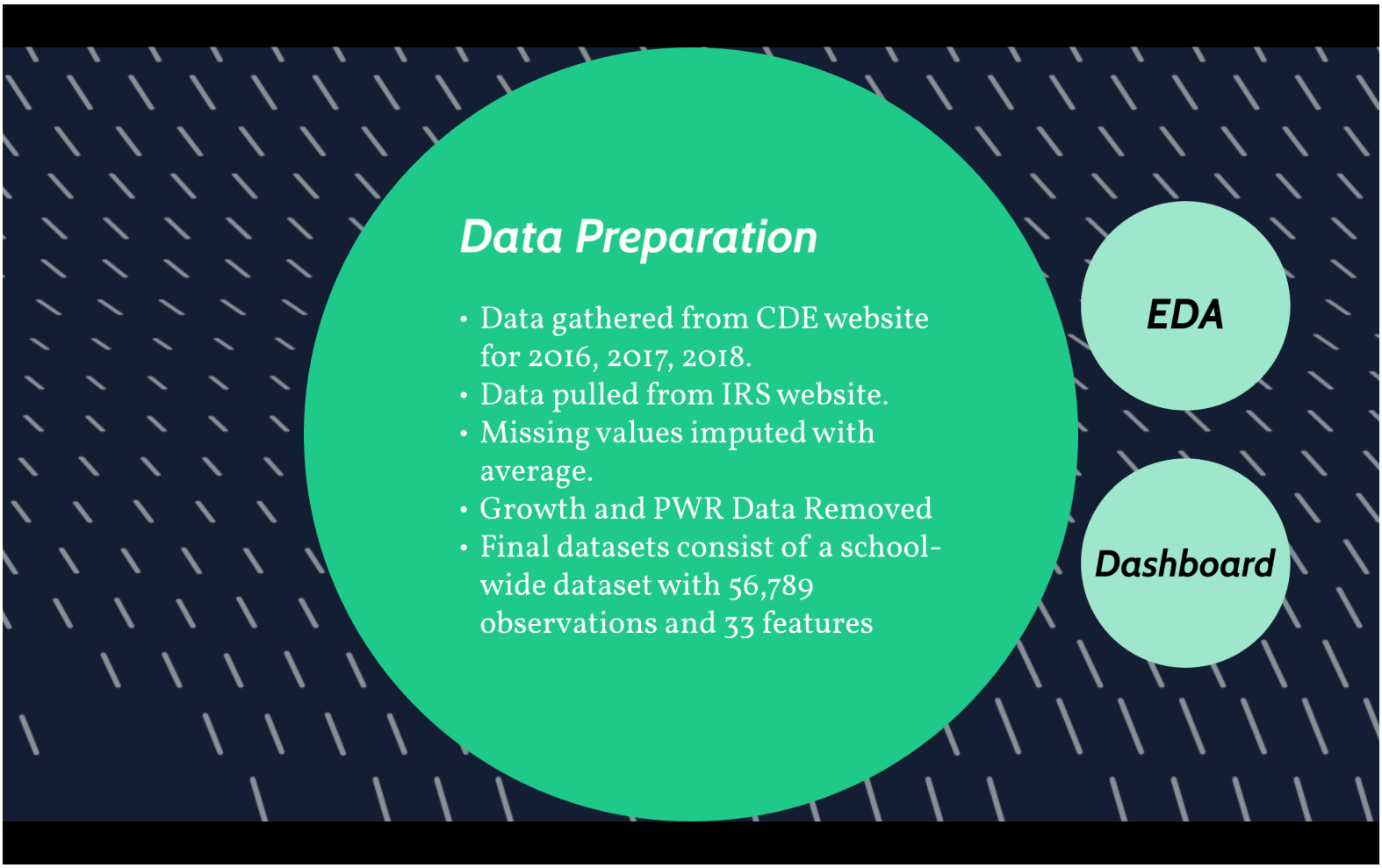
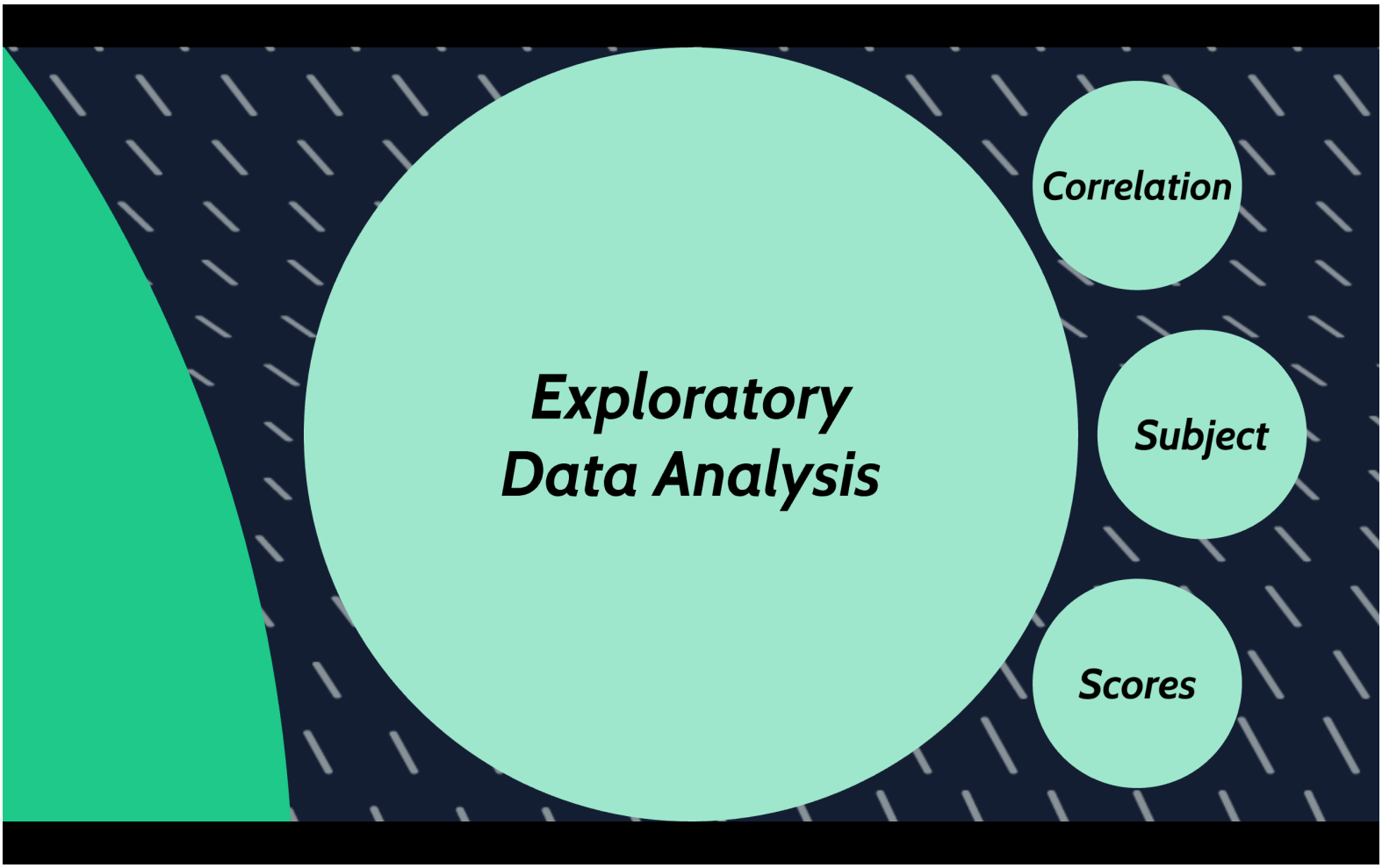
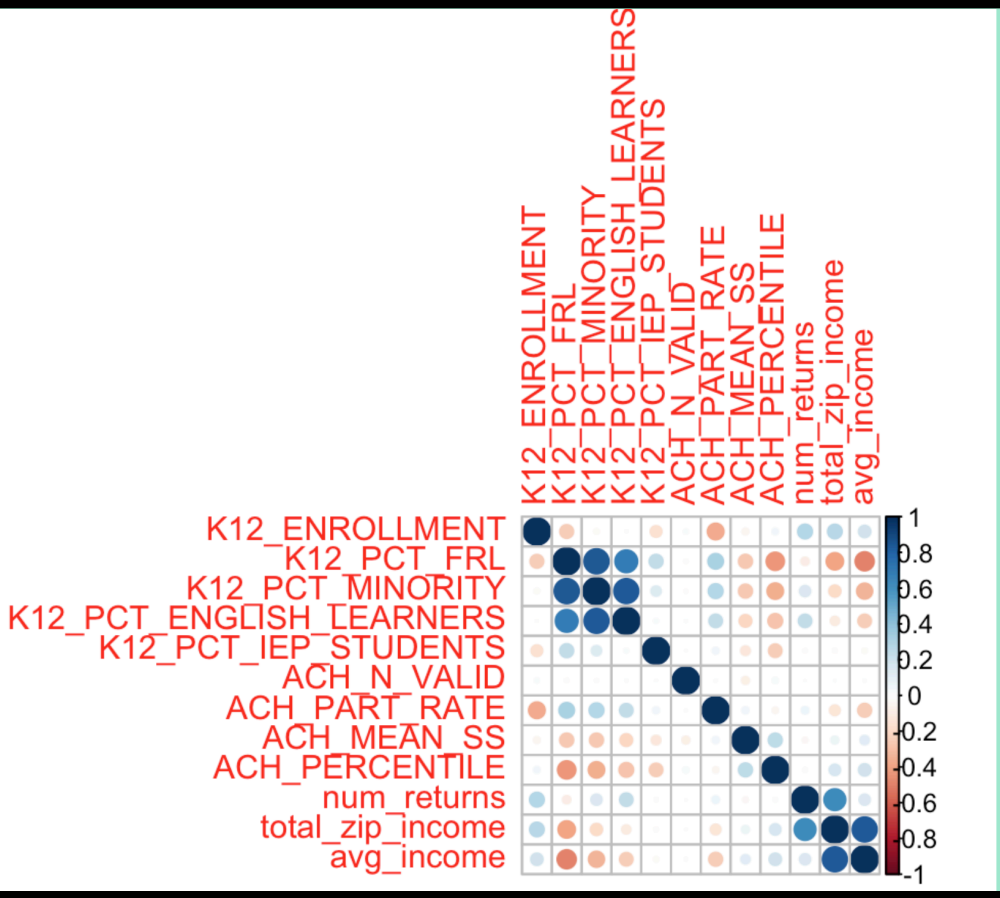
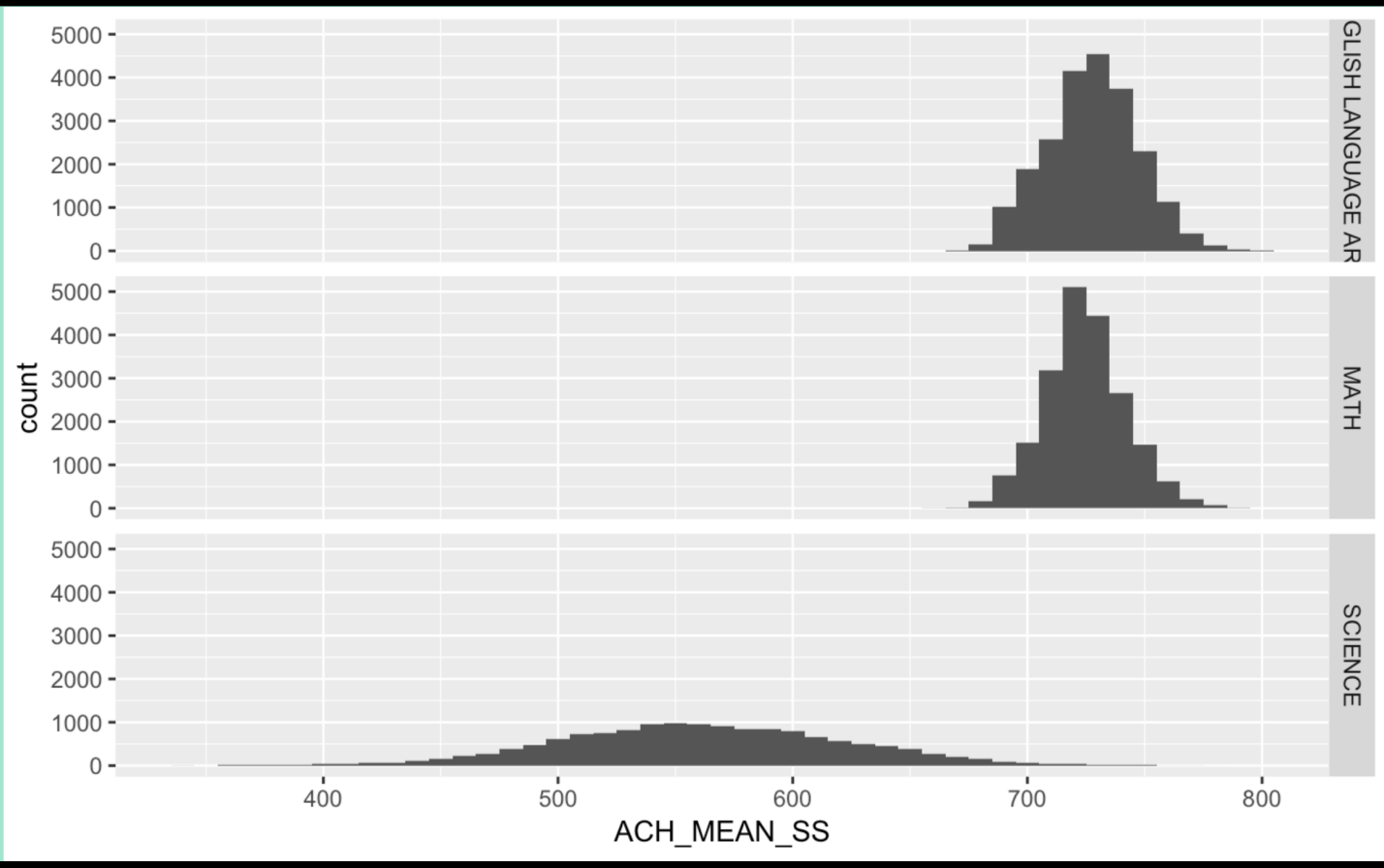
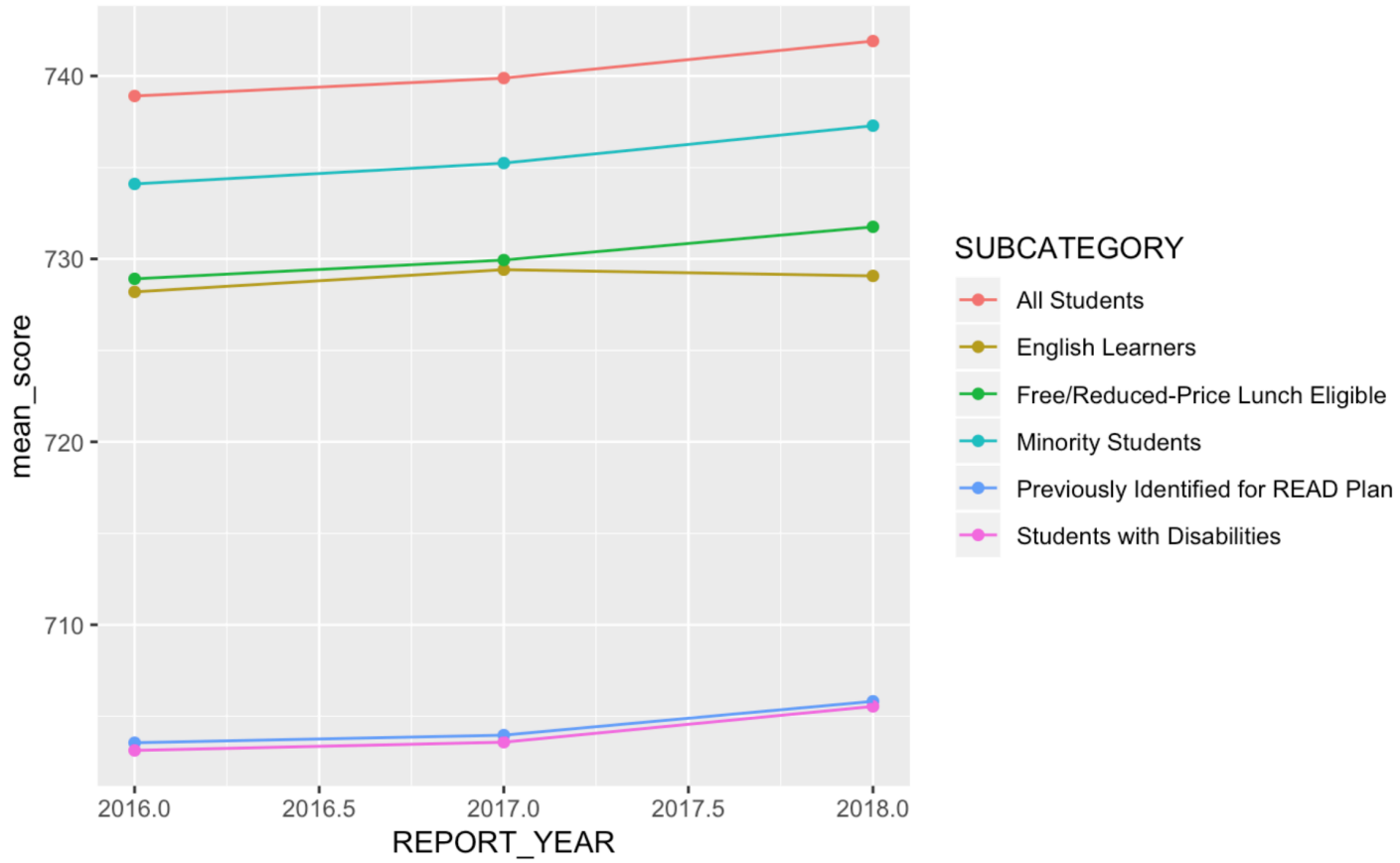

Overview
Clustering
Machine Learning
EDA and Visualization
Hypothesis Testing
Conclusions
Practicum I
Colorado Department of Education Data Analysis
Lee Wise

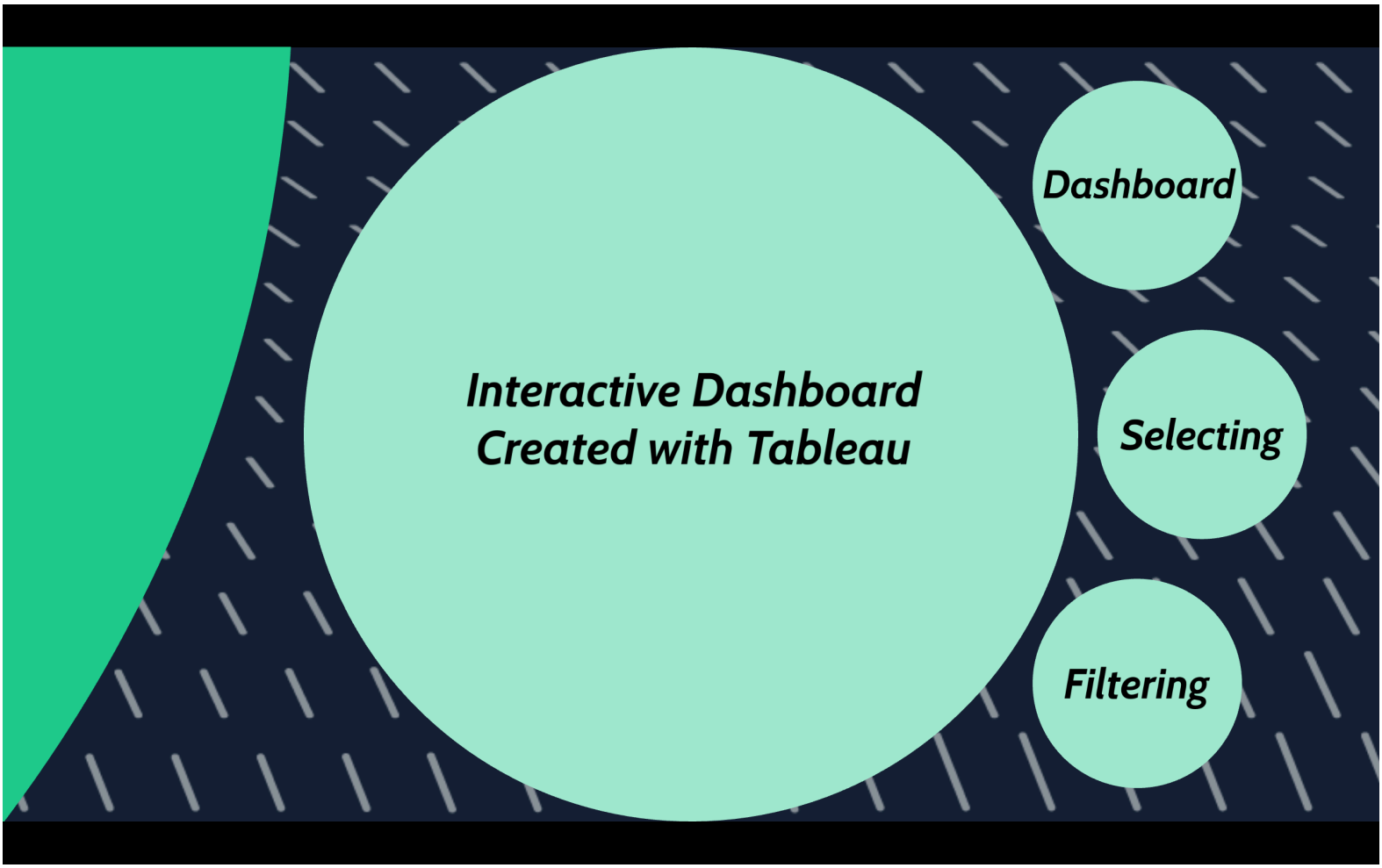

Data Preparation
• Data gathered from CDE website for 2016, 2017, 2018.
• Data pulled from IRS website.
• Missing values imputed with average.
• Growth and PWR Data Removed
• Final datasets consist of a school-wide dataset with 56,789 observations and 33 features
EDA
Dashboard


Exploratory
Data Analysis
Correlation
Subject
Scores

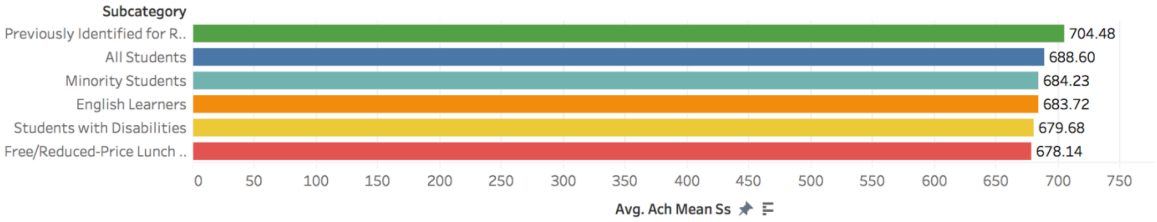

K12_ENROLLMENT
K12_PCT_FRL
K12_PCT_MINORITY
K12_PCT_ENGLISH_LEARNERS
K12_PCT_IEP_STUDENTS
ACH_N_VALID
ACH_PART_RATE
ACH_MEAN_SS
ACH_PERCENTILE
num_returns
total_zip_income
avg_income
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


GLISH LANGUAGE AR
MATH
SCIENCE
count
5000
4000
3000
2000
1000
0
400
500
600
700
800
ACH_MEAN_SS

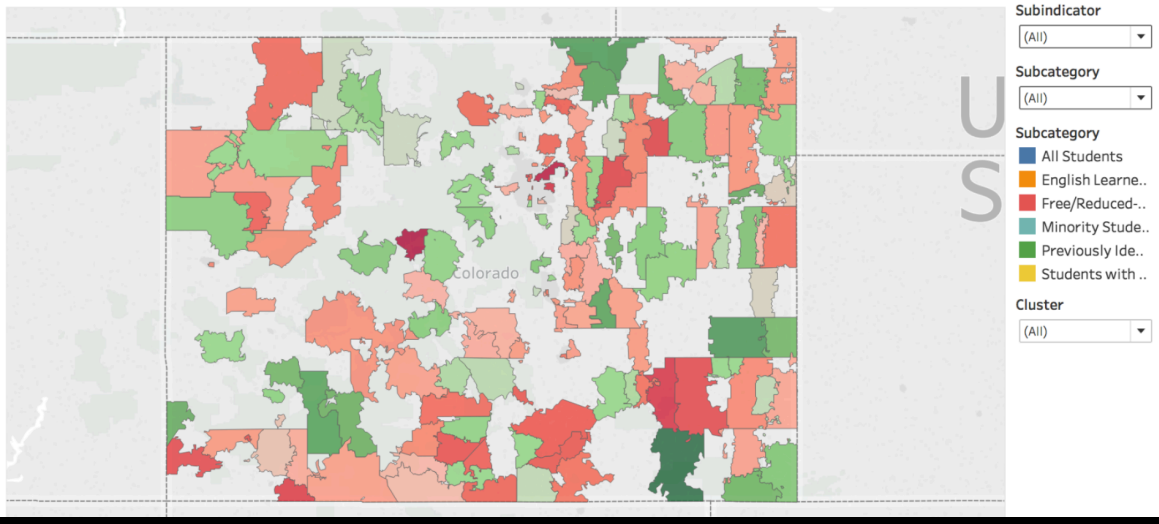

740
730
720
710
mean_score
2016.0
2016.5
2017.0
2017.5
2018.0
REPORT_YEAR
SUBCATEGORY
All Students
English Learners
Free/Reduced-Price Lunch Eligible
Minority Students
Previously Identified for READ Plan
Students with Disabilities


Interactive Dashboard
Created with Tableau
Dashboard
Selecting
Filtering

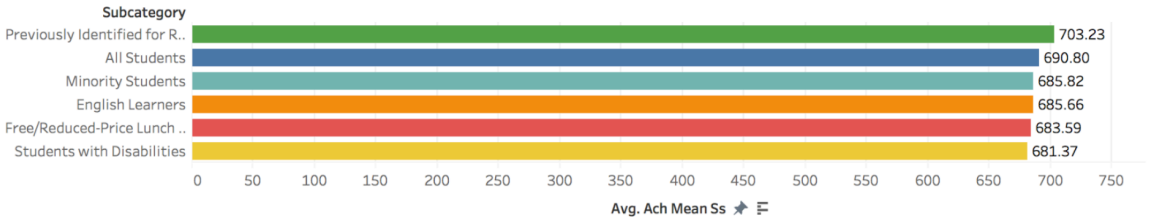
Average CMAS Score



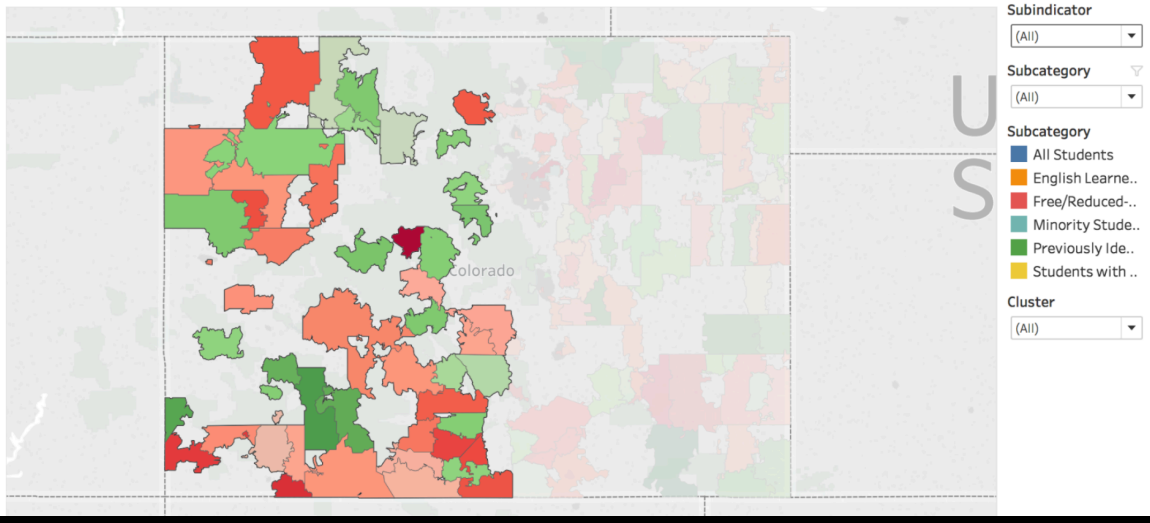
Average CMAS Score By Zip Code for All
Subject: All



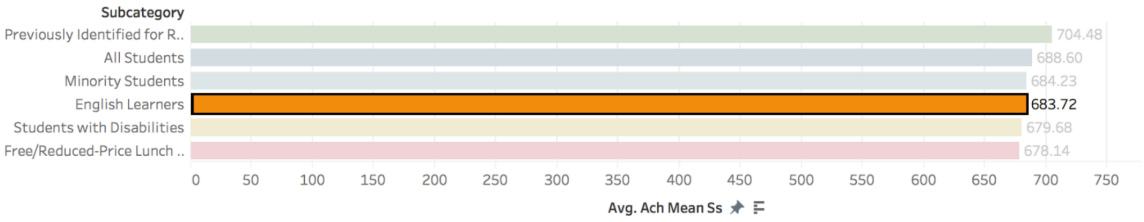
Average CMAS Score



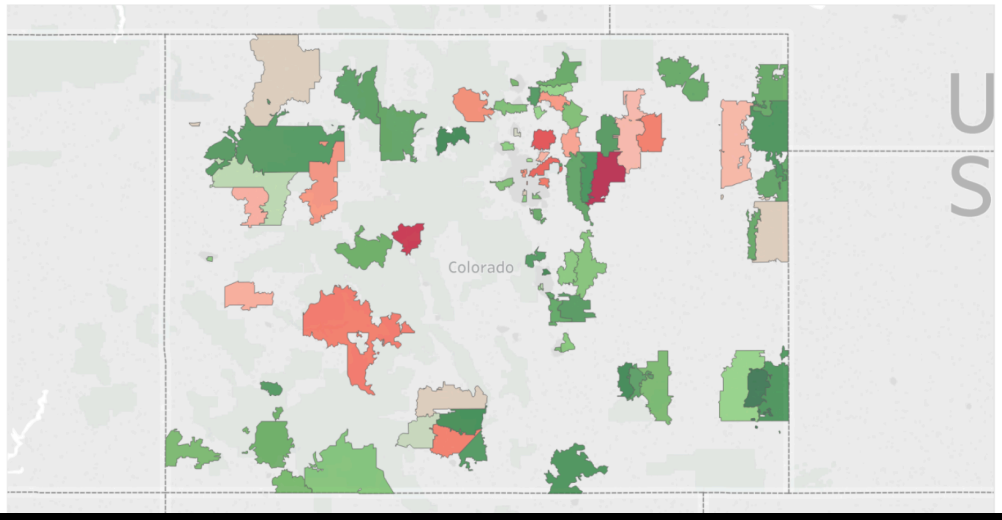
Average CMAS Score By Zip Code for All Subject: All

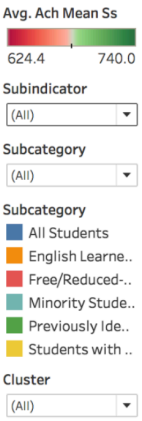


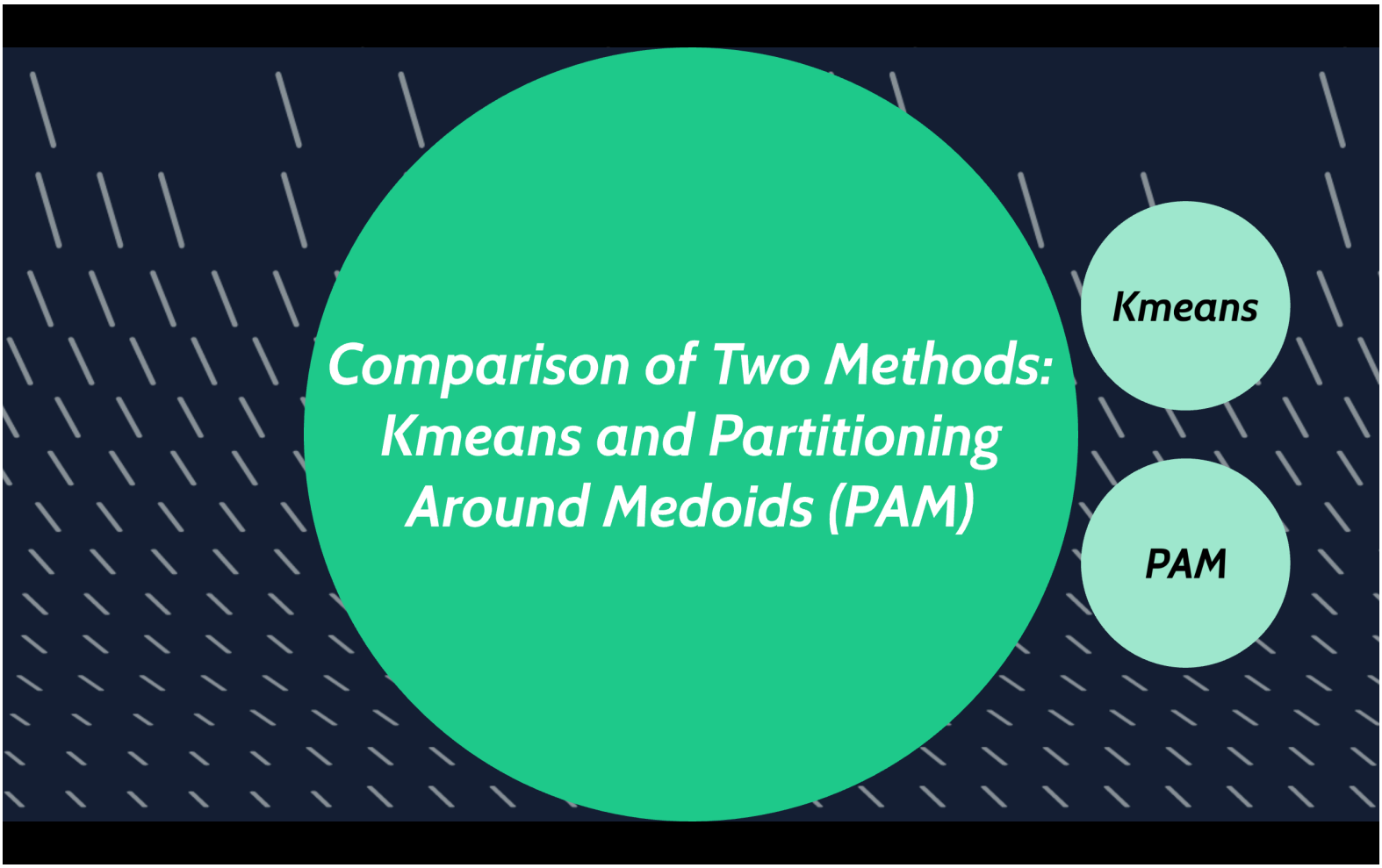
Average CMAS Score

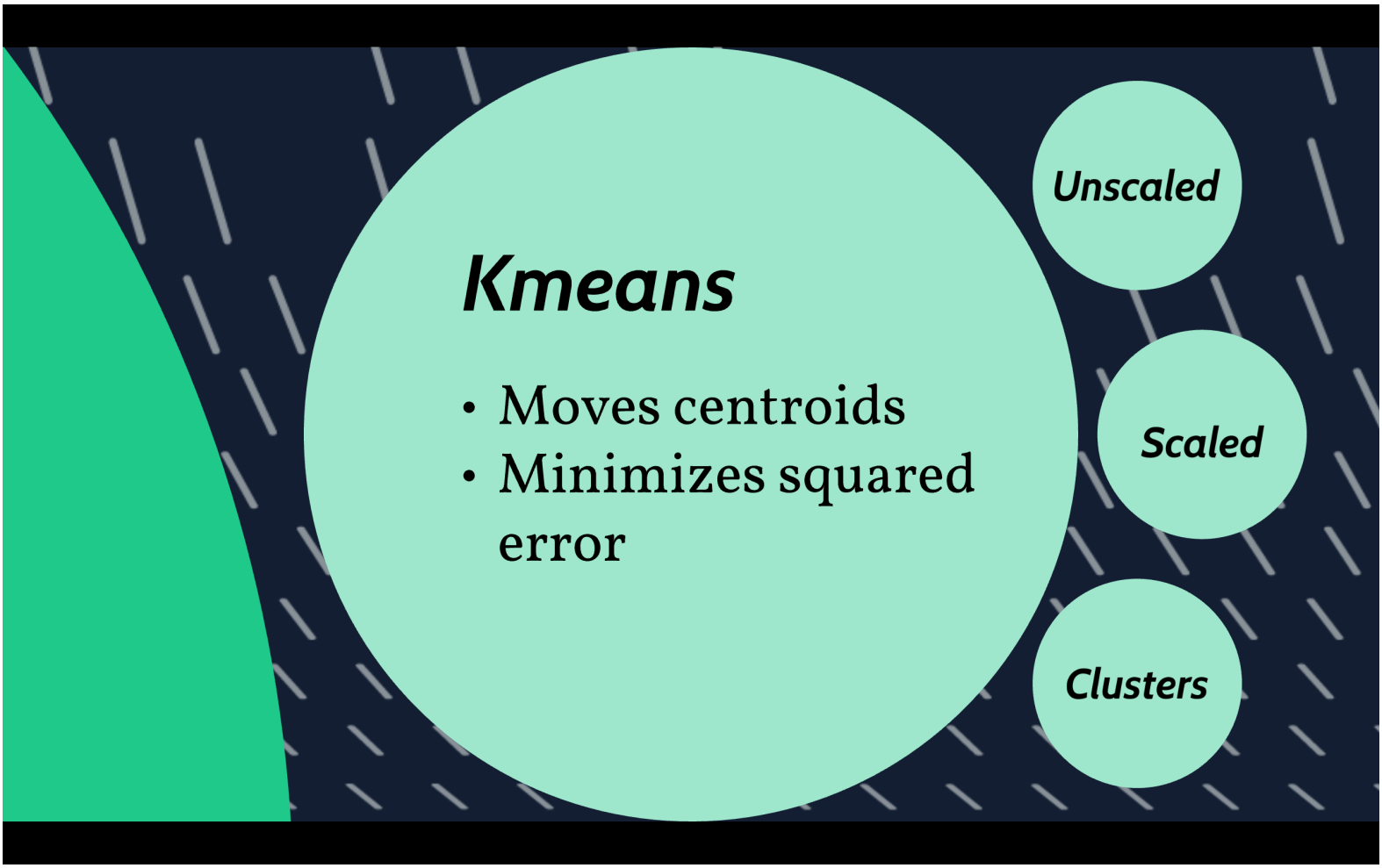


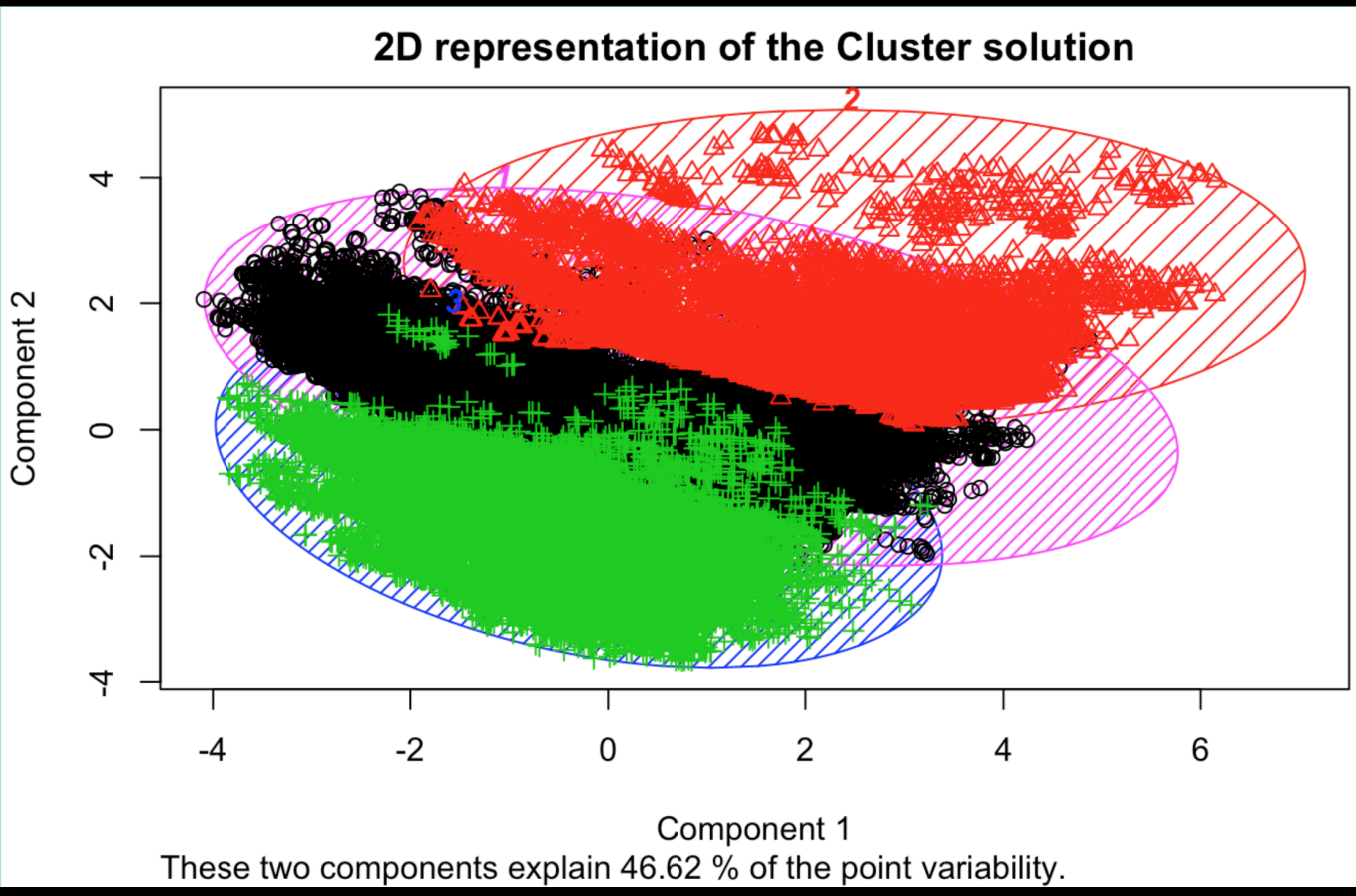
Average CMAS Score By Zip Code for None
Subject: All

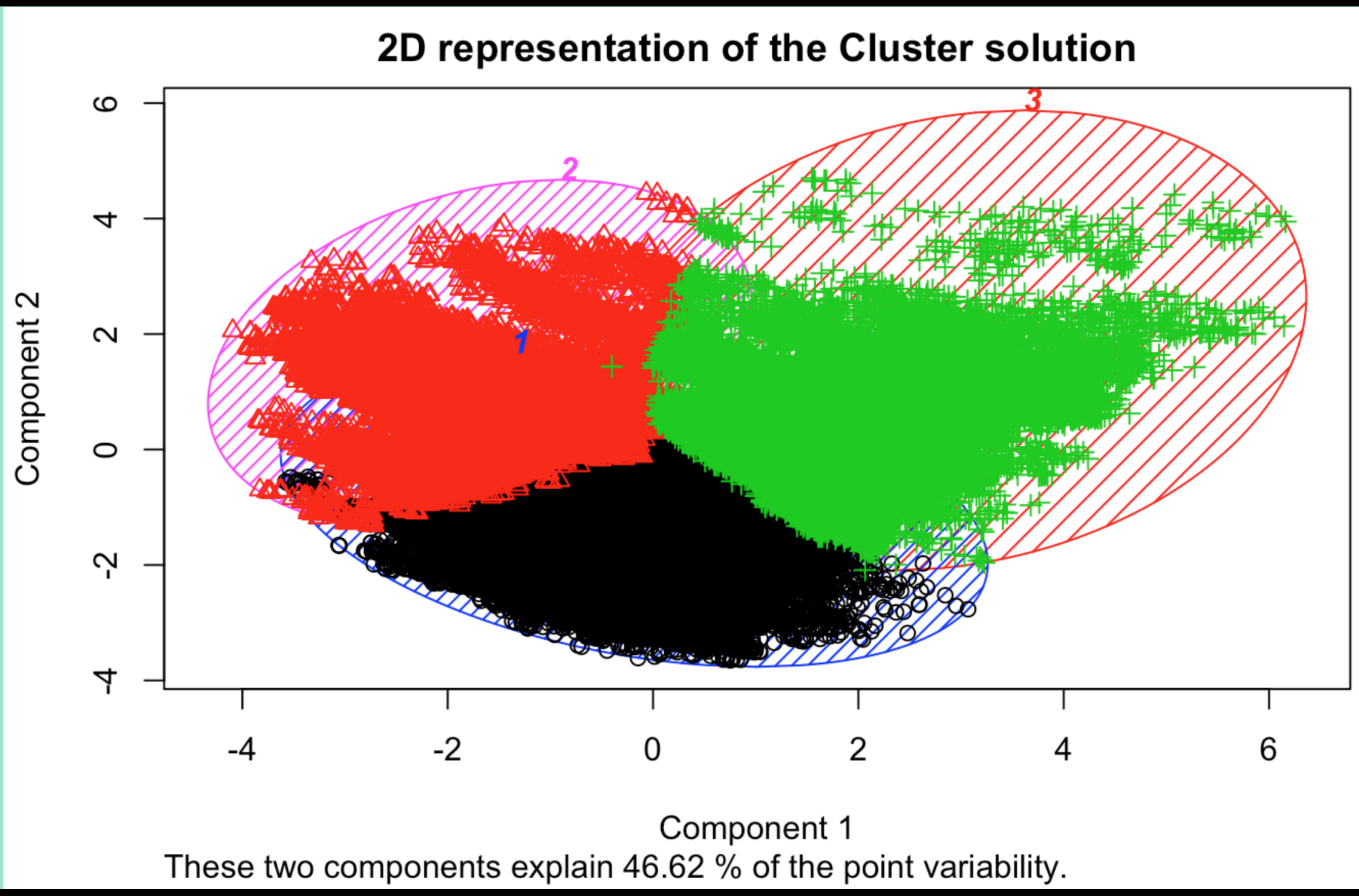





Comparison of Two Methods:
Kmeans and Partitioning
Around Medoids (PAM)
Kmeans
PAM

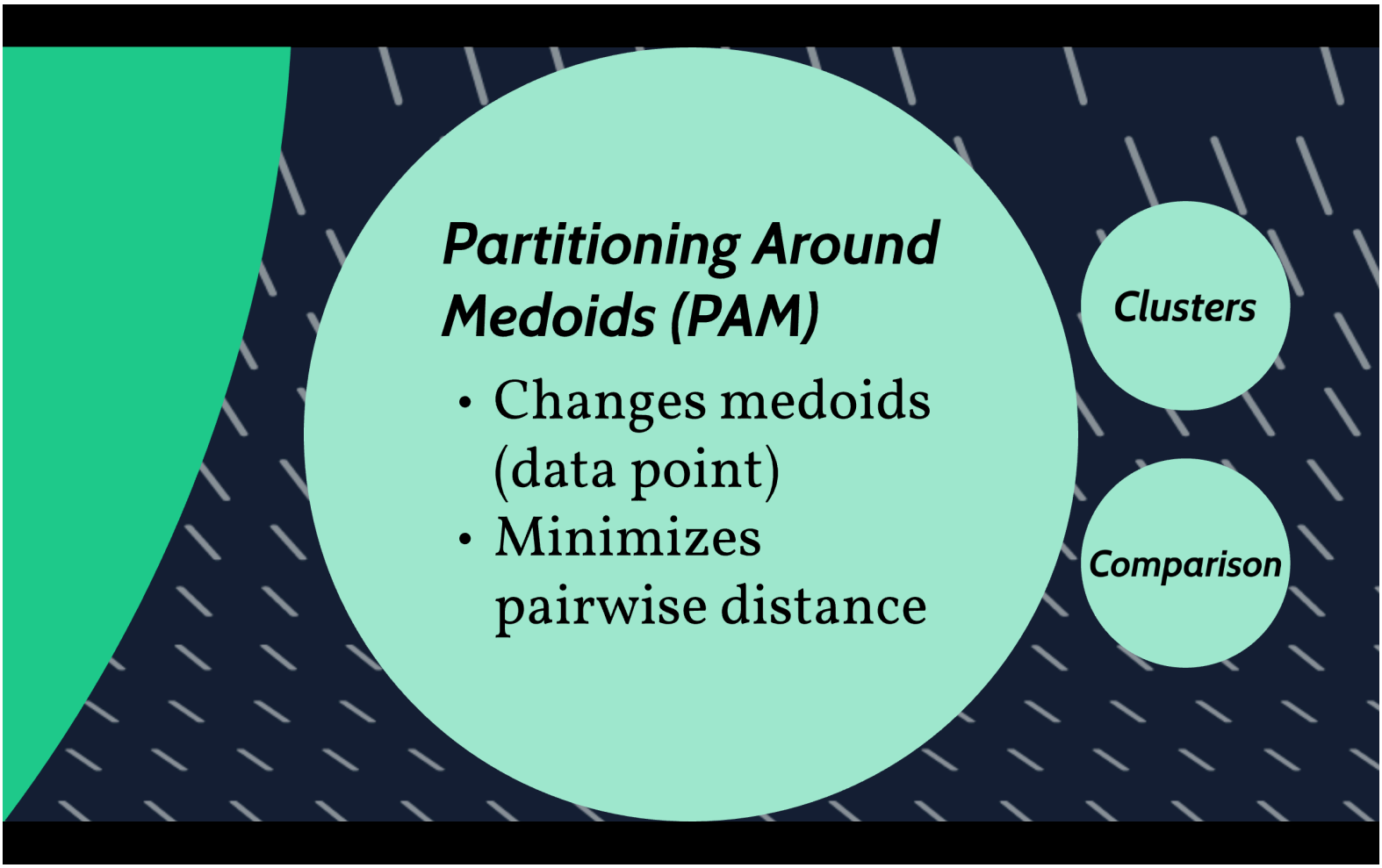

Kmeans
• Moves centroids
• Minimizes squared error
Unscaled
Scaled
Clusters

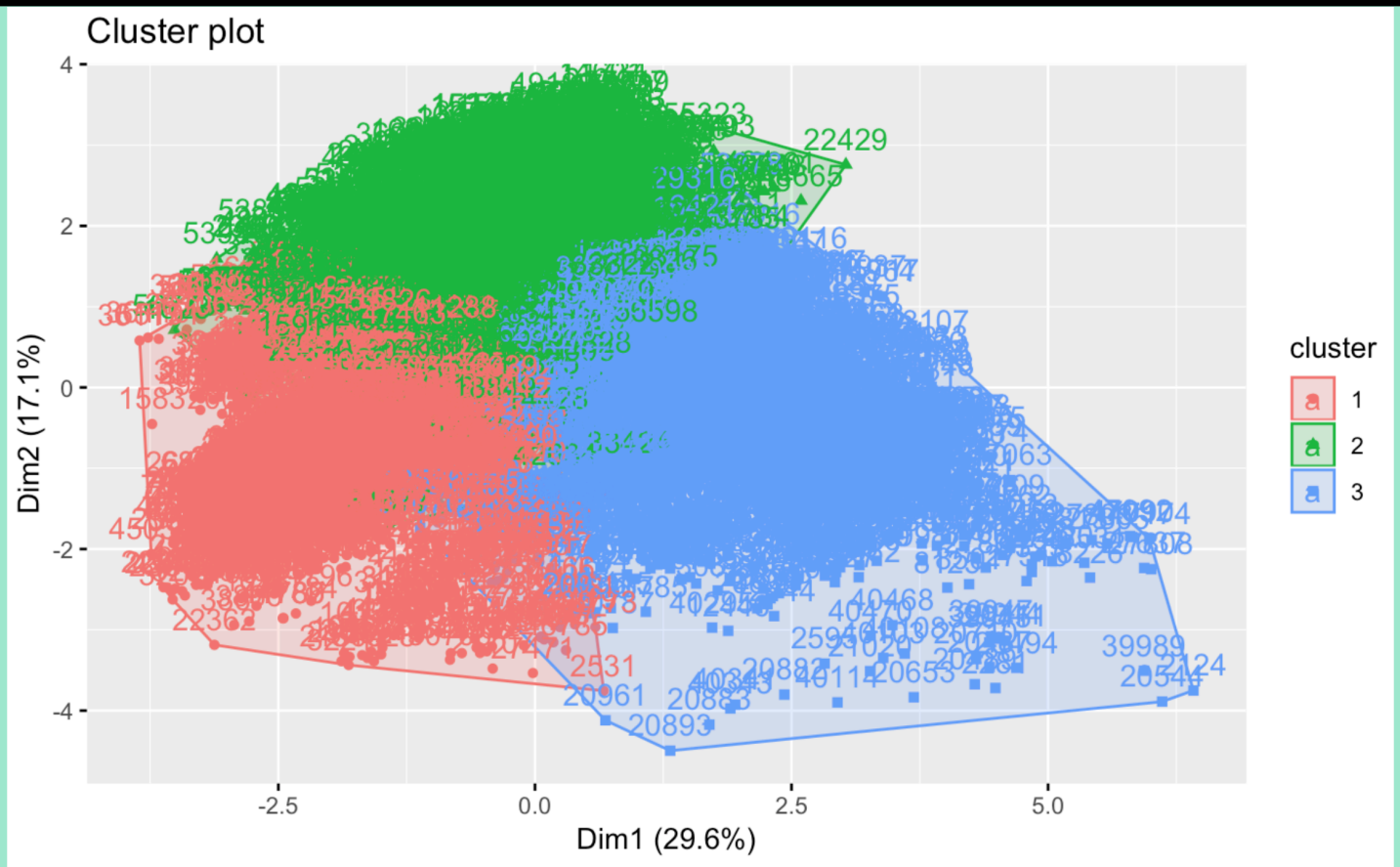

2D representation of the Cluster solution
2
3
Component 2
4
2
0
-2
-4
-4
-2
0
2
4
6
Component 1
These two components explain 46.62 % of the point variability.


2D representation of the Cluster solution
Component 2
6
4
2
0
-2
-4
-4
-2
0
2
4
6
Component 1
These two components explain 46.62 % of the point variability.
1
2
3

Characteristics

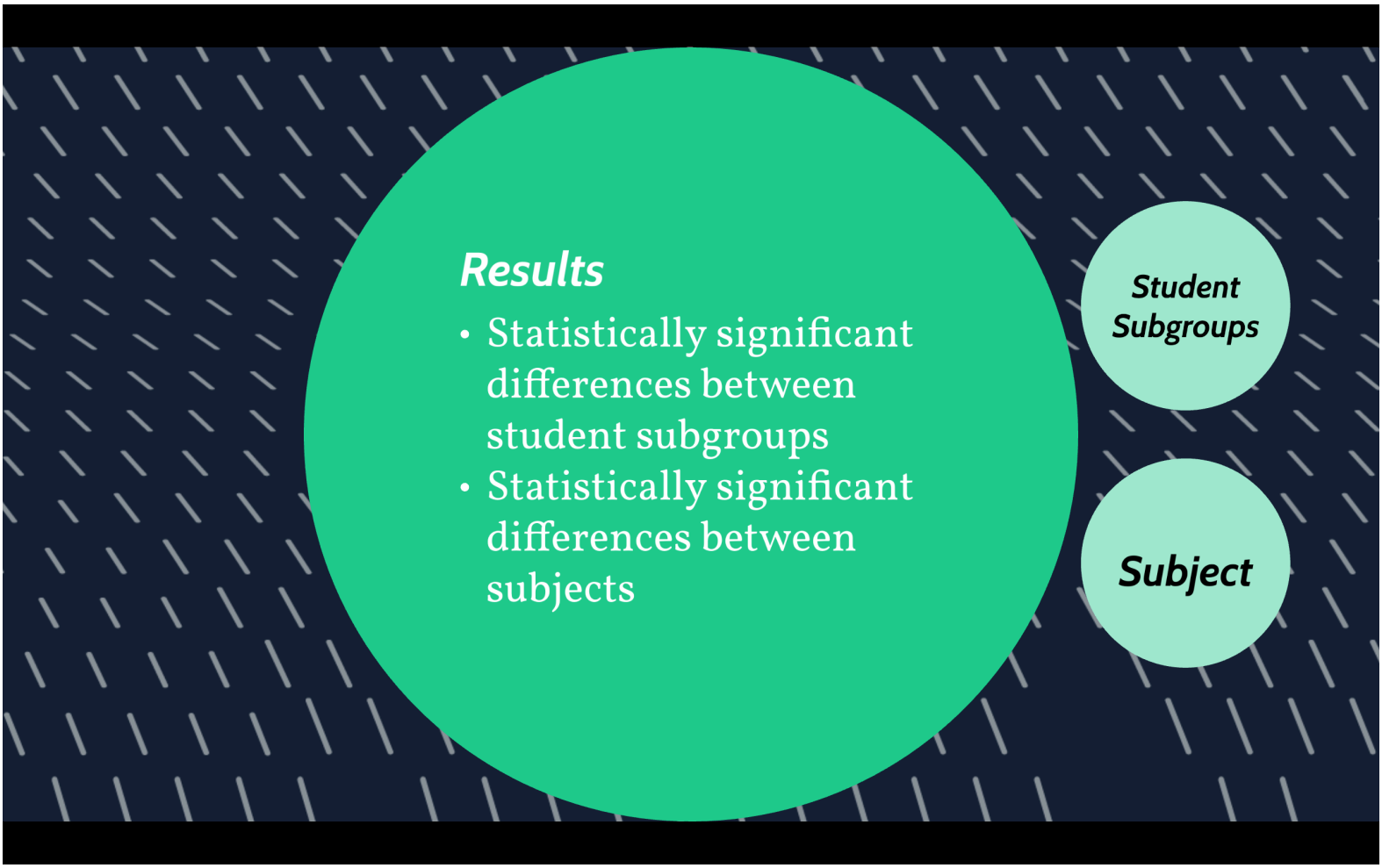
- *Custer 1: High enrollment, high minority and percent of English learners, lowest test scores.*
- *Cluster 2: High enrollment, low free/reduced lunch/ minority/English learner percentages, highest average income by a significant margin, highest test scores.*
- *Cluster 3: Low enrollment, high percent free/reduced lunch, medium test scores, low income.*

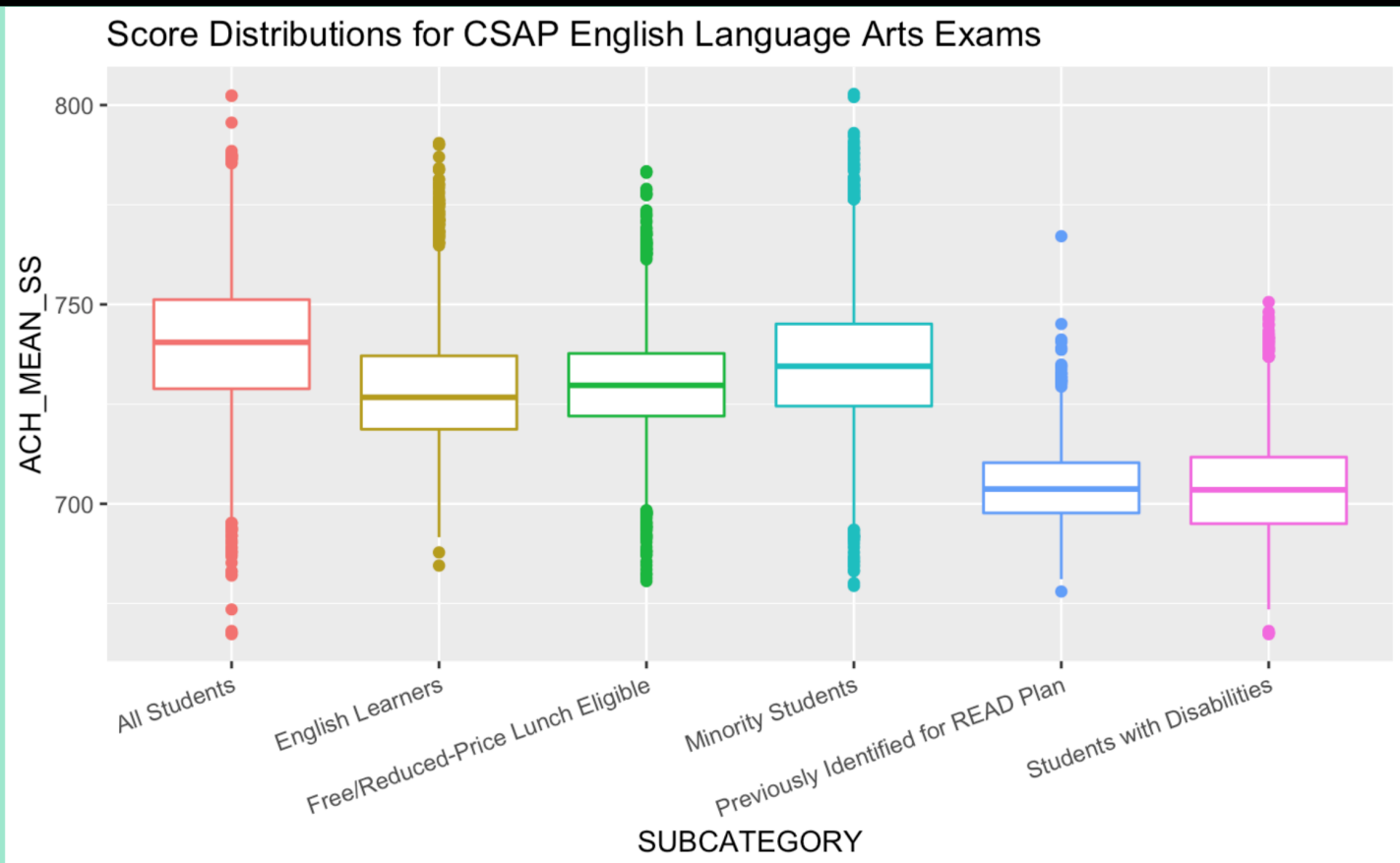

Partitioning Around Medoids (PAM)
• Changes medoids (data point)
• Minimizes pairwise distance
Clusters
Comparison

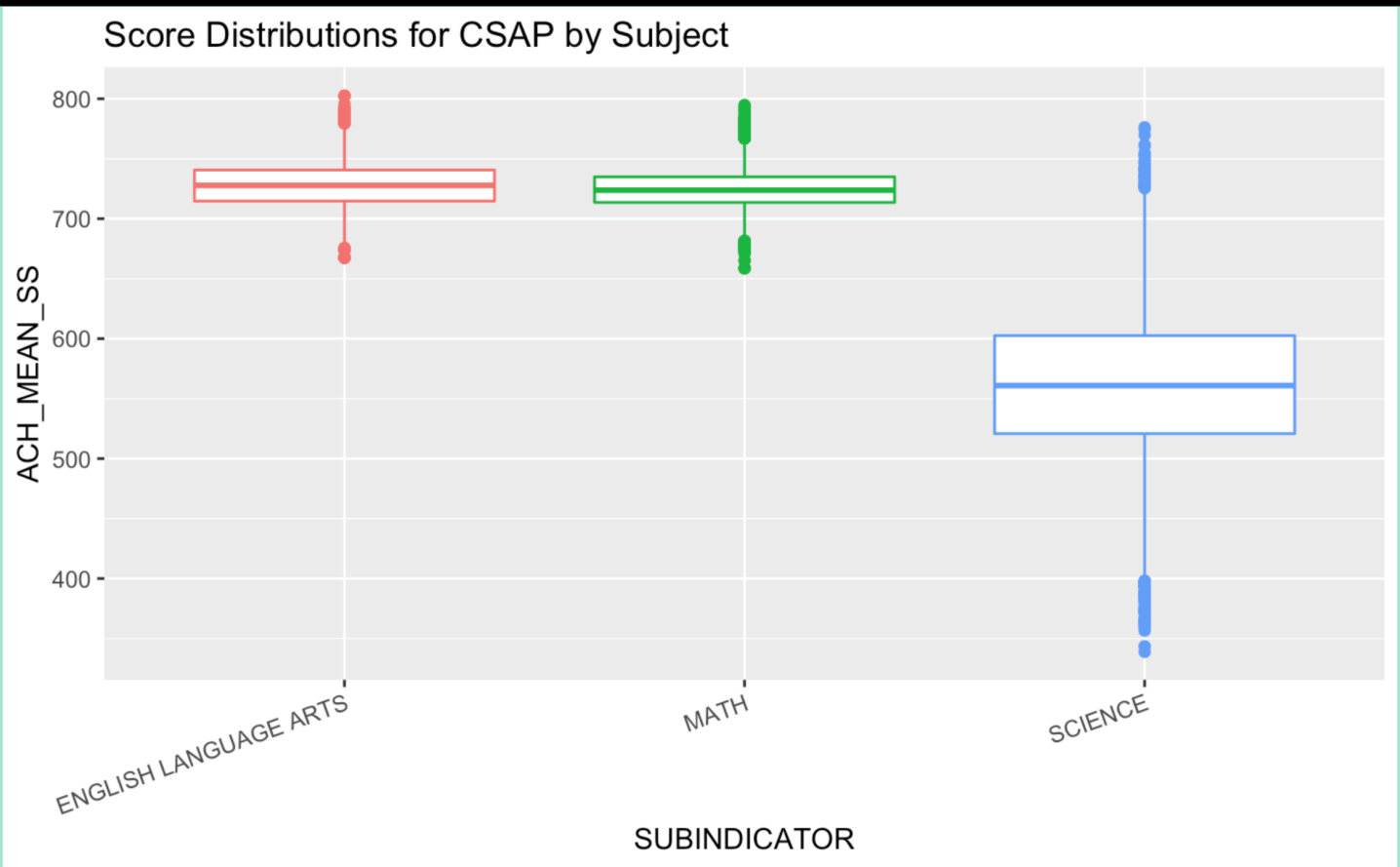

Cluster plot
Dim2 (17.1%)
Dim1 (29.6%)
4
2
0
-2
-4
-2.5
0.0
2.5
5.0
cluster
1
2
3
22429
2531
20961
20893
39989

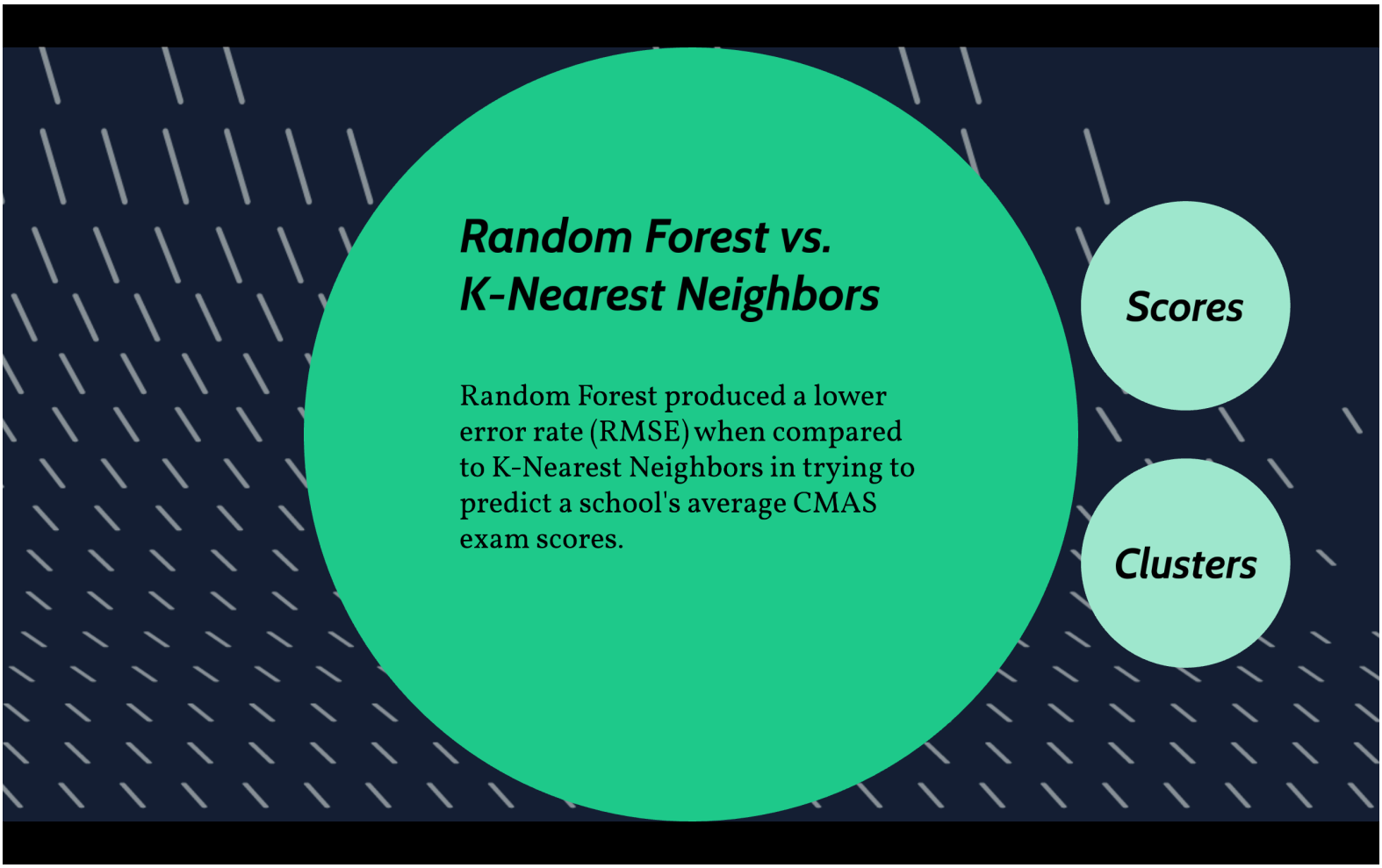
Agreement Between Kmeans and PAM

```
          Reference
Prediction    1    2    3
         1 1679   51   33
         2   96 1737  262
         3    6    9 1806

Overall Statistics

                Accuracy : 0.9195
```


Results
• Statistically significant differences between student subgroups
• Statistically significant differences between subjects
Student Subgroups
Subject


Score Distributions for CSAP English Language Arts Exams
800
750
700
ACH_MEAN_SS
All Students
English Learners
Free/Reduced-Price Lunch Eligible
Minority Students
Previously Identified for READ Plan
Students with Disabilities
SUBCATEGORY


Score Distributions for CSAP by Subject
800
700
600
500
400
ACH_MEAN_SS
ENGLISH LANGUAGE ARTS
MATH
SCIENCE
SUBINDICATOR

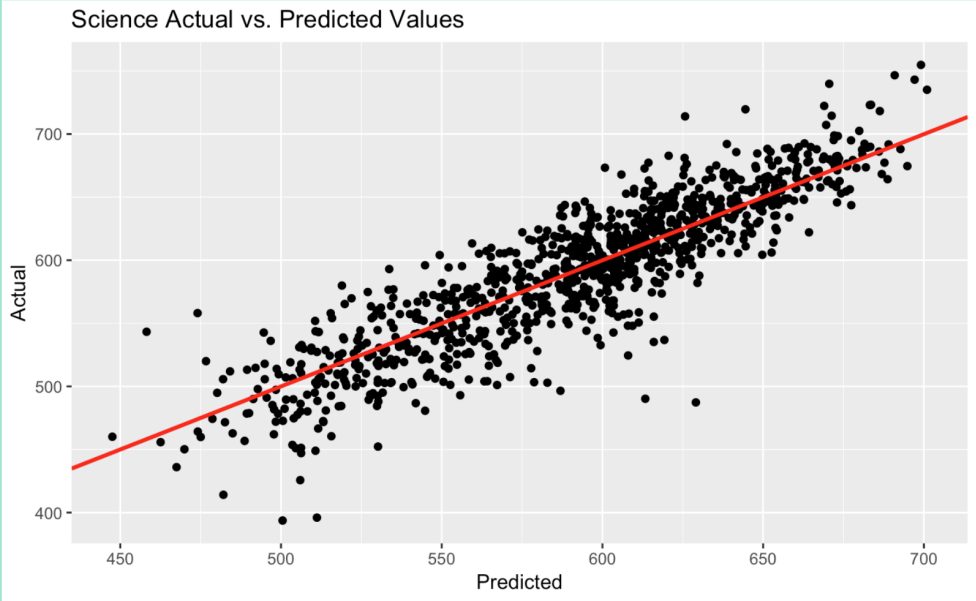

Random Forest vs.
K-Nearest Neighbors
Random Forest produced a lower error rate (RMSE) when compared to K-Nearest Neighbors in trying to predict a school's average CMAS exam scores.
Scores
Clusters

Predicting CMAS Scores Using RF

Science
- RMSE: 28.1
- Range: 393-754

Math:
- RMSE: 7.6
- Range: 679-777

English:
- RMSE: 8.3
- Range: 667-795

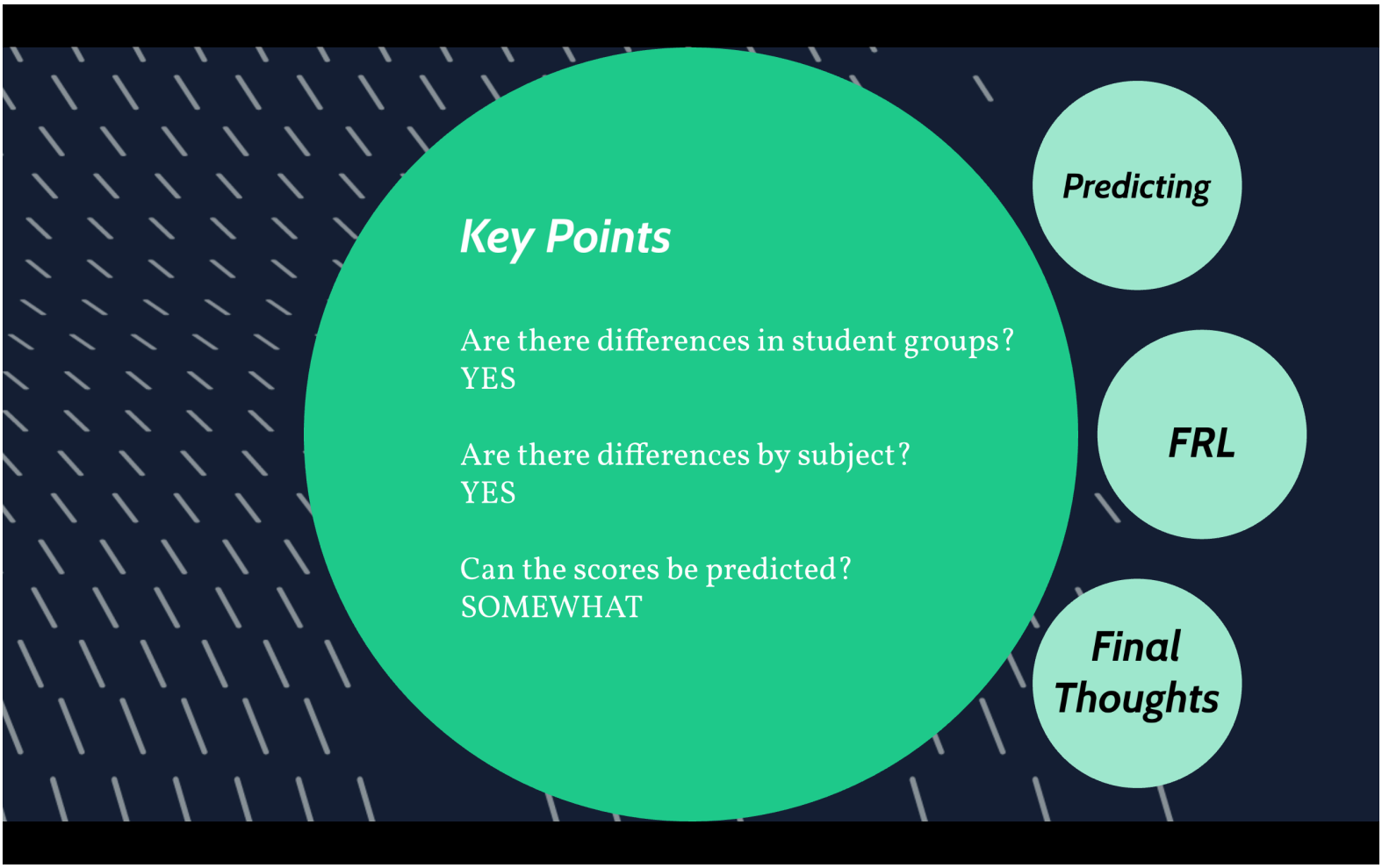


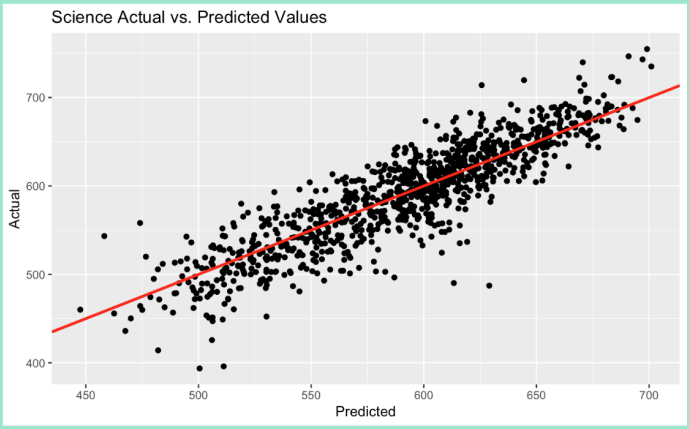
Predicting Clusters

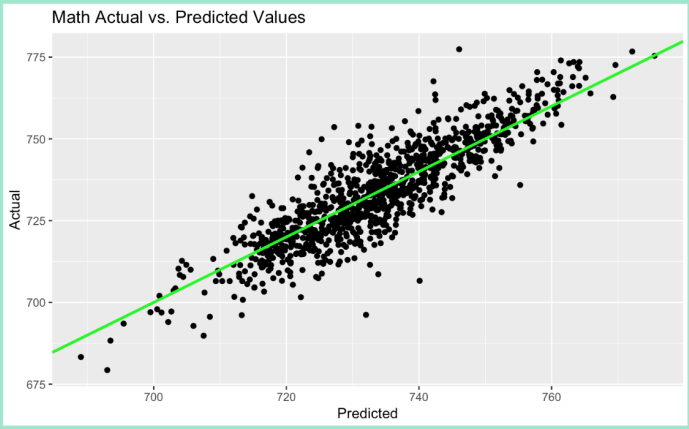
Accuracy: 97%
Kappa: 96.6%

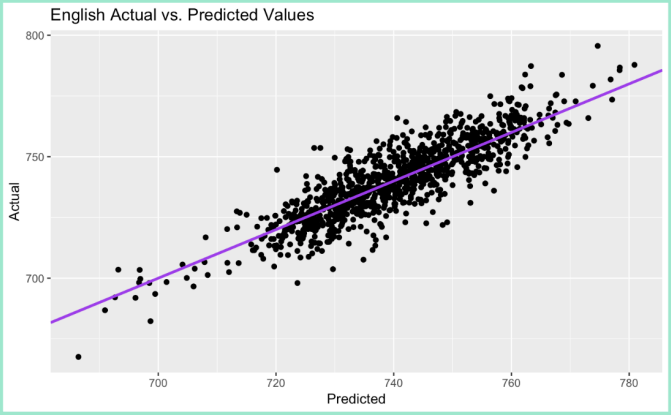
The clusters are reproducible and can be predicted using machine learning.

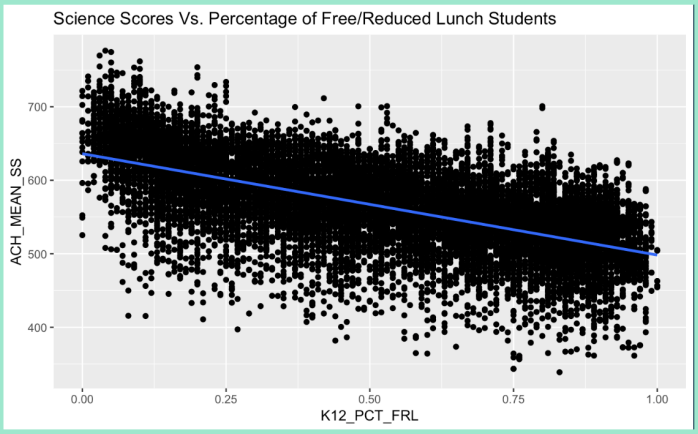
```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1 798  13  11
         2   9 690   5
         3   6   7 731

Overall Statistics

               Accuracy : 0.9775
                 95% CI : (0.9706, 0.9832)
    No Information Rate : 0.3581
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9662
 Mcnemar's Test P-Value : 0.4697

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Precision              0.9708   0.9801   0.9825
Recall                 0.9815   0.9718   0.9786
F1                     0.9761   0.9760   0.9805
```

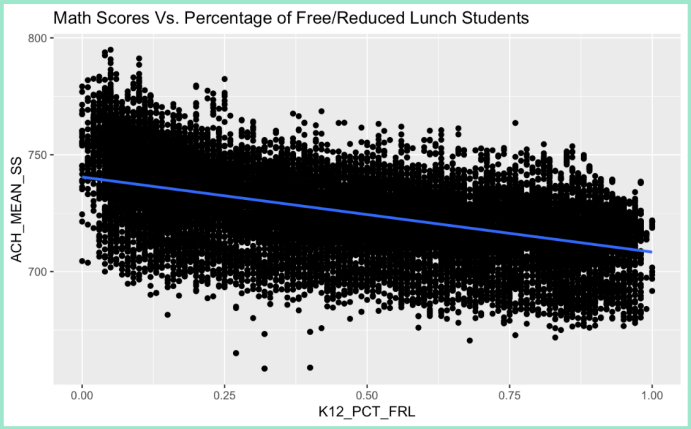

Key Points
Are there differences in student groups?
YES
Are there differences by subject?
YES
Can the scores be predicted?
SOMEWHAT
Predicting
FRL
Final Thoughts

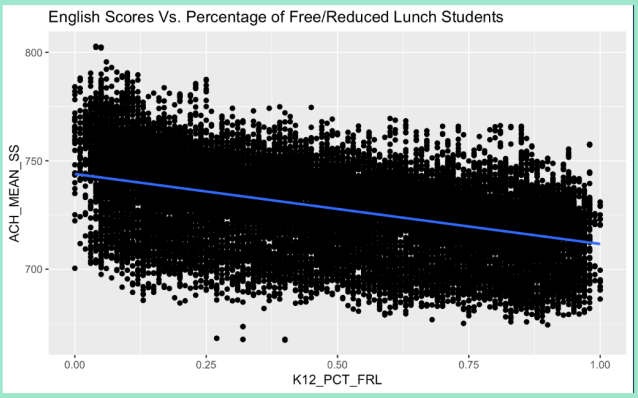

Science Actual vs. Predicted Values
Actual
Predicted
400
500
600
700
450
500
550
600
650
700


Math Actual vs. Predicted Values
Actual
Predicted
675
700
725
750
775
700
720
740
760


English Actual vs. Predicted Values
Actual
Predicted
700
750
800
700
720
740
760
780


Science Scores Vs. Percentage of Free/Reduced Lunch Students
ACH_MEAN_SS
700
600
500
400
0.00
0.25
0.50
0.75
1.00
K12_PCT_FRL


Math Scores Vs. Percentage of Free/Reduced Lunch Students
ACH_MEAN_SS
800
750
700
0.00
0.25
0.50
0.75
1.00
K12_PCT_FRL


English Scores Vs. Percentage of Free/Reduced Lunch Students
ACH_MEAN_SS
800
750
700
0.00
0.25
0.50
0.75
1.00
K12_PCT_FRL

How can we help these students in ways that go beyond what we do in the walls of the school building?