PREPARED FOR
Professor – Eyyub Kibis

PREPARED BY
GROUP –03

FINAL PROJECT

AMAZON – INDIA

2023

Amazon's Internal market share (Amazon vs Merchant) in India.

PREPARED BY,

Harsh Saxena

Mariana Tapieros Arias

Olid Sarker

Chirag Mahayan

Abstract

Developed a dashboard of visualization that identified and showed the comparison of the internal market apparel share between Amazon fulfillment (FBA) and its merchants. We observed and analyzed that FBA (Fulfillment by Amazon), has almost 70% market share based on revenues. On the other hand, all merchants are contributing 30% of revenues or market share. Our team justified the market share based on Price, place, product, and promotion. Most of the time FBA holds the maximum number or percentage on the above-mentioned criteria, although Merchants were dominated in particular products and states.

Overall, the performance of the FBA was good for almost all marketing criteria with many fluctuations. It is clear that merchants are struggling with their performance because of using external shipping services and delivery methods. We noticed that 22% of orders were canceled by merchants whereas it was just 12% for FBA.

Introduction

Dataset:

Rows: 128975
Columns: 24

Target: Amazon Sale Report.csv (Category)

Categorical Variables: OrderID, Date, Status, Fulfillment, Sales Channel, ship-service-level, SKU, Category, currency, ship-city, ship-state, ship-country, promotion-ids, B2B, fulfilled-by

Numerical/Continuous Variables: index, Amount, ship-postal-code

- Data Collection
- Data Cleaning
- Data Preprocessing
- Data Visualization
- Analysis & Results

Our team worked on the project to find out the percentage of internal apparel market share and performance of the operations between FBA and Merchants. According to Kaggle, the E-Commerce Sales dataset released by THE DEVASTATOR Includes variables rderID, Date, Status, Fulfillment, Sales Channel, ship-service-level, SKU, Category, currency, ship-city, ship-state, ship-country, promotion-ids, B2B, fulfilled-by.

The primary objective of this project is to develop future marketing positioning strategies by identifying and comparing the market share of individuals. So, we visualized all measurement criteria as per the instruction of our superior so that he/she can make a final decision.

Data used:

https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data

Variables Description:

- **OrderID**: It is a unique identifier for each order of the category of dresses.
- **Date**: It is a type of variable that represents an unknown date value. A date value is any value that we can express using any of the combinations of the date, month, year in any sequence of our choice.
- **Status**: It tells the status to the customer whether the item is Shipped, or Canceled.
- **Fulfillment**: Tells if the item will be fulfilled by another merchant OR called as a third-party company or Amazon.
- **Sales Channel**: Showing the sales of the category/item where it is sold that is: Amazon.in
- **ship-service-level**: It shows whether the category/item of service is Standard or Expedited.
- **SKU**: Shows SKU number of each of the unique Category and it is in String datatype.
- **Category**: Shows different types of category like Set, kurta, Western-Dress, Top, Blouse, and bottom.
- **currency**: The currency used here is Indian Rupees (INR) or ₹
- **ship-city**: Shows the city name for India,.
- **ship-state**: Shows the state name for India
- **ship-country**: Shows where the category/item is being shipped to and to which country.
- **promotion-ids**: It is a unique id for different types of promotion for each category or an item.
- **B2B**: Shows if the category/item is Business-to-Business either False (0) or True (1)
- fulfilled-by: It shows which company has fulfilled the product or Category, such as East Ship.
- **index**: It is an integer
- **Amount**: Shows the amount in Indian Rupee (INR) in decimals, and integers.
- **ship-postal-code**: It shows which type of category is going to be shipped by postal code/zip code.
- **Category**: Type of products like Set, Kurta, Blouse, Top, Bottom, and Western-Dress.

Methodology: The present study uses Data Wrangling and Data Warehousing analysis in order to clean and visualize the data. The objective of this project was to compare FBA vs Merchant by 4'p of marketing to identify the power they have in the market for future positioning strategies .

Data Collection

We collected our dataset to conduct a research about features, decision-making and strategy of Dresses the category like Western Dress, Kurta, Set, Ethnic Dress, Top, Blouse, Bottom with sizes of S, L, M, XL, XXL, 3XL, 4XL, 6XL.

The dataset was collected from Kaggle website (Amazon Sale Report.csv), and it contains two types of data: quantitative and qualitative.

With the use of Python and libraries such as pandas, and numpy, for data visualization we used Seaborn, Matplotlib, statistics, Sklearn and Tableau.

```
[1]  import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
     import statistics as stats
     import matplotlib.pyplot as plt
     import statistics as stats
     from sklearn import preprocessing
     from sklearn.linear_model import LinearRegression
     import statsmodels.api as sm
     from sklearn.linear_model import LogisticRegression
     import scipy as sp
     from sklearn.preprocessing import StandardScaler
     from sklearn.metrics import confusion_matrix
     from sklearn.metrics import roc_curve,auc


[2]  df = pd.read_csv('Amazon Sale Report.csv')
```

Exploring the Dataset:

As we mentioned previously, the dataset contains 128975 rows and 24 Columns.
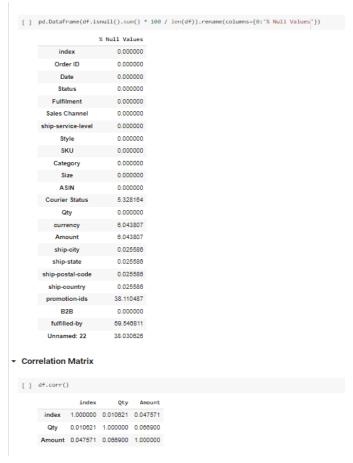
The data type we will be working on are numerical and categorical.

Data Preprocessing

Our data reprocessing was s step in the data mining and data analysis process that took raw data and transformed into a format that is simple to understand and analyze. We proceeded to clean our data, verifying the null vales:

Percentage of null value with matrix:

```
[ ] pd.DataFrame(df.isnull().sum() * 100 / len(df)).rename(columns={0:'% Null Values'})
```

	% Null Values
index	0.000000
Order ID	0.000000
Date	0.000000
Status	0.000000
Fulfilment	0.000000
Sales Channel	0.000000
ship-service-level	0.000000
Style	0.000000
SKU	0.000000
Category	0.000000
Size	0.000000
ASIN	0.000000
Courier Status	5.328164
Qty	0.000000
currency	6.043807
Amount	6.043807
ship-city	0.025586
ship-state	0.025586
ship-postal-code	0.025586
ship-country	0.025586
promotion-ids	38.110487
B2B	0.000000
fulfilled-by	69.546811
Unnamed: 22	38.030626

▼ **Correlation Matrix**

```
[ ] df.corr()
```

	index	Qty	Amount
index	1.000000	0.010621	0.047571
Qty	0.010621	1.000000	0.066900
Amount	0.047571	0.066900	1.000000

Null and unique value:

	data_type	null_count	unique_count
index	int64	0	128975
Order ID	object	0	120378
Date	object	0	91
Status	object	0	13
Fulfilment	object	0	2
Sales Channel	object	0	2
ship-service-level	object	0	2
Style	object	0	1377
SKU	object	0	7195
Category	object	0	9
Size	object	0	11
ASIN	object	0	7190
Courier Status	object	6872	3
Qty	int64	0	10
currency	object	7795	1
Amount	float64	7795	1410
ship-city	object	33	8955
ship-state	object	33	69
ship-postal-code	object	33	9459
ship-country	object	33	1
promotion-ids	object	49153	5787
B2B	object	0	2
fulfilled-by	object	89698	1
Unnamed: 22	object	49050	1

Drop Numerical Variables with Zero Variance:

We proceed to analyze numerical variables with zero variance however we did observe any remarkable numerical variable with zero variance.

```
[ ] df[Numerical_Variables].std()

    index     37232.019822
    Qty           0.313354
    Amount      281.211687
    dtype: float64

[ ] df = df.drop(df[Numerical_Variables].std()[df[Numerical_Variables].std()== 0].index, axis = 1)

[ ] df[Numerical_Variables].std()

    index     37232.019822
    Qty           0.313354
    Amount      281.211687
    dtype: float64

[ ] df = df.drop('index', axis = 1)

[ ] df.head()
```

	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Style	SKU	Category	Size	...	currency	Amount	ship-city	ship-state	ship-postal-code	ship-country	promotion-ids	B2B	fulfilled-by	Unnamed: 22
0	405-8078784-5731545	04-30-22	Cancelled	Merchant	Amazon.in	Standard	SET389	SET389-KR-NP-S	Set	S	...	INR	647.620000	MUMBAI	MAHARASHTRA	400081.000000	IN		NaN	False	Easy Ship	NaN
1	171-9198151-1101146	04-30-22	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	JNE3781	JNE3781-KR-XXXL	kurta	3XL	...	INR	406.000000	BENGALURU	KARNATAKA	560085.000000	IN	Amazon PLCC Free-Financing Universal Merchant ...	False	Easy Ship	NaN
2	404-0687676-7273146	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3371	JNE3371-KR-XL	kurta	XL	...	INR	329.000000	NAVI MUMBAI	MAHARASHTRA	410210.000000	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	True	NaN	NaN
3	403-9615377-8133951	04-30-22	Cancelled	Merchant	Amazon.in	Standard	J0341	J0341-DR-L	Western Dress	L	...	INR	753.330000	PUDUCHERRY	PUDUCHERRY	605008.000000	IN		NaN	False	Easy Ship	NaN
4	407-1069790-7240320	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3671	JNE3671-TU-XXXL	Top	3XL	...	INR	574.000000	CHENNAI	TAMIL NADU	600073.000000	IN		NaN	False	NaN	NaN

5 rows × 23 columns

Drop Categorical Variables with Zero Variance:

In the same way we also did not observe any remarkable categorical variable with zero variance.

```
[ ] Categorical_Variables = list(df.select_dtypes(object).columns)
    Categorical_Variables

    ['Order ID',
     'Date',
     'Status',
     'Fulfilment',
     'Sales Channel ',
     'ship-service-level',
     'Style',
     'SKU',
     'Category',
     'Size',
     'ASIN',
     'Courier Status',
     'currency',
     'ship-city',
     'ship-state',
     'ship-postal-code',
     'ship-country',
     'promotion-ids',
     'B2B',
     'fulfilled-by',
     'Unnamed: 22']
```

```
[ ] len(Categorical_Variables)

    21
```

```
[ ] zero_cardinality = []

    for i in Categorical_Variables:
        if len(df[i].value_counts().index) == 1:
            zero_cardinality.append(i)
    zero_cardinality

    ['currency', 'ship-country', 'fulfilled-by', 'Unnamed: 22']
```

Dropping Categorical Variables with Many Levels:

```
[ ] Categorical_Variables = list(df.select_dtypes(object).columns)
    Categorical_Variables

    ['Order ID',
     'Date',
     'Status',
     'Fulfilment',
     'Sales Channel ',
     'ship-service-level',
     'Style',
     'SKU',
     'Category',
     'Size',
     'ASIN',
     'Courier Status',
     'ship-city',
     'ship-state',
     'ship-postal-code',
     'ship-country',
     'promotion-ids',
     'B2B']

[ ] len(Categorical_Variables)

    18

[ ] high_cardinality = []

    for i in Categorical_Variables:
        if len(df[i].value_counts().index) > 200:
            high_cardinality.append(i)

    print(high_cardinality)

    ['Order ID', 'Style', 'SKU', 'ASIN', 'ship-city', 'ship-postal-code', 'promotion-ids']
```

We exclued 'Order ID', 'ship-city', 'ship-postal-code', 'promotion-ids' variable, as we will apply these for visualization on tableau.

Data Imputation:

For numerical variable:

```
[ ] df.isnull().sum()

    Order ID                0
    Date                    0
    Status                  0
    Fulfilment              0
    Sales Channel           0
    ship-service-level      0
    Category                0
    Size                    0
    Courier Status       6872
    Qty                     0
    Amount               7795
    ship-city              33
    ship-state             33
    ship-postal-code       33
    ship-country           33
    promotion-ids       49153
    B2B                     0
    dtype: int64
```

```
[ ] Numerical_Variables = list(df.select_dtypes(exclude = object).columns)
    Numerical_Variables

    ['Qty', 'Amount']
```

```
[ ] df['Amount'].median()

    605.0
```

```
[ ] df['Amount'] = df['Amount'].fillna(df['Amount'].median(), inplace = False)
```

For categorical variable:

```
[ ] df.isnull().sum()

    Order ID            0
    Date                0
    Status              0
    Fulfilment          0
    Sales Channel       0
    ship-service-level  0
    Category            0
    Size                0
    Courier Status   6872
    Qty                 0
    Amount              0
    ship-city          33
    ship-state         33
    ship-postal-code   33
    ship-country       33
    promotion-ids   49153
    B2B                 0
    dtype: int64
```

```
[ ] Categorical_Variables = list(df.select_dtypes(object).columns)
    Categorical_Variables

    ['Order ID',
     'Date',
     'Status',
     'Fulfilment',
     'Sales Channel ',
     'ship-service-level',
     'Category',
     'Size',
     'Courier Status',
     'ship-city',
     'ship-state',
     'ship-postal-code',
     'ship-country',
     'promotion-ids',
     'B2B']
```

```
[ ] df[Categorical_Variables].mode()
```

	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size	Courier Status	ship-city	ship-state	ship-postal-code	ship-country	promotion-ids	B2B
0	171-5057375-2831560	05-03-22	Shipped	Amazon	Amazon.in	Expedited	Set	M	Shipped	BENGALURU	MAHARASHTRA	201301.000000	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	False
1	403-4984515-8861958	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
[ ] df = df.fillna({'Courier Status' : 'Shipped','promotion-ids': 'No Promotion', 'ship-country': 'IN' })
```

As we know our data, we cannot impute the location of the orders with the mode, since a state may not match with the city or zip code, for these reasons we can delete the rows or filter the data to not take into account missing values for map graphs.

For our 'promotion-ids' variable, we can not impute the data with its mode since orders with this missing value mean that they do not apply for promotion ids.

```
[ ]  df.isnull().sum()
```

```
     Order ID               0
     Date                   0
     Status                 0
     Fulfilment             0
     Sales Channel          0
     ship-service-level     0
     Category               0
     Size                   0
     Courier Status         0
     Qty                    0
     Amount                 0
     ship-city             33
     ship-state            33
     ship-postal-code      33
     ship-country           0
     promotion-ids          0
     B2B                    0
     dtype: int64
```

```
[ ]  df = df.dropna()
```

```
[ ]  df.isnull().sum()
```

```
     Order ID               0
     Date                   0
     Status                 0
     Fulfilment             0
     Sales Channel          0
     ship-service-level     0
     Category               0
     Size                   0
     Courier Status         0
     Qty                    0
     Amount                 0
     ship-city              0
     ship-state             0
     ship-postal-code       0
     ship-country           0
     promotion-ids          0
     B2B                    0
     dtype: int64
```
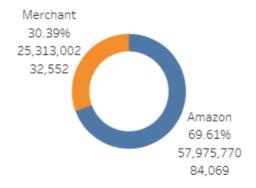
Data Visualization:

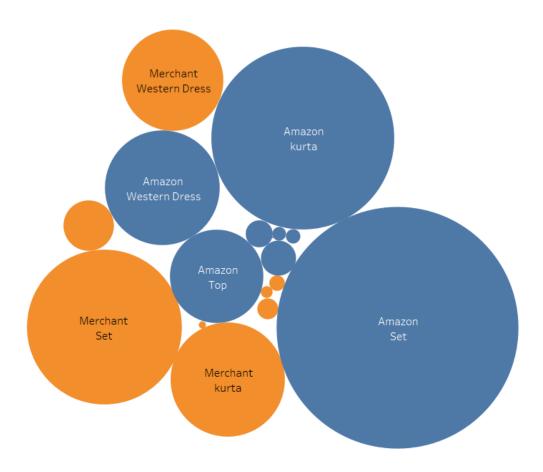**Fix this chart you have to visualize it on a laptop. Sales channel below to the map.**

Market Share
Market Propotion by sale amount and quantity.

Merchant
30.39%
25,313,002
32,552

Amazon
69.61%
57,975,770
84,069

Here we considered the market share based on sales. It is clear from the pie-chat that almost two-third percentage of the market share is captured by BFA.
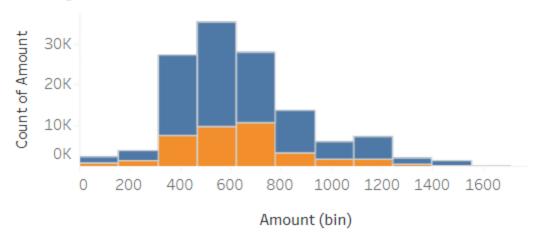
Product popularity:

If we notice the western dress catalog, we will see that Merchants hold a 6% market share whereas FBA is 1% above but which should be 3 times higher than Merchants if we consider the overall market share. But, the popularity was acceptable for other categories based on the overall market share.

Price position:

Price
Sale range of Amazon vs Merchant.
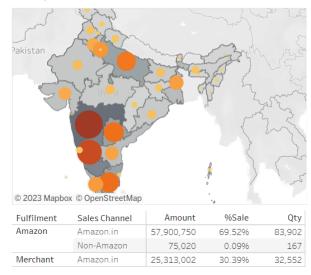
Top states:

Place
Amazon vs. Merchant Presence in the Indian Market. Indentifying the state's purchase amount and sale channel.

© 2023 Mapbox © OpenStreetMap

Fulfilment	Sales Channel	Amount	%Sale	Qty
Amazon	Amazon.in	57,900,750	69.52%	83,902
	Non-Amazon	75,020	0.09%	167
Merchant	Amazon.in	25,313,002	30.39%	32,552

Top State

Ship-State	
KARNATAKA	11,044,974
MAHARASHTRA	14,053,669
TELANGANA	7,335,881

Among all states we ranked top 3 states for both and we also were able to identify the areas where sales were at what amount of market share.

Shipping services:

Ship-Services

Fulfilment	Ship-Ser..	% Sale
Amazon	Expedited	69.44%
	Standard	0.17%
Merchant	Standard	30.39%

Merchants
22.20% Cancelled
1.50% Pending
76.29% Shipped

AMAZON (FBA)
12.52% Cancelled
0.47% Pending
87.00% Shipped

Although market share was lower for the merchants they experienced a higher percentage of order cancellations and pending.

Analysis & Results:

Internal Market Share Report
Amazon vs Merchant Sale of Clothes in India.
4'P of Marketing Comparation.

Market Share
Market Propotion by sale amount and quantity.

Merchant
30.39%
25,313,002
32,552

Amazon
69.61%
57,975,770
84,069

Price
Sale range of Amazon vs Merchant.

Fulfilment
■ Amazon
■ Merchant

Fulfilment
☑ (All)
☑ Amazon
☑ Merchant

Status (group)
☑ (All)
☑ Cancelled
☑ Pending
☑ Shipped

Amount
83,288,772

Amount
521   14M

Qty
0   20,328

Qty
0
5,000
10,000
15,000
20,328

Promotion
Amazon vs. Merchant promotion strategies. Comparing % sale by orders status.

Fulfilment	Promotion-Ids ..	Cancelled	Pending	Shipped
Amazon	Coupon			0.39%
	Free Shipping	1.39%	29.42%	45.46%
	Free-Financing	0.00%	0.19%	0.52%
	No Promotion	54.97%	12.33%	25.95%
Merchant	Free-Financing	10.76%	56.18%	27.68%
	No Promotion	32.87%	1.89%	0.01%

Product Category
Amazon vs Merchant top products by sale amount.

Place
Amazon vs. Merchant Presence in the Indian Market. Indentifying the state's purchase amount and sale channel.
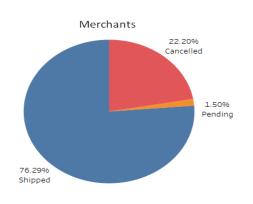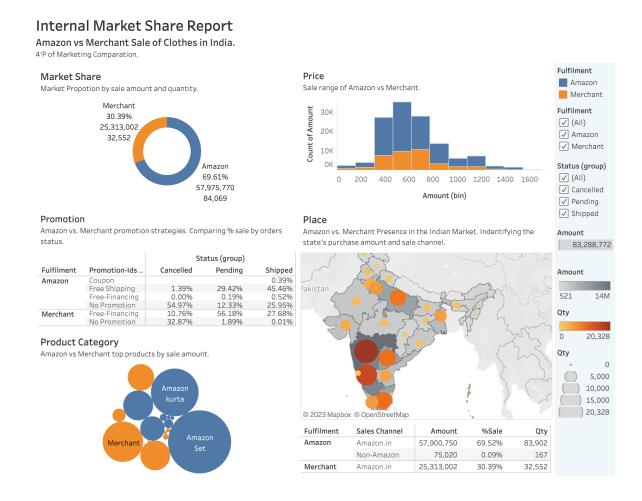

© 2023 Mapbox © OpenStreetMap

Fulfilment	Sales Channel	Amount	%Sale	Qty
Amazon	Amazon.in	57,900,750	69.52%	83,902
	Non-Amazon	75,020	0.09%	167
Merchant	Amazon.in	25,313,002	30.39%	32,552

Our main intention of the project was to find out the areas where we will increase and 4 Marketing propositions (Price, Product, Place, Promotion).

To conduct the whole project properly we cleaned the data with python code and based on the clean data we visualized our requirements.

From the snapshot of the 4Ps, it was clear to our team that the 2/3 market share was captured by the FBA. We noticed Merchants always suffering from shipment, delivery, increasing sales, and ranking products. Merchants shipped their product directly to customers without quality checks and with the help of the 3rd party shipping company.

But Amazon always would use its own shipping service and must keep the products in its own warehouse for quality checks and fast shipping. That is why customers always trusted FBA service.

Conclusion:

To sum up, it is clear to our team we have to focus on Merchants operational performance to increase the market share and acceptance to customers. Now it is the time to develop a 4Ps marketing proposition strategy.

Public Tableau Link

https://public.tableau.com/app/profile/mariana2012/viz/ProjectAmazonv1/4PofMarketingComparationAmazonvsMerchant?publish=yes