INFO3370-Final-Project

```r
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 becom
```

```r
library(scales)
```

```
Attaching package: 'scales'

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

    discard

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

    col_factor
```

```r
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 and overall job satisfaction after graduation. Our project hopes to provide insight into how different graduate degrees impact one's life graduation. 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.

The choice of target population is motivated by the fact that it represents the people around us the best, as college students who will be graduating in the near future. This makes the observations from our study very applicable and therefore easily share-able with others around us.

Predictors:

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

Outcome Variable:

Our overall outcome variable is job satisfaction as a way to gain an understanding of individuals' overall contentment and fulfillment 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, specifically from National Surveys of College Graduates, Recent College Graduates and Doctorate Recipients.

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

Graph 1: Job Satisfaction and Salary:

Filtering for Sample Restrictions

Prior to any sort of graphing, as we were looking through the data-set, we noticed quite a few values that looked out of place. We first needed to drop those cases, since they were essentially unusable data points.

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.

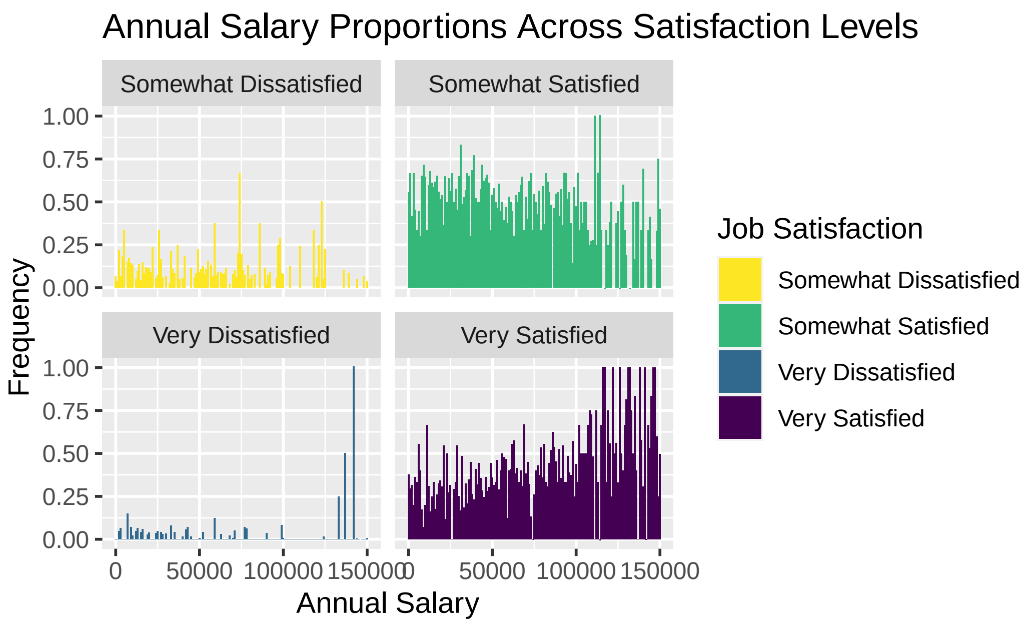
Then, we also filtered out the cases where `jobsatis <= 4` , since IPUMS tells us that a code of 98 refers to a logical skip. For the same reason, cases with `salary` values of 9,999,998 and 9,999.999 were also dropped. This brings us down to 139,656 cases of usable data, meaning we dropped 1,000,909 cases. Using these values would make them skew our statistics, as they are intentionally huge to represent skips.

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

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

```r
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)
```

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



Data Sourced from IPUMS Higher Education

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 also `ndgmemg` this time, since we are observing majors now. In addition, we also filter out cases where the `ndgmemg` variable was 99, since it also refers to a logical skip for the survey and the `ndgmemg` is crucial because it determines a participant's major.

This brings our data from 1,140,565 to 139,656 cases again, meaning we dropped 1,000,909 cases.

```
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"
    )
  )
```

We notice that our data set does not included specific data about non-stem fields. It is strange since this data set is not specific to only STEM Related Higher Education. Moreover, there is a great number of people in the United States who pursue higher education in non-STEM related fields. Hence, we acknowledge that this data set has some bias.

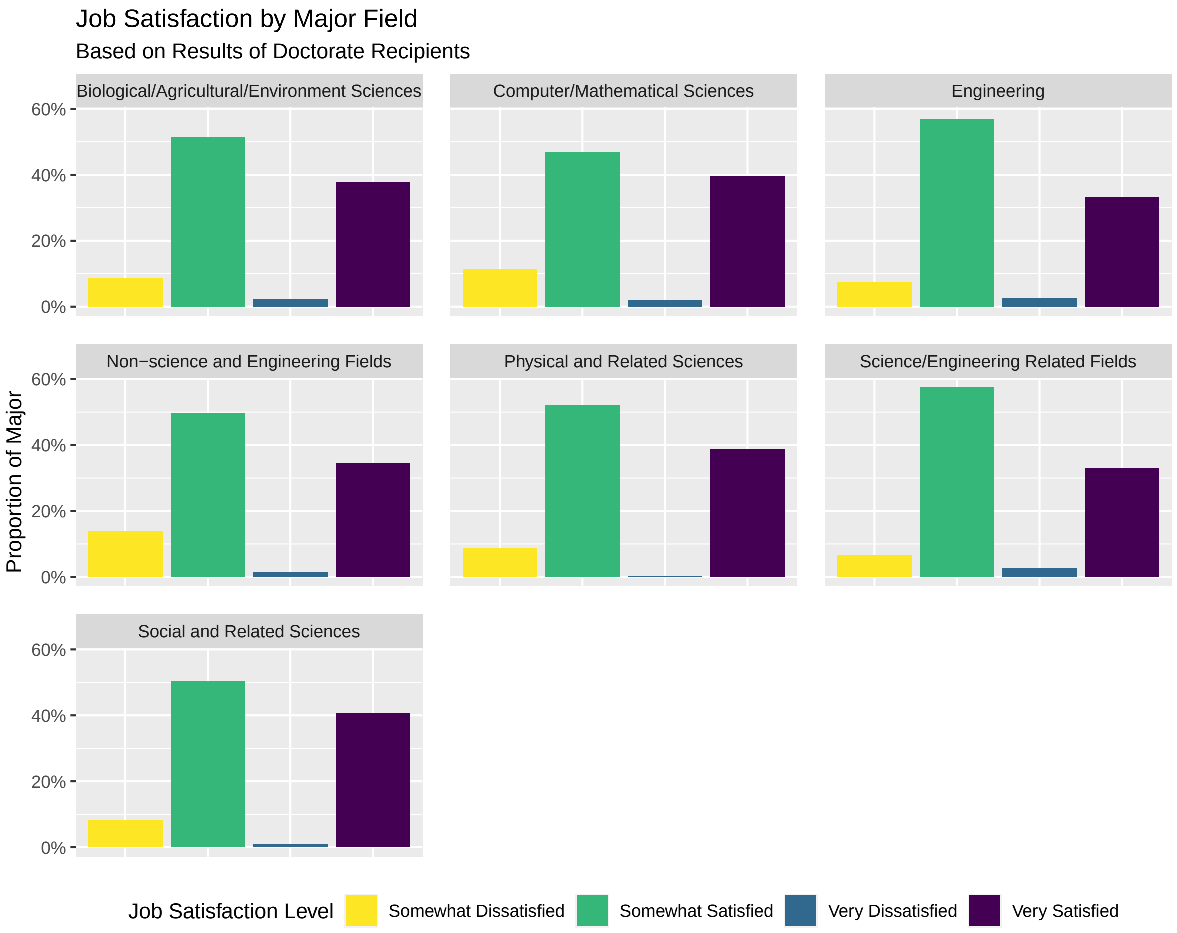
We now explore job satisfaction markers and overall job satisfaction in these specified majors.

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

```r
  mutate(proportion = weight_sum / sum(weight_sum))
```

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

```r
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())
```



Data Sourced from IPUMS Higher Education

Interpretation of Results:

Biological/Agricultural/Environmental Sciences: 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.

Computer/Mathematical Sciences: Again, somewhat satisfied is the most common response, with very satisfied as the second most. Very dissatisfied individuals are nearly absent, and somewhat dissatisfied respondents are few.

Engineering: This field has the highest proportion of very satisfied respondents among the displayed fields. Somewhat satisfied is next, and there are very few somewhat dissatisfied and very few very dissatisfied individuals.

Non-science and Engineering Fields: A high proportion of respondents are very satisfied, with somewhat satisfied coming in second. The somewhat dissatisfied and very dissatisfied categories have a small presence.

Physical and Related Sciences: Most respondents are somewhat satisfied, with very satisfied close behind. The somewhat dissatisfied category is more prominent here compared to the other fields, though the very dissatisfied group remains small.

Science/Engineering Related Fields: Somewhat satisfied is the most common response, followed by very satisfied. There are visible proportions of both somewhat dissatisfied and very dissatisfied respondents.

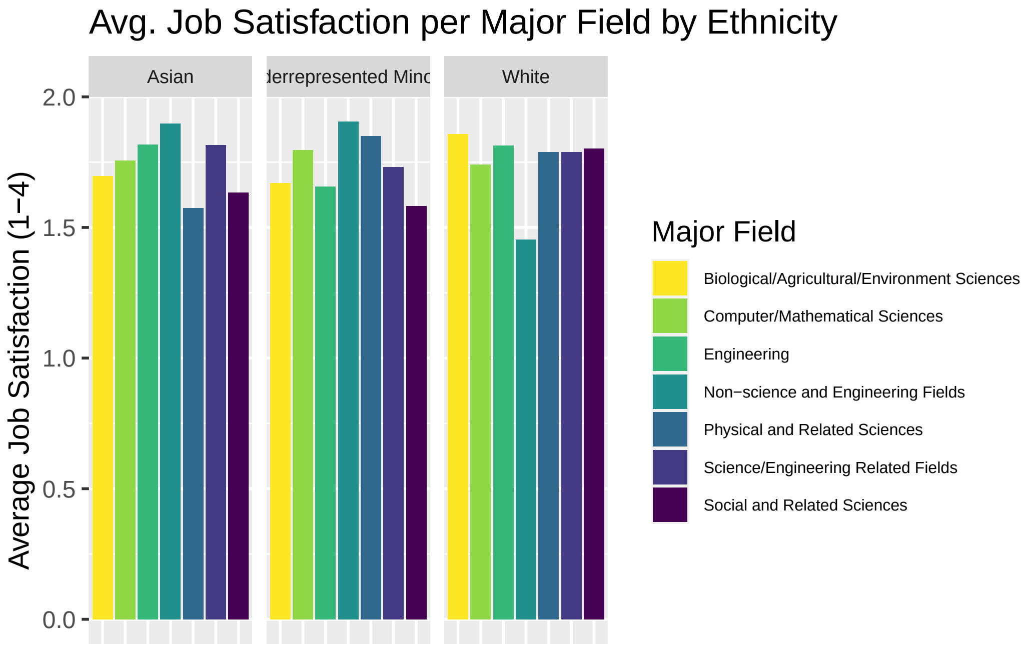
Social and Related Sciences: In this category, the somewhat satisfied group is the largest, followed by very satisfied, with smaller but noticeable percentages of both somewhat dissatisfied and very dissatisfied respondents.

Across all fields, "Somewhat Satisfied" tends to be the most common response, with "Very Satisfied" being the second most common. The "Very Dissatisfied" response is consistently the least common across all fields. The visual suggests that job satisfaction levels are relatively 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 subjectively 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 a bit.

```r
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)
```


Avg. Job Satisfaction per Major Field by Ethnicity
Asian
derrepresented Minc
White
Average Job Satisfaction (1−4)
Major Field
Biological/Agricultural/Environment Sciences
Computer/Mathematical Sciences
Engineering
Non−science and Engineering Fields
Physical and Related Sciences
Science/Engineering Related Fields
Social and Related Sciences