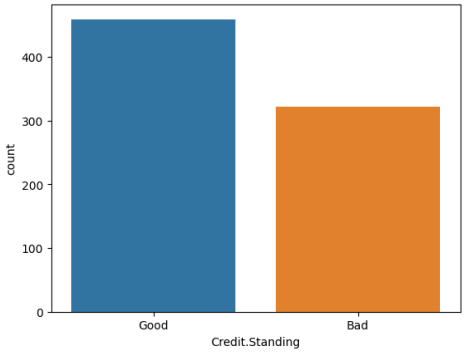
Group Name: Tech Geeks
Name: Dhanashri Jagadale
Email: dhanashri.jagadale1998@gmail.com
College: Munster Technological University
Specialization: Data Science

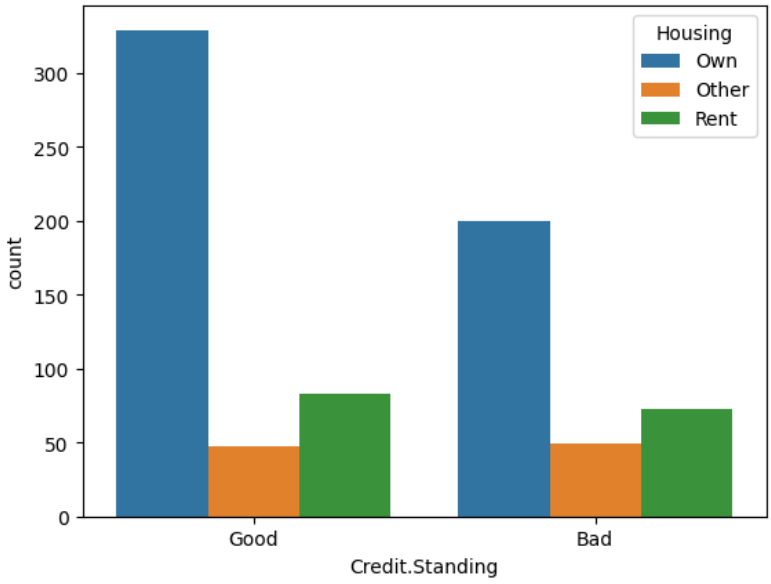
Problem Description: The problem is to develop a predictive model that can assess the credit worthiness of potential future customers of a financial institution. The available data set consists of 807 past loan customer cases, each with 14 attributes including financial standing, reason for the loan, employment, demographic information, foreign national status, years of residence in the district, and the outcome/label variable Credit Standing, which classifies each case as either a good loan or bad loan. The objective is to build a model that accurately predicts the credit standing of new loan applications, using the available data as the training set. The model should be able to identify the key factors that determine creditworthiness and provide insights to help the financial institution make better lending decisions.

EDA Performed on the data:

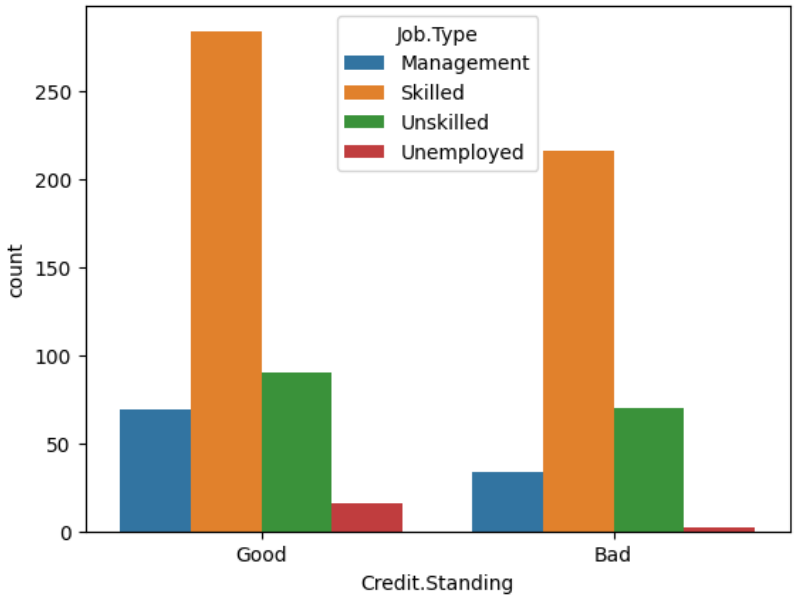
1. Checked for the first five rows with the .head() function.
2. Checked the number of rows and columns with .shape.
3. Gained information about data types with .info() function.
4. Used .describe() function to gain statistical summary of the data.
5. Checked for missing values and dropped them.
6. Plotted histogram for numerical variables to check the distribution of data and found that there is a skewness in the data which I handled through log transformation.
7. Plotted boxplots to check the presence of outliers. There were very few outliers present in the data so dropped them to avoid errors in the result.
8. Used sns.countplot function to plot the count of good and bad credit standings and found that the dataset has more good credit standing records.



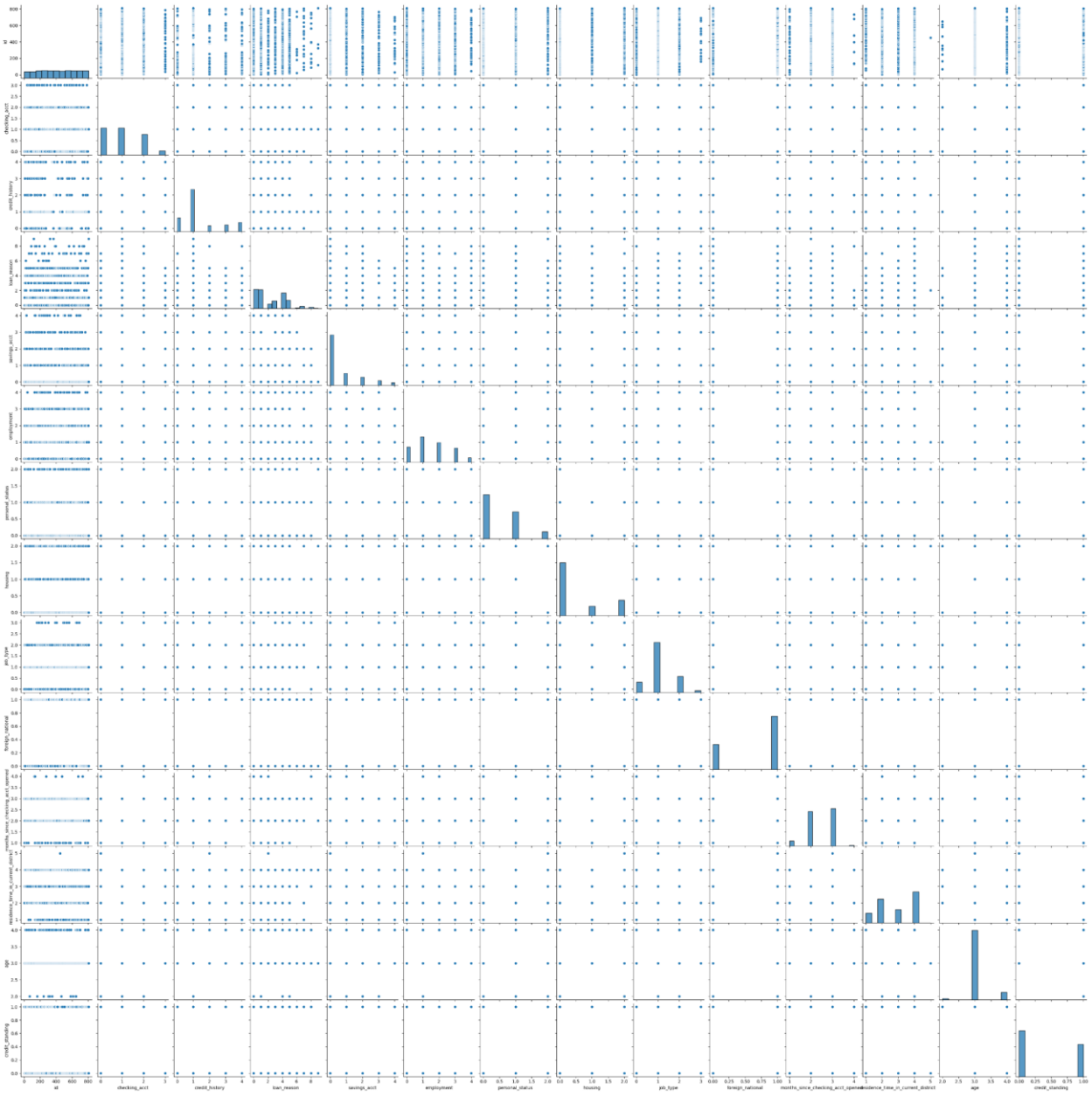
9. Next plotted the credit standing by using housing type as the differentiating category. The result shows that people with their own house are more likely to have good credit standing.



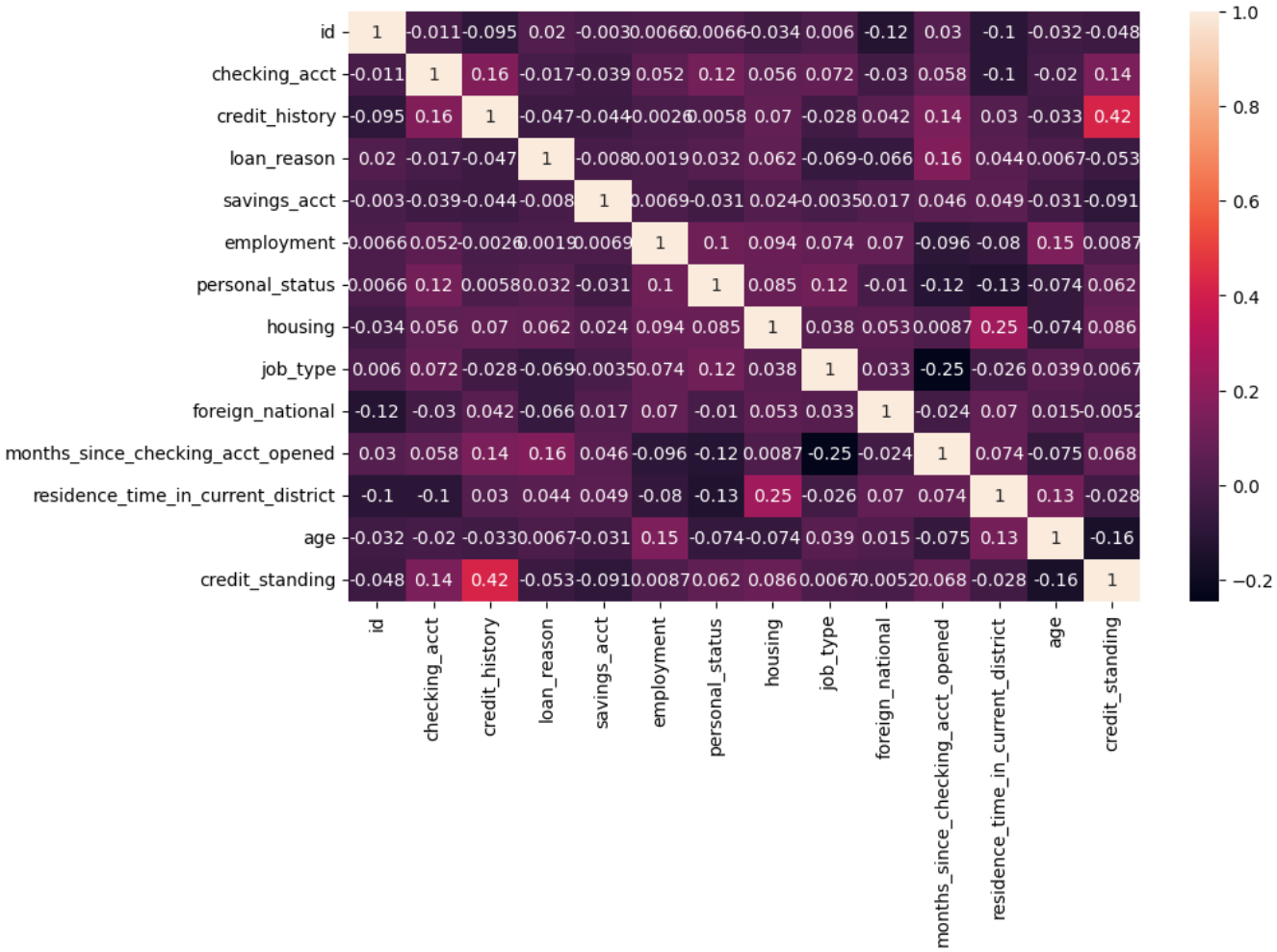
10. Next, plotted the credit standing count with the job type as a differentiating variable. Skilled people are more likely to have good credit standing. To my surprise even if the people are unemployed they are most likely to have good credit standing.



11. Next plotted the pairwise relationship among the numerical variables using sns.pairplot function and there is mostly no linear relationship between any of the two variables



12. Plotted heatmap to verify the correlation and it is evident that there is no strong negative or positive correlation between these variables and hence no linear relationship.



It seems there is a non relationship between feature variables from the Exploratory data analysis. Decision tree or Random forest machine learning algorithms can be assumed to give best results in such scenarios. Overall, the best machine learning model for this dataset will depend on the specific data and It is important to carefully evaluate and compare the performance of different algorithms before selecting the best one for this dataset. So I will be building different models to compare the performance and select the best model at the end.