Topic Ideas

Roger Bukuru

2024-02-24

Business: Online Retail Data

This Online Retail II 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 gift-ware. Many customers of the company are wholesalers.

Research Question

1. **What items can be recommended with a given item ?**
    - What products are often purchased together ?
    - Are there cross-selling oportunities that exists ?

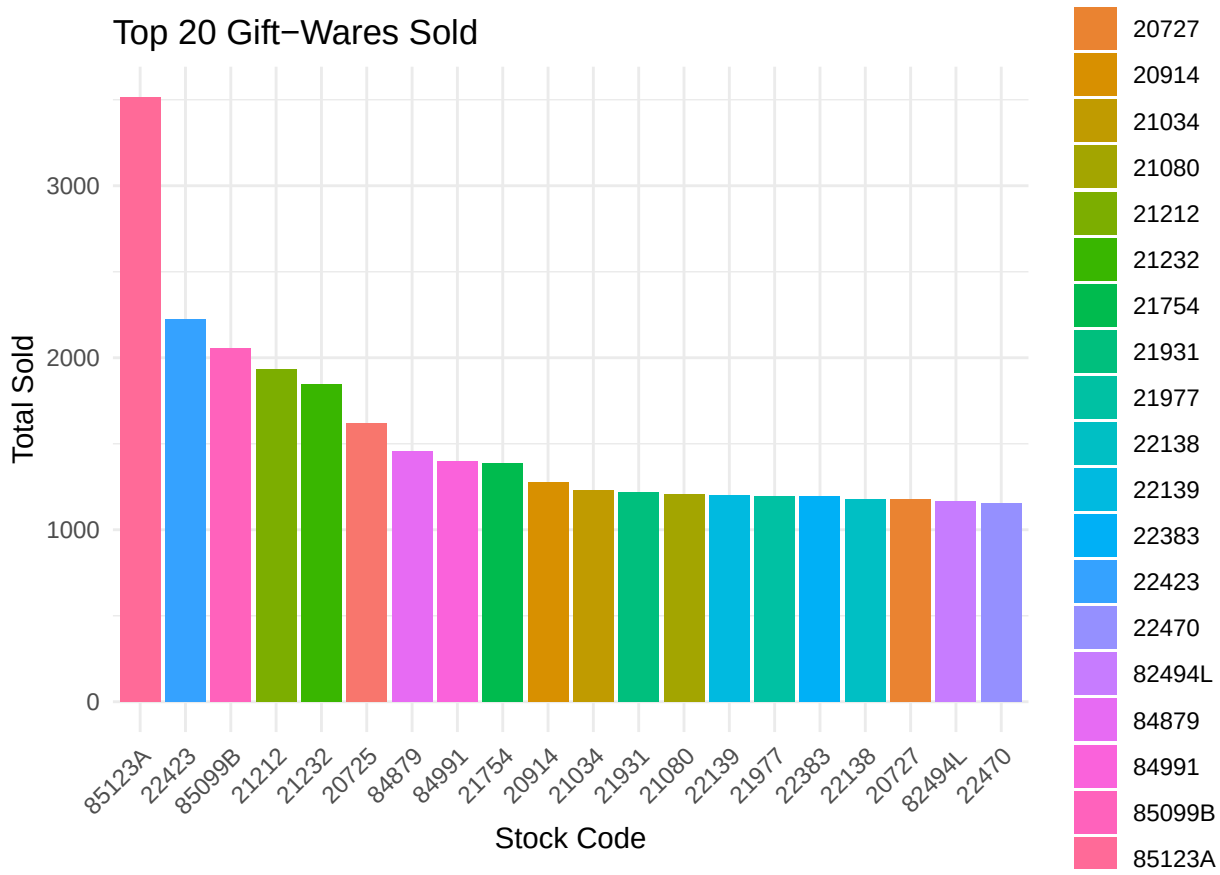
2. **Wholesaler customer retention over time ?**
    - Are the factors that contribute to possible customer churns ?

3. **Can wholesalers be grouped into various customer segements based on their purchasing patterns ?**
    - What are the trends of top wholesalers, how can they be targeted more effectively ?

4. **What gifts were are sold the most ?**
    - Top selling products ?
    - Probability of selling x product ?
    - Identify relationships between the sales of different products ?
    - What was the average basket size ?

Exploratory Data Analysis

There are 4631 unique gift-ware items, the top 20 selling items are as shown below. Gift ware with stock code 85123A was the highest selling item. Across the 2 year period, there was a total of 28816 sales across 4384 wholesalers, about 5229 sales(invoices) do not have associated customers.



The top wholesales

Customer ID	totalInvoices	totalItems	totalValue
NA	5229	107927	840616.41
14911	270	5710	40282.40
14063	13	44	39920.95
15760	5	5	33628.55
12918	3	3	32860.50
14156	138	2710	29888.03
17399	1	1	25111.09
17949	87	104	18585.93
17841	126	5114	15617.09
15202	8	8	14573.16
14606	135	3927	11247.37
12748	159	2665	10703.96
14527	108	1826	9625.22
15413	7	27	9416.82
17850	158	2515	7737.30
15311	158	2226	6436.08
18102	95	635	6408.17
15768	32	1213	6277.08
17017	18	181	6203.75
15480	6	172	5909.58

As seen in the table, it's worrying that the top performing wholesaler is unknown as they contribute to

about 71.78% of the total revenue generated over the 2 year period. This might be one or more different wholesalers.

E-Commerce: Amazon Sale Data

This dataset provides an in-depth look at the profitability of e-commerce sales. It contains data on a variety of sales channels, including Shiprocket and INCREFF, as well as financial information on related expenses and profits. The columns contain data such as SKU codes, design numbers, stock levels, product categories, sizes and colors. In addition to this we have included the MRPs across multiple stores like Ajio MRP , Amazon MRP , Amazon FBA MRP , Flipkart MRP , Limeroad MRP Myntra MRP and PaytmMRP along with other key parameters like amount paid by customer for the purchase , rate per piece for every individual transaction Also we have added transactional parameters like Date of sale months category fulfilledby B2b Status Qty Currency Gross amt . This is a must-have dataset for anyone trying to uncover the profitability of e-commerce sales in today's marketplace

```
file_path = "../Data/Assignment/Amazon Sale Report.csv"
amazon_sales_report = as_tibble(read_csv(file_path))
```

Research Questions

- Customer Segementation, what are various customer profiles
- Product Recommendation: Which products can be recommended together ?
- What are the top predictors for a sale to be fulfilled ?

Exploratry Data Analysis

The data consists of 20 variables, the variables are as described below. It describes sales data from

Feature Exploration

```
##  [1] "index"              "Order ID"           "Date"
##  [4] "Status"             "Fulfilment"         "Sales Channel"
##  [7] "ship-service-level" "Style"              "SKU"
## [10] "Category"           "Size"               "ASIN"
## [13] "Courier Status"     "Qty"                "currency"
## [16] "Amount"             "ship-city"          "ship-state"
## [19] "ship-postal-code"   "ship-country"       "promotion-ids"
## [22] "B2B"                "fulfilled-by"       "Unnamed: 22"
```

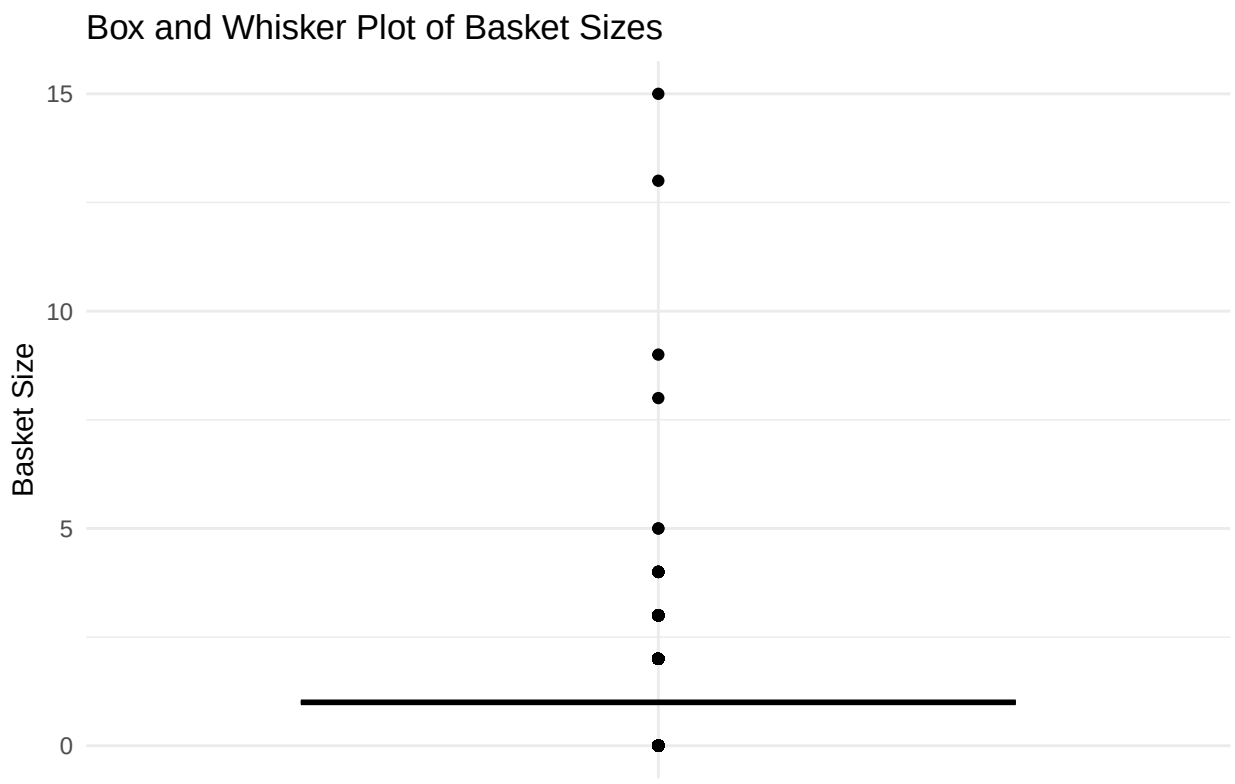
```
## [1] "03-31-22" "06-29-22"
```

Five number summary basket sizes

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

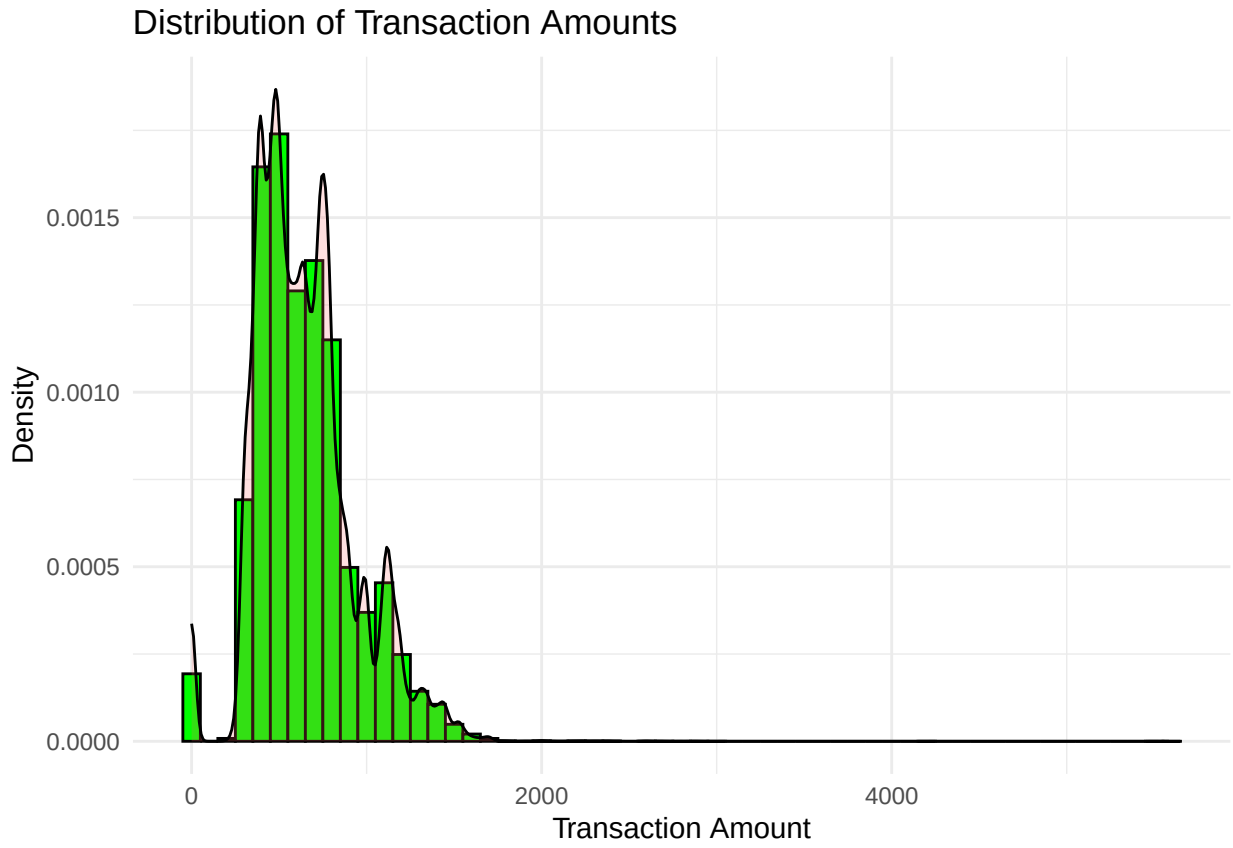
```
## 25%
##   1
```

```
## 50%
##   1

## 75%
##   1

## [1] 15
```



- Orders range from 0 to 15 units, with the average basket size consisting of 1 unit - The smallest quantity ordered in an any transaction is zero, this be be indicative of orders that were cancelled after having been placed. - The first quantile is 1 indicating that 25% of transactions have a basket size of 1 or less. - The median basket size is 1, indicating that half of the transactions have a single item - The third quantile also values matches that of the first quartile, 1, indicating that 75% of transactions have a single item,this shows that a very large majority of transactions consisted of a single item - The largest quantity in a single transaction is 15. Given the high frequency of single item transactions, this is indicative of an outlier(s).

Transaction Amount



- Transaction amounts from 0 to 5584, this indicates that there was a wide range of product prices.
- The average transaction amount was 648.56 INR
- In figure 1 we plot we show the frequency of different transaction amounts, with an estimated distribution overlayed. The distribution of transaction amounts is right-skewed, meaning that most transactions are of a relatively low value, with few high-value transactions.

Order distribution and successful orders ?

- We note the postal codes ranges from 110001 to 989898, indicating nationwide coverage across India.
- There are 2 ship channels namely Amazon.in and Non-Amazon.
- Order lifecyle consists of the following statuses
- About 22.06% of the total sales can be identified with the " " status, indicating that this orders have been successfully sold.

Business: Default of credit card clients

Additional Information

This research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown,

this study presented the novel Sorting Smoothing Method to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables: X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. X2: Gender (1 = male; 2 = female). X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). X4: Marital status (1 = married; 2 = single; 3 = others). X5: Age (year). X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = 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. X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005. X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Research Questions

1. Probability of customer defaulting
2. What other factors are more likely to effect a customer defaulting i.e do younger customers default more then older clients ?

Exploratory Data Analysis

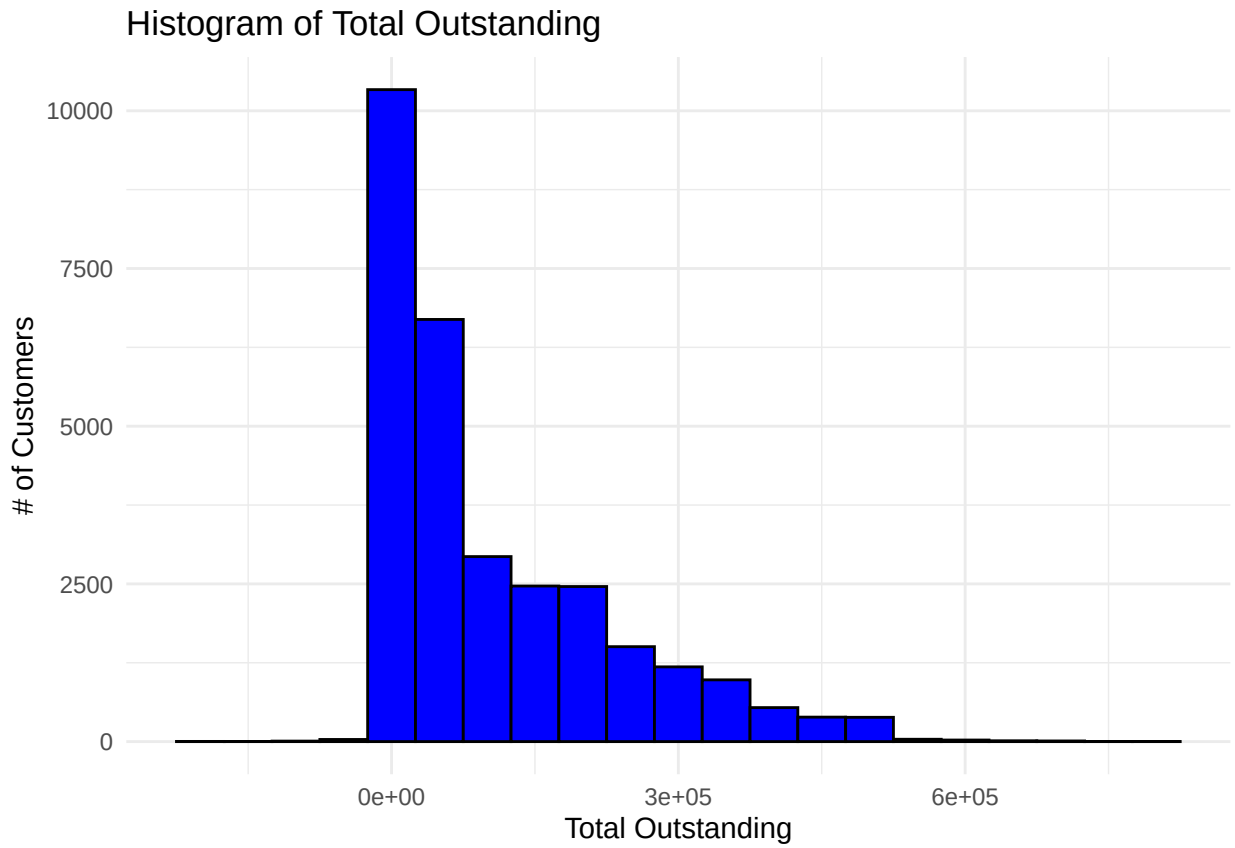
https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

```
##  [1] "ID"                         "LIMIT_BAL"
##  [3] "SEX"                        "EDUCATION"
##  [5] "MARRIAGE"                   "AGE"
##  [7] "PAY_0"                      "PAY_2"
##  [9] "PAY_3"                      "PAY_4"
## [11] "PAY_5"                      "PAY_6"
## [13] "BILL_AMT1"                  "BILL_AMT2"
## [15] "BILL_AMT3"                  "BILL_AMT4"
## [17] "BILL_AMT5"                  "BILL_AMT6"
## [19] "PAY_AMT1"                   "PAY_AMT2"
## [21] "PAY_AMT3"                   "PAY_AMT4"
## [23] "PAY_AMT5"                   "PAY_AMT6"
## [25] "default payment next month"
```

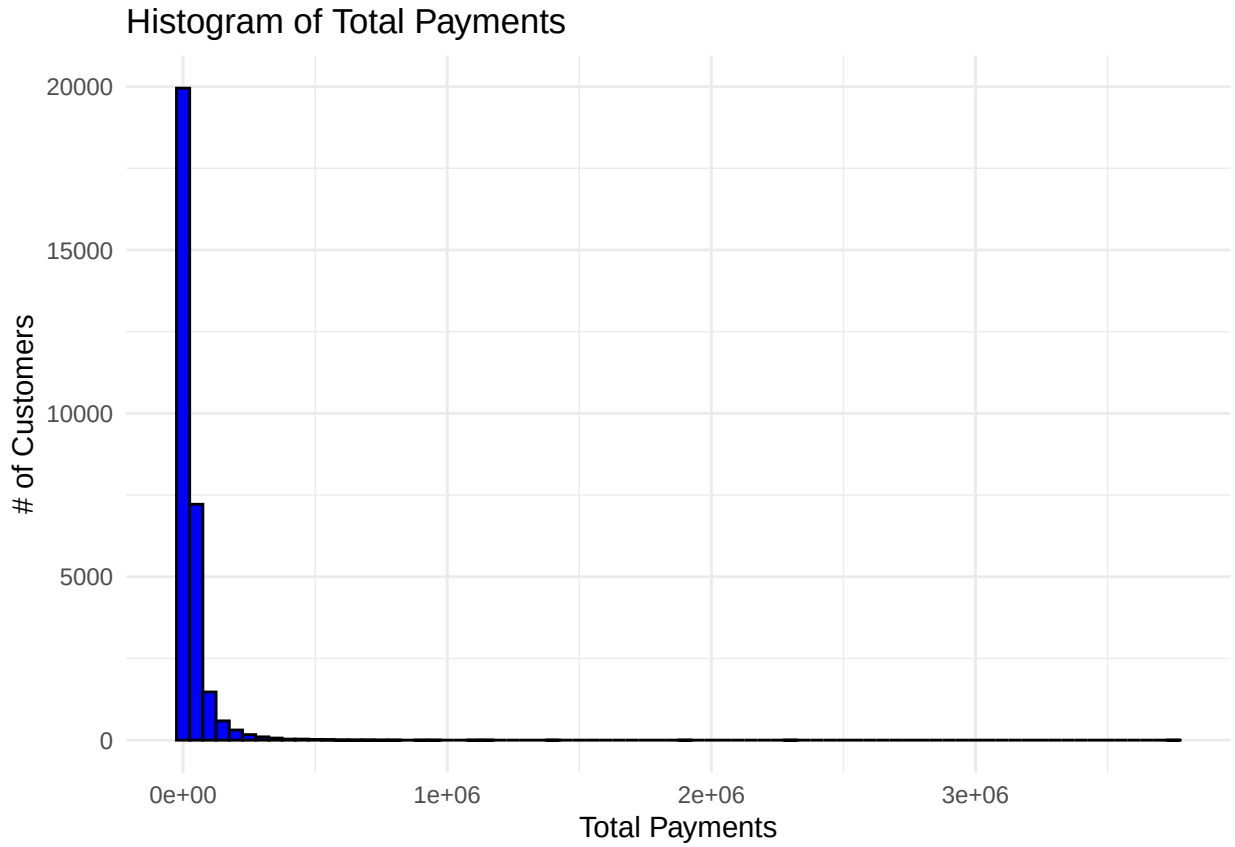
```
## [1] 1e+04 1e+06
```

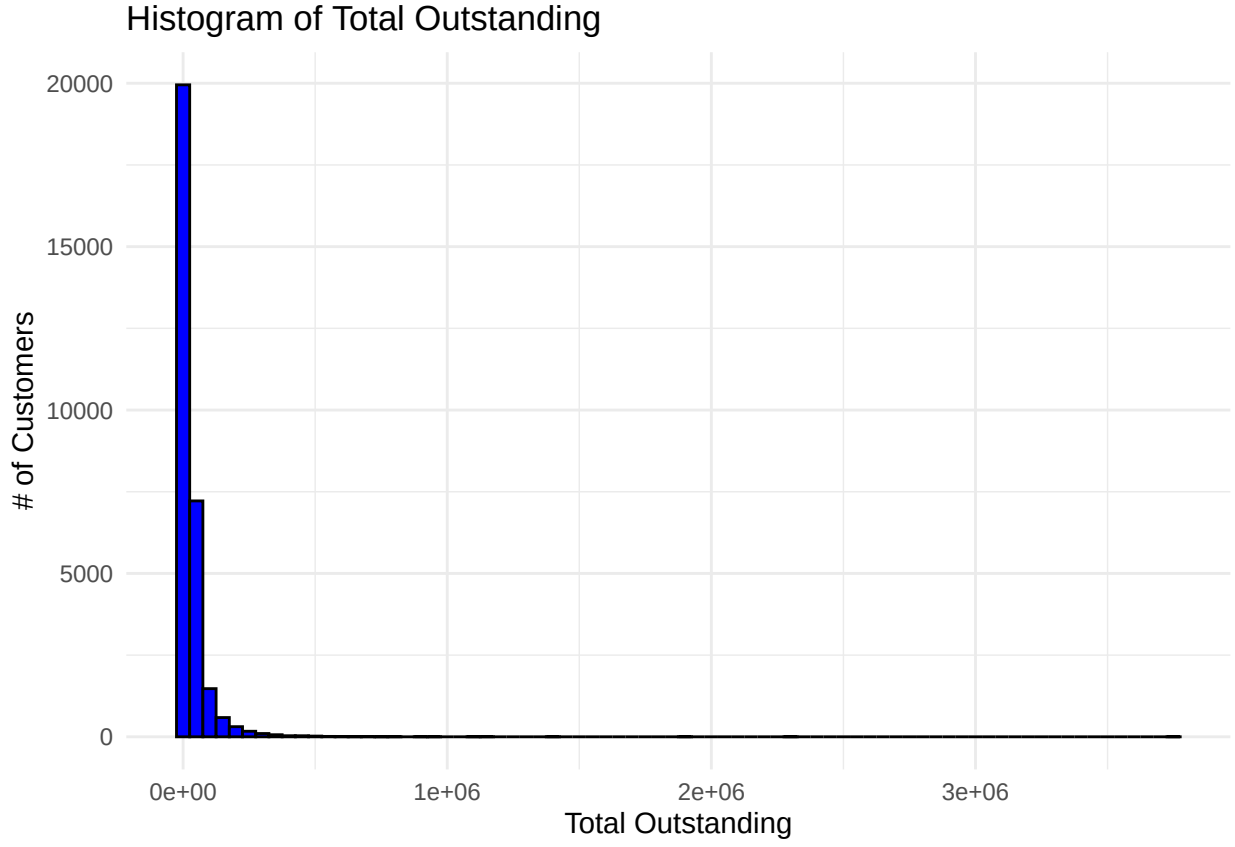
- The average credit limit amount is 167484.32
- Credits limits range from 10 000 to 1000 000
- Average male credit limit is 163519.82 and female is 170086.46

Credit Utilization Patterns



Payment Patterns





- Above we show the distribution of the credit utilization from April to Septemner. It's right skewed indicating that most clients tend to use a small portion of their allocoated credit limit # Education: Student Performance Dataset

Student Performance Data was obtained in a survey of students' math course in secondary school. It consists of 33 Column Dataset Contains Features like

- school ID
- gender
- age
- size of family
- Father education
- Mother education
- Occupation of Father and Mother
- Family Relation
- Health
- Grades
- This dataset can be used for Regression (as target variable Grade) as well as Analysis tasks. it might contain imbalanced category features.

Research Question

1. What top factors contribute to students performing well ?
2. How do male students perform in relation to female students?

Exploratory Data Analysis