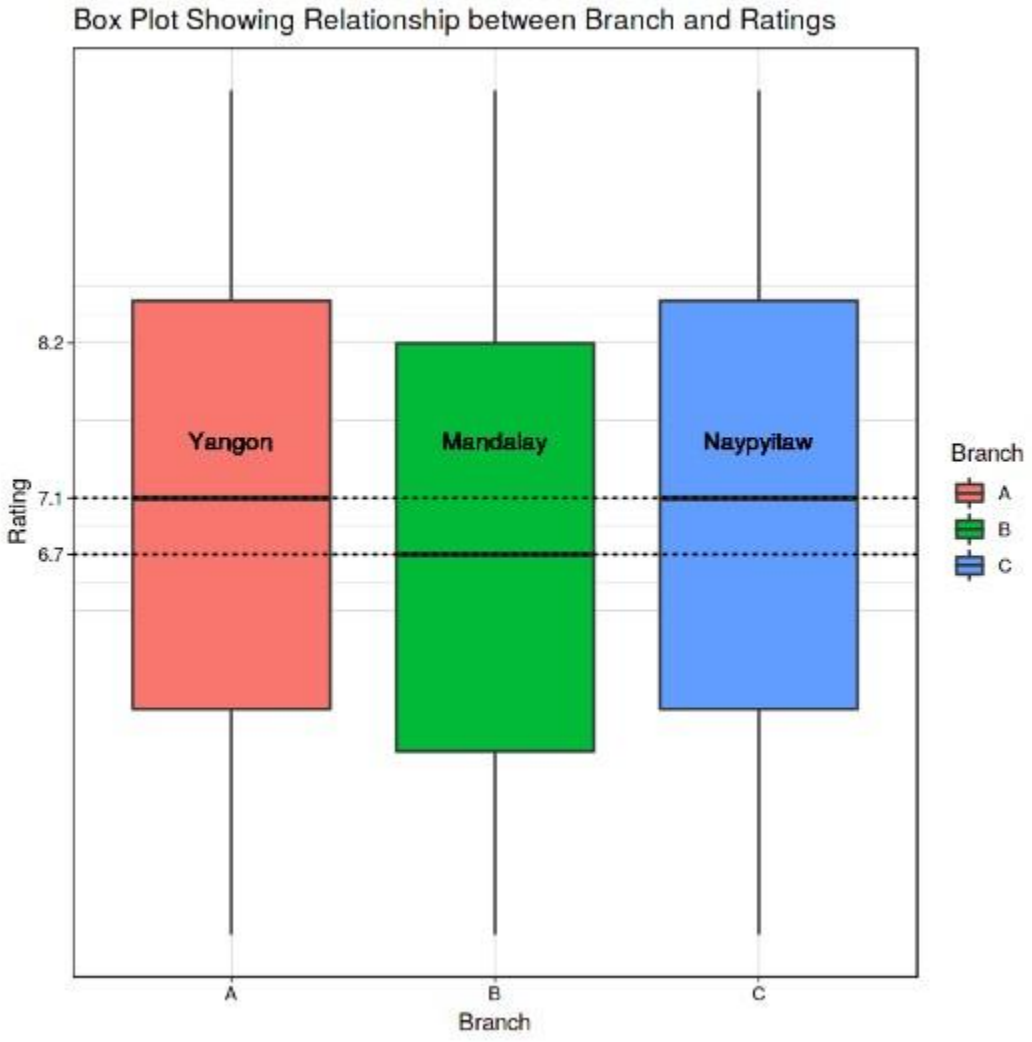
IST 707 Applied Machine Learning

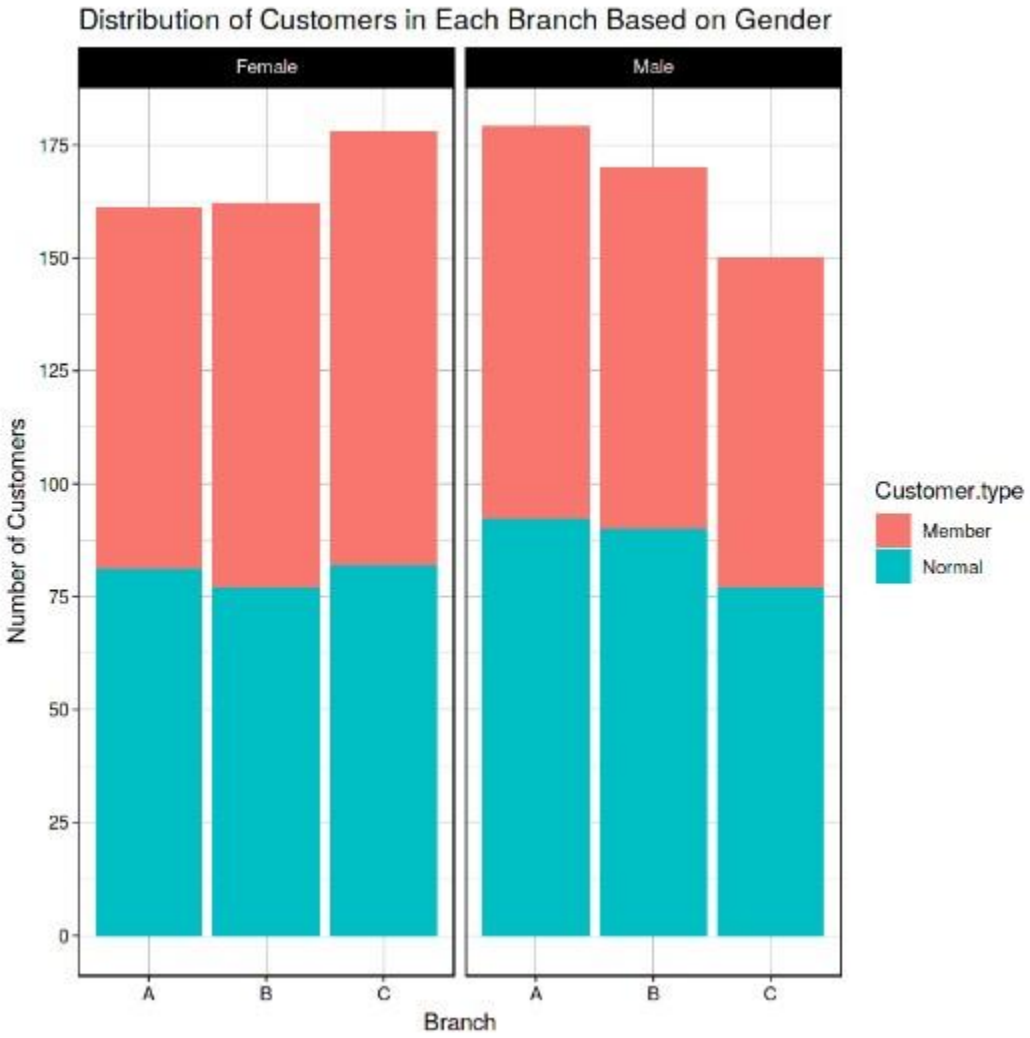
HW2: Tell a Data Story

Name: Jainish Savaliya

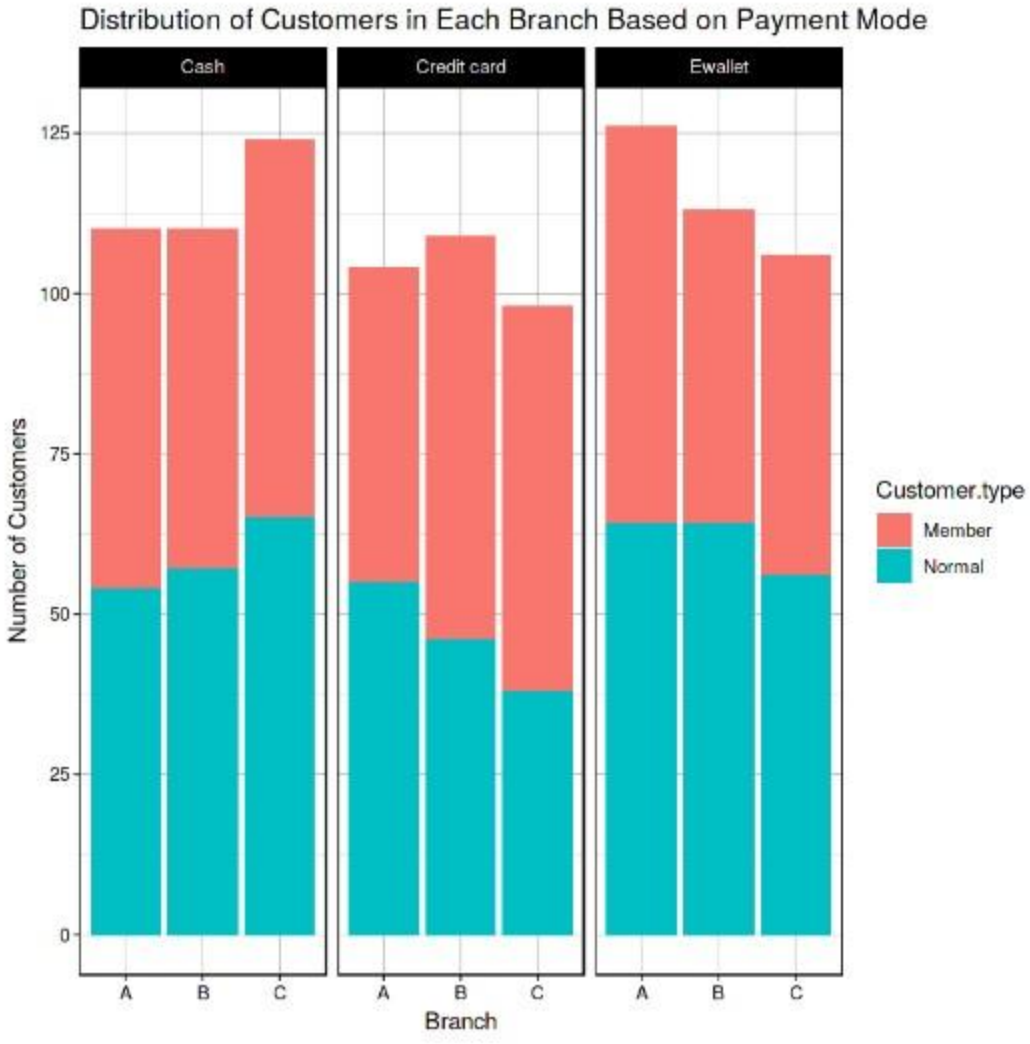
- First of all, I opened that dataset and tried to go through the data by summary and I found out that Invoice Id has nothing to do with this analysis so I removed that and looked for any missing values and tried to remove those. After looking at the summary I looked at all the means and medians and tried to figure out the percentile of the ratings of all three branches so I created the visualization of the box plots of branches A, B & C and got to know that A& C were almost similar and Branch B was little lower in the ratings.



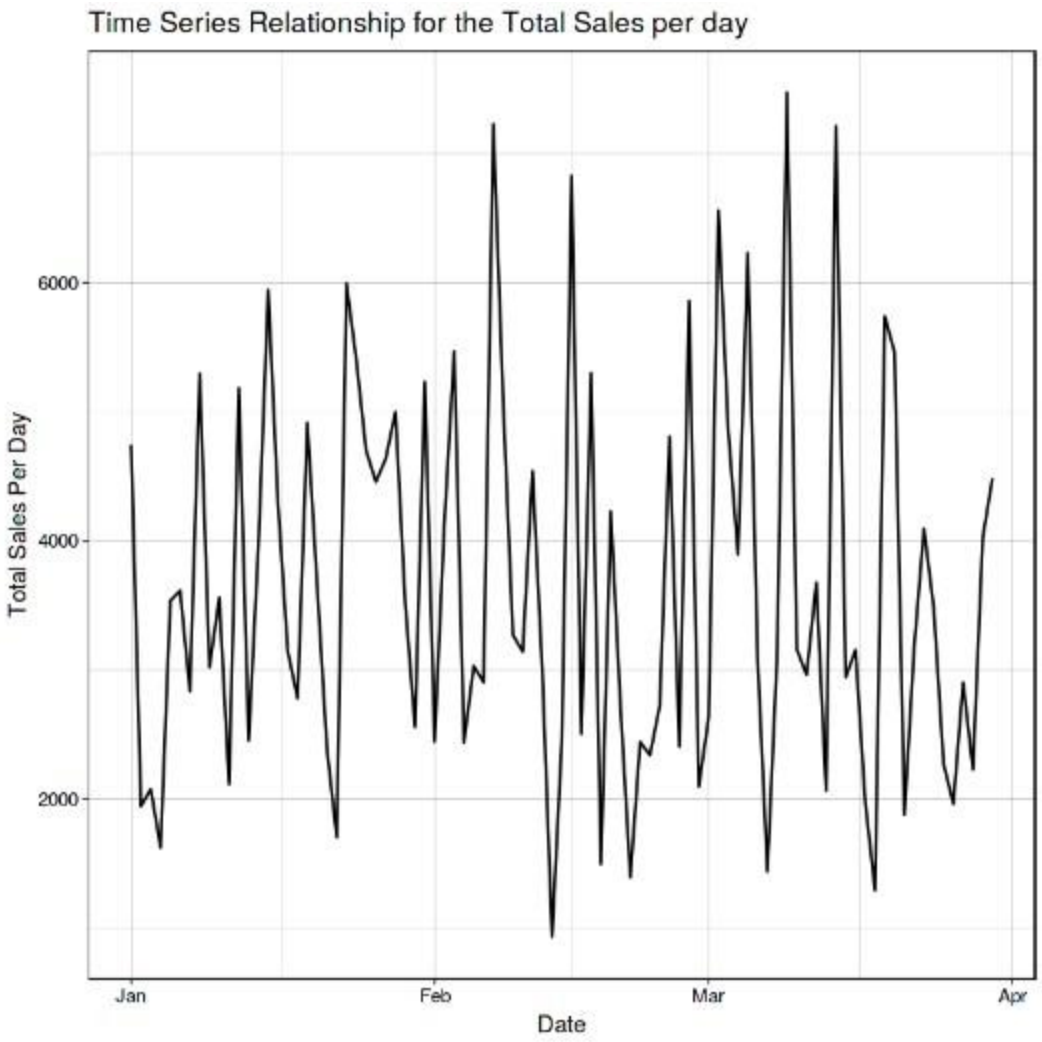
- Then I saw the types of customers there were customers who were members and rent of them were not members so and also I looked for this data alongside the gender of the customers. I divided the bar plot in such a way that can same female and male customers also inside this category there is a sub-category of members and normal customers. On average there are more female customers compared to male customers and the distribution between members and nonmember customers is almost 50-50.



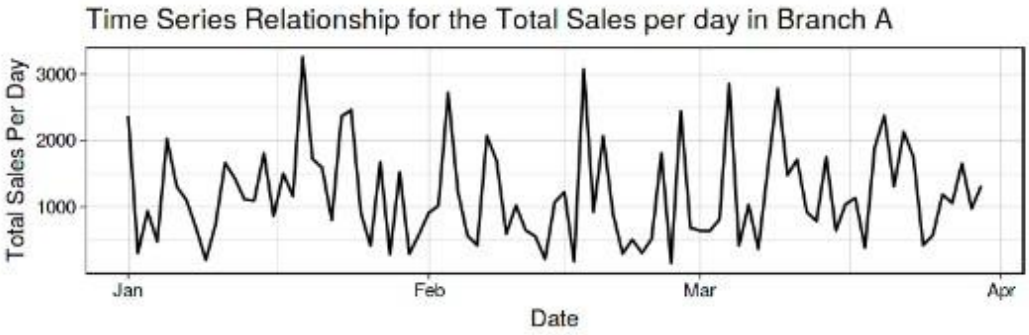
- After that, I tried to change the category a little bit to look for the most used payment methods between members & non Members so I used the same method as above and separated cash, credit_card &Ewallet purchases among all three branches. I found out that the most used payment method was Ewallet and branch A was the location where this method is the most preferred one and then comes cash and then a credit card. So we can conclude that all kinds of customers are preferring cashless transactions like Ewallet more over the rest of the methods.

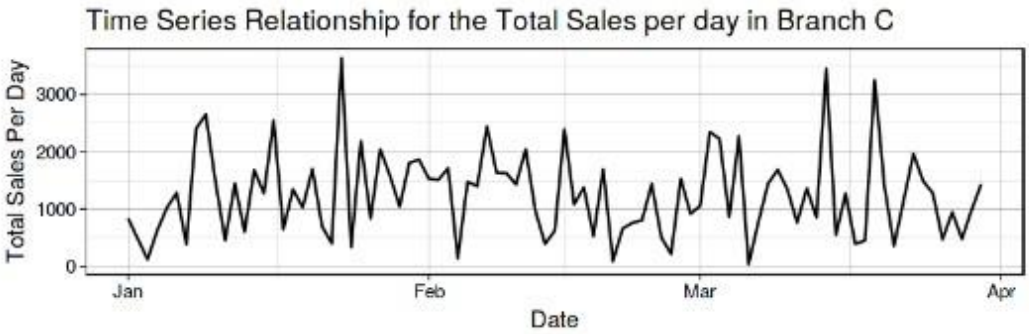


- The next thing that came to my mind is to check the sales now the best way to get insight from the sales data according to me is through time series analysis so I created the line graph to check the relationship between total sales per day over months and here I observed that sales data had a lot of fluctuation over the month and within the months during the week and during the march and April the sales went the highest and by the end of April it started decreasing.

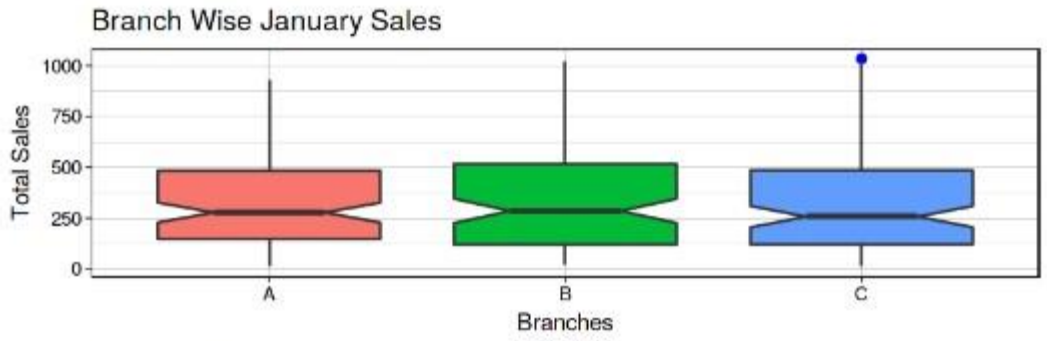


- While the above time series chart was for combined of all three branches over time I decided to look for individual branches to do the analysis of the pattern and I found out that While almost all the branches gave a similar pattern to each other There were mostly two highest peaks observed in all three branches first was during mid-Jan and second was during the mid march. However, there was another peak during mid fab but it wasn't that significant over the recent of the two peaks.



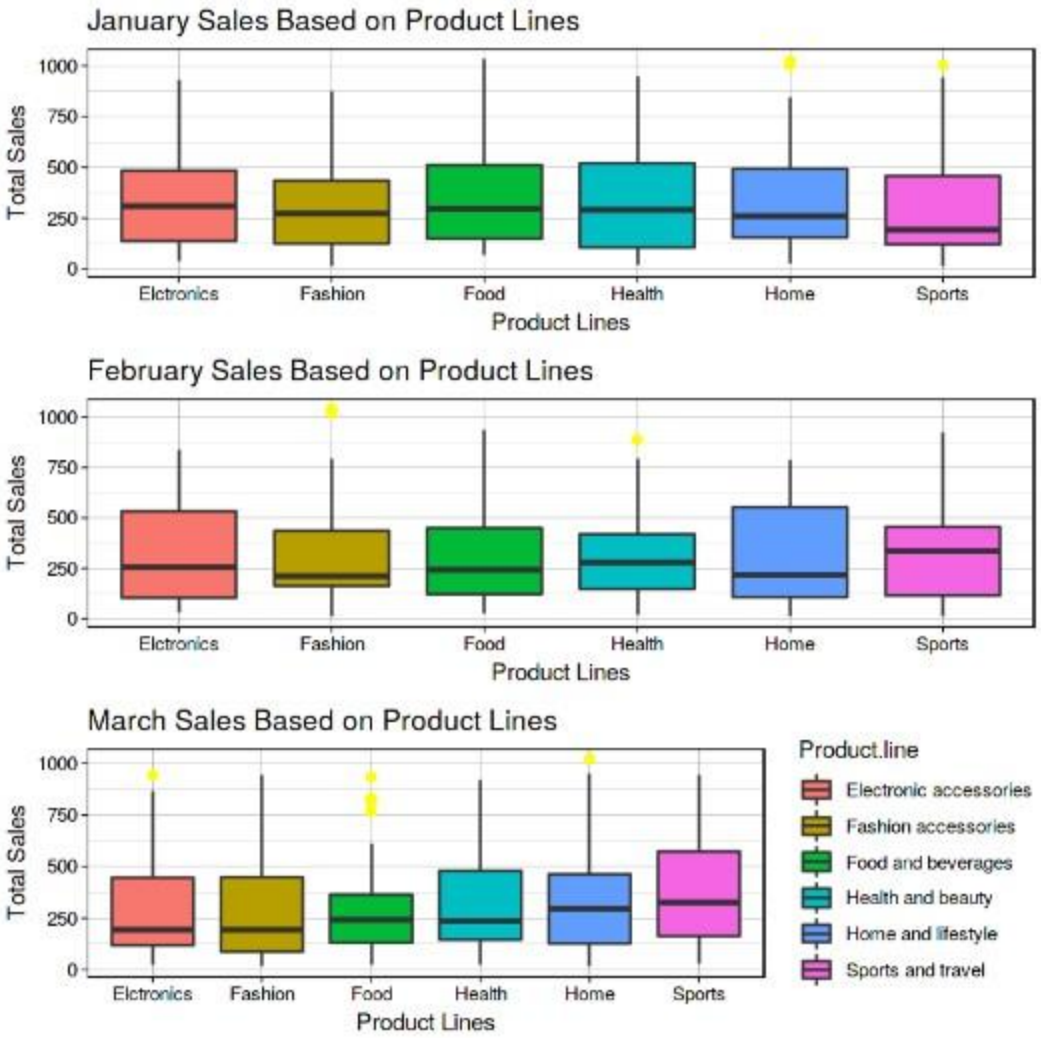


- To get more clarity on the distribution I also Visualized Monthwise sales of all three branches and with the notch to get an idea of the confidence percentages as well. And found similar insights as the time series.


Branch Wise January Sales
Total Sales
1000
750
500
250
0
A
B
C
Branches


Branch Wise February Sales
Total Sales
1000
750
500
250
0
A
B
C
Branches


Branch Wise March Sales
Total Sales
1000
750
500
250
0
A
B
C
Branches
Branch
A
B
C



- After all those categories I went to check all the sales patterns of the categories of products by creating box plots separately for all the categories for product lines and dividing those sales by months & highlighting all the outliers. I found out that the percentile wise or by looking at the means for Jan food and health was on the highest side and sport was the lowest. But During the Feb & March on average sports items got sold the most in the category sales-wise. Finally Looking at the highest sold categories Food was highest during Jan & Fab and then In march Food dropped drastically and other categories like home & Fashion increased in march. By Doing all this analysis It could be analyzed what changes should be made regarding target customers, Payment systems, inventory management, revenue generation & improvements to make.