Progress Report - Wholesale Customer Analysis

Johnny Lee

2025-05-01

Dataset Description

The dataset contains information from a wholesale distributor in Portugal, which captures the annual purchasing behaviour across 440 customers. It includes 8 variables **Channel**, **Region**, **Fresh**, **Milk**, **Grocery**, **Frozen**, **Detergents_Paper**, and **Delicassen**.

- **Channel**: Customer type (1 = Hotels/Restaurants/Cafes, 2 = Retailers like grocery stores)
- **Region**: Customer location (1 = Lisbon, 2 = Oporto, 3 = Other regions in Portugal)

The remaining 6 variables represent the annual spending of each customer (in monetary units) on each category (**Fresh**, **Milk**, **Grocery**, **Frozen**, **Detergents_Paper**, and **Delicassen**)

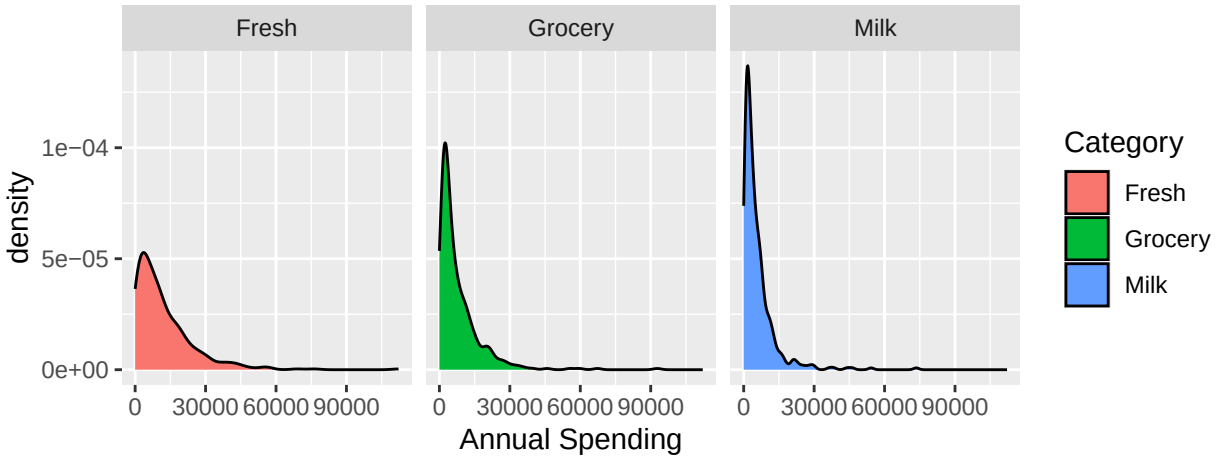
Exploratory Analysis

First, we examined the summary statistics of the annual spending variables:

- **Medians**: Highest - **Fresh** (8504), Lowest - **Delicassen** (1524.9)
- **Means**: Highest - **Fresh** (12000), Lowest - **Detergents_Paper** (816.5)
- **Bounds**: Max - **Fresh** (112151), Min - **Fresh**, **Grocery**, **Detergents_Paper**, **Delicassen** (3)
- **IQR**: Largest - **Fresh** (13806), Smallest - **Delicassen** (1412)

The summary statistics suggest that all annual spending variables are right skewed, as the means are larger than the medians. Overall, customers spend the most on **Fresh** foods and the least on **Delicassen** foods.

Next, we examined the density plots of the annual spending to further verify their right-skewed distributions. We generated plots for the overall data, as well as stratified by region, channel, and all region/channel combinations. Below are 3 density plots for the **Fresh**, **Grocery**, and **Milk** variables, which shows their right skewness.

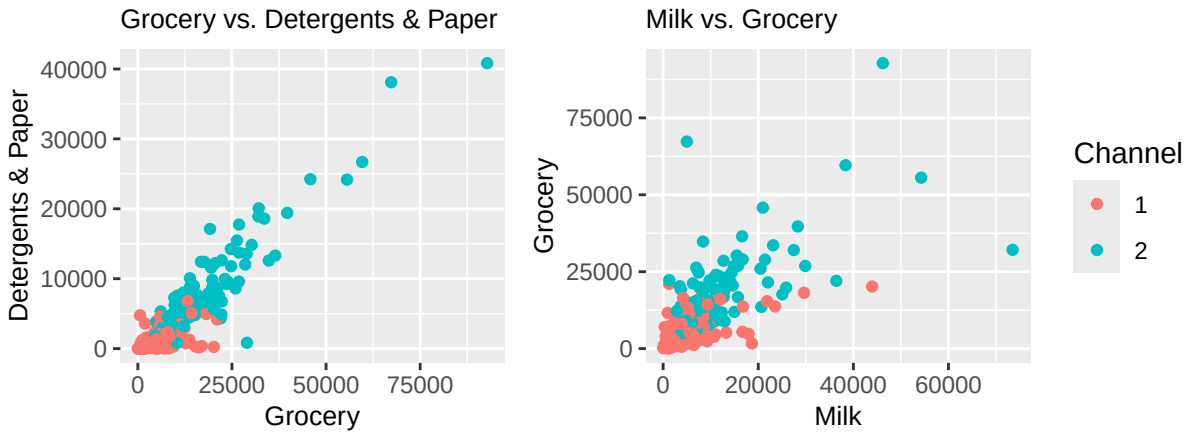


Afterward, we analyzed the Spearman correlations between the annual spending variables. Below are the moderate to strong correlations ($\rho \geq |0.3|$):

- **Frozen** and **Fresh** ($\rho = 0.38$), **Grocery** and **Milk** ($\rho = 0.77$) **Milk** and **Detergents_Paper** ($\rho = 0.68$),
- **Milk** and **Delicassen** ($\rho = 0.37$), **Grocery** and **Detergents_Paper** ($\rho = 0.80$),
- **Grocery** and **Delicassen** ($\rho = 0.30$), **Grocery** and **Detergents_Paper** for Channel 2 ($\rho = 0.85$)

The correlations suggest that customers tend to purchase **Milk**, **Grocery**, and **Detergent_Paper** together, as these categories are all strongly correlated. Additionally, the moderate correlation between **Fresh** and **Frozen** suggests customers purchase these items together.

To visualize these correlations, we generated scatter-plots for variable pairs where $\rho \geq |0.3|$. Below, we highlight the two most strongly correlated pairs.



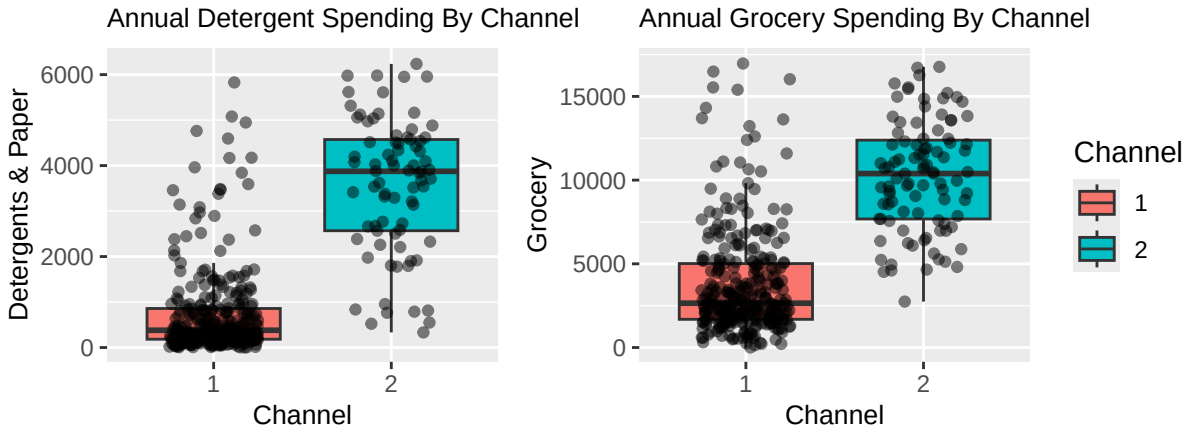
The scatter-plots reveal a strong linear relationship between **Grocery** and **Detergents_Paper** and a moderate association between **Grocery** and **Milk**, reinforcing the trends that were identified from the correlation analysis.

Hypothesis Testing:

Given the non-normality of the data, we used non-parametric hypothesis testing.

- **Regional Differences**: Kruskal-Wallis test were conducted for all six annual spending categories and total annual spending (the sum of all categories). No statistically significant differences in medians were found ($\alpha = 0.05$).
- **Channel Differences**: Mann-Whitney U tests revealed statistically significant differences in median spending between the two channels for all individual categories and for total annual spending ($\alpha = 0.05$), indicating distinct purchasing patterns between the channels.
- **Channel-Region Association**: A Chi-squared test found no statistically significant dependence between **Channel** and **Region** ($\alpha = 0.05$), suggesting these two variables are independent.

To visualize the difference in median spending between the two channels, we generated box-plots. Below, we highlight the two most statistically significant results from the Mann-Whitney U test.



The box-plots reinforce the Mann-Whitney U test results, showing a large difference in medians between the two channels.

Model Plans

Given the strong pairwise correlations and linear relationships among **Grocery**, **Detergents_Paper**, and **Milk**, we will use regularized regression (ridge or lasso) to model annual spending in these categories, addressing the multicollinearity between them.

- **Primary Model**: Predict **Grocery** spending using **Detergents_Paper**, **Milk**, and additional predictors (other spending categories, interaction terms, ratio terms, **Channel**, and **Region**)
- **Secondary Models**: Predict **Milk** and **Detergents_Paper** spending using the same framework as the **Primary Model**

Given the statistically significant differences in median spending across all spending categories between the two channels, we will create a random forest classifier to predict the purchasing channel. The model will incorporate all individual spending categories, total annual spending, interaction terms, and ratio terms.