Midterm Project 1

Brady Lamson

3/4/2022

```
# Read in data

dictionary <- readxl::read_xlsx("HEsegDataviz_Dictionary.xlsx")
four_year <- readr::read_csv("HEsegDataviz_CollegeData_4-year_v5.csv")
two_year <- readr::read_csv("HEsegDataviz_CollegeData_2-year_v5.csv")
```

Research Question 1

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

To answer this question I decided to calculate these differences separately for four and two year institutions. For this question the information we need is readily available from the data set itself so we can easily select out the difference columns specifically and get an average.

Code

```
# Calculate average differences for each demographic column
# Then create a summarised data frame of those averages

# Four year average differences -----
average_four_year_diffs <- four_year %>%
    select(starts_with("dif")) %>%
    summarise(
        across(
            .cols = everything(),
            .fns = mean, na.rm = TRUE
        )
    )

# Two year average differences -----
average_two_year_diffs <- two_year %>%
    select(starts_with("dif")) %>%
    summarise(
        across(
            .cols = everything(),
            .fns = mean, na.rm = TRUE
        )
    )
```

```r
# Combine the two difference data frames together
# Add a new column for institution type

average_diff_df <- dplyr::bind_rows(
    average_four_year_diffs,
    average_two_year_diffs
) %>%
    dplyr::mutate(
        institution_type = c("Four Year", "Two Year")
    ) %>%
    dplyr::relocate(
        institution_type
    )
```

```r
# Turn data frame into a clean HTML table using kableExtra.
q1_tbl <- average_diff_df %>%
    kableExtra::kbl(
        booktabs = T,
        caption = "College versus Geographic Demographic Differences (Percent)",
        col.names = c(
            "Institution Type", "White", "Hispanic",
            "Black", "Asian", "American Indian",
            "Pacific Islander", "Multiracial"
        )
    ) %>%
    kableExtra::kable_styling(latex_options = c("scale_down", "hold_position"))
```

```r
# Plot averages of four-year and two-year institutions

# year-four difference plot ---------------
q1_y4_plot <- average_four_year_diffs %>%
    # pivot longer to make data easier to work with
    tidyr::pivot_longer(
        cols = dplyr::starts_with("dif"),
        names_to = "Demographic",
        values_to = "Difference"
    ) %>%
    # Add in boolean difference column for coloring purposes
    dplyr::mutate(pos = Difference >= 0) %>%
    # Generate plot, reordering columns by magnitude of difference value
    ggplot(
        aes(x = reorder(Demographic, -abs(Difference)), y = Difference, fill = pos)
    ) +
    geom_bar(stat = "identity") +
    # Create horizontal line at x-axis for readability of visual
    geom_abline(slope=0, intercept=0, col="black", lty = 2) +
    labs(title = "Differences Between College and Geographic Demographic Representation",
         subtitle = "Four-Year Institutions") +
    ylab("Difference (%)") +
    xlab("Demographic") +
    scale_x_discrete(labels = c("Hispanic", "White", "Multiracial", "Pacif. Isl.",
                                "Amer. Indian", "Asian", "Black")
                     ) +
```

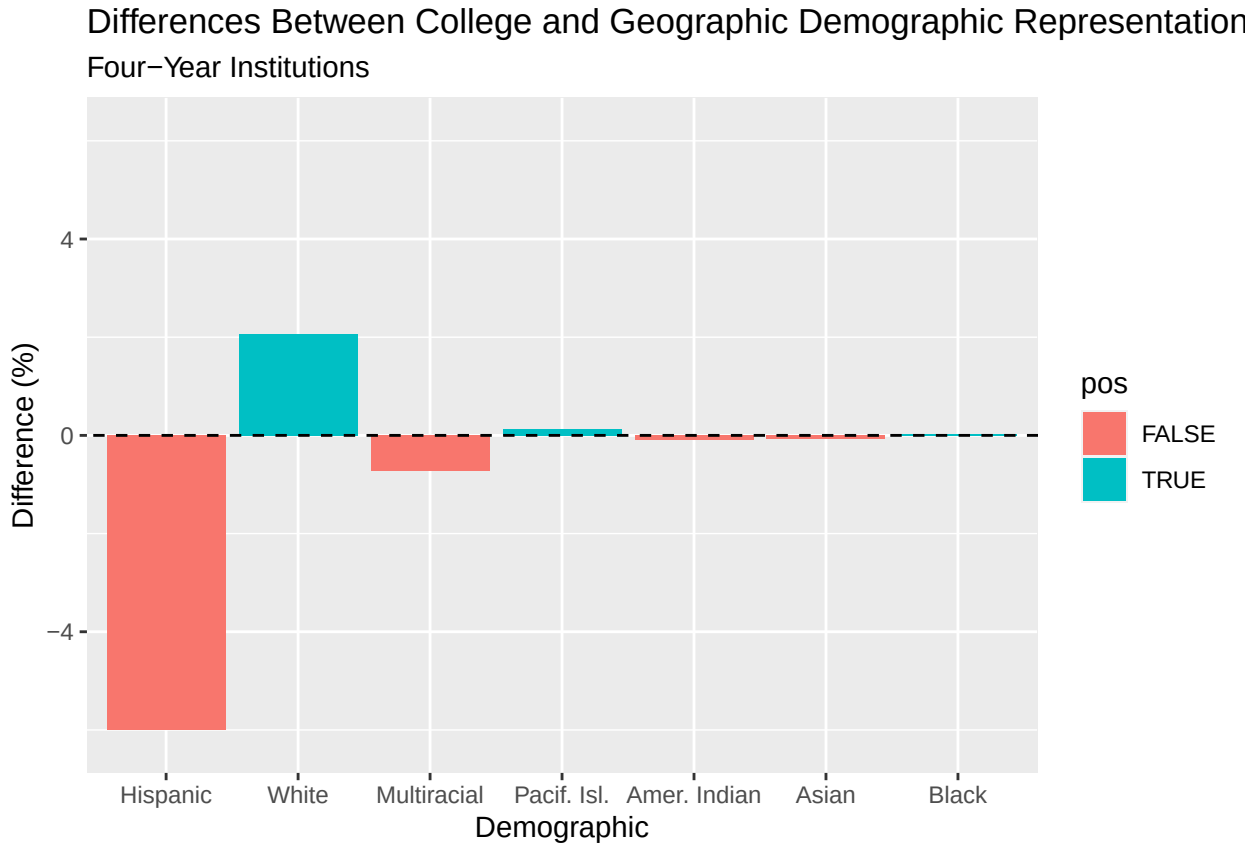
```
    ylim(c(-6.25,6.25))

# year-two difference plot ---------------------
q1_y2_plot <- average_two_year_diffs %>%
    # pivot longer to make data easier to work with
    tidyr::pivot_longer(
        cols = dplyr::starts_with("dif"),
        names_to = "Demographic",
        values_to = "Difference"
    ) %>%
    # Add in boolean difference column for coloring purposes
    dplyr::mutate(pos = Difference >= 0) %>%
    # Generate plot, reordering columns by magnitude of difference value
    ggplot(
        aes(x = reorder(Demographic, -abs(Difference)), y = Difference, fill = pos)
    ) +
    geom_bar(stat = "identity") +
    # Create horizontal line at x-axis for readability of visual
    geom_abline(slope=0, intercept=0, col="black", lty = 2) +
    labs(title = "Differences Between College* and Geographic Demographic Representation",
         subtitle = "*Two-Year Institutions") +
    ylab("Difference (%)") +
    xlab("Demographic") +
    ylim(c(-8.5,8.5))
```

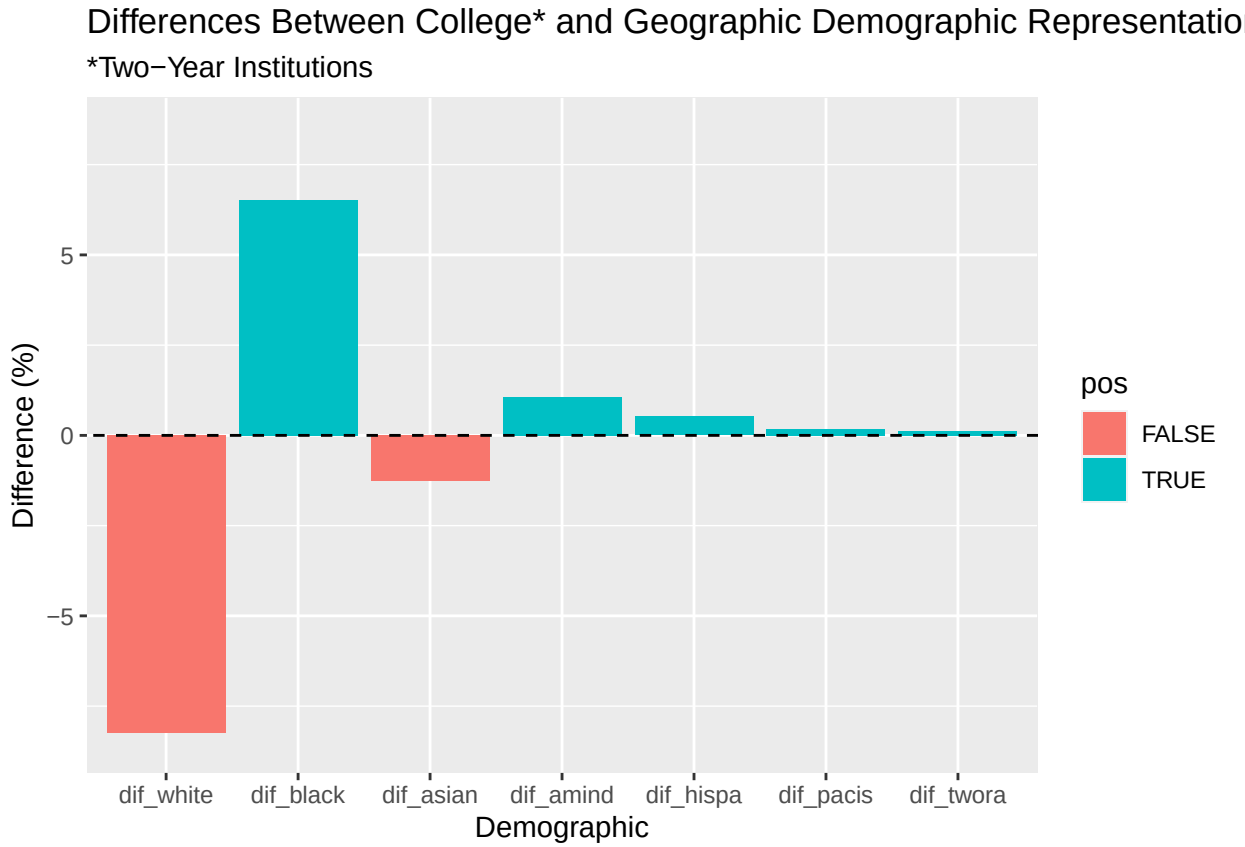
```
    # scale_x_discrete(labels = c("Hispanic", "White", "Multiracial", "Pacif. Isl.",
    #                             "Amer. Indian", "Asian", "Black")
    #                  ) +
```

Output

Table 1: College versus Geographic Demographic Differences (Percent)

Institution Type	White	Hispanic	Black	Asian	American Indian	Pacific Islander	Multiracial
Four Year	2.069017	-6.0057497	0.0310581	-0.0728726	-0.0896958	0.1319228	-0.7278570
Two Year	-8.230215	0.5250744	6.5192186	-1.2671895	1.0679816	0.1675113	0.1160342


Differences Between College and Geographic Demographic Representation
Four−Year Institutions
Difference (%)
4
0
−4
Hispanic
White
Multiracial
Pacif. Isl.
Amer. Indian
Asian
Black
Demographic
pos
FALSE
TRUE


Differences Between College* and Geographic Demographic Representatio
*Two−Year Institutions
Difference (%)
5
0
−5
dif_white
dif_black
dif_asian
dif_amind
dif_hispa
dif_pacis
dif_twora
Demographic
pos
FALSE
TRUE