ND Skills Assessment

AUTHOR
Cristian T

Assumptions to Document

1. IPEDS Data Accuracy and Completeness

- It was assumed that the IPEDS Completions and Institutional Characteristics data are accurate, consistent across years, and comparable between institutions. Any discrepancies in reporting by institutions were considered to be minimal or non-systematic.

2. CIP Code Stability and Mapping

- Changes in CIP codes over time were handled using the 2020 CIP crosswalk. It was assumed that redirected or updated CIP codes retained the same general disciplinary meaning and that mapped fields accurately represented consistent academic domains.

3. Nonresident Doctorate Definition

- The `cnralt` variable (nonresident alien recipients) was assumed to represent all international students who were on student visas at the time of degree completion. It was believed that all institutions applied this classification consistently.

4. Representative Year Selection

- The periods 2011–2013 and 2021–2023 were assumed to be representative of long-term trends, despite potential short-term fluctuations.

5. Uniform Award Level

- Only award level 17 (Doctorate – Research/Scholarship) was included. It was assumed that this category consistently reflects research-based doctoral degrees (e.g., PhD, EdD) and excludes professional doctorates (e.g., MD, JD).

6. Institutional Consistency

- It was assumed that the same set of 31 AAU private institutions operated with similar definitions and classifications across all selected years. Mergers, name changes, or program reclassifications were assumed not to affect comparability.

7. Nonresident vs. Race/Ethnicity Overlap

- Since IPEDS does not disaggregate race/ethnicity by residency status, it was assumed that overlap exists (e.g., Asian or Hispanic nonresident students may also be counted under those racial categories). This limitation was acknowledged in interpreting demographic data.

Static Institutional Control and Classification

- Institutions were assumed to have remained private and AAU-affiliated throughout the study period. Any institutional changes in control or AAU membership were not expected to affect the validity of the comparison set.

The Association of American Universities (AAU) includes 31 private research institutions that are widely recognized for their academic excellence and global impact. This analysis focuses on the University of Notre Dame. It examines how its patterns of awarding doctorates to U.S. nonresidents compare to those of its AAU private peers.

The primary goal of this report is to evaluate the percentage of doctorates awarded to nonresident aliens at Notre Dame and assess how that share has evolved over the past decade. By placing Notre Dame's trends in the context of peer institutions, we aim to identify whether Notre Dame has expanded or narrowed its international doctoral footprint, and how it aligns with broader patterns in elite private research universities.

To do this, we analyze data from the Integrated Postsecondary Education Data System (IPEDS) Completions files for the years 2011–2013 and 2021–2023, focusing specifically on award level 17 (Doctorate - Research/Scholarship). The analysis includes institution-level summaries, outlier detection, discipline-level comparisons, and year-over-year trends.

This work supports institutional benchmarking and strategic planning by providing insight into how Notre Dame's doctoral education profile compares nationally within a defined set of peer institutions.

Setup

I approached this using a structured data science workflow: data acquisition, cleaning, transformation, analysis, and communication. To ensure precise control, I downloaded data directly from the IPEDS Data Center, selecting Completions data files ending in _A for the years 2011–2013 and 2021–2023. I referenced the IPEDS data dictionaries and used the CIP 2020 crosswalk to resolve legacy codes.

Exploratory Data Analysis (EDA)

Summary

```r
# Extract award level 17 (IPEDS system defined as a Doctorate — Research/Scholarship),
#     filter to AAU schools, group by institution, and calculate the summary stats by
#     year.
process_yearly_data <- function(df, year, aau_ids) {
  df %>%
    filter(awlevel == 17, unitid %in% aau_ids) %>%
    group_by(unitid) %>%
    summarise(
      year = year,
      total_docs = sum(ctotalt, na.rm = TRUE),
      nonresident = sum(cnralt, na.rm = TRUE),
```

```r
    male_docs = sum(ctotalm, na.rm = TRUE),
    female_docs = sum(ctotalw, na.rm = TRUE),
    pct_nonresident = round(nonresident / total_docs * 100, 1),
    pct_male = round(male_docs / total_docs * 100, 1),
    pct_female = round(female_docs / total_docs * 100, 1),
    pct_asian = round(sum(casiat, na.rm = TRUE) / total_docs * 100, 1),
    pct_white = round(sum(cwhitt, na.rm = TRUE) / total_docs * 100, 1),
    pct_black = round(sum(cbkaat, na.rm = TRUE) / total_docs * 100, 1),
    pct_hispanic = round(sum(chispt, na.rm = TRUE) / total_docs * 100, 1),
    .groups = "drop"
  ) %>%
    filter(total_docs > 0)
}

# Apply function to all years
aau_ids <- aau_private$unitid

summary_2011 <- process_yearly_data(c2011, 2011, aau_ids)
summary_2012 <- process_yearly_data(c2012, 2012, aau_ids)
summary_2013 <- process_yearly_data(c2013, 2013, aau_ids)
summary_2021 <- process_yearly_data(c2021, 2021, aau_ids)
summary_2022 <- process_yearly_data(c2022, 2022, aau_ids)
summary_2023 <- process_yearly_data(c2023, 2023, aau_ids)

# Combine all summaries
summary_all_years <- bind_rows(
  summary_2011,
  summary_2012,
  summary_2013,
  summary_2021,
  summary_2022,
  summary_2023
)

# Add Institution Names
summary_all_named <- summary_all_years %>% left_join(hd2023 %>% select(unitid, instnm),
        by = "unitid")

# Create short aliases for long institution names
summary_all_named <- summary_all_named %>%
  mutate(alias = case_when(
    instnm == "Boston University" ~ "BostonU",
    instnm == "Brandeis University" ~ "Brandeis",
    instnm == "Brown University" ~ "Brown",
    instnm == "California Institute of Technology" ~ "Caltech",
    instnm == "Carnegie Mellon University" ~ "Carnegie Mellon",
    instnm == "Case Western Reserve University" ~ "Case Western",
    instnm == "Columbia University in the City of New York" ~ "Columbia",
    instnm == "Cornell University" ~ "Cornell",
    instnm == "Dartmouth College" ~ "Dartmouth",
    instnm == "Duke University" ~ "Duke",
```

```
    instnm == "Emory University" ~ "Emory",
    instnm == "George Washington University" ~ "GWU",
    instnm == "Harvard University" ~ "Harvard",
    instnm == "Johns Hopkins University" ~ "Johns Hopkins",
    instnm == "Massachusetts Institute of Technology" ~ "MIT",
    instnm == "New York University" ~ "NYU",
    instnm == "Northwestern University" ~ "Northwestern",
    instnm == "Princeton University" ~ "Princeton",
    instnm == "Rice University" ~ "Rice",
    instnm == "Stanford University" ~ "Stanford",
    instnm == "Tufts University" ~ "Tufts",
    instnm == "Tulane University of Louisiana" ~ "Tulane",
    instnm == "University of Chicago" ~ "Chicago",
    instnm == "University of Miami" ~ "Miami",
    instnm == "University of Notre Dame" ~ "Notre Dame",
    instnm == "University of Pennsylvania" ~ "Penn",
    instnm == "University of Rochester" ~ "Rochester",
    instnm == "University of Southern California" ~ "USC",
    instnm == "Vanderbilt University" ~ "Vanderbilt",
    instnm == "Washington University in St Louis" ~ "WashU",
    instnm == "Yale University" ~ "Yale",
    TRUE ~ instnm
  ))

# Select final columns for output
summary_all_named <- summary_all_named %>%
  select(instnm, alias, year, total_docs, nonresident, male_docs, female_docs,
         pct_nonresident, pct_male, pct_female, pct_asian, pct_white, pct_black,
         pct_hispanic)

summary_all_named
```

instnm	alias	year
<chr>	<chr>	<dbl>
California Institute of Technology	Caltech	2011
University of Southern California	USC	2011
Yale University	Yale	2011
George Washington University	GWU	2011
University of Miami	Miami	2011
Emory University	Emory	2011
University of Chicago	Chicago	2011
Northwestern University	Northwestern	2011
University of Notre Dame	Notre Dame	2011
Tulane University of Louisiana	Tulane	2011

1-10 of 186 rows | 1-3 of 14 columns   Previous **1** 2 3 4 5 6 … 19 Next

```
# write_csv(summary_all_named, "/Users/cristian/Downloads/ND
        Data/summary_all_named.csv")
```

As part of the institutional summaries, we calculated the percentage of doctorates awarded to men, women, and individuals from major racial and ethnic groups (Asian, White, Black, and Hispanic) using variables such as `ctotalm`, `ctotalw`, `casiat`, `cwhitt`, `cbkaat`, and `chispt`. These variables reflect total awards conferred regardless of citizenship status.

It is important to note that IPEDS does not disaggregate gender or race/ethnicity data by residency status. In other words, the counts of male/female and racial/ethnic categories include both U.S. residents and nonresident aliens. As a result, we cannot isolate demographic percentages specifically for U.S. citizens or for international students. While we report these figures for descriptive context, they should not be interpreted as mutually exclusive from the nonresident category (`cnralt`), which is tracked separately.

This limitation highlights an important consideration when interpreting equity and inclusion metrics from IPEDS Completions data: overlap may exist between the race/ethnicity and nonresident fields, particularly for students who identify as Asian or Hispanic but hold non-U.S. citizenship.
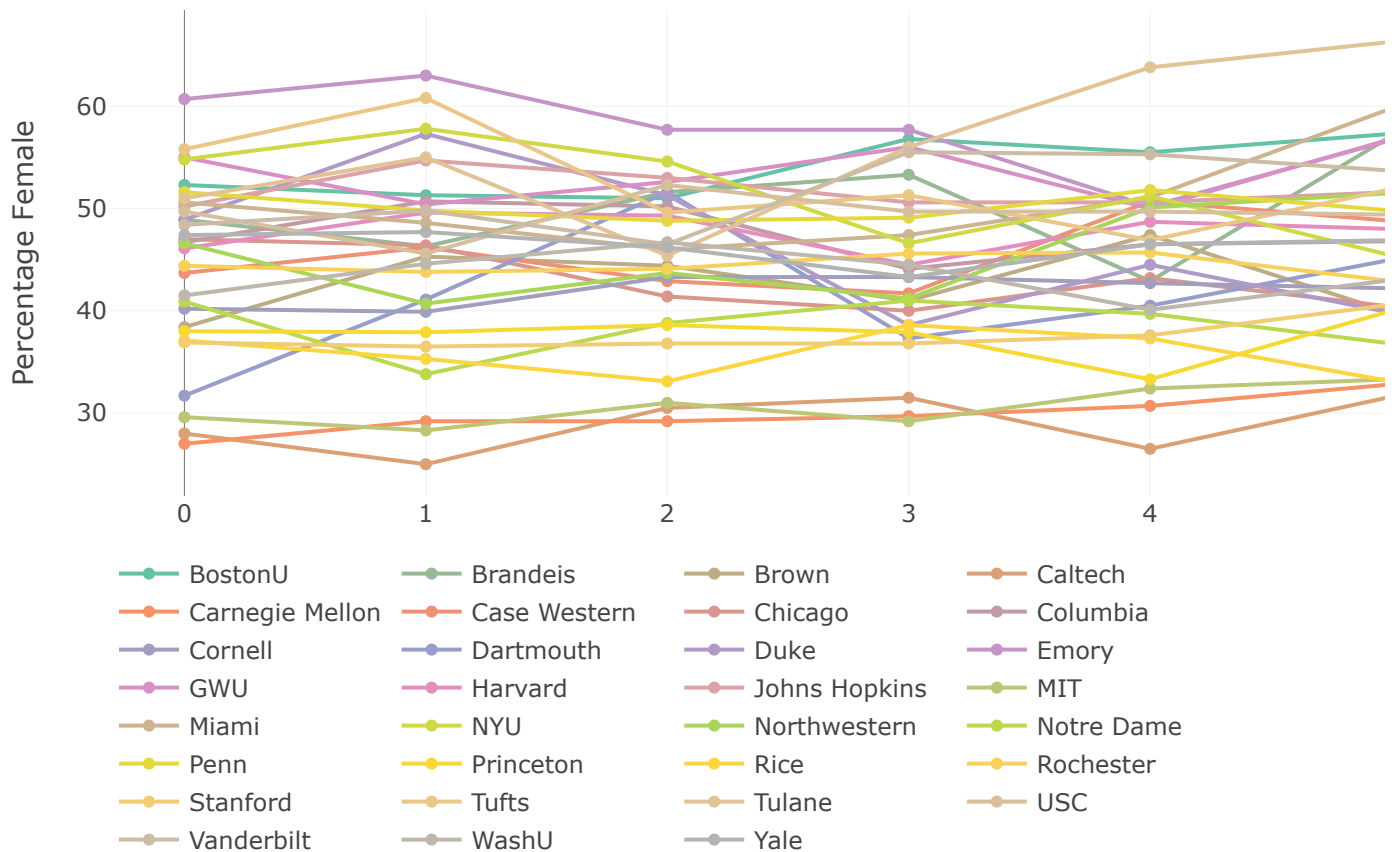
Gender

```
# Convert male/female percent columns to long format
# Select institution name, year, and gender percentage columns
# Create gender and percent column
# Rename variables
gender_long <- summary_all_named %>%
  select(alias, year, pct_male, pct_female) %>%
  pivot_longer(cols = starts_with("pct_"), names_to = "gender", values_to = "percent")
        %>%
  mutate(gender = recode(gender, pct_male = "Male", pct_female = "Female"))

# Use the Plotly charts with wide format
gender_wide <- summary_all_named %>%
  select(alias, year, pct_male, pct_female)

# Plot line chart for percentage female
plot_ly(gender_wide,
        y = ~pct_female,
        color = ~alias,
        type = 'scatter',
        mode = 'lines+markers',
        text = ~paste(alias, "<br>Percentage Female:", pct_female),
        hoverinfo = 'text') %>%
  layout(title = "Percentage Female Doctorate Recipients (2011–2023)",
          caption = "Source: IPEDS Completions Data, NCES",
          xaxis = list(),
          yaxis = list(title = "Percentage Female"),
          legend = list(orientation = 'h'))
```
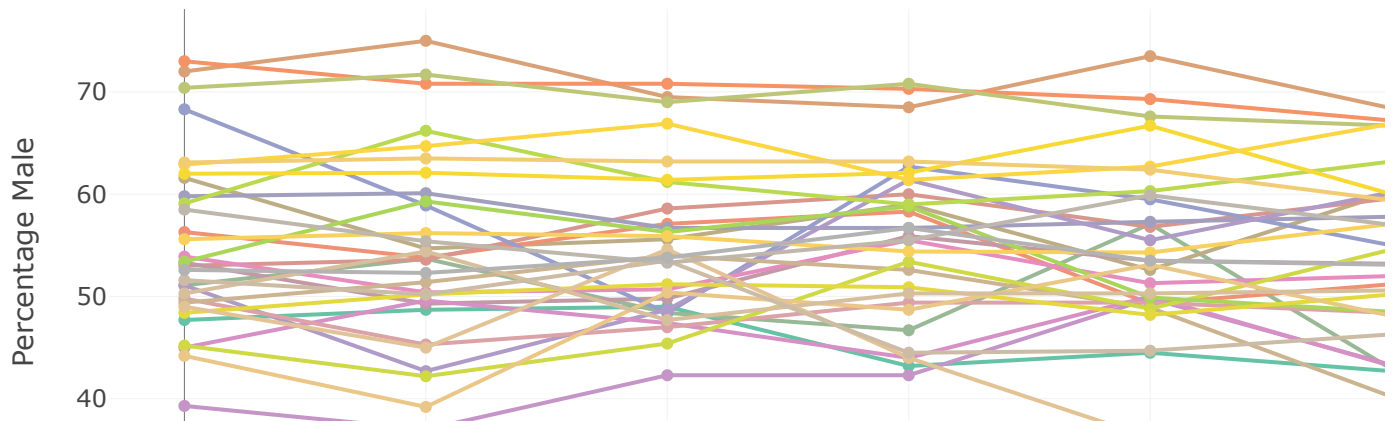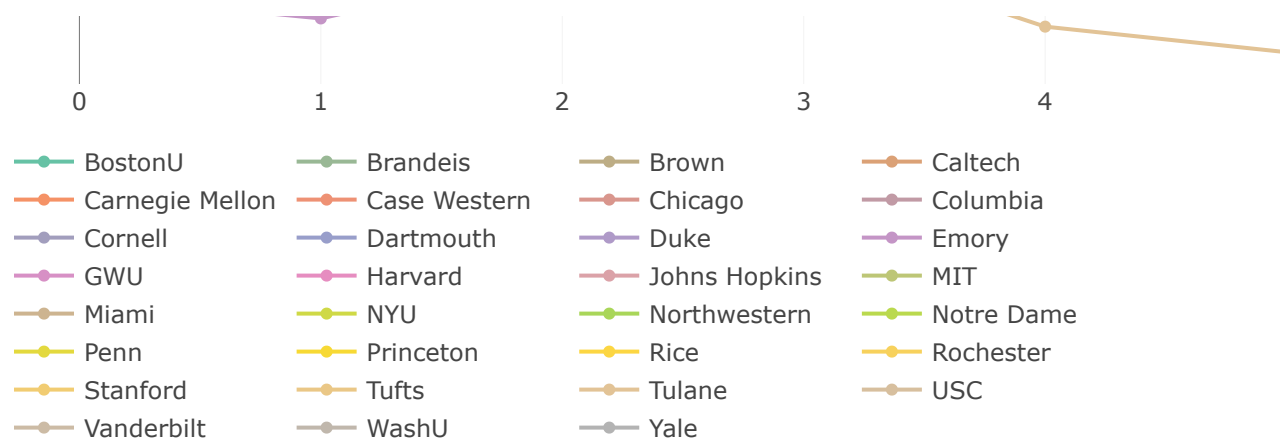
Percentage Female Doctorate Recipients (2011–2023)

Percentage Female

```r
# Plot interactive line chart for percentage male
plot_ly(gender_wide,
        y = ~pct_male,
        color = ~alias,
        type = 'scatter',
        mode = 'lines+markers',
        text = ~paste(alias, "<br>Percentage Male:", pct_male),
        hoverinfo = 'text') %>%
  layout(title = "Percentage Male Doctorate Recipients (2011–2023)",
         caption = "Source: IPEDS Completions Data, NCES",
         xaxis = list(),
         yaxis = list(title = "Percentage Male"),
         legend = list(orientation = 'h'))
```

Percentage Male Doctorate Recipients (2011–2023)

Percentage Male

Legend:
- BostonU
- Brandeis
- Brown
- Caltech
- Carnegie Mellon
- Case Western
- Chicago
- Columbia
- Cornell
- Dartmouth
- Duke
- Emory
- GWU
- Harvard
- Johns Hopkins
- MIT
- Miami
- NYU
- Northwestern
- Notre Dame
- Penn
- Princeton
- Rice
- Rochester
- Stanford
- Tufts
- Tulane
- USC
- Vanderbilt
- WashU
- Yale

```
# write_csv(gender_wide, "/Users/cristian/Downloads/ND Data/gender_wide.csv")
```

We examined the gender distribution, focusing on the percentage of degrees awarded to male and female students, with particular attention to whether institutions maintain a gender balance and how this has shifted over time.

Several universities consistently awarded an approximately equal share of doctorates to men and women. These include Boston University, George Washington University, New York University, Tufts University, and Brandeis University. A few institutions have awarded more than 50% of doctorates to women in recent years. Notably, Emory University was the highest institution, with 63.8% of doctorates awarded to women in 2022. Tulane University's representation increased from 51% in 2011 to 66.3% in 2023, and Boston University and George Washington University also consistently showed strong female representation.

On the other hand, several institutions consistently showed low female representation (<40%) throughout the period, with the California Institute of Technology averaging around 70–75% male representation. The Massachusetts Institute of Technology and Carnegie Mellon University consistently showed a male-to-female ratio of over 65%. Stanford, Rice, and Princeton also maintained male dominance above 60%.

We did observe that some universities demonstrated measurable gains in gender equity, with Tulane showing an increase of 15.3 percent from 2011 to 2023. Vanderbilt showed a 5.3 percent increase, and Boston University showed a 5 percent increase. Stanford shows a modest increase of 3.7 percent. While institutions like Harvard, Yale, and Columbia have remained relatively flat near gender parity, showing little movement over the six years examined.

Specifically looking at Notre Dame, they consistently awarded a higher proportion of doctorates to male students than to female students across all years observed (2011–2023). The percentage of female doctorate recipients at Notre Dame ranged from a low of 33.8% in 2012 to a high of 41.0% in 2011, never reaching gender parity. Notre Dame's downward trend in female doctoral representation contrasts with that of many of its peers, suggesting an opportunity for further investigation and potential intervention in specific programs or departments. Compared to many of its AAU private peers, Notre Dame remained below the AAU private median in most years. Suggesting room for improvement in achieving gender equity among doctoral recipients, particularly when compared to peer institutions that exhibit more balanced outcomes.

Race and Ethnicity

```r
# Reshape race/ethnicity data from wide to long format
# Select institution name, year, and race/ethnicity percentage columns
# Consolidate race percentage columns
# Rename variables
race_long <- summary_all_named %>%
  select(alias, year, pct_asian, pct_white, pct_black, pct_hispanic) %>%
  pivot_longer(cols = starts_with("pct_"), names_to = "race", values_to = "percent")
        %>%
  mutate(race = recode(race,
                  pct_asian = "Asian",
                  pct_white = "White",
                  pct_black = "Black",
                  pct_hispanic = "Hispanic"))

# Calculate average percentages by race and year
# Plot the results
race_long %>%
  group_by(year, race) %>%
  summarise(mean_percent = mean(percent, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = year, y = mean_percent, color = race)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  labs(title = "Average Percentage of Doctorates by Race/Ethnicity (2011–2023)",
    x = "",
    y = "Average Percentage of Doctorates",
    caption = "Source: IPEDS Completions Data, NCES",
    color = "Race/Ethnicity"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "right"
  )
```

Average Percentage of Doctorates by Race/Ethnicity (2011–2023)
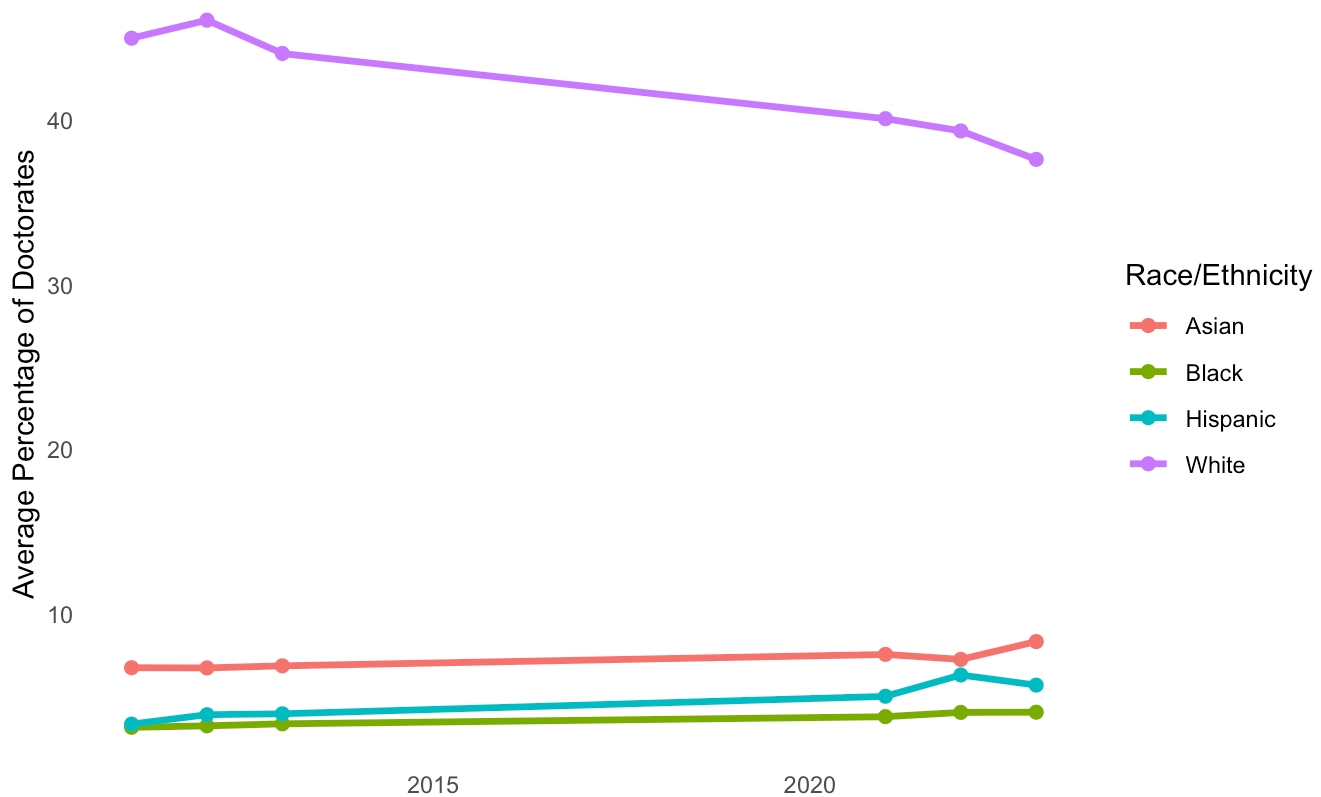
Source: IPEDS Completions Data, NCES

```
# write_csv(race_long, "/Users/cristian/Downloads/ND Data/race_long.csv")
```

White students consistently received the majority of doctorates, averaging between 55 and 60% of all awards throughout the period. However, their share has shown a gradual decline, suggesting modest diversification across institutions. Asian students have maintained a stable and notable presence, accounting for approximately 10–15% of doctorate recipients each year. Their representation remained relatively flat, with no significant upward or downward trend. Hispanic students showed a slight upward trend, increasing from approximately 5% in 2011 to around 7–8% by 2023. While modest, this increase indicates some progress in the representation of Hispanic students in doctoral education. Black students remained consistently underrepresented, with average percentages hovering around 4–6% across the years. There was little change over time, underscoring persistent equity gaps for Black doctoral candidates.

Despite some positive signs, such as incremental gains among Hispanic students, the data reveal that racial and ethnic diversity among doctorate recipients remains limited across AAU private institutions. The dominant representation of White and Asian students, particularly in STEM-heavy universities, continues to shape the overall demographic profile. Moreover, the stagnant percentages for Black students suggest that systemic barriers persist and that institutional efforts may not yet be translating into significant changes at the doctoral level.

```
# Using the same data, filter to Notre Dame
# Calculate average percentages by race and year
```

```r
notre_dame_race <- summary_all_named %>%
  filter(instnm == "University of Notre Dame") %>%
  select(year, pct_asian, pct_white, pct_black, pct_hispanic) %>%
  pivot_longer(cols = starts_with("pct_"),
               names_to = "race",
               values_to = "percent") %>%
  mutate(
    race = recode(race,
                  pct_asian = "Asian",
                  pct_white = "White",
                  pct_black = "Black",
                  pct_hispanic = "Hispanic")
)

Plot the results
ggplot(notre_dame_race, aes(x = year, y = percent, color = race)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  labs(
    title = "Notre Dame Doctorates by Race/Ethnicity (2011–2023)",
    x = "Year",
    y = "Percentage of Doctorates",
    caption = "Source: IPEDS Completions Data, NCES",
    color = "Race/Ethnicity"
) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "right"
)
```

# Notre Dame Doctorates by Race/Ethnicity (2011–2023)
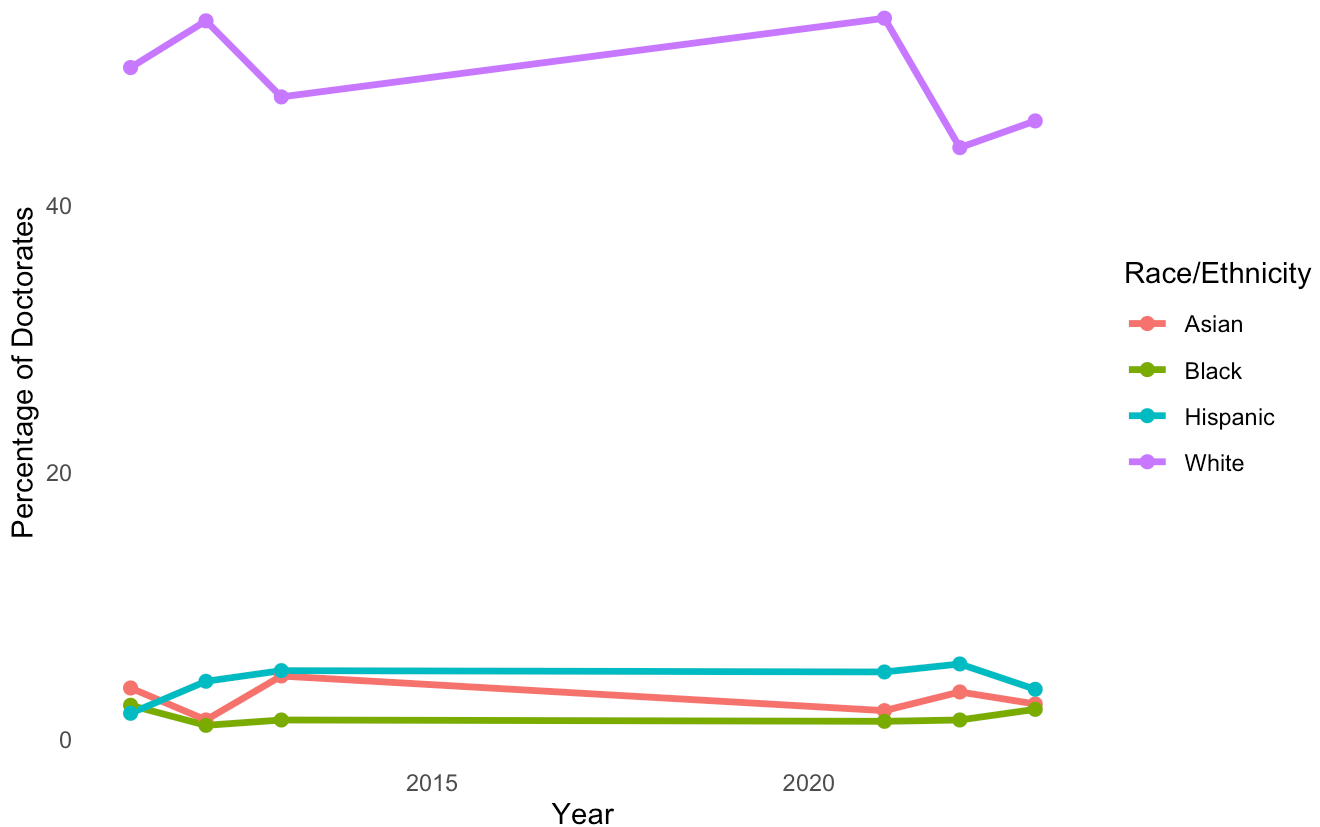


Source: IPEDS Completions Data, NCES

Between 2011 and 2023, the racial and ethnic composition of doctorate recipients at the University of Notre Dame remained predominantly White, with White students consistently comprising over 44% of recipients each year. However, there were modest shifts during this period. Asian students represented between 1.4% and 4.7% of recipients, peaking in 2013 and reaching their lowest point in 2012. Hispanic representation fluctuated between 1.9% and 5.6%, with the highest percentage observed in 2022. While this marks some progress, the trend was not consistently upward. Black students remained a small share of doctorate recipients, ranging from 1.0% to 2.5% across the observed years.

These figures underscore a limited increase in the representation of underrepresented minority groups (Black and Hispanic). However, the modest rise in Hispanic representation in recent years is notable. Additionally, the proportion of White doctorate recipients declined slightly from 54% in 2021 to 46.3% in 2023, which may signal a gradual shift toward greater demographic diversity.

## Identifying Outliers in Non-resident Doctorate Awards

```
# Identify outliers using z-score
# Calculate Z-score for pct_nonresident per year
# Flag an outlier if absolute Z-score > 3
summary_all_named <- summary_all_named %>%
  group_by(year) %>%
  mutate(z_nonresident = scale(pct_nonresident),
    is_outlier_nonresident = abs(z_nonresident) > 3
```

```
  ) %>%
  ungroup()

# View flagged outliers
summary_all_named %>%
  filter(is_outlier_nonresident) %>%
  select(year, instnm, pct_nonresident, z_nonresident)
```

year	instnm	pct_nonresident	z_nonresident
<dbl>	<chr>	<dbl>	<dbl[,1]>
2011	Duke University	3	-3.011539

1 row

```
# Identify outliers using IQR method
# Calculate quartiles for pct_nonresident
# Define lower and upper bounds for outliers
# Filter institutions falling outside the IQR boundaries
summary_outliers <- summary_all_named %>%
  group_by(year) %>%
  mutate(
    Q1 = quantile(pct_nonresident, 0.25, na.rm = TRUE),
    Q3 = quantile(pct_nonresident, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1,
    lower_bound = Q1 - 1.5 * IQR,
    upper_bound = Q3 + 1.5 * IQR
  ) %>%
  filter(pct_nonresident < lower_bound | pct_nonresident > upper_bound) %>%
  select(year, instnm, alias, pct_nonresident, lower_bound, upper_bound) %>%
  arrange(year, pct_nonresident)

# View Box Plots to identify outliers
# Add red points for outliers
# Label outlier points with institution name
ggplot(summary_all_named, aes(x = factor(year), y = pct_nonresident)) +
  geom_boxplot(fill = "#69b3a2", outlier.shape = NA, width = 0.5) +
  geom_point(
    data = summary_outliers,
    aes(x = factor(year), y = pct_nonresident),
    color = "red",
    size = 2
  ) +
  geom_text_repel(
    data = summary_outliers,
    aes(x = factor(year), y = pct_nonresident, label = alias),
    nudge_y = 2,
    direction = "y",
    hjust = 0.5,
    size = 3.5,
```

```
      segment.color = NA,
      box.padding = 0.25,
      point.padding = 0.3,
      max.overlaps = Inf
    ) +
    labs(
      title = "Outliers in Percentage of Nonresident Doctorates by Year",
      subtitle = "Private AAU Institutions, Award Level 17",
      x = NULL,
      y = "Percentage Nonresident Doctorates",
      caption = "Source: IPEDS Completions Data, NCES",
    ) +
    theme_minimal(base_size = 12) +
    theme(
      plot.title = element_text(face = "bold", size = 14),
     plot.subtitle = element_text(size = 10, margin = margin(b = 10)),
      axis.title.x = element_text(margin = margin(t = 10)),
      axis.title.y = element_text(margin = margin(r = 10)),
      panel.grid = element_blank(),
      panel.border = element_blank()
    )
```

Outliers in Percentage of Nonresident Doctorates by Year

Private AAU Institutions, Award Level 17
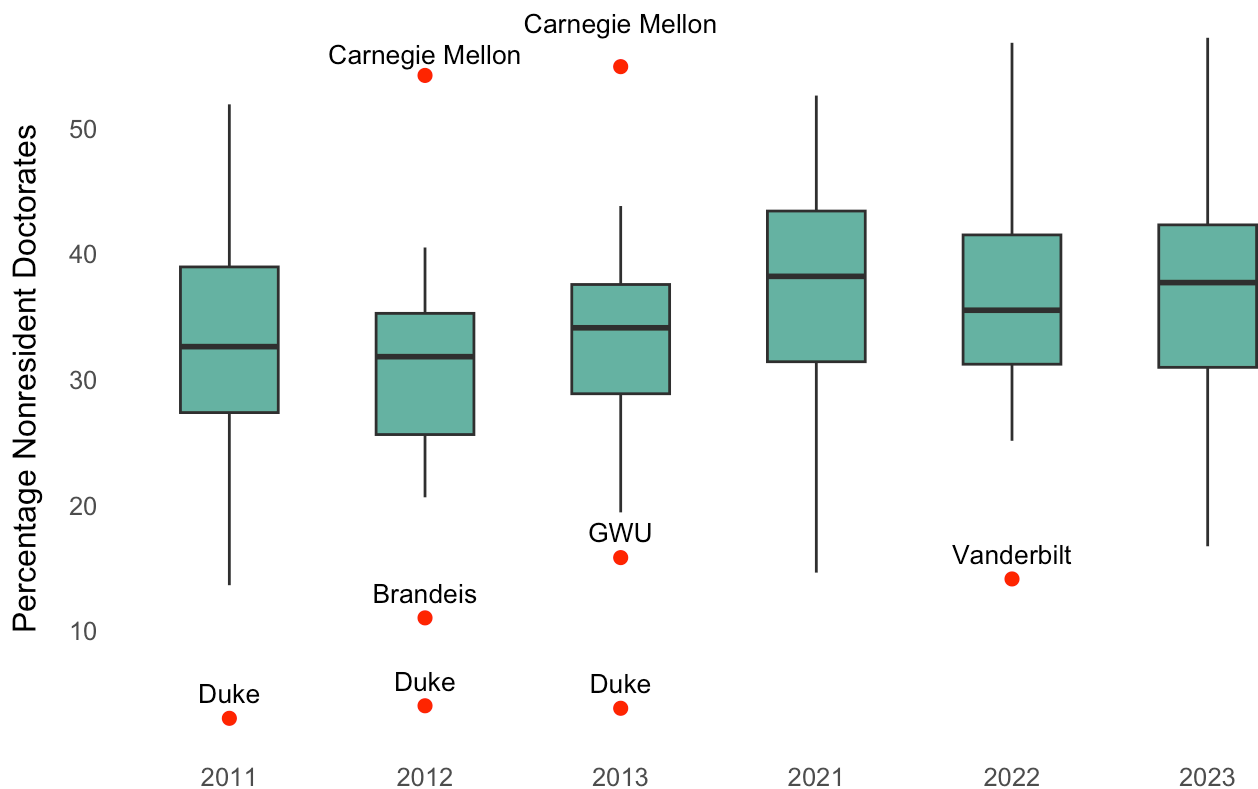
Source: IPEDS Completions Data, NCES

To identify institutions with unusually high or low percentages of doctorates awarded to non-resident students, I examined the data using both an IQR-based and z-score-based outlier detection method

across the years 2011-2013 and 2021-2023.

We were able to see that Duke University was a consistent low-end outlier in the early period, with just 3% of its doctorates awarded to non-residents in 2011 and 4% in 2012. These values fell significantly below the distribution of peer institutions and were flagged by both IQR and z-score methods. Notably, Duke's 2011 value yielded a z-score of −3.01, more than three standard deviations below the mean. No other institutions exceeded this z-score threshold. Carnegie Mellon University was a high-end outlier in 2012 and 2013, with non-resident doctorate shares exceeding 54%, well above the upper bounds of the distribution in those years. Brandeis University, George Washington University (2013), and Vanderbilt University (2022) were also flagged for values on or just beyond the IQR thresholds, suggesting marginal but notable deviations from peer norms.

Classification of Doctoral Degrees

```r
# Clean CIP Lookup Table
# Remove formatting and keep relevant columns
cip_lookup_clean <- cip_lookup %>%
  mutate(
    cipcode = str_replace_all(cip_code2020, '="|"', ""),
    cip_title = cip_title2020
  ) %>%
  select(cipcode, cip_title)

# Create redirection map from 2010 to 2020 CIP codes for matching legacy data
cip_redirect <- cip_lookup %>%
  filter(action == "Moved to") %>%
  mutate(
    from_cip = str_replace_all(cip_code2010, '="|"', ""),
    to_cip = str_replace_all(cip_code2020, '="|"', "")
  ) %>%
  select(from_cip, to_cip)

# Extract award level 17 (IPEDS system defined as a Doctorate - Research/Scholarship),
#        filter to AAU schools, join cip code.
process_year <- function(df, year) {
  df %>%
    filter(awlevel == 17, unitid %in% aau_private$unitid) %>%
    mutate(
      year = year,
      cipcode = as.character(cipcode)
    ) %>%
    left_join(cip_redirect, by = c("cipcode" = "from_cip")) %>%
    mutate(
      cipcode_final = if_else(!is.na(to_cip), to_cip, cipcode)
    )
}

# Apply function to all years
c2011_clean <- process_year(c2011, 2011)
```

```r
c2012_clean <- process_year(c2012, 2012)
c2013_clean <- process_year(c2013, 2013)
c2021_clean <- process_year(c2021, 2021)
c2022_clean <- process_year(c2022, 2022)
c2023_clean <- process_year(c2023, 2023)

# Combine all summaries
all_detailed <- bind_rows(
  c2011_clean, c2012_clean, c2013_clean,
  c2021_clean, c2022_clean, c2023_clean
)

# Create CIP 2-digit to category title mapping for general categories
cip_2digit_categories <- tibble::tibble(
  cip_2digit = c(
    "01", "03", "04", "05", "09", "10", "11", "12", "13", "14",
    "15", "16", "19", "21", "22", "23", "24", "25", "26", "27",
    "28", "29", "30", "31", "32", "33", "34", "35", "36", "37",
    "38", "39", "40", "41", "42", "43", "44", "45", "46", "47",
    "48", "49", "50", "51", "52", "53", "54", "55", "60", "61"
  ),
  cip_category_title = c(
    "Agriculture & Related Sciences", "Natural Resources & Conservation", "Architecture
        & Related Services", "Ethnic/Cultural/Gender Studies", "Communication &
        Journalism", "Communications Technologies", "Computer & Information Sciences",
        "Culinary/Entertainment/Personal Services", "Education", "Engineering",
        "Engineering Technologies", "Foreign Languages & Linguistics", "Family &
        Consumer Sciences", "Reserved", "Legal Professions", "English Language &
        Literature", "Liberal Arts & Humanities", "Library Science", "Biological &
        Biomedical Sciences", "Mathematics & Statistics", "Military Science", "Military
        Technologies", "Multi/Interdisciplinary Studies", "Parks, Recreation &
        Fitness", "Basic Skills & Remedial Education", "Citizenship Activities",
        "Health-Related Knowledge", "Interpersonal & Social Skills", "Leisure &
        Recreational Activities", "Personal Awareness & Self-Improvement", "Philosophy
        & Religious Studies", "Theology & Religious Vocations", "Physical Sciences",
        "Science Technologies", "Psychology", "Homeland Security & Firefighting",
        "Public Administration & Social Services", "Social Sciences", "Construction
        Trades", "Mechanic & Repair Technologies", "Precision Production",
        "Transportation & Materials Moving", "Visual & Performing Arts", "Health
        Professions", "Business, Management, Marketing", "Secondary Diplomas &
        Certificates", "History", "Reserved", "Health Professions Residency", "Medical
        Residency Programs"))

# Add institution name, cleaned CIP title, and 2-digit category info
all_detailed_named <- all_detailed %>%
  left_join(hd2023 %>% select(unitid, instnm), by = "unitid") %>%
  left_join(cip_lookup_clean, by = c("cipcode_final" = "cipcode")) %>%
  mutate(
    cip_title = if_else(
      is.na(cip_title) & cipcode_final == "99",
      "All Disciplines (Total Doctorates)",
      cip_title
    ),
    cip_category = if_else(
```

```r
      str_detect(cipcode_final, "^99"),
      "Other",
      "Non-Other"
    )
  ) %>%
  mutate(cip_2digit = str_sub(cipcode_final, 1, 2)) %>%
  left_join(cip_2digit_categories, by = "cip_2digit")

# Summarize calculate the summary stats by year.
summary_withCip <- all_detailed_named %>%
  filter(cnralt > 0) %>%
  group_by(instnm, year, cipcode_final, cip_title, cip_category_title) %>%
  summarise(
    total_docs = sum(ctotalt, na.rm = TRUE),
    nonresident = sum(cnralt, na.rm = TRUE),
    pct_nonresident = round(nonresident / total_docs * 100, 1)
  ) %>%
  filter(total_docs > 0)

# Create short aliases for long institution names
summary_withCip <- summary_withCip %>%
  mutate(alias = case_when(
    instnm == "Boston University" ~ "BostonU",
    instnm == "Brandeis University" ~ "Brandeis",
    instnm == "Brown University" ~ "Brown",
    instnm == "California Institute of Technology" ~ "Caltech",
    instnm == "Carnegie Mellon University" ~ "Carnegie Mellon",
    instnm == "Case Western Reserve University" ~ "Case Western",
    instnm == "Columbia University in the City of New York" ~ "Columbia",
    instnm == "Cornell University" ~ "Cornell",
    instnm == "Dartmouth College" ~ "Dartmouth",
    instnm == "Duke University" ~ "Duke",
    instnm == "Emory University" ~ "Emory",
    instnm == "George Washington University" ~ "GWU",
    instnm == "Harvard University" ~ "Harvard",
    instnm == "Johns Hopkins University" ~ "Johns Hopkins",
    instnm == "Massachusetts Institute of Technology" ~ "MIT",
    instnm == "New York University" ~ "NYU",
    instnm == "Northwestern University" ~ "Northwestern",
    instnm == "Princeton University" ~ "Princeton",
    instnm == "Rice University" ~ "Rice",
    instnm == "Stanford University" ~ "Stanford",
    instnm == "Tufts University" ~ "Tufts",
    instnm == "Tulane University of Louisiana" ~ "Tulane",
    instnm == "University of Chicago" ~ "Chicago",
    instnm == "University of Miami" ~ "Miami",
    instnm == "University of Notre Dame" ~ "Notre Dame",
    instnm == "University of Pennsylvania" ~ "Penn",
    instnm == "University of Rochester" ~ "Rochester",
    instnm == "University of Southern California" ~ "USC",
    instnm == "Vanderbilt University" ~ "Vanderbilt",
```

```r
      instnm == "Washington University in St Louis" ~ "WashU",
      instnm == "Yale University" ~ "Yale",
      TRUE ~ instnm
  ))


# Aggregate summary by category
# CIP code "99" represents a roll-up category ("All Disciplines (Total Doctorates)").
# It is relabeled for clarity and excluded from discipline-level summaries to avoid
#        inflating totals or duplicating counts across categories
categories <- summary_withCip %>%
  filter(!is.na(cip_category_title)) %>%
  group_by(cip_category_title) %>%
  summarise(total_docs = sum(total_docs, na.rm = TRUE)) %>%
  arrange(desc(total_docs))

# Total nonresident counts by field per institution and year
summaryYears <- summary_withCip %>%
  filter(!is.na(cip_category_title)) %>%
  group_by(instnm, year, cip_category_title) %>%
  summarise(total_nonresident = sum(nonresident, na.rm = TRUE), .groups = "drop") %>%
  arrange(instnm, year, desc(total_nonresident))

# Top 10 fields by total doctorate count
top_fields <- categories %>%
  filter(!is.na(cip_category_title)) %>%
  group_by(cip_category_title) %>%
  summarise(total_docs = sum(total_docs, na.rm = TRUE), .groups = "drop") %>%
  arrange(desc(total_docs)) %>%
  slice_head(n = 10)

# Plot top doctorates
ggplot(top_fields, aes(x = reorder(cip_category_title, total_docs), y = total_docs)) +
  geom_col(fill = "#0C2340") +
  coord_flip() +
  labs(
    title = "Top Fields by Total Doctorates Awarded (2011–2023)",
    x = NULL,
    y = NULL
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.y = element_text(size = 12),
    axis.text.x = element_text(size = 12),
    plot.title = element_text(face = "bold", size = 12, hjust = 0)
  )
```
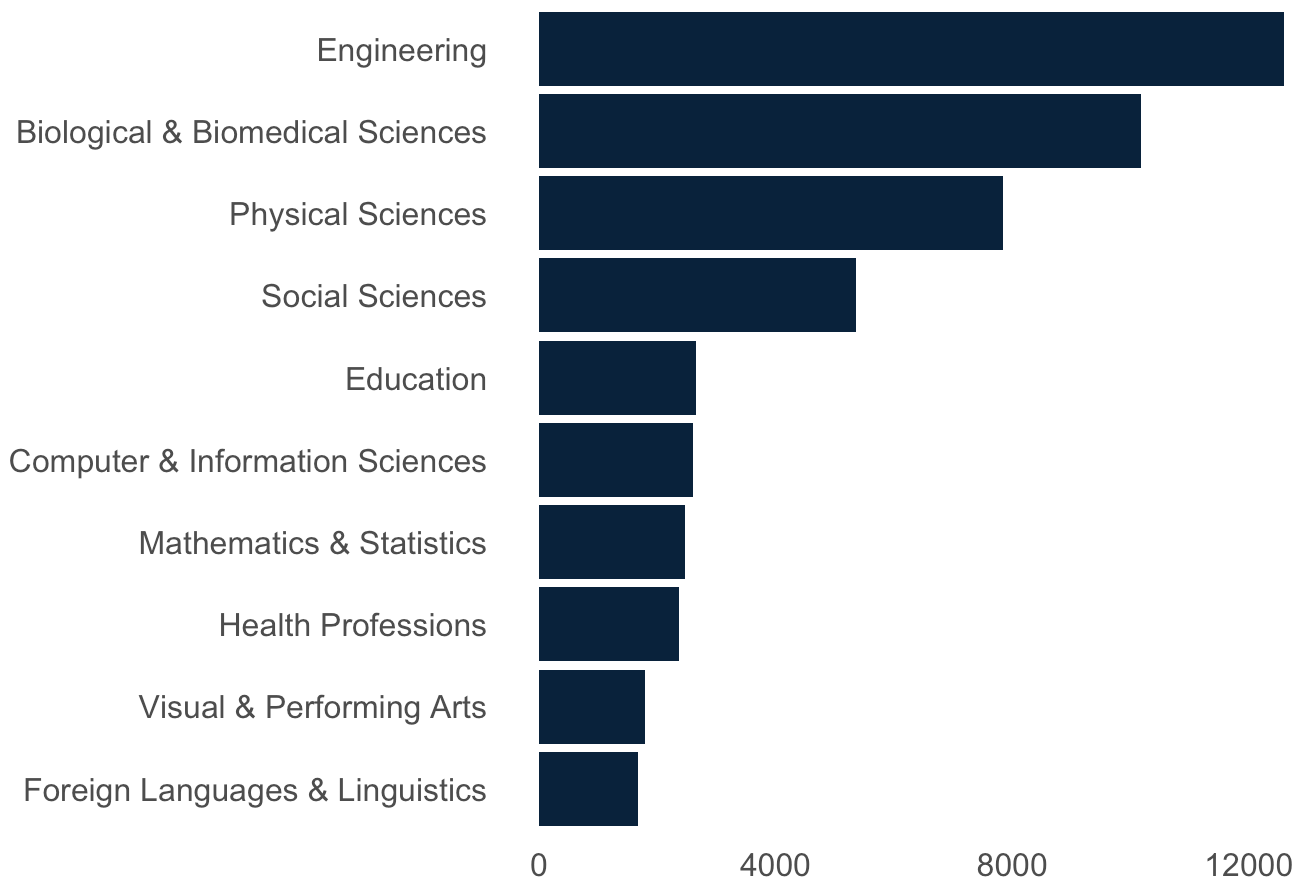
Top Fields by Total Doctorates Awarded (2011–2023)

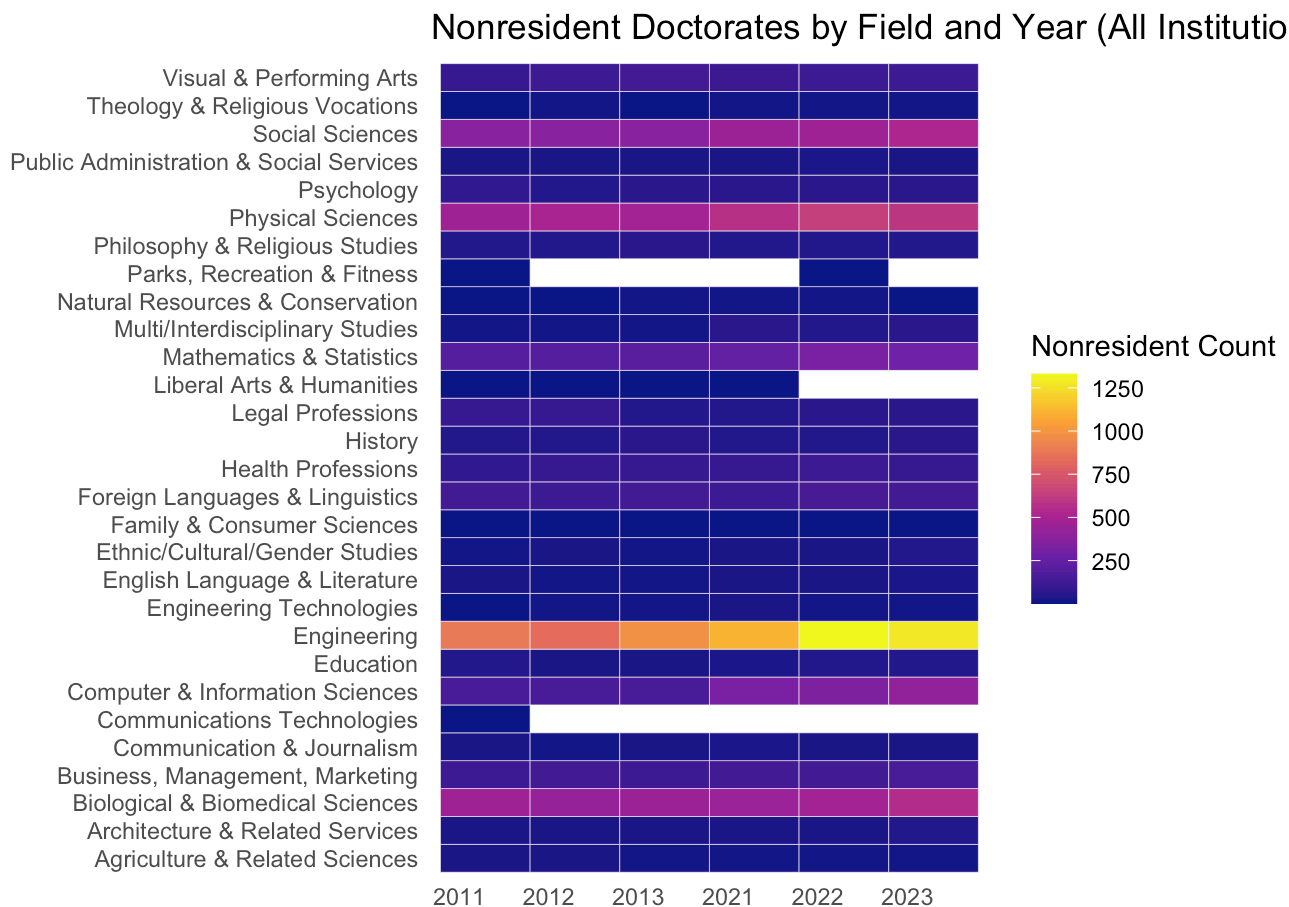| Field	
Engineering	
Biological & Biomedical Sciences	
Physical Sciences	
Social Sciences	
Education	
Computer & Information Sciences	
Mathematics & Statistics	
Health Professions	
Visual & Performing Arts	
Foreign Languages & Linguistics	

(0   4000   8000   12000)

```
# write_csv(summary_withCip, "/Users/cristian/Downloads/ND Data/summary_withCip.csv")
```

To conclude the exploratory data analysis, we examined the fields with the highest volume of doctorate awards from 2011 to 2023 and found that STEM fields dominated in terms of total degrees conferred. Engineering, Biological Sciences, and Physical Sciences collectively accounted for over 30,000 doctoral degrees, underscoring the strong institutional focus on scientific and technological research. The social sciences also emerged as a major contributor. Though not classified as a STEM field, it ranked fourth in total awards, reflecting robust research activity in disciplines such as economics, psychology, sociology, and political science.

Public-facing fields remain prominent, with Education and Health-related disciplines both appearing in the top ten, illustrating the continued importance of training professionals who directly serve societal needs. Additionally, computing and data-driven fields are growing rapidly. Computer & Information Sciences surpassed even Mathematics & Statistics in total awards, highlighting increasing interest in artificial intelligence, machine learning, and data science.

A balance between technical and cultural disciplines was also observed. The inclusion of Visual & Performing Arts, along with Foreign Languages & Linguistics, signals that humanities and creative fields remain vital components of doctoral education at elite private institutions. While STEM continues to lead in volume, these findings show that AAU private universities support a broad and diverse doctoral portfolio that spans applied, theoretical, artistic, and humanistic domains.

```r
# Create a heat map for doctorate categories and years
summary_withCip %>%
  filter(!is.na(cip_category_title)) %>%
  group_by(year, cip_category_title) %>%
  summarise(total_nonresident = sum(nonresident, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = factor(year), y = cip_category_title, fill = total_nonresident)) +
  geom_tile(color = "white") +
  scale_fill_viridis_c(option = "C", name = "Nonresident Count") +
  labs(
    title = "Nonresident Doctorates by Field and Year (All Institutions)",
    x = "",
    y = ""
  ) +
  theme_minimal(base_size = 11) +
  theme(
    axis.text.x = element_text(hjust = 1),
    panel.grid = element_blank()
  )
```


To further understand changes over time, we examined trends in doctoral degree conferrals across fields from 2011 to 2023. Engineering led all fields consistently each year, with the number of nonresident recipients increasing from 884 in 2011 to 1,275 in 2023. Computer & Information Sciences saw particularly sharp growth, more than doubling from 160 to 409 recipients during the same period.

Biological Sciences maintained steady output, ranging between 430 and 548 awards annually, while Physical Sciences increased from 457 to 600 awards. Mathematics & Statistics also showed growth, rising from 189 to 282 nonresident doctorates—aligning with broader trends in data science and quantitative research.

Social Sciences continued to be a strong contributor year over year, with totals ranging from 370 to 518 nonresident awards annually. Business, Management, and Marketing showed modest but consistent growth, increasing from 121 to 144 recipients.

Although smaller in absolute numbers, disciplines such as Visual & Performing Arts, Philosophy & Religious Studies, Foreign Languages & Linguistics, and Theology consistently attracted international scholars. For example, Visual & Performing Arts rose from 96 to 116 recipients, and Foreign Languages peaked at 143 in 2022.

These trends reflect the growing global demand for U.S. doctoral education, particularly in STEM and data-driven fields, while also affirming the enduring international appeal of humanities and social science programs. Overall, the increasing number of doctorate recipients across most fields highlights the expansion of international recruitment and the sustained value placed on doctoral-level research across a broad range of academic disciplines.

```r
# Summarize doctoral degree data for the University of Notre Dame
notre_dame_summary <- summary_withCip %>%
  filter(instnm == "University of Notre Dame") %>%
  summarise(
    total_docs = sum(total_docs, na.rm = TRUE),
    nonres_docs = sum(nonresident, na.rm = TRUE),
    pct_nonres = round(nonres_docs / total_docs * 100, 1)
  )

# Convert summary data into a longer format for plotting
notre_dame_summary_long <- notre_dame_summary %>%
  pivot_longer(cols = c(total_docs, nonres_docs),
               names_to = "type",
               values_to = "count") %>%
  mutate(type = recode(type,
                       total_docs = "Total Doctorates",
                       nonres_docs = "Nonresident Doctorates"))


# Create a bar chart comparing total and nonresident doctorates at Notre Dame
ggplot(notre_dame_summary_long, aes(x = type, y = count, fill = type)) +
  geom_col(width = 0.6, show.legend = FALSE) +
  scale_fill_manual(values = c("#0c2340", "#c99700")) +
  labs(
    title = "Doctorates Awarded at Notre Dame (2011–2023)",
    subtitle = paste0("Nonresident students accounted for ",
        notre_dame_summary$pct_nonres, "% of all doctorates"),
    x = NULL,
    y = "Number of Doctorates"
```
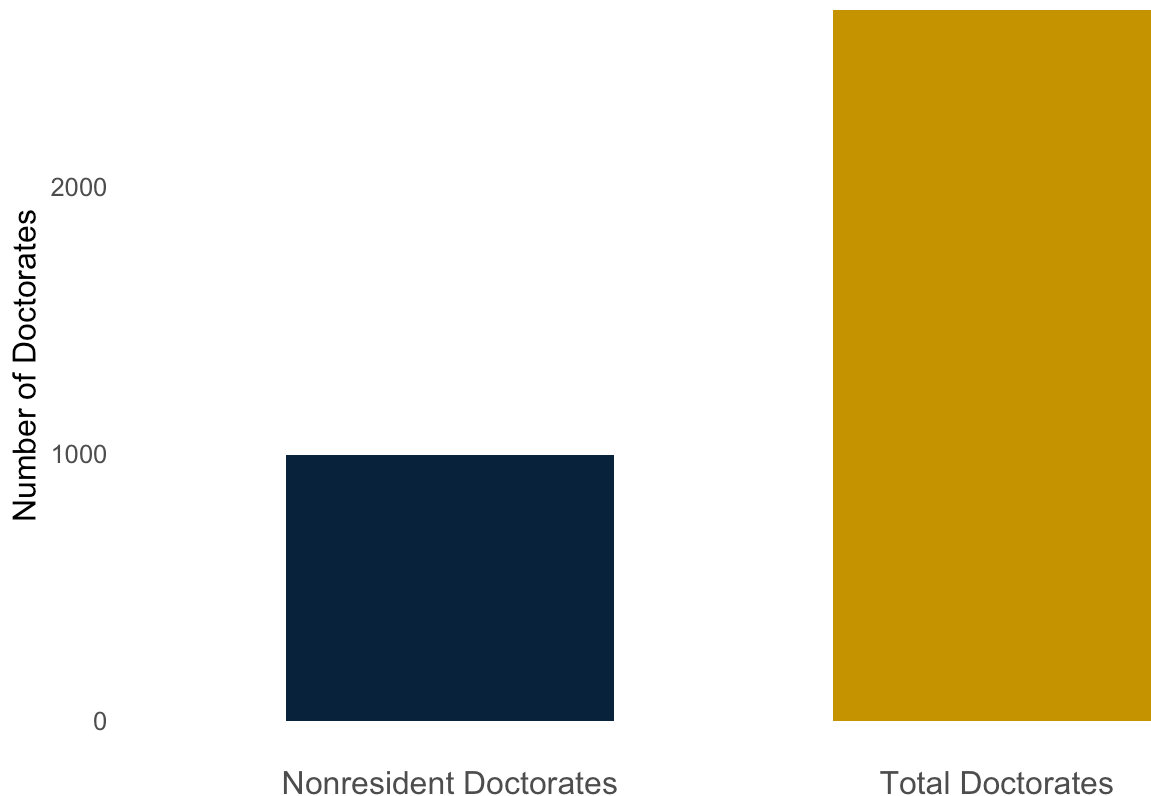
```
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid = element_blank(),
    axis.text.x = element_text(size = 12),
    plot.title = element_text(face = "bold"),
    plot.subtitle = element_text(size = 10, margin = margin(b = 10))
  )
```

Doctorates Awarded at Notre Dame (2011–2023)

Nonresident students accounted for 33.3% of all doctorates

Between 2011 and 2023, the University of Notre Dame awarded a total of 2,161 research doctorates. Of these, 735 were conferred to non-resident students, representing 33.3% of all doctoral recipients during this period. This means that roughly one in every three doctoral graduates was an international student, highlighting the university's global academic reach and sustained ability to attract talent from abroad.

While the total volume of doctorates awarded has increased over time, the proportion of non-resident recipients has remained relatively stable. This consistency suggests that international engagement is a persistent feature of Notre Dame's doctoral programs, rather than a recent or fluctuating trend. The university's level of international participation is comparable to that of other private institutions within the AAU, reinforcing Notre Dame's position as a competitive destination for global scholars across a wide range of academic disciplines.

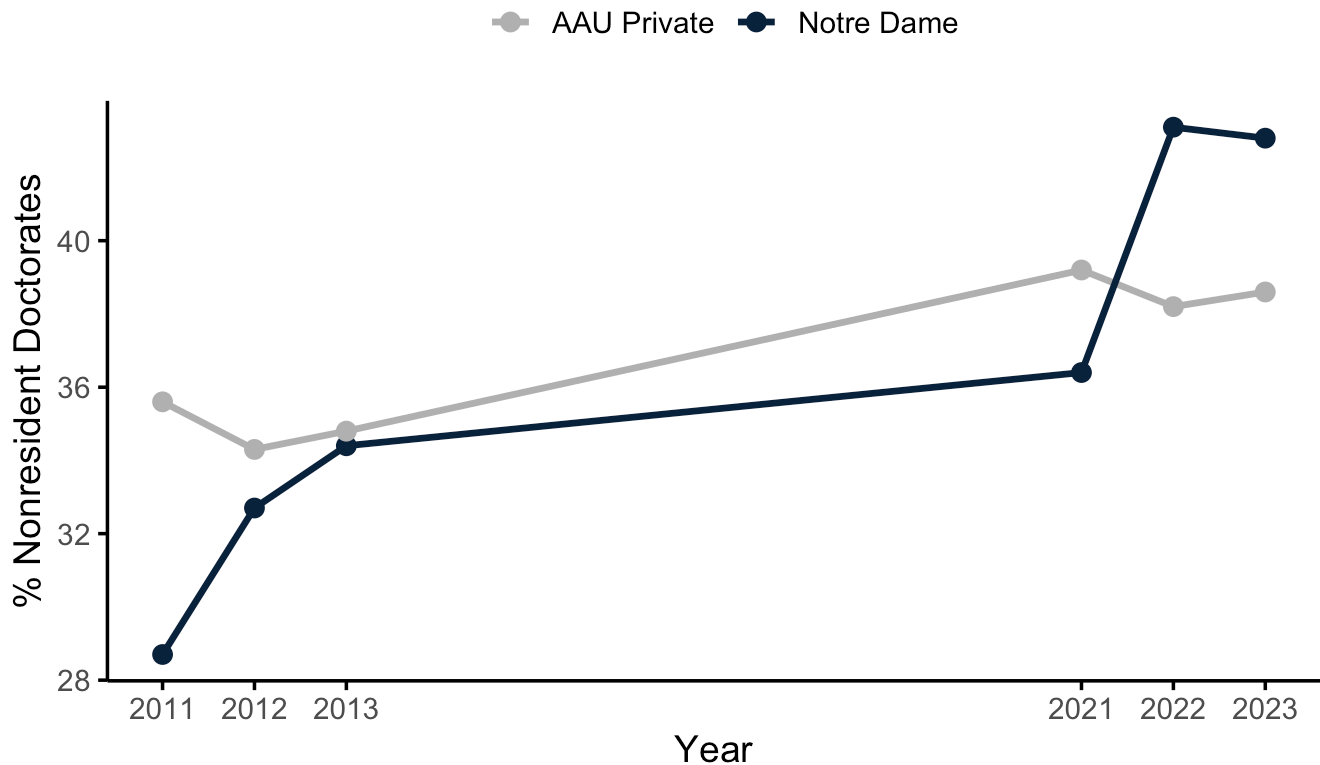# Notre Dame Compared to AAU Private Peers

```r
# Summary dataset comparing Notre Dame to AAU Private peers
nd_vs_peers <- summary_withCip %>%
  filter(!is.na(instnm), year %in% c(2011:2013, 2021:2023)) %>%
  mutate(group = if_else(instnm == "University of Notre Dame", "Notre Dame", "AAU
         Private")) %>%
  group_by(group, year) %>%
  summarise(
    total_docs = sum(total_docs, na.rm = TRUE),
    nonres_docs = sum(nonresident, na.rm = TRUE),
    pct_nonres = round(nonres_docs / total_docs * 100, 1),
    .groups = "drop"
  ) %>%
  arrange(year, desc(group))

# nd_vs_peers

# Create a line chart comparing Notre Dame to AAU Private peers
ggplot(nd_vs_peers, aes(x = year, y = pct_nonres, color = group)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  scale_color_manual(values = c("Notre Dame" = "#0c2340", "AAU Private" = "#B3B3B3")) +
      # Colorblind-safe
  scale_x_continuous(breaks = c(2011:2013, 2021:2023)) +
  labs(
    title = "Share of Doctorates Awarded to Nonresident Aliens",
    subtitle = "Comparison of Notre Dame vs. AAU Private Institutions (2011–2013, 2021–
        2023)",
    x = "Year",
    y = "% Nonresident Doctorates",
    caption = "Source: IPEDS Completions Data, NCES",
    color = NULL
  ) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.major = element_blank(),   # Remove major gridlines
    panel.grid.minor = element_blank(),   # Remove minor gridlines
    legend.position = "top",
    axis.line = element_line(color = "black"),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold") +
  geom_text(aes(label = pct_nonres), vjust = -0.7, hjust = -0.2, size = 3, show.legend
        = FALSE))
```

Share of Doctorates Awarded to Nonresident Aliens
Comparison of Notre Dame vs. AAU Private Institutions (2011–2013, 2021-

Source: IPEDS Completions Data, NCES

```
# write_csv(nd_vs_peers, "/Users/cristian/Downloads/ND Data/nd_vs_peers.csv")
```

A year-by-year comparison between Notre Dame and its AAU private peers reveals a clear upward trajectory in the university's share of doctorates awarded to non-resident students. From 2011 to 2013, Notre Dame awarded between 28.7% and 34.4% of its research doctorates to non-resident recipients—consistently trailing the AAU private university average, which remained relatively stable between 34.3% and 35.6%.

This dynamic shifted notably in the most recent period. Starting in 2021, Notre Dame's share of non-resident doctorates rose to 36.4%, then jumped to 43.1% in 2022 and remained high at 42.8% in 2023. These gains outpaced the average for private AAU institutions, which hovered just below 40% in the same years.

By 2022 and 2023, Notre Dame had not only closed the historical gap with its peers but had surpassed the group average. This shift signifies a strengthened international recruitment effort and a growing global academic presence, aligning with the university's broader strategic priorities in research, reputation, and global engagement.

Broader Trends

```r
# Sort and calculate year-over-year change
yoy_summary <- summary_withCip %>%
  arrange(alias, year) %>%
  group_by(alias) %>%
  mutate(pct_nonresident_yoy_change = pct_nonresident - lag(pct_nonresident)
  ) %>%
  ungroup()

# Filter for 2011 and 2021 to calculate 10-year change
change_yr <- yoy_summary %>%
  filter(year %in% c(2011, 2021)) %>%
  group_by(alias, year) %>%
  summarise(pct_nonresident = mean(pct_nonresident, na.rm = TRUE), .groups = "drop")
      %>%
  pivot_wider(names_from = year, values_from = pct_nonresident, names_prefix = "y") %>%
  mutate(
    pct_nonresident_10yr_change = if_else(
      !is.na(y2011) & y2011 != 0,
      (y2021 - y2011) / y2011 * 100,
      NA_real_
    )
  )
# change_yr


# Sort institutions by year change
change_yr_plot <- change_yr %>%
  filter(!is.na(pct_nonresident_10yr_change)) %>%
  arrange(desc(pct_nonresident_10yr_change)) %>%
  mutate(alias = factor(alias, levels = unique(alias)))

# Plot
ggplot(change_yr_plot, aes(x = pct_nonresident_10yr_change, y = alias)) +
  geom_col(fill = "#0c2340") +
  geom_text(aes(label = paste0(round(pct_nonresident_10yr_change, 1), "%")),
            hjust = -0.8, size = 2.5, color = "black") +
  labs(
    title = "Percentage Change in Nonresident Doctorate Awards",
    subtitle = "",
    x = "Percentage Change (2011 to 2021)",
    y = NULL,
    caption = "Source: IPEDS Completions Data, NCES"
  ) +
  scale_x_continuous(expand = expansion(mult = c(0, 0.15))) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid = element_blank(),
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(face = "bold", size = 12),
```
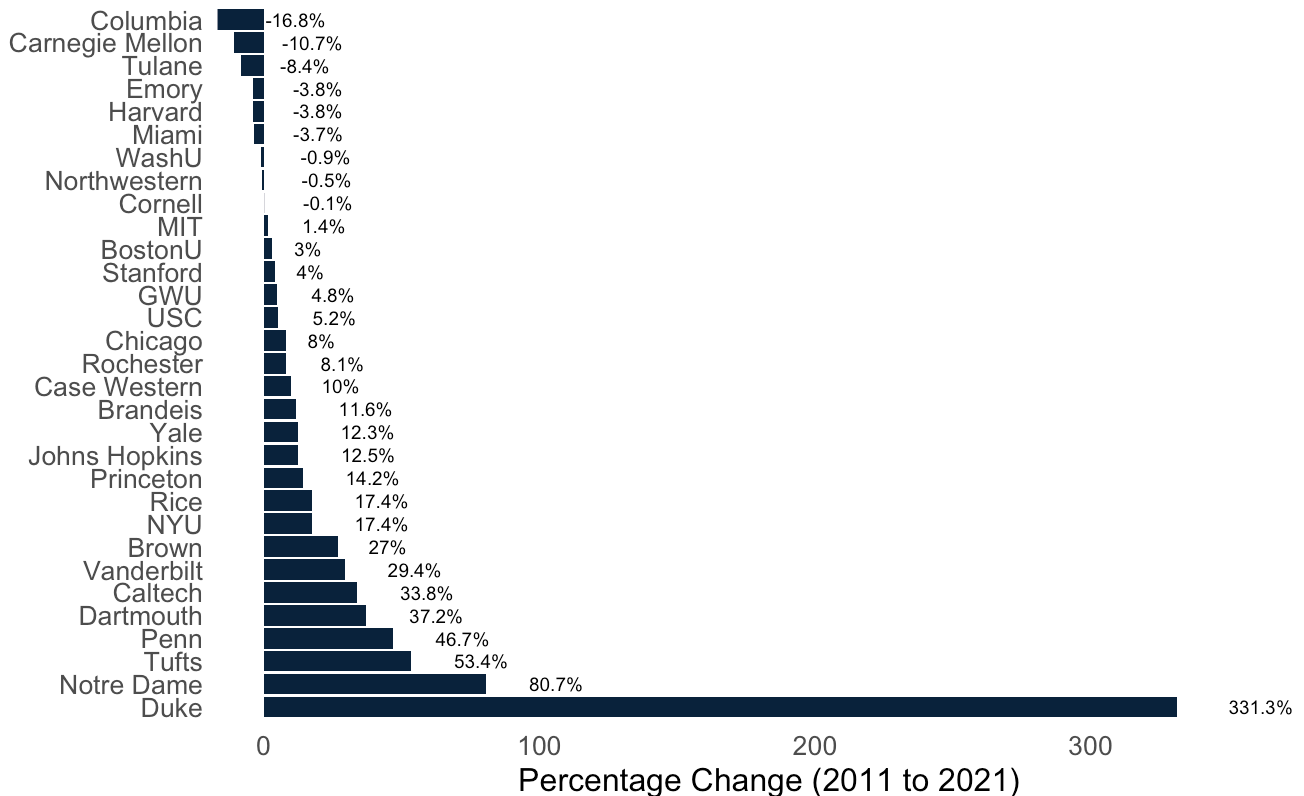
```
    plot.subtitle = element_text(size = 10, margin = margin(b = 10))
)
```

Percentage Change in Nonresident Doctorate Awards

Percentage Change (2011 to 2021)

Source: IPEDS Completions Data, NCES

```
# write_csv(change_yr, "/Users/cristian/Downloads/ND Data/change_yr.csv")
```

Looking more broadly at AAU private institutions, we observe substantial variation in how non-resident doctoral enrollment has changed over time. While some universities maintained steady shares, others experienced sharp increases or declines in the proportion of doctorates awarded to international students.

For instance, Duke University more than tripled its international doctoral population and recorded a dramatic increase, from just 9.9% in 2011 to 42.7% in 2021, a change of over 331.3%. Similarly, Tufts University's acceptance rate rose from 35.2% to 54.0%, and Dartmouth's climbed from 36.7% to 50.4%. Caltech, Penn, and Vanderbilt also posted increases exceeding 25 percentage points, indicating substantial gains in international engagement.

In contrast, other institutions remained relatively unchanged. Cornell and Northwestern saw little to no movement in their non-resident share, while a few, such as Columbia University (−16.8%) and Carnegie Mellon (−10.7%), experienced notable declines.

These trends illustrate that while the overall average among AAU private institutions remained stable, individual institutions have followed divergent paths. The substantial growth at Notre Dame stands out

as particularly strong compared to the broader landscape, reinforcing the significance of its recent increases in international doctoral participation.

Conclusion

Notre Dame has demonstrated a significant increase in the share of doctorates awarded to non-resident students over the past decade. While previously below the AAU private average, Notre Dame has now surpassed many of its peer institutions, indicating strong global engagement. These trends indicate sustained institutional success in attracting international talent, positioning Notre Dame competitively among elite private research universities.

Key Findings

- Growth in International Participation: Notre Dame increased its nonresident doctorate share from ~30% in 2011 to over 43% in 2023.
- Surpassing Peers: By 2022–2023, Notre Dame outpaced the average nonresident share among AAU private institutions.
- Institutional Variation: Across the AAU private cohort, trends varied widely; some schools saw dramatic increases (e.g., Duke, Tufts), while others (e.g., Columbia, Carnegie Mellon) declined.
- STEM Dominance: Most nonresident doctorates were concentrated in STEM fields, especially Engineering, Biological Sciences, and Computer Science.
- Limited Demographic Disaggregation: IPEDS data do not allow clear separation of gender/race by residency status, limiting targeted equity analyses.

Implications

- Strategic Recruitment: Notre Dame's rising nonresident share signals effective international recruitment and global academic branding.
- Equity Considerations: Continued gender imbalance and modest racial/ethnic diversity, especially for Black and Hispanic students, highlight equity gaps that need targeted support, regardless of citizenship.
- Benchmarking Opportunities: The variation across AAU institutions offers opportunities for Notre Dame to learn from peer strategies in recruitment and diversity.

Recommendations

- Sustain International Outreach: Maintain and refine current international recruitment pipelines, especially in high-demand fields like Engineering and Computer Science.
- Disaggregate Future Data: Explore supplemental internal data sources to analyze gender and race among nonresident students more precisely.
- Targeted Support for Women and underrepresented minorities (URMs): Address Notre Dame's persistent gender imbalance and underrepresentation of Black and Hispanic students through program-level reviews and targeted initiatives.
- Use Peer Comparisons for Policy Planning: Leverage insights from high-performing peers (e.g., Duke's international surge or Emory's gender equity) to inform strategic priorities.