College Major Backgrounds and Effects

Dylan Holliday and Max Collins

2025-05-07

Research Topic: College Majors

Our Research is focused on finding trends and insights from post-secondary majors: how one's socioeconomic status plays a role in their major, as well as their major's effect on their income after graduation. This is a topic of interest for our team because we are in college and are thus interested in how different majors contain different groups of people and how the choice of a college major can affect one's future career. We will explore these relationships using a series of data visualizations and provide insights on income and diversity trends of recent college graduates. Our underlying purpose is to expose societal trends that may be hindering diversity initiatives in order to become a basis for possible change.

Research Questions

We plan to conduct an exploratory data analysis from the information given by our data sets. Our first question we will explore is what is the highest earning major category. We will find the said category and perform a two sample t test to show if the highest earning category is significantly different from the 2nd highest. We then want to know about the distribution of the highest earning major category to see if there is any skew or outliers that would affect the mean. We also inquire whether there is a significant difference in college race proportions compared to the national proportions. To do this, we intend to conduct a one sample proportion test comparing primarily the white proportion of college students to the national to expose a possible bias.

Provenance of Our Data

Our primary data source was found on Kaggle. Kaggle is a data science focused website that contains many data sets that learners can find and use. The author of this particular data set is Bojan Tunguz. The data is described as being from the American Community Survey 2010-2012 Public Use Microdata Series. The data set contains multiple data frames, but our study will focus upon the recent-grads data, where each case represents a unique major. Each case contains information on majors, earnings post-graduation, employment, employment type, major categories, and the amount of people in each major. Since these data are from a specific survey, they have not been updated and will not. The data, however, still maintains integrity following that all these majors still exist and no abnormal world events would lead us to deny insight from the data. Our analysis here will mainly focus on median salary, major categories (majors grouped within a similar field), unemployment rate, and total amount of graduates.

Our secondary data source is from the 2022 American Community Survey, from the U.S. Census Bureau. Each case for this data set is also a major, so that is how we intend to join the data frames together. This data contains information on popular majors and the percentage of each race and sex, as well as age and income categories. With all this said, our research with this data set will solely focus on the race proportions and poverty proportions. These data are not to be updated since they are also from a specific survey. Nevertheless, the data provides useful insights on socioeconomic background's role in college majors.

FAIR and CARE Principles

Analyses

Analysis of Salaries among Major Categories

Below is an initial summary table that depicts the average salary of each quartiles 1-3, as well as the standard deviation of the median.

Table 1: Salary Summary Table per Major Category

Major_category	Count	Average First Quartile	Average Median Salary	Std. Dev. of Median Salary	Average Third Quartile
Agriculture & Natural Resources	10	25,400	36,900	6,935	48,010
Arts	8	21,963	33,063	7,223	43,663
Biology & Life Science	14	26,614	36,421	4,529	46,086
Business	13	33,462	43,538	7,774	54,846
Communications & Journalism	4	26,250	34,500	1,000	44,975
Computers & Mathematics	11	29,291	42,745	5,109	58,091
Education	16	26,591	32,350	3,893	38,563

		Average	Average	Std. Dev.	Average
		First	Median	of Median	Third
Major_category	Count	Quartile	Salary	Salary	Quartile
Engineering	29	41,555	57,383	13,626	70,448
Health	12	26,167	36,825	5,776	50,250
Humanities & Liberal Arts	15	23,493	31,913	3,393	42,073
Industrial Arts & Consumer Services	7	26,771	36,343	7,291	45,143
Law & Public Policy	5	32,640	42,200	9,066	55,000
Physical Sciences	10	28,350	41,890	8,252	57,290
Psychology & Social Work	9	25,333	30,100	5,382	38,778
Social Science	9	28,356	37,344	4,751	50,111

At first sight, one can see that engineering has quite a high salary for all quartiles. Even compared to the second highest earning category of major, business. We will perform a hypothesis test (two sample t test) to determine whether the average engineering median salary is significantly different from that of the business category.

$H_0 : \mu_1 = \mu_2$ There is no significant difference in average median salary for engineering majors and business majors.
$H_a : \mu_1 \neq \mu_2$ There exists a significant difference in average median salary for engineering majors and business majors.

It is important to note, assuming engineering is sample 1 and business is sample 2:

$\bar{x}_1 = 57383$ $s_1 = 13626$ $n_1 = 29$ $\bar{x}_2 = 43538$ $s_2 = 7774$ $n_2 = 13$

From these numbers, the $t-value$ can be found to be:

$$t = 4.16$$

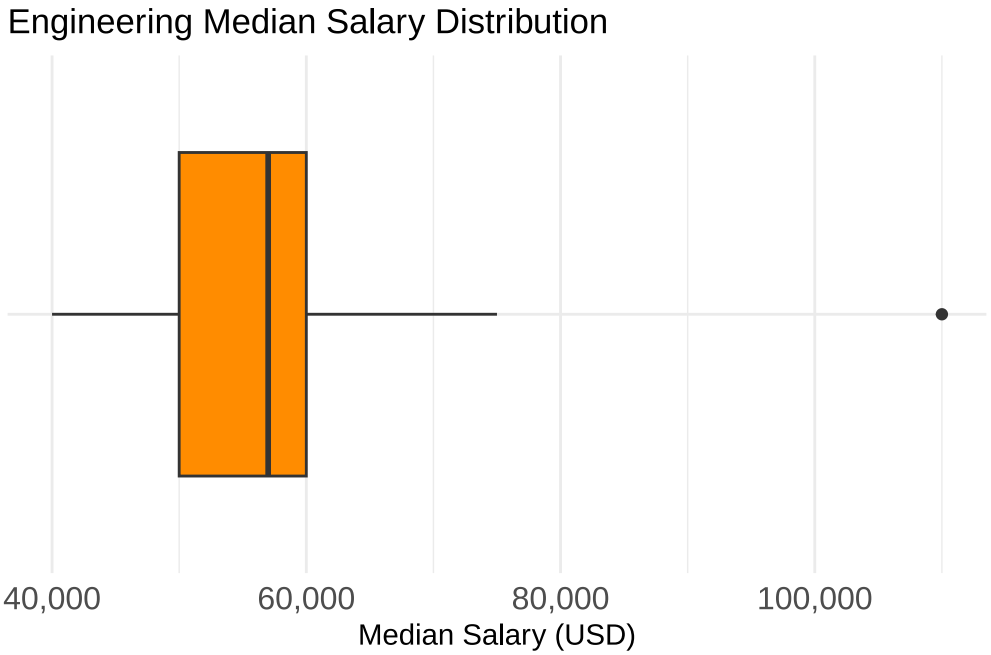
And the $p-value$ is found to be:

$$p = 0.0002$$

Because p is less than 0.05, the null hypothesis is rejected. Therefore, there is reason to believe that the median salary of engineering majors differs from that of business majors.

Now that we have this information, let's take a look at the shape of the engineering median salary data to search for outliers and any skew. See below.

Figure 1: Exploring the Shape of the Engineering Median Salary Distribution



Although this plot does not seem to have a large skew one way or another, there is a noticeable outlier that lies well beyond the other median salaries. The outlier major would be PETROLEUM ENGINEERING with a median salary of $110,000. The adjusted mean of the median salary without the outlier is $55,503.57 while the standard deviation is now $9,292.19

An Analysis of Race Frequency among Major Categories

This analysis will try to determine if there is a statistically significant bias towards people of white ethnicity in a college setting by comparing the collegiate white race percentage with the national white race percentage. We will be conducting a One Sample Proportion Z-Test to make our conclusion.

Hypotheses:

$H_0 : p = p_0$ There is no significant difference in the white collegiate race proportion and national (population) white race proportion.
$H_a : p \neq p_0$ There exists a significant difference in the proportion of white people between college students and the national (population).

To start, we must find the proportion of white college students. The proportion lies in the table below.

Below is a table depicting the percent frequency of each race in each major category as a portion of the sample population.

Table 2: Frequency of Each Major Category Based on Race

Major Category/Race	Asian	Black	Hispanic	Other	White	Total
Health	1.72%	1.43%	1.35%	0.60%	8.78%	13.88%
Biology & Life Science	1.51%	0.90%	1.03%	0.54%	6.78%	10.76%
Business	1.18%	0.93%	1.09%	0.46%	7.04%	10.70%
Psychology & Social Work	0.54%	1.06%	1.14%	0.47%	5.85%	9.06%
Education	0.35%	0.64%	0.88%	0.34%	6.03%	8.24%
Computers & Mathematics	2.38%	0.63%	0.70%	0.38%	4.04%	8.14%
Communications & Journalism	0.41%	0.75%	0.81%	0.38%	5.74%	8.10%
Law & Public Policy	0.19%	1.08%	1.03%	0.32%	4.15%	6.76%
Humanities & Liberal Arts	0.42%	0.39%	0.51%	0.27%	4.24%	5.83%
Social Science	0.70%	0.52%	0.57%	0.26%	3.19%	5.25%
Engineering	1.11%	0.22%	0.44%	0.19%	2.57%	4.54%
Arts	0.32%	0.20%	0.32%	0.17%	2.32%	3.34%
Physical Sciences	0.60%	0.23%	0.25%	0.13%	1.79%	3.00%
Industrial Arts & Consumer Services	0.20%	0.22%	0.26%	0.10%	1.63%	2.41%
Total	11.63%	9.22%	10.39%	4.61%	64.14%	100.00%

As one can tell, there is a clear disparity between the proportion of white students and the proportion of students of different races, with the total percentage of white college students being 64.14%. Since we do not yet have the population proportion, no assumptions relating to our hypotheses can be made.

Another noteworthy finding is that 20.46% of Asian students have a major in the Computers and Mathematics Category, as opposed to just 8.13% of the whole collegiate population. This percentage can be obtained by taking the 2.38% of Asians in Computers & Mathematics and divide by the 11.63% of total Asian students in the collegiate population. This statistic tells us that Asians are more likely to have a major in Computers and Mathematics than the college population as a whole. Even still, everyone should avoid stereotyping so that all people have free choice to study whatever they choose.

Back to our analysis, the college proportion of white students is 64.14%. Let's call this p_0. According to the US Census, 63.44% of the total United States population identifies as white. Let's call this p. According to the One Sample Proportion Z-Test:

$$z = \frac{p - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Where p_0 is the white college students proportion, p is the total white US population proportion, and n is the total number of college students in the sample:
3.36972×10^7

As such, our z value is calculated to be:

$$z = -84.73$$

Since our z value is so large (in absolute value), our p-value is most nearly 0, leading us to reject our null hypothesis. This leads us to believe there is a statistical difference in the proportion of white college students and the proportion of the US Population that is white. There are faults with this analysis, however; since the college data came from the US Census as well, such a large "sample" size would indicate a difference in the smallest of proportions. Therefore, Any insights to be made from this test are rendered insignificant. Conducting tests on a smaller scale (such as a school by school basis) will help individual school leaders find if their school is meeting diversity standards.

Conclusions

References

https://www.neilsberg.com/insights/united-states-population-by-race/#:~:text=The%20percent%20distributio

```
### Read in data and format to appropriate data frames

#1 Open Necessary Packages
library(tidyr)
library(readr)
library(dplyr)
library(rvest)
library(tidyverse)
library(ggplot2)
library(readxl)
library(janitor)
library(knitr)
library(kableExtra)

#2 Read in earnings data from url, diversity data from xlsx path

# Read in income data
recentGrads <- read_csv(
  # Separate string for visibility
  str_c("https://raw.githubusercontent.com/DJH6655",
        "/STAT184-Final-Project/refs/heads/main/recent-grads.csv"))

# Read in secondary data source
sexRaceFreqRaw <- read_excel("SexRaceFrequency.xlsx")


#3 Clean Excel Data, Selecting necessary attributes and renaming them
sexRaceFrequency <- sexRaceFreqRaw %>%
  slice_head(n=41) %>%
  slice_tail(n = 37) %>%
  select(c(1,2,9,10,11,12,13,21,22)) %>%
  rename(
    Major = 1,
    Total = 2,
    Race_White_Percent = 3,
    Race_Black_Percent = 4,
    Race_Asian_Percent = 5,
    Race_Hispanic_Percent = 6,
    Race_OtherMulti_Percent = 7,
    Poverty_Percent = 8,
    Not_Poverty_Percent = 9
  ) %>%
  filter(!is.na(Total)) %>%
  slice(c(-14,-20,-23)) %>%
  mutate(Major = toupper(Major),
         Total = as.numeric(Total),
         Race_White_Percent = as.numeric(Race_White_Percent),
         Race_Black_Percent = as.numeric(Race_Black_Percent),
         Race_Asian_Percent = as.numeric(Race_Asian_Percent),
         Race_Hispanic_Percent = as.numeric(Race_Hispanic_Percent),
         Race_OtherMulti_Percent = as.numeric(Race_OtherMulti_Percent),
         Poverty_Percent = as.numeric(Poverty_Percent),
         Not_Poverty_Percent = as.numeric(Not_Poverty_Percent)
         )

# Count Table
sexRaceCount <- sexRaceFrequency %>%
  mutate(
    Race_White_Count = floor((Race_White_Percent / 100)*Total),
    Race_Black_Count = floor((Race_Black_Percent / 100)*Total),
```

```r
    Race_Asian_Count = floor((Race_Asian_Percent / 100)*Total),
    Race_Hispanic_Count = floor((Race_Hispanic_Percent / 100)*Total),
    Race_OtherMulti_Count = floor((Race_OtherMulti_Percent / 100)*Total),
    Poverty_Count = floor((Poverty_Percent / 100)*Total),
    Not_Poverty_Count = floor((Not_Poverty_Percent / 100)*Total)
  )

# Join our two dataframes together
joinedTables <-
  inner_join(
    sexRaceCount,
    recentGrads,
    by = join_by(Major == Major)
  )

# Make a Summary Table
majorCategories <- joinedTables %>%
  group_by(Major_category) %>%
  summarize(
    "Count" = sum(Total.x),
    "Average First Quartile" = mean(P25th),
    "Average Median Salary" = mean(Median),
    "Average Third Quartile" = mean(P75th),
    "Average Unemployment Rate" = mean(Unemployment_rate),
    "White" = floor(mean(Race_White_Count)),
    "Black" = floor(mean(Race_Black_Count)),
    "Asian" = floor(mean(Race_Asian_Count)),
    "Hispanic" = floor(mean(Race_Hispanic_Count)),
    "Other" = floor(mean(Race_OtherMulti_Count)),
    "Poverty Count" = floor(mean(Poverty_Count))
  )



### Provides a Table on major category summary statistics

salarySummary <- recentGrads %>%
  filter(Major_category != 'Interdisciplinary') %>%
  group_by(Major_category) %>%
  summarize(
    "Count" = n(),
    "Average First Quartile" = mean(P25th),
    "Average Median Salary" = mean(Median),
    "Std. Dev. of Median Salary" = sd(Median),
    "Average Third Quartile" = mean(P75th),
  )
salarySummary %>%
  kable(
    caption = "Salary Summary Table per Major Category",
    format = "simple",
    booktabs = TRUE,
    align = "lrrrrr",
    format.args = list(big.mark = ',', digits = 2)
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 10,
  )


### Make a box plot depicting engineering salary distribution

#1 Make the necessary table, filters only engineering majors
engineerStats <- recentGrads %>%
  filter(Major_category == 'Engineering')

ggplot(engineerStats) +
  aes(x = Median, y = "" ) +
  geom_boxplot(fill = "#FF8C00") +
  labs(
    x = "Median Salary (USD)",
    y = "",
    title = "Engineering Median Salary Distribution",
  ) +
  scale_x_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 12L),
    axis.text.x = element_text(size = 12L)
  )

### Find out info about the highest earning engineering major

#1 Find Major Name
```

```
highestEarning <- engineerStats %>%
  filter(Median == max(Median)) %>%
  select(Major)
#2 Find Major Median Salary
highestMedSalary <- engineerStats %>%
  filter(Major == highestEarning[[1]]) %>%
  select(Median)
#3 Create table without outlier
withoutOutlier <- engineerStats %>%
  filter(Major != 'PETROLEUM ENGINEERING')

### Make a Race/Major Category frequency table

#1 Select Necessary attributes and pivot longer
freqTablePrep <- majorCategories %>%
  select(1,7,8,9,10,11) %>%
  pivot_longer(!Major_category, names_to = "Race", values_to = "Count")

#2 Use uncount to have a case be one individual
byStudent <- uncount(freqTablePrep, Count)

#3 Formats frequency table and relative frequencies
raceFreq <- byStudent %>%
  tabyl(Major_category, Race) %>%
  adorn_totals(where = "col") %>%
  arrange(desc(Total)) %>%
  adorn_totals(where = "row") %>%
  adorn_percentages(denominator = "all") %>%
  adorn_pct_formatting(digits = 2) %>%
  adorn_title(
    placement = "combined",
    row_name = "Major Category",
    col_name = "Race"
  )

#4 Outputs a stylized table
raceFreq %>%
  kable(
    caption = "Frequency of Each Major Category Based on Race",
    booktabs = TRUE,
    align = "lrrrrrr"
  )
```