Wholesale Customer Dataset Evaluation

W. Kerney

Introduction

The UCI Wholesale customers dataset is provided by the University of California at Irvine, it refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories. What we are doing today is using this data to analyze some customer spending patterns and derive some insights.

Step 1: Load the Dataset

First, we load the "Wholesale customers" dataset into R. This dataset has information on annual spending in various product categories by different customers. You can download it from Kaggle:

```r
# Install and load necessary packages
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("scales")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("knitr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
library(readr)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(knitr)
# Load the dataset
wholesale_data <- read_csv("Wholesale customers data.csv")
```

```
## Rows: 440 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, De...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(wholesale_data)  # Check the first few rows
```

```
## # A tibble: 6 x 8
##   Channel Region Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
##     <dbl>  <dbl> <dbl> <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
## 1       2      3 12669  9656    7561    214             2674       1338
## 2       2      3  7057  9810    9568   1762             3293       1776
## 3       2      3  6353  8808    7684   2405             3516       7844
## 4       1      3 13265  1196    4221   6404              507       1788
## 5       2      3 22615  5410    7198   3915             1777       5185
## 6       2      3  9413  8259    5126    666             1795       1451
```

Step 2: Data Cleaning and Transformation

Next, let's clean the data to ensure it is ready for analysis:

```
# Check for missing values
sum(is.na(wholesale_data))  # Check for any NA values
```

```
## [1] 0
```

```
# Transform data if necessary
# In this dataset, ensure relevant columns are numeric
wholesale_data$Channel <- as.factor(wholesale_data$Channel)
wholesale_data$Region <- as.factor(wholesale_data$Region)

# Create a new variable for total annual spending
wholesale_data$TotalSpending <- rowSums(wholesale_data[, c("Fresh", "Milk", "Grocery", "Frozen", "Deter
```

Step 3: Data Aggregation

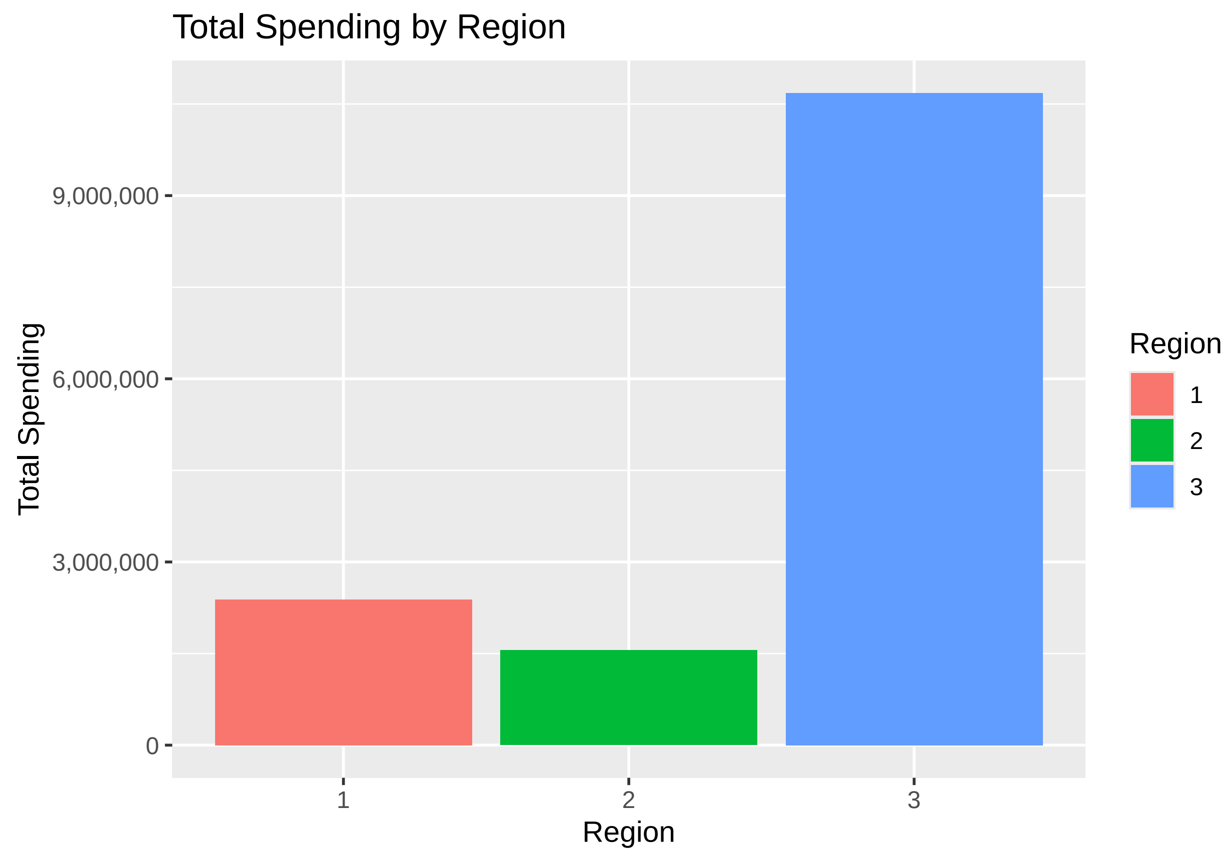
Now, let's aggregate the data to get insights into customer spending patterns:

```
# Aggregate total spending by region
spending_by_region <- aggregate(TotalSpending ~ Region, data = wholesale_data, sum)

# Aggregate total spending by channel
spending_by_channel <- aggregate(TotalSpending ~ Channel, data = wholesale_data, sum)

# Get the average spending per category
average_spending_per_category <- sapply(wholesale_data[, c("Fresh", "Milk", "Grocery", "Frozen", "Deter
```

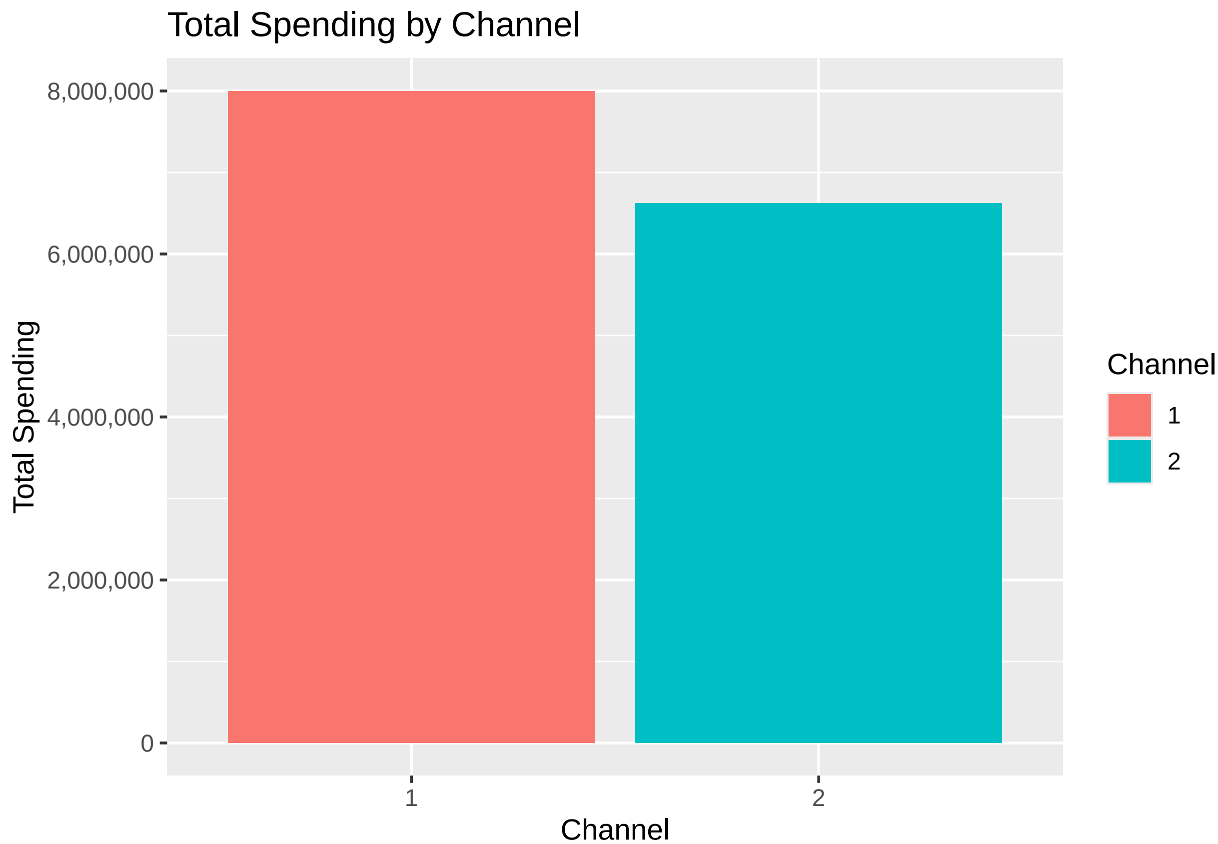
Step 4: Data Visualization

Then we will use ggplot2 to create visualizations to better understand the data:

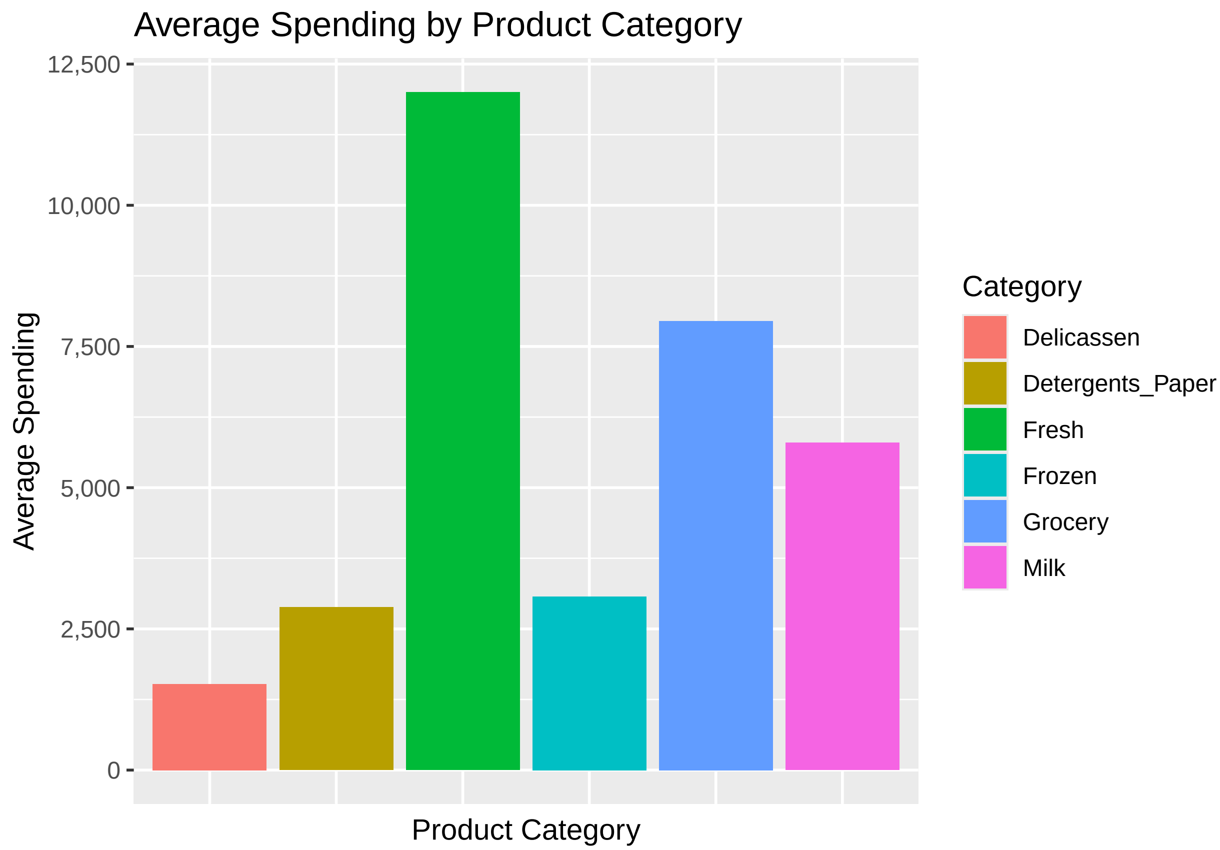
```
# Bar plot of total spending by region
ggplot(spending_by_region, aes(x = Region, y = TotalSpending, fill = Region)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Spending by Region", x = "Region", y = "Total Spending")+
  scale_y_continuous(labels = comma)
```




```r
# Bar plot of total spending by channel
ggplot(spending_by_channel, aes(x = Channel, y = TotalSpending, fill = Channel)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Spending by Channel", x = "Channel", y = "Total Spending")+
  scale_y_continuous(labels = comma)
```




```r
# Bar plot of average spending by product category
average_spending_df <- data.frame(Category = names(average_spending_per_category), AverageSpending = ave

ggplot(average_spending_df, aes(x = Category, y = AverageSpending, fill = Category)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Spending by Product Category", x = "Product Category", y = "Average Spending") +
   theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  scale_y_continuous(labels = comma)
```



Conclusion

Looking at the first plot, we can see that there is a substantial amount more sales in Region 3, classified as “other” in the dataset. Meaning we should probably focus our marketing effort more on the Lisnon and Oporto regions in an attempt raise sales in those regions. What we do not know is how many regions are being initially aggregated to form the “other” category, we may want to investigate that further to understand more about how individual regions are performing. Looking at the second chart, we see that there is about 20% more sales via Channel 1 (hotel/restaurant/cafe) than Channel 2 (retail). Investigating this further we can see that Channel 1 is aggregating hotel, restaurant and cafe sales into one category so that may be a reason why the sales for that channel is higher in aggregate sales. The third chart is showing us that our average fresh sales are significantly more than any other category of product that we sell. Delicatessen products have the least average sales by a significant amount. Delicatessen product sales are at about 1000 average sales units, while fresh product sales are at about 11,500 average sales units. Using this data I would recommend that the company focus on delicatessen, detergent, and frozen product sales in their marketing efforts in an attempt to raise those sales to at least the average.

Data Citation Cardoso,Margarida. (2014). Wholesale customers. UCI Machine Learning Repository. https://doi.org/10.24432/C5030X.