Midterm Project 1

Brady Lamson

March 14, 2022

Data Work

To accomplish this project a large amount of data wrangling was necessary. Thankfully this wrangling was *largely* consistent across visualizations, with some slight differences depending on the specific analysis being done. The first step I took was combining the two provided csv files. Since the four year and two year files shared the exact same format it was trivial to combine them using **dplyr::bind_rows()** and saved me a great deal of trouble in the long run.

Next I began creating a couple simple helper functions, one to calculate average differences for a data frame and the other to create a simple table utilizing the kableExtra package. The first of these two is used in nearly every single visual made and is the meat of the data wrangling being done. The function selects out only the columns starting with *"dif"*, computes a summarized mean for every one of those columns and then renames all of them to be a more presentable format for visualizations. Finally it converts this new data frame to tidy format using **tidyr::pivot_longer()** to allow the ggplot2 package to work its magic. Any other data wrangling was largely smaller in scale, typically involving **dplyr::group_by()** to prep the data frame before passing it into the difference calculator function I wrote.

Plots

Figure 1

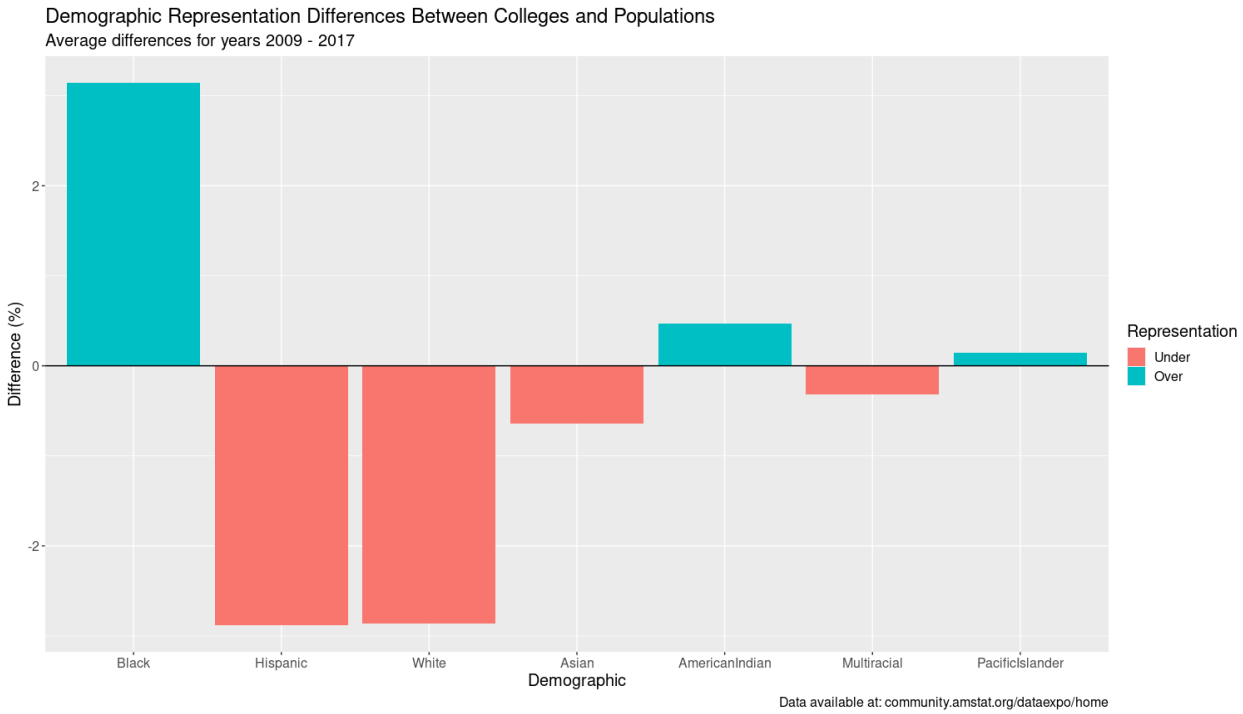


Figure 2

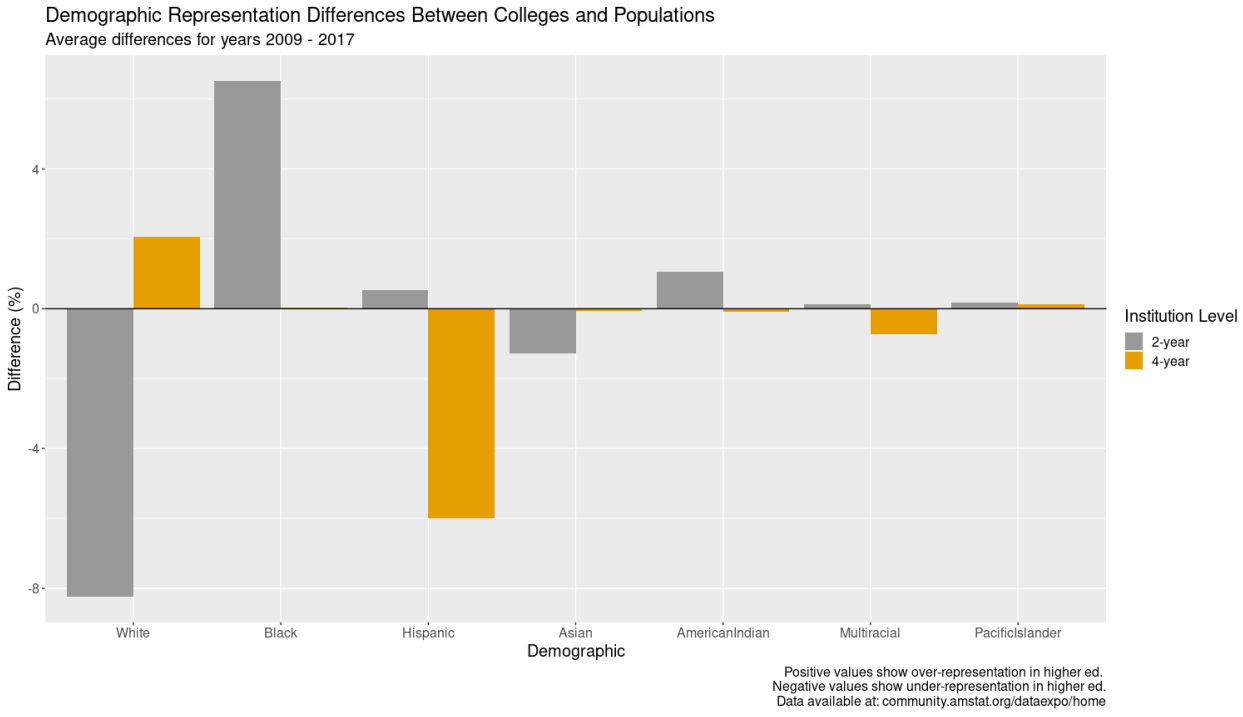


Figure 3

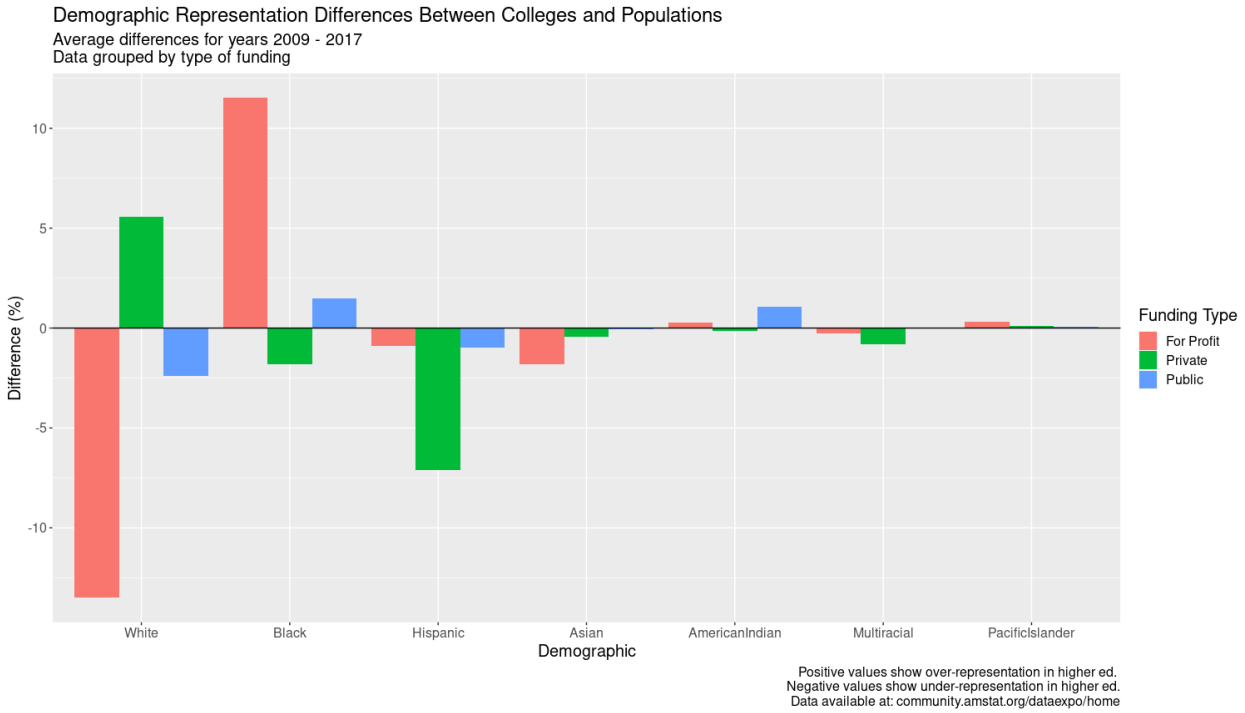


Figure 4

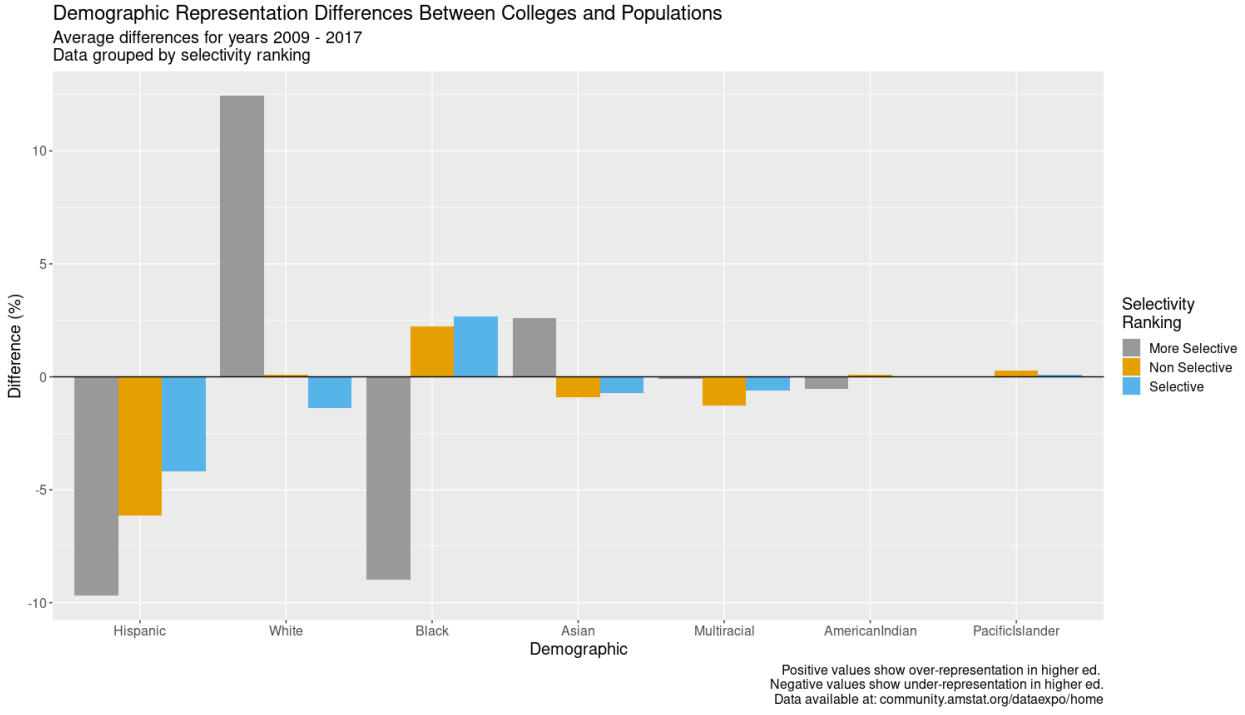
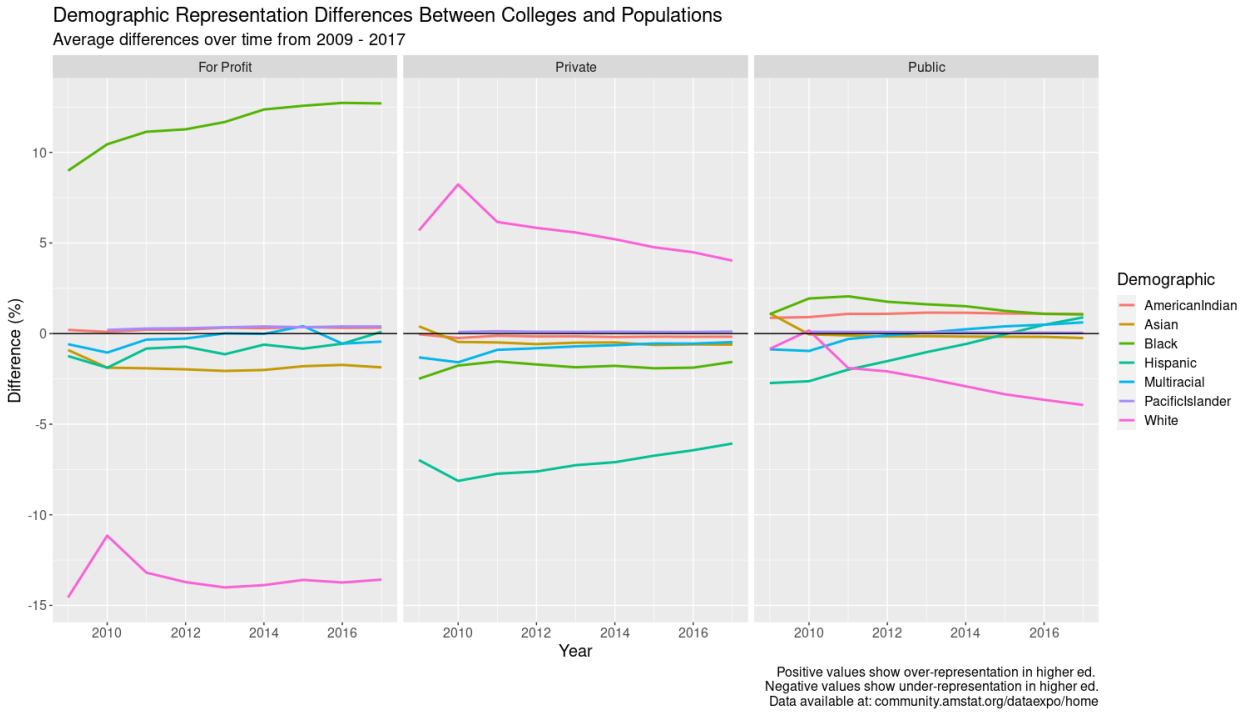


Figure 5


Demographic Representation Differences Between Colleges and Populations
Average differences over time from 2009 - 2017
For Profit
Private
Public
Difference (%)
10
5
0
-5
-10
-15
2010
2012
2014
2016
Year
Demographic
AmericanIndian
Asian
Black
Hispanic
Multiracial
PacificIslander
White
Positive values show over-representation in higher ed.
Negative values show under-representation in higher ed.
Data available at: community.amstat.org/dataexpo/home

Tables

Figure 6

Overall Representation Differences (%) by Demographic

White	Hispanic	Black	Asian	AmericanIndian	PacificIslander	Multiracial
-2.864503	-2.877366	3.139005	-0.6449722	0.4648527	0.149238	-0.3151498

Figure 7

Representation Differences (%) Grouped by School Profit Status

profit_status	White	Hispanic	Black	Asian	AmericanIndian	PacificIslander	Multiracial
For Profit	-13.471702	-0.8954512	11.512145	-1.8178387	0.2574803	0.3274676	-0.2851833
Private	5.554934	-7.1200504	-1.832079	-0.4409861	-0.1580736	0.1010500	-0.7986164
Public	-2.376334	-0.9636711	1.495623	-0.0450153	1.0658531	0.0716155	0.0191217

Figure 8

Representation Differences (%) by Institution Level

slevel	White	Hispanic	Black	Asian	AmericanIndian	PacificIslander	Multiracial
2-year	-8.230215	0.5250744	6.5192186	-1.2671895	1.0679816	0.1675113	0.1160342
4-year	2.069017	-6.0057497	0.0310581	-0.0728726	-0.0896958	0.1319228	-0.7278570

Figure 9

Representation Differences (%) by Selectivity Ranking

how_selective	White	Hispanic	Black	Asian	AmericanIndian	PacificIslander	Multiracial
More Selective	12.4301347	-9.668411	-8.963038	2.6038388	-0.5214001	-0.0350654	-0.0899183
Non Selective	0.0994982	-6.143813	2.230361	-0.9046117	0.0838646	0.2782288	-1.2771006
Selective	-1.3827663	-4.185463	2.675240	-0.7323396	-0.0116869	0.1042069	-0.6255491

Research Questions

Q1: Overall, to what degree do college racial and ethnic compositions differ from the racial and ethnic compositions of the institutions' geographic "markets"?

Figure 1 and **Figure 6** both do a good job showcasing general differences between college demographics versus geographic representations. At the most general we see that these representations are fairly accurate, with the worst cases being over or under representation by a few percent.

Q2: For which specific racial or ethnic groups are the discrepancies between their representations in colleges and their representations in the "markets" largest?

Across every single figure in the report you will see the same three groups appear in various orders. **Black White** and **Hispanic** are, across the board, the groups with the largest representation issues. In general, that is, without grouping by the type of school, the **black** population is over represented by around **3.1%** and the **Hispanic** and **white** communities are under represented by around **2.8%** (Figure 6). It is of note that, despite groups such as multi-racial appearing well represented, that this data is still subject to any bias that stems from its collection. Multiracial populations have historically been under-counted, so it is difficult to draw conclusions about them from this data alone. Similar critiques may also apply to any of the other demographics concerned and should be taken into consideration.

Q3: If colleges are grouped by institution level, degree of selectivity, and/or public/private/for profit status, do the discrepancies between college and "market" racial and ethnic compositions vary across groups? In other words are the discrepancies larger for some types of colleges than others? If so, for which types of colleges are the discrepancies largest?

This question is where I find the meat of the analysis to be. As mentioned in my answer to **Q2**, the same three groups appear at the top of every visualization. **Black**, **white**, and **Hispanic** populations are consistently mis-represented, but the specific type and degree of mis-representation varies a **ton** between types of school.

Let us first examine the representation discrepancies for one of those three previously mentioned groups. Looking at figure 2 and figure 8 is where we immediately see that the general answer to **Q1** isn't showing the full picture. Here we see that for two-year institutions, the white population is severely under-represented by around **8%**. This isn't the only type of school where this under-representation happens. Jumping over to figure 3 and figure 7 shows an under-representation of over **13%** for the white population but it also shows an over-representation at private schools by over **5%** and figures 9 and shows an enormous over-representation of over **12%** for schools with the highest selectivity ranking.

We see this type of flip-flopping with the representation of the black population as well. A sizable over-representation at 2-year institutions and for profit schools but a substantial under-representation at schools with the highest selectivity ranking. Strangely enough, the Hispanic population does not experience such flip flopping. Across every type of school this group is either represented fairly or is severly under-represented. This is very apparent when looking at selectivity ranking, where Hispanic individuals are under-represented across every type of that ranking, with around a **9.7%** under-representation at schools with the highest selectivity ranking.

Overall, the types of schools that see the largest degree of mis-representation are schools with the highest selectivity ranking and for-profit institutions. 2-year institutions come close but don't quite experience this to the same degree as the others. We can also see that public schools tend to be the most representative of their geographic populations with the most extreme mis-representation being approximately **-2.48%** for the white population.