Understanding MFIs in India using Multivariate Statistics

31 January 2021
Rohith Krishna, Ishita Gupta

Abstract

Microfinance Institutions (MFIs) offer financial products and services to low income populations driven by the objective of financial inclusion. In India RBI is the main regulatory authority that set norms for margins caps, interest rate caps, and pricing components (such as interest rate, processing fees, insurance premium etc.), and lay other prudential guidelines. MFIs procure funding through traditional methods like bank lending, equity as well through funds and grants from donors. It is therefore essential for banks to evaluate MFIs on their financial performance; MFIs too carry out a self-regulatory and evaluatory analysis. Our objective is to cluster MFIs on the basis of various financial performance indicators (such as debt-to-equity ratio, gross loan portfolio, financial revenue etc.) in order to derive insights on regulatory as well as evaluatory framework for MFIs based on their clustering structure. We collect the required data on the financial performance indicators from the MIX Market dataset made available by the World Bank. We reduce the dimensionality of the chosen dataset using Fisher's Linear Discriminant Analysis. We validate our clustering methodology using Multivariate ANOVA technique.

> **Objective**
>
> - Clustering Analysis on MFI based on Financial Variables
> - Dimensionality Reduction using Fisher's LDA
> - Multivariate ANOVA on clusters and groups within clusters

Introduction

Microfinance, also called microcredit, is the provision of small credit to the low-income individuals or groups who otherwise would have no other access to financial services. The term ‘missing middle’ is often used for this section of our society who lost out on formal credit for want of collateral support. It is here that Microfinance steps in, providing ‘micro’ credit with no collateral whatsoever.

The last few years saw the microcredit industry mainstreaming with the broader financial sector. Some of the largest NBFC-MFIs became Banks or Small Finance Banks (SFBs) and some were bought over by Banks and large NBFCs. Banks and NBFCs also started building their own micro-credit portfolio, through Business Correspondent partnerships. As a result of mainstreaming, the micro-credit sector today is competitive and served by a diverse set of players - Banks, SFBs, NBFC-MFIs, BCs and NBFCs.

Outside of these microfinance providers, there exists a strong ecosystem of other stakeholders including the regulators, the Government, financial institutions, credit bureau, employee bureau, rating agencies and others which play an important role in the delivery of Microfinance.

Microfinance includes a broad range of financial services such as loans, deposits and payment services and insurance to the poor and low-income households and their micro enterprises. Microfinances seek to promote financial inclusion and create social benefits. They are regulated by the Reserve Bank of India (RBI).

Dataset - `MIX Market`

- The data chosen for analysis is referred to as the `MIX Market` dataset and is available for public access in the World Bank website here.
- `MIX Market` is collected and reported by Financial Service Providers across the globe. It is a panel data that comprises of a wide variety of financial and social variables pertaining to MFIs all over the world.
- We select a subset of this data particular to:
    - MFIs located and registered in India
    - Pertaining to the second quarter of 2019 (As of 30 September 2019).
- The chosen quarterly data comprises of 35 financial variables pertaining to the MFI and has 78 distinct MFI records after thorough data pre-processing and cleaning.

Data Preprocessing

- The original dataset contains about 200 variables with some of them having incomplete data.
- We select only those variables for which 90% of the data is available (devoid of missing values).
- For those MFIs for which the latest quarter data is not available, we filled in the missing values with data from the previous quarter.
- If data from the previous quarter is unavailable as well, we impute missing values with the median values for the variable.
- Furthermore, the MFIs which had no data available for a majority of the variables are

removed.
- This activity gives us a final dataset of 78 MFIs with data across 35 variables.

Selected Variable Definitions

- **Gross Loan Portfolio** - All outstanding principals due for all outstanding client loans. This includes current, delinquent, and renegotiated loans, but not loans that have been written off.
- **Cash and Cash Equivalents** - Cash on hand, near cash, and other liquid instruments, including bank balances and deposits. It may include money market investments or treasuries. Cash equivalents are held for the purpose of meeting short-term cash commitments (liquidity management) rather than for investments or other purposes.
- **Liabilities** - Total value of present obligations of the financial institution arising from past events, the settlement of which is expected to result in an outflow from the financial institution of resources embodying economic benefits. For calculation purposes, liabilities are the sum of each individual liability account listed.
- **Equity** - The residual interest in the assets of the financial institution after deducting all its liabilities. For calculation purposes, equity is the sum of each equity account listed.
- **Financial Revenue** - Includes all financial income and other operating revenue which is generated from non-financial services. Operating income also includes net gains (losses) from holding financial assets (changes on their values during the period and foreign exchange differences).
- **Profit and Loss** - The total of income less expenses, excluding the components of other comprehensive income.
- **Operating Expense** - Includes expenses not related to financial and credit loss impairment, such as personnel expenses, depreciation, amortization and administrative expenses.
- **Debt to Equity Ratio** - Total Liabilities / Total Equity
- **Return on Equity Ratio** - (Net Operating Income - Taxes) / Average Total Equity
- **Cost per Loan** - Operating Expense / Average Number of Loans Outstanding
- **Staff turnover rate** - Percentage of staff (permanent and contract) having left the financial institution during the last reporting year, as calculated by the number of staff exiting the organization during the period divided by the average number of permanent and contract staff for the period

Other Variables

- **Personnel** - The number of individuals who are actively employed by an entity. This number includes contract employees or advisors who dedicate a substantial portion of their time to the entity, even if they are not on the entity's employees roster.
- **Loan Officers** - The number of employees whose main activity is to manage a portion of the gross loan portfolio. A loan officer is a staff member of record who is directly

responsible for arranging and monitoring client loans

- **Offices** - The number of staffed points of service and administrative sites / branches used to deliver or support the delivery of financial services and wide array of face-to-face and automated services to clients.
- **Borrowers per loan officer** - Number of Active Borrowers / Number of Loan Officers
- **Number of active borrowers** - The number of individuals who currently have an outstanding loan balance with the financial institution or are primarily responsible for repaying any portion of the gross loan portfolio. Individuals who have multiple loans with a financial institution should be counted as a single borrower.
- **Number of loans outstanding** - The number of loans in the gross loan portfolio. For financial institutions using a group lending methodology, the number of loans should refer to the number of individuals receiving loans as part of a group or as part of a group loan.
- **Loan Portfolio Disbursed** - The value of all loans disbursed in cash during the period, regardless of whether they are performing, non-performing, or written off.
- **Number of loans disbursed** - The number of loans disbursed during the period.
- **Net Loan Portfolio** - Value of loan portfolio net of impairment loss allowance and unearned income and discount (when applicable).
- **Assets** - Total value of resources controlled by the financial institution as a result of past events and from which future economic benefits are expected to flow to the financial institution. For calculation purposes, assets are the sum of each individual asset account listed.
- **Borrowings** - The principal balance for all funds received through a loan agreement. It may include bonds or similar debt securities issued and credit lines.
- **Net Operating Income** - Total operating revenue less all expenses related to the financial institution's core financial service operation including total financial expense, impairment loss and operating expense.
- **Financial expense on funding liabilities** - All costs incurred in raising funds from third parties including deposits, borrowings, subordinated debt and other financial obligations in addition to fee expenses from non-financial services
- **Personnel Expense** - Includes wages and salaries, other short-term employee benefits as bonuses and compensated absences, post-employment benefit expense, termination benefit expense, share-based payment transactions, other long-term benefits and other employee benefits.
- **Administrative expense** - Non-financial expenses excluding personnel directly related to the provision of financial services or other services that form an integral part of a financial institution's financial services relationship with clients.
- **Tax expense** - Income tax for the period comprising current and deferred tax.
- **Capital /asset ratio** - Total Equity / Total Assets
- **Gross loan portfolio to total assets ratio** - Gross Loan Portfolio / Total Assets
- **Return on assets** - (Net Operating Income - Taxes) / Average Total Assets
- **Operational self sufficiency** - Financial Revenue / (Financial Expense + Net Impairment Loss + Operating Expense)

- **Financial revenue / assets** - Financial Revenue / Average Total Assets
- **Profit margin** - Net Operating Income / Financial RevenueYield on gross portfolio (nominal) - Financial Revenue from Loan Portfolio / Average Gross Loan Portfolio
- **Financial expense / assets** - Financial Expense / Average Total Assets
- **Operating expense / assets** - Operating Expense / Average Total Assets

Methodology

> We follow the following steps in our analysis:
>
> - PCA on the original (78x35) dataset
> - KMeans Clustering on data subsets as well as reduced dataset.
> - Fisher's LDA for dimensionality reduction.
> - MANOVA for cluster validation.

Algorithm

- Split the original dataset variables into various heads based on their nature:
  - Employee, Loan Portfolio, Balance Sheet, Income & Expense, Financial Ratios
- Perform PCA to identify optimal number of PCs required for lower dimensional data representation.
- Obtain KMeans clusters from the Principal Components. Use this as a baseline model and compare the clusters formed using Fisher's LDA to those obtained here.
- Identify optimal number of clusters using elbow plot and perform K-Means Clustering on individual data subsets.
- Reduce each data subset into a single variable using Fisher's LDA. Here, the clusters produced by KMeans form the corresponding groups.
- Perform clustering based on reduced dataset and compare the results with the clustering on the original dataset.
- Remove outlier clusters and select only those clusters with greater than 5 observations.
- Use MANOVA to validate cluster analysis:
  - Use MANOVA to compare the means of different clusters obtained. If $H0$ is rejected then the clusters are significantly different to each other.
  - Select a cluster. Use MANOVA to compare the means of randomly formed groups within chosen cluster. If $H0$ is accepted then the observations in the cluster have significantly similar means.

Custom Function Definitions

> The following functions are defined for use in the analysis:
>
> - `getElbowPlot()` - To perform multiple KMeans clustering with variable K's and determine the optimal number of clusters (K)
> - `getKMeansClusters()` - To obtain KMeans clusters using the `sclearn` package
> - `plotKMeansClusters2D()` - Plot clusters in the 2D space of selected variables.
> - `plotKMeansClustersOnPCs()` - Plot clusters on 2D principal component space.
> - `getPCA(df_passed)` - Obtain Principal components using `sclearn` package.
> - `eigenfn(matrix)` - Obtain decreasingly-sorted eigenvalues and eigenvectors of a matrix.
> - `fishersLDA(df)` - Perform Fisher's LDA and obtain samples on a lower dimensional projected space, based on specified classes (clusters)
> - `manova(df)` - Test the null hypothesis that group means are equal across a multivariate dataset.

- `getElbowPlot(df_passed, n_clus, rs = 42)`
    - Inputs the dataframe, maximum number of clusters to be tried and optional random state.
    - Outputs the Elbow plot between number of clusters and within cluster sum of squares.
- `getKMeansClusters(df_passed, n_clus, rs=42)`
    - Inputs the dataframe, number of clusters and optional random state.
    - Outputs the KMeans Cluster labels as a numpy array.
- `plotKMeansClusters2D(df_passed, n_clus, df_plotted, x_var, y_var, rs=42)`
    - Inputs the dataframe on which clustering is performed (df_passed), number of clusters, dataframe for visualizing clusters on 2D scatterplots (df_plotted), x variable name, y variable name and optional random state.
    - Outputs a scatter plot between the chosen variables that can be saved as
    - Note: df_passed and df_plotted need not be the same. Their variables may differ, but their sizes must be same to run error-free. In addition, for sake of interpretation, the observations must be the same and their order preserved.
    - Note: x and y variable names are passed as strings. These columns must be present in df_plotted but not necessarily be in df_passed.
- `plotKMeansClustersOnPCs(df_passed, n_clus, rs = 42)`
    - Inputs the dataframe, number of clusters and optional random state

    - Plots the KMeans Clusters on Principal Component Dimensions in 2D.
    - Note: It does not perform the clustering over PCs, rather the PCs are simply used for visualization.
    - It outputs the PC plot as well as the explained variance ratio as a numpy array.
- `getPCA(df_passed)`
    - Inputs the dataframe for which Principal components need to be computed.
    - Outputs the Principal Components and the Explained Variance Ratio
    - Note: The StandardScaler is implemented within this function: thus the input dataframe is scaled before PCA is applied.
- `eigenfn(matrix)`
    - Inputs the matrix whose eigenvalues and eigenvectors are to be computed.
    - Outputs the 'descending'-sorted eigenvalues and eigenvectors.
- `fishersLDA(df)`
    - Inputs the grouped dataframe. Note that one of the columns must contain these groups and must be renamed as `Group` before passing the dataframe.
    - Outputs a dataframe in a new feature-space using Fisher's Linear Discriminant Analysis.
    - The selected features can be used in dimensionality reduction.
- `manova(df)`
    - Inputs the grouped dataframe. Note that one of the columns must contain these groups and must be renamed as `Group` before passing the dataframe.
    - Outputs the Wilks Lambda Test statistic, Chi-square value and checks for acceptance/rejection of the Null Hypothesis.

Results and Discussion

- The dataset is split into subsets with the following heads (features):
    - **Employees** - personnel, loan-officers, staff-turnover-rate, offices
    - **Loan Portfolio** - gross-loan-portfolio, number-of-active-borrowers, number-of-loans-outstanding, loan-portfolio-disbursed, number-of-loans-disbursed, borrowers-per-loan-officer, net-loan-portfolio
    - **Balance Sheet** - cash-and-cash-equivalents, assets, borrowings, liabilities, equity
    - **Income and Expenses** - financial-revenue, net-operating-income, profit-loss, financial-expense-on-funding-liabilities, operating-expense, personnel-expense, administrative-expense
    - **Financial Ratios** - capital-toasset-ratio, debt-to-equity-ratio, gross-loan-portfolio-to-total-assets, return-on-assets, return-on-equity, operational-self-sufficiency, financial-revenue-to-assets, profit-margin, yield-on-gross-portfolio-nominal, financial-expense-to-assets, operating-expense-to-assets, cost-per-loan

- The Elbow Plot provides a method to determine the optimal number of clusters, using within cluster sum of squares. Further the Principal Component Analysis finds the percentage of variance explained by each component in a lower dimensional feature space. These results are summarised below:

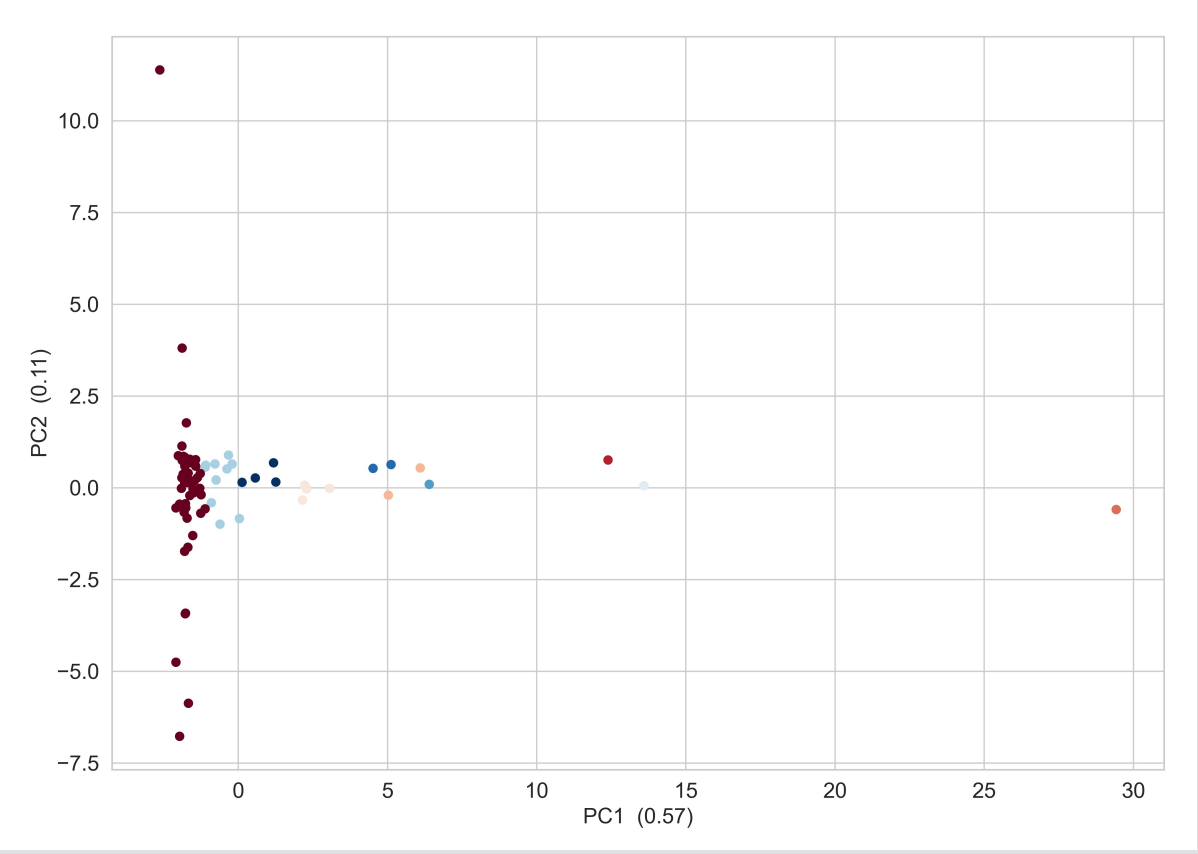
Data Subset	K (from elbow plot)	PC 1 (% var explained)	PC 2 (% var explained)	% var explained by PC1 and PC2
All Variables (original)	10	0.57	0.11	0.68
Employees	6	0.74	0.25	0.99
Loan Portfolio	6	0.8	0.14	0.94
Balance Sheet	6	0.94	0.04	0.98
Income & Expenses	4	0.954	0.04	0.994
Financial Ratios	11	0.31	0.25	0.56
Reduced Dataset	10	0.36	0.30	0.66

- With the optimal number of clusters from above, we perform KMeans clustering on each of data subset. Then the clusters are treated as groups and Fisher's LDA is performed on these data subsets. We then select the largest eigenvalue for a given subset and compute a new feature by taking the linear combination of all old features in the subset with the eigenvectors corresponding to the largest eigenvalue as weights.
- On repeating this across the 5 subsets we obtain the following features:

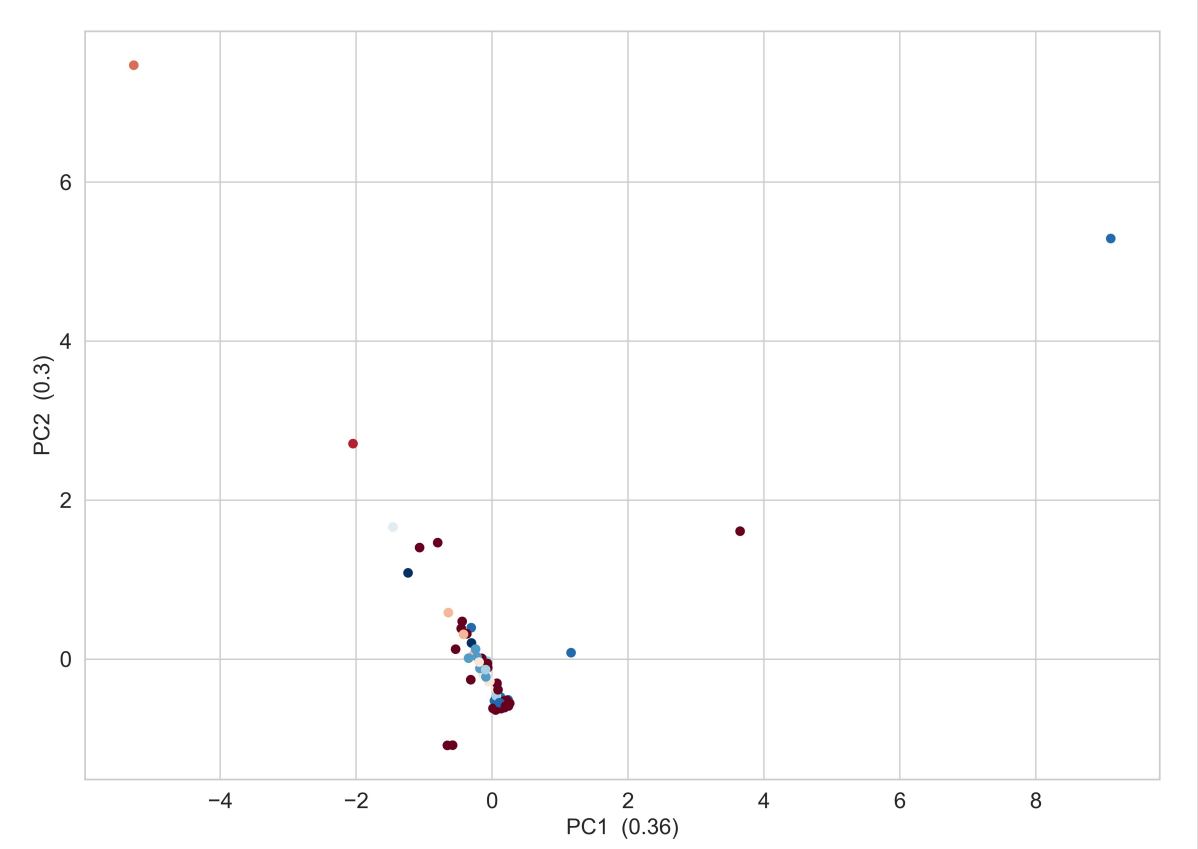
MFI #	Employee	Loan Portfolio	Balance Sheet	Income & Expense	Financial Ratios
0	-3.797004	-201.60931	5.36E+08	-3.30E+06	-0.130223
1	-1.540299	-463.724195	9.77E+06	-2.21E+05	-0.121085
2	-1.291945	-232.237392	2.96E+08	-1.80E+05	-0.136981
3	-2.130679	-12.545701	2.07E+09	-2.36E+06	-0.156978
4	-0.487372	-318.53529	4.86E+08	-2.13E+06	-0.190488
...	...	...	...	...	...
73	-0.94654	-1188.556761	4.85E+09	-3.75E+07	-0.130729
74	-0.667621	-930.628743	2.82E+09	-4.66E+07	-0.142187
75	-0.637256	-212.837467	3.57E+09	-4.26E+07	-0.143837
76	-11.944348	-1027.830542	2.33E+09	-1.85E+07	-0.143297
77	-8.379749	-1375.806047	1.58E+09	-5.37E+07	-0.16165

- We compare the cluster plots in PC space produced by the two methods:
  - Clustering on dimensionality-reduced features: 3 PC - 85% variance explained
  - Clustering on Original dataset with 35 features: 3 PC - 74% variance explained
  - These plots are presented in the figures below:

Clustering on the 35-feature original dataset



Clustering on the dimensionality reduced dataset



- The 10 clusters formed have the following number of MFIs per cluster:

CLUSTER	# MFIs	
Cluster 6	44	considered for further analysis
Cluster 9	12	considered for further analysis
Cluster 1	9	considered for further analysis
Cluster 8	5	outliers --> removed.
Cluster 7	2	outliers --> removed.
Cluster 4	2	outliers --> removed.
Cluster 5	1	outliers --> removed.
Cluster 3	1	outliers --> removed.
Cluster 2	1	outliers --> removed.
Cluster 0	1	outliers --> removed.

- We perform Multivariate ANOVA on the selected dataset with just 3 clusters: number 6, 9 and 1. The other clusters are removed for they are outliers in the dataset. For instance, Bharat Finance is tremendously large MFI with its Gross Loan Portfolio spanning to several times the average value for an MFI. We also know that this MFI has since merged with IndusBank.
- In MANOVA, we treat each cluster to be a group and check the null hypothesis that, across all $k = 5$ variables, $H_0 : \mu_{6k} = \mu_{9k} = \mu_{1k}$. We find the following:

```
Wilks' Lambda   : 0.015
Degrees of Freedom  : 10
Chi-sq (calculated) : 251.96935
Chi-sq (critical) : 20.48318

Since Chi-sq-calc > Chi-sq-crit, H0 is rejected.
```

  - This validates our hypothesis that the clusters so formed have significantly different means. Therefore from a regulatory point of view, we understand that these MFIs are to treated differently.
- In the second MANOVA, we pick up each cluster and within a cluster create random groups. On testing a similar hypothesis as before, we find that:

```
Multivariate ANOVA Results
--------------------------
H0: U_1k = U_2k = ... = U_gk
H0 is rejected when Chi-sq-calc is greater than Chi-sq-crit.
This occurs when the Wilk's Lambda is relatively small.

Number of groups in each cluster = 2

Cluster 6
Wilks' Lambda   : 0.8855
Degrees of Freedom  : 5
Chi-sq (calculated) : 4.80344
Chi-sq (critical) : 12.8325

Since Chi-sq-crit > Chi-sq-calc, H0 is accepted.
--------------------------------------------------
--------------------------------------------------

Cluster 9
Wilks' Lambda   : 0.67677
Degrees of Freedom  : 5
Chi-sq (calculated) : 2.92817
```

```
Chi-sq (critical) : 12.8325

Since Chi-sq-crit > Chi-sq-calc, H0 is accepted.
--------------------------------------------------
--------------------------------------------------

Cluster 1
Wilks' Lambda   : 0.48241
Degrees of Freedom  : 5
Chi-sq (calculated) : 3.28028
Chi-sq (critical) : 12.8325

Since Chi-sq-crit > Chi-sq-calc, H0 is accepted.
```

    - This validates our cluster analysis, since we see that for different groups within each cluster across all variables are not significantly different.
- In order to confirm this further, we also perform the MANOVA on the original 10 clusters formed from original 35-variable dataset. We obtain that cluster 6 maintains almost the same number of MFIs - appoximately 48 in number from the earlier 44. The results of MANOVA on 2 random groups within this cluster 6 is:

```
Cluster 6
Wilks' Lambda   : 0.92042
Degrees of Freedom  : 5
Chi-sq (calculated) : 3.27549
Chi-sq (critical) : 12.8325

Since Chi-sq-crit > Chi-sq-calc, H0 is accepted.
```

    - Thus the results on clusters formed are robust as validated by MANOVA.
- Another component of analysis features MANOVA on tertile groups of MFIs based on their size (measured by the Gross Loan Portfolio). As expected the results show a singnificant difference in the group means across all variables. This is because the groups thus created reflect small, medium and large MFIs respectively.

```
Multivariate ANOVA Results
--------------------------
H0: U_1k = U_2k = ... = U_gk
H0 is rejected when Chi-sq-calc is greater than Chi-sq-crit.
This occurs when the Wilk's Lambda is relatively small.

Wilks' Lambda   : 0.46242
Degrees of Freedom  : 8
Chi-sq (calculated) : 56.6895
Chi-sq (critical) : 17.53455
```

```
Since Chi-sq-calc > Chi-sq-crit, H0 is rejected.

--------------- BSS ---------------
[228350785.256, 140925540.705, 1826.655, 24672022.756]
[140925540.705, 86972224.333, 1131.159, 15224284.282]
[1826.655, 1131.159, 0.035, 187.244]
[24672022.756, 15224284.282, 187.244, 2670749.564]
--------------- WSS ---------------
[402195133.423, 256257012.154, -642.298, 36828237.538]
[256257012.154, 165815113.885, -514.126, 23491156.808]
[-642.298, -514.126, 0.468, -87.962]
[36828237.538, 23491156.808, -87.962, 3548863.885]
```

Insights on MFIs

- ***Staff turnover rate*** - Shows the **stability of the MFI**. We understand whether people look at an MFI as a short term job based on this attrition level.
    - It is also linked to Personnel Expense. More people quitting implies that more money is spent on hiring & training new people.
- ***Gross Loan Portfolio*** - Gives an idea about the **size of the MFI**. A larger MFI such as Bharat Finance has a larger Gross Loan Portfolio when compared to any other MFI.
- ***Cash and Cash Equivalents***: Throws insight into the **liquidity of the MFI**, and short term cash requirements. An MFI doing badly on this variable would suffer from liquidity crunch if depositors wish to cash in short term.
- ***Liabilities*** - How much the **MFI owe its lenders?** Indebtedness factor. An MFI with a high liability faces a greater risk of insolvency. It also goes into the reliability of the company
- ***Equity*** - Gives the **perception of MFI to potential Investors.** Well performing MFI can attract investors and eventually could merge with banks. We have seen this in the case of Bharat Financial. Another MFI Spandana Sphoorty Financial raised Rs 1.25 billion from existing investors.
- ***Financial Revenue*** - Talks about the **Growth of the MFI**.
- ***Profit/ Loss*** - Goes into the financial **performance of an MFI**.
- ***Operating expense*** - Talks about the **cost of running the MFI - maintenance**.

Conclusion

We have found that there exists three major clusters of MicroFinance Institutions in India. From our analysis, these clusters have a strength of 44, 12 and 9 respectively, after removing for outlier cluster MFIs. These clusters appear to be robust across the different iterations of analysis we have conducted. KMeans clustering on both a PCA of the dataset as well as a groupwise Fisher's LDA dimensionality reduction yield the same set of MFIs grouped to the three clusters.

Further we also validate these clusters using Multivariate ANOVA by testing for statistically different group means across clusters and similar means within subgroups of a single cluster. Further scope for research involves a closer study of the MFIs clusters, their cluster averages for different variables and its implications from a regulatory and supervisory point of view. We believe that effective regulation of MFIs would go a long way in properly incentivizing banks to give credit on reasonable terms. This would also facilitate MFIs in self-regulation and grant credit with more caution, so as to prevent households from falling into the over-indebtedness trap.

References

- Chasnow Johnson - Equity Investment for Microfinance.
- Equity of a microfinance insitution
- Key drivers of MFI pricing and valuation https://www.dvara.com/wp-content/uploads/2011/03/IFMRTrustDiscussionNote-MFIPricingandValuation.pdf
- Malegam Committee MFI Sector report https://www.dvara.com/research/Files/Response-to-the-Malegam-Committee-report-on-Issues-and-Concerns-in-the-MFI-Sector.pdf
- https://mfinindia.org/microfinance/industryOverview
- https://www.arohan.in/Industry_Code_of_Conduct.pdf
- https://m.rbi.org.in/Scripts/FAQView.aspx?Id=102