Individual Project
Final Report

Author: Bruno Santana Sereicikas de Azevedo

Student number: 2116218 Email: b.santanasereicikasdeazevedo@student.han.nl

MDD001 – IB Minor of data driven decision making for business,
HAN University of Applied Sciences

Project M3DMiB

January 15, 2023

Table of Contents

1. Introduction

This is the report about the individual project developed during the IB Minor MDD01 – Data-driven decision making for business at HAN University of Applied Sciences.

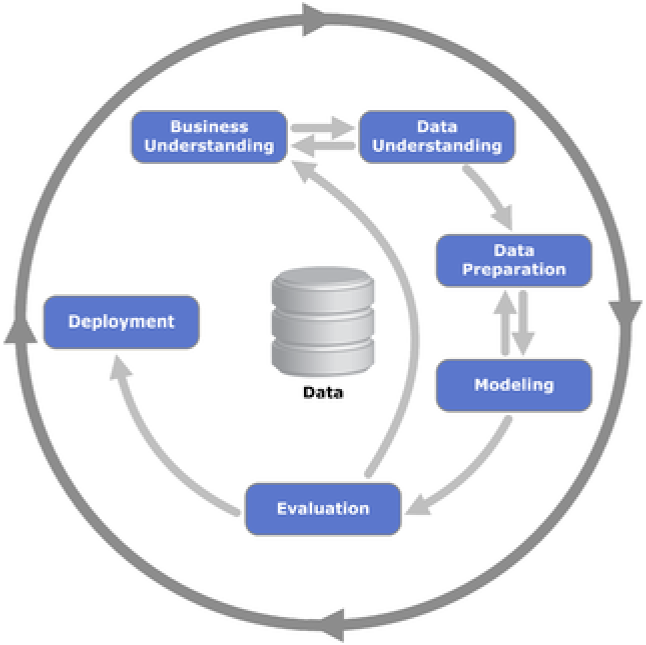
The chosen project is a multiclass classification problem in the context of an automobile company. The project and dataset are available on Kaggle. However, a few changes were made from the original project suggested. The problem solved in this project is the one described in section Business Case.

The project (script and dataset) is available in this GitHub Repository.

1.1. Objectives

The objectives proposed for this project were:

- Learn about and practice Python, as this programming language offers several tools that are useful for data analysis and is highly demanded in the labor market
- Understand the main Python packages used for Data Science
- Create a machine learning model using Python
- Apply Crisp-DM model steps in a practical project



2. Business Case

The business case to be solved in this project is described below:

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4 and P5). After intensive market research, they've deduced that the behavior of new market is similar to their existing market.

In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). Then, they performed segmented outreach and communication for different segments of customers. This strategy has worked exceptionally well for them. They plan to use the same strategy on new markets.

You are required to help the manager to predict the right group of new customers. The demand is to create a machine learning model that can allocate each customer to the adequate segment.

3. Setting up environment

Python offers several packages to support data scientists in data cleaning, visualization, analysis, and modeling. All packages used are described as following:

- Pandas: import the dataset and perform calculations and transformations in the data.
- Numpy: perform calculations and transformations in the data
- Missingno: identify missing values in the dataset
- Matplotlib and Seaborn: plot graphs for data visualization
- Scikit-learn: create machine learning model and provide evaluation tools to understand the model performance.

4. Data Understanding

Following the methodology suggested in the Crisp-DM Model, the first step after the business understanding is data understanding.

4.1. Dataset variables' description

After importing the dataset and a first look into it revealed that it consisted of 11 columns with 8068 rows with information about the customers of the company. The table below describe the content of each column:

Column	**Description**
ID	Unique identifier code for each customer (numerical value)
Gender	Gender of each customer (Male or Female)
Ever_Married	Classify each customer if they have ever married or not (Yes or No)
Age	Gender of each customer (Numerical value)
Graduated	Classify each customer if they are graduated or not (Yes or No)
Profession	Occupation field of each customer (text variable with 9 different fields)
Work_Experience	Years of working experience of each customer in years (numerical value)
Spending_Score	Spending score of each customer (High, Average or Low)
Family_Size	Number of people in the customer family (numerical value)
Var_1	Anonymized category for the customer (text variable with 7 categories)
Segmentation	Segment classified for each customer by the sales team (A, B, C, D)

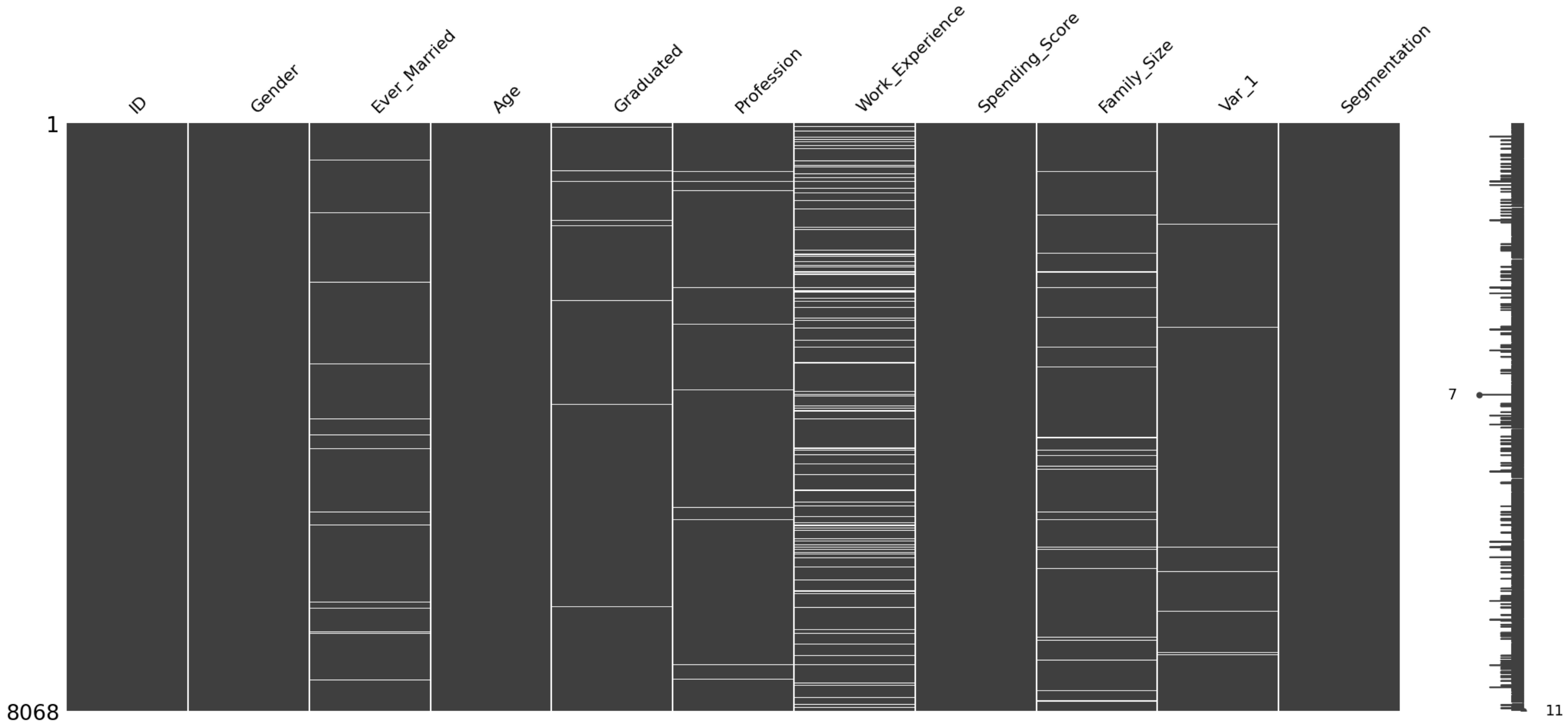
The first look into the dataset shows that it contains 9 variables, being 3 numerical variables and 6 categorical variables, and the column 'Segmentation' which have the segments to be predicted for each customer. The column 'ID' is not considered, as it doesn't represent useful data for the analysis.

4.2. Missing data

Some columns have missing data. The table below shows the number of occurrences of missing data on each column:

Variable	Number of missing values
Work_Experience	829
Family_Size	335
Ever_Married	140
Profession	124
Graduated	78
bVar_1	76

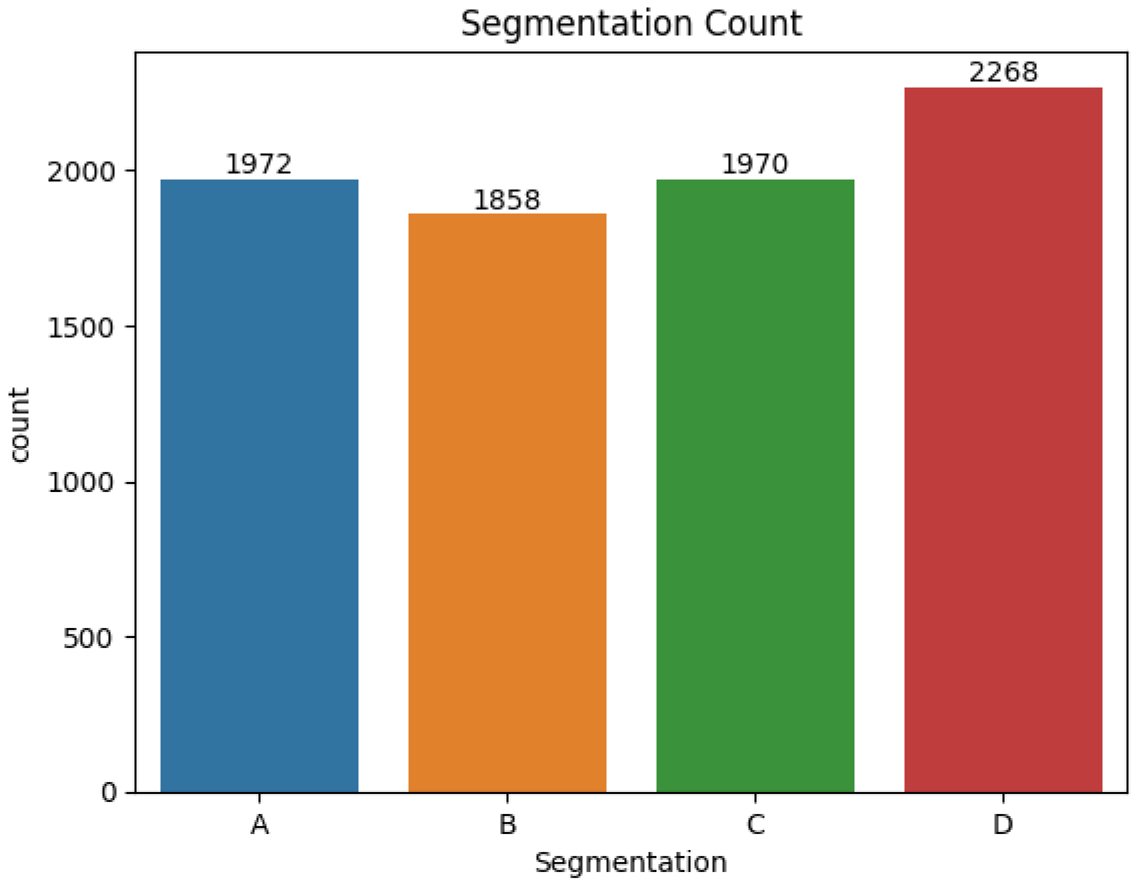
After identifying these missing values, it is important to understand the behavior to know if these are missing randomly or for a specific reason, as this affects the decision of how to deal with it. The package 'missingno' creates a visual that helps visualizing it. The black columns represent columns of the dataset, and the white rows are the missing values in each column.



After analyzing the visuals, the conclusion is that the missing data doesn't follow a regular pattern and is randomly missing. Therefore, the decision made for dealing with these missing registers was to replace it with the string “No info” for categorical values and to replace it with the median for numerical variables.

4.3. Segmentation distribution

The next step of the data understanding was to know the segmentation distribution between customers. The graph below shows that, although segment D is slightly bigger than the other segments, the number of customers in each segment is similar.



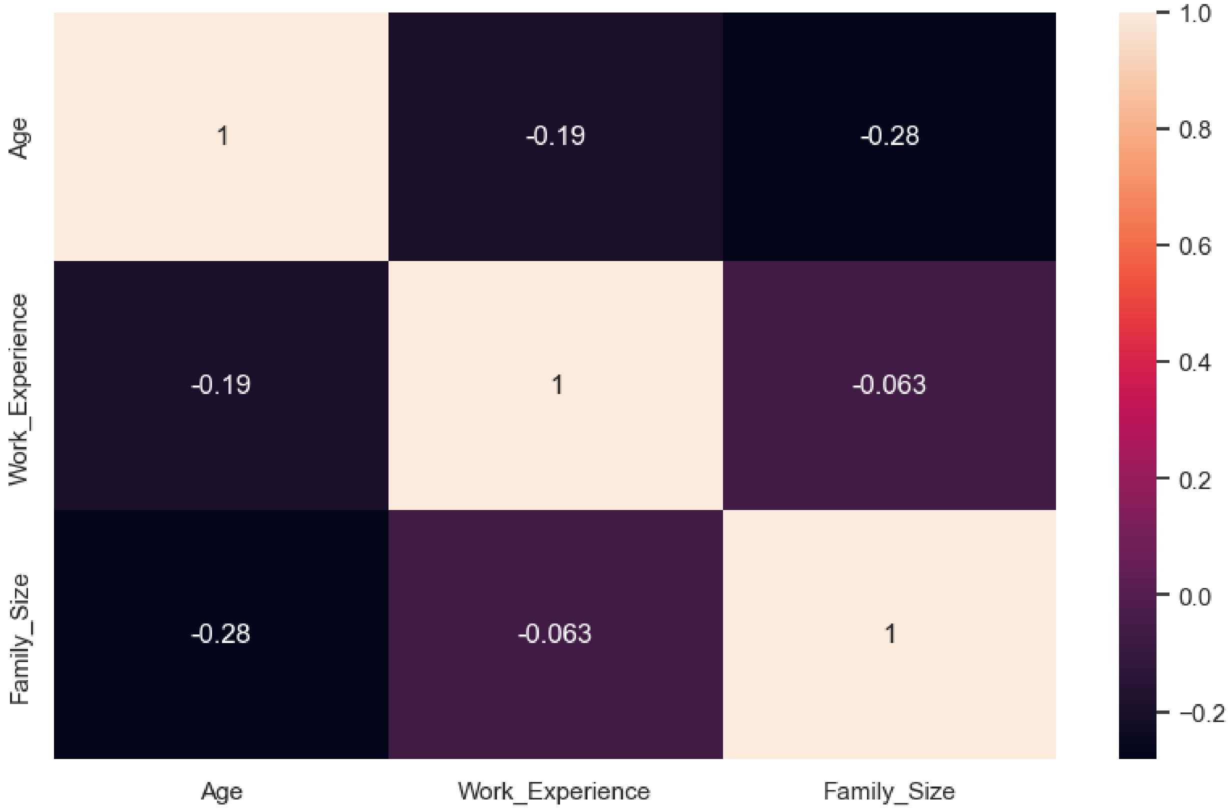
The next step is to understand the features of each group and how they differ from each other.

4.4. Numerical variables understanding

For the numerical variables, there are two main aspects to be analyzed. The first one is to check the correlation between variables. The second one is to visualize each variable behavior according to each segment and in general.

4.4.1. Correlation Matrix

If two or more of them have a high correlation, this would mean that both cause the effect of multicollinearity in the modeling step later and affect the model performance. To check if this is happening, a correlation matrix was used as seen below.



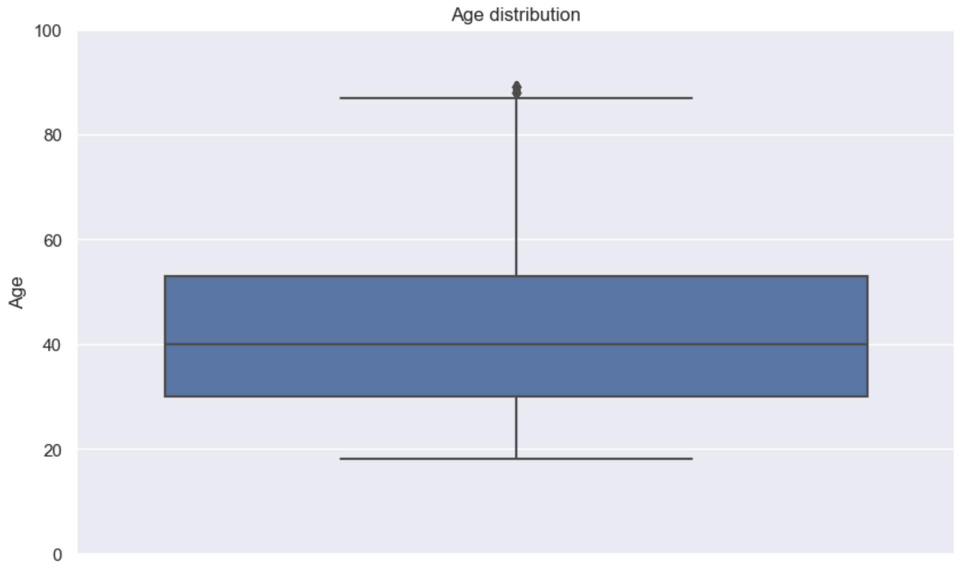
The result is that the correlation between variables is low. This means that all of them can be kept.

4.4.2. Data visualization per variable

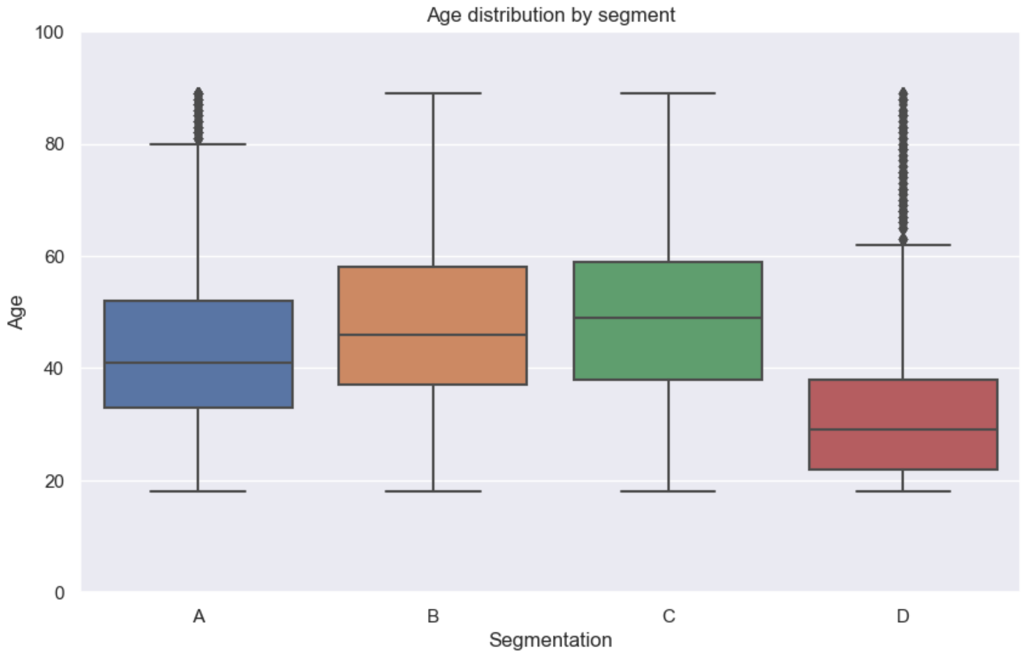
For numerical variables, a very effective visual for analyzing them is the boxplot graph. Following, it was plotted a boxplot for a general view of the dataset and a boxplot for each segment so that the view of the features of each segment is clear.

- **Age**

The visual below shows the age distribution for the whole dataset. Customers' age varies between 18 to around 90 years old, with most of them having less than 50 years old.

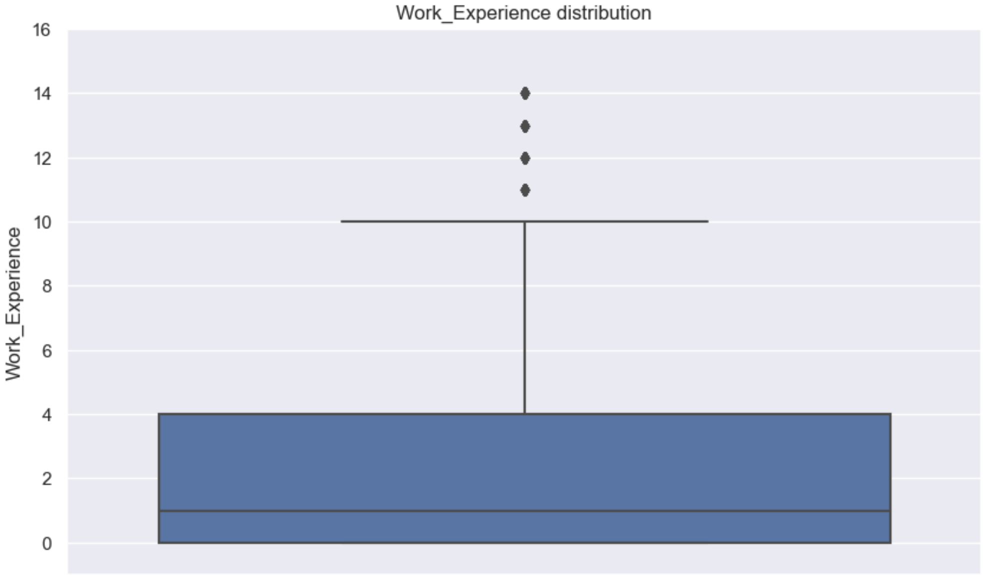


When looking at the breakdown by segment, it's easy to notice that segment D concentrates on customers from a considerably younger age than the other segments. Segment A also seems to consist of customers a bit younger than segments B and C, which are segments that look similar. However, it is also interesting to notice that although there are these differences in trends of age between segments, all of them contain customers of the age of 18 years old and people with more than 80 years old.

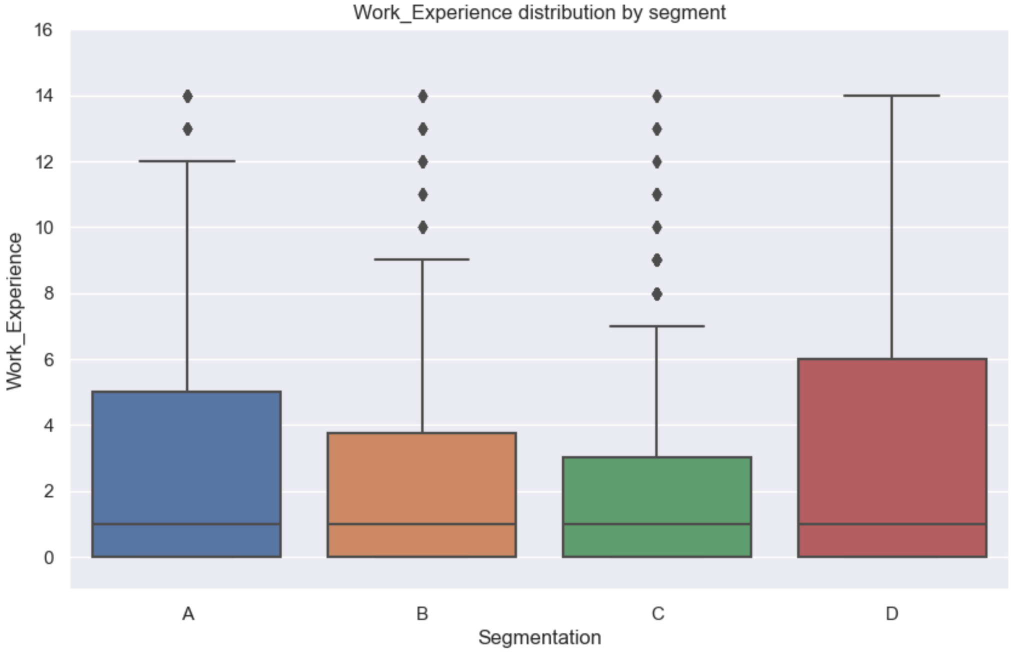


- **Work Experience**

The visual below shows the work experience distribution for the whole dataset. It shows that most customers have up to 4 years of experience and half of them have up to 1 year. Even though there are some with more than 5 years of professional experience.

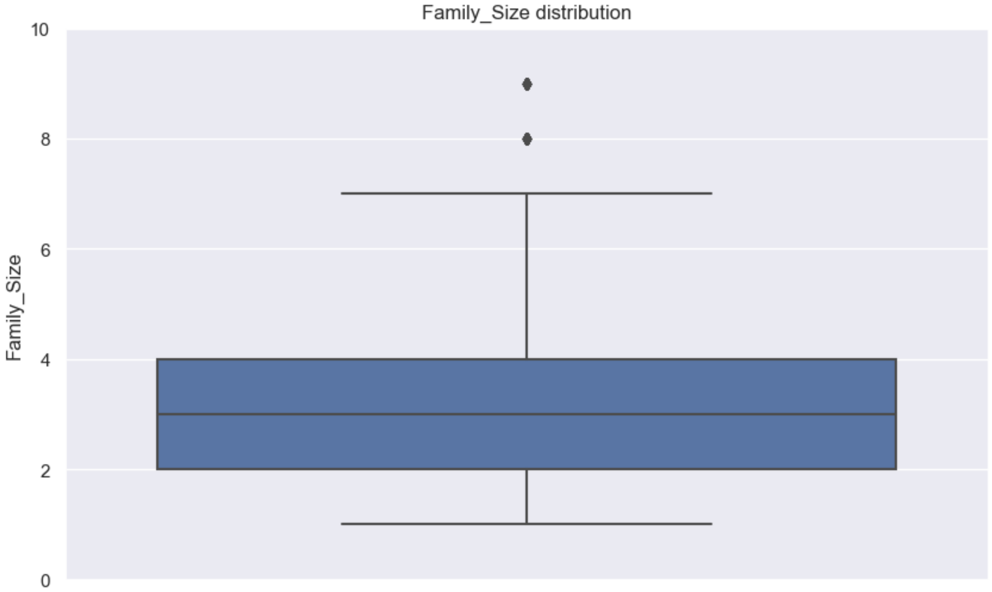


The breakdown by segment reveals that all segments have the same median of 1 year of experience. However, segments D, A, B and C in that order concentrates people with more years of experience. That is, segment D concentrates more customers with more years of experience and segment C concentrates customers with less years of experience.

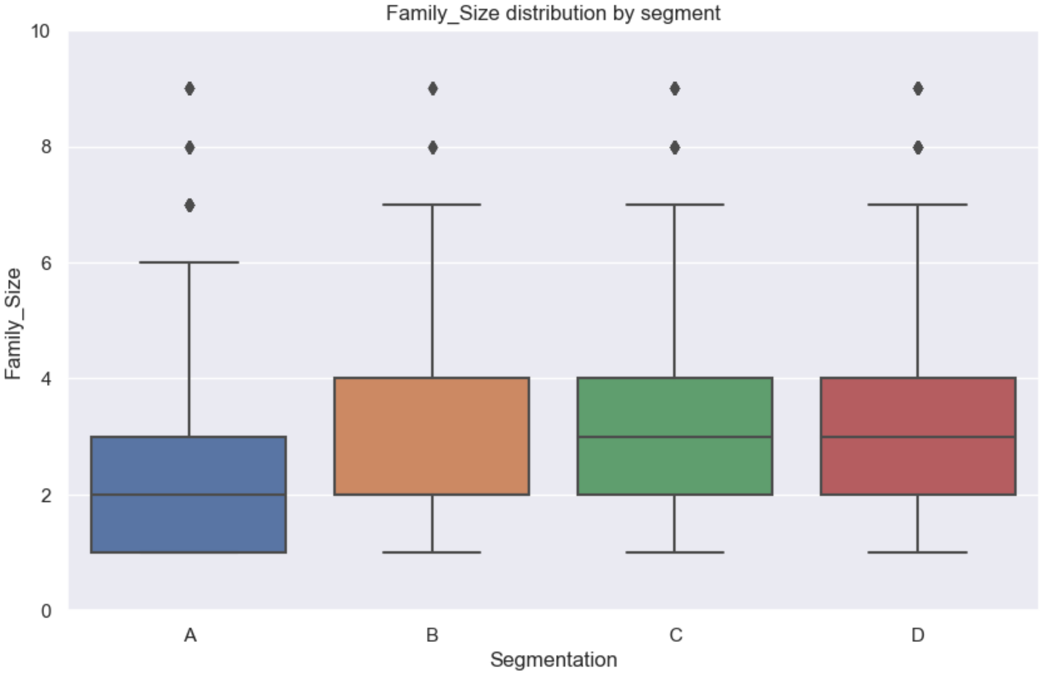


- **Family size**

The visual below shows the work experience distribution for the whole dataset. Most customers have 2 or more family members. It's also interesting to notice that all customers have at least 1 family member.



The breakdown by segment reveals that segment A consists of a group of customers with smaller families than the other groups. Segment B also concentrates customers with smaller family size than groups C and D (the median of segment B is 2, which is smaller than the median of groups C and D). Segments C and D looks to have really similar behavior in this variable.



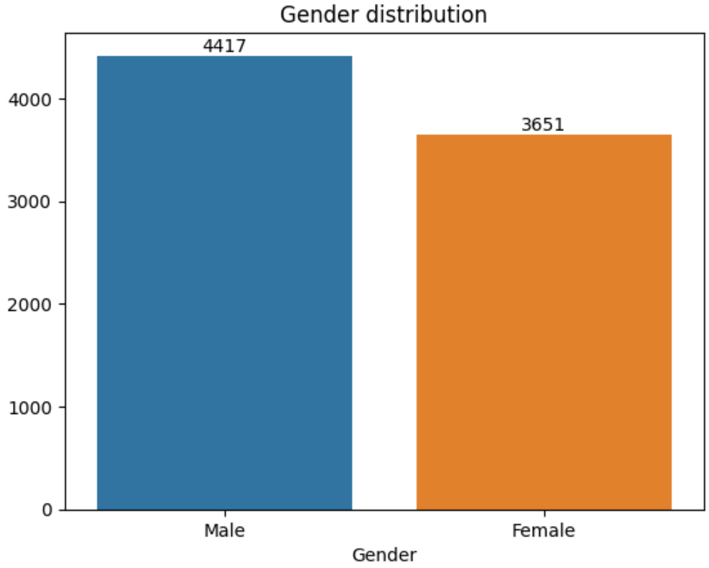
4.5. Categorical variables understanding

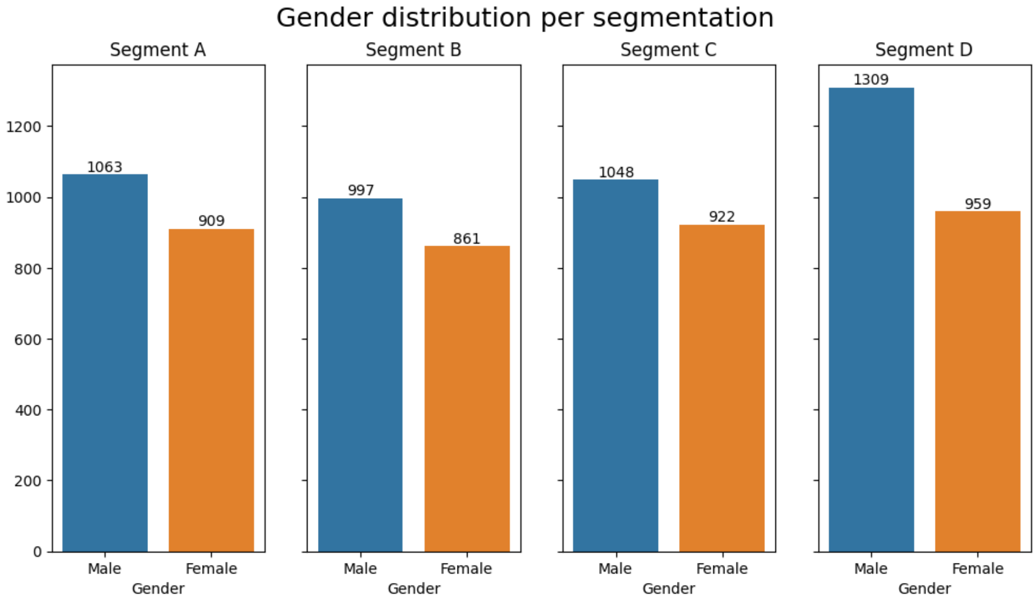
For the categorical variables, the main aspect to be analyzed is to visualize each variable behavior according to each segment and in general.

4.5.1. Data visualization per variable

- **Gender**

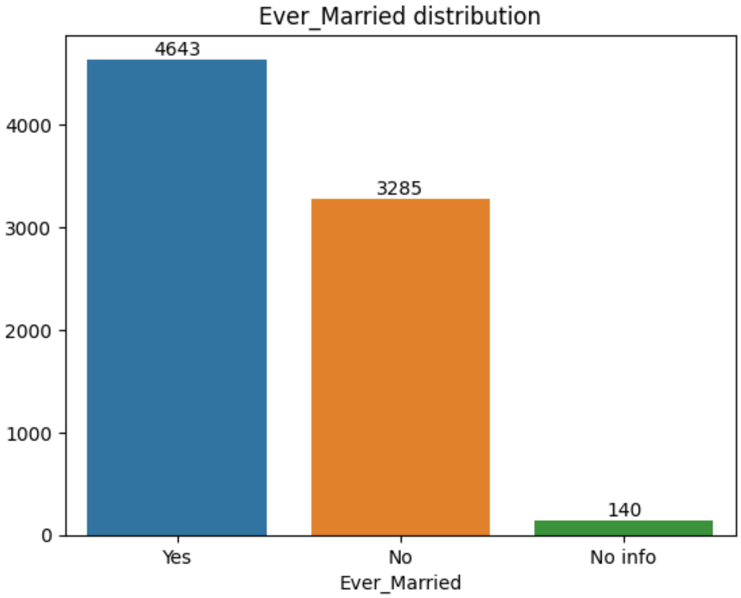
Both the general view and the view by segment reveals that there are more male than female customers. The only different difference that is interesting to highlight is that segment D has a higher proportion of male customers than the other segments.



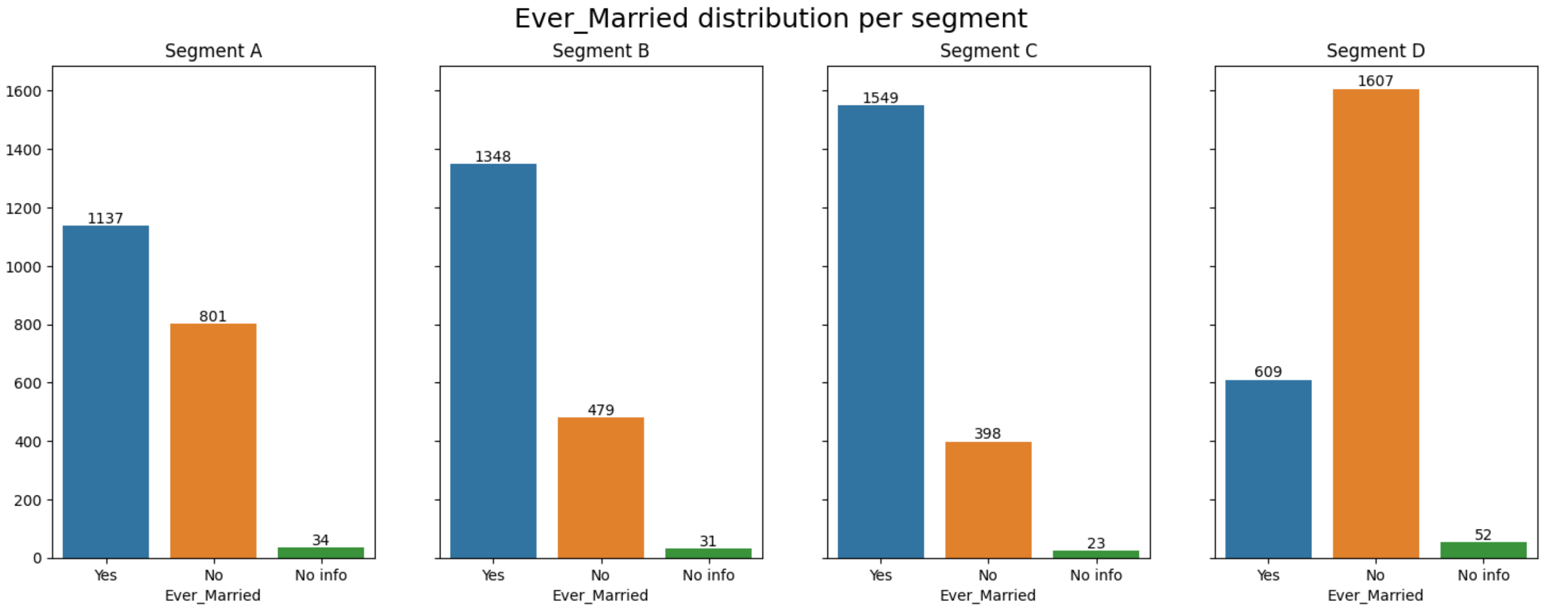


- **Ever married**

The general view reveals that most customers have already married. Although, the number of customers that are not married stills relevant.

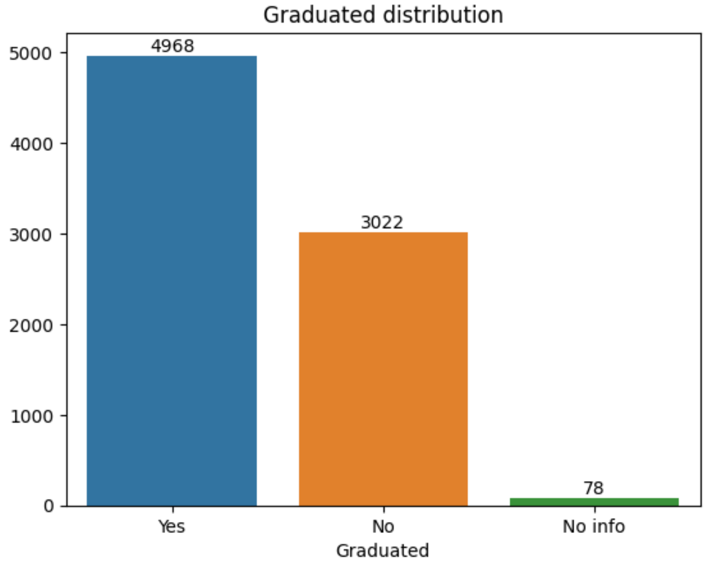


The breakdown by segment reveals that segments B and C concentrate on a higher number of customers that have already married. Segment D follows the opposite trend, concentrating on a big number of customers that are not married. Finally, segment A is the one that shows a more balanced proportion than the other groups, although the number of people who have already married is higher than the ones that are not married.

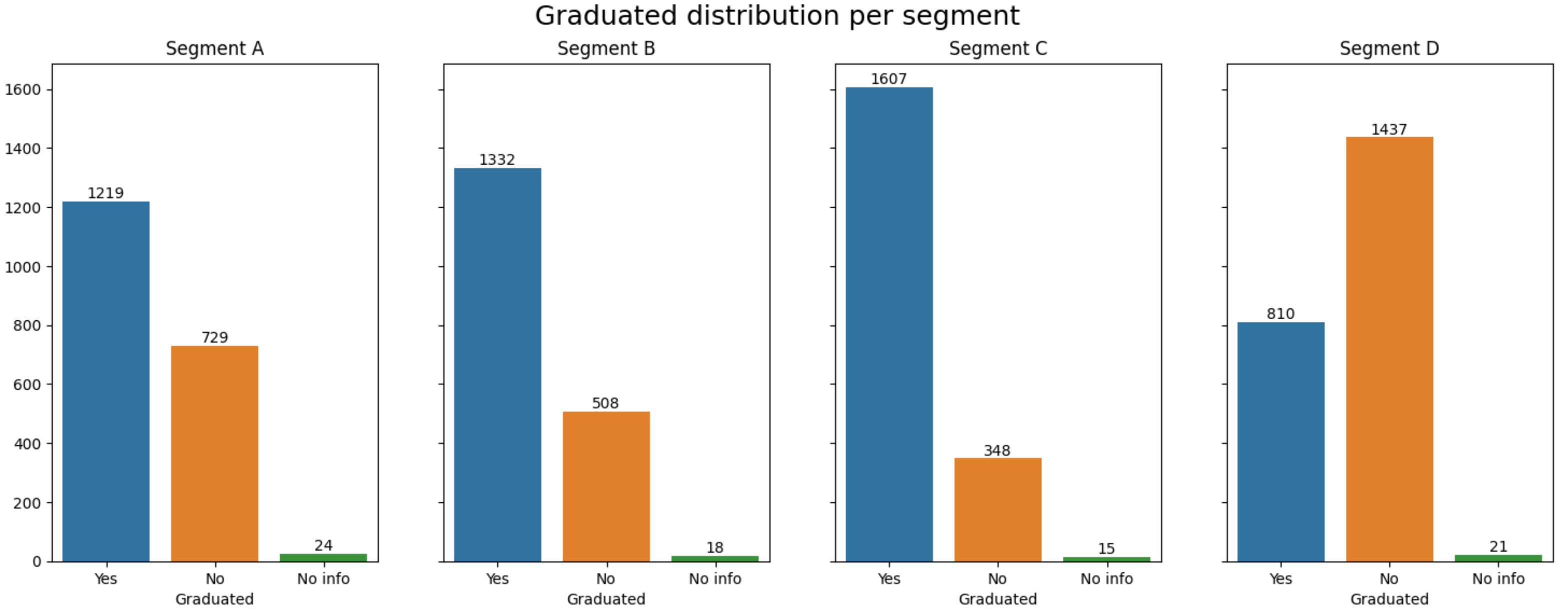


- **Graduated**

The general view reveals that most customers are graduated, although there is a big number of customers who are not graduated.

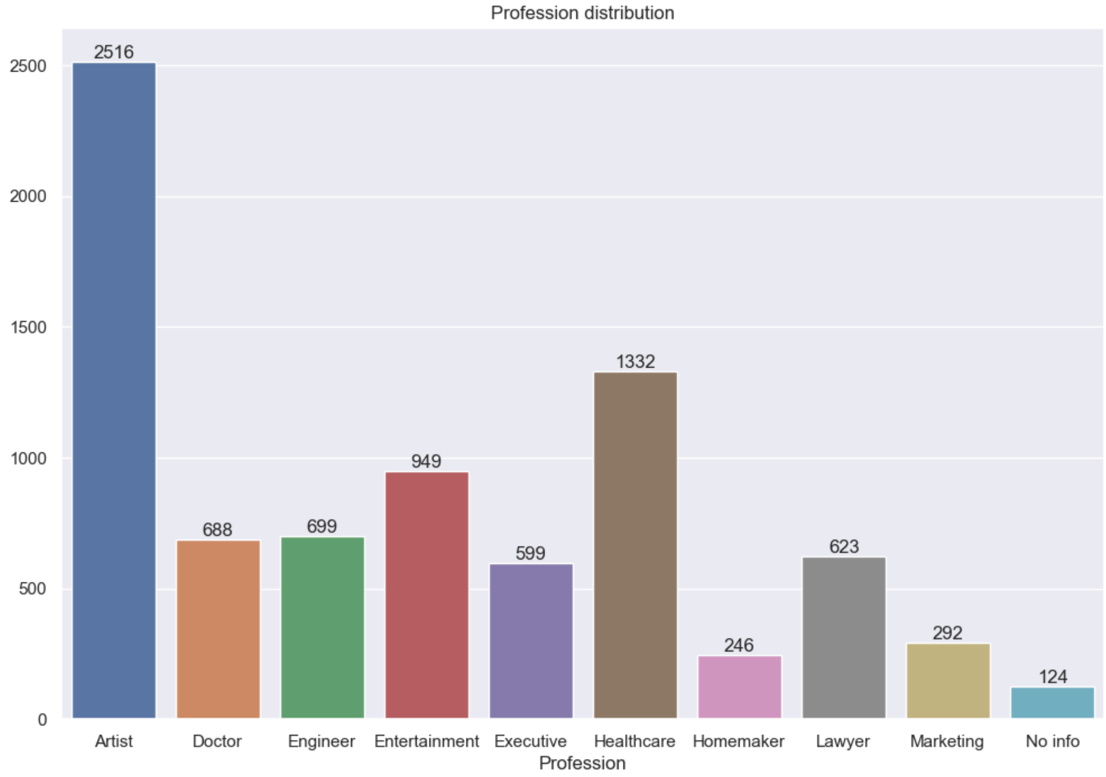


The breakdown by segment shows that segments A, B and C have a bigger proportion of graduated people. Segment C shows the highest proportional difference between graduated and non-graduated, followed by segment B and A in that order. Segment D in the other hand has the exact opposite behavior: the number of people who are not graduated is a lot higher than graduated.

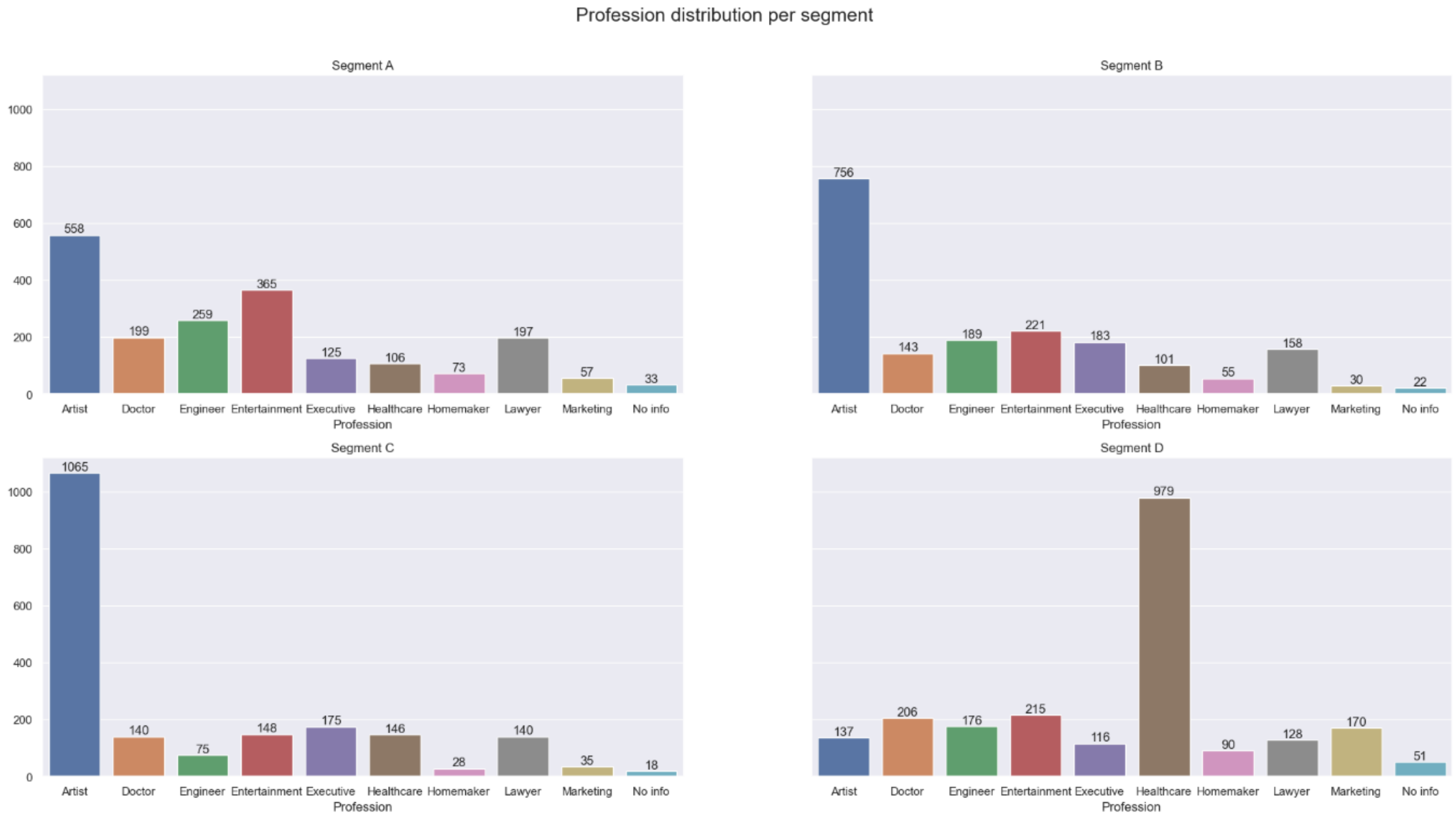


- **Profession**

The general view reveals that the biggest category is Artist, followed by Healthcare and Entertainment. The categories Doctor, Engineer, Executive and Lawyer looks similar in numbers and the two categories Homemaker and Marketing that are less representative.

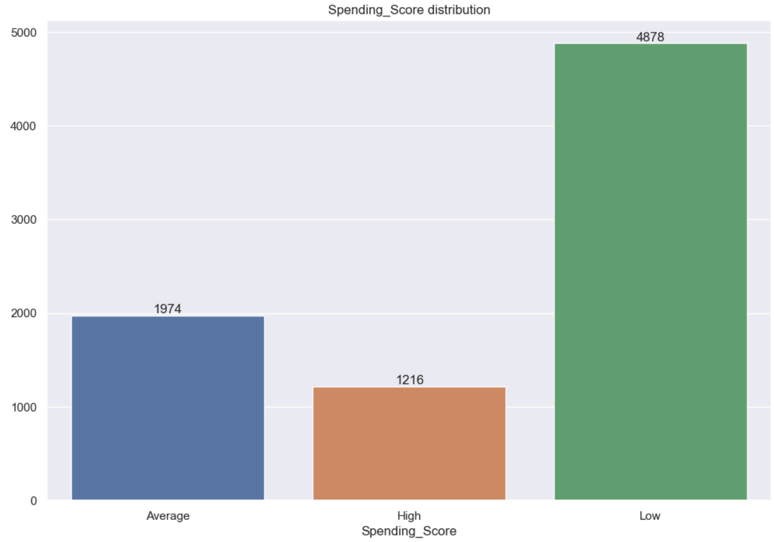


A breakdown by segment shows that segments A, B and C concentrate most customers that work as artists and includes little customers that work in healthcare. Segment D is the exact opposite, as it contains little artists and a lot of healthcare professionals. The other categories seen to be more evenly distributed except for marketing: even though is a small category, most customers that work in marketing are in segment D.

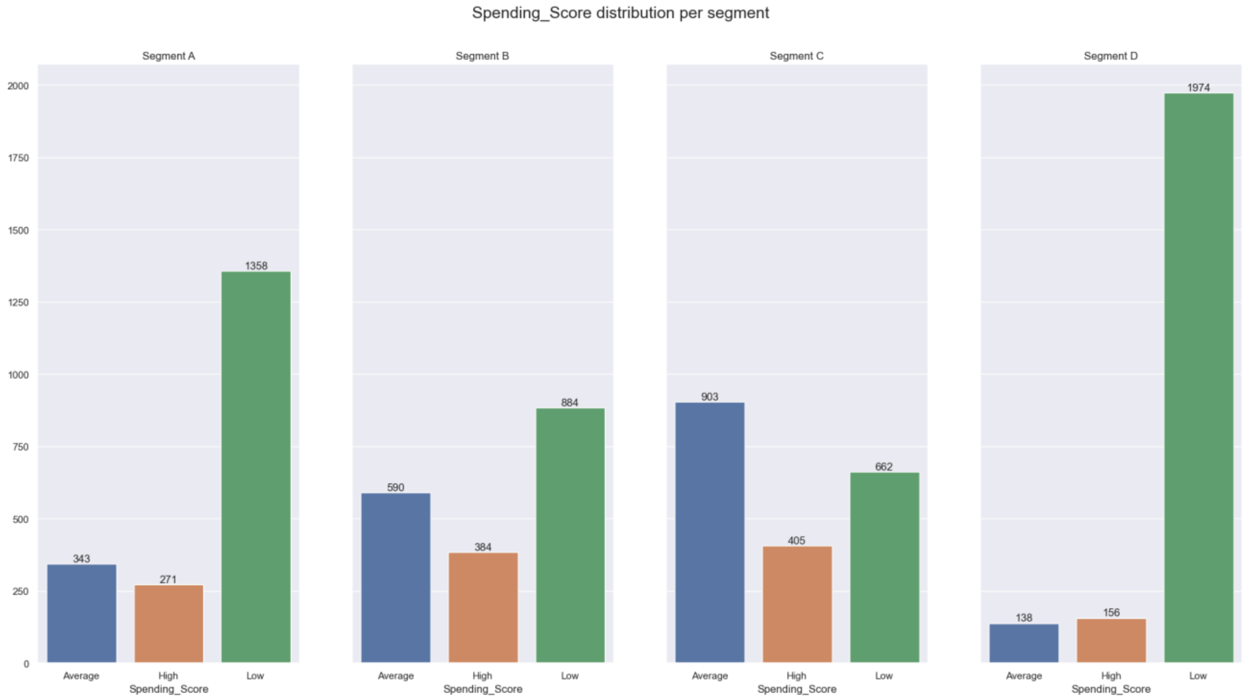

Profession distribution per segment
Segment A
1000
800
600
400
200
0
558
199
259
365
125
106
73
197
57
33
Artist
Doctor
Engineer
Entertainment
Executive
Healthcare
Homemaker
Lawyer
Marketing
No info
Profession
Segment B
756
143
189
221
183
101
55
158
30
22
Artist
Doctor
Engineer
Entertainment
Executive
Healthcare
Homemaker
Lawyer
Marketing
No info
Profession
Segment C
1000
800
600
400
200
0
1065
140
75
148
175
146
28
140
35
18
Artist
Doctor
Engineer
Entertainment
Executive
Healthcare
Homemaker
Lawyer
Marketing
No info
Profession
Segment D
979
137
206
176
215
116
90
128
170
51
Artist
Doctor
Engineer
Entertainment
Executive
Healthcare
Homemaker
Lawyer
Marketing
No info
Profession

- **Spending score**

The general view reveals that most customers have a low spending score, followed by average and high spending score in that order.

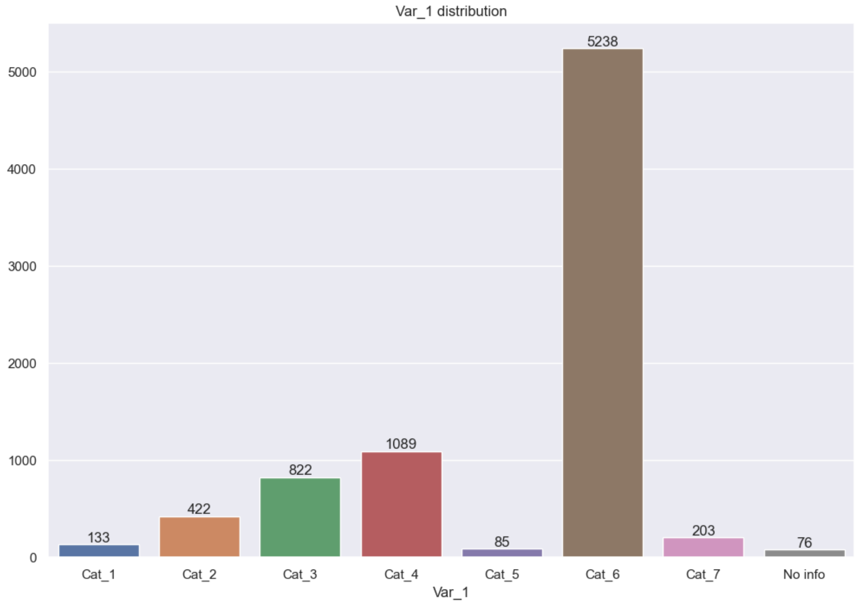


The breakdown by segment reveals that segments C and B concentrates more customers with high and average spending score. And segment C is the only one that contains more customers with an average spending score than low spending score. All others have the opposite behavior. Segments D and A concentrate more customers that have a low spending score.

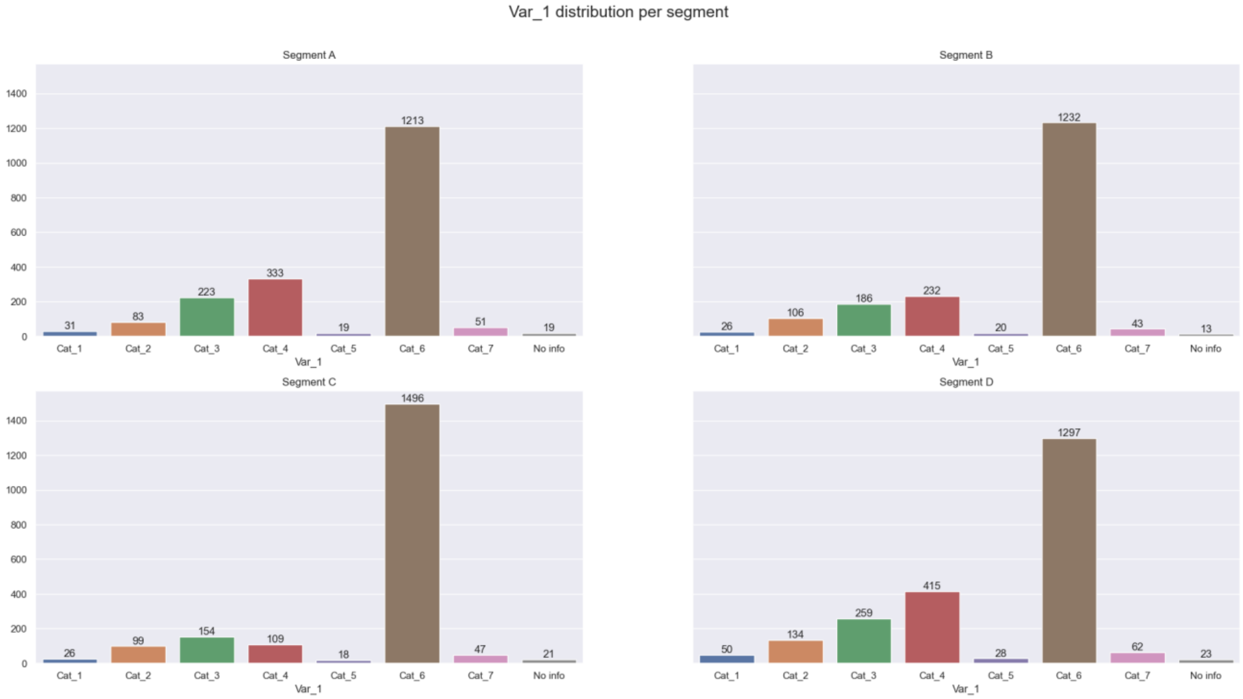


- **Var_1**

The general view shows that most customers are allocated to Cat_6, followed by Cat_4 with almost 4000 customers of difference. It draws attention to how disproportional the distribution is between the categories.



The breakdown by segment shows that Cat_6 is strongly present in all segments. And the second biggest category Cat_4 is also heavily present in all segments except for segment C.



4.6. Conclusion: segments description

After analyzing all variables in general and by segment, it is possible to describe the profile of each segment.

- **Segment A:** people mostly between 30 and 50 years old, with up to 5 years of professional experience, usually with 3 or less family members, graduated, low spending score and, although there are multiple professional fields in this groups, predominantly artists and entertainers.
- **Segment B:** people mostly between 35 and 60 years old, with up to 4 years of professional experience, usually with 3 or less family members, mostly graduated, average spending score, married and, although there are multiple professional fields in this groups, predominantly artists.
- **Segment C:** people mostly between 35 and 60 years old, with up to 3 years of professional experience, usually with 3 to 4 family members, married, mostly graduated, average to high spending score and, although there are multiple professional fields in this groups, predominantly artists.
- **Segment D:** people mostly between 20 and 40 years old, with up to 6 years of professional experience, usually with 3 to 4 family members, not graduated or married, low spending score and, although there are multiple professional fields in this groups, predominantly healthcare.

5. Data Preparation

This step consists of data preparation for the modeling. The model that will be used for this project is K-Nearest Neighbor (KNN).

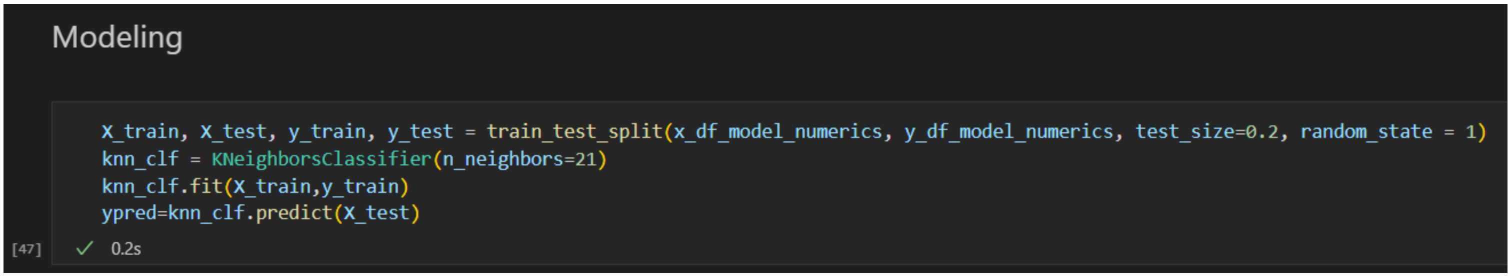
For this model, it is necessary that all input variables are numbers as it will use Euclidean distance to classify each customer to a specific segment. This way, transformations made were:

- Segmentation: A, B, C and D were replaced by 0, 1, 2 and 3 respectively
- Gender: Male and female were replaced by 0 and 1 respectively
- Spending_Score: Low, Average and High were replaced by 0, 1, 2 respectively

- Columns Ever_Married, Profession, Graduated and Var_1 were replaced by dummy columns. To avoid the dummy column trap, the dummy column ‘column_no info’ for each categorical variable was dropped.
- All columns were scaled by using the minmax method so that all columns have values between 0 and 1.

6. Modeling

Having all data prepared and using python, the model is created by just these few lines of code.



```
X_train, X_test, y_train, y_test = train_test_split(x_df_model_numerics, y_df_model_numerics, test_size=0.2, random_state = 1)
knn_clf = KNeighborsClassifier(n_neighbors=21)
knn_clf.fit(X_train,y_train)
ypred=knn_clf.predict(X_test)
```

The choice of using KNN was because this is a simple model and we studied it in class during the Minor. And the main objective of this project was to learn and practice programming in Python, using a simple and already known model that can fit in the proposed problem seen to be the most suitable.

7. Evaluation

The evaluation of the model was made by using a confusion matrix and accuracy.

- **Confusion matrix**

	A	**B**	**C**	**D**
A	169	103	44	74
B	107	119	98	39
C	41	77	232	56
D	100	45	15	295

- **Accuracy**

The accuracy achieved by this model was 50.49%. This is not high enough to make this model reliable for use, so it needs to be improved.

8. Conclusion

To conclude, the model was not very accurate, and it surely needs improvement. However, as a first try of writing a python script and modeling by myself, I can say that the main objective of the project was to learn and practice python and apply the steps of the CRISP-DM model was achieved.