Sirak Jraghatspanyan

Final Report
Customer lifetime value (LTV) prediction in E-commerce

Project Proposal

In ecommerce, your customers have all the power. They're armed with reliable product reviews from their peers, the ability to quickly compare prices between online stores, and the opportunity to vote with their wallets. In order to give them a reason to buy from you, you need to offer them the best possible online shopping experience. No matter how you slice it, there's only one thing that will help you do it right: data.

That means gathering data from all areas that have an impact on your online store and using this information to understand the trends and the shift in consumers' behavior to make data-driven decisions that will drive more online sales.

For this capstone project we've used e-commerce sales data in order to uncover the profitability of e-commerce sales in today's marketplace. The dataset used for the project provides a comprehensive overview of e-commerce sales data from different channels covering a variety of products. It contains data on a variety of sales channels such as Amazon and international sales data.

In the scope of this project I've built a predictive model using Machine Learning to predict customer lifetime value (LTV). We invest in customers (acquisition costs, offline ads, promotions, discounts & etc.) to generate revenue and be profitable. Naturally, these actions make some customers super valuable in terms of lifetime value but there are always some customers who pull down the profitability. Therefore, we need to identify these behavior patterns, segment customers and act accordingly.

Data Wrangling

For the project I used two datasets: Amazon sales and International sales.

Amazon sales

I started the data wrangling process with the Amazon sales data, which had 24 columns and 129k rows. I started the data cleaning process by creating a dataframe presenting the number of unique and null values for every column in the data. I changed the 'Date' column's data type to datetime, then dropped the 'ship_country' and 'currency' columns because we had 1 unique value for each of them(India for country and Indian rupee for currency). I also dropped the 'Unnamed: 22' column which didn't contain any valuable information. There were 2 columns presenting the orders fulfillment

information: 'fulfilled-by' and 'Fulfilment'. The latter helped us to impute the missing values of the former: all of the orders were fulfilled by 2 companies: Amazon and Easy ship. I dropped the 'fulfilled-by' column and replaced the 'Merchant' values in 'Fulfilment' with 'Easy ship'. 50k null values in the ‘promotion-id’ column were replaced by ‘no-promotion’ values. There were also 33 missing values for 'ship-postal-code', 'ship-city' and 'ship-state' columns which were dropped after ensuring that the rows with missing values are the same. Next we had about 7k missing values in the 'Courier Status' column which were replaced by ‘Canceled’ values after being compared with the values in the ‘Status’ column which provided the same information but had 13 unique values. And the last column with the missing values was the ‘Amount’ column which I decided to impute using the values with the same SKU code. But it appeared that a lot of SKU codes had more than one unique amount value, not only because in some cases the amount was calculated for more than one item but also they were sold for different prices. And in some cases there were 0 values, which needed to be dealt with the same way as the missing values. So I decided to replace null values with 0s and save all those rows in a separate dataset. Then I added a new column ‘Price’ by dividing the amount by quantity. A new dataset was created containing information about the unique SKU codes and the average prices for them which later was used for imputing the 0 values in the ‘Price’ column.

After dealing with the missing values in all the columns of Amazon sales data I changed the data type of 'ship-postal-code' from ‘float’ to categorical and dropped the 6 duplicate rows.

International sales

Next we had the International sales data of the same company. The dataset had 9 columns with 37k rows. About 1000 rows were missing all the values in 6 columns which were dropped. Then I noticed that starting from row 19675 we had different order of columns and instead of 'size' we had 'stock' column.So I divided the data into two datasets and brought them to the same format and order. After comparing the datasets I found out that all of the rows were present in both datasets, some of them more than once. The only difference is the first one had the size information of the order and the second one had the number of items in stock. After joining those sets and dropping the duplicate rows we were left with 12k rows.

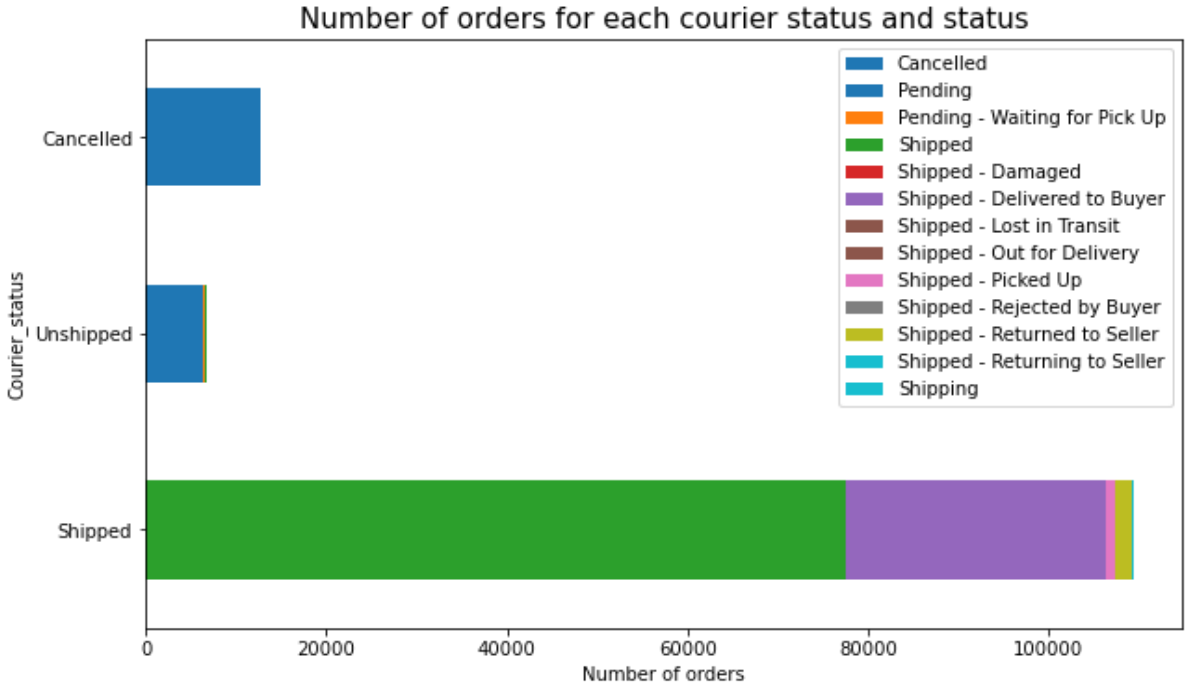
Then we converted the data types for 'DATE', 'PCS', 'RATE' and 'GROSS AMT' columns and took a look at the 'stock' and 'Size' columns, which had a lot of null and 0 values after joining the two datasets. The stock information was missing for all the rows because the rows with that information were dropped as duplicate values. But I saved the information for every SKU and customer combination in a separate dataset before dropping the rows. This dataset was used for imputing the missing values in the ‘stock’ column. While trying to impute the missing values in the ‘Size’ column I found out that

more than 200 rows were just shipping cost information for different orders which were saved in a different dataset and later joined with the main dataset using the order date and the customer name. And lastly some of the columns were renamed.

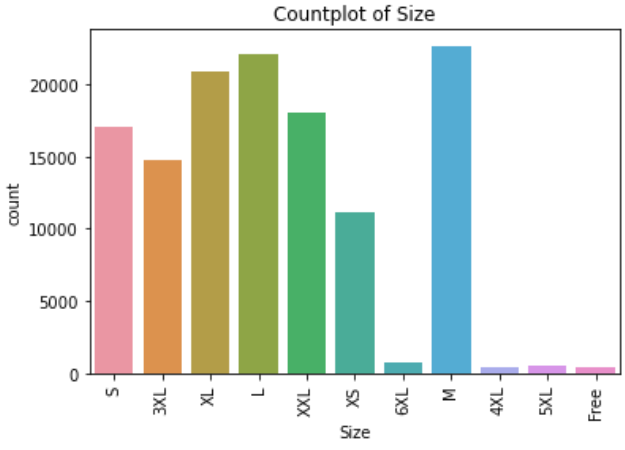
Exploratory Data Analysis

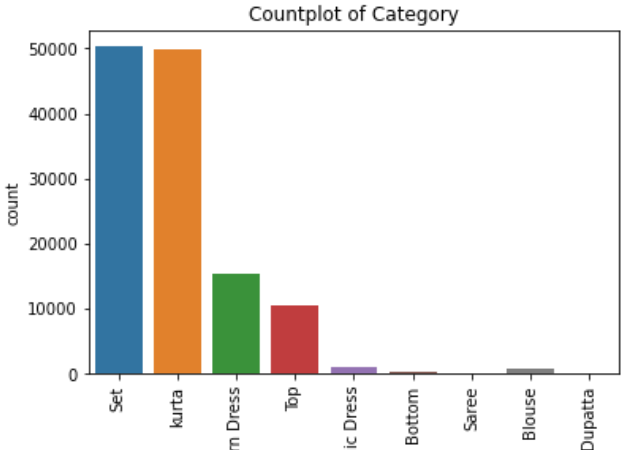
Amazon sales

I started the EDA again with the Amazon sales data. First I captured the time period when the orders were placed. The data represented the company's sales from the end of March till the end of June of 2022. Then I compared the ‘Status’ column with the ‘Courier Status’ column by creating a stacked bar plot showing the number of orders for each courier status and status.



Most of the orders with the courier status 'Shipped' were either on the way or already delivered, there were also about 2000 returned orders. But We also had a lot of cancelled orders for both 'Cancelled' and 'Unshipped' courier statuses: about 14% of all orders from our data were cancelled.

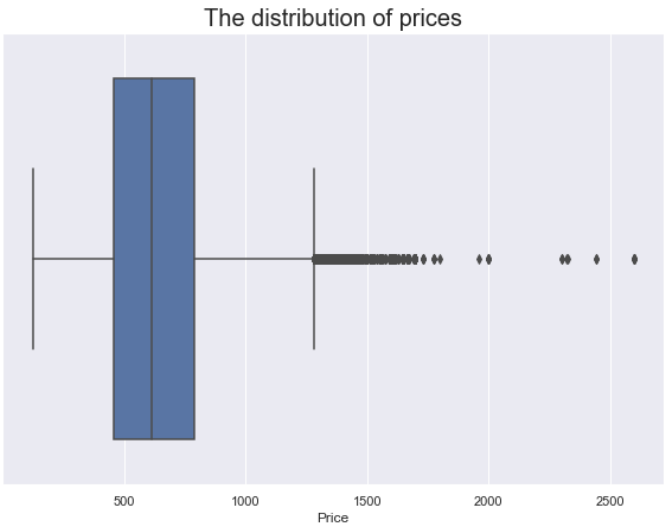




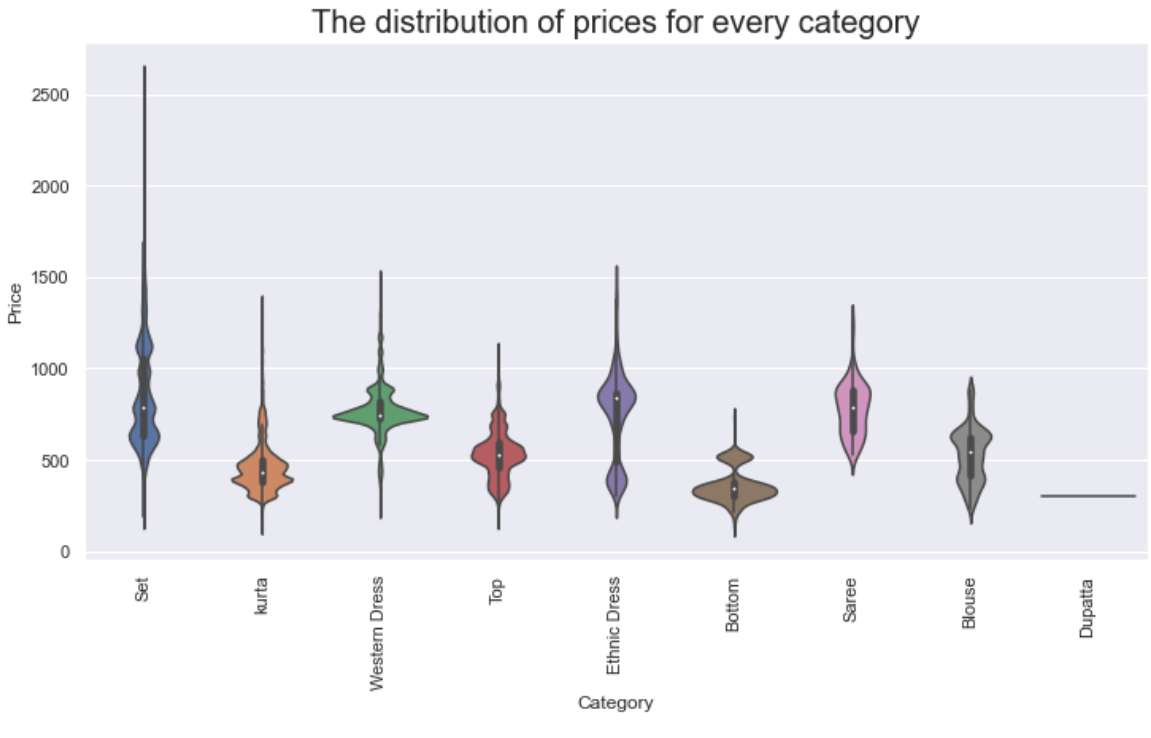
After creating bar plots for all the categorical columns I noticed the following insights:

- Most of the orders are fulfilled by Amazon but Easy Ship also has a significant amount of orders.
- Very little percentage of orders are sold via non Amazon sales channels.
- Majority of the orders are delivered by expedited shipping.
- Sets and Kurtas are the most sold item categories: they are being sold 3-5 times more than items in other categories.
- Although products with M and L sizes are the most sold ones, there is also a huge demand for extra large sized items.
- And the quantity is limited to 1 item in the vast majority of cases.
- Most of the orders don't have any discount, but at the same time a lot of orders have free shipping and PLCC discount from Amazon.

Then I created a boxplot showing the distribution of 'Price' column:



The median price was around 600 rupees and most of the values were between 400-750 rupees. We also had a lot of outliers which were explained by the different product categories in the next violinplot. As we can see sets can be more expensive than all the other categories and have a wider range of values.

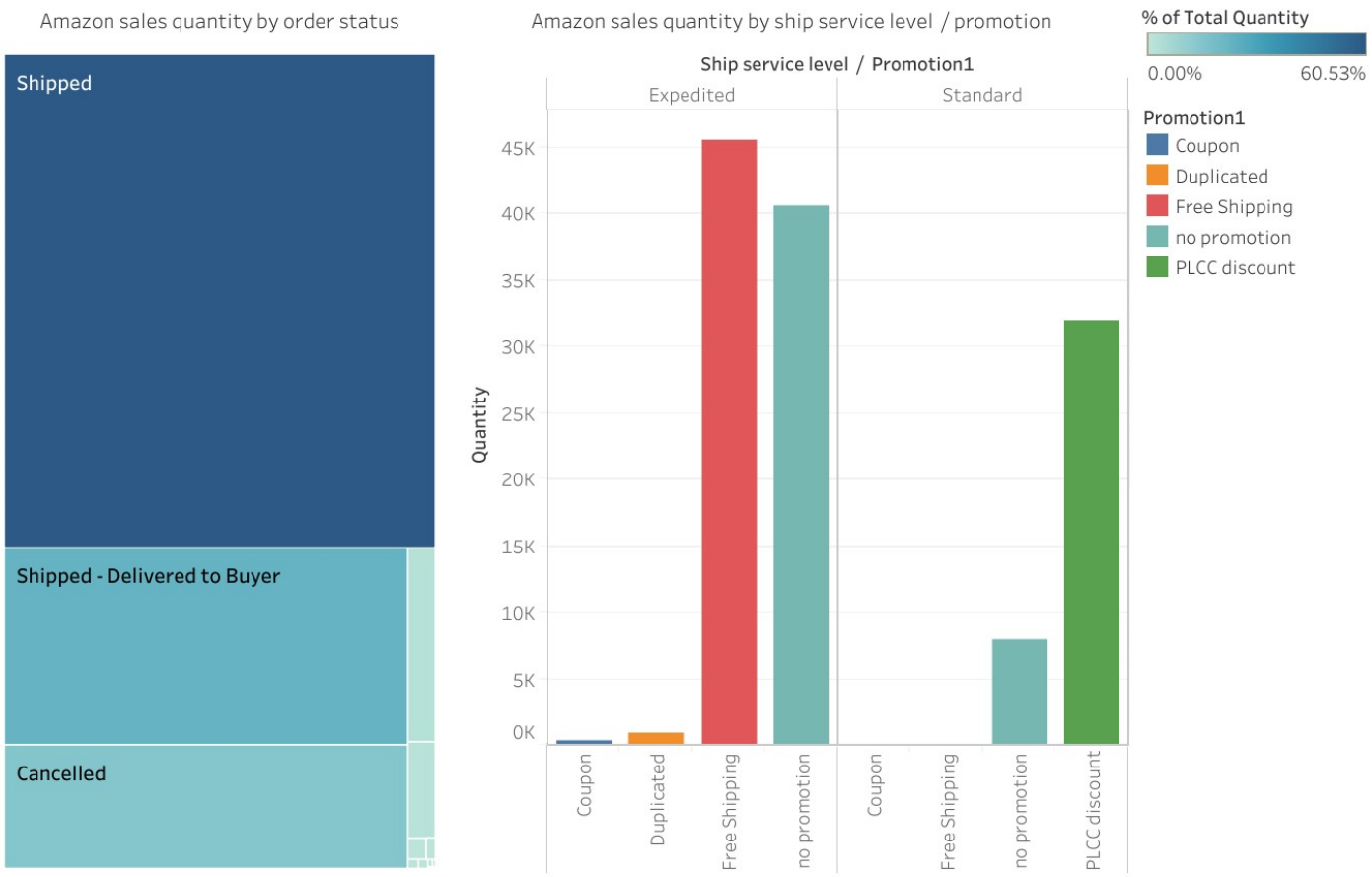


Next we have some exploratory data analysis in Tableau.
(Amazon sales quantity)

Amazon sales quantity

If we take a look at our order quantity for different shipping service levels and promotions we'll notice that most of the customers that choose expedited shipping are using free shipping and in canceled orders, the free shipping is used rarely. So, maybe making free shipping available for more products and customers will reduce the number of canceled orders and increase sales overall.

Amazon sales quantity by order status and service level / promotion

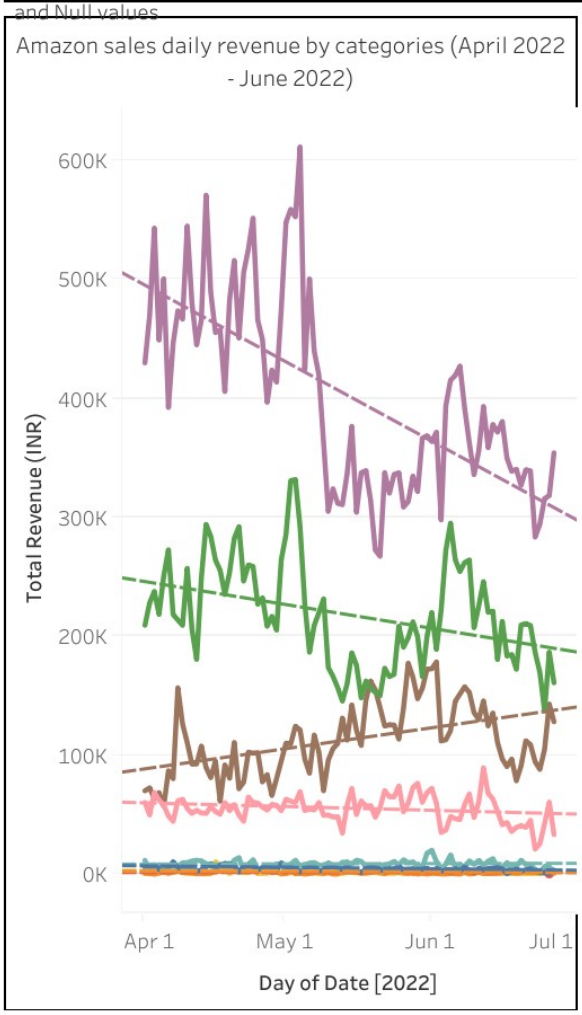


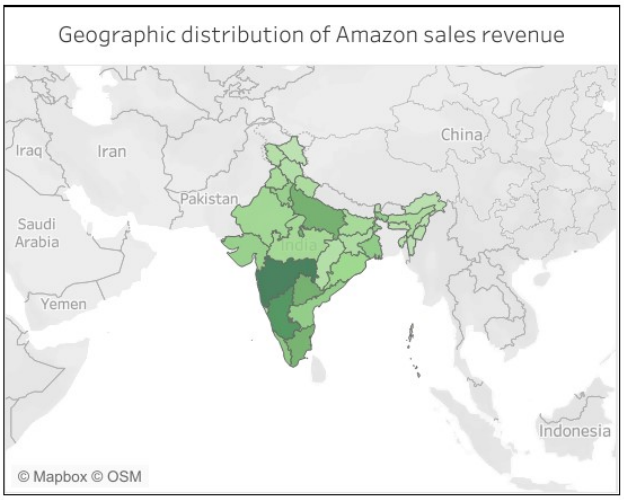
[Amazon sales revenue]

Amazon sales revenue

1) The company's revenue decreases from April 2022 to June 2022, but it might be because of the seasonality. Unfortunately, our data for a limited period is insufficient to check it.
2) Our bestseller categories show a decline in sales (Set & Kurta): the decline is sharper in the Set category. On the other hand, there is a revenue incline for Western dresses which might potentially be a new market segment company can target.
3) The central and southern states of India show higher demand for the company's products.

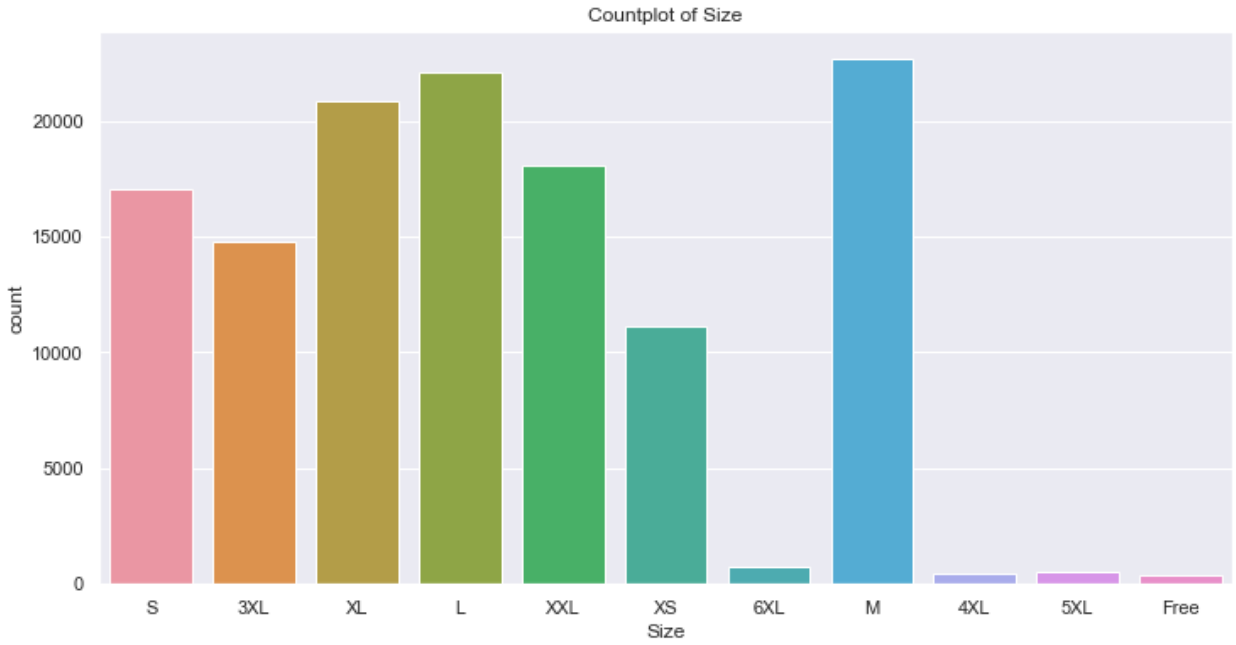
April 1, 2022 to June 27, 2022
and Null values





International sales

First of all, I checked the time period of orders : the orders were placed between 2021-06-05 and 2022-05-11. So we had orders data for 11 months. Next I looked at the distribution of item sizes and prices and compared them with the Amazon sales data.



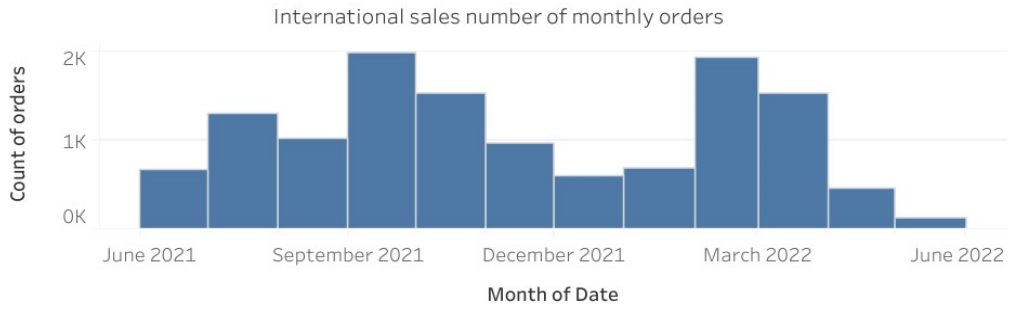
The distribution of sizes is very similar to the one in Amazon sales: products with M and L sizes are the most sold ones, but there is also a huge demand for extra large sized items. And for the price, again, the distribution is very similar to the one in previous data: the boxplot shows that the median price is around 600 rupees and most of the values are between 450-850 rupees. I also created a boxplot for ‘stock’ and ‘shipping cost’ columns. Most of the products have less than 50 items in stock but there are a lot of outliers with larger

quantities and most of the orders have shipping costs between 2000-6000 and the median value is around 4000 rupees. I also added an order ID column based on the Customer name and date in order to investigate the dataset further in Tableau.

[International sales number of monthly active customers, orders and average order amounts]

International sales number of monthly active customers, orders and average order amounts.

There are significant increases in number of active customers and orders throughout the periods September 2021-October 2021 and February 2022 - March 2022. The average order amount is relatively stable throughout the whole time period. The only exceptions are June and July 2021 with the highest and the lowest values respectfully.



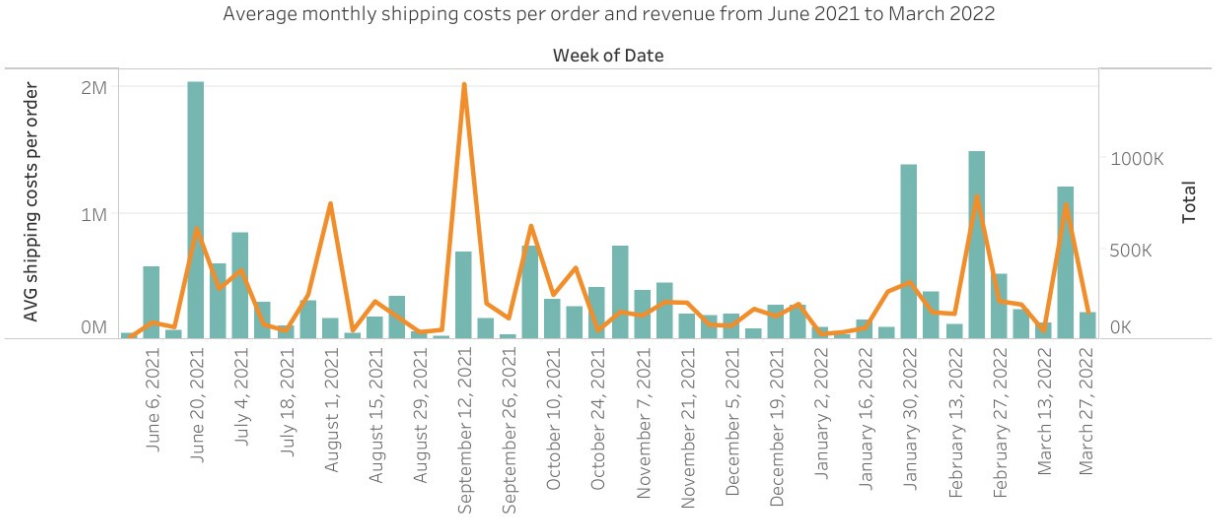
[The average number of items in stock vs sold items for each size (International sales)]

[International Sales total revenue and shipping costs]

International Sales total revenue and shipping costs

Like the Amazon sales data, there is a total revenue decline in International sales but at the same time, we can notice the presence of seasonality: the revenue increases in the spring months and goes down in the summer. This is something that we saw in Amazon sales as well.

In the comparison of shipping costs and revenue, we can notice the positive correlation between them, which was observed in exploratory data analysis, most of the time. But there are some cases when shipping costs are higher whereas the revenue is o..



Preprocessing and Modeling

Our goal in this project was to predict the customer lifetime value (LTV). Since we didn't have customer information in our Amazon sales data, I used the International sales data. But before predicting the LTW I needed to segment the customers to adopt our actions depending on their different needs and behavior. In order to do that I used the RFM method which stands for Recency - Frequency - Monetary Value. So for each of these metrics I calculated their value and gave a metrics score. For the scores I used 3 clusters.The number was calculated using the elbow method. After that all the scores were summarized in a new column called 'Overall_score'. And then we divided the customers into 3 groups based on the overall score: low value, mid value and high value.

After plotting the values in 3 different groups I noticed how the segments are clearly differentiated from each other in terms of RFM.

- For the low value and mid value segments the revenue and frequency levels were quite similar to each other and needed to be improved. The biggest difference between these segments was in recency: mid value customers have much lower recency rates.
- Some of the high value customers had high recency values which should be improved. In terms of frequency and revenue the high value segment had clearly better results and it still could be improved.

In order to calculate the LTV first we needed to select a time window. Considering the fact that we had data for 11 months I decided to predict the LTV for the last 5 months using the first 6 months of data. I used RFM scores as features for the model. Then I splitted the data and calculated the RFM metrics for the first 6 months. After that two datasets were merged and the revenue for the last 5 months was chosen as the target column. And finally the dataset was split into training and test sets.

For the model I used the XGBoost Regressor algorithm. For the first model the root mean squared error (RMSE) on the training set was approximately 0.003, and on the test set, it was around 11771.1.The large difference in the RMSE for the training set and the test set suggested that the model might be overfitting to the training data. In order to deal with this I tried to tune the hyperparameters using RandomizedSearchCV.

We've also tried Linear Regression which showed better results on the test set than XGBoost Regressor with tuned hyperparamters. As a final model I chose Linear Regression even though it had worse results on the training set.

This improved the rmse on the test set (8516.3) but the results were not practical considering the fact that the rmse on the test set was close to the mean value of the target column. This could be due to several reasons:

- Limited Data: With more data, the model might be able to generalize better.

- Outliers: If there are outliers in the target variable (i.e., some customers have extremely high future monetary values), this could be influencing the model's performance.

Takeaways

So the Linear Regression showed better results than the XGBoost Regressor with tuned hyperparameters. Overall, while the model's performance is not perfect, it could still provide valuable insights for making business decisions. For example, it could help to identify customers who are predicted to have high future monetary values, so that marketing efforts can be focused on retaining these customers.