

**Master Analyse et politique économique**

**Parcours Statistique pour l'évaluation et la prévision**

**Cours de modèles linéaires et d'économétrie**

---

**Modélisation et prédiction du montant à prêter à un emprunteur pour un projet de lancement d'une plateforme de peer-to-peer lending.**

**Ismael Djoulde DIALLO**

---

**-Chargés du cours et des travaux dirigés : Mme Emmanuelle GAUTHERAT et Mr FATIH Zakaria**

**-Responsable du parcours : Mme Emmanuelle GAUTHERAT, Enseignant chercheur et maître de conférences.**

**Année universitaire 2021-2022**

## TABLES DES MATIERES

I- Objectif de l'étude .....	3
II- Présentation de Bondora.....	4
III- Nettoyage et préparation des données.....	5
a. Aperçu de la base de données brutes avant traitement et nettoyage .....	6
a. Description des variables retenues pour l'étude .....	11
b. Aperçu des données nettoyées .....	13
IV- Analyse exploratoire des données retenues pour le projet.....	14
b. Représentation graphique de la distribution de certaines variables .....	16
c. Analyse de la corrélation entre les variables.....	21
V- Modélisation de la variable Amount .....	23
Valeurs prédites de Amount vs valeurs observées.....	28
d. Statistiques des valeurs prédites de Amount.....	28
VI- Conclusion.....	29

## **I- Objectif de l'étude**

Pour de nombreux experts financiers, l'économie de plateforme (marchés biface) est le futur de l'industrie financière. De ce fait, nous souhaitons profiter du potentiel financier de ce secteur en y développant notre business.

Face à la difficulté pour de nombreuses personnes d'obtenir des prêts, nous souhaitons mettre en place une plateforme de prêts entre particuliers «peer-to-peer lending : P&P lending» afin de leur permettre d'obtenir des financements pour leurs différents projets. Pour ce faire, nous nous inspirons du modèle d'affaires de Bondora.

Par ailleurs, rappelons que le principal obstacle des plateformes à leur création est la circularité. En effet, dans le cas d'une plateforme de P&P lending, une « face » des utilisateurs (emprunteurs) ne sera attirée que si l'autre (prêteurs) est présente et vice-versa. Avant d'avoir suffisamment de particuliers prêteurs puisqu'il ne manque pas de personnes en besoin de financement (emprunteurs), il nous faut mobiliser des capitaux suffisants pour jouer le rôle de prêteur dans le but de faire fonctionner notre plateforme et accroître sa popularité. Ainsi, avec l'augmentation de l'audience sur notre plateforme, de nombreux particuliers prêteurs viendront y faire des investissements et nous titriserons les prêts dans notre portefeuille (les céder à d'autres particuliers prêteurs) pour se prémunir des créances en souffrances qui pourraient survenir. A la longue, nous miserons plus sur les commissions que nous obtiendrons de nos utilisateurs et aussi l'accord de certains prêts que nous jugerons avantageux. Les identités des P&P n'étant pas dévoilées au public, revendre des prêts sur la plateforme ne pourrait avoir d'obstacle.

Il est donc primordial pour la viabilité de notre projet de connaître le montant prévisionnel à mobiliser. Nous visons une clientèle ayant les mêmes habitudes d'emprunt de celle de Bondora et environ 3310 emprunteurs, avec une croissance annuelle de 50%. C'est pourquoi, nous utiliserons la base de données des prêts de Bondora 'Loan Dataset' pour construire un modèle prédictif du montant (variable Amount) du prêt à accorder quel que soit le montant de l'emprunt désiré (variable AppliedAmount) par tout emprunteur potentiel. Il est aussi à préciser que les frais préliminaires et d'exploitation (budget prévisionnel) du projet sont déjà réunis et donc qu'ils ne sont l'objet de cette étude.

## **II-Présentation de Bondora**

Basée en Estonie, Bondora est une plateforme de financement participatif ‘Crowdfunding’ pionnière en Europe dans les prêts entre particuliers (peer-to-peer lending P&P) qui offre de nombreux produits de la finance participative.

Avec son service de P&P, elle assure l’intermédiation (moyennant des commissions) entre d’un côté des particuliers à la recherche de financement dits emprunteurs et d’autre côté des particuliers investisseurs ‘prêteurs’. Elle met à la disposition de ses utilisateurs de nombreux outils d’analyse de portefeuilles afin d’optimiser leurs investissements pour les uns et trouver les meilleurs prêts en termes de coûts, d’échéances entre autres pour les autres.

Il est à noter que les taux d’intérêts proposés sur Bondora sont généralement plus élevés que ceux pratiqués par les banques traditionnelles. Cette particularité chez Bondora est compréhensible compte tenu des risques auxquels s’exposent ses investisseurs d’autant plus que les emprunteurs qui se tournent vers sa plateforme ont généralement peu de chances d’obtenir des prêts auprès des institutions financières traditionnelles.

Bondora dispose d’outils permettant de déterminer le montant optimal qu’il faut prêter compte tenu des données recueillies sur l’emprunteur. Nous supposons donc que le montant accordé (Amount) à l’emprunteur tient compte des scores de crédits (EsEquifaxRisk, FiAsiakasTietoRiskGrade, EeMini, etc.), des modèles de notation (Rating) et de toutes les variables que nous retiendrons.

### **III- Nettoyage et préparation des données**

Nous avons téléchargé notre base de données « Loan dataset » le 01/01/2022 sur la page web <https://www.bondora.com/en/public-reports> via le lien <https://www.bondora.com/marketing/media/LoanData.zip> sous format .csv. Ainsi, on peut suivre l'état d'évolution de tout prêt et ce de sa souscription à la date de téléchargement du jeu de données. Rappelons que la variable LoanNumber est l'identifiant principal de l'emprunt.

Au 1<sup>er</sup> Janvier 2022, date de téléchargement du jeu de données, elle contient 208872 observations (lignes) et 112 variables (colonnes).

Notre clientèle cible se constituera de personnes de 60 ans au plus ayant fait des études supérieures, secondaires ou professionnelles et qui ne sont pas des retraités, des sans-emplois ou des sans-abris. De ce fait, nous excluons de notre jeu de données toutes observations ne respectant pas nos critères. Aussi, dans notre modèle nous tiendrons compte du montant total des revenus de notre emprunteur potentiel sans s'atteler à leur différente source puisque nous supposons qu'il est en règle avec les services fiscaux. Aussi, nous privilégions le statut de l'emploi 'EmploymentStatus' et donc au lieu de traiter la variable EmploymentStatus comme catégorielle nous la considérons comme un poids sur le dossier de demande d'emprunt (priorité pour les entrepreneurs : 5 points, travailleurs indépendants :4 Self-employed, les employés à temps plein : 3 points Fully employed.

De plus, vu que notre plateforme aura une dimension internationale, nous ne ferons aucune discrimination en ce qui concerne la langue et le pays de l'emprunteur.

Contrairement à Bondora, nous souhaitons avoir affaire qu'avec des emprunteurs pouvant fournir des justificatifs authentiques sur leur situation ainsi, nous allons nous restreindre aux observations dont les revenus et les dépenses ont été vérifiés (modalité 4 de la variable VerificationType). Aussi, compte tenu de la précision des informations souhaitées, nous excluons la catégorie Others comme nous l'avons fait pour les sans-abris de la variable HomeOwnerShipType qui renseigne sur l'adresse et l'appartenance ou non de la résidence de l'emprunteur.

## a. Aperçu de la base de données brutes avant traitement et nettoyage

ReportAsOfEOD	LoanId	LoanNumber	ListedOnUTC	BiddingStartedOn	BidsPortfolioManager	BidsApi	BidsManual	PartyId	NewCreditCustomer
01/01/2022	66AE108B-532B-4BB3-BAB7-0019A46412C1	483449	23/03/2016 16:07	23/03/2016 16:07		970	1150	5 {EBF05573-554D-4A3B-BC77-A2CF00B7D110}	False
01/01/2022	D152382E-A50D-46ED-8FF2-0053E0C86A70	378148	25/06/2015 11:02	25/06/2015 11:02		1295	0	1705 {46C6CBA4-0FBE-44AD-9304-A3EF0111A5FB}	False
01/01/2022	87342E13-66CB-483F-833A-007953E50C78	451831	14/01/2016 10:00	14/01/2016 10:00		2700	565	5835 {CA64DA9B-8E95-450E-9EFE-A58601016DB2}	True
01/01/2022	87227056-6BF9-410C-98D1-008F788E122A	349381	24/03/2015 15:55	24/03/2015 15:55		1115	0	385 {F08F654D-DB2E-4C4B-8C90-A46100FCE7B6}	True
01/01/2022	2DDE6336-E466-4624-A337-00A0ED1A1468	443082	17/12/2015 10:12	17/12/2015 10:12		305	0	785 {DAFC5C08-8201-4654-9D06-A56D00084659}	True
01/01/2022	BA1FC89D-44B5-4481-9FCD-00C4BBC174B0	430905	13/11/2015 14:08	13/11/2015 14:08		600	0	175 {49167251-7475-4968-A719-A3B500C71600}	True
01/01/2022	932B0F92-8B44-499F-A056-00C6D6D1E312	473276	29/02/2016 14:24	29/02/2016 14:24		635	0	0 {64856EDA-8FFC-43FE-A896-A59D001FC539}	True
01/01/2022	CDC5127D-AE6E-4BEC-9508-00FE1FB8FFE4	348441	25/03/2015 16:39	25/03/2015 16:39		295	0	705 {E676162F-F160-45E2-A6F4-A2EB0094E068}	True
01/01/2022	BF536CB8-9221-475C-A4F1-010347DD5A83	337054	09/03/2015 15:09	09/03/2015 15:09		4000	0	0 {E4B0FC8F-20D7-486D-B07E-A3D200D99BBB}	True
01/01/2022	612D7DC3-EA4D-4D16-9C9E-010D4B72B766	362787	29/04/2015 12:19	29/04/2015 12:19		3790	0	1210 {5436932A-E1A2-4F44-99DB-A41A01302623}	False
01/01/2022	437A52D1-3EBF-4926-82B1-0120B11F26CA	360579	21/04/2015 17:06	21/04/2015 17:06		1715	0	285 {AFD342B3-EA7C-4595-A0F5-A2BC00C6B1D9}	False
01/01/2022	8AEA1291-D508-4812-BFB2-012DFCEA8A0D	420503	19/10/2015 13:34	19/10/2015 13:34		530	0	0 {7EA79A59-EC75-4A6E-9FEC-A43C009B3588}	True
01/01/2022	737A1EC7-839D-4DDE-B3F5-01444AD58B16	399202	26/08/2015 13:33	26/08/2015 13:33		3295	0	2205 {29356315-4FD1-4184-8725-A4FE016DC352}	True
01/01/2022	8F12E924-1D8C-4829-BBC1-0162F32EC8BF	368068	19/05/2015 15:08	19/05/2015 15:08		5080	0	1820 {831E4C8E-6C84-4ECD-9485-A21B00EA05A8}	False
01/01/2022	ED86263F-A2FB-4FFA-9862-018A6144F5E3	446706	29/12/2015 08:40	29/12/2015 08:40		1375	175	1105 {5576DA44-1DE7-4927-80E1-A57C00D26F2C}	True
01/01/2022	B694CD01-088B-465F-88A3-01A98527382D	449125	06/01/2016 11:12	06/01/2016 11:12		120	3070	0 {2BF895F4-55C4-4248-8F0B-A4430124D8B0}	False
01/01/2022	F7289542-C48E-4FE1-8205-01E155E06A3B	452329	15/01/2016 08:54	15/01/2016 08:54		455	0	75 {BDFE1CC8-2944-4D10-8543-A34800F27FBA}	False
01/01/2022	C586E40F-CBE3-4F37-A232-01E686573DA2	360182	20/04/2015 15:49	20/04/2015 15:49		0	0	500 {FEA4E67D-0355-4538-BBB4-A2E700AF3BF4}	False
01/01/2022	27AD9179-AE1B-4493-BC6D-021D37749264	450980	12/01/2016 13:32	12/01/2016 13:32		3665	0	55 {CC4275F1-4331-4850-AE04-A56E009DD7E0}	True
01/01/2022	7AE48BA6-997C-43E4-BC2C-025BCD89B235	388355	27/07/2015 10:16	27/07/2015 10:16		760	0	1740 {10974855-3E84-4D66-97BC-A2DC00C8A994}	False

LoanApplicationStartedDate	LoanDate	ContractEndDate	FirstPaymentDate	MaturityDate_Original	MaturityDate_Last	ApplicationSignedHour	ApplicationSignedWeekday	VerificationType	LanguageCode	Age	DateOfBirth	Gender	Country
17/03/2016 12:39	23/03/2016	26/06/2020	12/05/2016	12/04/2021	26/06/2020	17		4	4	1	53	NA	1 EE
24/06/2015 12:36	25/06/2015		17/08/2015	17/07/2020	17/07/2020	11		5	1	1	50	NA	1 EE
07/01/2016 15:37	19/01/2016	24/10/2019	22/02/2016	20/01/2021	20/01/2021	22		3	4	1	44	NA	0 EE
20/03/2015 15:20	27/03/2015		04/05/2015	01/04/2020	01/04/2020	15		3	3	6	42	NA	0 ES
13/12/2015 00:30	22/12/2015		01/02/2016	02/01/2020	02/01/2020	20		3	4	6	34	NA	1 ES
12/11/2015 12:58	19/11/2015		04/01/2016	01/12/2020	01/12/2020	0		5	4	6	31	NA	1 ES
29/02/2016 10:25	29/02/2016	13/07/2018	15/04/2016	15/03/2021	12/06/2019	15		2	3	1	22	NA	0 EE
17/03/2015 08:25	01/04/2015	26/04/2016	25/05/2015	25/04/2016	25/04/2016	16		4	1	6	47	NA	0 ES
02/03/2015 12:36	13/03/2015	14/09/2016	15/04/2015	15/03/2018	20/05/2021	10		2	4	4	60	NA	0 FI
23/04/2015 18:38	05/05/2015	14/05/2020	22/06/2015	20/05/2020	20/05/2020	12		4	1	1	39	NA	0 EE
21/04/2015 11:28	21/04/2015	28/03/2017	25/05/2015	23/04/2020	23/04/2020	17		3	4	1	31	NA	0 EE
14/10/2015 09:31	19/10/2015	20/10/2016	08/12/2015	09/11/2020	09/11/2020	14		2	4	1	45	NA	1 EE
24/08/2015 22:11	02/09/2015	08/01/2019	15/10/2015	15/09/2020	15/09/2020	22		4	3	1	49	NA	0 EE
19/05/2015 09:55	20/05/2015	21/05/2015	15/07/2015	15/06/2020	21/05/2015	15		3	4	1	31	NA	0 EE
28/12/2015 12:46	29/12/2015	27/09/2024	25/02/2016	25/01/2021	27/09/2024	10		3	4	1	35	NA	1 EE
05/01/2016 13:18	06/01/2016	10/10/2022	10/02/2016	11/01/2021	10/10/2022	14		4	4	1	22	NA	1 EE
14/01/2016 12:36	16/01/2016	19/07/2018	07/03/2016	08/02/2021	08/02/2021	12		7	1	6	42	NA	1 ES
20/04/2015 14:16	20/04/2015	27/12/2023	20/05/2015	20/04/2020	27/12/2023	15		2	4	3	33	NA	0 EE
12/01/2016 10:38	14/01/2016	05/09/2016	04/03/2016	04/02/2021	05/09/2016	11		5	4	4	67	NA	0 FI
26/07/2015 08:22	27/07/2015		10/09/2015	10/08/2020	10/08/2020	10		2	4	1	23	NA	1 EE

City	UseOfLoan	Education	MaritalStatus	NrOfDependants	EmploymentStatus	EmploymentDurationCurrentEmployer	EmploymentPosition	WorkExperience	OccupationArea	HomeOwnershipType	IncomeFromPrincipalEmployer
NA	2	4	2	0	6	MoreThan5Years	NA	15To25Years	1	1	0
NA	3	5	2	0	5	MoreThan5Years	NA	MoreThan25Years	7	1	900
NA	3	4	4	1	5	UpTo3Years	NA	MoreThan25Years	8	8	600
NA	2	2	1	0	3	UpTo5Years	NA	5To10Years	1	2	863
NA	7	4	4	2	6	UpTo1Year	NA	5To10Years	1	3	0
NA	7	4	1	0	3	UpTo5Years	NA	10To15Years	7	4	970
NA	8	2	3	0	3	UpTo1Year	NA	2To5Years	9	4	745
NA	6	3	2	2	6	MoreThan5Years	NA	5To10Years	1	1	0
NA	2	4	1	0	3	MoreThan5Years	NA	MoreThan25Years	11	1	2590
NA	0	4	4	1	3	UpTo5Years	NA	15To25Years	8	6	1000
NA	7	4	3	0	3	MoreThan5Years	NA	2To5Years	8	3	605
NA	2	4	1	0	3	MoreThan5Years	NA	15To25Years	1	9	633
NA	2	4	4	0	5	MoreThan5Years	NA	MoreThan25Years	13	1	550
NA	7	2	3	1	3	MoreThan5Years	NA	5To10Years	19	1	833
NA	2	4	3	0	3	MoreThan5Years	NA	5To10Years	17	4	341
NA	7	3	3	0	3	UpTo1Year	NA	LessThan2Years	3	2	430
NA	7	5	3	0	3	UpTo1Year	NA	LessThan2Years	1	1	1000
NA	5	2	1	3	3	UpTo5Years	NA	15To25Years	1	1	500
NA	0	3	4	0	6	UpTo2Years	NA	MoreThan25Years	6	3	0
NA	6	4	2	0	3	UpTo1Year	NA	2To5Years	3	7	277

IncomeFromPension	IncomeFromFamilyAllowance	IncomeFromSocialWelfare	IncomeFromLeavePay	IncomeFromChildSupport	IncomeOther	IncomeTotal	ExistingLiabilities	LiabilitiesTotal	RefinanceLiabilities	DebtToIncome	FreeCash
301		0	53	0	0	354	8	485.09	6	26.29	10.92
0		0	0	0	0	900	4	736.45	0	30.58	78.8
0		0	0	0	600	1200	7	905	3	26.71	349.43
0		0	0	0	0	863	1	350	0	7.36	449.47
697		0	0	0	0	697	5	940	2	36.04	95.81
0		0	0	0	0	970	5	960	2	47.96	154.81
0		0	0	0	0	745	1	250	0	3.64	467.88
1126		0	0	0	0	1126	2	560	0	29.04	449.05
0		0	0	0	0	2590	8	2068	0	48.51	336.65
0		0	0	0	0	1000	9	1293.18	3	51.52	14.81
0		0	0	0	0	605	7	722.4	1	53.04	34.11
0		0	0	0	0	633	7	666	1	30.32	41.08
0		0	0	0	0	550	1	250	0	29.57	137.36
0		0	0	0	0	833	10	754	4	45.75	201.87
0		0	0	0	0	341	6	403	5	25.4	4.37
0		0	0	0	0	430	7	595	5	40.21	7.11
0		0	0	0	0	1000	9	1097.32	2	58.9	61.05
0	247		0	0	240	987	6	769.33	0	50.22	195.29
1287		0	104	0	0	1391	12	1637	3	60.21	53.48
237		0	40	0	0	554	4	535	1	48.6	34.78

MonthlyPaymentDay	ActiveSchedule	FirstPaymentReached	PlannedPrincipalTillDate	PlannedInterestTillDate	LastPaymentOn	CurrentDebtDaysPrimary	DebtOccuredOn	CurrentDebtDaysSecondary	DebtOccuredOnForSecondary	ExpectedLoss
12	True		630.22	1251.98	06/12/2021	717	14/01/2020	795	28/10/2019	0.0685118666998432
17	True		1333.51	3000	19/06/2019	2083	18/04/2016	2144	17/02/2016	0.0307991298909077
20	True		3348.5	9100	23/10/2019	1533	20/10/2017	1593	21/08/2017	0.0231768239099692
1	True		573.39	1500	07/08/2020	2222	01/12/2015	2222	01/12/2015	0.2208097125
1	True		304.9	1090	01/02/2016	2160	01/02/2016	2160	01/02/2016	0.56867833493864
1	True		103.16	775	23/07/2021	2188	04/01/2016	2188	04/01/2016	0.5302226475
15	True		220.45	617.98	18/06/2018	1509	13/11/2017	1509	13/11/2017	0.242447184307918
25	True		1000	403.77	26/04/2016	NA	NA	NA		0.407863575
20	True		435.75	3236.28	14/09/2016	NA	NA	NA		0.146732058007789
20	True		2175.04	5000	14/05/2020	NA	NA	NA		0.0808357810992321
23	True		2000	575.44	28/03/2017	NA	NA	NA		0.0307165572053542
8	True		530	130.03	20/10/2016	NA	NA	NA		0.0873923894989163
15	True		2208.1	5500	08/01/2019	NA	NA	NA		0.0652203959750599
15	True		6900	3.74	21/05/2015	NA	NA	NA		0.065277411240252
25	True		782.1	598.42	28/12/2021	NA	NA	NA		0.106410786919366
8	True		0	1731.15	11/12/2017	1438	23/01/2018	1514	08/11/2017	0.354228358703779
6	True		98.94	530	19/07/2018	NA	NA	NA		0.356179642641465
20	True		31.7	138.26	08/12/2021	4	27/12/2021	4	27/12/2021	0.212058971191784
4	True		2297.61	124.37	05/09/2016	NA	NA	NA		0.13643105594287
10	True		890.19	2500	27/12/2021	2213	10/12/2015	2213	10/12/2015	0.12954322630376

LossGivenDefault	ExpectedReturn	ProbabilityOfDefault	DefaultDate	PrincipalOverdueBySchedule	PlannedPrincipalPostDefault	PlannedInterestPostDefault	EAD1	EAD2	PrincipalRecovery	InterestRecovery	RecoveryStage
0.58	0.141144933565886	0.122215942282687	14/01/2020	1141.84	1251.98	77.68	1251.98	64.07	110.14	0	2
0.65	0.140435615979576	0.0364486744271097	02/06/2016	2436.41	2658.82	1078.96	2730.84	2370.77	294.43	0	2
0.58	0.113484142045691	0.0413443321532396	06/12/2017	0	6456.37	1537.37	6723.01	5014.54	4160.01	0	2
0.9	0.183229267830455	0.18872625	19/02/2016	1035.27	1406.56	1520.11	1434.68	1078.76	399.41	0	1
0.75	0.115240123985128	0.749511220774706	18/04/2016	1089.99	1075.51	1931.04	1089.99	1089.99	0	0	1
0.9	0.207060390408129	0.45318175	16/03/2016	738.14	770.57	2021.41	774.96	774.96	36.82	0	2
0.58	0.184104040542068	0.408035908768402	29/01/2018	0	487.68	170.43	546.19	123.44	383.78	0	2
0.9	0.207060390408129	0.45318175		NA	NA	NA	NA	NA	NA	NA	NA
0.9	0.16338742368407	0.125412015391273	06/09/2016	NA	3236.28	3201.9	3236.28	1937.62	3236.28	0	NA
0.65	0.16436811971466	0.0956636462712806		0	NA	NA	NA	NA	NA	NA	NA
0.65	0.140409097024032	0.0363509552726085		NA	NA	NA	NA	NA	NA	NA	NA
0.65	0.169450552465178	0.103422946152564		NA	NA	NA	NA	NA	NA	NA	NA
0.65	0.150993428972526	0.077183900562201		0	NA	NA	NA	NA	NA	NA	NA
0.65	0.151045865020019	0.077251374248819		NA	NA	NA	NA	NA	NA	NA	NA
0.58	0.162421297262494	0.189822511323025		0	NA	NA	NA	NA	NA	NA	1
0.58	0.161262827264702	0.631895680669448	23/01/2018	1731.15	1731.15	4771.06	2865.72	196.44	0	0	2
0.75	0.15202178519752	0.406100917788613	20/06/2017	0	484.94	592.38	495.79	196	495.79	0	2
0.65	0.22763371754237	0.250957362357141		6.73	NA	NA	NA	NA	NA	NA	1
0.68	0.132986724180193	0.198665921183098		NA	NA	NA	NA	NA	NA	NA	NA
0.65	0.196207260498538	0.153305593258888	26/02/2016	2144.29	2390.71	2193.53	2447.28	2213.96	302.99	0	1



StageActiveSince	ModelVersion	Rating	EL_V0	Rating_V0	EL_V1	Rating_V1	Rating_V2	Status	Restructured	ActiveLateCategory	WorseLateCategory	CreditScoreEsMicroL	CreditScoreEsEquifaxRisk
2020-03-03 09:27:48.493000000		2 C	NA		NA		C	Late	False	180+	180+		
01/08/2019 14:18		1 B	NA		0.0307991298909077	B	B	Late	False	180+	180+		
2018-02-28 14:43:37.670000000		2 A	NA		NA		A	Repaid	False	180+	180+		
27/11/2020 00:00		1 F	NA		0.2208097125	F	HR	Late	False	180+	180+	M3	B
27/11/2020 00:00		2 HR	NA		NA		HR	Late	False	180+	180+	M5	C
26/11/2019 00:00		1 HR	NA		0.5302226475	HR	HR	Late	False	180+	180+	M5	C
2018-03-21 17:56:17.197000000		2 F	NA		NA		F	Repaid	False	180+	180+		
		1 HR	NA		0.407863575	HR	HR	Repaid	False		01-juil	M5	C
		1 E	NA		0.146732058007789	E	D	Repaid	True		31-60		
		1 C	NA		0.0808357810992321	C	C	Repaid	False				
		1 B	NA		0.0307165572053542	B	B	Repaid	False		01-juil		
		1 C	NA		0.0873923894989163	C	C	Repaid	False				
		1 C	NA		0.0652203959750599	C	D	Repaid	False		01-juil		
		1 C	NA		0.065277411240252	C	D	Repaid	False				
2019-11-28 09:00:35.287000000		2 D	NA		NA		D	Current	True		août-15		
2019-07-12 16:51:54.713000000		2 HR	NA		NA		HR	Late	True	180+	180+		
11/07/2018 11:19		2 HR	NA		NA		HR	Repaid	False		180+	M5	C
2021-12-28 02:19:09.677000000		1 F	NA		0.212058971191784	F	F	Late	True	01-juil	16-30		
		2 E	NA		NA		E	Repaid	False				
02/01/2019 15:10		1 D	NA		0.12954322630376	D	C	Late	False	180+	180+		

CreditScoreFiAsiakasTietoRiskGrade	CreditScoreEeMini	PrincipalPaymentsMade	InterestAndPenaltyPaymentsMade	PrincipalWriteOffs	InterestAndPenaltyWriteOffs	PrincipalBalance	InterestAndPenaltyBalance	NoOfPreviousLoansBeforeLoan
	1000	983.16	1187.91	0	0	1141.84	507.21	1
	1000	563.59	360.07	0	0	2436.41	2429.7	1
	1000		6537 1708.47	2303.33	0.88	0	0	0
	NA	464.73	355.92	0	0	1035.27	2972.09	0
	NA	0.01		0	0	1089.99	4461.78	0
	NA	36.86		0	0	738.14	3443.63	0
	800	472.59	422.75	96.77	2.87	0	0	0
	NA		1000 403.89	0	0	0	0	0
RL1	NA		4000 1567.36	0	0	0	0	0
	800		5000 3770.58	0	0	0	0	1
	1000		2000 575.44	0	0	0	0	2
	800		530 130.03	0	0	0	0	0
	1000		5500 3071.51	0	0	0	0	0
	1000		6900 3.74	0	0	0	0	1
	1000	1222.97	3427.38	0	0	1432.03		0
	1000	324.28	2669.28	0	0	2865.72	5609.26	2
	NA		530 434.68	0	0	0	0	1
	800	232.18	1183.87	0	0	267.82	9.85	1
RL1	NA		3720 512.48	0	0	0	0	0
	1000	355.71	233.32	0	0	2144.29	4315.66	2

AmountOfPreviousLoansBeforeLoan	PreviousRepaymentsBeforeLoan	PreviousEarlyRepaymentsBeforeLoan	PreviousEarlyRepaymentsCountBeforeLoan	GracePeriodStart	GracePeriodEnd	NextPaymentDate	NextPaymentNr	NrOfScheduledPayments
500	590.95	0	0	28/10/2019	27/01/2020		NA	NA
1800	445.26	3000	1				NA	NA
0	0	0	0				NA	NA
0	0	1500	1				NA	NA
0	0	0	0				NA	NA
0	0	0	0				NA	NA
0	0	0	0	10/01/2017	12/07/2017		NA	NA
0	0	1500	1				NA	NA
0	0	0	0	20/06/2016	20/06/2017		NA	NA
1000	106.17	5000	1				NA	NA
2100	1574.57	2000	1				NA	NA
0	0	0	0				NA	NA
0	0	0	0				NA	NA
1500	1678.75	15900	2				NA	NA
0	0	0	0			27/01/2022	28	60
1600	1901.16	0	0	08/11/2017	08/11/2018		0	60
600	344.09	0	0				NA	NA
3500	884.71	10000	7			27/01/2022	37	60
0	0	0	0				NA	NA
3800	4456.41	18000	8				NA	NA

ReScheduledOn	PrincipalDebtServicingCost	InterestAndPenaltyDebtServicingCost	ActiveLateLastPaymentCategory
	0	59.26	16-30
	0	47.08	180+
259.67		1659.6	180+
	0	215.05	180+
	0		0 180+
	0	19.83	151-180
65.64		147.67	180+
	0		0
	0		0
	0		0
	0		0
	0		0
	0		0
	0		0
27/09/2019	0		0
17/10/2017	0		0 180+
	0	106.93	
13/01/2019	0		0 16-30
	0		0
	0	163.14	01-juil

Nous excluons toutes les variables générées après l'octroi du prêt sauf la variable ExpectedReturn car des variables sur l'amortissement de l'emprunt, des pénalités, et de nombreuses autres ont été générées pour permettre de faire un suivi des prêts.

Enfin, pour la constitution de notre base de données d'exploration de l'activité de Bondora et de construction de notre modèle prédictif du montant optimal à prêter, nous allons nous restreindre aux observations ayant des informations complètes pour toutes les variables que nous avons retenues.

Après croisement de toutes nos exigences, nous retenons 15 variables pour 11033 observations.

### **a. Description des variables retenues pour l'étude**

Les variables retenues sont :

**1-Age:** l'âge de l'emprunteur à la date de souscription de l'emprunt

**2-Amount :** le montant du prêt accordé à l'emprunteur, la variable expliquée que nous souhaitons modéliser et prédire.

**3-AmountOfPreviousLoansBeforeLoan:** le montant des prêts précédemment accordés à l'emprunteur.

**4-AppliedAmount:** Le Montant qu'il désire emprunter qui n'est pas toujours celui qui lui est accordé (Amount)

**5-DebtToIncome :** La proportion du revenu mensuel de l'emprunteur consacré au remboursement de ses emprunts.

**6-EmploymentStatus:** Cette variable fournit des informations de la catégorie socioprofessionnelle de l'emprunteur.

**7-ExpectedLoss:** C'est un pourcentage indiquant la perte financière attendue d'un prêt dans un horizon temporel spécifié. En multipliant le pourcentage par l'exposition, on obtient la perte attendue en termes monétaires.<sup>1</sup> Il varie selon le modèle de notation (rating) et est basé sur des données historiques

**8-ExpectedReturn:** Mesure le rendement anticipé du prêt et est également basé sur des données historiques.<sup>2</sup>

**9-IncomeTotal :** Somme des revenus mensuels des différentes sources.

**10-Interest :** Le taux d'intérêt maximal accepté pour le prêt dont il est question.

**11-LoanDuration :** Durée du prêt.

---

<sup>1</sup> [https://www.openriskmanual.org/wiki/Expected\\_Loss](https://www.openriskmanual.org/wiki/Expected_Loss)

<sup>2</sup> <https://www.investopedia.com/terms/e/expectedreturn.asp>

**12-LossGivenDefault :** Le montant d'argent qu'un prêteur perd lorsque l'emprunteur ne rembourse pas un prêt, après avoir pris en compte tout recouvrement, représenté en pourcentage de l'exposition totale (le montant du prêt) au moment de la perte.<sup>3</sup>

**13-NrOfDependants :** Le nombre de personnes à la charge de l'emprunteur, nous en avons retenu des emprunteurs n'ayant pas plus 10 à leur charge.

**14-ProbabilityOfDefault :** C'est la probabilité que l'emprunteur ne s'acquitte pas de sa dette.<sup>4</sup>

**15-UseOfLoan :** L'usage qu'en fait l'emprunteur du prêt qui lui a été accordé, le motif de l'emprunt.

---

<sup>3</sup> <https://corporatefinanceinstitute.com/resources/knowledge/credit/loss-given-default-lgd/>

<sup>4</sup> <https://www.garp.org/risk-intelligence/credit/probability-of-default-the-pluses-and-minuses-of-transition-matrices/>

## b. Aperçu des données nettoyées

Le jeu de données que nous retenons pour l'études est de la forme suivante

	Age	Amount	AmountOfPreviousLoansBeforeLoan	AppliedAmount	DebtToIncome	EmploymentStatus	ExpectedLoss	ExpectedReturn	IncomeTotal	Interest	LoanDuration	LossGivenDefault	NrOfDependants	ProbabilityOfDefault	UseOfLoan
44	9100	0	10630	0.2671	5	0.023176823	0.113484142	1200	0.1367	60	0.58	1	0.041344332	3	
31	775	0	3720	0.4796	3	0.530222647	0.207060390	970	0.7373	60	0.9	0	0.45318175	7	
60	4000	0	4000	0.4851	3	0.146732058	0.163387423	2590	0.3101	36	0.9	0	0.125412015	2	
31	2000	2100	2000	0.5304	3	0.030716557	0.140409097	605	0.1711	60	0.65	0	0.036350955	7	
45	530	0	530	0.3032	3	0.087392389	0.169450552	633	0.2568	60	0.65	0	0.103422946	2	
35	2655	0	2655	0.254	3	0.106410786	0.162421297	341	0.2688	60	0.58	0	0.189822511	2	
22	3190	1600	3190	0.4021	3	0.354228358	0.161262827	430	0.5155	60	0.58	0	0.631895680	7	
23	2500	3800	2500	0.486	3	0.129543226	0.196207260	554	0.3258	60	0.65	0	0.153305593	6	
33	3000	0	3000	0.1003	3	0.103109419	0.180534413	1038	0.2836	60	0.65	0	0.122022981	0	
34	4000	0	4000	0.5916	3	0.163470071	0.169719681	1256	0.3332	60	0.9	0	0.139718009	2	
37	2125	0	2125	0.4557	3	0.175821534	0.181471373	496	0.3573	60	0.58	0	0.313641936	0	
32	500	500	500	0.149	3	0.024397844	0.137759207	476	0.1622	60	0.65	1	0.028873188	2	
25	9565	0	9565	0.6945	3	0.075086167	0.145586502	850	0.2207	60	0.58	0	0.133943610	0	
23	1595	0	1595	0.362	3	0.165824905	0.179948147	430	0.3458	60	0.58	0	0.295809295	6	
50	1400	3000	1400	0.2904	3	0.017290295	0.132105082	610	0.1494	36	0.65	0	0.020461887	2	
54	2655	0	2655	0.6065	3	0.273840521	0.238586378	640	0.5124	60	0.65	0	0.324071622	0	
31	4000	1500	4000	0.3739	3	0.026215452	0.138673686	895	0.1649	60	0.65	1	0.031024203	7	
44	2655	0	2655	0.4928	3	0.201659126	0.144008944	2127	0.3457	60	0.68	1	0.293648654	6	
44	6910	4200	6910	0.3916	3	0.020117262	0.134854325	1300	0.155	60	0.65	1	0.023807410	7	
47	2500	0	2500	0.4574	5	0.101365359	0.179375411	475	0.2807	60	0.65	1	0.119959005	3	
49	1060	6000	1060	0.6702	3	0.060632591	0.135316046	800	0.1959	60	0.58	0	0.108160376	7	
26	635	1000	635	0.089	3	0.128610448	0.170853580	785	0.2995	60	0.58	0	0.229423717	1	
24	500	1500	500	0.5624	2	0.530222647	0.207060390	850	0.7373	60	0.9	0	0.45318175	7	
26	2125	0	2125	0.3191	3	0.264930168	0.150900019	621	0.4158	60	0.75	0	0.331701207	4	
32	1060	0	1060	0.4954	3	0.044780730	0.121608969	1091	0.1664	60	0.58	1	0.079882791	6	

Nous n'avons retenu que des variables numériques pour notre étude.

## IV- Analyse exploratoire des données retenues pour le projet

Il est important pour notre étude d'avoir une idée précise des emprunteurs de Bondora pour mieux implémenter notre modèle. C'est pourquoi, nous allons observer la distribution de certaines variables prises individuellement et ainsi faire un croisement entre certaines d'entre elles.

donnees\_c1eaned

1	Variables	11033	Observations										
-----													
Age													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	
	11033	0	42	0.999	36.25	11.26	23	24	28	35	43	51	
	.95												
	54												
Lowest : 19 20 21 22 23, highest: 56 57 58 59 60													
-----													

1	Variables	11033	Observations										
-----													
Amount													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	
	11033	0	920	0.999	3082	2477	530	700	1275	2500	4100	6375	
	.95												
	8000												
lowest : 100 120 130 150 190, highest: 10100 10310 10410 10415 10630													
-----													

1	Variables	11033	Observations										
AmountOfPreviousLoansBeforeLoan													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	
	11033	0	888	0.67	1093	1804	0	0	0	0	1060	4000	
	.95												
	6000												
lowest :		0.0000	31.9558	63.9100	63.9118	95.8675, highest: 22648.2382 23178.2382 23630.0000 23708.2382							
30000.0000													

1 Variables		11033 Observations										
-----												
AppliedAmount												
n	missing	distinct		Info	Mean	Gmd	.05	.10	.25	.50	.75	.90
11033	0	261		0.999	3342	2740	530	700	1400	2655	4500	7120
.95												
9565												
lowest : 100 120 130 150 200, highest: 10205 10310 10415 10420 10630												
-----												

1	Variables	11033 Observations											
-----													
IncomeTotal													
	n	missing	distinct		Info	Mean	Gmd	.05	.10	.25	.50	.75	.90
	11033	0	2217		1	1211	718.7	472	550	725	1007	1500	2090
	.95												
	2500												
-----													
Lowest :	200	300	302	303	304,	highest:	10018	11265	14016	16000	16400		
-----													

Pour chacun des tableaux ci-dessus, on constate qu'il n'y a pas de valeurs manquantes (missing=0). Les 5 plus petites valeurs (lowest) de notre variable cible Amount sont 100 120 130 150 et 190 et les 5 plus grandes valeurs (highest) du montant en Euro de

l'emprunt désiré (pas forcément obtenu : AppliedAmount) des emprunteurs sont 10205 10310 10415 10420 et 10630.

On peut aussi voir que le revenu moyen des emprunteurs est de 1211 alors que le revenu le plus élevé est de 16400€. Dans l'ensemble, la moitié des emprunteurs a un revenu supérieur à 1007€ et l'autre moitié touche moins de 1007€.

Nous pouvons aussi voir le nombre valeurs distinctes pour chacune de nos variables de même que la différence de la moyenne de Gini qui est un indicateur alternatif à l'écart type (la dispersion des observations autour de leur valeur moyenne).

-Tableau résumé de certaines statistiques descriptives

	Nombres d'observations	Moyenne		Ecart-type	Valeur minimale	1er quartir	3em quartile	Valeur maximale
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max	
Age	11033	36,255	9,886	19	28	43	60	
Amount	11033	3081,925	2336,108	100	1275	4100	10630	
AmountOfPreviousLoansBeforeLoan	11033	1092,971	2362,602	0	0	1060	30000	
AppliedAmount	11033	3341,931	2572,836	100	1400	4500	10630	
DebtToIncome	11033	0,351	0,176	0	0,21	0,481	0,8	
EmploymentStatus	11033	3,073	0,447	2	3	3	5	
ExpectedLoss	11033	0,163	0,16	0,006	0,052	0,225	0,999	
ExpectedReturn	11033	0,121	0,139	-0,799	0,101	0,195	0,695	
IncomeTotal	11033	1211,171	758,061	200	725	1500	16400	
Interest	11033	0,299	0,186	0,06	0,2	0,31	2,636	
LoanDuration	11033	42,517	17,759	1	24	60	60	
LossGivenDefault	11033	0,677	0,097	0,267	0,65	0,68	0,9	
NrOfDependants	11033	0,774	1,009	0	0	1	8	
ProbabilityOfDefault	11033	0,201	0,171	0,013	0,073	0,297	0,854	
UseOfLoan	11033	3,316	2,857	0	0	6	8	

Pour le jeu de données que nous avons retenu :

-Le plus jeune des emprunteurs a 19 ans et le plus vieux a 60 ans, la moyenne d'âge des emprunteurs est de 36,25 ans et la dispersion des âges autour de cette moyenne est de 9,886 ans. Aussi, 75 % des emprunteurs ont au plus 43 ans et seulement 25 % des emprunteurs ont au plus 28 ans.

-Le taux d'intérêt le plus élevé appliqué est de 2,636 soit plus de 250% du montant du prêt.

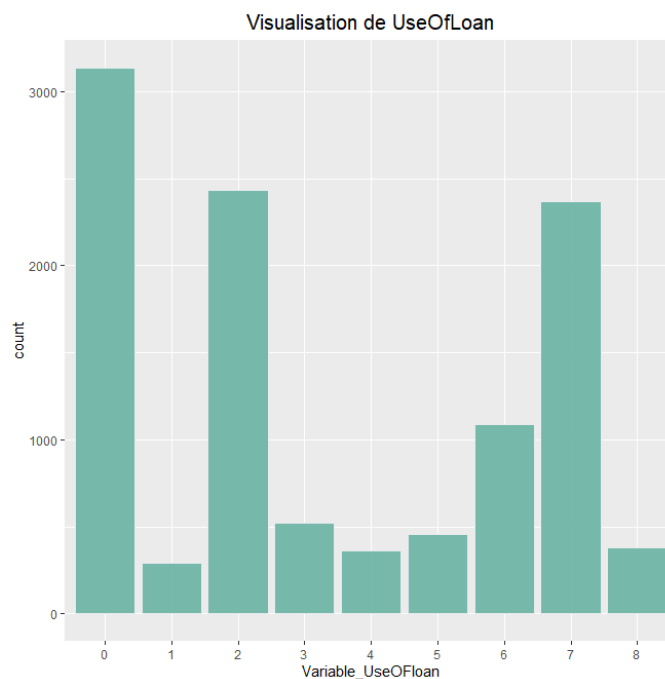
-Le taux de la perte financière anticipée la plus élevée des prêts accordés est de 99,9% alors que le rendement le plus faible est -79,9% ( il ne sera même pas possible de récupérer le montant du prêt à plus forte raison celui des intérêts).

-Le prêt accordé le plus élevé était de 10630€ il en est de même du montant de l'emprunt demandé et 100€ comme valeur minimale pour chacune des deux variables. La moyenne des prêts accordés est de 3081,925€ et 3341,931€ pour celui du montant de l'emprunt désiré.

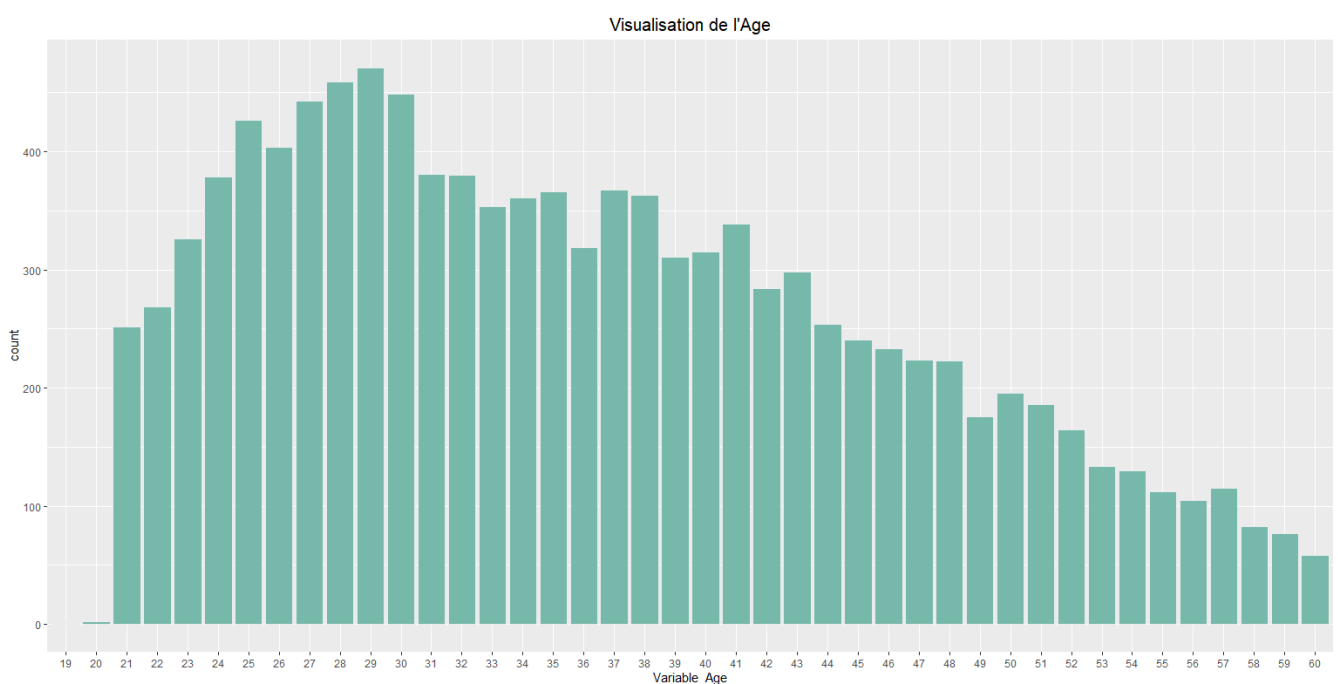
## b. Représentation graphique de la distribution de certaines variables

-Diagramme en barre de la distribution des emprunts selon la raison de la souscription. Avant de faire le graphique, rappelons que les modalités [0, 1, 2, 3, 4, 5, 6, 7, 8] de la variable UseOfLoan représente respectivement les motifs suivants [Consolidation de prêts, Immobilier, Amélioration de l'habitat, Affaires, Éducation, Voyage, Véhicule, Autre, Santé].

Ainsi, nous remarquons sur le graphique que le motif le plus donné est celui de la consolidation des prêts (emprunt pour rembourser un autre), ensuite c'est celui de l'amélioration de l'habitat puis vient le motif autre.



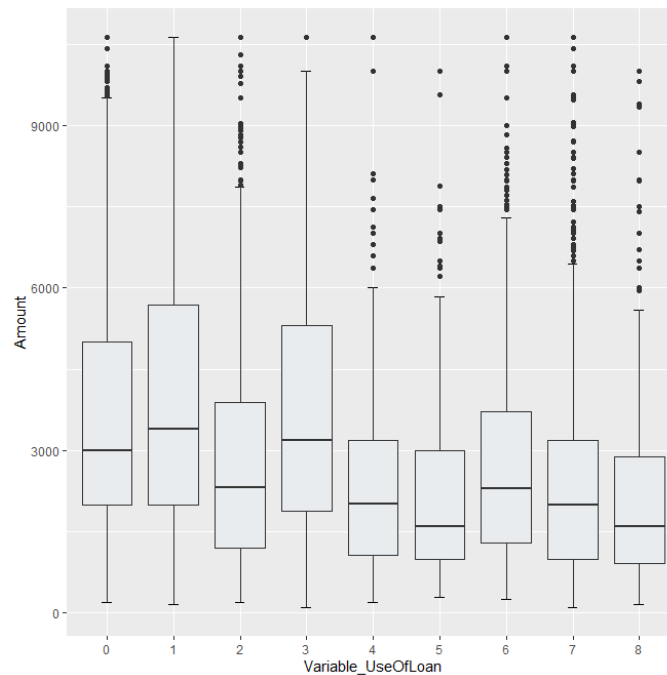
-Diagramme en barre de la distribution de l'âge des emprunteurs





Nous pouvons constater sur le graphe ci-dessus que les personnes ayant 29 ans sont ceux qui empruntent le plus.

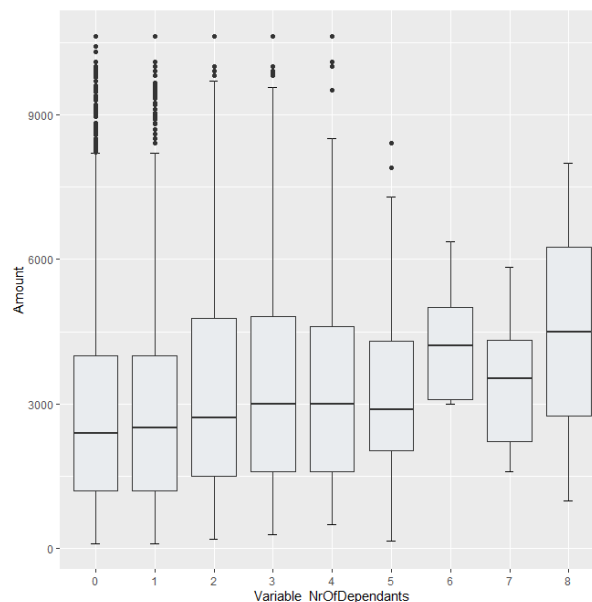
-Représentation graphique avec des boîtes à moustache 'boîte de Tukey' des caractéristiques de la variable que nous cherchons à prédire Amount sur les sous-populations définies par les modalités de UseOfLoan.



Les longueurs des boîtes traduisent la variabilité des valeurs des montants prêtés selon les différents groupes. La ligne du milieu est la médiane. Pour toutes les modalités, la médiane est dans le bas de la boîte de Tukey, ce qui traduit une distribution asymétrique qui penche vers les valeurs basses du montant prêté.

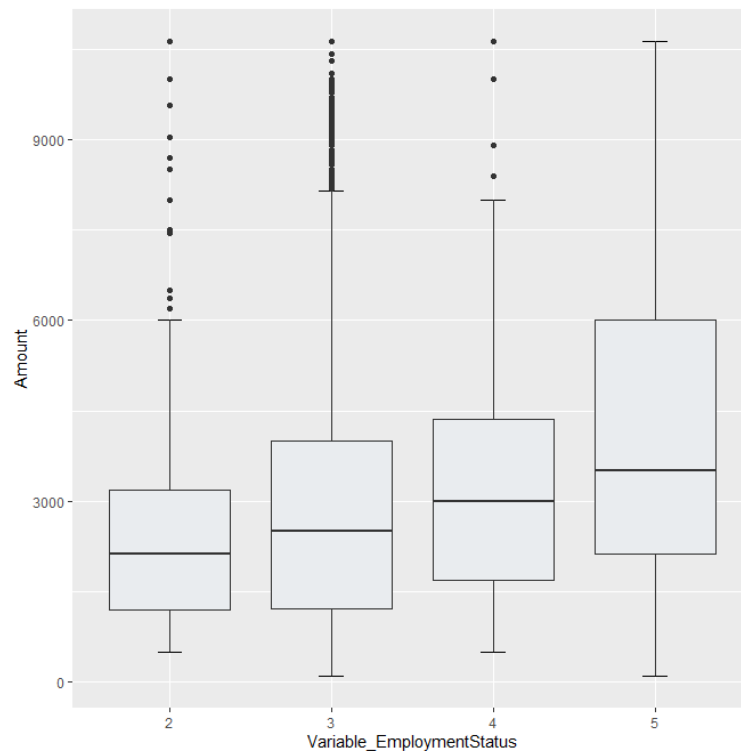
Par exemple, pour la modalité 4 de UseOfLoan (motif éducation) 50% des emprunteurs ont bénéficié d'un prêt supérieur à 2022,5€ et les 50% autres restant en ont obtenu moins de 2022,5€.

-Représentation graphique de la distribution de Amount en fonction du nombre de personnes à la charge de l'emprunteur à l'aide du diagramme de Tukey.



Pour les emprunteurs n'ayant qu'une seule personne à leur charge, 25% n'obtiennent pas plus de 1200€ alors qu'il existe une même proportion qui obtient plus de 4000€ de prêt.

-Représentation graphique de la distribution de Amount en fonction du nombre de personnes à la charge de l'emprunteur.

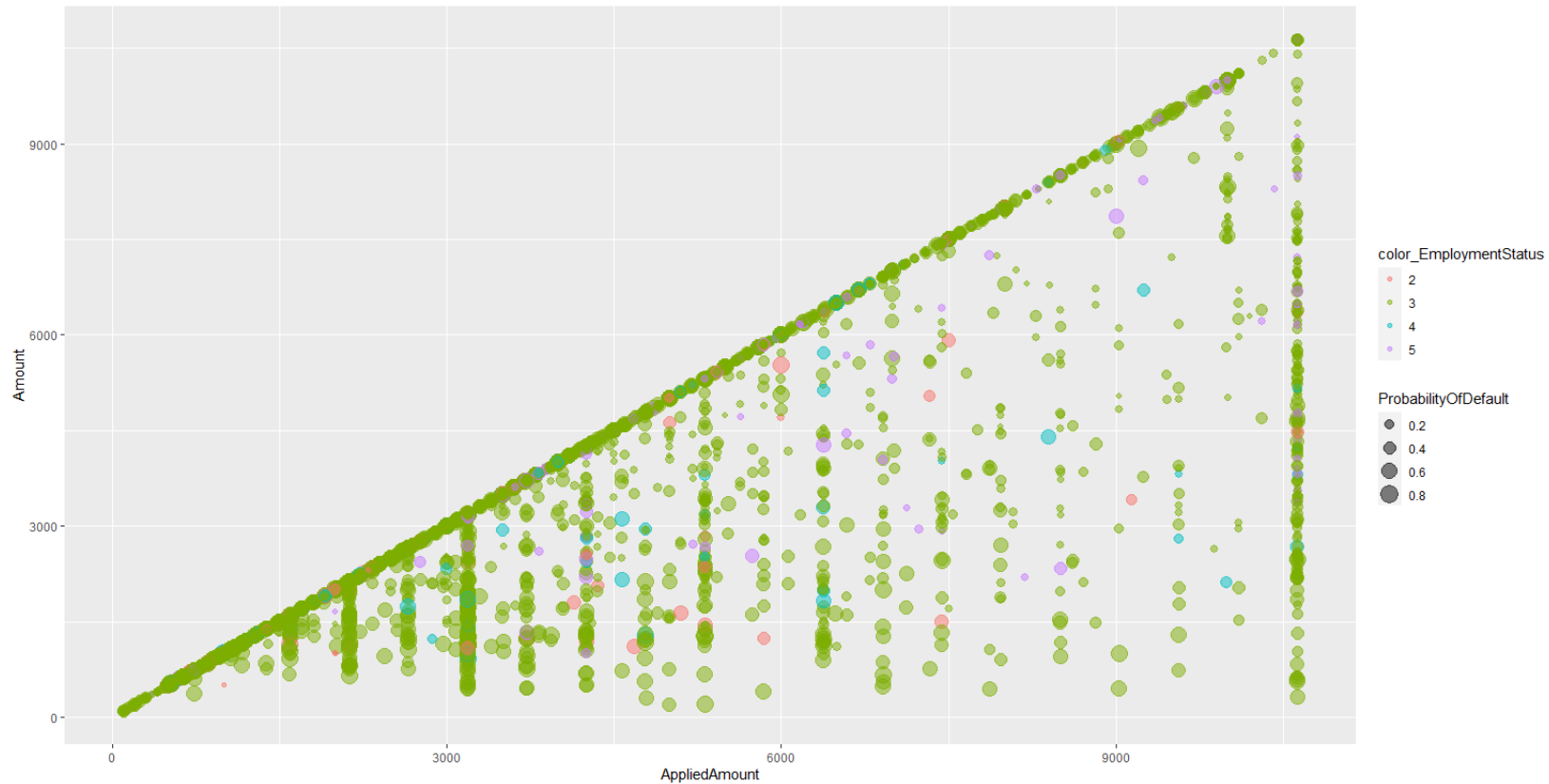


Le graphique montre la présence de valeurs aberrantes (très éloignées des autres valeurs), ce sont les points noirs situés en dehors des moustaches. Toutefois, il faut préciser que même si ces points sont éloignés ils ne constituent pas des erreurs, puisqu'ils ont été générés par le modèle utilisé par Bondora et donc qu'il n'y a pas besoin de les exclure de notre base de données. De fait, Ce n'est pas parce qu'une personne gagne 16000€ qu'elle ne demande pas de prêt, ce n'est pas non plus parce que le rendement anticipé d'un prêt à sa signature est très élevé qu'il continuera de l'être durant toute sa durée de vie, c'est le cas des rendements négatifs que nous avons dans notre jeu de données. D'ailleurs avoir des valeurs négatives de la variable ExpectedReturn nous permettra de ne pas trop être optimiste sur la rentabilité de notre projet. Dans la pratique, on constitue des réserves financières destinées à se prémunir de potentielles pertes de l'investissement (couverture du risque auquel on est exposé) et le coût de ces réserves est supporté indirectement par l'emprunteur sous forme de charges d'intérêt. La couverture des risques des investissement est une vraie contrainte juridique pour les institutions financières conventionnelles. Voir Bâle 1, 2 et 3 <sup>5</sup>.

---

<sup>5</sup> <chrome-extension://efaidnbmninnbpcajpcgglefindmkaj/viewer.html?pdfurl=https%3A%2F%2Facpr.banque-france.fr%2Fsites%2Fdefault%2Ffiles%2Fmedias%2Fdocuments%2F20170125-bale.pdf&clen=1045587&chunk=true>

# -Graphique de la distribution de la variable Amount selon EmploymentStatus



La taille de tout point du graphique (AppliedAmount, Amount) est proportionnelle à la probabilité de défaut de l'emprunteur et sa couleur dépend de la catégorie socioprofessionnelle de l'emprunteur. Il en ressort donc visuellement que les employés à temps plein (3 : couleur verte) sont les plus nombreux parmi les emprunteurs ensuite vient la modalité 5 des entrepreneurs. La taille de nos points est généralement grande ce qui signifie que la probabilité de défaut des emprunteurs est élevée (plus de 0,2).

Nombre d'emprunteurs selon la modalité du statut de l'emploi des emprunteurs (EmploymentStatus)			
2	3	4	5
281	10117	186	449

On peut aussi constater sur ce nuage de points que les deux variables ont tendance à évoluer dans le même sens. Pour une bonne partie des points, plus le montant de l'emprunt souhaité est élevé plus le montant du prêt accordé est grand.

## -Nuage de points de la distribution de la variable Amount selon le rendement anticipé du prêt



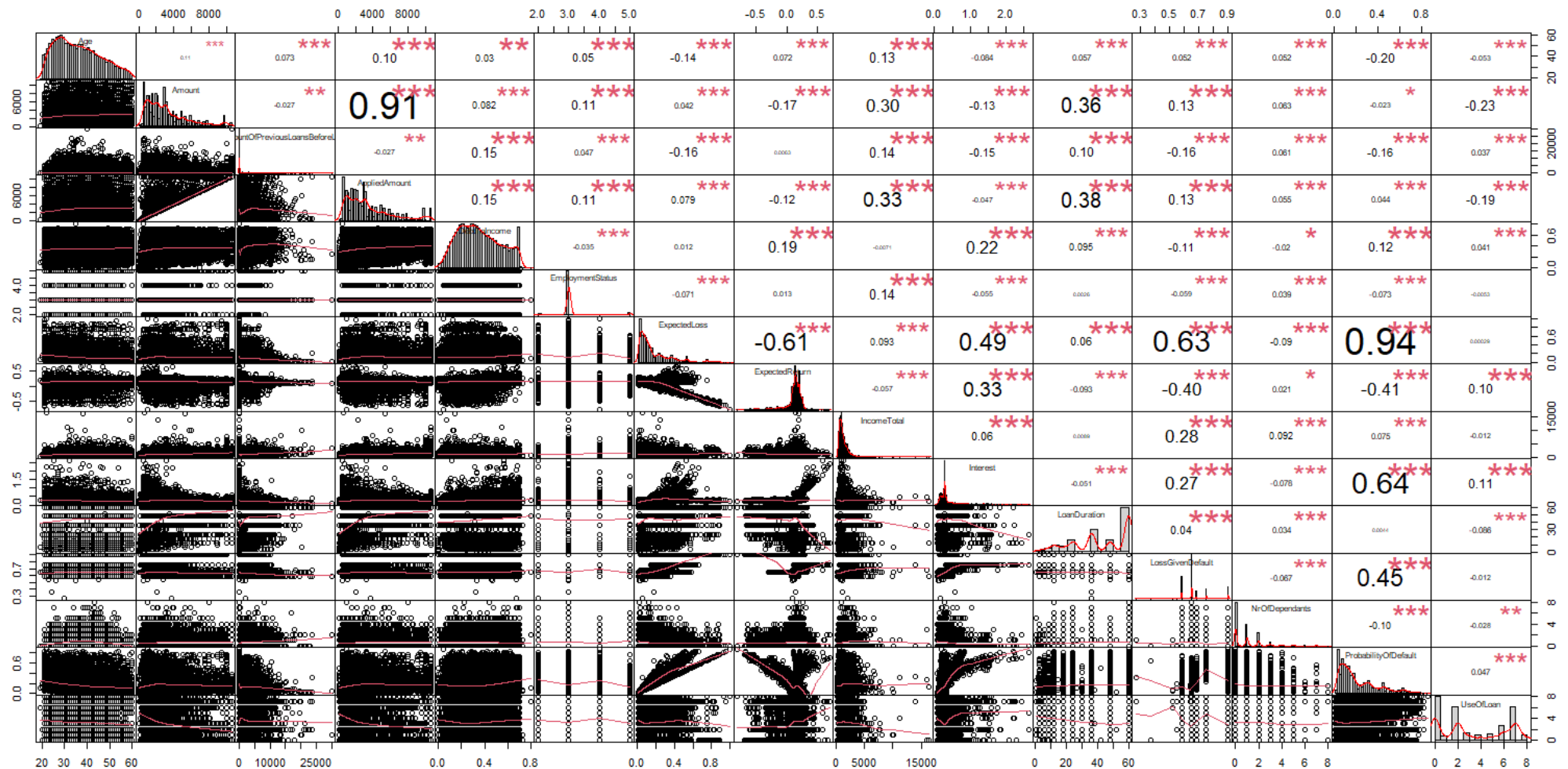
On observe que le rendement des prêts à forte probabilité de défaut est élevé (les points sont plus grands et sont éloignés de la valeur 0 de ExpectedReturn). En effet, le rendement est la rémunération du risque, tout prêteur rationnel exigerait un rendement élevé pour se prémunir des risques auquel il s'expose.

On remarque aussi que les prêts à court terme sont concentrés autour de 0 et les points y sont de petite taille. De fait, les prêts à court terme sont moins rentables ce qui signifie que la probabilité de réaliser un gain s'accroît avec la durée de placement.

### c. Analyse de la corrélation entre les variables

Nous souhaitons vérifier s'il existe bien une relation linéaire entre les variables de notre jeu de données. En d'autres termes, avant de procéder à la modélisation de la variable Amount, nous allons d'abord examiner graphiquement la distribution de Amount en fonction de chacune des variables prédictives afin d'avoir une idée de leur relation. Compte tenu du nombre élevé des variables, nous allons utiliser l'un des outils que propose PerformanceAnalytics pour l'analyse de performance et des risques d'un placement.


Nous allons nous focaliser sur la deuxième ligne et la deuxième colonne du graphique correspondant à la description de la nature statistique du lien entre notre variable cible et chacune des autres variables (même si nous avons supposé au début que les variables retenues justifiaient le montant à prêter).



La colonne 2 contient pour chacune de ses cellules sauf les 2 premières les nuages de points de la variable Amount avec chacune des autres. Dans la quatrième cellule de la colonne 2 du quadrillage par exemple nous avons les nuages de points de Amount avec AppliedAmount (le même que celui qu'on a réalisé supra avec des points aux couleurs et tailles différentes) avec une ligne des valeurs ajustés de la relation des 2 variables. A partir de la troisième cellule de la deuxième ligne nous avons les coefficients de corrélation de Pearson qui reflète la relation linéaire ainsi que le résultat du test de cette relation sous forme d'étoiles entre Amount et chacune des autres variables.

Les valeurs négatives (lien négatif) du coefficient de corrélation de Amount avec les variables (ExpectedReturn :-0.17 et ProbabilityOfDefault : -0.023) par exemple signifient que lorsque ces variables augmentent le montant du prêt lui diminue. En effet, lorsque les chances qu'un emprunteur fasse défaut sont élevées il a moins de chance contracté un prêt de grand montant, aussi un rendement élevé est associé à un risque de perte élevé.

En revanche, les valeurs positives (corrélation positive) de l'indice de corrélation indiquent que le montant du prêt accordé avec la valeur des variables comme AppliedAmount, le revenu total ou le statut de l'emploi varient ensemble dans le même sens.

Le nombre d'étoiles traduit le degré de significativité du lien entre les deux variables. Par exemple AppliedAmount a trois étoiles  ce qui signifie que sa relation linéaire avec Amount est assez pertinente contrairement à la variable AmountOfPreviousLoansBeforeLoan ou la probabilité de défaut.

L'analyse exploratoire des données nous a permis d'approfondir nos connaissances sur notre base de données. Nous avons pu voir qu'il existe bel et bien une relation linéaire entre notre variable cible et chacune des autres variables quand bien même nous l'avions affirmé implicitement en disant que nous voudrions prédire Amount en fonction des autres variables.

En conséquence, nous pouvons passer à la construction du modèle pour la prédiction du montant des prêts que nous devons accorder à nos emprunteurs compte tenu des valeurs que prendront nos variables explicatives.

## V-Modélisation de la variable Amount

Nous souhaitons construire un modèle de prévision du montant optimal à prêter à tout emprunteur potentiel en fonction de nos variables explicatives et ensuite nous allons vérifier qu'il est performant avant de l'utiliser sur notre plateforme pour éviter des pertes financières. Pour ce faire il nous faut diviser notre jeu de données en deux échantillons distincts, l'un servira à la construction et l'ajustement du modèle (entraînement du modèle) et l'autre à tester la performance (qualité) du modèle et à la prévision. Nous nous imposons cette démarche afin d'éviter des conclusions hâtives et optimistes des performances de notre modèle qui résulteraient de l'entraînement et du test de la qualité de notre modèle sur l'ensemble des données et qui pourraient s'avérer dangereuses pour notre projet de plateforme.

L'échantillon pour l'entraînement du modèle s'appellera `donnees_train` et contiendra 70 % des données de l'étude (`Bondora_donnees_cleaned.csv`) alors que celui pour le test de la performance du modèle obtenu de l'entraînement `donnees_test` et contiendra 30% soit 3310 observations des données de l'étude. Précisons aussi que vu qu'il n'existe aucun tri sur notre jeu de données il n'y aura pas de biais dans la composition de nos deux échantillons.

Nous retenons la régression linéaire pour la construction de notre modèle, même s'il existe de nombreuses méthodes de construction d'un modèle de prévision. Nous supposons que Bondora favorise les emprunteurs selon le statut de leur emploi et aussi selon le motif de l'emprunt (valeur de la variable `UseOfLoan`) et donc que les emprunts pour motif de santé sont les plus favorisés (8)

Avec notre régression, nous souhaiterions vérifier que notre variable à expliquer, `Amount`, s'écrive en fonction des autres variables de notre jeu de données que nous avons supposé étant déterministes de la variable `Amount`. Soit :

***Amount*** = *f*(Age, AmountOfPreviousLoansBeforeLoan, AppliedAmount, DebtToIncome, EmploymentStatus, ExpectedLoss, ExpectedReturn, IncomeTotal, Interest, LossGivenDefault, NrOfDependants, ProbabilityOfDefault, UseOfLoan )

En effectuant notre régression nous obtenons l'équation suivante :

```
> modele <- lm(Amount ~ ., donnees_train)
> summary(modele)

Call:
lm(formula = Amount ~ ., data = donnees_train)

Residuals:
    Min       1Q   Median       3Q      Max
-8819.4  -145.1    34.6   289.8  1764.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.767e+02  1.211e+02   8.067  8.3e-16 ***
Age          -2.514e+00  9.391e-01  -2.677  0.007449 **
AmountOfPreviousLoansBeforeLoan -1.052e-02  4.473e-03  -2.351  0.018763 *
AppliedAmount  8.756e-01  4.401e-03 198.979 < 2e-16 ***
DebtToIncome  -2.145e+02  6.191e+01  -3.465  0.000533 ***
EmploymentStatus  3.993e+01  2.095e+01   1.906  0.056687 .
ExpectedLoss   9.481e+03  1.514e+04   0.626  0.531210
ExpectedReturn  6.404e+03  1.514e+04   0.423  0.672276
IncomeTotal   -1.318e-02  1.455e-02  -0.906  0.365196
Interest      -7.812e+03  1.514e+04  -0.516  0.605845
LoanDuration   7.248e-01  5.373e-01   1.349  0.177379
LossGivenDefault -2.392e+02  1.474e+02  -1.623  0.104553
NrOfDependants   9.915e+00  8.825e+00   1.123  0.261268
ProbabilityOfDefault -3.091e+03  2.588e+02 -11.941 < 2e-16 ***
UseOfLoan      -3.080e+01  3.167e+00  -9.726 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 766.1 on 7708 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.8972
F-statistic: 4817 on 14 and 7708 DF,  p-value: < 2.2e-16
```

**Amount= -2.51380Age -0.01051AmountOfPreviousLoansBeforeLoan +0.87561AppliedAmount -214.526DebtToIncome +39.92906EmploymentStatus +9481.495ExpectedLoss +6404.222ExpectedReturn -0.01318IncomeTotal -7812.221Interest +0.72477LoanDuration -239.2475LossGivenDefault +9.9145NrOfDependants -3090.721ProbabilityOfDefault -30.7978UseOfLoan + 976.72.**

Selon cette équation, lorsque pour une quelconque raison une des variables a coefficient négatif (Interest, ProbabilityOfDefault par exemple) augmente en gardant toutes les autres variables du jeu de données constantes le montant du prêt devant être octroyé à l'emprunteur diminuerait. De même, l'augmentation de toute variable à coefficient positif en gardant les autres constantes aura tendance à augmenter le montant du prêt à accorder. Toutefois, l'importance du pouvoir explicatif des variables que nous avons décidé d'utiliser pour déterminer le montant de l'emprunt obtenu par l'emprunteur ne pourra être confirmée que lorsque nous aurons effectué un test de significativité de ces variables. Pour cela un test de Student est habituellement utilisé.



Le test de Student nous permet de savoir lesquelles des variables explicatives sont significatives (significativement différent de 0). Dans les sorties de l'estimation des paramètres en R, la valeur de  $\Pr(> |t|)$  donne directement le résultat de ce test. Plus la valeur 'p' est faible, plus la variable est significative. En effet, les \*\*\* signifient que la variable est hautement significative car la valeur p correspondante ( $\Pr(> |t|)$ ) est inférieure à 1/1000, très significative quand elle inférieure à 1/100 soit \*\* et significative dès qu'elle est inférieure à 5/100.

Nous constatons ainsi que seules les variables AppliedAmount, DebtToIncome, ProbabilityOfDefault, UseOfLoan sont hautement significatives dans la justification du montant à prêter. De ce fait, nous allons construire un modèle final en excluant les autres variables non significatives puisque leurs variations n'affectent pas significativement le montant qu'on devrait prêter à notre emprunteur sur notre plateforme.

```
> summary(modele_final)

Call:
lm(formula = Amount ~ AppliedAmount + DebtToIncome + ProbabilityOfDefault +
    UseOfLoan, data = donnees_train)

Residuals:
    Min       1Q   Median       3Q      Max
-9000.4  -111.2    66.3   292.3  1369.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.062e+02  2.676e+01   22.66  <2e-16 ***
AppliedAmount  8.894e-01  3.717e-03  239.25  <2e-16 ***
DebtToIncome  -6.086e+02  6.002e+01  -10.14  <2e-16 ***
ProbabilityOfDefault -5.797e+02  5.390e+01  -10.76  <2e-16 ***
UseOfLoan     -3.644e+01  3.242e+00  -11.24  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 793.5 on 7718 degrees of freedom
Multiple R-squared:  0.8898,    Adjusted R-squared:  0.8898
F-statistic: 1.558e+04 on 4 and 7718 DF,  p-value: < 2.2e-16
```

-Modèle final

**Amount= 606.2133+ 0.8893676 \*AppliedAmount -608.5902 \* DebtToIncome -579.7249 \* ProbabilityOfDefault -36.43787 \* UseOfLoan.**

Le coefficient de détermination ( $R^2$  ou encore R-squared) est la métrique qui permet de juger la précision de la prédiction d'une régression linéaire. Il varie de 0 à 1 et mesure l'adéquation entre le modèle obtenu de la régression et les données observées, c'est à dire à quel point l'équation de régression est adaptée pour décrire la distribution des points dans nos données. Plus il est proche de 1, mieux cela vaut.

Dans notre analyse, nous avons obtenu un coefficient de détermination de 0,8898 on en conclut donc que le modèle obtenu est de meilleure qualité en termes de précision de la prédiction de notre variable Amount d'autant plus que la part de variation dans la variable du montant de l'emprunt qui est expliquée par des variations dans les variables explicatives qui est exprimée par le coefficient de détermination est proche de 1.

Cependant, même si ce coefficient nous indique qu'il y a une forte relation entre le montant de l'emprunt et les variables explicatives, nous n'en savons rien de sa significativité statistique à moins d'analyser la significativité globale de notre modèle à travers le test de Fisher.

#### -Test de significativité globale du modèle

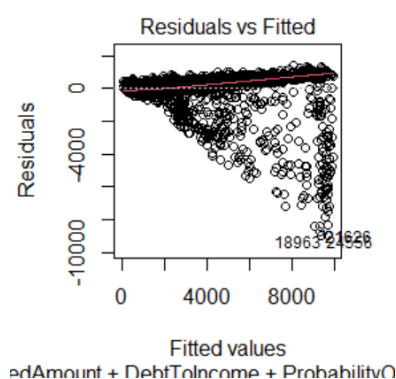
Ce test repose sur la statistique de Fisher. Si la p-value qui lui est associée est inférieure à 1%, on conclut que notre modèle est globalement significatif.

La p-value de la statistique de Fisher de notre modèle vaut  $2.2 \cdot 10^{-16}$  notre modèle est donc significatif.

Pour valider notre modèle afin qu'il nous serve d'outil de prédiction il nous faut absolument réaliser une étude des écarts (résidus) entre les valeurs de la variable à expliquer se trouvant dans `donnees_train` et celles obtenues à l'aide de notre modèle.

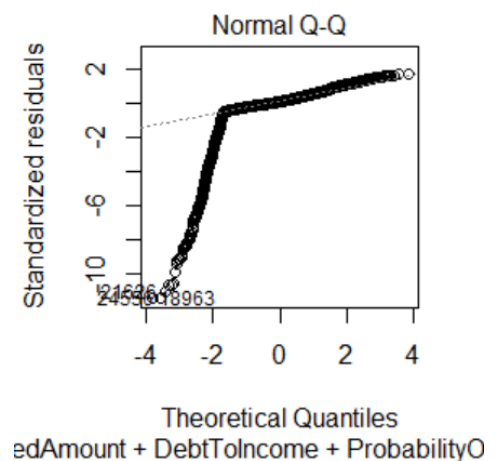
Si les hypothèses de linéarité, de normalité, d'homogénéité et d'indépendance des résidus sont satisfaites alors les résultats de la régression sont valides et on pourra ainsi utiliser notre modèle pour la prédiction du montant optimal à prêter.

Les informations absentes du modèle sont résumées par le terme de l'erreur. Le tracé des résidus (la différence de la valeur de Amount se trouvant dans `donnees_train` et celle prouite par le modèle) doivent être de nature aléatoire et il ne doit pas y avoir de modèle dans le graphique. La moyenne du des résidus (l'écart qu'on espère obtenir) doit être proche de zéro. Dans le graphique ci-dessous, nous pouvons voir que la ligne de tendance rouge est presque à zéro, sauf au point de départ.

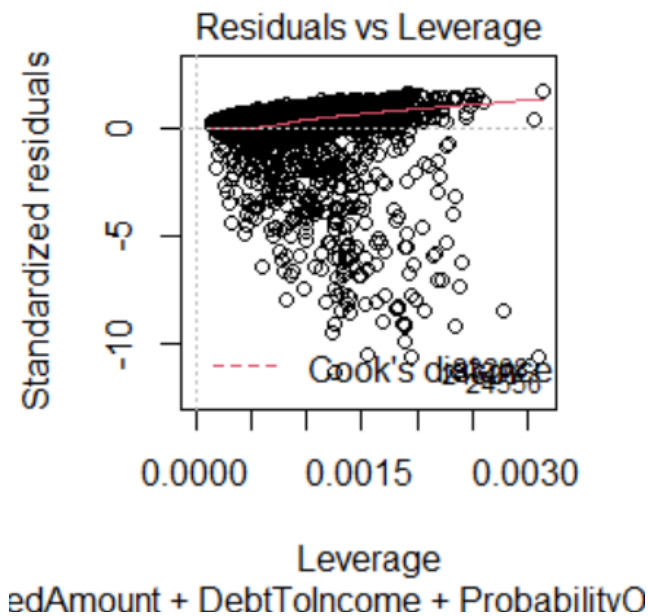


#### Test de normalité des résidus – Droite de Henry

La droite de Henry est un diagramme quantile-quantile (Q-Q plot) confrontant les quantiles observés avec ceux que l'on aurait sous hypothèse de normalité (quantiles théoriques). S'ils sont cohérents, c.-à-d. forment une droite, l'hypothèse est crédible. Dans le graphique ci-dessous, nous voyons que la plupart des graphiques sont sur la ligne, sauf vers la fin.



Ce graphique permet de trouver les observations influentes. Ici, nous devons vérifier les points qui sont en dehors de la ligne en pointillée. Un point à l'extérieur de celle-ci sera un point influent et sa suppression affectera les coefficients de régression



Aussi, pour s'assurer de l'indépendance des résidus, on effectue le test de Durbin-Watson. La statistique qu'on obtiendra du test devra être proche de 2 pour conclure que les résidus ne sont pas dépendants.

Pour notre modèle, la statistique de Durbin-Watson vaut  $1.84 \sim 2$  ce qui signifie donc qu'il n'y a pas de corrélation entre nos résidus et l'écart entre les valeurs de Amount dans données\_train et celles observées n'est qu'un effet de hasard.

Il est assez fréquent dans un modèle que certaines des variables explicatives fournissent la même information sur la variable à expliquer ou qu'elles soient liées. Les conséquences d'une telle situation est la difficulté d'interprétation des coefficients des dites variables dans le modèle. Pour s'assurer que ce n'est pas le cas de notre modèle, nous allons évaluer le facteur d'inflation de la variance (FIV) de nos variables explicatives afin de savoir de combien augmenterait la variance d'un des coefficients de la régression en raison d'une potentielle relation linéaire avec un autre. L'idéal pour nous serait que les valeurs des FIV soient suffisamment proches de 1 ainsi, on en déduira d'une absence de relation linéaire entre les variables explicatives de notre modèle et donc que chacune apporte une information distincte sur le montant à prêter.

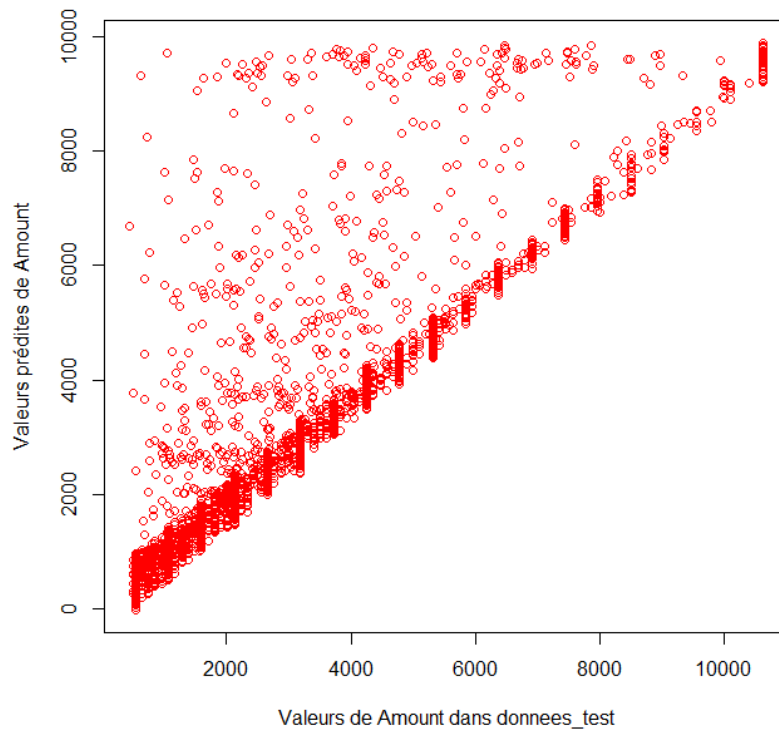
```
vif(modele_final)
```

AppliedAmount	DebtToIncome	ProbabilityOfDefault	UseOfLoan
1.079163	1.017617	1.008021	1.055421

Nous constatons que toutes les FIV des variables explicatives de notre modèle sont très proches de 1, alors il n'y a pas de relation linéaire entre elles.

Toutes les hypothèses et les conditions de stabilité d'un modèle de régression étant vérifiées pour notre modèle, nous pouvons procéder à présent à la prédiction de notre variable de réponse Amount avec l'équation de notre modèle.

## Valeurs prédites de Amount vs valeurs observées.



### d. Statistiques des valeurs prédites de Amount

```
> summary(prediction_IC)
```

fit	lwr	upr
Min. : -9.848	Min. : -90.24	Min. : 70.55
1st Qu.: 1333.235	1st Qu.: 1292.85	1st Qu.: 1369.82
Median : 2487.260	Median : 2433.19	Median : 2527.62
Mean : 3121.678	Mean : 3077.29	Mean : 3166.06
3rd Qu.: 4152.032	3rd Qu.: 4111.19	3rd Qu.: 4197.05
Max. : 9881.264	Max. : 9815.87	Max. : 9946.66

La moyenne des données prédites du montant optimal à prêter est de 3121,678. Pour tout emprunt futur, cette moyenne variera dans l'intervalle [3077.29 ; 3166,06].

La moyenne des montants des prêts accordés que nous avons calculé au début de notre étude était de 3081,925. On constate donc que la distribution de nos valeurs prédites ne s'écarte pas trop des observées. Il en est de même pour les 1<sup>er</sup> et 3<sup>em</sup> quartiles.

## VI- Conclusion

Pour tout emprunteur qui voudra emprunter sur notre plateforme de P&P Lending, le montant du prêt qui lui sera proposé dépendra du montant qu'il souhaite emprunter, de la part de son revenu qu'il alloue au remboursement de ses dettes, de sa probabilité de défaut et du motif qu'il fournira.

$$\text{Amount} = 606.2133 + 0.8893676 * \text{AppliedAmount} - 608.5902 * \text{DebtToIncome} - 579.7249 * \text{ProbabilityOfDefault} - 36.43787 * \text{UseOfLoan}.$$

Selon notre modèle, la sensibilité du montant du prêt accordé à un emprunteur au montant de l'emprunt qu'il souhaitait initialement emprunter quand il a formulé sa demande sur la plateforme est de 88,93%.

Pour une personne qui n'a aucune dette à sa charge (**DebtToIncome=0**) et qui souhaiterait emprunter 500€ (**AppliedAmount**) dans le but de s'acheter une formation professionnelle (**UseOfLoan=4**), le montant maximal du prêt qui lui sera accordé au cas où nous estimerions qu'il y a 50 % qu'il fasse défaut (**ProbabilityOfDefault**) sera de 615,28€ (**Amount**). En revanche, si nous estimons qu'il y a 80% de chances qu'il ne soit pas en mesure de nous rembourser nous ne le prêterions que 441.36€.

Pour la constitution d'un capital prévisionnel afin d'assurer le fonctionnement de notre plateforme en jouant le rôle de prêteur jusqu'à sa popularité et éviter les problèmes de circularité dont nous craignons pour notre projet, nous allons nous baser sur les données que nous avons prédites.

Si nous espérons toucher 3310 emprunteurs au lancement de notre plateforme, il nous faudra absolument constituer une réserve financière au moins égale à 10 185 842€ sans dépasser les 10 479 665 € la première année. Avec une croissance de 50% chaque année, pour la seconde année il nous faudra constituer un montant prévisionnel minimal de 15 278 763 € et un montant prévisionnel maximal de 15 719 498 €.