Demo Analysis Script

Josie Griffin

Load libraries. (Including installing the {reedr} package from Github.)

```
# Do the following line only once ever. So you can uncomment it, run it, and then delete it if you want
# remotes::install_github("data-at-reed/reedr")

# load libraries
library(tidyverse) # mega-package
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(reedr)     # for the data
library(broom)     # for making outputs into nicer tables
```

Load the data and view it in different ways.

```
# load data into object 'raw_grad'
raw_grad <- grad_rates_2011_2024

# shows all columns and their data types and first entries
glimpse(raw_grad)
```

```
## Rows: 13,020
## Columns: 19
## $ unitid                    <dbl> 112260, 112260, 112260, 112260, 112260, 11~
## $ year                      <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, ~
## $ fips                      <chr> "California", "California", "California", ~
## $ cohort_year               <dbl> 2006, 2006, 2006, 2006, 2006, 2006, 2006, ~
## $ institution_level         <chr> "Four or more years", "Four or more years"~
## $ subcohort                 <chr> "Bachelor's or equivalent subcohort of fou~
## $ race                      <chr> "American Indian or Alaska Native", "Ameri~
## $ sex                       <chr> "Female", "Male", "Total", "Female", "Male~
## $ cohort_rev                <dbl> 1, 0, 1, 17, 14, 31, 4, 6, 10, 26, 20, 46,~
## $ exclusions                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```
## $ cohort_adj_150pct           <dbl> 1, 0, 1, 17, 14, 31, 4, 6, 10, 26, 20, 46,~
## $ completers_150pct           <dbl> 1, 0, 1, 17, 14, 31, 4, 5, 9, 24, 19, 43, ~
## $ transfers_out               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ still_enrolled_long_program <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ completers_100pct           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ still_enrolled              <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, ~
## $ no_longer_enrolled          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 3, 0, 0, ~
## $ completion_rate_150pct      <dbl> 1.0000000, NA, 1.0000000, 1.0000000, 1.000~
## $ name                        <chr> "Claremont McKenna College", "Claremont Mc~
```

```
# shows summary stats and missing data for columns
summary(raw_grad)
```

```
##      unitid            year          fips            cohort_year
##  Min.   :112260   Min.   :2011   Length:13020       Min.   :2006
##  1st Qu.:150455   1st Qu.:2014   Class :character   1st Qu.:2009
##  Median :197133   Median :2018   Mode  :character   Median :2012
##  Mean   :184352   Mean   :2018                      Mean   :2012
##  3rd Qu.:211273   3rd Qu.:2021                      3rd Qu.:2016
##  Max.   :239017   Max.   :2024                      Max.   :2019
##
##  institution_level   subcohort            race               sex
##  Length:13020       Length:13020       Length:13020       Length:13020
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    cohort_rev       exclusions     cohort_adj_150pct completers_150pct
##  Min.   :  0.00   Min.   :0.0000   Min.   :  0.00    Min.   :  0.00
##  1st Qu.:  2.00   1st Qu.:0.0000   1st Qu.:  2.00    1st Qu.:  2.00
##  Median : 13.00   Median :0.0000   Median : 15.00    Median : 13.00
##  Mean   : 60.49   Mean   :0.2141   Mean   : 61.82    Mean   : 52.29
##  3rd Qu.: 40.00   3rd Qu.:0.0000   3rd Qu.: 42.00    3rd Qu.: 37.00
##  Max.   :808.00   Max.   :5.0000   Max.   :807.00    Max.   :743.00
##  NA's   :6510     NA's   :11292                      NA's   :15
##  transfers_out     still_enrolled_long_program completers_100pct
##  Min.   :  0.000   Mode:logical                Mode:logical
##  1st Qu.:  0.000   NA's:13020                  NA's:13020
##  Median :  1.000
##  Mean   :  6.108
##  3rd Qu.:  4.000
##  Max.   :119.000
##  NA's   :11310
##  still_enrolled    no_longer_enrolled completion_rate_150pct     name
##  Min.   : 0.0000   Min.   :  0.000    Min.   :0.0000         Length:13020
##  1st Qu.: 0.0000   1st Qu.:  0.000    1st Qu.:0.7500         Class :character
##  Median : 0.0000   Median :  1.000    Median :0.8614         Mode  :character
##  Mean   : 0.3168   Mean   :  7.254    Mean   :0.8211
##  3rd Qu.: 0.0000   3rd Qu.:  5.000    3rd Qu.:0.9355
##  Max.   :10.0000   Max.   :163.000    Max.   :1.0000
##  NA's   :9182      NA's   :6540       NA's   :2272
```

Clean & Wrangle the data to remove data we don't need and create new clean data object. Remove columns that are empty (all missing data). Remove columns that aren't relevant (things we won't use at all). Rename columns to be easily understandable. Show all the values options that exist in categorical variables in columns (show what all is listed for race & sex). Make sure we're only comparing similar schools (remove any non 4-year schools). Filter for only Reed College to create an only Reed data set.

```
# remove empty columns and create new object for data
# use select with - to remove things
grad <- raw_grad |>
  select(-c(still_enrolled_long_program, completers_100pct))

# keep only relevant columns by selecting them
# rearrange with select to have 'name' be the first column
grad <- grad |>
  select(name, year, fips, institution_level, subcohort, race, sex,
         cohort_adj_150pct, completers_150pct, completion_rate_150pct)

# show all column names
colnames(raw_grad)
```

```
##  [1] "unitid"                    "year"
##  [3] "fips"                      "cohort_year"
##  [5] "institution_level"         "subcohort"
##  [7] "race"                      "sex"
##  [9] "cohort_rev"                "exclusions"
## [11] "cohort_adj_150pct"         "completers_150pct"
## [13] "transfers_out"             "still_enrolled_long_program"
## [15] "completers_100pct"         "still_enrolled"
## [17] "no_longer_enrolled"        "completion_rate_150pct"
## [19] "name"
```

```
# rename columns, new = old
grad <- grad |>
  rename(state = fips,
         total_students = cohort_adj_150pct,
         completers = completers_150pct,
         completion_rate = completion_rate_150pct)

# view all unique categories in columns
unique(grad$name)
```

```
##  [1] "Claremont McKenna College" "Occidental College"
##  [3] "Pitzer College"            "Pomona College"
##  [5] "Colorado College"          "Wesleyan University"
##  [7] "Agnes Scott College"       "Earlham College"
##  [9] "Grinnell College"          "Bates College"
## [11] "Carleton College"          "Macalester College"
## [13] "Bard College"              "Hamilton College"
## [15] "Sarah Lawrence College"    "Vassar College"
## [17] "Davidson College"          "Denison University"
## [19] "Kenyon College"            "Oberlin College"
## [21] "Lewis & Clark College"     "Reed College"
## [23] "Willamette University"     "Bryn Mawr College"
```

```
## [25] "Haverford College"         "Swarthmore College"
## [27] "Trinity University"        "University of Puget Sound"
## [29] "Whitman College"           "Beloit College"
## [31] "Lawrence University"
```

```
unique(grad$institution_level)
```

```
## [1] "Four or more years"
```

```
# this is an alternate way to do the same thing
grad |> distinct(subcohort)
```

```
## # A tibble: 2 x 1
##   subcohort
##   <chr>
## 1 Bachelor's or equivalent subcohort of four-year institutions
## 2 Degree/certificate nonbachelor's seeking subcohort of four-year institutions
```

```
# remove 'institution level'
# either run the code below or preferably, add it to the columns that we removed in the code above
grad <- grad |>
  select(-c(institution_level))

# filter to remove Total and nonbachelors from 'subcohort'
grad <- grad |>
  filter(subcohort == "Bachelor's or equivalent subcohort of four-year institutions")

# show all unique values in other columns
unique(grad$race)
```

```
##  [1] "American Indian or Alaska Native"
##  [2] "Asian"
##  [3] "Black"
##  [4] "Hispanic"
##  [5] "Native Hawaiian or other Pacific Islander"
##  [6] "Nonresident alien"
##  [7] "Total"
##  [8] "Two or more races"
##  [9] "Unknown"
## [10] "White"
```

```
unique(grad$sex)
```

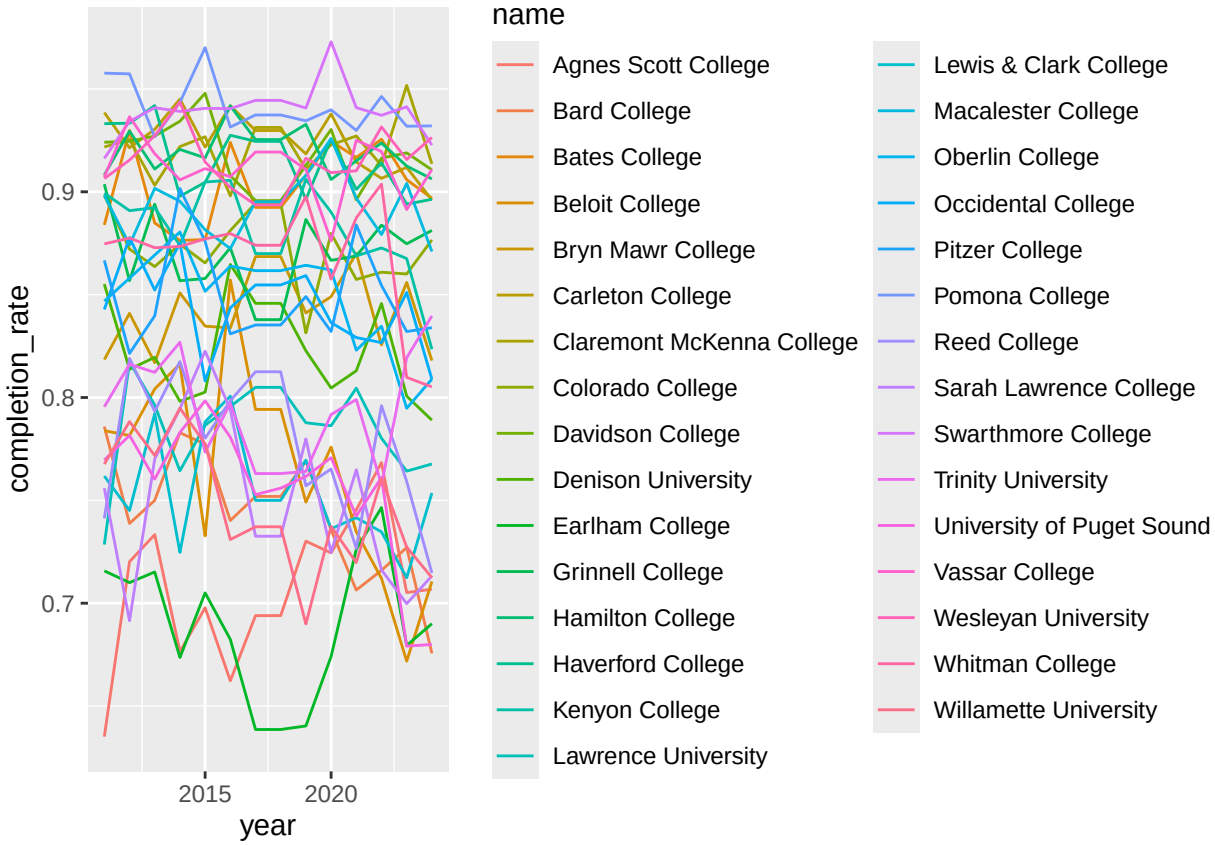
```
## [1] "Female" "Male"   "Total"
```

```
# separate Reed into its own new data object
reed <- grad |>
  filter(name == "Reed College")

# recode data to pooled races, change things with small numbers to be included only as Other
reed <- reed |>
```

```
  mutate(race = case_when(
  race %in% c("Native Hawaiian or other Pacific Islander",
              "American Indian or Alaska Native",
              "Two or more races",
              "Nonresident alien",
              "Unknown") ~ "Other",
  .default = race))
```

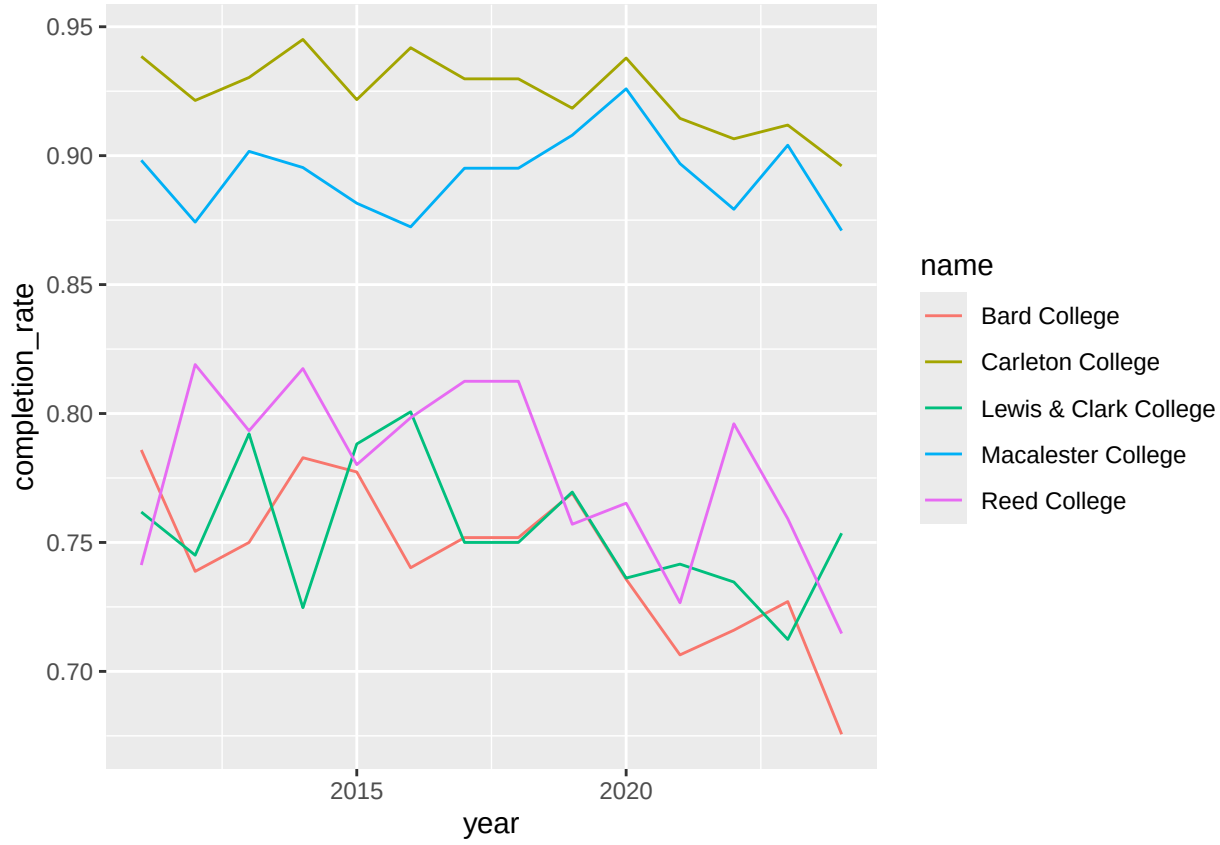
Find stats to compare all schools' completion rates across years. This will involve making multiple new objects to hold different types of data. For example, in one dataset we will only want the total number of students, not things broken down by sex and race.

```
# get single values for grad rates for each school for each year (just totals, not broken down)
grad_totals <- grad |>
  filter(race == "Total",
         sex == "Total")

# plot completion over time as line plot, separate by college (name)
ggplot(data = grad_totals, aes(x = year, y = completion_rate, color = name)) +
  geom_line()
```




```
# too many colleges to make sense!
# pick selected colleges to compare to reed (sub out names for whatever you want)
subset_grad <- grad_totals |>
  filter(name == "Reed College" |
```

```
           name == "Lewis & Clark College" |
           name == "Macalester College" |
           name == "Bard College" |
           name == "Carleton College")

# replot with just the subsetted data
ggplot(data = subset_grad, aes(x = year, y = completion_rate, color = name)) +
  geom_line()
```




```
# show the data for just the subset
subset_grad
```

```
## # A tibble: 70 x 9
##    name              year state  subcohort race  sex   total_students completers
##    <chr>            <dbl> <chr>  <chr>     <chr> <chr>          <dbl>      <dbl>
##  1 Carleton College  2011 Minne~ Bachelor~ Total Total            504        473
##  2 Carleton College  2012 Minne~ Bachelor~ Total Total            509        469
##  3 Carleton College  2013 Minne~ Bachelor~ Total Total            488        454
##  4 Carleton College  2014 Minne~ Bachelor~ Total Total            528        499
##  5 Carleton College  2015 Minne~ Bachelor~ Total Total            511        471
##  6 Carleton College  2016 Minne~ Bachelor~ Total Total            516        486
##  7 Carleton College  2017 Minne~ Bachelor~ Total Total            527        490
##  8 Carleton College  2018 Minne~ Bachelor~ Total Total            527        490
##  9 Carleton College  2019 Minne~ Bachelor~ Total Total            527        484
## 10 Carleton College  2020 Minne~ Bachelor~ Total Total            515        483
```

```
## # i 60 more rows
## # i 1 more variable: completion_rate <dbl>
```

```
# or
View(subset_grad)
```

Now we'll break down our big dataset to show trends within sex. We'll get summary stats for each sex over each year and we'll plot these with a boxplot. Then we'll test to see if there are signficant differences between sexes with a t-test.

```
# show all variables in the column 'sex'
unique(grad$sex)
```

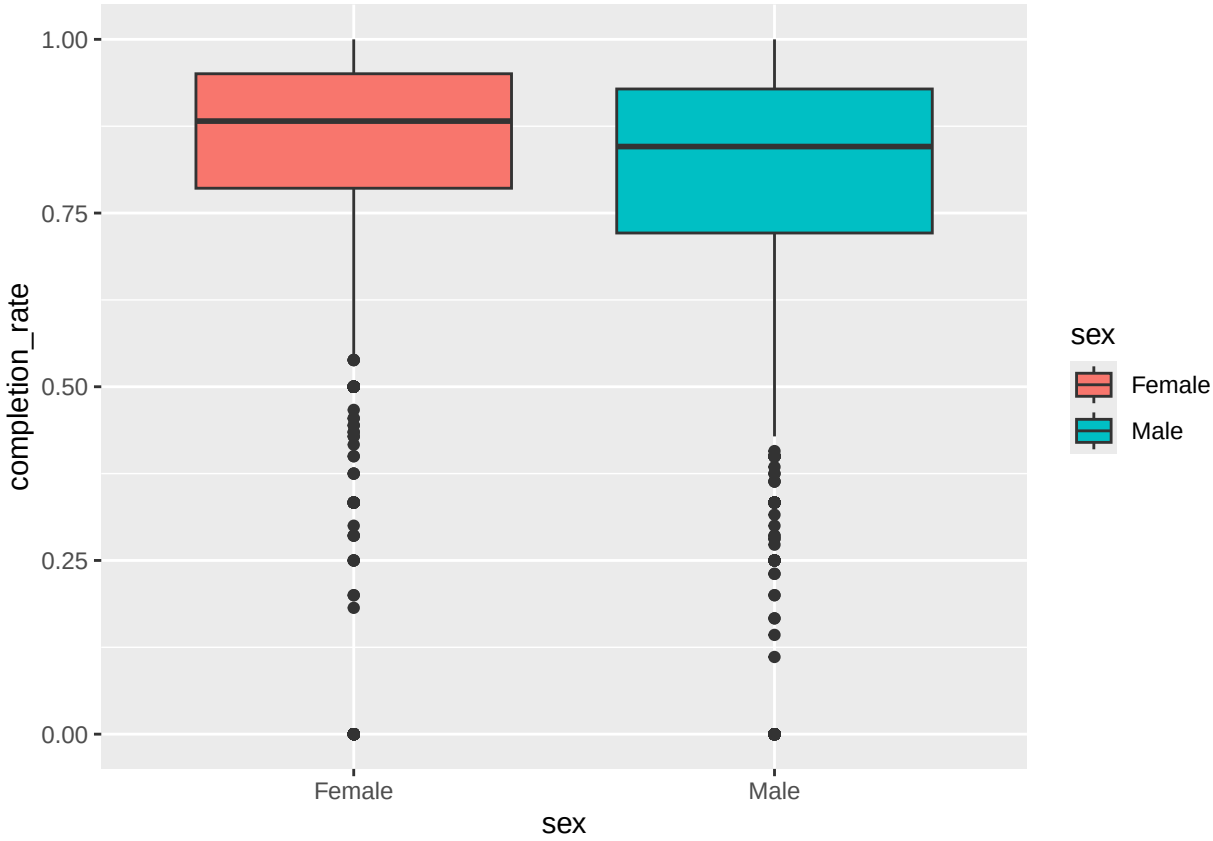
```
## [1] "Female" "Male"   "Total"
```

```
# get summary statistics divided into groups by sex and year
# need "na.rm = TRUE" to remove missing values
sex_stats <- grad |>
  group_by(sex, year) |>
  summarize(mean = mean(completion_rate, na.rm = TRUE),
            stdev = sd(completion_rate, na.rm = TRUE),
            median = median(completion_rate, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

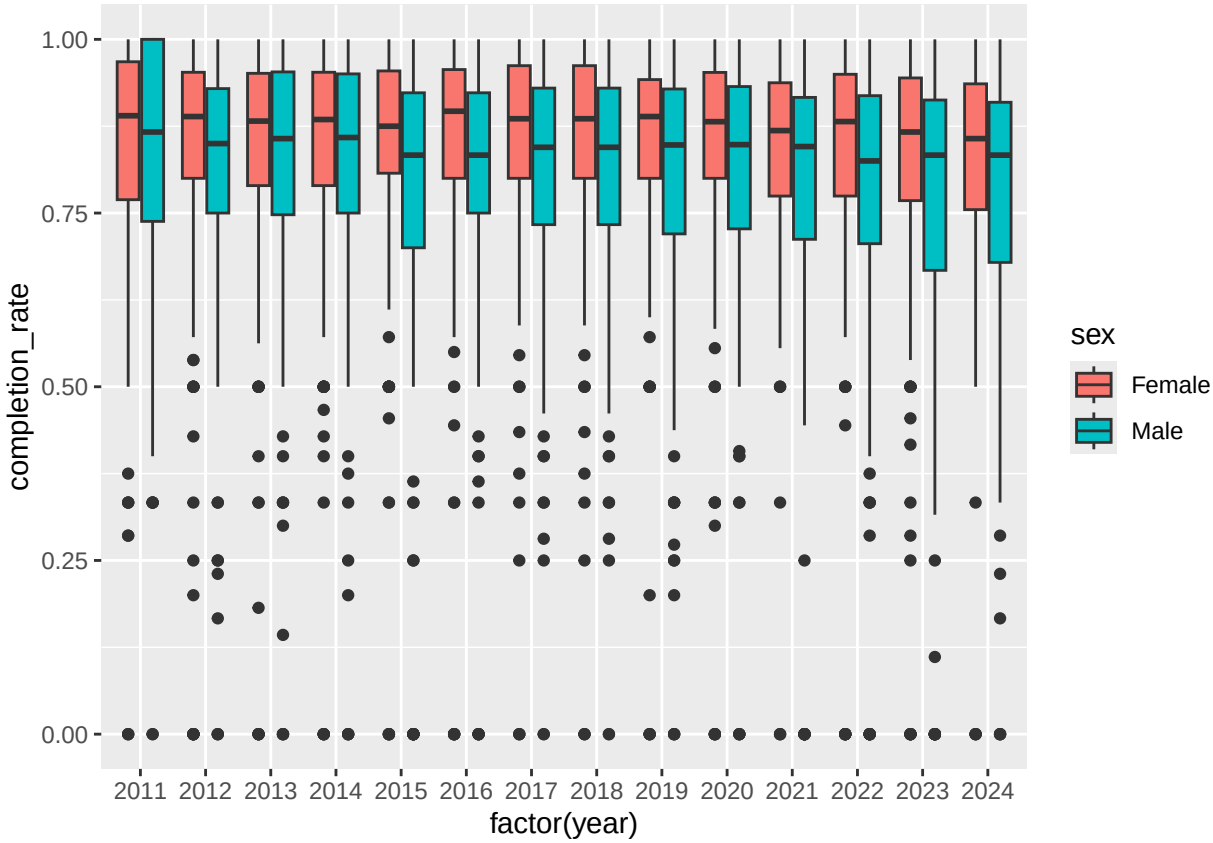
```
# to make the graph compare just male vs female, we need to remove "Total"
# we can do that as we have before, with making a new dataset, or we can do it within the graph plot
ggplot(data = grad |> filter(sex != "Total"), aes(x = sex, y = completion_rate, fill = sex)) +
  geom_boxplot() # this graph also just plots overall male vs female
```

```
## Warning: Removed 1635 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
# this graph breaks that down over year
# we need to add the 'factor' function so that R treats year like a category instead of a number
ggplot(data = grad |> filter(sex != "Total"), aes(x = factor(year), y = completion_rate, fill = sex)) +
  geom_boxplot()
```

```
## Warning: Removed 1635 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
# t-test for significant differences
# we're running it with '|> filter(sex != "Total")' because a t-test only works for two variables
# so it will only work for male vs female, not male, female, and total
t.test(completion_rate ~ sex, data = grad |> filter(sex != "Total"))
```

```
##
##  Welch Two Sample t-test
##
## data:  completion_rate by sex
## t = 9.5024, df = 6700.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  0.03324490 0.05052678
## sample estimates:
## mean in group Female   mean in group Male
##            0.8423380            0.8004521
```

```
# the above output is okay, but if we add the library {broom}, we can use the tidy function
# tidy makes our output into a more visually understandable table
tidy(t.test(completion_rate ~ sex, data = grad |> filter(sex != "Total")))
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic  p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl>
## 1   0.0419     0.842     0.800      9.50 2.79e-21     6700.   0.0332    0.0505
```

```
## # i 2 more variables: method <chr>, alternative <chr>
```

```
# the previous code compares male vs female combined over all years
# if we want to do this t-test for each individual year we can add a group & a summarize function
grad |>
  filter(sex != "Total") |>
  group_by(year) |>
  summarize(tidy(t.test(completion_rate ~ sex))) |>
  ungroup() # we add ungroup at the end just to make sure we return 'grad' to normal and don't have hid
```

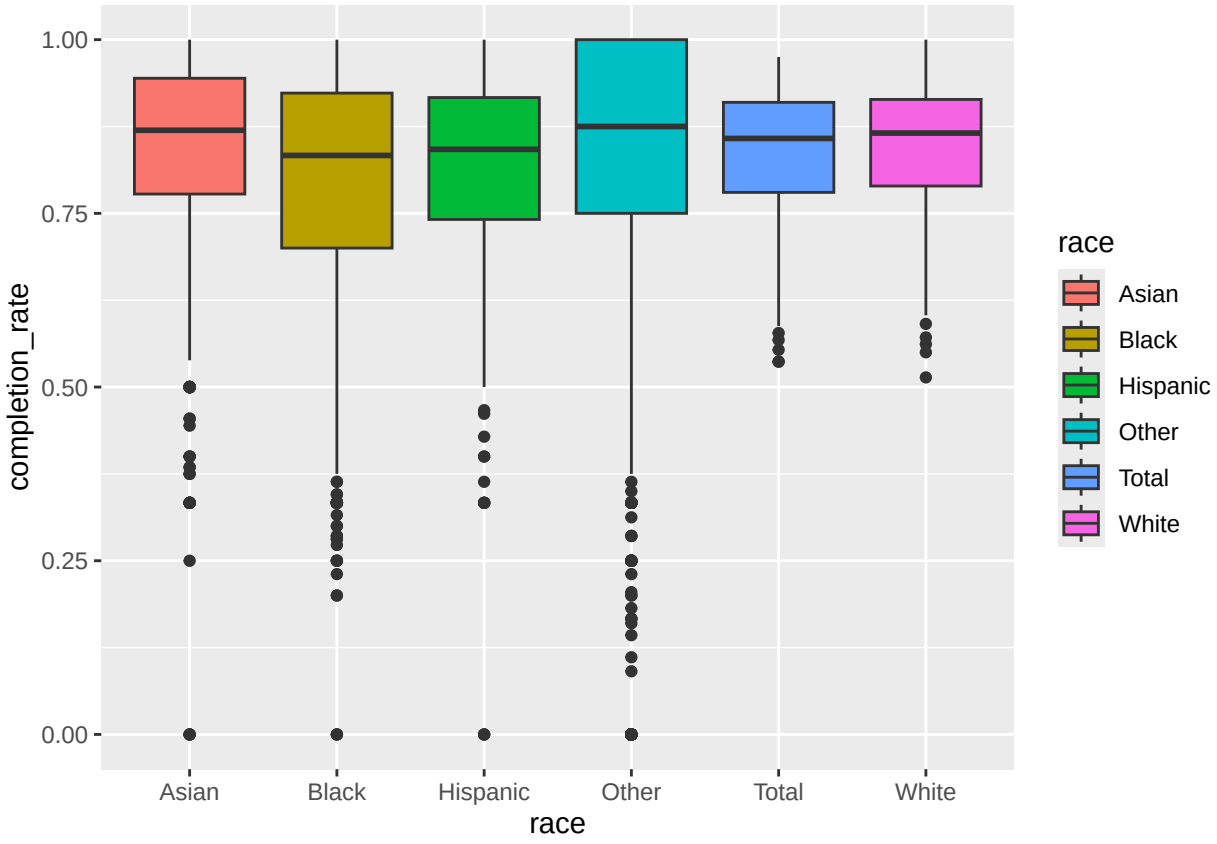
```
## # A tibble: 14 x 11
##     year estimate estimate1 estimate2 statistic  p.value parameter conf.low
##    <dbl>    <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
##  1  2011   0.0226     0.842     0.820      1.38 0.168         470. -0.00952
##  2  2012   0.0269     0.839     0.812      1.58 0.115         489. -0.00654
##  3  2013   0.0292     0.842     0.813      1.75 0.0810        477. -0.00361
##  4  2014   0.0200     0.841     0.821      1.22 0.222         482. -0.0121
##  5  2015   0.0630     0.843     0.780      3.54 0.000435      463.  0.0280
##  6  2016   0.0409     0.849     0.808      2.46 0.0144        475.  0.00819
##  7  2017   0.0398     0.852     0.812      2.64 0.00858       477.  0.0102
##  8  2018   0.0395     0.852     0.812      2.62 0.00898       478.  0.00992
##  9  2019   0.0521     0.852     0.800      3.24 0.00130       464.  0.0204
## 10  2020   0.0365     0.846     0.810      2.32 0.0205        498.  0.00565
## 11  2021   0.0558     0.846     0.790      3.45 0.000625      423.  0.0240
## 12  2022   0.0483     0.832     0.784      2.81 0.00519       495.  0.0145
## 13  2023   0.0552     0.826     0.771      3.08 0.00220       480.  0.0200
## 14  2024   0.0531     0.833     0.780      3.30 0.00106       471.  0.0215
## # i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

Now we'll break our code down by race. We'll combine some racial categories because they are small and their trends will likely be skewed due to small sample size. Then we'll get summary stats for each race over each year and we'll plot these with a boxplot. Then we'll use an ANOVA to see if there are signficant differences between completion rates by race. Then we'll use a Tukey test to find individual paired differences.

```
# combine races that have few students into category called "Other"
# case_when() finds matches and replaces them with whatever is after the ~
# %in% is a way of determining if something matches any of the listed conditions
grad_other <- grad |>
  mutate(race = case_when(
    race %in% c("Native Hawaiian or other Pacific Islander",
                "American Indian or Alaska Native",
                "Two or more races",
                "Nonresident alien",
                "Unknown") ~ "Other",
    .default = race))

# find the summary statistics for the different races
# note that you can do this with 'grad_other' or 'grad'
race_stats <- grad_other |>
  group_by(race) |>
  summarize(mean = mean(completion_rate, na.rm = TRUE),
            stdev = sd(completion_rate, na.rm = TRUE),
            median = median(completion_rate, na.rm = TRUE))
```

```
# make a boxplot for overall racial differences
ggplot(data = grad_other, aes(x = race, y = completion_rate, fill = race)) +
  geom_boxplot()
```

```
## Warning: Removed 2197 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```




```
# use an ANOVA to see if there is a difference between our racial groups
# this creates the ANOVA model object
anova <- aov(completion_rate ~ race, data = grad)
# this shows you the summary of the ANOVA
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## race            9   12.7  1.4076   45.07 <2e-16 ***
## Residuals   10721  334.8  0.0312
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 2197 observations deleted due to missingness
```

```
# this shows the same thing but a little bit nicer, tidy() is from the {broom} library
tidy(anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic   p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>     <dbl>
## 1 race          9  12.7 1.41        45.1  3.21e-80
## 2 Residuals 10721 335.  0.0312      NA   NA
```

```
# this shows all the pairwise comparisons that exist within our ANOVA model
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = completion_rate ~ race, data = grad)
##
## $race
##                                                                                 diff
## Asian-American Indian or Alaska Native                                     0.118663633
## Black-American Indian or Alaska Native                                     0.071370805
## Hispanic-American Indian or Alaska Native                                  0.096786112
## Native Hawaiian or other Pacific Islander-American Indian or Alaska Native 0.010226093
## Nonresident alien-American Indian or Alaska Native                         0.130466905
## Total-American Indian or Alaska Native                                     0.117369509
## Two or more races-American Indian or Alaska Native                         0.110627189
## Unknown-American Indian or Alaska Native                                   0.078206331
## White-American Indian or Alaska Native                                     0.122510779
## Black-Asian                                                               -0.047292828
## Hispanic-Asian                                                            -0.021877521
## Native Hawaiian or other Pacific Islander-Asian                           -0.108437541
## Nonresident alien-Asian                                                    0.011803271
## Total-Asian                                                               -0.001294124
## Two or more races-Asian                                                   -0.008036445
## Unknown-Asian                                                             -0.040457303
## White-Asian                                                                0.003847145
## Hispanic-Black                                                             0.025415307
## Native Hawaiian or other Pacific Islander-Black                           -0.061144713
## Nonresident alien-Black                                                    0.059096099
## Total-Black                                                                0.045998704
## Two or more races-Black                                                    0.039256383
## Unknown-Black                                                              0.006835525
## White-Black                                                                0.051139973
## Native Hawaiian or other Pacific Islander-Hispanic                        -0.086560019
## Nonresident alien-Hispanic                                                 0.033680793
## Total-Hispanic                                                             0.020583397
## Two or more races-Hispanic                                                 0.013841077
## Unknown-Hispanic                                                          -0.018579781
## White-Hispanic                                                             0.025724667
## Nonresident alien-Native Hawaiian or other Pacific Islander                0.120240812
## Total-Native Hawaiian or other Pacific Islander                            0.107143416
## Two or more races-Native Hawaiian or other Pacific Islander                0.100401096
## Unknown-Native Hawaiian or other Pacific Islander                          0.067980238
## White-Native Hawaiian or other Pacific Islander                            0.112284686
## Total-Nonresident alien                                                   -0.013097395
## Two or more races-Nonresident alien                                       -0.019839716
## Unknown-Nonresident alien                                                 -0.052260574
```

```
## White-Nonresident alien                                                          -0.007956126
## Two or more races-Total                                                          -0.006742321
## Unknown-Total                                                                    -0.039163179
## White-Total                                                                       0.005141269
## Unknown-Two or more races                                                        -0.032420858
## White-Two or more races                                                           0.011883590
## White-Unknown                                                                     0.044304448
##                                                                                           lwr
## Asian-American Indian or Alaska Native                                            0.091490896
## Black-American Indian or Alaska Native                                            0.044169086
## Hispanic-American Indian or Alaska Native                                         0.069613375
## Native Hawaiian or other Pacific Islander-American Indian or Alaska Native -0.031390496
## Nonresident alien-American Indian or Alaska Native                                0.103254230
## Total-American Indian or Alaska Native                                            0.090207554
## Two or more races-American Indian or Alaska Native                                0.083015384
## Unknown-American Indian or Alaska Native                                          0.050454429
## White-American Indian or Alaska Native                                            0.095345234
## Black-Asian                                                                      -0.069572947
## Hispanic-Asian                                                                   -0.044122247
## Native Hawaiian or other Pacific Islander-Asian                                  -0.147017465
## Nonresident alien-Asian                                                          -0.010490222
## Total-Asian                                                                      -0.023525678
## Two or more races-Asian                                                          -0.030815425
## Unknown-Asian                                                                    -0.063405901
## White-Asian                                                                      -0.018388793
## Hispanic-Black                                                                    0.003135188
## Native Hawaiian or other Pacific Islander-Black                                  -0.099745056
## Nonresident alien-Black                                                           0.036767290
## Total-Black                                                                       0.023731736
## Two or more races-Black                                                           0.016442839
## Unknown-Black                                                                    -0.016147383
## White-Black                                                                       0.028868628
## Native Hawaiian or other Pacific Islander-Hispanic                               -0.125139944
## Nonresident alien-Hispanic                                                        0.011387300
## Total-Hispanic                                                                   -0.001648156
## Two or more races-Hispanic                                                       -0.008937903
## Unknown-Hispanic                                                                 -0.041528380
## White-Hispanic                                                                    0.003488729
## Nonresident alien-Native Hawaiian or other Pacific Islander                       0.081632747
## Total-Native Hawaiian or other Pacific Islander                                   0.068571085
## Two or more races-Native Hawaiian or other Pacific Islander                       0.061510676
## Unknown-Native Hawaiian or other Pacific Islander                                 0.028990226
## White-Native Hawaiian or other Pacific Islander                                   0.073709827
## Total-Nonresident alien                                                          -0.035377746
## Two or more races-Nonresident alien                                              -0.042666322
## Unknown-Nonresident alien                                                        -0.075256448
## White-Nonresident alien                                                          -0.030240851
## Two or more races-Total                                                          -0.029508438
## Unknown-Total                                                                    -0.062099010
## White-Total                                                                      -0.017081492
## Unknown-Two or more races                                                        -0.055887692
## White-Two or more races                                                          -0.010886809
## White-Unknown                                                                     0.021364367
##                                                                                           upr
```

```
## Asian-American Indian or Alaska Native                                          0.1458363707
## Black-American Indian or Alaska Native                                          0.0985725249
## Hispanic-American Indian or Alaska Native                                       0.1239588493
## Native Hawaiian or other Pacific Islander-American Indian or Alaska Native      0.0518426817
## Nonresident alien-American Indian or Alaska Native                              0.1576795798
## Total-American Indian or Alaska Native                                          0.1445314649
## Two or more races-American Indian or Alaska Native                              0.1382389939
## Unknown-American Indian or Alaska Native                                        0.1059582316
## White-American Indian or Alaska Native                                          0.1496763229
## Black-Asian                                                                    -0.0250127093
## Hispanic-Asian                                                                  0.0003672039
## Native Hawaiian or other Pacific Islander-Asian                                -0.0698576155
## Nonresident alien-Asian                                                         0.0340967644
## Total-Asian                                                                     0.0209374295
## Two or more races-Asian                                                         0.0147425354
## Unknown-Asian                                                                  -0.0175087046
## White-Asian                                                                     0.0260830834
## Hispanic-Black                                                                  0.0476954252
## Native Hawaiian or other Pacific Islander-Black                                -0.0225443694
## Nonresident alien-Black                                                         0.0814249084
## Total-Black                                                                     0.0682656718
## Two or more races-Black                                                         0.0620699279
## Unknown-Black                                                                   0.0298184328
## White-Black                                                                     0.0734113186
## Native Hawaiian or other Pacific Islander-Hispanic                             -0.0479800941
## Nonresident alien-Hispanic                                                      0.0559742858
## Total-Hispanic                                                                  0.0428149510
## Two or more races-Hispanic                                                      0.0366200568
## Unknown-Hispanic                                                                0.0043688168
## White-Hispanic                                                                  0.0479606048
## Nonresident alien-Native Hawaiian or other Pacific Islander                     0.1588488763
## Total-Native Hawaiian or other Pacific Islander                                 0.1457157483
## Two or more races-Native Hawaiian or other Pacific Islander                     0.1392915155
## Unknown-Native Hawaiian or other Pacific Islander                               0.1069702488
## White-Native Hawaiian or other Pacific Islander                                 0.1508595448
## Total-Nonresident alien                                                         0.0091829548
## Two or more races-Nonresident alien                                             0.0029868904
## Unknown-Nonresident alien                                                      -0.0292647009
## White-Nonresident alien                                                         0.0143285990
## Two or more races-Total                                                         0.0160237969
## Unknown-Total                                                                  -0.0162273480
## White-Total                                                                     0.0273640306
## Unknown-Two or more races                                                      -0.0089540244
## White-Two or more races                                                         0.0346539890
## White-Unknown                                                                   0.0672445288
##                                                                                      p adj
## Asian-American Indian or Alaska Native                                          0.0000000
## Black-American Indian or Alaska Native                                          0.0000000
## Hispanic-American Indian or Alaska Native                                       0.0000000
## Native Hawaiian or other Pacific Islander-American Indian or Alaska Native      0.9988829
## Nonresident alien-American Indian or Alaska Native                              0.0000000
## Total-American Indian or Alaska Native                                          0.0000000
## Two or more races-American Indian or Alaska Native                              0.0000000
## Unknown-American Indian or Alaska Native                                        0.0000000
```

```
## White-American Indian or Alaska Native                                    0.0000000
## Black-Asian                                                               0.0000000
## Hispanic-Asian                                                            0.0584099
## Native Hawaiian or other Pacific Islander-Asian                           0.0000000
## Nonresident alien-Asian                                                   0.8096008
## Total-Asian                                                               1.0000000
## Two or more races-Asian                                                   0.9831852
## Unknown-Asian                                                             0.0000011
## White-Asian                                                               0.9999380
## Hispanic-Black                                                            0.0114636
## Native Hawaiian or other Pacific Islander-Black                           0.0000240
## Nonresident alien-Black                                                   0.0000000
## Total-Black                                                               0.0000000
## Two or more races-Black                                                   0.0000024
## Unknown-Black                                                             0.9951057
## White-Black                                                               0.0000000
## Native Hawaiian or other Pacific Islander-Hispanic                        0.0000000
## Nonresident alien-Hispanic                                                0.0000772
## Total-Hispanic                                                            0.0975227
## Two or more races-Hispanic                                                0.6534681
## Unknown-Hispanic                                                          0.2360507
## White-Hispanic                                                            0.0095220
## Nonresident alien-Native Hawaiian or other Pacific Islander               0.0000000
## Total-Native Hawaiian or other Pacific Islander                           0.0000000
## Two or more races-Native Hawaiian or other Pacific Islander               0.0000000
## Unknown-Native Hawaiian or other Pacific Islander                         0.0000016
## White-Native Hawaiian or other Pacific Islander                           0.0000000
## Total-Nonresident alien                                                   0.6961692
## Two or more races-Nonresident alien                                       0.1540378
## Unknown-Nonresident alien                                                 0.0000000
## White-Nonresident alien                                                   0.9817451
## Two or more races-Total                                                   0.9952593
## Unknown-Total                                                             0.0000030
## White-Total                                                               0.9993106
## Unknown-Two or more races                                                 0.0005260
## White-Two or more races                                                   0.8224944
## White-Unknown                                                             0.0000000
```

```
# you can also 'tidy' this table
tidy(TukeyHSD(anova))
```

```
## # A tibble: 45 x 7
##    term  contrast             null.value estimate conf.low conf.high adj.p.value
##    <chr> <chr>                     <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 race  Asian-American Indi~          0   0.119    0.0915    0.146     0
##  2 race  Black-American Indi~          0   0.0714   0.0442    0.0986    0
##  3 race  Hispanic-American I~          0   0.0968   0.0696    0.124     0
##  4 race  Native Hawaiian or ~          0   0.0102  -0.0314    0.0518    9.99e- 1
##  5 race  Nonresident alien-A~          0   0.130    0.103     0.158     0
##  6 race  Total-American Indi~          0   0.117    0.0902    0.145     0
##  7 race  Two or more races-A~          0   0.111    0.0830    0.138     0
##  8 race  Unknown-American In~          0   0.0782   0.0505    0.106     0
##  9 race  White-American Indi~          0   0.123    0.0953    0.150     0
## 10 race  Black-Asian                  0  -0.0473  -0.0696   -0.0250    2.50e-10
```

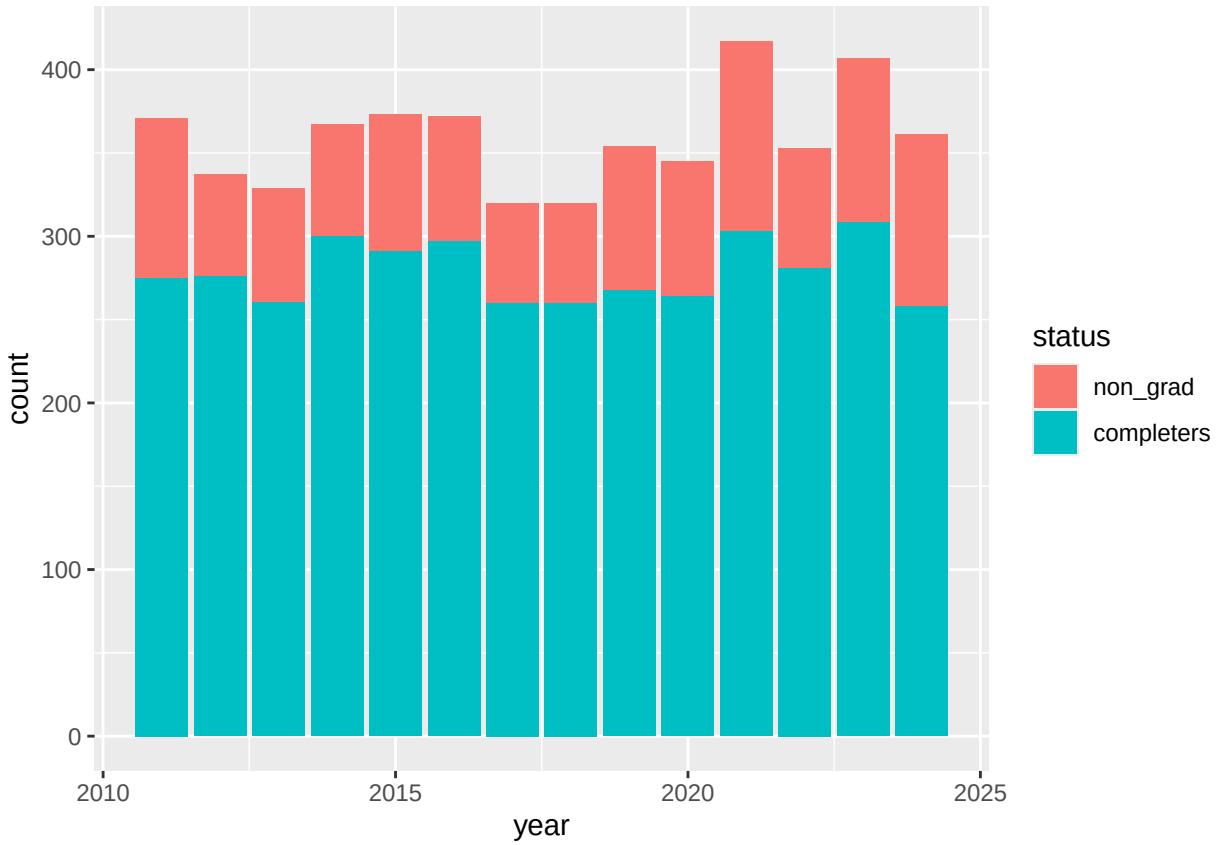
```
## # i 35 more rows
```

```
# you can also assign it to a new object to see the full table
tukey_tidy <- tidy(TukeyHSD(anova))
tukey_tidy # or View(tukey_tidy)
```

```
## # A tibble: 45 x 7
##    term  contrast             null.value estimate conf.low conf.high adj.p.value
##    <chr> <chr>                     <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 race  Asian-American Indi~          0   0.119    0.0915    0.146     0
##  2 race  Black-American Indi~          0   0.0714   0.0442    0.0986    0
##  3 race  Hispanic-American I~          0   0.0968   0.0696    0.124     0
##  4 race  Native Hawaiian or ~          0   0.0102  -0.0314    0.0518    9.99e- 1
##  5 race  Nonresident alien-A~          0   0.130    0.103     0.158     0
##  6 race  Total-American Indi~          0   0.117    0.0902    0.145     0
##  7 race  Two or more races-A~          0   0.111    0.0830    0.138     0
##  8 race  Unknown-American In~          0   0.0782   0.0505    0.106     0
##  9 race  White-American Indi~          0   0.123    0.0953    0.150     0
## 10 race  Black-Asian                   0  -0.0473  -0.0696   -0.0250    2.50e-10
## # i 35 more rows
```

Now we'll do calculations for Reed specific statistics. We'll look at number of students that graduated each year and number that didn't out of the total number of students. We'll make these numbers into a stacked bar graph to show the actual total number of students. (To do this, we'll pivot our data.)

```
# get just Reed college totals
reed_totals <- reed |>
  filter(race == "Total",
         sex == "Total")

# if we want a stacked bar plot of grads vs non-grads, we need a non-grads column
reed_totals <- reed_totals |>
  mutate(non_grad = total_students - completers)

# we can pivot our data so completers and non-grads are in same column, this make them easier to put in
pivot_reed <- reed_totals |>
  pivot_longer(cols = c(completers, non_grad), # put the columns where you want to pull data from
               names_to = "status",            # put the name of the new column where the categoricals
               values_to = "count")            # put the name of the column where the values will go

# re-level to put statuses in correct order
pivot_reed$status <- factor(pivot_reed$status, levels = c("non_grad", "completers"))

# make bar plot of the pivoted reed data
ggplot(pivot_reed, aes(x = year, y = count, fill = status)) +
  geom_col()
```

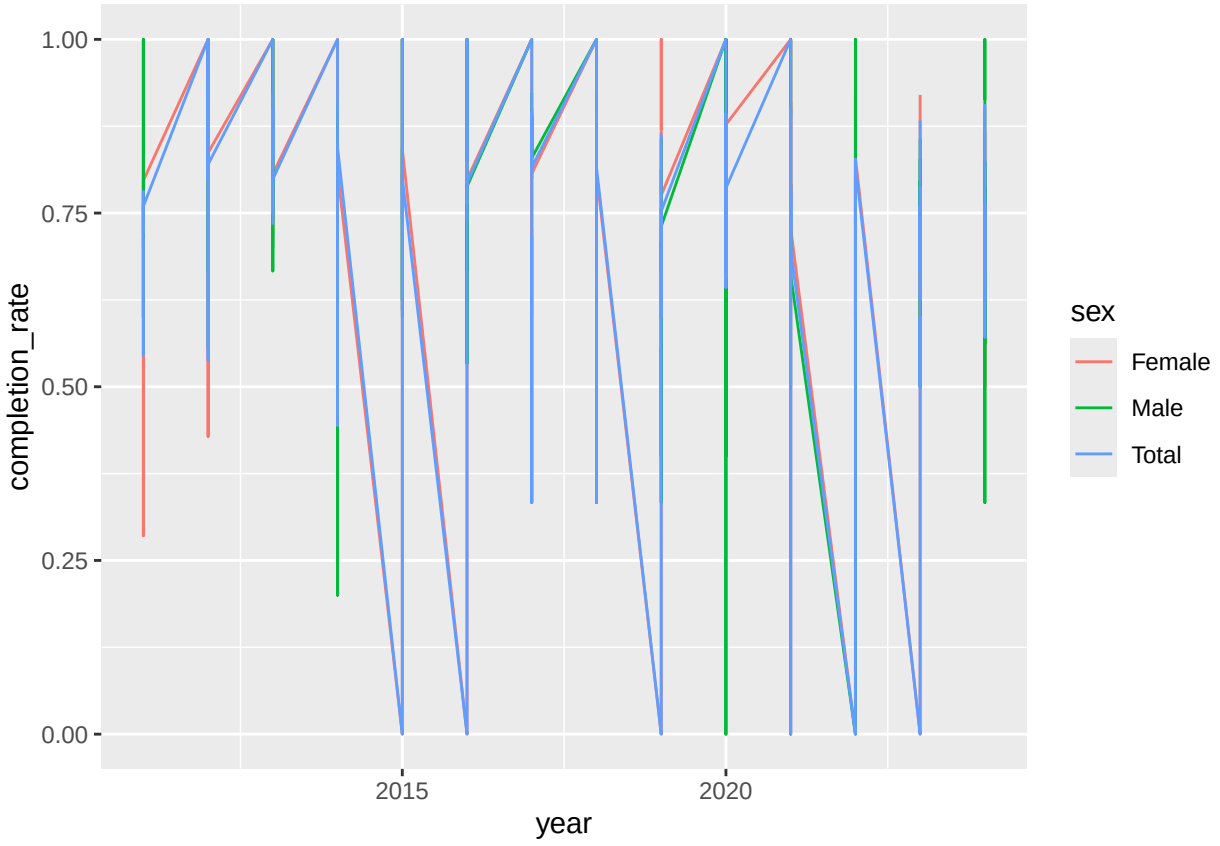


Now we'll calculate summary statistics with the "Other" group classification. We'll plot the completion rate over time broken down by demographic group. We'll show how to break the graph into multiple plots to show differences across both race and sex.

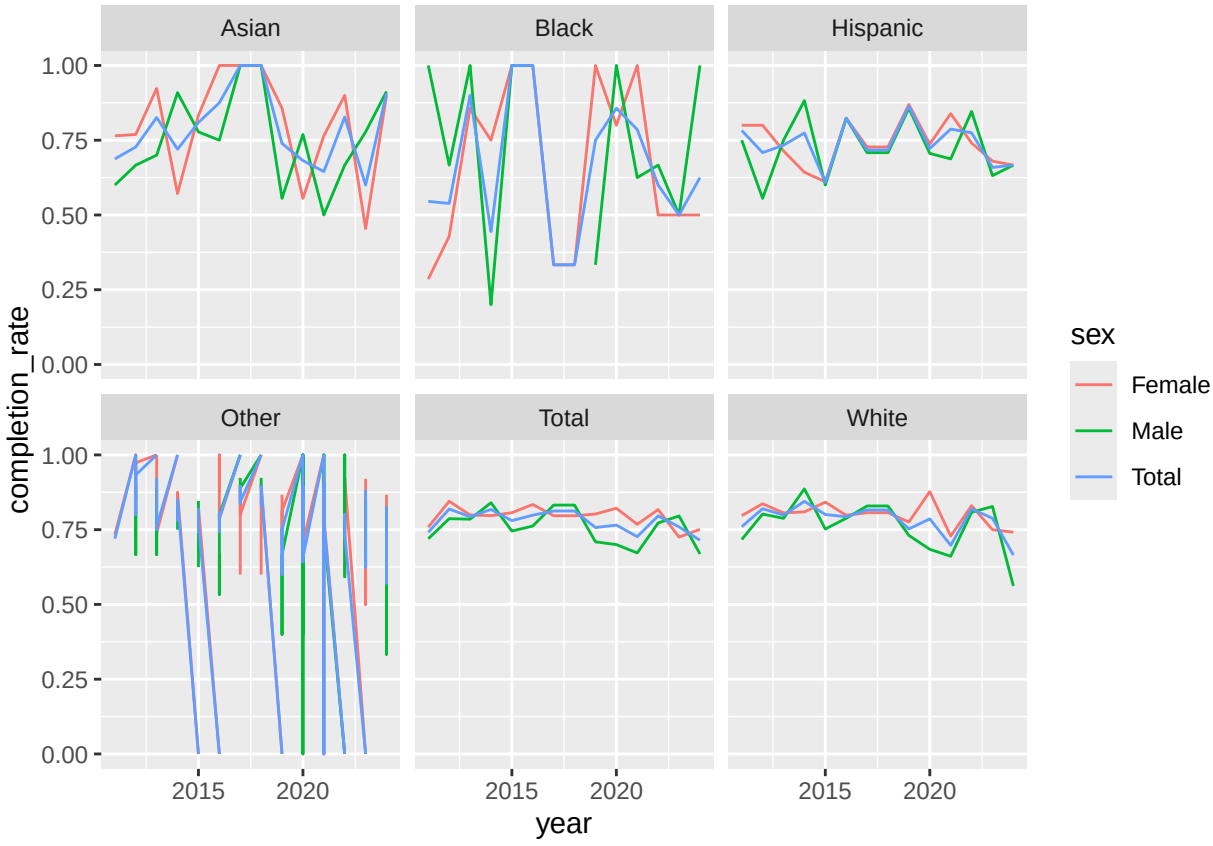
```
# calculate summary statistic based on year and race or sex (or both, just change what is in group_by)
reed_race_stats <- reed |>
  group_by(year, race) |>
  summarize(mean = mean(completion_rate, na.rm = TRUE),
            stdev = sd(completion_rate, na.rm = TRUE),
            median = median(completion_rate, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

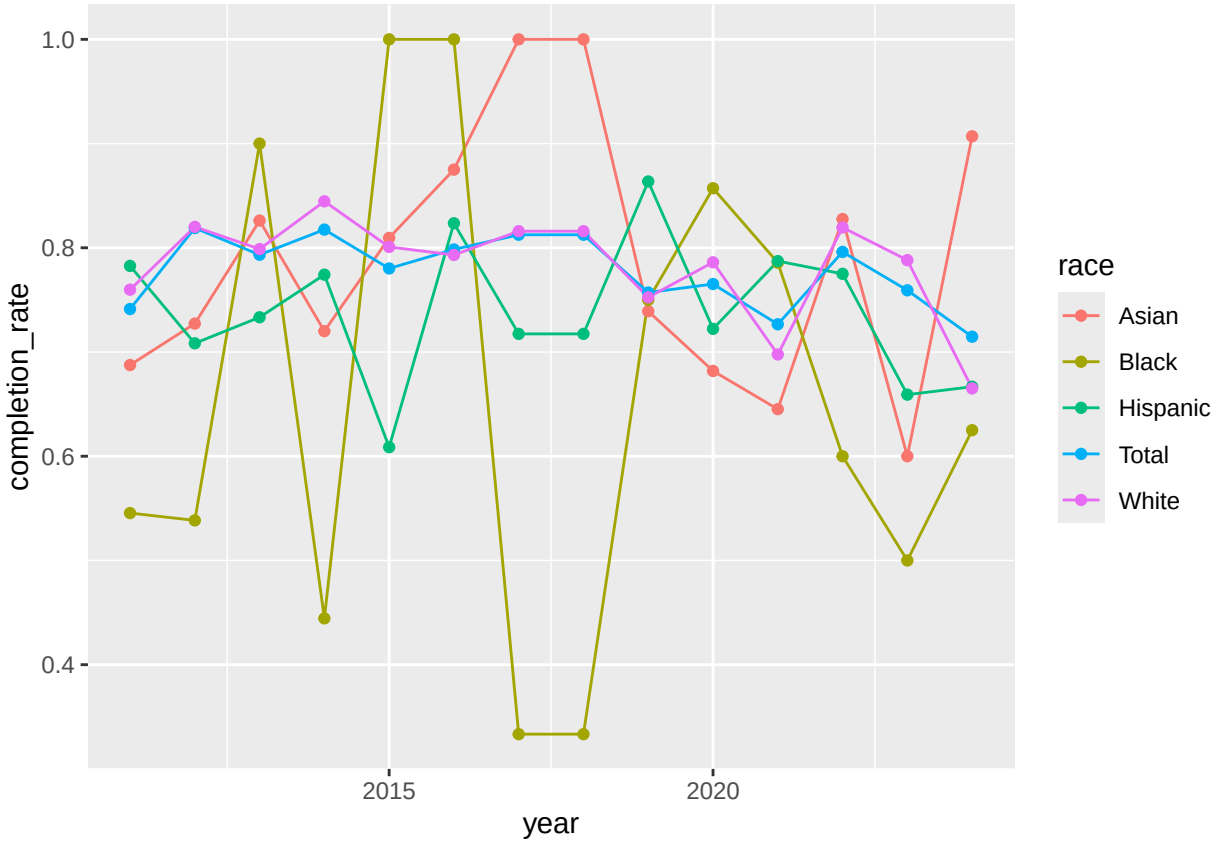
```
# plot completion rate over time
ggplot(data = reed, aes(x = year, y = completion_rate, color = sex)) +
  geom_line()
```




```
# this looks bad because "Other" contains multiple values per year
# to fix it we need to either remove "Other" or we can add race data to the graph

# we separate the plots by both sex and race by using 'facet_wrap()' to separate based on a variable
ggplot(data = reed, aes(x = year, y = completion_rate, color = sex)) +
  geom_line() +
  facet_wrap(~race)
```



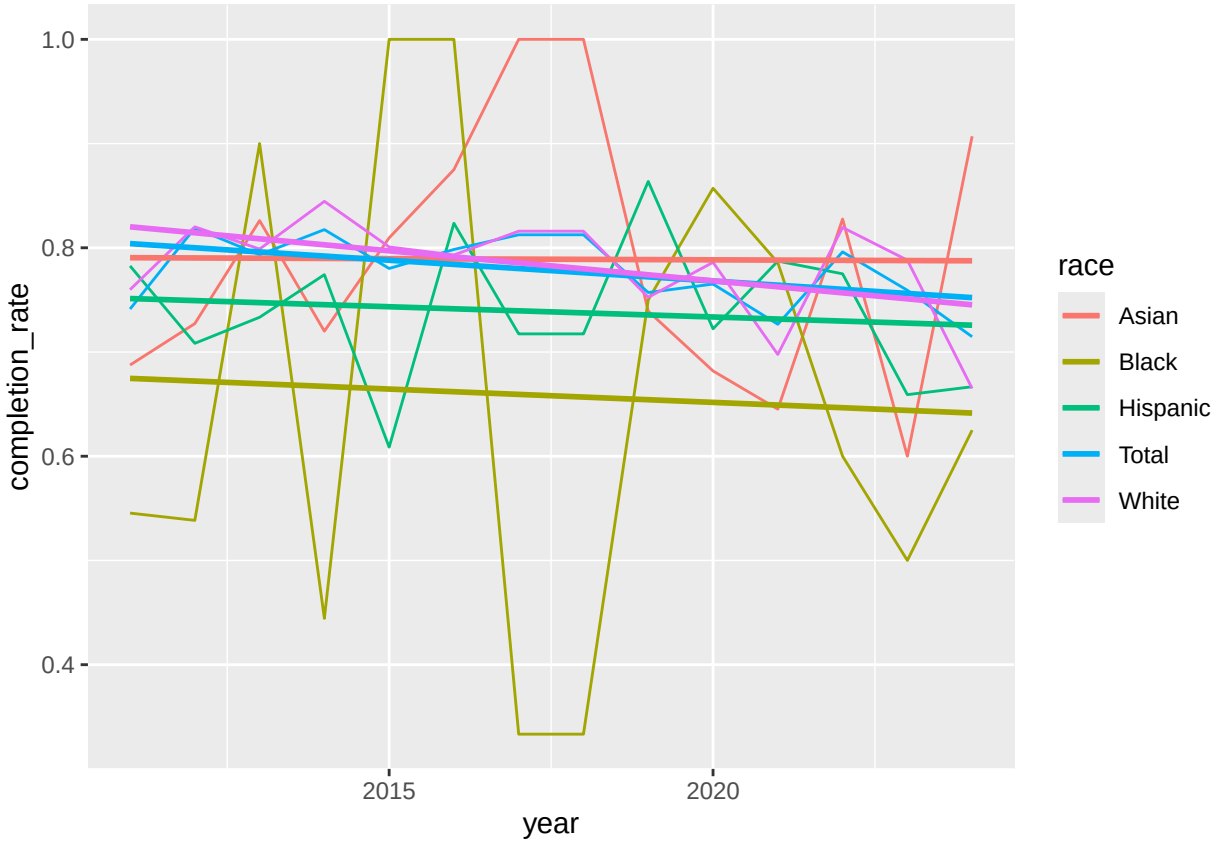

```
# yes, other still looks bad, we'd need to average it or remove it, but we'll move on for now

# we can also look at trends for races without the "Other" category
reed_completion <- reed |>
  filter(sex == "Total",
         race != "Other")

# now we can graph completion rates by year separated by race
# by adding both 'geom_line()' and 'geom_point()' we can put points within our line graph
ggplot(data = reed_completion, aes(x = year, y = completion_rate, color = race)) +
  geom_line() +
  geom_point()
```



Now we'll find a line of best fit to our graph of completion rates. Then we'll use a linear model to understand how our different variables influence completion rate. Then we'll use the `predict()` function with our linear model to see what completion rate might be in future years.

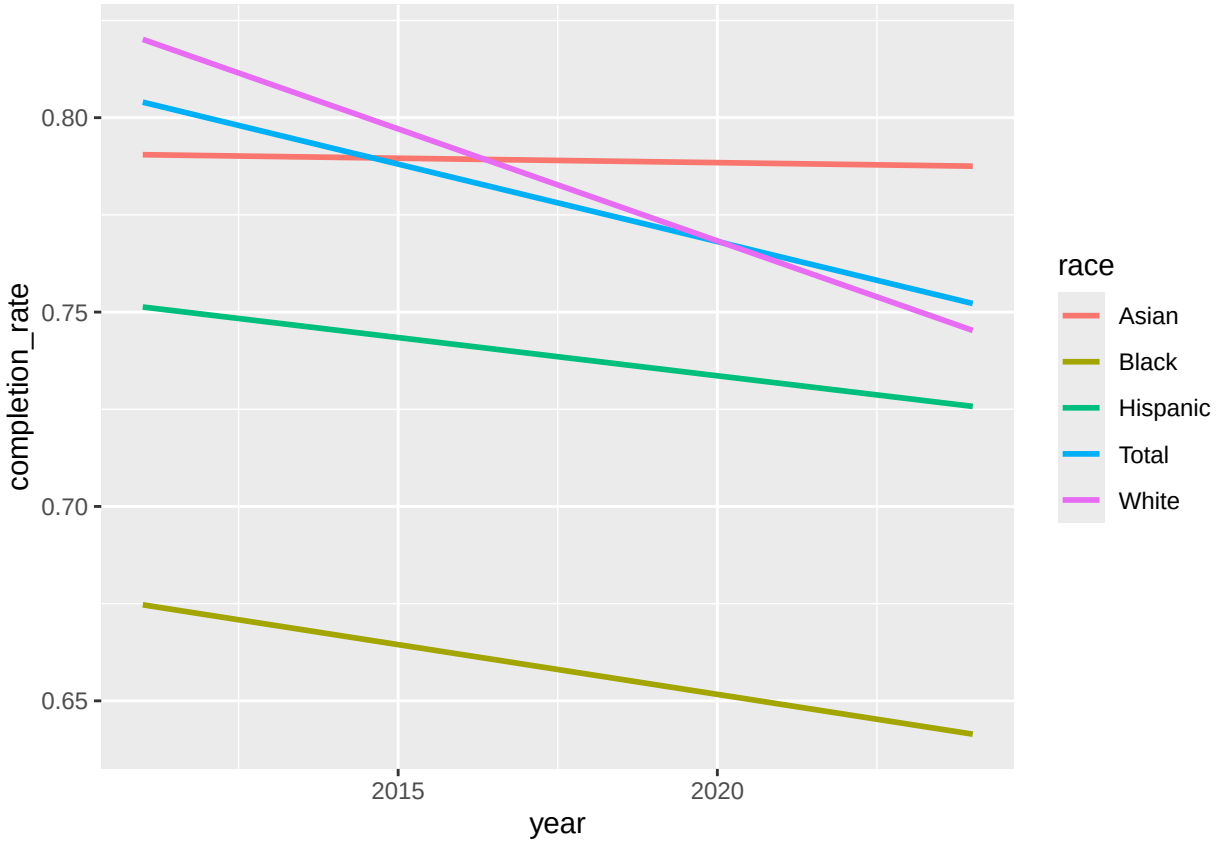
```
# add a trend line to our race completion graph with 'geom_smooth'
# method = "lm" means linear model, so a straight line
# se = FALSE means don't show the gray shade for standard error around the line, change it to TRUE to s
ggplot(data = reed_completion, aes(x = year, y = completion_rate, color = race)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```

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




```
# its' a little messy with both the actual numbers and the trend lines
# we can remove 'geom_line()' to just plot the trends
ggplot(data = reed_completion, aes(x = year, y = completion_rate, color = race)) +
  geom_smooth(method = "lm", se = FALSE)
```

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




```
# we'll make a linear model to describe the trend lines
# this model is for all data combined (not split by race)
# this creates the model
lm_model <- lm(completion_rate ~ year, data = reed)
# this shows the estimates (the y = mx + b and the R^2)
summary(lm_model)
```

```
##
## Call:
## lm(formula = completion_rate ~ year, data = reed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76846 -0.04751  0.03276  0.11360  0.30463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.132729   5.361962   3.195  0.00152 **
## year        -0.008121   0.002658  -3.056  0.00241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2037 on 367 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.02481,    Adjusted R-squared:  0.02215
## F-statistic: 9.338 on 1 and 367 DF,  p-value: 0.002409
```

```r
# this shows the same thing a little neater
tidy(lm_model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 17.1       5.36         3.20 0.00152
## 2 year        -0.00812   0.00266     -3.06 0.00241
```

```r
# before we add race and sex to our models, let's relevel them so that "Total" comes first
# (the first thing is what everything will be compared to, so the coefficients are how different each r
reed$race <- factor(reed$race, levels = c("Total", "Asian", "Black", "Hispanic", "White", "Other"))
# do the same for sex
reed$sex <- factor(reed$sex, levels = c("Total", "Male", "Female"))


# add another independent term (race) to a linear model to make it more complex
# note that this is lm_model2
lm_model2 <- lm(completion_rate ~ year + race, data = reed)
summary(lm_model2)
```

```
##
## Call:
## lm(formula = completion_rate ~ year + race, data = reed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76125 -0.05931  0.01584  0.10564  0.36539
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.135678   5.345803   3.205  0.00147 **
## year         -0.008108   0.002650  -3.060  0.00238 **
## raceAsian     0.006948   0.044321   0.157  0.87552
## raceBlack    -0.089679   0.044871  -1.999  0.04640 *
## raceHispanic -0.041490   0.044321  -0.936  0.34983
## raceWhite     0.004670   0.044321   0.105  0.91613
## raceOther    -0.036016   0.035191  -1.023  0.30679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2031 on 362 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.04395,    Adjusted R-squared:  0.02811
## F-statistic: 2.774 on 6 and 362 DF,  p-value: 0.01195
```

```r
tidy(lm_model2)
```

```
## # A tibble: 7 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  17.1       5.35         3.21  0.00147
```

```
## 2 year         -0.00811   0.00265    -3.06  0.00238
## 3 raceAsian     0.00695   0.0443      0.157 0.876
## 4 raceBlack    -0.0897    0.0449     -2.00  0.0464
## 5 raceHispanic -0.0415    0.0443     -0.936 0.350
## 6 raceWhite     0.00467   0.0443      0.105 0.916
## 7 raceOther    -0.0360    0.0352     -1.02  0.307
```

```
# add a diferent independent term (sex) to a linear model (lm_model3)
lm_model3 <- lm(completion_rate ~ year + sex, data = reed)
summary(lm_model3)
```

```
##
## Call:
## lm(formula = completion_rate ~ year + sex, data = reed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77428 -0.04884  0.03373  0.10885  0.30331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.129845   5.374916   3.187  0.00156 **
## year        -0.008123   0.002664  -3.049  0.00246 **
## sexMale      0.008342   0.026287   0.317  0.75117
## sexFemale    0.012822   0.025626   0.500  0.61715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2042 on 365 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.0255, Adjusted R-squared:  0.01749
## F-statistic: 3.184 on 3 and 365 DF,  p-value: 0.02397
```

```
tidy(lm_model3)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 17.1       5.37        3.19  0.00156
## 2 year        -0.00812   0.00266    -3.05  0.00246
## 3 sexMale      0.00834   0.0263      0.317 0.751
## 4 sexFemale    0.0128    0.0256      0.500 0.617
```

```
# add an interaction effect between the two independent variables (lm_model4)
lm_model4 <- lm(completion_rate ~ year + race * sex, data = reed)
summary(lm_model4)
```

```
##
## Call:
## lm(formula = completion_rate ~ year + race * sex, data = reed)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.76536 -0.05662  0.01034  0.10741  0.36498
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            17.158493   5.416598   3.168  0.00167 **
## year                   -0.008119   0.002685  -3.024  0.00268 **
## raceAsian               0.010907   0.077772   0.140  0.88855
## raceBlack              -0.120034   0.077772  -1.543  0.12363
## raceHispanic           -0.039574   0.077772  -0.509  0.61118
## raceWhite               0.004615   0.077772   0.059  0.95272
## raceOther              -0.046937   0.061376  -0.765  0.44494
## sexMale                -0.019393   0.077772  -0.249  0.80323
## sexFemale               0.016088   0.077772   0.207  0.83624
## raceAsian:sexMale      -0.013481   0.109986  -0.123  0.90251
## raceBlack:sexMale       0.110636   0.112253   0.986  0.32502
## raceHispanic:sexMale    0.007515   0.109986   0.068  0.94556
## raceWhite:sexMale      -0.001478   0.109986  -0.013  0.98929
## raceOther:sexMale       0.036172   0.087706   0.412  0.68028
## raceAsian:sexFemale     0.001603   0.109986   0.015  0.98838
## raceBlack:sexFemale    -0.010715   0.109986  -0.097  0.92244
## raceHispanic:sexFemale -0.013264   0.109986  -0.121  0.90408
## raceWhite:sexFemale     0.001645   0.109986   0.015  0.98808
## raceOther:sexFemale    -0.002185   0.086799  -0.025  0.97993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2058 on 350 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.05127,    Adjusted R-squared:  0.002473
## F-statistic: 1.051 on 18 and 350 DF,  p-value: 0.402
```

```
tidy(lm_model4)
```

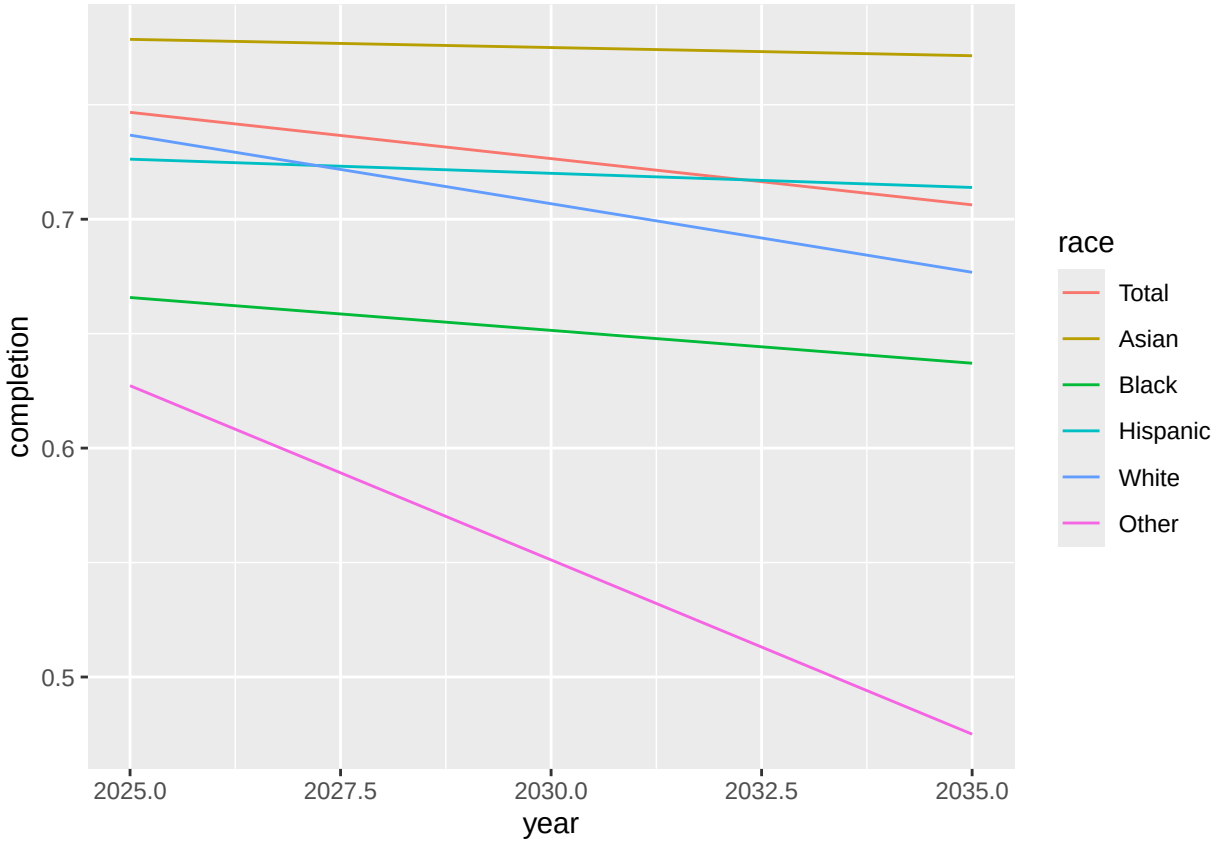
```
## # A tibble: 19 x 5
##    term                   estimate std.error statistic p.value
##    <chr>                     <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)            17.2       5.42       3.17   0.00167
##  2 year                   -0.00812   0.00268   -3.02   0.00268
##  3 raceAsian               0.0109    0.0778     0.140  0.889
##  4 raceBlack              -0.120     0.0778    -1.54   0.124
##  5 raceHispanic           -0.0396    0.0778    -0.509  0.611
##  6 raceWhite               0.00461   0.0778     0.0593 0.953
##  7 raceOther              -0.0469    0.0614    -0.765  0.445
##  8 sexMale                -0.0194    0.0778    -0.249  0.803
##  9 sexFemale               0.0161    0.0778     0.207  0.836
## 10 raceAsian:sexMale      -0.0135    0.110     -0.123  0.903
## 11 raceBlack:sexMale       0.111     0.112      0.986  0.325
## 12 raceHispanic:sexMale    0.00751   0.110      0.0683 0.946
## 13 raceWhite:sexMale      -0.00148   0.110     -0.0134 0.989
## 14 raceOther:sexMale       0.0362    0.0877     0.412  0.680
## 15 raceAsian:sexFemale     0.00160   0.110      0.0146 0.988
## 16 raceBlack:sexFemale    -0.0107    0.110     -0.0974 0.922
## 17 raceHispanic:sexFemale -0.0133    0.110     -0.121  0.904
```

```
## 18 raceWhite:sexFemale     0.00164   0.110      0.0150 0.988
## 19 raceOther:sexFemale    -0.00219   0.0868    -0.0252 0.980
```

```r
# compare all the models with Akaike Information Criterion (a measure of error)
# the lowest number represents the model with the least error
AIC(lm_model, lm_model2, lm_model3, lm_model4)
```

```
##           df        AIC
## lm_model   3 -122.97859
## lm_model2  8 -120.29296
## lm_model3  5 -119.23872
## lm_model4 20  -99.12654
```

```r
# create a new dataset of empty years and races to fill with predicted values
predictions <- expand.grid(
  year = c(2025:2035),
  race = c("Total", "Asian", "Black", "Hispanic", "White", "Other")
)

# create a prediction model that allows race to vary across year
predict_lm <- lm(completion_rate ~ year * race, data = reed)

# use the predict() function with our predict_lm to find predicted completion rates
predictions$completion <- predict(predict_lm, newdata = predictions)

# plot just the predictions
ggplot(data = predictions, aes(x = year, y = completion, color = race)) +
  geom_line()
```




```
# add predictions to our existing data by seeing what columns exist in the predicted data
# selecting just those columns from the existing data
# making sure they are named the same
# then using rbind() to bind the rows into one new dataset called all_reed
all_reed <- reed_completion |>
  select(year, race, completion_rate) |>
  rename(completion = completion_rate) |>
  rbind(predictions)

# plot all of the data together
ggplot(data = all_reed, aes(x = year, y = completion, color = race)) +
  geom_line()
```


1.0
0.8
0.6
0.4
completion
2010
2015
2020
2025
2030
2035
year
race
Asian
Black
Hispanic
Other
Total
White