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. And finally, we will explore how the percentage of college graduates in each major category by race is affected by the average median salary. This will show if there are any discrepancies in the opportunity to be in these high-earning fields. To do this, we will do separate linear regression models to see the R-Squared between the variables to see how well participation in a major category correlates to median salary by race.

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

The data that we used through this project satisfies the FAIR principles of findability, accessibility, interoperability, and reusability. The first dataset we found came from a publicly accessible repository from FiveThirtyEight, a reputable source on data analysis and reporting, and the second dataset we used came from the publicly accessible US Census database. These data sets can be accessed by anyone on the internet as they have been freely provided by the US Census Bureau and FiveThirtyEight. The data was also in compatible formats such as .csv files and .xlsx files, and all languages used throughout our report and the wrangling and visualization process are formal and applicable to the analysis we conducted. And finally, the data we used contains plenty of relevant pieces of economic and socioeconomic data, which, as outlined above, have sufficient provenance.

Our data also satisfies the CARE principles of collective benefit, authority to control, responsibility, and ethics. These standards are especially relevant to our project as we are working with socioeconomic data that could potentially be sensitive to some people. Our analysis aims to discuss socioeconomic impacts and trends associated with a choice in college major and give people insight into our potential findings on their impact on economic outcomes. The data we used also came from the US Census, which occurs every 10 years and people willingly participate in. We are trying to use this data to expand people's understanding on how college

major choices relate to income, poverty, and unemployment. And finally, we are not trying to do any harm through this project and are doing this to explore trends that may allow people to make better decisions for their future.

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	21,962	33,062	7,223	43,662
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	38,562
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:

$$t = 4.16$$

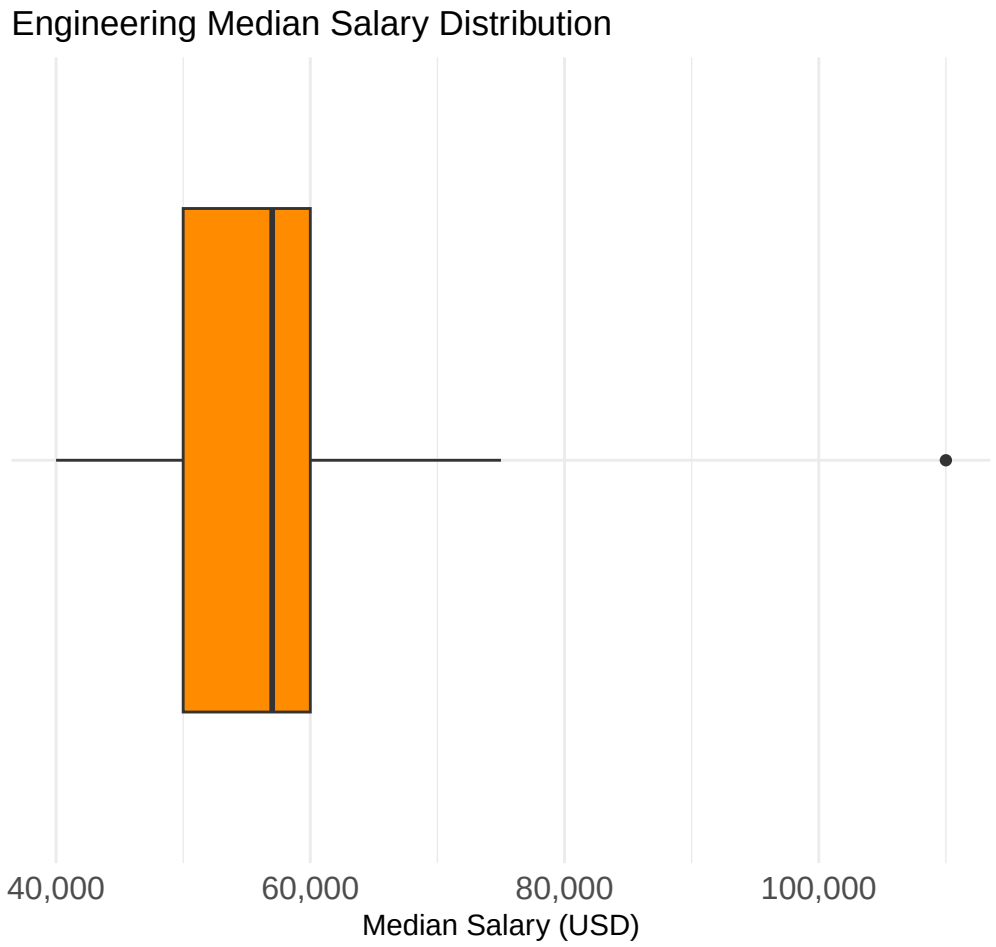
And the $p - value$ is found to be:

$$p = 0.0002$$

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.

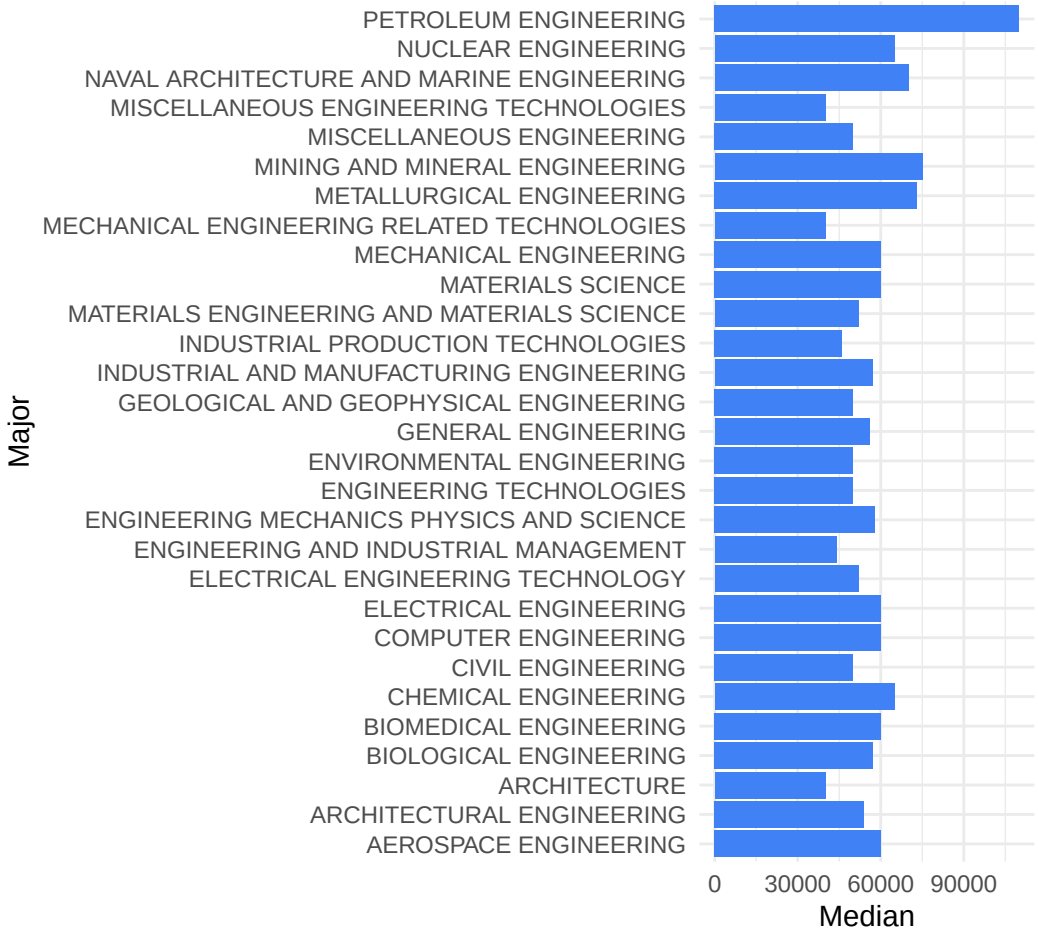
Now that we have this information, let's take a look at the shape of the engineering median salary data to search for outliers and any skew. See below.

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

Figure 2: Median Salary per Engineering Major



Here is another visualization showing just how significant the disparity between earnings for petroleum engineering majors is compared to other engineering graduates. In a field full of high earners, petroleum engineers make far more than the others, in some cases making close to double what other engineering majors make.

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%

Frequency of Each Major Category Based on Race

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:

$$z = -84.73$$

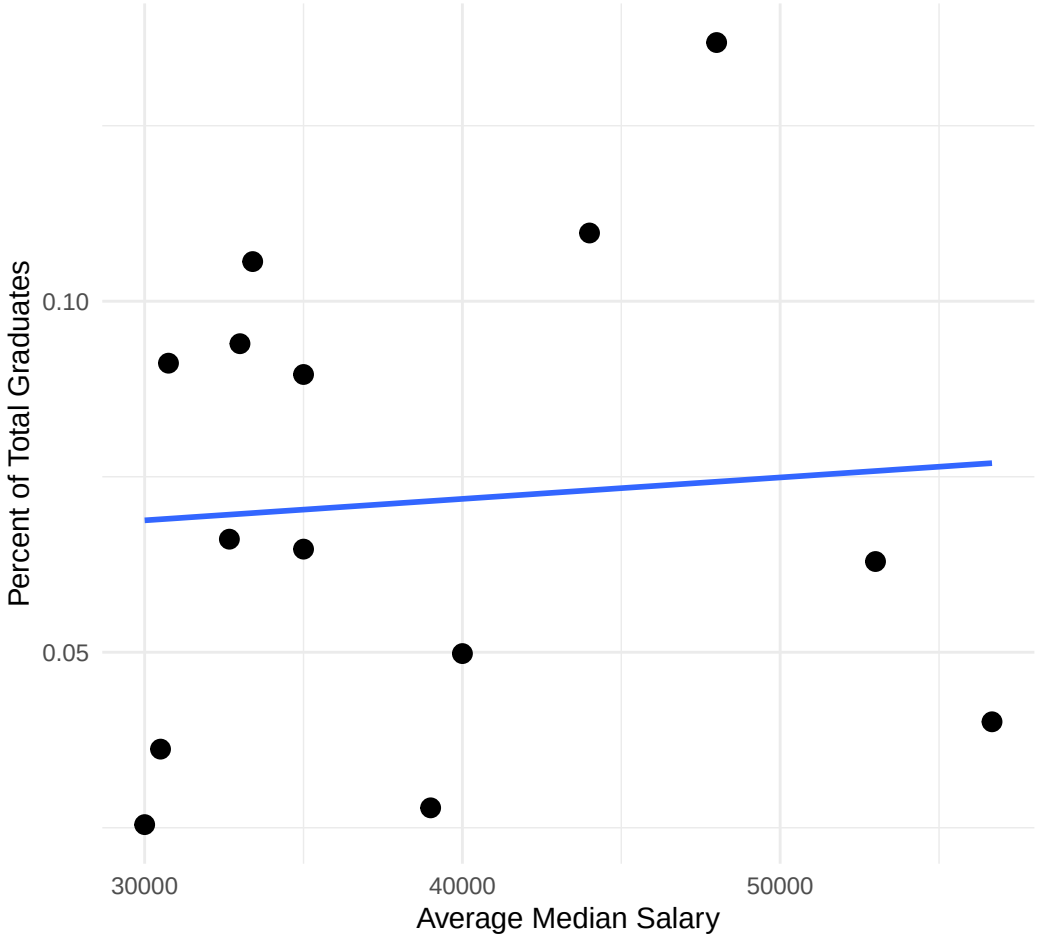
Since our z value is so large (in absolute value), our p-value is most nearly 0, leading us to reject our null hypothesis. This leads us to believe there is a statistical difference in the proportion of white college students and the proportion of the US Population that is white. There are faults with 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.

Analysis of Percentage of Graduates and Average Median Salary by Major Category

For this analysis, we took the percentage of college graduates who graduated with a degree in each major category within each race and plotted their relation to the average median salary. This shows the relative frequency of a person of each race to graduation with a degree in a certain category of major when grouped with those of their same race, not the whole sample size of all races. We made linear regressions of people with the races of white, black, asian, and hispanic for this part of our research and calculated the R-Squared of each regression to find how much of the variance in percentage of people within each race choosing a certain category of major can be explained by the average median salary.

To find what differences in the R-Squared coefficients exist between racial groups, we'll examine our linear regressions.

Figure 3: Linear Regression Dot Plot and Model of Percentage of White College Graduates by Average Median Salary

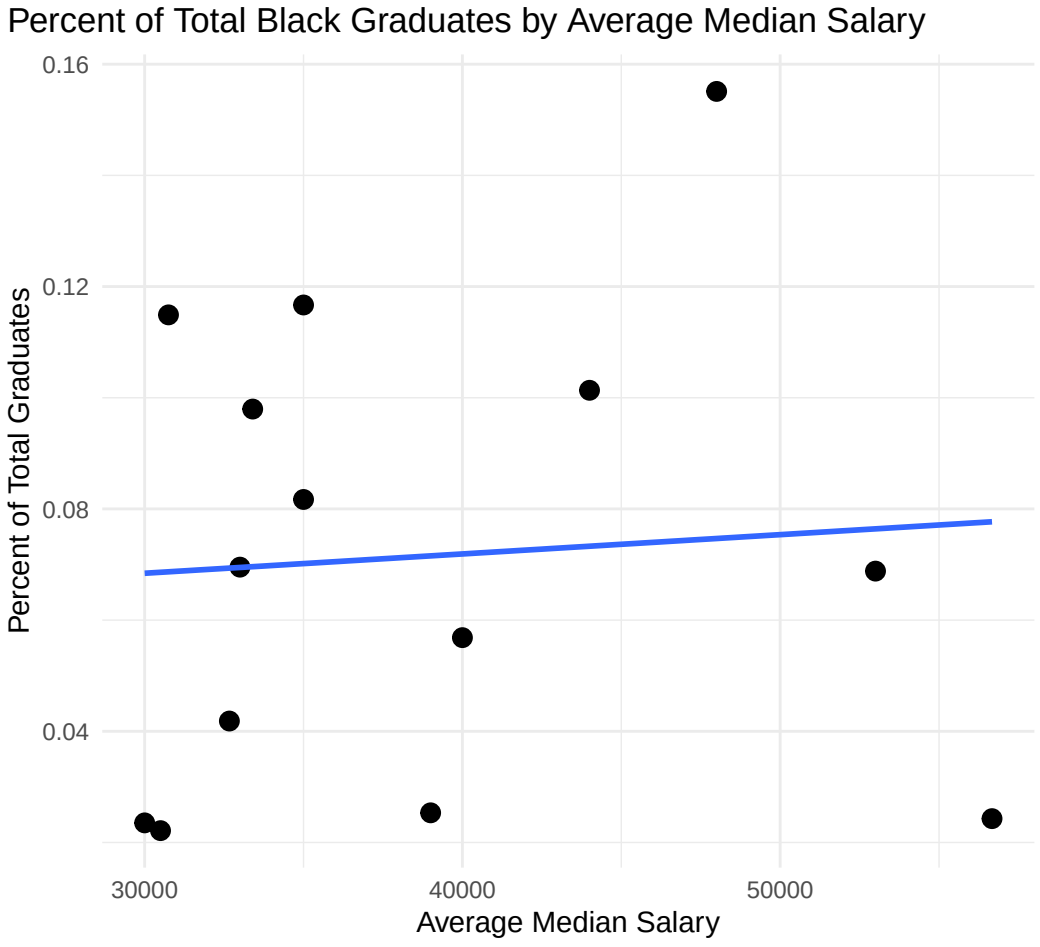



```
Call:
lm(formula = Pct ~ `Average Median Salary`, data = dotPlotWhite)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043702 -0.030065 -0.004564  0.023726  0.062561

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.962e-02  4.464e-02   1.336    0.206
`Average Median Salary` 3.055e-07  1.129e-06   0.271    0.791
```

```
Residual standard error: 0.03524 on 12 degrees of freedom
Multiple R-squared:  0.006065,  Adjusted R-squared:  -0.07676
F-statistic: 0.07322 on 1 and 12 DF,  p-value: 0.7913
```

With this first regression, we can see that the percentage of total white graduates from any given major category has almost no correlation to the average median salary. The R-Squared for this regression was only 0.006, which is less than 1% of the variance being described by the relationship between the two variables. This indicates that the average median salary is not a good predictor of what percentage of white graduates will obtain a certain category of degree.

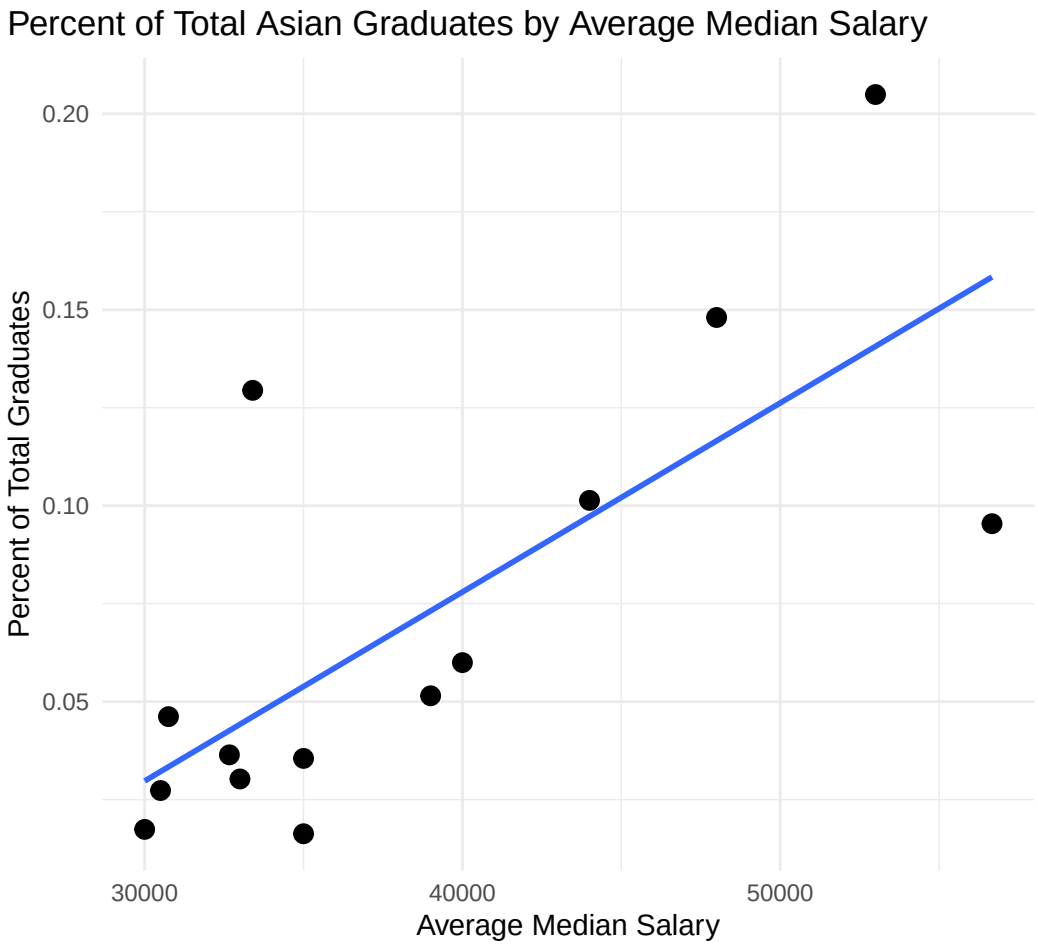
Figure 4: Linear Regression Dot Plot and Model of Black College Graduates by Average Median Salary



```
Call:
lm(formula = Pct ~ `Average Median Salary`, data = dotPlotBlack)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053398 -0.040555 -0.003772  0.028278  0.080426

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.801e-02  5.509e-02   1.053    0.313
`Average Median Salary` 3.472e-07  1.394e-06   0.249    0.807

Residual standard error: 0.04349 on 12 degrees of freedom
Multiple R-squared:  0.005148,  Adjusted R-squared:  -0.07776
F-statistic: 0.06209 on 1 and 12 DF,  p-value: 0.8074
```

And yet again, with this regression, there is almost no correlation between average median salary and the percentage of black college graduates obtaining a certain category of degree. The resulting R-Squared is 0.005, which is lower than for white graduates, but such a low number that we can't make a determination from it.

Figure 5: Linear Regression Dot Plot and Model of Asian College Graduates by Average Median Salary

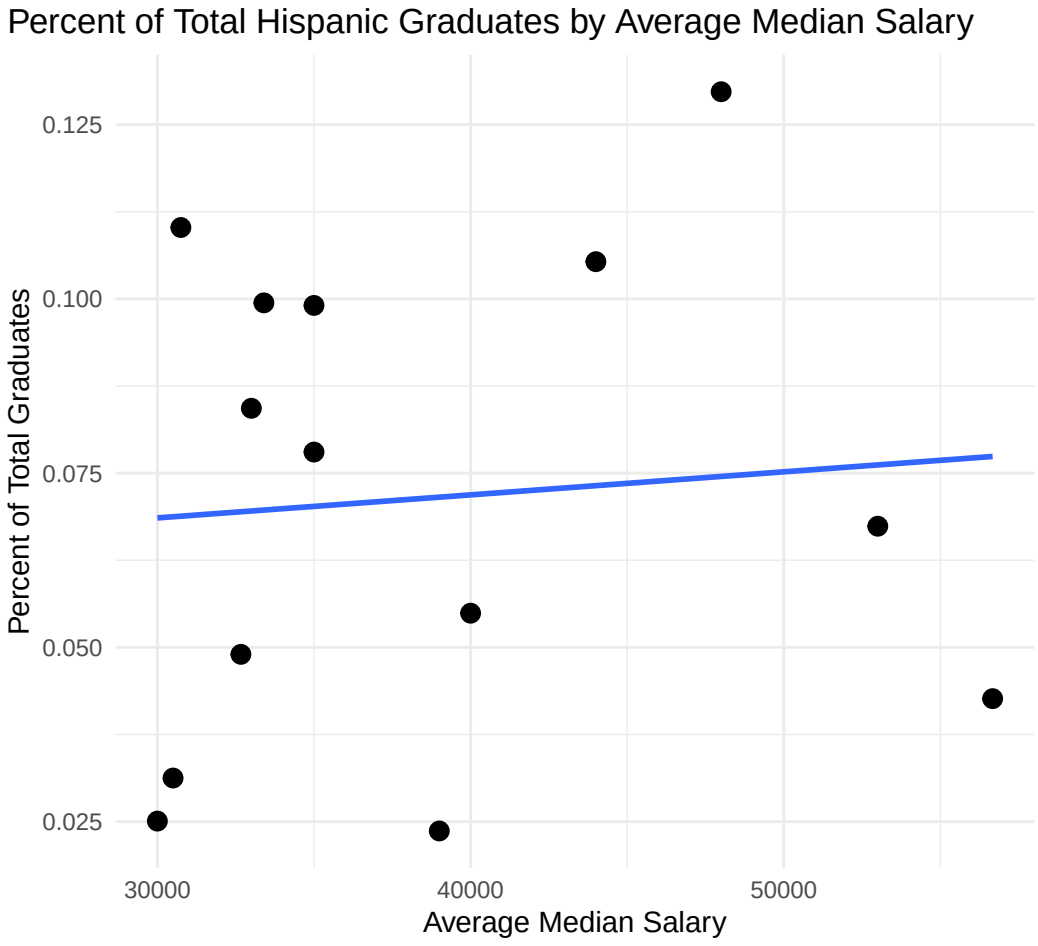



```
Call:
lm(formula = Pct ~ `Average Median Salary`, data = dotPlotAsian)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06294 -0.01827 -0.00927  0.01063  0.08328

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.149e-01  5.060e-02  -2.271  0.04234 *
`Average Median Salary`  4.822e-06  1.280e-06   3.768  0.00268 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03995 on 12 degrees of freedom
Multiple R-squared:  0.5419,    Adjusted R-squared:  0.5038
F-statistic:  14.2 on 1 and 12 DF,  p-value: 0.002682
```

For the first time, we see a clear correlation between average median salary and percentage of graduates from a particular race obtaining a degree in that category. Not only is the apparent through the R-Squared of 0.542, but the visual trend is far more obvious than any regression before it. This shows an almost undeniable correlation between the two variables.

Figure 6: Linear Regression Dot Plot and Model of Hispanic College Graduates by Average Median Salary



```
Call:
lm(formula = Pct ~ `Average Median Salary`, data = dotPlotHispanic)

Residuals:
     Min       1Q   Median       3Q      Max
-0.04789 -0.03118 -0.00050  0.02951  0.05520

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.865e-02  4.533e-02   1.294    0.220
`Average Median Salary` 3.306e-07  1.147e-06   0.288    0.778

Residual standard error: 0.03579 on 12 degrees of freedom
Multiple R-squared:  0.006878,  Adjusted R-squared:  -0.07588
F-statistic: 0.0831 on 1 and 12 DF,  p-value: 0.7781
```

To wrap it up, we have yet another graph with almost no correlation. The R-squared for this regression model for Hispanic students was a mere 0.007. There is pretty much no evidence to say that the major they are graduating with is related to how much the average median salary of that category of graduates is.

Based solely on the logic that people would be attracted to more lucrative degrees, we assumed that we would examine a positive relationship between these two variables throughout the different races, but this trend was only shown amongst Asian graduates. This also still could be a case of systemic barriers to certain high-earning degrees and jobs, but the trends among the group of white graduates that have not had to deal with the same racial oppression in this country being almost identical to black and Hispanic graduates lessens that theory. However, with our last analysis showing that white students make up a larger portion of the college graduating population than they do the total population, it is still entirely plausible that a more thorough analysis would uncover some systemic barriers to these higher degrees. Perhaps if we even just showed how the total graduations compared to the population at large, but that is beyond the scope of this study.

Conclusions

In this data analysis of the trends related to socioeconomic status and economic factors relating to college majors, we made many interesting findings.

In our first research question, we uncovered that engineering majors are the highest earning major category by average median salary and have a statistically significant difference in average median salary compared to the second highest earning category. Engineers also had a

significant outlier, with Petroleum engineering graduates making over $20,000 more than the second-highest earning engineering major.

In our second research question, our data analysis uncovered that there is a statistically significantly higher percentage of white college students than that of the overall national population. This shows a potential bias in college opportunities for white people compared to other races. However, faults in our study make it so we can't draw any real conclusions from this data.

And finally, our last research question tried to find how the average median salary affected the percentage of college graduates per race for a given major category. Our linear regressions revealed next to no correlation for every race, except for Asian graduates, who had an R-Squared of 0.542. While this may lead some to point towards there not being systemic barriers to these degrees for certain groups of people, our findings from our previous question directly refute those assumptions.

Overall, our findings show that patterns between major choices and socioeconomic factors aren't incredibly different from overall national patterns.

References

Global B2B Market Research & Advisory Solutions. (2025, February 21). *United States population by Race & Ethnicity - 2025 update.* Neilsberg. https://www.neilsberg.com/insights/united-states-population-by-race/#:~:text=The%20percent%20distribution%20of%20United%20States%20population%20by,are%20some%20other%20race%20and%2010.71%25%20are%20multiracial.

Tunguz, B. (2021, April 20). *College majors.* Kaggle. https://www.kaggle.com/datasets/tunguz/college-majors

U.S. Census Bureau (2022). *American Community Survey.* Census.gov. www.census.gov/acs


```
### 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)
```

```r
#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
```

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

```r
#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')

### Creates a sideways bar chart of Engineering median salaries
library(ggplot2)
#1 Select data frame to make graph from and pick which variables correspond to each axis
ggplot(engineerStats) +
  aes(x = Major, y = Median) +
  geom_col(fill = "#4081F6") +
  coord_flip() +
  theme_minimal()
### 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"
  )

### Wrangle data to make dot plots and create dot plot for white graduates

#1Wrangle data, create percent variable to be used in regression, and filter by race
dotPlotRacePrep <- majorCategories %>%
  select(1,4,7,8,9,10,11) %>%
  pivot_longer(#makes it so each major category is split by race of graduates
    cols = -c(Major_category, `Average Median Salary`),
    names_to = "Race",
    values_to = "Count"
  ) %>%
  group_by(Race) %>%
  mutate (Pct = Count / sum(Count))#creates new column

dotPlotWhite <- dotPlotRacePrep %>%
  filter(Race == "White")
#seperated to fit figure caption
#2 Creates dot plot
ggplot(dotPlotWhite)+
  aes(
    x = `Average Median Salary`,
    y = Pct
  ) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE)+#makes line of best fit
  labs(
    x = "Average Median Salary",
    y = "Percent of Total Graduates",
  )+
  theme_minimal()

#3 Creates linear regression for average median salary and percent of graduates
rSquareWhite<- lm(
```

```
  formula = Pct ~ `Average Median Salary`,
  data = dotPlotWhite
)

#4 Put summary into pdf
summary(rSquareWhite)


#Makes another linear regression model, this time for black graduates

#1 Filter out data by race
dotPlotBlack <- dotPlotRacePrep %>%
  filter(Race == "Black")

#Same as before and same for rest of linear regressions
#2 Creates dot plot
ggplot(dotPlotBlack)+
  aes(
    x = `Average Median Salary`,
    y = Pct
  ) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE)+
  labs(
    x = "Average Median Salary",
    y = "Percent of Total Graduates",
    title = "Percent of Total Black Graduates by Average Median Salary"
  )+
  theme_minimal()+
  theme(
    plot.title.position = "plot"
  )

#3 Creates linear regression for average median salary and percent of graduates
rSquareBlack<- lm(
  formula = Pct ~ `Average Median Salary`,
  data = dotPlotBlack
)

#4 Put summary into pdf
summary(rSquareBlack)
```

```r
#Makes another linear regression model, this time for asian graduates

#1 Filter out data by race
dotPlotAsian <- dotPlotRacePrep %>%
  filter(Race == "Asian")
#2 Creates dot plot
ggplot(dotPlotAsian)+
  aes(
    x = `Average Median Salary`,
    y = Pct
  ) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE)+
  labs(
    x = "Average Median Salary",
    y = "Percent of Total Graduates",
    title = "Percent of Total Asian Graduates by Average Median Salary"
  )+
  theme_minimal()+
  theme(
    plot.title.position = "plot"
  )

#3 Creates linear regression for average median salary and percent of graduates
rSquareAsian<- lm(
  formula = Pct ~ `Average Median Salary`,
  data = dotPlotAsian
)

#4 Put summary into pdf
summary(rSquareAsian)
#Makes another linear regression model, this time for Hispanic graduates

#1 Filter out data by race
dotPlotHispanic <- dotPlotRacePrep %>%
  filter(Race == "Hispanic")
#2 Creates dot plot
ggplot(dotPlotHispanic)+
  aes(
    x = `Average Median Salary`,
    y = Pct
  ) +
```

```
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE)+
  labs(
    x = "Average Median Salary",
    y = "Percent of Total Graduates",
    title = "Percent of Total Hispanic Graduates by Average Median Salary"
  )+
  theme_minimal()+
  theme(
    plot.title.position = "plot"
  )

#3 Creates linear regression for average median salary and percent of graduates
rSquareHispanic<- lm(
  formula = Pct ~ `Average Median Salary`,
  data = dotPlotHispanic
)

#4 Put summary into pdf
summary(rSquareHispanic)
```