Ved Redkar

2024-10-28

1. Introduction

This report analyzes customer spending patterns in the wholesale dataset, utilizing R for descriptive and inferential statistics. It explores customer preferences, identifies high-spending segments, and examines regional purchasing behaviors. Through statistical tests and visualizations, the analysis aims to inform targeted marketing strategies and optimize resource allocation, ultimately enhancing revenue potential for the wholesaler.

2. Exploratory Data Analysis

2.1 Data Preprocessing

A check for Structure of the data

The command str(my.wholesale) was executed in order to check the structure of the wholesale dataset my.wholesale. The checking of the structure has confirmed that integers present key product categories, while regions and channels are represented as factors or categorical variables, which is appropriate for further analysis and preprocessing.

A check for missing values

Count of missing values: 0

2.2 Descriptive Statistics

Summary statistics for numerical variables

Table 1: Summary Statistics of Spending Categories

Statistic	Fresh	Milk	Grocery	Frozen	Detergent	Delicatessen
Min.	23	201	3	36	3.0	3.0
1st Qu.	3408	1608	2062	647	231.8	343.5
Median	9392	3479	3822	1409	634.5	883.0
Mean	12739	5230	7169	2767	2594.1	1207.5
3rd Qu.	16782	6641	8928	3045	3613.2	1576.0
Max.	112151	73498	67298	35009	38102.0	8550.0

There is a huge variability of spending within each category, particularly extremely wide ranges in "Fresh" and "Grocery" items, which support diverse purchasing behavior. Another point to be made is that average spending in "Milk" and "Detergent" is much higher than the median; that might indicate there were a few customers who spent extremely high amounts on those categories.

Region Wise Average Customer Spending

Table 2: Region-wise Customer Spending Comparison

Region	Fresh	Milk	Grocery	Frozen	Detergent	Delicatessen
Lisbon	11102	5486	7403	3000	2651	1355
Porto	9888	5088	9219	4045	3687	1160
Rest Of Portugal	12533	5977	7896	2945	2818	1621

It is indicative in the data that the average spending of Rest of Portugal customers is the highest in a majority of the categories, especially in "Fresh" and "Delicatessen," which evidence a focus on fresh and specialty products. However, in Porto, the spending on "Frozen" and "Detergents" is higher, possibly due to different consumption patterns or business needs.

Regional Distribution of Customers by Channel

Table 3: Regional Distribution of Customers by Channel

	Hotel/Restaurant/Cafe	Retail
Lisbon	31	8
Porto	12	6
Rest of Portugal	100	43

The segment "Rest of Portugal" leads, with 100 Hotel/Restaurant/Cafe and 43 Retail customers, showing good reaches even outside main cities. Lisbon follows with 31 Hotel/Restaurant/Cafe and 8 Retail customers. Porto has the least, with 12 in Hotel/Restaurant/Cafe and 6 in Retail, indicating regional focuses in some sectors.

2.3 Data Visualization

2.3.1 Univariate Analysis

Histograms for numerical variables

Right-skewed distributions: The following spending occurs in each category: Fresh, Milk, Grocery, Frozen, Detergent, and Delicatessen. They are all right-skewed, indicating that the majority of their customers are spending at the lowest range of each category.

Category-Specific Spending Patterns Spending on fresh and grocery categories is mostly clustered in low values but spikes extremely high, to a value as high as 90,000 euros in the Fresh category and as high as 60,000 euros in the Grocery category.*(Refer to figure 1)*

Figure 1 – Distribution of Spending Categories

2.3.2 Bivariate Analysis

Correlation Matrix of Spending Categories

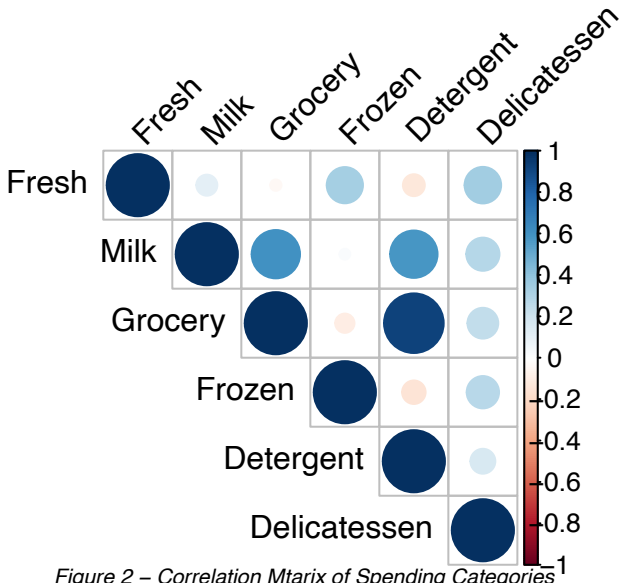


Figure 2 – Correlation Matrix of Spending Categories

There is a very strong positive correlation between Groceries & Detergent, indicating that when customers spend more on grocery items, they also tend to spend more on detergents, and vice versa.

2.4 Discussion

1. Spending variability is huge, especially in "Fresh" and "Grocery" categories, which indicates diverse customer needs and also the presence of outliers with high expenditures.
2. The "Rest of Portugal" has the biggest market share and is leading in the "Fresh" and "Milk" purchases, followed by Lisbon. The smallest is Porto; hence, it can provide an expansion opportunity.
3. In the HoReCa sector, the wholesaler is well-represented in Lisbon but under-retailed, which would suggest some potential for growth. "Rest of Portugal" also has both HoReCa and retail well-served, but there is again some potential in terms of retail growth across the regions.
4. Data for each of the spend categories is right-skewed, suggesting that subsequent analysis using probability distributions such as log-normal or gamma may be appropriate.
5. It also shows that the most important features of the later discussed logistic regression model are highly correlated since the occurrence of independent variables that are highly correlated could lead to problems of multicollinearity, thus impacting model accuracy.

3. Probablity Distribution Fit

The study of the variable "Fresh" will be of great relevance for probability distribution analysis, as it presents high variability in customer spending and a great market share in the "Rest of Portugal." This segment is directly linked to revenues and, simultaneously, customer satisfaction, hence may be an object of targeted marketing or inventory optimization. Since the data is right-skewed, the probability distributions would be log-normal or gamma. This would give more accuracy to the forecast and hence enable strategic decision-making in this matter. Then we plot a box to analyze the disparities in fresh product spending between the two channels, through which we can understand customer behavior and spending trends better.

Average Spending by Channel

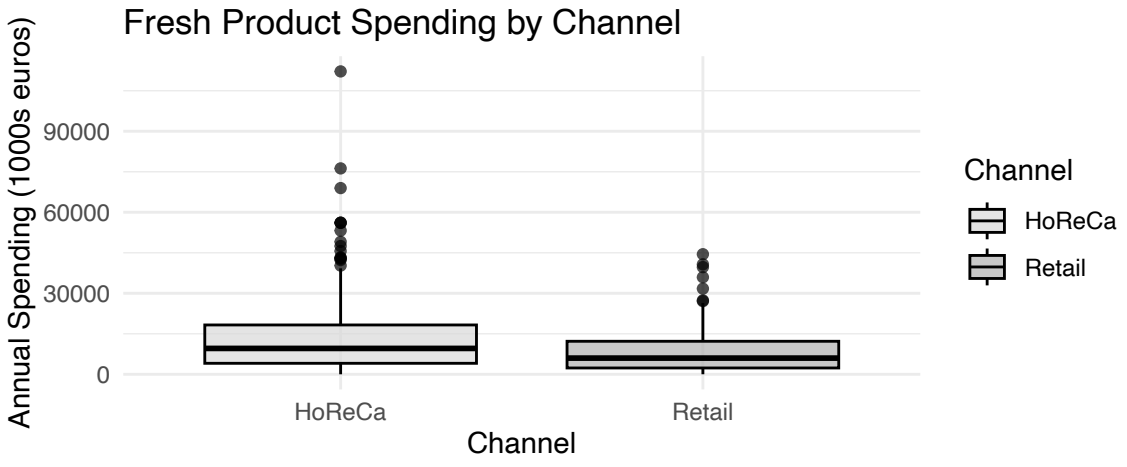


Figure 3 – Box Plot for Fresh spending by Channel

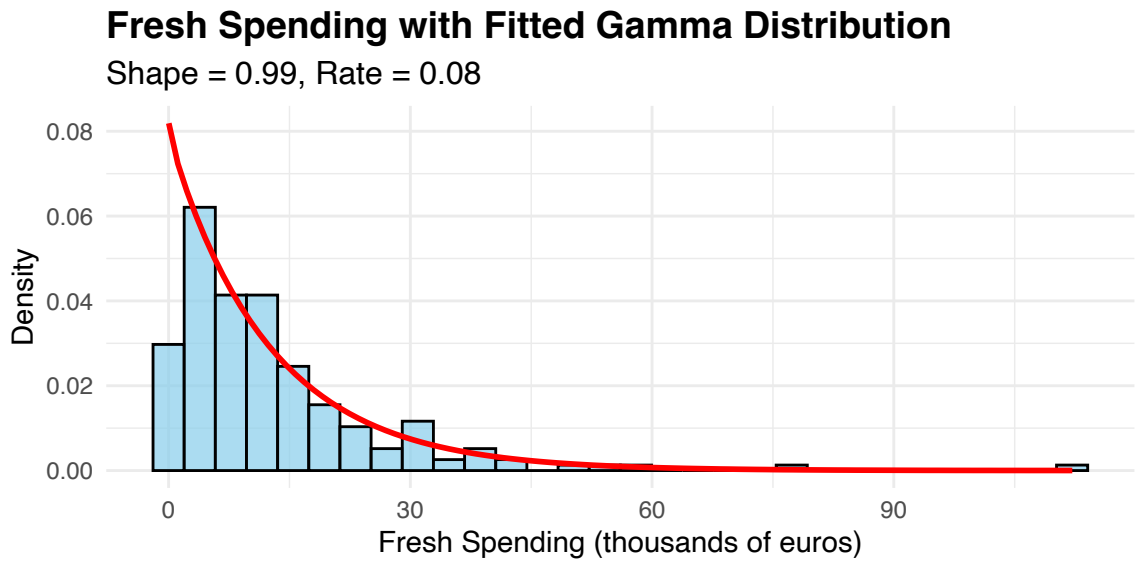
This plot shows the variability in the expenditure of fresh products via two sales channels. Channel 1 has higher variability and a higher median compared with Channel 2. Outliers were contained in both channels,

while Channel 1 contained some customers who spent almost 90,000 euros. In summary, Channel 1 is the most dominant channel with respect to customer spending on fresh product sales. The variability and outliers in Channel 1's spending distribution is insightful and makes a compelling case for considering a gamma distribution.

Below are the results for gamma distribution: Using probability calculations, the wholesaler can predict the likelihood of customers exceeding specific spending thresholds, enabling better inventory management and targeted marketing strategies for high-spending segments.

```
## Summary of Fresh values (in thousands of euros):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.023   3.408   9.393  12.739  16.782 112.151
##
## Gamma Distribution Parameters:
##      shape       rate
## 0.99068961 0.07776781
##
## Standard Errors:
##       shape        rate
## 0.087149931 0.008792423
```

```
##
## Comparison of Actual vs Theoretical Quantiles (in thousands of euros):
##     Probability   Actual Theoretical Difference
## 25%        0.25  3.40775    3.629904 -0.2221536
## 50%        0.50  9.39250    8.797180  0.5953202
## 75%        0.75 16.78150   17.660915 -0.8794152
## 90%        0.90 29.64180   29.400472  0.2413282
## 95%        0.95 37.12385   38.289293 -1.1654432
##
## Example probability calculations:
## Probability of spending more than 10 thousand euros: 0.455
## Probability of spending more than 20 thousand euros: 0.208
## Probability of spending more than 30 thousand euros: 0.095
```



Rationale for Employing the Gamma Distribution

It would be appropriate to model the spending of Fresh products using the Gamma distribution since this model has a sharp peak near zero and a long tail reflecting the right-skewed nature of this dataset. Unlike the log-normal distribution starting from zero with a rate parameter of 0.08, it does not misestimate the low values.

4. Channel Prediction

The channel prediction analysis applied logistic regression to identify factors influencing customer type/channel based on spending, using a training set of 161 observations and a test set of 39.

Model Summary

Formula: Channel=Fresh+Milk+Grocery+Frozen+Detergent+Delicatessen Family: Binomial

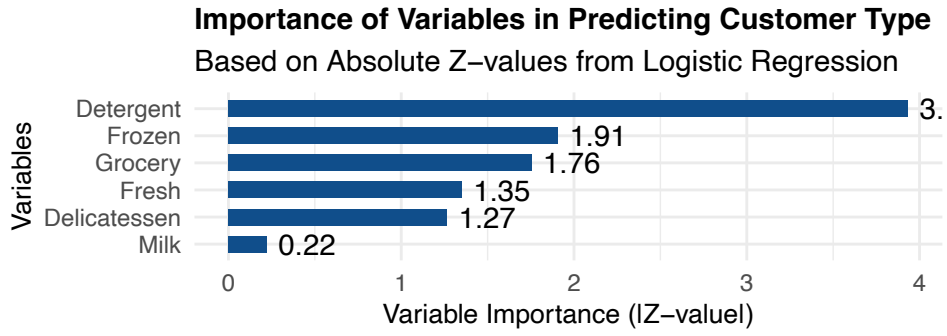


Figure 4 – Variable Importance Plot

Detergent spending has the largest coefficient of 0.001711, with a standard error of 0.0004347. This gives a z-value of 3.935, which is very high and means a very strong relationship between Detergent spending and customer channel likelihood. The associated p-value is 0.0000831, which is below the 0.05 significance level, so detergent spending is significant to the model.

Then we perform logistic regression on the significant variables

We could see that the previous model was with all the variables as dependent variables and showed an accuracy of 87.18%. However, developing the model with only the most significant variable, Detergent, massively improved the accuracy to 89.74%.

```
##
## Call:
## glm(formula = Channel ~ Detergent, family = "binomial", data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.741174   0.782082  -6.062 1.34e-09 ***
## Detergent    0.001484   0.000250   5.938 2.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 192.643  on 160  degrees of freedom
## Residual deviance:  51.111  on 159  degrees of freedom
## AIC: 55.111
##
## Number of Fisher Scoring iterations: 8
```

```
## [1] "Test Data Confusion Matrix (Single Variable Model):"
```

```
##           Actual
## Predicted  1  2
##         1 27  3
##         2  1  8
```

```
## [1] "Test Data Accuracy (Single Variable Model): 89.74 %"
```

The logistic regression analysis shows that there is a significant relationship between Detergent spending and customer channel classification. The coefficient is positive, 0.001484; that means with higher spend, it is more likely to become a member of a certain channel. The intercept of -4.741 and the Detergent coefficient were significant based on z values of -6.062 and 5.938, respectively, with p-values less than 0.001. The deviance, from the model, reduced from 192.643 to 51.111, representing the null deviance and the residual deviance, respectively. This ensured that the model was well fitted. In contrast, using the confusion matrix, the accuracy of the model stood at 89.74%, where out of the 30 instances in channel 1 and 9 in channel 2, only four were misclassified and thus classified 27 and 8 correctly. These results, in other words, ensure that Detergent spending is a very powerful variable in identifying customer channels.

5. Regional Spending Analysis

Summary of Fresh Spending by Region

```
## # A tibble: 3 x 4
##   Region   Mean     SD Count
##   <fct>   <dbl>  <dbl> <int>
## 1 1      11800. 13400.    39
## 2 2      12270.  7216.    18
## 3 3      13054. 14516.   143
```

ANOVA Results

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Region        2 5.252e+07  26261744   0.137  0.872
## Residuals   197 3.763e+10 191024855
```

Although ANOVA was conducted, its important to make sure the assumptions of ANOVA are first met, in which this case it is not. One-way ANOVA assumes normality in each region and homogeneity of variances.

Shapiro-Wilk test: The test shows significant departures from normality for all three regions, with p-values below 0.05. Thus, the ANOVA **assumption of normally distributed residuals is not met.**

Bartlett's Test: The test yields a p-value of 0.005638, indicating sufficient evidence to reject the null hypothesis of equal variances. Thus, the **assumption of homogeneity of variances is not satisfied.**

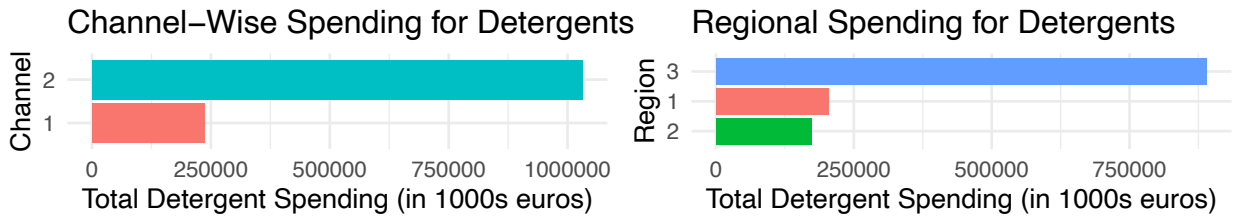


Figure 5 – Distribution of Detergent Spending by Channel and Region

Pairwise T-test

From the pairwise t-tests with Bonferroni-adjusted p-values, we can interpret the following comparisons among regions based on their grocery spending:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  my.wholesale$Fresh and my.wholesale$Region
##
##   1 2
## 2 1 -
## 3 1 1
##
## P value adjustment method: bonferroni
```

Matrix Results:

The matrix displays pairwise comparisons between the regions: Region 1 vs. Region 2: No result (indicated by -), suggesting either no significant difference or that this comparison is not applicable. Region 1 vs. Region 3: There is a significant difference (indicated by 1). Region 2 vs. Region 3: There is a significant difference (indicated by 1).

In summary, the pairwise t-tests with Bonferroni adjustments show no significant differences in Fresh spending across regions, suggesting any observed differences are minimal or consistent.

6. Conclusion

Targeted Marketing and Inventory: Large variability in fresh product spending points to the need for a targeted marketing approach. By basing their analysis on probability distribution, the wholesaler will be able to more accurately predict the spending of its customer segments and adjust inventory levels to meet the needs of high-spending segments, in particular.

Key Product Focus: Detergent spend is strongly related to customer channel type, and therefore targeted promotions in detergent could drive acquisition and retention within key channels to increase overall spend.

Region-Specific Strategies: Regional spending differences, suggest that rest of Portugal has a propensity for higher spend on fresh products and would, therefore, be the best area in which focused promotions could be used to maximize revenue.

Spending Probability for Forecasting: With a 45.5% probability of customers spending over 10,000 euros annually on fresh products, the wholesaler can better plan inventory and promotions, targeting high-value customers and enhancing forecasting accuracy.