CS 506 Deliverable 1

MA Comp Sci. Representation - Team A

Problem Statement

In an era marked by rapid technological advancements and digital integration, computer science education stands as a critical pillar for economic opportunity and social equity. The goal is to identify the gaps and barriers faced by these communities in the realm of computer science education and AP test-taking.

The insights from this analysis will be crucial for the NCF to determine where to channel investments and which strategic partnerships with non-profits or corporations could be cultivated to promote racial equity and social justice in education.

The endeavor will culminate in a proposal that not only sheds light on the current state of educational disparity but also recommends actionable steps for the NCF to effect tangible change, ensuring that computer science education serves as a tool for empowerment, rather than a gatekeeper of opportunity.

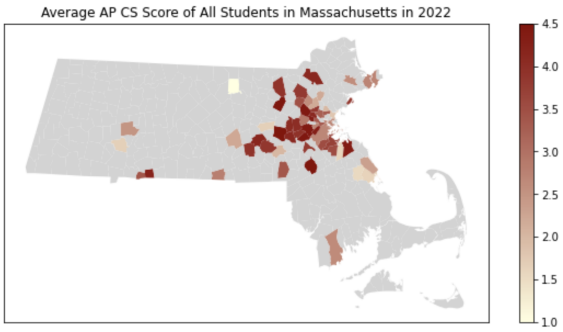
Data Collection and Cleaning

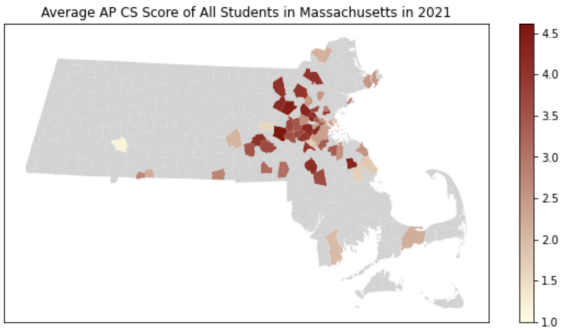
- **Data Retrieval**: Downloaded the datasets for the years 2021 and 2022, focusing on different student groups at the district level from the DESE website.
- **Initial Inspection**: Conducted an initial review of both datasets to understand the structure, content, and any apparent inconsistencies or missing data.
- **Combining Datasets**: Merged the 2021 and 2022 data files into a single dataset, ensuring that similar categories were aligned and that data types were consistent across both years.
- **Categorical Data Consistency**: Ensured that categorical data (e.g., race, ethnicity, gender) were consistent across the dataset, with no variations in spelling or categorization.
- **Verification Against Source**: Verified a random sample of the cleaned data against the original source to check for accuracy in the cleaning process.

Exploratory Data Analysis

AP CS Performance

- **HeatMap: Average AP CS Scores of All Students in Massachusetts**

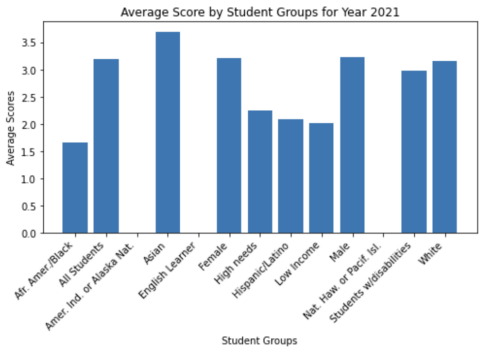


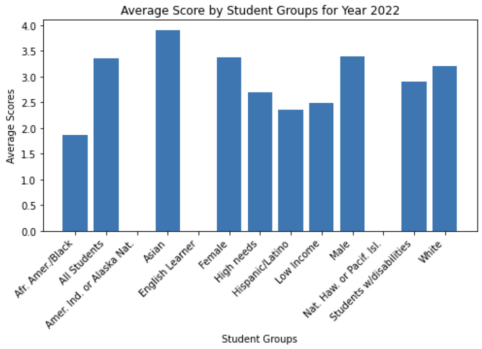


  - When analyzing the average AP CS scores of all students in Massachusetts, we observe that the available data is primarily concentrated in Greater Boston. When examining the color scale, it becomes evident that students in these regions achieve notably high average CS scores, with most of them being represented by darker shades of red. This implies that a significant majority of students in these areas attain scores of at least 4.0.

  - Furthermore, even though there is missing data, we can still observe a clear trend: areas further from Boston exhibit relatively lower scores, often around 2.0 or 2.50. From this analysis, it can be inferred that regions outside of Boston, apart from the city itself, may not prioritize computer science education to the same extent.

- **Histogram: Average AP CS Scores of Different Student Groups**

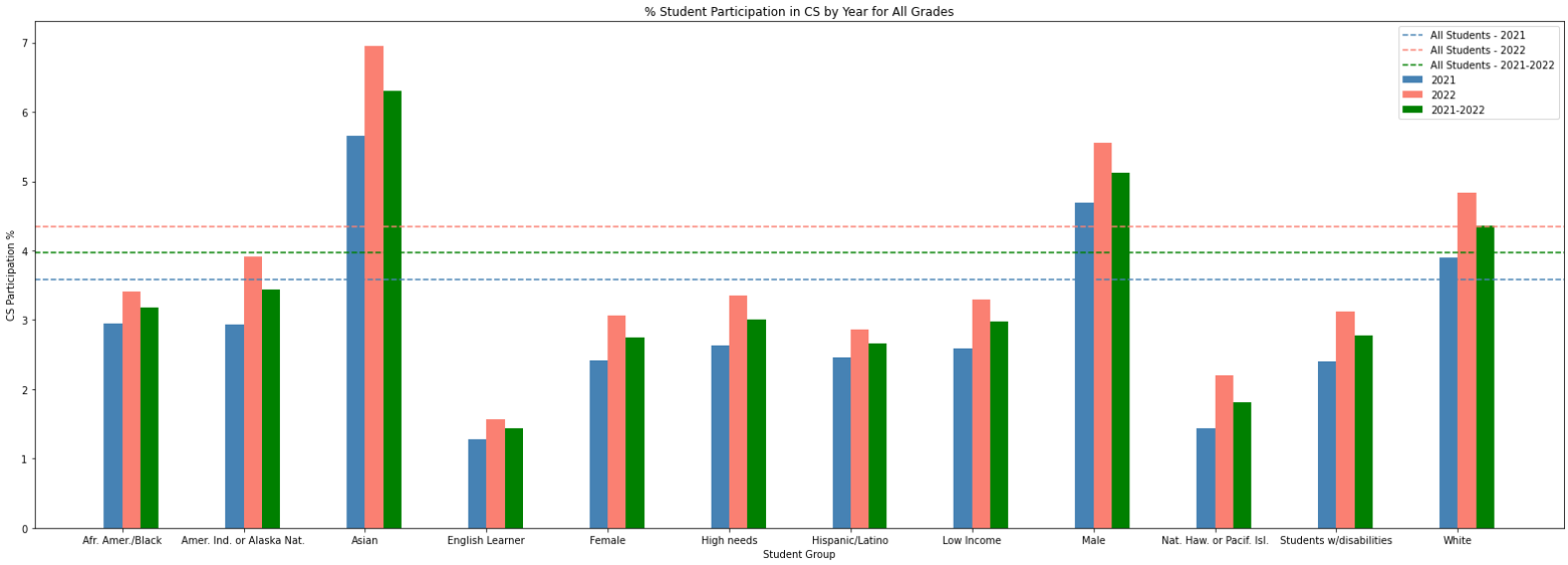




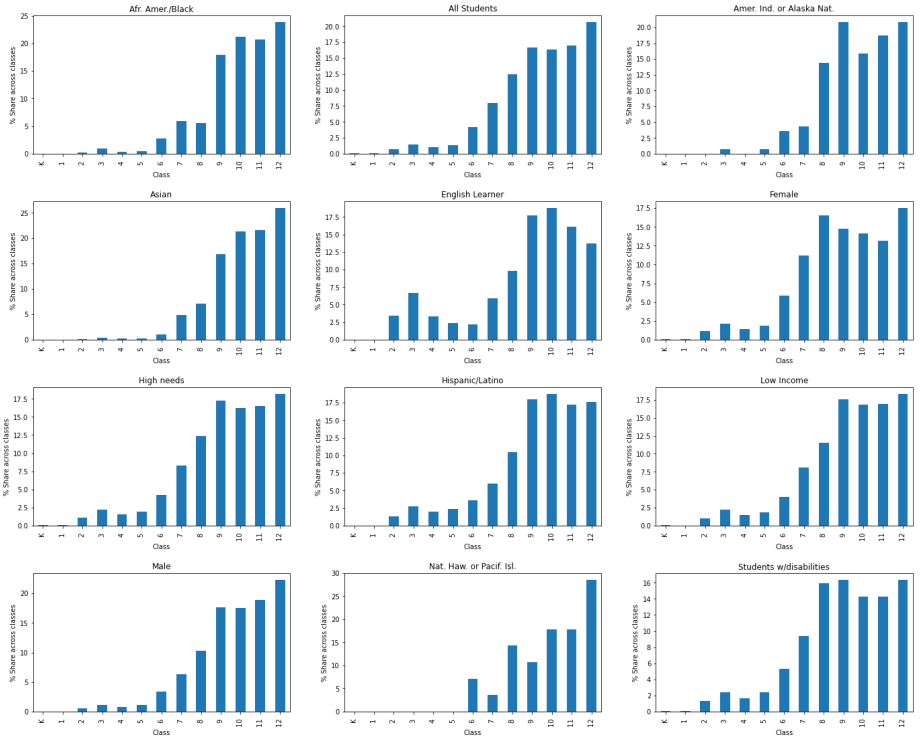
  - From the graphs depicting scores across various student groups, it is evident that the scores for female and male students are generally comparable, with minimal differences. However, when considering racial demographics, it becomes apparent that African American students tend to have the lowest scores, while Asian students achieve the highest scores in AP Computer Science.
  - It's essential to acknowledge that this disparity may be influenced by missing data in the dataset. The dataset contains much more scores for Asian and white students, while data for African American and Hispanic Latino students is notably lacking. As a result, the overall average score for all students may appear inflated, primarily because the scores of Asian students have a more significant impact on the aggregate due to their higher representation in the dataset.

CS Participation

1. CS Participation % by Student Group and Year:
   a. Asian and Male student groups above average
   b. Increase in CS participation for all student groups
   c. English Learners having the least participation followed by Nat.Haw or Pacf.Isl



2. Distribution in CS Participation across grades for all Student Groups:
   a. For Asian and Male Student groups, the share of classes keeps increasing with the class
   b. English Learners have the opposite trend
   c. Nat.Haw or Pacf.Isl seem to start the participation late

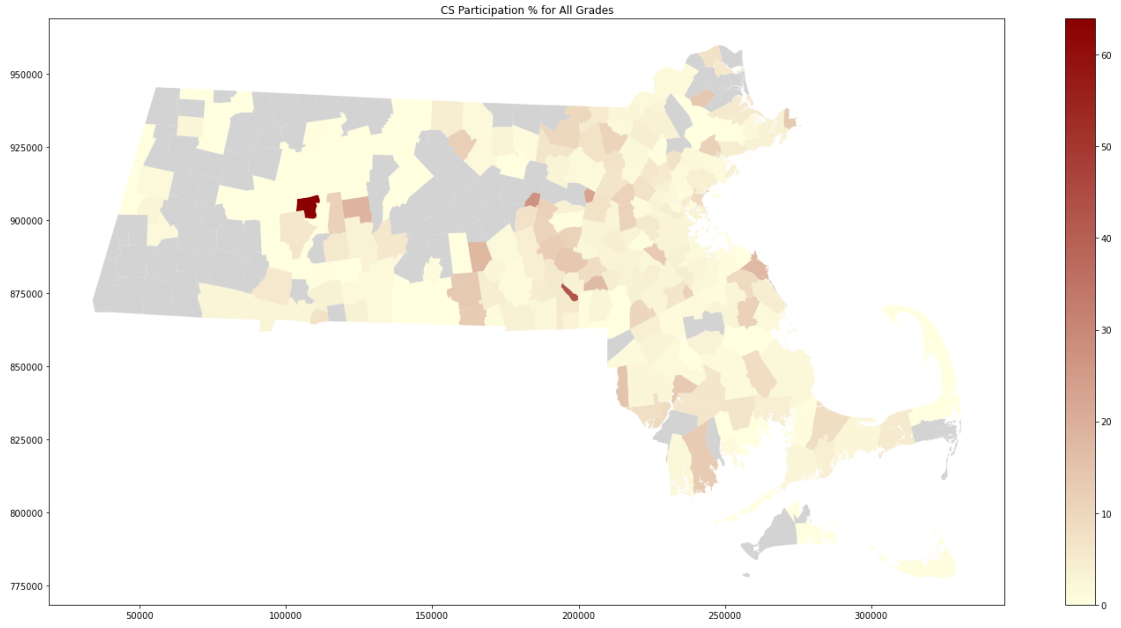


3. Geographical Distribution of CS Participation:
   a. Few towns with much better performance
   b. YOY difference between the top towns is huge

Top 5 Towns in 2021 and 2022

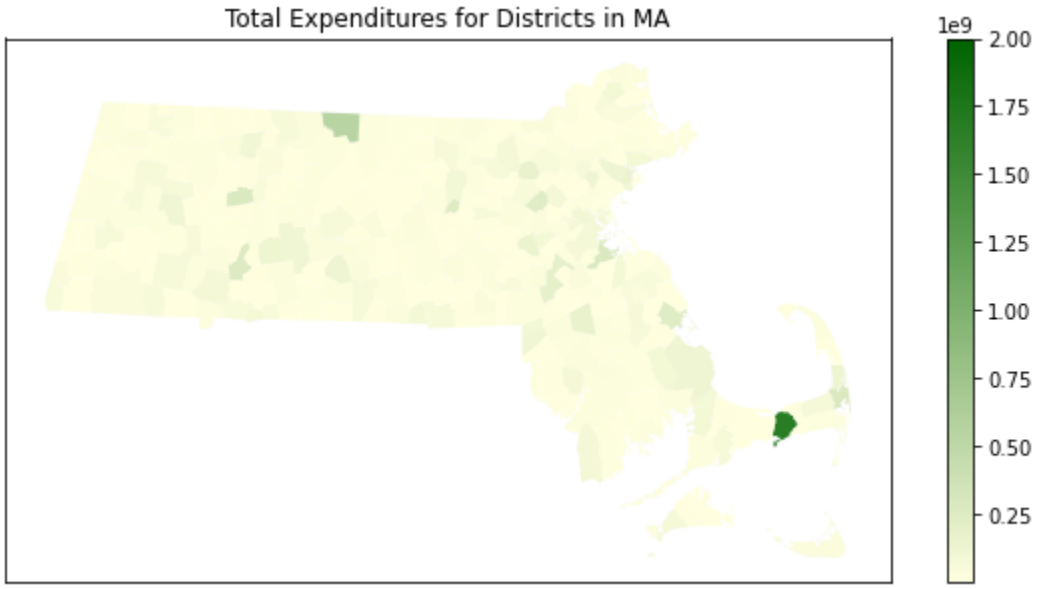
Town	CS Participation %
HATFIELD	59.366755
MAYNARD	31.499556
HOPEDALE	31.034483
ROCKPORT	19.736842
SCITUATE	18.501420

Town	CS Participation %
HATFIELD	68.838527
HOPEDALE	54.813360
CLINTON	39.045093
PELHAM	27.446809
GARDNER	21.337580

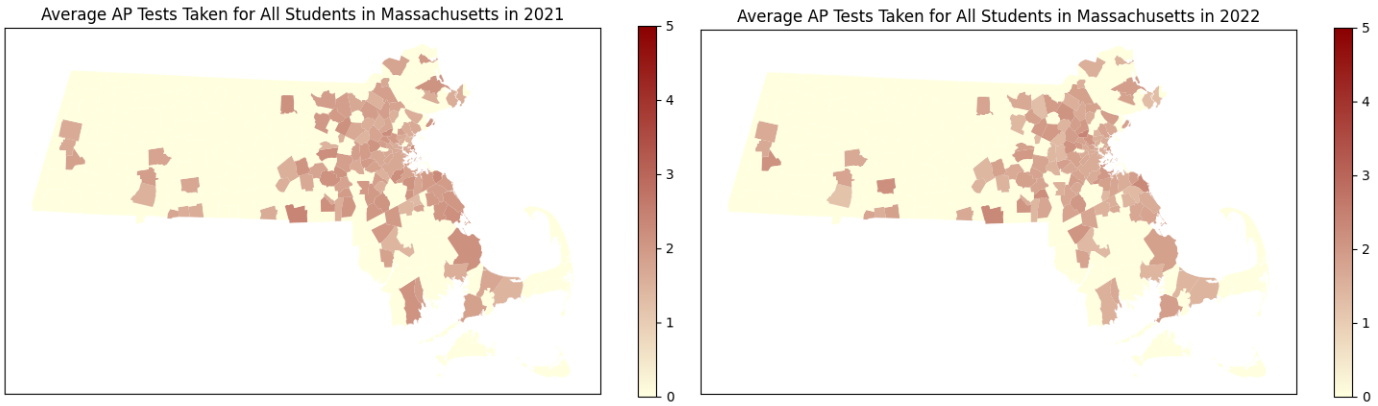


4. Expenditures for Districts

This shows the total expenditures for the districts in MA. From the data, we can see that the districts all spend over millions of dollars on their schools. The districts also look like they spend the same amount of money, except for the one district in the south east.



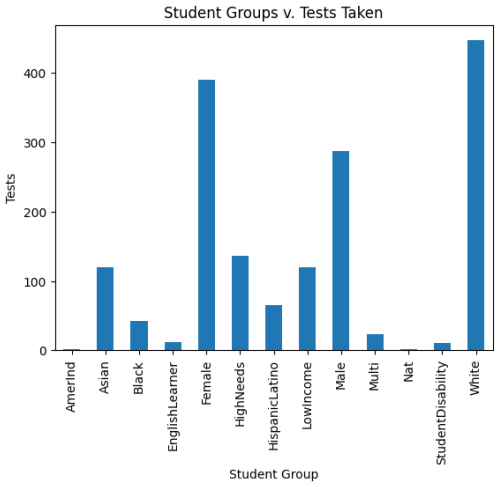
AP Participation (# of Tests Taken)

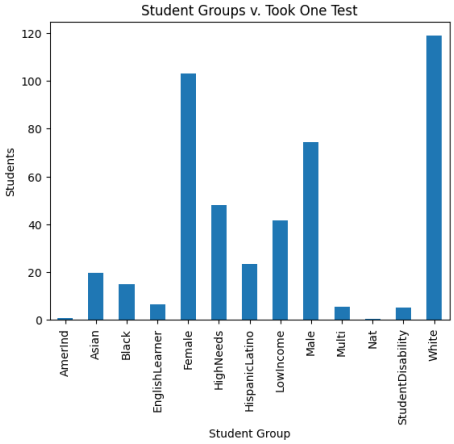


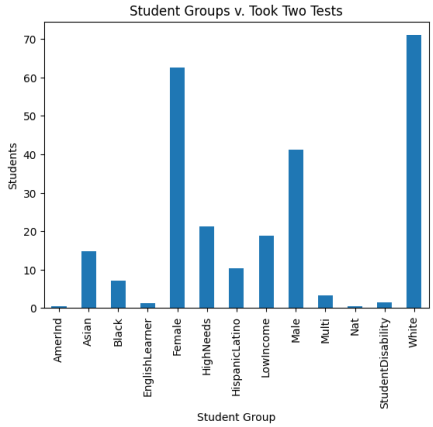
When examining the heat density maps, we can see that much like the results above, the most tests tend to be taken in the Greater Boston area. From the graph we can see that anywhere outside of the Greater Boston Area has no tests, or is missing data. However, it seems that there seems to be little deviation in the number of tests taken for the areas that are filled in, as there is not any stark contrast between school districts for the most part. This information applies to both years 2021 and 2022.

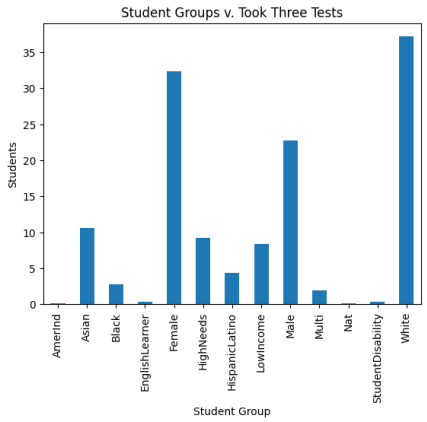
Bar Plots and Analysis

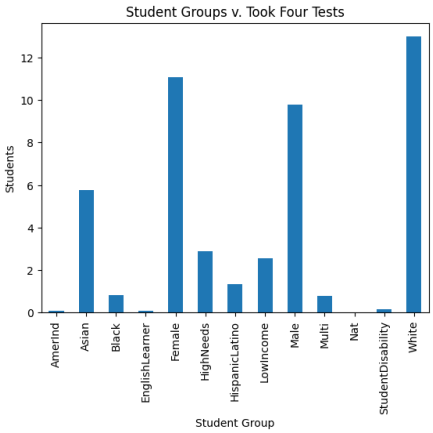
As we can see in this chart that compares the student groups against the total AP tests taken, we can observe stark contrasts in the number of tests taken. The overwhelming majority seems to be white students, while other races are much less. This overall trend can be observed in the student groups vs amount of tests taken as well. These results demonstrate a high deviation between groups.

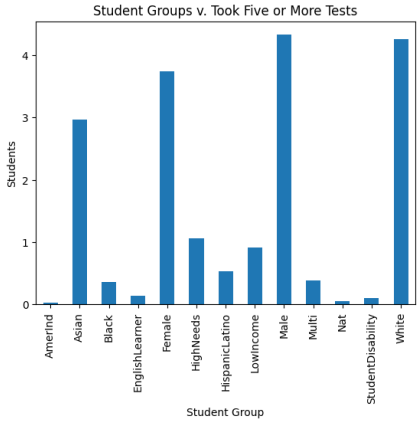



Student Groups v. Took One Test
Students
Student Group
AmerInd
Asian
Black
EnglishLearner
Female
HighNeeds
HispanicLatino
LowIncome
Male
Multi
Nat
StudentDisability
White

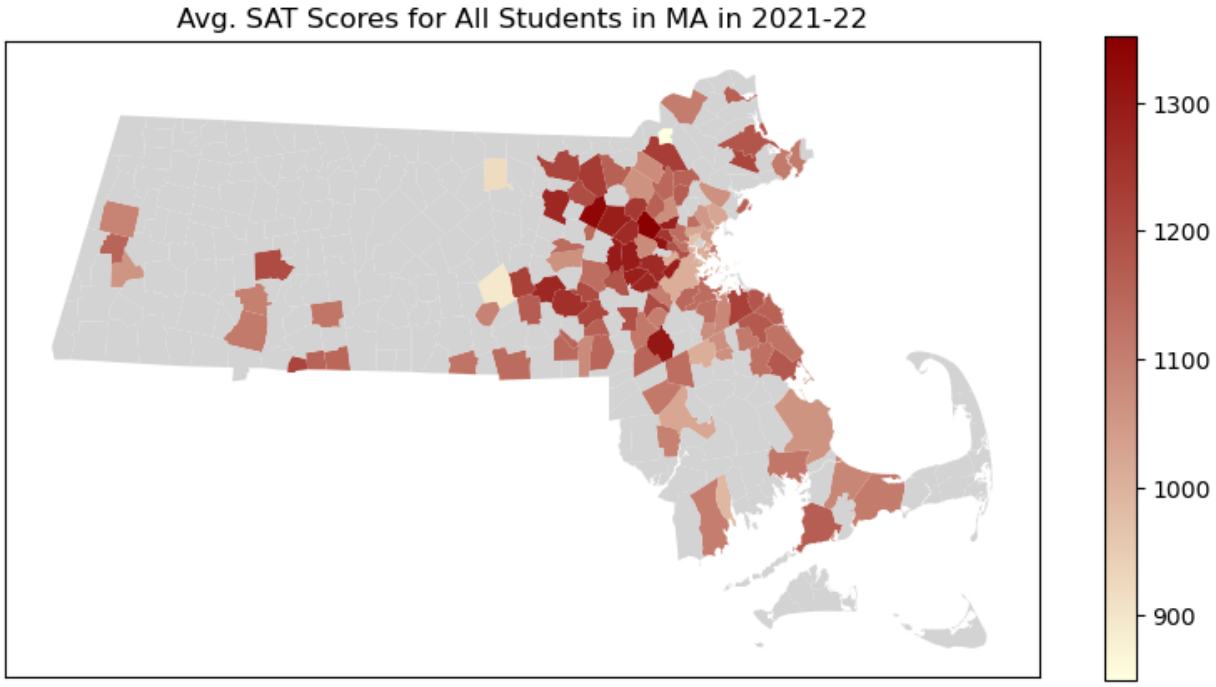

Student Groups v. Took Two Tests
Students
Student Group
AmerInd
Asian
Black
EnglishLearner
Female
HighNeeds
HispanicLatino
LowIncome
Male
Multi
Nat
StudentDisability
White


Student Groups v. Took Three Tests
Students
Student Group
AmerInd
Asian
Black
EnglishLearner
Female
HighNeeds
HispanicLatino
LowIncome
Male
Multi
Nat
StudentDisability
White


Student Groups v. Took Four Tests
Students
Student Group
AmerInd
Asian
Black
EnglishLearner
Female
HighNeeds
HispanicLatino
LowIncome
Male
Multi
Nat
StudentDisability
White


Student Groups v. Took Five or More Tests
Students
Student Group
AmerInd
Asian
Black
EnglishLearner
Female
HighNeeds
HispanicLatino
LowIncome
Male
Multi
Nat
StudentDisability
White

SAT Performance (across MA)



Using the data retrieved from the MA Comp Sci. Representation database, extracted & cleaned by our team, we take a deeper dive analyzing SAT Scores for all Students in MA for the year of 2021-22. We aren't exploring both the years separately; both the plots are visually identical showing little to no change across years. While we had three distinct features available (the Reading Score, Math Score & the number of tests taken), per the exploratory data analysis, it seems that the total SAT score may be the best representative feature to create a heat/density map. The number of tests taken showing a small delta is indicative of a lack of big differences between district & may just be representative of the sheer number of students, the districts having a larger population, not of the performance of their students. We have removed outliers, data for districts with too few students taking tests & replaced the missing data in the map with the 'light gray' color for better visualization.

Individual Contributions

1. **Sai Tejaswini Junnuri**: Individually worked on extracting the CS Participation dataset and analyzing it. Analyzed the difference in CS participation across student groups, across grades and geography.
2. **Changxuan Fan**: Individually worked on extracting the CS Performance dataset and analyzing it. Analyzed the difference in CS participation across student groups, across grades and geography.
3. **Kelvin Lin**: Individually worked with the AP participation dataset. Worked to graph and visualize the differences in numbers of tests taken for AP students. Analyzed the differences in geography in total AP tests taken in Massachusetts.
4. **Pratham Shroff**: Individually extracted & cleaned the SAT Performance dataset, conducted exploratory data analysis to choose the best features, merged datasets, compared them year over year, created meaningful visualizations (such as that of the Heat Density Map) that the client can use to further their key objectives along with detailing the why & how behind these steps.
5. **Kenise Neal**: I extracted and cleaned the data for the finances of each district. Used the data to create a heat density map to see the distribution of funding in MA. By doing this, it allows us to see which districts receive the most resources and in correlation the most help.