THE CURIOSITY CUP 2024
A Global SAS® Student Competition

Model Comparison for Analysing Maven Coffee Shop by SAS Viya®

Kenneth Wong Meng Yong, Tan Ru Poh, Hew Kar Eun,
Yu Yung Jun and Lim Siew Mooi

Tunku Abdul Rahman University of Management and Technology

AnalyticaTitans

Abstract

This study explores the nuanced landscape of the coffee shop industry, harnessing the insights gleaned from historical sales data sourced from two distinct datasets. It delves into the intricate analysis of seasonal trends, discernment of customer preferences, and the development of robust demand forecasting methodologies within this specialized domain. Employing a meticulous blend of rigorous data analysis techniques such as regression analysis and predictive analytics, the research endeavors to unearth actionable insights that propel sales growth and optimize inventory management practices. Central to the study's objectives is a comprehensive understanding of historical sales patterns, coupled with the foresight to anticipate future demand fluctuations. By elucidating these dynamics, the aim is to enhance operational efficiency, minimize waste, and ultimately bolster profitability within coffee shop operations. Moreover, the research seeks to uncover innovative strategies tailored to augment sales performance and capitalize on emerging opportunities amidst the competitive coffee market landscape. Through the synthesis of empirical data and strategic insight generation, this study aspires to provide a roadmap for informed decision-making and sustainable business growth within the coffee shop industry. By bridging the gap between data-driven analysis and actionable strategies, stakeholders are empowered to navigate the complexities of the market landscape with confidence. Ultimately, this research aims to equip coffee shop proprietors and industry professionals with the requisite tools and knowledge to thrive in an ever-evolving market environment, driving toward long-term success and resilience.

Problem Statement

Current coffee demand?

Maven Roaster faces significant inventory management challenges due to frequent stockouts, overstocks, and reliance on a manual tracking system prone to errors. These shortcomings disrupt operations, impact revenue, and hinder strategic product associations, leading to customer dissatisfaction. To address these issues, Maven Roaster must adopt a data-driven inventory management system to streamline processes, reduce discrepancies, and make informed decisions. Embracing advanced tracking and forecasting tools will enable Maven Roaster to meet customer needs, maximise revenue, and stay competitive in the coffee industry.

Data Preparation

The data is from an open database obtained from Kaggle, and the author indicated that it was sourced from the Maven Analytics website. We performed cleaning on the data before using it for analysis, then it resulted in a dataset with 49,894 records with 13 features, which are outlined in Appendix Table 1. The original 201904 sales reciepts.csv data file was transformed into an SAS data file using SAS Studio® from SAS Viya for Learner®. Appendix Table 1 shows the variables, their values, and summary statistics. This newly formatted data file was then uploaded into SAS Viya for Learners to be used for the study.

Introduction

Coffee Sales Trends and Forecasting Strategies

Introduction - The Material and Method/Algorithm - Results and Discussion – Conclusion.

Coffee is a beverage brewed from roasted coffee beans. It can be served hot or cold. It is dark in colour and tastes bitter and slightly acidic. Coffee also has a stimulating effect on humans due to its caffeine leading to an increase in the activity of the brain and the spinal cord. This can improve the alertness of humans and increase their cognitive performance and physical strength. The coffee comes in 2 main varieties which are arabica and robusta. The coffee shops usually use high-quality arabica beans in speciality coffees such as espresso, latte, cappuccino and americano. In another way, robusta beans are more likely to appear in commercial or instant coffee since it has a longer harvest season and is cheaper than arabica beans.

Maven Coffee is a company that provides customised coffee service where they will control the beans from beginning to end to produce a special or brand-new taste. Maven Coffee also manage a coffee shop called Maven Roaster. It provides many types of coffee with different coffee beans. This becomes one of the characteristics of Maven Roaster.

One effective strategy for making informed decisions about finding sales trends and forecasting strategies is to consult the historical data of the sales transaction and product such as coffee type or pastries. Through this approach, the individual can obtain advanced insight into various aspects of Maven Roaster such as customer all-time favourites and inventory optimisation. This could help the decision maker to make the right decision and lead Maven Roaster to another level. Regression analysis is a method that

Maven Coffee

In the coffee shop market, analyse the inventory, pricing and sales of the product and transactions of customers. This research is about Maven Roaster a relatively new or niche coffee roaster. Coffee Roaster freshly roasted and expertly brewed. It enjoys specialty coffee in the comfort of your own home. The dataset comprises essential information, including product id, sales outlet id, frequency, transaction date and more. Our team has 3 different dataset and each offers unique insights into various aspects of association sales.

By combining these datasets, we aim to :

- Analysing historical sales data to understand seasonal trends and customer preferences.
- Utilising predictive analytics for demand forecasting and inventory optimization.
- Generate idea to increase the sales

Our team has 4 Objectives which are Analyze Historical Sales Data, Implement Anomaly Detection for Quality Control, Optimise Inventory Level and Facilitate Data-Driven Decision Making.

Data Preprocessing

The data cleaning process is confusing and hard to follow. Very likely, some important steps of the data cleaning process were omitted.

Figure 1, 2 and 3 in **Appendix A** shows the results of missing values for each dataset. As you can see Product and Pastry Inventory didn't have any missing values, but there are 168 of missing values in the Sales_Receipt dataset. After we found this, we didn’t plan to impute or replace the missing values as the attribute instore_yn is talking about whether the customers bought the products through physical or online purchases.

Hypothesis Testing 1

There is no significant difference in coffee demand across different day.Testing this hypothesis can help Maven Roaster better anticipate seasonal fluctuations in demand and adjust **inventory levels** accordingly.

Data Models

Based on **Appendix B**, we can conclude that the hypothesis is rejected. Which means that there is a significant difference in coffee demand across different days. Besides, based on the visualised graph we can find that the coffee sales for each day is different. So, by using the results, we can do forecasting on the inventory levels for one week or one month in order to avoid waste of inventory and also the quality.

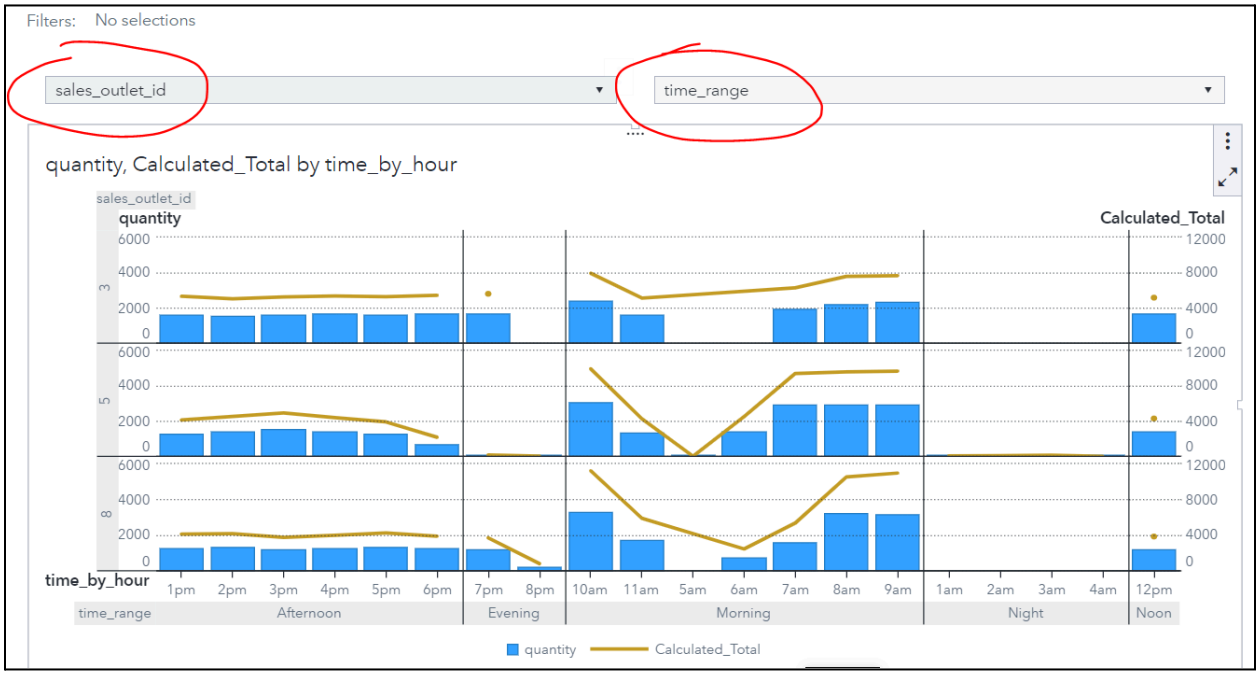
Hypothesis Testing 2

Transaction times exhibit a normal distribution throughout the day. Visualising transaction times throughout the day can reveal patterns indicative of peak customer activity, staff scheduling issues, or system glitches.

Through visualising transaction time patterns on graphs or plots, Maven Roaster can discern peak sales periods, enabling the definition of targeted business strategies to optimise sales performance.

Data Models

Figure below shows the quantity, calculated total (quantity * unit price) by the time by hour. Currently is visualisation for all outlets with all time ranges (morning, noon, afternoon, …). Choose the outlet's ID from the option list or the desired time frame to better visualise a certain outlet or certain range. (Red circle)

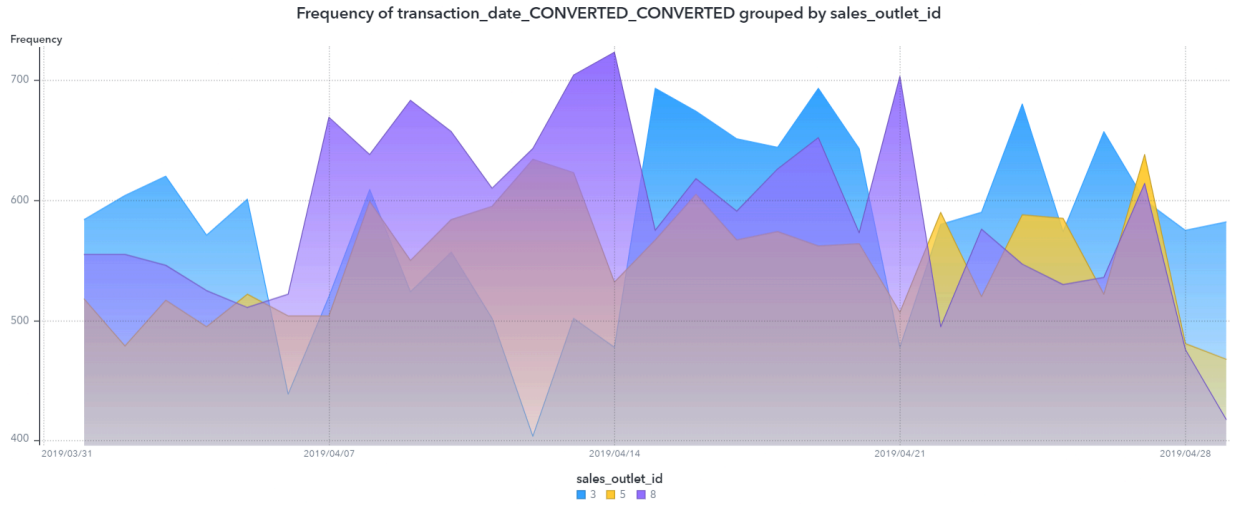


The findings indicate that sales were lowest at night across all shops, while the highest sales occurred in the morning. These allow companies to suggest various business plans for raising sales and setting up inventory stock levels.

Hypothesis Testing 3

Transaction volumes at **all sales outlets** follow a consistent distribution over time. Anomalies in transaction volumes at specific sales outlets may indicate operational issues, popularity variations, or potential fraudulent activities

Data Models



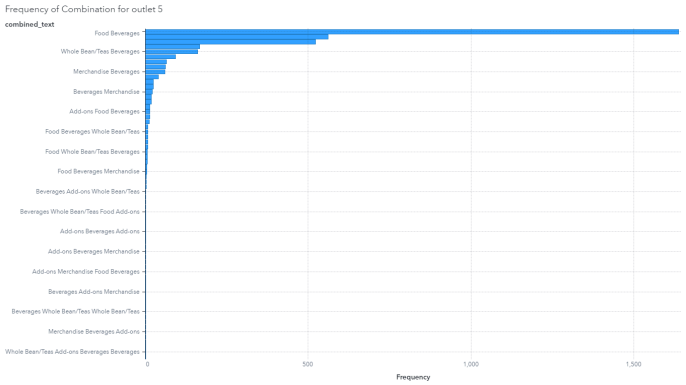
As the result showed, outlet 8 is the most frequent sales outlet which is 17071, second is outlet 3 which is 16829 and the third is outlet 5 which is 15994.Between all the outlets, the most frequency outlet sales is outlet 8 which is 723 and the least frequency outlet sales is outlet 3 which is 404.Comparison of all outlets, the most frequent outlet sales is outlet 8 which is 723 at 2019/04/14 and the least frequent outlet sales is outlet 3 which is 404 at 2019/04/12.

Hypothesis Testing 4

Most of the customers will buy beverages and food as a combination of meals. **Anomalies in the pastries sold during transactions** involve analysing transactional data and comparing it with inventory records to identify discrepancies or irregularities.

Data Models

As a result, the text combined will show the combination of the customers usually and the Food and beverage are the most combined. It means that the marketing team can make a combination promotion to increase sales and generate more profits.



CONCLUSION AND FUTURE STUDIES

This paper outlines a standard approach for analysing a dataset of a coffee shop using SAS Viya. The study found that the analysis conducted on Maven Roaster's data reveals several significant findings regarding its operations and customer behaviours. Firstly, there is no notable difference in coffee demand across various days, aiding in the anticipation of seasonal fluctuations and inventory management. Secondly, the normal distribution of transaction times suggests the possibility of peak customer activity, staffing concerns, or system irregularities, requiring closer monitoring and adjustment. Thirdly, consistent transaction volumes across sales outlets imply stable operations, yet anomalies may signal operational challenges or popularity shifts, necessitating further investigation. Lastly, the observation that customers often purchase beverages and food together indicates an opportunity for targeted promotions and increased profitability, reinforcing the importance of understanding customer preferences and optimising product offerings across all outlets. In short, this project enables Maven Roaster to analyse its operations, understand customer behaviour, optimise inventory management, and implement targeted strategies to enhance sales and customer satisfaction. Future studies could focus on understanding the drivers behind coffee demand patterns, exploring factors influencing anomalies in transaction times, analysing performance metrics at individual sales outlets, and investigating the impact of promotional strategies on customer purchasing behaviour. These insights would facilitate informed decision-making and foster sustainable growth for Maven Roaster in the competitive coffee market.

Acknowledgements

We would like to acknowledge Dr. Lim Siew Mooi for her supervision and support throughout our work.

Appendix A

Find missing value

Input Variable Statistics

Obs	Input Variable	Measurement Level	Number of Missing Values	Percentage Missing	Imputable	Minimum	Maximum	Mean	Midrange	Standard Deviation	Skewness	Kurtosis	Label
1	instore_yn	BINARY	168	0.56118	0	.	.	.	.	.	.	.	
2	line_item_amount	INTERVAL	0	0.00000	0	0	360	4.660316665	180.0	3.700921333	32.365287804	2859.7506571	
3	line_item_id	NOMINAL	0	0.00000	0	.	.	.	.	.	.	.	
4	order	NOMINAL	0	0.00000	0	.	.	.	.	.	.	.	
5	product_id	BINARY	0	0.00000	0	.	.	.	.	.	.	.	
6	promo_item_yn	BINARY	0	0.00000	0	.	.	.	.	.	.	.	
7	quantity	NOMINAL	0	0.00000	0	.	.	.	.	.	.	.	
8	sales_outlet_id	NOMINAL	0	0.00000	0	.	.	.	.	.	.	.	
9	staff_id	INTERVAL	0	0.00000	0	6	45	25.344222868	25.5	12.486652124	0.3379712875	-1.268101223	
10	transaction_date	BINARY	0	0.00000	0	.	.	.	.	.	.	.	
11	transaction_id	INTERVAL	0	0.00000	0	1	4199	868.94999499	2100.0	858.7012834	1.3326656231	1.4109861494	
12	transaction_time	NOMINAL	0	0.00000	0	.	.	.	.	.	.	.	
13	unit_price	INTERVAL	0	0.00000	0	0.8	45	3.3869258777	22.9	2.6723443375	8.4297316419	97.711125564	

Figure 1 : Sales_Receipt Dataset

Input Variable Statistics

Obs	Input Variable	Measurement Level	Number of Missing Values	Percentage Missing	Imputable	Minimum	Maximum	Mean	Midrange	Standard Deviation	Skewness	Kurtosis	Label
1	current_retail_price	INTERVAL	0	0	0	1	23	6.5471698113	12.00	5.8130027913	1.5277563902	1.2342433533	
2	current_wholesale_price	INTERVAL	0	0	0	0.04	16.8	3.7249056604	8.42	4.7402919509	1.611932679	1.5671661314	
3	new_product_yn	BINARY	0	0	0	.	.	.	.	.	.	.	
4	product	NOMINAL	0	0	0	.	.	.	.	.	.	.	
5	product_category	NOMINAL	0	0	0	.	.	.	.	.	.	.	
6	product_description	NOMINAL	0	0	0	.	.	.	.	.	.	.	
7	product_group	NOMINAL	0	0	0	.	.	.	.	.	.	.	
8	product_type	NOMINAL	0	0	0	.	.	.	.	.	.	.	
9	promo_yn	BINARY	0	0	0	.	.	.	.	.	.	.	
10	tax_exempt_yn	BINARY	0	0	0	.	.	.	.	.	.	.	
11	unit_of_measure	NOMINAL	0	0	0	.	.	.	.	.	.	.	

Figure 2 : Product Dataset

Input Variable Statistics

Obs	Input Variable	Measurement Level	Number of Missing Values	Percentage Missing	Imputable	Minimum	Maximum	Mean	Midrange	Standard Deviation	Skewness	Kurtosis	Label
1	% waste	NOMINAL	0	0	0	.	.	.	.	.	.	.	
2	product_id	NOMINAL	0	0	0	.	.	.	.	.	.	.	
3	quantity_sold	INTERVAL	0	0	0	0	30	9.2445652174	15	5.3826246291	1.3704708024	2.1906234831	
4	start_of_day	BINARY	0	0	0	.	.	.	.	.	.	.	
5	transaction_date	NOMINAL	0	0	0	.	.	.	.	.	.	.	
6	waste	INTERVAL	0	0	0	0	46	15.201086957	23	11.594600162	1.4023735527	0.8563869688	

Figure 3 : Pastry Inventory Dataset

Appendix B

Hypothesis 1

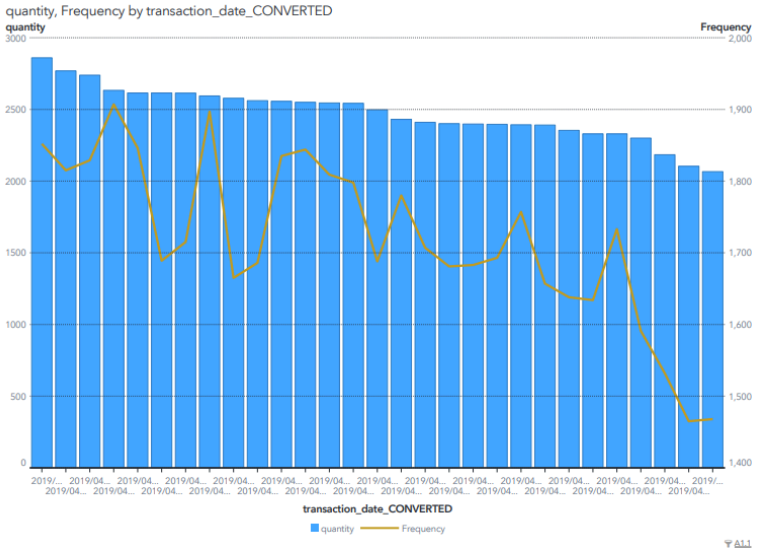


Figure 1

2/5/24, 5:41 PM Results:Program.sas

The REG Procedure
Model: MODEL1
Dependent Variable: quantity_numeric

Number of Observations Read	29
Number of Observations Used	29

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	88625	88625	2.76	0.1085
Error	27	868236	32157		
Corrected Total	28	956861			

Root MSE	179.32343	R-Square	0.0926
Dependent Mean	2474.41379	Adj R-Sq	0.0590
Coeff Var	7.24711		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-140602	86184	-1.63	0.1144
transaction_date_CONVERTED	1	6.60739	3.98005	1.66	0.1085

The REG Procedure
Model: MODEL1
Dependent Variable: quantity_numeric

References

www.kaggle.com. (n.d.). Coffee shop sample data (11.1.3+). [online] Available at: https://www.kaggle.com/datasets/ylchang/coffee-shop-sample-data-1113/data [Accessed 21 Feb. 2024].

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

WONG MENG YONG KENNETH
kennethwmy-wm21@student.tarc.edu.my

KAR EUN HEW
hewke-wm20@student.tarc.edu.my

RU POH TAN
tanrp-wm21@student.tarc.edu.my

YUNG JUN YU
yuyj-wm20@student.tarc.edu.my

Dr Lim Siew Mooi

Department of Mathematical And Data Science, Faculty of Computing And Information Technology
BSc (Hons) (UTM), MSc, PhD (UPM)
siewmooi@tarc.edu.my
Major of Study/Specialization: Intelligent Computing
Area of Interest: Machine Learning, Forecasting and Predictive Modelling, Natural Language Processing