# INFO3370-Final-Project

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(haven)
```

# JOB SATISFACTION AND HIGHER EDUCATION:

**The Tired Folks**

Aditya Kakade, Richie Sun, Gabby Fite, Tiffany Pan, Jacqueline Hui, Brittanie Chen

### Why are we interested in this topic and why is it important?

As undergraduate students, we are interested in the impact of higher education on overall job satisfaction after graduation. Our project hopes to provide insight into how different graduate degrees impact one's life. This can help prospective high school and undergrad students shape their education path since they can examine the results and see what degrees are expected to yield the highest reward for them after graduation.

### Unit of Analysis:

Our unit of analysis is an individual. We are using the IPUMS Higher Education data set, which contains survey information gathered from survey participants who have at least a bachelors degree.

### Target Population:

Our target population are individuals who have graduated with at least a bachelor's degree and have entered the work force. Ideally, our population has an equal proportion of bachelor, master, and doctorate recipients across various different majors.

This target populations makes observations from our study applicable to students from any population.

### Predictors:

Our two predictors are salaries (in USD), major fields (for doctorate recipients), and race/ethncity.

### Outcome Variable:

Our overall outcome variable is job satisfaction as a way to gain an understanding of individuals' overall contentment with their employment.

### Summary Statistic:

The summary statistic is the proportion of respondents in each salary range that fall into each job satisfaction category, adjusted by their survey weight (`wtsurvy`). This weighted proportion accounts for the sample design and aims to make the results more representative of the population from which the sample was drawn.

### Data Source:

We sourced our data from the IPUMS Higher Education database, which draws from the National Surveys of College Graduates, Recent College Graduates and Doctorate Recipients.

```
data = read_dta("data/highered_00001.dta")
```

## Graph 1: Job Satisfaction and Salary:

### Filtering for Sample Restrictions

During our exploratory analysis, we noticed that the variables were inconsistent across data points, we worked to select data points that had the same variables present.

We first drop all cases of NA in our `wtsurvy` variable, which brings our data variable from 1,140,565 cases down to 190,611 cases.

Then, we also filtered out the cases where `jobsatis > 4` , since 98 refers to a logical skip. For the same reason, cases with `salary` values of 9,999,998 plus were also dropped. This brings us down to 139,656 usable cases .

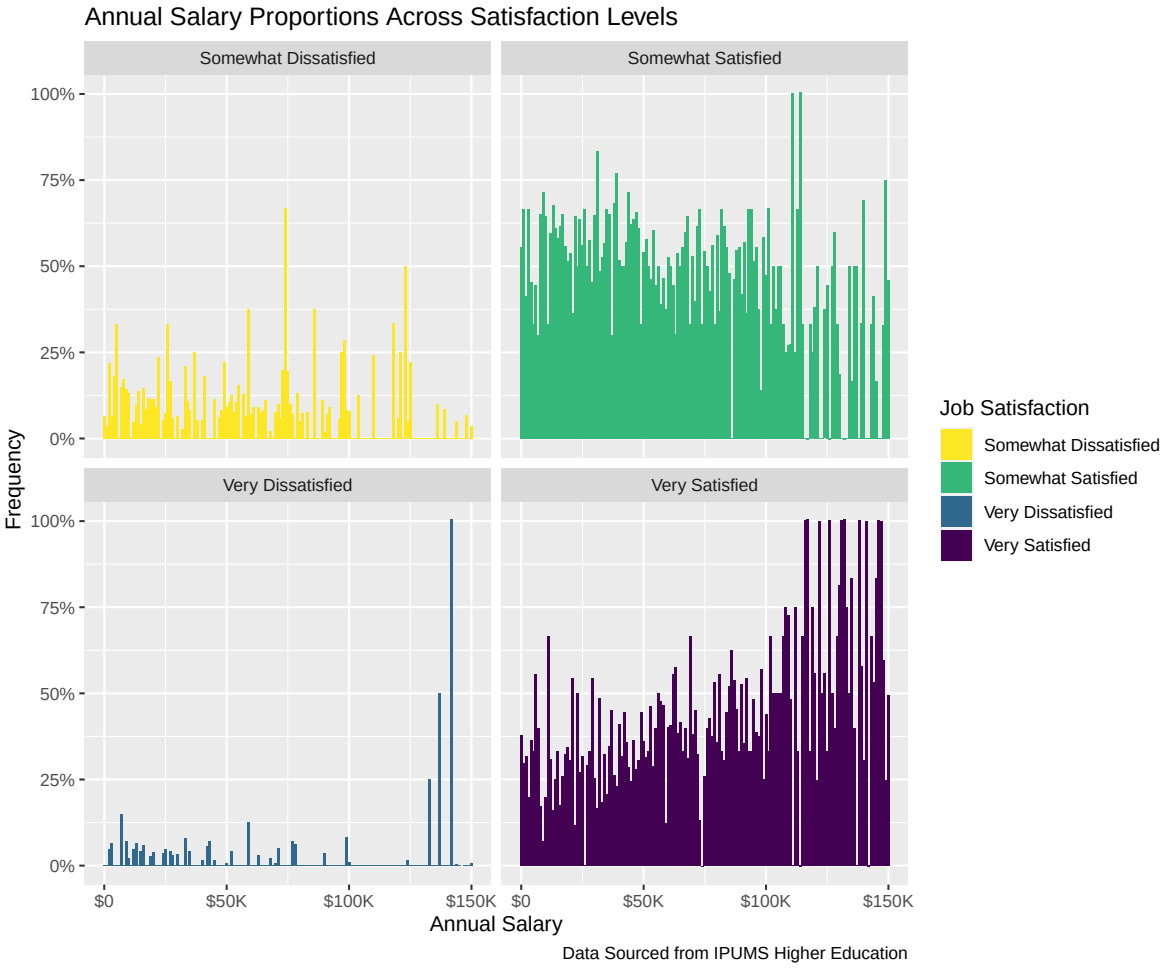
```
filtered <- data |> drop_na(wtsurvy)

filtered_sat <- filtered |> drop_na(jobsatis) |>
  filter(
    jobsatis <= 4
  ) |>
  filter(
    salary != 9999998 & salary != 9999999
  )
```

```
filtered_sat |>
  group_by(salary, jobsatis) |>
  summarise(weight_sum = sum(wtsurvy)) |>
  mutate(proportion = weight_sum / sum(weight_sum)) |>
```

```
    mutate(
      jobsatis = case_when(
        jobsatis == 1 ~ "Very Satisfied",
        jobsatis == 2 ~ "Somewhat Satisfied",
        jobsatis == 3 ~ "Somewhat Dissatisfied",
        jobsatis == 4 ~ "Very Dissatisfied"
      )
    ) |>
    mutate(proportion = weight_sum / sum(weight_sum)) |>
    ggplot(mapping = aes(x = salary, y = proportion, fill = jobsatis)) +
    geom_bar(stat = "identity") +
    facet_wrap("jobsatis") +
    labs(
      title = "Annual Salary Proportions Across Satisfaction Levels",
      x = "Annual Salary",
      y = "Frequency",
      caption = "Data Sourced from IPUMS Higher Education",
      fill = "Job Satisfaction"
    )+
    scale_fill_viridis_d(direction = -1) +
    scale_y_continuous(labels = label_percent()) +
    scale_x_continuous(labels = label_dollar(scale_cut = cut_long_scale()))
```

```
`summarise()` has grouped output by 'salary'. You can override using the
`.groups` argument.
```




**Interpretation of Results:**

Individuals who rated themselves as very dissatisfied with their jobs often reported salaries exceeding $125,000. Similarly, among those who expressed high levels of job satisfaction, a majority earned salaries surpassing $100,000. Interestingly, both the very satisfied and very dissatisfied groups exhibited a similar pattern of higher frequency at higher salary levels. Conversely, individuals who described themselves as somewhat satisfied or somewhat dissatisfied tended to report more frequently at lower salary ranges.

## Graph 2: Job Satisfaction Across Different Majors

**Filtering for Sample Restrictions**

For this graph, we also needed to drop cases that were NA in our original data set. We dropped the NA cases in `wtsurvy`, `jobsatis`, and `ndgmemg`. In addition, we also filter out cases where the `ndgmemg` variable was 99, since this means the respondent skipped the question.

This brings our data from 1,140,565 to 139,656.

```
data_major <- data |>
  drop_na(wtsurvy)|>
  drop_na(jobsatis)|>
  drop_na(ndgmemg)
```

```
data_major <- data_major |>
  filter(ndgmemg != 99)|>
  mutate(
    ndgmemg = case_when(
      ndgmemg == 1 ~ "Computer/Mathematical Sciences",
      ndgmemg == 2 ~ "Biological/Agricultural/Environment Sciences",
      ndgmemg == 3 ~ "Physical and Related Sciences",
      ndgmemg == 4 ~ "Social and Related Sciences",
      ndgmemg == 5 ~ "Engineering",
      ndgmemg == 6 ~ "Science/Engineering Related Fields",
      ndgmemg == 7 ~ "Non-science and Engineering Fields",
    )
  )|>
  filter(
    salary != 9999998 & salary != 9999999
  ) |>
  mutate(
    jobsatis = case_when(
      jobsatis == 1 ~ "Very Satisfied",
      jobsatis == 2 ~ "Somewhat Satisfied",
      jobsatis == 3 ~ "Somewhat Dissatisfied",
      jobsatis == 4 ~ "Very Dissatisfied"
    )
  )
```

```
data_major_job <- data_major |>
  select(wtsurvy,jobsatis, ndgmemg)|>
  group_by(ndgmemg, jobsatis)|>
  summarise(weight_sum = sum(wtsurvy)) |>
  mutate(proportion = weight_sum / sum(weight_sum))
```

```
`summarise()` has grouped output by 'ndgmemg'. You can override using the
`.groups` argument.
```

```
ggplot(data = data_major_job,
       mapping = aes(x = jobsatis,y = proportion, fill = jobsatis)) +
  geom_bar(stat = "identity")+
  facet_wrap("ndgmemg")+
  theme(panel.spacing = unit(1, "lines"))+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  theme(legend.position = "bottom") +
  labs(
    y = "Proportion of Major",
    title = "Job Satisfaction by Major Field",
    subtitle = "Based on Results of Doctorate Recipients",
    caption = "Data Sourced from IPUMS Higher Education",
    fill = "Job Satisfaction Level"
  )+
  scale_fill_viridis_d(direction = -1)+
  scale_y_continuous(labels = label_percent())
```


## Graph 3: Exploring Major and Race/Ethnicity and Job Satisfaction:

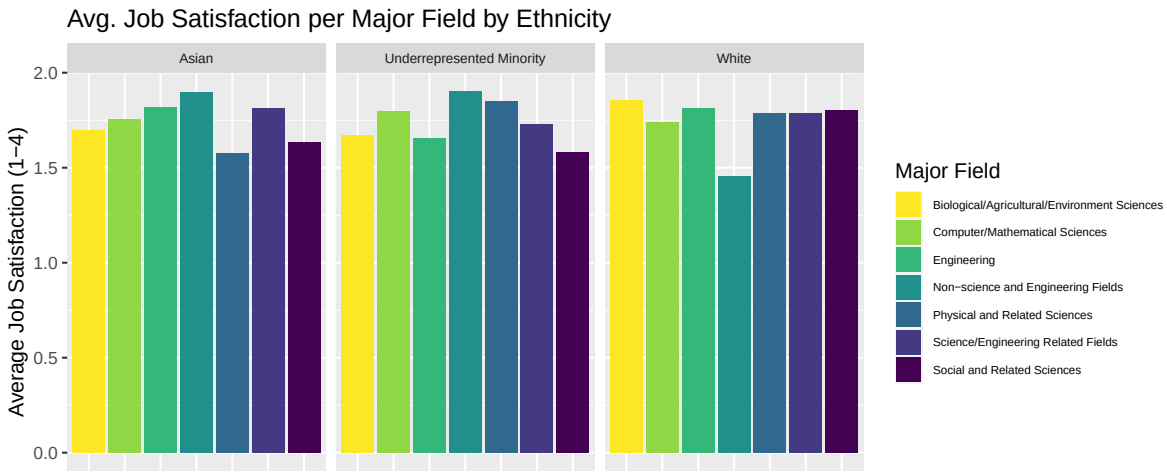
Our findings from graph 2 made us curious to how adding the variable of race/ethnicity would impact out results.

```
raceth_job_sat <- filtered_sat |>
  group_by(raceth, ndgmemg) |>
  summarise(mean_jobsatis = weighted.mean(jobsatis, wtsurvy))|>
  mutate(
    raceth = case_when(
      raceth == 1 ~ "White",
      raceth == 2 ~ "Asian",
      raceth == 3 ~ "Underrepresented Minority",
```

```
      raceth == 4 ~ "Other"
    ),
    ndgmemg = case_when(
      ndgmemg == 1 ~ "Computer/Mathematical Sciences",
      ndgmemg == 2 ~ "Biological/Agricultural/Environment Sciences",
      ndgmemg == 3 ~ "Physical and Related Sciences",
      ndgmemg == 4 ~ "Social and Related Sciences",
      ndgmemg == 5 ~ "Engineering",
      ndgmemg == 6 ~ "Science/Engineering Related Fields",
      ndgmemg == 7 ~ "Non-science and Engineering Fields",
    ))
```

```
`summarise()` has grouped output by 'raceth'. You can override using the
`.groups` argument.
```

```
ggplot(data = raceth_job_sat, mapping = aes(x = ndgmemg,
  y = mean_jobsatis, fill = ndgmemg)) +
  facet_wrap("raceth") +
  geom_bar(stat = "identity") +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        strip.text = element_text(size = 7),
        legend.key.size = unit(0.5, 'cm'),
        legend.text = element_text(size = 6)) +
  ylab("Average Job Satisfaction (1-4)") +
  ggtitle(("Avg. Job Satisfaction per Major Field by Ethnicity")) +
  guides(fill=guide_legend(title="Major Field")) +
  scale_fill_viridis_d(direction = -1)
```




**Interpretation of Results:**

We observed similar trends in the Biological/Agricultural/Environmental Sciences, Computer/Mathematical Sciences, Science/Engineering Related Fields, and Social and Related Sciences:

These majors had a majority of respondents are somewhat satisfied with their jobs, followed by those who are very satisfied. Very few are somewhat dissatisfied, and a negligible proportion is very dissatisfied.

Engineering: Highest proportion of very satisfied respondents among the displayed fields: Very Satisfied > Somewhat satisfied > Somewhat dissatisfied and Very dissatisfied

Non-science and Engineering Fields: Very satsified > Somewhat satisfied > Somewhat dissatisfied and Vert Dissatisfied.

Physical and Related Sciences: The somewhat dissatisfied category is more prominent here compared to the other fields: Somewhat satisfied > Very satisfied > Somewhat dissatisfied and Very Dissatisfied

**Across all fields,** "Somewhat Satisfied" is the most common response, followed by "Very Satisfied". The "Very Dissatisfied" response is universally less frequent. The visual suggests that job satisfaction levels are high among doctorate recipients, but varies depending on the field of study. It's important to note that satisfaction is subjective and may be influenced by a variety of factors not captured by this data. In addition, it may be harder for participants to acknowledge that they are dissatisfied with their fields, as it'd undermine their time and effort spent in that field which may also skew the results.

**Findings on Ethnicity and Race**

Underrepresented Minorities and White respondents have the highest job satisfaction in Non-Science Fields, while Asian respondents have the highest job satisfaction in Bio/Agricultural/Environmental sciences. Non-science fields exhibit the widest satisfaction difference among ethnicities (~0.45), whereas CS/math had a narrower difference (~0.05). Major Fields with the lowest satisfaction also differ across ethnicities, with the lowest satisfaction among Asian, White, and Underrepresented Minority respondents being in the Non-Science, Physical Science, and Social Science fields respectively. This visualization allows us to see that a difference in job satisfaction across Major Fields is present when we consider ethnicities as a subgroup. We cannot say for sure that this is the sole factor, as multiple aspects of an individual such as financial background and gender might influence satisfaction in a field. Additionally, because the data groups all ethnicities except for White and Asian into the 'Minorities' category, all differences across ethnic groups are not visible.