Exploratory Data Analysis on

Supply Chain Data

Through exploring the supply chain dataset with python, the objective is to understand the business landscape through data along with solving the 05 questions as below:

1. What are the product categories that generate the most profit?

2. What are product categories that are highly on demand in each country?

3. How does time reflect on sales?

4. What should the discount rate be good for the profit margin?

5. What is the distribution of customer by RFM score?

At the end, the analysis will come up with insights in accordance with problem statements. It is supposed to tap into the bottleneck and figure out the tipping point for the business in term of generating profit and expanding potential.

Table of content

A. DATASET

The given dataset includes 53 columns and 180,549 rows such as:

Type, Days for shipping (real), Days for shipment (scheduled), Benefit per order, Sales per customer, Delivery Status, Late_delivery_risk, Category Id, Category Name, Customer City, Customer Country, Customer Email, Customer Fname, Customer Id, Customer Lname, Customer Password, Customer Segment, Customer State, Customer Street, Customer Zipcode, Department Id, Department Name, Latitude, Longitude, Market, Order City, Order Country, Order Customer Id, Order Date, Order Item Cardprod Id, Order Item Discount, Order Item Discount Rate, Order Item Id, Order Item Product Price, Order Item Profit Ratio, Order Item Quantity, Sales, Order Item Total, Order Profit Per Order, Order Region, Order State, Order Status, Order Zipcode' Product Card Id, Product Category Id, Product Description, Product Image, Product Name, Product Price' Product Status, Shipping Date (Date Order), Shipping Time, Shipping Mode.

In general, the data is collected within two countries, the US and Puerto Rico, with the customer and purchase details, sales indicators as well as the logistics schedule in the period of 2015 to at the beginning of 2018.

B. DATA CLEANING

After investigating the dataset, there are 03 tasks required to done include:

1. Null / Not a number (NaN) value cleaning:

Method: Via Python - Apply function df.isnull().sum()

Status: There are 4 features containing null values including: Customer Lname with 08 values, Customer Zipcode with 03 values, Order Zipcode with 155,679 values, Product Description with 180,519 values. [Appendix 1]

Action: Replace every NaN value with 0 since detecting it.

2. Checking for duplicate

Method: Via Python – Check the value in the describe table with the mean, minimum and maximum value of each feature. Then, apply df[col1 name'].equals(df[col2 name’]) to testify.

Status: Column ‘Benefit per order’ shares same values with column ‘Order Profit Per Order’ [Appendix 2]

Action: Permanently drop column ‘Benefit per order’ .

3. Remove irrelevant columns to the upcoming analysis:

Method: Investigate columns with strange character, empty of value or the data is not necessary to explore.

Status: There are 20 features considered as irrelevant including:

Customer Email, Customer Fname, Customer Lname, Customer Password, Customer Street, Customer Zipcode, Latitude, Longitude, Order Time, Order Zipcode, Product Card Id, Product Category Id, Product Description, Product Image, Product Status, Shipping Time, Order Item Cardprod Id, Order Item Id, Category Id, Order Customer Id. [Appendix 3]

Action: Permanently drop 20 irrelevant columns

C. EXPLORATORY DATA ANALYSIS

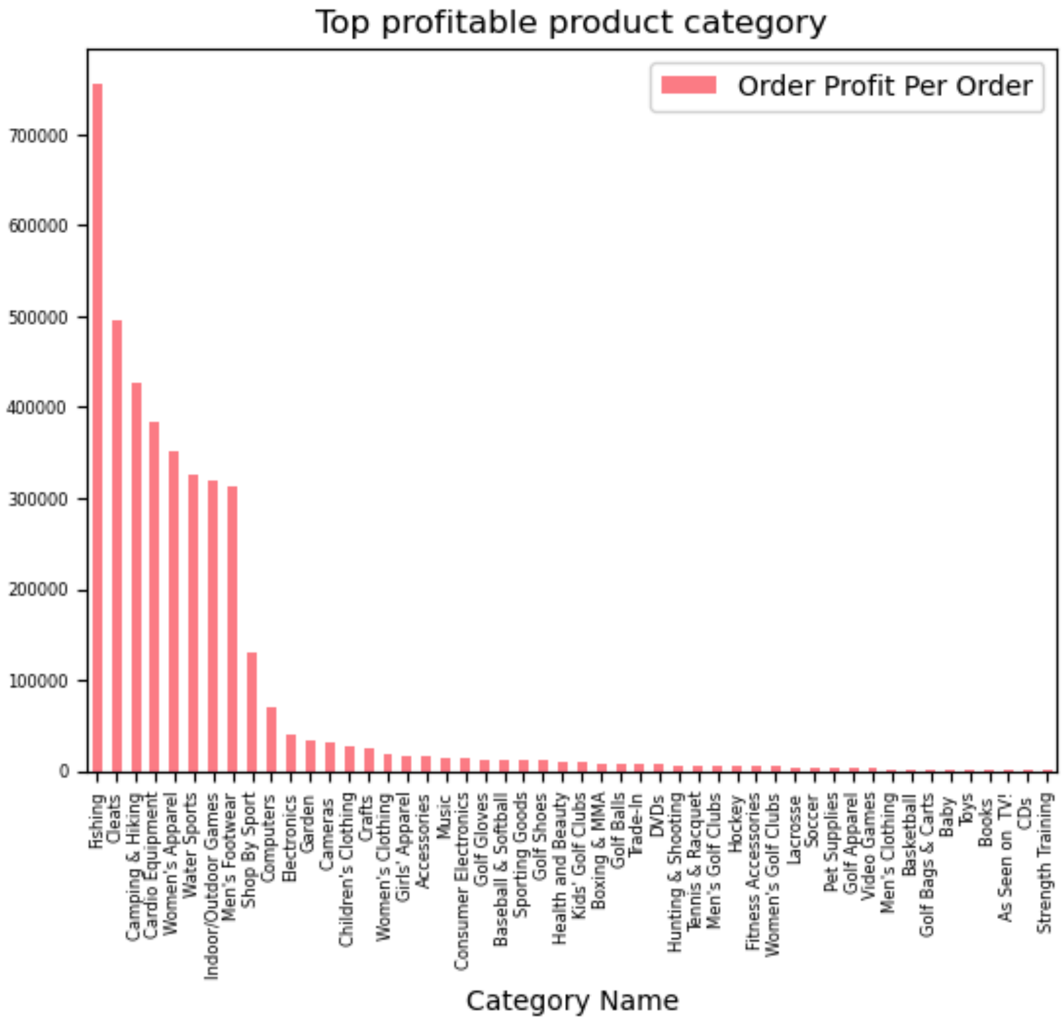
# An overall view of the dataset

To get an overall understanding to the every feature available in the dataset, the heat map is performed between numeric values [Appendix 4]. The chart demonstrates how the features are related to each other through the positive and negative or even neutral correlation indicated by the different color. Most of the findings is derived from this chart to further investigation.

1. What are the product categories that generate the most profit?

The chart is created by the profit summation of 50 product categories [Appendix 5]. The top profitable categories which have the sum of profit above $100,000 including: Fishing, Cleats, Camping and Hiking, Cardio Equipment, Women’s Apparel, Water Sports, Indoor/Outdoor Games, Men’s Footwear and Shop By Sport. It can be seen as the profitable category most distributes at the category for sport or workout purpose.

On the other hand, the top 05 categories that have not been doing well in term of profit are Strength Training, CDS, As Seen on TV!, Books and Toys. They both have the sum of profit of the 03-year-period below $1,000. This can be seen as the business has the setback in conducting some of products for recreation.

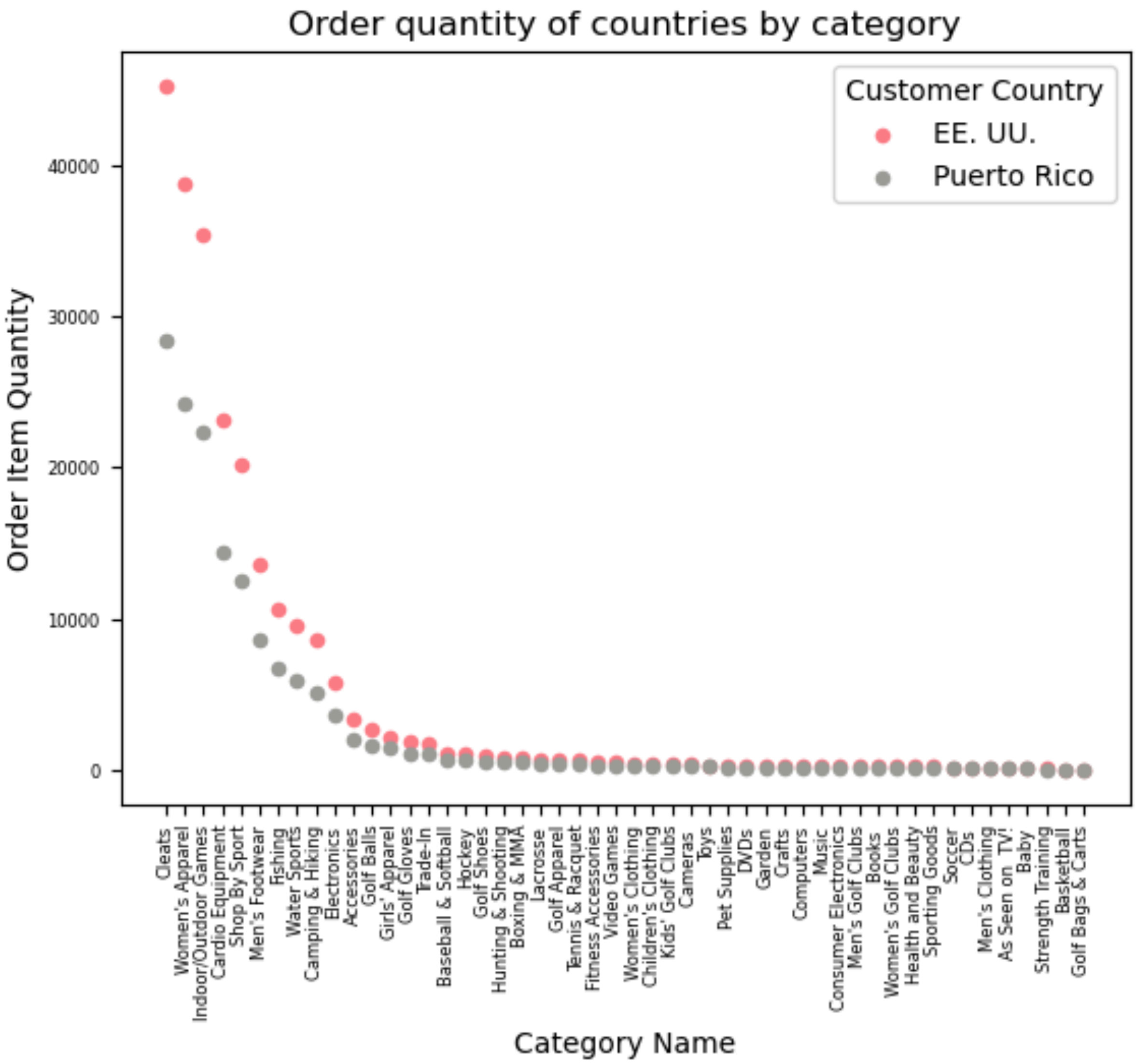


Appendix 5: Ranking of the 50 product categories basing on the sum of profit from 2015 to 2018.

2. What are product categories that are highly on demand in each country?

The scatter is plotted by the summation of order quantity of the US and Puerto Rico. The reason to utilise the order quantity in term of answer the question about demanding in each country is the transformation of discount rate in term of time, location. Basing on observation, the discount rate is vacillating by the product itself, days, cities that there is no evidence for the rule of discount rate allocation. In a simple word, if the discount is low or not available and the purchase is still carried on, the demand clearly exists.

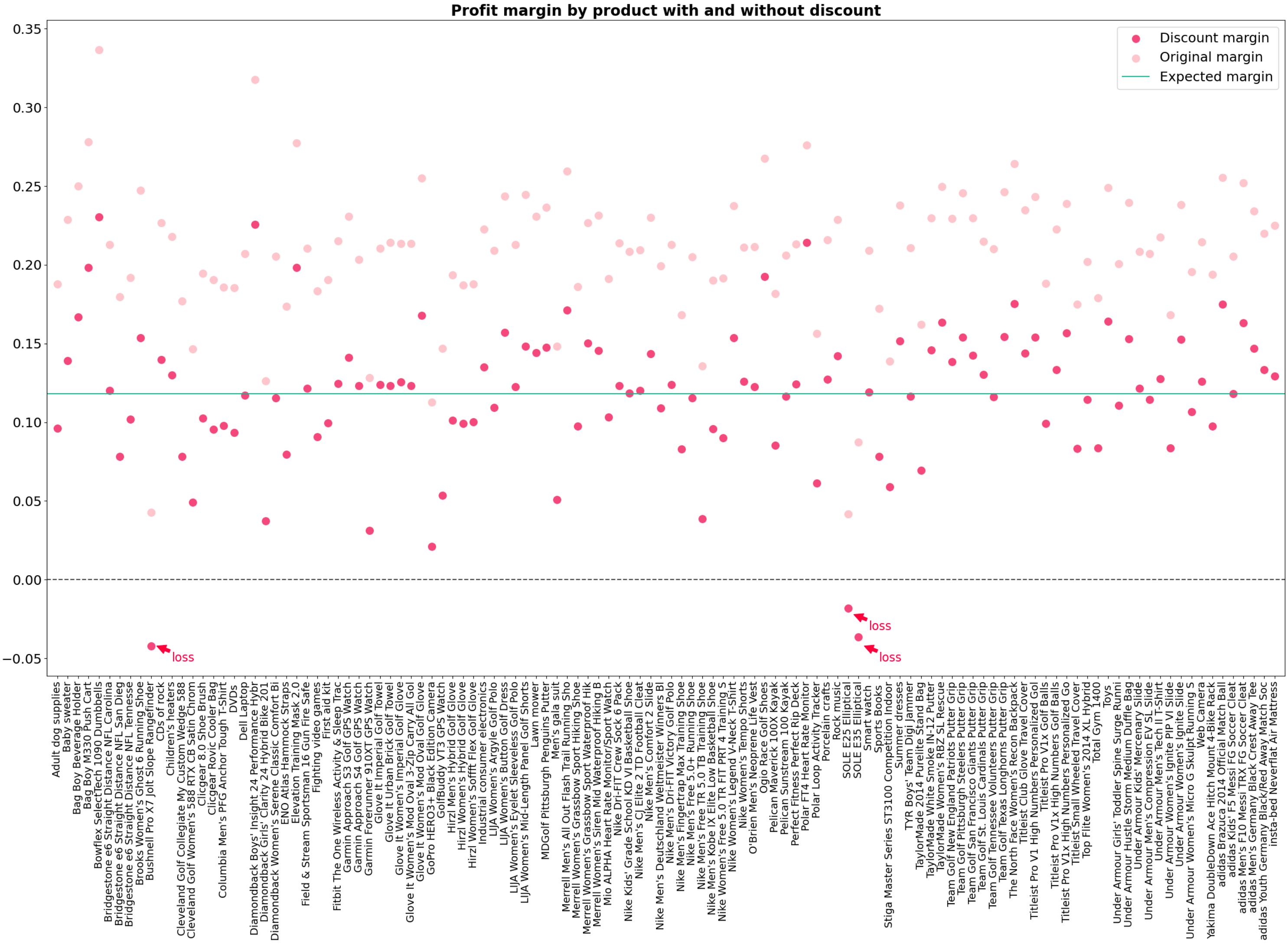
The product categories , in general, are mostly the one profitable in question 1 [Appendix 6]. Both countries share the top 05 categories which have over 10,000 numbers of order, involving: Cleats, Women's Apparel, Indoor/Outdoor Games, Cardio Equipment and Shop By Sport. Respectively, the consumption in the US is higher about double as in Puerto Rico. However, it can be affected as the the number of customer making a purchase in Puerto Rico is just more than a half of the one in the US [Appendix 7]. Meanwhile, the rest of categories do not have much of gap in the volume for both the US and Puerto Rico.



Appendix 6: The purchase quantity of the two countries by product category.

3. What should the discount rate be good for the profit margin?

The plotted data rings the alarm that there are 52 products among 118, having the profit margin below the average (the cyan line) in which 03 of them are suffering loss including Bushnell Pro X7 Jolt Slope Rangefinder, SOLE E25 Elliptical, SOLE E35 Elliptical and 01 product - GoPro HERO3+ Black Edition Camera, is on the edge. It indicates that with the discount rate, nearly half of the goods will end up drag the total margin down [Appendix 8].



Appendix 8: Profit margin of 118 product with and without discount.

The chart is founded by the discount and original profit margin. The original profit margin is the one involving sales and profit without discount rate and conversely for the discount one. The idea to combine the original margin into the chart is to give the general view when there is no discount application. However, basing on observation, there is still a few case below the expected level.

For that reason, to improve the sale performance, the business should be sensitive about the discount rate and price strategy for each product. In fact, the discount rate varies with different location and time so this analysis will take the mean of the discount rate of a specific product to be that product's discount rate.

There is a suggestion that the new discount rate should drag the new profit above or at the average one which is equals to 0.118056 [Appendix 10]. Hence, for this dataset, it is safe to say that new discount rate should be less than or equal to the result of the calculation for the profit margin be at average.

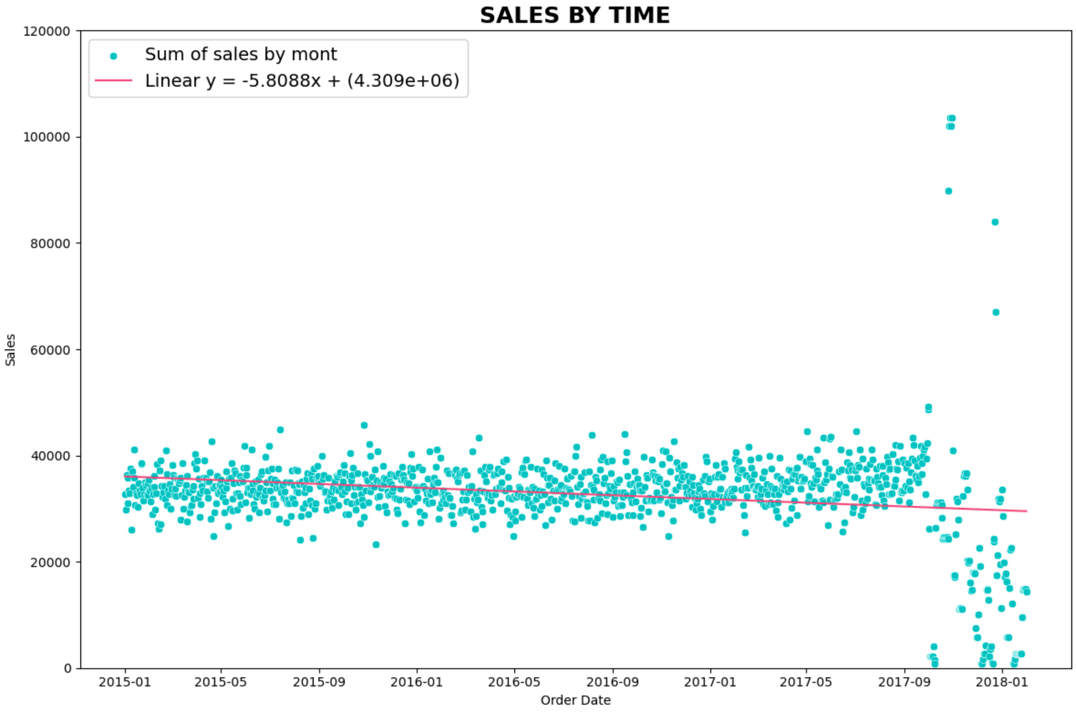
However, there are 04 products (Bushnell Pro X7 Jolt Slope Rangefinder, SOLE E25 Elliptical, SOLE E35 Elliptical, GoPro HERO3+ Black Edition Camera) must be alert about the price strategy as the given new discount rates are negative [Appendix 11]

** All the calculation of this part is completely performed in Jupyter Notebook. Please refer Appendix 9 to Appendix 11 for the calculation and formulas.*

4. How does time reflect on sales?

Basing on the given statistic, the chart demonstrates a prediction of sales basing on historical sales record in the past 03 years from January 2018 [Appendix 12]. Overall, the performance is quite plain as there is no trend or seasonality although a few peak at the end of 2017 and the sudden drop at the transition between 2017 and 2018. However, this cannot assure for the decrement in sales due to the lacking of data collection of some product at the transition between 2017 and 2018 in the given raw data.

Regardless the shortage in data, the major performance has no peak while the business should have several peak sales basing on seasonality thanks to the advantage of product diversity. The company should invest more occasional sale campaigns to increase the revenue efficiency.



Appendix 12: Prediction sales throughout periods of time

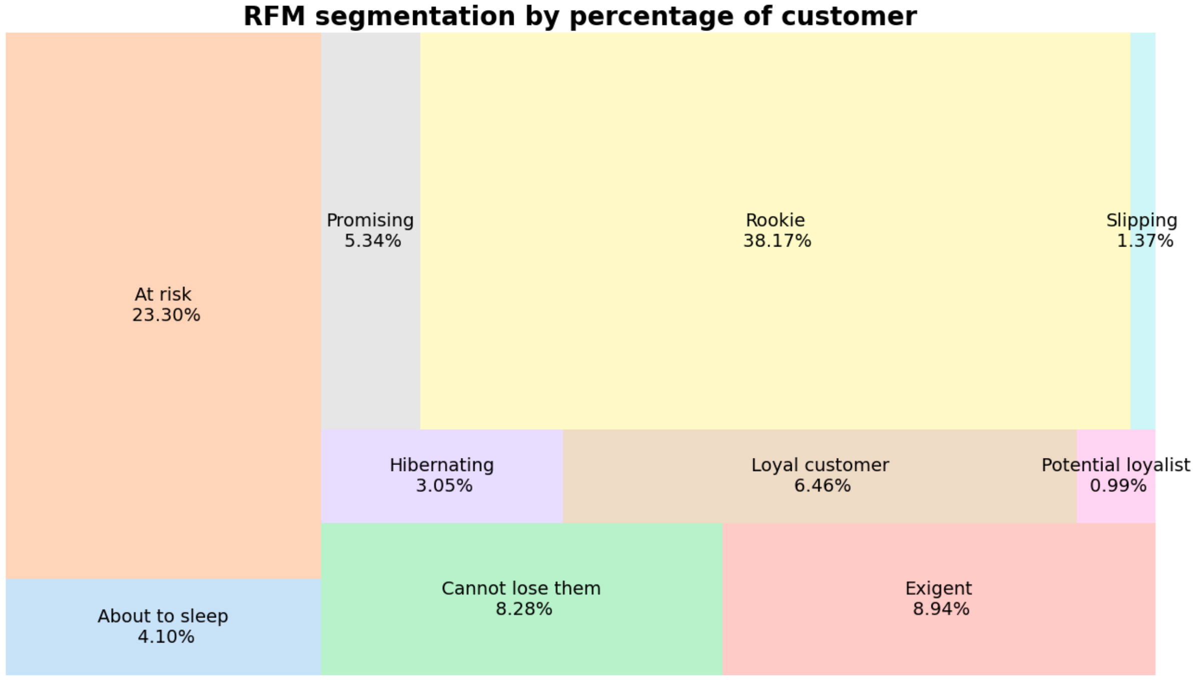
5. What is the distribution of customer by RFM score?

RFM (Recency - Frequency - M (Monetary Value): is a part of Marketing Analysis and is used to analyse customer value, thereby helping businesses to analyse each customer group that they existing, from which there are marketing campaigns or special care (Murphy 2022). The attributes can be briefly explained as:

- Recency (R): The period of time when the customer made the latest purchase.
- Frequency (F): The rate of customer making a purchase in a certain period of time.
- Monetary (M): The value per purchase that customer to spend.

Each attribute is calculated from the given data and categorised by value-based indicators. It is stated that there is 11 customer segmentation with description concluding from the variety of R-F-M values (Connectif, 2022) [Appendix 13]. Among 20,652 customers, the distribution of the dataset is categorised by 10 out of 11 groups in term of RFM score [Appendix 14].

It can be seen the Top valuable customer who owns the high score for both attributes missing out. The distribution is dominated by Rookie and At risk group [Appendix 15]. The rookie is the one has purchased lately while the At risk is the now that the brand would not see them in period of time. Furthermore, the high value-based group is occupied at the low rate consisting of 6.46% of Loyal customer and 0.99% of Potential loyalist. Meanwhile the percentage of the low value-based including About to sleep, At risk, Hibernating and Slipping occupy up to more than 30% of the segmentation.



Appendix 15: RFM segmentation by percentage of customer

The unequal distribution can cause the huge loss in number of customer. It is believed that the business should invest in marketing campaign in order to build the brand value as well as customer relationship strategy for each specific group. For instance, with Rookie, they are new customer to the brand, the campaign with on-board special deal would be an impeccable tactic. Besides, last but not least, the business should put some effort to build up to Top valuable customer group because they are considered to be the one who would effectively contributed on the way of generating profit in a long term.

D. EVALUATION

The analysis explores the dataset at the aspect of profit, sales and customer. Respectively, the business has seemed to suffer some setbacks in customer retention, discount and price strategy as well as optimisation of sales by time. Therefore, there are 03 recommendation in accordance with the 03 setbacks including:

1. **Sales:** The business has some profitable (category) products. However, the majority is mostly covered by the low one. Therefore, the company should create the opportunities to increase the sales potential by seasonality or trend. Moreover, despite doing fine in the US market, the business should not ignore the market expansion and growth in the Puerto Rico.

2. **Discount and price strategy:** Discount is a perfect tactic to improve sales and attract consumption. However, it should not be the cause of the loss in conduct. To specific products such as Bushnell Pro X7 Jolt Slope Rangefinder, SOLE E25 Elliptical, SOLE E35 Elliptical, GoPro HERO3+ Black Edition Camera, there must a revision in pricing and cost evaluation. Since the financial detail is not available in the give dataset, this report is unable to detect the underlying cause.

3. **Customer building relationship:** The business is owning the customer profile in which majorly are customer from medium to low value-based segmentation. Therefore, the upcoming marketing strategies should focus on the priority of nourishing the exisiting customer and limiting the increment of low segment.

   - For the high segments (Top valuable customer, Loyal Customer, Potential Loyalist), loyalty programs should be actively pushed throughout a year.

   - For the medium segments (Rookie, Promising, Need Attention, Cannot Lose Them), discount and special deal will effectively attract them. And for sure, investment in customer acquisition and retention is inevitable.

   - For the low segments (About To Sleep, At Risk, Hibernating, Slipping), those are the red flag groups as they need a unique hook to stay with the brand. Special offer and communication strategy (such as influencer marketing, endorsement) are considered to be helpful in this case.

E. REFERENCE

Connectif, 2022, ‘What Are RFM Scores and How To Calculate Them’, Connectif site, accessed 22 March 2023, available from: <https://connectif.ai/en/what-are-rfm-scores-and-how-to-calculate-them/>

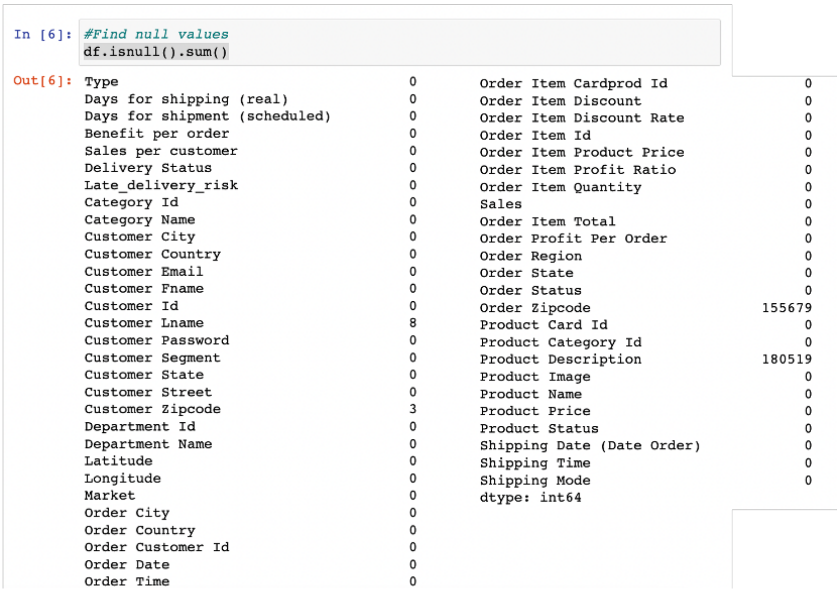
Girardin, M., 2023, ‘How to Calculate Profit Margins: Definition and Examples’, Forage site, accessed 25 March 2023, available from: <https://www.theforage.com/blog/skills/how-calculate-profit-margin#:~:text=Net%20Profit%20Margin%20%3D%20(Net%20Profit,100%20to%20create%20a%20percentage.>

Mucha, M. and Pál, T., 2023, ‘Margin With Discount Calculator’, Omni Calculator site, accessed 22 March 2023, available from: <https://www.omnicalculator.com/finance/margin-discount#:~:text=The%20formula%20expressing%20the%20new,to%20be%20expressed%20as%20decimals.>

Murphy, C., 2022, ‘What Is Recency, Frequency, Monetary Value (RFM) in Marketing?’, Investopedia site, accessed 25 March 2023, available from: <https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp>

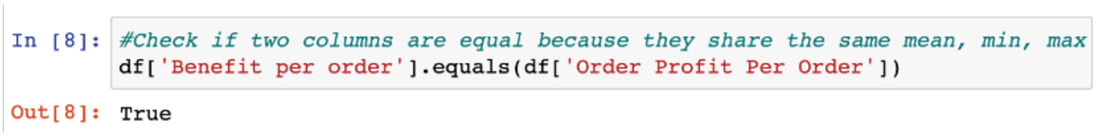
F. APPENDIX

Appendix 1: Number of null values in each feature



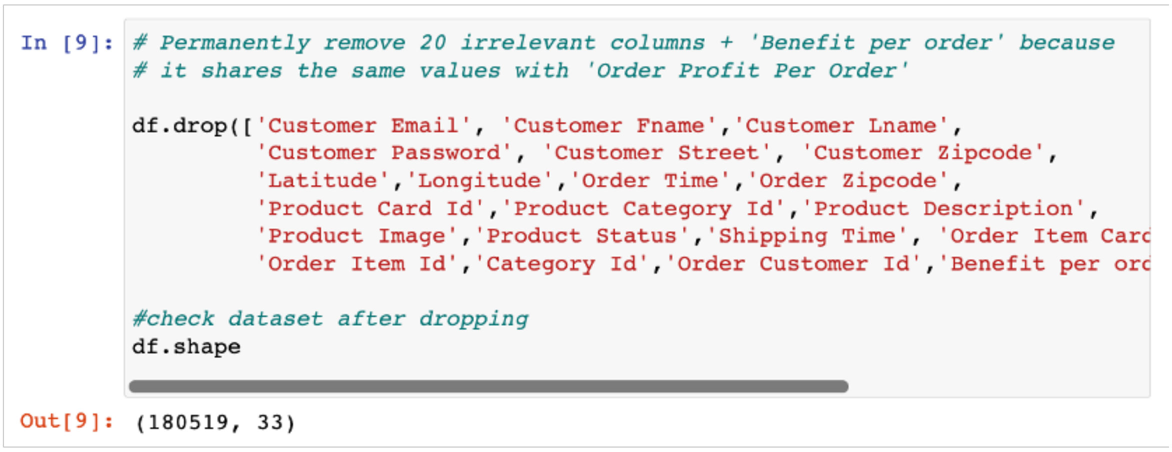
```
In [6]: #Find null values
        df.isnull().sum()

Out[6]: Type                                 0
        Days for shipping (real)             0
        Days for shipment (scheduled)        0
        Benefit per order                    0
        Sales per customer                   0
        Delivery Status                      0
        Late_delivery_risk                   0
        Category Id                          0
        Category Name                        0
        Customer City                        0
        Customer Country                     0
        Customer Email                       0
        Customer Fname                       0
        Customer Id                          0
        Customer Lname                       8
        Customer Password                    0
        Customer Segment                     0
        Customer State                       0
        Customer Street                      0
        Customer Zipcode                     3
        Department Id                        0
        Department Name                      0
        Latitude                             0
        Longitude                            0
        Market                               0
        Order City                           0
        Order Country                        0
        Order Customer Id                    0
        Order Date                           0
        Order Time                           0
        Order Item Cardprod Id                    0
        Order Item Discount                       0
        Order Item Discount Rate                  0
        Order Item Id                             0
        Order Item Product Price                  0
        Order Item Profit Ratio                   0
        Order Item Quantity                       0
        Sales                                     0
        Order Item Total                          0
        Order Profit Per Order                    0
        Order Region                              0
        Order State                               0
        Order Status                              0
        Order Zipcode                        155679
        Product Card Id                           0
        Product Category Id                       0
        Product Description                  180519
        Product Image                             0
        Product Name                              0
        Product Price                             0
        Product Status                            0
        Shipping Date (Date Order)                0
        Shipping Time                             0
        Shipping Mode                             0
        dtype: int64
```

Appendix 2: Testify the columns are identical similar



```
In [8]: #Check if two columns are equal because they share the same mean, min, max
        df['Benefit per order'].equals(df['Order Profit Per Order'])

Out[8]: True
```

Appendix 3: Irrelevant columns in given dataset



```
In [9]:  # Permanently remove 20 irrelevant columns + 'Benefit per order' because
          # it shares the same values with 'Order Profit Per Order'

          df.drop(['Customer Email', 'Customer Fname','Customer Lname',
                   'Customer Password', 'Customer Street', 'Customer Zipcode',
                   'Latitude','Longitude','Order Time','Order Zipcode',
                   'Product Card Id','Product Category Id','Product Description',
                   'Product Image','Product Status','Shipping Time', 'Order Item Carc
                   'Order Item Id','Category Id','Order Customer Id','Benefit per orc

          #check dataset after dropping
          df.shape

Out[9]:   (180519, 33)
```

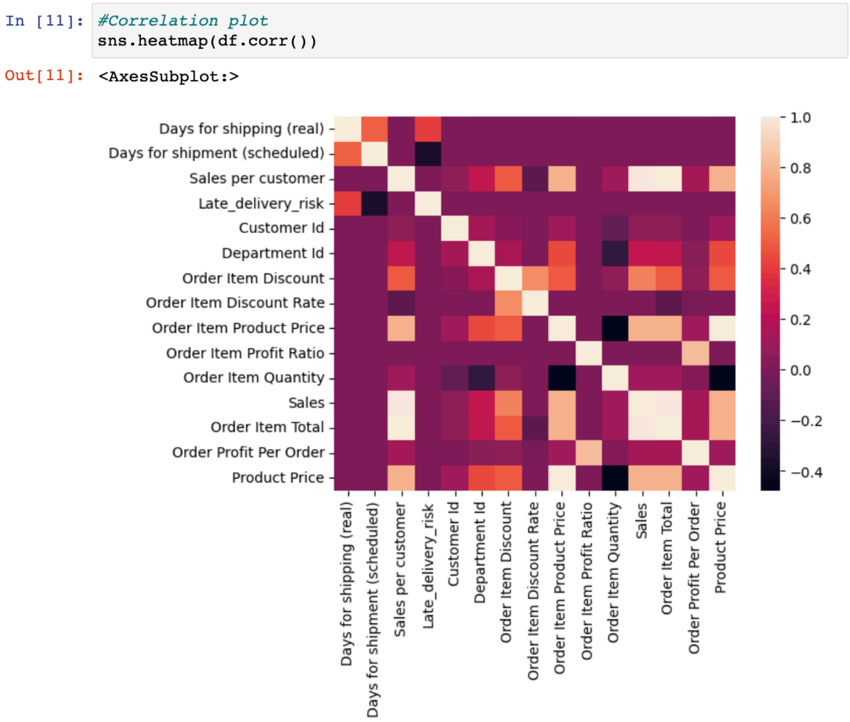
Appendix 4: Heat map performs the correlation between columns available on the dataset.

The correlation is explained as:

- A Positive correlation (greater than 0): Both value change in the same direction.

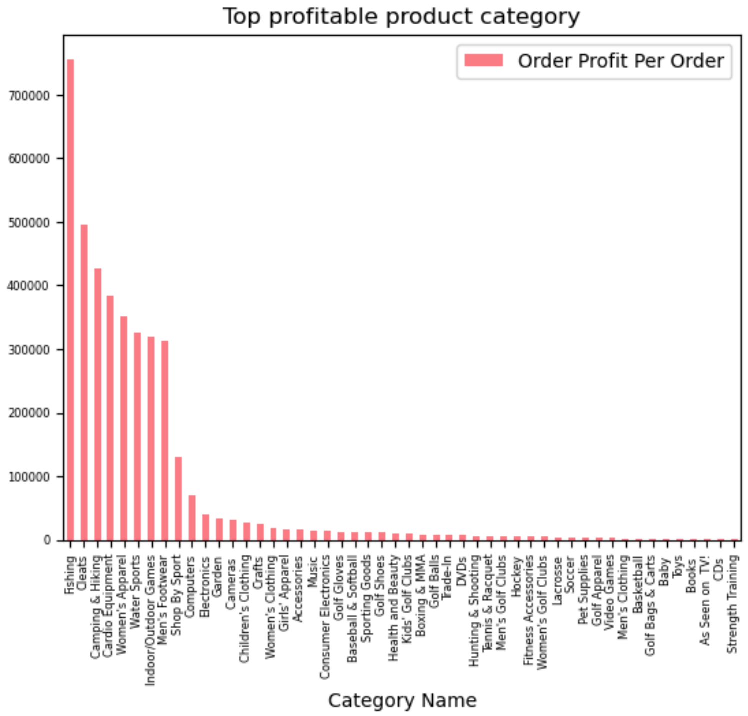
- A Neutral correlation (0): The two values have no relationship

- A Negative correlation (less than 0): The two value change in the contrast direction.

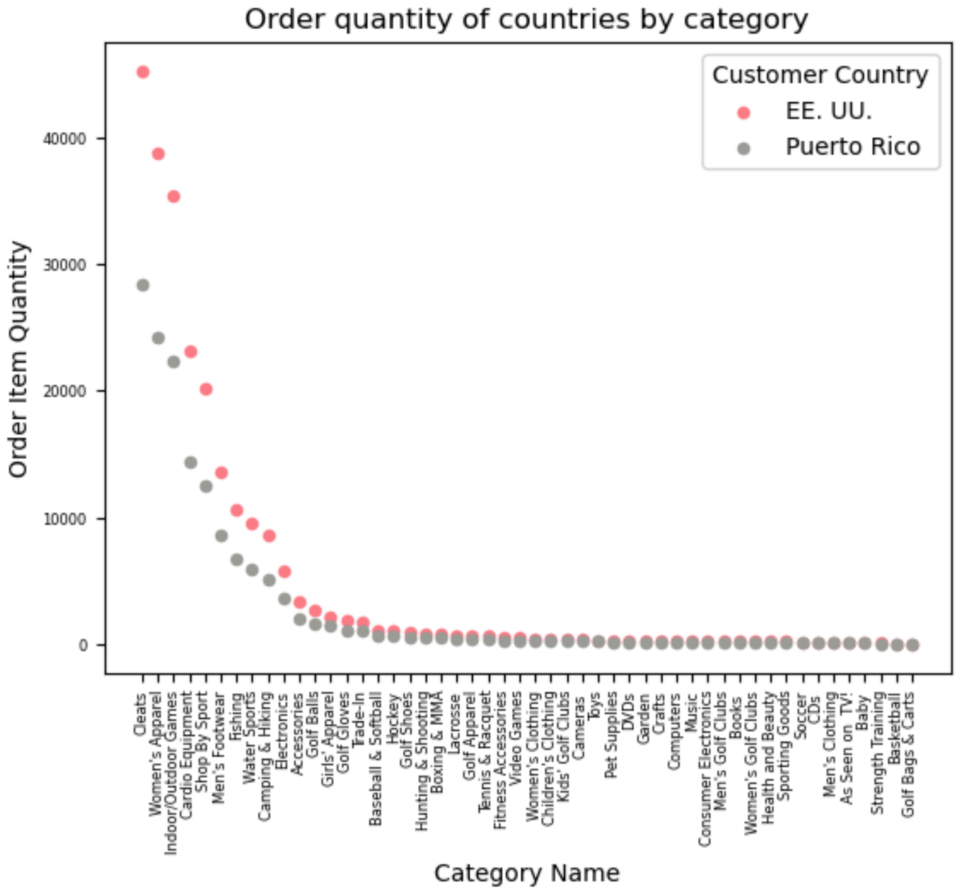


```
In [11]:  #Correlation plot
          sns.heatmap(df.corr())

Out[11]:  <AxesSubplot:>
```

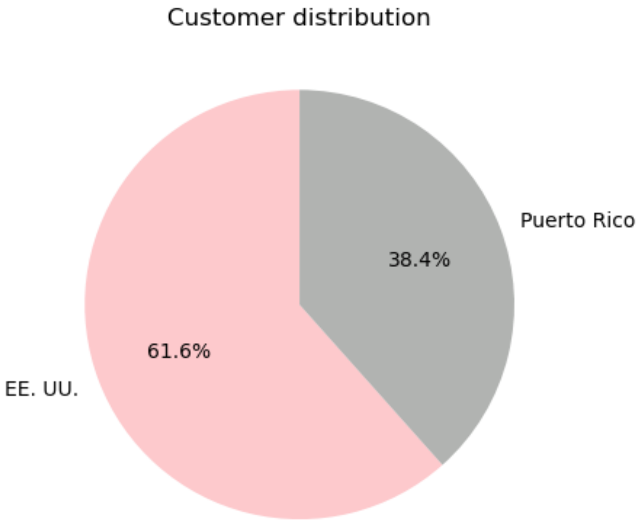
Appendix 5: Ranking of the 50 product categories basing on the sum of profit from 2015 to 2018.



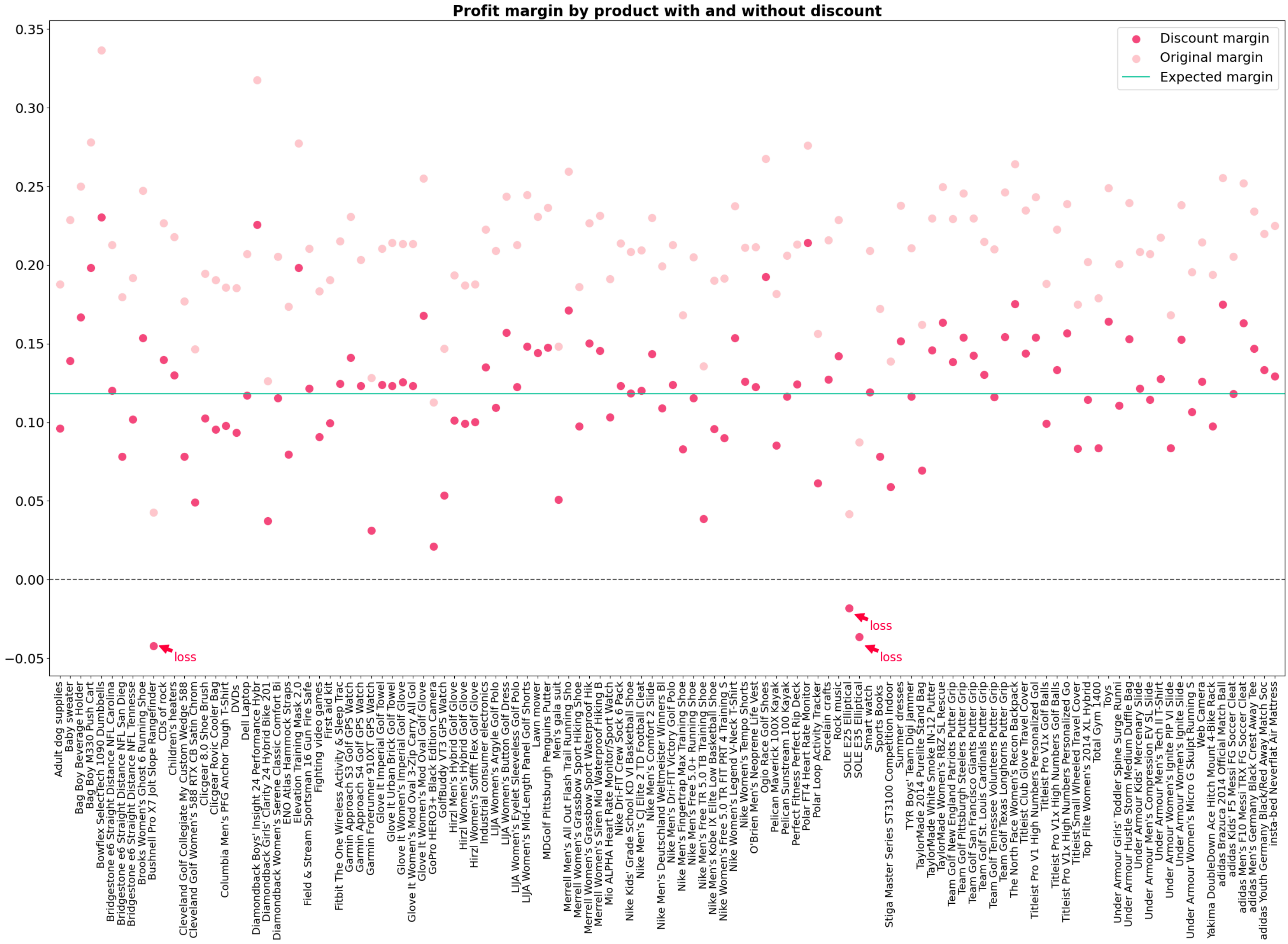
Appendix 6: The purchase quantity of the two countries by product category.



Appendix 7: The distribution of customer by country that makes the purchase



Appendix 8: Profit margin of 118 product with and without discount.



Appendix 9: Calculation of Mean of discount rate, New mean of discount rate, Discount, Original and New profit margin through Jupyter Notebook.

# Formula 9.1: Applied for the Discount profit margin

Profit margin = Profit / Sales

(Girardin 2023)

# Formula 9.2: Applied for the Mean of discount rate, New mean of discount rate

Profit Discount margin = Discount profit / Discount sales [in which discount sale = original sale (1- discount rate)]

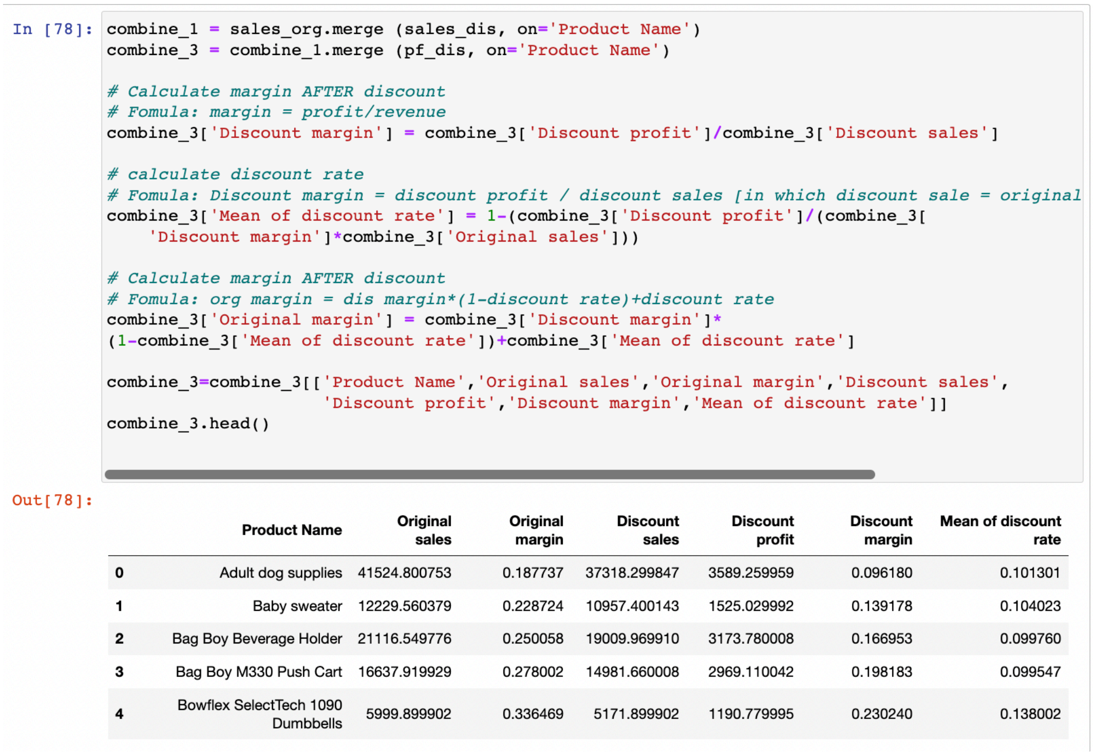
<=> Discount margin = Discount profit / Original sale (1- Discount rate)

<=> Discount rate = 1 - [Discount profit / (Discount margin * Original sale)]

# Formula 9.3: Applied for Original margin, New margin

Discount margin = (Original margin - Discount rate) / (1 - Discount rate)

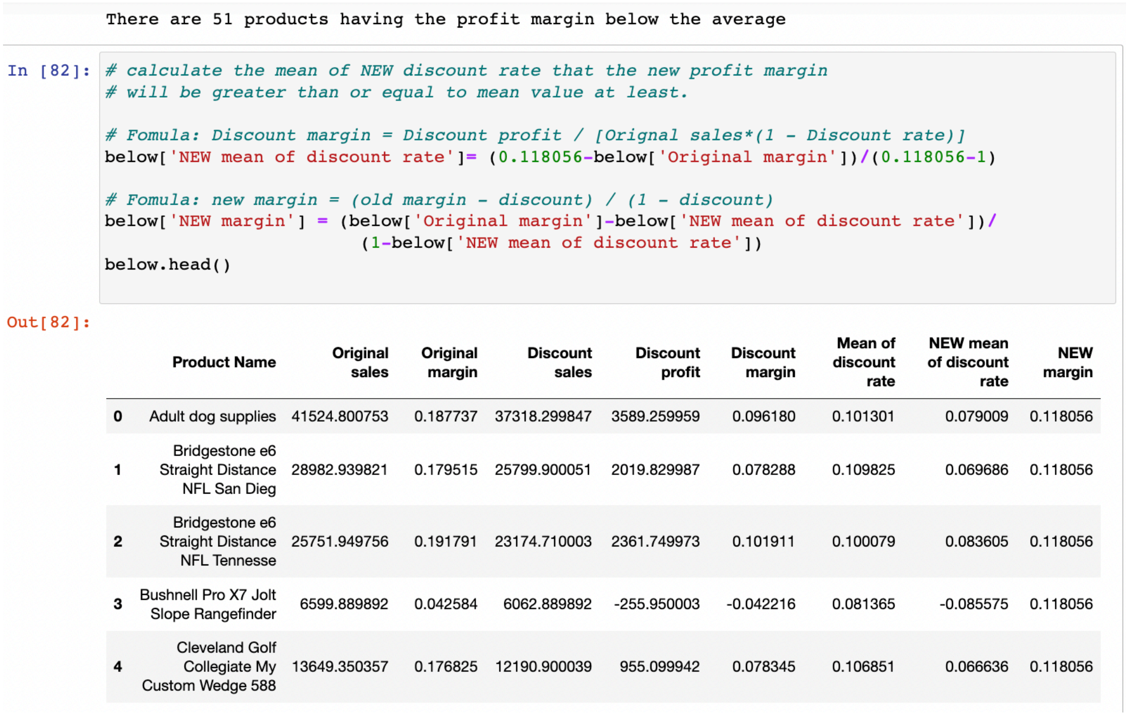
(Much and Pál, 2023)



```
In [78]: combine_1 = sales_org.merge (sales_dis, on='Product Name')
         combine_3 = combine_1.merge (pf_dis, on='Product Name')

         # Calculate margin AFTER discount
         # Fomula: margin = profit/revenue
         combine_3['Discount margin'] = combine_3['Discount profit']/combine_3['Discount sales']

         # calculate discount rate
         # Fomula: Discount margin = discount profit / discount sales [in which discount sale = original
         combine_3['Mean of discount rate'] = 1-(combine_3['Discount profit']/(combine_3[
             'Discount margin']*combine_3['Original sales']))

         # Calculate margin AFTER discount
         # Fomula: org margin = dis margin*(1-discount rate)+discount rate
         combine_3['Original margin'] = combine_3['Discount margin']*
         (1-combine_3['Mean of discount rate'])+combine_3['Mean of discount rate']

         combine_3=combine_3[['Product Name','Original sales','Original margin','Discount sales',
                              'Discount profit','Discount margin','Mean of discount rate']]
         combine_3.head()
```

Out[78]:

	Product Name	Original sales	Original margin	Discount sales	Discount profit	Discount margin	Mean of discount rate
0	Adult dog supplies	41524.800753	0.187737	37318.299847	3589.259959	0.096180	0.101301
1	Baby sweater	12229.560379	0.228724	10957.400143	1525.029992	0.139178	0.104023
2	Bag Boy Beverage Holder	21116.549776	0.250058	19009.969910	3173.780008	0.166953	0.099760
3	Bag Boy M330 Push Cart	16637.919929	0.278002	14981.660008	2969.110042	0.198183	0.099547
4	Bowflex SelectTech 1090 Dumbbells	5999.899902	0.336469	5171.899902	1190.779995	0.230240	0.138002



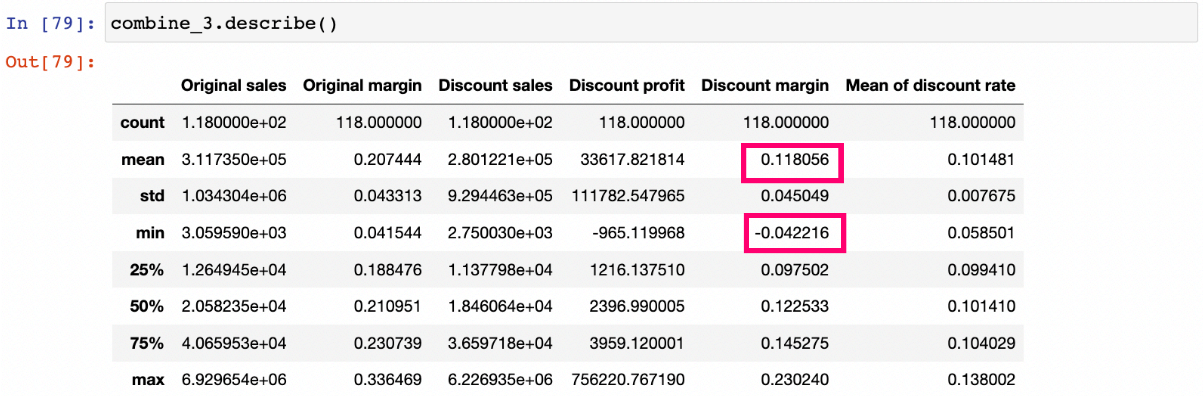
```
There are 51 products having the profit margin below the average
```

```
In [82]: # calculate the mean of NEW discount rate that the new profit margin
         # will be greater than or equal to mean value at least.

         # Fomula: Discount margin = Discount profit / [Orignal sales*(1 - Discount rate)]
         below['NEW mean of discount rate']= (0.118056-below['Original margin'])/(0.118056-1)

         # Fomula: new margin = (old margin - discount) / (1 - discount)
         below['NEW margin'] = (below['Original margin']-below['NEW mean of discount rate'])/
                                (1-below['NEW mean of discount rate'])
         below.head()
```

Out[82]:

	Product Name	Original sales	Original margin	Discount sales	Discount profit	Discount margin	Mean of discount rate	NEW mean of discount rate	NEW margin
0	Adult dog supplies	41524.800753	0.187737	37318.299847	3589.259959	0.096180	0.101301	0.079009	0.118056
1	Bridgestone e6 Straight Distance NFL San Dieg	28982.939821	0.179515	25799.900051	2019.829987	0.078288	0.109825	0.069686	0.118056
2	Bridgestone e6 Straight Distance NFL Tennesse	25751.949756	0.191791	23174.710003	2361.749973	0.101911	0.100079	0.083605	0.118056
3	Bushnell Pro X7 Jolt Slope Rangefinder	6599.889892	0.042584	6062.889892	-255.950003	-0.042216	0.081365	-0.085575	0.118056
4	Cleveland Golf Collegiate My Custom Wedge 588	13649.350357	0.176825	12190.900039	955.099942	0.078345	0.106851	0.066636	0.118056

Appendix 10: Description table of data frame to check the min and mean value.

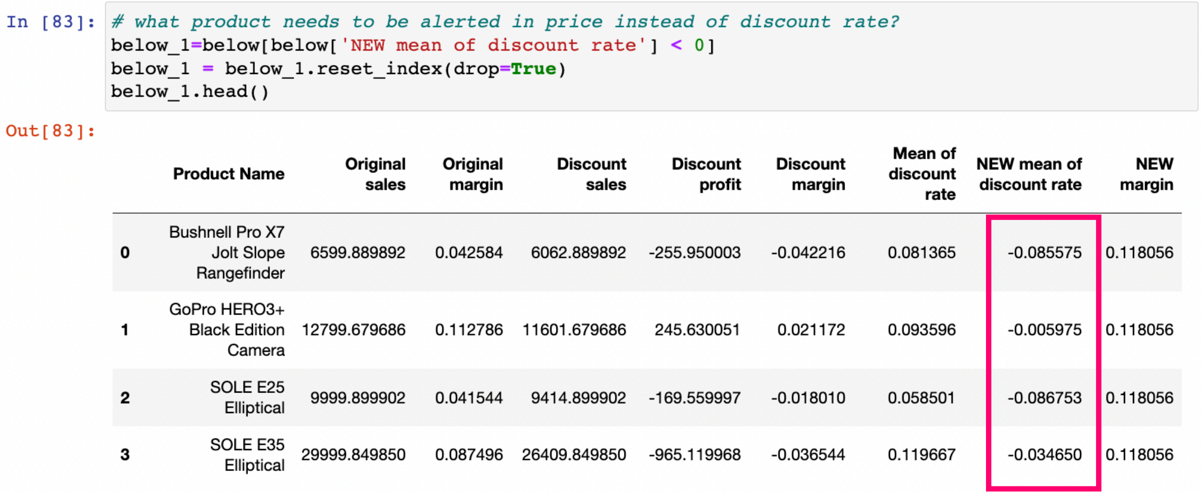


```
In [79]: combine_3.describe()
```

Out[79]:

	Original sales	Original margin	Discount sales	Discount profit	Discount margin	Mean of discount rate
count	1.180000e+02	118.000000	1.180000e+02	118.000000	118.000000	118.000000
mean	3.117350e+05	0.207444	2.801221e+05	33617.821814	0.118056	0.101481
std	1.034304e+06	0.043313	9.294463e+05	111782.547965	0.045049	0.007675
min	3.059590e+03	0.041544	2.750030e+03	-965.119968	-0.042216	0.058501
25%	1.264945e+04	0.188476	1.137798e+04	1216.137510	0.097502	0.099410
50%	2.058235e+04	0.210951	1.846064e+04	2396.990005	0.122533	0.101410
75%	4.065953e+04	0.230739	3.659718e+04	3959.120001	0.145275	0.104029
max	6.929654e+06	0.336469	6.226935e+06	756220.767190	0.230240	0.138002

Appendix 11: Table to check any product need to be focused on price rather than just discount rate

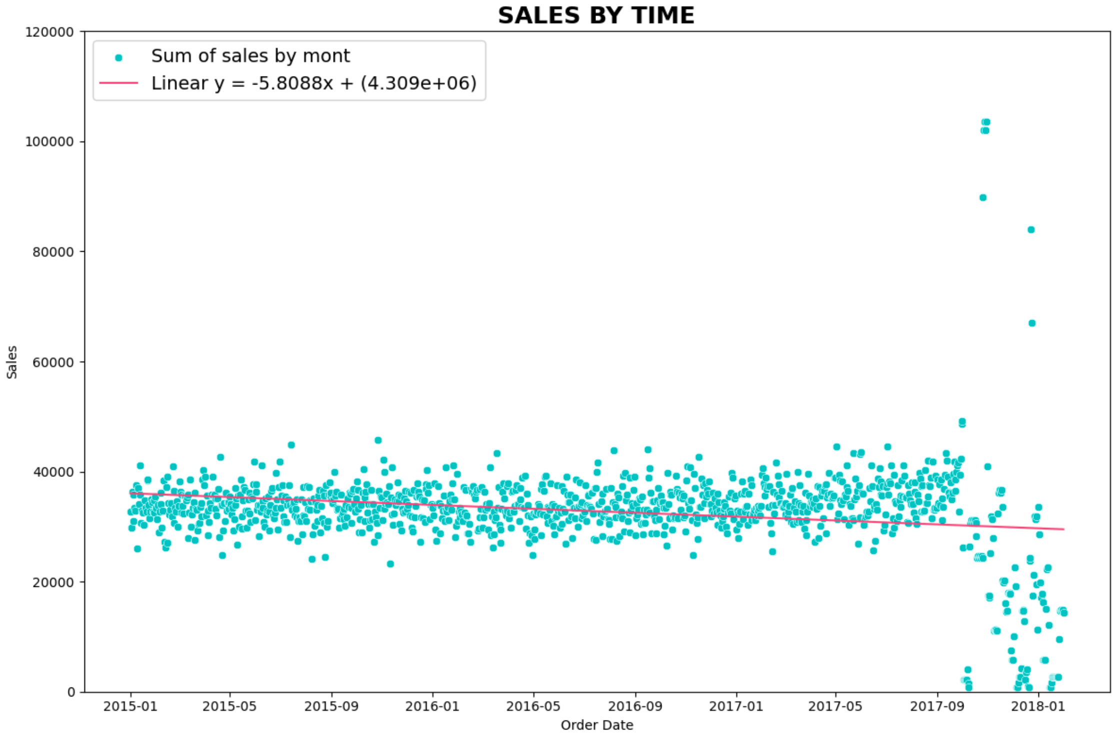


```
In [83]: # what product needs to be alerted in price instead of discount rate?
         below_1=below[below['NEW mean of discount rate'] < 0]
         below_1 = below_1.reset_index(drop=True)
         below_1.head()
```

Out[83]:

	Product Name	Original sales	Original margin	Discount sales	Discount profit	Discount margin	Mean of discount rate	NEW mean of discount rate	NEW margin
0	Bushnell Pro X7 Jolt Slope Rangefinder	6599.889892	0.042584	6062.889892	-255.950003	-0.042216	0.081365	-0.085575	0.118056
1	GoPro HERO3+ Black Edition Camera	12799.679686	0.112786	11601.679686	245.630051	0.021172	0.093596	-0.005975	0.118056
2	SOLE E25 Elliptical	9999.899902	0.041544	9414.899902	-169.559997	-0.018010	0.058501	-0.086753	0.118056
3	SOLE E35 Elliptical	29999.849850	0.087496	26409.849850	-965.119968	-0.036544	0.119667	-0.034650	0.118056

Appendix 12: Prediction sales throughout periods of time



Appendix 13: Description of customer segmentation in term of RFM score.

- Top valuable customer: Customers who bought most recently, most frequently, and spent the most

- Loyal customer: Most recently purchased customer

- Potential loyalist: They often make a purchase but just with a fair amount of spending.

- Rookie: They have a recent purchase but not frequently.

- Promising: Have a recent purchase but with a small amount of spending

- Exigent: They perform at the average.

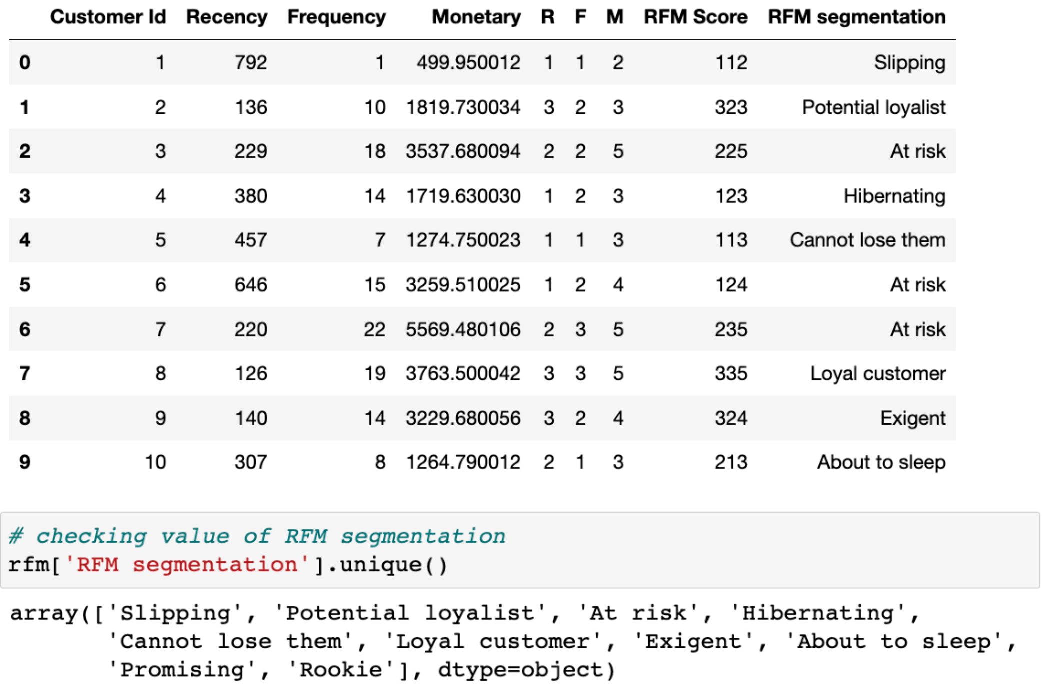
- Cannot lose them: They used to have the high frequency since the last purchase.

- About to sleep: The RFM score is below the average. This group tends to be at alerting level due to the opportunity to contribute to the churn rate.

- At risk: The only problem with this group is the time they making purchase

- Hibernating: Has low recency, frequency and amount of spending.
- Slipping: The last purchase was a long time ago, the transaction amount is small and the number of transactions is small.

Appendix 14: Part of RFM segmentation table and the labels available in it

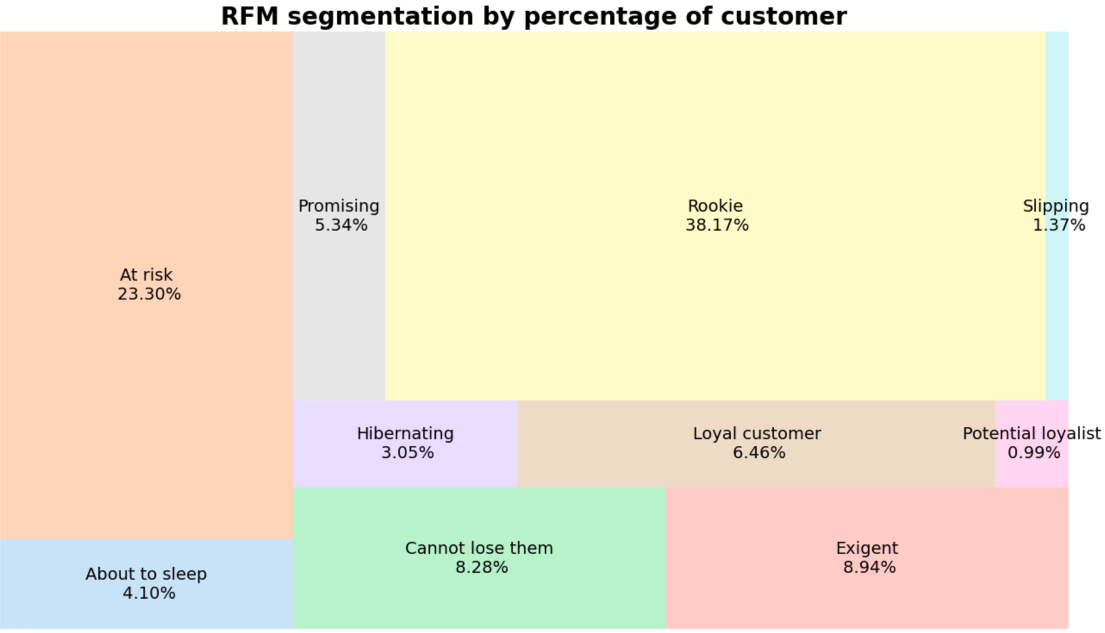


	Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Score	RFM segmentation
0	1	792	1	499.950012	1	1	2	112	Slipping
1	2	136	10	1819.730034	3	2	3	323	Potential loyalist
2	3	229	18	3537.680094	2	2	5	225	At risk
3	4	380	14	1719.630030	1	2	3	123	Hibernating
4	5	457	7	1274.750023	1	1	3	113	Cannot lose them
5	6	646	15	3259.510025	1	2	4	124	At risk
6	7	220	22	5569.480106	2	3	5	235	At risk
7	8	126	19	3763.500042	3	3	5	335	Loyal customer
8	9	140	14	3229.680056	3	2	4	324	Exigent
9	10	307	8	1264.790012	2	1	3	213	About to sleep

```
# checking value of RFM segmentation
rfm['RFM segmentation'].unique()
```

```
array(['Slipping', 'Potential loyalist', 'At risk', 'Hibernating',
       'Cannot lose them', 'Loyal customer', 'Exigent', 'About to sleep',
       'Promising', 'Rookie'], dtype=object)
```

Appendix 15: RFM segmentation by percentage of customer



END OF REPORT.