PREDICTING CATALOG DEMAND REPORT

## Person Holding Blue and Clear Ballpoint Pen

JENS GAULKE

13.01.2021

UDACITY BERTELSMANN SCHOLARSHIP

# Executive Summary

Last year our company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. We have 250 new customers in our their mailing list that we want to send the catalog to. Can we predict the expected profit from this new customers? We only want to send out the catalog to these customers if the expected profit contribution exceeds $10,000.

## Recommendation:

Our model predicts a revenue of **$21,987** taking into account a gross margin of 50% and deducting the printing cost of $6.5 dollars per catalog copy. This is more than the required $10,000. We should definitely send out the catalog to the new customers.

The following pages show the steps which led to the conclusion. In detail, these steps are

* Business and Data Understanding
* Analysis, Modeling, and Validation
* Presentation/Visualization

# Business and Data Understanding

We need to make a decision whether or not our company should invest in a marketing campaign which includes sending catalogues out to 250 new customers. The management states that the campaign has to exceed the expected profit contribution of $10,000 to be executed. The prediction therefore needs to calculate the expected profit in order to make the decision of printing and sending the catalogues to the new customers or not.

To calculate the profit, we use the equation ***profit = revenue – cost***. To perform the calculation, we have to go some steps:

* Calculate the expected revenue from the new 250 customers considering the probability that a customer will buy if we send the catalog to him.
* Calculate the costs for the catalog and set 50% for the gross margin
* Subtract the costs from the revenue.
* If the profit is greater than $10,000, send the catalog to the new customrs

We will need this data for our prediction:

|  |  |
| --- | --- |
| Expected profit | expected profit from catalog-induced sales – costs of printing and distributing |
| costs of printing and distributing | given as $6.50 per catalog |
| expected profit from catalog-induced sales | Expected total revenue from catalog-induced sales \* average gross margin |
| average gross margin | Given as 50% (0.5) |
| expected total revenue from catalog-induced sales | expected sales volume per customer \* probability of buying per customer |
| probability of buying per customer | given as [Score\_Yes] in the dataset |
| expected sales volume per customer | to be predicted in the linear regression model |

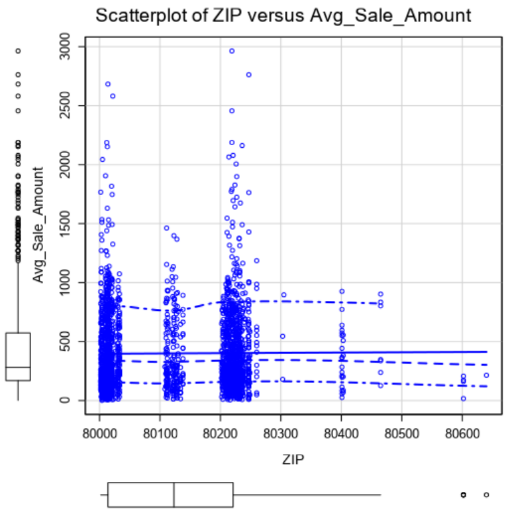
# Analysis, Modeling, and Validation

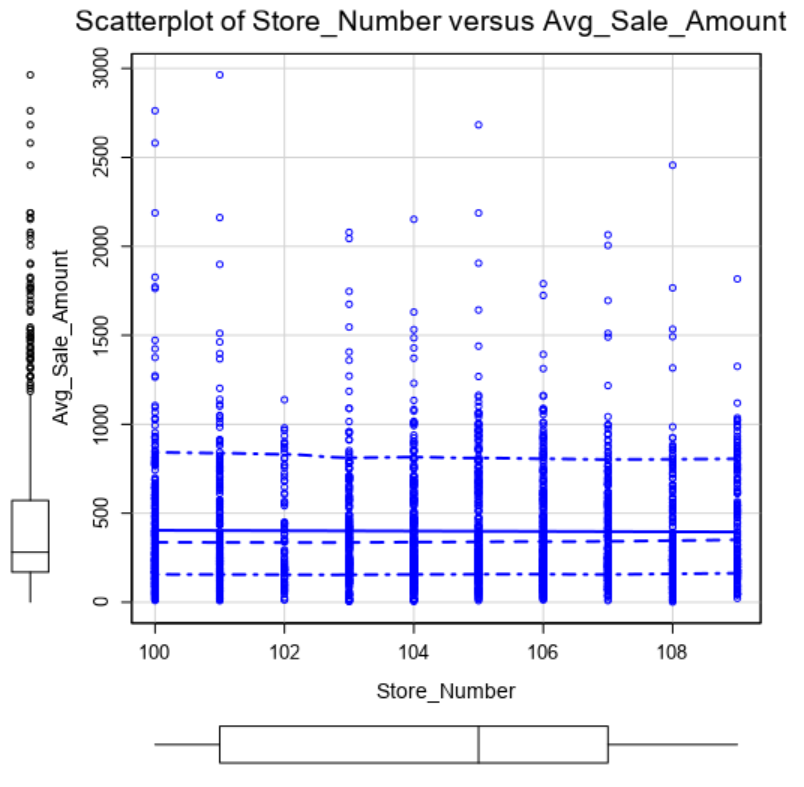
We will use a linear regression model to predict the expected sales volume for each customer. Thus, the target variable (y-axis) will be **Avg\_Sale\_Amount**.

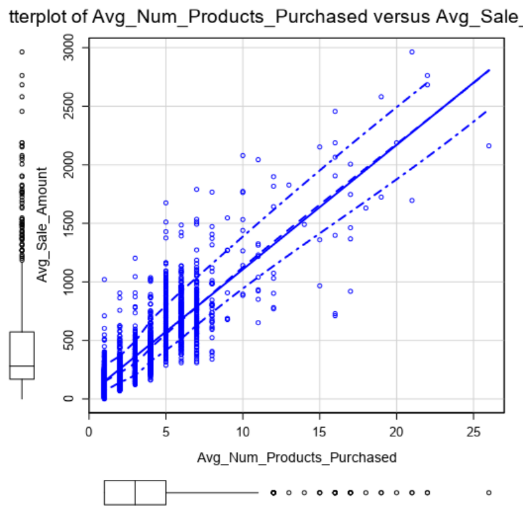
To make a prediction, the appropriate predictor variables from the dataset “p1-customers.xlsx” have to be chosen. The target variable **Avg\_Sale\_Amount** can be excluded, leaving us with eleven data columns to be considered. During the initial assessment, the following columns can directly be excluded:

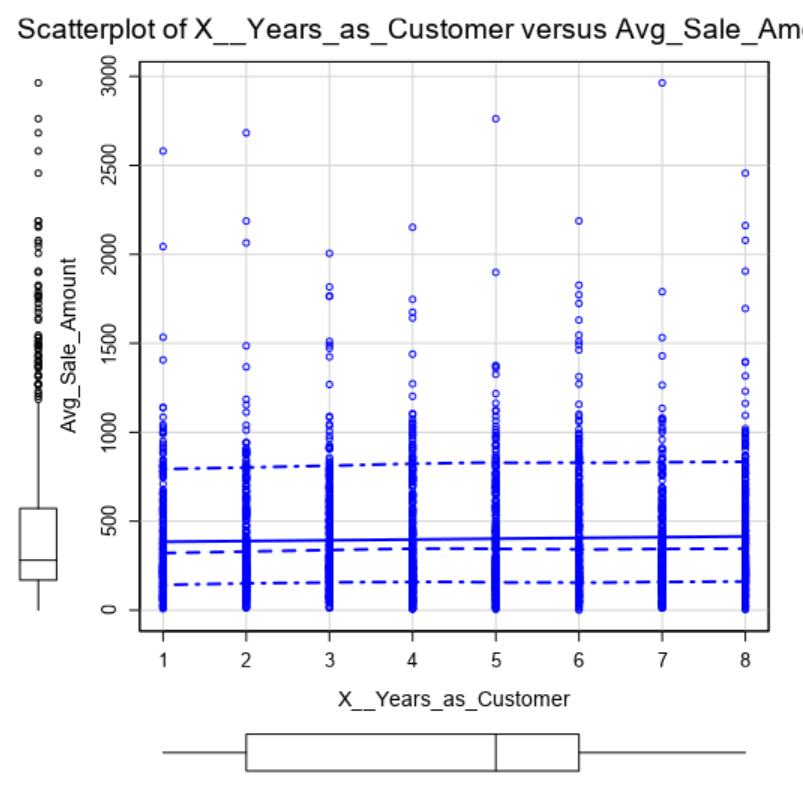
* Name, Address, ID unique by definition, so not of predictive nature
* State all customers are from CO
* City we have a ZIP also
* Responded\_to\_last\_catalog not available in new dataset used for the prediction

This leaves us with **#\_Years\_As\_Customer**, **ZIP**, **Avg\_Num\_Products\_Purchased** and **Store\_number** as possible candidates for the linear regression. The next step was to create scatterplots to identify if a linear relationship is likely to be statistically significant.



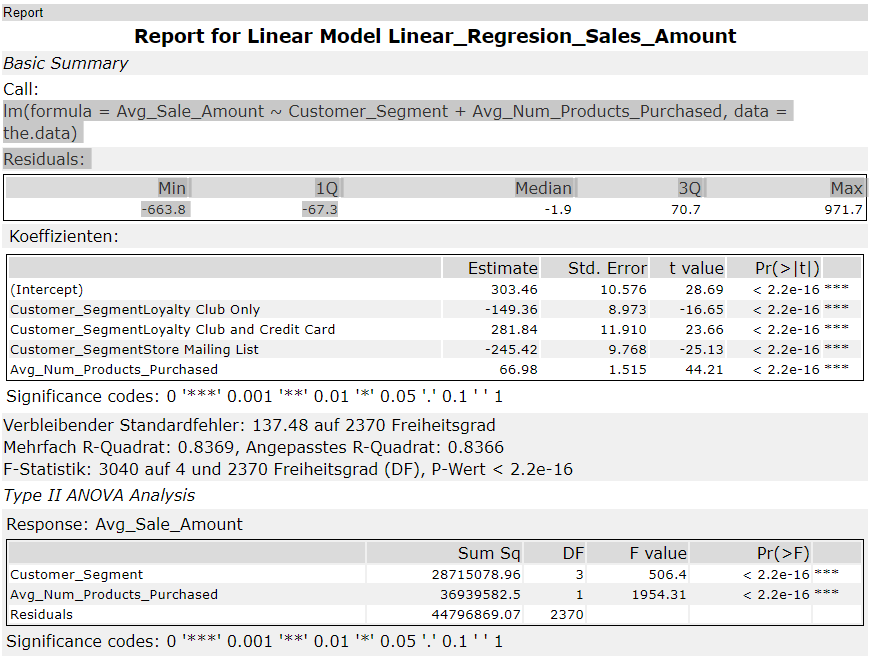






Considering the scatterplots, we see that only for **Avg\_Num\_Products\_Purchased** there seems to be a linear relationship. So we will assume that only for this variable a statistically significant correlation exists.

Setting up a linear regression with alteryx using **Avg\_Num\_Products\_Purchased** gave a good result showing that our assumptions were correct. We tested the nominal/categorical variables like City and Customer\_Segment as well alteryx’ linear regression and found out that **Customer\_Segment** is relevant for our prediction, while City is not.



With these result we built the model. Both predictor variables Customer\_Segment as well as Avg\_Num\_Products\_Purchased have p-values below 2.2e-16, making them statistically significant. The adjusted R-squared-value is .8366, so this can be considered a strong model.

Using the coefficients from alteryx’ output, the final linear regression equation is

*Avg\_Sales\_Amount = 303.46 + 66.98 \* (Avg\_Num\_Products\_Purchased) – 149.36 \* (Loyalty\_Club\_Only) + 281.84 \* (Loyalty Club and Credit Card) - 245 \* (Store Mailing List) + 0 \* (Credit Card Only)*

# Presentation/Visualization

GO!

Our Recommendation is to conduct the marketing campaign as planned as the predicted outcome is **$21,987** which is more than double the required $10,000 considered to be a success.

We tested our model with the data provided in the file p1-mailingslist.xlsx. The expected total revenue from the campaign is predicted at **$47.224** in our model.

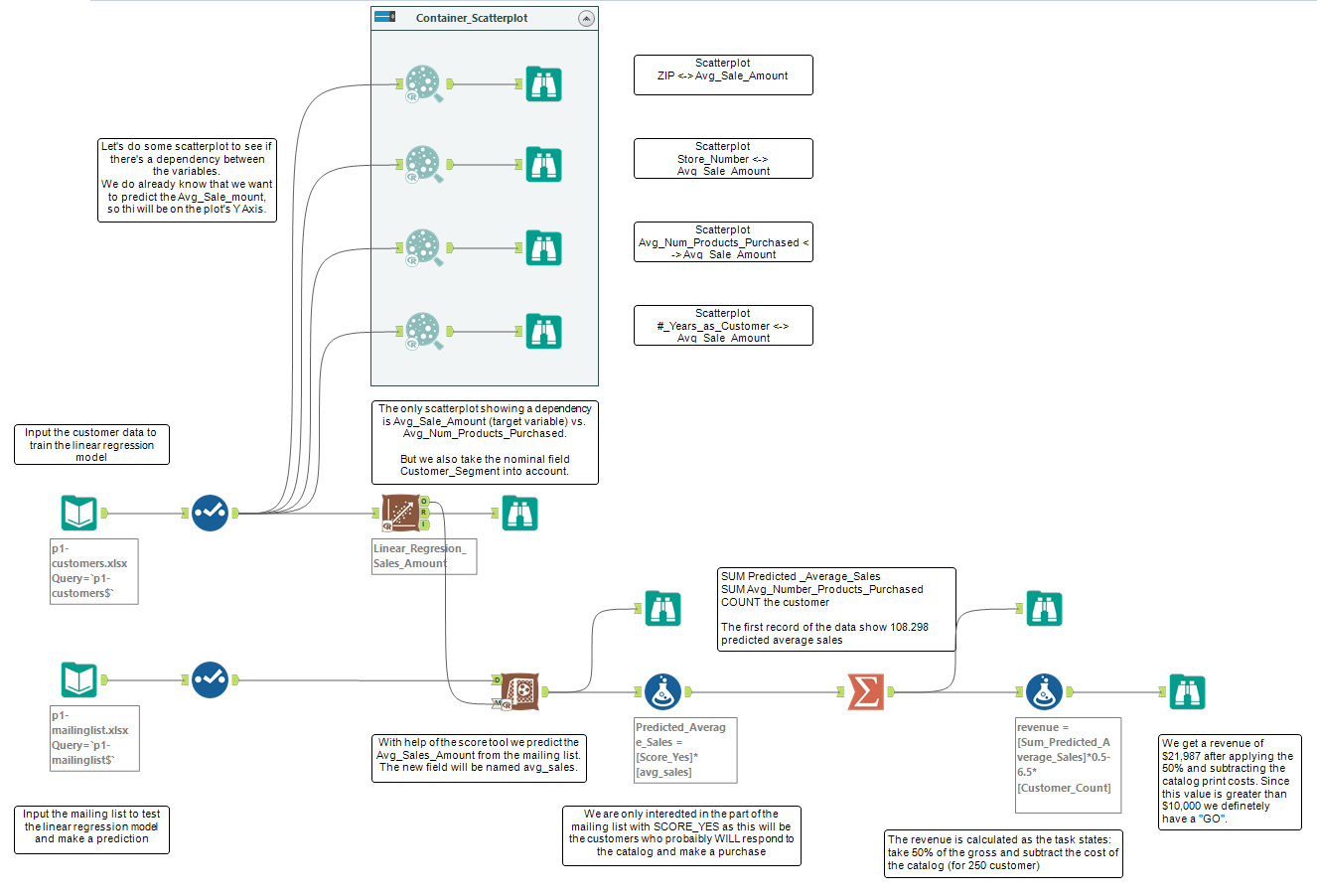
We are supposed to set an average gross margin of 50% (0.5), so the expected profit would be **$23,612**.

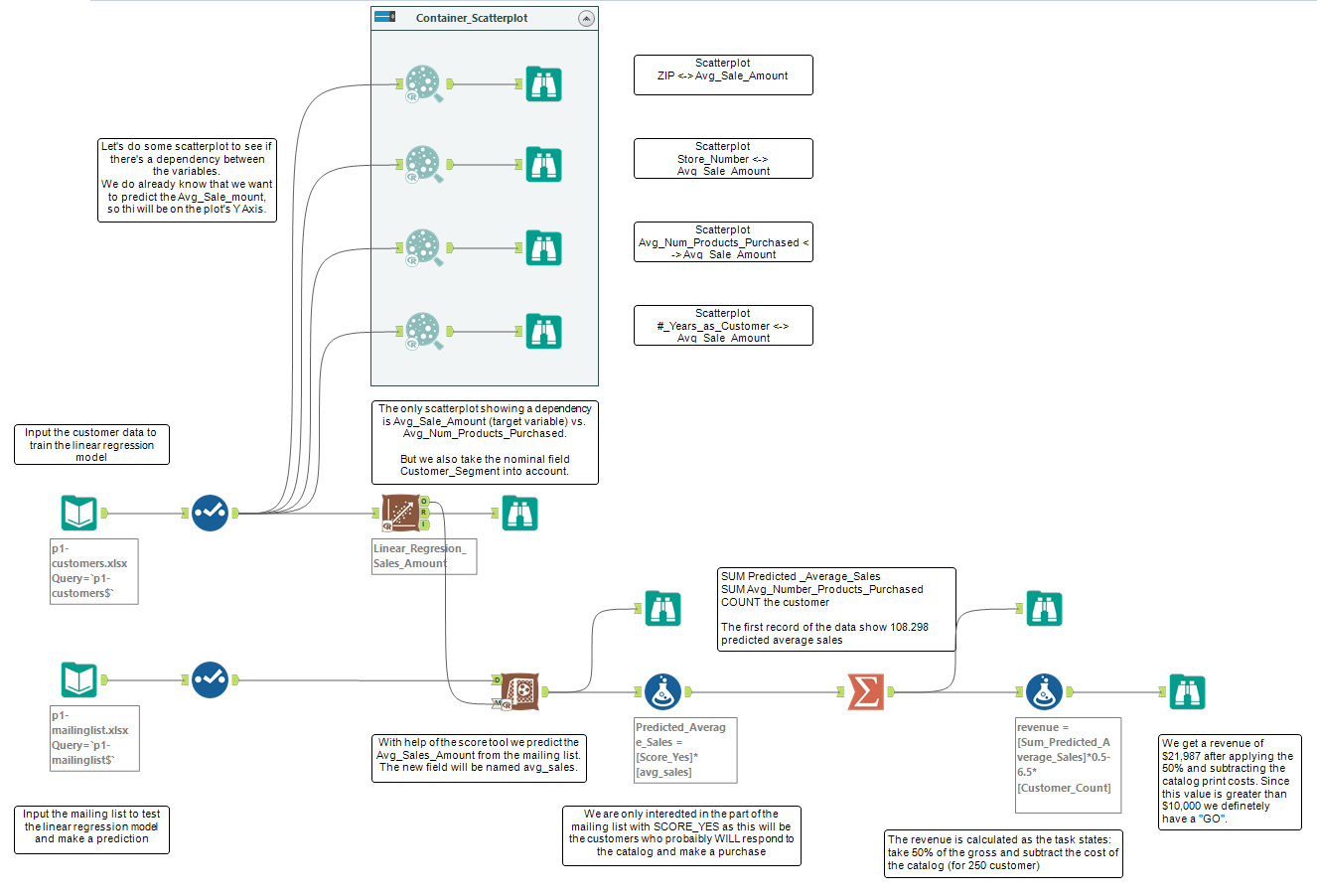
Less the cost of printing and shipping 250 catalogs to the customer which add up to 250 \* $6.5 = **$1,625**,wecalculate the expected profit to be **$21,987.**

If you want to see more details on the model, please have a look at the annexes. Annex 1 shows the alteryx workflow, Annex 2 show the modeling done with Python. As you may see, both models deliver the same result

# Annex 1: Alteryx workflow

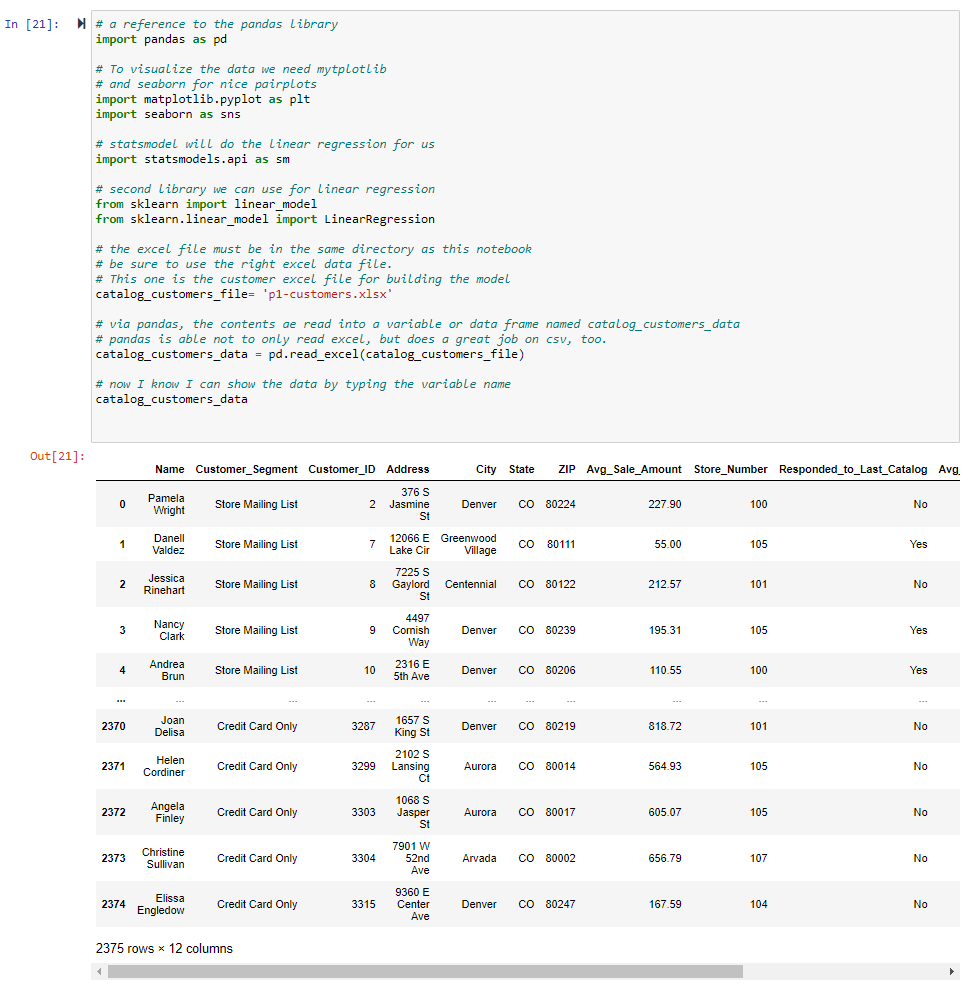
We include the alteryx workflow for your convenience in this annex. It is split up in two illustrations to make it more readable. Every information needed to build your own model is in the comments in the illustrations.

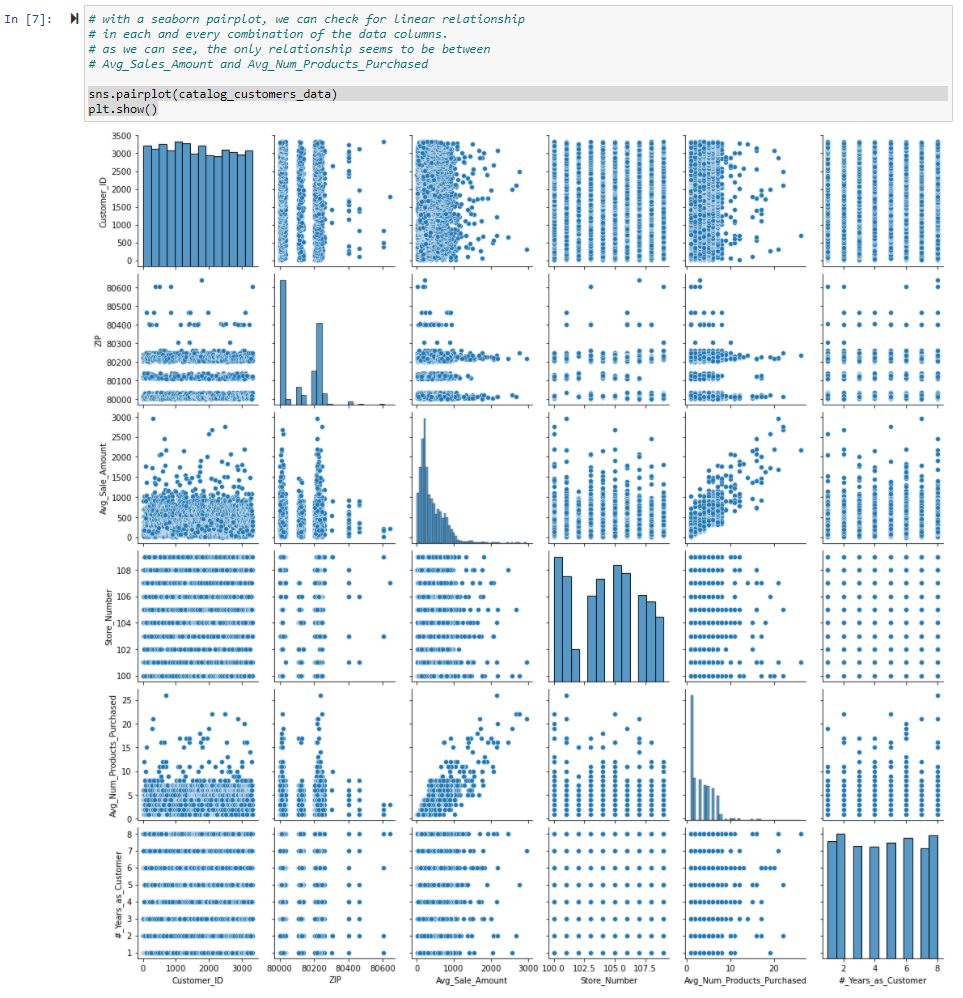


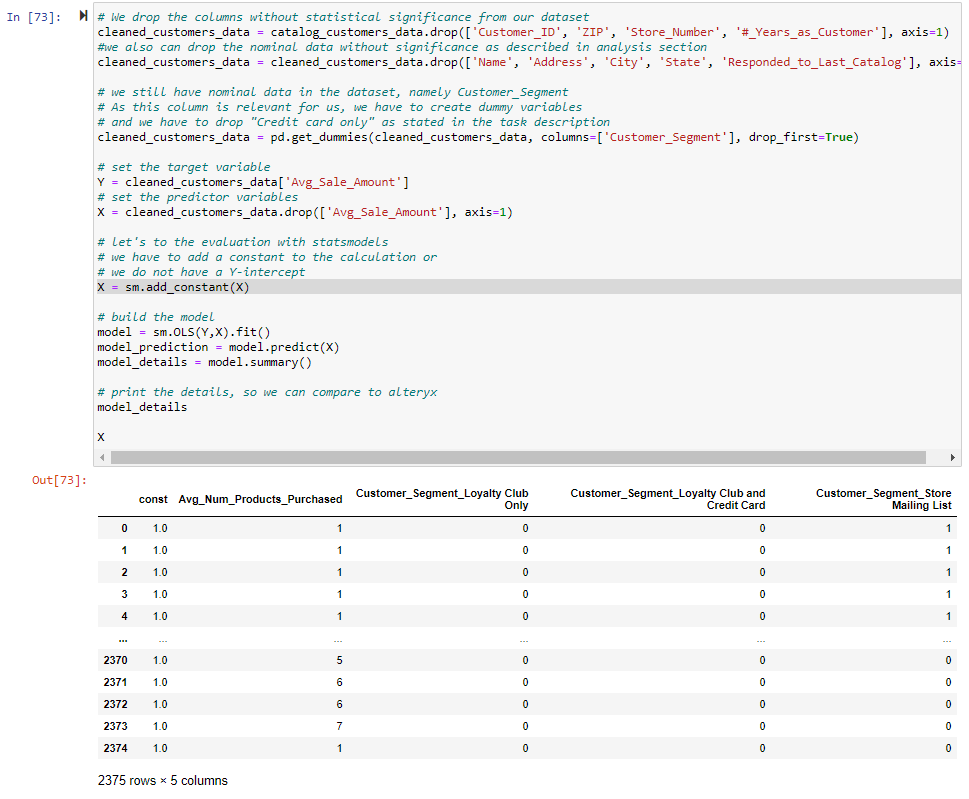


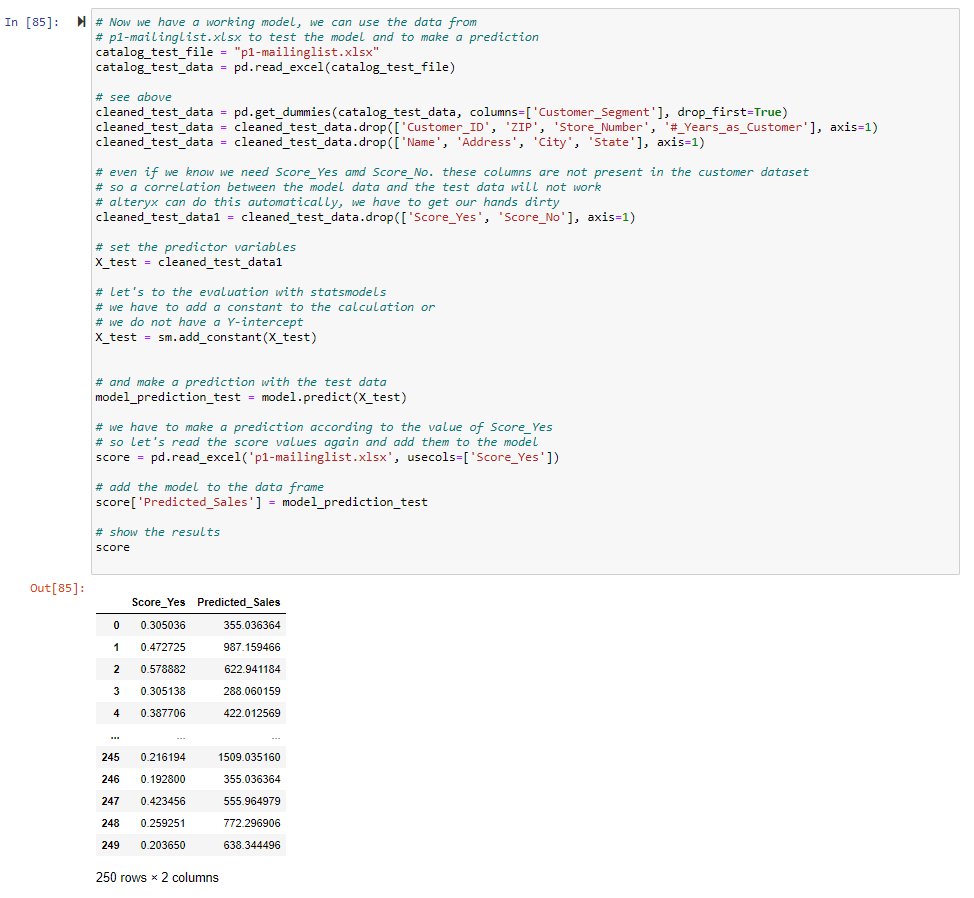
# Annex 2: Python / Jupyter notebook workflow

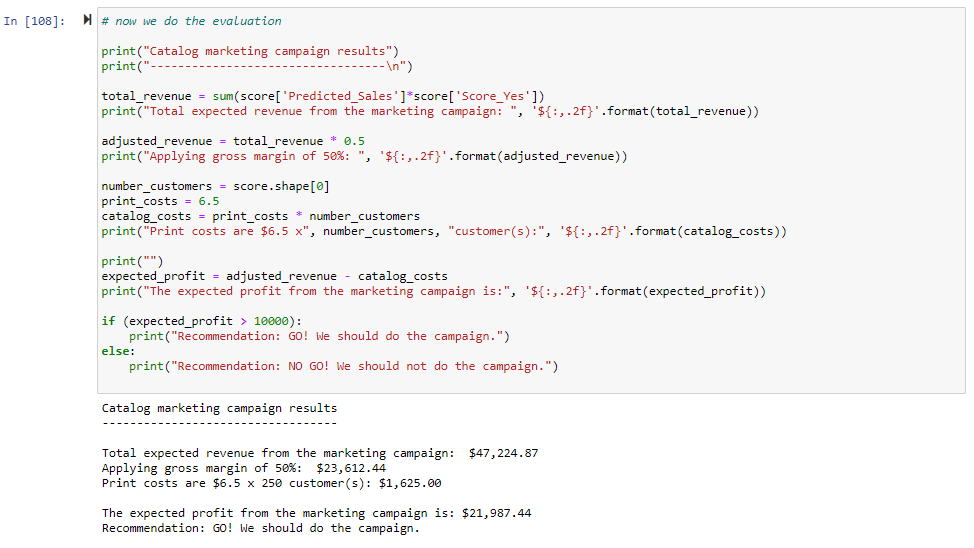
To check if the alteryx prediction is correct, we did another regression with python to prove or disprove the result. This is our workflow:











As expected, Python computed the same results as alteryx.