Customer Response Predict & Personality Analysis

By: Mahmoud Khaled

Content

- Data Preprocessing
- Exploratory Data Analysis
  - 2.1 Univariate Analysis
  - 2.2 Bivriate Analysis
  - 2.3 Multivriate Analysis
- Feature Selection & Dimension Reduction

Task Description

This dataset gives 2240 different customers basic information, their product purchasing preferences as well as their reactions to some marketing compaigns. I want to perform 2 tasks on this dataset:

1) Supervised Learning Task - Predict Response: As the data description says, the column 'Response' stands for if certian customer accepted the offer in the last campaign. So the question is whether we can use some customers' responses to this compaign to predict someone else's reactions ? If we can achieve this, a business could promote the compaign to customers that are more likely to accept the offer, which could help it make more efficient marketing plan.

2) Unsupervised Learning Task - Customer Personality Segmentation: This dataset also gives us some information about customers(Including their basic information and purchasing preference). So we could perform Customer Personality Analysis to help find a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to make proper market plans, like modifying and promoting products for different customers according to the specific needs, behaviors and concerns.
But before doing the tasks above, some data preprocessing work need to be done. Also, we need to do some exporatory data analysis(EDA) to help people better understand the dataset.

Task Description

This dataset gives 2240 different customers basic information, their product purchasing preferences as well as their reactions to some marketing compaigns. I want to perform 2 tasks on this dataset:

1) Supervised Learning Task - Predict Response: As the data description says, the column 'Response' stands for if certain customer accepted the offer in the last campaign. So the question is whether we can use some customers' responses to this compaign to predict someone else's reactions

If we can achieve this, a business could promote the compaign to customers that are more likely to accept the offer, which could help it make more efficient marketing plan.

Dataset Description

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Dataset Description

Promotion

- - NumDealsPurchases: Number of purchases made with a discount
- - AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- - AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- - AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- - AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- - AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- - **Response**: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- - NumWebPurchases: Number of purchases made through the company's web site
- - NumCatalogPurchases: Number of purchases made using a catalogue
- - NumStorePurchases: Number of purchases made directly in stores
- - NumWebVisitsMonth: Number of visits to company's web site in the last monthspent on gold in last 2 years

Data Cleaning

Removing Null Values

There are 24 NA rows in 'Income' columns, so we fill these NA with the average income of all people

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

Data Cleaning

Removing non important columns

- **Z_Revenue & Z_CostContact have Constant value, which don't provide any information so we should drop them.**
- **Response - AcceptedCmp5 are all Binary Variables.**

	Unique Values
Z_Revenue	1
Z_CostContact	1
Response	2
AcceptedCmp3	2
AcceptedCmp4	2
AcceptedCmp2	2
Complain	2
AcceptedCmp1	2
AcceptedCmp5	2
Kidhome	3
Teenhome	3
Education	5
Marital_Status	8
NumCatalogPurchases	14
NumStorePurchases	14
NumDealsPurchases	15
NumWebPurchases	15
NumWebVisitsMonth	16

Data Cleaning

Renaming

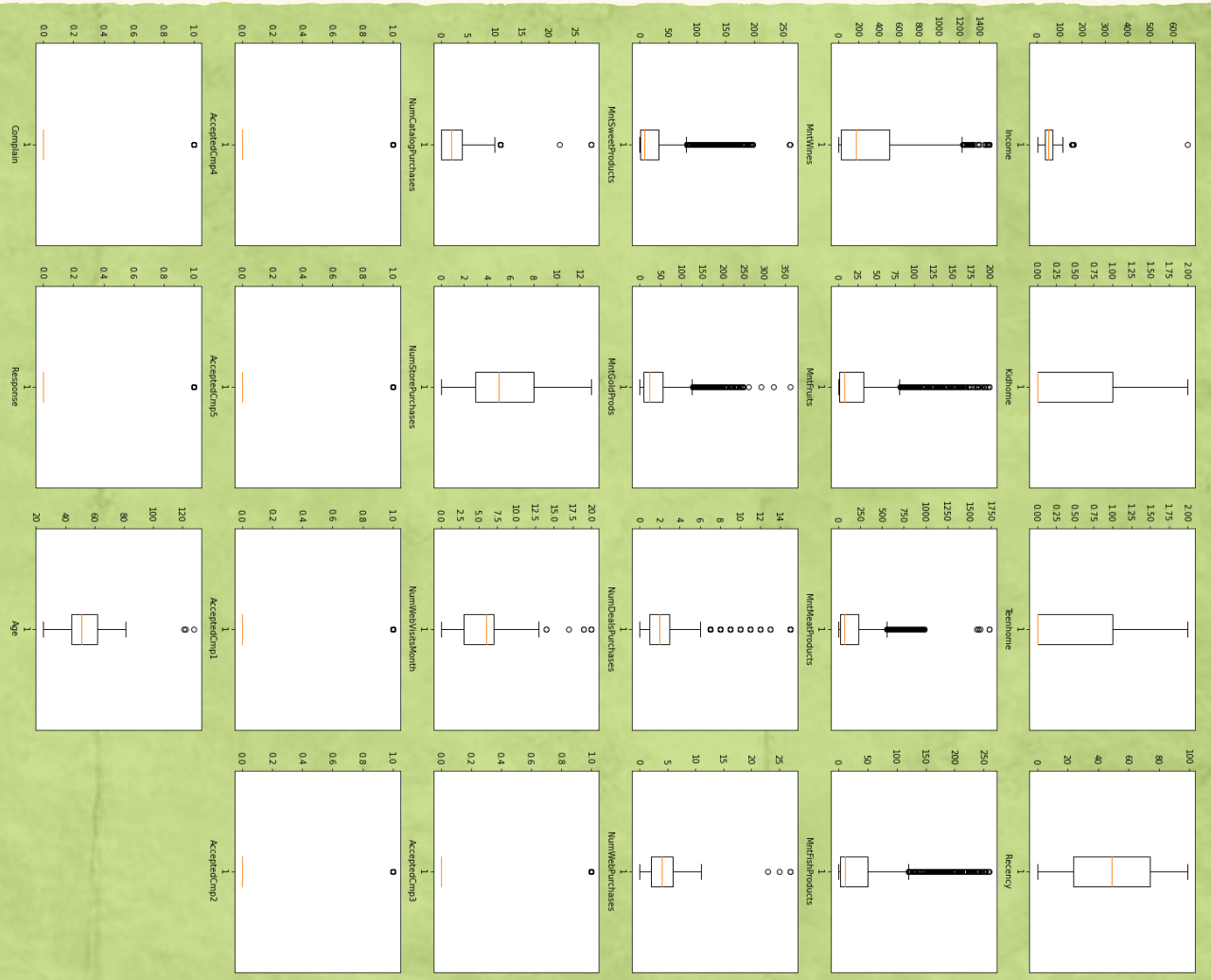
- There are too many maritial status, which might affects the efficiency of classification algorithms.
- Alone, Adsurd and YOLO are all specific situations of 'Single', so replace all of them with 'Single'

Married	864
Together	580
Single	480
Divorced	232
Widow	77
Alone	3
Absurd	2
YOLO	2

Data Cleaning

Detecting outliers

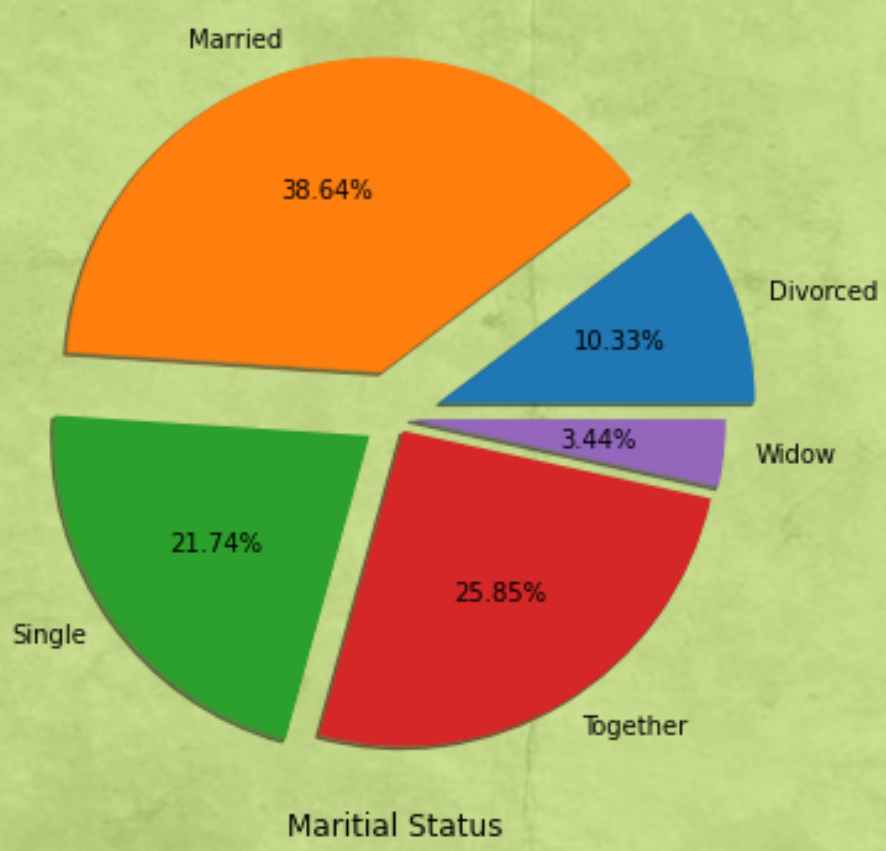
- **We can see some clear outliers in Income and Age. We will remove the rows where the Income is greater than 200K and birth year is less than 1920.**
- **For other columns, we cannot blindly remove these outliers as there could be cases where the requirement for these products is high by the user. Maybe the consumer is hosting a party or an event or is more comfortable getting his products from a particular channel.**

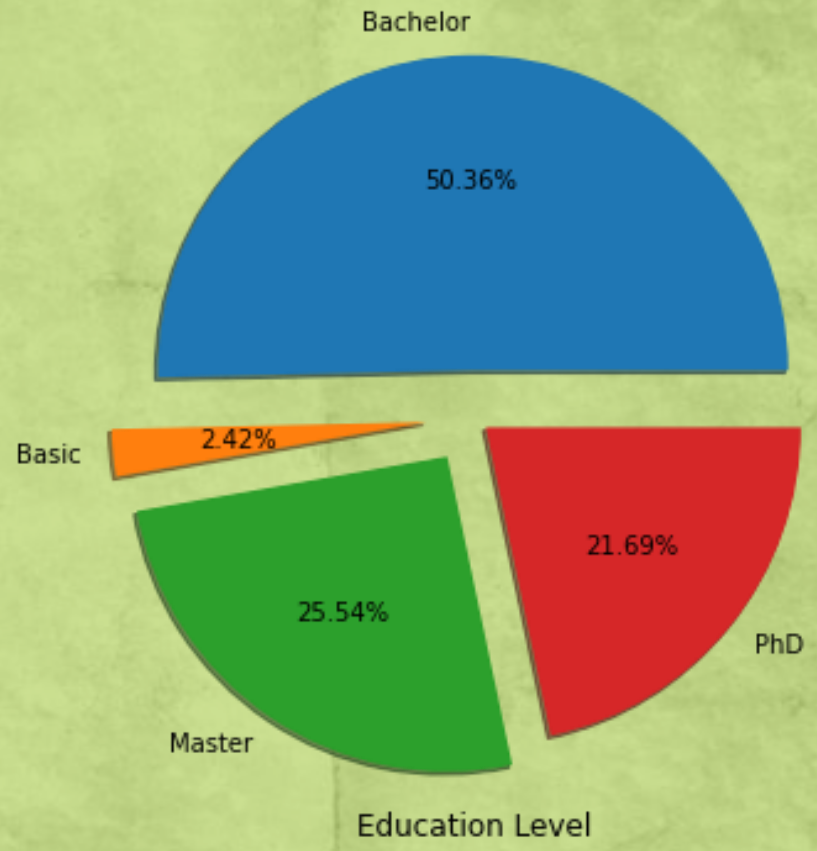


Exploratory Data Analysis

Univariate Analysis

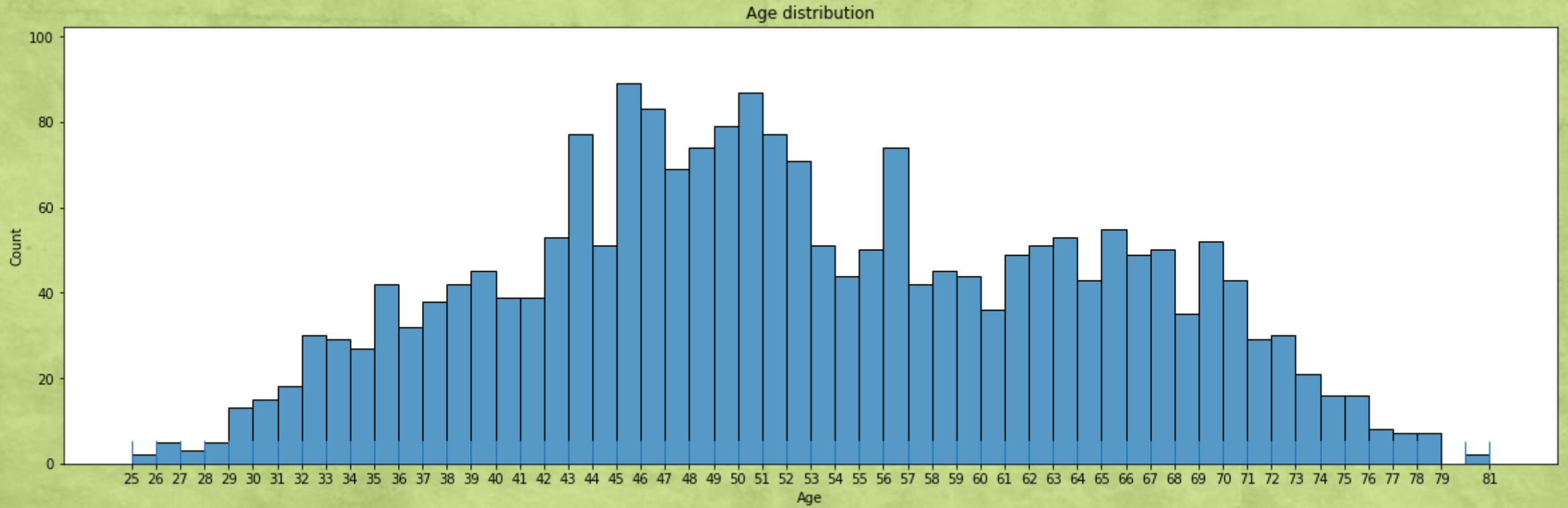
- **These 2 figures give a quick look of the customer distribution, we could see the most our customer(64%) are in relationships(Married or Together) and most(97%) are at least bachelor degrees.**





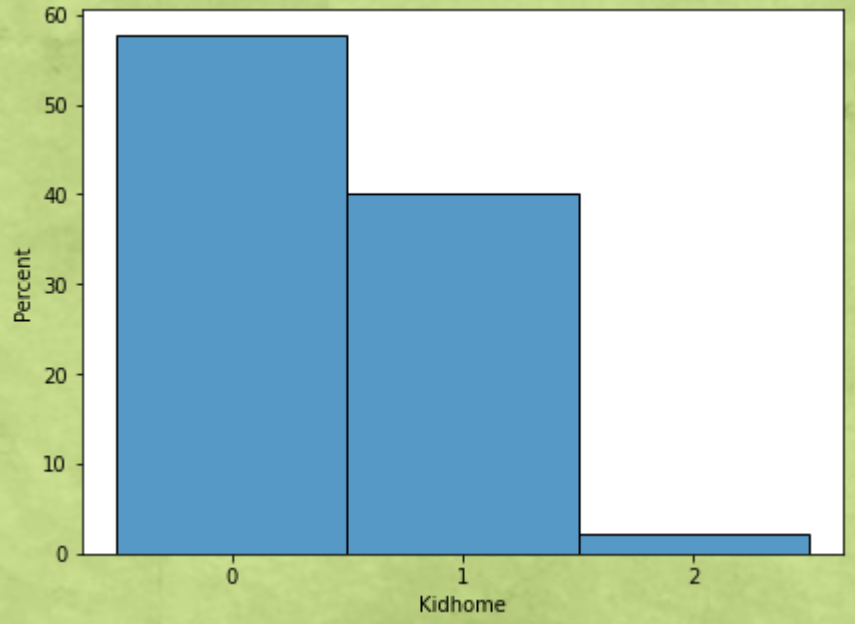
Exploratory Data Analysis

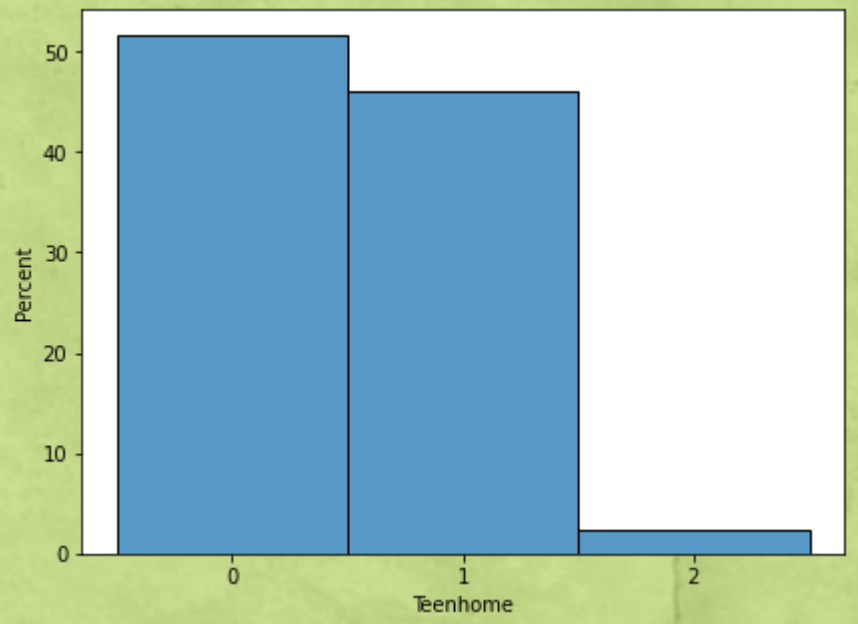
- **The age of the customers are mainly clustering in 40s or 60s, the young people(under 30) are very few.**
- **These people are at their middle ages or old ages, so their family condition should be further taken care of.**



Exploratory Data Analysis

- **Most customers have 1 kid or 0 kid at home, very few have 2 kids, and no one have kids over 2.**
- **The number of teens at home is very similar to kid num.**

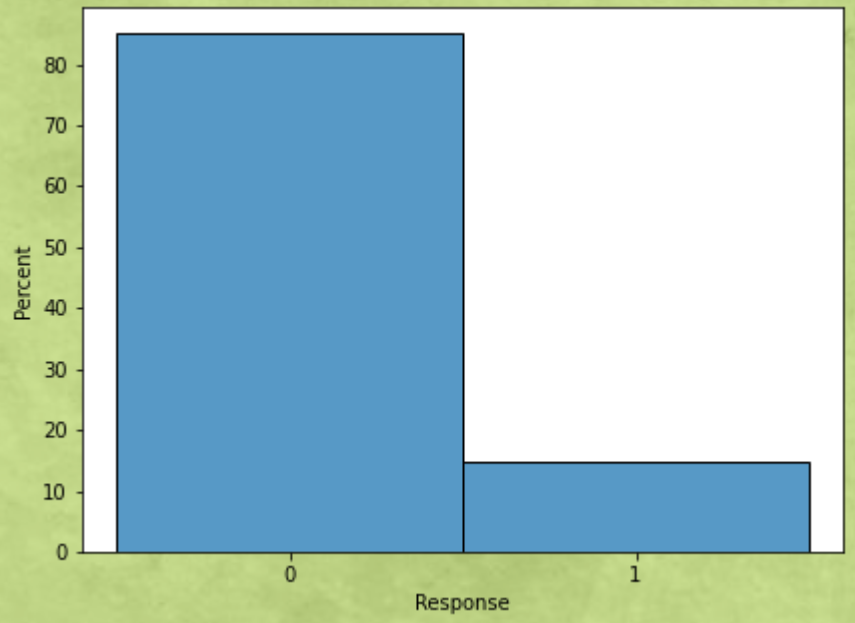


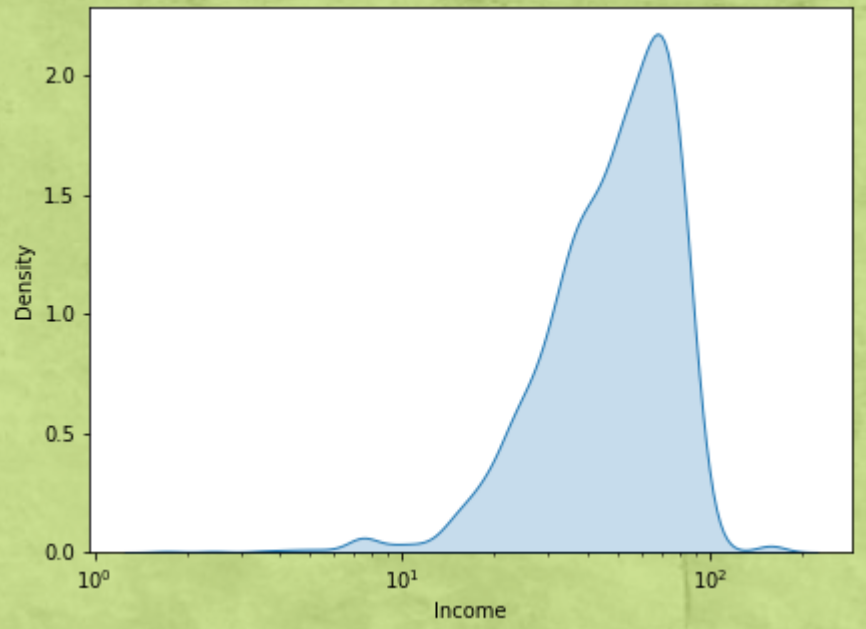


Exploratory Data Analysis

•According to Response, this is an unbalanced dataset, over 80% customers say no to the last compaign. So we should take care and use more comprehensive and accurate indicators(like F-1 or MCC) to evaluate the classification models.

•Most customers income levels are in the range [10K, 100K] per year.

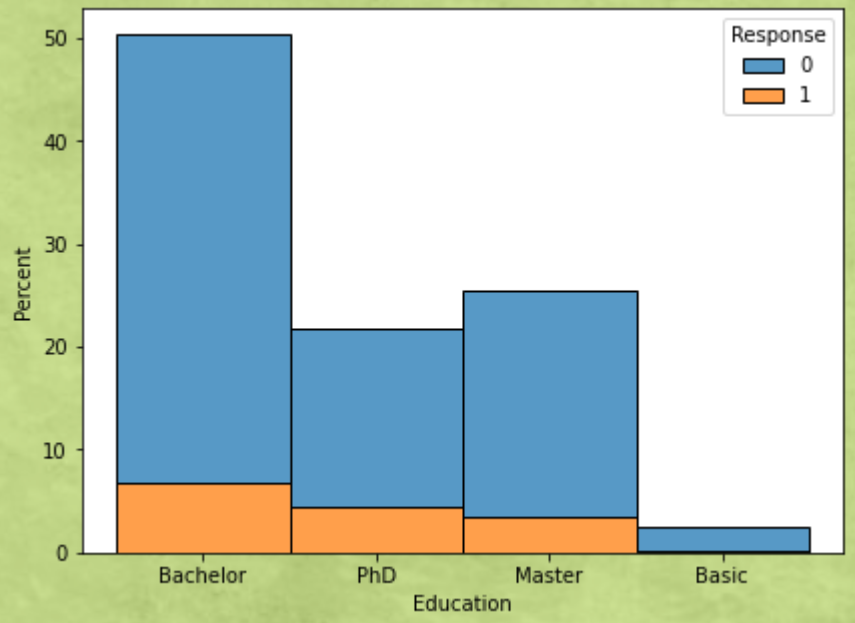


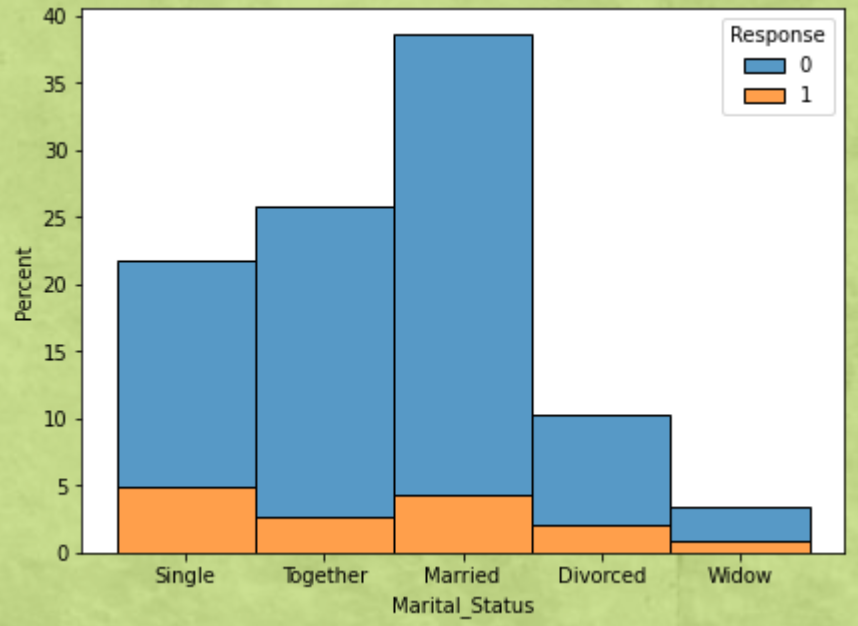


Exploratory Data Analysis

Bivriate Analysis

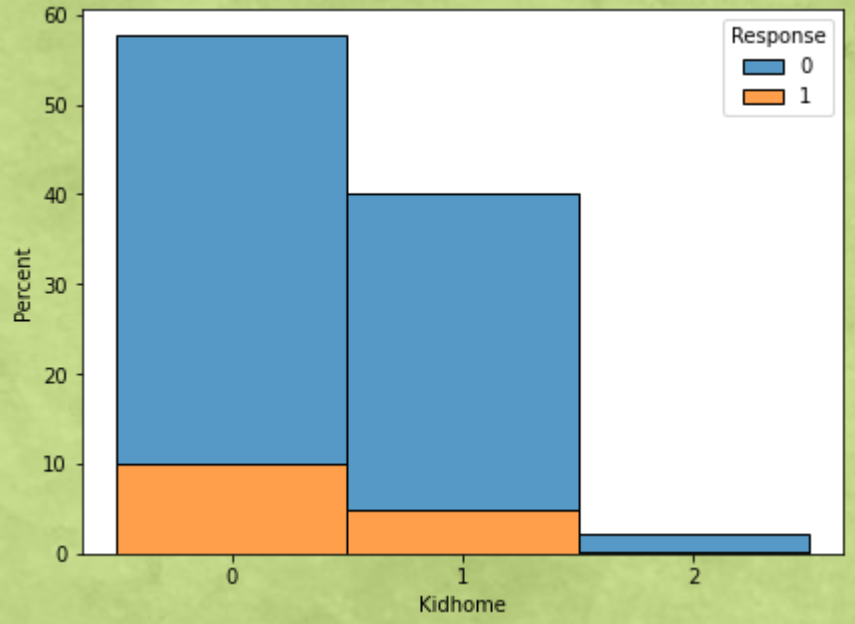
•From the left figure, we can find that the compaign acceptance rate in high education groups(Master and PhD) are higher than that in low education groups.
•From the right plot, we find that the single people tend to say yes to this compaign.

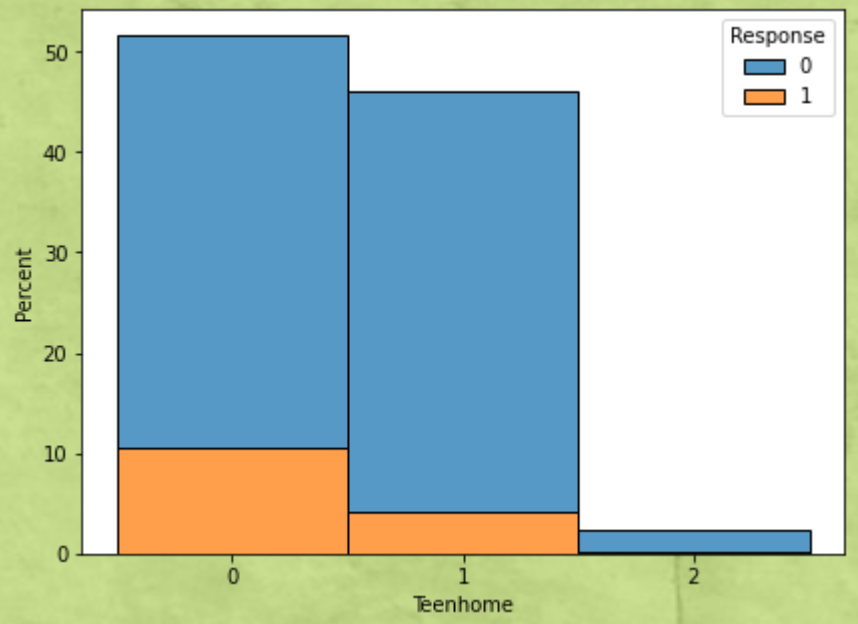




Exploratory Data Analysis

•It seems that customers with no kids and no teens at home are more likely to accept the offer in this compaign.





Preparations for Prediction Models

- **Converting Catg data to numerical data**

Education have orders, so we change Basic-Phd to scale 0-3

Change Marital_Status to dummies

Splitting data Train and Test

```
x_train = final_data[:2000]
y_train = data['Response'].values[:2000]
x_test = final_data[2000:]
y_test = data['Response'].values[2000:]
```

SMOTE

Why SMOTE (Dataset Extended Trick) & MCC Scorer (Matthews correlation coefficient) ?

•As we find in the data exploration phase, this is an unbalanced dataset(over 80% say no to the compaign). So the models are easy to learn some traits about negative samples, but it might be hard to get from positive samples.

•While SMOTE alleviate the problem by offering us more positive training samples.

•At the same time, MCC scorer takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. In this learning task, MCC is a more efficient measure than accuracy in test period, because there are only a few positive samples in the test set.