The 21st of February, 2022

GROUP 9
PROJECT REPORT

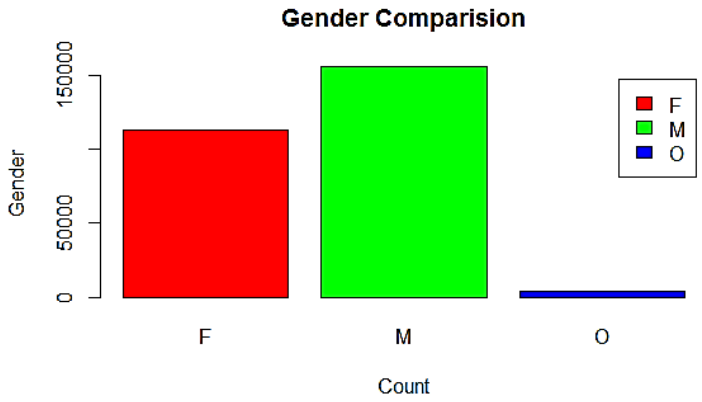
PROJECT BACKGROUND

The purpose of this project is to apply customer segmentation techniques to an unknown dataset, after it has been cleaned and studied. We would like to know which type of customers responded best to the discounts offered by the company.

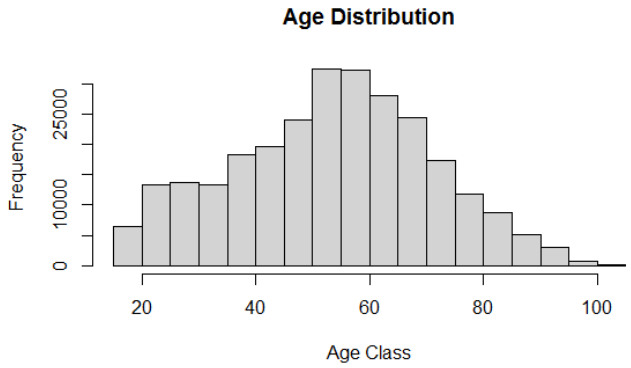
Most of this month was dedicated to exploratory analysis. This report is a summary of our findings.

EXPLORATORY ANALYSIS

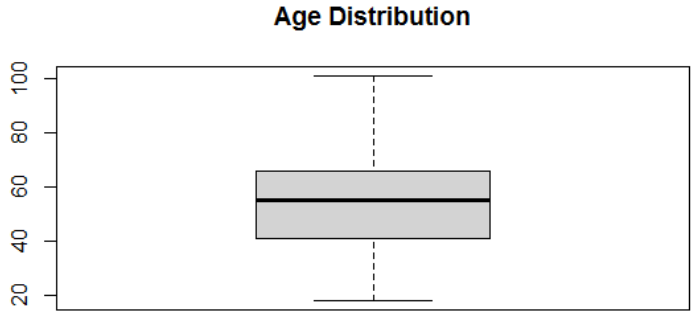
We have started our analysis by exploring the dataframes and producing plots.



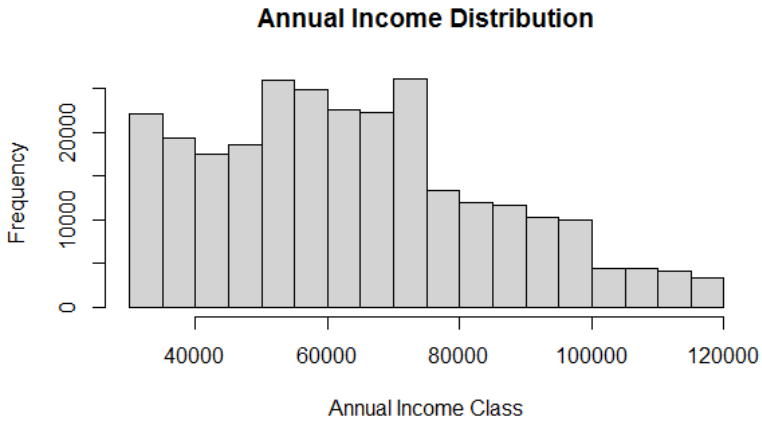
The plot above shows the distribution of gender among the customers. We can see that we have a majority of men in the dataset. We have decided to keep the missing values in the gender column (renamed “O”) : we didn’t want to assume they were indeed missing values and not deliberate – therefore valid – answers.



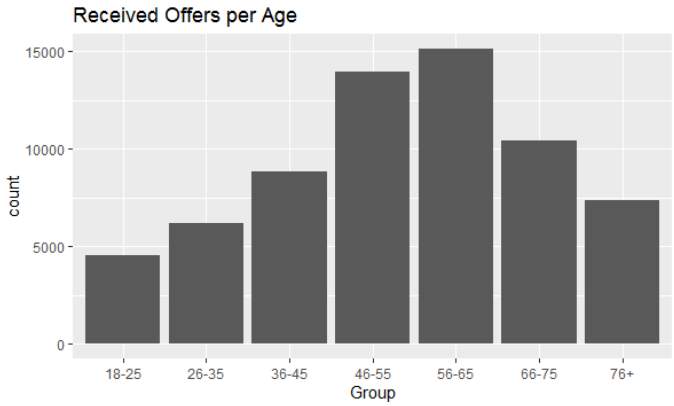
The age variable is an interesting insight into the customer base as well. It seems that the median age is around 50 years old, with 50% of the customers being between 40 and 65 years old. We have removed the number '118' as it is seemingly a code for missing values, and not a relevant age.

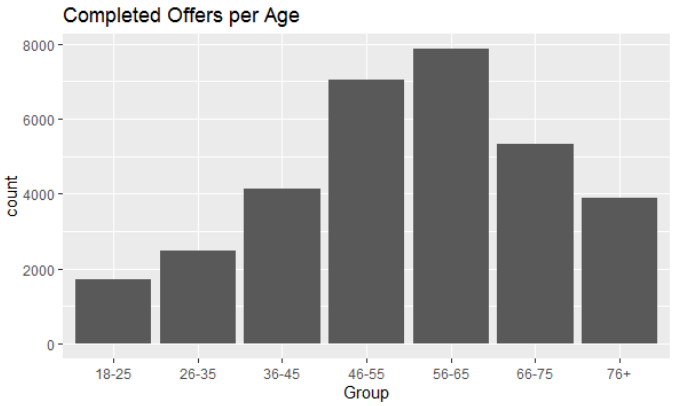


We have then focused our attention on the *income* column. On average, incomes are rather high, with most customers earning between 50k and 80k per year (in pounds).

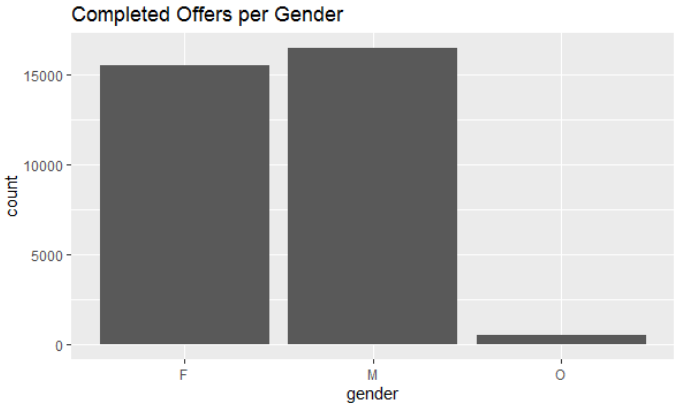


We have then compared the amount of received offers vs completed offers by age, gender and income. We wanted to analyze how groups react to the offers, in other words if we can detect difference between them.

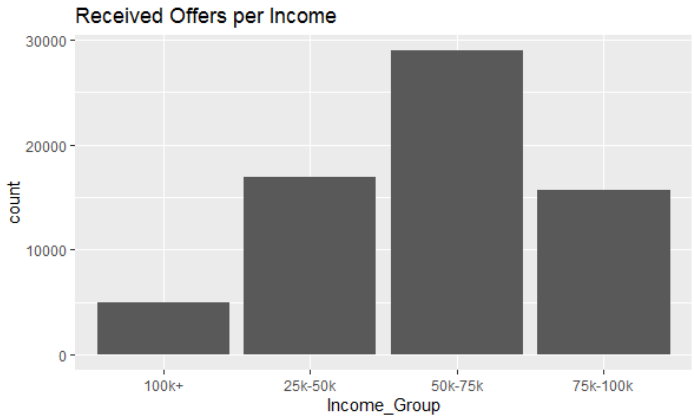


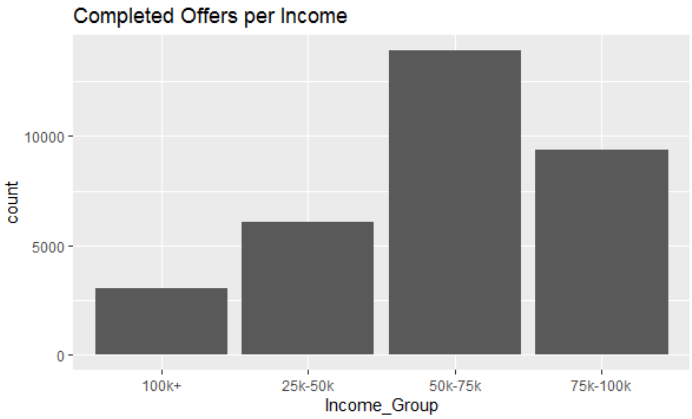


For the *age* variable, the number of offers received corresponds roughly to the number of completed offers. However, if we focus on the *gender* variable, we can see that if men receive more offers, in proportion women seem to complete offers more often.



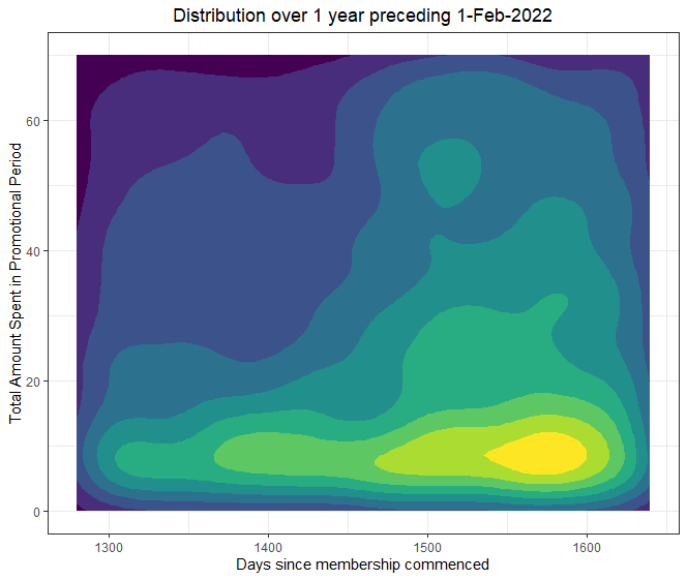
Similarly, the customers in the *income* bracket 50k-70k seem to receive and complete offers the most, however the 29k-50k bracket completes offers less often in proportion, and the 79k-100k bracket completes more offers than received, which is due to a certain amount of errors in the data. We have tried to fix this issue by filtering valid transactions.



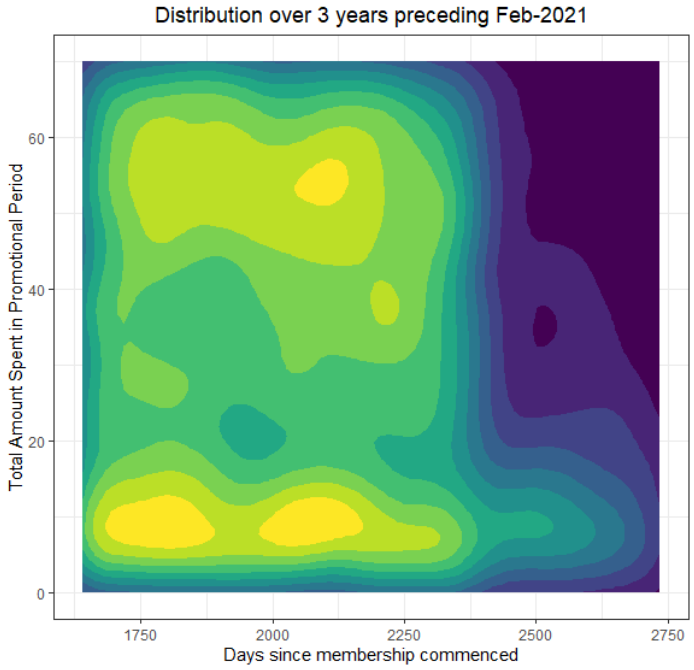


The following two plots explore how the population is distributed considering the membership commencement dates of the customers who made a purchase with a possible redemption of an offer. Also represented is the amount spent by the customers over the promotional period.

Over a period spanning more than 5 years, most of the people who have attempted to redeem offers have joined in the year preceding 1-Feb-2022. As a result, it is not possible to point out whether recent members spend more money the longer they hold membership. Filtering the customers who have become members in the one year preceding 1-Feb-2022, likelihood that recent customers have a better response to offers the longer they remain customers, is high.



In the 3 years preceding Feb-2021, a distinct trend cannot be observed. A possible explanation could be that long-time members are less receptive to the offers, they spend less, but would spend regardless of the offers.



INFORMATIVE VARIABLES

To be able to perform customer segmentation on this dataset, we have decided to primarly use the *profile* database, then add variables to it, namely the number of transactions per account, the number of offers received, viewed and completed per account, the total amount of money spent per account, as well as several proportion variables.

Our main concern was to avoid redundancy and to merge the recorded events together.

As of today, our dataframe contains 17,000 rows, corresponding to the number of customers. We have decided to focus the segmentation on 1) demographics and 2) amount spent and profitability.

The first point regroups the variables linked directly to the person, beyond their customer profile: age, gender and income will help us identify clusters.

The second point focuses on transactions: we intend on studying the amount of the transactions and their number. The aim of this approach is to link profitability with customer segmentation to bring insight from a financial perspective.