REVA UNIVERSITY
Bengaluru, India

A Project Report on

Sales Analytics to drive Profitability - A case study of a Fashion E-Commerce Retailer

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration

In Business Analytics

Submitted By

Tharuka Gallekankanamge

R19MBA82

Under the Guidance of

Dr. JB Simha

CTO, ABIBA Technologies

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

race.reva.edu.in

August, 2022

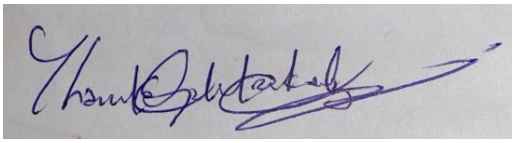
REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Ms. Tharuka Gallekankanamge hereby declare that I have completed the project work towards the second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **"Sales Analytics to drive profitability - A case study of a Fashion E-Commerce Retailer"** under the supervision of Dr. J. B. Simha. This report embodies my original work in partial fulfilment of the requirements for the award of degree for the academic year 2022.

Place: Bengaluru
Date: 27.08.2022

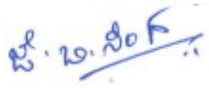
Name of the Student: Tharuka Gallekankanamge
Signature of Student:



REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled "**Sales Analytics to drive profitability - A case study of a Fashion E-Commerce Retailer**" carried out by Tharuka Gallekankanamge with R19MBA82, is a bonafide student at REVA University, is submitting the second-year project report in fulfilment for the award of Master of Business Administration in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.



Signature of the Guide	Signature of the Director
Name of the Guide: Dr. J.B. Simha	Name of the Director: Dr. Shinu Abhi
Guide	Director

External Viva

Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru

Date: Date: 27.08.202

Acknowledgment

I would like to convey a heartfelt thanks to all my mentors at RACE, Dr. J. B. Simha and Mr. Mithun D J for their continuous support throughout the learning journey. A special mention to Dr. J. B. Simha for the valuable feedback and guidance as a Guide and Mentor throughout the project lifecycle.

I would like to express a special thanks to Dr. Shinu Abhi, Director of REVA Academy of Corporate Excellence for her cordial support, valuable guidance, and information at various stages, which helped in completing the project.

I would like to acknowledge the support provided by the founder and Hon'ble Chancellor, Dr. P Shayma Raju, Vice-Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N Ramesh.

I thank my family and my daughter for keeping up with my busy schedule and supporting me in this skill-upgrading journey.

Place: Bengaluru

Date: Date: 27.08.2022

REVA UNIVERSITY
Bengaluru, India

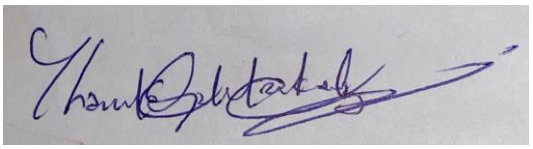
Similarity Index Report

This is to certify that this project report titled "**Sales Analytics to drive profitability - A case study of a Fashion E-Commerce Retailer"** in India was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitn
Date of Report Generation: 25/08/2022
Similarity Index in %: 6%
Total word count: 7427
Name of the Guide: Dr. J.B. Simha

Place: Bengaluru
Date: 27/08/2022

Name of the Student: Tharuka Gallekankanamge
Signature of Student:



Verified by: M N Dincy Dechamma

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	CODB	Cost of Doing Business
2	D2C	Direct to Consumer
3	EDA	Exploratory Data Analysis
4	CRISP-DM	Cross-Industry Standard Process for Data Mining
5	CAGR	Compound annual growth rate
6	ARPU	The average revenue per User
7	ARIMA	Autoregressive Integrated Moving Average
8	SARIMA	Seasonal Autoregressive Integrated Moving Average
9	IT	Information Technology
10	COGS	Cost of Goods Sold
11	RTO	Return to Origin
12	COD	Cost of Goods Sold
13	UPI	Unified Payments Interface
14	BAU	Business As Usual
15	MRP	Maximum Retail Price
16	GST	Goods and Service Tax
17	Ho	Null Hypothesis

18	H1	Alternative Hypothesis
19	ACF	Auto Correlation Function
20	PACF	Partial Auto Correlation Function
21	LSTM	Long Short-Term Memory
22	SKU	Stock Keeping Units

List of Figures

List of Tables

Abstract

The value fashion industry in India is booming and India has become a very promising market for both local and international fashion retailers. Due to the vast diversity in the market, it has become vital that retailers have an Omni presence to cater to the needs of Indian consumers. Today, retailers are thriving both offline and online. Indian retail has shown that being a D2C brand is important. Also, social commerce is booming, and this is due to India having a younger population compared to the rest of the world.

The data set acquired for the project belongs to a global fashion retailer that entered India in 2021. This retailer specializes in value fashion format in women's wear, men's wear, kids,' footwear, and accessories. The sales data gathered are specific to one online portal which started its operations in January 2022. The brand is operating on "a marketplace model" with one of the major online giants in the country. This means the inventory is owned by the brand and the products are only listed on the online platform. This model is the most preferred mode of operating in e-commerce due to its easy scalability and high turnaround time of working capital.

Forecasting techniques such as ARIMA and SARIMAX are utilized for sales prediction. A comprehensive approach is taken where most of the weightage was given to the Exploratory Data Analysis (EDA) to derive meaningful and insightful information. The reason for giving more weightage to EDA is to ease the analytical adaptation to ensure concepts are simple and practical to the targeted audience.

This study will be a stepping-stone in creating an analytical culture in the organization. The key insights derived from this study will be passed on to the e-commerce and operations teams.

Keywords: E-commerce, Value Fashion, Trend Analysis, profitability, Sales Analytics, Retail, Sales Forecasting.

Table of Contents

Chapter 1: Introduction

This study, titled “Sales Analytics to drive profitability - A case study of a Fashion E-commerce Retailer” will be laying the foundation for introducing an analytical culture to the organization. The data acquired belongs to a global value fashion retailer that started its operations in India in the year 2021. Hence the sales data acquired is of a shorter time. The sales data belongs to one e-commerce channel that started its operations in January 2022.

The Indian value fashion market is heavily penetrated by competition. Hence it is vital to keep an eye on the Cost of Doing Business (CODB) and overall profitability. Value fashion brands work on seasonal collections and the SKUs in a season are large to ensure all categories are served. A typical brand follows four seasons which are Spring, Summer, Autumn, and Winter in a year. Especially when the brand is new to the target market, it is important to find its pros and cons, and USPs and create a niche in the industry.

A thorough analysis is conducted to find out what are the key drivers of E-commerce profitability. How are discounts run on the platform affecting the net profit? What are the top categories contributing to a healthy net profit? This study aims to derive meaningful practical insights keeping in mind that there is a CODB in the online channel. Industry and business research are carried out to ensure the insights derived from this study and well thought through. It is vital to know the domain and gather enough knowledge of the industry to carry out this study. This was conducted by speaking to the department heads and business heads of the organization’s Indian branch.

The organization is in a growth phase since it is new to operating in India. The focus is given to the expansion plan and setting up operations. Hence process improvement and technology enhancement has taken a back seat. This study has given focused on the importance of sales forecasting which will directly help in better inventory management. Lack of knowledge in analytical forecasting techniques is leading to inaccurate inventory pileups and working

capital mismanagement in the previous seasons. Forecasting techniques such as ARIMA and SARIMAX are used to predict future sales. The main aim is to develop a framework for predicting online sales and convincing the management to use analytical tools and techniques for the betterment of the organization.

Chapter 2: Literature Review

Several research papers and blogs are thoroughly read to proceed with this study. Both industry and subject-related research has been carried out to acquire complete knowledge on the main subject of this study.

State of Fashion 2022: To succeed in formulating an effective social media strategy which is an exceedingly challenging task since the number of platforms the strategy needs to be adapted to has changed. The target audience on Facebook vs. Instagram is completely different. Generating relevant and timely content for these platforms is a tedious task and it involves a lot of workforce and investment. So, by default, big organizations have the upper hand when it comes to formulating a more impactful social media strategy (McKinsey & Company, 2022).

Apparel Trends: 2025 what new business models will emerge? The Brands can observe how consumers are interacting with their content which directly influences their buying patterns. New-gen companies leverage emerging tech such as Instagram and WhatsApp to sell and generate more revenue. Currently, the fashion industry is run on a year's old trends, and this is changing fast since social influencers are impacting near-time purchase choices via these mediums (Deloitte Digital, 2022).

The Road to 2025, Five market, trade, and investment trends that will pave the way for the international textile and apparel industry: Online domain has emerged strongly in the past few years due to the digital revolution that is shaping up in India. India will be the world's most tech-savvy e-commerce market with exponential growth due to the rapid growth of internet users in the country (Wazir Advisors, 2022).

Application of Predictive Analytics to Sales Forecasting in Fashion Business: The fashion industry in India is a heavily penetrated competitive market. Due to this reason, it is important to ensure accurate sales forecasting is done. It will also set the expectation for to supply chain department to get the products. The forecast will have to consider certain aspects that are unique to the fashion industry. There is a variety of forecasting methods out there to take

care of these needs of ever challenging fashion industry. Computer-based predictive analytics is one of them. Various forecasting modelling techniques were evaluated and their application to the fashion industry is thoroughly examined. Even though there is a visible benefit of using predicting analytics models in sales forecasting in the fashion industry it is not widely accepted due to the inbuilt nature of the business. This study gives a good understanding of the fashion domain and provides vital insights and the future bottlenecks of predictive analytics in the fashion industry (Bug, 2016).

Forecasting of demand using (Autoregressive Integrated Moving Average) ARIMA model: This paper help me in understanding the practical application of the ARIMA model in a bona fide business problem. Even though it was on-demand forecasting of a company in the Food domain the theoretical aspect of modelling could be directly applied to this study's objectives. It highlights how historical data can be used to forecast the future implications and how it affects the downstream verticals of an organization. The model has pitched well with ARIMA (1, 0, 1) and it was validated by another historical demand data under similar situations. The results achieved agrees that this model can be used to predict future demand in the food industry under the same conditions. The results obtained prove that the model could be utilized to forecast the future demand in this food manufacturing. Hence the approach taken in this study can be adapted to develop a solid model to predict sales in the coming months (Fattah et al., 2018).

Inventory Management using Demand Sales Forecasting: This research paper has focused on finding a solution for issues dealt with in inventory management. How much inventory to keep? How to ensure working capital is not stuck and optimum level of inventory is maintained without going into out-of-stock situations. Due to a lack of experience in forecasting methods companies have settled for simplistic approaches. This study covers the statistical models used to accurately forecast demand. The approach and the vocabulary used is non-technical in this which helps in understanding the content easily (Belgamwar, 2021).

Annual Automobile Sales Prediction Using ARIMA Model: Sales forecasting is one of the most important predictive analytics tools utilized around the globe. The common approach to forecasting is to learn from historic data and predict the future. The assumption is if certain patterns are inbuilt into the data for a long time, they will be appropriate for the future as well. Since it is a generic approach, it can be easily applied to weather prediction, Sales forecasting, etc. Sales prediction will be influenced by quantity sold, inventory, cost of the goods, and the time considered for prediction. This study has predicted the sales quantity for ten years' time series data. The data belongs to Mahindra Tractors Company. The output of the ARIMA model predicts the sales quantity for the next five years (Shakti et al., 2017).

The Research review was conducted in a few stages. The first stage was to refer to industry white papers to gain domain knowledge. Since the company is new to the business and the analytical journey ARIMA & SARIMAX were utilized in forecasting sales to ease the business into forecasting techniques. Post-reading several research articles on this subject there wasn't any ready material a new company in the Indian fashion retail domain could straight away adapt to understand the forecasting techniques. Also, it is crucial for the company to efficiently manage this area due to the many Stock Keeping Units (SKUs). This will directly help in increasing profitability.

Chapter 3: Problem Statement

This study is conducted to Analyse e-commerce sales data acquired by an Indian fashion retailer and suggest areas for profitability improvement and guide the business in data-driven decision-making via concrete recommendations.

Even though online is becoming a strong channel of business, it is also becoming a costly affair. There is an in-built CODB in the online channel. Hence, most offline, and conventional retailers become unsuccessful in this area since they try to apply one formula to all the sales avenues. Retailers who understand the online CODB well easily succeed by offering the right product at the right price and at the right time. The retailer should understand the psyche of the online consumer.

The problem at hand is to derive concrete suggestions based on the profitability analysis and give actionable input to the organization. Also, predict future sales by using forecasting techniques in business analytics to build confidence in the organization.

Chapter 4: Objectives of the Study

This study focuses on three main objectives. Those are,

- Derive specific insights related to CODB and drivers of profitability.
- Suggest improvement on specific categories where the e-retailer is lagging.
- Utilize forecasting techniques to develop a model that can be used in the future.

The sales data collected for this study are specific to one online channel which started its operations in January 2022. The data collected are from February 2022 to July 2022 (Six months). The data is fetched from the organization's Order Management System (OMS) directly. Hence a lot of focus is given to "Data Preparation" to ensure data is all prepped to derive the best output.

This study will be the foundation for building an analytics-driven culture in the organization. Hence a lot of focus will be given to deriving meaningful and actionable insights that will help in identifying key drivers of profitability.

Chapter 5: Project Methodology

This study has been conducted as per Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. This method has been adopted to derive insights from online sales data, conduct an extensive Exploratory Data Analysis (EDA) and conduct forecasting modelling techniques. The CRISP-DM methodology is a well-adapted process model across various industries, and it has six phases which is a step-by-step approach to any data science problem.

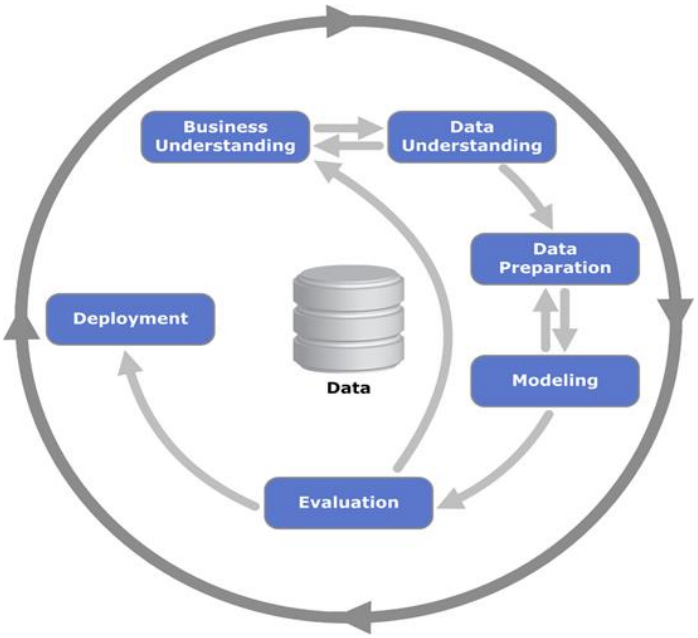


Figure No. 5.1: Cross-industry standard process for data mining Lifecycle (IBM, n.d.)

The six phases mentioned above are thoroughly explained below.

Business Understanding:

The first step of the Business Understanding phase helps in identifying the objectives and purpose of the project. Once the objective and the scope are framed, it helps in assessing the ground reality, defining the set of goals for sales forecasting, understanding the project plan, assessing how the industry and the domain react, and getting a pulse of the business. This is by far the most crucial step before initializing any data science problems.

Data Understanding:

After thoroughly understanding the business, the next step is data understanding. At this stage, needed information is collected about the data. The goal is to identify, gather, and deeply examine the available data. This means observing the available data, exploring the data, and understating various data points, labels, and features. It also includes analysing the data for potential quality issues which may later hamper the quality of the results and the objective of this study. This includes looking at data types and sources and whether it is structured or unstructured data and planning for the approach.

Data Preparation:

This phase is the most time-consuming and vital as it sets the correct precedent for the next steps and prepares the data for the evaluation. The data gathered are pre-processed, cleaned, and observed for any outliers, anomalies, biases, or missing g information are prepped properly based on the domain. Also, the data is either standardized to get rid of biases and feature engineering is conducted to gather fruitful insights. This would also include combining various other data sources and points and restructuring the data frame to perform the next steps.

Modeling:

Forecasting techniques such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) are applied in the modelling phase, where the data models are created based on data and the business goals. These models could be based on Forecasting model techniques.

Model Evaluation:

At this phase, the strength of the model is determined. Depending on the desired output and business goal, the results and findings of the model are tweaked if

needed. This also helps in preparing for the next steps in the approach. The results of the model are assessed against business goals and the desired insights.

Deployment:

This is the final stage of CRISP-DM methodology, where on the successful completion of the evaluation stage, the business deploys the model into live projects to examine the results. This also helps in planning for future projects, looking at the business benefits as they meet the desired goals and the roadmap for the scope. It assesses what is the need of the hour and learns and plans the growth.

Chapter 6: Business Understanding

For any data science project, before assessing the data it is especially important to thoroughly understand the industry and the pulse of the business. This will help in formulating a better result. This study is based on a data set that comes from the Indian fashion retail domain. Below are a few facts from the Indian fashion retail industry which was acquired by a fashion eCommerce report (Statista Digital Market Outlook, 2021).

- Projected revenue in the fashion segment is US$19.69 billion in 2022.
- Expected CAGR is (2022-2025) 18.92%.
- Projected market volume is US$33.11 billion by 2025.
- Expected number of users in the fashion segment is expected to be 446.2m users by 2025.
- User penetration will be at 22.8% in 2022 and hit 30.9% by 2025.
- ARPU (The average revenue per user) will amount to US$61.46.

Market Definition:

The Online sales market share of fashion includes D2C sale of apparel (menswear, womenswear, and Kidswear), footwear, luggage, and bags, as well as accessories (hats and caps, watches, and jewellery) by a medium which is online. The mode of sales in this market share includes e-commerce retailers such as Myntra, AJIO, amazon, etc.

In Scope:

- Apparel and footwear
- Watches, jewellery, and other accessories (e.g., hats, scarves)
- Eyewear
- Luggage and bags
- Leather goods (e.g., leather bags, shoes, and belts)

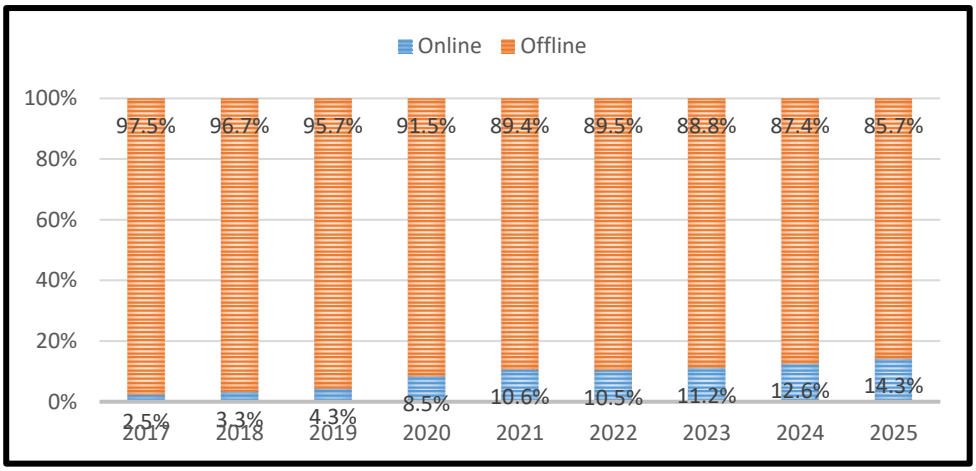


Figure No. 6.1: Sales Market Share by Chanel, Source: (Statista Digital Market Outlook, 2021).

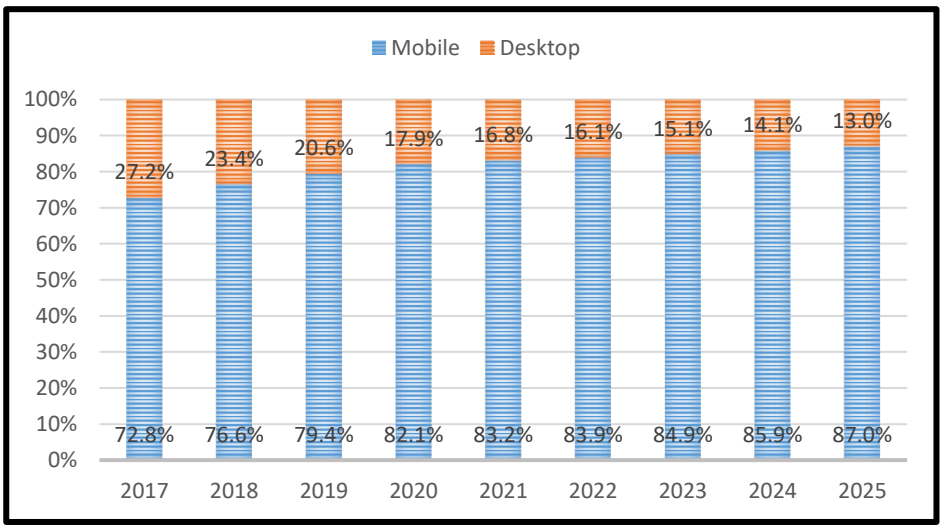


Figure No. 6.2: Desktop vs. Mobile Market share, Source: (Statista Digital Market Outlook, 2021).

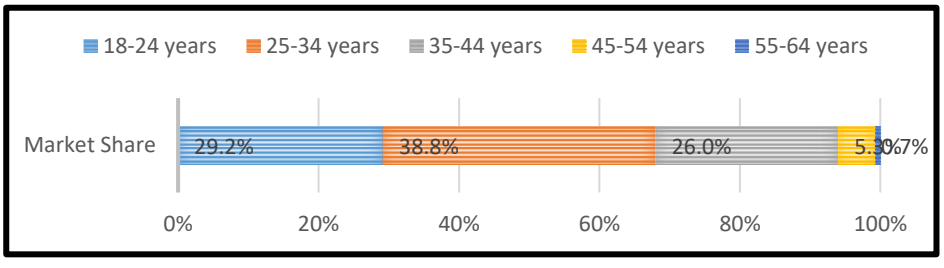


Figure No. 6.3: Market share by age group, Source: (Statista Digital Market Outlook, 2021).

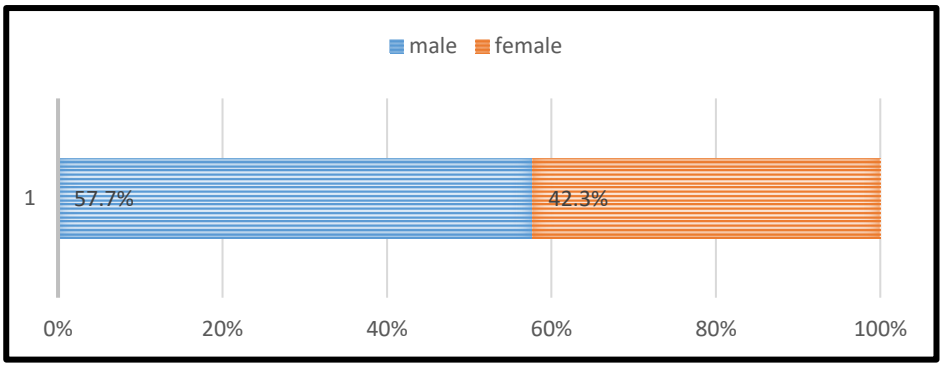


Figure No. 6.4: Market Share by Gender, Source: (Statista Digital Market Outlook, 2021).

The Indian fashion retail domain is expected to achieve great heights by 2025. As per Figure No. 6.1, the Online market share is exponentially growing and continues to grow. Sales via mobile devices are increasing hence the desktop share will keep reducing and this is depicted in Figure No. 6.2. This is a promising figure for the online sales channel.

Sixty-eight percent of the Indian shoppers are below the age of 34 years as per Figure No. 6.3 and most shoppers are male which is 57.7% and this is shown in Figure No. 6.4. This generation is tech-savvy and is most of the workforce hence the focus should be given to these age groups and the male segment.

Post understanding the industry and the domain it is also important to understand the organization. The data belongs to an international fashion retailer operating their business in India since 2021. The Brand operates via offline stores which are in south India and there is a rigorous expansion plan of opening 10+ stores in the South by end of 2022. The Brand's online presence is significant, and it contributes to 32% of the current business. The Brand is doing business on several well-known online aggregators such as Amazon, Ajio, and Myntra via the "marketplace" model.

Chapter 7: Data Understanding

Data understanding and transforming the data to fit the requirement is the most time-bound activity in the entire project. This is a vital activity since the data frame is used in consecutive activities to derive insights and forecast sales. It is best to review the data set thoroughly and understand it before moving on to any data transformation.

So, what does it mean when we say "understand" the data? The objective of this step is to understand the attributes of the data and summarize and derive the essence of the data by identifying key characteristics, such as the volume of data and the total number of variables/attributes in the data. Understanding if there are any issues with the data, such as missing values, inaccuracies, and outliers is vital at this stage.

This study has considered the below sub-sections under "Data understanding"

- Data Source
- Data Exploration

Data Source:

The dataset is acquired from the organization's OMS. The OMS is capturing the B2C sales. There were 110 columns in the original data set. The data set is thoroughly examined to identify the important attributes and omit the rest. Also, the timeline considered for this study is six months (February 2022 – July 2022). New attributes were also derived from the existing data set to ensure more meaningful insights are captured throughout this study. The final data set has twenty-two columns with all missing values imputed.

Number of variables	22
Sample size	76008
Missing cells	1749
Missing cells (%)	0.10%
Duplicate rows	80
Duplicate rows (%)	0.10%
Numeric Variables	9
Categorical Variables	13

Table No. 7.1: Dataset Statistics

Table No. 7.1 depicts a quick snapshot of the collected data. The total no. of records is 76,008. The total missing cells are 1,749. This dataset is a healthy dataset to consider since the missing value percentage is minuscule at 0.1%.

The data source will be further pre-processed in the “Data Preparation” phase. This study identified two main missing elements in the data that could be easily captured. Geographical data and customer contact details. This request has been passed on to the Information Technology (IT) department of the organization and capturing customer and geographical data has been implemented since August 2022.

Data Exploration:

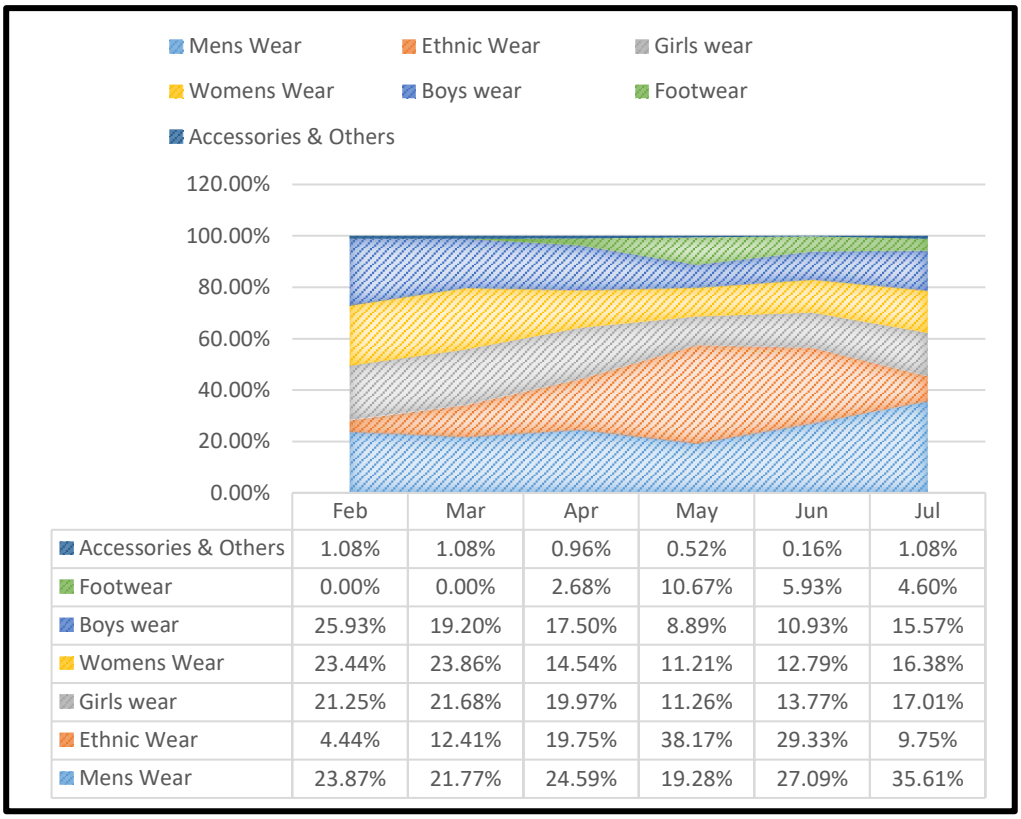
This is the most important section of this study. The insights derived from this will shape the analytical culture of the organization. The aim is to derive actionable and practical insights. The dataset belongs to an organization where processes are new and being framed as you read this report. Hence it is critical to understand each element of the data and where it gets generated to provide meaning to the research carried forward by this study.

Python’s panda’s package was extensively used in this section. Before getting into deep exploration pandas profiling report has been generated to identify important variables and their correlation with other variables. A few important aspects captured in the pandas profiling report are depicted in Table 7.2.

No.	Variable	Highly Correlated Field
1.	Disc %	COGS
2.	Gross Sales	Discount %
3.	Net Profit	Gross Sales
4.	Category	Discount %

Table No. 7.2: Correlation overview of important variables

Correlation depicts the relationship between two variables. As per the pandas profiling report the four variables mentioned in Table No: 7.2 have a high correlation with the mentioned variable. This is an important insight to carry out further analysis and deriving insights.



	Feb	Mar	Apr	May	Jun	Jul
Accessories & Others	1.08%	1.08%	0.96%	0.52%	0.16%	1.08%
Footwear	0.00%	0.00%	2.68%	10.67%	5.93%	4.60%
Boys wear	25.93%	19.20%	17.50%	8.89%	10.93%	15.57%
Womens Wear	23.44%	23.86%	14.54%	11.21%	12.79%	16.38%
Girls wear	21.25%	21.68%	19.97%	11.26%	13.77%	17.01%
Ethnic Wear	4.44%	12.41%	19.75%	38.17%	29.33%	9.75%
Mens Wear	23.87%	21.77%	24.59%	19.28%	27.09%	35.61%

Figure No. 7.1: Category-wise Gross Sales

Figure No. 7.1 illustrates the sales share owned by each category of the organization. Few insights derived from the Figure No. 7.1 are,

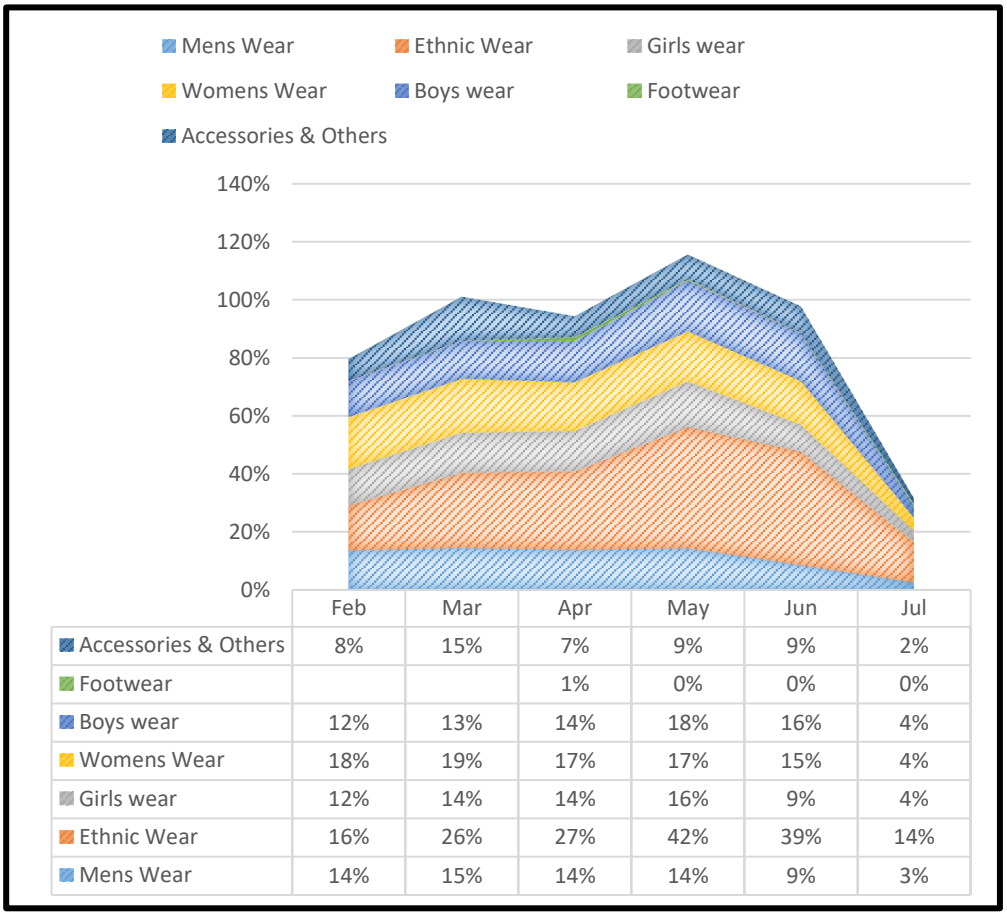
- Men's wear category is the most stable and the biggest contributor to sales.

- Women's wear share has drastically dropped since April 2022.

- Kids wearing both Girl's and Boy's categories have performed well in the first three months and there is a dip in these categories since April 2022.

- Ethnic wear had a slow start in the begging and contributed to the top line of the business in May and June 2022. This volatility is due to other factors and needs more exploration.

- Non-Apparel categories such as footwear, accessories, and others are contributing not more than 6%.



	Feb	Mar	Apr	May	Jun	Jul
Accessories & Others	8%	15%	7%	9%	9%	2%
Footwear			1%	0%	0%	0%
Boys wear	12%	13%	14%	18%	16%	4%
Womens Wear	18%	19%	17%	17%	15%	4%
Girls wear	12%	14%	14%	16%	9%	4%
Ethnic Wear	16%	26%	27%	42%	39%	14%
Mens Wear	14%	15%	14%	14%	9%	3%

Figure No. 7.2: Category-wise Average discount percentage

As per Figure No. 7.2, it is evident that the more discounts you run on the online platforms more top line it will deliver.

- Men's wear is achieving a steady share of business with minimum discounts run online,

- Even though Ethnic wear highlighted promising top-line numbers in Figure No. 7.2, it is evident in Figure 7.2 that it is by running more discounts.

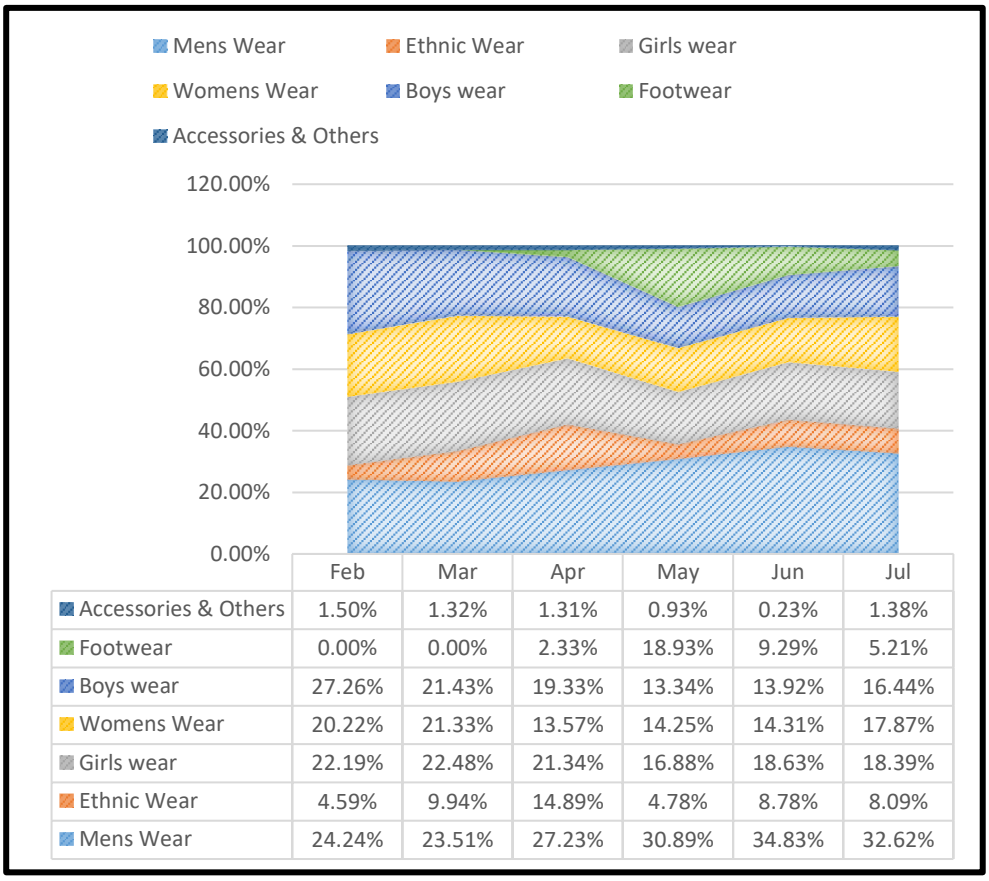
- Other categories such as Boy's, Girl's, and Women's wear categories have maintained a healthy discount percentage.

Another important aspect of profitability is the Cost of Goods Sold (COGS). The lesser the COGS better it is for the sales teams. This gives them more room to adapt offers, Promotions, etc. The simple logic of COGS is derived in Figure No. 7.3.

Profit = Sales – Discounts - COGS

Figure No. 7.3: Profit Formula

As per the main subject of this study, it is especially important to understand how Net profit is distributed across six months. Figure No. 7.4 depicts the same.



	Feb	Mar	Apr	May	Jun	Jul
Accessories & Others	1.50%	1.32%	1.31%	0.93%	0.23%	1.38%
Footwear	0.00%	0.00%	2.33%	18.93%	9.29%	5.21%
Boys wear	27.26%	21.43%	19.33%	13.34%	13.92%	16.44%
Womens Wear	20.22%	21.33%	13.57%	14.25%	14.31%	17.87%
Girls wear	22.19%	22.48%	21.34%	16.88%	18.63%	18.39%
Ethnic Wear	4.59%	9.94%	14.89%	4.78%	8.78%	8.09%
Mens Wear	24.24%	23.51%	27.23%	30.89%	34.83%	32.62%

Figure No. 7.5: Category-wise Net Profit

As per Figure No. 7.5, it is evident that even though online is a discount-led channel it is important to keep an eye on the bottom line to ensure profitability is kept intact. As per Figure No. 7.5, this study has found the below insights.

Men's wear is achieving the highest share of business with low discounts and keeping the bottom line healthy.

- Ethnic wear category is the least profitable category now. The main reason for the net profit of ethnic wear to drop drastically is the high discounts run in May and June.

- All other categories have maintained a healthy Net profit margin.

Post analysing Gross Sales in Figure No. 7.2, Average discount percentage in Figure 7.3, and Net profit in Figure No. 7.5 This study have derived some very insightful information that is affecting profitability. Apart from gross sales, Average discount percentage, and Net profit, some categorical variables directly influence Sales.

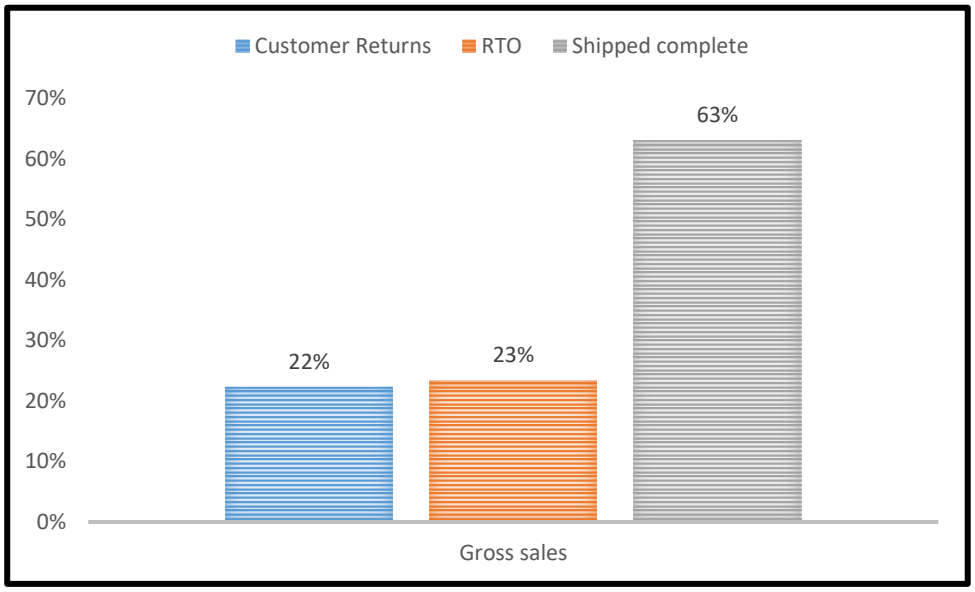


Figure No. 7.6: Online Gross sales as per Order Status

Figure No. 7.6 is capturing the actual sales that finally impacted the top line. Return To Origin (RTO) is Orders that got shipped out of the Organisation's

warehouse but returned without reaching the customer and orders that were cancelled before it was shipped out from the warehouse. Customer returns are genuine returns where the customer has returned it post receiving the order. The numbers in Figure 7.6 are alarming. Forty-five percent of confirmed orders have become returns during the past three months. A thorough root cause analysis needs to be conducted to reduce this number as per the industry standard of <25%.

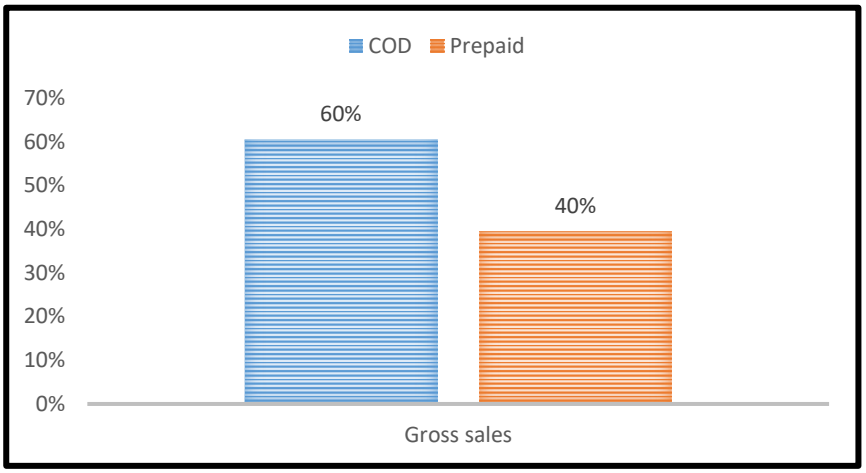


Figure No. 7.7: Online Gross sales as per the mode of payment.

As per Figure No. 7.7, it is evident that Cash of Delivery (COD) is the most preferred mode of payment by the customer. To understand this point further, domain experts were consulted, and it was told that it is the most preferred payment mode due to a lack of trust in order delivery and the ease of payment to the delivery partner via Unified Payments Interface (UPI) payments.

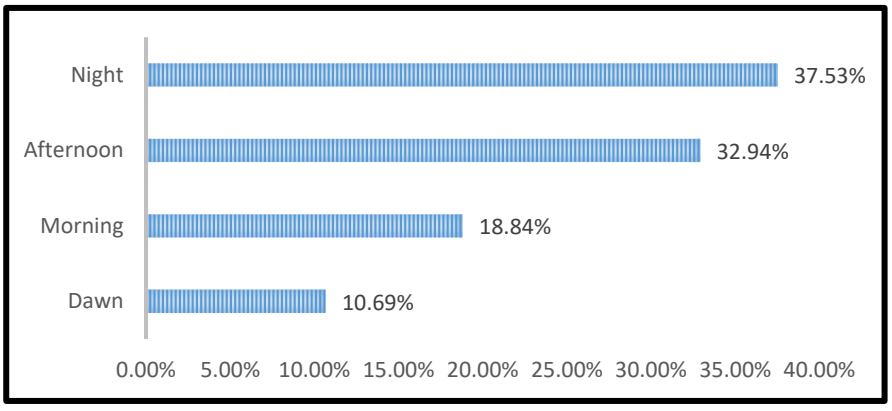


Figure No. 7.8: Online Order flow as per the time of the day.

Figure No. 7.8 depicts the order flow per the four timelines in a day: Dawn, Morning, Afternoon, and Night. Seventy percent of the orders flow in from noon to midnight. Also, 10% of the orders come in during twelve midnight to six in the morning which is a considerable number of orders.

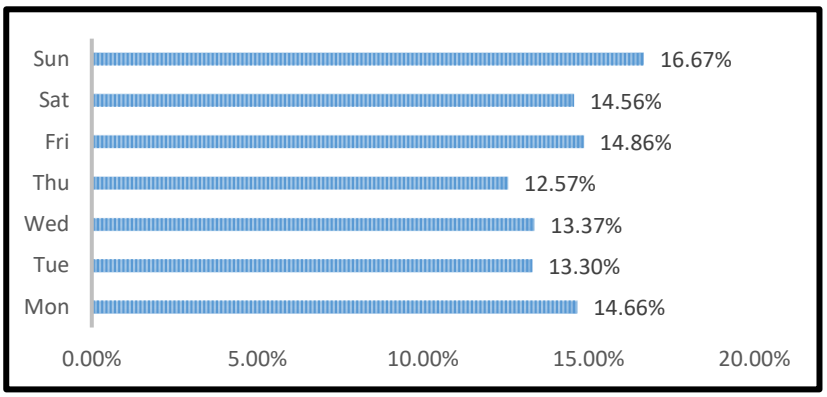


Figure No. 7.9: Online Order flow as per the days of the week.

Figure No. 7.9 illustrates the order flow that comes in each day of the week. His highest order flow is on Sundays followed by Fridays, Mondays, and Saturdays. This is evidence that most of the shopping happens on and around the weekend. Thursdays are the lowest order flow of the week followed by Tuesdays and Wednesdays.

As mentioned at the beginning of this study “Data Exploration” was given a lot of weightage in this project to ensure correct and directional insights are fetched via the dataset. All these insights are passed on to the e-commerce team to carry forward course correction immediately to ensure an optimum level of profitability is achieved.

Chapter 8: Data Preparation

Data preparation is vital in sales forecasting, and it addresses the inconsistencies in data. Before starting pre-processing the data, the source must be identified and reviewed. Then the attributes are labelled as numeric, categorical and character, etc. to understand how to proceed with data preparation. The objective of this phase is to eliminate outliers, and discrepancies while amending the data to derive insights.

This study has directly gathered the data from the OMS backend. Even though it is a directly downloaded file there are a lot of calculated variables that are newly constructed for this operation. The date for the time series dataset is captured via the date and time stamp that gets recorded in the OMS system. The target variable for this study is "Gross Sales" which is "Maximum Retail Price (MRP) – Product discounts offered." Please note that Gross sales are considered which is inclusive of Goods and Service Tax (GST).

The below inconsistencies in data have been dealt with before fitting the ARIMA model.

- Dataset has 80 (0.1%) duplicate rows. The Duplicate rows have been removed from the data set.

- No. of missing data attributes to 1,749 in the data set. The missing values were COGS. This information was available with the business in a separate file and the data has been mapped to all the 1,749 missing COGS values.

All Numerical and Categorical variables are utilized extensively in EDA in the "Data Understanding " chapter. For time series sales forecasting, only two variables are considered out of twenty-two variables.

Converting Month to Datetime:

As the first step, the month is converted into “Datetime” format with the help of python’s Datetime package. This module helps in allowing variables to work with date and time. Datetime is treated as an object in Python.

Visualize the Data:

After prepping the dataset as per the time series equivalent procedure, it is always good to visualize the data. Figure No. 8.1 depicts the first instance of how the “Gross sales” are spread across six months before converting to Datetime format.

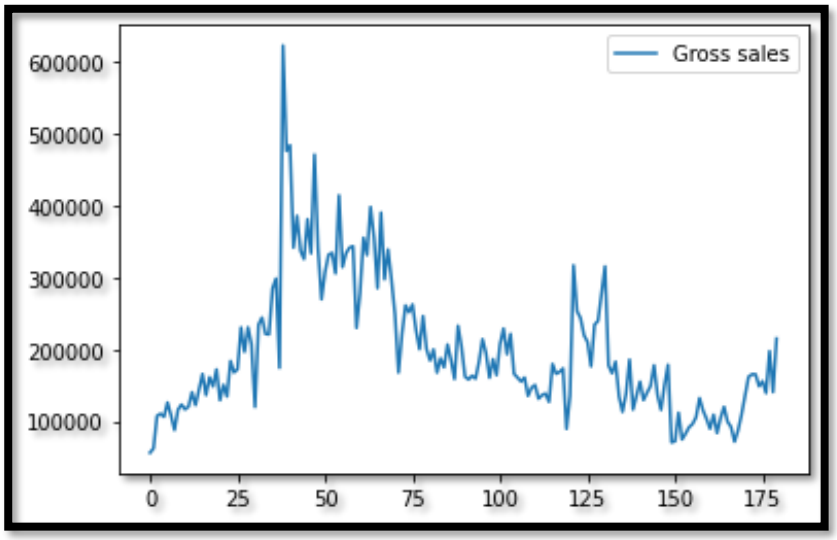


Figure No. 8.1: Gross sales before converting to Datetime

Figure No. 8.2 depicts Gross sales post application of Datetime conversion. This transformation makes the data time series equivalent.

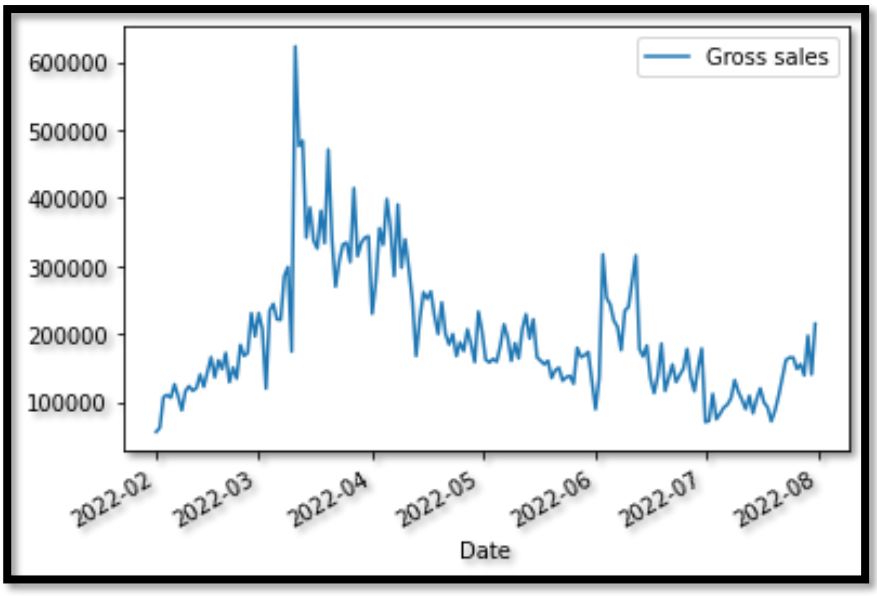


Figure No. 8.2: Gross sales post converting to Datetime format

Testing For Stationarity and Differencing:

If the data is stationary, it means that the statistical properties in the time series data do not change with time. The data need to be stationary since theoretical statistical models and tests depend on the data being stationary. Hence it is vital to evaluate if the data is stationary or non-stationary. “The Dickey-Fuller Test” have been conducted to check if the data is stationary or non-stationary.

The Dickey-Fuller test was the original statistical test formed to test the null hypothesis that a unit root is present in an autoregressive model of a given time series and that the process is thus not stationary. The first test treats the case of a simple lag-1 autoregression model. “Dicky Fuller test” on python was conducted with the below hypothesis built to conduct Dicky-fuller test.

- Null Hypothesis (Ho): It is non-stationary
- Alternative Hypothesis (H1): It is stationary

```
ADF Test Statistic : -1.9091341775456034
p-value : 0.3277763816273437
#Lags Used : 12
Number of Observations Used : 167
weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary
```

Figure No. 8.3: Hypothesis testing for Stationarity

As per Figure No. 8.3 results, P-value is 32% which is greater than 5% hence the data is non-stationary. Statistical models operate with the assumption that the data is stationary hence Differencing is adapted to ensure data becomes stationary. “Sales First difference” is adapted and the “Dicky Fuller” test was re-conducted.

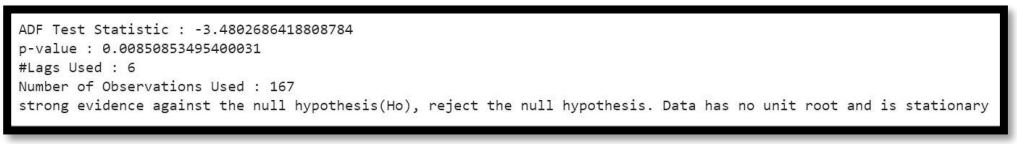


Figure No. 8.4: Hypothesis testing for Stationarity post-Differencing

As per Figure No. 8.4 results, P-value is 0.8% which is lesser than 5% hence the data is stationary. The sales data attracts a lot of noise and volatility. There can be a lot of reasons for the data to be non-station such as inventory, market conditions, seasonality, etc. In the Indian context, the buying patterns will differ even per the festivals of India. Due to the nature of the data and the domain, this study must make the data stationary before carrying out the ARIMA model.

Chapter 9: Modeling

Post getting the data set prepped for conducting the modelling phase this study has fitted ARIMA and SARIMA models. ARIMA is a naïve model used across various business domains. The main reason to select the ARIMA sales forecasting technique is the easy understandability by the business experts. As explained at the beginning of the study, The organization is new to the analytical journey and operations in India. Hence to keep the objective simple yet powerful ARIMA forecasting techniques are utilized. This study has utilized the "stat models" package of python extensively in this chapter.

ARIMA (Auto Regressive Integrated Moving Average):

ARIMA is a group of models that predicts the target variable by utilizing its historical records. Auto-Regressive means that it utilizes "lag values" to forecast. And the Moving Average (MA) component utilizes "lagged forecast errors" for its prediction. Integrated (I) combines both "AR" & "MA" components together (Tony Yiu, 2020).

Due to the nature of the time series data where sales data captured contains seasonality SARIMA model is utilized for predictions.

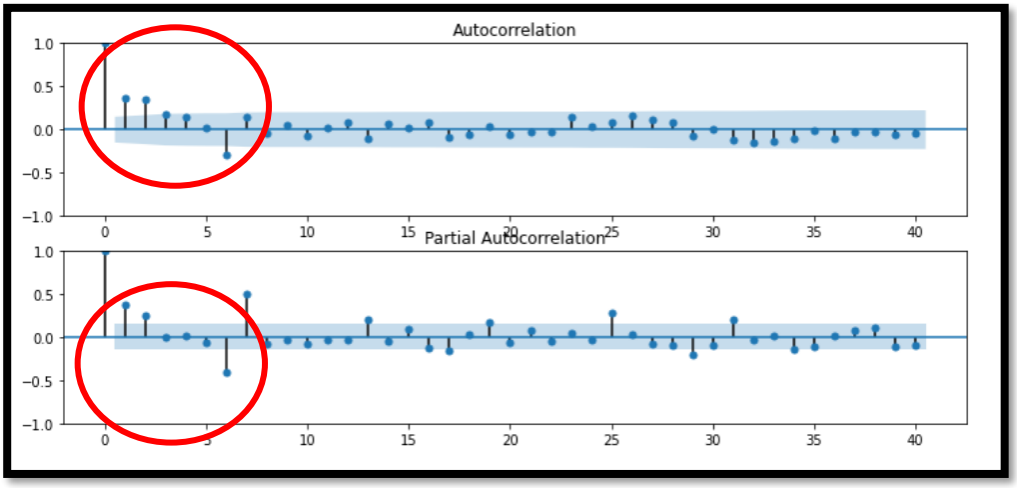


Figure No. 9.1: ACF and PACF plots

Figure No. 9.1 depicts the Auto Correlation Function (ACF) and the Partial Auto Correlation Function (PACF) plots post-fitting the ARIMA model. The ACF and PACF plots are utilized to determine the order of AR, MA, and ARMA models. The order of autoregression in the PACF plot is two. This means the lags do not cross the threshold post the second lag. This means the model's "AR" component has fitted well. And the order of autoregression in ACF is two as well. This translates that the model's "MA" component is fitting well too. In the "Model Evaluation" phase an upgraded version of ARIMA is fitted and evaluated to make the results more adaptable.

Chapter 10: Model Evaluation

"Model Evaluation is the last phase" where all modelling and technical approaches are deeply examined to carry out and conclude the findings. This stage would help in assessing how well the model has performed and how adaptable it is to a bona fide business problem. Hence fashion industry sales data captured seasonality, SARIMAX was further explored to make the approach more adaptable.

SARIMAX:

SARIMAX is an uplifted version of the ARIMA model. Since the data has a seasonal aspect to it SARIMAX would be the best approach to predicting sales. Even external noise can be dialled down by this approach. Hence, the SARIMAX forecasting technique was used to forecast sales in this study (YUGESH VERMA, 2021).

SARIMAX Results

Dep. Variable:	Gross sales	**No. Observations:**	180
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-2195.861
Date:	Tue, 23 Aug 2022	**AIC**	4397.722
Time:	05:28:36	**BIC**	4407.284
Sample:	0	**HQIC**	4401.599
	- 180		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0846	0.080	1.060	0.289	-0.072	0.241
ma.L1	-0.6246	0.079	-7.878	0.000	-0.780	-0.469
sigma2	2.724e+09	6.9e-11	3.95e+19	0.000	2.72e+09	2.72e+09

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	3032.76
Prob(Q):	0.95	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.35	**Skew:**	2.59
Prob(H) (two-sided):	0.00	**Kurtosis:**	22.49

Figure No. 10.1: SARIMAX results

As per Figure No. 10.1, AR lag one is non-significant since the p-value is greater than 5%. However, MA lag one is significant since the p-value is lesser than 5%. As per the "Ljung-Box" Statistical approach, this study was able to achieve a probability score of 0.95 Hence it translates that the model has fitted well.

Forecasting the sales:

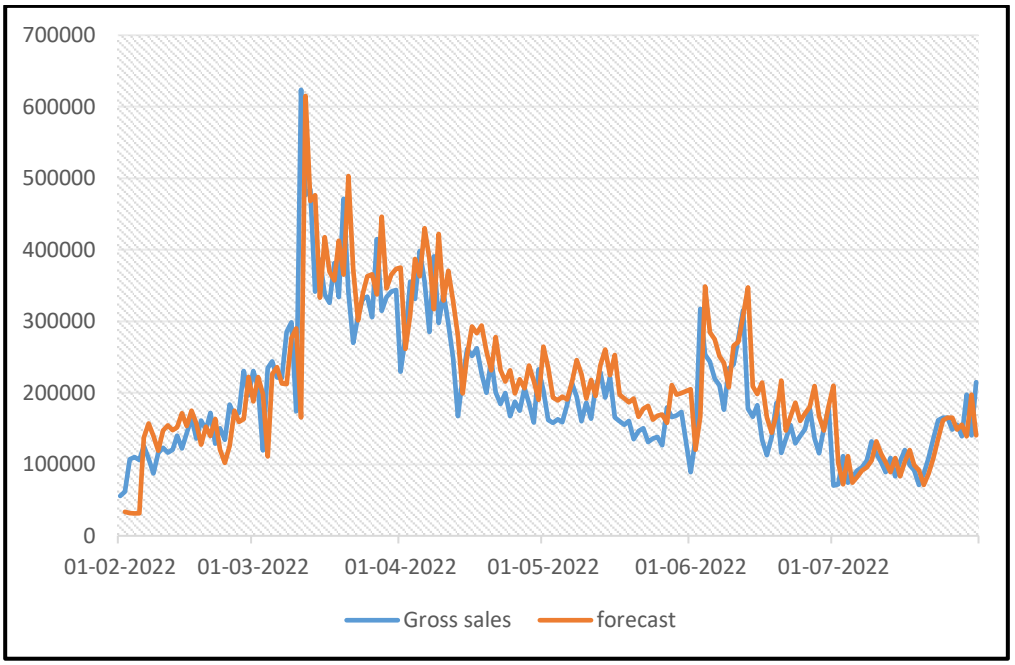


Figure No. 10.2: SARIMAX (Actual vs. Predicted Sales).

As per Figure No. 10.2, this study concludes that SARIMAX is a good forecasting technique to be adapted in the fashion industry. Also, it is a gateway forecasting technique that non-technical business owners and leaders can also adapt.

Chapter 11: Analysis and Results

As mentioned at the beginning of this study the organization the data has derived from is new in adapting to an analytical culture. Their forte lies in operations such as “Product Sourcing,” “Sales and marketing,” etc. Hence, this study’s main objective was to derive meaningful and directional insights the business can seamlessly apply.

- At the beginning of the “Data Understanding” phase, it was identified that the Customer Information and where the order is generated were not captured. Immediate action was taken to capture these aspects.

Men’s wear:

- The men’s wear category is the most stable and the biggest contributor to sales with a healthy top-line as well as a bottom line.
- Men’s wear is achieving a steady share of business with minimum discounts run online.

Women’s western and ethnic wear:

- Women’s wear share has drastically dropped since April 2022.
- Ethnic wear had a slow start in the begging and contributed to the top line of the business in May and June 2022 by funding drastic discounts and the net profit was drastically affected due to this.

Kid’s wear:

- Kid’s wear both Girl’s and Boy’s categories have performed well in the first three months and there is a dip in these categories since April 2022.
- Kid’s wear has maintained a healthy discount and the bottom-line has been healthy as well.

Operational insights:

- Forty-five percent of confirmed orders have become returns during the past three months. A detailed root cause analysis needs to be conducted to reduce this number as per the industry standard of <25%.

- Cash of Delivery (COD) is the most preferred mode of payment by the customer. This is due to a lack of trust in order delivery and the ease of payment to the delivery partner via Unified Payments Interface (UPI) payments.
- Seventy percent of the orders flow in from noon to midnight. Also, 10% of the orders come in during twelve midnight to six in the morning which is a considerable number of orders.
- The highest order flow, 16% is on Sundays followed by Fridays, Mondays, and Saturdays. This is evidence that most of the shopping happens on and around the weekend.

All the above insights and recommendations are by-products of the comprehensive EDA conducted to understand what drives profitability in Indian online retail in the fashion domain. This study extensively helped in introducing sales forecasting techniques to the buying and planning teams of the compa

Chapter 12: Conclusions and Future Scope

This study concludes that SARIMAX is the best sales forecasting technique to adopt since it smoothens out most of the volatility in the data and most of all takes care of the seasonality aspect. This Study's main approach is to convince the organization's management of the advantages of adapting an analytical culture in the organization. The best way is to provide meaningful insights which this study was able to achieve to a larger extent.

Since the Business is new to its operations, The dataset captured for this study is limited to six months of data. Hence, In the future scope of this study, it is recommended to conduct this activity every month by adding new data so the model and learn better.

Once the data set is more mature and the business is also admirably adapting to the recommendations and start seeing the positive impact the model is bringing in forecasting sales, more forecasting techniques such as Long Short-Term Memory (LSTM), Linear Regression, and Random Forest can be explored. This study was able to provide meaningful and actionable insights to the sales team to achieve better profitability. Better forecasting will directly lead to better inventory management, and this influences a positive topline as well as a positive bottom line

Bibliography

Belgamwar, T. (2021). Inventory Management using Demand Sales Forecasting. In *International Journal of Operations Management and Services* (Vol. 11, Issue 1). http://www.ripublication.com

Bug, J. E. P. (2016). *Application of predictive analytics to sales forecasting in the fashion business*. https://www.researchgate.net/publication/325100494

Deloitte Digital. (2022). *Apparel Trends 2025*. https://www.deloittedigital.com/content/dam/deloittedigital/us/documents/blog/blog-20200610-apparel-trends.pdf

Fattah, J., Ezzine, L., Aman, Z., el Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, *10*. https://doi.org/10.1177/1847979018808673

IBM. (n.d.). *Cross-industry standard process for data mining Lifecycle*. Retrieved August 18, 2022, from https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

McKinsey & Company. (2022). *The State of Fashion 2022*. https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/state%20of%20fashion/2022/the-state-of-fashion-2022.pdf

Shakti, S. P., Hassan, M. K., Zhenning, Y., Caytiles, R. D., & N.Ch.S.N, I. (2017). Annual Automobile Sales Prediction Using ARIMA Model. *International Journal of Hybrid Information Technology*, *10*(6), 13–22. https://doi.org/10.14257/ijhit.2017.10.6.02

Statista Digital Market Outlook. (2021). *Fashion eCommerce report 2021*. https://www.statista.com/study/38340/ecommerce-report-fashion/

Tony Yiu. (2020, April 26). *Understanding ARIMA (Time Series Modeling)*. https://towardsdatascience.com/understanding-arima-time-series-modeling-d99cd11be3f8

Wazir Advisors. (2022). *Wazir Report - The Road to 2025*. 1–32. https://wazir.in/pdf/Wazir%20Report%20-%20The%20Road%20to%202025.pdf

YUGESH VERMA. (2021, July 30). *Complete Guide to SARIMAX in Python for Time Series Modeling*. https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/

Appendix

Plagiarism Report[1]

Sales Analytics to drive Profitability - A case study of a Fashion E-Commerce Retailer

by Tharuka G

Submission date: 25-Aug-2022 10:25AM (UTC+0530)
Submission ID: 1886751460
File name: Sales_Analytics_to_drive_Profitability_-_Tharuka.docx (450.2K)
Word count: 6210
Character count: 32116

[1] Turnitn report to be attached from the University.

Sales Analytics to drive Profitability - A case study of a Fashion E-Commerce Retailer

ORIGINALITY REPORT

6%	4%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	global.oup.com Internet Source	1%
2	www.statista.com Internet Source	1%
3	Submitted to University of Aberdeen Student Paper	1%
4	Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab. "Forecasting of demand using ARIMA model", International Journal of Engineering Business Management, 2018 Publication	1%
5	Submitted to University of College Cork Student Paper	<1%
6	www.science.gov Internet Source	<1%
7	Submitted to Higher Education Commission Pakistan Student Paper	<1%

8	Submitted to University of East London Student Paper	<1 %
9	www.miljo.fi Internet Source	<1 %
10	kmps.vse.cz Internet Source	<1 %
11	mdpi-res.com Internet Source	<1 %
12	store.ectap.ro Internet Source	<1 %
13	www.sersc.org Internet Source	<1 %
14	Submitted to SP Jain School of Global Management Student Paper	<1 %
15	www.projectpro.io Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publications in a Conference

Paper submitted: Tharuka Gallekankanamge, J.B. Simha, Rashmi Agarwal, “Sales Analytics to drive Profitability - A case study of a Fashion E-Commerce Retailer”, BAICONF2022 - Ninth International Conference on Business Analytics and Intelligence, 15-17 December, 2022

Submission ID: 2711

Sales Analytics to drive profitability - A case study of a Fashion E-Commerce Retailer

Tharuka Gallekankanamge
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
tharuka.ba07@race.reva.edu.in

JB Simha
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
jb.simha@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
rashmi.agarwal@reva.edu.in

***Abstract*— This research is based on a global fashion retailer that entered India in 2021. This retailer specializes in value fashion format. The value fashion industry in India. Due to the vast diversity in the market, it has become vital that retailers have an Omni presence to cater to the needs of Indian consumers. Today, retailers are thriving both offline and online. Indian retail has shown that being a Direct to Consumer (D2C) brand is important. Also, social commerce is booming, and this is due to India having a younger population compared to the rest of the world.**

This study lays the foundation for introducing an analytical culture to the organization. Various aspects of profitability have been covered in this study. The Indian value fashion market is heavily penetrated by competition. Hence it is vital to keep an eye on the Cost of Doing Business (CODB) and overall profitability. Value fashion brands work on seasonal collections and the Stock Keeping Units (SKUs) in a season are large to ensure all categories are served. Especially when the brand is new to the target market, it is important to find its pros and cons, and Unique Selling Prepositions and create a niche in the industry.

A thorough analysis is conducted to find out what are the key drivers of e-commerce profitability. How are discounts run on the platform affecting the net profit? What are the top categories contributing to a healthy net profit? This study aims to answer all these questions and derive meaningful insights keeping in mind that there is a CODB in the online channel. Industry and business research are carried out to ensure the insights derived from this study and well thought through. It is vital to know the domain and gather enough knowledge of the industry to carry out this study. This is conducted by speaking to the department heads and business heads of the organization's Indian Branch.

The main aim is to develop a framework for predicting online sales and convincing the management to use analytical tools and techniques for the betterment of the organization. This study will be a stepping-stone in creating an analytical culture in the organization. The key insights derived from this study will be passed on to the e-commerce and operations teams that will deploy the suggestions derived from this study.

Forecasting techniques such as (Auto Regressive Integrated Moving Average) ARIMA and Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) are utilized for sales prediction. Overall, a comprehensive approach is taken where most of the weightage is given to the Exploratory Data Analysis (EDA) to derive meaningful and insightful information. The reason for giving more weightage to EDA is to ease the analytical adaptation to ensure concepts are simple and practical to the targeted audience.

Keywords*— *E-commerce, Value Fashion, Trend Analysis, profitability, Sales Analytics, Retail, Sales Forecasting.

I. INTRODUCTION

The retailer considered for this study is a private limited company that specializes in value fashion format in women's wear, men's wear, kids', footwear, and accessories. The value fashion industry in India is booming and India has become a very promising market for both local and international fashion retailers. Due to the vast diversity in the market, it has become vital that retailers have an Omni presence to cater to the needs of Indian consumers. Today, retailers are thriving both offline and online.

The sales data gathered are specific to one online portal which started its operations in January 2022. The brand is operating on "a marketplace model" with one of the major online giants in the country. This means the inventory is owned by the brand and the products are only listed on the online platform. This model is the most preferred mode of operating in e-commerce due to its easy scalability and high turnaround time of working capital.

The focus is given to creating a process for online sales forecasting and encouraging the management to adopt analytical ways of seeing data for the betterment of the organization. The insights and recommendations from this study will be passed on to the e-commerce and operations teams that will deploy the suggestions derived from this study.

ARIMA and SARIMAX forecasting techniques are thoroughly explored for predicting sales. A holistic approach is followed, and more focus and weightage are given to the EDA section. This helps in introducing analytical concepts within the company.

II. LITERATURE REVIEW

Both industry and subject-related research has been carried out to acquire complete knowledge on the main subject of this study.

A. Retail industry research

To succeed in formulating an effective social media strategy which is an exceedingly challenging task since the number of platforms the strategy needs to be adapted to has changed. So, by default, big organizations have the upper hand when it comes to formulating a more impactful social media strategy [1].

New-gen companies leverage emerging tech such as Instagram and WhatsApp to sell and generate more revenue. Currently, the fashion industry is run on a year's old trends, and this is changing fast since social influencers are impacting near-time purchase choices via these mediums [2].

The online domain has emerged strongly in the past few years due to the digital revolution that is shaping up in India. India will be the world's most tech-savvy e-commerce market with exponential growth due to the rapid growth of internet users in the country [3].

B. Subject study-related research

This paper helps in understanding the practical application of the ARIMA model in a bona fide business problem. Even though it is on-demand forecasting of a company in the Food domain the theoretical aspect of modeling could be directly applied to this study's objectives. It highlights how historical data can be used to forecast future implications and how it affects the downstream verticals of an organization. The model has pitched well with ARIMA (1, 0, 1) and it is validated on historical demand data under similar situations. The results achieved agrees that this model can be used to predict future demand in the food industry under the same conditions. The results obtained prove that the model could be utilized to forecast the future demand in this food manufacturing[4]

The fashion industry in India is a heavily penetrated competitive market. Due to this reason, it is important to ensure accurate sales forecasting is done. It will also set the expectation for to supply chain department to get the products. The forecast will have to consider certain aspects that are unique to the fashion industry. There is a variety of forecasting methods out there to take care of these needs of ever challenging fashion industry. Computer-based predictive analytics is one of them. Various forecasting modelling techniques are evaluated and their application to the fashion industry is thoroughly examined. Even though there is a visible benefit of using predicting analytics models in sales forecasting in the fashion industry it is not widely accepted due to the inbuilt nature of the business. This study gives a good understanding of the fashion domain and provides vital insights and the future bottlenecks of predictive analytics in the fashion industry [5].

Sales forecasting is one of the most important predictive analytics tools utilized around the globe. The common approach to forecasting is to learn from historic data and predict the future. The assumption is if certain patterns are inbuilt into the data for a long time, they will be appropriate for the future as well. Since it is a generic approach, it can be easily applied to weather prediction, Sales forecasting, etc. Sales prediction will be influenced by quantity sold, inventory, cost of the goods, and the time considered for prediction. This study has predicted the sales quantity for ten years of time series data. The data belongs to Mahindra Tractors Company. The output of the ARIMA model predicts the sales quantity for the next five years [6].

I. Methodology

A. Business Understanding

This study is based on a data set that comes from the Indian fashion retail domain. Below are a few facts from the Indian fashion retail industry which is acquired by a fashion eCommerce report [7]. The projected revenue in the fashion segment is US$19.69 billion in 2022. The expected CAGR is (2022-2025) 18.92%. The projected market volume is US$33.11 billion by 2025. The expected number of users in the fashion segment is expected to be 446.2 million users by 2025. User penetration will be at 22.8% in 2022 and hit 30.9% by 2025. The average revenue per user (ARPU) will amount to US$61.46.

The Online sales market share of fashion includes D2C sales of apparel (menswear, womenswear, and Kidswear), footwear, luggage, and bags, as well as accessories (hats and caps, watches, and jewellery) by a medium that is online. The mode of sales in this market share includes e-commerce retailers such as Myntra, AJIO, amazon, etc.

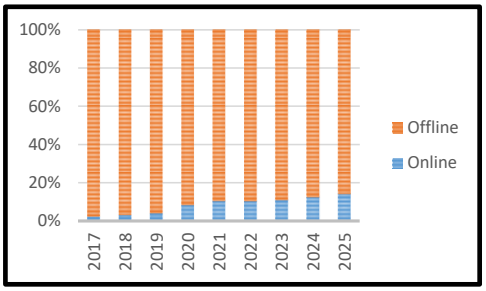


Fig. 1 Sales Market Share by Chanel [7].

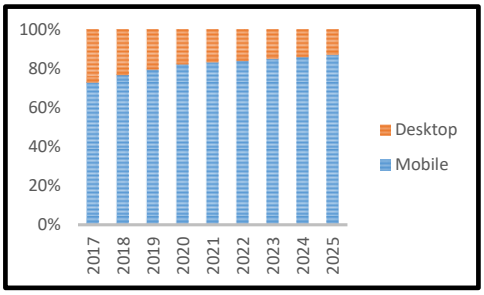


Fig. 2 Desktop vs. Mobile Market share [7].

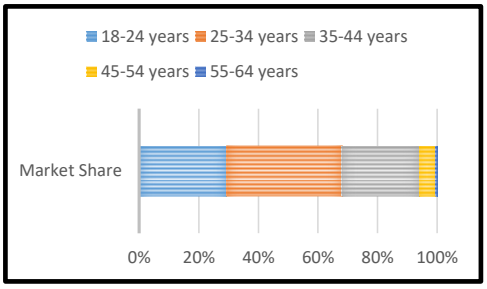


Fig. 3 Market share by age group [7].

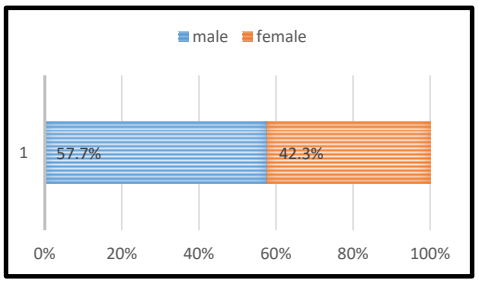


Fig. 4 Market Share by Gender [7].

The Indian fashion retail domain is expected to achieve great heights by 2025. As per Fig. 1, the Online market share is exponentially growing and continues to grow. Sales via mobile devices are increasing hence the desktop share will keep reducing and this is depicted in Fig. 2. This is a promising figure for the online sales channel.

Sixty-eight percent of the Indian shoppers are below the age of 34 years as per Fig. 3 and most shoppers are male which is 57.7% and this is shown in Fig. 4. This generation is tech-savvy and is most of the workforce hence the focus should be given to these age groups and the male segment.

Post understanding the industry and the domain it is also important to understand the organization. The data belongs to an international fashion retailer operating their business in India since 2021. The Brand operates via offline stores which are in south India and there is a rigorous expansion plan of opening 10+ stores in the South by end of 2022. The Brand's online presence is significant, and it contributes to 32% of the current business. The Brand is doing business with several well-known online aggregators such as Amazon, Ajio, and Myntra via the "marketplace" model.

B. Data Understanding

The objective of this step is to understand the attributes of the data and summarize and derive the essence of the data by identifying key characteristics, such as the volume of data and the total number of variables/attributes in the data. Understanding if there are any issues with the data, such as missing values, inaccuracies, and outliers is vital at this stage.

1). Data Source

The dataset is acquired from the organization's Order Management System (OMS). The OMS is capturing the B2C sales. There are 110 columns in the original data set. The data set is thoroughly examined to identify the important attributes and omit the rest. Also, the timeline considered for this study is six months (February 2022 – July 2022). New attributes are also derived from the existing data set to ensure more meaningful insights are captured throughout this study. The final data set has twenty-two columns with all missing values imputed.

TABLE I DATASET STATISTICS

Statistics	Figures
Number of variables	22
Sample size	76008
Missing cells	1749
Missing cells (%)	0.10%
Duplicate rows	80
Duplicate rows (%)	0.10%
Numeric Variables	9
Categorical Variables	13

Table I depicts a quick snapshot of the collected data. The total no. of records is 76,008. The total missing cells are 1,749. This dataset is a healthy dataset to consider since the missing value percentage is minuscule at 0.1%.

The data source will be further pre-processed in the "Data Preparation" phase. This study identified two main missing elements in the data that could be easily captured. Geographical data and customer contact details. This request has been passed on to the Information Technology (IT) department of the organization and capturing customer and geographical data has been implemented since August 2022.

2). Data Exploration

The insights derived from this will shape the analytical culture of the organization. The aim is to derive actionable and practical insights. The dataset belongs to an organization where processes are new and being framed as you read this report. Hence it is critical to understand each element of the data and where it gets generated to provide meaning to the research carried forward by this study. Python's pandas package is extensively used in this section. Before getting into deep exploration pandas profiling report has been generated to identify important variables and their correlation with other variables. A few important aspects captured in the pandas profiling report are depicted in Table II.

TABLE II CORRELATION OVERVIEW OF IMPORTANT VARIABLES

No.	Variable	Highly Correlated Field
1.	Disc %	COGS
2.	Gross Sales	Discount %
3.	Net Profit	Gross Sales
4.	Category	Discount %

Correlation depicts the relationship between two variables. As per the pandas profiling report, the four variables mentioned in Table II have a high correlation with the mentioned variable. This is an important insight to carry out further analysis and deriving insights.

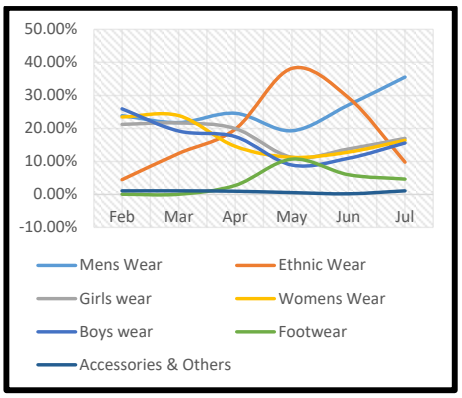


Fig. 5 Category-wise Gross Sales (Share)

Fig. 5 illustrates the sales share owned by each category of the organization. A few insights derived from Fig. 5 are,

a. Men's wear category is the most stable and the biggest contributor to sales.

b. Women's wear share has drastically dropped since April 2022.

c. Kids' wear in both Girl's and Boy's categories have performed well in the first three months and there is a dip in these categories since April 2022.

d. Ethnic wear had a slow start in the begging and contributed to the top line of the business in May and June 2022. This volatility is due to other factors and needs more exploration.

e. non-Apparel categories such as footwear, accessories, and others are contributing not more than 6%.

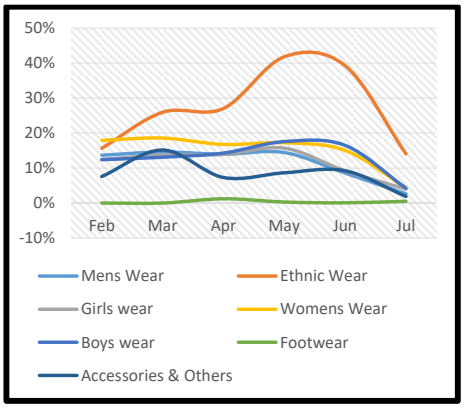


Fig. 6 Category-wise Average discount percentage

As per Fig. 6, it is evident that the more discounts you run on the online platforms more top line it will deliver. Also, Men's wear is achieving a steady share of business with minimum discounts run online, even though Ethnic wear highlighted promising top-line numbers in Fig. 5, it is evident in Fig. 6 that it is by running more discounts. All other categories have maintained a healthy discount percentage.

Another important aspect of profitability is the Cost of Goods Sold (COGS). The lesser the COGS better it is for the sales teams. This gives them more room to adapt offers, Promotions, etc. The simple logic of COGS is explained in equation (1).

$$Profit = Sales - Discounts - COGS \quad (1)$$

As per the main subject of this study, it is especially important to understand how Net profit is distributed across six months. Fig. 8 depicts the same.

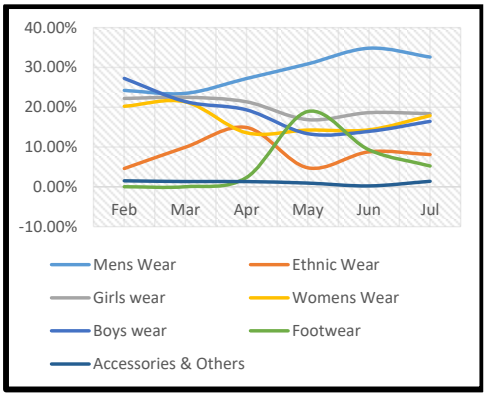


Fig. 8 Category-wise Net Profit

As per Fig. 8, it is evident that even though online is a discount-led channel it is important to keep an eye on the bottom line to ensure profitability is kept intact. As per Fig. 8, this study has found the below insights.

a. Men's wear is achieving the highest share of business with low discounts and keeping the bottom line healthy.

b. The ethnic wear category is the least profitable category now. The main reason for the net profit of ethnic wear to drop drastically is the high discounts run in May and June.

c. All other categories have maintained a healthy Net profit margin.

Post analysing Gross Sales in Fig. 5, Average discount percentage in Fig. 6, and Net profit in Fig. 8, this study has derived some very insightful information that is affecting profitability. Apart from gross sales, Average discount percentage, and Net profit, some categorical variables directly influence Sales.

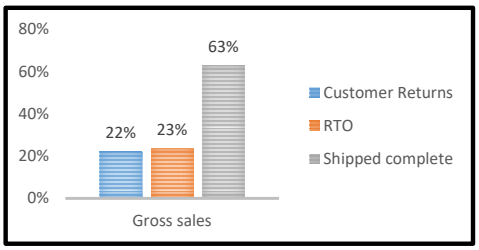


Fig. 9 Online Gross sales as per Order Status

Fig. 9 is capturing the actual sales that finally impacted the top line. Return To Origin (RTO) is Orders that got shipped out of the Organisation's warehouse but returned without reaching the customer and orders that are cancelled before it is shipped out from the warehouse. Customer returns are genuine returns where the customer has returned it post receiving the order. The numbers in Fig. 10 are alarming. Forty-five percent of confirmed orders have become returns during the past three months. A thorough root cause analysis needs to be conducted to reduce this number as per the industry standard of <25%.

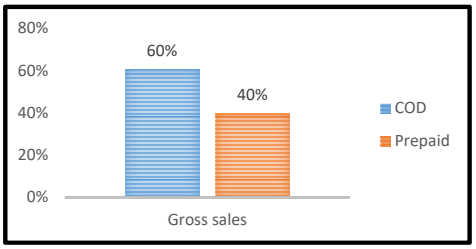


Fig. 10 Online Gross sales as per the mode of payment.

As per Fig. 10, it is evident that Cash of Delivery (COD) is the most preferred mode of payment by the customer. To understand this point further, domain experts are consulted, and it is told that it is the most preferred payment mode due to a lack of trust in order delivery and the ease of payment to the delivery partner via Unified Payments Interface (UPI) payments.

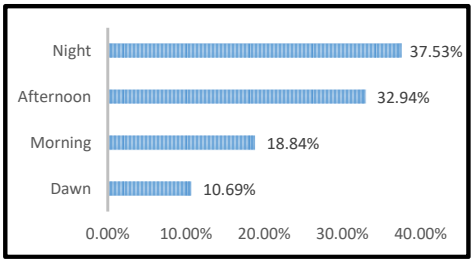


Fig. 11 Online Order flow as per the time of the day.

Fig. 11 depicts the order flow per the four timelines in a day: Dawn, Morning, Afternoon, and Night. Seventy percent of the orders flow in from noon to midnight. Also, 10% of the orders come in between twelve midnight to six in the morning which is a considerable number of orders.

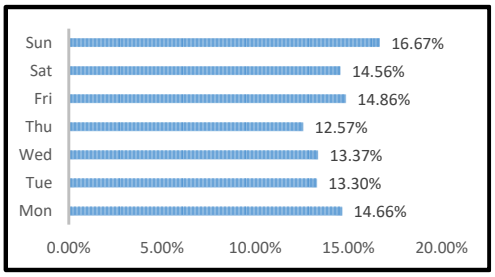


Fig. 12 Online Order flow as per the days of the week.

Fig. 12 illustrates the order flow that comes in each day of the week. The highest order flow is on Sundays followed by Fridays, Mondays, and Saturdays. This is evidence that most of the shopping happens on and around the weekend. Thursdays are the lowest order flow of the week followed by Tuesdays and Wednesdays.

As mentioned at the beginning of this study "Data Exploration" is given a lot of weightage in this project to ensure correct and directional insights are fetched via the dataset. All these insights are passed on to the e-commerce team to carry forward course correction immediately to ensure an optimum level of profitability is achieved.

C. Data Preparation

Data preparation is vital in sales forecasting, and it addresses the inconsistencies in data. Before starting pre-processing the data, the source must be identified and reviewed. Then the attributes are labelled as numeric, categorical and character, etc. to understand how to proceed with data preparation. The objective of this phase is to eliminate outliers, and discrepancies while amending the data to derive insights.

This study has directly gathered the data from the OMS backend. Even though it is a directly downloaded file there are a lot of calculated variables that are newly constructed for this operation. The date for the time series dataset is captured via the date and time stamp that gets recorded in the OMS system. The target variable for this study is "Gross Sales" which is "Maximum Retail Price (MRP) – Product discounts offered." Please note that Gross sales are considered which is inclusive of Goods and Service Tax (GST).

The below inconsistencies in data have been dealt with before fitting the ARIMA model.

1. Dataset has 80 (0.1%) duplicate rows. The Duplicate rows have been removed from the data set.
2. No. of missing data attributes to 1,749 in the data set. The missing values are COGS. This information is available with the business in a separate file and the data has been mapped to all the 1,749 missing COGS values.

All Numerical and Categorical variables are utilized extensively in EDA in the "Data Understanding " section. For time series sales forecasting, only two variables are considered out of twenty-two variables.

1). Converting Month to Datetime

As the first step, the month is converted into "Datetime" format with the help of python's Datetime package. This module helps in allowing variables to work with date and time. Datetime is treated as an object in Python.

2). Visualize the Data

After prepping the dataset as per the time series equivalent procedure, it is always good to visualize the data. Fig. 13 depicts the first instance of how the "Gross sales" are spread across six months before converting to Datetime format.

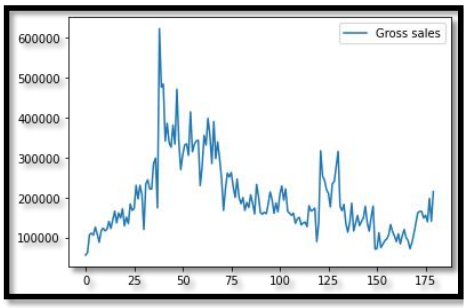


Fig. 13 Gross sales before converting to Datetime

Fig. 14 depicts Gross sales post-application of Datetime conversion. This transformation makes the data time series equivalent.

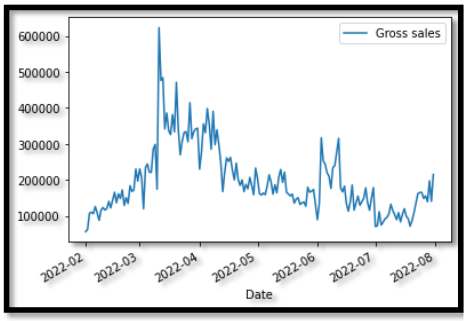


Fig. 14 Gross sales post converting to Datetime format

3). Testing For Stationarity and Differencing

If the data is stationary, it means that the statistical properties in the time series data do not change with time. The data need to be stationary since theoretical statistical models and tests depend on the data being stationary. Hence it is vital to evaluate if the data is stationary or non-stationary. "The Dickey-Fuller Test" have been conducted to check if the data is stationary or non-stationary.

The Dickey-Fuller test is the original statistical test formed to test the null hypothesis that a unit root is present in an autoregressive model of a given time series and that the process is thus not stationary. The first test treats the case of a simple lag-1 autoregression model. The "Dicky Fuller test" on python is conducted with the below hypothesis built to conduct Dicky-fuller test.

1. Null Hypothesis (Ho): It is non-stationary
2. Alternative Hypothesis (H1): It is stationary

```
ADF Test Statistic : -1.9091341775456034
p-value : 0.3277763816273437
#Lags Used : 12
Number of Observations Used : 167
```

Fig. 15 Hypothesis testing for Stationarity

As per Fig. 15 results, P-value is 32% which is greater than 5% hence the data is non-stationary. Statistical models operate with the assumption that the data is stationary hence Differencing is adapted to ensure data becomes stationary. "Sales First difference" is adapted and the "Dicky Fuller" test is re-conducted.

```
ADF Test Statistic : -3.4802686418808784
p-value : 0.00850853495400031
#Lags Used : 6
Number of Observations Used : 167
```

Fig. 16 Hypothesis testing for Stationarity post-Differencing

As per Fig. 16 results, P-value is 0.8% which is lesser than 5% hence the data is stationary. The sales data attracts a lot of noise and volatility. There can be a lot of reasons for the data to be non-station such as inventory, market conditions, seasonality, etc. In the Indian context, the buying patterns will differ even per the festivals of India. Due to the nature of the data and the domain, this study must make the data stationary before carrying out the ARIMA model.

D. Modeling

Post getting the data set prepped for conducting the modelling phase this study has fitted ARIMA and SARIMA models. ARIMA is a naïve model used across various business domains. The main reason to select the ARIMA sales forecasting technique is the easy understandability by the business experts. As explained at the beginning of the study, The organization is new to the analytical journey and operations in India. Hence to keep the objective simple yet powerful ARIMA furcating techniques are utilized. This study has utilized the "stat models" package of python extensively in this chapter.

1). ARIMA

ARIMA is a group of models that predicts the target variable by utilizing its historical records. Auto-Regressive

Average (MA) component utilizes "lagged forecast errors" for its prediction. Integrated (I) combines both "AR" & "MA" components together [8]. Due to the nature of the time series data where sales data captured contains seasonality SARIMA model is utilized for predictions.

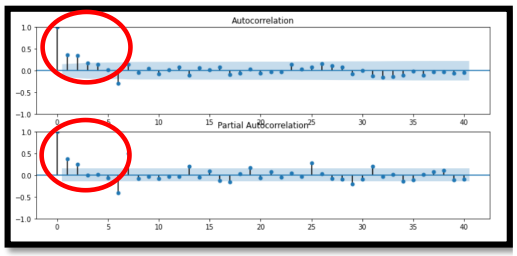


Fig. 17 ACF and PACF plots

Fig. 17 depicts the Auto Correlation Function (ACF) and the Partial Auto Correlation Function (PACF) plots post-fitting the ARIMA model. The ACF and PACF plots are utilized to determine the order of AR, MA, and ARMA models. The order of autoregression in the PACF plot is two. This means the lags do not cross the threshold post the second lag. This means the model's "AR" component has fitted well. And the order of autoregression in ACF is two as well. This translates that the model's "MA" component is fitting well too. In the "Model Evaluation" phase an upgraded version of ARIMA is fitted and evaluated to make the results more adaptable.

2). SARIMAX

SARIMAX is an uplifted version of the ARIMA model. Since the data has a seasonal aspect to it SARIMAX would be the best approach to predicting sales. Even external noise can be dialed down by this approach. Hence, the SARIMAX forecasting technique is used to forecast sales in this study [9].

SARIMAX Results

Dep. Variable:	Gross sales	**No. Observations:**	180
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-2195.861
Date:	Tue, 23 Aug 2022	**AIC**	4397.722
Time:	05:28:36	**BIC**	4407.284
Sample:	0	**HQIC**	4401.599
	- 180		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0846	0.080	1.060	0.289	-0.072	0.241
ma.L1	-0.6246	0.079	-7.878	0.000	-0.780	-0.469
sigma2	2.724e+09	6.9e-11	3.95e+19	0.000	2.72e+09	2.72e+09

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	3032.76
Prob(Q):	0.95	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.35	**Skew:**	2.59
Prob(H) (two-sided):	0.00	**Kurtosis:**	22.49

Fig. 18 SARIMAX results

As per Fig. 18, AR lag one is non-significant since the p-value is greater than 5%. However, MA lag one is significant since the p-value is lesser than 5%. As per the "Ljung-Box" Statistical approach, this study can achieve a probability score of 0.95. Hence, it translates that the model has fitted well.

3). Forecasting the sales

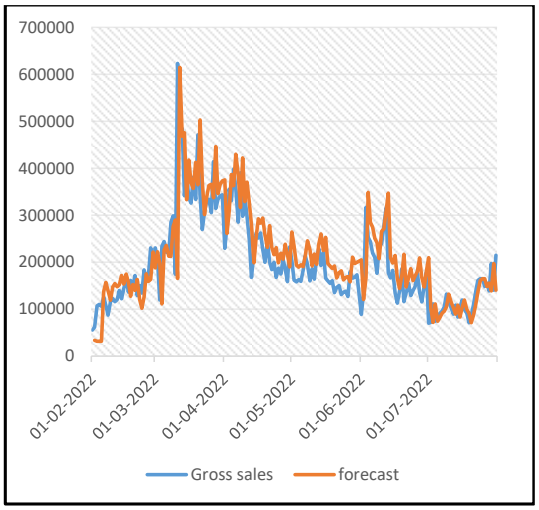


Fig. 19 SARIMAX (Actual vs. Predicted Sales).

As per Fig. 19, this study concludes that SARIMAX is a good forecasting technique to be adapted in the fashion industry. Also, it is a gateway forecasting technique that non-technical business owners and leaders can also adapt easily.

I. CONCLUSION

As mentioned at the beginning of this paper the organization the data has derived from is new in adapting to an analytical culture. Their forte lies in operations such as "Product Sourcing," "Sales and marketing," etc. Hence, this study's main objective is to derive meaningful and directional insights the business can seamlessly apply. At the beginning of the "Data Understanding" phase, it is identified that the customer information and where the order is generated are not captured.

This study concludes that SARIMAX is the best sales forecasting technique to adopt since it smoothens out most of the volatility in the data and most of all takes care of the seasonality aspect. This Study's main approach is to convince the organization's management of the advantages of adopting an analytical culture in the organization. The best way is to provide meaningful insights which this study can achieve to a larger extent.

II. FUTURE SCOPE

Since the business is new to its operations, The dataset captured for this study is limited to six months of data. Hence, In the future scope of this study, it is recommended to conduct this activity every month by adding new data so the model and learn better.

Once the data set is more mature and the business is also admirably adapting to the recommendations and start seeing

the positive impact the model is bringing in forecasting sales more forecasting techniques such as Long Short-Term Memory (LSTM), Linear Regression, and Random Forest can be explored.

ACKNOWLEDGMENT

We would like to convey heartfelt gratitude to all the mentors at Reva Academy of Corporate Excellence, Mithun Dolthody Jayaprakash, and Ratnakar Pandey for their continuous support throughout the learning journey. We would like to express a special thanks to Dr. Shinu Abhi, Director of REVA Academy of Corporate Excellence for her amiable support, helpful guidance, and information at various phases of this study, which helped in completing it.

REFERENCES

McKinsey & Company, "The State of Fashion 2022," 2022. Accessed: Aug. 15, 2022. [Online]. Available: https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/state%20of%20fashion/2022/the-state-of-fashion-2022.pdf

Deloitte Digital, "Apparel Trends 2025," 2022. Accessed: Aug. 15, 2022. [Online]. Available: https://www.deloittedigital.com/content/dam/deloittedigital/us/documents/blog/blog-20200610-apparel-trends.pdf

Wazir Advisors, "Wazir Report - The Road to 2025," pp. 1–32, 2022, Accessed: Aug. 15, 2022. [Online]. Available: https://wazir.in/pdf/Wazir%20Report%20-%20The%20Road%20to%202025.pdf

J. Fattah, L. Ezzine, Z. Aman, H. el Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, vol. 10, Oct. 2018, doi: 10.1177/1847979018808673.

J. E. P. Bug, "Application of predictive analytics to sales forecasting in fashion business," 2016. [Online]. Available: https://www.researchgate.net/publication/325100494

S. P. Shakti, M. K. Hassan, Y. Zhenning, R. D. Caytiles, and I. N.Ch.S.N, "Annual Automobile Sales Prediction Using ARIMA Model," *International Journal of Hybrid Information Technology*, vol. 10, no. 6, pp. 13–22, Jun. 2017, doi: 10.14257/ijhit.2017.10.6.02.

Statista Digital Market Outlook, "Fashion eCommerce report 2021," Jul. 2021. Accessed: Aug. 18, 2022. [Online]. Available: https://www.statista.com/study/38340/ecommerce-report-fashion/

Tony Yiu, "Understanding ARIMA (Time Series Modeling)," Apr. 26, 2020. https://towardsdatascience.com/understanding-arima-time-series-modeling-d99cd11be3f8 (accessed Aug. 23, 2022).

YUGESH VERMA, "Complete Guide To SARIMAX in Python for Time Series Modeling," Jul. 30, 2021. https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/ (accessed Aug. 23, 2022).

GitHub Link:

https://github.com/TharukaG/ARIMA_2nd_year_V3.ipynb.git