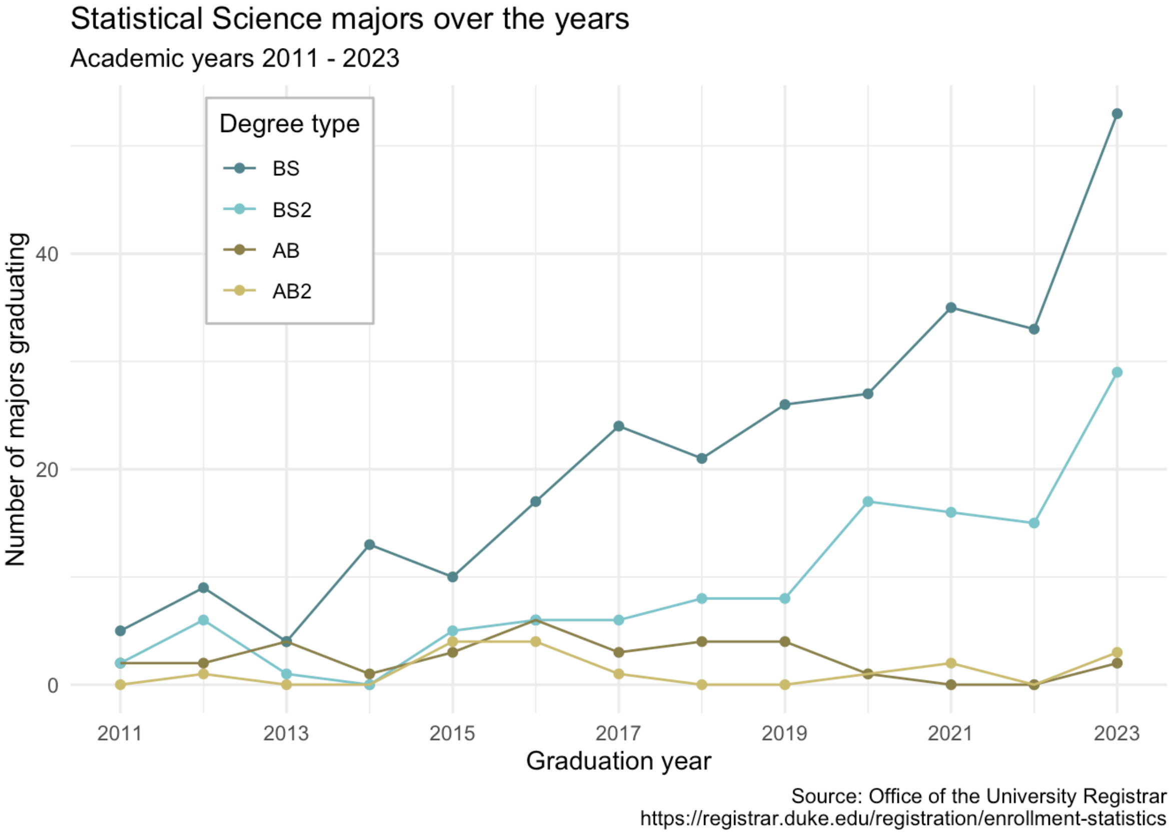
AE 05: Tidying StatSci Majors

Goal

Our ultimate goal in this application exercise is to make the following data visualization.



Data

The data come from Duke's Office of the University Registrar. The data were downloaded from Duke as a PDF file. The data have been exported to a CSV file for you. Let's load that in.

```
library(tidyverse)

statsci <- read_csv("data/statsci.csv")
```

And let's take a look at the data.

```
statsci
```

```
# A tibble: 4 x 14
  degree   `2011` `2012` `2013` `2014` `2015` `2016` `2017` `2018` `2019` `2020`
  <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Statist~     NA      1     NA     NA      4      4      1     NA     NA      1
2 Statist~      2      2      4      1      3      6      3      4      4      1
3 Statist~      2      6      1     NA      5      6      6      8      8     17
4 Statist~      5      9      4     13     10     17     24     21     26     27
# i 3 more variables: `2021` <dbl>, `2022` <dbl>, `2023` <dbl>
```

Pivoting

- **Demo:** Pivot the `statsci` data frame *longer* such that each row represents a degree type / year combination and `year` and `n`umber of graduates for that year are columns in the data frame.

```
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
)
```

```
# A tibble: 52 x 3
   degree                    year      n
   <chr>                     <chr> <dbl>
 1 Statistical Science (AB2) 2011     NA
 2 Statistical Science (AB2) 2012      1
 3 Statistical Science (AB2) 2013     NA
 4 Statistical Science (AB2) 2014     NA
 5 Statistical Science (AB2) 2015      4
 6 Statistical Science (AB2) 2016      4
 7 Statistical Science (AB2) 2017      1
 8 Statistical Science (AB2) 2018     NA
 9 Statistical Science (AB2) 2019     NA
10 Statistical Science (AB2) 2020      1
# i 42 more rows
```

- **Question:** What is the type of the `year` variable? Why? What should it be?

Add your response here.

It is a chr variable since the column headings were coded as chr variables.It should be a numeric type instead.

- **Demo:** Start over with pivoting, and this time also make sure `year` is a numerical variable in the resulting data frame.

```r
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
names_transform = as.numeric,
values_to = "n"
)
```

```
# A tibble: 52 x 3
   degree                       year     n
   <chr>                       <dbl> <dbl>
 1 Statistical Science (AB2)  2011    NA
 2 Statistical Science (AB2)  2012     1
 3 Statistical Science (AB2)  2013    NA
 4 Statistical Science (AB2)  2014    NA
 5 Statistical Science (AB2)  2015     4
 6 Statistical Science (AB2)  2016     4
 7 Statistical Science (AB2)  2017     1
 8 Statistical Science (AB2)  2018    NA
 9 Statistical Science (AB2)  2019    NA
10 Statistical Science (AB2)  2020     1
# i 42 more rows
```

- **Question:** What does an NA mean in this context? *Hint:* The data come from the university registrar, and they have records on every single graduates, there shouldn't be anything "unknown" to them about who graduated when.

Add your response here. The NA means that value is empty in the dataframe since no student graduated with that degree that year

- **Demo:** Add on to your pipeline that you started with pivoting and convert NAs in n to 0s.

```
#| label: convert-na

# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n),0,n))
```

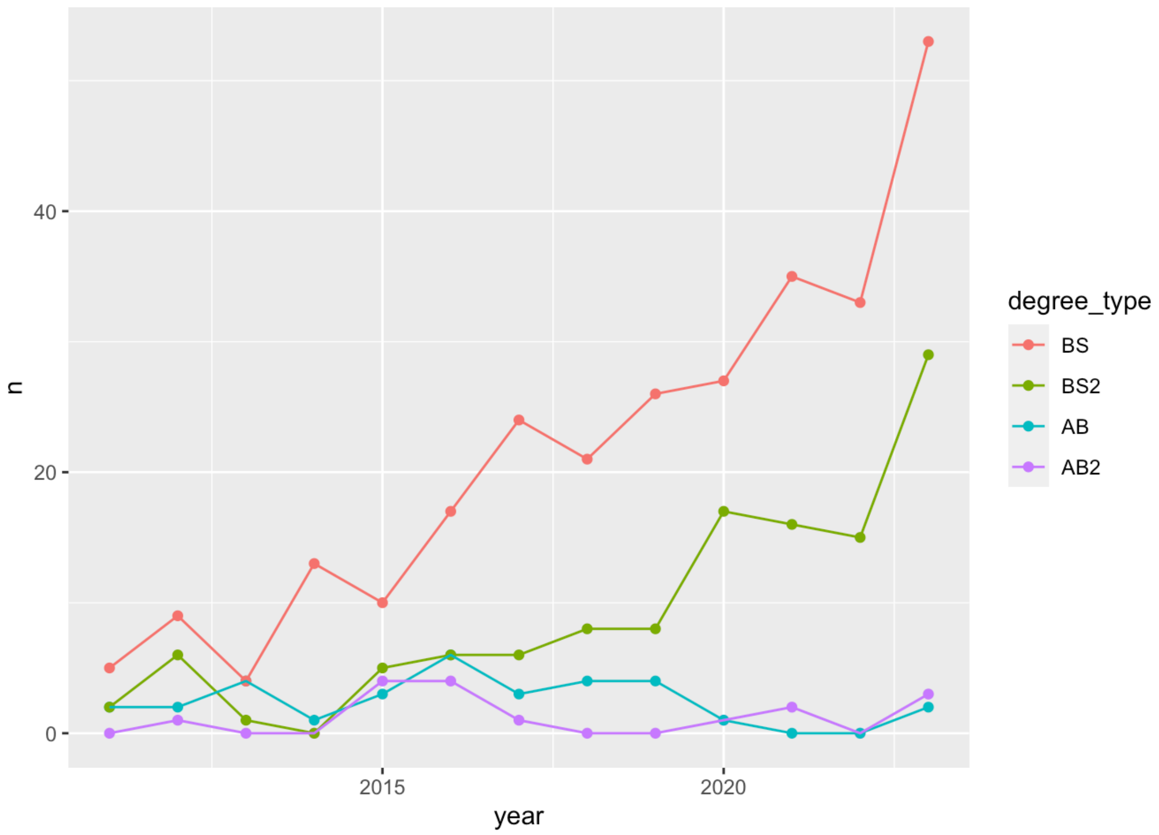
```
# A tibble: 52 x 3
   degree                       year      n
   <chr>                        <chr> <dbl>
 1 Statistical Science (AB2) 2011      0
 2 Statistical Science (AB2) 2012      1
 3 Statistical Science (AB2) 2013      0
 4 Statistical Science (AB2) 2014      0
 5 Statistical Science (AB2) 2015      4
 6 Statistical Science (AB2) 2016      4
 7 Statistical Science (AB2) 2017      1
 8 Statistical Science (AB2) 2018      0
 9 Statistical Science (AB2) 2019      0
10 Statistical Science (AB2) 2020      1
# i 42 more rows
```

- **Demo:** In our plot the degree types are BS, BS2, AB, and AB2. This information is in our dataset, in the degree column, but this column also has additional characters we don't need. Create a new column called degree_type with levels BS, BS2, AB, and AB2 (in this order) based on degree. Do this by adding on to your pipeline from earlier.

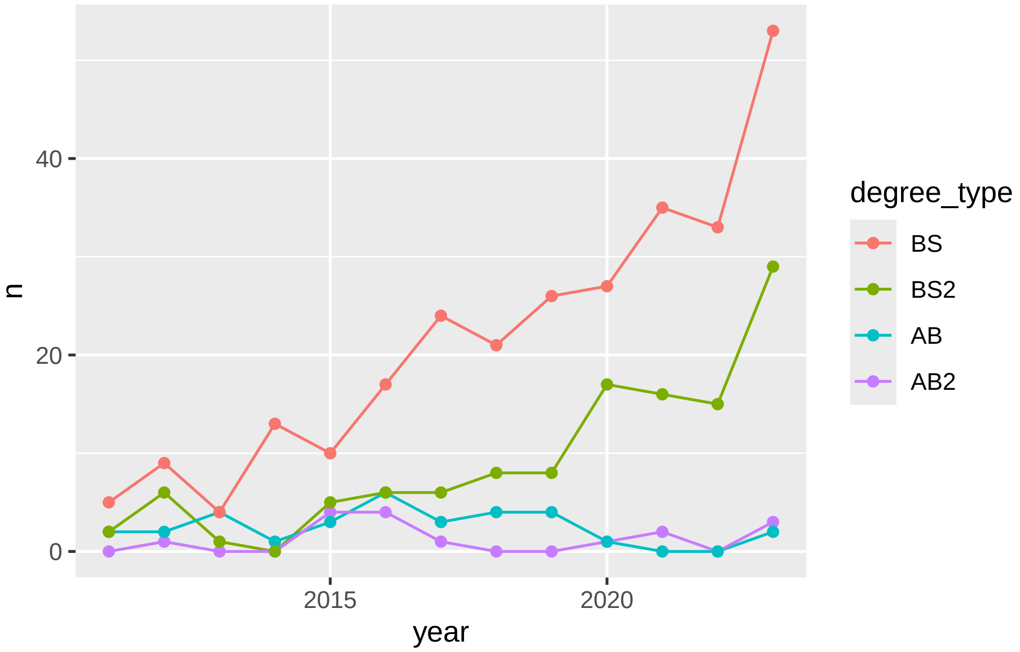
```r
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
)
```

```
# A tibble: 52 x 4
   major                   degree_type year      n
   <chr>                   <fct>       <chr> <dbl>
 1 "Statistical Science "  AB2         2011      0
 2 "Statistical Science "  AB2         2012      1
 3 "Statistical Science "  AB2         2013      0
 4 "Statistical Science "  AB2         2014      0
 5 "Statistical Science "  AB2         2015      4
 6 "Statistical Science "  AB2         2016      4
 7 "Statistical Science "  AB2         2017      1
 8 "Statistical Science "  AB2         2018      0
 9 "Statistical Science "  AB2         2019      0
10 "Statistical Science "  AB2         2020      1
# i 42 more rows
```

- **Your turn:** Now we start making our plot, but let's not get too fancy right away. Create the following plot, which will serve as the "first draft" on the way to our Goal. Do this by adding on to your pipeline from earlier.



```
# add your code here.
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
) %>%
ggplot(mapping = aes(x = year, y = n, color = degree_type))+
  geom_line(aes(group=degree_type))+
  geom_point()+
  scale_x_discrete(breaks=c(2015,2020))
```

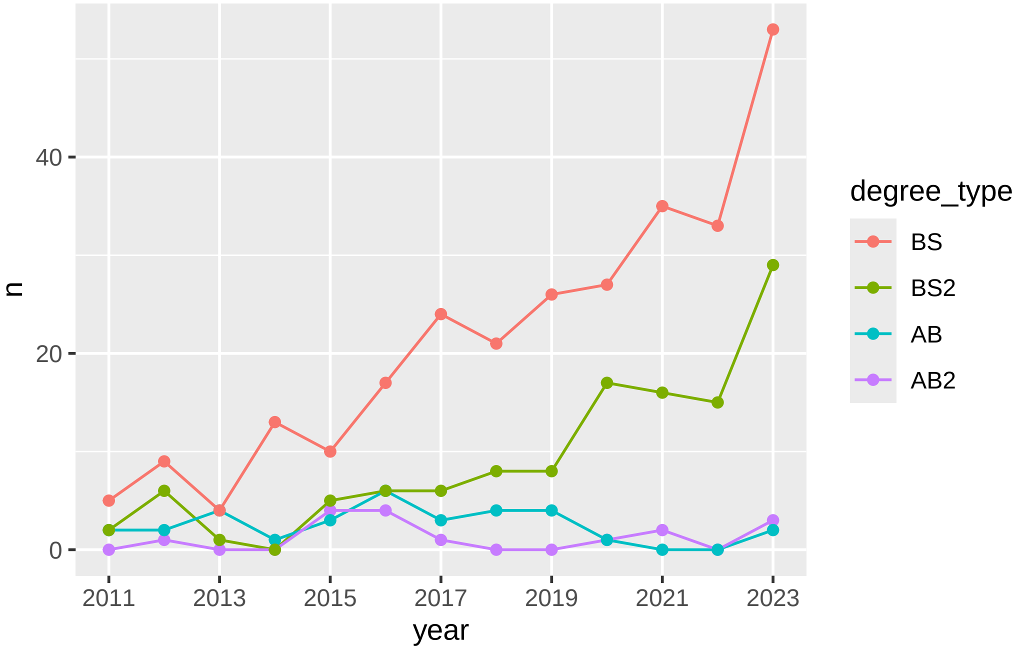


- **Your turn:** What aspects of the plot need to be updated to go from the draft you created above to the Goal plot at the beginning of this application exercise.

Add your response here. We need to change the theme, axis labels, the x axis ticks, the line colors,position of the legend, add a title and subtitle and add the source at the bottom

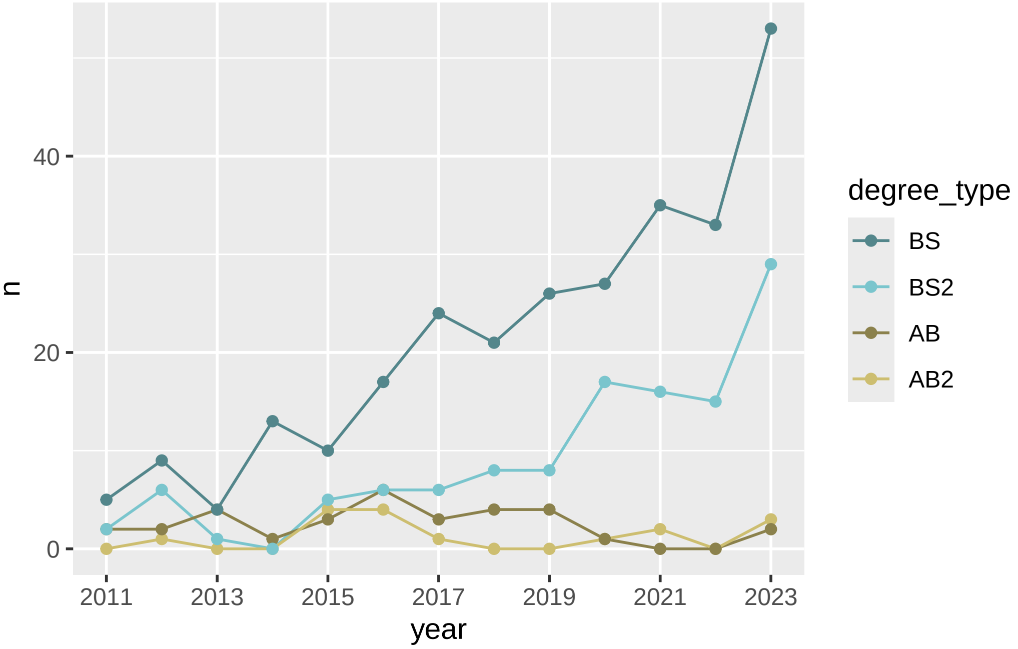
- **Demo:** Update x-axis scale such that the years displayed go from 2011 to 2023 in increments of 2 years. Do this by adding on to your pipeline from earlier.

```
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
) %>%
ggplot(mapping = aes(x = year, y = n, color = degree_type))+
  geom_line(aes(group=degree_type))+
  geom_point()+
  scale_x_discrete(breaks=seq(2011,2023,2))
```



- **Demo:** Update line colors using the following level / color assignments. Once again, do this by adding on to your pipeline from earlier.
    - "BS" = "cadetblue4"
    - "BS2" = "cadetblue3"
    - "AB" = "lightgoldenrod4"
    - "AB2" = "lightgoldenrod3"

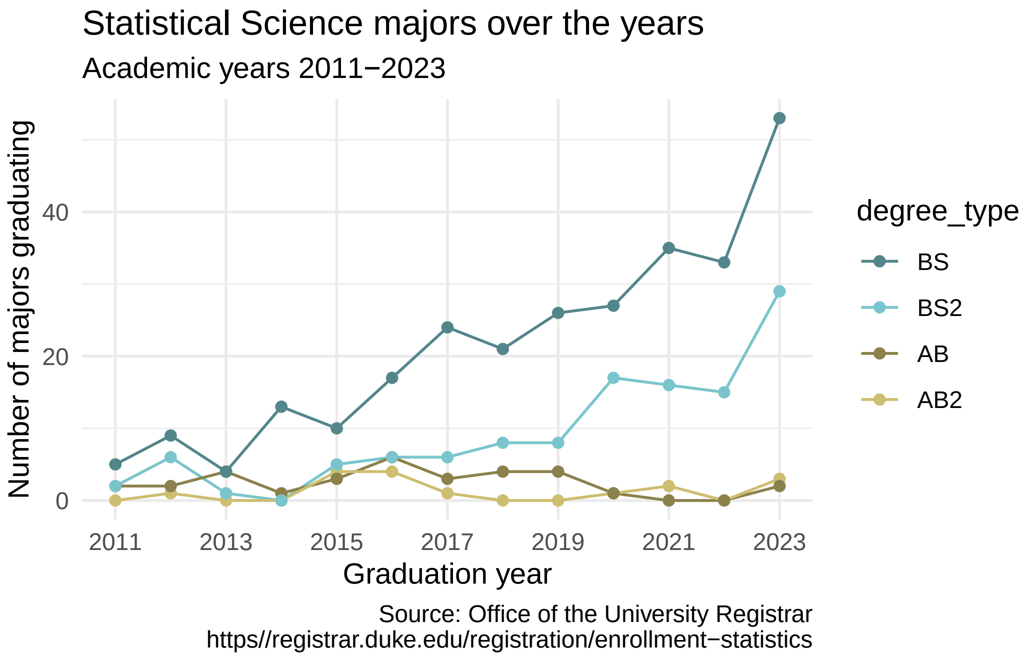
```
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
) %>%
ggplot(mapping = aes(x = year, y = n, color = degree_type))+
  geom_line(aes(group=degree_type))+
  geom_point()+
  scale_x_discrete(breaks=seq(2011,2023,2))+
  scale_color_manual(
    values = c(
      "BS" = "cadetblue4",
      "BS2" = "cadetblue3",
      "AB" = "lightgoldenrod4",
      "AB2" = "lightgoldenrod3"
    )
  )
```



- **Your turn:** Update the plot labels (`title`, `subtitle`, `x`, `y`, and `caption`) and use `theme_minimal()`. Once again, do this by adding on to your pipeline from earlier.

```
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
```

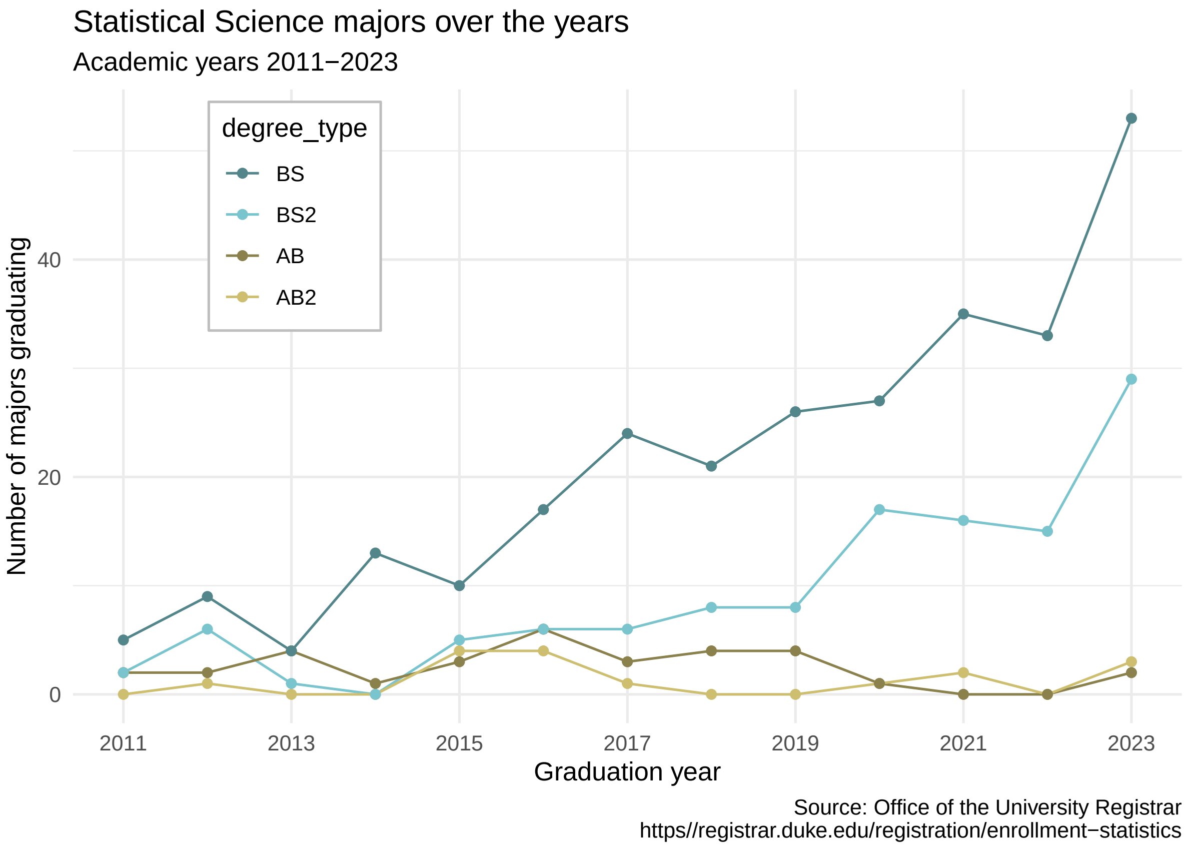
```r
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
) %>%
ggplot(mapping = aes(x = year, y = n, color = degree_type))+
  geom_line(aes(group=degree_type))+
  geom_point()+
  scale_x_discrete(breaks=seq(2011,2023,2))+
  scale_color_manual(
    values = c(
      "BS" = "cadetblue4",
      "BS2" = "cadetblue3",
      "AB" = "lightgoldenrod4",
      "AB2" = "lightgoldenrod3"
    )
  )+
  labs(
    title = "Statistical Science majors over the years",
    subtitle = "Academic years 2011-2023",
    x = "Graduation year",
    y = "Number of majors graduating",
    caption ="Source: Office of the University Registrar\n https//registrar.duke.edu/registra
  )+
  theme_minimal()
```



- **Demo:** Finally, adding to your pipeline you've developed so far, move the legend into the plot, make its background white, and its border gray. Set `fig-width: 7` and `fig-height: 5` for your plot in the chunk options.

```
# add your code here
statsci %>%
pivot_longer(
cols = -degree,
names_to = "year",
values_to = "n"
) %>%
mutate(n = if_else(is.na(n), 0, n)) %>%
separate_wider_delim(degree, delim = "(", names= c("major","degree_type")) %>%
mutate(
degree_type = str_remove(degree_type, "\\)"),
degree_type = fct_relevel(degree_type, "BS",
"BS2", "AB", "AB2")
) %>%
ggplot(mapping = aes(x = year, y = n, color = degree_type))+
  geom_line(aes(group=degree_type))+
  geom_point()+
  scale_x_discrete(breaks=seq(2011,2023,2))+
  scale_color_manual(
    values = c(
      "BS" = "cadetblue4",
      "BS2" = "cadetblue3",
      "AB" = "lightgoldenrod4",
      "AB2" = "lightgoldenrod3"
    )
  )+
  labs(
    title = "Statistical Science majors over the years",
    subtitle = "Academic years 2011-2023",
    x = "Graduation year",
    y = "Number of majors graduating",
    caption ="Source: Office of the University Registrar\n https//registrar.duke.edu/registra
  )+
  theme_minimal()+
  theme(
    legend.position = c(0.2, 0.8),
    legend.background = element_rect(fill = "white", color = "gray")
  )
```

```
Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2
3.5.0.
i Please use the `legend.position.inside` argument of `theme()` instead.
```


Statistical Science majors over the years

Academic years 2011−2023

Source: Office of the University Registrar
https//registrar.duke.edu/registration/enrollment−statistics