

Week3 Presentation

John Hulse

Week 3

- joining datasets.
- Going from basic plots to more complex and attractive plots.

Data

For our first part of class, we're going to work with the Educational Attainment Data from Eurostat. I've also created a dataset that has full names for the countries and some basic information about them.

```
1 EUEduc <- get_eurostat_data("edat_lfse_03")  
2 ESNames <- read_csv('ESNames.csv', col_types = "ffffnf") # You may n
```

EUEduc

```
1 head(EUEduc)
```

Key: <sex, age, unit, isced11, geo>

	sex	age	unit	isc11	geo	time	values
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<num>
1:	F	Y15-64	PC	ED0-2	AT	1995	41.4
2:	F	Y15-64	PC	ED0-2	AT	1996	39.6
3:	F	Y15-64	PC	ED0-2	AT	1997	36.6
4:	F	Y15-64	PC	ED0-2	AT	1998	36.0
5:	F	Y15-64	PC	ED0-2	AT	1999	35.3
6:	F	Y15-64	PC	ED0-2	AT	2004	28.4

ESNames

```
1 head(ESNames)
```

```
# A tibble: 6 × 6
  CODE      countryname Membership    EA20 AccYr ColdWar
  <fct>      <fct>      <fct>      <fct> <dbl> <fct>
1 EU27_2020 European Union Group      No      NA <NA>
2 BE        Belgium    EUMember_2020 Yes    1957 NATO
3 BG        Bulgaria    EUMember_2020 No     2007 WP
4 CZ        Czechia      EUMember_2020 No     2004 WP
5 DK        Denmark      EUMember_2020 No     1973 NATO
6 DE        Germany      EUMember_2020 Yes    1957 NATO
```

Joining the datasets

```
1 EUEduc <- EUEduc %>% # specifies that we are storing the result of
2   left_join(ESNames, by=c("geo"="CODE"))# left join keeps all of th
3
4 head(EUEduc)
```

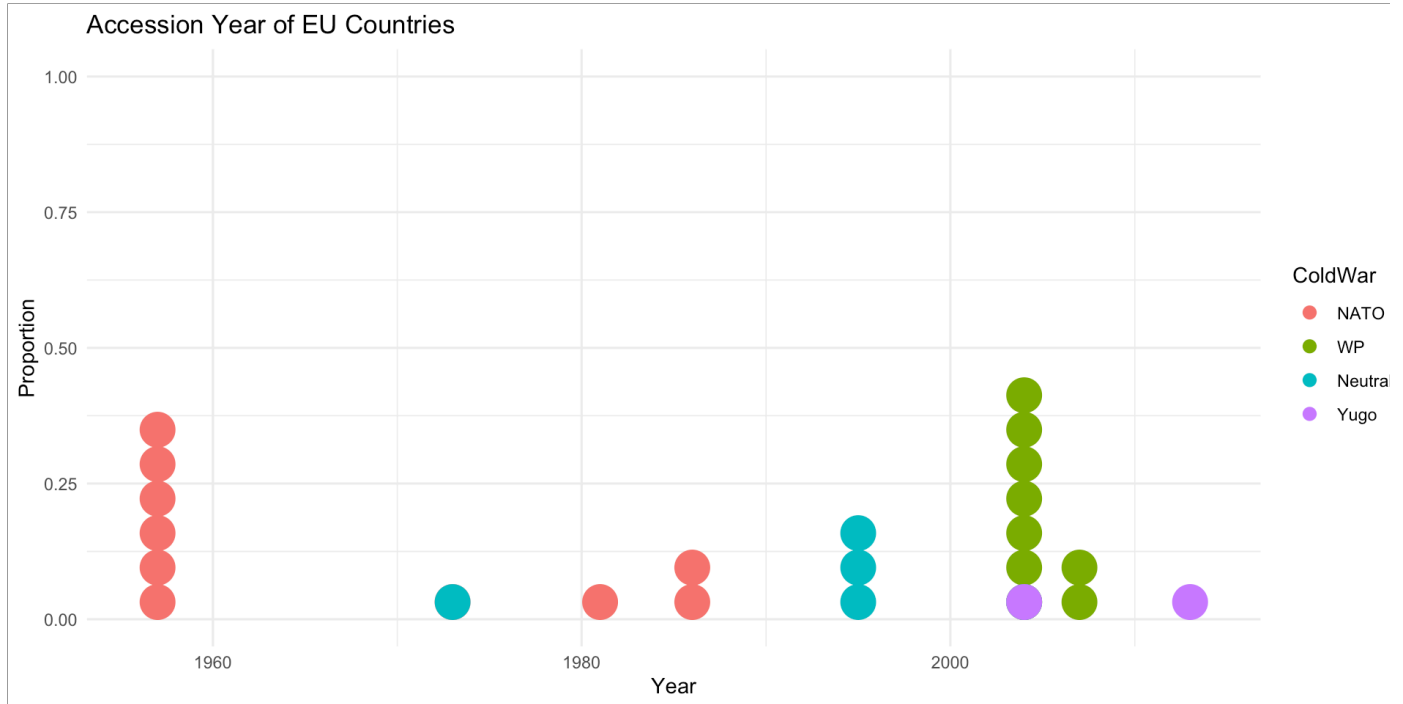
Key: <sex, age, unit, isced11, geo>

	sex	age	unit	iscd11	geo	time	values	countryname
Membership								
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<num>	<fctr>
<fctr>								
1:	F	Y15-64	PC	ED0-2	AT	1995	41.4	Austria
EUMember_2020								
2:	F	Y15-64	PC	ED0-2	AT	1996	39.6	Austria
EUMember_2020								
3:	F	Y15-64	PC	ED0-2	AT	1997	36.6	Austria
EUMember_2020								
4:	F	Y15-64	PC	ED0-2	AT	1998	36.0	Austria
EUMember_2020								
5:	F	Y15-64	PC	ED0-2	AT	1999	35.3	Austria
EUMember_2020								
6:	F	Y15-64	PC	ED0-2	AT	2004	28.4	Austria

It puts the same information from the ESNames dataset for every instance of the country code in the EUEduc dataset.

Dotplot of Accession Year

```
1 ESNames %>%  
2   ggplot(aes(x = AccYr)) +  
3   geom_dotplot(aes(color=ColdWar, fill=ColdWar)) +  
4   labs(title = "Accession Year of EU Countries",  
5         y = "Proportion",  
6         x = "Year")+  
7   theme_minimal()
```



Grammar of Graphics: ggplot2

ggplot2 is a package that implements the Grammar of Graphics. This is a way of thinking about plots that is based on the idea that a plot is a combination of data, aesthetics, and geoms. So, every plot is built out of a layered series of commands that control what is displayed and how.

First, the data

One of the main things we've practiced so far is using the pipe operator to pass data from one function to another, so that we can only pass the data that we need to the plotting function. Here we're only passing the Total population, ages 25-64, for three different levels of education, in four countries plus the EU as a whole, in 2023.

```
1 EUeduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"| isced11=="ED3_4")
```

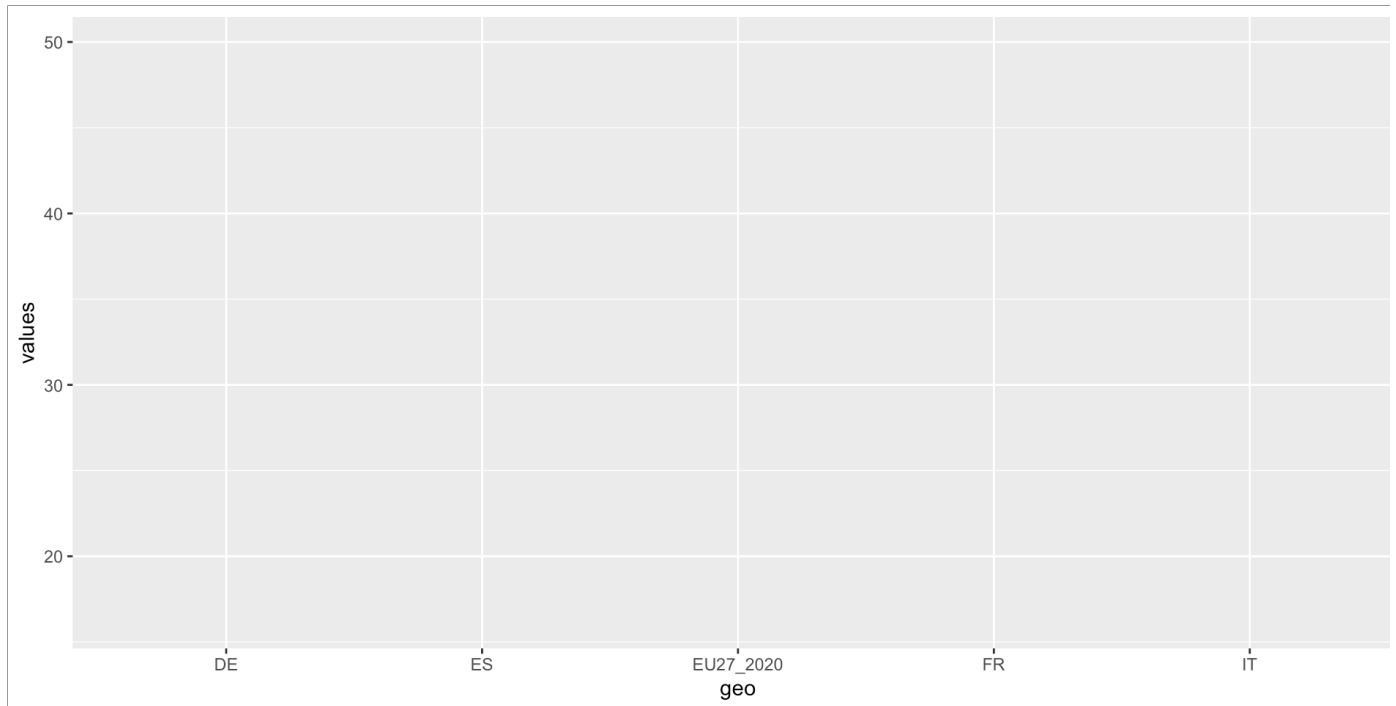
Key: <sex, age, unit, isced11, geo>

	sex	age	unit	isced11	geo	time	values
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<num>
1:	T	Y25-64	PC	ED0-2	DE	2023	16.9
Germany							
2:	T	Y25-64	PC	ED0-2	ES	2023	35.8
Spain							
3:	T	Y25-64	PC	ED0-2	EU27_2020	2023	20.2 European
Union							
4:	T	Y25-64	PC	ED0-2	FR	2023	16.3
France							
5:	T	Y25-64	PC	ED0-2	IT	2023	34.5
Italy							
6:	T	Y25-64	PC	ED3_4	DE	2023	49.8

Aesthetics

In defining the aesthetics, we're identifying the variables that will be on the x and y axis, as well as the variables that will define the fill colors for the bars. However, we haven't yet specified what kind of plot we want to make.

```
1 EUEduc %>%  
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8" | isced11=="ED3_4")  
3   ggplot(aes(x=geo, y=values, fill=iscd11))
```

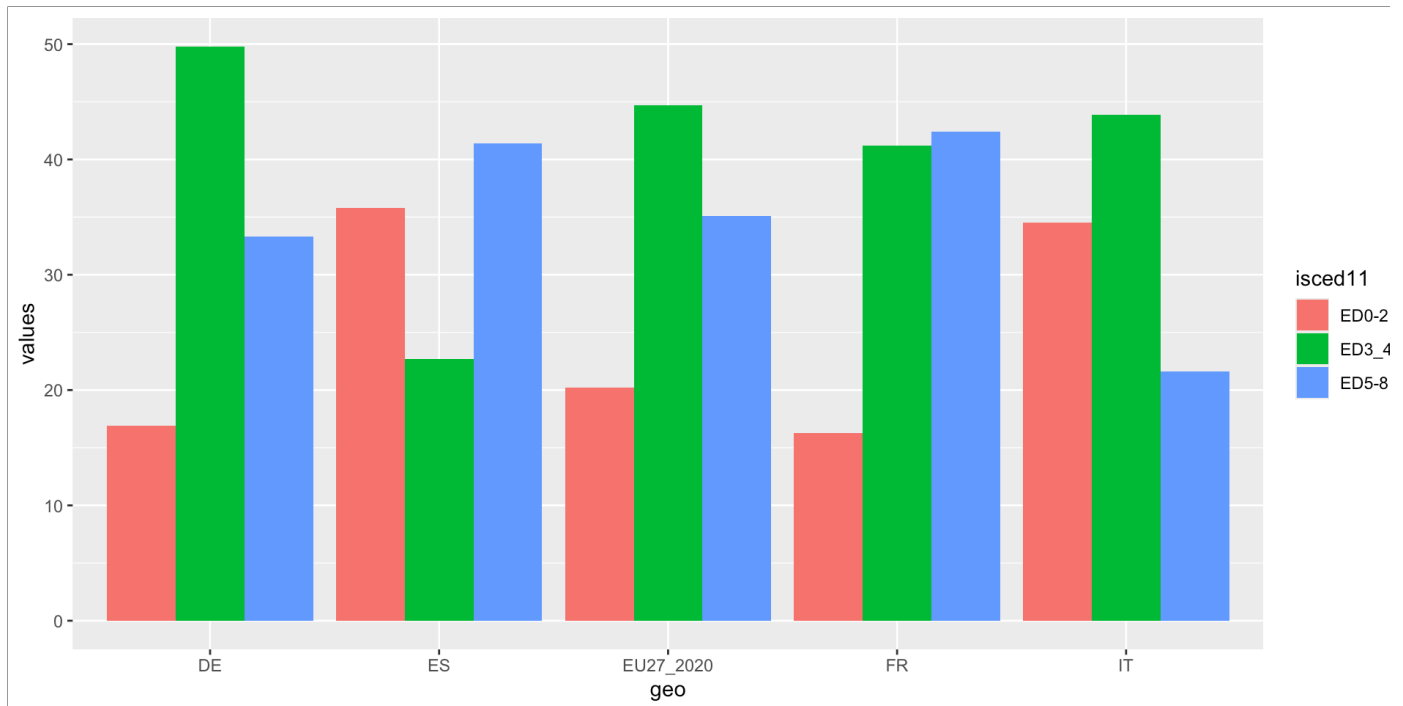


Adding Geoms

By adding `geom_bar`, we're specifying that we want to make a bar plot. The `stat="identity"` argument specifies that we want the height of the bars to be determined by the values in the dataset, rather than by some other transformation. The `position="dodge"` argument specifies that we want the bars to be displayed side by side, rather than stacked.

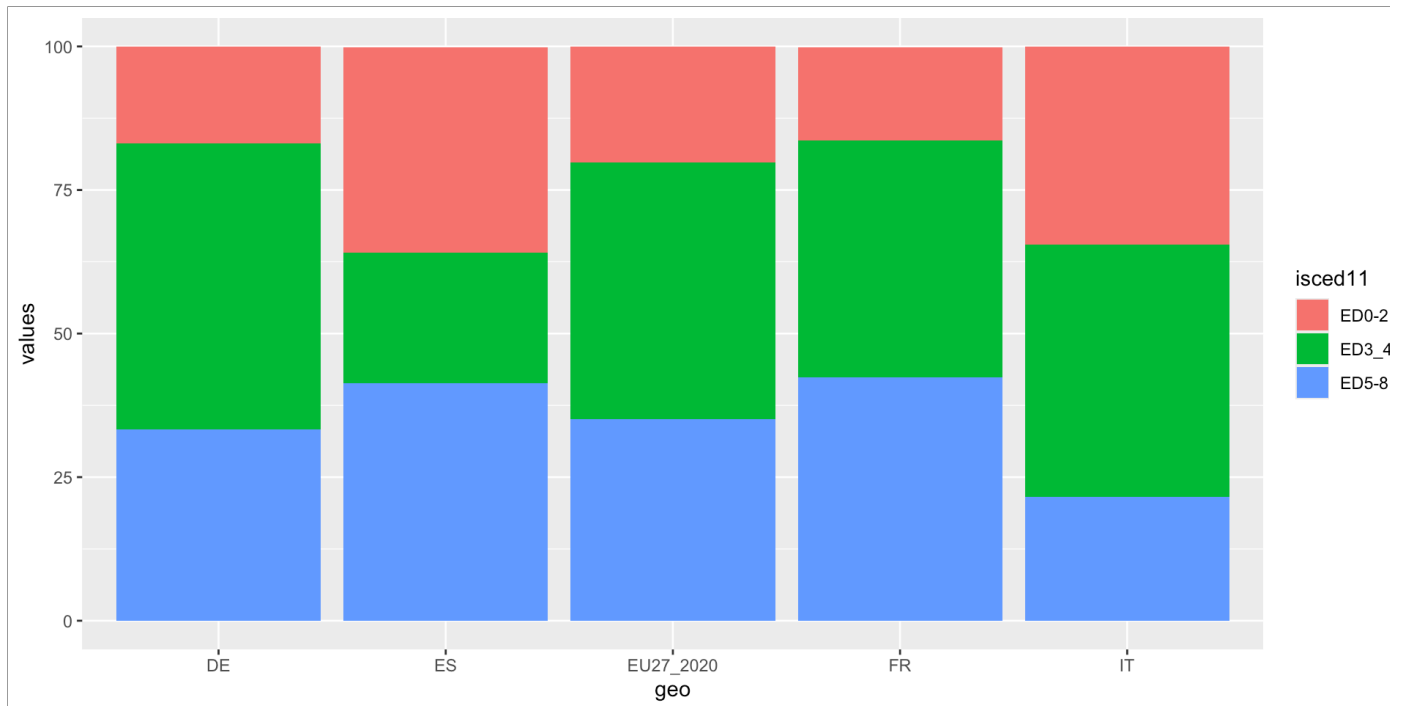
Dodged bar

```
1 EUEduc %>%  
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"| isced11=="ED3_4")  
3   ggplot(aes(x=geo, y=values, fill=iscd11)) +  
4   geom_bar(stat="identity", position="dodge")
```

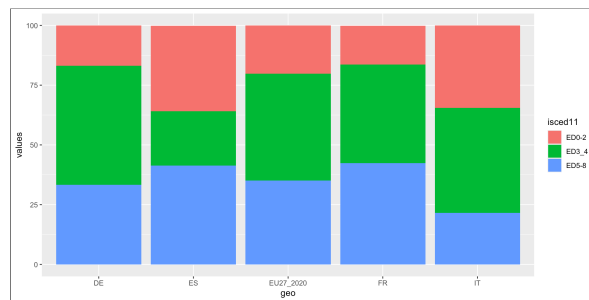
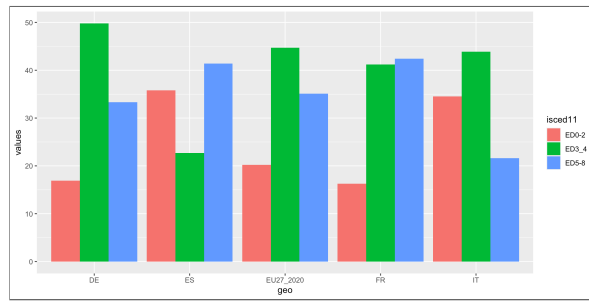


Stacked bar

```
1 EUEduc %>%  
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|iscd11=="ED3_4"  
3     ggplot(aes(x=geo, y=values, fill=iscd11)) +  
4     geom_bar(stat="identity", position="stack")
```



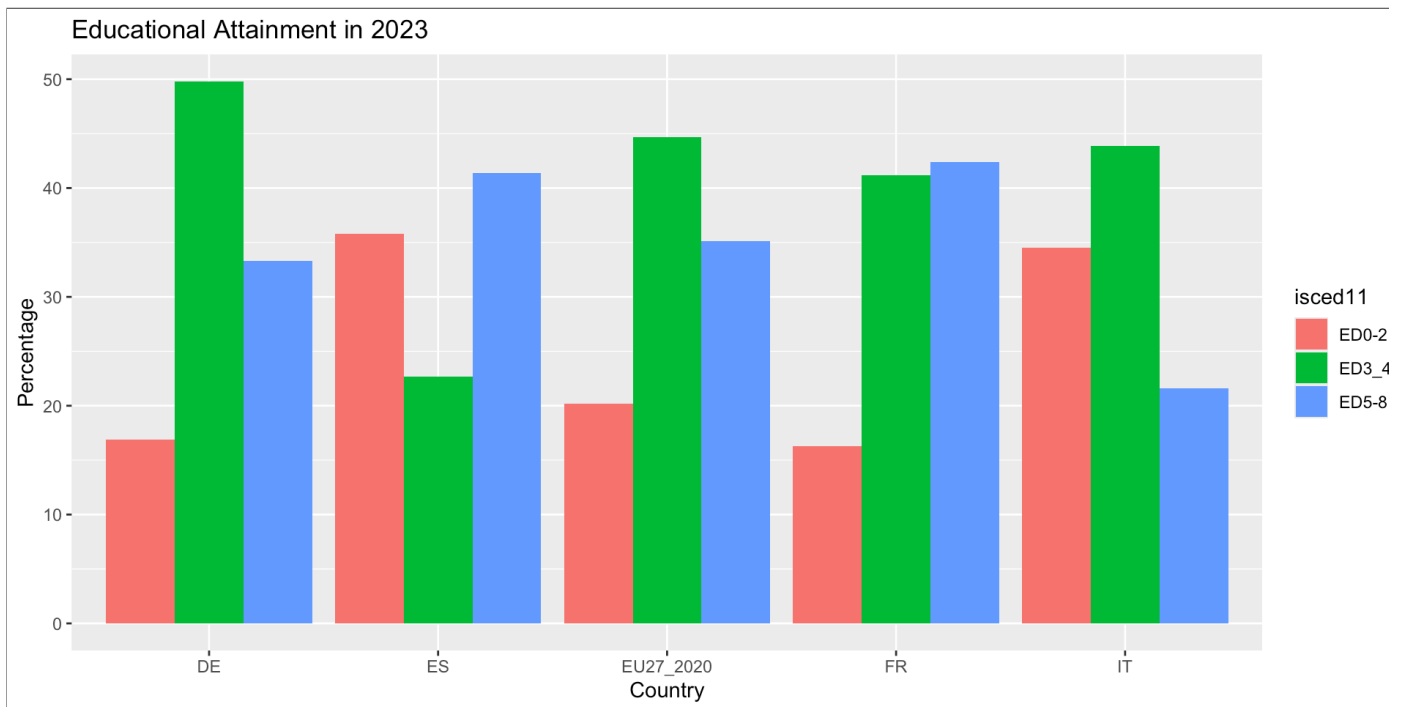
Which is better?



Labels

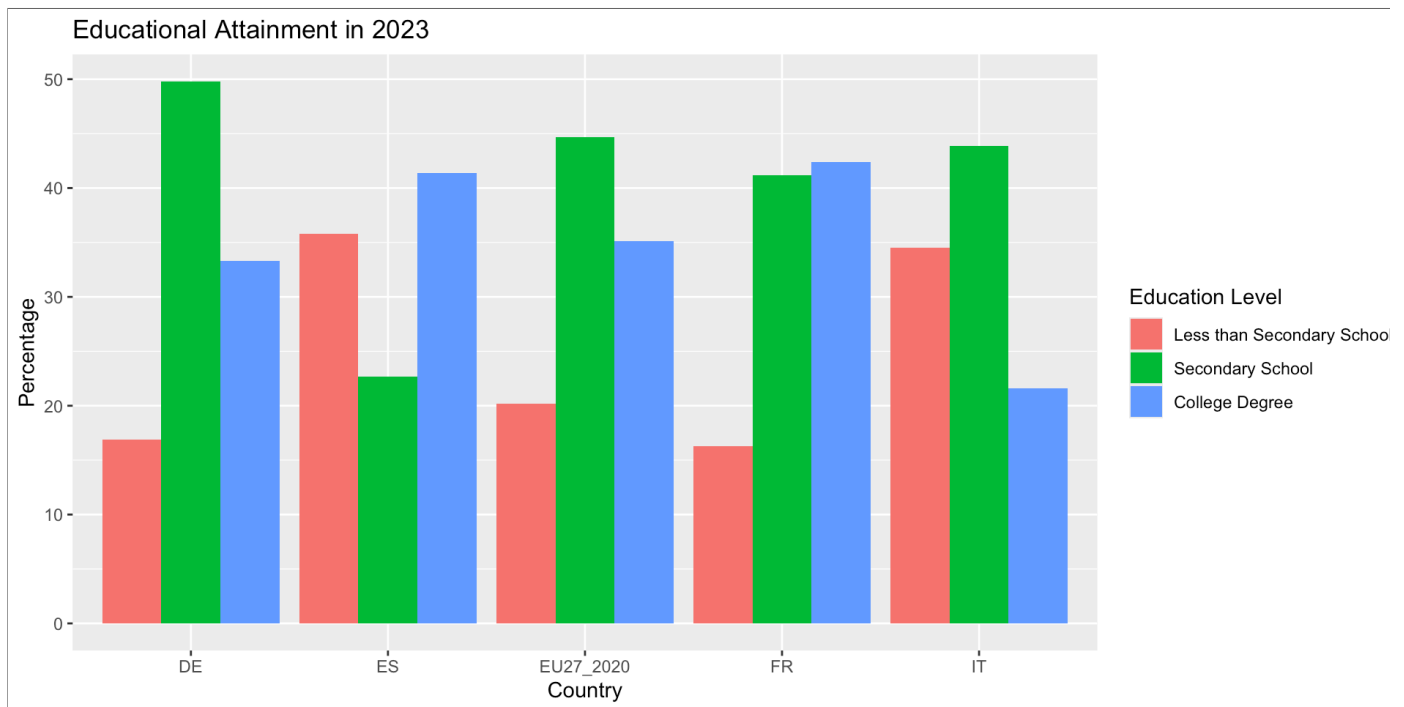
The labs function allows us to add labels to the plot. Here we're adding a title, and labels for the x and y axes.

```
1 EUEduc %>%  
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"| isced11=="ED3_4")  
3   ggplot(aes(x=geo, y=values, fill=iscd11)) +  
4   geom_bar(stat="identity", position="dodge") +  
5   labs(title="Educational Attainment in 2023",  
6        x="Country",  
7        y="Percentage")
```



Changing the legend

```
1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|iscd11=="ED3_4"
3   mutate(iscd11 = recode(iscd11, "ED5-8"="College Degree", "ED3_4"
4   ggplot(aes(x=geo, y=values, fill=iscd11)) +
5   geom_bar(stat="identity", position="dodge") +
6   labs(title="Educational Attainment in 2023",
7        x="Country",
8        y="Percentage")+
9   guides(fill=guide_legend(title="Education Level"))
```

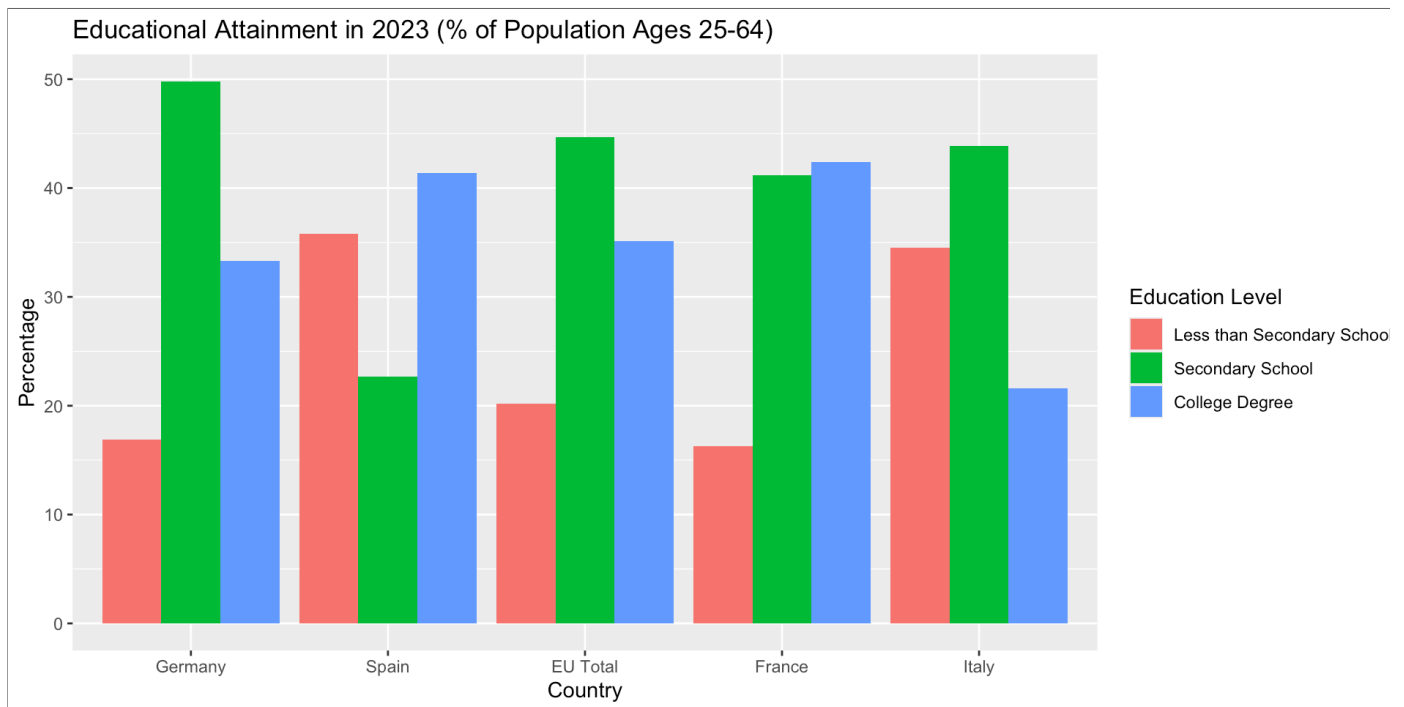


Better country names

```

1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|iscd11=="ED3_4"
3   mutate(iscd11 = recode(iscd11, "ED5-8"="College Degree", "ED3_4"
4   ggplot(aes(x=geo, y=values, fill=iscd11)) +
5   geom_bar(stat="identity", position="dodge") +
6   labs(title="Educational Attainment in 2023 (% of Population Age
7     x="Country",
8     y="Percentage")+
9   guides(fill=guide_legend(title="Education Level"))

```

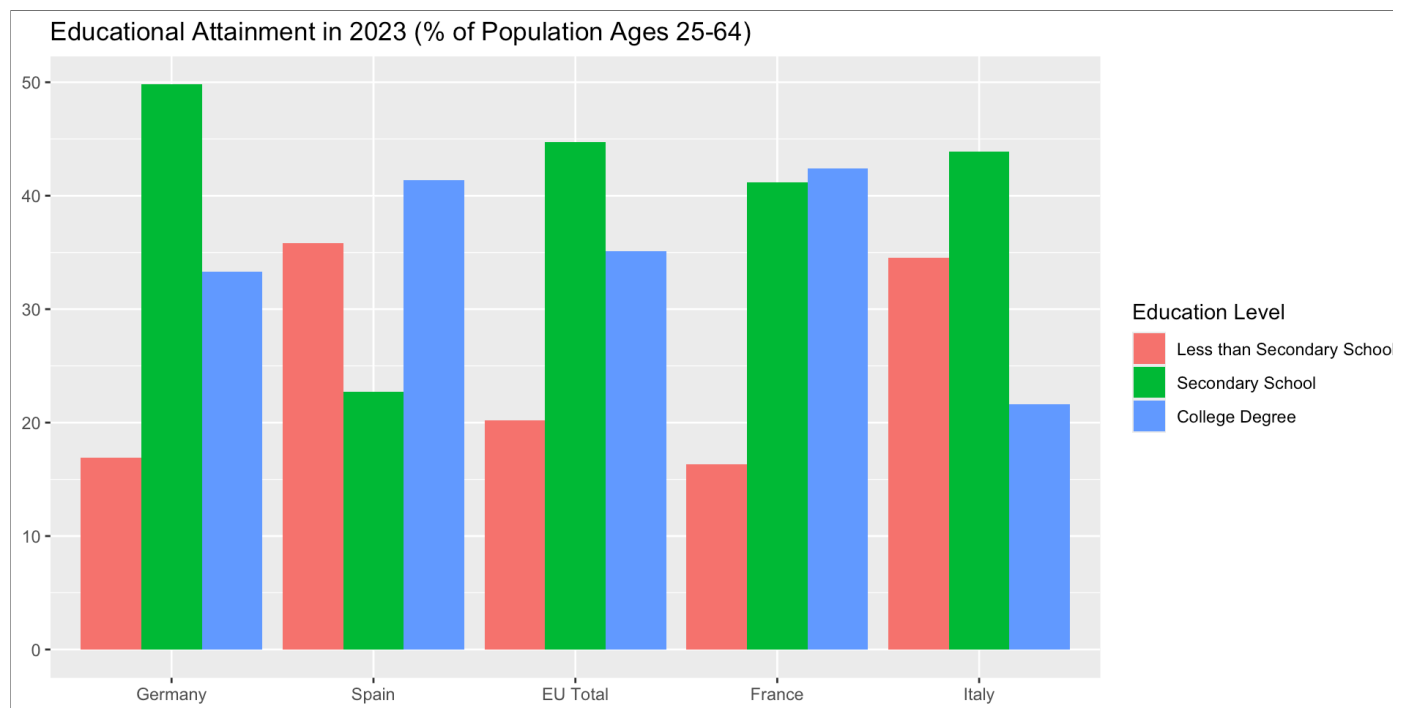


Simplifying labels

```

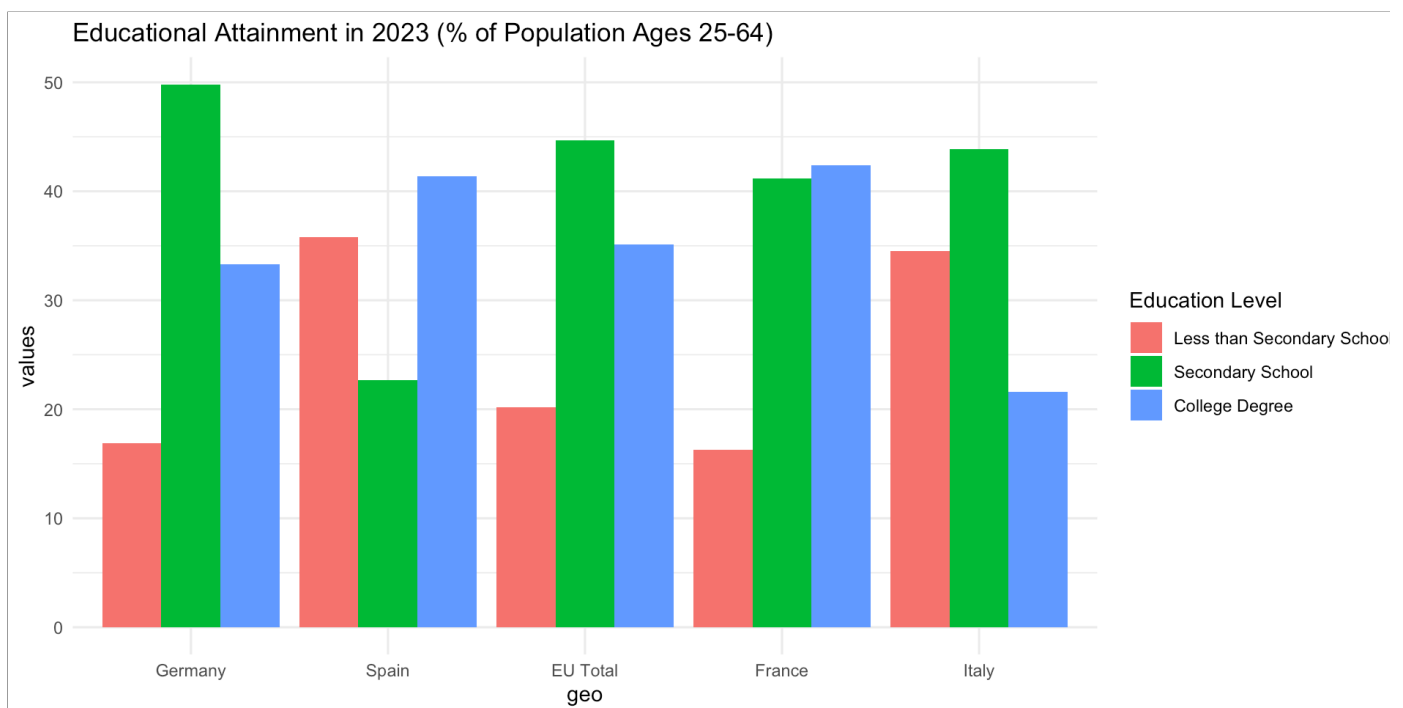
1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"| isced11=="ED3_4")
3   mutate(isced11 = recode(isced11, "ED5-8"="College Degree", "ED3_4"
4     ggplot(aes(x=geo, y=values, fill=isc11)) +
5     geom_bar(stat="identity", position="dodge") +
6     labs(title="Educational Attainment in 2023 (% of Population Age
7     guides(fill=guide_legend(title="Education Level"))+
8     theme(axis.title.x=element_blank(),
9           axis.title.y=element_blank())
10    )

```



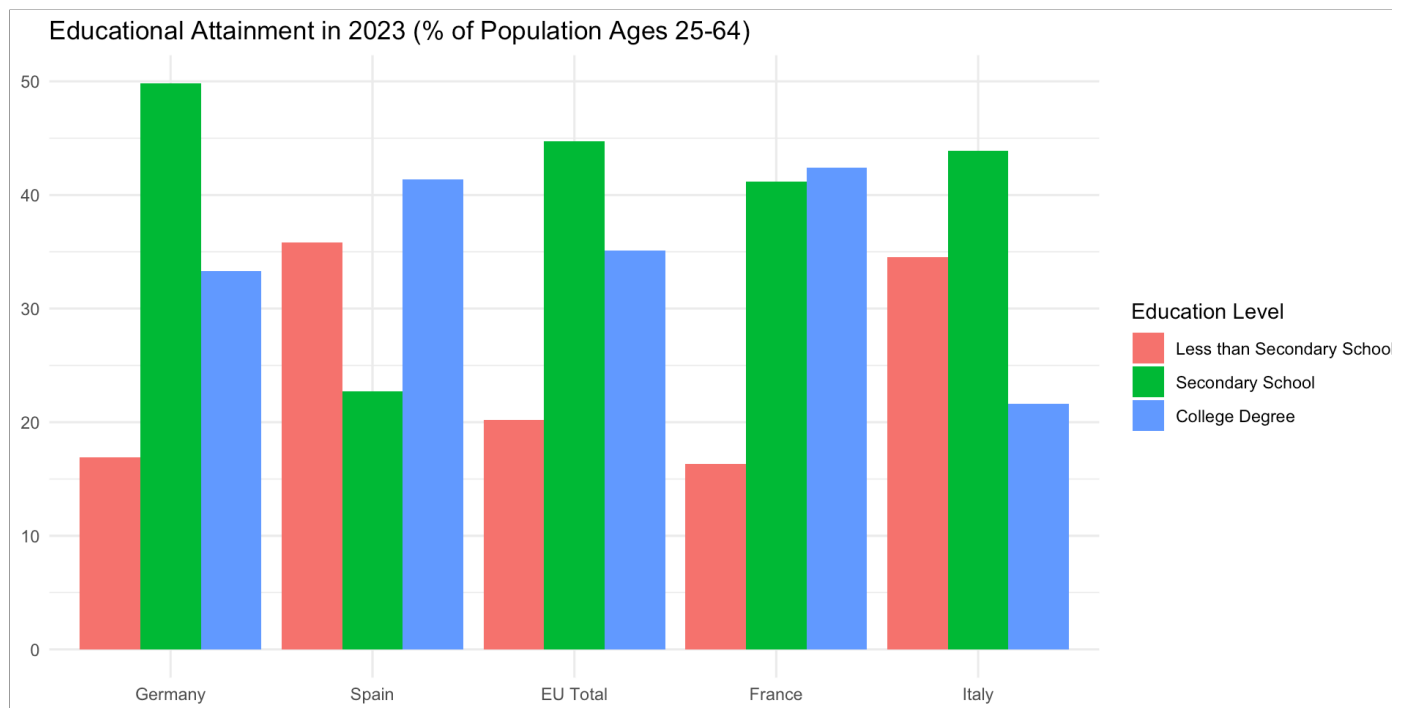
That background?

```
1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|iscd11=="ED3_4"
3   mutate(iscd11 = recode(iscd11, "ED5-8"="College Degree", "ED3_4"
4   ggplot(aes(x=geo, y=values, fill=iscd11)) +
5   geom_bar(stat="identity", position="dodge") +
6   labs(title="Educational Attainment in 2023 (% of Population Age
7   guides(fill=guide_legend(title="Education Level"))+
8   theme(axis.title.x=element_blank(),
9         axis.title.y=element_blank(),
10        )+
11   theme_minimal()
```



That background?

```
1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|isced11=="ED3_4"
3   mutate(isced11 = recode(isced11, "ED5-8"="College Degree", "ED3_4"
4   ggplot(aes(x=geo, y=values, fill=isced11)) +
5   geom_bar(stat="identity", position="dodge") +
6   labs(title="Educational Attainment in 2023 (% of Population Age
7   guides(fill=guide_legend(title="Education Level"))+
8   theme_minimal()+
9   theme(axis.title.x=element_blank(),
10         axis.title.y=element_blank(),
11         )
```



More countries?

```

1 EUEduc %>%
2   filter(sex=="T", age=="Y25-64", isced11=="ED5-8"|iscd11=="ED3_4"
3   mutate(iscd11 = recode(iscd11, "ED5-8"="College Degree", "ED3_4
4   ggplot(aes(x=iscd11, y=values, fill=iscd11)) +
5   facet_wrap(~geo)+
6   geom_bar(stat="identity") +
7   labs(title="Educational Attainment in 2023 (% of Population Age
8   guides(fill=guide_legend(title="Education Level"))+
9   theme_minimal()+
10  theme(axis.title.x=element_blank(),
11         axis.title.y=element_blank(),
12         axis.text.x=element_blank(),
13         )

```





Plot design

Using ggplot2 you can control all aspects of the plot design, but the options can be overwhelming. To deal with this, use a design thinking process.

1. Identify the main comparisons you need in order to address the question.
2. Brainstorm different chart types that could be used to show these comparisons.
3. Sketch out the features you think work best.
4. Make a simple plot that shows the main comparisons.
5. Iterate your design to incrementally improve it.

Group Activity

We're going to break up into groups. Each group will be assigned a different question. The group will work together to produce the best chart they can, including all of the components we've discussed (and maybe more!).

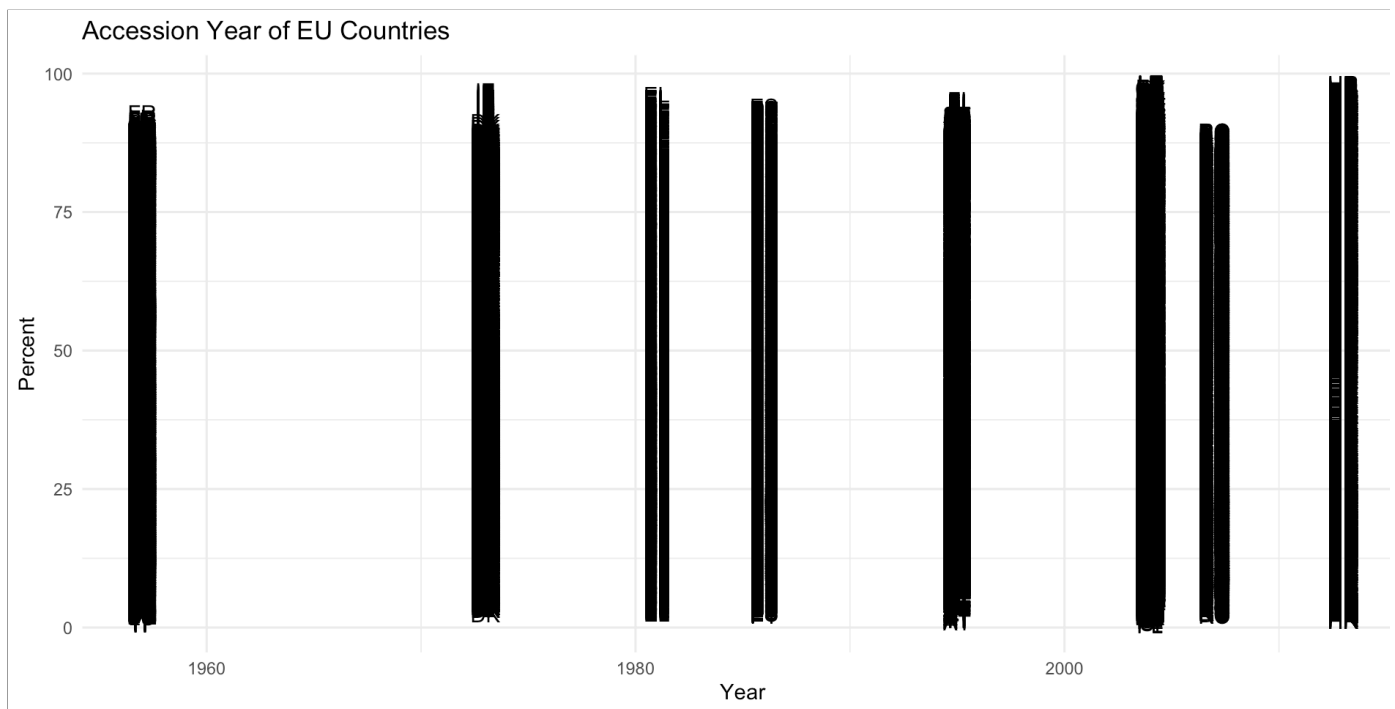
The groups will send their finished code back to me and present the chart to the class.

Group 1 Annika, Lillie, Adrienne, Olivia

Do the countries that joined the EU earlier have more or fewer people with college degrees than those that joined later? (Hint:

https://ggplot2.tidyverse.org/reference/geom_text.html)

```
1 EUEduc %>%
2   filter() %>%
3   ggplot(aes(x = AccYr, y=values, label = geo)) +
4   geom_text()+
5   labs(title = "Accession Year of EU Countries",
6         y = "Percent",
7         x = "Year")+
8   theme_minimal()
```



Group 2 George, Camryn, Rachel, Sophia

Has the percentage of women attaining college degrees surpassed the percentage of men in all member states of the EU?

Group 3 Cade, William, Emma, Sattie

How do the candidate countries for EU membership compare to the EU member states in terms of the percentage of the population with college degrees? (Hint: the Membership variable says which countries are candidates.)

Group 4 Lauren, Zaria, Carter, Jackie (sp?)

Are older people more likely to have a college degree than younger people and does the pattern vary by country? (Hint: use the decade levels of the age variable to compare the percentage with a college degree for 25-34, 35-44, 45-54, and 55-64 year olds.)

Group 5 Michael, Miles, Harrison, Prema

Do countries with a higher percentage of women with a college degree also have a higher percentage of men with a college degree? (Hint: https://ggplot2.tidyverse.org/reference/geom_text.html)