Credit Risk

David Yang

This is a dataset that I found online about credit risk since it is my first time seeing a dataset that's related to risk, I decided to do a quick analysis on it. I am going to do some explorations through the Credit Risk to understand the distributions and patterns within.

Introduction

The original dataset contains 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. The link to the original dataset can be found below.

Due to its complicated system of categories and symbols, several columns are simply ignored, because in my opinion either they are not important or their descriptions are obscure. The selected attributes are:
Age (Numeric)
Sex (Text: male, female)
Job (Numeric:
0 - unskilled and non-resident,
1 - unskilled and resident,
2 - skilled,
3 - highly skilled)
Housing (Text: own, rent, or free)
Saving accounts (Text - little, moderate, quite rich, rich)
Checking account (Numeric, in DM - Deutsch Mark)
Credit amount (Numeric, in DM)
Duration (Numeric, in month)
Purpose(Text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others
Risk (Value target - Good or Bad Risk)

Setting up Library/Import Dataset

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

#Importing the data
df_credit = pd.read_csv("credit_data.csv",index_col=0)
```

Quick look at the data

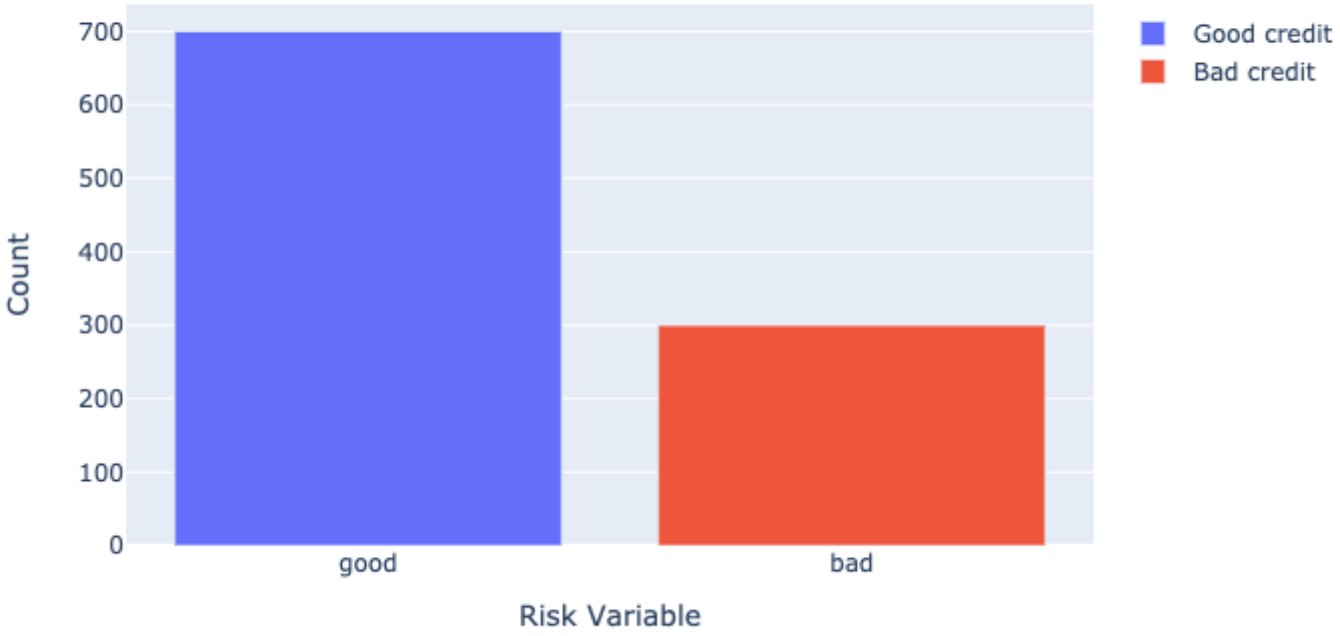
- Data type for each columns
- Null
- Unique Values
- Overview

```
#Checking info of the data
print(df_credit.info())
#Looking unique values
print(df_credit.nunique())
#Looking the data
print(df_credit.head())
```

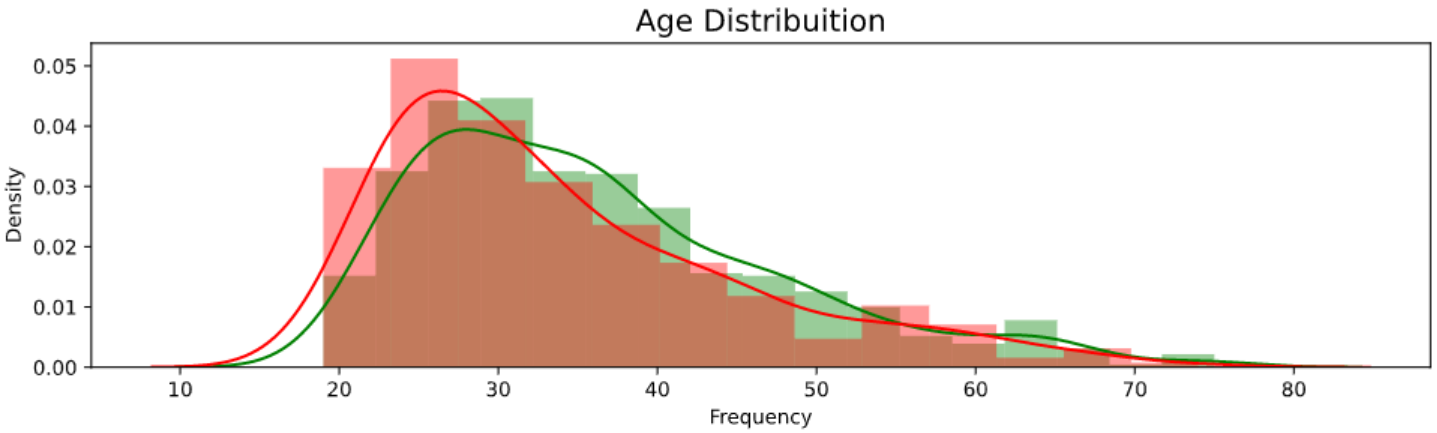
Exploring the data

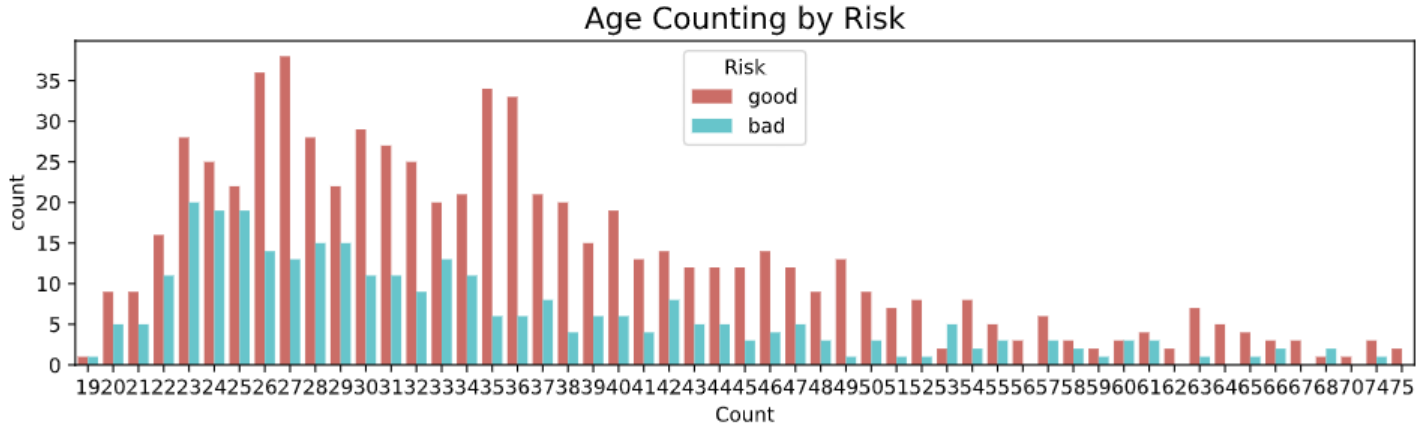
Let's start looking through target variable and their distribuition.

Target variable distribution



Age Distribuition



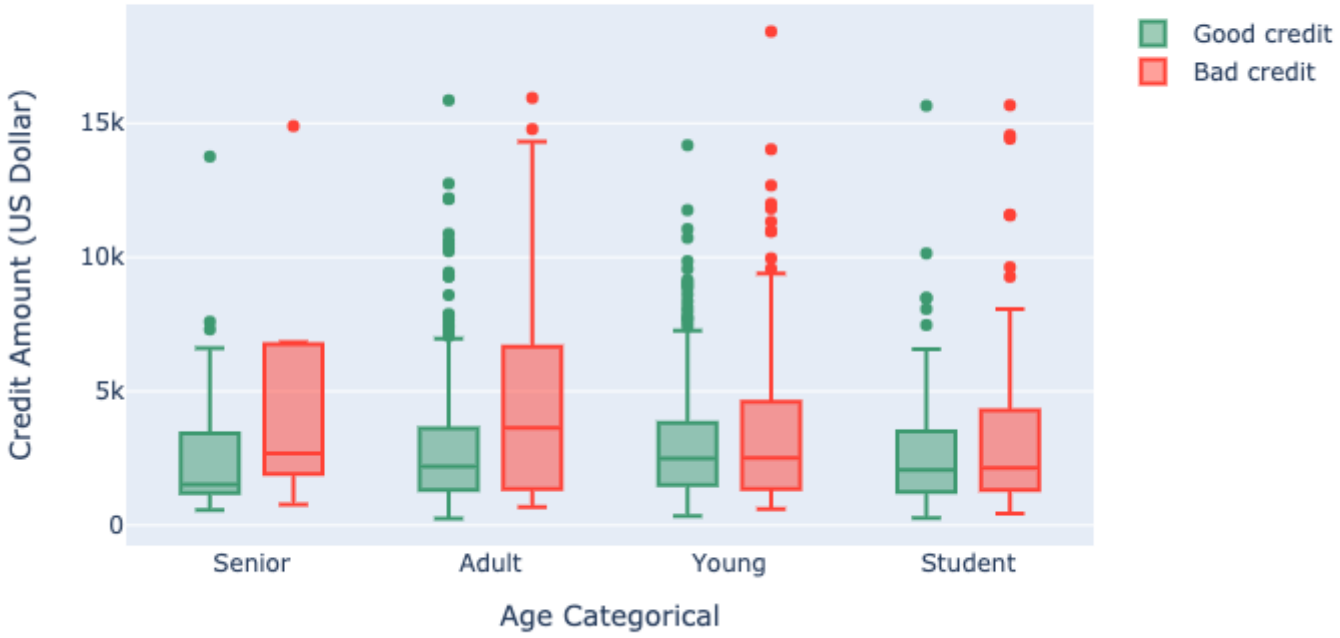


Creating an categorical variable to handle with the Age variable.

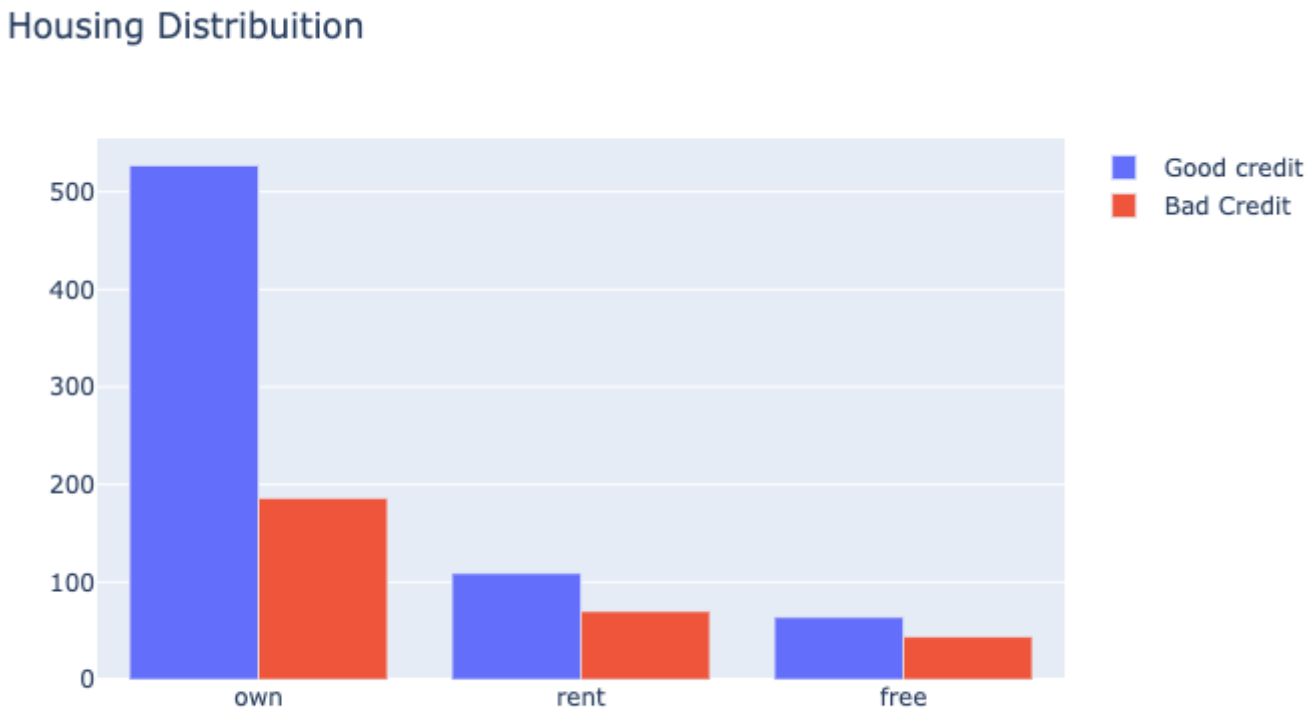
```
#Let's look the Credit Amount column
interval = (18, 25, 35, 60, 120)

cats = ['Student', 'Young', 'Adult', 'Senior']
df_credit["Age_cat"] = pd.cut(df_credit.Age, interval, labels=cats)

df_good = df_credit[df_credit["Risk"] == 'good']
df_bad = df_credit[df_credit["Risk"] == 'bad']
```

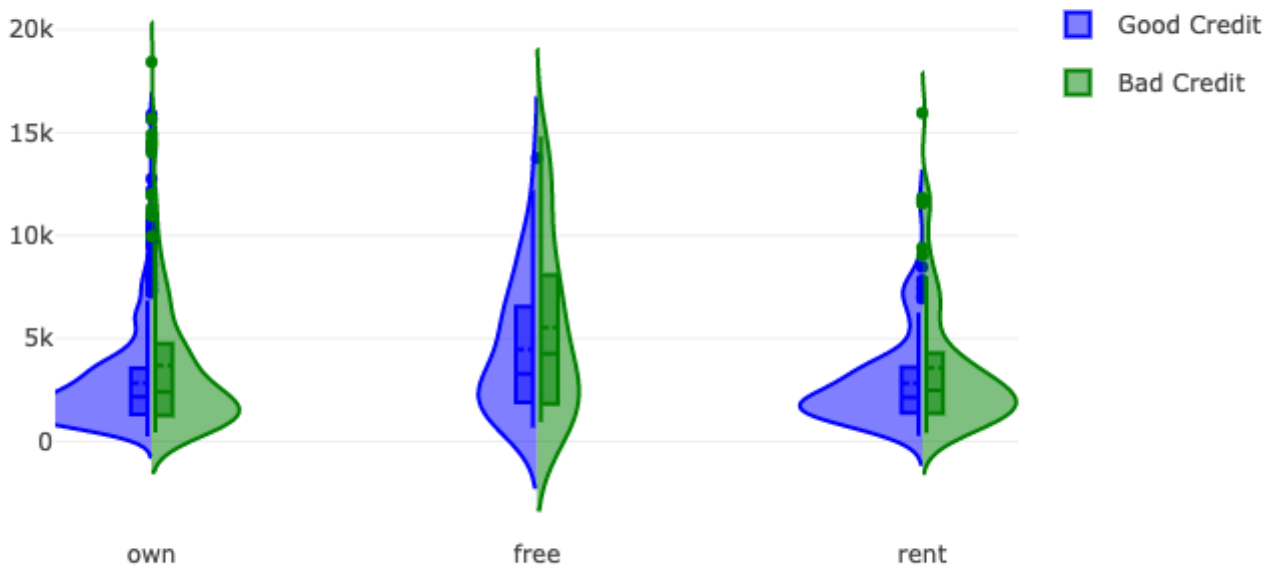


I will now look at the distribuition of Housing own and rent by Risk.



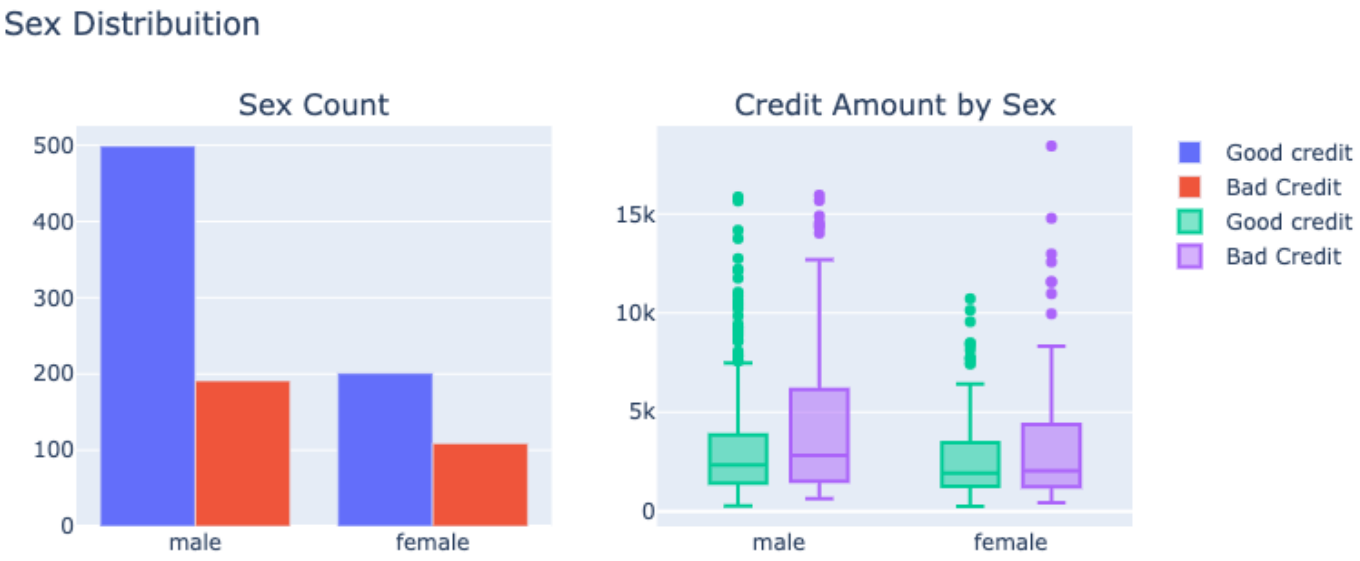
Now we can see that the own and good risk have a high correlation.

Distribuition of Credit Amount by Housing



Interesting moviments! Highest values come from category "free" and we have a different distribuition by Risk.

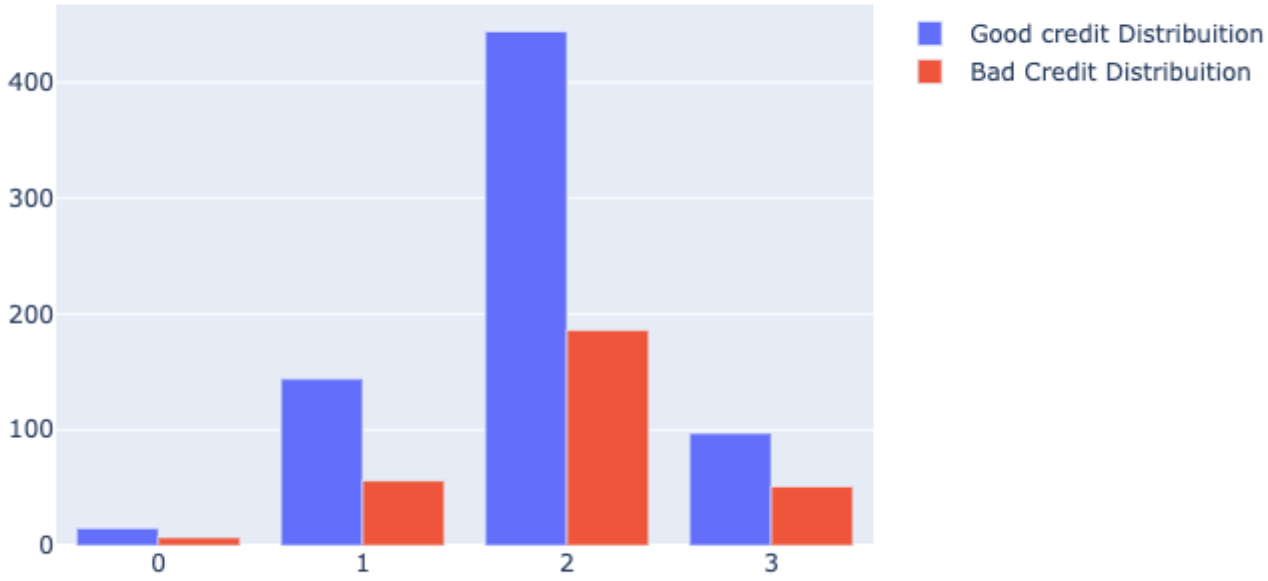
Looking the diference by Sex

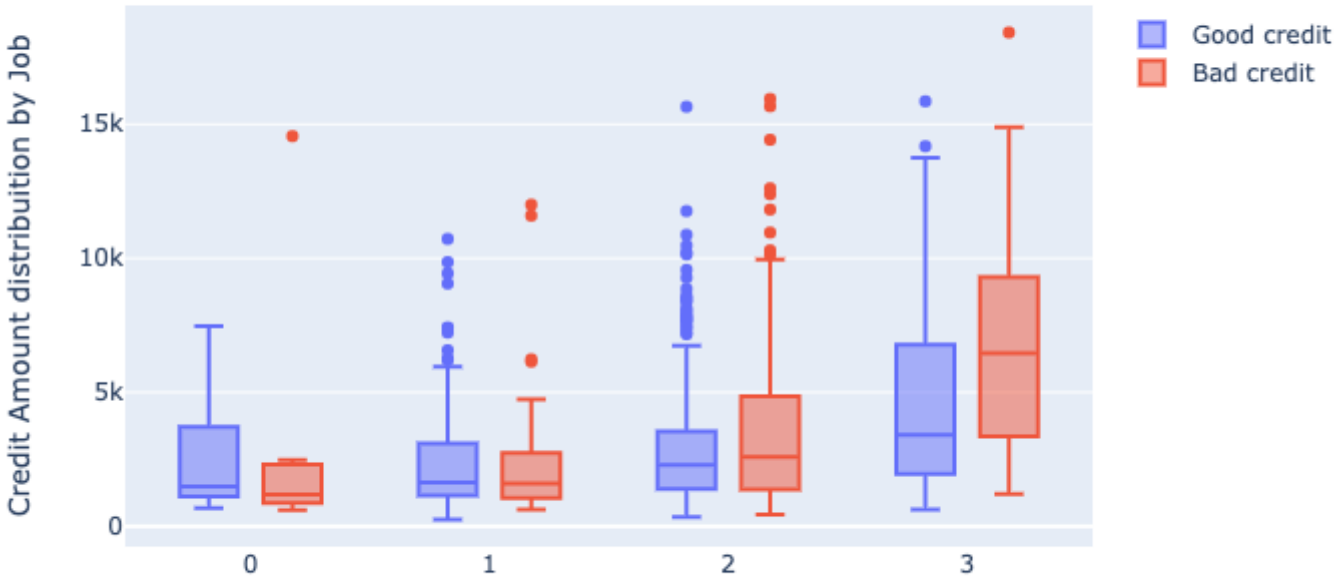


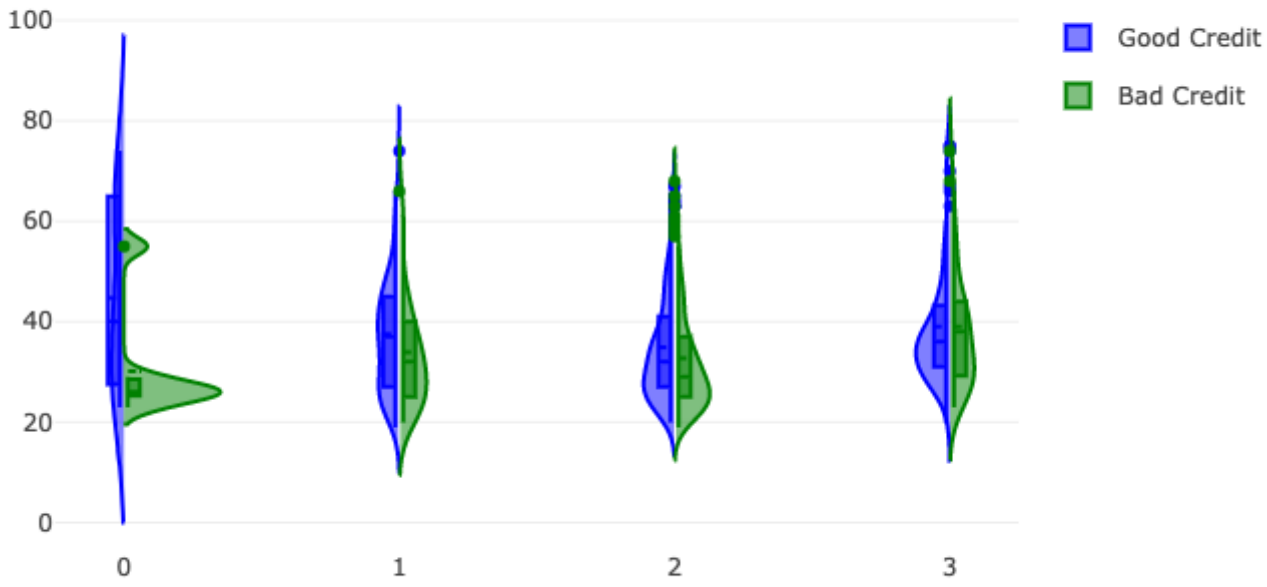
I am going to create categories of Age and look the distribuition of Credit Amount by Risk... and do some explorations through the Job.

- Distribuition
- Crossed by Credit amount
- Crossed by Age

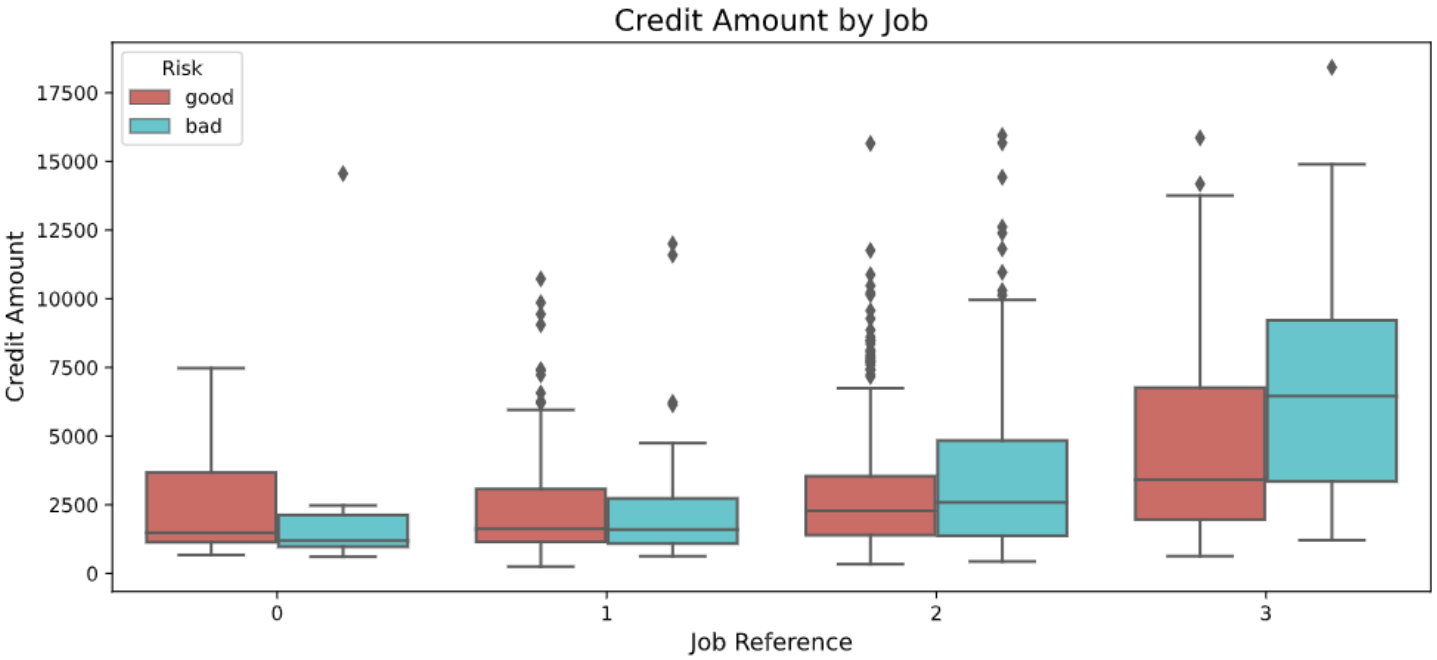
Job Distribuition



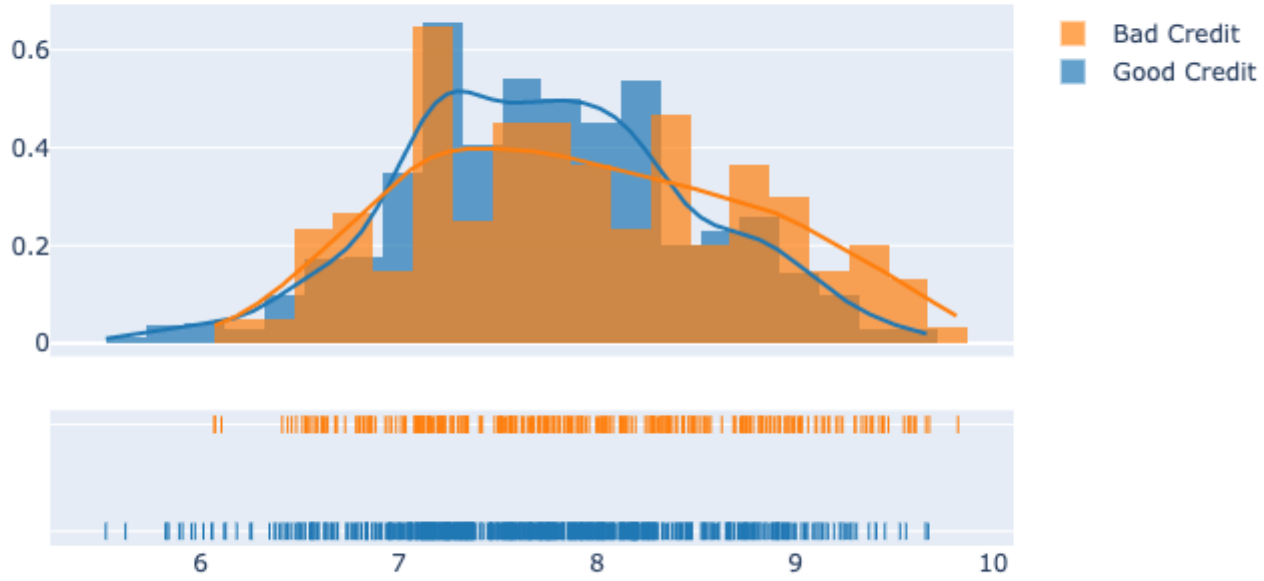


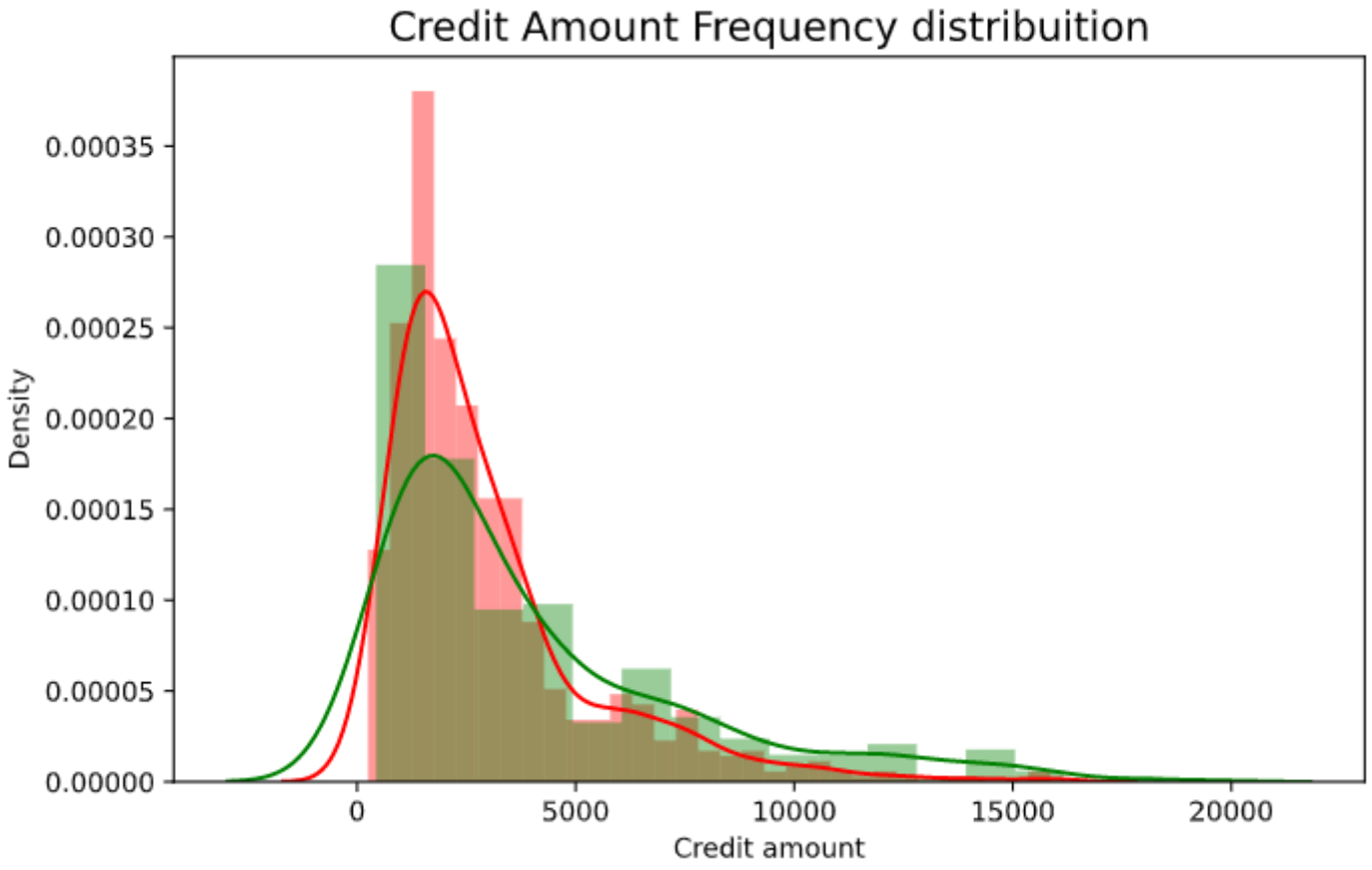


Job Type reference x Age



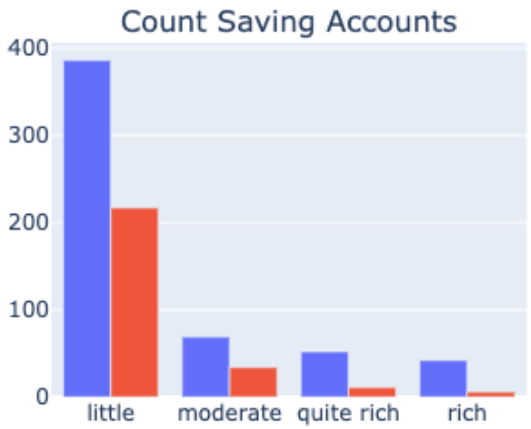
Distrubution of Credit Amount

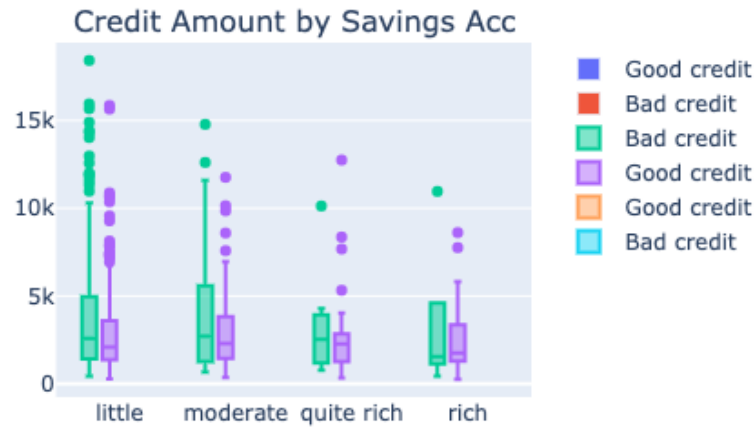


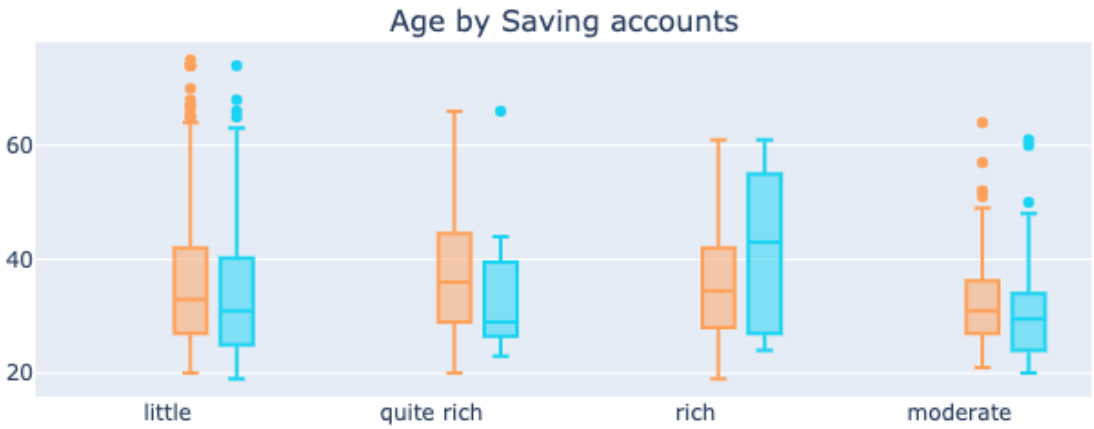


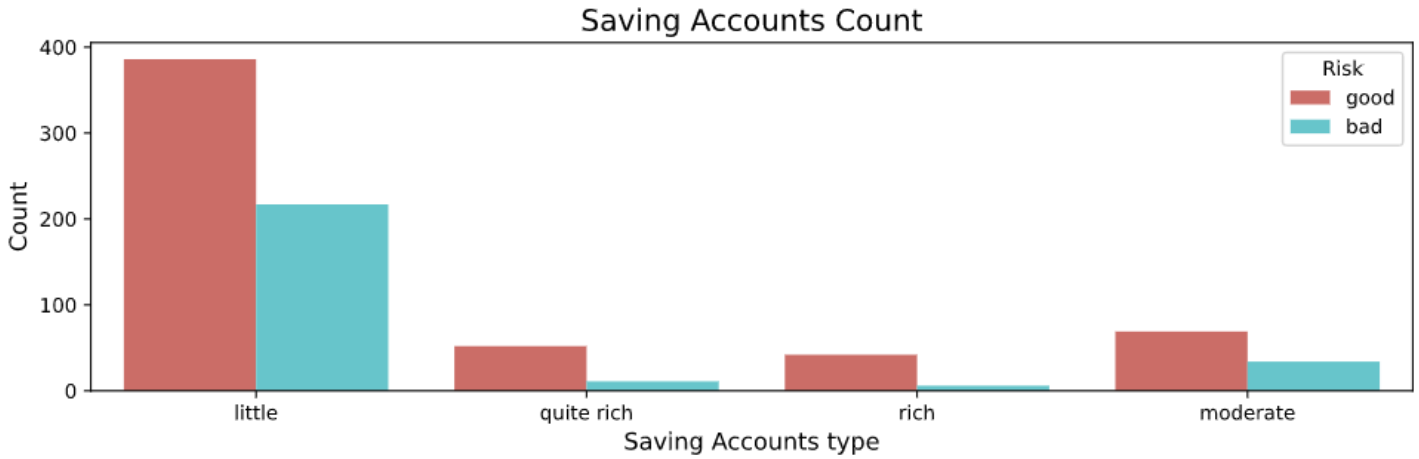
Distrubution of Saving accounts by Risk

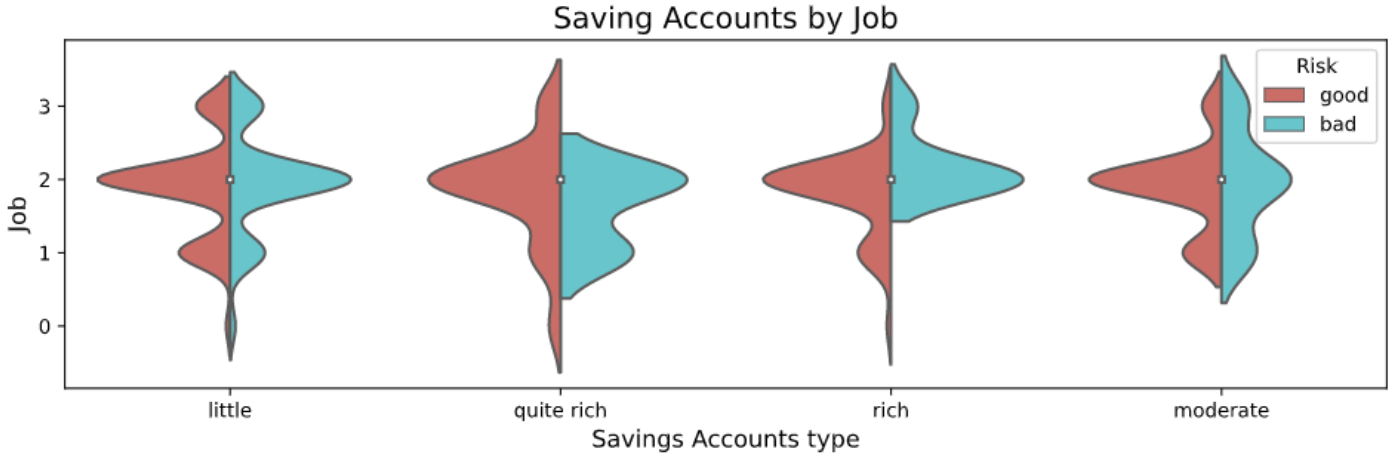
Saving Accounts Exploration

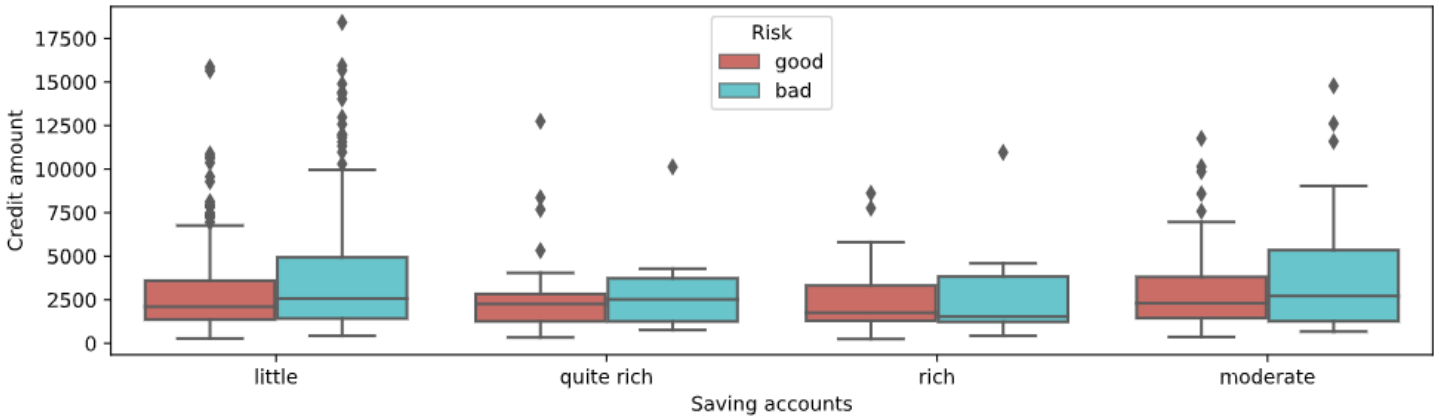


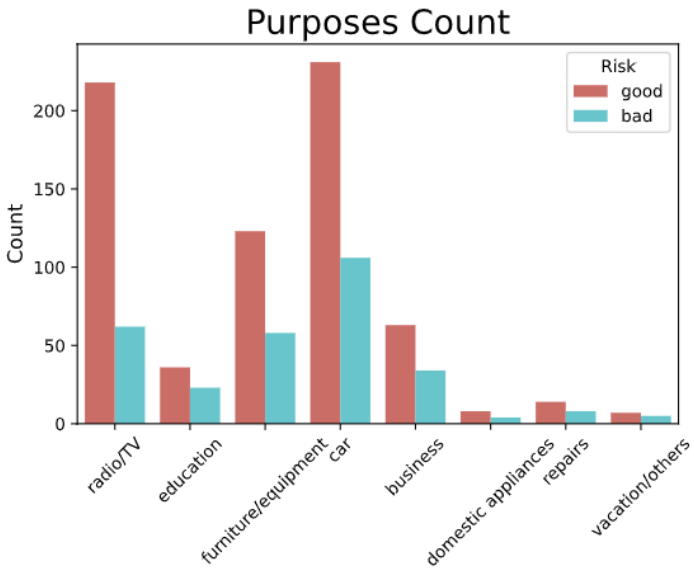


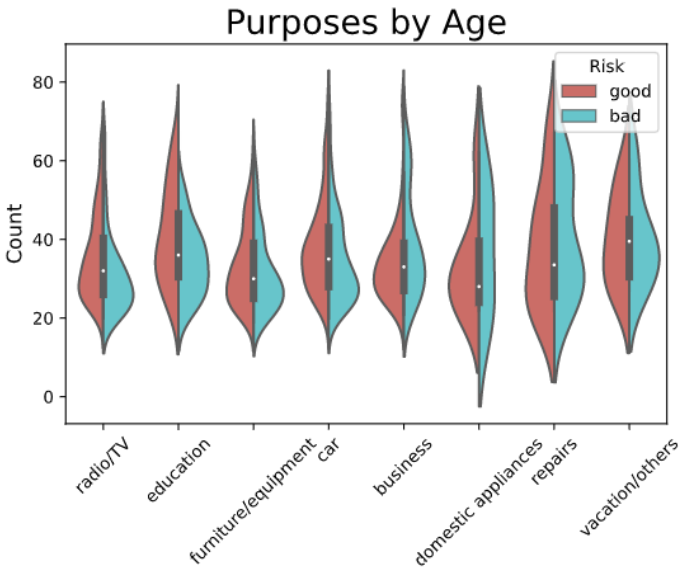


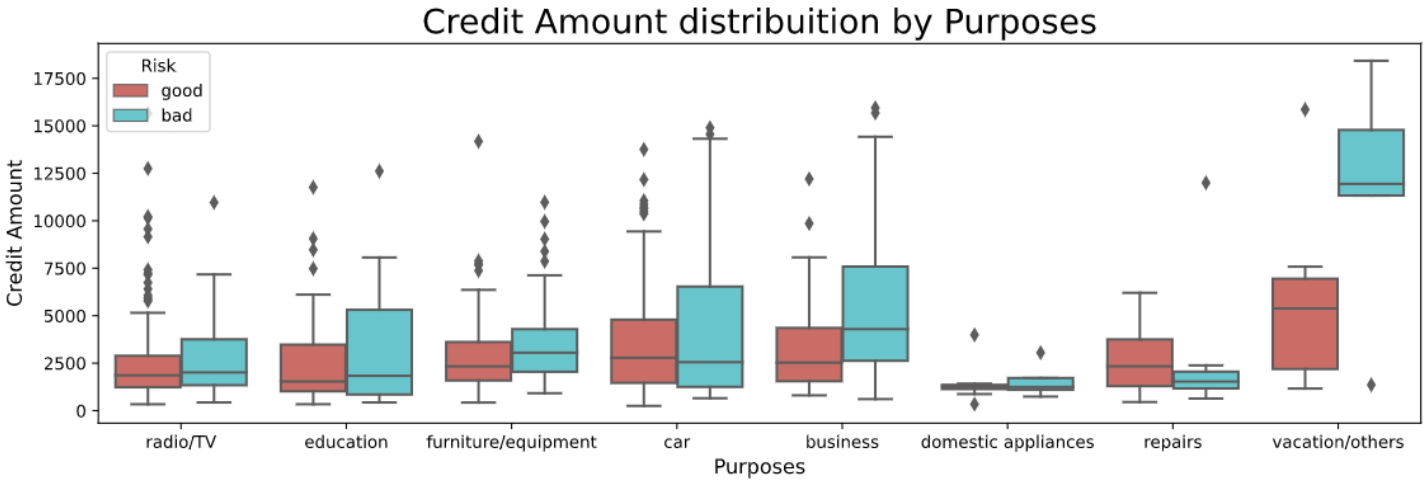

Saving Accounts Count
Count
400
300
200
100
0
Risk
good
bad
little
quite rich
rich
moderate
Saving Accounts type


Saving Accounts by Job
Job
3
2
1
0
Risk
good
bad
little
quite rich
rich
moderate
Savings Accounts type


Credit amount
17500
15000
12500
10000
7500
5000
2500
0
Risk
good
bad
little
quite rich
rich
moderate
Saving accounts

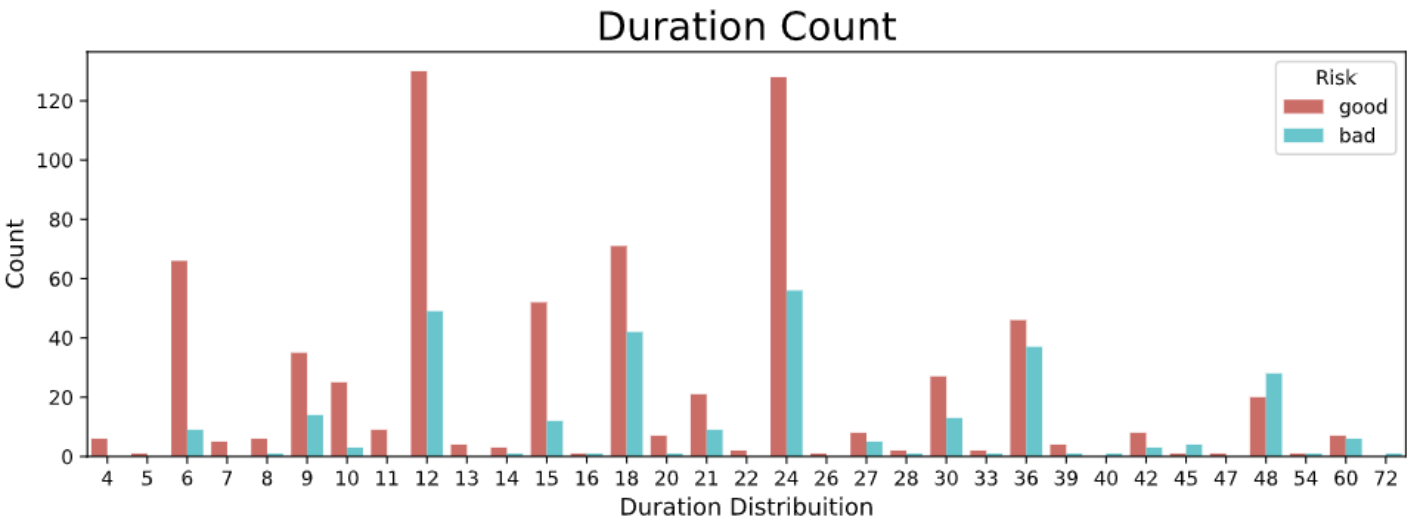


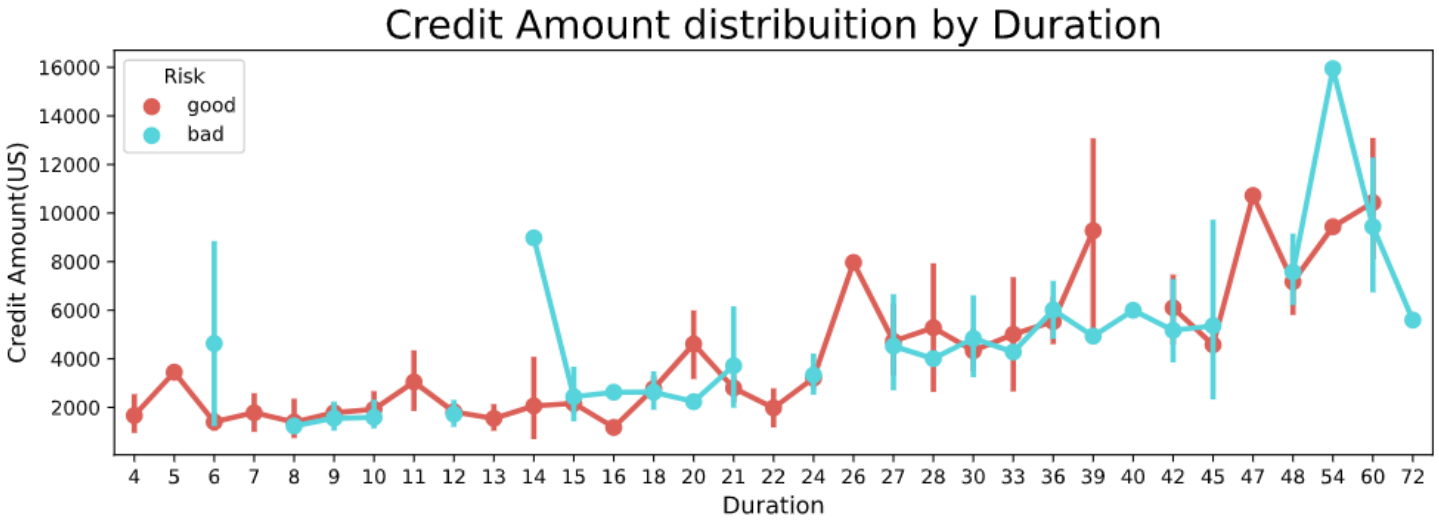


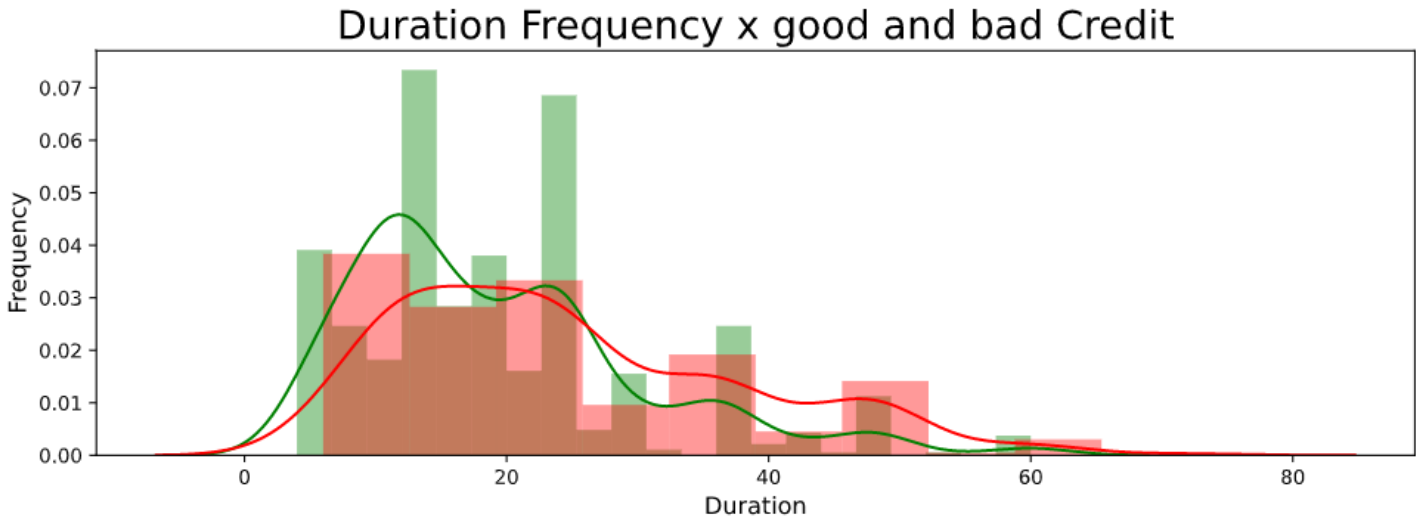



```
print("Values describe: ")
print(pd.crosstab(df_credit.Purpose, df_credit.Risk))
```

Duration of the loans distribuition and density







we can see that the highest duration have the high amounts.

The highest density is between [12 ~ 18 ~ 24] months

It all make sense.