Walmart Sales Analysis

Dauren Omarov

Demetrios Daniel

Yougender Chauhan

California State University,

Los Angeles

Submitted to Dr. Shilpa Balan

TABLE OF CONTENTS

INTRODUCTION

Walmart is one of the world's biggest retail corporations, with approximately 10,500 stores and clubs in 20 countries. It employs 2.1 million associates worldwide, with roughly 1.6 million in the United States alone. Walmart serves millions of customers daily with its diverse product offering and competitive pricing. As a consequence, the business collects massive amounts of data on sales, inventory, and customer behavior.

Data analysis is essential for understanding its performance based on insights and trends found in its data. In this assignment, we will use Tableau, a powerful data visualization tool, to explore and analyze the Walmart Sales dataset. The Walmart dataset includes weekly sales data for 45 Walmart stores in the United States over the course of the year 2019. It also includes profit information for the retail giant, including data on product categories, discounts, stores, regions, and time periods. By analyzing this data using Tableau, we can gain valuable insights into Walmart's sales patterns, trends, and performance over time. By using Tableau, we can create meaningful visualizations that help us communicate our findings effectively.

We chose this Walmart dataset because working with such a large and successful company is always fascinating. It teaches the analyst and, more importantly, the reader about how this company still retains a strong competitive spot within its industry. We wanted a dataset we could explore in such a way that describes current patterns, shows contributing factors, and overall allows for descriptive analysis.

Descriptive analysis is a useful tool for gaining an overall understanding of the Walmart Sales dataset. This type of analysis involves summarizing and visualizing the data to reveal key characteristics, such as the distribution of sales and profits, the most popular product categories, and the top-performing stores and regions. Not only that, but we can also group similar elements

into one category, make calculations such as average, and use the newly formed variables for further analysis. By creating interactive visualizations using Tableau, we can explore this data in greater detail and uncover insights that might not be apparent through simple numerical summaries.

Overall, the objective of this project is to provide a deeper understanding of Walmart's sales and profitability trends in a way that the findings can be effectively communicated to stakeholders and people who have the need to interact with the data. Another goal is to showcase the importance of data analysis in the retail industry and how it can be used as a reference while making informed decisions. We plan to do that by exploring the sales trends, identifying factors that influence sales, and creating actionable insights. By using Tableau to create interactive visualizations and reports, we can enable users to explore the data and gain insights that are relevant to their needs.

DATA DESCRIPTION

Dataset Url: https://www.kaggle.com/datasets/shinjihirako/walmart-dataset-2019

The data set contains ‘24022’ rows and ‘32’ columns from the Walmart sales year 2019 and has the following data attributes:

order_id	This column contains the unique order identification number.
order_info_id	It is part of order_id where it has the two-letter code followed by the year.
order_id_number	It is part of order_id where it contains the unique 4-7 digit number.
return_reason	The Reason selected for returning a product. It has two values, either “not given” or “not returned”.
order_date	The full date of the time the order was billed.
order_weekday	The day the order was billed.
order_month	The month the order was billed.
ship_date	The full date the order was shipped.

ship_mode	Method of shipping, for example, same day, standard, second class, etc.
product_id	Unique id of the product in Walmart database.
days_to_ship	The number of days it took for the product to be shipped from the date of order.
category	The category in which the product falls under. In this dataset there are three, i.e., technology, furniture, and office supplies.
sub_category	Further classification of the categories within the three main categories.
product_name	The name of the product listed on the website.
unit_cost	The price of the product listed on the website.
sales	Total Sale amount for the products Walmart sold over the year 2019.
sales_volume	This field has two values, High and low, where high is if profit > 300$.
profit	The amount of profit made per total sales of product sold

	over the year 2019.
profit_margin	The percentage of profit made per sale.
quantity	The number of the product customer ordered.
discount	The discount percentage for the product.
discount_over_30	Marked as ‘1’ if the discount percentage is over 30% for the product.
region_id	It consists of the IDs of the region
postal_code	The postal code where the order was set for.
city	The city for which the order was set.
state	The state for which the order was set.
sub_region	The subregion is the central US.
salesperson	The name of the Walmart employee who processed the order placed.
customer_id	It is a unique string with 2 characters followed by numbers assigned to the customer account.
customer_name	It consists of the name of the customer who has placed the

	order.
segment	It is divided into three customer segments i.e. consumer, corporate, and home office.
Country	The country for this dataset is the US.

DATA CLEANING

1. Filling in the null values:

   For the column “Country” we have few null values. Since this dataset is for the country United States only, we went ahead and filled all the empty cells with “ United States”.

   Before Data cleaning:

AE	AF
Consumer	United States
Consumer	United States
Consumer	
Home Office	
Home Office	United States
Consumer	
Corporate	United States
Consumer	United States
Corporate	
Consumer	United States

After Data cleaning**:**

AE	AF
Consumer	United States
Consumer	United States
Consumer	United States
Home Office	United States
Home Office	United States
Consumer	United States
Corporate	United States
Consumer	United States
Corporate	United States
Consumer	United States

2. Filling null values with Average:

For the column profit we have a few null values since our dataset has about 24k rows in which the null value for this is 2611. we went ahead and filled it with the average of the whole column which is “$1.24”.

Before Data cleaning:

R	S
profit	profit_margin
$ 0.19	0.02%
$ -	0.00%
$ 0.05	0.28%
$ 0.05	0.10%
$ 0.10	0.75%
$ -	0.00%
$ -	0.00%
$ 0.05	0.22%
$ -	0.00%
$ 0.24	0.10%
$ 0.10	0.15%
$ -	0.00%
$ 0.05	0.09%

After Data cleaning**:**

R	S
profit	profit_margin
$ 0.19	0.02%
$ 1.24	7.09%
$ 0.05	0.28%
$ 0.05	0.10%
$ 0.10	0.75%
$ 1.24	1.01%
$ 0.10	0.63%
$ 0.05	0.22%
$ 1.24	4.97%
$ 0.24	0.10%
$ 0.10	0.15%
$ 1.24	2.18%
$ 0.05	0.09%

3. Changing Date Format: For the "order_date" column change the order of date from dd/mm/yy. We changed the format to mm/dd/yyyy as per the standard format.

Before Data cleaning:

order_date	order_month
18/12/2019	December
26/12/2019	December
18/12/2019	December
27/12/2019	December
30/11/2019	November
18/11/2019	November
6/12/2019	December
26/12/2019	December
25/3/2019	March

After Data cleaning**:**

return_reason	order_date
Not Returned	12/18/2019
Not Returned	12/26/2019
Not Returned	12/18/2019
Not Returned	12/27/2019
Not Returned	11/30/2019
Not Returned	11/18/2019
Not Returned	12/6/2019
Not Returned	12/26/2019
Not Returned	3/25/2019
Not Returned	1/13/2019

4. Changing 1,0 to yes and no:

For the column " discount_over_30" it had a '1' if the discount percentage for that product was over 30% else a '0'. We changed the 1 to a "yes" and 0 to a "no"

Before Data cleaning:

discount	discount_over_30
0.4	1
0.4	1
0.4	1
0.4	1
0.35	1
0.4	1
0.35	1
0.35	1
0.3	0
0.35	1
0.3	0
0.3	0

After Data cleaning**:**

discount	discount_over_30
0.4	yes
0.4	yes
0.4	yes
0.4	yes
0.35	yes
0.4	yes
0.35	yes
0.35	yes
0.3	no
0.35	yes
0.3	no
0.3	no

5. Changing months number to text:

For the column “order_month” it was in the numeric formate i.e 1-12 for each month which is a bit confusing. we changed the 1-12 as January to December respectively for better understanding and readability of data.

Before Data cleaning:

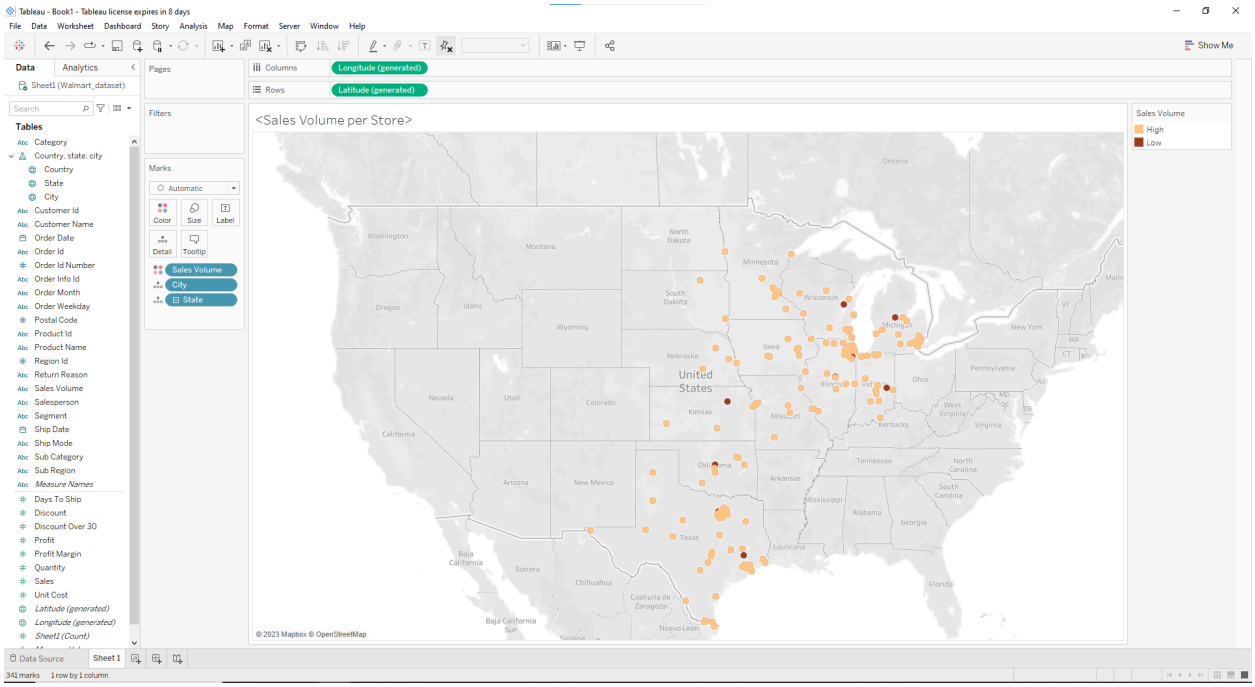
order_month	ship_date
12	12/23/2019
12	12/31/2019
12	12/23/2019
12	12/31/2019
11	12/5/2019
11	11/19/2019
12	12/11/2019
12	12/31/2019
3	3/29/2019
1	1/18/2019
3	3/23/2019

After Data cleaning**:**

order_month	ship_date
December	12/23/2019
December	12/31/2019
December	12/23/2019
December	12/31/2019
November	12/5/2019
November	11/19/2019
December	12/11/2019
December	12/31/2019
March	3/29/2019
January	1/18/2019
March	3/23/2019
March	3/26/2019
December	12/20/2019
December	12/26/2019

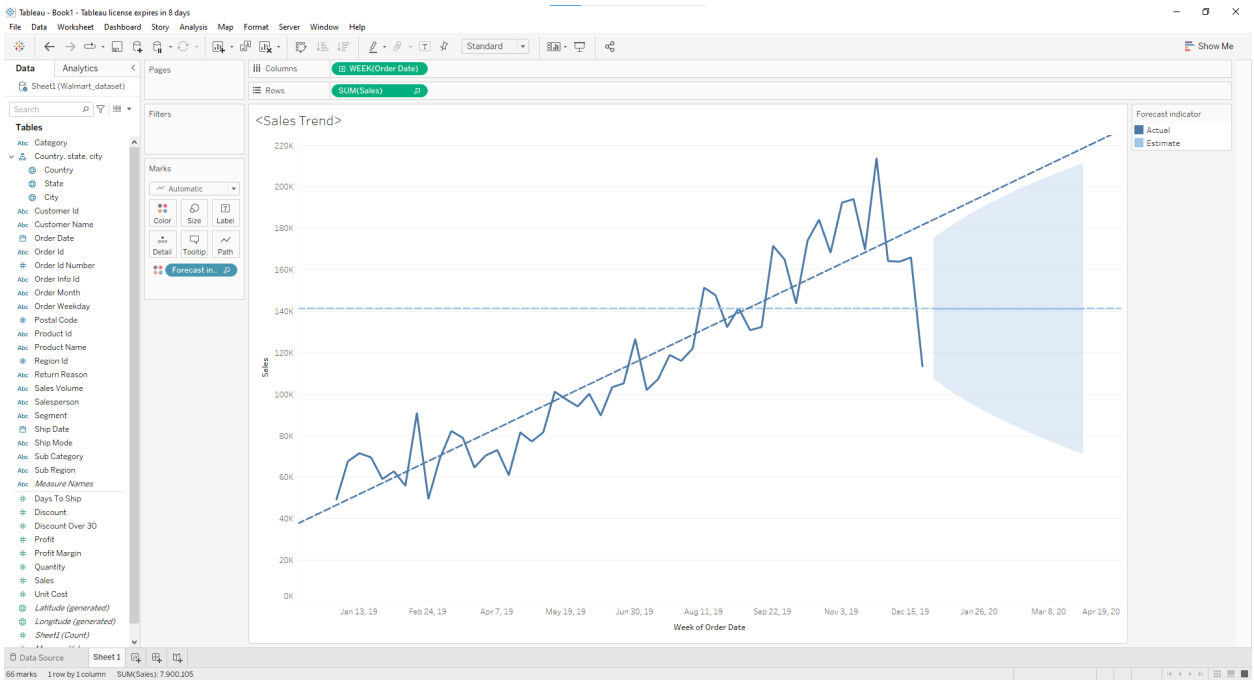
DATA VISUALISATION:

1) Which stores had the highest volume of sales across cities? (Geographical Maps)



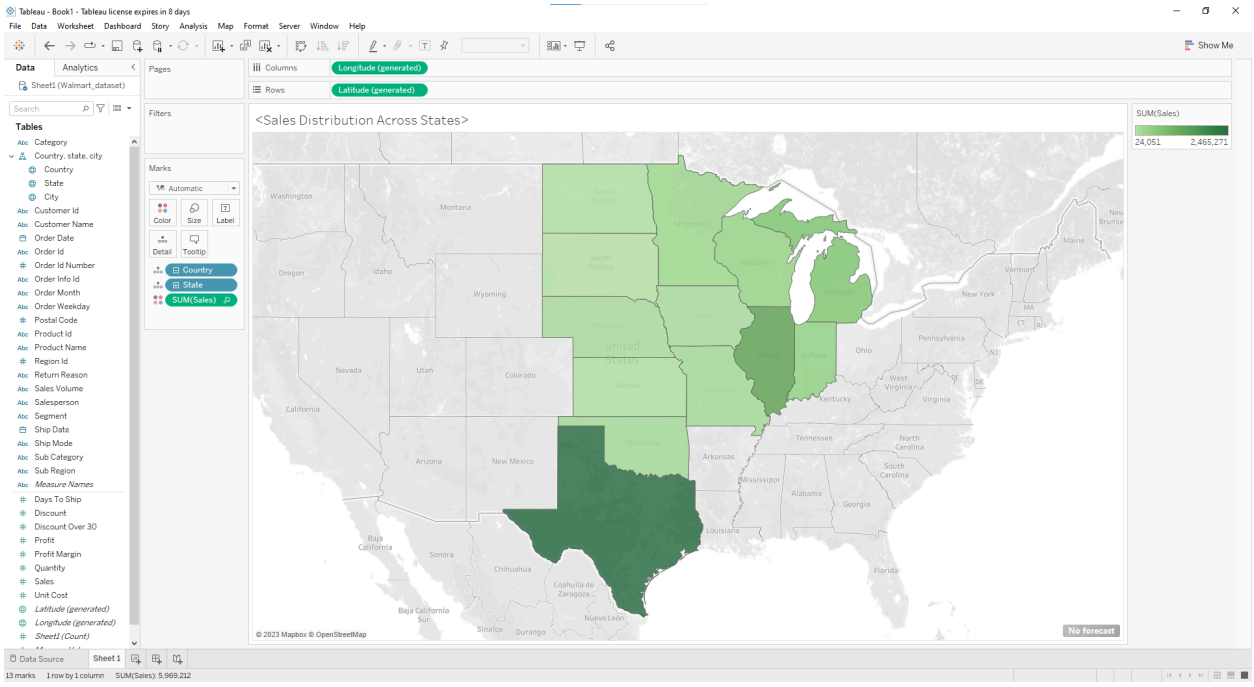
The visualization shows the store locations in every city in the form of dots on the geographical map. In addition to the store location, each store is marked a different color according to the sales volume. We can see that some stores had a higher sales volume than others, with the yellow dots representing high sales volume, and the red dots representing low sales volume. Namely, the low-sales stores appear to be in the following states/cities: Wisconsin (Appleton), Indiana (New Castle), Michigan (Glenview, Orland Park, Tinley Park, Normal), Kansas (Manhattan), Oklahoma (Edmond), and Texas (Conroe, Keller).

2) How do the sales of Walmart change over time? (Trend Line)



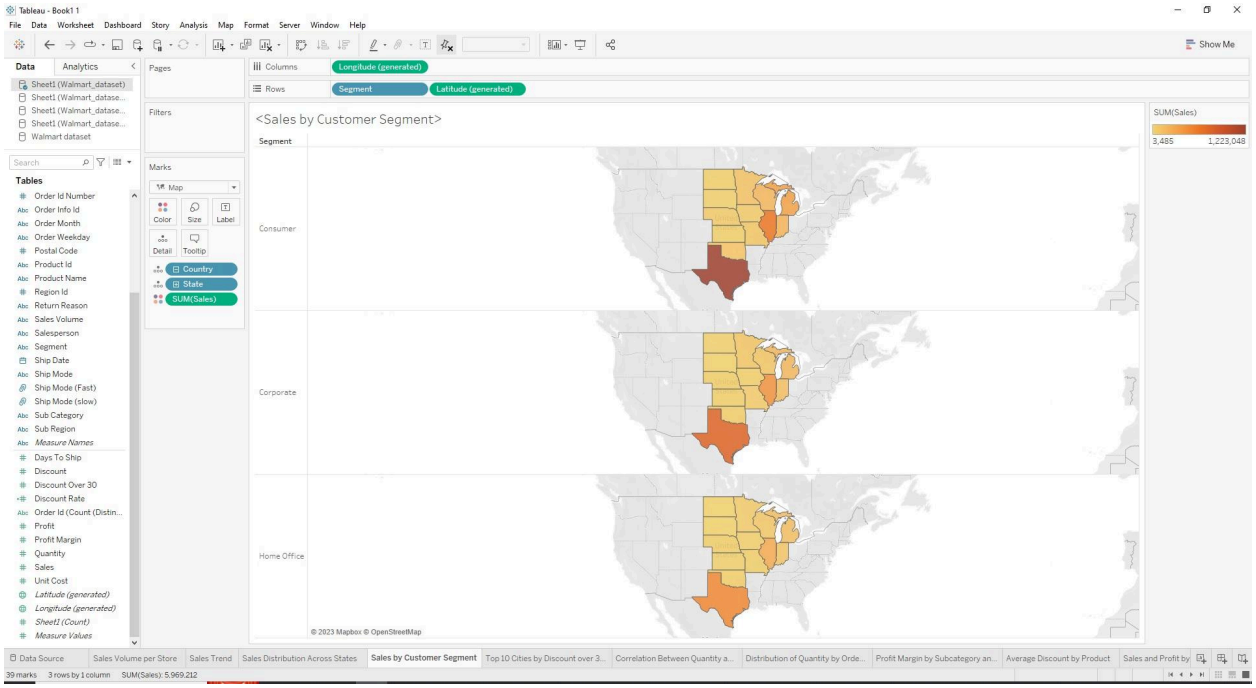
This visualization shows the sales sum each month of the year 2019. We can see that although there are some fluctuations between the early months of the year, the sum of sales is always increasing with the difference between September - November being the highest ($213,738). The peak of sales appears to be in the month of November, likely due to the Christmas holiday in the month after. In this graph, we also have a reference line to compare our sales with which is a benchmark value of $141,327. The highlighted area represents the confidence level and idens up accordingly. We see the further into the future, the wider the line, hence the more range of values, hence the more uncertain it becomes.

3) What is the distribution of sales across stores in different states? (Geographical Maps)



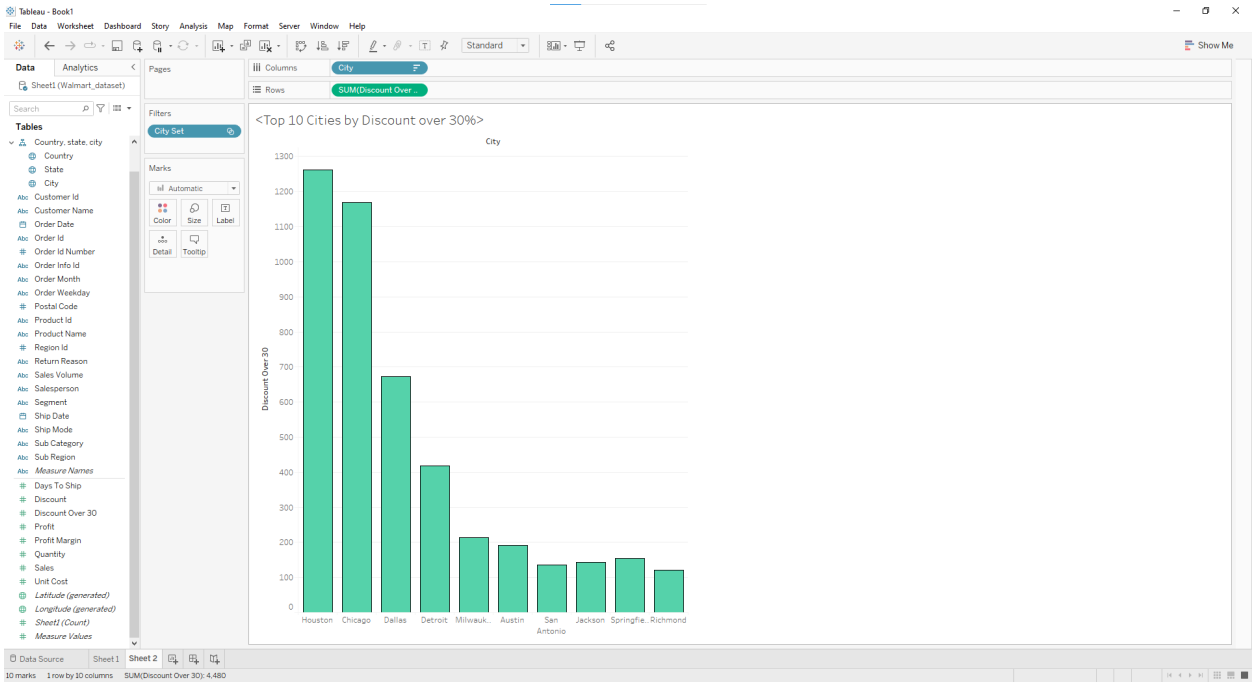
In the above visualization, a geographical map is utilized to showcase the weekly sales distribution by state. The map is used in conjunction with the colour function from the marked pane to show the reader instantly which states experience the most weekly sales with a more intense shade of the colour while the ones with the least amount of weekly sales with less shade of the colour. In this instance, we can see that the state of South Dakota generated the least amount of sales with a result of $24,051, while the state of Texas generated the highest amount of sales with a result of $2,465,271.

4) What is the distribution of sales by state and customer segment? (Dual Axis)



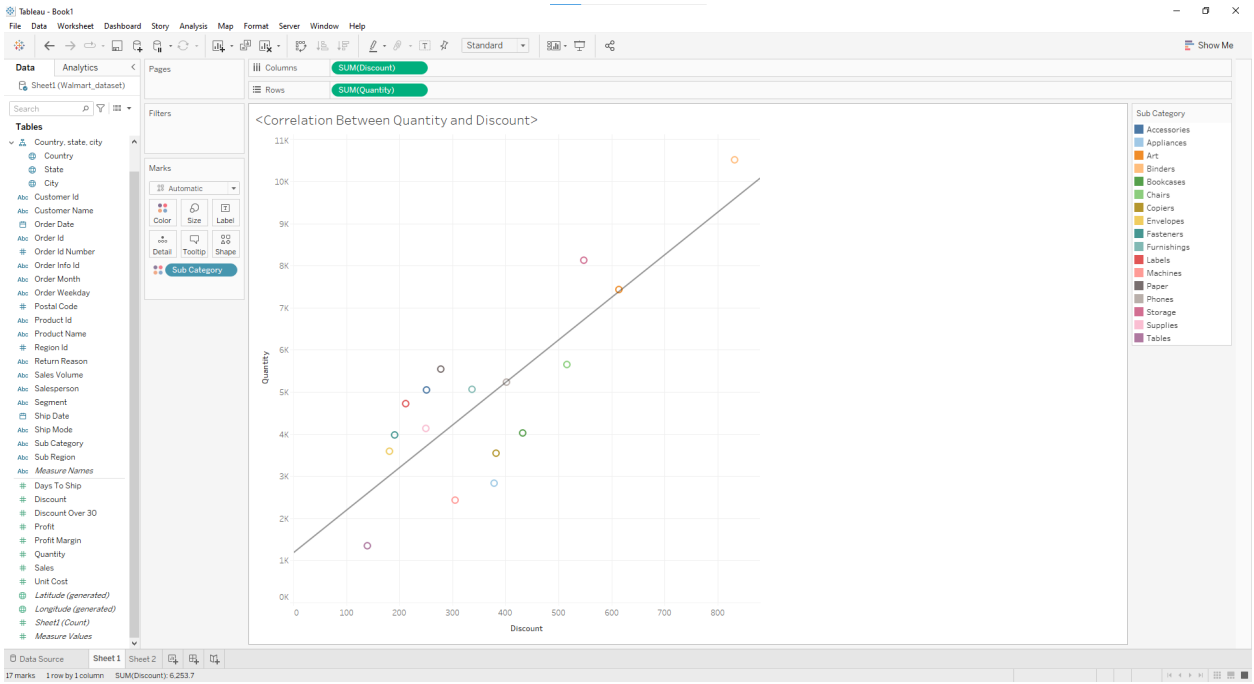
In this visualization, we have a dual-axis representation of sales by state and customer segment. As we can see the consumer segment is by far the one where most sales are distributed to and specifically, the state of Texas with 1,223,048 sales. The least amount of sales is distributed to the state of South Dakota in the Home Office segment with 3,485 sales.

5) Which stores had the highest number of transactions where a discount over 30% was offered? (Ranks)



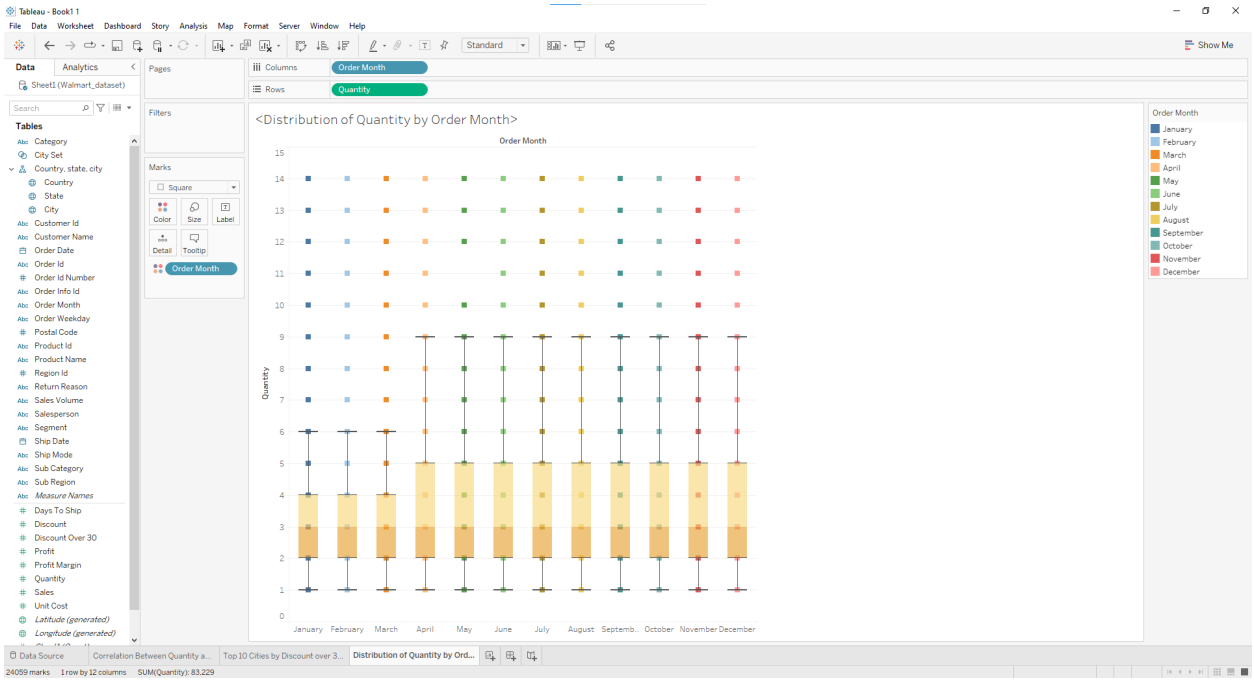
As we can see from the above visualization, the Houston and Chicago stores had the highest number of transactions, where a discount of over 30% was offered. This result came after we ranked them by a list of the Top 10 in descending order. With 1261 and 1168 transactions, respectively, this could be due to the population of the area. In any case, Walmart can use these numbers to its advantage and push more deals at those stores or include these stores in its list of promoted products at all times.

6) Is there any correlation between product quantity and discount offered? (Scatterplot, Trend Line)



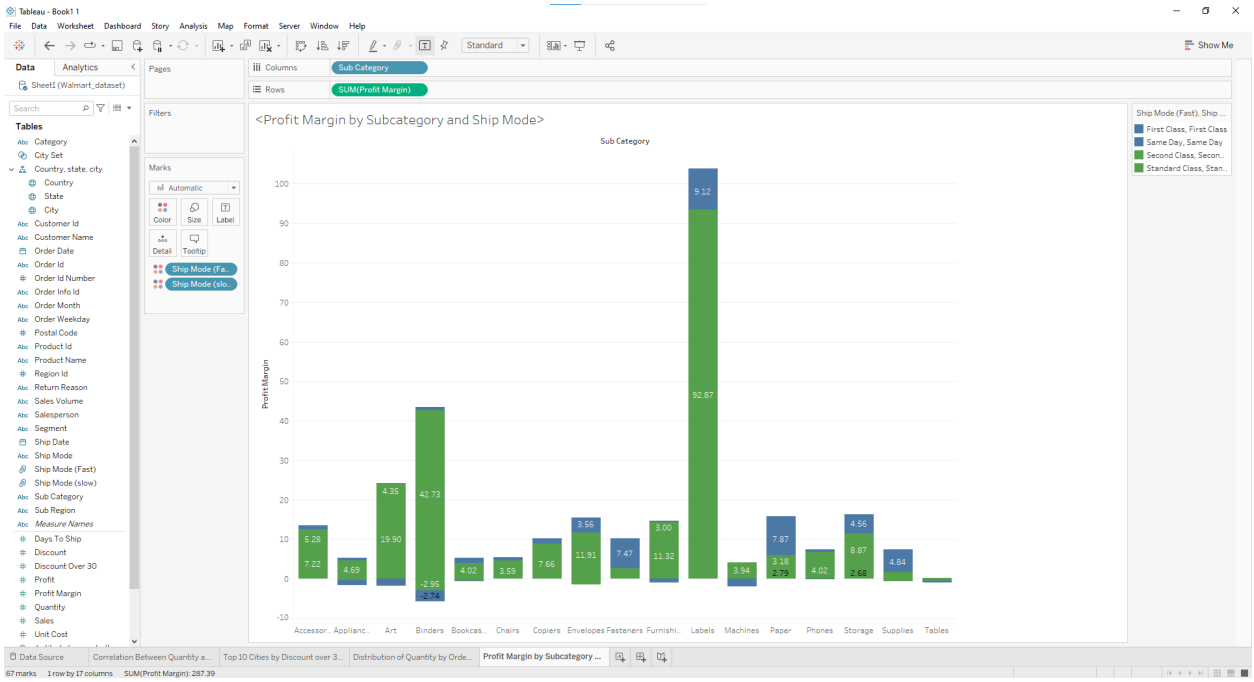
Based on the above visualization, we utilize a scatter plot to see that the discount offered increases alongside the product quantity. This fact is even more evident by our trend line that showcases this linear relationship. We have also added our product sub-categories in color to pinpoint exactly which ones get discounted the most and in which quantities. In this case, the Binders subcategory has the most discount applied at 832.3 with a quantity total of 10,512. The table product sub-category has the least discount applied with 139.9 and a total quantity of 1,351. This visualization should be used in conjunction with others, such as the highest profit margin by subcategory to reveal other aspects of its profitability that Walmart can capitalize on by adjusting its discount strategies.

7) What is the distribution of quantity by order month? (Box-and-Whiskers)



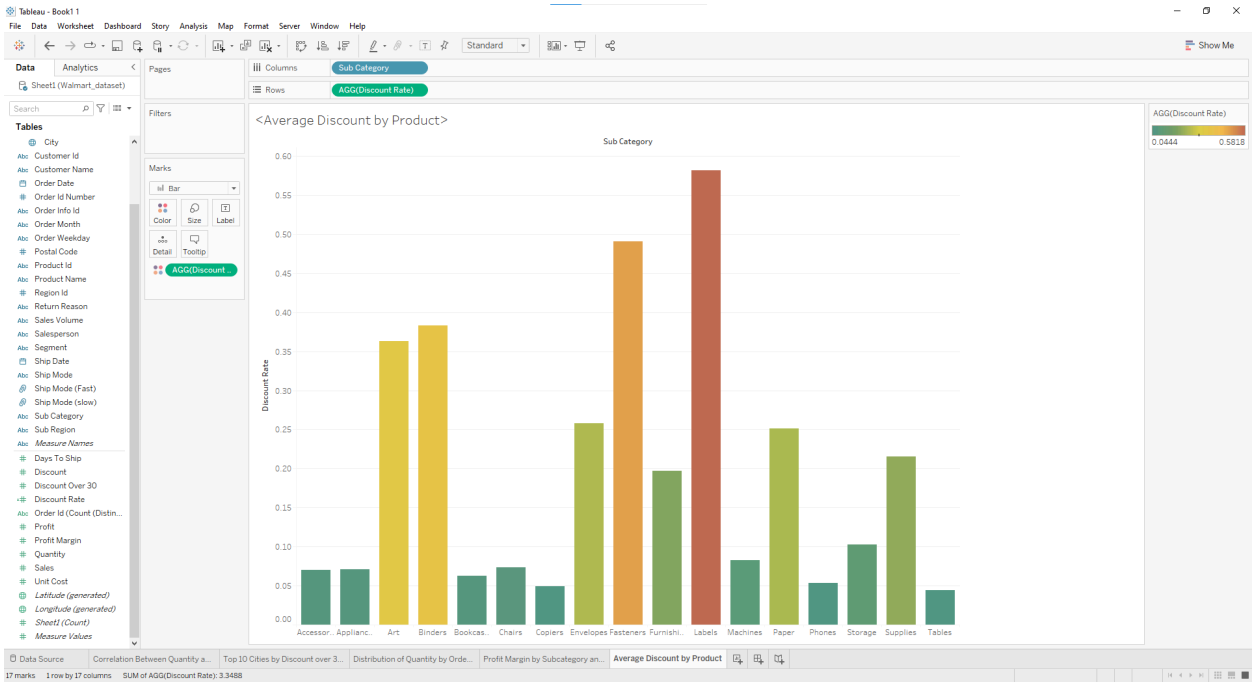
Here, in the visualization, we have utilized a box-and-whiskers box to show the distribution of quantity per order month. We can see that in the first 3 months of the year (Jan-Feb-Mar), the product quantity ordered is a lesser amount than the rest of the year from April to December. More specifically, we can see that most of the ordered quantity during the first 3 months falls within the 2-4 range, with less being 1,5 and 6, and the rest being outliers. As for April to December, we see that most of the quantity ordered is within the 2-5 range, less being ordered at 1,6,7,8,9, and the rest of the product quantity being outliers. This is likely because last year's Christmas holiday season was right before this sales analysis took place. Consuming entities may need time to save a disposable budget for the new year.

8) What is the profit margin by product subcategory, and how does it differ between orders shipped through faster and slower shipping modes? (Groups)



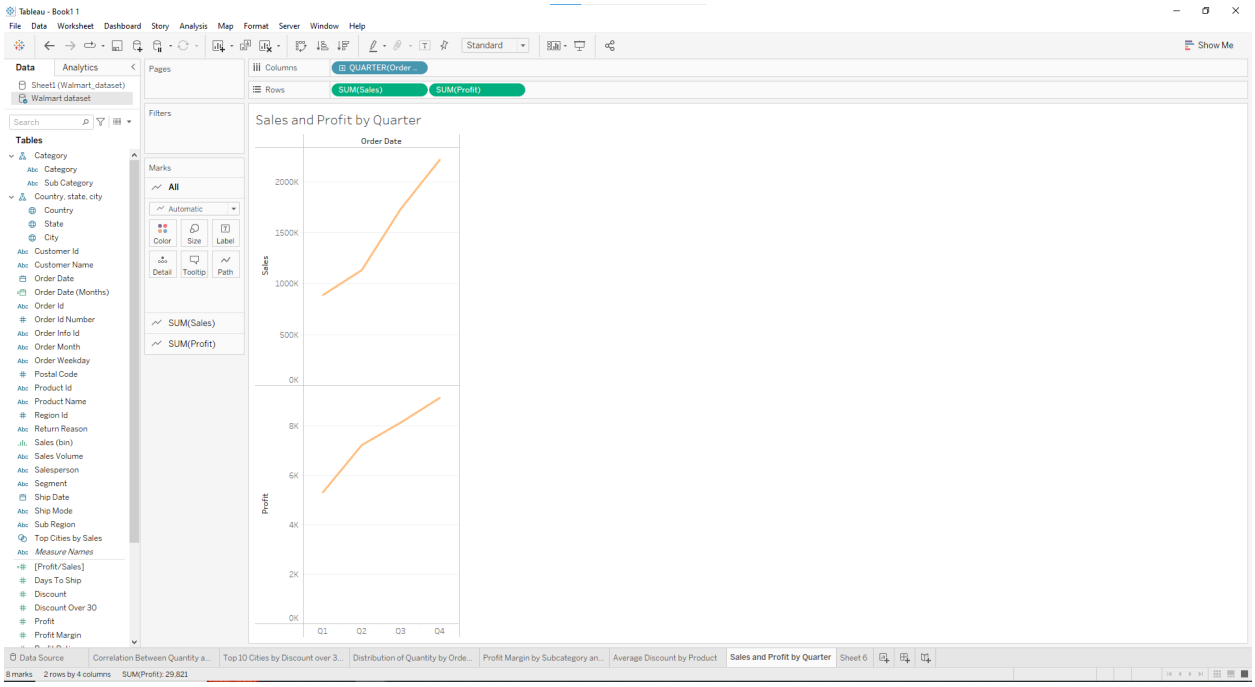
To explore the above question, we created this bar chart visualization by applying groups to the Ship Mode measure variable. To make the visualization easier to understand, we grouped the faster shipping modes (same day and first class) and colored them blue. Similarly, we grouped the slower shipping modes (second class and standard class) and colored them green. We can see that labels are the most profitable subcategory and that, by far, standard class is the most preferred shipping mode and accounts for most of the product's profit margin at 92.7.

9) What is the average discount rate for each product subcategory? (Calculated Field - Average)



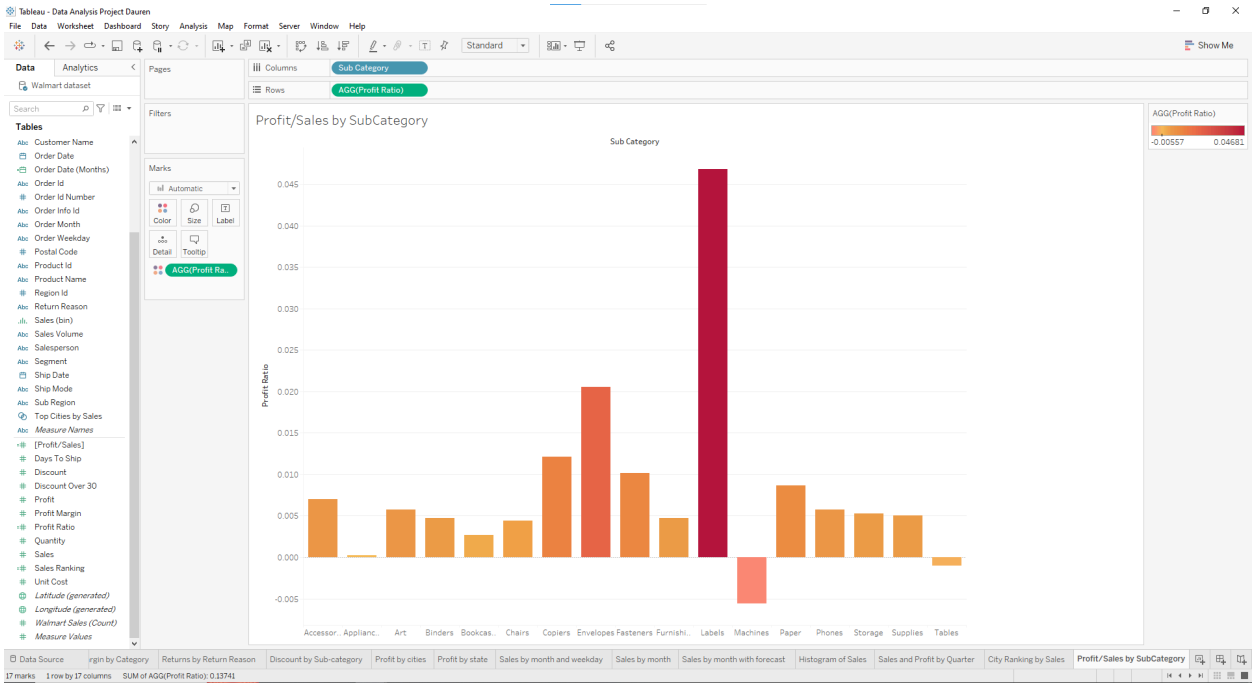
To answer the above question, we created a calculated field for the discount measure to create an average discount number represented in the new measure called ‘Discount Rate.’ By using our subcategory data we can see which products had the most discount applied to them so far. In this case, the Labels products seem to have the most discount applied, with 0.5818 applied on average, with fasteners not too far behind with 0.4908 applied on average. Tables had the least discount rate applied on average, with just 0.0444. This helps Walmart understand which products get most of the discount rate applied, however, it becomes really meaningful when it gets combined with other visualizations to truly make decisions and understand relationships.

10) What is the trend of profit and sales over the quarters of the year 2019? (Dual Axis, Trend Line)



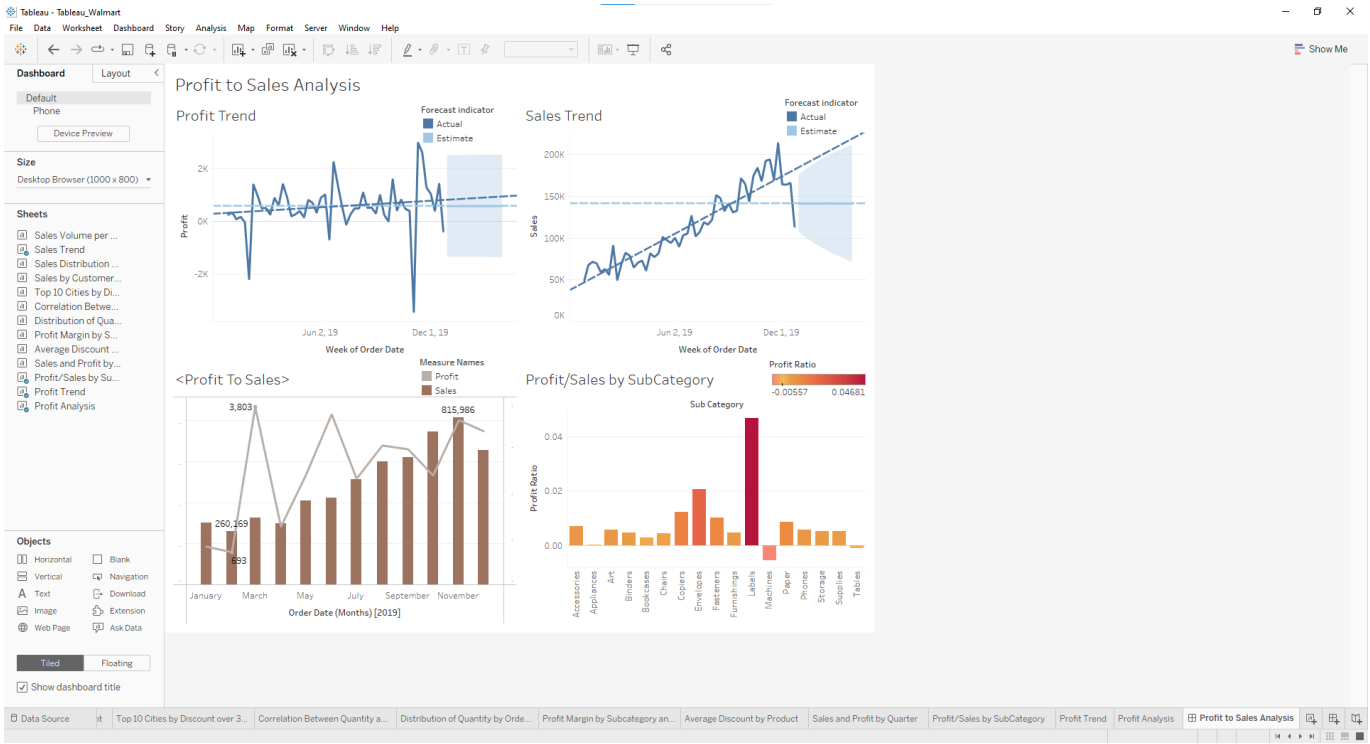
In the above visualization, we can see profit and sales represented with a trend line in a dual axis. We can see that both profit and sales increase as the quarters pass by, with the most profit being recorded at the end of the year with the amount of $9,138. Since this is a linear relationship, when sales increase, so does profit. Profit is the dependent variable in this relationship, so when the highest profit number is recorded at the end of the year at the 4th quarter, so does the highest sales number with the amount of $2,215,316.

11) What is the profit ratio across different subcategories of products? (Calculated Field: Ratio)



To explore the above question, we created another calculated measure to form a ratio of sales: profit across all product subcategories. This is an extremely useful visualization, not just for Walmart but for any retailer. This really shows the organization whether or not selling a current product currently yields any profit and takes action accordingly. In this case, we can see that selling Labels yields the most profit, with envelopes far behind in terms of this ratio. What Walmart should take not of, however, is the loss of profits by selling Machines and Tables and investigating why that is the case.

DASHBOARD



Our Dashboard is a representation of Walmart's sales and profit analysis. To accurately represent that, we used the Sales Trend sheet from earlier and a new Forecast Trend Line for Profit. By placing them next to each other, we gain valuable insight into how they fluctuate over the order weeks. As we see, sales increase with relatively small fluctuations but always keep on increasing. Profit, however, fluctuates a lot more by looking at the actual trend line in the Profit Trend chart. In fact, Walmart has been in the negative in terms of Profit for 3 weeks in the year 2019. Namely, February 3rd, June 16th, July 14th, and November 3rd are the biggest losses (-$3,439). The biggest profit mark was achieved on November 10th, for the amount of $2,996, which is a week after the biggest loss. This could be due to a big expense in the form of a sizeable marketing investment that ultimately ended up paying off shortly after. Considering the

holidays are right around the corner, it could have something to do with pushing holiday-exclusive products or deals.

Looking at the Sales Trend line, we can see that sales take a small dip between the 2 last weeks of December, likely signifying the transition from December 15th to December 22nd. This brings us closer to Christmas and the holiday season, which reflects the drop in sales by the end of December 22nd, just 3 days before Christmas, when most of the shopping has concluded.

Another graph we used is the Dual Axis Chart of Sales/Profit to SubCategory, where it was evident that Machine and Table sales have negatively impacted profit. We did that by calculating the sales-to-profit ratio and gained valuable insights about profit that we were not able to pinpoint otherwise with our previous graphs.

Lastly, we created another Dual Axis Chart utilizing a bar chart for Sales and a Trend Line for Profit by examining it with the Order Month dimension. We wanted to use Order Month and not Order Week for this graph to get an overall view of the months of the year. We can see that the lowest sales ($260,169) go along with the lowest profit marks ($693), as both seem to be in the month of February. In terms of the highest sales they achieved in the month of November ($815,986) and the highest profit in March ($3,803).

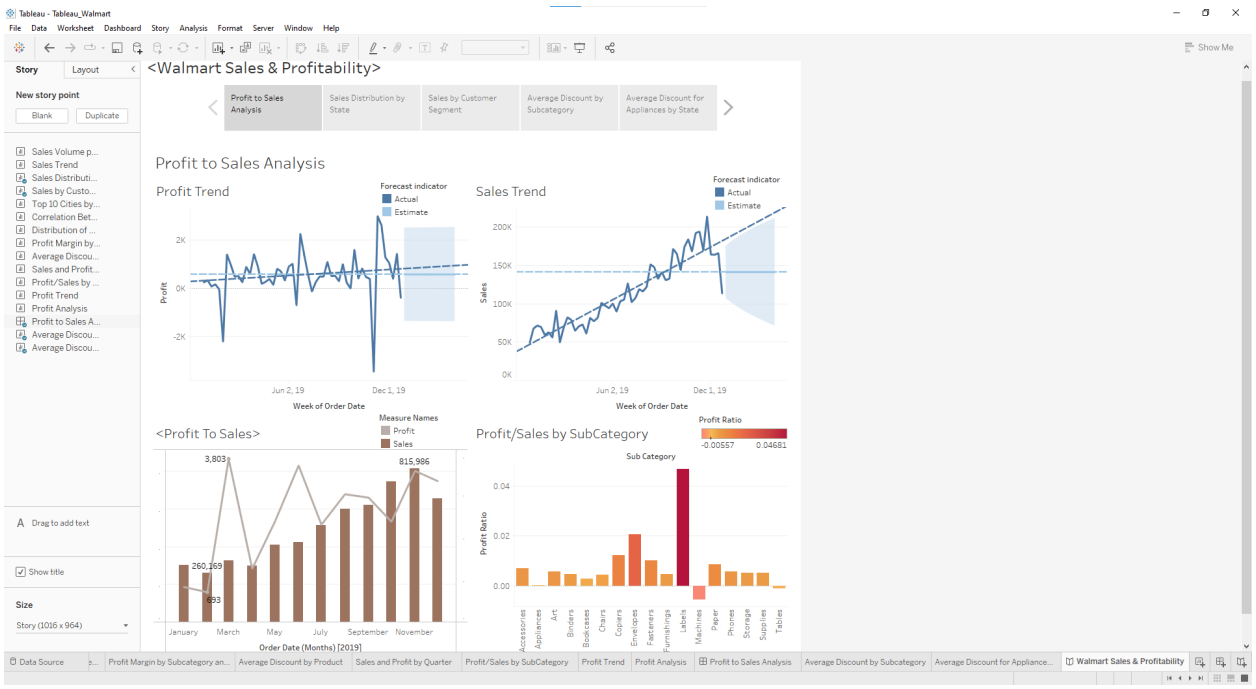
Overall, the Profit Trend additional visualization helps us see how profit fluctuates through the weeks of the year, and the additional Sales and Profit by Order Month helps us see how sales and profit get shaped monthly. The insights given when using all those graphs together on our Dashboard were that sales seem to gradually increase with small fluctuations while Profit follows a more uneasy distribution on the graph. This could be due to various investments and expenses on the organization’s part. The ratio visualization helps us understand which products yield the most profit through sales and which do not. In this way, we were able to identify Tables

and Machines as not being profitable. Finally, the last visualization completes the picture by binding the most important elements together in a bar/line graph chart. We have sales, profit, and order month and how they fluctuate through the year of 2019, which can help further explore relationships and variable dynamics and adjust strategies for effective competition.

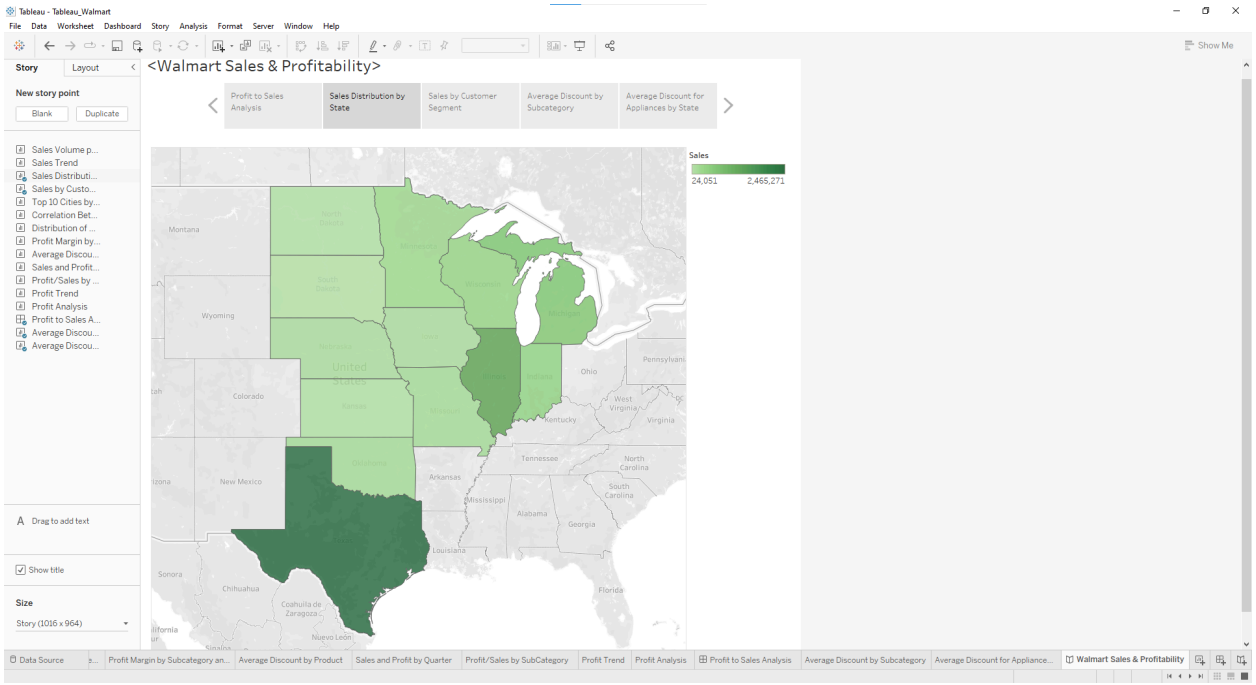
STORY TELLING

In Tableau, the story function allows the user to arrange visualizations in a sequence. This can be done in such a way that it allows for a story to be told. However, it does not necessarily need to serve a narrative purpose only, but it can also help provide context, showcase how certain actions lead to specific outcomes, lay out a strong case, or help understand relationship dynamics. The story is a sheet made up of other sheets generated throughout the analysis, with each individual sheet being a story point.

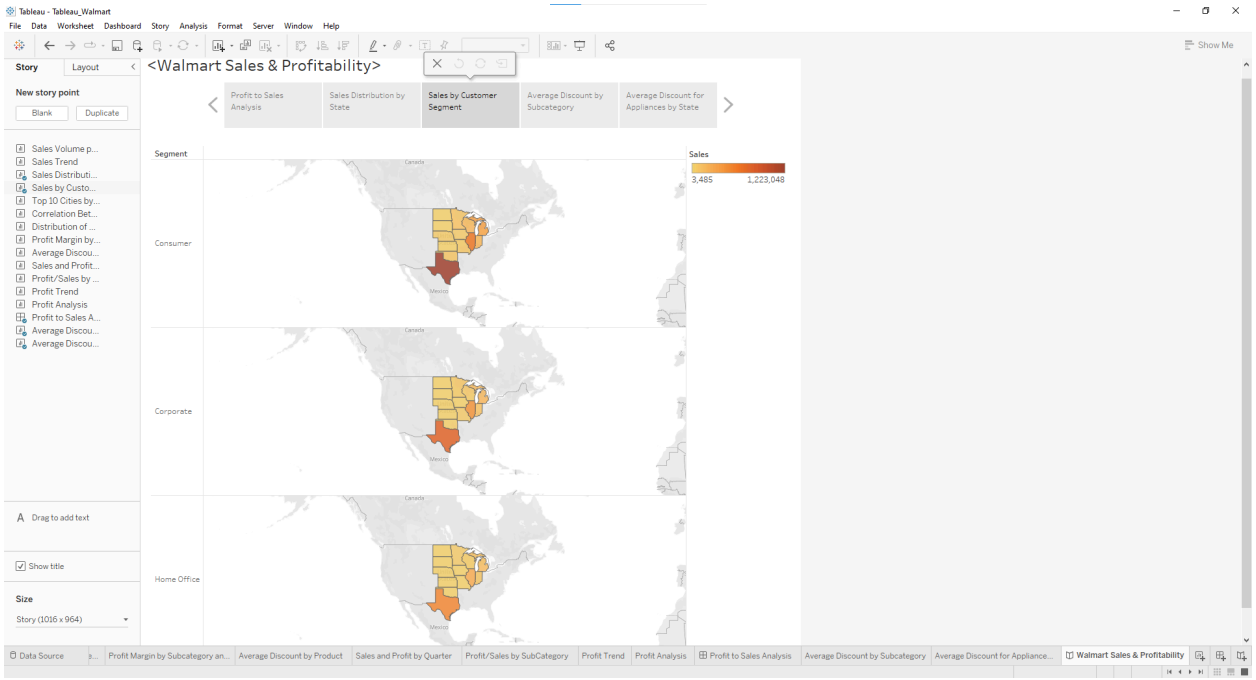
The sequence of sheets on our story section depends on the purpose and objectives of the project. For our project, the purpose is to provide an understanding of Walmart's sales and profitability trends in a meaningful way so they can be conveyed to stakeholders or people of interest who need the venture to keep on profiting by competing in the market. As such, our story consists of our Dashboard sheet and 4 more visualization sheets that provide a better understanding of Walmart's sales and profitability relationship, as well as identify potential areas of interest and attention that could represent opportunity, growth, or improvement.



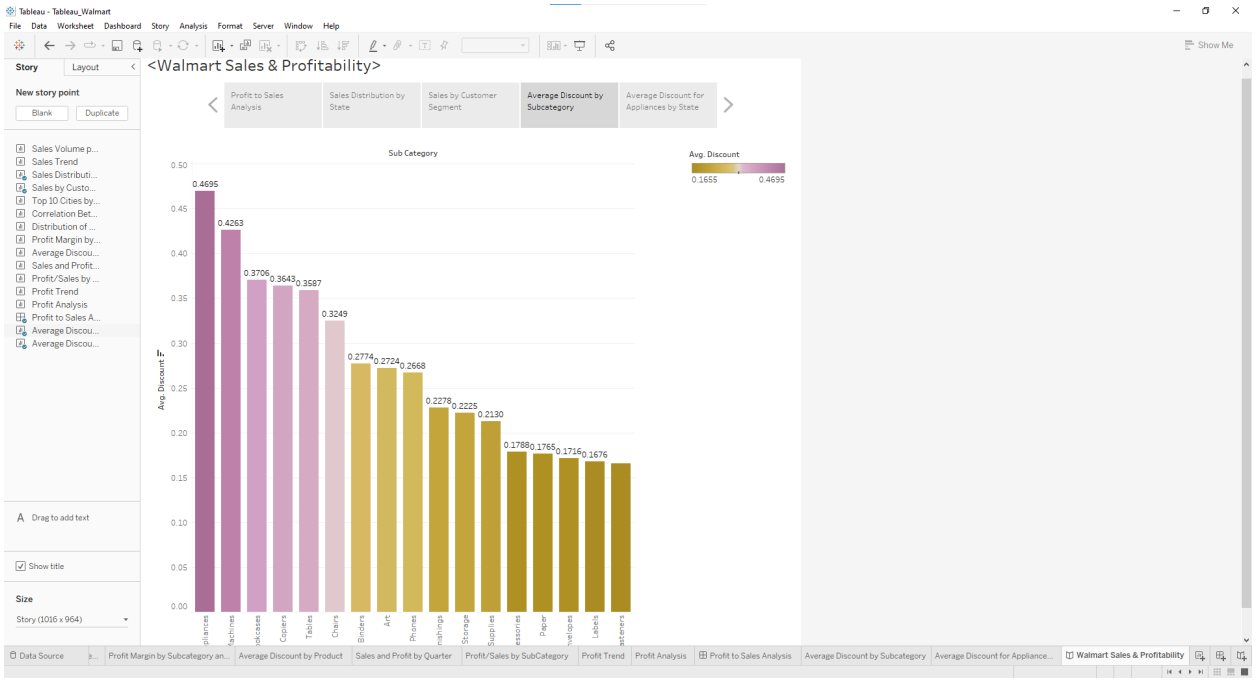
The first visualization we used is the Dashboard which provides a great summary of profit and sales over the year of 2019 as well as a forecast trend going forward. Since the Dashboard has been explained above, we will not go into detail here, but will focus instead on what the rest of our visualizations tell us together, along with the dashboard. Our main takeaway from the dashboard however, is the way profit stands alongside sales over the months of the year 2019.



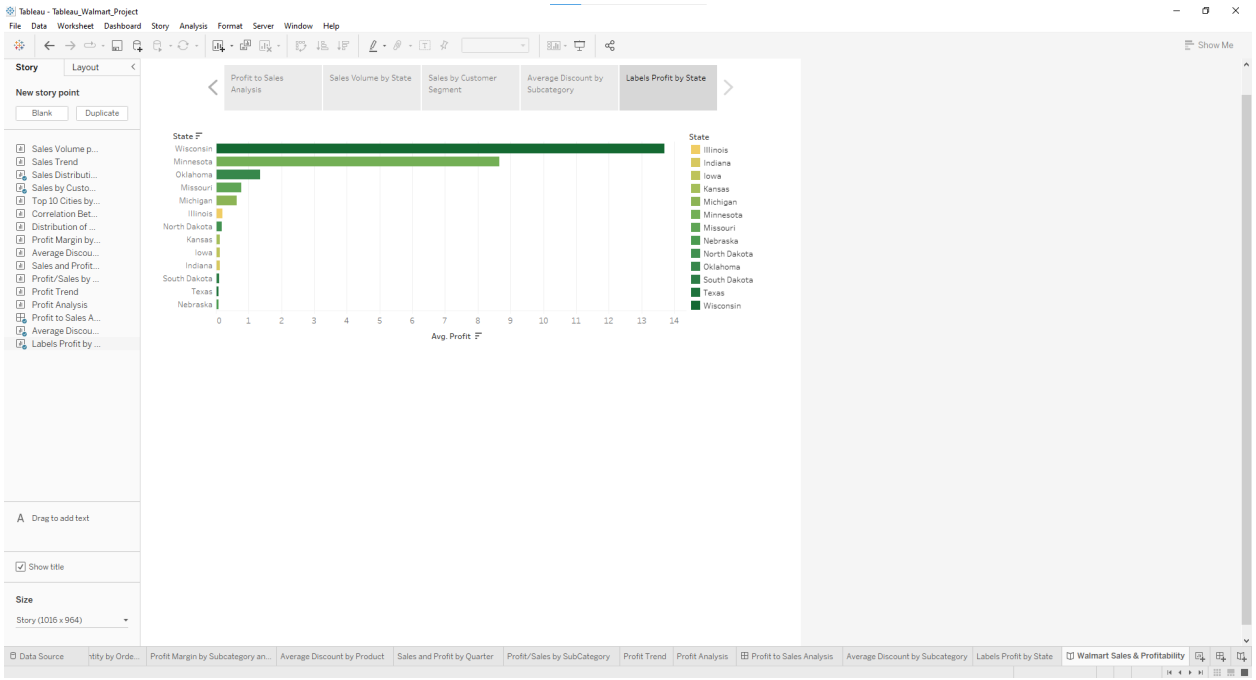
The second story point we used is a Sales Volume Distribution by State which was explained earlier in the paper. This can identify which states experience the highest sales volume, in this case Texas, and Illinois. For any company, maps like this one are always useful in seeing geographically which areas their organization sells at the most.



Another visualization we used for our third story point was Walmart's sales by Customer Segment. Again, this dual axis helps the company see which one out of the three segments in conjunction with the geographical area, in this case state level, has the most amount of sales for Walmart. We see that among all segments, it is always Texas and Illinois who hold most of the sales, with the Consumer segment being the one that outmatches the rest in sales amount.



Moving on to some new visualizations we created for our Story Telling section, first there is the Average Discount by Product Subcategory . To generate that, we used ranking in descending order based on the average discount that product subcategories receive at Walmart. The purpose of this story point is to see which products get discounted the most and as we can see Appliances and Machines are the top discounted product subcategories. This gives us some insight when we combine this information with the Profit/Sales ratio by Subcategory visualization that is part of the Dashboard. Machines seem to negatively impact profit based on that graph, which could help Walmart scrap any plans they currently have about discounting Machines further. Instead, we see that the most profitable subcategory product, which based on our Dashboard is Labels, hardly receives any discount since it appears second from the end in the ranking chart. Being by far the most highly sought after product out of all, Walmart can maybe focus its efforts into promoting bundles or discounts of Labels instead of Machines since the latter presents loss of money while being the most discounted product at the same time.



Four our final story point, we wanted to shed some more light into the profitability of the Label product subcategory by state. To do that we used a ranking bar chart through descending order, while filtering out the subcategory products and just keeping Labels. We used that in conjuction with the average profit that comes by selling Labels across each state. What this filtration can accomplish is to help Walmart's strategy decision team pinpoint exactly which states the most profitable product yields the most returns. In this case, it appears to be Wisconsin, followed by Minnessota. As we mentioned previously, our story telling aims to help Walmart understand the relationship between their profits and sales. Although there is the potential for more in depth analysis, for the purposes of this project we wanted to at least touch upon some aspects that we found interesting. In the Labels case, for example, it is very interesting to see that almost all the profit comes from Wisconsin and Minnesota, followed by Oklahoma far behind. Considering it receives very little discount compared to other products, Walmart could come up with offers and discount deals that could drive up Label sales even more, thus increasing profits. Walmart can also reduce the discounting on machines. Although machines are heavily

discounted, they seem to negatively impact Walmart's profit. Hence, Walmart could consider reducing the discounts on Machines or eliminating them altogether to improve their profit margins. Lastly, per the final story point, Wisconsin and Minnesota are the most profitable states for Walmart to sell Labels. Therefore, Walmart could consider expanding their operations in these states to increase their profits further.

Overall, it is crucial for Walmart to strike a balance between discounts and profits while identifying and promoting the most profitable products in different regions to optimize their sales and profits.

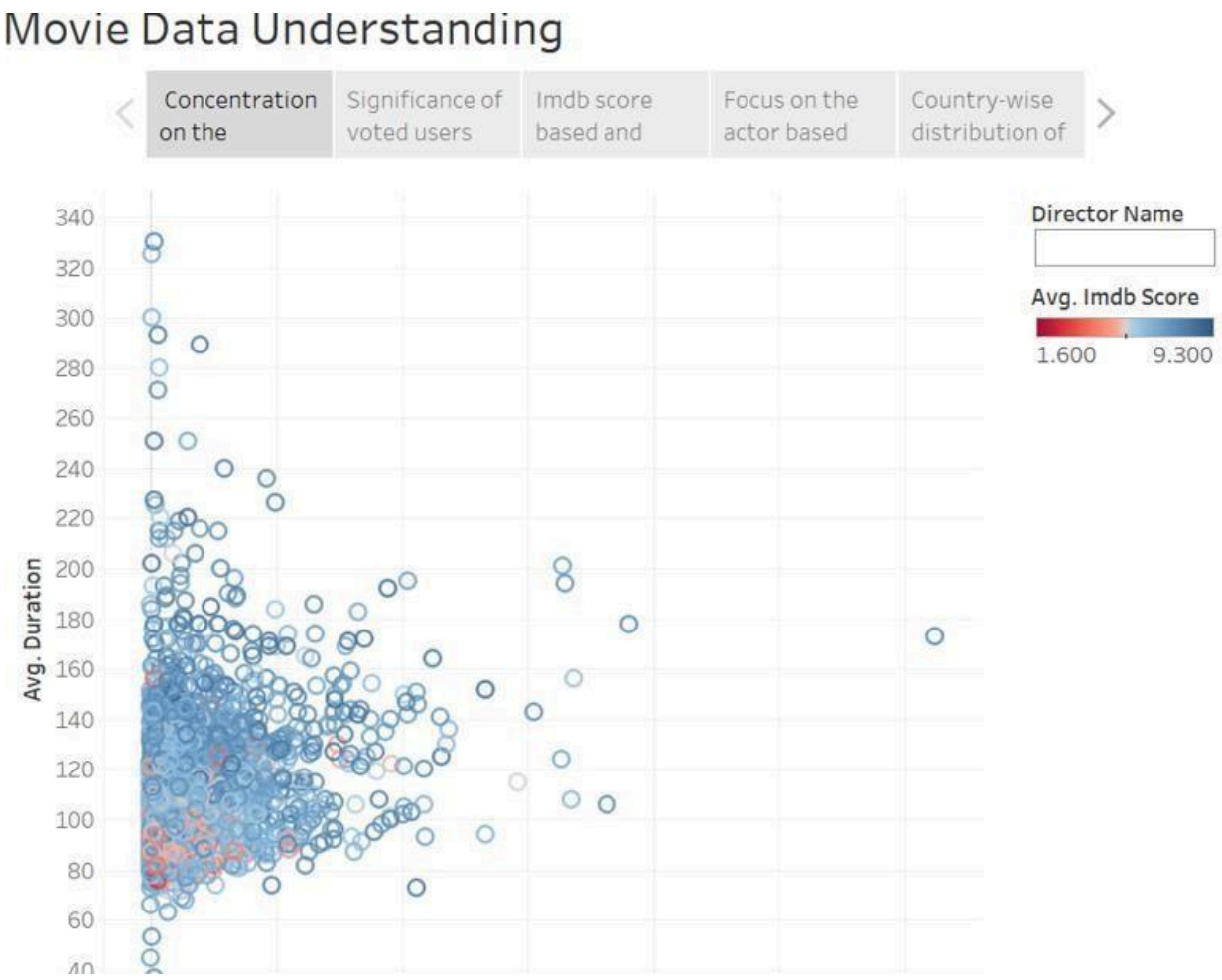
Story Telling:

Movies are the most loved side interest all over the world. There are a wide range of sorts of movies, and individuals lean toward various ones. Regardless of whether it is dramatization, satire, or tension, it appears they are for the most part similarly favored. Movies are viewed as a form of entertainment and to have a good time with your loved ones. The majority of movies are made with the intention of being seen on large screens in theaters and at home. After a few weeks or months on the big screen, movies can be promoted through a number of other media. They are

broadcast on pay television or digital television, and DVD discs or videocassette recordings are sold or leased to viewers so that they can watch them at home.

Majority of the film studios are still successful most years because they have worked out how to monetize the ancillary stream by selling pay-TV and overseas rights, as well as creating tie-in computer games and amusement park rides. Filmmakers have created a large amount of movie- hits and continue to do so focusing on budget and an extensive amount of film making based on the region and films. The plot summary and description of a film and the assessment of the director's and screenwriters' work that makes up most of film reviews can still have an important impact on whether people decide to see a film. Movies are an important part of our lives and create a lot of learning and entertainment for the people around.

Concentration on the duration and gross revenue: An understanding of the total revenue of movies during a duration can be interpreted and is a major factor in understanding the high rated and low rated movies according to their imdb scores and budget. This gives us an overall idea with respect to budget and even tell us the rated movies according to their respective directors. The imdb ratings are from the range of 6.2 – 8.9 with the colour shade specifying the ratings. If we want to find out the rating for a movie that was made by a specific director a director is searched for and the number of movies and their ratings are displayed. The imdb score gives an idea about the various factors and the duration required for the high rated and low rated movies. If we want to find out the rating for a movie that was made by a specific director a director is searched for and the number of movies and their ratings are displayed. The gross revenue and the imdb score for those movies can be displayed giving a significant amount of information which will be useful.

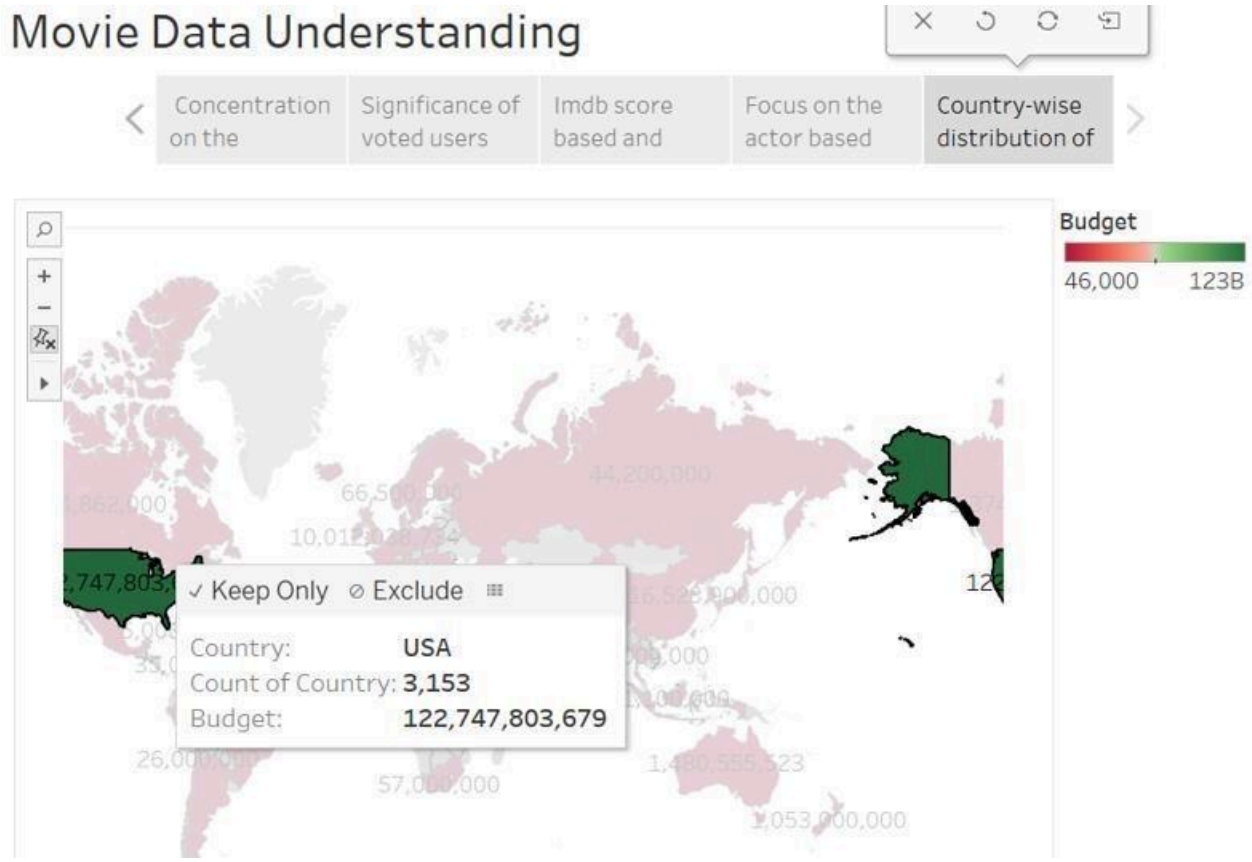


Significance of voted users: This is number of voted users in the specific year for various movies helps identify, the number of voters, the year at which these movies were released and the increase in the voted users according to the year. The number of voted users are shown that helps us understand the number of people watching those movies. It gives us an idea of the impact that is created on the people with these movies. The various time frames and the number of people being open to different kind of movies is significant in understanding the various kinds of genres they would be interested in, the movies to focus on and the impact of the user base that is interested in these movies.

Imdb score based and revenue based criteria: The imdb score according to the various genre is specified and differs according to the genre. The imdb score could be low for a particular genre and but can have a high gross revenue This helps us understand the various genres and their significance to various kinds of movies with their imdb and their gross revenues. The genre shows which kind of movies have a greater gross revenue and the impact they are creating on the people with the imdb score rating. A numerous other factors can be classified but a genre based evaluation important way to identify a user-base based on the genre.

Focus on the actor based revenue: The ranks show various actors from movies that have generated the highest gross revenue. The data here signifies various data with respect to gross revenue considering actors that have worked. Ranks is a way where this can be specifically understood from the bar chart above. A data filter with the name of the director can be mentioned to identify the actors that have worked with these various directors in the high revenue movies.

Country-wise distribution of movie budget: A geographic map helps understand the country- wise distribution of various movie budgets all across the world. This is an important way to analyze not just the budget but the high revenue movies and gives a large amount of user data which can be useful to make different kind of movies all across the world. The filter shows the colour which specifies the budget that is lowest in the countries to the budget that is the highest in the countries. Various movies can be depicted and the different genres that have created an impact across the globe is also identified. The largest budget is created by the US and is one of the leading countries in film making and film profits as well.



References:

1) **Imdb movies :** IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows

2) **Movie gross budget:** The Numbers - Movie Budgets, Most Expensive Movies, Most Profitable

Movies, Biggest Money-Losing Movies (the-numbers.com)