Bank Loans

Team 6: Zinan Chen, Qiuhao Chengyong, Qiaoling Huang, Melissa Putur, Weifu Shi

Introduction

Dataset: Loan Data for Dummy Bank

- 800,000 rows and 30 columns
- Columns included:
  - **Information about the Loan**
    - **Example:** Reason for Loan, Loan Amount, Interest Rate & Installments.
  - **Information about the customer**
    - **Example:** Income, Employment Length, Region, Housing Status (Rent or Own)

Project Goals

1. **Identify customer segmentation**
   a. By studying the characteristics of each cluster, bank can adopt different marketing and management strategies to different clusters.
   b. Different level of customized services.
2. **Prediction whether a loan is good or bad**
   a. distinguish what are the good loans and bad loans with supervised machine algorithm
   b. Increase the efficiency of examining the loan status.

Data Cleaning & Feature Selection

As part of our data cleaning process we converted all categorical data columns to dummy variables. We converted the following columns:

- Region: Converted to 5 dummy region columns; Munster, Leinster, Cannught, Ulster and Northern Ireland
- Home Ownership: Converted to 3 dummy columns; Rent, Own, and Mortgage
- Purpose: Converted to 5 dummy columns; Credit Card, Medical, House, Small Business and Vacation
    - Note: The original dataset contained 14 purpose categories, we knew we did not want to keep every category because the original dataset was so large we were not able to successfully upload it to github. When choosing which purposes to keep, we chose the categories that were both interesting to us and were included in at least a couple thousand rows. We removed categories like education, renewable energy and wedding that did not make up a large portion of the total loan reasons.
- Loan Grade: Converted values “A”,”B”,”C” etc. to 7,6,5 etc.

Dimension Reduction

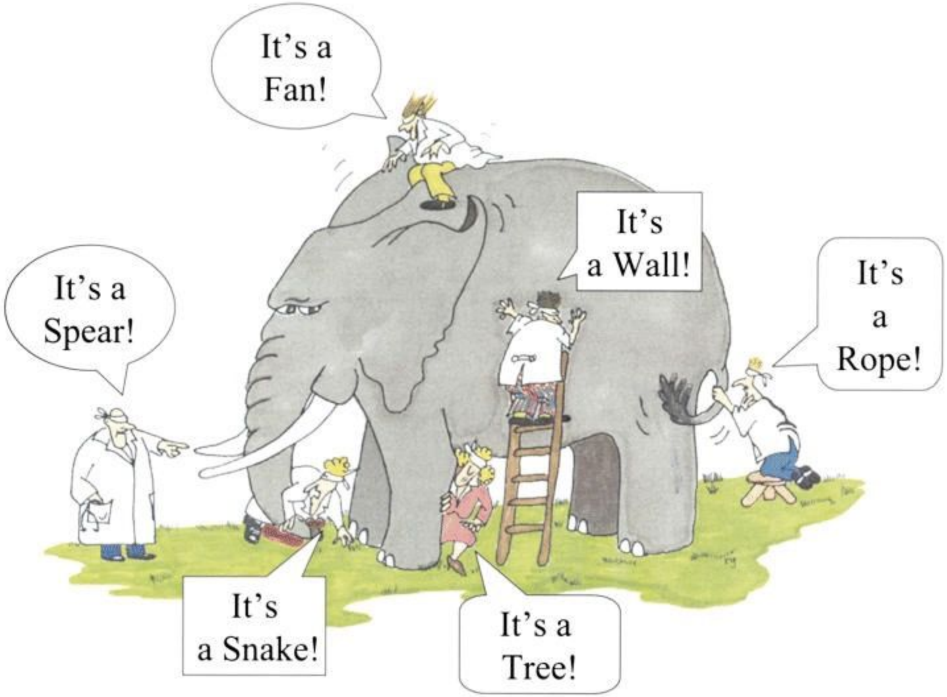
1. PCA

   Is it good with mixed data?

Our data:

50% Binary and Categorical Data

50% Numerical Data



PCA Analysis with Different method

1. Only using the numerical data in the dataset
   (4 principal components and 72% variance percent)

2. Scale all of data including categorical data
   (12 principal components, 86% of the variance percent)

3. scaled with a max/min methodology.
   (7 principal components explains 70% of the variance percent)

4. Scale numerical data and bind with categorical data
   (4 principal components, 64.2% of the variance percent)

Interesting Pattern

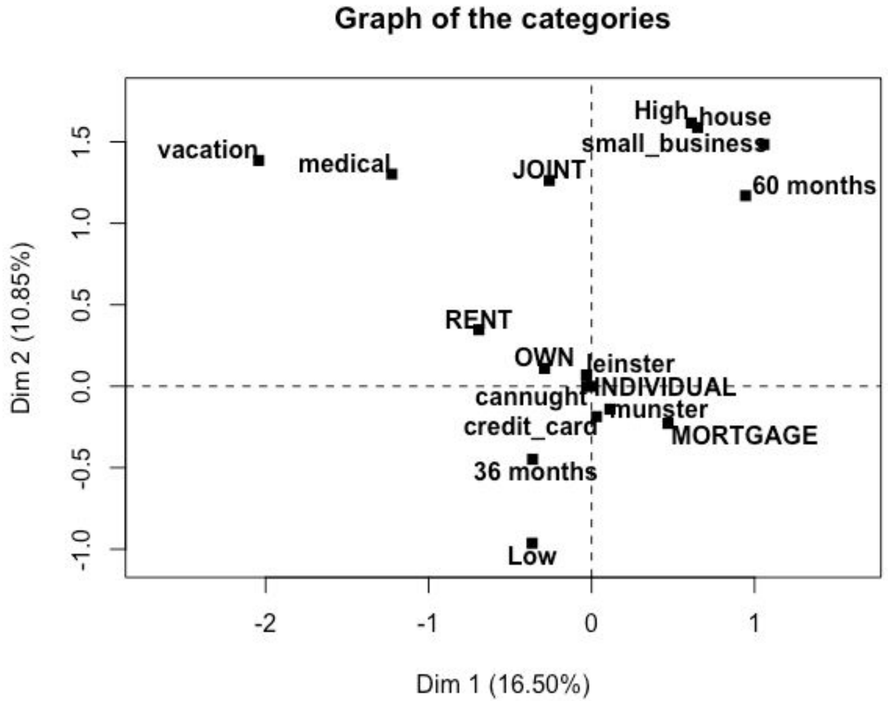
Our guess: Factor Analysis with Mixed Data

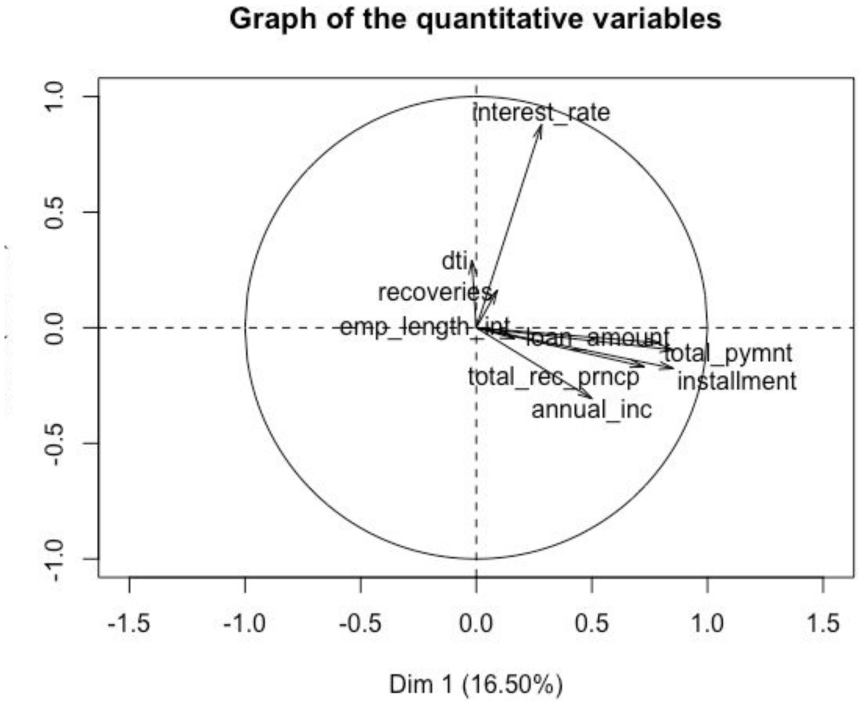
What is FAMD?

FAMD can be seen as a mixed between principal component analysis (PCA) and multiple correspondence analysis (MCA). It acts as PCA for quantitative variables and as MCA for qualitative variables.

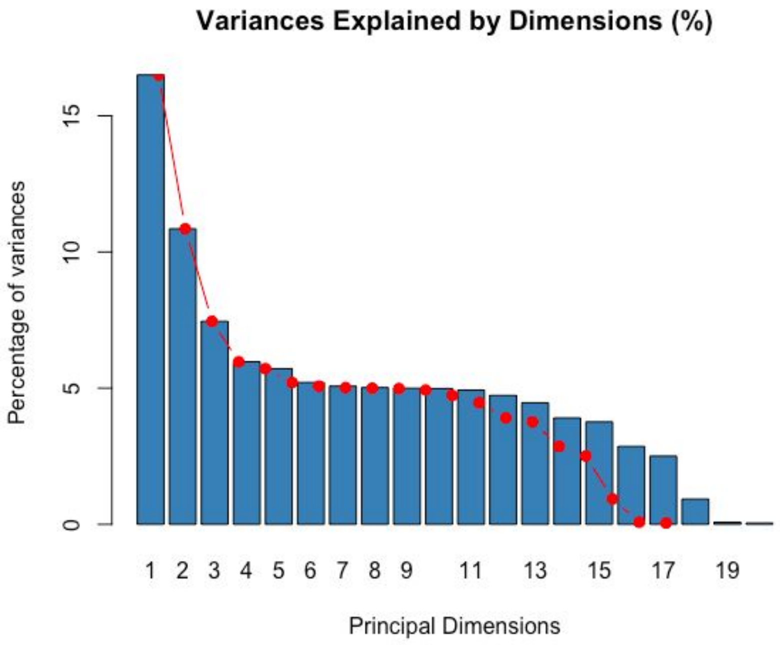
Our data: Around 50% binary or categorical data / 50% numerical data

Graph and Results





Graphs and Results



We choose 12 dimension with around 80% of variance explained percent.

Eigenvalue > 1

Our original dataset: 17 variables

Dimension reduction: 12 Dimensions

Introduce Gower Distance and Daisy Function

Distance is a numerical measurement of far apart individuals are.

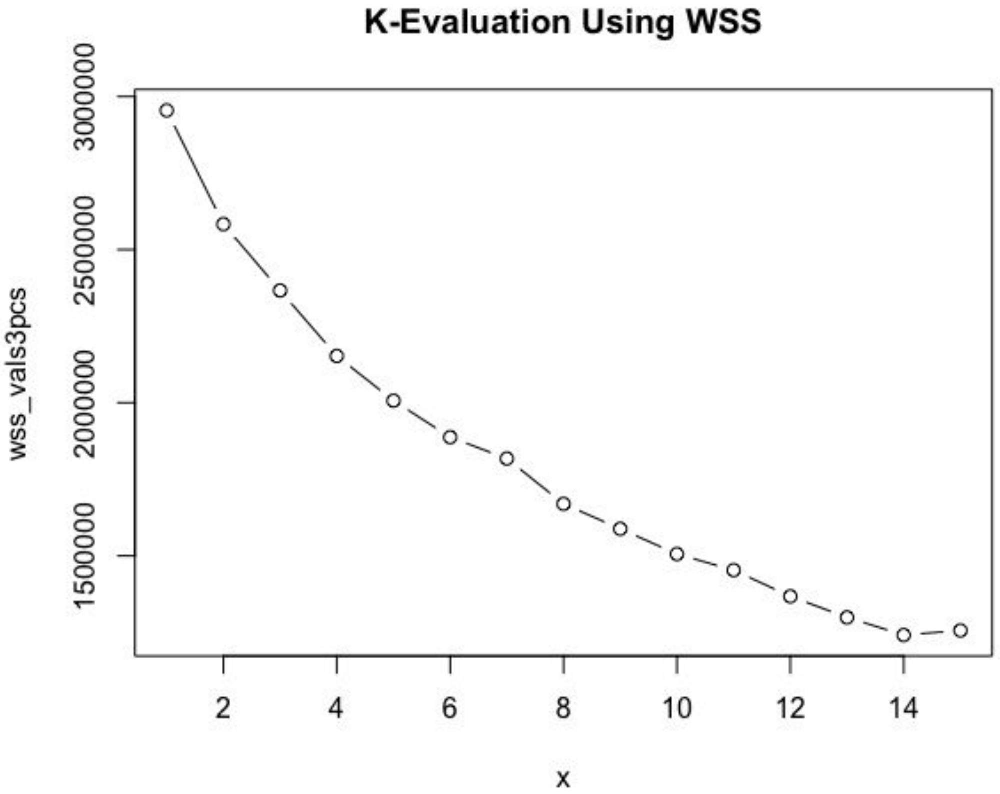
Gower distance is computed as the average of partial dissimilarities across individuals.

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$$

Gower distance is available in R using daisy() function from the cluster package.

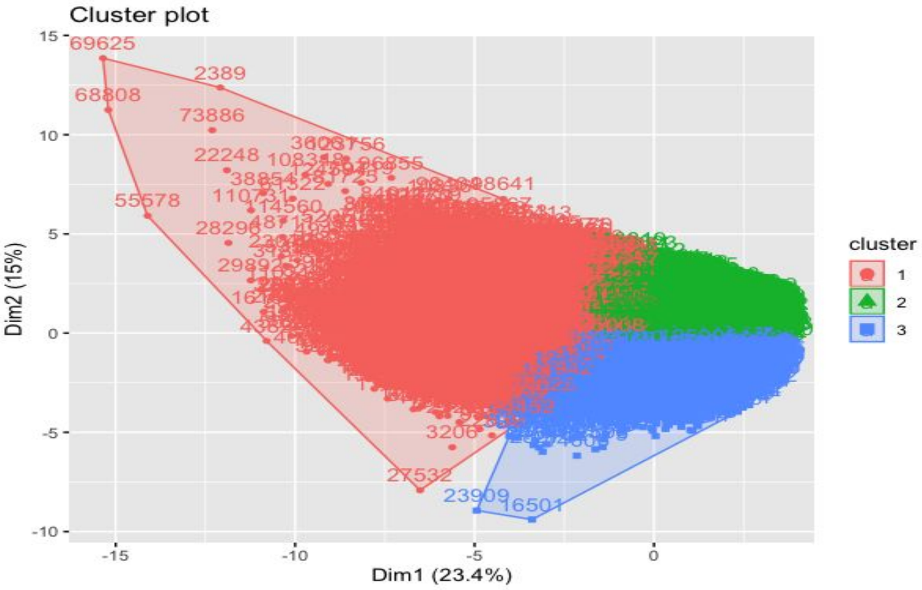
Clustering Analysis

- K-means clustering using original clean data
- K-means clustering using first 12 principle components
- K-means and Hierarchical clustering on the dimensions from T-sne.

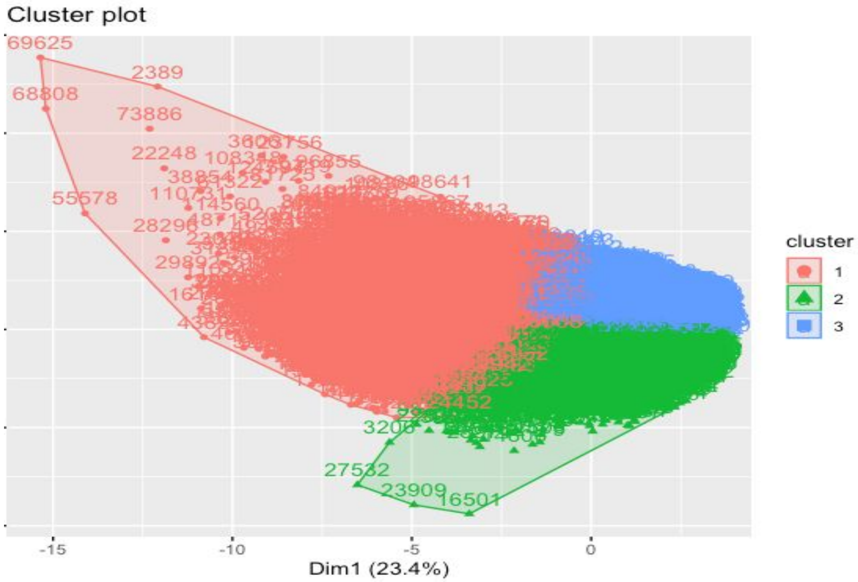


Clustering Results

- Clustering on the original data



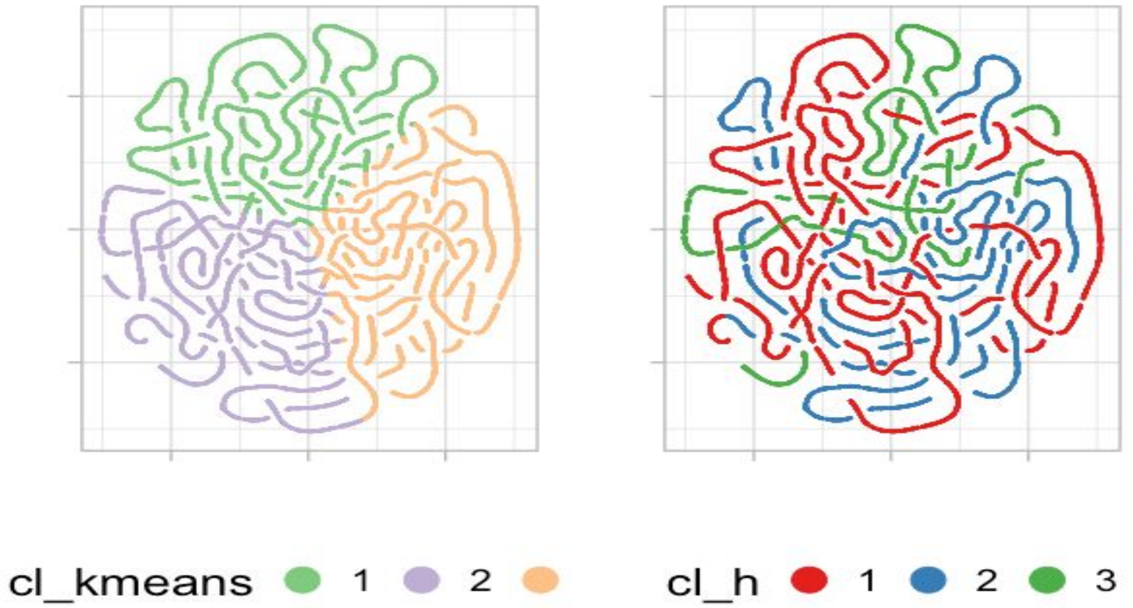
- Clustering on the first 12 principal components



Clustering Results

- Dimension reduction using T-sne

- Clustering using K-means and Hierarchical



Cluster Characteristics

	Cluster 1	Cluster 2	Cluster 3
# of Loans	24852	37208	62961
Avg Salary	$113,574	$57,821	$65,871
Avg Loan Amount	$25,869	$12,912	$11,847
Top Loan Purpose	Credit Card (91%)	Credit Card (77%)	Credit Card (94%)
Avg Total Payment	$17,172	$5,534	$4,214
Avg Installment	763	369	351
Debt to Income	17.7%	20.8%	18.0%
Interest Rate	12.7%	16.5%	9.5%
Employment Length (Yrs)	6.73	5.81	5.77
Home Status	Mortgage (75%)	Split b/w Mortgage and Rent	Split b/w Mortgage and Rent
36 Month	65%	54%	86%
60 Month	35%	46%	14%

PCA Applying to Binomial Classify

```
> get_eigenvalue(train_p)
        eigenvalue variance.percent cumulative.variance.percent
Dim.1  3.481527684      23.21018456                    23.21018
Dim.2  2.229700285      14.86466857                    38.07485
Dim.3  1.519426724      10.12951150                    48.20436
Dim.4  1.344317913       8.96211942                    57.16648
Dim.5  1.113997206       7.42664804                    64.59313
Dim.6  1.008946768       6.72631179                    71.31944
Dim.7  0.981950144       6.54633429                    77.86578
Dim.8  0.834932969       5.56621979                    83.43200
Dim.9  0.762301363       5.08200909                    88.51401
Dim.10 0.695210070       4.63473380                    93.14874
Dim.11 0.565561594       3.77041062                    96.91915
Dim.12 0.249700797       1.66467198                    98.58382
Dim.13 0.187754450       1.25169633                    99.83552
Dim.14 0.015286560       0.10191040                    99.93743
Dim.15 0.009385471       0.06256981                   100.00000
```

```
> head(mod_df)
              PC1        PC2        PC3        PC4        PC5          PC6          PC7
88257 -3.94770744 -0.2026756  2.2768093  0.8849609 -0.3095799  0.259530297 -0.040448365
44811 -0.36850317  1.3527910 -1.2153041 -0.2689789  2.0934925  0.078690176 -0.119494035
6823   0.44542564 -0.8341367 -0.6940127  0.4732847 -0.3388397 -0.003668879 -0.088827374
17650  0.09039107  0.6634471 -0.9606995 -0.4880383 -1.5518184 -0.044088915  0.256789140
59114  1.56349042 -0.6281341 -0.1435591  1.1343528 -1.3091405 -0.260950114 -0.005604597
68070 -0.69711691 -0.1236215  2.0314753 -2.2705064 -0.4272817 -0.259630131 -0.247476692
              PC8         PC9       PC10 loan_condition
88257  0.88042464 -0.69679105  0.6108885            Bad
44811  0.73831907  0.07887952  0.5128414            Bad
6823  -0.07829916 -1.86396698 -0.6210185            Bad
17650 -0.29402377 -0.14642756 -0.4212260            Bad
59114 -1.79673162  0.11165076 -1.0016070            Bad
68070  0.55565436  0.04294642  1.0213098            Bad
```

```
     mod1_pred loan_condition co
1          Bad            Bad
2          Bad            Bad
3          Bad            Bad
4          Bad            Bad
5          Bad            Bad
6          Bad            Bad
7          Bad            Bad
8          Bad            Bad
9          Bad            Bad
10         Bad            Bad
11         Bad            Bad
12         Bad            Bad
13         Bad           Good
14         Bad            Bad
15         Bad            Bad
16         Bad            Bad
17         Bad            Bad
18         Bad            Bad
19         Bad            Bad
20         Bad            Bad
21        Good           Good
22         Bad            Bad
23         Bad            Bad
24         Bad            Bad
25         Bad            Bad
26         Bad            Bad
27         Bad            Bad
28         Bad            Bad
```

TP / TP + FP

```
> Accuracy
[1] 0.9568359
```

SML- Loan Condition Prediction

1. **Model 2**
   a. **Seperate the dataset before PCA and then predict**
      i. **The Accuracy is 96%**

```
> Accuracy
[1] 0.9568359
```

2. **Model 1**
   a. **Dataset - using PCA predict to get a new dataset to do the SML**
      i. **Xg Boosting with PCA data (separate Numeric and Binary) --- 94.1%**

3. **Model 2**
   a. **Database- using original dataset which after cleaning**
      i. **Random Forest - MSE**
      ii. **Xg Boosting with Original Data ----- 96.1%**

```
> paste("Random Forest Train MSE",mse_rf_train)
[1] "Random Forest Train MSE 0.00855293926191827"
> paste("Random Forest Test MSE",mse_rf_test)
[1] "Random Forest Test MSE 0.0349383437673393"
```

Conclusion:
PCA dimension is not necessary for our dataset. We decided to use original dataset to do the supervised machine learning which has better performance.

Conclusions

- Original Dataset was Better than the PCA dimension dataset.
  - Prediction with best PCA----Model ---- 94.1% --- xgboosting
  - Prediction with original data (no PCA) ---- Model. ----- 96% xgboosting.
  - Because the correlation between variables are relatively weak, PCA dimension reduction may not be a good fit for this data
- 3 Clusters
  - Cluster 1: Highest earners, largest loan amounts (more than double average of other two groups), typically own a house
  - Cluster 2: Riskiest customers, lowest earners, larger percent of renters, relatively even split between short and long term loans
  - Cluster 3: Middle of the road customers, Middle earners, larger percent of renters, mostly 36 month loans
- Recommendations
  - Cluster 1: Financially stable customers. We will send our greetings or small gifts for holiday.
  - Cluster 2: We will set up more frequent alert for them to remind their loan status: installment deadlines, current balance, and penalties if the bill is not paid up.
  - Cluster 3: Largest group, middle of the road in terms of performance, study their complaints and concerns to understand how to increase/maintain loyalty

Questions?