# Lab 02 - Exploring college majors

Add the date here. Due Thu, Jan 30 at 11:59p

You may knit this document to see what the template looks like. ***When turning a document in on Gradescope, remember to knit to .pdf** and turn in that .pdf document. For Lab 02, we are not telling you when to commit – it is up to you to commit at appropriate intervals with meaningful commit comments. Be sure to commit **at least three times** during this lab.

Delete these comments in your final version of the lab you turn in.

## Packages

```r
library(tidyverse)
library(fivethirtyeight)
```

## Exercise 1

Using options is a better option than mutating as mutating manipulates the data itself whereas options simply manipulates the way the original data is displayed.

```r
options(digits = 2)
```

```r
college_recent_grads %>%
  arrange(unemployment_rate) %>%
  select(rank, major, unemployment_rate)
```

```
## # A tibble: 173 x 3
##     rank major                                      unemployment_rate
##    <int> <chr>                                                  <dbl>
##  1    53 Mathematics And Computer Science                     0
##  2    74 Military Technologies                                0
##  3    84 Botany                                               0
##  4   113 Soil Science                                         0
##  5   121 Educational Administration And Supervision           0
##  6    15 Engineering Mechanics Physics And Science            0.00633
##  7    20 Court Reporting                                      0.0117
##  8   120 Mathematics Teacher Education                        0.0162
##  9     1 Petroleum Engineering                                0.0184
## 10    65 General Agriculture                                  0.0196
## # ... with 163 more rows
```

## Exercise 2

```r
college_recent_grads %>%
  arrange(desc(sharewomen)) %>%
```

```r
  select(major, women, sharewomen) %>%
  slice(1:3)
```

```
## # A tibble: 3 x 3
##   major                                         women sharewomen
##   <chr>                                         <int>      <dbl>
## 1 Early Childhood Education                     36422      0.969
## 2 Communication Disorders Sciences And Services 37054      0.968
## 3 Medical Assisting Services                    10320      0.928
```

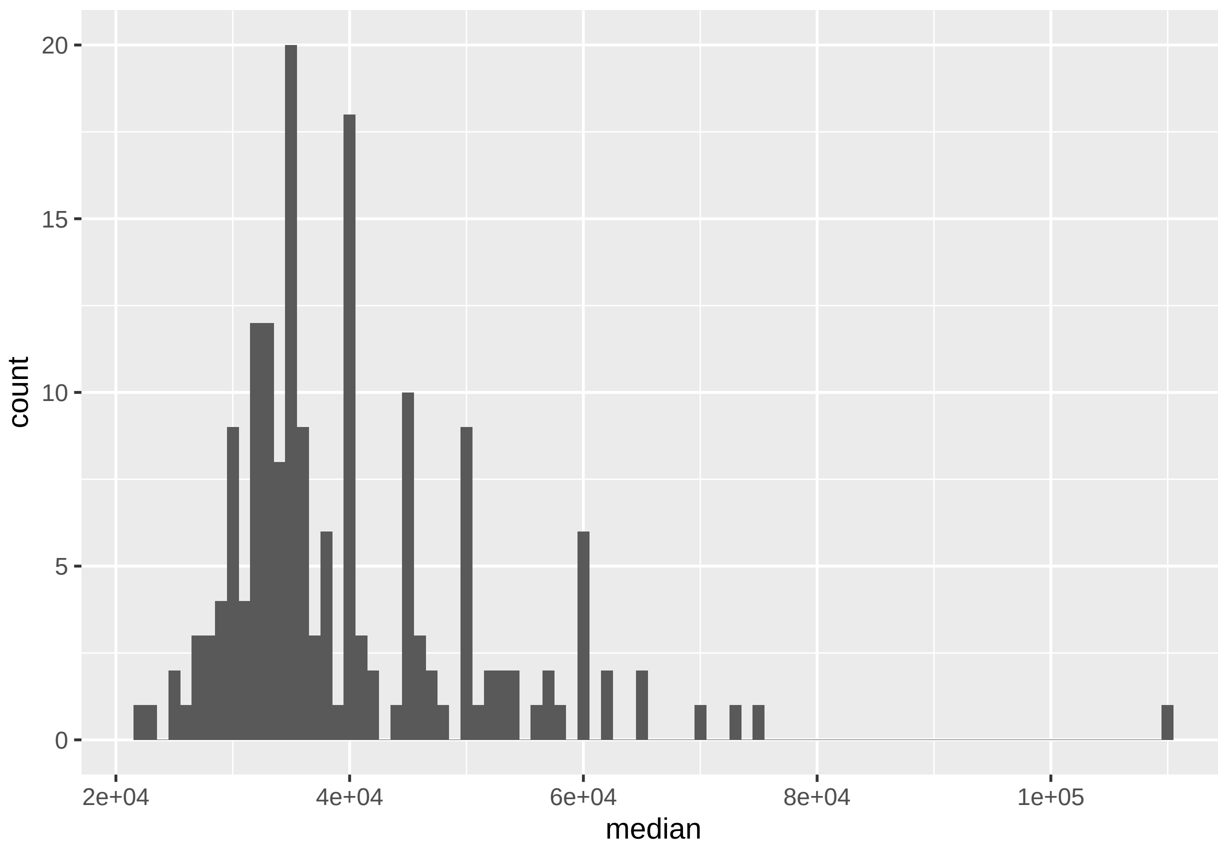
### Exercise 3

Using mean to describe the typical income of a group of people could possibly be drastically skewed. This is because one outlier on either end can drastically increase or decrease the mean. Using the median gives a more solid and accurate look at the middle of your set of data.

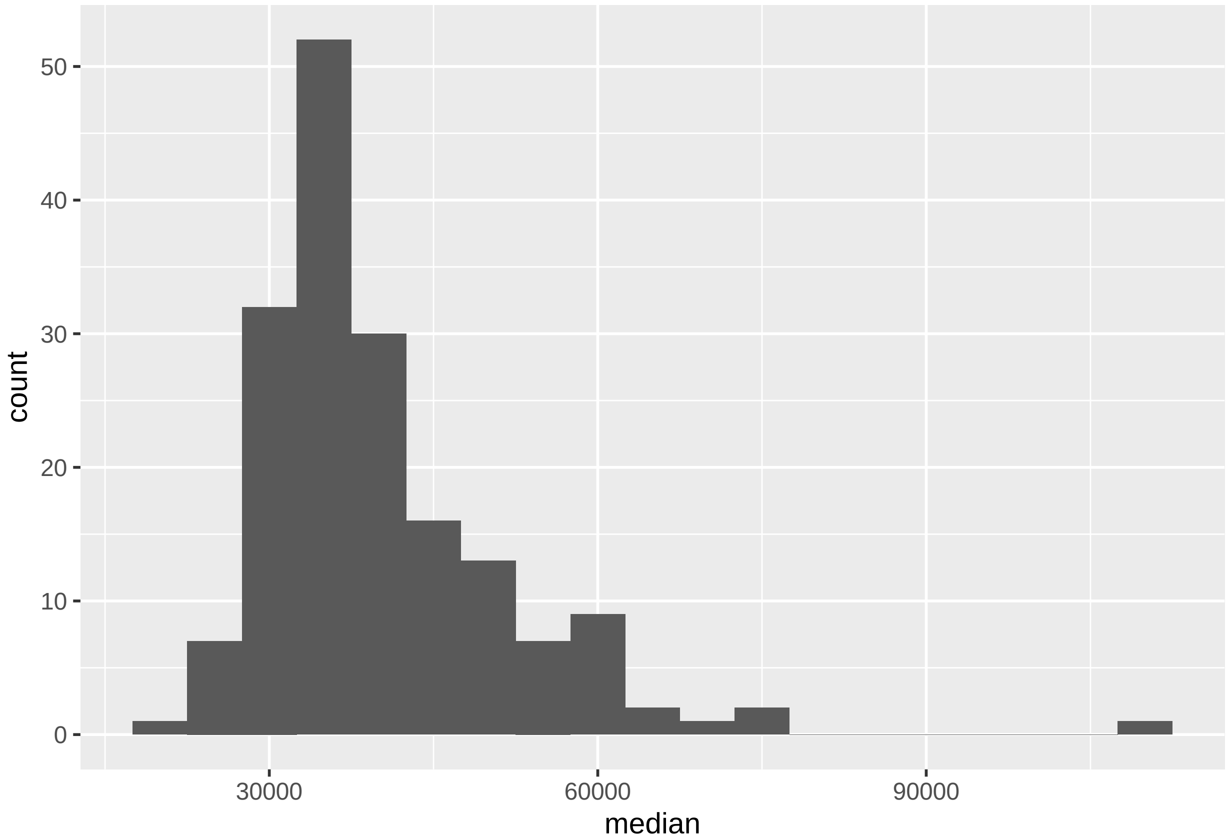
### Exercise 4

Using binwidth of 5000 would be more effective as we're looking for the best summary of medians rather than specific data, as shown in the histogram with binwidth 1000. People typically look for ranges when looking at income levels, rather than by each thousand. This is best shown with the histogram of binwidth 5000.

```r
ggplot(data = college_recent_grads, mapping = aes(x = median)) +
  geom_histogram(binwidth = 1000)
```



```r
ggplot(data = college_recent_grads, mapping = aes(x = median)) +
  geom_histogram(binwidth = 5000)
```




### Exercise 5

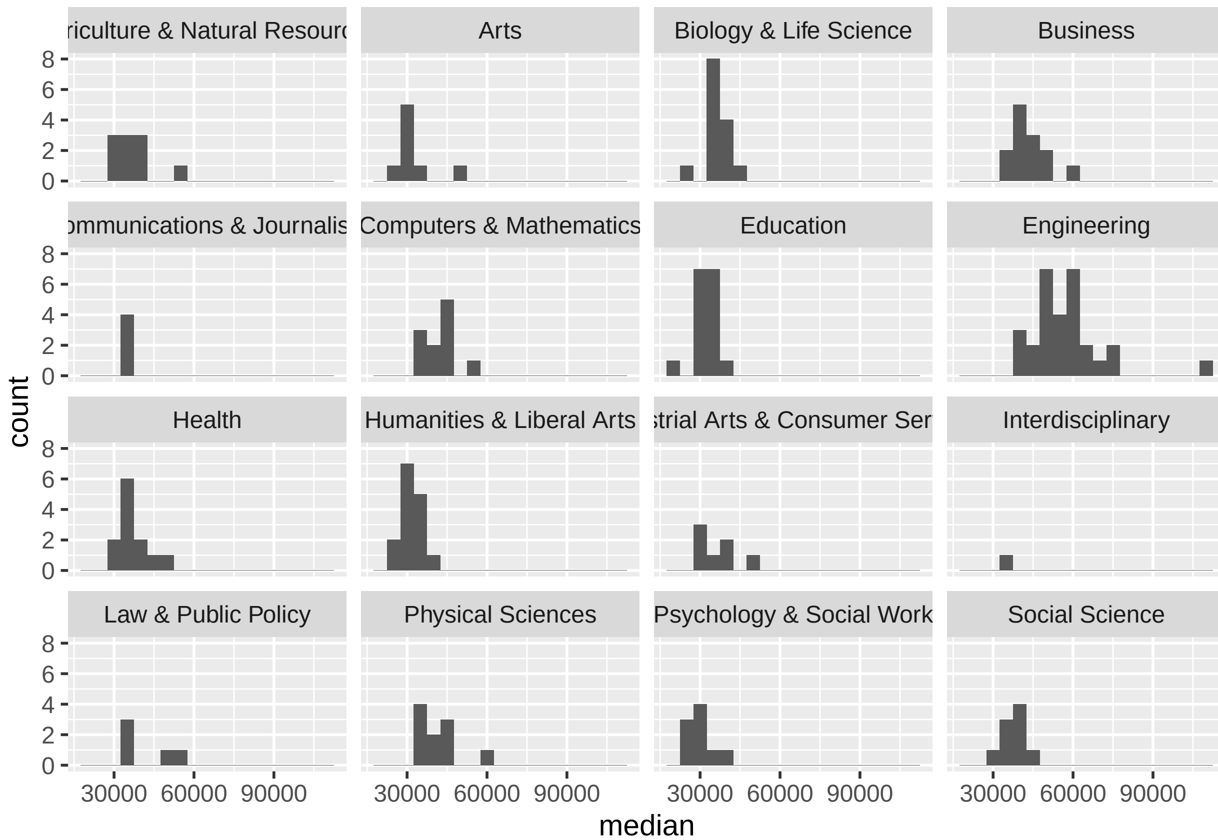
The median would be the best to describe this data as there is data above 100,000 that could potentially skew any other data representations. Median gives the most “unbiased” snapshot at income.

```r
college_recent_grads %>%
  summarise(min  = min(median),
            max  = max(median),
            mean = mean(median),
            med  = median(median),
            sd   = sd(median),
            q1   = quantile(median, probs = 0.25),
            q3   = quantile(median, probs = 0.75))
```

```
## # A tibble: 1 x 7
##     min    max   mean   med     sd    q1    q3
##   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 22000 110000 40151. 36000 11470. 33000 45000
```

### Exercise 6

```r
ggplot(data = college_recent_grads, mapping = aes(x = median)) +
  geom_histogram(binwidth = 5000) +
  facet_wrap( ~ major_category, ncol = 4)
```




### Exercise 7

```r
college_recent_grads %>%
  group_by(major_category) %>%
  summarise(median2 = median(median)) %>%
  arrange(desc(median2))
```

```
## # A tibble: 16 x 2
##    major_category                     median2
##    <chr>                                <dbl>
##  1 Engineering                          57000
##  2 Computers & Mathematics              45000
##  3 Business                             40000
##  4 Physical Sciences                    39500
##  5 Social Science                       38000
##  6 Biology & Life Science               36300
##  7 Law & Public Policy                  36000
##  8 Agriculture & Natural Resources      35000
##  9 Communications & Journalism          35000
## 10 Health                               35000
## 11 Industrial Arts & Consumer Services  35000
```

```
## 12 Interdisciplinary                     35000
## 13 Education                             32750
## 14 Humanities & Liberal Arts             32000
## 15 Arts                                  30750
## 16 Psychology & Social Work              30000
```

## Exercise 8

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

```
## # A tibble: 16 x 2
##    major_category                          n
##    <chr>                               <int>
##  1 Interdisciplinary                       1
##  2 Communications & Journalism             4
##  3 Law & Public Policy                     5
##  4 Industrial Arts & Consumer Services     7
##  5 Arts                                    8
##  6 Psychology & Social Work                9
##  7 Social Science                          9
##  8 Agriculture & Natural Resources        10
##  9 Physical Sciences                      10
## 10 Computers & Mathematics                11
## 11 Health                                 12
## 12 Business                               13
## 13 Biology & Life Science                 14
## 14 Humanities & Liberal Arts              15
## 15 Education                              16
## 16 Engineering                            29
```

## Exercise 9

```
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)) %>%
  select(major, median, p25th, p75th) %>%
  arrange(desc(median))
```

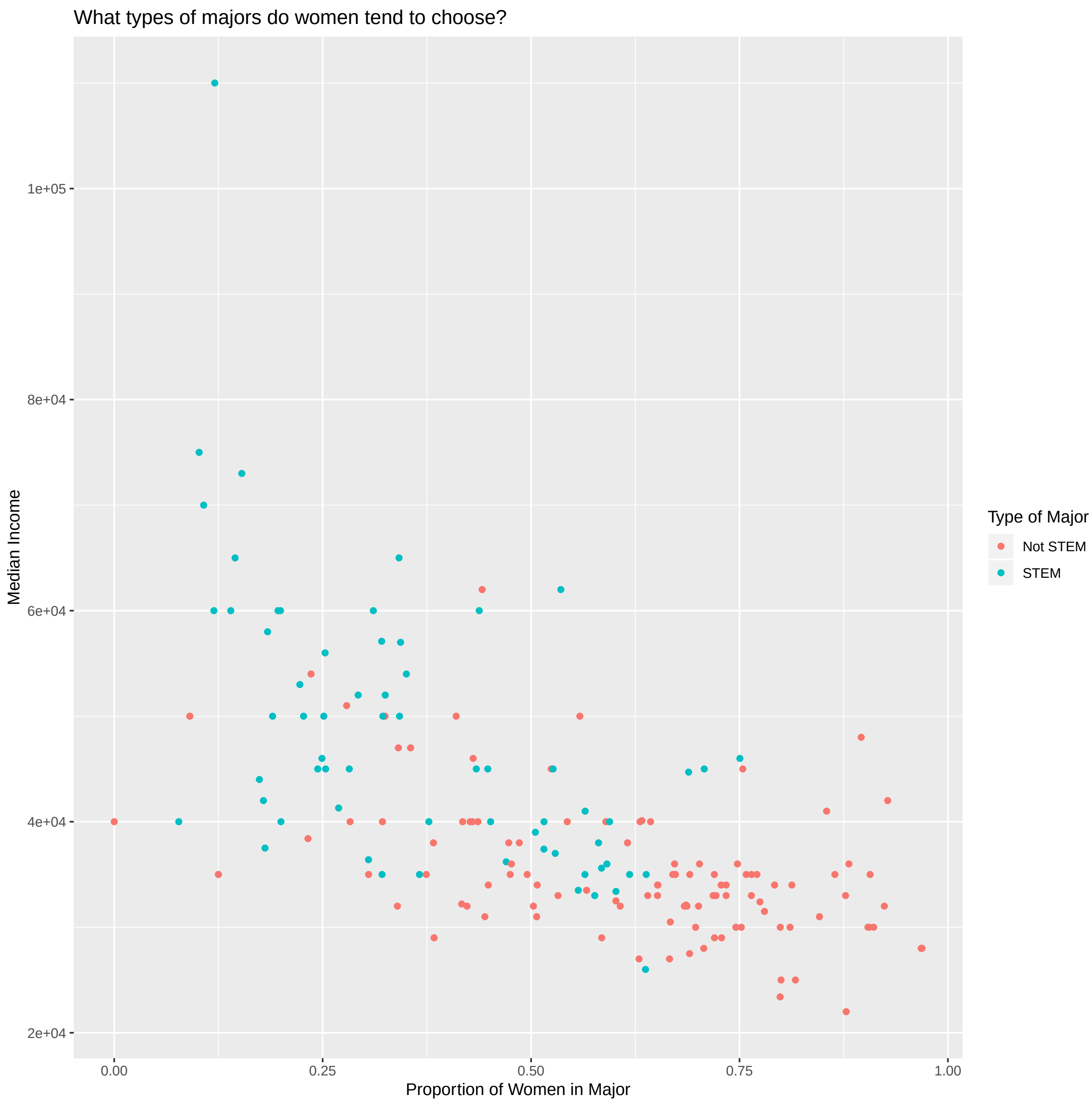
```
## # A tibble: 11 x 4
##    major                                  median p25th p75th
##    <chr>                                   <dbl> <dbl> <dbl>
```

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

### Exercise 10

```
ggplot(data = college_recent_grads, mapping = aes(x = sharewomen, y = median, color = major_type)) +
  geom_point()+
  labs(title = "What types of majors do women tend to choose?",
       color = "Type of Major", x = "Proportion of Women in Major", y = "Median Income")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




This graph is able to convey a lot of information with little complexity. First, this scatterplot shows that STEM jobs have lower proportions of women compared to non-STEM jobs. Secondly, those in STEM jobs typically have a higher median income than those not in STEM jobs. When taking both of these factors into consideration, women tend to have lower median incomes than men, seeing as men are in higher proportions for higher paying STEM jobs.