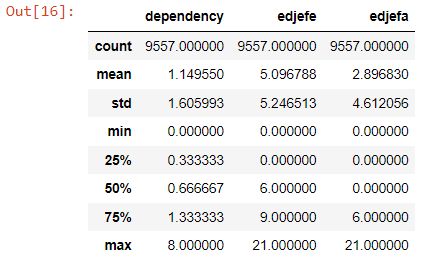
**Income Qualification**

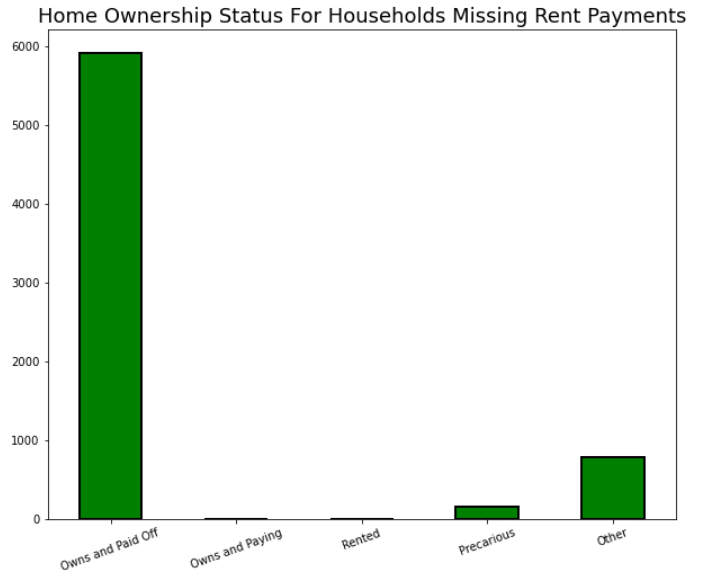
* The important piece of information here is that we don’t have ‘Target’ feature in Test Dataset.
* There are 3 Types of the features:
* 5 object type
* 130(Train set)/ 129 (test set) integer type
* 8 float type
* **Columns for different Datatypes:**
* Looking at the different types of data and null values for each feature. We found the following:
* No null values for Integer type features.
* No null values for object type features.
* For float64 types below features has null value:-

|  |  |
| --- | --- |
| v2a1 | 6860 |
| v18q1 | 7342 |
| rez\_esc | 7928 |
| meaneduc | 5 |
| SQBmeaned | 5 |

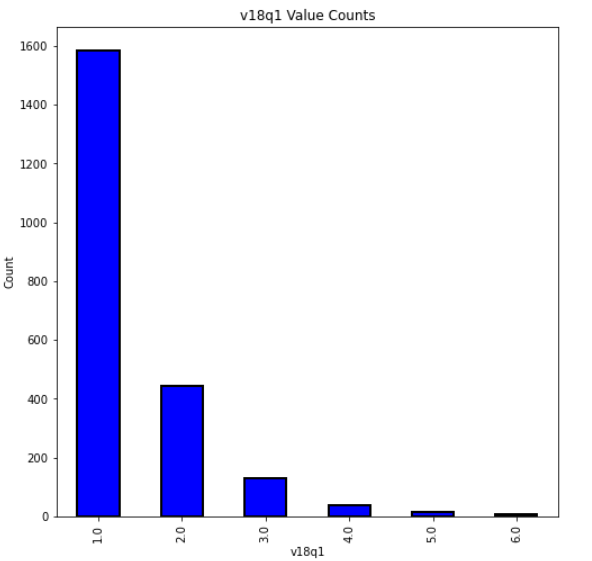
* For these three variables, it seems "yes"=1 and "no"=0. We can correct the variables using a mapping and convert to floats.



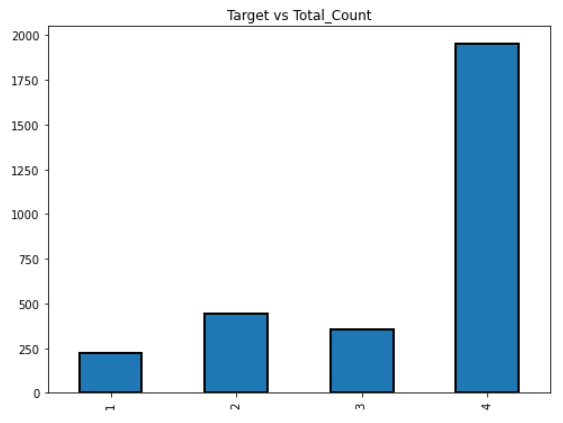
* Home Ownership Status For Household Missing Rent Payments:



* Value Counts:



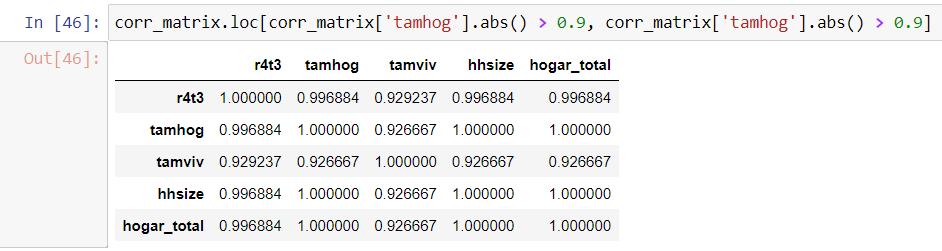
* Let’s look at the dataset and plot head of household and Target:-

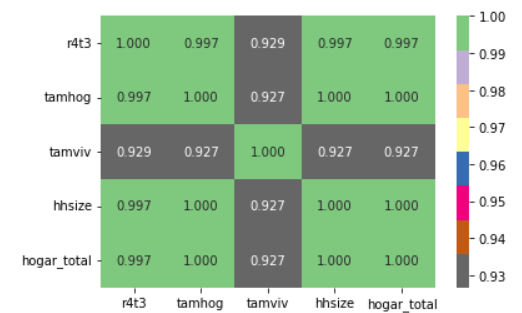


* Note :- extreme poverty is the smallest count in the train dataset. The dataset is biased.

Lets look at the Squared Variables

* ‘SQBescolari’
* ‘SQBage’
* ‘SQBhogar\_total’
* ‘SQBedjefe’
* ‘SQBhogar\_nin’
* ‘SQBovercrowding’
* ‘SQBdependency’
* ‘SQBmeaned’
* ‘agesq’
* Correlation Matrix:-





* Note:-

There are several variables here having to do with the size of the house:

* r4t3, Total persons in the household
* tamhog, size of the household
* tamviv, number of persons living in the household
* hhsize, household size
* hogar\_total, # of total individuals in the household
* These variables are all highly correlated with one another.
* Predict the accuracy using random forest classifier.

