Insights Data Report

Laura Wangari, Melissa Michuki, Paul Ndirangu,
Collins Kemboi & Peter Kiragu

Moringa School

Table of Contents

Business Understanding

Problem Statement

Every day, supermarkets in Kenya process thousands of transactions. In each transaction, there is a lot of data about the customers that supermarkets can aggregate and try to make sense of so as to improve their services and product offering. The data available includes the amount of money that a consumer spends on their shopping, the type of products they buy and the method of payment used. When this data is collected and analyzed, some important insights can be acquired and used to improve the products and services at the supermarket.

Objective

The objective of this analysis is to investigate consumer spending patterns in selected Kenyan supermarkets. We are interested in understanding how much consumers spend across different times of the week and for different product categories. Our claim is that consumers spend more money in supermarkets on the weekends as opposed to weekdays. As such, our hypothesis will be testing whether the average spending in supermarkets during weekdays is similar to the average spending during the weekends.

The null hypothesis is: H_0: $\mu 1 = \mu 2$

The alternative Hypothesis is: H_a: $\mu 1 \neq \mu 2$

This hypothesis is important because it will help us understand when consumers spend more money at supermarkets. This information can be used by the supermarkets to make decisions

around their promotion activities, staffing and stocking to ensure enough products are available to meet demand.

Data Understanding

The data used for this analysis was sourced from Kaggle. The data was downloaded in excel form which makes it easier for us to load and conduct further analysis.

The data used has 27 features and 1464 entries. The features of the data include:

Feature	Explanation	Feature	Explanation
Supermarket	Name of the supermarket where the purchase was made	Mall	Whether or not the supermarket is in a mall
No_of_items	Number of items bought in the transaction	Time	The exact time when the transaction was completed
Variation		Time type	Classifies time as morning, afternoon or night.
Total	Total amount in Ksh spent by the buyer on the transaction	Type market	Describing the type of market that the supermarket operates.
Paid	Amount paid by buyer	Location	Place where the supermarket is located.
Change	Change returned to the buyer	Loc_category	Category in which the location lies. This is either a mid-level, or high-end location.
Type	This describes the payment method used	Day	Day of the week the purchase was made

Food	Whether the product was food or not	Day_type	Classifies day as either weekend or weekday.
Snack	Whether the product was a snack or not	24hr	Describes whether the supermarket operates 24 hours or not
Beverage	Describes whether the product was a beverage or not	Day_1	Describes day of the month
Consumables	Whether the product was a is a consumable or not	Month	Describes the month the purchase was made
High_end	Whether the purchase made was for a high-end product or not	Year	Describes the year the purchase was made
Asset	Whether the purchase made was for an asset or not		
Fixed asset	Whether the product made is a fixed asset or not		
Date	The data of the purchase was made		

Data Preparation

Data preparation included cleaning the dataset and selecting the variables that would be used for univariate analysis, bivariate analysis, multivariate analysis and hypothesis testing.

Data cleaning was the first data preparation step taken. Data cleaning involved checking for duplicated values and removing these values. The dataset had 3 null values which were dropped. The dataset had outliers in the numerical columns which were also dropped. It is

important to note that the outliers in the numerical columns were less than 2% of all the entries thus dropping them did not significantly affect our dataset.

To ensure that the data was ready for univariate analysis, the dataset was divided into numerical and categorical variables. This allowed for numerical and categorical variables to be analyzed and explored effectively. By generating the frequency table for the categorical variable, it was possible to plot frequency distributions and pie charts for specific variables. A frequency distribution table was also developed from the numerical variables.

Only the numeric variables were used in the bivariate analysis. The data frame that was separated for univariate analysis was used for bivariate analysis. Similarly, only the numeric variables were used to plot the correlation matrix as well as compute the Pearson correlation coefficient.

For multivariate analysis, both numeric and non-numeric variables were used. The categorical variables used were encoded with ones and zeros to make sure that they could be used for dimensionality reduction.

Data Exploration

The data was explored at three levels namely univariate, bivariate and multivariate.

Univariate Analysis

For univariate analysis, a frequency table, pie charts and frequency distributions were used to summarize categorical variables. The categorical variables in the dataset include a

supermarket, type, food, snack, beverages, consumables, High_end, asset, fixed_asset, mall, time type, type_market, location and loc_category.

The data collected was from 23 supermarkets and the top five supermarkets were Karrymart, Tumaini, Nakumatt, Cleanshelf, and Tuskys. These top five supermarkets had more than a hundred entries in the data with Karrymart being the top supermarket with 510 entries.

The frequency table reveals that cash is the most preferred payment method with 1332 customers paying with cash. Mpesa is the second most popular payment method with 50 customers paying using Mpesa. Only 25 people used debit cards while other customers paid with a credit card, vouchers, and redeem points to pay for their shopping.

Food and non-food items were equally purchased in the supermarkets. Snacks were the most popular food type bought in supermarkets. Among all the purchases made, food items or consumables were the majority. Also, non-beverages were more than beverages. Very few customers purchased assets or fixed assets.

Only about 6 per cent of the purchases analyzed were from supermarkets are located in malls and just over 5 per cent are from high-end locations. Most of the purchases analyzed were from supermarkets located within the CBD, Ongata Rongai, Saika and Umoja.

The total number of purchases made in the evening, afternoon, morning and night was 390, 389, 334 and 299 respectively. This means that there was no significant difference between purchases made in the evening and the afternoon and only a slight difference between purchases made in the morning. Only two purchases were made during midnight which begs the question of whether operating 24 hours makes economic sense.

The frequency distribution for numeric variables reveals they follow a normal distribution and are skewed to the right.

Bivariate Analysis

The bivariate analysis was conducted on numeric variables which included the number of items, variation, the total amount for the purchase, the amount paid and change.

The bivariate analysis included a pair plot and correlation matrix. From the pair plots, only the number of items variable and total variable have a relationship. This can be explained by the fact that the items in a shopper`s basket will determine how much they will spend in the supermarket. The correlation matrix also shows that number of items is highly correlated with the total amount spent as well as the amount paid.

Multivariate Analysis

Discriminant analysis was used to implement the multivariate analysis. In the multivariate analysis, day type was the dependent variable while the independent variables were the number of items, variation, total, paid, change, food, snack, beverage, consumables, high-end, and fixed asset. The observation made was that the model is 78 per cent accurate in predicting the day of the week when the purchase was made.

Modelling

The data was modelled using classification techniques. Classification is a process of predicting the class of given data points. This type of predictive modelling allows data scientists to approximate a mapping function from input variables to output variables. Two

classification techniques were used and these are discriminant analysis and Naive Bayes probabilistic classifier.

Discriminant Analysis

Discriminant analysis was used to try and create a prediction of the day of the week a purchase was made. To achieve this, the dataset was divided into dependent and the independent variables. The dependent variable that was being predicted is the day of the week represented by the variable day type. The independent variables used to predict the dependent is the number of items, variation, total, paid change, food, snack, beverage, consumables, high-end, and fixed asset.

Naive Bayes

Naive Bayes classifier is different from discriminant analysis because it looks at independent features separately to determine how these features contribute to the probability of getting the dependent variable. The dependent variable that was being predictedThe first step in hypothesis testing was to group the dataset by day type. This generated two datasets, one representing the shopping done on weekdays and another representing shopping done on the weekend. using Naive Bayes is the day type. The features that were analysed to determine the type of day include the number of items, variation, total, paid change, food, snack, beverage, consumables, high-end, and fixed asset.

Hypothesis Testing

After separating the dataset by day type, it was observed that the dataset for the weekday had more entries than that for the weekend. The shape of the weekday dataset was transformed to be equal.

The sample size was selected from 265 entries and this was 10 per cent of the population. The sample size used was 26. This means that the ideal test statistic is the t-test.

The Levene test was implemented to check whether the variances for the two groups are the same. The alpha used for the test is 5%.

Evaluation

The models built to predict the day type by inputting the independent variables were able to predict whether the day was a weekday or a weekend. Linear Discriminant Analysis (LDA) was able to accurately predict 78.09 per cent of the day type in the test sample. On the other hand, Naive Bayes classification model was able to accurately predict 79.94 per cent of the day types in the test sample. This means that the Naive Bayes classification model is much more effective in determining the type of day given the independent variables used. There is however concern about the effectiveness of the models in helping us determine what customers buy across different days of the week. The conclusion from this is that the number of items, variation, total, paid, change, food, snack, beverage, consumables, high-end, and fixed asset are not effective in telling us the type of day when a purchase was made. This model does not have an economically viable business application but it allows supermarkets to understand consumption patterns of their customers across the week.

The hypothesis test reveals that the p-value is 0.3632. This means the null hypothesis should not be rejected at a 5 percent significance level. The test statistic is 0.8420. The implication for this in the business sense is that supermarkets understand that their sales are different during the weekdays as opposed to the weekends.

Deployment

Analysis conducted will be implemented by providing a summary of the variables analyzed to help the supermarkets understand consumer behaviour and preferences. The information on payment methods can help supermarkets plan accordingly to ensure that they have enough change to give back to customers. The information on popular locations for shopping can be used by the management to determine the location for the next chain store. The hypothesis tests also show that the average spending on weekdays is not the same as average spending on weekends. This information can be used to create campaigns that encourage people to shop more during the weekends. Such information could also be used to determine staffing needs during different times of the week.