Supermarket sales

Exploratory Data Analysis for Machine Learning - Course Project

Caio Mescouto Terra de Souza

May 14, 2021

Brief Description of the Data Set and a Summary of its Attributes

Supermarket sales data set [1] is a historical record of sales data in 3 different Branches for 3 months. Each observation has 17 features that indicates where the purchase happen (Branch, City), when (Date and Time), the Product Line, Unit Price, Quantity, Payment method, Cost of goods sold, margins, Taxes and Customer informations (Gender and Type) in addition to stratification rating on their overall shopping experience.

The data set has 1000 observations and no missing values. The variables are presented below (data dictionary and types of attributes):

- **Invoice id** (categorical - object): Computer generated sales slip invoice identification number
- **Branch** (categorical - object): Branch of supercenter (3 branches are available identified by A, B and C).
- **City** (categorical - object): Location of supercenters
- **Customer type** (Categorical - object): Type of customers, recorded by Members for customers using member card and Normal for without member card.
- **Gender** (categorical - object): Gender type of customer
- **Product line** (Categorical - object): General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
- **Unit price** (numerical - float64): Price of each product in $
- **Quantity** (numerical - int64): Number of products purchased by customer
- **Tax** (numerical - float64): 5% tax fee for customer buying
- **Total** (numerical - float64): Total price including tax
- **Date** (categorical - object): Date of purchase (Record available from January 2019 to March 2019)
- **Time** (categorical - object): Purchase time (10am to 9pm)
- **Payment** (numerical - float64): Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)
- **COGS** (numerical - float64): Cost of goods sold
- **Gross margin percentage** (numerical - float64): Gross margin percentage
- **Gross income** (numerical - float64): Gross income

- **Rating** (numerical - float64): Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Initial Plan for Data Exploration

At a first glimpse on the data set we found some constraints that will guide the plan for Data Exploration.

Main constraints:

- Branch and City is the same information, meaning that we have only one information about geographical position;
- Tax and gross margin have no variance, meaning that all product lines have the same tax and gross margin, so all analysis based on Cost of Goods, Income, Margins and Taxes are meaningless;
- The record available has only 3 months of data, so the seasonality that can be explored are just day of the week and hour of the day.

Considering the restrictions presented above, the **promising starting points** for exploratory data analysis are highlighted below. These starting points were thought as actionable tasks that can improve the company result.

- Does the rating differ between branches? If so, what hypotheses can explain and guide a solution?
- Do branches have different mean tickets? If so, why?
- Does the mean ticket differ between day of the week or hour?
- Does the rating differ between day of the week or hour?

Finally, the **Roadmap for exploring the data** is:

- Data cleaning: drop all meaningless variables and looking for outliers that can problematic attributes;
- Data transformation: At the first moment only Date and Time will be merged;
- Data exploration: Get to know the numerical variables distributions and highlight the transformations needed. Get to know the categorical variables, the number of categories in each one and if it can be meaningful;
- Correlations: The next step is to analyse the correlations between numerical variables before the feature engineering process;
- Feature Engineering;
- Draw the insights and hypotheses;

- Test the most promising hypothesis;
- Suggest next steps for further investigations according with was highlighted during the exploration;
- Final comments and conclusions.

Data Cleaning and Feature Engineering

Data Cleaning

The data set does not have missing data or outliers that could indicate problematic attributes or observation. As stated before some attributes were dropped and Date and Time was merged as timestamps. After the data cleaning process the data set has the following attributes (Table 1)

DatetimeIndex: 1000 entries, 2019-01-05 13:08:00 to 2019-02-18 13:28:00

Column	**Dtype**	**Total var. or range**
Branch	object	3
Customer type	object	2
Gender	object	2
Product line	object	6
Unit price	float64	10.08 - 99.96
Quantity	int64	1 - 10
Total	float64	10.68 - 1042.65
Payment	object	3
Rating	float64	4 - 10

Table 1: Data Set Attributes

Feature Engineering

First of all, as we are mainly interested in the differences between branches, we observed that the distribution between them is almost a third part each, and the same behavior is observed for all categorical attributes. Besides that, as presented before, the number of variables of each attribute is small. In addition neither of them are ranked, meaning that all are good candidates for One Hot Encoding as feature engineering for machine learning preprocessing.

Considering the numerical attributes, only a few transformations seem worth it. "Total" is the total invoice price and it is clearly right skewed (Figure 1) and a log transformation is a good way to better fit the variable for machine learning algorithms. The rest of the numerical variables are equally distributed, so only a standardization is required to improve the fit in the models.

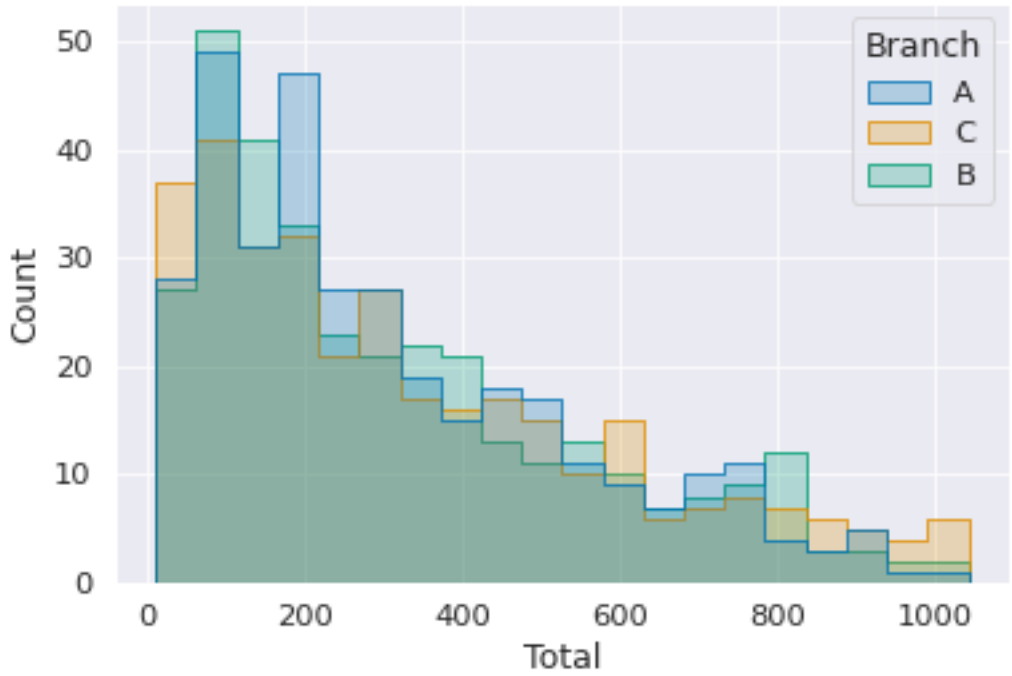

Figure 1: Histograms per Branches

Key Findings and Insights

Besides the *Total* distribution (Figure 1) per branches that present a similar behavior, the mean and the median are higher in C followed by B and A (Table 2). This difference is most explained by the *Unit price* of the products as de mean and median quantity are fairly the same through the branches. However, these differences do not seem significant as the standard deviation is very high.

Statistic	**A**	**B**	**C**
mean	312.35	319.87	337.10
median	240.83	252.88	271.42
std	231.64	242.45	263.16

Table 2: Total per Branches

There is no correlation between *Total* and *Rating* neither in all data sets nor by branches, as the A and B are similarly rated and the B has the lowest *Rating* between them (Figure 2). Again the significance seems not large enough, but these first findings are good guides for the next steps.

Analyzing*Branch* per *Gender* was observed that A has more men invoices and C women ones. We also know that *Female* spend more than *Male*, knowing that this relationship of

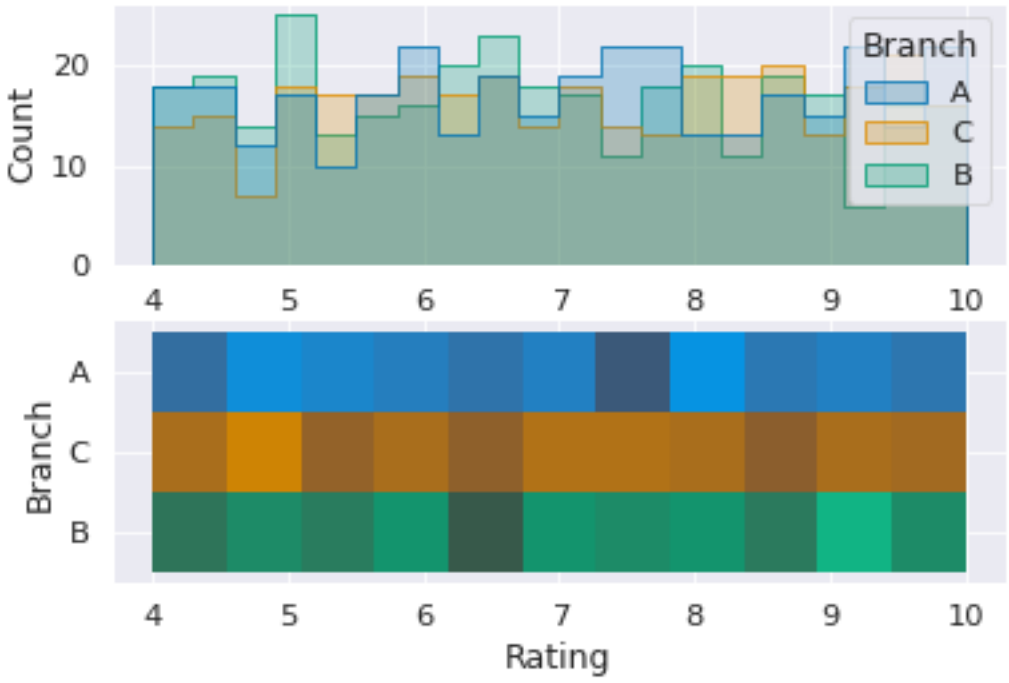

Figure 2: Ratings per Branches

invoice price is the same for *Branch*, this is meaningful, but we cannot explain causation. In addition, we also observed that Membership and payment method follow same distributions. In brief ***Female* are majority among customer members and use more with *Cash*, the same relationship was observed for *C*. *Male* are the majority among normal customers and use more *Ewallet*.**

The exploration of the *Product line* brought interesting insights because we can observe variance between *Product line* and *Branch* (Figure 3) and *Gender* (Figure 4). First of all **Women are the major buyers of *Sports and travel***, but *C*, where they are the majority of clients performs badly in this line. however, looking further the reality is that, **this *Branch* performs badly with men in this line, in quantity of invoices and average ticket compare others.** But as the rating is high (Table 3), one hypothesis is that this *Branch* doesn't have more expensive products for men. We need to investigate further to confirm this hypothesis, but a suggested action could be to expand the range of these products in this *Branch*.

The second insight is ***Male* are the major buyer of *Health and beauty***, but they buy more in *B* and *C* and with high ticket, even if they are the majority of clients in *A*. The same hyphotesis and suggestion presented before is applied here (Table 5).

The third insight is that **men buy less *Fashion* articles in *A* than *C* and the rating is also higher in *C* and this is also true for women.** It would be related to the quality of the products in *A*.

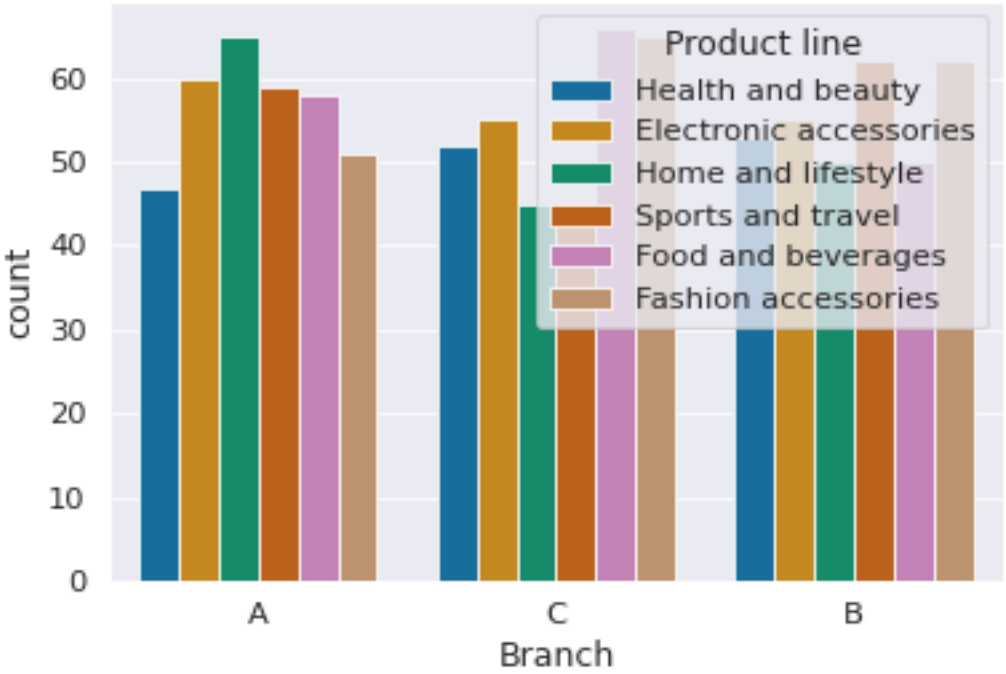

Figure 3: Product lines per Branches

Branch	Gender	Count	Mean	Rating
A	Female	29	279.83	7.16
	Male	30	375.25	7.35
B	Female	30	307.09	6.37
	Male	32	336.73	6.64
C	Female	29	387.82	6.79
	Male	16	282.20	7.47

Table 3: Sports and travel per Branch and Gender

Finally, we can infer women spend more in *Food and Bevarage* and *Home and Lifestyle* because usually they are responsible for maintaining the home, and men spend more in *Health and beauty* because they use more popular products that are found in supermarkets.

Moving on the seasonality aspect, we can observe daily variation in all branches, however, *A* seems less extreme (Figure 5). In fact, *A* seems more stable than *B* (Figure 6). In addition, we also analyze seasonality per hour(Figure 7. ***B* has a high traffic at 13h and 19h during working days** .We also noticed that the movement is high on Saturdays and is the branch that sells the most products in *Sports and Travel*. It can indicate that it is close to workplaces and people go there after work or at lunch to buy a few things, but they also have a good line of *Sport and Travel*, which motivates customers to move around on

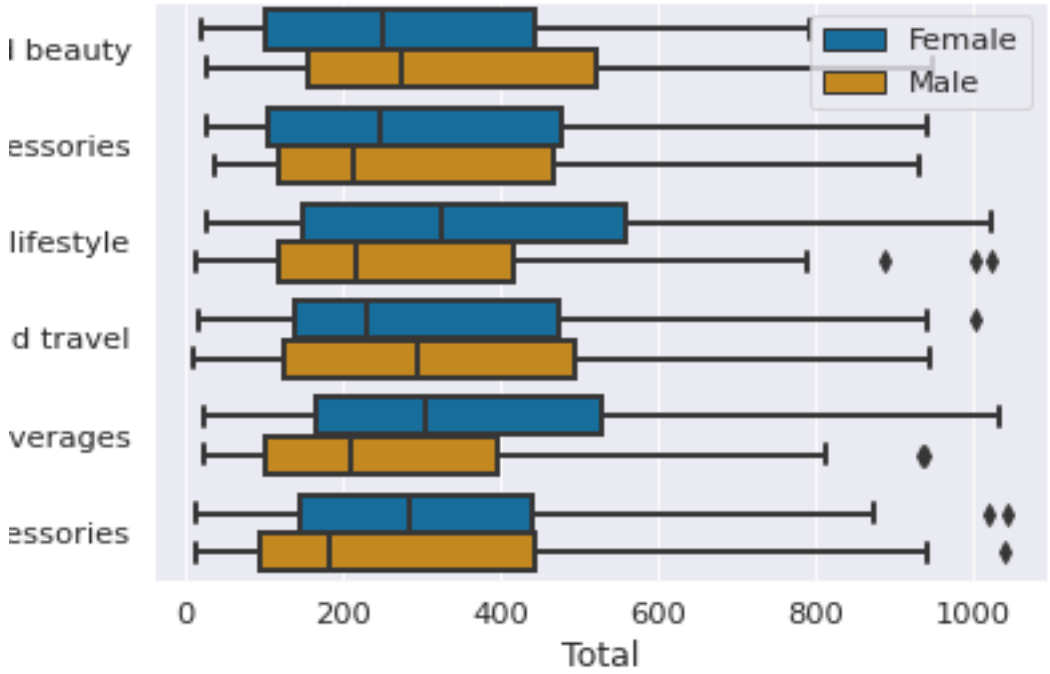

Figure 4: Boxplot Product line per Gender

Branch	Gender	Count	Mean
A	Female	21	272.14
	Male	26	264.73
B	Female	20	320.02
	Male	33	411.52
C	Female	23	280.25
	Male	29	350.68

Table 4: Health and beauty per Branch and Gender

Branch	Count	Rating
A	51	6.88
C	65	7.44

Table 5: Health and beauty per Branch and Gender

Saturdays .

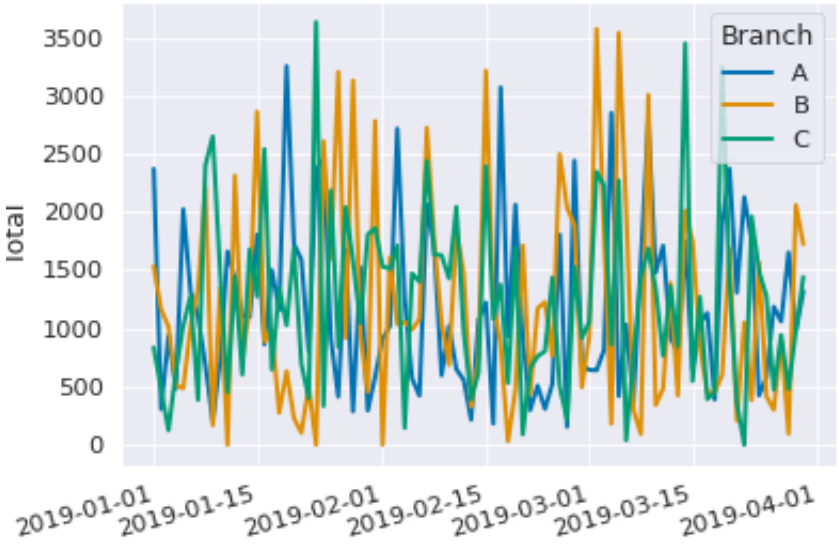

Figure 5: Sum of Total per Day

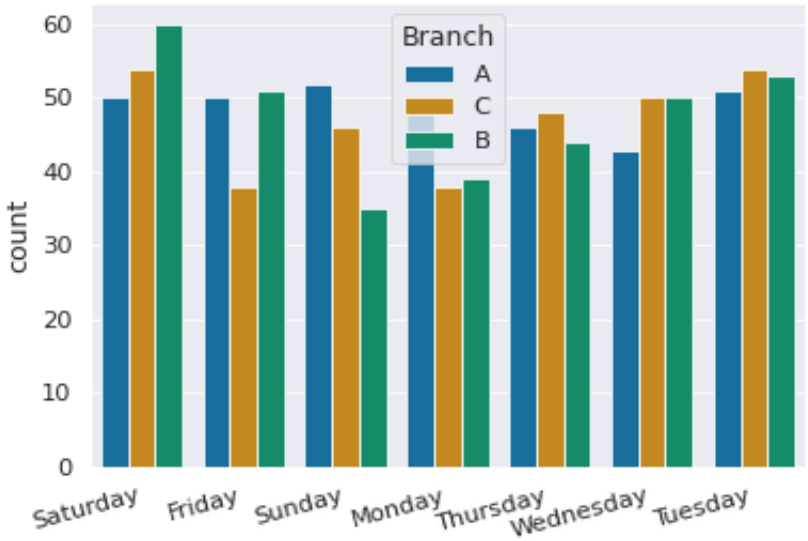

Figure 6: Invoices per day of the week

Main Hypotheses

Considering the insights presented in the preceding section, we can sintetize some hypotheses.

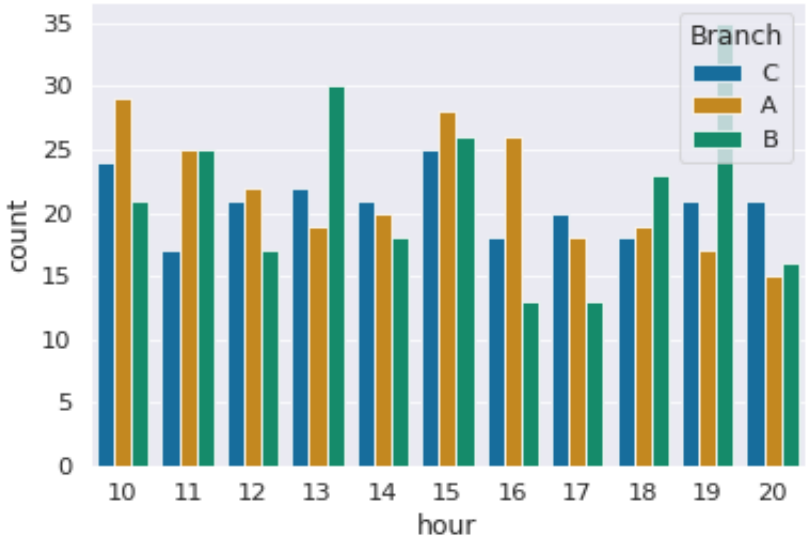

Figure 7: Invoices per hours (working days)

- Considering *Sport and Travel*, male invoices in C are significantly lower than invoices in A;

- Men are more satisfied buying *Fashion accessories* in C than in A;

- B sells more during working days at lunch time (13:00) and 19:00 than A and C.

Significance Test

To test one of those hypotheses, a formal significance test was carried out for the third hypothesis raised. As the distributions are not normal, but have similar shape (Figure 8), the *Mann-Whitney* non-parametric test was applied and the result is presented below.

$\alpha = 0.05$
H_0 : The sum of Total during working days and during 13:00 and 19:00 is the same at all three Branch
H_1 : B sells more during working days at lunch time (13:00) and 19:00 than A and C.
$p - value = 0.01 < \alpha$
$statistic = 28009.0$

As the $p - value$ is smaller than α, the null hypothesis can be rejected.

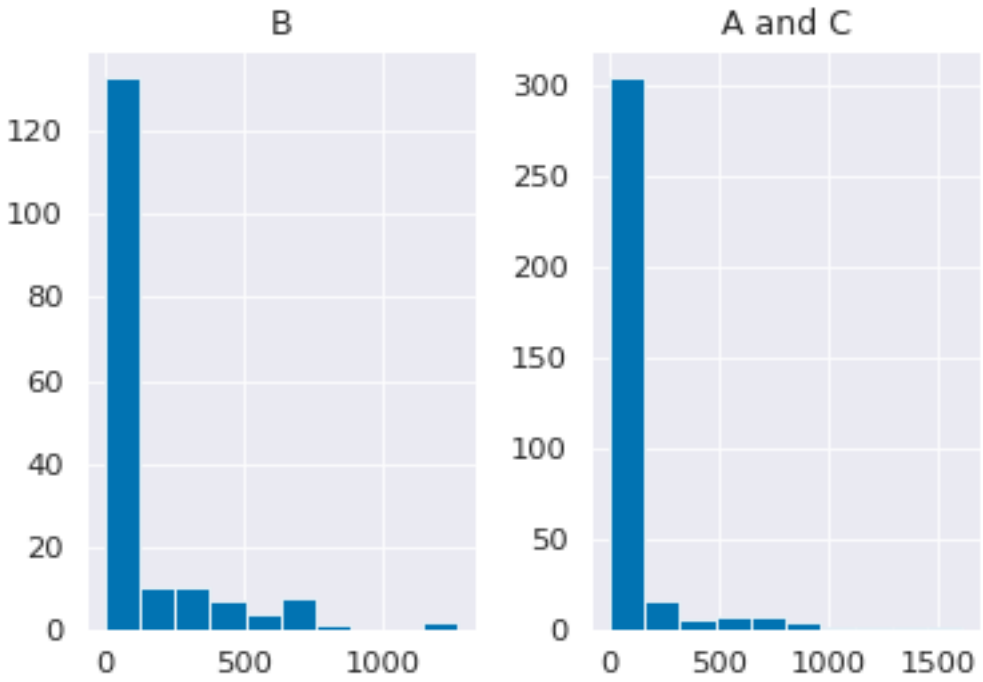

Figure 8: Total histogram

Next Steps

Seasonalities can be a good topic to investigate more, the relationship with ratings on days of the week and hours can indicate some issues, for example few employees helping customers during rush hours, or a specific employee group that are not well trained in customer assistance.

Other interesting topic is the *Product line*, as we found some differences between *Gender* and *Branch* that can indicate the necessity of adjust in some lines depending on the *Branch*.

Coments About the Data Set

The data set has a time constraint for example for forecasting demand, another problem is the level of information related to products, only the big lines are available and further investigations on consumptions are impossible. In addition, Customer ID can help with suggestion algorithms. Finally, the information about Tax and Cost are meaningless and any analysis on costs and earnings is impossible from this data set.

References

[1] Supermarket sales: Historical record of sales data in 3 different supermarkets, `https://www.kaggle.com/aungpyaeap/supermarket-sales`