Master’s Capstone

Understanding Public Schools System in MA

Yuanyuan Yang 12/18/2023

Motivations

K-12 education is the most basic and significant education for kids. I hope that large/big data could give us a wide as well as detailed picture about our public school districts. A better, evidence-based understanding of what are the key factors and how much these factors affect student academic outcomes could help our policy-makers effectively deploy fundings and benefit shaping our public thinking, so that it would improve the K-12 education system to build a more effective, supportive and equitable future.

The target audience and users

- Education institutions and administrators of state, city/town and school district.
- Educators and caregivers.

Outline

- Data (source, assumption, preprocessing)
- Visualization
- Statistic Analysis
- Machine Learning Model
- Conclusion & Discussion

Data - Source

- Massachusetts Department of Elementary and Secondary Education (DESE)

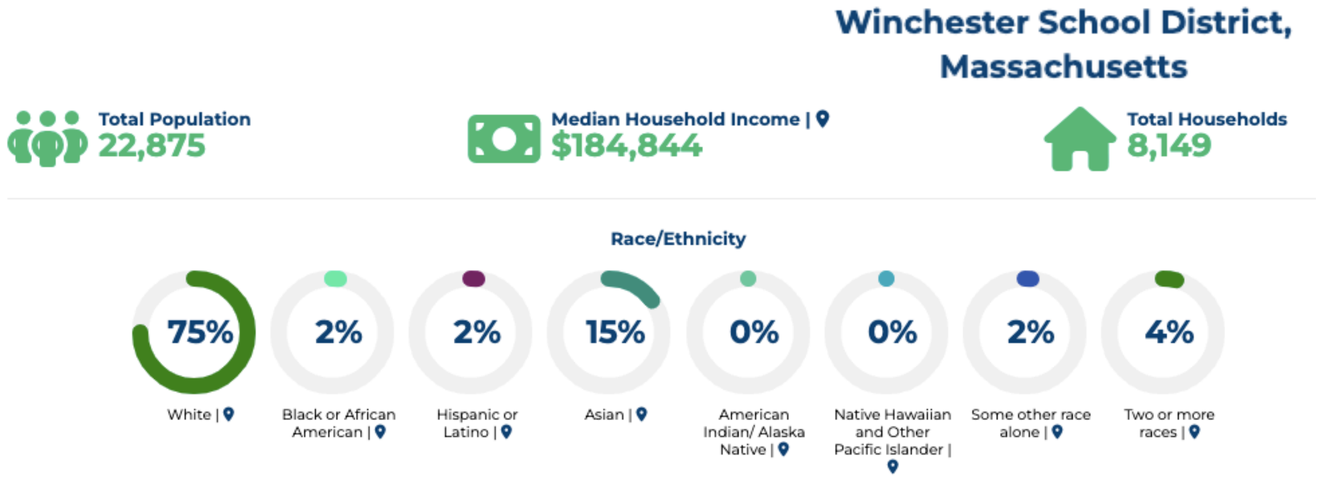
dese MASSACHUSETTS Department of Elementary and Secondary Education

School and District Profiles

School Name	School Code	# in Cohort	% Graduated	% Still in School	% Non-Grad Completers	% H.S. Equiv.	% Dropped Out
Abby Kelley Foster Charter Public (District) - Abby Kelley Foster Charter Public School	04450105	94	96.8	0.0	0.0	0.0	3.2
Abington - Abington High	00010505	156	93.6	1.9	0.0	0.0	4.5
Academy Of the Pacific Rim Charter Public (District) - Academy Of the Pacific Rim Charter Public School	04120530	56	98.2	0.0	0.0	0.0	1.8
Acton-Boxborough - Acton-Boxborough Regional High	06000505	452	98.0	0.9	0.0	0.2	0.9

- National Center for Education Statistics (NCES)

IES : NCES National Center for Education Statistics



Data - Assumption

- **Data categories**:
  - Situational factors: features that describes the school's basic information.
  - Controllable factors: features that school authorities have some control over.
  - Outcome: metrics that describe student academic performance.
- **School transition over time**:
  - School closes, new school open, change code, etc.
  - Use 2023 public school and district as target list.
- **Variable minimum threshold**:
  - Certain variables have limited sample size.
  - Schools are excluded if certain variable contains less than 10 samples.
- **Outliers**: Boston city could be treated as a outlier in most of analysis, because the student population is so complicate that the data deviates significantly from the overall pattern of the dataset.

Data - Processing Pipeline

Data Sources → Metadata Study → Sample Study → Data Selection → Web Scraping → Data Cleaning/ Organizing

dese MASSACHUSETTS Department of Elementary and Secondary Education

IES NCES National Center for Education Statistics

Profiles Help - About the Data
Public Schools & Districts
General Students Teachers Finances Assessment

Massachusetts Public Schools Data
Data Card Code (8) Discussion (0) kaggle

MA_Public_Schools_2017.csv (1.61 MB)

scraping_assessment.py

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
@author: yuanyuan
"""
from bs4 import BeautifulSoup
import requests
import pandas as pd
import os
import time
import random
# import validators
```

School Name	School Code	School Type	Town	State	Zip	Grade	District Name	District Code	District Type
Abby Kelley Foster	4450105	Charter School	Worceste	MA	1606	K,01,02,(	Abby Kelley Foste	4450000	Charter District
Abington: Abington	10001	Public School	Abington	MA	2351	PK	Abington	10000	Public School District
Abington: Abington	10505	Public School	Abington	MA	2351	09,10,11,	Abington	10000	Public School District
Abington: Abington	10405	Public School	Abington	MA	2351	05,06,07,	Abington	10000	Public School District
Abington: Beaver B	10020	Public School	Abington	MA	2351	K,01,02	Abington	10000	Public School District
Abington: Woodsda	10015	Public School	Abington	MA	2351	03,04	Abington	10000	Public School District

Visualization

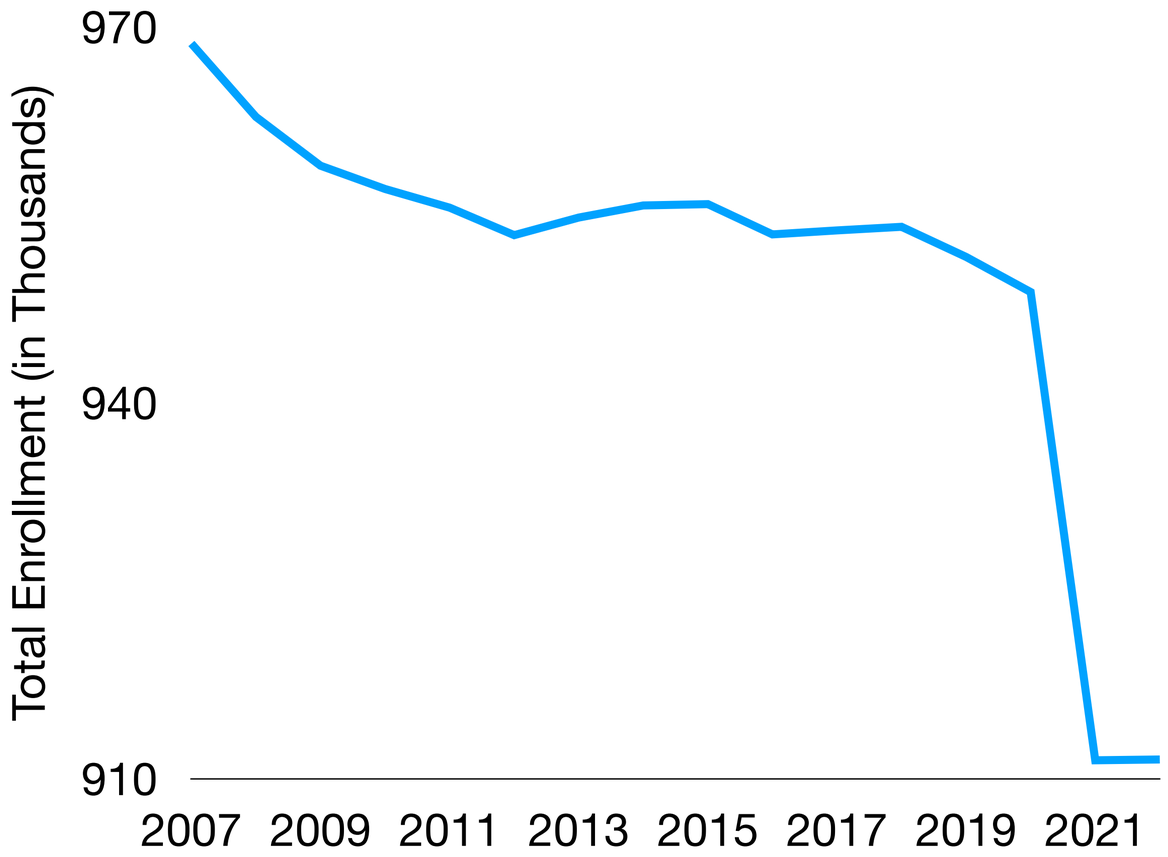
- Data visualization is the practice of designing and creating easy-to-communicate and easy-to-understand graphic and visual representations of a large amount of complex data with the help of static, dynamic or interactive visual items such as charts, graphs, and maps.
- In this project, data visualization covers from raw data mining to result visualization, and it includes following three parts:
  - Part 1: Temporal visualization
  - Part 2: Spatial visualization
  - Part 3: In-depth visualization for selected areas of interest

Part 1 - Temporal Visualization

- Visualize data from 2007 to 2022 (16 years of data), reveal interesting trends at both state and school/district levels.
- **State Level** explores trends in:
  - Enrollment: e.g., total, race, and gender, etc.
  - Socialeconomic status of students: e.g., economic disadvantaged percentage
  - Evaluation metric: e.g., high school graduation rate, SAT, etc.
- **School/District Level** selects 10 traditional public schools (top 5 and bottom 5 based on 2023 ranking), and explores the differences through socioeconomic and academic aspects.
  - *Top 5*: Lexington high, Weston High, Brookline High, Newton North High, Belmont High
  - *Bottom 5*: Southbridge High School, Chelsea High School, Monson High, Hoosac Valley High School, Athol High

Part 1 - Temporal Visualization (cont.)

State Level Enrollment by *Total Number*, 2007 - 2022

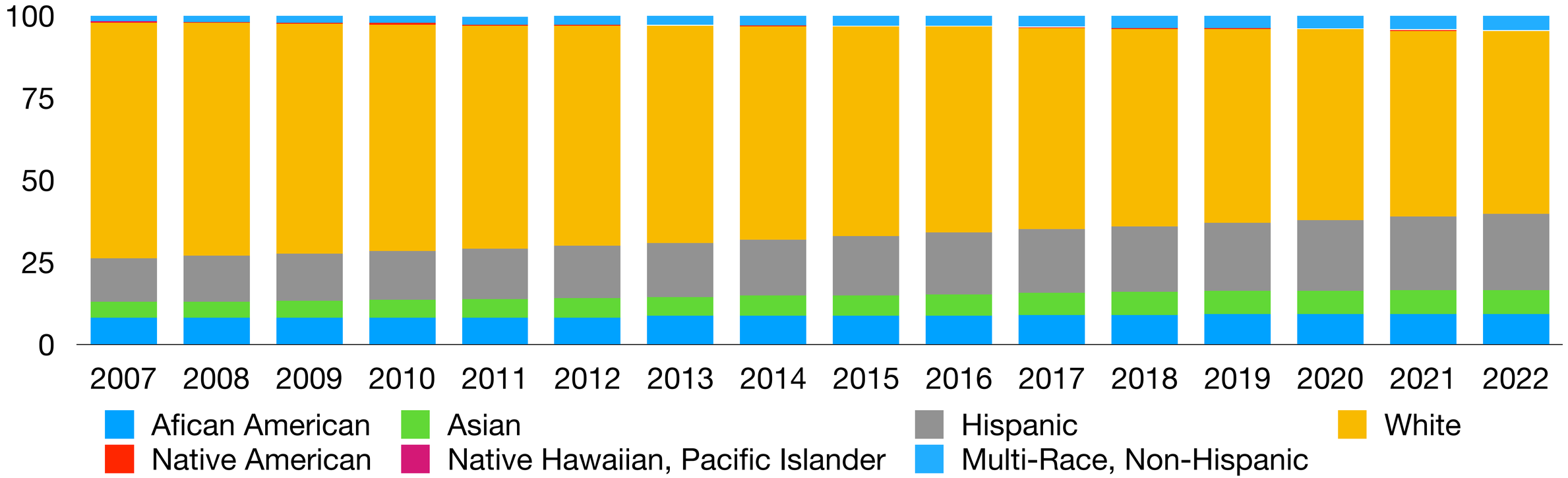


- 2007 - 2012, total enrollment decrease slightly.
- 2012 - 2019, enrollment stable
- 2020 - 2022, significant drop with a reduction of more than 40,000 students. One possible explanation is during COVID-19 pandemic, part of the students switched to private schools or home-schooling due to the inefficient online courses provided by many public schools.
- The level of reduction since 2020 in Massachusetts echos the trend observed in National data, for whose enrollment reduced more than 1 million.

Part 1 - Temporal Visualization (cont.)

State Level Enrollment by *Race*, 2007 - 2022

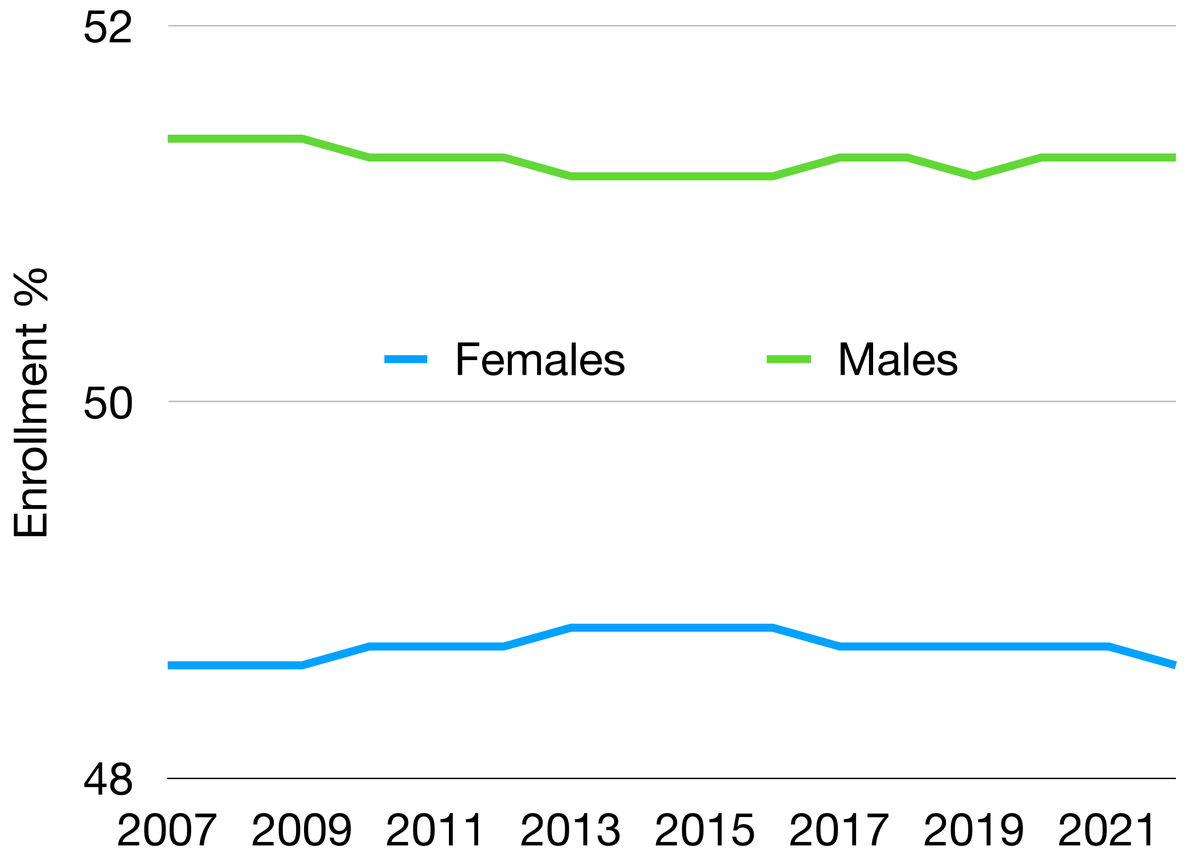
- From 2007 to 2022, public school students in MA became more diverse.
- One important change is the White contribution reduced while Hispanic increased, followed by Asian and African American. This could because of more students from other races enrolled into public school, which dilutes the percentage of the White.



Part 1 - Temporal Visualization (cont.)

State Level Enrollment by *Gender*, 2007 - 2022

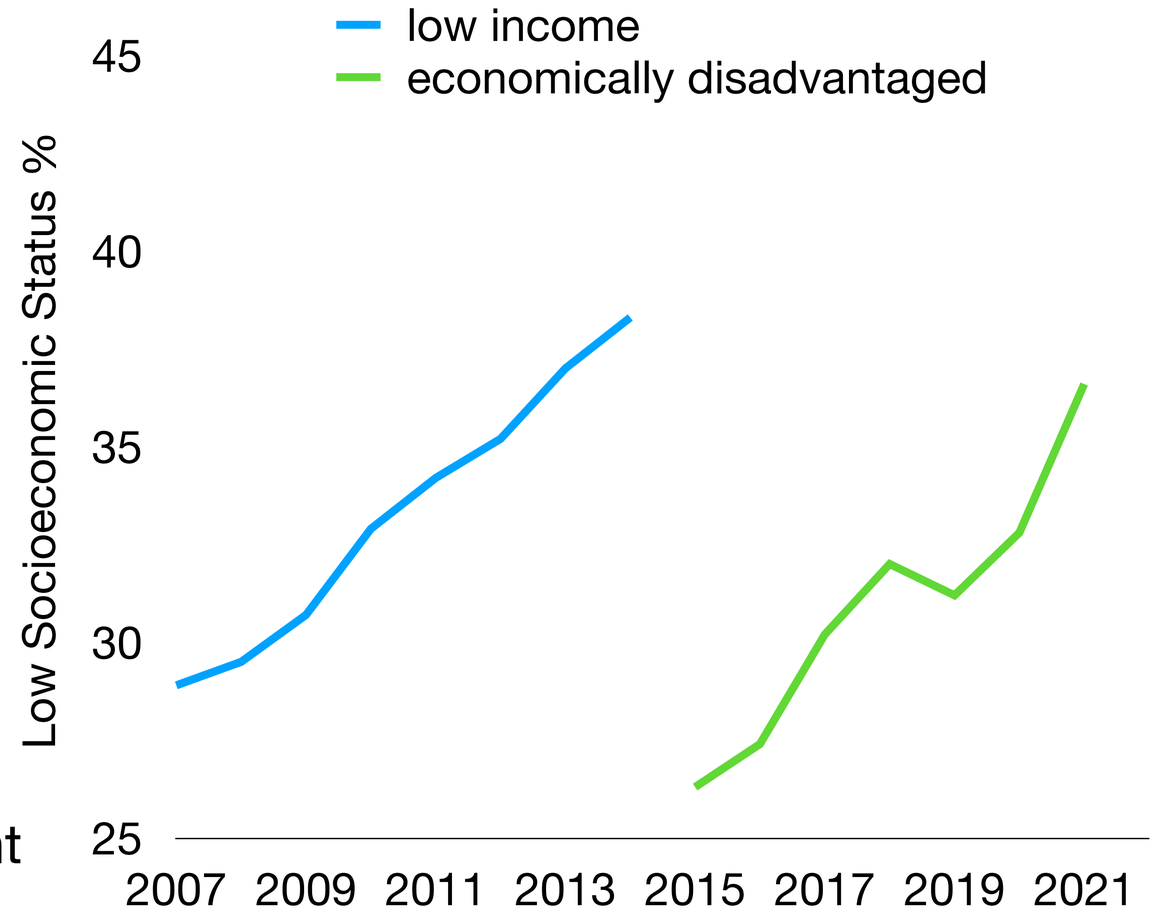
- The gender ratio between male and female is relatively stable through out the years, they are in line with childbirth gender ratio of US, which is 1.05 male vs. 1.0 female. This indicates that there is no gender inequity in the enrollment in MA.



Part 1 - Temporal Visualization (cont.)

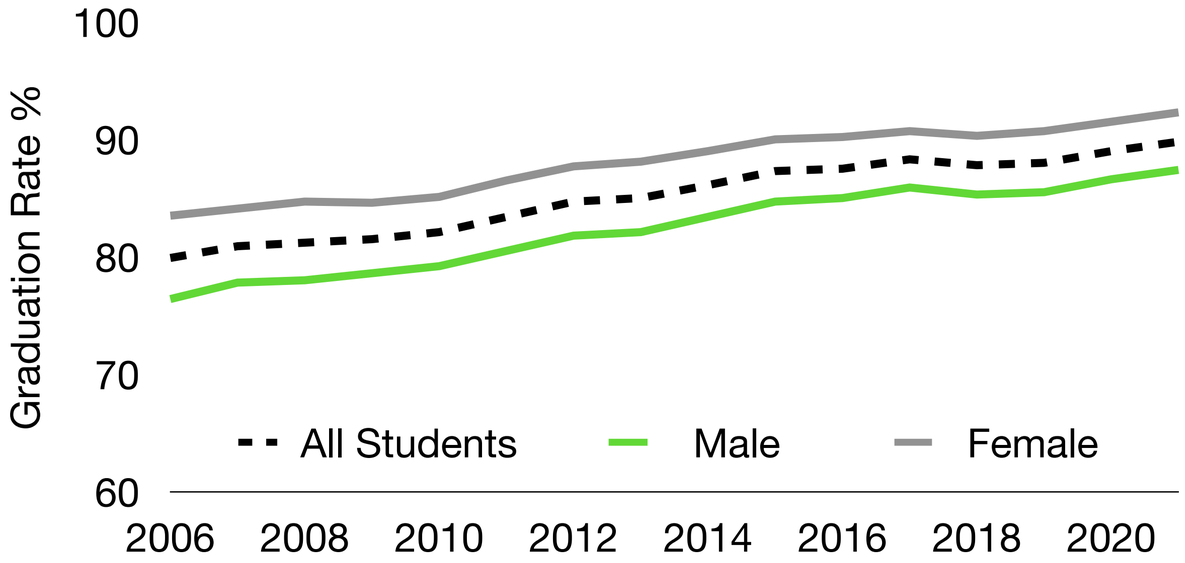
State Level Enrollment by *Socioeconomic*, 2007 - 2022

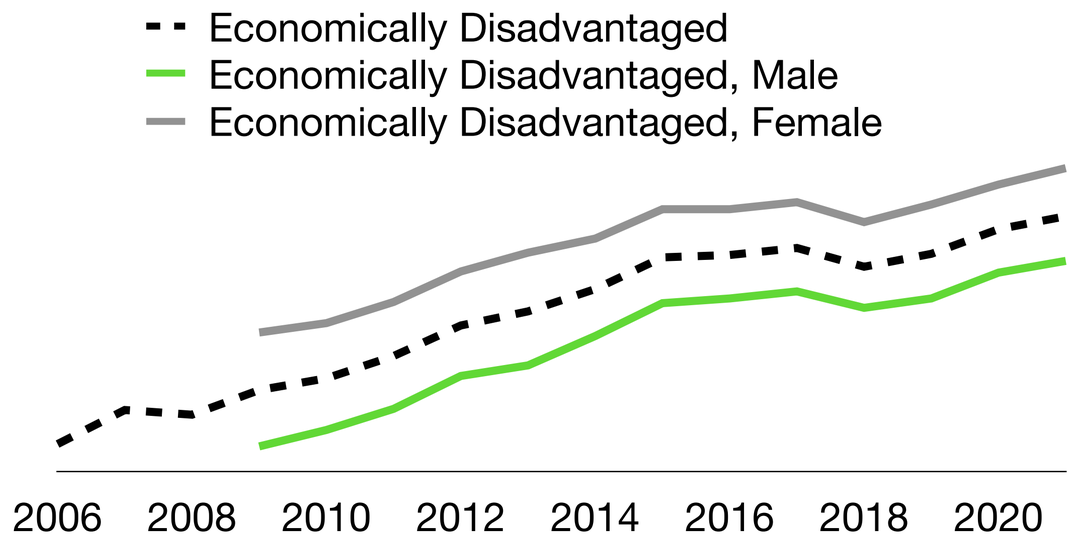
- Two criteria has been used to capture the socioeconomic status:
  - 2007 - 2014, and 2022, *low-income*
  - 2015 - 2021, *economically-disadvantaged*
- Note both criteria are calculated based on one student participates in one or more state-administered programs. For the *low-income*, it includes one more program and therefore it is generally higher than the *economically-disadvantaged*.
- Regardless of the differences between the two criteria, both indicate an increase of the student with low socioeconomic status over the years.



Part 1 - Temporal Visualization (cont.)

State Level Graduation by *Gender and Socioeconomic*, 2006 - 2021



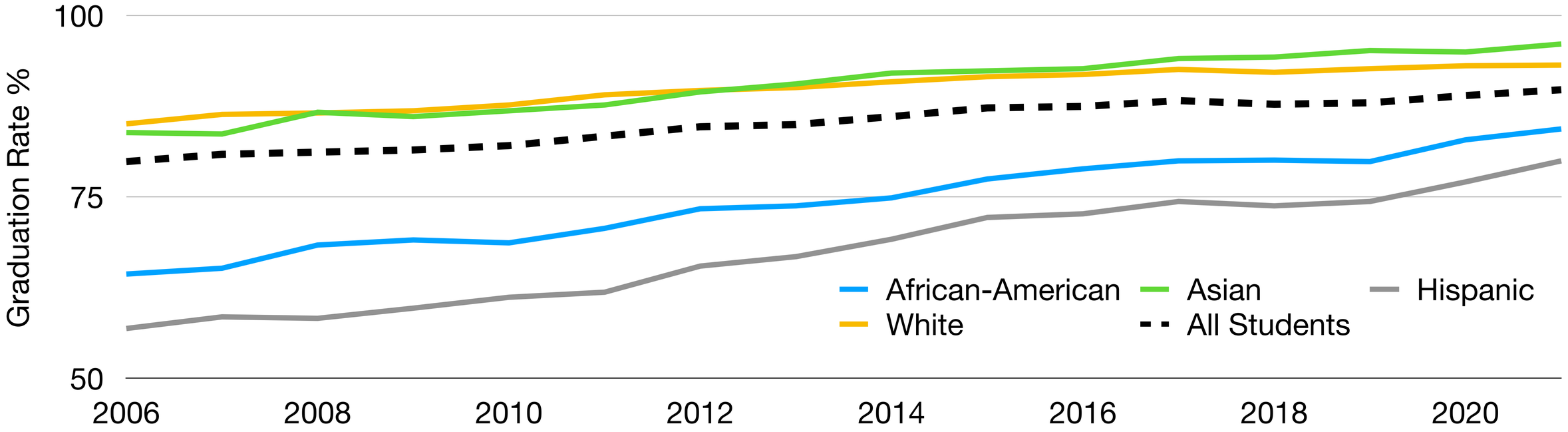


- **Left** figure shows graduation rates for all students, also separates by male and female. Notably, female always has higher graduation rate than male.
- **Right** figure focus on group with low socioeconomic status. While female still outperforms male, all three categories are lower than their all-student counterparts. This indicates that socioeconomic status does have significant impact on student graduation, regardless of gender.
- The good part is that the gap between two groups is decreasing over those years.

Part 1 - Temporal Visualization (cont.)

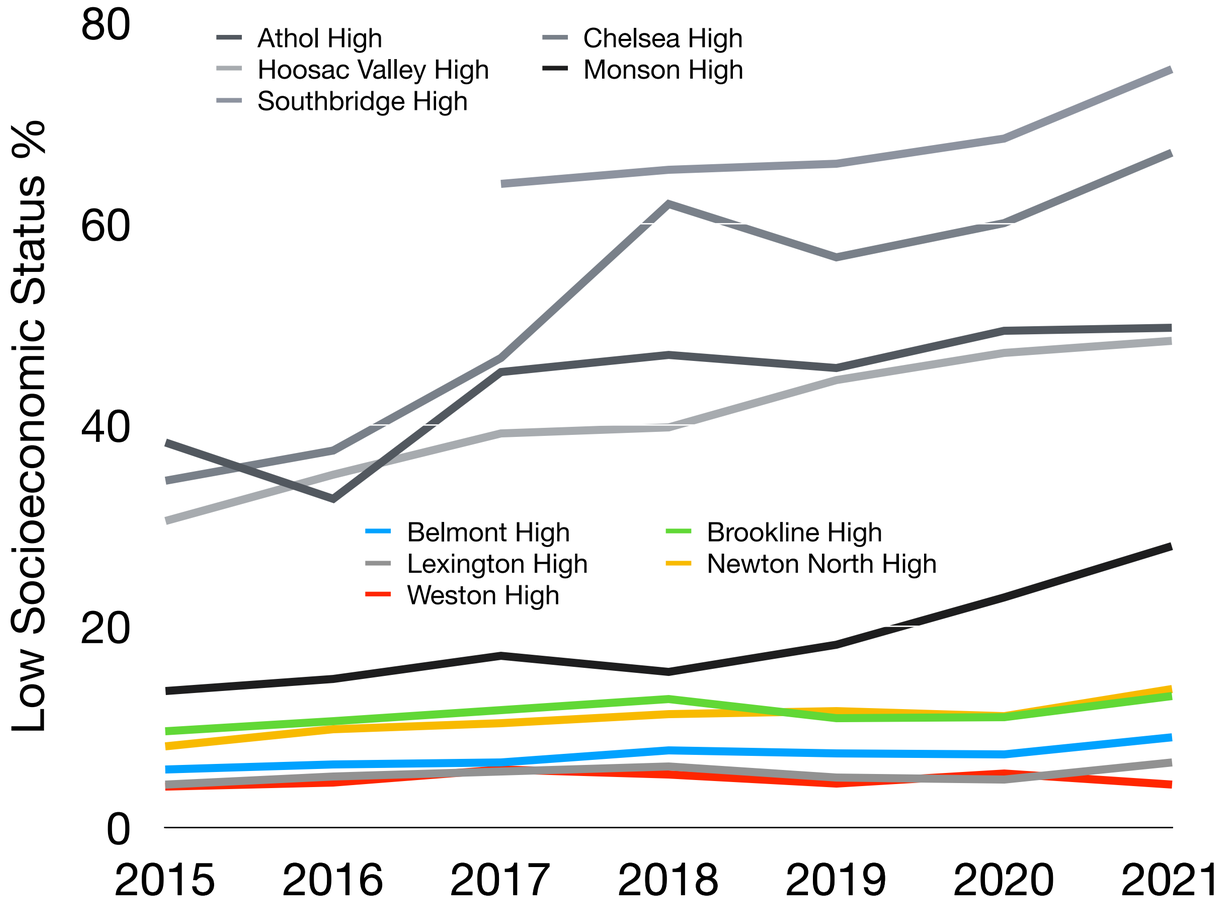
State Level Graduation by *Race*, 2006 - 2021

- Overall, graduation rate improves for all races, indicating enhancement of education quality in general.
- African-American and Hispanic increase faster than average (all students), and the gap between the four groups are much smaller in 2021 than in 2006.



Part 1 - Temporal Visualization (cont.)

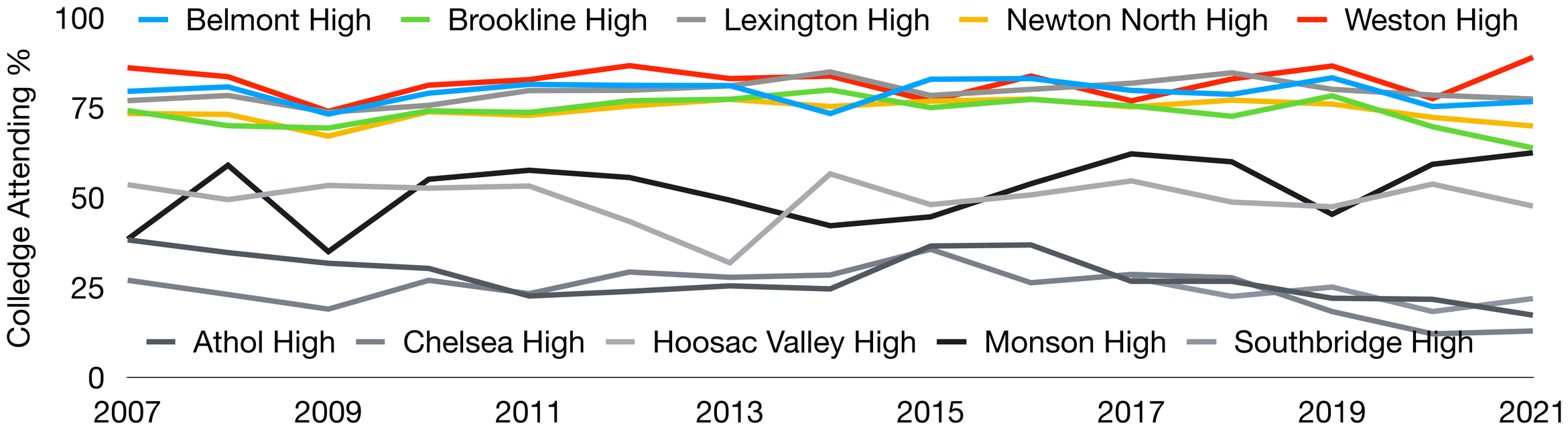
School Level Enrollment by *Socioeconomic*, 2007 - 2021



- We selected 10 schools based on their academic performance, within 5 on top of the list (colored lines) and 5 at the bottom (gray to black lines).
- Overall, most of 10 schools see an increase of students with low socioeconomic status. However, bottom 5 schools have more rapid increase, indicating enlarging demands of government supports for public education in these towns.
- This is also positively correlated with the median housing prices of these towns.

Part 1 - Temporal Visualization (cont.)

School Level *College Attending Rate*, 2007 - 2021

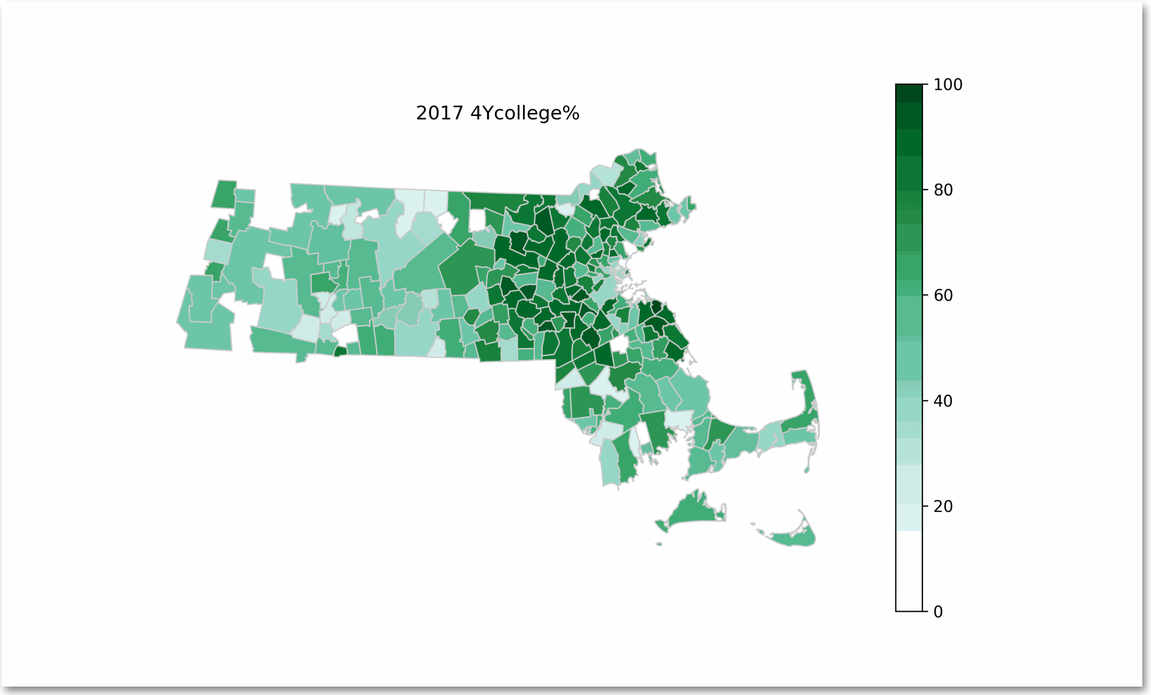


- Academic performance metrics such as AP, SAT, 4-year Colledge Attending Rates have been explored and similar patterns have been observed.
- Top figure shows College Attending Rates for top 5 and bottom 5 schools. Not surprisingly, top 5 schools have much higher College Attending Rate than bottom 5. The average of top 5 doubles the average of bottom 5 (78% vs 38%). Indicating huge schools and education inequity existing in MA.

Part 2 - Spatial Visualization

State Level Map by *College Attending Rate*, 2017 - 2021

- Visualize data on maps could reveal some patterns that are not easily to discover through charts, and could provide a quick overview.
- Several academic performance metrics aforementioned (AP, SAT, College Attending) have been evaluated through maps.
- School Districts with high *College Attending Rate* are concentrated in Northeast of MA, where the traditional “good” school districts are. The geographical distribution remains stable throughout years. Indicating the regional educational inequity exists for a fairly long time.



Interactive map contains information of school districts. file:///Users/yuanyuan/Downloads/Tufts/master's capstone/final report/school_district_map.html

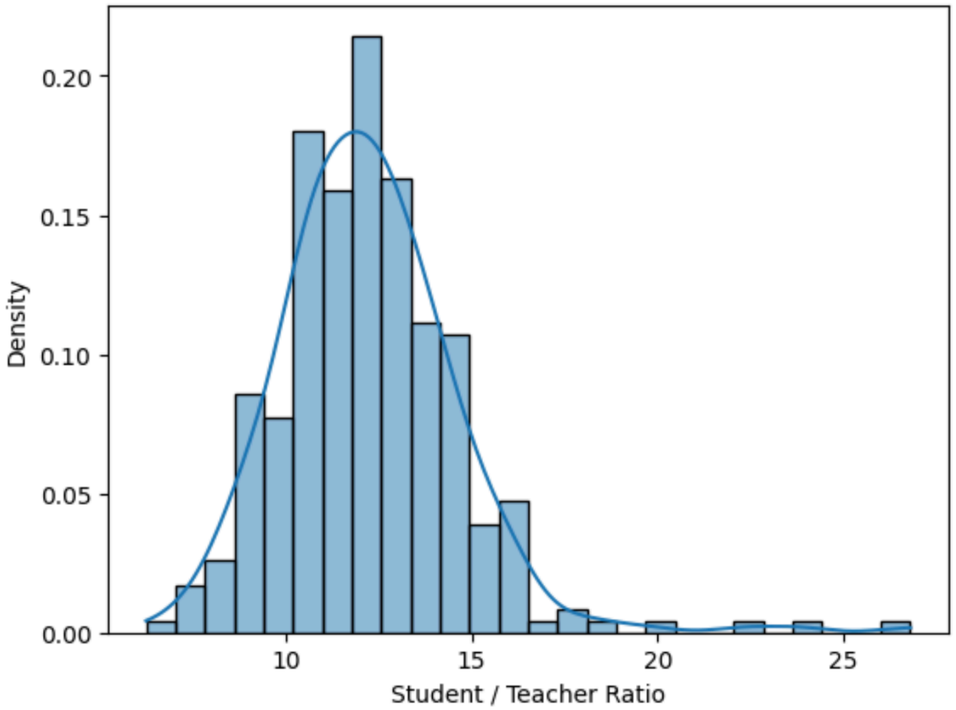
Part 2 - Spatial Visualization (cont.)

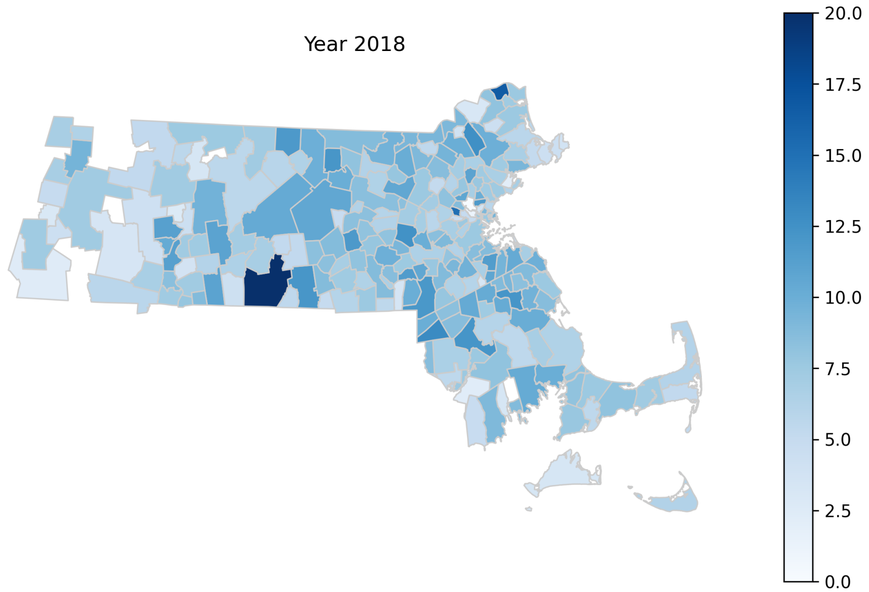
Exploring More Features with geographical distribution

- More features have been visualized and analyzed through maps to preliminarily screen out which features may have correlation with the academic performance, so that we can do in-depth visualizations and study later.
  - Student-Teacher Ratio
  - Experienced Teacher Percentage
  - Exemplary/Proficient Teacher Percentage
  - Socioeconomic Status
  - Teacher Average Salary
  - In-district Expenditure Per Pupil

Part 2 - Spatial Visualization (cont.)

Student-Teacher Ratio



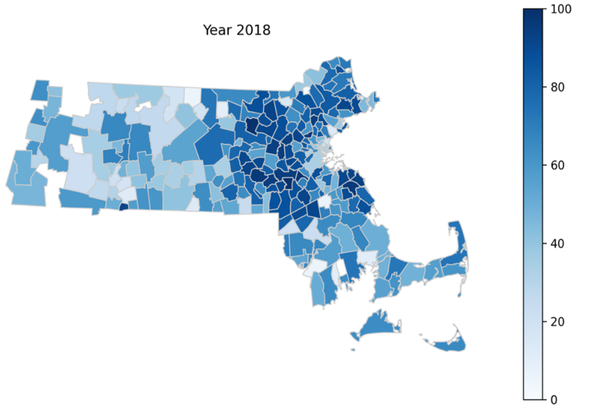


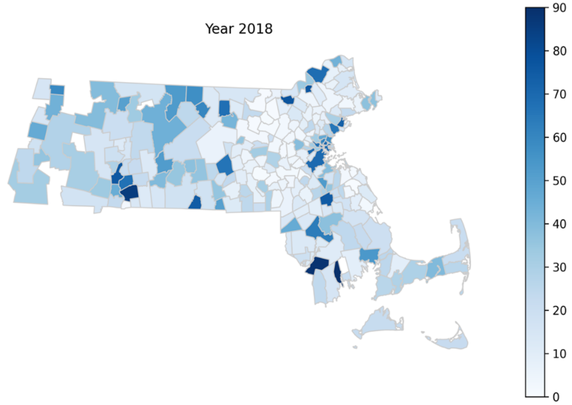
- Student-teacher ratio is using student number divided by teacher number, and lower ratio means each student could get more attention from the teacher.
- National average Public School Student/Teacher Ratio is about 15:1. Massachusetts average ratio is about 12:1.
- Student-teacher ratio in Massachusetts has a relatively even distribution.

Part 2 - Spatial Visualization (cont.)

Socioeconomic status, Average salary of teacher, and In-district expenditures per pupil show some correlation with the academic performance in their geographic distributions.

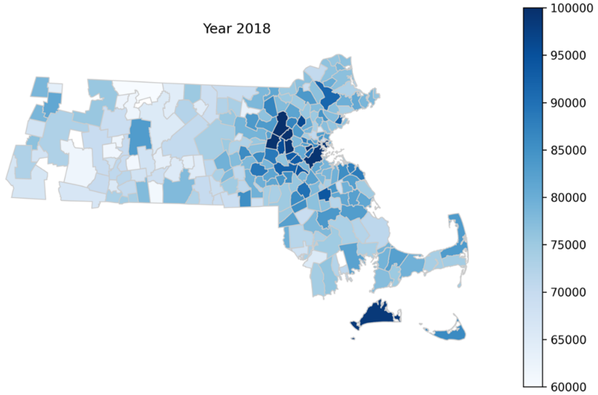
College attending %

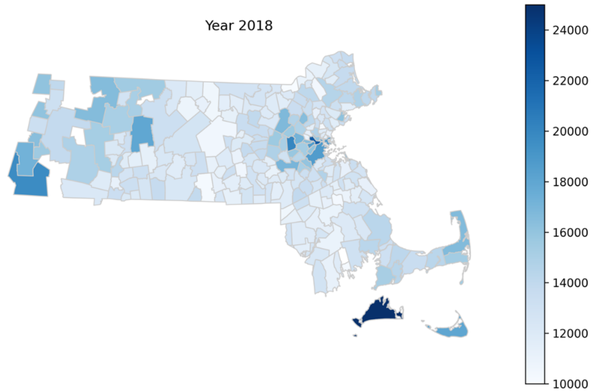




Economically Disadvantaged %

Average Salary





Expenditures Per Pupil

Part 3 - In-depth visualization for selected areas of interest

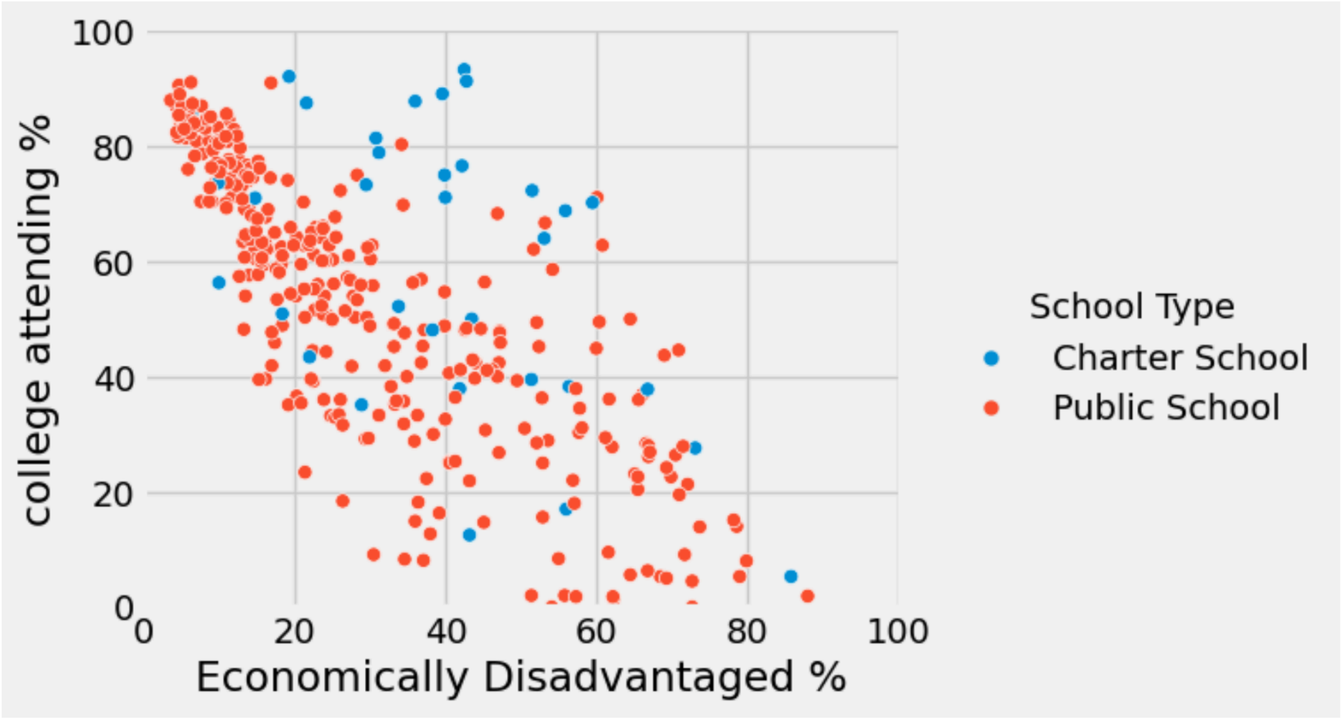
Focus on certain features: socioeconomic status, average salary of teacher, In-district expenditures per pupil, gender, and races that may have correlation with academic performance.

Conduct deep dive investigations that are manageable and may yield meaningful insights.

- For **Grades 3-8**: Academic performance metric is **Massachusetts Comprehensive Assessment System** (MCAS).
- For **Grades 9-12**: Academic performance metric is **College Attending Percentage**.

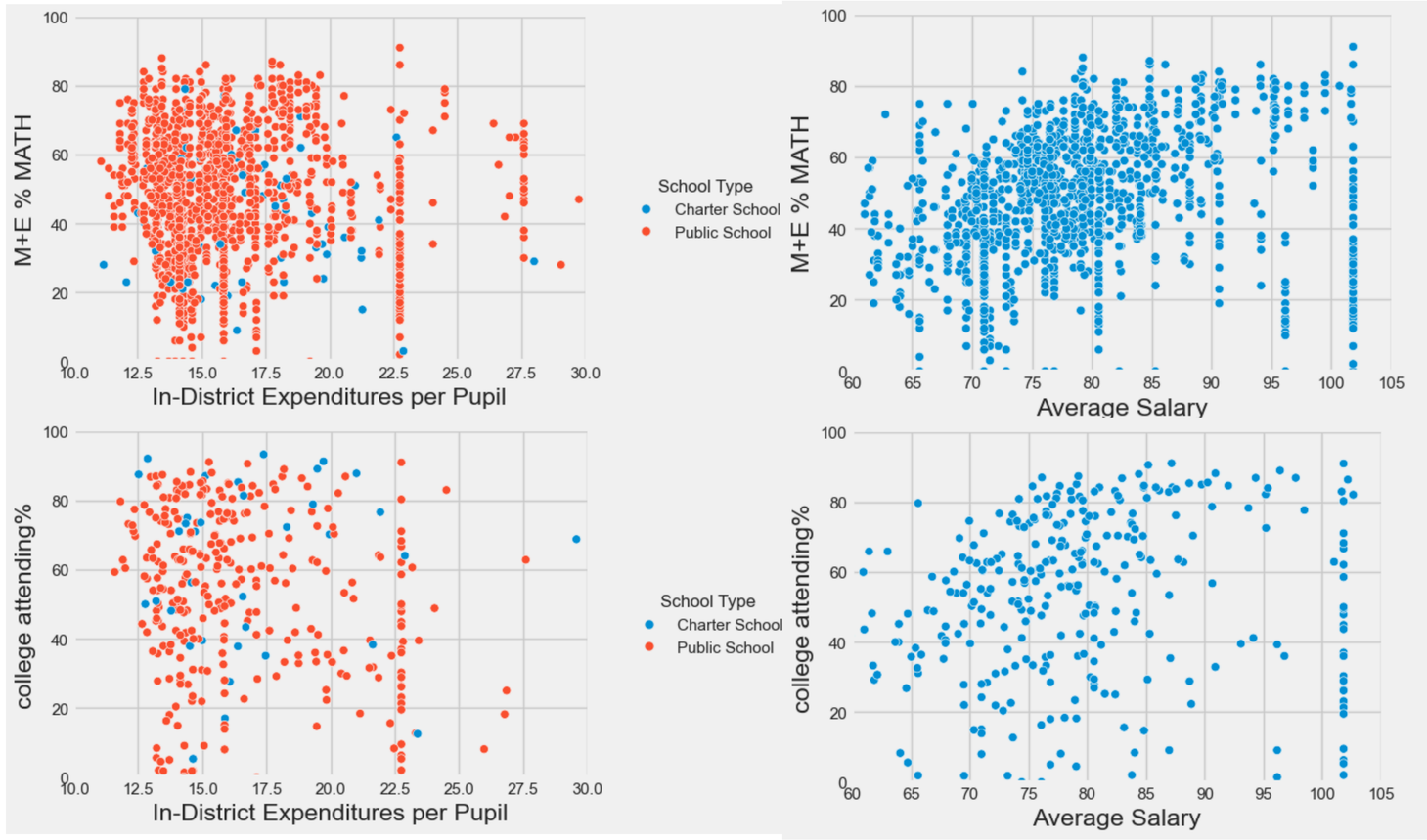
Part 3 - In-depth visualization for selected areas of interest (cont.)

- Compared to traditional public schools, the academic performance for students in Charter schools could be affected less by their economic situation.
- Unique public schools (exam, magnet, pilot) have better outcome than traditional ones.

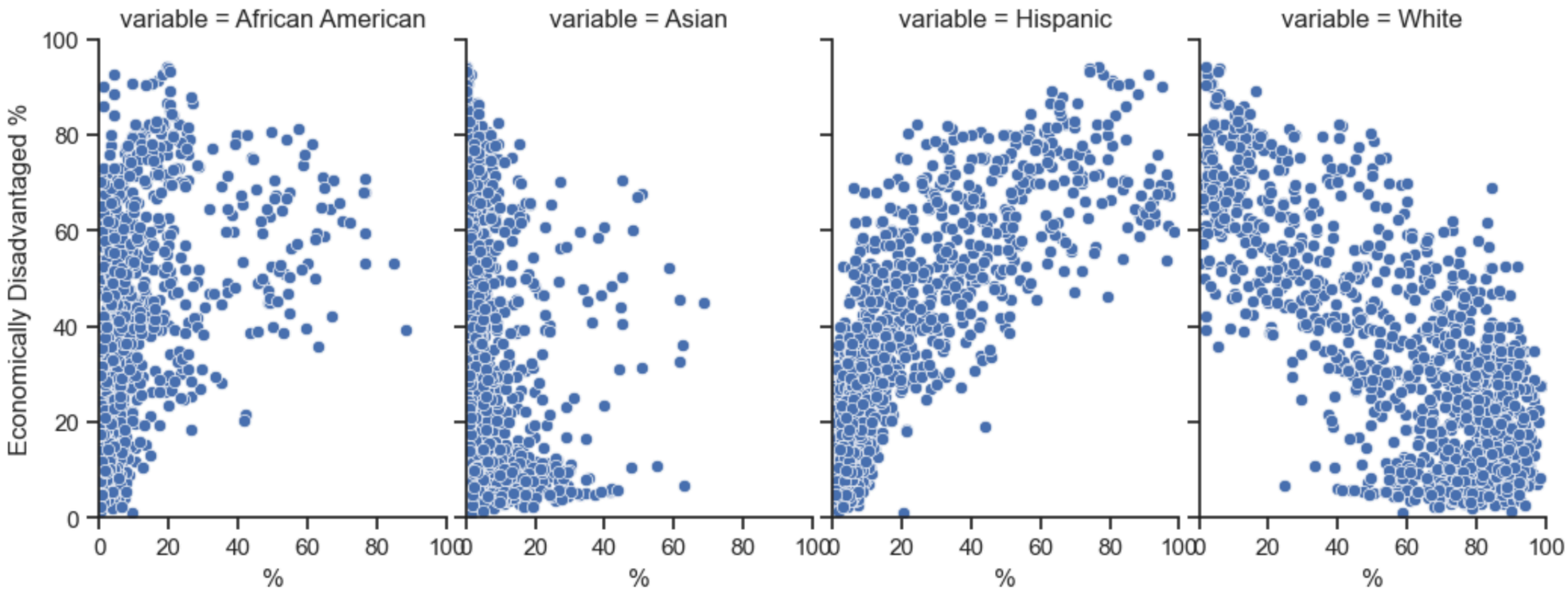


Part 3 - In-depth visualization for selected areas of interest (cont.)

- Expenditures do not have obvious impacts on the academic performance.
- Average Salary may have some correlation with the performance. we could take schools in range: 100k to 105k, which are schools in Boston city, as outlier.



Part 3 - In-depth visualization for selected areas of interest (cont.)



Except Asia, other three races have strong correlations with economic status of students.

Statistic analysis

Correlation study

Correlation is a statistical measure that tells us about the association between the two variables. It describes how one variable behaves if there is some change in the other variable. Therefore, after looking at the visual trends for the main features, I decided to dive deeper and study the potential correlations between them.

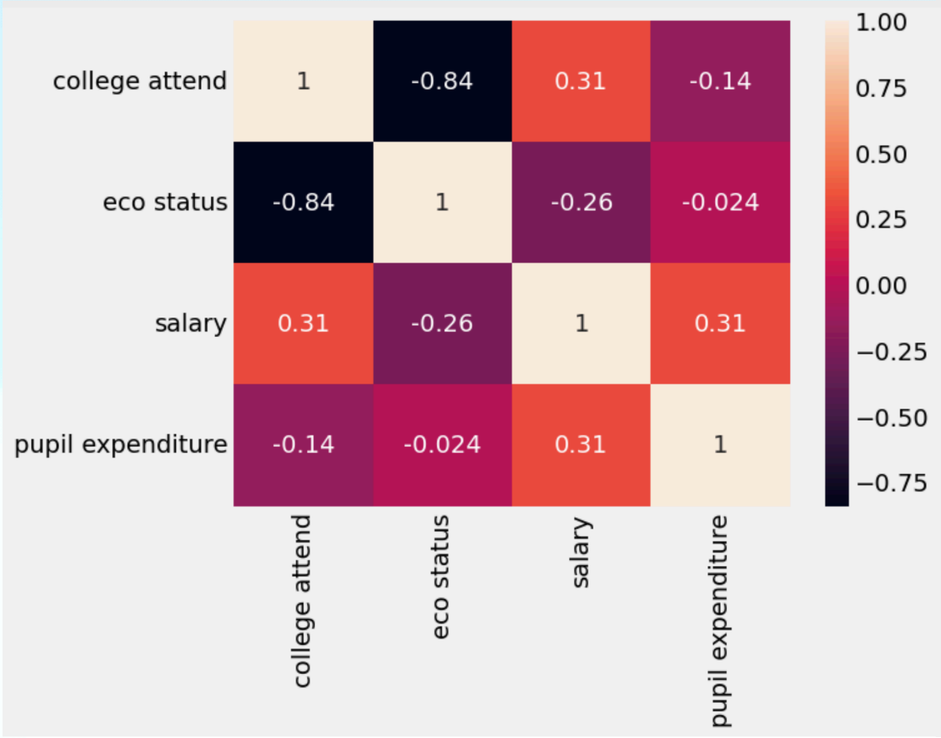
Two different studies regarding correlation:

- The Pearson Correlation Coefficient (for linear relationships between variables).
- The Spearman Correlation Coefficient (for monotonic relationships between variables).

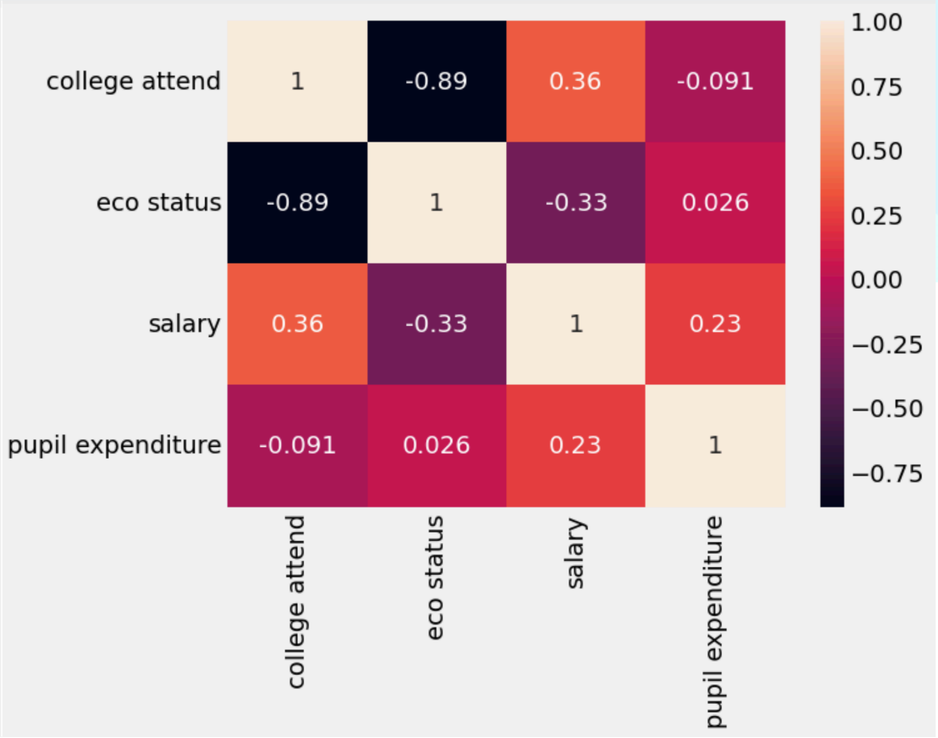
Statistic analysis

The Pearson and Spearman correlation

- Both correlation methods show that the **economic status** is the most significant factor that correlates with academic performance.
- Compared to per pupil expenditure, **average salary** does have more impact on the academic outcome.



Pearson correlation



Spearman correlation

Machine Learning model

Decision Trees

There are three reasons to implement Decision Trees model in our problem:

- The tree structure is intuitive and allows for an easy understanding of the decision-making process.
- Decision Trees are non-parametric, meaning they do not make assumptions about the underlying distribution of the data. This makes them useful for our data that is not normally distributed and there are outliers.
- Decision Trees can capture interactions between features, making them useful when there are complex relationships between the input variables.

Machine Learning model

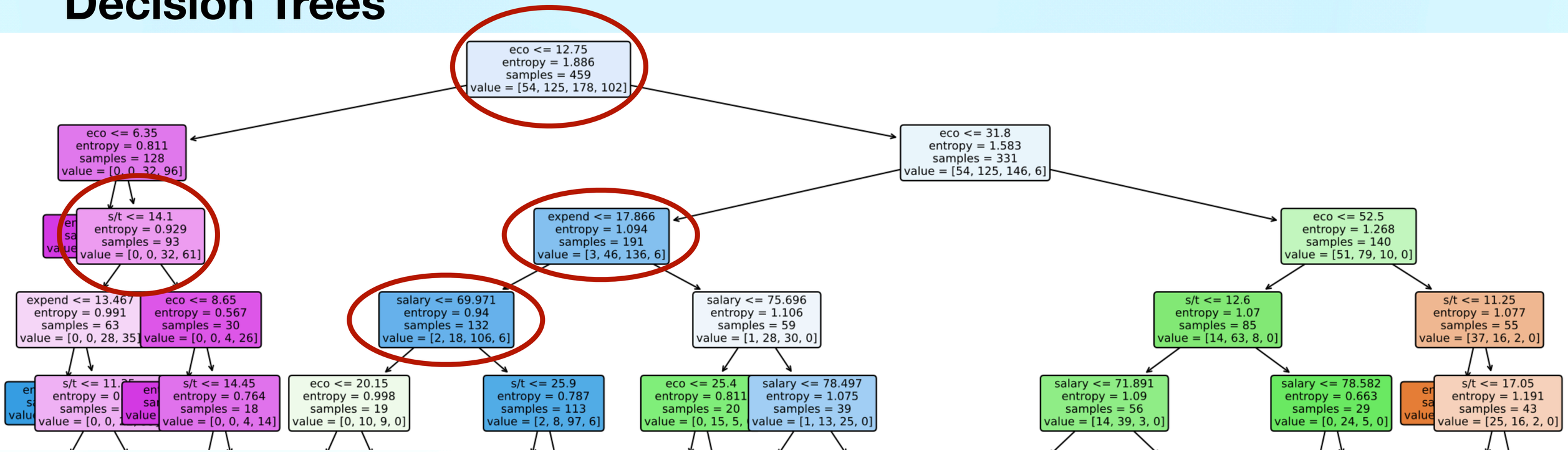
Decision Trees

- Dependent variable is “college attending %”, which has been assigned to one of the four levels (see range detail below).
- Build a multi-class *DecisionTreeClassifier*.
- Criterion is using ‘**entropy**' (better than criterion with “gini”)
- Train size 459, test size 115 (all high school for year 2018 and 2019).

Classification Report			
College Attending %	**Precision**	**Recall**	**F1_score**
Level 1 (0 - 25%)	0.69	0.69	0.69
Level 2 (25% - 50%)	0.75	0.70	0.72
Level 3 (50% - 75%)	0.67	0.76	0.71
Level 4 (75% - 100%)	0.88	0.79	0.83
Accuracy	Train: 1.000, Test: 0.739		

Machine Learning model

Decision Trees



- The above tree structure is from our fitted decision tree model (the first several levels). From this tree, we can see “**economic status**” is the most important feature (root node) through entropy criterion. Followed by “**salary**”, “**expenditure per pupil**” and “**student teacher ratio**”.
- This inline with the observation from data visualization as well as statistical analysis.

Conclusion

- **Socioeconomic status** is the most important factor that could affect academic performance.
- **COVID-19** did impact the total enrollment, and MA has no bias in gender enrollment.
- MA **teacher-student ratio** outperforms national average, and is relative even within state.
- **Race discrepancy** for academic performance is decreasing over years.
- **Charter schools** and specially designed schools are less influenced by economic status. Therefore, Government should encourage more innovative educational practices and education reforms to provide parents and students with more greater choices in public education, and to provide models for replication in traditional public schools.
- Reasonably increase **teacher salaries** to motivate more talent educators joining public education system.
- The decision-maker should improve the policy of **resource allocations**, because the current higher funding does not cause an increase in student performance.
- We need to redesign **educator evaluation framework** since the current evaluation results do not have relations with student academic performance.