E-COMMERCE TRENDS AND PREDICTIONS USING BIG DATA TOOLS

Manoj Suggala
Computer Science
University of Missouri,
Kansas City
Kansas City Missouri
USA
msqxh@umkc.edu

Praveen Kumar Reddy Kadapala
Computer Science
University of Missouri,
Kansas City
Kansas City Missouri
USA
pkhmf@umkc.edu

Sreevardhan Reddy Soma
Computer Science
University of Missouri,
Kansas City
Kansas City Missouri
USA
ssfwk@umkc.edu

Rahul Karthik Arunachalam Usharani
Computer Science
University of Missouri
Kansas City
Kansas City Missouri
USA
ra6zr@umkc.edu

1. ABSTRACT

In today's dynamic e-commerce landscape, big data analytics is emerging as a powerful tool for businesses to gain a competitive edge. E-commerce is a strong catalyst for economic development. The rapid growth in usage of Internet and Web-based applications is decreasing operational costs of large enterprises, extending trading opportunities and lowering the financial barriers for active e-commerce participation. Many companies are restructuring their business strategies to attain maximum value in terms of profits as well as customer satisfaction. The objective of this paper is to evaluate the trend and analysis of Flipkart sales. We have used Hadoop for data storage, spark for data analysis and Matplotlib and Seaborn for data visualization.

2. INTRODUCTION

With the increase of e-commerce platforms, vast amounts of data are being generated which can be used to provide key insights for user satisfaction. This paper explores several big data tools and concepts to extract the trends and predictions. The dataset used in this paper is taken from Flipkart, one of the leading e-commerce platforms in India. By using big data tools, patterns in customer purchase behavior, product choices and preferences are analyzed to identify major trends which include seasonal buying patterns and customer sentiment patterns.

This project's primary goal is to exploit data from e-commerce sites, including purchase history, to aid businesses with logistics, profitability optimization, and brand value enhancement through advertising. To enhance business products or services, companies employ data extracted from online product purchase history. These records are monitored to provide insights into customer behavior and market trends.

3. RELATED WORK

The author “Shaikh, Eman” conducted a review on Apache Spark to determine the platform's salient characteristics, advantages, and drawbacks. This assisted them in gaining a thorough understanding of Apache Spark and pinpointing the areas in which more research is required. The writers compared Hadoop and Hive with Apache Spark, this assisted them in highlighting Apache Spark's advantages and disadvantages as well as determining when it is the best option for a certain task. The authors provided the development and implementation of actual big data applications using Apache Spark. This aided in showcasing Apache Spark's capabilities and offering suggestions for applying it to practical problem-solving.

Piyush Sewal and Hari Singh, A sizable dataset was gathered by the writers from an e-commerce site. Data on product reviews, consumer transactions, and other pertinent variables were included in this dataset. After cleaning and preparing the data for analysis, the writers preprocessed it. Using the preprocessed data, the authors created and assessed several machine learning models. These models were utilized for a variety of functions, including fraud detection, product recommendation, and consumer segmentation.

Khadija Aziz, Dounia Zaidouni, the authors included several instances demonstrating how the machine learning models they had created may be applied to enhance e-commerce decision-making. The authors employed a range of machine learning algorithms, including random forests, decision trees, and logistic regression. Additionally, they employed various hyper-parameter optimization strategies to raise the models' performance.

Hong Li, The application of Spark Streaming for real-time data processing in e-commerce is the main topic of research. The authors use Spark Streaming to assess consumer behavior and

generate customized product recommendations. The data is shown using a web application, which provides real-time insights for improving customer engagement and sales.

Penglin Gao, Zhaoming Han,suggests a strategy for using Spark to analyze e-commerce data. The authors use Spark to process customer transaction data and generate insights on product sales, revenue patterns, and consumer behavior. The data is shown using Tableau, which provides insights for improving product recommendations and marketing strategies.

4. PROPOSED TECHNIQUES

Our goal in this research is to investigate how Flipkart data analysis might benefit from distributed computing using Hadoop and Apache Spark. We want to investigate how combining Hadoop with Apache Spark might speed up and increase the precision of dataset analysis and to process the Flipkart dataset faster and more effectively by merging these two systems. We are also making use of data visualization tool Matplotlib to assist in converting challenging data into readable and visually appealing graphs, charts, and interactive dashboards.

5. DATA AND METHODOLOGY

We collected our dataset of 4.5GB of data from Kaggle. https://www.kaggle.com/datasets/iyumrahul/flipkartsalesdataset/?select=products.csv.We have two datasets in this project. Sales dataset and Products dataset.

Sales dataset contains the records of all the sales in the month of April in India. It consists of more than 40 million records, and it has 13 columns which include.

- Date
- Order_id,
- Product_id,
- City_name,
- Unit_selling
- Customer_id

and all the required sales columns.

Figure 1 contains the details for the Sales dataset. Figure 2 contains the number of records in the sales dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Unnam	Unnam	date_	city_name	order_id	cart_id	dim_custom	procured_	unit_sellin	total_discou	product_id	total_weighted_	landing_price	
2	0	0	0	4/1/2022	Mumbai	112246974	173273802	17995199	1	234	0	344107	202.51303		
3	1	1	1	4/1/2022	Bengaluru	112246976	173273597	18259433	1	64	0	389676	48.714375		
4	2	2	2	4/1/2022	Bengaluru	112247019	173123717	5402601	1	1031	0	39411	975.996		
5	3	3	3	4/1/2022	HR-NCR	112247045	172547459	15649744	1	57	0	369742	25		
6	4	4	4	4/1/2022	Mumbai	112247123	173081820	10127605	2	30	0	12872	57.980004		
7	5	5	5	4/1/2022	HR-NCR	112247149	173274225	479129	1	65	0	99407	61.5		
8	6	6	6	4/1/2022	Bengaluru	112247170	173273989	8658748	1	115	0	439935	103.05988		
9	7	7	7	4/1/2022	Mumbai	112247182	173273662	16541671	1	20	0	424394	15.399999		
10	8	8	8	4/1/2022	Delhi	112247233	173269263	21466	1	24	0	470636	18		
11	9	9	9	4/1/2022	Delhi	112247239	173274446	13522664	1	36	0	311	32.38837		
12	10	10	10	4/1/2022	Delhi	112247317	173274278	13497358	1	20	0	339152	12		
13	11	11	11	4/1/2022	HR-NCR	112247321	173273059	4480518	1	50	0	18546	46.477013		
14	12	12	12	4/1/2022	Delhi	112247317	173274278	13497358	1	23	0	4927	14		
15	13	13	13	4/1/2022	Bengaluru	112247328	173269880	18049405	2	19	0	37083	36.1		
16	14	14	14	4/1/2022	Delhi	112247393	173273514	5933479	1	130	0	366189	113.04337		
17	15	15	15	4/1/2022	Bengaluru	112247425	173273350	17213227	1	13	0	3928	8		
18	16	16	16	4/1/2022	HR-NCR	112247431	173274695	1778240	1	35	0	424982	29.799997		

Figure 1: Sales

```
Number of rows in the sales: 46706387
```

Figure 2: Records in Sales

The product dataset contains the data of all the products available for sale in Flipkart, it consists of more than 32000 products, and it has 12 columns which include.

- Product_id
- Product_name,
- Product_type,
- Brand_name,
- Manufacturer_name

and all the required products columns.

Figure 3 contains the details for the Products dataset. Figure 4 contains the number of records in the Products dataset.

	A	B	C	D	E	F	G	H
1		product_id	product_name	unit	product_typ	brand_name	manufacturer_	l0_category
2	0	476763	Christmas - Carc	1 unit	Card		HOT	Specials
3	1	483436	Plum BodyLovin	20 ml	Sample	Plum BodyLovi	Pureplay Skin S	Specials
4	2	476825	Diwali Gift Card	1 unit	Sample		HOT	Specials
5	3	483438	Plum BodyLovin	20 ml	Sample	Plum BodyLovi	Pureplay Skin S	Specials
6	4	480473	Flipkart Valentin	1 unit	Card	Flipkart	Dummy Manuf	Specials
7	5	483694	Dabur Vita Choc	75 g	Sample	Dabur	Dabur India Pv	Specials
8	6	486016	Plum Green Tea	15 ml	Sample	Plum	Pureplay Skin S	Specials
9	7	486017	Plum Green Tea	15 ml	Sample	Plum	Pureplay Skin S	Specials
10	8	486124	Kari kari Salt & P	22 g	Sample	Kari kari	LT Foods	Specials
11	9	486125	Maggi Liquid Co	180 ml	Sample	Maggi	Nestle India	Specials
12	10	486376	Orion Mango Ch	28 g	Sample	Orion	ORION FOOD V	Specials
13	11	486377	Tata Tea Gold C	25 g	Sample	Tata Tea	Tata consumer	Specials
14	12	486559	Conscious Food	50 g	Sample	Conscious Foo	Conscious Foo	Specials
15	13	487659	The Laughing Co	15 g	Sample	The Laughing C	Bel Vietnam Co	Specials
16	14	487667	Sunfeast Dark F	20 g	Sample	Sunfeast Dark	ITC Limited	Specials
17	15	487668	Sunfeast All Rou	28.2 g	Sample	Sunfeast	ITC Limited	Specials
18	16	488061	Sleepy Owl Cold	5 x 50 g	Sample	Sleepy Owl	Sleepy Owl Cof	Specials

Figure 3: Products

```
Number of rows in the products: 32226
```

Figure 4: Records in Products

6. DATA STORAGE

We are using Hadoop Hadoop Distributed File System (HDFS) as storage in our project so we loaded our datasets into HDFS by creating a directory named Flipkart shown in figure 5 and figure 6.

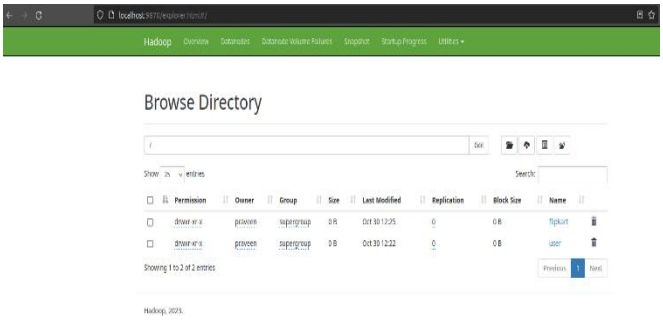

Figure 5: HDFS

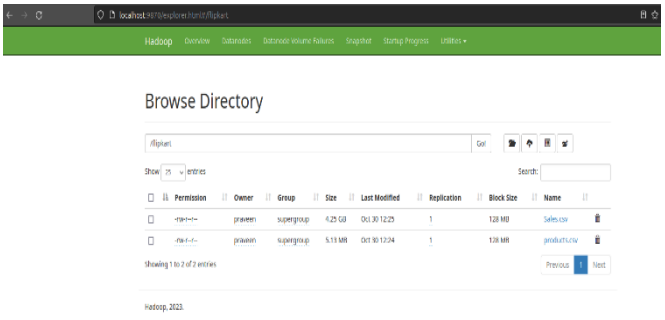

Figure 6: Flipkart Directory

7. DATA PRE-PROCESSING

In the dataset we identified there are some columns that are not required for our analysis, we also found that some rows had no meaning at all, and some other rows contained null values. Consequently, we cleaned and pre-processed the data such as removing the unwanted columns and a statistical metric such as the variable's mean has been used in place of missing values.

Figure 7 contains the preprocessed Sales dataset. Figure 8 contains the number of records in the Preprocessed sales dataset.

_c0	date_	city_name	order_id	cart_id	dim_customer_key	procured_quantity	unit_selling_price	total_discount_amount	product_id	total_weighted_landing_price
0	2022-04-01	Mumbai	112246974	173273882	17995199	1	234.0	0.0	344107	202.51303
1	2022-04-01	Bengaluru	112246976	173273597	18259433	1	64.0	0.0	389676	48.714375
2	2022-04-01	Bengaluru	112247019	173123717	5402601	1	1031.0	0.0	39411	975.996
3	2022-04-01	HR-NCR	112247045	173547459	15649744	1	57.0	0.0	369742	25.0
4	2022-04-01	Mumbai	112247123	173081820	10127605	2	30.0	0.0	12872	57.980004
5	2022-04-01	HR-NCR	112247149	173274225	479129	1	65.0	0.0	99407	61.5
6	2022-04-01	Bengaluru	112247170	173273989	8658748	1	115.0	0.0	439935	103.059875
7	2022-04-01	Mumbai	112247182	173273662	16541671	1	20.0	0.0	424394	15.399999
8	2022-04-01	Delhi	112247233	173269263	21466	1	24.0	0.0	470636	18.0
9	2022-04-01	Delhi	112247239	173274446	13512664	1	36.0	0.0	311	32.38837
10	2022-04-01	Delhi	112247317	173274278	13497358	1	20.0	0.0	339152	12.0
11	2022-04-01	HR-NCR	112247321	173273659	4486518	1	50.0	0.0	18546	46.477013
12	2022-04-01	Delhi	112247317	173274278	13497358	1	23.0	0.0	4927	14.0
13	2022-04-01	Bengaluru	112247328	173269880	18049405	2	19.0	0.0	37083	36.1
14	2022-04-01	Delhi	112247393	173273514	5933479	1	130.0	0.0	366189	113.04337
15	2022-04-01	Bengaluru	112247425	173273350	17213227	1	13.0	0.0	3928	8.0
16	2022-04-01	HR-NCR	112247431	173274695	1778240	1	35.0	0.0	424982	29.799997
17	2022-04-01	Delhi	112247443	169126739	7258345	1	76.0	0.0	432764	85.41983
18	2022-04-01	HR-NCR	112247447	172994313	4246317	2	91.0	0.0	13	150.02224
19	2022-04-01	HR-NCR	112247464	173273818	13982586	1	16.0	0.0	3913	10.0

Figure 7: Preprocessed Sales Dataset

```
Number of rows in the sales: 46627032
```

Figure 8: Sales Records after Preprocessing

Figure 9 contains the preprocessed Products dataset. Figure 10 contains the number of records in the Preprocessed Products dataset.

_c0	product_id	product_name	unit	product_type	brand_name	manufacturer_name	l0_category	l1_category	l0_category_id	l1_category_id
1	483436	Plum BodyLovin' H...	20 ml	Sample	Plum BodyLovin'	Pureplay Skin Sci...	Specials	Free Store	343	1493
3	483438	Plum BodyLovin' T...	20 ml	Sample	Plum BodyLovin'	Pureplay Skin Sci...	Specials	Free Store	343	1493
4	480473	Flipkart Valentin...	1 unit	Card	Flipkart	Dummy Manufacturer	Specials	Bill Breaker	343	1741
5	483694	Dabur Vita Chocol...	75 g	Sample	Dabur	Dabur India Pvt Ltd	Specials	Free Store	343	1493
6	486016	Plum Green Tea Po...	15 ml	Sample	Plum	Pureplay Skin Sci...	Specials	Free Store	343	1493
7	486017	Plum Green Tea Oi...	15 ml	Sample	Plum	Pureplay Skin Sci...	Specials	Free Store	343	1493
8	486124	Kari kari Salt & ...	22 g	Sample	Kari kari	LT Foods	Specials	Freebie	343	1013
9	486125	Maggi Liquid Coco...	180 ml	Sample	Maggi	Nestle India	Specials	Freebie	343	1013
10	486376	Orion Mango Choco...	28 g	Sample	Orion	ORION FOOD VINA C...	Specials	Freebie	343	1013
11	486377	Tata Tea Gold Car...	25 g	Sample	Tata Tea	Tata consumer pro...	Specials	Freebie	343	1013
12	486559	Conscious Food Or...	50 g	Sample	Conscious Food	Conscious Food	Specials	Freebie	343	1013
13	487659	The Laughing Cow ...	15 g	Sample	The Laughing Cow	Bel Vietnam Co., ...	Specials	Freebie	343	1013
14	487667	Sunfeast Dark Fan...	20 g	Sample	Sunfeast Dark Fan...	ITC Limited	Specials	Freebie	343	1013
15	487668	Sunfeast All Roun...	28.2 g	Sample	Sunfeast	ITC Limited	Specials	Freebie	343	1013
16	488061	Sleepy Owl Cold B...	5 x 50 g	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013
17	488062	Sleepy Owl Cold B...	5 x 50 g	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013
18	488063	Sleepy Owl Cold B...	5 x 50 g	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013
19	488059	Sleepy Owl Cinnam...	1 pack (5 x 50 g)	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013
20	488058	Sleepy Owl Carame...	5 x 50 g	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013
21	488060	Sleepy Owl Cinnam...	3 x 50 g	Sample	Sleepy Owl	Sleepy Owl Coffee...	Specials	Freebie	343	1013

Figure 9: Preprocessed Products Dataset

```
Number of rows in the DataFrame: 29030
```

Figure 10: Products Records after Preprocessing

8. Data Analysis

We started a spark session which connects to HDFS and reads data from the data frame which we loaded in the Flipkart directory in HDFS. Using this data frame, we started performing analysis.

Figure 11 shows the analysis of orders placed in each city. where we plotted the graph for it.

Figure 11: Count of orders recorded in each city.

Figure 12 shows the analysis of product categories where we plotted the graph of the top 10 product categories which recorded more orders.

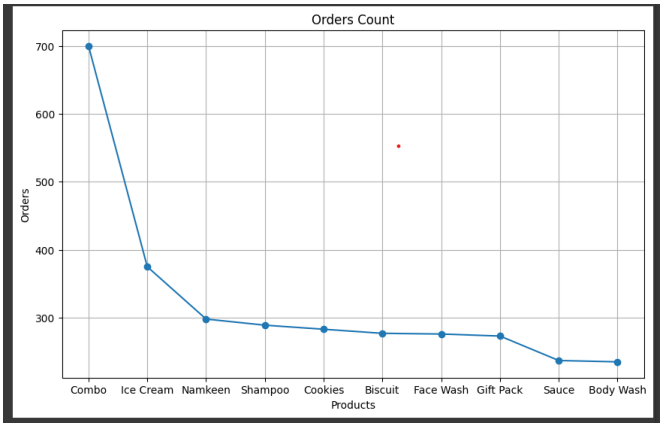

Figure 12: Top 10 Product Categories which recorded more orders.

These are some of our preliminary analyses. With these preliminary analyses we started analyzing sales based on multi categories like date-city, city-month, and many more.

In our data analysis, at first, we started with the total sales happening in all the four cities namely Mumbai, Delhi, Bengaluru, and HR-NCR available in the data set for April, May, June and July months.

Figure 13: Sales recorded in each City in each Month.

From the sales bar graph for cities and months, we can obtain a few insights.

- Delhi has recorded a higher number of sales than the other three cities in all the four months.
- Sales are higher in April than in the other three months in all the cities.

This might be because of the start of summer holiday season in India. This observation helps the e-commerce companies to plan their sales well ahead and come up with attractive marketing strategies to attract more people and they can even stock up more things during this time so that they can maintain customer satisfaction well.

While coming to the analysis of average daily sales in all categories in each city.

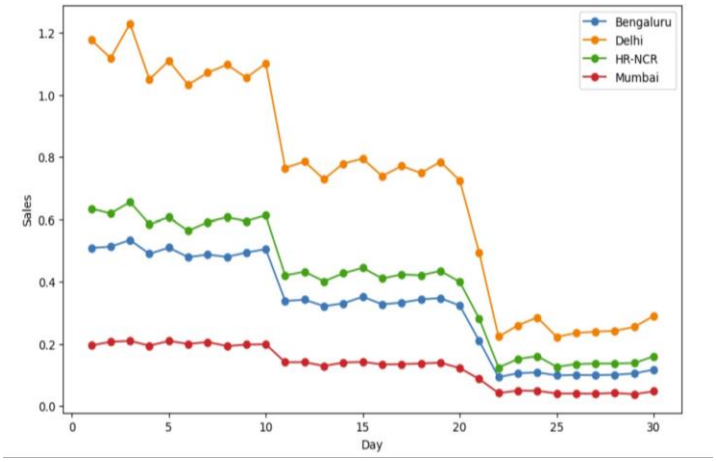

Figure 14: Daily sales in millions in each City.

Here we can observe an interesting pattern that maximum sales in all the cities are happening only in the first 10 days of the month and after that we can observe a steep decrease in the daily sales and towards the end of the month, there were little ups and downs in the pattern. This shows an interesting observation that people are more likely to purchase an item at the beginning of the month rather than during the middle of the month. This analysis will give an idea of the people's purchase patterns and they can increase the procured items at a particular city warehouse so that they can deliver products to the customers in a faster and secure manner.

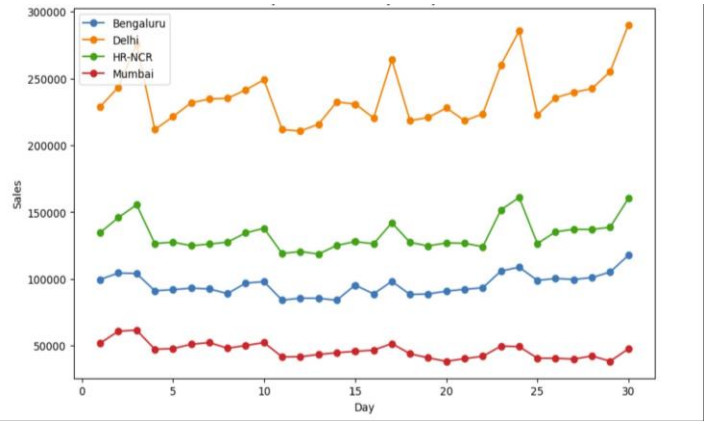

Figure 15: Daily sales in April Month in each City.

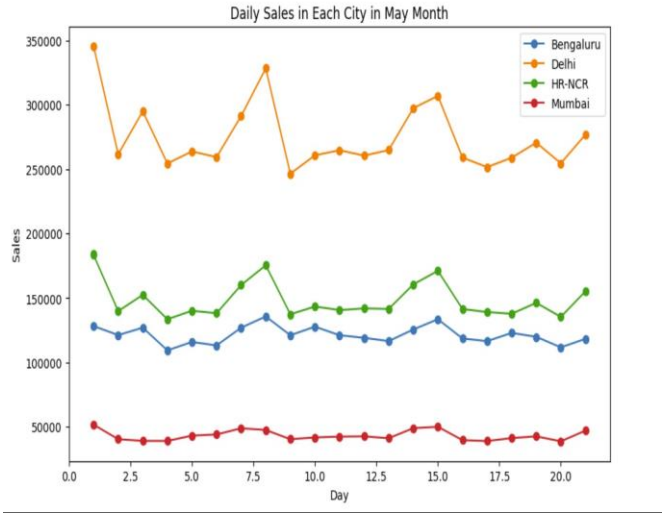

Figure 16: Daily sales in May month in each City.

When we break down the trends of the April and May month, In April month daily sales we can see contradicted patterns than what we have discussed above. The daily sales in all the cities have begun to rise more during mid-April than at the start of the month. From our observation, we found that it might be because of the attractive discounts e-commerce platforms have offered to make use of the holiday season. Coming to the May month, our starting trend continues to follow. It is a challenging task to describe people's purchase patterns without giving any particular details about discounted sales and the special holiday seasons.

We have obtained a list of several products ranging from grocery to cleaning essentials which were sold more in all the cities. From that list we have selected to see the trends of Onion daily sales in April month as this is the most used cooking product in India.

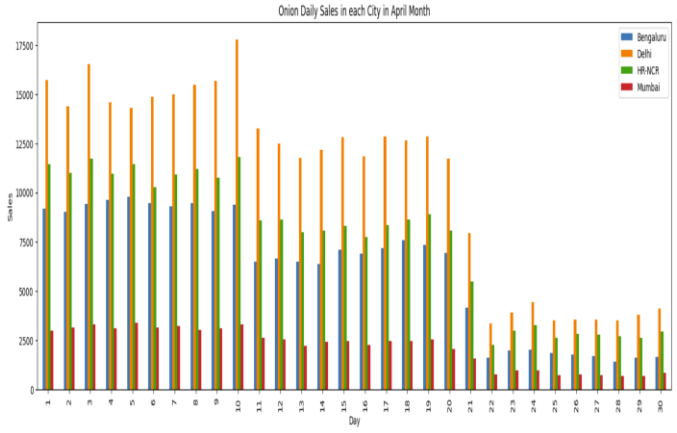

Figure 17: Onion April Daily sales in each City.

In Onion daily sales, Delhi recorded the most and Mumbai the least. Also, the sales were higher in the first 20 days of the month and later there was a gradual decrease in the sales towards the month-end. This analysis helps the business vendors to procure more stock during the initial days of the month so that the fresh products will attract customers more and which in turn maximize the profits.

The top 10 brands sold across all the cities are GHH, GMC, Aplus, GHD, Amul, Town Grocer, Haldiram's, Apple, SaveMore, Grocery.

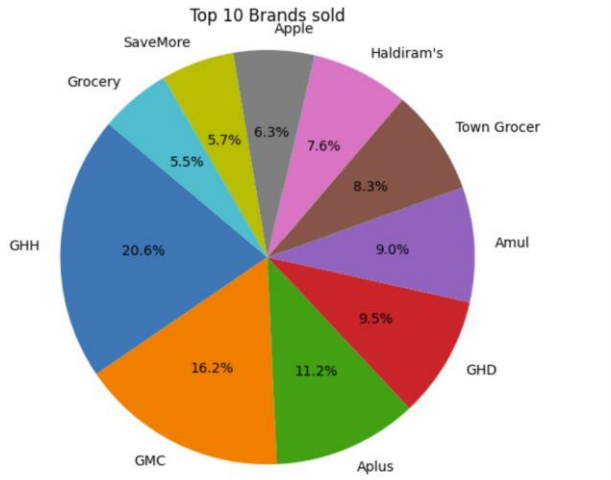

Figure 18: Top 10 Brands Sold.

Most of the brands from the list are related to household items which are used in everyday cooking. This pie-chart also tells how people are showing interest in purchasing products for their daily use from e-commerce platforms.

Post COVID-19 season, we identified that people are keener in maintaining personal hygiene and household hygiene. So we tried to analyze how the sales of cleaning essentials products are happening.

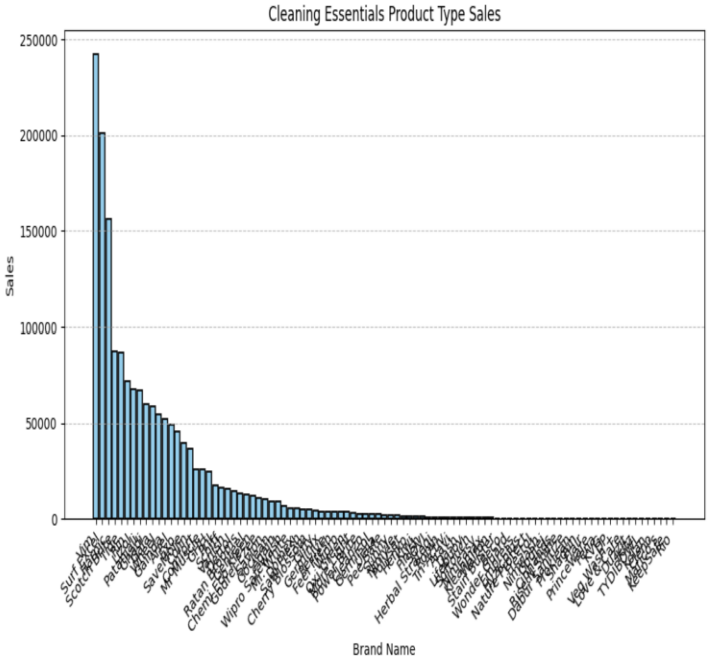

Figure 19: Cleaning Essentials Brands sales.

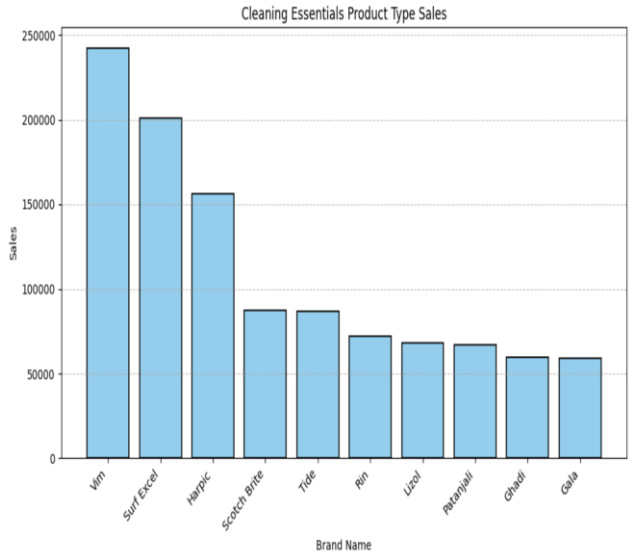

Figure 20: Top 10 Cleaning Essentials Brand sales.

We obtained a clumsy graph as there are numerous brands, so we cut down our analysis into top 10 cleaning essential product brands.

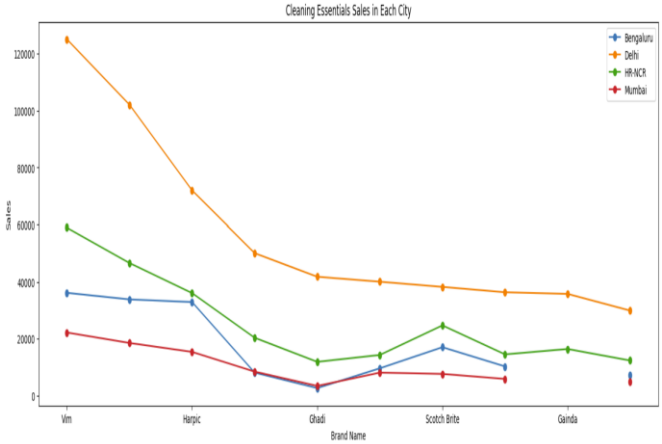

Figure 21: Cleaning Essentials Brand sales in each City.

Even in the sales of cleaning essentials, Delhi has recorded more and Mumbai the least.

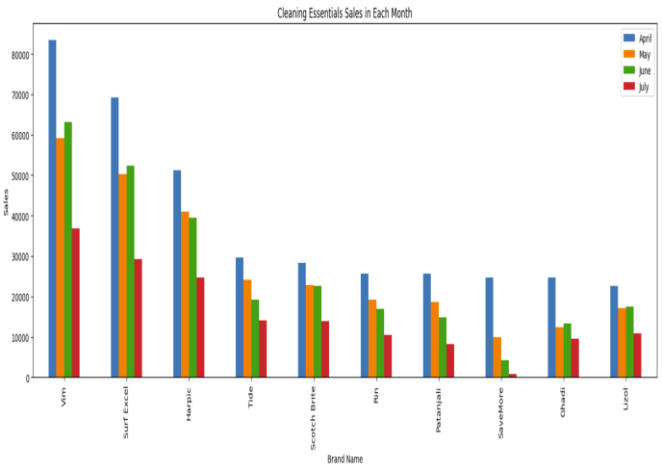

Figure 22: Cleaning Essentials Brand sales in each Month.

Sales in April are more when compared to the other months. This also might have an influence on the holiday which we have discussed earlier.

9. CONCLUSION

The report represents a comprehensive approach of E-commerce trends data analysis and trends using big data tools, The proposed methodology was applied to the transaction data acquired from Kaggle, which included approximately 467,06,387 records of users and the items they purchased, viewed with their respective location.

Initially we performed data cleaning and then pre-processing it using spark. We have come up with some outcomes like the total sales occurred in all the cities with respective consecutive months. We also come across the analysis of average daily sales of all kinds of products in each city, while we break down the trends of sales of products during each month and get to know that maximum sales are going to happen in the first ten days and the top 10 types of products which are going to sell more. Effective dynamic charts and graphs were made to show the outcomes, which helped other users understand them clearly about all the results and findings.

Based on our research, the recommended method provides a solid and scalable means for businesses to enhance their business and optimize their operations by gaining insights from e-commerce data. This can lead to better business outcomes. Businesses can use this approach to analyze e-commerce data and gain valuable insights into consumer behavior, market dynamics, and trends by harnessing the power of big data. By personalizing the customer experience, companies can increase customer satisfaction and loyalty with dynamic pricing strategies, customized recommendations, and focused advertising. When companies have a better grasp of their customers and the market than their rivals, they may develop more effective marketing campaigns and product offers.

10. AUTHOR CONTRIBUTIONS

Our project team is Praveen Kumar Reddy Kadapala, Manoj Suggala, Sreevardhan Reddy Soma, Rahul Karthik Arunachalam Usharani. The team first got together to brainstorm project ideas. Later, after deliberating over the professor's suggestions in class, we decided to investigate online competitions to find engaging tasks. We researched for more than a week and came up with our e-commerce project.

Praveen Kumar Reddy: Data set has been obtained from multiple locations, analyzed, and pre-processed it. Done some analysis on data. Outlined the initial concept and objectives of the project. Defined the scope and key goals for the analysis of e-commerce trends. Implemented data cleaning and preprocessing techniques to ensure data quality. Addressed missing values and outliers for a comprehensive dataset.

Manoj: Created the research plan and came up with the paper idea. Completed the comparison analysis and literature review. Examined and found appropriate datasets for trends and activity

in e-commerce. Managed the integration of diverse data sources into the big data tools. Designed and developed visualizations for presenting trends to stakeholders.

Sreevardhan Reddy: Helped in designing the research plan. Provided feedback on case studies and literature studies. Using Spark, written logic and feature evaluation code. Researched and selected appropriate big data algorithms for trend analysis. Collaborated with the team to choose the best-fit algorithms for specific tasks. Also contributed in performing deeper analysis of the results.

Rahul Karthik: Is responsible for composing sections of the paper on the analysis and contributing to the comparative analysis and literature review. Conducted rigorous evaluations to ensure the accuracy and reliability of the analysis. Created comprehensive reports summarizing key findings and actionable insights.

11. REFERENCES

[1]Khadija Aziz, Dounia Zaidouni, and Mostafa Bellafkih. 2018. Real-time data analysis using Spark and Hadoop. IEEE Xplore, 1–6. DOI:https://doi.org/10.1109/ICOA.2018.8370593

[2]Zhanchi Dong. 2022. Research of Big Data Information Mining and Analysis : Technology Based on Hadoop Technology. IEEE Xplore, 173–176. DOI:https://doi.org/10.1109/BDICN55575.2022.00041

[3]Penglin Gao, Zhaoming Han, and Fucheng Wan. 2020. Big Data Processing and Application Research. International Conference on Artificial Intelligence (October 2020). DOI:https://doi.org/10.1109/aiam50918.2020.00031

[4]Hong Li. 2021. Research on Big Data Analysis Data Acquisition and Data Analysis. IEEE Xplore, 162–165. DOI:https://doi.org/10.1109/CAIBDA53561.2021.00041

[5]Piyush Sewal and Hari Singh. 2021. A Critical Analysis of Apache Hadoop and Spark for Big Data Processing. IEEE Xplore, 308–313. DOI:https://doi.org/10.1109/ISPCC53510.2021.9609518

[6]Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan, and Irum Nahvi. 2019. Apache Spark: A Big Data Processing Engine. 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM) (November 2019). DOI:https://doi.org/10.1109/menacomm46666.2019.8988541