DATA MINING: SHOULD THIS LOAN BE APPROVED OR DENIED?

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

INDEX

- WHAT IS OUR DATABASE ABOUT? WHICH IS OUR OBJECTIVE?
- DATA MINING PROCESS SCHEMA
- DESCRIPTIVE ANALYSIS
- UNIVARIATE DESCRIPTIVE ANALYSIS
- ADDITIONAL DESCRIPTIVE ANALYSIS ISSUES
- PREPROCESSING
- PCA SPECIFICATIONS
- FIRST FACTORIAL PLANE FOR PCA
- PCA CONCLUSIONS
- CLUSTERING AND DENDROGRAM
- CLASS INTERPRETATION TOOLS
- WHAT DEFINES OUR CLUSTERS?
- FINAL CLASS PROFILING
- COMPARISON BETWEEN PCA AND CLUSTERING
- CONCLUSIONS
- INITIAL AND FINAL SCHEDULING

WHAT IS OUR DATABASE ABOUT? WHICH IS OUR OBJECTIVE?

Name: “Should This Loan be Approved or Denied?”

Subset of the larger dataset from the U.S. Small Business Administration (SBA).

Important variables: State, NAICS, ApprovalFY, Term, NoEmp, NewExist, CreateJob, DisbursementGross, MIS_Status.

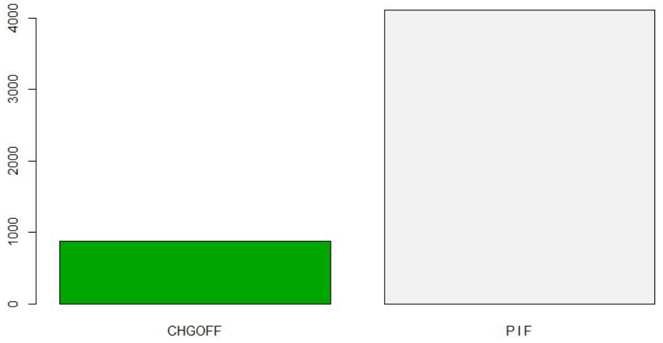
Objective: For this case-study assignment, assume the role of loan officer at a bank and are asked to approve or deny a loan by assessing its risk of default using logistic regression.

DATA MINING PROCESS SCHEMA

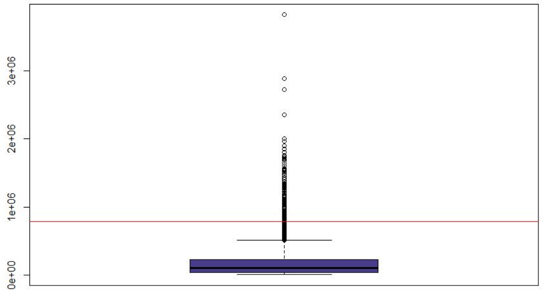
1. Data and metadata analysis
2. Premature data preprocessing
3. Initial univariate & bivariate data description without preprocessing
4. Preprocessing
5. Univariate & bivariate data description
6. PCA
7. Clustering
8. Profiling of clusters

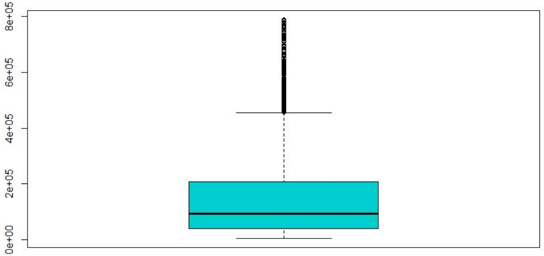
DESCRIPTIVE ANALYSIS

MIS_Status



DisbursementGross





Min.	1st Qu.	Median
0	40000	91713

Mean	3rd Qu.	Max.
154139	204863	790000

UNIVARIATE DESCRIPTIVE ANALYSIS

Categorical

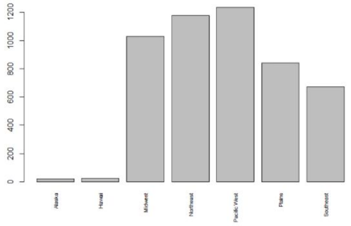


Image 55: Barplot of State

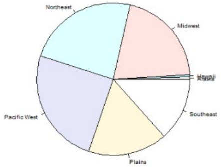


Image 56: Pie of State

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

Quantitative

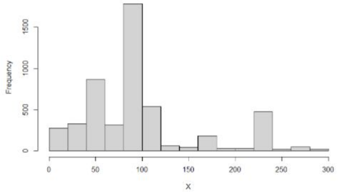


Image 59: Histogram of Term

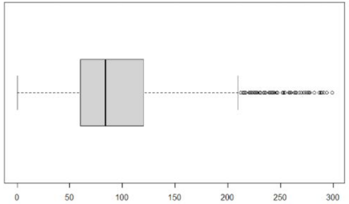


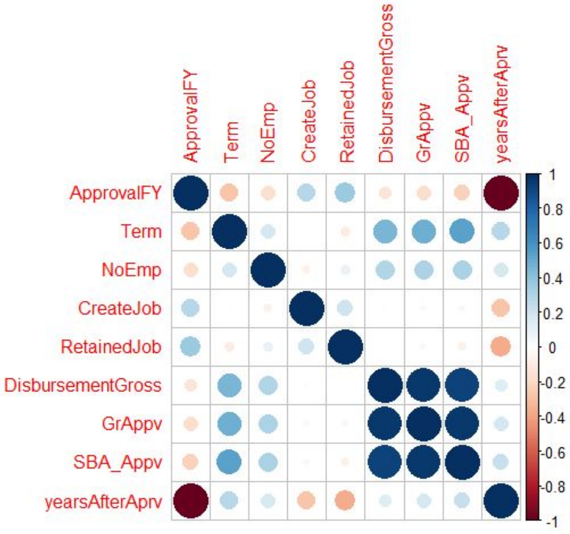
Image 60: Boxplot of Term

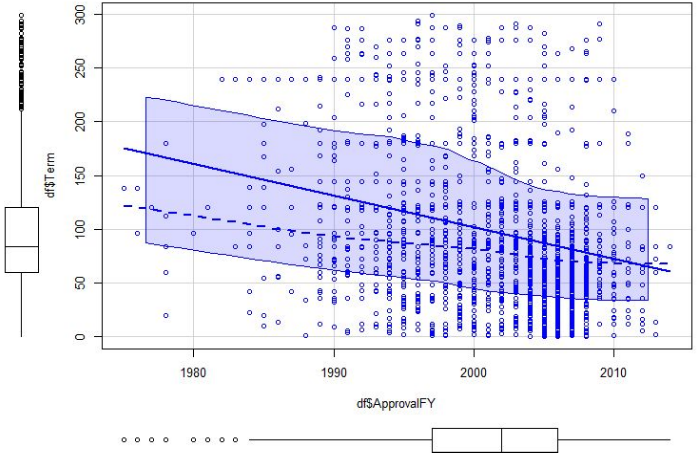
Extended Summary Statistics:

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

"sd: 63.29"
"vc: 0.642"

ADDITIONAL DESCRIPTIVE ANALYSIS ISSUES

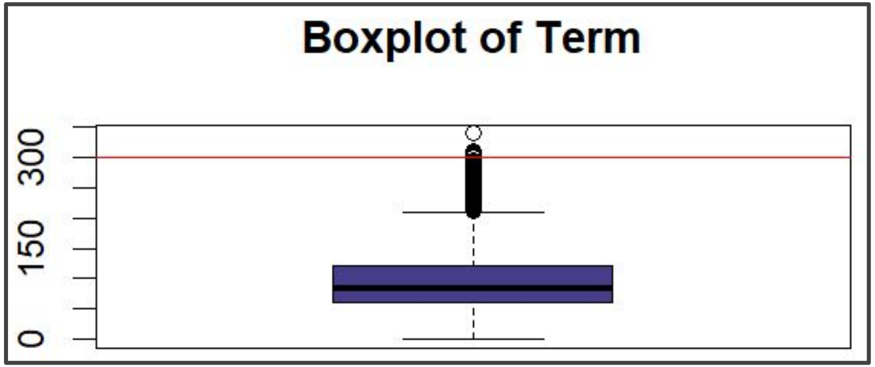




PREPROCESSING

- Elimination of variables: *Name*, *Bank*, *BalanceGross*
- Factorization of numericals
- From [char] to numeric
- Reorganization *State* and *BankState*
- *NAICS* to *WhichCompany*

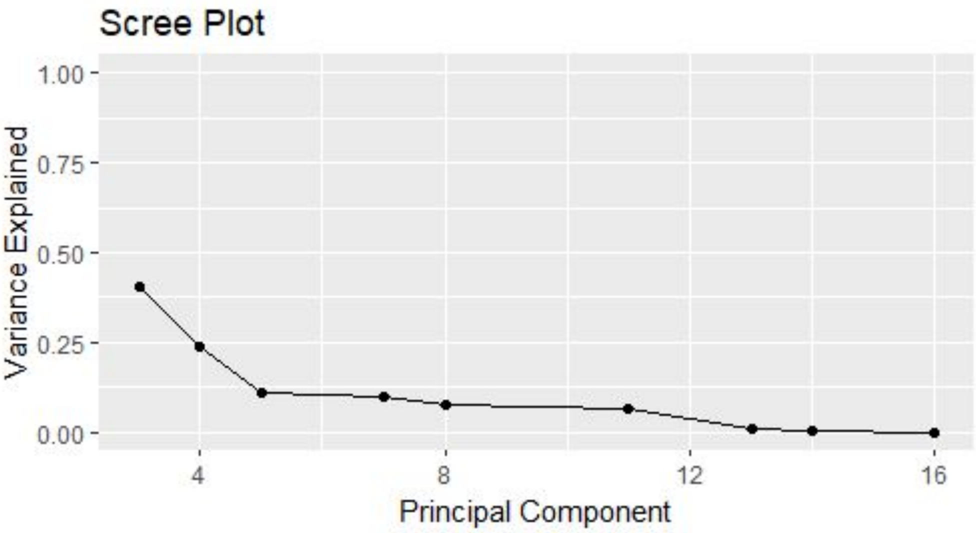
- Finding outliers
- Imputation by Knn



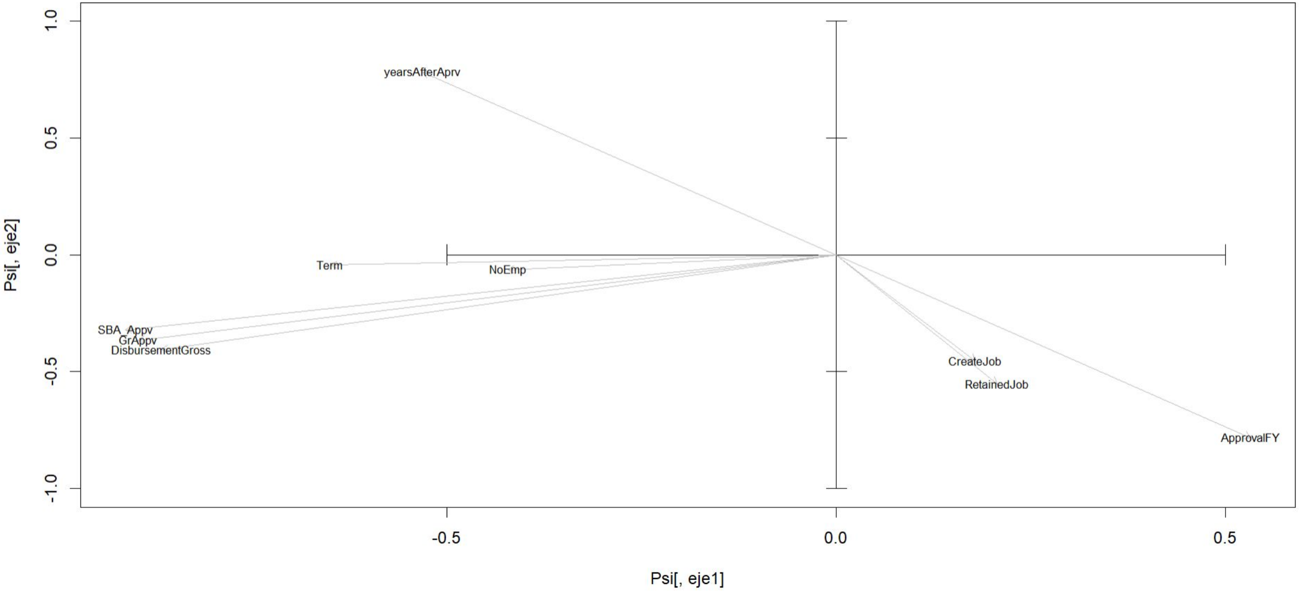
Outliers found in variable Term

PCA SPECIFICATIONS

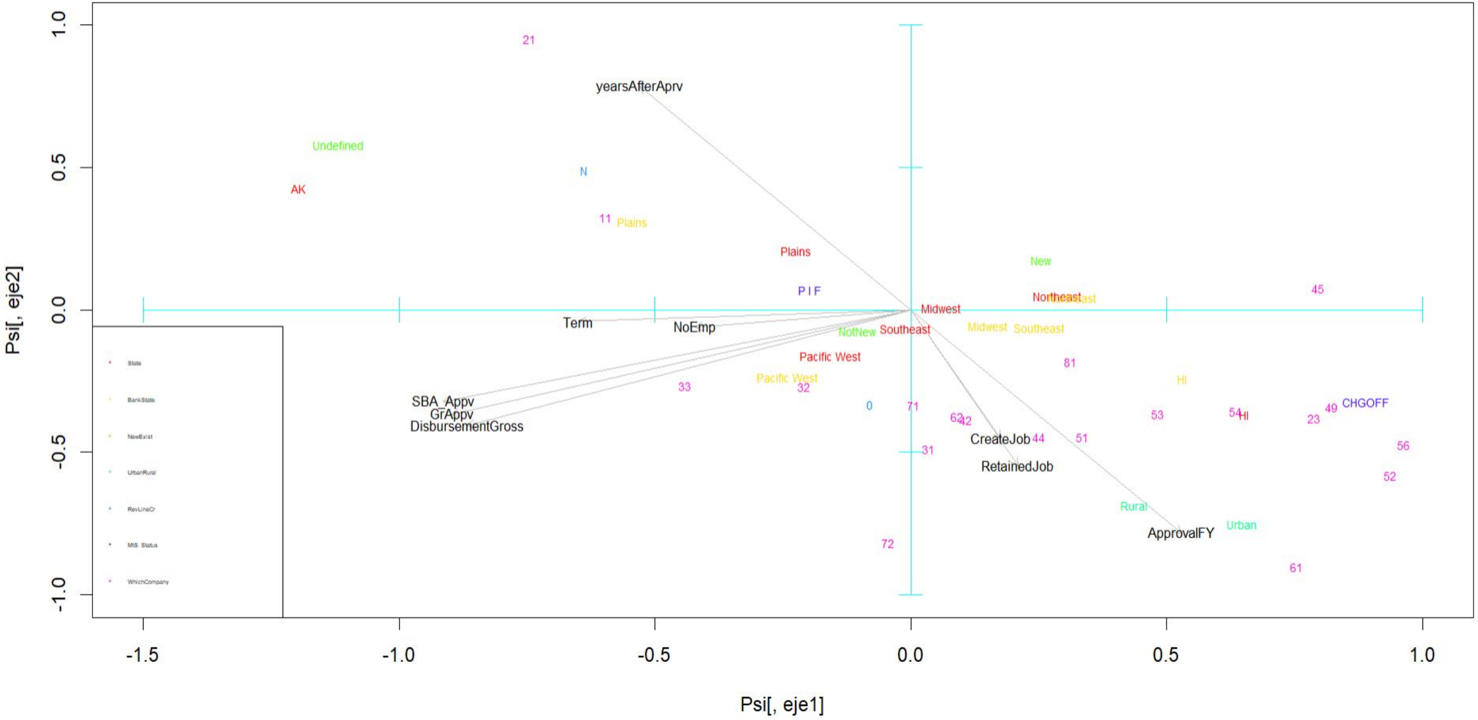
- Principal Component Analysis
- Reduce the dimension of data set.



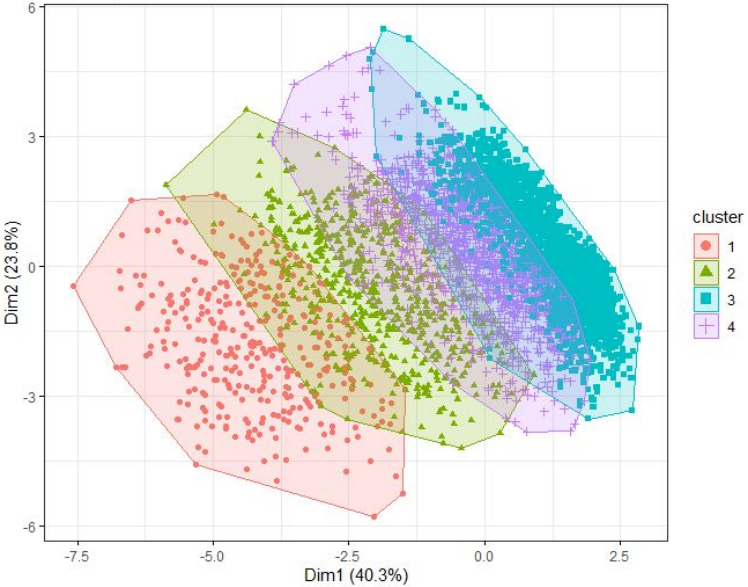
FIRST FACTORIAL PLANE FOR PCA

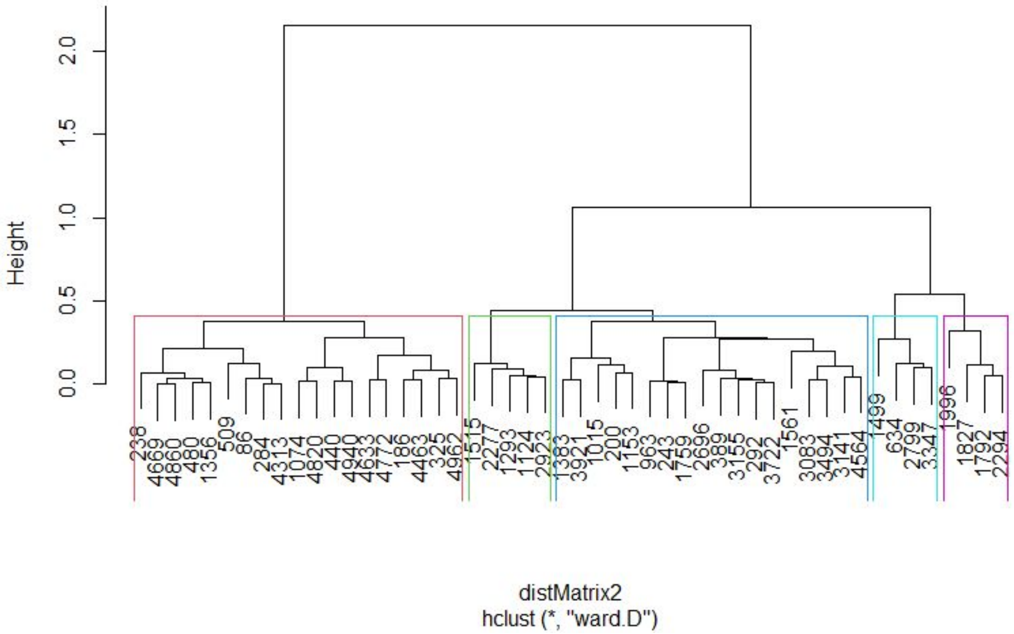


PCA CONCLUSIONS



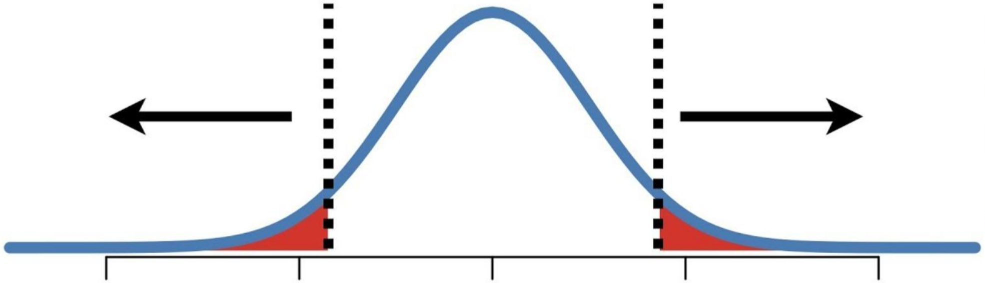
CLUSTERING AND DENDROGRAM



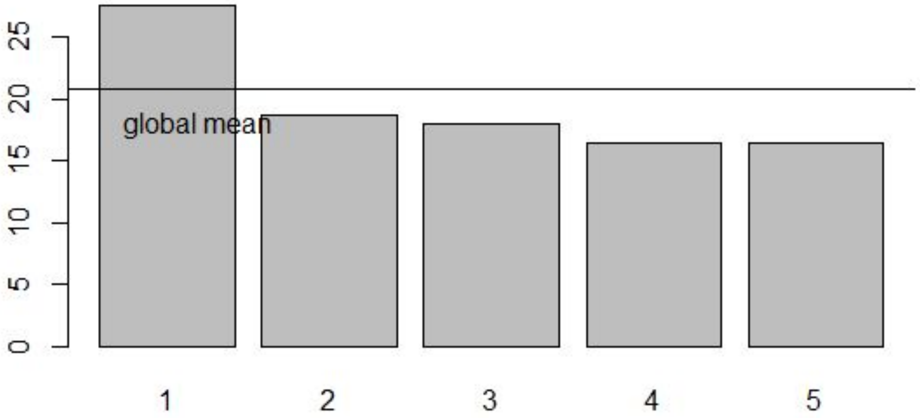


CLASS INTERPRETATION TOOLS

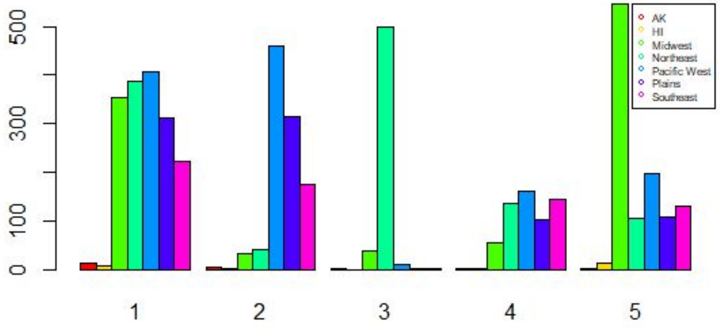
P-values

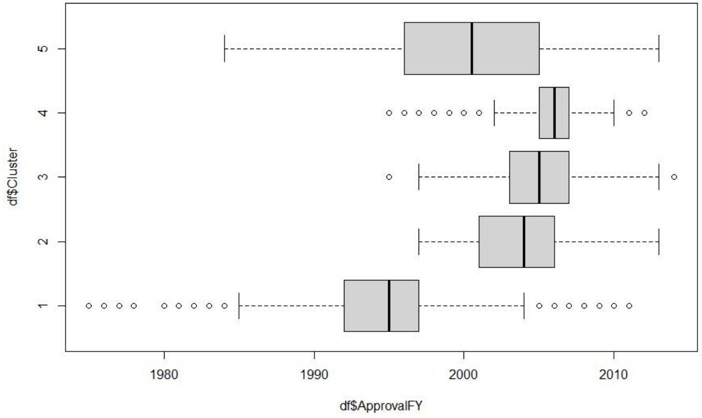


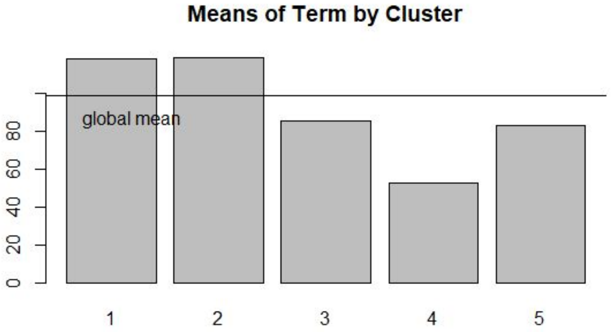
Graphs: Barplot,

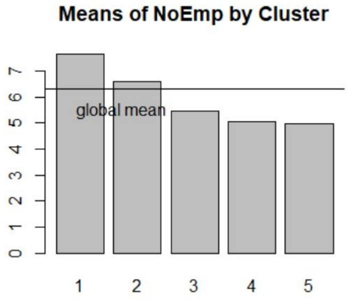


Profiling graphs or numerical information about clusters to be highlighted

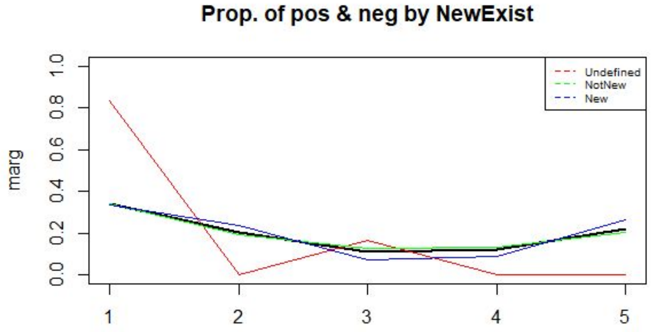


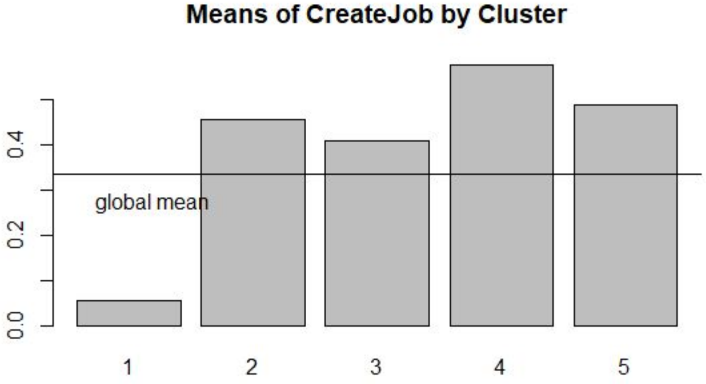


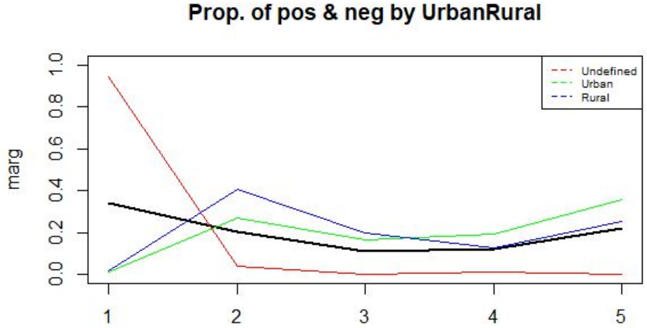

Means of Term by Cluster


Means of NoEmp by Cluster

Profiling graphs or numerical information about clusters to be highlighted

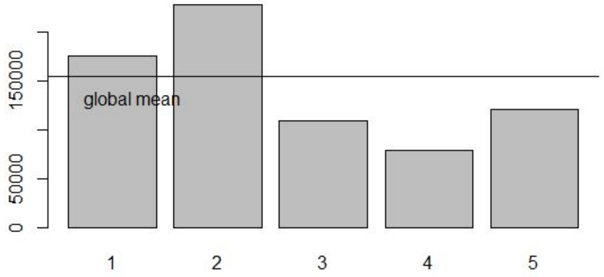




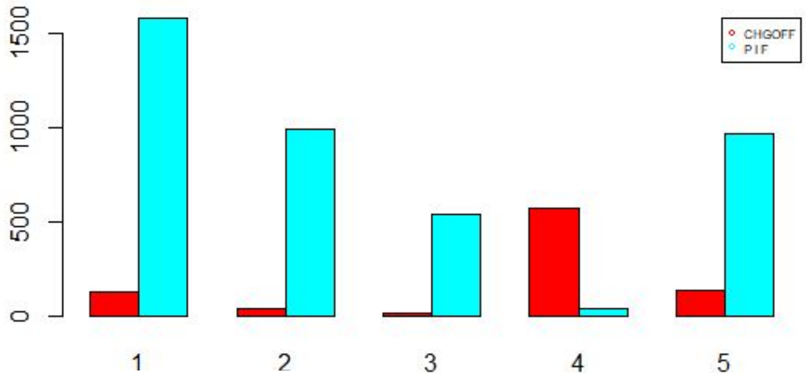


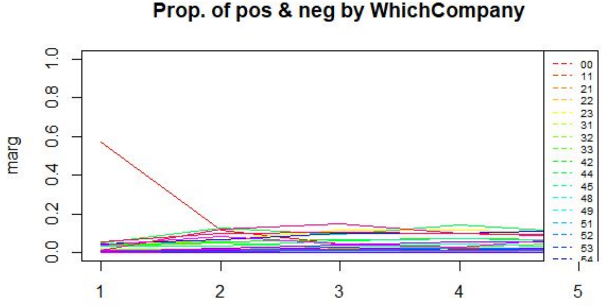
Profiling graphs or numerical information about clusters to be highlighted

DisbursementGross



MIS_Status





FINAL CLASS PROFILING

Cluster 1: We have seen that they have more employees, don’t have many created jobs and they have older loans. That means that they are old companies bigger than the new ones.

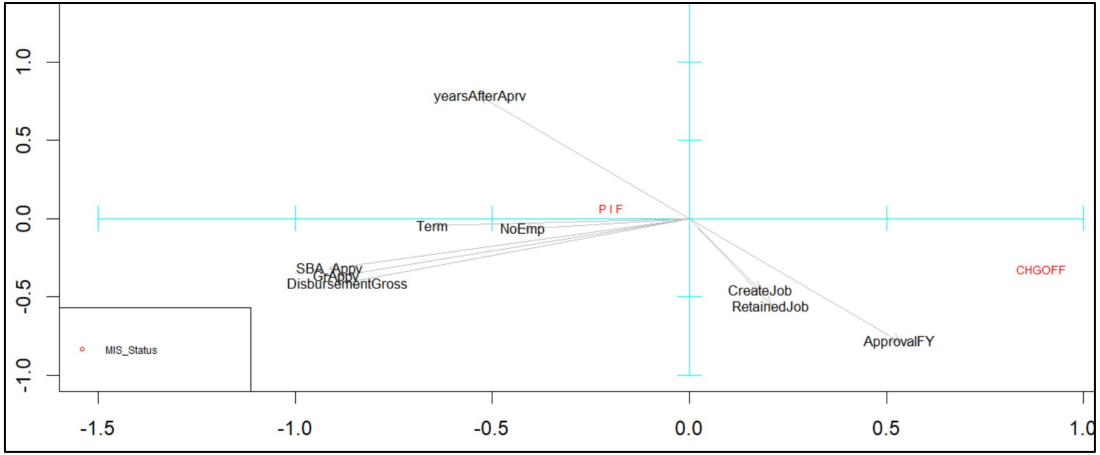
Cluster 2: The companies from this cluster are very similar to the companies in cluster 1 but they are younger and hired more employees with their loan.

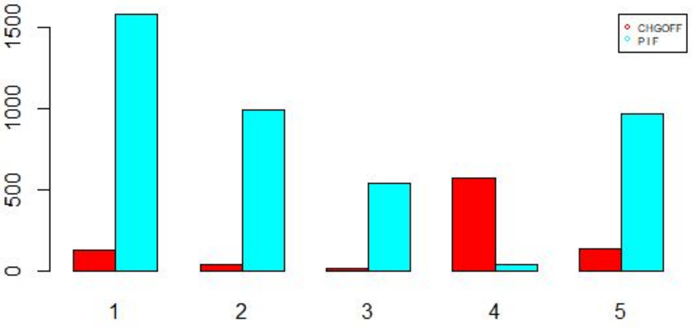
Cluster 3: Is a small cluster with that only highlights because many of its companies are from NorthWest.

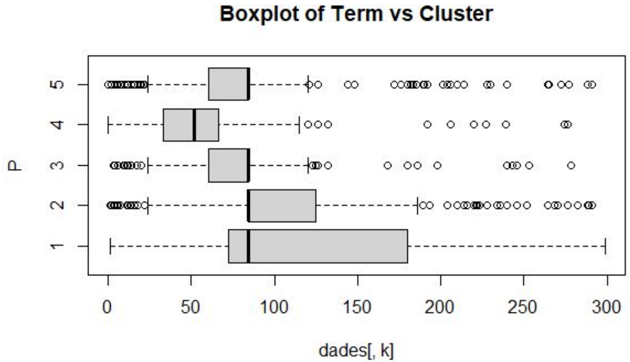
Cluster 4: Have less time to return the loan. It is defined by the big amount of charged off loans (it groups the companies that could not return the loan).

Cluster 5: Its companies are situated in MidWest and Hawaii, and have revolve line credits.

COMPARISON BETWEEN PCA AND CLUSTERING





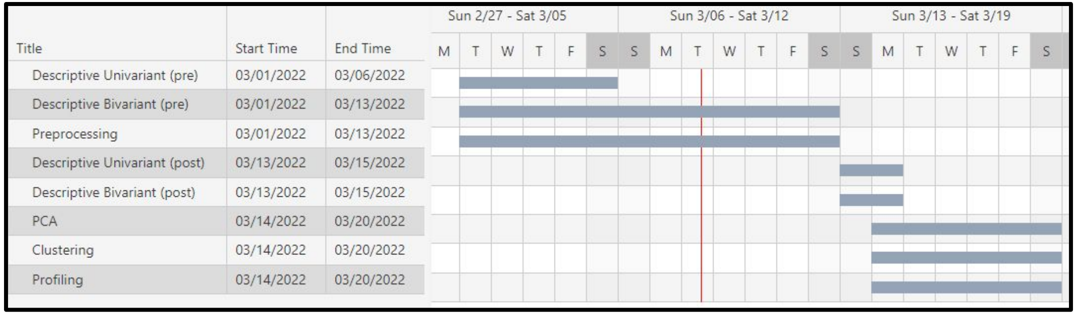


CONCLUSIONS

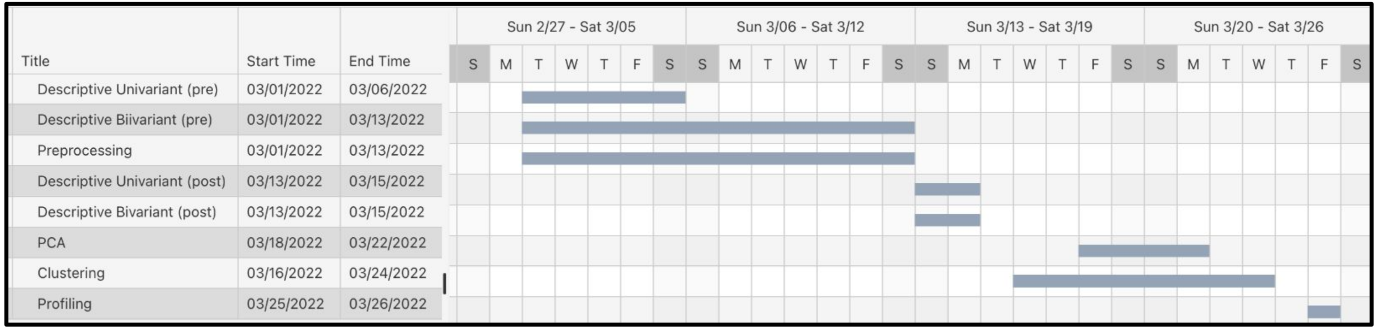
- Some variables define a loan more than others

- More knowledge about R
- Statistical methods
- Importance of cleaning a dataset
- More in-depth knowledge about loans and banks

Initial and final scheduling



Title	Start Time	End Time
Descriptive Univariant (pre)	03/01/2022	03/06/2022
Descriptive Bivariant (pre)	03/01/2022	03/13/2022
Preprocessing	03/01/2022	03/13/2022
Descriptive Univariant (post)	03/13/2022	03/15/2022
Descriptive Bivariant (post)	03/13/2022	03/15/2022
PCA	03/14/2022	03/20/2022
Clustering	03/14/2022	03/20/2022
Profiling	03/14/2022	03/20/2022



Title	Start Time	End Time
Descriptive Univariant (pre)	03/01/2022	03/06/2022
Descriptive Biivariant (pre)	03/01/2022	03/13/2022
Preprocessing	03/01/2022	03/13/2022
Descriptive Univariant (post)	03/13/2022	03/15/2022
Descriptive Bivariant (post)	03/13/2022	03/15/2022
PCA	03/18/2022	03/22/2022
Clustering	03/16/2022	03/24/2022
Profiling	03/25/2022	03/26/2022