Processing missing values: In this data type replacement step, I started my handling the 'days_employed' column. I believe that this column is quite corrupted as there are many empty values or absurdly high values. To handle this column, I changed the type of the column from float to int, then I changed any negative value for this column to 0 (if you have been employed for a negative amount of time, then you have not been employed for that amount of time and therefore, your days employed are and then I went through each value to change the number from days to years for ease of readibility then stored this new years value in a new column, years_employed. Even after this cleanup, there are still very strange values - looking at the first 30 rows in this column yields either 0 years employed or between 932 to 1096 years of employment, which is not possible.

For the 'total_income' column, I changed the type of the column from float to int to make it easier to read the numbers. For the 'education' column, there were many of the same values but with different casing (for example, 'secondary education' and 'Secondary Education' are the same value but because of the capitilization difference, are stored differently). I used the str.lower() method to lower case all of the values which removes this unnecessary repetition.

The 'purpose' column was very problematic because there were many purposes that were essentially the same but because of the different wording, was stored differently. For example, 'to have a wedding' and 'wedding ceremony' are the same purpose but worded differently so will be handed as different reasons. We don't want this so I used lemmatization to find all unique identifiable words in the raw purpose output and from that, categorized the purpose accordingly.

Processing duplicates: In this processing duplicates step, I identified 408 duplicates in the data. I decided to drop all of these duplicates using the drop_duplicated() method.

Is there a relation between having kids and repaying a loan on time?

I started by grouping the data by the 'children' column. There are 13949 rows with 0 children, 4751 rows with 1 child, 2039 rows with 2 children, 329 rows with 3 children, 40 rows with 4 children and 9 rows with 5 children. Looking at the 'debt' column within this data grouping, there are 1070 rows with debt and 0 children, 444 rows with debt and 1 child, 194 rows with debt and 2 children, 27 rows with debt and 3 children, 4 rows with debt and 4 children and 0 rows with debt and 5 children. Putting these two numbers together: 13.0% of rows with 0 children have debt, 10.7% of rows with 1 child have debt, 10.5% of rows with 2 children have debt, 12.1% of rows with 3 children have debt, 10% of with 4 children have debt and 0% of rows with 5 children have debt. From this data, we can conclude that there is a relation between having kids and paying debt on time. The rows with more children are more likely to repay their loans on time, with the exception of the rows with 3 children.

Is there a relation between marital status and repaying a loan on time?

I started by grouping the data by the 'family_status' column. Looking at the data: 388/4124 = 14.9% of those who are widows have debt, 14.0% of those who are divorced have debt, 13% of those who are married have debt,10.6% of those in a civil partnership have debt, 10.1% of those who are unmarried have debt. From this data, we can conclude that there is a relation between martial status and repaying the debt on time. Those who either have been married (widows, divorced) or are married are more likely to have debt than those who are unmarried (either single or in a civil partnership).

Is there a relation between income level and repaying a loan on time?

I started by grouping the data by the 'income_type' column. Looking at the data: 376/5032 = 13.3% of those who own a business have debt, 86/1450 = 16.8% of those who are civil servants have debt, 1059/10899 = 10.3% of those who are employees have debt, neither of the two entrepreneurs in the data have debt, the 1 row on paternity/maternity leave have debt, 216/3730 = 17.2% of retirees have debt, the 1 student in the data does not have debt, and only 1 of the 2 unemployed in the data have debt. From this data, we can conclude that there is a relation income level and repaying the loan on time. Those who either were employed (retired) or are employed by an external entity (business, civil servent, employee) have high percentage of debt than those who are either self employed (entrepreneur), on paternity / maternity leave, a student or unemployed.

How do different loan purposes affect on-time repayment of the loan?

I started by grouping the data by the 'purpose' column.Looking at the data: 781/10576 = 13.5% of those with debt took it out for real estate purpoes, 186/2306 = 12.4% of those with debt took it out for wedding purposes, 370/3964 = 10.7% of those with debt took it out for education purposes and 402/4271 = 10.6% of those with debt took it out for car purchase purposes. From this data, we can conclude that most people take out loans for real estate purposes (50% of people who took out debt took it out for real estate purposes). As a result, the number of those who have debt and took out that debt for real estate purposes is going to be the highest. Those who took out loans for real estate or wedding purposes are less likely to repay their loans on time than those who took out loans for car or education purposes.

General conclusion

From this data, we can conclude that there is a relation income level and repaying the loan on time. Those who either were employed (retired) or are employed by an external entity (business, civil servent, employee) have high percentage of debt than those who are either self employed (entrepreneur), on paternity / maternity leave, a student or unemployed.
In conclusion:
- Those who have children are generally more likely to repay their loans on time, compared to those who do not have children.

- Those who have been married or are currently maried are generally more likely to have debt than those who have never been married or are not married.
- Most people take out loans for real estate purposes. Those who take out loans for real estate or wedding purposes are less likely to repay their loans on time.
- Those who either were employed (retired) or are currently employed by an external entity (business, civil servant,employee) are more likely to have debt than those who are either self-employed (entrepreneur), on paternity / maternity leave, a student or unemployed.