College Enrollments in Illinois

Collin Jaeger

3/30/2018

About the data

Source: IBHE Data Book

Table I-2: Total Fall Enrollments by Gender, Race or National Origin, and Type of Institution, and Level of Instruction

http://legacy.ibhe.org/IBHEDatabook/ChapterI/Table%20I-2.aspx

One file (.csv) was downloaded for each year (1996 - 2016, except 2013).

The data are very untidy . . . but at least they're *consistently* untidy!

Load a few libraries

```
library(tidyverse)
library(readxl)
```

Import and tidy the data

```
setwd('~/Documents/R_Data/IBHE_Enrollments/Data')

year = 1996

file = paste0(year, '.csv')

df = read.csv(file)

df %>% tbl_df
```

```
## # A tibble: 725 x 23
##    Level.of.Instruction  Black.M Black.F Indian.M Indian.F White.M White.F
##    <fct>                 <fct>   <fct>   <fct>    <fct>    <fct>   <fct>
##  1 " Public Universitie~ ""      ""      ""       ""       ""      ""
##  2 Chicago State Univer~ ""      ""      ""       ""       ""      ""
##  3 Undergraduate         1,784   4,564   0        7        112     107
##  4 Graduate              431     1,223   1        5        306     435
##  5 Total                 2,215   5,787   1        12       418     542
##  6 Eastern Illinois Uni~ ""      ""      ""       ""       ""      ""
##  7 Undergraduate         229     293     12       8        3,941   5,322
##  8 Graduate              14      51      1        1        546     916
##  9 Total                 243     344     13       9        4,487   6,238
## 10 Governors State Univ~ ""      ""      ""       ""       ""      ""
## # ... with 715 more rows, and 16 more variables: Asian.M <fct>,
## #   Asian.F <fct>, Hawaiian.M <int>, Hawaiian.F <int>, Hisp..M <fct>,
## #   Hisp..F <fct>, Two.or.more.races.M <int>, Two.or.more.races.F <int>,
## #   Alien.M <fct>, Alien.F <fct>, Other.M <fct>, Other.F <fct>,
## #   Total.M <fct>, Total.F <fct>, Grand.Total <fct>, X <lgl>
```

```
# Remove the commas from the data
no.commas = function(x){
  dx = as.character(df[ , x])
  dx2 = as.numeric(gsub(',', '', dx))
  dx2
}

df2 = data.frame(df[, 1], sapply(2:23, no.commas))

names(df2) = names(df)

df2$Level.of.Instruction = as.character(df2$Level.of.Instruction)

df2 %>% tbl_df
```

```
## # A tibble: 725 x 23
##    Level.of.Instruction  Black.M Black.F Indian.M Indian.F White.M White.F
##    <chr>                   <dbl>   <dbl>    <dbl>    <dbl>   <dbl>   <dbl>
##  1 " Public Universitie~    NA       NA     NA       NA         NA      NA
##  2 Chicago State Univer~    NA       NA     NA       NA         NA      NA
##  3 Undergraduate          1784    4564       0        7.00     112     107
##  4 Graduate                431    1223       1.00     5.00     306     435
##  5 Total                  2215    5787       1.00    12.0      418     542
##  6 Eastern Illinois Uni~    NA       NA     NA       NA         NA      NA
##  7 Undergraduate          229     293     12.0        8.00    3941    5322
##  8 Graduate                 14.0    51.0     1.00     1.00     546     916
##  9 Total                  243      344    13.0        9.00    4487    6238
## 10 Governors State Univ~    NA       NA     NA       NA         NA      NA
## # ... with 715 more rows, and 16 more variables: Asian.M <dbl>,
## #   Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>, Hisp..M <dbl>,
## #   Hisp..F <dbl>, Two.or.more.races.M <dbl>, Two.or.more.races.F <dbl>,
## #   Alien.M <dbl>, Alien.F <dbl>, Other.M <dbl>, Other.F <dbl>,
## #   Total.M <dbl>, Total.F <dbl>, Grand.Total <dbl>, X <dbl>
```

```
categories = df2 %>%
  tbl_df %>%
  filter(grepl('^ [[:blank:]]*', Level.of.Instruction)) %>%
  select(Level.of.Instruction)

categories$rownum = grep('^ [[:blank:]]*', df2$Level.of.Instruction)

l1 = split(df2, cumsum(1:nrow(df2) %in% categories$rownum))

names(l1) = categories$Level.of.Instruction

df3 = bind_rows(l1, .id='Category')

df3 %>% tbl_df
```

```
## # A tibble: 725 x 24
##    Category    Level.of.Instruc~ Black.M Black.F Indian.M Indian.F White.M
##    <chr>       <chr>               <dbl>   <dbl>    <dbl>    <dbl>   <dbl>
##  1 " Public U~ " Public Univers~    NA      NA      NA       NA         NA
##  2 " Public U~ Chicago State Un~    NA      NA      NA       NA          NA
##  3 " Public U~ Undergraduate      1784    4564       0        7.00      112
##  4 " Public U~ Graduate            431    1223       1.00     5.00      306
##  5 " Public U~ Total              2215    5787       1.00    12.0       418
##  6 " Public U~ Eastern Illinois~    NA      NA      NA       NA          NA
##  7 " Public U~ Undergraduate       229     293      12.0      8.00    3941
##  8 " Public U~ Graduate             14.0    51.0     1.00     1.00     546
##  9 " Public U~ Total               243     344      13.0      9.00    4487
## 10 " Public U~ Governors State ~    NA      NA      NA       NA          NA
## # ... with 715 more rows, and 17 more variables: White.F <dbl>,
## #   Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>,
## #   Hisp..M <dbl>, Hisp..F <dbl>, Two.or.more.races.M <dbl>,
## #   Two.or.more.races.F <dbl>, Alien.M <dbl>, Alien.F <dbl>,
## #   Other.M <dbl>, Other.F <dbl>, Total.M <dbl>, Total.F <dbl>,
## #   Grand.Total <dbl>, X <dbl>
```

```
rownum2 = grep('Undergraduate', df3$Level.of.Instruction) - 1

l2 = split(df3, cumsum(1:nrow(df3) %in% rownum2))

names(l2) = df3$Level.of.Instruction[c(1, rownum2)]

df4 = bind_rows(l2, .id='School')
df4$Year = year

df4 %>% tbl_df
```

```
## # A tibble: 725 x 26
##    School    Category  Level.of.Instruc~ Black.M Black.F Indian.M Indian.F
##    <chr>     <chr>     <chr>               <dbl>   <dbl>    <dbl>    <dbl>
##  1 " Public~ " Public~ " Public Univers~    NA      NA      NA       NA
##  2 Chicago ~ " Public~ Chicago State Un~    NA      NA      NA       NA
##  3 Chicago ~ " Public~ Undergraduate      1784    4564        0       7.00
##  4 Chicago ~ " Public~ Graduate            431    1223        1.00    5.00
##  5 Chicago ~ " Public~ Total              2215    5787        1.00   12.0
##  6 Eastern ~ " Public~ Eastern Illinois~    NA      NA      NA       NA
##  7 Eastern ~ " Public~ Undergraduate       229     293      12.0      8.00
##  8 Eastern ~ " Public~ Graduate             14.0    51.0      1.00    1.00
##  9 Eastern ~ " Public~ Total               243     344      13.0      9.00
## 10 Governor~ " Public~ Governors State ~    NA      NA      NA       NA
## # ... with 715 more rows, and 19 more variables: White.M <dbl>,
## #   White.F <dbl>, Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>,
## #   Hawaiian.F <dbl>, Hisp..M <dbl>, Hisp..F <dbl>,
## #   Two.or.more.races.M <dbl>, Two.or.more.races.F <dbl>, Alien.M <dbl>,
## #   Alien.F <dbl>, Other.M <dbl>, Other.F <dbl>, Total.M <dbl>,
## #   Total.F <dbl>, Grand.Total <dbl>, X <dbl>, Year <dbl>
```

```
df5 = df4 %>%
  tbl_df %>%
  filter( ! is.na(Grand.Total)) %>%
  filter( ! grepl('Total', School, ignore.case = T)) %>%
  rename(Level = Level.of.Instruction) %>%
  filter(Level %in% c('Undergraduate', 'Graduate')) %>%
  mutate(Category = trimws(Category)) %>%
  select(Year, Category, School, everything(),
         -X, -Total.M, -Total.F, -Grand.Total)

df5 %>% tbl_df
```

```
## # A tibble: 348 x 22
##     Year Category  School  Level Black.M Black.F Indian.M Indian.F White.M
##    <dbl> <chr>     <chr>   <chr>   <dbl>   <dbl>    <dbl>    <dbl>   <dbl>
##  1  1996 Public U~ Chicag~ Unde~  1784    4564       0        7.00     112
##  2  1996 Public U~ Chicag~ Grad~   431     1223      1.00     5.00     306
##  3  1996 Public U~ Easter~ Unde~   229     293      12.0      8.00    3941
##  4  1996 Public U~ Easter~ Grad~    14.0    51.0     1.00     1.00     546
##  5  1996 Public U~ Govern~ Unde~   187     561       3.00     5.00     708
##  6  1996 Public U~ Govern~ Grad~   173     597       0        2.00     587
##  7  1996 Public U~ Illino~ Unde~   572     868      29.0     24.0     6338
##  8  1996 Public U~ Illino~ Grad~     67.0  103       2.00     6.00     951
##  9  1996 Public U~ Northe~ Unde~   324     620       7.00    11.0     1525
## 10  1996 Public U~ Northe~ Grad~     69.0  166       1.00     3.00     702
## # ... with 338 more rows, and 13 more variables: White.F <dbl>,
## #   Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>,
## #   Hisp..M <dbl>, Hisp..F <dbl>, Two.or.more.races.M <dbl>,
## #   Two.or.more.races.F <dbl>, Alien.M <dbl>, Alien.F <dbl>,
## #   Other.M <dbl>, Other.F <dbl>
```

So, that code works to tidy *one* year.

Turn it into a *function*, and use it on a *different* year.

```
clean.data(1997)
```

```
## # A tibble: 342 x 22
##     Year Category  School  Level Black.M Black.F Indian.M Indian.F White.M
##    <dbl> <chr>     <chr>   <chr>   <dbl>   <dbl>    <dbl>    <dbl>   <dbl>
##  1  1997 Public U~ Chicag~ Unde~  1677    4340       0        7.00     107
##  2  1997 Public U~ Chicag~ Grad~   367     1099      1.00     6.00     252
##  3  1997 Public U~ Easter~ Unde~   241     309       9.00     9.00    3848
##  4  1997 Public U~ Easter~ Grad~    17.0    50.0     1.00     0        470
##  5  1997 Public U~ Govern~ Unde~   189     595       3.00     5.00     690
##  6  1997 Public U~ Govern~ Grad~   184     641       0        1.00     566
##  7  1997 Public U~ Illino~ Unde~   567     895      31.0     30.0     6497
##  8  1997 Public U~ Illino~ Grad~     81.0  103       2.00     8.00     894
##  9  1997 Public U~ Northe~ Unde~   353     659       4.00     9.00    1494
## 10  1997 Public U~ Northe~ Grad~     72.0  152       3.00     1.00     655
## # ... with 332 more rows, and 13 more variables: White.F <dbl>,
## #   Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>,
## #   Hisp..M <dbl>, Hisp..F <dbl>, Two.or.more.races.M <dbl>,
## #   Two.or.more.races.F <dbl>, Alien.M <dbl>, Alien.F <dbl>,
## #   Other.M <dbl>, Other.F <dbl>
```

```
clean.data(2002)
```

```
## # A tibble: 368 x 22
##     Year Category  School  Level Black.M Black.F Indian.M Indian.F White.M
##    <dbl> <chr>     <chr>   <chr>   <dbl>   <dbl>    <dbl>    <dbl>   <dbl>
##  1  2002 Public U~ Chicag~ Unde~  1127    3275       2.00     4.00    77.0
##  2  2002 Public U~ Chicag~ Grad~   341    1018       0        2.00   246
##  3  2002 Public U~ Easter~ Unde~   298     391       9.00    10.0   3476
##  4  2002 Public U~ Easter~ Grad~    21.0    49.0     2.00     1.00   486
##  5  2002 Public U~ Govern~ Unde~   198     649       3.00     3.00   531
##  6  2002 Public U~ Govern~ Grad~   206     697       1.00     3.00   503
##  7  2002 Public U~ Illino~ Unde~   401     682      22.0     25.0   6887
##  8  2002 Public U~ Illino~ Grad~    48.0   117       1.00     9.00   753
##  9  2002 Public U~ Northe~ Unde~   389     769       7.00     7.00  1427
## 10  2002 Public U~ Northe~ Grad~    73.0   169       5.00     8.00   583
## # ... with 358 more rows, and 13 more variables: White.F <dbl>,
## #   Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>,
## #   Hisp..M <dbl>, Hisp..F <dbl>, Two.or.more.races.M <dbl>,
## #   Two.or.more.races.F <dbl>, Alien.M <dbl>, Alien.F <dbl>,
## #   Other.M <dbl>, Other.F <dbl>
```

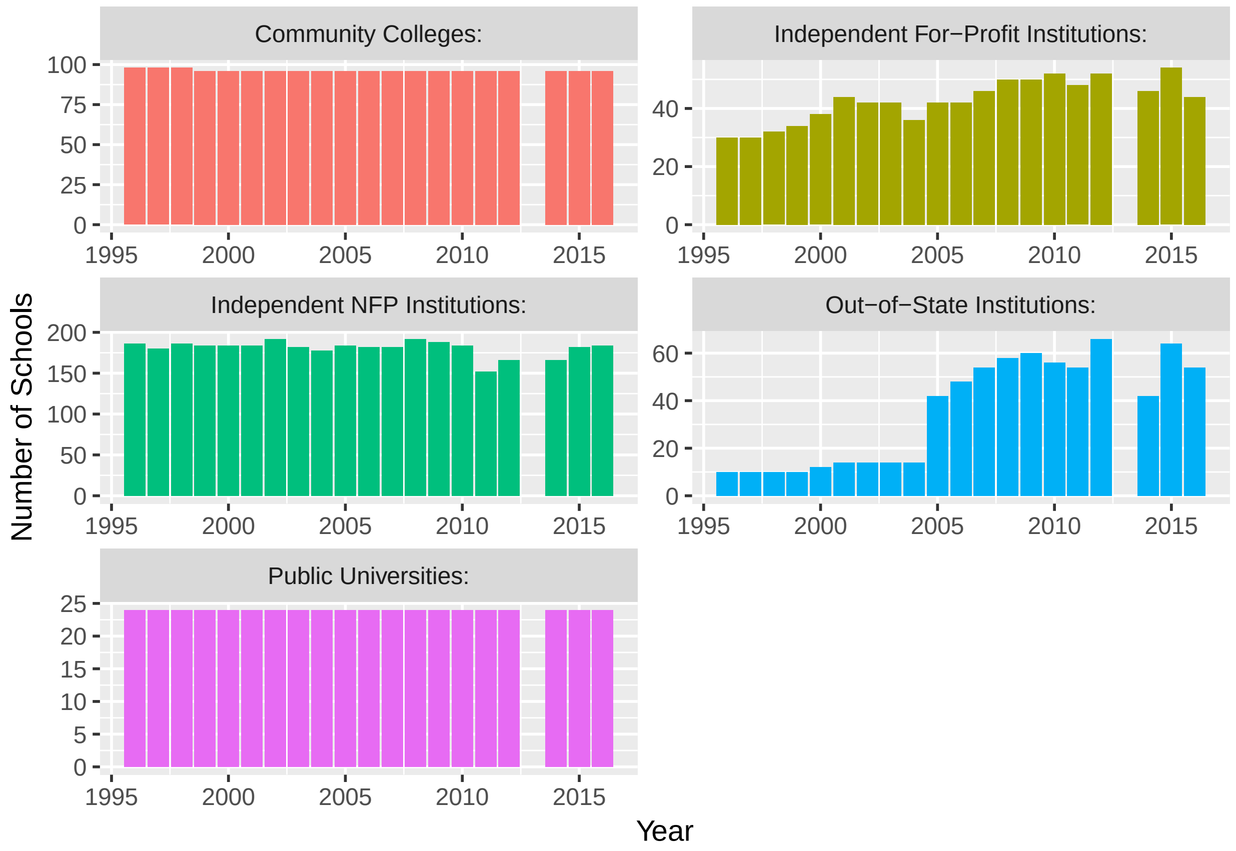
Now, tidy and consolidate all of the data.

```r
yrs = c(1996:2012, 2014:2016)

out.list = lapply(yrs, clean.data)

out.df = bind_rows(out.list)

out.df
```

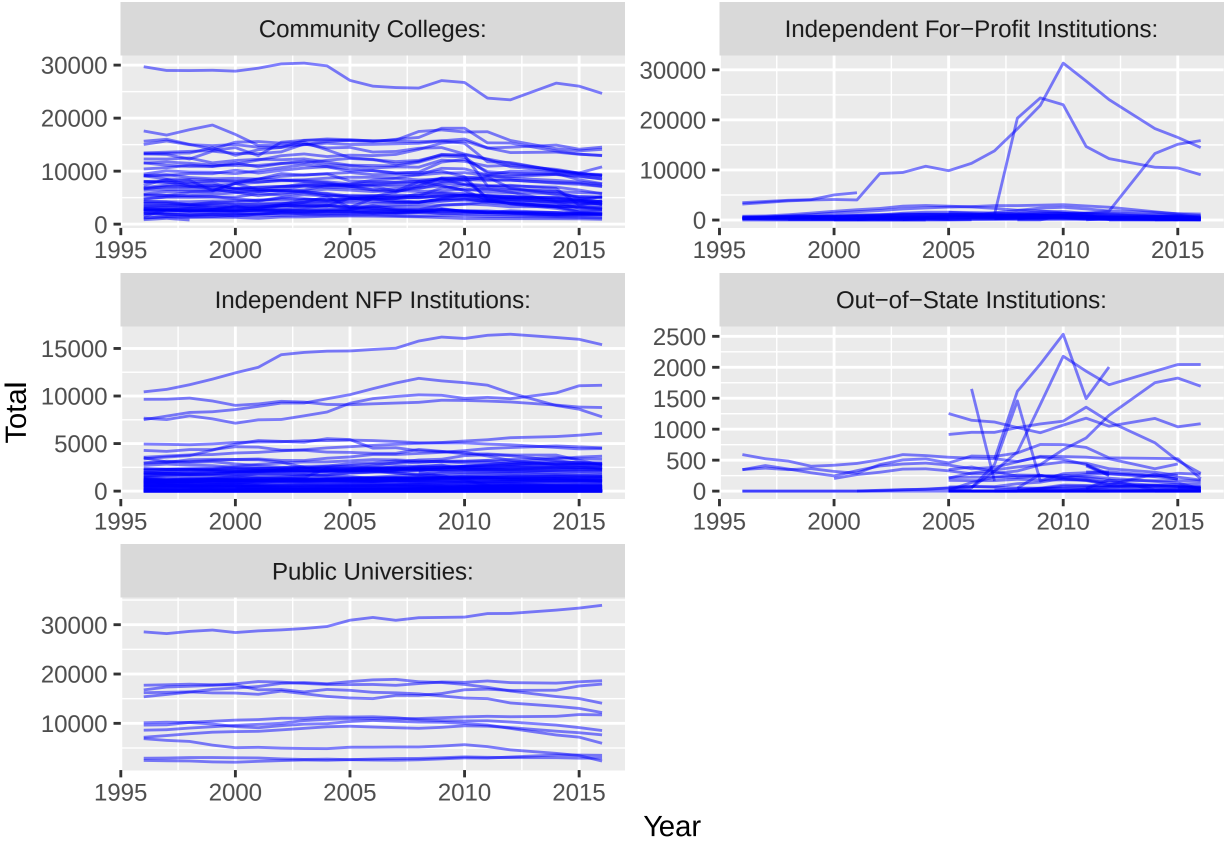
```
## # A tibble: 7,584 x 22
##     Year Category School  Level Black.M Black.F Indian.M Indian.F White.M
##    <int> <chr>    <chr>   <chr>   <dbl>   <dbl>    <dbl>    <dbl>   <dbl>
##  1  1996 Public U~ Chicag~ Unde~  1784    4564       0        7.00     112
##  2  1996 Public U~ Chicag~ Grad~   431    1223       1.00     5.00     306
##  3  1996 Public U~ Easter~ Unde~   229     293      12.0      8.00    3941
##  4  1996 Public U~ Easter~ Grad~    14.0    51.0     1.00     1.00     546
##  5  1996 Public U~ Govern~ Unde~   187     561       3.00     5.00     708
##  6  1996 Public U~ Govern~ Grad~   173     597       0        2.00     587
##  7  1996 Public U~ Illino~ Unde~   572     868      29.0     24.0     6338
##  8  1996 Public U~ Illino~ Grad~    67.0   103       2.00     6.00     951
##  9  1996 Public U~ Northe~ Unde~   324     620       7.00    11.0     1525
## 10  1996 Public U~ Northe~ Grad~    69.0   166       1.00     3.00     702
## # ... with 7,574 more rows, and 13 more variables: White.F <dbl>,
## #   Asian.M <dbl>, Asian.F <dbl>, Hawaiian.M <dbl>, Hawaiian.F <dbl>,
## #   Hisp..M <dbl>, Hisp..F <dbl>, Two.or.more.races.M <dbl>,
## #   Two.or.more.races.F <dbl>, Alien.M <dbl>, Alien.F <dbl>,
## #   Other.M <dbl>, Other.F <dbl>
```

Make a few plots

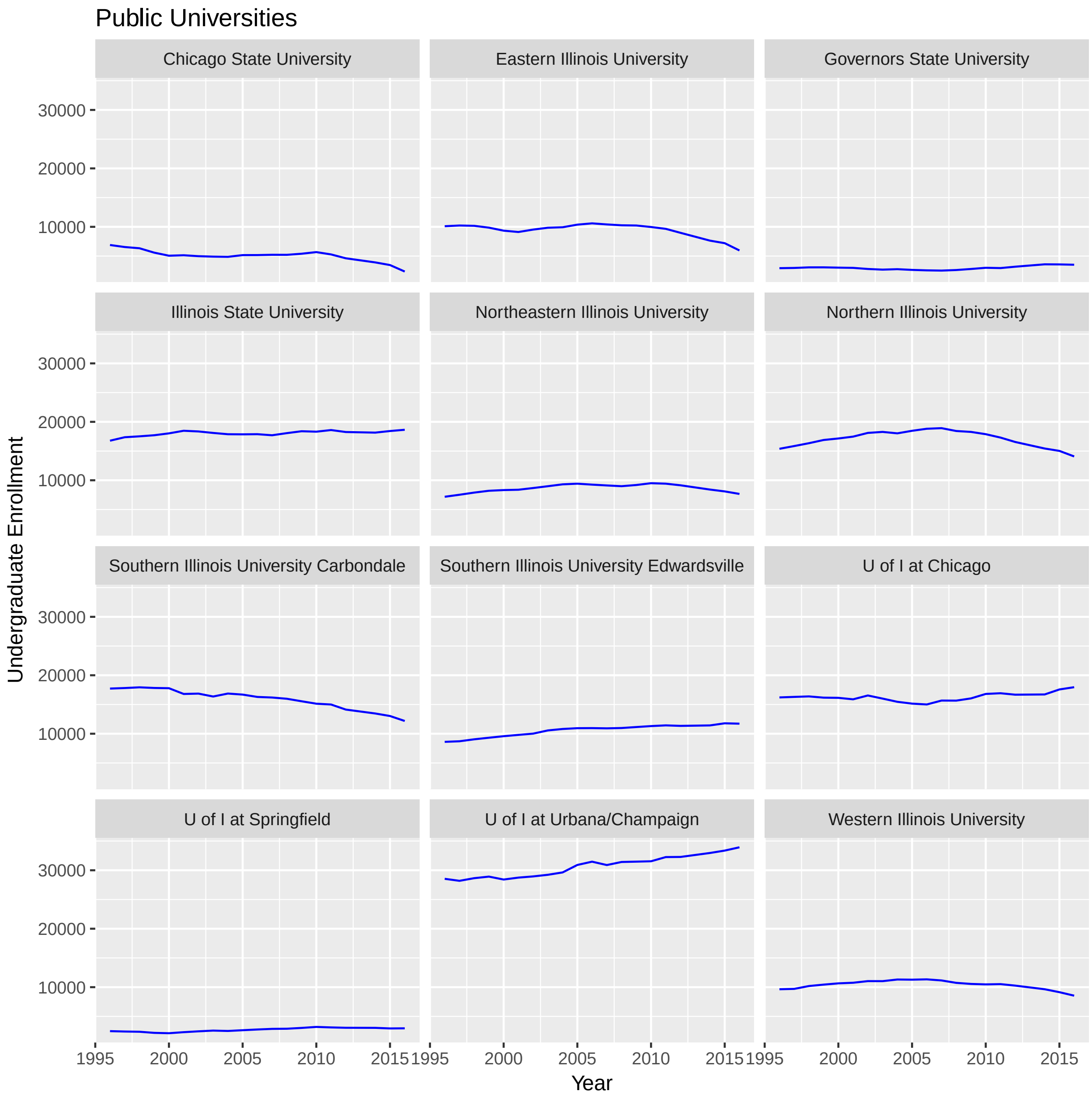
```r
out.df %>%
  ggplot() +
  geom_bar(aes(Year, fill=Category), show.legend=F) +
  labs(y='Number of Schools') +
  facet_wrap(~Category, scales='free', ncol=2)
```



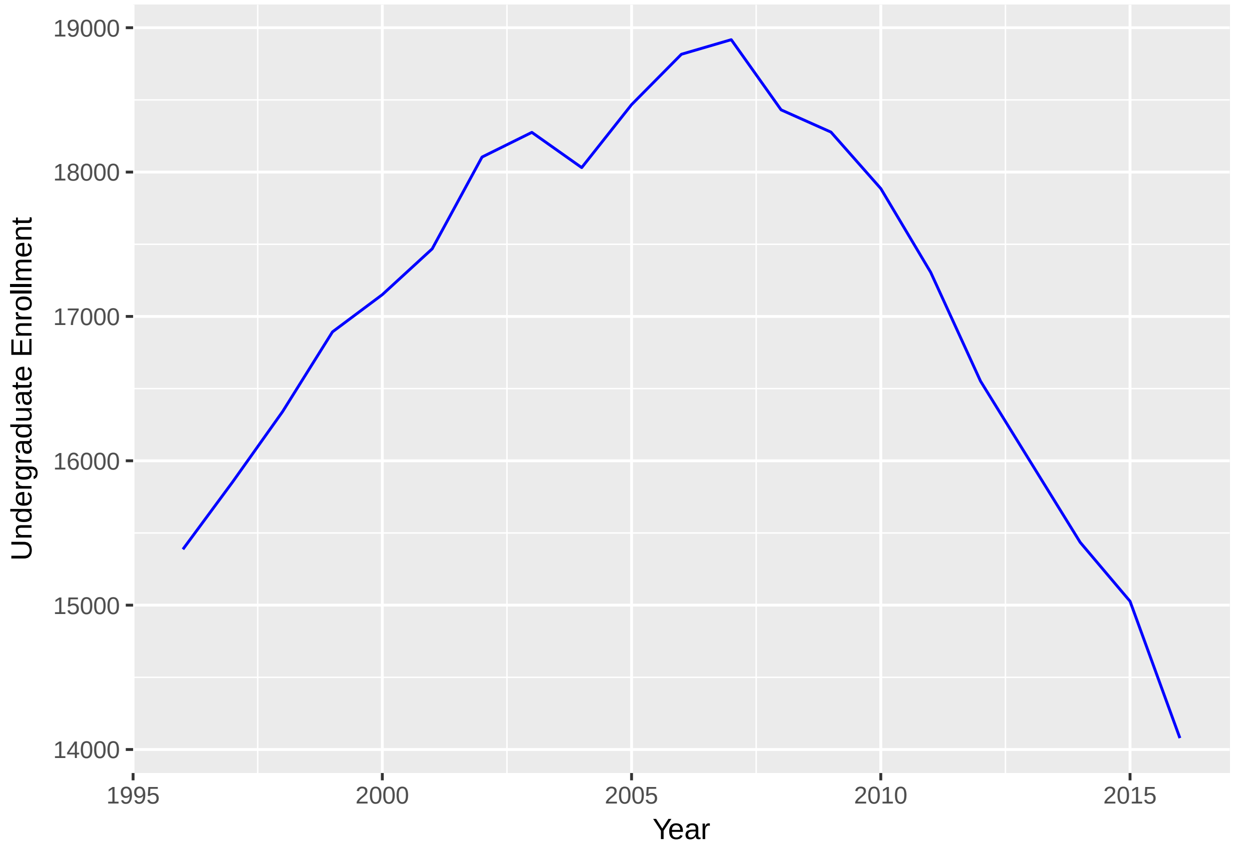
```
out.df %>%
  mutate(Total = out.df %>%
           select(contains(".")) %>% rowSums()) %>%
  filter(Level == 'Undergraduate') %>%
  ggplot() +
  geom_line(aes(Year, Total, group=School),
             show.legend=F, alpha=0.5, col='blue') +
  facet_wrap(~Category, scales='free', ncol=2)
```



```
out.df %>%
  mutate(Total = out.df %>%
           select(contains(".")) %>% rowSums()) %>%
  filter(Category == 'Public Universities:',
         Level == 'Undergraduate') %>%
  ggplot() +
  geom_line(aes(Year, Total, group=School),
            show.legend=F, col='blue') +
  labs(y='Undergraduate Enrollment',
       title='Public Universities') +
  facet_wrap(~School, ncol=3)
```



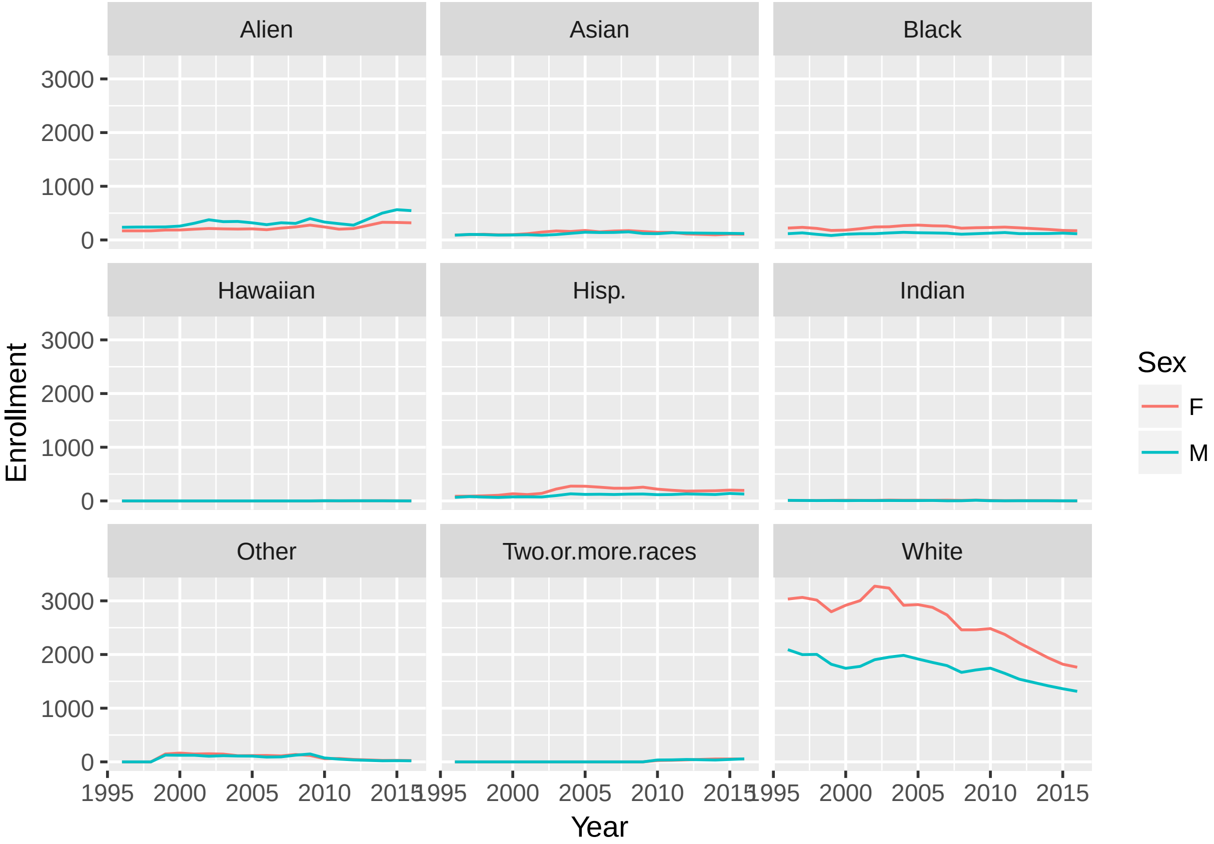
```r
out.df %>%
  mutate(Total = out.df %>%
           select(contains(".")) %>% rowSums()) %>%
  filter(School == 'Northern Illinois University',
         Level == 'Undergraduate') %>%
  ggplot() +
  geom_line(aes(Year, Total), col='blue') +
  labs(y='Undergraduate Enrollment')
```



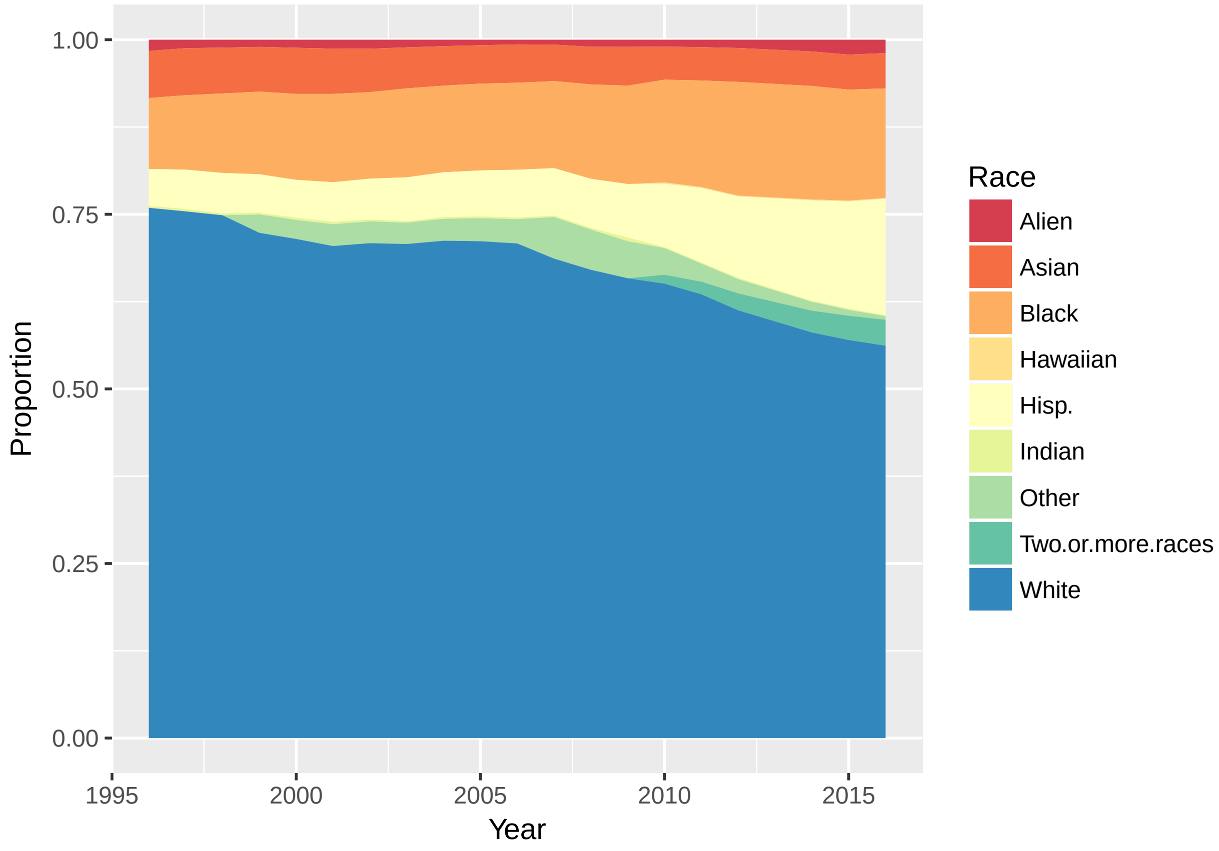
```
out.df.long = out.df %>%
  gather(Group, Enrollment,
         -Year, -Category, -School, -Level) %>%
  mutate(Group = gsub('.M', '_M', Group),
         Group = gsub('.F', '_F', Group)) %>%
  separate(Group, into=c('Race', 'Sex'),
           sep='_', extra='merge')

out.df.long
```

```
## # A tibble: 136,512 x 7
##     Year Category             School        Level   Race  Sex   Enrollment
##   * <int> <chr>                <chr>         <chr>   <chr> <chr>      <dbl>
##   1  1996 Public Universities: Chicago Stat~ Underg~ Black M          1784
##   2  1996 Public Universities: Chicago Stat~ Gradua~ Black M           431
##   3  1996 Public Universities: Eastern Illi~ Underg~ Black M           229
##   4  1996 Public Universities: Eastern Illi~ Gradua~ Black M            14.0
##   5  1996 Public Universities: Governors St~ Underg~ Black M           187
##   6  1996 Public Universities: Governors St~ Gradua~ Black M           173
##   7  1996 Public Universities: Illinois Sta~ Underg~ Black M           572
##   8  1996 Public Universities: Illinois Sta~ Gradua~ Black M            67.0
##   9  1996 Public Universities: Northeastern~ Underg~ Black M           324
## 10  1996 Public Universities: Northeastern~ Gradua~ Black M            69.0
## # ... with 136,502 more rows
```

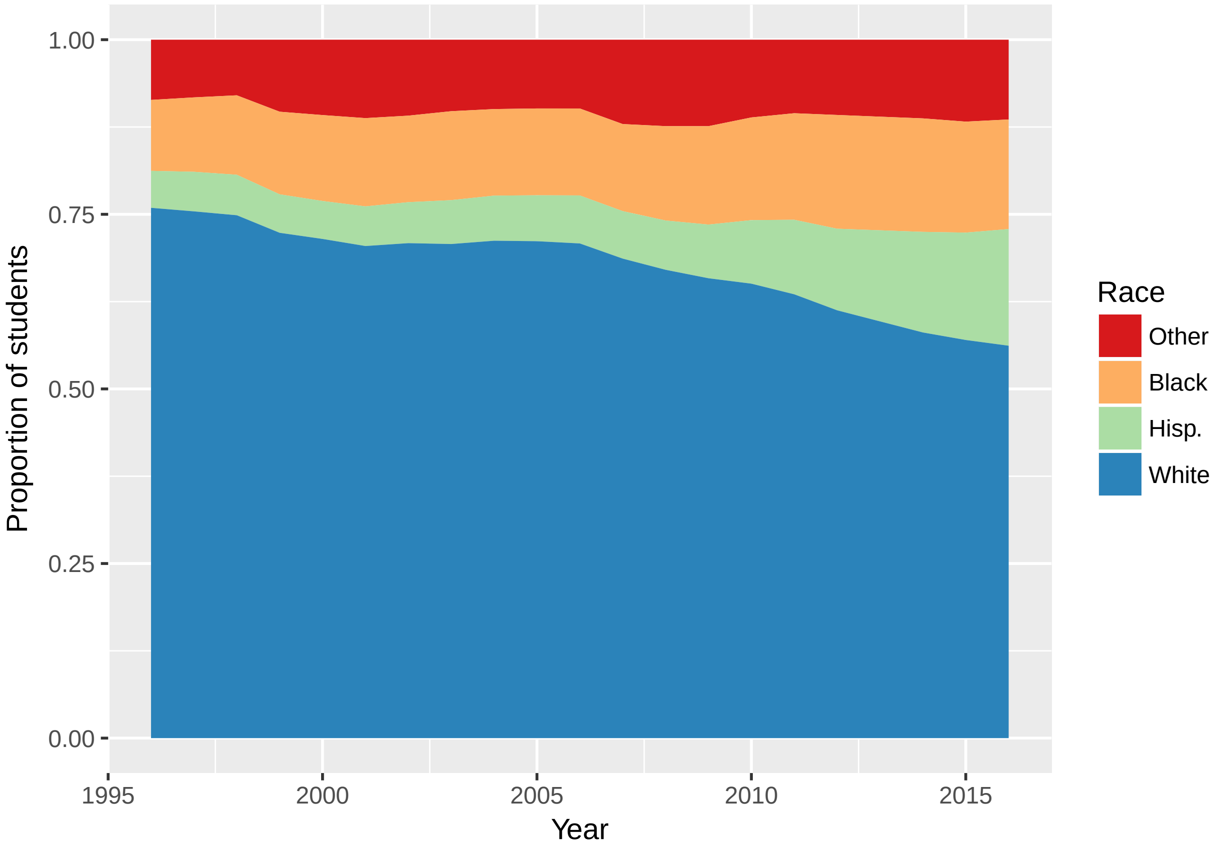
```
out.df.long %>%
  filter(grepl('Northern Illinois', School),
         Level == 'Graduate') %>%
  ggplot() +
  geom_line(aes(Year, Enrollment, color = Sex)) +
  facet_wrap(~Race, ncol=3)
```



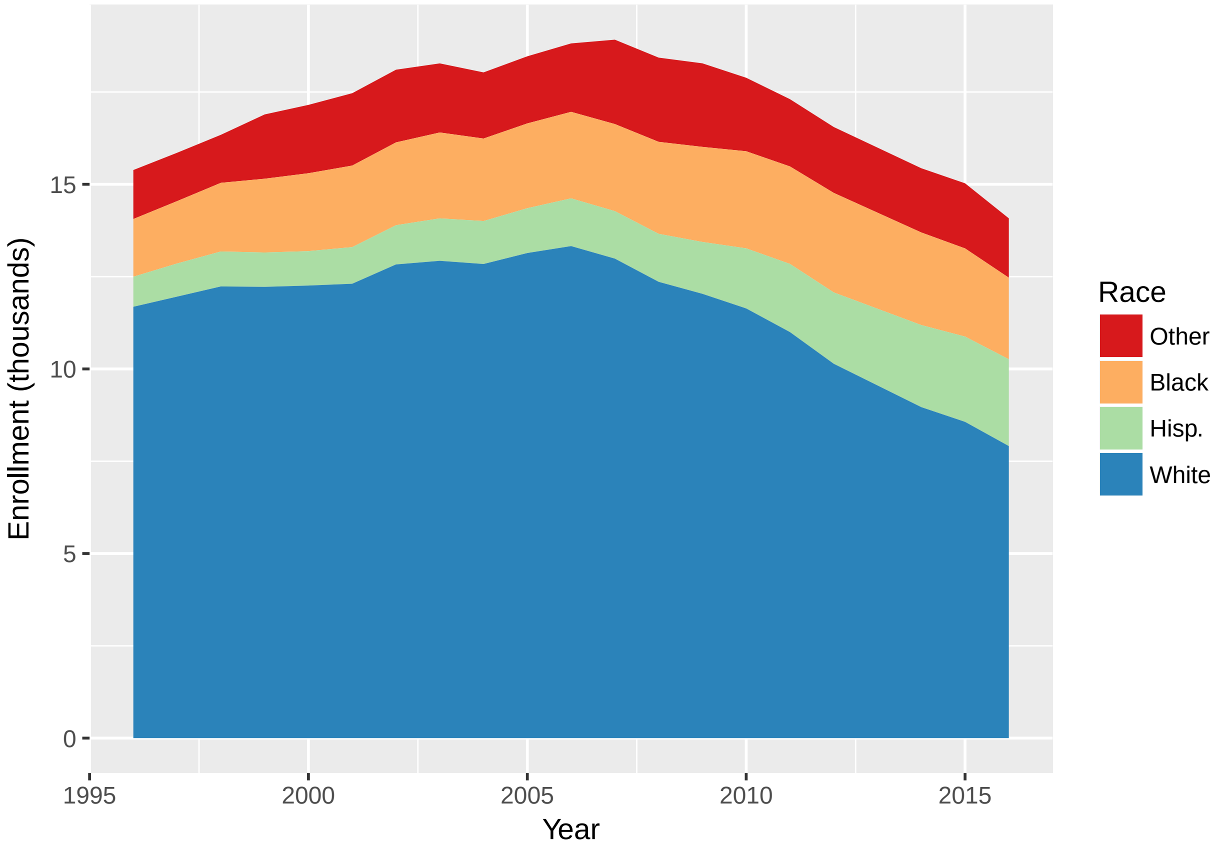
```
out.df.long %>%
  filter(School == 'Northern Illinois University',
         Level == 'Undergraduate') %>%
  group_by(Year, Race) %>%
  summarize(n = sum(Enrollment)) %>%
  mutate(Proportion = prop.table(n)) %>%
  ggplot() +
  geom_area(aes(Year, Proportion, fill=Race, group=Race)) +
  scale_fill_brewer(palette = 'Spectral')
```



```
out.df.long %>%
  filter(School == 'Northern Illinois University',
         Level == 'Undergraduate') %>%
  mutate(Race = fct_collapse(Race,
                             Other = c('Alien',
                                       'Asian',
                                       'Hawaiian',
                                       'Indian',
                                       'Other',
                                       'Two.or.more.races'))) %>%
  group_by(Year, Race) %>%
  summarize(n = sum(Enrollment)) %>%
  ggplot() +
  geom_area(aes(Year, n/1000, fill=Race, group=Race),
            position='fill') +
  labs(y='Proportion of students') +
  scale_fill_brewer(palette = 'Spectral')
```



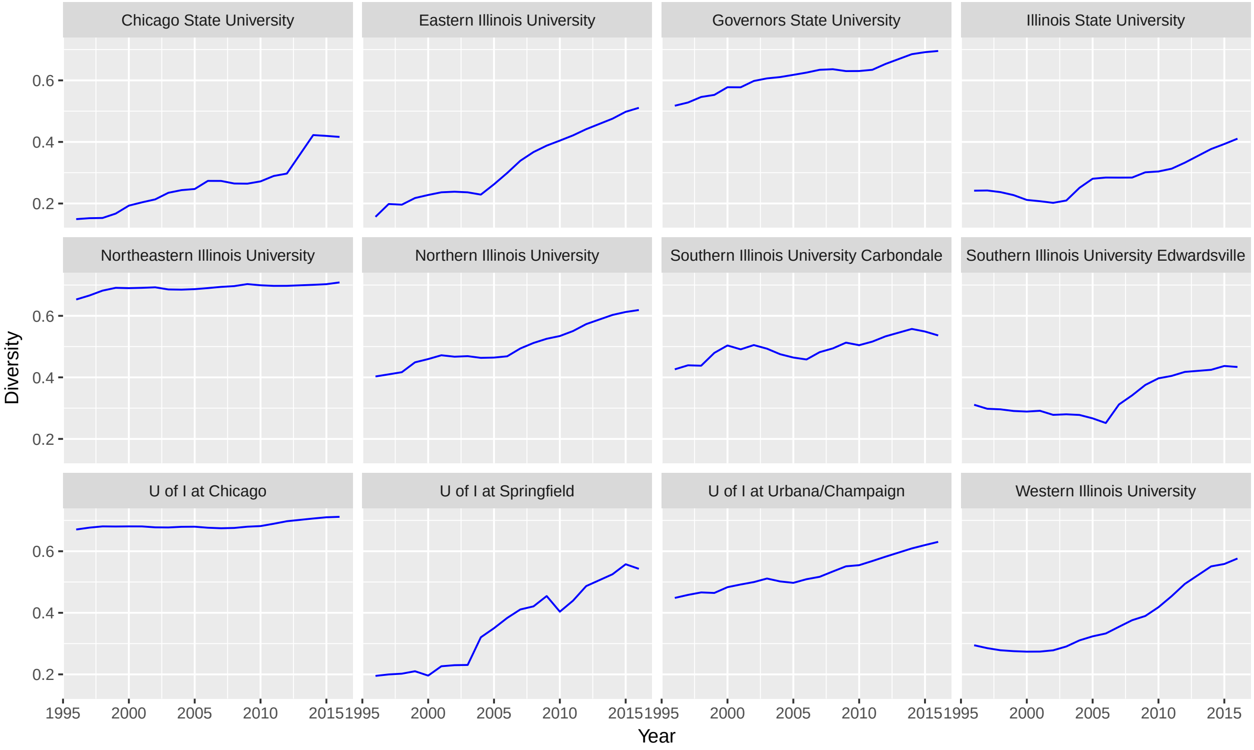
```
out.df.long %>%
  filter(School == 'Northern Illinois University',
         Level == 'Undergraduate') %>%
  mutate(Race = fct_collapse(Race,
                             Other = c('Alien',
                                       'Asian',
                                       'Hawaiian',
                                       'Indian',
                                       'Other',
                                       'Two.or.more.races'))) %>%
  group_by(Year, Race) %>%
  summarize(n = sum(Enrollment)) %>%
  ggplot() +
  geom_area(aes(Year, n/1000, fill=Race, group=Race)) +
  labs(y = 'Enrollment (thousands)') +
  scale_fill_brewer(palette = 'Spectral')
```



Racial diversity within a school over time (M and F combined)

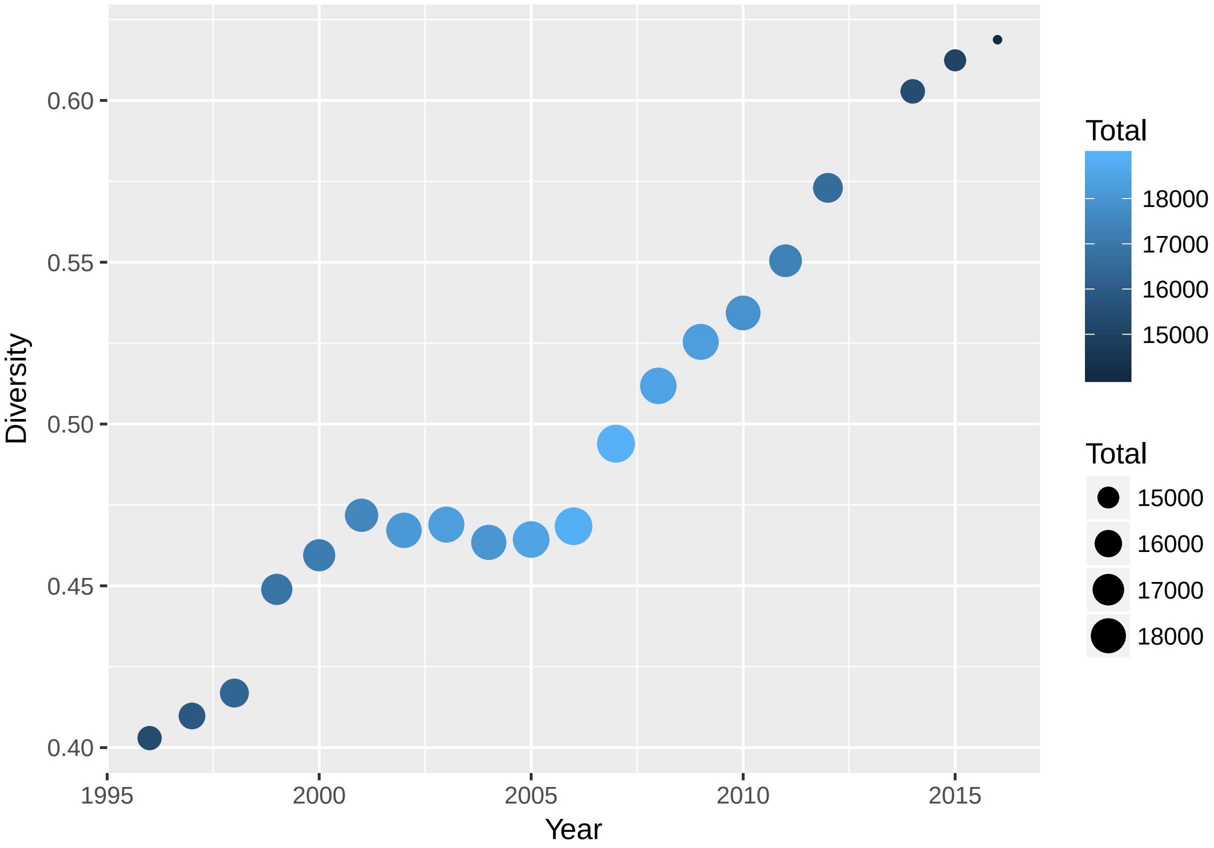
```
library(vegan)

out.df2 = out.df %>%
  gather(Group, Enrollment,
         -Year, -Category, -School, -Level) %>%
  mutate(Group = gsub('.M', '_M', Group),
         Group = gsub('.F', '_F', Group)) %>%
  separate(Group, into=c('Race', 'Sex'),
           sep='_', extra='merge') %>%
  mutate(Race = fct_collapse(Race,
                             Other = c('Alien',
                                       'Asian',
                                       'Hawaiian',
                                       'Indian',
                                       'Other',
                                       'Two.or.more.races'))) %>%
  group_by(Category, School, Year, Level, Race) %>%
  summarize(Enrollment = sum(Enrollment)) %>%
  filter(Level == 'Undergraduate') %>%
  spread(Race, Enrollment) %>%
  mutate(Total = Other + Black + Hisp. + White)
```

```r
out.df2$Diversity = diversity(out.df2[ , 5:8], 'simpson')

out.df2 %>%
  filter(Category == 'Public Universities:') %>%
  ggplot() +
  geom_line(aes(Year, Diversity, group=School), col='blue') +
  facet_wrap(~School)
```

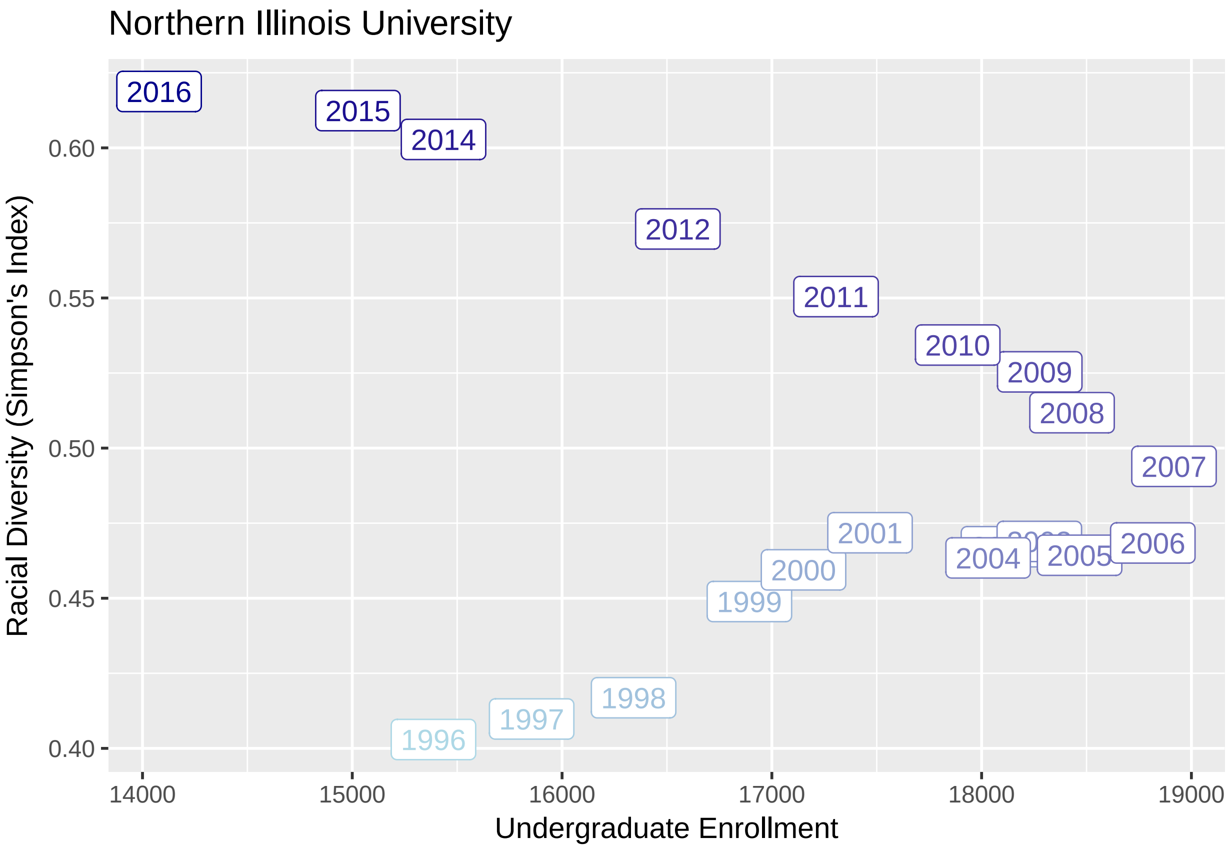


Diversity and Enrollment (over time)

```r
out.df2 %>%
  filter(School == 'Northern Illinois University') %>%
  ggplot() +
  geom_point(aes(Year, Diversity,
                 group=School, size=Total, col=Total))
```



```
sch = "Northern Illinois University"

out.df2 %>%
  filter(School == sch) %>%
  ggplot() +
  geom_point(aes(Total, Diversity), show.legend=F) +
  geom_label(aes(Total, Diversity, label=Year, col=Year),
             show.legend=F) +
  scale_color_gradient(low = 'lightblue', high = 'darkblue') +
  labs(x='Undergraduate Enrollment',
       y="Racial Diversity (Simpson's Index)",
       title=sch)
```


Northern Illinois University
Racial Diversity (Simpson's Index)
0.60
0.55
0.50
0.45
0.40
Undergraduate Enrollment
14000
15000
16000
17000
18000
19000
2016
2015
2014
2012
2011
2010
2009
2008
2007
2001
2006
2004
2005
2000
1999
1998
1997
1996