OPERATION VARSITY BLUES

Kevin Dackow, Joshua Levin,
Dao Han Lim, Peter Mao

Exploring the College Scorecard: Racial Diversity and What Makes an Ivy

Introduction

THE IVY LEAGUE

The College Scorecard Dataset is an online tool provided by the US Government for people to "compare the cost and value of higher education". The dataset contains college statistics from the last 20 years, formatted and cleaned.

In the wake of several college scandals in the past few years, we wanted to take a deep dive into diversity across US colleges. We also wanted to find out what schools were most similar to the Ivy League.

With these interests in mind, we chose to explore the following topics:

- Racial diversity and distribution across the country
  - Looking by state and academic rigor
  - Considering outside factors like school size
- What schools are most similar to an Ivy
  - Which are closest to all ivies
  - Which are closest to each ivy

Hypotheses

We wanted to test:

- How does diversity change with respect to a college's SAT scores and admissions rate as indicators for college rigor?
- What are the factors that define an Ivy?

We wanted to see if there is a significant difference in the levels of diversity across colleges with different scores, and quantify these changes.

Challenges

1. Dealing with many null/missing values while maintaining and using as much of the dataset as possible
2. Normalizing values across different scales and units when comparing school feature vectors

Methodology

We considered schools over a 4 year period that reported their demographic data, and filtered out schools that historically had a demographic of only one race.

For our diversity analysis, we first divided the colleges into distinct groups based on the mean SAT score of their students.

To consider the underlying population distribution of the US, we took into account the population distribution by state. The figure on the right compares the average distribution of college students in each state to the population distribution of the state.

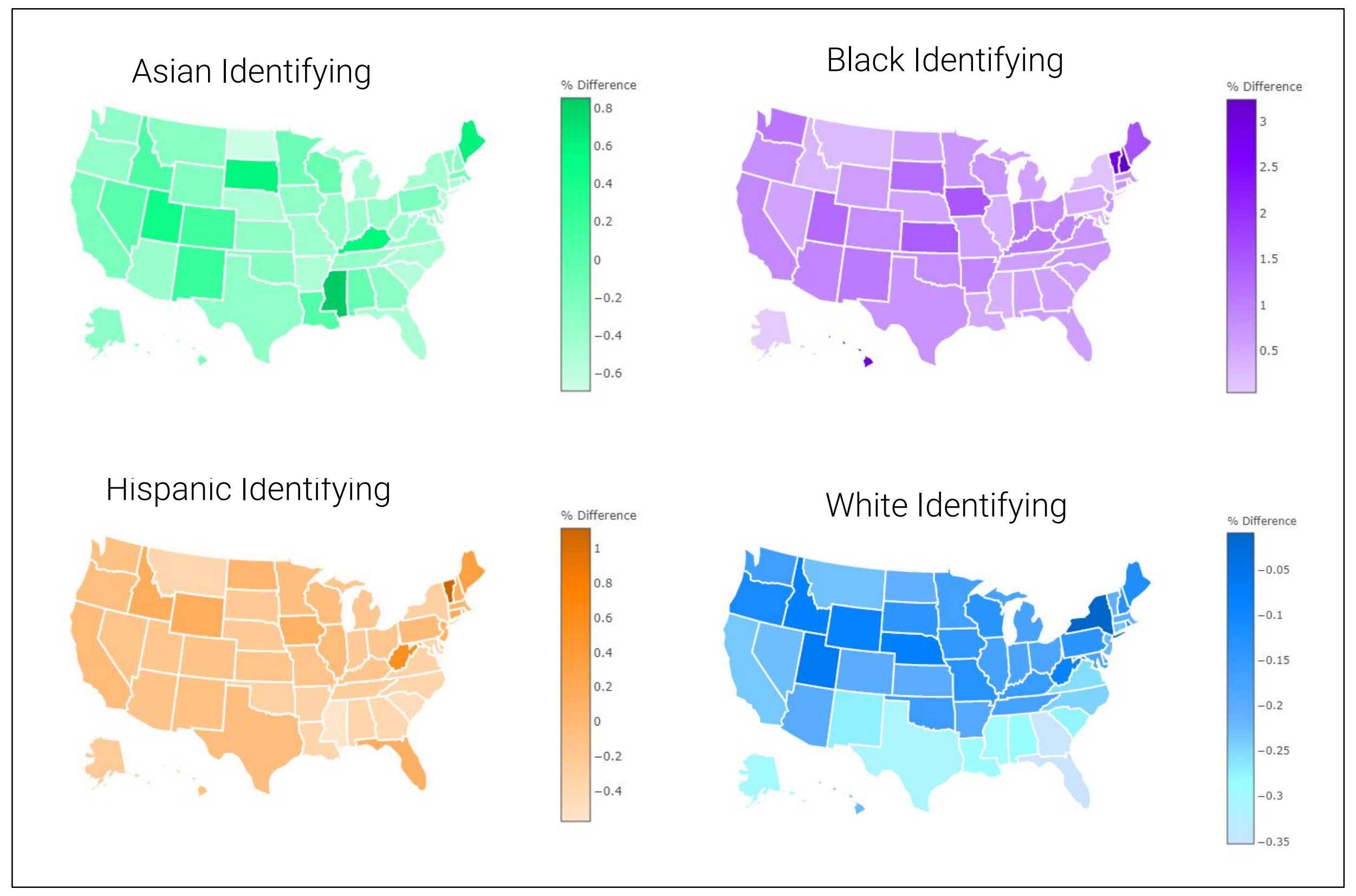
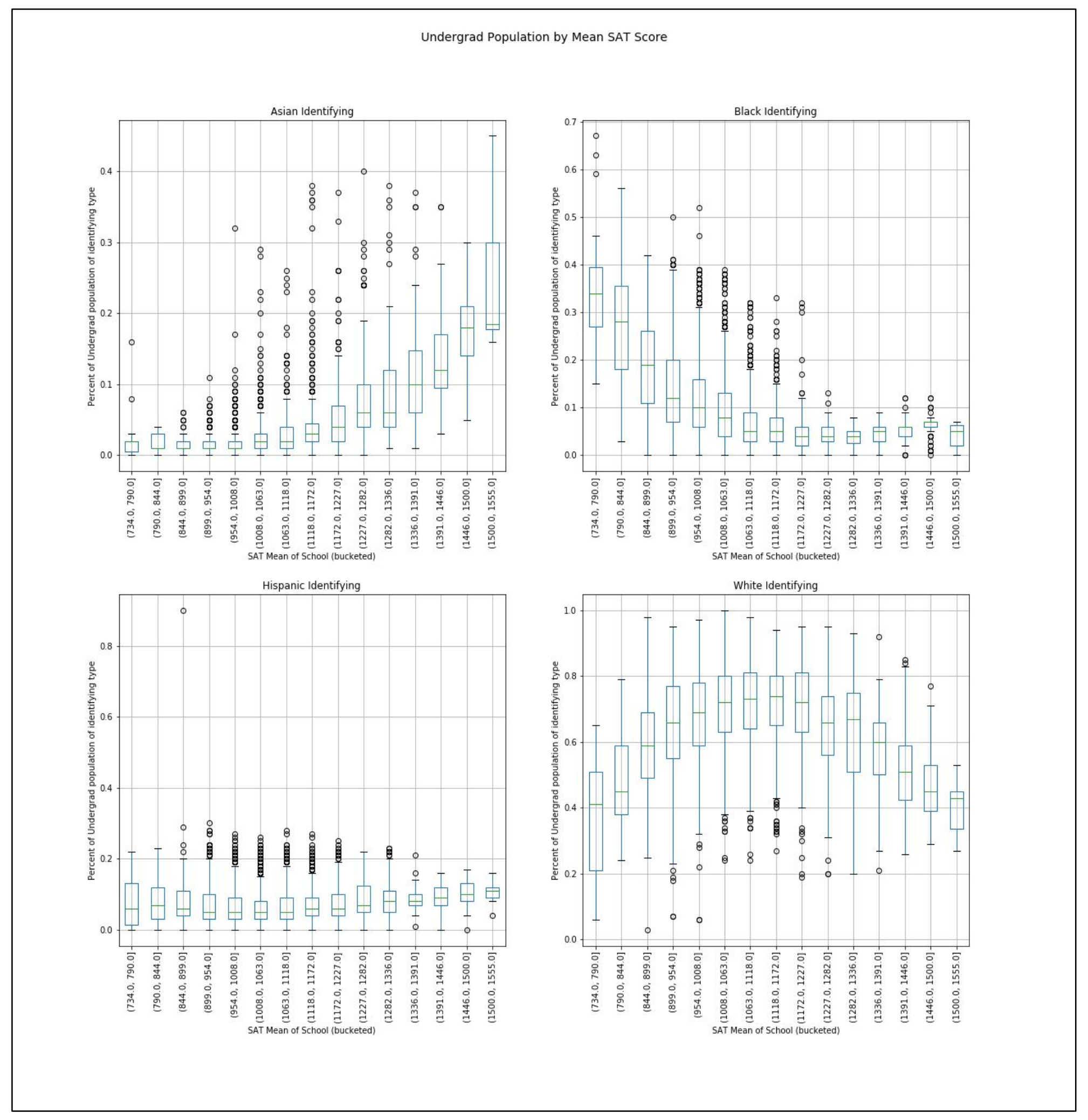
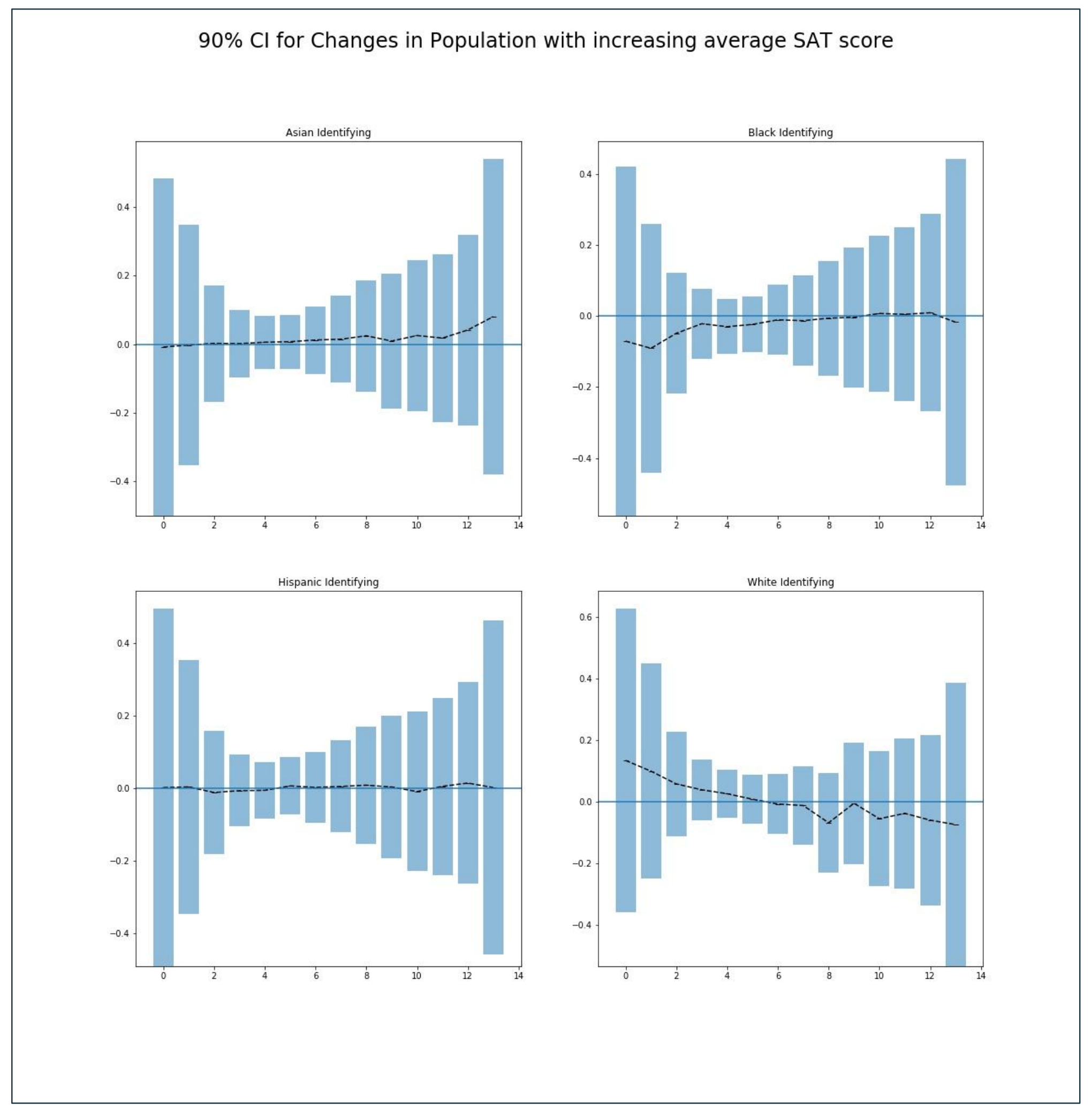


Figure 1: Percent difference between university and state populations.
A negative percentage indicates that the proportion of college students of a given race is smaller than the proportion of the state population of said race.

Diversity and SAT scores





We plotted over 4000+ of these data points to see the median, first and third quartile, as well as outliers. To support our plots, we then used an ANOVA test to see if the mean of each bucket was changing, and by how much.

We then performed a linear regression to fit the population representation to (1) median SAT scores and (2) size of the school.

The t-statistic is (1) ~42 median SAT scores and (2) ~7 for the schools size. Indicating a stronger correlation with SAT scores than school population.

To test if the mean of each SAT score bucket is changing, we used a two-way ANOVA test with the null hypothesis that the means of each bucket is the same.

We arrived with the following p-values:

Asian Identifying:	~0
Black Identifying:	2.37 e -280
Hispanic Identifying:	6.47 e -24
White Identifying:	5.44 e -160

What makes an Ivy?

We selected and normalized a wide range of school-level statistics across student body economics, diversity, and school size, faculty salary and mode. A few examples of the 24 attributes we selected were admissions rate, undergraduate size, proportion white/black/hispanic/asian, academic year cost, average faculty salary, 4 year completion rate, median debt for graduating students, and average family income.

On the left are the results when we looked at which schools were most similar to each college in the Ivy League. On the right are the colleges most similar to all colleges in the Ivy League. Unsurprisingly, Stanford is number one!

Ivy	Closest	2nd Closest	3rd Closest
Brown	Georgetown	Tufts	Colgate
Columbia	Gallaudet	University of the Pacific	LIU Post
Cornell	Vanderbilt	Duke	UChicago
Dartmouth	Sarah Lawrence	Georgetown	Rice
Harvard	Pomona	Stanford	Amherst
Penn	Life University	USC	Pomona
Princeton	Pomona	Amherst	Stanford
Yale	Amherst	Wellesley	Stanford

The Next 10 Ivys
Stanford
Amherst
UChicago
Northwestern
MIT
Vanderbilt
Rice
Duke
Wellesley
Emory

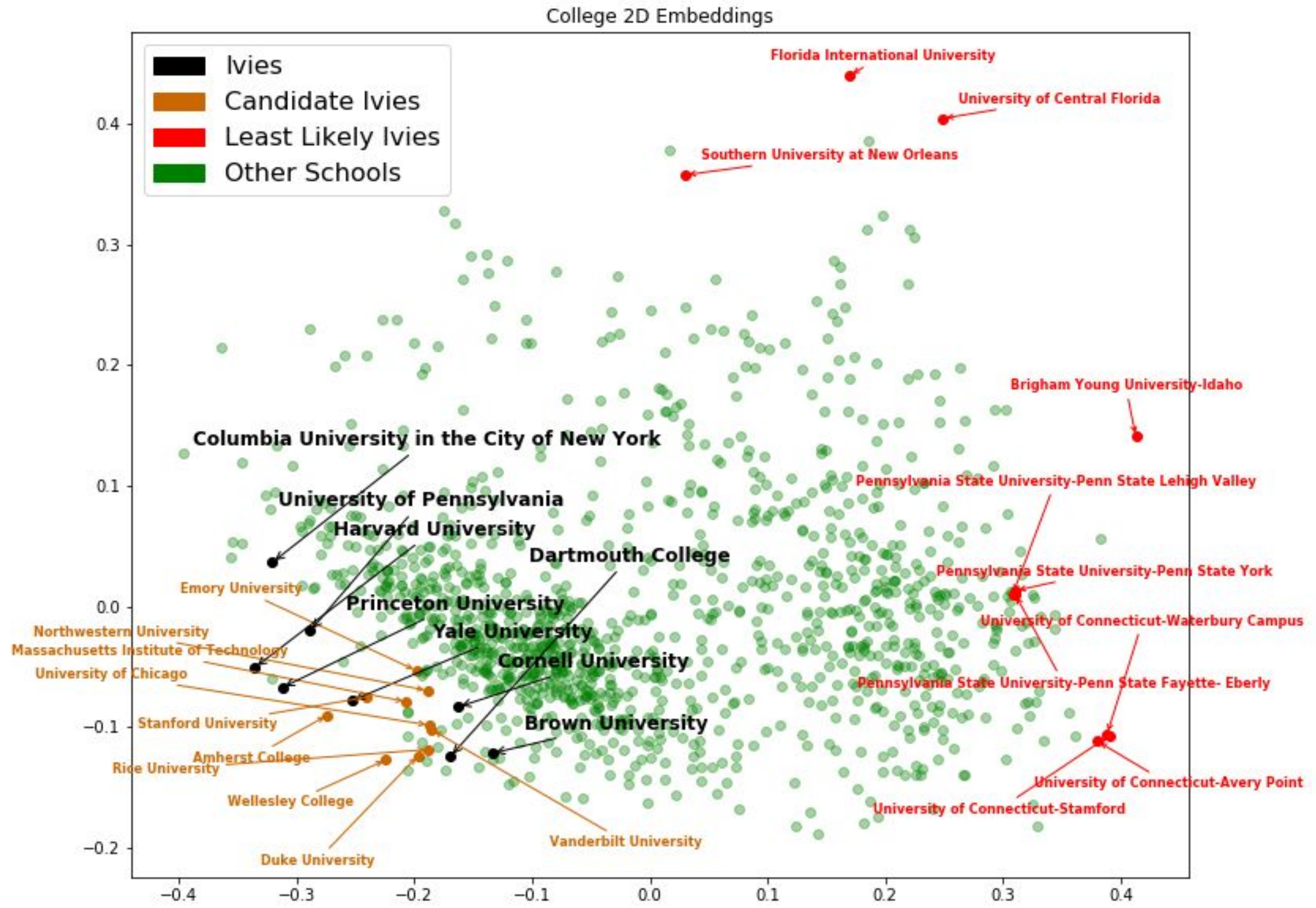


Figure 2: A 2 Dimensional (PCA Reduction) Representation of the School Vectors

Tools:

scikit learn matplotlib StatsModels Statistics in Python Pandas jupyter python NumPy SQLite d3.js