DATA VISUALIZATION FOR ANALYTICS

E-commerce Trend Analysis

Yashvi Pipaliya (yjp5162@psu.edu)

Shatakshi Singh(sqs7319@psu.edu)

School of Graduate Professional Studies

MPS in Data Analytics

Oct, 2023

Document Control

Work carried out by:

Name	Email Address	Other
Yashvi Pipaliya	yjp5162@psu.edu	
Shatakshi Singh	sqs7319@psu.edu	

Revision Sheet

Release No.	Date	Revision Description
1	10-29-2023	Formulated the business and research goals for the project. Further, identified the different questions that can be addressed from the dataset selected.
2	11-02-2023	Identified the required dataset and further understood the dataset
3	11-12-2023	Building a dashboard that answers various analytical questions
4	11-19-2023	Creating a rationale for the different design principles and techniques we will be using in creating the interactive dashboard

Note: For the project assignment deadline submission refer to the syllabus. Each week's assignment is for 10 points (except for project assignment 1) and the project assignment is worth 25% of your final grade. The project presentation is worth 5% of your final grade.

TABLE OF CONTENTS

General Guidelines

1. Please use this template document to complete each deliverable assignment.
2. Each assignment must be submitted by the due date in the Course Schedule.
3. All figures should be followed by a brief description about the figure.
4. Figures can be hand-drawn and scanned in some circumstances, but the hand-drawn figure should be clear and legible to obtain full credit. Unclear hand-drawn figures will receive partial credit. For constructing figures and diagrams it is advised to use tools.
5. Figures and tables should have appropriate captions. For documenting and referencing styles please follow the APA or MLA writing style.
6. Please make sure that you provide a reference section.
7. Any material text or figure taken from books, journals or Internet should be referenced. If you have a sentence or a figure that does not belong (authorship) to you, they must be clearly referenced. If you fail to do so your report will be considered as a case for plagiarism. It is your duty to make sure that your report is free from any activity related to plagiarism. Please see the section on Academic Integrity found below. The penalty for plagiarism will be a "0" awarded to your report. Thus, it is good to keep it simple, always have the principle to acknowledge people for their contributions.

Academic Integrity

Academic integrity — scholarship free of fraud and deception — is an important educational objective of Penn State. Academic dishonesty can lead to a failing grade or referral to the Office of Student Conduct.

Academic dishonesty includes, but is not limited to:

- cheating
- plagiarism
- fabrication of information or citations
- facilitating acts of academic dishonesty by others
- unauthorized prior possession of examinations
- submitting the work of another person or work previously used without informing the instructor and securing written approval

In cases where academic integrity is questioned, procedure requires an instructor to notify a student of suspected dishonesty before filing a charge and recommended sanction with the college. Procedures allow a student to accept or contest a charge. If a student chooses to contest a charge, the case will then be managed by the respective college or campus Academic Integrity Committee. If a disciplinary sanction also is recommended, the case will be referred to the Office of Student Conduct.

All Penn State colleges abide by this policy, but review procedures may vary by college when academic dishonesty is suspected. Information about Penn State's academic integrity policy and college review procedures is included in the information that students receive upon enrolling in a course.

Penn State students are expected to act with civility and personal integrity; respect other students' dignity, rights, property, and help create and maintain an environment in which all can succeed through the fruits of their own efforts. An environment of academic integrity is requisite to respect for oneself and others, and a civil community.

For More Information on Academic Integrity at Penn State

Please see the Academic Integrity Chart for specific college contact information or visit one of the following URLs:

- Penn State Senate Policy on Academic Integrity
- iStudy for Success! — learn about plagiarism, copyright, and academic integrity through an educational module
- Turnitin a web-based plagiarism detection and prevention system

Project Assignment 1

Introduction to Tableau (0 points)

Tableau software is one of the leading software tools available for data visualization and exploratory data analysis. In this course, *Tableau will be used as the primary tool* for visualizing one-dimensional data (e.g., temporal data); two-dimensional data (e.g., geospatial data); multidimensional data (e.g., mapping relational data in n-dimensional space); as well as hierarchical tree data. Tableau provides a powerful toolset for identifying patterns and trends in complex datasets as well as creating spectacular visualizations of data from variety of domains (e.g., science, business, engineering, social media, etc.). In addition, Tableau has an intuitive user interface; it is compatible with a multitude of platforms; and easily handles large scale datasets.

While your hands-on work using Tableau will begin in future weeks, your assignment for this week is to download and install Tableau Desktop and begin watching the Tableau training videos to familiarize yourself with the basic features and functionality of Tableau. The Tableau software company provides an extensive suite of tutorial resources that you may find valuable throughout the course. Prior to future weeks assignments, please make sure that you have viewed the training videos listed below.

Instructions for Downloading and Installing Tableau Desktop

Each student should go to the landing page (see link below) to download the most recent version of Tableau Desktop and enter the product key noted below. This key will activate Tableau for the duration of the course. Note: if you already have the latest version of Tableau Desktop loaded on your computer and an active license, there is no need to re-install Tableau Desktop.

1. Download the latest version of Tableau Desktop and Tableau Prep Builder here
2. Click on the link above and select "Download Tableau Desktop" and "Download Tableau Prep Builder". On the form, enter your school email address for Business E-mail and enter the name of your school for Organization.
3. Activate with your product key: **Get the code from the Instructor.**
4. Already have a copy of Tableau Desktop installed. Update your license in the application: Help menu → Manage Product Keys

Note from Tableau: Students can continue using Tableau after the class is over by individually requesting their own one-year license through the Tableau for Students program here . Tableau has also provided you with access to their 'Tableau Prep' software, which is a powerful tool for data processing and cleaning. Although we will not use Tableau Prep in this course, you are welcome to explore the features and functionality of this powerful software tool on your own. If you choose to do so, the activation key above can also be used to activate Tableau Prep.

View the following Tableau training videos at:

Getting Started section, please watch the following 3 videos

(http://www.tableau.com/learn/training#getting-started**):**

- Getting Started (25 min)
- The Tableau Interface (4 min)
- Distributing and publishing (4 min)

- **Under Connecting to Data, please watch the following video:**
  - Data Prep with Text and Excel Files (5 min)

Project Assignment 2

Research Goals and Business Understanding

For modern businesses, comprehending e-commerce data is essential because it offers insightful information that can facilitate well-informed decision-making, improve operational effectiveness, and increase profitability. The capacity to gather, process, and evaluate data is a competitive advantage that can result in substantial economic growth in a time when online purchasing is becoming more and more common. E-commerce data may be utilized for a number of benefits since it provides a window into consumer behavior, product performance, and the effects of different advertising or marketing techniques.

For this project, we have utilized e-commerce operations data for 2020 available on kaggle . We will be able to see which product categories are profitable, how different delivery options impact sales in different areas, which customer segments are based on how they make purchases, and how discount rates impact sales and profitability by using data visualization. We hope to have a thorough grasp of our e-commerce data by responding to these inquiries.

We have identified four business/research queries we would try to analyze using data visualization.

- In 2020, which product categories and subcategories yielded the highest and lowest profits?
- What effects do shipping methods have on sales across different areas?
- Which specific consumer segments may be identified based on their purchasing patterns, and how do they vary in terms of average buy value and discount usage?
- When does giving larger discounts have a detrimental impact on profitability, and does it actually enhance sales?

We plan to use the United States E-Commerce records 2020 dataset from kaggle.[1]

Project Assignment 3

1. **Title:** E-commerce Trend Analysis
2. **Link**: https://www.kaggle.com/datasets/ammaraahmad/us-ecommerce-record-2020/data
3. **Sources**:
    a. Creator: Ammara Ahmad
    b. Donors: Ammara Ahmad
    c. Date: 17 June 2021
4. **Relevant Information Paragraph**: The dataset contains detailed information about United States e-commerce records in 2020. The dataset provided represents an e-commerce transaction log with various columns containing information about each order. It includes details such as the order date, order ID, shipping mode, customer ID, market segment, and geographic information like the country, city, state, postal code, and region. Product-related information consists of the product ID, category, sub-category, product name, sales, quantity, discounts applied, and profit margins. The dataset's columns offer a comprehensive view of the e-commerce business, making it suitable for various analyses, such as sales trends, customer segmentation, and profitability assessments.
5. **Number of Attributes**: 19
6. **Number of Instances:** 3313

Data Dictionary for United States e-commerce records

Variable Name	Data Type	Data Classification	Description	Missing Values
Order Date	Date	Interval	Date when the order was placed.	0
Row ID	Integer	Interval	Unique identifier for each row.	0
Order ID	String	Nominal	Unique identifier for each order.	0
Ship Mode	Categorical	Nominal	Shipping mode for the order (e.g., Standard Class).	0

Customer ID	String	Nominal	Unique identifier for each customer.	0
Segment	Categorical	Nominal	Customer segment (e.g., Consumer, Home Office).	0
Country	Categorical	Nominal	Country where the order was placed	0
City	Categorical	Nominal	City where the order was delivered.	0
State	Categorical	Nominal	State where the order was delivered.	0
Postal Code	Integer	Interval	Postal code of the delivery location.	0
Region	Categorical	Nominal	Geographic region of the order (e.g., East, West).	0
Product ID	String	Nominal	Unique identifier for each product.	0
Category	Categorical	Nominal	Category of the product (e.g., Furniture, Office Supplies).	0
Sub-Category	Categorical	Nominal	Sub-category of the product (e.g., Binders, Furnishings).	0
Product Name	String	Nominal	Name of the product.	0
Sales	Numeric	Interval	Total sales amount for the order.	0
Quantity	Integer	Interval	Quantity of products ordered.	0

Discount	Numeric	Interval	Discount applied to the order.	0
Profit	Numeric	Interval	Profit generated from the order.	0

Project Assignment 4

Dashboard Name: Sales and Profit Analysis Dashboard

Purpose of the Dashboard

The objective of the Sales and Profit Analysis Dashboard is to provide an ecommerce organization with significant findings about sales and profitability patterns. By examining different types of products, delivery options, customer demographics, and the effect of discounts on sales and profitability, it helps users make well-informed choices.

Users

- Executives and Decision-Makers: These users will be able to make data driven decisions based on the dashboard visualizations and insights provided.
- Sales and marketing team: They will utilize the dashboard to enhance their sales approaches and comprehend the effects of discounts.
- Data analysts: These users will go deeply into the data to produce insightful analysis that can be put to use.

Data Sources for Visualizations

The primary data source is the provided dataset with columns such as Order Date, Sales, Profit, Product Category, Shipping Method, Customer Segment, and Discount. Each visualization will utilize different combinations of these variables.

Analytic Questions

Query 1: Profitable and Unprofitable Product Categories

- What product categories and subcategories had the highest and lowest profits in 2020?

Query 2: Impact of Shipping Methods on Sales

- How do different shipping methods impact sales across different regions or areas?
- What effects do shipping methods have on sales across different areas?

Query 3: Identifying Consumer Segments

- Which consumer segments can be identified based on purchasing patterns?
- How do these segments vary in terms of average purchase value and discount utilization?

Query 4: The Impact of Discounts on Profitability

- When does offering large discounts have a detrimental impact on profitability?
- Does providing larger discounts actually result in increased sales?

Tableau visualizations

- **Profitable and Unprofitable Product Categories**
  - Visualization: Bar chart showing product categories and subcategories with profits in 2020.
  - Relevant Columns: Order Date, Product Category, Sub-Category, Profit.
  - Queries Answered:
    - What product categories and subcategories had the highest and lowest profits in 2020?
- **Impact of Shipping Methods on Sales**
  - Visualization: Geographic map or heat map showing regions with sales data, colored by shipping methods.
  - Relevant Columns: Country, Region, Shipping Method, Sales.
  - Queries Answered:
    - How do different shipping methods impact sales across different regions or areas?
    - What effects do shipping methods have on sales across different areas?
- **Identifying Consumer Segments**
  - Visualization: Segmentation chart (e.g., clustered bar chart) showing purchasing patterns by customer segments.
  - Relevant Columns: Customer Segment, Sales, Discount.
  - Queries Answered:

- Which consumer segments can be identified based on purchasing patterns?
- How do these segments vary in terms of average purchase value and discount utilization?

❖ **The Impact of Discounts on Profitability:**

➢ Visualization: A line chart showing the relationship between discount levels and profit margins.

➢ Relevant Columns: Discount, Profit.

➢ Queries Answered:

- When does offering large discounts have a detrimental impact on profitability?
- Does providing larger discounts actually result in increased sales?

Project Assignment 5

1. **<u>Profitable and Unprofitable Product Categories</u>**

Queries addressed

- What product categories and subcategories had the highest and lowest profits in 2020?

Visualization: A bar chart showing product categories and subcategories with profits in 2020.

Rationale

- Bar Chart: When comparing profitability between several product categories and subcategories, a bar chart is effective.
- Sorted Bars: To show which categories have the biggest and lowest profits, we will sort the bars according to profit.

- Time Filter (2020): We will apply a time filter to focus specifically on profits in 2020.
- Color Coding: Color will be used to differentiate between categories that are profitable and those that are not.

Design Principles

- Tufte's Principles:
    - Data Density: Presenting as much information as possible without sacrificing clarity.
    - Maximize Data-Ink Ratio: In order to reduce visual clutter and concentrate on effectively communicating information, we shall remove any excess ink or build the visualization with minimal ink usage.
- Few's Principles:
    - Data-Ink Ratio: Prioritizing the data-ink ratio to ensure that the ink used on the visualization conveys meaningful information.
    - Avoid Chartjunk: Eliminating non-data ink and embellishments that do not contribute to understanding.

2. Geographic Sales Data Colored by Shipping Methods

Queries addressed

- How do different shipping methods impact sales across different regions or areas?
- What effects do shipping methods have on sales across different areas?

Visualization: Geographic map or heat map showing regions with sales data, colored by shipping methods.

Rationale

- Map Visualization: Maps are useful for regional analysis because they give a clear geographical overview.
- Color by Shipping Method: We plan to utilize varied colors to represent various shipping methods. This allows for quick identification of the impact of shipping methods on sales.
- Interactive Tooltip: When users hover over a particular area, we want to include a tooltip with information about sales and shipping options.
- Legend: Adding a legend to clarify the meaning of different colors.

Design Principles

- Tufte's Principles:
    - Small Multiples: We would think about comparing regional sales using small multiples for various shipping options
- Few's Principles:
    - Clarity and Simplicity: We would prioritize clarity and simplicity in the visualization design to enhance comprehension

3. Identifying Consumer Segments

Queries addressed

- Which consumer segments can be identified based on purchasing patterns?
- How do these segments vary in terms of average purchase value and discount utilization?

Visualization: Segmentation chart (e.g., clustered bar chart) showing purchasing patterns by customer segments

Rationale

- Clustered Bar Chart: Appropriate for comparing various metrics among various user groups.
- Color-coded Segments: We plan to allocate distinct colors to each customer segment for clarity.
- Interactive Filters: To enable users to interactively filter data based on product categories, time periods, or additional relevant characteristics, we would implement interactive filters.
- Annotations: We plan to add annotations to highlight specific observations or trends

Design Principles

- Tufte's Principles:
    - Sparklines: Sparklines or other tiny visuals would be added to the dashboard to convey patterns and context in an easy-to-read manner.
- Few's Principles:
    - Avoid Distortion: In order to prevent distortion, we would make sure that the data representation appropriately reflects the underlying values.

4. **<u>The Impact of Discounts on Profitability</u>**

Queries addressed

- When does offering large discounts have a detrimental impact on profitability?
- Does providing larger discounts actually result in increased sales?

Visualization: Line chart showing the relationship between discount levels and profit margins.

Rationale

- Line Chart: Useful for showing patterns across a continuous variable, like the correlation between profit margins and discount rates
- Dual-Axis: For ease of comparison, we want to show profit margins and discount levels on the same chart using a dual-axis.
- Annotations: We will use annotations to emphasize key points or highlight specific data.
- Interactive Tooltip: When users hover over data points, an interactive tooltip will appear, offering further information.
- Reference Lines: Reference lines will be there to indicate break-even points or thresholds.

Design Principles

- Tufte's Principles:
    - Data Density: With a detailed representation of the data, we hope to let users quickly and effectively draw conclusions.
- Few's Principles:
    - Show the relevant data: Emphasize the importance of displaying the actual data rather than relying solely on unnecessary statistics.

General Design Principles to be followed:

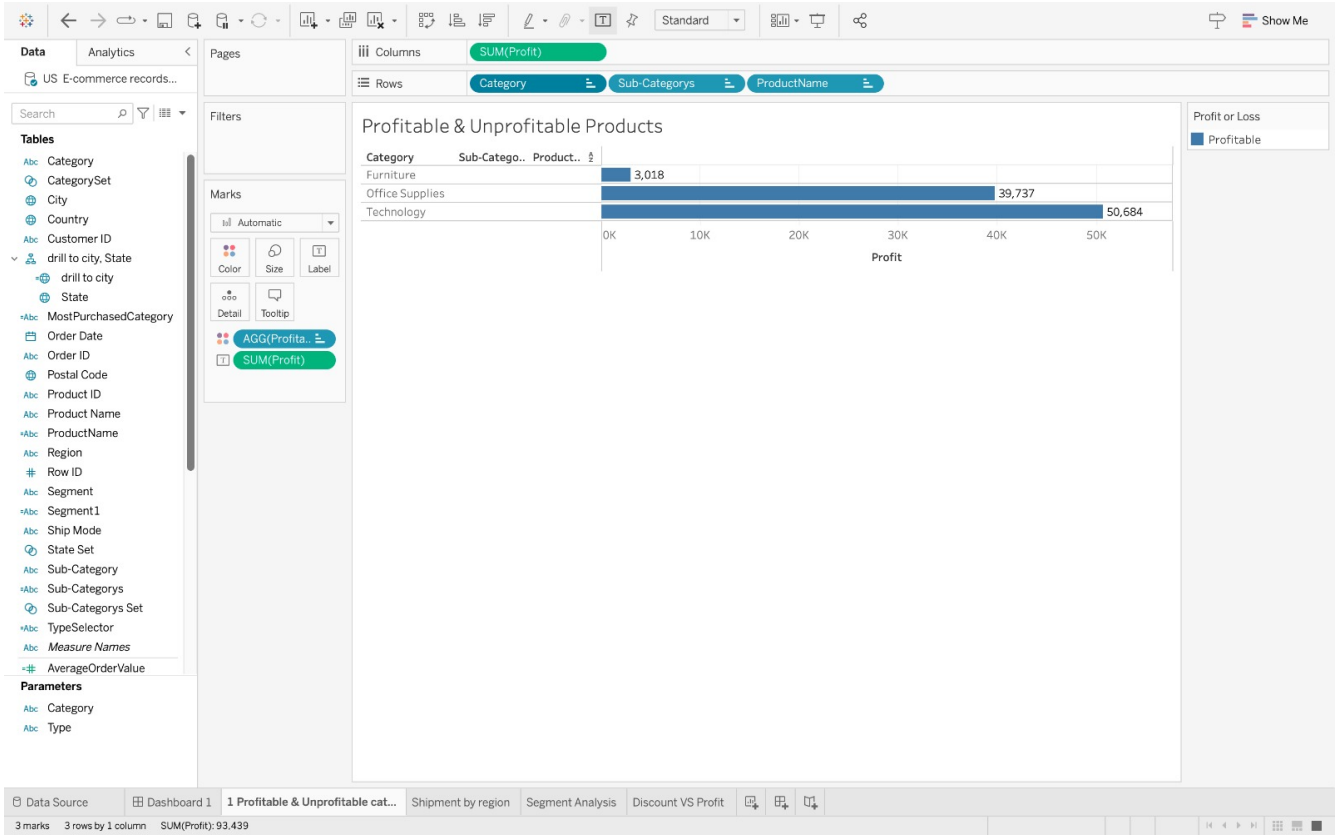
- Consistent Color Scheme: For uniformity and readability, we would keep the color scheme of each visualization the same.
- Interactive Filters: We aim to include filters so that users may drill down into particular facts or tailor the view according to their preferences.
- Clear Titles and Labels: To deliver the desired message, we will make sure that each visualization includes titles and labels that are succinct and clear.
- Data-Driven Storytelling: Our goal is to construct a narrative inside the dashboard that leads consumers through the story the data is conveying.
- Context is Key: Through comments, reference lines, or more textual material, we would provide context to improve understanding.

- Effective Use of Color: We will purposely use color to emphasize important ideas and communicate information while avoiding obtrusive or distracting color schemes.
- Consistency and Standardization: We will adhere to standards and maintain uniformity in design components to increase customer familiarity.

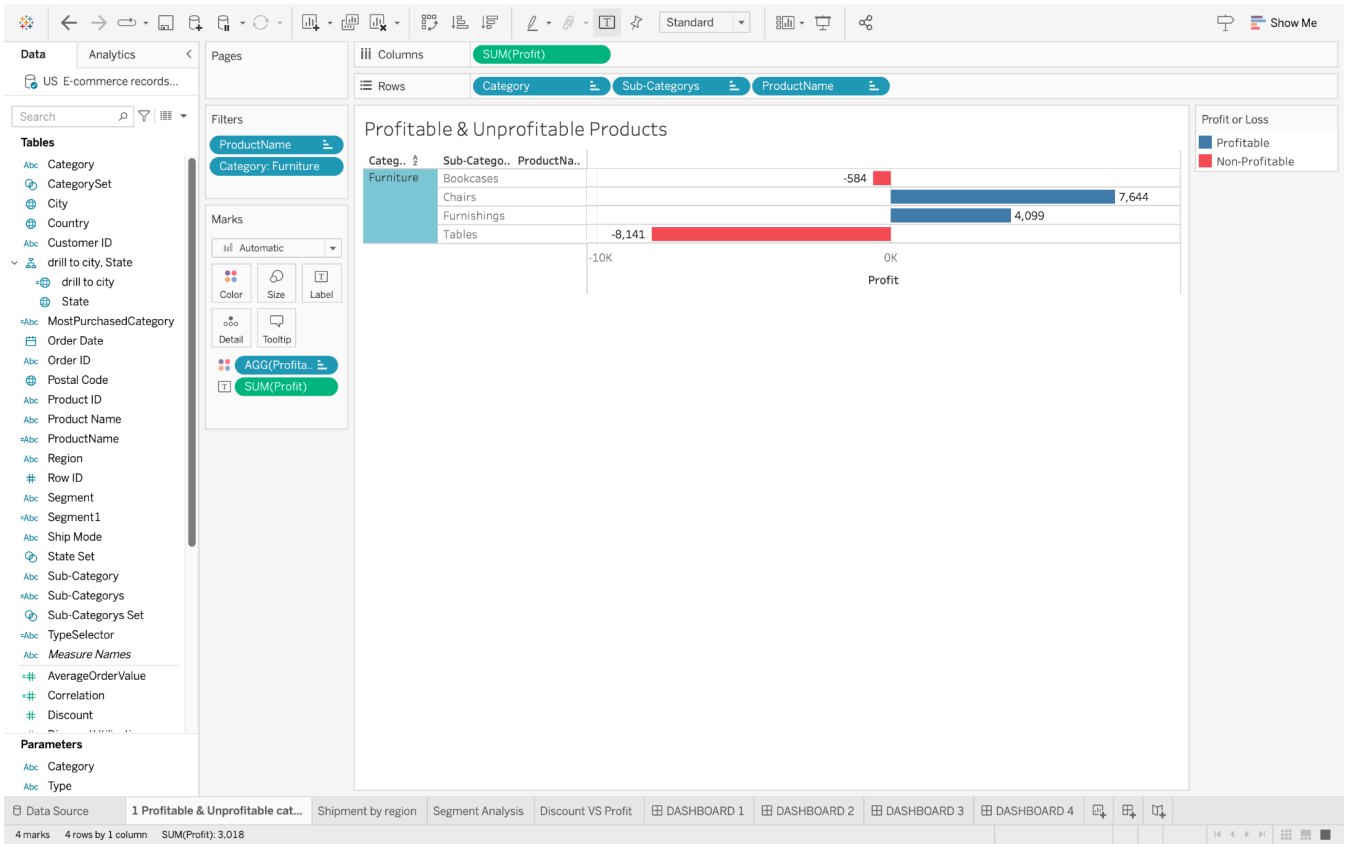
Project Assignment 7

USE CASE 1:

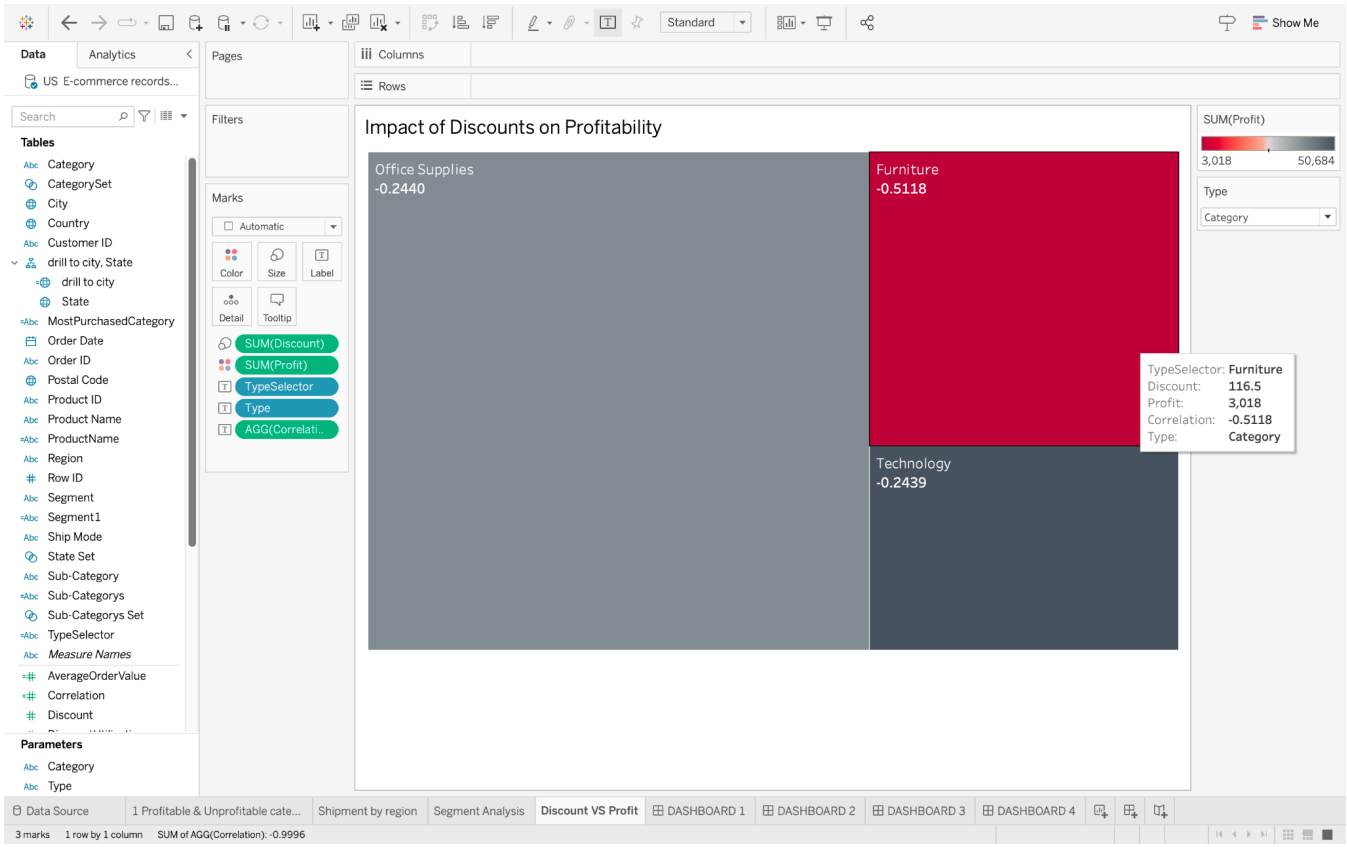
Assessing Category-Wise Profitability and Analyzing the Impact of Discounts on profitability for a specific category - **Furniture**



- In the visualization above we can see profit value for all the three categories and by selecting the category furniture, we can see the profit and non profit values for all the different sub-categories under the same category.



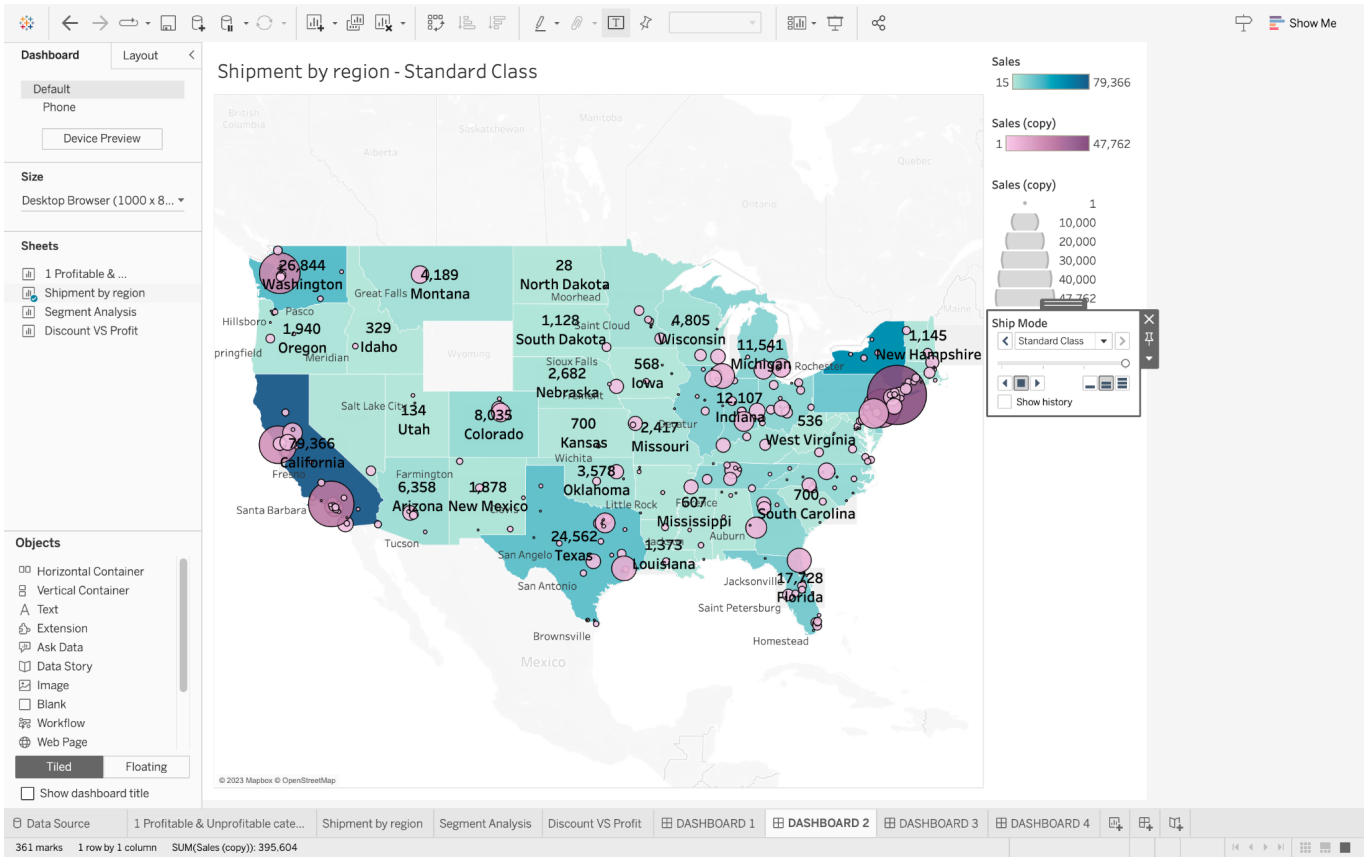
- We can observe that two of the sub-categories under furniture are profitable, which is indicated by a blue bar and two are non-profitable, which is indicated by a red bar.
- Now we can see further how discount impacts the profitability of this category from the last dashboard.



- In this visualization also we can choose the type from the drill down menu to choose category
- We can see that in the category wise tree map between discount and profit there is a negative correlation for the category furniture, which means that on increasing discount the profit rates decrease.
- The lowest range of profit is represented by red colour and higher profits are represented by grey colour.

USE CASE 2:

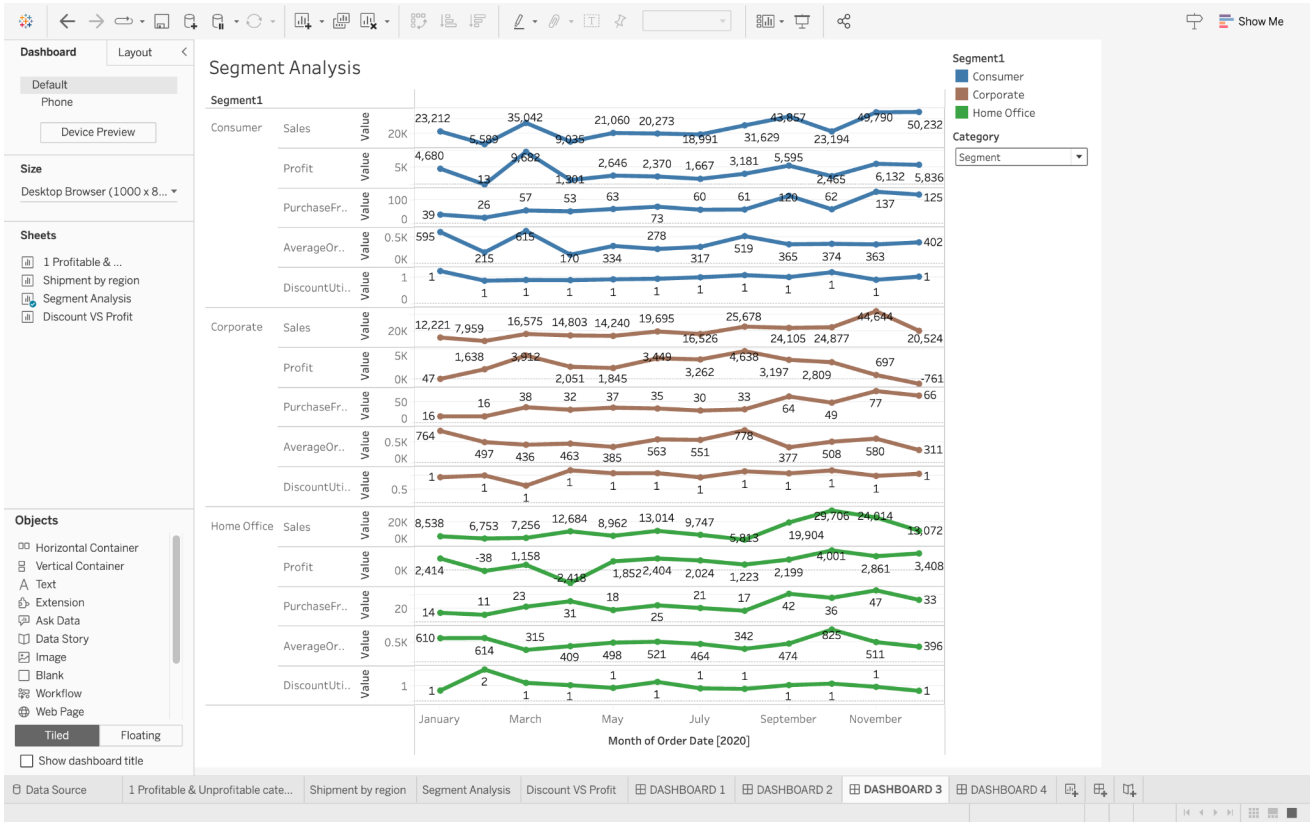
Analyzing Product Shipments and Sales by Region for a specific shipment category - Standard Class



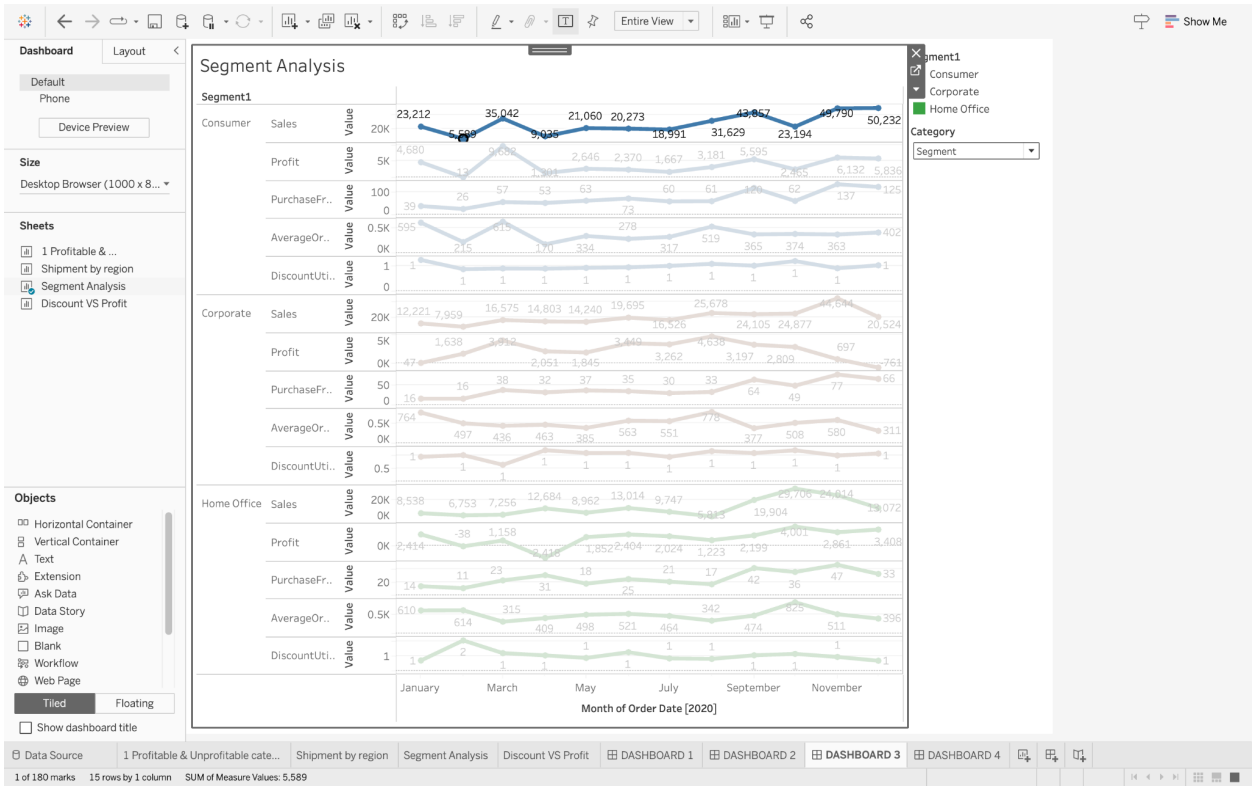
- In the above visualization we can see a drop down menu from which we can select different shipping modes and as we select one mode we can see the sales on the spatial chart.
- Here we have selected the shipping mode as Standard Class and the graph shows the sales based on states as well as cities.The blue gradient filter is used to represent sales for states and we can see that California has the highest sales as it is coded in dark blue shade, followed by New York and so on when the shipping mode is standard class.
- The pink gradient bubbles are used to represent sales for various cities and we can clearly see from the chart that New York City has the highest sales in this shipment mode as is represented by the large pink bubble and so on.The size of the bubbles also indicates a gradient relationship.

USE CASE 3:

Analyzing various consumer segments across multiple metrics



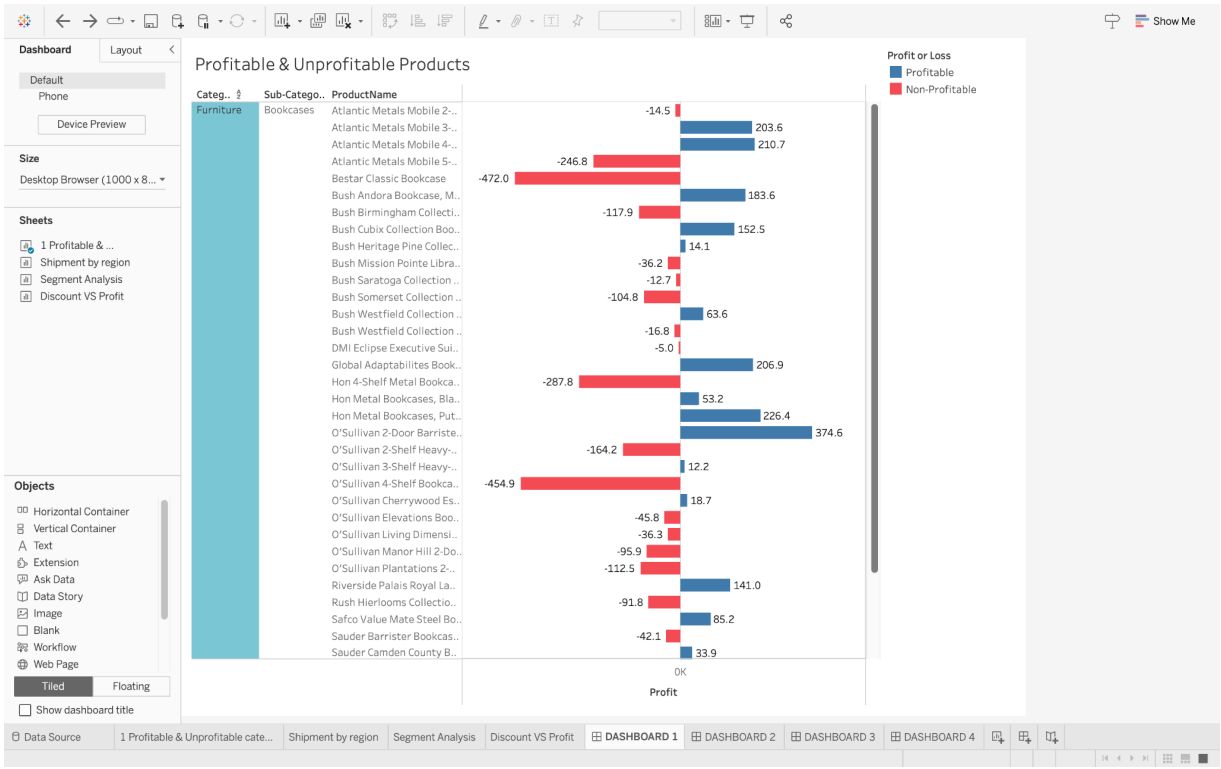
- In the above visualization we can see a drop down menu which allows users to filter the data by segment and region. Here we are analysing three customer segments - consumer, corporate and Home Offices.
- The line charts represent the trends for five different categories- sales, purchase frequency, average order value, discount utilization, and profit, with respect to month of order date.
- All the three sectors are coded in different colour like blue represents consumer, brown represents corporate and green represents Home office.



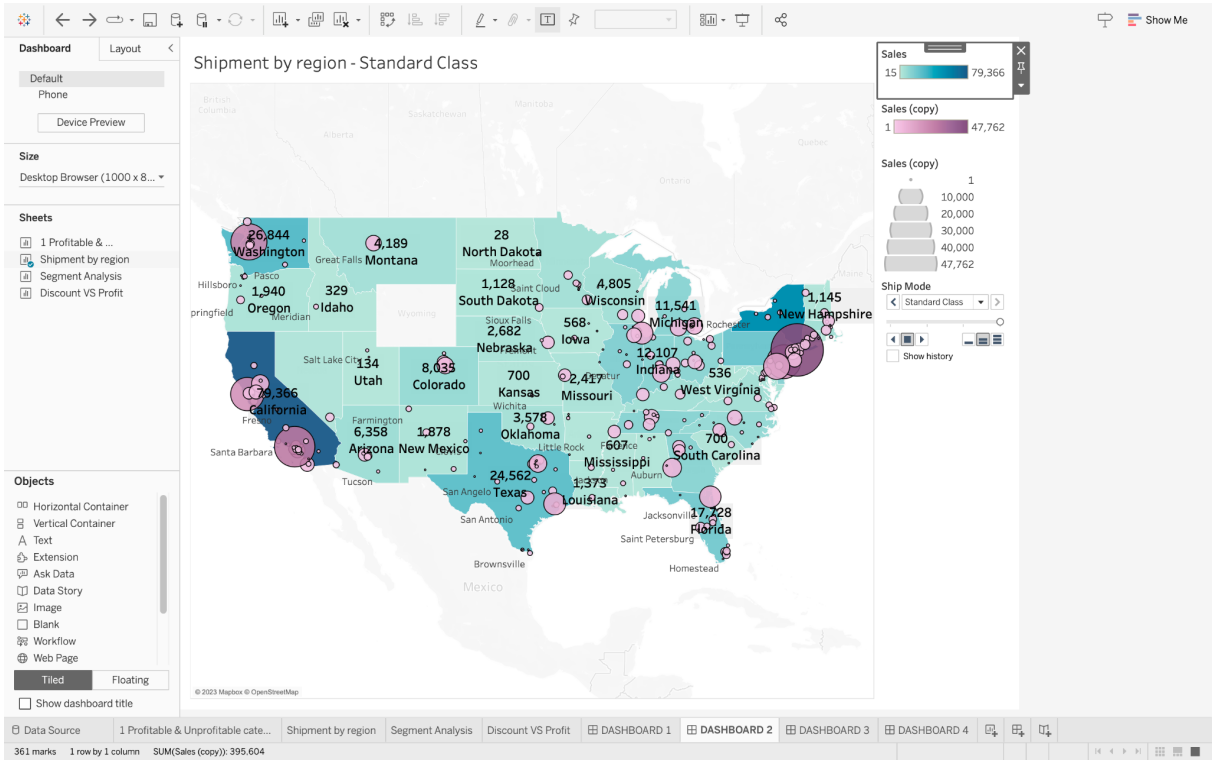
- Now if we want to check the sales trend for the consumer segment over time, this chart is very useful to observe trends.
- In the above visualization we can see that the sales are highest (50,252) for the month of December 2020 and lowest (5589) in the month of February 2020.

- Business users can use this dashboard to identify the impact of discounts on profitability. This information can be used to optimize pricing and discount strategies.
- We can observe the comparison for any specific category (for example- sales) for all the three segments.

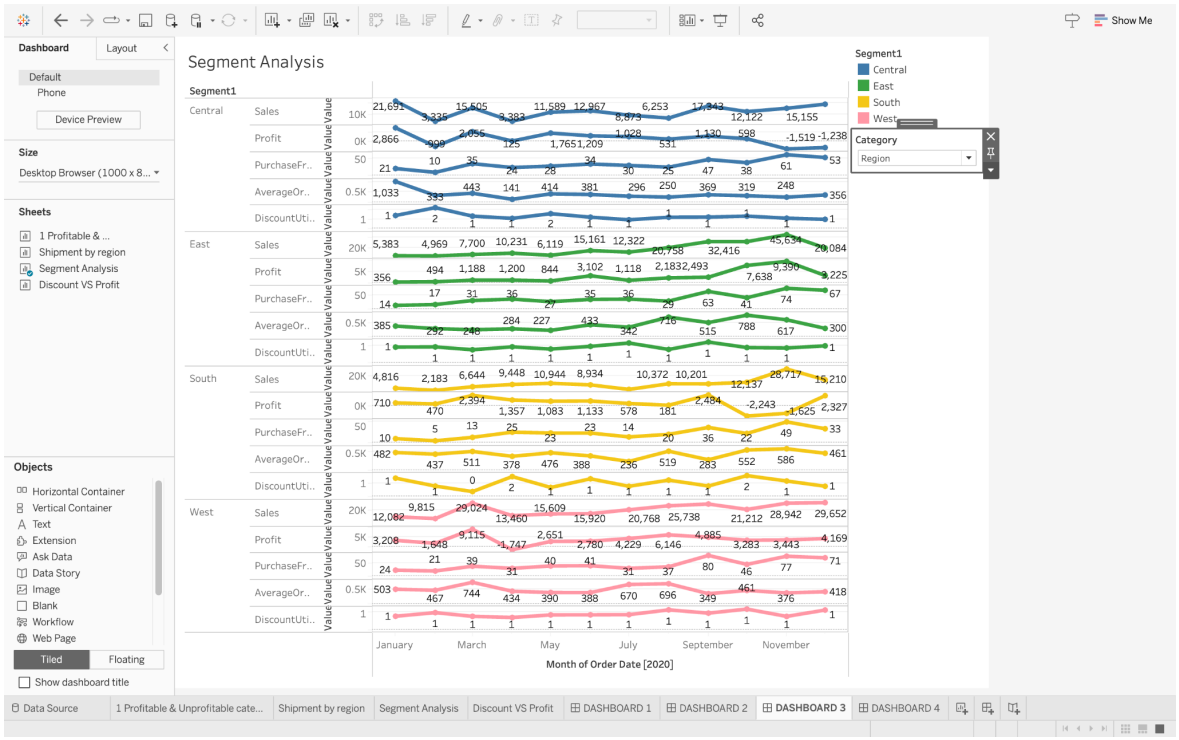
Dashboard 1:



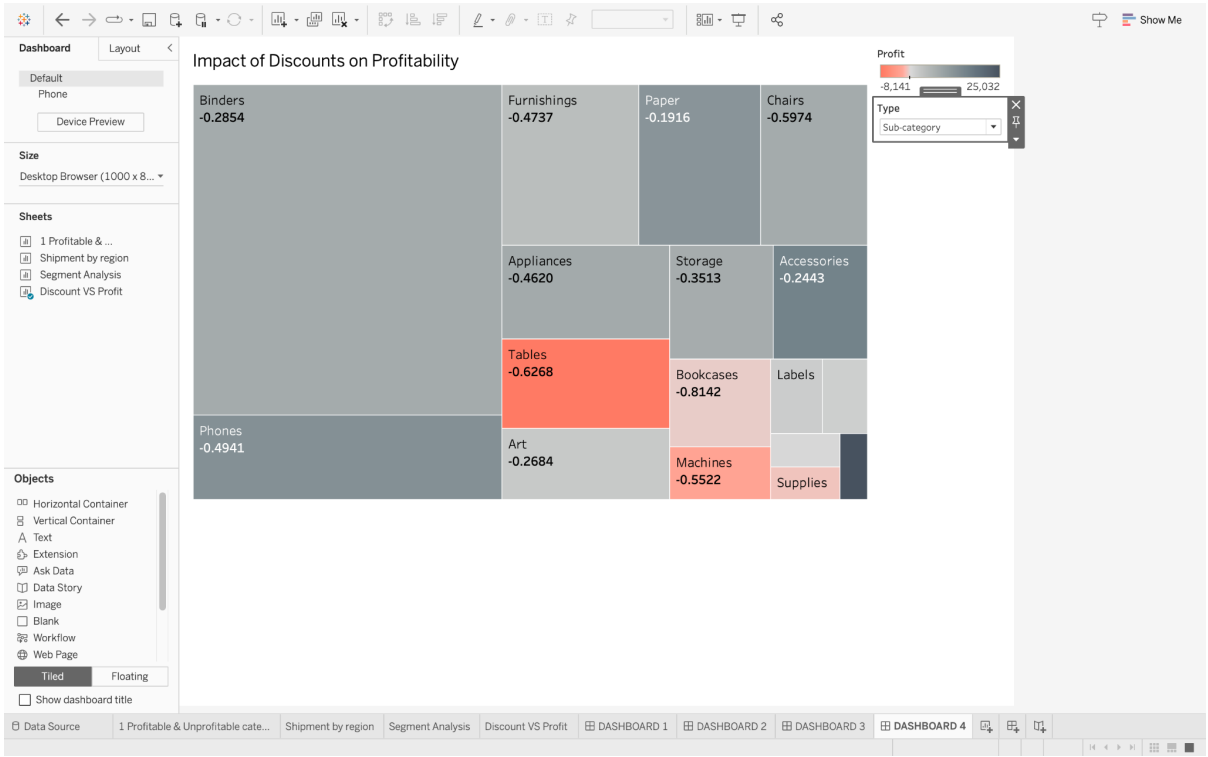
Dashboard 2:



Dashboard 3:



Dashboard 4:



References

- AMMARAAHMAD. “United States E-Commerce records 2020.” *Kaggle*, https://www.kaggle.com/datasets/ammaraahmad/us-ecommerce-record-2020. Accessed 29 October 2023.