Clustering of U.S. Universities

Project 2 - ANLT 212 Fall 2020

By Cameron Swanson and Marisol Hernandez

Case Study: San Francisco County Office of Education

Problem: College applicants have many factors to consider when deciding where to apply, including location, available majors, and standardized test scores -- these factors often make for a difficult choice.

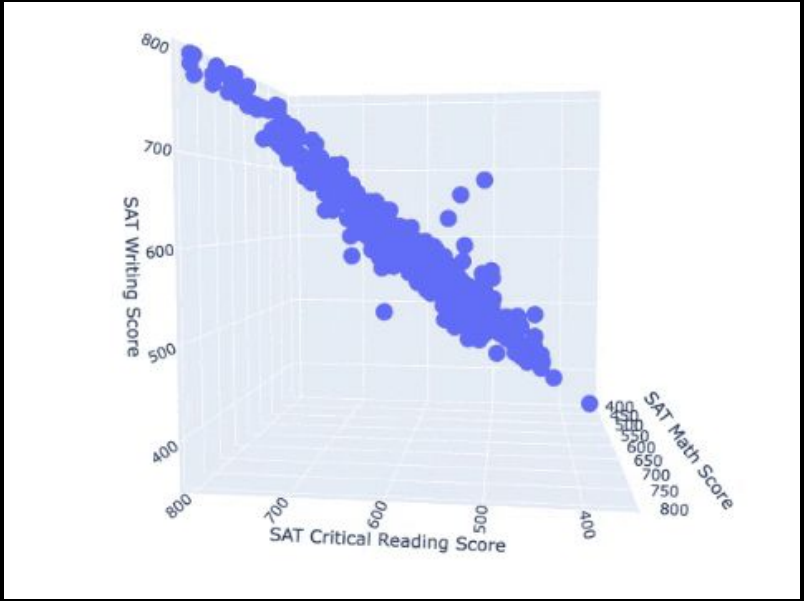
Goal: We built a recommendation engine that suggests schools to students based on their SAT component scores. High school counselors can use this model to help their students choose where to apply.

Data Collection

- IPEDS - International Postsecondary Education Data System
    - Part of the National Center for Education Statistics
- 1534 U.S. universities, 145 features
- After cleaning: 705 universities, 3 features

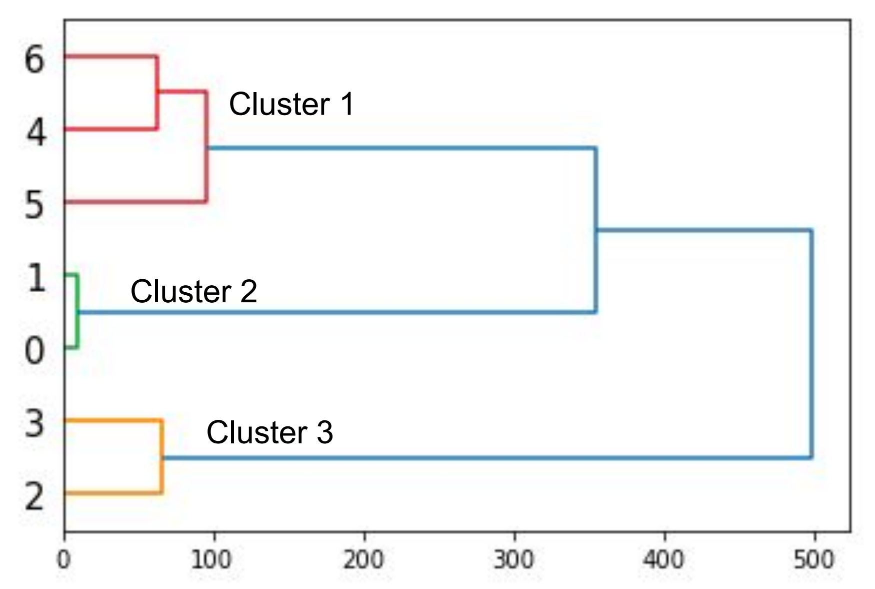
Descriptive Visualization

- 3D Scatter Plot
- Depicts universities by 75th percentile of SAT component scores (Reading, Writing, Math)
- Each data point:
  - University name
  - Accepted scores for each component



Cluster Analysis

- Hierarchical clustering allows us to determine optimal number of clusters (k)
- Assign clusters to each point, iteratively combine according to shortest distance
- Plot results with dendrogram
- For our data, k = 3

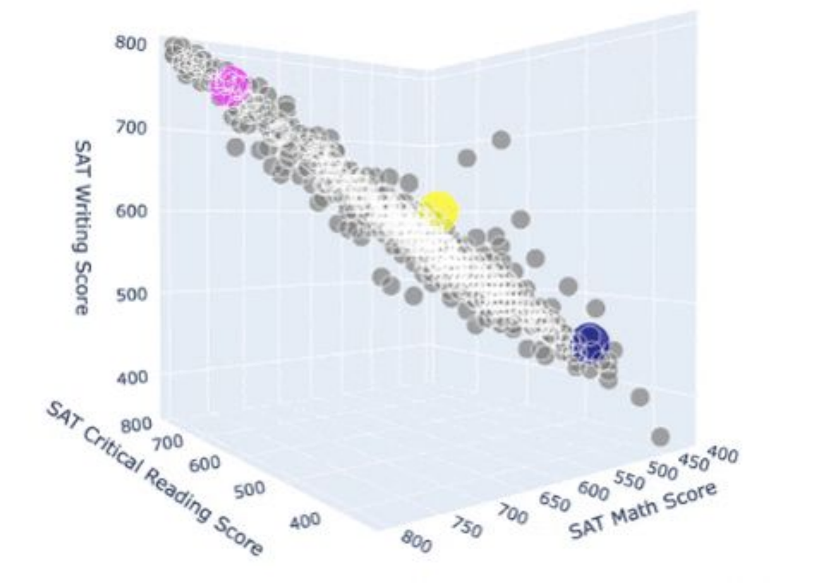


K-means Clustering

- K-Means clustering – groups data into clusters based on similarities

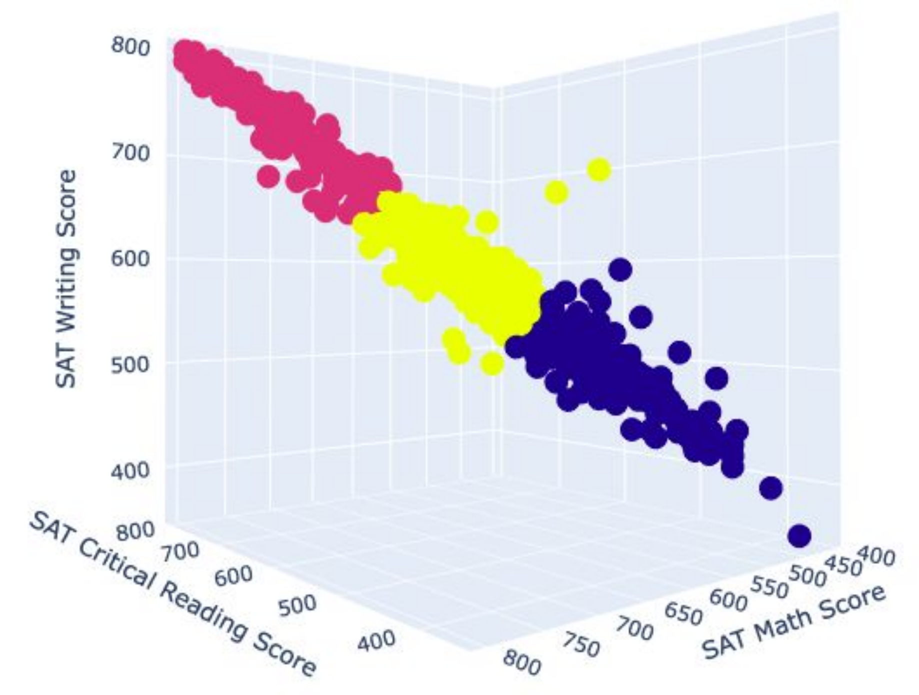
- To do this, we define *k* number of centroids
    - Centroid = the center of a cluster

- We established *k = 3* random centroids



K-means Algorithm

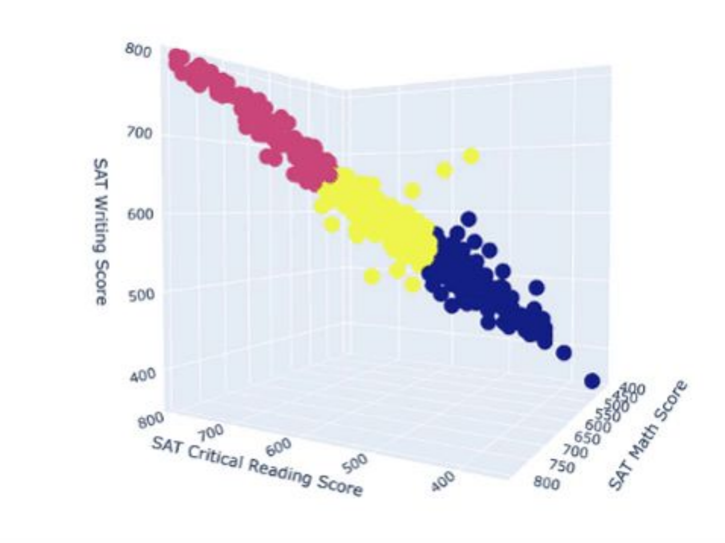
- K-Means algorithm iteratively assigns each point to a centroid until the positions of the centroids are optimized.

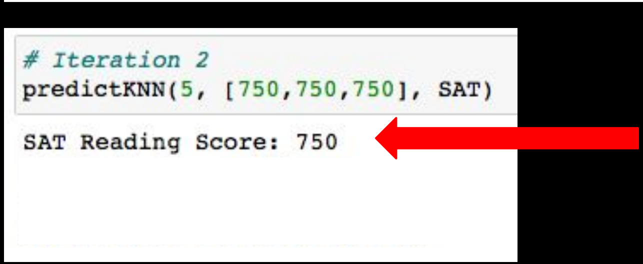


K-Nearest Neighbors Function

- predictKNN( ) takes in a data point and classifies it based on the classification of its 5 nearest neighbors

- Data point = SAT scores in each component



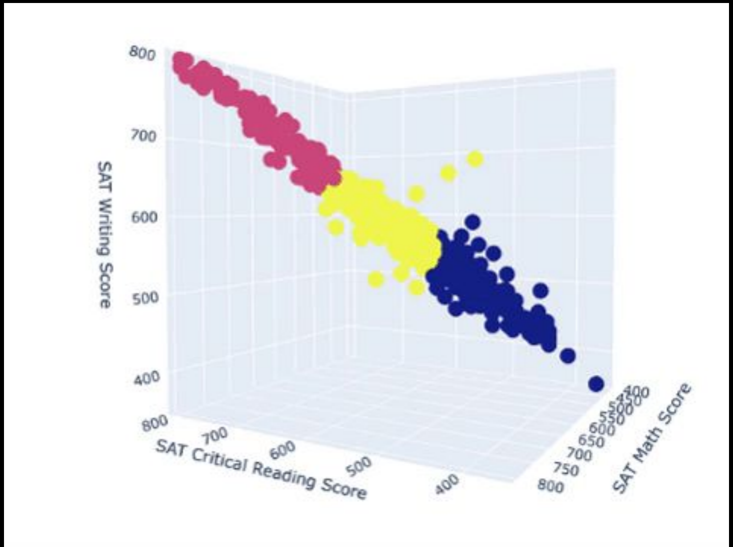


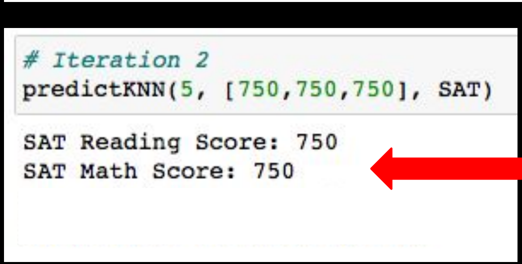
```
# Iteration 2
predictKNN(5, [750,750,750], SAT)

SAT Reading Score: 750
```

K-Nearest Neighbors Function

- predictKNN( ) takes in a data point and classifies it based on the classification of its 5 nearest neighbors

- Data point = SAT scores in each component



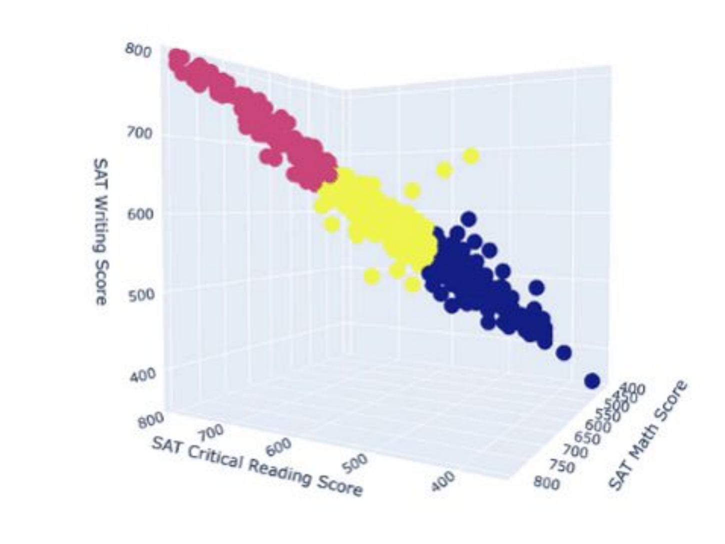


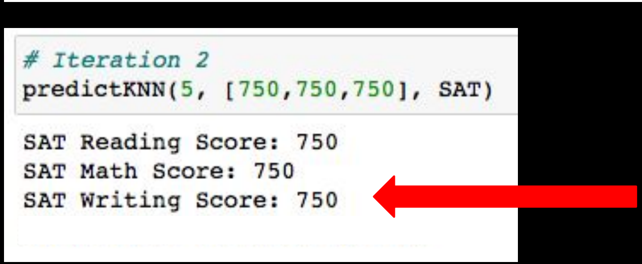
```
# Iteration 2
predictKNN(5, [750,750,750], SAT)

SAT Reading Score: 750
SAT Math Score: 750
```

K-Nearest Neighbors Function

- predictKNN( ) takes in a data point and classifies it based on the classification of its 5 nearest neighbors

- Data point = SAT scores in each component



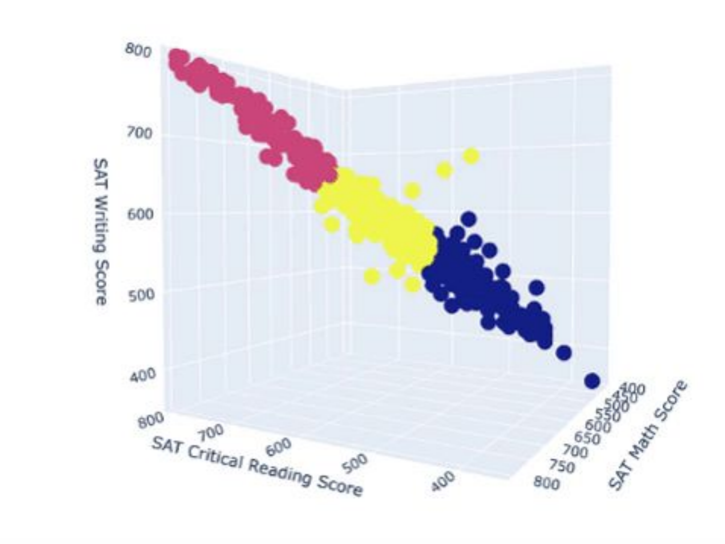


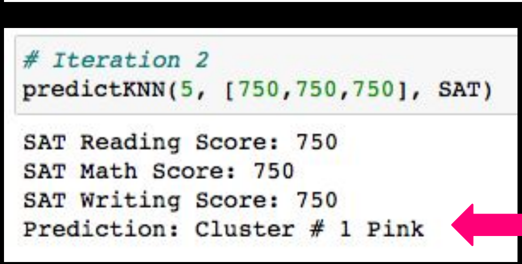
```
# Iteration 2
predictKNN(5, [750,750,750], SAT)

SAT Reading Score: 750
SAT Math Score: 750
SAT Writing Score: 750
```

K-Nearest Neighbors Function

- predictKNN( ) takes in a data point and classifies it based on the classification of its 5 nearest neighbors

- Data point = SAT scores in each component





```
# Iteration 2
predictKNN(5, [750,750,750], SAT)

SAT Reading Score: 750
SAT Math Score: 750
SAT Writing Score: 750
Prediction: Cluster # 1 Pink
```

Conclusion

- We recommend the SF County Office of Education to utilize our recommendation engine to better assist students during the college application process

- Benefits:
  - Determine the schools a student should apply to
  - Keep for future use - can account for changes in SAT components
    - Universities can easily be added to the model