Mariama SOUMAHORO

Store Analysis

ABSTRACT

Making business decisions to increase sales and overall experience of customers are some of the most important parts of running a business. For the longest, businesses have been making those decisions based on guesses, intuitions, and sometimes luck instead of using all the information provided in the data that has been accumulated over the years. Fortunately, a shift has been made since the introduction of big data. Nowadays, businesses rely on information retrieved from their data to make driven business decisions. This applies to any type of business, in particular stores. Indeed, to keep customers satisfied, stores utilize information collected on their clients to categorize them. The department store's data presented to us has for purpose to find patterns and similarities among the customers' behavior as well as recognizing important variables. Using statistical methods, we identified different hidden patterns and groups based on the given information. The main appearing trend involves customers who previously visited the store the most tend to have the highest number of items bought, the highest net_sales as well as gross sales regardless of their gender, marital status, or type. Saturday is the day of the week with the highest total gross sales. We also noticed that Net_Sales, Gross_Sales, Previous_spent, Items, Previous_Visits were important variables that help us differentiate customers. All the information found will help the store to increase their sales while catering to the needs of their customers.

1. DATA CHARACTERISTICS

Data Collection

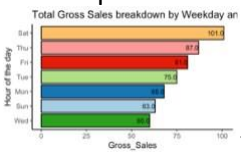
The goal is to be able to identify trends and patterns that are appearing and might be relevant for the Department store. The dataset used for this analysis contains 1000 observations with 19 variables. Some variables are duplicates. The covariates are:

- Customer: Id's customer; each unique
- Type of Customer: The customer is either promotional or regular. The Duplicate is Customer Dummy Promo: 1 meaning the customer is Promotional and 0 meaning the customer is regular
- Items: Number of items bought by the customer
- Net Sales: Company's gross sale minus deductions
- Gross Sales: total sales received by the store before any deductions
- Total Discount: Discount's percentage on products for customers
- Method Payment: American Express, Discover, MasterCard, RazorCard, Visa
- Gender: Female, Male
- Marital Status: Married, Single
- Age: Age of the customer
- Birthdate: month/day/year
- Date Purchase: date of purchase. Variable: Day Purchase is a duplicate since it just shows the day of the customer's purchase.
- Time Purchase: Time of purchase; Hours: Minutes: Seconds. Variable: Time_Purchase_1 is a duplicate.

• Loyalty Member: 0 represents customers who are not loyal and 1 represents customers who are loyal to the store.
• Previous Visits: number of previous visits made to the department store by the customer.
• Previous Spent: Total amount of money spent on previous purchases.

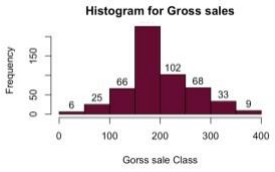
2. ***DATA VISUALISATION AND EXPLORATION***

We take a closer look at the information given to have a general idea of the data. Variables: Customer_Dummy_Pro, Day_purchase, Time_Purchase_1 are duplicates; therefore, we remove them from the database. Since Age variable is present, we remove the Date of Birth variable as well from the data. We extract the hours from the time variable and weekdays from variable: Day of purchase. We also remove observations with missing values which leaves us with a total of 535 observations that we will be using for our study. All the categorical variables appeared as non-numerical. We proceed by recoding them into categorical (Gender, Marital Status, Type of Customer, Method payment, and loyalty member) variables giving us a total of 6 categorical and 8 continuous variables. We continue to go through each variable in the input data to gain necessary insights. Our first step consists of visualizing gross sales

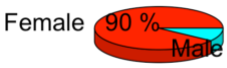


throughout the week. Looking at the bar chart, Saturday looks like the busiest day of the week since it has the highest total gross sale. We go further by checking exactly at what time there is a pick in the sales. The plot below is observed.

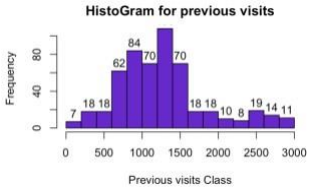
Sunday and Monday have a peak at 12pm. Tuesday, Wednesday and Thursday have peaks at 11am. Friday has an increase in sales at 5pm. As for Saturday, the sales increase both at 11am and 2pm.
Looking at variable: Age, we observe the minimum age to be 12 years old and the maximum to be 70 years old. The average age of a customer is 42. Regarding gross sales:



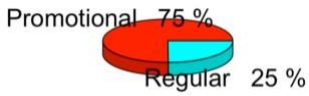
Looking at the histogram, we conclude that the minimum Gross sale of the customers is 35.12 and the maximum Gross sales is 374.02. Customers between classes 150 and 200 have the highest gross sales. The average gross sales of all the customers is 196.91.
Looking at the gender variable,



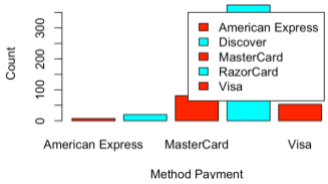
90% of the customers are female and only 10% are males. The minimum net sales is 28.1, with a maximum of 299.2 and an average of 163.9.
We then take a look at the histogram for the previously spent variable.



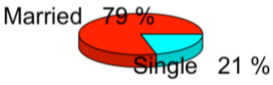
Indeed, we observe that people spending between $1000 and $1500 have the highest frequency (108). The minimum spent is 106.5 and the maximum spent is equal to 2987.8 with only a few people (11) in this category.
1 item is the minimum item purchased and 15 items represent the maximum items bought by customers.



The Pie chart shown above represents the pie chart for the type of customers we have in the data. the percentage of promotional customers is 75%, whereas the percentage of the regular customer dataset is 25%. Next, we look at the histogram plot for the method of Payment used by customers.



RazorCard was the most used card followed by MasterCard. American Express is the least used.



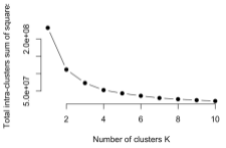
The pie chart for the Marital Status variable shows that 79% of customers are married and 21% are single.

After viewing the data, the next step into the analysis is to identify relevant patterns.

3. UNSUPERVISED LEARNING AND INTERPRETATION

This section aims to summarize the hidden patterns and trends in our unlabeled data. To do so, we perform what we call unsupervised learning. Indeed, unsupervised learning uses machine learning algorithms that analyze and cluster unlabeled datasets. With this process, we let undetected information guide us into discovering meaningful clusters and trends to better understand the data.

For our analysis, we first start with the method called K-means clustering. This algorithm is one of the most popular "clustering" algorithms. K-means stores k centroids that it uses to define clusters. A point is in a particular cluster if it is closer to that cluster's centroid than any other centroid. Working with clusters requires us to specify the number of clusters. To help determine the optimal clusters, we use the "Elbow Method". With this method, we first calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. The plot obtained denotes the appropriate number of clusters required in our model.



From the above graph, we conclude that k= 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot. Silhouette method and Gap statistic are two other methods used to determine the optimum clusters. When using those methods, we get the same answers confirming that we have four optimal clusters. Using this information, we proceed by computing the gap statistic for the k-means. The results are shown below:

```
Within cluster sum of squares by cluster:
[1]  9430996 11820315 17294927 14333982
 (between_SS / total_SS =  77.3 %)

Available components:

[1] "cluster"      "centers"      "totss"
[4] "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

In the output of our K-means operation, we observe a list with several key information. From this, we conclude the useful information being:

Cluster: This is a vector of several integers that denote the cluster which has an allocation of each point.

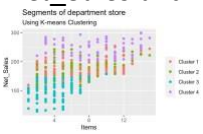
Totss: This represents the total sum of squares.

Withinss: This is a vector representing the intra-cluster sum of squares having one component per cluster.

Betweenss: This is the sum of between-cluster squares.

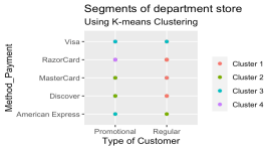
After checking the proportions for each cluster, we realize that cluster1 accounts for

14% of the data, cluster2 and cluster3 have both 28%, and cluster4 accounts for 30% of the data.
We visualize the clustering results for different variable used in the process. We first start with K-means clustering for Net_Sales and Items.



From the visualization, cluster 3 represents the customer data with the lowest number of items bought and the lowest net_sales. Cluster 1 and 3 represent customers with medium net_sales and medium item bought. Cluster 4 has customers with a high number of items bought and high net_sales.
We continue by visualizing the relation between Items and the Type of Customer.

Next, we look at the Method_Payment and Type of Customer.

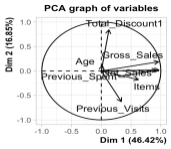


Cluster1 represents customers with only regular customers using both Discover, MasterCard RazorCard, and Discover. Cluster2 has both promotional customers using Discovery card and MasterCard; and regular customers using American Express. Regarding Cluster 3, Promotional customers use American Express and Visacard. Regular customers use only Visacard. Cluster4 contains only promotional customers only using RazorCard.
We take a look at the relationship between the Type of customers and their gender.

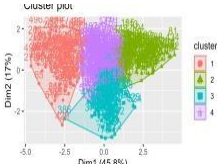
The K-means Clustering plot above shows that cluster1 contains Male regular customers. Cluster3 has male promotional customers and female regular customers. Cluster4 contains only female promotional customers.

The above plot represents the visualization for Variables: Previous_Spent and Previous_Visits. Cluster1 and Cluster2 contain customers with medium previous spent who made a previous visit ranging from 0 to 20. Cluster3 represents customers with the lowest amount spent and lowest number of previous visits. Cluster4 contains customers with a high amount of money spent with a high number of previous visits made.

We continue by performing a Principal Component Analysis (PCA) to identify the important variables. Indeed, PCA is a dimensionality-reduction method used to reduce the dimension of the data into a small data still containing most of the information in order tote the analysis of the data.

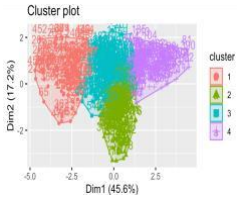


After performing the analysis, 4 variables were considered as most contributing variables: Net_Sales, Gross_Sales, Previous_spent, Items, Previous_Visits.
Continuing with our analysis, using the levels of the categorical variables, we group the data and observe the different clusters and trends forming in the data using the statistics from the K-means clustering. We focus on the main characteristics of the main clusters. We start off by analyzing the results for variable: Loyalty member.



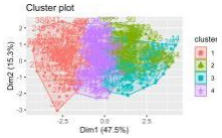
Based on the above plot and summary statistic, cluster 1 which accounts for 19% of the data contains the highest number of purchased items, net sales, gross sales, previous money spent, and total discount. Cluster2 represents 24% of the data and contains customers who have the lowest number of items and net sales, as well as the smallest amount of money previously spent. Cluster3 represents customers with the highest age and highest previous visits.

We look at Gender (Female) next.



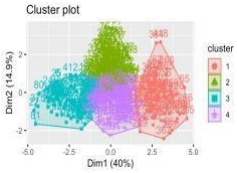
Cluster1 contains customers with the highest number of items purchased, highest net sales and gross sales, and highest amount of previous spent money. Cluster2 has more mature customers with higher number of previously visits made.

Variable: Type of Customer (Promotional) is the next categorical variable we take a closer look at. The results are as follow:



Cluster1 has mature customers who bought the most items, who have highest net sales and gross sales, and who spent the most during previous visits of the store. Cluster4 represents promotional customers with the highest of previous visits.

We continue with the marital status (Married) variable.



With this case, cluster1 contains married customers with the highest number of items bought, highest net sales and gross sales. It also contains customers who spent the most during their previous visits. Cluster2 mainly contains customers with the highest number of previous visits.

It is important to note that with each variable, cluster1 contains the most important variables.

4. LIMITATION

At the end of the unsupervised analysis, the statistical methods used allowed us to find hidden patterns in the data. With the help of clustering, we can understand the variables much better, helping us make careful decisions. Identifying customers is an important step for stores since it helps them make business decisions that target customers based on several parameters like income, age, spending patterns, etc. Despite the results, more work can be done. Indeed, after finding the trends, a model can be put in place to predict a specific outcome, depending on the department focus. More complex patterns can be taken into consideration to improve the findings as well.