**Data Analysis use case explain the R script**

**Task 1.** According to the diagram we see that there is a relationship between the product sold to the customer and his age.

The age range of the clients is wide and it starts from a young age and ends in old age (13-90).

All 3 products have a relationship to the different ages of the customers but it is important to note that the stronger relationship is towards natural products. This can be seen by looking at the density of the red dots in all 3 products the density is higher than the density of the blue dots. You can also see a wide bulk that indicates consumption by customers in a large age range.

According to the density of the red dots we can see that customers of all ages consume more natural masks, then natural soaps and finally a natural bath lotion.

Graphical user interface, qr code

Description automatically generated

**Task 2**. The models obtained for transactions with customers from the USA and for transactions with Swedish customers

Text

Description automatically generated

Based on the characteristics of the models obtained, it can be seen that the CFO's claim is incorrect.

The model that best explains the relationship is the model for transactions with **Swedish customers.**

This can be seen according to the Adjusted R-squared index. In the Swedish model, the value of the index is 0.789 compared to 0.7834, this value is closer to 1, so it can be said that the explanatory ability of the Swedish model is slightly better than the American model.

According to these data, it can be seen that these are good models, with a high ability to explain that indicates that there is a direct and strong connection between the variables. The value received by the index tends to be 1 and higher than 0.6, which indicates a strong bond strength and a high correlation between the variables.

In the Swedish model it can be seen that the point of intersection with the Y-axis indicates a loss created as a result of the transaction and is very statistically significant in that the value obtained is a number very close to 0 (\*\*\* - 0). We can therefore be sure that this figure is very accurate and should be relied upon.

It is also advisable to rely on the variable Machine\_482 which indicates a profit generated as a result of the transaction. Here too you can see an excellent level of significance because it is very close to 0. This is a very significant variable which contributes greatly to the explanatory capacity of the model.

The significance level of the variable Tan\_39 is given the value 0.00585, which indicates a good level of significance and a number that tends to 0. (\*\* - 0.001)

The other variables tend to be 1 and are not completely statistically significant.

The model as a whole is very statistically significant. We can see this by looking at the p-value parameter which gets a very low value close to 0.

In the American model on the other hand it can be seen that the point of intersection with the Y-axis and only the variable Machine\_482 get an excellent level of significance very close to 0. The other variables tend to be 1 and are not completely statistically significant. Thus it can be said that the American model is less good because there are fewer significant variables. The model as a whole is very significant and the p-value parameter gets a very low value close to 0.

**Task 3.** **Predictive models**

A. First, we tried to understand the business problem that means predicting as much as possible.

We started from the premise that we need to choose one goal variable, which we will have to predict.

The reason we chose to predict the **"usage"** variable is because we wanted to predict the number of natural products, because in section 1 we came to the conclusion that customers prefer to purchase natural products over products that contain artificial materials.

Based on the claim we have made, if we can predict the amount of products that contain natural ingredients we can indicate an indirect relationship in the potential increase or decrease, in the future profit / loss of the company.

In the second step, we performed some operations in the code in order to understand the data we are going to work on, and immediately afterwards we used commands from different libraries such as: (caret, dplyr) to prepare the data for the model. (We decided to predict the usage variable using 3 independent variables).

Before we chose the models we will work on, we divided the data as follows: 70% went to the **training group** where we will perform all the various calculations.

And the remaining 30% that we will not "touch" at all that belong to the **test group** and we would like to "predict" them later.

Next, we selected two algorithms of forecasting models: "logistic regression" and "decision trees". And in each of them we tried to predict whether a product is natural.

.(The prediction of course was made on the test group)

B. The best model we have built and more accurate in predicting the right cases (TP) is: a logistical regression model. (Better than the decision tree model we built).

The reason we chose this particular model for the more successful is that, based on the accepted prediction metrics: specificity, and sensitivity it yielded better results. For example: in the **sensitivity index** that shows how much the model was able to predict the positive cases (in our case an identification of some of the natural products that the model predicted to be natural and in fact they are natural) **the result was 1.**

Which means that surprisingly, the model was able to predict the positive cases perfectly. However, the model was unable to predict the negative cases at all. (specificity=0)

C.Text, letter

Description automatically generated

Text

Description automatically generated

D. The meaning of the index values ​​in the chosen model:

**The accuracy index** according to the error matrix gained an accuracy of almost 89% (0.8846), which means that the overall level of accuracy of the model was particularly high.

**The sensitivity index** (parallel to the precision index we mentioned earlier reached the result 1. (Maximum level of accuracy for the positive cases)

**The Specificity Index** (which we also mentioned earlier) reached a result of 0. This means that the model was unable to predict the non-natural products at all.

(It can be said that because the **Specificity Index** came out particularly low, the model itself is not perfect and not even close to it. But since most of the products in the dataset are so natural, and our main goal in the first place was to identify the natural products we can say we succeeded in this task).