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)
```

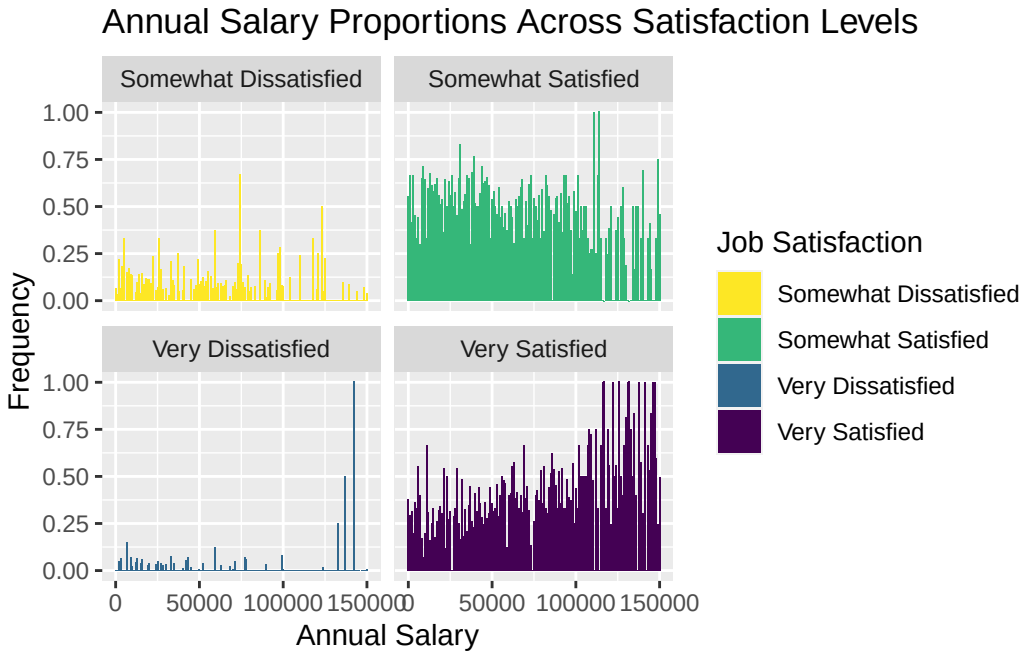
```r
data = read_dta("data/highered_00001.dta")
```

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

filtered_sat <- filtered |> drop_na(jobsatis) |>
  filter(
    jobsatis <= 4
  )


filtered_sat |>
  filter(
    salary != 9999998 & salary != 9999999
  ) |>
  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",
    fill = "Job Satisfaction"
  )+
  scale_fill_viridis_d(direction = -1)
```

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



Major Specific Data:

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