Q4. Find potential issues with categorical variables. How to deal with missing values in categorical data.

Answer:

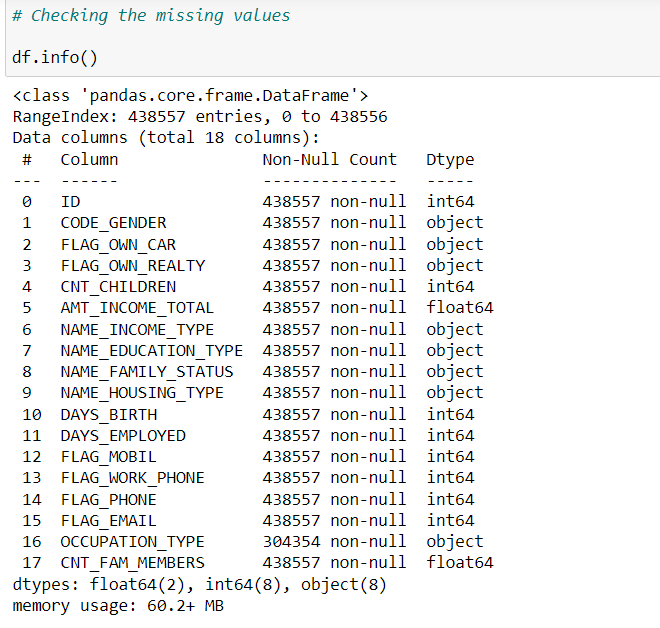
Definition:

Categorical variables represent types of data which may be divided into groups.

Examples of categorical variables are race, sex, age group, and educational level.

While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.

Below diagram shows information related to the application\_record dataset:



Following issues have been observed with the categorical variables present in the given sample data:

1. Categorical data is represented using text or numbers that fall under certain groups
2. Some categorical variables consist of integers which provide a discrete set of data which lacks continuity. Example a series can be represented as integers between 1-9 but the values lie between two integers cannot be shown.
3. Textual categories require to create dummy variables with binary or numeric values. E.g. Entries consist of colors Red, Green Blue etc.
4. It is difficult/expensive to fill the missing values for categorical variable to accurately reflects the data in trye sense
5. Almost 31% of categorical data is missing from “OCCUPATION\_TYPE” i.e.

|  |  |  |  |
| --- | --- | --- | --- |
| **Total Records** | **Data present** | **Data Missing** | **% missing** |
| 438557 | 304354 | 134203 | 31% |

Ways to handle missing categorical data.

Below are the ways in which missing categorical data can be populated:

1. The easiest way is to delete the records from where data is missing. This is recommended if the data is huge and missing values are very few.
2. Replace the missing values with most common value that appeared in the column. Again, this may have impact on the overall analysis and may not provide accurate results.
3. Create a new category e.g. “UNKNOWN” for missing data
4. Populate the missing data based on the dependent columns/functions within the dataset.

Example of steps implemented to replace the missing data:

