Should this loan be approved or denied?

Data Mining (GEI)

Spring semester, course 2021/22

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Index

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Motivation

When we started to think of a topic we all liked and were interested in, we couldn't find a specific topic that appealed to all of us. That's why we decided to start scrolling through different data sets until we found one that seemed interesting enough to make a project about. After almost an entire laboratory session we ended up with a topic that, at first sight, might not be much interesting, but we all agreed that it was striking enough to choose it. In the end, loans is something that we all know a little about and, looking to the future, is a topic that will be useful to increase our knowledge in. The rest of the sessions, we were reading the metadata and finished understanding in more detail the subject of the topic we picked.

Moreover, loans and financial topics are always interesting subjects to learn about and, despite being centered in the US, we are sure that we will be able to understand how loans work depending on the type of company, how large it is, if it is new… Probably, if we took the same samples in Spain, the results would vary slightly but the general idea and concept will be applicable in the same way. Although we know that the learning curve at the beginning will be complicated, due to banking concepts, terms and other peculiarities, we are sure that the subject is captivating enough to do a work about it.

Data Source presentation

We got our data by searching in the kaggle website datasets with a theme that was interesting to all the members of our group and was big enough so that we could work properly with it. Once we found a dataset that fulfilled our needs we downloaded it and loaded it using Rstudio so that we could work with it and make the necessary changes. Using Rstudio we reduced the size of the dataset to obtain a similar structure of data matrix as the one the assignment requires.

The dataset chosen gathers information about businesses in the United States that applied for a bank loan. For every observation of the dataset there is information about the name and location of the business and also its type. We also have information about the back that supplied the loan and the year of approval. About the business we also have information about the location (if it's rural or not) and the number of business employees. Finally we have information about the loan: the quantity of money lent and whether the loan was paid in full or the business failed to pay.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Metadata

Data Description

The dataset of this project is a real dataset from the U.S. Small Business Administration (SBA). The U.S. SBA was founded in 1953 on the principle of promoting and assisting small enterprises in the U.S. credit market. Small businesses have been a primary source of job creation in the United States; therefore, fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment. One way SBA assists these small business enterprises is through a loan guarantee program which is designed to encourage banks to grant loans to small businesses. SBA acts much like an insurance provider to reduce the risk for a bank by taking on some of the risk through guaranteeing a portion of the loan. In the case that a loan goes into default, SBA then covers the amount they guaranteed.

There have been many success stories of start-ups receiving SBA loan guarantees such as FedEx and Apple Computer. However, there have also been stories of small businesses and/or start-ups that have defaulted on their SBA-guaranteed loans. The rate of default on these loans has been a source of controversy for decades. Conservative economists believe that credit markets perform efficiently without government participation. Supporters of SBA guaranteed loans argue that the social benefits of job creation by those small businesses receiving government guaranteed loans far outweigh the costs incurred from defaulted loans.

Since SBA loans only guarantee a portion of the entire loan balance, banks will incur some losses if a small business defaults on its SBA-guaranteed loan. Therefore, banks are still faced with a difficult choice as to whether they should grant such a loan because of the high risk of default. One way to inform their decision making is through analyzing relevant historical data such as the datasets provided here.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Description of variables

Variable	Modalities	Meaning	Type	Measuring unit	Missing code	Mesuring procedure	Range	Role
Name		Borrower name	Qualitative		NA			None
State	States of the US	State of the borrower	Qualitative		NA			Explanatory
Bank		Bank name	Qualitative		NA			None
BankState	States of the US	The state of the bank	Qualitative		NA			Explanatory
NAICS	Economic sectors in Annex 1	North American industry classification system code. The first two digits represent the economic sector.	Qualitative		NA			Explanatory
ApprovalFY		Fiscal year of commitment	Date		NA			Explanatory
Term		Loan term in months	Qualitative		NA			Explanatory
NoEmp		Number of business employees	Numeric		NA			Explanatory
NewExist	1=Existing business, 2 = New business	Type of business	Binary		NA			Explanatory

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

CreateJob		Number of jobs created	Numeric		NA			Explanatory
RetainedJob		Number of jobs retained	Numeric		NA			
UrbanRural	1 = Urban, 2 = rural, 0 = undefined	Location of business	Qualitative		NA			Explanatory
RevLineCr	Y = Yes, N = No	Revolving line of credit	Qualitative		NA			Explanatory
DisbursmentGross		Currency amount disbursed	Currency	Dollars($)	NA			Explanatory
Balance Gross		Gross amount outstanding	Currency	Dollars($)	NA			Explanatory
MIS_Status	CHGOFF, PIF	Loan status charged off	Qualitative		NA			Explanatory
GrAppv		Gross amount of loan approved by bank	Currency	Dollars($)	NA			Explanatory
SBA_Appv		SBA'sguaranteed amount of approved loan	Currency	Dollars($)	NA			Explanatory
WhichCompany		Economic sector of the company	Qualitative		NA			Explanatory

Annex 1: First 2 digits of the NAICS representing the economic sector

Sector	Description
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts,entertainment and recreation
72	Accommodation and food services
81	Other services (except public administration)
92	Public administration

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Preprocessing

In this section it is explained the decisions that we have made in the preprocessing of the data frame. If a variable is not in this section is because any changes have been made to it because it was not necessary.

Name

We have eliminated this variable due to not providing any significant information.

Bank

We have eliminated this variable due to not providing any significant information.

BalanceGross

We have eliminated this variable due to not providing any significant information.

DisbursementGross

It was a character variable with a dollar symbol and we have eliminated the symbol and converted it to a numeric variable. We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

GrAppv

It was a character variable with a dollar symbol and we have eliminated the symbol and converted it to a numeric variable. We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

SBA_Appv

It was a character variable with a dollar symbol and we have eliminated the symbol and converted it to a numeric variable. We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

UrbanRural

It was a numerical variable and we have factored it and changed its values to autoexplicative ones.

WhichCompany

It is a new variable created by us from the first two digits of the variable NAICS. These numbers indicate which type of company it is.

NewExist

It was a numerical variable and we have factored it and changed its values to autoexplicative ones.

RevLineCr

It was a character variable and we have factored it. We have imputed a '0' in the variables with errors or missings and then eliminated the levels which had 0 individuals with that value.

MIS_Status

It was a character variable and we have factored it. We have eliminated the individuals with missing or errors instead of imputing their values because it is our response variable.

State

It was a character variable and we have factored it. We have changed their values to indicate the zone where the state is instead of the state itself.

BankState

It was a character variable and we have factored it. We have changed their values to indicate the zone where the state is instead of the state itself.

NAICS

The little information this variable had, we extracted it and put it in a new variable, the *WhichComapy* one.

yearsAfterAprov

It is a new variable created by us from the variable ApprovalFY, it indicates the years since the loan was approved. We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

NoEmp

We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

CreateJob

We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

RetainedJob

We have imputed NAs into the extreme outliers. Then we have imputed the values in the individuals with NA with the KNN method.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Basic statistical descriptive analysis

Initial Univariate data description

After deciding the dataset we were going to use, we did the following descriptive of the data. At this moment we have 18 variables and 5000 observations.

State

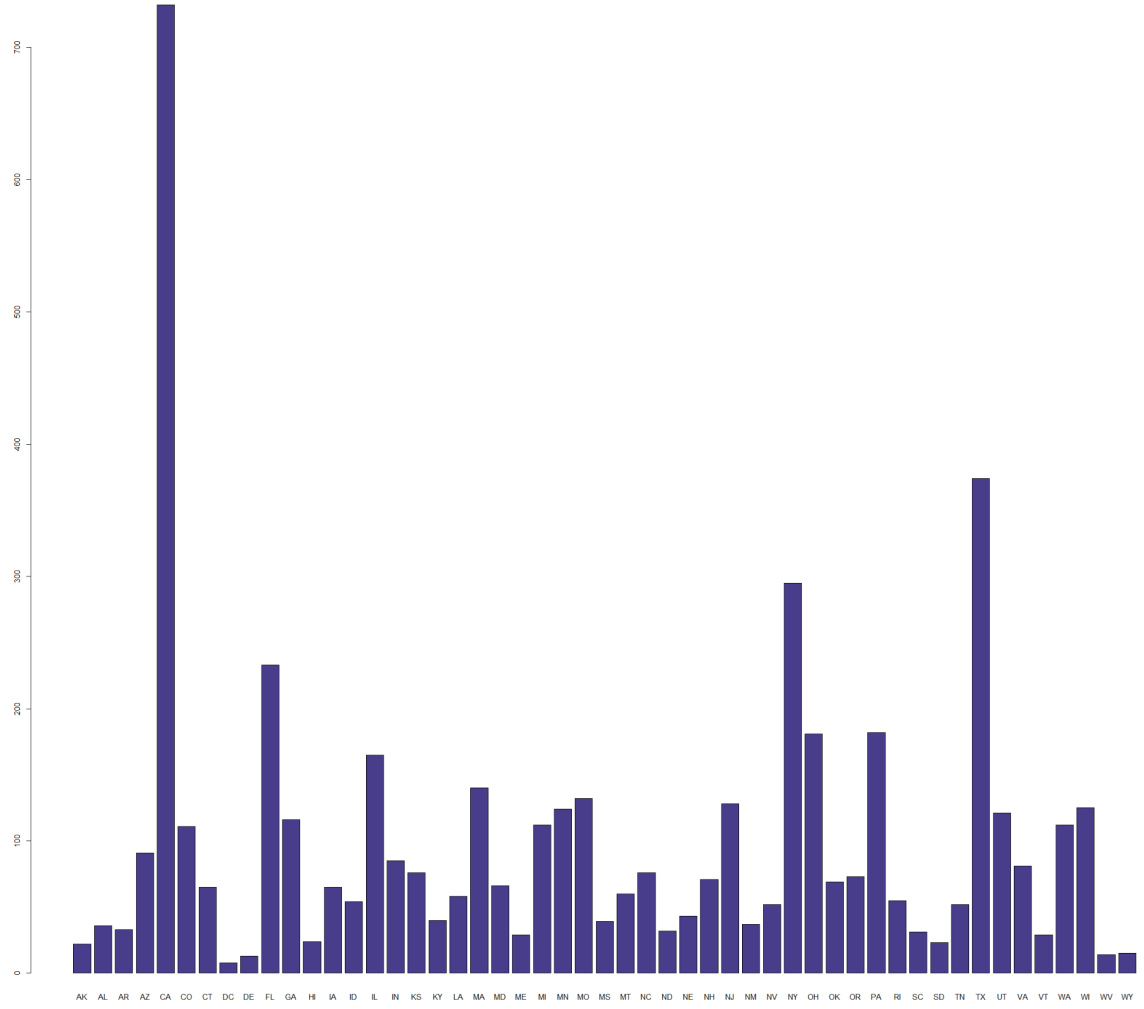


Image 1: State barplot

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

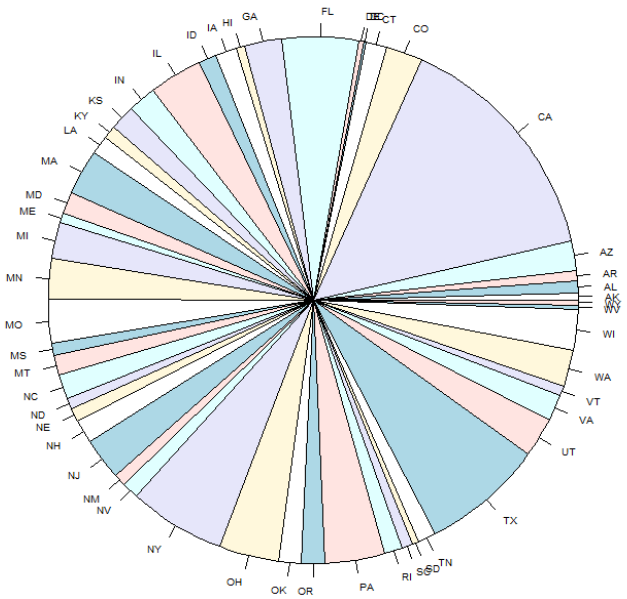


Image 2: Pie of State

Statistics of variable “State”	
Number of modalities	50
Frequency table of the modalities	AK AL AR AZ CA CO CT DC DE FL GA 22 36 33 91 732 111 65 8 13 233 116 HI IA ID IL IN KS KY LA MA MD ME 24 65 54 165 85 76 40 58 140 66 29 MI MN MO MS MT NC ND NE NH NJ NM 112 124 132 39 60 76 32 43 71 128 37 NV NY OH OK OR PA RI SC SD TN TX 52 295 181 69 73 182 55 31 23 52 374 UT VA VT WA WI WV WY 121 81 29 112 125 14 15
Proportions of modalities (out of 1)	AK AL AR AZ CA 0.0044 0.0072 0.0066 0.0182 0.1464 CO CT DC DE FL 0.0222 0.0130 0.0016 0.0026 0.0466 GA HI IA ID IL 0.0232 0.0048 0.0130 0.0108 0.0330 IN KS KY LA MA 0.0170 0.0152 0.0080 0.0116 0.0280 MD ME MI MN MO 0.0132 0.0058 0.0224 0.0248 0.0264 MS MT NC ND NE 0.0078 0.0120 0.0152 0.0064 0.0086 NH NJ NM NV NY 0.0142 0.0256 0.0074 0.0104 0.0590 OH OK OR PA RI 0.0362 0.0138 0.0146 0.0364 0.0110 SC SD TN TX UT 0.0062 0.0046 0.0104 0.0748 0.0242 VA VT WA WI WV 0.0162 0.0058 0.0224 0.0250 0.0028 WY 0.0030

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

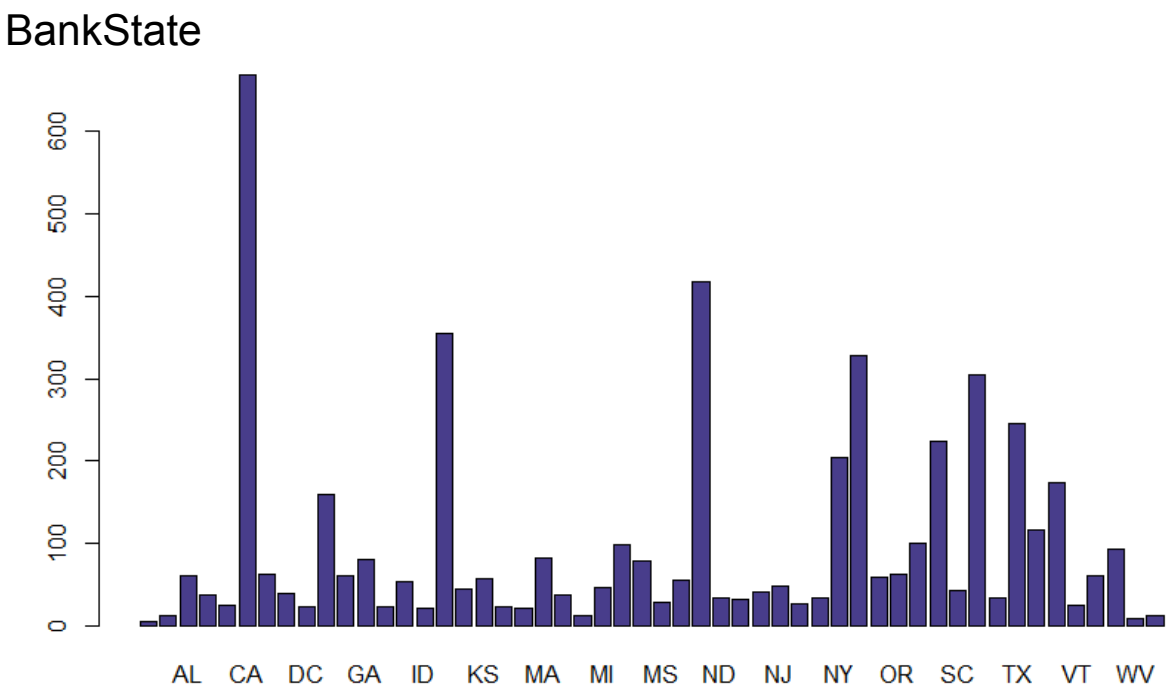


Image 3: BankState barplot

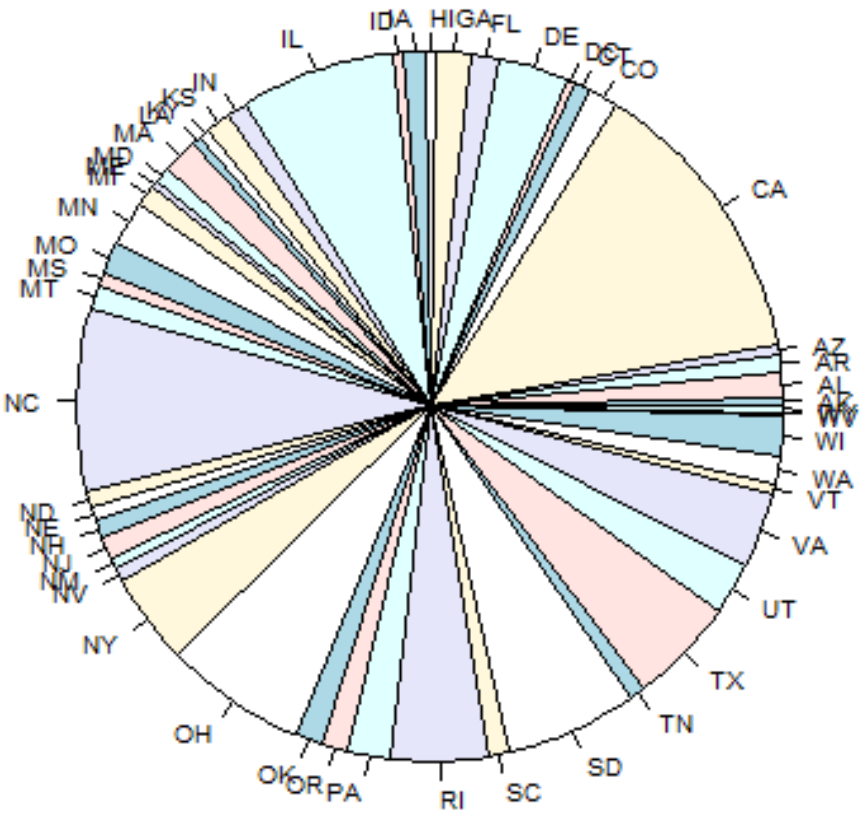


Image 4: Pie of BankState

Statistics of variable “BankState”	
Number of modalities	52
Frequency table of the modalities	AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND 5 13 61 37 25 668 63 39 23 159 61 81 23 54 22 355 45 57 23 22 82 38 13 47 99 78 28 56 418 34 NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY 32 41 48 27 34 204 327 58 62 100 224 43 305 33 245 116 173 25 60 93 9 12
Proportions of modalities (out of 1)	AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN 0.0010 0.0026 0.0122 0.0074 0.0050 0.1336 0.0126 0.0078 0.0046 0.0318 0.0122 0.0162 0.0046 0.0108 0.0044 0.0710 0.0090 KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM 0.0114 0.0046 0.0044 0.0164 0.0076 0.0026 0.0094 0.0198 0.0156 0.0056 0.0112 0.0836 0.0068 0.0064 0.0082 0.0096 0.0054 NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV 0.0068 0.0408 0.0654 0.0116 0.0124 0.0200 0.0448 0.0086 0.0610 0.0066 0.0490 0.0232 0.0346 0.0050 0.0120 0.0186 0.0018 WY 0.0024

ApprovalFY

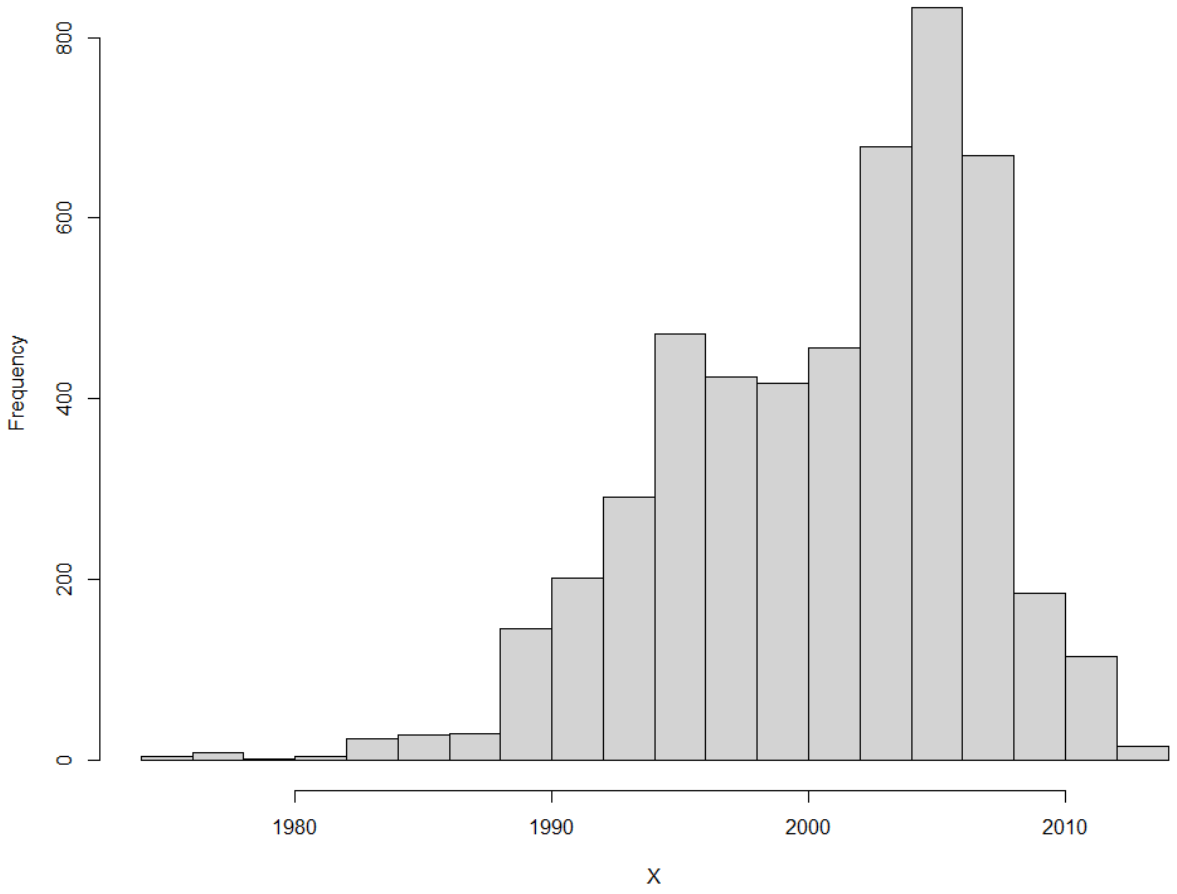


Image 5: Histogram of ApprovalFY

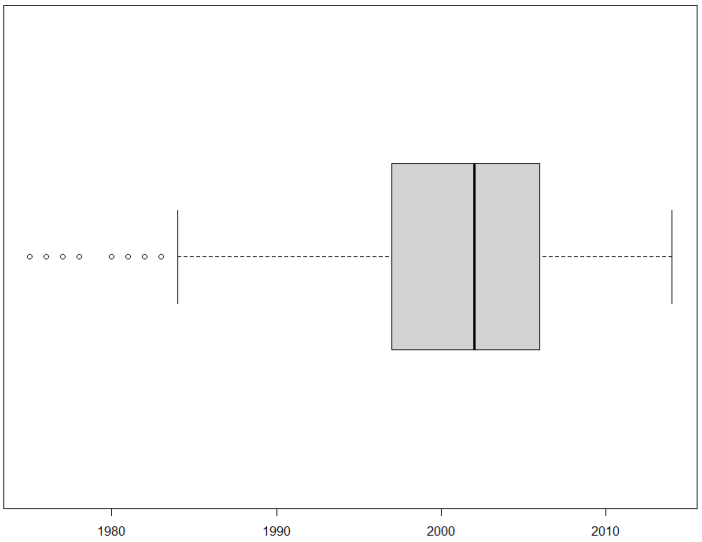


Image 6: Boxplot of ApprovalFY

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1975	1997	2002	2001	2006	2014

"sd: 5.98153782558898"
"vc: 0.00298902182953513"

Term

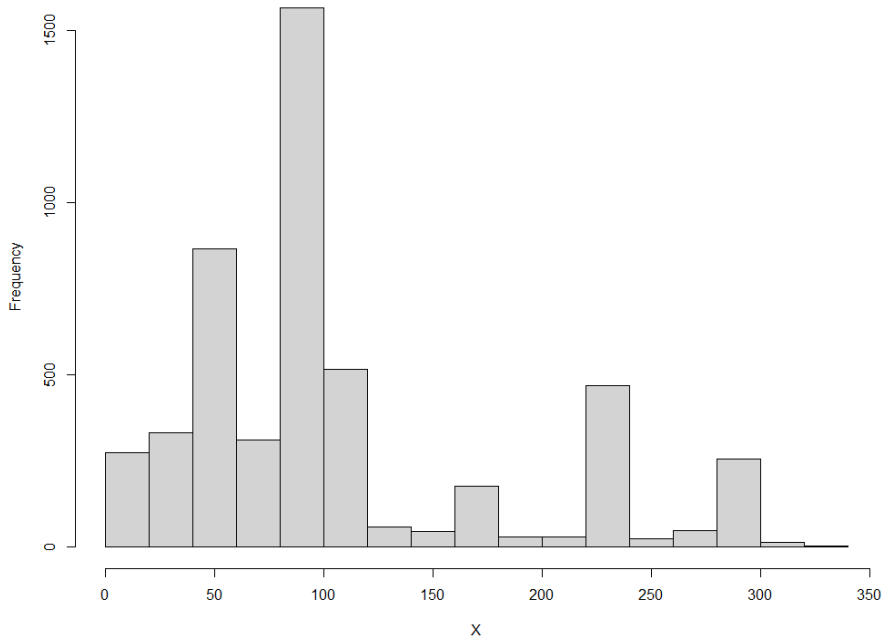


Image 7: Histogram of Term

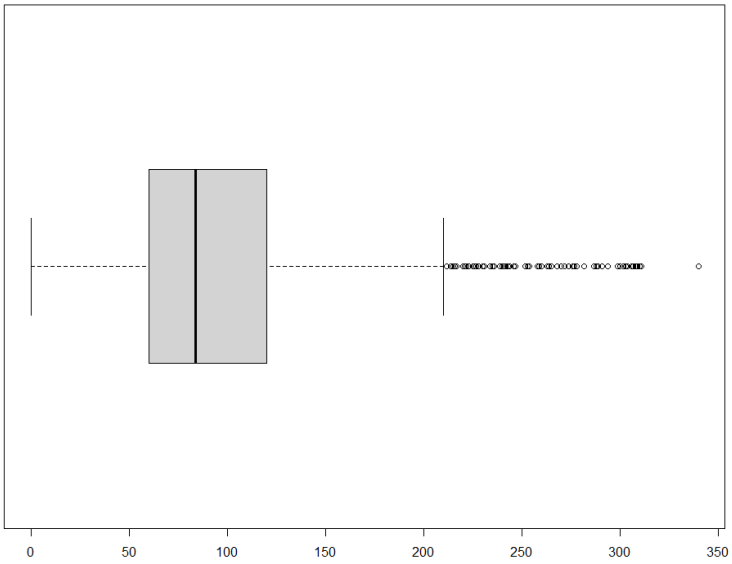


Image 8: Boxplot of Term

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	60	84	108.8	120	340

"sd: 76.5755441139746"
"vc: 0.703649894179661"

NoEmp

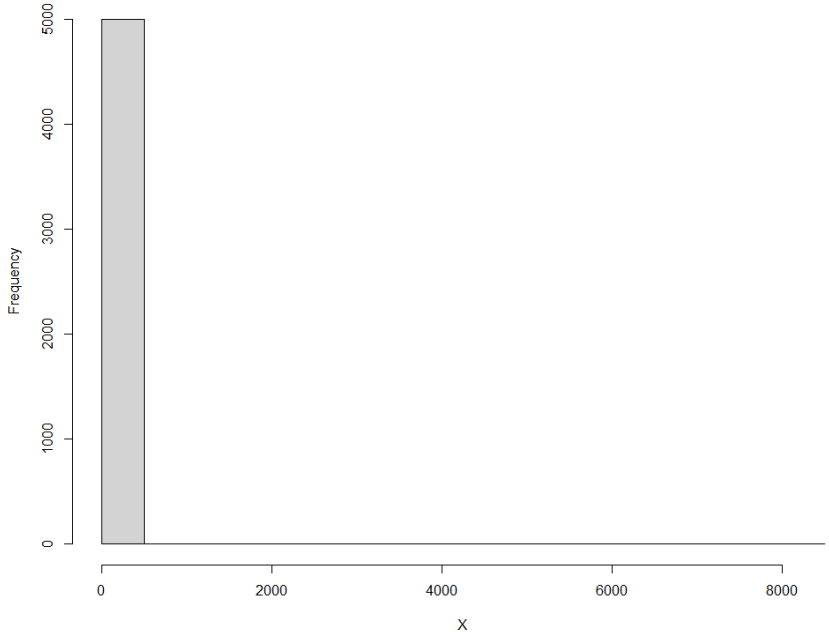


Image 9: Histogram of NoEmp

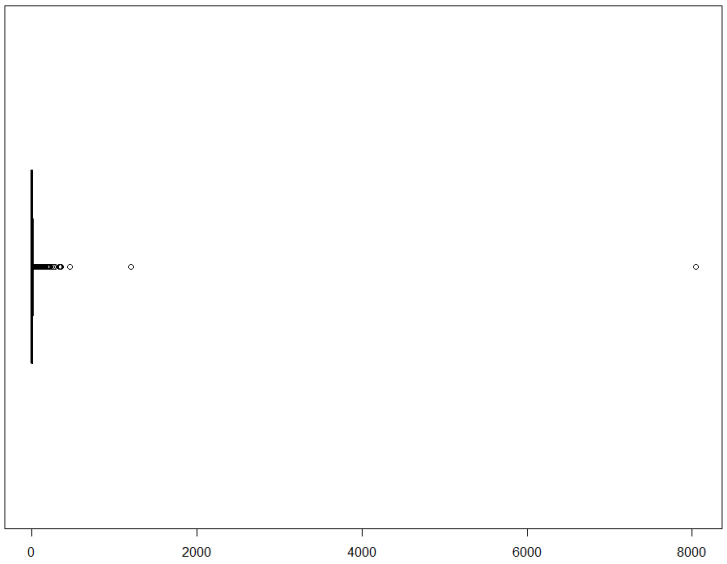


Image 10: Boxplot of NoEmp

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	2	4	11.96	10	8041

"sd: 116.721665974324"
"vc: 9.75591063124351"

NewExist

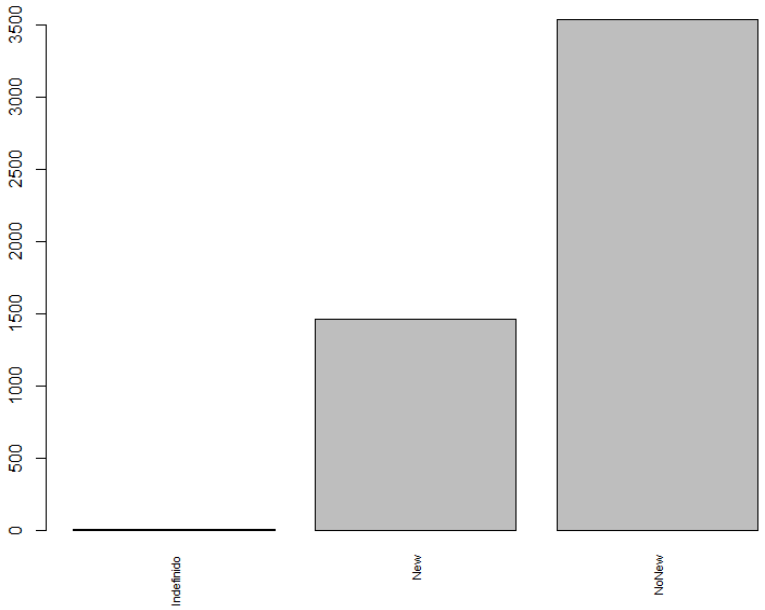


Image 11: Barplot of NewExist

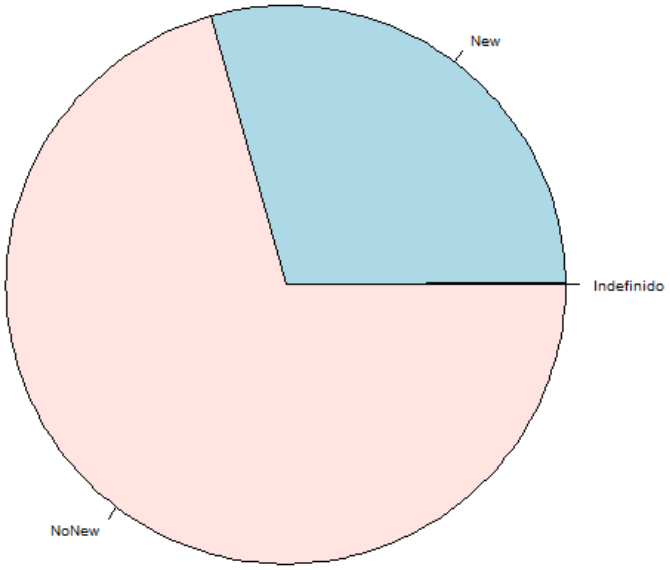


Image 12: Pie of NewExist

Statistics of variable “NewExist”			
Number of modalities	3		
Frequency table of the modalities	Undefined	NoNew	New
	6	3534	1460
Proportions of modalities (out of 1)	0.0012	0.7068	0.2920

CreateJob

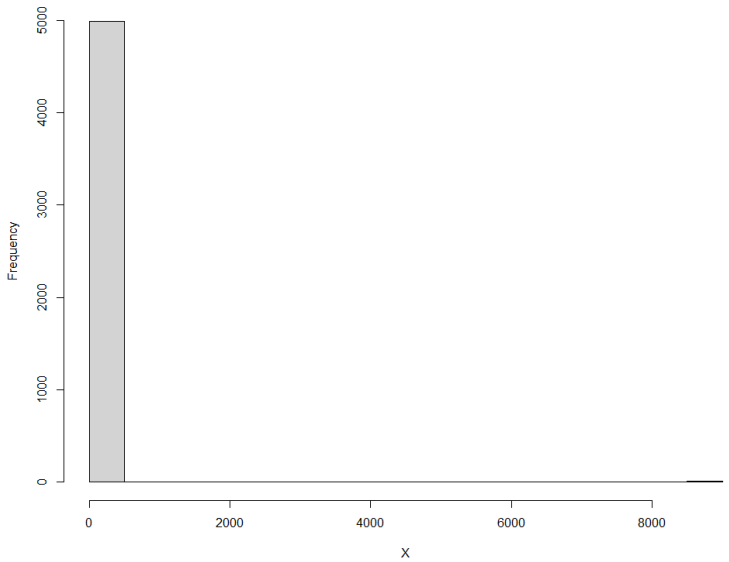


Image 13: Histogram of CreateJob

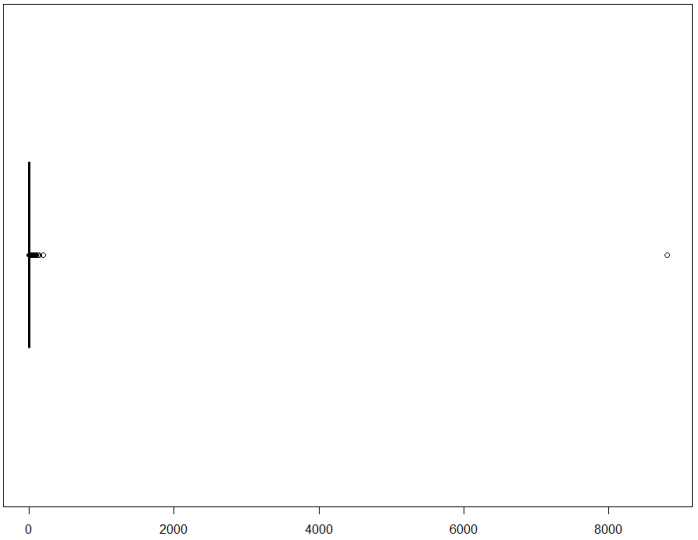


Image 14: Boxplot of CreateJob

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	12.49	1	8800

"sd: 304.701505775676"
"vc: 24.4022797058988"

RetainedJob

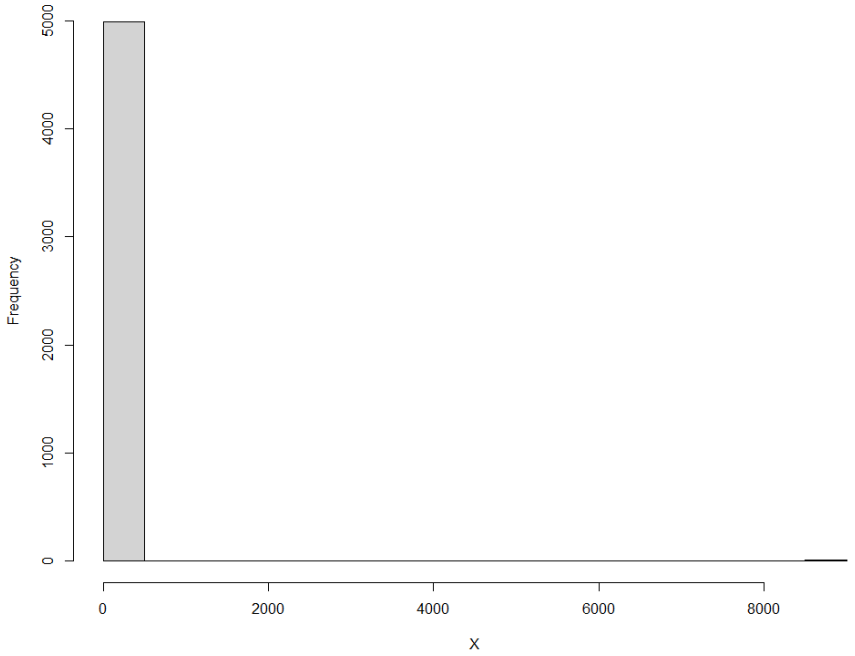


Image 15: Histogram of RetainedJob

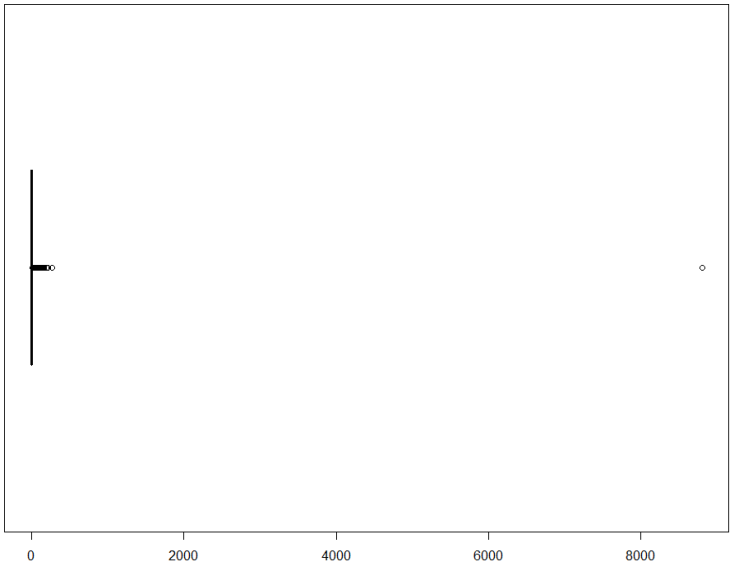


Image 16: Boxplot of RetainedJob

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	1	14.8	4	8800

"sd: 304.773662762816"
"vc: 20.5903108245495"

UrbanRural

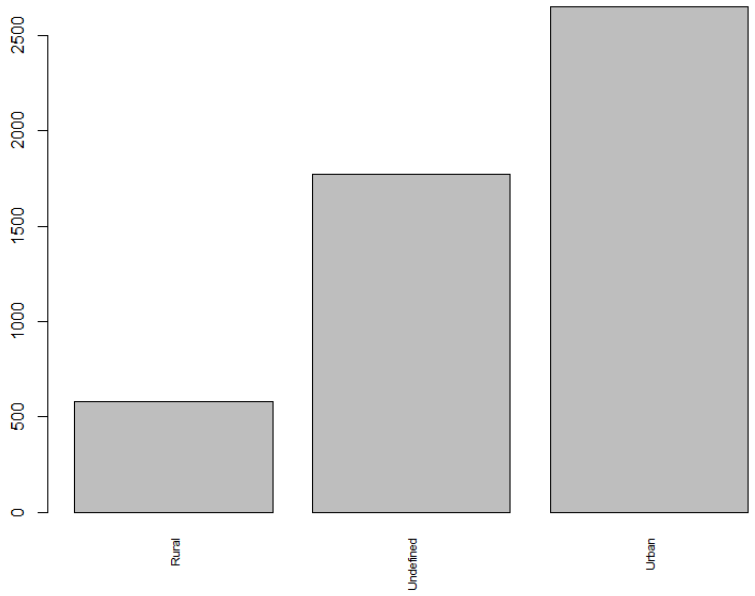


Image 17: Barplot of UrbanRural

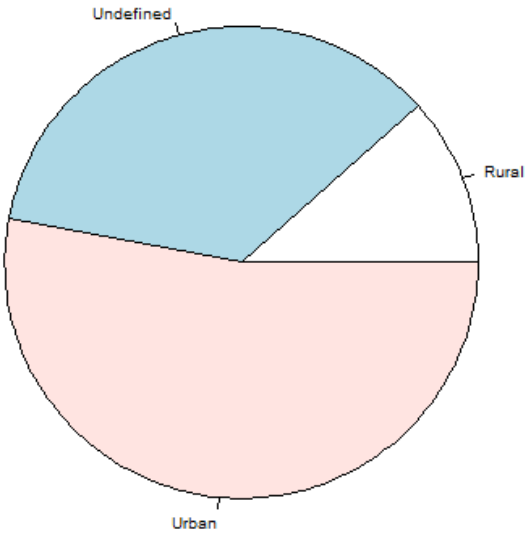


Image 18: Pie of UrbanRural

Statistics of variable “UrbanRural”			
Number of modalities	3		
Frequency table of the modalities	Undefined	Urban	Rural
	1773	2648	579
Proportions of modalities (out of 1)	0.3546	0.5296	0.1158

RevLineCr

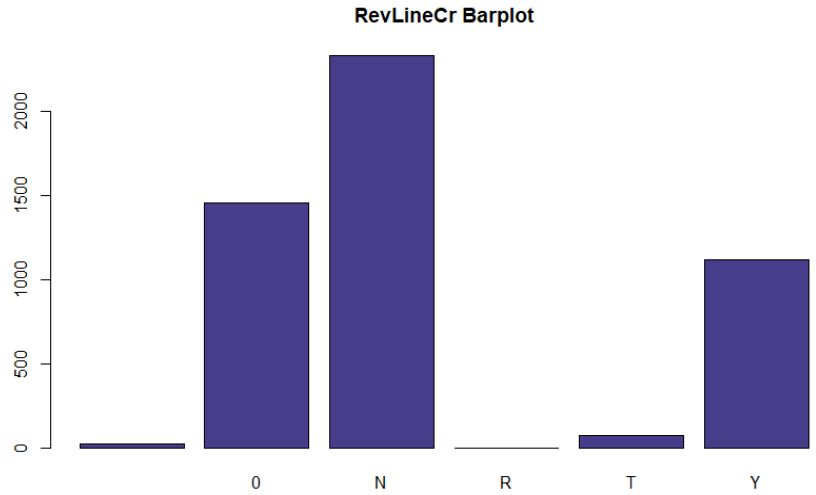


Image 19: RevLineCr Barplot

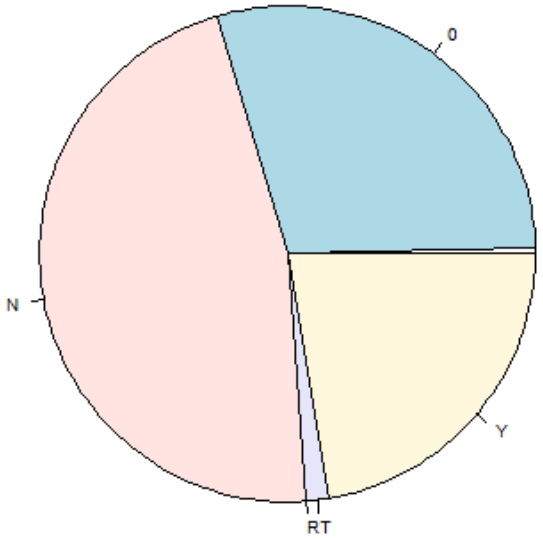


Image 20: Pie of RevLineCr

Statistics of variable “RevLineCr”						
Number of modalities	6					
Frequency table of the modalities	NA	N	R	T	Y	NA
	1458	2331	1	72	1117	21
Proportions of modalities (out of 1)	0.2916	0.4662	0.0002	0.0144	0.2234	0.0042

DisbursementGross

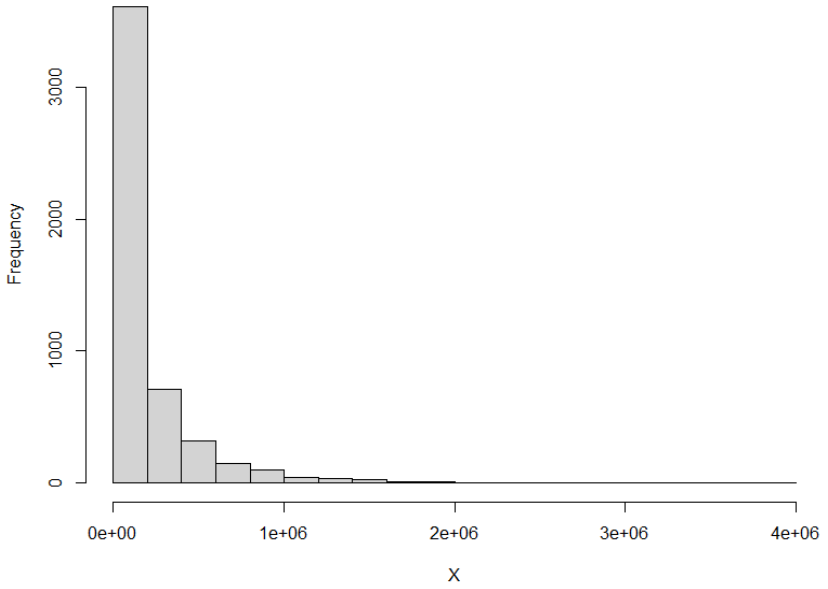


Image 21: Histogram of DisbursementGross

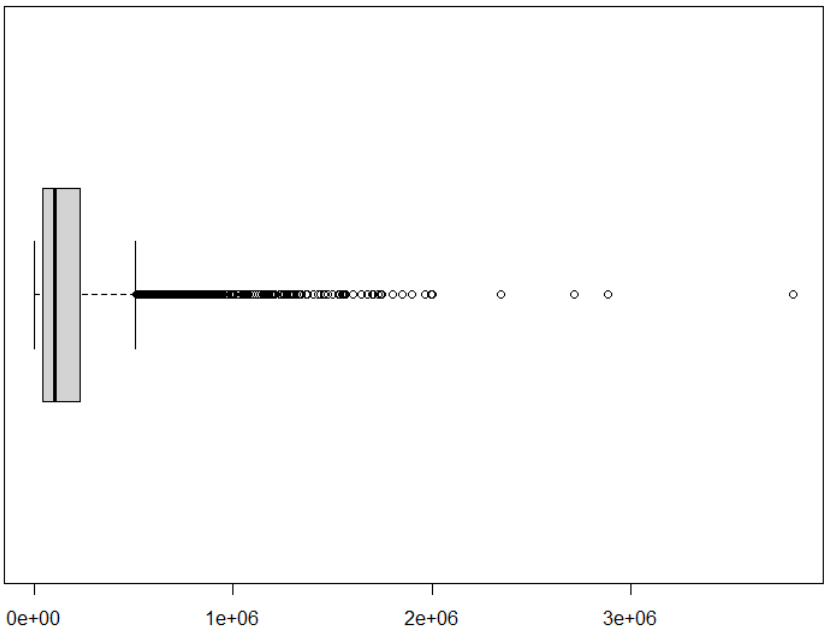


Image 22: Boxplot of DisbursementGross

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	40000	99495	196700	227700	3820000

"sd: 280292.345936634"
"vc: 1.42497067935564"

MIS_Status

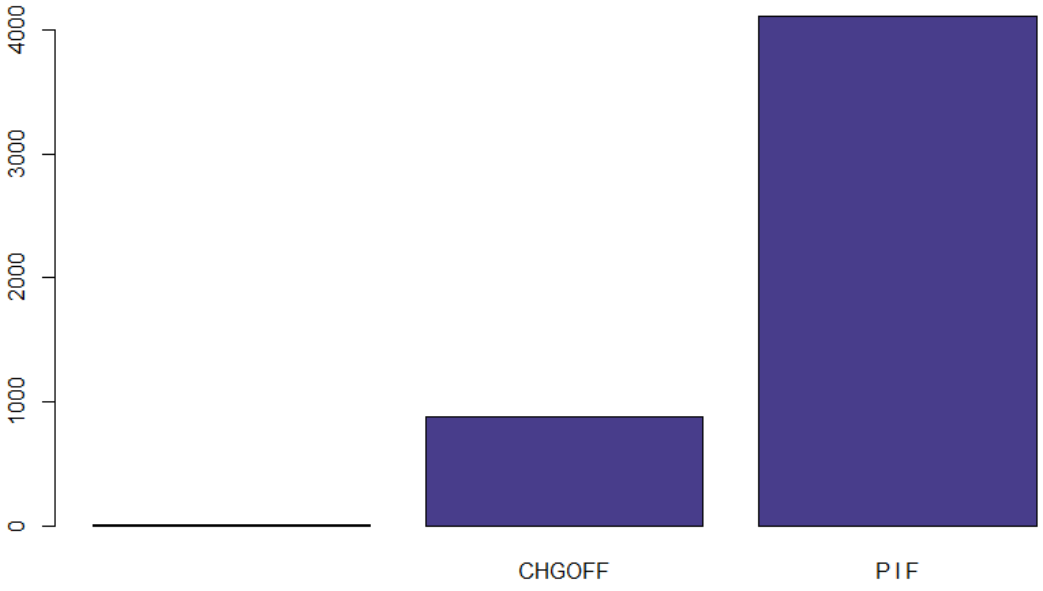


Image 23: MIS_Status Barplot

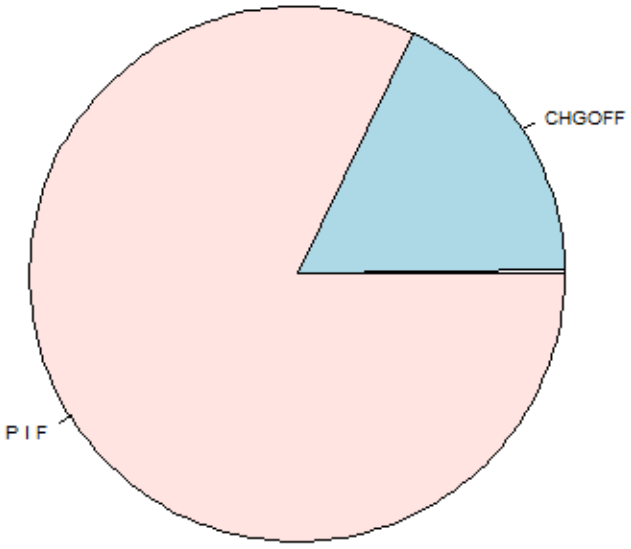


Image 24: Pie of MIS_Status

Statistics of variable “MIS_Status”			
Number of modalities	3		
Frequency table of the modalities	PIF	CHGOFF	NA
	4109	878	13
Proportions of modalities (out of 1)	0.8218	0.1756	0.0026

GrAppv

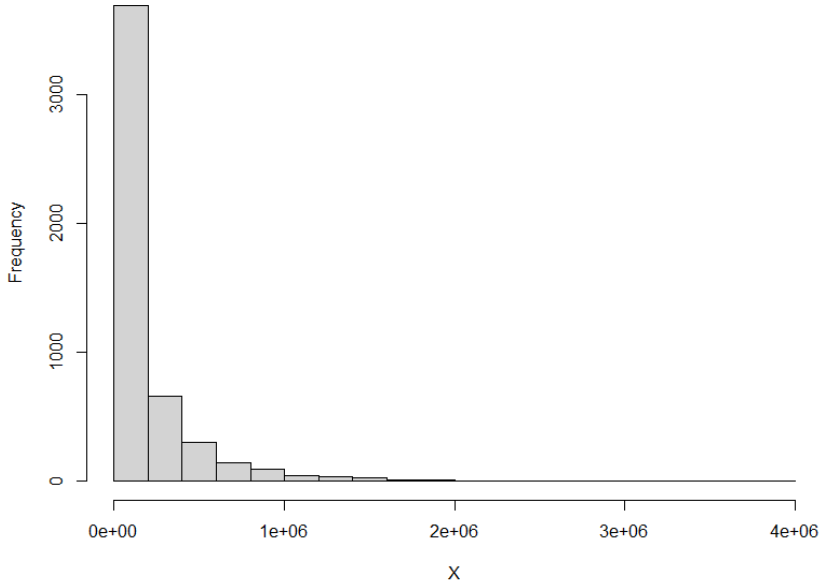


Image 25: Histogram of GrAppv

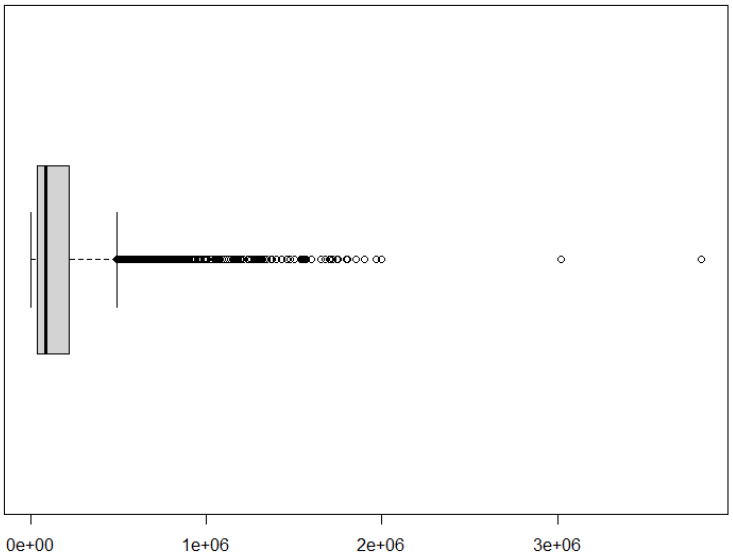


Image 26: Boxplot of GrAppv

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1000	35000	85000	187670	216764	3820000

"sd: 276220.74145269"
"vc: 1.47184123897391"

SBA_Appv

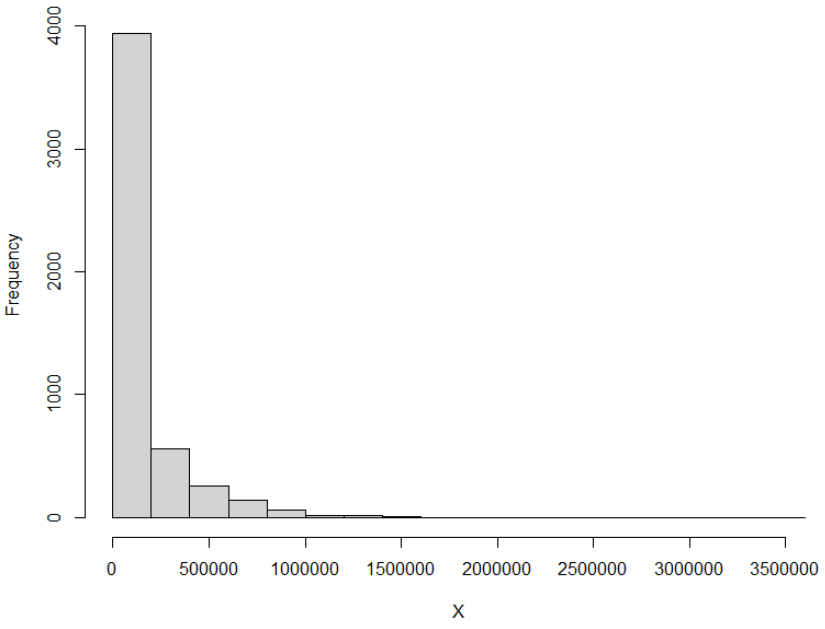


Image 27: Histogram of SBA_Appv

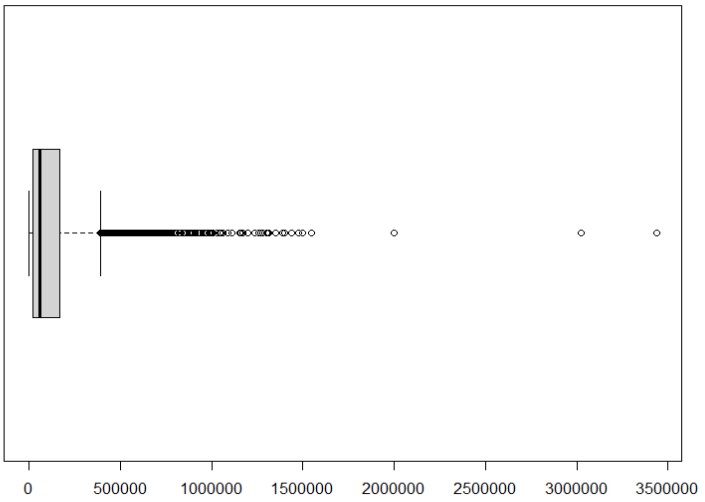


Image 28: Boxplot of SBA_Appv

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
500	21250	60000	144820	168750	3438000

"sd: 221224.993111245"
"vc: 1.52758459426255"

WhichCompany

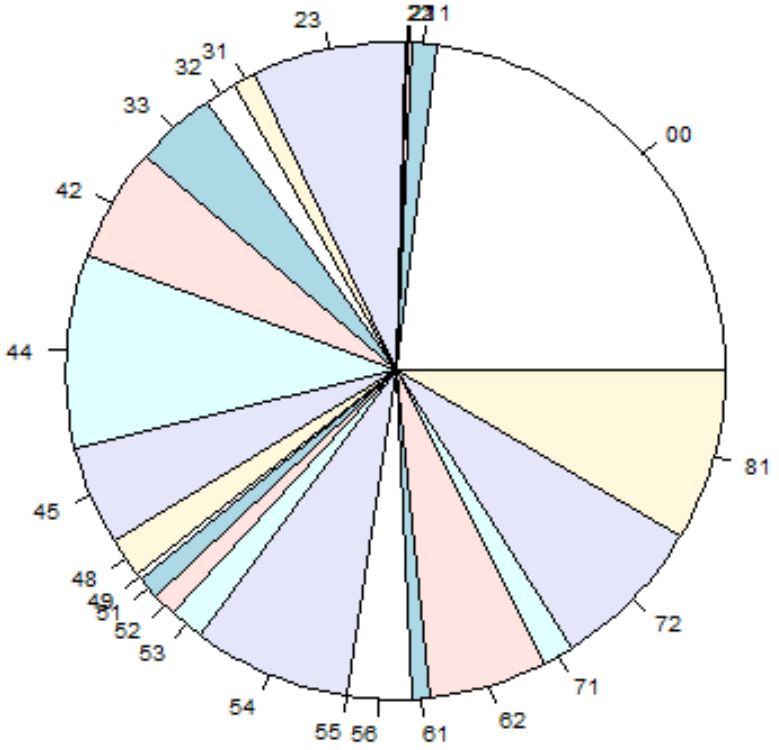


Image 29: Pie of WhichCompany

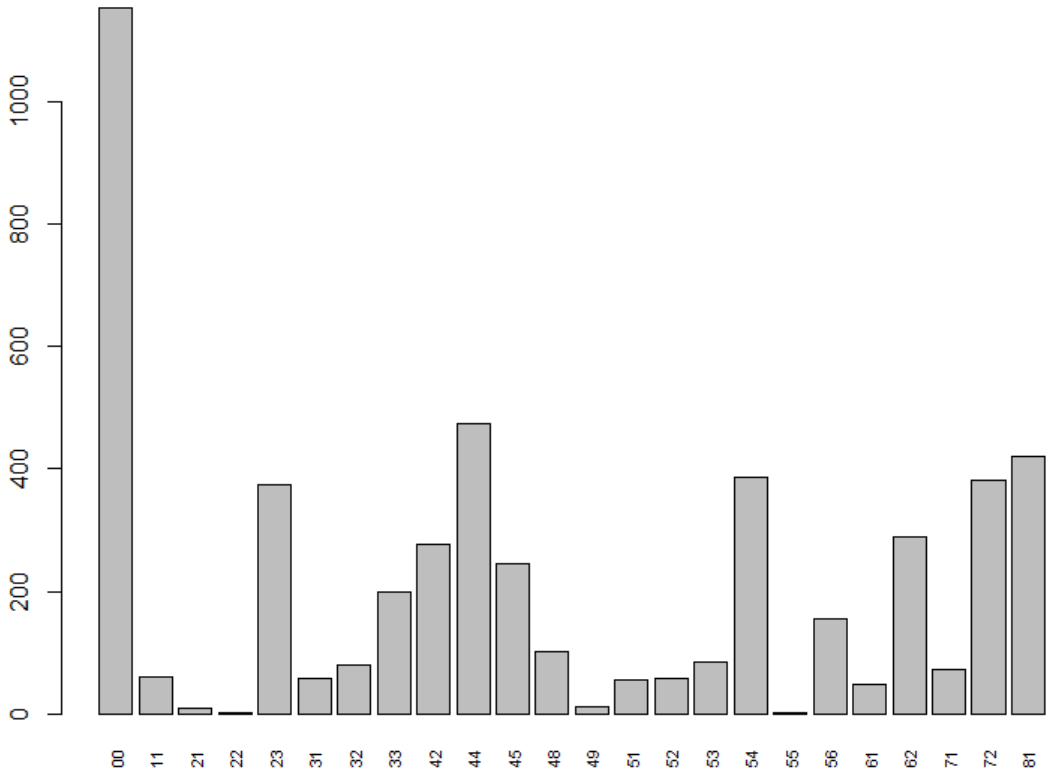


Image 30: Barplot of WhichCompany

Statistics of variable “State”	
Number of modalities	24
Frequency table of the modalities	00 11 21 22 23 31 32 33 42 44 45 48 49 51 52 53 1152 60 10 3 375 58 80 199 278 473 246 101 12 56 57 85 54 55 56 61 62 71 72 81 386 1 156 48 289 73 382 420
Proportions of modalities (out of 1)	00 44 81 54 72 23 62 42 45 33 56 48 0.2304 0.0946 0.0840 0.0772 0.0764 0.0750 0.0578 0.0556 0.0492 0.0398 0.0312 0.0202 53 32 71 11 31 52 51 61 49 21 22 55 0.0170 0.0160 0.0146 0.0120 0.0116 0.0114 0.0112 0.0096 0.0024 0.0020 0.0006 0.0002

Initial Bivariate data description

Bivariate description of the quantitative variables

Making use of the library corrplot, we will graph a correlation plot between all the numeric variables in our dataset to analyze how they are correlated with each other.

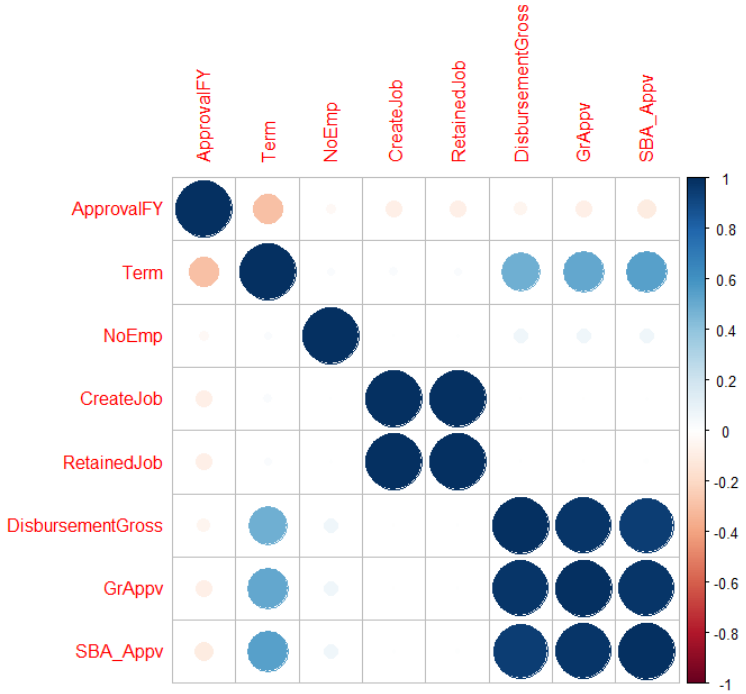


Image 31: Correlation between all the numeric variables

- Correlation between **Term** & **ApprovalFY**:

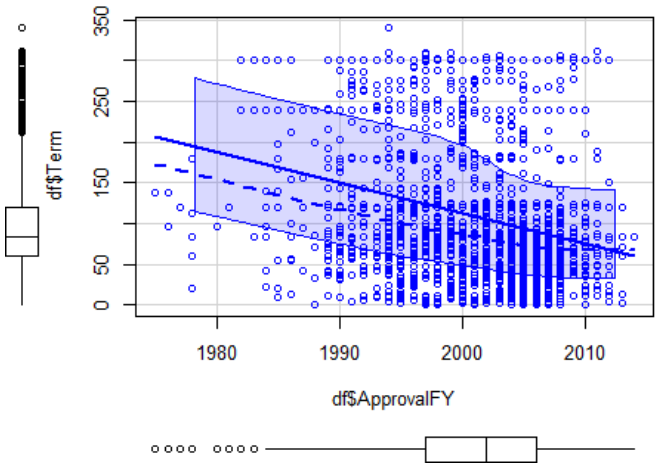


Image 32: Correlation between Term and ApprovalFY

  - They have a light negative correlation (-0.3), as we can see in the graph, the more recent the approval year of commitment is the smaller the loan term in months is.

- Correlation between **Term** & **Disbursement Gross**:

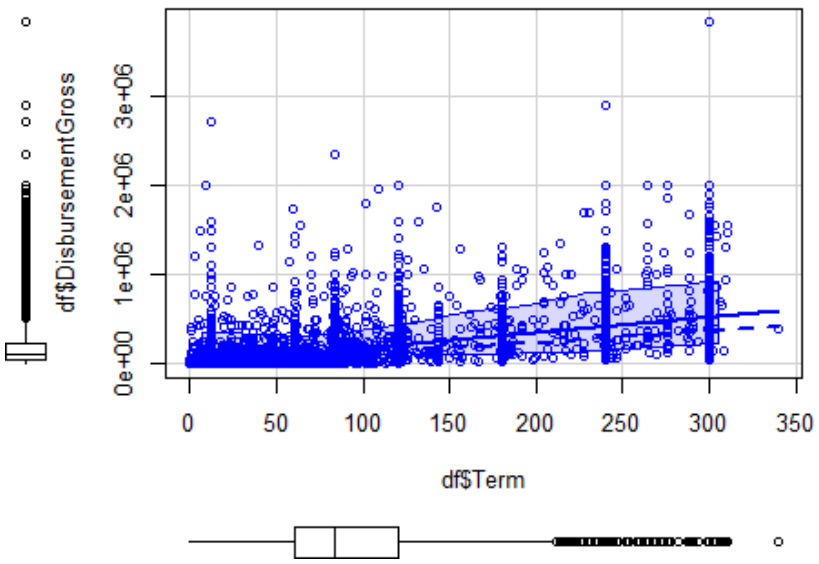


Image 33: Correlation between Term and DisbursementGross

  - Direct correlation; as the term increases the gross disbursement increases. This is due to the fact that banks give a greater amount of time to pay larger loans to businesses.

- Correlation between **Term** & **GrAppv**:

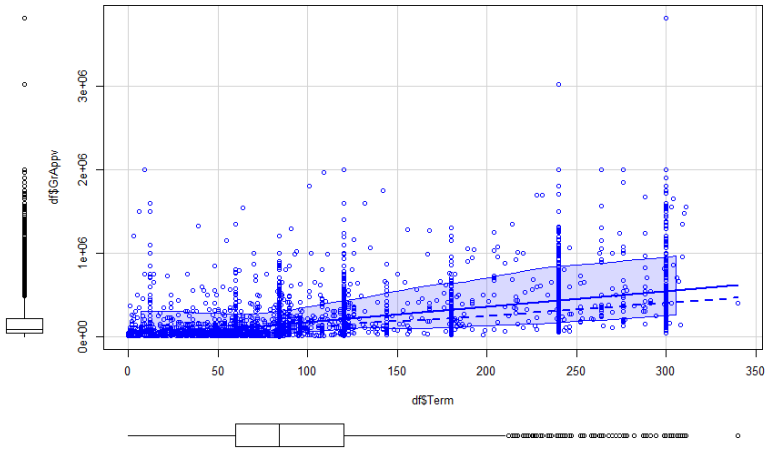


Image 34: Correlation between Term and GrAppv

  - They have a light direct correlation (0.4) since the bigger the loan term in months is the bigger the gross amount of loan approved by the bank is. This happens because banks understand that bigger loans are related to bigger projects that take a greater amount of time than projects that require smaller loans.

- Correlation between **Term** & **SBA_Appv**:

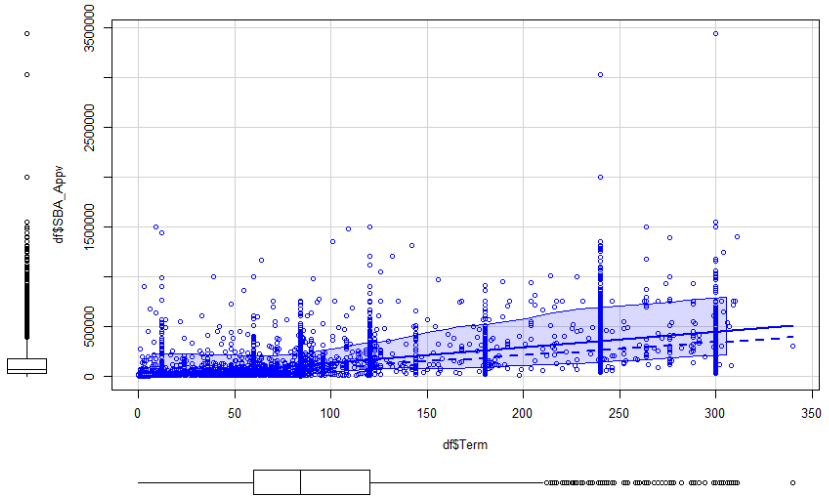


Image 35: Correlation between Term and SBA_Appv

  - They have a light direct correlation (0.4) since the bigger the loan term in months is the bigger SBA's guaranteed amount of approved loan is. This is explained by the same reasoning behind the relationship between gross amount of loan and loan term in months since, as it is explained later in the document, the gross amount of loan and the SBA's guaranteed amount of approved loan are highly correlated variables.

- Correlation between **Retained Job** & **Create Job:**

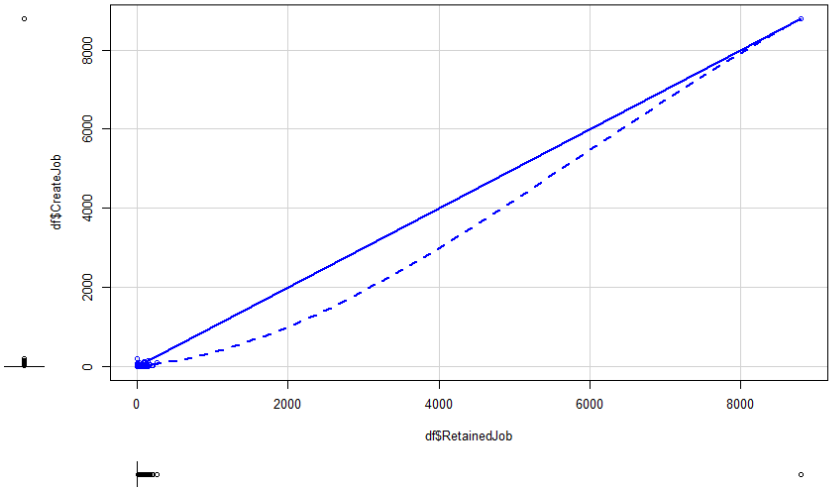


Image 36: Correlation between RetainedJob and CreateJob

  - Direct correlation of almost 1. The graph demonstrates that the number of jobs created is highly correlated to the number of jobs retained since the bigger the number of created jobs is the bigger the number of retained jobs is. This happens because when a loan is given to a business, this business has just started so it is logical that it creates new job positions that will be retained for a decent amount of time meaning that a lot of jobs are being created and retained simultaneously.

- Correlation between **Disbursement Gross** & **GrAppv**:

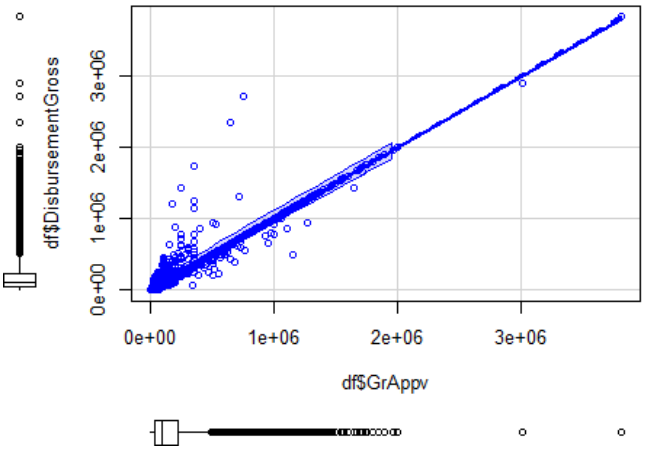


Image 37: Correlation between DisbursementGross and GrAppv

  - Direct correlation of almost 1. As the total amount of disbursement given to the business increases, the gross amount of loan granted by the bank increases. This is due to the fact that a very high percentage if not all of the total disbursement gross is the gross amount approved by the bank, (the bank gives all of the loan).
- Correlation between **Disbursement Gross** & **SBA_Appv**:

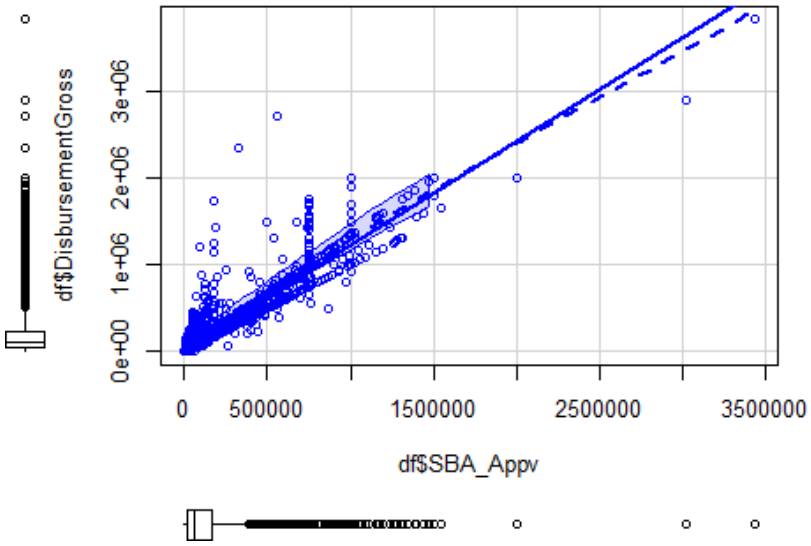


Image 38: Correlation between DisbursementGross and SBA_Appv

  - Direct correlation of almost 1. As the total amount of disbursement given to the business increases, the amount of loan guaranteed by the SBA to the bank increases. This is due to the fact the SBA tends to guarantee a percentage of the loan given by the bank, therefore if the quantity of the loan increases, the guaranteed loan also increases.

- Correlation between **GrAppv** & **SBA_Appv**

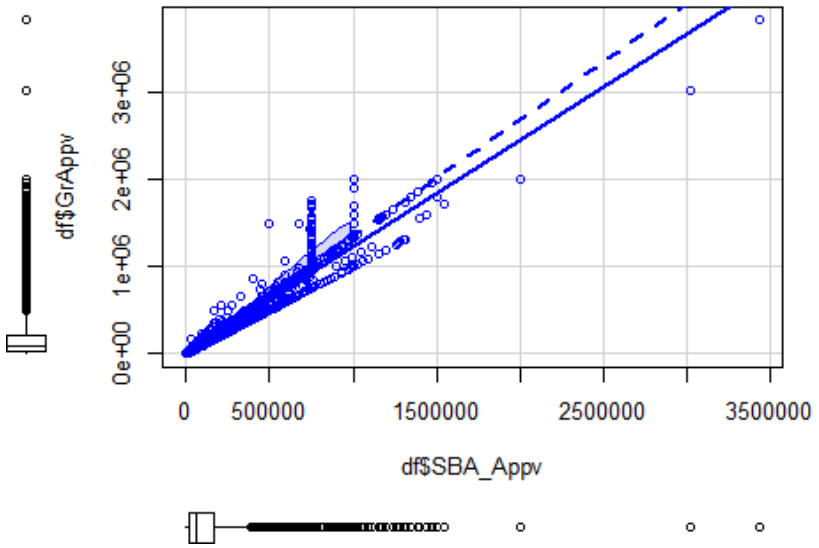


Image 39: Correlation between GrAppv and SBA_Appv

  - Direct correlation of almost 1. As the total amount of disbursement given to the business increases, the amount of loan guaranteed by the SBA to the bank increases. This is due to the fact the SBA tends to guarantee a percentage of the loan given by the bank, therefore if the quantity of the loan increases, the guaranteed loan also increases.

Bivariate description of categorical variables with quantitative variables

In this section we will analyze the correlation between our categorical and numerical variables. To keep this analysis short we will only be analyzing the correlation of all our numerical variables with our response variable that happens to be a categorical variable .

- Correlation between **ApprovalFY** & **MIS_Status**

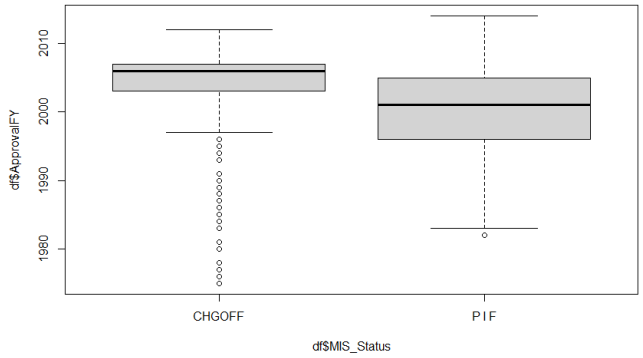


Image 40: Correlation between ApprovalFY and MIS_Status

  - For CHGOFF → v.test = 12.798456
    For PIF → v.test = -12.933671

- Correlation between **Term** & **MIS_Status**

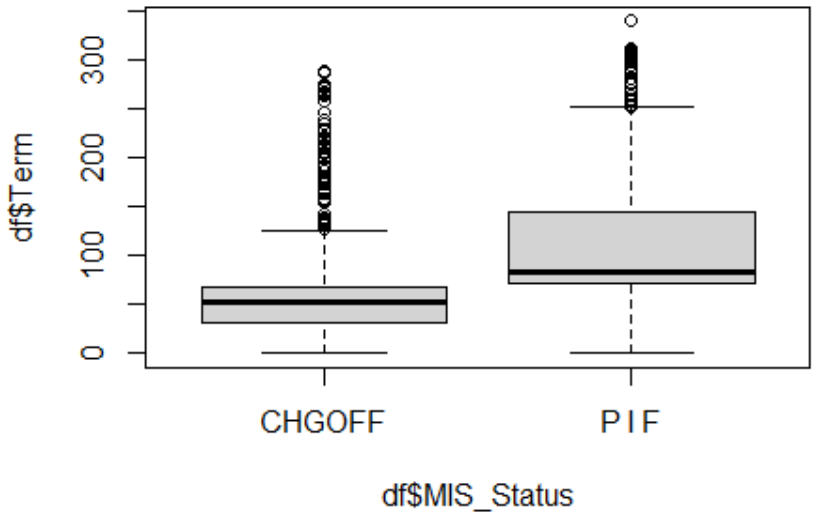


Image 41: Correlation between Term and MIS_Status

For CHGOFF → v.test = -21.5
For PIF → v.test = 21.57

- Correlation between **NoEmp** & **MIS_Status**

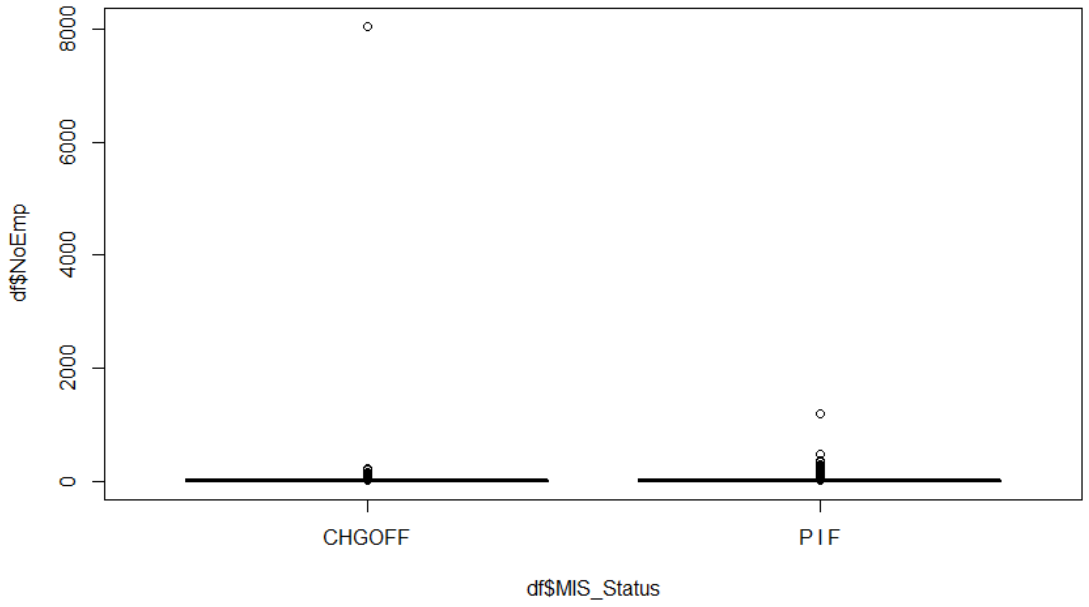


Image 43: Correlation between NoEmp and MIS_Status

  - This graph can not be interpreted as it has too many outliers to be relevant to our study.
  - For CHGOFF → v.test =
    For PIF → v.test =

- Correlation between **CreateJob** & **MIS_Status**

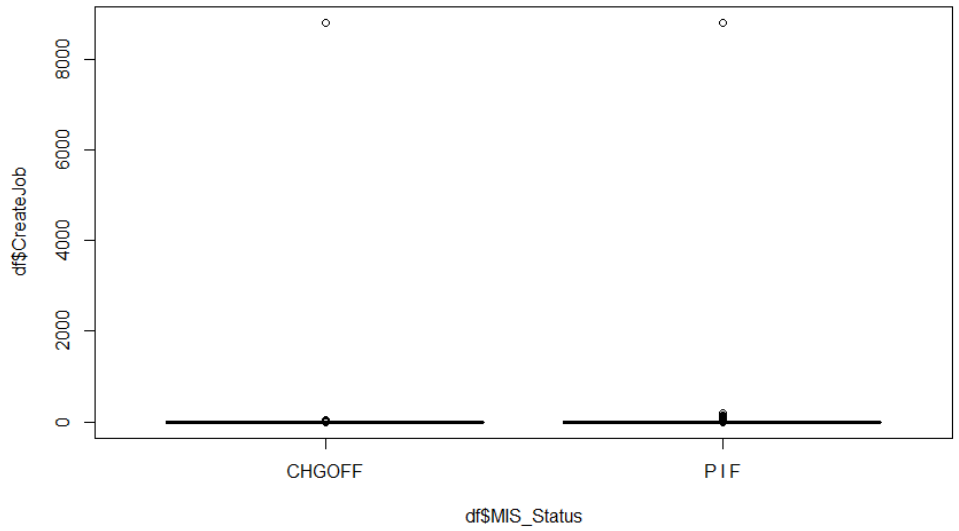


Image 44: Correlation between CreateJob and MIS_Status

  - For CHGOFF → v.test = 3.128052
    For PIF → v.test = -3.091218

- Correlation between **RetainedJob** & **MIS_Status**

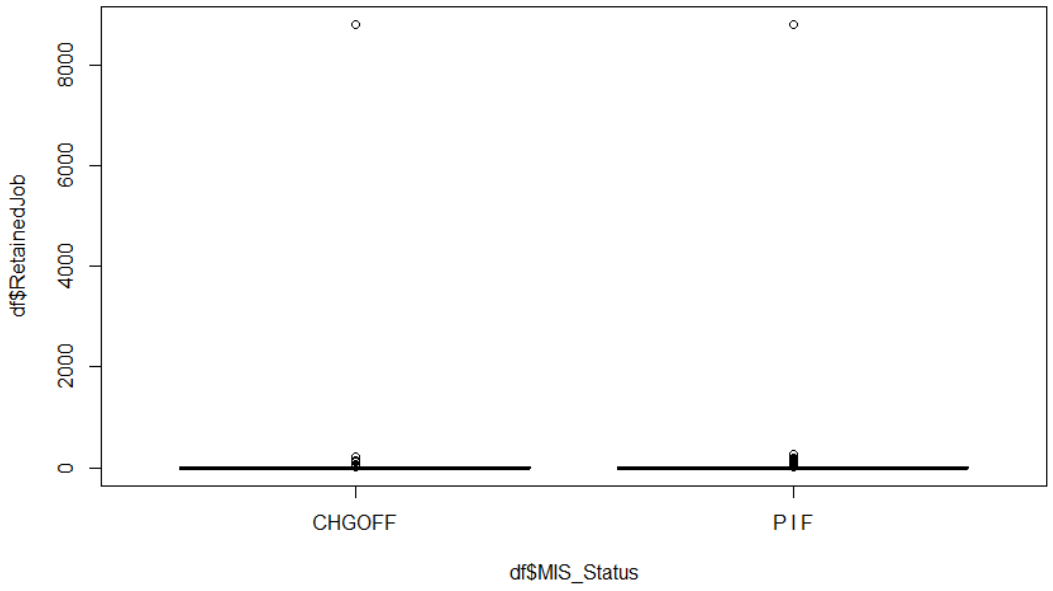


Image 45: Correlation between RetainedJob and MIS_Status

  - For CHGOFF → v.test = 3.231738
    For PIF → v.test = -3.196483

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

- Correlation between **DisbursementGross** & **MIS_Status**

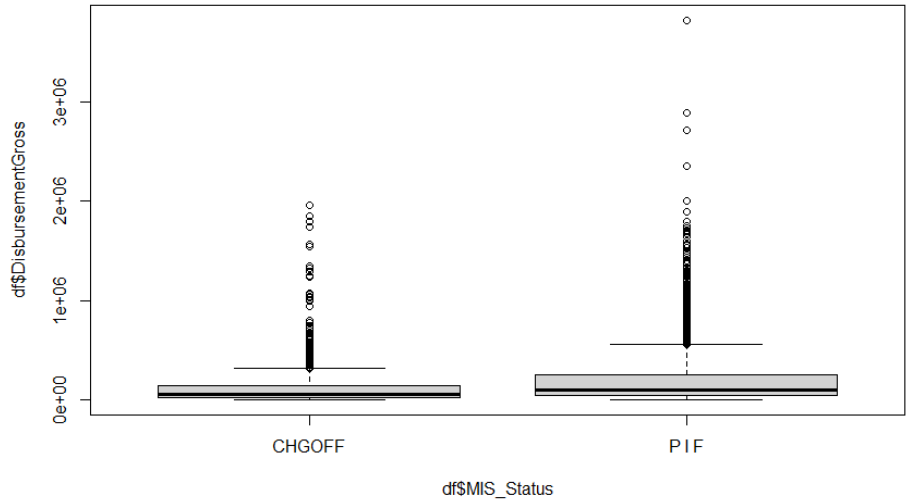


Image 46: Correlation between DisbursementGross and MIS_Status

  - For CHGOFF → v.test = -6.625149
    For PIF → v.test = 6.921161

- Correlation between **GrAppv** & **MIS_Status**

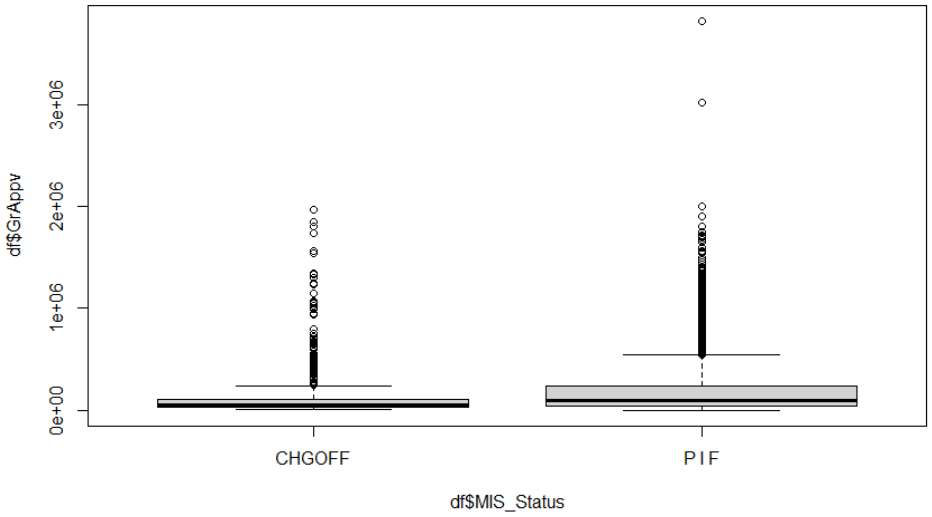


Image 47: Correlation between GrAppv and MIS_Status

  - For CHGOFF → v.test = -6.625149
    For PIF → v.test = -7.096743

- Correlation between **SBA_Appv** & **MIS_Status**

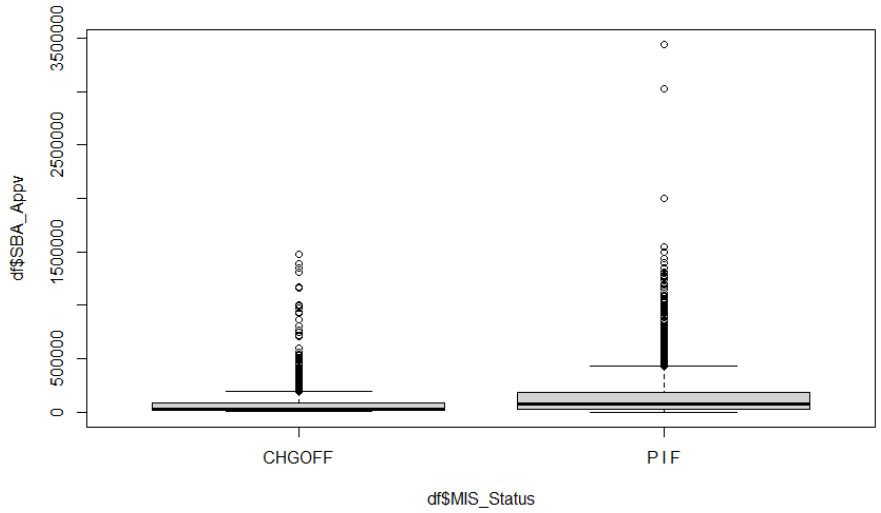


Image 48: Correlation between SBA_Appb and MIS_Status

  - For CHGOFF → v.test = -7.866118
    For PIF → v.test = 8.124746

BIvariate description of categorical variables

In this section we will analyze the correlation between our categorical variables. To keep this analysis short we will only be analyzing the correlation of all our categorical variables with our response variable that is also categorical variable.

- Correlation between **State** & **MIS_Status**

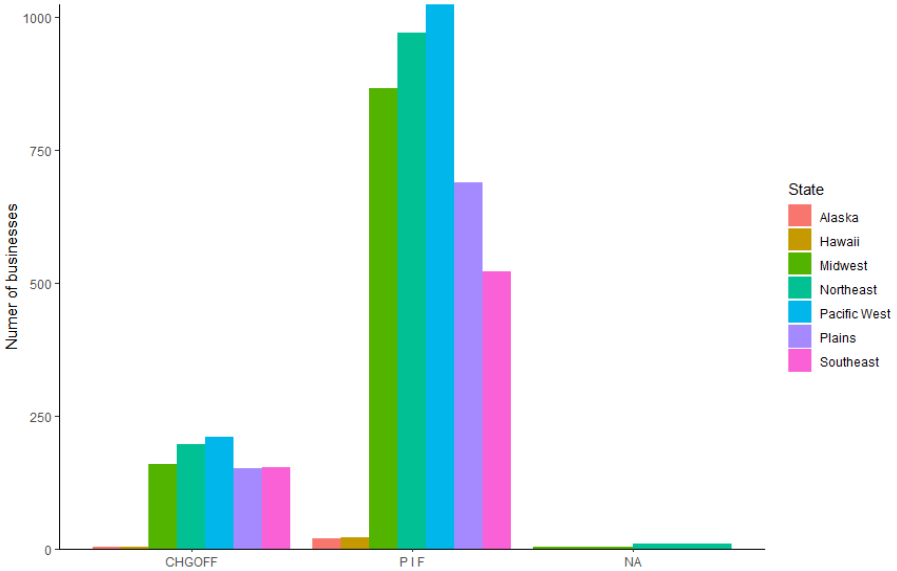


Image 49: Correlation between State and MIS_Status

Row Percentages	Column Percentages

	CHGOFF	P I F
Alaska	0.1363636	0.8636364
Hawaii	0.1666667	0.8333333
Midwest	0.1551220	0.8448780
Northeast	0.1688089	0.8311911
Pacific West	0.1708502	0.8291498
Plains	0.1797619	0.8202381
Southeast	0.2270030	0.7729970

	CHGOFF	P I F
Alaska	0.003416856	0.004623996
Hawaii	0.004555809	0.004867364
Midwest	0.181093394	0.210756875
Northeast	0.224373576	0.236067170
Pacific West	0.240318907	0.249209053
Plains	0.171981777	0.167680701
Southeast	0.174259681	0.126794841

- Correlation between **BankState** & **MIS_Status**

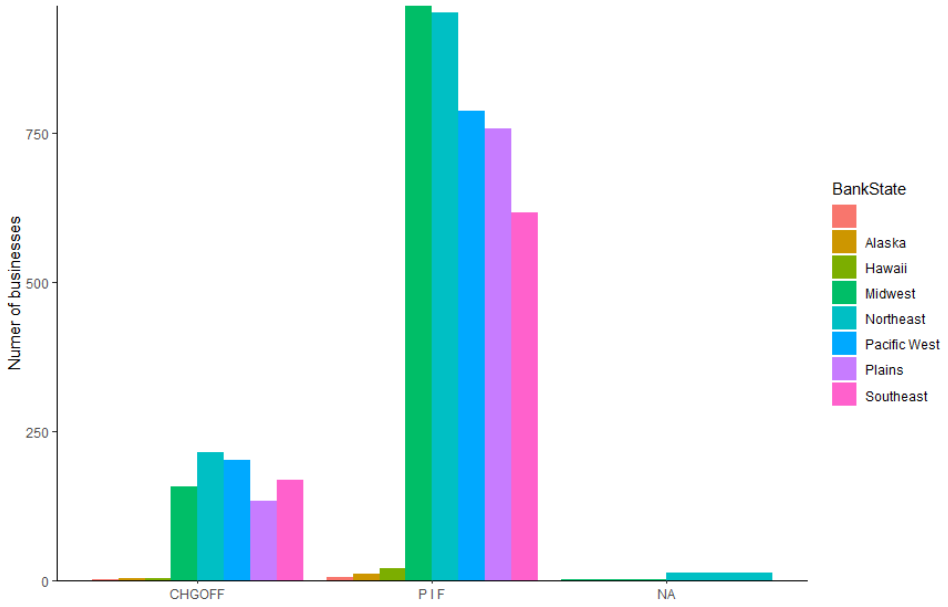


Image 50: Correlation between BankState and MIS_Status

Row Percentages	Column Percentages

	CHGOFF	P I F
	0.2000000	0.8000000
Alaska	0.2307692	0.7692308
Hawaii	0.1304348	0.8695652
Midwest	0.1401786	0.8598214
Northeast	0.1835334	0.8164666
Pacific West	0.2036474	0.7963526
Plains	0.1484814	0.8515186
Southeast	0.2130102	0.7869898

	CHGOFF	P I F
	0.0011389522	0.0009734729
Alaska	0.0034168565	0.0024336822
Hawaii	0.0034168565	0.0048673643
Midwest	0.1788154897	0.2343635921
Northeast	0.2437357631	0.2316865417
Pacific West	0.2289293850	0.1912874179
Plains	0.1503416856	0.1842297396
Southeast	0.1902050114	0.1501581893

- Correlation between **WhichCompany** & **MIS_Status**

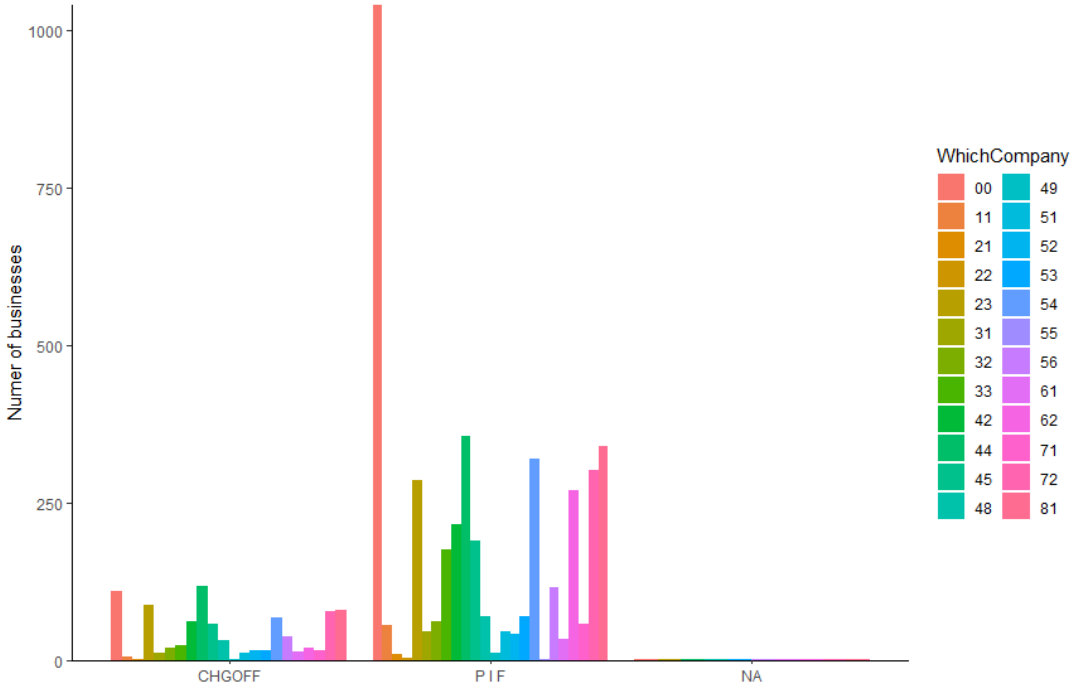


Image 51: Correlation between WhichCompany and MIS_Status

Row Percentages

	CHGOFF	P I F
00	0.09478261	0.90521739
11	0.08333333	0.91666667
21	0.10000000	0.90000000
22	0.00000000	1.00000000
23	0.23529412	0.76470588
31	0.20689655	0.79310345
32	0.23750000	0.76250000
33	0.12060302	0.87939698
42	0.22302158	0.77697842
44	0.24788136	0.75211864
45	0.23170732	0.76829268
48	0.31000000	0.69000000
49	0.08333333	0.91666667
51	0.19642857	0.80357143
52	0.26315789	0.73684211
53	0.17857143	0.82142857
54	0.17357513	0.82642487
55	0.00000000	1.00000000
56	0.24675325	0.75324675
61	0.27659574	0.72340426
62	0.06597222	0.93402778
71	0.21917808	0.78082192
72	0.20526316	0.79473684
81	0.19093079	0.80906921

Column Percentages

	CHGOFF	P I F
00	0.1241457859	0.2533463130
11	0.0056947608	0.0133852519
21	0.0011389522	0.0021903139
22	0.0000000000	0.0007301046
23	0.1002277904	0.0696033098
31	0.0136674260	0.0111949379
32	0.0216400911	0.0148454612
33	0.0273348519	0.0425894378
42	0.0706150342	0.0525675347
44	0.1332574032	0.0863957167
45	0.0649202733	0.0459965928
48	0.0353075171	0.0167924069
49	0.0011389522	0.0026770504
51	0.0125284738	0.0109515697
52	0.0170842825	0.0102214651
53	0.0170842825	0.0167924069
54	0.0763097950	0.0776344609
55	0.0000000000	0.0002433682
56	0.0432801822	0.0282307131
61	0.0148063781	0.0082745193
62	0.0216400911	0.0654660501
71	0.0182232346	0.0138719883
72	0.0888382688	0.0734972013
81	0.0911161731	0.0825018253

- Correlation between **NewExist** & **MIS_Status**

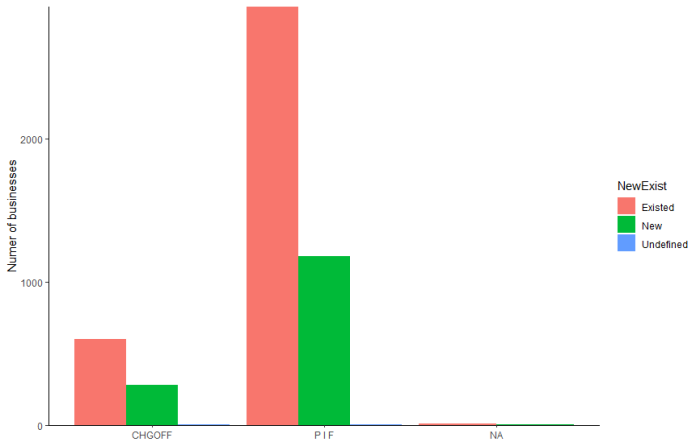


Image 52: Correlation between NewExist and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F Existed 0.1700255 0.8299745 New 0.1906722 0.8093278 Undefined 0.1666667 0.8333333	CHGOFF P I F Existed 0.682232346 0.711608664 New 0.316628702 0.287174495 Undefined 0.001138952 0.001216841

- Correlation between **UrbanRural** & **MIS_Status**

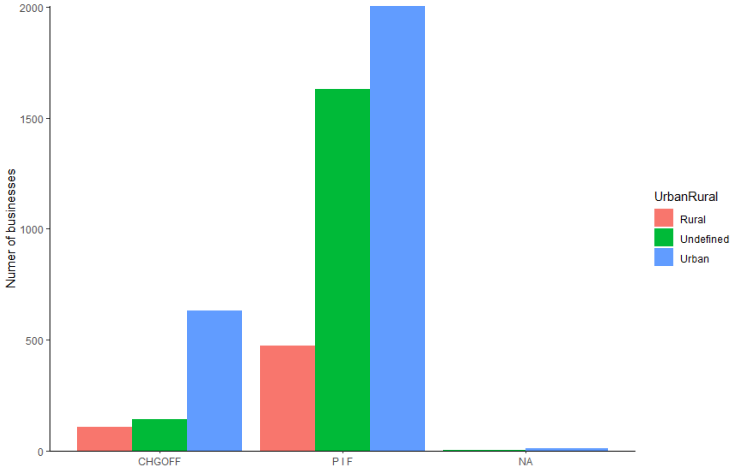


Image 53: Correlation between UrbanRural and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F Rural 0.18134715 0.81865285 Undefined 0.08018069 0.91981931 Urban 0.23928707 0.76071293	CHGOFF P I F Rural 0.1195900 0.1153565 Undefined 0.1617312 0.3964468 Urban 0.7186788 0.4881966

- Correlation between **RevLineCr** & **MIS_Status**

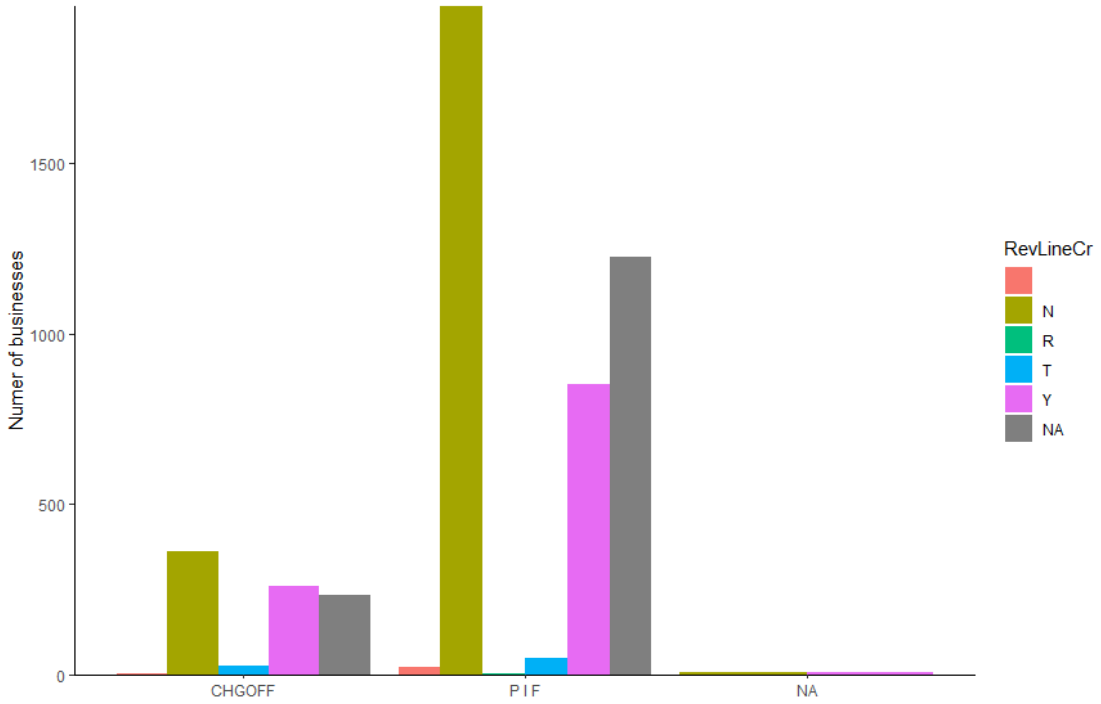


Image 54: Correlation between RevLineCr and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F 0.04761905 0.95238095 N 0.15497202 0.84502798 R 0.00000000 1.00000000 T 0.36111111 0.63888889 Y 0.23381295 0.76618705	CHGOFF P I F 0.0015455951 0.0069396253 N 0.5564142195 0.6811242193 R 0.0000000000 0.0003469813 T 0.0401854714 0.0159611381 Y 0.4018547141 0.2956280361

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Final Univariate data description

After preprocessing, the following variables were modified:

State

- We grouped the states of the business in their corresponding geographical region of the US, to better analyze grouped regions instead of states and to have a more clear picture of the data.

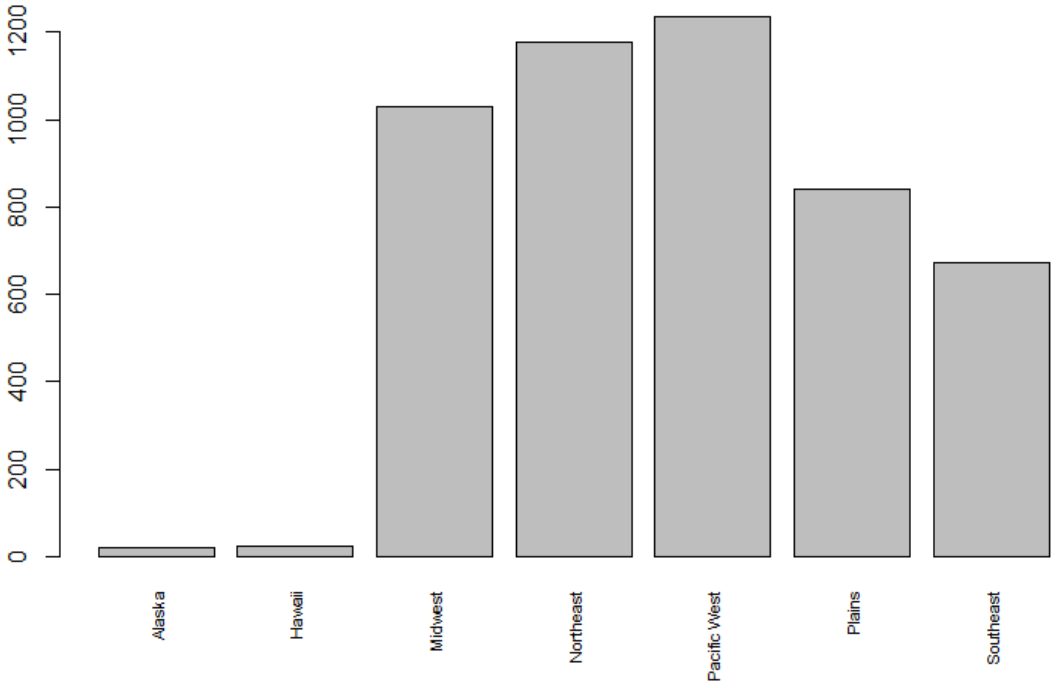


Image 55: Barplot of State

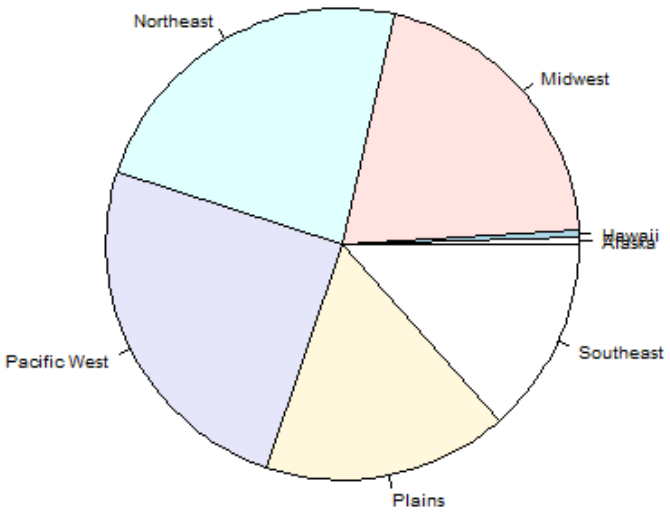


Image 56: Pie of State

Statistics of variable “State”	
Number of modalities	7
Frequency table of the modalities	Alaska 22 Hawaii 24 Midwest 1029 Northeast 1176 Pacific West 1235 Plains 840 Southeast 674
Proportions of modalities (out of 1)	Alaska 0.0044 Hawaii 0.0048 Midwest 0.2058 Northeast 0.2352 Pacific West 0.2470 Plains 0.1680 Southeast 0.1348

BankState

- We grouped the states of the bank in their corresponding geographical region of the US, to better analyze grouped regions instead of states and to have a more clear picture of the data.

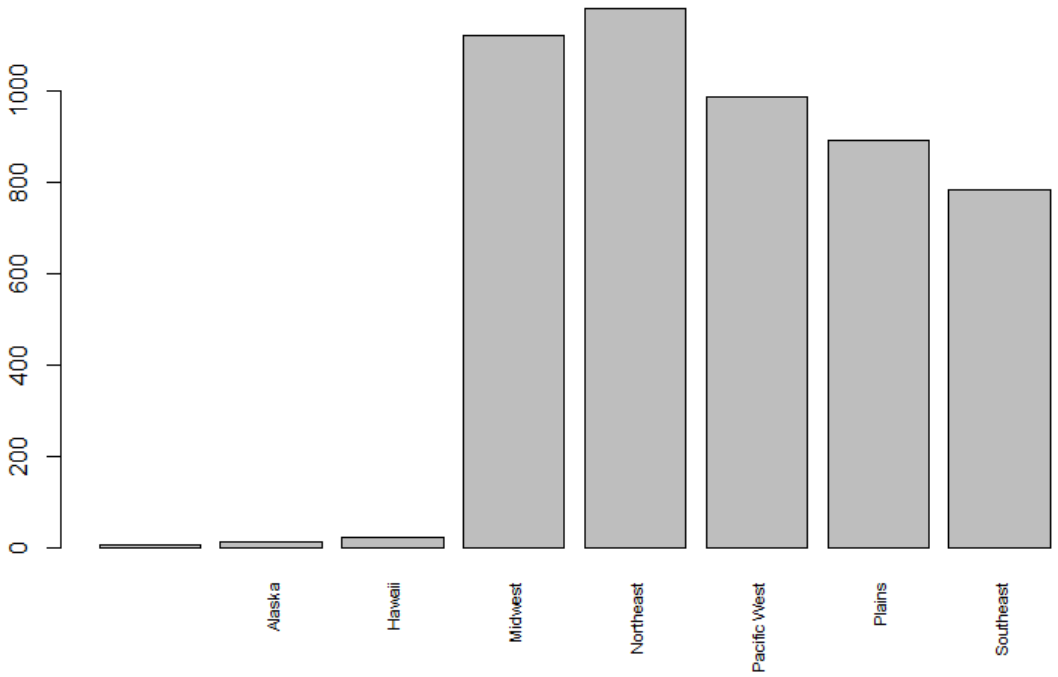


Image 57: Barplot of BankState

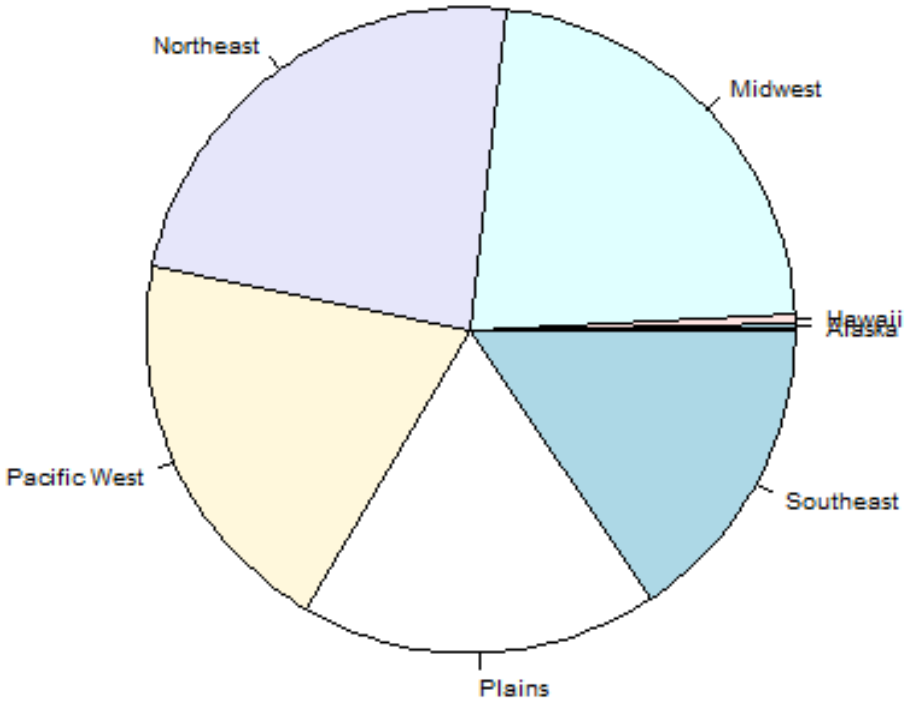


Image 58: Pie of BankState

Statistics of variable “State”	
Number of modalities	7
Frequency table of the modalities	5 Alaska 13 Hawaii 23 Midwest 1121 Northeast 1178 Pacific West 987 Plains 889 Southeast 784
Proportions of modalities (out of 1)	0.0010 Alaska 0.0026 Hawaii 0.0046 Midwest 0.2242 Northeast 0.2356 Pacific West 0.1974 Plains 0.1778 Southeast 0.1568

ApprovalFY

- This variable did not change during preprocessing

Term

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.

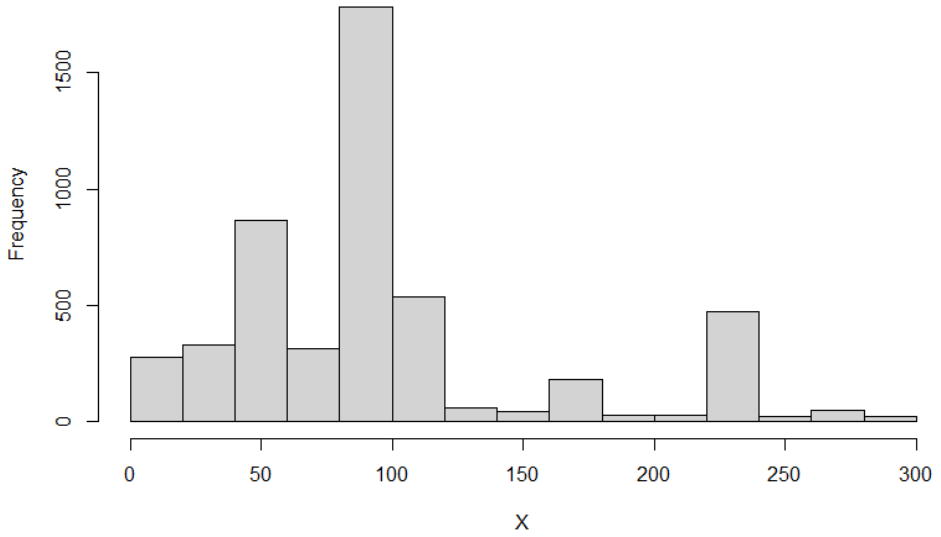


Image 59: Histogram of Term

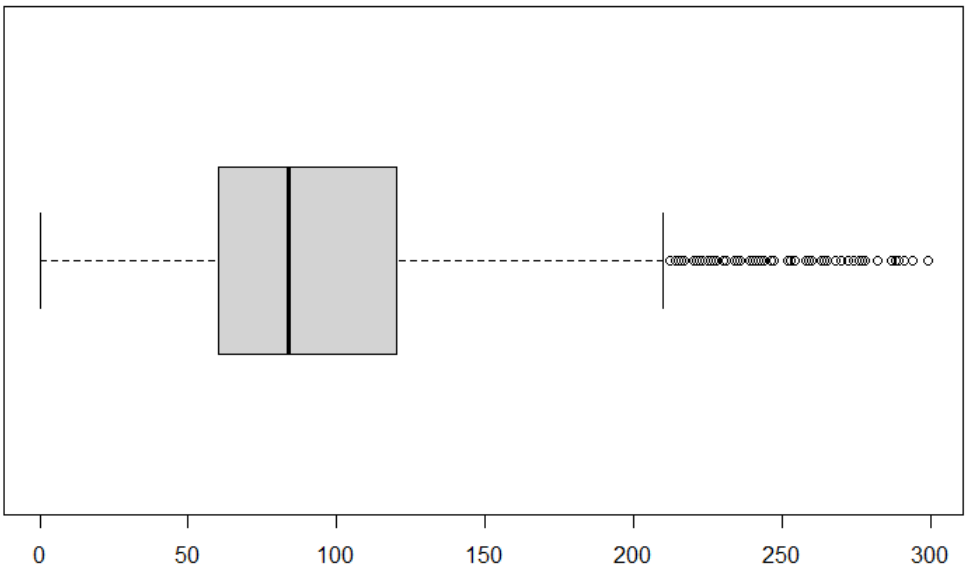


Image 60: Boxplot of Term

Extended Summary Statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	60	84	98.56	120	299

"sd: 63.29"
"vc: 0.642"

NoEmpl

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.
- Before preprocessing, we had some big business with a lot of employees and those were acting as extreme outliers. Those outliers were distorting our graphics and analysis, but now we can clearly see the employer distribution in our set of businesses.

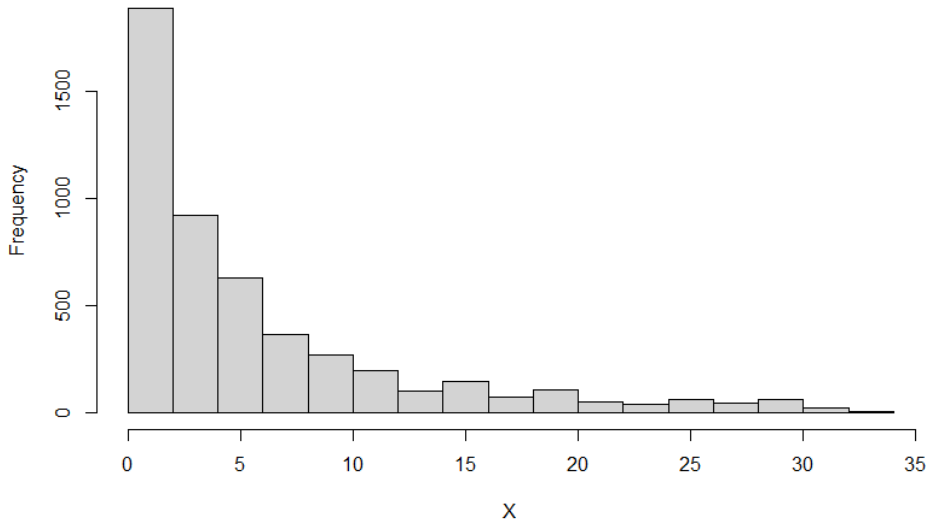


Image 61: Histogram of NoEmp

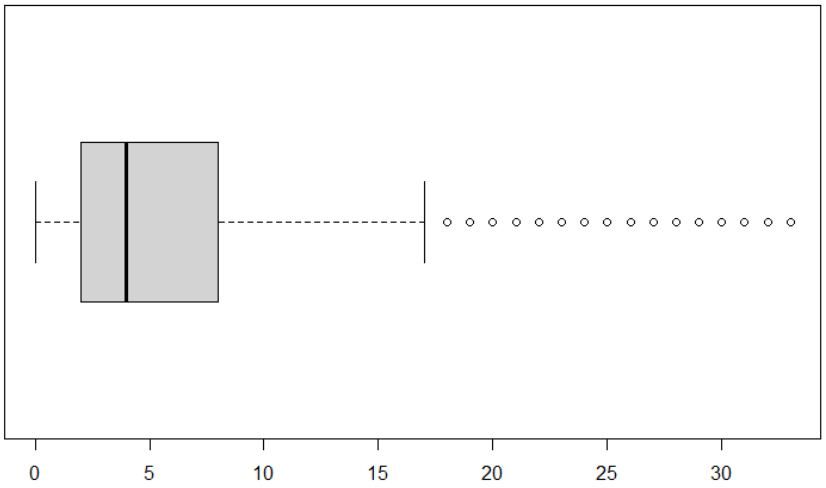


Image 62: Boxplot of NoEmp

Extended Summary Statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	2	4	6.33	8	33

"sd: 6.74"
"vc: 1.06"

NewExist

- This variable did not change during preprocessing.

CreateJob

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.
- Before preprocessing, we had some big business that hired a lot of employees with the loan and those were acting as extreme outliers. Those outliers were distorting our graphics and analysis, but now we can clearly see the employer distribution in our set of businesses.
- As we are dealing with small businesses, those will not hire a lot of employers with their loan.

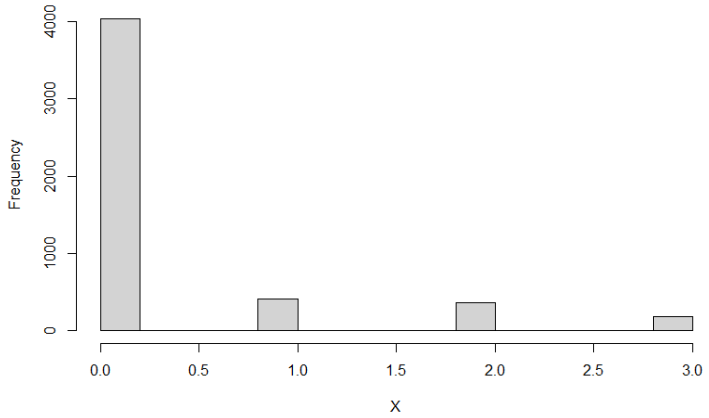


Image 63: Histogram of CreateJob

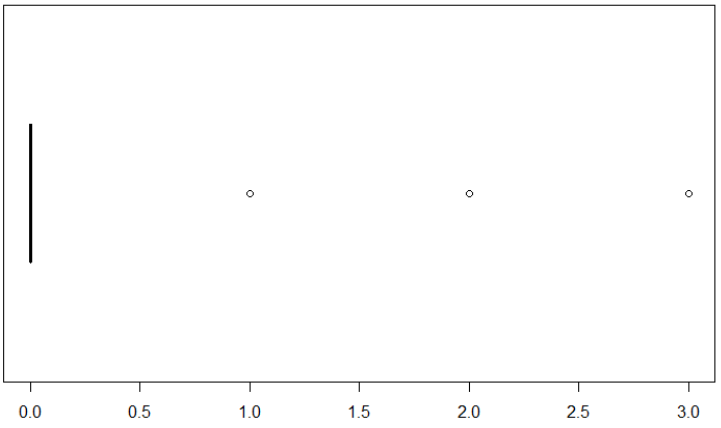


Image 64: Boxplot of CreateJob

Extended Summary Statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	0	0	0.3374	0	3

"sd: 0.765"
"vc: 2.2689"

RetainedJob

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm

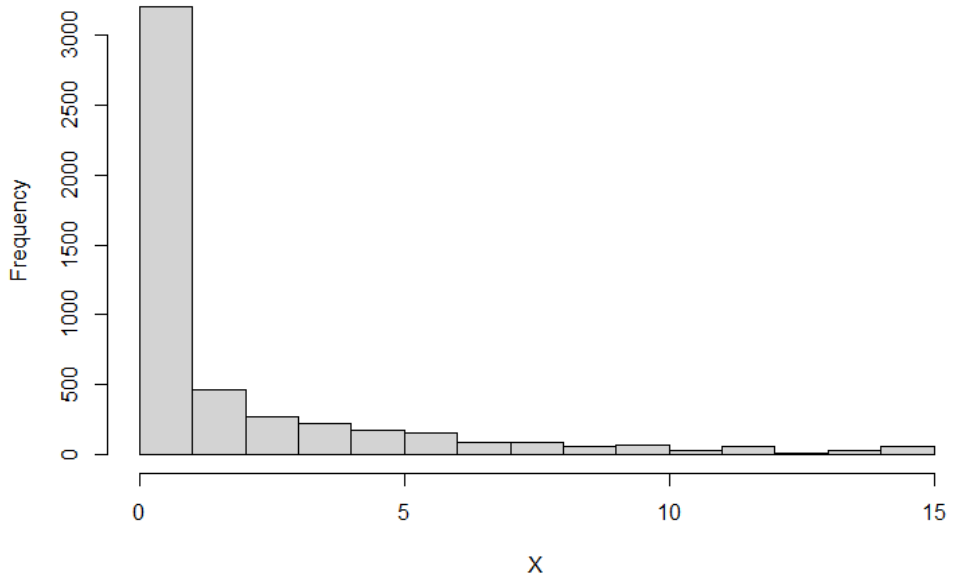


Image 65: Histogram of RetainedJob

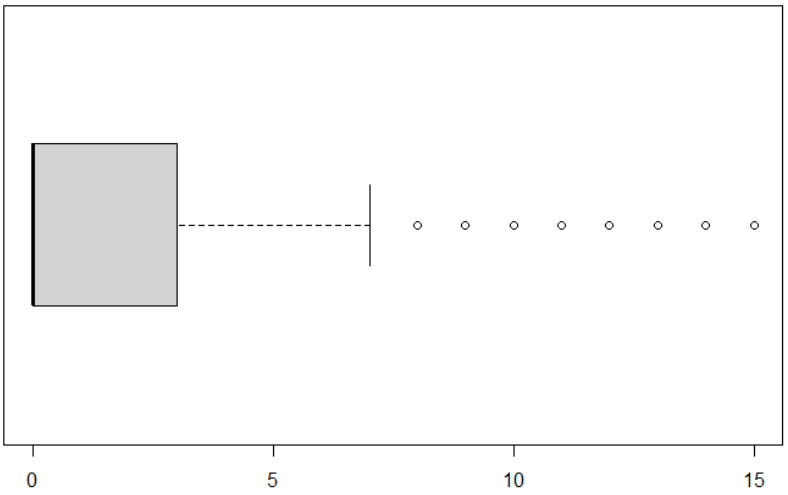


Image 66: Boxplot of RetainedJob

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	2.029	3	15

"sd: 3.30296290514904"
"vc: 1.62819821805631"

UrbanRural

- This variable experienced no changes after the preprocessing.

RevLineCr

- We reduced the number of modalities adding their share to the NA values.

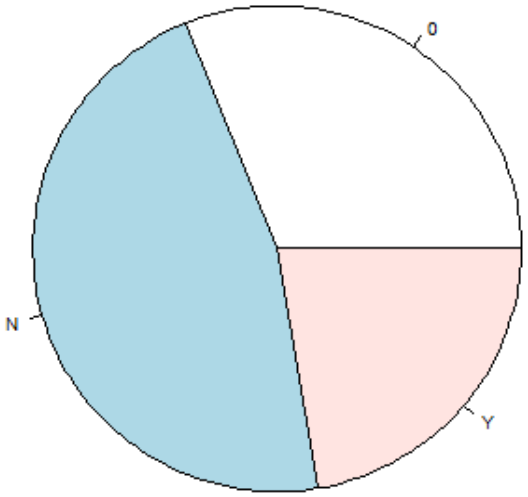


Image 67: Pie of RevLineCr

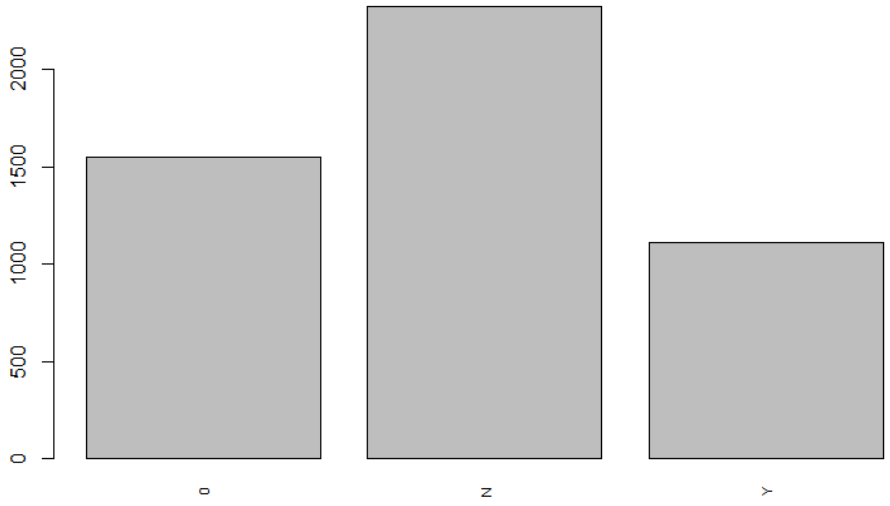


Image 68: Barplot of RevLineCr

Statistics of variable “RevLineCr”			
Number of modalities	3		
Frequency table of the modalities	0	N	Y
	1552	2323	1112
Proportions of modalities (out of 1)	0.3112091	0.4658111	0.2229797

DisbursementGross

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.

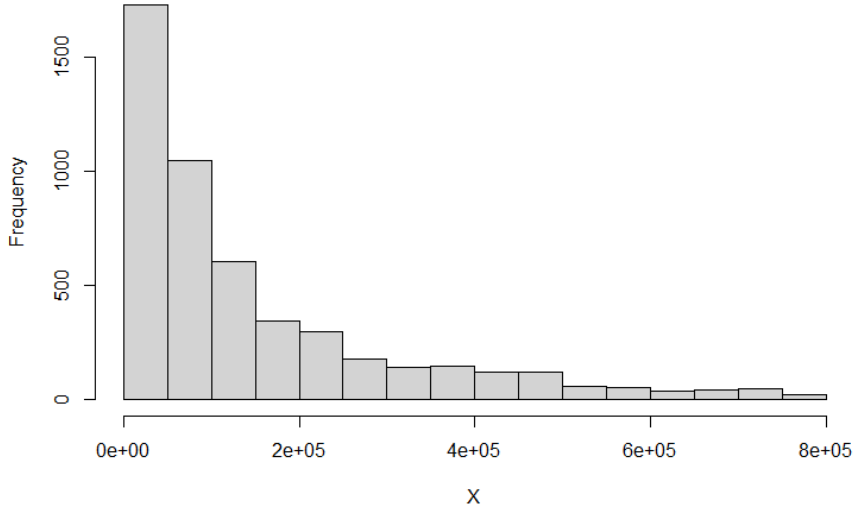


Image 69: Histogram of DisbursementGross

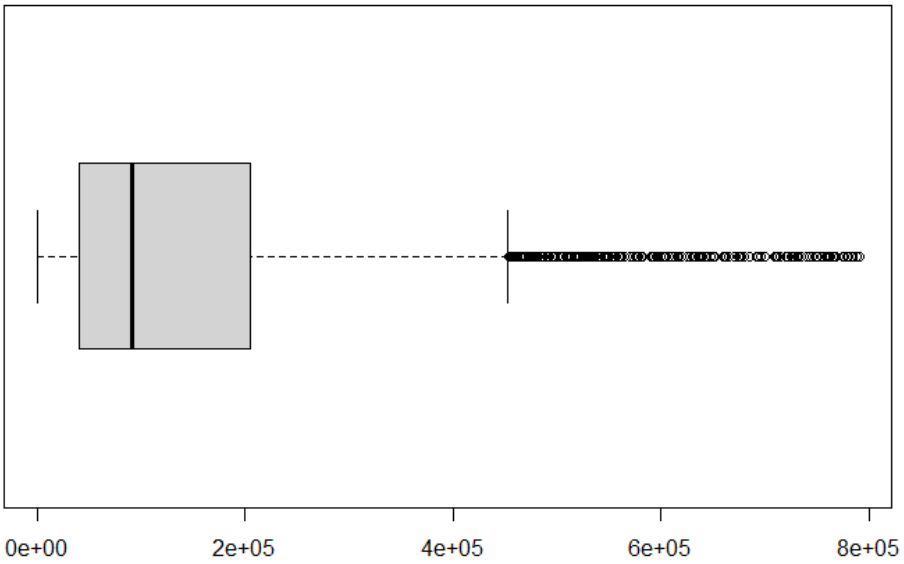


Image 70: Boxplot of DisbursementGross

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	40000	91713	154139	204863	790000

"sd: 165401.031639747"
"vc: 1.07306151329542"

MIS_Status

- We deleted all the observations which contained outliers. This is our response variable so we chose not to impute new values for those NAs.

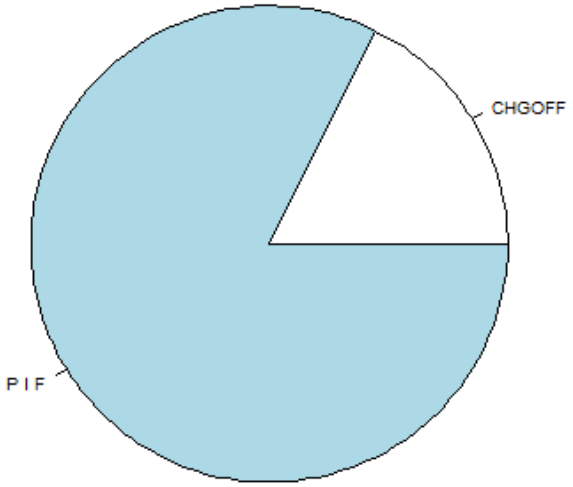


Image 71: Pie of MIS_Status

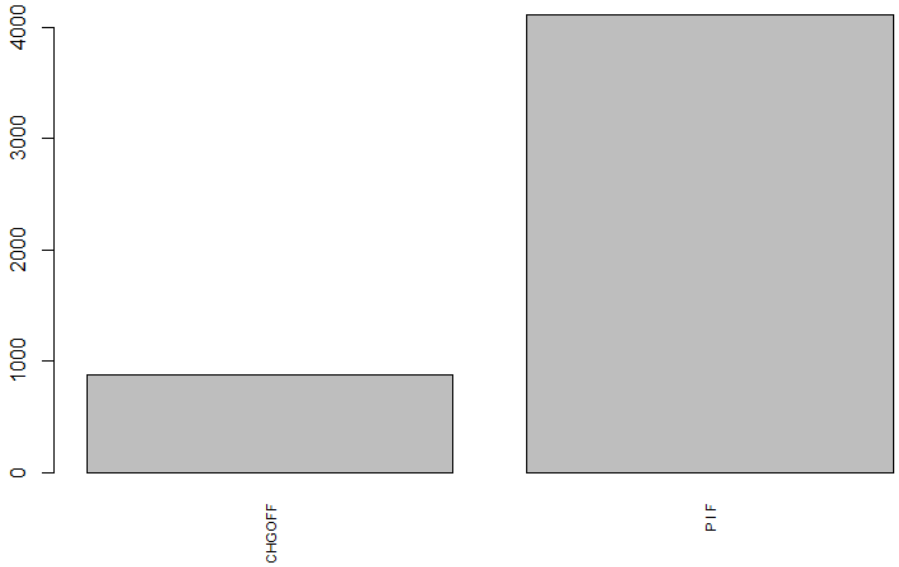


Image 72: Barplot of MIS_Status

Statistics of variable “MIS_Status”		
Number of modalities	2	
Frequency table of the modalities	PIF	CHGOFF
	4109	878
Proportions of modalities (out of 1)	0.8239422	0.1760578

GrAppv

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.

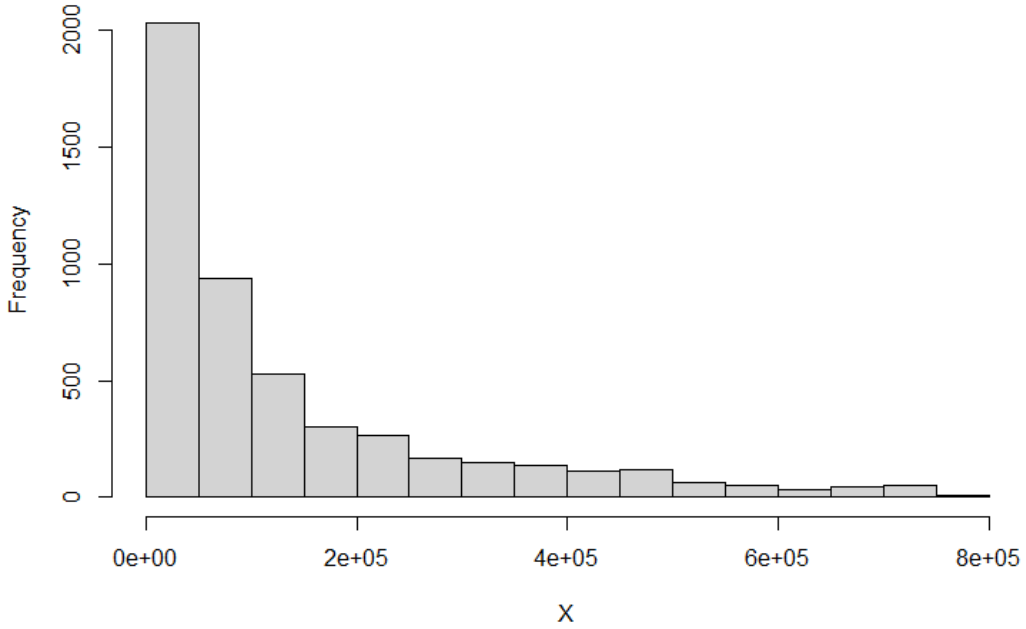


Image 73: Histogram of GrAppv

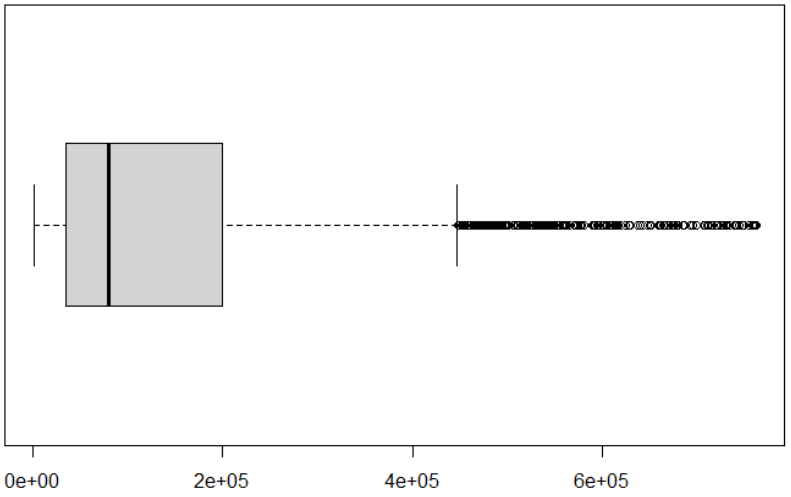


Image 74: Boxplot of GrAppv

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1000	35000	80000	146771	200000	762000

"sd: 164717.741360349"
"vc: 1.12227804604241"

SBA_Appv

- We placed as NA all the extreme outliers and then we imputed new values using the Knn algorithm.

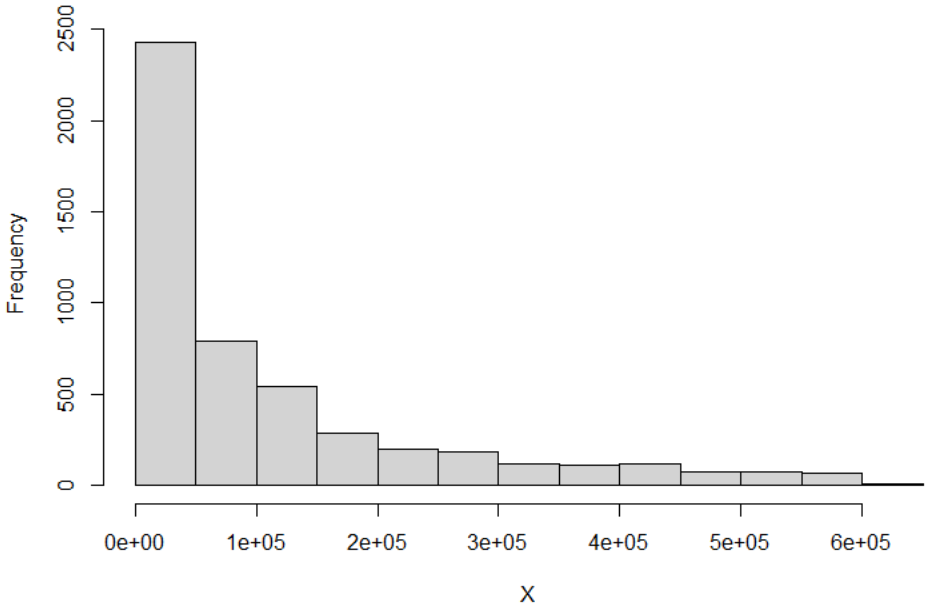


Image 75: Histogram of SBA_Appv

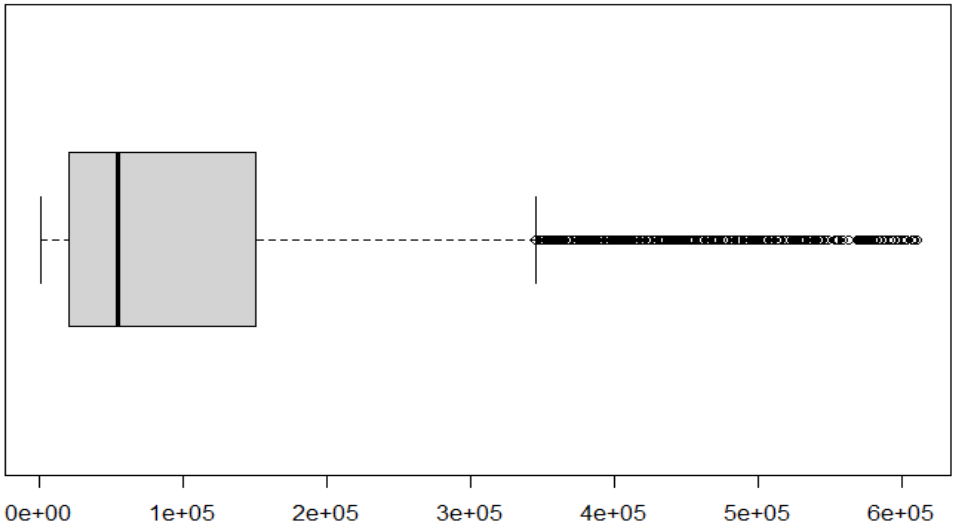


Image 76: Boxplot of SBA_Appv

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
500	20000	54114	114574	150000	610000

"sd: 136700.126459863"
"vc: 1.19312216227449"

WhichCompany

- This variable experienced no changes after the preprocessing.

yearsAfterAprv

- This variable is created after the preprocessing and represents the number of years that have passed since the loan was approved.

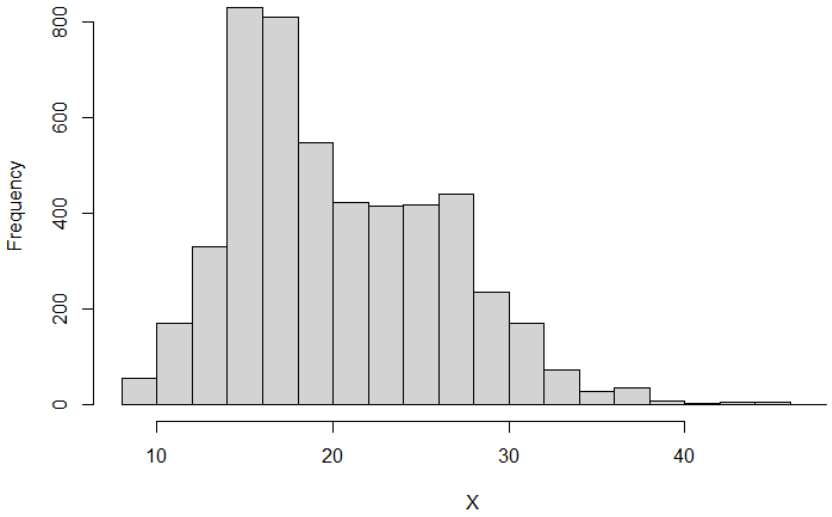


Image 77: Histogram of yearsAfterAprv

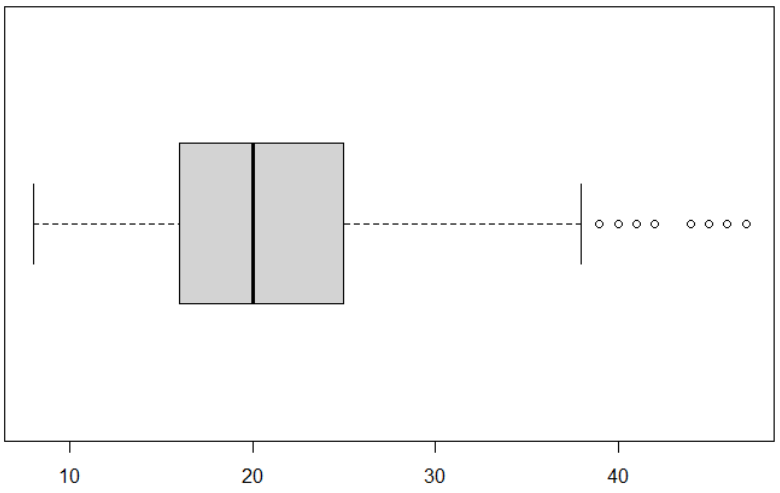


Image 78: Boxplot of yearsAfterAprv

"Extended Summary Statistics"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8	16	20	20.83	25	47

"sd: 5.98153782558898"
"vc: 0.287145975977581"

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Final Bivariate data description

Bivariate description of the quantitative variables

Making use of the library corrplot, we will graph a correlation plot between all the numeric variables in our dataset to analyze how they are correlated with each other. The variable yearsAfterAprv comes from the variable ApprovalFY so it has a perfect (inverse) correlation, also as a result of that, all the variables that correlate with ApprovalFY will also correlate with yearsAfterAprv.

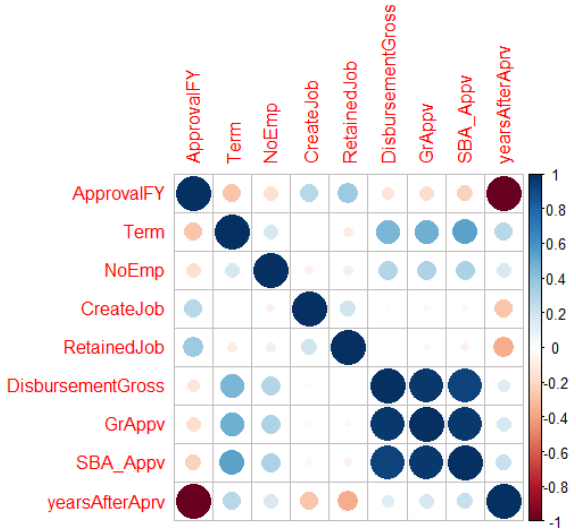


Image 79: Correlation plot between all the numeric variables

After preprocessing, the following variables were modified:

- Correlation between **Term** & **ApprovalFY**:

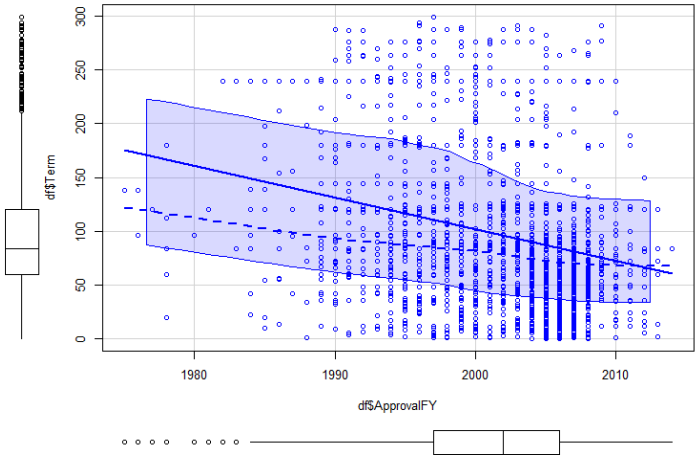


Image 80: Correlation between Term and ApprovalFY

  - Negative correlation (-0.3) The correlation between these two variables is the same as before preprocessing

- Correlation between **Term** & **Disbursement Gross**:

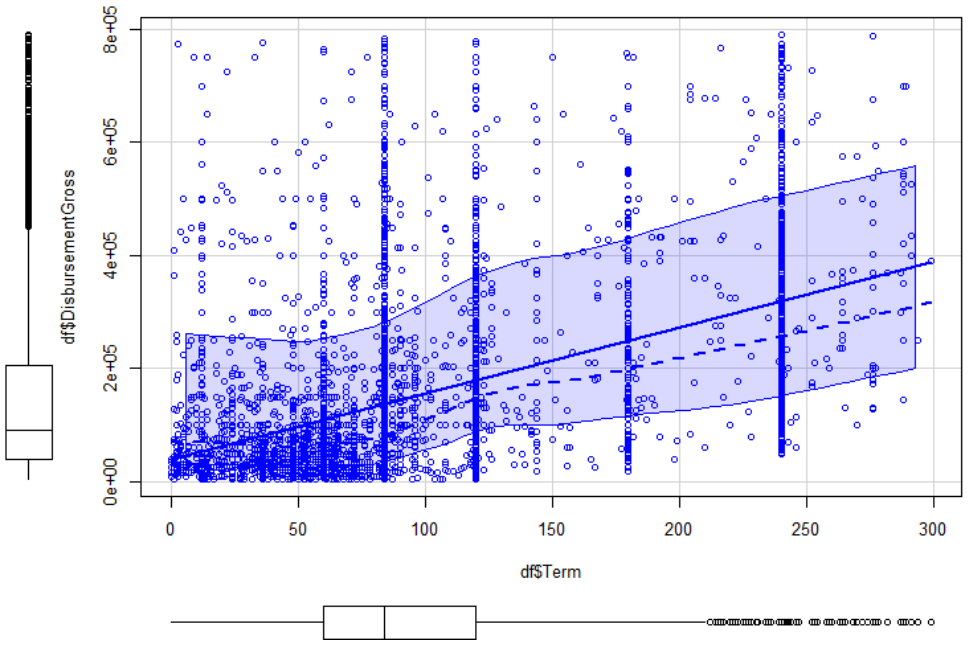


Image 81: Correlation between Term and DisbursementGross

  - Direct correlation (0.4); the correlation stays the same as before the preprocessing

- Correlation between **Term** & **GrAppv**:

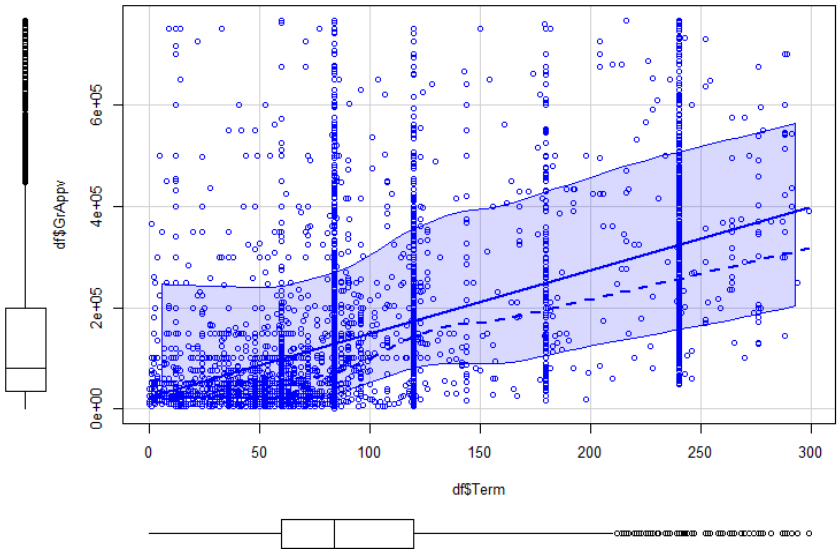


Image 82: Correlation between Term and GrAppv

  - Direct correlation (0.4); the correlation stays the same as before the preprocessing

- Correlation between **Term** & **SBA_Appv**:

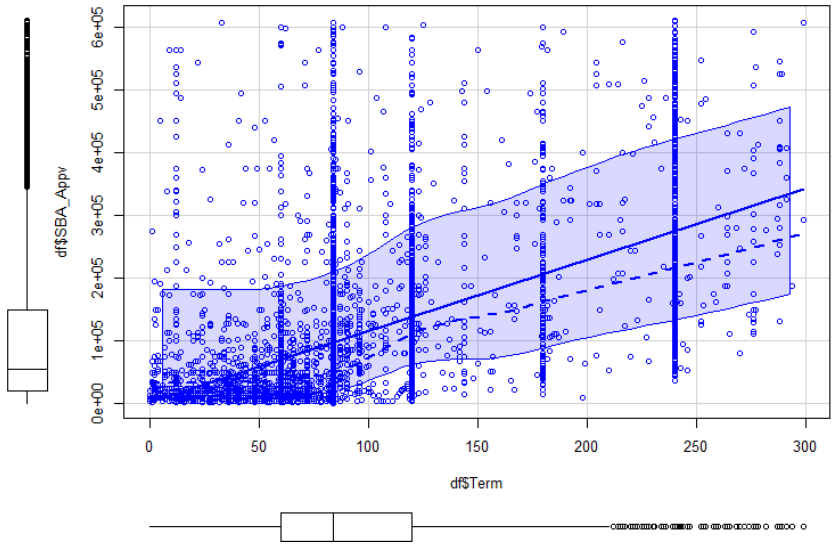


Image 83: Correlation between Term and SBA_Appv

  - Direct correlation (0.4); the correlation stays the same as before the preprocessing

- Correlation between **Retained Job** & **Create Job:**

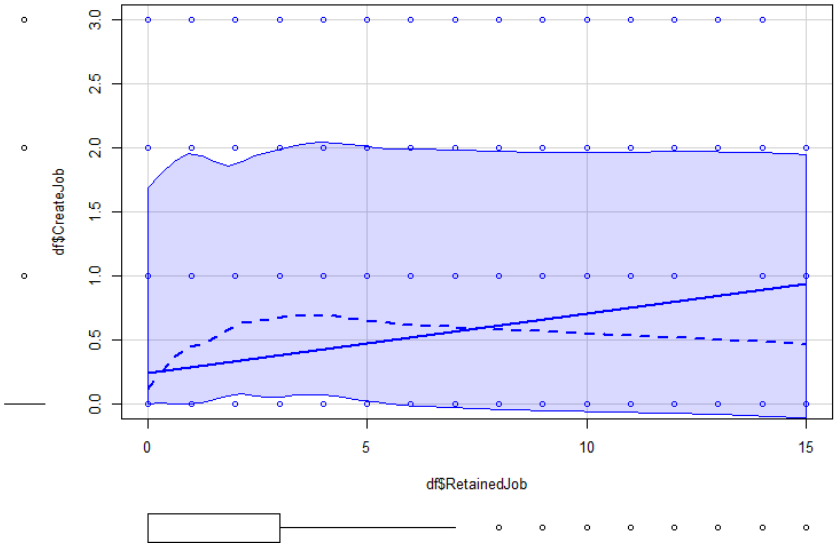


Image 84: Correlation between RetainedJob and CreateJob

  - The correlation between these two variables after the preprocessing has changed from a correlation of 1 to a correlation of 0.2. The reason why this has happened is because of the deletion of outliers.

- Correlation between **Disbursement Gross** & **GrAppv**:

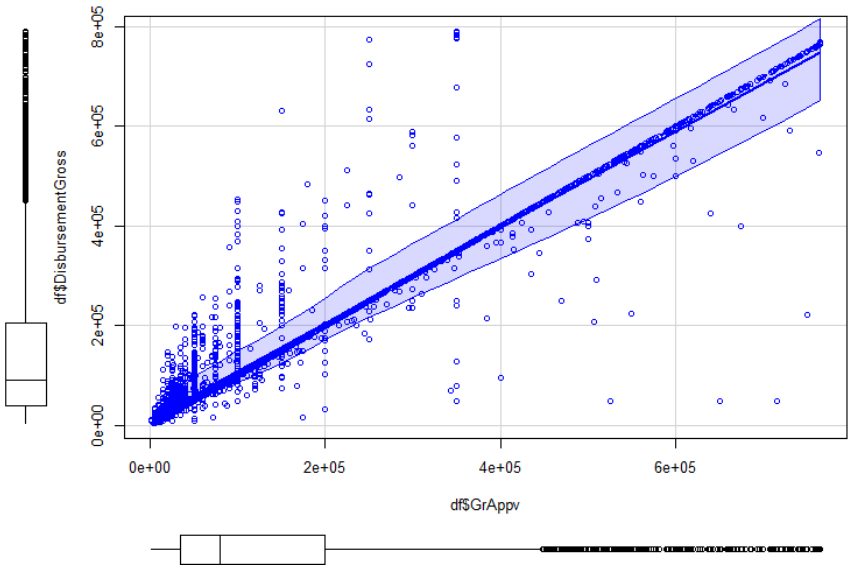


Image 85: Correlation between DisbursementGross and GrAppv

  - Direct correlation (1); the correlation stays the same as before the preprocessing

- Correlation between **Disbursement Gross** & **SBA_Appv**:

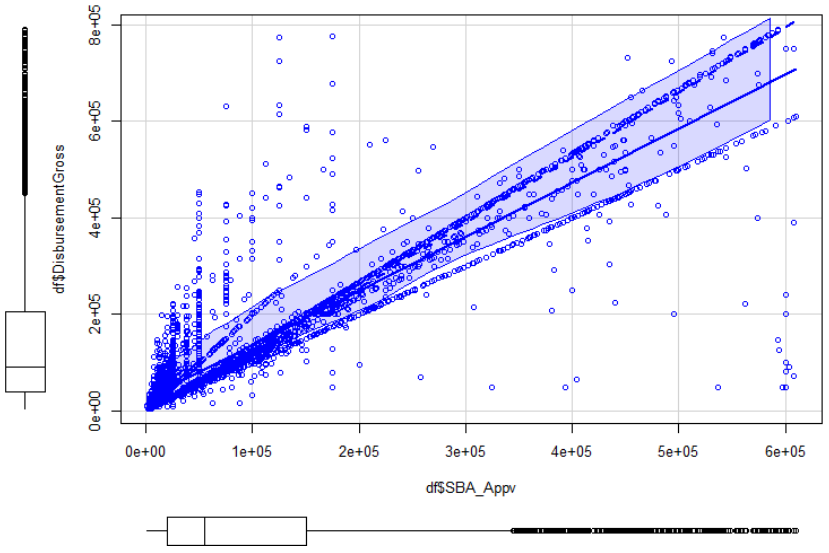


Image 86: Correlation between DisbursementGross and SBA_Appv

  - Direct correlation (1); the correlation stays the same as before the preprocessing

- Correlation between **GrAppv** & **SBA_Appv**

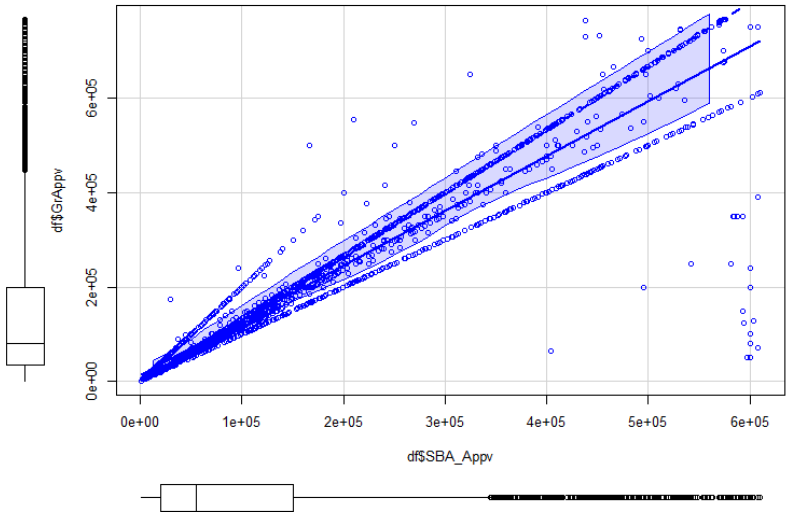


Image 87: Correlation between GrAppv and SBA_Appv

  - Direct correlation (1); the correlation stays the same as before the preprocessing

- Correlation between **ApprovalFY** & **RetainedJob**:

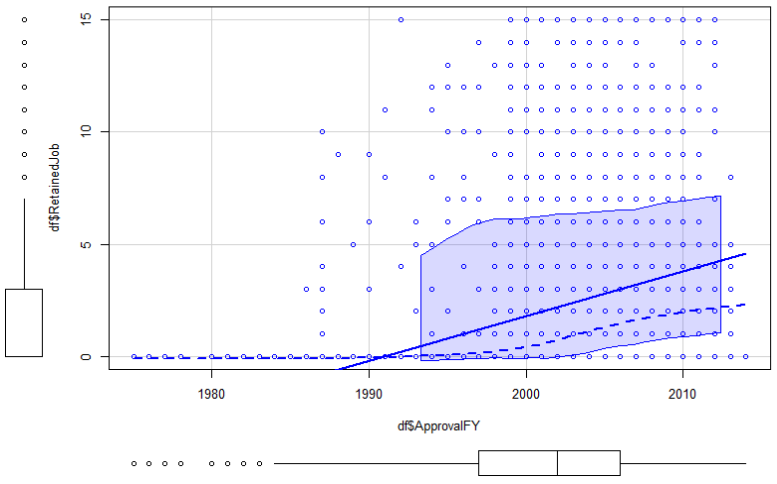


Image 88: Correlation between ApprovalFY and RetainedJob

  - Now that we have processed the outliers of Retained Job, we discover a new correlation between this variable and the year of approval (ApprovalFY). From the scatterplot we can observe that as the year of approval of the loan increases, the jobs retained in the business also increases.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Bivariate description of categorical variables with quantitative variables

In this section we will analyze the correlation between our categorical and numerical variables. To keep this analysis short we will only be analyzing the correlation of all our numerical variables with our response variable that happens to be a categorical variable.

- Correlation between **ApprovalFY** & **MIS_Status**

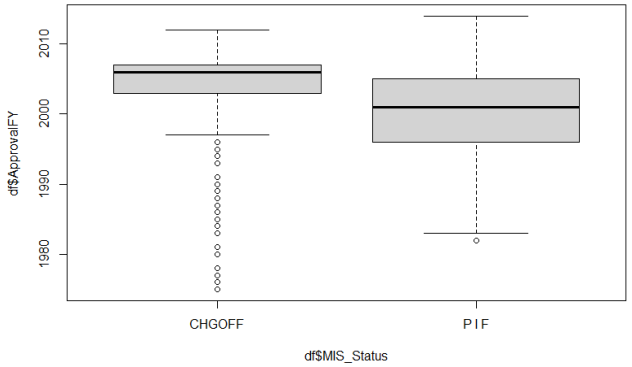


Image 89: Correlation between ApprovalFY and MIS_Status

	Mean in category	Overall mean
CHGOFF	2.003515e+03	2.001162e+03
PIF	2.000660e+03	2.001162e+03

p-value → 1.078211e-37

- Correlation between **Term** & **MIS_Status**

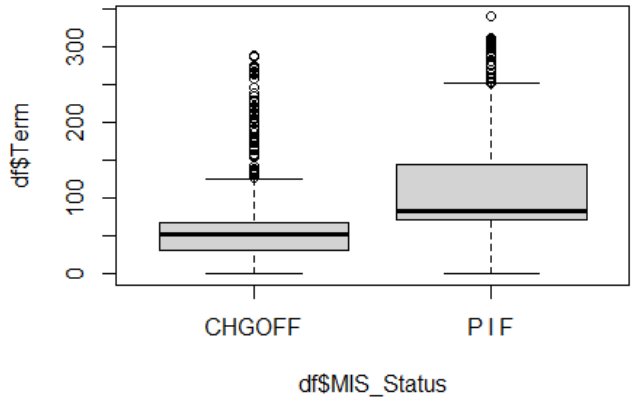


Image 90: Correlation between Term and MIS_Status

	Mean in category	Overall mean
CHGOFF	58.37927	98.60497
PIF	107.2003	98.60497

p-value → 1.914108e-95

- Correlation between **NoEmp** & **MIS_Status**

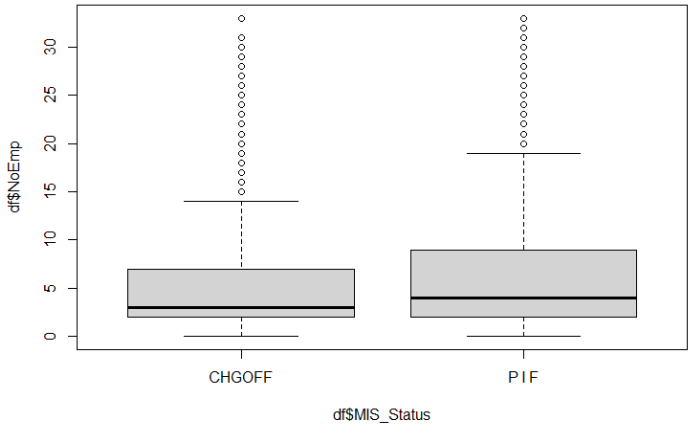


Image 91: Correlation between NoEmp and MIS_Status

	Mean in category	Overall mean
CHGOFF	5.403	6.345
PIF	6.54	6.345

p-value → 5.115900e-06

- Correlation between **CreateJob** & **MIS_Status**

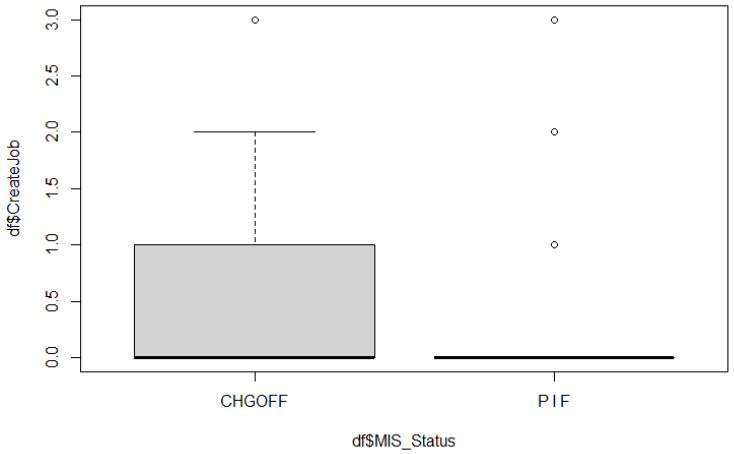


Image 92: Correlation between CreateJob and MIS_Status

	Mean in category	Overall mean
CHGOFF	0.469	0.335
PIF	0.306	0.335

p-value → 1.132097e-08

- Correlation between **RetainedJob** & **MIS_Status**

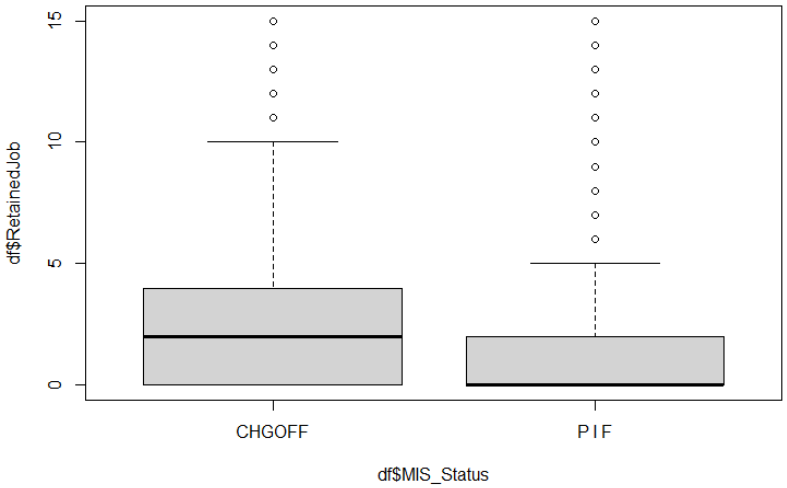


Image 93: Correlation between RetainedJob and MIS_Status

	Mean in category	Overall mean
CHGOFF	2.702	2.0228
PIF	1.877	2.0228

p-value → 1.734078e-11

- Correlation between **DisbursementGross** & **MIS_Status**

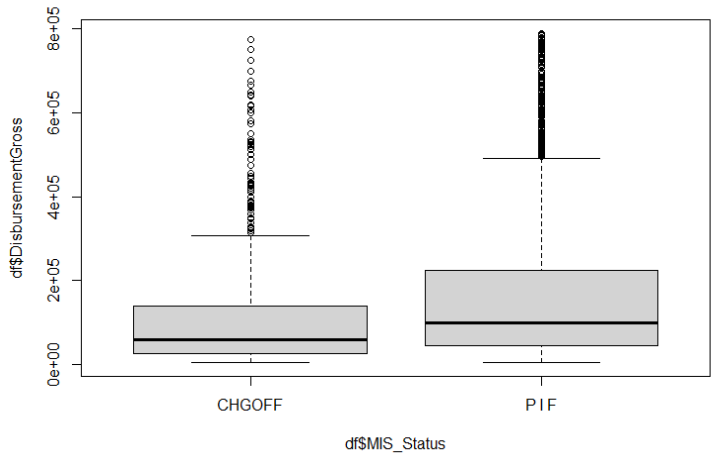


Image 94: Correlation between DisbursementGross and MIS_Status

	Mean in category	Overall mean
CHGOFF	1.141854e+05	1.539883e+05
PIF	1.624932e+05	1.539883e+05

p-value → 2.955844e-15

- Correlation between **GrAppv** & **MIS_Status**

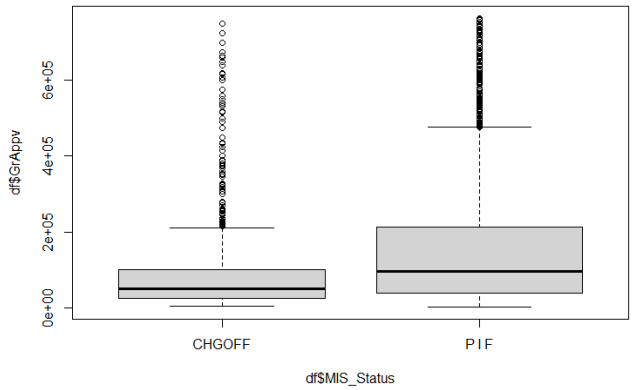


Image 95: Correlation between GrAppv and MIS_Status

	Mean in category	Overall mean
CHGOFF	1.012537e+05	1.468330e+05
PIF	1.565722e+05	1.468330e+05

p-value → 1.454018e-19

- Correlation between **SBA_Appv** & **MIS_Status**

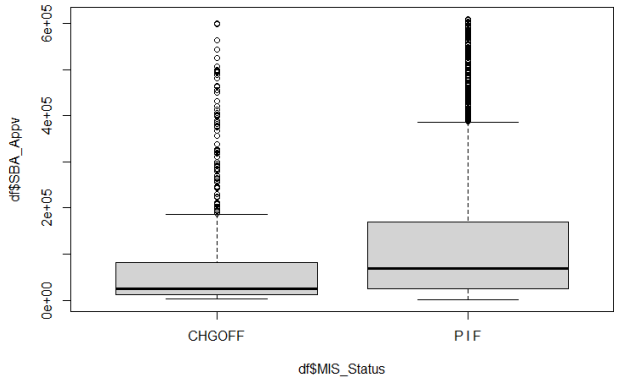


Image 97: Correlation between SBA_Appv and MIS_Status

	Mean in category	Overall mean
CHGOFF	7.307423e+04	1.145397e+05
PIF	1.233999e+05	1.145397e+05

p-value → 3.388829e-23

All the p-values are below 0.05, therefore we reject the null hypothesis of not being any differences in the values in CHGOFF and PIF.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Bivariate description of categorical variables

After preprocessing, we no longer have NA values in our response variable, therefore we only have two categories in our response variable.

- Correlation between **State** & **MIS_Status**

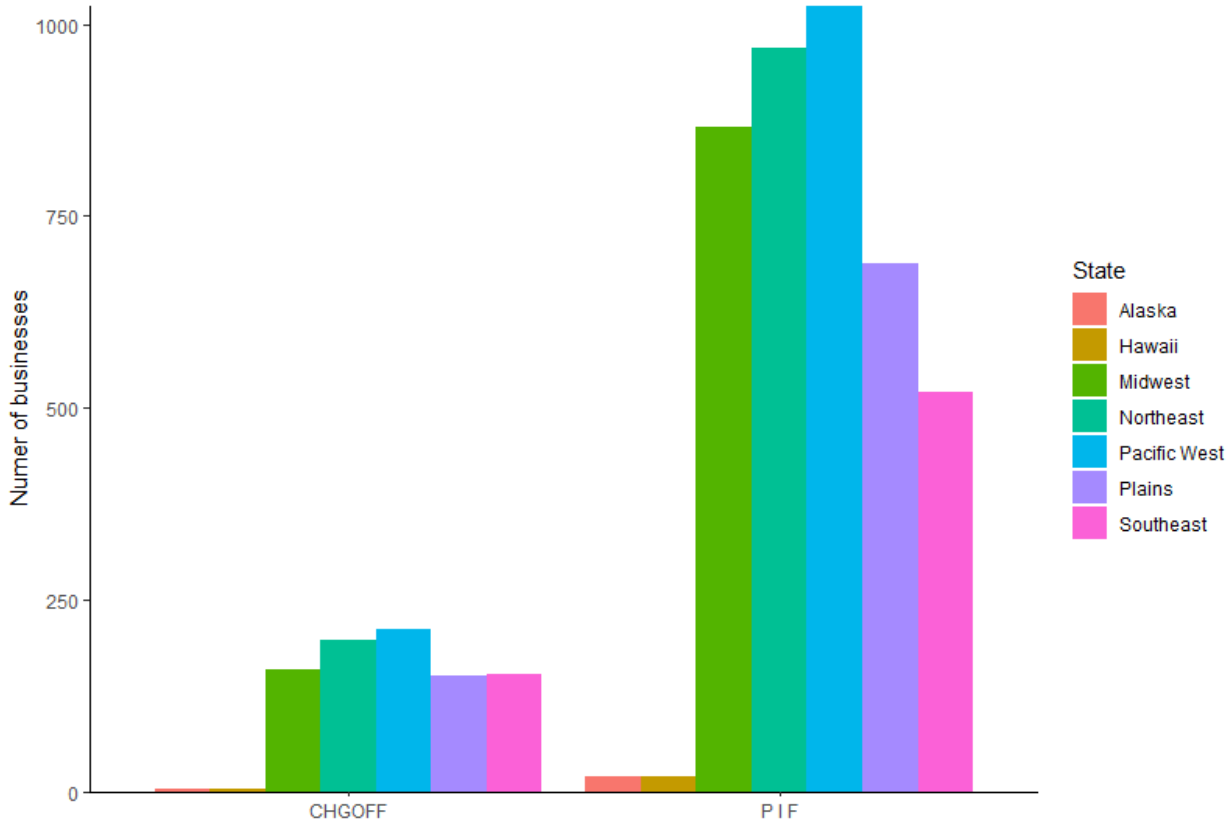


Image 98: Correlation between State and MIS_Status

Row Percentages	**Column Percentages**

```
                CHGOFF     P I F
Alaska       0.1363636 0.8636364
Hawaii       0.1666667 0.8333333
Midwest      0.1551220 0.8448780
Northeast    0.1688089 0.8311911
Pacific West 0.1708502 0.8291498
Plains       0.1797619 0.8202381
Southeast    0.2270030 0.7729970
```

```
                CHGOFF     P I F
Alaska       0.1363636 0.8636364
Hawaii       0.1666667 0.8333333
Midwest      0.1551220 0.8448780
Northeast    0.1688089 0.8311911
Pacific West 0.1708502 0.8291498
Plains       0.1797619 0.8202381
Southeast    0.2270030 0.7729970
```

- Correlation between **BankState** & **MIS_Status**

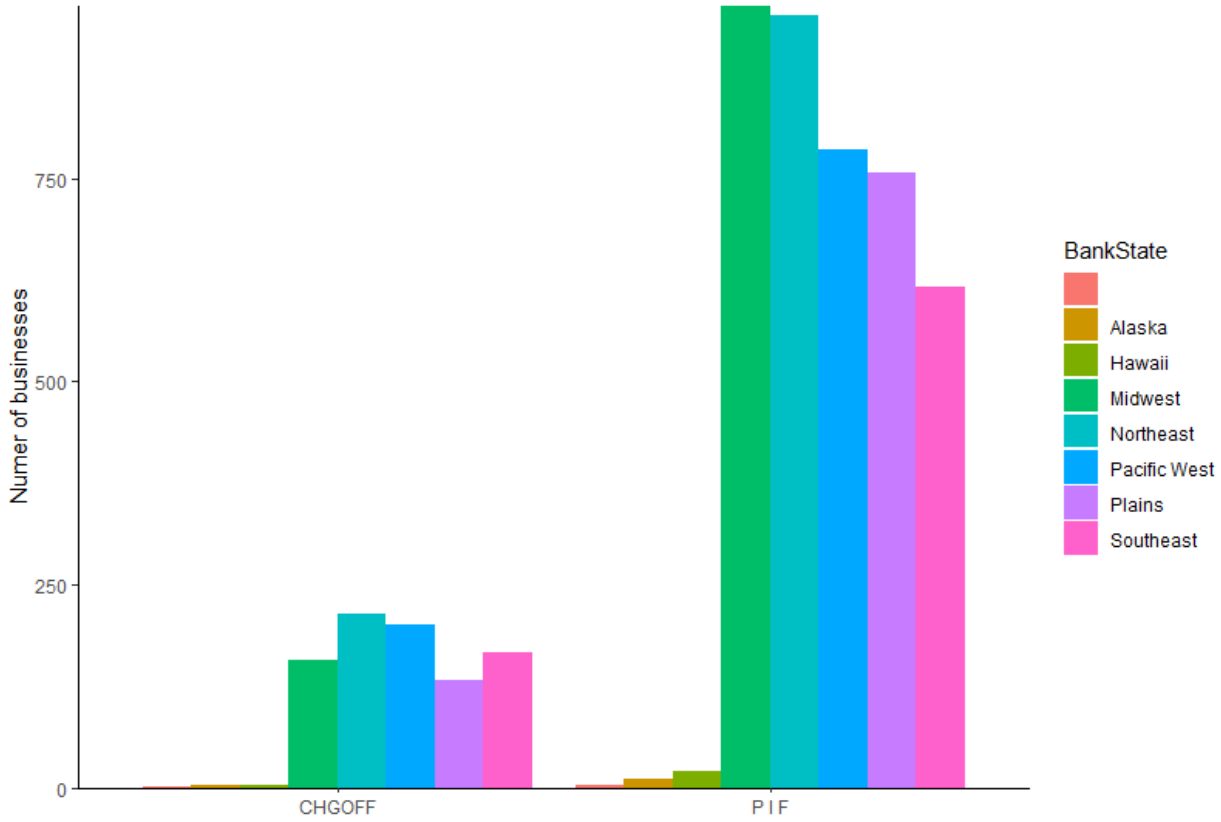


Image 99: Correlation between BankState and MIS_Status

Row Percentages

	CHGOFF	P I F
	0.2000000	0.8000000
Alaska	0.2307692	0.7692308
Hawaii	0.1304348	0.8695652
Midwest	0.1401786	0.8598214
Northeast	0.1835334	0.8164666
Pacific West	0.2036474	0.7963526
Plains	0.1484814	0.8515186
Southeast	0.2130102	0.7869898

Column Percentages

	CHGOFF	P I F
	0.0011389522	0.0009734729
Alaska	0.0034168565	0.0024336822
Hawaii	0.0034168565	0.0048673643
Midwest	0.1788154897	0.2343635921
Northeast	0.2437357631	0.2316865417
Pacific West	0.2289293850	0.1912874179
Plains	0.1503416856	0.1842297396
Southeast	0.1902050114	0.1501581893

- Correlation between **WhichCompany** & **MIS_Status**

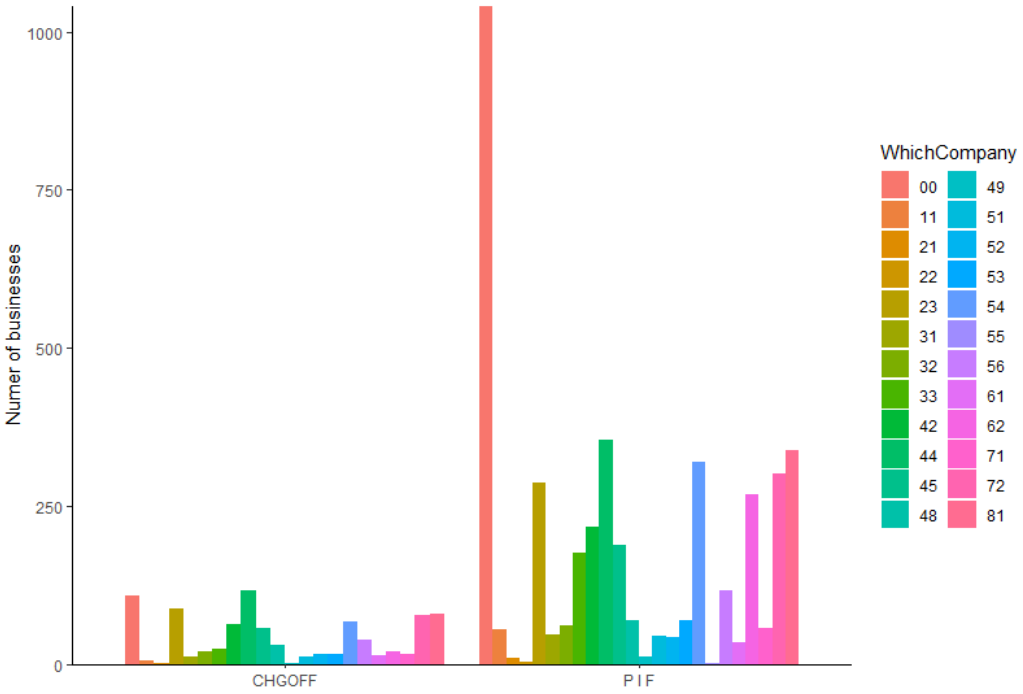


Image 100: Correlation between WhichCompany and MIS_Status

Row Percentages	Column Percentages

```
        CHGOFF      P I F
00 0.09478261 0.90521739
11 0.08333333 0.91666667
21 0.10000000 0.90000000
22 0.00000000 1.00000000
23 0.23529412 0.76470588
31 0.20689655 0.79310345
32 0.23750000 0.76250000
33 0.12060302 0.87939698
42 0.22302158 0.77697842
44 0.24788136 0.75211864
45 0.23170732 0.76829268
48 0.31000000 0.69000000
49 0.08333333 0.91666667
51 0.19642857 0.80357143
52 0.26315789 0.73684211
53 0.17857143 0.82142857
54 0.17357513 0.82642487
55 0.00000000 1.00000000
56 0.24675325 0.75324675
61 0.27659574 0.72340426
62 0.06597222 0.93402778
71 0.21917808 0.78082192
72 0.20526316 0.79473684
81 0.19093079 0.80906921
```

```
           CHGOFF         P I F
00 0.1241457859 0.2533463130
11 0.0056947608 0.0133852519
21 0.0011389522 0.0021903139
22 0.0000000000 0.0007301046
23 0.1002277904 0.0696033098
31 0.0136674260 0.0111949379
32 0.0216400911 0.0148454612
33 0.0273348519 0.0425894378
42 0.0706150342 0.0525675347
44 0.1332574032 0.0863957167
45 0.0649202733 0.0459965928
48 0.0353075171 0.0167924069
49 0.0011389522 0.0026770504
51 0.0125284738 0.0109515697
52 0.0170842825 0.0102214651
53 0.0170842825 0.0167924069
54 0.0763097950 0.0776344609
55 0.0000000000 0.0002433682
56 0.0432801822 0.0282307131
61 0.0148063781 0.0082745193
62 0.0216400911 0.0654660501
71 0.0182232346 0.0138719883
72 0.0888382688 0.0734972013
81 0.0911161731 0.0825018253
```

- Correlation between **NewExist** & **MIS_Status**

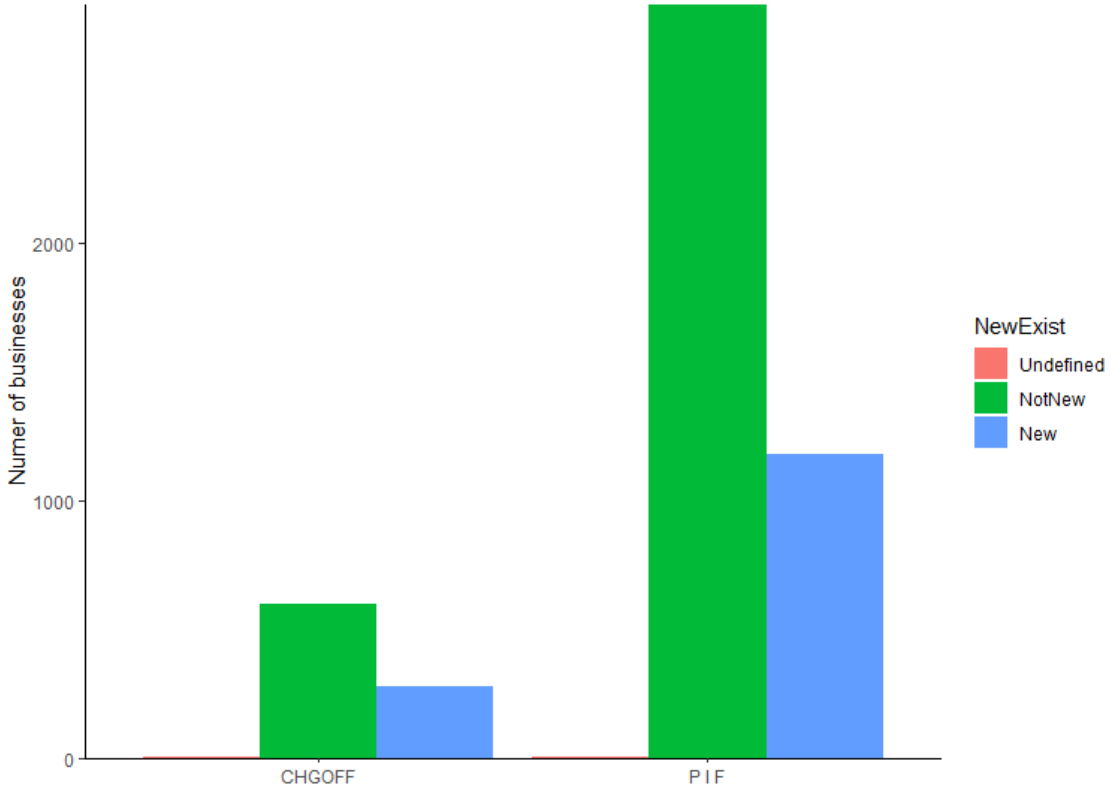


Image 101: Correlation between NewExist and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F Undefined 0.1666667 0.8333333 NotNew 0.1700255 0.8299745 New 0.1906722 0.8093278	CHGOFF P I F Undefined 0.001138952 0.001216841 NotNew 0.682232346 0.711608664 New 0.316628702 0.287174495

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

- Correlation between **UrbanRural** & **MIS_Status**

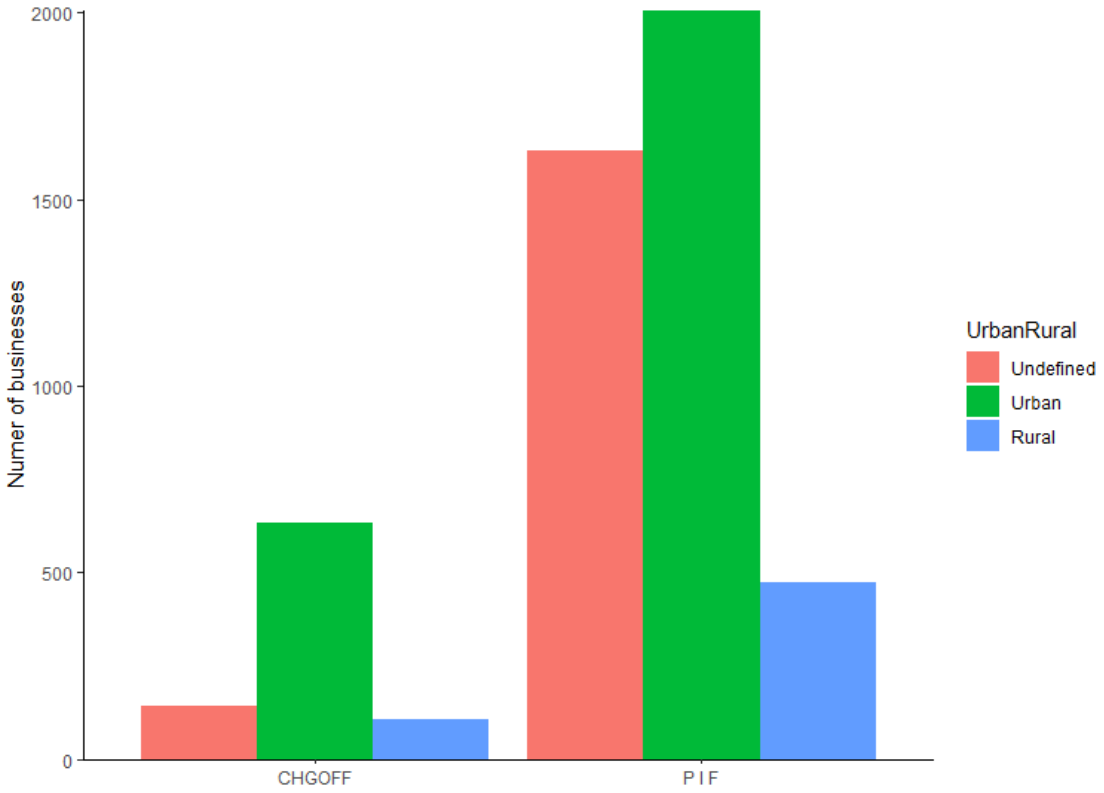


Image 102: Correlation between UrbanRural and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F Undefined 0.08018069 0.91981931 Urban 0.23928707 0.76071293 Rural 0.18134715 0.81865285	CHGOFF P I F Undefined 0.1617312 0.3964468 Urban 0.7186788 0.4881966 Rural 0.1195900 0.1153565

- Correlation between **RevLineCr** & **MIS_Status**

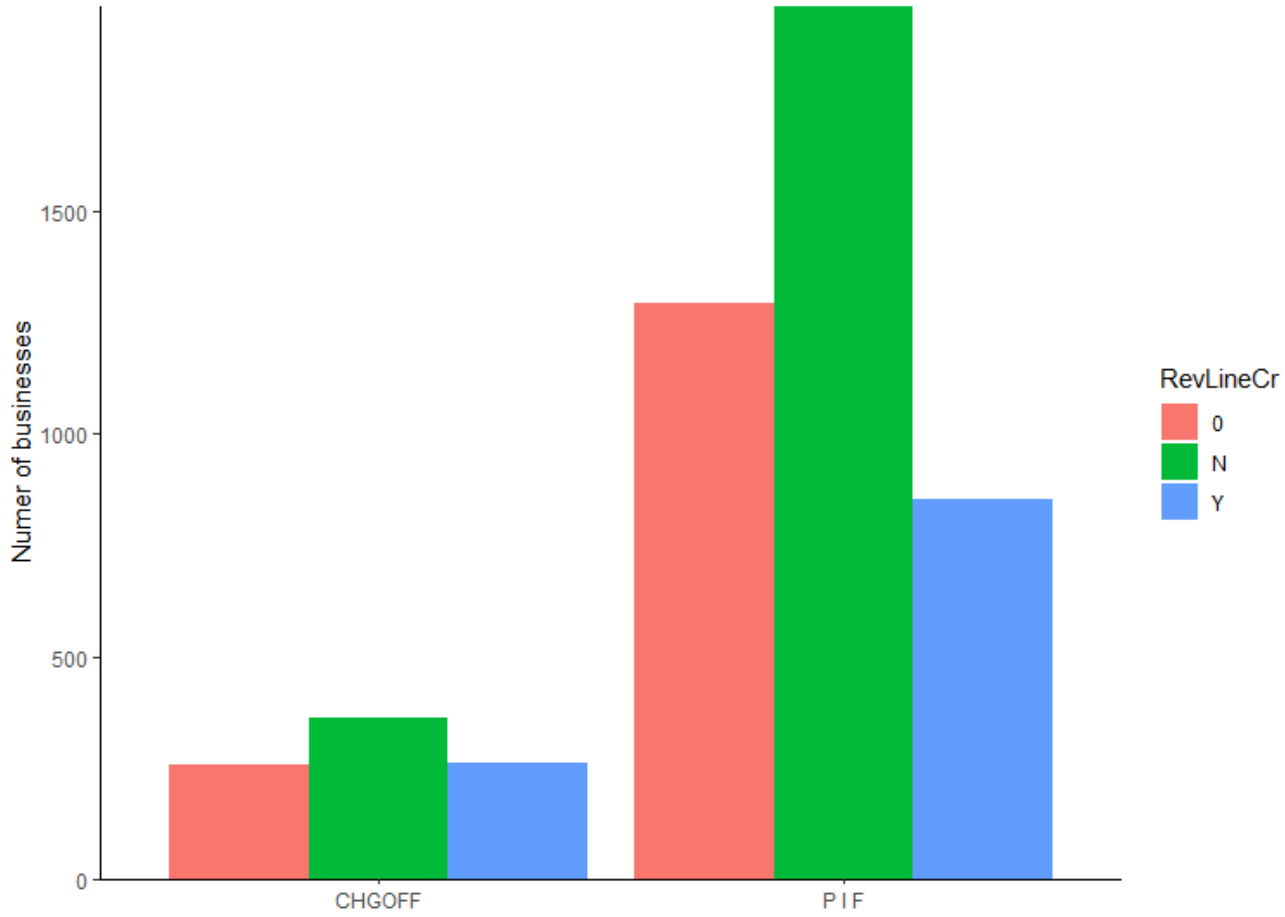


Image 103: Correlation between RevLineCr and MIS_Status

Row Percentages	Column Percentages
CHGOFF P I F 0 0.1662371 0.8337629 N 0.1549720 0.8450280 Y 0.2338129 0.7661871	CHGOFF P I F 0 0.2938497 0.3149185 N 0.4100228 0.4777318 Y 0.2961276 0.2073497

After the preprocessing the number of modalities has decreased from 6 to 3.

PCA

Principal Component Analysis, or PCA, is a method that consists in reducing the dimensionality of large data sets, to achieve that reduction PCA takes a few data sets of variables that still contains most of the information of the data set. This reduction, naturally, comes with a loss in precision and accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. After achieving the reduction, exploring and analyzing the dataset becomes much easier and faster.

For this part of the project, initially, we will be working with the numerical variables we have in our data set, this variables are: *ApprovalFY*, *Term*, *NoEmp*, *CreateJob*, *RetainedJob*, *DisbursementGross*, *GrAppv*, *SBA_Appv* and *YearsAfterAprv*. In order to work with the Principal Component Analysis method, we have to check which level of dimensions we want so we can properly process the data and analyze correctly the results we obtain.

In order to understand what percentage of the total variation in the dataset can be explained by each principal component, we make use of scree plot. Scree plot is one of the easiest ways of visualizing and understanding the relationship between variation and the principal components. First step consists of executing the *prcomp()* function to perform principal component analysis. This function returns the standard deviation, the center and the scale in each attribute, the rotation between each pair of attributes and the score for each individual component.

With the standard deviation we obtained through the *prcomp()* function, we can calculate the total variance explained by each principal component executing the script we see in the Image 104.

```
inerProj<- results$sdev^2
totalIner<- sum(inerProj)
pinerEix <- inerProj/totalIner
```

Image 104: Script to obtain the total variance of all the principal components

The value *pinerEix* represents that total variance that we just explained. Once we have that value, we can obtain the scree plot that will allow us to visualize the variations and the principal components. Executing the script of the Image 105, we obtain the scree plot we see in the Image 106. We explain in the next section how we interpret the plot obtained and what information it gives us.

```
#SCREE PLOT
library(ggplot2)

qplot(numeriques, pinerEix) + geom_line() +
  xlab("Principal Component") + ylab("Variance Explained") +
  ggtitle("Scree Plot") + ylim(0,1)
```

Image 105: Script to obtain the scree plot

Scree Plot

Once we have the plot generated, we can analyze it and extract information from it. The horizontal axis represents the principal component and the vertical axis the percentage of total variance explained by each individual principal component. In the x-axis, we have the 9 numeric variables we have in our data set, each one has a dot to represent it in the scree plot. It's clearly visible that the first 3 principal components are the most representative and give us important information. Despite that, we can not choose how many principal components we take for just by looking at the plot.

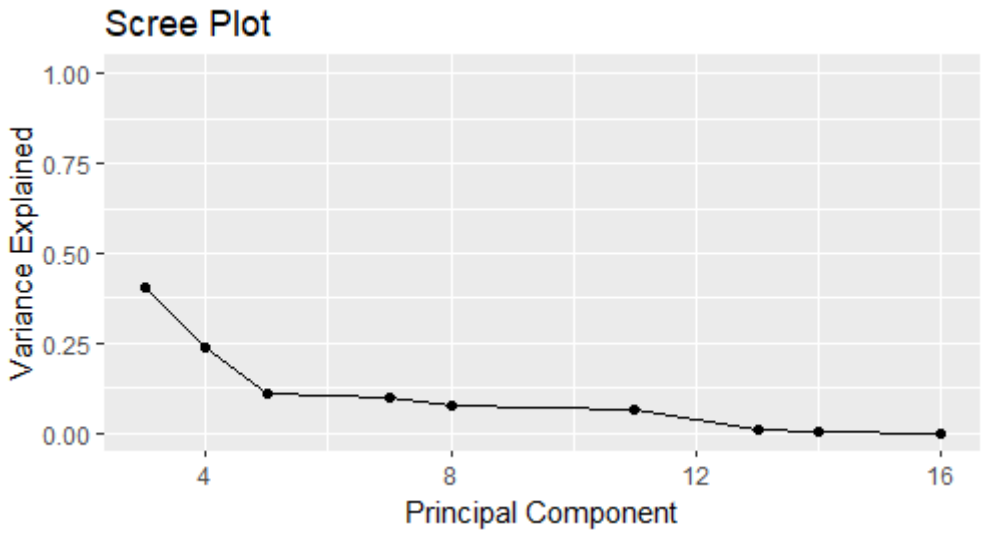


Image 106: Scree plot that represents the relationship between variation and the principal components

We should take the principal components that allow us to represent as much information as possible but try to minimize the number of components we take. Knowing that, we take as many principal components as we can until we arrive at the 0.8 of variance explained. Looking just at the plot it is kinda difficult to know how many we should take, in order to make things easier we obtain the plot represented in the Image 107, through a script.

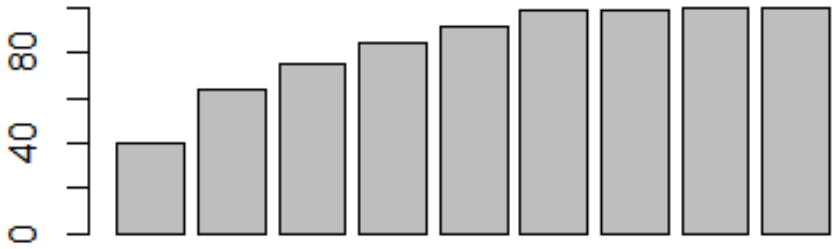


Image 107: Barplot that represent each one of the principal components

This plot gives us the same information we obtained through the scree plot we saw before but in a more explicit way. In the Image 107, we clearly see that we need to take the first 4 component analysis in order to have more than the 80 percent of variance explained. From now on, we will only work with this first principal component.

Factorial map visualization

In this section, we will explore and analyze the different factorial maps we generated with the new dimensionality obtained from the principal component analysis method. All the scripts used in this section can be found at the end of the report, *PCA script*. Through the different plots we can see the correlation between the numeric variables of the dataset and the existing relationship between the numeric variable and the qualitative variables. All the images will be attached at the end of this section.

Correlation between the numeric variables

The first factorial map generated is the one represented in the Image 108. In this plot are represented all the numerical variables of the dataset and we are able to visualize the correlation that exists between some of them. From a more mathematical point of view, the correlation between 2 variables is represented by the cosine of the angle generated by the 2 vectors, with that definition clear, a better analysis of the plot we see in the Image 108 can be done.

The two variables *ApprovalFY* and *yearsAfterAprv* have, naturally, an inverted correlation. The second one is generated from the first one, *yearsAfterAprv* is the value of doing 2022 minus the *ApprovalFY* value. So, as we can see their vectors are completely inverted from each other. They don't have any kind of correlation with some of the other numerical variables. The *CreateJob* and *RetainedJob* values are clearly correlated with each other, both of their vectors represented in the plot have almost the same direction and magnitudes. This means that the correlation between both of them is positive, in other words, they move in the same direction. The conclusion we extract is that *CreateJob* and *RetainedJob* have a strong relationship.

Between *SBA_Appv*, *GrAppv* and *DisbursementGross* seems to have a strong relationship between them with a positive correlation. Although at the beginning, we can see a pattern where the more they grow, the more they move away from each other so we can not confirm that those variables have a completely positive correlation. On the other hand, they don't have either a negative correlation, in conclusion, *SBA_Appv*, *GrAppv* and *DisbursementGross* are not entirely correlated.

The last pair of variables, *Term* and *NoEmp*, is a similar case like the previous one, where at the beginning it seems like there is a strong relationship between the variables but at the end there is no correlation at all. The difference between this case and the last one, is related to the knowledge we have about the variables. All three variables we saw previously were money related so it makes sense that there was a little correlation between them. *Term* and *NoEmp*, on the other hand, refers to the time to pay the loan and the number of employees of the company, respectively, which apparently don't make sense that are correlated. Despite that, we can see in the plot that there exists a little relationship between. Probably because bigger companies have more employees and also ask for more money, consequently, they have more time to return the loan, due to the amount of money being larger.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

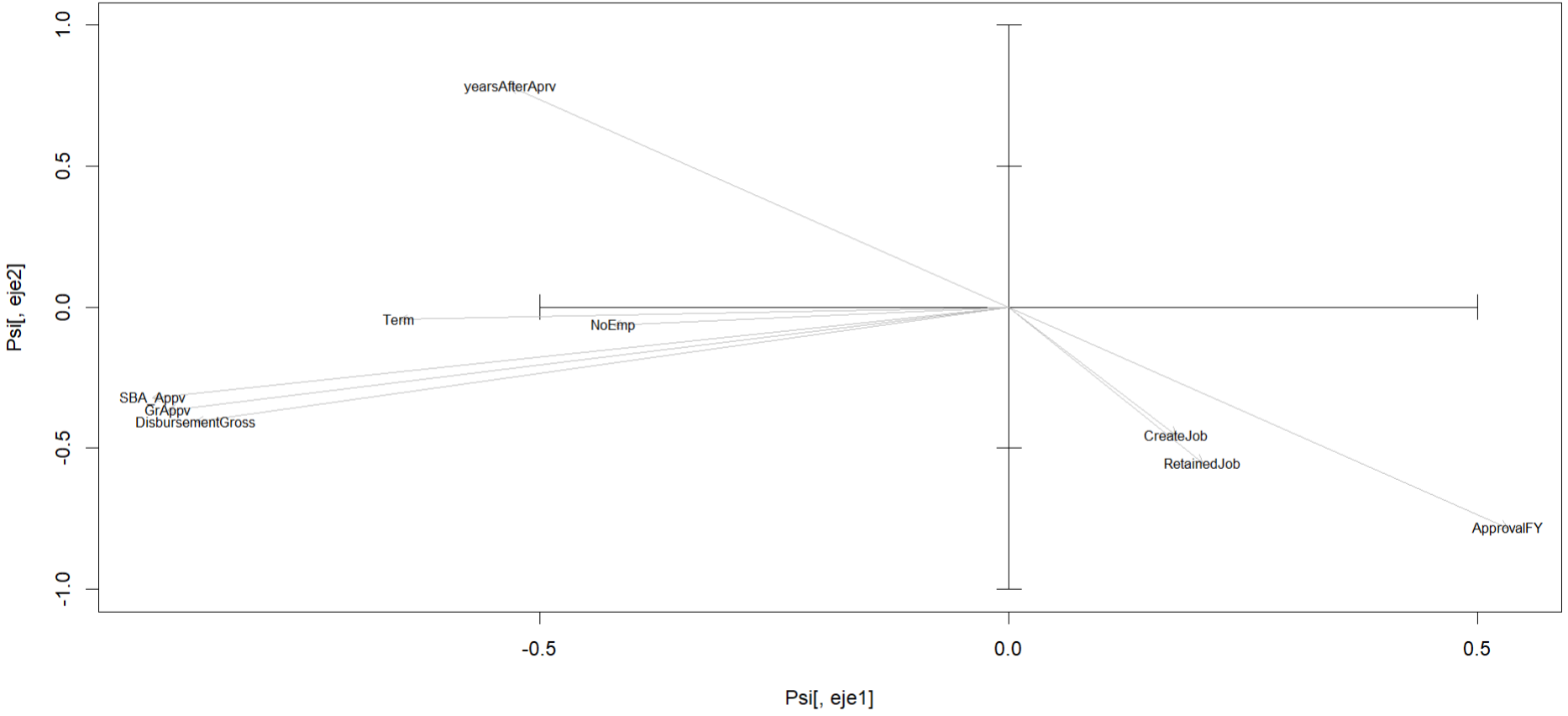


Image 108: Factorial map with the representation of the correlation between numerical variables of the dataset

State / BankState

This factorial map seen in Image 109, comprehends the correlation that exists between the variables *State* and *BankState* and the numerical variables. The first we notice is that each *State* category is near the same category represented in the *BankState* variable, seeing that we can conclude that companies tend to ask loans to banks located in the same state they are. Focusing on the *year sAfterAprv* and *ApprovalFY* on how the categories are distributed among them, we see the companies located in *AK* and *Plains* are much older, or nowadays, not that many companies in those states ask for loans. On the other hand, there seem to be new companies in *HI* and the *Northeast* asking for loans. To conclude this section, it can be seen that companies located in the *PacificWest* are companies that, normally, ask for more money from the bank.

NewExist

After analyzing the factorial map we see in Image 110, we can see that there is a fairly logical and expected trend that gives us the next information. The previously established companies, or the not new ones, are usually asking to get more money from the bank and tend to have a higher number of employees. About the new ones, we see the next pattern, they ask for less quantity of money, they have a lower number of employees and, generally, seem to be more recent loans, compared to the not new ones. Additionally, because they ask for less money, they also have less time to pay it back.

Urban Rural

In Image 111, there is a quite interesting trend that we did not expect, despite this and after analyzing, it makes sense that this pattern appears. Both urban and rural companies are located in the same area as the new companies. Reasoning this pattern, we can conclude that this variable didn't exist before and that's why old companies appear as undefined and all the new ones, whether urban or rural in the same area. Due to this rare situation we cannot conclude anything clearly about this variable.

RevLineCr

Variable *RevLineCr*, seems to be in a situation similar to the one we saw with the *UrbanRural* variable. Doing a little reminder, the revolving line of credit is a type of credit, established not so long ago, that consists in allowing the company to borrow another equal loan once it has paid the previous one. We can see in the factorial map displayed in the Image 112, that the only companies situated, once again, in the area to which the most recent companies belong. That's probably due to when older companies asked for a loan, the revolving line of credit didn't exist.

MIS_Status

In the factorial map that displays the *MIS_Status* variable correlation with the numerical variables, Image 113, we see a pattern not surprising at all. *MIS_Status* gives the information of which companies have already paid their loan and which don't. Relating again our factorial map with the areas where the new and old companies belong to, we can observe how the new ones usually have not paid their loans. On the other hand, those companies that requested the loan longer ago, naturally, have already paid the agreed amount.

WhichCompany

In Image 114, we can see the factorial map of the correlation between the variable that represents which type of company is one individual, *WhichCompany*, and all numerical variables of the data set. We observe that those companies related with mining, agriculture, fishing, oil and this type of industries, are companies that have a higher number of employees, asks for more money to loan from the bank and are companies that have been asking for loans for longer, probably because they have been on the industrial market for more time. We could say that they are older companies with a more solid base and experience. On the other hand, companies that are related with transportation, administration, finance and different services like education or construction, are newer companies that ask for less money and, due to being recent companies, provide a lot of new hires and retained jobs.

Correlation between all variables

In the last factorial map, Image 115, we can see all qualitative variables we commented represented in the same map.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

State and BankState

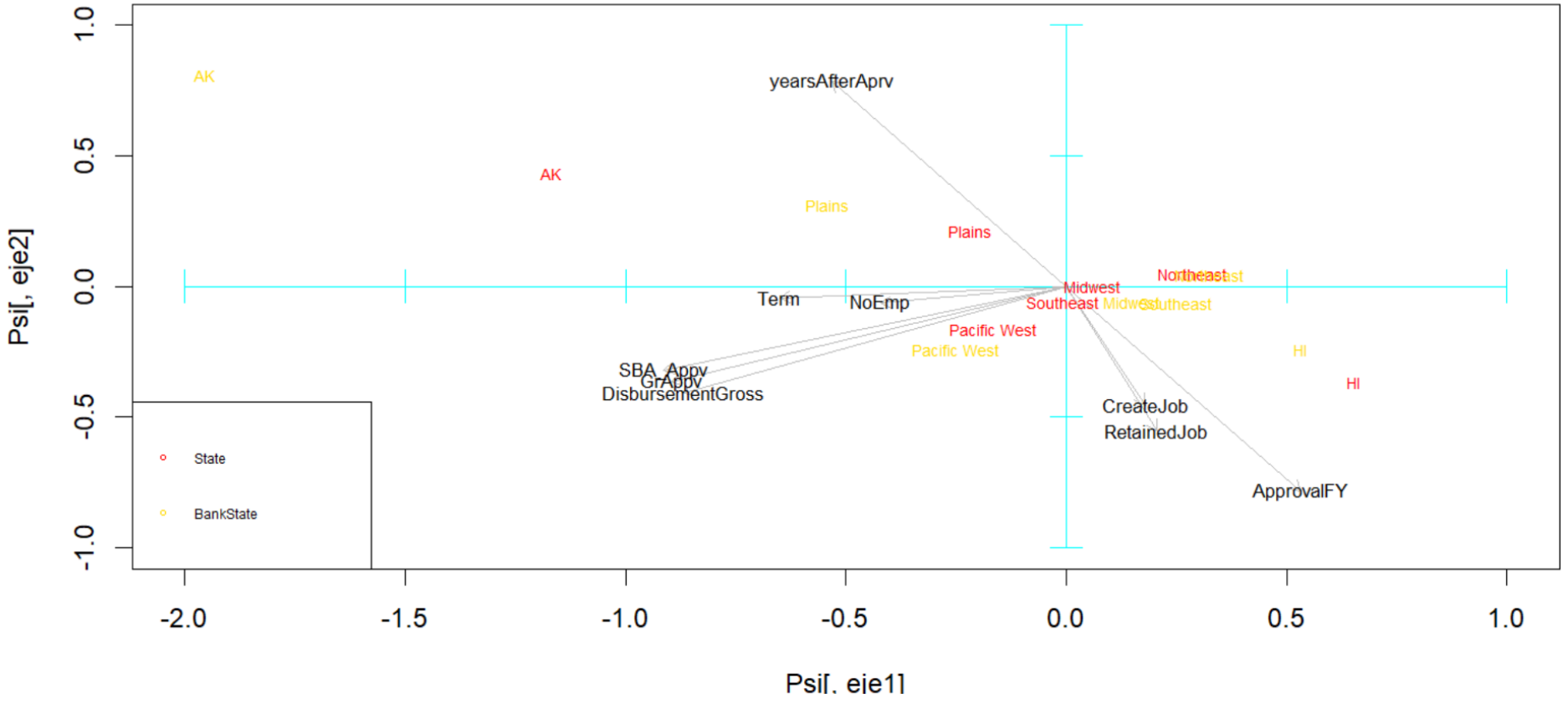


Image 109: Correlation between State and BankState variables with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

NewExist

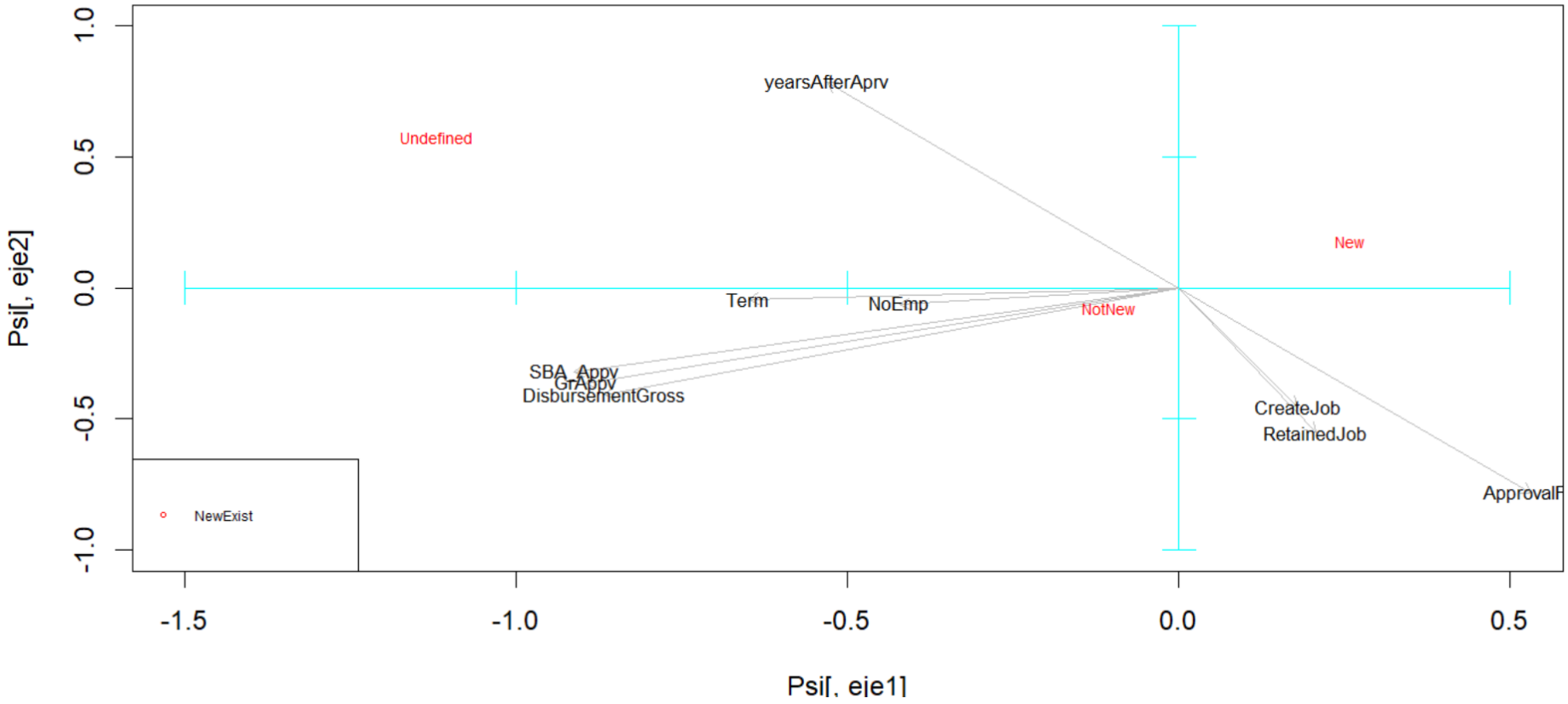


Image 110: Correlation between NewExist variable with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

UrbanRural

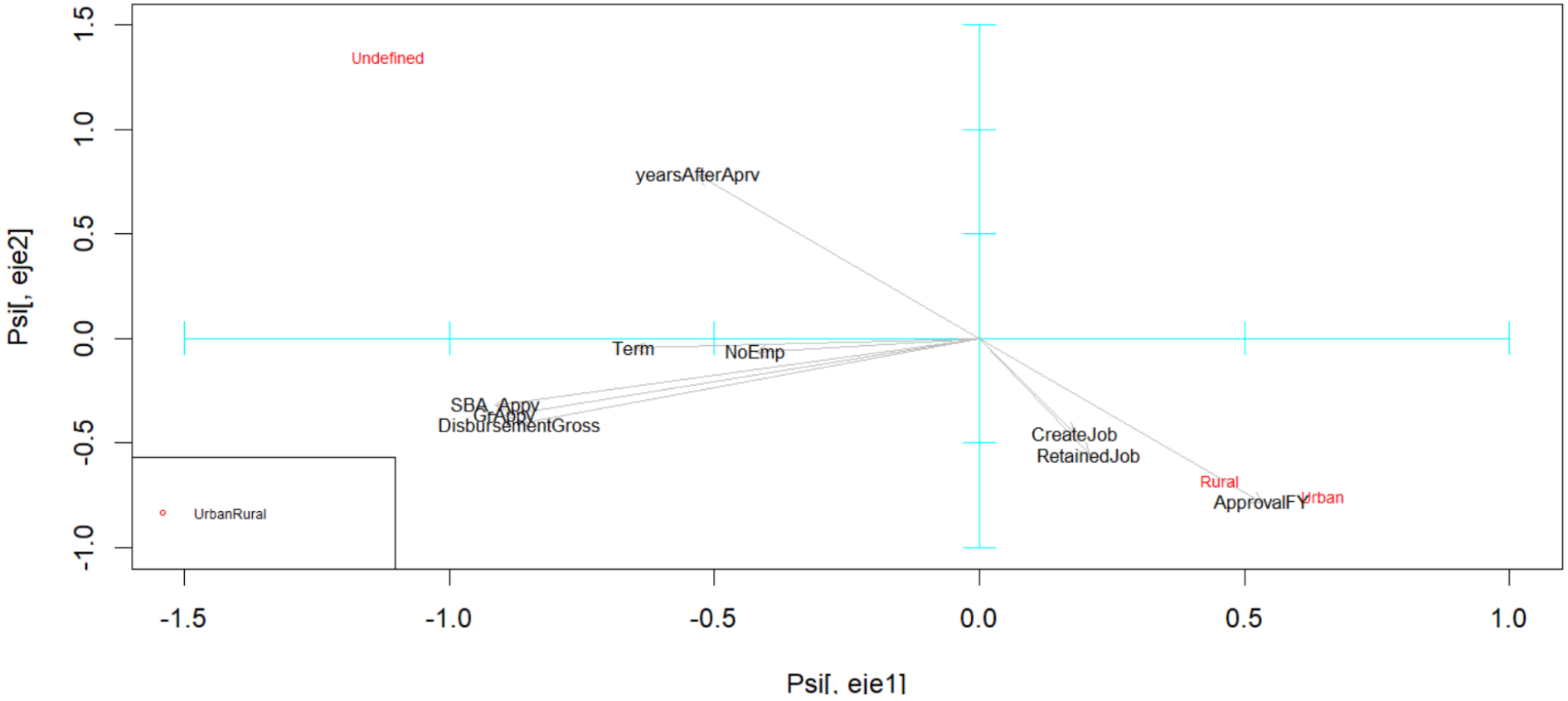


Image 111: Correlation between UrbanRural variable with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

RevLineCr

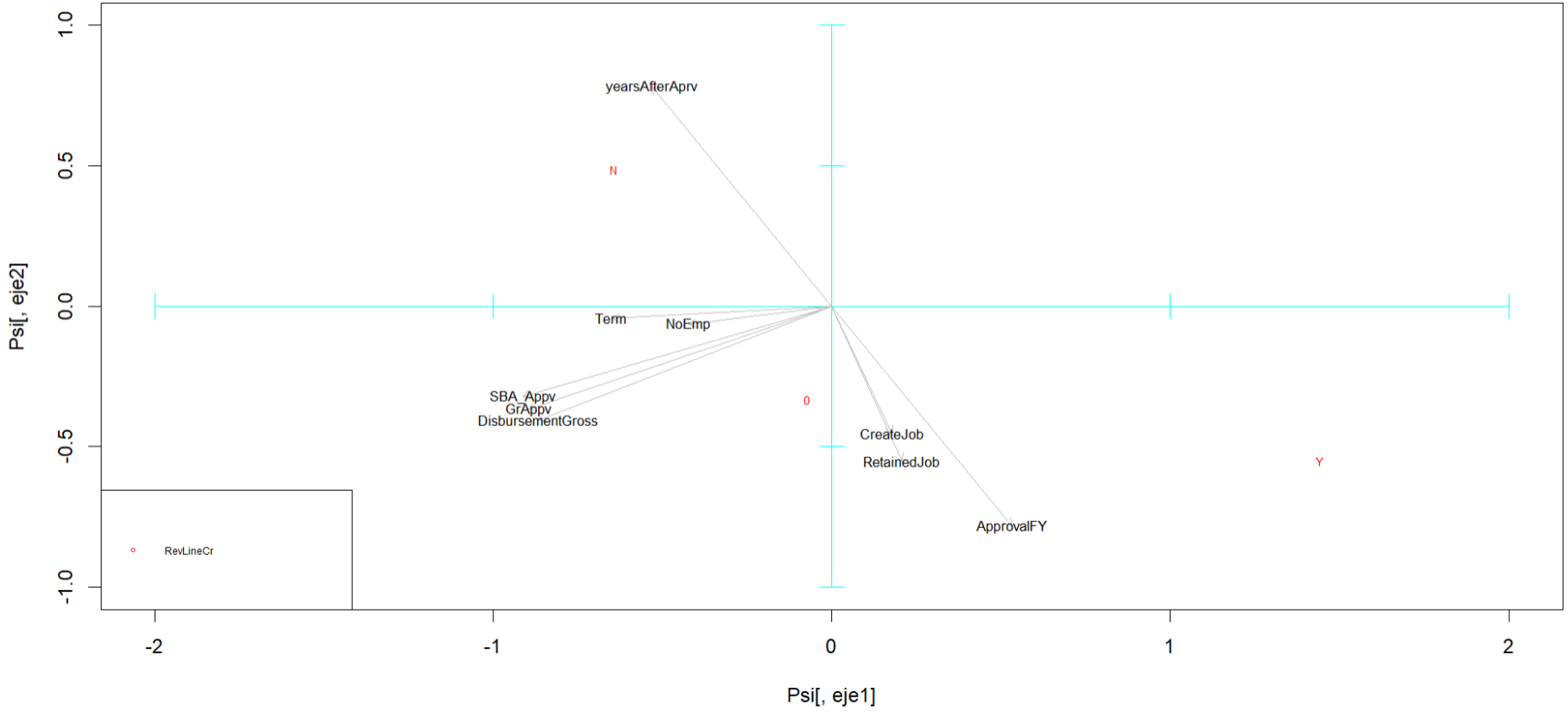


Image 112: Correlation between RevLineCr variable with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

MIS_Status

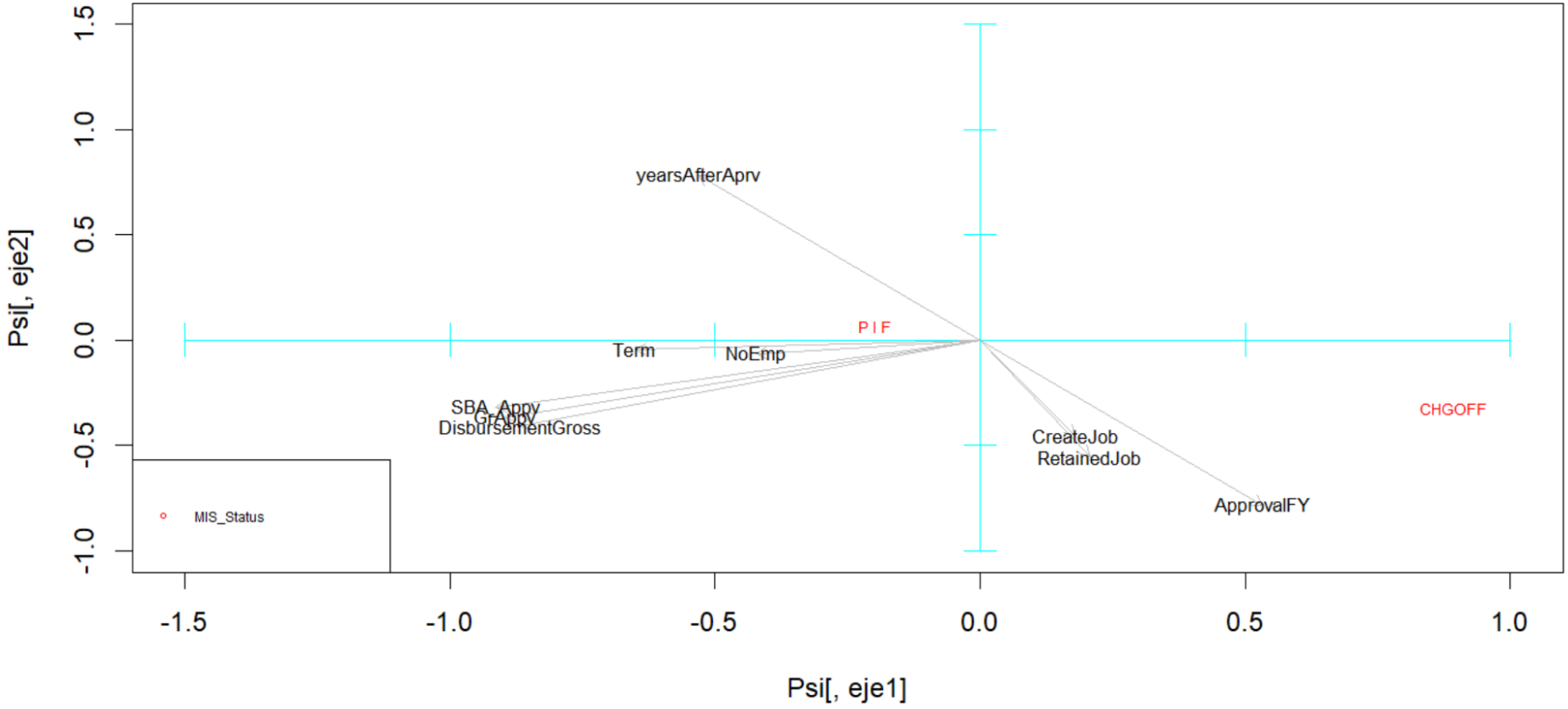


Image 113: Correlation between MIS_Status variable with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

WhichCompany

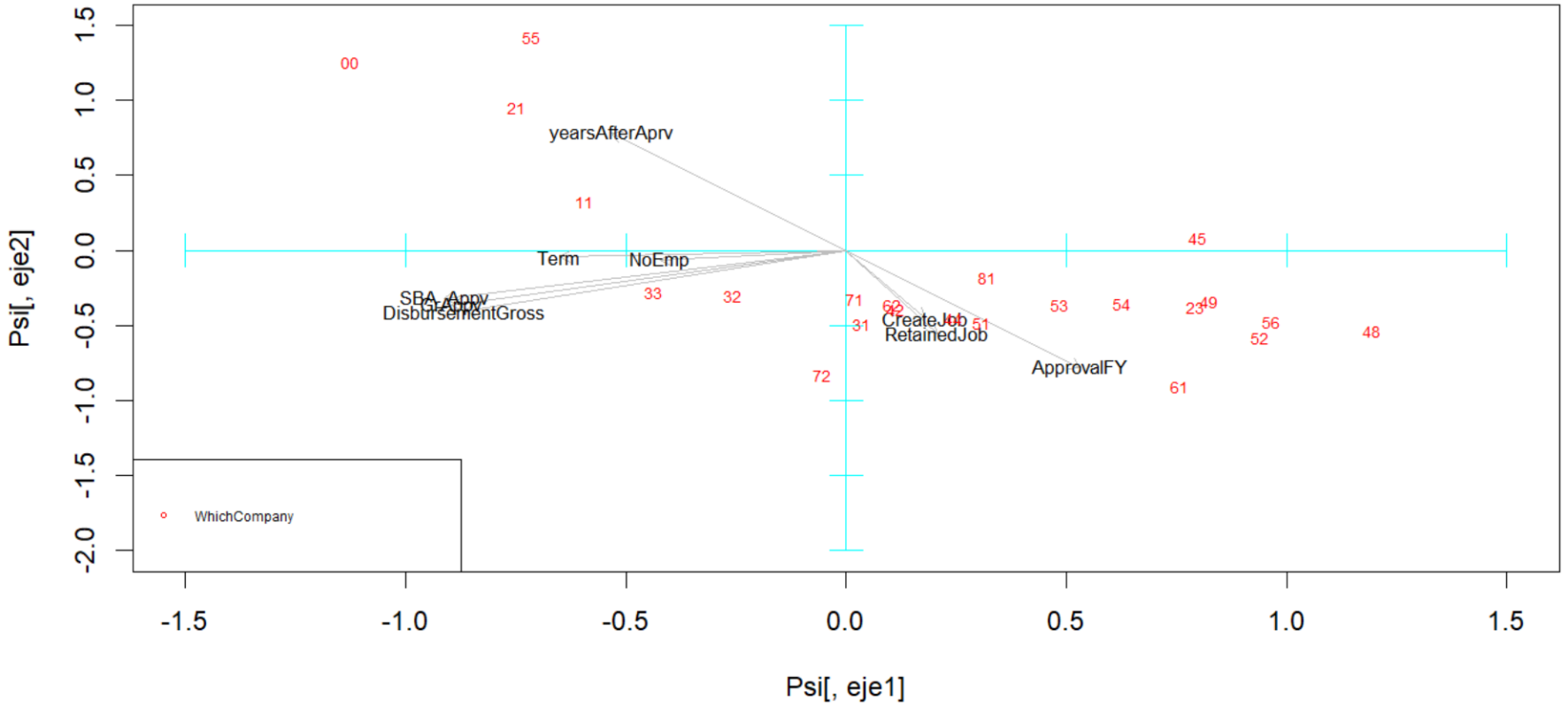


Image 114: Correlation between WhichCompany variable with the numerical variables of the dataset

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Correlation between all the variables

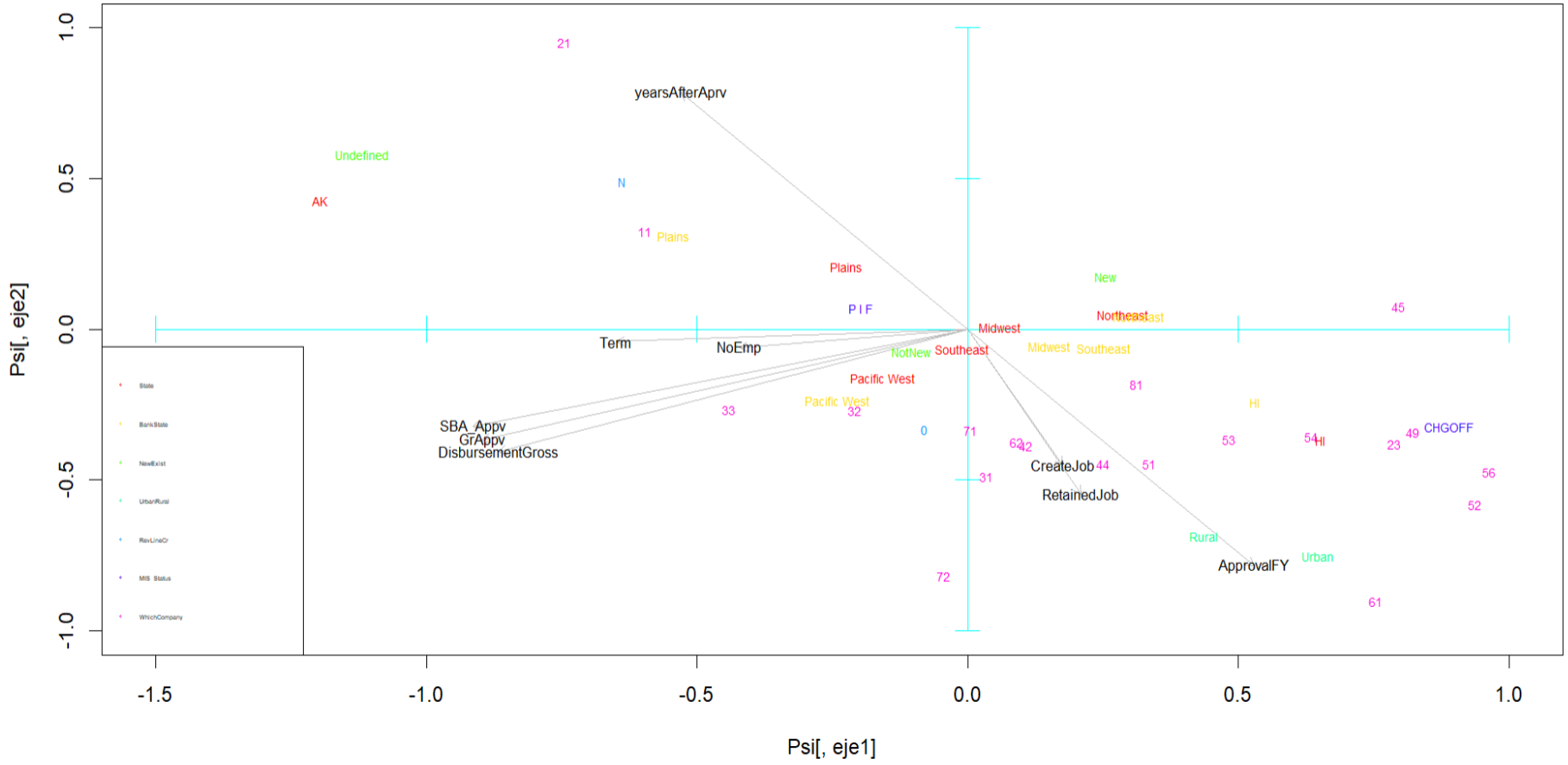


Image 115: Correlation between all the variables of the datase

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Clustering

Precise description of the data

We decided to include all of our numerical variables in this analysis. Our main goal here is to determine in which state a business has the highest probability of getting a loan.

Clustering method and metrics

We used the hierarchical method with Gower metrics because we have a mixed clustering with categorical and numerical variables. We also used the Kmeans method of clustering. We then compared the quality of the clusters produced by the 2 methods using the Calinski-Harabasz index.

Dendrogram with all variables and observations

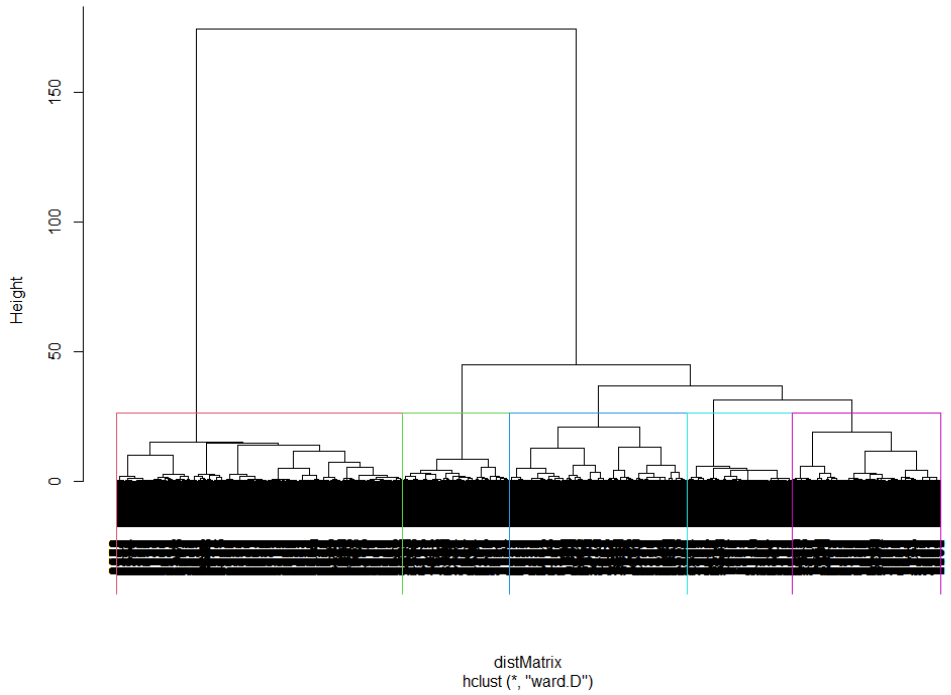


Image 116: Cluster dendrogram

Clusters				
1	2	3	4	5
1733	1077	638	644	895

As we can see in the table above, we obtained a pretty unbalanced cluster size.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Dendrogram with all variables and a random sample of 50 observations

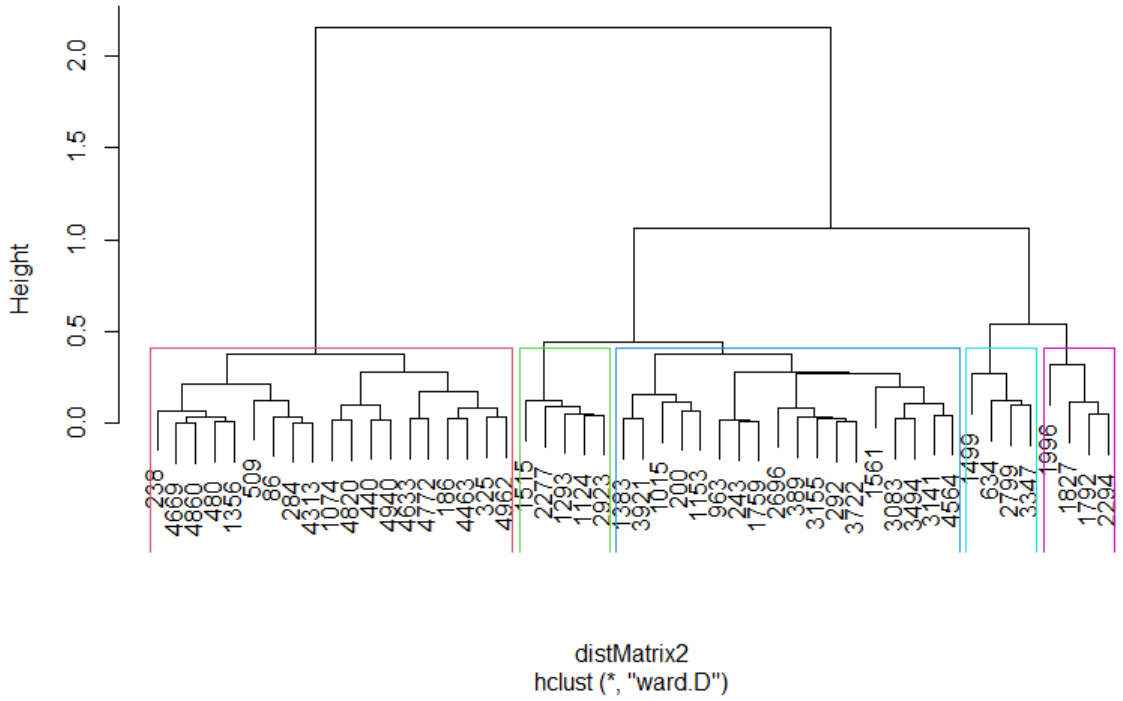


Image 117: Cluster dendrogram

Clusters				
1	2	3	4	5
19	18	4	5	4

As we can see in the table above, we obtained a pretty unbalanced cluster size even with a sample of the dataset. Calinski-Harabasz index for estimating the quality of the clusters = 122.26

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

KMeans cluster for only numeric variables and all observations

To begin the kmean clustering, we performed a series of iterations with the kmeans with different k values to evaluate the optimal number of clusters using the elbow method.

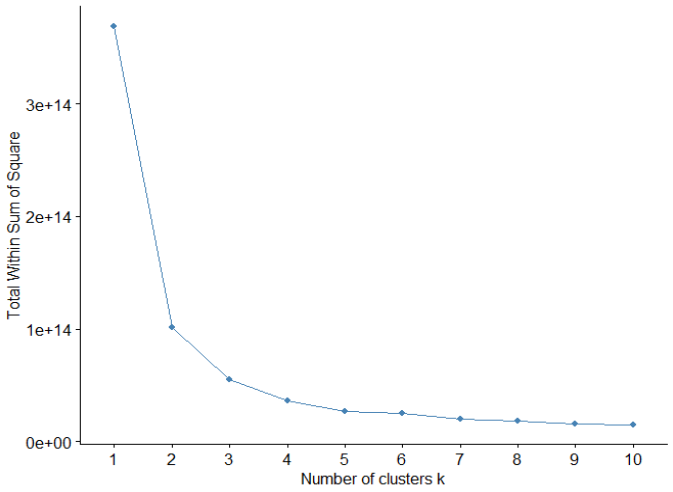


Image 118: Optimal number of clusters

The results suggest that 4 is the optimal number of clusters as it appears to be the bend in the knee (or elbow).

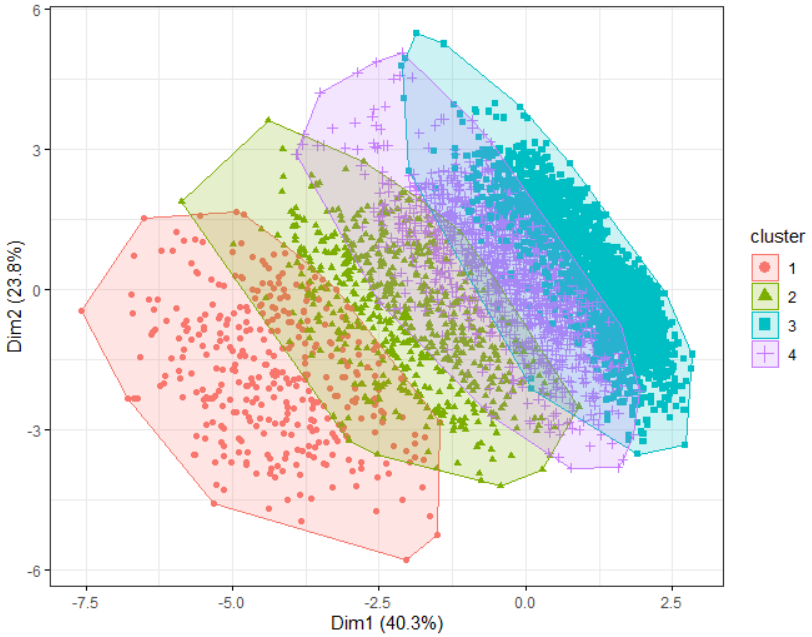


Image 119: Cluster plot

Clusters			
1	2	3	4
380	640	2685	1282

As we can see in the table above, we obtained a pretty unbalanced cluster size even with a sample of only numeric variables.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

KMeans cluster for only numeric variables and a random sample of 50 observations

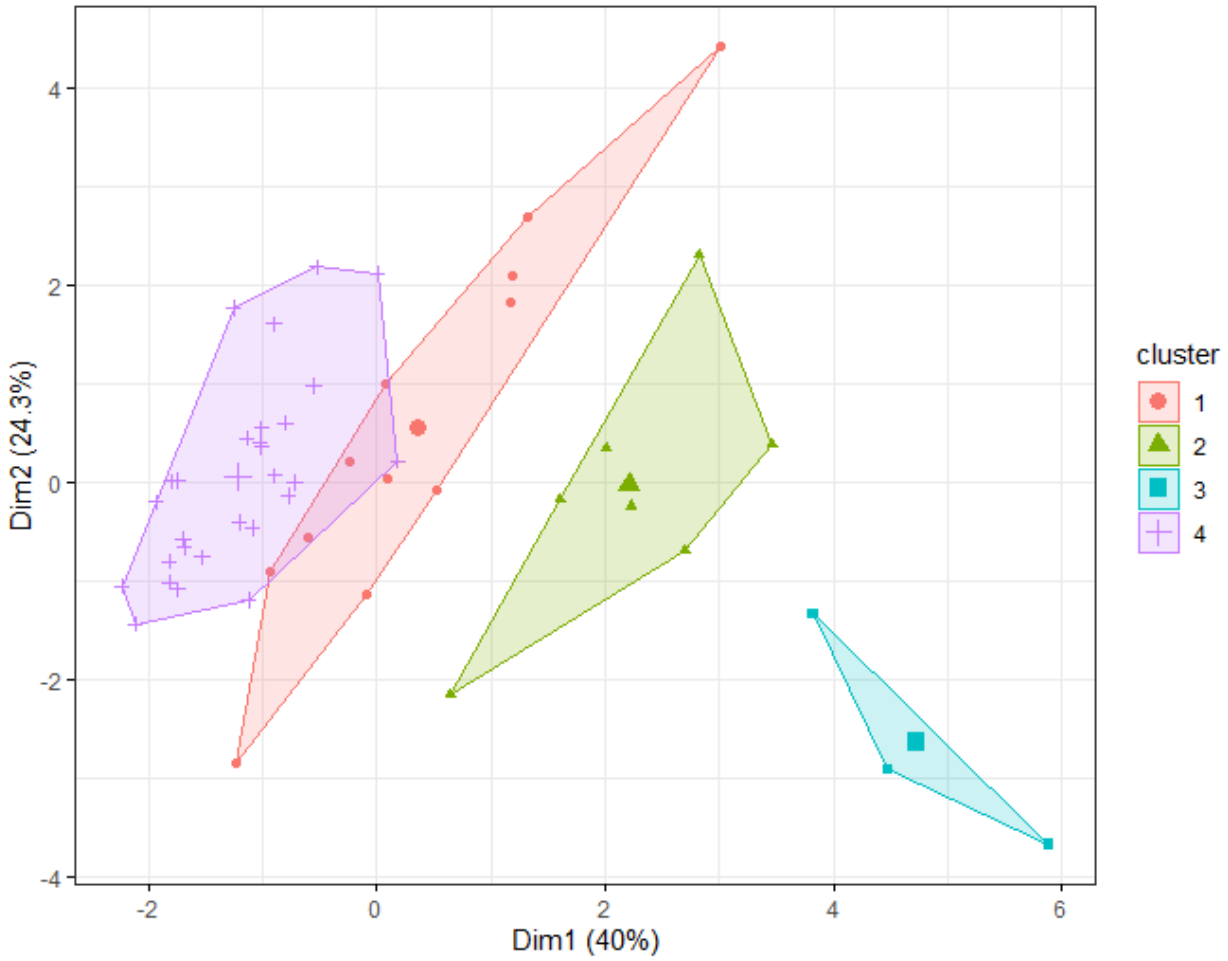


Image 120: Cluster plot

Calinski-Harabasz index for estimating the quality of the clusters = 117.41

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

Profiling of clusters

Data profiling, or profiling, is a statistical method that consists in examining the data comprehend in an information source and analyzing and collecting the relationships between them, their structure and the information each variable provides. Before jumping into the profiling section, we have treated the data set with the clustering method, for this reason, now we have all the individuals of the data grouped in different clusters. There's a new variable, named cluster, which indicates to which group each of the individuals belongs, we have a sum of 5 different clusters.

For this part of the assignment, we will be working with all the variables in the dataset and we are going to observe and analyze how each of them is related to the cluster variable that, as we said, we generated in the clustering process. Focusing on the process of treating the data, we are going to divide it into 2 sections, the treatment of the numerical variables and the treatment of the qualitative variables. For the numerical ones, we calculate the test values for all modalities of the P factor with the script we see in Image 121. For the qualitative ones, using the script in Image 122, we make a similar procedure as the numerical variables to obtain the value test from each variable.

```
ValorTestXnum <- function(Xnum,P){
  #freq dis of fac
  nk <- as.vector(table(P));
  n <- sum(nk);
  #mitjanes x grups
  xk <- tapply(Xnum,P,mean);
  #valors test
  txk <- (xk-mean(Xnum))/(sd(Xnum)*sqrt((n-nk)/(n*nk)));
  #p-values
  pxk <- pt(txk,n-1,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){if (pxk[c]>0.5){pxk[c]<-1-pxk[c]}}
  return (pxk)
}
```

Image 121: Script to obtain the value test from the numerical variables

```
ValorTestXquali <- function(P,Xquali){
  taula <- table(P,Xquali);
  n <- sum(taula);
  pk <- apply(taula,1,sum)/n;
  pj <- apply(taula,2,sum)/n;
  pf <- taula/(n*pk);
  pjm <- matrix(data=pj,nrow=dim(pf)[1],ncol=dim(pf)[2], byrow=TRUE);
  dpf <- pf - pjm;
  dvt <- sqrt(((1-pk)/(n*pk))%*%t(pj*(1-pj)));
  #i hi ha divisions iguals a 0 dona NA i no funciona
  zkj <- dpf
  zkj[dpf!=0]<-dpf[dpf!=0]/dvt[dpf!=0];
  pzkj <- pnorm(zkj,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){for (s in 1:length(levels(Xquali)))
    {if (pzkj[c,s]> 0.5){pzkj[c,s]<-1- pzkj[c,s]}}}
  return (list(rowpf=pf,vtest=zkj,pval=pzkj))
}
```

Image 122: Script to obtain the value test from the qualitative variables

State

The first variable we are going to observe and analyze its relationship with the clusters, is the *State* one, which represents in which state an individual is located. The first plot we generated is the one we see in Image 123, which explains to us how the different states are distributed in all the 5 clusters. For example, we can see that most of the companies located in *NorthEast* states, are grouped in the cluster number 3. And the majority of *AK* ones are grouped in the first cluster.

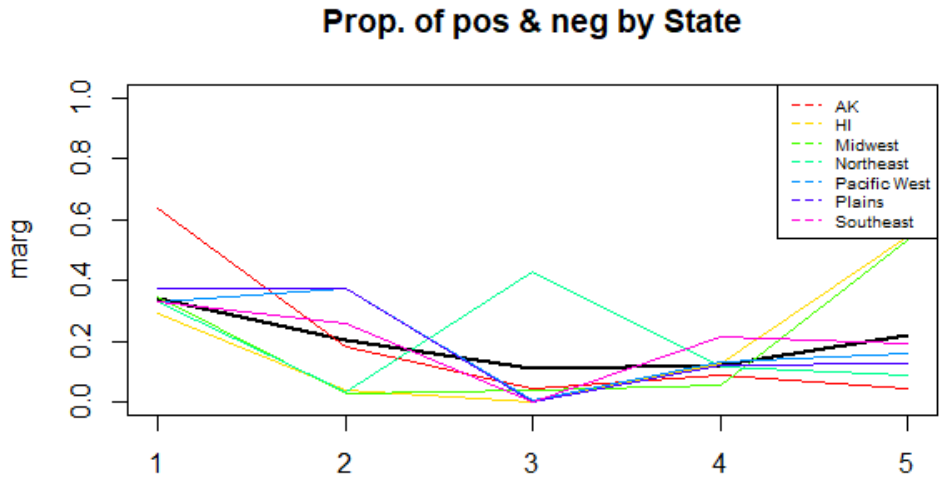


Image 123: Proportion of distribution of the State variable in the different clusters

Even so, the last graph only tells us the proportion of individuals but not the exact quantity. That's why we also generated the graph we see in Image 124, which represents the amount of individuals of each state in all the clusters. Referring to the examples described above, we can see, for example, that even though most *AK* companies are in the first cluster, there should be no more than 10 or 20 individuals in all the data sets. On the other hand, there a a lot of occurrences of companies located in *MidWest* and *NorthEast*, and those are clearly differentiated in different clusters.

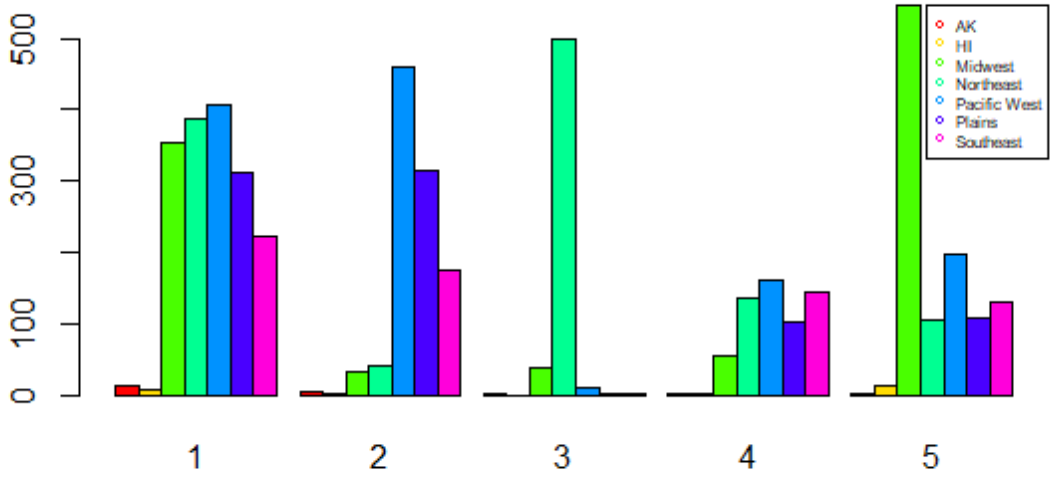


Image 124: Quantity of individuals with an specific State value in each cluster

BankState

The next variable we are going to observe and analyze its relationship with the clusters, is the *BankState* one, which represents in which state is located the bank whom companies ask their loan to. The first plot we generated is the one we see in Image 125, which explains to us how the different bank states are distributed in all the 5 clusters. In a similar way that before, we can see that *NorthEast* banks are also grouped in the third cluster, the *AK* ones in the first one and the *MidWist* ones in the fifth cluster. Adiotionaly, we can observe that the *Undefined* bank states are also grouped in the first cluster.

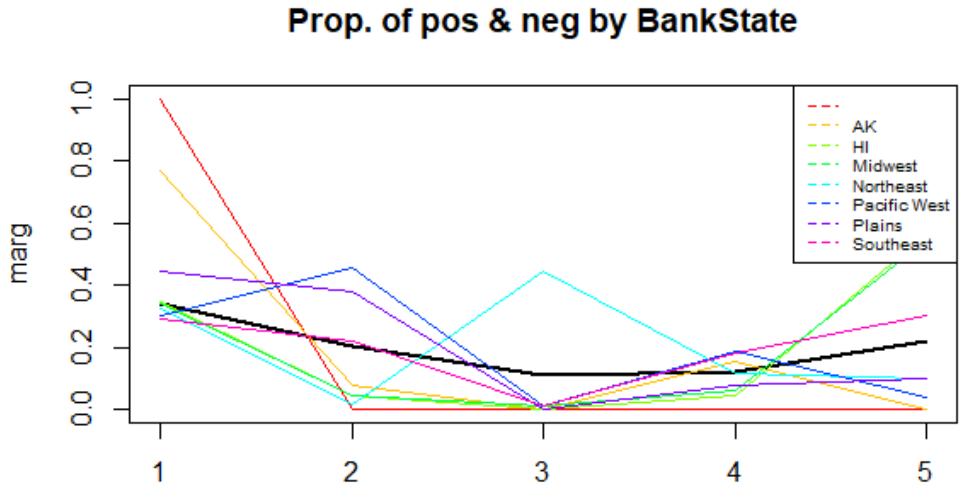


Image 125: Proportion of distribution of the BankState variable in the different clusters

Equally as how we did with the *State* variable, we have also generated a graph which represents the amount of individuals of each bank state in all the clusters, the one we see in Image 126. In the exact same way it happens with the *State* variable, we have few occurrences of *AK*, even so, they are grouped in the first cluster, same happens with the undefined ones. Furthermore, we can observe that the majority of *PacificWest* individuals are grouped in the second cluster. Analyzing together this variable with the *State* one, we can conclude that there exists a positive correlation between them.

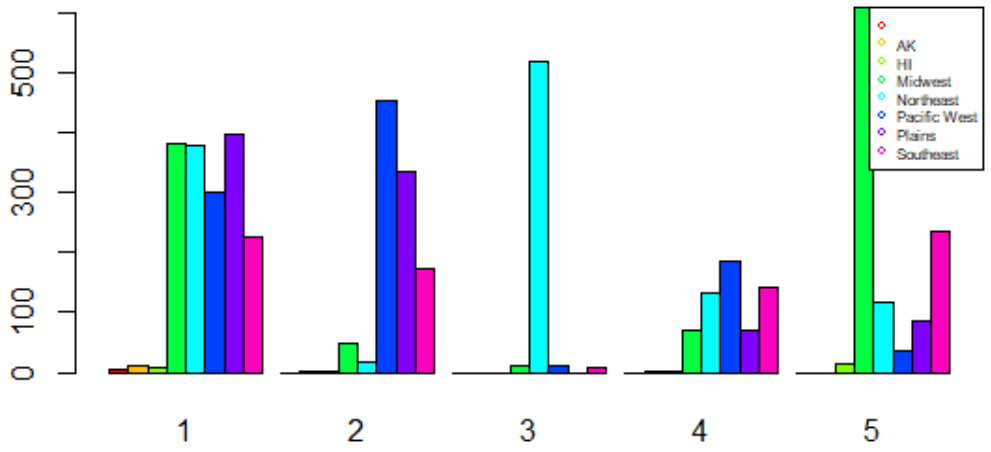


Image 126: Quantity of individuals with an specific BankState value in each cluster

ApprovalFY

The next variable we are going to observe and analyze its relationship with the cluster, is the *ApprovalFY* one, which represents the year when the loan was approved by the bank. With this variable we can see that, unlike was seen previously, a rare behavior appears in the distribution between the clusters. As we see in Image 127, the distribution of *ApprovalFY* doesn't seem to follow any particular pattern, the means of this variable of all the clusters have almost an identical value. After analyzing the graph seen in Image 127, we can conclude that this variable is not a decisive one when deciding the distribution of the individuals in the clusters.

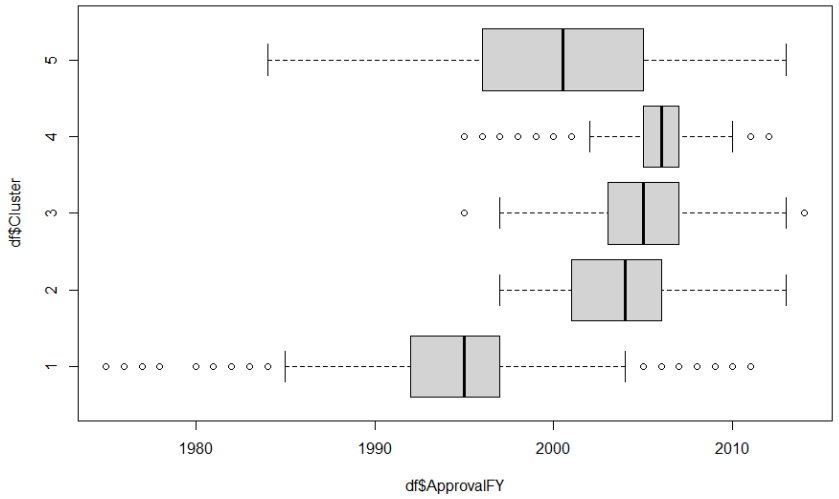


Imagen 127: Boxplot of ApprovalFY by Cluster

Term

The next variable we are going to observe and analyze is the relationship between the cluster and variable *Term*, which represents the length of time (months) it will take for a loan to be paid off. The image 128 shows the boxplot generated, we can observe the clusters are overlapped in the range between 50 and 100, we can also observe that all the median of clusters is approximately 80 except cluster.

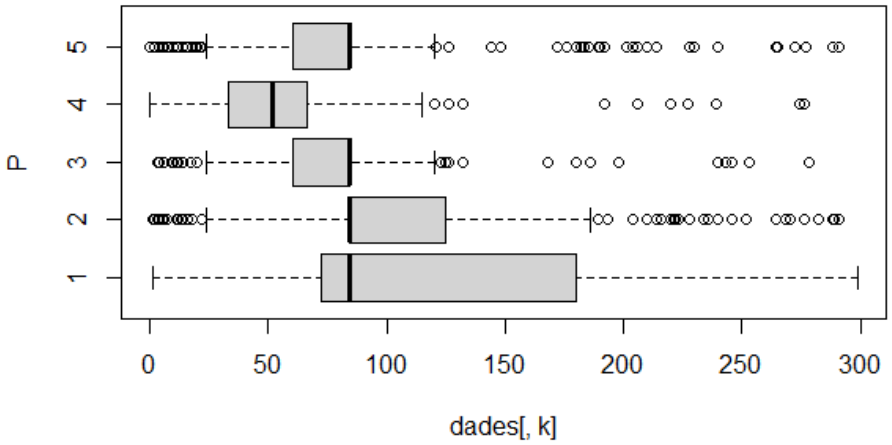


Image 128: Boxplot of Term by Cluster

In the Image 129, explains the means of *Term* in all 5 clusters. We can also see the means of cluster 1 and cluster 2, cluster 3 and 4 are similar, which we can deduce that it is quite distributed for all clusters. After analyzing the plot seen in Image 128 and the graph seen in Image I, We can observe that the companies in the cluster 1 and 2 have more time to return the loan, and the companies in the cluster 4 less.

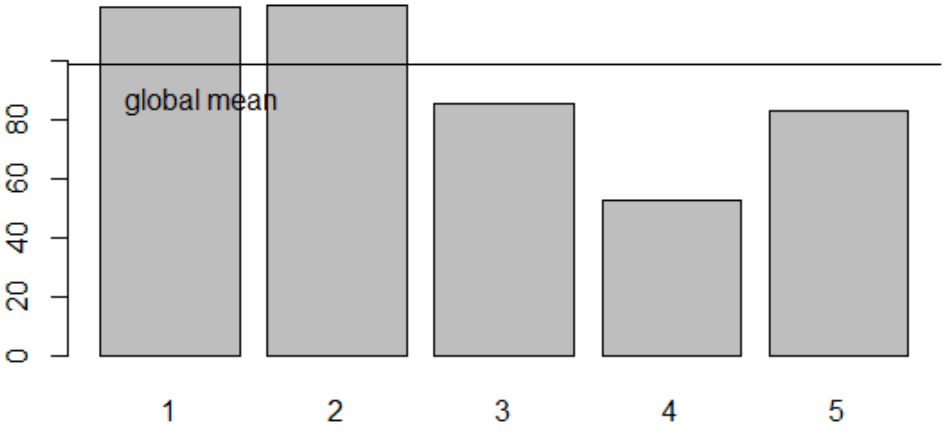


Image 129: Means of Term by Cluster

NoEmp

The next variable we are going to observe and analyze is the relationship between the cluster and variable *NoEmp*, which represents the number of employees of the business. In image 130 we can see the means of *NoEmp* in all 5 clusters. We can also see the means of cluster 3, 4 and 5 are similar, while in clusters 1 and 2 is above the global mean. This means that the companies in clusters 1 and 2 have more employees who could indicate that they are bigger.

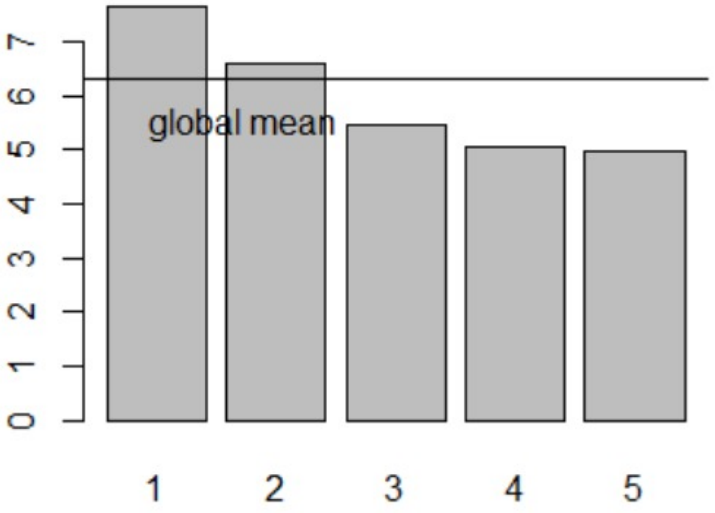


Image 130: Means of NoEmp by Cluster

NewExist

The next variable we are going to observe and analyze its relationship with the clusters, is the *NewExist* one, which represents if the company is a new business or it's an already created one. In order to study the relation we generated the graph we see in Image 131, which explains to us how the *NewExist* variable is distributed in all the 5 clusters. Between the *NotNew* and the *New* one, we can not conclude anything because it seems that they are equally distributed among the clusters. Still, we can observe that those companies with an undefined value in the variable, are mostly grouped in the first of the clusters.

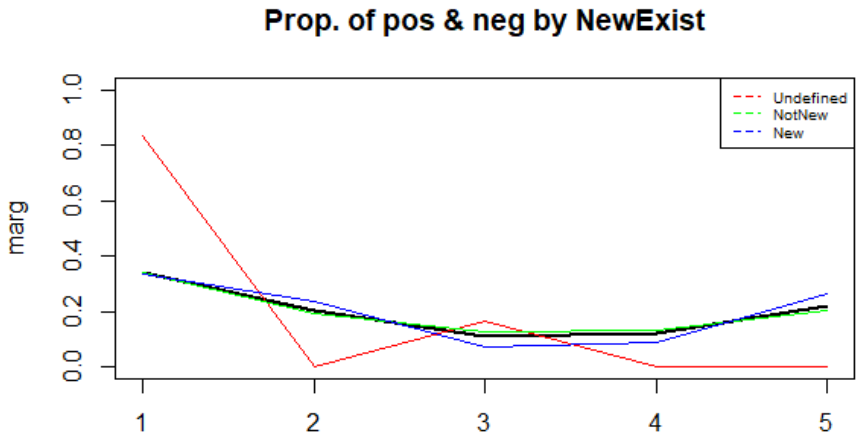


Image 131: Proportion of distribution of the NewExist variable in the different clusters

CreateJob

The next variable we are going to observe and analyze its relationship with the clusters, is the *CreateJob* one, which represents the number of new jobs a business is going to offer. In order to study the relation between the variable and the clusters, we generated the graph we see in Image 132, which explains the means of *CreateJob* in all 5 clusters. After studying and analyzing the graph, we can see that the means of the cluster 2, 3, 4 and 5 are quite similar, which can mean that they are somehow equally distributed, in relation to the *CreateJob* variable. Regarding the first cluster, it's visible that it's means is smaller than the other clusters, the information we can extract from this success is that the first cluster contains all the companies that provide a small number of new job positions.

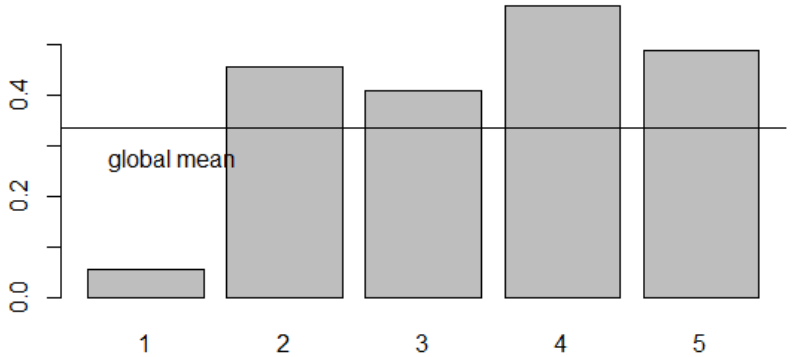


Image 132: Means of CreateJob variable by Cluster

RetainedJob

The next variable we are going to observe and analyze its relationship with the clusters, is the *RetainedJob* one, which represents the number of retained jobs a business has. In order to study the relation between the variable and the clusters, we generated the graph we see in Image 133, which explains the means of *RetainedJob* in all 5 clusters. After studying and analyzing the graph, we can see the same pattern we saw with the NewJob, where the lower mean is held in the first cluster. The conclusion that we can extract is that both variables are related when it comes to being distributed in the clusters. Furthermore, we can see how the higher mean is held in the forth cluster.

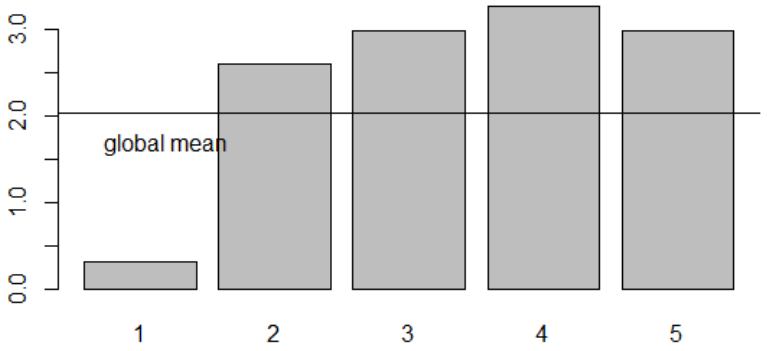


Imagen 133: Means of RetainedJob variable by Cluster

UrbanRural

The next variable we are going to observe and analyze its relationship with the clusters, is the *UrbanRural* one, which represents if the company is an urban or a rural one. In order to study the relation we generated the graph we see in Image 134, which explains to us how the *UrbanRural* variable is distributed in all the 5 clusters. The first thing we notice is a pattern that we had already seen in the variable *NotNew*, the individuals with an undefined value in the variable we are analyzing, are mostly grouped in the first cluster. Beyond this, we notice that the rural companies are generally held in the second cluster and the rural ones in the fifth cluster.

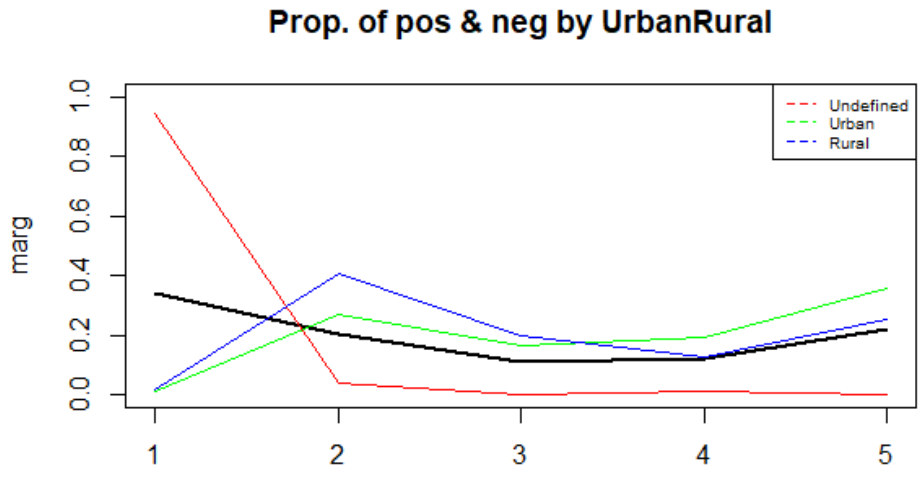


Image 134: Proportion of distribution of the NewExist variable in the different clusters

For this variable, we also generated a graph which represents how the possible values of the variable are distributed in all the clusters we have, this graph is the one we see in Image 135. We see how almost all the undefined values are held in the first cluster, as we explained in the PCA section, there are a lot of occurrences of undefined values probably due to the recent implementation of the variable. Following the information we extract from the graph in Image AN, we can see how between the *Urban* occurrences most of them are in the last cluster. Referring to the *Rural* apparitions, we can observe how they are distributed, mostly, in the second group.

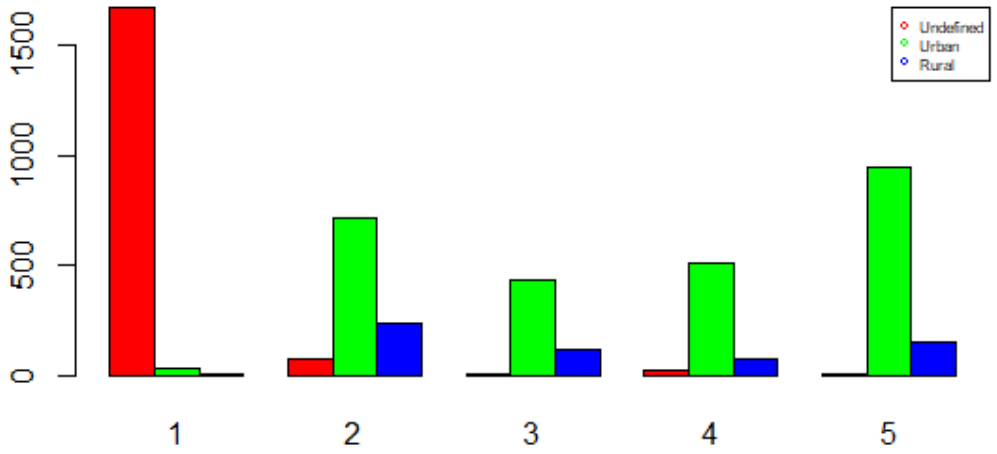


Image 135: Quantity of individuals with an specific UrbanRural value in each cluster

RevLineCr

The next variable we are going to observe and analyze its relationship with the clusters is the *RevLineCr*, which indicates if a line of credit is or not revolving. For this variable we generate a graph which represents how the possible values of the variable are distributed in all the clusters we have, this graph is the one we see in Image 136. We can highlight that the major part of the companies in cluster 1 doesn't have a revolving line of credit and the opposite in cluster 5.

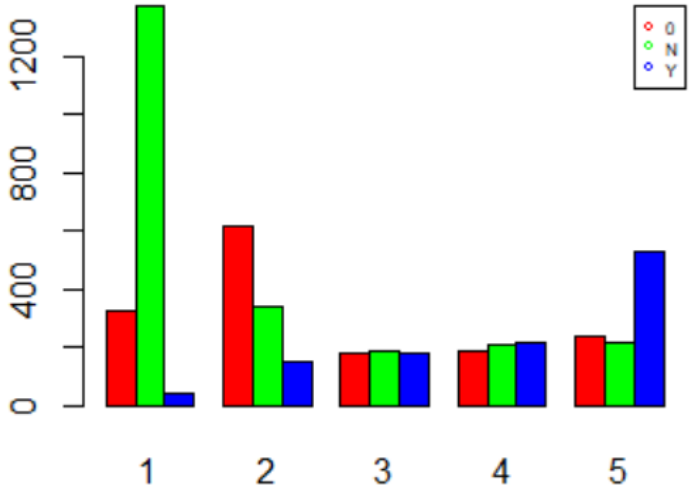


Image 136: Quantity of individuals with an specific RevLineCr value in each cluster

DisbursementGross

The next variable we are going to observe and analyze its relationship with the clusters, is the *DisbursementGross* one, which represents the amount of money disbursed. In order to study the relation between the variable and the clusters, we generated the graph we see in Image 137, which represents the means of *DisbursementGross* distribution in all 5 clusters. After studying and analyzing the graph, what we can conclude is that those businesses that were with a more amount of money disbursed are grouped mostly in the second cluster, and those with a less amount of money disbursed are located in the third and fourth cluster.

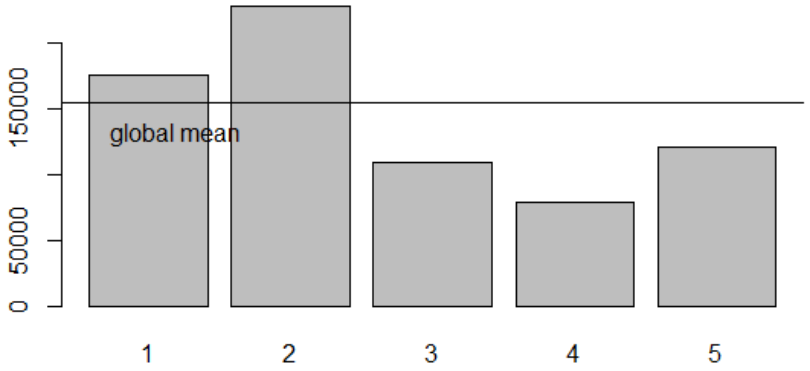


Image 137: Means of DisbursementGross variable by Cluster

MIS_Status

The next variable we are going to observe and analyze its relationship with the clusters is the *MIS_Status* one, which represents the type of loan a bank gives to a certain company. In order to study the relation we generated the graph we see in Image 138, which lets us see how the *MIS_Status* variable is distributed in all the 5 clusters. The first thing we notice is a pattern that stands out clearly on the graph, the majority of companies with a type of loan *CHGOFF*, are situated together in the fourth cluster. Furthermore, this cluster is the one with less occurrences of *PIF* loans type. Moreover, it seems that most of the *PIF* values are held in the first cluster.

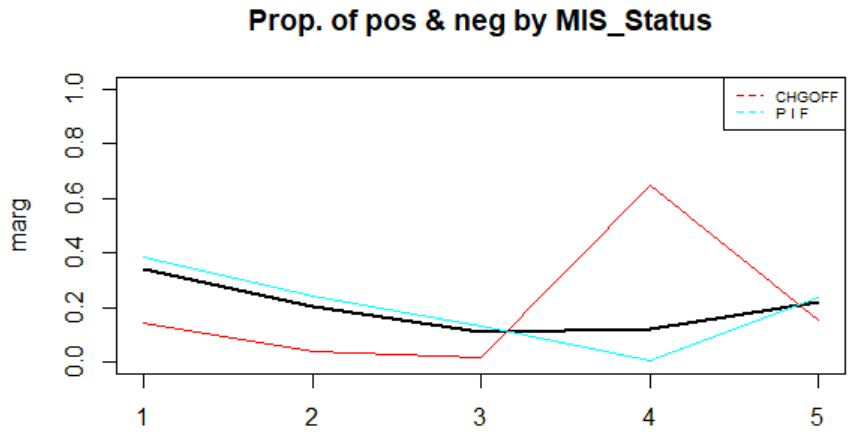


Image 138: Proportion of distribution of the MIS_Status variable in the different clusters

For this variable, we also generated a graph which represents how the possible values of the variable are distributed in all the clusters we have, this graph is the one we see in Image 139. As we saw in the last graph, we can confirm that most of *CHGOFF* loans are held in the fourth cluster and, despite not having such a clear differentiation, most of the *PIF* loans are grouped in the first cluster.

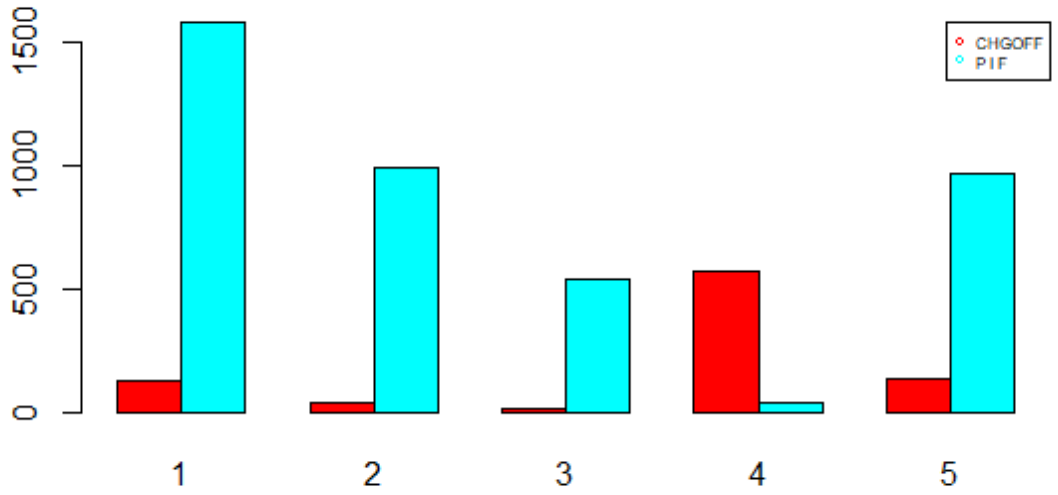


Image 139: Quantity of individuals with an specific MIS_Status value in each cluster

GrAppv

The next variable we are going to observe and analyze its relationship with the clusters, is the *GrAppv* one, which represents the gross amount of loan approved by the bank. In order to study the relation between the variable and the clusters, we generated the graph we see in Image 140, which represents the means of *GrAppv* distribution in all 5 clusters. After studying and analyzing the graph, what we can conclude is that those businesses that were paid more are grouped mostly in the second cluster, and those who were paid less are located in the third and fourth cluster. If we compare this conclusion with the *DisbursementGross* ones, we can see that they are equally distributed, which means that those two variables have a positive correlation.

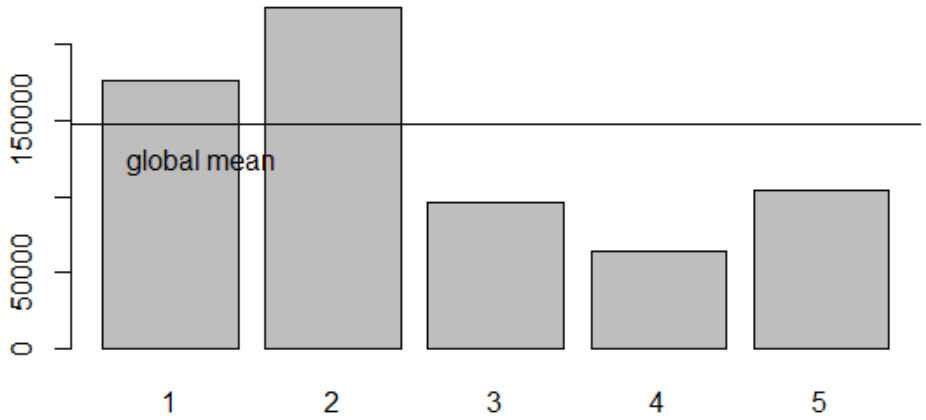


Image 140: Means of GrAppv variable by Cluster

SBA_Appv

The next variable we are going to analyze is the *SBA_Appv* one, which represents the SBA's guaranteed amount of approved loan. In order to study the relation between the variable and the clusters, we generated the graph we see in Image 141, which represents the means of *GrAppv* distribution in all 5 clusters. After studying and analyzing the graph, we can conclude that this variable is positively correlated with the *GrAppv* and the *DisbursementGross* ones. The three graphs represent exactly the same distribution, where the higher mean is the second cluster and the lower one are the third and the fourth ones.

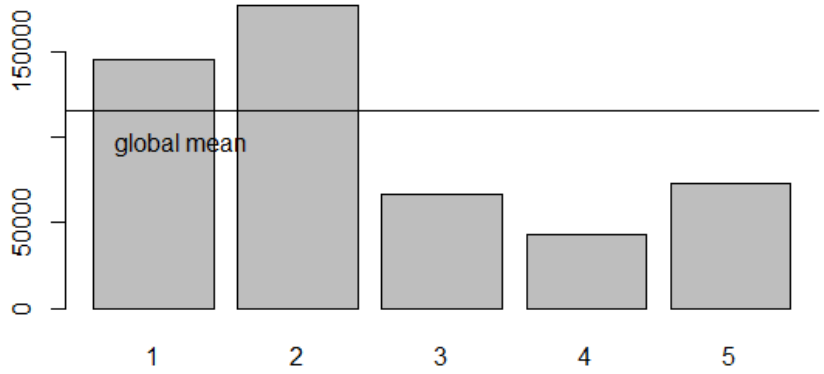


Image 141: Means of SBA_Appv variable by Cluster

WhichCompany

The next variable we are going to observe and analyze its relationship with the clusters is the *WhichCompany* one, which represents what type of company is one individual. In order to study the relation we generated the graph we see in Image 142. About this graph we can observe one pattern that has already emerged in some other variables, the undefined ones, or *00*, are mostly situated in the first cluster. About the other types of companies we can not conclude any specific pattern, because there are not any types that distinguish amongst the others. We could say that all the companies are equally distributed among all the five clusters that define our data set.

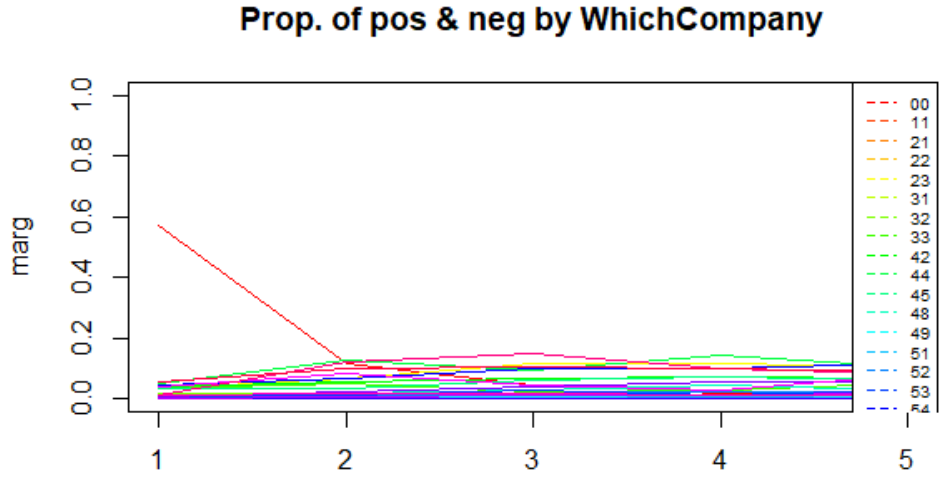


Image 142: Proportion of distribution of the WhichCompany variable in the different clusters

yearsAfterAprv

The last variable we are going to observe and analyze its relationship with the clusters, is the *yearsAfterAprv* one, which represents how many years have passed since the loan was accepted. In order to study the relation we generated the graph we see in Image 143. After studying and analyzing the graph, we can observe that the only pattern that stands out in this graph is that the oldest companies, the ones with a high value *yearsAfterAprv*, are grouped in the first cluster, that's why it has a higher mean. The rest of the individuals seem to be equally distributed among the clusters.

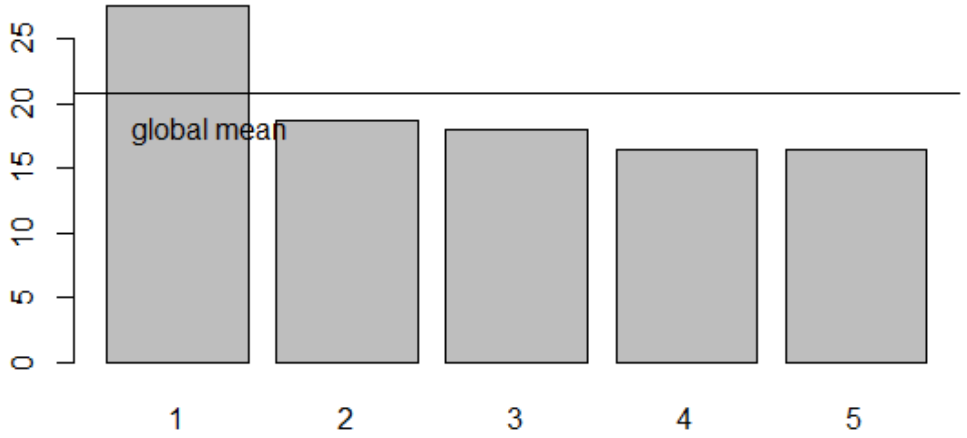


Image 143: Means of yearsAfterAprv variable by Cluster

P-values

After analysis the all clusters for each variable, we can use P-value to check if really exist a correlation between each cluster and each variable, if p-value is more than 0.05 implicate is no relationship between the variable and the indicated cluster. In this case only one independence of cluster 2 with the variable NoEmp, where the p-value obtained is 0.468 more than 0.05, more details see the following image.

```
[1] "P.values per class: 2"
            State         BankState          NewExist        UrbanRural
         0.00e+00          0.00e+00          0.00e+00          0.00e+00
        RevLineCr       MIS_Status      WhichCompany    yearsAfterAprv
         0.00e+00          0.00e+00          0.00e+00          0.00e+00
          Cluster             GrAppv          SBA_Appv DisbursementGross
         0.00e+00          5.14e-61          5.45e-58          7.65e-55
        ApprovalFY              Term      RetainedJob         CreateJob
         5.03e-39          4.89e-29          2.18e-10          7.82e-09
             NoEmp
         4.68e-01
```

Analyzing the association between categorical variables and numerical variables, we can observe an Independence between the variable State and NoEmp. We can conclude that all the numerical variables are closely related for each cluster and categorical variables, excepting the variable NoEmp for the cluster 2 and categorical variable State.

```
> res.catdes$quanti.var  # Global association to numeric variables
                          Eta2      P-value
Cluster           0.065754245 3.809917e-70
SBA_Appv          0.011769190 7.244900e-11
GrAppv            0.009480056 1.518218e-08
Term              0.009241198 2.635461e-08
ApprovalFY        0.007833914 6.592355e-07
yearsAfterAprv    0.007833914 6.592355e-07
DisbursementGross 0.007721335 8.507676e-07
RetainedJob       0.006545950 1.187461e-05
CreateJob         0.002614910 4.233403e-02
NoEmp             0.001792791 1.769979e-01
```

Conclusions

We established different objectives along the duration of the work, like learning about the loans, how different aspects of a company are important when asking for a loan, acquire the knowledge about analyzing and managing large amounts of data.

During the first dates we spent our time learning more about the topic and about each of the variables that were on the data set, what information they gave us and if it was important or not. All this process went along with the descriptive univariate and bivariate and the preprocessing. Once we had our data established and fixed and we knew enough to work and understand the behavior of the variables, we began treating, analyzing and applying statistical methods to our data set, like PCA or clustering.

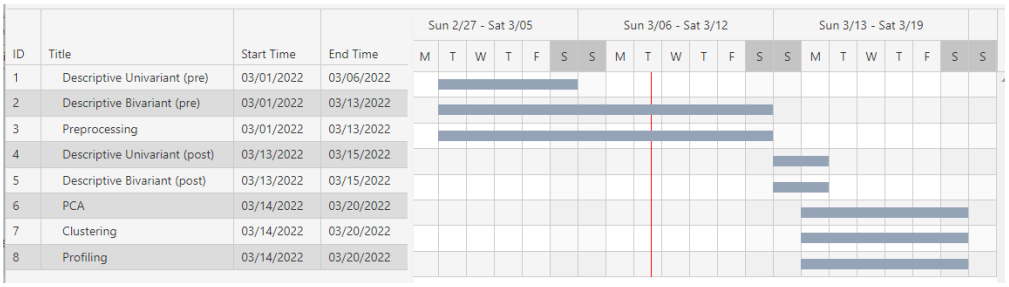
After doing the project, we can assure that now we have tools related to statistics that we did not that even existed before. Like how preprocessing a data set can be the most important task of a data analysis and can be crucial when it comes to getting or not good results or how performing clustering allows us to group individuals and analyze them more accurately.

To compare the results of the PCA and Clustering profiling we looked at all the graphs and data from both analyses. We observed a lot of coincidences with the results of both analyses. For example we observed that in both analyses, the variable Term is directly correlated and explains the variable MIS_Status, with low terms the company tends to fail in its loan. And in contraposition, with high terms the company tends to be more successful. All the variables present in both analyses have the same behavior when profiling the variable MIS_Status. Despite all the similarities, we encountered some divergences. For instance, we observe from the PCA that the locations Urban and Rural are present in the 4th quadrant of the PCA graph, the same quadrant where the CHGOFF companies are. This means that the companies that fail tend to be Urban and Rural and not Undefined. We get a different result from the profiling of clusters where companies that are Urban and Rural are evenly distributed among clusters.

Pere Arnau Alegre
Andrés Jiménez González
Victor Teixidó López
Max Vives Ribera
You Wu

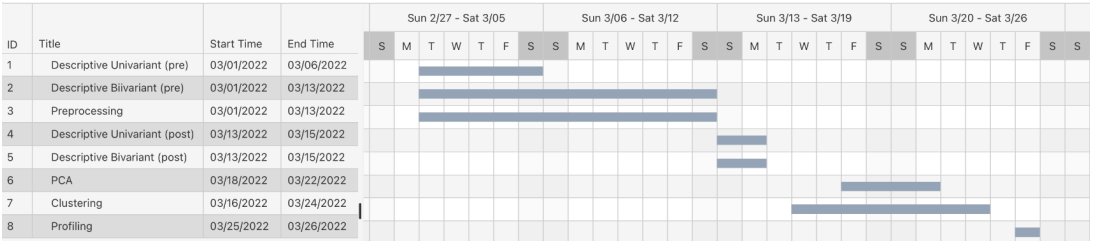
Working plan

Gantt diagrams



ID	Title	Start Time	End Time
1	Descriptive Univariant (pre)	03/01/2022	03/06/2022
2	Descriptive Bivariant (pre)	03/01/2022	03/13/2022
3	Preprocessing	03/01/2022	03/13/2022
4	Descriptive Univariant (post)	03/13/2022	03/15/2022
5	Descriptive Bivariant (post)	03/13/2022	03/15/2022
6	PCA	03/14/2022	03/20/2022
7	Clustering	03/14/2022	03/20/2022
8	Profiling	03/14/2022	03/20/2022

Image 144: Initial Gantt diagram



ID	Title	Start Time	End Time
1	Descriptive Univariant (pre)	03/01/2022	03/06/2022
2	Descriptive Biivariant (pre)	03/01/2022	03/13/2022
3	Preprocessing	03/01/2022	03/13/2022
4	Descriptive Univariant (post)	03/13/2022	03/15/2022
5	Descriptive Bivariant (post)	03/13/2022	03/15/2022
6	PCA	03/18/2022	03/22/2022
7	Clustering	03/16/2022	03/24/2022
8	Profiling	03/25/2022	03/26/2022

Image 145: Final Gantt Diagram

Task Assignments

Nombres	DB	P	DA	PCA	Clustering	Profiling
Max	X		X		X	
Arnau	X		X		X	
Andres		X			X	X
Victor		X		X		X
You		X		X		X

Critical discussion

As we can see in the planned Gantt diagram and the real one, there have been delays especially in the last sections of the project. Probably, due to other courses' existence and that were sections more complicated and laborious than the initial ones. Despite that, we managed to deliver on time all the deliverables that we had to do and more importantly, we achieved the good work and enthusiasm we were seeking at the beginning. Moreover, we had enough time to do all the documentation in a proper and nice way. Fortunately, we have not encountered any unforeseen problems that forced us to change the distribution of the work or similar cases we took in count as risk situations.