Preprocessing

```r
table(df$Default)
```

```
## 
##     0     1 
## 29653 29653
```

After thoroughly inspecting our data, there seem to be no issues with missing values and duplicates. There no obvious data entry errors and typos and the categorical variables have acceptable frequencies across all levels deeming feature engineering unnecessary. However, there was one glaring issue which was the strong imbalance in the outcome variable for Defaulted loans. There were 225694 loans that didn't default as opposed to 29653 that did. As a result, leaving our data like this made our models predict everything as a 0 which is why we decided to balance our data with undersampling. We randomly sampled 29653 out of the 225694 non-defaulted loans and combined it with our 29653 defaulted observations to fit the models on. While we are losing information, the sample size was quite large so we still have a lot of data to work with.