Again, not much of a new significant trend can be recognized here, other than the small observation that initially, in 2018, central and eastern Europe generated more revenue the SEMEA region.

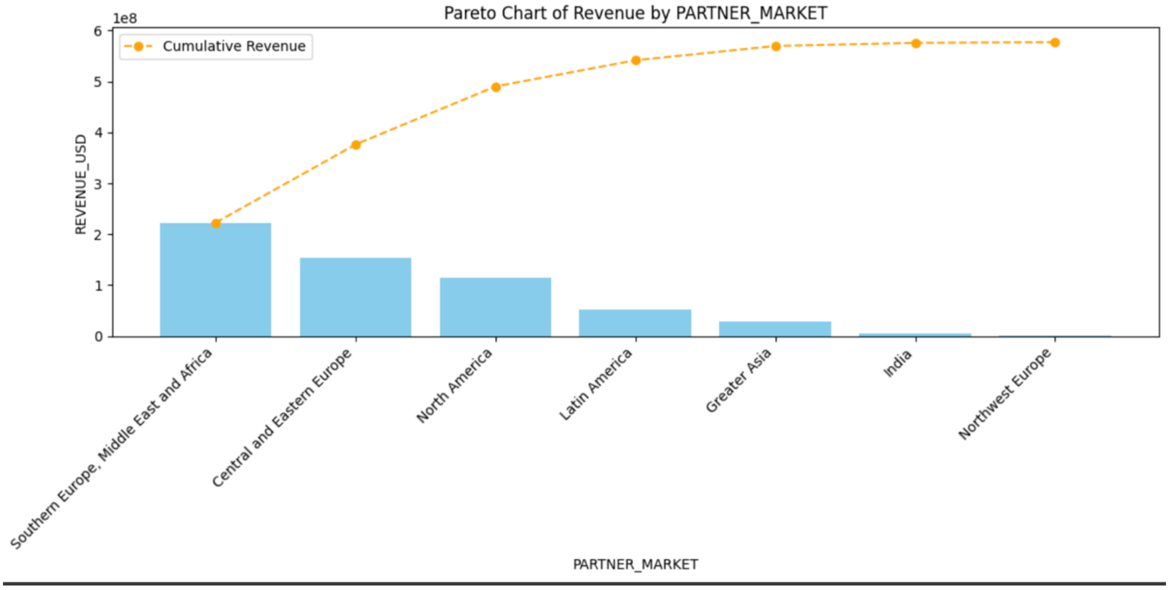
(iii) <u>Pareto Analysis:</u>

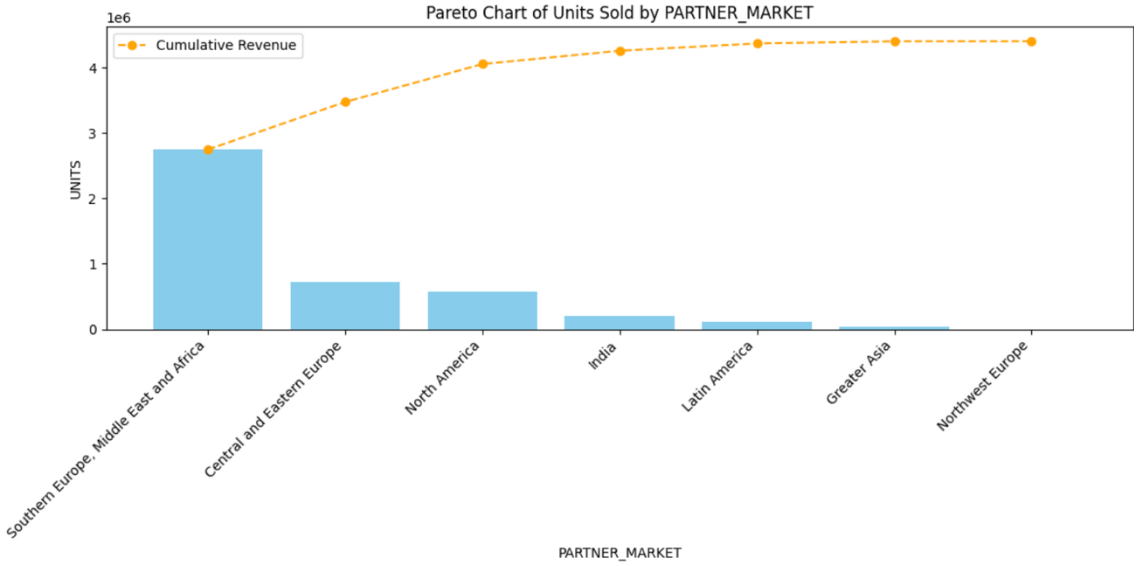
- Pareto analysis involves identifying and prioritizing the most significant factors contributing to a particular outcome or problem. Pareto analysis or the 80/20 rule, is a decision-making and problem-solving technique that is based on the observation that a small number of factors or inputs often have a disproportionately significant impact on outcomes or results. The principle is named after Vilfredo Pareto, an Italian economist who observed that approximately 80% of the wealth in Italy was owned by 20% of the population.

- To make this analysis, we have to create a Pareto chart, which is a bar chart that displays the categories in descending order of revenue generated or the number of units sold for each product type or geographical region. The most significant categories are on the left, and the less significant ones are on the right.

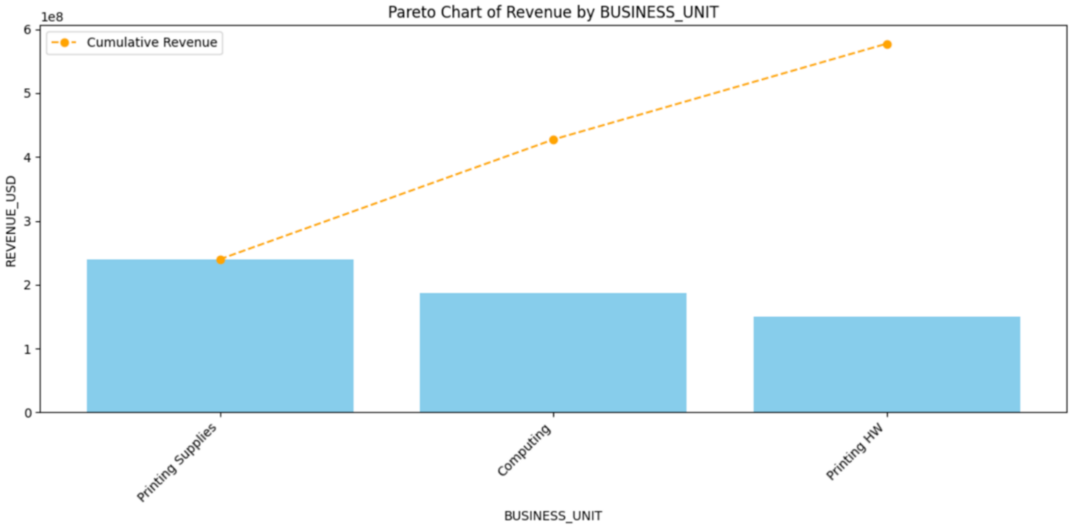
- Next, we should identify the product types/geographical regions that represent the most significant factors contributing to the problem. These are often referred to as the "vital few." In the context of the 80/20 rule, they represent the 20% of factors that have an 80% impact on the problem.

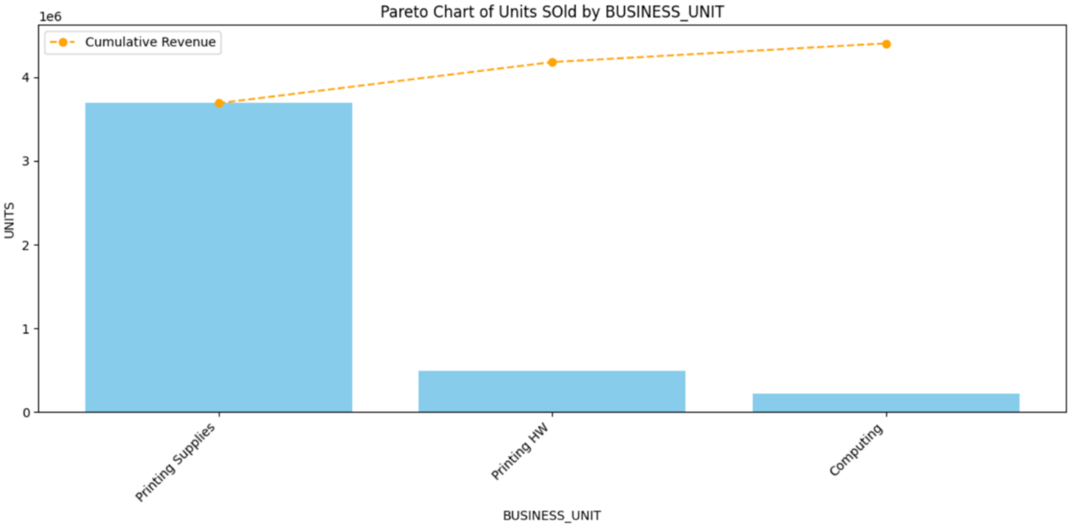
- Here are the pareto charts for different geographical locations. Both revenue and units sold, indicate SEMEA and Central & Eastern European regions dominate as the “vital few”.



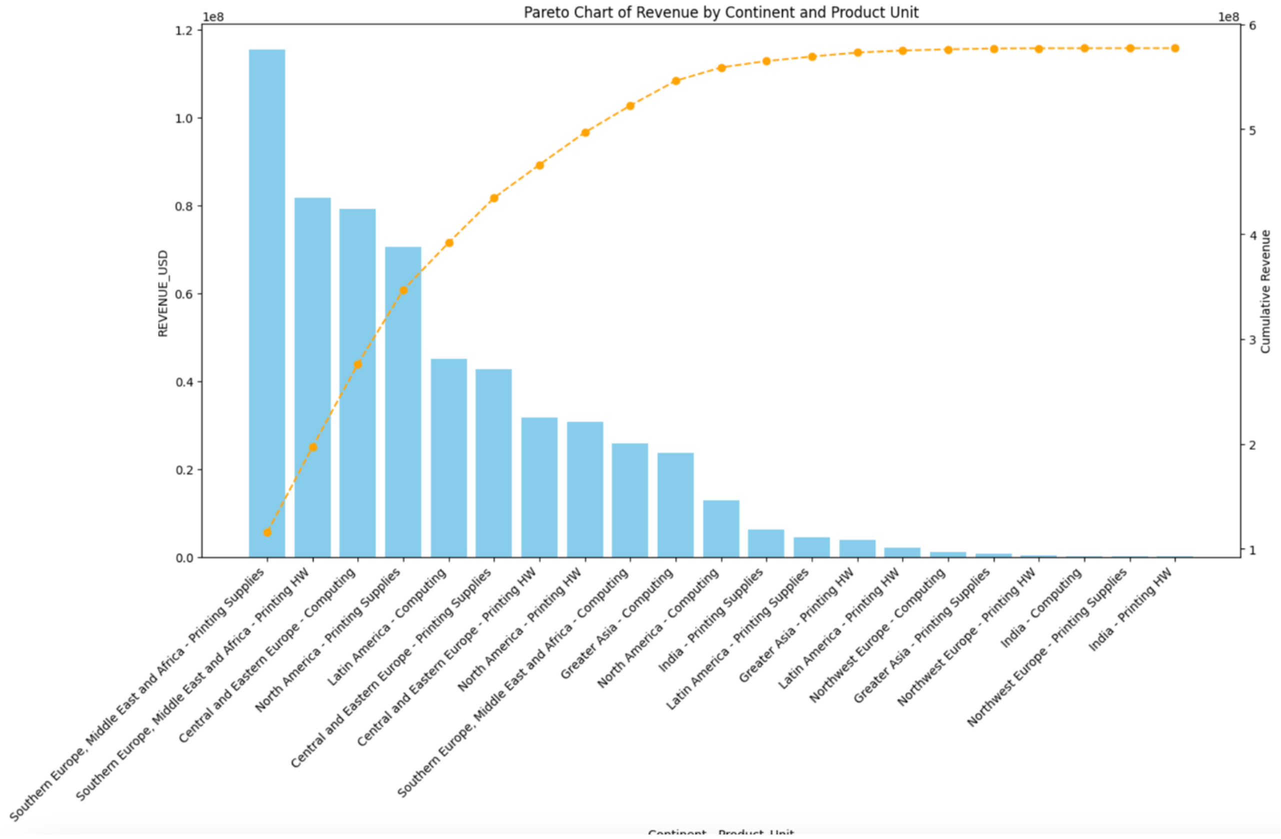


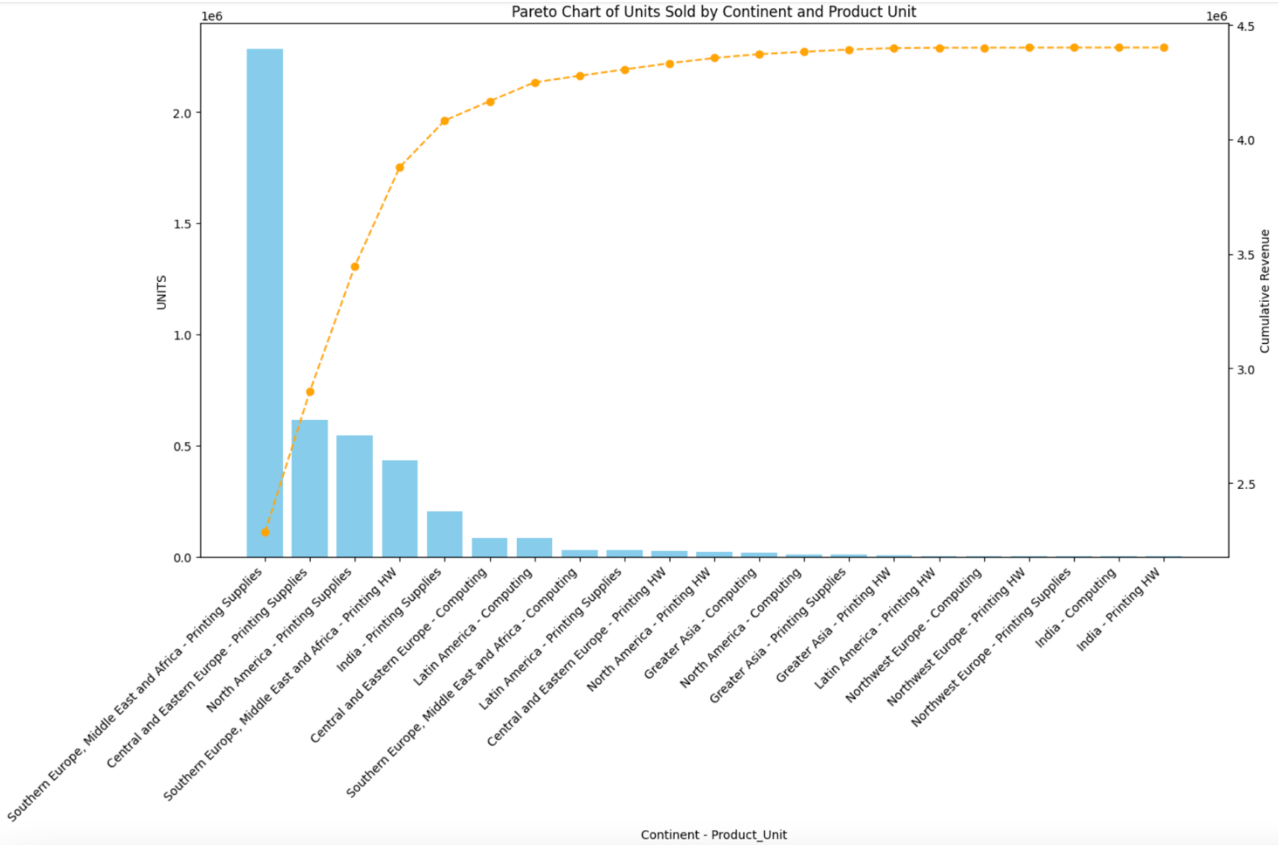
- Here are the pareto charts for different product types. Both revenue and units sold, indicate printing supplies should be considered as “vital few”





- Here are the pareto charts for different product type-geographical area combination combined. Based on revenue and units sold, the top four categories are chosen as the "vital few":



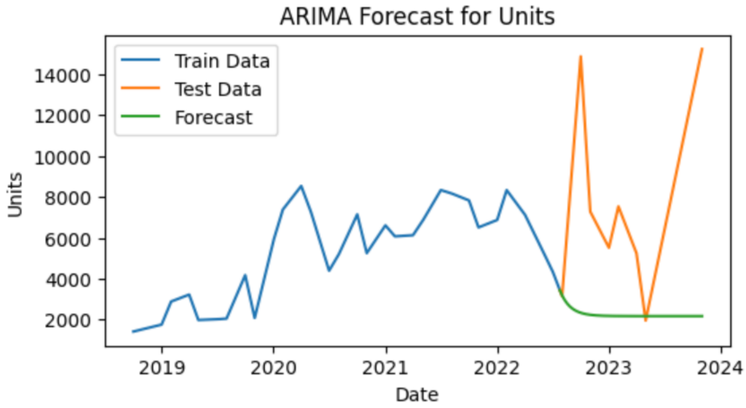


II. Demand and Revenue Forecasting:

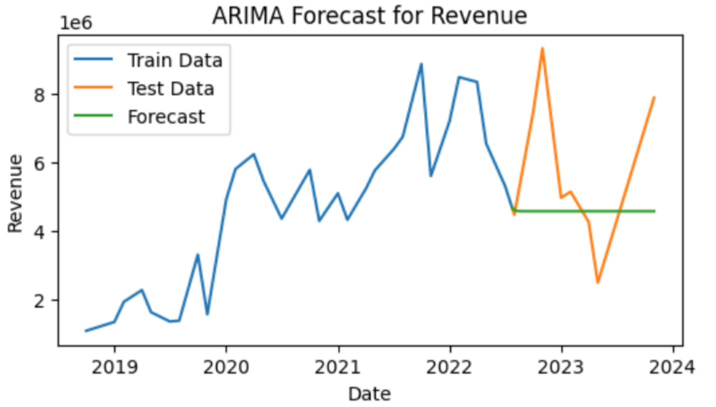
- Here, the forecasting has been done on units sold for each purchase, thereby predicting the future demand for each type of product and products sold in different geographical locations. The accuracy for each model/prediction is calculated with Mean Absolute Percentage Error or MAPE.
- MAPE is a commonly used metric for evaluating the accuracy of time forecasting models, especially in the context of forecasting time series data. It measures the percentage difference between the forecasted values and the actual observed values. MAPE is a measure of the relative accuracy of a forecast and is expressed as a percentage.
- MAPE remains a widely used metric for evaluating time forecasting models, as it provides a simple and interpretable measure of forecast accuracy in terms of percentage error.
- The accuracies of each of the predicted models are printed below the graphs.

(i) For different product types (BUSINESS UNIT):

```
--------------------FOR THE BUSINESS UNIT: Computing ---------------------------
```

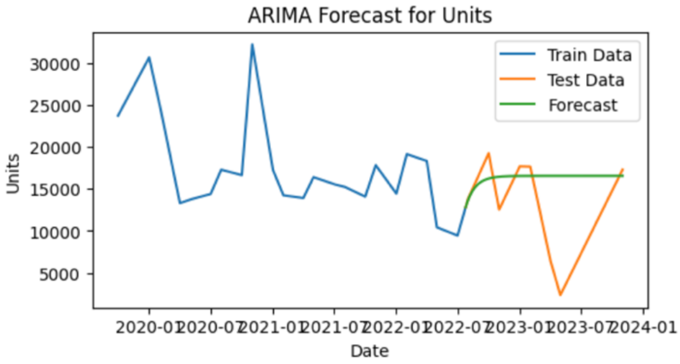


```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 63.64%
```

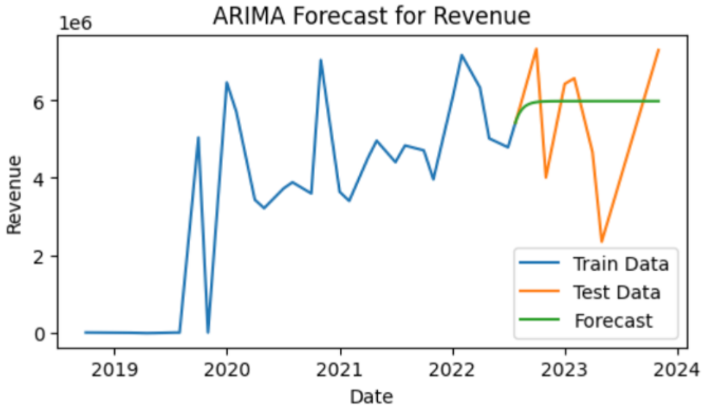


```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 25.38%
```

```
--------------------FOR THE BUSINESS UNIT: Printing HW ----------------------------
```

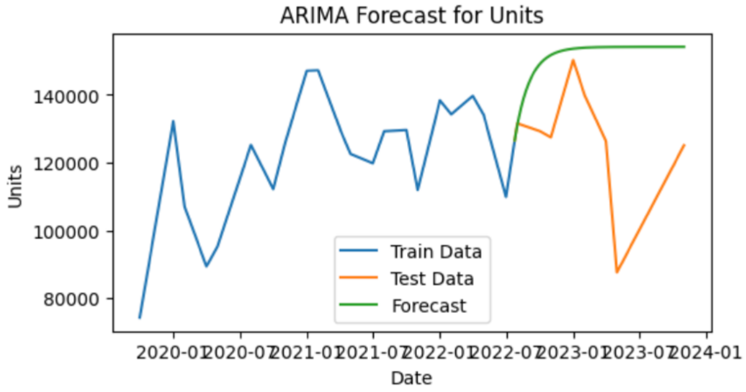


```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 77.45%
```

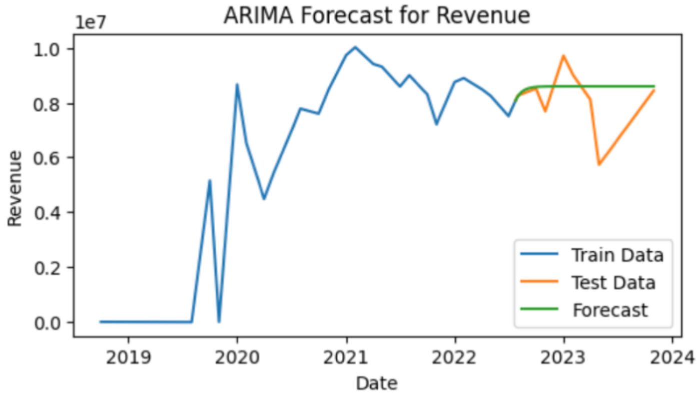


```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 28.28%
```

```
--------------------FOR THE BUSINESS UNIT: Printing Supplies ----------------------------
```



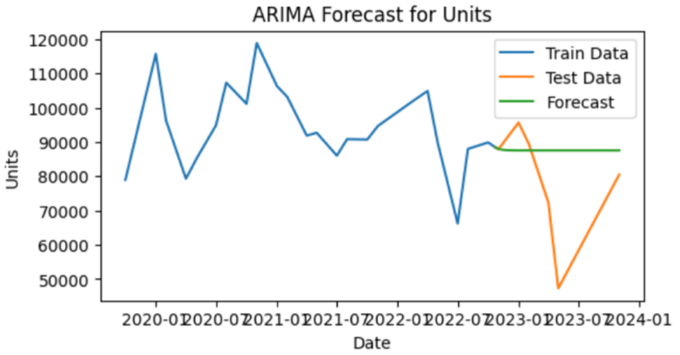
```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 27.48%
```



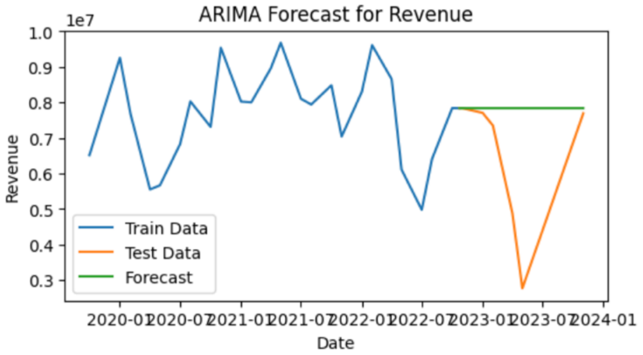
```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 12.84%
```

(ii) For different geographical areas (PARTNER MARKET):

```
--------------------FOR THE PARTNER MARKET: Southern Europe, Middle East and Africa ----------------------------
```

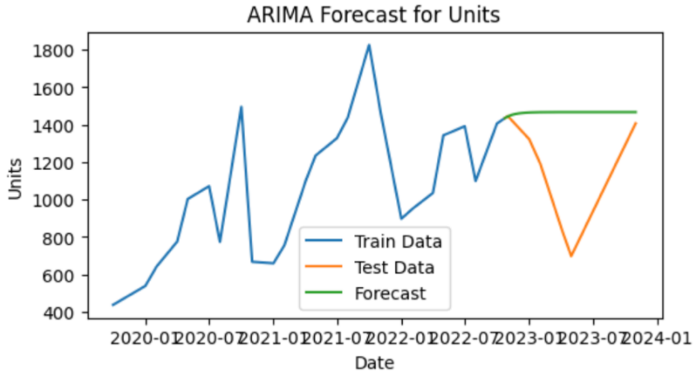


```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 26.23%
```

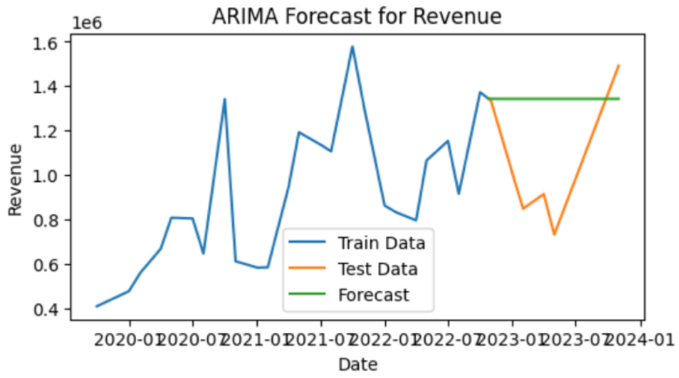


```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 44.91%
```

```
--------------------FOR THE PARTNER MARKET: Greater Asia ----------------------------
```

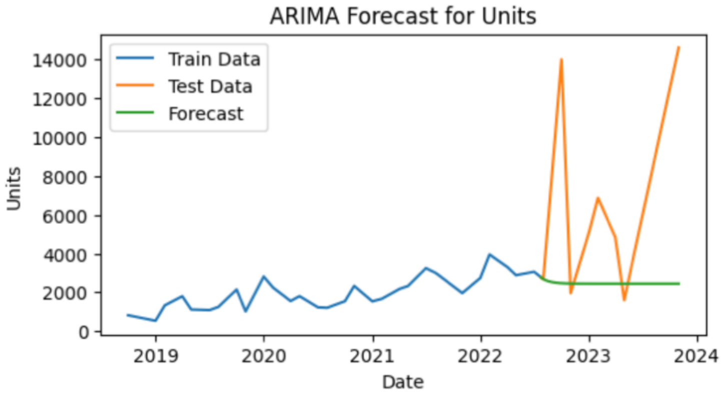


```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 38.58%
```

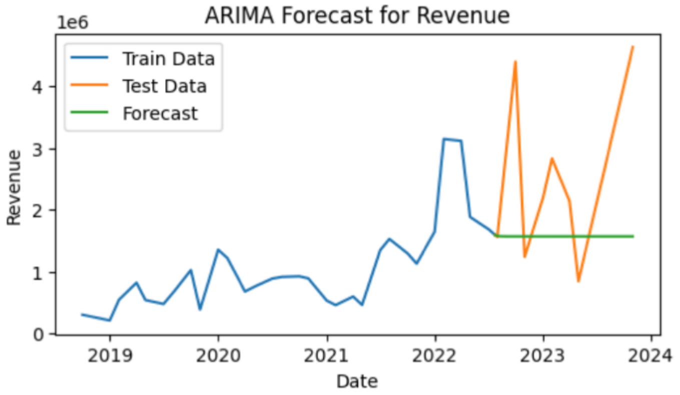


```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 33.48%
```

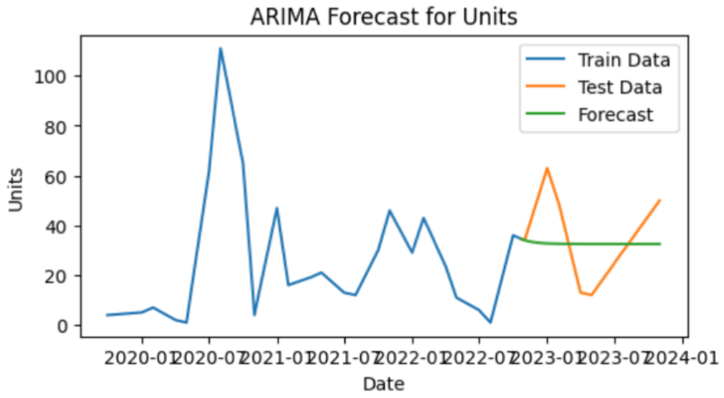
```
--------------------FOR THE PARTNER MARKET: Latin America ---------------------------
```


ARIMA Forecast for Units

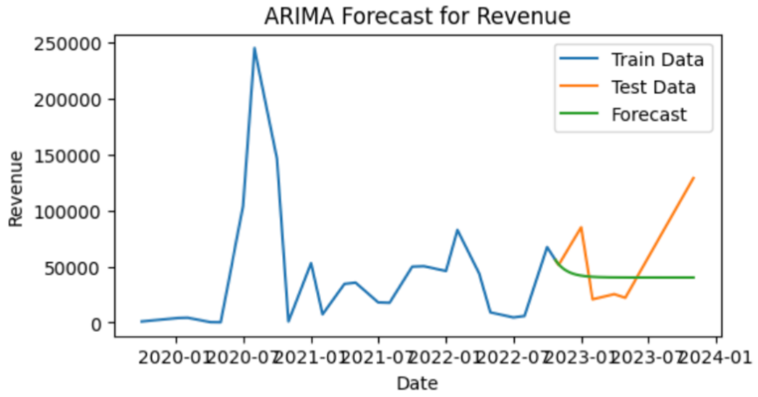
```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 53.50%
```


ARIMA Forecast for Revenue

```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: [illegible]
--------------------FOR THE PARTNER MARKET: Northwest Europe ---------------------------
```


ARIMA Forecast for Units

```
Mean Absolute Percentage Error (MAPE) for FUTURE DEMAND: 46.95%
```


ARIMA Forecast for Revenue

```
Mean Absolute Percentage Error (MAPE) for FUTURE REVENUE: 47.64%
```

III. <u>Customer Lifetime Value Analysis (CLV):</u>

- Calculating Customer Lifetime Value (CLV) involves estimating the total value a customer is expected to bring to your business over their entire relationship with your company.

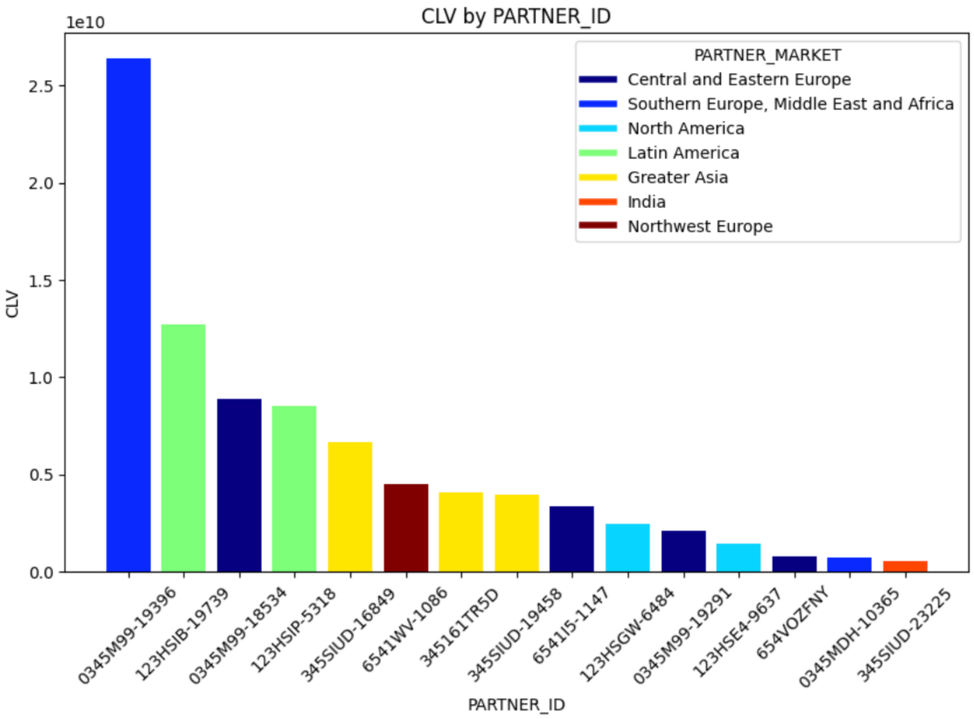
- To calculate CLV, you'll need to compute a few key metrics:

  - Average Purchase Value (APV): This is the average amount a customer spends on each purchase. You can calculate this by dividing the total revenue by the total number of purchases for each customer.
    APV = Total Revenue / Total Number of Purchases
  - Average Purchase Frequency (APF): This represents how often, on average, a customer makes a purchase. It's calculated by dividing the total number of purchases by the number of unique customers.
    APF = Total Number of Purchases / Number of Unique Customers
  - Average Customer Lifespan (ACL): This is the average number of months or years a customer stays with your company. It's calculated as the average time between a customer's first purchase and their last purchase.
    ACL = (Sum of Customer Lifespans) / Number of Unique Customers

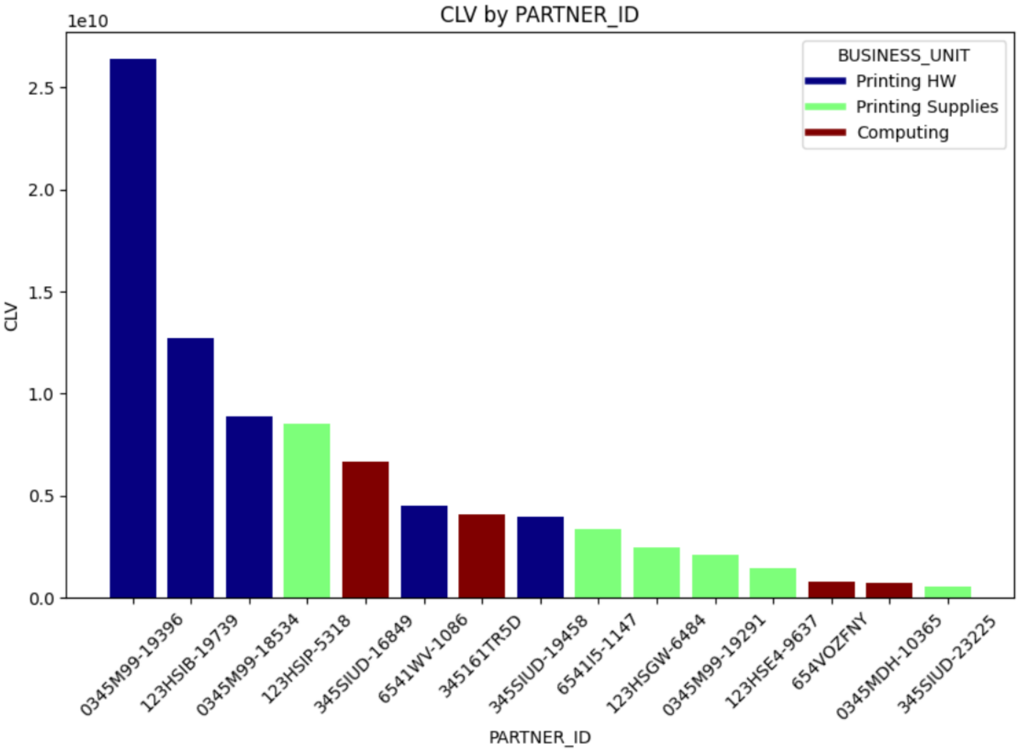
- Customer Lifetime Value (CLV): Finally, you can calculate CLV using the following formula: *CLV = APV x APF x ACL*

- Identify high CLV segments: Determine which customer segments have the highest CLV. These are your most valuable customers.
  - Tailor marketing strategies: Use the CLV insights to tailor marketing strategies for each customer. High CLV customers may receive premium offers or loyalty program invitations, while low CLV customer may receive targeted promotions to increase their value.
  - Budget allocation: Allocate your marketing budget more effectively by focusing resources on high CLV customers.
  - Measure effectiveness: Over time, measure how changes in your marketing and customer retention strategies impact CLV.
  - Monitor over time: CLV is not static. Regularly update your CLV calculations to track changes in customer behavior and value.

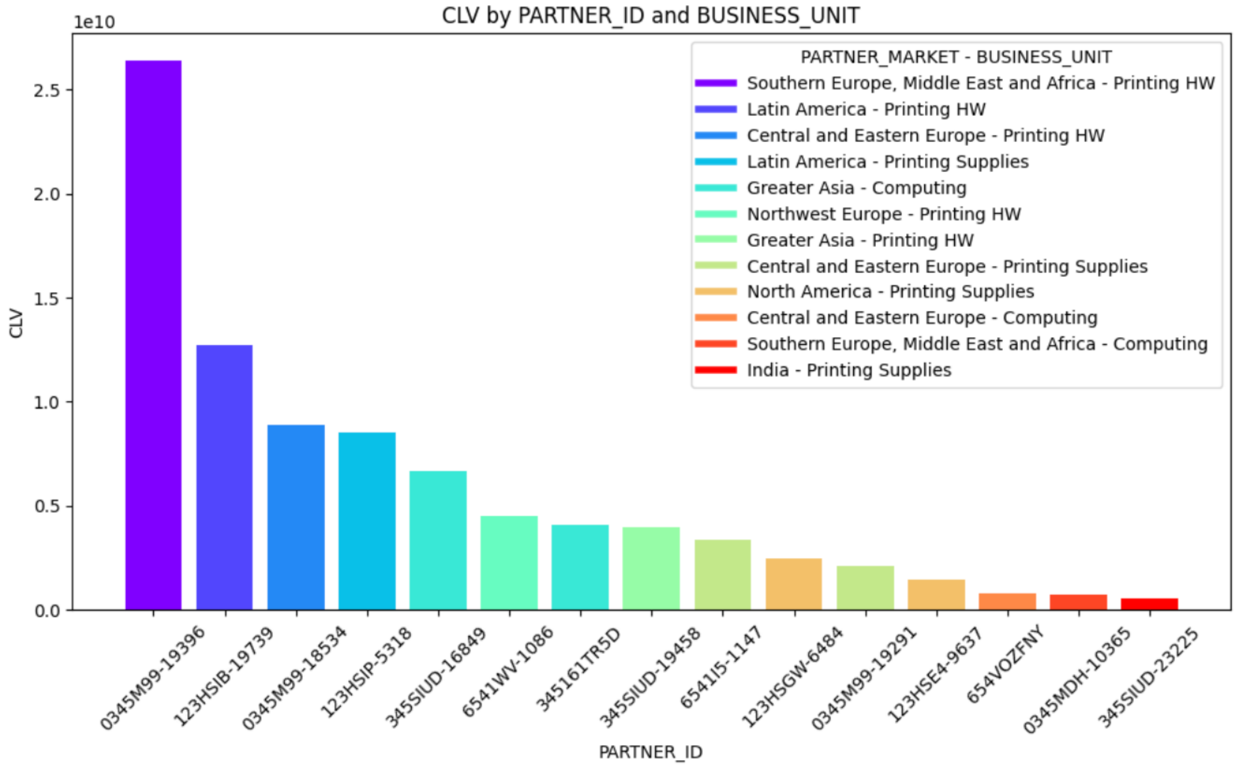
- CLV is a powerful metric for making data-driven decisions to maximize customer profitability and retention. It's an ongoing process that requires monitoring and adjustment as your business evolves.



- This bar graph depicts the CLV values for all the 15 major individual partner customers who have bought products over a 5-year time period. This graph segments, which geographical location each partner customer belongs to. Here, we can observe with no surprises that a partner from SEMEA region is the most valuable customer. We can also observe that two major partners hail from Latin America.



- This CLV graph segments, which product type each partner buys. Here, we can observe that the top three clients, all buy printing hardware, surprisingly as based on pareto analysis, printing supplies and computing products generated more revenue.



- This CLV graph segments, which product type-geographical area combination does each of the customer belongs to. As we can observe, SEMEA-Printing Hardware combination is the highest. We can also observe that throughout Europe, there are many partners requiring printing hardware.

4 Interpretation of Results and Recommendation:

I. <u>Customer and product segmentation:</u>

1) On visualizing how sum of Revenue of the Products and Number of Units sold are distributed among the product types and the related pareto analysis chart, we can make an inference that printing supplies revenue generated per item, is very less than computing or printing hardware. This inference can be drawn by analyzing the corresponding pareto charts, where the cumulative revenue line is way steeper than the cumulative number of units sold for types of products. Since producing more number of units might require more cost and effort to produce, pack, store and deliver to the customers, a good recommendation here would be to concentrate more company resources on producing printing hardware and computing products (assuming cost/effort to make all product units are similar), because even though printing supplies generate more revenue, the other

product types generate more revenue for lesser number of units sold, thus potentially saving up on costs and efforts.

2) On visualizing how sum of Revenue of the Products and Number of Units sold are distributed among the geographical areas and the related pareto charts, a very simple inference can be drawn, which is that customers situated in SEMEA and central & eastern European countries are the "vital few" (Top 20%) of all the geographical regions. Hence these two regions are buying more units and generate more revenue. We can notice that these two regions are of close proximity. Hence, a simple recommendation would be to allocate more company resources to an area common to these two regions, like setting up a large-scale production factory/warehouse in say a country like Turkey, which is kind of a center point when the two regions are combined.

II. Demand and Revenue Forecasting:

3) These Time forecasting outputs, help forecast the upcoming demand (number of units going to be sold) and revenue for the future. While the accuracy of these predictions needs some working, accurate demand forecasting helps a company in numerous ways: -
    a. Help the company plan its inventory and production processes more efficiently. By knowing what products are likely to be in demand, the company can reduce overproduction and minimize excess inventory costs while ensuring that popular items are readily available.
    b. When a company can accurately forecast demand, it can reduce operational costs by avoiding rush orders, excessive warehousing, and transportation costs associated with underestimating or overestimating demand.
    c. Accurate demand forecasting ensures that products are available when customers want them, reducing instances of stockouts and backorders, which can frustrate customers.
    d. Accurate demand and ad revenue forecasting enable better resource allocation. Companies can allocate their workforce, raw materials, and capital more effectively, ensuring that they are used where they are needed most.

III. Customer Lifetime Value Analysis (CLV):

4) By looking at the CLV bar chart which classifies each partner customer by their product type-geographical location combination, we can observe that all the top valued partners prefer to buy printing hardware. Thus, by extending the 2nd point, which suggested to allocate more resources on SEMEA and Central & Eastern European region, we can further increase profits by also concentrating on producing/distributing printing hardware supplies in this region. That is, the previously discussed large-scale production factory/warehouse in Turkey could specialize in producing printing hardware. Thus, increasing revenue for the company, as well as keeping the top values customers satisfied

by giving extra attention to their needs. This setup in Turkey would get more products that the top customers want and in faster time.

5 Conclusion:

In conclusion, the above set of interpretations and recommendations are based upon the data which was available. All methods used to produce these results are fairly easy to implement, fast and easy to understand such as Time-Forecasting, Pareto Analysis, Customer Lifecyle Value calculation, etc. These methods' outcomes are easy to study, analyze and interpret, without the use of complex methods such as 'Propensity to Buy' attribute which are usually calculated by the company's business analysts.