# 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.

**Analysis of other race variables – menzionare velocemente**

The dataset presents 4 other variables indicating other ethnic traits of borrowers, such as Borrower1Race2Type, Borrower1Race3Type, etc. We will now inspect their values.

As we can see, these variables are not meaningful for our analysis, since most observations fall under the categories: 5=White; 6=Information not provided; 7=Not Applicable (borrower is an institution, corporation or partnership). Hence, we can exclude these variables from our data.

A comparison of a bar graph

Description automatically generated

**BORROWER 2 RACE**

Considering the variables about the second borrower’s ethnicity, we find once again a majority of white applicants (~~category 2 and 5 respectively~~), and a consistent number of loans for which there is no co-borrower (~~category 5 and 8 respectively~~). However, we can find the same patterns as before, that is we can hypothesize that all races other than white are classified as "Hispanic or Latino" ethnicity (category 1) and other race variables do not provide any additional information. We will therefore proceed in the same way by grouping minorities into one category and dropping the other variables.

A screenshot of a graph

Description automatically generatedA graph of different sizes and shapes

Description automatically generated with medium confidence**Numerical variables:**

* Histogram oppure boxplot?

Outlier detection and **Winsorization**

We decide to recode the most extreme values using **winsorization**, in order to obtain a more balanced dataset while keeping the information content of the extreme values. The threshold used is obtained using the interquartile range, a measure of the spread of the middle 50% of the data. In particular, the lower threshold is then established by subtracting three times the IQR from the first quartile and the upper threshold by adding three times the IQR to the third quartile. Any data points falling below the lower threshold or above the upper threshold are considered outliers.

QUESTE???? METTERLE?

MinorityRatio winsorized on the right side

CensusTractMedFamIncome winsorized on the right side

LocalAreaMedianIncome winsorized on both sides

MonthlyIncome winsorized on the right side

HUDMedIncome winsorized on the right side

UPB winsorized on the right side

LTV winsorized on both sides

HousingExpense winsorized on the right side

DebtExpense winsorized on the right side

NoteAmount winsorized on the right side

Two exceptions to this process are the variables *PaymentCount* and *PMI.* The variable *PaymentCount* represents the Term of the Mortgage in Months, and we will prevent winsorization because of the little meaning it could have and also because of the highly skewed distribution, which will imply that we would lose all information content of the variable.

The variable *PMI* represents the percent of mortgage balance at origination covered by loan level PMI (Private Mortgage Insurance), and it is also highly skewed towards 0. By the same reasoning, modifying this variable will lead to loss of information therefore it will not be subject to winsorization.

**Correlation Analysis – da qui**

Insert *corr\_heatmap.png* oppure *corr\_heatmap2.png*

PRENDERE COMMENTI !!!!

CLEANING: COSE DA DIRE

* We first dropped observations with null credit score for the main borrower, since it has an important impact on the assigned interest rate.
* Then, we assessed the proportion of missing values on the dataset and decided to drop them in cases where they constituted a small percentage of our dataset. For instance, the percentage of missing values in the variable Age1 was 0.002%.

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.