Qs 1. C.

	Methodology used	Underlying rationale
Handline Outliers	Removing the values that are outside the quartile range	Easiest and fastest way to identify outliers for each column
Handling missing values	Fill with mean	Assuming that we would proceed with normalization later

Qs1 e.

	Patterns observed	Respective insights
Pattern 1	A few histograms for attributes are skewed	Remove the outliers
Pattern 2	Amount spend on Wines, Fish,Gold and Meat reduce as the number of kids in the house increase	Could use feature engineering that captures the interaction between the number of kids in the household and the amount spent on each product.
Pattern 3	Number of catalog purchases is zero for people who belong to alone under Marital_Status	Select models that can handle missing data

Qs 2.a

Column	Data type	Data Processing Technique
ID	nominal	Dropped as no relevance to learning process given it is just a unique identifier for each of customer
Year_Birth	Discrete	Left as is as can be used to determining relationship between age profiles and prediction variable
Education	ordinal	Rank Replacement
Marital Status	nominal	One hot encoding
Income	Continuous	Normalization and Missing values(already handled)
Kidhome	discrete	Normalization should have been followed but given a very small range of values (0,1,2) columns can be left as is.
Teenhome	discrete	Normalization should have been followed but given a very small range of values (0,1,2) columns can be left as is.

Dt_Customer	ordinal (as there might be relationships to explore between prediction and old/new customers)	Seperated into day of week, year , month
Recency	discrete	Left as is it can be further used for finding relationships between last days of purchase and the prediction variable
MntWines	Continuous	Normalization
MntFruits	Continuous	Normalization
MntMeatProducts	Continuous	Normalization
MntFishProducts	Continuous	Normalization
MntSweetProducts	Continuous	Normalization
MntGoldProds	Continuous	Normalization
NumDealsPurchases	discrete	Normalization
NumWebPurchases	discrete	Normalization
NumCatalogPurchases	discrete	Normalization
NumStorePurchases	discrete	Normalization
NumWebVisitsMonth	discrete	Normalization
AcceptedCmp3	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed
AcceptedCmp4	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed
AcceptedCmp5	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed
AcceptedCmp1	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed
AcceptedCmp2	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed
Complain	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y

Z_CostContact	discrete	No preprocessing required as value constant through data so can remove them under the pretext of feature engineering
Z_revenue	discrete	No preprocessing required as value constant through data so can remove them under the pretext of feature engineering
Response	nominal	No modifications made as all rows contain 0s and the column would eventually be part of Y hence,removed

****FOR ALL THE NUMERICAL COLUMNS OUTLIERS HAVE BEEN REMOVED**

Q4b.

The final model that gives the best results is Gradient Boosting Classifier(GBC). Though random forests and gradient boosting have very similar accuracy scores the F1-score for GBC was higher than random forests capturing both accuracy and precision. Furthermore, Gradient Boosting was chosen because it is a more flexible model with many parameters, which tuned accurately can lead to better results.

Qs6 .

1. Besides the variables used to make up the predictor, the most influential/important features are the amount spent on wine,meat,gold products and income.
2. From the SHAP summary plot we can also see that all the variables that went in final prediction have a positive impact when they are lower in value (closer to 0, this is due to the fact that the model is training with data which is zero in these columns)
3. Precision is also higher for predicting 0 class

Limitations and Future recommendations:

1. The model will not be able to correctly classify those that will accept the offer given that most of the data is 0 in columns which are used for the predictor variable
2. In future a dataset with an equal distribution between positive and negative classes , for variables used for prediction can lead to a more robust model