Final Report

Customer Rating Prediction

Joo Yong Yoon – Group 5

MI 353

Dr. Cevikparmak

16 April 2023

I. Executive summary

The purpose of this project is to predict customer ratings as the growth of supermarkets in most populated cities is increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data.

I found a supermarket sales dataset with 1000 records that came from 3 different branches within 3 months period. Data contains several information such as Branch and City of the supermarket, Customer type, Customer gender, Product line, Unit price, Quantity, Tax 5%, Total price, Date and Time of the sale, Payment type, Cost of goods (cogs), gross margin percentage, gross income, and Customer Rating (variable description is explained on III – 1. Data Description / Variables). My goal was to predict the customer rating using other parameters. It is important to understand the factors impacting customer rating which will help with marketing campaigns. I first looked at the dataset and excluded variables that are directly related to others. For example, I selected City and excluded the branch. I selected total price and excluded tax and cogs. I found that none of the factors in the dataset are statistically significant in predicting the customer rating from the regression models I design. This means there are other unmeasured factors that can impact customer satisfaction. For example, some factors I can think of are wait time in payment queues, availability of the products, approachability, and customer service of sales associates.

II. Project motivation/background

I wanted to find the factors impacting customer rating which can lead to an increase in sales. I work as a quality assurance manager at a manufacturing facility for beauty products. And even though I am not directly related to sales, I think to understand the factors leading to an increase in sale can be a valuable information for the company I work for as our products are distributed to the national retailers such as Walmart and Target.

III. Data description

1. Variables

- Invoice id: Computer generated sales slip invoice identification number
- Branch: Branch of supercenter (3 branches are available identified by A, B and C).
- City: Location of supercenters
- Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.
- Gender: Gender type of customer
- Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
- Unit price: Price of each product in $
- Quantity: Number of products purchased by customer
- Tax: 5% tax fee for customer buying
- Total: Total price including tax
- Date: Date of purchase (Record available from January 2019 to March 2019)
- Time: Purchase time (10am to 9pm)
- Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)
- COGS: Cost of goods sold
- Gross margin percentage: Gross margin percentage
- Gross income: Gross income
- Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

2. Data Exploration

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	7
INPUT	NOMINAL	7
TARGET	INTERVAL	1

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Branch	INPUT	3	0	A	34.00	B	33.20
TRAIN	City	INPUT	3	0	Yangon	34.00	Mandalay	33.20
TRAIN	Customer_type	INPUT	2	0	Member	50.10	Normal	49.90
TRAIN	Gender	INPUT	2	0	Female	50.10	Male	49.90
TRAIN	Payment	INPUT	3	0	Ewallet	34.50	Cash	34.40
TRAIN	Product_line	INPUT	6	0	Fashion accessories	17.80	Food and beverages	17.40

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Quantity	INPUT	5.51	2.923431	1000	0	1	5	10	0.012941	-1.21555
Tax_5_	INPUT	15.37937	11.70883	1000	0	0.5085	12.08	49.65	0.89257	-0.08188
Total	INPUT	322.9667	245.8853	1000	0	10.6785	253.68	1042.65	0.89257	-0.08188
Unit_price	INPUT	55.67213	26.49463	1000	0	10.08	55.07	99.96	0.007077	-1.21859
cogs	INPUT	307.5874	234.1765	1000	0	10.17	241.6	993	0.89257	-0.08188
gross_income	INPUT	15.37937	11.70883	1000	0	0.5085	12.08	49.65	0.89257	-0.08188
Rating	TARGET	6.9727	1.71858	1000	0	4	7	10	0.00901	-1.15159

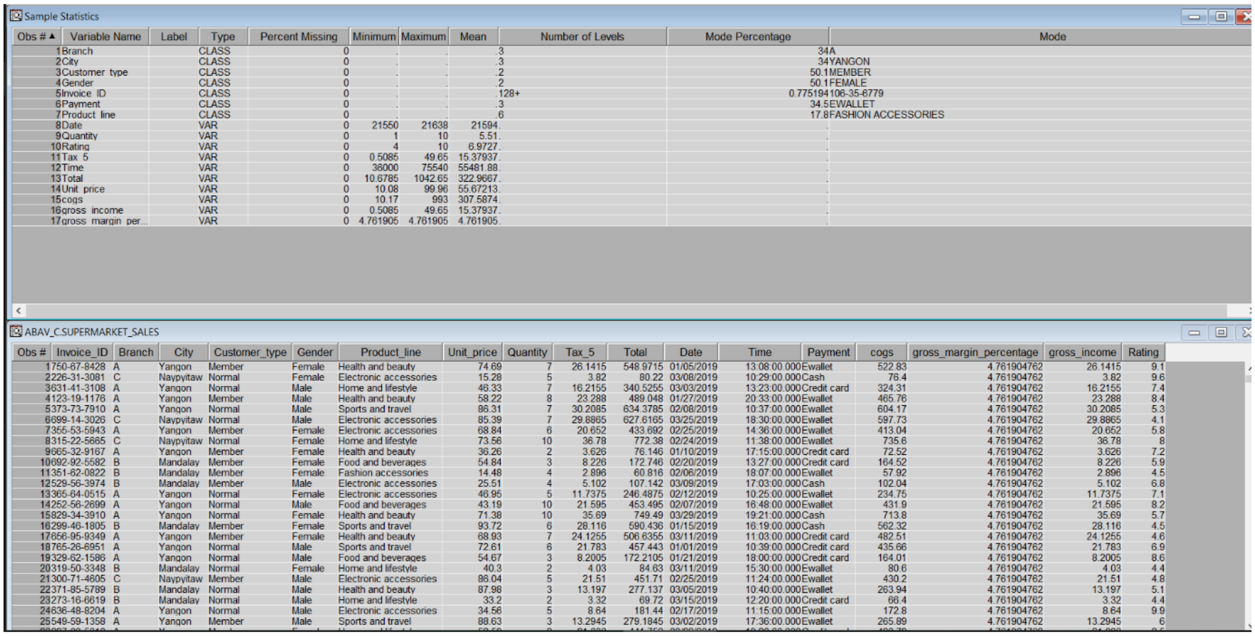
- There are no missing values detected for character and numerical columns.
- However, the skewness for Tax_5%, Total, cogs, and gross income have a slight right-skewed distribution.

Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Rating

Input	Correlation
Unit_price	-0.008778
Quantity	-0.015815
cogs	-0.036442
Total	-0.036442
Tax_5_	-0.036442
gross_income	-0.036442

- All continuous variables in the dataset do not correlate with the rating because the correlation value is close to 0 even though the correlation value is negative. This was a bit concerning and needed to be investigated further.
- As is obvious, quantity and gross income have very high correlation of 70%. Unit price is positively correlated to cogs with 63% correlation.



Sample Statistics

Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number of Levels	Mode Percentage	Mode
1	Branch		CLASS	0	.	.	.	3	34	A
2	City		CLASS	0	.	.	.	3	34	YANGON
3	Customer type		CLASS	0	.	.	.	2	50.1	MEMBER
4	Gender		CLASS	0	.	.	.	2	50.1	FEMALE
5	Invoice ID		CLASS	0	.	.	.	128+	0.775194	106-35-6779
6	Payment		CLASS	0	.	.	.	3	34.5	EWALLET
7	Product line		CLASS	0	.	.	.	6	17.8	FASHION ACCESSORIES
8	Date		VAR	0	21550	21638	21594	.	.	
9	Quantity		VAR	0	1	10	5.51	.	.	
10	Rating		VAR	0	4	10	6.9727	.	.	
11	Tax 5		VAR	0	0.5085	49.65	15.37937	.	.	
12	Time		VAR	0	36000	75540	55481.88	.	.	
13	Total		VAR	0	10.6785	1042.65	322.9667	.	.	
14	Unit price		VAR	0	10.08	99.96	55.67213	.	.	
15	cogs		VAR	0	10.17	993	307.5874	.	.	
16	gross income		VAR	0	0.5085	49.65	15.37937	.	.	
17	gross margin per...		VAR	0	4.761905	4.761905	4.761905	.	.	

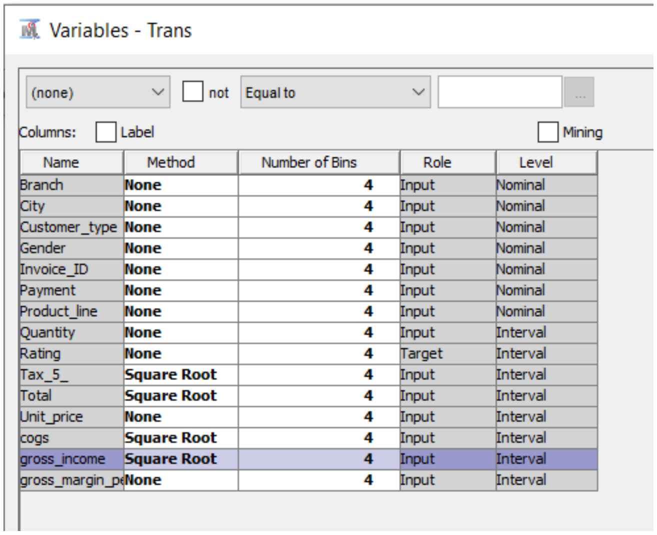
ABAV_C.SUPERMARKET_SALES

Obs #	Invoice_ID	Branch	City	Customer_type	Gender	Product_line	Unit_price	Quantity	Tax_5	Total	Date	Time	Payment	cogs	gross_margin_percentage	gross_income	Rating
1	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	01/05/2019	13:08:00.000	Ewallet	522.83	4.761904762	26.1415	9.1
2	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	03/08/2019	10:29:00.000	Cash	76.4	4.761904762	3.82	9.6
3	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	03/03/2019	13:23:00.000	Credit card	324.31	4.761904762	16.2155	7.4
4	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	01/27/2019	20:33:00.000	Ewallet	465.76	4.761904762	23.288	8.4
5	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	02/08/2019	10:37:00.000	Ewallet	604.17	4.761904762	30.2085	5.3
6	699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	03/25/2019	18:30:00.000	Ewallet	597.73	4.761904762	29.8865	4.1
7	355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.652	433.692	02/25/2019	14:36:00.000	Ewallet	413.04	4.761904762	20.652	5.8
8	315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	02/24/2019	11:38:00.000	Ewallet	735.6	4.761904762	36.78	8
9	665-32-9167	A	Yangon	Member	Female	Health and beauty	36.26	2	3.626	76.146	01/10/2019	17:15:00.000	Credit card	72.52	4.761904762	3.626	7.2
10	692-92-5582	B	Mandalay	Member	Female	Food and beverages	54.84	3	8.226	172.746	02/20/2019	13:27:00.000	Credit card	164.52	4.761904762	8.226	5.9
11	351-62-0822	B	Mandalay	Member	Female	Fashion accessories	14.48	4	2.896	60.816	02/06/2019	18:07:00.000	Ewallet	57.92	4.761904762	2.896	4.5
12	529-56-3974	B	Mandalay	Member	Male	Electronic accessories	25.51	4	5.102	107.142	03/09/2019	17:03:00.000	Cash	102.04	4.761904762	5.102	6.8
13	365-64-0515	A	Yangon	Normal	Female	Electronic accessories	46.95	5	11.7375	246.4875	02/12/2019	10:25:00.000	Ewallet	234.75	4.761904762	11.7375	7.1
14	252-56-2699	A	Yangon	Normal	Male	Food and beverages	43.19	10	21.595	453.495	02/07/2019	16:48:00.000	Ewallet	431.9	4.761904762	21.595	8.2
15	829-34-3910	A	Yangon	Normal	Female	Health and beauty	71.38	10	35.69	749.49	03/29/2019	19:21:00.000	Cash	713.8	4.761904762	35.69	5.7
16	299-46-1805	B	Mandalay	Member	Female	Sports and travel	93.72	6	28.116	590.436	01/15/2019	16:19:00.000	Cash	562.32	4.761904762	28.116	4.5
17	656-95-9349	A	Yangon	Member	Female	Health and beauty	68.93	7	24.1255	506.6355	03/11/2019	11:03:00.000	Credit card	482.51	4.761904762	24.1255	4.6
18	765-26-6951	A	Yangon	Normal	Male	Sports and travel	72.61	6	21.783	457.443	01/01/2019	10:39:00.000	Credit card	435.66	4.761904762	21.783	6.9
19	329-62-1586	A	Yangon	Normal	Male	Food and beverages	54.67	3	8.2005	172.2105	01/21/2019	18:00:00.000	Credit card	164.01	4.761904762	8.2005	8.6
20	319-50-3348	B	Mandalay	Normal	Female	Home and lifestyle	40.3	2	4.03	84.63	03/11/2019	15:30:00.000	Ewallet	80.6	4.761904762	4.03	4.4
21	300-71-4605	C	Naypyitaw	Member	Male	Electronic accessories	86.04	5	21.51	451.71	02/25/2019	11:24:00.000	Ewallet	430.2	4.761904762	21.51	4.8
22	371-85-5789	B	Mandalay	Normal	Male	Health and beauty	87.98	3	13.197	277.137	03/05/2019	10:40:00.000	Ewallet	263.94	4.761904762	13.197	5.1
23	273-16-6619	B	Mandalay	Normal	Male	Home and lifestyle	33.2	2	3.32	69.72	03/15/2019	12:20:00.000	Credit card	66.4	4.761904762	3.32	4.4
24	636-48-8204	A	Yangon	Normal	Male	Electronic accessories	34.56	5	8.64	181.44	02/17/2019	11:15:00.000	Ewallet	172.8	4.761904762	8.64	9.9
25	549-59-1358	A	Yangon	Member	Male	Sports and travel	88.63	3	13.2945	279.1845	03/02/2019	17:36:00.000	Ewallet	265.89	4.761904762	13.2945	6

- It can be seen that the variables do not have missing values in the dataset.
- The rating distribution looks uniform and there seems to be no skewness on the left or right side of the distribution.
- There is not much difference in sales across the 3 branches.
- Dataset contained similar number of customers coming from each city/branch 33.2% from Mandalay, 32.8% from Naypyitaw and 34.0% from Yangon.
- There were similar number of males and female customers in the dataset.
- Average of total sales was 322.97 (standard deviation=245.89) and median 253.8. In general these supermarkets have received higher customer rating. Average customer rating was 7 out of 10 (standard deviation=1.7). Median rating was 7 (min=4, max=10).

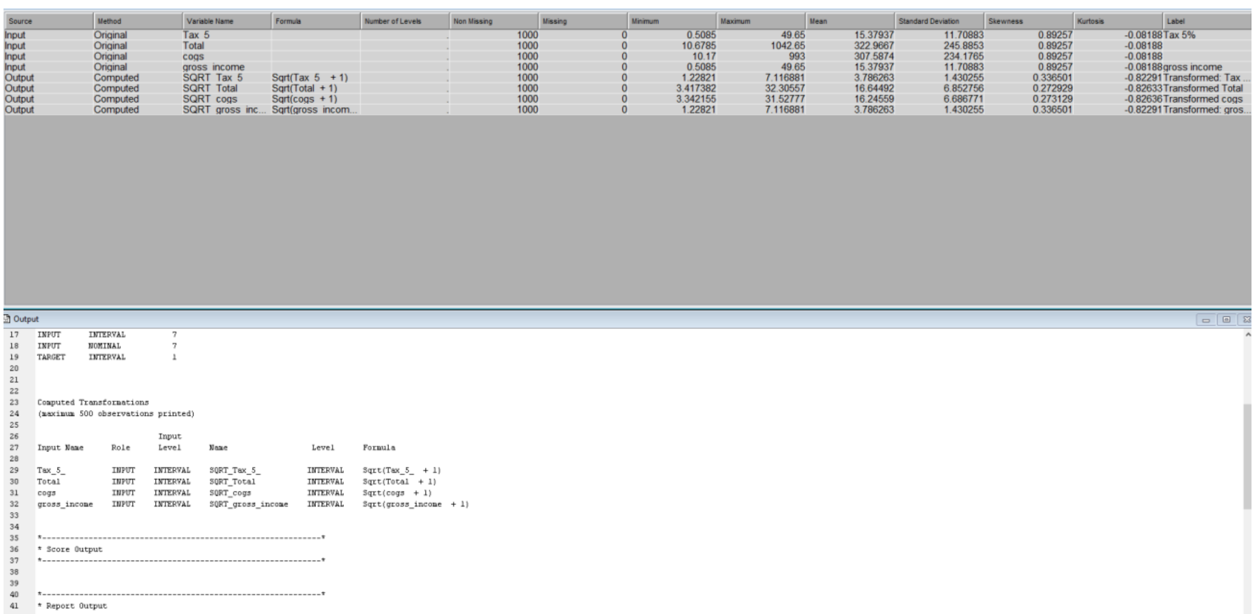
IV. Data preparation activities

1. Data Transformation



Name	Method	Number of Bins	Role	Level
Branch	None	4	Input	Nominal
City	None	4	Input	Nominal
Customer_type	None	4	Input	Nominal
Gender	None	4	Input	Nominal
Invoice_ID	None	4	Input	Nominal
Payment	None	4	Input	Nominal
Product_line	None	4	Input	Nominal
Quantity	None	4	Input	Interval
Rating	None	4	Target	Interval
Tax_5_	Square Root	4	Input	Interval
Total	Square Root	4	Input	Interval
Unit_price	None	4	Input	Interval
cogs	Square Root	4	Input	Interval
gross_income	Square Root	4	Input	Interval
gross_margin_pe	None	4	Input	Interval

- Configuration of dtata transformation node.



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	Tax 5		.	1000	0	0.5085	49.65	15.37937	11.70883	0.89257	-0.08188	Tax 5%
Input	Original	Total		.	1000	0	10.6785	1042.65	322.9667	245.8853	0.89257	-0.08188	
Input	Original	cogs		.	1000	0	10.17	993	307.5874	234.1765	0.89257	-0.08188	
Input	Original	gross income		.	1000	0	0.5085	49.65	15.37937	11.70883	0.89257	-0.08188	gross income
Output	Computed	SQRT Tax 5	Sqrt(Tax 5 + 1)	.	1000	0	1.22821	7.116881	3.786263	1.430255	0.336501	-0.82291	Transformed: Tax ...
Output	Computed	SQRT Total	Sqrt(Total + 1)	.	1000	0	3.417382	32.30557	16.64492	6.852756	0.272929	-0.82633	Transformed Total
Output	Computed	SQRT cogs	Sqrt(cogs + 1)	.	1000	0	3.342155	31.52777	16.24559	6.686771	0.273129	-0.82636	Transformed cogs
Output	Computed	SQRT gross inc...	Sqrt(gross incom...	.	1000	0	1.22821	7.116881	3.786263	1.430255	0.336501	-0.82291	Transformed: gros...

Output

```
INPUT      INTERVAL        7
INPUT      NOMINAL         7
TARGET     INTERVAL        1



Computed Transformations
(maximum 500 observations printed)

                          Input
Input Name      Role      Level       Name                 Level       Formula

Tax_5_          INPUT     INTERVAL    SQRT_Tax_5_          INTERVAL    Sqrt(Tax_5_  + 1)
Total           INPUT     INTERVAL    SQRT_Total           INTERVAL    Sqrt(Total  + 1)
cogs            INPUT     INTERVAL    SQRT_cogs            INTERVAL    Sqrt(cogs  + 1)
gross_income    INPUT     INTERVAL    SQRT_gross_income    INTERVAL    Sqrt(gross_income  + 1)


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

- Data transformation output.
- The skewness value of the new variables has a skewness value of < 0.5, means that the distribution of these variables is normal.

2. Dropping Variables

Variables - Drop

(none) | not | Equal to

Columns: Label | Mining

Name	Drop	Role	Level
Branch	No	Input	Nominal
City	No	Input	Nominal
Customer_type	No	Input	Nominal
Date	Yes	Time ID	Interval
Gender	No	Input	Nominal
gross_margin_percentage	Yes	Input	Interval
Invoice_ID	Yes	Input	Nominal
Payment	No	Input	Nominal
Product_line	No	Input	Nominal
Quantity	No	Input	Interval
Rating	No	Target	Interval
SQRT_cogs	No	Input	Interval
SQRT_gross_income	No	Input	Interval
SQRT_Tax_5_	No	Input	Interval
SQRT_Total	No	Input	Interval
Time	Yes	Time ID	Interval
Unit_price	No	Input	Interval

- Dropping some variables in the dataset. I considered Date, Gross Margin Percentage, Invoice ID, and Time are irrelevant.

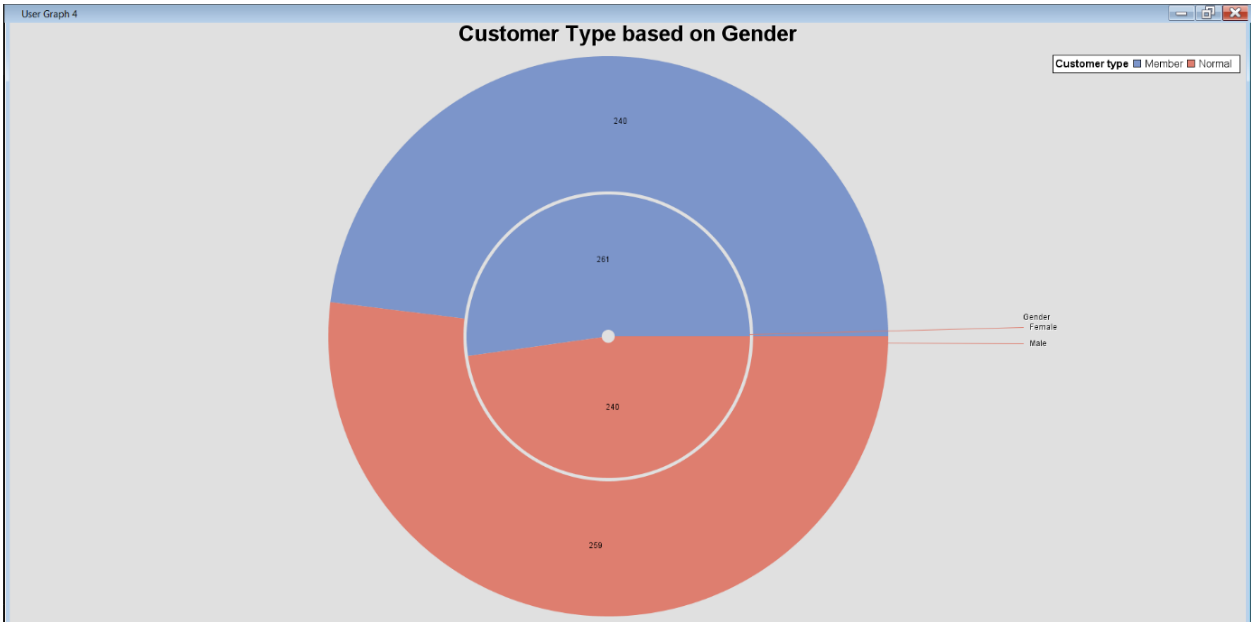
V. EDA

I understand that EDA does not determine models on the data, but I believe it helps business owners to explore possible data analysis models that best suit the data based on the business problems and goals.

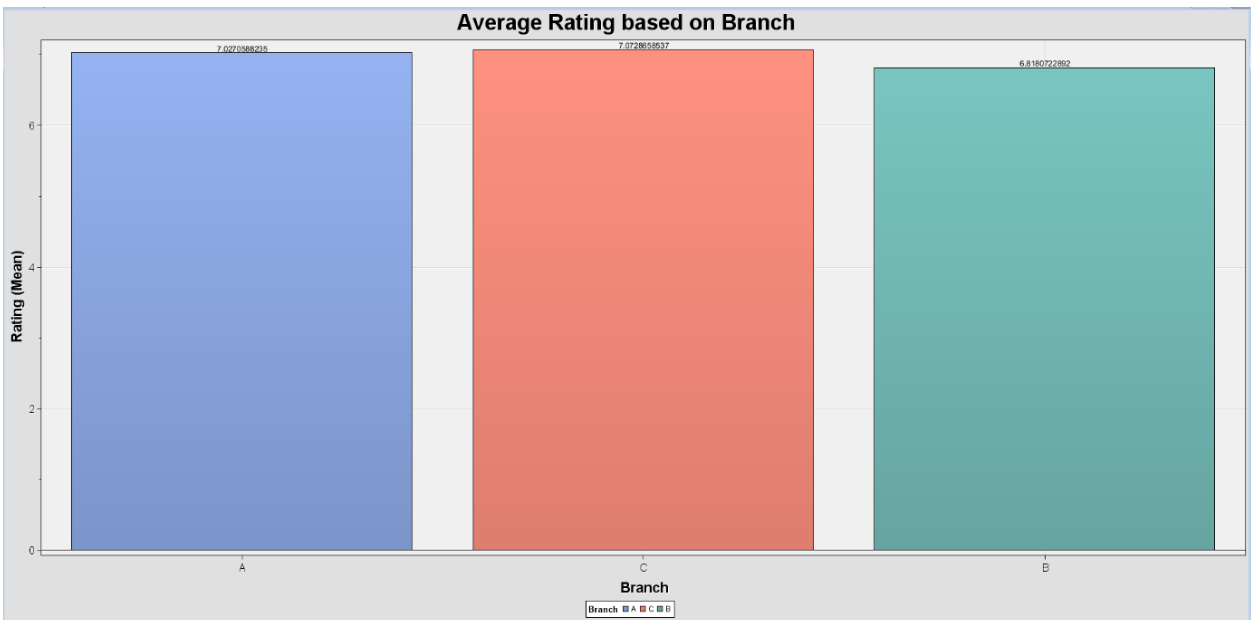
1. Payment Type based on Branch

- The most popular payment method is in fact E-wallet and surprisingly not credit cards. Cash payment is also popular.

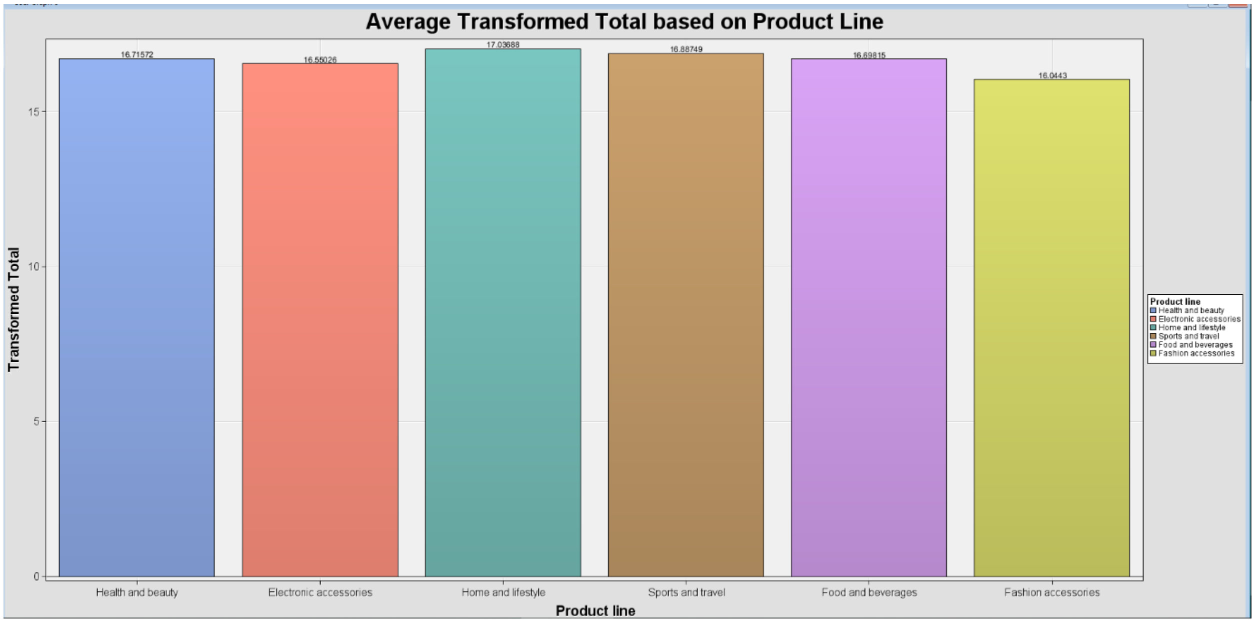
2. Customer Type based on Gender



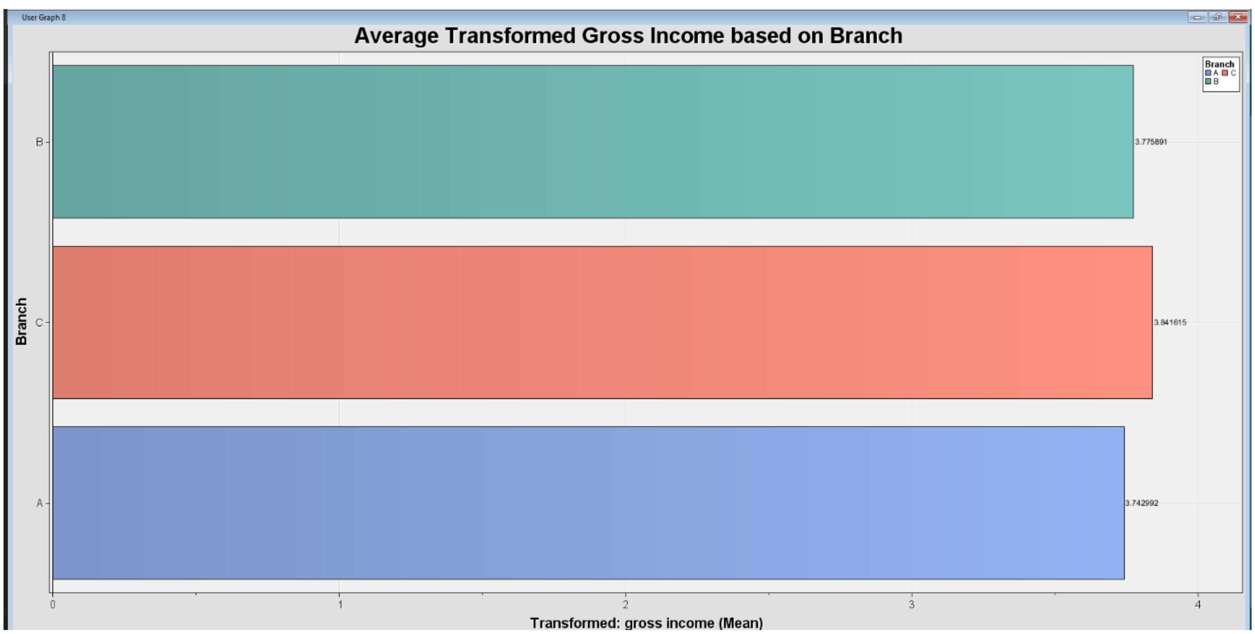
3. Average Rating based on Branch



4. Average Transformed Total based on Product Line



5. Average Transformed Gross Income based on Branch



Findings from EDA are:

- There is not much difference in gross income by branch at an average level.
- Gross income is similar for both male and female, though female customers spend a bit higher at the 75th percentile.
- No particular time trend is observed.
- At an overall level, ‘Sports and Travel’ generates the highest gross income.

VI. Model(s)/Enterprise Miner diagrams used

1. Data Preparation

.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	15.0
Test	15.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	8/5/21 3:20 PM
Run ID	5e47529d-1fd1-2e4a-b58e-86cf07

- The dataset split into 70% train, 15% validation, and 15% test.

Summary Statistics for Interval Targets

Data=DATA

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	10	6.9727	4	1000	0	1.7185802944	

Data=TEST

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	10	6.976	4	150	0	1.7285475146	

Data=TRAIN

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	10	6.97	4	700	0	1.7146386482	

Data=VALIDATE

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	10	6.982	4.1	150	0	1.7384352294	

2. Linear Regression

2-1. Standard Linear Regression

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	785.15	.	.
ASE	Average Squared Error	2.88	2.929	3.050
AVERR	Average Error Function	2.88	2.929	3.050
DFE	Degrees of Freedom for Error	678.00	.	.
DFM	Model Degrees of Freedom	22.00	.	.
DFT	Total Degrees of Freedom	700.00	.	.
DIV	Divisor for ASE	700.00	150.000	150.000
ERR	Error Function	2018.01	439.414	457.438
FPE	Final Prediction Error	3.07	.	.
MAX	Maximum Absolute Error	3.42	3.160	3.292
MSE	Mean Square Error	2.98	2.929	3.050
NOBS	Sum of Frequencies	700.00	150.000	150.000
_NW_	Number of Estimate Weights	22.00	.	.
RASE	Root Average Sum of Squares	1.70	1.712	1.746
RFPE	Root Final Prediction Error	1.75	.	.
RMSE	Root Mean Squared Error	1.73	1.712	1.746
SBC	Schwarz's Bayesian Criterion	885.28	.	.
SSE	Sum of Squared Errors	2018.01	439.414	457.438
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	37.037056	1.763669	0.59	0.9250
Error	678	2018.012944	2.976420		
Corrected Total	699	2055.050000			

Model Fit Statistics

R-Square	0.0180	Adj R-Sq	-0.0124
AIC	785.1518	BIC	788.5774
SBC	885.2756	C(p)	22.0000

- RMSE value for training is 1.73, for validation is 1.712, whereas the RMSE value for testing is 1.746.
- The p-value from ANOVA table is 0.9250 (> 0.05), which can be concluded that the independent variables do not exhibit a statistically significant connection with the dependent variable, or the independent variables do not predict the dependent variable dependably.
- The value of R-square is 0.0180, which means this result shows that the independent variable can be used to predict just 1.8% of the variation in ratings (dependent variable).

2-2. Forward Linear Regression

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	755.88	.	.
ASE	Average Squared Error	2.94	3.002	2.968
AVERR	Average Error Function	2.94	3.002	2.968
DFE	Degrees of Freedom for Error	699.00	.	.
DFM	Model Degrees of Freedom	1.00	.	.
DFT	Total Degrees of Freedom	700.00	.	.
DIV	Divisor for ASE	700.00	150.000	150.000
ERR	Error Function	2055.05	450.323	445.199
FPE	Final Prediction Error	2.94	.	.
MAX	Maximum Absolute Error	3.03	3.030	3.030
MSE	Mean Square Error	2.94	3.002	2.968
NOBS	Sum of Frequencies	700.00	150.000	150.000
_NW_	Number of Estimate Weights	1.00	.	.
RASE	Root Average Sum of Squares	1.71	1.733	1.723
RFPE	Root Final Prediction Error	1.72	.	.
RMSE	Root Mean Squared Error	1.71	1.733	1.723
SBC	Schwarz's Bayesian Criterion	760.43	.	.
SSE	Sum of Squared Errors	2055.05	450.323	445.199
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	699	2055.050000	2.939986		
Corrected Total	699	2055.050000			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	755.8826	BIC	757.9099
SBC	760.4337	C(p)	-7.5565

- The RMSE value for training is 1.71, for validation is 1.733, whereas the RMSE value for testing is 1.723.
- The p-value and r-square obtained error or did not appear.

2-3. Stepwise Linear Regression

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	755.88	.	.
ASE	Average Squared Error	2.94	3.002	2.968
AVERR	Average Error Function	2.94	3.002	2.968
DFE	Degrees of Freedom for Error	699.00	.	.
DFM	Model Degrees of Freedom	1.00	.	.
DFT	Total Degrees of Freedom	700.00	.	.
DIV	Divisor for ASE	700.00	150.000	150.000
ERR	Error Function	2055.05	450.323	445.199
FPE	Final Prediction Error	2.94	.	.
MAX	Maximum Absolute Error	3.03	3.030	3.030
MSE	Mean Square Error	2.94	3.002	2.968
NOBS	Sum of Frequencies	700.00	150.000	150.000
_NW_	Number of Estimate Weights	1.00	.	.
RASE	Root Average Sum of Squares	1.71	1.733	1.723
RFPE	Root Final Prediction Error	1.72	.	.
RMSE	Root Mean Squared Error	1.71	1.733	1.723
SBC	Schwarz's Bayesian Criterion	760.43	.	.
SSE	Sum of Squared Errors	2055.05	450.323	445.199
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	699	2055.050000	2.939986		
Corrected Total	699	2055.050000			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	755.8826	BIC	757.9099
SBC	760.4337	C(p)	-7.5565

- RMSE value for training is 1.71, for validation is 1.733, whereas the RMSE value for testing is 1.723.
- The p-value and r-square obtained error or did not appear.

2-4. Backward Linear Regression

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	F Value	Pr > F
1	Gender	1	14	0.00	0.9479
2	Quantity*Quantity	1	13	0.01	0.9055
3	Product_line	5	12	0.42	0.8325
4	SQRT_Tax_5_	1	11	0.12	0.7281
5	SQRT_Total	1	10	0.01	0.9184
6	Payment	2	9	0.46	0.6301
7	Quantity*SQRT_Total	1	8	0.81	0.3679
8	Unit_price	1	7	0.70	0.4017
9	Branch	2	6	1.19	0.3039
10	Quantity	1	5	1.32	0.2507
11	Quantity*SQRT_Tax_5_	1	4	0.27	0.6009
12	Quantity*Unit_price	1	3	0.54	0.4615
13	SQRT_Tax_5_*Unit_price	1	2	0.65	0.4187
14	Unit_price*Unit_price	1	1	0.83	0.3615
15	Customer_type	1	0	1.85	0.1737

The selected model is the model trained in the last step (Step 15). It consists of the following effects:

Intercept

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	699	2055.050000	2.939986		
Corrected Total	699	2055.050000			

The DMREG Procedure

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	755.8826	BIC	757.9099
SBC	760.4337	C(p)	-7.5565

- RMSE value for training is 1.71, for validation is 1.733, whereas the RMSE value for testing is 1.723.
- There are 15 steps performed in this model, and based on the resulting p-value, there are no independent variables that affect the rating value (dependent variable) since the p-value is more than 0.05.
- The p-value and r-square obtained error or did not appear.

2-5. Linear Regression Summary

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Train: Root Mean Squared Error	Valid: Root Mean Square Error	Test: Root Mean Square Error
Y	Reg	Reg	None Regression	Rating		1.72523	1.711556	1.746307
	Reg2	Reg2	Forward Regression	Rating		1.714639	1.732672	1.722787
	Reg3	Reg3	Backward Regression	Rating		1.714639	1.732672	1.722787
	Reg5	Reg5	Stepwise Regression	Rating		1.714639	1.732672	1.722787

2-6. Improvements

- There is no relationship between gross income and customer ratings.
- Using the correlation analysis, one interesting observation has emerged that customer ratings is not related to any variable.
- Due to the findings I mentioned above, the linear regression models are not unfortunately statistically significant. This was unexpected but could often happen with not “textbook” dataset.
- RMSE measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).
- I concluded that this happens when two variables providing same/very similar information, I need to select one variable. For example I selected city and drop branch from the analysis. I selected the total sales and dropped tax, cogs. I also dropped the variable gross margin percentage since it has the same value.

2-7. Final model & Interpretation

```
                                 Analysis of Variance

                                       Sum of
Source                   DF           Squares     Mean Square
  F Value     Pr > F

Model                    13         26.171715        2.013209
      0.68     0.7854
Error                   986       2924.392995        2.965916
Corrected Total         999       2950.564710


                  Model Fit Statistics

R-Square          0.0089      Adj R-Sq       -0.0042
AIC            1101.0869      BIC          1103.4841
SBC            1169.7955      C(p)           14.0000

```

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
City	2	12.1275	2.04	0.1300
Customer_type	1	1.2056	0.41	0.5239
Gender	1	0.0430	0.01	0.9041
Payment	2	0.6248	0.11	0.9000
Product_line	5	6.8063	0.46	0.8069
Total	1	4.3836	1.48	0.2244
Unit_price	1	0.9026	0.30	0.5813

New variable selections as follows:

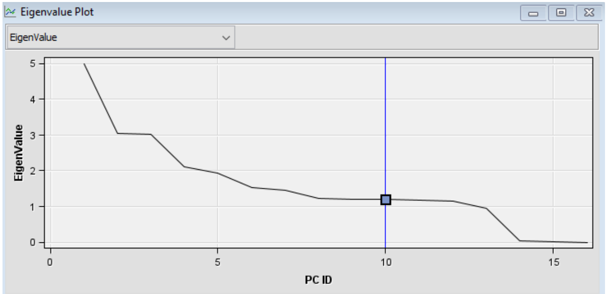
- Target/Outcome: Rating
- Predictors: City, Customer type, Gender, Product line, Unit price, Quantity, Total, Payment

Variable selection criteria didn't select any variables. Therefore, I just explored a model without any selection using the variables we thought as important. This is the final model I obtained. R-square of the model is only 0.0089 which indicates a poor fit. Only less than 1% of the variability of customer satisfaction is explained by the variables in the model. This implies that there should be other variables that impact customer satisfaction than the variables collected in the dataset. Predictors in the final model are interpreted as follows. Compared to Yangon, Mandalay city has 0.155 lower customer satisfaction when adjusted for all the other factors in the model. Even though it is not significant, when total price increase by 100 units, customer satisfaction decreases by 0.04 units adjusting for other covariates in the model.

Variable	Estimate	P-value
City		
Mandalay vs Yangon	-0.155	0.045
Naypyitaw vs Yangon	0.096	0.217
Customer_type		
Member vs Normal	-0.035	0.524
Gender		
Female vs Male	-0.007	0.904
Payment		
Cash vs Ewallet	-0.006	0.935
Credit card vs Ewallet	0.034	0.668
Product_line		
Electronic accessories vs Sports and travel	-0.048	0.692
Fashion accessories vs Sports and travel	0.052	0.666
Food and beverages vs Sports and travel	0.131	0.275
Health and beauty vs Sports and travel	0.034	0.791
Home and lifestyle vs Sports and travel	-0.127	0.307
Total	-0.0004	0.224
Unit_price	0.001	0.581

2-8. PCA

Minimizing the variable selection has proven not very successful. This might show that the linear regression model are not suitable for this dataset. Before I found the other options, I decided to try PCA (85%) and regression model accordingly.



For the original PCA model, there are 15 principal components that correspond to accumulative value of 1.0. With 85% similarity, there are 10 principal components selected. I think 7 principal components can be a reasonable number. It honestly seems that deciding the number of principal components can be subjective to the goal of the model. Lowering the number of principal components by maintaining low similarity can be a aggressive reduction, but it could also be useful if the goal if to reduce the number of principal components. On the other hand, maintaining the similarity, cumulative variance, in the data set could be another goal. I believe reducing the number of variables should be the priority for this dataset due to the results of Model 1. Additionally, principal components are not co-related to each other. Regression can be very sensitive to correlation, and reducing correlation among variables could also be effective.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	20.764198	2.076420	0.70	0.7242
Error	989	2929.800512	2.962387		
Corrected Total	999	2950.564710			

Model Fit Statistics

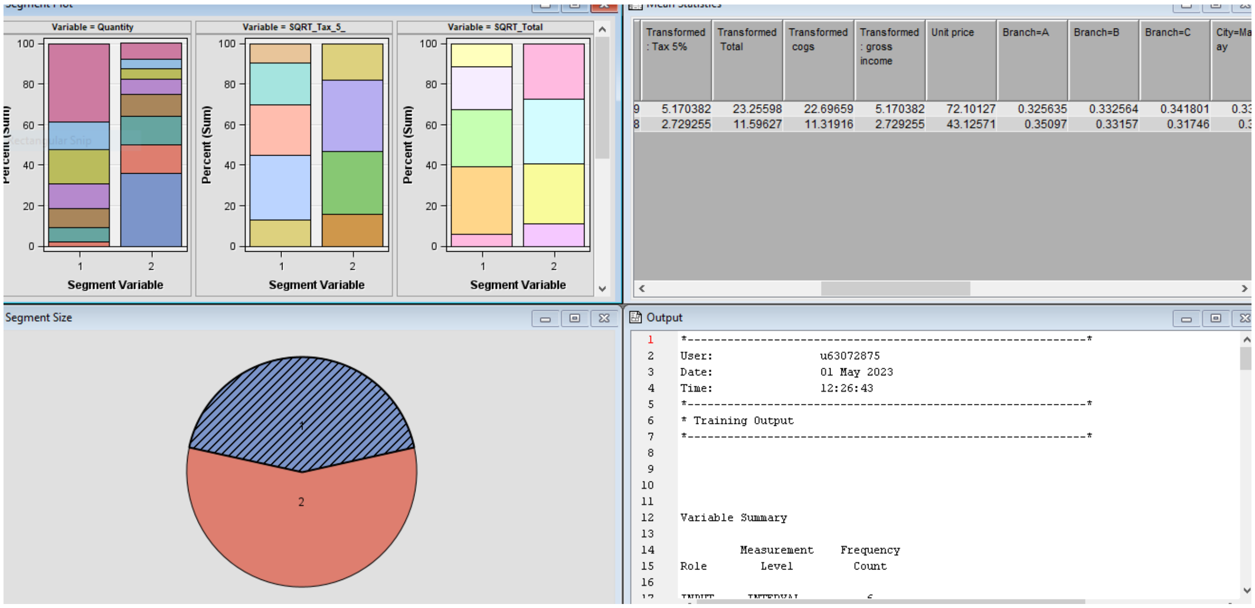
R-Square	0.0070	Adj R-Sq	-0.0030
AIC	1096.9343	BIC	1099.1788
SBC	1150.9196	C(p)	11.0000

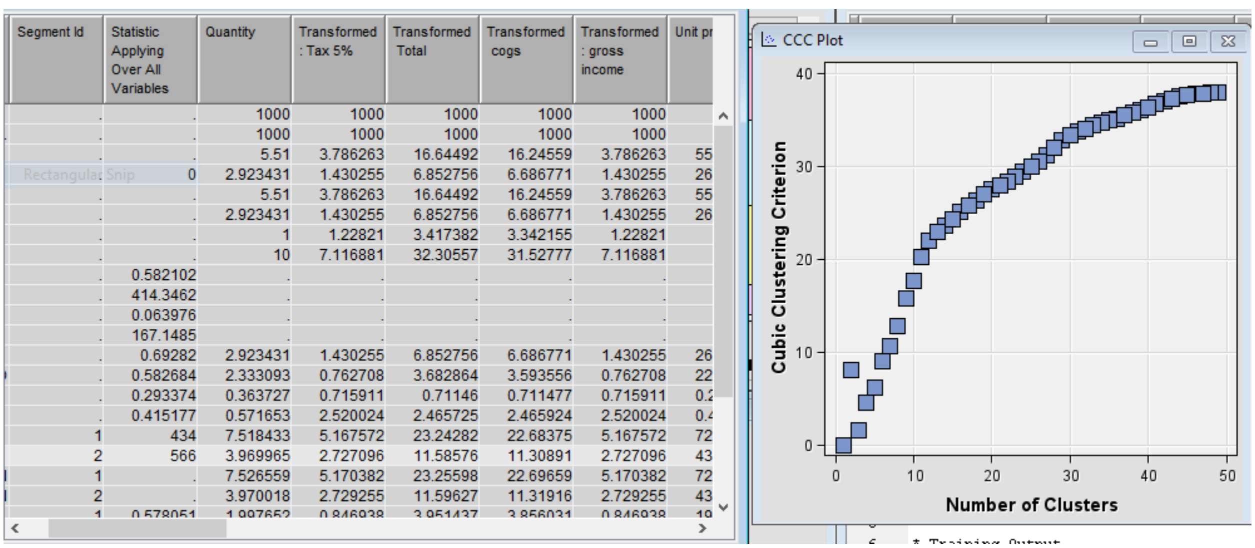
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.9727	0.0544	128.11	<.0001
PC_1	1	-0.0182	0.0244	-0.74	0.4569
PC_10	1	0.0237	0.0499	0.48	0.6346
PC_2	1	0.000235	0.0312	0.01	0.9940
PC_3	1	-0.0644	0.0314	-2.05	0.0406
PC_4	1	0.0190	0.0375	0.51	0.6130
PC_5	1	-0.00593	0.0392	-0.15	0.8795
PC_6	1	0.00106	0.0439	0.02	0.9808
PC_7	1	0.0414	0.0451	0.92	0.3590
PC_8	1	0.0273	0.0492	0.56	0.5790
PC_9	1	-0.0385	0.0499	-0.77	0.4401

Unfortunately, the result of regression model using PCA was not statistically significant, similar to the ones of other regression models.

3. Clustering





Findings/Interpretation from clustering:

- Fashion accessories and food and beverages are the most sold product in Naypyitaw and these products should be focused on along with electronic accessories.
- Most of the customers buy 10 quantities and busiest time of the day is afternoon i.e. around 2 pm which records highest sales. Sales is higher on Tuesdays and Saturdays compared to the rest of the week.
- Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but the quantity purchased is low. Hence, supply for these products need to be increased.

- No matter if it is the weekdays or the weekends, majority of the customers will only spend around $0 to $200. On the weekends however, more customers spend $200 to $400 and $400 to $600 compared to when on the weekdays.
- The data shows that there is a trend on the number of customers at certain days and time, which provides information for business decisions such as targeted time to offer discounts.
- There is weak causal relationship observed between average sales on weekdays vs weekends.

4. Neural Network

Neural networks are a class of parametric models that can accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable than can logistic regression.

3-1	Generalized Linear Model	Number of hidden units: - Max Iterations: 100
3-2	Multilayer Perceptron 1	Number of hidden units: 2 Max Iterations: 100
3-3	Multilayer Perceptron 2	Number of hidden units: 4 Max Iterations: 200
3-4	Multilayer Perceptron 3	Number of hidden units: 8 Max Iterations: 300

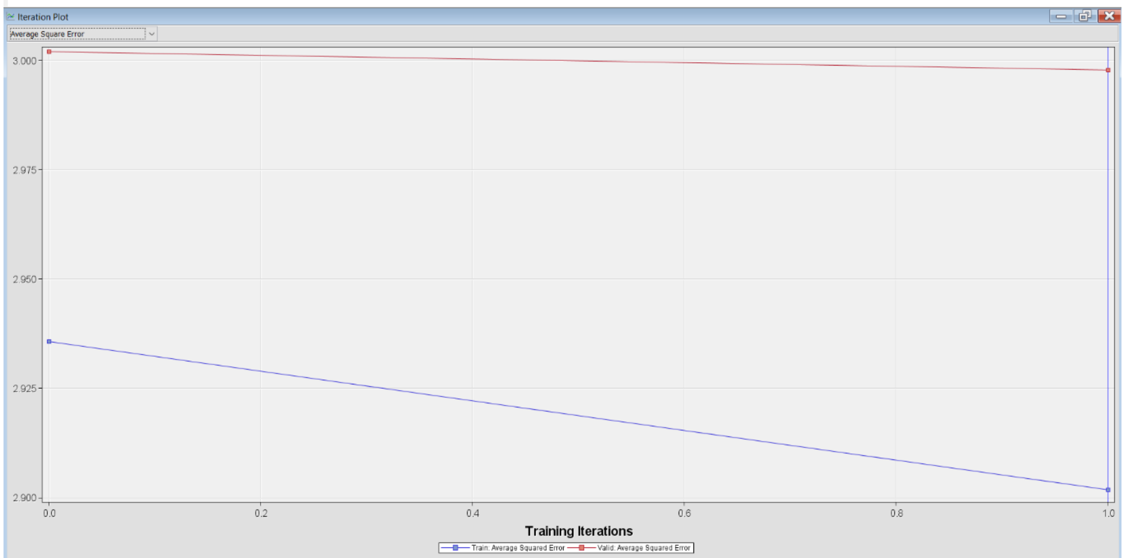
A hidden layer in an artificial neural network is a layer in between input layers and output layers, where artificial neurons take in a set of weighted inputs and produce an output through an activation function. I stopped at maximizing the number of hidden units at 8 because if I have too many hidden units, I may get low training error but still have high generalization error due to overfitting and high variance. In addition, more iterations are almost always better, but each additional iteration provides a smaller gain. If I add more iterations, I will get better matches, but the increase in overlapping index becomes smaller and smaller. At some point, I should expect adding iterations to make no practical change in the quality of the matching. This makes sense - as the model add samples from population, the chance that any observation will be the maximum decreases over time.

3-1 Generalized Linear Model

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
DFT	Total Degrees of Freedom	700.00	.	.
DFE	Degrees of Freedom for Error	679.00	.	.
DFM	Model Degrees of Freedom	21.00	.	.
_NW_	Number of Estimated Weights	21.00	.	.
AIC	Akaike's Information Criterion	787.73	.	.
SBC	Schwarz's Bayesian Criterion	883.31	.	.
ASE	Average Squared Error	2.90	2.998	3.037
MAX	Maximum Absolute Error	3.37	3.191	3.272
DIV	Divisor for ASE	700.00	150.000	150.000
NOBS	Sum of Frequencies	700.00	150.000	150.000
RASE	Root Average Squared Error	1.70	1.731	1.743
SSE	Sum of Squared Errors	2031.27	449.679	455.594
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000
FPE	Final Prediction Error	3.08	.	.
MSE	Mean Squared Error	2.99	2.998	3.037
RFPE	Root Final Prediction Error	1.76	.	.
RMSE	Root Mean Squared Error	1.73	1.731	1.743
AVERR	Average Error Function	2.90	2.998	3.037
ERR	Error Function	2031.27	449.679	455.594
MISC	Misclassification Rate	.	.	.
WRONG	Number of Wrong Classifications	.	.	.



Optimization Results

Iterations	1	Function Calls	4
Jacobian Calls	3	Active Constraints	0
Objective Function	2.9018076035	Max Abs Gradient Element	5.9957799E-6
Lambda	1.729127E-9	Actual Over Pred Change	1
Radius	382.8211993		

Convergence criterion (ABSGCONV=0.00001) satisfied.

- RMSE value for training is 1.73, for validation is 1.731, whereas the RMSE value for testing is 1.743.

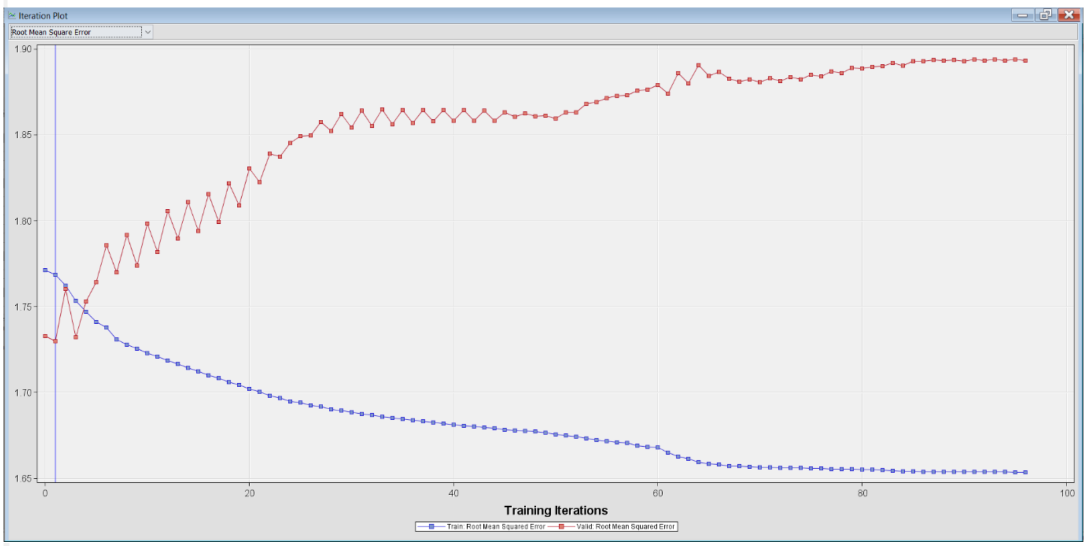
- There is only one iteration with the initial RMSE train 1.739708 and ending at 1.729611
- The initial RMSE validation is 1.732672 and ends at 1.731434.

3-2 Multilayer Perceptron 1 (2 Hidden Units, Iterations 100)

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
DFT	Total Degrees of Freedom	700.00	.	.
DFE	Degrees of Freedom for Error	655.00	.	.
DFM	Model Degrees of Freedom	45.00	.	.
_NW_	Number of Estimated Weights	45.00	.	.
AIC	Akaike's Information Criterion	841.59	.	.
SBC	Schwarz's Bayesian Criterion	1046.39	.	.
ASE	Average Squared Error	2.93	2.993	2.960
MAX	Maximum Absolute Error	3.13	3.021	3.150
DIV	Divisor for ASE	700.00	150.000	150.000
NOBS	Sum of Frequencies	700.00	150.000	150.000
RASE	Root Average Squared Error	1.71	1.730	1.721
SSE	Sum of Squared Errors	2048.33	448.944	444.039
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000
FPE	Final Prediction Error	3.33	.	.
MSE	Mean Squared Error	3.13	2.993	2.960
RFPE	Root Final Prediction Error	1.82	.	.
RMSE	Root Mean Squared Error	1.77	1.730	1.721
AVERR	Average Error Function	2.93	2.993	2.960
ERR	Error Function	2048.33	448.944	444.039
MISC	Misclassification Rate	.	.	.
WRONG	Number of Wrong Classifications	.	.	.



Optimization Results

Iterations	96	Function Calls	110
Jacobian Calls	99	Active Constraints	0
Objective Function	2.5586420217	Max Abs Gradient Element	0.0053277272
Lambda	0.0769617646	Actual Over Pred Change	0.1476621464
Radius	0.0532422292		

Convergence criterion (FCONV=0.0001) satisfied.

- RMSE value for training is 1.77, for validation is 1.730, whereas the RMSE value for testing is 1.721.
- There were 96 iterations in this model with the initial RMSE train of 1.771294 and continued to decrease to 1.65361

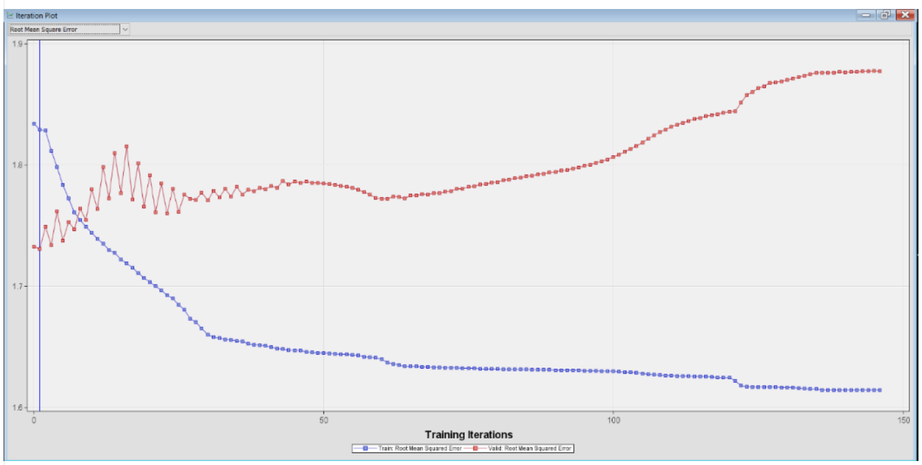
- The initial RMSE validation value is 1.732672 and continues to increase to 1.893184 (overfitting).
- The best iteration is the first iteration with RMSE training, which is 1.768396, and RMSE validation is 1.730018.

3-3. Multilayer Perceptron 3 (8 Hidden Units, Iterations 300)

Fit Statistics

Target=Rating Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
DFT	Total Degrees of Freedom	700.00	.	.
DFE	Degrees of Freedom for Error	611.00	.	.
DFM	Model Degrees of Freedom	89.00	.	.
_NW_	Number of Estimated Weights	89.00	.	.
AIC	Akaike's Information Criterion	928.13	.	.
SBC	Schwarz's Bayesian Criterion	1333.18	.	.
ASE	Average Squared Error	2.92	2.995	2.967
MAX	Maximum Absolute Error	3.16	3.026	3.145
DIV	Divisor for ASE	700.00	150.000	150.000
NOBS	Sum of Frequencies	700.00	150.000	150.000
RASE	Root Average Squared Error	1.71	1.731	1.723
SSE	Sum of Squared Errors	2044.07	449.245	445.059
SUMW	Sum of Case Weights Times Freq	700.00	150.000	150.000
FPE	Final Prediction Error	3.77	.	.
MSE	Mean Squared Error	3.35	2.995	2.967
RFPE	Root Final Prediction Error	1.94	.	.
RMSE	Root Mean Squared Error	1.83	1.731	1.723
AVERR	Average Error Function	2.92	2.995	2.967
ERR	Error Function	2044.07	449.245	445.059
MISC	Misclassification Rate	.	.	.
WRONG	Number of Wrong Classifications	.	.	.



Optimization Results

Iterations	146	Function Calls	166
Jacobian Calls	150	Active Constraints	0
Objective Function	2.2747722168	Max Abs Gradient Element	0.004794685
Lambda	0.0696572397	Actual Over Pred Change	0.1587145378
Radius	0.0513073796		

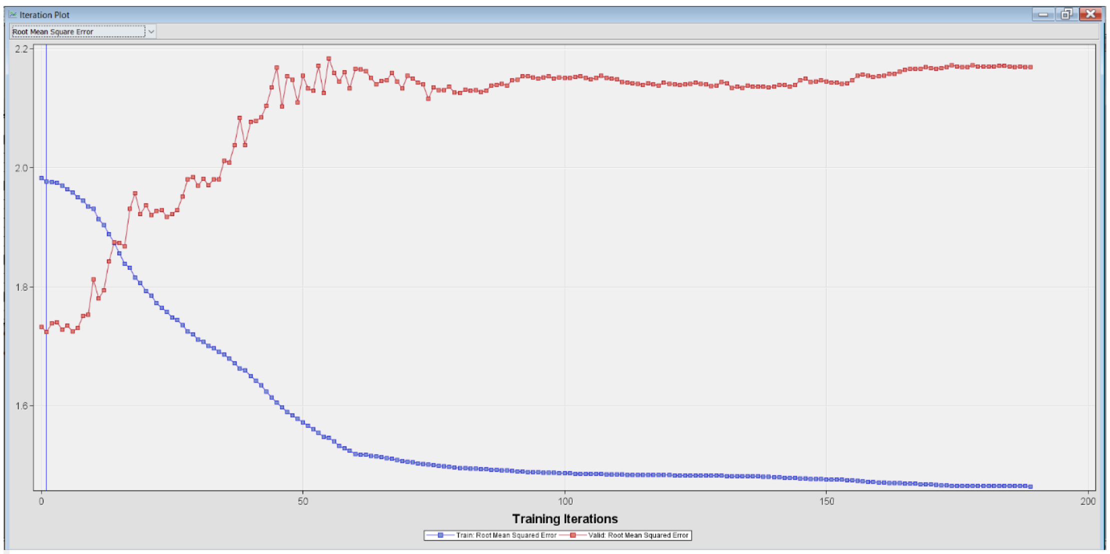
Convergence criterion (FCONV=0.0001) satisfied.

NOTE: At least one element of the gradient is greater than 1e-3.

- RMSE value for training is 1.83, for validation is 1.731, whereas the RMSE value for testing is 1.723.

- There were 146 iterations in this model with the initial RMSE train of 1.833963 and continued to decrease to 1.614349.
- The initial RMSE validation value is 1.732672 and continues to increase to 1.877191 (overfitting).
- The best iteration is the first iteration with RMSE training is 1.828181, and for RMSE validation is 1.730597.

3-4. Multilayer Perceptron 2 (4 Hidden Units, Iterations 200)




```
                              Optimization Results

Iterations                              189  Function Calls                        490
Gradient Calls                          223  Active Constraints                      0
Objective Function             1.6030190462  Max Abs Gradient Element     0.0019964216
Slope of Search Direction      -0.000025832

Convergence criterion (FCONV=0.0001) satisfied.

NOTE: At least one element of the gradient is greater than 1e-3.
```

- RMSE value for training is 1.98, for validation is 1.724, whereas the RMSE value for testing is 1.725.
- There were 189 iterations in this model with the initial RMSE train of 1.982259 and continued to decrease to 1.464764.
- The initial RMSE validation value is 1.732672 and continues to increase to 2.169641.
- The best iteration is the first iteration with RMSE training, which is 1.976888 and RMSE validation, which is 1.72416.

3-5. Neural Network Summary

Selected Model	Predecessor Node	Model Node	Model Description	Target	Target Label	Train: Root Mean Squared Error	Valid: Root Mean Squared Error	Test: Root Mean Squared Error
Y	Neural4	Neural4	MLP 8H Neural Network	Rating		1.976888	1.72416	1.724924
	Neural	Neural	MLP 2H Neural Network	Rating		1.768396	1.730018	1.720541
	Neural3	Neural3	MLP 4H Neural Network	Rating		1.829058	1.730597	1.722516
	Neural2	Neural2	GLM Neural Network	Rating		1.729611	1.731434	1.742782

VII. Findings

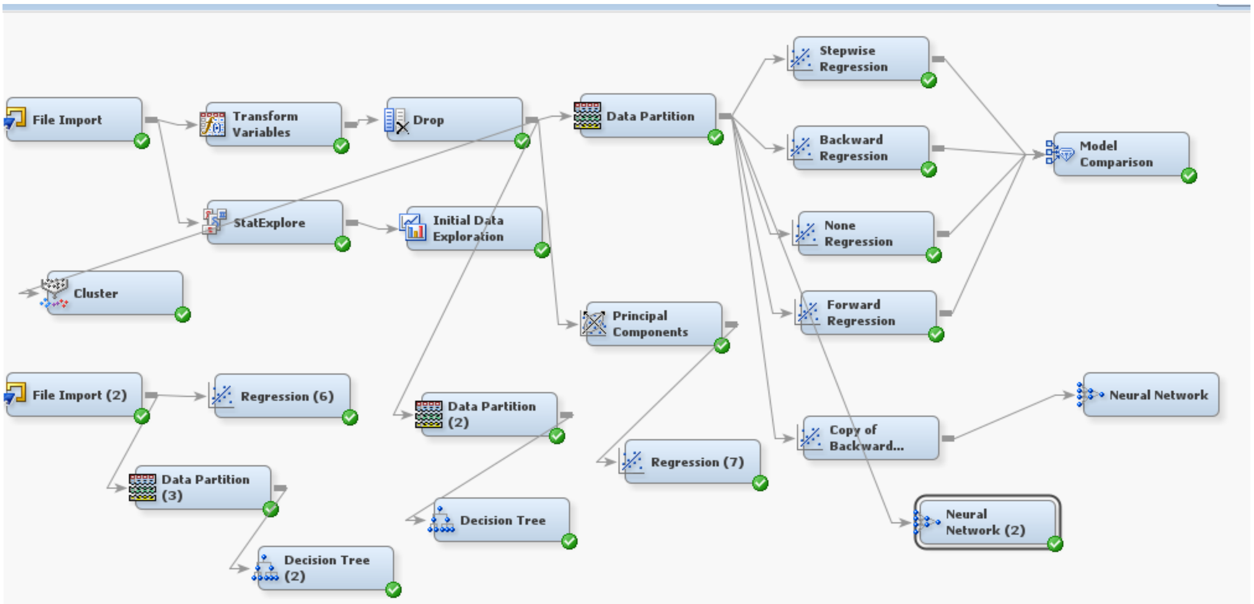
I found that none of the factors in the dataset are statistically significant in predicting the customer rating. The final regression model's Model R-square was 0.0089 means variables included in the model are not predicting the customer satisfaction well. Even though I could not build an ideal model that suits my initial purpose, I think I found some valuable observations in the process. This could be possible, for example, because EDA and Clustering does not necessarily determine models on the data, but it helps business owners to explore possible data analysis models that best suit the data based on the business problems and goals. Below are the findings:

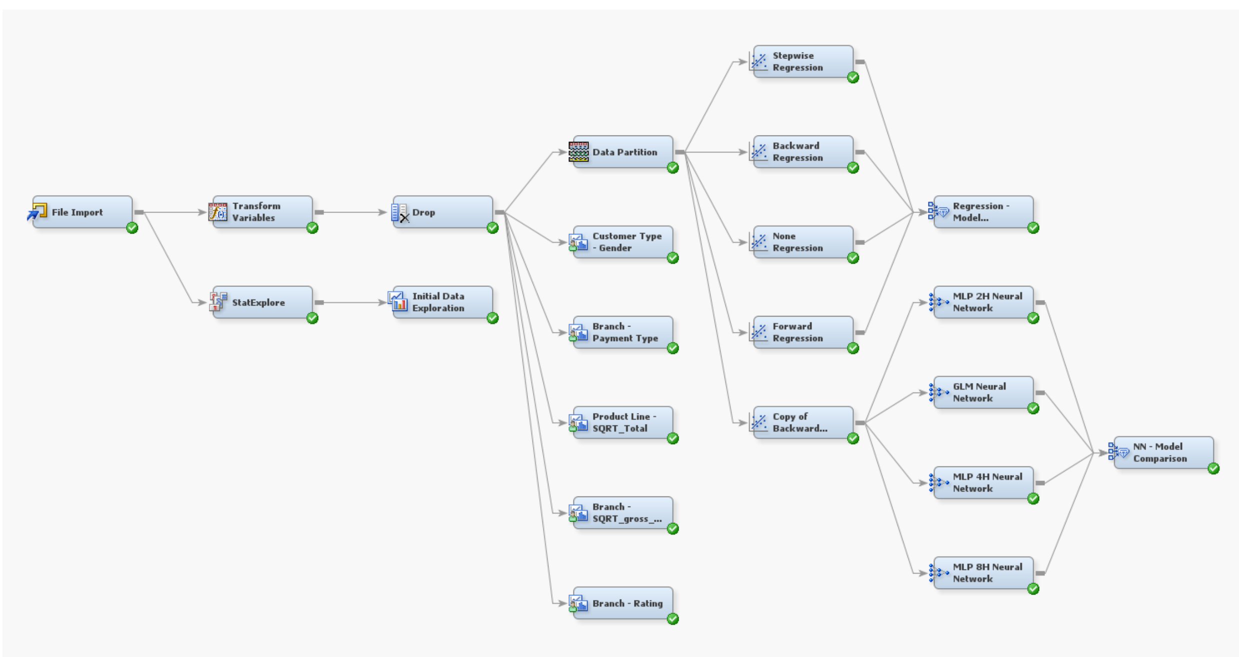
- Customers prefer to use cash when they bought electronic accessories items.
- Customers prefer to use their credit cards more when they bought for food and beverages items.
- Customers prefer to use e-wallet when they pay for home and lifestyle products or fashion accessories items.
- Overall, the most popular payment method is E-wallet and cash payment is also on the higher side.

- The customer rating is more or less uniform with the mean rating being around 7 and there is no relationship between gross income and customer ratings.
- Fashion accessories and food and beverages are the most sold product in Naypyitaw and these products should be focused on along with electronic accessories.
- Gross income is similar for both male and female, though female customers spend a bit higher at the 75th percentile. Females spend on ‘fashion accessories’ the most and for males surprisingly it is ‘Health and beauty’. Females also spend more on ‘Sports and travel’ which generates highest income overall.
- Using the correlation analysis, one interesting observation has emerged that customer ratings is not related to any variable.

VIII. Managerial implications/conclusions

Since I didn’t find any variables that are significant, I can’t say much based on the regression models. This is unfortunate, but I think this is also a information and the further insights could be made. There should be other unmeasured factors that are predicting the customer rating. For example, wait time in payment queues, availability of the products, approachability and customer service of sales associates can be some of those unmeasured confounders impacting the customer rating. Besides, I believe some of the observations I found in the process of building an ideal model serves some of the initial purpose of this project. For instance, Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but the quantity purchased is low. Hence, supply for these products need to be increased.


File Import
Transform Variables
Drop
Data Partition
Stepwise Regression
Backward Regression
Model Comparison
StatExplore
Initial Data Exploration
None Regression
Cluster
Forward Regression
Principal Components
File Import (2)
Regression (6)
Data Partition (2)
Neural Network
Copy of Backward...
Regression (7)
Data Partition (3)
Decision Tree
Neural Network (2)
Decision Tree (2)


Stepwise Regression
Data Partition
Backward Regression
File Import
Transform Variables
Drop
Regression - Model...
Customer Type - Gender
None Regression
StatExplore
Initial Data Exploration
MLP 2H Neural Network
Branch - Payment Type
Forward Regression
GLM Neural Network
Product Line - SQRT_Total
Copy of Backward...
NN - Model Comparison
MLP 4H Neural Network
Branch - SQRT_gross_...
MLP 8H Neural Network
Branch - Rating