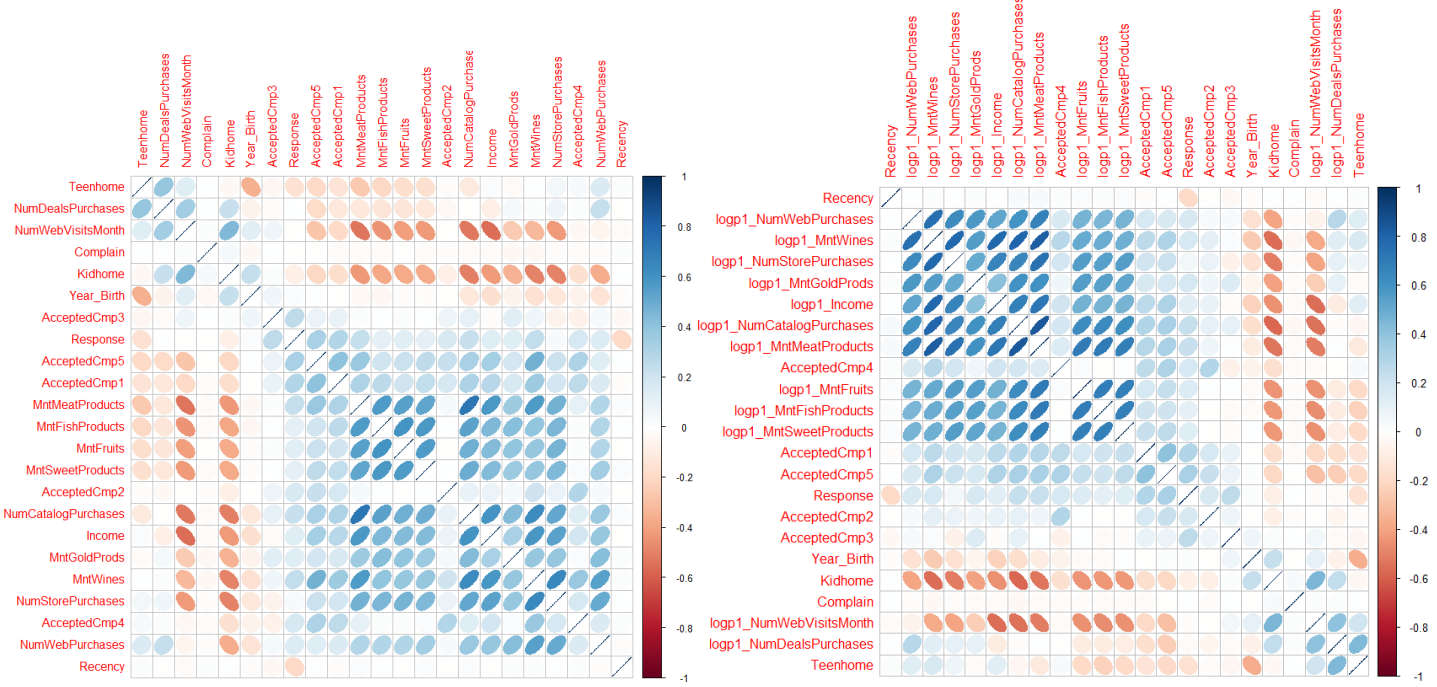
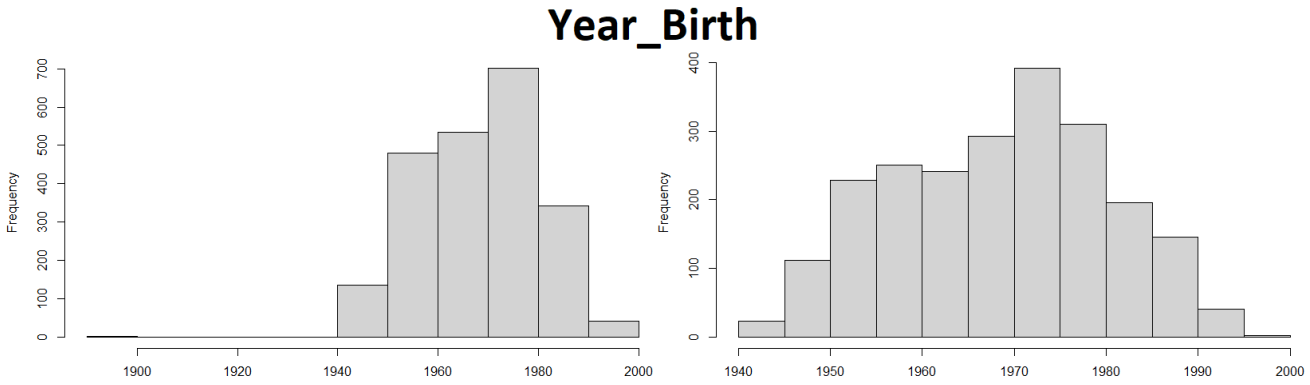
Exploratory data analysis

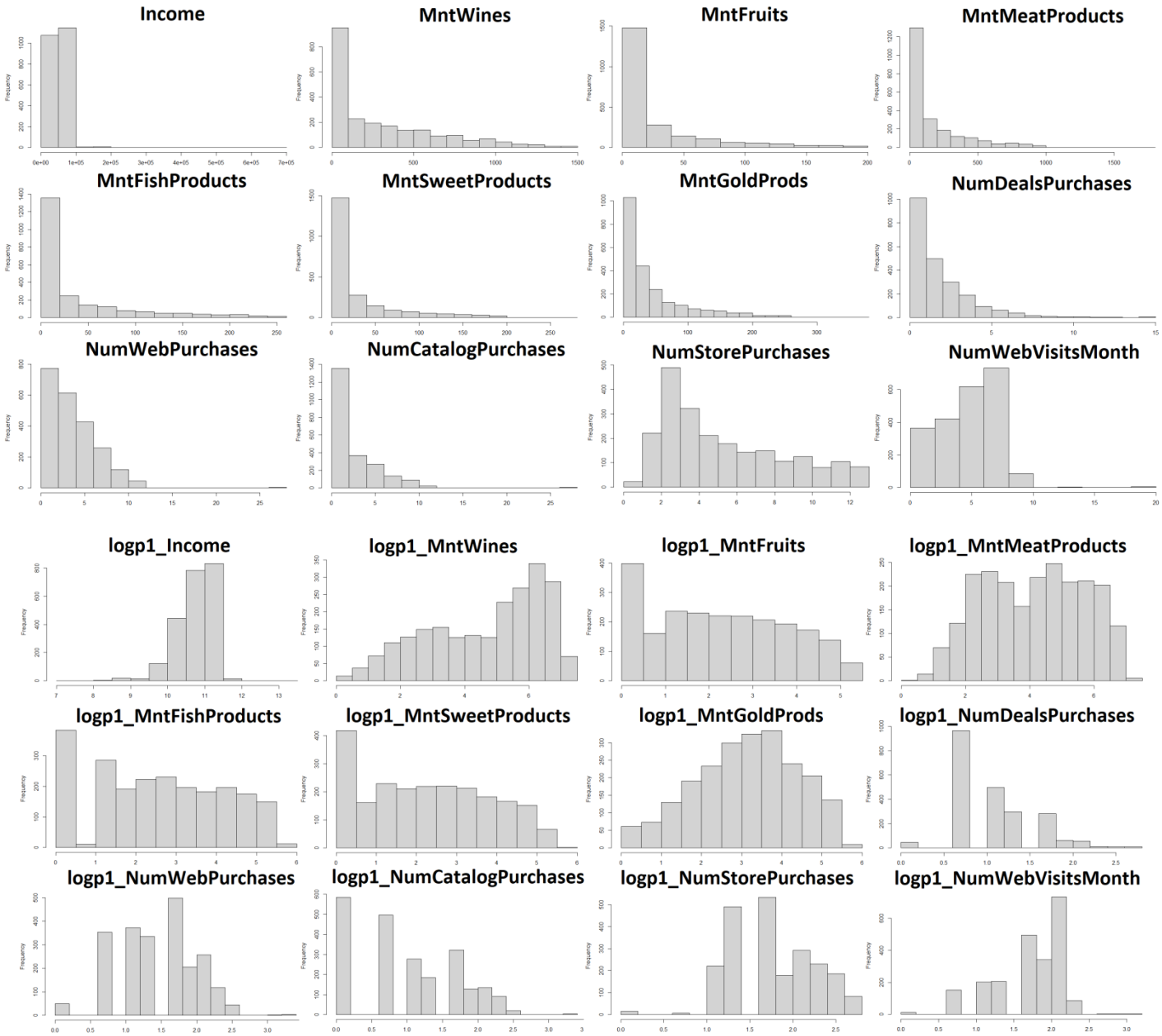
As can be seen on the correlation matrix on the left, many variables were significantly correlated even before performing variable transformation. However, as can be seen on the correlation matrix on the right, the correlation between many of the variables strengthened significantly with the transformations performed.



The first variable to be analyzed was “Year_Birth”, which was left skewed due to the presence of three outliers. After removing said outliers, without changing any relationships in the dataset, “Year_Birth” was centered as can be seen below. Variables “Marital Status” and “Education” had a few values combined into one. For the first, “Single” and “Alone” were combined. For the latter “Masters”, “PhD” and “2n cycle” were combined, due to the understanding that this dataset is of an unknown international precedence, where “Graduation” is equivalent to an undergraduate degree in the US and “Second Cycle”(“ 2n cycle”) is the equivalent to a graduate degree in the US.



There are over two thousand observations in the dataset, which is way more than the needed amount of observations to approximate a normal distribution. Nonetheless, as can be seen on the figures below, the transformations performed contributed significantly to normalize the variables involved. The chosen transformation was the log plus one, considering that the immense majority of the numeric variables had some value equal to zero and that these variables were all right skewed.



Variables "AcceptedCmp" one through five, as well as "Response" were binary and received zero for when a customer did not accept an offer on a given stage or one when a customer did accept an offer on a given stage. All the six variables were combined into the variable "Accepted", which receives a zero if a customer did not accept any offers at any stage, or an integer equivalent to the stage in which a customer accepted an offer. The distribution of the variable can be seen on the figure below, showing that most customers did not accept any of the offers and the remaining offers had a somewhat evenly distributed acceptance throughout the six stages. Variables Marital Status (Single, Living together, Married, Divorced or Widow, from 0 to 4, in this sequence) and Education (High school, undergraduate or graduate, from 0 to 2, in this sequence) were transformed into dummy variables and their distribution can be seen below as well.