Erasmus University Rotterdam

Erasmus School of Economics

Credit risk assessment in Peer-to-Peer Lending during the COVID-19 pandemic

Master Thesis

Author:

Jan Mukherjee

573207

Supervisor:

Dr. X.Ma

Abstract

This paper studies the effect of loan and personal characteristics based on current dataset consisting of 211,552 loan observation from the P2P lending platform Bondora.com. Available determinants are evaluated individually and further investigated through logistic regression models enabling a deep understanding for potential drivers of default probability. Besides the importance of rather established covariates such as *Rating*, Interest, Amount, Income, Liabilities, further drivers for default rate such as *NewCustomer*, *Country*, *Gender* and *HomeOwner* show an important impact. Additionally, the rather young P2P lending market, which evolved after the great financial crisis in 2008, experience its first economic downturn induced by the COVID-19 pandemic. With more than two years since the beginning of the pandemic this study explores possible structural effects on the estimation of P2P lending credit risk. COVID appears to have a negative correlation on default rates in P2P lending while acting as a significant interaction term for covariates.

Keywords: P2P lending, credit risk, COVID-19, logistic regression, interaction effect

Table of Contents

I List of Tables

1. Introduction

Continuous digitalisation of the financial sector led to an increasing use of fintech applications. Looking at the lending sector Peer-to-Peer (P2P), lending platforms stand out with about EUR 18 billion in volume (https://p2pmarketdata.com/p2p-lending-funding-volume-eu/). As a debt instrument it is crucial for debtors, creditors, regulators and platform providers to accurately estimate default rates for P2P credit in order to account for risks, costs, returns and lending volume. Furthermore, as a rather new financial instrument that emerged after the financial crisis 2007-2008 effects of high market volatility and changing economic conditions on P2P credits are still unclear. With the COVID-19 pandemic lasting for two years causing major changes on financial markets the current characteristics of P2P lending must be revised. However, this topic remains unfocused in P2P lending. The importance of this topic and lack of empirical evidence gives rise to the following research question:

How does the credit risk assessment in Peer-to-Peer Lending perform during COVID-19 pandemic?

To infuse characteristics of P2P lending this study will provide an empirical analysis of the P2P lending market by building a model to predict default rates. Here I will motivate and provide information on the relevance of the topic. Furthermore, this section will give an overview on the study itself. Continuous digitalisation of the financial sector led to an increasing use of fintech applications. Looking at the lending sector Peer-to-Peer (P2P) lending platforms stand out with about EUR 18 billion in volume (https://p2pmarketdata.com/p2p-lending-funding-volume-eu/). As a debt instrument it is crucial for debtors, creditors, regulators, and platform providers to accurately estimate default rates for P2P credit to account for risks, costs, returns and lending volume. Furthermore, as a rather new financial instrument that emerged after the financial crisis 2007-2008 effects of high market volatility and changing economic conditions on P2P credits are still unclear. With the COVID-19 pandemic lasting for two years causing major changes on financial markets the current characteristics of P2P lending must be revised. To infuse characteristics of P2P lending this study will provide an empirical analysis of the P2P lending market by building a model to predict default rates.

The rest of this study is structured in the following way. Based on related literature Section 2 an introduction into the market of P2P lending is made by explaining the unique characteristics as well as the differences to traditional lending with a focus on credit risk. Additionally, the development of the COVID-19 pandemic and its relation to P2P credit risk is derived. Section

3 describes the dataset used in the analysis and combines the previous conclusions to form X hypotheses. In the following Section 4, logistic regression modelling and adequate validation measures are introduced to develop the models to be analysed. Section 5 investigates and summarizes the results to answer the proposed hypotheses. Finally, Section 6 concludes the study providing an outlook for further research.

2. Related Literature

In this chapter a thorough literature review is conducted to inform on existing research in P2P lending, credit risk measuring and the impact due to the COVID-19 pandemic.

2.1 Peer-to-peer lending

The development of information technology fostered the creation of digital marketplaces reducing the need for intermediaries (Berger & Gleisner, 2009). Inevitably this dynamic also affected financial markets and led to the development of financial technology applications. For the credit market in particular P2P Lending developed as a means of more "decentralized" financing. With the first platforms started in 2005 P2P Lending is recent financial innovation providing alternative funding (Ölvedi, 2020). It provides a new way in the loan-origination process for both lenders and borrowers based on the concept of crowdfunding (Beáta Gavurová et al., 2018) Here, on the one side the borrower can directly list a loan request on an online platform (i.e., Bondora, Prosper, LendingClub). On the other side, lenders can freely provide fractional funding to the borrower. As an intermediary P2P-lending platforms are deeply embedded in this process not only connecting both parties but gathering and processing data as well as facilitating payments and secondary transactions. Based on this concept P2P lending is often referred to as "marketplace lending", as funds are provided by peers (Najaf et al., 2021). The innovative technology underlying this form of funding makes it possible to provide loans with decreasing intermediation costs (ibid.). Hence, this provides the possibility for borrowers to receive lending at more attractive conditions compared to the traditional banking sector. Additionally, the rather new concept of funding comes with less regulatory hurdles. In fact, the majority of borrowers served through P2P lending platforms appear to be a portion of the customer credit market abandoned by the regular banking sector (Roure et al., 2016). With tightened regulatory conditions and increased risk aversion of banks after the financial crisis, this suggests that currently P2P lending markets are mostly driven by high-risk borrowers (ibid.).

Furthermore, P2P lending provides investors the possibility to realise higher returns matching the increased risk of default for the funded loans (Bachmann et al., 2011). De Roure et al. (2016) show that these loans are comparable to traditional loans when adjusting higher interest rates for the increased risk.

Even though P2P lending offers multiple advantages it also comes with concerns and unique problems. Especially the risk of default plays a vital role in the funding process and creates concerns due to limited default consequences and untested screening standards creating information asymmetry problems (Ölvedi, 2020).

2.2 Estimating credit risk

Besides the advantages of P2P lending there are also difficulties. Lacking a personal relationship and engagement between borrowers and lenders the accurate prediction of credit rating is aggravated. As investors may not have sufficient information on individual borrowers to evaluate credibility correctly adverse selection effects arise (Akerlof, 1970). This relates to the problem that lenders have less information to accurately predict borrowers credit risk while the borrower knows leading to mispricing and information asymmetries (Emekter et al., 2015). This problem is already well known in traditional credit markets where adverse selection can cause credit constraints for borrowers (Stiglitz & Weiss, 1981). Even though this phenomenon is already frequently researched for the traditional market the differing lending setup creates unique challenges (Ölvedi, 2020). For instance, this is important as the investment decision is not delegated to the intermediary, in this case the P2P lending platform, but lies in the hands of each investor alone. Therefore, traditional financial intermediaries such as banks take over the credit risk and optimize the screening process to identify the individual credibility. Paired with protective regulation and laws it safeguards lenders money and protect against potential loses (Aveni et al., 2015). As Fintech innovations are still unregulated certain protective measures for lenders are not yet applicable to P2P lending platforms leaving investors at higher risk of loses (Philippon, 2016). With more and more data being gathered by these platforms additional screening methods and credibility measures are introduced (Bachmann et al., 2011). As data plays a vital role in the development of financial innovations data generated by these platforms create more opportunities to find patterns and increase credit risk measurement. Especially as high-risk borrowers are major customer base for P2P lending platforms a thorough underwriting process is important for efficient lending (Aveni et al., 2015). For example, the P2P lending platform Bondora launched its first credit risk assessment model in 2012 after four years of operation (Bondora.com, 2022). Still the rather young P2P lending industry often struggles to inform and publicize accurate information regarding the unique risk and return profile (Aveni et al., 2015).

Overall, as credit risk plays a key role for investors lending through P2P lending platforms it is highly important to further analyse the factors leading to loan default (Serrano-Cinca et al., 2015). Stressing macroeconomic environments reinforce these uncertainties leaving the question on how robust P2P lending markets are and may provide improvements for credit risk measurement. (Najaf et al., 2021). Hence in that sense the hypothesis is created stating that "*Including available loan and borrower characteristics improves default rate prediction for P2P lending.*" (Hypothesis 1)

2.3 Credit risk during COVID-19

This chapter will introduce the impact of the current COVID-19 pandemic on credit markets and the corresponding risk. Here, a changing market environment and use of P2P lending through COVID-19 will be explored. Beginning with the first case of COVID-19 discovered at the end of 2019 a pandemic has emerged all over the world (Augustin et al., 2021). The induced humanitarian crisis with over 400 million cases of infection and more than 6 million deaths due to COVID-19 (John Hopkins Coronavirus Resource Centre, 2022) also put the global economy into a rigid state. Major disruptions in global economic activity such as global trade caused enormous economic growth shocks (Augustin et al., 2021). The COVID-19 pandemic caused major risk levels on financial markets.(Zhang et al., 2020). The financial and economic consequences affects markets both on micro and macroeconomic levels (Goodell, 2020). It is suspected that this causes a severe, wide-scale economic crisis that is expected to continue even after the pandemic (ibid.). The large impact of COVID-19 on the global economy has led to vast research on the effects on financial markets (Dubinova et al., 2021). Especially economically stressed companies and individuals with increased liquidity problems drives the loan demand (World Bank, 2020). Still, there is rather few academic research on the effects of the COVID-19 pandemic on P2P lending. Najaf et al., (2021) provides a first and thorough analysis on this matter testing the influence of the COVID-19 pandemic on P2P determinants. The analysis shows that P2P loan volume highly increased during the pandemic, which is consistent with the observations on the overall credit market (World Bank, 2020). Still, there seems to be an edge compared to the traditional banking sector, creating more relative demand for P2P lending platforms. With the setup of P2P lending platforms being fully digitalised this result is reasonable in the sense that fast and easy access to funding can be provided to the traditional banking sector lacking online loan verification processes. As banking facilities are prone to strong borrowing rules, borrowers prefer P2P credit during the pandemic. Even though the asset class is largely regarded as high risk, a shifting customer base in need for faster and easier credit might attract more financially potent borrowers. This is even more influenced by the fact that impaired credit quality through the pandemic causes banks to shift to safer assets to mitigate increased losses due to distressed debt (Najaf et al., 2021). This leads to the hypothesis that “*the harsh economic environment induced by the COVID-19 pandemic is significantly correlated with P2P lending default rates”* (Hypothesis 2). Furthermore, P2P lending can be leveraged to tackle financial challenges arising by the pandemic (ibid.). With coronavirus-related constraints affecting lending determinants such as interest rate and verification status a last hypothesis is developed to discuss the impact of the COVID-19 pandemic on credit risk assessment. “*The correlation of loan and borrower characteristics on default probability is influenced by the COVID-19 pandemic”* (Hypothesis 3).

3. Data & Summary Statistics

3.1 Data & Variables

For this study public data from the P2P lending platform Bondora is used. With more than EUR 500 million Loans issued since being established in 2008 the Estonian platform is one of the biggest P2P lending providers in Europe (Bondora.com, 2022). Furthermore, Bondora enables borrowers from Estonia, Finnland and Spain to receive funding from lenders all over Europe. The platform provides investors with the opportunity to invest in unsecured consumer loans containing principal amounts from €500 up to €10,000 with repayment periods from three to 60 months (Bondora, 2020). Additionally, the platform comes with a secondary market and the opportunity for automated investments. On the secondary market, investors have the opportunity to trade loans directly with each other. Investors can exit the investment before maturity, where they will cash out on the investment. Finally, Bondora publicly discloses its data records which makes it a suitable data source to investigate different effects on credit default prediction. Overall, Bondora is well suited to be used for the analysis of P2P lending default rates as one of the biggest platforms existing since the establishment of P2P lending offering public data on multiple countries.

The initial dataset was extracted from Bondora.com on March 7th March 2022, containing 221,155 observations over the period from February 2009 till March 2022. All included loans are issued in the EU-countries Estonia, Finland, and Spain.

Bondora offers a wide range of information on loan and borrower characteristics (112 variables). Here, 15 variables are chosen based on preliminary research on P2P determinants as well as continuous availability (Beáta Gavurová et al., 2018). For the matter of data merging the date of loan issuance is included but omitted in the subsequent analysis. To ensure consistency and interpretability, categorical variables are transformed into dummy variables. The following provides an overview on all variables included in the final data set.

- *Default.* Bondora lists the variable DefaultDate which occurs when loan payment is overdue for at least 60 days (Bondora, 2014). Based on this variable the proxy for credit risk measurement defined as *Default* is constructed. The variable will be used as the dependent variable in the following analysis. Put differently, in case of the used logistic regression the dependent variable will be one if the probability of default is larger than 50% and will be zero, if the probability of default is smaller than 50%.

- *Rating.* Bondora creates a custom credit scoring system to assess possible expected losses after recoveries ranging from best to worst: AA, A, B, C, D, E, F and HR (Bondora.com, 2022). The rating is mainly based on externally validated data including behavioural data from trusted third parties. For this analysis the Rating is transformed to a numerical scale ranging from one to eight equally to a diminishing rating quality (i.e., AA to HR).

Risk Rating	Min EL%	Max EL%
AA	0.0%	2.0%
A	2.0%	3.0%
B	3.0%	5.5%
C	5.5%	9.0%
D	9.0%	13.0%
E	13.0%	18.0%
F	18.0%	25.0%
HR	25.0%	>25.0%

Table 1: Classification - Rating

(Source: Bondora)

Loan characteristics

- *Amount*. When a loan is issued initially the loan amount is paid to the borrower in EUR.
- *Interest.* This variable describes the maximum interest rate in percentage per annum accepted in the loan application process.
- *Duration.* This variable includes the loan duration in months.

Personal characteristics

- *Age.* The age of the borrower when signing the loan application.
- *Estonia/Spain [Country]. Country* represents the residency of the borrower. It is represented by the two dummy variables *Estonia* and *Spain* which take on the value of one when applicable for the given observation. To avoid the dummy variable trap, which is represented by perfect multicollinearity between dummy variables, the country

Finland is not included but inherently defined in the other two dummy variables (*Estonia*=0 and *Spain*=0 indicating *Finland)*.

- *Estonia*. Equal to one in case residency is in Estonia and zero otherwise.
- *Spain*. Equal to one in case residency is in Estonia and zero otherwise.

- *NewCustomer*. To differentiate customer types, the binary variable *NewCustomer* is equal to one in case the customer is new and zero in case of an existing customer.
- *Male/Female[Gender]*. Gender is split up into two dummy variables called *Male* and *Female* being one when applicable. As there also exists the Gender "*Undefinied*" the same concept as for the country variables applies to avoid perfect multicollinearity [*Male*=0 and *Female*=0 indicating *Undefinied]*.
  - *Male*. Equal to one when gender is male and zero otherwise.
  - *Female*. Equal to one when gender is female and zero otherwise.
- *Income*. Income describes the total monthly income of the borrower in EUR.
- *VerifiedIncome*. The binary variable *VerifiedIncome* is based on the method used for loan application data verification. According to the following classification the variable equals one when Income is verified and zero otherwise.

Value	Classification
0	0 - Not set, 1 - Income unverified, 2 - Income unverified, cross-referenced by phone
1	3 - Income verified, 4 - Income and expenses verified

Table 2: Classification - VerifiedIncome

- *Education*. Education dummy variable classifies borrowers that received higher education as one and the rest that did not receive higher education as zero based on the following classes.

Value	Classification
0	1 - Primary education, 2 - Basic education, 3 - Vocational education, 4 - Secondary education
1	5 – Higher Education

Table 3: Classification - Education

- *Employment*. Employment dummy assigns zero to borrowers that are currently working less than five years with their current employer and one to those that work more than five years with the current employer.

- *ExistingLiabilities*. Borrower's number of existing liabilities
- *Liabilities*. Based on the total monthly liabilities in EUR for each borrower.
- *HomeOwner*. The dummy variables *HomeOwner takes the value one when the borrowers is classified as a home owner based on the following classification and zero otherwise.*

Value	Classification
0	0 Homeless, 2 Living with parents, 3 Tenant, pre-furnished property, 4 Tenant, unfurnished property 5 Council house 6 Joint tenant
1	1 Owner, 7 Joint ownership, 8 Mortgage, 9 Owner with encumbrance, 10 Other

Table 4: Classification - HomeOwner

External

- *COVID.* Finally, the data set is merged with external data on COVID-19 cases. Here a proxy for the pandemic is created based on existing cases. Hence, a dummy variable was created that holds the value of 1 in the case of existing COVID-19 cases in the country at the time when the loan was issued. The COVID-19 data is originated from the John Hopkins Coronavirus Resource Centre (2022) as of 12th April 2022. Both data sets are merged based on the date and country.

For analysis reasons all observations are dropped that incorporate non-interpretable values for at least one of the variables. Additionally, as the unique rating by Bondora was introduced in the end of 2012, all observations for loans before November 2012 are dropped. Hence the final data set ranges from 10th November 2012 till 7th March 2022. This leads to a final data set incorporating 211,552 observations.

Lastly, to allow for thorough model validation, the main data set is randomly split into a training set and a testing set with a ratio of 80% (training) to 20% (testing) as the empirically optimal split (Gholamy et al., 2018).

3.2 Descriptive statistics

The following provides an overview on the final dataset used in the analysis. To receive first insight into the dataset, both the overall dataset including all observations as well as the population for all defaulted and non-defaulted loans are summarized in *Table 6*. For continuous variables the mean and standard deviation is displayed. For binary variables the number of positive observations equal to zero is described in relation to the total sample size.

Overall, the average borrower is a 40-year-old male customer from Estonia. Moreover, an average income of €1,470 per month, which is slightly below the EU average of approximately €1,700 (Clark, 2020). Then again, this is expected since P2P platforms service underbanked clients (Tang, 2019). Additionally, the average customer did not receive higher education yet and works less than five years for the current employer. As a rather new form of lending experiencing strong growth, it is intuitive for the average customer to be new to the platform. Moreover, the average loan amount is €2,595 with a duration of 49 months which corresponds to the industry average of 36 to 60 months (Faia & Paiella, 2017). The largest portion of the loans have a rating between four and five which translates to a C - D rating with an average maximum interest rate of 32%. While this might sound high, it is in line with the platform's focus on providing smaller uncollateralized loans to borrowers with less credibility (Tang, 2018). Additionally, borrowers usually have two to three liabilities in total of €380 per month.

To make a first assessment on the relevance of each variable regarding loan defaults a new variable *DefaultRatio* is constructed creating a subset for each variable. *DefaultRatio* is defined in equation (I) as follows:

$$\text{(I)} \qquad DefaultRatio = \frac{Number\ of\ Defaults\ at\ value\ x}{Number\ of\ Observations\ at\ value\ x} \sim [0\ ,1]$$

For every variable a new subset of the initial dataset is created with the DefaultRatio being grouped at all possible values x for each variable.

Consequently, a simple linear regression model is applied for continuous variables to estimate the correlation between each variable and *DefaultRatio*. The 211,552 loans issued on the P2P lending platform are separated into 136,396 (64%) loans with no default associated leaving 75,156 (36%) defaulted loans. On average the *Amount* slightly increases to 2,688 compared to 2,544 for non-defaulted loans. At the same time duration remains constant at 49 months for all groups. Furthermore, borrowers that defaulted on their loan must pay 15 percentage points more with an interest rate of 42% p.a. on average. The rather large difference is confirmed by a significant correlation between *Interest* and DefaultRatio as described in Figure 1. Due to the natural risk-return relationship this result is in line with intuition.

Similarly, we observe a shift in the *Rating* distribution with the median falling into the category 6 for defaulted loans and 4 for non-defaulted loans. Also, for monthly *Liabilities* an increase of 86 EUR to an average of 435 is observed. Again, both relationships are confirmed through a significant linear relationship to *DefaultRatio* visualised see Appendix 3.

Variable	**Overall, N = 211,552**[1]	**Default**		**p-value**[2]
		0, N = 136,396[1]	**1, N = 75,156**[1]	
Amount	2,595 (2,182)	2,544 (2,185)	2,688 (2,175)	<0.001
Interest	32 (23)	27 (15)	42 (29)	<0.001
Duration	49 (16)	49 (18)	49 (13)	<0.001
Rating				<0.001
1	8,126 / 211,552 (3.8%)	6,814 / 136,396 (5.0%)	1,312 / 75,156 (1.7%)	
2	9,511 / 211,552 (4.5%)	7,447 / 136,396 (5.5%)	2,064 / 75,156 (2.7%)	
3	28,998 / 211,552 (14%)	23,719 / 136,396 (17%)	5,279 / 75,156 (7.0%)	
4	43,239 / 211,552 (20%)	34,146 / 136,396 (25%)	9,093 / 75,156 (12%)	
5	46,613 / 211,552 (22%)	33,494 / 136,396 (25%)	13,119 / 75,156 (17%)	
6	34,844 / 211,552 (16%)	18,800 / 136,396 (14%)	16,044 / 75,156 (21%)	
7	25,808 / 211,552 (12%)	8,524 / 136,396 (6.2%)	17,284 / 75,156 (23%)	
8	14,413 / 211,552 (6.8%)	3,452 / 136,396 (2.5%)	10,961 / 75,156 (15%)	
Age	40 (12)	40 (12)	41 (13)	<0.001

Variable	Overall, N = 211,552[1]	Default		p-value[2]
		0, N = 136,396[1]	1, N = 75,156[1]	
Estonia	126,143 / 211,552 (60%)	92,041 / 136,396 (67%)	34,102 / 75,156 (45%)	<0.001
Spain	26,286 / 211,552 (12%)	6,896 / 136,396 (5.1%)	19,390 / 75,156 (26%)	<0.001
NewCustomer	117,093 / 211,552 (55%)	69,456 / 136,396 (51%)	47,637 / 75,156 (63%)	<0.001
Male	123,812 / 211,552 (59%)	77,558 / 136,396 (57%)	46,254 / 75,156 (62%)	<0.001
Female	75,358 / 211,552 (36%)	55,535 / 136,396 (41%)	19,823 / 75,156 (26%)	<0.001
Income	1,470 (782)	1,478 (803)	1,456 (744)	0.2
VerifiedIncome	159,676 / 211,552 (75%)	107,864 / 136,396 (79%)	51,812 / 75,156 (69%)	<0.001
Employment	77,933 / 211,552 (37%)	49,743 / 136,396 (36%)	28,190 / 75,156 (38%)	<0.001
Education	49,167 / 211,552 (23%)	31,447 / 136,396 (23%)	17,720 / 75,156 (24%)	0.007
ExistingLiabilities	2.84 (3.09)	2.80 (2.96)	2.90 (3.31)	<0.001
Liabilities	380 (442)	349 (396)	435 (511)	<0.001
HomeOwner	121,199 / 211,552 (57%)	83,332 / 136,396 (61%)	37,867 / 75,156 (50%)	<0.001
COVID	79,445 / 211,552 (38%)	72,584 / 136,396 (53%)	6,861 / 75,156 (9.1%)	<0.001

[1] Mean (SD); n / N (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

Table 5: Summary Statistics

To assess the relevance of binary variables, a different approach must be considered as compared to the continuous variables previously analysed. This is caused as binary variables are inherent to a Bernoulli distribution. The Bernoulli distribution is as special case of the binomial distribution where only a single trial is conducted. Under this distribution the variable takes the value of one with the probability p and the value of zero with the probability $q = 1 - p$. To assess the relevance of the variable on the default variable the following hypothesis is tested via a Pearson's Chi-squared test for comparing response probabilities with the following null-hypothesis:

H_0: The frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

For this case the H_0 can be understood as such that the difference in the Bernoulli distribution for the binary variable regarding Default is zero. The following table provides the results for the test including *DefaultRatio* for both groups as well as the p-value for significance assessment. *DefaultRatio* corresponds to the resulting probabilities of the Pearson's chi-square test.

Table 5 shows that all variables are significant at the 0.01 level indicating a significant change for the probability distribution of the variable based on *Default*. For the further analysis the relative difference between *DefaultRaio* is calculated in the following way (II):

(II) $$Relative\Delta = \frac{DefaultRatio_0 - DefaultRatio_1}{DefaultRatio_0}$$

Here, $DefaultRatio_0$ corresponds to the DefaultRatio in case the binary variable is equal to 0 and $DefaultRatio_1$ in case of the binary variable equal to 1. For this sample of variables, the highest $Relative\Delta$ is associated with the Country variables *Estonia* and *Spain* with a value of -43% and +145% respectively. Additionally, the *DefaultRatios* for *NewCustomer (+40%), Female (-35%),* VerifiedIncome (-28%), *HomeOwner(-24%)* and COVID (-83%) indicate that a connection to the probability of default for P2P loans might exist. Hence, these variables in addition to the above-mentioned continuous variables will be the focus of the further analysis.

Variable	DefaultRatio		Relative Δ	p-value[1]
	0	**1**		
Estonia	48,1%	27,0%	-43%	<0.01
Spain	30.1%	73.8%	145%	<0.01
NewCustomer	29.1%	40.6%	40%	<0.01
Male	32.9%	37.4%	14%	<0.01
Female	40.6 %	26.3%	-35%	<0.01
VerifiedIncome	45.0%	32.4%	-28%	<0.01
Employment	35.1%	36.2%	3%	<0.01
Education	35.4%	36.0%	2%	<0.01
HomeOwner	41.3%	31.2%	-24%	<0.01
COVID	51.7%	8.6%	-83%	<0.01
[1]Chi squared statistic				

Table 6: Binary Variables

4. Methodology and Empirical Setting

This chapter will describe the methodology used for the analysis of the previously defined hypothesises. In a first step logistic regression is introduced to enable the modelling of the binary variable default.

In a second step, validation measures and their theoretical framework as well as the intended use are introduced. The last step refers to the model development, where the logistic regression models based on the selected variables are set up in order to evaluate the stated hypothesis.

4.1 Logistic regression

To analyse the hypotheses, default probability has to be estimated based on given variables. As this describes a binary response process this can be facilitated by performing logistic regression. This is necessary as the constructed dependent variable Default is binary rather than continuous. The logit model requires a linear relationship between the independent variable and each dependent variable. As the dependent variable cannot be estimated by using the ordinary-least-squares estimation used for linear regression models, the maximum-likelihood estimation method is applied. The likelihood function measures how well the values used as input for the model fit in the model. Here, the regression coefficients are used to maximise the likelihood function of explaining the dependent variable. It is worth noting that the Maximum likelihood estimation differs from normal linear models where coefficients are estimated through iteratively reweighted least squares (Murphy, 2012).

Moreover, a logit model assumes no multicollinearity among input variables and allows for a straightforward econometric interpretation based on coefficient estimates and variable significance.

In a logistic regression model the interest primarily lies in the response probability. Logistic models are a form of generalized linear models (GLMs). In general, a GLM consists of three components. First, is the random component which considers the probability distribution of the dependent variable (Y). In comparison to normal linear models GLMs allow for the dependent variable to be distributed in different ways (i.e., Binomial, Poisson, …) compared to normal linear models. The logit model requires a linear relationship between the independent variable and each dependent variable (Wooldridge, 2016).

This corresponds to the second component, which is unique for GLM models, the linear link function. The link function is a non-linear function which connects the linear model to the non-normally distributed dependent variable.

In the case of the logistic regression used in this analysis the dependent variable is binary variable following a Bernoulli distribution.

(III) $y_i \sim Bernoulli(\theta_i), for\ i \epsilon\ 1, \ldots , n$

The used model must be robust to only allowing for each θ_i to only take values between 1 and 0. As this cannot be guaranteed by normal linear models, a GLM model is used with a link function that transforms each possible value of the linear model ranging from -∞ till ∞ in the R dimension to the interval [0,1].

(IV) $y_i \sim Bernoulli(\theta_i), \theta = f(\emptyset), \emptyset\ = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\, for\ i\ \in 1, \ldots, n$

This equation clearly shows that to estimate the θ for the Bernoulli distributed independent variable, a different variable is first estimated which is linked to θ via the logit function defined as:

(V) $\emptyset = f(\theta) =\ logit(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$

As the estimation of the dependent variable is not performed directly this influences the interpretation for the regression coefficients. As we receive an estimator for the $\emptyset$ the estimation coefficients must be transformed first. It is important to understand that the coefficients do not directly indicate the correlation between dependent and independent variables, but the difference in the log-odds of the probability of the outcome variable. The odds ratio represents the ratio of the odds that a default will occur (*Default* = 1) given the presence of the predictor x (x = 1) in case of a dummy variable or the increase of the predictor for non-binary variables, compared to the odds of a default occurring in the absence of that predictor (x = 0) or when there is no change in the predictor (non-binary). Now, for a fitted model the beta coefficient of a unique predictor corresponds to the log of the specific odds ratio.

This paper utilizes a four-stage logit model to assess the prediction power of loan and personal characteristics as well as the impact of the COVID-19 pandemic. Next to assessing the outcome, errors are also determined. If the value "1" or "0" is not equal to the value, which comes out of the regression for Default, an error value will be created. Consequently, the average is taken of the errors resulting in the absolute error, which shows how predictive the logistic regressions are. This absolute error is useful in determining the best model out of the two.

In general, each coefficient in the fitted model can be interpreted independently based on the assumption that all other coefficients stay the same. To get a better understanding for the impact of logistical regression models interaction terms can be calculated to infer how the effect of one independent variable is dependent on the magnitude of another independent variable. As this odds ratio can be defined as an interaction effect can

4.2 Validation Methods

To further assess the best model fitted validation methods will be performed to measure the model with the most reliable predictors. Therefore, multiple measures which are advised in the literature are utilized (Kuha, 2004). Firstly, the Likelihood-Ratio-Test (LRT) is performed to assess the goodness of fit of the models. This measure indicates if a correlation between the independent variables and the dependent variables exists. The formula is:

(VI) $$LRT = RD_1 - RD_0 = -2(lnL_1 - lnL_0)$$

Here, *RD* stands for Residual Deviance with *L* describing the resulting likelihood from the Maximum-Likelihood estimation for the logit models.

In addition to that, a coefficient of determination is calculated for each model. Even though the classic R^2 cannot be computed for logistical regressions a similar measure is used to evaluate the explanatory strength for each model. For binary response models such as the logit model, McFadden's Pseudo R^2 as well as the adjusted R^2 allow for the evaluation of the models. When comparing two models on the same data, McFadden's would be higher for the model with the greater likelihood. The two measures are defined by the following formulas:

(VII) McFadden's Pseudo R^2

$$R^2 = 1 - \frac{RD_1}{RD_0}$$

(VIII) McFadden's adjusted Pseudo R^2

$$R^2_{adj} = 1 - \frac{RD_1 - K}{RD_0}$$

Here, K is equal to the number of parameters added to the intercept model and hence penalises for additional parameters included in the model. This provides the possibility to make goodness-of-fit comparable between models with varying parameter numbers (Long & Freese, 2006). R^2 has the problem, that additional variables do not increase the R^2 which may lead to overfitting the model. What is Overfitting? Adjusted R^2 provides a solution as this measure penalizes the loss of degrees of freedom from adding variables to the model (Greene, Econometric Analysis) Book Econometric Analysis: p.183 Ch. 5.8

To receive robust results, further methods of assessing model fit are introduced. Accounting for overfitting by penalizing for the number of estimated parameters, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are introduced. On the one hand the Bayesian information criterion measures the parameterized models' efficiency in predicting the data. A lower BIC score constitutes a better model fit. On the other hand, the Akaike information criterion is like the BIC since they both penalize for the number of variables. However, the degree of penalty for additional variables differs between the AIC and BIC (i.e., $2k$ for AIC and $k\ln(n)$ for BIC). It is argued that the AIC is asymptotically optimal for selecting the model with the least mean squared error (Yang, 2005). This is based on the fact, that the best model is not in the candidate set. As for the AIC a lower score also constitutes a better model fit. The AIC and BIC are defined as:

(IX) $BIC = k * ln(n) - 2\ln(\hat{L})$

(X) $AIC = 2k - 2\ln(\hat{L})$

Here, *k* relates to the number of predictors used as independent variables and *n* for the number of observations. *L* again stands for the maximum value of the likelihood function.

In addition to that, the phenomenon of overfitting is considered in this analysis. When the corresponding phenomenon described by the model is fully known it is best to use all data available (Gholamy et al., 2018). Unfortunately, in most cases perfect adequacy of models is not given which might lead to the divergence between accuracy and adequacy. By splitting the data into a training and testing set, first the models' parameters are determined and then tested out-of-sample (ibid.). In this analysis, the empirically optimal split of 80% for the training and 20% for the testing set is applied. After fitting the logistical model on the training set the testing set is used to compare predictions on *Default* of the fitted model. Based on the results compared to the testing set the accuracy is determined in the following way:

(XI) $Accuracy = \frac{TN+TP}{TP+FP+TN+FN}$

Here, TN stands for true negative, TP stands for true positive, FP for false positive and FN for false negative.

4.3 Model Development

To empirically analyse the stated hypotheses a four-stage logit model is developed. First, a baseline model is created that calculates the explanatory power of the *Rating*. Rating is a major indicator used by borrowers on the Bondora platform to assess credit risk. As the Rating is mainly based on externally validated data related to the individual borrower the fitted model not only serves as a benchmark for the extended models but also provides the possibility to explore the power of the credit risk measure directly available to lenders.

For the following models

Baseline Model

$$\text{(XII) } Default = \beta_0 + \beta_1 Rating_i + \varepsilon_i$$

Subsequently, the second-stage logit model adds the loan characteristics and personal determinants from the internal Bondora dataset. These variables are aggregated in the described formula but for a detailed overview please refer to Appendix A. Generally, based on the preliminary variable analysis in section 3.2 interesting variables are analysed based on this model.

Extended Model

$$\text{(XIII) } Default = \beta_0 + \beta_1 Rating_i + \beta_2 Loan_i + \beta_3 Personal_i + \varepsilon_i$$

In a next step the COVID dummy variable is added to create the third model including all variables from the dataset. As described in chapter 3.1 the variable was created and merged with the data set by matching corona cases based on the country and (loan) date available. As used by recent studies this seems to be a promising way of approximating the status of COVID-19 (Najaf et al., 2021).

Full Model

$$\text{(XIV) } Default = \beta_0 + \beta_1 Rating_i + \beta_2 Loan_i + \beta_3 Personal_i + \beta_4 COVID_i + \varepsilon_i$$

Lastly, interaction terms are added to the third model. The term displays the multiplication of the two variables, here an internal variable and COVID, and the corresponding coefficient provides information if the effect of the two variables combined is significantly larger than the sum of the individual effects.

Interaction Model

$$\text{(XV) } Default = \beta_0 + \beta_1 Rating_i + \beta_2 Loan_i + \beta_3 Personal_i + \beta_4 COVID_i + \sigma_1(\mathrm{Personal_i} * COVID_i) + \sigma_2(\mathrm{Personal_i} * COVID_i) + \varepsilon_i$$

5. Main Results

This section discusses the results from the developed models. The results for all logistic regression models are displayed in Appendix 3. For interpretability the odds-ratio for each variable based on the coefficient is calculated as follows.

$$\text{(XVI) } odds\ ratio = \exp(\beta)$$

5.1 Baseline model

First, in the baseline model *Default* is regressed on the *Rating*. Firstly, the significant negative intercept (0.044; $p<0.01$) suggests that for the null model the probability that a default will occur without knowing the Rating is 4,4%. The variable *Rating* itself has a significant positive effect on *Default* (1.654; $p<0.01$). Hence, in case of a one-unit increase in the Bondora rating (i.e., a worse credit score) the chance of default increases by 65.40%. Looking at the McFadden R^2 of the baseline model a value of 0.109 indicates that the *Rating* accounts for only 10.9% of the variation in *Default*. As a main proxy for credit risk used by borrowers on P2P platforms this confirms the intuition that additional predictors may be necessary to receive a more robust result. Here, it has to be noted that in general more predictors always increase the R^2 value and would suggest an improvement. To still receive robust results the R^2 is supplemented by additional measures that account for overfitting which are displayed in chapter 5.4.

5.2 Extended model

After regressing the default rate on the Bondora rating, the loan and demographic variables are added to logistic regression. Firstly, as mentioned already in chapter 3 the observations for borrowers residing in Finland and the gender undefined are inherently included in the intercept. As can be seen, all variables are significant at the 1% level. In line with the baseline model, the Rating has a significant and positive coefficient (1.341; $p<0.01$). Still, as to be expected due to more predictors being added, a one-unit increase in *Rating* leads to the Default increasing relatively by 34,1%, ceteris paribus.

For the loan characteristics the variable significant correlation between *Interest* and *DefaultRatio* suggests the importance for *Default*. As indicated *Interest* has a significant positive coefficient (1.011, $p<0{,}01\%$) with the default probability changing relatively by 1,1% ceteris paribus for a one-unit increase. For the *Amount* a significant positive coefficient is fitted. The default probability increases relatively by 0,008% ceteris paribus per additional Euro borrowed.

Looking at the borrower characteristics the country variables *Estonia* and *Spain* have a negative and positive fitted coefficient *(Estonia: 0.693* and *Spain*: 1.998). The model shows that for borrowers from Estonia the default probability decreases by 30,07% ceteris paribus while opposingly for Spanish borrowers defaulted loans have a 99,80% ceteris paribus higher chance to occur. As for *NewCustomre* lending to a new borrower on Bondora the probability of incurring default increase by 28,4% ceteris paribus. Also increasing the probability of default are the variables *ExistingLiabilities* and *Liabilities* with relative increases in default probability by 5,06% ceteris paribus and 0,04% ceteris paribus respectively for every additional liability engaged in and for every Euro paid more monthly. Lastly, the variables *VerifiedIncome Eduaction* and *HomeOwner* are all individually related to a decrease in default probability. Intuitively, a borrower that verifies its Income is associated with a relative decrease in default probability by 54,96% ceteris paribus (0.4504). Similarly, the variable *HomeOwner* is associated with a relative decrease in default probability of 20,56% ceteris paribus (0.7944). Moreover, McFadden's R^2 improves to 0.166 and hence increases the model fit on a first take compared to the base model.

5.3 Full model

After regressing the default rate on the Bondora rating, loan characteristics and personal information, the COVID-19 parameter *COVID* is added to the logistic regression. COVID has a significant negative coefficient that indicates a relative decrease in the default probability 87,63% (0,1236; $p<0.01$). At the same time other variables change

For the full model the McFadden's R^2 improves to 0.248 and ensures a good model fit. The explained variation in the model increases strongly relative to the change from the base model to the extended model suggesting a certain explanatory power for adding the COVID dummy.

5.3 Interaction model

Finally, the interaction model is computed. Here, the focus lies on the interaction terms which are set up between COVID and the variables used in the full model. Overall, the interaction terms *Liabilities*COVID*, *NewCustomer*COVID* and *HomeOwner*COVID* are not significant and will hence be eliminated from the model. This implies that the coefficients of *Liabilities*, *NewCustomer* and *HomeOwner* are not affected by the magnitude of the COVID variable. In contrast to that, the variables *Rating*, *Amount*, *Interest*, *ExistingLiabilities*, *Estonia*, *Spain*, *Male* and *Female* are significantly influenced by the magnitude of the *COVID* variable. In order to interpret the magnitude of the interaction effect the odds-ratio for the interaction term can be calculate by multiplying with the odds ratio of the interacted variable. Here, for *Rating* the odds ratio considering the interaction effect with *COVID* is equal to

$$\text{(XVII)} \quad \exp\left(\beta_{Rating}\right) * \exp\left(\beta_{Rating*COVID}\right) = 1.224 * 0.827 = 1.012$$

Hence, the relative change in default probability for a one unit jump in *Rating* decreases from 22,4% increase to a 1,2% in times of COVID. Based on the odds-ratios displayed in the Appendix 4 interaction effects can be computed. While the COVID variable dampens the effect of *Rating*, the odds ratio for variables such as *Spain* and *Male* are largely magnified for loans during COVID compared to pre-COVID increasing from 26,8% to 749% and -9,4% to 149% respectively.

5.4 Model comparison

Additional to the McFadden's R^2 measure used to describe model relevance other measures are introduced specially to correct for overfitting the model. By using the same training and testing dataset for all four logistic regression models it is possible to make a comparison based on their accuracy. All measures for model comparison are included in Table 7. The accuracy as well as the LRT, AIC, and BIC of the performed empirical models are summarized in Table 6. It becomes clear that the full model, as expected, outperforms the Base model and Extended model across all validation measures. The full model can correctly predict loan default in 75% of all cases and is therefore economically significant. Hence, loan characteristics, demographics and controlling for the COVID pandemic help to better understand customer default risk.

The interaction terms

Measure	**Base Model**	**Extended Model**	**Full Model**	**Interaction Model**
McFaddens R^2	0.109	0.166	0.248	0.262
LRT	-98,150	-91,835	-82,783	-81,250
AIC	196,304	183,704	165,600	162,557
BIC	196,324.9	183,864.6	165,771.3	162,838.6
Accuracy	0.592	0.735	0.753	0.757

Table 7: Model Comparison

6. Conclusion

This section provides a summary of the previous results in regard to the hypotheses formed in chapter 2. Furthermore, the limitations as well as an outlook on future research will be provided.

Generally, the accurate prediction of borrower default is crucial for both P2P lending platforms and borrowers seeking to dampen the economic setbacks by the COVID-19 pandemic. The goal of this study was to identify whether internally available default covariates exist to improve the assessment of the inherent credit risk in P2P lending. Additionally, and at least as important, this study displayed a first approach on analysing the extend of ongoing COVID-19 pandemic on credit risk predictors for P2P lending markets. Hence, I would like to come back to the initially formulated research question.

"Is the credit risk assessment in Peer-to-Peer Lending affected by the COVID-19 pandemic?"

The answer can be provided by relating the results from the previous chapter to the formulated hypotheses. The results of the extended model show that *Internally available Loan and borrower characteristics improve default rate prediction in P2P lending.* Additional variables to the baseline model showed to be significant increase in the predictive power as well as goodness of fit. Similarly, analysing the COVID-19 variable show a significant correlation to the default rate for P2P loans with a relative decrease ceteris paribus of 87.63% for the default probability in a COVID environment compared to pre-COVID. Hence *the harsh economic environment induced by the COVID-19 pandemic is significantly correlated with P2P lending default rates.* Lastly, the significant interaction terms computed in the Interaction model show that t*he correlation of loan and borrower characteristics on default probability is influenced by the COVID-19 pandemic.* Overall, the performed analysis can confirm that like many asset classes the inherent risk for P2P is affected to a certain extend by the COVID-19 pandemic.

With the COVID-19 pandemic still unfolding the current results provide a first empirical approach on the impact credit risk measurement. With an average loan duration of 49 months and the pandemic persisting for about 30 months now it is still unclear how the effects will unfold in the future. In general, this paper also stresses the importance of including borrower and loan characteristics to assess default probability for P2P loans. Especially in times of crisis when structural changes in credit risk occur, reliable and transparent access to capital can be an important source of economic growth.

Lastly, this research can serve as a reference point for future researchers. With continuously new data available the results from this study could be relativize or strengthened by similar research on future data and different platforms. Furthermore, additional external data can and should be analysed in this context to receive a broader picture on P2P lending drivers and the impact of changing economic circumstances.

.

II References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488. https://doi.org/10.2307/1879431

Augustin, P., Sokolovski, V., Subrahmanyam, M. G., & Tomio, D. (2021). In sickness and in debt: The COVID-19 impact on sovereign credit risk. *Journal of Financial Economics.* Advance online publication. https://doi.org/10.1016/j.jfineco.2021.05.009

Aveni, Qu, C., Hsu, K., Zhang, A., Lei, X., & Hemrika, L. (2015). *New Insights Into An Evolving P2P Lending Industry: how shifts in roles and risk are shaping the industry*. https://www. findevgateway. org/sites/default/files/publication_files/new_insights_into_an_evolving_p2p_lending_industry_positiveplanet2015. pdf

Bachmann, Becker, Buerckner, Hilker, & Funk (2011). Online Peer-to-Peer Lending--A Literature. *Journal of Internet Banking and Commerce*, *16*. https://www.researchgate.net/profile/burkhardt-funk/publication/236735575_online_peer-to-peer_lending--a_literature

Beáta Gavurová, Martin Dujcak, Viliam Kovál\vc, & Anna Kotásková (2018). Determinants of Successful Loan Application at Peer-to-Peer Lending Market. *Economics & Sociology*, *11*, 85–99.

Berger, S. C., & Gleisner, F. (2009). Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. *Business Research*, *2*(1), 39–65. https://doi.org/10.1007/BF03343528

Bondora.com. (2022). *Public Statistics*. https://www.bondora.com/en/public-statistics

Dubinova, A., Lucas, A., & Telg, S. (2021). COVID-19, Credit Risk and Macro Fundamentals. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3875628

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Faia, E., & Paiella, M. (2017). P2P Lending: Information Externalities, Social Networks and Loans' Substitution. *Social Networks and Loans' Substitution (August 2017). CEPR Discussion Paper No. DP12235*.

Gholamy, Kreinovich, & Kosheleva (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. In

Goodell, J. W. (2020). Covid-19 and finance: Agendas for future research. *Finance Research Letters*, *35*, 101512. https://doi.org/10.1016/j.frl.2020.101512

Johns Hopkins University (2022): Corona Virus Resource Center. Online verfügbar unter https://coronavirus.jhu.edu/.

Long, J. S., & Freese, J. (2006). *Regression models for categorical and limited dependent variables using data* (2nd ed.). *Advanced quantitative techniques in the social sciences: Vol. 7*. State Press.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. *Adaptive computation and machine learning series*. MIT Press.

Najaf, K., Subramaniam, R. K., & Atayah, O. F. (2021). Understanding the implications of FinTech Peer-to-Peer (P2P) lending during the COVID-19 pandemic. *Journal of Sustainable Finance & Investment*, 1–16. https://doi.org/10.1080/20430795.2021.1917225

Ölvedi, T. (2020). An overview of peer-to-peer lending. *Economy & Finance*, *7*(2), 218–232. https://doi.org/10.33908/EF.2020.2.6

Philippon, T. (2016). The FinTech Opportunity. *National Bureau of Economic Research.* Advance online publication. https://doi.org/10.3386/w22476

Roure, C. de, Pelizzon, L., & Tasca, P. (2016). How Does P2P Lending Fit into the Consumer Credit Market? *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.2848043

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PloS One*, *10*(10), e0139427. https://doi.org/10.1371/journal.pone.0139427

Stiglitz, J., & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *American Economic Review*, *71*(3), 393–410. https://EconPapers.repec.org/RePEc:aea:aecrev:v:71:y:1981:i:3:p:393-410

Tang, H. (2019). Peer-to-Peer Lenders Versus Banks: Substitutes or Complements? *Review of Financial Studies*, *32*(5), 1900–1938. https://doi.org/10.1093/rfs/hhy137

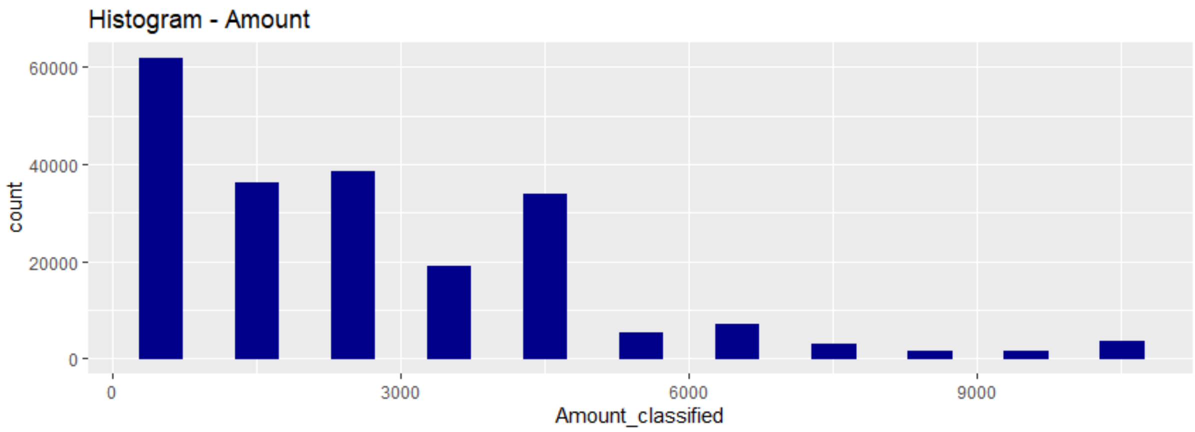
Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th edition). Cengage Learning.

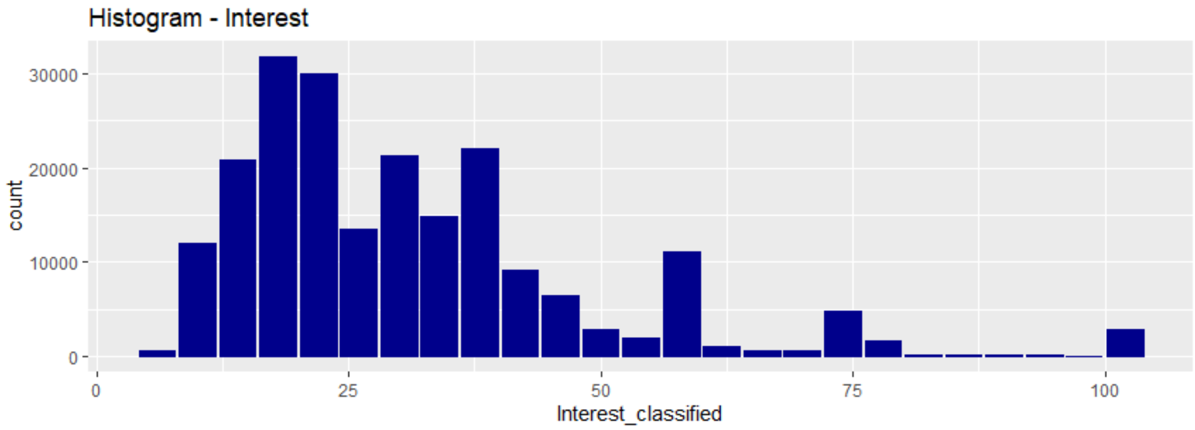
World Bank (2020): World Bank Database. Hg. v. World Bank. Online verfügbar unter https://data.worldbank.org/topic/7.

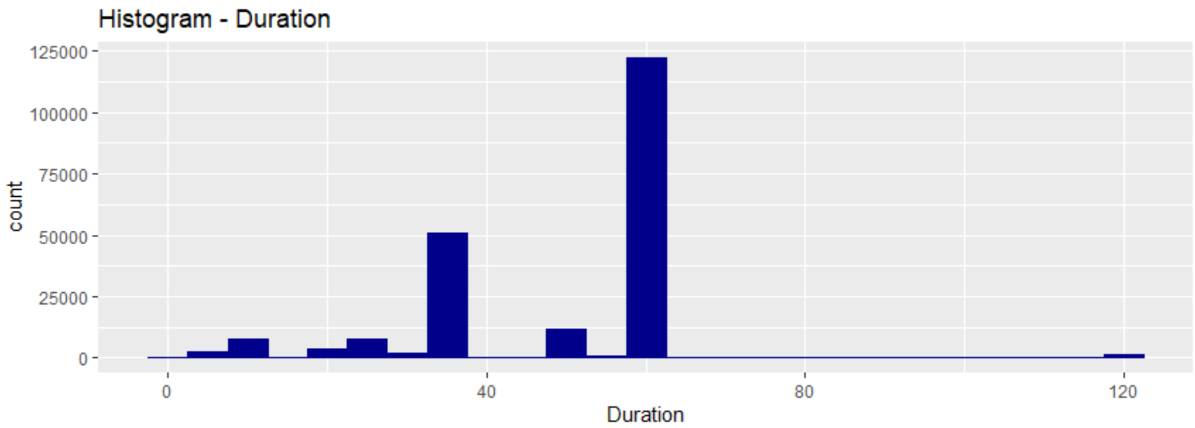
Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters*, *36*, 101528. https://doi.org/10.1016/j.frl.2020.101528

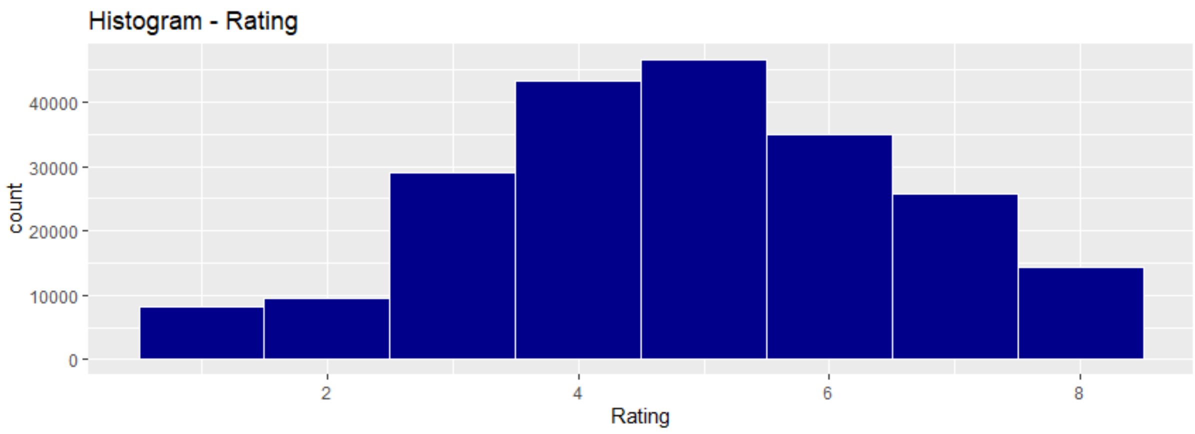
III Appendix

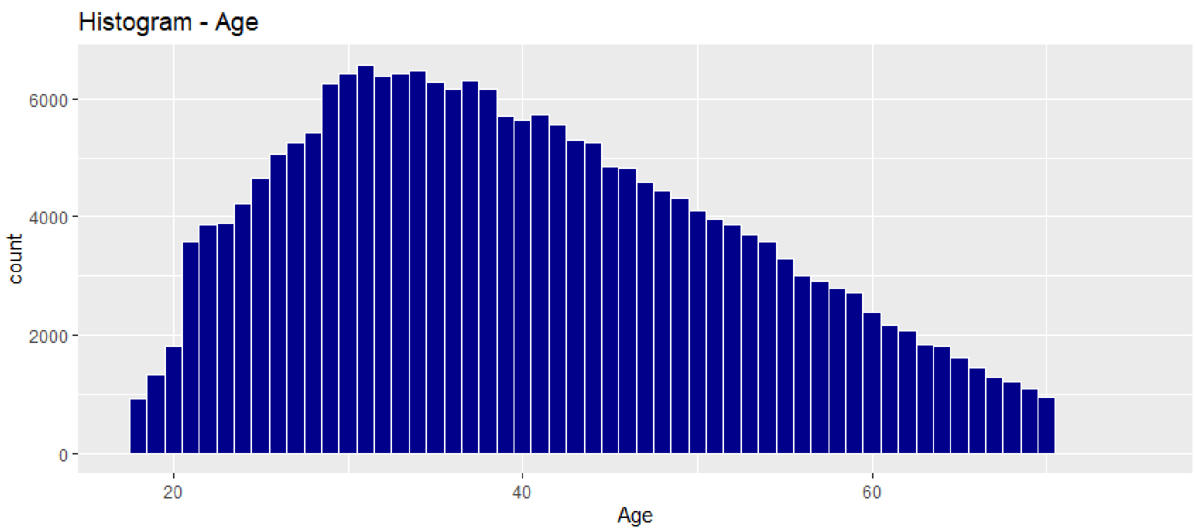
Appendix 1: Variable Discussion - Histograms

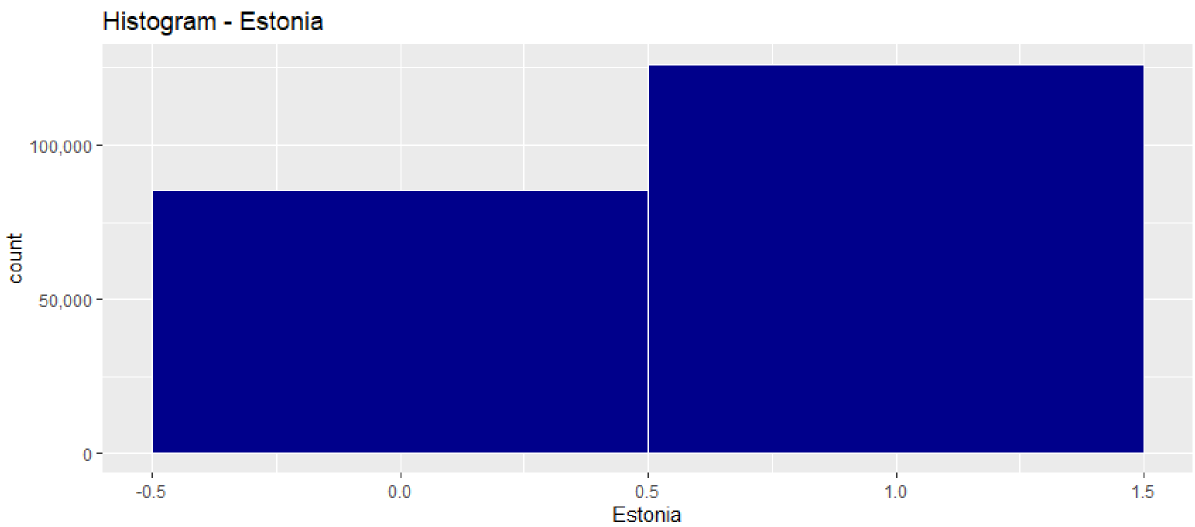


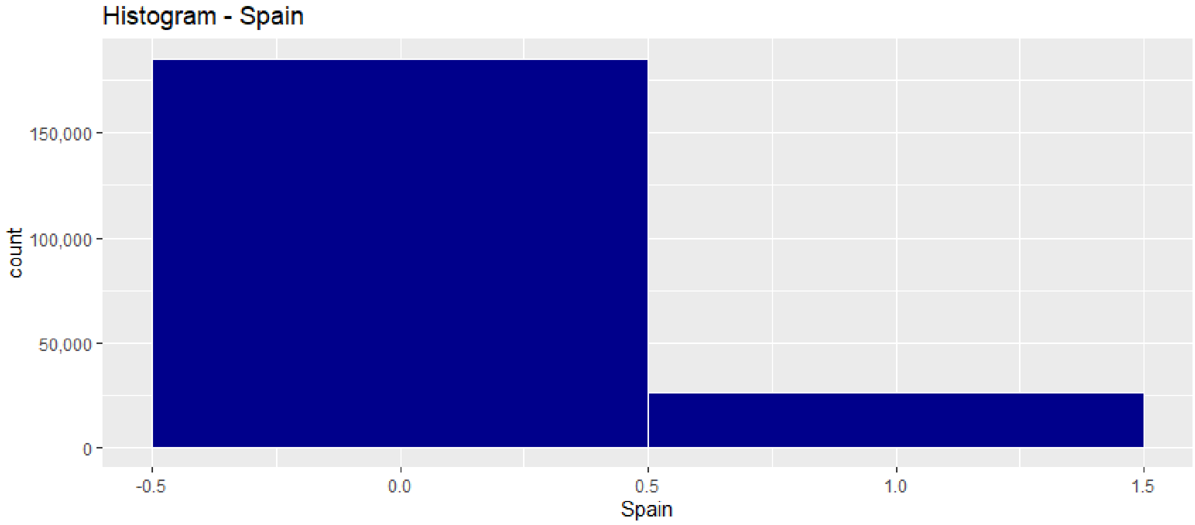


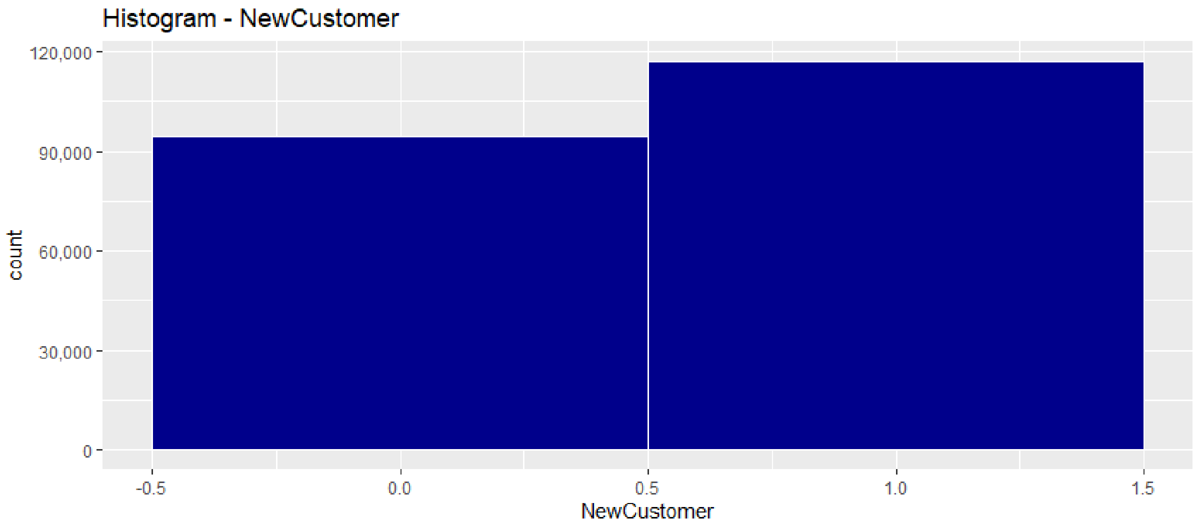


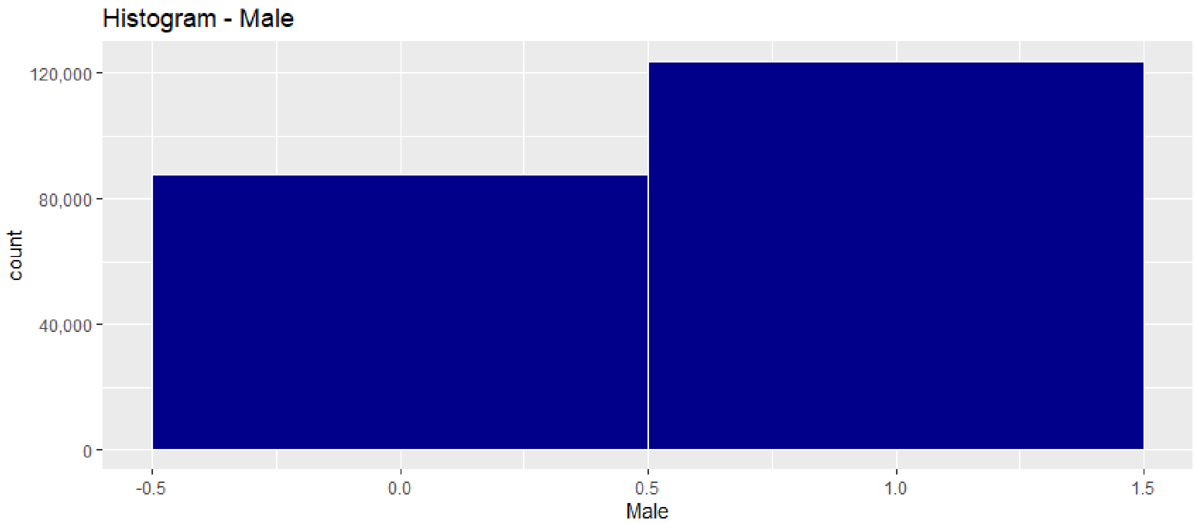

Histogram - Rating
count
40000
30000
20000
10000
0
2
4
6
8
Rating

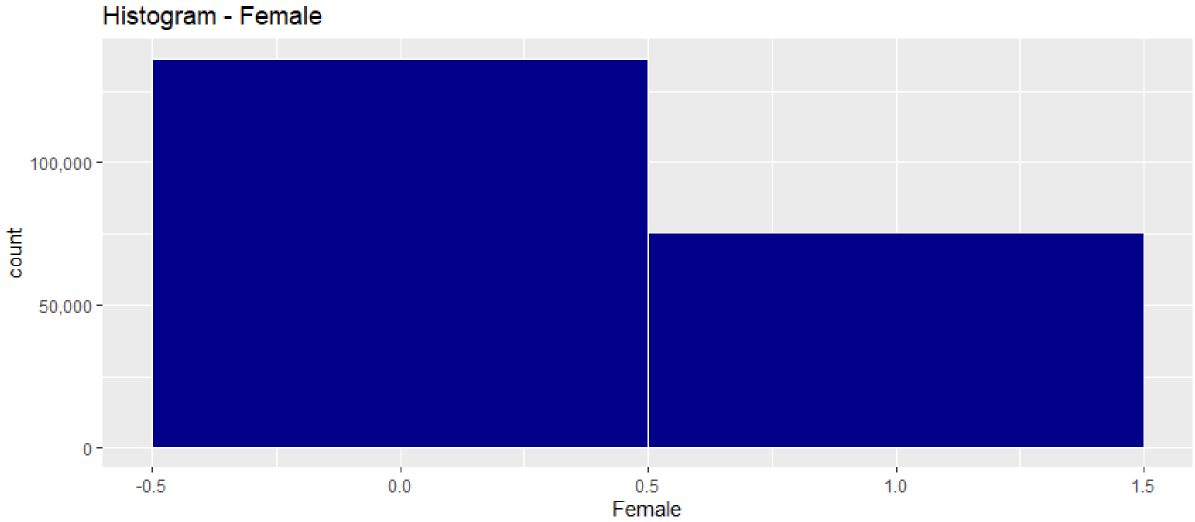

Histogram - Age
count
6000
4000
2000
0
20
40
60
Age

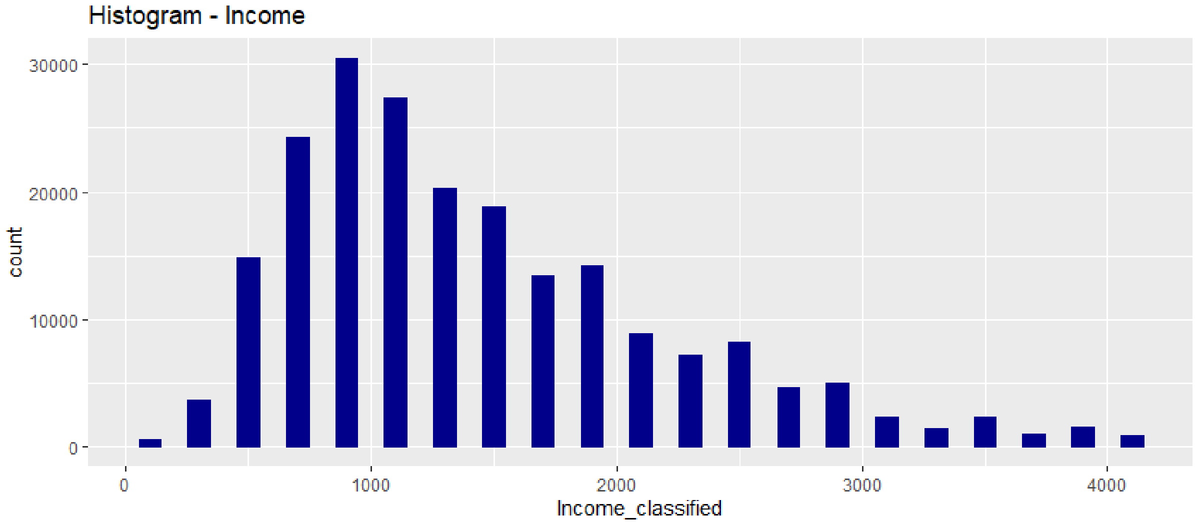

Histogram - Estonia
count
100,000
50,000
0
-0.5
0.0
0.5
1.0
1.5
Estonia

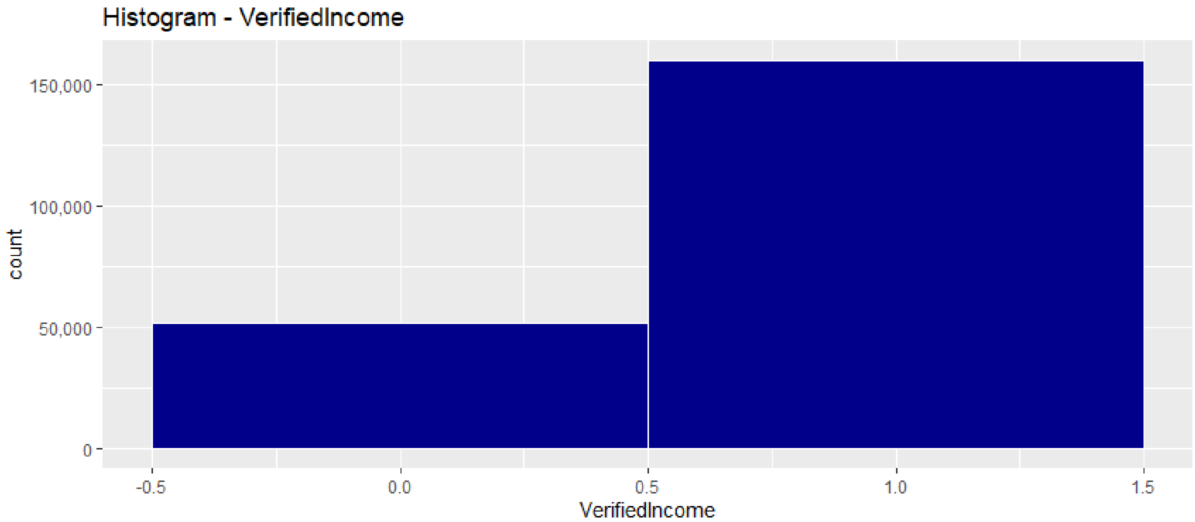

Histogram - Spain
150,000
100,000
50,000
0
count
-0.5
0.0
0.5
1.0
1.5
Spain

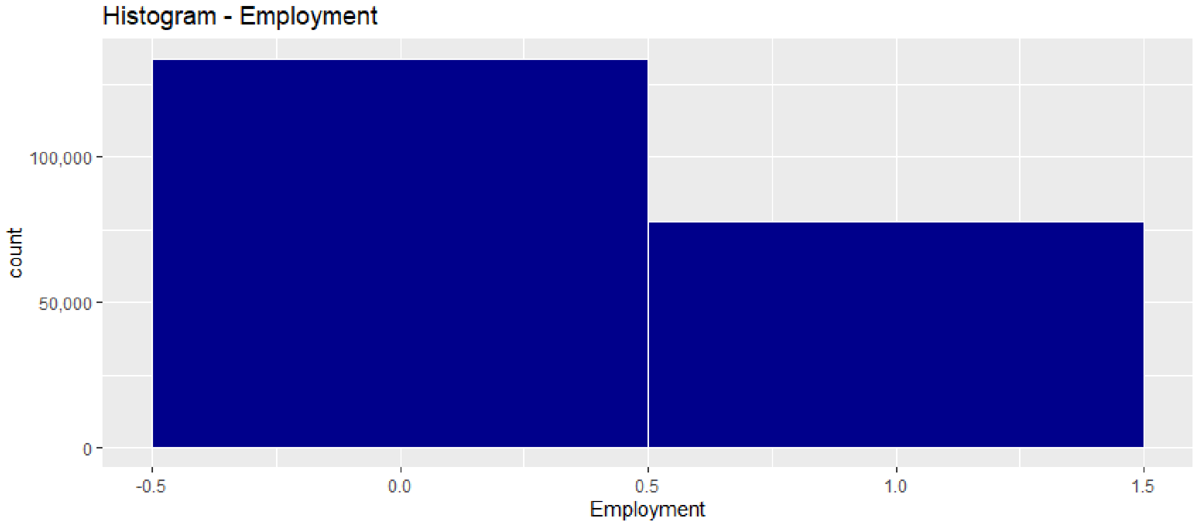

Histogram - NewCustomer
120,000
90,000
60,000
30,000
0
count
-0.5
0.0
0.5
1.0
1.5
NewCustomer

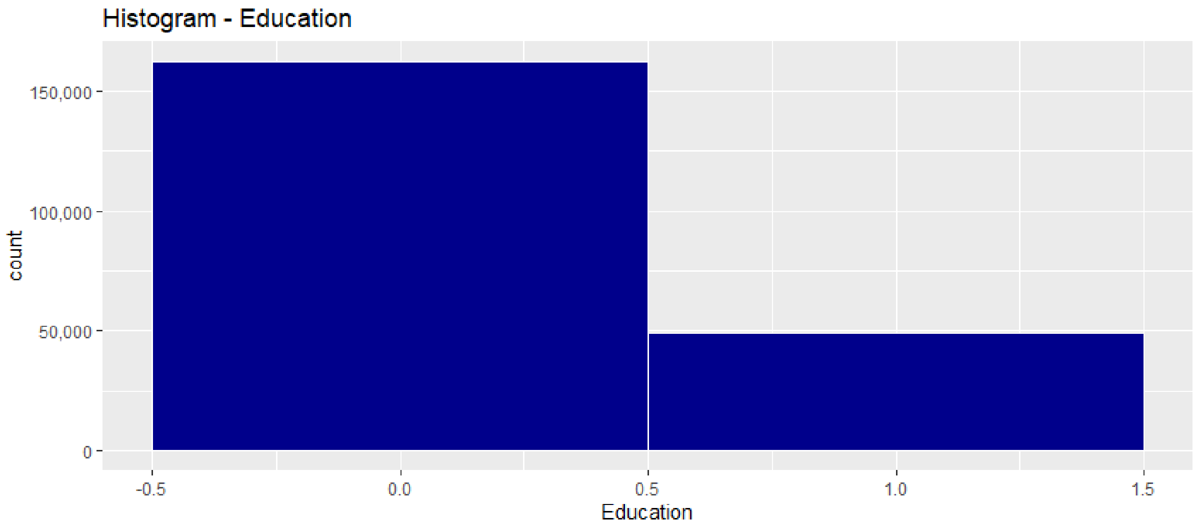

Histogram - Male
120,000
80,000
40,000
0
count
-0.5
0.0
0.5
1.0
1.5
Male

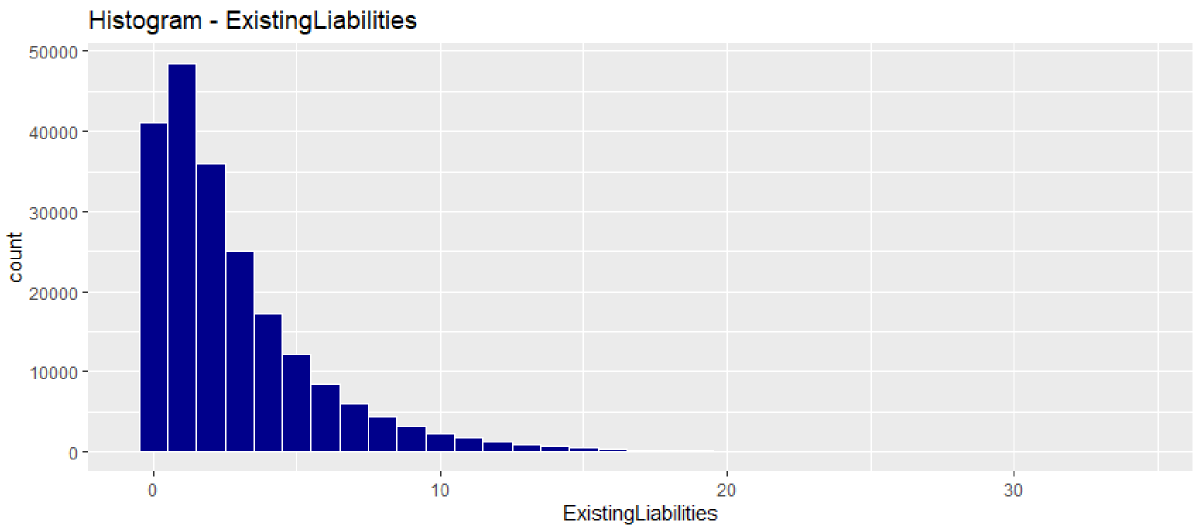

Histogram - Female
count
100,000
50,000
0
-0.5
0.0
0.5
1.0
1.5
Female

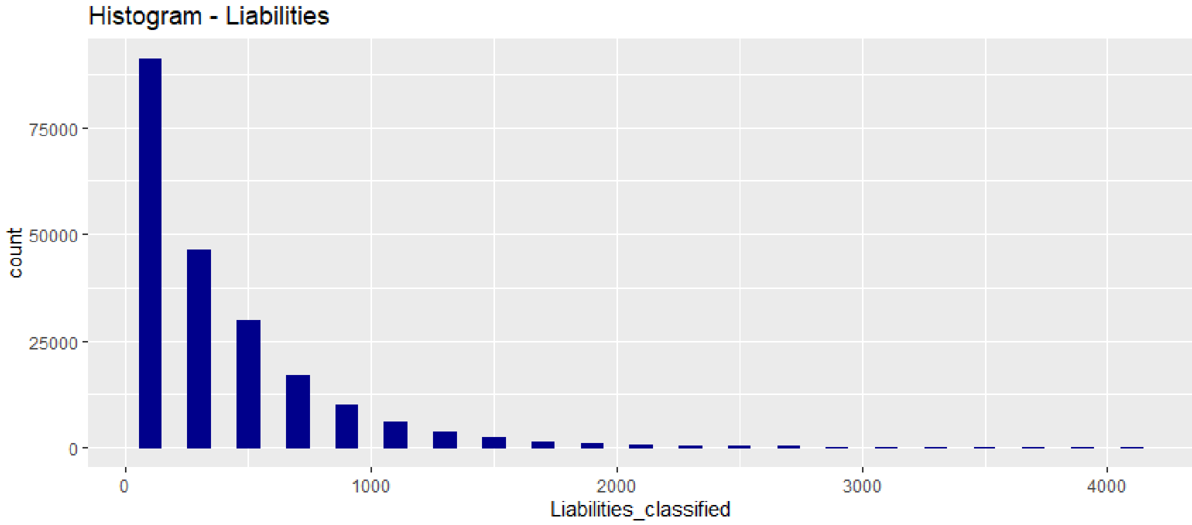

Histogram - Income
count
30000
20000
10000
0
0
1000
2000
3000
4000
Income_classified

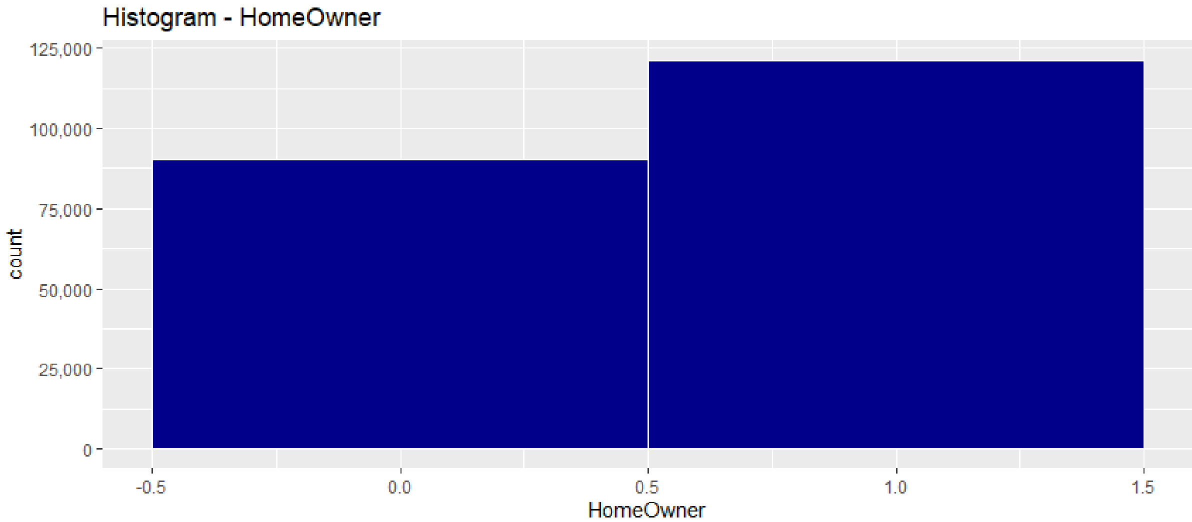

Histogram - VerifiedIncome
count
150,000
100,000
50,000
0
-0.5
0.0
0.5
1.0
1.5
VerifiedIncome

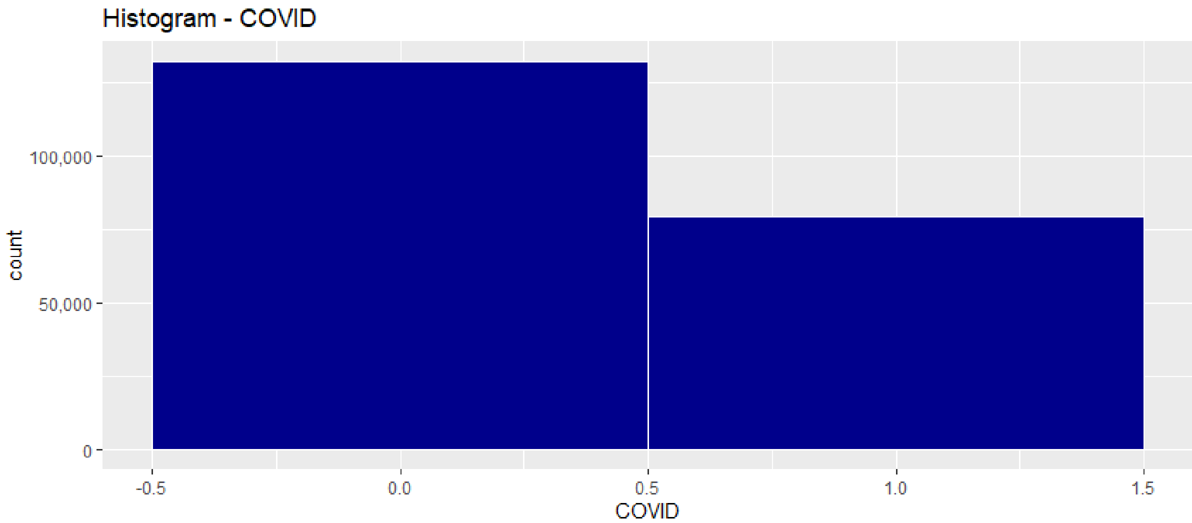

Histogram - Employment
100,000
50,000
0
count
-0.5
0.0
0.5
1.0
1.5
Employment


Histogram - Education
150,000
100,000
50,000
0
count
-0.5
0.0
0.5
1.0
1.5
Education

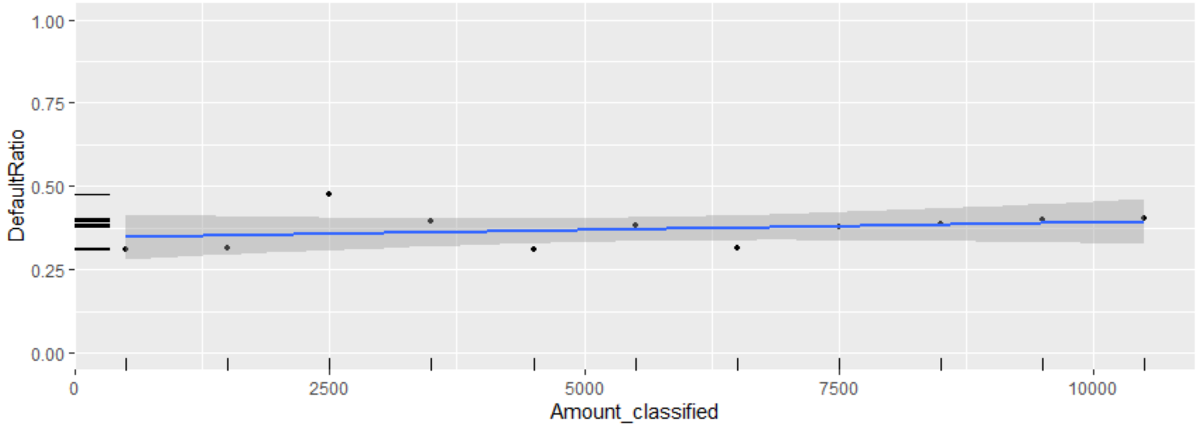

Histogram - ExistingLiabilities
50000
40000
30000
20000
10000
0
count
0
10
20
30
ExistingLiabilities

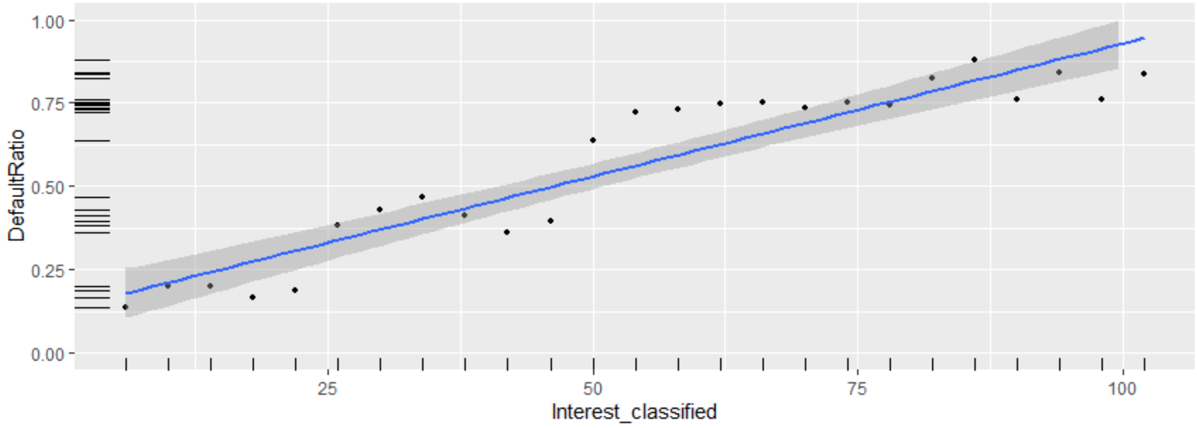

Histogram - Liabilities
count
75000
50000
25000
0
0
1000
2000
3000
4000
Liabilities_classified

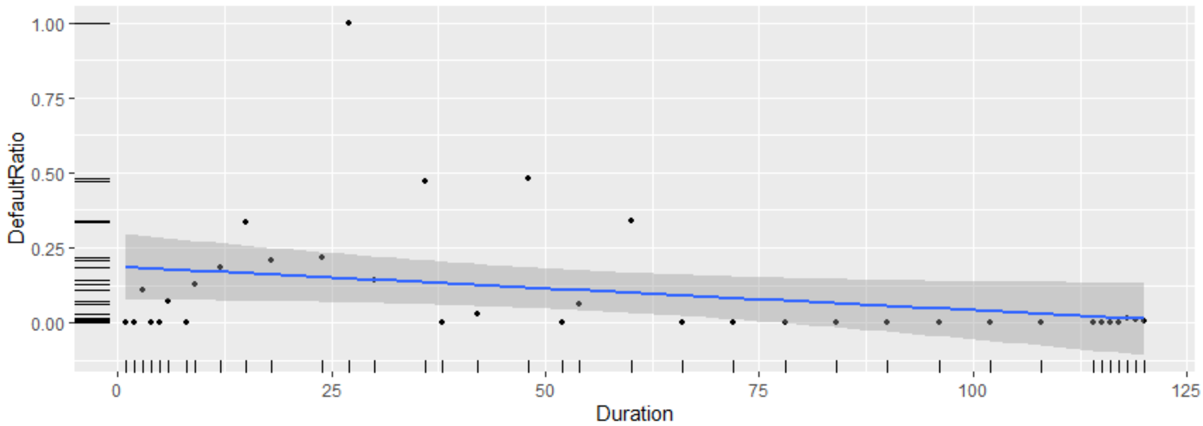

Histogram - HomeOwner
count
125,000
100,000
75,000
50,000
25,000
0
-0.5
0.0
0.5
1.0
1.5
HomeOwner

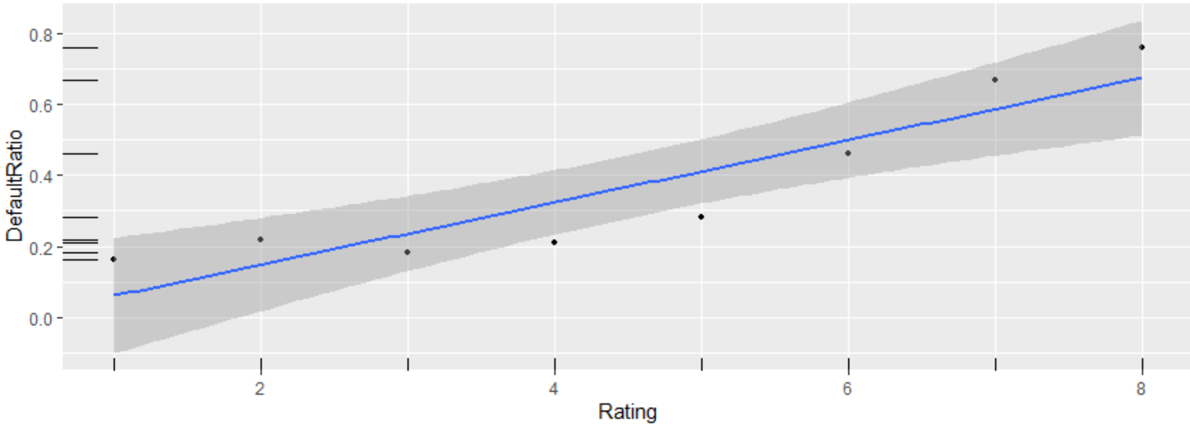

Histogram - COVID
count
100,000
50,000
0
-0.5
0.0
0.5
1.0
1.5
COVID

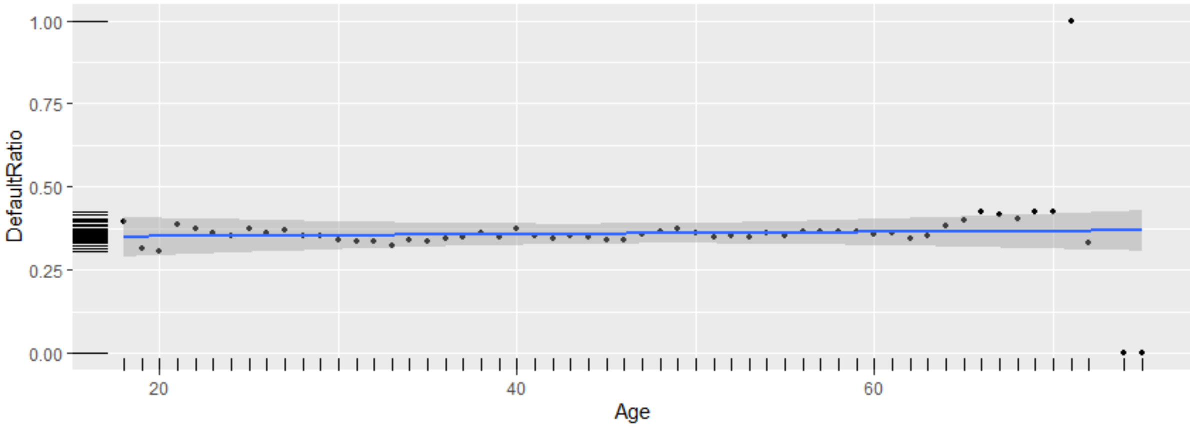
Appendix 2: Variable Discussion - DefaultRatio

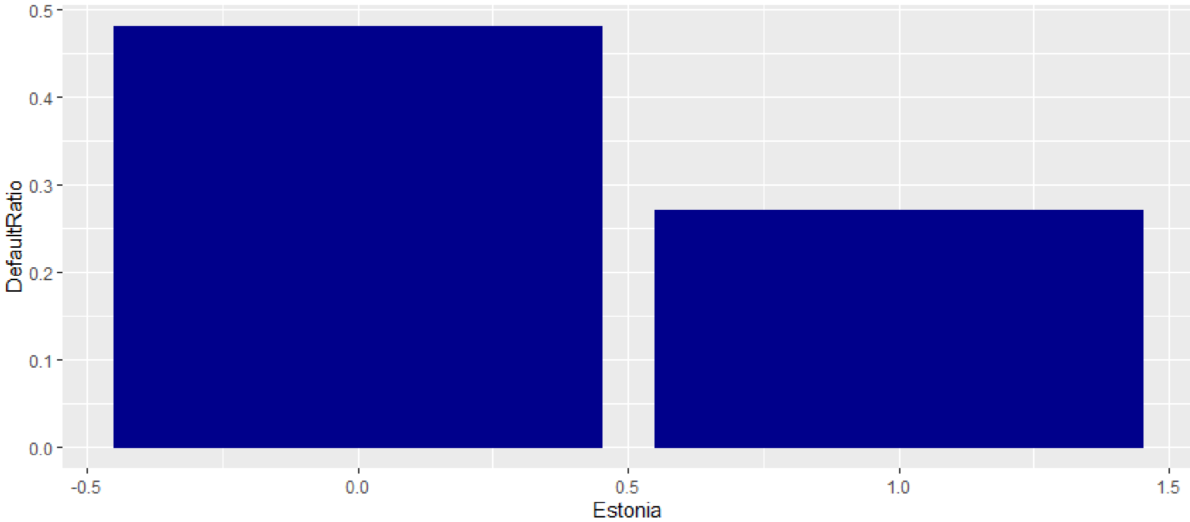


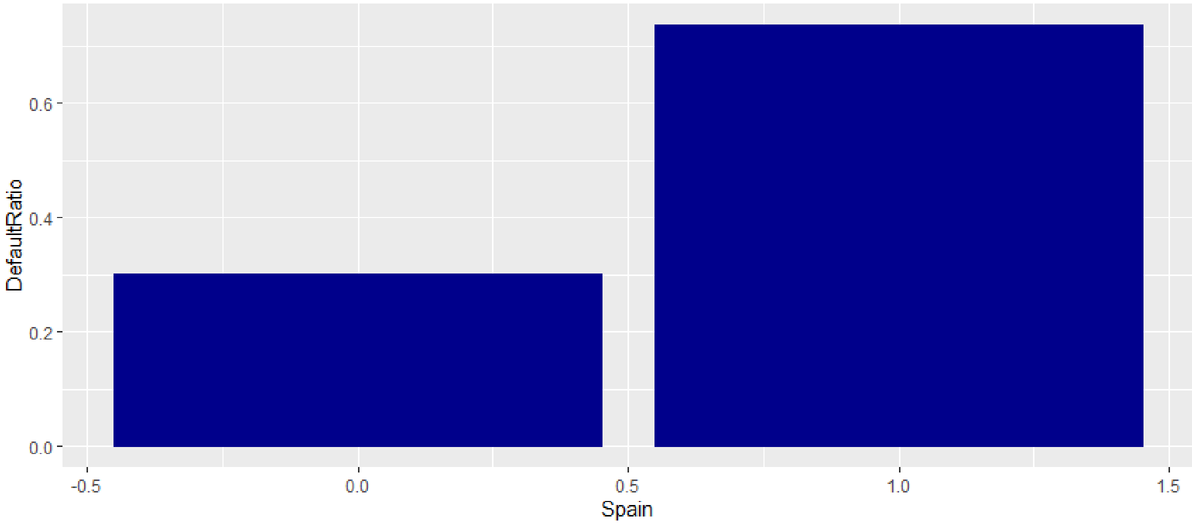




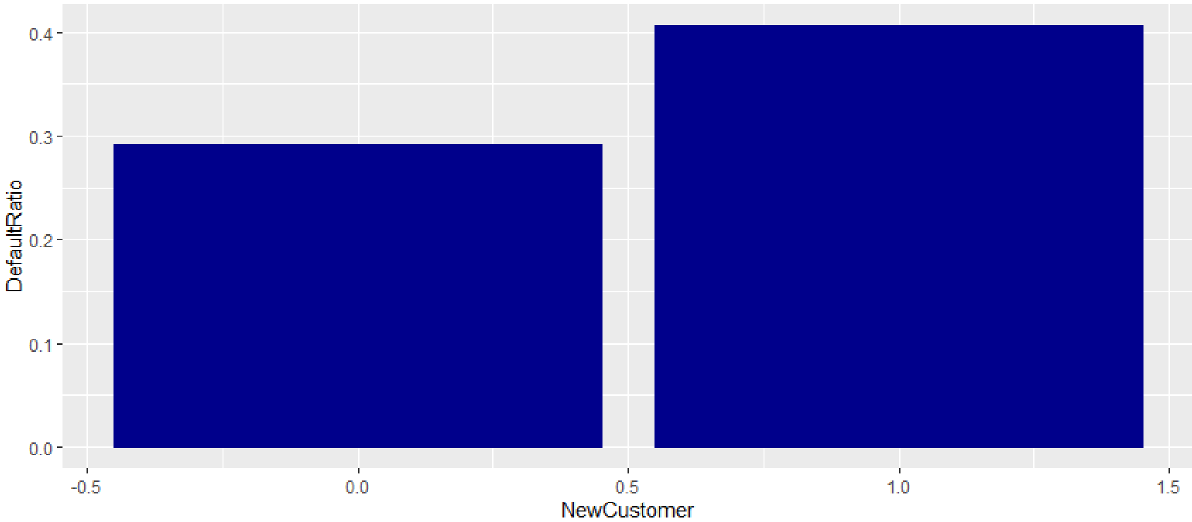

0.8
0.6
0.4
0.2
0.0
DefaultRatio
2
4
6
8
Rating


1.00
0.75
0.50
0.25
0.00
DefaultRatio
20
40
60
Age

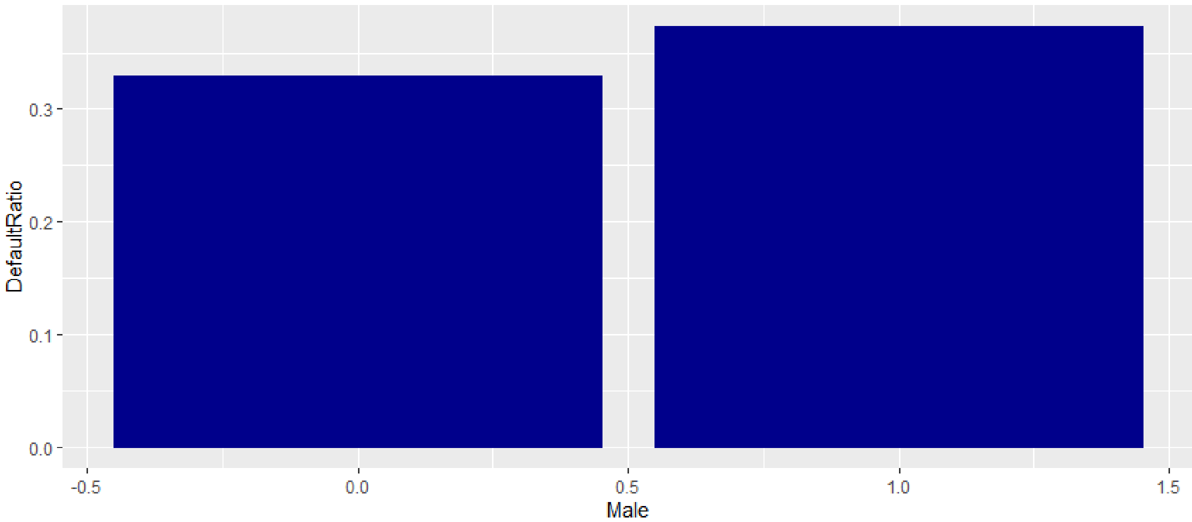

0.5
0.4
0.3
0.2
0.1
0.0
DefaultRatio
-0.5
0.0
0.5
1.0
1.5
Estonia



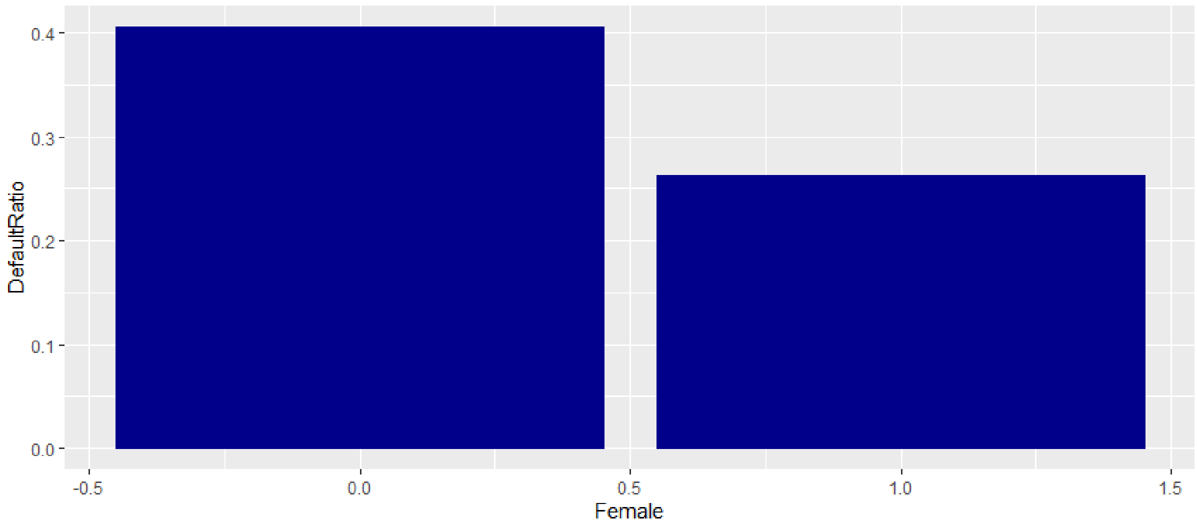
DefaultRatio
0.0
0.2
0.4
0.6
-0.5
0.0
0.5
1.0
1.5
Spain

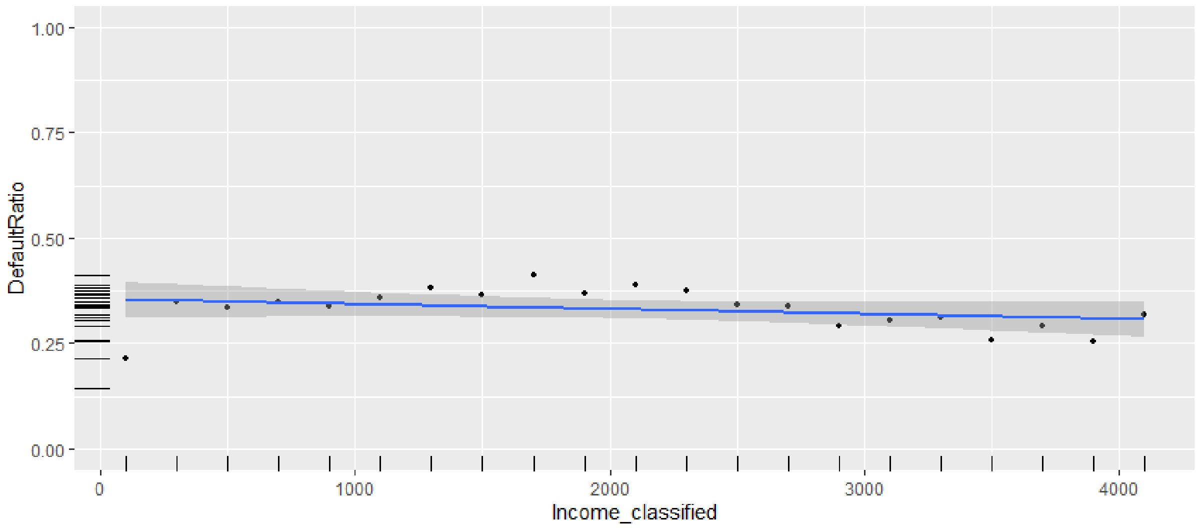


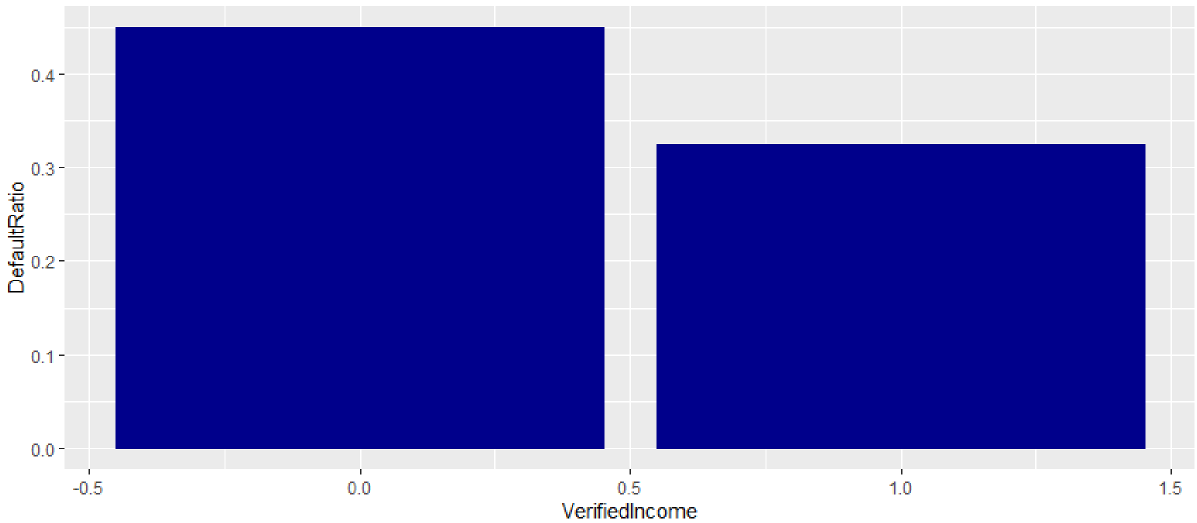
DefaultRatio
0.0
0.1
0.2
0.3
0.4
-0.5
0.0
0.5
1.0
1.5
NewCustomer

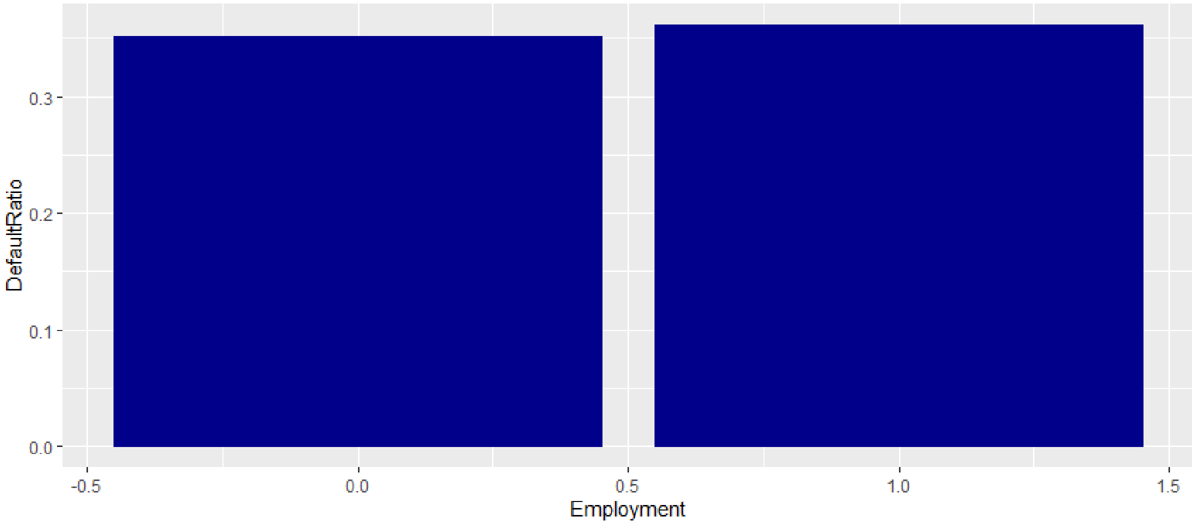


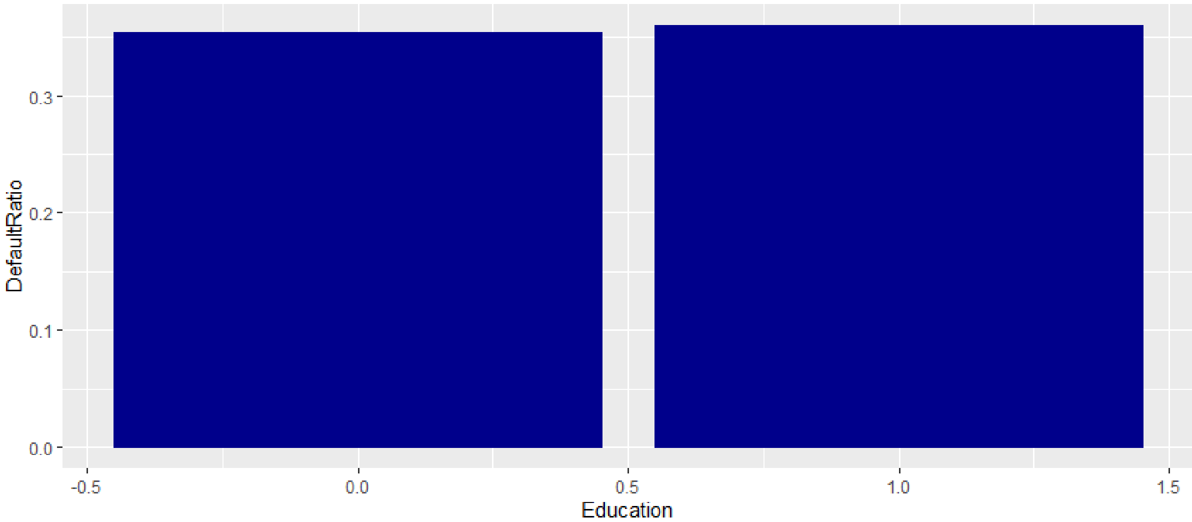
DefaultRatio
0.0
0.1
0.2
0.3
-0.5
0.0
0.5
1.0
1.5
Male

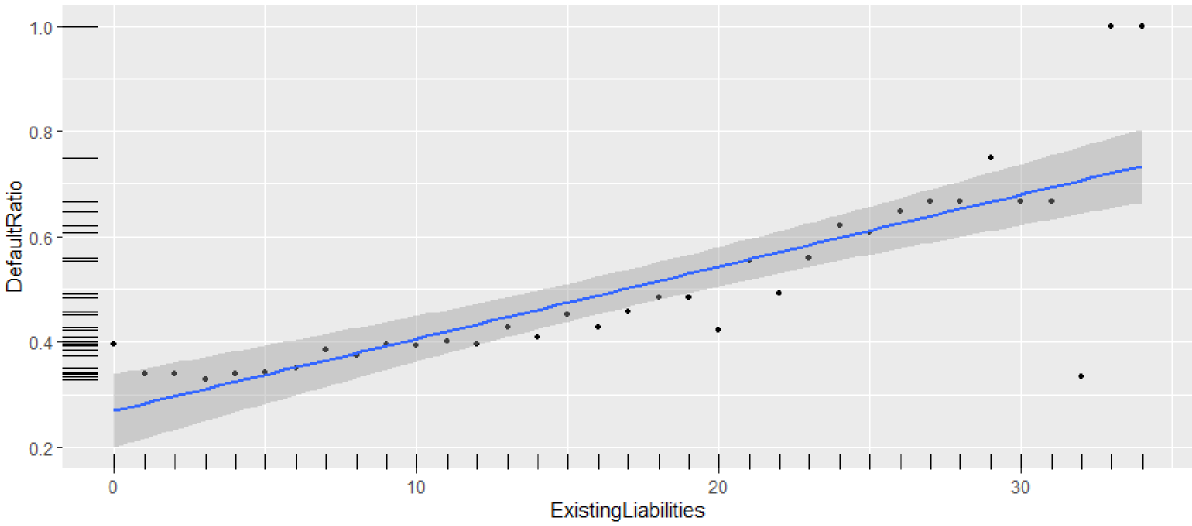

DefaultRatio
0.0
0.1
0.2
0.3
0.4
-0.5
0.0
0.5
1.0
1.5
Female

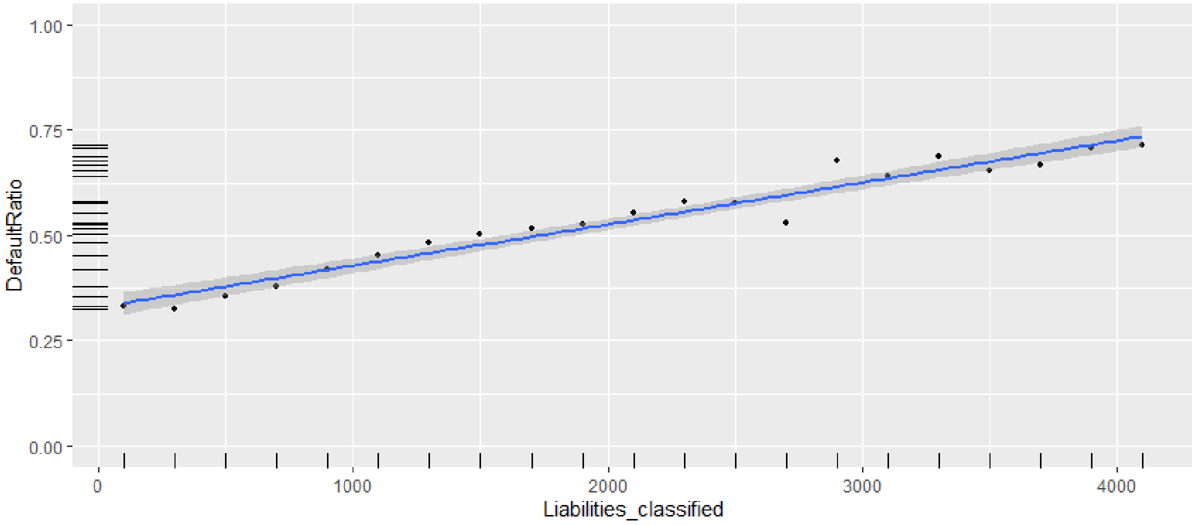

DefaultRatio
0.00
0.25
0.50
0.75
1.00
0
1000
2000
3000
4000
Income_classified

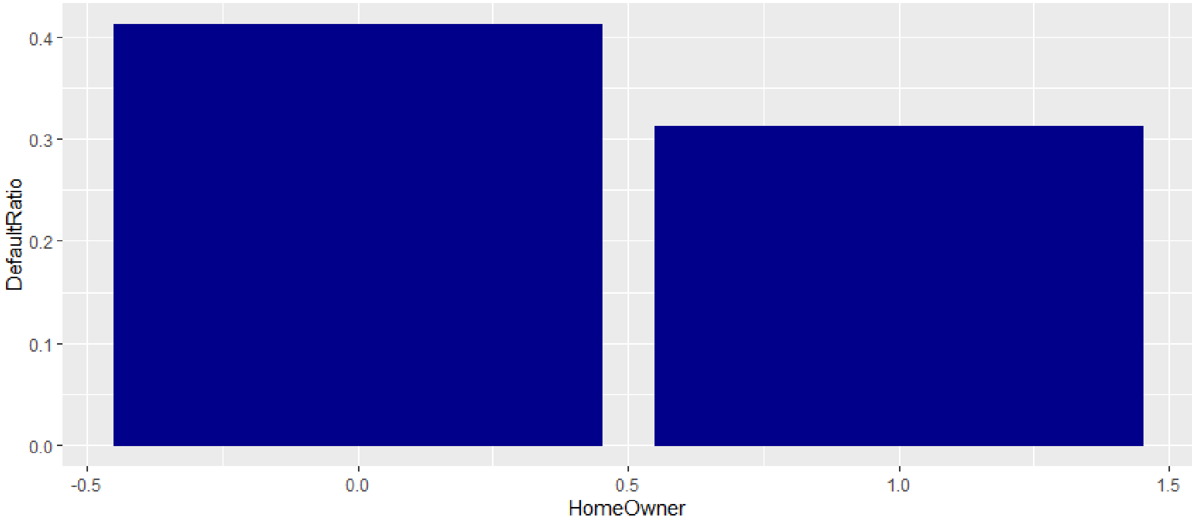

DefaultRatio
0.0
0.1
0.2
0.3
0.4
-0.5
0.0
0.5
1.0
1.5
VerifiedIncome

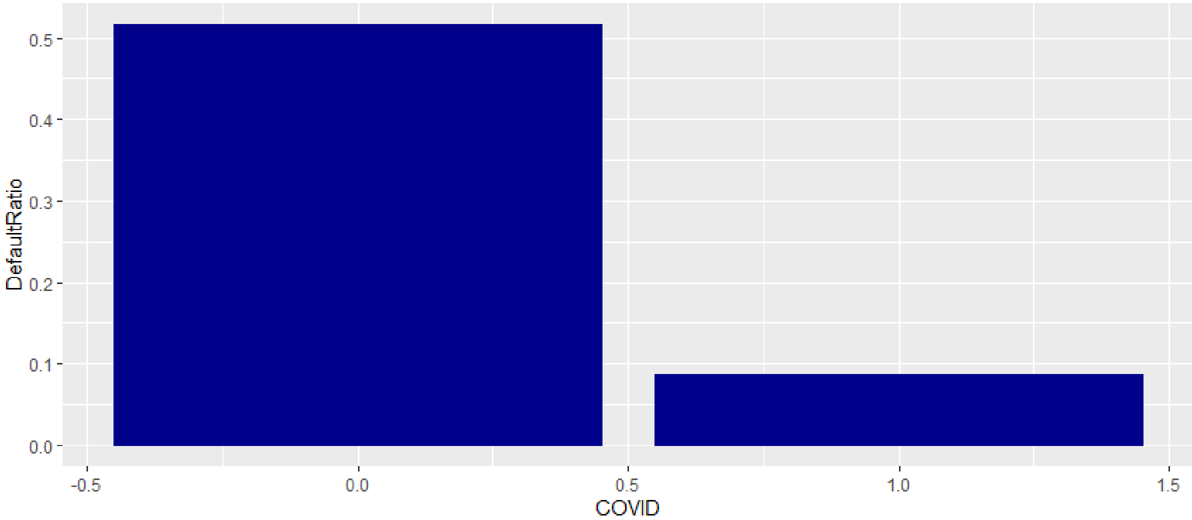

DefaultRatio
0.0
0.1
0.2
0.3
-0.5
0.0
0.5
1.0
1.5
Employment


DefaultRatio
0.0
0.1
0.2
0.3
-0.5
0.0
0.5
1.0
1.5
Education


DefaultRatio
0.2
0.4
0.6
0.8
1.0
0
10
20
30
ExistingLiabilities


1.00
0.75
0.50
0.25
0.00
DefaultRatio
0
1000
2000
3000
4000
Liabilities_classified


0.4
0.3
0.2
0.1
0.0
DefaultRatio
-0.5
0.0
0.5
1.0
1.5
HomeOwner


0.5
0.4
0.3
0.2
0.1
0.0
DefaultRatio
-0.5
0.0
0.5
1.0
1.5
COVID

Appendix 3: Logistic Regression Models - Results

	Dependent variable:			
	Default			
	(1)	(2)	(3)	(4)
Rating	0.503***	0.293***	0.244***	0.203***
	(0.004)	(0.006)	(0.006)	(0.006)
Amount		0.0001***	0.0001***	0.0001***
		(0.00000)	(0.00000)	(0.00000)
Interest		0.011***	0.007***	0.003***
		(0.001)	(0.001)	(0.0005)
Estonia		-0.366***	-0.379***	-0.683***
		(0.017)	(0.019)	(0.022)
Spain		0.692***	0.339***	0.236***
		(0.028)	(0.029)	(0.029)
NewCustomer		0.250***	0.034**	0.094***
		(0.013)	(0.014)	(0.016)
Male		-0.064*	-0.305***	-0.100***
		(0.033)	(0.035)	(0.034)
Female		-0.516***	-0.561***	-0.363***
		(0.035)	(0.036)	(0.037)
Income		-0.0004***	-0.0002***	-0.0002***
		(0.00001)	(0.00001)	(0.00001)
VerifiedIncome		-0.798***	-0.095***	-0.077***
		(0.013)	(0.014)	(0.014)
Employment		-0.002	-0.067***	-0.064***

		(0.012)	(0.013)	(0.014)
Education		-0.172***	-0.304***	-0.334***
		(0.015)	(0.015)	(0.016)
ExistingLiabilities		0.049***	0.019***	0.023***
		(0.003)	(0.003)	(0.003)
Liabilities		0.0004***	0.0002***	0.0001***
		(0.00002)	(0.00002)	(0.00002)
HomeOwner		-0.230***	-0.226***	-0.229***
		(0.012)	(0.013)	(0.014)
COVID			-2.097***	-4.613***
			(0.017)	(0.215)
Rating:COVID				-0.188***
				(0.029)
Amount:COVID				0.00004***
				(0.00001)
Interest:COVID				0.073***
				(0.003)
Income:COVID				-0.0001**
				(0.00003)
VerifiedIncome:COVID				-0.182***
				(0.069)
Employment:COVID				-0.084**
				(0.040)
Education:COVID				0.116**
				(0.050)

ExistingLiabilities:COVID				-0.093***
				(0.012)
Liabilities:COVID				-0.00001
				(0.0001)
Estonia:COVID				0.398***
				(0.057)
Spain:COVID				1.922***
				(0.154)
NewCustomer:COVID				0.060
				(0.040)
Male:COVID				1.025***
				(0.165)
Female:COVID				1.234***
				(0.168)
HomeOwner:COVID				-0.006
				(0.036)
Constant	-3.134***	-1.462***	-0.679***	-0.343***
	(0.019)	(0.050)	(0.053)	(0.056)
Observations	169,242	169,242	169,242	169,242
Log Likelihood	-98,150.440	-91,851.380	-82,755.850	-81,006.160
Akaike Inf. Crit.	196,304.900	183,734.800	165,545.700	162,076.300
Note:				*p**p***p<0.01

Appendix 4: Odds Ratio for Interaction Effects

	Estimate	Std. Error	z value	Pr(> | z|)
(Intercept)	0.708	1.058	0.002	1
Rating	1.224	1.006	1,014,475,646,263,525.000	1
Amount	1.000	1.000	59,304,832.000	1
Interest	1.003	1.000	321.289	1
Income	1.000	1.000	0.00000	1
VerifiedIncome	0.924	1.015	0.004	1.000
Employment	0.938	1.014	0.011	1.000
Education	0.716	1.016	0	1
ExistingLiabilities	1.024	1.003	3,275.563	1
Liabilities	1.000	1.000	909.010	1
Estonia	0.505	1.022	0	1
Spain	1.268	1.030	3,266.143	1
NewCustomer	1.109	1.015	1,267.093	1
Male	0.906	1.035	0.057	1.004
Female	0.696	1.037	0.00005	1
HomeOwner	0.794	1.013	0	1
COVID	0.011	1.232	0	1
Rating:COVID	0.828	1.029	0.001	1
Amount:COVID	1.000	1.000	340.144	1
Interest:COVID	1.076	1.003	24,062,323,848.000	1
Income:COVID	1.000	1.000	0.093	1.018
VerifiedIncome:COVID	0.833	1.071	0.070	1.008
Employment:COVID	0.917	1.040	0.109	1.027
Education:COVID	1.121	1.051	9.890	1.022

ExistingLiabilities:COVID	0.905	1.009	0.00001	1
Estonia:COVID	1.471	1.058	951.330	1
Spain:COVID	6.703	1.166	245,287.100	1
Male:COVID	2.753	1.179	469.811	1
Female:COVID	3.391	1.183	1,428.522	1