R Notebook

```
library(tidyverse)
```

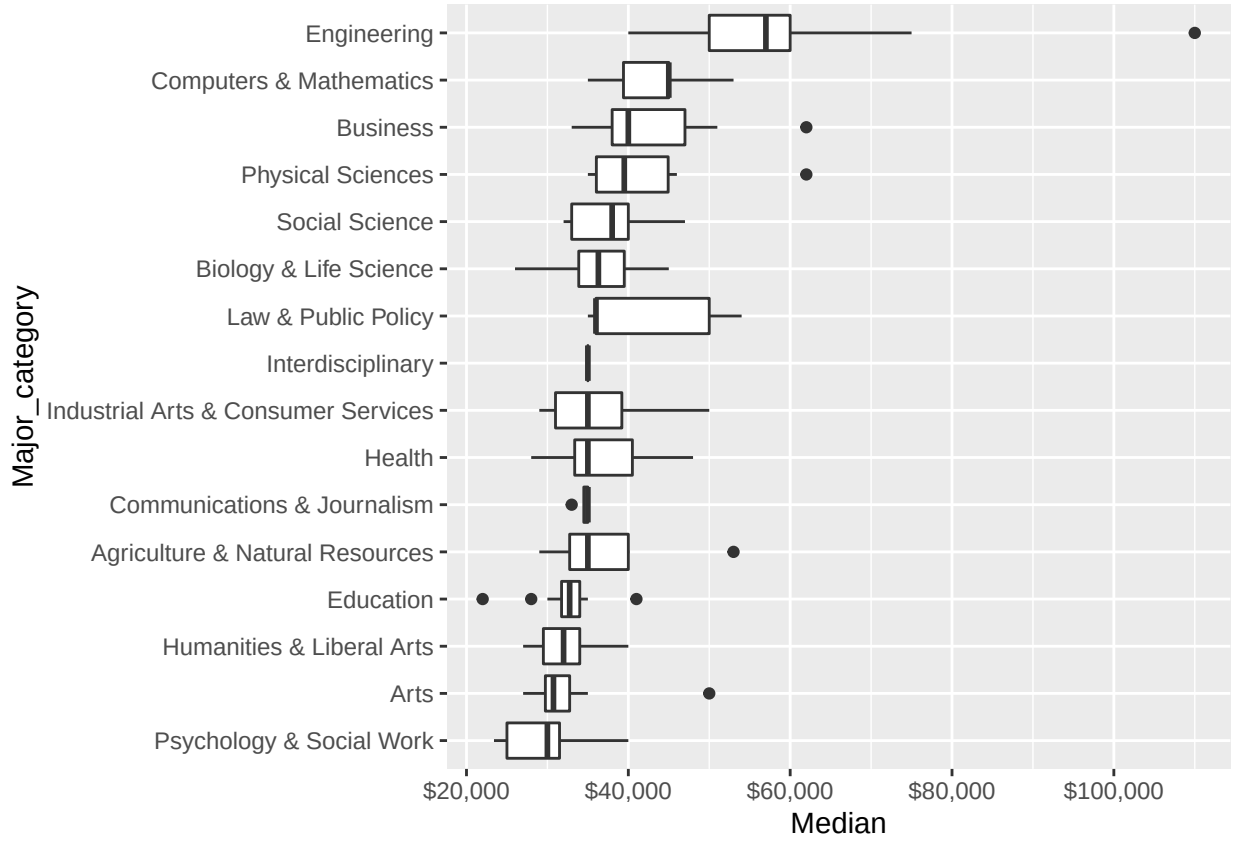
```
## -- Attaching packages ------------------------------------------------------------------ tidyverse 1.3

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts --------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
college_grad <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Major = col_character(),
##   Major_category = col_character()
## )

## See spec(...) for full column specifications.
```

```
#view(college_grad)
```

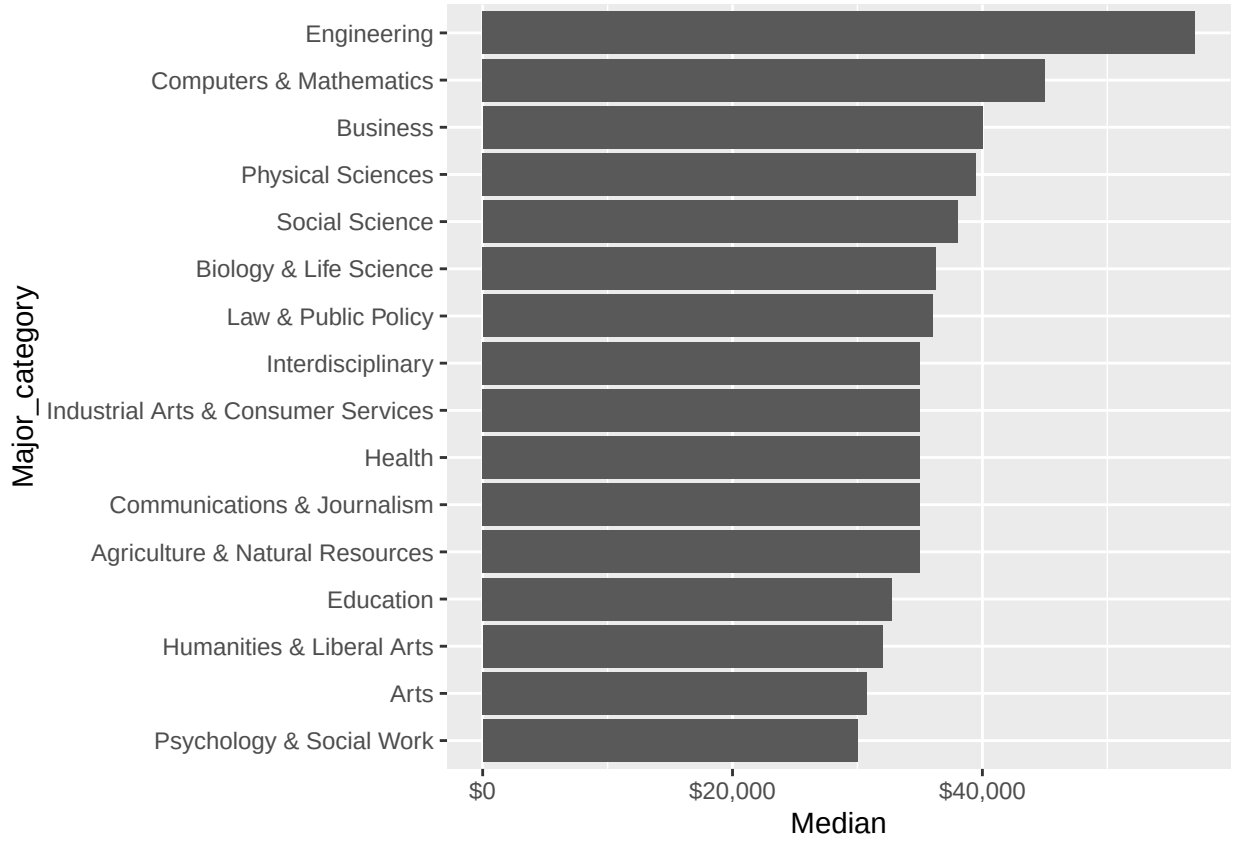
```
college_grad %>%
    mutate(Major_category = fct_reorder(Major_category, Median)) %>%
    ggplot(aes(Major_category, Median)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```



Here is the categories that make a lot money upon graduation

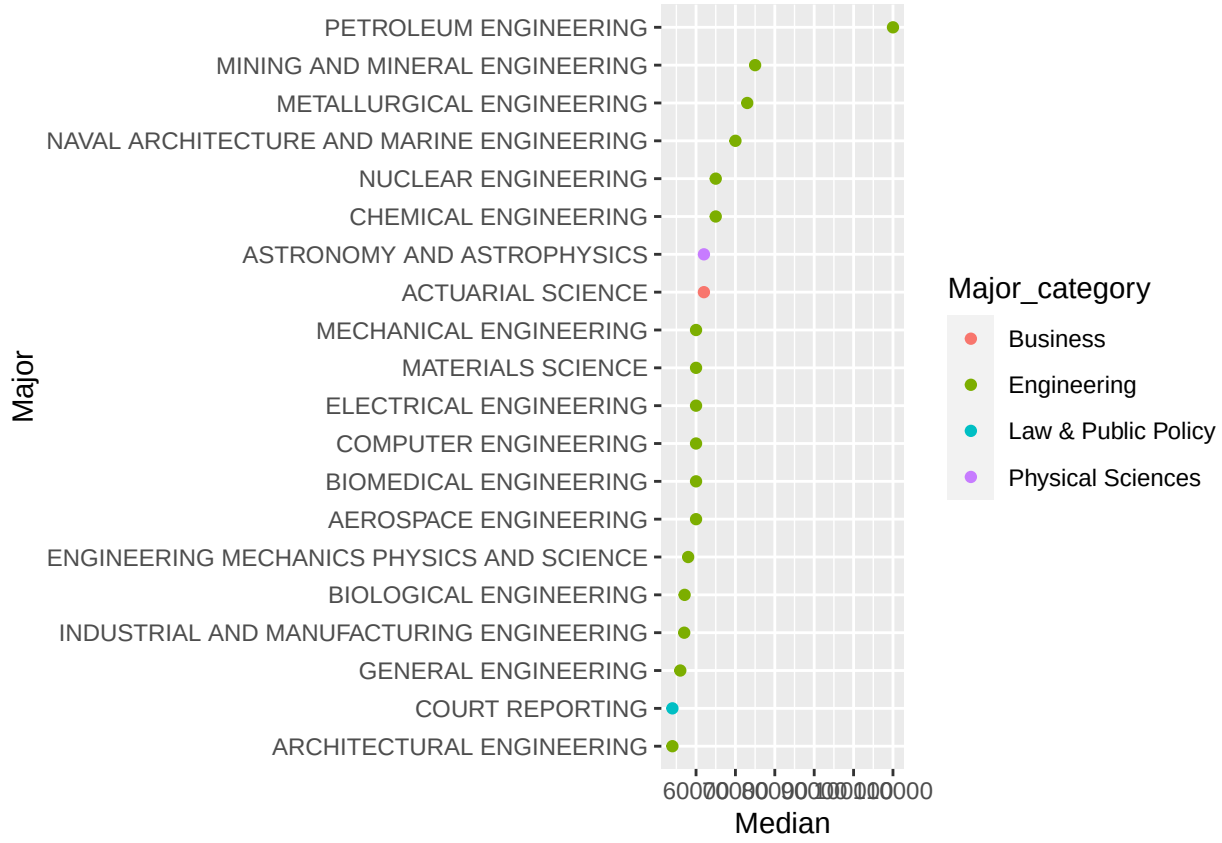
```
college_grad %>%
    group_by(Major_category) %>%
    summarize(Median = median(Median)) %>%
    mutate(Major_category = fct_reorder(Major_category, Median)) %>%
    ggplot(aes(Major_category, Median)) +
    geom_col() +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



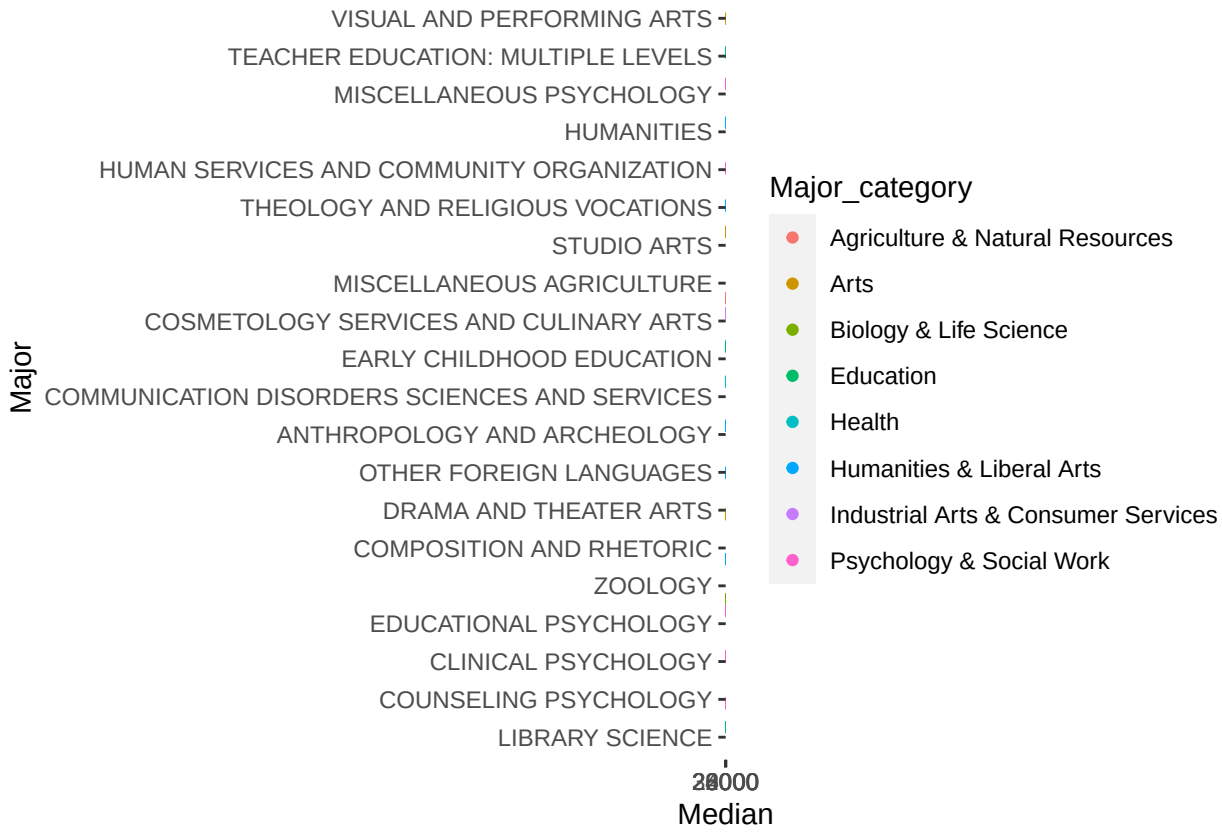
Here, we will find the highest top earning majors

```
Majors <- college_grad %>%
    arrange(desc(Median)) %>%
    select(Major, Major_category, Median, P25th, P75th, Sample_size)  %>%
    mutate(Major = fct_reorder(Major, Median))

Majors %>% head(20) %>%
    ggplot(aes(Major, Median,, color = Major_category)) +
    geom_point() +
    coord_flip()
```



The lowest earning Majors

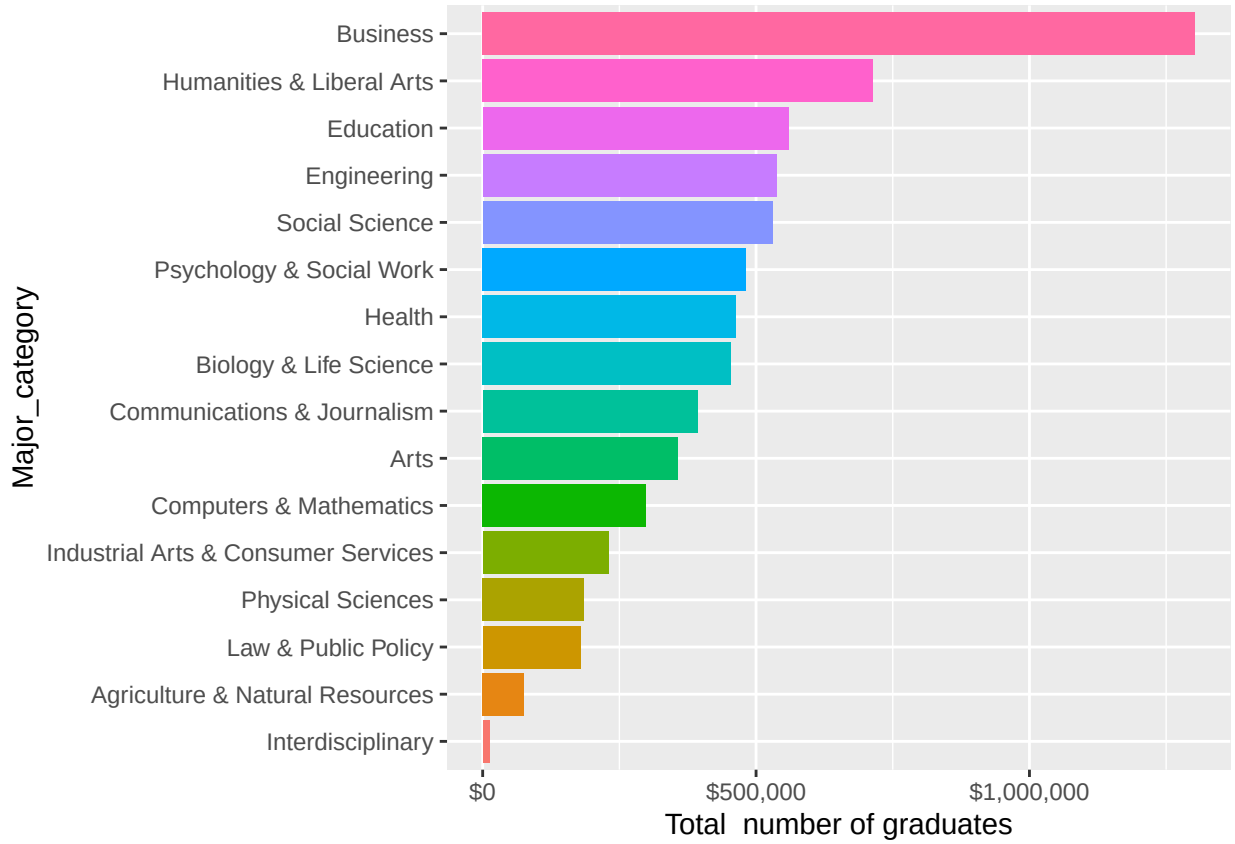
```
college_grad %>%
    arrange(desc(Median)) %>%
    select(Major, Major_category, Median, P25th, P75th)  %>%
    tail(20) %>%

    mutate(Major = fct_reorder(Major, Median)) %>%
    ggplot(aes(Major, Median,, color = Major_category)) +
    geom_jitter() +
    coord_flip()
```




```
# Majors  %>%
#     ggplot(aes(Sample_size, Median)) +
#     geom_point() +
#     scale_x_log10()

# install.packages("tm")           # for text mining
# install.packages("SnowballC")    # for text stemming
# install.packages("wordcloud")    # word-cloud generator
# install.packages("RColorBrewer") # color palettes

# Load the packages
# library("tm")
# library("SnowballC")
# library("wordcloud")
# library("RColorBrewer")
# wordcloud(words = Majors$Major_category,
#           freq = Majors$Median,
#           min.freq = 1,
#           max.words =200,
#           random.order = TRUE,
#           rot.per = 0.35,
#           colors = brewer.pal(8, "Dark2"))
```

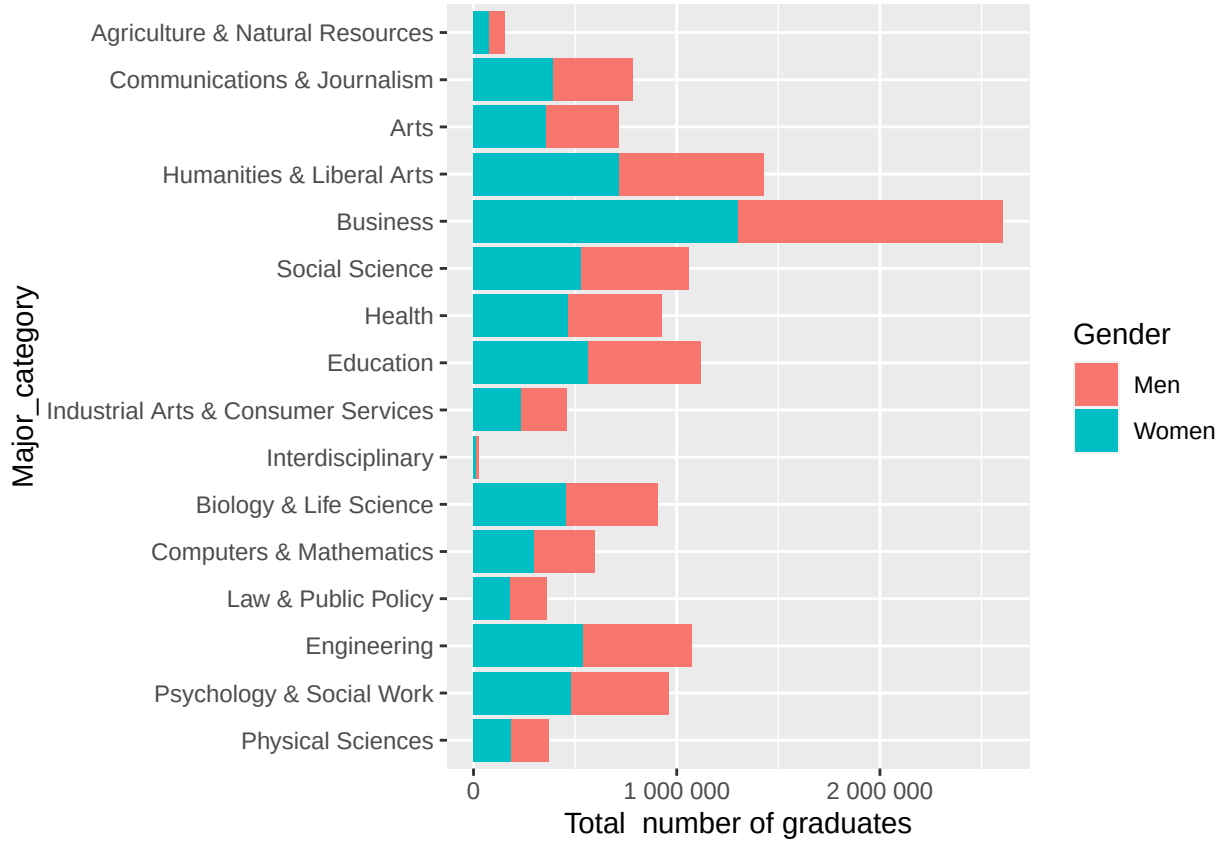
Most common majors :

```
college_grad %>%
    count(Major_category, wt = Total, sort = TRUE) %>%
    mutate(Major_category = fct_reorder(Major_category,n)) %>%
    ggplot(aes(Major_category, n, fill = Major_category)) +
    theme(legend.position = "none") +
    geom_col() +
    coord_flip() +
    scale_y_continuous(labels = scales::dollar_format()) +
    labs(y = "Total  number of graduates")
```



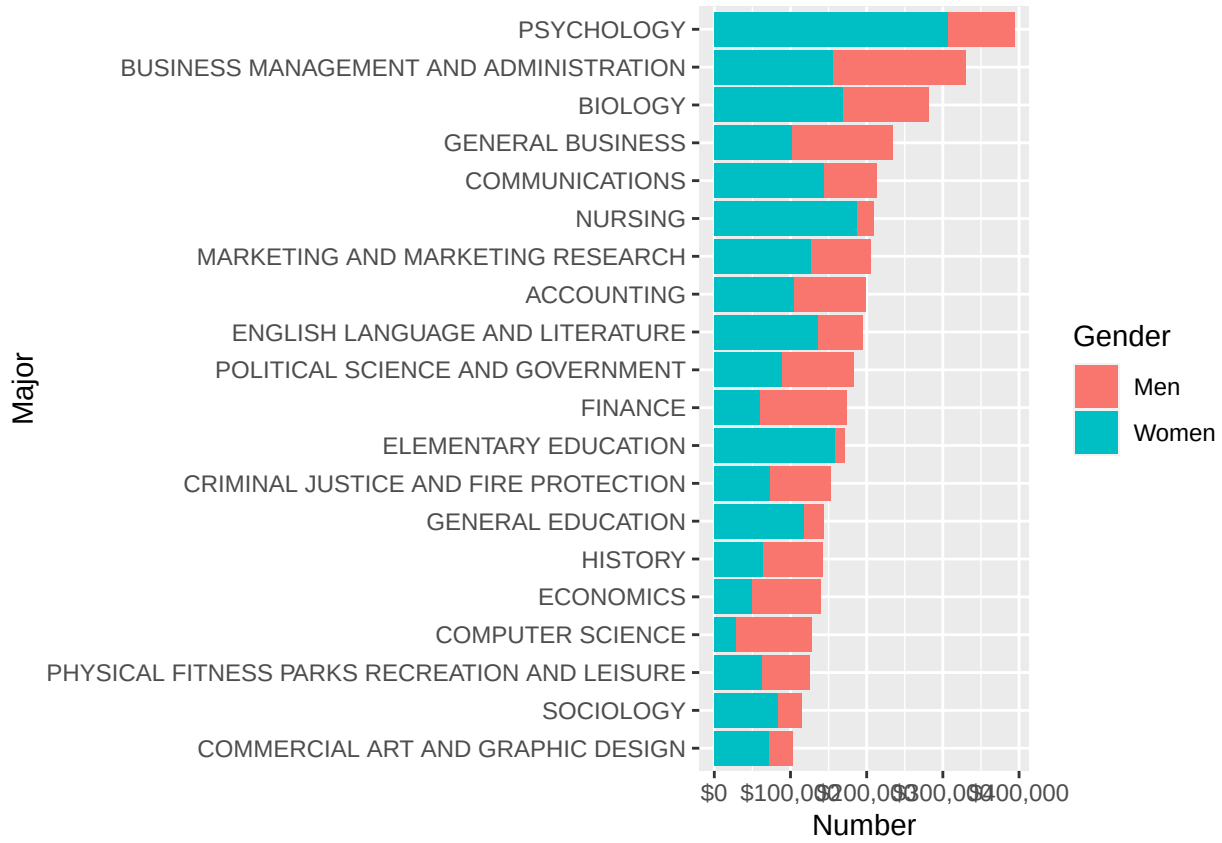

```
college_grad %>%
    mutate(Major_category = fct_reorder(Major_category, Total)) %>%
    #count(Major_category, wt = Total, sort = TRUE) %>%
    gather(Gender, Number, Men, Women) %>%
    arrange(desc(Major_category)) %>%

    ggplot(aes(Major_category,Total, fill = Gender)) +
    #theme(legend.position = "none") +
    geom_col() +
    coord_flip() +
    scale_y_continuous(labels = scales::number_format()) +
    labs(y = "Total  number of graduates")
```

```
## Warning: Removed 2 rows containing missing values (position_stack).
```



which gender earn more money based on top Major

```
college_grad %>%
    arrange(desc(Total)) %>%
    head(20) %>%
    mutate(Major = fct_reorder(Major, Total)) %>%
    gather(Gender, Number, Men, Women) %>%
    select(Major, Gender, Number) %>%


    ggplot(aes(Major, Number, fill = Gender)) +
    geom_col() +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```




```
college_grad %>%
    group_by(Major_category, Median) %>%
    summarize_at(vars(Total, Men, Women), sum, na.rm=TRUE) %>%
    mutate(ShareWomen = Women/Total) %>%
    arrange(desc(ShareWomen)) %>%

    ggplot(aes(ShareWomen, Median, color = Major_category)) +
    geom_jitter() +
    scale_y_continuous(labels = scales::dollar_format())
```

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

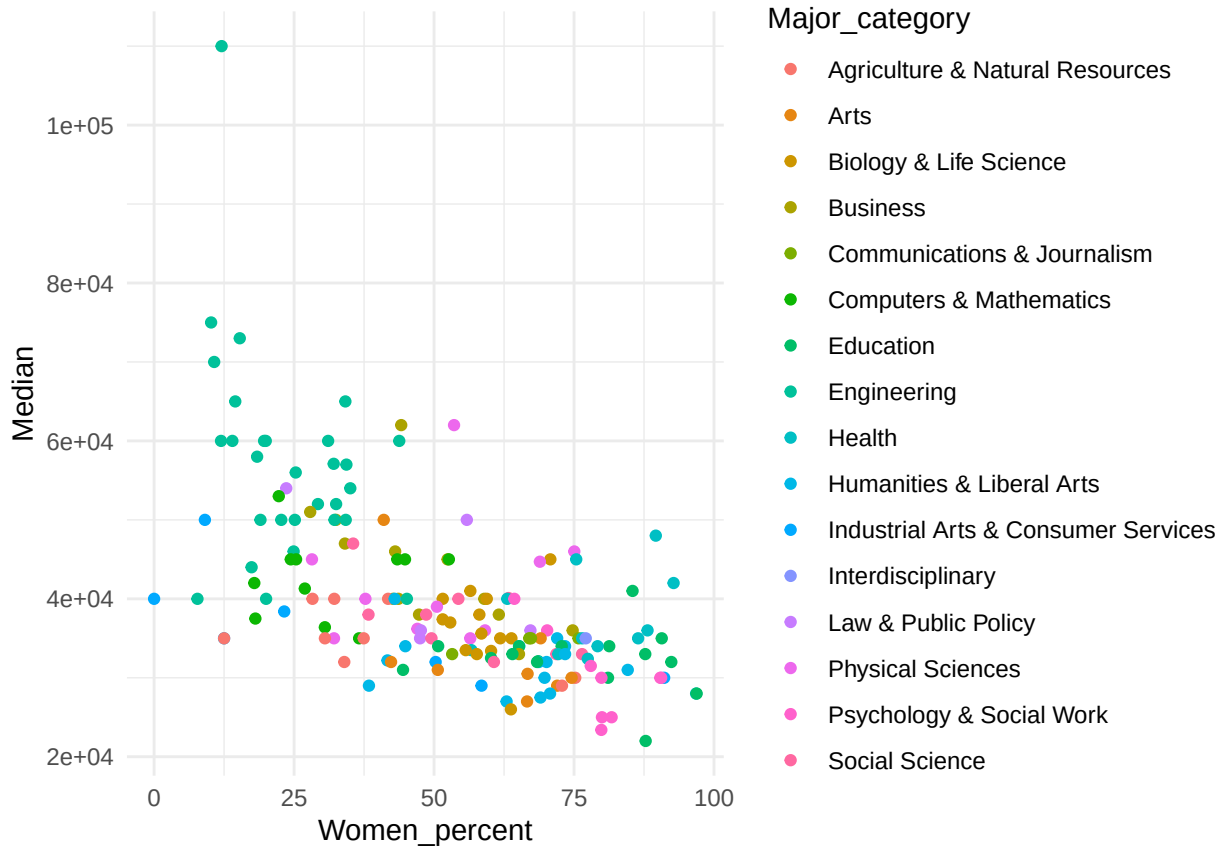
Total Major in each major's category

```
college_grad %>%
    select(Major, Major_category, Total, ShareWomen, Sample_size, Median) %>%
    add_count(Major_category) %>%
    filter(n >= 10) %>%
    count(Major_category) %>%
    arrange(desc(n))
```

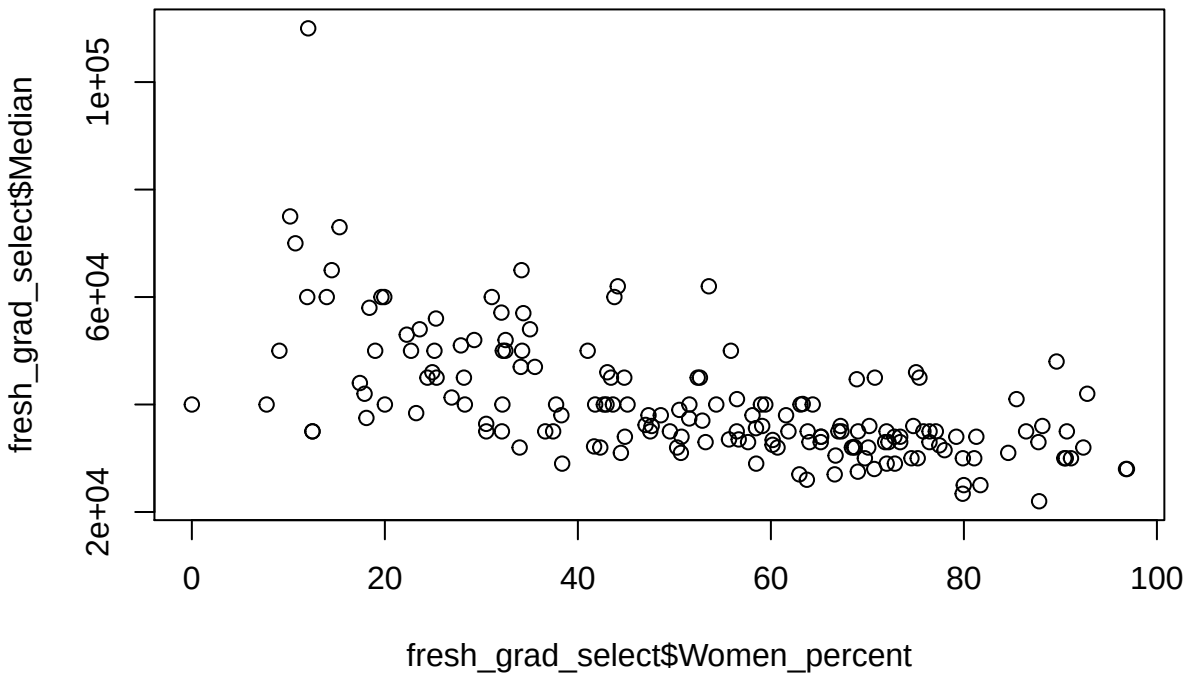
```
## # A tibble: 9 x 2
##   Major_category                     n
##   <chr>                          <int>
## 1 Engineering                       29
## 2 Education                         16
## 3 Humanities & Liberal Arts         15
## 4 Biology & Life Science            14
## 5 Business                          13
## 6 Health                            12
## 7 Computers & Mathematics           11
## 8 Agriculture & Natural Resources   10
## 9 Physical Sciences                 10
```

```r
#install.packages("esquisse")
#library(esquisse)
fresh_grad_select<-college_grad %>%
  select (Major, Total, Men, Women, Major_category,Median, Unemployment_rate)

#View(fresh_grad_select)

fresh_grad_select<- fresh_grad_select %>%  mutate(Men_percent = (Men/Total)*100)

fresh_grad_select<-fresh_grad_select %>% mutate(Women_percent=(Women/Total)*100)

#esquisse::esquisser(fresh_grad_select)

ggplot(data = fresh_grad_select) +
  aes(x = Women_percent, y = Median, color = Major_category) +
  geom_point() +
  theme_minimal()
```

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




```r
plot(fresh_grad_select$Women_percent, fresh_grad_select$Median)
```



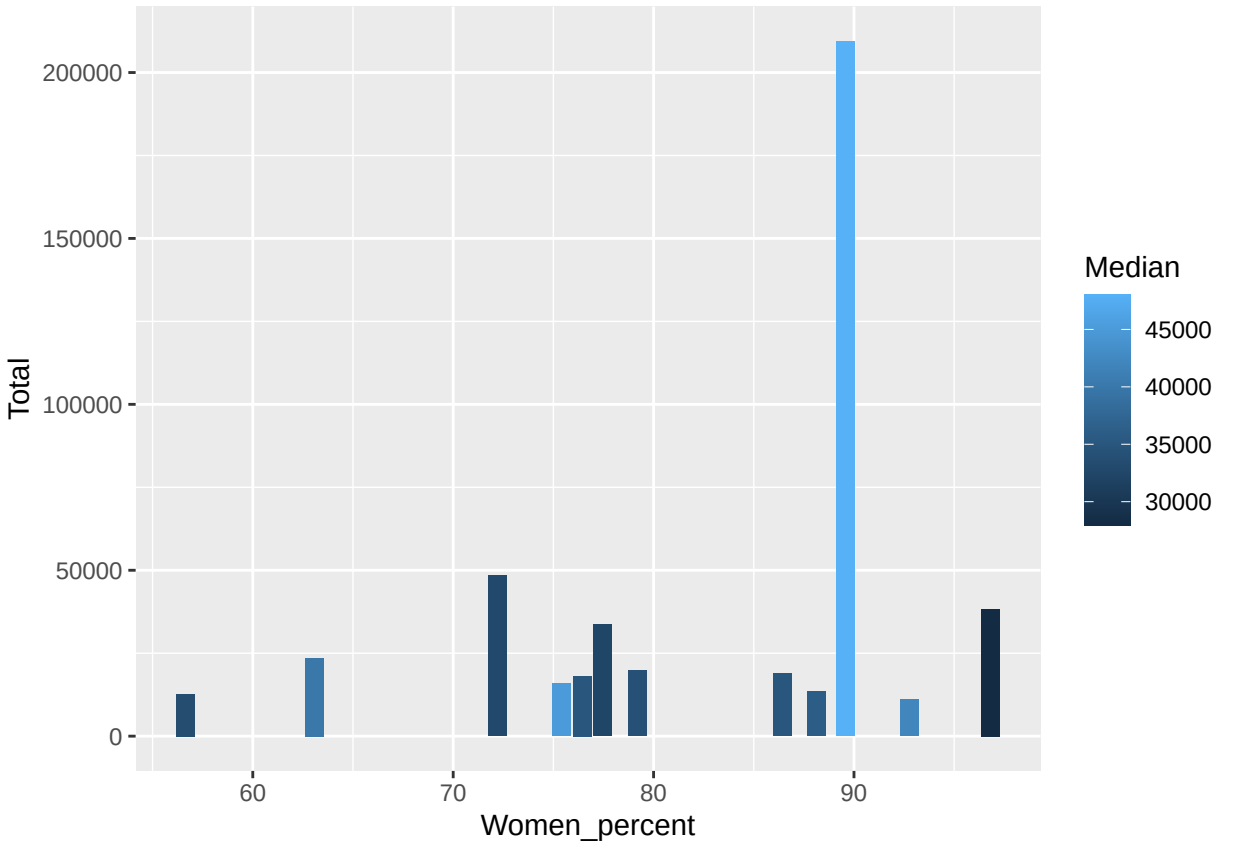

```
#View(fresh_grad_select)
```

#========== CREATING A DATA WHERE MEN_PERCENTAGE< WOMEN_PERCENTACE==========

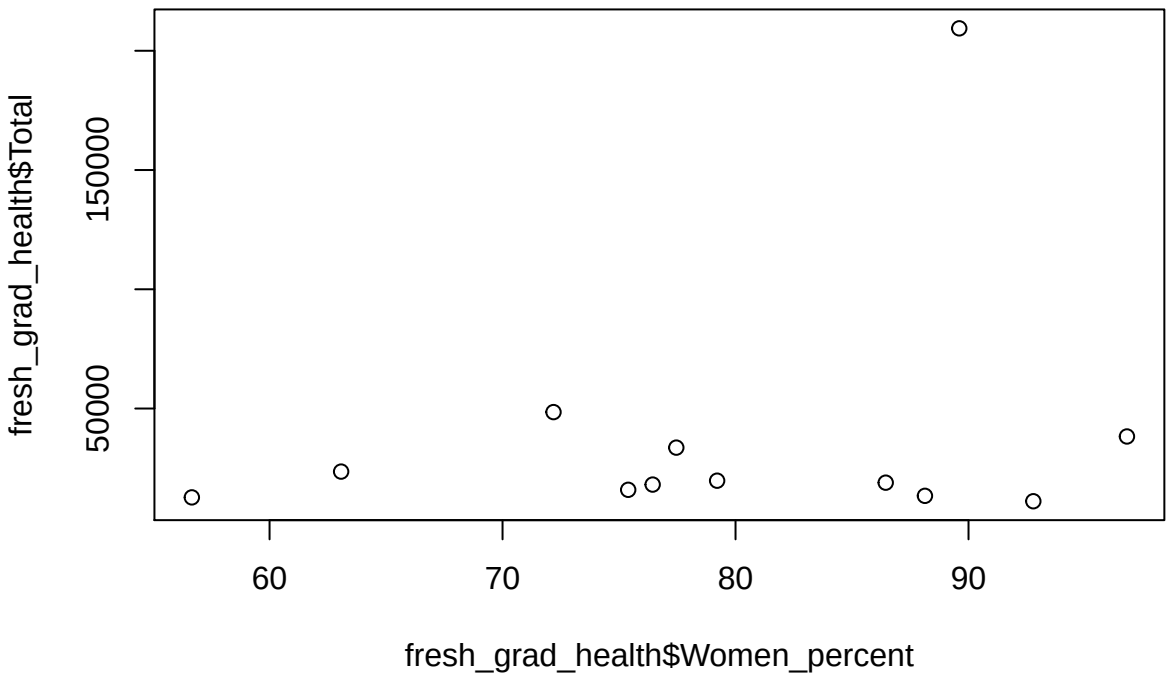
#————————FILTERING ————————————

The below graphs tell us how much money women make in health care profession

```
fresh_grad_health<- fresh_grad_select %>% filter(Major_category == 'Health')
View(fresh_grad_health)
ggplot(fresh_grad_health, aes(Women_percent, Total, fill=Median))+geom_bar(stat="identity")
```




```
plot(fresh_grad_health$Women_percent, fresh_grad_health$Total) #an outlier above 10000
```



"Majors that are either in the engineering category or have over 1,000 graduates"

```
fresh_grads_science<- fresh_grad_select %>% filter(Major_category == "Biology & Life Science" | Major_ca
```

```
## Rows: 24
## Columns: 9
## $ Major             <chr> "ASTRONOMY AND ASTROPHYSICS", "NUCLEAR, INDUSTRIA...
## $ Total             <dbl> 1792, 2116, 32142, 1762, 2418, 3831, 18300, 3635,...
## $ Men               <dbl> 832, 528, 23080, 515, 752, 1667, 7426, 1761, 894,...
## $ Women             <dbl> 960, 1588, 9062, 1247, 1666, 2164, 10874, 1874, 5...
## $ Major_category    <chr> "Physical Sciences", "Physical Sciences", "Physic...
## $ Median            <dbl> 62000, 46000, 45000, 45000, 44700, 41000, 40000, ...
## $ Unemployment_rate <dbl> 0.02116741, 0.07154047, 0.04822450, 0.08553157, 0...
## $ Men_percent       <dbl> 46.42857, 24.95274, 71.80636, 29.22815, 31.10008,...
## $ Women_percent     <dbl> 53.57143, 75.04726, 28.19364, 70.77185, 68.89992,...
```

```
#View(fresh_grads_science)
```

"Recent graduates must have a median salary above 40,000 USD More than 40 percent of graduates must be women"

```
MyMajors<-fresh_grads_science %>% filter(Median >=40000 & Women_percent >40)
View(MyMajors)

My_majors<-MyMajors %>% arrange(Unemployment_rate, desc(Median))
View(My_majors)
```

‘‘‘

Predict the median of ShareWomen