University of Cape Town

STA50697Z

Multivariate Statistics

Topic Ideas

Author:
Roger Bukuru

Student Number:
BKRROG001

February 25, 2024

Contents

1 Introduction

We are tasked to perform topic idea analysis. The aim is to obtain 2-3 data sets, from a specific application area. Using these data sets we will describe a research question and perform various exploratory data analysis techniques. Upon completion of the above analysis, one data set will be selected to implement and analyze various multivariate statistics techniques within the application area of the said data set.

For this analysis, we will be analyzing 2 data sets within the following application areas **Online E-Commerce** and **Banking**. The data sets that we will be analyzing consist of the following

- Online Retail Data
- Default of credit card clients

For a majority of these data sets the aim will be to research multivariate techniques that can help one achieve the following various business goals:

- Customer Segmentation
- Product Analysis
- Customer Recommendations

2 Online Retail Data

The Online Retail data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers [2].

2.1 Research Question

For this data set our research question of interest can be broken into two parts:

1. What products are often purchased together, with the business objective of assessing if cross-selling opportunities exist?
2. What are the trends of top wholesalers and how can they be targeted more effectively, essentially analyzing various customer segments?

2.2 Exploratory Data Analysis

The data consists of 8 feature variables, these features are described in Table 1 below:

Features	**Description**
Invoice	Transaction Invoice Number
StockCode	Unique gift-ware product stock code
Quantity	The number of units purchased on a transaction
InvoiceDate	The date an invoice was issued
Customer ID	Unique wholesaler customer id
Country	The country where the customer operates

Table 1: Online Retail Data: Features

Analysing the data set we observed the following exploratory results:

1. There were a total of 4631 unique gift-ware items
2. Across the 2 years, there were a total of 28816 sales across 4385 customers, **about 5229 sales do not have an associated customer id.**

Product Analysis

Figure 1 highlights the top 20 gift-ware sold, we observe that giftware with stock code 21212 (PACK OF 72 RETRO SPOT CAKE CASES) was the most sold giftware with almost 60,000 units sold over the 2 years.

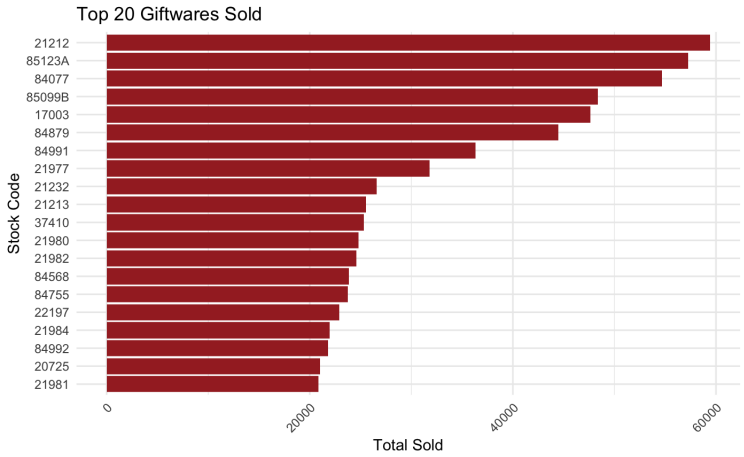

Figure 1: Top 20 Gift-Wares Sold

Customer and Sales Analysis

From the data, we observe that there was a total of 4384 unique customers, the customers are distributed across 40 countries, and Figure 2 shows the top 5 countries where customers originate from.

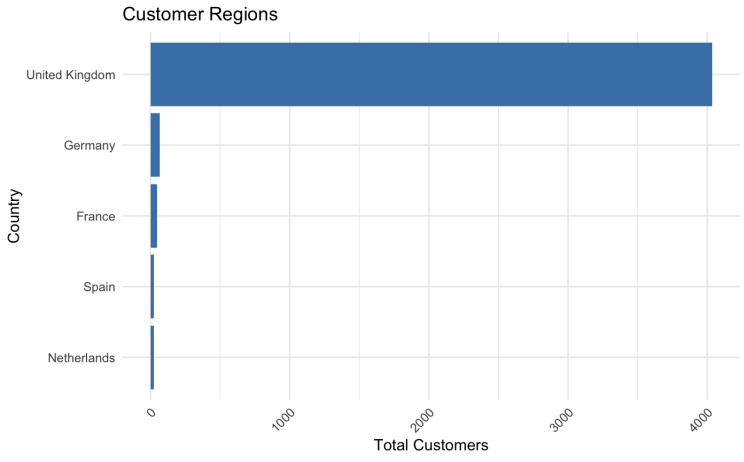

Figure 2: Customer Regions

We note that the country with the majority of wholesalers is the United Kingdom with over 4000 wholesalers.

Assessing revenues generated, figure 3 shows the top revenue-generating customers.

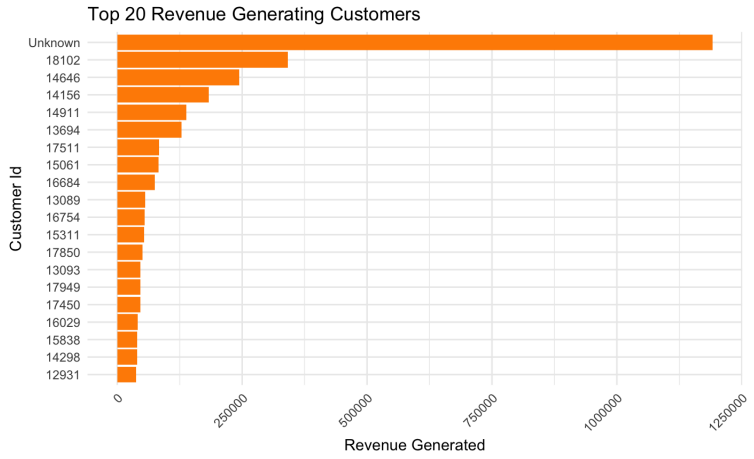

Figure 3: Caption

We observe that the customer ID(s) of the top performing customer(s) is unknown, this is concerning, as this customer cannot be targeted, say example for promotions, etc. We also observed that this accounts for 5229 sales (invoices) that do not have associated customers.

3 Default of credit card clients

The dataset contains records of credit card clients in Taiwan, focusing on the prediction of default payments. With 23 explanatory variables ranging from demographic information (like gender, education, marital status, and age) to financial behaviors (including payment history, bill statement amounts, and previous payment amounts), the dataset provides a comprehensive view of factors that could influence credit card default risk[1].

3.1 Research Question

For this data set we wish to explore what factors are more likely to affect a customer defaulting e.g. are younger customers more likely to default than older clients? With this ability, the bank would be able to more effectively cluster various customer risk profiles.

3.2 Exploratory Data Analysis

We start the exploratory analysis by reviewing the feature variables that form part of the data set, this is shown in table 2 below.

Feature	**Type**	**Description**
Limit Balance	Quantitative	A customers credit limit
Sex	Categorical	1 - Male; 2 - Female
Education	Categorical	1 = graduate school; 2 = university; 3 = high school; 4 = others
Marriage	Categorical	1 - Married; 2 - Single; 3 - Other
Age	Quantitative	Customer age
Pay0 - Pay11	Quantitative	History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: Pay0 = the repayment status in September 2005; Pay2 = the repayment status in August 2005; . . .; Pay6 = the repayment status in April 2005. The measurement scale for the repayment status is -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
BillAmt1 - BillAmt6	Quantitative	Amount of bill statement (NT dollar). BillAmt1 = amount of bill statement in September 2005; BillAmt2 = amount of bill statement in August 2005; . . .; BillAmt6 = amount of bill statement in April 2005.
PayAmt1 - PayAmt6	Quantitative	Amount of previous payment (NT dollar). PayAmt1 = amount paid in September, 2005; PayAmt2 = amount paid in August, 2005; . . .; PayAmt6 = amount paid in April 2005.

Table 2: Default Credit Card Clients: Features

Credit Limit Overwiew

We performed a 5-number summary shown in Figure 4, and we observed the following.

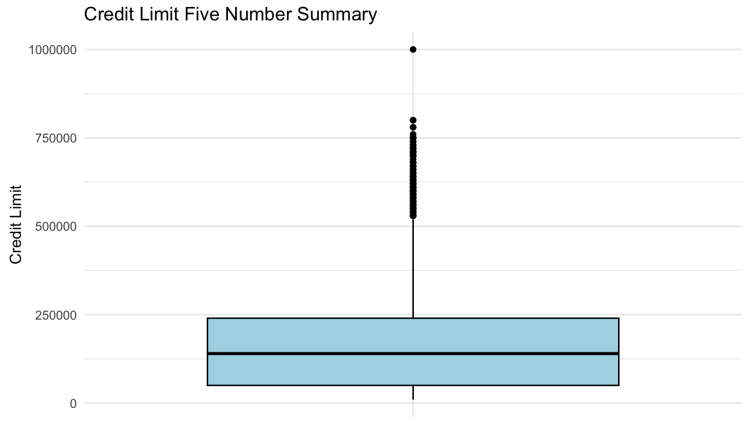

Figure 4: Credit Limit: Five Number Summary

- The minimum credit limit balance was NT$ 10 000
- 25% of customers had credit limit balances of NT$ 50 000 or less, indicating that 75% of clients had credit limit balances above NT$ 50,000.
- The median credit limit was NT$ 140 000, indicating that 50% of customers have credit limit balances less than NT$ 140 000 whilst another 50% of customers have credit limit balances above NT$ 140 000.
- 75% of customers had credit limit balances below NT$ 240 000, whilst 25% of customers had credit limit balances above NT$ 240 000
- The maximum credit limit amount was NT$ 1 000 000
- The average male credit limit balance was NT$ 163519.82 and the average female credit limit balance was NT$ 170086.46

The summary above suggests the following; there is significant variability in the credit limit balances that are given to customers as the limits range from NT$ 10 000 to NT$ 1 000 000. Furthermore, we note that the median value is closer to the third quantile than it is to the first, indicating that the data is right-skewed, which means that there is a large number of customers who have credit limits on the lower end of the range, with fewer individuals having higher credit limits. The significant

difference between the maximum value and the third quantile finally indicates that outliers exist on the higher end of the credit limit balances.

Credit Utilization Patterns

Analysing overall credit utilization we preliminary noted the following:

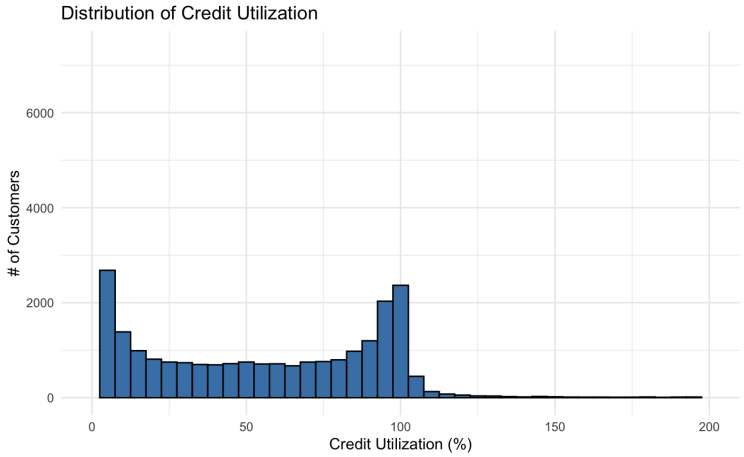

Figure 5: Credit Utilization

- There are a few customers that have exceeded their credit limit facility. Indicating several default risk customers.
- Overall the distribution is left-skewed, indicating that more customers have utilized a large portion of their credit facility than less. This could indicate that several customers could be at risk of defaulting as they owe more money.

4 Conclusion

From the above exploratory analysis and the stated research questions, we expect the implementation of multivariate techniques to include but not limited to the following techniques:

- Clustering: For customer segmentation
- Principal Component Analysis: To assist with mapping factors that most contribute to a customer defaulting

References

[1] Default credit card clients, 2019.

[2] Online retail data, 2019.