Steps to Data Wrangling

Introduction:

The two most critical questions in the banking industry are:

1) How risky is the borrower?
2) Given the borrower's risk, should we lend him/her?

The answer to the first question determines the credit score the borrower would have.Credit score measures among other things the riskiness of the borrower, i.e. the riskier the borrower, the lower the credit score. With credit score in mind,credit score varies how timely the borrower repays the monthly payment. we can then determine if the borrower is eligible for the loan. Investors/Bank (lenders) provide loans to borrowers in exchange for the promise of repayment with interest. That means the lender only makes profit (interest) if the borrower pays off the loan. However, if he/she doesn't repay the loan, then the lender loses money.

1) What kind of cleaning steps did you perform?

Features as "Years in current job","Home Ownership", "Number of Credit Problems" and "Years of Credit History" required cleaning as they may impact the dependent variable loan status.

as say Feature(categorical) as "Years in current job" has values like "10 years,2 years and 3 years" and so on so its been converted to tidy as "10 ,2 ,3 ,1 and 0".

Dataset:
https://www.kaggle.com/zaurbegiev/my-dataset

2) How did you deal with missing values, if any? Why missing value treatment is required ? Why data has missing values? Which are the methods to treat missing value ?

Almost always real world data sets have missing values. This can be due, for example, users didn't fill some part of the forms or some transformations happened while collecting and cleaning the data.Missing data in the training data set can reduce to fit of a model or can lead to a biased model.

a) delete all records that have any missing value if records are less than 1 percentage of total volume or if missing values are more than 50 % of total volume then drop the column from the dataset.

b) Mean/ Mode/ Median Imputation:

3) Were there outliers, and how did you handle them?

I checked the data for customers loan records and found out the mean annual income and current loan amount.Noticed that mean of both the fields is extremely high.
On further analysis I came to know that dataset has presence of customers who has loan amount and annual income above 9 crore and 16 crores and which is extremely high and act as outlier.On checking the loan amount records having value 9 crores i found there around 11 thousand records.Deleting these numbers of records will lose significant information what we are looking for.In order to deal with this type of situation i break records into new dataframe having values as low income , medium income , high income and very high income and finally use pandas get dummies to overcome it.