Target Audience for New Marketing Campaign

October, 2023

This is a case study for a Data Scientist position.

Note: all the scripting codes can be found here.

Goals

1. **Provide meaningful analysis of customers' historical data**
2. **Suggest different approaches to help the Marketing Manager target the right audience for his new campaign**
3. **Raise some questions to improve further recommendations**

1. Data Overview

1. Samples

The data frame contains 537,577 samples from 5,891 unique customers.

2. Data Types

We have a total of 12 variables, 9 of which are categorical, 1 numeric continuous, and 2 ID variables.

3. Variables

It has data about customers' order history, including some demographics, product information, and total revenue.

# user_id	unique identifier of our customers
# prod_id	unique identifier of the product
# sex	F-female, M-male sex
# age_cat	customer age category
# credit_status_cd	determine the credit status code 0-20. These are internal codes with no further explanation.
# education_cat	education category (high-school degree, undergrad of high school, university/college degree)
# years_in_residence	how many years customer stayed at the current residence i.e. 1,2,3, 4+ years
# car_ownership	flag of car ownership
# prod_cat_1	product category class 1 (internal product category classification)
# prod_cat_2	product category class 2 (internal product category classification)
# prod_cat_3	product category class 3 (internal product category classification)
# revenue_usd	total revenue purchase

2. Feature Engineering

1. Missing Values

Two variables presented missing values:

- prod_cat_2: 31,1%
- prod_cat_3: 69,4%

In this case, I simply imputed ‘N/A’ to substitute the missing values.

2. Correcting Data Types

Some variables needed data type changing to be in accordance with their nature:

- prod_cat_1: integer → string
- predit_status_cd: integer → string

3. Transforming

In total, four variables suffered some level of transformation, all of them created to remove any numeric/ordinal effect on these categorical variables:

- prod_cat_1, credit_status_cd: both received an incremental “x”
- car_ownership: 1:'Y', 0:'N'

4. Feature Creation

I created some new variables in order to enrich the analysis (each line being 1 single customer):

- total_orders: the total amount of purchases per customer
- average_ticket: average amount spent per purchase
- total_revenue: gross margin revenue per customer
- revenue_proportion: participation of each customer on total revenue purchase
- acumulative_proportion: revenue_proportion accumulated sum
- revenue_usd_scaled: revenue_usd in min-max scale
- total_orders_scaled: total_orders in min-max scale
- revenue_range: revenue_usd categorized in 5 equal bins

3. EDA

I made some exploratory analyses to understand the variables' behavior and distribution, and also to understand the existence of any correlation regarding the revenue_usd feature, our target/dependent variable of interest.

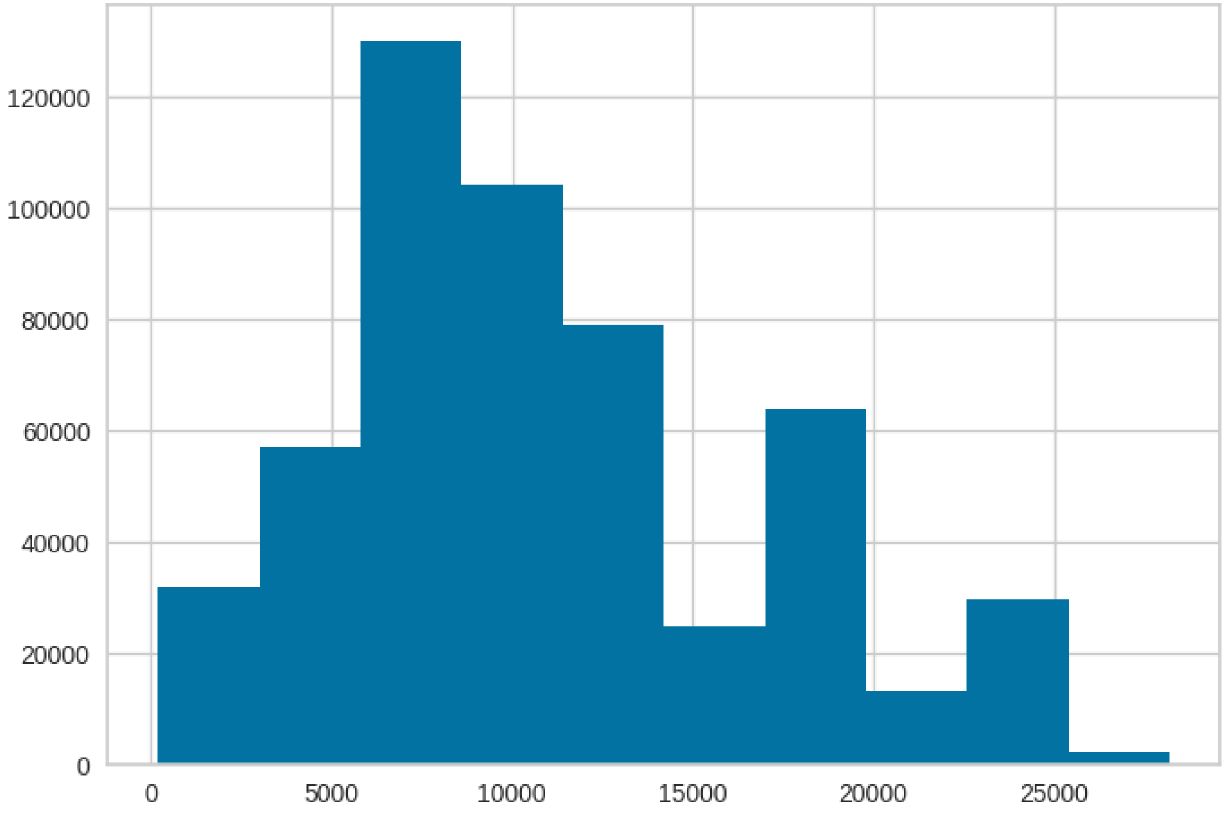
Revenue Distribution

This distribution falls in non-normality given its shape. Besides that, clearly, we have some outliers to the right of the distribution. This is reinforced by looking at the standard deviation being very high compared to the mean/median values, confirming high variance between the values of this particular variable.

Mean: U$ 10.981

Median: U$ 9.484

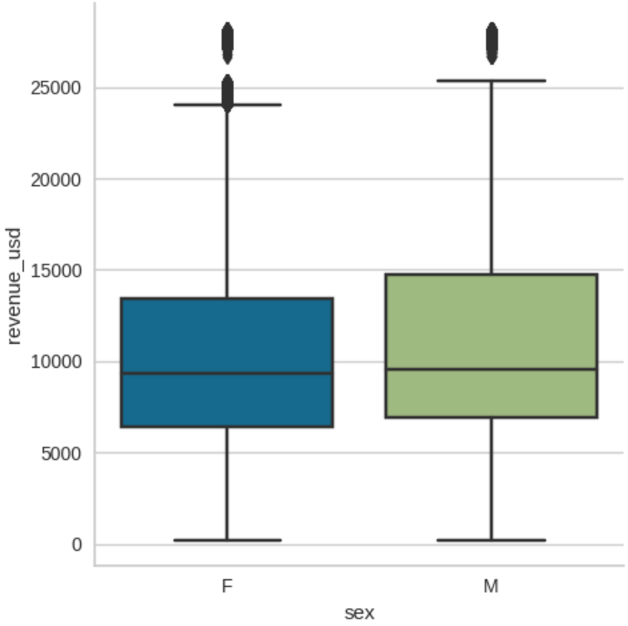
Standart Deviation: U$ 5.860



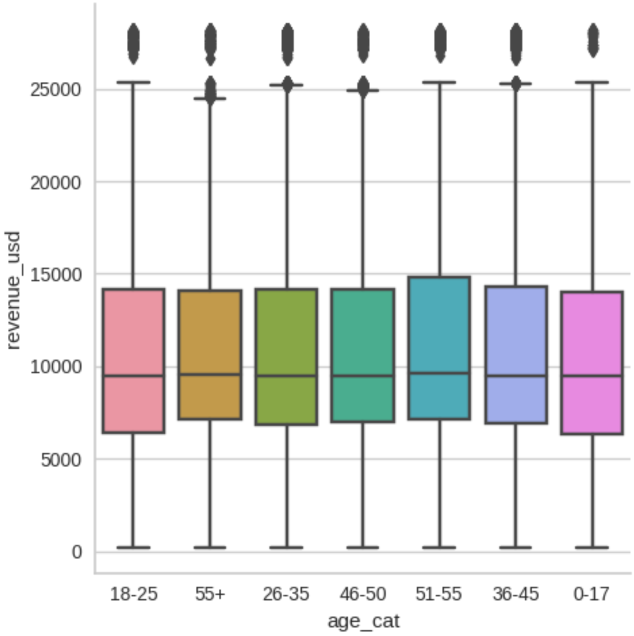
Revenue Correlation

I also looked at the relationship between the dependent variable and the independent variables (revenue vs. the other variables). The goal was to find a high positive or negative correlation between the data.

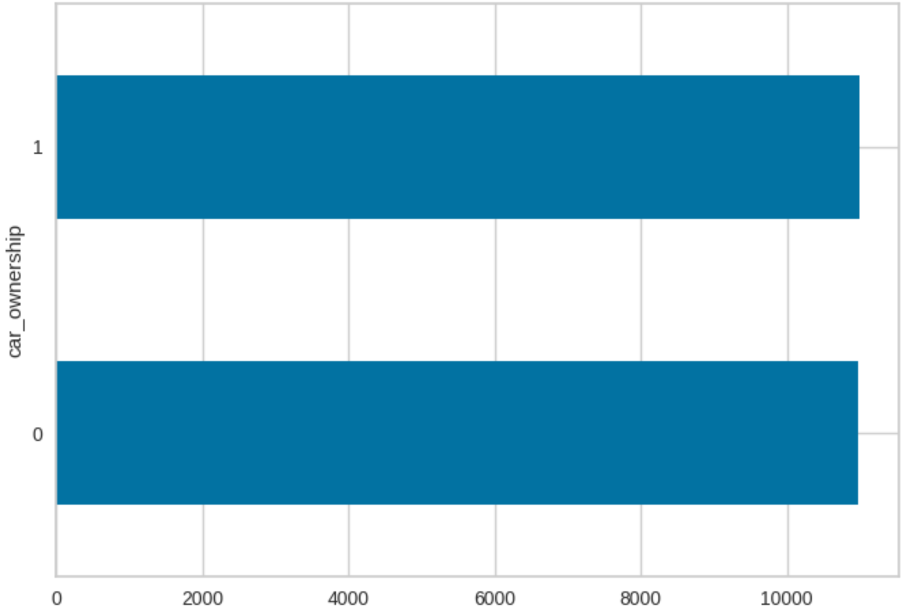
Sex vs. Revenue



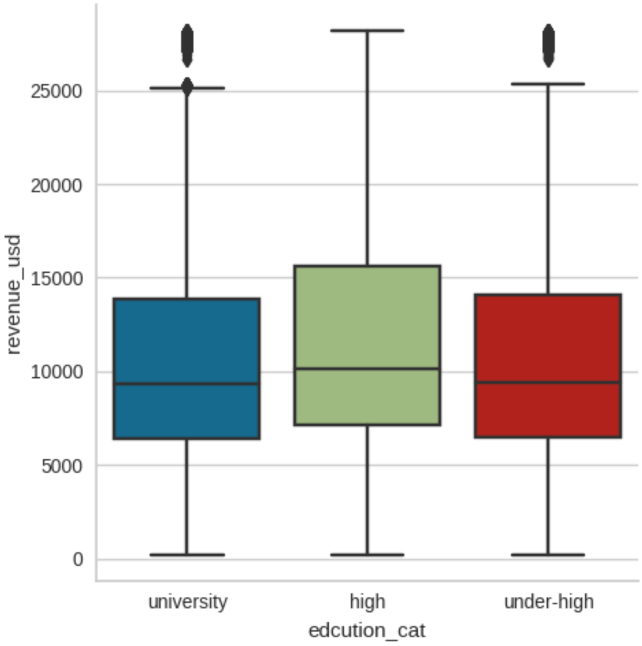
Age vs. Revenue



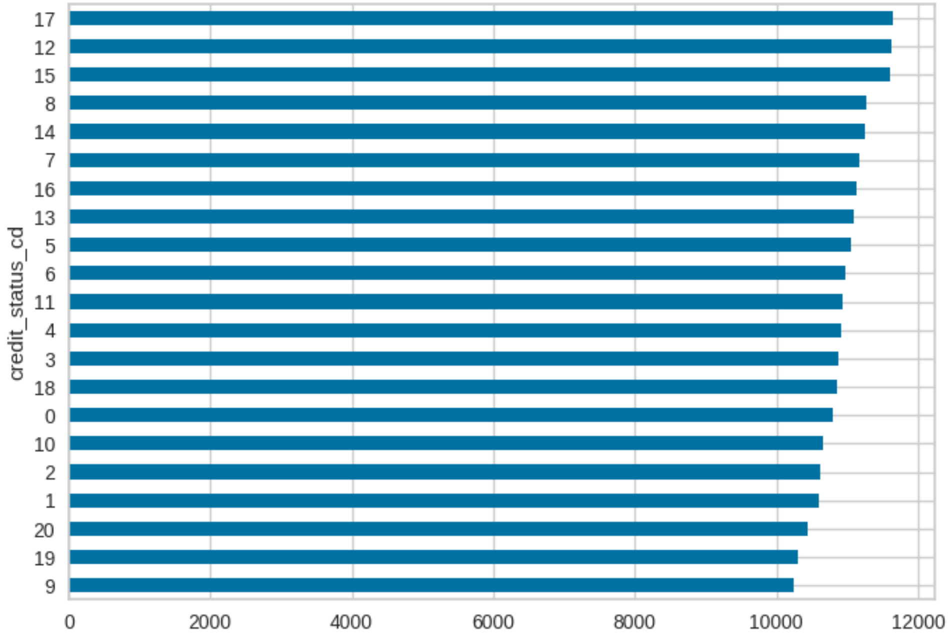
Car ownership vs. Revenue

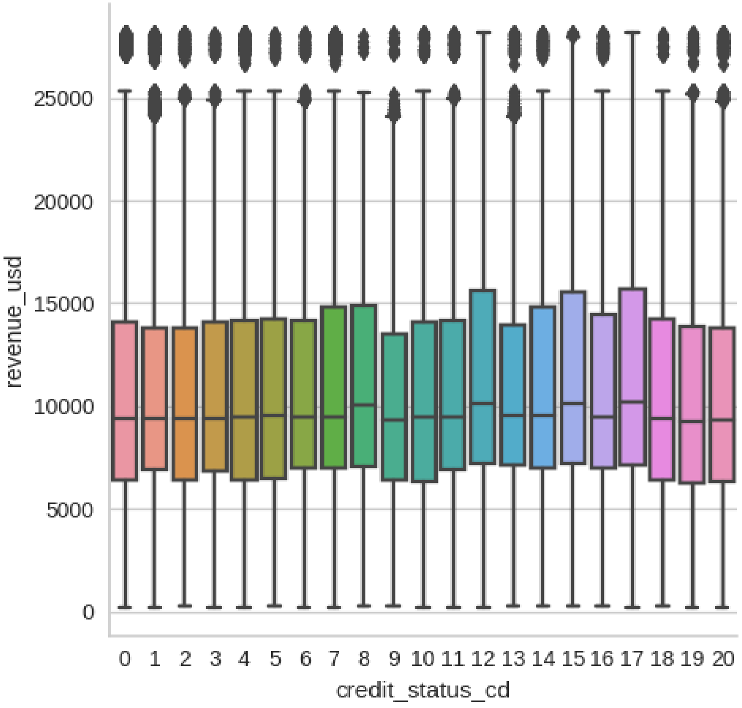


Education vs. Revenue

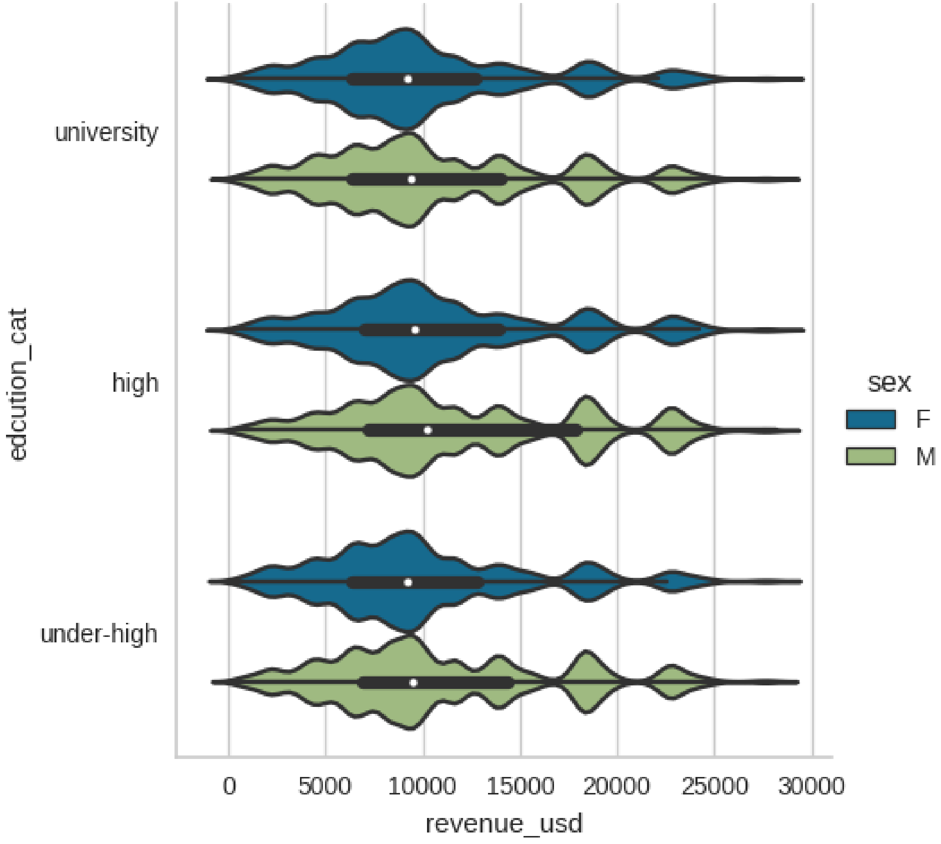


Credit Status vs. Revenue



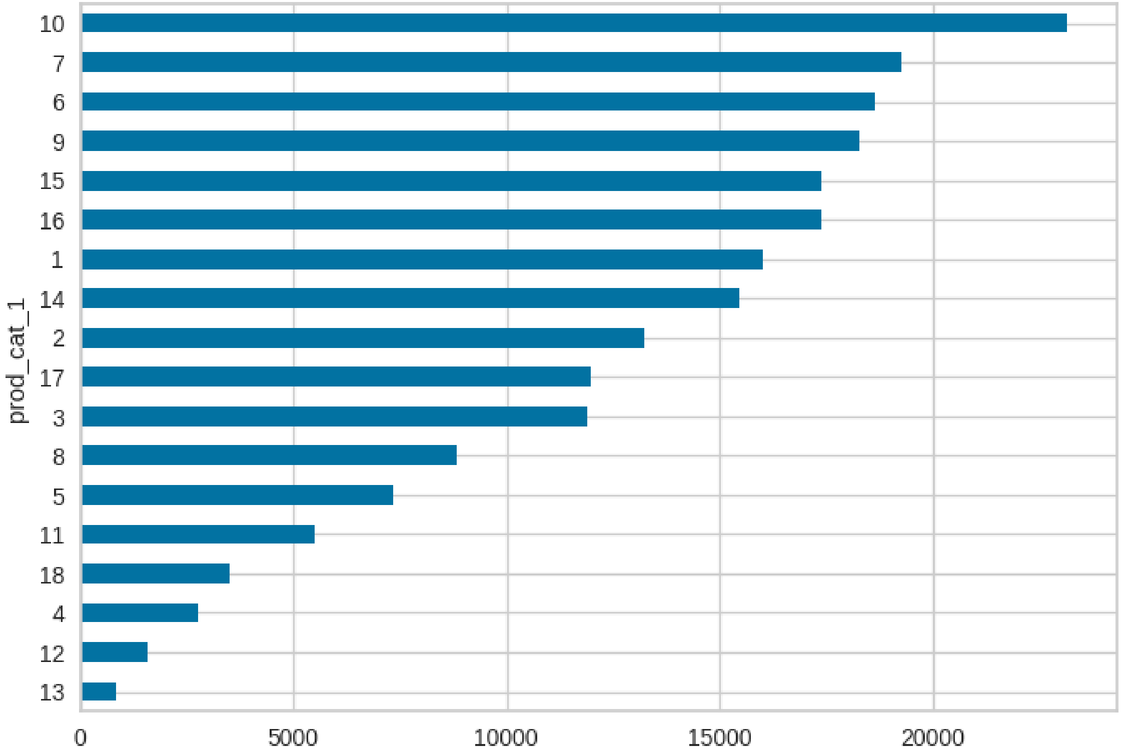


Education + Sex vs. Revenue



Taking a look at these graphics is hard to notice a high correlation between them. I also tried to combine some variables, but it didn’t give too much clarity at all. The exception is the variable prod_cat_1 which appears to have some categories very correlated to the high revenue values:

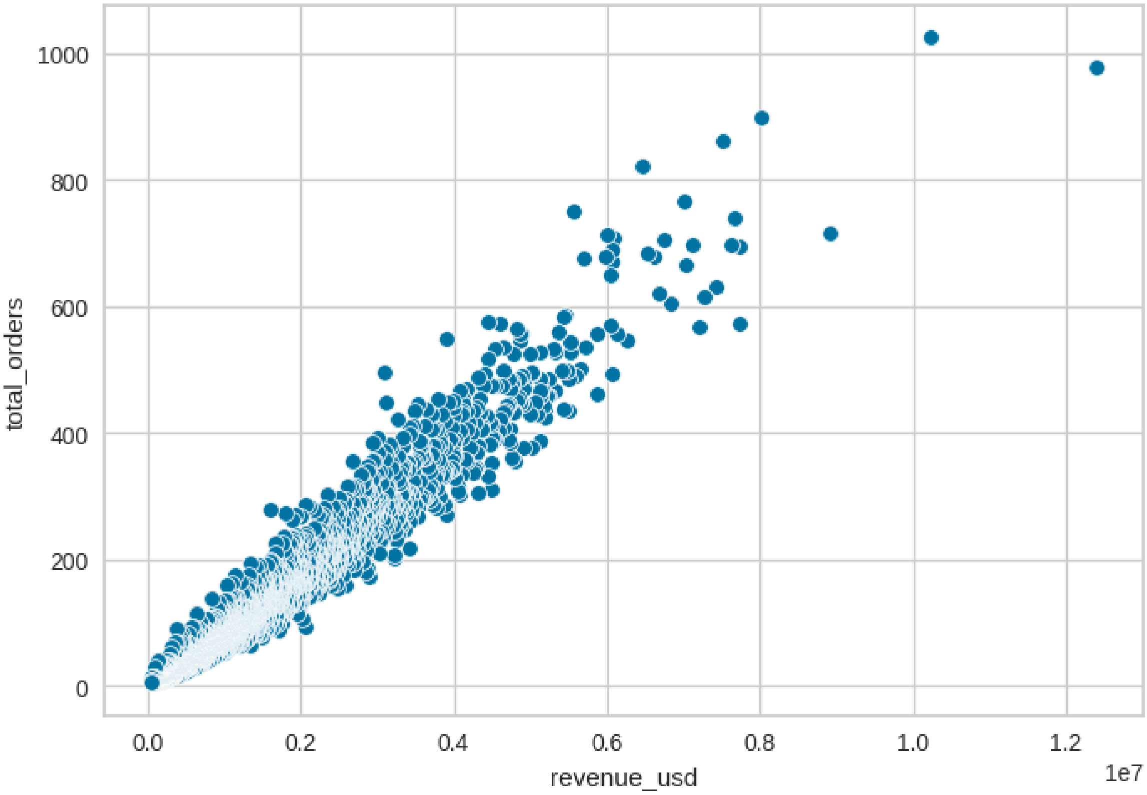
Product Category 1 vs. Revenue



4. RFV (recency, frequency and value)

Although it was hard to find any sign of good parameters to build a regression model to find out the path to reach the best customers, we still have a perfect linearity between orders and total revenue, as I was expecting:

Linear relationship between purchases and total revenue



I decided to use this to make an approach based on RFV techniques, but a partial one, since the data frame doesn’t have the purchase date to build the recency variable.

Even working only with the F (frequency of purchases) and M (total revenue) of the equation, I thought it would be a good approach to help the Marketing Manager in this case.

Approach #1:

We can simply sort the customers based on their total purchases, total revenue, or average amount per purchase. This could work in campaigns with general goals like “Increase revenue/sales”.

I propose to make some specific slices in the sorted data, for example:

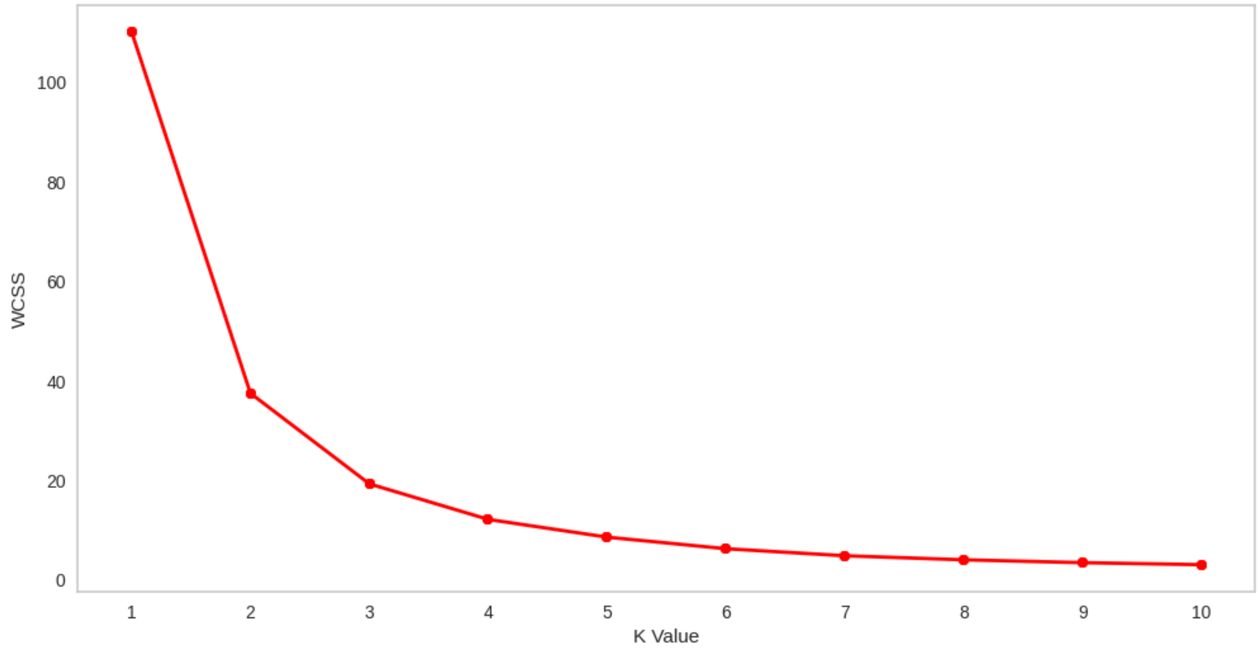
- Get the list of customers that together are responsible for 60% of the total revenue
- Select customers with average purchase values above a certain number (this could improve sales of products with high value, for example)
- We can also select the top customers in terms of revenue and purchases

Approach #2:

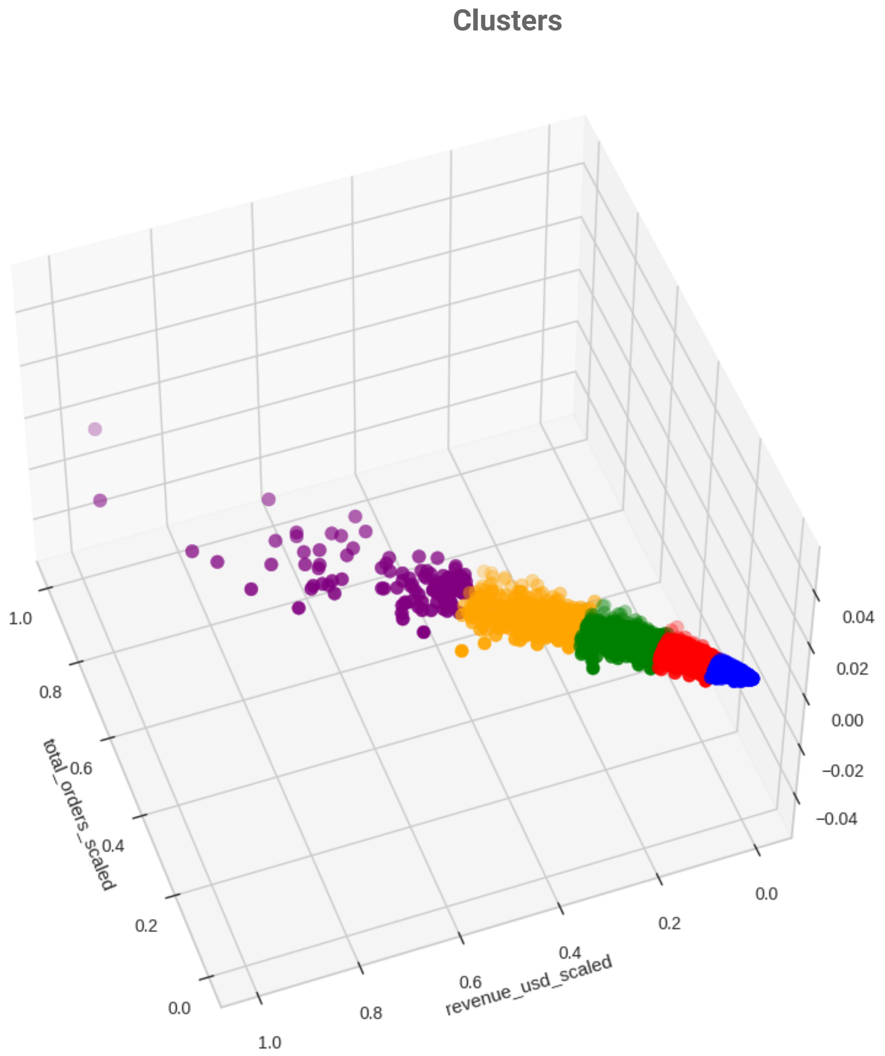
Another good approach would be clustering these customers into different groups based on their frequency and value information.

I used the KMeans algorithm and first had to find the optimum value of clusters, in this case 5:

Optimal values for clusters

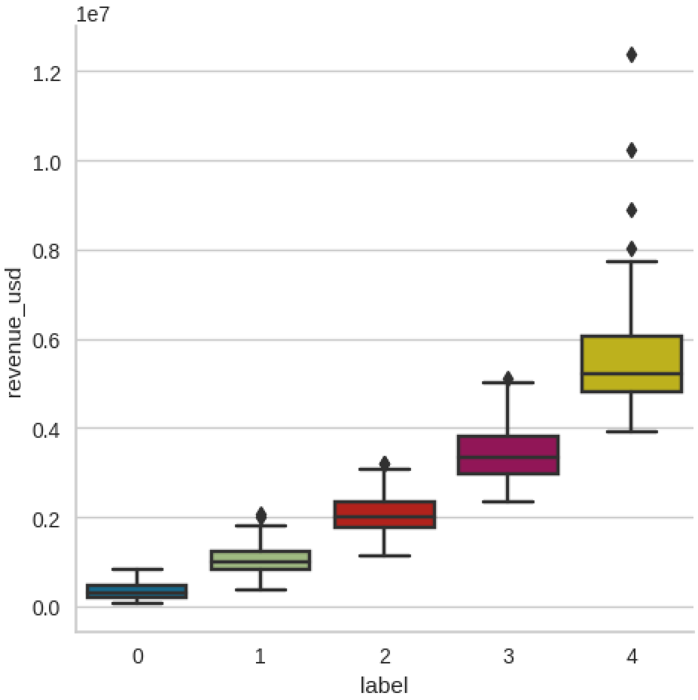


So I sliced the customers into 5 groups, each one representing a different combination of frequency and value:

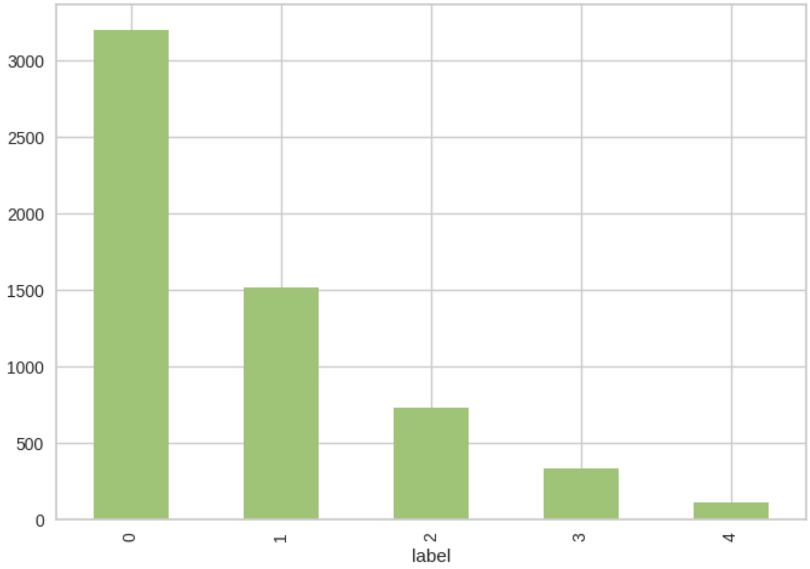


Group #4 for example, has the customers with the highest frequency and value, but it’s the smallest group. This is the kind of trade-off I believe is important to make it transparent to the Marketing Manager.

Clusters vs. Revenue



Quantity of unique customers in each cluster

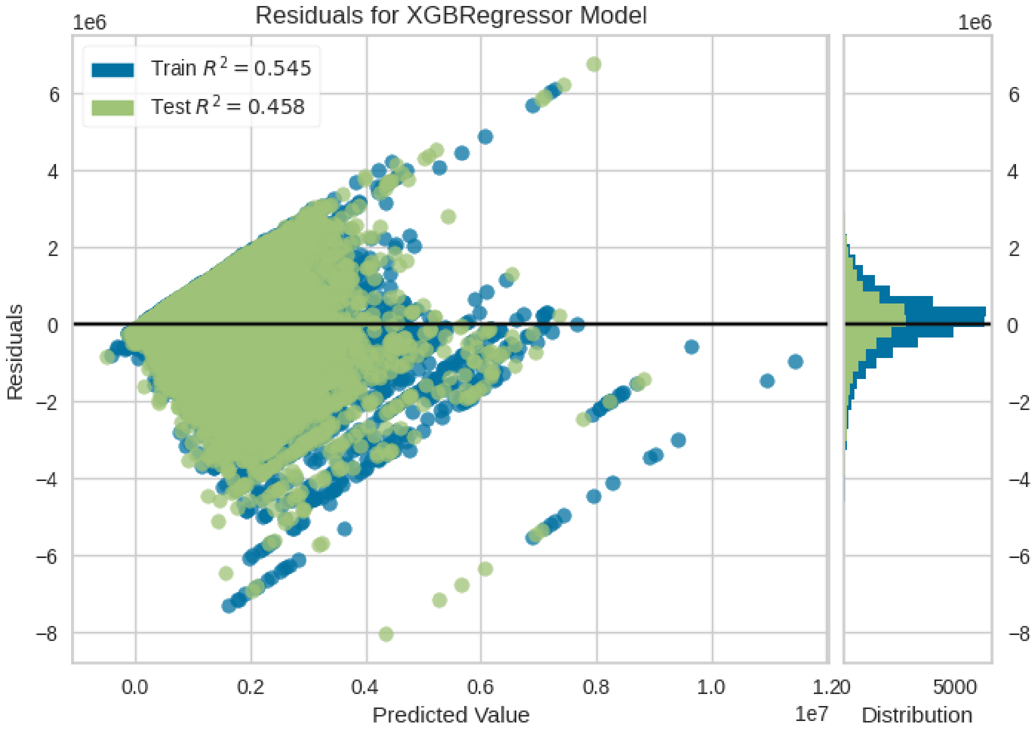


5. Regression

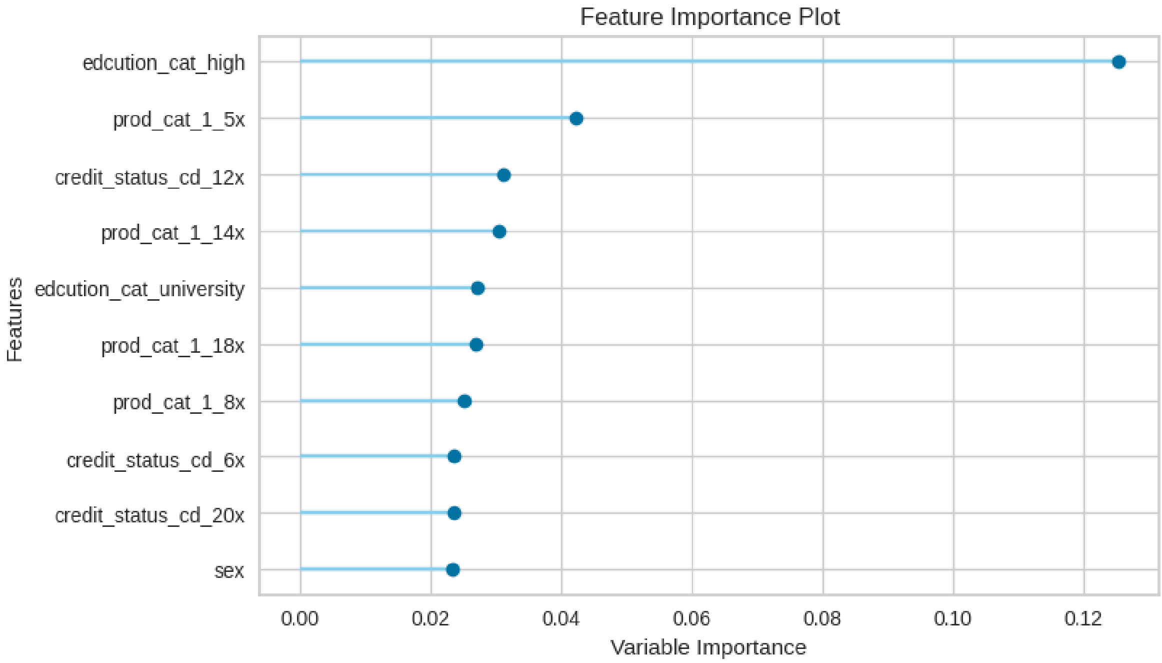
Another good approach would be to run a Linear Regression to understand which interaction of variables can explain revenue the most. This could be great for creating personas based on customer demographics and behavior. Sadly we don’t have good parameters to do this since a large percentage of the variables are not correlated with the revenue itself as I demonstrated before.

After running some different machine learning models to solve this regression problem, I got to XGBRegressor, but with very poor results and a lot of errors:

Residuals for XGBRegressor model



Feature Importance



Would be great to have some more data or different variables to improve this model and start to build these personas for the marketing team.

6. Conclusion

I believe part of the problem can be solved with approaches #1 and #2, but to bring more complexity and have the opportunity to recommend other solutions, these are the questions I would ask the Marketing Manager:

1) Is it possible to collect more data (samples and variables) about these customer's purchases?
2) Do you have more directions about the strategy of the new marketing campaign?
3) In this case, I'm recommending approaches based on the customers, but we can also develop solutions based on products if this is the case (we can develop a Next Best Offer model, for example)