Summary and Analysis

Summary:

First of all, the data cleaning process was done where a lot of empty and unwanted values were cleaned. Specifics on this are written in comments in the code as well.

Our main target variable ‘Personal Loan’ had values 0 and 1 but, they weren’t all integer, so typecasting was done.

Standard Scaling is a common step in ML algorithms to ensure all features have the same scale, which are normalized.

Label encoding data because 'Gender' and 'Home Ownership' are not numerical values. It basically assigns a unique integer to each unique category in a column, which is good for Random Forest Classifiers.

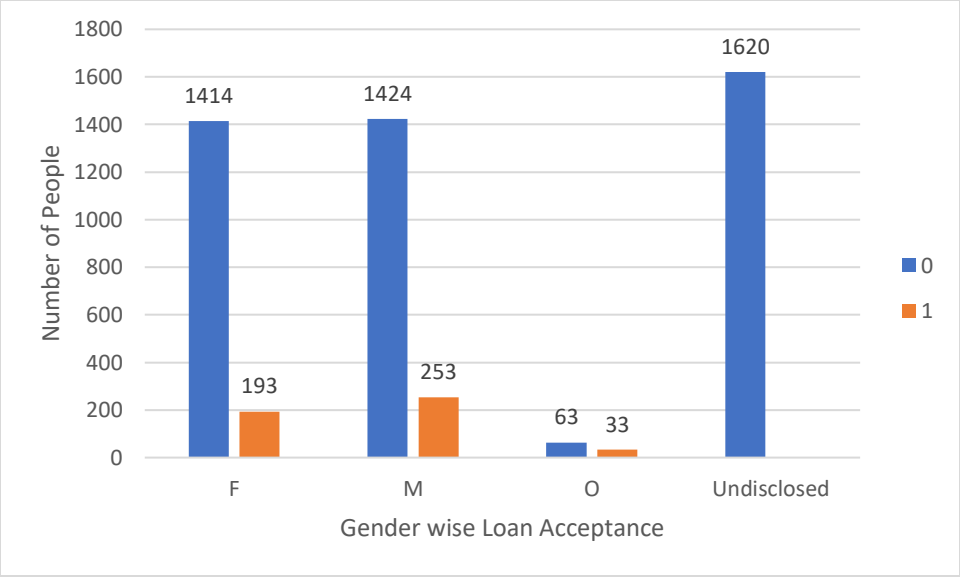
Random Forest Classifier (RFC) was chosen as the classifier due to its advantage of maintaining its accuracy despite missing data fields. In this code, I have tried my best to handle the missing

data in preprocessing steps but even in case of any missing values RFC will not be severely affected. RFC can also easily deal with large datasets like this.

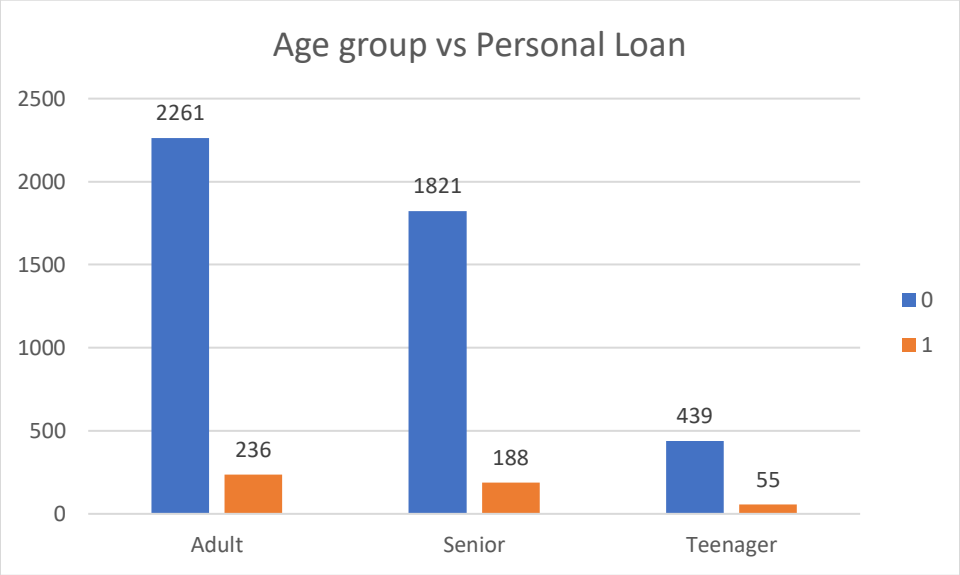
Random state is defined to maintain consistency split in each run and make it replicable. So, our example data predict1 and predict2 will always be predicted the same.

Distribution of the numeric variables are shown as a part of EDA. And the following correlation plots are achieved through pivot table for analysis:

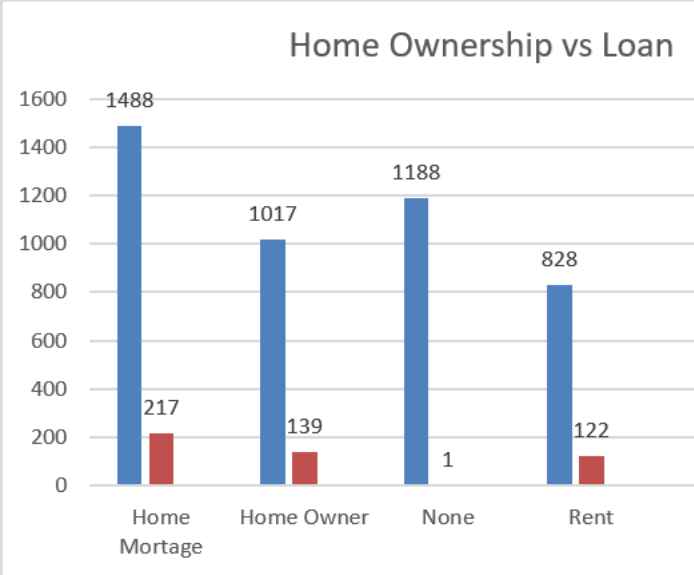
Analysis:



1) **Personal Loan's correlation with Gender:**
Empty gender fields are renamed to 'Undisclosed' for not assuming their genders.
Looks like there is no discrimination between the genders for loan acceptance.



2) **Personal Loan's correlation with Age Group:**
Ages 50+ are grouped as senior, 30+ as adults, and the rest as teenager. It seems that the probability of a person's loan being accepted is more if they are adults.
Although this value maybe skewed due to the high number of adults among the given data.



3) **Personal Loan's correlation with Home Ownership:**
Empty values of Home Ownership are replaced with 'None'. It seems that the probability of a person's loan being accepted is more if they have mortgaged their home.
Although this value maybe skewed due to the high number of people with Home Mortgage.