College Major & Income Analysis

Guy Gaash & Michael Rubinfeld

7/24/2020

Background

In our paper we will examine this dataset describing a survey conducted in the USA in 2011 - 2012 checking 173 college majors in terms of income, gender representations and employemnt characteristics.

Goals

in our research we will examine the following:

1. fields of study and the predicted nature of future employment and how that influences income in those fields.
2. How is gender correlated to preferences in choosing a major and the following career's characteristics.

Data Import and Tidying

first of all, we want to tidy the database by dropping NA's, majors without observations and the "Interdisciplinary" category because it's insufficiently explanetory.

Secondly, the majors are devided into 15 major categories, here they are:

```
CMI %>%
  group_by(Major_category) %>%
  summarise()
```

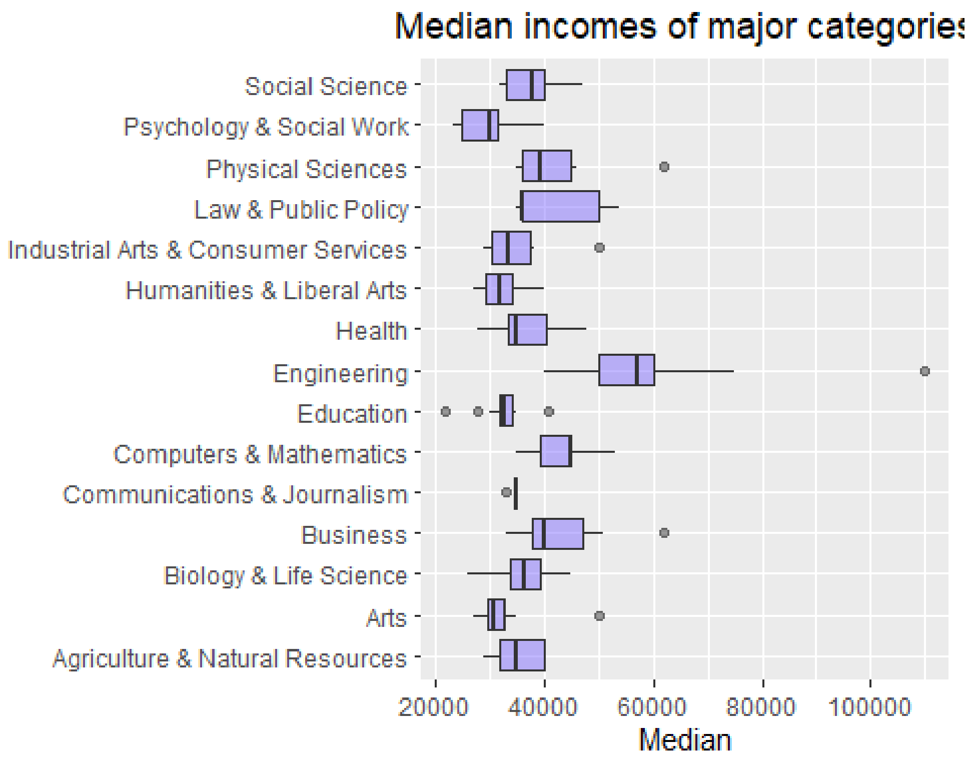
```
## # A tibble: 15 x 1
##    Major_category
##    <chr>
##  1 Agriculture & Natural Resources
##  2 Arts
##  3 Biology & Life Science
##  4 Business
##  5 Communications & Journalism
##  6 Computers & Mathematics
##  7 Education
##  8 Engineering
##  9 Health
## 10 Humanities & Liberal Arts
## 11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

Transformations and visualizations

Incomes

now, let's take a look at the median incomes (of year round workers) and their variances of each major category.

```
options(scipen = 999)

CMI %>%
  ggplot(aes(x = Major_category, y = Median, las = 2)) +
    geom_boxplot(fill = "slateblue1", alpha = 0.5) +
    labs(x = NULL, y = "Median", title = "Median incomes of major
categories") +
    theme(plot.title = element_text(hjust = 0.5)) +
    coord_flip()
```



as we can see, by far the biggest median income is in the field of Engineering. other than that, most of the others are quite similliar.

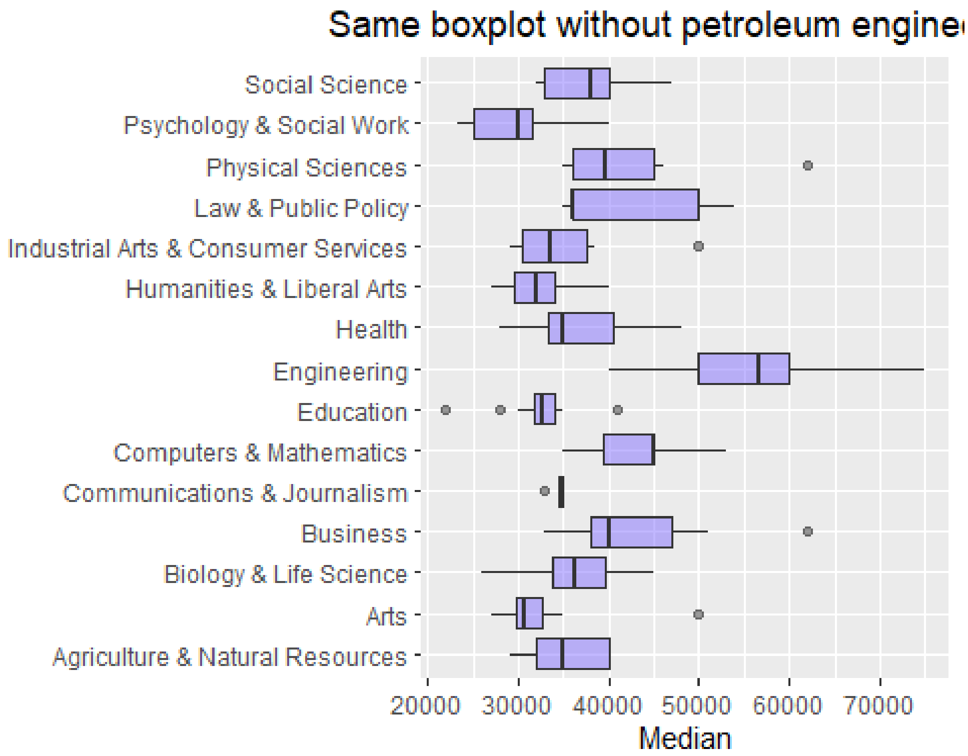
wow, Engineering has a crazy outlier! let’s check out what that major is:

```
crazy <- select(CMI, Major, Median) %>%
  arrange(desc(Median))

head(crazy, 1)

## # A tibble: 1 x 2
##   Major                     Median
##   <chr>                      <dbl>
## 1 PETROLEUM ENGINEERING 110000
```

seems like petroleum engineering has a way better median income than the other engineering majors and that single observation distorts the whole graph!

let’s see what happens to the representation of the data if we ommit it.

```
no_PE <- CMI %>%
  filter(Major != "PETROLEUM ENGINEERING")

no_PE %>%
  ggplot(aes(x = Major_category, y = Median, las = 2)) +
    geom_boxplot(fill = "slateblue1", alpha = 0.5) +
    labs(x = NULL, y = "Median", title = "Same boxplot without petroleum
engineering") +
    theme(plot.title = element_text(hjust = 0.5)) +
    coord_flip()
```

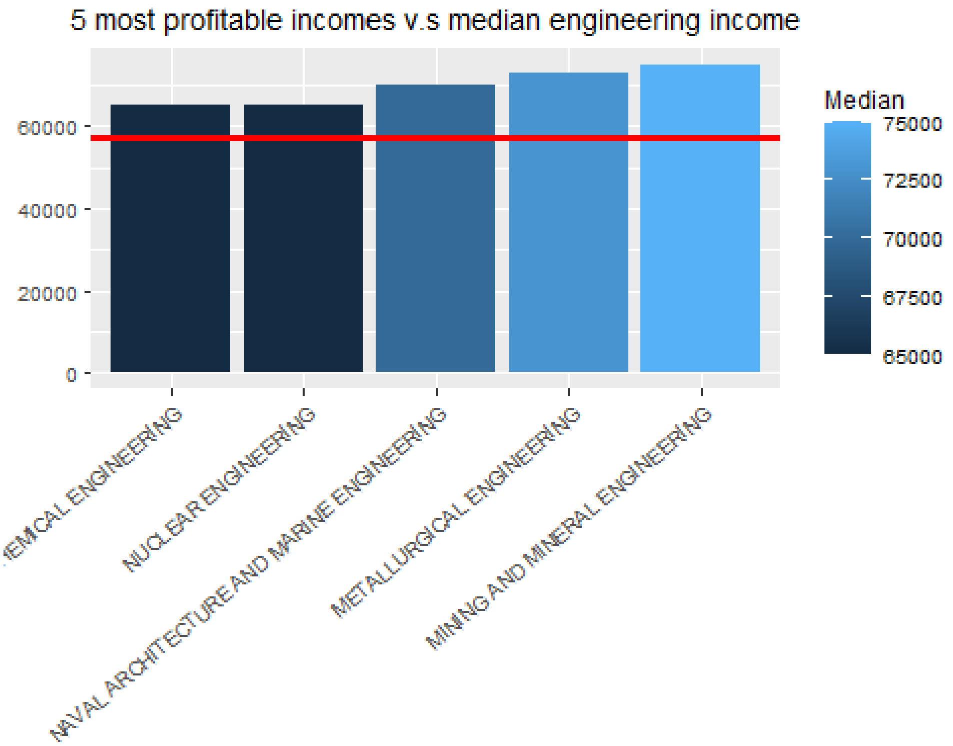


ahhh, much better.

We know that the bigger the size of the box, the bigger the variance and there are big differences amongst the different fields of study. A possible explination for the differences is the more diverse the field, we can expect more diverse majors with different incomes, which increases the variance.

We can see that engineering has the biggest median income. let's check out the top 5 majors by income, other than petroleum engineering which we saw before was "off the chart":

```
arranged_top <- no_PE %>%
  arrange(desc(Median))
top_5 <- head(arranged_top, 5)

Engineerings <- CMI %>%
  filter(Major_category == "Engineering")

median_income_engineering <- median(Engineerings$Median)

top_5 %>%
  ggplot(aes(x = reorder(Major, Median), y = Median, las = 2, fill = Median))
+
    geom_bar(stat = "identity") +
  
    geom_hline(yintercept = median_income_engineering, color = "red", size = 
1.2) +
  
    theme(axis.text.x = element_text(angle = 40, hjust = 1),
          text = element_text(size = 9), plot.title = element_text(hjust = 
0.5)) +
  
    labs(x = NULL, y = NULL, title = "5 most profitable incomes v.s median 
engineering income")
```

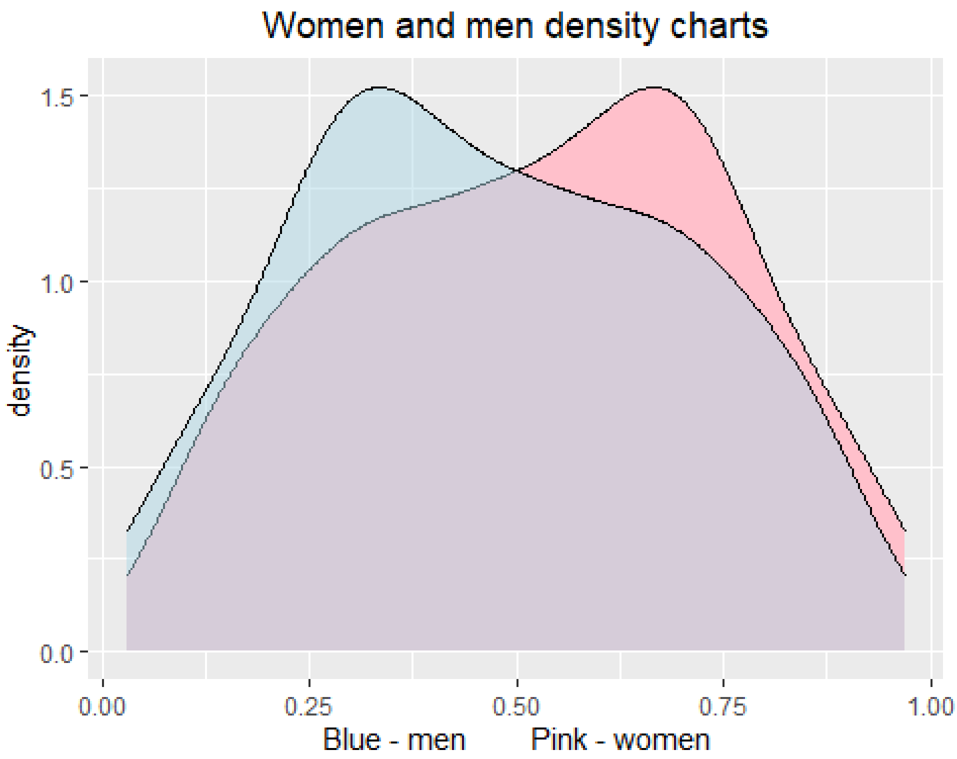


Interesting to see that in the USA, where the servey was taken, 5 of the 6 most profitable majors are chimestry related engineering, which is very much not the case in Israel.

Gender

Another aspect we will examine is the different correlations of the gender of the participants and the data. let's look at the distribution of the precentage of men and women in the different majors:

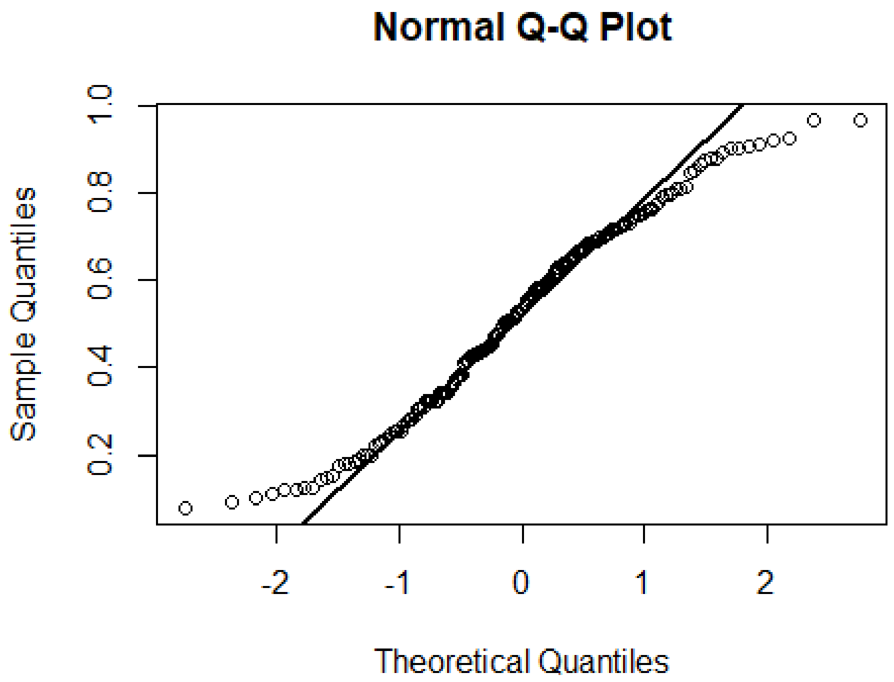
```
men_and_women <- transmute(CMI,
          men = 1 - ShareWomen,
          women = ShareWomen)

ggplot(data=men_and_women) +
  geom_density(aes(x=women), fill="pink") +
  geom_density(aes(x=men), fill="lightblue", alpha = 0.5) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Blue - men        Pink - women", title = "Women and men density
charts")
```



we can see that the density graphs that the representations are relatively normal-looking. the women's density chart has a "hump" between 50% to 75% and men between 25% to 50%, of course. That means women are better represented than men in the different fields of study, that's great!

To further check if the women's chart is normal let's take a look at the qqplot:

```
qqnorm(CMI$ShareWomen)
qqline(CMI$ShareWomen, lwd = 2)
```



By looking at the qqplot, which can give us a good graphical assessment of "goodness of fit", rather than reducing to a numerical summary, we can see most values behave pretty well so we will assume normality.

Modeling

Incomes

let's get down to the real stuff.

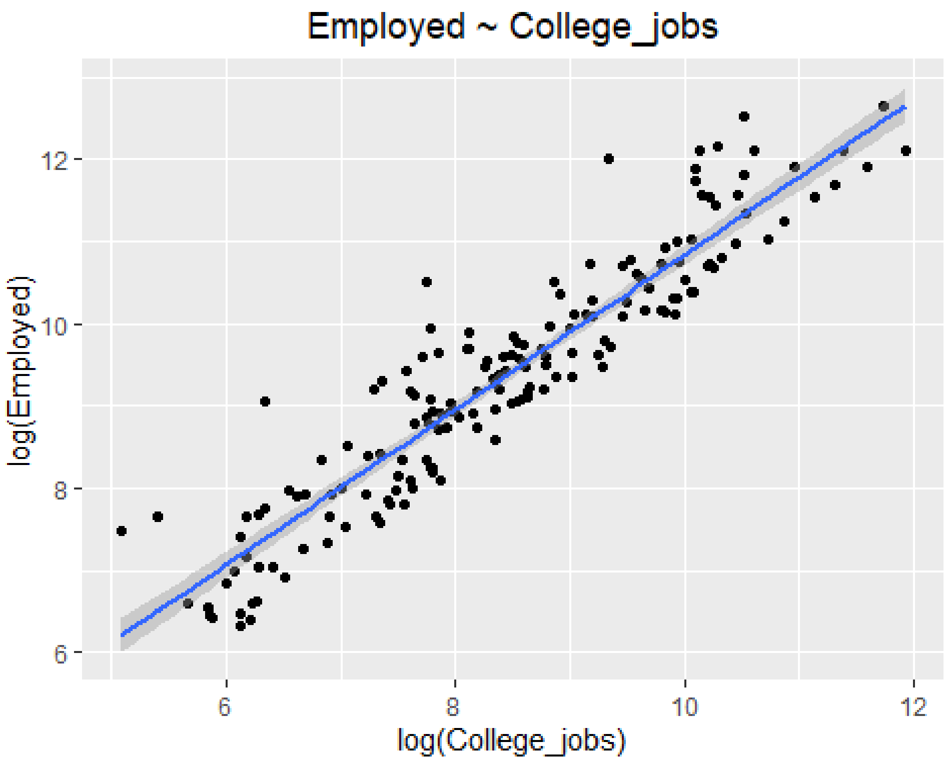
first of all, let's check if the participants (who have college degrees in various fields) work in a job that requires a degree or not. we assume that going to school will get you a better job - in this case, college level job. what we're interested in finding out is how strong is that connection. so, our null hypothesis is that there is no correlation between employment and being employed at a college level job (given that you went to college).

```
employed_and_college_jobs <- lm(formula = Employed ~ College_jobs, data =
CMI)

summary(employed_and_college_jobs)

##
## Call:
## lm(formula = Employed ~ College_jobs, data = CMI)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -114356   -8688   -6711    -827  198689
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  7981.4366  2746.8530   2.906              0.00416 **
## College_jobs    1.8944     0.1109  17.081 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30920 on 168 degrees of freedom
## Multiple R-squared:  0.6346, Adjusted R-squared:  0.6324
## F-statistic: 291.8 on 1 and 168 DF,  p-value: < 0.00000000000000022
```

In the representation of the regression that follows, we used the logarithmic values to represent the data to scale.

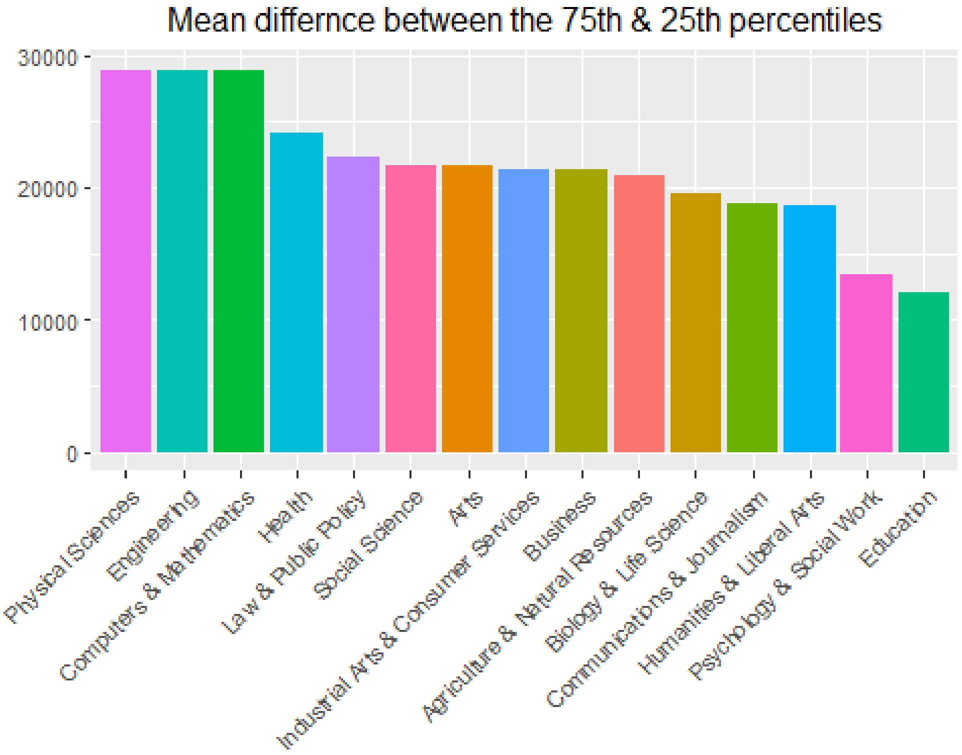
```
 ggplot(CMI, aes(x = log(College_jobs), y = log(Employed))) +
  geom_point() +
  labs(title = "Employed ~ College_jobs") +
  theme(plot.title = element_text(hjust = 0.5)) +
  stat_smooth(method = "lm")

## `geom_smooth()` using formula 'y ~ x'
```



As we can see in the linear regression above, we got a very small p-value which suggests strong correlation as expected, so with significance level of 0.05 we can say that there is a correlation. We reject our null hypothesis - if you're an employed college grad, you'll most likely work at a college level job.

As college students, obviously, we are interested in having a high-income job. But also, as aspiring data scientists, when conducting a research on this topic, we have to think not only on the gross income but also on the certainty of earning that ammount, hence, the variance of the median income. Let's see those side by side in decreasing order:

```r
CMI_percentilse <- CMI %>%
  as_tibble() %>%

  transmute(
    Major = Major,
    Major_category = Major_category,
    diff_75p_25p = P75th - P25th) %>%

  arrange(diff_75p_25p)

precentile_differences <- aggregate(diff_75p_25p ~ Major_category, data =
CMI_percentilse, mean)


ggplot(data = precentile_differences, cex.axis = 1.5) +

  geom_bar(aes(x = reorder(Major_category, -diff_75p_25p),
               y = diff_75p_25p, fill = Major_category), stat="identity") +

  labs( x = NULL,y = NULL, title = "Mean differnce between the 75th & 25th
percentiles") +

  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        text = element_text(size = 10), axis.title.y = element_text(size =
8),
        plot.title = element_text(hjust = 0.5)) +

  theme(legend.position = "none")
```



We can conclude from looking at the graph that the 3 major categories with the worst income security are scientific and the 3 with the best are non-scientific. So, if you want high certainty of the ammount you will earn at your job, we'd stay away from science.

Gender

Next, we want to challenge the notion that women mostly choose to study in the fields of humanities rather than the scientific fields. in order to do that we will distinguish between scientific and non-scientific major categories:

```
humanities <- CMI %>% filter(Major_category == "Arts" |
                             Major_category == "Communications & Journalism"
|
                             Major_category == "Education" |
                             Major_category == "Humanities & Liberal Arts" |
                             Major_category == "Health" |
                             Major_category == "Industrial Arts & Consumer
Services" |
                             Major_category == "Law & Public Policy" |
                             Major_category == "Psychology & Social Work" |
                             Major_category == "Social Science")
```

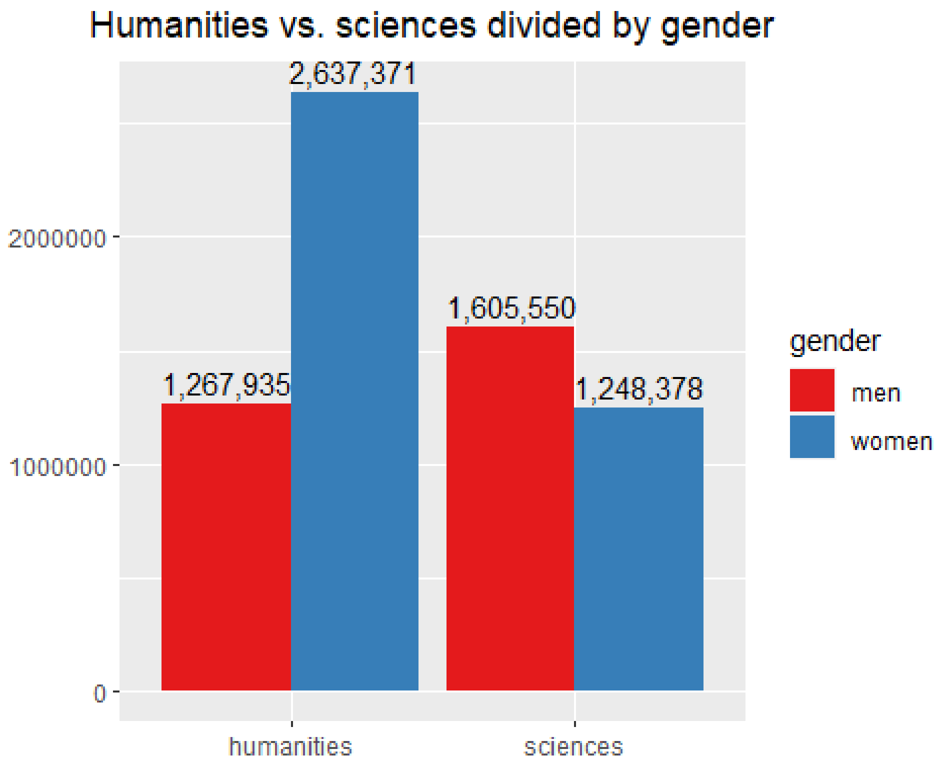
Now we will check how many men and women study in each:

```
women_in_humanities <- sum(humanities$Women)
men_in_humanities <- sum(humanities$Men)
```

```
women_in_sciences <- sum(CMI$Women) - women_in_humanities
men_in_sciences <- sum(CMI$Men) - men_in_humanities
```

Finally, you gotta get all that data in a nice bar chart:

```
sciences_vs_humanities_gender <- tribble(
  ~gender, ~typeof, ~amount,
  "women", "sciences", women_in_sciences,
  "women", "humanities", women_in_humanities,
  "men",   "humanities", men_in_humanities,
  "men",   "sciences", men_in_sciences)

ggplot(sciences_vs_humanities_gender, aes(x = typeof, y = amount, fill =
gender)) +

  geom_bar(stat="identity", position = "dodge") +

  geom_text(aes(label = prettyNum(amount, big.mark = ",", scientific =
FALSE)),
            position = position_dodge2(width = 0.9), vjust = -0.4) +

  theme(plot.title = element_text(hjust = 0.5, vjust = 1.5)) +
  labs(x = NULL, y = NULL, title = "Humanities vs. sciences divided by
gender") +
  scale_fill_brewer(palette = "Set1")
```



With men, we can see that there is a preference towards scientific majors but relatively not by a lot. on the other hand, in the fields of humanities, the ammount of women doubles the men. because of the size of the data in the survey (~ 6.76 million) and the massive difference in humanities, we can say that as a whole, women probably prefer studying non-scientific majors.

Secondly, another way we can explore the connection between gender and career characteristics is by looking at part-time jobs. Because of the lifestyle a lot of women choose - raising children, some prefer having part-time jobs in order to be more with their children.

Because of that, we assume that there will be a positive correlation between the precentage of women who pursued a degree in a specific field and the precentage of part-time jobs (less than 35 hours weekly) offered in that field.

```
women_WF <- transmute(CMI,
                   prec_women = ShareWomen,
                   prec_part_time = Part_time / Employed,
                   field = Major_category)

part_time <- lm(formula = prec_part_time ~ prec_women*field, data = women_WF)
summary(part_time)

##
## Call:
## lm(formula = prec_part_time ~ prec_women * field, data = women_WF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.234869 -0.036150  0.001884  0.035060  0.183892
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                          0.237028   0.051693
## prec_women                                           0.038336   0.114988
## fieldArts                                            0.181237   0.126526
## fieldBiology & Life Science                         -0.343289   0.214434
## fieldBusiness                                       -0.136086   0.087599
## fieldCommunications & Journalism                     0.175700   0.280000
## fieldComputers & Mathematics                        -0.310617   0.079506
## fieldEducation                                       0.027018   0.099964
## fieldEngineering                                    -0.100266   0.060943
## fieldHealth                                          0.120611   0.145342
## fieldHumanities & Liberal Arts                       0.095527   0.097137
## fieldIndustrial Arts & Consumer Services            -0.107244   0.070050
## fieldLaw & Public Policy                            -0.097280   0.117674
## fieldPhysical Sciences                               0.202647   0.093585
## fieldPsychology & Social Work                        0.627971   0.231204
## fieldSocial Science                                  0.040319   0.109406
## prec_women:fieldArts                                -0.043374   0.219716
## prec_women:fieldBiology & Life Science               0.742530   0.371347
## prec_women:fieldBusiness                             0.117892   0.182095
## prec_women:fieldCommunications & Journalism         -0.257942   0.430495
## prec_women:fieldComputers & Mathematics              0.895403   0.215753
## prec_women:fieldEducation                           -0.083104   0.160580
## prec_women:fieldEngineering                          0.281603   0.169578
## prec_women:fieldHealth                              -0.009838   0.204481
## prec_women:fieldHumanities & Liberal Arts            0.075370   0.171505
## prec_women:fieldIndustrial Arts & Consumer Services  0.220042   0.148818
## prec_women:fieldLaw & Public Policy                  0.233499   0.239047
## prec_women:fieldPhysical Sciences                   -0.269906   0.187084
## prec_women:fieldPsychology & Social Work            -0.695637   0.304796
## prec_women:fieldSocial Science                       0.043217   0.204667
##                                                     t value   Pr(>|t|)
## (Intercept)                                           4.585 0.00000996 ***
## prec_women                                            0.333   0.739339
## fieldArts                                             1.432   0.154254
## fieldBiology & Life Science                          -1.601   0.111651
## fieldBusiness                                        -1.554   0.122557
## fieldCommunications & Journalism                      0.628   0.531354
## fieldComputers & Mathematics                         -3.907   0.000145 ***
## fieldEducation                                        0.270   0.787343
## fieldEngineering                                     -1.645   0.102164
## fieldHealth                                           0.830   0.408043
## fieldHumanities & Liberal Arts                        0.983   0.327094
## fieldIndustrial Arts & Consumer Services             -1.531   0.128036
## fieldLaw & Public Policy                             -0.827   0.409821
## fieldPhysical Sciences                                2.165   0.032052 *
## fieldPsychology & Social Work                         2.716   0.007439 **
## fieldSocial Science                                   0.369   0.713042
## prec_women:fieldArts                                 -0.197   0.843795
## prec_women:fieldBiology & Life Science                2.000   0.047483 *
## prec_women:fieldBusiness                              0.647   0.518422
## prec_women:fieldCommunications & Journalism          -0.599   0.550024
## prec_women:fieldComputers & Mathematics               4.150 0.00005740 ***
## prec_women:fieldEducation                            -0.518   0.605607
## prec_women:fieldEngineering                           1.661   0.099029 .
## prec_women:fieldHealth                               -0.048   0.961696
## prec_women:fieldHumanities & Liberal Arts             0.439   0.661002
## prec_women:fieldIndustrial Arts & Consumer Services   1.479   0.141495
## prec_women:fieldLaw & Public Policy                   0.977   0.330359
## prec_women:fieldPhysical Sciences                    -1.443   0.151340
## prec_women:fieldPsychology & Social Work             -2.282   0.023979 *
## prec_women:fieldSocial Science                        0.211   0.833071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06712 on 140 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.5981
## F-statistic: 9.671 on 29 and 140 DF,  p-value: < 0.0000000000000022
```

We can see that The adjusted R squared is sufficient and the model's p-value is very small, so we reject our null hypothesis with a confidence level of 0.05. There is a correlation between the precentage of women in a specific field and the precentage of part-time jobs as a whole with 3 fields which have significant correlations, most of which are postive as expected.

Connecting the dots

Let's put it all together…

We've seen before that income varies greatly between the different majors. there were more female graduates in the survey than men so the representation of women in different fields should be more spread out than men so there is no reason for us to expect there will be a correlation between the precentage of women ("ShareWomen") represented in a field and it's income. That will be our null hypothesis - beta_Women = 0.

(we know that in linear regression the null hypothesis is that all the coefficients are 0 but we want to focus specifically on the precentage of women later on).

Let's challenge that by using stepwise regression. We took into account factors from the dataset which could influence the income (besides "ShareWomen"), and are not strongly correlated to each other. let's look at what factors influence the median incomes in the different majors:

```
Medianstep <- lm(formula = Median ~ Men + ShareWomen + College_jobs +
Employed + Unemployment_rate, data = CMI)
step(Medianstep, direction = "backward")

## Start:  AIC=3097.95
## Median ~ Men + ShareWomen + College_jobs + Employed + Unemployment_rate
##
##                      Df  Sum of Sq         RSS    AIC
## - Men                  1   10788897 13013805770 3096.1
## - Employed             1   27291498 13030308371 3096.3
## - Unemployment_rate  1  111045815 13114062688 3097.4
## <none>                            13003016873 3098.0
## - College_jobs         1  271556012 13274572885 3099.5
## - ShareWomen           1 6879474567 19882491440 3168.1
##
## Step:  AIC=3096.09
## Median ~ ShareWomen + College_jobs + Employed + Unemployment_rate
##
##                      Df  Sum of Sq         RSS    AIC
## - Unemployment_rate  1  116229607 13130035377 3095.6
## <none>                            13013805770 3096.1
## - Employed             1  276582469 13290388239 3097.7
## - College_jobs         1  392508752 13406314522 3099.1
## - ShareWomen           1 8846607125 21860412895 3182.3
##
## Step:  AIC=3095.6
## Median ~ ShareWomen + College_jobs + Employed
## 
##                 Df  Sum of Sq         RSS    AIC
## <none>                        13130035377 3095.6
## - Employed      1  332396302 13462431679 3097.9
## - College_jobs  1  453843910 13583879287 3099.4
## - ShareWomen    1 8975694437 22105729813 3182.2

## 
## Call:
## lm(formula = Median ~ ShareWomen + College_jobs + Employed, data = CMI)
## 
## Coefficients:
##  (Intercept)    ShareWomen  College_jobs      Employed  
##   57002.9121   -32547.3243        0.1275       -0.0455
```

The stepwise suggested that "ShareWomen" does explain the median income! other than that, the number of employed and the number of people employed at college-level jobs also explain income best.

Let's model that:

```
Medianlm <- lm(formula = Median ~ ShareWomen + College_jobs + Employed, data
= CMI)
summary(Medianlm)

## 
## Call:
## lm(formula = Median ~ ShareWomen + College_jobs + Employed, data = CMI)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -17961  -5386  -1342   3757  56815 
## 
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)    
## (Intercept)  57002.91206   1726.23095  33.022 <0.0000000000000002 ***
## ShareWomen  -32547.32430   3055.34436 -10.653 <0.0000000000000002 ***
## College_jobs     0.12753      0.05324   2.395              0.0177 *  
## Employed        -0.04550      0.02220  -2.050              0.0419 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8894 on 166 degrees of freedom
## Multiple R-squared:  0.4148, Adjusted R-squared:  0.4042 
## F-statistic: 39.22 on 3 and 166 DF,  p-value: < 0.00000000000000022
```

Let's break down what we have here: first of all the model and the coefficients are all significant which means they explain the income well.

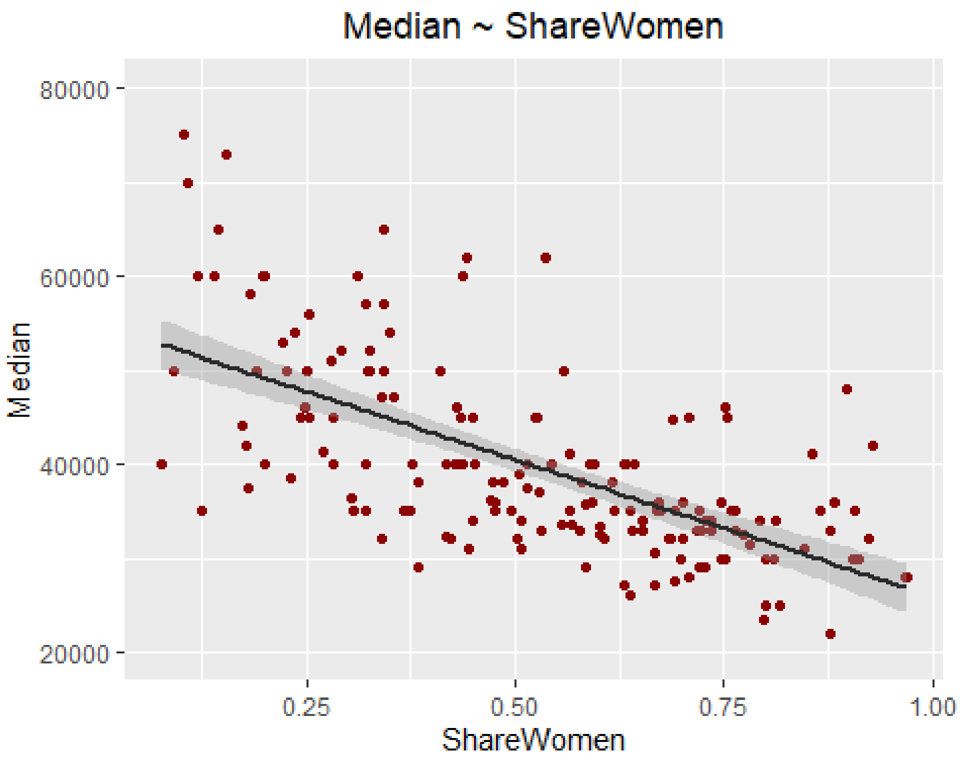
1. we can see that “college_jobs” has a positive correlation with income and “Employment” has a negative correlation. This means that, according to the results, in fields of work with higher income we would expect to see less people employed as a whole, but more skilled workers in college-level jobs.

A possible explination is that fields providing higher income do require a more skilled work force. We showed before that the most prominent jobs with high incomes are chemistry-related engineerings, which are definately high-skill jobs. Those hard skills are difficult to obtain so we can expect a smaller, more skilled work-force in those fields which have higher incomes.

2. There is a strong negative correlation between precentage of women and income. this means that the higher the income in a field, we would expect to see a much lower precentage of women. The p-value is very small so with significance level of 0.05 we can say that there is a correlation. We reject our null hypothesis - there is a negative correlation between precentage of women in a field and it’s median income.

```r
ggplot(data = CMI, aes(x = ShareWomen, y = Median)) +
  geom_point(color = "darkred") +
  ylim(c(20000,80000)) +
  labs(title = "Median ~ ShareWomen") +
  theme(plot.title = element_text(hjust = 0.5)) +
  stat_smooth(method = "lm", color = "gray16")
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

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



And why is that? actually, the answer shouldn’t be a surprise at all!

As we’ve shown before, the survey suggested that out of the ~ 6.76 million people the survey took into account, women prefer non-scientific pursuits. let’s see what it has to do with income.

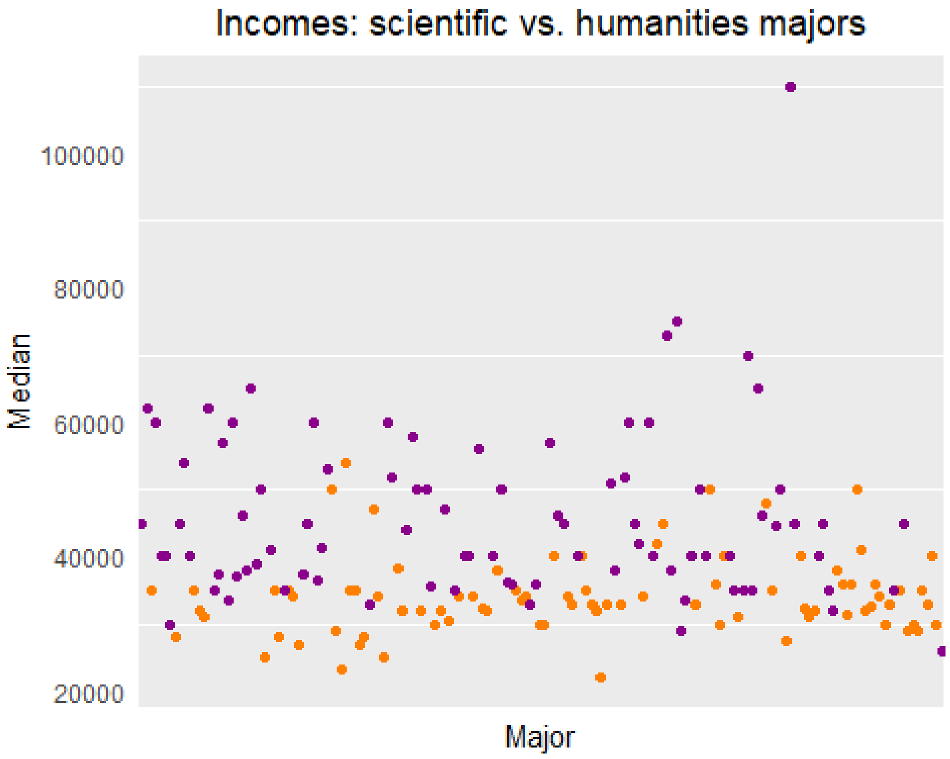
first of all let’s look at the 3 non-scientific majors with the highest income:

```
non_scientific <- humanities %>%

  arrange(desc(Median)) %>%
  select(Rank, Major)

head(non_scientific,3)

## # A tibble: 3 x 2
##    Rank Major
##   <dbl> <chr>
## 1    20 COURT REPORTING
## 2    27 CONSTRUCTION SERVICES
## 3    30 PUBLIC POLICY
```

Between the top 30 majors with the highest incomes only 3 are non-scientific! let’s see what happens with the rest.

To do that, we will use the same method as before to distinguish between the two types of majors and see how they behave together.

```
sciences <- CMI %>%
  filter(Major_category != "Arts" & Major_category != "Communications &
Journalism" & Major_category != "Education" & Major_category != "Humanities &
Liberal Arts" & Major_category != "Health" & Major_category != "Industrial
Arts & Consumer Services" & Major_category != "Law & Public Policy" &
Major_category != "Psychology & Social Work" & Major_category != "Social
Science")

scientifics_incomes <- sciences
humanities_incomes <- humanities

ggplot() +
  geom_point(data = humanities_incomes, aes(x = Major, y = Median), color =
"darkorange1") +
  geom_point(data = scientifics_incomes, aes(x = Major, y = Median), color =
"darkmagenta") +
  labs(title = "Incomes: scientific vs. humanities majors") +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank(),
        panel.grid.major = element_blank(), plot.title = element_text(hjust =
0.5))
```



Clearly, the purple dots, which represent the scientific majors, have higher y values. this suggests that scientific majors provide higher incomes. let's conduct a two-sampled t-test to check if they provide a higher income with 95% confridence level. our null hypothesis is that there is no difference in incomes.

```
check_h_s <- t.test(scientifics_incomes$Median, humanities_incomes$Median,
alternative = "greater")
check_h_s

##
##  Welch Two Sample t-test
##
## data:  scientifics_incomes$Median and humanities_incomes$Median
## t = 7.6991, df = 121.63, p-value = 0.000000000002031
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  9157.398      Inf
## sample estimates:
## mean of x mean of y
##  45873.26  34203.57
```

The t-test suggested, with a very small p-value, that the mean income in scientific majors is greater. Which means that we reject our null hypothesis: scientific majors do provide higher incomes.

In conclusion, women prefer non-scientific majors which provide lower incomes and that's why the higher the income in a major the less women we will see in it, hence the strong negative correlation in the linear model.

Conclusion

In our report we set two goals to check.

The first one was: fields of study and the predicted nature of future employment and how that influences income in those fields.

We found that the fields of study in the survey have distinct incomes and income variances, both of which differ greatly between scientific and non-scientific fields.

We've seen that in scientific fields you can expect a higher income than non-scientific fields, but you'll have a low level of certainty of earning that high ammount. Also, we've found out that in fields with higher income, we would expect to find a relatively small, high skilled workforce, and sadly, less women.

The second goal was: how is gender correlated to preferences in choosing a major and the following career's characteristics.

We found that women tend to prefer non-scientific career fields. Furthermore, the bigger the precentage of women pursuing a career in a specific field, the higher our expectation to find smaller salaries and more part time jobs in that field.

Thank you for reading!