Clustering Toastmasters Clubs with Machine Learning

By Ricky Soo

August 23rd, 2020

INTRODUCTION

Toastmasters is a non-profit organization that trains members in communication and leadership skills. Members organize regular meetings in "clubs" where they practise public speaking, presentation and impromptu speaking skills. Members follow an education program called Pathways, featuring experiential learning, self-paced learning, peer evaluation and mentoring.

Founded in California in 1924, Toastmasters has spread its wing to 143 countries. Now there are more than 358,000 members in more than 16,800 clubs worldwide. To manage the many clubs around, the clubs are divided hierarchically into regions, districts, divisions, and areas.

Typically, each club consists of around 20 members. 4 to 6 nearby clubs are organized into an "area". 3 to 6 areas constitute a "division". 6 to 12 divisions form a "district". About 10 districts make up a "region". There are now 14 regions worldwide.

The regions are numbered from Region 1 to Region 14. Districts are also numbered such as District 102 where I am active in. Divisions are named with alphabetical letters such as Division A and Division B. Areas are identified with a number after the Division name, such as Area A1 and Area A2. The information of all regions and districts can be found at https://www.toastmasters.org/~/media/35503AED4D20498FBDA2AA75559FF2E0.ashx

The fiscal year of Toastmasters starts in July and ends in June the next year. The year is more commonly called a "term". Each term, officers are elected or appointed to serve as District Officers to manage the clubs and develop their leadership skills. At the end of each term, each district has the option to "realign" clubs to group them in a way that helps to manage, market and strategize for the clubs, areas, divisions in the district.

THE PROJECT

There are Toastmasters clubs of various sizes and conditions. There are bigger clubs with more than 50 members, and there are clubs with only a few members. There are restricted clubs such as corporate clubs whose membership is open only to the employees of a sponsoring company, or university clubs which are open for students of a university, and there are community clubs where anyone 18 years old and above can join as a member. There are clubs which produce good results in membership growth and members' achievements, and there are clubs having challenges to recruit members or hold regular meetings. There are clubs that are close to one another geographically, and there are clubs which are located far from the others.

This project seeks to help District Officers to gain insights into clubs for the purpose of formulating strategies to grow and support the clubs, promoting the clubs to the general public and assisting in the realignment exercise at the end of every term. Toward this end, machine learning techniques are used to group similar clubs into clusters to learn the similarities and dissimilarities among the clubs.

The specific areas of benefits include:

- **Management and support** – District Officers might need different strategies to support clubs of different sizes and conditions to help them to be effective clubs serving their members.
- **Marketing** – Clubs need to formulate marketing strategies to recruit new members. Understanding the neighbourhood where a club is located can help gain insights on the potential market out there for new members.
- **Realignment** – The yearly realignment exercise at the end of term typically does not seek to group clubs of similar nature together to be fair to give District Officers a variety of experiences in leading them. But clustering the clubs could help identify the similar clubs and avoid aligning them together. However, the realignment does seek to group clubs that are near to one another for easier logistics.

The scope of this clustering project includes 89 Toastmasters clubs in District 102 located in the state of Selangor in Malaysia. This is the district where I served as a District Officer twice before and so is familiar to me. These clubs currently belong to Division B, C, D, E and H in District 102. A summary of all clubs in District 102 for the term 2019-2020 can be found at https://dashboards.toastmasters.org/2019-2020/Club.aspx?id=102

THE DATA

The data for this project comes mainly from 3 sources – Toastmasters web site, Foursquare data and domain knowledge.

Toastmasters web site – The Toastmasters web site publishes a public dashboard showing the performance reports of all clubs in all areas, all divisions, districts and regions. The web site also contains a club page for each club showing its name, location and contact information.

Foursquare data – With the location information, Foursquare API is used to understand the neighbourhood where the clubs are located. Some neighbourhoods are popular with places with many people checking in. The popularity can give an indication on how active people in the vicinity of the Toastmasters clubs, as such the potential members who might be able to visit the clubs and join as members.

Domain knowledge – As a two-time District Officer myself for the terms 2017-18 and 2019-2020, I have gained sufficient knowledge into the working of Toastmasters clubs. This helps me to identify possible errors in the data, the potential club features that go into explaining the nature of the clubs, the similarities and dissimilarities, and the strategies that the clubs might need to be effective.

In this project, a dataset of 89 clubs is drawn up completed with data below (field name in bracket):

- **Club number (ClubNum)** – The club identification number of a club assigned by Toastmasters. Not used in modelling.
- **Club name (ClubName)** – The name of a club used to identify a club. Not used in modelling.
- **Members (Members)** – The number of members in the club. This gives an indication of the club membership strength.
- **New members (NewMembers)** – The number of new members joining the club in the current term. This gives an indication of the club marketing effort.
- **Increase in membership (Increase)** – The net increase or decrease in the membership number. This gives an indication of the club marketing effectiveness.
- **Goals achieved (Goals)** – The performance of each club is measured by the number of goals achieved in the Distinguished Club Program. The program consists of 10 goals and hence, this field can range from 0 to 10. This gives an indication of club quality of a club in serving its members.
- **Education awards (Education)** – The number of educational awards achieved by members of the club. This gives an indication of the level of activities of the members of a club.
- **Distinguished club status (Distinguished)** – A club is considered distinguished if it fulfils certain performance criteria. Specifically, a club is distinguished if it has achieved at least 5 goals, and it has at least 20 members (or has an increase of at least 5 members). This field on takes the value 0 or 1.
- **Open to public (Open)** – A community club opens its membership to the public to join. But a restricted club is only for the employees or students of a sponsoring company or university. A community club might have bigger market potential, but a restricted club might better align the club to the purpose of the sponsoring company or university. This field takes on the value 0 or 1.
- **Online attendance (Online)** – A club might allow meeting attendance by online means. A club with online attendance might attract members regardless of geographical boundaries. This field takes on the value 0 or 1.
- **Popular venues (Venues)** – The number of popular places within walking distance of 200 metres from the club location according to Foursquare. This gives an indication of the level of activities of the neighbourhood and hence the potential visitors to the club.

- **Location (Longitude, Latitude)** – The longitude and latitude of a club location to be used in clustering nearby clubs together.

All data is taken from the term-end result of the 2019-2020 term as on July 13th, 2020. The club data is taken from the publicly available data on Toastmasters web site.

DATA PREPARATION

The data collected from Toastmasters web site is not yet ready for the purpose of this project. Two more steps are required to build the complete dataset.

1. Fixing club locations – Some club locations are found to be incorrect. This is because some clubs have not indicated their location, or they have marked their location incorrectly.
2. Adding Venues column – The number of popular venues around the clubs need to come from Foursquare data. Foursquare API is used to extract the numbers to be added to the dataset.

To fix the club locations, first the original club locations are visualized using the Folium library as below. But the map is not showing the map of Selangor state of Malaysia.

Figure 1 - Club locations before correction

There are 14 known location errors in the data. To fix these, the Nominatim function is called for each correct club locations to get the correct coordinates. Then the correct latitudes and longitudes are updated into the dataset.

Below is the map of the clubs with location data corrected.

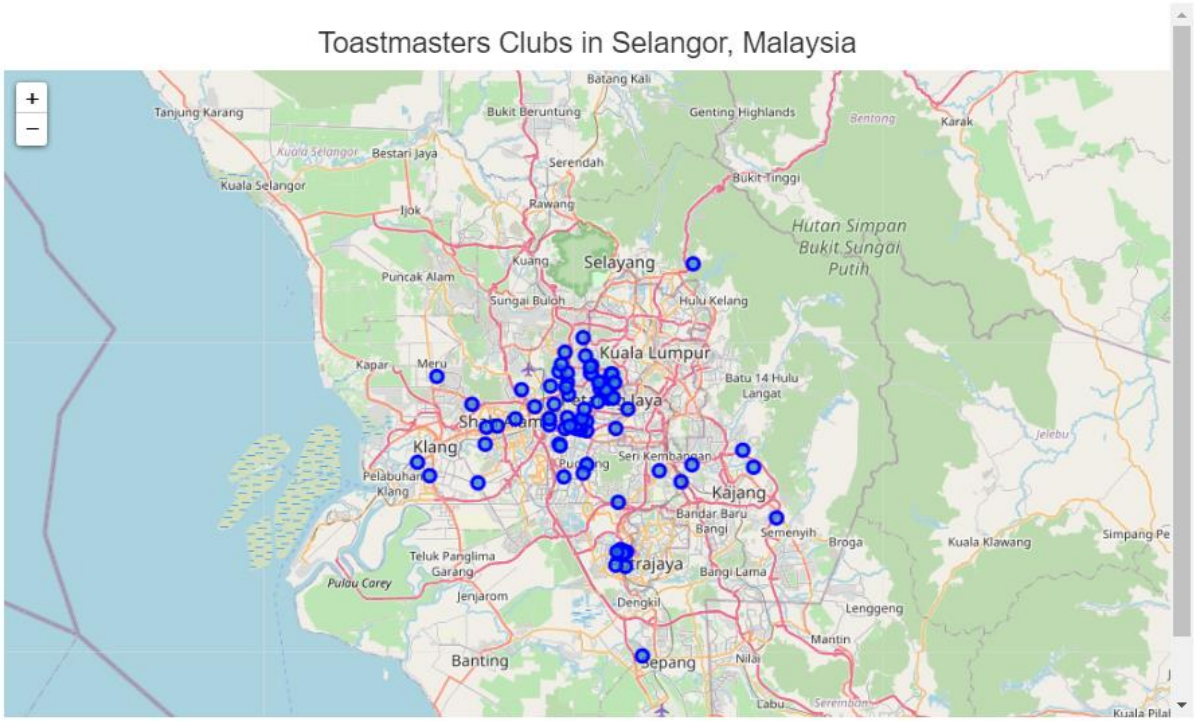

Figure 2 - Club locations after correction

Next, the Foursquare API is used to retrieve the number of popular venues within 200 metres of each club. The *explore* endpoint in Foursquare API is used. The API returns the data in JSON format. Then the "totalResults" is extracted from the data to indicate the number of popular venues.

There are altogether 1,093 popular venues found around 89 clubs. This shows the vibrant level of activities in the neighbourhoods where the clubs are located. These numbers are added into the dataset in the new Venues column. The dataset is now complete and ready for exploratory data analysis.

Insert into dataset as Venues column.

```
In [18]: df_clubs.insert(9, 'Venues', clubs_venues)
         df_clubs.head()
```

Out[18]:

ClubNum	ClubName	Members	NewMembers	Increase	Goals	Education	Distinguished	Open	Online	Venues	Latitude	Longitude
989051	Phoenix Toastmasters Club	25	11	1	10	15	1	1	1	25	3.136364	101.622000
1517575	Standard Chartered GBS Toastmasters Club	20	17	0	7	7	1	0	1	16	3.103077	101.638906
1558120	Taylor's Toastmasters Club	11	5	-5	6	7	0	0	0	16	3.062739	101.617100
7086903	Roche Malaysia Toastmasters Club	17	7	3	3	1	0	0	1	6	3.070388	101.610026
7511521	SIRIM Toastmasters Club	15	9	-5	4	0	0	0	1	3	3.067967	101.514620

Figure 3 - Venues column inserted

METHODOLOGY

First, exploratory data analysis is performed to understand the data. Then the feature set is determined. Multiple machine learning models is trained using k-means algorithm and the evaluation is done to decide on the right model.

Exploratory Data Analysis

The data prepared consists of 89 rows and 12 columns. Descriptive statistics techniques are performed to get a basic understanding of the data. There is no missing data found. However, some variables are found to have strong correlations with one another. These correlations might indicate strong dependencies among variables in the data features.

- "Distinguished" and "Members": This is expected. One criterion of being a distinguished club is to have least 20 members, or a net increase of at least 5 members in the club.
- "Distinguished" and "Goals": This is expected. One criterion of being a distinguished club is to have least 5 goals achieved.

There are also 3 categorical variables in the feature set. The analysis of variance (ANOVA) is performed to check whether they are important in the model or not. Each categorical variable is analysed against other variables and the means are calculated. If a categorical variable is

important, then the mean should differ significantly due to the difference in the categorical variable.

```
Distinguished and Members: F = 45.70, P = 0.00
Distinguished and NewMembers: F = 23.08, P = 0.00
Distinguished and Increase: F = 16.46, P = 0.00
Distinguished and Goals: F = 118.76, P = 0.00
Distinguished and Education: F = 8.58, P = 0.00
Distinguished and Open: F = 5.93, P = 0.02
Distinguished and Online: F = 10.29, P = 0.00
Distinguished and Venues: F = 0.29, P = 0.59
Distinguished and Latitude: F = 1.88, P = 0.17
Distinguished and Longitude: F = 0.49, P = 0.48

Open and Members: F = 0.20, P = 0.66
Open and NewMembers: F = 1.09, P = 0.30
Open and Increase: F = 0.40, P = 0.53
Open and Goals: F = 17.02, P = 0.00
Open and Education: F = 3.79, P = 0.05
Open and Distinguished: F = 5.93, P = 0.02
Open and Online: F = 15.63, P = 0.00
Open and Venues: F = 0.26, P = 0.61
Open and Latitude: F = 0.27, P = 0.61
Open and Longitude: F = 0.20, P = 0.65

Online and Members: F = 0.77, P = 0.38
Online and NewMembers: F = 13.08, P = 0.00
Online and Increase: F = 0.77, P = 0.38
Online and Goals: F = 19.90, P = 0.00
Online and Education: F = 3.57, P = 0.06
Online and Distinguished: F = 10.29, P = 0.00
Online and Open: F = 15.63, P = 0.00
Online and Venues: F = 1.07, P = 0.30
Online and Latitude: F = 3.17, P = 0.08
Online and Longitude: F = 0.00, P = 1.00
```

Figure 4 - ANOVA on Categorical Variables

All 3 categorical variables are found not significant in relation to "Venues" as the p-value is large. This is expected as most Toastmasters clubs typically operate independently from what is happening in the neighbourhood.

All 3 categorical variables are also found not significant in relation to "Latitude" and "Longitude". This is expected as the nature of clubs does not typically correlate with the club locations.

Both "Open" and "Online" are found not significant in relation to other variables in the feature set too. As such, they are candidates to be discarded from the feature set.

Then, histograms and box plots are used to visualize the data distributions.

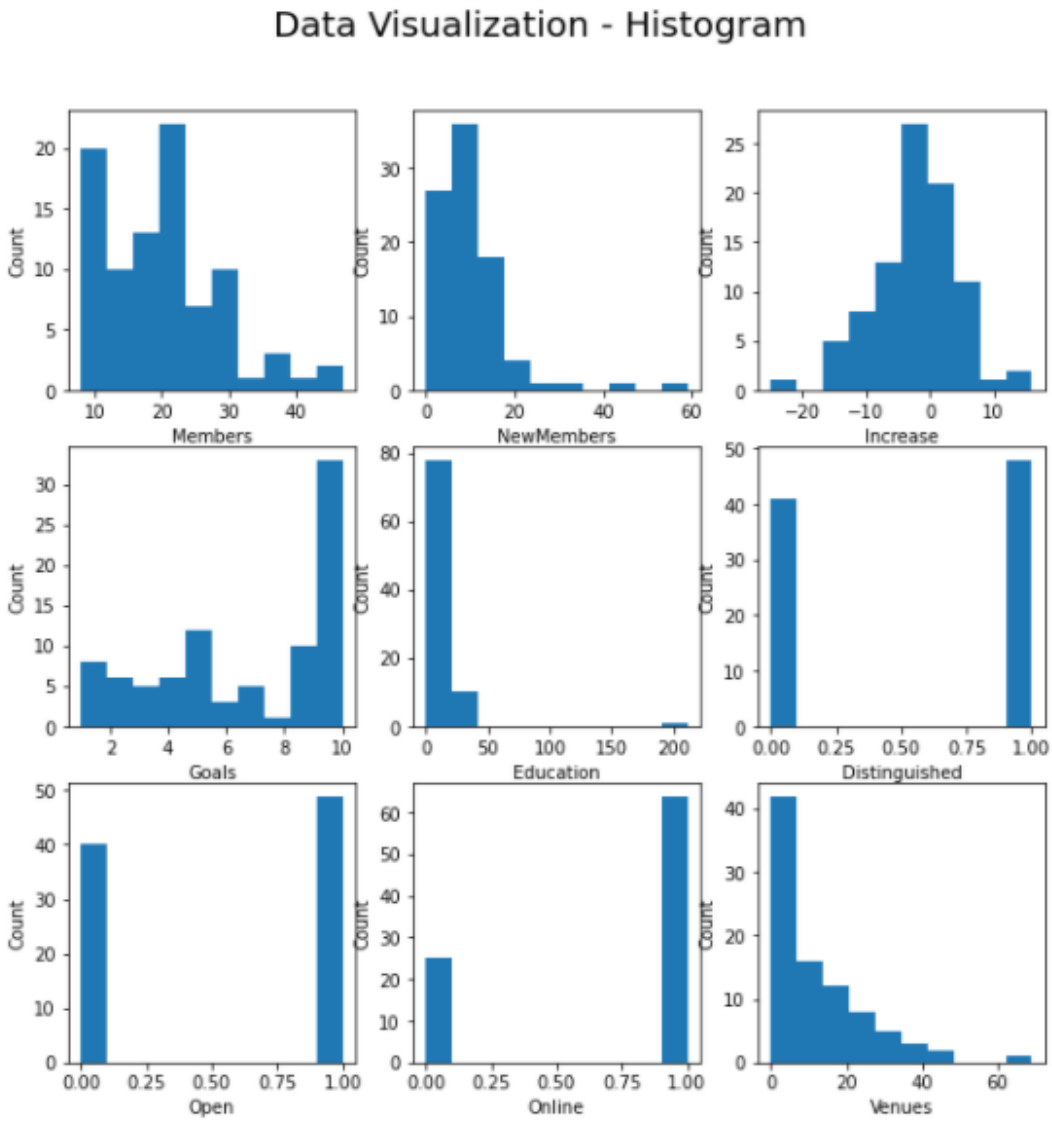

Figure 5 - Data Visualization using Histograms

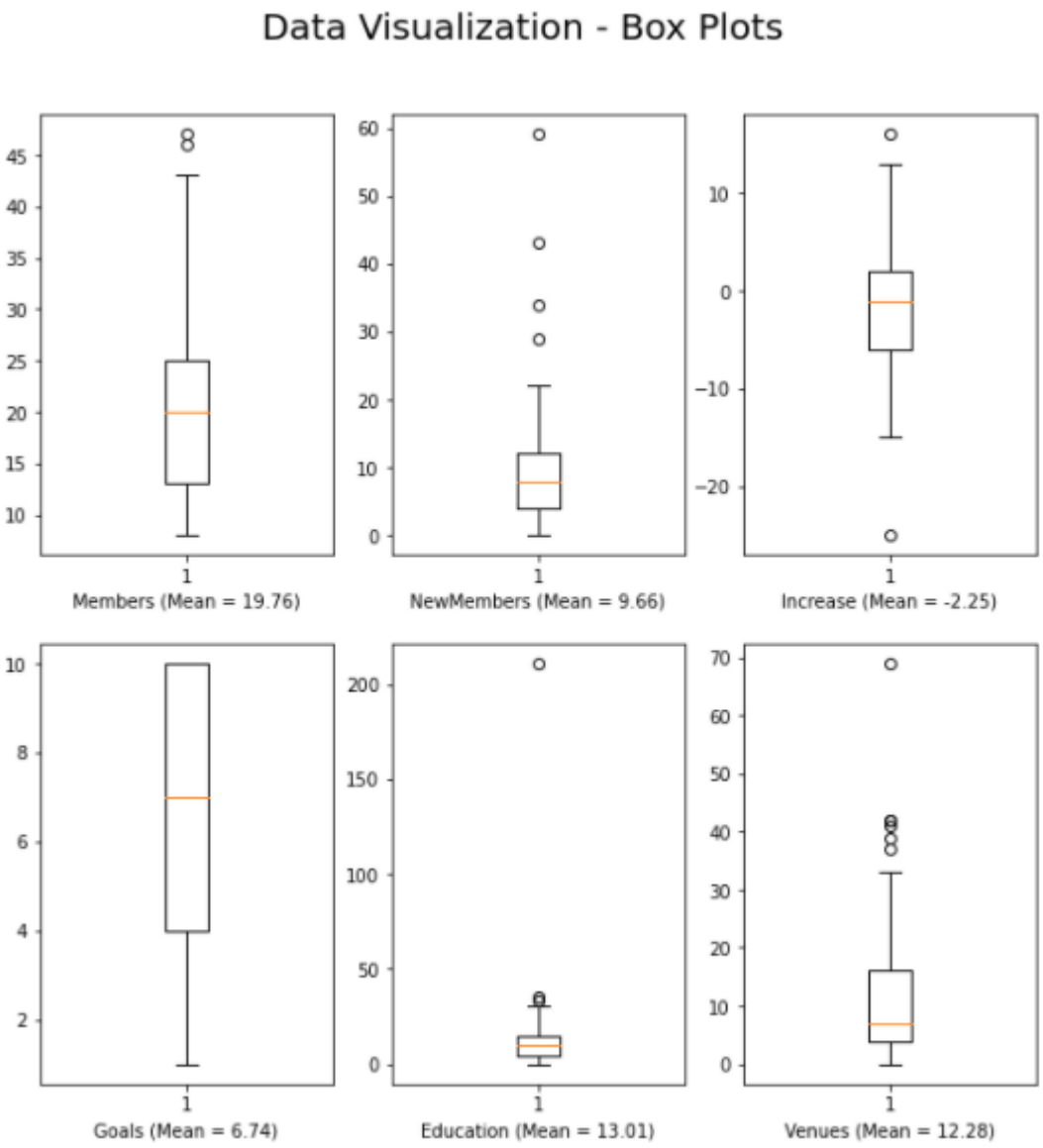

Figure 6 - Data Visualization using Box Plots

Understanding the nature of the data above will be useful in feature selection and model training next.

Feature Selection

The next question is whether the complete feature set should be used to train the model. Or some variables should be removed from the set? There are 3 possibilities:

- Complete feature set
- Without categorical variables

- Minimum feature set

We will try to build the model using k-means algorithm for each possibility. First, we loop k from 2 through 10. Then, the sum of squared errors (SSE) is calculated. Next, the elbow method is used to determine the best k. Lastly, Silhouette analysis is used to validate the clusters generated.

The k-means algorithm is used because it is simple to implement, it can scale to large dataset, and it guarantees convergence. Its weakness of having to find the best k (number of clusters) can be mitigated by using the elbow method and Silhouette analysis.

From the results, then we make a sensible decision on the feature set, as well as the best k to use.

Option 1 – Complete Feature Set

This feature set consists of all 11 variables collected in the data, namely, 'Members', 'NewMembers', 'Increase', 'Goals', 'Education', 'Distinguished', 'Open', 'Online', 'Venues', 'Latitude', 'Longitude'.

To test this feature set, we use the k-means algorithm to train the data using the k value from 2 to 10. Then, we examine the sum of squared errors (SSE), use the elbow method to determine the best k, and then use Silhouette analysis to draw insights into the feature set.

Below are the analysis results:

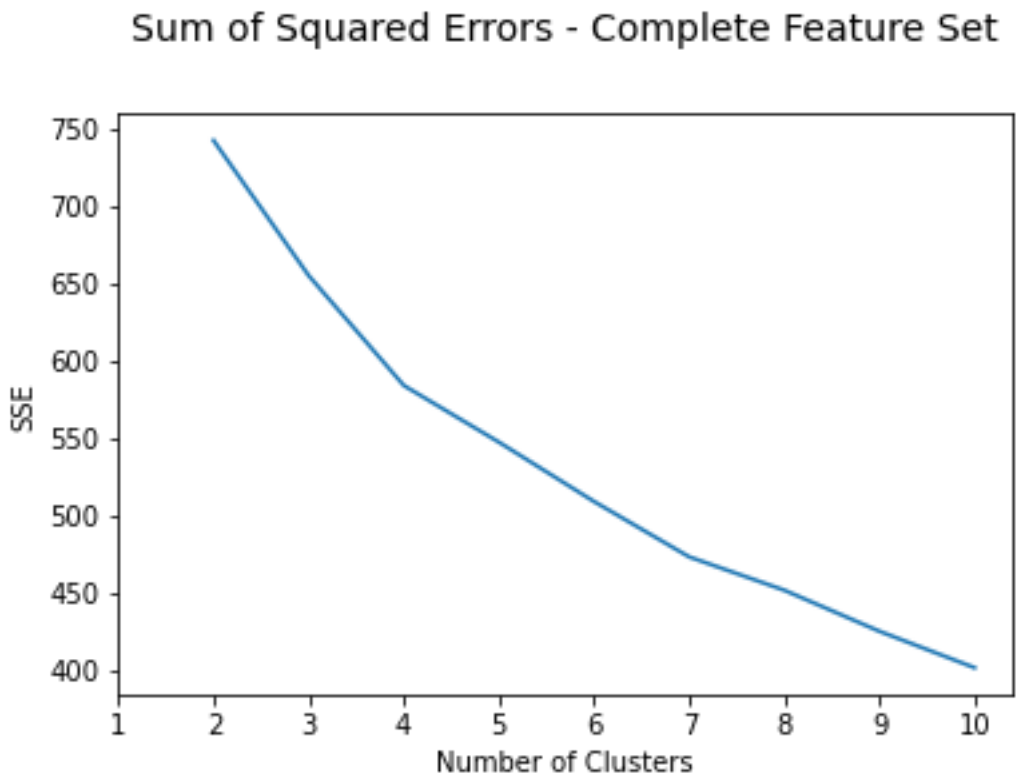

Figure 7 - Sum of Squared Errors - Complete Feature Set

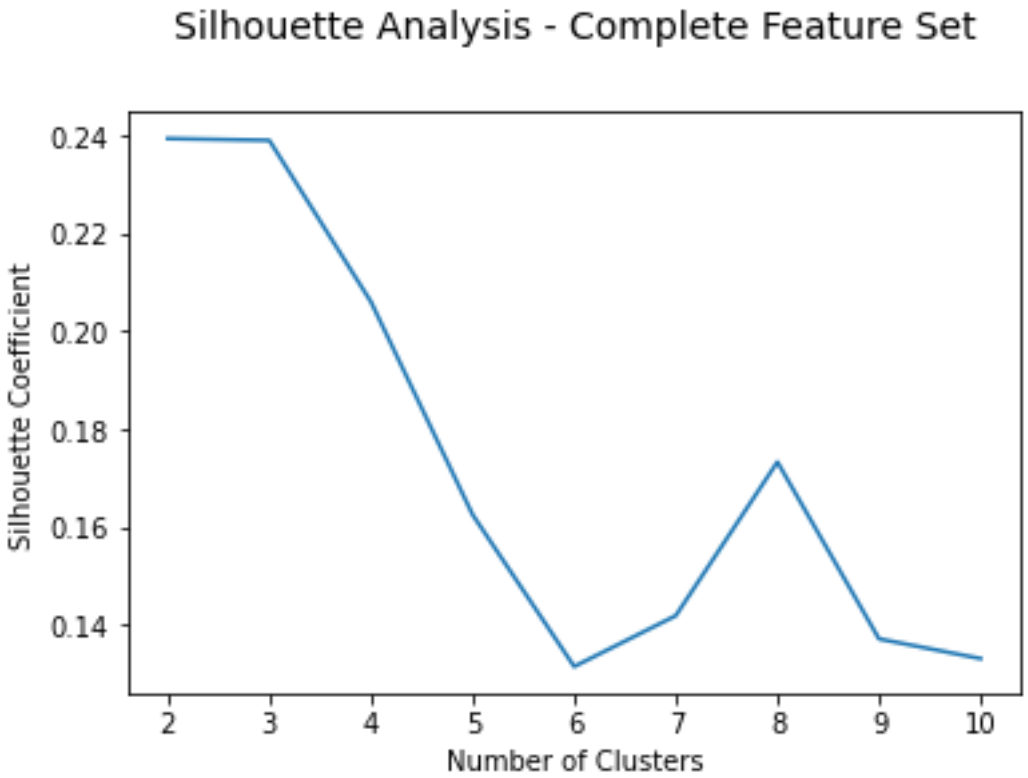

Figure 8 - Silhouette Analysis - Complete Feature Set

The best number of clusters (k) is found to be 4 but its Silhouette coefficient is not the best in graph above.

Option 2 – Without Categorical Variables

Now the dataset without the 3 categorical variables are to be tested.

- "Distinguished" is removed because it is highly correlated to "Members" and "Goals", according to correlation analysis above. The nature of a distinguished club might already be well-explained by these variables.
- "Open" and "Online" are removed because they are not significant against some other variables, according to ANOVA analysis above.

The resulting feature set consists of 8 variables, that is, 'Members', 'NewMembers', 'Increase', 'Goals', 'Education', 'Venues', 'Latitude', 'Longitude'. The same tests are done, and these are the results produced.

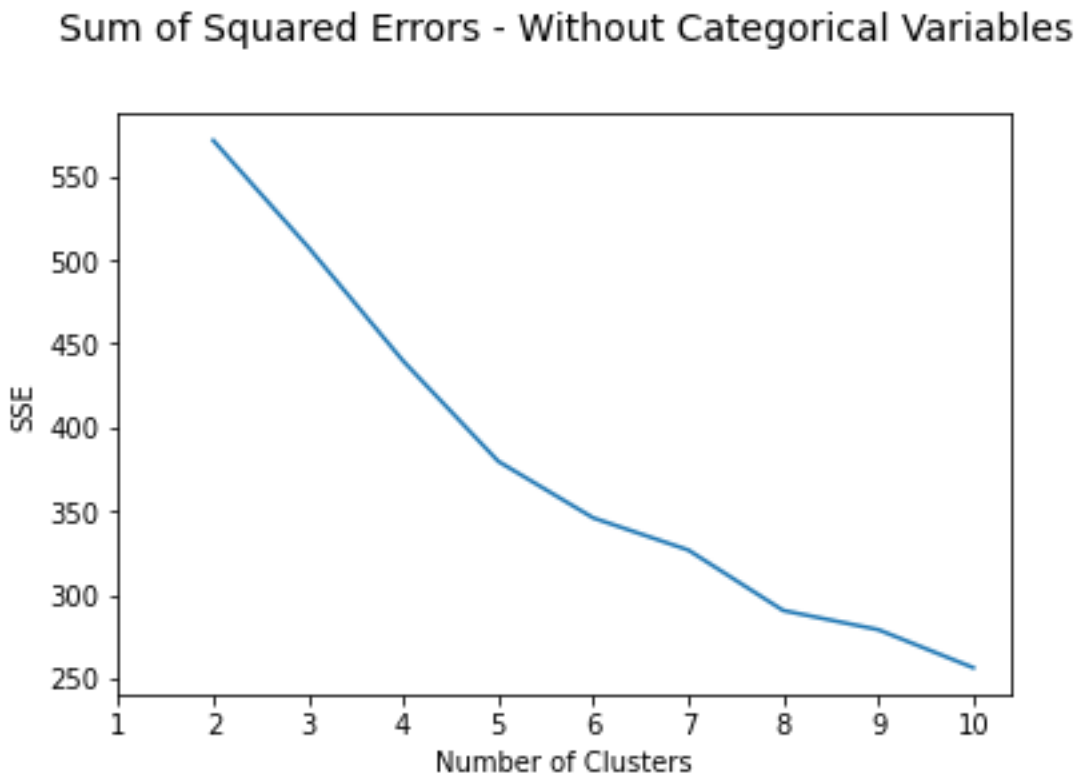

Figure 9 - Sum of Squared Errors - Without Categorical Variables

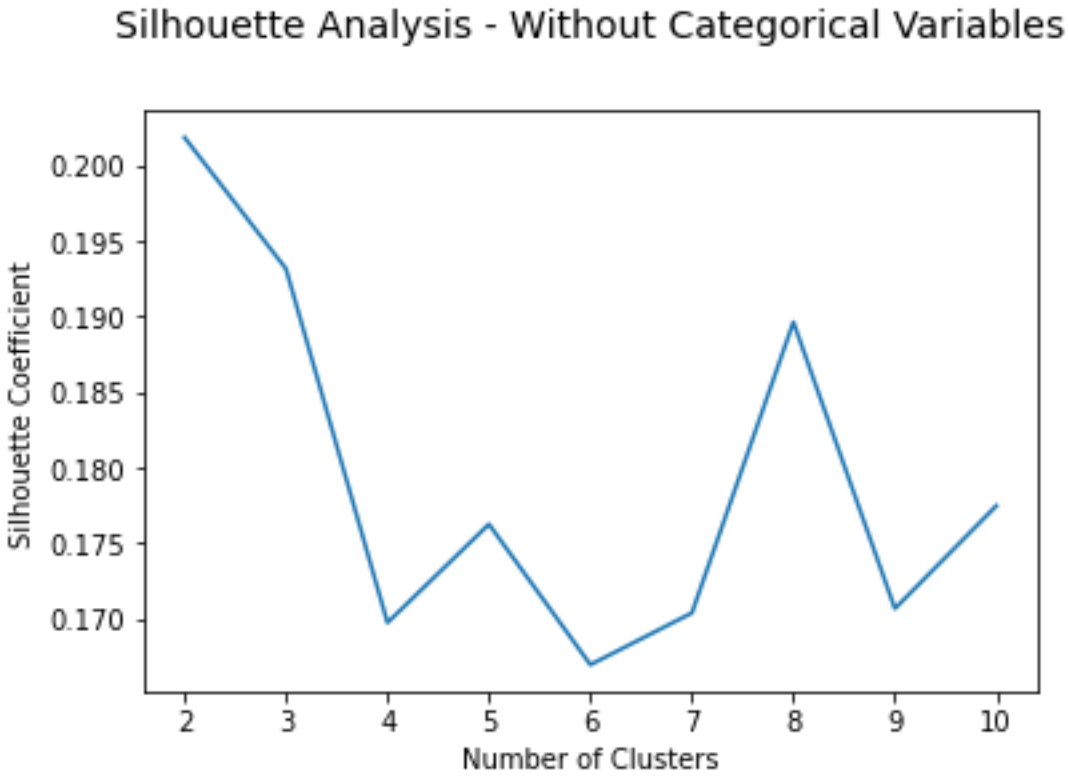

Figure 10 - Silhouette Analysis - Without Categorical Variables

The best number of clusters (k) is 6 but its Silhouette coefficient is not the best in graph above.

Option 3 – Minimum Feature Set

It is suggested these variables be removed to produce the minimum feature set:

- "NewMembers": The number of new members could already be explained by the current number of members (Members).
- "Increase": The net increase of membership could already be explained by the current number of members (Members).
- "Education": The educational awards achieved by members could already be explained by the number of goals achieved (Goals).

The resulting feature set consists of 5 variables, that is, 'Members', 'Goals', 'Venues', 'Latitude', 'Longitude'. The same tests are done, and these are the results produced.

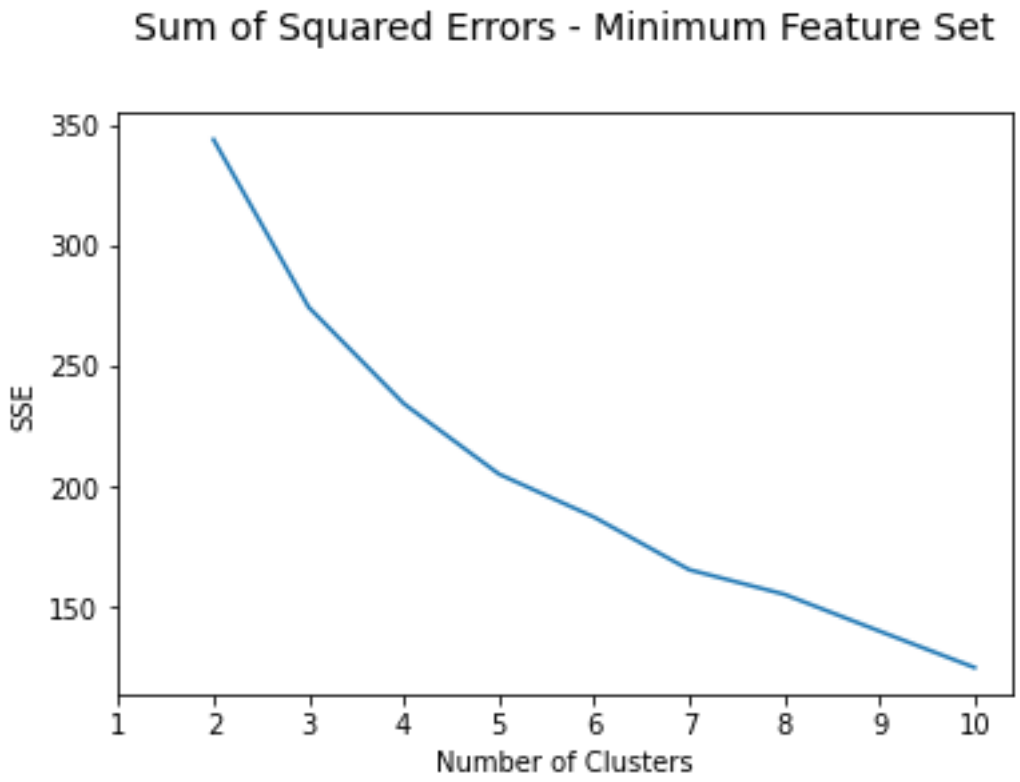

Figure 11 - Sum of Squared Errors - Minimum Feature Set

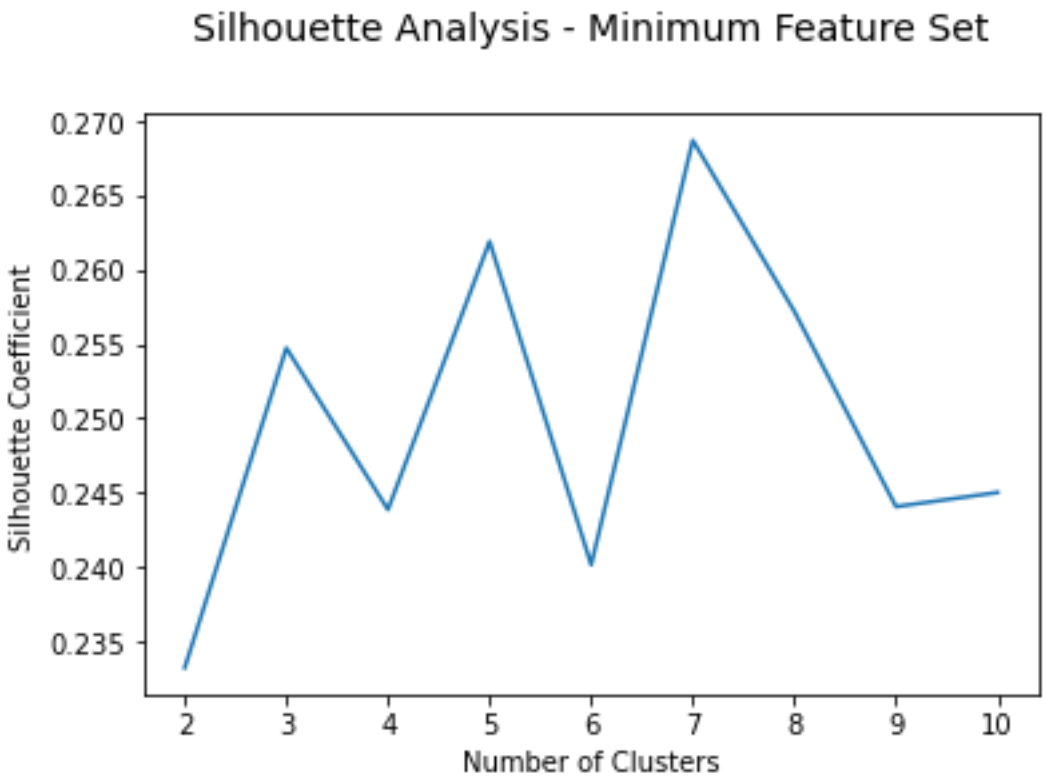

Figure 12 - Silhouette Analysis - Minimum Feature Set

The best number of clusters (k) is 5 and its Silhouette coefficient looks good in graph above.

Model Training

Option 3 looks good so the minimum set of features below will be used as the feature set to train the model.

The results are 89 clubs being grouped into 5 clustered as below:

Cluster	Club Count
1	19
2	25

3	8
4	8
5	29
Total	89

Figure 13 - Clusters of Toastmasters Clubs

The clubs are visualized in map with different colour markers below:

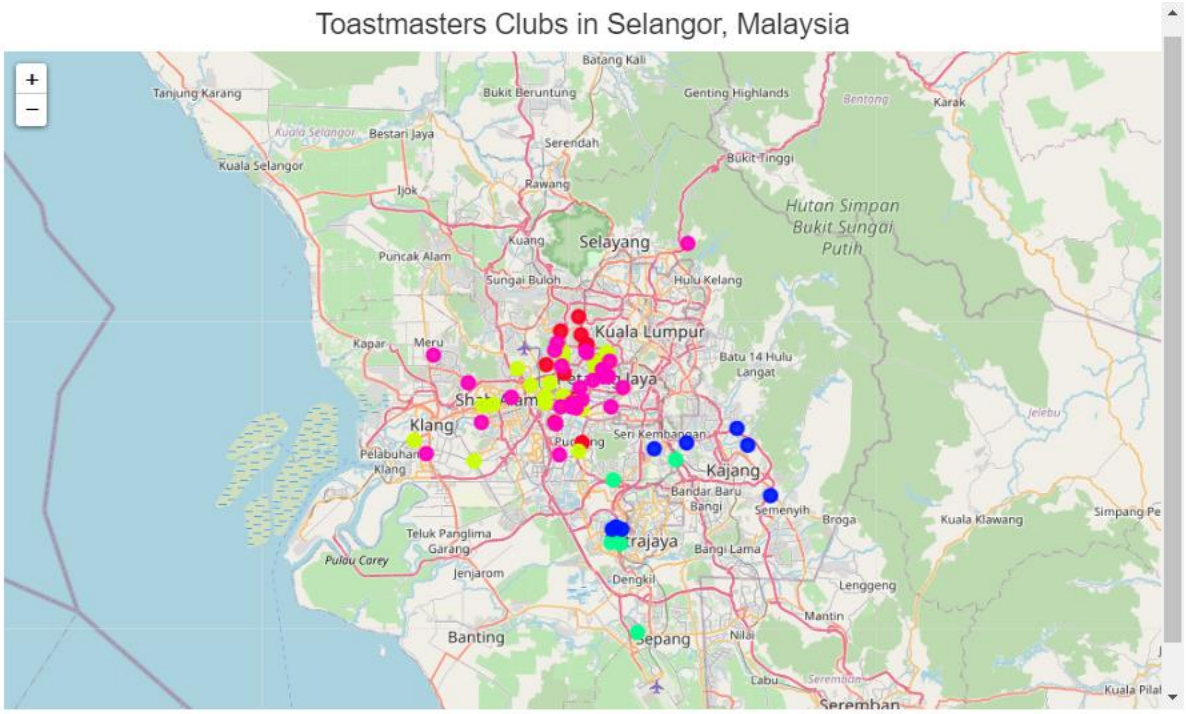

Figure 14 - Toastmasters Clubs in Clusters

RESULTS

The 89 Toastmasters clubs in the state of Selangor, Malaysia have been successfully grouped into 5 clusters. The club listing according to each cluster is included in the Appendix. Here, each cluster is analysed in relation to each variable in the data set.

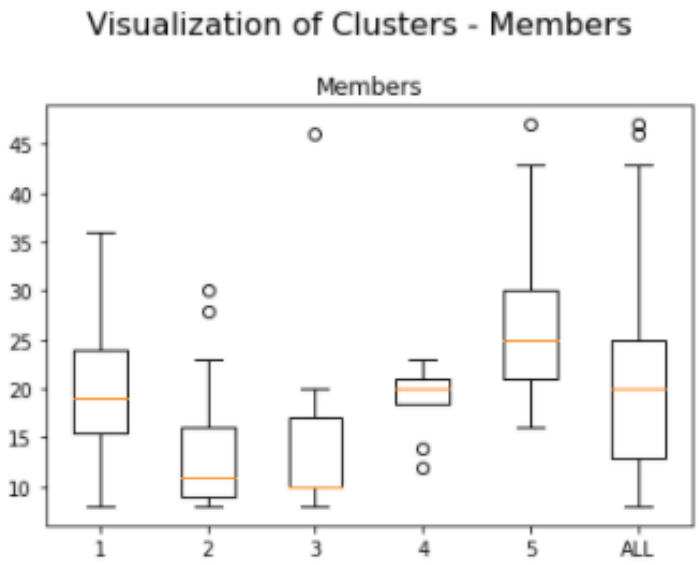

Figure 15 - Visualization of Clusters – Members

Cluster 1 and 4 tend to have average number of members. Cluster 2 and 3 are below average, while Cluster 5 is above average.

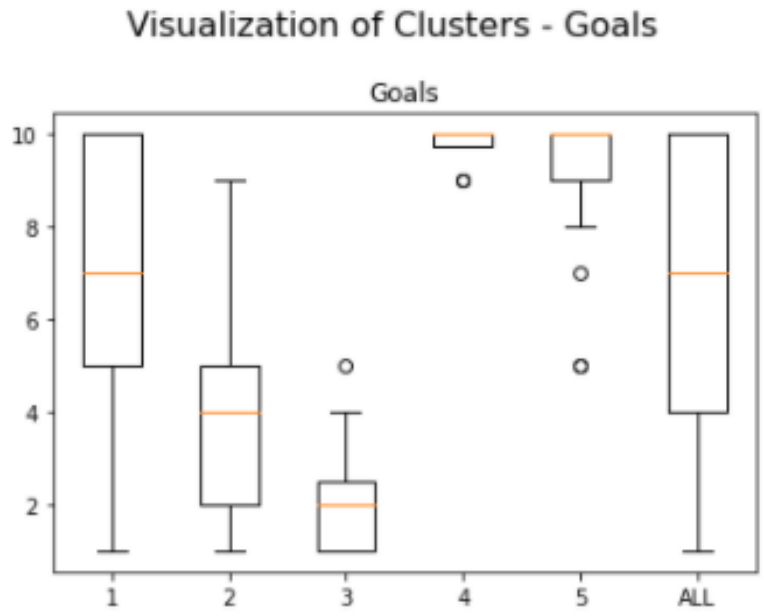

Figure 16 - Visualization of Clusters – Goals

Cluster 1 tend to have achieved average number of club goals. Cluster 2 and 3 tend to perform below average. Cluster 4 and 5 tend to perform above average. In fact, at least half of Cluster 4 and 5 achieved perfect 10 club goals!

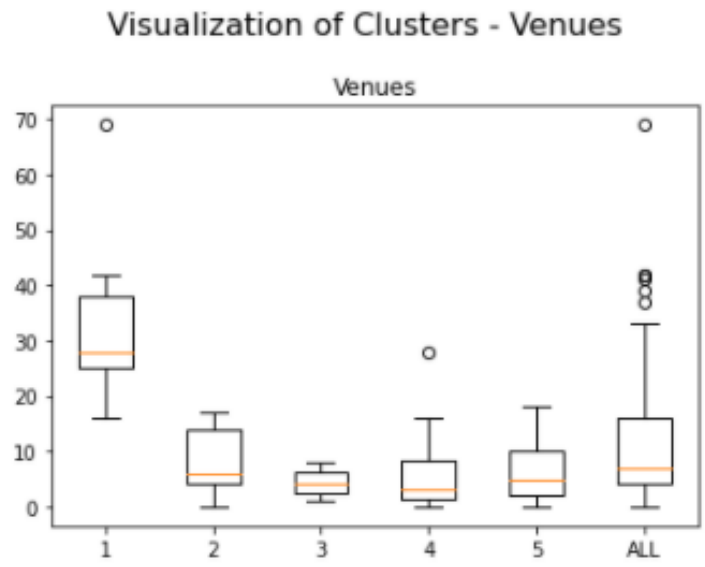

Figure 17 - Visualization of Clusters – Venues

Cluster 1 tends to have than double the average number of popular venues within walking distance of club locations. The other clusters tend to have average number of such venues.

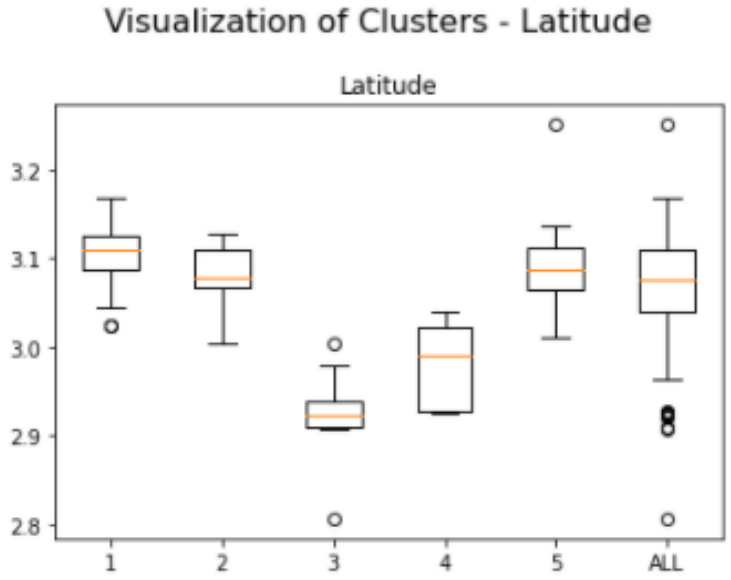

Figure 18 - Visualization of Clusters – Latitude

Cluster 3 followed by Cluster 4 tend to lie toward the southern part of Selangor state.

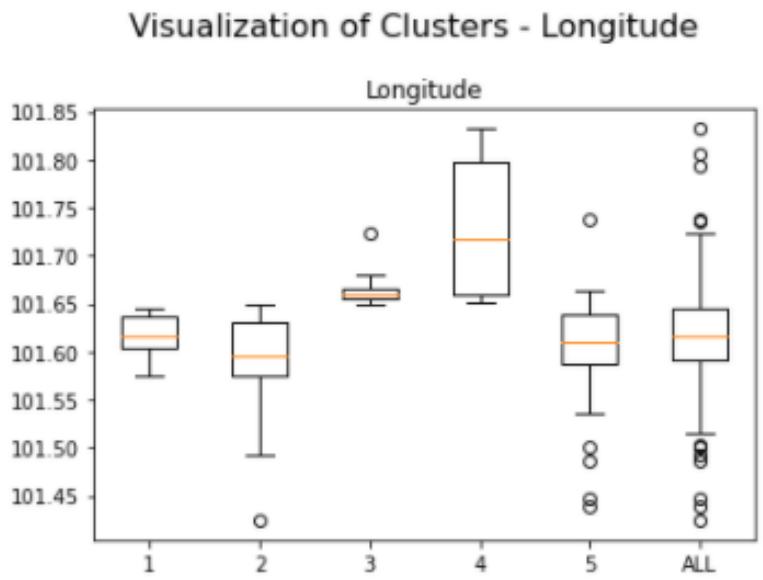

Figure 19 - Visualization of Clusters - Longitude

Cluster 4 followed by Cluster 3 tend to lie toward the eastern part of Selangor state.

Interpretation of Clusters

Based on the characteristics of each cluster, the following insights can be drawn:

Cluster 1 (19 clubs) - These clubs tend to have average number of members and goals achieved but tend to have high number of popular venues around them.

Cluster 2 (25 clubs) - These clubs tend to have number of members and goals achieved below the averages.

Cluster 3 (8 clubs) - These clubs tend to have number of members and goals achieved below the averages. They are located more toward the southern part of Selangor state.

Cluster 4 (8 clubs) - These clubs tend to have average number of members, but above-average goals achieved. They are located more toward the south-eastern part of Selangor state.

Cluster 5 (29 clubs) - These clubs tend to have above-average number of members and goals achieved.

DISCUSSION

With these results, District Officers can gain insights as to the similarities and dissimilarities among the clubs under their fold. Then they can make use of these insights to formulate strategies in management and support, marketing, realignment as outlined at the start of this project:

Management and support – Clubs perform at different levels due to the condition, opportunities and challenges faced by the individual clubs. One role of District Officers is to lead and support these clubs to excellence so that the members can fully enjoy the benefits of Toastmasters. For the better performing clubs such as those in Cluster 4 and 5, their strengths can be harnessed to help other clubs in the district. For the temporarily less performing clubs such as those in Cluster 2 and 3, District Officers need to help review each club's situation and come up with valid strategies to help the clubs grow both in terms of membership and club goals. Different programs can be strategized for the different needs of clubs in different clusters.

Marketing – One main feature of this clustering exercise is to gain insights into the level of activities in the neighbourhoods where the clubs are located. It is astounding to find there are 1,093 popular venues in the vicinity of the clubs, which translates to over 12 venues within just 100 metres of each club! Furthermore, Clubs in Cluster 1 has an average of over 33 venues per club. These indicate huge untapped opportunities in what clubs can do to promote themselves to the people around their clubs. Both online and offline marketing activities can be planned and executed.

Realignment – This clustering exercise can help give inputs to the yearly realignment performed by district at the end of each term. The 89 clubs can be realigned equitably into 5 divisions (currently they are also in 5 divisions) in a way that each division consists of clubs of different strengths and are close to one another geographically as much as possible. For example, Cluster 1 (19 clubs) can form a division by itself. Cluster 3 and 4 (16 clubs) in the southern part of Selangor state can be grouped into one division. The remaining 54 clubs in Cluster 2 and 5 can be divided into 3 divisions of 18 clubs each consisting of clubs of different strengths. Then each division will have about the same number of clubs and have equal collective strength due to the mixture of clubs.

CONCLUSION

This project has started by identifying the needs of the main stakeholder in mind, that is to help District Officers gain insights into clubs for strategic purposes. As data on location and venues is added to the dataset, more insights have been gained and more opportunities are found to grow and promote the clubs.

Multiple techniques have been used in the project, including descriptive statistics, data visualization, feature selection, model training, and model evaluation. All these are done with the end in mind, that is the needs of the District Officers and what the insights can help them. It is now incumbent to them to make good use of the knowledge gained and perhaps do further study into the model to make it even more useful.

This project is completed as the capstone project for the Applied Data Science Capstone course on Coursera at https://www.coursera.org/learn/applied-data-science-capstone, which is the final course for the IBM Data Science Professional Certificate course at https://www.coursera.org/professional-certificates/ibm-data-science

Learn more about this project at https://github.com/rickysoo/clustering-toastmasters/
Ideas and comments are welcome. Please email to ricky [at] rickysoo [dot] com.

APPENDIX – CLUB LISTING

Disclaimer – The clubs are clustered using machine learning techniques, some of which are like black boxes and are hard to explain how they work internally. Best possible efforts have been put in to generate the most accurate results, as per explained in previous sections. Each cluster of clubs may **tend to** exhibit certain characteristics, but it does not mean all clubs in a cluster behave uniformly the same way.

Cluster 1 (19 clubs)

1. Apple Mentors Toastmasters Club
2. CIMA Malaysia Toastmasters Club
3. Damansara Toastmasters Club
4. Gamuda Toastmasters Club
5. HILTI ASIA IT SERVICES
6. Kelab Pidato Perdana Toastmasters Club
7. MDA Kuala Lumpur and Selangor Toastmasters Club
8. MIM Club of Petaling Jaya
9. MY Puchong Toastmasters Club
10. Nielsen Malaysia Toastmasters Club
11. Novartis Malaysia
12. Phoenix Toastmasters Club
13. Puchong English Toastmasters Club
14. S&P Industries Toastmasters Club
15. School Of Hard Knocks Toastmasters Club
16. Standard Chartered GBS Toastmasters Club
17. Subang Toastmasters Club
18. Trailblazer Toastmasters Club
19. eLawyer Toastmasters Club

Cluster 2 (25 clubs)

1. Anbu Tamil Bilingual Toastmasters Club
2. BAT Toastmasters Club
3. DRB-HICOM Toastmasters Club
4. Extol Toastmasters Club
5. KPMG Toastmasters Club
6. Kota Anggerik Toastmasters Club
7. Lavangam Tamil Toastmasters Club
8. Law UiTM Toastmasters Club
9. MAS United Toastmasters Club

10. MC JO Toastmasters Club
11. MIEA Toastmasters Club
12. MWKA Toastmasters Club
13. Macfood Toastmasters Club
14. REAL Toastmasters Club
15. Roche Malaysia Toastmasters Club
16. SIRIM Toastmasters Club
17. SJKT Castlefield Toastmasters Club
18. Shopper360 Toast and Roast Toastmasters Club
19. Siemens Malaysia Toastmasters Club
20. Subang Jaya Medical Centre Toastmasters Club
21. Summit Toastmasters Club
22. Tamil Toastmasters Club, Petaling Jaya
23. Taylor's Toastmasters Club
24. Unipac Toastmasters Club
25. Wah Seong Community Toastmasters Club

Cluster 3 (8 clubs)

1. AIG Movers & Shakers
2. AMD Cyberjaya Toastmasters Club
3. DHL Cyberjaya Toastmasters Club
4. Dell Technologies Malaysia Toastmasters Club
5. Hartalega Toastmasters Club
6. Hasil Toastmasters Club
7. Putra Toastmasters Club
8. Toastmasters@Leap

Cluster 4 (8 clubs)

1. Bukit Serdang Toastmasters Club
2. Cyberjaya Community Toastmasters Club
3. Kajang Toastmasters Club
4. Master Jaya Toastmasters Club
5. Shell Cyberjaya Toastmasters Club
6. T-Systems Cyberjaya Toastmasters Club
7. UNM Toastmasters Club
8. UTAR Sungai Long Toastmasters Club

Cluster 5 (29 clubs)

1. Crystal Toastmasters Club
2. D'Utama Advanced Toastmasters Club
3. D'Utama Toastmasters Club

4. EcoWorld Toastmasters Club
5. Friendship Toastmasters Club
6. Fusion Inspired Toastmasters Club
7. Gasing Hills Toastmaster Club
8. IEM Toastmasters
9. IIUM Toastmasters Club
10. IJM Toastmasters Club
11. INTEC Deutsch
12. JCorp 2 Toastmasters Club
13. Klang Bilingual Toastmasters Club
14. MAD Toastmasters Club
15. Monash University Toastmasters Club
16. Money Mastery-KL Toastmasters Club
17. OUM Toastmasters Club
18. PJ Toastmasters Club
19. Roche SS Toastmasters Club
20. SHINE Toastmasters Club
21. Sai Masters Toastmasters Club
22. Shah Alam Toastmasters Club
23. Shaklee Dynamic Family International Club
24. Sunway Toastmasters Club
25. Sunway University Toastmasters Club
26. Super Speakers Toastmasters Club
27. Taman Indrahana Toastmasters Club
28. Top Glove Toastmasters Club
29. USJ Toastmasters Club