# HW 04 - What should I major in?

Elijah Russell

10/03/25

## Load packages and data

```
library(tidyverse)
library(scales)
library(fivethirtyeight)
library(moments)
```

## Exercises

### Exercise 1

```
college_recent_grads %>%
  arrange(desc(sharewomen)) %>%
  select(major,total,sharewomen) %>%
  top_n(3)
```

```
## Selecting by sharewomen
```

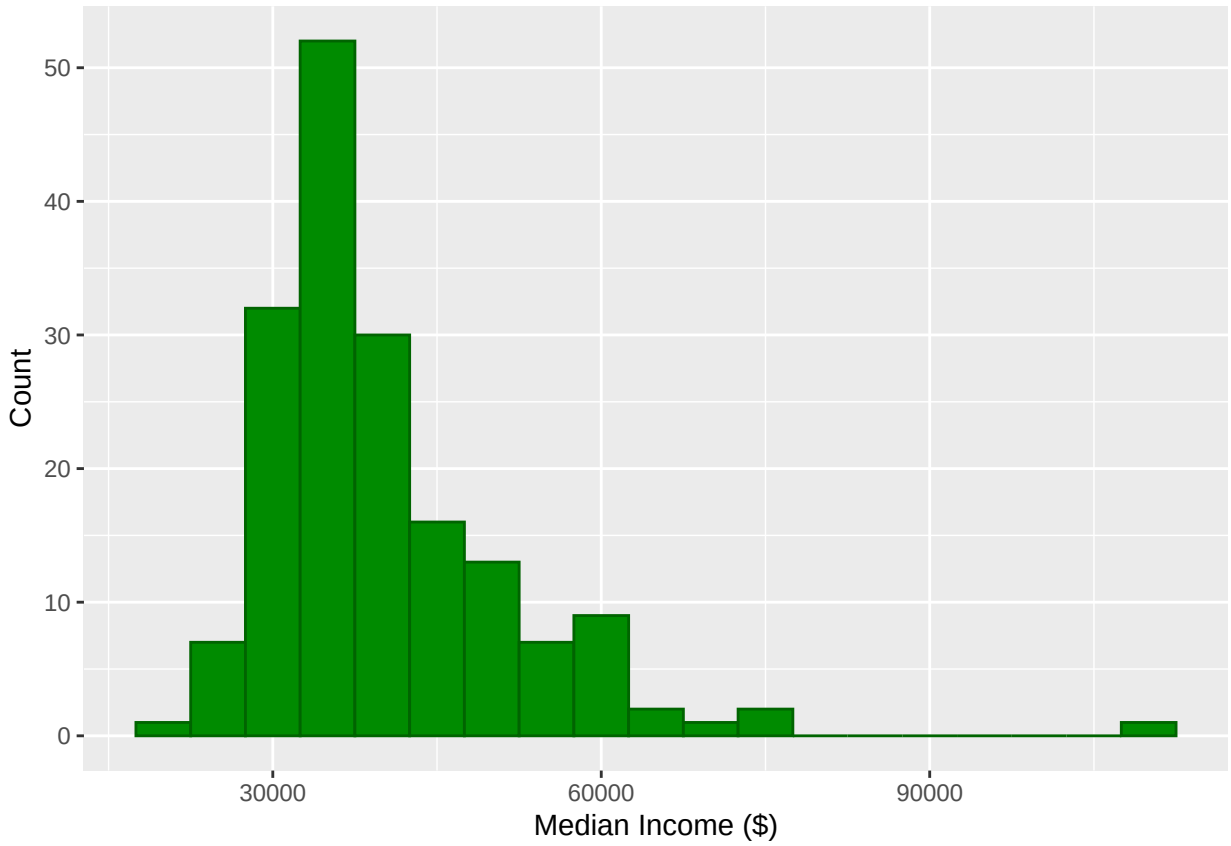
```
## # A tibble: 3 x 3
##   major                                          total sharewomen
##   <chr>                                          <int>      <dbl>
## 1 Early Childhood Education                      37589      0.969
## 2 Communication Disorders Sciences And Services 38279      0.968
## 3 Medical Assisting Services                     11123      0.928
```

### Exercise 2

We often choose the median, rather than the mean, to describe the typical income of a group of people because the median is not as affected by outliers, extreme values, and skewed data as the mean is.

### Exercise 3

```
ggplot(data = college_recent_grads, mapping = aes(x = median)) +
  geom_histogram(fill="green4",color="darkgreen",binwidth=5000) +
  labs(y="Count",x="Median Income ($)")
```




I went with a binwidth of 5000 because while a binwidth of 1000 showed the data with more precision it was also harder to read and interpret.

**Exercise 4**

```
college_recent_grads %>%
  summarise(min = min(median),
            q1 = quantile(median, probs = 0.25),
            med = median(median),
            q3 = quantile(median, probs = 0.75),
            max = max(median),
            iqr = q3 - q1,
            lwr = q1 - 1.5 * iqr,
            upr = q3 + 1.5 * iqr,
            outliers = sum(median < lwr | median > upr),
            mean = mean(median),
            sd = sd(median),
            skew = skewness(median))
```

```
## # A tibble: 1 x 12
##     min    q1   med    q3    max   iqr   lwr   upr outliers   mean     sd  skew
##   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <int>  <dbl>  <dbl> <dbl>
## 1 22000 33000 36000 45000 110000 12000 15000 63000        6 40151. 11470.  2.02
```

This distrubution shows a moderately wide distribution of values, with a mean of about 40,151 and noticeable right skew due to high outliers. The median (36,000) is lower than the mean, reinforcing the skew. Overall, the data are centered around the mid-30,000s but stretched upward by a few extreme values.

**Exercise 5**

```
ggplot(data = college_recent_grads, mapping = aes(x = median)) +
  geom_histogram(fill="green4",color="darkgreen",binwidth=5000) +
  facet_wrap(~major_category) +
  labs(y="Count",x="Median Income ($)")
```


**Exercise 6**

```
college_recent_grads %>%
  group_by(major_category) %>%
  summarise(median = median(median)) %>%
  arrange(desc(median)) %>%
  top_n(1)
```

```
## Selecting by median
```

```
## # A tibble: 1 x 2
##   major_category median
##   <chr>           <dbl>
## 1 Engineering     57000
```

Engineering has the highest typical median income.

**Exercise 7**

```
college_recent_grads %>%
  count(major_category) %>%
```

```
  arrange(n) %>%
  slice_min(order_by = n, n=1)
```

```
## # A tibble: 1 x 2
##   major_category        n
##   <chr>             <int>
## 1 Interdisciplinary     1
```

**Exercise 8**

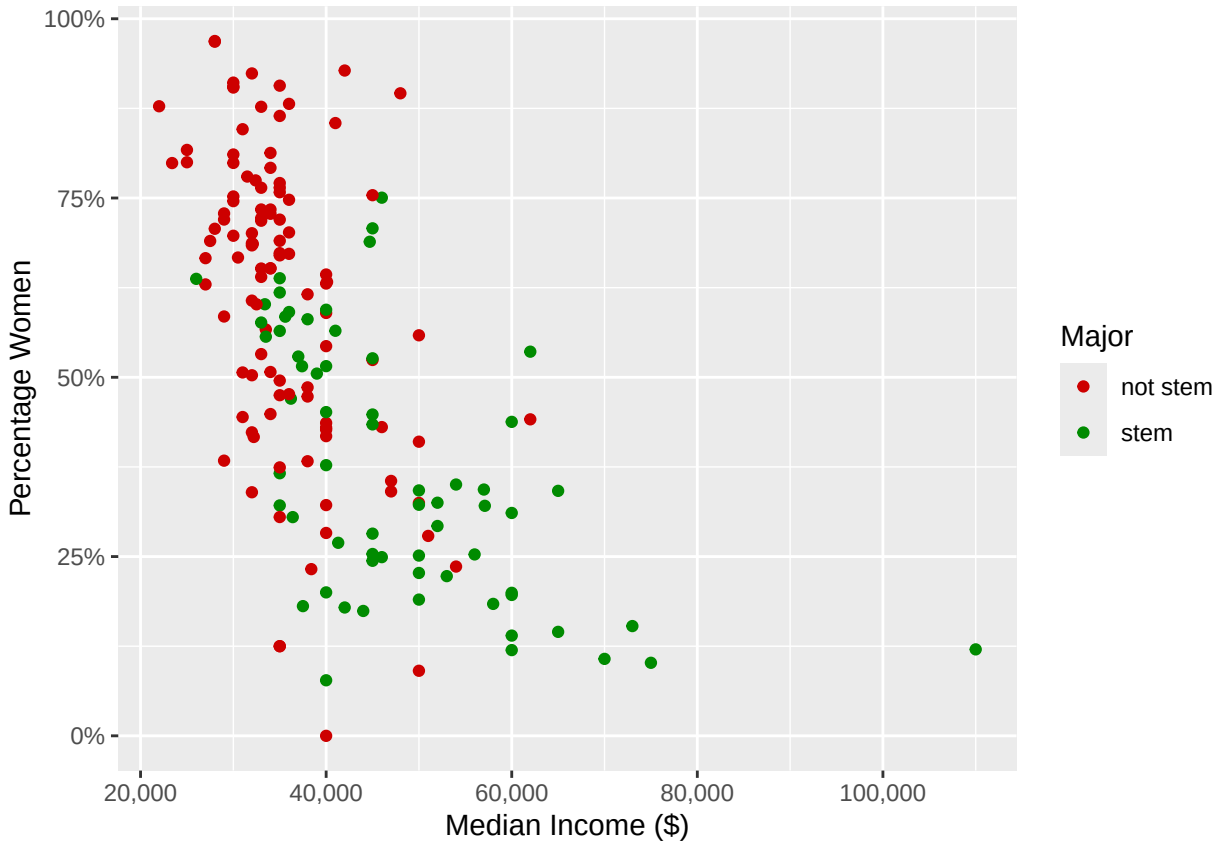
```
stem_categories <- c("Biology & Life Science",
                     "Computers & Mathematics",
                     "Engineering",
                     "Physical Sciences")

college_recent_grads <- college_recent_grads %>%
  mutate(major_type = ifelse(major_category %in% stem_categories, "stem", "not stem"))

college_recent_grads %>%
  filter(
    major_type == "stem",
    median <= median(median)
  ) %>%
  arrange(desc(median)) %>%
  select(major,p25th,median,p75th)
```

```
## # A tibble: 11 x 4
##    major                                p25th median p75th
##    <chr>                                <dbl>  <dbl> <dbl>
##  1 Geosciences                          21000  36000 41000
##  2 Environmental Science                25000  35600 40200
##  3 Multi-Disciplinary Or General Science 24000  35000 50000
##  4 Physiology                           20000  35000 50000
##  5 Communication Technologies           25000  35000 45000
##  6 Neuroscience                         30000  35000 44000
##  7 Atmospheric Sciences And Meteorology 28000  35000 50000
##  8 Miscellaneous Biology                23000  33500 48000
##  9 Biology                              24000  33400 45000
## 10 Ecology                              23000  33000 42000
## 11 Zoology                              20000  26000 39000
```

**Exercise 9**

```
ggplot(college_recent_grads,aes(x=median,y=sharewomen,color=major_type)) +
  geom_point() +
  scale_color_manual(values=c("stem"="green4","not stem"="red3")) +
  scale_x_continuous(labels=comma) +
  scale_y_continuous(labels=percent) +
  labs(x="Median Income ($)",y="Percentage Women",color="Major")
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

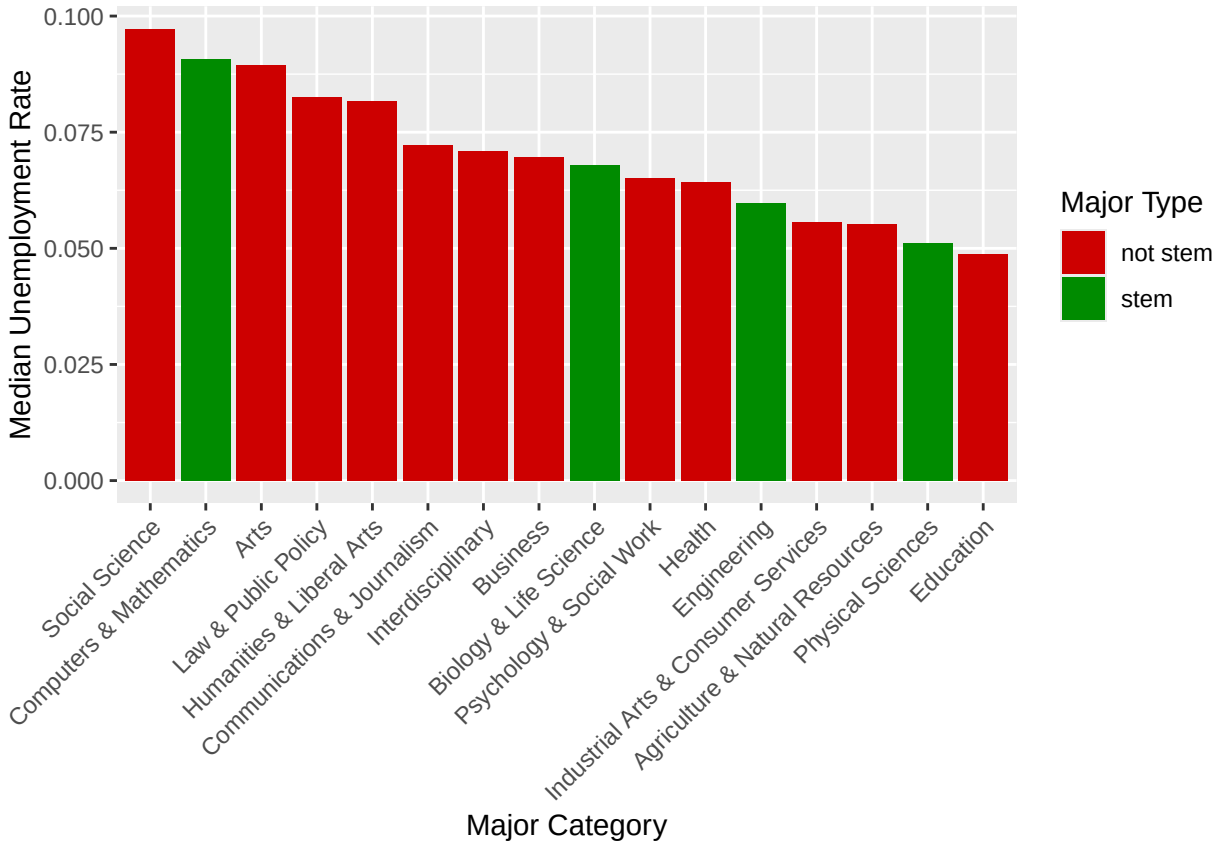



The majors that are not stem majors typically have a low median income. The higher percentage of women a major has the more likely it is that it is not a stem major and has a low median income.

**Exercise 10**

Which categories of majors have the highest and lowest unemployment rates?

```
college_recent_grads %>%
  group_by(major_category, major_type) %>%
  summarise(median_unemp = median(unemployment_rate), .groups = "drop") %>%
  ggplot(aes(x = reorder(major_category, -median_unemp),
             y = median_unemp,
             fill = major_type)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("stem" = "green4", "not stem" = "red3")) +
  labs(x = "Major Category", y = "Median Unemployment Rate", fill = "Major Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


0.100
0.075
0.050
0.025
0.000
Median Unemployment Rate
Social Science
Computers & Mathematics
Arts
Law & Public Policy
Humanities & Liberal Arts
Communications & Journalism
Interdisciplinary
Business
Biology & Life Science
Psychology & Social Work
Health
Engineering
Industrial Arts & Consumer Services
Agriculture & Natural Resources
Physical Sciences
Education
Major Category
Major Type
not stem
stem