**ABSTRACT**

Data analysis helps data analysts to pull out meaningful results from the raw data and help their clients/users to make important business decisions by identifying various findings and patterns. It can be achieved by preparing reports, supporting analysis and presenting to management.

When it comes to processing and analyzing vast amounts of data to make sense of it, information technology in the twenty-first century is reaching new heights. Retailers now require a 360-degree perspective of their customers because, without one, they risk missing a competitive market advantage. To achieve their sales and marketing objectives, retailers must develop strong promotions and offers; otherwise, they risk losing sight of the significant potential that the current market presents. Retailers frequently find it difficult to understand the state of the market because their locations vary geographically. These retail companies may utilize data from previous years to more accurately forecast and predict sales for the upcoming year thanks to big data applications.

This research about Global Superstore data analysis will help us identify the various impacts of different aspects on the sales and profit of different products by identifying and highlighting the relationships. Further, the investigation examined the impact of different scenarios in different regions in order to better understand customer behavior trends, identify important drivers of sales performance and provide in-depth hidden insights. For visualizing the data, we have used IBM Cognos Analytics tool and then deployed ML algos with the help of Python Colab notebook.

**TABLE OF CONTENTS**

| **CHP. NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
|  | **ABSTRACT** | **7** |
|  | **LIST OF FIGURES** | **10-12** |
|  | **LIST OF TABLES** | **13** |
|  | **LIST OF ABBREVIATIONS** | **14** |
| **1** | **INTRODUCTION**  **1.1 General Introduction**  **1.2 Tools and Technology Used** | **15-16**  **15-16**  **16** |
| **2** | **LITERATURE SURVEY**  **2.1 Summary of research papers studied**  **2.2 Problem Statement** | **17-22**  **17-21**  **21-22** |
| **3** | **REQUIREMENT ANALYSIS**  **3.1 Functional Requirements**  **3.2 Non-Functional Requirements** | **23-26**  **23-24**  **25-26** |
| **4** | **REPORT ON PRESENT INVESTIGATION**  **4.1 Ideation and Proposed Solution**  **4.2 Description of Project**  **4.3 Algorithms Used** | **27-31**  **27-28**  **29-30**  **30-31** |
| **5** | **RESULTS AND DISCUSSION**  **5.1 Implementation Details and Issues**  **5.1.1 Implementation Details**  **5.1.2 Implementation Issues**  **5.2 Evaluation Parameters**  **5.3 Observations and Results** | **32-58**  **32-28**  **32-37**  **38**  **38**  **39-58** |
| **6**  **7** | **CONCLUSION AND FUTURE WORK**  **6.1 Conclusion**  **6.2 Future Work**  **REFERENCES** | **59-60**  **59**  **59-60**  **61-63** |

**LIST OF FIGURES**

| **Fig. No.** | **Name of Figure** | **Page No.** |
| --- | --- | --- |
| 4.2.1 | Block diagram showing the lifecycle of project | 29 |
| 5.1.1.1 | 1. CSV File Responses 2. CSV File Responses   (c) CSV File Responses  (d) CSV File Responses | 32  32  33  33 |
| 5.1.1.2 | 1. Updated CSV File Responses 2. Updated CSV File Responses 3. Dataframe uploaded in Project 4. Final Dataframe used in Project 5. Final Dataframe used in Project | 33  34  34  34  35 |
| 5.1.1.3 | (a) Describing the dataset (Statistics)  (b) Describing the dataset (Statistics) | 35  36 |
| 5.1.1.4 | 1. Correlation Table 2. Correlation Table | 36  37 |
| 5.1.1.5 | Data Relationship between the columns of dataset | 37 |
| 5.3.1 | Profit and Sales for Country regions | 39 |
| 5.3.2 | Discount representation points over the World | 39 |
| 5.3.3 | Country by Profit sized by Sales | 40 |
| 5.3.4 | Sales by Region | 40 |
| 5.3.5 | Profit by Segment | 41 |
| 5.3.6 | Profit by Region | 41 |
| 5.3.7 | Sales for Sub-Category hierarchy | 42 |
| 5.3.8 | Profit, Sales and Quantity by Segment and Category | 42 |
| 5.3.9 | Sales and Profit by Order Date | 43 |
| 5.3.10 | Region-wise Quantity and Sales | 43 |
| 5.3.11 | Region-wise Quantity to Sales Ratio | 44 |
| 5.3.12 | Relationship between Shipping Cost and Sub-Category | 44 |
| 5.3.13 | Month-wise Sales in India | 45 |
| 5.3.14 | Profit (Decision Tree) | 45 |
| 5.3.15 | Region (Decision Tree) | 46 |
| 5.3.16 | (a) Profit and Sales by Country colored by Order Priority for the months of January - April  (b) Profit and Sales by Country colored by Order Priority for the months of May - August  (c) Profit and Sales by Country colored by Order Priority for the months of September - December | 46  47  47 |
| 5.3.17 | Discount for Category | 48 |
| 5.3.18 | Discount for Sub-Category | 48 |
| 5.3.19 | (a) Different segments of data based on Order Priority of Haryana  (b) Different segments of data based on Order Priority of Haryana  (c) Different segments of data based on Order Priority of Haryana | 49  49  50 |
| 5.3.20 | Year-wise Profit Rate on Sales | 50 |
| 5.3.21 | Sales based on Order Priority | 51 |
| 5.3.22 | Sales by Discount for Profit | 51 |
| 5.3.23 | Sales by Discount for Profit for Furniture Category | 52 |
| 5.3.24 | Profit ordered by Ship Month | 52 |
| 5.3.25 | Region-wise Quantity segmented by Segment | 53 |
| 5.3.26 | Country-wise measure of Sales and Discount | 53 |
| 5.3.27 | Profit and Sales by Segment and Category | 54 |
| 5.3.28 | Sub-Category hierarchy colored by Profit and sized by Sales | 54 |
| 5.3.29 | Correlation heat map between all numerical fields of dataframe (in python) | 55 |
| 5.3.30 | Profit by Order Priority colored by Ship Mode | 56 |
| 5.3.31 | Country-wise Quantity colored by Segment | 56 |
| 5.3.32 | Profit by Country colored by Segment | 57 |
| 5.3.33 | Plot representing accuracy of different Classification models (in python) | 58 |

**LIST OF TABLES**

| **Table No.** | **Name of Table** | **Page No.** |
| --- | --- | --- |
| 2.1 | Summary of the research papers studied | 17-21 |

**LIST OF ABBREVIATIONS**

| * Algos | * Algorithms |
| --- | --- |
| * IBM | * International Business Machines |
| * Colab | * Google Colaboratory |
| * CSV | * Comma Separated Values |
| * DL | * Deep Learning |
| * i.e. | * That is |
| * AI | * Artificial Intelligence |
| * ML | * Machine Learning |
|  |  |
|  |  |

**CHAPTER 1**

**INTRODUCTION**

**1.1 General Introduction**

The Global Superstore dataset serves as the foundation for this comprehensive analysis, aiming to uncover valuable insights and trends within the retail industry. This dataset comprises a wide range of attributes related to customer orders, shipping details, product information, and sales performance. By exploring the vast array of data available, we seek to gain a deeper understanding of the market dynamics, customer behavior, and factors that impact profitability.

In today's competitive business landscape, understanding customer preferences and optimizing operational efficiency are crucial for sustained success. The Global Superstore dataset provides a rich and diverse collection of information, encompassing various geographical regions, product categories, and market segments. By examining this dataset, we can uncover hidden patterns, identify areas of improvement, and make informed decisions to drive business growth.

The dataset includes attributes such as order ID, order date, ship date, ship mode, customer ID, customer name, segment, city, state, country, postal code, market, region, product ID, category, subcategory, product name, sales, quantity, discount, profit, shipping cost and order priority. Each attribute offers unique insights into different aspects of the business, enabling us to analyze the relationships between various variables and derive meaningful conclusions.

Throughout this report, we had conducted an in-depth exploration of the dataset, employing various analytical techniques to extract valuable insights. By examining the relationships between different attributes, we aim to identify key factors that contribute to sales performance, uncover trends in customer behavior, and highlight potential areas for improvement in operational efficiency.

Ultimately, the findings of this analysis will provide actionable recommendations and insights to enhance decision-making processes within the retail industry and we can optimize resource allocation, refine marketing strategies, and improve customer satisfaction, thereby driving long-term success and growth in the ever-evolving global market.

**1.2** **Tools and Technology Used**

* **Python**

Python is a high-level and general-purpose programming language. Python is dynamically-memory allocated. It supports multiple programming patterns, including structured , object-oriented and functional programming. Python is often seen as a "cell included" language due to its huge standard library [[1]](#1fob9te).

* **Machine learning**

Machine learning is a sub-part of AI that gives systems the capability to learn on their own and improve from experiences without the interference of humans. Machine learning focuses on the development of computer algorithms and accessing data and performing various functions [2].

* **IBM Cognos Analytics**

The web-based integrated business intelligence suite from IBM is called Cognos Business Intelligence. It offers a toolkit for analytics, scorecarding, reporting, and keeping track of events and data. The software is made up of a number of parts that are intended to satisfy the various information needs of a business [3].

* **Google Colab Notebook**

Colab is a product from Google Research. It performs the same as a jupyter notebook on the desktop [[4]](#2et92p0).

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 Summary of the research papers studied**

**Table 2.1 Summary of Literature Survey**

| **S.No** | **Name of Paper** | **Authors** | **Summary** | **Tools / Technologies** |
| --- | --- | --- | --- | --- |
| 1 | Big Data and Analytics in Retailing [5] | Venky [Shank](https://www.researchgate.net/scientific-contributions/Venky-Shankar-2163866383)ar | The importance of big data analytics in the retail industry, focusing on its role in shaping marketing strategies, understanding customer preferences, and making informed business decisions. | Big Data Analytics and ML |
| 2 | Retail Analytics: Driving Success in Retail Industry with Business Analytics [6] | Sudeep B. Chandramana | Big data in retail revolutionizes decision-making, enhances profitability, enables real-time insights, addresses data privacy challenges, leverages omnichannel behavior, and drives an AI-driven future for improved customer experiences, operations, and profitability. | Big Data Analytics, ML and AI |
| 3 | Retailing and retailing research in the age of big data analytics [7] | Marnik G. Dekimpe | This paper describes the allure of the retail sector as a research domain, emphasizing its size, multifaceted and dynamic nature, the availability of high-quality data, and coverage by business analysts. It also discusses the potential of big data analytics in the retail industry, including its impact on retail managers, retailing researchers, public policy makers, investors, and retailing educators. | Business Analytics |
| 4 | Big Data Analytics: A Literature Review Paper [8] | Elgendy, Nada & Elragal, Ahmed | This paper discusses that big data analytics allows organizations to analyze large volumes of data from various sources and uncover patterns, trends, and customer behaviors. Applications of big data analytics include customer intelligence, supply chain management, performance and quality management, risk mitigation, and fraud detection. It offers opportunities for organizations to enhance their decision-making processes, improve efficiency, and gain a competitive edge in various industries. | Big Data and its Analysis |
| 5 | Data Science - Cosmic Infoset Mining, Modeling and Visualization [[9]](#2s8eyo1) | Mr. Subhashish Kumar, Dr. Namarata Dhanda and Mr. Ashutosh Pandey | This research paper found that they were working on a statistical model.They presented the dataset mining, modeled and visualized marketing big data using user friendly open source python library.They experimented on the real time data sheet that focused on predicting job and incidentals that was needed by the organization. | Machine Learning using Python (Statistical Model) |
| 6 | Emergence of Data Analytics in the Information Systems Curriculum [10] | Musa J. Jafar,Jeffry Babb,Amjda Abdullat | It discusses the importance of integrating data analytics into the curriculum of Information Systems education. It explores the challenges and opportunities presented by data analytics, emphasizing the need for interdisciplinary collaboration and diverse delivery approaches in educational institutions. | Data Analytics |
| 7 | Walmart's Sales Data Analysis - A Big Data Analytics Perspective [11] | Singh, Manpreet & Ghutla, Bhawick & Jnr, Reuben & Mohammed, Aesaan & Rashid, Mahmood | The purpose is to learn more about consumer behavior, comprehend the variables that influence sales, and forecast future sales. By analyzing, we can improve processes, distribute resources wisely, and increase revenue. | Big Data  and  Analytics |
| 8 | Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities [12] | Shahbaz, Muhammad & Gao, Changyuan & Zhai, Lili & Shahzad, Fakhar & Luqman, Adeel & Zahid, Rimsha | This paper suggested the impact of big data analytics (BDA) on sales performance and customer relationship management (CRM) capabilities. The findings suggest that individual and organizational characteristics influence the salesforce's perception of BDA, leading to improved person-technology fit, CRM capabilities, and sales performance. | Big Data Analytics |
| 9 | Visual Exploratory Data Analysis of COVID-19 Pandemic [[13]](#35nkun2) | Sumindar Kaur Saini,  Vishal Dhull, Sarbjeet Singh, Akashdeep Sharma | The dataset used was from 22nd January 2020 to 12th June 2020. The results of the graphical k-means clustering algorithm by dividing the 200+ countries in the groups of 3 namely (0, 1, 2). In this analysis, timely patterns of the rise and fall of confirmed deaths and recovery cases had been presented visually wherein the top 15 countries for such types of cases have been highlighted. | k-means Clustering |

**2.2 Problem Statement**

**Inconsistent and Missing Data:** Identify and address inconsistencies and missing values across the dataset attributes to ensure data accuracy and reliability.

**Sales Performance and Profitability:** Analyze sales metrics, including sales, profit, and discount, to identify trends, patterns, and areas of improvement.

**Operational Efficiency:** Evaluate the order management and shipping processes to identify bottlenecks, delays, or inefficiencies that impact customer satisfaction and operational costs.

**Customer Satisfaction and Retention:** Analyze customer behavior, including purchase frequency, feedback, and complaints, to identify factors affecting customer satisfaction and loyalty.

**Market Analysis and Expansion:** Explore market dynamics and regional performance to identify untapped markets, growth opportunities, and market saturation.

**CHAPTER 3**

**REQUIREMENT ANALYSIS**

**3.1 Functional Requirements**

**Order Management:**

⦁ The system should allow the creation and storage of new orders with unique order IDs.

⦁ The system should capture order details, including order date, customer information, and product details.

⦁ The system should validate and enforce order priority based on predefined rules.

**Shipping Management:**

⦁ The system should track and update the shipping status of each order, including ship date and ship mode.

⦁ The system should calculate and record shipping costs for each order based on the selected shipping mode.

⦁ The system should provide notifications or alerts for delayed shipments beyond predefined thresholds.

**Customer Management:**

⦁ The system should maintain customer records, including customer ID, customer name, and contact information.

⦁ The system should enable searching and filtering of customer records based on various criteria, such as segment or region.

⦁ The system should support customer segmentation and categorization for targeted marketing strategies.

**Product Management:**

⦁ The system should store and manage product information, including product ID, category, subcategory, and product name.

⦁ The system should enable product search and filtering based on various attributes, such as category or subcategory.

⦁ The system should support product performance analysis, including sales, quantity, discount, and profit.

**Sales Analysis and Reporting:**

⦁ The system should provide analytics capabilities to analyze sales performance based on different dimensions, such as market, region, or product category.

⦁ The system should generate reports and visualizations summarizing sales metrics, trends, and patterns.

⦁ The system should support customized reports or ad-hoc queries for specific analysis requirements.

**Discount Management:**

⦁ The system should calculate and apply appropriate discounts based on predefined rules and criteria.

⦁ The system should validate discount eligibility based on customer segment, product category, or order amount.

⦁ The system should track and analyze the impact of discounts on sales, profitability, and customer behavior.

**Market Analysis:**

⦁ The system should enable market analysis by aggregating and analyzing sales data based on market attributes.

⦁ The system should provide insights into market trends, market share, and opportunities for expansion.

⦁ The system should support geographical analysis by analyzing sales performance across different cities, states, or countries.

**3.2 Non-Functional Requirements**

**Performance:**

⦁ The system should provide fast response times when querying or analyzing large volumes of data.

⦁ The system should be capable of handling concurrent user requests without significant performance degradation.

⦁ Data retrieval and processing should be optimized to ensure efficient and timely analysis.

**Scalability:**

⦁ The system should be scalable to accommodate increasing data volume, user load, and future business growth.

⦁ It should support the addition of new markets, regions, and product categories without compromising performance.

⦁ The system should allow for horizontal or vertical scalability to meet changing demands.

**Reliability:**

⦁ The system should ensure data integrity and accuracy, minimizing the occurrence of data errors or inconsistencies.

⦁ It should provide backup and recovery mechanisms to prevent data loss in the event of system failures or disruptions.

⦁ The system should have a high level of availability, minimizing downtime for maintenance or upgrades.

**Security:**

⦁ The system should enforce access control and user authentication to protect sensitive data and prevent unauthorized access.

⦁ It should encrypt data transmissions and ensure secure storage of customer information and other sensitive data.

⦁ The system should comply with relevant data protection regulations and industry best practices.

**Usability:**

⦁ The system should have an intuitive and user-friendly interface, allowing users to navigate and interact with the data easily.

⦁ It should provide clear and meaningful visualizations, reports, and dashboards for effective data analysis.

The system should support customization and personalization features

**Maintainability:**

⦁ The system should be modular and well-documented, facilitating ease of maintenance and future enhancements.

⦁ It should support efficient data updates and modifications without disrupting ongoing operations.

⦁ The system should be designed with a clear separation of concerns, allowing for easier troubleshooting and debugging.

**Compatibility:**

⦁ The system should be compatible with various operating systems, browsers, and devices to support diverse user environments.

⦁ It should integrate with other existing systems or third-party applications, such as CRM or ERP systems, for seamless data exchange.

⦁ The system should support data interoperability standards and formats for easy integration with external systems.

**CHAPTER 4**

**REPORT ON PRESENT INVESTIGATION**

**4.1 Ideation and Proposed Solution**

Based on the Global Superstore dataset, here are some ideation areas and proposed solutions that can be explored:

**Sales Optimization:**

⦁ Identify top-selling products or categories and develop targeted marketing strategies to further boost their sales.

⦁ Analyze customer segments and tailor promotions or discounts to specific groups to encourage repeat purchases.

⦁ Implement cross-selling or upselling techniques to increase average order value.

**Supply Chain Efficiency:**

⦁ Analyze shipping modes and identify the most cost-effective and timely options for different regions.

⦁ Improve delivery times and cut shipping costs by streamlining the order fulfilment process.

⦁ Implement inventory management techniques to minimize stockouts and excess inventory.

**Customer Satisfaction Enhancement:**

⦁ Analyze customer feedback and ratings to identify areas for improvement in product quality or customer service.

⦁ Develop personalized customer experiences by leveraging customer segmentation and purchase history.

⦁ Implement a customer loyalty program to incentivize repeat purchases and enhance customer retention.

**Geographical Expansion:**

⦁ Identify untapped markets or regions with growth potential based on sales performance analysis.

⦁ Develop market entry strategies for new regions, considering factors such as local preferences and competition.

⦁ Customize marketing campaigns and product offerings to cater to specific regional demands.

**Operational Cost Reduction:**

⦁ Analyze profit margins and identify areas where costs can be optimized.

⦁ Evaluate shipping costs and negotiate favorable contracts with logistics providers.

⦁ Streamline internal processes to reduce inefficiencies and operational expenses.

**Fraud Detection and Risk Management:**

⦁ Implement fraud detection algorithms to identify suspicious patterns or anomalies in customer orders.

⦁ Monitor and analyze indicators of potential fraudulent activities, such as high-value orders with unusual shipping details.

⦁ Develop risk management strategies to mitigate potential losses associated with fraudulent transactions.

**Sales Forecasting and Inventory Planning:**

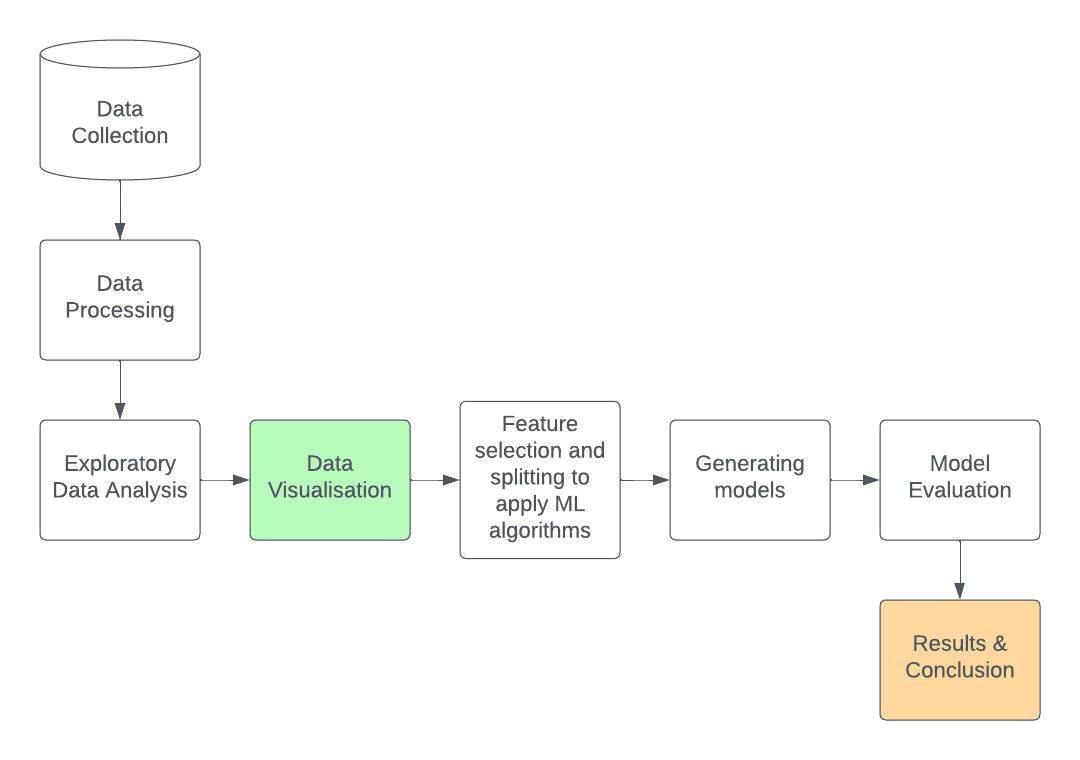
⦁ Utilize historical sales data and predictive analytics to forecast future sales.

⦁ Optimize inventory levels based on demand patterns to minimize stockouts and excess inventory costs.

⦁ Implement a dynamic inventory replenishment system to ensure optimal stock availability.

**4.2 Description of Project**

**Data Collection** : We searched and downloaded the dataset of Global Superstore which captured data from all over the world from the year 2016-2020 and the records were saved in the form of a csv file. The file consisted of 51290 rows and 23 columns namely, ‘Row ID', 'Order ID', 'Order Date', 'Ship Date',’Ship Mode', 'Customer ID', 'Customer Name', 'Segment', 'City', 'State', 'Country', 'Postal Code', 'Market', 'Region', 'Product ID', 'Category','Sub-Category', 'Product Name', 'Sales', 'Quantity', 'Discount', 'Profit', 'Shipping Cost' and 'Order Priority'. We extracted a new column named 'NumOfDays' from Ship Date and Order Date, which depicted the total number of days it took to get shipped and reach the destination from the date of order. So, our updated dataset has 24 columns and 51290 records.



**Fig 4.2.1 Block Diagram showing the Lifecycle of Project**

**Data Processing (in python):**

1. Importing python libraries
2. Loading the data (.csv) in dataframe
3. Dropping the irrelevant columns
4. Renaming the columns
5. Mapping the values using Label Encoding

**Exploratory Data Analysis and Data Visualisation:** We performed analysis through visualizations using IBM Cognos Analytics Tool.

**Feature Selection and Splitting to apply ML algorithms :** In this, we selected the and splitted the categorical and numerical features required for analysis.

**Generating Models:** In this, we applied different machine learning algorithms on the field - Order Priority (‘Critical’, ‘High’ , ‘Medium’ and ‘Low’) and generated the models for evaluation.

**Model Evaluation:** In this, we checked the performance of different models (including ensemble models)and which model has the highest accuracy.

**Results and Conclusion:** We have discussed the outcomes drawn from our analysis and concluded our project.

**4.3 Algorithms Used**

* **Random forest:** Random forest is a type of supervised ML algo. that is used in classification as well as regression problems. It works on making decision trees on different samples and takes majority votes for classification problems and average for regression problems [[14]](#44sinio).
* **K-nearest neighbour:** KNN or K-Nearest Neighbour, is a type of supervised machine learning algo. that can be used to solve both classification and regression problems. It is also known as “Lazy Learning”. On the basis of “K” , the new data point is given the position after calculating distance to the nearest neighbourhood. It is used for non linear data [[15]](#2jxsxqh).
* **Gradient boosting:** It is one of the most powerful boosting algo. of ML. It works on the principle of comparing previous models and giving the best possible next model to minimise errors [[16].](#z337ya)
* **Decision tree:** It is a supervised learning algo. that can be used for both classification and regression problems, but mostly preferred for classification problems. It makes a tree-based structure based on splitting of data in two or more homogeneous sets/groups based on the most significant splitter input variable [[17].](#3j2qqm3)
* **SVM:** Support Vector Machine or SVM is a classification algorithm (in most cases) which works on a mathematical technique called kernel functions (these allow to map data into higher dimensional space) and then find a hyperplane that separates the two classes of data. This hyperplane is called a decision boundary [[18].](#1y810tw)
* **Logistic regression:** This is a process of fitting data into a logit function and predicting probability of an event. The most common logistic regression model gives a binary output. Examples- Yes/No, 0/1 [[19].](#4i7ojhp)
* **CATBoost:** This is a variant of gradient boosting that can handle both categorical and numerical features. It does not require any feature encodings techniques like One-Hot Encoder or Label Encoder to convert categorical features into numerical features [20].
* **XGBoost:** This combines the predictions of multiple weak models to produce a stronger prediction. XGBoost can efficiently handle missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing [21].
* **Light GBM:** Thisextends the gradient boosting algorithm by adding a type of automatic feature selection as well as focusing on boosting examples with larger gradients. This can result in a dramatic speedup of training and improved predictive performance [22].

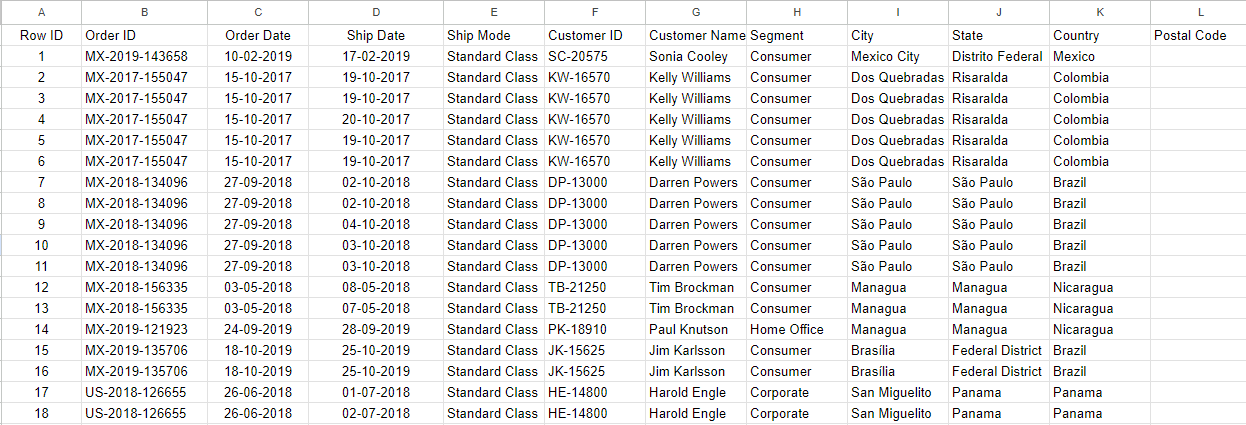
**CHAPTER 5**

**RESULTS AND DISCUSSIONS**

**5.1 Implementation Details and Issues**

**5.1.1 Implementation Details**

**Step -1 Gathering of data :** We gathered and stored the Global Superstore data in the form of .csv file format.Fig 5.1.1.1(a) to Fig 5.1.1.1(d) represents the first 18 responses stored in the original CSV , sorted by Row ID.

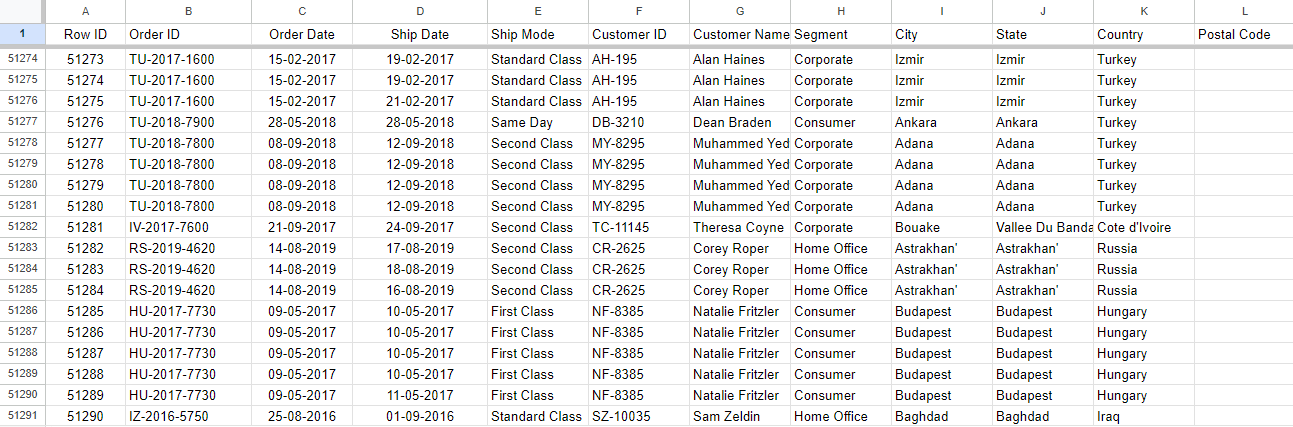
****

**Fig 5.1.1.1(a) CSV File Responses**

****

**Fig 5.1.1.1(b) CSV File Responses**

Fig 5.1.1.1(c) and Fig 5.1.1.1(d) represent the last 18 records of the responses collected.

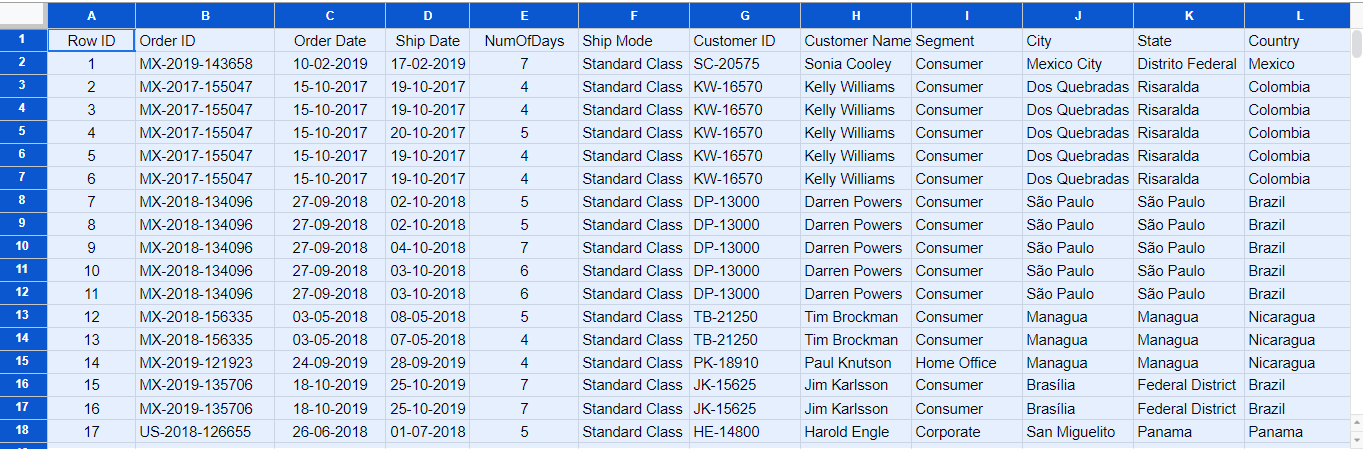


**Fig 5.1.1.1(c) CSV File Responses**

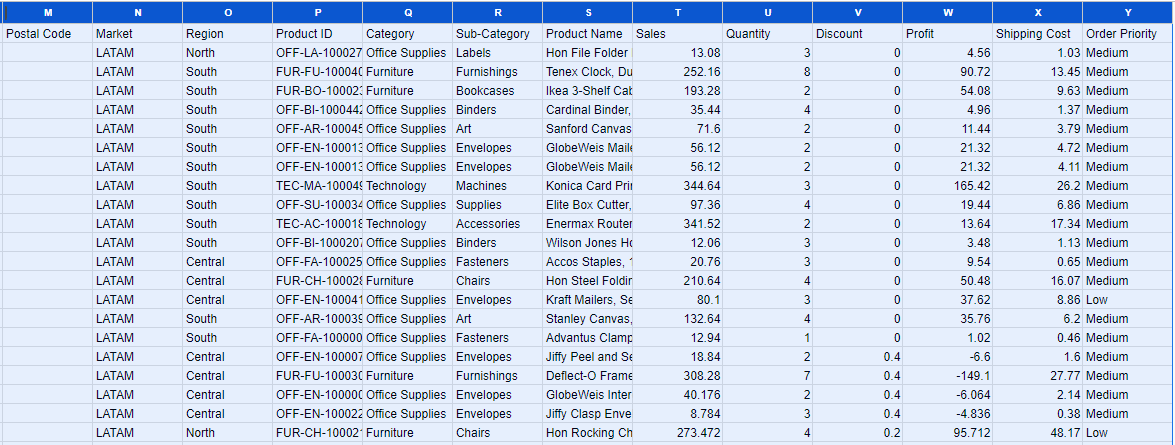
****

**Fig 5.1.1.1(d) CSV File Responses**

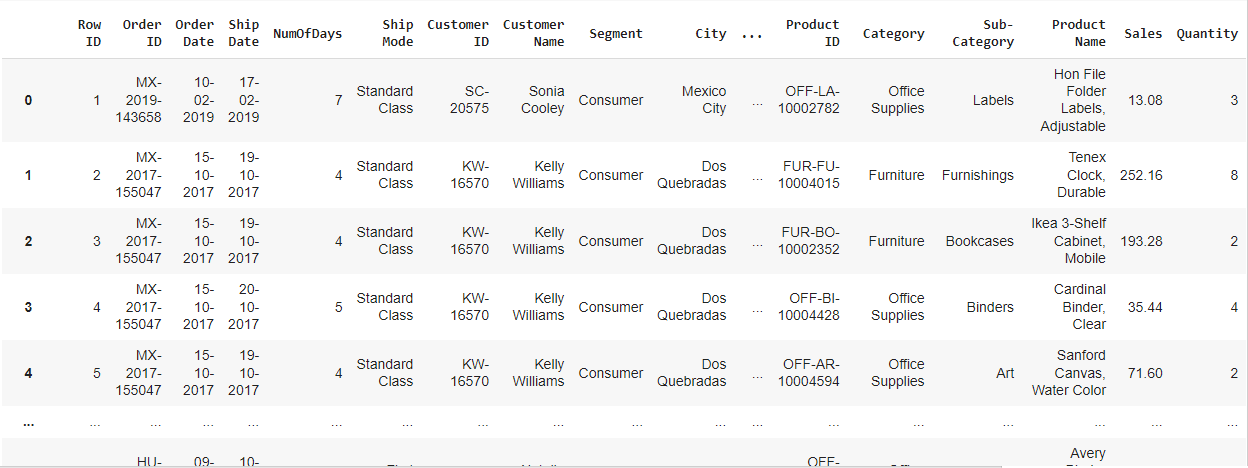
**Step -2 Processing of Data :** Fig 5.1.1.2(a) and Fig 5.1.1.2(b) represent the updated CSV dataset and Fig 5.1.1.2(c) represent that after preprocessing the data, we uploaded the file in a dataframe (python) and Fig 5.1.1.2(d) and Fig 5.1.1.2(e) represent dataframe after dropping the columns as per the requirement.

****

**Fig 5.1.1.2(a) Updated CSV File Responses**



**Fig 5.1.1.2(b) Updated CSV File Responses**

****

**Fig 5.1.1.2(c) Dataframe uploaded in project**

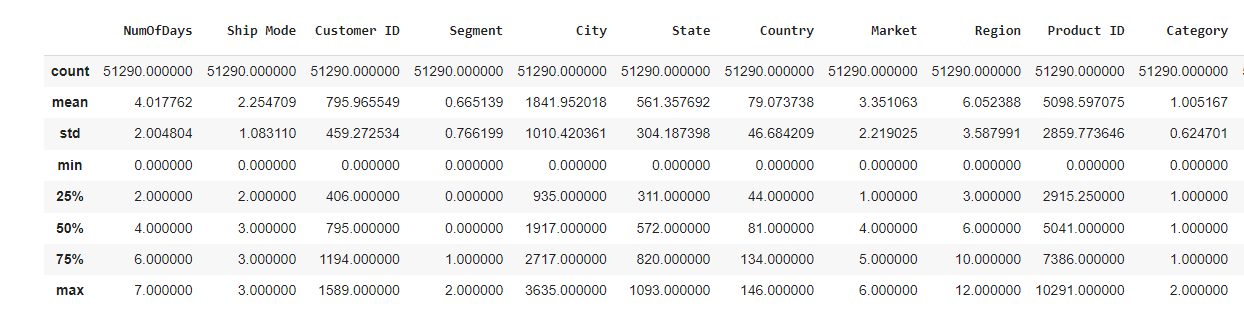
****

**Fig 5.1.1.2(d) Final Dataframe used in project**

****

**Fig 5.1.1.2(e) Final Dataframe used in project**

**Step -3 Describing the Dataframe :** After necessary replacement of data columns using Label Encoding, Fig 5.1.1.3(a) and Fig 5.1.1.3(b) depicts the measures of central tendency in statistics between the attributes. We observed that the mean number of days for the delivery of a product is 4 days, minimum discount is 0 and the maximum profit is around 8400 units. The mean quantity for products is 3.47 that is nearly 4. Mean number of sales is 246.49 or nearly 250 units.

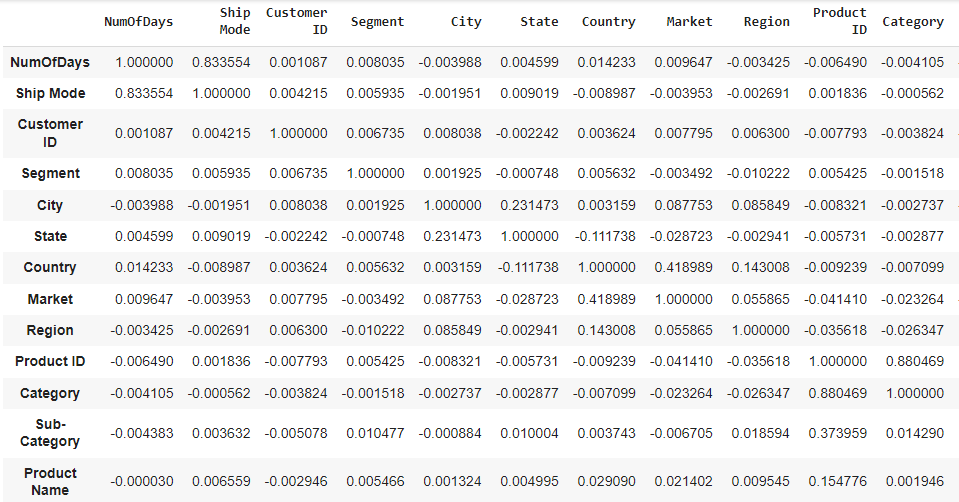


**Fig 5.1.1.3(a) Describing the dataframe (Statistics)**

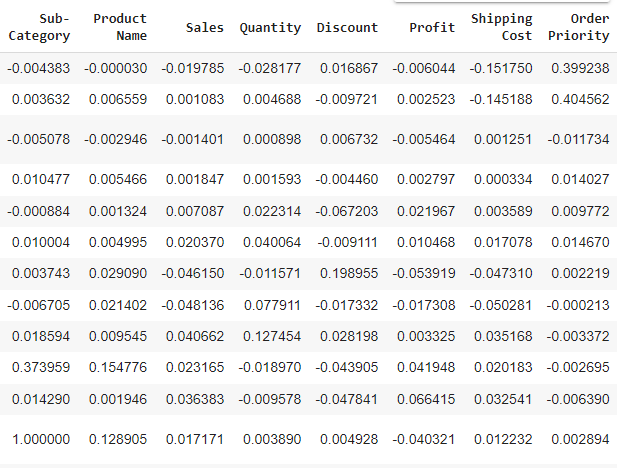
****

**Fig 5.1.1.3(b) Describing the dataframe (Statistics)**

**Step -4 Finding out Correlation of columns :** Fig 5.1.1.4(a) and Fig 5.1.1.4(b) depicts the first few records of the correlation table which has the values of coefficient of correlation (Karl Pearson’s method) and shows how the fields are correlated to each other. It is observed that there is a strong positive correlation between the Number of Days and Shipment Mode columns.

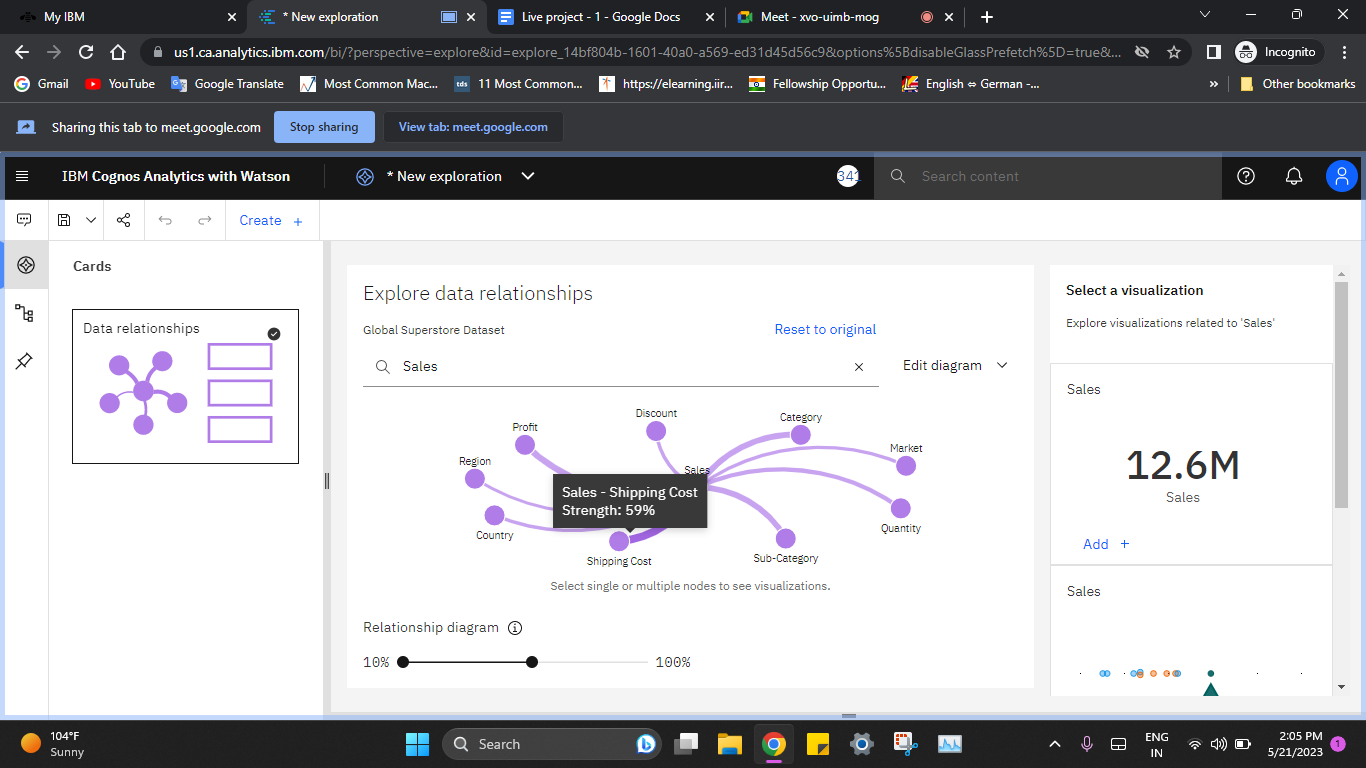


**Fig 5.1.1.4(a) Correlation table**



**Fig 5.1.1.4(b) Correlation table**

**Step - 5 Visualising data in IBM Cognos Analytics :** Fig 5.1.1.5 depicts the Cognos Data exploration platform.



**Fig 5.1.1.5 Data Relationship between the columns of dataset**

**Step - 6 Generating Models and Evaluation :** After analysing graphically on IBM Cognos Analytics, we used some classification models in python and predicted “Order Priority” of the orders. We then evaluated the models and finally came up with the conclusion that our “LightGBM” model has the highest accuracy.

**5.1.2 Implementation Issues**

While working with the Global Superstore dataset, there are several implementation issues that need to be considered. These issues may impact the practical application and utilization of the dataset for analysis and decision-making.

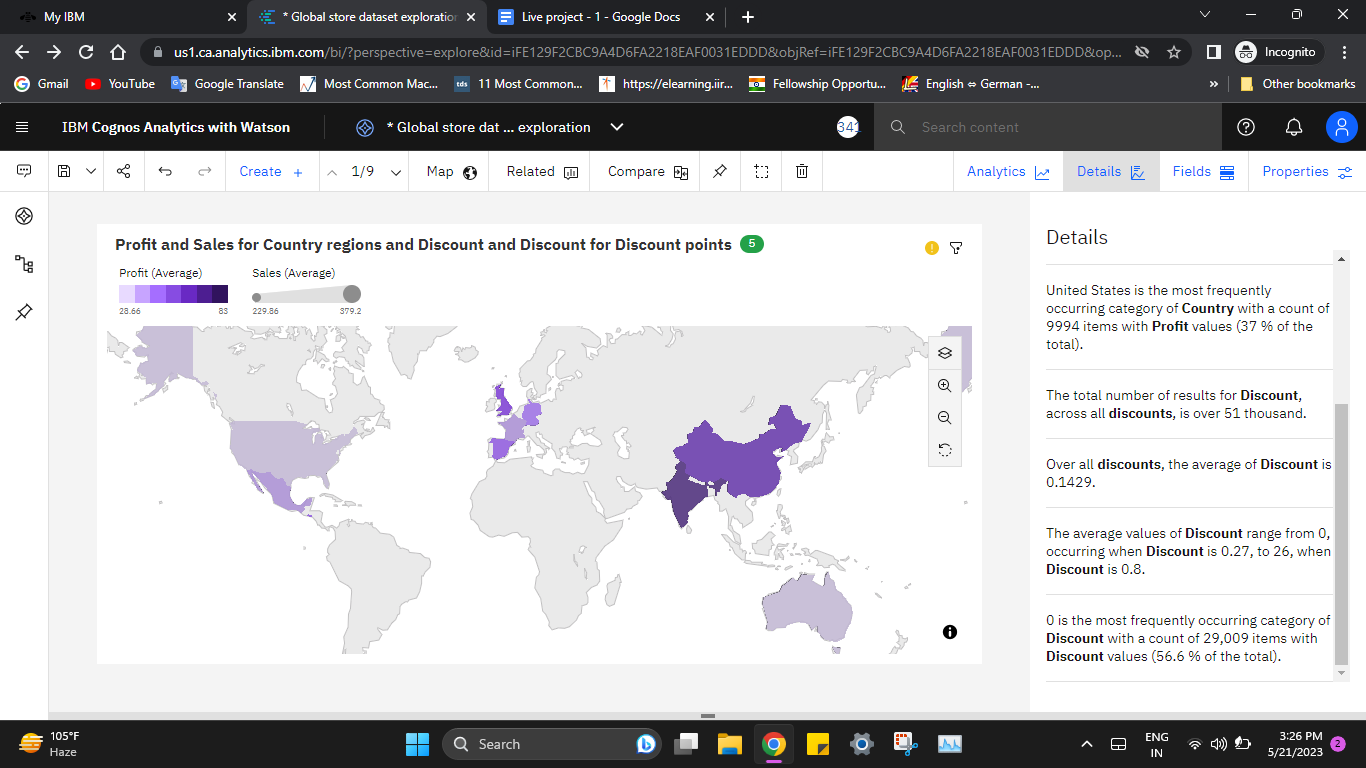
* **Data Quality and Cleansing:** The dataset may contain inconsistencies, errors, or missing values that need to be identified and addressed through data cleansing techniques to ensure accurate and reliable analysis results.
* **Data Integration and Preprocessing:** The dataset may consist of multiple sources or formats, requiring data integration and preprocessing to consolidate and transform the data into a usable format to ensure data consistency and compatibility.
* **Data Security and Privacy:** The dataset may contain sensitive information such as customer details or order transactions that need to be protected from unauthorized access or breaches.
* **System Infrastructure and Resources:** Adequate computing resources, including processing power and storage capacity, should be available to handle the volume and complexity of the dataset.
* **Analytics Tools and Technologies:** Selecting appropriate analytics tools and technologies based on the dataset size, analysis complexity, and organizational capabilities is crucial. Consideration should be given to the availability of tools for data exploration, visualization, statistical analysis, and machine learning.

**5.2 Evaluation Parameters (for Models)**

* Test/Train split of data - 85% training and 15% testing
* Classification metrics - Confusion matrix, Accuracy and Mean Square Error
* Detailed study of plots and correlation heat map

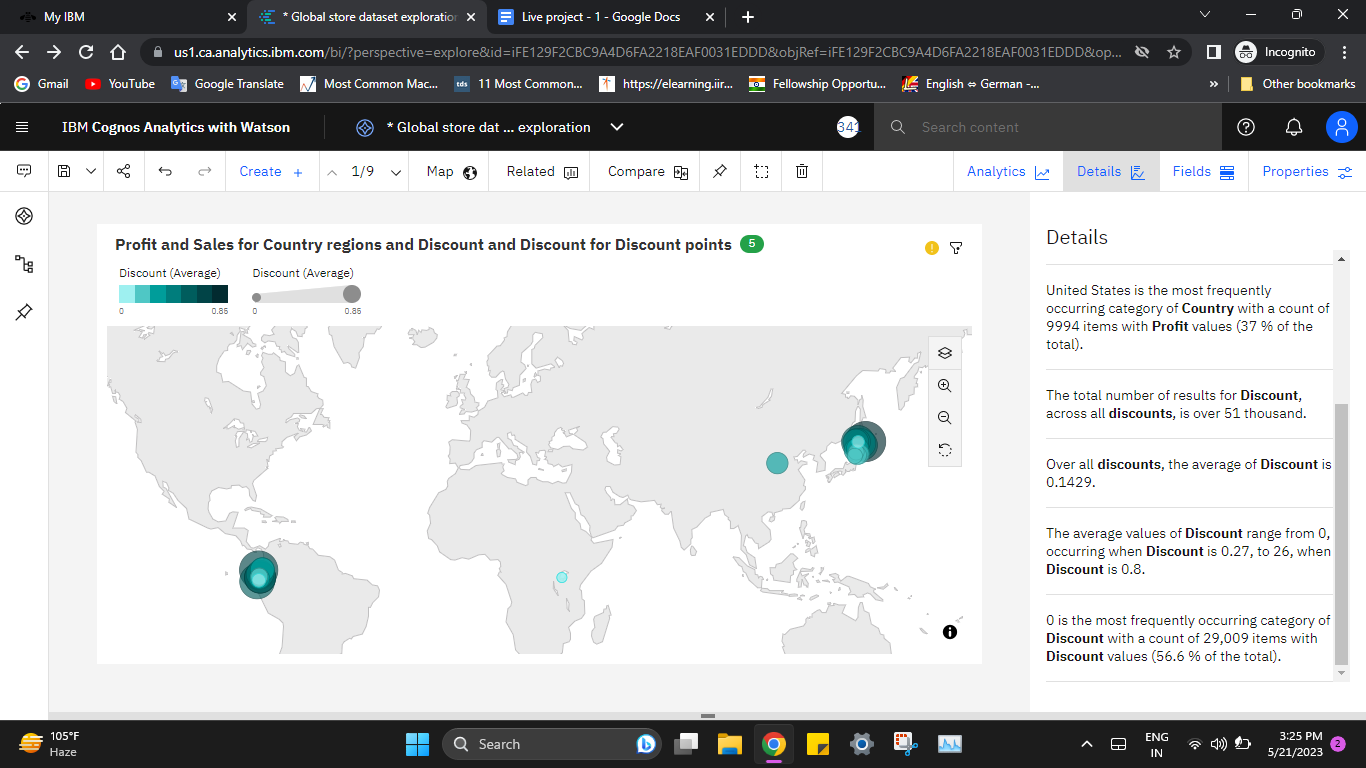
**5.3 Observations and Results**

After visualizing in the IBM Cognos Analytics, we came up with the following observations and results :



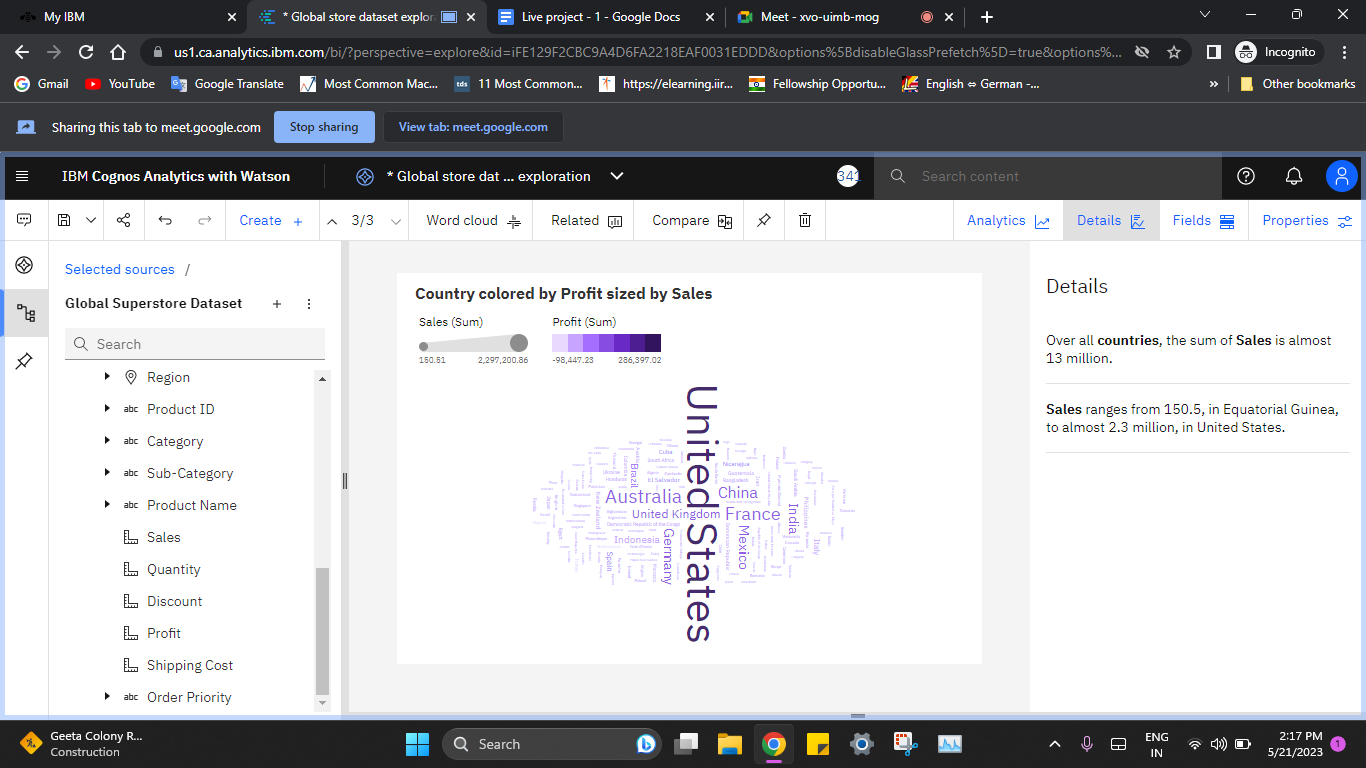
**Fig 5.3.1 Profit and Sales for Country regions**

Fig 5.3.1 depicts that average values of Profit range from 28.66, occurring in the United States, to 83, in India. Average of Profit is 44.3.



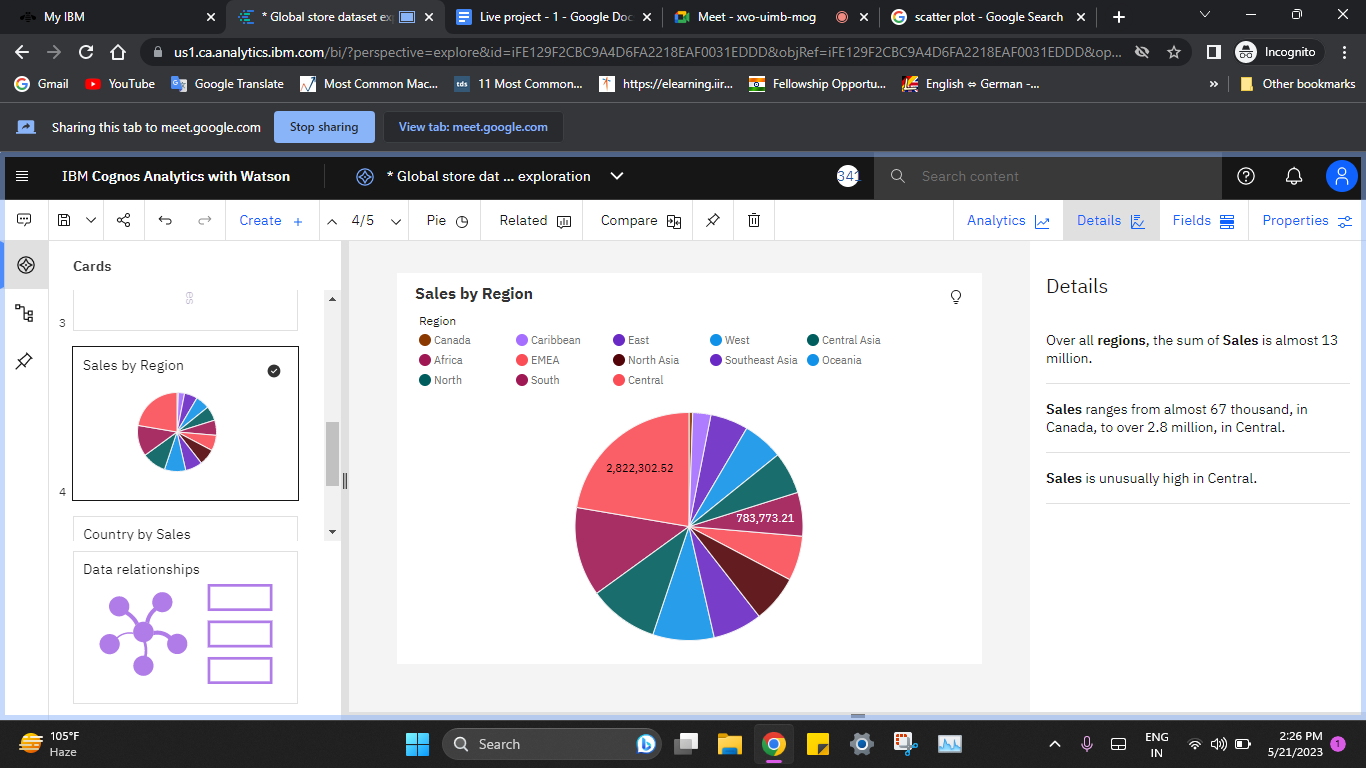
**Fig 5.3.2 Discount representation points over the World**

Fig 5.3.2 represents the discount value point over the World market region. Average Discount value is 0.1429 .



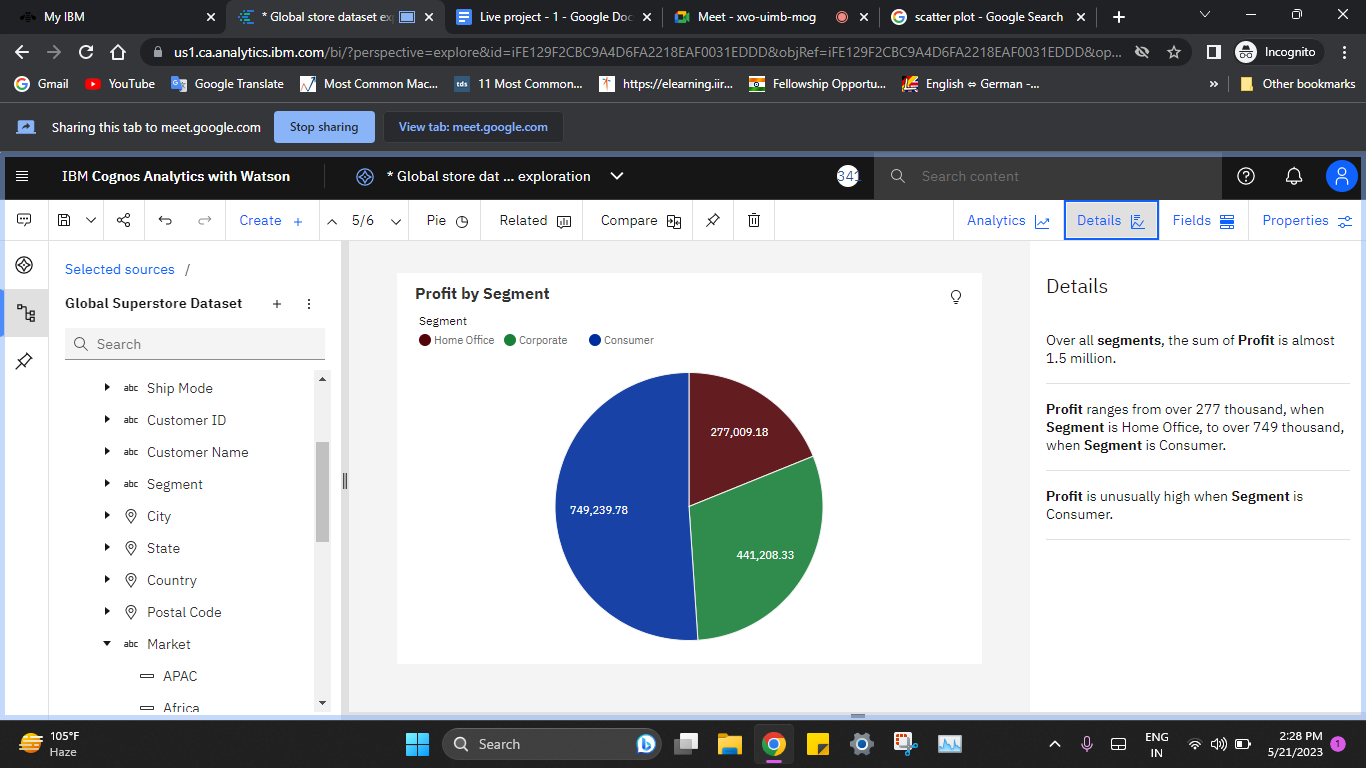
**Fig 5.3.3 Country by Profit sized by Sales**

Fig 5.3.3 depicts the Sales ranges from 150.5, in Equatorial Guinea, to almost 2.3 million, in the United States.



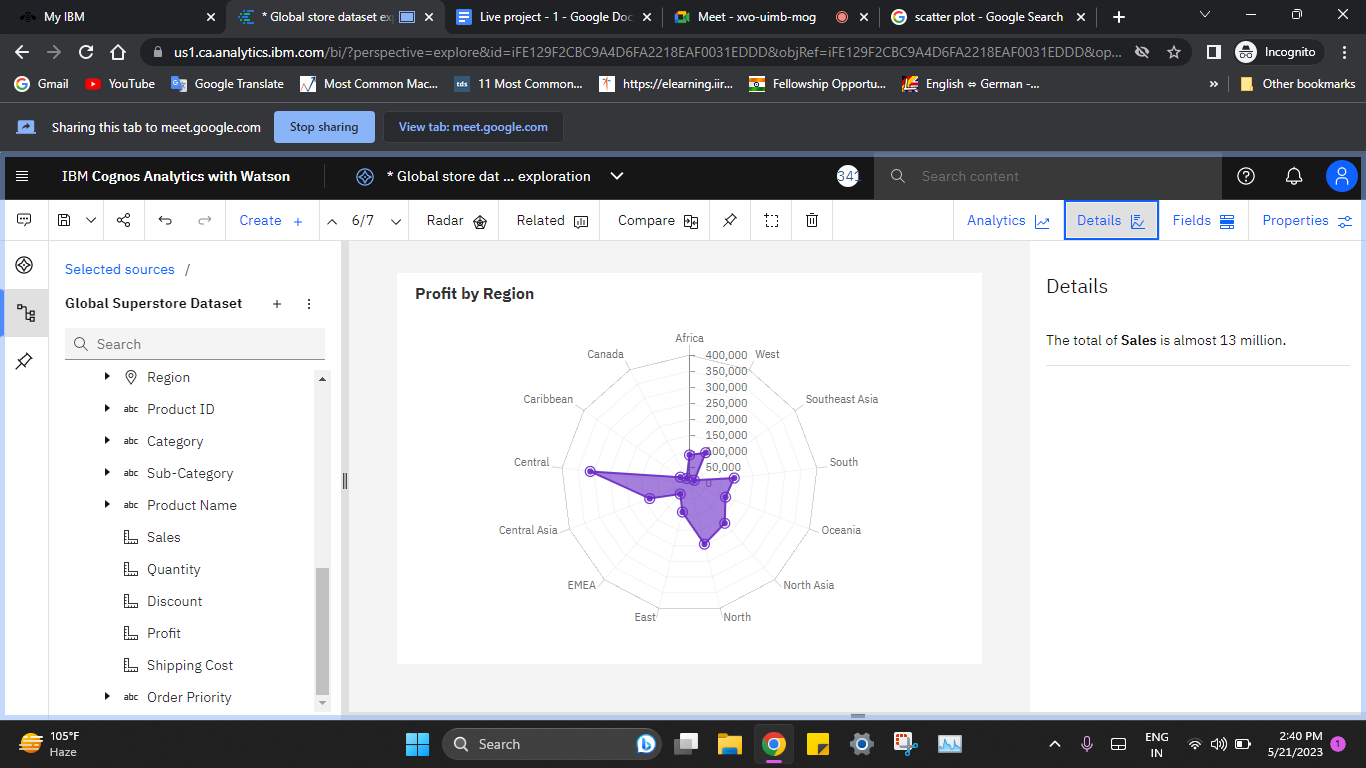
**Fig 5.3.4 Sales by Region**

Fig 5.3.4 depicts over all regions, the sum of Sales is almost 13 million. Sales ranges from almost 67 thousand, in Canada, to over 2.8 million, in Central. Sales are unusually high in Central.



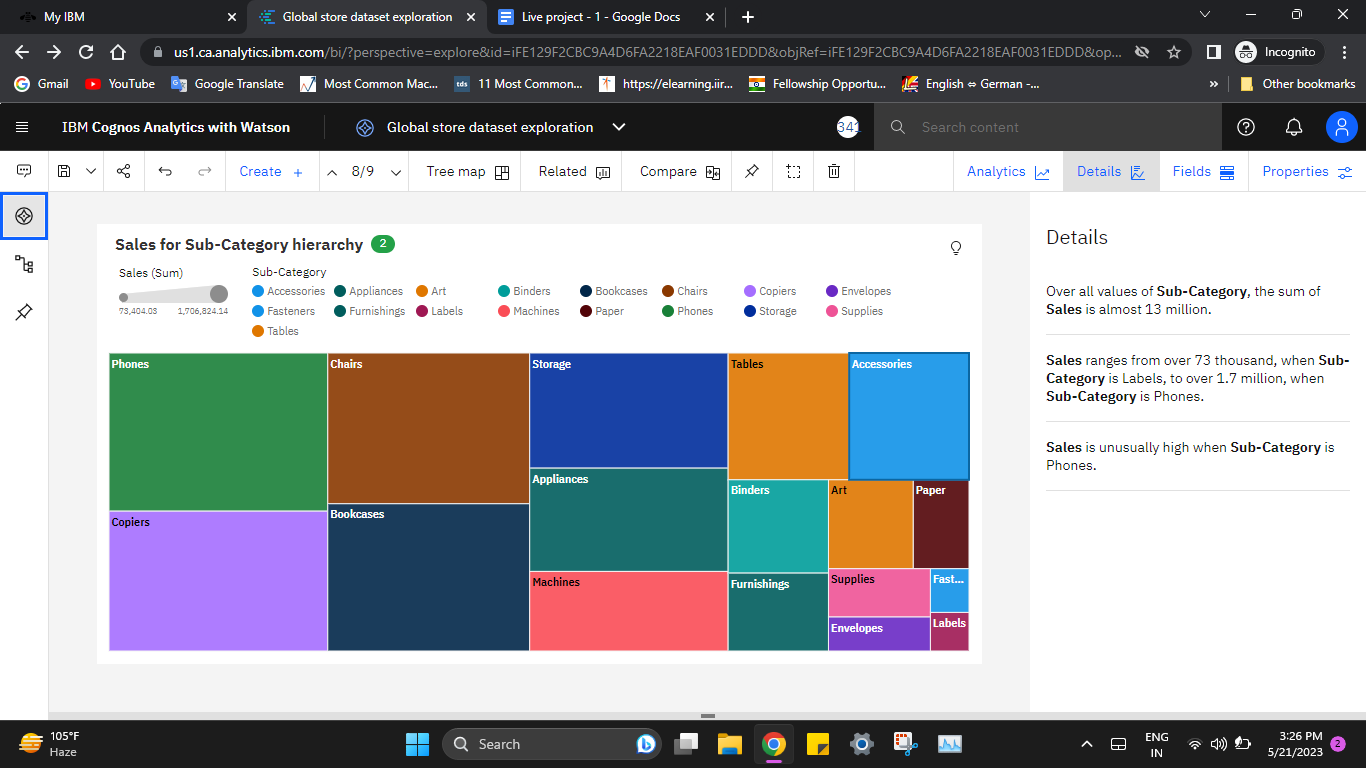
**Fig 5.3.5 Profit by Segment**

Fig 5.3.5 depicts over all segments, the sum of Profit is almost 1.5 million. Profit ranges from over 277 thousand, when Segment is Home Office, to over 749 thousand, when Segment is Consumer.Profit is unusually high when Segment is Consumer.



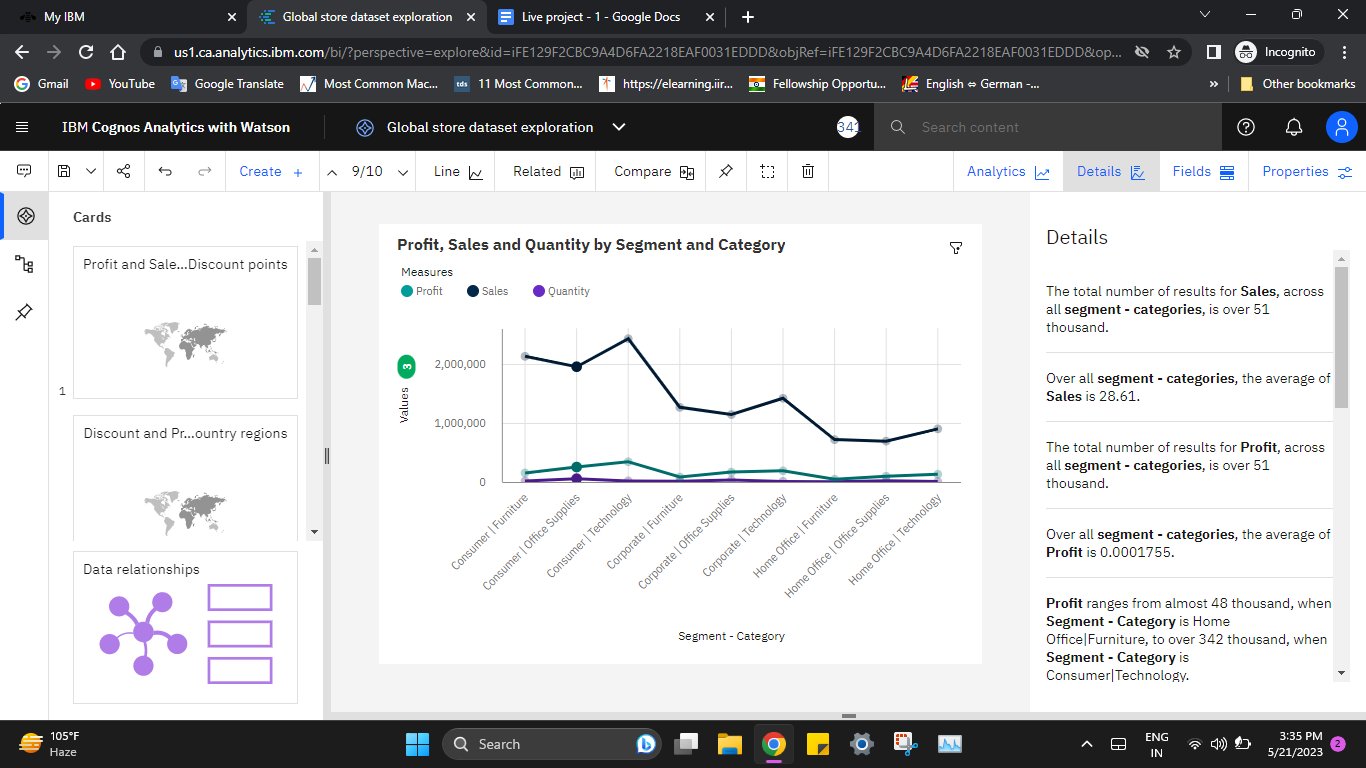
**Fig 5.3.6 Profit by Region**

Fig 5.3.6 depicts the total of Sales is almost 13 million. with highest in the Central Region



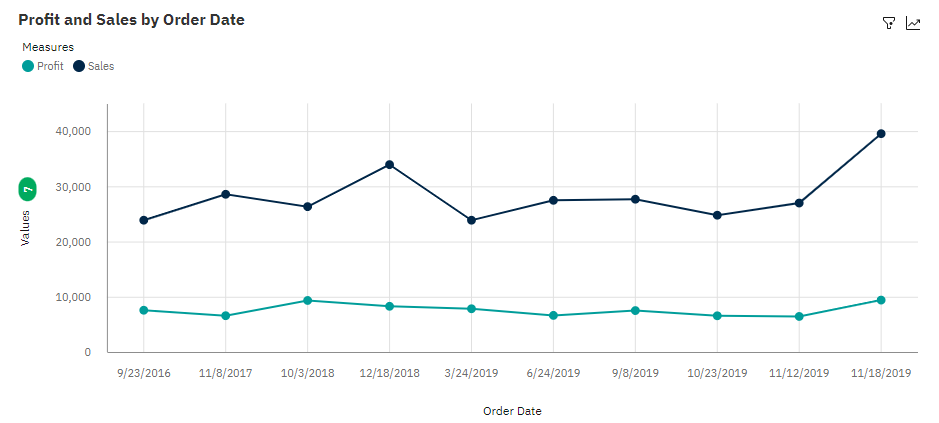
**Fig 5.3.7 Sales for Sub-Category hierarchy**

Fig 5.3.7 depicts the over all values of Sub-Category, Sales ranges from over 73 thousand, when Sub-Category is Labels, to over 1.7 million, when Sub-Category is Phones, Sales is unusually high when Sub-Category is Phones.



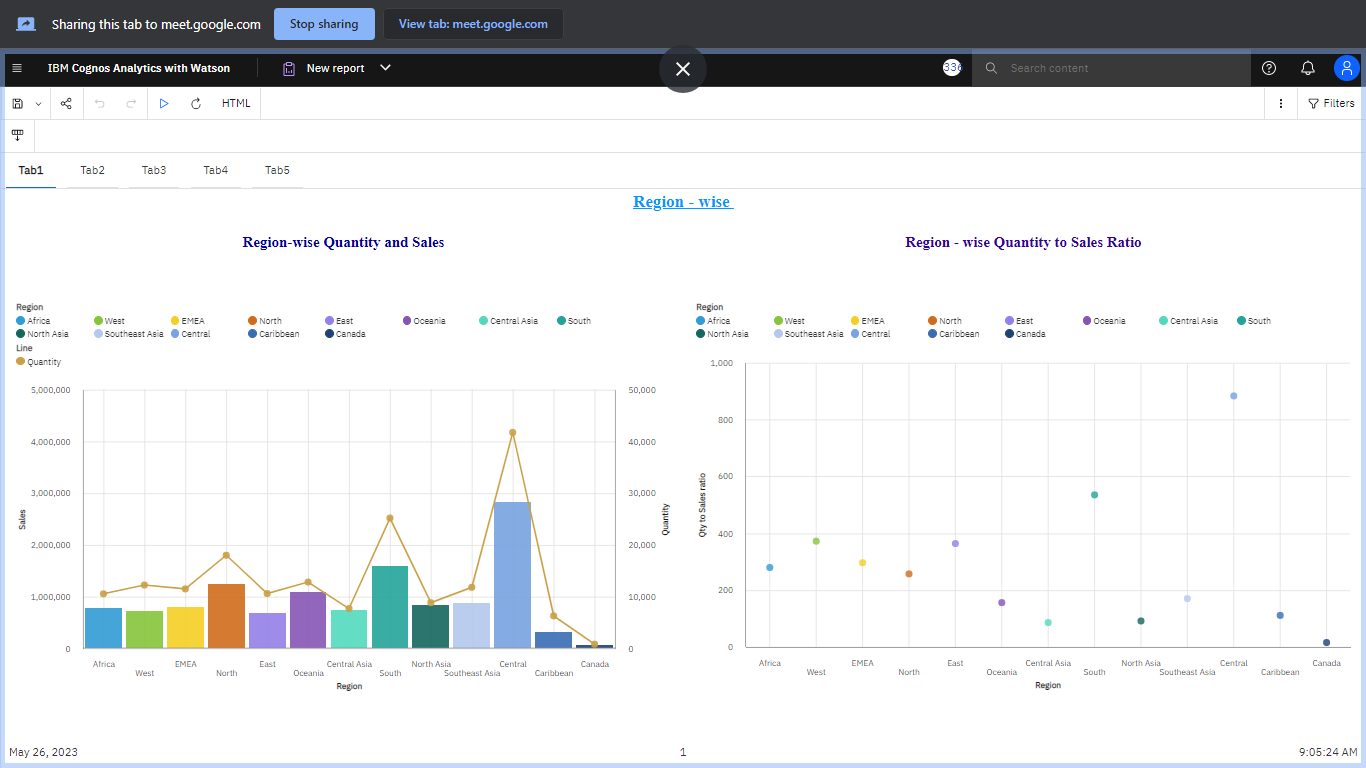
**Fig 5.3.8 Profit, Sales and Quantity by Segment and Category**

Fig 5.3.8 represents that over all the segments and categories, the average of Sales is 28.61. Profit is unusually high when Segment and Category is Consumer and Technology. The total number of results for Profit, across all segments and categories, is over 51 thousand. Over all segments and categories, the average of Profit is 0.0001755. Over all segments and categories, the average of Quantity is 246.5. Home Office segment and Technology Category is the most frequently occurring category with a count of 26,518 items with Sales values (51.7 % of the total).



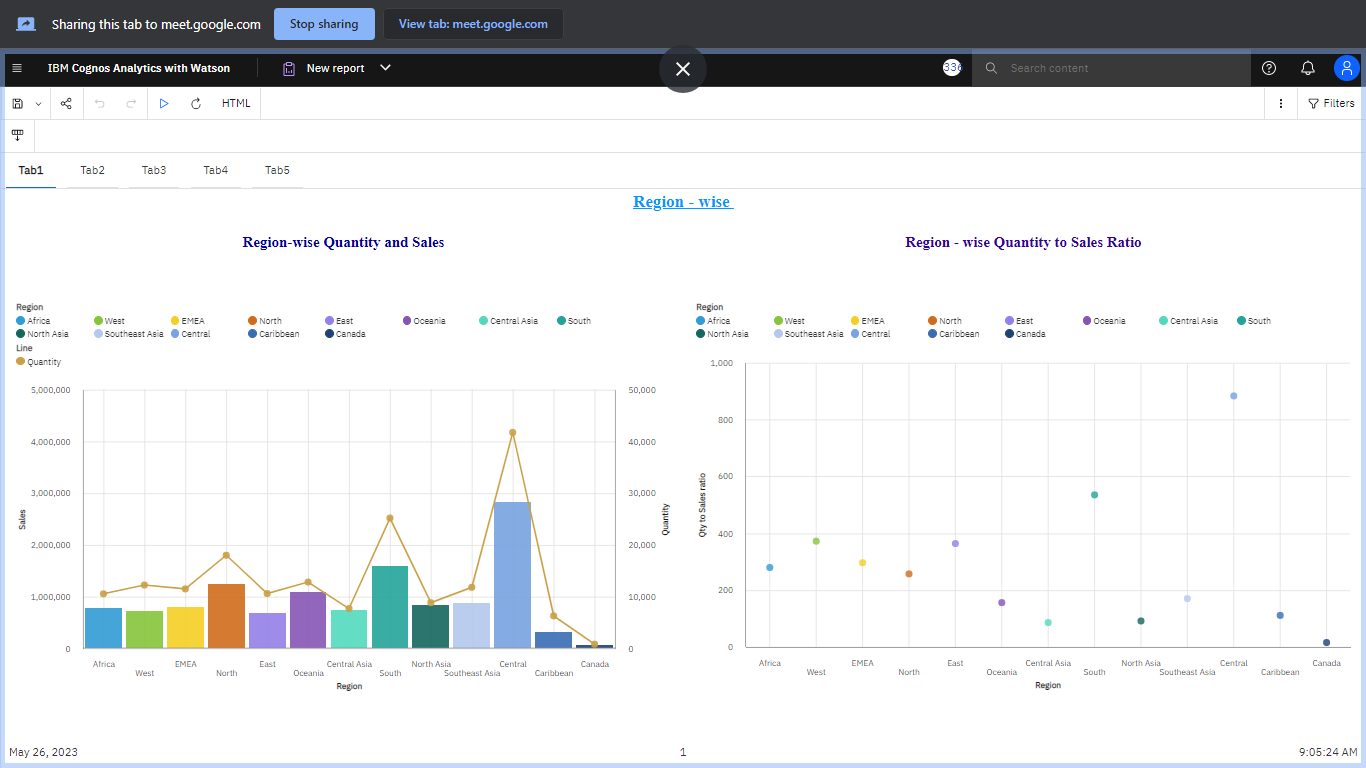
**Fig 5.3.9 Sales and Profit by Order Date**

Fig 5.3.9 depicts the total number of results for Sales, across all order dates, is 749. The average of Sales is 377.8 and average of Profit is 101.6 over all the order dates. 2019-11-18 is the most frequently occurring category of Order Date with a count of 127 items with Sales and Profit values (17 % of the total).



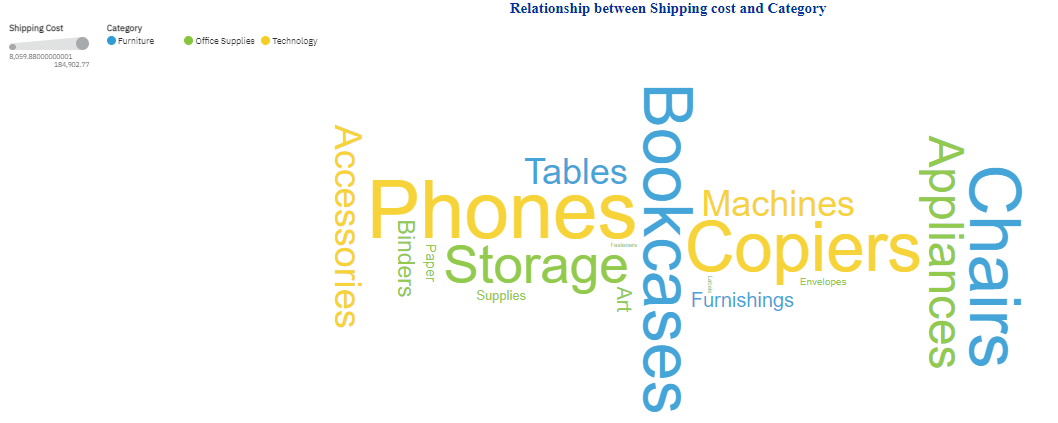
**Fig 5.3.10 Region-wise Quantity and Sales**

Fig 5.3.10 depicts the Region-wise bar and line plot of Sales and Quantity. It shows that the Central region has the highest sales and quantity.



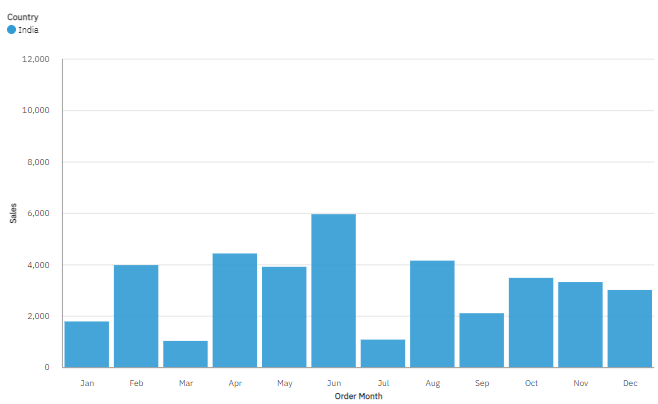
**Fig 5.3.11 Region-wise Quantity to Sales Ratio**

Fig 5.3.11 depicts the Central region has the highest sale and quantity



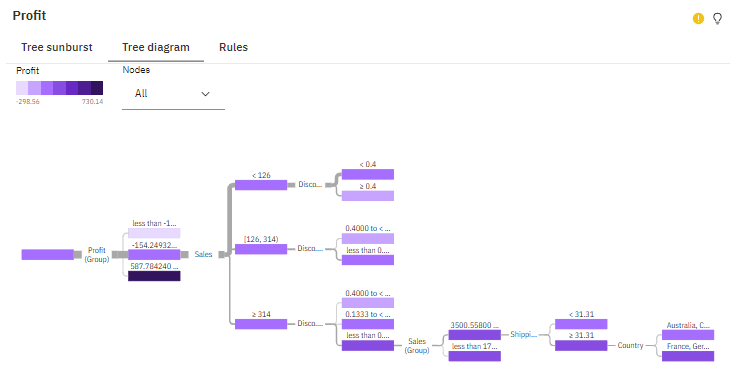
**Fig 5.3.12 Relationship between Shipping Cost and Sub-Category**

Fig 5.3.12 depicts the all values of Sub-Category and Category. The most significant value of Sub-Category is Phones that add up to 13.7 % of the total and Technology for Category that add up to 37.5 % of the total.



**Fig 5.3.13 Month-wise Sales in India**

Fig 5.3.13 depicts the month-wise sales over India. It shows that Sales are highest in the month - June which shows that sales after the Financial Year increases and last month of Financial Year has decreased sales.



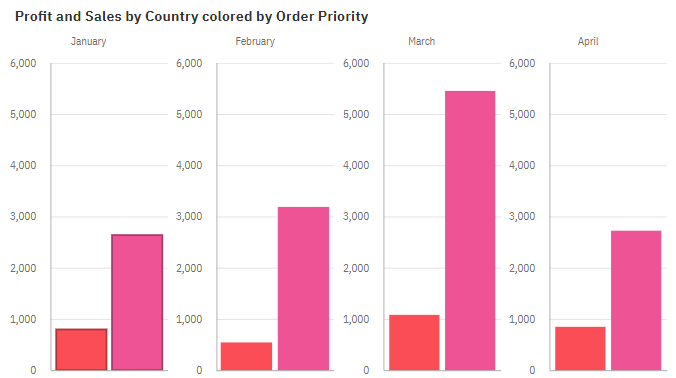
**Fig 5.3.14 Profit (Decision Tree)**

Fig 5.3.14 depicts the decision tree based on Discount (Group), Profit (Group), Sales, Shipping Cost and others , which predict Profit with a strength of 59.8%.



**Fig 5.3.15 Region (Decision Tree)**

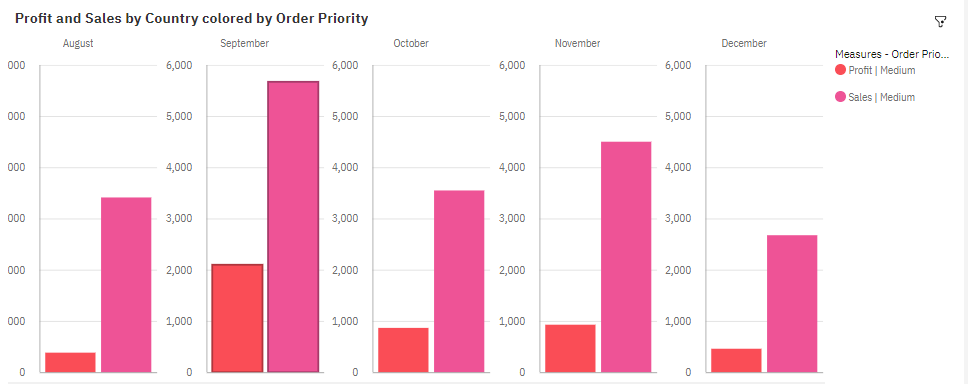
Fig 5.3.15 depicts the decision tree based on Country, Postal Code, and Market which predict Region with a strength of 89.4%.



**Fig 5.3.16(a) Profit and Sales by Country colored by Order Priority for the months of January-April**

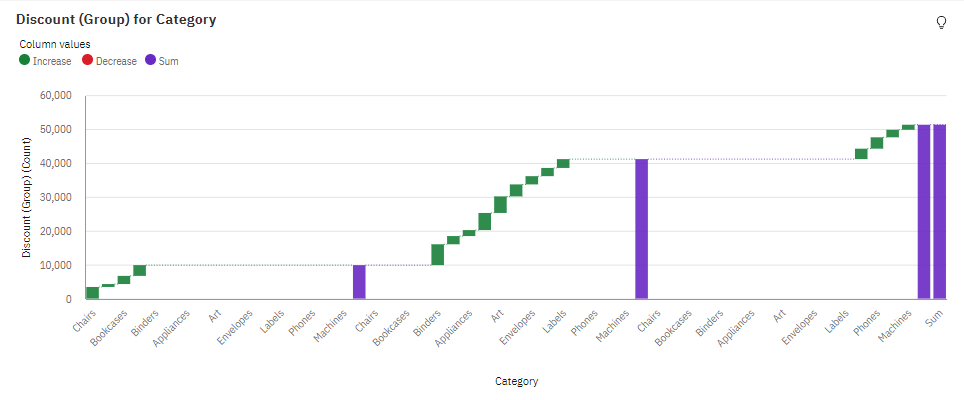


**Fig 5.3.16(b) Profit and Sales by Country colored by Order Priority for the months of May-August**



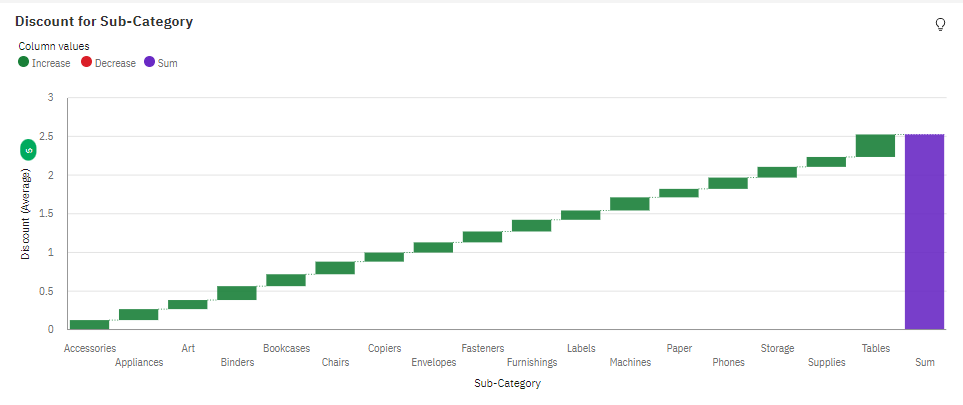
**Fig 5.3.16(c) Profit and Sales by Country colored by Order Priority for the months of September-December**

Fig 5.3.16(a) , Fig 5.3.16(b) and Fig 5.3.16(c) depict the total number of results for Sales, across all countries, is 936. The largest value of Profit is 2818, Sales is 5752 and occurs in India and June. August (11.9 %), October (11.3 %), September (11 %), June (10.8 %), and December (10.8 %) are the most frequent Month with Profit values (55.8 % of the total).



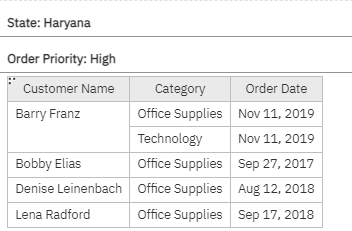
**Fig 5.3.17 Discount for Category**

Fig 5.3.17 depicts the Discount is unusually high when Category is Furniture and Sub-Category is Tables. Office Supplies is the most frequently occurring category of Category with Discount values (61 % of the total). Binders is the most frequently occurring category of Sub-Category with Discount values (12 % of the total).



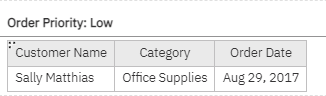
**Fig 5.3.18 Discount for Sub-Category**

Fig 5.3.18 depicts the average values of Discount range from 0.1095, occurring when Sub-Category is Paper, to 0.2907 (high), when Sub-Category is Tables. Binders is the most frequently occurring category of Sub-Category with Discount values (12 % of the total).



**Fig 5.3.19 (a) Different segments of data based on Order Priority of Haryana**

Fig 5.3.19 (a) depicts the subset of data based on ‘High’ category of Order Priority segmented on State - Haryana. There are 7 customers under this category.



**Fig 5.3.19 (b) Different segments of data based on Order Priority of Haryana**

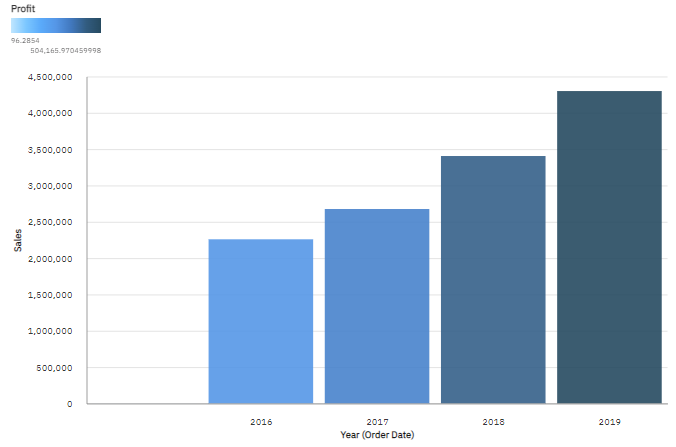
Fig 5.3.19 (b) depicts the subset of data based ‘Low’ category of Order Priority segmented on State - Haryana. There is only 1 customer under this category.





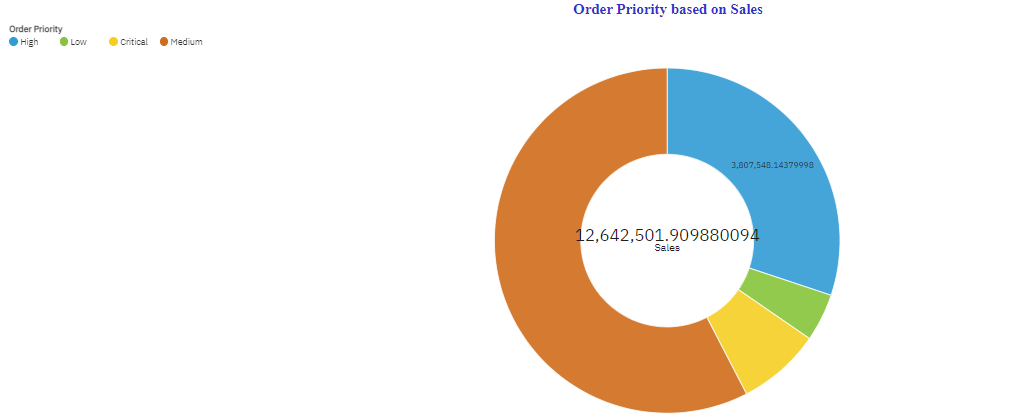
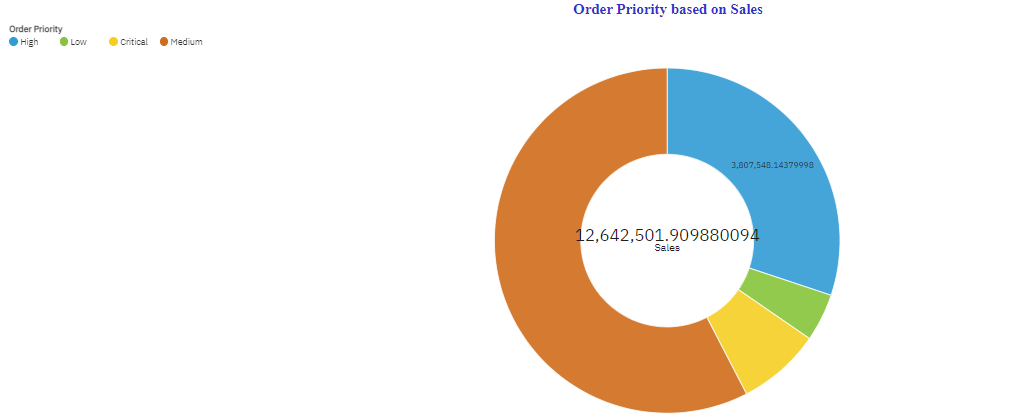
**Fig 5.3.19 (c) Different segments of data based on Order Priority of Haryana**

Fig 5.3.19 (c) depicts the subset of data based ‘Medium’ category of Order Priority segmented on State - Haryana. There are 13 customers under this category.



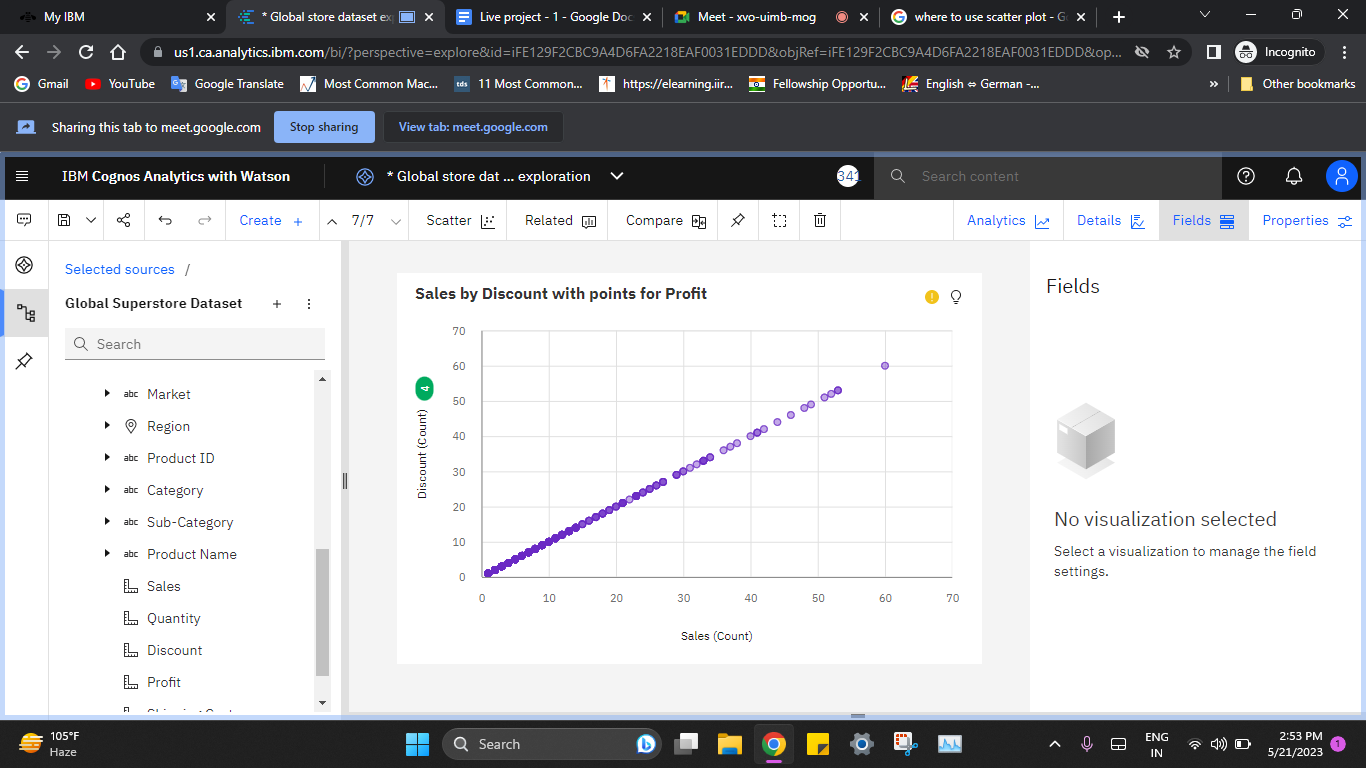
**Fig 5.3.20 Year-wise Profit Rate on Sales**

Fig 5.3.20 depicts the profit intensity based on sales during each year. This shows 2019 has the highest profit.



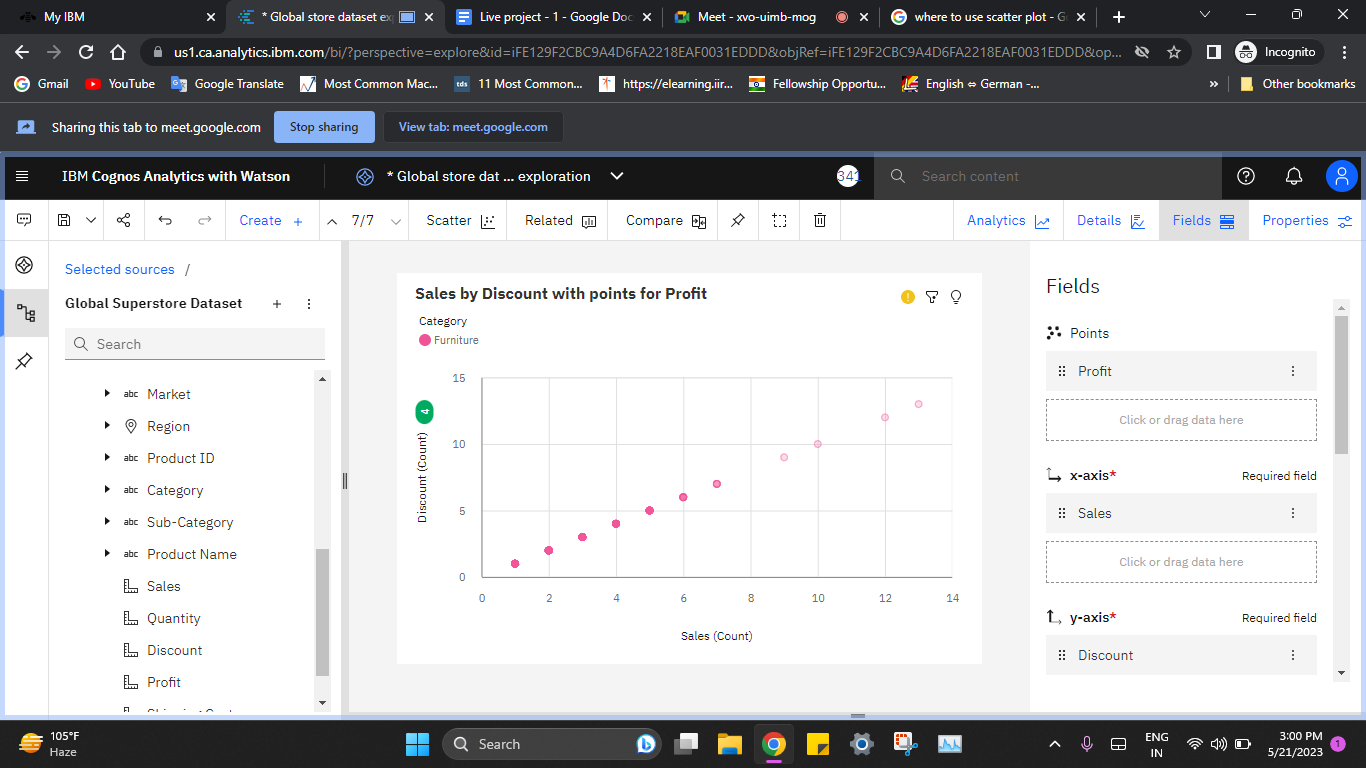
**Fig 5.3.21 Sales based on Order Priority**

Fig 5.3.21 depicts the sales according to different categories in Order Priority. This shows that Medium Order priority has the highest sales followed by High order priority sales.



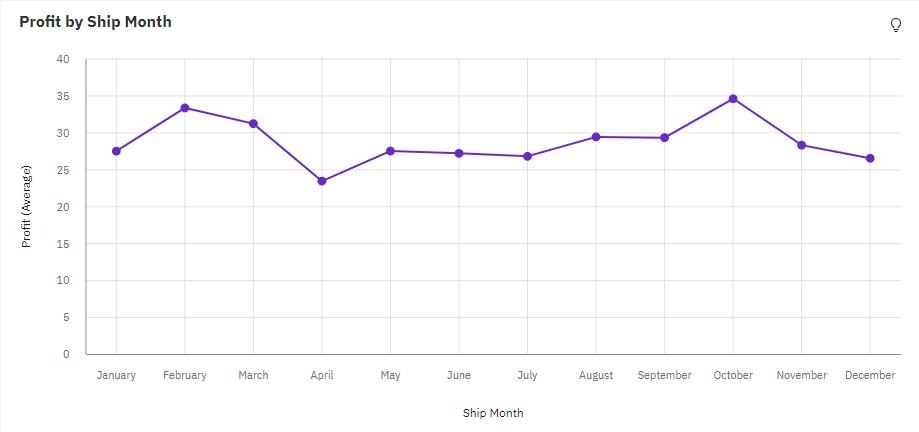
**Fig 5.3.22 Sales by Discount for Profit**

Fig 5.3.22 depicts the rising discount rates can lead to increased consumer spending resulting in a significant boost to sales and profits.



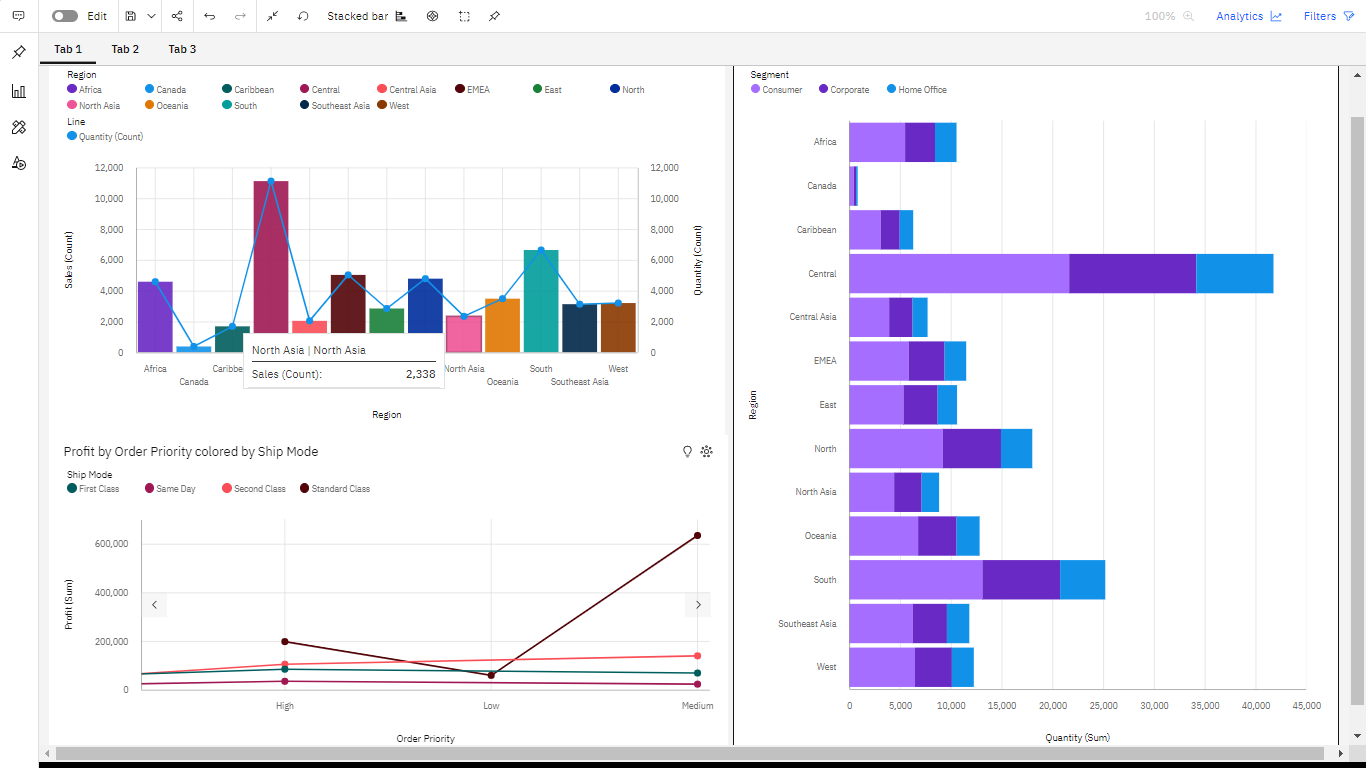
**Fig 5.3.23 Sales by Discount for Profit for Furniture Category**

Fig 5.3.23 depicts the discount rates rise, sales and profits soar in the category Furniture.

****

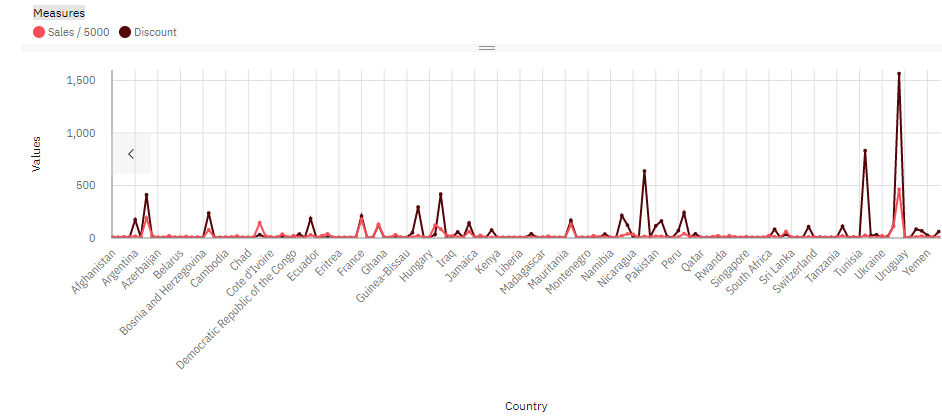
**Fig 5.3.24 Profit ordered by Ship Month**

Fig 5.3.24 shows that over all ship months, the average of Profit is 28.61.December (12.3 %), November (12 %), September (11.5 %), June (10.1 %), and August (9.4 %) are the most frequently occurring categories of Ship Month with a combined count of 28,345 items with Profit values (55.3 % of the total) .



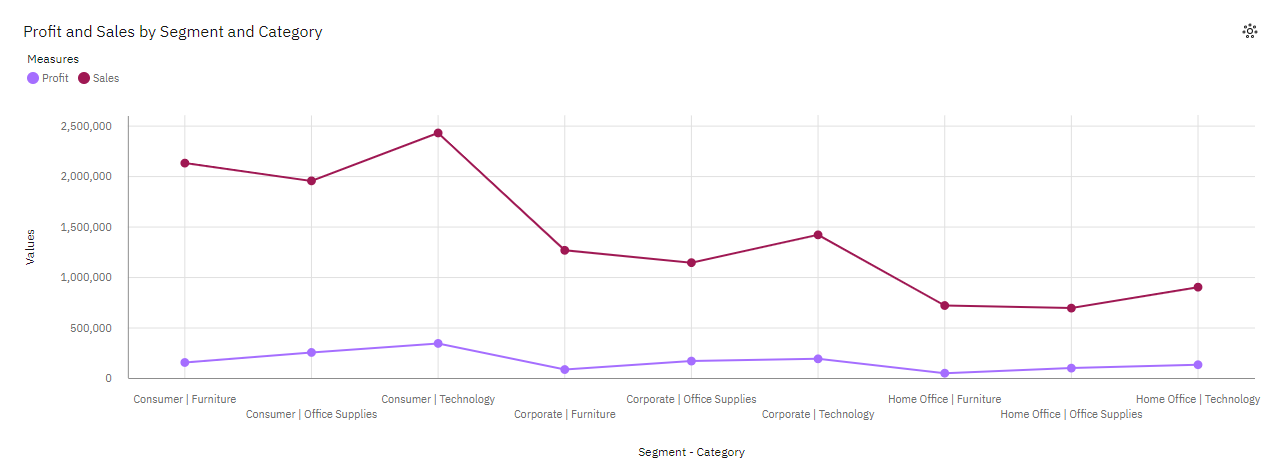
**Fig 5.3.25 Region-wise Quantity segmented by Segment**

Fig 5.3.25 depicts Quantity of products in different regions based on different segments. It shows that the Central Region has the highest values of both Quantity and Profit. Consumer segment has the highest values of both Quantity and Profit.



**Fig 5.3.26 Country-wise measure of Sales and Discount**

Fig 5.3.26 depicts that the United States has the highest values of both Discount and Quantity and is the most frequently occurring category of Country with a count of 9994 items with Discount values (19.5 % of the total). For all countries, the average Discount value is 0.1429.

****

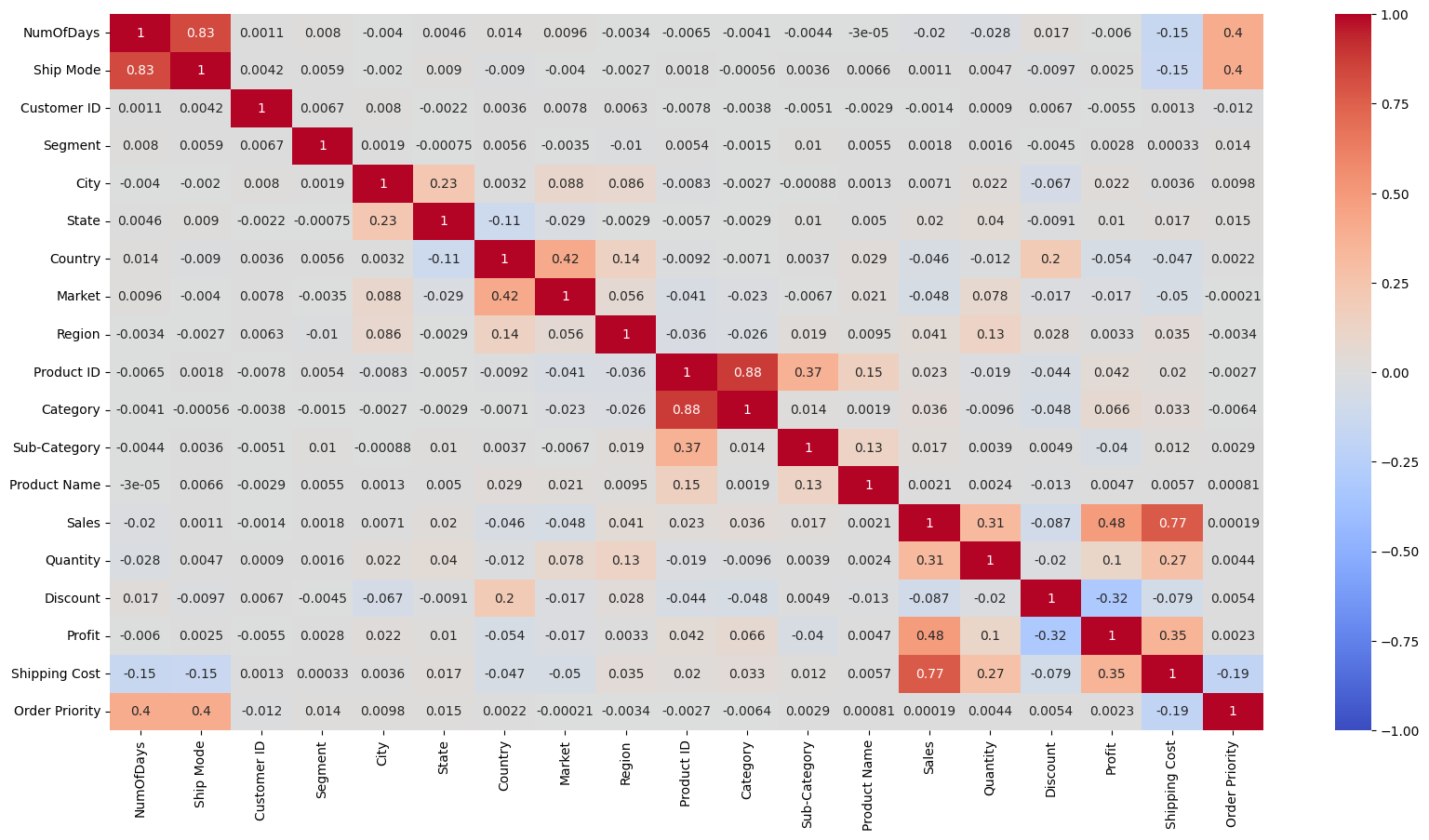
**Fig 5.3.27 Profit and Sales by Segment and Category**

Fig 5.3.27 depicts that the Consumer segment has the highest values of both Profit and Shipping Cost. Profit is most unusual when Segment and Category is Consumer and Technology and Home Office and Furniture. Office Supplies category has the highest Total Quantity but is ranked 2 in Total Profit.



**Fig 5.3.28 Sub-Category hierarchy colored by Profit and sized by Sales**

Fig 5.3.28 depicts the tree map showing Sub-Category Copiers has the highest Total Profit but is ranked 14 in Total Quantity. Sub-Category Binders has the highest Total Quantity but is ranked 8 in Total Profit. Sum of Sales is almost 13 million among all the Sub-Category.



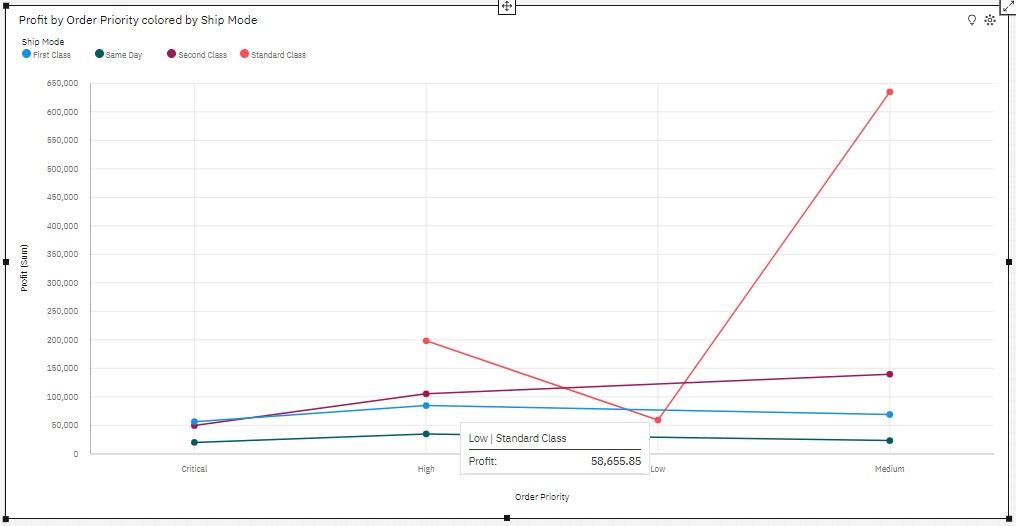
**Fig 5.3.29 Correlation heat map between all numerical fields of dataframe (in python)**

Fig 5.3.29 depicts that there is a positive correlation between the following fields :

1. Category and Product ID (0.88)
2. NumofDays and Ship Mode (0.83)
3. Shipping Cost and Sales (0.77)
4. Profit and Sales (0.48)
5. Order Priority and Ship Mode , NumofDays and Ship Mode (0.4)
6. Sub-Category and Product ID (0.37)
7. Profit and Shipping Cost (0.35)
8. Quantity and Sales (0.31)
9. Shipping Cost and Quantity (0.27)
10. Country and Discount (0.2)
11. Quantity and Profit (0.1)

A slight correlation between Market and Quantity , Region and Quantity , Category and Profit was also seen.

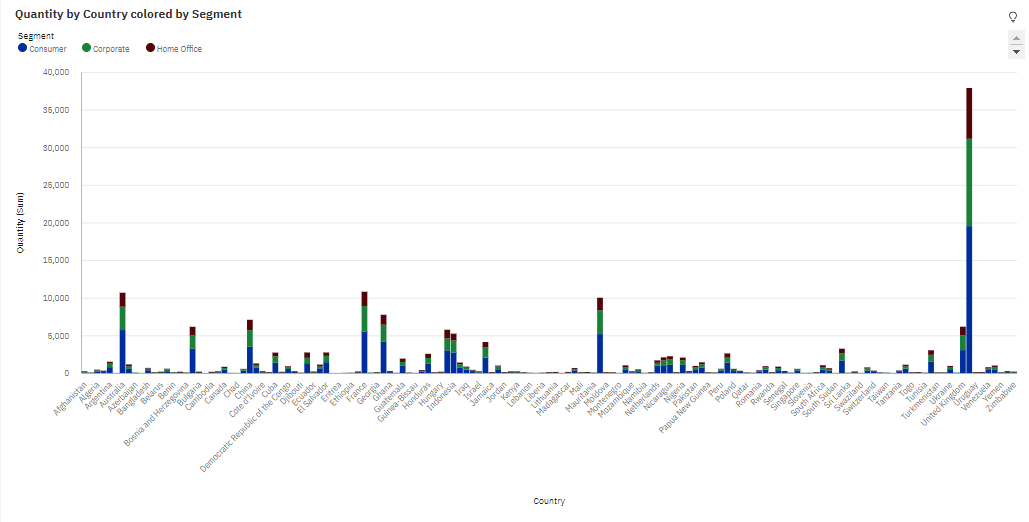
There is a negative correlation seen in the fields : Profit and Discount (-0.32), Discount and Sales (-0.087), Quantity and Discount (-0.02), , Shipping Cost and Discount (-0.079), Shipping Cost and NumOfDays , Shipping Cost and Ship Mode (-0.15), Shipping Cost and Order Priority (-0.19) and many more.

****

**Fig 5.3.30 Profit by Order Priority colored by Ship Mode**

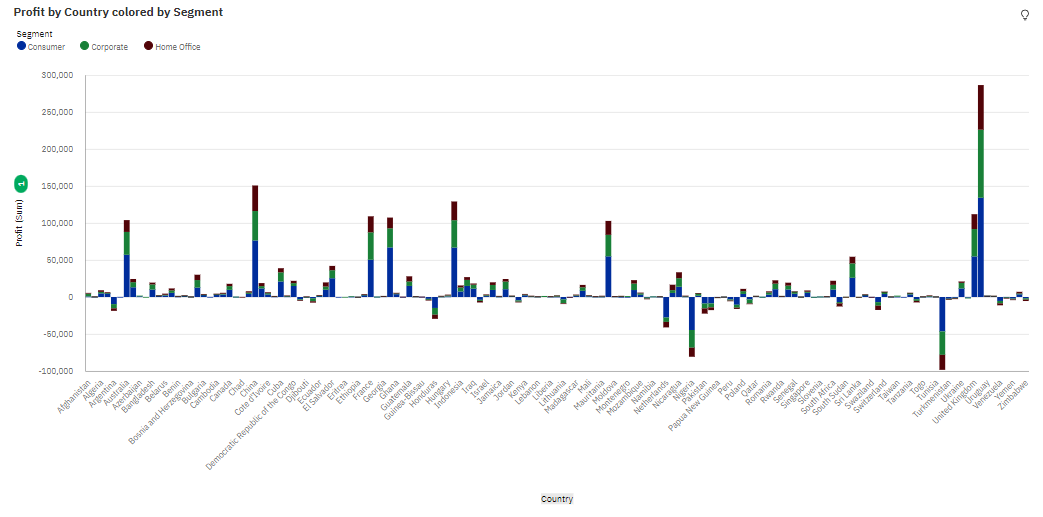
Fig 5.3.30 depicts the Profit is unusually high when Ship Mode is Standard Class and Profit is unusually high when Order Priority is Medium add up to 58.9 % of the total.

Profit is unusually high when the combination of Order Priority and Ship Mode is Medium and Standard Class.

****

**Fig 5.3.31 Country - wise Quantity colored by Segment**

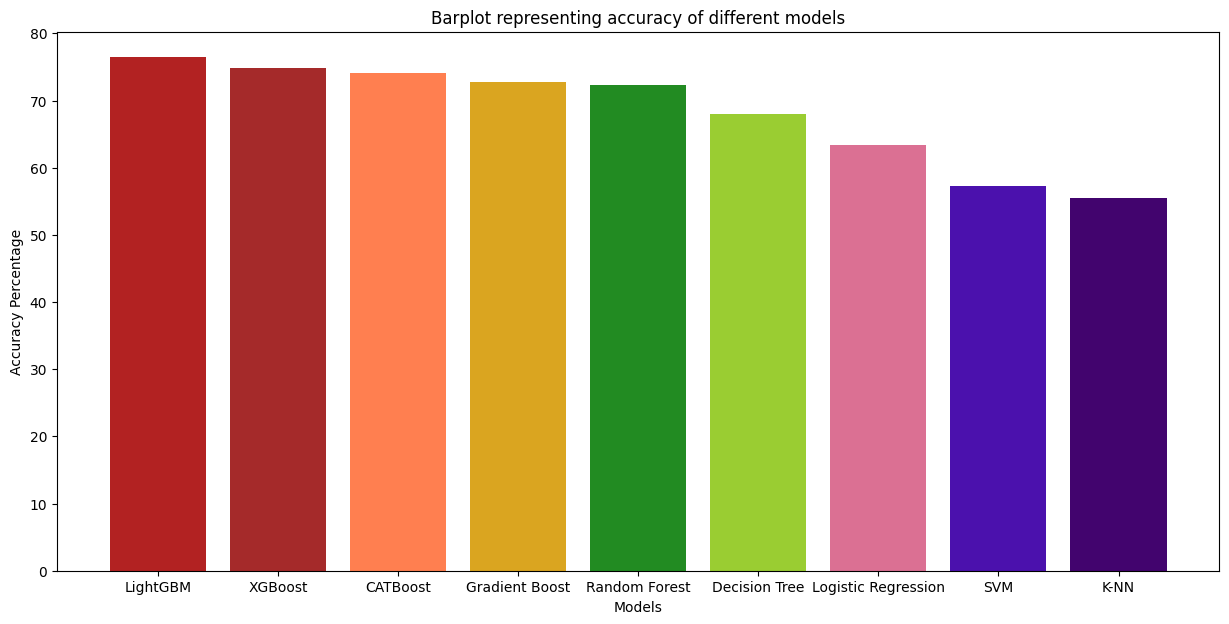
Fig 5.3.31 depicts the summed values of Quantity range from 1 to nearly twenty thousand. Quantity is unusually high when the combination of Country and Segment is United States and Consumer. For Quantity, the most significant Country is United States (add up to nearly 21.2 % of the total) and the most significant Segment is Consumer (add up to 51.7 % of the total).

****

**Fig 5.3.32 Profit by Country colored by Segment**

Fig 5.3.32 depicts the summed values of Profit ranges from nearly minus 98 thousand, in Turkey and -80 thousand in Nigeria (mostly with the consumer segment), to over 286 thousand, in the United States. Profit is unusually high when the combination of Country and Segment is the United States (add up to 15 % of the total) and Consumer.

For Profit, the most significant Segment is Consumer with 51.1 % of the total.



**Fig 5.3.33 Plot representing accuracy of different Classification models**

**(in python)**

Fig 5.3.33 depicts the accuracy of different classification models. The LightGBM model has the highest accuracy (76.42%) and the K-nearest Neighbour model has the least accuracy (55.47%).

This research focused on analyzing the Global Superstore dataset to identify the impacts of different aspects on sales and profit of different products, and to understand customer behavior trends. It involved tools like IBM Cognos Analytics for data visualization and Python with ML algorithms for deploying machine learning models. From our project, we found out that there was an unusual change (increase) in all the Sales, Profit and Quantity of the United States. But the US has the lowest profit as compared to the ratio of Sales to Profit whereas India has the highest Sales to Profit ratio. It was also seen that the correlation coefficient of Shipping Cost and Sales has a significant value, which means that more Sales means more Shipping Cost. Whereas, Shipping Cost and Discount have a negative relationship , which means that more discount imposes less Shipping Cost.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORK**

**6.1 Conclusion**

The United States has the lowest value of ratio of Sales to Profit whereas India has the highest Sales to Profit ratio. Acquiring new customers can be more expensive than retaining existing ones for the United States. Assessing the pricing structure and ensuring to have set the prices at a level that allows for an adequate profit margin is a must for the US. By suggesting some complementary or upgraded products, they can increase the average order value and boost their profits. For example, with a product - Copier, they can offer a cover and or some other accessories or add pages and other related items so as to gain customer satisfaction. Lowest Shipping cost has a positive impact on the Profit in the Sales. In India, Sales and profit will increase if discount value is increased. In Nigeria and Turkey , after giving discounts also, Profit is on the negative side, i.e., in loss. They can conduct a thorough review of their business operations and identify the areas where they can reduce costs without compromising the product quality or customer experience. In most countries, the Discount value is lowest in the Segment - Home Office which has a direct negative impact on Sales and their overall Profit. The most accurate predictive model was the LightGBM classifier with an accuracy percentage of 76.42%.

**6.2 Future Work**

* Further analysis: The research may suggest additional analyses that could provide deeper insights or address specific research questions that were not fully explored in the current study. We can perform hyperparameters tuning to increase the accuracy of the classification and ensemble models.
* Advanced modeling techniques: Using more sophisticated machine learning algorithms, deep learning techniques, or advanced statistical models to improve prediction accuracy or uncover hidden patterns in the data.
* Validation and generalization: The study might recommend validating the findings on independent datasets to ensure the robustness and generalizability of the results.
* Real-world implementation: Future work may involve implementing the research findings in a real-world setting and evaluating the practical implications of the proposed solutions.
* Data collection and expansion: Researchers may suggest collecting additional data or expanding the dataset to include more variables, time periods or diverse sources to enhance the comprehensiveness and reliability of the analysis.
* Integration with other systems: The research could propose integrating the developed models or insights with existing business systems, such as customer relationship management (CRM) systems .
* Product Performance: Identify the top-performing products in terms of sales and profitability. Analyze their characteristics, pricing strategies, and customer preferences to replicate their success in other products.

**CHAPTER 7**

**REFERENCES**

[1] ["About Python"](https://www.python.org/about). Python Software Foundation. [Archived](https://web.archive.org/web/20120420010049/http://www.python.org/about/) from the original on 20 April 2012. Retrieved 24 April 2012., second section "Fans of Python use the phrase "batteries included" to describe the standard library, which covers everything from asynchronous processing to zip files." Available : https://www.python.org/about

[2] Jay Selig. “What is Machine Learning? A definition. “ . Published 14 March 2022. [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/

[3] Wikipedia, “IBM Cognos Analytics,”. Last edited on 18 March 2023, at 05:03 (UTC). [Online]. Available: https://en.wikipedia.org/wiki/IBM\_Cognos\_Analytics.

[4] Google. “Colaboratory Frequently Asked Questions”. [Online]. Available: https://research.google.com/colaboratory/faq.html

[5] Shankar, Venky. “Big Data and Analytics in Retailing”. Published : 2019. NIM Marketing Intelligence Review. 11. 37-40. DOI : 10.2478/nimmir-2019-0006.

[6] Chandramana, Sudeep. “Retail Analytics: Driving Success in Retail Industry with Business Analytics”. Published : 2017. Volume 7. DOI : 10.6084/m9.figshare.13323179.

[7] Marnik G. Dekimpe. “Retailing and retailing research in the age of big data analytics”. International Journal of Research in Marketing, Volume 37, Issue 1, Published : 2020. Pages 3-14, ISSN 0167-8116. Available : https://doi.org/10.1016/j.ijresmar.2019.09.001.

[8] Elgendy, Nada & Elragal, Ahmed.” Big Data Analytics: A Literature Review Paper”.Published : 2014. Lecture Notes in Computer Science. 8557. 214-227. DOI : 10.1007/978-3-319-08976-8\_16.

[9] Mr. Subhashish Kumar, Dr. Namarata Dhanda and Mr. Ashutosh Pandey. Data Science - Cosmic Infoset Mining, Modelling and Visualization. Published at 2018 International Conference on Computational and Characterization Techniques in Engineering & Sciences (CCTES), Integral University, Lucknow, India. Published: Sep 14-15, 2018.

[10] Musa J. Jafar,Jeffry Babb,Amjda Abdullat.”Emergence of Data Analytics in the Information Systems Curriculum”. Published : 2017 . Information Systems Education Journal (ISEDJ). ISSN: 1545-679X . Available : https://files.eric.ed.gov/fulltext/EJ1151898.pdf

[11] Singh, Manpreet & Ghutla, Bhawick & Jnr, Reuben & Mohammed, Aesaan & Rashid, Mahmood. “Walmart's Sales Data Analysis - A Big Data Analytics Perspective”. Published : 2017. Pages : 114-119. DOI : 10.1109/APWConCSE.2017.00028.

[12] Shahbaz, Muhammad & Gao, Changyuan & Zhai, Lili & Shahzad, Fakhar & Luqman, Adeel & Zahid, Rimsha.” Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities”. Published : 2021. PLoS ONE. 16. 1-22. DOI : 10.1371/journal.pone.0250229.

[13] Sumindar Kaur Saini,Vishal Dhull, Sarbjeet Singh, Akashdeep Sharma.”Visual Exploratory Data Analysis of COVID-19 Pandemic”.Published 2020. 5th IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE 2020 (IEEE Record#51050). DOI: 10.1109/ICRAIE51050.2020.9358331

[14] [Sruthi E R](https://www.analyticsvidhya.com/blog/author/sruthi94/). “ Understanding Random Forest ”. Published 17 June 2021. [Online]. Available : https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[15] Shivam Sharma . “K-Nearest Neighbour: The Distance-Based Machine Learning Algorithm.”. Published 15 May 2021. [Online]. Available : <https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/>

[16] Nitesh Trabani.”How the Gradient Boosting Algorithm works?”. Published 19 April 2021. [Online]. Available : <https://www.analyticsvidhya.com/blog/2021/04/how-the-gradient-boosting-algorithm-works/>

[17] V[ijay Choubey](https://medium.com/@vijay-choubey?source=post_page-----adf1bc246254--------------------------------). “Decision Tree-End to End Implementation” . Published 28 October 2020. [Online]. Available : <https://medium.com/analytics-vidhya/decision-tree-end-to-end-implementation-adf1bc246254>

[18] Rohith Gandhi. “Support Vector Machine — Introduction to Machine Learning Algorithms ”. Published 7 June 2018. [Online]. Available : https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[19] ”Logistic Regression in Machine Learning”. Published [Online]. Available : https://www.javatpoint.com/logistic-regression-in-machine-learning

[20] Akshi Saxena. “CatBoost in Machine Learning”. Published [Online] . Last Updated : 12 May, 2023 . Available : https://www.geeksforgeeks.org/catboost-ml/

[21] Pawan. “XGBoost”. Published [Online]. Last Updated : 06 Feb, 2023.Available : https://www.geeksforgeeks.org/xgboost/

[22] Jason Brownlee in [Ensemble Learning](https://machinelearningmastery.com/category/ensemble-learning/). “How to Develop a Light Gradient Boosted Machine (LightGBM) Ensemble”. Published on : November 25, 2020. [Online]. Last Updated : April 27, 2021. Available : https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble/