**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

For this business analytics project, I have chosen a dataset on supermarket sales from the following link: https://www.kaggle.com/akshitmadan/complete-data-analysis-supermarket-dataset/data?fbclid=IwAR3QSKhR2I7lEezbzLTJwvdOVBRcASMmmi_FCpX3J5RQl7ZeqLCe0Bu3lnM

Workings used in the analyses can be found here: https://drive.google.com/drive/folders/1um3vUl2Zdc-YQ4CB9zeQzxJv8Kkwo4V3?usp=sharing

## About the dataset

The dataset consists of the following data:

- Invoice ID
- Branch of the supermarket
- City of the supermarket
- Member or normal customers
- Gender of the customers
- Product lines
- Unit price of each product in $
- Quantity or number of products purchased by a customer in each transaction
- 5% tax fee
- Total amount spent by the customer
- Date of purchase
- Time of purchase
- Mode of payment used by the customer
- Cost of goods sold
- 7% gross margin percentage
- Gross income of the store
- Customer rating

Of these, 5% tax fee and 7% gross margin percentage are constant throughout the data. The dataset contains sales data for only the months of January, February and March of 2019. The supermarket has **three** outlets in total, with one branch in each of the cities Yangon, Naypyidaw and Mandalay of **Myanmar**. Hence, we can ignore either the branch or the city field. The total amount spent by a customer is dependent on unit price (which in turn is dependent on cost of goods sold), quantity of products bought, and the 5% tax. I have also divided the time of purchase into 4 distinct times of the day: morning, noon, afternoon and evening.

## Key assumptions

The source link doesn't provide any metadata on the dataset. For this reason, I have made the following assumptions:

- The shops are all located in urban areas of the cities
- Since the dataset spans only 3 months, I initially thought it could be that of a new supermarket that has just been in operation for 3 months. But it is unlikely that a new supermarket will provide membership facilities right off the bat. I have therefore

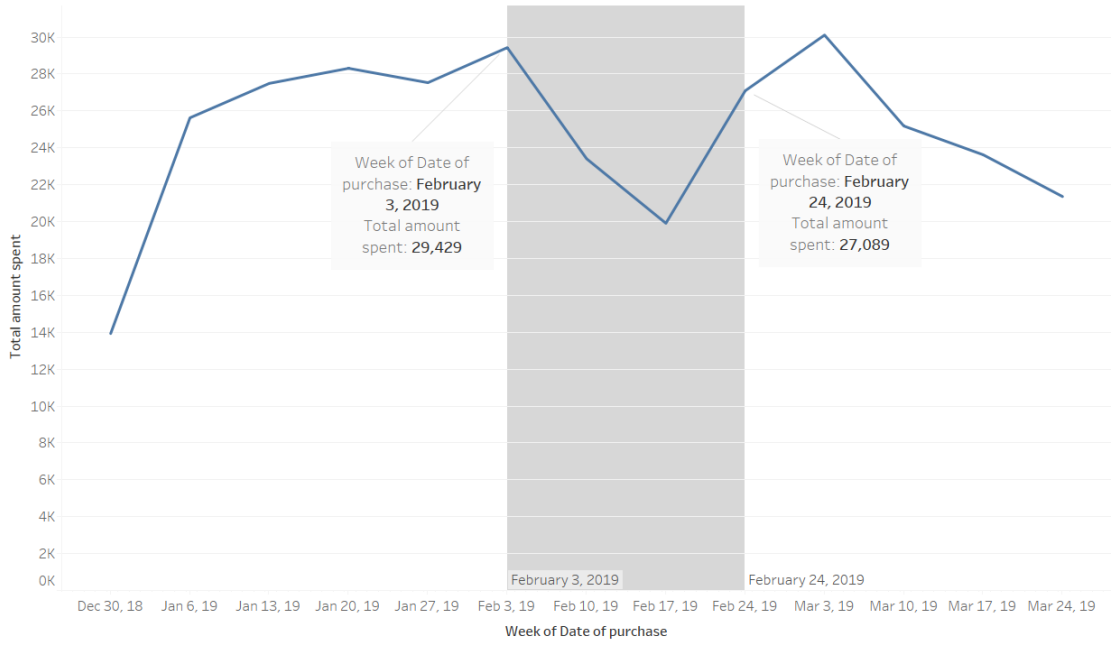
hypothesized that the supermarket has taken the data for analyzing something specific. I want to analyze what that could be.

- While we do not know if any of these customers are recurring, we can see that each transaction is limited to one type of product. This is not a practical real-world scenario, but could be a simplification by the supermarket itself for particular reasons. Maybe, for the simplicity of analysis, they have considered that a transaction belongs to a particular product line when the greatest number of goods bought in that transaction belongs to that product line.

## Exploring and analyzing the data

My first course of action was to go with assumption that **the supermarket has taken the data for analyzing something specific**. While exploring the dataset and going through visualizations, the first thing that caught my eye was that the sales for the month of February was low almost throughout the month.

### Sales increased and were steady in January, but then it started free-falling from the beginning of February




The trend of sum of Total amount spent for Date of purchase Week. The data is filtered on Product line, which keeps 6 of 6 members.

As we can see, sales started decreasing from the first week of February. Sales hit the lowest point of the month in the 3rd week and then reach levels similar to that of January around the last week of February. Identifying the reason behind **what went wrong around February** could be a potential reason behind a company's curiosity behind analyzing the three-month sales data.

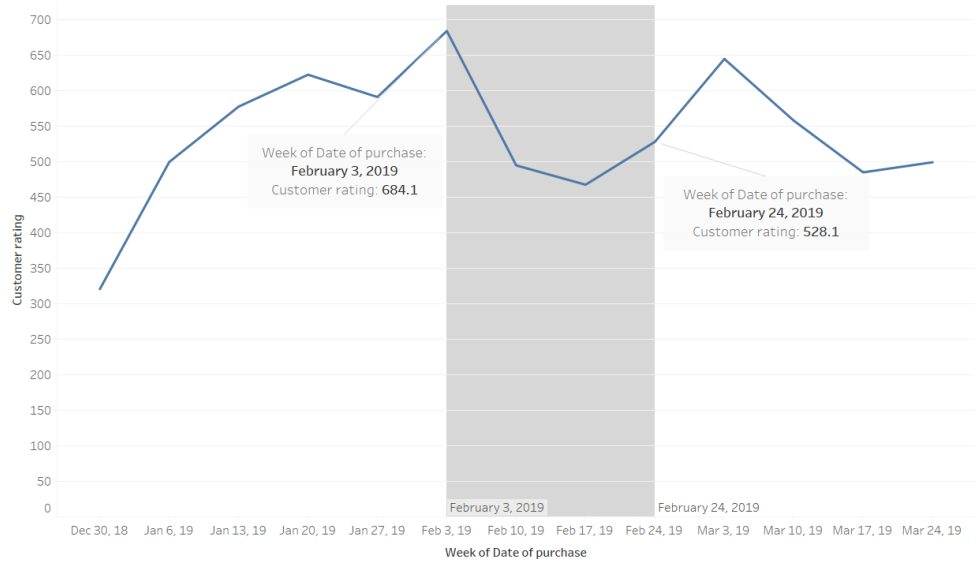
### What could have caused this decrease in sales?

In order to find out if any internal factor of the supermarket might have caused this decrease in sales, I have tried to analyze the change in customer ratings and price of products sold.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

Firstly, as we can see from the following graph, the fall in ratings seem to correspond to the trend of decreasing sales in February. So internal quality control, poor customer service or stockouts could be a possibility. Stockouts would simultaneously reduce sales and cause customers to have a bad experience, thus bringing the customer ratings down.

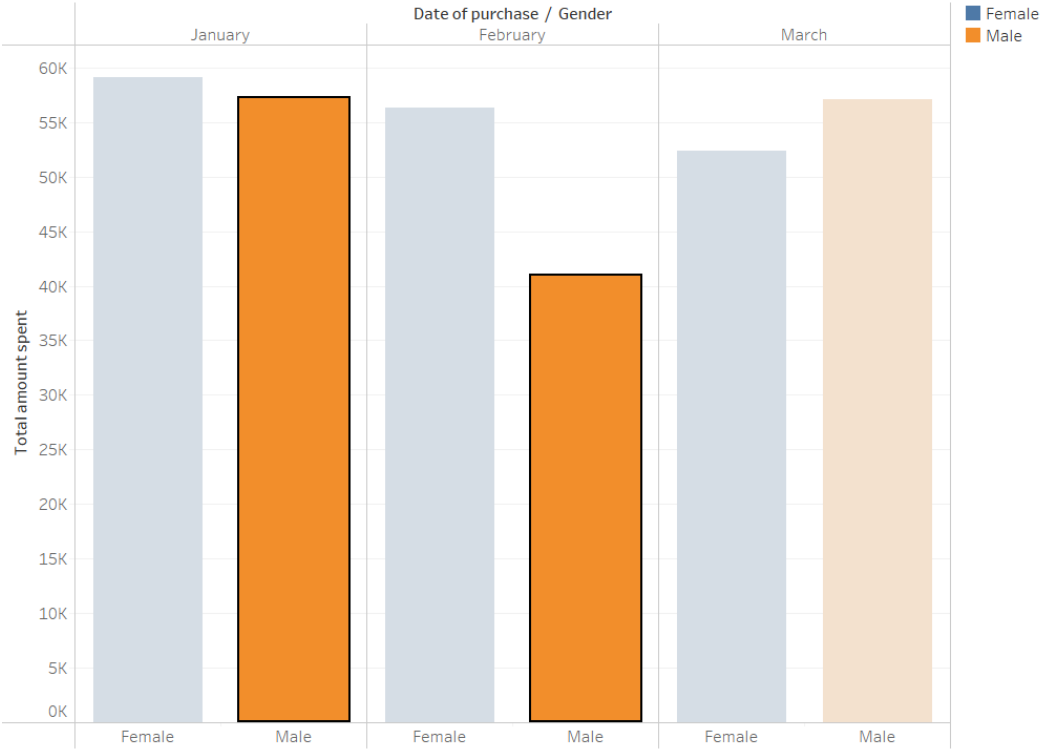
## Customer ratings plummeted at the same time of the February freefall




The trend of sum of Customer rating for Date of purchase Week. The data is filtered on Product line, which keeps 6 of 6 members.

While the analysis of ratings could give a plausible reason about the scenario, I wanted to explore further into this.

## Number of male customers were significantly low during February




Sum of Total amount spent for each Gender broken down by Date of purchase Month. Color shows details about Gender.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

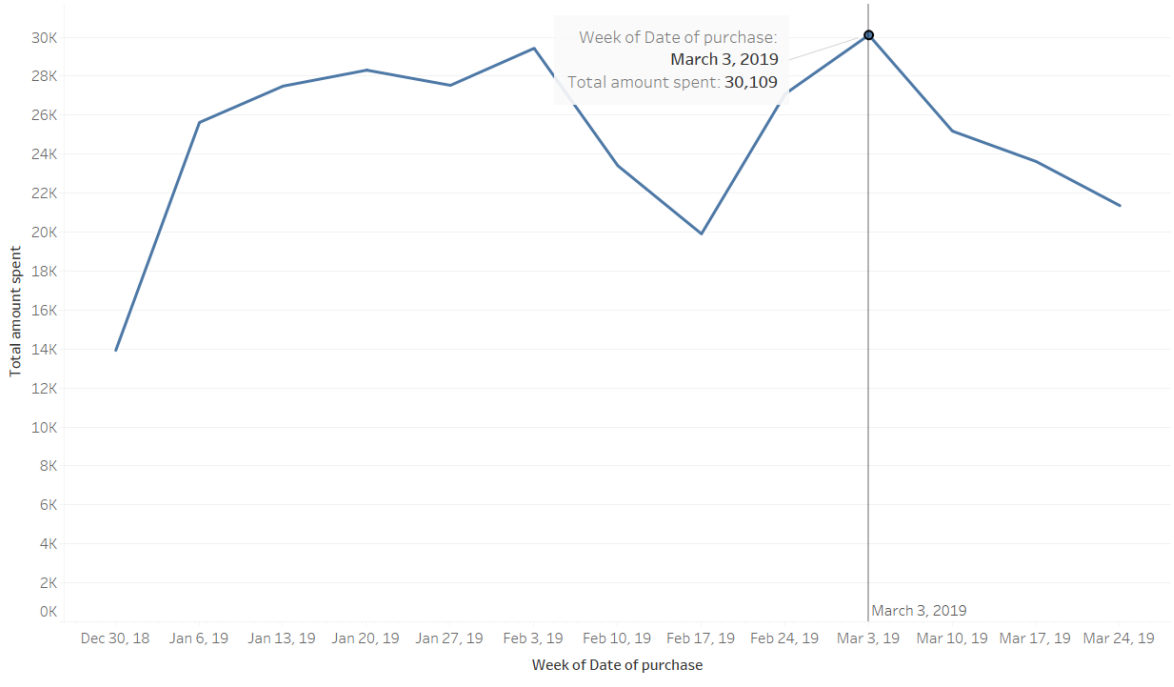
A significant finding from this analysis is that, February had the least number of male customers. I wanted to try to see if there were any significant correlation of sales or any other variables with that of male customers.

## No significant correlation between gender or any other variable

| | | | |
|---|---|---|---|
| 1 | +0.085 | Gender=Male | Time |
| 2 | -0.074 | Gender=Male | Quantity |
| 3 | -0.068 | Gender=Male | Time Horizon=Morning |
| 4 | +0.068 | Gender=Male | Product line=Health and beauty |
| 5 | -0.058 | City=Naypyidaw | Gender=Male |
| 6 | +0.054 | Gender=Male | Payment=Ewallet |
| 7 | -0.049 | Gender=Male | Total |
| 8 | -0.049 | Gender=Male | gross income |
| 9 | -0.049 | Gender=Male | cogs |
| 10 | -0.049 | Gender=Male | Tax 5% |
| 11 | +0.048 | Gender=Male | Time Horizon=Evening |
| 12 | +0.043 | Date | Gender=Male |
| 13 | +0.040 | Customer type=Normal | Gender=Male |
| 14 | +0.039 | City=Yangon | Gender=Male |
| 15 | -0.037 | Gender=Male | Time Horizon=Noon |
| 16 | -0.036 | Gender=Male | Product line=Fashion accessories |
| 17 | -0.031 | Gender=Male | Payment=Credit card |
| 18 | -0.026 | Gender=Male | Product line=Sports and travel |
| 19 | +0.015 | Gender=Male | Unit price |
| 20 | -0.015 | Gender=Male | Product line=Food and beverages |
| 21 | +0.006 | Gender=Male | Product line=Home and lifestyle |
| 22 | +0.005 | Gender=Male | Rating |

At this point into the analysis, I was a bit stumped at the sudden decrease in male customers and the lack of any apparent reason that could be influencing this fall. I decided to move on and see what happened afterwards and how the supermarket started recovering.

## Sales recovers and reaches a three-month high in March



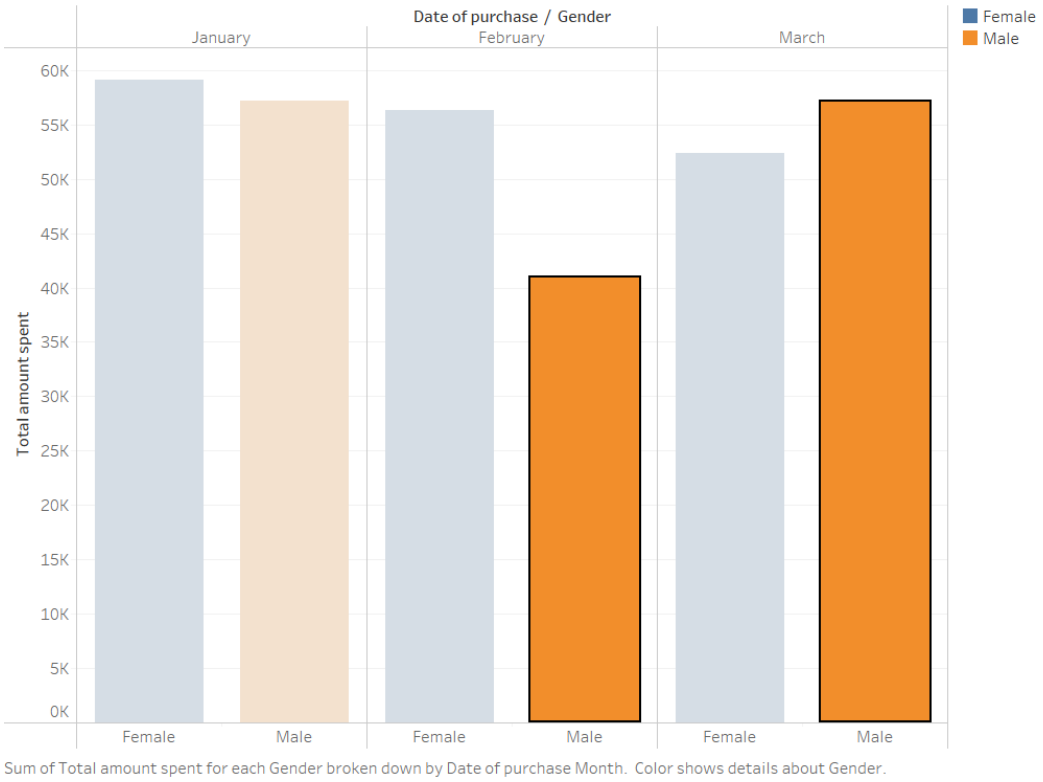

The trend of sum of Total amount spent for Date of purchase Week. The data is filtered on Product line, which keeps 6 of 6 members.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

This increase in sales was certainly a positive for the company. It could be because they had improved the quality of their service, because the customer ratings also rose following the same trend during this time.

**But what about the male customers who brought the lowest sales in February?**

## In March, sales from male customers reached levels similar to that of January




Sales from male customers increased in March as much as it had decreased from January to February. But the time series shows that sales again started decreasing after peaking in March.

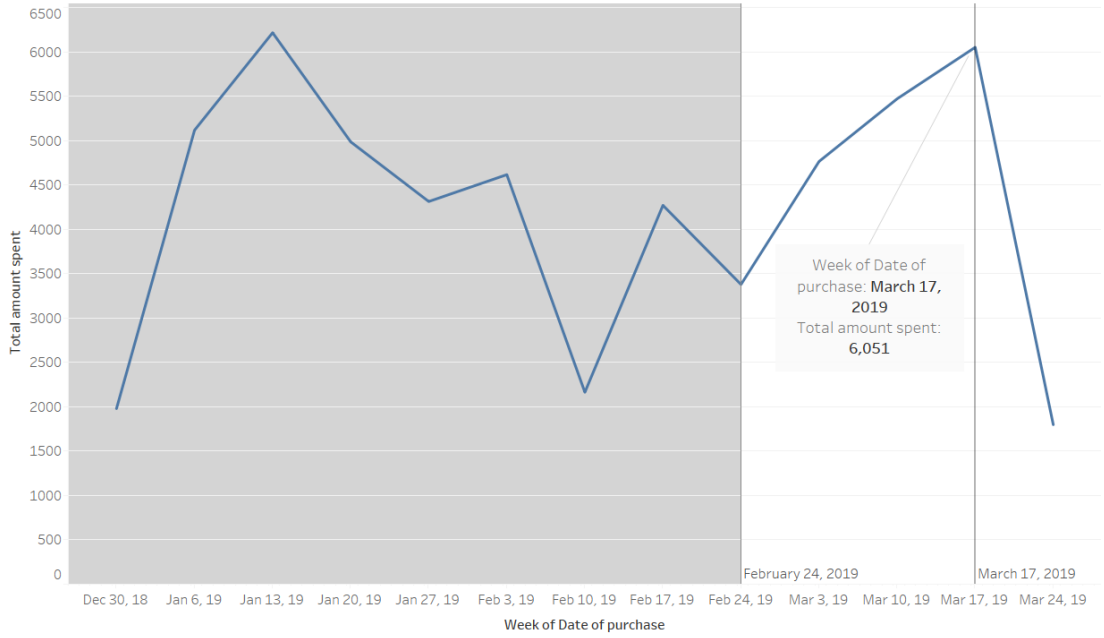
**Could this mean that this increase in sales in March is a bubble?**

This is where I consulted the internet. And I found out that Myanmar has a series of temple festivals starting from the end of February to the end of March. And end of February is exactly when sales started recovering. March is when the people of Myanmar celebrate the ***Shwedagon Pagoda Festival,*** the biggest pagoda festival of the year and festive occasion.

The pattern of less spending by males in February and more spending in March might suggest that, **the male customers preserved their earnings initially and held back on shopping in order to buy for their loved ones in March, when it's the festival occasion**. For the ***Shwedagon Pagoda Festival,*** people from all over Myanmar travel to Yangon. So, a general assumption can be that travel accessories or such type of items would be in high demand during this time and hence their sales would be high.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

## Sales of Sports and Travel started increasing since end of February and peaked right at the beginning of *Shwedagon Pagoda Festival*

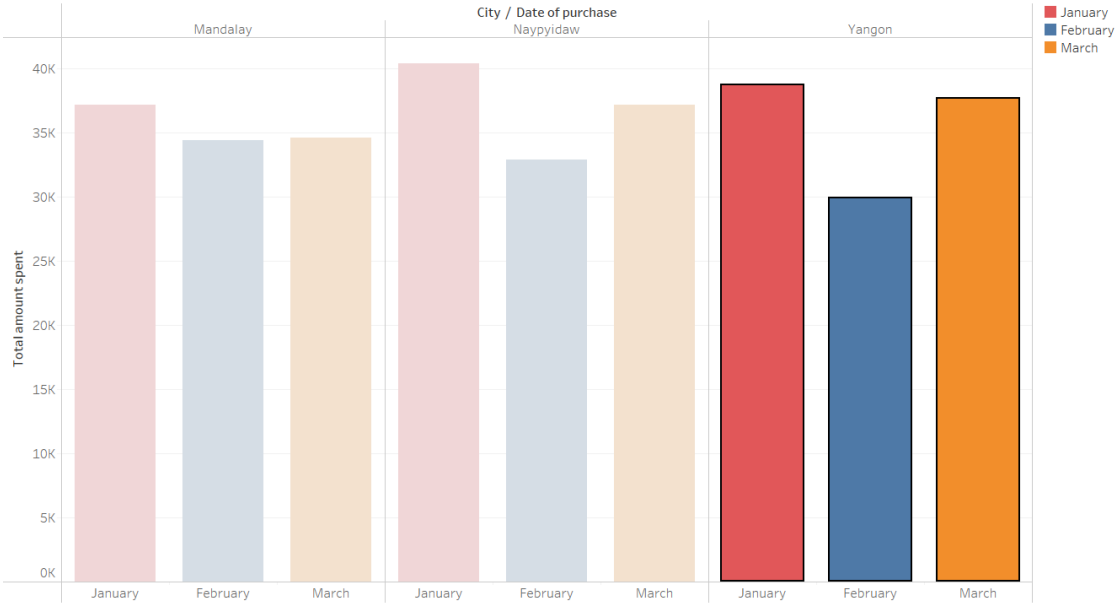



The trend of sum of Total amount spent for Date of purchase Week. The data is filtered on Product line and City. The Product line filter keeps Sports and travel. The City filter keeps Mandalay, Naypyidaw and Yangon.

Sales of Sports and Travel goods started increasing right at such a time that indicates that people buying them are most likely doing so in preparation of attending the country's biggest Pagoda festival. But this Pagoda and the festival is in Yangon.

So, **how did Yangon's sales look during this period?**

## Yangon branch of the supermarket recovered in March during the Festival period after February's decline in sales




Sum of Total amount spent for each Date of purchase Month broken down by City. Color shows details about Date of purchase Month.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

The sales recovery in Yangon could be driven by the influx of people coming from all over the country. But sales data shows a very steep fall after the end of the festival period. **This could mean that the increase in sales experienced by the shop was indeed a bubble.** Moving forward, further analyses of the months would reveal more details.

The difference in purchasing pattern of male during the months of February and March made me interested to look into how exactly males differed from females.

## Decline in purchases were much more prominent among males in February, but they recovered at much greater rates in March

| Gender | City | Date of purchase: January | February | March |
|---|---|---|---|---|
| Female | Mandalay | | 6.36% | -25.04% |
| | Naypyidaw | | -9.95% | -16.01% |
| | Yangon | | -9.58% | 27.59% |
| Male | Mandalay | | -21.03% | 34.58% |
| | Naypyidaw | | -30.14% | 63.28% |
| | Yangon | | -33.64% | 24.47% |

-63.3% 63.3%

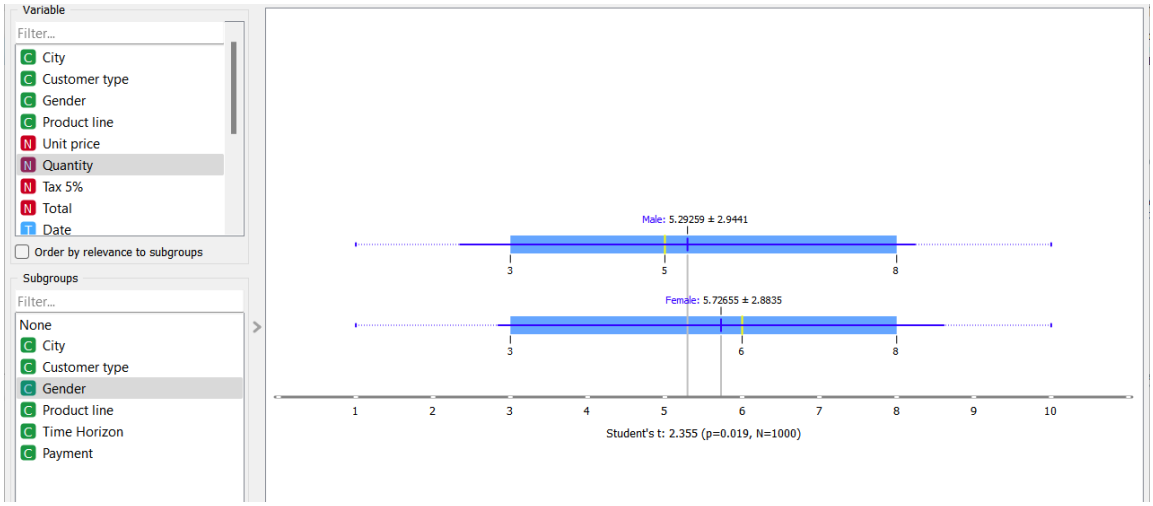
% Difference in Total amount spent broken down by Date of purchase Month vs. Gender and City. Color shows % Difference in Total amount spent.

This insight made me hypothesize that there is indeed prominent difference in the purchase patterns of male and female customers of the supermarket.

**But is this difference statistically significant?**

In order test my hypothesis, I considered two outcomes – either male and female customers don't behave differently from each other (the null hypothesis) or they are very different from each other (alternate hypothesis).

## Female customers tend to buy in higher quantities compared to male customers



**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

After t-test and ANOVA analyses across variables, the only statistically significant finding allows me to say that the average quantity of goods bought by female customers is different from the average quantity of goods bought by male customers.

## Using predictive modeling

While on the topic of male and female customers having different purchase patterns, I tried to see if predictive modeling techniques would allow me to identify when a customer would be a female and when would they be a male.

## Evaluating model fitness for predicting gender of the customer

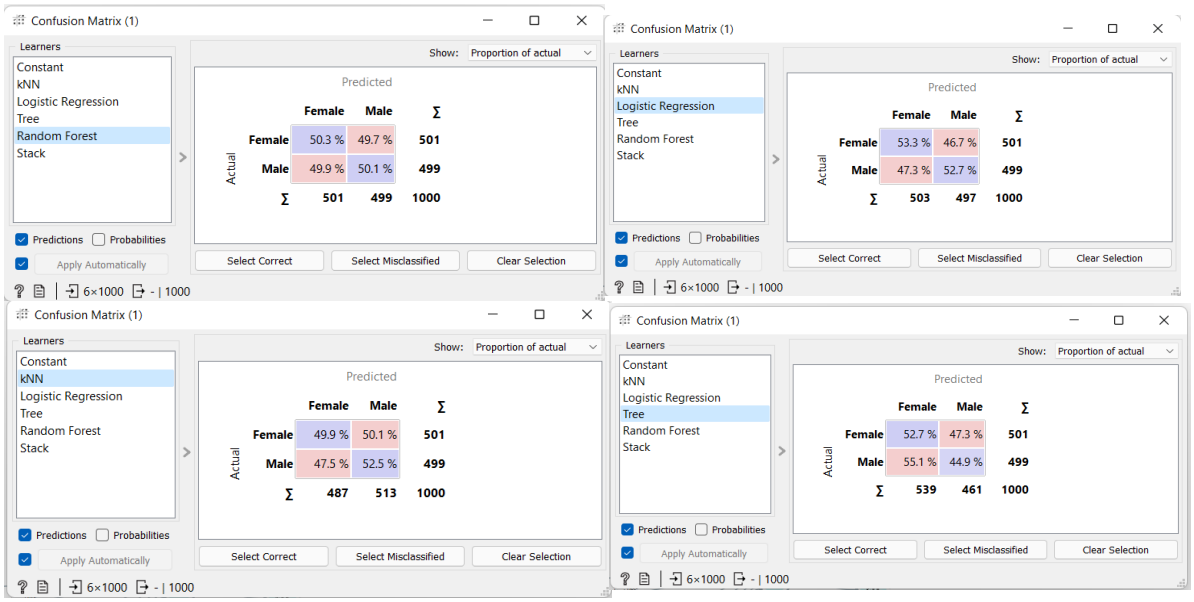
Sampling
- Cross validation
  - Number of folds: 5
  - Stratified
- Cross validation by feature
- Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %

Evaluation Results

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Logistic Regression | 0.533 | 0.530 | 0.530 | 0.530 | 0.530 |
| kNN | 0.526 | 0.512 | 0.512 | 0.512 | 0.512 |
| Random Forest | 0.509 | 0.497 | 0.497 | 0.497 | 0.497 |
| Stack | 0.504 | 0.490 | 0.490 | 0.490 | 0.490 |
| Tree | 0.499 | 0.488 | 0.487 | 0.488 | 0.488 |
| Constant | 0.498 | 0.499 | 0.449 | 0.498 | 0.499 |

I tried using different types of modelling techniques – logistic regression, k nearest neighbor, random forest, tree and even constant model as a bar to judge the rest. Here, precision tells me the accuracy of the prediction being correct. That is, it shows the possibility that my customer is in reality a female given that the model has predicted them to be female. Recall tells me the probability of the model identifying a customer as a female. The F1 score is an average of these two metrics. Although we can see that logistic regression is the best model out of these, we should still look at the confusion matrix.

## Evaluating model fitness for predicting gender of the customer using confusion matrix



**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

Observations show that the confusion matrix gives around half of the results as either false positives, or false negatives. Multiple trials with a different ratio of male and female customers give results in the same ratio. Hence, we can conclude that our dataset isn't a good fit for predictive modelling.

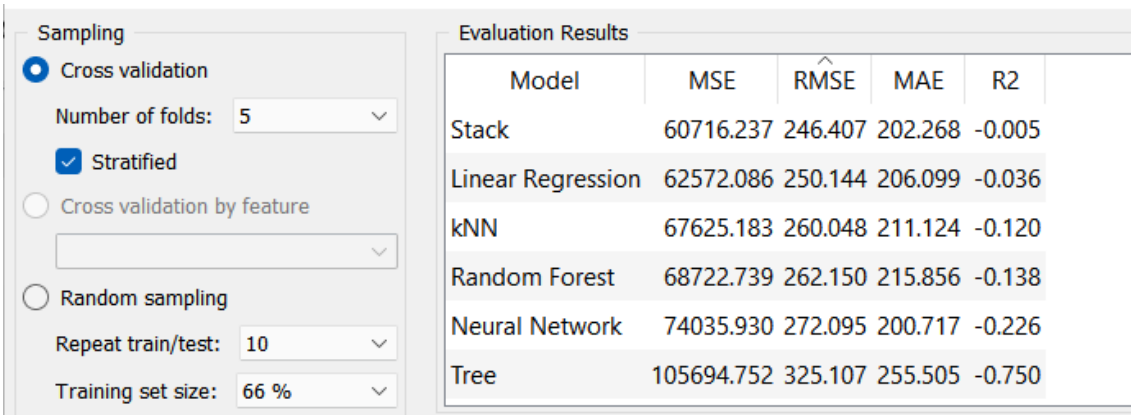
**But what if we tried to predict the amount spent my customers?**

In order to predict the total amount spent by a customer, I chose the following predictors:

- Customer type
- Gender
- Product line
- Payment method
- Time of the day
- Rating
- City

Firstly, I have tried to use different models to see which has a better accuracy by cross-validating across models according to their RMSE.

## Evaluating model fitness for predicting total amount spent by a customer



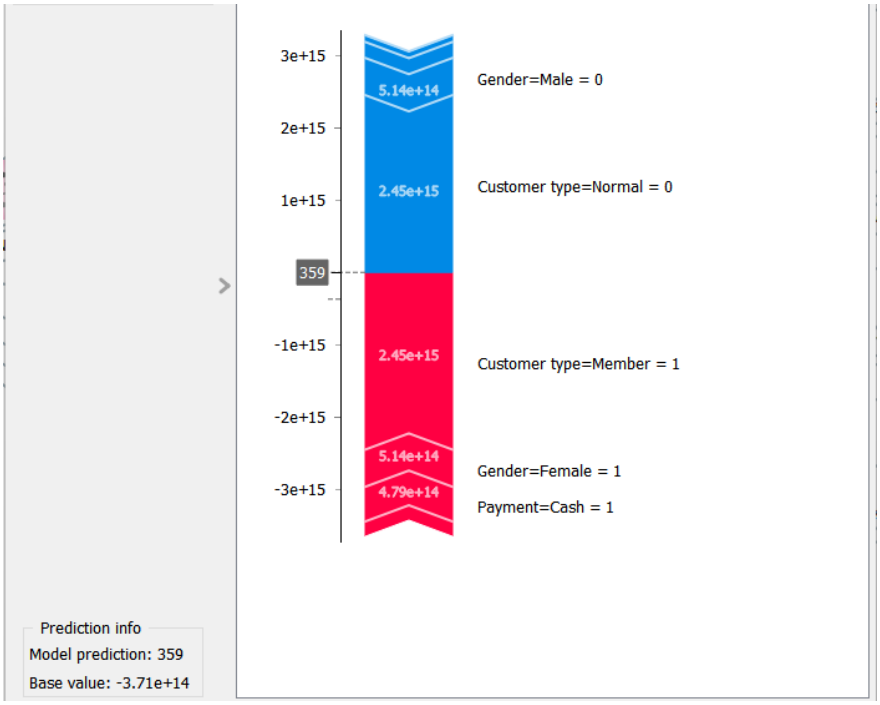

Evaluation Results

| Model | MSE | RMSE | MAE | R2 |
|---|---|---|---|---|
| Stack | 60716.237 | 246.407 | 202.268 | -0.005 |
| Linear Regression | 62572.086 | 250.144 | 206.099 | -0.036 |
| kNN | 67625.183 | 260.048 | 211.124 | -0.120 |
| Random Forest | 68722.739 | 262.150 | 215.856 | -0.138 |
| Neural Network | 74035.930 | 272.095 | 200.717 | -0.226 |
| Tree | 105694.752 | 325.107 | 255.505 | -0.750 |

The linear regression gives the best accuracy here. Although the $R^2$ suggests that this dataset is not fit for predictions, the modelling technique led me to collect an interesting finding.

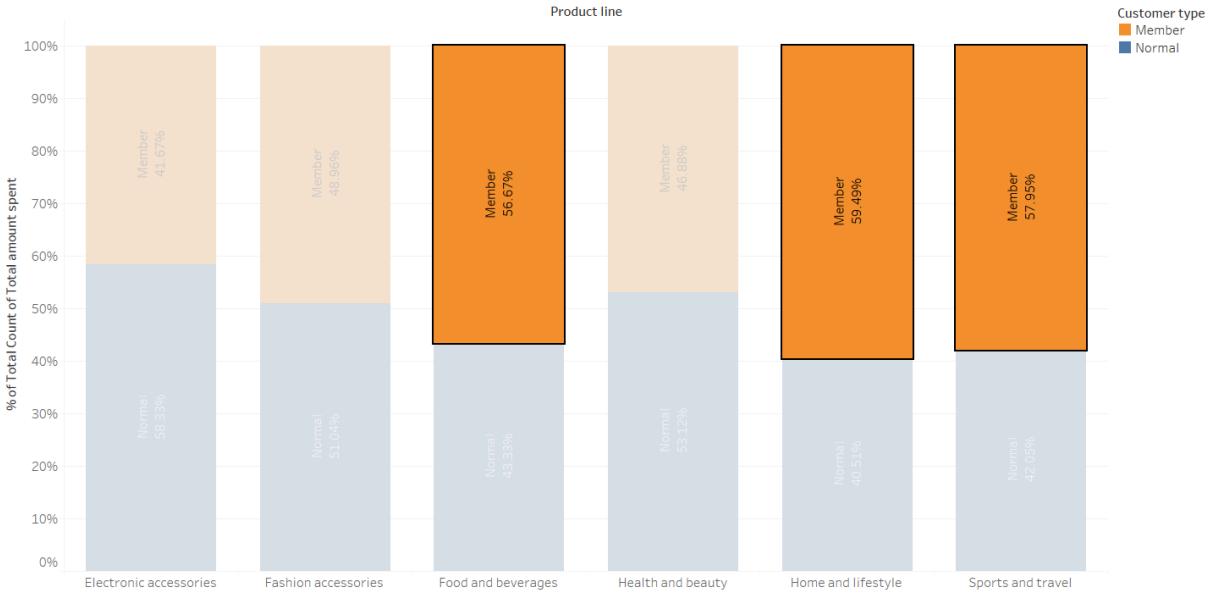
**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

## Membership of customers is the most important factor to determine increased spending, followed by female customers




This esentially means that, members are the more likely to increase average customer spending to 359 or above. Members who are female have the highest propensity to spend. This can be an important insight for the supermarket, since they can incentivize female members more so that they are willing to spend more. Or, they could also incentivize more female customers so that they become members.

## Members who are female dominate the sales of Food & Beverages, Home & Lifestyle, and Sports & Travel product lines




% of Total Count of Total amount spent for each Product line. Color shows details about Customer type. The marks are labeled by Customer type and % of Total Count of Total amount spent. The data is filtered on City, Gender and Payment method. The City filter keeps Mandalay, Naypyidaw and Yangon. The Gender filter keeps Female. The Payment method filter keeps Cash, Credit card and Ewallet.

**Sadman Sakib**
**Roll 05, Section A**
**IBA BBA 27th**

If the supermarket wants to target higher spenders, customers interested in these product lines should be more incentivized.

## Recommendations and the way forward

Considering that the company wanted to analyze the time period of 3-months to identify whether the change in sales between February and March is significant, and considering the quality of the data provided, I would like to make the following recommendations –

- The sales trend is likely an annual occurance and the supermarket should be better prepared for handling the sales during these times every year. Moreover, customers need to be incentivized to prevent low sales in February, preferably by discounts or lower cost goods so that males don't reduce their shopping frequency too much.
- The data is only on 3-months and hence isn't appropriate for predictive modelling. More data across more demographic and behavioral parameters like location of customers, earnings, family size etc. could be better predictors. The shop should collect such data for future analyses.