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:

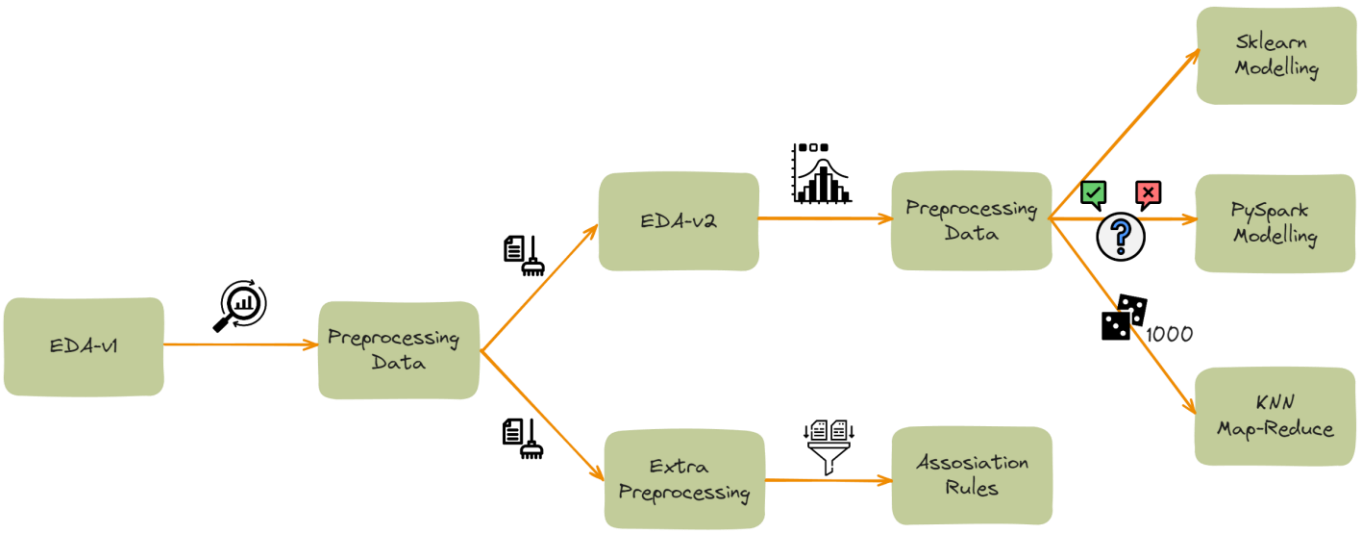
In this project we will be doing credit risk modelling of peer to peer lending Bondora systems, Credit risk modeling involves analyzing data to assess the likelihood that a borrower will default on a loan or fail to meet their financial obligations. Peer-to-peer lending platforms like Bondora facilitate 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 🔗:

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

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

Project Pipeline:

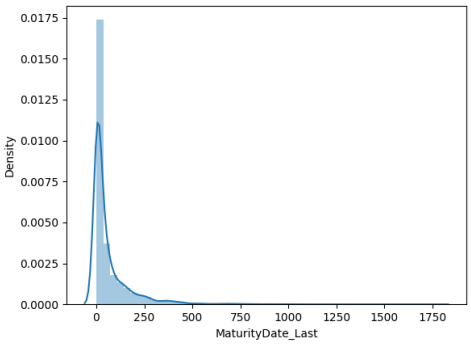


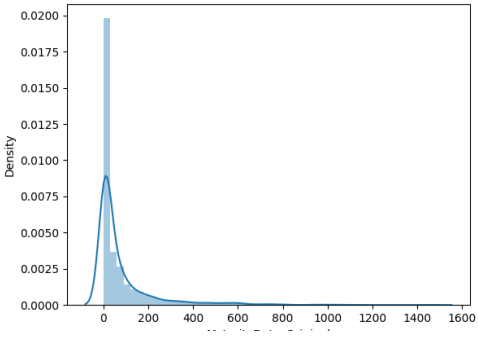
Analysis and Solution of the Problem

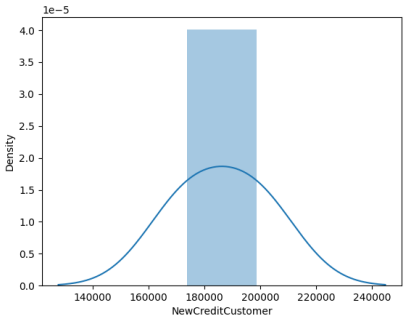
in EDA1:

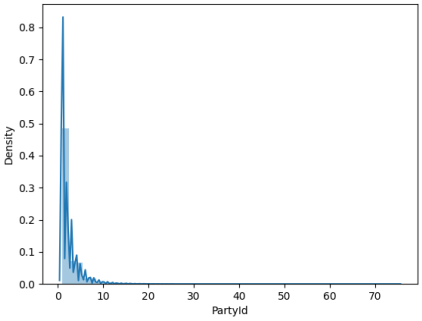
1. **Exploring null values:**
2. **Correlation among features**
3. **Checking distribution of each feature**

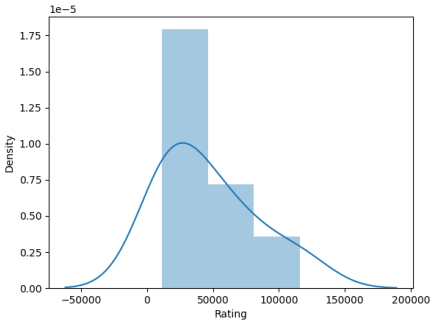
DebtToIncome	0.013421
FreeCash	0.013421
MonthlyPaymentDay	0.000000
ActiveScheduleFirstPaymentReached	0.000000
PlannedPrincipalTillDate	85.380401
PlannedInterestTillDate	1.246843
LastPaymentOn	2.651789
CurrentDebtDaysPrimary	63.977119
DebtOccuredOn	63.977119
CurrentDebtDaysSecondary	61.063346
DebtOccuredOnForSecondary	61.063346
ExpectedLoss	0.708378
LossGivenDefault	0.708378
ExpectedReturn	0.708378
ProbabilityOfDefault	0.708378
PrincipalOverdueBySchedule	3.987749
PlannedPrincipalPostDefault	67.622893
PlannedInterestPostDefault	67.622893
EAD1	67.623429
EAD2	67.623429
PrincipalRecovery	67.622893
InterestRecovery	67.622893
RecoveryStage	40.398775
StageActiveSince	36.699585
ModelVersion	0.708378
Rating	0.733611
EL_V0	98.773558
Rating_V0	98.773558
EL_V1	96.530852
Rating_V1	96.530852

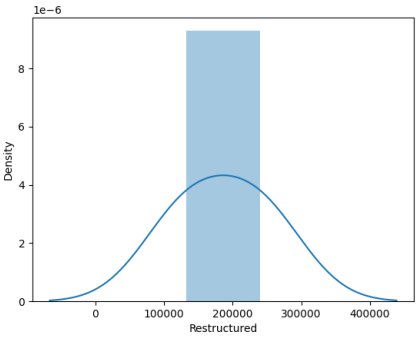

Density
MaturityDate_Last

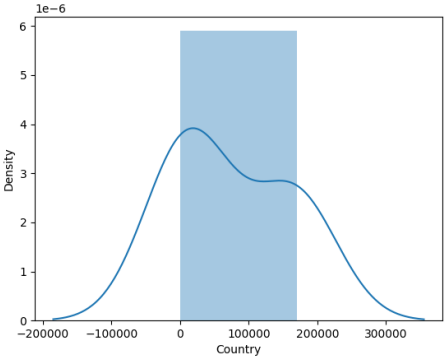

Density

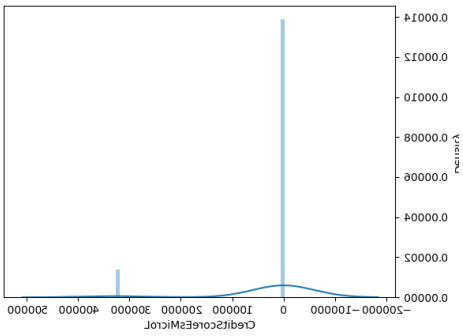

1e−5
Density
NewCreditCustomer

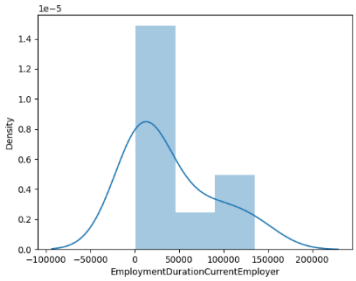

Density
PartyId

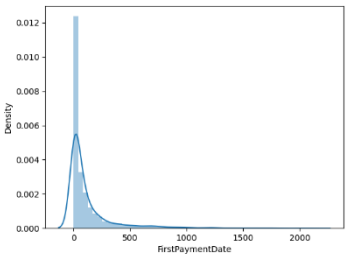

1e−5
Density
Rating

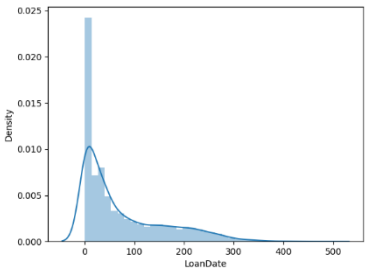

1e−6
Density
Restructured

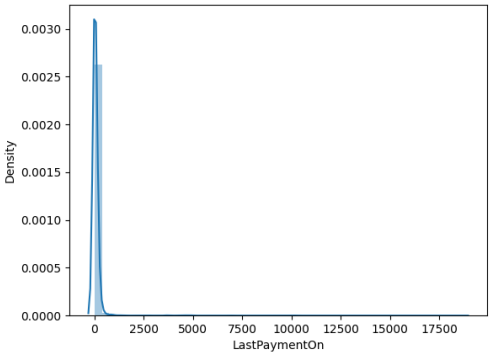

1e−6
Density
Country



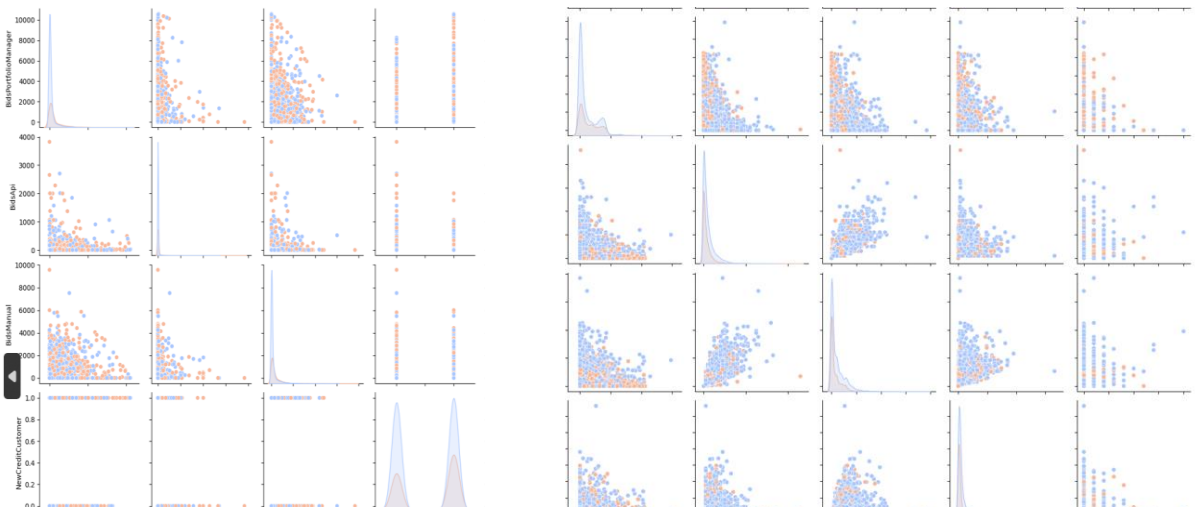




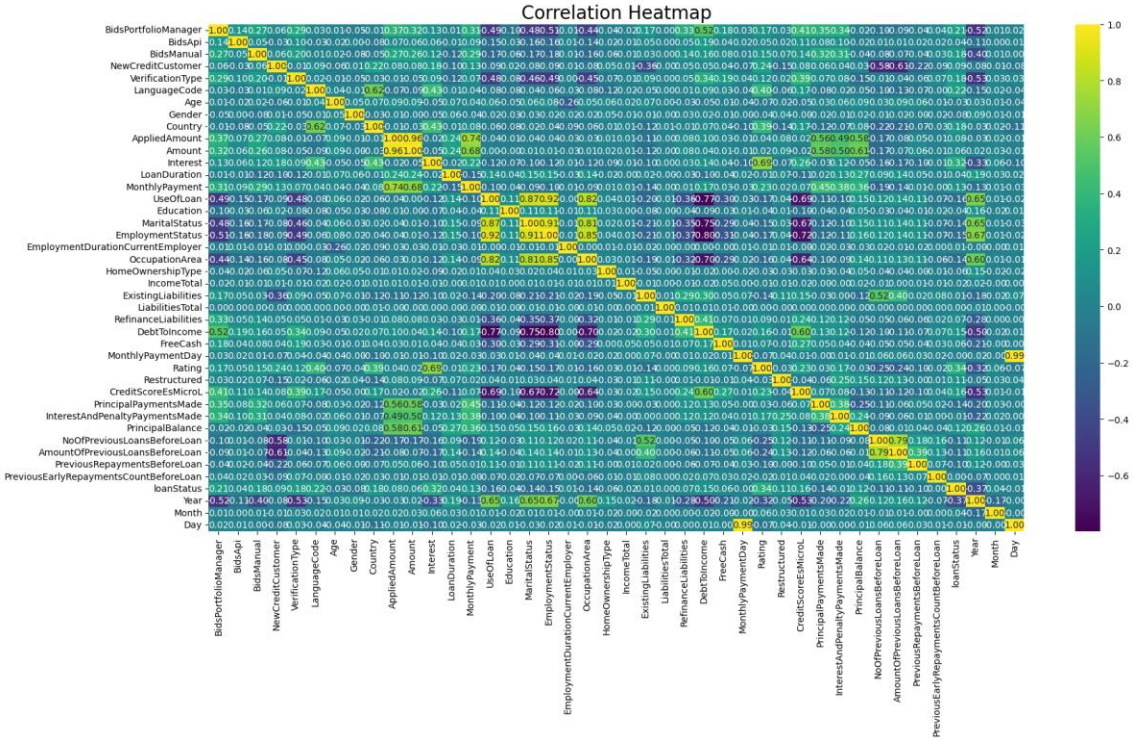


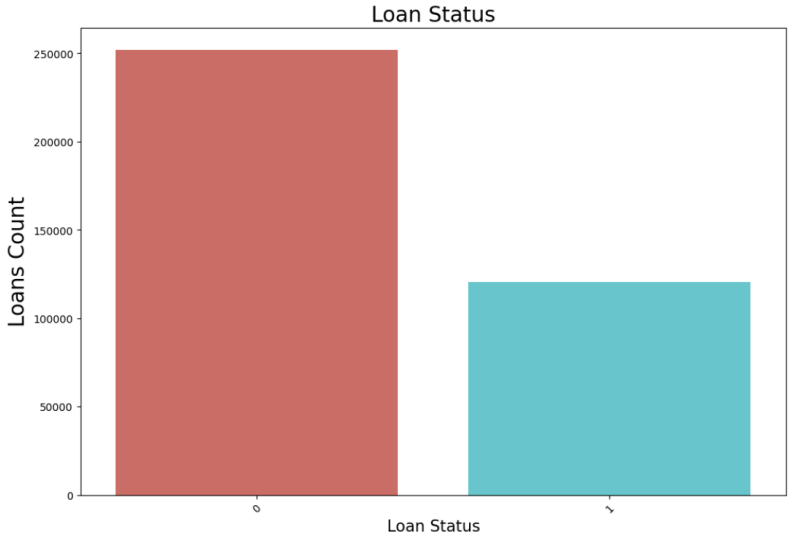


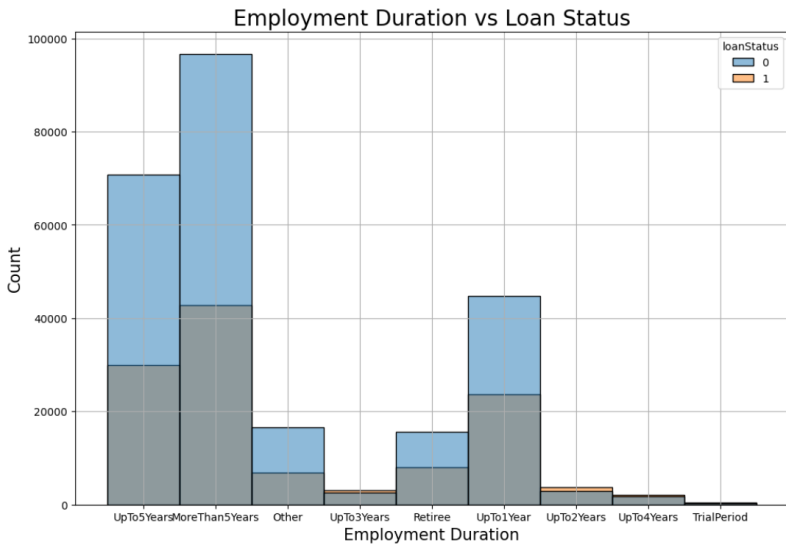
4. **Dataset preparation:**
    - Drop columns having more than 40% missing values except for some columns are kept for visualizations
    - Drop features that will have no rule in default predictions like (LoanId, LoanNumber, ApplicationSignedWeekday)
    - Create target variable (Current loan **Status**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



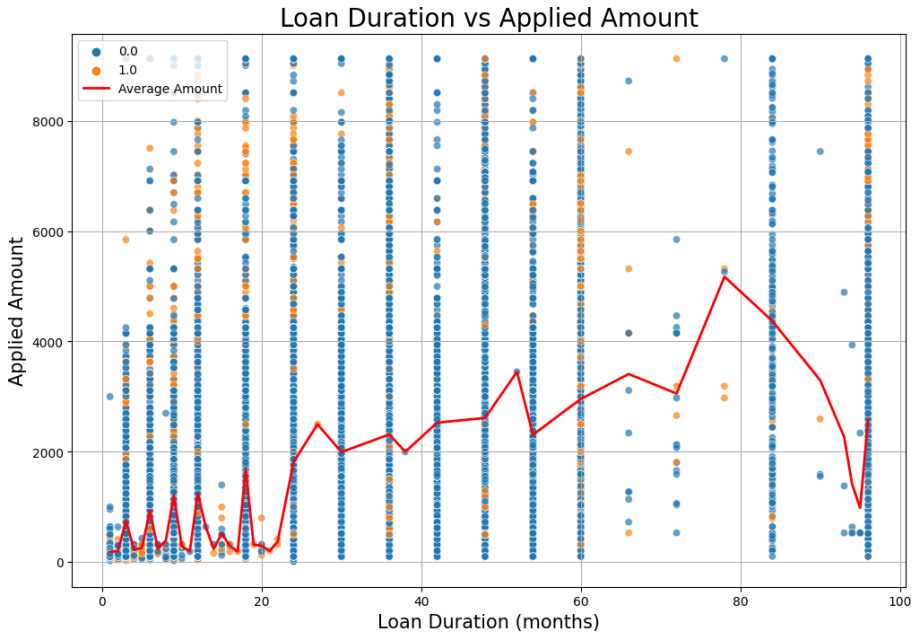
- **Correlation matrix:**



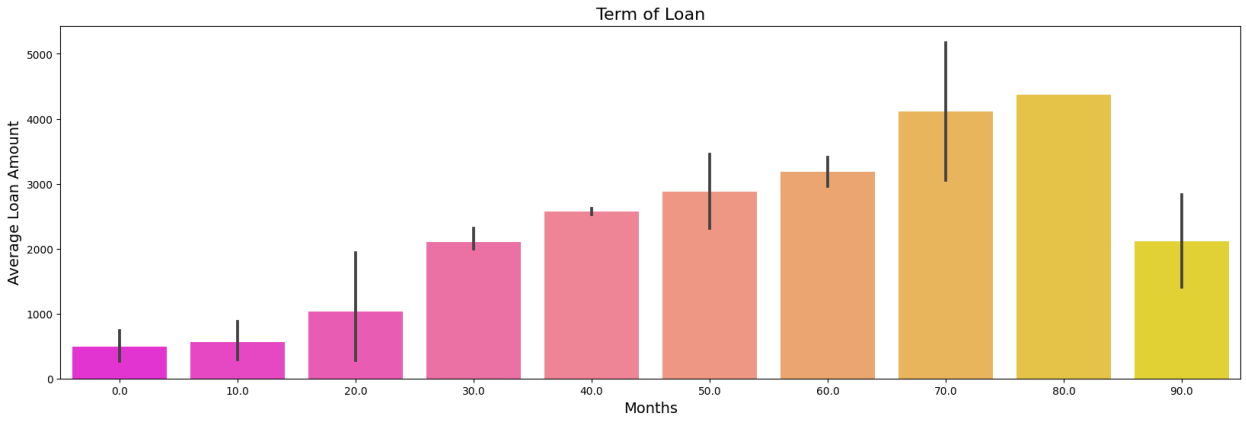




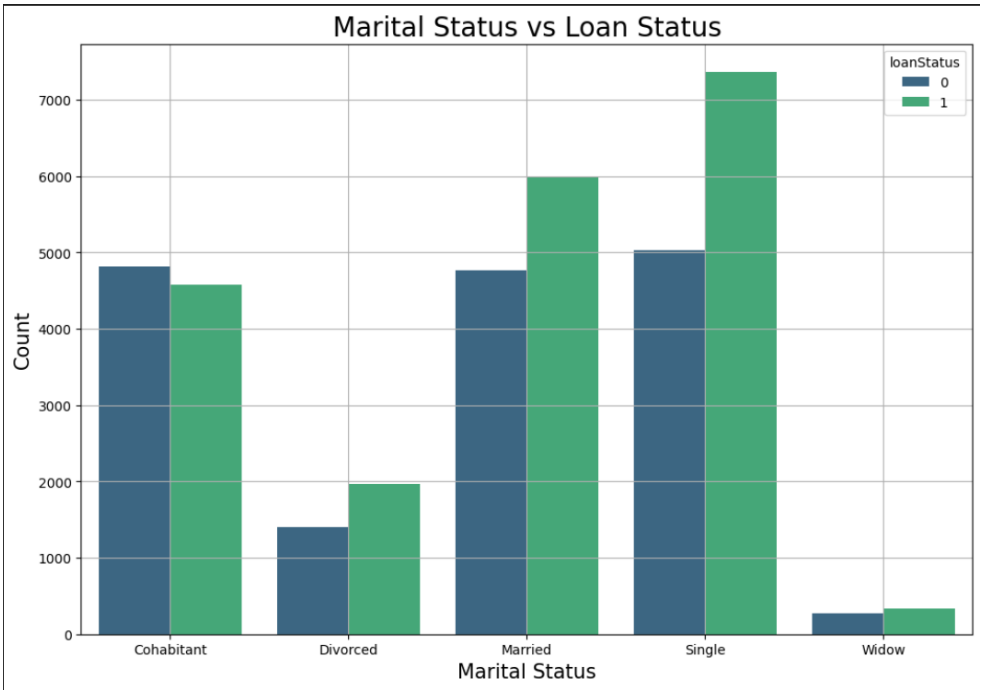
more than 5 years employment have high acceptance among all employee status but also highest loans rejected while employees up to 3 years, 2 years up to 4 years employment the loan is most likely be defaulted (rejected) → you need 5 years of experience to have high probability of your loan be accepted



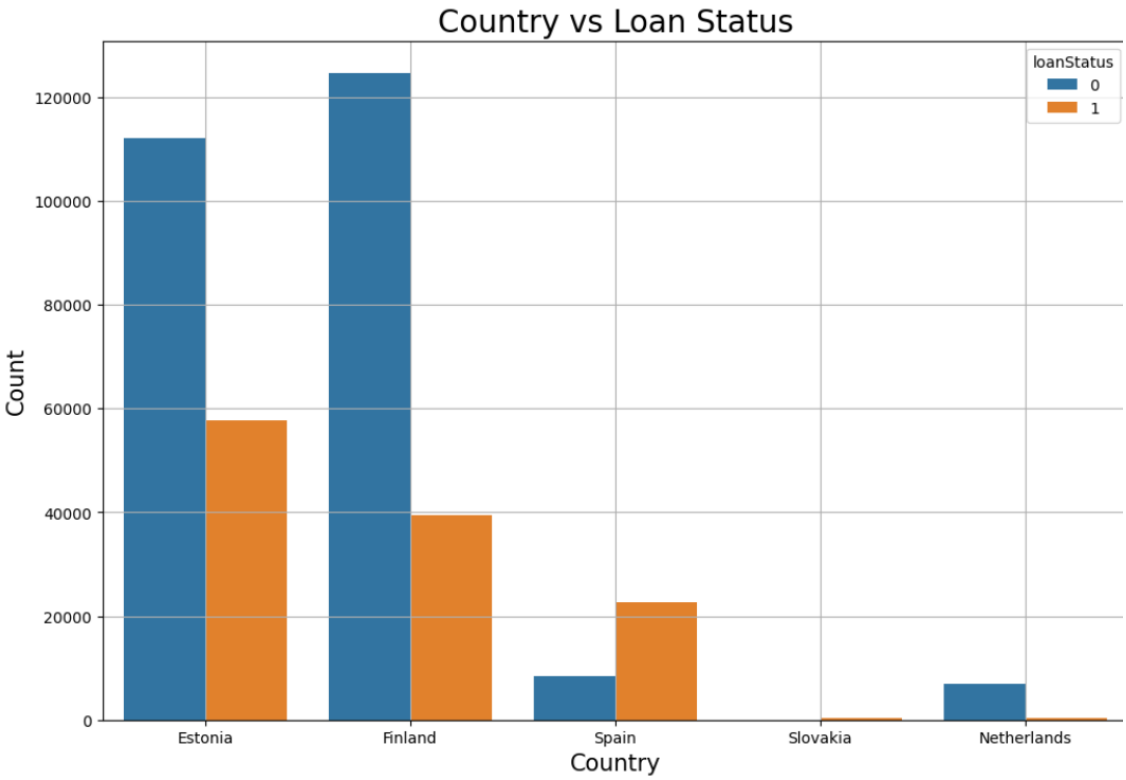
- **As Duration increases Applied amount also increases despite the fact that most loans are applied in the range 100-120 months**



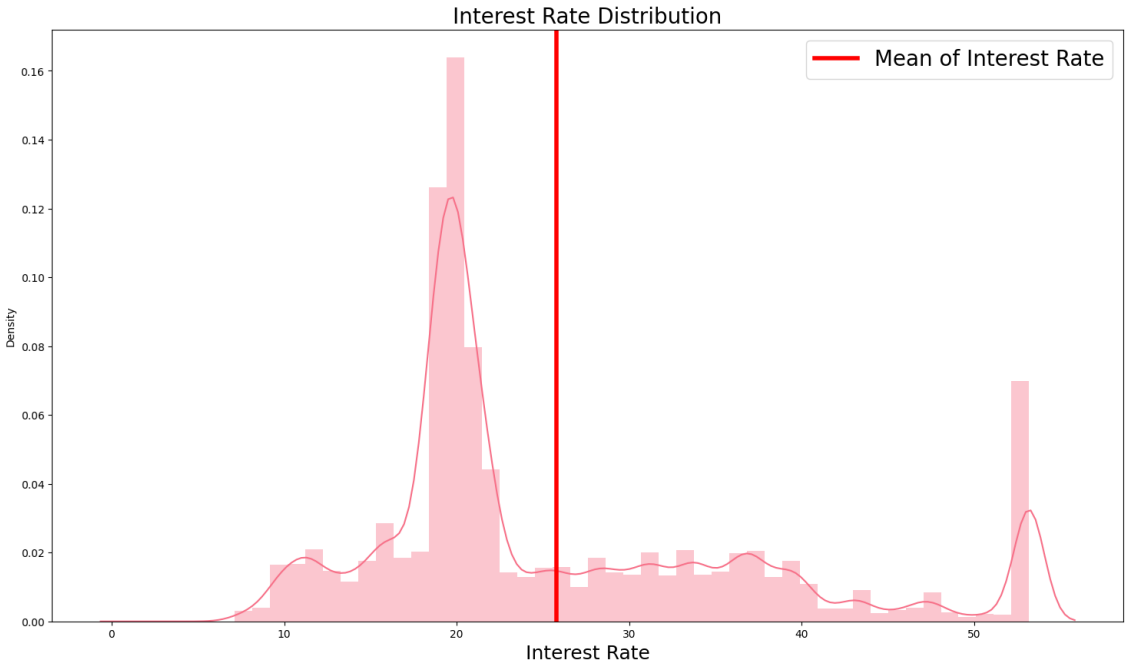
- **Large loan amounts are within range of 70-80 months duration**

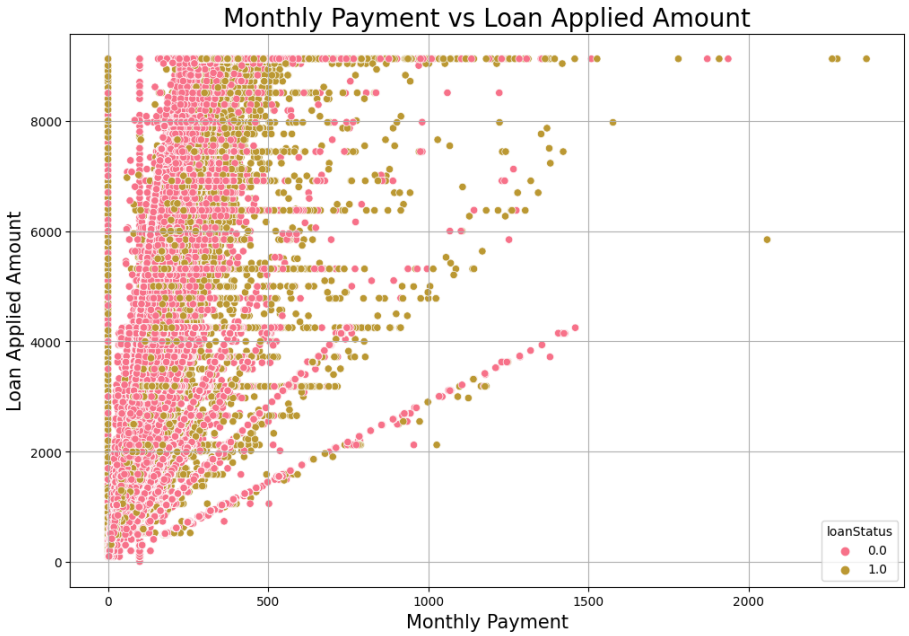


- **being single is more likely your loan will be rejected ☹**

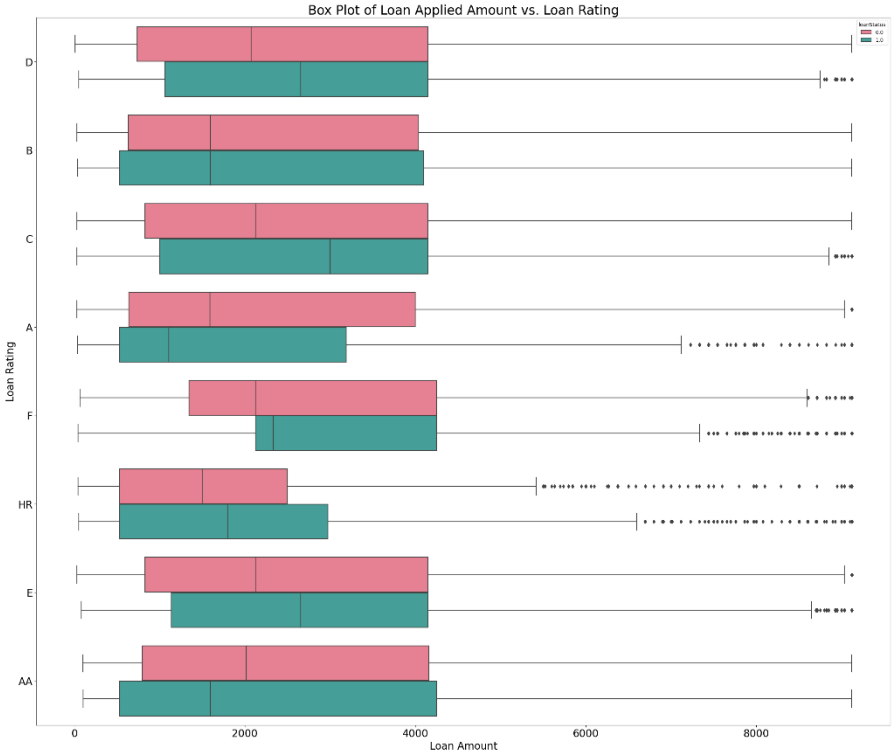


Most loans in Estonia and Finland are accepted while in Spain most loans are rejected



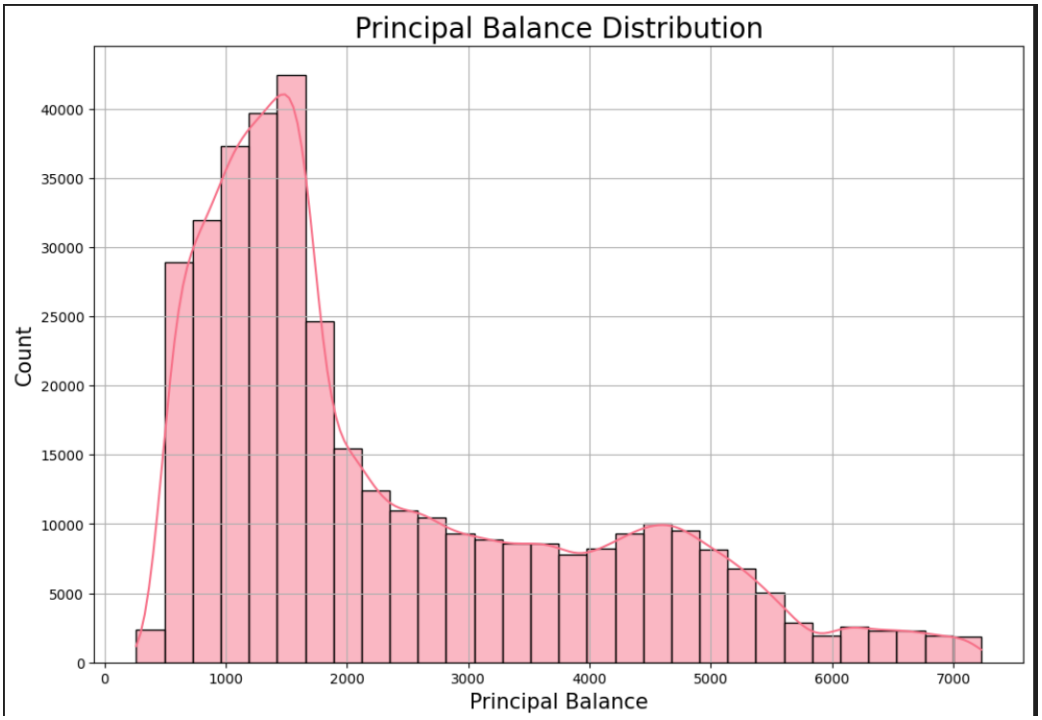


- **we can see here that as monthly payment increase with Lan applied amount obviously if you want a bigger loan you should pay more**

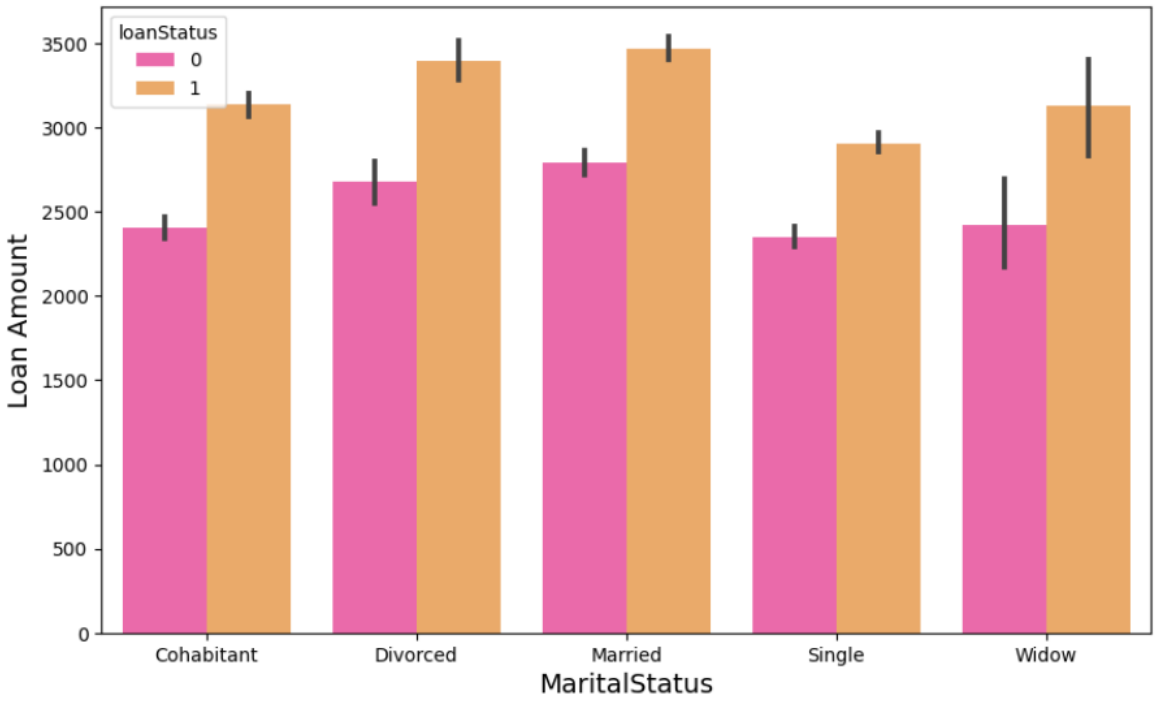


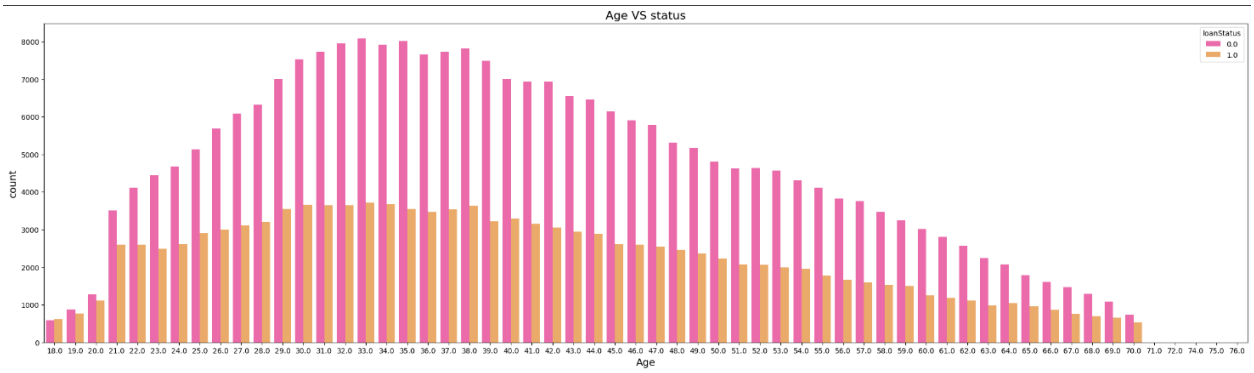
	Expected Loss (EL)		Interest rate	
	Min	Max	Min	Max
AA	0.90%	2.00%	13.10%	15.90%
A	2.00%	3.00%	14.10%	17.30%
B	3.00%	5.50%	15.20%	21.20%
C	5.50%	9.00%	18.20%	27.90%
D	9.00%	13.00%	22.40%	34.40%
E	13.00%	18.00%	28.30%	41.30%
F	18.00%	25.00%	35.20%	50.40%
HR	25.00%		44.00%	

Note: The overlaps are mainly caused by different country risk factors, but also becau
combinations within the same risk grade.

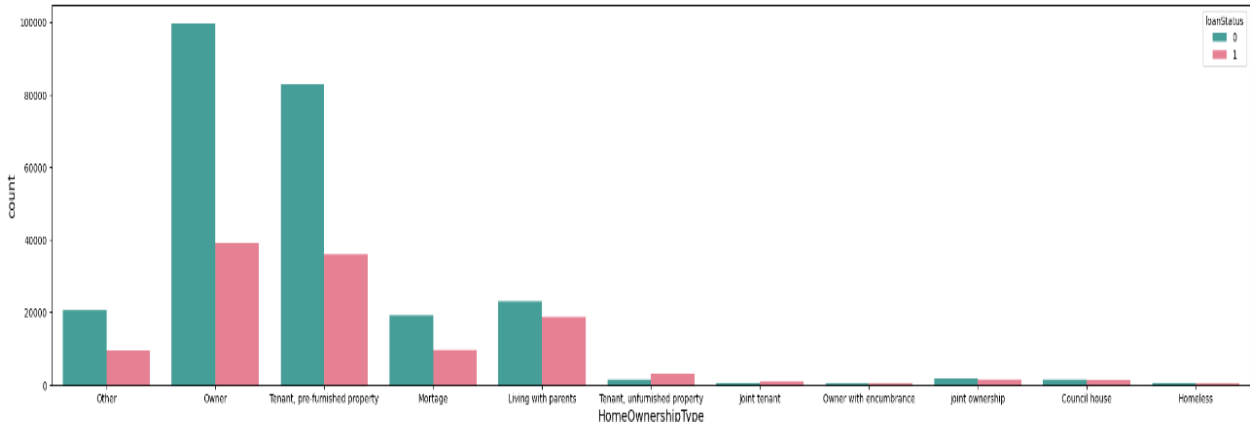


- **The principal balance in loans refers to the amount of money you originally borrowed, excluding any interest or fees that have been added on since the loan's inception. It represents the initial debt that you are obligated to repay to the lender**

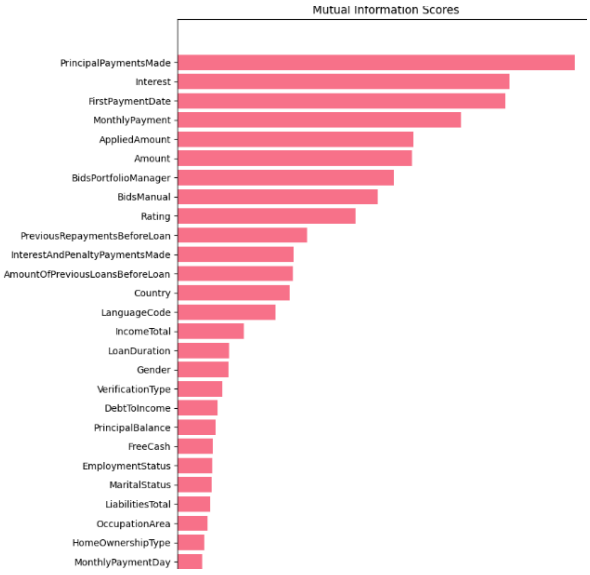


- 

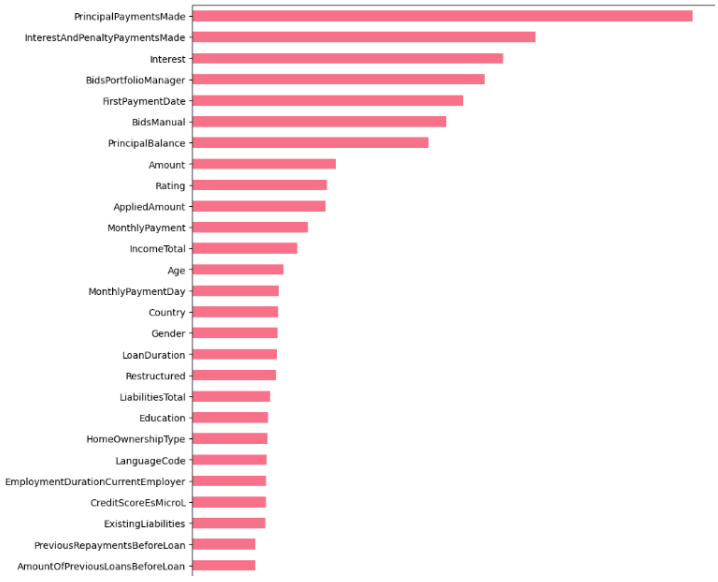
Normal distribution as being in age 25-45 is high probably you will apply to a loan (starting a new business or something like that)



- **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:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
103	(Fully employed)	(defaulted)	0.828066	0.554186	0.453259	0.547371	0.987703	-0.005643	0.984943	-0.067525
56	(Estonia)	(not defaulted)	0.610247	0.445814	0.348748	0.571487	1.281895	0.076691	1.293276	0.564216
70	(Estonian)	(not defaulted)	0.525811	0.445814	0.308819	0.585798	1.313997	0.073605	1.337961	0.503940
296	(Estonian)	(not defaulted, Estonia)	0.525811	0.348748	0.305647	0.581288	1.666785	0.122272	1.555370	0.843635
293	(Estonia, Estonian)	(not defaulted)	0.521978	0.445814	0.305647	0.585557	1.313455	0.072942	1.337181	0.499243
295	(Estonia)	(not defaulted, Estonian)	0.610247	0.308819	0.305647	0.500859	1.626064	0.117680	1.386343	0.987852
338	(Estonia, Fully employed)	(not defaulted)	0.524073	0.445814	0.299084	0.570692	1.280112	0.065445	1.290882	0.459773
130	(Male)	(defaulted)	0.526445	0.554186	0.289240	0.549421	0.991402	-0.002508	0.989425	-0.017984
413	(Estonian)	(not defaulted, Fully employed)	0.525811	0.374807	0.266987	0.507762	1.354729	0.069909	1.270102	0.552195
411	(Estonian, Fully employed)	(not defaulted)	0.457396	0.445814	0.266987	0.583710	1.309313	0.063073	1.331250	0.435383
743	(Estonia, Fully employed)	(not defaulted, Estonian)	0.524073	0.308819	0.264670	0.505025	1.639591	0.103246	1.398012	0.819647
744	(Estonian, Fully employed)	(not defaulted, Estonia)	0.457396	0.348748	0.264670	0.578646	1.659209	0.105154	1.545617	0.732215
746	(Estonian)	(not defaulted, Estonia, Fully employed)	0.525811	0.299084	0.264670	0.503356	1.682991	0.107408	1.411305	0.855818
742	(Estonia, Estonian)	(not defaulted, Fully employed)	0.521978	0.374807	0.264670	0.507053	1.352837	0.069029	1.268276	0.545608
738	(Estonia, Estonian, Fully employed)	(not defaulted)	0.453728	0.445814	0.264670	0.583323	1.308445	0.062392	1.330014	0.431533
518	(Male, Fully employed)	(defaulted)	0.434012	0.554186	0.237563	0.547366	0.987694	-0.002960	0.984934	-0.021538
112	(Income and expenses verified)	(not defaulted)	0.418018	0.445814	0.218288	0.522198	1.171336	0.031930	1.159865	0.251338
146	(Woman)	(defaulted)	0.416584	0.554186	0.217351	0.521745	0.941462	-0.013514	0.932168	-0.096311
133	(MoreThan5Years)	(defaulted)	0.380019	0.554186	0.209298	0.550758	0.993815	-0.001303	0.992370	-0.009939
136	(Secondary Education)	(defaulted)	0.404009	0.554186	0.202763	0.501877	0.905611	-0.021133	0.894988	-0.148849

Sorted by support

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(HR, Spain)	(defaulted)	0.155306	0.554186	0.123015	0.792081	1.429269	0.036946	2.144171	0.355563
(HR, Spanish)	(defaulted, Spain)	0.153844	0.157953	0.121829	0.791898	5.013512	0.097529	4.046323	0.946089
(HR, Spain, Spanish)	(defaulted)	0.153844	0.554186	0.121829	0.791898	1.428939	0.036571	2.142288	0.354758
(HR, Spanish)	(defaulted)	0.153844	0.554186	0.121829	0.791898	1.428939	0.036571	2.142288	0.354758
(HR, Spain)	(defaulted, Spanish)	0.155306	0.156602	0.121829	0.784446	5.009183	0.097508	3.912702	0.947522
(Spain)	(defaulted)	0.204032	0.554186	0.157953	0.774159	1.396929	0.044881	1.974014	0.356979
(Spanish)	(defaulted, Spain)	0.202405	0.157953	0.156602	0.773706	4.898335	0.124631	3.721027	0.997810
(Spain, Spanish)	(defaulted)	0.202405	0.554186	0.156602	0.773706	1.396112	0.044432	1.970063	0.355726
(Spanish)	(defaulted)	0.202405	0.554186	0.156602	0.773706	1.396112	0.044432	1.970063	0.355726
(98.7738557226849, Estonian)	(not defaulted)	0.144165	0.445814	0.111378	0.772571	1.732944	0.047107	2.436740	0.494193
(98.7738557226849, Estonia, Estonian)	(not defaulted)	0.142952	0.445814	0.110357	0.771991	1.731643	0.046628	2.430542	0.492987
(Spain, Fully employed)	(defaulted)	0.156574	0.554186	0.120588	0.770166	1.389724	0.033817	1.939718	0.332492
(Spain, Spanish, Fully employed)	(defaulted)	0.155554	0.554186	0.119623	0.769013	1.387644	0.033417	1.930036	0.330813
(Spanish, Fully employed)	(defaulted, Spain)	0.155554	0.157953	0.119623	0.769013	4.868623	0.095053	3.645425	0.940975
(Spanish, Fully employed)	(defaulted)	0.155554	0.554186	0.119623	0.769013	1.387644	0.033417	1.930036	0.330813
(98.7738557226849, Estonia)	(not defaulted)	0.158146	0.445814	0.121498	0.768265	1.723286	0.050994	2.391465	0.498558
(Spain)	(defaulted, Spanish)	0.204032	0.156602	0.156602	0.767536	4.901203	0.124650	3.628084	1.000000
(98.7738557226849, Estonia, Fully employed)	(not defaulted)	0.141077	0.445814	0.108096	0.766224	1.718707	0.045202	2.370581	0.486850
(98.7738557226849, Estonian)	(not defaulted, Estonia)	0.144165	0.348748	0.110357	0.765493	2.194976	0.060080	2.777117	0.636120
(Spain, Fully employed)	(defaulted, Spanish)	0.156574	0.156602	0.119623	0.764001	4.878631	0.095103	3.573743	0.942613

Sorted by confidence

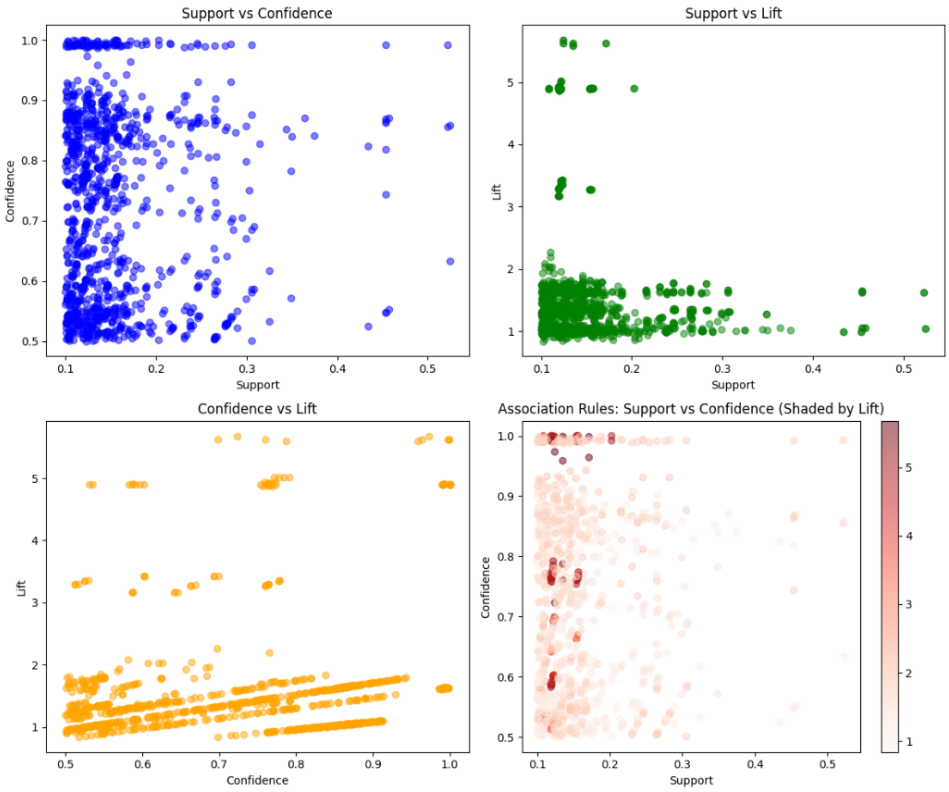
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(Finnish)	(defaulted, Finland)	0.171492	0.127344	0.123980	0.722946	5.677113	0.102141	3.149767	0.994383
(Finland)	(defaulted, Finnish)	0.177614	0.124145	0.123980	0.698028	5.622678	0.101930	2.900453	0.999712
(HR, Spanish)	(defaulted, Spain)	0.153844	0.157953	0.121829	0.791898	5.013512	0.097529	4.046323	0.946089
(HR, Spain)	(defaulted, Spanish)	0.155306	0.156602	0.121829	0.784446	5.009183	0.097508	3.912702	0.947522
(Spain)	(defaulted, Spanish, Fully employed)	0.204032	0.119623	0.119623	0.586295	4.901203	0.095216	2.128034	1.000000
(Spain)	(defaulted, Spanish)	0.204032	0.156602	0.156602	0.767536	4.901203	0.124650	3.628084	1.000000
(Spain)	(defaulted, HR, Spanish)	0.204032	0.121829	0.121829	0.597108	4.901203	0.096972	2.179667	1.000000
(Spanish)	(defaulted, Spain, Fully employed)	0.202405	0.120588	0.119623	0.591008	4.901057	0.095215	2.150195	0.997953
(Spanish)	(defaulted, Spain)	0.202405	0.157953	0.156602	0.773706	4.898335	0.124631	3.721027	0.997810
(Spanish)	(defaulted, Spain, HR)	0.202405	0.123015	0.121829	0.601907	4.892977	0.096930	2.202968	0.997530
(Spain, Fully employed)	(defaulted, Spanish)	0.156574	0.156602	0.119623	0.764001	4.878631	0.095103	3.573743	0.942613
(Spanish, Fully employed)	(defaulted, Spain)	0.155554	0.157953	0.119623	0.769013	4.868623	0.095053	3.645425	0.940975
(Spain)	(defaulted, HR)	0.204032	0.175877	0.123015	0.602919	3.428076	0.087130	2.075455	0.889848
(Spain, Spanish)	(defaulted, HR)	0.202405	0.175877	0.121829	0.601907	3.422322	0.086231	2.070179	0.887418
(Spanish)	(defaulted, HR)	0.202405	0.175877	0.121829	0.601907	3.422322	0.086231	2.070179	0.887418
(HR)	(defaulted, Spain)	0.232021	0.157953	0.123015	0.530188	3.356622	0.086366	1.792306	0.914193
(HR)	(defaulted, Spain, Spanish)	0.232021	0.156602	0.121829	0.525077	3.352950	0.085494	1.775864	0.913768
(HR)	(defaulted, Spanish)	0.232021	0.156602	0.121829	0.525077	3.352950	0.085494	1.775864	0.913768
(98.7738557226849, Estonia)	(not defaulted, Estonian)	0.158146	0.308019	0.110357	0.697820	2.265511	0.061645	2.289966	0.663534
(98.7738557226849, Estonian)	(not defaulted, Estonia)	0.144165	0.348748	0.110357	0.765493	2.194976	0.060080	2.777117	0.636120

Sorted by lift

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(Spain)	(defaulted, Spanish)	0.204032	0.156602	0.156602	0.767536	4.901203	0.124650	3.628084	1.000000
(Spanish)	(defaulted, Spain)	0.202405	0.157953	0.156602	0.773706	4.898335	0.124631	3.721027	0.997810
(Estonian)	(not defaulted, Estonia)	0.525811	0.348748	0.305647	0.581288	1.666785	0.122272	1.555370	0.843635
(Estonia)	(not defaulted, Estonian)	0.610247	0.308019	0.305647	0.500859	1.626064	0.117680	1.386343	0.987852
(Estonian)	(not defaulted, Estonia, Fully employed)	0.525811	0.299084	0.264670	0.503356	1.682991	0.107408	1.411305	0.855818
(Estonian, Fully employed)	(not defaulted, Estonia)	0.457396	0.348748	0.264670	0.578646	1.659209	0.105154	1.545617	0.732215
(Estonia, Fully employed)	(not defaulted, Estonian)	0.524073	0.308019	0.264670	0.505025	1.639591	0.103246	1.398012	0.819647
(Finnish)	(defaulted, Finland)	0.171492	0.127344	0.123980	0.722946	5.677113	0.102141	3.149767	0.994383
(Finland)	(defaulted, Finnish)	0.177614	0.124145	0.123980	0.698028	5.622678	0.101930	2.900453	0.999712
(HR, Spanish)	(defaulted, Spain)	0.153844	0.157953	0.121829	0.791898	5.013512	0.097529	4.046323	0.946089
(HR, Spain)	(defaulted, Spanish)	0.155306	0.156602	0.121829	0.784446	5.009183	0.097508	3.912702	0.947522
(Spain)	(defaulted, HR, Spanish)	0.204032	0.121829	0.121829	0.597108	4.901203	0.096972	2.179667	1.000000
(Spanish)	(defaulted, Spain, HR)	0.202405	0.123015	0.121829	0.601907	4.892977	0.096930	2.202968	0.997530
(Spain)	(defaulted, Spanish, Fully employed)	0.204032	0.119623	0.119623	0.586295	4.901203	0.095216	2.128034	1.000000
(Spanish)	(defaulted, Spain, Fully employed)	0.202405	0.120588	0.119623	0.591008	4.901057	0.095215	2.150195	0.997953
(Spain, Fully employed)	(defaulted, Spanish)	0.156574	0.156602	0.119623	0.764001	4.878631	0.095103	3.573743	0.942613
(Spanish, Fully employed)	(defaulted, Spain)	0.155554	0.157953	0.119623	0.769013	4.868623	0.095053	3.645425	0.940975
(Spain)	(defaulted, HR)	0.204032	0.175877	0.123015	0.602919	3.428076	0.087130	2.075455	0.889848
(HR)	(defaulted, Spain)	0.232021	0.157953	0.123015	0.530188	3.356622	0.086366	1.792306	0.914193
(Spain, Spanish)	(defaulted, HR)	0.202405	0.175877	0.121829	0.601907	3.422322	0.086231	2.070179	0.887418

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.



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

Model Training and Evaluation

model	dataset	Accuracy		Precision		Recall		F1	
Logistic Regression	Train	0.814		0.810		0.814		0.819	
	Test	0.816	81.62	0.812		0.816		0.817	
Decision Tree	Train	0.832		0.835		0.832		0.822	
	Test	0.830		0.833		0.830		0.891	
Random Forest	Train	0.815		0.824		0.815		0.799	
	Test	0.817		0.826		0.817		0.801	
Linear SVC	Train	0.813		0.809		0.813		0.806	
	Test	0.815		0.812		0.815		0.808	
Gradient Boosted Tree	Train	0.892		0.895		0.892		0.888	
	Test	0.891		0.894		0.891		0.887	
KNN (Map-Reduce) 1000	Test	0.66	0.66 sklearn						

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**