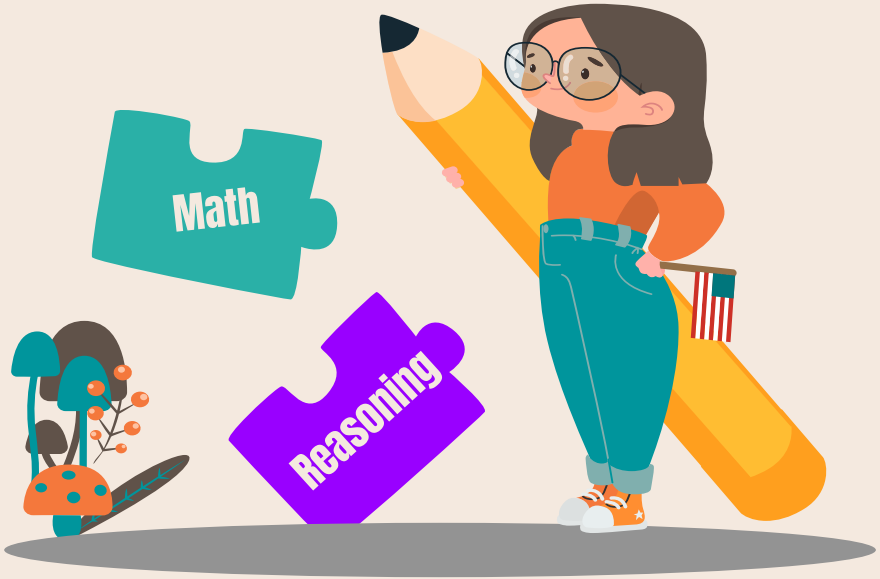
XYZ
Testing Services

Identifying Patterns and Drivers for Testing Performance Disparity



Team 2
Jhan-Syuan Lin | Niharika Agnihotri
Saloni Jadhav | Yash Bhirud
Aditya Tomar

Outline

BACKGROUND

- Business Problem
- Business Objective
- About the Datasets

MODELLING & ANALYSIS

- Regression Model
- Persistance Analysis
- Clustering Analysis

KEY TAKEAWAYS

- Insights
- Recommendations

Business Problem

XYZ
Testing Services

Help adults gain their high school equivalency diploma

- Is there a disparity in the performance of test takers based on regions?
- If so, can we identify potential reasons for this performance disparity?
- Is there any relationship between performance and the usage of XYZ resources?

Business Objective

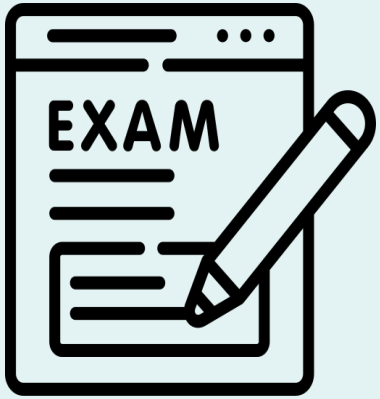
Identify patterns of regional level testing performance disparity in the XYZ test marketplace and its potential drivers.

Develop actionable steps that could proactively help achieve higher persistence rates with XYZ Testing Service.

About the Candidate Dataset

Data timeline	Earliest account creation: 1st January, 2020 Latest account creation: 29th September, 2022
Age	Mean age of all candidates: 26 years old Median age of all candidates: 22 years old
Enrollment Status	6% of total candidates are enrolled in XYZ services
Ethnicity	45% of the candidates are non-hispanic latino

The Dataset - Testing Data



Testing Reason

- 30% of total candidates are taking it for educational reasons
- 25% are taking it for their personal gain

Attempt Metrics

- 8% of the tests are taken online
- 40% of the candidates clear their subject test on the first attempt
- 40% of the tests are given on Fridays and Saturdays

Scope of Analysis for all our models

Geographic

Jurisdictions of USA

Candidates

291k

Only those who took the actual test

Subjects

- Social Studies
- Maths
- Reasoning
- Science

Score

Between the range of 100-200

XYZ Resources

Included all available XYZ resources

Linear Regression

- Determine if there is a relationship between the predictors and -
  - time taken to get credentialed
  - candidate score

- Coefficients: describes the associate impact of each predictor holding the others constant

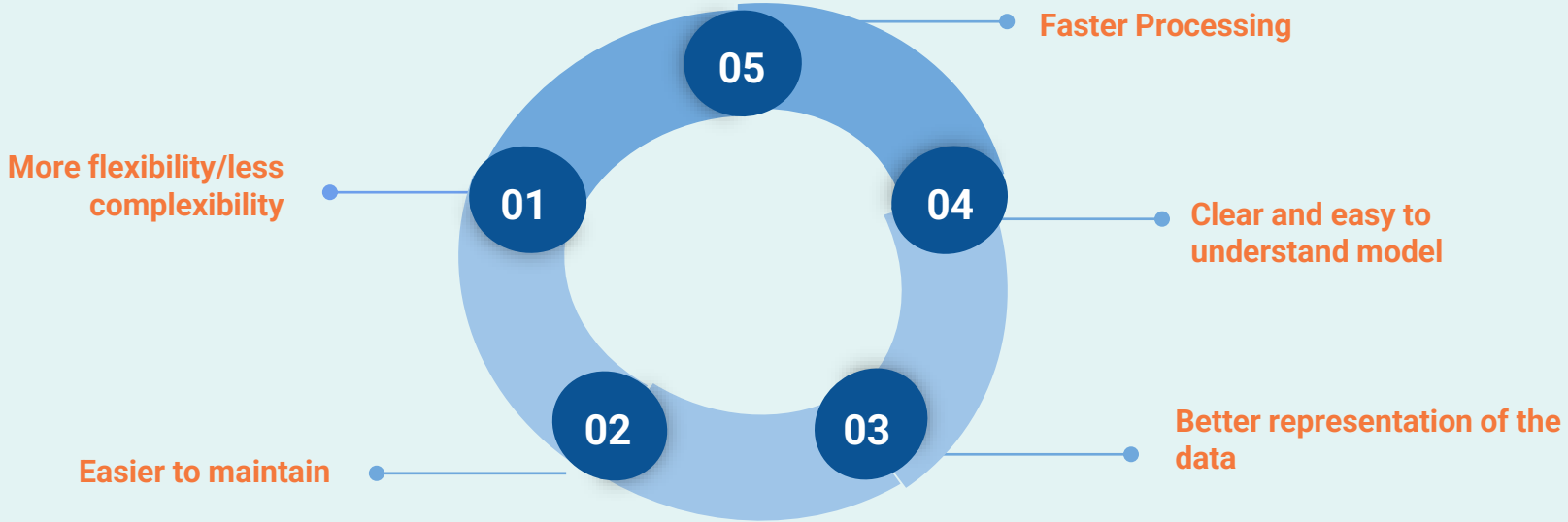
  - For every additional unit in the given predictor:
    - Positive: causes an increase in performance
    - Negative: causes a decrease in the performance

  - The larger the coefficient, the bigger the impact

Feature Engineering for our Regression Model

Process of manipulating and mutating our data to improve model training and overall performance.

So what are its various advantages?



New features generated for our Regression Model

Credential Time

Number of days taken for a candidate to get fully credentialed on passing all 4 test

Exam Show Up

Binary variable highlighting if the candidate showed up for the test

Geo Zip code Distance

Distance between the test center and the candidate's location

Test Day Difference

Gap in days between booking the test and taking it

Practice Test Evaluators

Evaluated each candidates count of practice tests and average practice score per subject

Running a Regression Model on Credentialed Candidates

Why we do this?

To get a better understanding of what resources helped candidates get credentialed

How we do this?

Taking the time taken to get credentialed as the target variable and XYZ resources as the independent variables

What can we try to infer from this?

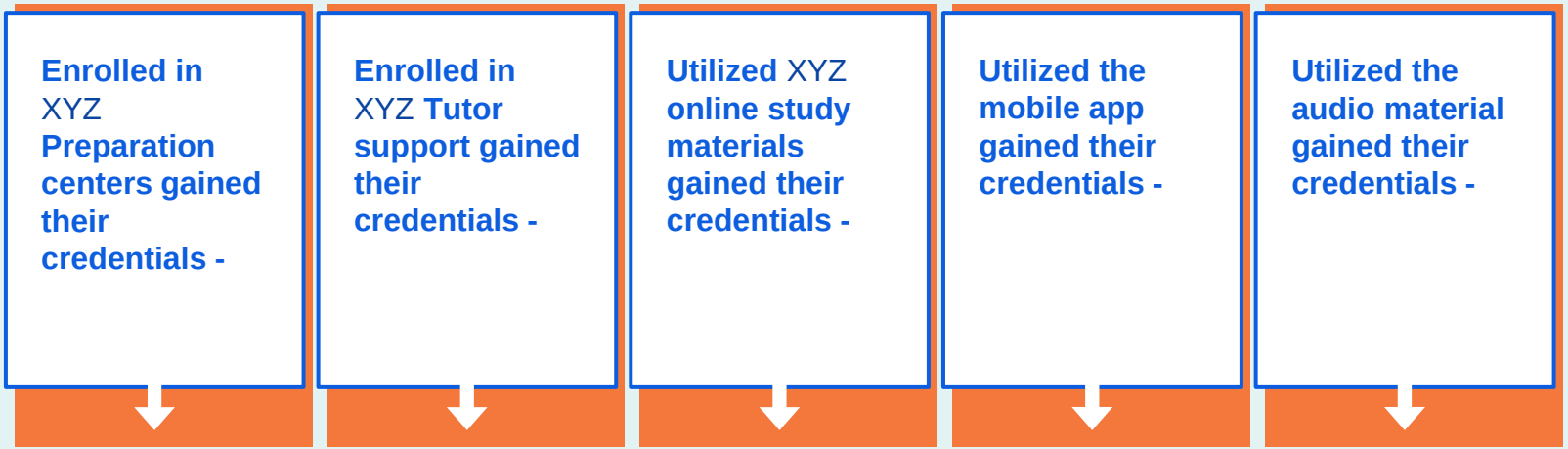
We can find out which resources help the candidate get credentialed quicker and to what degree

Independent Variables Taken

- All XYZ Help Resources
- Test Day Difference
- Practice Tests Mean Score
- Distance to Test Center
- Study-Test Center Location

Insights from our Credentialed Candidates Model

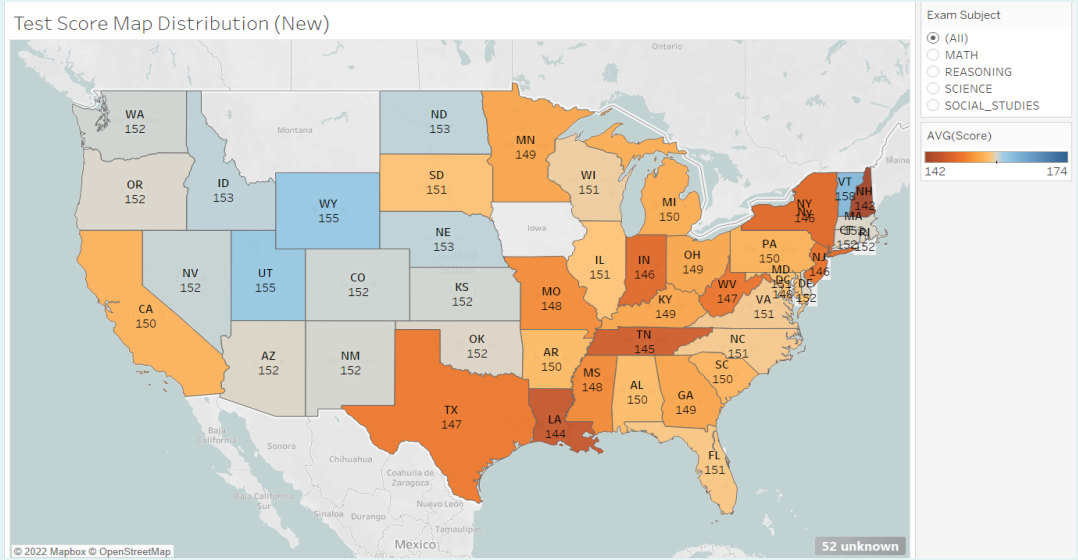
Average number of days to get credentialed: **60 days**
On an average candidates who:



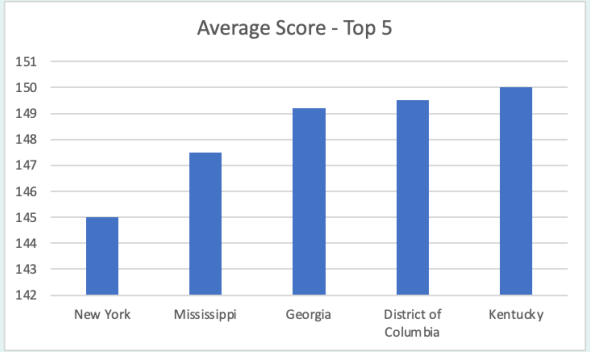
Running a Regression Model on a Jurisdiction Level

- To inspect the disparity between the jurisdiction, we decide to group the jurisdictions into a top and bottom 5 based on their average scores (coefficients)
- To understand how the scores are varying the most due to resources, we take the best vs the worst states

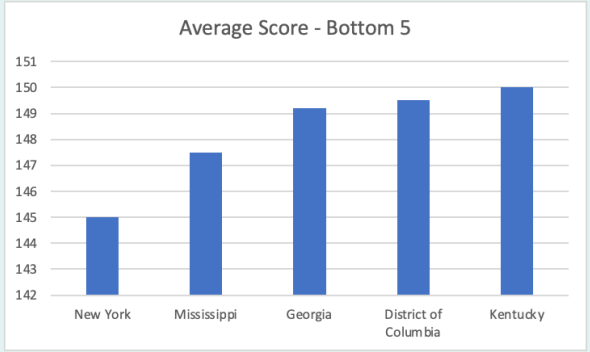
Test score distribution →

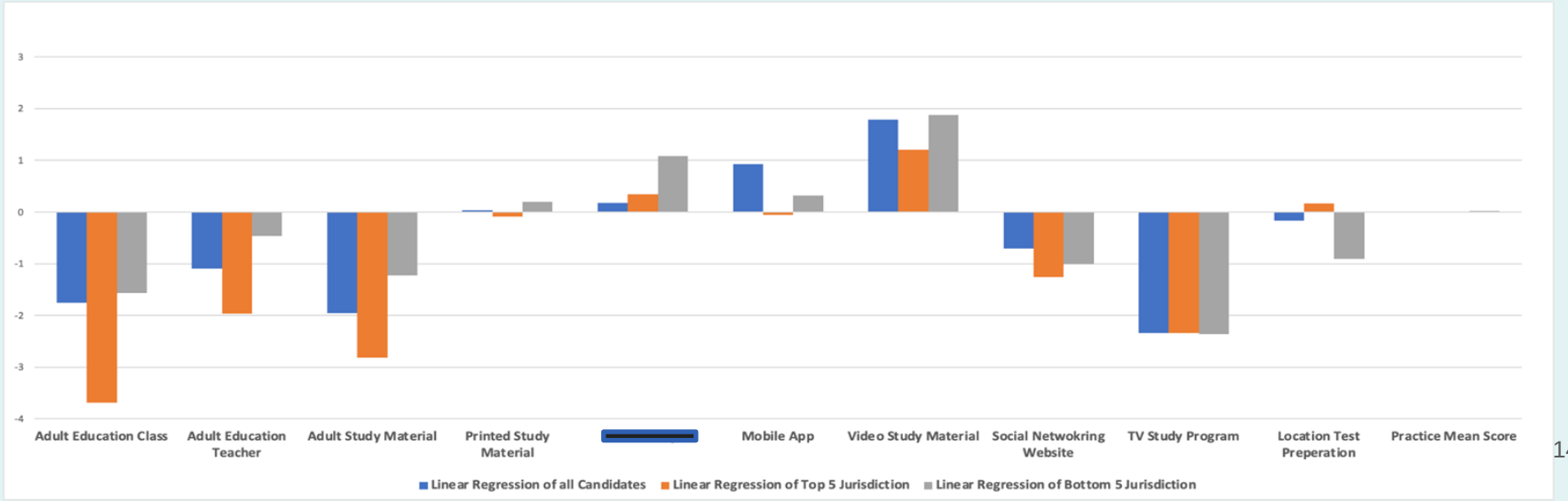


Top 5 vs Bottom 5: Average Scores and Coefficients



.....





Jurisdiction level Observations - Top 5

- *What resources cause a positive impact in good scoring jurisdictions?*
  - Online Course Video Study Material: increases the score by nearly 2 points
  - Taking more practice tests: increases score by a point

- *What resources cause a negative impact in good scoring jurisdictions?*
  - Enrolling for the adult education class: brings down the score by an average of 4 points
  - TV Study Program: causes a negative impact and reduces score by 4 points

Jurisdiction level Observations- Bottom 5

- *What resources cause a "positive impact in low scoring jurisdictions"?*
  - Enrolling in the Online Video Study Material: causes an increase in 2 points on average
  - Using the audio materials: increases the score by 2 points on average

- *What resources cause a "negative impact in low scoring jurisdictions"?*
  - The adult study material: bringing the scores down by 3 points
  - Using the TV study program: reduces score by 2 points on average

Analysing Dropout Estimate

Why we do this?

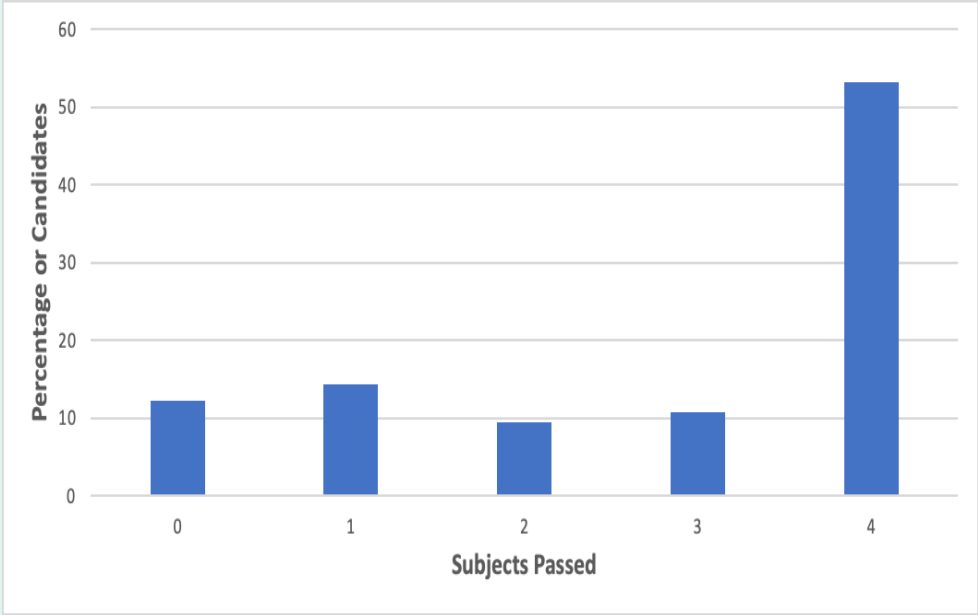
To understand where and on which subject do candidates find it hard to clear the last subject

How we do this?

By checking what proportion of candidates have their last subject remaining and mapping it on a state level

What can we try to infer from this?

We can find out which factors are contributing to this drop in persistence which are causing candidates to drop out



Insights from our Dropout Analysis

Last Subject Struggle

Most Candidates find it hard clearing the last subject - which causes the drop in persistence

Majority Struggle in clearing Maths

States such as Michigan and Indiana have the most candidates who have maths left as their last subject

Provide Recommendations directed towards Maths

We can tailor needs specific to those states by offering help in material related to Maths after running a regression model on it

Clustering

Why we do this?

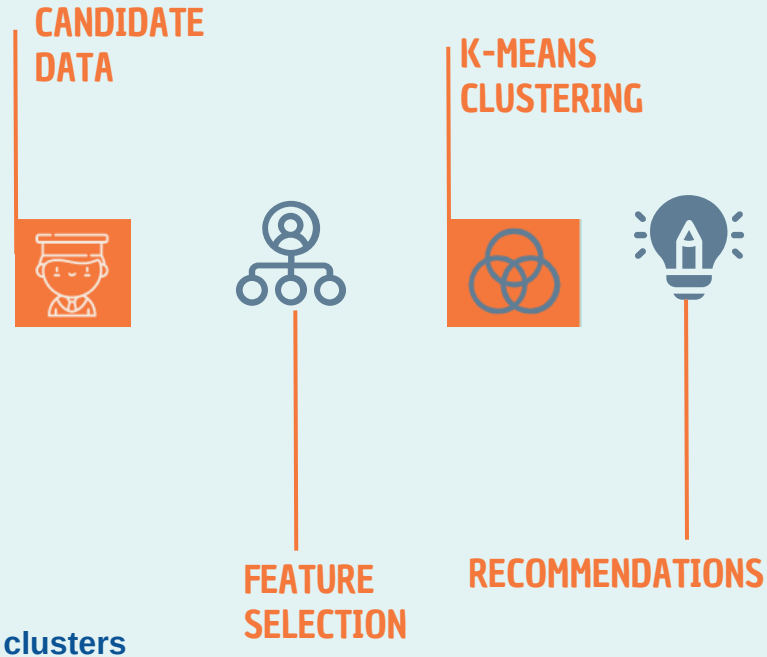
To identify **similar cohorts of data** and study their characteristics

How we do this?

This will be done at a **candidate level** to identify similar sets of candidates in our data

What can we try to infer from this?

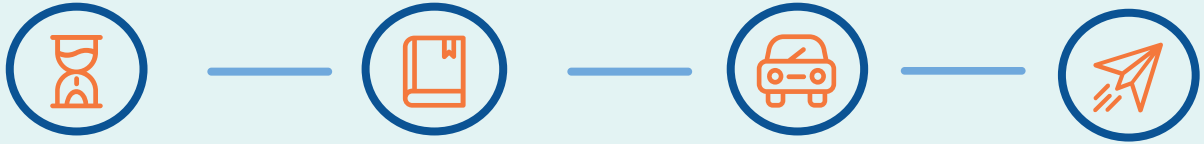
We can identify the **distribution of candidates across the clusters on a jurisdiction level** and perform profiling - giving state wise recommendations



Feature Selection for Clustering

XYZ Resource Usage	Usage of resources like adult education class, audio study material and so on.
Testing Metrics	Test metrics like number of online tests attempted, credential earned and so on
Ethnicity	Ethnicity like whites, asians, african american, etc

Overall Cluster Metrics for Candidates



Total Candidates	Average Score	Exam Show Up Rate	Average Test Centre Difference
289k	142	97%	371 kms

Cluster Descriptions-1

Cluster 1

Candidates : 93k
Avg Score : 152
Show up rate : 99%

- Highest proportion of online test takers (possibly because of above point)
- Frequent users of every XYZ study material
- High proportion of whites

Cluster 3

Candidates : 57k
Avg Score : 150
Show up rate : 99%

- Have a high proportion of online test takers
- Frequent users of every XYZ study material
- Consist of African_Americans, Indians, etc

Cluster Descriptions-2

Cluster 2

Candidates : 62k
Avg Score : 132
Show up rate : 95%

- Low number of tests given per candidate
- Lowest users of every XYZ study material
- Have asians, hawaiians, etc

Cluster 4

Candidates : 45k
Avg Score : 132
Show up rate : 93%

- Least number of tests given per candidate
- Lowest users of every XYZ study material
- Mostly consists of whites

Cluster Descriptions-3

Cluster 5

Candidates : 30k
Avg Score : 128
Show up rate : 94%

- Use moderate amount of study materials (3rd rank among resource usage)
- Lowest proportion of online test takers
- Have a high proportion of African_Americans

Key Takeaways - Cluster 4

- **Cluster 4** has lower performance possibly due to low show up rate i.e. the proportion of candidates cancelling or postponing tests is high.

- Jurisdictions like **Kentucky, Wisconsin** has a fairly significant proportion of students from Cluster 4, hence that has led to a drop in the overall score of those district.

Key Takeaways - Cluster 1 & Cluster 3

- Candidates from **Cluster 1** and **Cluster 3** have high scores. These students have used XYZ resources like adult education class, adult education teacher, audio study materials, etc more than students from the other clusters and hence the usage of these resources should be promoted.

- Jurisdictions of **Wyoming, Utah** etc have a significant proportion of students from these clusters, hence they have higher average scores

Key Takeaways - Cluster 5

- Attempting tests online seem to have a better performance instead of offline tests
- **Cluster 5** has similar resource usage like **Cluster 1** and **Cluster 3,** however, their average scores are the lowest and primarily this could be due to the lower proportion of candidates giving tests online.
- Hence, it is necessary to identify the reasons for better performance on online tests
- New York has 29% candidates from Cluster 5, hence this is one reason for a drop in their average score

Recommendations

To improve candidate credential rate

- Run campaigns targeted towards the Expert Tutor resource - reduces time to get credentialed by 9 days
- Encourage usage of mobile apps - reduces time to get credentialed by 4 days

To improve scores in bottom 5 states such as New York & Kentucky

- Encourage candidates to enroll in online study material - increases score by 2.3 points on average

To maintain persistence

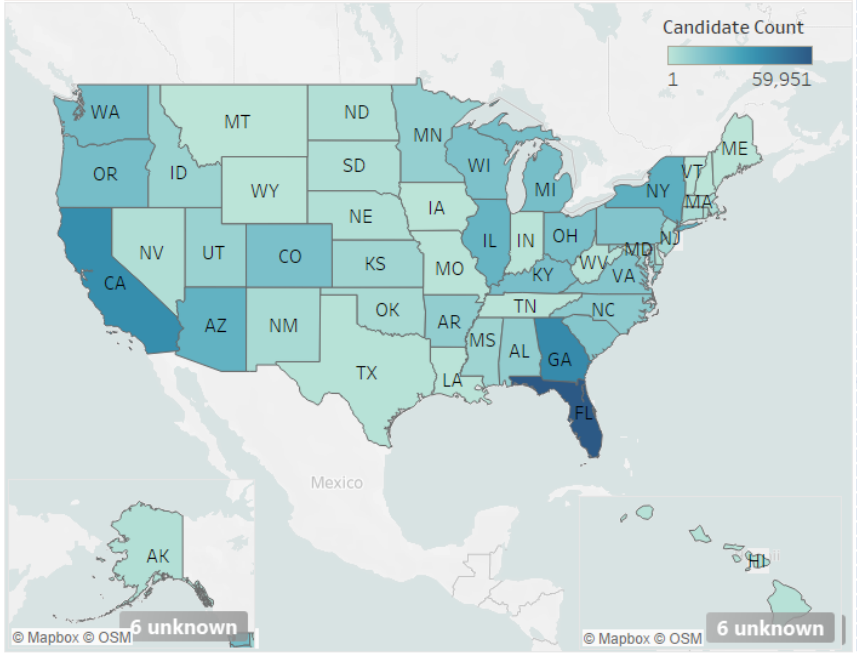
- Encourage candidates in Michigan to use online study course material resource as this boosts Maths score by 1.8 points on average - which could help them improve their score

Tableau Dashboard Example

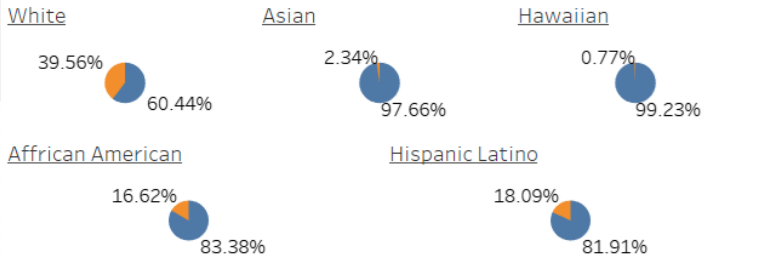
Candidate Count	Credential Percentage
503,619	31.51%

Candidate Map Distribution

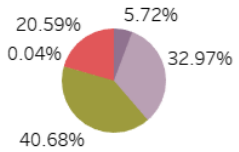
Click on the state to see the detail.



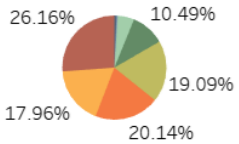
Ethnicity Percentage



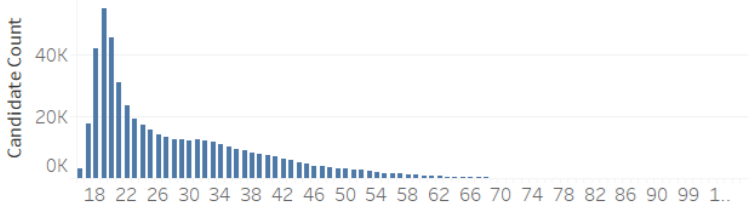
Candidate by Gender

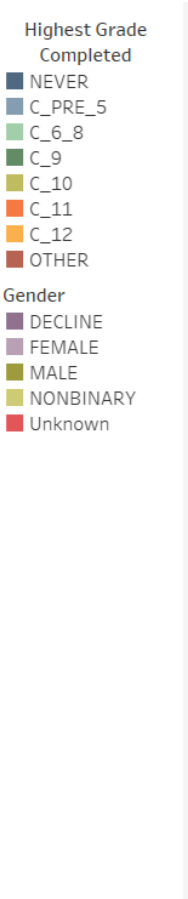


Candidate by Highest Grade



Age Distribution





THANK YOU :)