# Exploratory analysis

1. TODO: Rimuovere Dummy2Borrower !!!

IntRate\_Distribution.png

**Data Cleaning**

1. remove redundant or Non-Informative columns
2. clean up numerical codes that indicate "non applicable" or "unknown"

**Minority status analysis**

We first need to create a dummy indicating whether a person comes from a minority group or not, using the variable *Borrower1Race1Type:*

* 1=American Indian or Alaska Native,
* 2=Asian,
* 3=Black or African American,
* 4=Native Hawaiian or other Pacific Islander,
* 5=White,
* 6=Information not provided by Borrower,
* 7=Not Applicable (First or primary borrower is an institution, corporation or partnership)

*INSERIRE Race\_Dist.png*

We can see that the distribution is not balanced, with a vast majority of loan applicant belonging to the category "White" (5), and much fewer applicants in the other categories.

Therefore, we will group categories 1, 2, 3 and 4 into a single minority category.

*Minority\_Dummy\_Dist.png*

Since this classification presents a higher number of minority observations and less (in fact, 0) "Not Applicable" data points, we will keep this classification to determine whether a loan applicant is from a minority or not. We therefore pass this mapping down to the dataframe.

It must be noted that the dataset is clearly unbalanced since we have a vast majority of observations that belong to the non-minority category, and this will need to be taken into account during the analysis.

We will drop the Ethnicity information since we can consider it redundant as we don't have any additional information on how it was recorded, but we will keep the Race varibale in order to inspect possible differences among different ethinc groups among non-White individuals.

Riprendere da **Analysis of other race variables**

Fine

NO: We have two different variables that we can use, so we explore them more in detail.

**Variable 1: Borrower1EthnicityType**

* 1=Hispanic or Latino;
* 2=Not Hispanic or Latino;
* 3=Information not provided;
* 4=Not applicable (First or primary borrower is an institution, corporation or partnership)

*Inserire Borrower1EthnicityType\_Dist.png*

We can see that the distribution is not balanced, with a vast majority of loan applicant belonging to the category "Not Hispanic or Latino" (2), and too many applicants in the "Hispanic or Latino" category (1), which would be our minority. We now investigate our second option.