# Online Shopping Analytics

Hedyeh Erfani

DSC 680

Abstract

The retail space has been changing for quite some time. We’ve seen a shift from in person shopping to online shopping. This trend has increased as COVID started impacting business operations. More and more people have started shifting towards online retail. While in this project I will not be analyzing current online retail trends, I’ll be looking at online retail in general to learn more about online purchases.



Data

The dataset contains data which was collected between 2009 and 2011. It contains data points which reflect the purchases for an online store in the United Kingdom. Below are the attributes of the dataset

InvoiceNo: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.  
StockCode: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.  
Description: Product (item) name. Nominal.  
Quantity: The quantities of each product (item) per transaction. Numeric.  
InvoiceDate: Invice date and time. Numeric. The day and time when a transaction was generated.  
UnitPrice: Unit price. Numeric. Product price per unit in sterling (Â£).  
CustomerID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.  
Country: Country name. Nominal. The name of the country where a customer resides.

These measurements were recorded by the owner of the store. There are over 1M rows of data in this dataset with giftware being the main item sold

Walkthrough

This excel file had two sheets. The first sheet contained years 2009-2010 and the second sheet contained the years 2010-2011. The first step I had to take was combining both sheets into my dataframe so that I could analyze all of the data.

Table

Description automatically generated

Once this dataset was created, I was able to proceed with further analysis. I took a deeper dive to make sure that the data was cleaned and ready. It quickly came to my attention, however, from looking at df.describe(), that the minimum value was incredibly inaccurate. Since it was showing as a negative number, I knew that there were negative values in the Quantity and Price column which would need to be removed, so I then removed them using a filter.Table

Description automatically generated

I then created a new column titled ‘revenue’ so that it would be easily identifiable what the shop’s overall revenue is based on quantity and price. Histograms were subsequently created so that we could see the distribution of our variables.

Chart

Description automatically generated

I also did a log transformation on the Quantity column. I did this to take a closer look at the skew we see above.

Chart, histogram

Description automatically generated

Histograms were also created for Price and Revenue. These histograms were practically identical to the histogram we saw for Quantity.

I wanted to take a closer look at how Quantity over time appeared. Here is the plot for that.

Histogram

Description automatically generated with medium confidence

From there I moved onto bivariate analysis. I plotted Quantity against Price with a scatter plot to see if there is a relationship between these two columns.

Chart

Description automatically generated with low confidence

A relationship isn’t too evident from this scatterplot.

To continue my analysis, I looked at things such as value count. The store had 41,713 unique invoices (orders). This can be compared with the original number of rows that existed. From those 41,173 unique invoices, there were 5,873 unique customers (retrieved from Customer ID). From that analysis, I saw that orders had come in from 41 unique countries. The most purchased item was a ‘White Hanging Heart T-Light Holder’, which was purchased 5,648 different times. In terms of revenue, the top revenue came from the U.K and Netherlands.

In the last step of my analysis, I started my focus on Customer ID. I calculated recent purchases, days since last purchase, and frequency of purchase. From there I also closely investigated revenue.

Conclusion

Even in 2009-2011 when online retail was not as common as it is today, seeing the numbers behind the profitability and the consumer drive from that time period to go toward online retail makes this shift seem even more driven for today and the future. Companies can touch many different markets that they hadn’t thought possible, even in regions such as the West Indies. The growth in this sector had started a while before 2009 but I can see the growth for the future after this analysis.