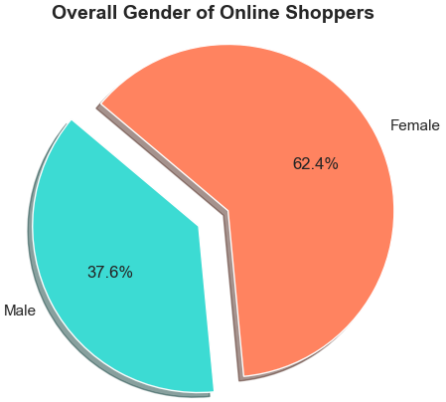
Final Write Up

In today's world, everything is available at the tip of your fingers. Online shopping has created a space for smaller companies to make their product available to the world. This new method of shopping has shaped a new world for physical retail stores. In order for companies to stay profitable and relevant, they need to understand what makes people want to buy certain products. The retail business is of interest to our group because of the joy online shopping brings to our lives. The amount people spend is also an indicator of the economy.

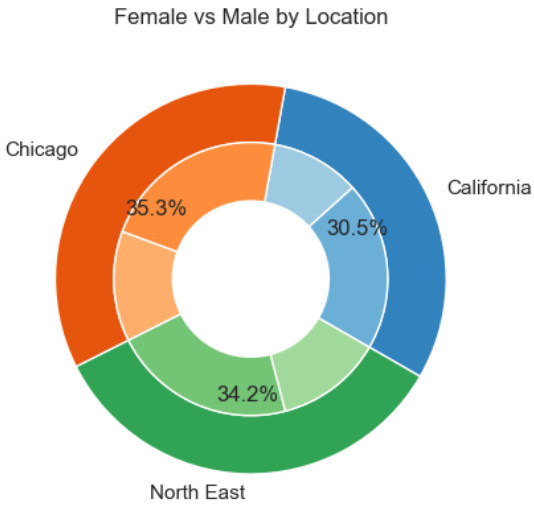
The objective of this project is to look at a few questions that overall help answer how people are spending their money online. This topic was chosen because of the relevance it holds in our lives and it interests us to see how healthy our economy is. To accomplish this, we took an online shopping dataset from Kaggle datasets named “Online Shopping Dataset.” To make this dataset usable, we dropped some columns that were not relevant to our analysis. Rows with null values were also dropped to ensure a clean dataset. To make the analysis easier, we combined like categories of “Nest” since they are the same company but different locations. Finally, we combined New Jersey, New York and DC and renamed it as “Northeast.” This made the three locations within the dataset similar in size to each other. This completed our data cleaning.

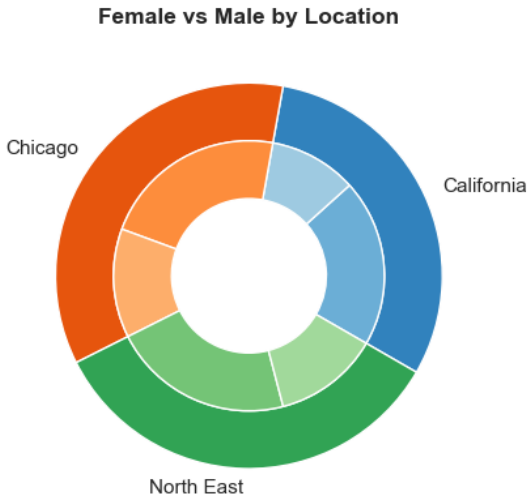
Ambers written portion

Online consumerism is an extremely important factor in today's economy. The accessibility and ease of a customizable shopping experience that is catered towards oneself, is a driving factor for most. Because of this, online shopping attracts a wide range of individuals with many different backgrounds. For merchandisers to achieve success, it is important that they possess an understanding of the overall market. That raises the questions, who makes up the majority of the market? Who is spending more, and what are they purchasing? To answer these questions, I created a pie chart that visualizes the gender of the total amount of Online shoppers.



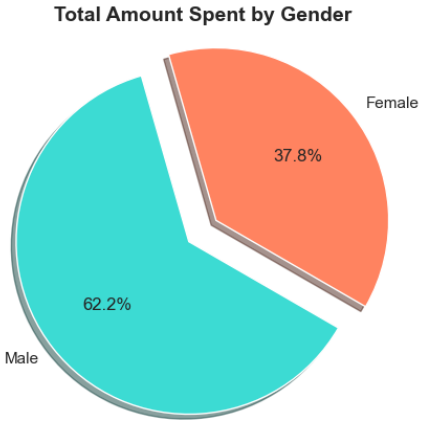
In this pie chart it is clear to see that across online shoppers gathered in this dataset of the United States, women make up more than half of the consumers. To understand this concept with greater detail, I then created a nested donut chart based on region in comparison to gender. The outer layer represents the comparison of both female and male online shoppers, indicating the percentage that each region contributes to the total online shopping population. The inner layer divides that percentage by “male” and “female” in order to compare the difference within all regions.





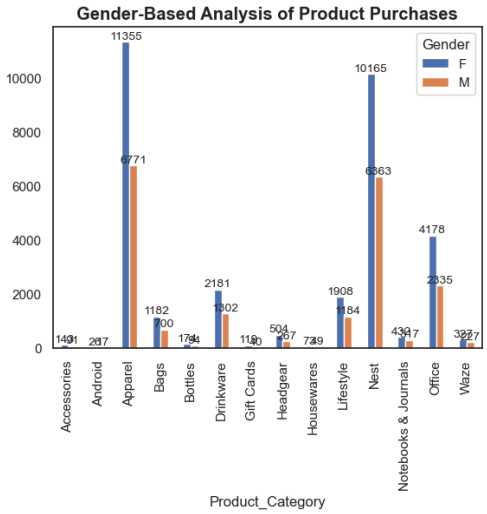
I first created the chart with the percentage visible in order to be able to distinguish each section by region. I then added the labels to name each section and drop the percentages in order to better visualize the size differences between sections. By analyzing the chart it is clear to see that the differences between male and female are nearly identical across all regions. This indicates that assuming women make up around at least sixty percent of the population of online shoppers across the country is accurate and unlikely to vary based on region.

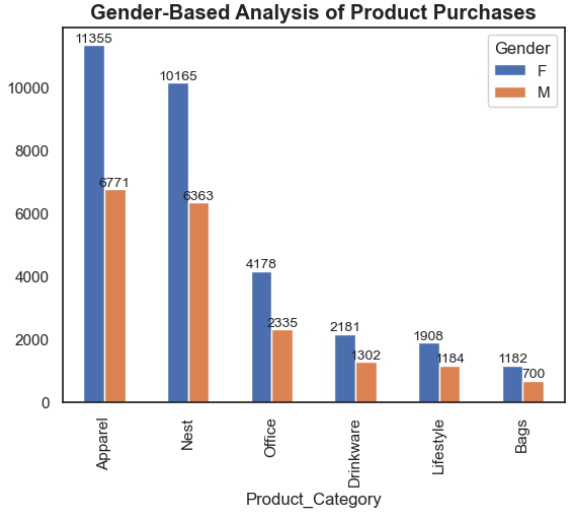
The next question that is raised is what gender is likely to spend more on a transaction? To answer this question I created another pie chart to visualize and compare with ease and clarity. Because there are only two factors, I felt a pie chart would be best to represent the data clearly.



In the chart it is evident that although Females make up the majority of the online shopper population in the United States, males tend to spend more on transactions. This would mean there is a higher chance of a man spending more than a woman.

Finally the last question, what is being purchased most frequently? This question is important because understanding the frequencies of product purchases and who is buying them will give an important insight that will allow merchandisers a better understanding of who to sell to in order to be most successful. This question is important because understanding the frequencies of product purchases and who is buying them will provide crucial insights, enabling merchandisers to better identify their target market. To achieve this insight I created a bar chart that compares between gender, product category, and quantity of product purchases. The y axis is the amount of products purchased by gender, the x axis is the category of products purchased by gender.

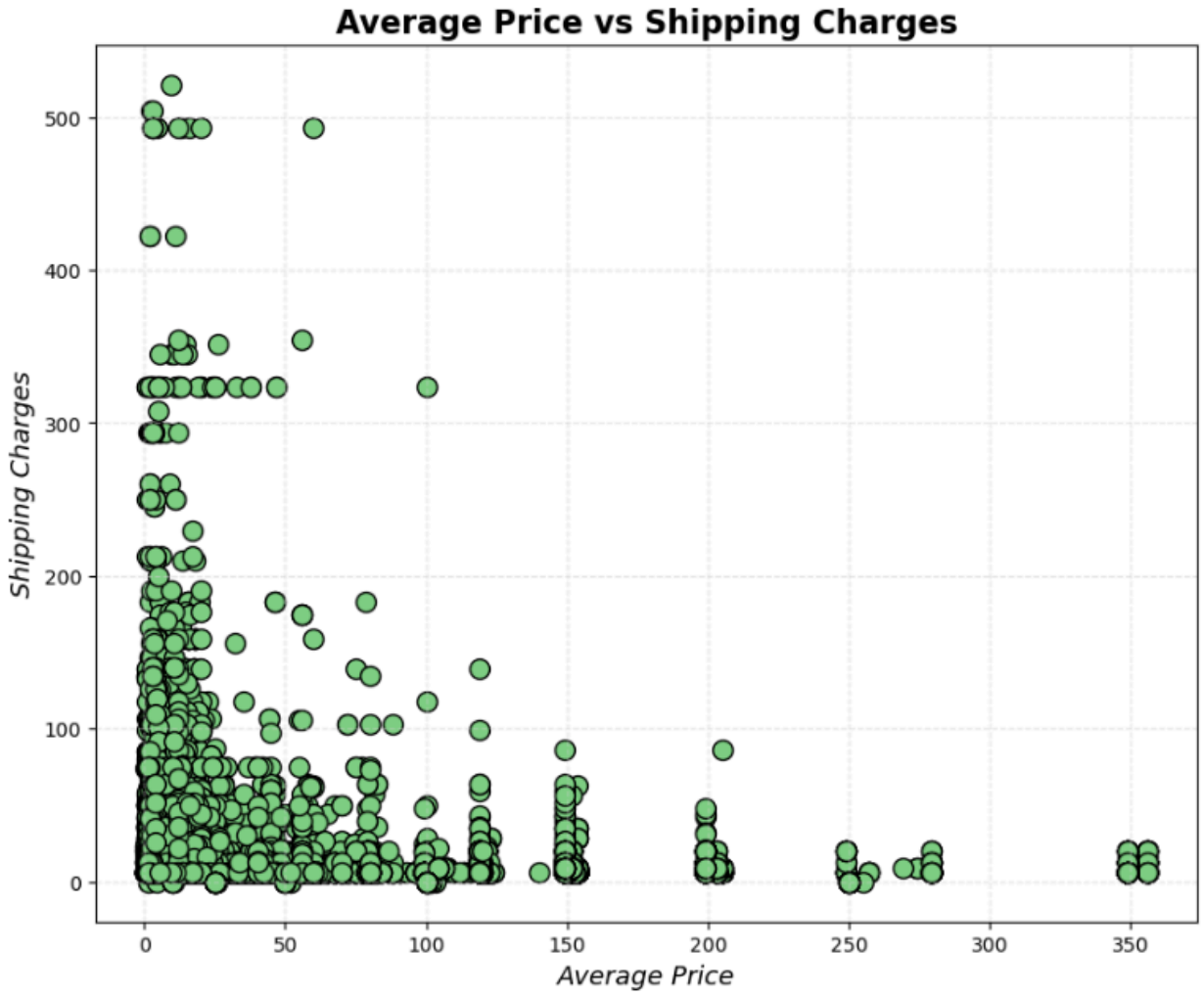




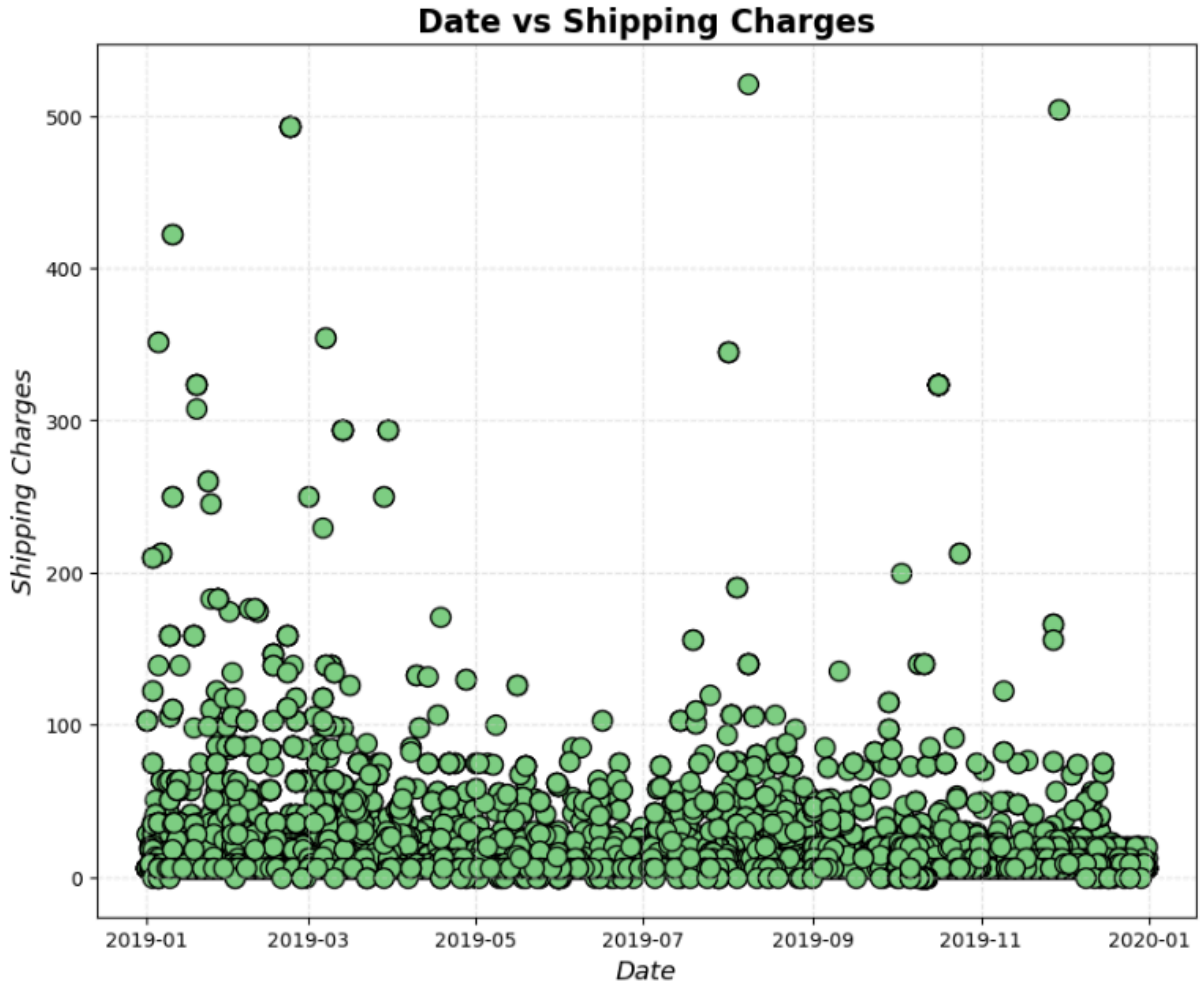
Initially I created the chart with all the categories listed and visible as well as not in a quantitative descending order. I decided to change this because it is not as easily understandable and comparable as my final version, which only shows the top six and is organized in descending order. By analyzing the chart one can compare the differences between product purchases and gender. With this information, it can be inferred which product category will be most desired based on gender. In conclusion, there is a trending difference of the online shopping population in the United States based mostly on gender.

Sams portion

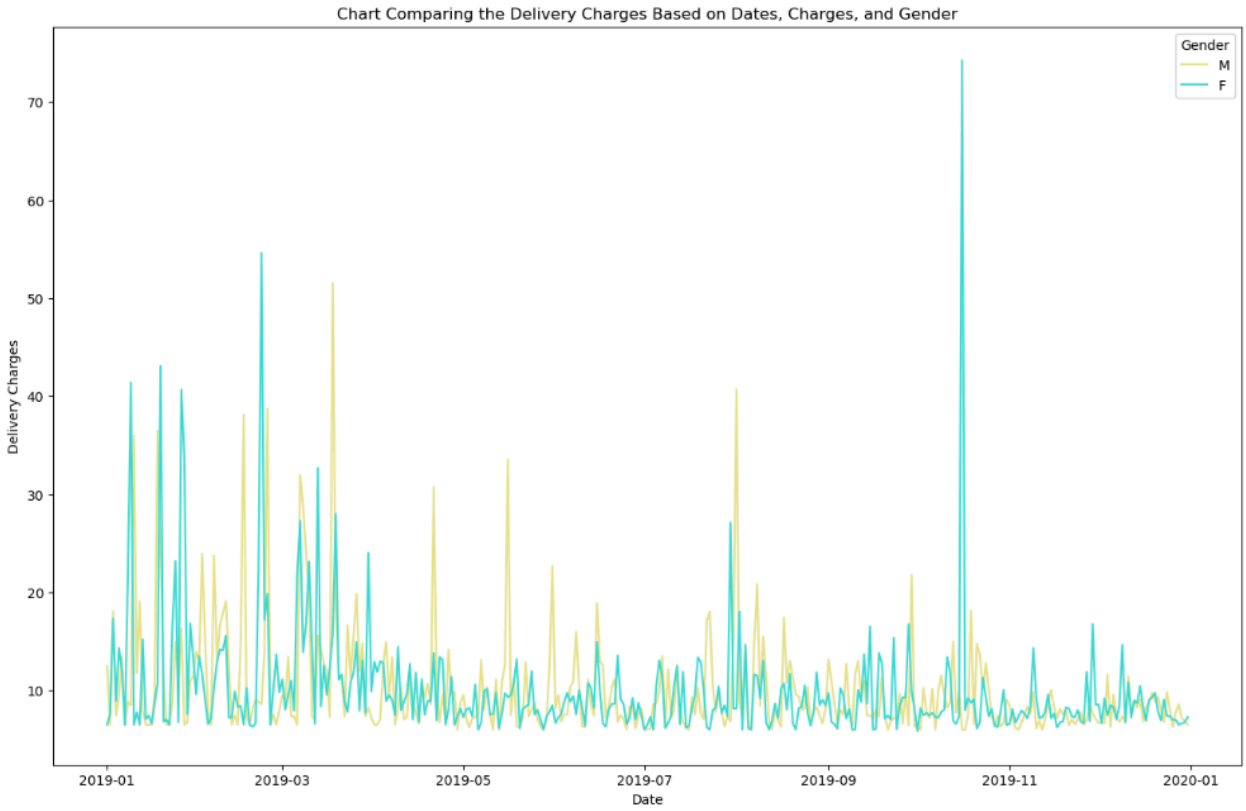
Is there a relationship between the cost of a product and the delivery charges people are willing to spend? Are people willing to spend more money on delivery on a larger purchase since they are already spending a large amount? Or are people willing to spend on cheaper items in relation to delivery? According to the following scatter plot, we find no relationship between the item cost and the shipping cost. It brings into question some of the accuracy of the delivery charges. Are delivery charges being used as the source of revenue for some of these companies? This plot raises more questions than it answers.



Since no questions were not answered in the previous plot, a new scatter plot was made to answer if the time of year affected the delivery charges. Again, the outliers are seen throughout the year except a dip in the delivery charges during the warmer months of the year. Are these outliers freight charges? What are people buying during the cooler months that would prompt a willingness to spend more on delivery? Again, more analysis is required to make sense of the data.



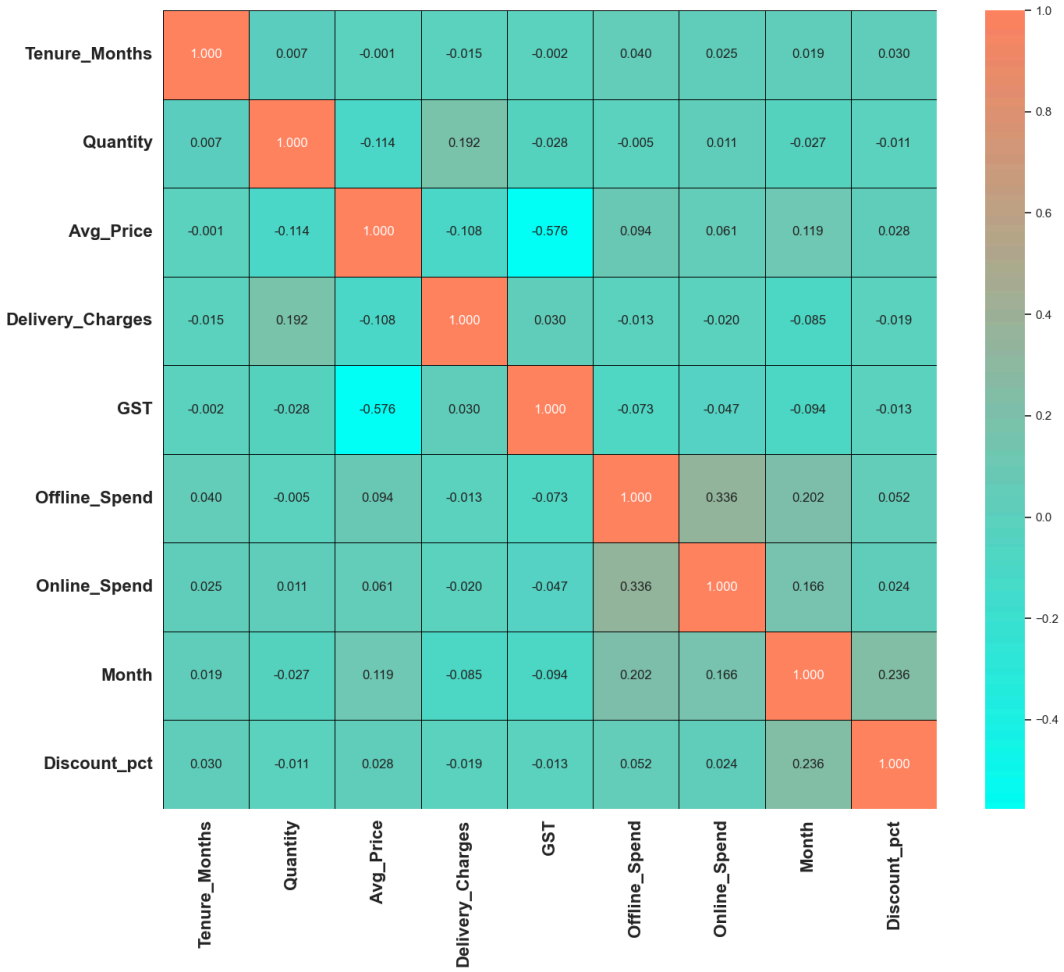
The final graph in relation to delivery charges is in relation to the time of year and if there was a difference between gender. Overall, men and women tend to spend the same amount on delivery charges throughout the year. There is a time of year when men will spend more on delivery. This is generally in the summer months. What is this in relation to? Are these outdoor items that require special delivery methods? Or are these last minute anniversary gifts that have to be expedited? Women will spend significantly more just before the “official” shopping season begins. Overall, there don’t seem to be a relationship between any delivery charges and the time of year, gender and product cost.



Alex's Regression Analysis:

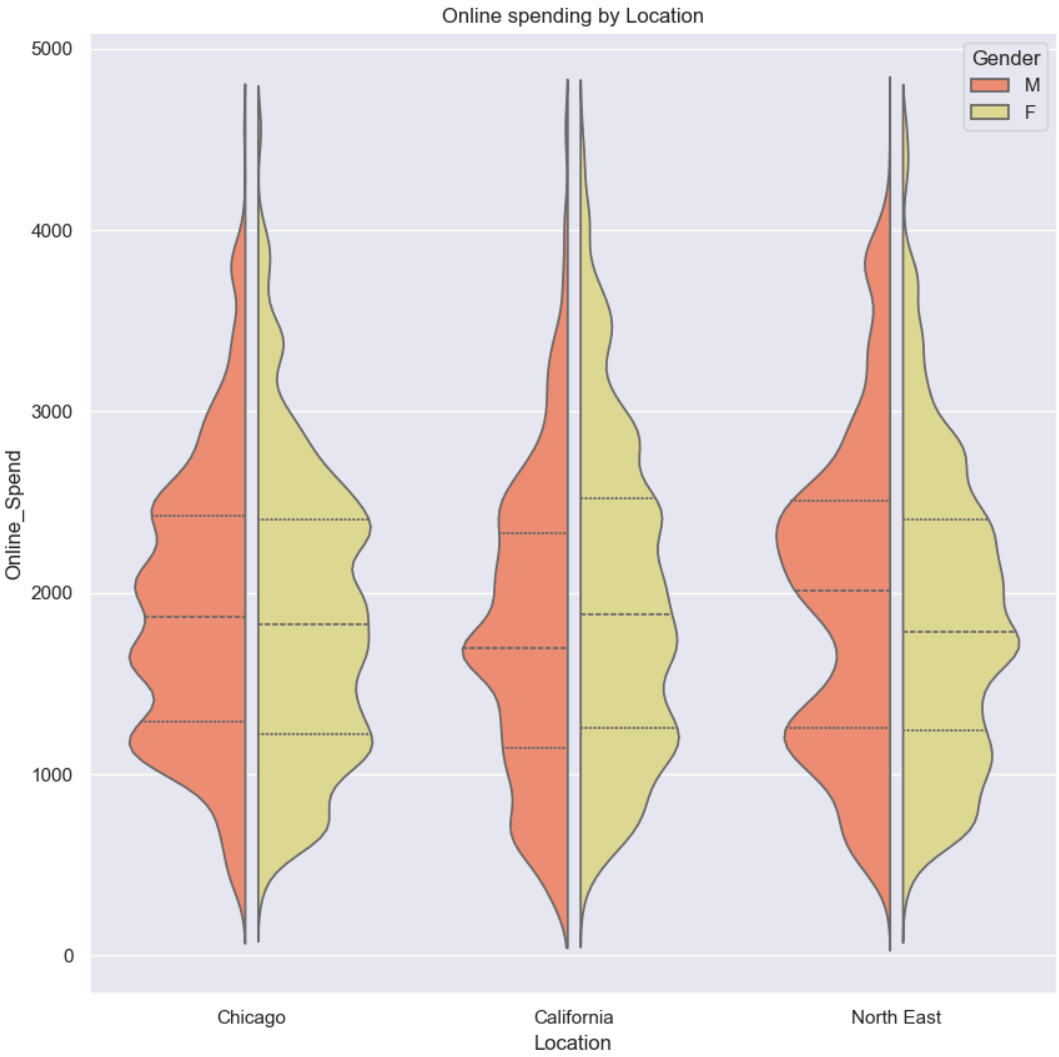
When I began trying to figure out what types of visuals I wanted to use with our regression analysis I first needed to get a feel for our data. I needed to see what kinds of questions we could ask, outside of our main research questions. I did several different exploratory plots to see

Before taking a deep dive into the relationships within our data I made this heat map to get a feel of what relationships I might see between our numeric categories.



At first glance there are no significant relationships between any of the categories. The GST or goods and services tax associated with the product and the average price of the product are the least correlated values in the heatmap, this could be because different products might have different countries of origin and different import tax rates.There are no values in the data that appear to be significantly correlated, but, the most correlated values are the offline and online spending, with about a 34% correlation. While this is the strongest correlation we have, it may not be a true correlation. This data is based on online shopping data, after a more in depth look we are unsure how the offline data was gathered or if there was a mistake when the data was created.

Next I created a violin plot to show the relationships between how men and women spend money online based on their location.

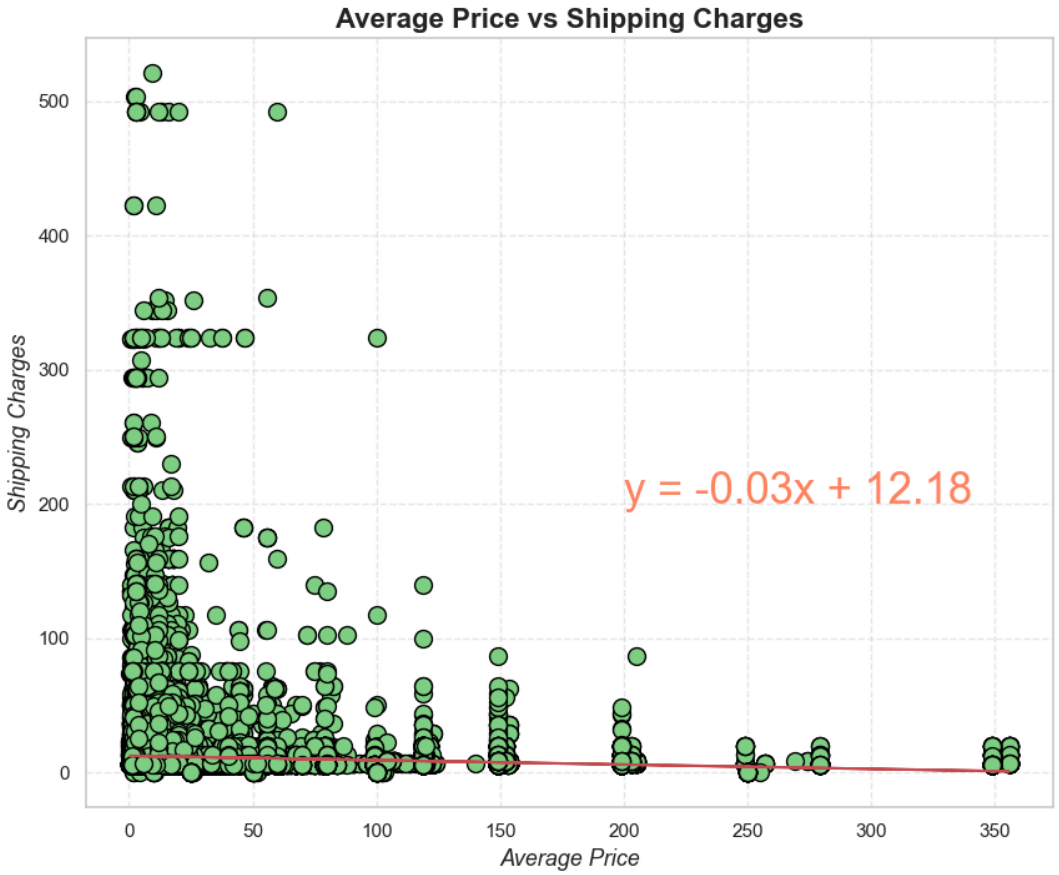


These plots are really interesting to look at in terms of who is spending more online. Many people would think that women may be spending more than men but, that is only true in the California location. In chicago the average spending for men is slightly higher than women and in the NorthEast region men on average are spending $77 more than women. The following leaderboard goes more in depth on the values associated with each location.

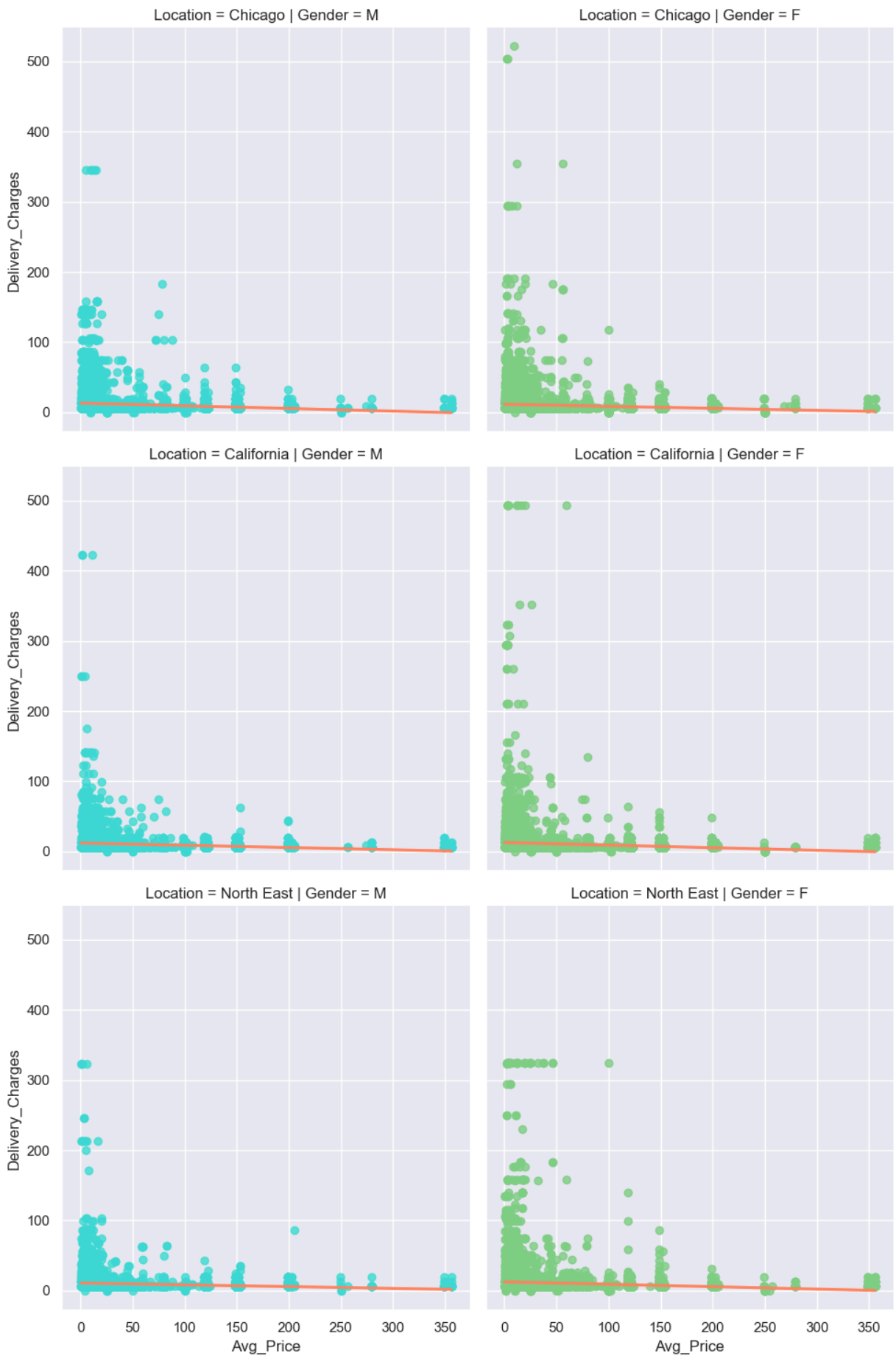
	Location	Gender	Online Avg	Online Med	Online Var	Online Stand Dev	Online SEM
0	California	F	1955.178420	1880.22	736505.498605	858.198985	8.668234
1	California	M	1776.553688	1696.64	625993.779132	791.197687	10.043362
2	Chicago	F	1867.223676	1827.02	608654.840454	780.163342	7.307220
3	Chicago	M	1898.182701	1870.67	541209.958268	735.669735	8.894532
4	North East	F	1888.501010	1783.56	649381.096948	805.841856	7.493043
5	North East	M	1965.651091	2011.57	715694.418111	845.987245	10.327675

It's also interesting to look at the highs and lows of these plots.In Chicago there are four distinct groups of spending among the men while women only have three. In California more men are spending the average $1776.55 while it looks like more women are spending around $1955.18. Overall these violin plots illustrate some very interesting relationships between how men and women are spending their money online.

I then began to look at the relationship between average price of an item and the shipping costs.



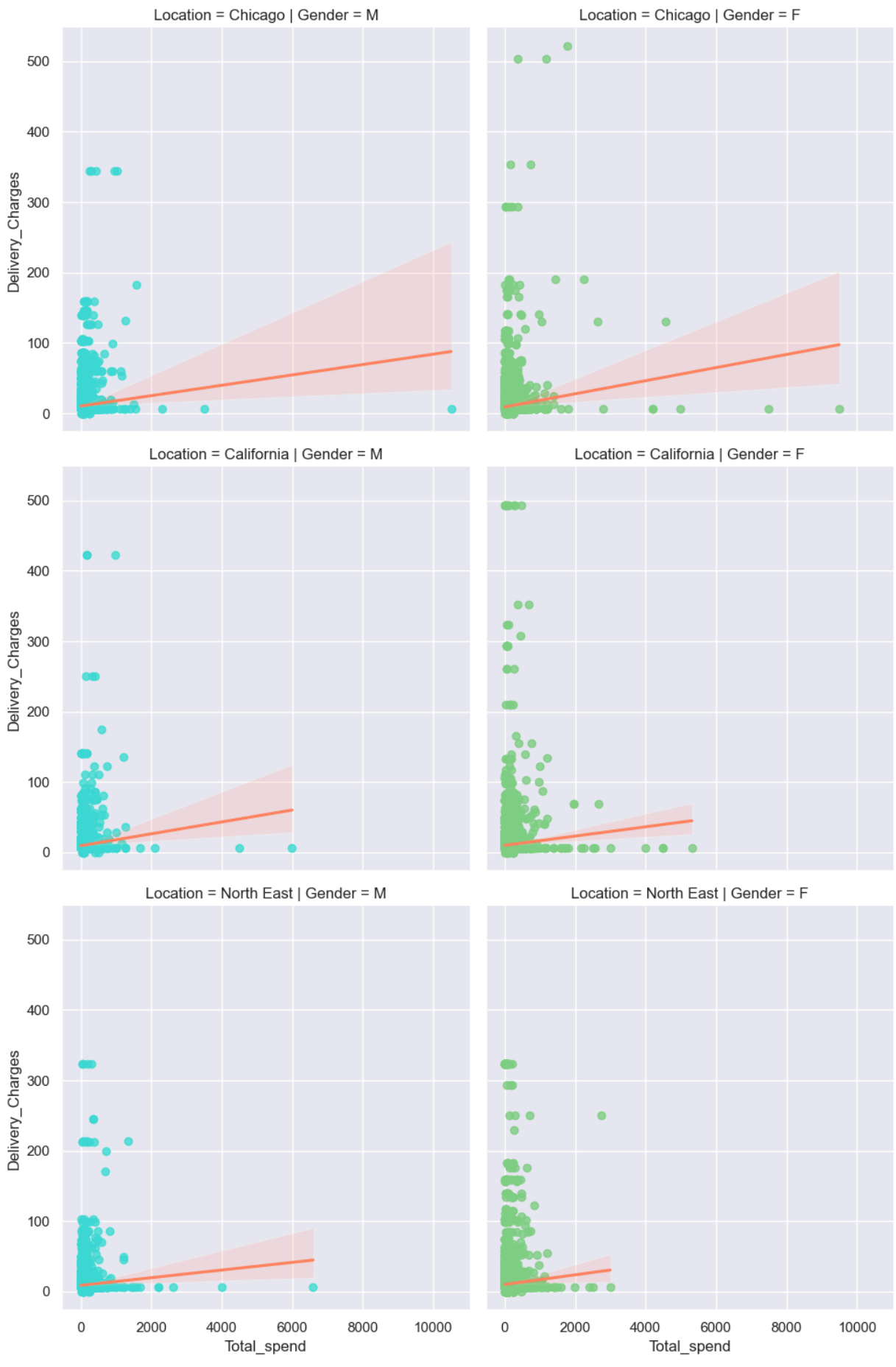
With the r-squared value at 0.011 it only explains about 1% of our data, because only about one percent of the data is correlated or significant there isn't a strong correlation between the average price of the item and the associated shipping charge. There are some interesting outliers on the plot, several items are close to 0 for the average purchase price but have shipping charges close to 500 dollars. This again could be due to the validity of the data. After this initial look at how the average item price and shipping charges were related I wondered how the location and gender affected the relationship. The following visual shows the relationship between gender and location.


Location = Chicago | Gender = M
Location = Chicago | Gender = F
Location = California | Gender = M
Location = California | Gender = F
Location = North East | Gender = M
Location = North East | Gender = F
Delivery_Charges
Avg_Price
0
100
200
300
400
500
0
50
100
150
200
250
300
350

Initially I thought that the r-squared value would change depending on the location and the gender but the r-squared value remained the same at 0.011. This makes sense because this is the same data from the first scatter plot, but just broken down a little bit more. After this realization that the r-squared was the same I did notice some interesting differences about the distribution of the points. I was curious about the outliers on the initial scatter plot, who was spending so little on the product but so much on shipping? By looking at these plots we can see that it seems like only a few women from California and Chicago are spending those prices on shipping. These outliers still seemed outrageous to me, so I had to dig a little bit deeper. The following leaderboard shows the outliers and the relevant info associated with them.

	Gender	Location	Product_Description	Quantity	Avg_Price	Total_spend	Delivery_Charges	Month	Coupon_Status	Discount_pct
1954	M	California	Google Laptop and Cell Phone Stickers	88.0	1.98	174.24	422.24	1	Clicked	10.0
1955	M	California	Google Ballpoint Pen Black	88.0	1.98	174.24	422.24	1	Clicked	10.0
1956	M	California	Leatherette Journal	88.0	10.95	963.60	422.24	1	Clicked	10.0
14306	F	Chicago	Google Spiral Leather Journal	185.0	9.60	1776.00	521.36	8	Used	20.0
28342	F	Chicago	Ballpoint LED Light Pen	475.0	2.50	1187.50	504.00	11	Not Used	20.0
28343	F	Chicago	Four Color Retractable Pen	125.0	2.99	373.75	504.00	11	Not Used	20.0
31975	F	California	Google Stylus Pen w/ LED Light	24.0	4.40	105.60	492.84	2	Used	20.0
32541	F	California	Google Men's Quilted Insulated Vest Black	1.0	59.99	59.99	492.84	2	Clicked	20.0
33265	F	California	Google Canvas Tote Natural/Navy	8.0	12.79	102.32	492.84	2	Clicked	20.0
33266	F	California	Google Zipper-front Sports Bag	16.0	15.99	255.84	492.84	2	Clicked	20.0
33437	F	California	26 oz Double Wall Insulated Bottle	24.0	19.99	479.76	492.84	2	Used	20.0
33723	F	California	Google Sunglasses	7.0	2.80	19.60	492.84	2	Clicked	20.0
33724	F	California	Google Sunglasses	5.0	2.80	14.00	492.84	2	Clicked	20.0
33725	F	California	Google Sunglasses	7.0	2.80	19.60	492.84	2	Clicked	20.0
33726	F	California	Google Sunglasses	5.0	2.80	14.00	492.84	2	Used	20.0
33861	F	California	Google Hard Cover Journal	24.0	11.99	287.76	492.84	2	Clicked	20.0

I filtered the data for all the shipping greater than $400 because I was unsure what the lowest outlier was. But, this leaderboard does somewhat show the reason for them. The scatter plot shows the average price rather than the new column I created called the total spend. Some of the purchases look to be for bulk items with a low average price but the shipping charge shows for the whole order. For fun I re-ran the scatter plots that showed the location based on gender. The r-squared value did go up to 0.232, but still with just two percent of data being correlated there is not a significant relationship between delivery charges and average price or total spending.


Location = Chicago | Gender = M
Location = Chicago | Gender = F
Location = California | Gender = M
Location = California | Gender = F
Location = North East | Gender = M
Location = North East | Gender = F
Delivery_Charges
Total_spend
0
100
200
300
400
500
0
2000
4000
6000
8000
10000

Call to action:
Because of the lack of correlation to be able to draw any meaningful conclusions from the data we would need data from other sources to try and prove/disprove some of our theories on the data.

Bias limitations:
The data comes from several different kaggle datasets making it hard to verify the validity of the data. Because of this there are several things we came across that raised some red flags like several date columns and an offline spend column. These limitations make it hard for us to have a solid call to action. This data didn't have very many locations, and of the locations that it did have it wasn't consistent, we had states and cities, this makes it hard to really compare location spending. The lack of locations also makes it hard to draw conclusions about how the country spends money online.

Future Work:
There is a lot of opportunity for future work for the topic in general, but maybe not with this dataset. All of our data came from 2019, it would be interesting to look at data from the years during the pandemic and post pandemic to see how the average consumer online shopping habits have changed. Using data similar to this dataset you could also better identify online shopping trends and who you should be targeting based on your product.

Works Cited (MLA please)

Divakar, Jackson R. "🛒 Online Shopping Dataset 📊📉📈." *Kaggle*, 12 Nov. 2023, www.kaggle.com/datasets/jacksondivakarr/online-shopping-dataset.

"Xpert Learning Assistant." *Bootcamp Spot*, bootcampspot.com/. Accessed June 2024.

AndrewAndrew 5, et al. "How to Adjust Padding with Cutoff or Overlapping Labels." *Stack Overflow*, 1 Feb. 1957, stackoverflow.com/questions/6774086/how-to-adjust-padding-with-cutoff-or-overlapping-labels.

GrazebrookMichael Grazebrook 5, Michael, et al. "How to Change the Figure Size of a Seaborn Axes or Figure Level Plot." *Stack Overflow*, 1 Mar. 1961, stackoverflow.com/questions/31594549/how-to-change-the-figure-size-of-a-seaborn-axes-or-figure-level-plot.

"How to Change the Size of Axis Labels in Matplotlib?" *GeeksforGeeks*, GeeksforGeeks, 3 Jan. 2021, www.geeksforgeeks.org/how-to-change-the-size-of-axis-labels-in-matplotlib/.

tylerthemilertylerthemiler 5, et al. "Savefig Outputs Blank Image." *Stack Overflow*, 1 Sept. 1957, stackoverflow.com/questions/9012487/savefig-outputs-blank-image.

user3287545user3287545 2, et al. "Seaborn Lmplot with Equation and R2 Text." *Stack Overflow*, 1 Apr. 1960, stackoverflow.com/questions/25579227/seaborn-lmplot-with-equation-and-r2-text.