Project II: Marketing Campaign Analysis

1. Objectives

The purpose of this project is to address two key questions. **Firstly, which customer segments spend the most money in our store?** Secondly, **which customer segments are more inclined to revisit for another shopping experience, following their recent purchases?**

2. Data collection

This dataset was sent by a particular client who wanted to optimize their business.

3. Data preparation

i. Data cleaning

Initially, we took an overview of our dataset and selected the appropriate columns for our analysis. As usual, we analyse any inconsistencies within the dataset through discussion. Furthermore, we will employ boxplots to visually identify any outlier values.

ii. Normalization and Dummy indexing

We will also normalize and convert categorical variables into numerical ones using dummy indexing prior to the final stage. This will lead to a more accurate prediction. Finally, we will standardize the remaining features to eliminate any novel and insignificant outliers.

4. Data exploration

i. Simple graph

We will conduct a thorough data exploration using clear and concise charts such as "boxplot" or "scatterplot" after we have purified our dataset.

ii. Non-graphical exploration

Eventually, we may employ the "describe" function for a non-graphical analysis of the dataset.

5. Data modelling

i. Models and variable selection

Once the data is ready for processing, we can begin to build the model. Finally, we execute the model. We select our target based on unique features and emphasise that our project requires a supervised perspective. Additionally, this problem could be addressed with neural networks.

ii. Execution model

Due to the nature of our work, regression models will be utilised. We will select from a range of models in order to obtain the superior one. The models will then undergo diagnostic and comparative analyses.

iii. Model diagnosis and comparison

Following model execution and identification of the optimal one, "cross_val_score" or "GridSearchCV" functions may be employed to assess the chosen model. In another scenario, such as in a classification problem, a confusion matrix could be used.

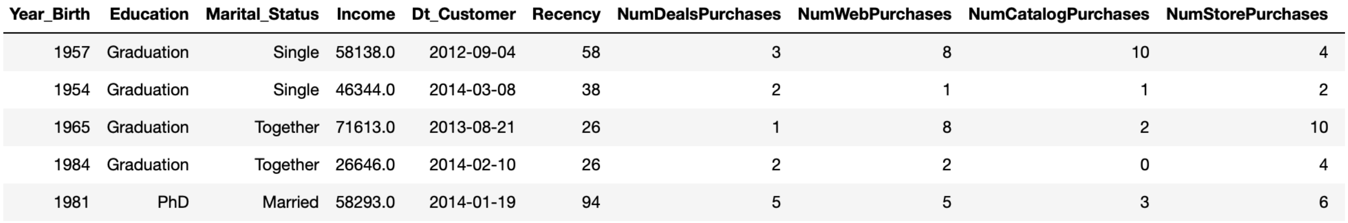
5. Presentation

i. Presentation

After selecting and refining the model, results can be visually depicted through a graph or table. The most crucial aspect is to convey the disparities between the anticipated and actual values.

Abstract

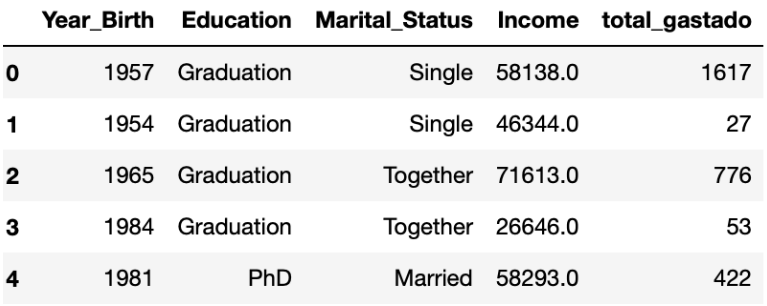
Our project consists of a marketing campaign comprising various columns of purchase information, subscriptions, and customer data, among others.



Year_Birth	Education	Marital_Status	Income	Dt_Customer	Recency	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
1957	Graduation	Single	58138.0	2012-09-04	58	3	8	10	4
1954	Graduation	Single	46344.0	2014-03-08	38	2	1	1	2
1965	Graduation	Together	71613.0	2013-08-21	26	1	8	2	10
1984	Graduation	Together	26646.0	2014-02-10	26	2	2	0	4
1981	PhD	Married	58293.0	2014-01-19	94	5	5	3	6

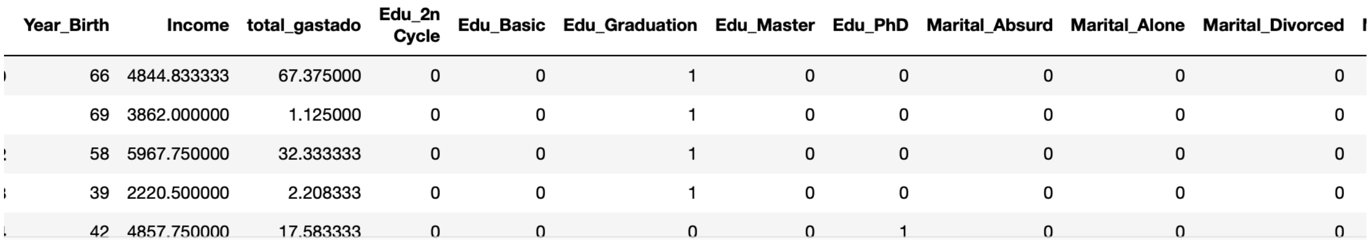
The objective of this study is to address two of the following queries: Firstly, **which customer segments spend the most money in our store?** Secondly, **which customer segments are more inclined to revisit for another shopping experience, following their recent purchases?**

We will proceed by selecting the relevant columns to address the first query.



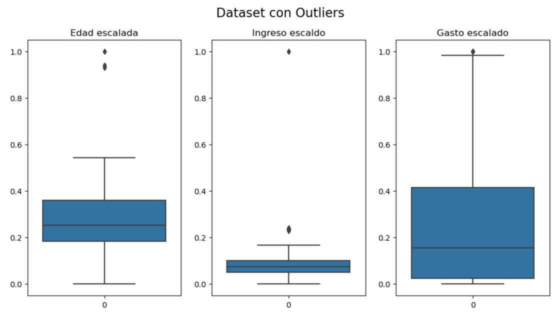
	Year_Birth	Education	Marital_Status	Income	total_gastado
0	1957	Graduation	Single	58138.0	1617
1	1954	Graduation	Single	46344.0	27
2	1965	Graduation	Together	71613.0	776
3	1984	Graduation	Together	26646.0	53
4	1981	PhD	Married	58293.0	422

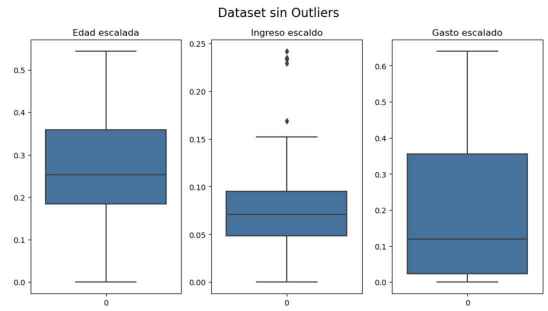
Afterwards, we proceed with data cleaning, converting categorical columns to dummy variables, and filtering the dataset.



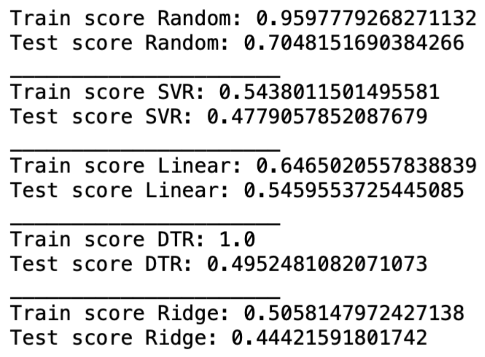
Year_Birth	Income	total_gastado	Edu_2n Cycle	Edu_Basic	Edu_Graduation	Edu_Master	Edu_PhD	Marital_Absurd	Marital_Alone	Marital_Divorced
66	4844.833333	67.375000	0	0	1	0	0	0	0	0
69	3862.000000	1.125000	0	0	1	0	0	0	0	0
58	5967.750000	32.333333	0	0	1	0	0	0	0	0
39	2220.500000	2.208333	0	0	1	0	0	0	0	0
42	4857.750000	17.583333	0	0	0	0	1	0	0	0

We will use libraries such as Seaborn or Matplotlib to identify outliers in our variables.





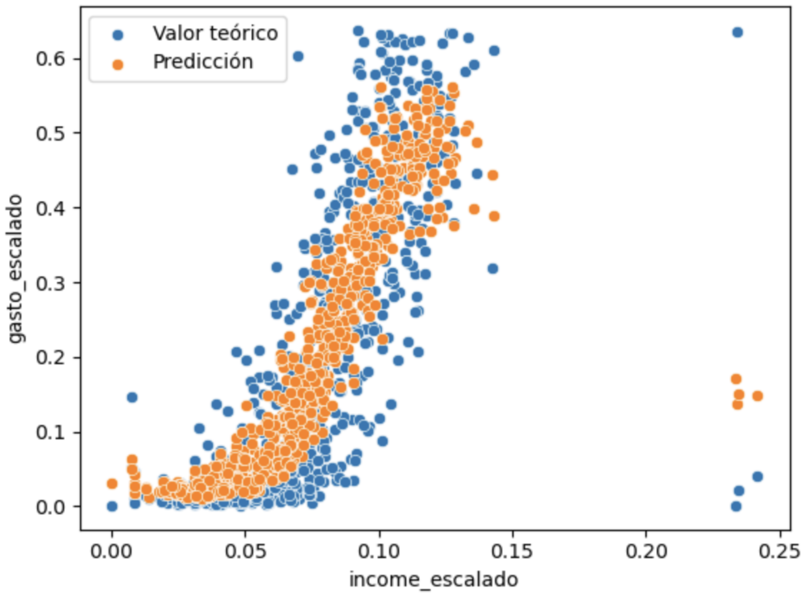
It's easy to visualize the dispersion of each variable, but this may affect the accuracy of our final prediction. Therefore, we can establish a threshold to constrain this dispersion and improve the prediction's reliability.
After implementing these changes, we can evaluate various regression models using our dataset to determine the best performer.



```
Train score Random: 0.9597779268271132
Test score Random: 0.7048151690384266
______________________
Train score SVR: 0.5438011501495581
Test score SVR: 0.4779057852087679
______________________
Train score Linear: 0.6465020557838839
Test score Linear: 0.5459553725445085
______________________
Train score DTR: 1.0
Test score DTR: 0.4952481082071073
______________________
Train score Ridge: 0.5058147972427138
Test score Ridge: 0.44421591801742
```

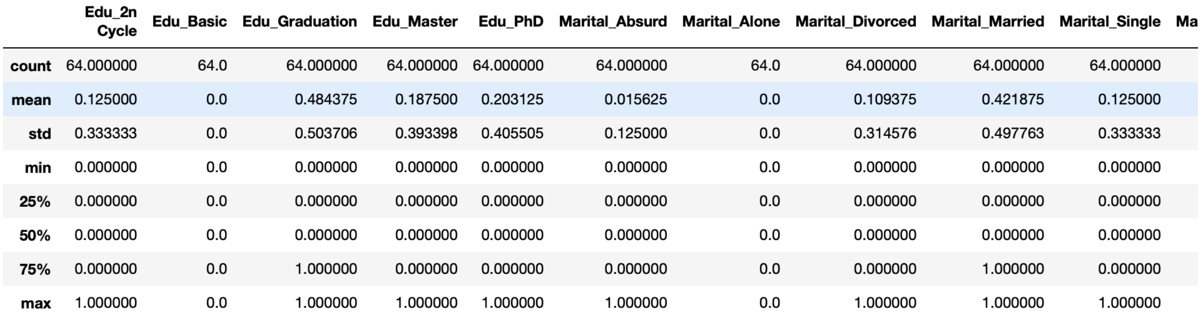
The Random Forest Regressor emerges as the top-performing option. However, I would like to emphasise that the Decision Tree Regressor model achieved an excellent prediction for the train set, yet only a mediocre result for the test set, possibly due to insufficient data.

Moving forward, we can visualise the predicted values against the actual values.



The overall prediction seems satisfactory. With a 70% level of performance, we can draw conclusions that may support our future marketing campaign. However, there is no doubt that each value is extremely separated from the overall result.

To reach our conclusions, we need to re-scale the dataset. We create a smaller sample dataset, containing just 10% of the original data. Our objective is to identify the customers that are the highest consumers. We have selected the features "Marital Status" and "Graduation". We are not interested in columns with null values.



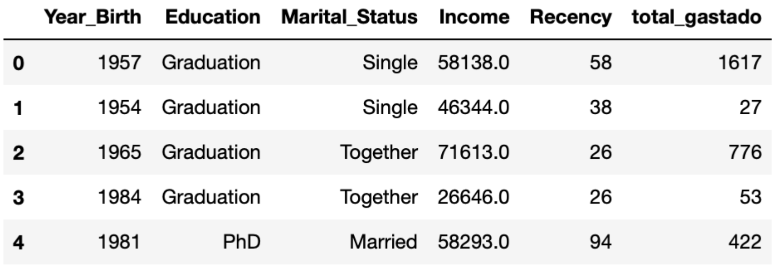
	Edu_2n Cycle	Edu_Basic	Edu_Graduation	Edu_Master	Edu_PhD	Marital_Absurd	Marital_Alone	Marital_Divorced	Marital_Married	Marital_Single	Ma
count	64.000000	64.0	64.000000	64.000000	64.000000	64.000000	64.0	64.000000	64.000000	64.000000	
mean	0.125000	0.0	0.484375	0.187500	0.203125	0.015625	0.0	0.109375	0.421875	0.125000	
std	0.333333	0.0	0.503706	0.393398	0.405505	0.125000	0.0	0.314576	0.497763	0.333333	
min	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
25%	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
50%	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
75%	0.000000	0.0	1.000000	0.000000	0.000000	0.000000	0.0	0.000000	1.000000	0.000000	
max	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	0.0	1.000000	1.000000	1.000000	

Next, we create an algorithm to analyze the dataset. **The algorithm reveals that widowed individuals and those with a PhD tend to spend more.**

```
Los gastos generados por la columna Edu_2n Cycle son 203

Los gastos generados por la columna Edu_Graduation son 1081

Los gastos generados por la columna Edu_Master son 1394

Los gastos generados por la columna Edu_PhD son 1758

Los gastos generados por la columna Marital_Divorced son 1946

Los gastos generados por la columna Marital_Married son 2683

Los gastos generados por la columna Marital_Single son 2905

Los gastos generados por la columna Marital_Together son 3436

Los gastos generados por la columna Marital_Widow son 3484
```

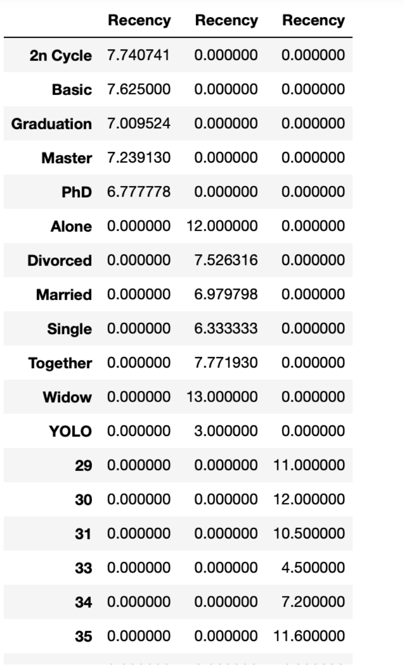
Moving on, we address the final question: **which customer segments are more likely to make a repeat purchase following their most recent one?**

To answer this, we create a sub-dataset with an additional column, "Recency."



	Year_Birth	Education	Marital_Status	Income	Recency	total_gastado
0	1957	Graduation	Single	58138.0	58	1617
1	1954	Graduation	Single	46344.0	38	27
2	1965	Graduation	Together	71613.0	26	776
3	1984	Graduation	Together	26646.0	26	53
4	1981	PhD	Married	58293.0	94	422

We are focused on identifying the minimum value. Following the refinement of our new dataset, we filtered ages with the aim of selecting individuals under 60 years old. Our approach to exploring the dataset was akin to that of MySQL programming, without using "groupby," "value_counts," or "idxmin/idxmax."



	Recency	Recency	Recency
2n Cycle	7.740741	0.000000	0.000000
Basic	7.625000	0.000000	0.000000
Graduation	7.009524	0.000000	0.000000
Master	7.239130	0.000000	0.000000
PhD	6.777778	0.000000	0.000000
Alone	0.000000	12.000000	0.000000
Divorced	0.000000	7.526316	0.000000
Married	0.000000	6.979798	0.000000
Single	0.000000	6.333333	0.000000
Together	0.000000	7.771930	0.000000
Widow	0.000000	13.000000	0.000000
YOLO	0.000000	3.000000	0.000000
29	0.000000	0.000000	11.000000
30	0.000000	0.000000	12.000000
31	0.000000	0.000000	10.500000
33	0.000000	0.000000	4.500000
34	0.000000	0.000000	7.200000
35	0.000000	0.000000	11.600000

Our findings reveal that the first column corresponds to level of educational attainment, the second to marital status, and the third to age. We can infer that **customers most likely to make return purchases are PhD graduates who are 33 years old and unmarried (Yolo).**