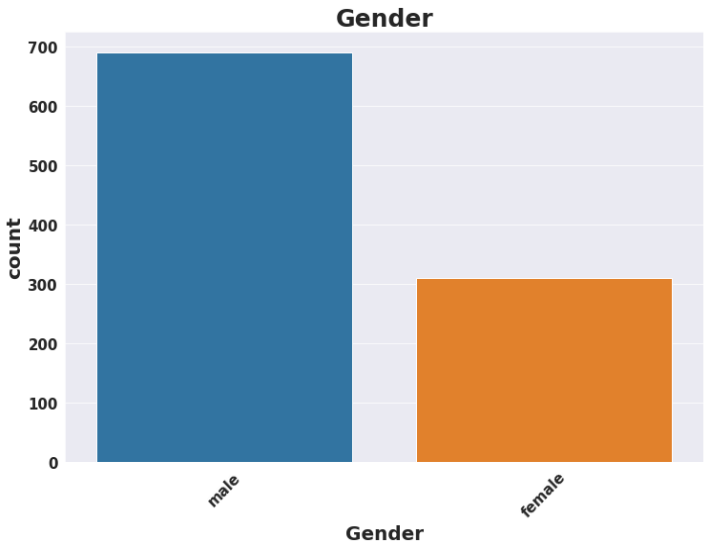
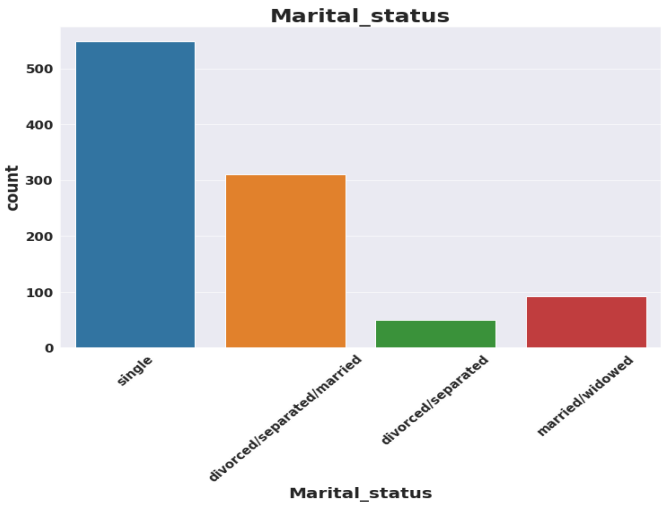
Reunion Data Science Assignment

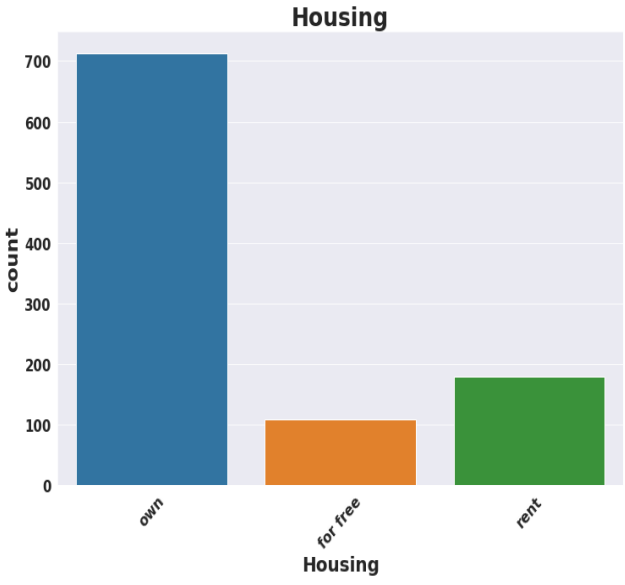
TASK – 1

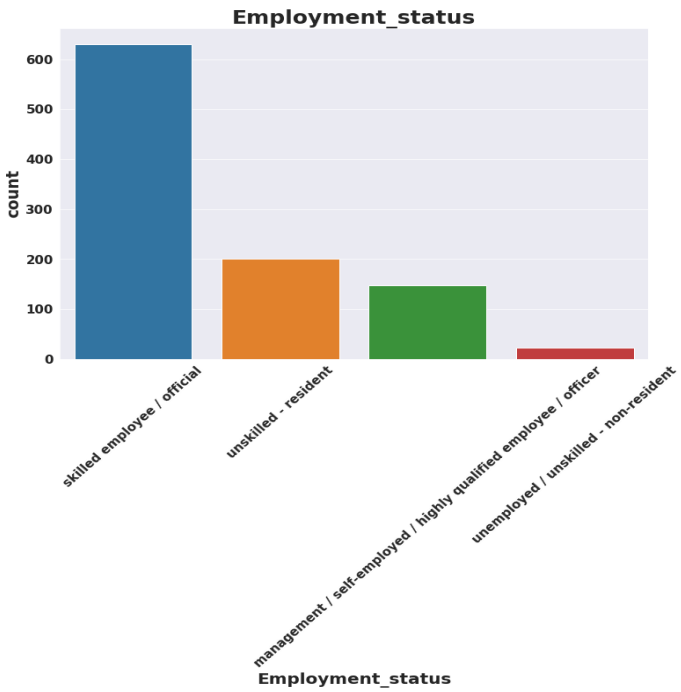
1. Do the Exploratory Data Analysis & share the insights.

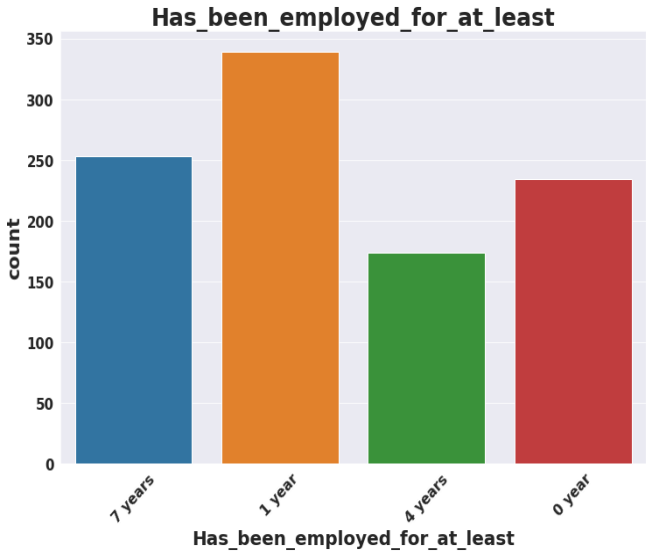
    a. Count of categories in categorical feature

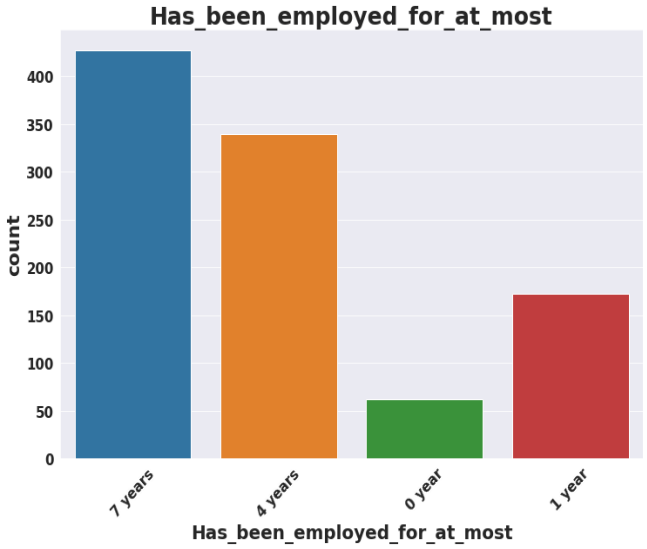


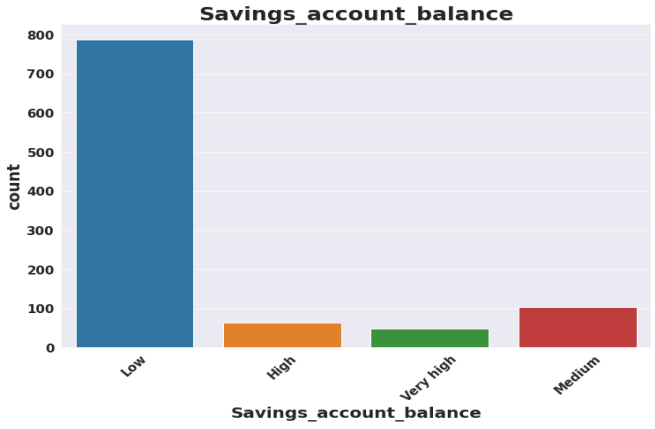


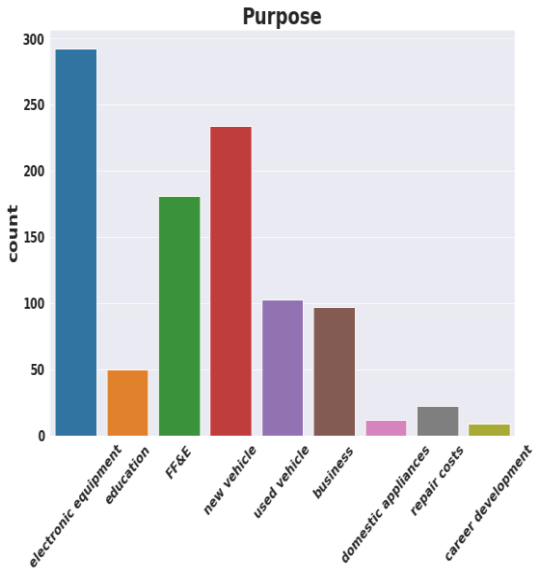


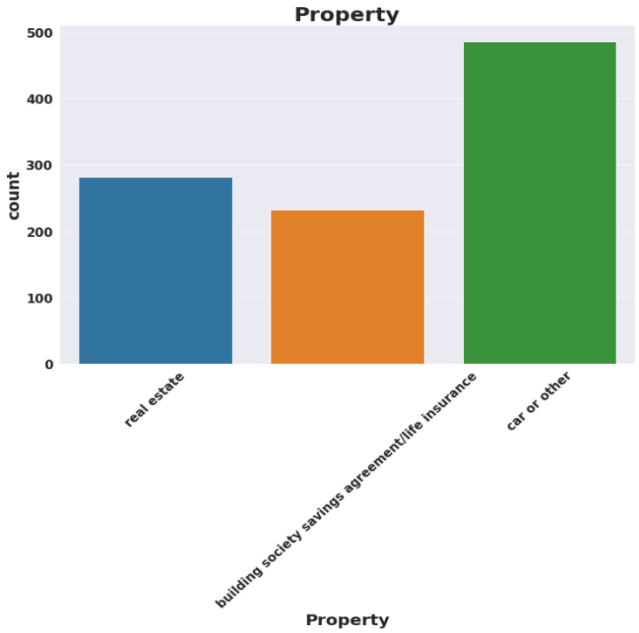


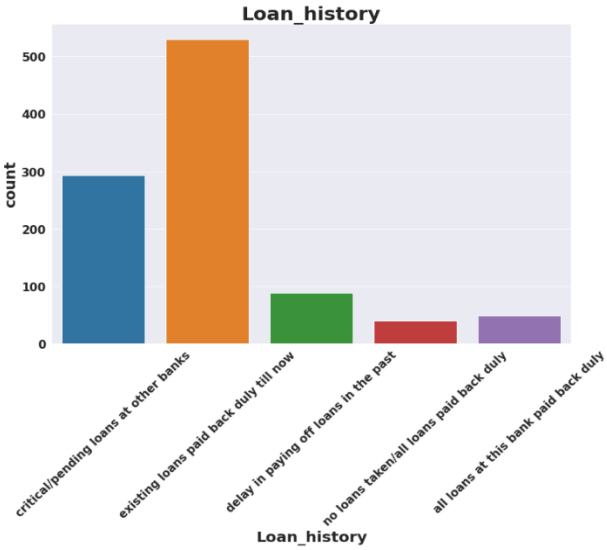

Has_been_employed_for_at_least
count
0
50
100
150
200
250
300
350
7 years
1 year
4 years
0 year
Has_been_employed_for_at_least


Has_been_employed_for_at_most
count
0
50
100
150
200
250
300
350
400
7 years
4 years
0 year
1 year
Has_been_employed_for_at_most


Savings_account_balance
count
0
100
200
300
400
500
600
700
800
Low
High
Very high
Medium
Savings_account_balance


Purpose
count
0
50
100
150
200
250
300
electronic equipment
education
FF&E
new vehicle
used vehicle
business
domestic appliances
repair costs
career development
Purpose

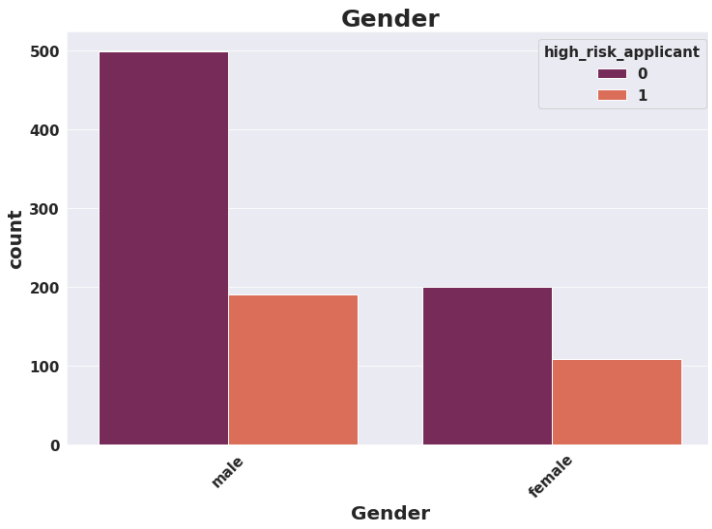

Property
count
0
100
200
300
400
500
real estate
building society savings agreement/life insurance
car or other
Property

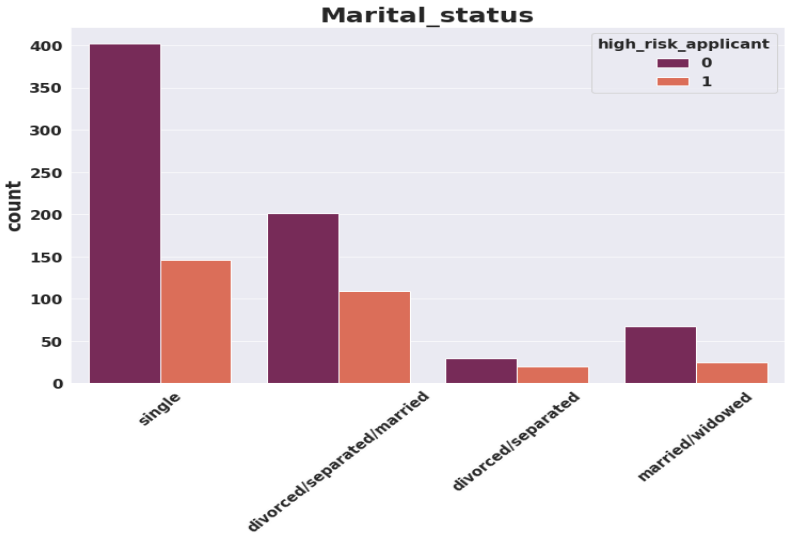

Loan_history
count
0
100
200
300
400
500
critical/pending loans at other banks
existing loans paid back duly till now
delay in paying off loans in the past
no loans taken/all loans paid back duly
all loans at this bank paid back duly
Loan_history

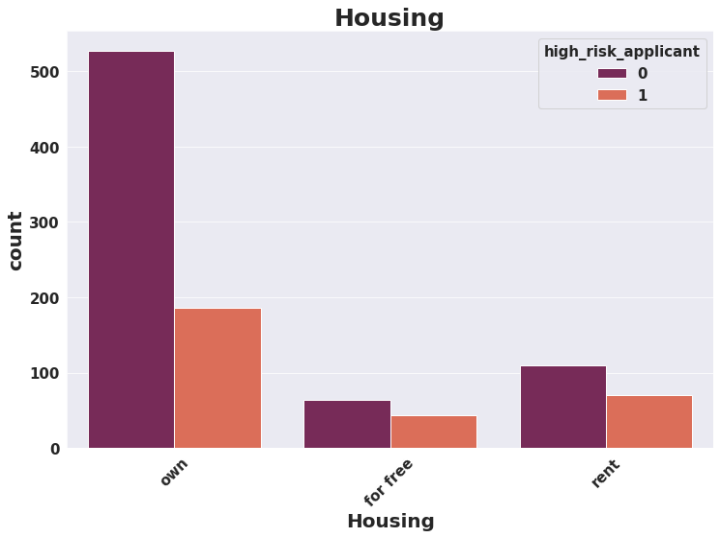
Inference

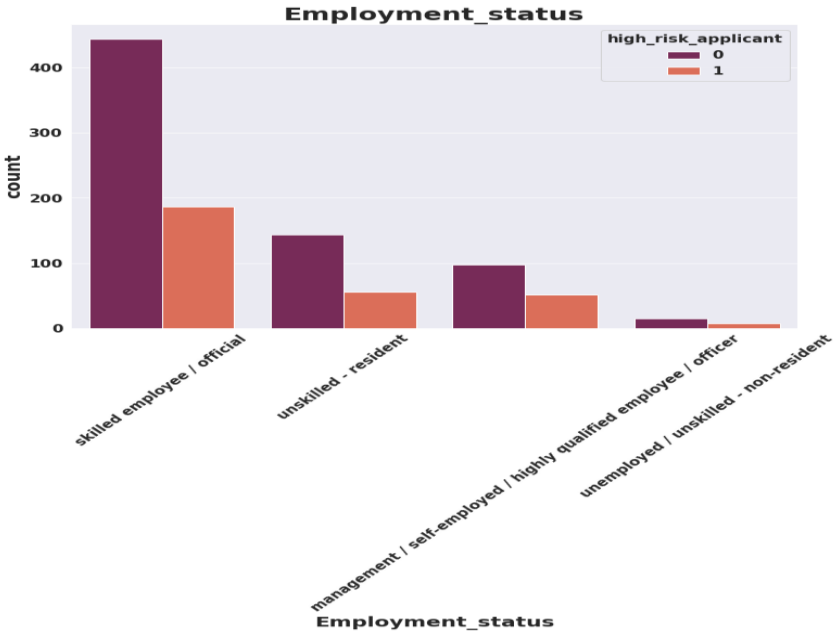
- Male gender have more data in our datset.
- More single people in our dataset than married/divorces/seperated and have least for divorces/seperated.
- People own house has more in our dataset than who live in rented house.
- We have skilled employee/ Officer maximun in our dataset and very low for unemployed/ unskilled & non resident.
- We have maximum number of people who is employed for at least 1 year and maximum number of people who is employed for maximum for 7 years.
- Maximum number of people have saving balance account at low category and very low amount of people have either medium, high or very high saving balance account.
- Purpose of taking a loan in most of the cases is for elctronic equipment, new vehicle or FF&E (Furnitures fixtures and equipment).
- Maximum number of people who owns some kind of property are for car or other than real estate.
- In more than 50% of the cases are for existing loans paid back duly till now and less than 5% of the cases are for no loan taken/ all loans paid back duly.

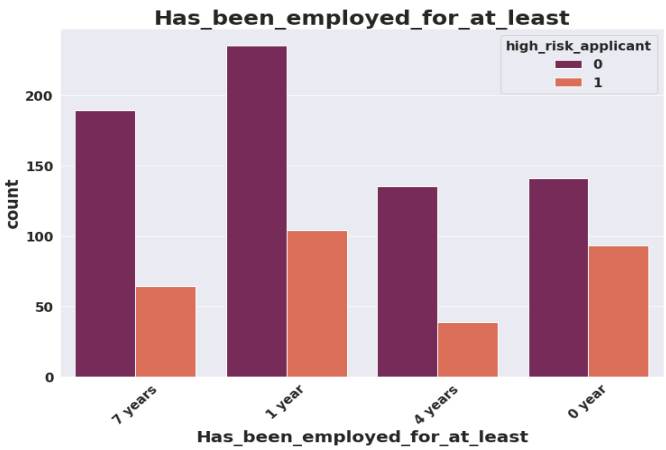
b. Relationship between categorical features and label

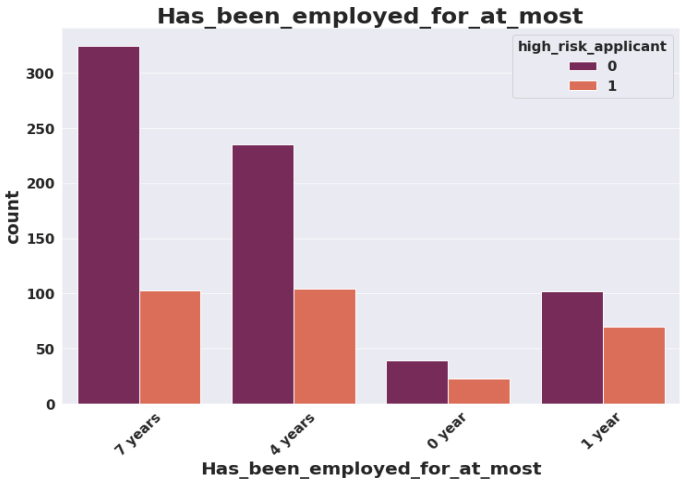


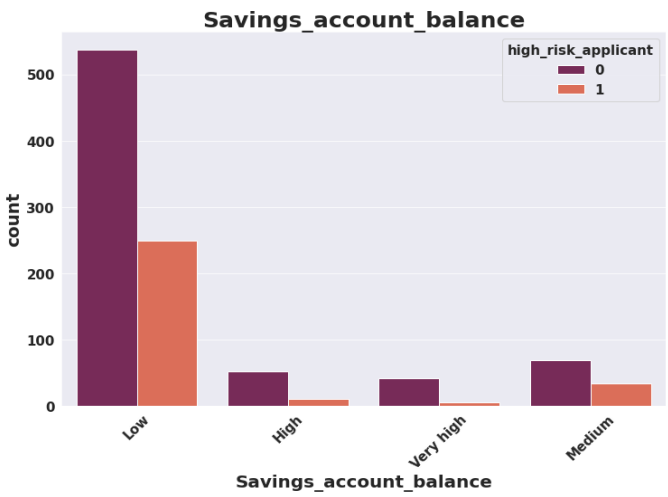


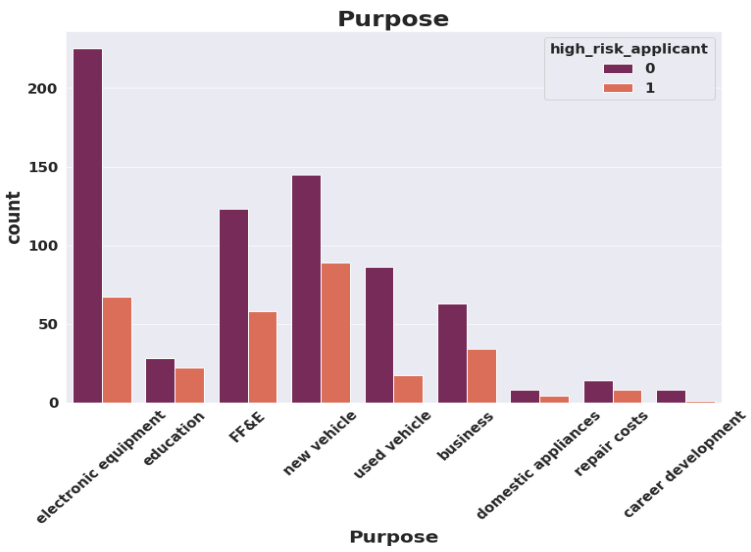


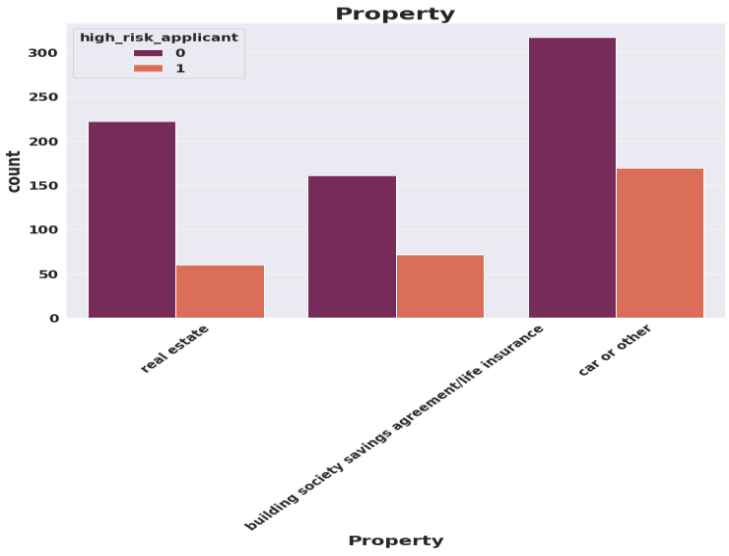


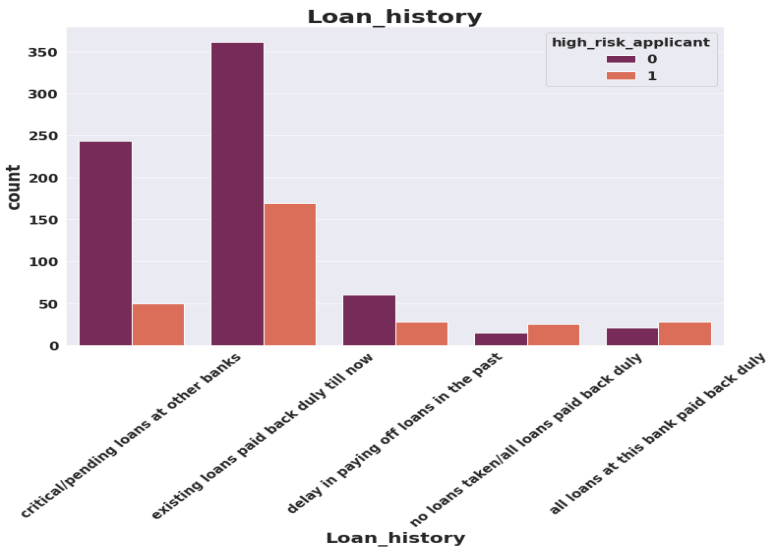








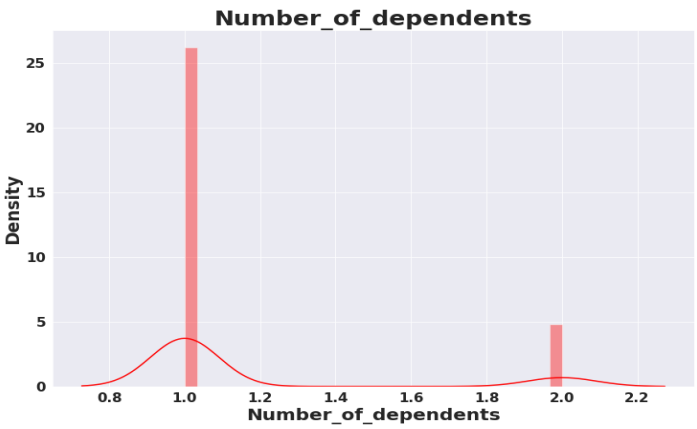


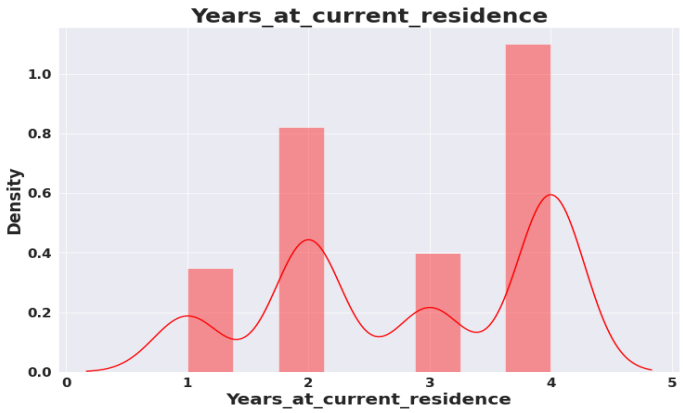


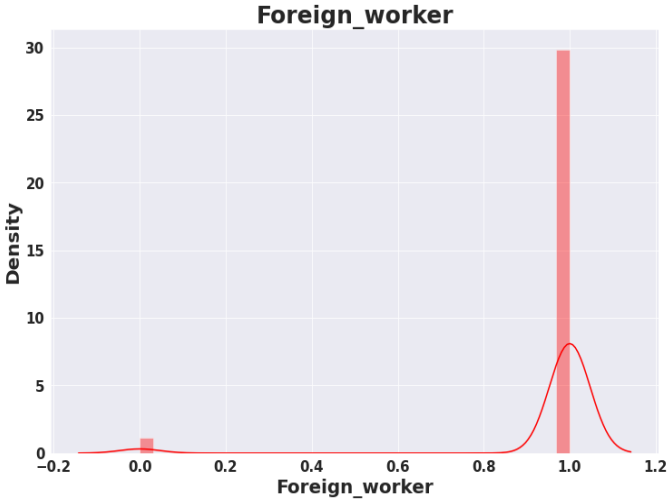
Inference

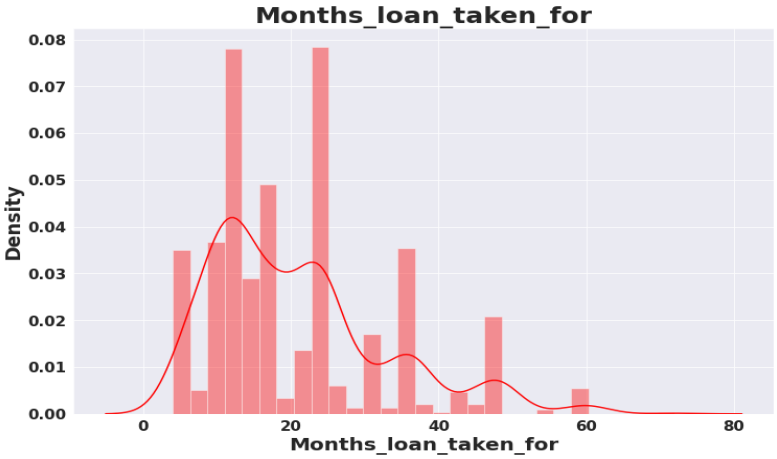
- High risk applicant is comparatively higher for female than male.
- Separated/divorced people have comparatively higher risk applicant.
- People who have housing status free or for rent have comparatively higher risk applicant.
- People who are employed for 0 or 1 year are comparatively high risk applicants.
- People who have saving balance account low or medium are comparatively higher risk applicant.
- Purpose of taking a loan for education or vehicle are comparatively higher risk applicant.
- People who own car or other property are at little higher risk.

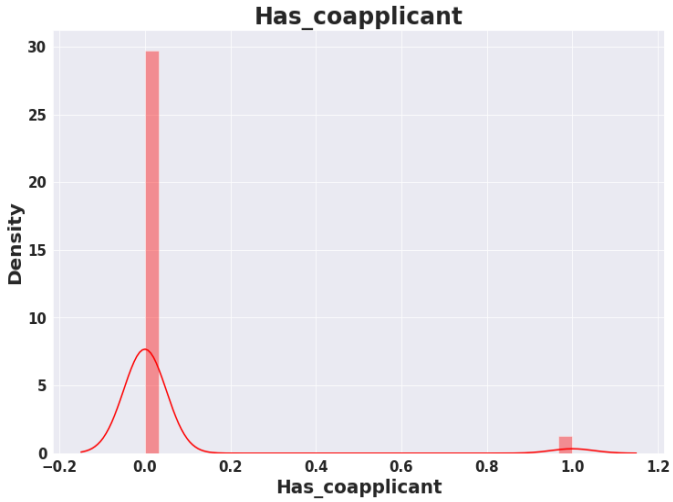
c. Distribution of numerical feature

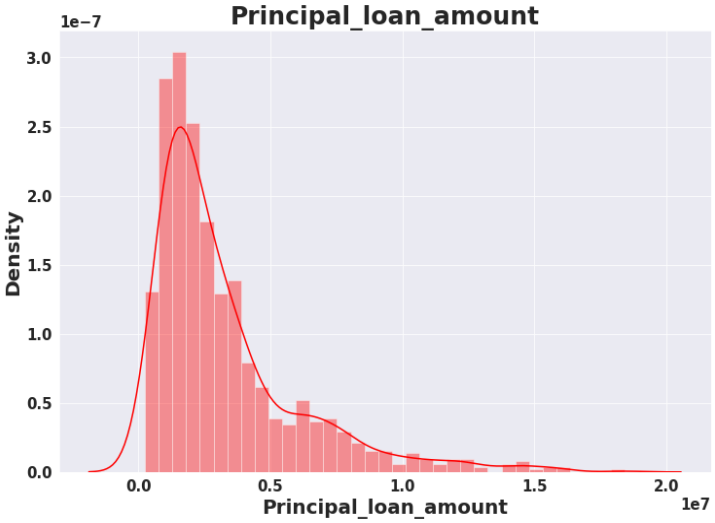


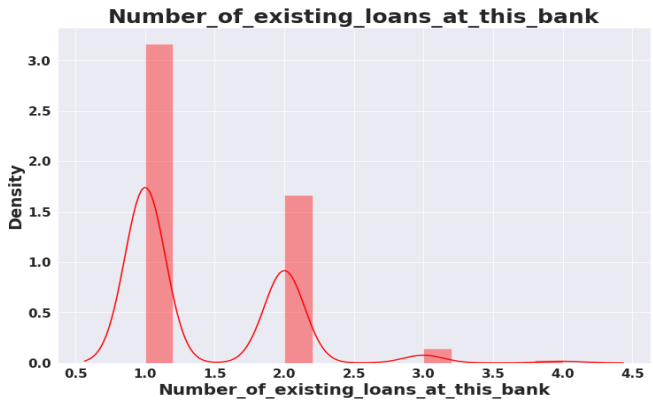


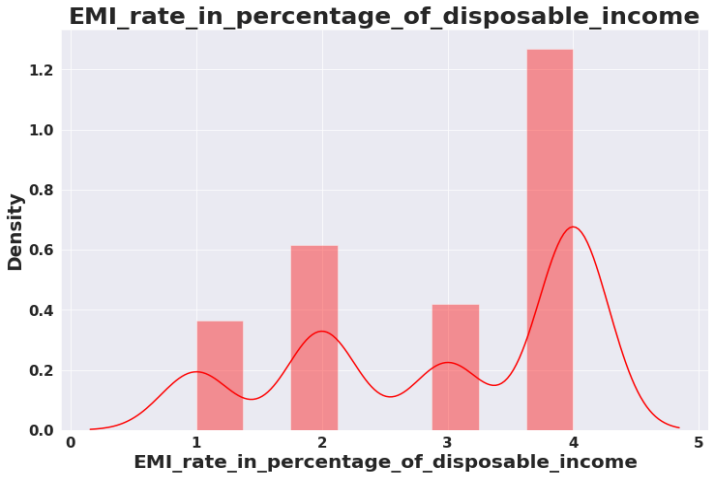


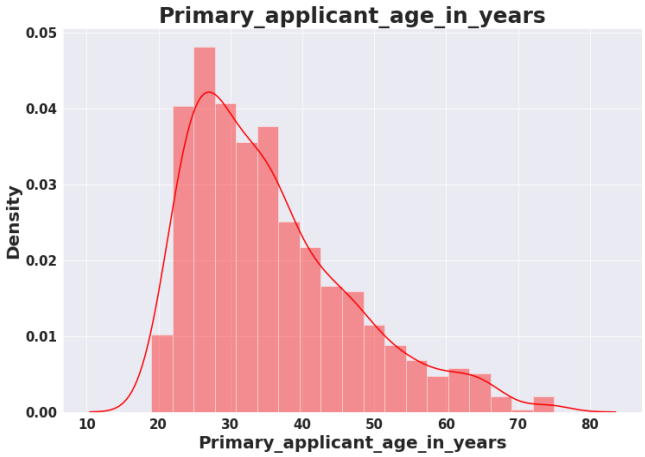








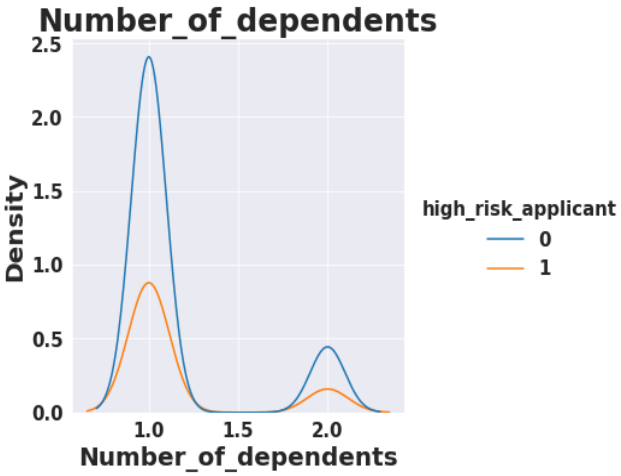


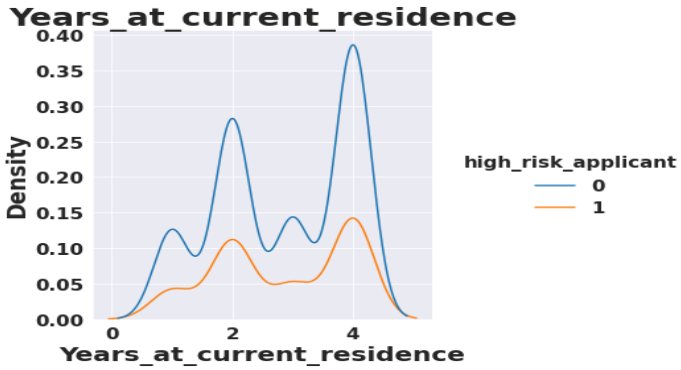


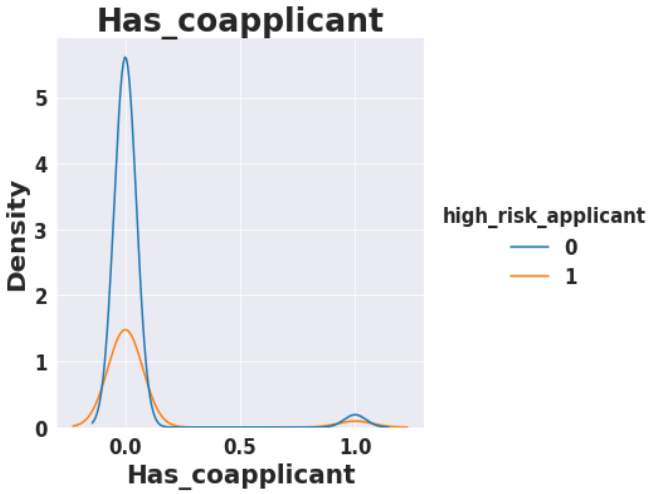
Inference

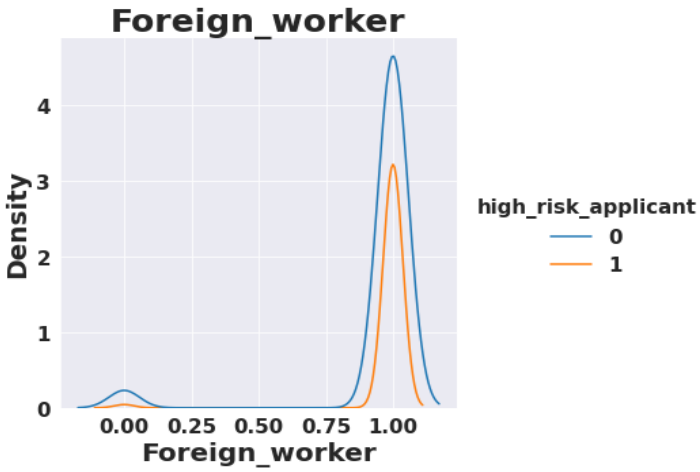
- Age column is right skewed which is also obvious senior people don't take much loan.
- Months loan taken for is for 10 to 25 months in most of the cases.
- Number_of_dependents, years_at_current_residence, foreign worker, EMI_rate_percenytage_for_disposable_income, has_guarantor, has_coapplicant, Number_of_exisiting_loan_at_this_bank is actually has some discrete values, which can be treated as an categorical variable for model training and prediction.

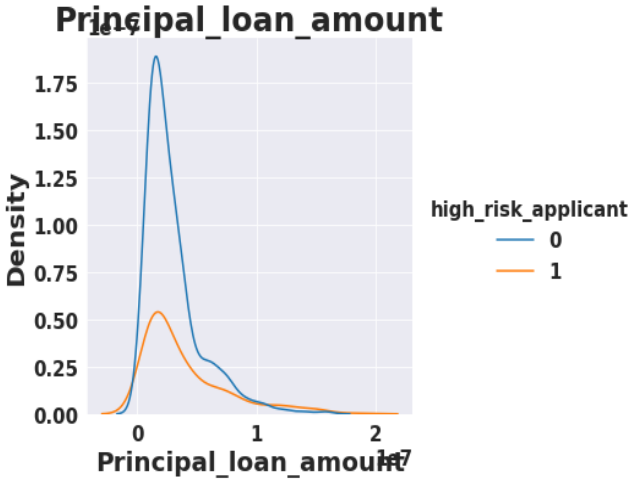
d. Trend of numerical features with target variable

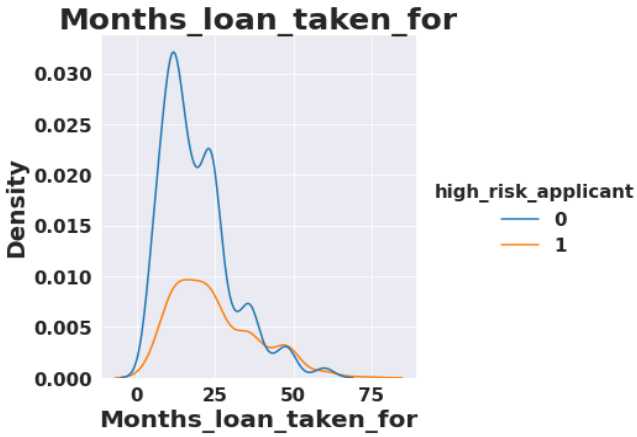


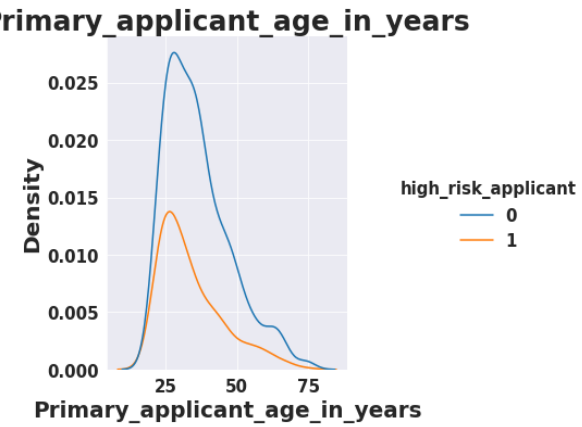


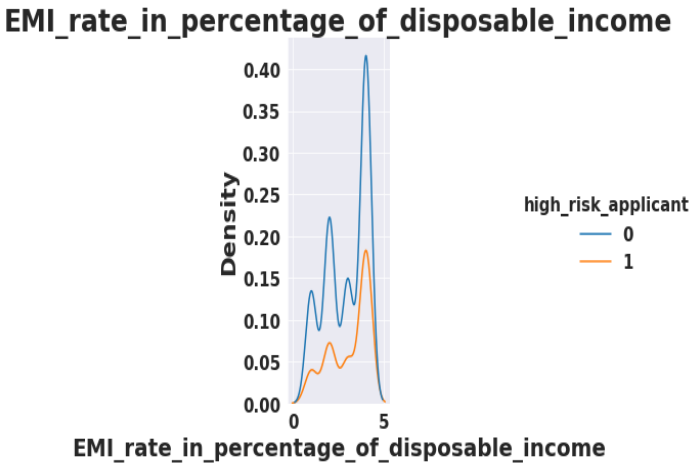


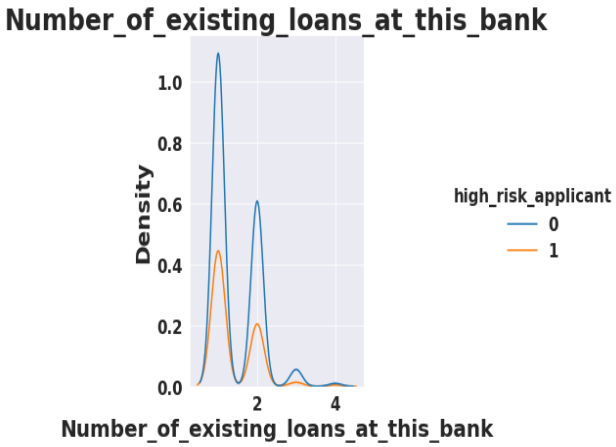








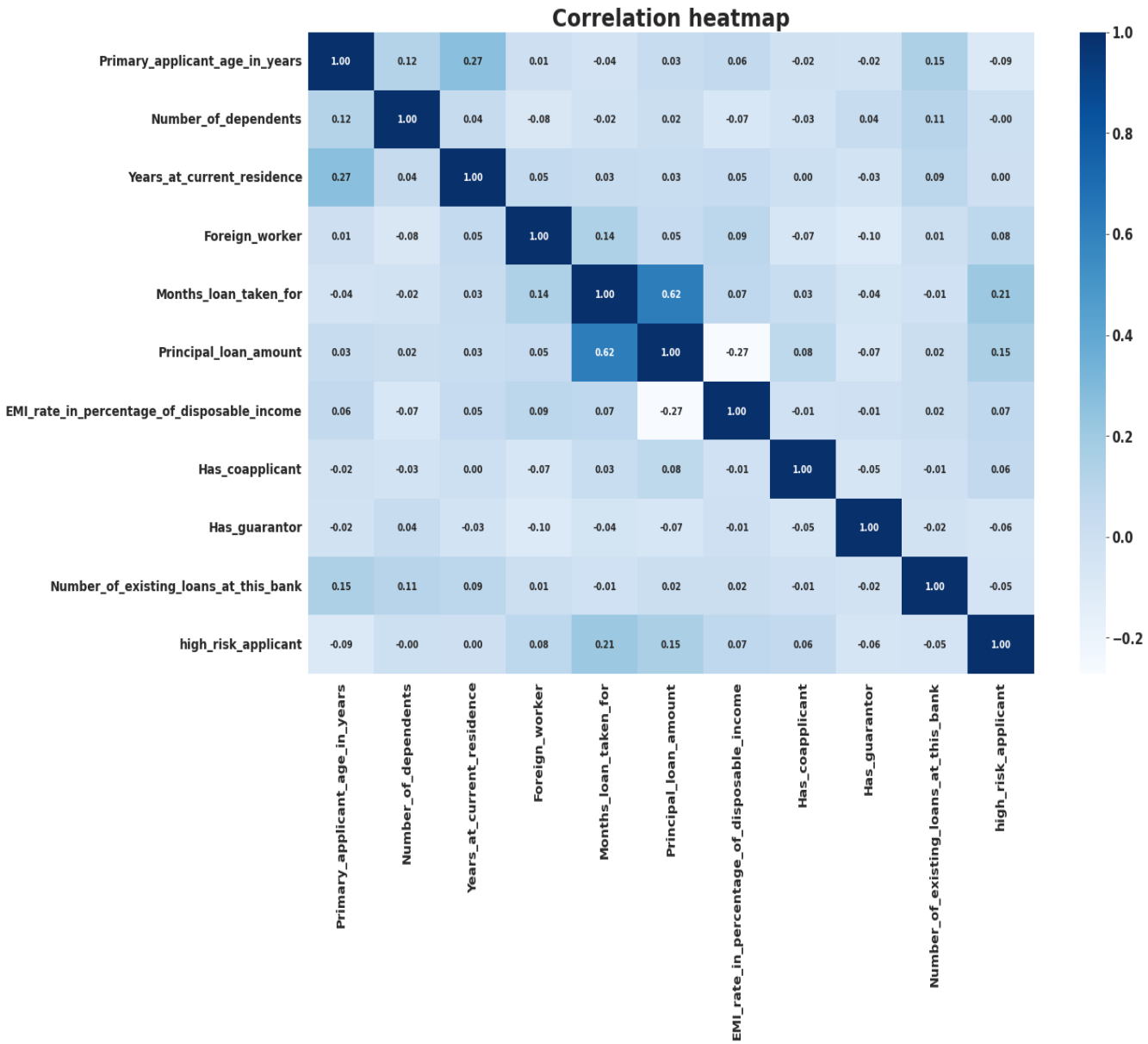




Inference

- People who have age between 25-30 are high risk applicant as they have maximum distribution in dataset compared to other range.
- Foreign workers are high risk applicant. people who have taken loan for less than 25 months are at little bit higher risk than others.
- People who have high EMI rate in percentage are at high risk.
- People who have no guarantor are high risk applicant.

e. Correlation



2. How would you segment customers based on their risk (of default)?

Segmenting customer

We can segment customer based on Houses owned, Employee experience, saving account balance, foreign worker and Guarantor.

I would propose the two segment to approve are-:

a. A person with more credit account with more account balance are more creditworthy.
b. Person who owns his house are more creditworthy.

3. Tell us what your observations were on the data itself (completeness, skews).

- The dataset is about knowing a person will they do default on loan taken. We had null values in few columns we handled them and started our data exploratory analysis.
- People are aged between 19-75 years.
- Age column is right skewed which is also obvious senior people don't take much loan.
- Females are comparatively at higher risk applicant, a well separated and divorced people.
- People who are between age 25-30 are high risk applicant and people who are taking loan for education or vehicle purpose are high risk applicant.
- Foreign workers are most likely to make default also who has no guarantor.
- People who owned their own house are comparatively low risk applicant.

TASK – 2

1. Explain your intuition behind the features used for modeling.

   I had first bin the age, number of month loan taken column into some part categories. Then I had converted categorical features into numerical feature using get dummies technique.

   Then I had used some scaling techniques to transform all the features on similar scale.

2. Are you creating new derived features? If yes explain the intuition behind them.

   No.

3. Are there missing values? If yes how you plan to handle it.

   Yes, there were missing values I handled it by using Pandas methods help us achieve this in a single line of code for every column. The fillna()method fills the empty fields with whatever parameter is given. Calculating the mean or mode of this array of values, and passing it to fillna() completes this step. Using mode works best in our case, as most columns are binary. Moreover, Mode will simply put the most occurring instance in place of empty fields, which, under the circumstances, would be the best guess.

4. How categorical features are handled for modeling.

   Using use the get_dummies function of Pandas for non-binary values like 'Property' to automatically one-hot encode them

5. Describe the features correlation using correlation matrix. Tell us about few correlated feature & share your understanding on why they are correlated.

   A correlation matrix is a table showing the correlation coefficients between different sets of variables

I've used seaborn heatmap for correlation in which the dark blue color denotes the stronger correlation and in our data-frame Principal_loan_amount and months_loan_taken for a strong corelation because if the we increase loan amount the duration of loan also increases and Emi rate and loan amount have inverse correlation.

6. Do you plan to drop the correlated feature? If yes then how.

   Yes, I plan to showcase this feature using Seaborn heatmap because cool warm colors really make it easy to identify the correlation in data frame.

7. Which ML algorithm you plan to use for modeling and there different parameters?

   I have used Linear Regression model because as we all know Linear Regression deals with continuous values and we have continuous data, and observations are independent of each other.

   Hyperparameters -: GridSearchCV on Linear Regression

```
      #hyperparameter tuning on Logistic regression
LR = LogisticRegression()
LR_Grid = GridSearchCV(LR,param_grid=parameter,cv=20,verbose=1)
LR_Grid.fit(x_trainr,y_trainr)
```

   I used a decision tree because it's a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). It is useful in cases where data set is very scattered in both X and Y axis in graph and in which multiple attributes are responsible for prediction.

   Hyperparameters -: GridSearchCV on Decision Tree

```
#selecting parameter for hyperparameter tuning
parameter = {"max_depth": [1,2,3,4,5],
      "min_samples_leaf": [1,2,3,4,5,6,7,8,9,10],
      "criterion": [ "gini","entropy"],
     "class_weight":["balanced"],
    "ccp_alpha":[0.01,0.1]}
```

I have used Random Forest classifier because Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. Decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Hyperparameters -: GridSearchCV on Random Forest Classifier

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(), n_jobs=-1,
     param_grid={'class_weight': ['balanced', 'balanced_subsample'],
                 'criterion': ['gini', 'entropy'],
                 'max_depth': [2, 6, 8, None],
                 'max_samples': [0.5, 0.75, 1.0],
                 'n_estimators': [20, 60, 100, 120]},
     verbose=1)
```

8. Train two (at least) ML models to predict the credit risk & provide the confusion matrix for each model.

   Done.

9. Which metric(s) you will choose to select between the set of models.

   I have evaluated the models using confusion matrix and classification report.

10. Explain how you will export the trained models & deploy it for prediction in production.

    import pickle #library for saving the model for future use or model deployment

    pickle.dump(LR_Grid,open('logistic_model.pkl','wb'))