employment_education

Nathan Grossman

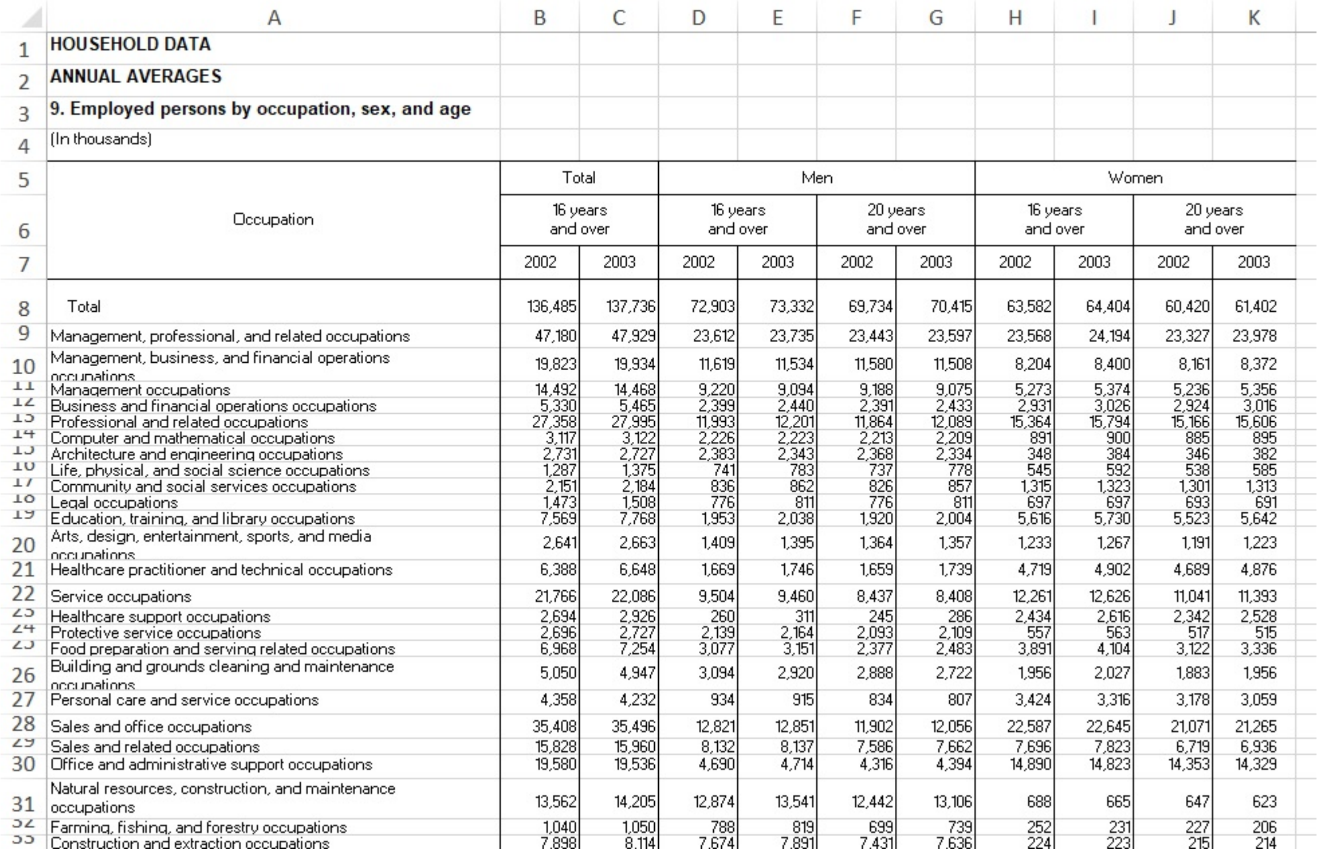
December 17, 2017

Introduction

The goal of this project is to gain insight regarding the under-representation of women in the hi-tech workforce. Specifically, we will explore whether this under-representation arises from an inadequete pipeline of women studying subjects related to hi-tech in school, or arises from an insufficient number of women being attracted to and retained retained in hi-tech jobs after graduation.

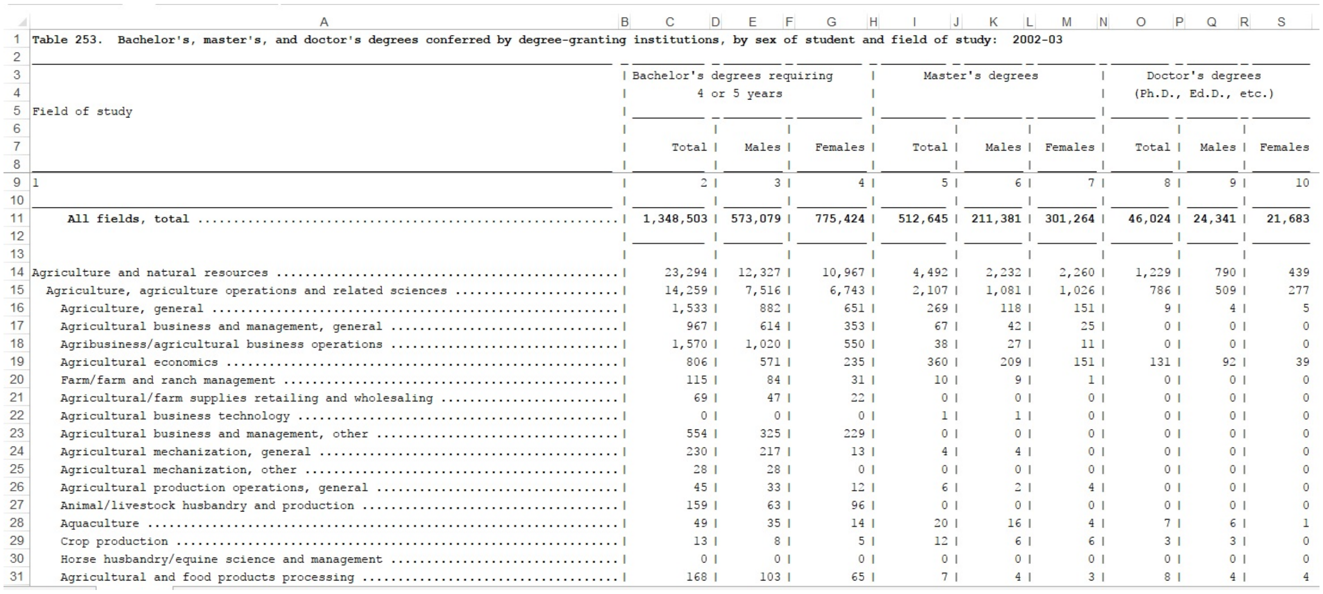
In order to accomplish this goal, we will examine two sources of data: (1) employment data collected by the Bureau of Labor Statistics (BLS) and located at: https://www.bls.gov/cps/tables.htm (https://www.bls.gov/cps/tables.htm); and (2) education data collected by the National Center for Educational Statistics (NCES) and located at: https://nces.ed.gov/programs/digest/current_tables.asp (https://nces.ed.gov/programs/digest/current_tables.asp).

The primary challenges result from the fact that both datasets are distributed across multiple XLS/CSV fles, and that neither dataset is in tidy format in its raw state. For example, the raw employment data is located in different files for different years, in the following format:



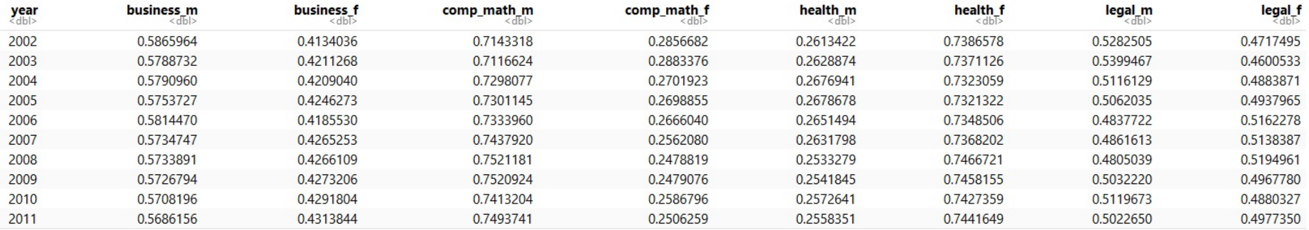
	A	B	C	D	E	F	G	H	I	J	K
1	**HOUSEHOLD DATA**										
2	**ANNUAL AVERAGES**										
3	**9. Employed persons by occupation, sex, and age**										
4	(In thousands)										
5	Occupation	Total		Men				Women			
6		16 years and over		16 years and over		20 years and over		16 years and over		20 years and over	
7		2002	2003	2002	2003	2002	2003	2002	2003	2002	2003
8	Total	136,485	137,736	72,903	73,332	69,734	70,415	63,582	64,404	60,420	61,402
9	Management, professional, and related occupations	47,180	47,929	23,612	23,735	23,443	23,597	23,568	24,194	23,327	23,978
10	Management, business, and financial operations occupations	19,823	19,934	11,619	11,534	11,580	11,508	8,204	8,400	8,161	8,372
11	Management occupations	14,492	14,468	9,220	9,094	9,188	9,075	5,273	5,374	5,236	5,356
12	Business and financial operations occupations	5,330	5,465	2,399	2,440	2,391	2,433	2,931	3,026	2,924	3,016
13	Professional and related occupations	27,358	27,995	11,993	12,201	11,864	12,089	15,364	15,794	15,166	15,606
14	Computer and mathematical occupations	3,117	3,122	2,226	2,223	2,213	2,209	891	900	885	895
15	Architecture and engineering occupations	2,731	2,727	2,383	2,343	2,368	2,334	348	384	346	382
16	Life, physical, and social science occupations	1,287	1,375	741	783	737	778	545	592	538	585
17	Community and social services occupations	2,151	2,184	836	862	826	857	1,315	1,323	1,301	1,313
18	Legal occupations	1,473	1,508	776	811	776	811	697	697	693	691
19	Education, training, and library occupations	7,569	7,768	1,953	2,038	1,920	2,004	5,616	5,730	5,523	5,642
20	Arts, design, entertainment, sports, and media occupations	2,641	2,663	1,409	1,395	1,364	1,357	1,233	1,267	1,191	1,223
21	Healthcare practitioner and technical occupations	6,388	6,648	1,669	1,746	1,659	1,739	4,719	4,902	4,689	4,876
22	Service occupations	21,766	22,086	9,504	9,460	8,437	8,408	12,261	12,626	11,041	11,393
23	Healthcare support occupations	2,694	2,926	260	311	245	286	2,434	2,616	2,342	2,528
24	Protective service occupations	2,696	2,727	2,139	2,164	2,093	2,109	557	563	517	515
25	Food preparation and serving related occupations	6,968	7,254	3,077	3,151	2,377	2,483	3,891	4,104	3,122	3,336
26	Building and grounds cleaning and maintenance occupations	5,050	4,947	3,094	2,920	2,888	2,722	1,956	2,027	1,883	1,956
27	Personal care and service occupations	4,358	4,232	934	915	834	807	3,424	3,316	3,178	3,059
28	Sales and office occupations	35,408	35,496	12,821	12,851	11,902	12,056	22,587	22,645	21,071	21,265
29	Sales and related occupations	15,828	15,960	8,132	8,137	7,586	7,662	7,696	7,823	6,719	6,936
30	Office and administrative support occupations	19,580	19,536	4,690	4,714	4,316	4,394	14,890	14,823	14,353	14,329
31	Natural resources, construction, and maintenance occupations	13,562	14,205	12,874	13,541	12,442	13,106	688	665	647	623
32	Farming, fishing, and forestry occupations	1,040	1,050	788	819	699	739	252	231	227	206
33	Construction and extraction occupations	7,898	8,114	7,674	7,891	7,431	7,636	224	223	215	214

Similarly, the raw education data is located in different files for different years, in the following format:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Table 253. Bachelor's, master's, and doctor's degrees conferred by degree-granting institutions, by sex of student and field of study: 2002-03																		
2																			
3		|	Bachelor's degrees requiring					|	Master's degrees					|	Doctor's degrees				
4		|	4 or 5 years					|						|	(Ph.D., Ed.D., etc.)				
5	Field of study	|						|						|				
6		|		|		|		|		|		|		|		|	
7		|	Total	|	Males	|	Females	|	Total	|	Males	|	Females	|	Total	|	Males	|	Females
8		|		|		|		|		|		|		|		|		|	
9	1	|	2	|	3	|	4	|	5	|	6	|	7	|	8	|	9	|	10
10		|		|		|		|		|		|		|		|		|	
11	**All fields, total ....**	|	**1,348,503**	|	**573,079**	|	**775,424**	|	**512,645**	|	**211,381**	|	**301,264**	|	**46,024**	|	**24,341**	|	**21,683**
12		|		|		|		|		|		|		|		|		|	
13		|		|		|		|		|		|		|		|		|	
14	Agriculture and natural resources	|	23,294	|	12,327	|	10,967	|	4,492	|	2,232	|	2,260	|	1,229	|	790	|	439
15	Agriculture, agriculture operations and related sciences	|	14,259	|	7,516	|	6,743	|	2,107	|	1,081	|	1,026	|	786	|	509	|	277
16	Agriculture, general	|	1,533	|	882	|	651	|	269	|	118	|	151	|	9	|	4	|	5
17	Agricultural business and management, general	|	967	|	614	|	353	|	67	|	42	|	25	|	0	|	0	|	0
18	Agribusiness/agricultural business operations	|	1,570	|	1,020	|	550	|	38	|	27	|	11	|	0	|	0	|	0
19	Agricultural economics	|	806	|	571	|	235	|	360	|	209	|	151	|	131	|	92	|	39
20	Farm/farm and ranch management	|	115	|	84	|	31	|	10	|	9	|	1	|	0	|	0	|	0
21	Agricultural/farm supplies retailing and wholesaling	|	69	|	47	|	22	|	0	|	0	|	0	|	0	|	0	|	0
22	Agricultural business technology	|	0	|	0	|	0	|	1	|	1	|	0	|	0	|	0	|	0
23	Agricultural business and management, other	|	554	|	325	|	229	|	0	|	0	|	0	|	0	|	0	|	0
24	Agricultural mechanization, general	|	230	|	217	|	13	|	4	|	4	|	0	|	0	|	0	|	0
25	Agricultural mechanization, other	|	28	|	28	|	0	|	0	|	0	|	0	|	0	|	0	|	0
26	Agricultural production operations, general	|	45	|	33	|	12	|	6	|	2	|	4	|	0	|	0	|	0
27	Animal/livestock husbandry and production	|	159	|	63	|	96	|	0	|	0	|	0	|	0	|	0	|	0
28	Aquaculture	|	49	|	35	|	14	|	20	|	16	|	4	|	7	|	6	|	1
29	Crop production	|	13	|	8	|	5	|	12	|	6	|	6	|	3	|	3	|	0
30	Horse husbandry/equine science and management	|	0	|	0	|	0	|	0	|	0	|	0	|	0	|	0	|	0
31	Agricultural and food products processing	|	168	|	103	|	65	|	7	|	4	|	3	|	8	|	4	|	4

To facilitate our analysis, we will merge the employment data from multiple files into a single dataframe, with the following tidy format:



year <dbl>	business_m <dbl>	business_f <dbl>	comp_math_m <dbl>	comp_math_f <dbl>	health_m <dbl>	health_f <dbl>	legal_m <dbl>	legal_f <dbl>
2002	0.5865964	0.4134036	0.7143318	0.2856682	0.2613422	0.7386578	0.5282505	0.4717495
2003	0.5788732	0.4211268	0.7116624	0.2883376	0.2628874	0.7371126	0.5399467	0.4600533
2004	0.5790960	0.4209040	0.7298077	0.2701923	0.2676941	0.7323059	0.5116129	0.4883871
2005	0.5753727	0.4246273	0.7301145	0.2698855	0.2678678	0.7321322	0.5062035	0.4937965
2006	0.5814470	0.4185530	0.7333960	0.2666040	0.2651494	0.7348506	0.4837722	0.5162278
2007	0.5734747	0.4265253	0.7437920	0.2562080	0.2631798	0.7368202	0.4861613	0.5138387
2008	0.5733891	0.4266109	0.7521181	0.2478819	0.2533279	0.7466721	0.4805039	0.5194961
2009	0.5726794	0.4273206	0.7520924	0.2479076	0.2541845	0.7458155	0.5032220	0.4967780
2010	0.5708196	0.4291804	0.7413204	0.2586796	0.2572641	0.7427359	0.5119673	0.4880327
2011	0.5686156	0.4313844	0.7493741	0.2506259	0.2558351	0.7441649	0.5022650	0.4977350

We will transform the education data in a similar fashion.

To summarize, the preparation of the data for analysis will comprise the following steps: loading the raw data from multiple files with slightly different data formats; extracting the rows and columns of interest into a dataframe; and transposing the dataframe, i.e. transforming rows to columns and vice versa, so that each row pertains to statistics from a different year.

Subsequently, we will link the employment and education datasets for ease of comparison.

Throughout this study, we will examine employment and education data across four areas: business, computers/ mathematics, health and legal. The hi-tech sector, as represented by employment and education data on the computers/ mathematics area, is of primary interest. Data on the business, health and legal areas is included for sake of comparison, and to baseline the representation of men and women in the educational pipeline and the professional workplace.

Initialize Environment

```
setwd('C:/Users/Nathan/RStudioProjects/employment_education')
rm(list=ls())
library(tidyverse)
library(datetime)
library(lubridate)
library(shiny)
library(rsconnect)
knitr::opts_chunk$set(echo = TRUE)
theme_update(plot.title = element_text(hjust = 0.5))
```

Define Functions

Put strings in snake case, and clean them up in general, using calls to gsub() and tolower() functions

```
snake_case = function(x) {
    y = gsub("\\s", "_", x)
    y = gsub("(.)([A-Z][a-z]+)", "\\1_\\2", y)
    y = tolower(gsub("([a-z0-9])([A-Z])", "\\1_\\2", y))
    y = gsub("__", "_", y)
    y = gsub("^[_, .]", "", y)
    y = gsub(' ', '', y)
    y = gsub(',', '', y)
    y = gsub('\\.', '', y)
    y = gsub('/', '_', y)
    y = gsub('-', '_', y)
    y = gsub('\\(', '_', y)
    y = gsub('\\)', '_', y)
    y = gsub("\\'", '', y)
    y = gsub('_$', '', y)
}
```

Put dataframe column names in snake case, and clean them up in general, using call to snake_case() function defined above

```
fix_column_names = function(x) {
    colnames(x) = snake_case(colnames(x))
    return(x)
}
```

Put dataframe row names in snake case, and clean them up in general, using call to snake_case() function defined above

```
fix_row_names = function(x) {
    x$X1 = snake_case(x$X1)
    return(x)
}
```

Find disciplines of interest in education data, using calls to filter() and str_detect() functions

```
find_disciplines = function(x) {
    x %>% filter((str_detect(X1, 'business_management_marketing_and_related_support_services')
                  | str_detect(X1, 'computer_and_information_sciences_and_support_services')
                  | str_detect(X1, 'health_professions')
                  | str_detect(X1, 'legal_professions_and_studies')
                  | str_detect(X1, 'mathematics_and_statistics'))
                  & !str_detect(X1, 'other'))
}
```

Rename disciplines of interest in education data, using calls rename() function

```
rename_disciplines = function(x) {
    x %>% rename(business = business_management_marketing_and_related_support_services) %>%
          rename(computers = computer_and_information_sciences_and_support_services) %>%
          rename(health = health_professions_and_related_clinical_sciences) %>%
          rename(legal = legal_professions_and_studies) %>%
          rename(mathematics = mathematics_and_statistics)
}
```

Transpose employment dataframe, i.e. transform rows to columns and vice versa, using calls to select(), t(), rownames() and colnames() functions

```
transpose_employment_dataframe = function(empX) {
    valuesX = empX %>% select(X02:X14) %>% t()
    rownames(valuesX) = c()

    labelsX = empX %>% select(X1) %>% t()
    rownames(labelsX) = c()
    labelsX[1] = 'Year'

    empX = as.tibble(valuesX)
    colnames(empX) = labelsX
    return(empX)
}
```

Adjust columns of employment dataframe, using calls to fix_column_names() function defined above and to select() function

```
adjust_employment_dataframe = function(empX) {
    empX = empX %>% fix_column_names() %>%
                    select(year = year,
                           business = management_business_and_financial_operations_occupations,
                           comp_math = computer_and_mathematical_occupations,
                           health = healthcare_practitioner_and_technical_occupations,
                           legal = legal_occupations)
    return(empX)
}
```

Transpose education dataframe, i.e. transform rows to columns and vice versa, using calls to select(), t(), rownames() and colnames() functions

```
transpose_education_dataframe = function(eduXY) {
    valuesXY = eduXY %>% select(X02:X14) %>% t()
    rownames(valuesXY) = c()

    labelsXY = eduXY %>% select(X1) %>% t()
    rownames(labelsXY) = c()

    eduXY = as.tibble(valuesXY)
    colnames(eduXY) = labelsXY
    return(eduXY)
}
```

Adjust columns of education dataframe, and add year column, using calls to rename_disciplines() function defined above, and to mutate(), row_number() and select() functions

```
adjust_education_dataframe = function(eduXY) {
    eduXY = eduXY %>% rename_disciplines() %>%
                      mutate(year = row_number() + 2001) %>%
                      select(year, business:mathematics) %>%
                      mutate(comp_math = computers + mathematics) %>%
                      select(year, business, comp_math, # computers, mathematics,
                             health, legal)
    return(eduXY)
}
```

Load Employment Data

Read data from multiple files

```
employment02 = read_csv('cpsaat09 2002-2003.csv', skip = 11, col_names = F) %>% slice(1:31)
employment03 = read_csv('cpsaat09 2003-2004.csv', skip = 11, col_names = F) %>% slice(1:31)
employment04 = read_csv('cpsaat09 2004-2005.csv', skip = 11, col_names = F) %>% slice(1:31)
employment05 = read_csv('cpsaat09 2005-2006.csv', skip = 11, col_names = F) %>% slice(1:31)
employment06 = read_csv('cpsaat09 2006-2007.csv', skip = 11, col_names = F) %>% slice(1:31)
employment07 = read_csv('cpsaat09 2007-2008.csv', skip = 11, col_names = F) %>% slice(1:31)
employment08 = read_csv('cpsaat09 2008-2009.csv', skip = 11, col_names = F) %>% slice(1:31)
employment09 = read_csv('cpsaat09 2009-2010.csv', skip = 11, col_names = F) %>% slice(1:31)
employment10 = read_csv('cpsaat09 2010-2011.csv', skip = 11, col_names = F) %>% slice(1:31)
employment11 = read_csv('cpsaat09 2011-2012.csv', skip = 11, col_names = F) %>% slice(1:31)
employment12 = read_csv('cpsaat09 2012-2013.csv', skip = 11, col_names = F) %>% slice(1:31)
employment13 = read_csv('cpsaat09 2013-2014.csv', skip = 11, col_names = F) %>% slice(1:31)
employment14 = read_csv('cpsaat09 2014-2015.csv', skip = 11, col_names = F) %>% slice(1:31)
```

Transform Employment Data

Combine data from multiple files into a single employment dataframe, i.e. one dataframe for each of the following cases: {males, females}

```
# Employment ("emp") dataframe for males ("M"): "empM"
empM = employment02 %>% select(X1, X02 = X6) %>%
                        add_column(X03 = employment03$X6) %>%
                        add_column(X04 = employment04$X6) %>%
                        add_column(X05 = employment05$X6) %>%
                        add_column(X06 = employment06$X6) %>%
                        add_column(X07 = employment07$X6) %>%
                        add_column(X08 = employment08$X6) %>%
                        add_column(X09 = employment09$X6) %>%
                        add_column(X10 = employment10$X6) %>%
                        add_column(X11 = employment11$X6) %>%
                        add_column(X12 = employment12$X6) %>%
                        add_column(X13 = employment13$X6) %>%
                        add_column(X14 = employment14$X6)

# Employment ("emp") dataframe for females ("F"): "empF"
empF = employment02 %>% select(X1, X02 = X10) %>%
                        add_column(X03 = employment03$X10) %>%
                        add_column(X04 = employment04$X10) %>%
                        add_column(X05 = employment05$X10) %>%
                        add_column(X06 = employment06$X10) %>%
                        add_column(X07 = employment07$X10) %>%
                        add_column(X08 = employment08$X10) %>%
                        add_column(X09 = employment09$X10) %>%
                        add_column(X10 = employment10$X10) %>%
                        add_column(X11 = employment11$X10) %>%
                        add_column(X12 = employment12$X10) %>%
                        add_column(X13 = employment13$X10) %>%
                        add_column(X14 = employment14$X10)
```

Transpose dataframe such that each row pertains to a particular year and each column pertains to the number of people in a particular discipline

```
empM = empM %>% transpose_employment_dataframe()
empF = empF %>% transpose_employment_dataframe()
```

Adjust columns of transposed dataframe

```
empM = empM %>% adjust_employment_dataframe()
empF = empF %>% adjust_employment_dataframe()
```

Compute dataframe of percentages of males and females working in each discipline. Then find average percentages across all years.

```
empP = empM %>% select(year) %>%
                add_column(business_m = empM$business / (empM$business + empF$business)) %>%
                add_column(business_f = empF$business / (empM$business + empF$business)) %>%
                add_column(comp_math_m = empM$comp_math / (empM$comp_math + empF$comp_math)) %>%
                add_column(comp_math_f = empF$comp_math / (empM$comp_math + empF$comp_math)) %>%
                add_column(health_m = empM$health / (empM$health + empF$health)) %>%
                add_column(health_f = empF$health / (empM$health + empF$health)) %>%
                add_column(legal_m = empM$legal / (empM$legal + empF$legal)) %>%
                add_column(legal_f = empF$legal / (empM$legal + empF$legal))
print(empP)
```

```
## # A tibble: 13 x 9
##     year business_m business_f comp_math_m comp_math_f  health_m  health_f
##    <dbl>      <dbl>      <dbl>       <dbl>       <dbl>     <dbl>     <dbl>
##  1  2002  0.5865964  0.4134036   0.7143318   0.2856682 0.2613422 0.7386578
##  2  2003  0.5788732  0.4211268   0.7116624   0.2883376 0.2628874 0.7371126
##  3  2004  0.5790960  0.4209040   0.7298077   0.2701923 0.2676941 0.7323059
##  4  2005  0.5753727  0.4246273   0.7301145   0.2698855 0.2678678 0.7321322
##  5  2006  0.5814470  0.4185530   0.7333960   0.2666040 0.2651494 0.7348506
##  6  2007  0.5734747  0.4265253   0.7437920   0.2562080 0.2631798 0.7368202
##  7  2008  0.5733891  0.4266109   0.7521181   0.2478819 0.2533279 0.7466721
##  8  2009  0.5726794  0.4273206   0.7520924   0.2479076 0.2541845 0.7458155
##  9  2010  0.5708196  0.4291804   0.7413204   0.2586796 0.2572641 0.7427359
## 10  2011  0.5686156  0.4313844   0.7493741   0.2506259 0.2558351 0.7441649
## 11  2012  0.5638406  0.4361594   0.7446138   0.2553862 0.2500629 0.7499371
## 12  2013  0.5662545  0.4337455   0.7388889   0.2611111 0.2556373 0.7443627
## 13  2014  0.5629707  0.4370293   0.7445068   0.2554932 0.2582703 0.7417297
## # ... with 2 more variables: legal_m <dbl>, legal_f <dbl>
```

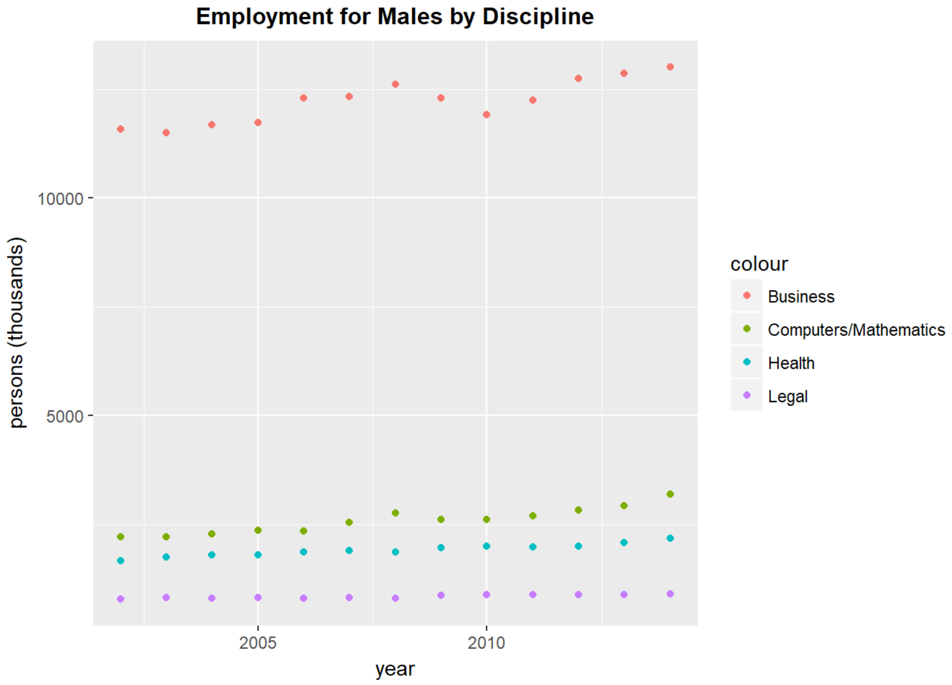
```
# Find average percentages across all years
empPAvg = empP %>% summarize(business_m_avg = mean(business_m),
                             business_f_avg = mean(business_f),
                             comp_math_m_avg = mean(comp_math_m),
                             comp_math_f_avg = mean(comp_math_f),
                             health_m_avg = mean(health_m),
                             health_f_avg = mean(health_f),
                             legal_m_avg = mean(legal_m),
                             legal_f_avg = mean(legal_f))
print(empPAvg)
```

```
## # A tibble: 1 x 8
##   business_m_avg business_f_avg comp_math_m_avg comp_math_f_avg
##            <dbl>          <dbl>           <dbl>           <dbl>
## 1      0.5733407      0.4266593       0.7373861       0.2626139
## # ... with 4 more variables: health_m_avg <dbl>, health_f_avg <dbl>,
## #   legal_m_avg <dbl>, legal_f_avg <dbl>
```

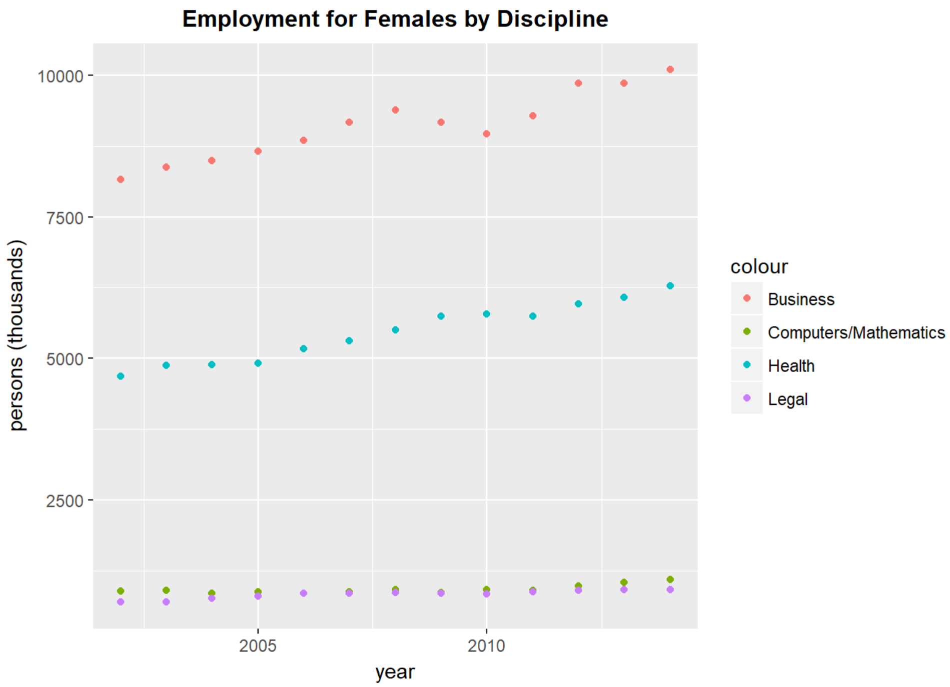
Explore Employment Data

Plot employment for the following cases: {males, females} X {business, computers/mathematics, health, legal}

```
# Employment for Males by Discipline
ggplot(data=empM) + geom_point(aes(x=year, y=business, colour='Business')) +
                    geom_point(aes(x=year, y=comp_math, colour='Computers/Mathematics')) +
                    geom_point(aes(x=year, y=health, colour='Health')) +
                    geom_point(aes(x=year, y=legal, colour='Legal')) +
                    ggtitle('Employment for Males by Discipline') +
                    theme(plot.title = element_text(size = 12, face = "bold")) +
                    ylab('persons (thousands)')
```



```
# Employment for Females by Discipline
ggplot(data=empF) + geom_point(aes(x=year, y=business, colour='Business')) +
                    geom_point(aes(x=year, y=comp_math, colour='Computers/Mathematics')) +
                    geom_point(aes(x=year, y=health, colour='Health')) +
                    geom_point(aes(x=year, y=legal, colour='Legal')) +
                    ggtitle('Employment for Females by Discipline') +
                    theme(plot.title = element_text(size = 12, face = "bold")) +
                    ylab('persons (thousands)')
```



The above plots of employment of males and females by discipline indicate that far more people work in business jobs than in any other category of jobs. For men, the computers/mathematics, health and legal sectors employ far fewer individuals than the business sector. For women, the health sector employs a substantial number of individuals, while the computers/mathematics and legal sectors employ far fewer individuals.

Moreover, the above plots appear to confirm the premise of this study, i.e. that women are under-represented in the hi-tech workforce relative to men.

Load Education Data

Read data from multiple files

```
education02 = read_csv('All Disciplines 2002-03.csv', skip = 13, col_names = F) %>% slice(1:1085)
education03 = read_csv('All Disciplines 2003-04.csv', skip = 12, col_names = F) %>% slice(1:1120)
education04 = read_csv('All Disciplines 2004-05.csv', skip = 6, col_names = F) %>% slice(1:1137)
education05 = read_csv('All Disciplines 2005-06.csv', skip = 6, col_names = F) %>% slice(1:1154)
education06 = read_csv('All Disciplines 2006-07.csv', skip = 7, col_names = F) %>% slice(1:1161)
education07 = read_csv('All Disciplines 2007-08.csv', skip = 7, col_names = F) %>% slice(1:1156)
education08 = read_csv('All Disciplines 2008-09.csv', skip = 6, col_names = F) %>% slice(1:1166) %>%
                                                                                   select(-X1, -X2)
colnames(education08) = c('X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8', 'X9', 'X10',
                          'X11', 'X12', 'X13', 'X14', 'X15')
education09 = read_csv('All Disciplines 2009-10.csv', skip = 5, col_names = F) %>% slice(1:1260)
education10 = read_csv('All Disciplines 2010-11.csv', skip = 5, col_names = F) %>% slice(1:1268)
education11 = read_csv('All Disciplines 2011-12.csv', skip = 5, col_names = F) %>% slice(1:1274)
education12 = read_csv('All Disciplines 2011-12.csv', skip = 5, col_names = F) %>% slice(1:1274)
education13 = read_csv('All Disciplines 2011-12.csv', skip = 5, col_names = F) %>% slice(1:1274)
education14 = read_csv('All Disciplines 2014-15.csv', skip = 5, col_names = F) %>% slice(1:1318)

eduMB02 = education02 %>% select(X1, X02 = X5) %>% fix_row_names() %>% find_disciplines()
eduMB03 = education03 %>% select(X1, X03 = X5) %>% fix_row_names() %>% find_disciplines()
eduMB04 = education04 %>% select(X1, X04 = X3) %>% fix_row_names() %>% find_disciplines()
eduMB05 = education05 %>% select(X1, X05 = X3) %>% fix_row_names() %>% find_disciplines()
eduMB06 = education06 %>% select(X1, X06 = X3) %>% fix_row_names() %>% find_disciplines()
eduMB07 = education07 %>% select(X1, X07 = X3) %>% fix_row_names() %>% find_disciplines()
eduMB08 = education08 %>% select(X1, X08 = X3) %>% fix_row_names() %>% find_disciplines()
eduMB09 = education09 %>% select(X1, X09 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMB10 = education10 %>% select(X1, X10 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMB11 = education11 %>% select(X1, X11 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMB12 = education12 %>% select(X1, X12 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMB13 = education13 %>% select(X1, X13 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMB14 = education14 %>% select(X1, X14 = X3) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)

eduFB02 = education02 %>% select(X1, X02 = X7) %>% fix_row_names() %>% find_disciplines()
eduFB03 = education03 %>% select(X1, X03 = X7) %>% fix_row_names() %>% find_disciplines()
eduFB04 = education04 %>% select(X1, X04 = X4) %>% fix_row_names() %>% find_disciplines()
eduFB05 = education05 %>% select(X1, X05 = X4) %>% fix_row_names() %>% find_disciplines()
eduFB06 = education06 %>% select(X1, X06 = X4) %>% fix_row_names() %>% find_disciplines()
eduFB07 = education07 %>% select(X1, X07 = X4) %>% fix_row_names() %>% find_disciplines()
eduFB08 = education08 %>% select(X1, X08 = X4) %>% fix_row_names() %>% find_disciplines()
eduFB09 = education09 %>% select(X1, X09 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFB10 = education10 %>% select(X1, X10 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFB11 = education11 %>% select(X1, X11 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFB12 = education12 %>% select(X1, X12 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFB13 = education13 %>% select(X1, X13 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFB14 = education14 %>% select(X1, X14 = X4) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)

eduMM02 = education02 %>% select(X1, X02 = X11) %>% fix_row_names() %>% find_disciplines()
eduMM03 = education03 %>% select(X1, X03 = X11) %>% fix_row_names() %>% find_disciplines()
eduMM04 = education04 %>% select(X1, X04 = X6) %>% fix_row_names() %>% find_disciplines()
eduMM05 = education05 %>% select(X1, X05 = X6) %>% fix_row_names() %>% find_disciplines()
eduMM06 = education06 %>% select(X1, X06 = X6) %>% fix_row_names() %>% find_disciplines()
eduMM07 = education07 %>% select(X1, X07 = X6) %>% fix_row_names() %>% find_disciplines()
eduMM08 = education08 %>% select(X1, X08 = X6) %>% fix_row_names() %>% find_disciplines()
eduMM09 = education09 %>% select(X1, X09 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMM10 = education10 %>% select(X1, X10 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMM11 = education11 %>% select(X1, X11 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMM12 = education12 %>% select(X1, X12 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMM13 = education13 %>% select(X1, X13 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMM14 = education14 %>% select(X1, X14 = X6) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)

eduFM02 = education02 %>% select(X1, X02 = X13) %>% fix_row_names() %>% find_disciplines()
eduFM03 = education03 %>% select(X1, X03 = X13) %>% fix_row_names() %>% find_disciplines()
eduFM04 = education04 %>% select(X1, X04 = X7) %>% fix_row_names() %>% find_disciplines()
eduFM05 = education05 %>% select(X1, X05 = X7) %>% fix_row_names() %>% find_disciplines()
eduFM06 = education06 %>% select(X1, X06 = X7) %>% fix_row_names() %>% find_disciplines()
eduFM07 = education07 %>% select(X1, X07 = X7) %>% fix_row_names() %>% find_disciplines()
eduFM08 = education08 %>% select(X1, X08 = X7) %>% fix_row_names() %>% find_disciplines()
eduFM09 = education09 %>% select(X1, X09 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFM10 = education10 %>% select(X1, X10 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFM11 = education11 %>% select(X1, X11 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFM12 = education12 %>% select(X1, X12 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFM13 = education13 %>% select(X1, X13 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFM14 = education14 %>% select(X1, X14 = X7) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
```

```
eduMD02 = education02 %>% select(X1, X02 = X17) %>% fix_row_names() %>% find_disciplines()
eduMD03 = education03 %>% select(X1, X03 = X17) %>% fix_row_names() %>% find_disciplines()
eduMD04 = education04 %>% select(X1, X04 = X9) %>% fix_row_names() %>% find_disciplines()
eduMD05 = education05 %>% select(X1, X05 = X9) %>% fix_row_names() %>% find_disciplines()
eduMD06 = education06 %>% select(X1, X06 = X9) %>% fix_row_names() %>% find_disciplines()
eduMD07 = education07 %>% select(X1, X07 = X9) %>% fix_row_names() %>% find_disciplines()
eduMD08 = education08 %>% select(X1, X08 = X9) %>% fix_row_names() %>% find_disciplines()
eduMD09 = education09 %>% select(X1, X09 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMD10 = education10 %>% select(X1, X10 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMD11 = education11 %>% select(X1, X11 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMD12 = education12 %>% select(X1, X12 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMD13 = education13 %>% select(X1, X13 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduMD14 = education14 %>% select(X1, X14 = X9) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)

eduFD02 = education02 %>% select(X1, X02 = X19) %>% fix_row_names() %>% find_disciplines()
eduFD03 = education03 %>% select(X1, X03 = X19) %>% fix_row_names() %>% find_disciplines()
eduFD04 = education04 %>% select(X1, X04 = X10) %>% fix_row_names() %>% find_disciplines()
eduFD05 = education05 %>% select(X1, X05 = X10) %>% fix_row_names() %>% find_disciplines()
eduFD06 = education06 %>% select(X1, X06 = X10) %>% fix_row_names() %>% find_disciplines()
eduFD07 = education07 %>% select(X1, X07 = X10) %>% fix_row_names() %>% find_disciplines()
eduFD08 = education08 %>% select(X1, X08 = X10) %>% fix_row_names() %>% find_disciplines()
eduFD09 = education09 %>% select(X1, X09 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFD10 = education10 %>% select(X1, X10 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFD11 = education11 %>% select(X1, X11 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFD12 = education12 %>% select(X1, X12 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFD13 = education13 %>% select(X1, X13 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
eduFD14 = education14 %>% select(X1, X14 = X10) %>% fix_row_names() %>% find_disciplines() %>% slice(1:5)
```

Transform Education Data

Combine data from multiple files into a single education dataframe, i.e. one dataframe for each of the following cases: {males, females} X {bachelors, masters, doctorate}

```r
# Education ("edu") dataframe for males-bachelors ("MB"): "eduMB"
eduMB = eduMB02 %>% add_column(X03 = eduMB03$X03) %>%
                    add_column(X04 = eduMB04$X04) %>%
                    add_column(X05 = eduMB05$X05) %>%
                    add_column(X06 = eduMB06$X06) %>%
                    add_column(X07 = eduMB07$X07) %>%
                    add_column(X08 = eduMB08$X08) %>%
                    add_column(X09 = eduMB09$X09) %>%
                    add_column(X10 = eduMB10$X10) %>%
                    add_column(X11 = eduMB11$X11) %>%
                    add_column(X12 = eduMB12$X12) %>%
                    add_column(X13 = eduMB13$X13) %>%
                    add_column(X14 = eduMB14$X14)

# Education ("edu") dataframe for females-bachelors ("FB"): "eduFB"
eduFB = eduFB02 %>% add_column(X03 = eduFB03$X03) %>%
                    add_column(X04 = eduFB04$X04) %>%
                    add_column(X05 = eduFB05$X05) %>%
                    add_column(X06 = eduFB06$X06) %>%
                    add_column(X07 = eduFB07$X07) %>%
                    add_column(X08 = eduFB08$X08) %>%
                    add_column(X09 = eduFB09$X09) %>%
                    add_column(X10 = eduFB10$X10) %>%
                    add_column(X11 = eduFB11$X11) %>%
                    add_column(X12 = eduFB12$X12) %>%
                    add_column(X13 = eduFB13$X13) %>%
                    add_column(X14 = eduFB14$X14)

# Education ("edu") dataframe for males-masters ("MM"): "eduMM"
eduMM = eduMM02 %>% add_column(X03 = eduMM03$X03) %>%
                    add_column(X04 = eduMM04$X04) %>%
                    add_column(X05 = eduMM05$X05) %>%
                    add_column(X06 = eduMM06$X06) %>%
                    add_column(X07 = eduMM07$X07) %>%
                    add_column(X08 = eduMM08$X08) %>%
                    add_column(X09 = eduMM09$X09) %>%
                    add_column(X10 = eduMM10$X10) %>%
                    add_column(X11 = eduMM11$X11) %>%
                    add_column(X12 = eduMM12$X12) %>%
                    add_column(X13 = eduMM13$X13) %>%
                    add_column(X14 = eduMM14$X14)

# Education ("edu") dataframe for females-masters ("FM"): "eduFM"
eduFM = eduFM02 %>% add_column(X03 = eduFM03$X03) %>%
                    add_column(X04 = eduFM04$X04) %>%
                    add_column(X05 = eduFM05$X05) %>%
                    add_column(X06 = eduFM06$X06) %>%
                    add_column(X07 = eduFM07$X07) %>%
                    add_column(X08 = eduFM08$X08) %>%
                    add_column(X09 = eduFM09$X09) %>%
                    add_column(X10 = eduFM10$X10) %>%
                    add_column(X11 = eduFM11$X11) %>%
                    add_column(X12 = eduFM12$X12) %>%
                    add_column(X13 = eduFM13$X13) %>%
                    add_column(X14 = eduFM14$X14)

# Education ("edu") dataframe for males-doctorate ("MD"): "eduMD"
eduMD = eduMD02 %>% add_column(X03 = eduMD03$X03) %>%
                    add_column(X04 = eduMD04$X04) %>%
                    add_column(X05 = eduMD05$X05) %>%
                    add_column(X06 = eduMD06$X06) %>%
                    add_column(X07 = eduMD07$X07) %>%
                    add_column(X08 = eduMD08$X08) %>%
                    add_column(X09 = eduMD09$X09) %>%
                    add_column(X10 = eduMD10$X10) %>%
                    add_column(X11 = eduMD11$X11) %>%
                    add_column(X12 = eduMD12$X12) %>%
                    add_column(X13 = eduMD13$X13) %>%
                    add_column(X14 = eduMD14$X14)

# Education ("edu") dataframe for females-doctorate ("FD"): "eduFD"
eduFD = eduFD02 %>% add_column(X03 = eduFD03$X03) %>%
                    add_column(X04 = eduFD04$X04) %>%
                    add_column(X05 = eduFD05$X05) %>%
                    add_column(X06 = eduFD06$X06) %>%
                    add_column(X07 = eduFD07$X07) %>%
                    add_column(X08 = eduFD08$X08) %>%
                    add_column(X09 = eduFD09$X09) %>%
                    add_column(X10 = eduFD10$X10) %>%
                    add_column(X11 = eduFD11$X11) %>%
                    add_column(X12 = eduFD12$X12) %>%
                    add_column(X13 = eduFD13$X13) %>%
                    add_column(X14 = eduFD14$X14)
```

Transpose dataframe such that each row pertains to a particular year and each column pertains to the number of people in a particular discipline

```
eduMB = eduMB %>% transpose_education_dataframe()
eduFB = eduFB %>% transpose_education_dataframe()
eduMM = eduMM %>% transpose_education_dataframe()
eduFM = eduFM %>% transpose_education_dataframe()
eduMD = eduMD %>% transpose_education_dataframe()
eduFD = eduFD %>% transpose_education_dataframe()
```

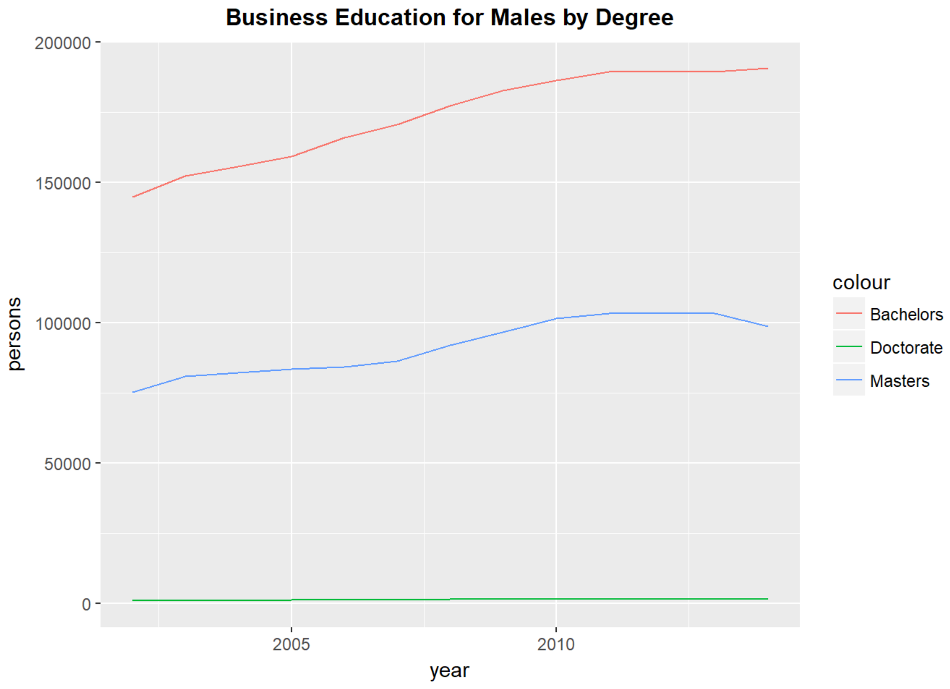
Adjust columns of transposed dataframe

```
eduMB = eduMB %>% adjust_education_dataframe()
eduFB = eduFB %>% adjust_education_dataframe()
eduMM = eduMM %>% adjust_education_dataframe()
eduFM = eduFM %>% adjust_education_dataframe()
eduMD = eduMD %>% adjust_education_dataframe()
eduFD = eduFD %>% adjust_education_dataframe()
```

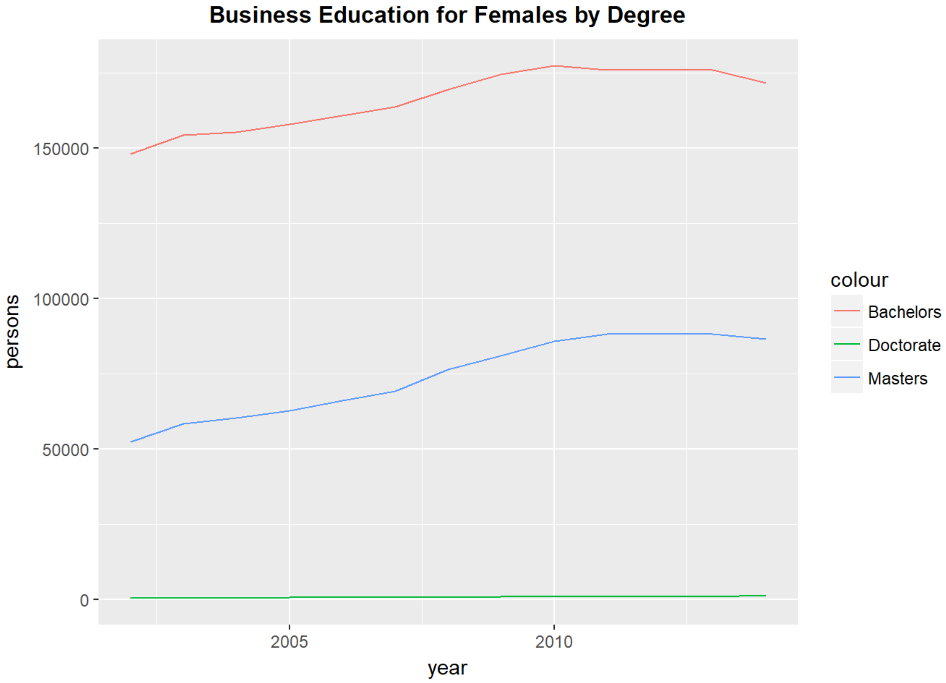
Explore Education Data

Plot education for the following cases: {males, females} X {business, computers/mathematics, health, legal} X {bachelors, masters, doctorate}

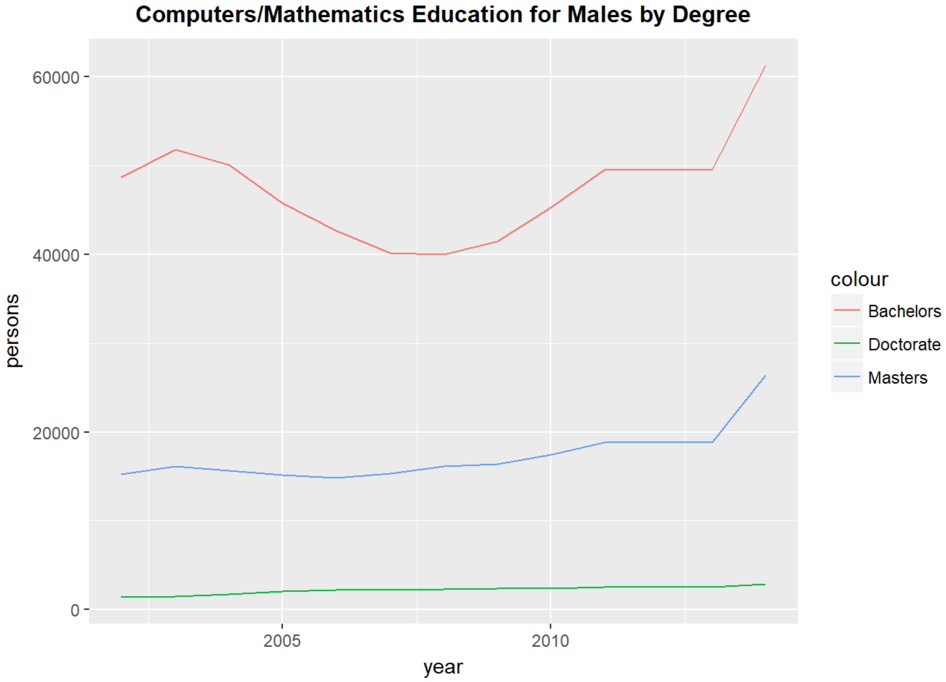
```
# Business Education for Males by Degree
ggplot() + geom_line(data=eduMB, aes(x=year, y=business, colour='Bachelors')) +
           geom_line(data=eduMM, aes(x=year, y=business, colour='Masters')) +
           geom_line(data=eduMD, aes(x=year, y=business, colour='Doctorate')) +
           ggtitle('Business Education for Males by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



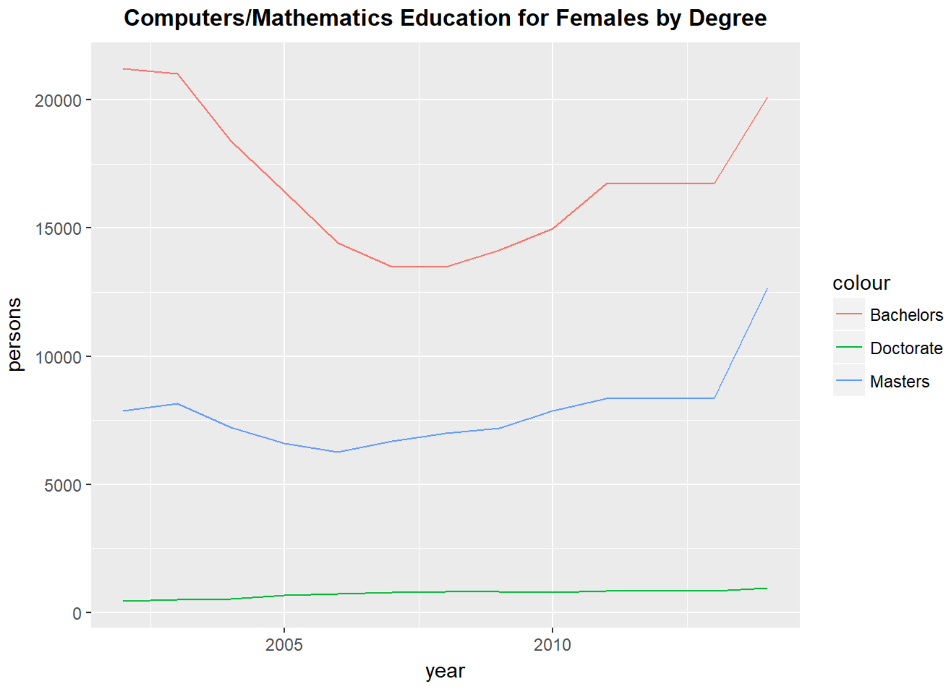
```
# Business Education for Females by Degree
ggplot() + geom_line(data=eduFB, aes(x=year, y=business, colour='Bachelors')) +
           geom_line(data=eduFM, aes(x=year, y=business, colour='Masters')) +
           geom_line(data=eduFD, aes(x=year, y=business, colour='Doctorate')) +
           ggtitle('Business Education for Females by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



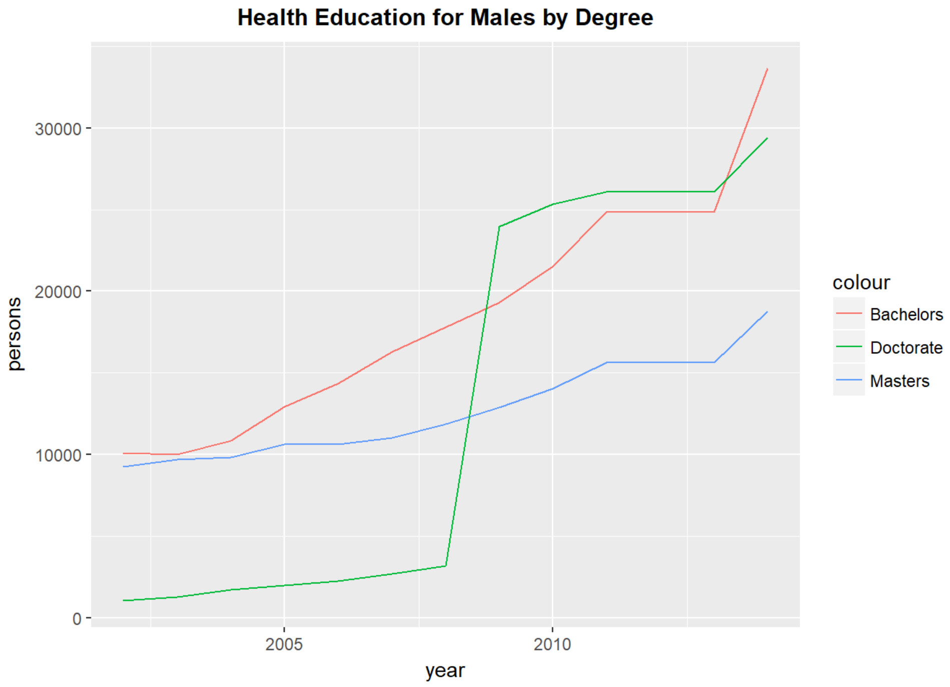
```
# Computers/Mathematics Education for Males by Degree
ggplot() + geom_line(data=eduMB, aes(x=year, y=comp_math, colour='Bachelors')) +
           geom_line(data=eduMM, aes(x=year, y=comp_math, colour='Masters')) +
           geom_line(data=eduMD, aes(x=year, y=comp_math, colour='Doctorate')) +
           ggtitle('Computers/Mathematics Education for Males by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



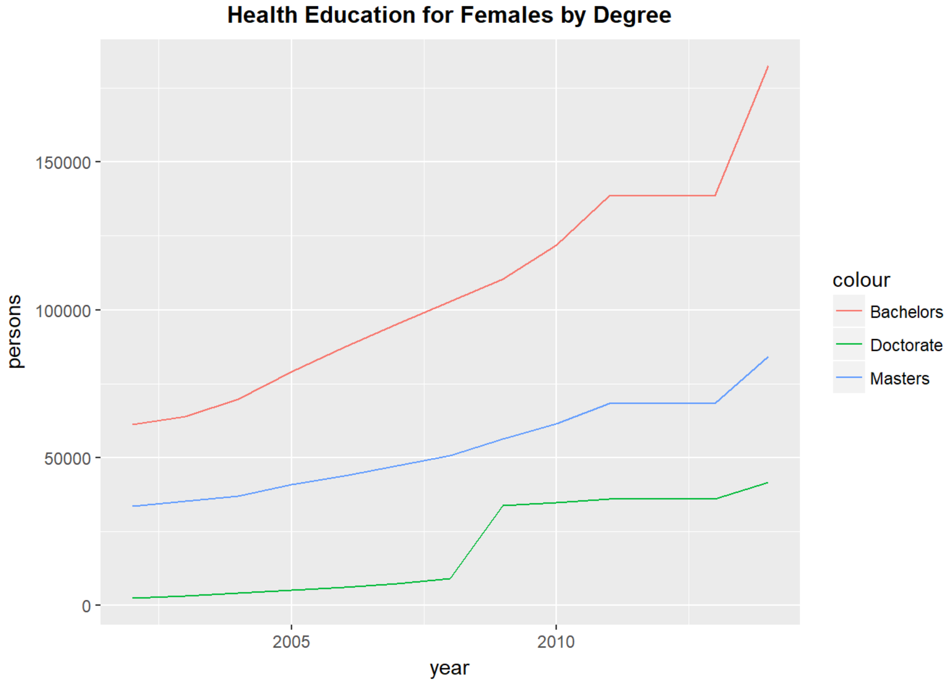
```
# Computers/Mathematics Education for Females by Degree
ggplot() + geom_line(data=eduFB, aes(x=year, y=comp_math, colour='Bachelors')) +
           geom_line(data=eduFM, aes(x=year, y=comp_math, colour='Masters')) +
           geom_line(data=eduFD, aes(x=year, y=comp_math, colour='Doctorate')) +
           ggtitle('Computers/Mathematics Education for Females by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



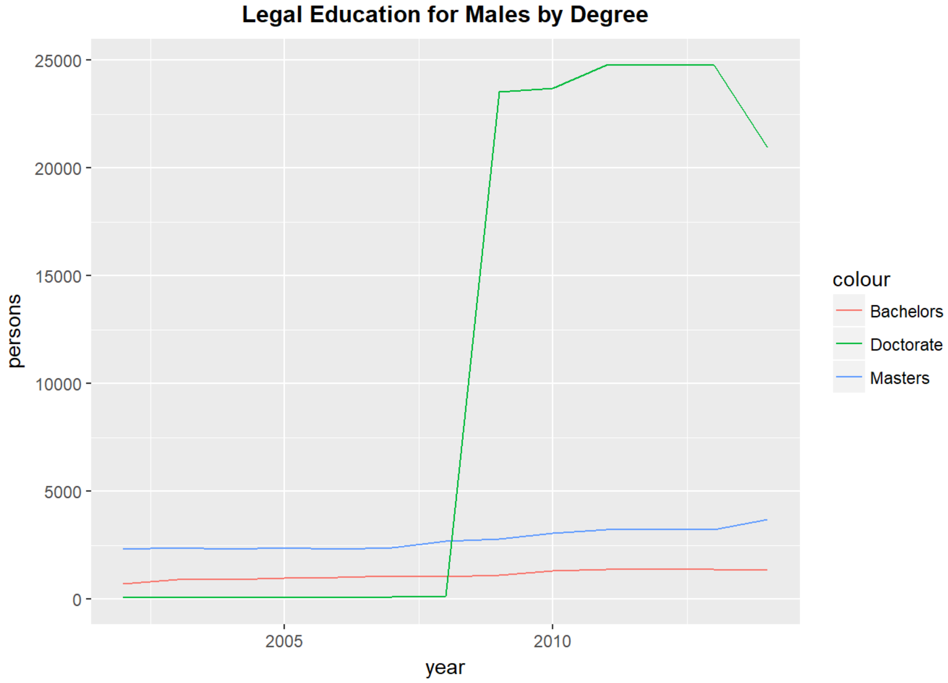
```
# Health Education for Males by Degree
ggplot() + geom_line(data=eduMB, aes(x=year, y=health, colour='Bachelors')) +
           geom_line(data=eduMM, aes(x=year, y=health, colour='Masters')) +
           geom_line(data=eduMD, aes(x=year, y=health, colour='Doctorate')) +
           ggtitle('Health Education for Males by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



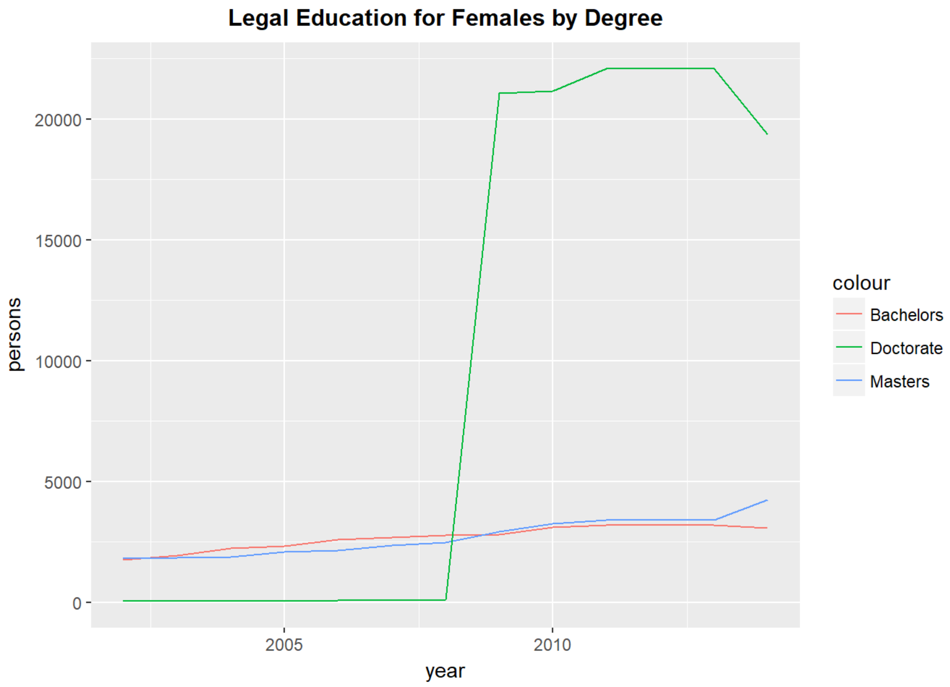
```
# Health Education for Females by Degree
ggplot() + geom_line(data=eduFB, aes(x=year, y=health, colour='Bachelors')) +
           geom_line(data=eduFM, aes(x=year, y=health, colour='Masters')) +
           geom_line(data=eduFD, aes(x=year, y=health, colour='Doctorate')) +
           ggtitle('Health Education for Females by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



```
# Legal Education for Males by Degree
ggplot() + geom_line(data=eduMB, aes(x=year, y=legal, colour='Bachelors')) +
           geom_line(data=eduMM, aes(x=year, y=legal, colour='Masters')) +
           geom_line(data=eduMD, aes(x=year, y=legal, colour='Doctorate')) +
           ggtitle('Legal Education for Males by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



```
# Legal Education for Females by Degree
ggplot() + geom_line(data=eduFB, aes(x=year, y=legal, colour='Bachelors')) +
           geom_line(data=eduFM, aes(x=year, y=legal, colour='Masters')) +
           geom_line(data=eduFD, aes(x=year, y=legal, colour='Doctorate')) +
           ggtitle('Legal Education for Females by Degree') +
           theme(plot.title = element_text(size = 12, face = "bold")) +
           ylab('persons')
```



The above plots of male and female graduates for different disciplines and degree levels indicate that, in general, bachelor degrees outnumber masters degrees, which in turn outnumber doctorates. This data agrees with intuition, as the pool of candidates for higher degrees are drawn (as a subset) from the pool of graduates with lower degrees.

(Of course, instances in which foreign graduates with lower degrees pursue higher degrees in the US constitute exceptions to this intuition.)

Note that the plots for the health and legal disciplines display sharp increases in the number of doctorates between 2007 and 2009. These increases appear highly suspect, as in some plots they are several orders of magnitude, and they show doctorates actually outnumbering bachelors and masters.

(We did check the raw data in the XLS/CSV files obtained from the NCES website to confirm that the increases in doctorates are present in the raw data, and are not artifacts of our processing of that raw data.)

Because of these suspect increases, when comparing the percentage of males and females in the educational pipeline and the professional workforce in subsequent sections, we either exclude data from the health and legal discplines, or exclude data on doctorates, from our analysis.

Link Employment and Education Data

Compute total degrees across all levels (bachelors, masters and doctorate) for each sex and for each discipline. Then compute percentages of males and females earning degrees in each discipline. Next find average percentages across all years.

```
# Compute dataframe of totals for males across all degree levels, i.e.
# education ("edu") dataframe for males ("M"): "eduM"
eduM = eduMB %>% select(year) %>%
                 add_column(business = eduMB$business + eduMM$business + eduMD$business) %>%
                 add_column(comp_math = eduMB$comp_math + eduMM$comp_math + eduMD$comp_math) %>%
                 add_column(health = eduMB$health + eduMM$health + eduMD$health) %>%
                 add_column(legal = eduMB$legal + eduMM$legal + eduMD$legal)

# Compute dataframe of totals for females across all degree levels, i.e.
# education ("edu") dataframe for females ("F"): "eduF"
eduF = eduFB %>% select(year) %>%
                 add_column(business = eduFB$business + eduFM$business + eduFD$business) %>%
                 add_column(comp_math = eduFB$comp_math + eduFM$comp_math + eduFD$comp_math) %>%
                 add_column(health = eduFB$health + eduFM$health + eduFD$health) %>%
                 add_column(legal = eduFB$legal + eduFM$legal + eduFD$legal)

# Compute dataframe of percentages of males and females across all degree levels, i.e.
# education ("edu") dataframe of percentages ("P"): "eduP"
# Cover the following cases:
# {males, females} X {business, computers/mathematics | bachelors + masters + doctorate}
eduP = eduM %>% select(year) %>%
                add_column(business_m = eduM$business / (eduM$business + eduF$business)) %>%
                add_column(business_f = eduF$business / (eduM$business + eduF$business)) %>%
                add_column(comp_math_m = eduM$comp_math / (eduM$comp_math + eduF$comp_math)) %>%
                add_column(comp_math_f = eduF$comp_math / (eduM$comp_math + eduF$comp_math))
print(eduP)
```

```
## # A tibble: 13 x 5
##     year business_m business_f comp_math_m comp_math_f
##    <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
##  1  2002  0.5236683  0.4763317   0.6889781   0.3110219
##  2  2003  0.5231401  0.4768599   0.7004681   0.2995319
##  3  2004  0.5244736  0.4755264   0.7204378   0.2795622
##  4  2005  0.5240880  0.4759120   0.7265286   0.2734714
##  5  2006  0.5246082  0.4753918   0.7361121   0.2638879
##  6  2007  0.5244620  0.4755380   0.7337961   0.2662039
##  7  2008  0.5230568  0.4769432   0.7327873   0.2672127
##  8  2009  0.5228394  0.4771606   0.7312888   0.2687112
##  9  2010  0.5228093  0.4771907   0.7334445   0.2665555
## 10  2011  0.5257617  0.4742383   0.7322516   0.2677484
## 11  2012  0.5257617  0.4742383   0.7322516   0.2677484
## 12  2013  0.5257617  0.4742383   0.7322516   0.2677484
## 13  2014  0.5283666  0.4716334   0.7284868   0.2715132
```

```
# Find average percentages across all years
eduPAvg = eduP %>% summarize(business_m_avg = mean(business_m),
                             business_f_avg = mean(business_f),
                             comp_math_m_avg = mean(comp_math_m),
                             comp_math_f_avg = mean(comp_math_f))
print(eduPAvg)
```

```
## # A tibble: 1 x 4
##   business_m_avg business_f_avg comp_math_m_avg comp_math_f_avg
##            <dbl>          <dbl>           <dbl>           <dbl>
## 1      0.5245229      0.4754771       0.7253141       0.2746859
```

Join employment data with education data

```
# Join employment dataframe ("empP") with education dataframe ("eduP"): "empEduP"
empEduP = empP %>% left_join(eduP, by = 'year') %>%
                   rename(business_m_emp = business_m.x,
                          business_f_emp = business_f.x,
                          comp_math_m_emp = comp_math_m.x,
                          comp_math_f_emp = comp_math_f.x,
                          business_m_edu = business_m.y,
                          business_f_edu = business_f.y,
                          comp_math_m_edu = comp_math_m.y,
                          comp_math_f_edu = comp_math_f.y) %>%
                   select(-health_m, -health_f, -legal_m, -legal_f)
print(empEduP)
```

```
## # A tibble: 13 x 9
##     year business_m_emp business_f_emp comp_math_m_emp comp_math_f_emp
##    <dbl>          <dbl>          <dbl>           <dbl>           <dbl>
##  1  2002      0.5865964      0.4134036       0.7143318       0.2856682
##  2  2003      0.5788732      0.4211268       0.7116624       0.2883376
##  3  2004      0.5790960      0.4209040       0.7298077       0.2701923
##  4  2005      0.5753727      0.4246273       0.7301145       0.2698855
##  5  2006      0.5814470      0.4185530       0.7333960       0.2666040
##  6  2007      0.5734747      0.4265253       0.7437920       0.2562080
##  7  2008      0.5733891      0.4266109       0.7521181       0.2478819
##  8  2009      0.5726794      0.4273206       0.7520924       0.2479076
##  9  2010      0.5708196      0.4291804       0.7413204       0.2586796
## 10  2011      0.5686156      0.4313844       0.7493741       0.2506259
## 11  2012      0.5638406      0.4361594       0.7446138       0.2553862
## 12  2013      0.5662545      0.4337455       0.7388889       0.2611111
## 13  2014      0.5629707      0.4370293       0.7445068       0.2554932
## # ... with 4 more variables: business_m_edu <dbl>, business_f_edu <dbl>,
## #   comp_math_m_edu <dbl>, comp_math_f_edu <dbl>
```

Compute total degrees across bachelors and masters levels (excluding doctorate) for each sex and for each discipline. Then compute percentages of males and females earning degrees in each discipline. Next find average percentages across all years.

```
# Compute dataframe of totals for males for bachelors and masters degrees, i.e.
# education ("edu") dataframe for males-bachelors-masters ("MBM"): "eduMBM"
eduMBM = eduMB %>% select(year) %>%
                 add_column(business = eduMB$business + eduMM$business) %>%
                 add_column(comp_math = eduMB$comp_math + eduMM$comp_math) %>%
                 add_column(health = eduMB$health + eduMM$health) %>%
                 add_column(legal = eduMB$legal + eduMM$legal)

# Compute dataframe of totals for females for bachelors and masters degrees, i.e.
# education ("edu") dataframe for females-bachelors-masters ("FBM"): "eduFBM"
eduFBM = eduFB %>% select(year) %>%
                 add_column(business = eduFB$business + eduFM$business) %>%
                 add_column(comp_math = eduFB$comp_math + eduFM$comp_math) %>%
                 add_column(health = eduFB$health + eduFM$health) %>%
                 add_column(legal = eduFB$legal + eduFM$legal)

# Compute dataframe of percentages of males and females for bachelors and masters degrees, i.e.
# education ("edu") dataframe for bachelors-masters-percentages ("BMP"): "eduBMP"
# Cover the following cases:
# {males, females} X {business, computers/mathematics, health, legal | bachelors + masters}
eduBMP = eduMBM %>% select(year) %>%
                add_column(business_m = eduMBM$business / (eduMBM$business + eduFBM$business)) %>%
                add_column(business_f = eduFBM$business / (eduMBM$business + eduFBM$business)) %>%
                add_column(comp_math_m = eduMBM$comp_math / (eduMBM$comp_math + eduFBM$comp_math)) %>%
                add_column(comp_math_f = eduFBM$comp_math / (eduMBM$comp_math + eduFBM$comp_math)) %>%
                add_column(health_m = eduMBM$health / (eduMBM$health + eduFBM$health)) %>%
                add_column(health_f = eduFBM$health / (eduMBM$health + eduFBM$health)) %>%
                add_column(legal_m = eduMBM$legal / (eduMBM$legal + eduFBM$legal)) %>%
                add_column(legal_f = eduFBM$legal / (eduMBM$legal + eduFBM$legal))
print(eduBMP)
```

```
## # A tibble: 13 x 9
##     year business_m business_f comp_math_m comp_math_f  health_m  health_f
##    <dbl>      <dbl>      <dbl>       <dbl>       <dbl>     <dbl>     <dbl>
##  1  2002  0.5232762  0.4767238   0.6876242   0.3123758 0.1699170 0.8300830
##  2  2003  0.5227248  0.4772752   0.6995234   0.3004766 0.1656137 0.8343863
##  3  2004  0.5242194  0.4757806   0.7194242   0.2805758 0.1622916 0.8377084
##  4  2005  0.5237596  0.4762404   0.7259118   0.2740882 0.1642379 0.8357621
##  5  2006  0.5243491  0.4756509   0.7355061   0.2644939 0.1596574 0.8403426
##  6  2007  0.5241415  0.4758585   0.7335710   0.2664290 0.1609453 0.8390547
##  7  2008  0.5226851  0.4773149   0.7327564   0.2672436 0.1619864 0.8380136
##  8  2009  0.5225380  0.4774620   0.7308790   0.2691210 0.1619229 0.8380771
##  9  2010  0.5225156  0.4774844   0.7326140   0.2673860 0.1624454 0.8375546
## 10  2011  0.5255296  0.4744704   0.7315608   0.2684392 0.1637185 0.8362815
## 11  2012  0.5255296  0.4744704   0.7315608   0.2684392 0.1637185 0.8362815
## 12  2013  0.5255296  0.4744704   0.7315608   0.2684392 0.1637185 0.8362815
## 13  2014  0.5282395  0.4717605   0.7278355   0.2721645 0.1643181 0.8356819
## # ... with 2 more variables: legal_m <dbl>, legal_f <dbl>
```

```
# Find average percentages across all years
eduBMPAvg = eduBMP %>% summarize(business_m_avg = mean(business_m),
                                 business_f_avg = mean(business_f),
                                 comp_math_m_avg = mean(comp_math_m),
                                 comp_math_f_avg = mean(comp_math_f),
                                 health_m_avg = mean(health_m),
                                 health_f_avg = mean(health_f),
                                 legal_m_avg = mean(legal_m),
                                 legal_f_avg = mean(legal_f))
print(eduBMPAvg)
```

```
## # A tibble: 1 x 8
##   business_m_avg business_f_avg comp_math_m_avg comp_math_f_avg
##            <dbl>          <dbl>           <dbl>           <dbl>
## 1      0.5242336      0.4757664       0.7246406       0.2753594
## # ... with 4 more variables: health_m_avg <dbl>, health_f_avg <dbl>,
## #   legal_m_avg <dbl>, legal_f_avg <dbl>
```

Join employment data with education data

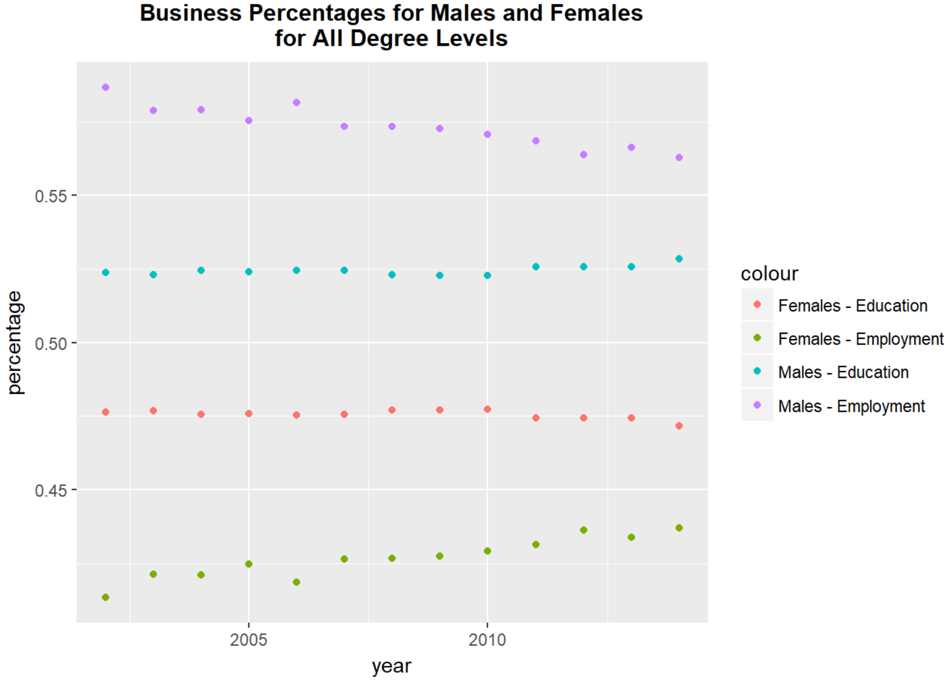
```
# Join employment dataframe ("empP") with education dataframe ("eduBMP"): "empEduBMP"
empEduBMP = empP %>% left_join(eduBMP, by = 'year') %>%
                     rename(business_m_emp = business_m.x,
                            business_f_emp = business_f.x,
                            comp_math_m_emp = comp_math_m.x,
                            comp_math_f_emp = comp_math_f.x,
                            health_m_emp = health_m.x,
                            health_f_emp = health_f.x,
                            legal_m_emp = legal_m.x,
                            legal_f_emp = legal_f.x,
                            business_m_edu = business_m.y,
                            business_f_edu = business_f.y,
                            comp_math_m_edu = comp_math_m.y,
                            comp_math_f_edu = comp_math_f.y,
                            health_m_edu = health_m.y,
                            health_f_edu = health_f.y,
                            legal_m_edu = legal_m.y,
                            legal_f_edu = legal_f.y)
print(empEduBMP)
```

```
## # A tibble: 13 x 17
##     year business_m_emp business_f_emp comp_math_m_emp comp_math_f_emp
##    <dbl>          <dbl>          <dbl>           <dbl>           <dbl>
##  1  2002      0.5865964      0.4134036       0.7143318       0.2856682
##  2  2003      0.5788732      0.4211268       0.7116624       0.2883376
##  3  2004      0.5790960      0.4209040       0.7298077       0.2701923
##  4  2005      0.5753727      0.4246273       0.7301145       0.2698855
##  5  2006      0.5814470      0.4185530       0.7333960       0.2666040
##  6  2007      0.5734747      0.4265253       0.7437920       0.2562080
##  7  2008      0.5733891      0.4266109       0.7521181       0.2478819
##  8  2009      0.5726794      0.4273206       0.7520924       0.2479076
##  9  2010      0.5708196      0.4291804       0.7413204       0.2586796
## 10  2011      0.5686156      0.4313844       0.7493741       0.2506259
## 11  2012      0.5638406      0.4361594       0.7446138       0.2553862
## 12  2013      0.5662545      0.4337455       0.7388889       0.2611111
## 13  2014      0.5629707      0.4370293       0.7445068       0.2554932
## # ... with 12 more variables: health_m_emp <dbl>, health_f_emp <dbl>,
## #   legal_m_emp <dbl>, legal_f_emp <dbl>, business_m_edu <dbl>,
## #   business_f_edu <dbl>, comp_math_m_edu <dbl>, comp_math_f_edu <dbl>,
## #   health_m_edu <dbl>, health_f_edu <dbl>, legal_m_edu <dbl>,
## #   legal_f_edu <dbl>
```

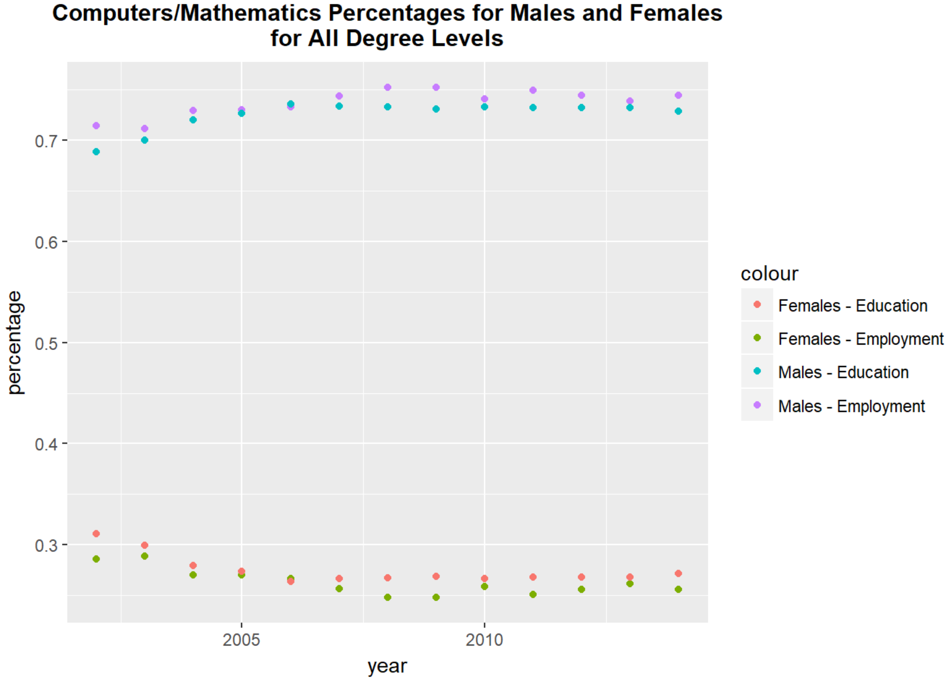
Analyze Linked Employment and Education Data - All Degree Levels Included

Plot employment and education degree levels percentages for the following cases: {males, females} X {business, computers/mathematics | bachelors + masters + doctorate}

```
# Business Percentages for Males and Females for All Degree Levels
ggplot(data=empEduP) + geom_point(aes(x=year, y=business_m_emp, colour='Males - Employment')) +
                       geom_point(aes(x=year, y=business_f_emp, colour='Females - Employment')) +
                       geom_point( aes(x=year, y=business_m_edu, colour='Males - Education')) +
                       geom_point(aes(x=year, y=business_f_edu, colour='Females - Education')) +
                       ggtitle('Business Percentages for Males and Females\nfor All Degree Levels') +
                       theme(plot.title = element_text(size = 12, face = "bold")) +
                       ylab('percentage')
```



```
# Computers/Mathematics Percentages for Males and Females for All Degree Levels
ggplot(data=empEduP) + geom_point(aes(x=year, y=comp_math_m_emp, colour='Males - Employment')) +
                       geom_point(aes(x=year, y=comp_math_f_emp, colour='Females - Employment')) +
                       geom_point(aes(x=year, y=comp_math_m_edu, colour='Males - Education')) +
                       geom_point(aes(x=year, y=comp_math_f_edu, colour='Females - Education')) +
                       ggtitle('Computers/Mathematics Percentages for Males and Females\nfor All Degree Levels')
+
                       theme(plot.title = element_text(size = 12, face = "bold")) +
                       ylab('percentage')
```



The above plots of the percentages of males and females in business and computers/mathematics in both the educational pipeline and the professional workforce show that: (1) the percentage of men in the workforce is larger than the percentage of men in the pipeline; (2) the percentage of women in the workforce is smaller than the percentage of women in the pipeline. This is consistent with the notion that more women pursue higher education than men, but that more women drop out of the workforce than men.

Moreover, comparison of the plot for business with the plot for computers/mathematics indicates that: (1) the percentage of women in the workplace is approximately equal to the percentage of in the women pipeline for computers/ mathematics; (2) the percentage of women in the workplace is significantly less than the percentage of women in the pipeline for business.

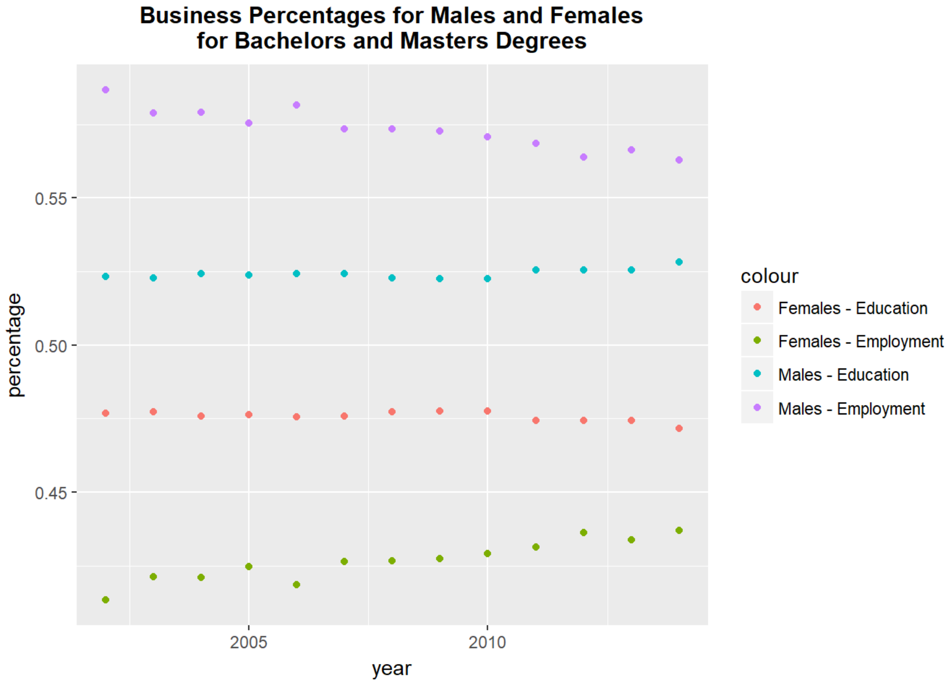
This data indicates that the under-representation of women in hi-tech arises from issues in the educational pipeline, rather than from issues in the professional workplace, since the percentage of women in the workforce closely tracks the percentage of women in the pipeline for computers/mathematics disciplines. In sum, the under-representation of women in hi-tech occupations arises in the educational pipeline rather than the professional workplace.

To put it another way, if the under-representation of women in computers/mathematics disciplines were caused by issues with workforce recruitment or retention, then the workforce percentage of women would be significantly lower than the pipeline percentage of women. We do not see this for computers/mathematics disciplines, while we do see this for business disciplines.

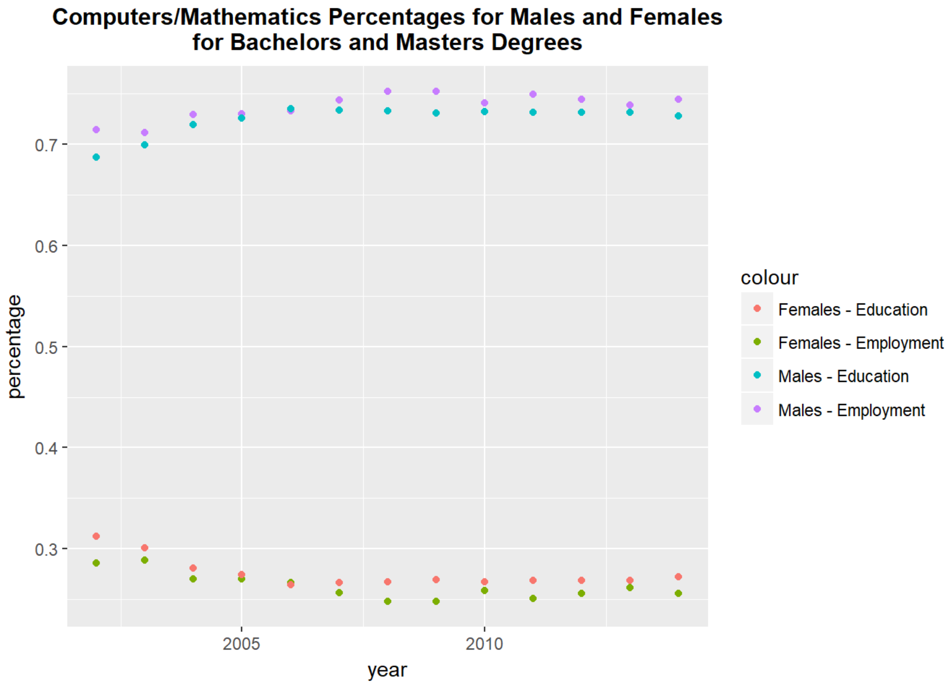
Analyze Linked Employment and Education Data - Doctorate Degrees Excluded

Plot employment and education degree levels percentages for the following cases: {males, females} X {business, computers/mathematics, health, legal | bachelors + masters}

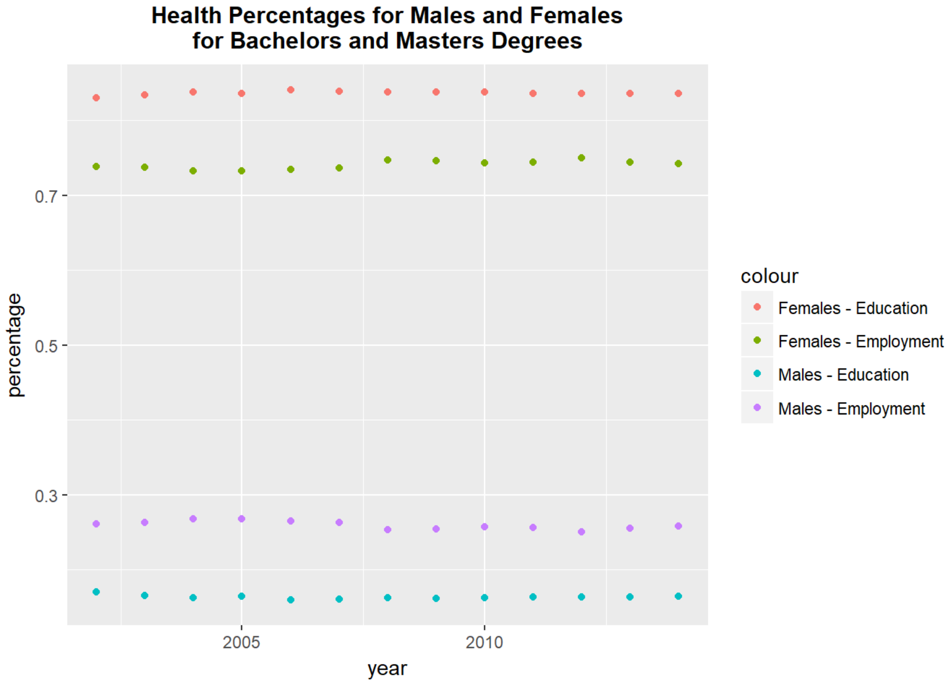
```
# Business Percentages for Males and Females for Bachelors and Masters Degrees
ggplot(data=empEduBMP) +
    geom_point(aes(x=year, y=business_m_emp, colour='Males - Employment')) +
    geom_point(aes(x=year, y=business_f_emp, colour='Females - Employment')) +
    geom_point(aes(x=year, y=business_m_edu, colour='Males - Education')) +
    geom_point(aes(x=year, y=business_f_edu, colour='Females - Education')) +
    ggtitle('Business Percentages for Males and Females\nfor Bachelors and Masters Degrees') +
    theme(plot.title = element_text(size = 12, face = "bold")) +
    ylab('percentage')
```



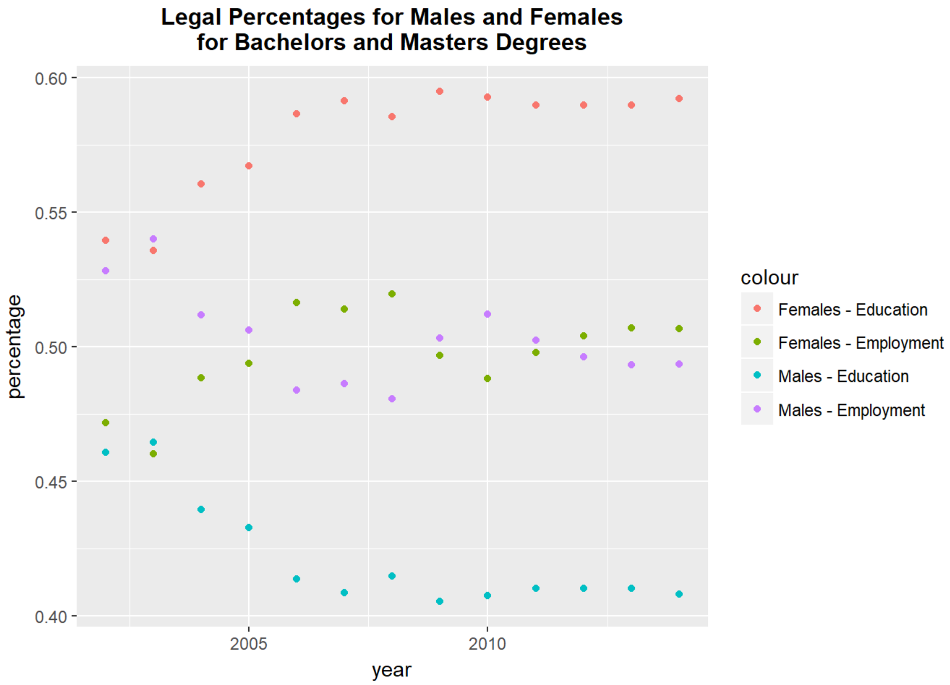
```
# Computers/Mathematics Percentages for Males and Females for Bachelors and Masters Degrees
ggplot(data=empEduBMP) +
    geom_point(aes(x=year, y=comp_math_m_emp, colour='Males - Employment')) +
    geom_point(aes(x=year, y=comp_math_f_emp, colour='Females - Employment')) +
    geom_point(aes(x=year, y=comp_math_m_edu, colour='Males - Education')) +
    geom_point(aes(x=year, y=comp_math_f_edu, colour='Females - Education')) +
    ggtitle('Computers/Mathematics Percentages for Males and Females\nfor Bachelors and Masters Degrees') +
    theme(plot.title = element_text(size = 12, face = "bold")) +
    ylab('percentage')
```



```
# Health Percentages for Males and Females for Bachelors and Masters Degrees
ggplot(data=empEduBMP) +
    geom_point(aes(x=year, y=health_m_emp, colour='Males - Employment')) +
    geom_point(aes(x=year, y=health_f_emp, colour='Females - Employment')) +
    geom_point(aes(x=year, y=health_m_edu, colour='Males - Education')) +
    geom_point(aes(x=year, y=health_f_edu, colour='Females - Education')) +
    ggtitle('Health Percentages for Males and Females\nfor Bachelors and Masters Degrees') +
    theme(plot.title = element_text(size = 12, face = "bold")) +
    ylab('percentage')
```



```
# Legal Percentages for Males and Females for Bachelors and Masters Degrees
ggplot(data=empEduBMP) +
    geom_point(aes(x=year, y=legal_m_emp, colour='Males - Employment')) +
    geom_point(aes(x=year, y=legal_f_emp, colour='Females - Employment')) +
    geom_point(aes(x=year, y=legal_m_edu, colour='Males - Education')) +
    geom_point(aes(x=year, y=legal_f_edu, colour='Females - Education')) +
    ggtitle('Legal Percentages for Males and Females\nfor Bachelors and Masters Degrees') +
    theme(plot.title = element_text(size = 12, face = "bold")) +
    ylab('percentage')
```



Comparison of the plots above for business and computers/mathematics indicates that: (1) the percentage of women in the workplace is approximately equal to the percentage of in the women pipeline for computers/mathematics; (2) the percentage of women in the workplace is significantly less than the percentage of women in the pipeline for business.

This data indicates that the under-representation of women in hi-tech arises from issues in the educational pipeline, rather than from issues in the professional workplace, since the percentage of women in the workforce closely tracks the percentage of women in the pipeline for computers/mathematics disciplines.

To put it another way, if the under-representation of women in computers/mathematics disciplines were caused by issues with workforce recruitment or retention, then the workforce percentage of women would be significantly lower than the pipeline percentage of women. We do not see this for computers/mathematics disciplines, while we do see this for business disciplines.

Furthermore, for health and legal disciplines, the workforce percentage of women is smaller than the pipeline percentage of women, even though the percentage of women is greater than the percentage of men in those fields. In other words, workforce recruitment and retention appear to be a bigger problem in the health and legal fields, in which women are over-represented, than in the computers/mathematics fields.

This implies that the under-representation of women in hi-tech occupations cannot be solved by focusing on the professional workplace, and can only be solved by focusing on the educational pipeline.