Risk Management VAC 1 Project Presentation

Group 4-

Himanshuraj Kashyap (202218006)

K Sethu Srivatsa (202218021)

Nidhi Somaiya (202218060)

Radhika Singhal (202218062)

Swayista Ahmed (202218035)

TABLE OF CONTENT

Brief Overview

- Aim to train a model that can predict whether a loan should be given or not to a customer.
- Our project aims to leverage collaborative filtering techniques to provide tailored loan recommendations to users.

- **Context of the data set-** The data contains 1000 entries with 10 categorical/symbolic attributes. In this dataset each entry represents a person who takes a credit by a bank. Each person is classified as a good or bad credit risk according to the set attributes.

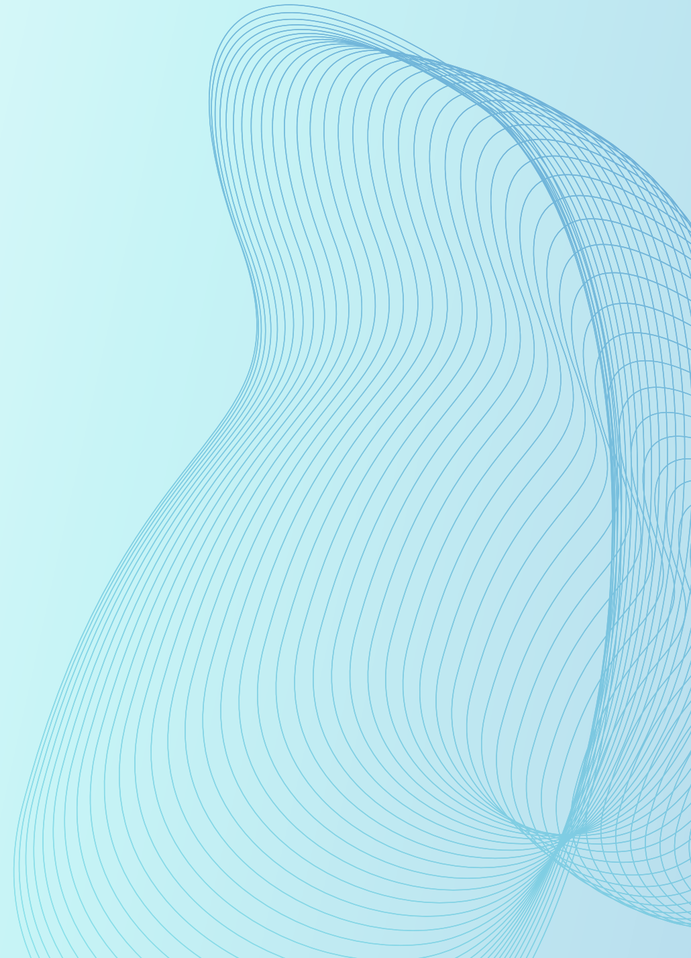
Attribute Information

Features	Type	**Description**	**Missing values**
Age	**Numeric**	Min age-19, Max age-75	No
Sex	Categorical	Male and Female	NO
Housing	Categorical	own ,rent, free	NO
Checking Account	Categorical	little, moderate, quite rich, rich	394
Credit Amount	**Numeric**	min-250 max-18,424	NO
Job	Categorical	0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled	NO
Duration	**Numeric**	given preferred months to repay the loan	NO
Purpose	Categorical	car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others	NO
Risk	Categorical	good, bad	NO
Savings account	Categorical	little, moderate, quite rich, rich	183

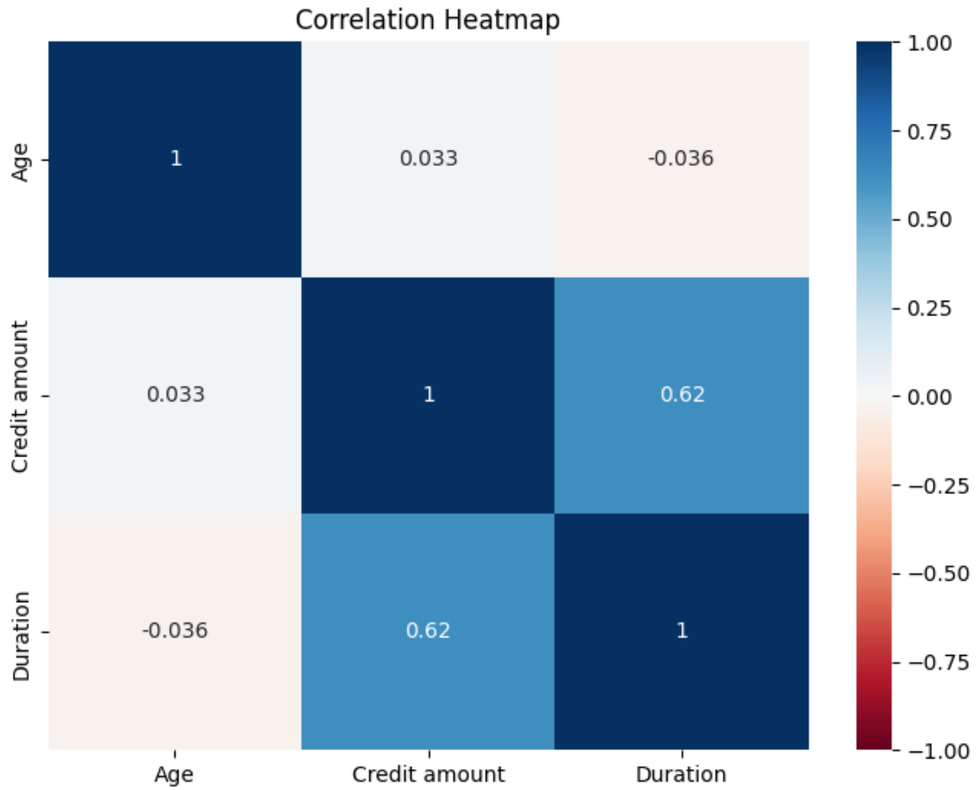
DATA VISUALIZATION AND ANALYSIS

Data Categorization

1. **Personal Information**
   - Sex
   - Gender
   - job
2. **Wealth Information**
   - Savings Account
   - Checking Account
   - Housing
3. **Loan Information**
   - Credit Amount
   - Duration
   - Purpose
   - Risk



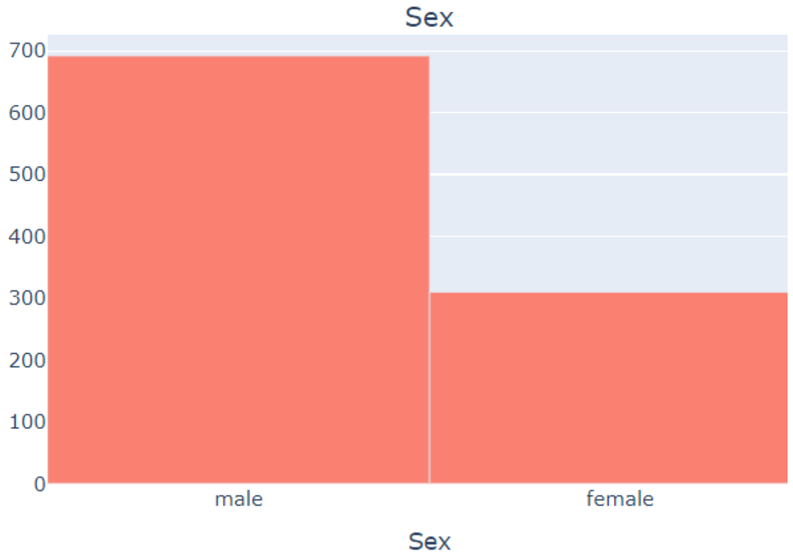
CORRELATION BETWEEN NUMERICAL VALUES

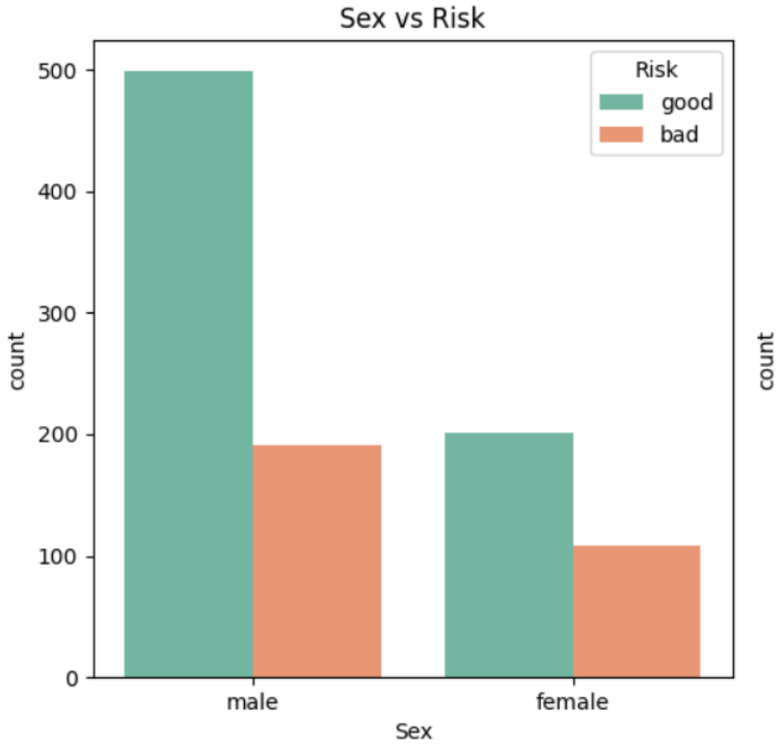


INFERENCE:

Credit amount and duration has correlation >0.5 . hence it shows a positive strong relation

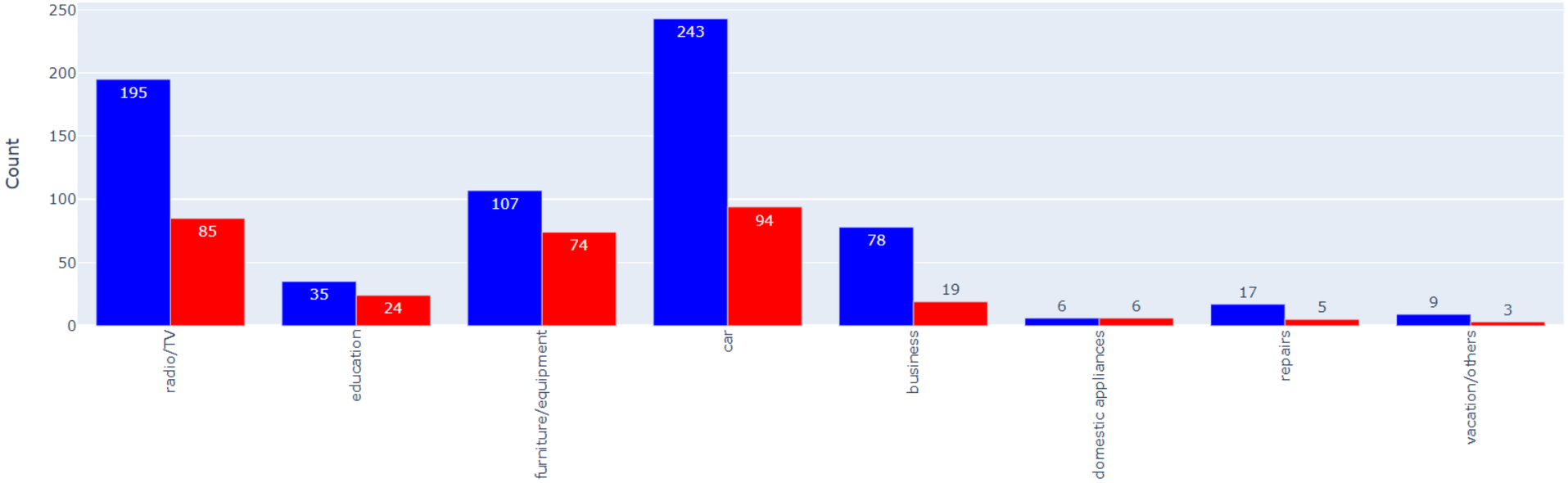
SEX ANALYSIS



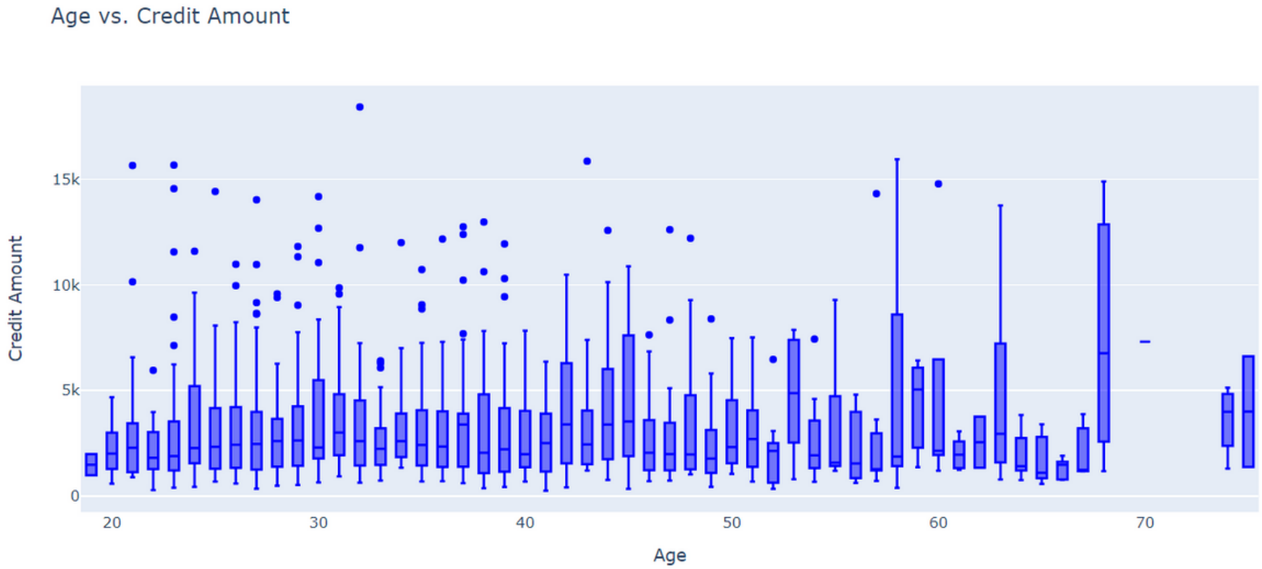


Inference:

- Males are twice as females
- Loan for car and business is mostly taken by males

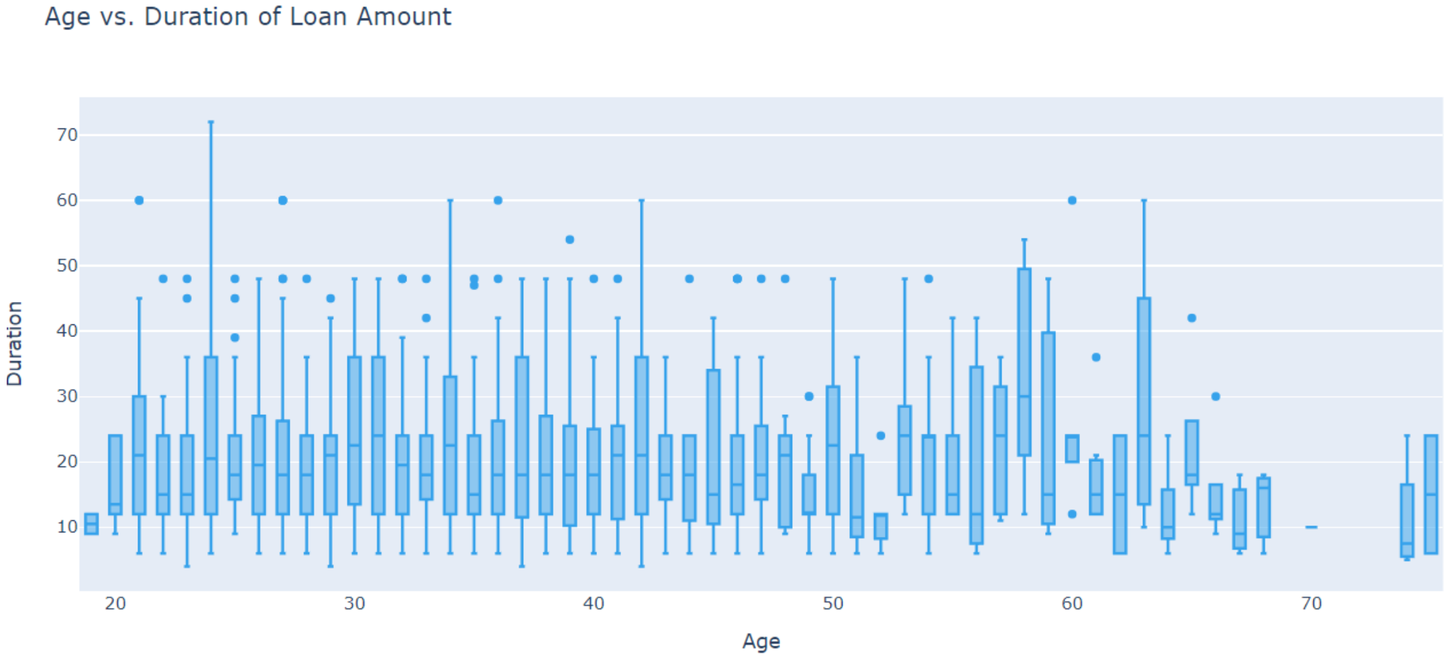


AGE ANALYSIS

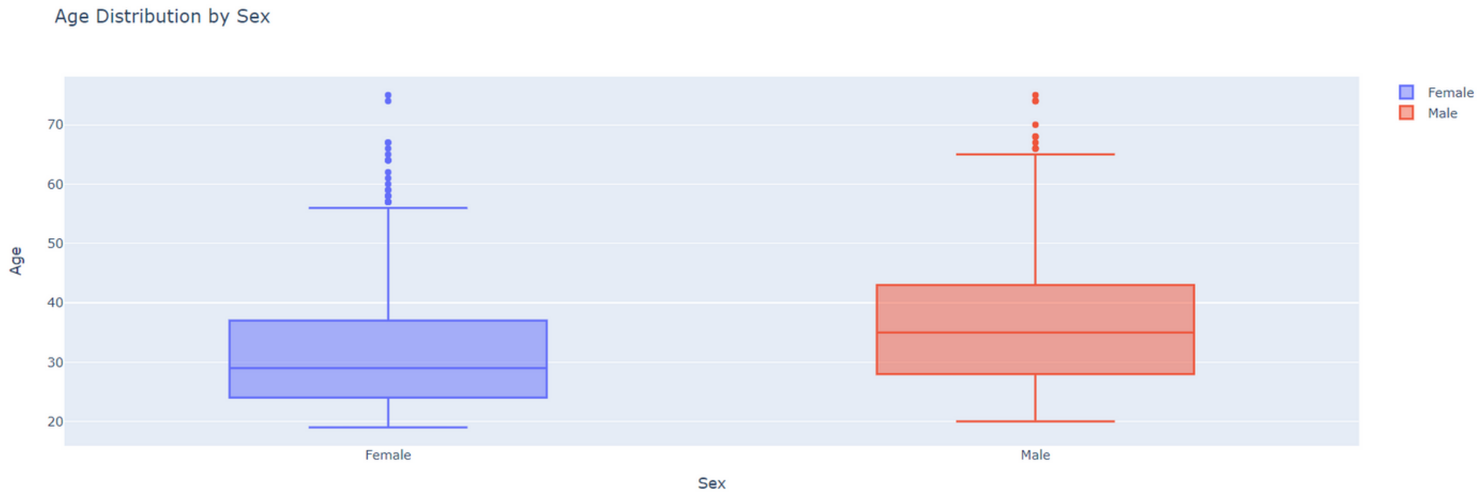


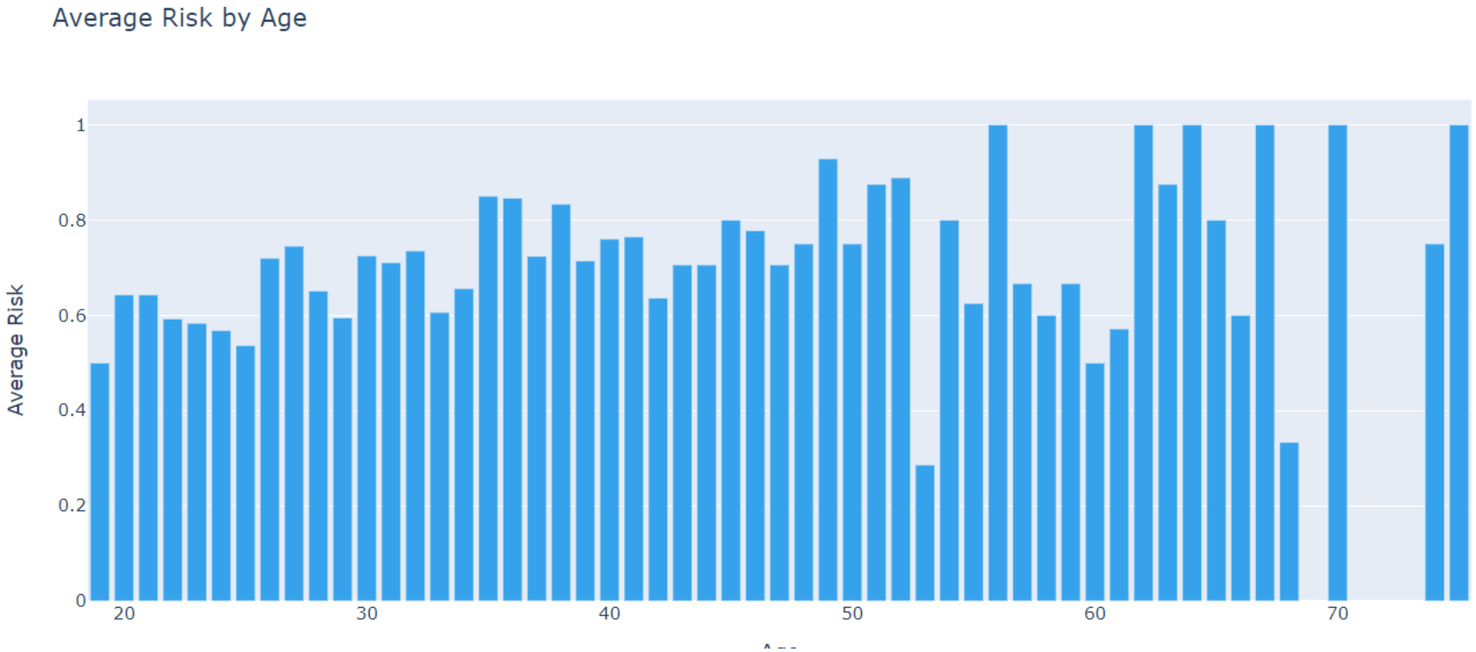
From the Credit Amount vs Age Plot, it can be seen that, majority of the customers below age 50 have the Credit Amount less than 10k.

From the Duration vs Age Plot, it can be seen that, majority of the customers have the average Loan Amount Duration of about 25 months. This shows, most of the customers do not want the loan to be repaid in large interval of time. Large interval time Loans increase interest debt on customers and also it is risky for the banks.



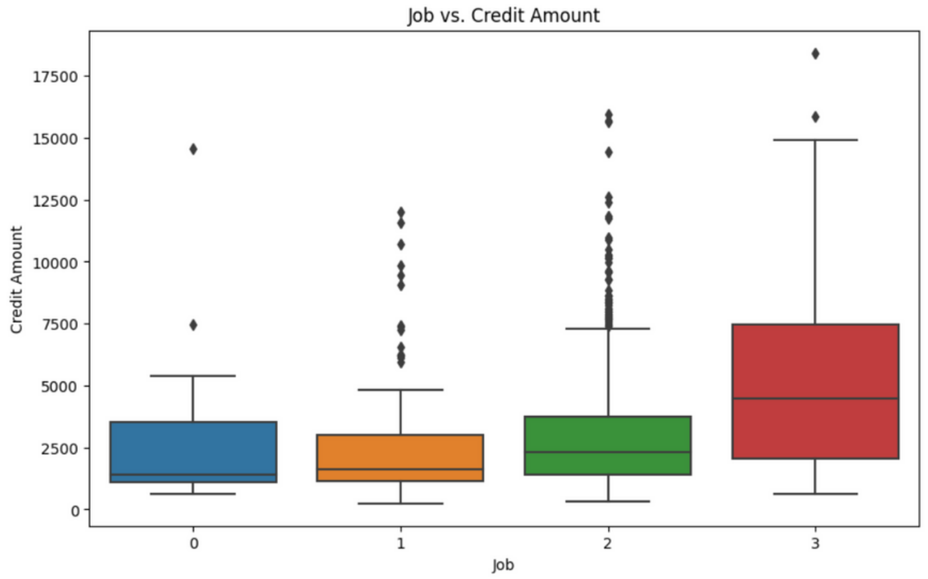
AGE ANALYSIS





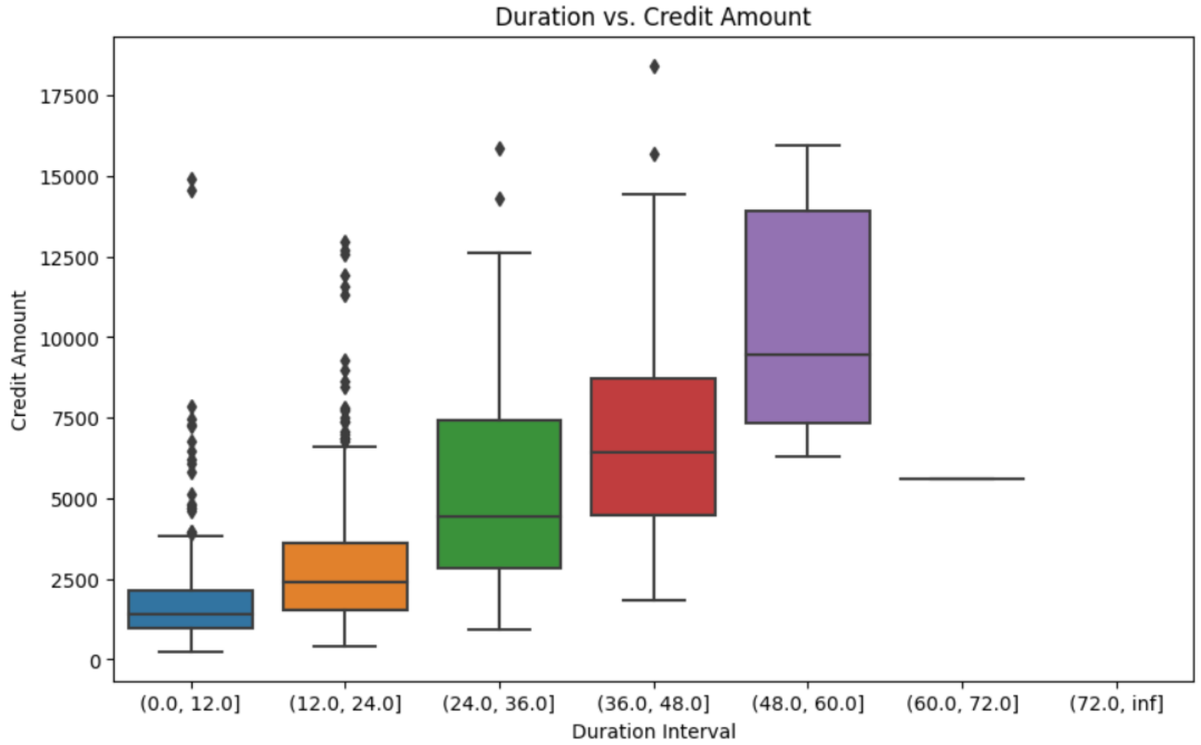
It can be seen from the analysis that Risk factor for the Customers, more than the age of 52 years is more and it is less for the customers with age less than or equal to approx. 35 years.

CREDIT AMOUNT ANALYSIS

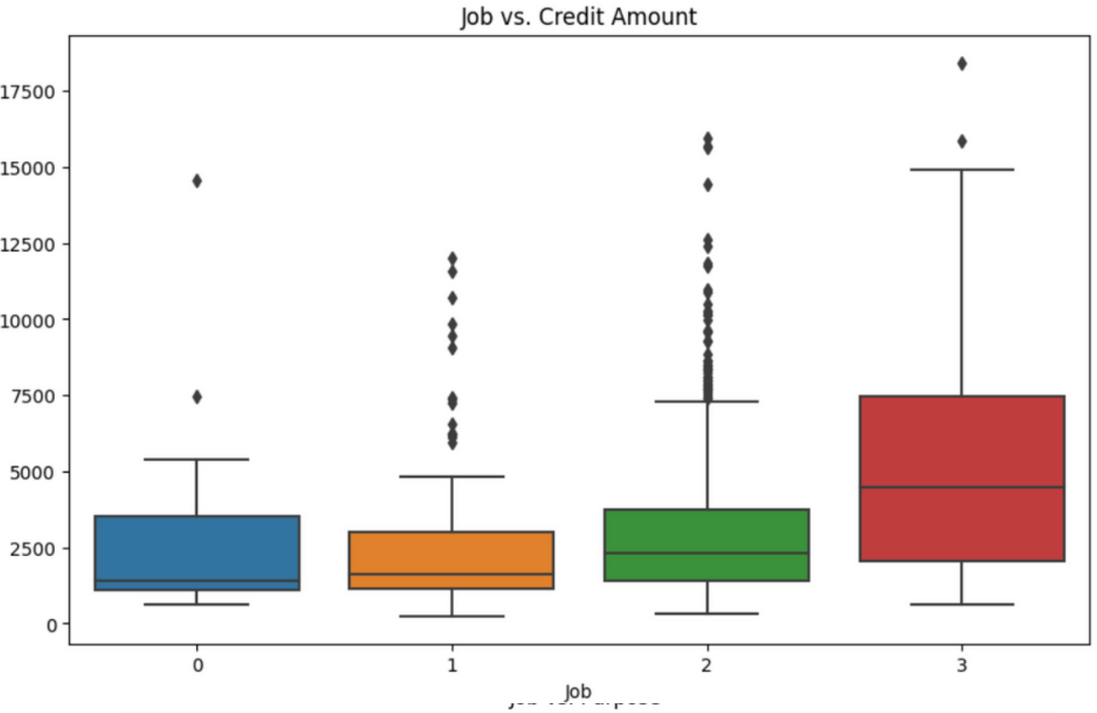


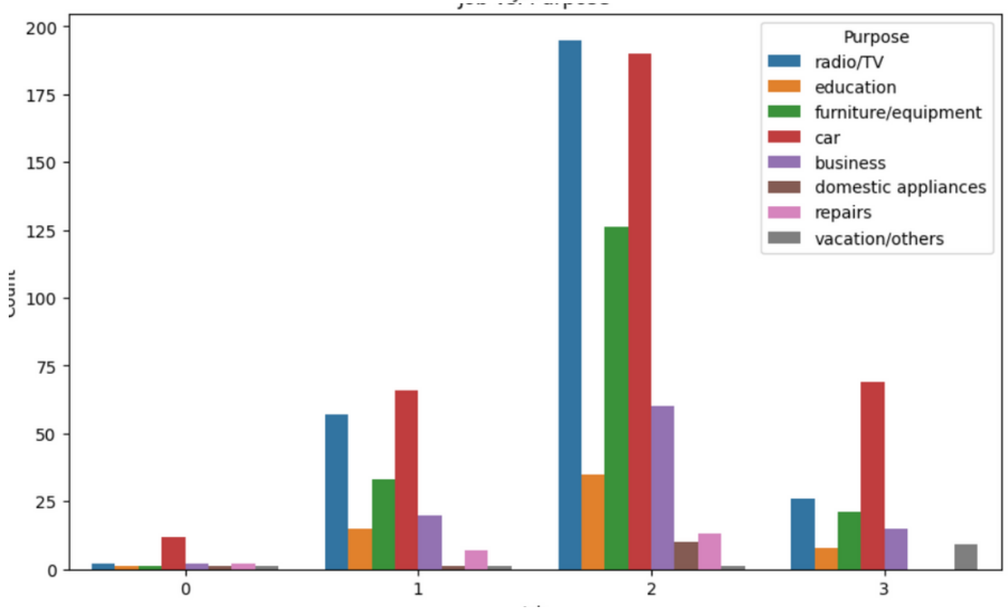
From the Analysis of the Credit Amount with different Features, it was obvserved that highly skilled customers had high credit amounts as compared to others. Also

It is also obverved that with the incease in Credit Amount, there duration of the Loan also increases as Customers ask for more time to repay the large debts



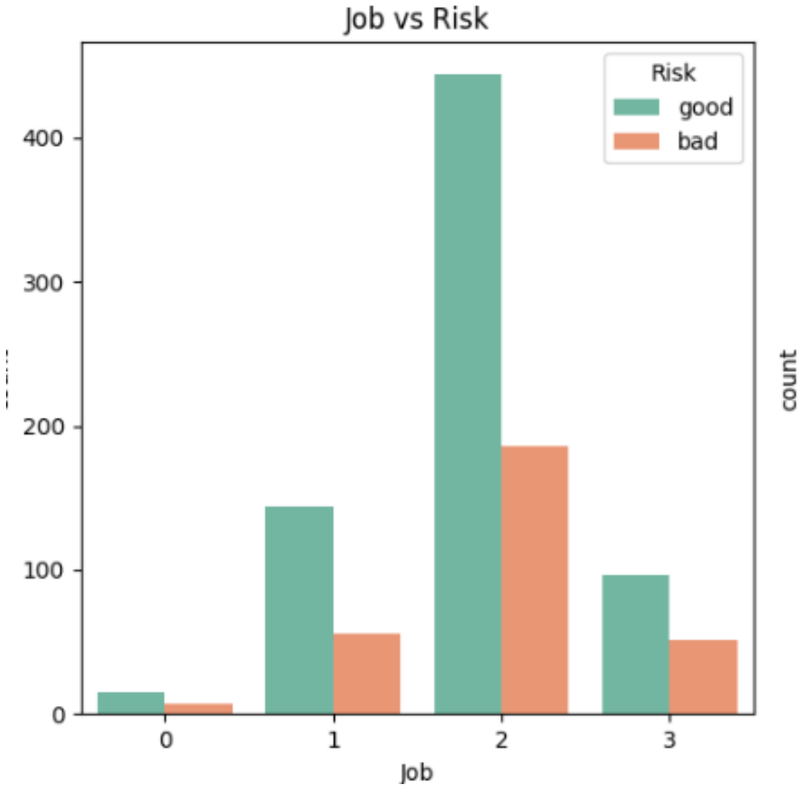
JOB ANALYSIS



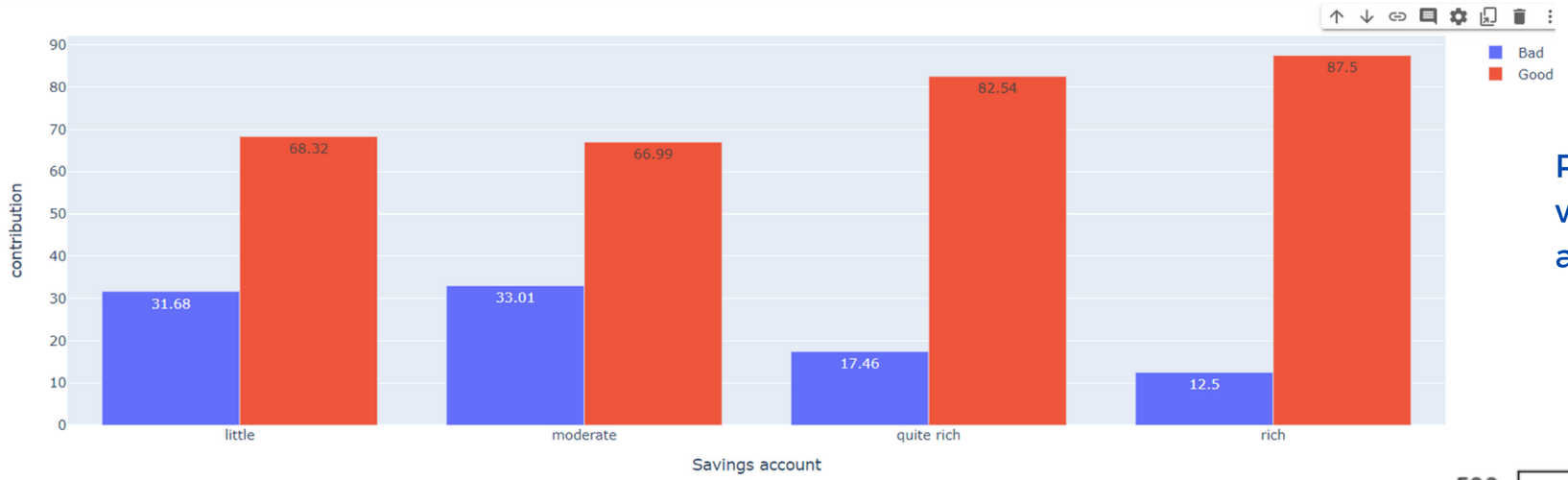


INFERENCE:

Category of unskilled non resident apply for loan car.
Highly skilled people applied for the high credit amount.
skilled people usually showed good risk



SAVINGS ACCOUNT



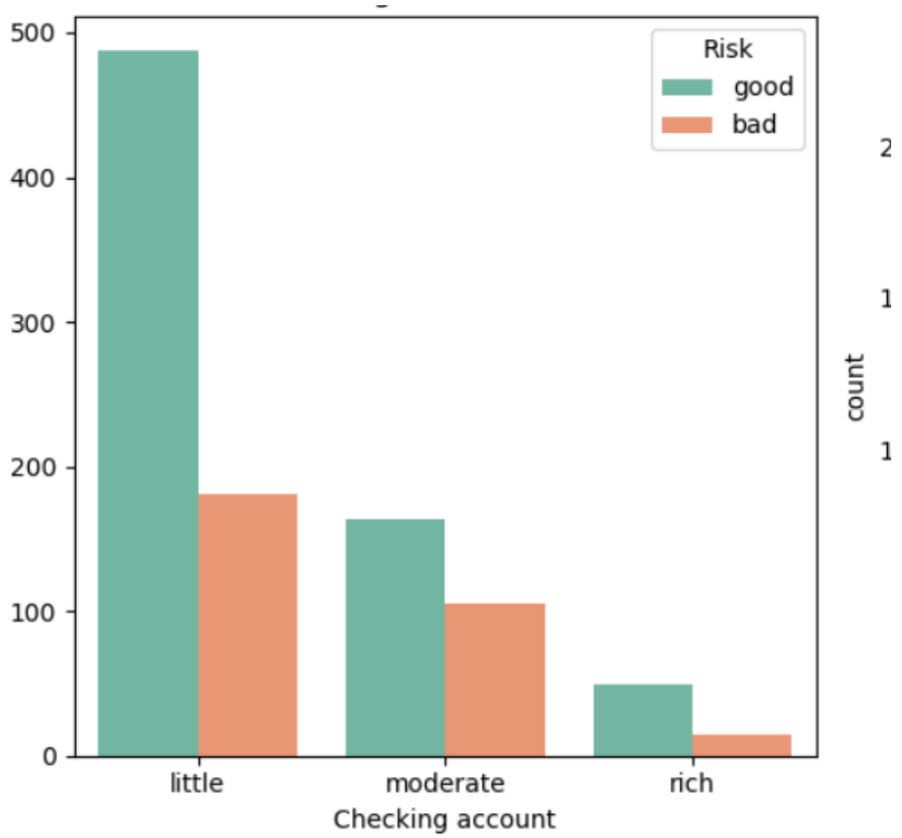
Inference:

Rich and quite rich have mostly good risk . whereas , little and moderate does not give good assurance of good risk

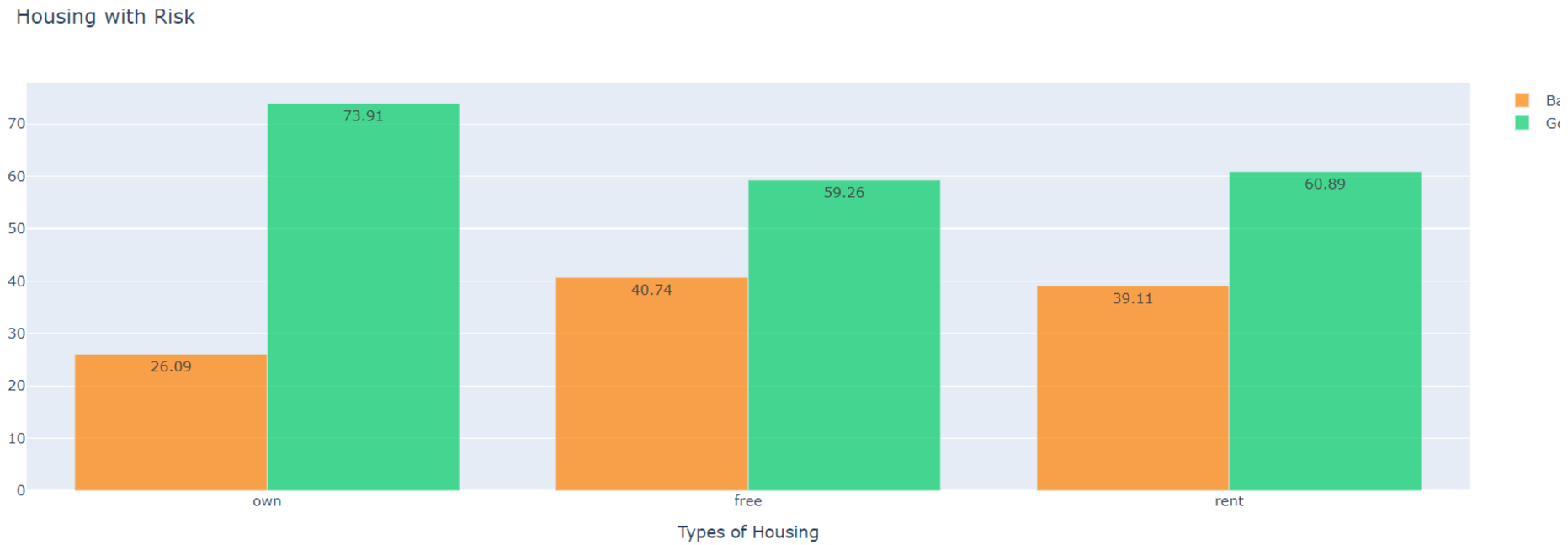
CHECKING ACCOUNT

Inference:

For little checking account , we get a discriminative difference between good and bad risk.

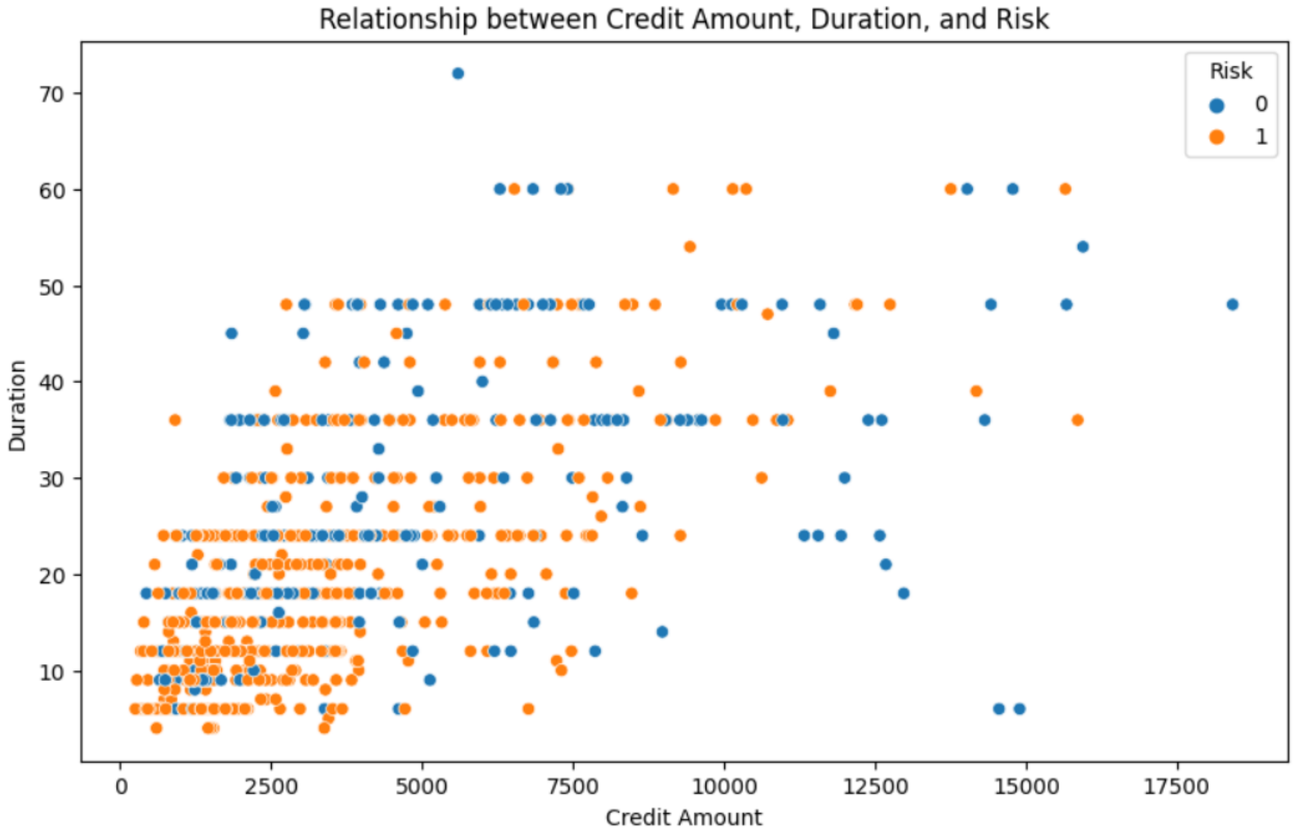


HOUSING



INFERENCE:

own housing promises good risk , and free housing type is ambigous



MULTIVARIATE ANALYSIS

Features like Credit Amount, Duration and Risk have high correlation between them. This multivariate analysis shows that risk factor increases if higher amount of loan is provided for less interval of time. Also risk factor appears good for Credit amount less than 10000 and duration less than 40 years.

SUMMARY OF EDA

Distribution of features:

- Age- Rightly skewed with maximum people belonging to age 23 -40.
- Gender-Males are more than twice than females.
- job data for highly skilled people is the most.
- saving account - the little category has the most data
- own housing contributes to the maximum part of data
- little category of checking account is the most
- Credit amount is rightly skewed and the maximum count lies in the range of less than 5000
- Most of the people wants loan for car.

Conclusion:

- Personal information (Age, sex, job) does not directly help to establish relation with risk. Instead it establishes a better relation with other loan features.
- Wealth information(Savings account, checking account, housing,duration) gives better understanding of the risk status.

DATA PREPROCESSING - IMPUTING NULL VALUES

Two columns have null values - Saving accounts and Checking Account.

Both the columns have null values less than 50%, hence we need to impute them.

Imputing using random method will be good since it will not make data skewed, and there is no specific pattern present in the dataset tested using MCAR curve.

FEATURE TRANSFORMATION

We performed feature transformation using One Hot encoding and Label encoding.

- One-hot encoding creates binary columns for each category, indicating the presence or absence of that category and hence is suitable for nominal categorical variables.
- Label encoding assigns a numerical label to each category, preserving the ordinal relationship between the categories and hence is suitable for ordinal categorical variables.
- One-hot encoding was applied to the 'sex', 'housing', and 'purpose' columns because they represent categorical variables with multiple distinct categories.
- On the other hand, label encoding was applied to the 'Saving accounts' and 'Checking account' columns because they represent ordinal variables with a specific order or hierarchy among categories.

FEATURE SELECTION

- We calculate the correlation matrix of the encoded DataFrame and select the top 11 columns that are strongly correlated with the 'Risk' variable by setting a threshold. It then creates a new DataFrame with only these selected columns.

- Selecting features that are strongly correlated with the target variable ('Risk' in this case) can be beneficial in machine learning tasks, especially for predictive modeling

- Features that have a high correlation with the target variable often contain valuable information related to the outcome of interest.

- By selecting only the top correlated features, we can effectively reduce the dimensionality of the dataset while retaining the most relevant information.

```
Duration                        0.237889
Credit amount                   0.159481
Housing_own                     0.150420
Purpose_radio/TV                0.132988
Saving accounts                 0.107867
Housing_rent                    0.104133
Age                             0.086629
Checking account                0.060083
Sex_male                        0.059970
Job                             0.049078
Purpose_education               0.047088
Purpose_car                     0.027087
Purpose_vacation/others         0.025361
Purpose_furniture/equipment     0.023914
Purpose_repairs                 0.014563
Purpose_domestic appliances     0.006522
Name: Risk, dtype: float64
```

MODELLING

Building a model that can predict whether a loan should be given to customer or not based on its features.

Our main aim is to reduce False Positives (Customer's Risk was bad, but model predicted it to be good).

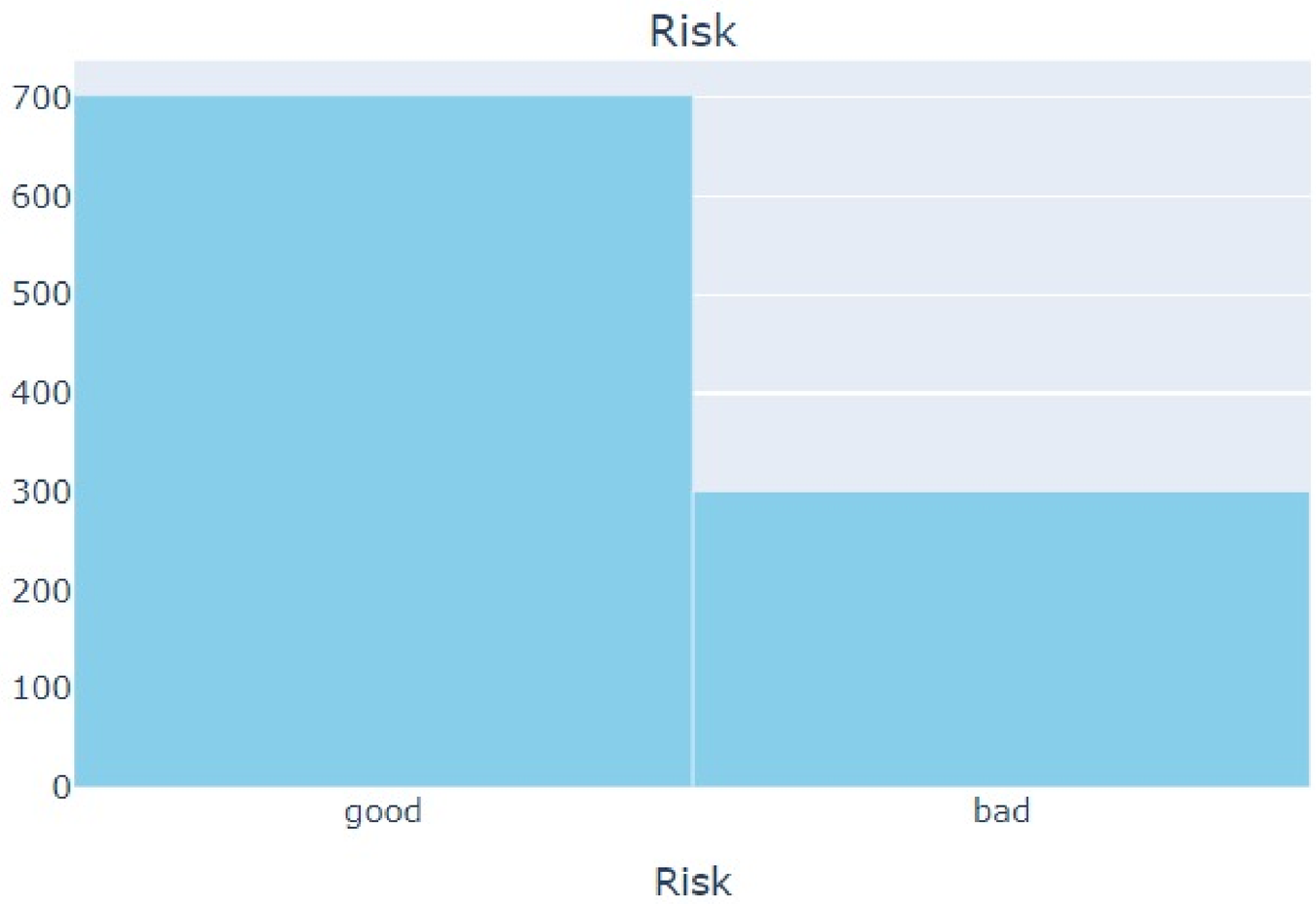
PIPELINE

1. Label Imbalance
2. Selecting Models
3. Hyperparameter Tuning
4. Selecting the best Model

LABEL IMBALANCE

Label imbalance occurs when the distribution of classes in a dataset is significantly skewed, with one or more classes having a disproportionately small number of instances. In our dataset, good risk samples are almost 2.5 times than bad risk samples

This poses challenges for classification models, as they tend to favor the majority class, leading to poor performance on minority classes



APPLYING SMOTE - FOR IMPROVED CLASSIFICATION

SMOTE, an acronym for Synthetic Minority Over-sampling Technique, is a popular technique for addressing label imbalance. Rather than relying solely on the existing data, SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances. By artificially increasing the number of minority class instances, SMOTE aims to balance the class distribution and alleviate the bias toward the majority class.

We balanced our data by applying SMOTE.

MODELS

LIGHTGBM AND CATBOOST ARE PERFORMING WELL BECAUSE THEY HAVE HIGHER ACCURACY AND LOW FALSE POSITIVE VALUES

Model	Accuracy	False Positive (out of 280)
XG BOOST	73.21	35
Logistic Regression	70.36	45
Random Forest	71.4	36
Decison Tree	68.57	34
Gradient Boosting	71.78	39
AdaBoost	70	40
CatBoost	75	29
LightGBM	76	30

WEIGHTS USED BY BEST MODEL

```
Feature: Duration, Importance: 0.1490824839786709
Feature: Credit amount, Importance: 0.11902359648569671
Feature: Housing_own, Importance: 0.12464546484542192
Feature: Saving accounts, Importance: 0.1103245346946578
Feature: Purpose_radio/TV, Importance: 0.15584376486636412
Feature: Checking account, Importance: 0.08070946743437567
Feature: Housing_rent, Importance: 0.04662124893193375
Feature: Age, Importance: 0.09392793822717944
Feature: Sex_male, Importance: 0.0564519210742581 7
Feature: Purpose_education, Importance: 0.0036803952981440064
Feature: Job, Importance: 0.05968918416329796
```

LOAN RECOMMENDATOIN USING COLLABORATIVE FILTERING

WHY LOAN RECOMMENDATI0N?

- In the lending industry, personalized loan recommendations are crucial for meeting customer needs and ensuring responsible lending practices.
- Our project aims to leverage collaborative filtering techniques to provide tailored loan recommendations to users.
- Banks recommend loans to help customers maximize the benefits associated with borrowing.
- By recommending loans, banks enable customers to make optimal choices that result in cost savings and improved loan terms.
- Banks have various risk management tools and frameworks in place to assess and mitigate credit risk. By recommending loans, banks can carefully evaluate borrowers' creditworthiness, analyze their financial profiles, and make informed decisions. Effective credit risk management helps banks maintain a healthy loan portfolio and minimize the risk of defaults and non-performing loans.

MODEL DEVELOPMENT AND RECOMMENDATION

- The loan recommendation model was built using collaborative filtering techniques.
- We used cosine similarity to calculate the similarity matrix among users, capturing their loan preference similarities.
- By identifying the top similar users, our model generated loan purpose recommendations based on their preferences.
- We developed a function to recommend loan purposes for new users.
- The function selected the top N similar users based on their loan preferences.
- The recommended loan purposes were determined by aggregating the loan preferences of the top users.

METHODOLOGY

- We employed collaborative filtering, a popular recommendation technique, for our loan recommendation system.
- To measure user similarity, we utilized cosine similarity, a metric that calculates the similarity between users based on their loan preferences.

DATA PREPARATION

- Our analysis was based on a dataset comprising customer Credit amount, Duration, Risk and loan purposes.
- We transformed the data into a user-item matrix, representing the loan preferences of users.
- To ensure data quality, we performed preprocessing tasks such as handling missing values and normalizing the data.

RESULTS

```
Recommended loan purposes for user 156 :
1 - business
2 - furniture/equipment
3 - car
```

```
Recommended loan purposes for user 1 :
1 - radio/TV
2 - business
3 - car
```

```
Recommended loan purposes for user 985 :
1 - furniture/equipment
```

BENIFITS AND IMPLICATOINS

- Our loan recommendation system offers several benefits, including personalized recommendations tailored to individual users.
- By providing accurate loan recommendations, we aim to enhance customer satisfaction and improve loan acceptance rates.
- Moreover, our system contributes to risk management efforts by aligning loan purposes with users' preferences and risk profiles.

LIMITATIONS AND FUTURE DIRECTIONS:

- It is important to acknowledge the limitations of our current model.
- The system heavily relies on historical loan data, which might not fully capture users' evolving preferences.
- In the future, additional risk analysis features can be incorporate to further improve our loan recommendation system.

THANK YOU!