Supermarket Sales Analysis

Group 8

- Chetan Bandi
- Soundarya Bachu
- Swati Kolpekwar
- Vishaka Sharma

Introduction & Overview

Purpose: Visualizing the Supermarket sales-data is important to understand the customer-satisfaction level which can impact the profit of the organizations drastically and can help in the development of the marketing strategies that can influence the purchasing behavior of different customer classes.

Dataset: 1000 observations capturing key metrics like Unit_price, Quantity, Total, Cogs, and gross_income.

Training set: 80%
Test set: 20%

Variables: Customer_type,Gender ,Product_line, Unit_price, Quantity, Tax, Total, Date, Time, Payment, cogs, gross_income, Rating

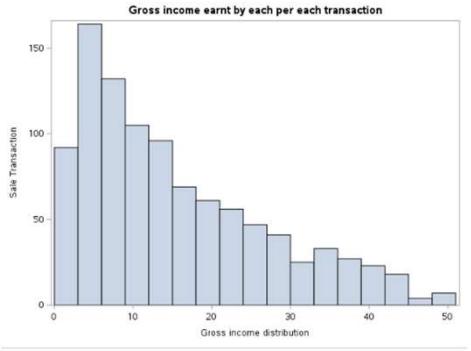
Rating Categorization: Rating classified into Low and High satisfaction levels to identify purchase patterns and build predictive models.
Includes quantitative (e.g., gross_income) and categorical (e.g., customer_type, gender) data.

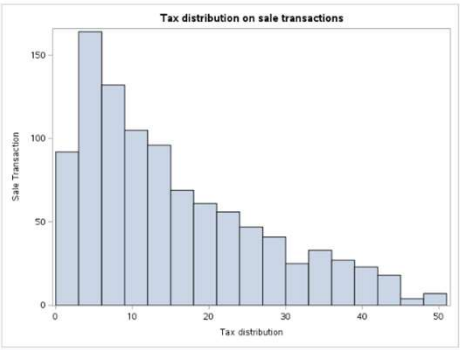
Goal: Understand and predict factors influencing customer satisfaction, as reflected in the rating provided by customers. By categorizing ratings into two classes (e.g., High and low satisfaction) .

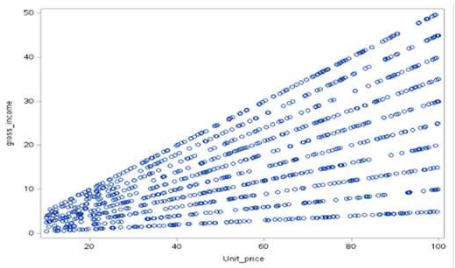
Data source: https://www.kaggle.com/code/aryantiwari123/supermarket-sales-prediction/input

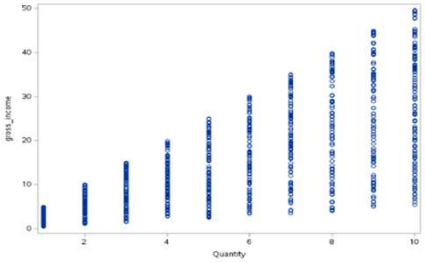
Plots

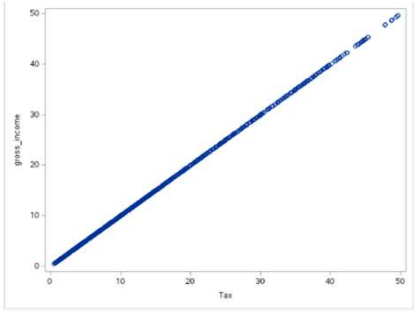
- The gross-income histogram is right-skewed and also significant number of sale transactions generate income between the range of 0-10 and the transactions generating income between the range of 40-50 is very low.
- The Tax distribution histogram is right-skewed and also most of the tax collected on the sales transaction is between 0-10 and the tax collected 40 on the transactions is merely low.

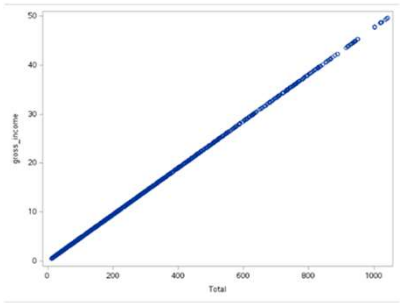






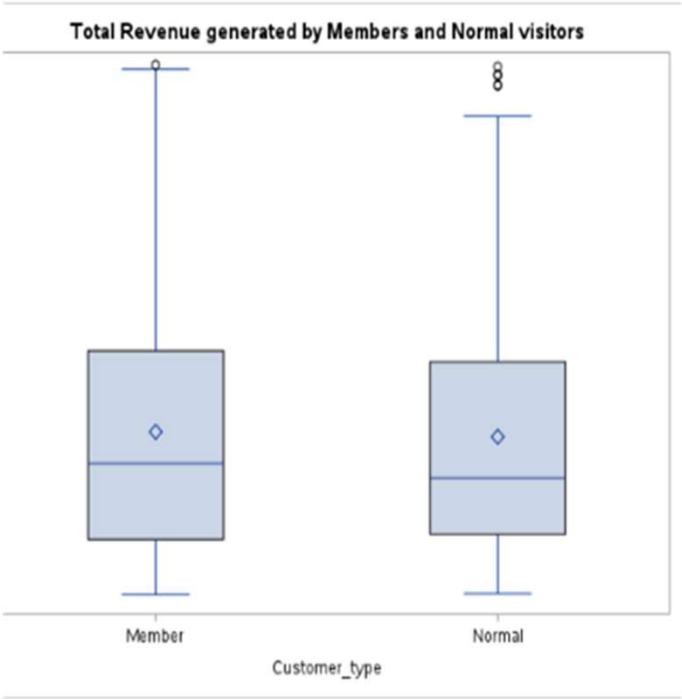




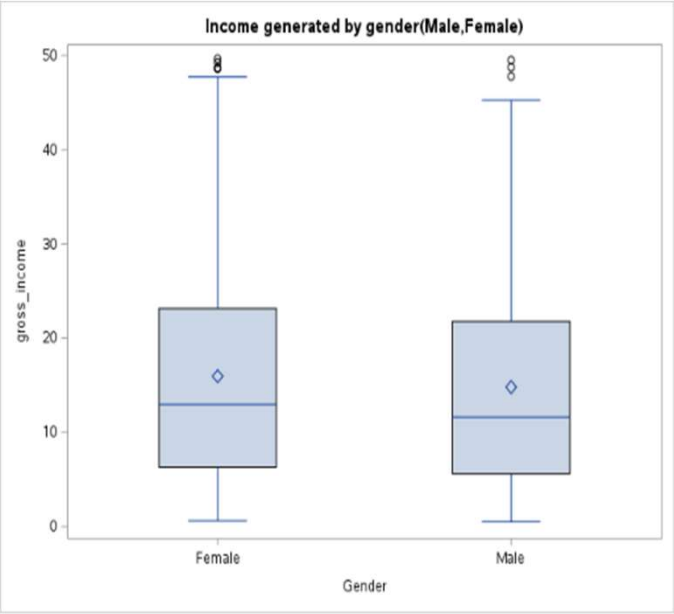


The scatter plots exhibit positive linear relationship between **Gross_income** and **Unit_price**, **Tax**, **Quantity** and **Total**, as expected. Increased Gross_income correlates with higher Unit_price, Tax, Quantity, and Total.

Plots



The revenue generated by the Members is high compared to Normal customers with the median value being slightly high . Also the mean total being high for members compared to normal visitors. The member and normal customers exhibit positive skewness with the total top whisker being high. But comparatively Normal customers have more **outliers** than members.
When looking at **IQR & Variability**, members demonstrates greater variability, which reflects more spending among its customers.



The median and mean value of females is slightly high compared to the median and mean value of males which slightly generates high income compared to males**.** Both the whiskers are right skewed with female whisker being slightly high. Both the categories also demonstrate very similar **IQR & Variability.**

Principal Component Analysis

Looking at the above PCA analysis Correlation would be better as the variables are measured in different units.

Looking at the Eigenvalues the first principal component accounts 61% of the data variability and the second Principal component accounts for 13% of the data variability and 86% of the variability is explained in the first three principal components only.

Looking at the most important features in each Principal component:

Prin1- Tax, Total, Cogs and Gross_income
Prin2 – Date and Time
Prin3 – Quantity
Prin4 – Date and Time
Prin5 – Quantity
Prin6 - Cogs
Prin7 – Gross_income
Prin8 – Total and Tax

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.92116498	3.86736475	0.6151	0.6151
2	1.05380024	0.06972484	0.1317	0.7469
3	0.98407540	0.03337384	0.1230	0.8699
4	0.95070156	0.86044373	0.1188	0.9887
5	0.09025782	0.09025782	0.0113	1.0000
6	0.00000000	0.00000000	0.0000	1.0000
7	0.00000000	0.00000000	0.0000	1.0000
8	0.00000000		0.0000	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Unit_price	Unit_price	0.291799	0.209941	-.712162	-.055160	0.600468	0.000000	0.000000	0.000000
Quantity	Quantity	0.324628	-.182152	0.640577	0.047066	0.669987	0.000000	0.000000	0.000000
Tax	Tax	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	-.408248	0.707107
Total	Total	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	-.408248	-.707107
Date	Date	-.013470	0.681666	0.143481	0.717318	0.004280	0.000000	0.000000	0.000000
Time	Time	-.002695	0.676742	0.248782	-.692897	-.003891	0.000000	0.000000	0.000000
cogs	cogs	0.449800	0.004934	0.001368	0.004786	-.218244	0.866025	0.000000	0.000000
gross_income	gross_income	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	0.816497	0.000000

Models

- Logistic Regression
- CART
- Neural Networks
- Discriminant Analysis

Logistic Regression

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	SALES
Random Number Seed	12345
Sampling Rate	0.8
Sample Size	800
Selection Probability	0.8
Sampling Weight	0
Output Data Set	SALES_PART

The LOGISTIC Procedure

Model Information		
Data Set	WORK.SALES_TRAIN	
Response Variable	Rating	Rating
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	800
Number of Observations Used	800

Response Profile		
Ordered Value	Rating	Total Frequency
1	0	412
2	1	388

Probability modeled is Rating='1'.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Is_Member	0.981	0.741	1.298
Is_Female	1.074	0.809	1.425
Is_Product_line_HB	1.127	0.688	1.845
Is_Product_line_EA	0.780	0.487	1.251
Is_Product_line_HL	0.845	0.525	1.361
Is_Product_line_ST	0.786	0.487	1.269
Is_Product_line_FB	1.154	0.720	1.848
Unit_price	1.005	0.994	1.017
Quantity	1.043	0.935	1.164
Tax	0.986	0.952	1.021
Date	0.999	0.993	1.004
Time	1.000	1.000	1.000
Is_Payment_Ewallet	1.073	0.767	1.502
Is_Payment_CC	1.151	0.813	1.631

Odds Ratio Estimate for Is_Member is 0.981, The odds of having rating one is 2% lower if the customer is member of the supermarket.

Odds Ratio Estimate for Is_Product_Line_HB is 1.127, The odds of having rating one is 12.7% higher for health and beauty compared to the default Fashion accessories product line.

Logistic Regression

- The model is statistically not significant.
- All the variables are not significant.
- AUC is acceptable.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.5439	14	0.8591
Score	8.5074	14	0.8613
Wald	8.4311	14	0.8657

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	56.2	Somers' D	0.124
Percent Discordant	43.8	Gamma	0.124
Percent Tied	0.0	Tau-a	0.062
Pairs	159856	c	0.562

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	24.6845	61.5697	0.1607	0.6885
Is_Member	1	-0.0194	0.1429	0.0184	0.8920
Is_Female	1	0.0710	0.1444	0.2414	0.6232
Is_Product_line_HB	1	0.1193	0.2517	0.2248	0.6354
Is_Product_line_EA	1	-0.2481	0.2408	1.0614	0.3029
Is_Product_line_HL	1	-0.1680	0.2429	0.4788	0.4890
Is_Product_line_ST	1	-0.2408	0.2443	0.9716	0.3243
Is_Product_line_FB	1	0.1431	0.2405	0.3540	0.5519
Unit_price	1	0.00543	0.00563	0.9274	0.3355
Quantity	1	0.0423	0.0560	0.5707	0.4500
Tax	1	-0.0139	0.0179	0.6029	0.4375
Total	0	0	.	.	.
Date	1	-0.00114	0.00285	0.1600	0.6892
Time	1	-8.31E-6	6.207E-6	1.7923	0.1806
Is_Payment_Ewallet	1	0.0709	0.1715	0.1708	0.6794
Is_Payment_CC	1	0.1410	0.1776	0.6298	0.4274
cogs	0	0	.	.	.
gross_income	0	0	.	.	.

	Training				Validation			
	AUC	Error Rate	Sensitivity	Specificity	AUC	Error Rate	Sensitivity	Specificity
Logistic Regression Model	0.56	45%	42%	67%	0.52	46%	46%	60%
Logistic Regression Model - Final	0.56	45%	42%	67%	0.52	46%	46%	60%

Fit Statistics for SCORE Data										
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC
WORK.SALES_TRAIN	800	-549.9	0.4525	1129.771	1130.384	1200.041	1200.041	0.010623	0.014168	0.561849
WORK.SALES_VALID	200	-138.4	0.4650	306.7487	309.3574	356.2235	356.2235	-0.00236	-0.00315	0.521757

The FREQ Procedure

Frequency
Row Pct

Table of Rating by I_Rating

Rating(Rating)	I_Rating(Into: Rating) 0	1	Total
0	275 66.75	137 33.25	412
1	225 57.99	163 42.01	388
Total	500	300	800

The FREQ Procedure

Frequency
Row Pct

Table of Rating by I_Rating

Rating(Rating)	I_Rating(Into: Rating) 0	1	Total
0	64 59.81	43 40.19	107
1	50 53.76	43 46.24	93
Total	114	86	200

CART

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	10
Number of Leaves Before Pruning	145
Number of Leaves After Pruning	138
Model Event Level	1

Number of Observations Read	1000
Number of Observations Used	1000
Number of Training Observations Used	800
Number of Validation Observations Used	200

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	0
Number of Leaves Before Pruning	122
Number of Leaves After Pruning	1
Model Event Level	1

Number of Observations Read	1000
Number of Observations Used	1000
Number of Training Observations Used	800
Number of Validation Observations Used	200

Gini Model: Complex with 138 leaves, captures more patterns but has high validation error, showing poor generalization.

Entropy Model: Too simple with 1 leaf, lacks complexity to capture meaningful patterns in the data.

CART Model: Gini vs. Entropy

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Training	0	368	44	0.1068
	1	43	345	0.1108
Validation	0	58	49	0.4579
	1	40	53	0.4301

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	138	0.0738	0.1088	0.8892	0.8932	0.3189	0.1476	118.0	0.9667
Validation	138	0.3721	0.4450	0.5699	0.5421	0.6496	0.3160	148.8	0.5590

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Training	0	412	0	0.0000
	1	388	0	1.0000
Validation	0	107	0	0.0000
	1	93	0	1.0000

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	1	0.2498	0.4850	0.0000	1.0000	0.9994	0.4996	399.6	0.5000
Validation	1	0.2492	0.4650	0.0000	1.0000	0.9965	0.4976	99.6700	0.5000

Gini Model

- Training Data:
  - Lower misclassification rates
  - Higher sensitivity and specificity
  - Performs well on training data
- Validation Data:
  - Struggles with generalization, indicating potential overfitting

Entropy Model

- High misclassification rates for both training and validation data
- Poor handling of class imbalances
- Training and validation AUC values near 0.5 shows poor predictive performance.

Conclusion

- The Gini model performs better overall, offering balanced metrics compared to the entropy model.

Neural Networks

- The misclassification rate is lower for both sets.
- The sensitivity is approx. 46% on training and 44% on validation set, suggesting that there is room for improvement in identifying true positives.
- The specificity is approx. 64% for both sets.

The model with 1 hidden layer of 12 neurons shows a moderate performance with no overfitting.

Adding another hidden layer – 12 neurons did not make much difference.

Model Information	
Data Source	WORK.SALES_PART
Architecture	MLP
Number of Input Variables	12
Number of Hidden Layers	1
Number of Hidden Neurons	12
Number of Target Variables	1
Number of Weights	253
Optimization Technique	Limited Memory BFGS

Number of Observations Read	1000
Number of Observations Used	1000
Number Used for Training	800
Number Used for Validation	200

Train: Misclassification Rate	Valid: Misclassification Rate
0.4450	0.4550

The FREQ Procedure

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	265 64.32	147 35.68	412
1	209 53.87	179 46.13	388
Total	474	326	800

The FREQ Procedure

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	68 63.55	39 36.45	107
1	52 55.91	41 44.09	93
Total	120	80	200

Model Information	
Data Source	WORK.SALES_PART
Architecture	MLP
Number of Input Variables	12
Number of Hidden Layers	2
Number of Hidden Neurons	24
Number of Target Variables	1
Number of Weights	409
Optimization Technique	Limited Memory BFGS

Train: Misclassification Rate	Valid: Misclassification Rate
0.4475	0.4850

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	290 70.39	122 29.61	412
1	236 60.82	152 39.18	388
Total	526	274	800

The FREQ Procedure

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	70 65.42	37 34.58	107
1	60 64.52	33 35.48	93
Total	130	70	200

Neural Networks

Reducing the neurons to 8

- The misclassification rate was reduced to 0.39 and 0.40 for the training & Validation set.
- Sensitivity remains at same levels for training & validation.
- Specificity improved for validation to 74%.
- There is still room for improvement in sensitivity, but overall the **model performed better with no sign of overfitting.**

Model Information	
Data Source	WORK.SALES_PART
Architecture	MLP
Number of Input Variables	12
Number of Hidden Layers	1
Number of Hidden Neurons	8
Number of Target Variables	1
Number of Weights	169
Optimization Technique	Limited Memory BFGS

Train: Misclassification Rate	Valid: Misclassification Rate
0.3900	0.4050

	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	317 76.94	95 23.06	412
1	217 55.93	171 44.07	388
Total	534	266	800

The FREQ Procedure

Frequency
Row Pct

Table of Rating by I_Rating			
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	79 73.83	28 26.17	107
1	53 56.99	40 43.01	93
Total	132	68	200

Discriminant Analysis

- 412 observations (52%) belong to Class 0 (unsatisfactory ratings).
- 388 observations (48%) belong to Class 1 (satisfactory ratings).
- Sensitivity: The model struggled to identify true positives effectively.
- Specificity: Achieved moderate accuracy in identifying true negatives.

Prior Proportion of 60-40

Class Level Information					
Rating	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	412	412.0000	0.515000	0.600000
1	1	388	388.0000	0.485000	0.400000

Model couldn't identify true positives.

Class Level Information					
Rating	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	412	412.0000	0.515000	0.515000
1	1	388	388.0000	0.485000	0.485000

Number of Observations and Percent Classified into Rating			
From Rating	0	1	Total
0	297 72.09	115 27.91	412 100.00
1	255 65.72	133 34.28	388 100.00
Total	552 69.00	248 31.00	800 100.00
Priors	0.515	0.485	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.2791	0.6572	0.4625
Priors	0.5150	0.4850	

Number of Observations and Percent Classified into Rating			
From Rating	0	1	Total
0	70 65.42	37 34.58	107 100.00
1	65 69.89	28 30.11	93 100.00
Total	135 67.50	65 32.50	200 100.00
Priors	0.515	0.485	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.3458	0.6989	0.5171
Priors	0.5150	0.4850	

Number of Observations and Percent Classified in			
From Rating	0	1	
0	412 100.00	0 0.00	
1	388 100.00	0 0.00	
Total	800 100.00	0 0.00	
Priors	0.6	0.4	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.0000	1.0000	0.4000
Priors	0.6000	0.4000	

Number of Observations and Percent Classified into Rating			
From Rating	0	1	Total
0	107 100.00	0 0.00	107 100.00
1	93 100.00	0 0.00	93 100.00
Total	200 100.00	0 0.00	200 100.00
Priors	0.6	0.4	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.0000	1.0000	0.4000
Priors	0.6000	0.4000	

Conclusion

	Training				Validation			
	AUC	Error Rate	Sensitivity	Specificity	AUC	Error Rate	Sensitivity	Specificity
Logistic Regression Model	0.56	45%	42%	67%	0.52	46%	46%	60%
CART Model (Gini)	**0.97**	**10%**	**89%**	**89%**	0.56	44%	57%	54%
Neural Network (hidden layer 8)		39%	44%	77%		**41%**	**43%**	**74%**
Discriminant Analysis		46%	34%	72%		51%	30%	65%

Model Comparison:

- The **CART model** performs well on the training with 89% sensitivity and specificity, but weaker error metrics on the validation set shows signs of overfitting.
- **Neural Network** Analysis: Outperforms other models on the validation set, with the highest sensitivity (43%) and specificity (74%), coupled with a relatively low error rate of 41%. This indicates its strong generalization ability, making it the most suitable model for predicting customer satisfaction.

Findings:

- Key factors influencing customer satisfaction through data-driven analysis - Total price, Date and Time.
- Models can predict satisfaction trends across similar customer demographics

Shortcomings:

- Potential biases in the data collection process could impact the generalizability of results.
- The dataset size is relatively small, which might limit the robustness of the findings.

Suggestions for Improvement:

- Increase the dataset size(2019) and diversity to capture a wider range of customer behaviors.
- Incorporate additional variables, such as customer feedback, to improve the accuracy and relevance of predictions.