Data Preprocessing

Daniel Sohm

July 2023

Contents

Introduction

For this project, we will work with a dataset acquired from kaggle.com. (link at the end)

The dataset encompasses sales data from three different branches over a span of three months. Additionally, it includes the customers' satisfaction ratings regarding the provided customer service and other data regarding the selling process (price, amount of goods sold, etc).

In this first part of our project, we will focus on preprocessing the data for further analysis, while also conducting a first exploratory data analysis.

So, this project will revolve around preprocessing the data, performing a first exploratory analysis, and later apply machine learning algorithms (supervised and non supervised), to uncover potentially interesting patterns or creating predictive models.

Objectives

As for objectives, the first one is to preprocess the data to match our requirements in the application of machine learning algorithms

Apply machine learning algorithms such as KNN, decision tree, random forest, etc..

Confirm if there is a relation between amount spent, gender or other factors and the overall satisfaction of the customer.

Exploratory Data Analysis

First, lets have a first glance at our data, and see what we are working with:

```
## 'data.frame':    1000 obs. of  17 variables:
##  $ Invoice.ID            : chr  "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" "373-73-791
##  $ Branch                : chr  "A" "C" "A" "A" "A" "C" "A" "C" "A" "B" ...
##  $ City                  : chr  "Yangon" "Naypyitaw" "Yangon" "Yangon" "Yangon" "Naypyitaw" "Yangon
##  $ Customer.type         : chr  "Member" "Normal" "Normal" "Member" "Normal" "Normal" "Member" "Nor
##  $ Gender                : chr  "Female" "Female" "Male" "Male" "Male" "Male" "Female" "Female" "Fe
##  $ Product.line          : chr  "Health and beauty" "Electronic accessories" "Home and lifestyle" "
##  $ Unit.price            : num  74.7 15.3 46.3 58.2 86.3 85.4 68.8 73.6 36.3 54.8 14.5 25.5 ...
##  $ Quantity              : int  7 5 7 8 7 7 6 10 2 3 4 4 5 10 10 6 7 6 3 2 5 3 2 5 3 ...
##  $ Tax.5.                : num  26.14 3.82 16.22 23.29 30.21 29.89 20.65 36.78 3.63 8.23 2.9 5.1 ..
##  $ Total                 : num  549 80.2 340.5 489 634.4 627.6 433.7 772.4 76.1 172.7 60.8 107.1 ..
##  $ Date                  : chr  "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" "2/8/2019" "3/25/2019"
##  $ Time                  : chr  "13:08" "10:29" "13:23" "20:33" "10:37" "18:30" "14:36" "11:38" "17
##  $ Payment               : chr  "Ewallet" "Cash" "Credit card" "Ewallet" "Ewallet" "Ewallet" "Ewall
##  $ cogs                  : num  522.8 76.4 324.3 465.8 604.2 597.7 413 735.6 72.5 164.5 57.9 102 ..
##  $ gross.margin.percentage: num  4.76 4.76 4.76 4.76 4.76 4.76 4.76 4.76 4.76 4.76 4.76 4.76 ...
##  $ gross.income          : num  26.14 3.82 16.22 23.29 30.21 29.89 20.65 36.78 3.63 8.23 2.9 5.1 ..
##  $ Rating                : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 4.5 6.8 7.1 8.2 5.7 4.5 4.6 6
```

Attribute information: Invoice id: Computer generated sales slip invoice identification number

Branch: Branch of supercenter (3 branches are available identified by A, B and C).

City: Location of supercenters

Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.

Gender: Gender type of customer

Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel

Unit price: Price of each product in $

Quantity: Number of products purchased by customer Tax: 5% tax fee for customer buying

Total: Total price including tax

Date: Date of purchase (Record available from January 2019 to March 2019)

Time: Purchase time (10am to 9pm)

Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)

COGS: Cost of goods sold

Gross margin percentage: Gross margin percentage

Gross income: Gross income

Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Check for NAs and modify data

After this first look, lets confirm if we have missing values in our data. As we have information regarding time, we can compact this in a new variable that indicates in which period during the day the sale was made.

```
##  Invoice.ID        Branch           City         Customer.type
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1000      FALSE:1000      FALSE:1000      FALSE:1000
##    Gender        Product.line    Unit.price       Quantity
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1000      FALSE:1000      FALSE:1000      FALSE:1000
##    Tax.5.          Total           Date            Time
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1000      FALSE:1000      FALSE:1000      FALSE:1000
##   Payment           cogs         gross.margin.percentage gross.income
##  Mode :logical   Mode :logical   Mode :logical           Mode :logical
##  FALSE:1000      FALSE:1000      FALSE:1000              FALSE:1000
##    Rating
##  Mode :logical
##  FALSE:1000
```

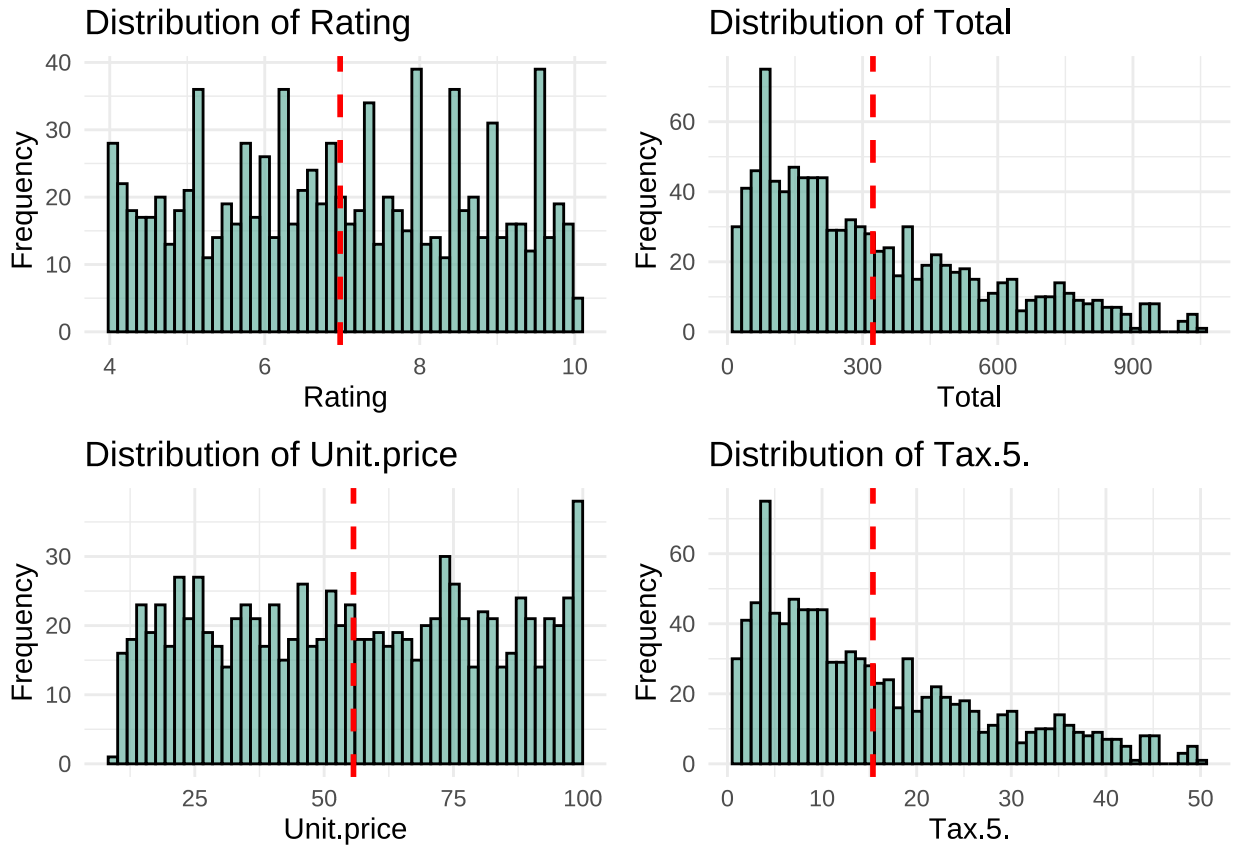
```
##   Rating            Datetime    Period
## 1    9.1 2019-01-05 13:08:00 Afternoon
## 2    9.6 2019-03-08 10:29:00   Morning
## 3    7.4 2019-03-03 13:23:00 Afternoon
## 4    8.4 2019-01-27 20:33:00   Evening
## 5    5.3 2019-02-08 10:37:00   Morning
```

We see that no variables contain missing values. We also transformed some of the variables to factors and binary based on their function. We can also see how the new column looks like after transformation.

Visual analysis

Lets begin our analysis with some graphical representations, as it gives an intuitive understanding of the data.

Numeric variables



For the rating, we can see that it is clearly not normally distributed, the lowest rating given is 4 and it averages around 7.

As for the total amount spent on purchases, it is also not normally distributed, but has a range from 0 to 1100, the mean is 300 and we can see that majority of purchases are valued around 300 or lower. For the other variables, the same happens and their distributions are not normal

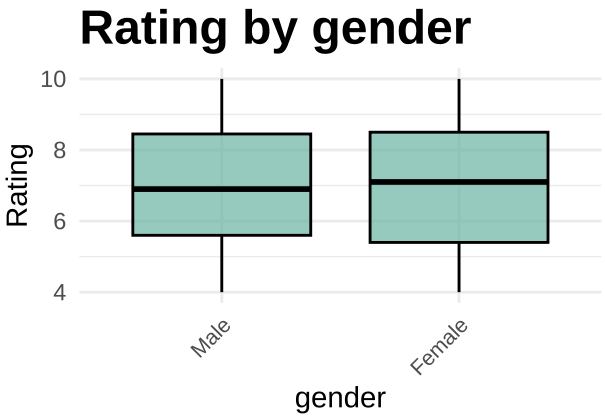
Categoric Variables

Branch and City have almost the same counts (makes sense as 1 city per branch).

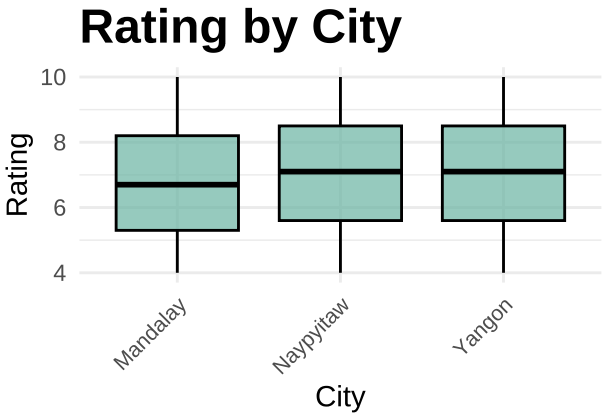
However, we can observer that majority of sales are made in the afternoon and evenings. Preferred payment methods are also pretty evenly distributed.

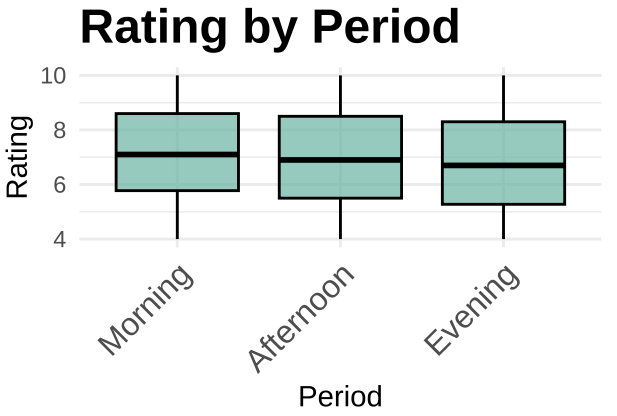
Comparison of variables graphically

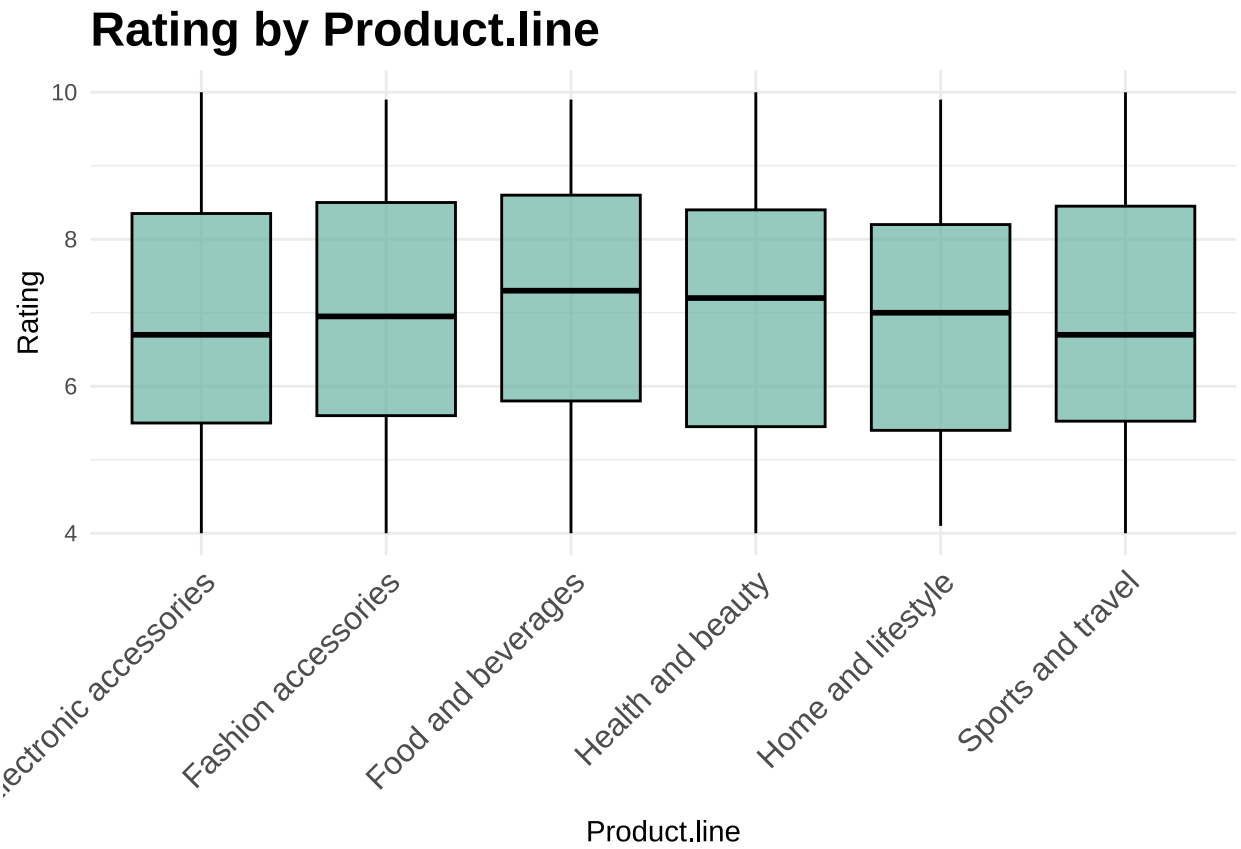
Lets take a closer look at the relationships our variables may present. For this, we can plot variables againts another


Rating by gender
Rating
10
8
6
4
Male
Female
gender

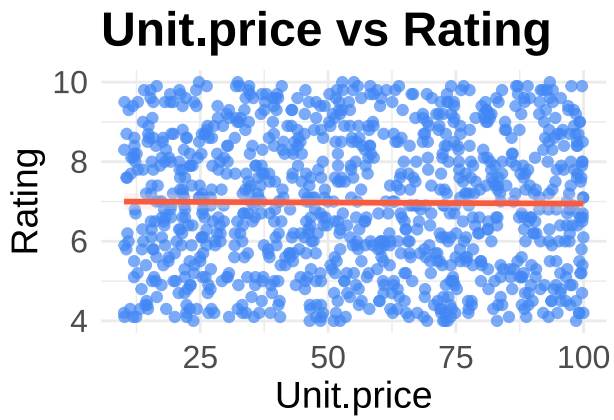

Rating by customer.typ
Rating
10
8
6
4
Member
Normal
customer.type

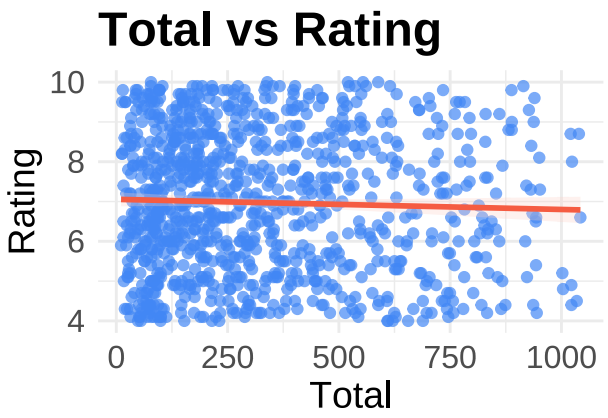

Rating by City
Rating
10
8
6
4
Mandalay
Naypyitaw
Yangon
City

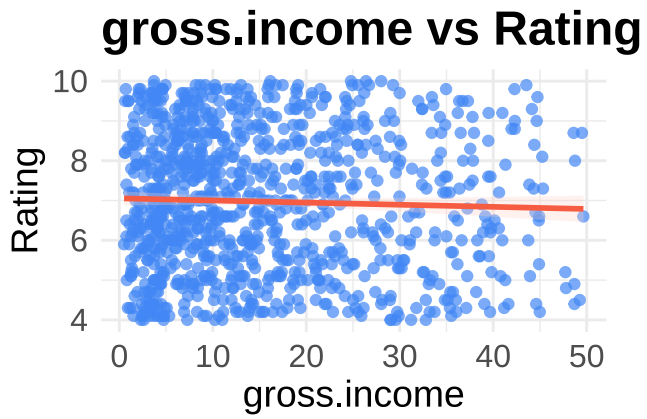

Rating by Period
Rating
10
8
6
4
Morning
Afternoon
Evening
Period



We already see a clear difference in satisfaction between purchases made in the morning, than the afternoon. However, satisfaction seems to not very between Male and Female. Rating means vary between servises purchased, being sports and travel the worst rated and food and beverages best rated. It also seems like customers are more prone to be stisfied in the morning than in the evenings. We can also see that there is a clear difference between ratings for Mandalay than the other 2





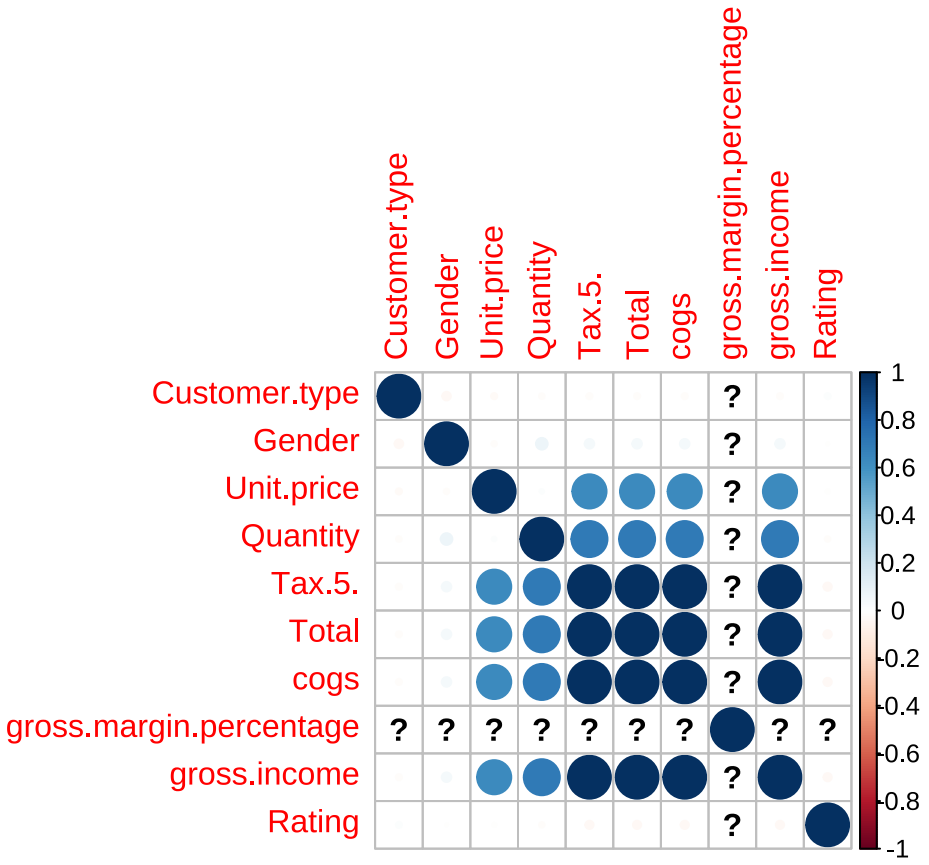


No visible correlation between the price of the item and satisfaction of the customer, which may seems strange because you want high paying customers to be treated better.

There is a slight negative relation between total paid and rating, this could be due to the fact that if a customer pays more, he expects better service. There is a slight negative realtion between the rating and the gross income, which goes down with the rating. which is to be expected

Correlation

Lets now check for correlation between our variables.



We can see that we have predictos with high correlation, but since we are not building regression models we will leave them as is for now, as clustering algorithms are not that sensible to colinearity, nor to data distribution. However, if we want to construct regression models, this issue might has to be addressed.

Normalizing and outlier removal

As ml algorithms are sensible to non normalized data and outliers, we will transform out data to fit our needs As for the outliers, we will use it on the normalized data, and remove outliers based on IQR.

```
##   Gender           Product.line  Unit.price   Quantity      Tax.5.       Total
## 1      1      Health and beauty  0.71780097  0.5096752  0.91914693  0.91914693
## 2      1 Electronic accessories -1.52454035 -0.1744526 -0.98723557 -0.98723557
## 3      0     Home and lifestyle -0.35260468  0.5096752  0.07141032  0.07141032
## 4      0      Health and beauty  0.09616553  0.8517391  0.67544187  0.67544187
## 5      0      Sports and travel  1.15638044  0.5096752  1.26649176  1.26649176
## 6      0 Electronic accessories  1.12165642  0.5096752  1.23899114  1.23899114
##       Payment        cogs gross.margin.percentage gross.income     Rating
## 1     Ewallet  0.91914693                4.761905   0.91914693  1.2378240
## 2        Cash -0.98723557                4.761905  -0.98723557  1.5287619
## 3 Credit card  0.07141032                4.761905   0.07141032  0.2486355
## 4     Ewallet  0.67544187                4.761905   0.67544187  0.8305111
## 5     Ewallet  1.26649176                4.761905   1.26649176 -0.9733034
## 6     Ewallet  1.23899114                4.761905   1.23899114 -1.6715541
##              Datetime    Period
## 1 2019-01-05 13:08:00 Afternoon
```

```
## 2 2019-03-08 10:29:00   Morning
## 3 2019-03-03 13:23:00 Afternoon
## 4 2019-01-27 20:33:00   Evening
## 5 2019-02-08 10:37:00   Morning
## 6 2019-03-25 18:30:00 Afternoon
```

Here we can see the data after the numeric variables have been normalized

```
## [1] 991
```

As we can see, we have only discarded 9 observations, so the data doesnt present a lot of outliers

Coding variables

To remove string variables, we will code each factor variable to only contain numbers. Coding will be done in factor level order, so first level is 1, and so on

```
##    Invoice.ID Branch City Customer.type Gender Product.line  Unit.price
## 1 750-67-8428      1    3             0      1            4  0.71780097
## 2 226-31-3081      3    2             1      1            1 -1.52454035
## 3 631-41-3108      1    3             1      0            5 -0.35260468
## 4 123-19-1176      1    3             0      0            4  0.09616553
## 5 373-73-7910      1    3             1      0            6  1.15638044
## 6 699-14-3026      3    2             1      0            1  1.12165642
##     Quantity      Tax.5.       Total Payment        cogs
## 1  0.5096752  0.91914693  0.91914693       3  0.91914693
## 2 -0.1744526 -0.98723557 -0.98723557       1 -0.98723557
## 3  0.5096752  0.07141032  0.07141032       2  0.07141032
## 4  0.8517391  0.67544187  0.67544187       3  0.67544187
## 5  0.5096752  1.26649176  1.26649176       3  1.26649176
## 6  0.5096752  1.23899114  1.23899114       3  1.23899114
##   gross.margin.percentage gross.income     Rating            Datetime Period
## 1                4.761905   0.91914693  1.2378240 2019-01-05 13:08:00      3
## 2                4.761905  -0.98723557  1.5287619 2019-03-08 10:29:00      2
## 3                4.761905   0.07141032  0.2486355 2019-03-03 13:23:00      3
## 4                4.761905   0.67544187  0.8305111 2019-01-27 20:33:00      4
## 5                4.761905   1.26649176 -0.9733034 2019-02-08 10:37:00      2
## 6                4.761905   1.23899114 -1.6715541 2019-03-25 18:30:00      3
```

Quick look at factors and the levels and count in each factor.

After these modifications, we have only some columns left that have not been adjusted to fit later algorithm application

#Single Value Decomposition

SVD plays a fundamental role in PCA, a technique used for dimensionality reduction and data visualization. It identifies the principal components (linear combinations of the original variables) that capture the maximum variance in the data.

Interpretation of Single, Left, and Right Vector Values:

Singular Values: The singular values in quantify the importance of each dimension in the reduced space. Larger singular values indicate more significant dimensions that capture a greater amount of variance or information in the data. They are often used to determine the rank or dimensionality of the reduced space.

Left Singular Vectors: The columns of the U matrix are the left singular vectors. Each column represents a vector that captures the relationships between the original variables i.Left singular vectors provide a basis for the reduced space and can be interpreted as a set of new variables that are linear combinations of the original variables.

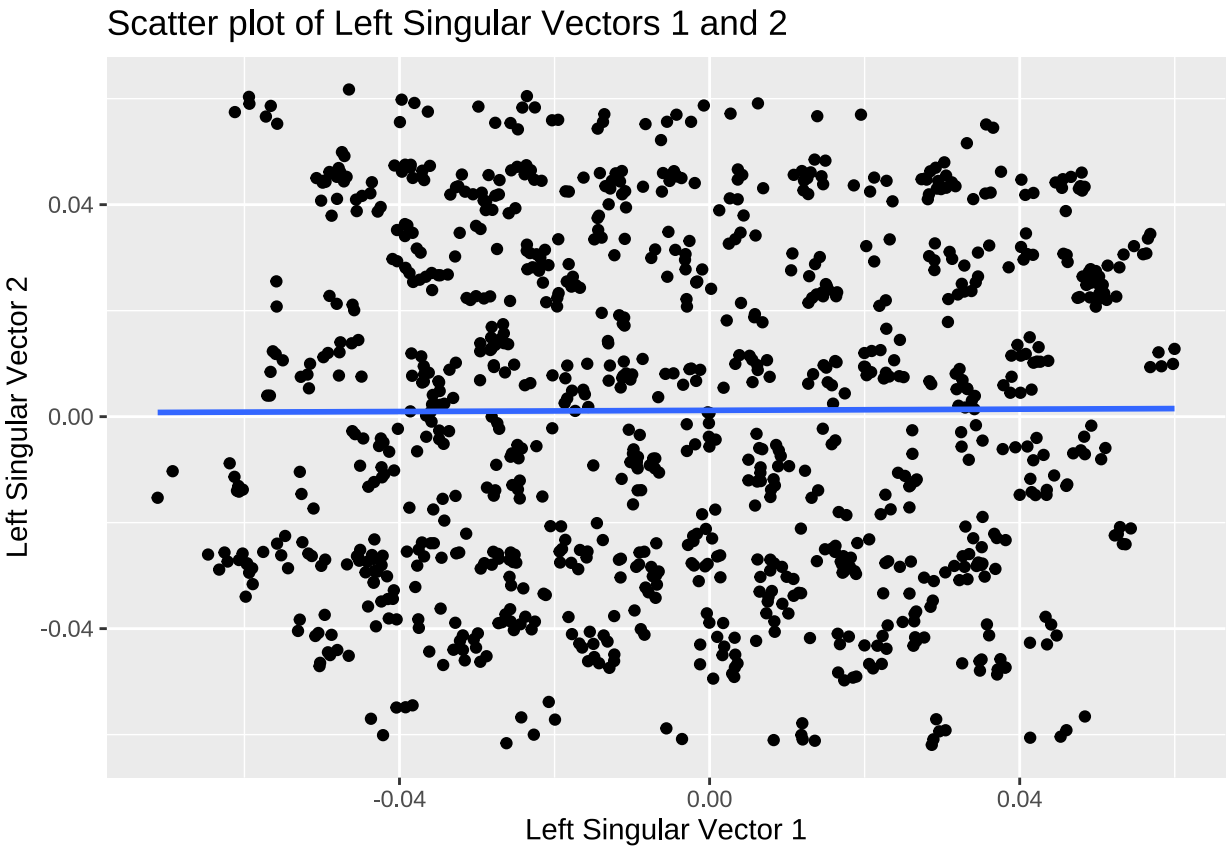
Right Singular Vectors: The columns of the V matrix are the right singular vectors. Each column represents a vector that captures the relationships between the observations in the data. ##Single vlaues vector

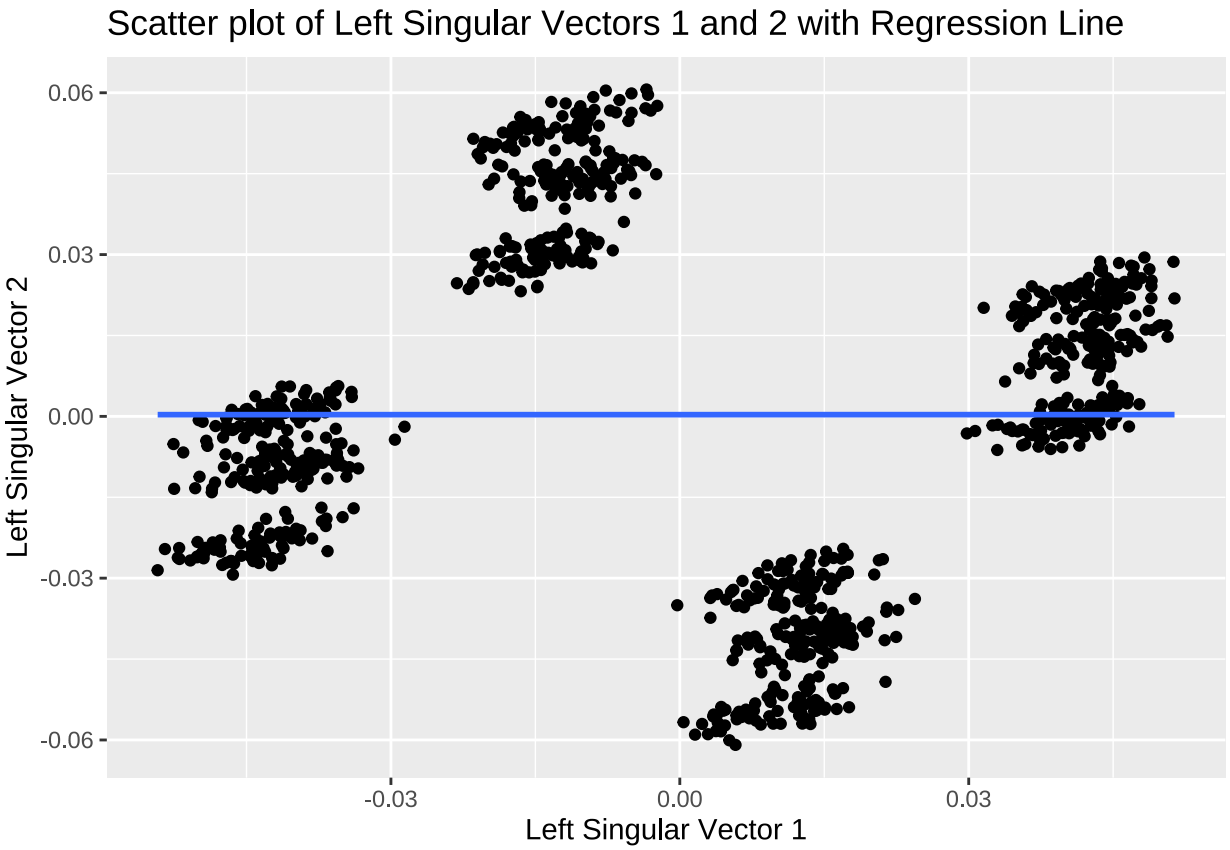
```
##  [1] 1.839674e+02 6.771097e+01 4.495199e+01 3.198840e+01 3.152142e+01
##  [6] 3.117291e+01 2.474571e+01 2.013265e+01 1.591885e+01 1.544843e+01
## [11] 9.347264e+00 1.563156e-14 3.828394e-15 3.694622e-15
```

Here we can see the values of the singular values vector, which amount of variability or importance is associated with the corresponding mode or component. Larger singular values represent more significant modes. Therefore, having higher singular values is generally considered more desirable. In summary, higher singular values indicate more important and significant modes, while smaller singular values correspond to less significant modes. Therefore, having higher singular values is generally preferred when analyzing the results of SVD. We can see that the last 3 variables which correspond to rating, gross income and period.

##Left Singular values vector

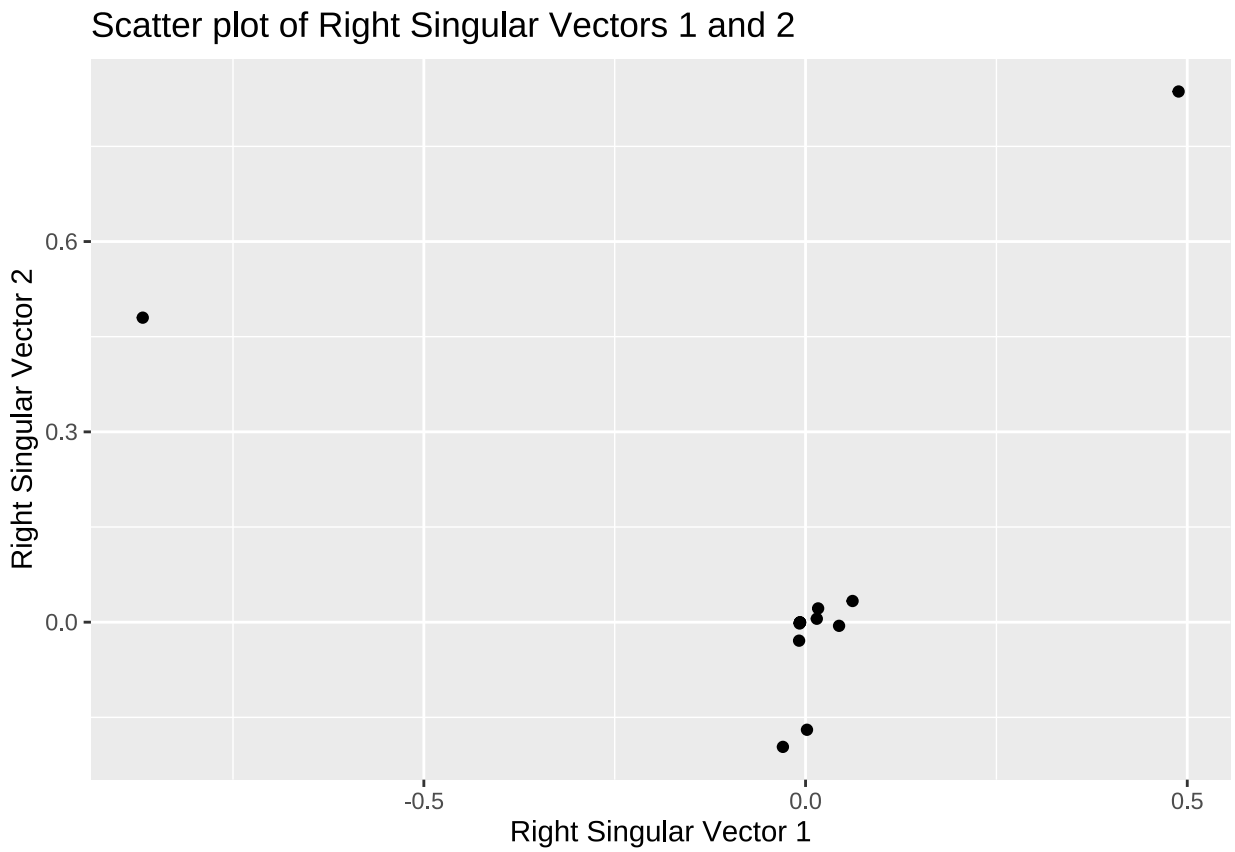
```
##            V1          V2           V3          V4           V5           V6
## 1 -0.03507823  0.03019925 -0.002308304 -0.05701202  0.020591056 -0.001984435
## 2 -0.02080653 -0.03400006 -0.039296389 -0.01444064  0.001527378  0.062852951
## 3 -0.03672318  0.00367489  0.022474678 -0.04275280 -0.015134282 -0.008117772
## 4 -0.03732420  0.02262421 -0.009875859 -0.05282684 -0.004090066 -0.001294731
## 5 -0.03922163  0.04234790  0.041327942 -0.02149087 -0.006135346 -0.053364059
## 6 -0.02652970  0.03978881 -0.062173484  0.03994740 -0.013277982 -0.036846797
##             V7           V8          V9         V10           V11          V12
## 1 -0.027647061  0.006949204  0.04700799  0.01371367  0.0086027152  0.008117705
## 2  0.034075003 -0.060112050 -0.01745689  0.02775072  0.0176574128  0.020148753
## 3  0.007205986  0.013242038 -0.04193348 -0.01087518 -0.0061799196  0.036941882
## 4 -0.016552819  0.050658579  0.01942815 -0.03990580  0.0009594506 -0.028214108
## 5 -0.042662254 -0.025901870 -0.04530500 -0.01198248  0.0077936069  0.033567573
## 6 -0.027341006 -0.029780385 -0.04345737 -0.02392328  0.0057513337  0.054974263
##            V13          V14
## 1  0.210196627 -0.148585629
## 2 -0.001543583  0.012614154
## 3 -0.151169523 -0.024262614
## 4  0.097947505  0.034920737
## 5 -0.123862886  0.008031899
## 6 -0.080817238  0.186418377
```


Scatter plot of Left Singular Vectors 1 and 2
Left Singular Vector 2
0.04
0.00
-0.04
-0.04
0.00
0.04
Left Singular Vector 1



The left singular vectors represent the patterns or directions in the original dataset that contribute the most to the variability captured by the singular value decomposition (SVD). Each column of the left singular vector matrix corresponds to a different mode or component. It also provides insights into the relationship. Each element in a left singular vector represents the weight or contribution of a specific variable to a mode. Looking for patterns of positive or negative weights across variables may help understanding how they interact or contribute to the variability in each mode. A pronounced slope indicates high correlation, while a random scatterplot indicates low relation. We can also discern clear clusters in the graphs, which is to be expected as it indicated groups with similar characteristics, given that there are several locations, so behavior differs.

##Right Singular vectors



It represents the relationships between the original variables or features in the dataset and the singular value components. Each right singular vector corresponds to a singular value and provides information about how the original variables contribute to that singular value component. Larger magnitudes indicate that the corresponding variables have a stronger influence on that singular value component. Exploring the patterns within each right singular vector may help identify any underlying structures or relationships between variables. Variables with larger contributions play a more significant role in explaining the variability captured by the corresponding singular value. Again, if we plot some of the variables, we can spot clustering, but this again is expected

No dimension reduction is going to happen however, as there are not a lot of variables to work with in the first place. However, it is interesting to have insight about how our data is structured and the relation between the variables we are working with

Source and credits for dataset: https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales