EDA

Ariel Wentworth

September 2020

```
library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

The Gender Gap: by Degree Level

We will first do some exploration of degrees conferred, separated by degree level.

Doctor's

```
doctors <- read_csv('data/doctors.csv')
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   Year = col_character(),
##   Total = col_double(),
##   White = col_double(),
##   Black = col_double(),
##   Hispanic = col_double(),
##   `Asian/Pacific Islander` = col_double(),
##   `American Indian/Alaska Native` = col_double(),
##   `Two or more races` = col_double(),
##   `Non-resident alien` = col_double()
## )
```

```r
head(doctors)
```

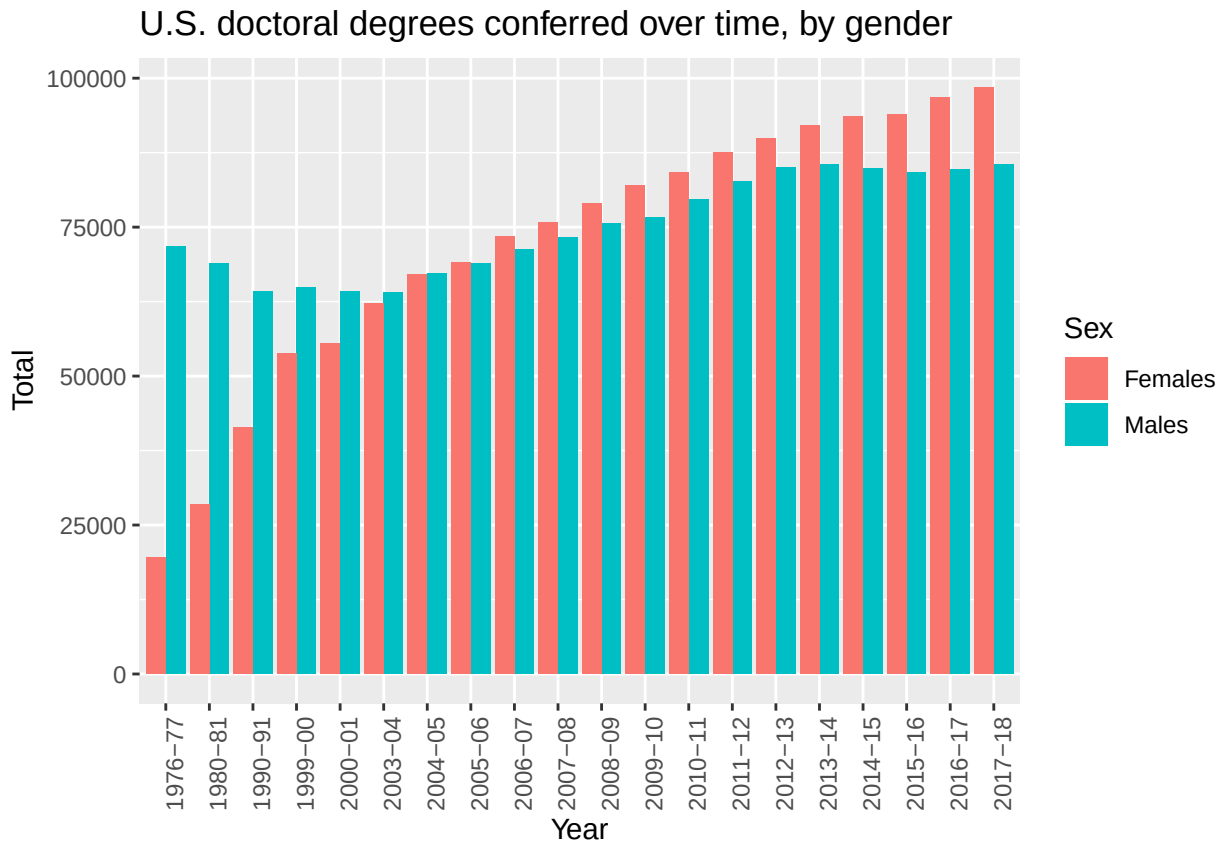
```
## # A tibble: 6 x 10
##   Sex   Year   Total White Black Hispanic `Asian/Pacific ~ `American India~
##   <chr> <chr>  <dbl> <dbl> <dbl>    <dbl>            <dbl>             <dbl>
## 1 Total 1976~  91218 79932  3575     1533             1674               240
## 2 Total 1980~  97281 84200  3893     1924             2267               312
## 3 Total 1990~ 105547 81791  4429     3210             5120               356
## 4 Total 1999~ 118736 82984  7078     5042            10682               708
## 5 Total 2000~ 119585 82321  7035     5204            11587               705
## 6 Total 2003~ 126087 84695  8089     5795            12371               771
## # ... with 2 more variables: `Two or more races` <dbl>, `Non-resident
## #   alien` <dbl>
```

```r
doctors_by_sex <- doctors %>%
  select(Sex, Year, Total) %>%
  group_by(Year) %>%
  filter(Sex!="Total")

head(doctors_by_sex)
```

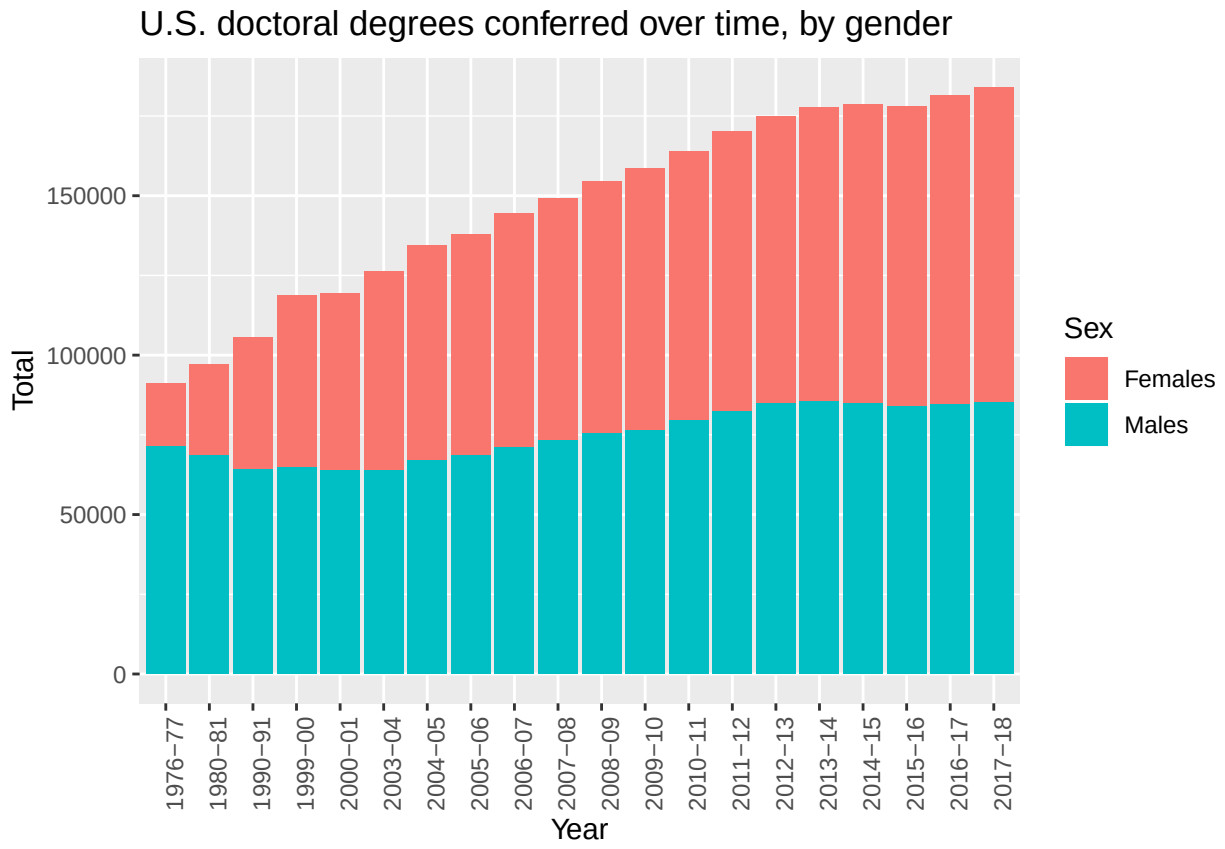
```
## # A tibble: 6 x 3
## # Groups:   Year [6]
##   Sex   Year    Total
##   <chr> <chr>   <dbl>
## 1 Males 1976-77 71709
## 2 Males 1980-81 68853
## 3 Males 1990-91 64242
## 4 Males 1999-00 64930
## 5 Males 2000-01 64171
## 6 Males 2003-04 63981
```

```r
ggplot(doctors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. doctoral degrees conferred over time, by gender") +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```

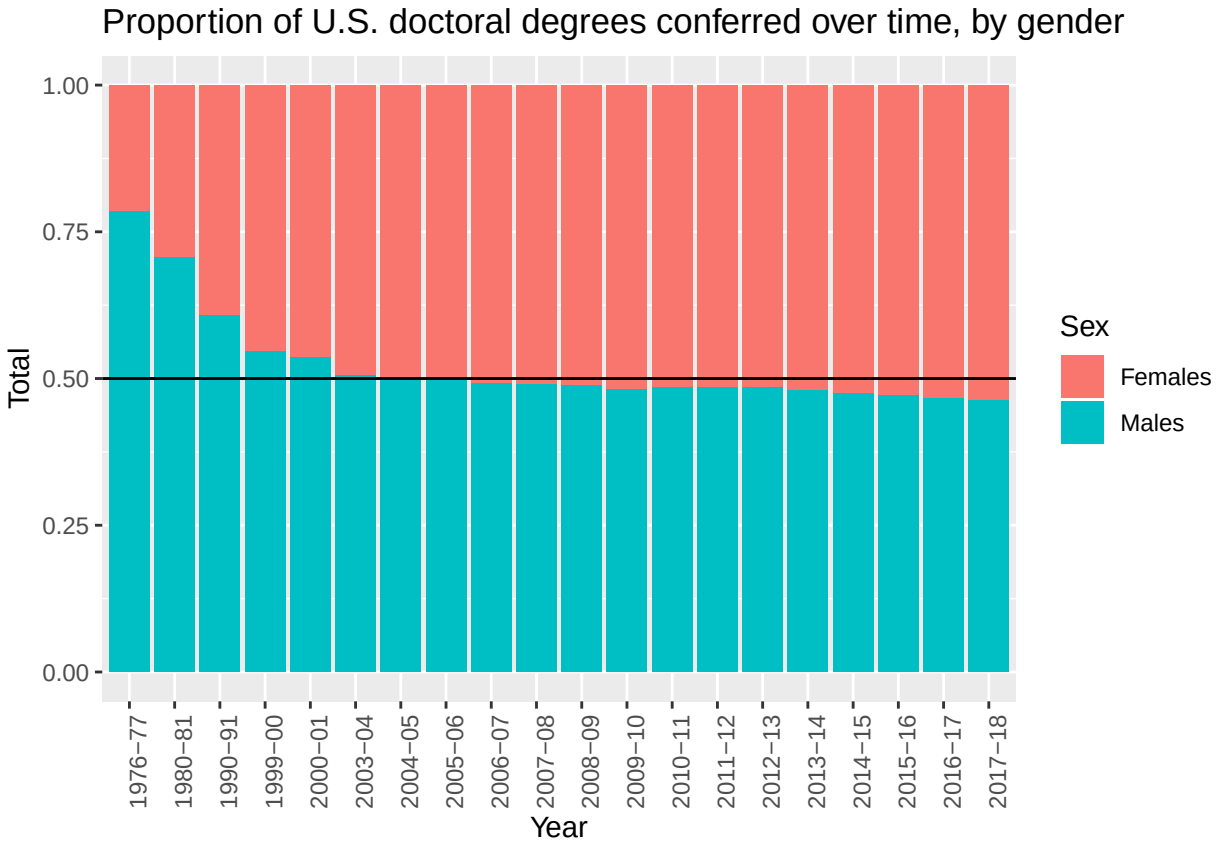


Though females seem to be steadily increasing over time, male degrees fluctate much more.

```
ggplot(doctors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. doctoral degrees conferred over time, by gender") +
  geom_bar(position="stack",stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ggplot(doctors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("Proportion of U.S. doctoral degrees conferred over time, by gender") +
  geom_bar(position="fill", stat="identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_abline(slope=0, intercept = 0.5)
```



Around 2005 was when females and males were earning doctorate degrees equally. Since then, females have been taking an increasing amount of the proportion, but still not startling more than 50%.

Masters

```
masters <- read_csv('data/masters.csv')
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   Year = col_character(),
##   Total = col_double(),
##   White = col_double(),
##   Black = col_double(),
##   Hispanic = col_double(),
##   `Asian/Pacific Islander` = col_double(),
##   `American Indian/Alaska Native` = col_double(),
##   `Two or more races` = col_double(),
##   `Non-resident alien` = col_double()
## )
```

```
head(masters)
```

```
## # A tibble: 6 x 10
```

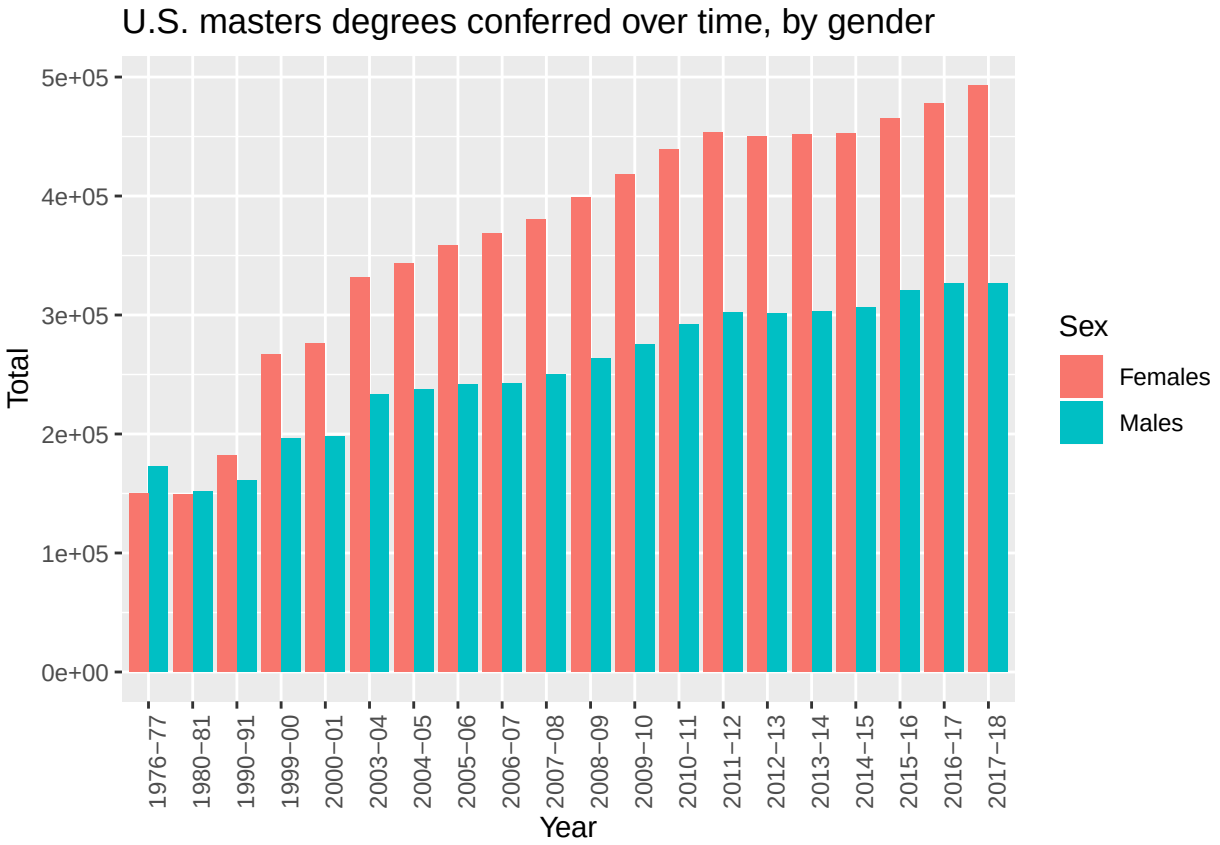
```
##   Sex   Year   Total  White Black Hispanic `Asian/Pacific ~ `American India~
##   <chr> <chr>  <dbl>  <dbl> <dbl>    <dbl>            <dbl>            <dbl>
## 1 Total 1976~ 322463 271402 21252     6136             5127             1018
## 2 Total 1980~ 301081 247475 17436     6534             6348             1044
## 3 Total 1990~ 342863 265927 17023     8981            11869             1189
## 4 Total 1999~ 463185 324990 36606    19379            23523             2263
## 5 Total 2000~ 473502 324211 38853    21661            24544             2496
## 6 Total 2003~ 564272 373448 51402    29806            31202             3206
## # ... with 2 more variables: `Two or more races` <dbl>, `Non-resident
## #   alien` <dbl>
```

```
masters_by_sex <- masters %>%
  select(Sex, Year, Total) %>%
  group_by(Year) %>%
  filter(Sex!="Total")

head(masters_by_sex)
```

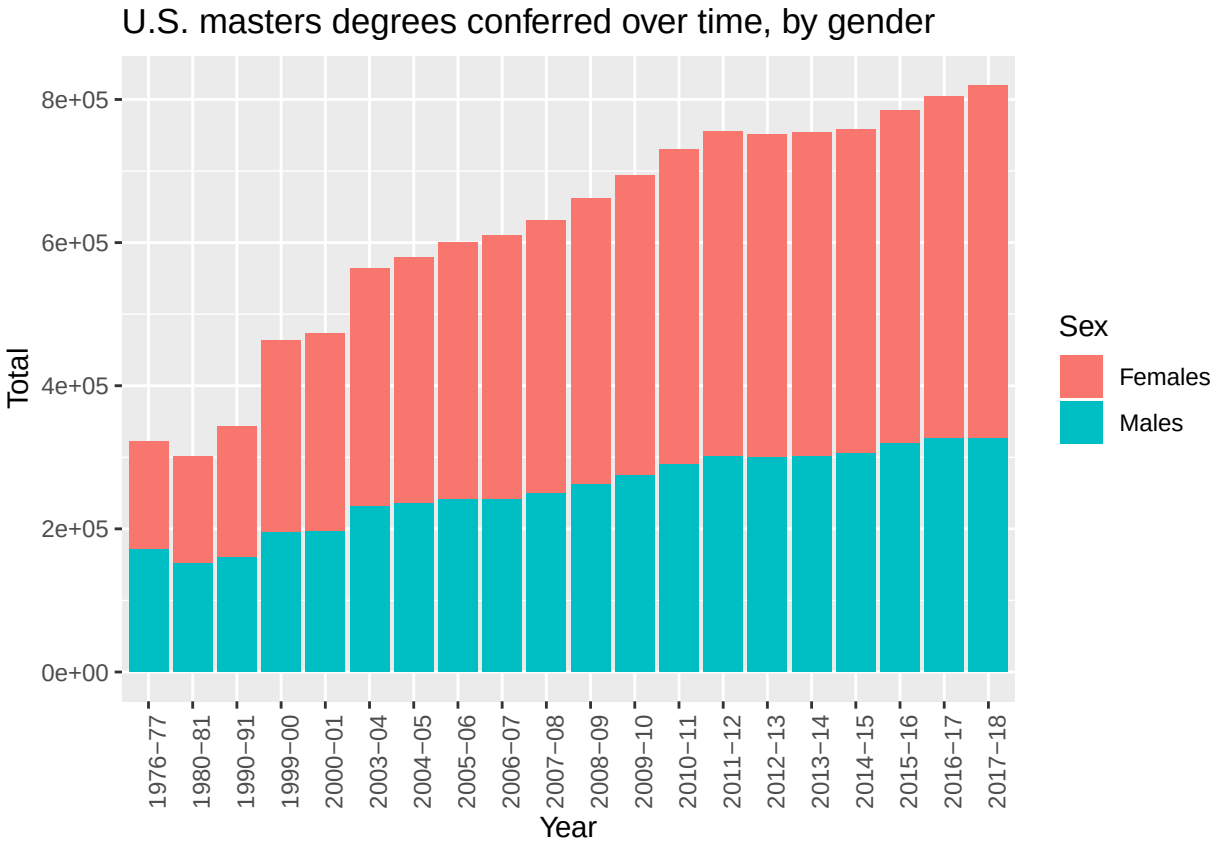
```
## # A tibble: 6 x 3
## # Groups:   Year [6]
##   Sex   Year     Total
##   <chr> <chr>    <dbl>
## 1 Males 1976-77 172703
## 2 Males 1980-81 151602
## 3 Males 1990-91 160842
## 4 Males 1999-00 196129
## 5 Males 2000-01 197770
## 6 Males 2003-04 233056
```

```
ggplot(masters_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. masters degrees conferred over time, by gender") +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```

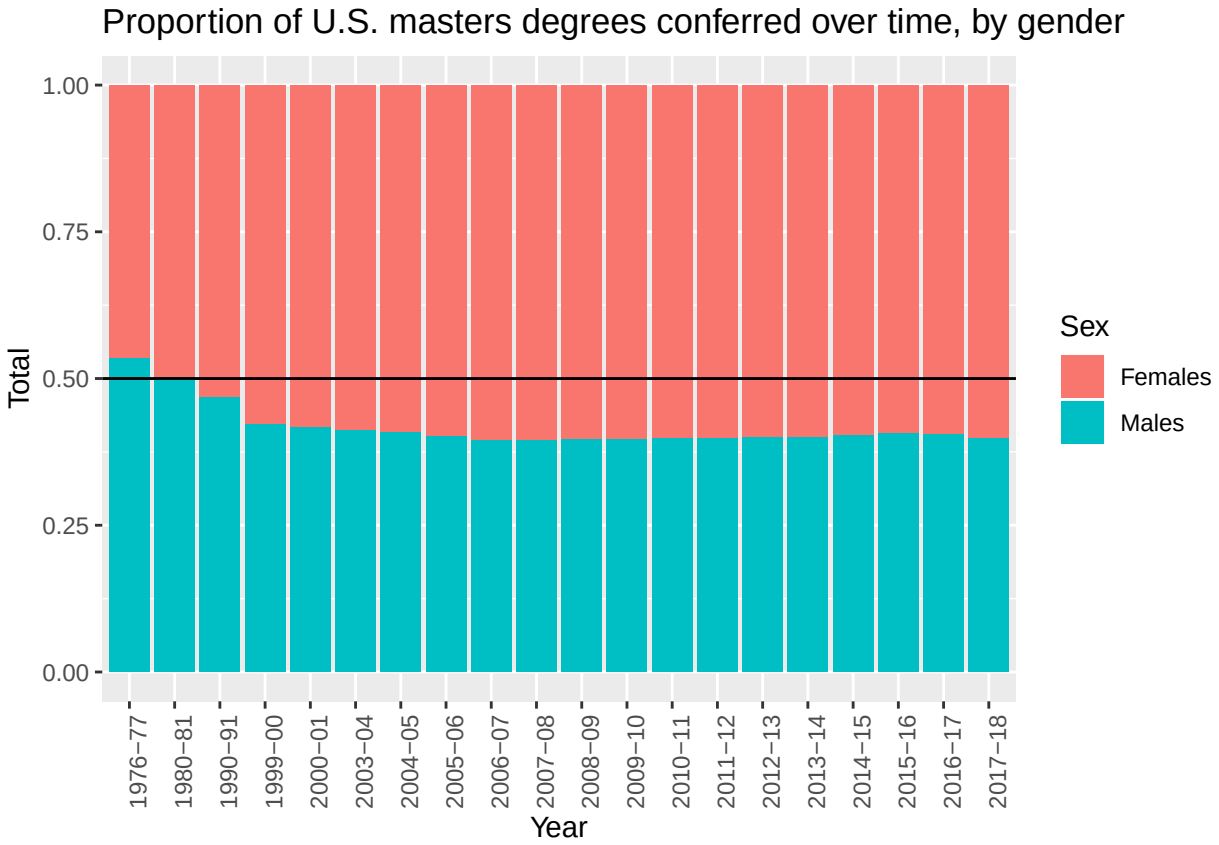


Though both male and female masters degrees seem to be gaining over time, there has been increasingly more degrees conferred to females, especially in the 21st century.

```
ggplot(masters_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. masters degrees conferred over time, by gender") +
  geom_bar(position="stack",stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ggplot(masters_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("Proportion of U.S. masters degrees conferred over time, by gender") +
  geom_bar(position="fill", stat="identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_abline(slope=0, intercept = 0.5)
```



Around 1981, women and men were earning bachelors degrees at similar rates, and into the 21st century females starting gaining on males much more rapidly. Since the start of the 21st century, only incremental changes have occured.

Bachelors

```
bachelors <- read_csv('data/bachelors.csv')
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   Year = col_character(),
##   Total = col_double(),
##   White = col_double(),
##   Black = col_double(),
##   Hispanic = col_double(),
##   `Asian/Pacific Islander` = col_double(),
##   `American Indian/Alaska Native` = col_double(),
##   `Two or more races` = col_double(),
##   `Non-resident alien` = col_double()
## )
```

```r
head(bachelors)
```

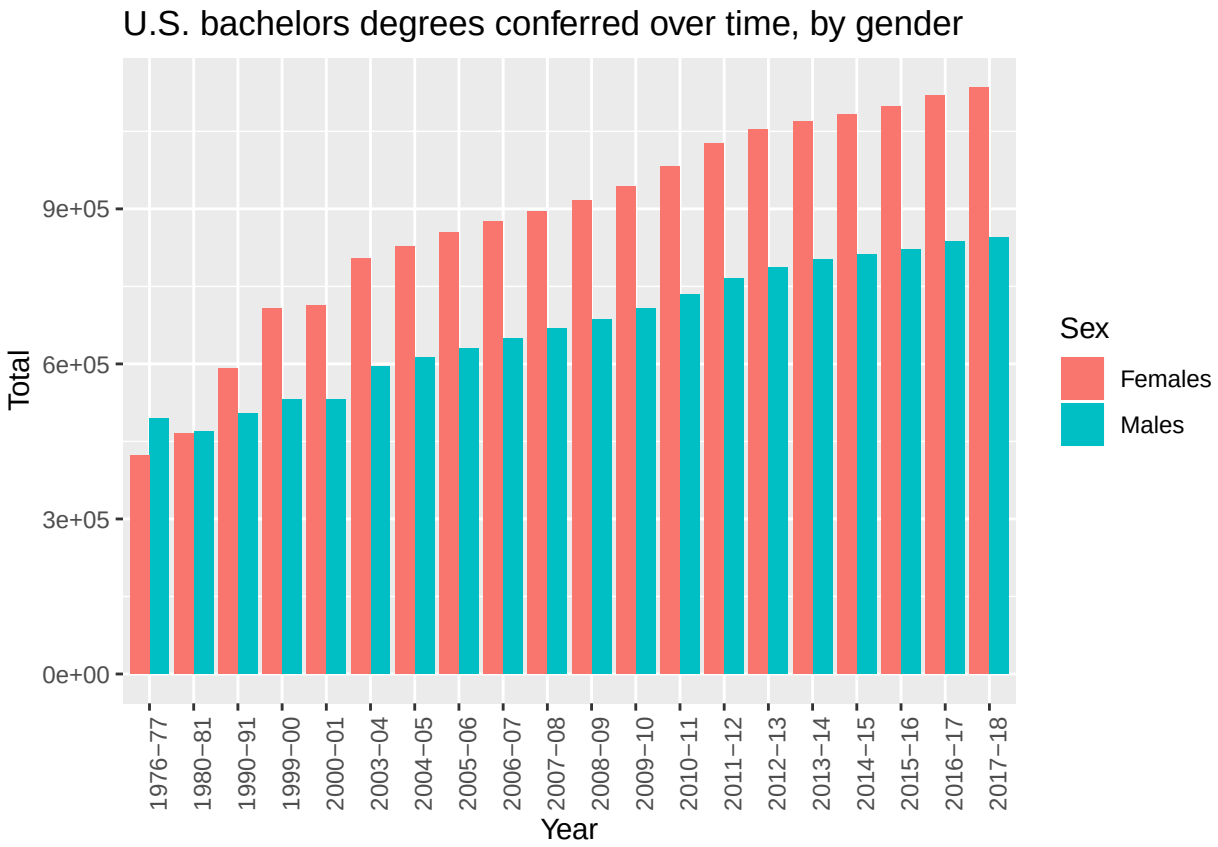
```
## # A tibble: 6 x 10
##   Sex   Year   Total  White  Black Hispanic `Asian/Pacific ~ `American India~
##   <chr> <chr>  <dbl>  <dbl>  <dbl>    <dbl>            <dbl>            <dbl>
## 1 Total 1976~ 9.18e5 8.08e5  58636    18743            13793             3326
## 2 Total 1980~ 9.35e5 8.07e5  60673    21832            18794             3593
## 3 Total 1990~ 1.09e6 9.14e5  66375    37342            42529             4583
## 4 Total 1999~ 1.24e6 9.29e5 108018    75063            77909             8717
## 5 Total 2000~ 1.24e6 9.27e5 111307    77745            78902             9049
## 6 Total 2003~ 1.40e6 1.03e6 131241    94644            92073            10638
## # ... with 2 more variables: `Two or more races` <dbl>, `Non-resident
## #   alien` <dbl>
```

```r
bachelors_by_sex <- bachelors %>%
  select(Sex, Year, Total) %>%
  group_by(Year) %>%
  filter(Sex!="Total")

head(bachelors_by_sex)
```

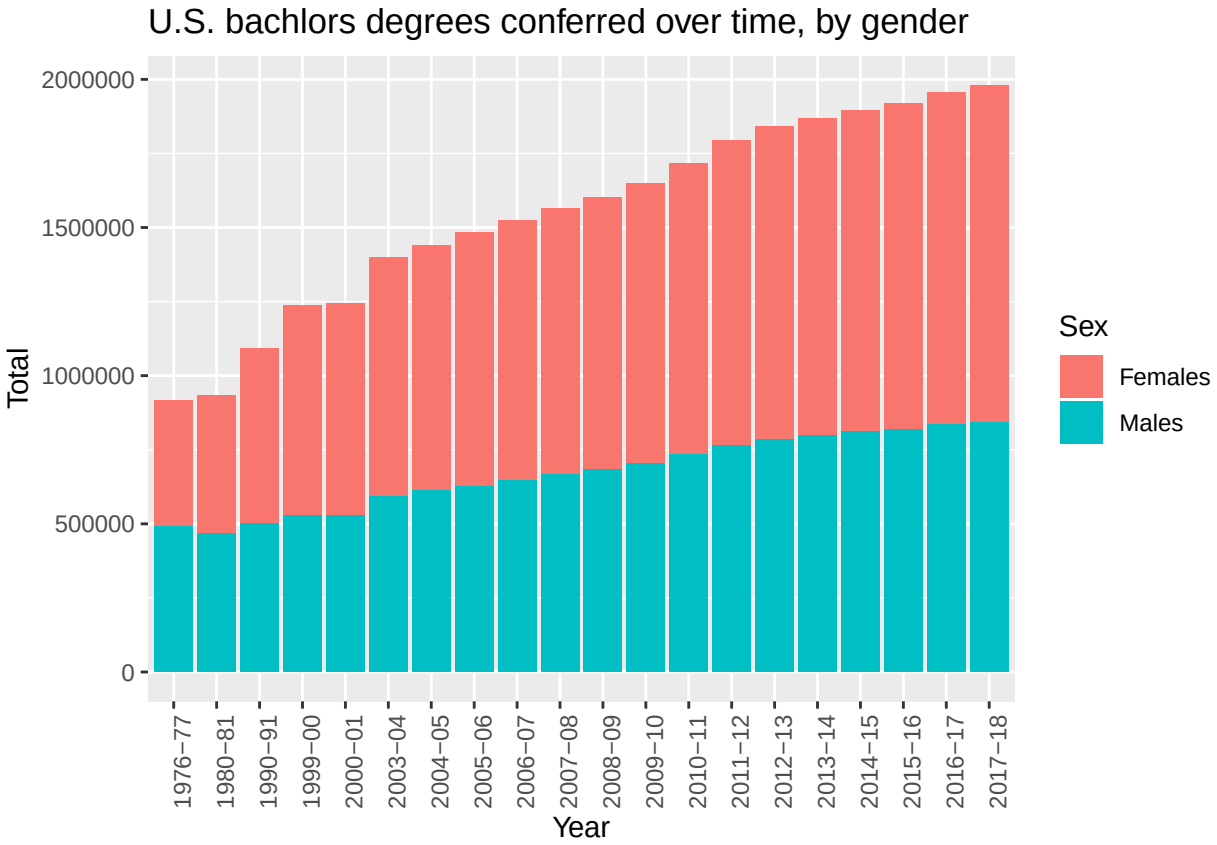
```
## # A tibble: 6 x 3
## # Groups:   Year [6]
##   Sex   Year     Total
##   <chr> <chr>    <dbl>
## 1 Males 1976-77 494424
## 2 Males 1980-81 469625
## 3 Males 1990-91 504045
## 4 Males 1999-00 530367
## 5 Males 2000-01 531840
## 6 Males 2003-04 595425
```

```r
ggplot(bachelors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. bachelors degrees conferred over time, by gender") +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```

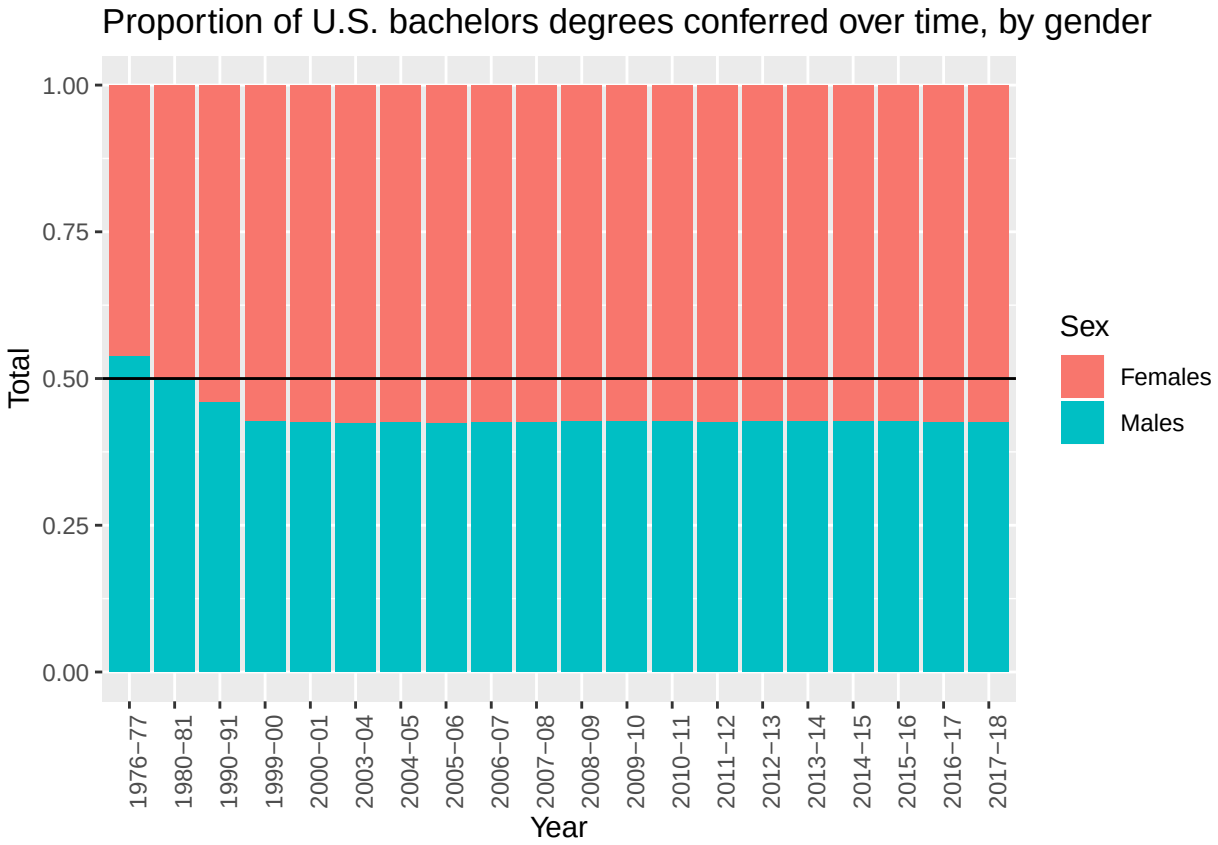


Both males and females are increasing over time, though it seems females may be increasing more rapidly.

```
ggplot(bachelors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. bachlors degrees conferred over time, by gender") +
  geom_bar(position="stack",stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ggplot(bachelors_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("Proportion of U.S. bachelors degrees conferred over time, by gender") +
  geom_bar(position="fill", stat="identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_abline(slope=0, intercept = 0.5)
```



Around 1981, women and men were earning equal amounts of bachelor's degrees. Since about 2000, it seems that the proportion of women vs men earning bachelor's degrees has stayed pretty consistent around a 60-40 split.

Associates

```
associates <- read_csv('data/associates.csv')
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   Year = col_character(),
##   Total = col_double(),
##   White = col_double(),
##   Black = col_double(),
##   Hispanic = col_double(),
##   `Asian/Pacific Islander` = col_double(),
##   `American Indian/Alaska Native` = col_double(),
##   `Two or more races` = col_double(),
##   `Non-resident alien` = col_double()
## )
```

```r
head(associates)
```

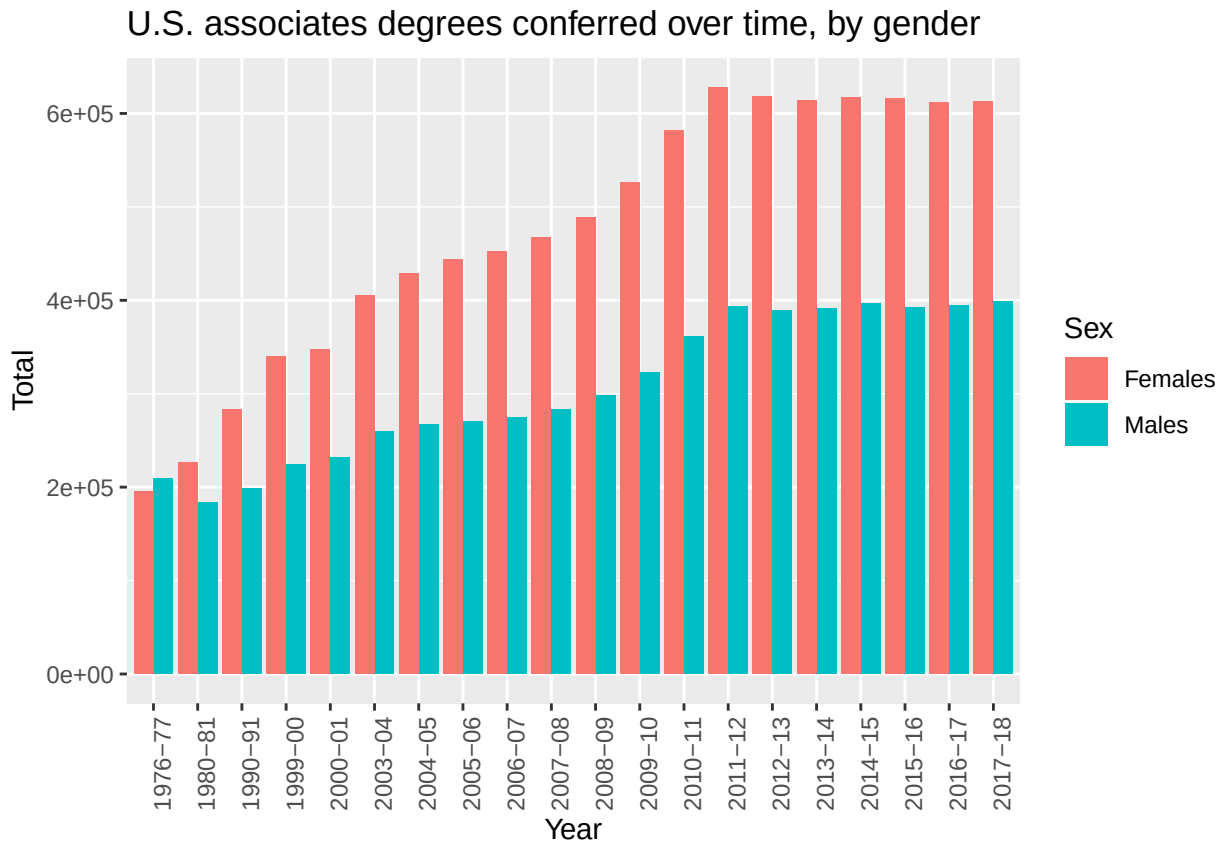
```
## # A tibble: 6 x 10
##   Sex   Year   Total  White Black Hispanic `Asian/Pacific ~ `American India~
##   <chr> <chr>  <dbl>  <dbl> <dbl>    <dbl>            <dbl>            <dbl>
## 1 Total 1976~ 404956 342290 33159    16636             7044             2498
## 2 Total 1980~ 410174 339167 35330    17800             8650             2584
## 3 Total 1990~ 481720 391264 38835    25540            15257             3871
## 4 Total 1999~ 564933 408822 60208    51563            27778             6474
## 5 Total 2000~ 578865 411075 63855    57288            28463             6623
## 6 Total 2003~ 665301 456047 81183    72270            33149             8119
## # ... with 2 more variables: `Two or more races` <dbl>, `Non-resident
## #   alien` <dbl>
```

```r
associates_by_sex <- associates %>%
  select(Sex, Year, Total) %>%
  group_by(Year) %>%
  filter(Sex!="Total")

head(associates_by_sex)
```

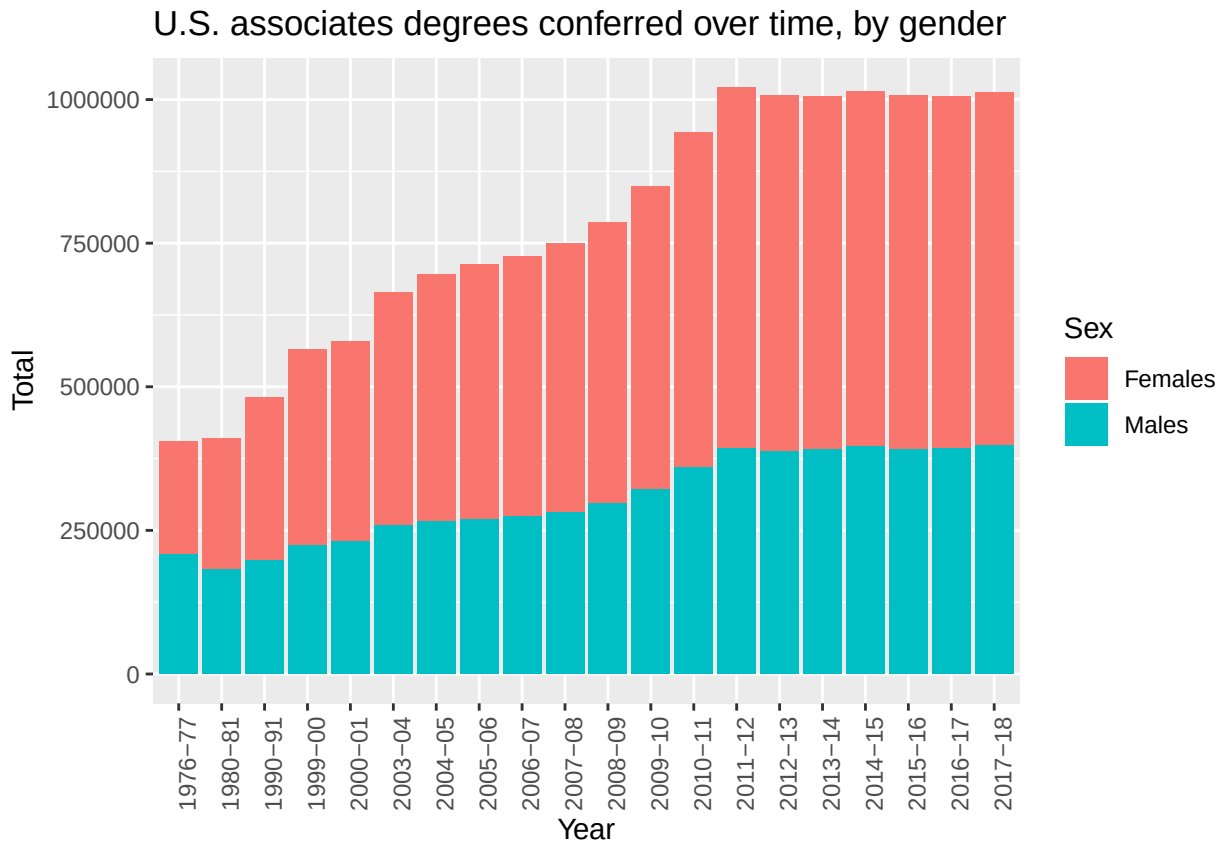
```
## # A tibble: 6 x 3
## # Groups:   Year [6]
##   Sex   Year     Total
##   <chr> <chr>    <dbl>
## 1 Males 1976-77 209672
## 2 Males 1980-81 183819
## 3 Males 1990-91 198634
## 4 Males 1999-00 224721
## 5 Males 2000-01 231645
## 6 Males 2003-04 260033
```

```r
ggplot(associates_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. associates degrees conferred over time, by gender") +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```

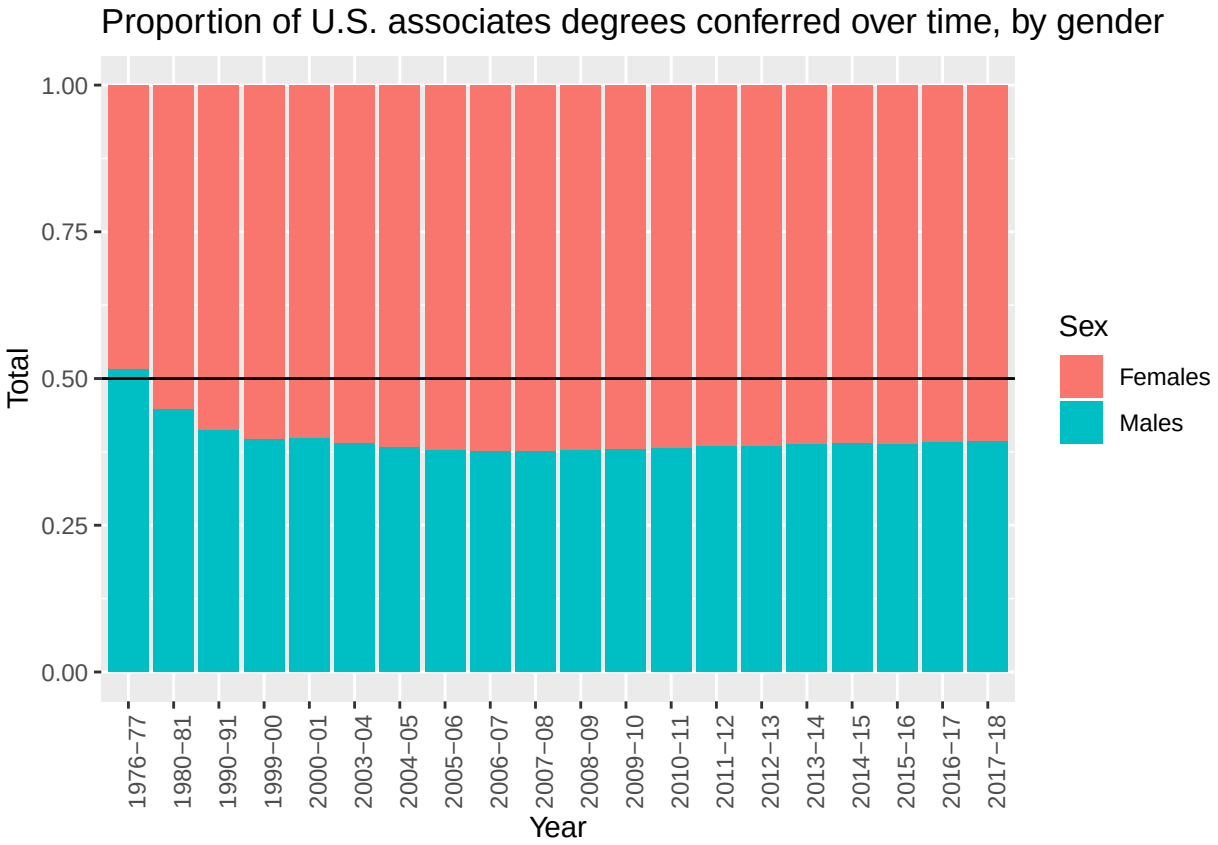


Both males and females are mostly increasing over time (potentially with a peak around 2011), though it seems females are increasing much more rapidly.

```
ggplot(associates_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("U.S. associates degrees conferred over time, by gender") +
  geom_bar(position="stack",stat="identity") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ggplot(associates_by_sex, aes(fill=Sex, y=Total, x=Year)) +
  ggtitle("Proportion of U.S. associates degrees conferred over time, by gender") +
  geom_bar(position="fill", stat="identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_abline(slope=0, intercept = 0.5)
```



Since the 80's women have been earning more associate's degrees than men. The difference here is much more pronouced than in higher levels of education, and is extremely persistent today.