Project document

CMP4011 – Big Data Course

|  |  |  |
| --- | --- | --- |
| **Name** | **Sec** | **BN** |
| **Khaled Hesham** | **1** | **21** |
| **Kirollos Samy** | **2** | **13** |
| **Abdelaziz Salah** | **2** | **2** |
| **Abdelrahman Noaman** | **2** | **4** |

**Problem Statement:**

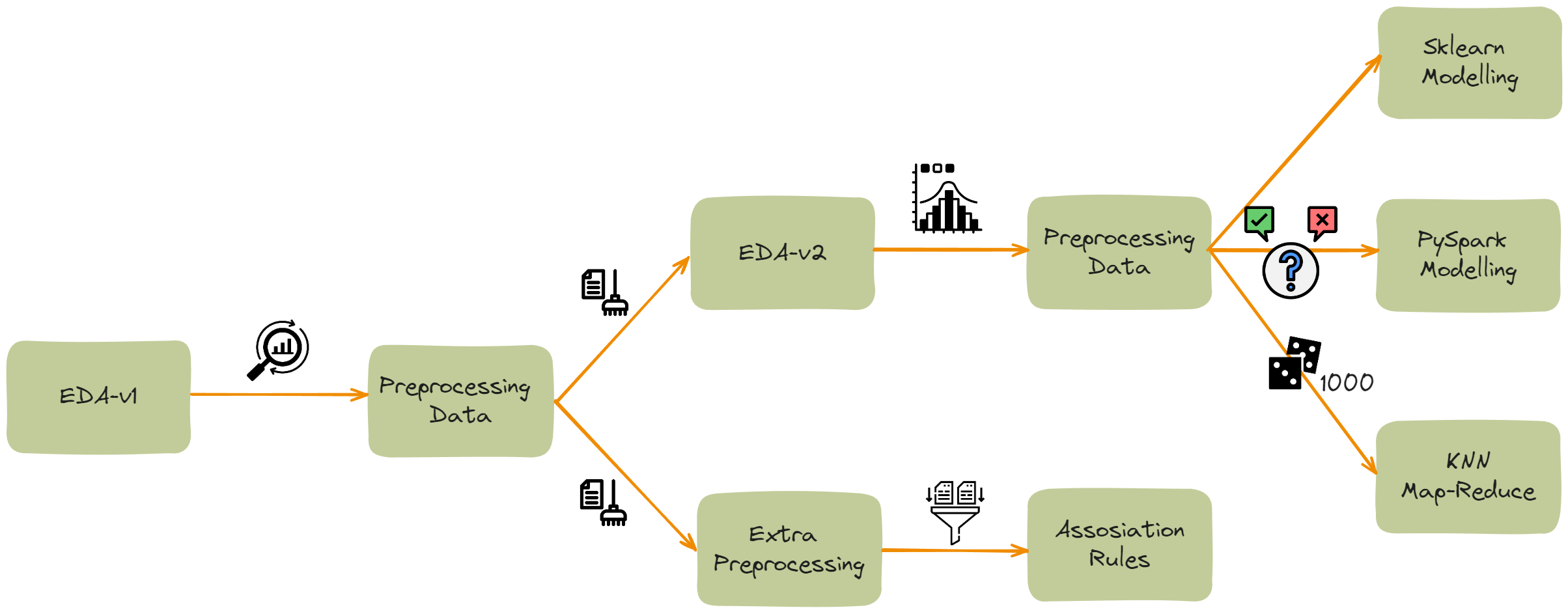
In this project we will be doing credit risk modelling of peer to peer lending Bondora systems, Credit risk modeling involves analyzing data to assess the likelihood that a borrower will default on a loan or fail to meet their financial obligations. Peer-to-peer lending platforms like Bondora facilitate lending directly between individuals, bypassing traditional financial institutions. In the context of Bondora's peer-to-peer lending system, credit risk modeling would likely involve analyzing borrower data, such as credit scores, income, employment history, and other relevant factors, to predict the likelihood of default for each borrower. This helps investors on the platform make informed decisions about which loans to fund and manage their risk exposure.

**Dataset** [**🔗**](https://www.bondora.com/en/public-reports)**:**

Link: https://www.bondora.com/en/public-reports

Rows: **372541**, Columns: **112**, Size: **303 MB**

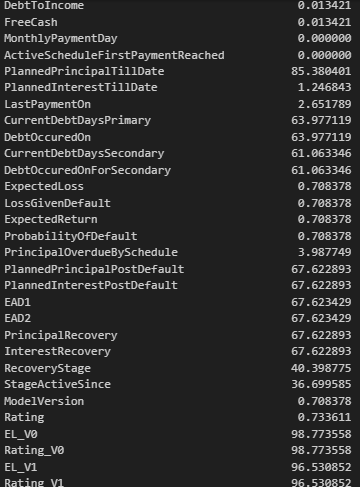
**Project Pipeline:**

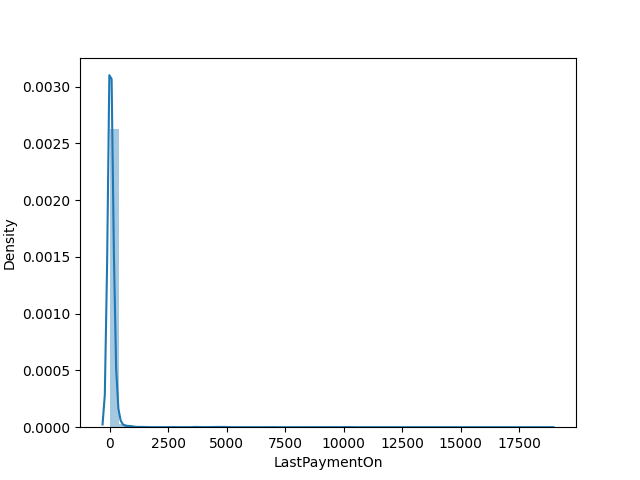
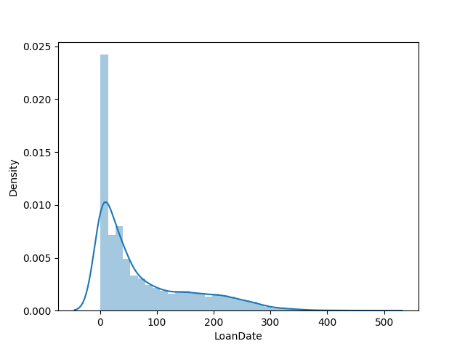
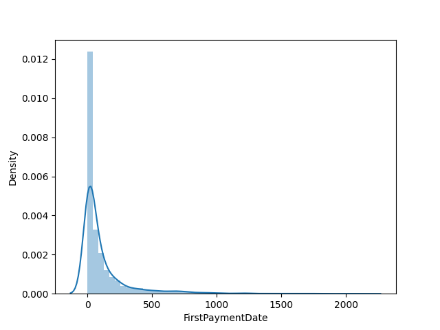
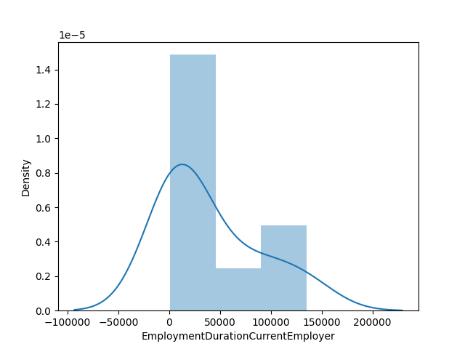
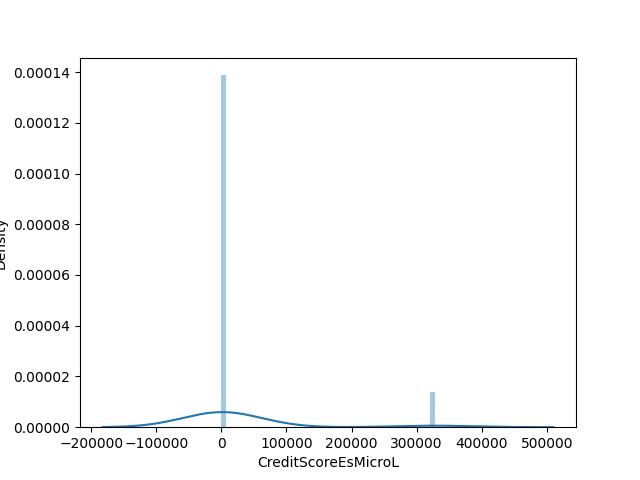
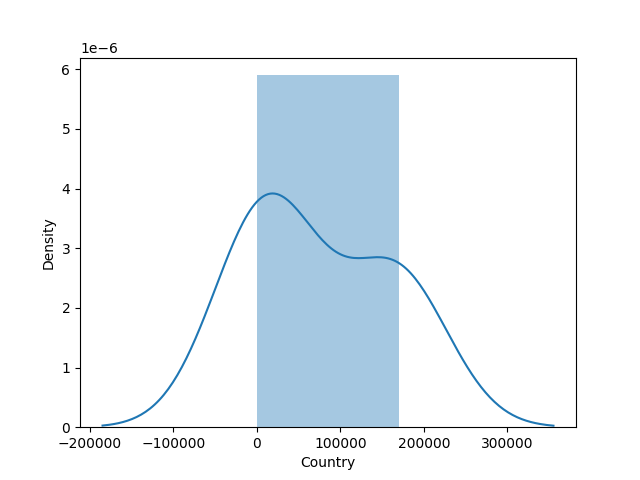
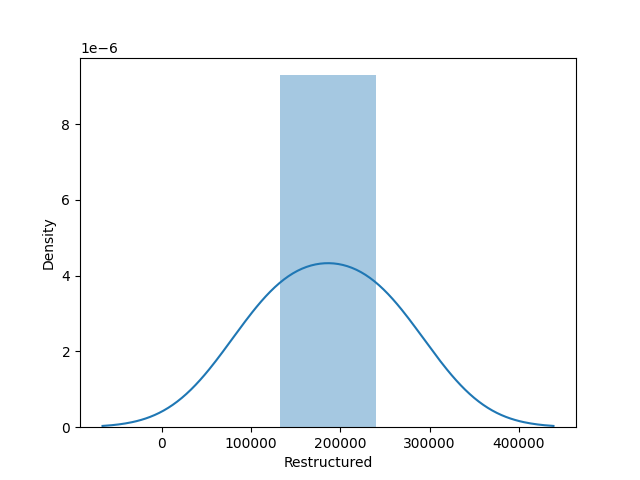
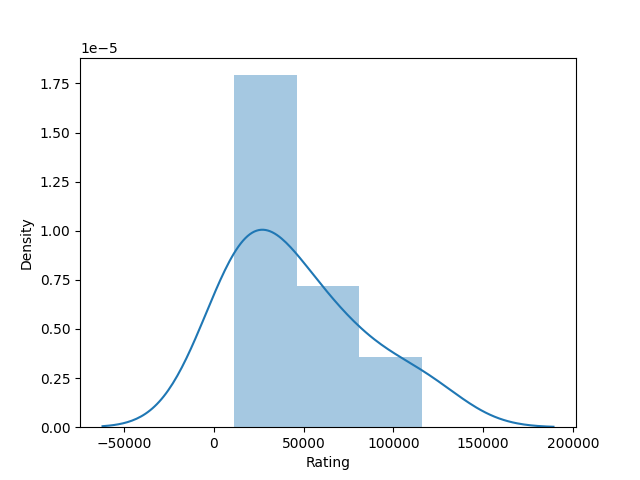
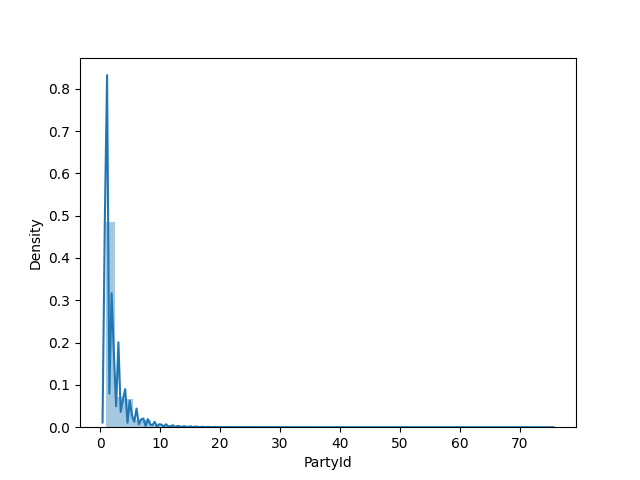
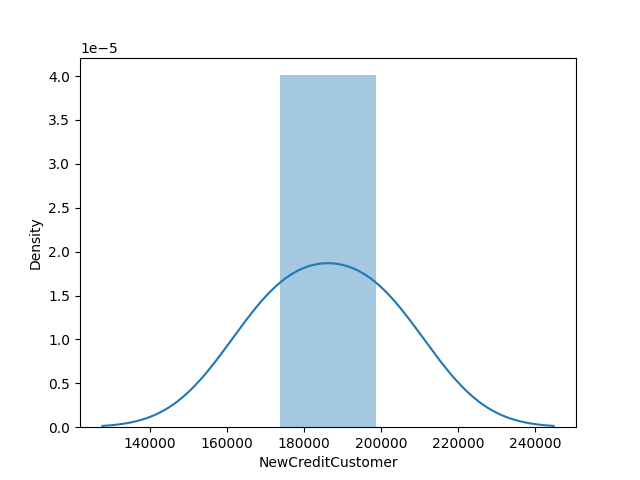
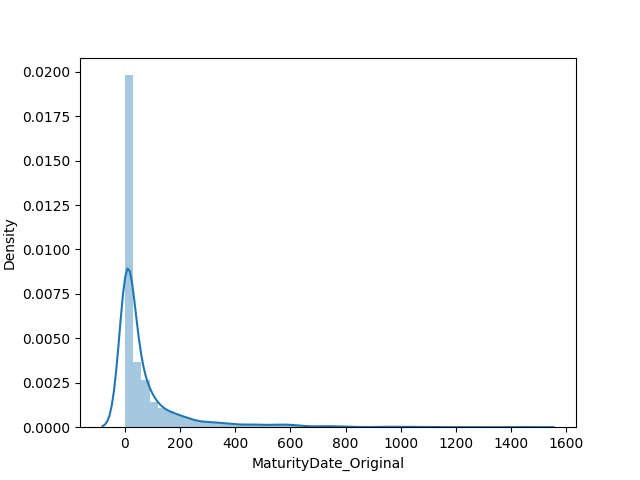
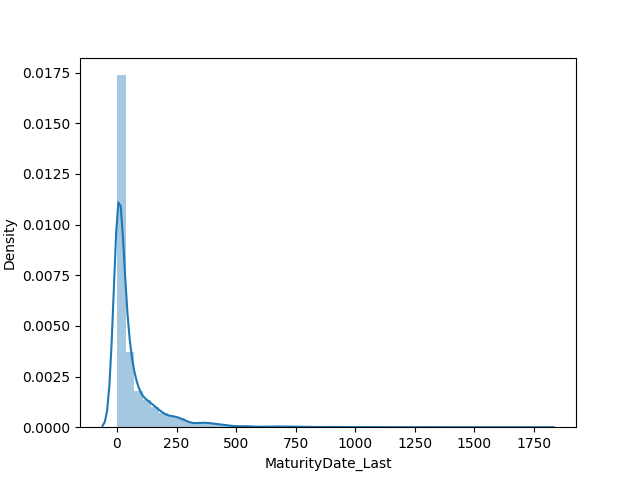
****

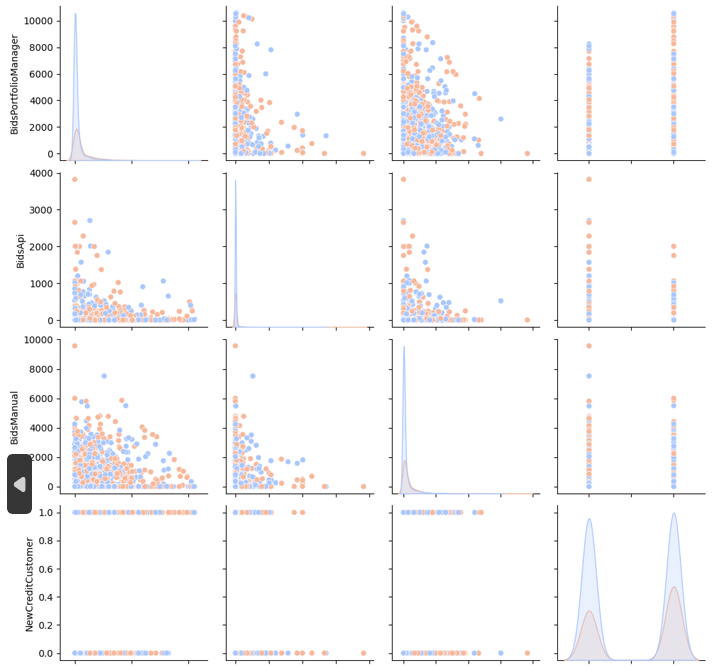
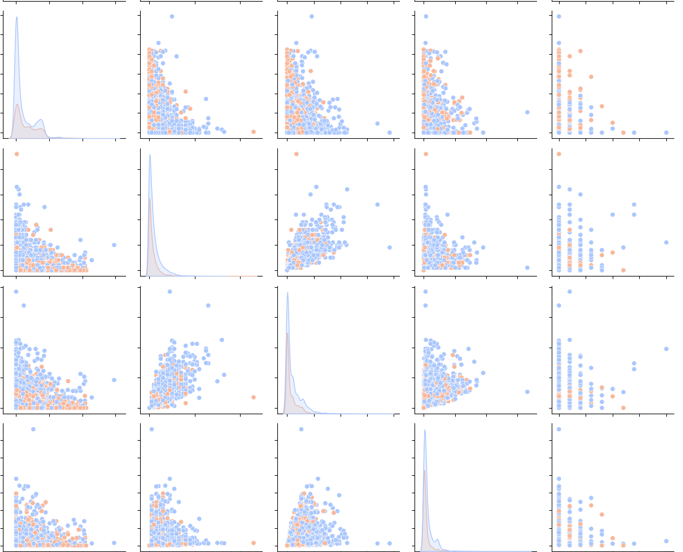
**Analysis and Solution of the Problem**

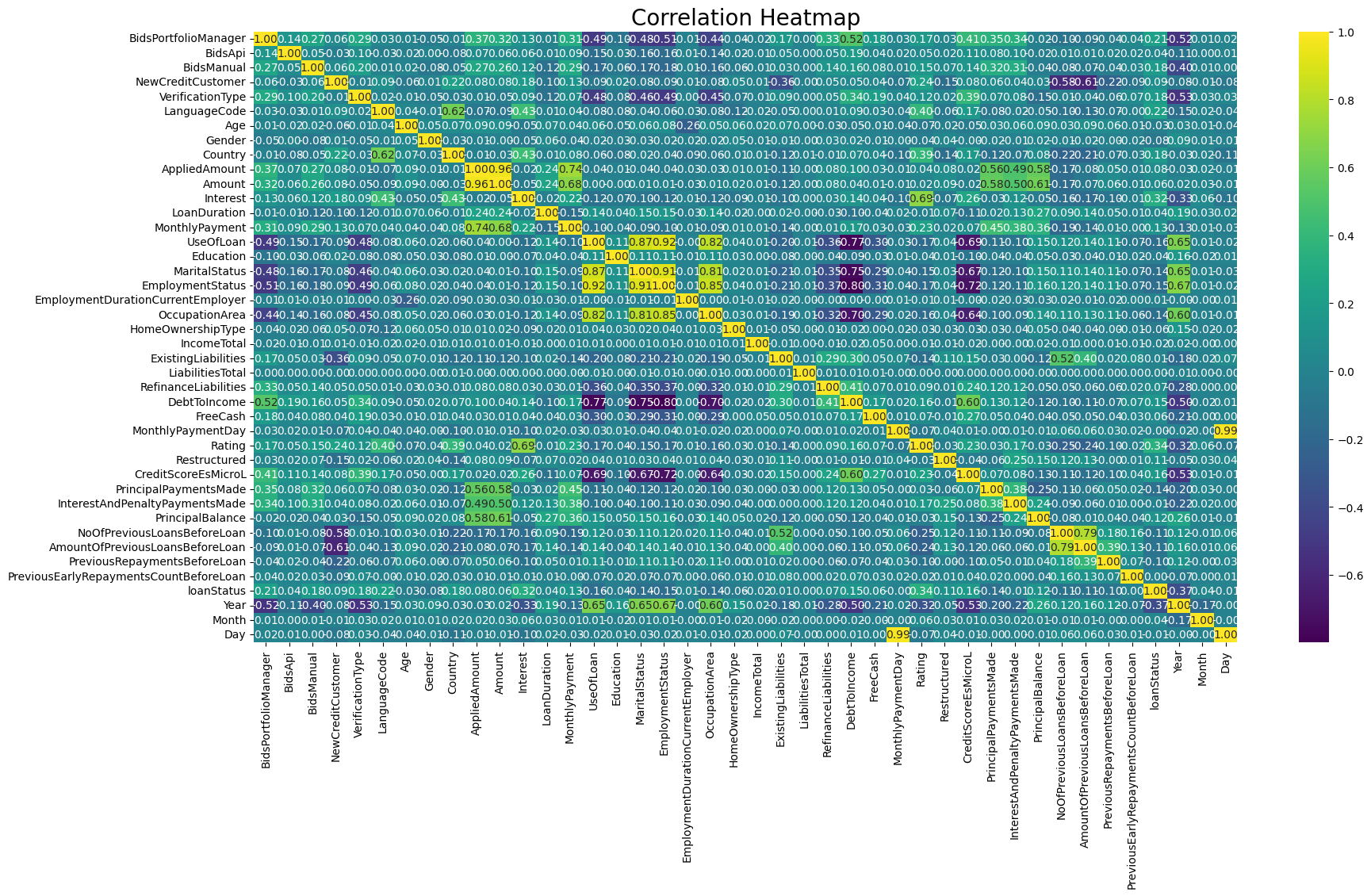
**in EDA1:**

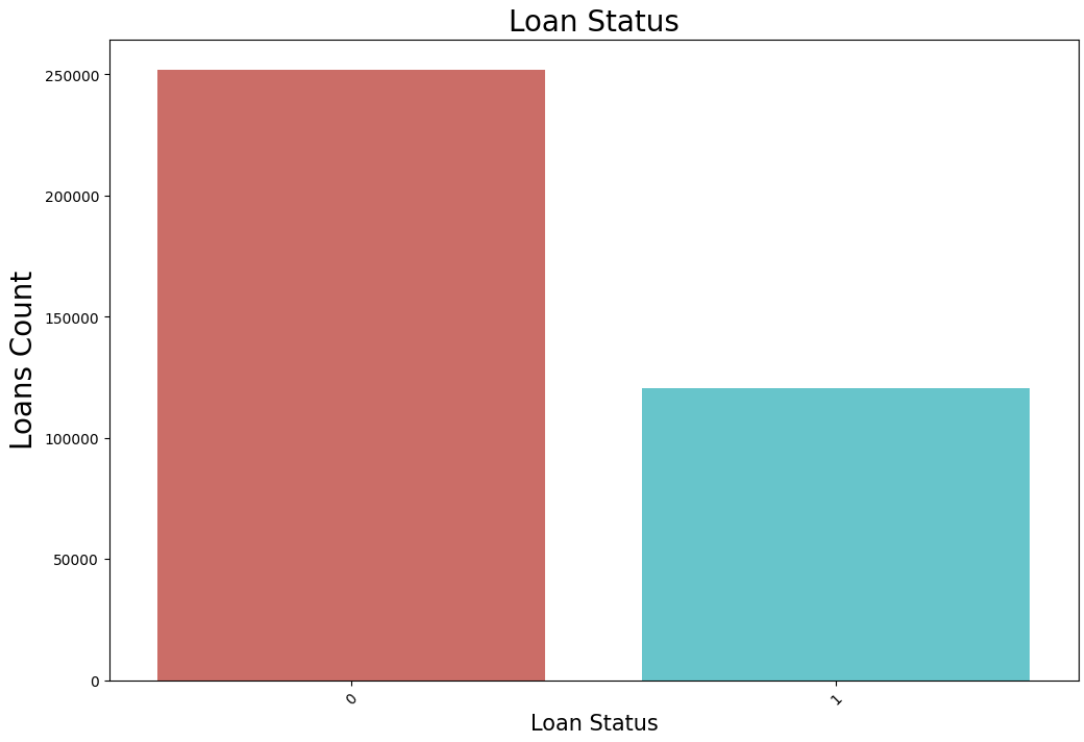
1. **Exploring null values:**
2. **Correlation among features**
3. **Checking distribution of each feature**

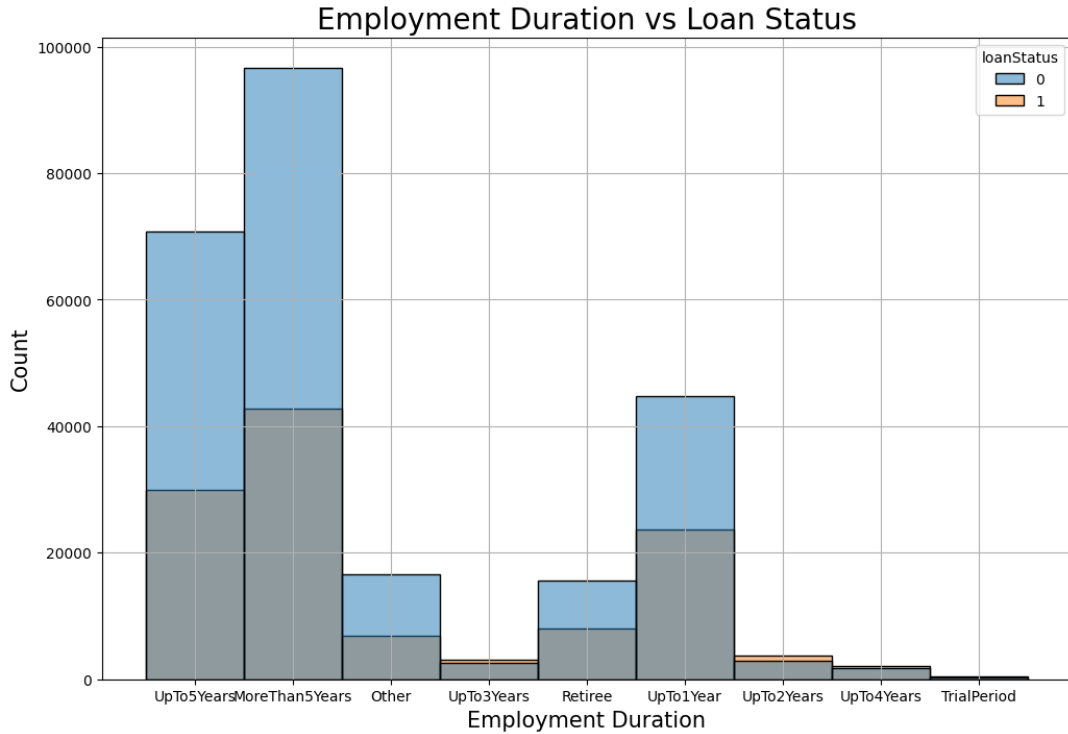
****



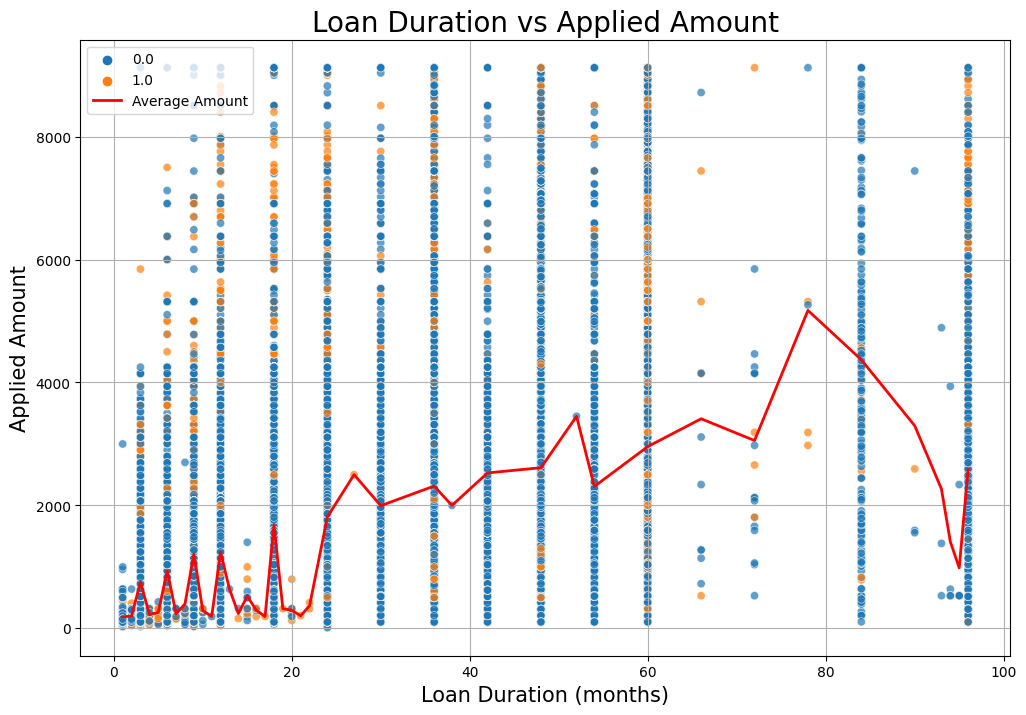
1. **Dataset preparation:**
   * Drop columns having more than 40% missing values except for some columns are kept for visualizations
   * Drop features that will have no rule in default predictions like (LoanId, LoanNumber, ApplicationSignedWeekday)
   * Create target variable (Current loan **Status** else not defaulted)
   * Handle Outliers in data
   * Restore labels for better visualizations
2. **Exploratory Data analysis Focusing mainly on visualizations**
   * ****Random snapshot of pair plot to show correlation of each feature with other

* **Correlation matrix:**
  + ****

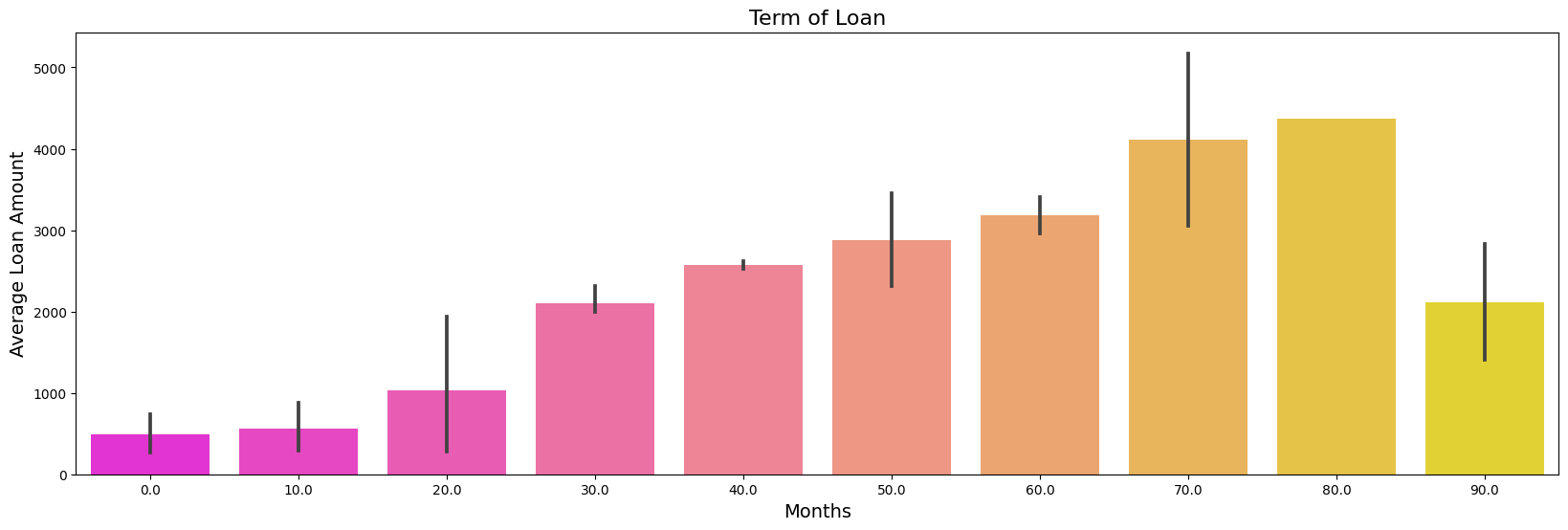


****

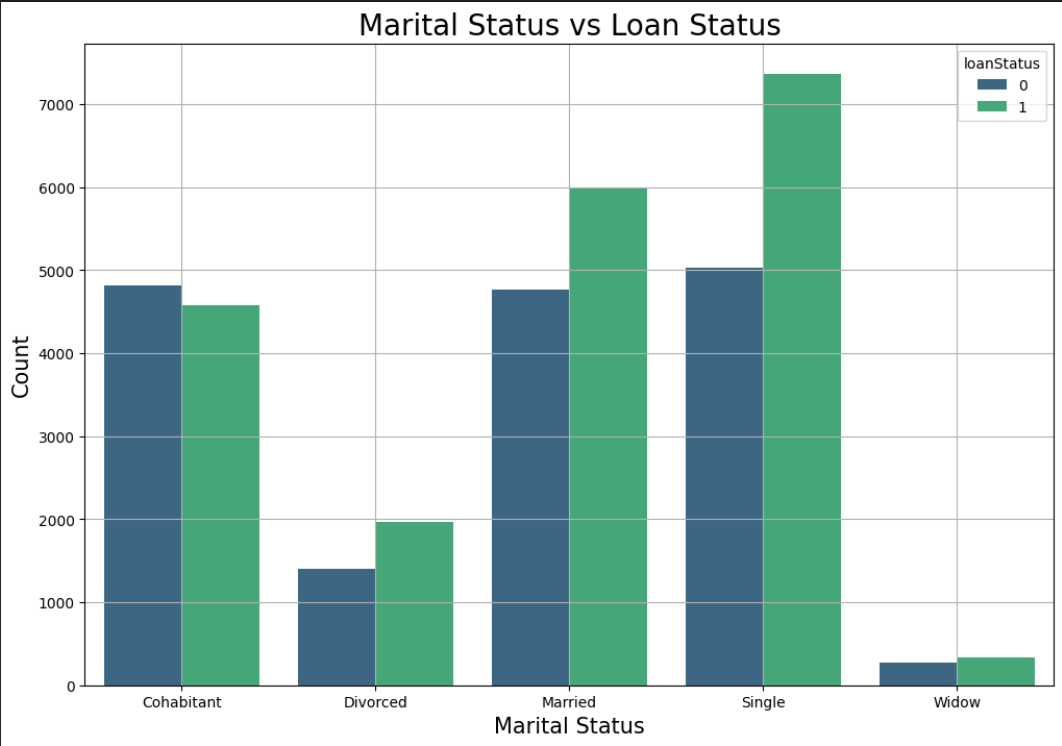
**more than 5 years employment have high acceptance among all employee status but also highest loans rejected while employees up to 3 years, 2 years up to 4 years employment the loan is most likely be defaulted (rejected) 🡪 you need 5 years of experience to have high probability of your loan be accepted**

****

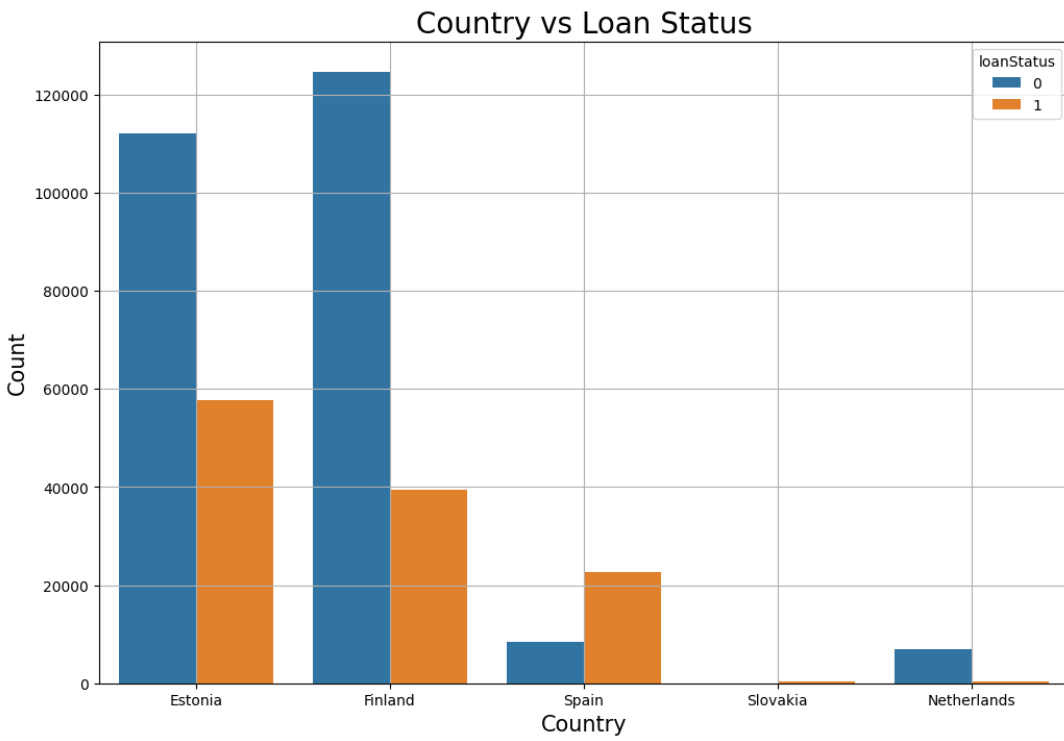
* **As Duration increases Applied amount also increases despite the fact that most loans are applied in the range 100-120 months**

****

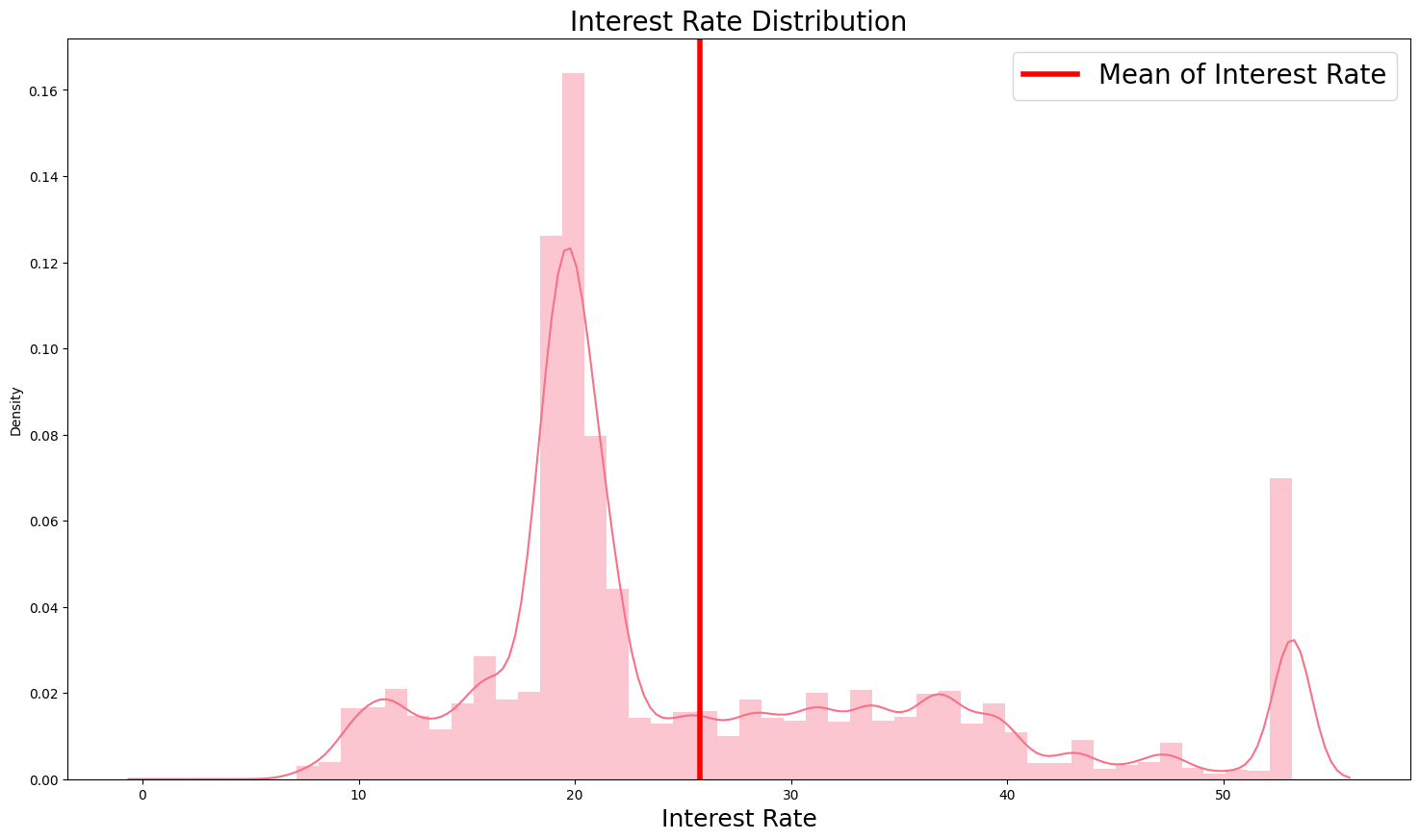
* **Large loan amounts are within range of 70-80 months duration**

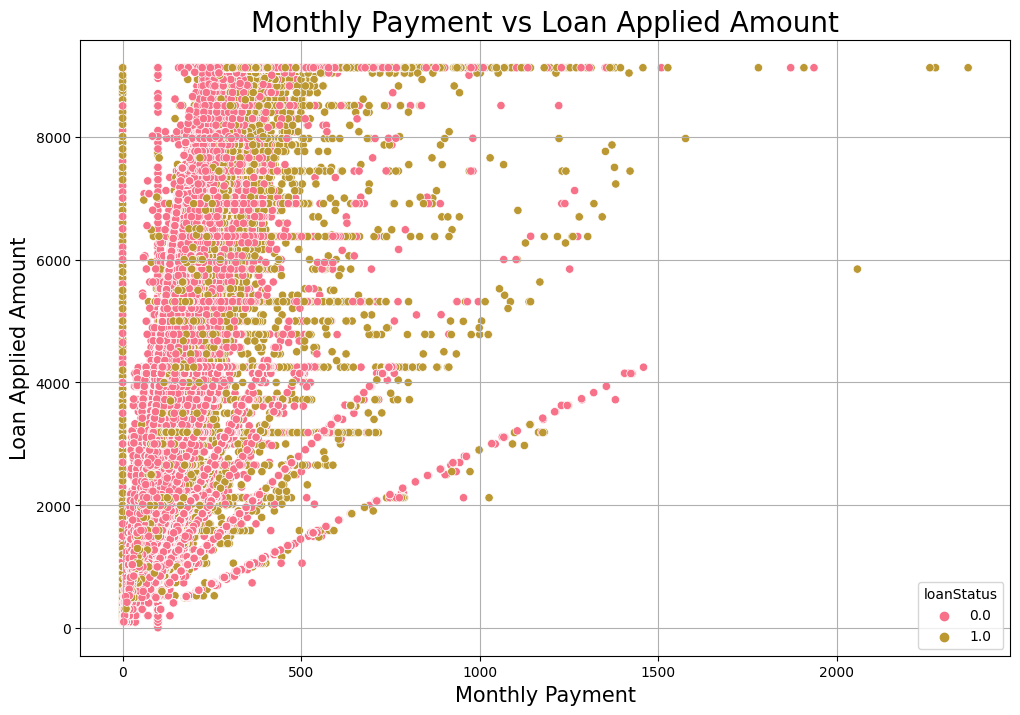
****

* **being single is more likely your loan will be rejected ☹**

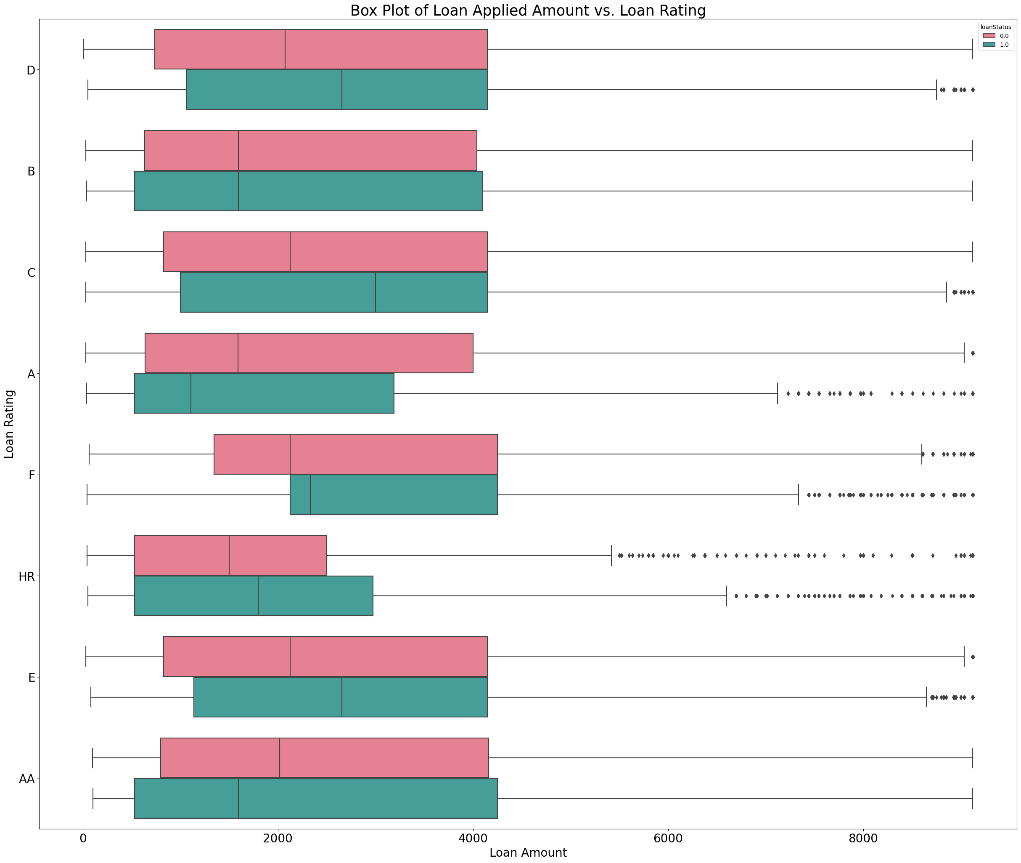
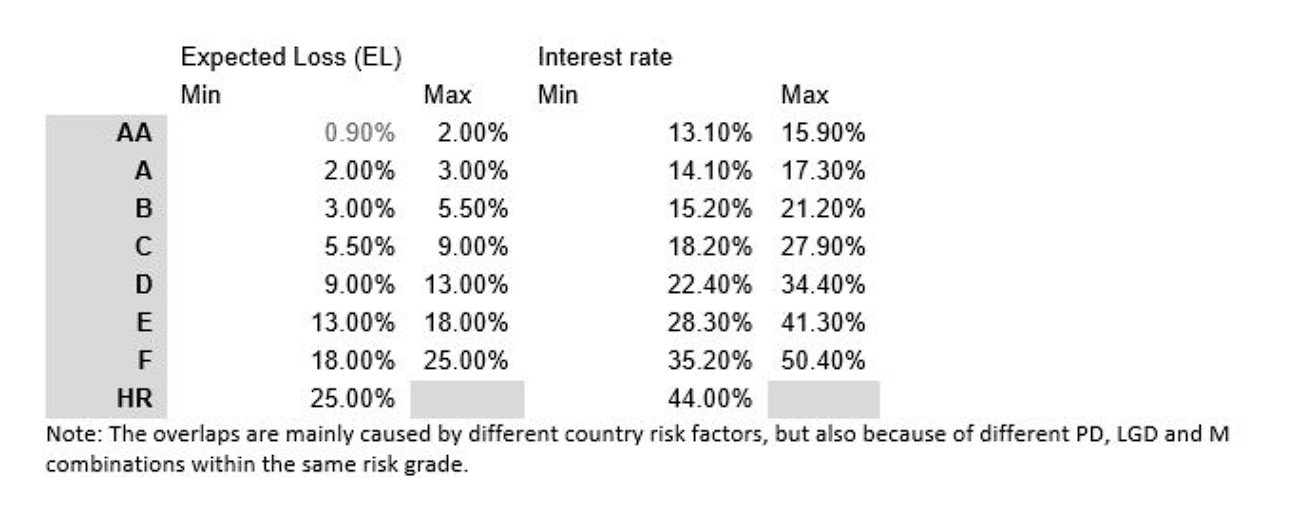
****

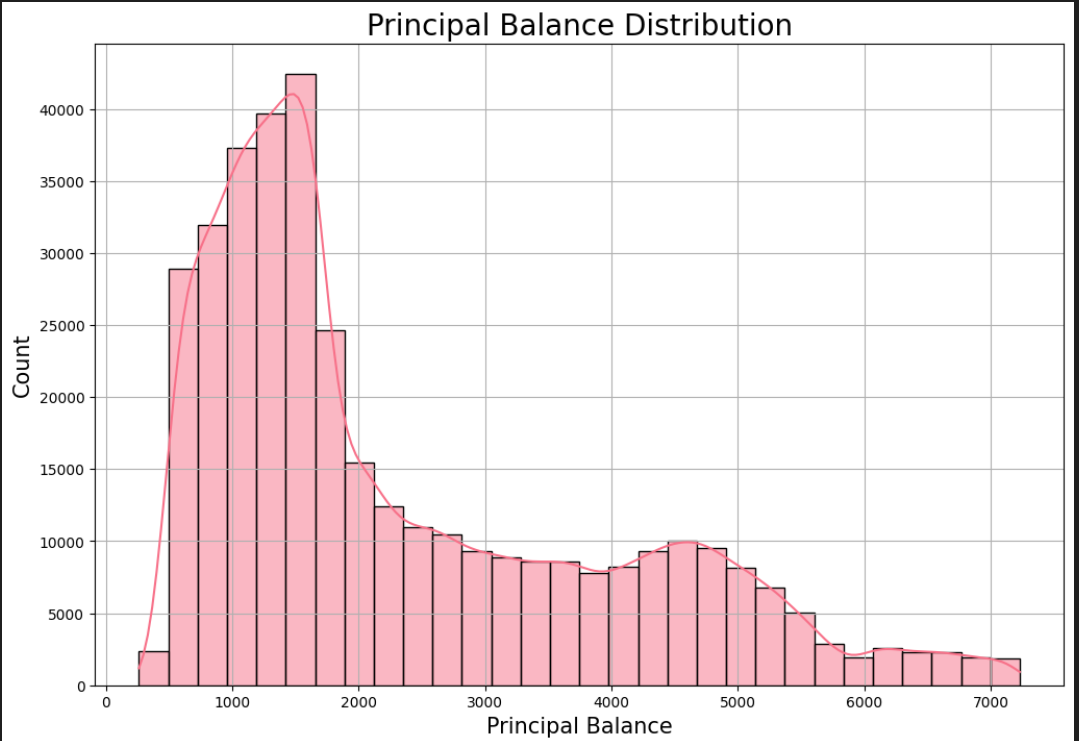
**Most loans in Estonia and Finland are accepted while in Spain most loans are rejected**

****

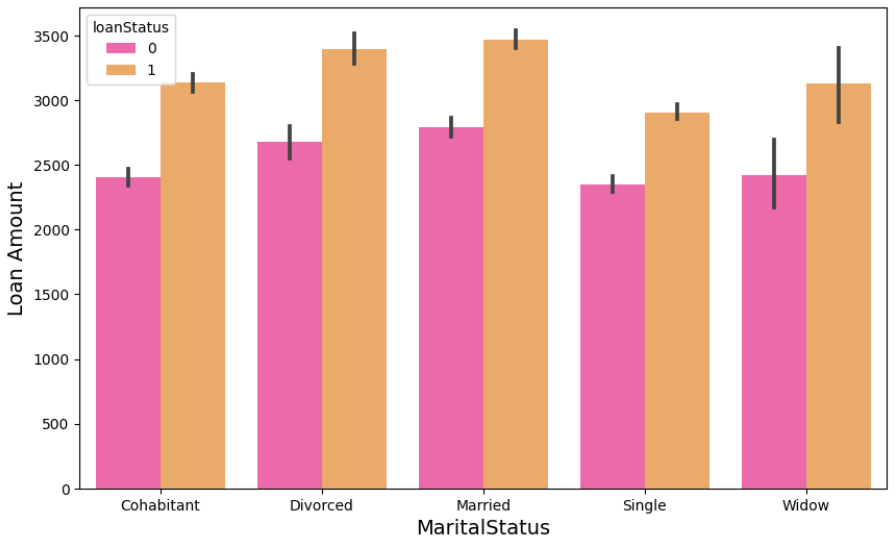
****

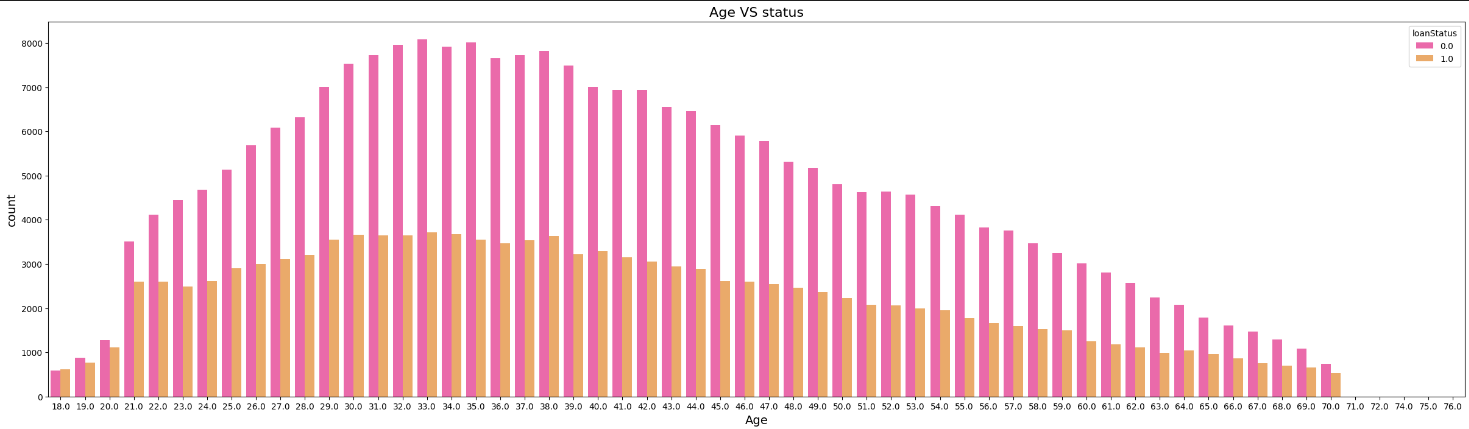
* **we can see here that as monthly payment increase with Lan applied amount obviously if you want a bigger loan you should pay more**

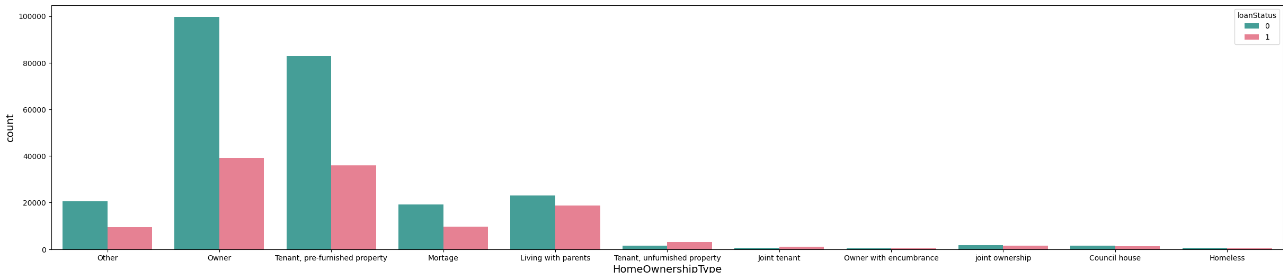
****

****

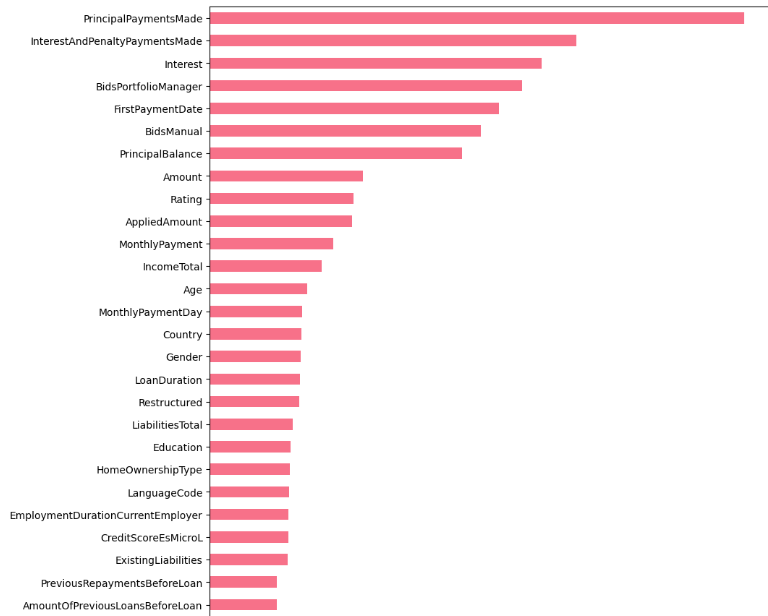
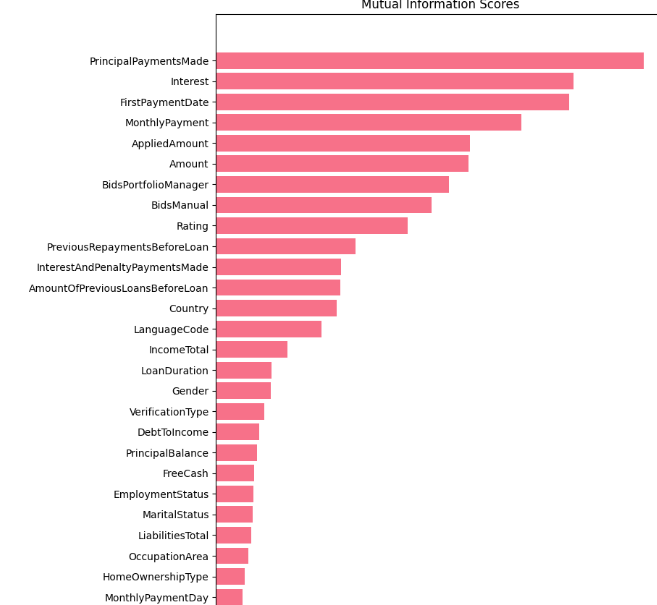
* **The principal balance in loans refers to the amount of money you originally borrowed, excluding any interest or fees that have been added on since the loan's inception. It represents the initial debt that you are obligated to repay to the lender**

****

* **Normal distribution as being in age 25-45 is high probably you will apply to a loan (starting a new business or something like that)**

****

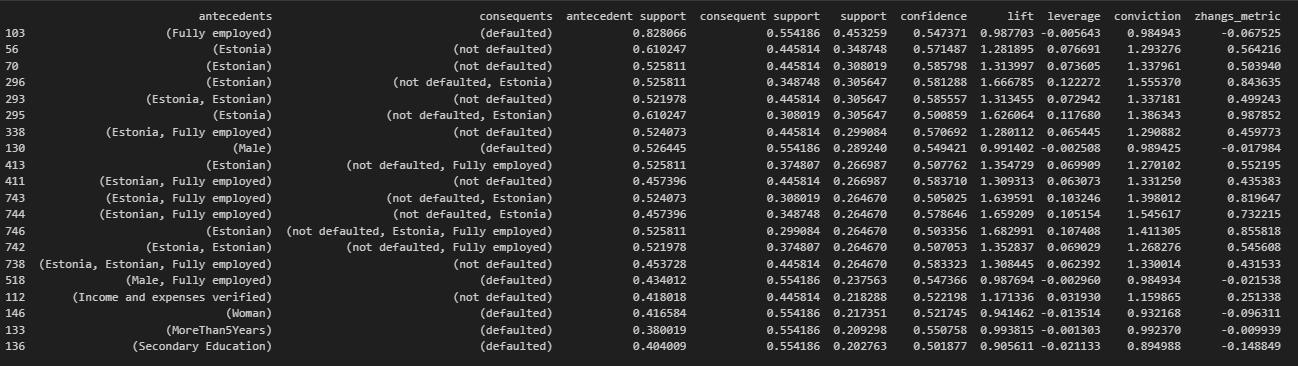
* **being an owner of house increases your chance of getting the loan also the people owning or tenanting an house are more likely to apply to a loan**

****

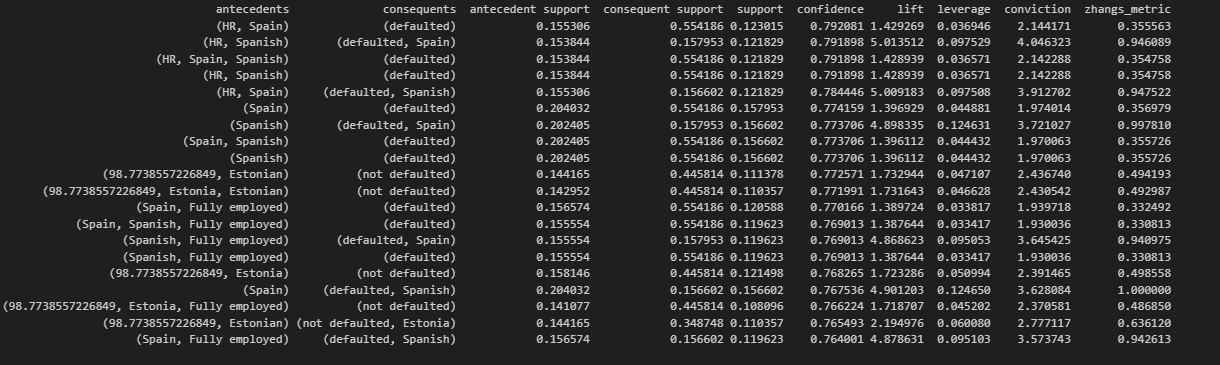
**Mutual information**

**Feature importance**

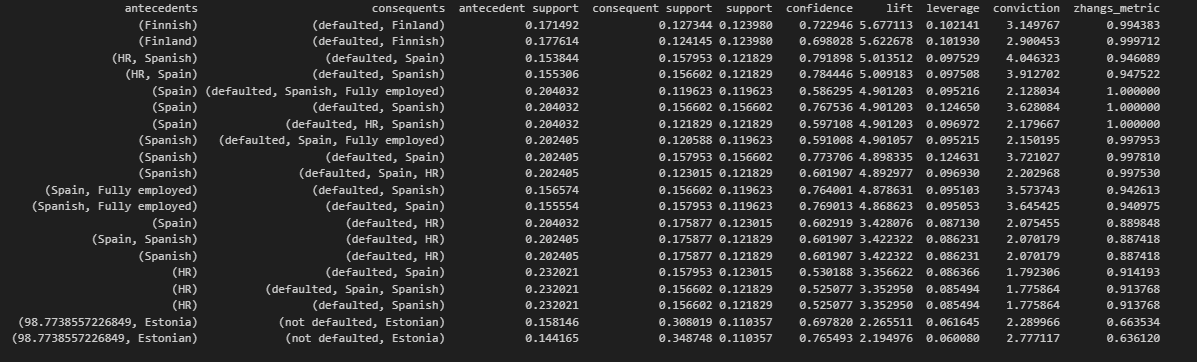
**Assisiation Rules:**

****

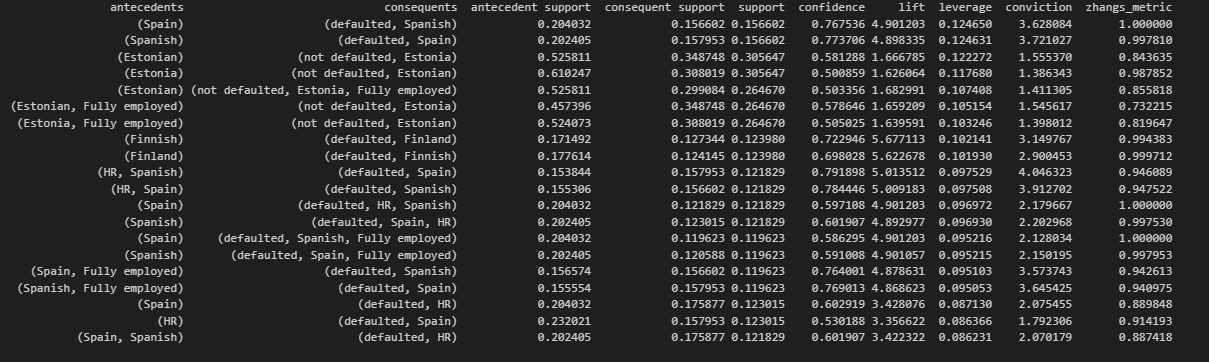
**Sorted by support**

****

**Sorted by confidence**

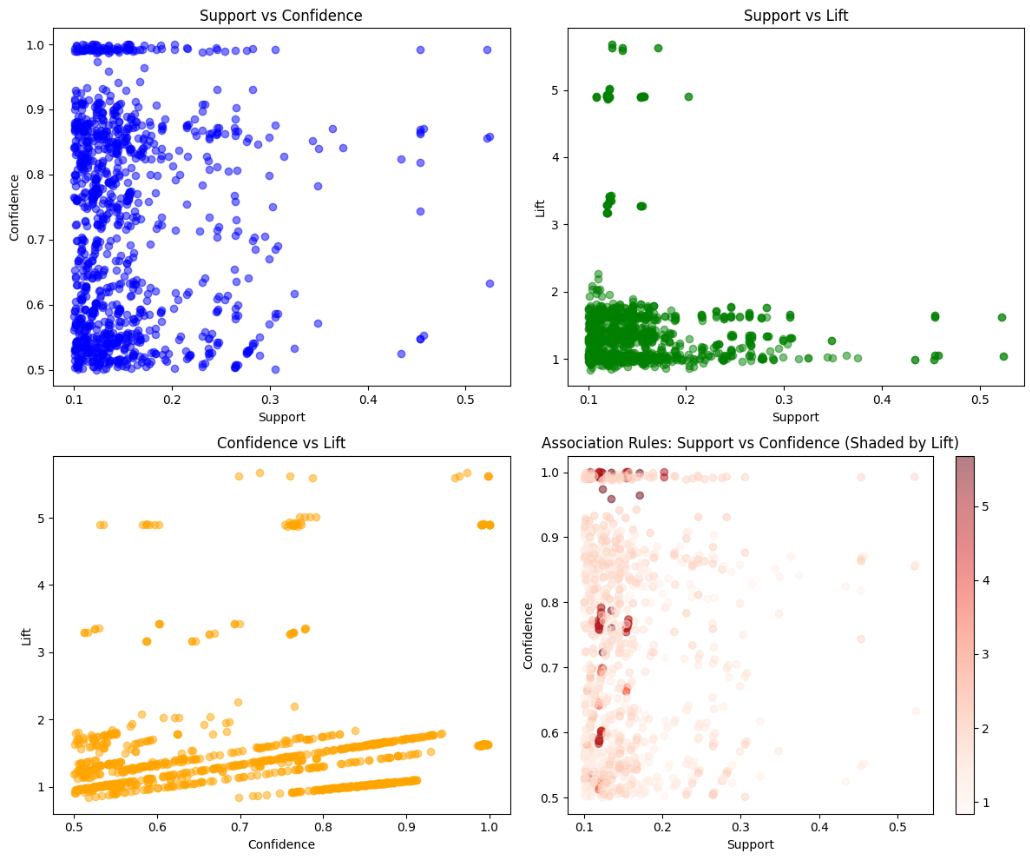
****

**Sorted by lift**

****

**Sorted by leverage**

From association rules we can notice that being a Spanish person and live in spain your loan has high lift to be defaulted but if you are Estonian and fully employed your loan have high leverage not to be defaulted, if your loan has a High Risk(HR) rating your loan is most probably defaulted.

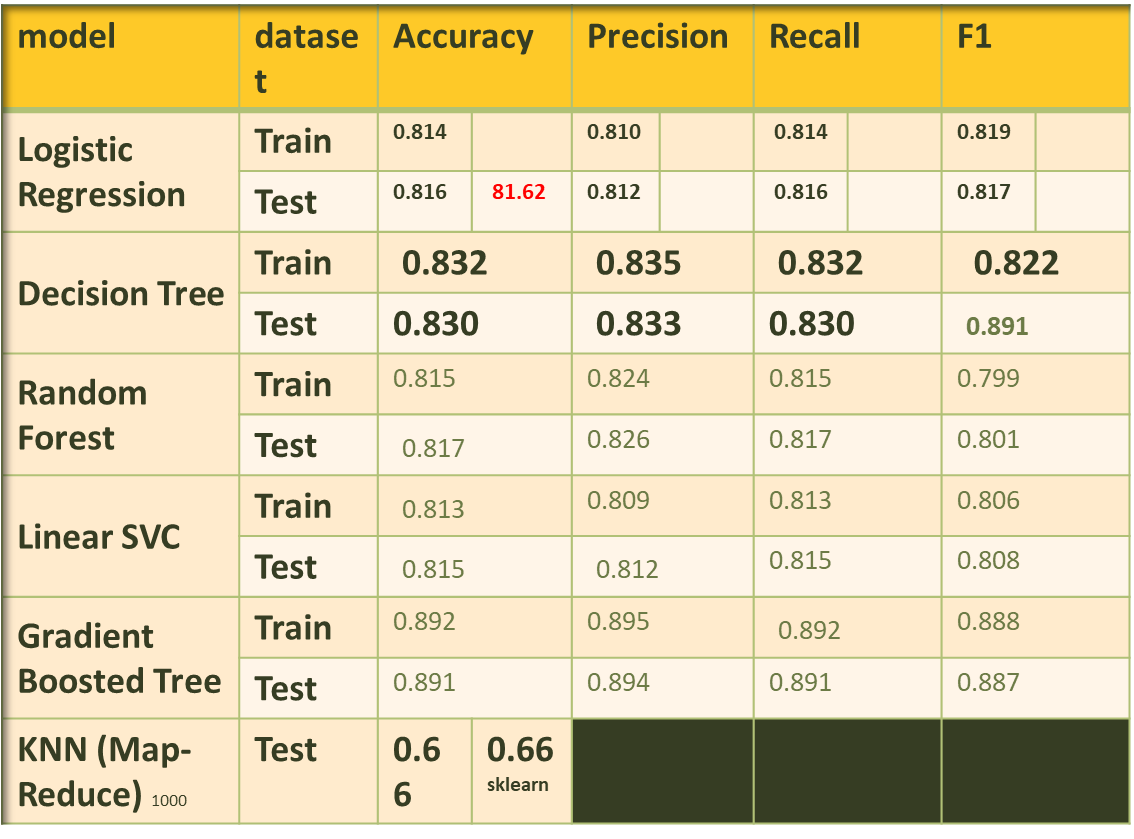
****

**Kmeans clustering**



* **Processing after EDA step:**
  + **Handel outliers**
  + **Impute missing values more than 5%**
  + **Drop columns with more than 40% null values**

**Model Training and Evaluation**

****

**Unsuccessful Trials:**

* **training with one hot encoding**
* **deploy in fully distributed mode**
* **Handling large data sizes with KNN**

**future work:**

* **Add more models**
* **Implement naïve bayes with MapReduce**
* **Add more visualizations**
* **Scale KNN algorithm**