# 2019
# Research Experience for Undergraduates

## Diversity and its Correlation to Group Performance

Faith Grice
Advisor: Dr. Rizk

Final Presentation
August 9th, 2019

UNIVERSITY of HOUSTON

# Goal

To determine if there is a correlation between gender diversity in groups and group performance in hopes of finding the best way to group students

# Objectives

- Form groups of students using data mining methods
- Calculate the diversity index and cohesion for each group and find correlation between them
- Evaluate how well the algorithm chooses groups

# Objective 1: Tasks

- Generate a random acquaintance matrix, randomly assign genders
- Compute distance matrix
- Apply hierarchical clustering, k-means, and dbscan on distance matrix

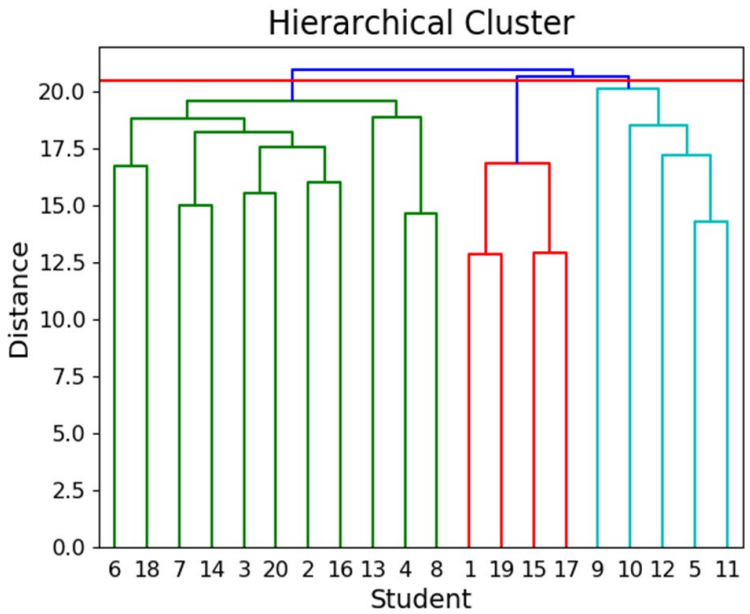
# Objective 1: Methodology

- Hierarchical Clustering and K-means
  - Chose optimal grouping by best silhouette score

- DBSCAN
  - Epsilon = 25, min # of points per cluster = 2

# Objective 1: Results

- Clusters from each method



# Objective 2: Tasks

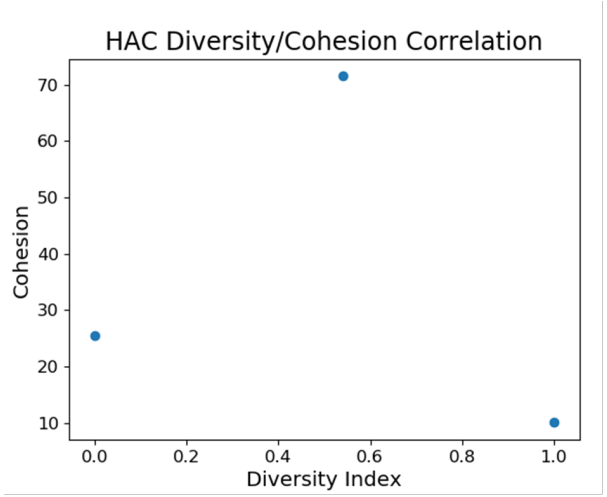
- For each clustering method
  - Calculate the cohesion for each cluster
  - Calculate the diversity for each cluster
  - Plot
  - Find correlation of diversity and cohesion

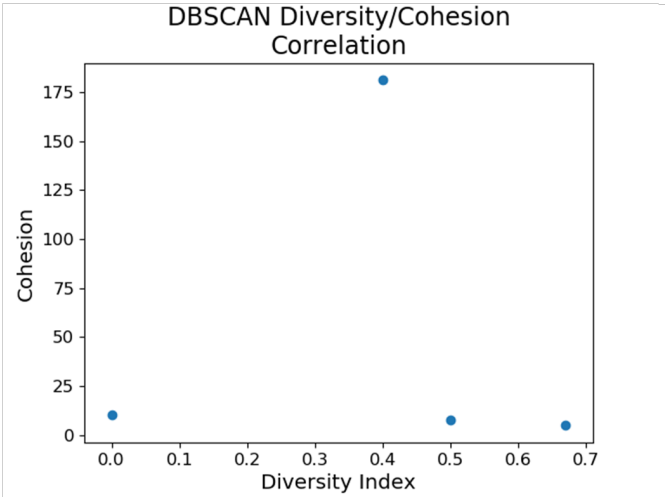
# Objective 2: Methodology

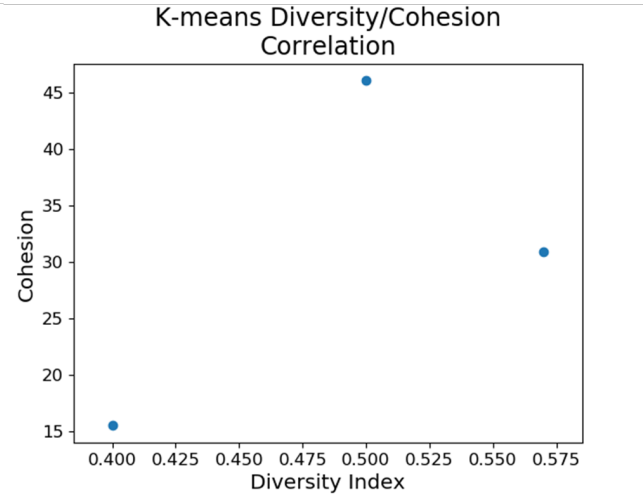
- Cohesion = Sum of Squared Error
- Diversity Index = $1 - \frac{|\#of\ males - \#of\ females|}{group\ cardinality}$
- Sci-py and matplotlib packages

# Objective 2: Results

## Diversity Index vs Cohesion




**Correlation Coefficients**

Hierarchical Clustering: -0.1968

DBSCAN: -0.00837

K-means: 0.5887

UNIVERSITY of HOUSTON

# Objective 3: Tasks

- Compute silhouette score for each method
- Determine best clustering method by silhouette score

# Objective 3: Methodology

- Compute the silhouette score

The **silhouette** ranges from −1 to +1, where a high value indicates that the object is well matched to its own **cluster** and poorly matched to neighboring **clusters**. If most objects have a high value, then the **clustering** configuration is appropriate.

# Objective 3: Results

**Silhouette Scores**

Hierarchical Clustering: 0.1243

DBSCAN: 0.0515

K-Means:0.1134

# Deliverables

**Cluster Validation** is used to evaluate the goodness of clustering algorithm results: using cohesion, correlation and silhouette score

- Cohesion of each cluster
- Correlation of diversity index and cohesion of each clustering method
- Silhouette score for each clustering method

# Limitations

- Randomized data
- Small matrix

# Future Work

- Apply this research to classroom data through surveys
- Further test clustering methods
- Use other factors than acquaintance

# Conclusions

- Silhouette scores are overall low for each method
- Hierarchical clustering is best method
  - Refine
- Little correlation between diversity index and group cohesion
- With large datasets and more courses =>better results

# Acknowledgements

The REU project is sponsored by NSF under award NSF-1659755. Special thanks to the following UH offices for providing financial support to the project: Department of Computer Science; College of Natural Sciences and Mathematics; Dean of Graduate and Professional Studies; VP for Research; and the Provost's Office. The views and conclusions contained in this presentation are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.