Description of the dataset:

That dataset what I will use, have information about current loans, completed loans, and defaulted loans. There are a lot of features which describe every loan. The description of these features you can find here.

Goal:

I want to do the preparing and cleaning of current data. And in the future, somebody can use it for building an ML model to predict if a loan will be paid off or not.

Implementation steps:

- In the beginning, I read our dataset from the second line because in the first line there is odd text and it will prevent to read the table correctly.
- I dropped two columns such as 'url' and 'desc'. 'url' I dropped because it contains the unnecessary link and 'desc' because there is a lot of explanation about every loan.
- I dropped columns which have a lot of missing cells. If there are, more then half of all data, missing values, I dropped this column.
- I dropped 'id','member_id','funded_amnt','funded_amnt_inv', 'int_rate','sub_grade','emp_title','issue_d','zip_code','out_prncp','out_prncp_inv', 'total_pymnt','total_pymnt_inv', 'total_rec_prncp','total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt'
- Now, let's take a look to FICO Score. FICO score shows how credit-worthy a person is. Now, if every borrower has low and high FICO score I think we should find an average between two this values
- Now when we have an average between low and high fico, so we can drop odd columns such as 'fico_range_low','fico_range_high','last_fico_range_low', 'last_fico_range_high'
- And now what I want to predict it's loan status. Loan status is in 'loan_status' column. Let's see which values there we have. To see this I used .value_counts(). I will choose only two of these values 'Fully Paid' and 'Charged Off' because only these values describe what I want to predict. So I should remove all other loans and transform our two main values to 1 and 0 by replace() method
- If we look at the dataFrame, we will see that there are columns where is only one value. We don't need such columns.
- Now I should deal with missing values. I use dmethod isnull() and sum(). isnull() to know if column has missing values and sum() to count them

- I used .dropna() to drop all rows with missing values and .drop() to drop wh ole column
- Like we see from method info there are a lot of column with object type. Th is means that we should transform all these strings to ints
- Column 'revol_unit' has numeric value but it has object type. So what we should do. At first, we should use str.rstrip() to strip the sign % and after that we will convert it to float type
- I dropped the columns which have to much unique values such as 'last_cred it_pull_d','addr_state','title','earliest_cr_line'
- I transform all columns with objects to numerical values by replace() metho d
- use pandas' get_dummies() method to return a new DataFrame containing a new column for each dummy variable
- use concat() method to add these dummy columns back to the Original Dat aFrame

Resalts:

Les's use method info to see if there are all non-null numerical values

```
loan_amnt                                38123 non-null float64
installment                              38123 non-null float64
grade                                    38123 non-null int64
emp_length                               38123 non-null int64
annual_inc                               38123 non-null float64
loan_status                              38123 non-null int64
dti                                      38123 non-null float64
delinq_2yrs                              38123 non-null float64
inq_last_6mths                           38123 non-null float64
open_acc                                 38123 non-null float64
pub_rec                                  38123 non-null float64
revol_bal                                38123 non-null float64
revol_util                               38123 non-null float64
total_acc                                38123 non-null float64
fico_average                             38123 non-null float64
home_ownership_MORTGAGE                  38123 non-null uint8
home_ownership_NONE                      38123 non-null uint8
home_ownership_OTHER                     38123 non-null uint8
home_ownership_OWN                       38123 non-null uint8
home_ownership_RENT                      38123 non-null uint8
verification_status_Not Verified         38123 non-null uint8
verification_status_Source Verified      38123 non-null uint8
verification_status_Verified             38123 non-null uint8
purpose_car                              38123 non-null uint8
purpose_credit_card                      38123 non-null uint8
purpose_debt_consolidation               38123 non-null uint8
purpose_educational                      38123 non-null uint8
purpose_home_improvement                 38123 non-null uint8
purpose_house                            38123 non-null uint8
purpose_major_purchase                   38123 non-null uint8
purpose_medical                          38123 non-null uint8
purpose_moving                           38123 non-null uint8
```

```
purpose_other                       38123 non-null uint8
purpose_renewable_energy            38123 non-null uint8
purpose_small_business              38123 non-null uint8
purpose_vacation                    38123 non-null uint8
purpose_wedding                     38123 non-null uint8
term_ 36 months                     38123 non-null uint8
term_ 60 months                     38123 non-null uint8
```

Now we can use this dataset for creating ML model to predict if a loan will be paid off or not.