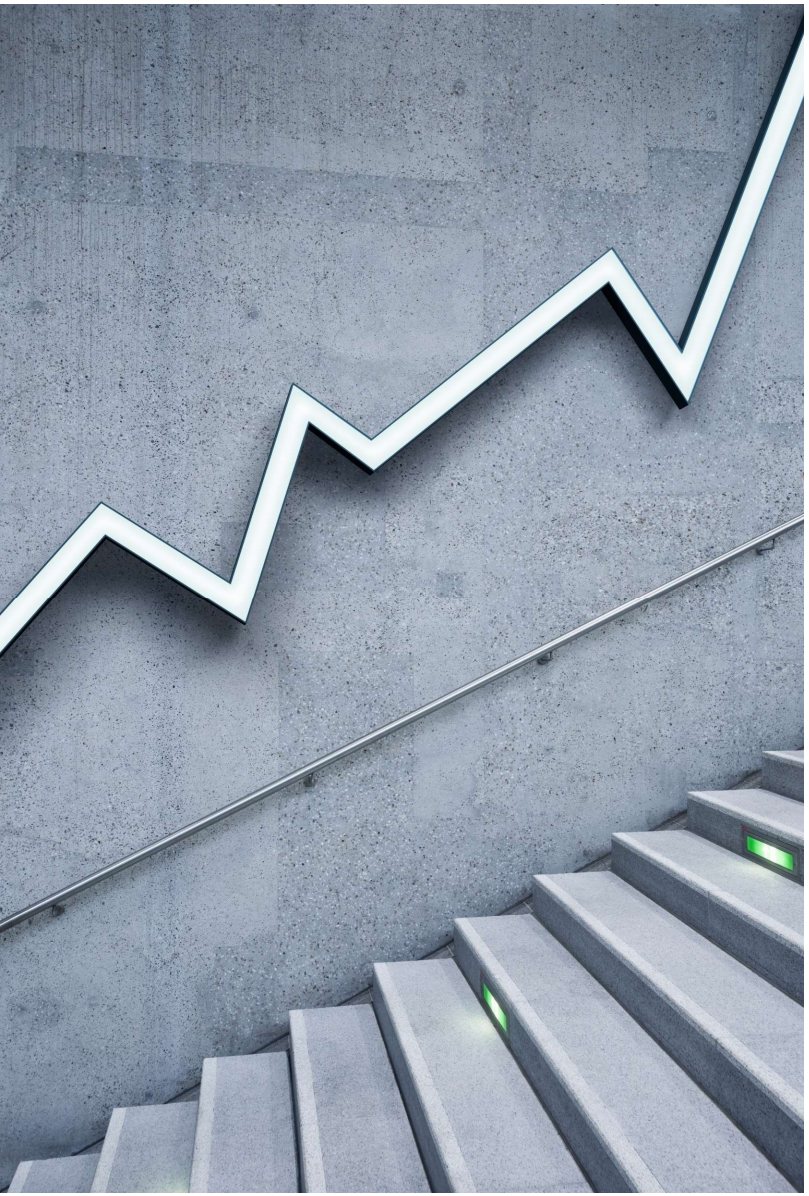


Clustering Toastmasters Clubs with Machine Learning

RICKY SOO

AUGUST 23RD, 2020

Machine Learning?

The data science magic that helps transform your clubs!

Note to Data Scientists

These slides are targeted to the main stakeholder – the Toastmasters district leaders. I've tried to use layman terms and minimize jargons as much as possible.

You may check https://github.com/rickysoo/clustering-toastmasters/ for some background knowledge on Toastmasters and more technical elaboration of the whole project.

Introduction

WHAT IS THIS PROJECT ABOUT?

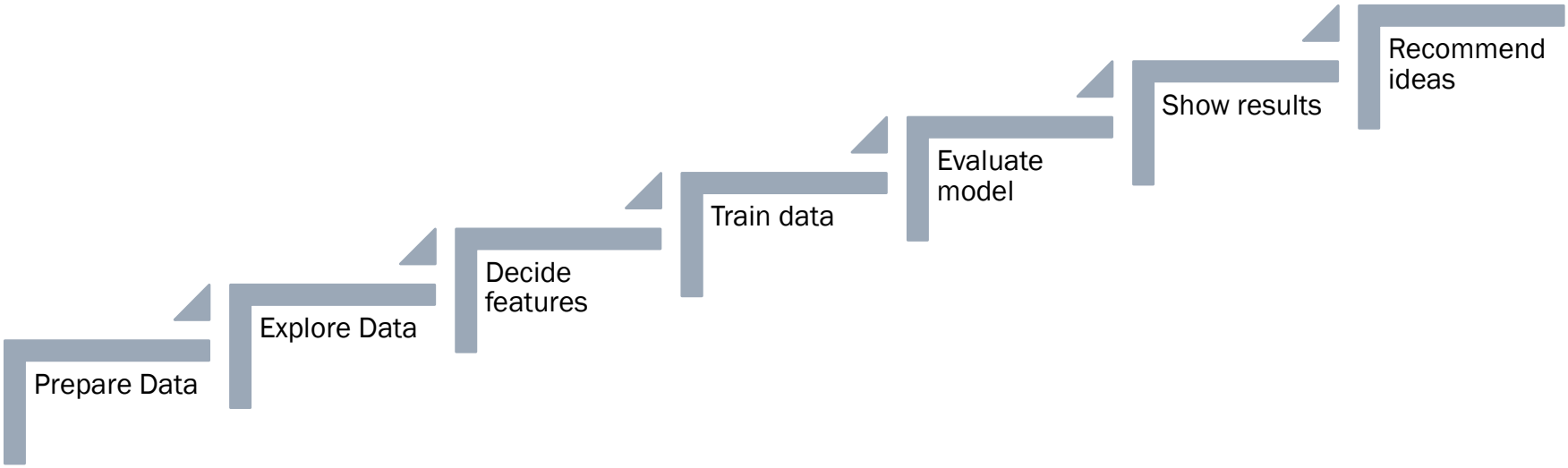
The Benefits for You

- **Management and Support** - Strategize your clubs to excellence, specific to their particular needs and conditions
- **Marketing** – Design marketing programs for your clubs, especially those located in popular places?
- **Realignment** – Gain new insights to the similarities and dissimilarities of your clubs to help in yearly realignment exercise

One-size solution does not fit all

- Make use of the latest technologies in data science
- Group your clubs into “clusters” of similar clubs
- Devise strategies according to the strengths and conditions of the clubs
- Explore new marketing opportunities for each cluster of clubs
- Lead your clubs to excellence, both membership and education goals!

The Process


Prepare Data
Explore Data
Decide features
Train data
Evaluate model
Show results
Recommend ideas

The Data

Data Sources

- Toastmasters web site
- Foursquare location data

Data Types

- Club info
- Club location

Data Subjects

- 89 active clubs in Selangor, Malaysia

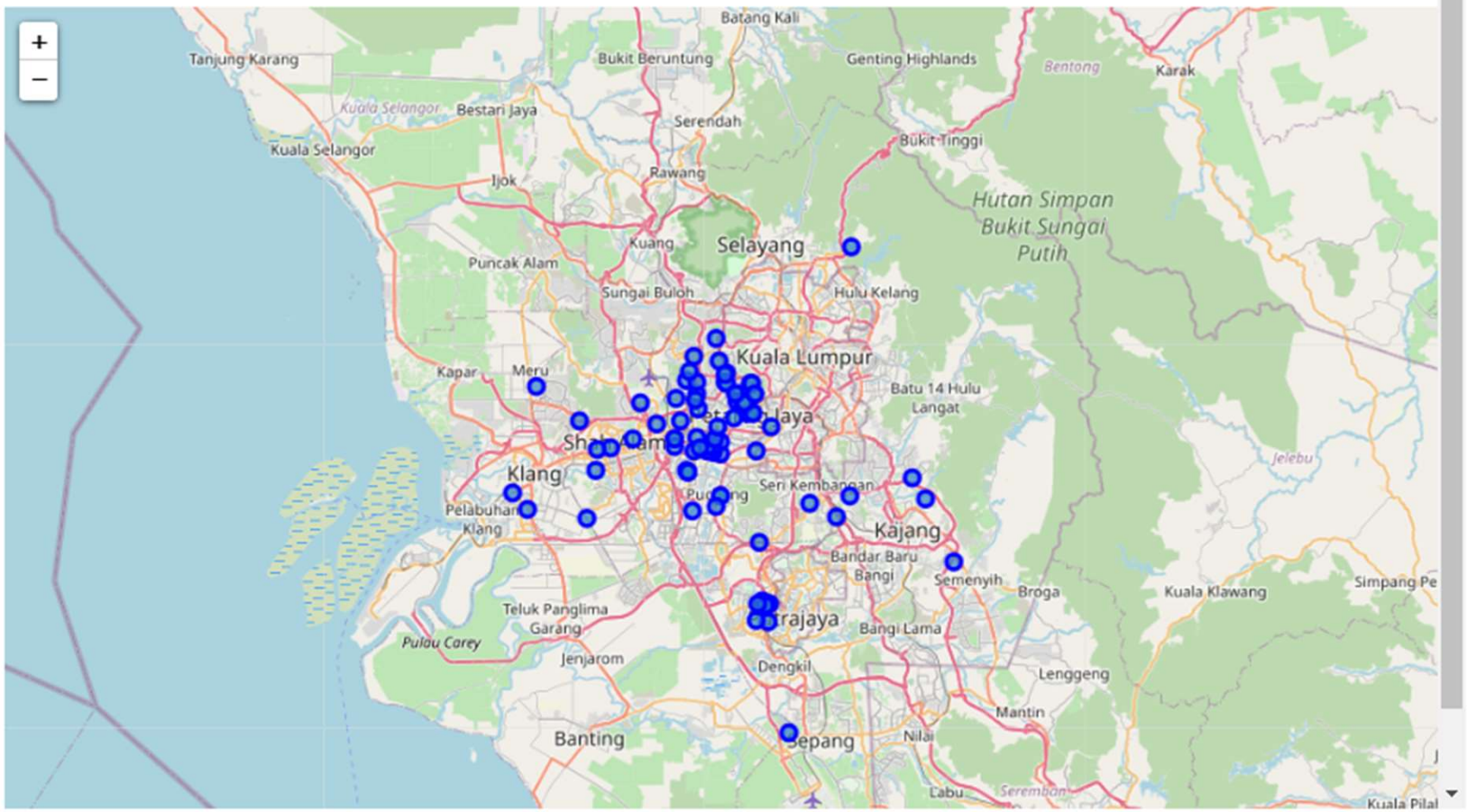
Data Prepared

1. Club number
2. Club name
3. Membership
4. New members
5. Increase or Decrease in membership
6. Club goals achieved
7. Education awards

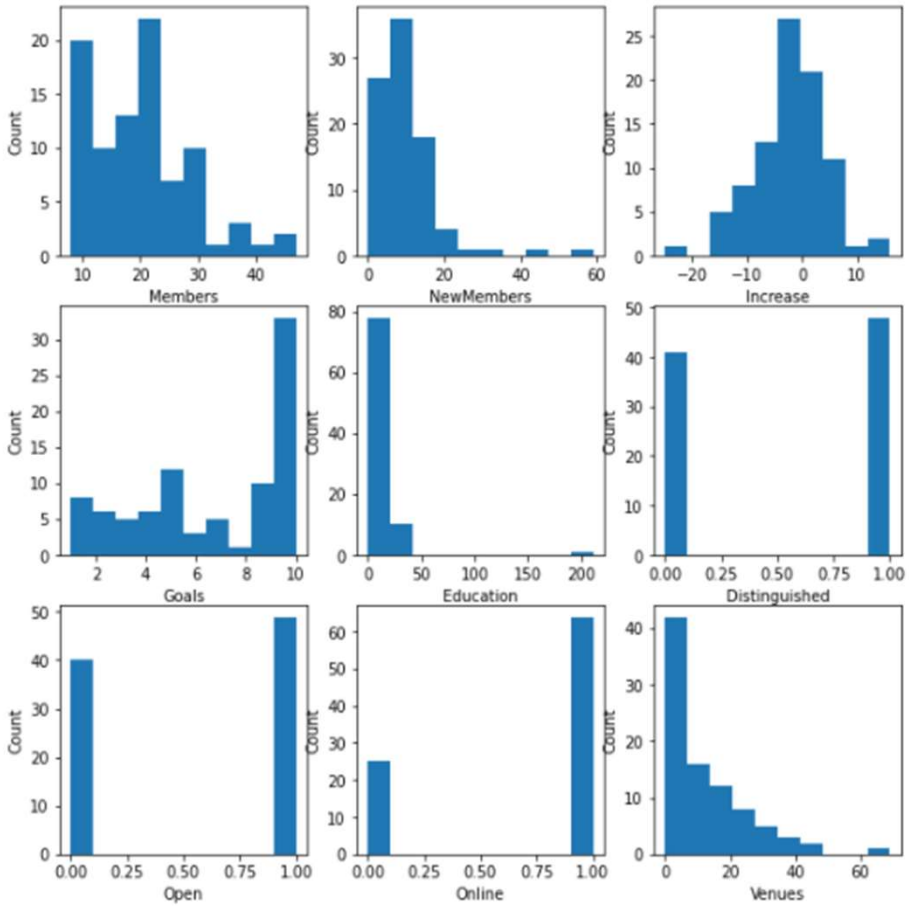
Data Prepared

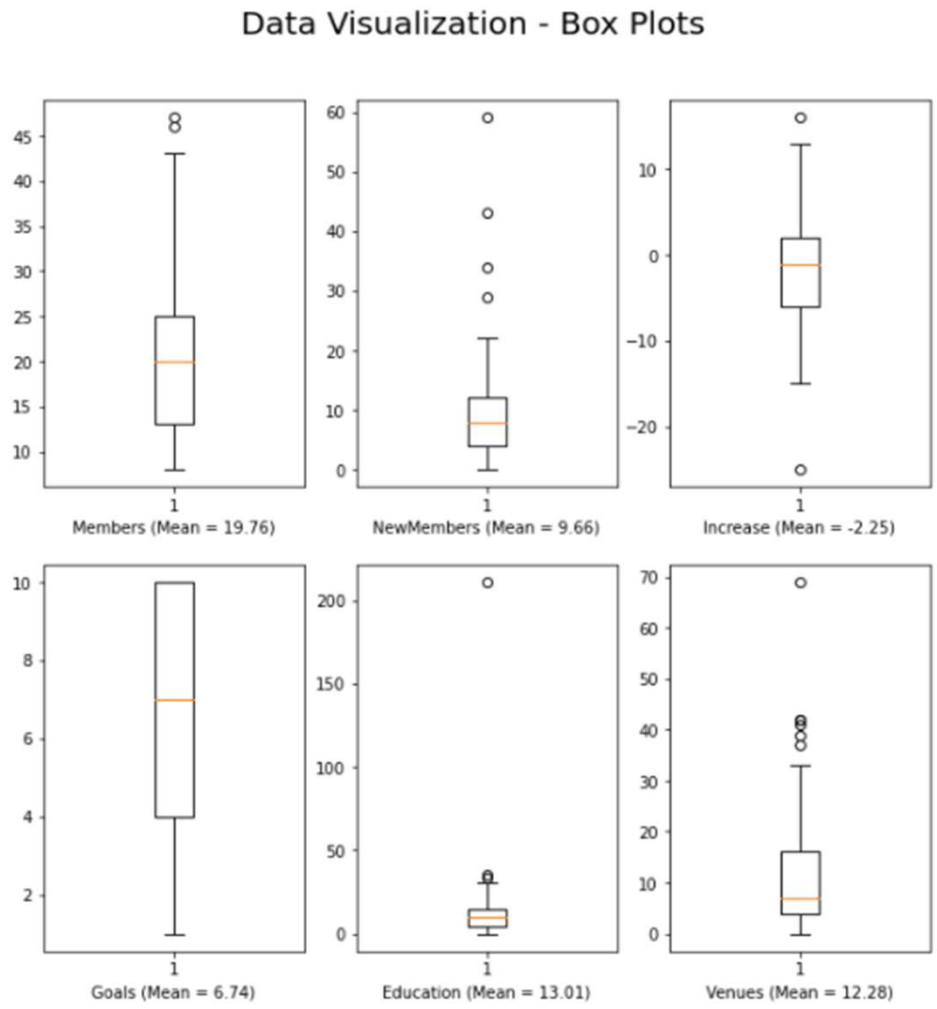
8. Distinguished club status (yes/no)
9. Open to public (yes/no)
10. Online attendance (yes/no)
11. Club location
12. Popular venues within 100m

Toastmasters Clubs in Selangor, Malaysia

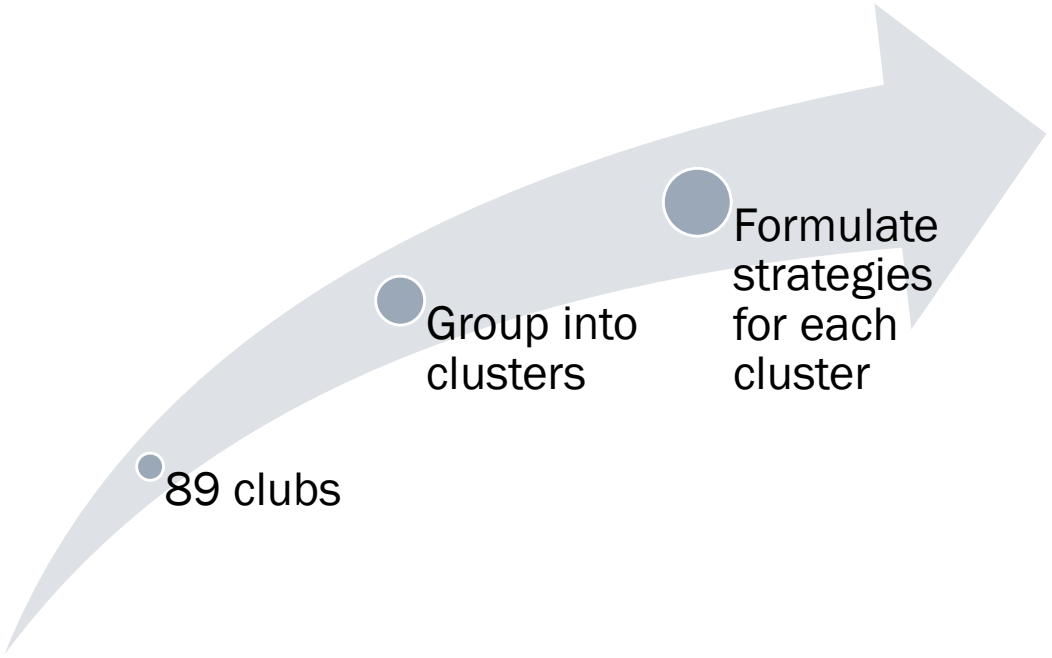


Data Visualization - Histogram




Data Visualization - Box Plots
Members (Mean = 19.76)
NewMembers (Mean = 9.66)
Increase (Mean = -2.25)
Goals (Mean = 6.74)
Education (Mean = 13.01)
Venues (Mean = 12.28)

The Challenge

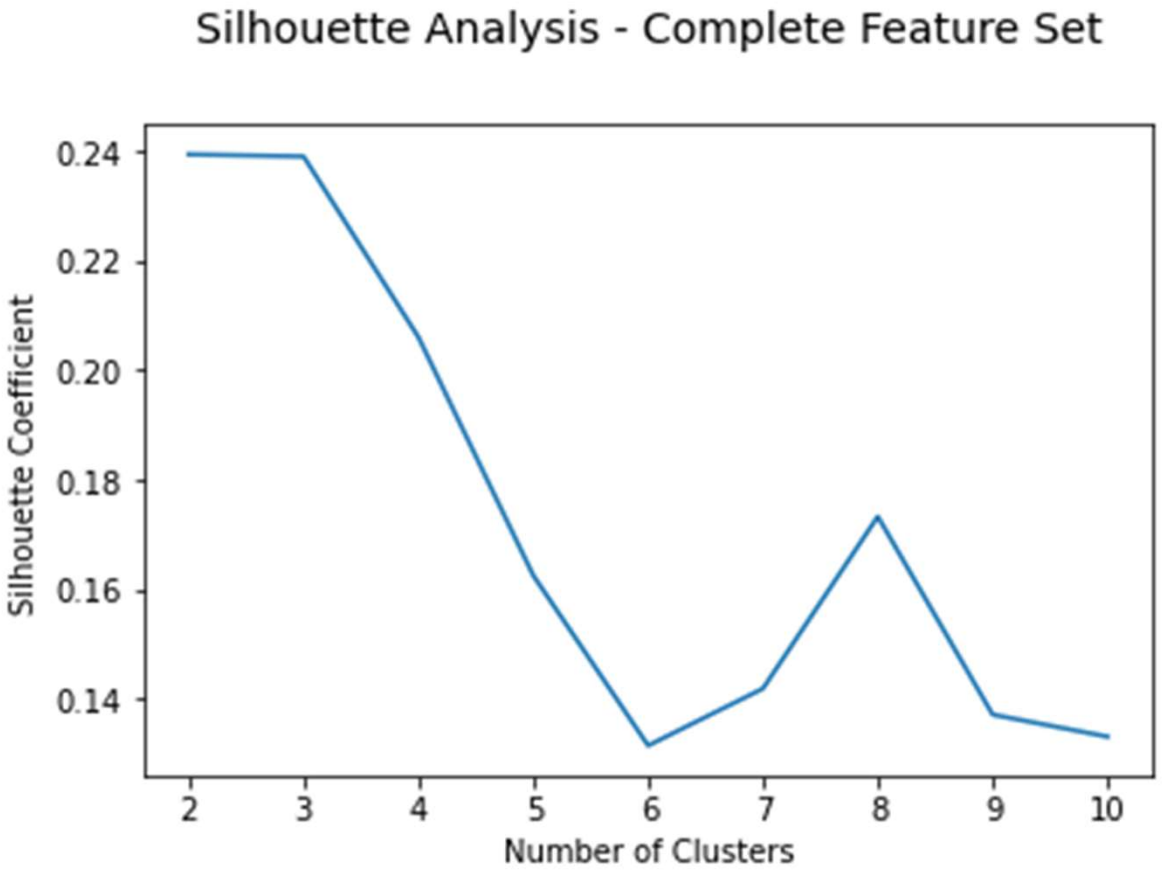


Model Training

LET THE TRAINING BEGIN!

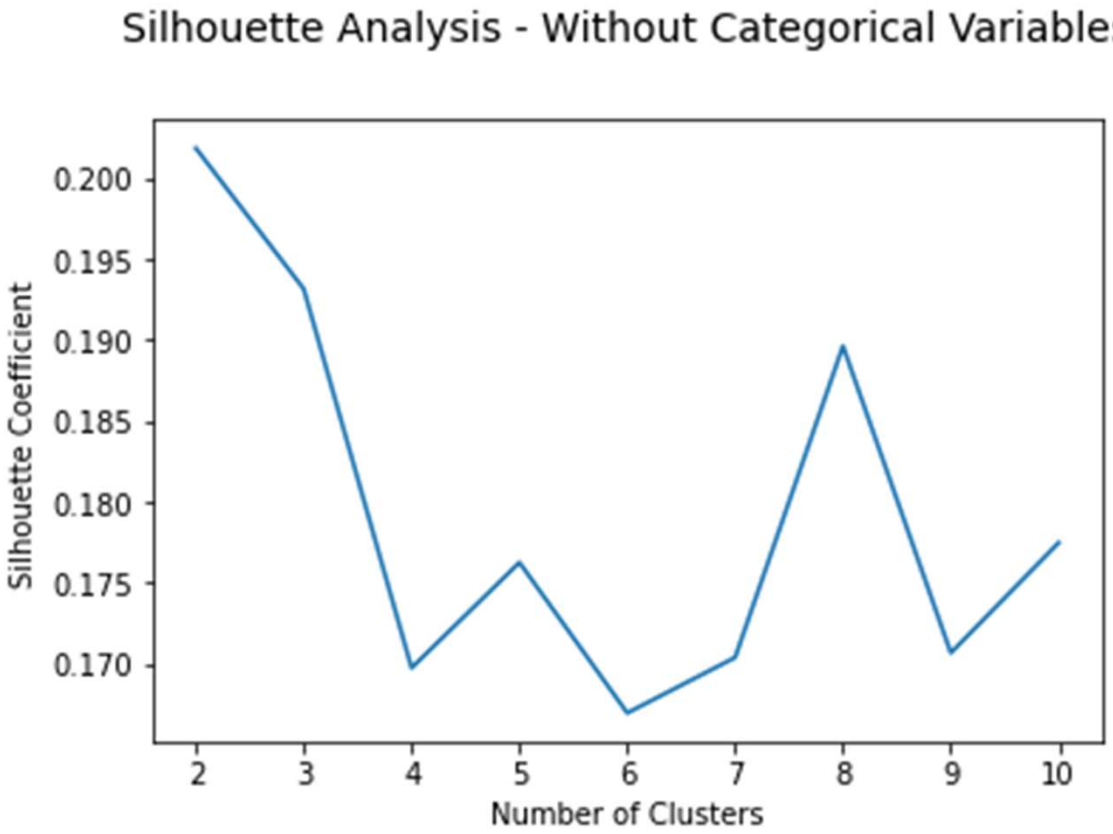
Option 1 – Complete Feature Set

- 11 variables used
- 4 clusters generated
- Not optimum ☹


Silhouette Analysis - Complete Feature Set
Silhouette Coefficient
0.24
0.22
0.20
0.18
0.16
0.14
2
3
4
5
6
7
8
9
10
Number of Clusters

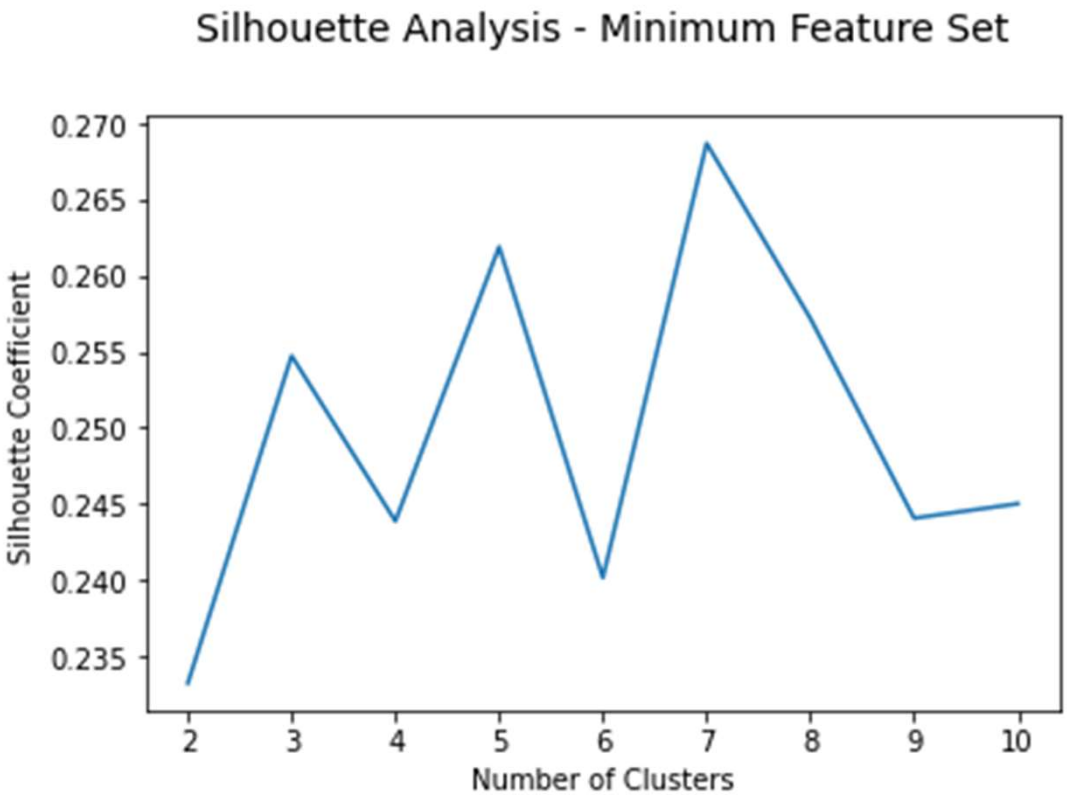
Option 2 – No Categorical Variables

- 8 variables used
- 6 clusters generated
- Not optimum ☹

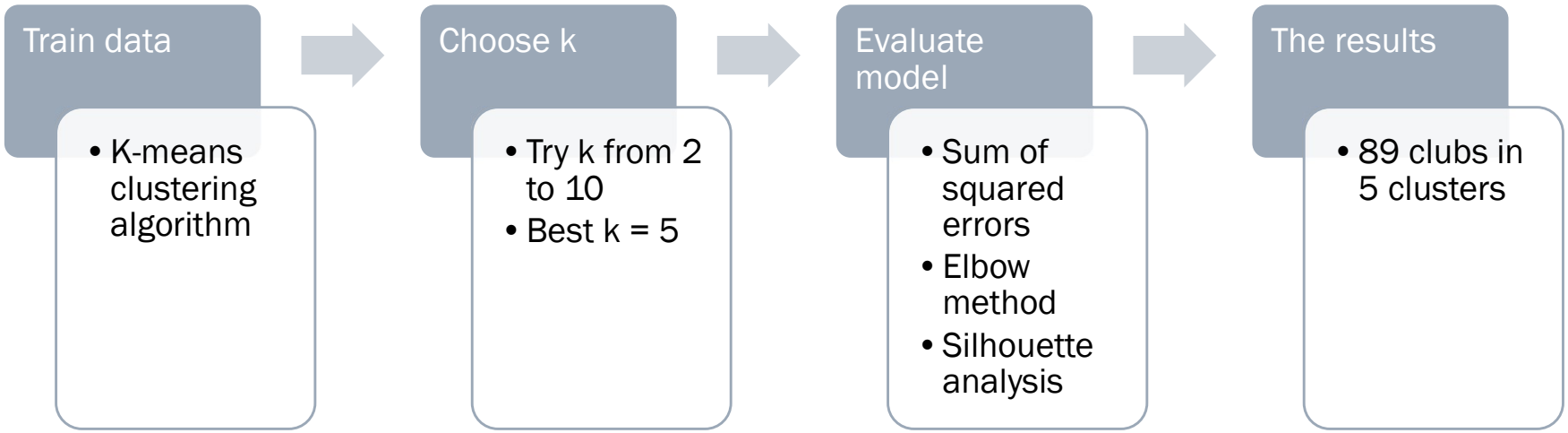

Silhouette Analysis - Without Categorical Variables
Silhouette Coefficient
0.200
0.195
0.190
0.185
0.180
0.175
0.170
2
3
4
5
6
7
8
9
10
Number of Clusters

Option 3 – Minimum Feature Set

- 5 variables used
- 5 clusters generated
- Optimum ☺


Silhouette Analysis - Minimum Feature Set
Silhouette Coefficient
0.270
0.265
0.260
0.255
0.250
0.245
0.240
0.235
2
3
4
5
6
7
8
9
10
Number of Clusters

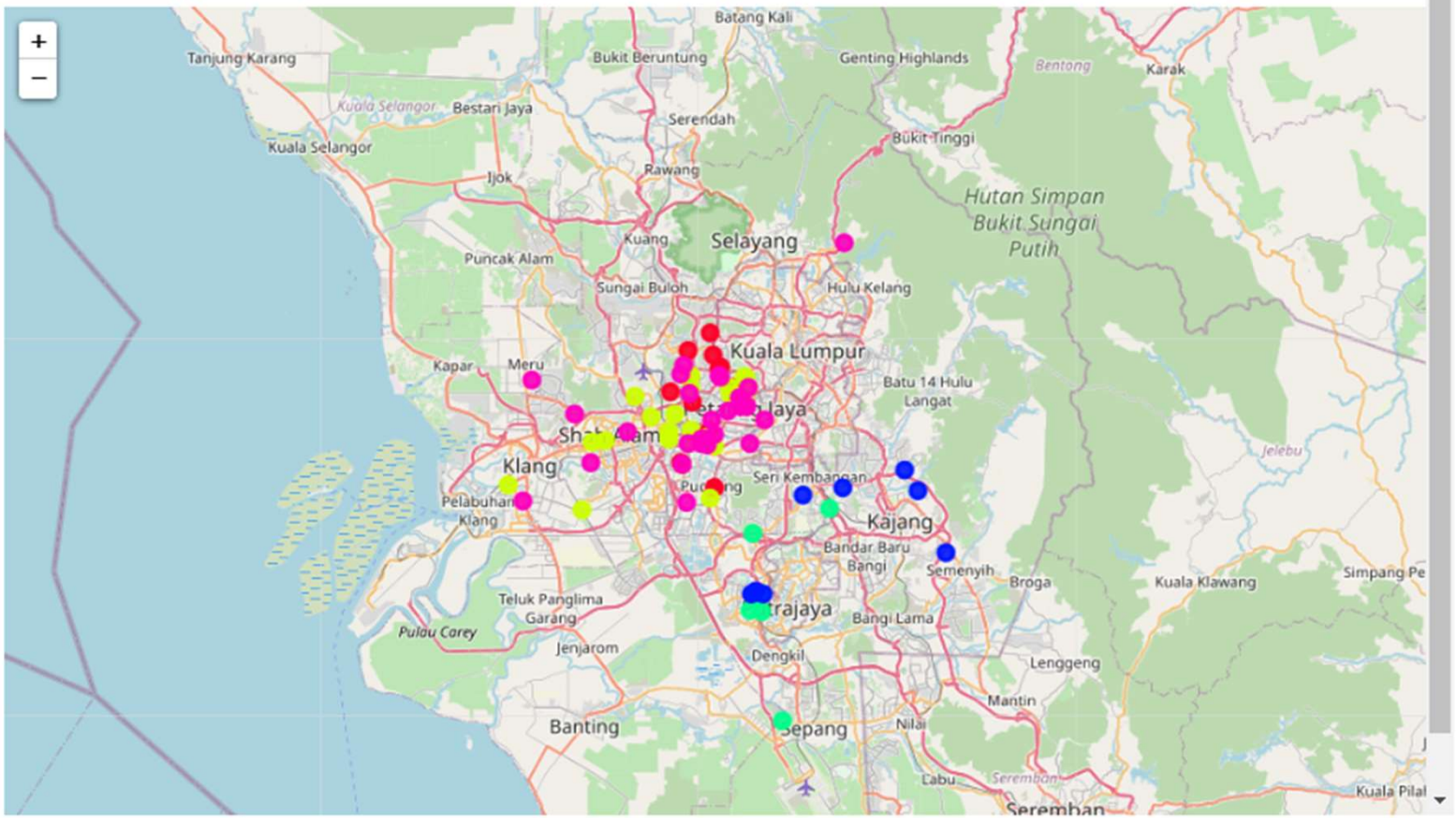
The Model

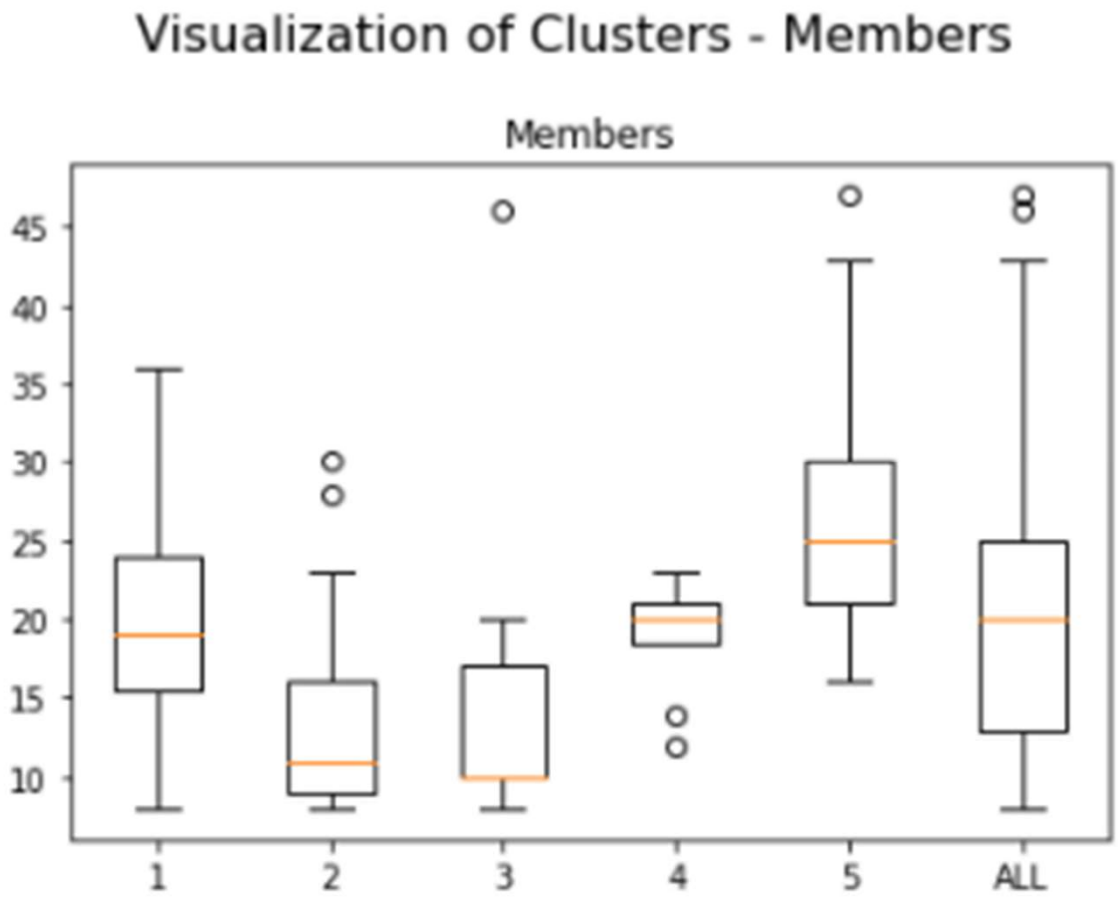

Train data
• K-means clustering algorithm
Choose k
• Try k from 2 to 10
• Best k = 5
Evaluate model
• Sum of squared errors
• Elbow method
• Silhouette analysis
The results
• 89 clubs in 5 clusters

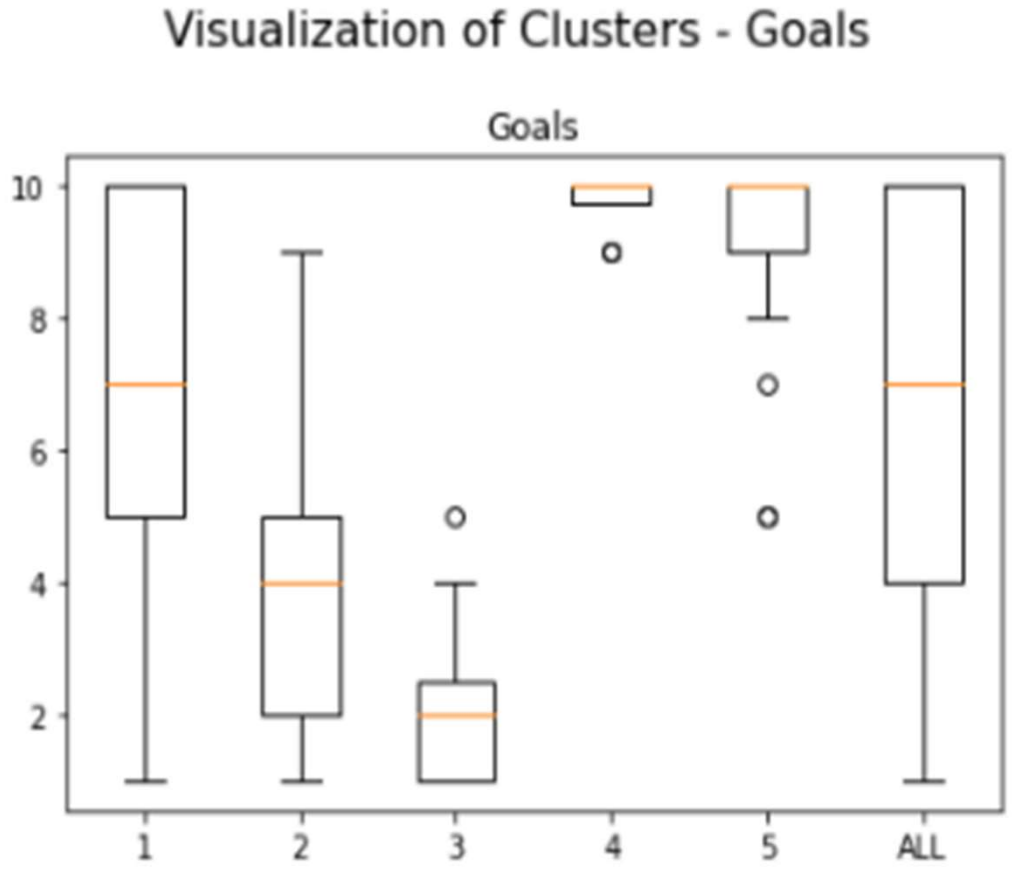
The Clusters

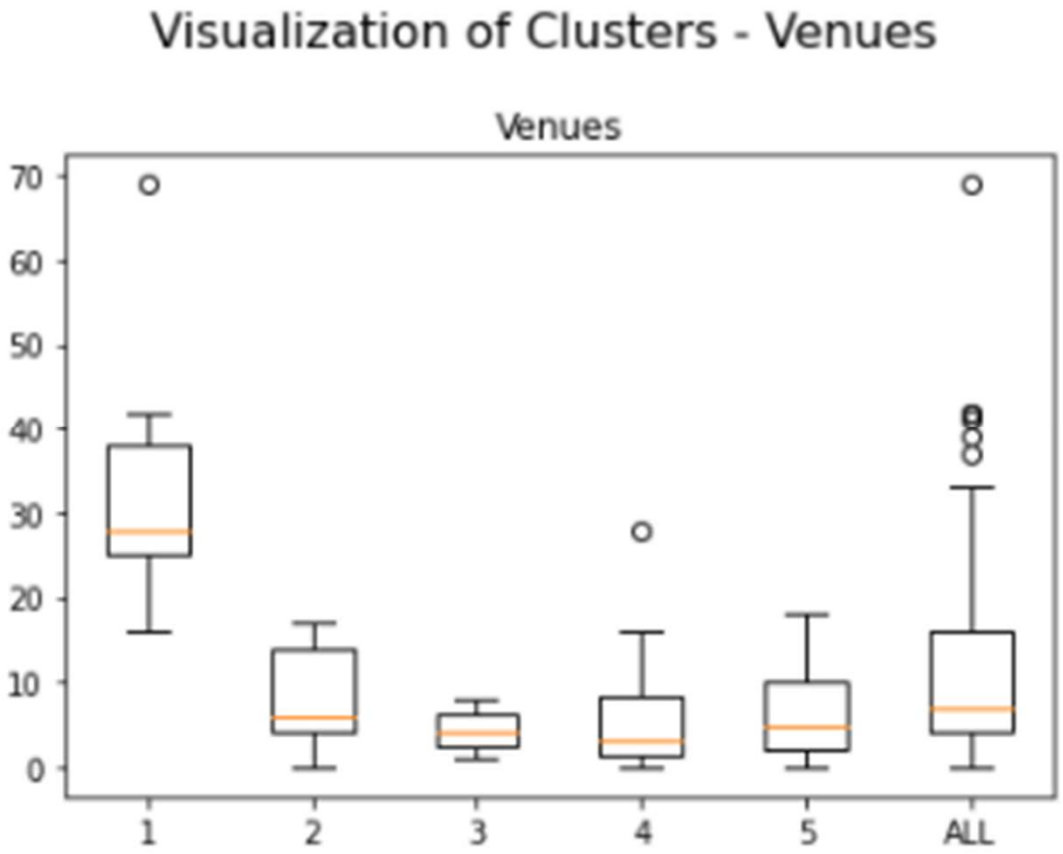
AND THE RESULTS ARE…

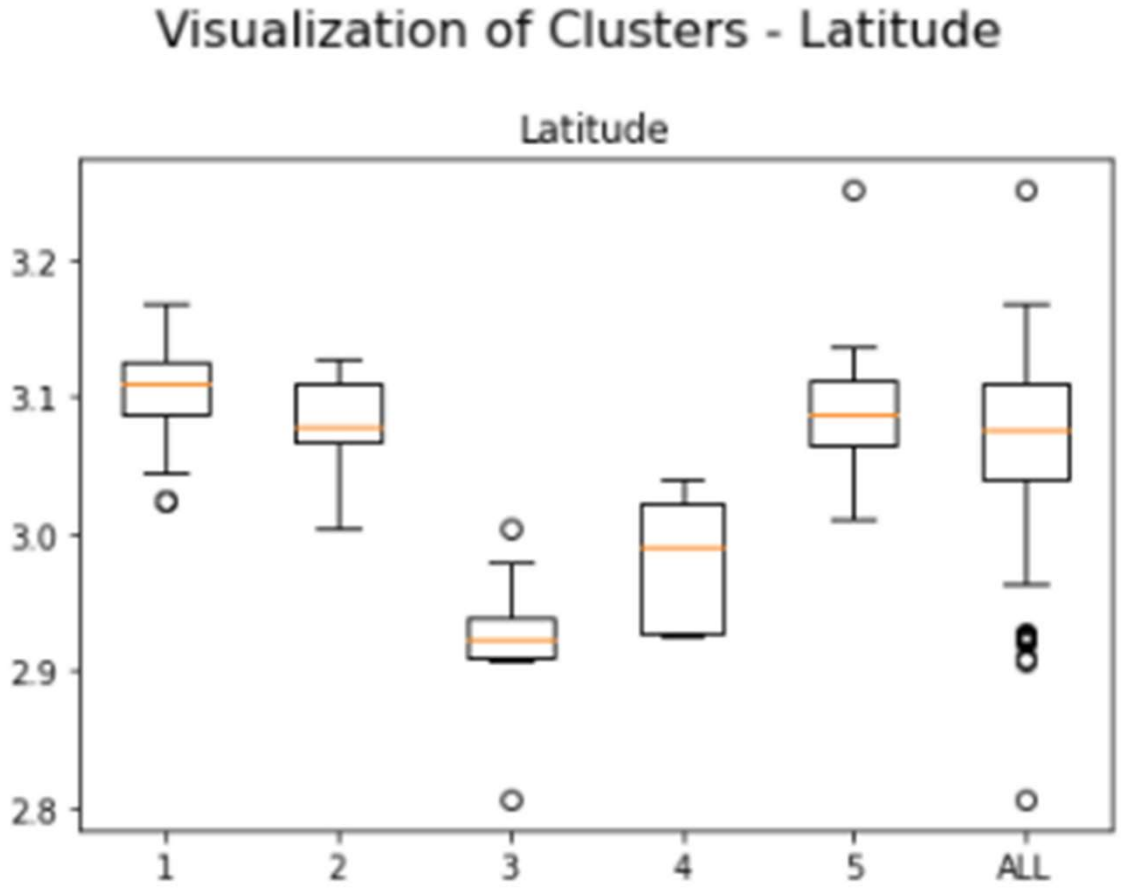
Toastmasters Clubs in Selangor, Malaysia

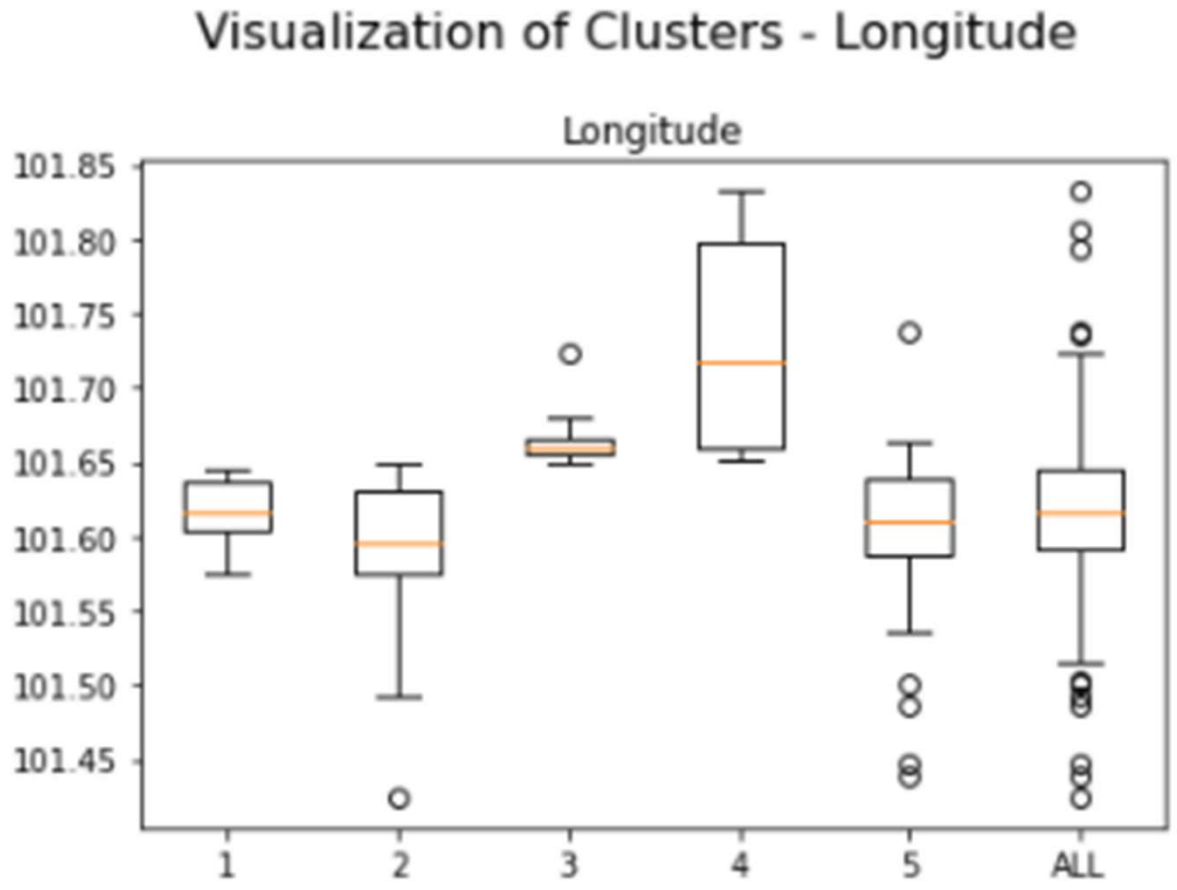



Visualization of Clusters - Members
Members
45
40
35
30
25
20
15
10
1
2
3
4
5
ALL


Visualization of Clusters - Goals
Goals
10
8
6
4
2
1
2
3
4
5
ALL


Visualization of Clusters - Venues
Venues
70
60
50
40
30
20
10
0
1
2
3
4
5
ALL


Visualization of Clusters - Latitude
Latitude
3.2
3.1
3.0
2.9
2.8
1
2
3
4
5
ALL


Visualization of Clusters - Longitude
Longitude
101.85
101.80
101.75
101.70
101.65
101.60
101.55
101.50
101.45
1
2
3
4
5
ALL

Interpretation

SO WHAT DOES ALL THESE MEAN?

Cluster 1 (19 clubs)

- Membership – Average
- Club goals – Average
- Location – Central
- Popular venues – A lot (more than double those of other clusters!)

Cluster 2 (25 clubs)

- Membership – Below average
- Club goals – Below average
- Location – Central
- Popular venues – Average

Cluster 3 (8 clubs)

- Membership – Below average
- Club goals – Below average
- Location – South
- Popular venues – Average

Cluster 4 (8 clubs)

- Membership – Average
- Club goals – Above average (at least half achieved perfect 10 goals!)
- Location – Southeast
- Popular venues – Average

Cluster 5 (29 clubs)

- Membership – Above average
- Club goals – Above average (at least half achieved perfect 10 goals!)
- Location – Central
- Popular venues – Average

Discussion

WHAT CAN WE DO NOW?

Management and Support

- Cluster 4 and 5 are performing – Use their strengths
- Cluster 2 and 3 need attention

Marketing

- 1,093 popular venues within 100m of 89 clubs (average 12 venues per club)
- Huge potential to promote clubs to people around!
- Cluster 1 has most popular neighbourhood (average 33 venues per club!)
- Devise promotion plans for clubs

Realignment

- Use insights on similarities and dissimilarities help in yearly realignment
- For example:
  - Cluster 1 clubs in one division (19 clubs)
  - Cluster 3 and 4 clubs in one division (16 clubs)
  - Remaining 54 clubs split among 3 divisions (18 clubs each)
- Benefits
  - About equal number of clubs per division
  - Good mix of strong and less strong clubs in each division
  - Geographical proximity in each division

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.

End Note

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

For a more complete elaboration of this project, refer to the full report at https://github.com/rickysoo/clustering-toastmasters/blob/master/report.pdf

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.

Thank You!