DSA 5900 Professional Practice- Final Report

Name: Nilam Reshim
Semester: Spring 2021
Course Number: DSA 5900
Credit hours: 4
Faculty supervisor: Dr. Browder, Russell E.

Introduction

The aim of the project is to provide useful analysis about survey responses from Nation of Makerspaces to the stakeholders.

The Nation of Makers (NOM) is a national non-profit supporting makerspaces and related organizations through community building, resource sharing, and advocacy.

A makerspace is a collaborative workspace inside a school, library, or separate public/private facility for making, learning, exploring, or creating new thing. These spaces are open to all age groups, entrepreneurs and provide variety of maker equipment/services such as laser cutters, CNC machines, sewing machine etc.

Apart from equipment they also provide programming like entrepreneurship/business support services, offsite education reach.

It is a place for creative people to have the space, tools, and community needed to create.

For many it is a place to work on their personal projects and for others, it is a place to get together with other like-minded individuals.

There are three major maker spaces: Nonprofit, for-profit, and other, including larger educational institutes, libraries, museums, and government entities. NOM has conducted multiple surveys over the last few years to help understand the landscape of U.S. makerspaces, their members, and their leaders. The surveys provide a directional perspective on importance of mission, entrepreneurship, innovation in maker spaces and economic activities at different levels.

The 2019 survey makerspace was composed of three descriptive surveys designed to provide an explanation of the nature of makerspaces.

The economy survey is targeted towards starting a makerspace, facility, and team, services, and tool offerings.

The makerspace leadership survey is targeted towards data pertaining to the demographics, making activities, and common practices of makerspace leaders.

The makerspace member survey is targeted towards data pertaining to the demographics, making activities, inclusion and diversity and motivations of makerspace members.

The goal of the survey is to analyze data to understand equipment offered in maker spaces, services offered in maker spaces, purpose of maker's activities in a maker spaces and income generating activities of individuals in a maker space.

The analysis will involve, connecting multiple survey files to identify opportunities for longitudinal and multi-level analysis.

Statistical and cross-sectional analysis of relationships between variables related to, for example, makerspace facilities, membership communities, leadership practices, and entrepreneurial activities and outcomes.

Objective

In this DSA project, the objective is to,

- Provide useful analysis that will help understand complex interrelation between different features of the questionnaire at group level. Changes in the pattern of the data in two different years, economical activities, and how these activities support small businesses.
- Explore potential opportunities for published research.
- Questions pointing to similar dimensions.

The detailed analysis at the individual level and makerspace level is already available for current and previous years. The objective is not to repeat this analysis and focus on cross referencing between different questions.
Measurement validity, scale reliability and method for minimizing missing data is not a primary focus.

Data

The dataset consists of responses from three surveys of the U.S. maker space population for two years.

Table 1. Dataset Description

Year	Type of makerspace population	Sample size	Features	Observations
2019	Economy	114	238	114
2019	Leader	332	252	332
2019	Member	860	178	860
2018	Economy	No unique identifier	155	91
2018	Leader	No unique identifier	236	103
2018	Member	No unique identifier	354	61

In all the questionnaires Personally, identifiable information (PII) data is not available. The data do not contain missing values as majority of the questions are mandatory and cannot be skipped. The survey was presented to participants on NOM website. The responses are captured using SurveyMonkey tool.

Connecting individuals was not possible due to the unavailability of common factor/ PII information. Analysis is performed mainly for 2019 dataset for individual makerspace population. Types of questions available are Open ended, Multiple choice, Single choice, Rating scale and Matrix questionnaire.
Multichoice questions are available in one-hot encoding format.

The variables can be split into the following groups:

- Economy (38 questions),
  - About Organization (4 questions)
  - The numbers: startup cost (5 questions)
  - The numbers: last full fiscal year (2 questions)
  - Access to space (6 questions)
  - Tools, capabilities, and programming offers (14 questions)
  - Background of the people served (7 questions)

- Leaders (25 questions),
  - Demographic (8 questions)
  - Relationship with makerspace (7 questions)
  - Your Activities (6 questions)
  - The challenges (4 questions)

- Members (39 questions),
  - Demographic (7 questions)
  - Relationship with makerspace (4 questions)
  - Your Making (9 questions)
  - Member's involvement in the makerspaces (5 questions)
  - Inclusion and diversity (10 questions)
  - The expectations (4 questions)

Data Preparation *(Represented only for Leader dataset due to page limitation)***,**

- Data Cleaning
  Feature selection,
  Select features which are point of interest and key factors as required by sponsor.
  Select features which explains more than 10% of the total data points.
  Identify and drop columns which are highly correlated.

  In the leader dataset, task performed by the leader is multichoice question and has 72 options. These 72 options are available in one hot encoded format and hence 72 columns. Each task is subcategorized into different levels. For instance,
  HR - Staff management, HR - Policy issues, HR - Tracking Hours, HR - Resolving disputes internally etc.
  By identifying the task which are highly correlated number of features can be reduced for efficient data exploration.

Final dataset after feature selection,
32 features and 332 entries

Fig. 1. Correlation plot for leader's task

The fig.1 heatmap highlights features that has correlation of more than 70%. The 70 tasks are reduced to 12 after removing highly correlated features. Group continuous variables into categorical variables and convert string data into numeric format.

- Label encoding:

Encode Categorical features using preprocessing. LabelEncoder() from sklearn.

- Remove Nan:

As all the questions are mandatory. Nans are not present in the data. Multichoice questions are represented into one-hot-encoding format. Null for these questions are replace by 0 using fillna(0).

Data pattern:

The count of number of questions answered,

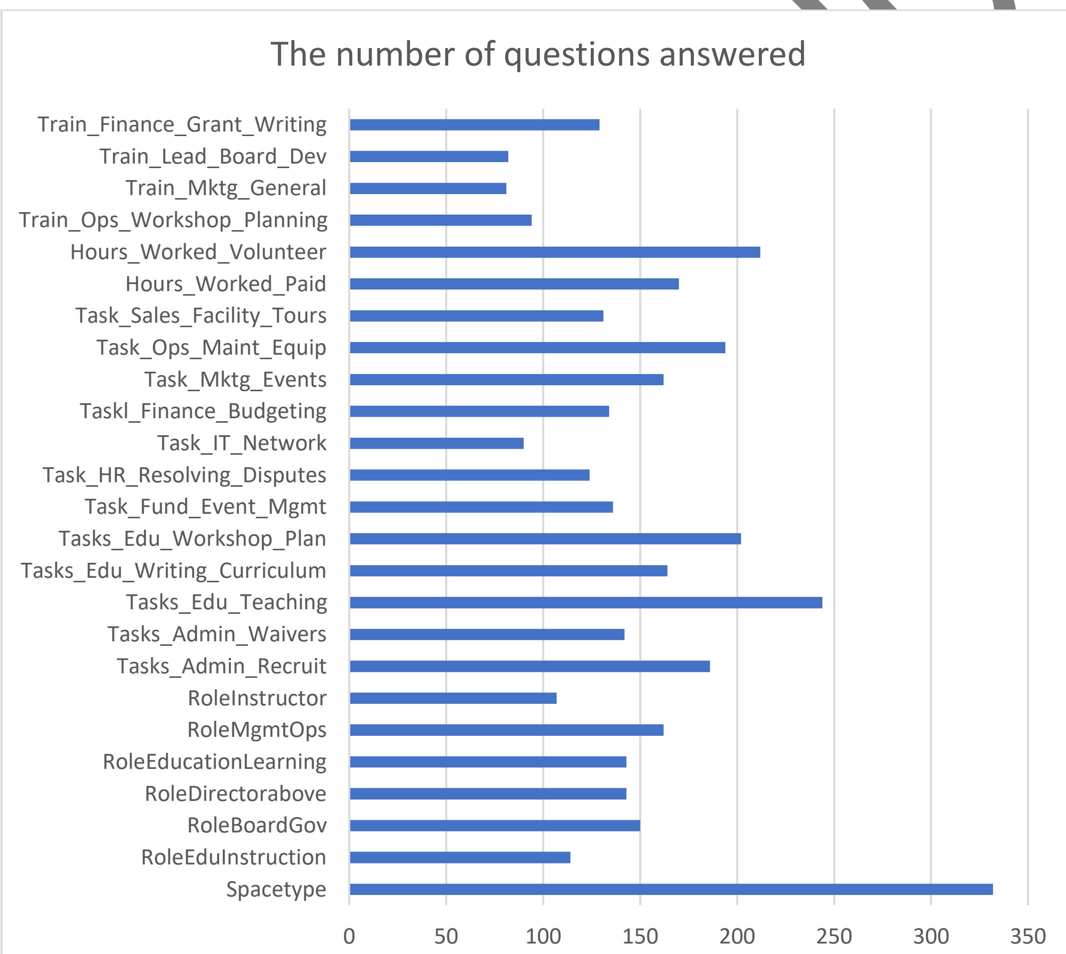

Fig. 2. Number of questions answered.

The fig. 2 gives overview of how each question is answered by the respondent.

Data distribution for all features,

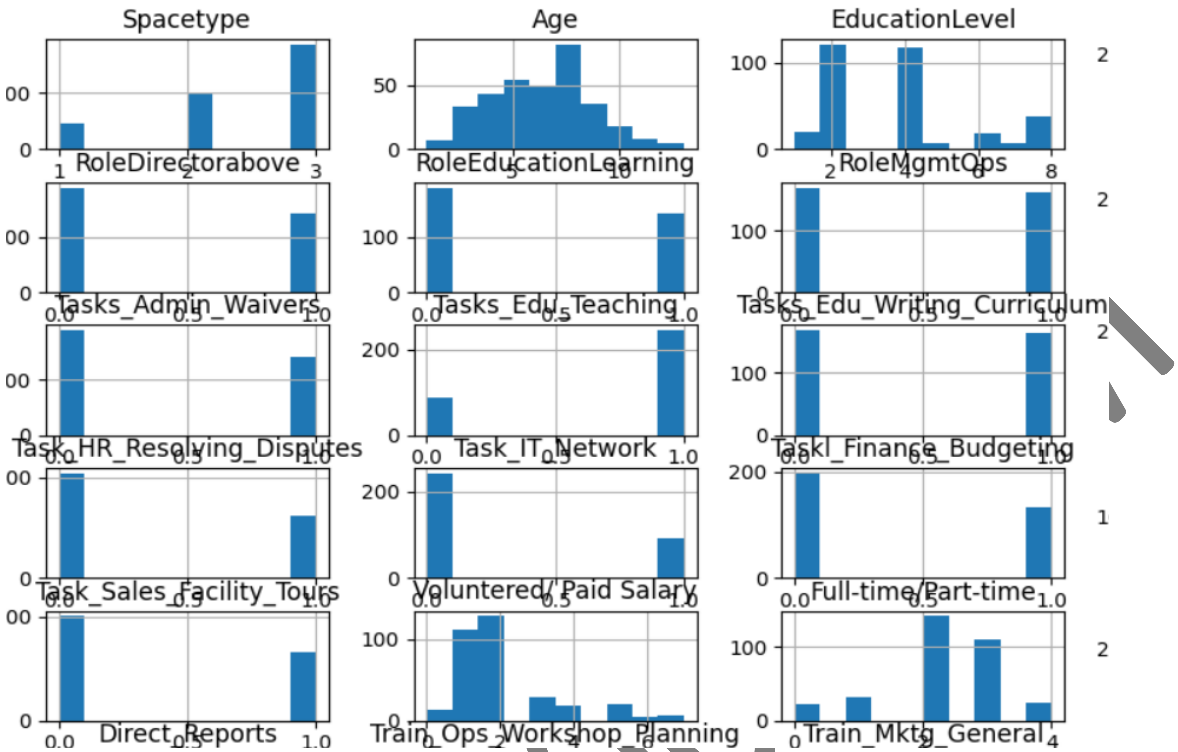

Fig. 3. Feature Distribution

fig.3 indicates most of the features are not in normalized format.

EDA

Summarize feature variables.

Table 2. Feature Summary Table

	Spacetype	Age	EducationLevel	RoleEduInstruction	RoleBoardGov	RoleDirectorabove
count	332.000000	332.000000	332.000000	332.000000	332.000000	332.000000
mean	2.421687	6.304217	3.731928	0.343373	0.451807	0.430723
std	0.722876	2.345062	2.057439	0.475552	0.498423	0.495925
min	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	2.000000	4.750000	2.000000	0.000000	0.000000	0.000000
50%	3.000000	6.000000	4.000000	0.000000	0.000000	0.000000
75%	3.000000	8.000000	4.000000	1.000000	1.000000	1.000000
max	3.000000	13.000000	8.000000	1.000000	1.000000	1.000000

8 rows × 30 columns

Table: Snippet of the dataset

To find correlation between features and remove highly correlated features to prepare final dataset.

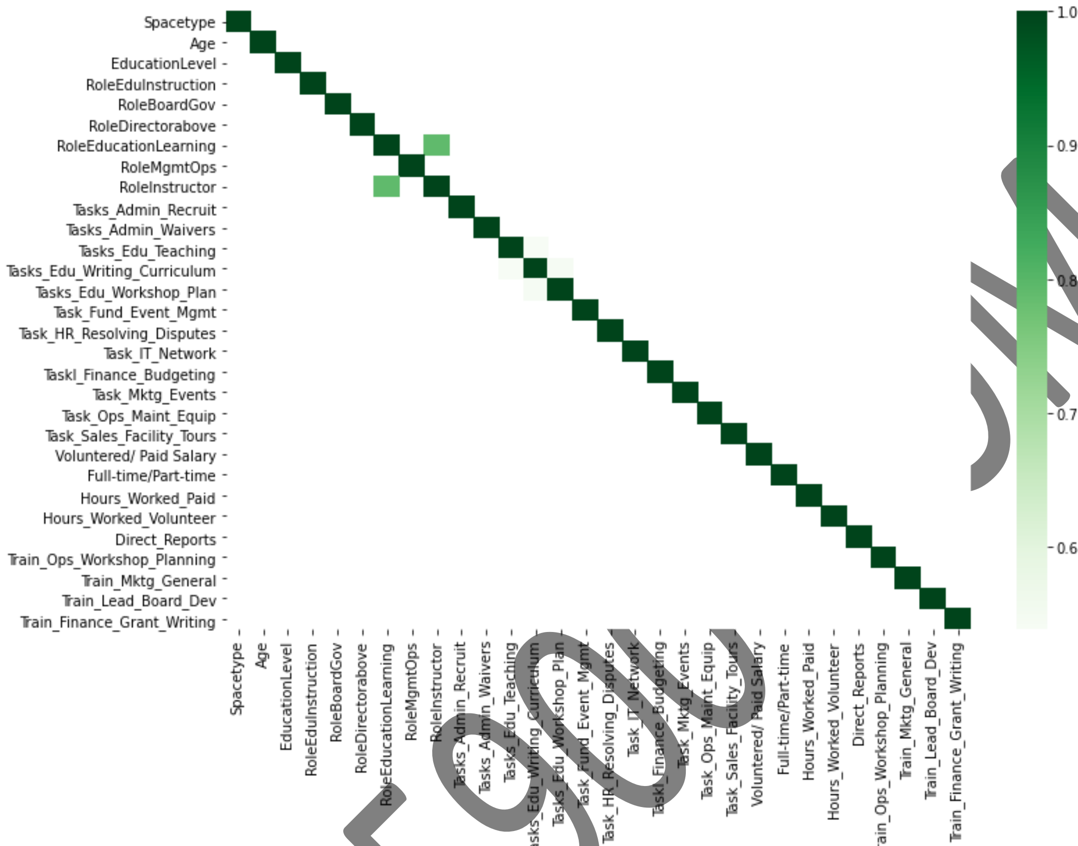

Fig. 4. Correlation between different leader questions

From fig 4. that Education Learning and Instructor roles are correlated. Hence, remove instructor role from the final dataset as that also presents less data points compared to educational learning.

1. Check outliers,

All continuous numeric variable is converted into ranges during data cleaning process. Hence, Outliers are not present in the data.

2. Identify skewness in the data.

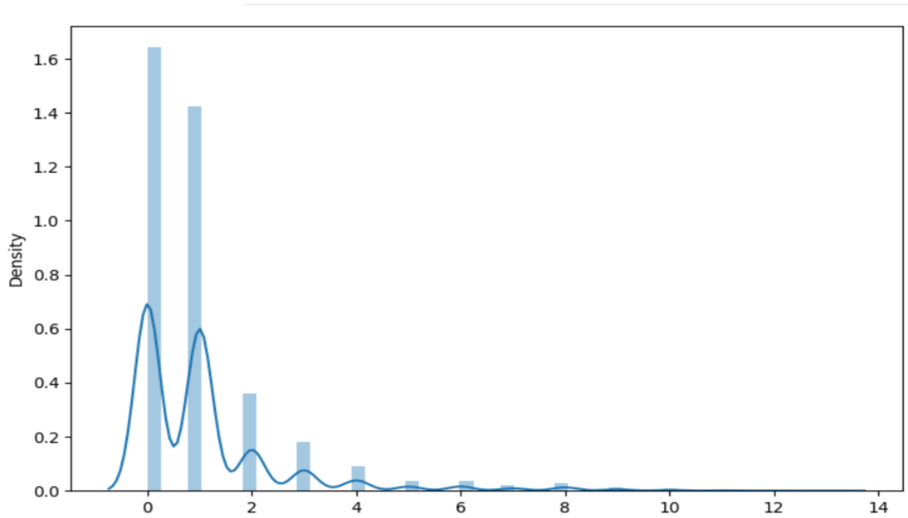

Fig. 5. Skewed Data

The data in fig.5 is right skewed. The transformation/ scaling techniques are required to normalize the data.

3. Transform/ Scale the data.

Scaling the features is a crucial step when the units of features differ from each other. In the given dataset it is evident that all attributes are in different ranges, scikit's Normalizer (norm = 'l2') is used for removing the mean and scaling to unit variance. There are several types of scaling available however maximum variance and normalized data are achieved using Normalizer (norm = 'l2') technique for this dataset.

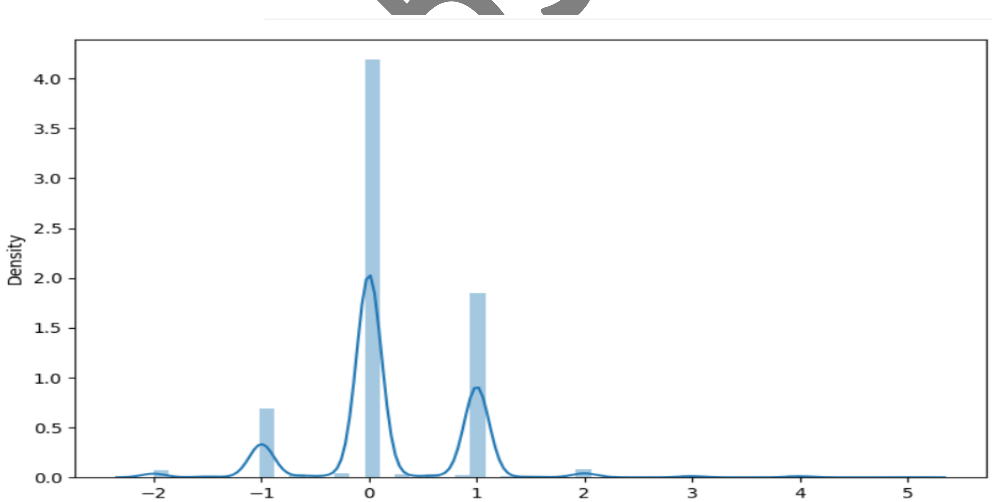

Fig. 6. Normalized Data

The plot in fig.6 represents normal distribution after scaling and transformation.

Methodology

a. Techniques,

The techniques include factor analysis to reduce dimensions from the data and clustering analysis to analyze patterns among different sections of the questionnaire.
Investigating patterns between different groups of questionnaires will help quickly understand best practices and avoid pitfalls in carrying out the mission of empowerment among different makerspaces.
The requirement of the sponsor is to analyze cross referencing and pattern among different sections of the questionnaire and understanding dimensions of different questions.
Cluster analysis is useful as it helps in detecting the groups that naturally occur in survey's data set. For instance, to find out how economic activities such as unpaid volunteering is conducive to the burden on leadership. This is done by analyzing involvement, and behavior qualities of the leaders by forming them in different clusters and how they are highly correlated. This approach allows for investigation of complex interrelations between different questions in the surveys from multiple perspectives – individual, contextual, and group level.
The other objective is to understand questions pointing to similar dimensions. This can be achieved using factor analysis. Factor Analysis is a linear model and is used to explain the variability in observed and correlation between variables and condenses the variables to smaller set called factors. Each factor describes certain variance in observed variables. This can help to understand significance of each questions which tend to answer similar question. This is helpful to reduce insignificant questions or options from the questionnaire to reduce survey response time.

b. Procedure,

Cluster Analysis

Clustering algorithm and is an Unsupervised Learning technique. It is used to divide a group of data points into clusters, where each point in the cluster is similar to each other. Clustering is mainly used for finding groups of data that are all similar. The goal of this analysis is to,

- Summarize data.
- Partition data
- Explore data.
- Find patterns in the data.

The type of clustering algorithms,

K-mean Clustering - K means is an iterative clustering algorithm that aims to find local maxima in each iteration. The important terminologies associated with K- mean clustering is,

- Centroid - This is the center point of a cluster. If K=3 then we would have 3 centroids, or centers, one for each cluster.
- Distance Measure - Distance measures how similar two elements are and will influence the shape of the clusters.

Hierarchical Clustering - Hierarchical clustering is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The important terminologies associated with Hierarchical clustering is,

- Distance Measure - It measures distance between two points.
- Linkage - It Measures the distance between two clusters.
- Dendrogram - Dendrograms can help decide the optimal number of clusters for the dataset by showing explicitly the hierarchy of the clusters.

K-mean clustering and Agglomerative Clustering are used in this study for analyzing economy, member, and leaders survey dataset. Goal is to group the responses based on the similarity of their answers on the survey. Initially we are unsure how many clusters (group) of makerspaces will be. Using different methods and hyper tuning of clustering we can decide the best "natural" number of groups of this dataset. The clusters will need to follow observation's pattern to be "natural".

The following steps are followed for cluster analysis,

1. Load the scaled dataset.
2. Reduce the dimensions using Principal Component Analysis (PCA).
3. 25 features will be in 25-dimensional space. It will be very hard to visualize (and understand) this many dimensions hence Principal Component Analysis (PCA) will be used for dimensionality reduction.
4. PCA can help to identify patterns based on the correlation between features. This algorithm aims to find maximum variance using fewer dimension than the original data.
5. Use PCA to visualize the clustering result.
6. Compare the clustering result with and without PCA.
7. Compare K-Mean and Agglomerative Clustering to obtain accurate groups of clusters.

K-Mean clustering using Elbow method and Silhouette score.

- k-Means starts by choosing k random centers.
- All data points are assigned to the closest center based on their Euclidean distance.
- New centers are calculated, and the data points are updated.
- This process continuous until clusters do not change between iterations.

Define the number of clusters using Elbow method and using Hyperparameter tuning using the silhouette score method.
KMeans has problem where number of clusters need to be mentioned beforehand to explicitly inform the KMeans model about the number of clusters needed in data to be categorized.
Silhouette score method and elbow method are used to determine best possible clusters.

Agglomerative Clustering using dendrograms.

- For Agglomerative Clustering, each individual data point starts as an individual cluster.
- Merge each two closest clusters into a new combined cluster.
- Eventually, all points are combined into a single cluster.

8. Compare the result of K-Means vs Agglomerative Clustering techniques.
9. Do deeper analysis with K-selected clustering techniques.

Factor Analysis

Factor Analysis is a linear model and is used to explain the variability and correlation between variables and condenses the variables to smaller set called factors. Each factor describes certain variance in observed variables. This can help to understand significance of each questions which tend to answer similar question. This is helpful to reduce insignificant questions or options from the questionnaire to reduce survey response time.

Step 1: Load the data
Step 2: Preprocess Data
Step 3: Adequacy Test to check factorability of the dataset,

a) Bartlett's Test - Bartlett's test checks whether the correlation is present in the given data. It tests the null hypothesis (H0) that the correlation matrix is an Identical matrix. If the p test statistic value is less than 0.05 then dataset is factorable.
b) Kaiser-Meyer-Olkin Test (Kmo) - KMO estimates the proportion of variance among all the observed variables. Lower proportion id is more suitable for factor analysis. KMO values range between 0 and 1 and value of KMO less than 0.6 is considered inadequate.

Step 4: Perform Factor Analysis

Results and Analysis.

Comparison of K-Means and Agglomerative Clustering

For Member dataset,

Silhouette score using Agglomerative,

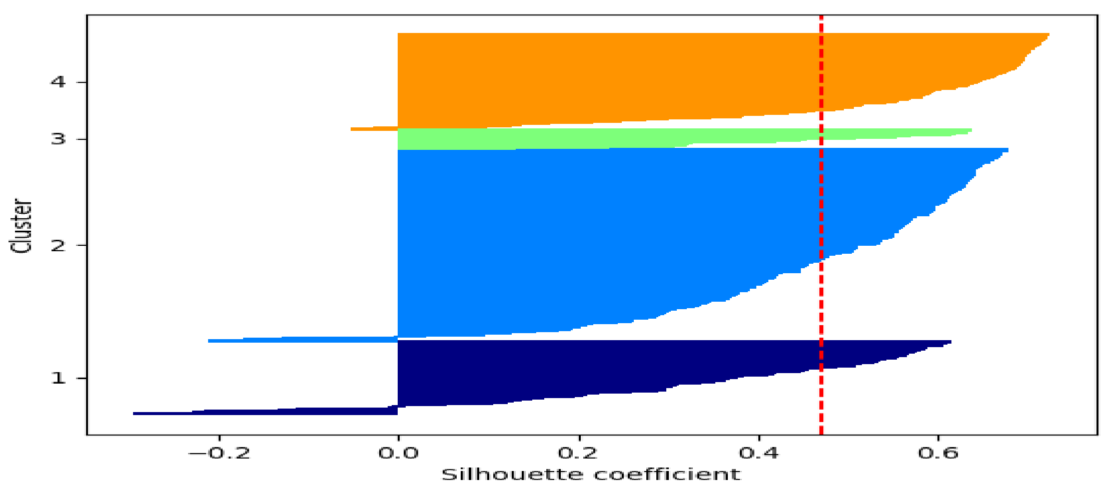

Fig. 7. Silhouette score - Agglomerative

Silhouette score using K-Mean,

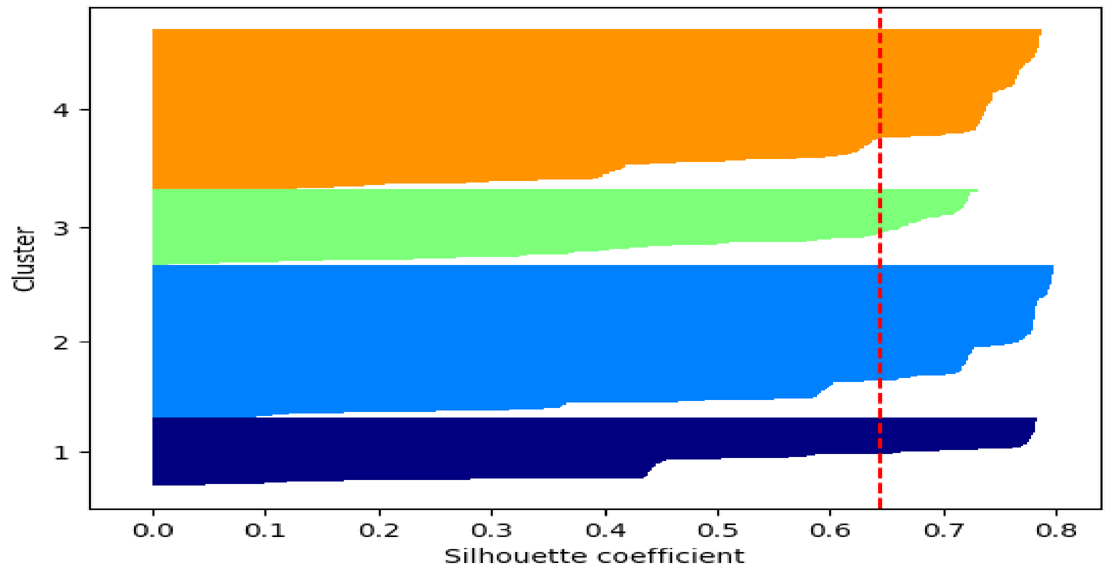

Fig. 8. Silhouette score - K mean.

From fig. 7 and fig.8, Silhouette score obtained using Agglomerative Clustering is lower than K mean hence, K-Mean clustering is selected for final analysis.

Results from K-mean cluster

1. Economy makerspace,

   Data: Features: 41, Observations: 104

   a. Finding optimal clusters using elbow method,

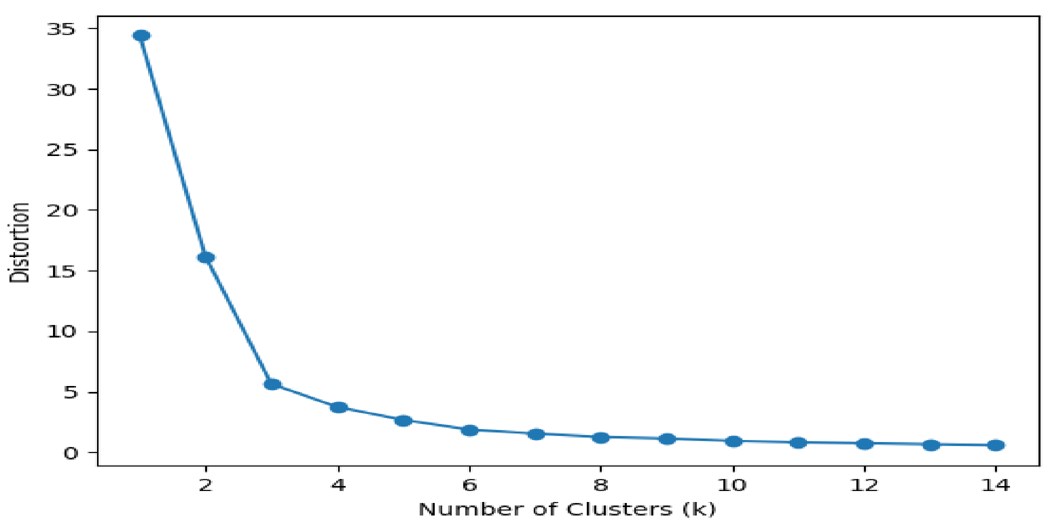

Fig 9. Economy - Elbow method to identify optimal cluster.

From fig 9. the optimal cluster is obtained at k= 3

b. Check Silhouette score for each cluster,

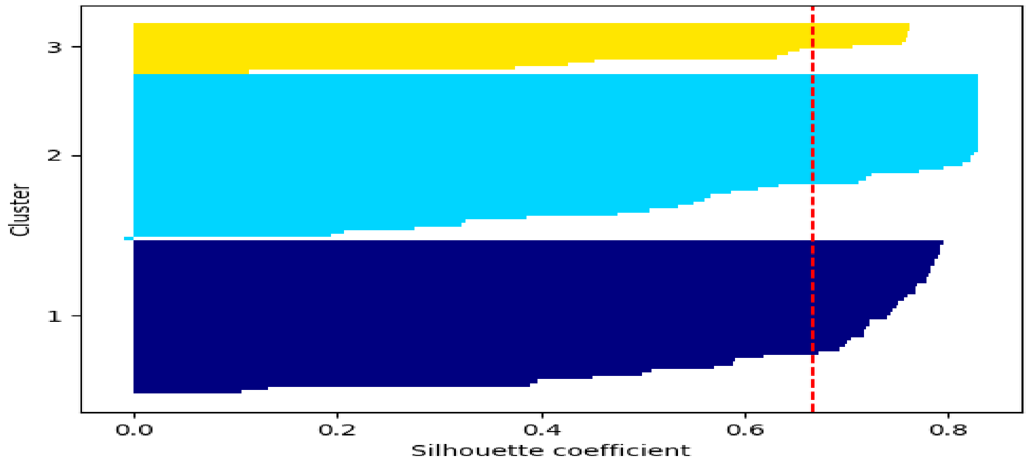

Fig 10. Economy – Silhouette for different cluster.

The Silhouette score is high for 2nd cluster.

c. Visualize clusters using PCA,

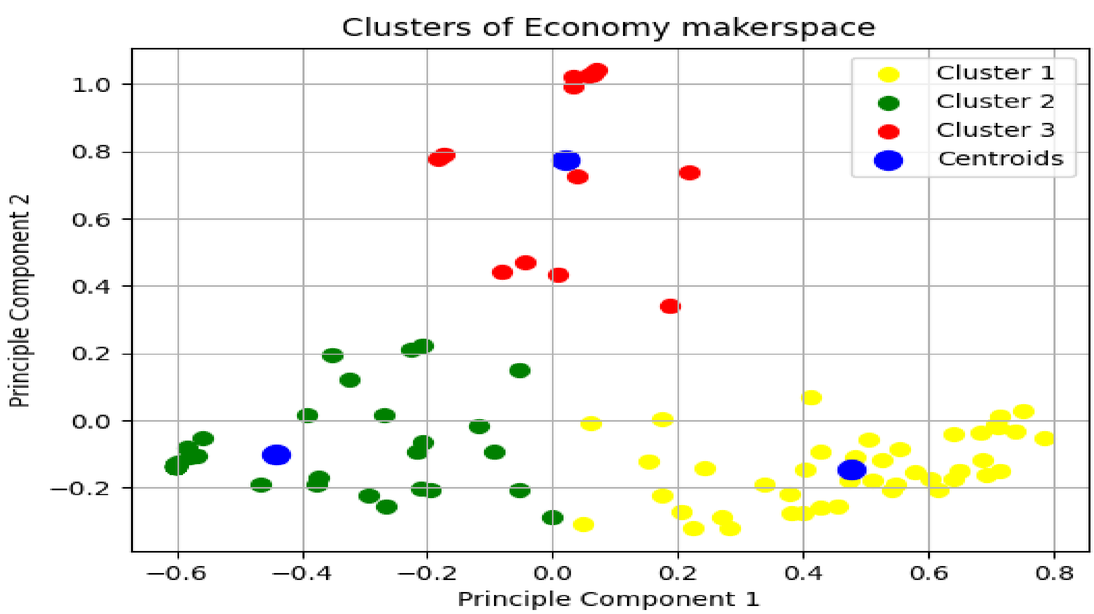

Fig. 11. Economy - Cluster Visualization

Statistics,
Variance Ratio: 0.63
63% of the variance is achieved which is not idle but due to less datapoints maximum variance is not possible to achieve.

K Means Result:
Cluster 0: 43, Cluster 1: 47, Cluster 2: 14
Final K Means Result (no PCA):
Cluster 0: 44, Cluster 1: 46, Cluster 2: 14
The results obtained from with PCA and without PCA are similar in distribution.
Cumulative explained variance PC1 - [0.41888447] PC2 - [0.21659333].
Ranking of the variables based on the variance from PC1,

Table 3. PC1 Variable Ranking - Economy

Ranking	Feature	Ranking	Feature
1	Num_Paid_Memberships	22	Programs_Adult_Classes
2	Response	23	Per_Rev_LY_Retail_Product_Sales
3	Per_Rev_LY_Membership	24	OpFunds_Y1_Founding_Members
4	Num_Free_Memberships	25	ServTools_Woodworking
5	Governance_Model	26	Programs_Tool_Orientation_Safety
6	Per_Rev_LY_Classes	27	ServTools_Electronics_Robotics
7	Funding_Started_With	28	ServTools_Sewing
8	Per_Rev_LY_Donations_Monetary	29	MemberLength_Annual
9	Value_Starting_Donations	30	ServTools_Large_Format_Printing_Vinyl
10	Per_Rev_LY_Donations_In_Kind	31	ServTools_Computer_Lab
11	Per_Rev_LY_DesignMake_For_Hire_Services	32	ServTools_Laser_Cutting
12	Per_Rev_LY_Studio_Table_Retals	33	ServTools_Fiber_Textiles
13	Per_Rev_LY_Grants	34	Memberships_Student
14	Spacetype_Grouped	35	ServTools_3D_Print_Scan
15	OpFunds_Y1_Grants	36	ServTools_Crafting
16	Per_Rev_LY_Loans	37	Programs_Club_Hosting
17	MemberLength_Monthly	38	Programs_Maker_Events
18	ServTools_Machine_Shop	39	OpFunds_Y1_Ind_Donations
19	ServTools_Welding	40	Per_Rev_LY_Event_Room_Rentals
20	Memberships_Individual	41	OpFunds_Y1_OpRevenue
21	Per_Rev_LY_Material_Fees		

2. Leader Makerspace

Data: Features: 30, Observations: 104

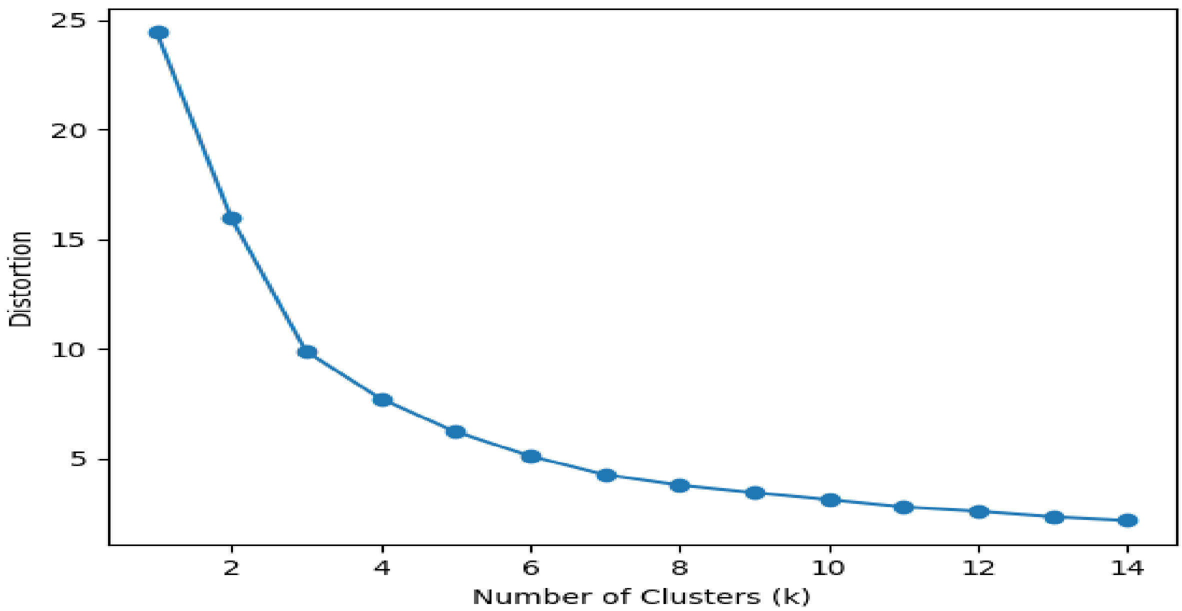

Fig 12. Leader - Elbow method to identify optimal cluster.

The optimal cluster is obtained at k= 6

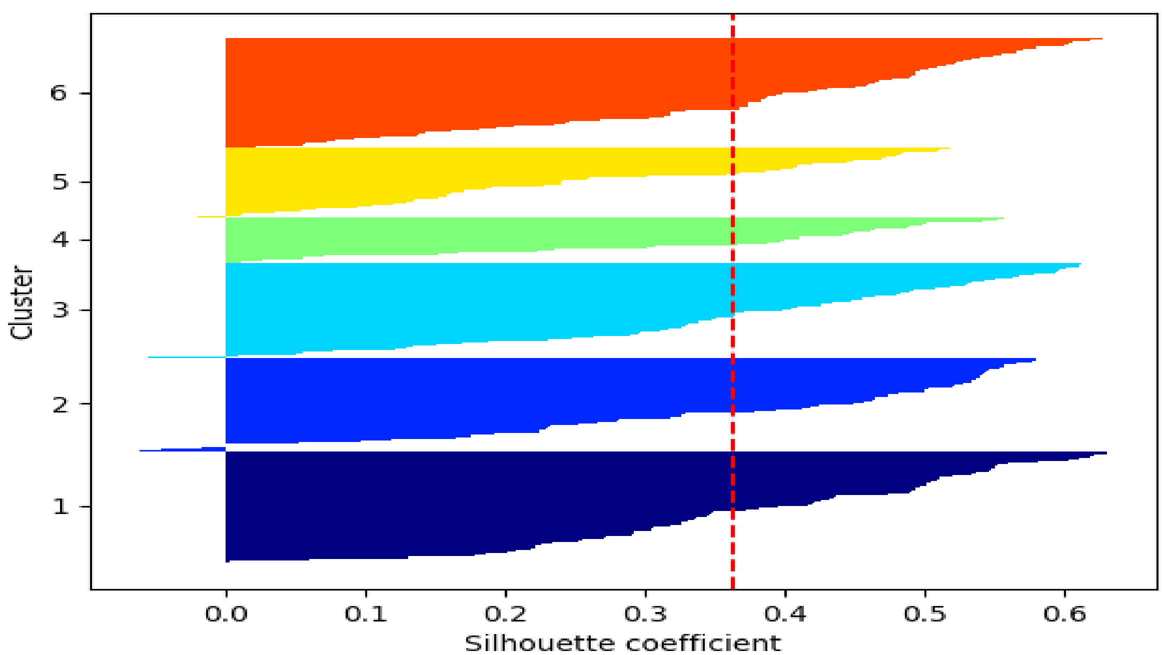

Fig 13. Leader – Silhouette for different cluster.

The Silhouette score is high for 1st, 2nd, and 6th cluster.

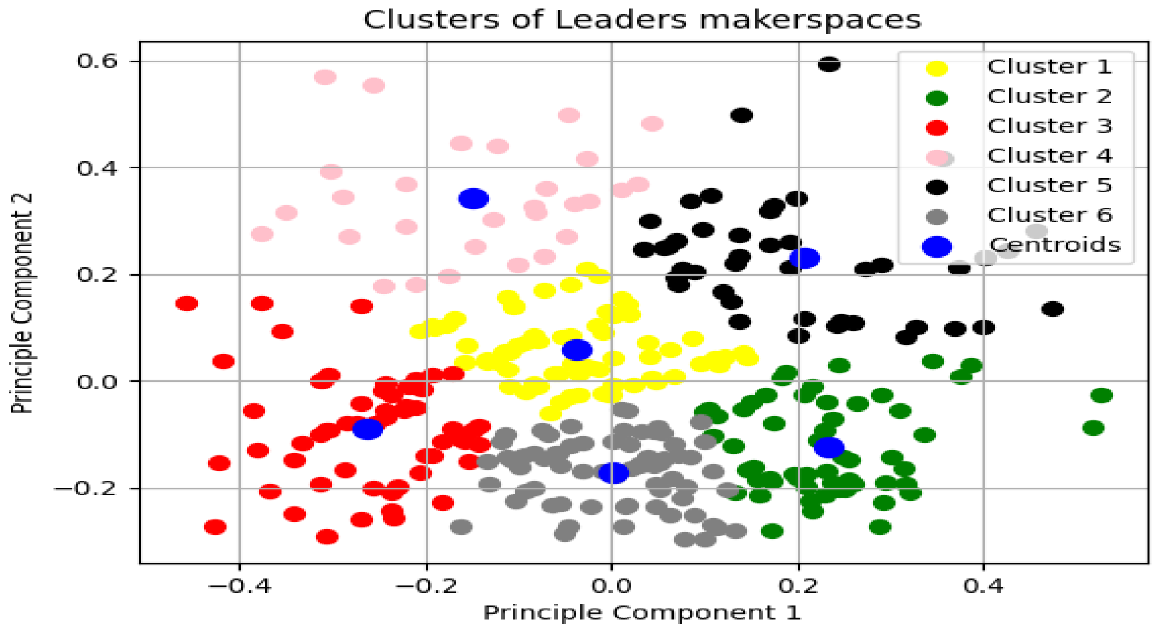

Fig. 14. Leader - Cluster Visualization

Statistics:
Results obtained from K-mean clustering with 6 clusters:
Explained Variance Ratio: 0.42
K Means Results:
Clusters ({0: 71, 5: 69, 2: 60, 1: 59, 4: 44, 3: 29})
Final K Means Result (no PCA):
Clusters ({5: 82, 1: 76, 0: 48, 2: 45, 3: 44, 4: 37})
Cumulative explained variance PC1 - [0.22385771] PC2- [0.20605753]
Ranking of the variables based on the variance from PC1,

Table 4. PC1 Variable Ranking - Leader

Ranking	Features	Ranking	Features
1	Age	16	Train_Finance_Grant_Writing
2	Direct_Reports	17	RoleDirectorabove
3	Voluntered/ Paid Salary	18	Task_HR_Resolving_Disputes
4	Full-time/Part-time	19	Train_Ops_Workshop_Planning
5	Hours_Worked_Paid	20	Taskl_Finance_Budgeting
6	Task_Ops_Maint_Equip	21	Task_IT_Network
7	RoleMgmtOps	22	Tasks_Admin_Recruit
8	Spacetype	23	RoleEducationLearning
9	Tasks_Edu_Teaching	24	Hours_Worked_Volunteer
10	Task_Fund_Event_Mgmt	25	EducationLevel
11	Tasks_Edu_Workshop_Plan	26	RoleInstructor
12	Task_Mktg_Events	27	Train_Mktg_General
13	Tasks_Admin_Waivers	28	Train_Lead_Board_Dev

14	Tasks_Edu_Writing_Curriculum	29	RoleBoardGov
15	Task_Sales_Facility_Tours	30	RoleEduInstruction

3. Member Makerspace,

   Data: Features: 31, Observations: 861

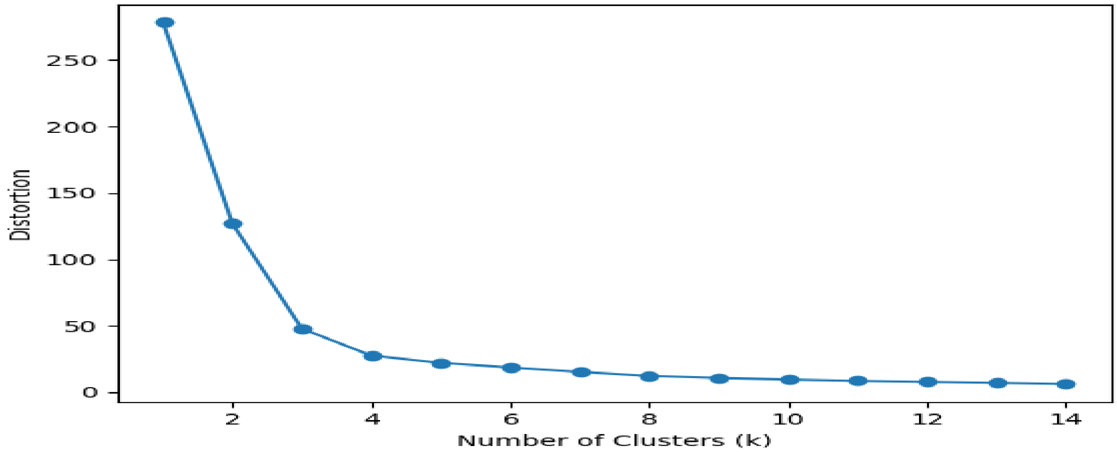

Fig 15. Member - Elbow method to identify optimal cluster.

The optimal cluster is obtained at k= 4

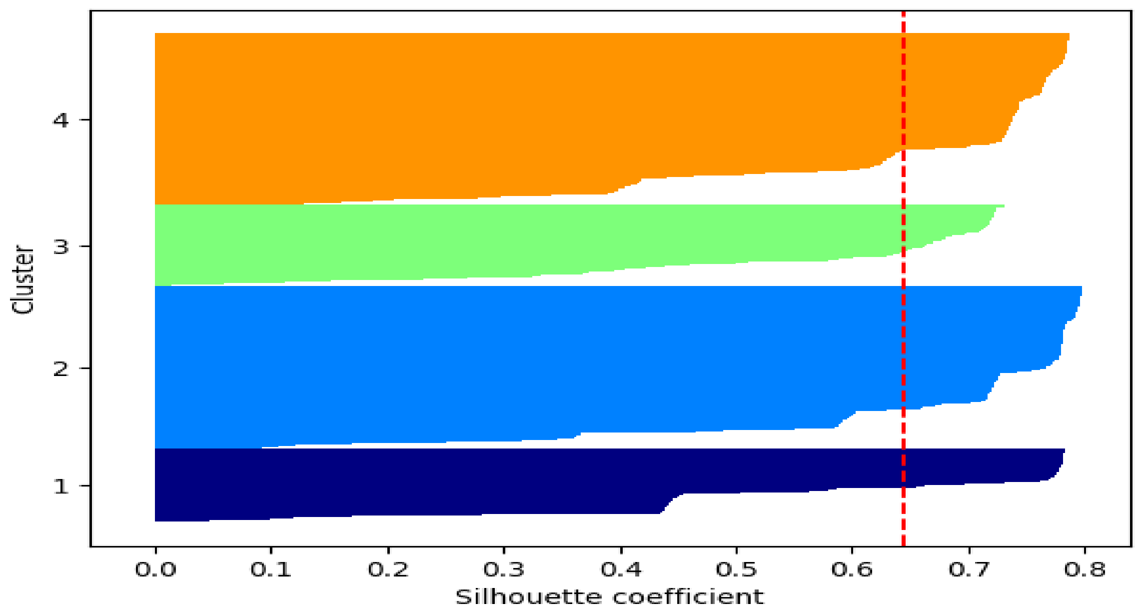

Fig 16. Member - Silhouette for different cluster.

The Silhouette score is high for 2nd and 4th cluster.

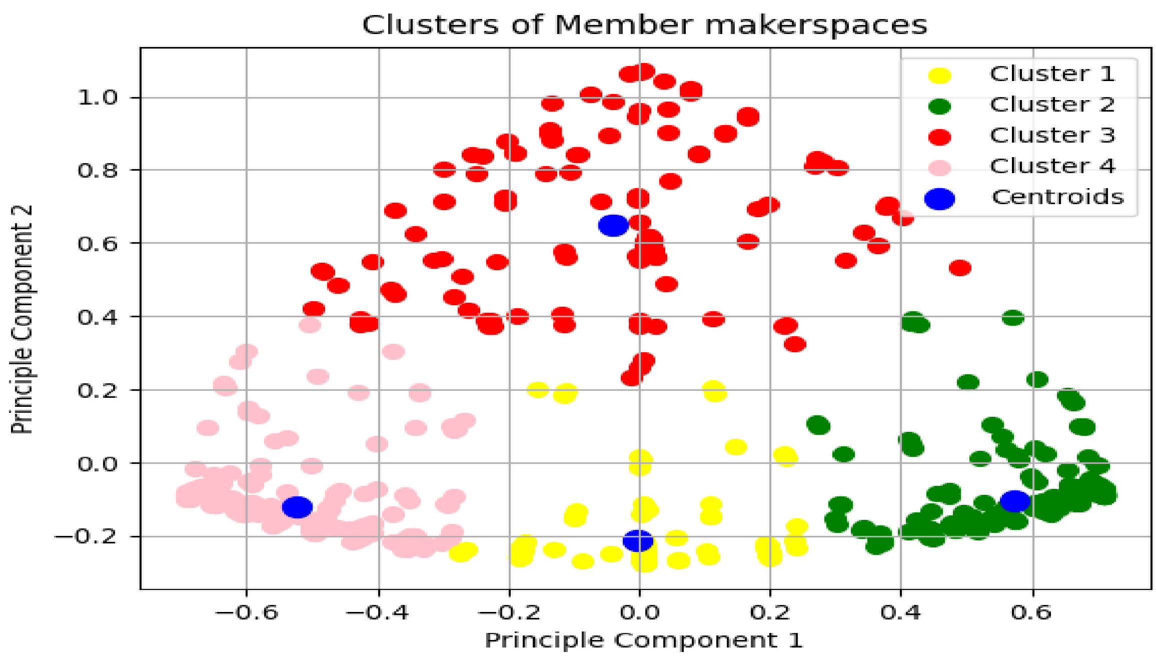

Fig. 17. Member - Cluster Visualization

Statistics:
Explained Variance Ratio: 0.80
K Means Result:
Clusters ({3: 302, 2: 287, 0: 143, 1: 128})
Final K Means Result (no PCA):
Clusters ({2: 398, 3: 309, 1: 109, 0: 44})
Cumulative explained variance PC1 - [0.55397388] PC2 [0.24848406]
Ranking of the variables based on the variance from PC1,

Table 5. PC1 Variable Ranking - Leader

Variables	Ranking	Variables	Ranking
Time_Home_Garage	1	ToolUse_3D_Printing	17
Time_Primary_Makerspace	2	ToolUse_Woodworking	18
Time_At_Work	3	ToolUse_Electronics_Robotics	19
UseFor_Project_Workspace	4	Make_To_Earn_Primary_Income	20
Age_Range	5	Bus_Read_Books	21
UseFor_Sharing_Knowledge	6	ToolUse_Machine_Shop	22
Frequency_Collab_In_Makerspace	7	Bus_Intend_To_Start	23
UseFor_Tool_Use	8	Make_Lo_Learn	24
UseFor_Learning_From_Others	9	Bus_Saving_For	25
UseFor_Learning_New_Things	10	Make_To_Express_Self	26
Highest_Edu	11	Gender	27
Spacetype_Grouped	12	ToolUse_Laser_Cutting	28

Bus_Plans_To_Launch	13	Frequency_Collab_Outside_Makerspace	29
Bus_Search_For_Opps	14	Make_Things_I_Need	30
Bus_Learning	15	Make_To_Earn_Extra_Income	31
ToolUse_CNC	16		

With PCA gives best result for membership cluster.
Best results were obtained using K-mean clustering for all three makerspaces.

Results from Factor analysis

For Member dataset
Bartlett's Test: p-value: 0
Kaiser-Meyer-Olkin Test: Kmo value: 0.82

```
                           0          1
Variance            4.719053   4.547674
Proportional Var    0.147470   0.142115
Cumulative Var      0.147470   0.289585
```

Fig. 18. Result snippet - Factor analysis

Analysis using Cluster technique,

The analysis is done by answering critical research questions from the sponsor.

Q. Burden on Leader who tend to perform a greater number of tasks,

Table 6. Leader - Cluster analysis

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Paid a Salary	22	40	0	5	2	7
Strictly a volunteer	5	2	0	11	0	26
Full- Time	21	34	8	5	1	6
Part-Time	0	4	6	2	1	6
I am not employed by the makerspace	0	4	2	3	4	24
Direct Report - No	6	1	4	8	1	20
Yes, 6-11 people	17	27	1	6	2	0
Yes, 1-5 people	0	4	9	3	0	2
Yes, 11-2 people	4	12	2	4	0	4
Count of all tasks	30	58	21	23	3	42

From the table 6,

Common Characteristics,

1. Leaders who perform more tasks are most likely to be paid salary, work full time and have more people reporting to them. Also, Volunteer, having no people reporting to them and not employed tend to perform a greater number of tasks.

Other Common characteristics from clusters,

Cluster 1: Leader who work full-time, and paid salary are most likely to have more people reporting to them.

Cluster 5: Leaders who strictly volunteer is more likely to be unemployed and have a smaller number of people reporting to them.

Factor Analysis to identify questions pointing to similar dimensions,

For member dataset,

Table 7. Member - Factor analysis

Features	Factor 0	Factor 1
UseFor_Tool_Use	0.07	0.61
UseFor Learning New Things	0.11	0.69
UseFor Learning From Others	0.17	0.72
UseFor Sharing Knowledge	0.19	0.70
UseFor Project Workspace	0.14	0.58
Frequency Collab In Makerspace	0.22	0.67
Business Intend To Start	0.84	0.18
Business Search For Opps	0.81	0.20
Business Saving For	0.72	0.09
Business Read Books	0.81	0.13
Business Plans To Launch	0.89	0.10
Business Learning	0.89	0.15

From the above table, Factor 0 has high loading for questions pointing to purpose of joining makerspaces while Factor 1 has high loading for questions pointing to business related activities.

These features/ questions point to similar dimensions and can be grouped into one category. Reducing number of questions/ options from survey can reduce survey response time.

Analysis using crosstab - python, PowerBI and SQL queries,

Q. Percentages of time spent by members at makerspaces and different location to make based on age,

Table 8. Cross Tabulation: Time spent vs age.

Age_Range	Primary Makerspace	Home Garage	Time at Work	Time Other Makerspace
13-16 years old	69.44	27.50	16.67	0.00
17-19 years old	71.66	30.00	7.00	2.00
20-24 years old	64.78	26.33	15.95	1.79
25-29 years old	47.53	44.39	24.07	4.20
30-34 years old	45.77	40.13	22.69	7.64
35-39 years old	45.99	48.67	18.58	6.85
40-44 years old	42.81	51.88	21.17	1.65
45-49 years old	40.01	53.53	24.00	4.08
50-54 years old	37.43	51.00	27.46	6.33
55-59 years old	43.01	51.25	28.70	6.80
60-64 years old	44.89	53.88	14.88	4.17
65-69 years old	36.02	61.76	18.88	7.29
70-74 years old	38.20	57.70	10.00	13.75
75-79 years old	33.33	60.00	80.00	20.00
80+ years old	30.00	30.00	0	0

From the above table we can conclude that younger people are more likely to spend their time at makerspaces as they do not have money to buy tools and other materials to make things. Older people between range of 40-80 years tend to spend more time at garage to make things.

Q. Membership and programming structure where membership is the number 1 revenue source for the economy spaces,

a. Paid vs Free Members,

Table 9. Membership fee structure

% of paid membership	94.56
% of free membership	5.44

Table 10. Membership classification

Member Classification	Count
Individual	36
Student	22
Family	20
Corporate	11
Other	10
Group	8
Senior	8
Household	3
No Membership	0

Table 11. Length of the membership

Length of membership	Count
Monthly	38
Annual	24
Daily	3
Other	2
Hourly	0
No membership offered	0

Economy makerspaces where number 1 revenue is from membership has majority of the paid members and likely to have long memberships.

Q. Hardest thing for making makerspace success - Open ended question,

Fig. 19. Word cloud for challenges in makerspaces

From the fig. 19-word cloud, it can be observed that,

Top challenges in making makerspace successful are 1. Getting people 2. Space 3. Keeping Members and 4. funding

Deliverables

The objective of this study is not to repeat the analysis which are already available with NOM and focus on cross referencing between different questions. Using clustering, patterns among different questions from the questionnaire can be analyzed. Without clustering it becomes harder to take individual features and analyze them for similar behavior/ pattern. The other deliverable factor analysis can help reduce questions/ responses from questionnaire which points to similar dimensions. The results of the clusters analysis, factor analysis and other analysis using crosstab will be shared with the sponsor for further analysis and answer any possible cross- referencing question they might have in the future. The code files, analysis done without clusters will be shared in the word format and visualizations in image format.

References

Questionnaire reference - https://makethedata.org/take-the-survey/

Cluster Techniques - https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-ofclustering/

Python libraries - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html