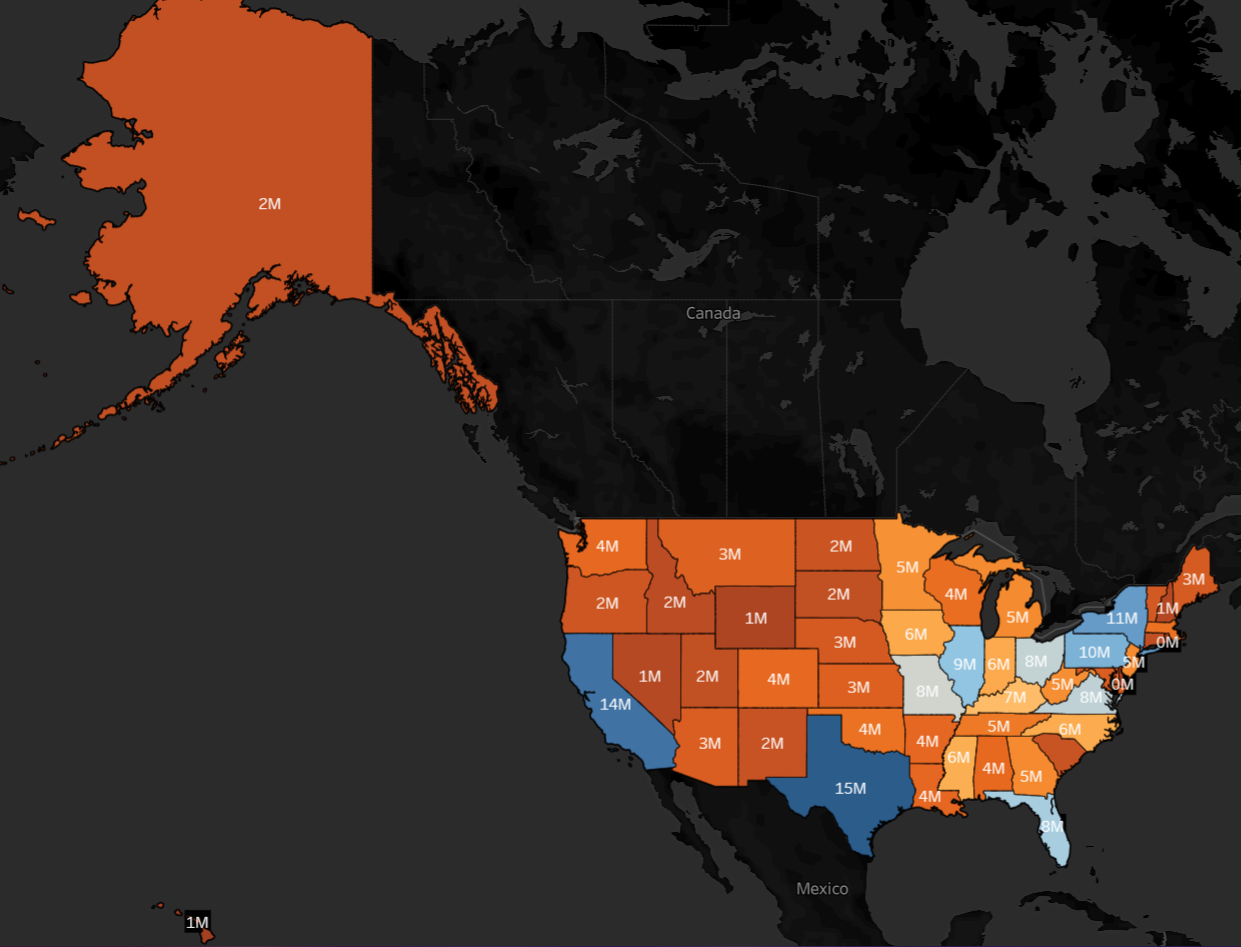
*Sales Analyst Case Study*

In this exercise, I utilize a large dataset containing descriptive characteristics of sales-related geospatial, time-series, and demographics data. The dataset includes ~287,000 observations pertaining to Order ID, Order Date, Order Status, Item ID, SKU, Quantity Ordered, Price, Value of Product, Discount Amount, Total Price, Category, Payment Method, Business Intelligence State, Customer ID, Year, Month, Reference Number, Name Prefix, First Name, Middle Initial, Last Name, Gender, Age, Full Name, E-Mail Address, Customer Since, Social Security Number, Phone Number, Place Name, County, City, State, Zip Code, Region, User Name, and Discount Percentage. The dataset can be found here: <https://github.com/stanley-george-joseph/Customer-Analysis-Tableau/blob/main/sales_06_FY2020-21.csv.zip>.

Nevertheless, the overall goal of this exercise is produce visualizations utilizing Tableau to provide business insights with respect to customer analysis. The business in this case study only operates throughout the territory of the United States.

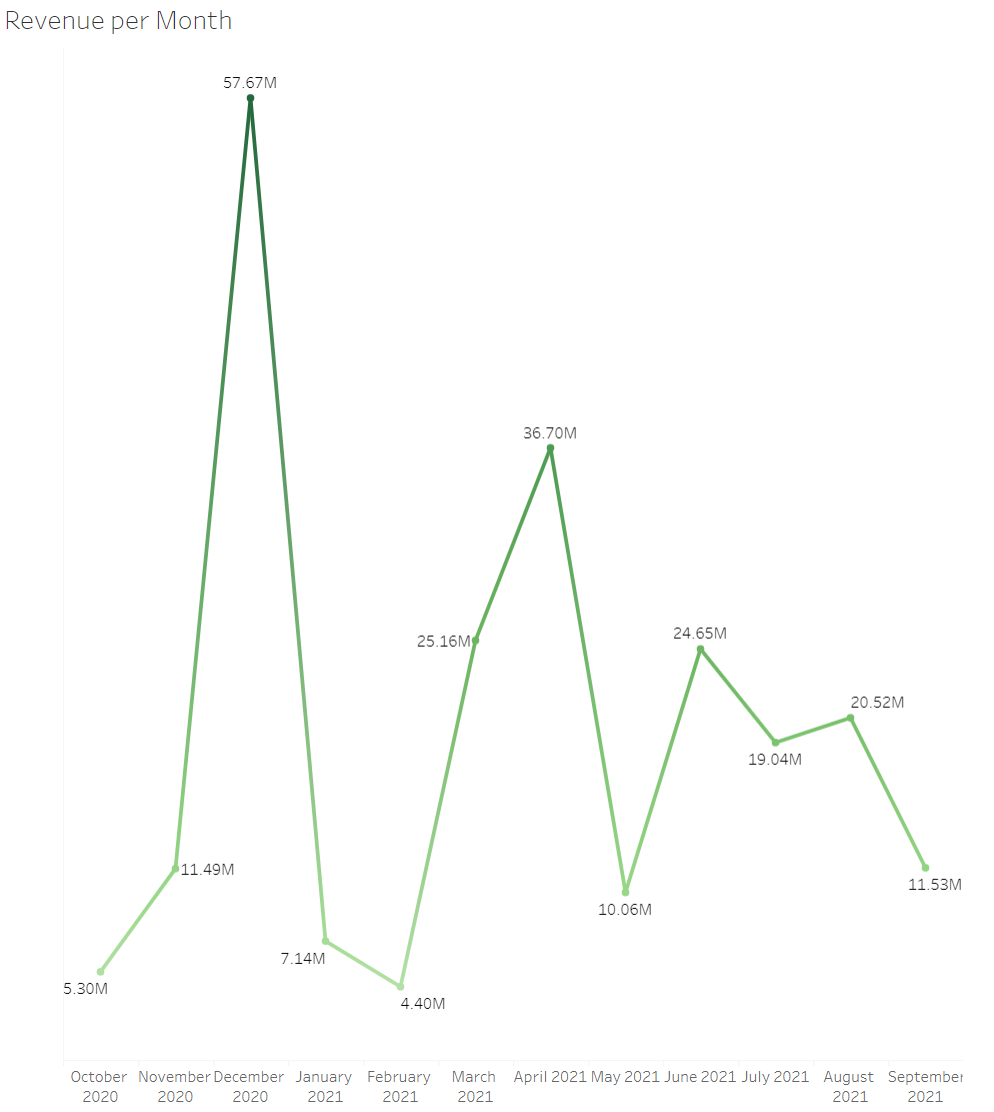
Given this information, the first thing that I would like to explore is revenue per state. To do so, I decide to produce a heatmap of the United States. This visualization is useful as it provides insight with respect to overall sales in specific geographic regions within the business’s operations.



In producing this visualization, you can clearly see that Texas is the state in which this business generates the most sales totals. Likewise, we can see that our sales are also focused in California as well as concentration of states in the Northeast. Similarly, we can see that Florida is a key state for our business as well. Through this visualization, we can also identify potential geographic territories where it would be beneficial to shift our business development focus. As an illustration, a relatively small portion of our sales totals occur in the Midwest and other western territories.

Something to note in this visualization as well is that I produced the color-scale to orange-blue diverging to take into account diversity and inclusion factors. People, particularly males, are relatively often colorblind. With that in mind, I decided to use a color-scale that could be seen and identified by everyone. I also changed the background to black to make the map pop a little more visually as well as maxed out the opacity of the color-scale for similar reasons. Additionally, I added totals to each individual state, so they can be directly identified in an exact manner rather than just a general manner using the color-scale.

For my next task, I would like to display business revenue based on month of the year. Naturally, for this type of task, I believe it would be most appropriate to utilize a time-series line graph.



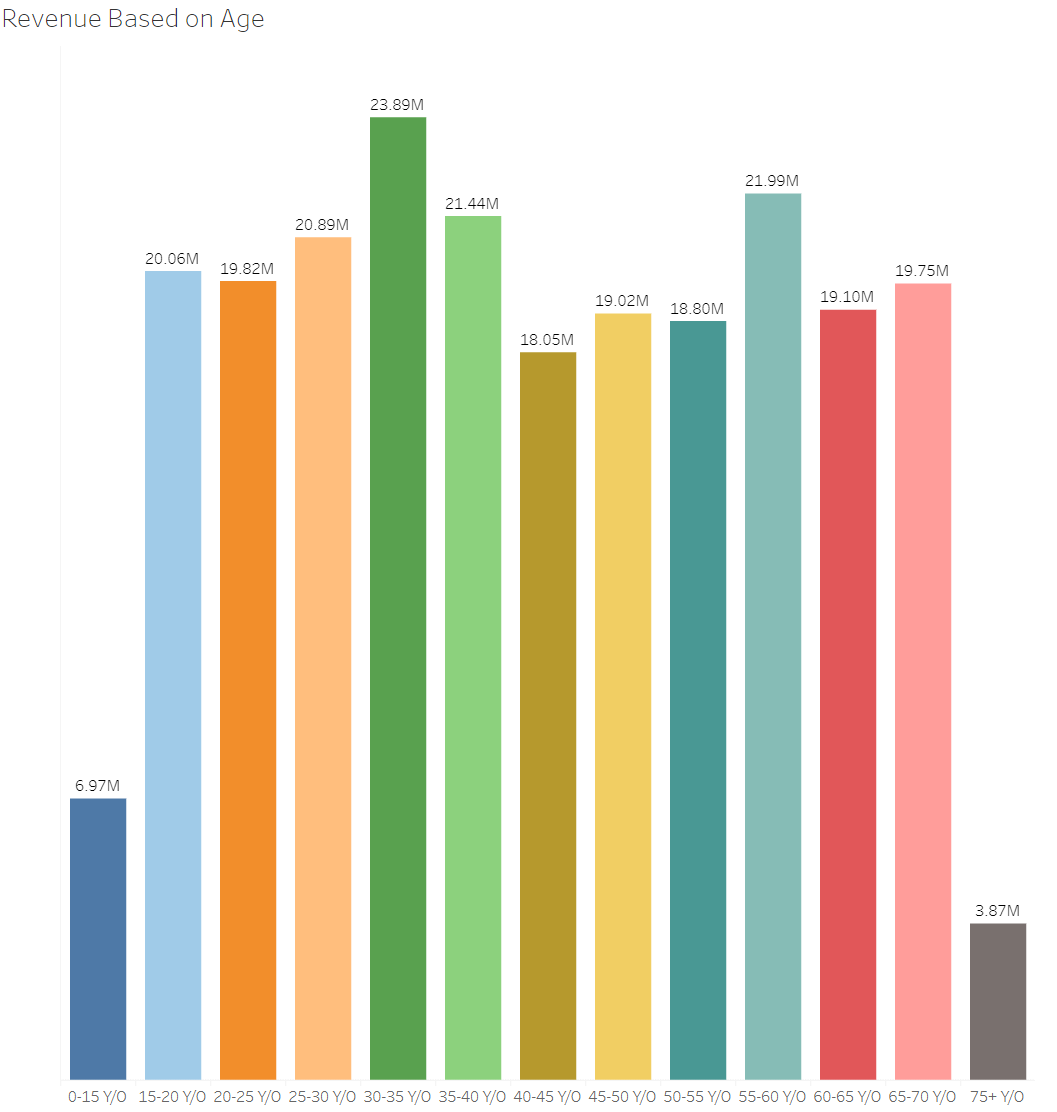
Using this visualization, we can clearly see that the business achieved the most revenue in December of 2020. Curiously, this is congruent with most market research indicators of consumer products. Intuitively, it has been well-established that a lot of businesses in the consumer products sector experience the most sales in the fourth quarter, particularly in December. This is driven by market demand via Holidays and an increase in discretionary income afforded to consumers via annual bonuses.

We can also see from the illustration that business revenues are not constant. There is a large degree of variation from month to month; however, there is a trend of increasing revenues over time.

There is something to note about this line graph visualization as well. I initially struggled to produce this visualization because Tableau, by default, identified the date columns of the dataset as a string variable. Nevertheless, I changed the variable to the dates datatype to get around this issue. Likewise, I was not initially provided month data either. To get around this problem, I utilized year data, which I did have, and broke down the years into month bins. I also changed the newly created month data into a discrete rather than continuous variable. Additionally, I removed distracting line grids, converted axis titles, and changed the color of the line graph to green considering all of these totals represented positive revenues as opposed to negative revenues, which could be interpreted as losses and are historically represented by the red color.

Next, I would likely to identify the relationship between age and revenues. To do so, I would like to construct bins of age ranges with respect to consumers. There isn’t necessarily an industry standard when it comes to how large these bins should be or how the exact age ranges should be defined, so this analysis is purely discretionary from my perspective as analyst.

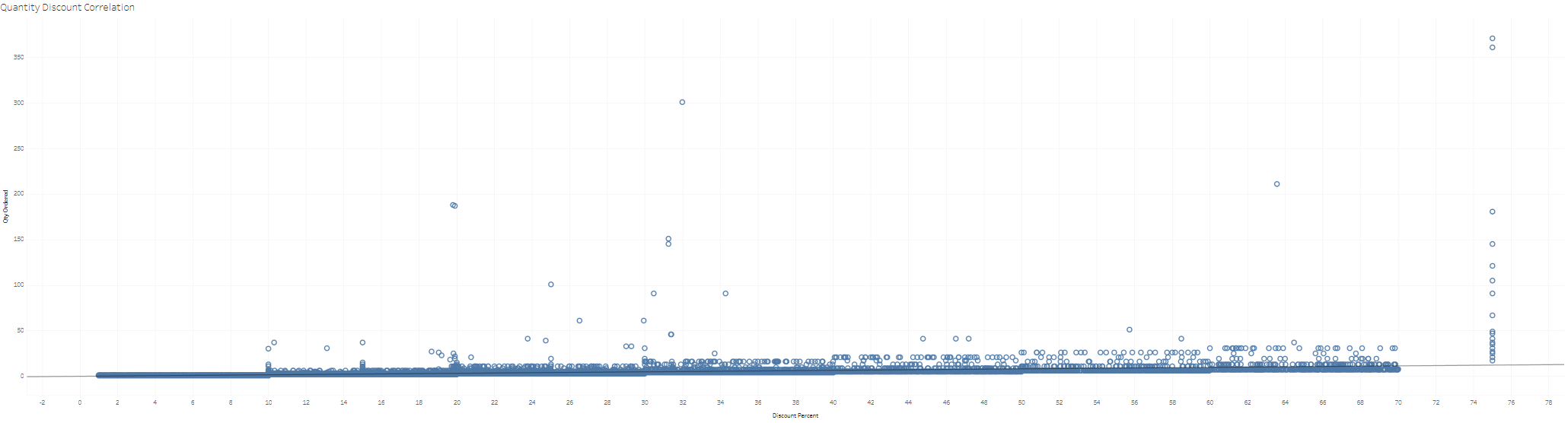
I first produce a bar chart that displays the relationship between exact ages and total revenues. However, upon producing this visualization, I immediately notice a major problem. Tableau, by default, views the age datatype as a measure, so it applies the sum function to the data. This is not what I want. So, I decide to follow through with my initial plan by creating bins with respect to age ranges. I also added to a color palette to make the bins more easily identifiable in the visualization as well as edited aliases to make the x-axis of the chart more understandable to a general audience.



Like to previous example, I also removed distracting grid lines, which make the visualization appear more messy in my opinion. Regardless, from the above visualization, we can clearly see that 30-35 year old individuals are the business’s largest segment of sales totals. Conversely, we can see 75+ year old individuals are the business’s smallest segment. And, as an overall picture, the business should endeavor to market their product to overall demographic of 15-70 year old individuals as a whole.

For my next task, I would like to explore the correlation between quantity and discount percentage. It has been well-established through market research that consumers are more likely to buy in increased quantities the higher a discount percentage may be. Given this fact and relying on my intuition, I would like to see if this relationship holds true for this particular business.

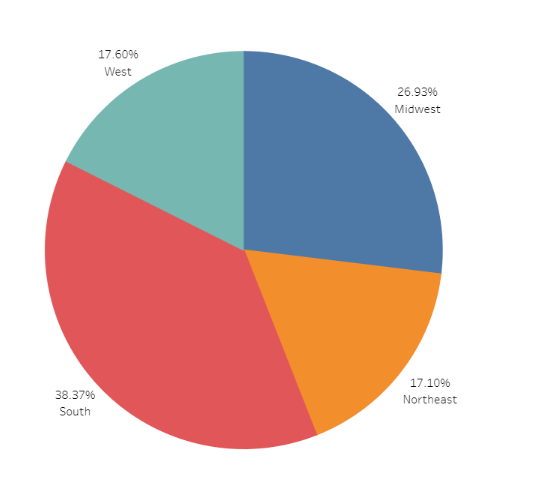
To produce the visualization, I identify that it would be most appropriate to produce a scatter plot. Likewise, I did encounter some issues upon my initial attempt to produce this visualization; however, these issues were easily solved by changing some variables into the dimension type as well adding a filter to the dataset via the discount percentage variable. Ultimately, below is the visualization I produced as an end product.



Naturally, given a dataset this large, there are some outliers. It is typically up to the analyst’s discretion on how to “deal with” outliers. In this instance, I decided to leave them in this analysis considering these were real transaction that took place.

Regardless, using the trend line, we can see that there is a positive correlation between discount percentage and quantity ordered, as initially suspected via intuition and general knowledge of how markets function. Essentially, we can conclude that if discount percentage is high, the quantity ordered is likely to be high as well.

For my next task, I would like to further explore geospatial-related data with respect to total sales. In particular, I would like to get a breakdown of percentage of revenues per region. Given this fact, I believe the most appropriate visualization would be a pie-chart or donut-chart. Ultimately, I decide that the former would be most appropriate.



This visualization is particularly important. Earlier via the heatmap I produced, I contended that there is concentration of sales in northeastern territories. However, this is not necessarily the case. As we can see from the above pie chart, northeastern represent a relatively small portion of totals sales, under 20%. Likewise, I earlier contended that the Midwest had little overall sales. However, this is not the case. The pie-chart, through interpretation, is telling us that, in fact, the Midwest represents over a quarter of total sales. Similarly, I contended that the west had little sales too. However, once again, the pie-chart is painting a conflicting narrative.

How can I reconcile and interpret these conflicting conclusions? The pie-chart and the heatmap are telling me two different stories. Upon further reflection, both visualizations are insightful and correct. The northeast is an aggregate of a small collection of states in total. Therefore, even though the northeast is under 20% of revenue, it is still an important territory that is under-represented in the sample. Moreover, the Midwest is a large collection of states, so, naturally, they should have a relatively large slice of total revenues as well in aggregate. The Midwest region is over-represented in the sample.

Ultimately, it would be beneficial to produce a larger number of regions to capture that over-representation/under-representation relationship I alluded to previously. Nevertheless, the pie-chart produced is insightful as displays the regional distribution of sales as opposed to state distribution of sales. Using this graph, I would recommend the business to either a) further capture market share in the south through targeted marketing of products or b) try to capture more market share in western regions of the United States.

I would now like to explore revenue per category per gender. To produce this visualization, I brainstormed up that a bar graph would be most appropriate. However, upon producing that visualization, I immediately noticed that the illustration was busy and difficult to interpret.

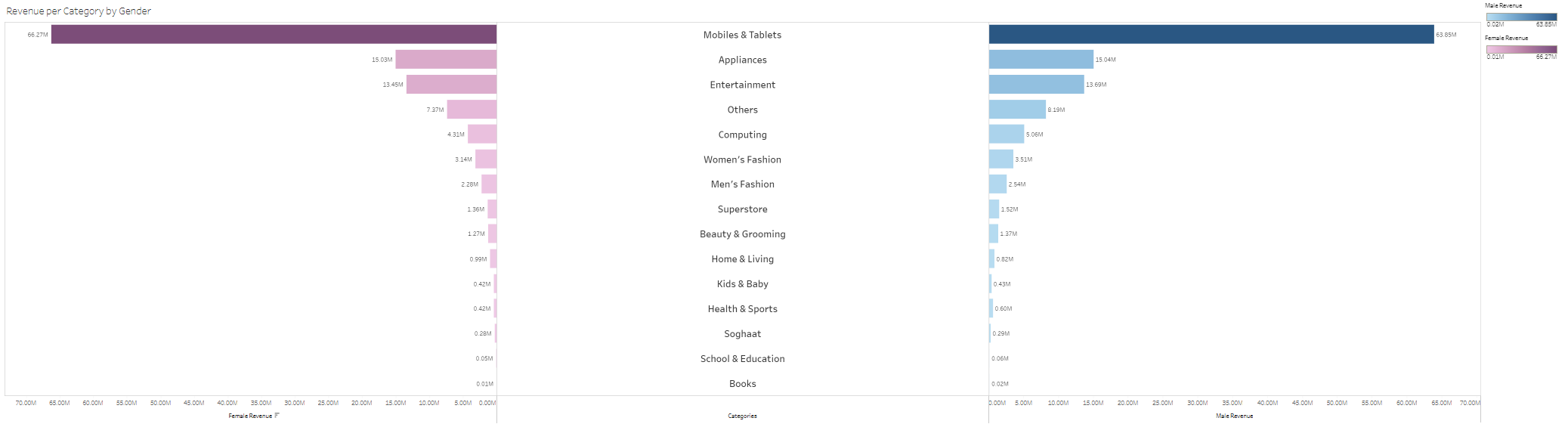
Upon this realization, I then envisioned that a butterfly graph would be most appropriate to produce. However, Tableau, by default, does not have an option for this particular type of visualization. I will need to manipulate my data appropriately to get the end of product I want.

To reconcile the issues I was having, I needed to create a new calculated field. In particular, I needed to separate total revenue by gender using the following statement: “IF [Gender] = ‘F/M’ THEN [Total] END.

Then, I create the bar chart I initially envisioned. However, it is still busy and hard to interpret, and it is not the end product of butterfly chart I wanted to produce. For instance, the charts are both reading left to right, and the charts are not sorted correctly either, varying in length from category to category.

To address this issue, I reverse the axis on the female revenue side of the chart. Likewise, I sort both charts into descending order. Even still, this visualization is still not what I want ultimately. I want the categories in the middle of the chart, not to the left for female revenues and to the right for male revenues.

To address this issue, I created a zero line with a new calculated field. Using this zero line, I defined categories as labels. And, I centered this new zero line in-between female revenues and male revenues. I then made distinct colors with respect to female and male revenues. Historically, females are represented by pink/purple, and males are represented by blue. So, this is the color-scale I ultimately went with. Below, I have attached the end product of the butterfly chart.



From the chart, we can see revenues per category per gender information in a visual format. We can also identify that movies & tablets are the largest revenue category for both females and males.

Overall, the purpose of this case study was to demonstrate my comfortability working with large datasets, my proficiency working with Tableau, and my ability to produce visually-appealing relevant illustrations that provide tangible business insights. To see the visualizations produced in a better format, please go to the following link: <https://public.tableau.com/views/SalesAnalystCaseStudy/OverallDashboard?:language=en-US&:display_count=n&:origin=viz_share_link>.